35

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

UN PANORAMA DEL ANALISIS DE CONGLOMERADOS

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I A

P R E S E N T A:

MIROSLAVA GODINEZ TREJO

DIRECTOR DE TESIS:

MAT. MARGARITA CHAVEZ CANO

FACULTAD DE CIENCIAS
UNAM

MEXICO, D.F. 2001

VNIVERSIDAD NACIONAL
AVFNOMA DI
MEXICO

M. EN C. ELENA DE OTEYZA DE OTEYZA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"Un Panorama del Análisis de Conglomerados"

realizado por Miroslava Godínez Trejo

con número de cuenta 9354843-8 , pasante de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Mat. Margarita Chávez Cano
Propietario	M. en D. María Teresa Velázquez Uribe
Propietario	M. en D. Alejandro Mina Valdés
Suplente	Dra. María Edith Pacheco Gómez Muñoz
Suplente	Act. Jaime Vázquez Alamilla

Consejo Departamental de Matemáticas

M. en C. José Antonio Flores Díaz

MATEMATICAS

Gracias :

A Dios por permitirme terminar este trabajo y ser mi guía día a día.

A mi mamá y a mi tía Soledad (q.e.p.d) por su cariño y apoyo incondicional siempre.

A Mat. Margarita Chávez Cano por su gran paciencia, comprensión, por estar siempre dispuesta a ayudarme, y darme parte de su valioso tiempo.

A M. en D. Alejandro Mina Valdés, Act. Jaime Vázquez Alamilla, M. en D. María Teresa Velázquez Uribe y Dra. Edith Pacheco Gómez Muñoz, por sus aportaciones y por el tiempo dedicado en la revisión de esta tesis.

A mis amigos y compañeros por sus palabras alentadoras.

ÍNDICE DE TABLAS Y GRÁFICAS

TABLAS

GRÁFICAS

ÍNDICE

CAPÍTULO 3
MÉTODOS JERÁRQUICOS Y NO JERÁRQUICOS

CAPÍTULO 4
APLICACIÓN

INTRODUCCIÓN

Este trabajo tiene como objetivo principal proporcionar un panorama sobre la teoría de conglomerados en cuanto a su metodología de agrupamiento y sus medidas de proximidad, así como mostrar algunas técnicas gráficas y métodos que puedan ayudar al análisis exploratorio de los datos que muchas veces es omitido porque no se conoce o porque no se les da la importancia que tiene.

Dentro del Análisis de Conglomerados se maneja el concepto de clasificación, lo primero que viene a la mente con esta palabra es agrupar objetos de acuerdo a un criterio que definamos. Un ejemplo muy simple de clasificación, es cuando a las personas se les puede agrupar de acuerdo a su nivel económico en diferentes grupos, como en clase baja, clase media y clase alta, o pueden ser clasificados respecto al consumo anual de alcohol: en bajo, medio y alto. Un esquema de clasificación puede representar simplemente un método conveniente para organizar un conjunto de datos para que el manejo de la información sea más eficiente. El punto importante que sugieren estos ejemplos es que la clasificación es una división de objetos en grupos basados en un conjunto de reglas.

El hombre siempre ha sido capaz de reconocer que muchos objetos comparten ciertas propiedades, en un sentido más amplio, la clasificación es necesaria para el desarrollo del lenguaje que consiste en palabras que ayudan a reconocer los diferentes tipos de eventos, objetos y personas que nos encontramos a nuestro paso, cada nombre en un lenguaje, por ejemplo, es una etiqueta que describe una clase de cosas las cuales tienen características en común, así los animales son llamados perros, gatos, vacas, estos nombres reúnen individuos en grupos. La clasificación además de ser una actividad humana, es fundamental en las ramas de las ciencias. En Biología, la clasificación de organismos ha sido una gran preocupación desde las primeras investigaciones biológicas. Aristóteles construyó un sistema para clasificar a las especies del reino animal dividiéndolos en vertebrados e invertebrados. La clasificación de los elementos en la tabla periódica producida por Mendeleyev en los 1860's, ha tenido un profundo impacto en el entendimiento de la estructura del átomo. La clasificación de estrellas en enanas y gigantes basadas en la gráfica de temperatura contra luminosidad de Herstprung-Russell ha afectado fuertemente las teorías de la evolución estelar.

El concepto de clasificación es fundamental en una técnica multivariada llamada Análisis de Conglomerados, de manera formal el propósito que tiene esta técnica es agrupar un conjunto de objetos (individuos, puntos, unidades), tomando en cuenta las características (variables, atributos, medidas) que poseen; a cada conjunto se le denomina conglomerado y cada objeto que hay dentro del conglomerado tiene un alto grado de asociación natural, es decir los elementos son muy similares entre sí de acuerdo a sus características. Los conglomerados resultantes deben mostrar homogeneidad dentro de ellos y heterogeneidad entre ellos. Por lo que cuando los conglomerados se representan gráficamente los objetos dentro de ellos estarán muy próximos entre sí y los distintos grupos alejados. Por otro lado, hay un criterio predeterminado de selección para decidir que objeto entra a que conglomerado.

En nuestros días el Análisis de Conglomerados es aplicado en muchas áreas: en las Ciencias de la Vida, como Biología, Botánica, Zoología, Ecología y Paleontología; en las Ciencias Sociales y del Comportamiento, como Psicología, Sociología, Criminología, Antropología, Lingüística y Arqueología; en las Ciencias de la Tierra, como Geología, Geografía, Estudios regionales y Ciencias del suelo; en la Medicina, como Psiquiatría, Citología y Diagnostico clínico; en la Ingeniería, como Inteligencia Artificial, Reconocimiento de patrones y Sistemas; en las Ciencias Políticas y de la Información, como Investigación de Operaciones, Ciencias Políticas, Economía, Investigación de Mercados y Recuperación de la Información. Debido a la diversidad de aplicaciones, el Análisis de Conglomerados tiene una variedad de nombres, se le denomina también análisis Q (en Psicología), construcción de tipología, análisis de clasificación, clasificación automática, análisis tipológico y taxonomía númerica (en Biología) entre otros.

A continuación se citan algunos ejemplos donde se especifica el uso del Análisis de Conglomerados que le han dado algunos investigadores:

En psiquiatría se puede utilizar para definir categorías de diagnóstico; Pilowsky et al (1969) agruparon 200 pacientes basándose en respuestas sobre la depresión, y conjuntando esta información con datos sobre su estado mental, sexo, longitud de su enfermedad y edad; de esta manera uno de los conglomerados resultantes estaba conformado con pacientes con depresión endógena. Otro de los usos en ésta rama fué encontrar una clasificación de individuos que habían intentado suicidarse; Paykel y Rassaby (1978), estudiaron 236 intentos de suicidios que llegaban a los principales servicios de emergencias, se basaron en 14 variables que fueron consideradas las más relevantes como la edad, el número de intentos de suicidios anteriores, la severidad de la depresión y hostilidad, y un número de características demográficas; se aplicaron algunas técnicas del Análisis de Conglomerados a los datos y se obtuvieron tres grupos con las siguientes características: dentro del primer grupo estaban los individuos que tomaban sobredosis, en general demostraban menor probabilidad a vivir, tenían menor número de trastornos psiquiátricos y una mayor evidencia a la motivación interpersonal que a la destrucción a sí mismos. En el segundo grupo se encontraban los individuos con intentos de suicidio más severos, con una mayor motivación de destrucción y el uso de métodos de violencia más que la sobredosis; y en el tercer grupo estaban los individuos que tenían una historia previa de intentos de suicidio, además de que su intento de suicidio más reciente había sido relativamente apacible y tenían un comportamiento demasiado hostil.

En Medicina, Wastell y Gray (1987) utilizaron el Análisis de Conglomerados para clasificar a pacientes con sindrome de disfunción del dolor temporomandibular, clasificando el dolor facial de acuerdo a su distribución espacial, con la finalidad de que esta clasificación pudiera ser útil en la identificación de las distintas etapas de la enfermedad, lo cual ayudaría a definir tratamientos más directos.

En la Investigación de Mercados; Green et al (1967) agruparon un conjunto de ciudades en 88 grupos, basándose en 14 características, como fue el tamaño de la ciudad, la circulación del periódico y el ingreso per capita entre otras, de tal manera que las ciudades dentro de cada grupo eran muy similares entre sí, posteriormente se eligió una ciudad de cada grupo de la cuál se obtendrían las pruebas de mercado.

En la Educación, Aitkin, Anderson y Hinde (1981) agruparon a profesores en distintos estilos con respecto algunas variables binarias a través de ciertas preguntas como: ¿ los alumnos tienen alguna preferencia en dónde sentarse?, ¿ se utiliza alguna tabla para organizar el trabajo ?, ¿ se realizan exámenes finales ?, ¿ se dan reconocimientos a los alumnos que realizan mejor su trabajo ?; de acuerdo a las respuestas dadas se describieron dos estilos de enseñanza: la 'formal' y la 'informal'.

El Análisis de Conglomerados se puede considerar como descriptivo; se utiliza fundamentalmente como una técnica de exploración. Los objetivos que persigue el Análisis de Conglomerados son la: exploración de datos, reducción de datos, generación de hipótesis y predicción basados en grupos

Hay muchos problemas prácticos que involucran al Análisis de Conglomerados, los resultados de tal análisis dependerán de las consideraciones que se tomen: como la elección del método de agrupación a utilizar, las variables que serán medidas y cuales son consideradas importantes. La adición o eliminación de variables relevantes puede tener un impacto fuerte en la solución resultante, por lo que el investigador debe de tener sumo cuidado en evaluar el impacto de cada decisión implicada en el desarrollo de un Análisis de Conglomerados. En el Análisis de Conglomerados no hay un camino completamente satisfactorio para definir los conglomerados, generalmente se tiene una idea intuitiva.

La mayoría de las veces cuando se aplican técnicas multivariadas se inicia directamente con la aplicación de alguna de ellas sin conocer previamente los datos, el realizar un análisis previo de estos, permite conocer su comportamiento y detectar observaciones que puedan alterar el conjunto de datos; es por ello que el *capitulo uno* da a conocer algunas técnicas gráficas más usuales en el análisis exploratorio, así como posibles problemas que se puedan encontrar dentro de los datos como es el caso de observaciones atípicas y ausentes, y sus soluciones alternativas. Asimismo también se dan a conocer los supuestos que debe de cumplir las técnicas multivariadas, en particular en el Análisis de Conglomerados. Por otro lado, es importante conocer que medidas son las que permiten saber cuales objetos son semejantes o similares y cuales no, y de esta manera poder iniciar la formación de grupos, es por ello que el *capitulo dos* introduce estas medidas llamadas proximidades, que permiten medir la cercanía o lejanía de los objetos, y que son divididas en disimilaridades, distancias y similaridades, y pueden ser usadas tanto en datos cuantitativos como cualitativos. También es necesario tener una metodología en donde se apliquen estas medidas de proximidad, por lo que el *capitulo tres* da una amplia explicación de los diferentes métodos que permitan agrupar a los objetos y son la base primordial del Análisis de Conglomerados, los cuales se dividen en jerárquicos y no jerárquicos, y donde a su vez los primeros son divididos en aglomerativos y divisivos. También se dan conocer algunas técnicas que permiten elegir el número óptimo de conglomerados y las propiedades y problemas de algunas técnicas jerárquicas aglomerativas. Por otro lado, existen algunos métodos multivariados como es el escalamiento multidimensional y el Análisis de Componentes Principales que pueden auxiliar al Análisis de Conglomerados; por esta razón en este capitulo se da una breve explicación del segundo y como se puede utilizar dentro de nuestro tema de interés. En el *capitulo cuatro* se utilizan los datos iris de Fisher donde se ejemplifican algunas técnicas gráficas bivariadas y multivariadas, y algunos métodos de agrupación jerárquicos y no

jerárquicos utilizando alguna medida de proximidad. También se realiza un Análisis de Conglomerados con el Análisis de Componentes Principales. Y finalmente, en dan las conclusiones generales y algunas recomendaciones.

CAPÍTULO 1

ANÁLISIS PREVIO DE LOS DATOS

INTRODUCCIÓN

El análisis de los datos es un paso importante y necesario, que generalmente lleva tiempo y que es descuidado por los analistas, uno de los problemas a los que se enfrentan es como evaluar y solucionar los problemas en el diseño de la investigación y en la recolección de los datos. El análisis de los datos nos permite identificar las tendencias en los datos y observaciones atípicas, revelar la formación de grupos y realizar comparaciones entre ellos así como establecer asociaciones entre variables. La exploración de los datos nos puede servir como guía para elegir después supuestos y modelos que puedan describir el conjunto de datos. El obtener resúmenes tanto estadísticos como gráficos complementan una interpretación de los resultados y por tanto una buena toma de decisiones.

En este capítulo se presentan técnicas gráficas univariadas, bivariadas y multivariadas para el análisis de las variables. También se muestran los procesos de evaluación para entender el impacto que puedan tener los datos ausentes sobre el análisis, y las posibles alternativas para el tratamiento de ellos; también la utilización de técnicas que mejor se ajustan para la identificación de datos atípicos. Por otro lado, también se muestran los supuestos que deben de cumplir las técnicas multivariadas y en particular el supuesto de multicolinealidad que es el único que se aplica al Análisis de Conglomerados.

1. 1 MÉTODOS GRÁFICOS UNIVARIADOS

El punto de partida para entender la naturaleza de las variables es caracterizar la forma de su distribución, esto es, a través de la frecuencia con que ocurren los valores que toman esas variables, por lo que el uso de tablas de frecuencia permiten visualizar la manera en que se distribuye el conjunto de datos, las tablas de frecuencia (distribuciones de frecuencias) contienen frecuencias absolutas *(fi)* y relativas *(pi)*, las primeras representan el número de veces en que se observó una variable y las segundas es el cociente de la frecuencia absoluta entre el número total de observaciones. La suma de las frecuencias absolutas debe ser igual al número de observaciones y la suma de las frecuencias relativas debe ser igual a uno.

Los métodos gráficos permiten describir características presentes de un conjunto de datos, ayudan a confirmar supuestos y pueden sugerir acciones correctivas.

Para las variables cualitativas existen dos tipos de métodos gráficos: el diagrama pie y el diagrama de barras. El primero compara las partes que componen una entidad con la entidad completa, expresándolas como porcentajes. Estos diagramas se construyen calculando el ángulo al que corresponde proporcionalmente la frecuencia relativa de la categoría a los 360 grados del círculo, esto es: *Angulo de la porción del círculo = 360 x pi.* El diagrama de barras representa gráficamente a las frecuencias relativas, el eje vertical

denota las frecuencias y el horizontal contiene las categorías de la variable, encima de cada categoría se alza una barra cuya altura es igual a la frecuencia relativa observada en esa categoría, el ancho de las barras debe ser el mismo y las barras deben encontrarse espaciadas entre sí.

Para las variables cuantitativas, se busca caracterizar la variabilidad presente en la población estadística a través de su distribución de frecuencias, los métodos gráficos que se utilizan son: el diagrama de punto y el diagrama de tallo y hoja. *En el diagrama de punto* se muestra el número de veces en que se presenta cada medición dentro del conjunto de datos, la construcción de estos diagramas se lleva a cabo colocando en el eje horizontal las diferentes observaciones de la variable y sobre cada valor se anotan tantos puntos como veces se repiten estos valores. En este diagrama se pueden observar ciertas características que se presentan en los datos como son las observaciones atípicas, que son valores muy grandes o pequeños respecto al conjunto de datos; los huecos que son grandes espacios entre el conjunto de puntos; y la distribución que son los valores frecuentes. *El diagrama de tallo y hoja* combina un método gráfico y otro de ordenación, este diagrama se forma con el (los) primer (os) dígito (s) del dato, mientras que la hoja se forma con los dígitos restantes, este tipo de diagrama nos permite detectar que tan alejados se encuentran los datos entre sí y alrededor de que valor se concentran, si existen datos atípicos o grupos de ellos y si existe simetría en la manera en que se distribuyen. Para la construcción de este diagrama se elige un par de dígitos subyacentes para dividir los valores, se escriben los dígitos del tallo de manera vertical y se les separa con una línea vertical de las hojas que se escriben en una secuencia de números enteros y por último se ordenan las hojas de manera creciente. La construcción de la distribución de frecuencias para las variables discretas es lo mismo que para las cualitativas, solo que las categorías son los valores discretos que toma la variable, se les llamarán clases en lugar de categorías.

El Histograma es la representación más útil para la distribución de frecuencia de datos continuos. Para las variables continuas se puede dar el caso de que ningún valor se repita, por lo que es necesario construir rangos o intervalos para clasificar a las observaciones; el procedimiento para formar los rangos consiste en determinar el máximo y el mínimo valor observado y sacar su diferencia (amplitud), este valor se divide en sub intervalos (intervalos de clase), este resultado indica la longitud del intervalo de clase que debe ser la misma para todas las que se formen, el número de clases se determina de forma arbitraria, aunque lo recomendable es emplear entre 5 y 20. El límite inferior del intervalo de clase debe ser menor al mínimo valor observado, el siguiente límite inferior se obtiene la sumarle la longitud que se obtuvo anteriormente. Ya formados todos los intervalos se efectúa el conteo del número de observaciones cuyos valores pertenecen a cada uno de ellos y se calculan las frecuencias absolutas y relativas. La tabla de frecuencias para las variables continuas incluye la frecuencia absoluta acumulada y la frecuencia relativa acumulada, la frecuencia absoluta acumulada para un intervalo dado se calcula sumando todas las frecuencias absolutas de intervalos anteriores a él, más la frecuencia absoluta que le corresponde. Esta misma se puede presentar en forma de proporción solamente dividiéndola entre el número total de observaciones, a la cual se le designará frecuencia relativa acumulada. El histograma es una gráfica de barras en la que el ancho de éstas es la longitud de los intervalos de clase y la altura es igual a la frecuencia observada. Desde el punto de vista visual no hay diferencia si se grafica la frecuencia relativa en lugar de la

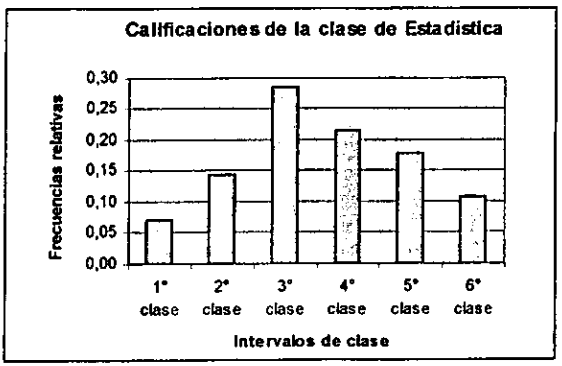
absoluta. En este tipo de gráficas se puede ver dónde se concentran las observaciones y que tan dispersos están (forma de la distribución).

Ejemplo:

A los alumnos de la clase de estadística se les promediaron las tareas, participaciones y exámenes obteniendo las siguientes calificaciones: 4.8, 5.0, 5.6, 6.0, 6.2, 6.4, 6.6, 6.6, 6.8, 6.9, 7.0, 7.0, 7.1, 7.2, 7.5, 7.8, 7.9, 8.0, 8.2, 8.4, 8.5, 8.8, 8.9, 9.2, 9.3, 9.5, 9.7, 10.0. se construirá la tabla de frecuencias agrupándolos en 6 intervalos de clase.

Clase	Intervalos de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
1	(4.5, 5.4)	2	2/28 = 0.07	2	0.07
2	(5.5, 6.4)	4	4/28 = 0.14	6	0.21
3	(6.5, 7.4)	8	8/28 = 0.29	14	0.50
4	(7.5, 8.4)	6	6/28 = 0.21	20	0.71
5	(8.5, 9.4)	5	5/28 = 0.18	25	0.89
6	(9.5, +)	3	3/28 = 0.11	28	1.00
		n = 28	1		

Fuente: Cálculos propios
Tabla de frecuencias
Tabla 1.1



Fuente: Cálculos propios
Histograma
Gráfica 1.1

Una variante del histograma es el *polígono de frecuencias* se construye uniendo los puntos medios de la parte superior de las barras del histograma y se cierran los extremos con el eje horizontal, esto puede ser útil para visualizar el perfil de la distribución de frecuencias.

También existe otra gráfica que es *el diagrama de caja y brazos* que involucra medidas de tendencia central y medidas de dispersión; este diagrama resulta ser muy útil cuando se desea comparar dos o más conjuntos de datos, además se emplea para analizar y presentar las características más importantes de un conjunto de observaciones como es la simetría, localización, dispersión y observaciones atípicas. Su forma es la siguiente:

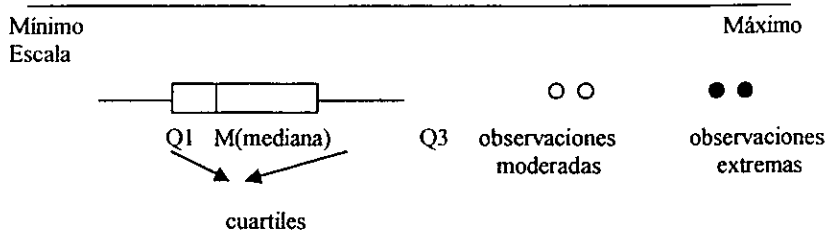


Diagrama de caja y brazos
Gráfica 1.2

En donde la longitud de la caja es la distancia entre el primero y tercer cuartil (rango intercuartil), la línea dentro de la caja es la mediana, sí esta cae cerca del final de la caja, se indica la presencia de asimetría, cuanto es mayor la caja, mayor es la extensión de las observaciones. El diagrama de caja es muy útil para comparar varios lotes de datos univariados cuando se grafican paralelamente, sin embargo es útil analizarlo de manera conjunta con el diagrama de tallo y hojas o un histograma para no incurrir en falsas interpretaciones.

1. 2 ANÁLISIS ENTRE VARIABLES

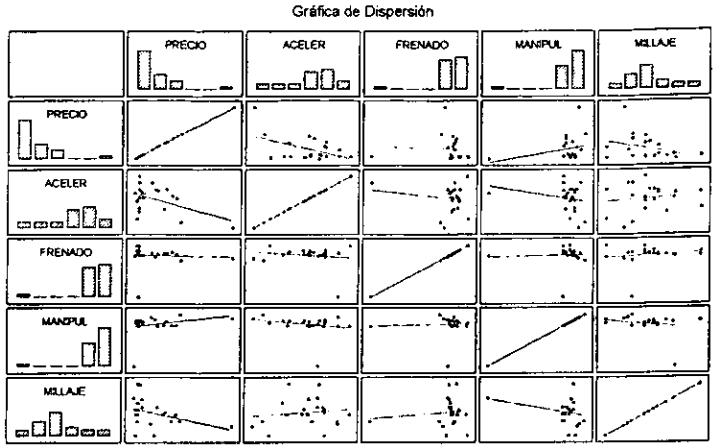
La relación entre las variables se puede obtener calculando el coeficiente de correlación (ver expresión matemática en el apéndice), el cual nos indica el grado de asociación que tienen las variables que se están comparando. Los valores que puede tomar este coeficiente están entre cero y uno. Cuando toma valores cercanos a uno se dice que las variables tienen un alto grado de asociación. Por otro lado, el método que se utiliza para representar dicha asociación es la gráfica de dispersión, que es una gráfica de puntos basados en dos variables, las cuales representan los correspondientes valores conjuntos de esas variables. Cuando los puntos se encuentran alrededor de una línea, se dice que tenemos una relación lineal de correlación, además el coeficiente de correlación nos daría un número

alto, cuando los puntos son curvados indica una relación no lineal, y cuando no existen patrones, es decir cuando sólo existe un conjunto de puntos aleatorios, indica que no hay relación alguna. Este tipo de diagramas puede dar evidencia de patrones o estructuras en los datos, en particular la presencia de conglomerados. La expresión para calcular el coeficiente de correlación de una población y de una muestra se encuentra en el apéndice.

A continuación se mostrará la gráfica de dispersión con datos relacionados a 22 marcas de carros con las siguientes características: aceleración (aceler), millas por galón (millaje), índice de arrendamiento de camino (manejo) frenado de 80 millas por hora (frenado).

	Precio	Aceleración	Frenado	Manejado	Millaje
Acura	-0,521072363	0,477252671	-0,006571039	0,381619066	2,078753555
Audi	0,865652474	0,208033216	0,31869537	-0,091373579	-0,677061608
BMW	0,495859184	-0,801539742	0,192202878	-0,091373579	-0,153805564
Buick	-0,613520685	1,688740221	0,933087475	-0,20962174	-0,153805564
Corvette	1,235445763	-1,8111127	-0,494470651	0,972859872	-0,677061608
Chrysler	-0,613520685	0,073423488	0,427117506	-0,20962174	-0,153805564
Dodge	-0,705969008	-0,195795968	0,481328574	0,145122743	-0,153805564
Eagle	-0,613520685	1,217606173	-4,198893635	-0,20962174	-0,677061608
Ford	-0,705969008	-1,541893245	0,987298543	0,145122743	-1,723573695
Honda	-0,42862404	0,409947807	-0,006571039	0,026874582	0,369450479
Isuzu	-0,79841733	0,409947807	-0,060782107	-4,230059224	1,067125204
Mazda	0,126065894	0,679167263	-0,133063531	0,499867227	-1,723573695
Mercedes	1,050549118	0,006118624	0,119921454	-0,091373579	-0,153805564
Mitsubichi	-0,613520685	-1,003454334	0,083780742	0,381619066	0,718287842
Nissan	-0,42862404	0,073423488	-0,006571039	0,263370905	0,997357732
Olds	-0,613520685	-0,734234878	0,40904715	0,381619066	2,113637291
Pontiac	-0,613520685	0,679167263	0,535539642	0,145122743	0,195031798
Porsche	3,454205501	-2,214941883	-0,295696735	0,618115388	-1,02589897
Saab	0,588307506	0,679167263	0,246413946	0,263370905	0,020613117
Toyota	-0,058830751	1,217606173	0,22834359	0,73636355	-0,851480289
VW	-0,705969008	-0,128491104	0,101851098	0,381619066	0,195031798
Volvo	0,218514217	0,611862399	0,13799181	-0,20962174	0,369450479

Fuente: Paquete Statistica
Datos de marcas de carros
Tabla 1.2



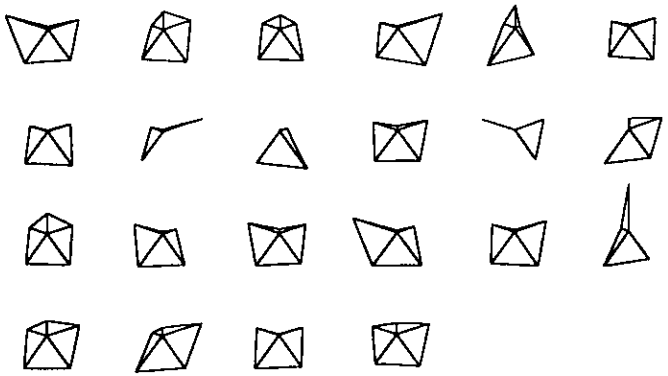
Fuente: Tabla 1.2
Gráfica de Dispersión
Gráfica 1.3

1. 3 MÉTODOS GRÁFICOS MULTIVARIADOS

Hasta ahora los métodos gráficos mencionados han sido útiles para datos con una o dos variables, pero cuando se desea comparar observaciones caracterizadas por la presencia de más de dos variables, se necesita un medio para representar una observación multivariada. Algunas gráficas utilizadas para representar este tipo de observaciones son: los glyphs o metroglyps, las estrellas, los perfiles multivariados, las caras de Chernoff y las Series de Fourier.

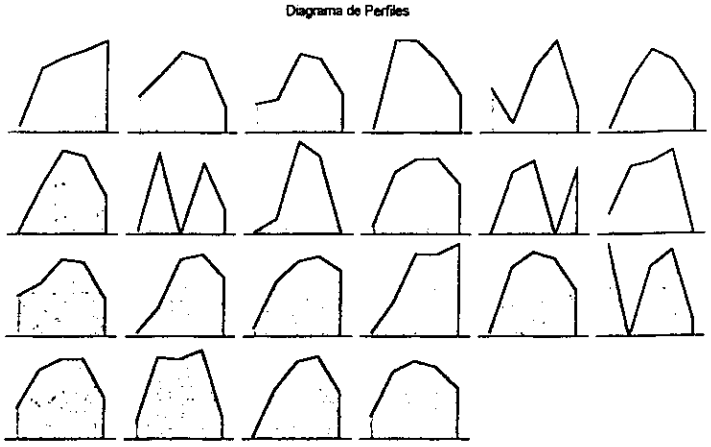
Anderson (1960) desarrolló una técnica para presentar los datos sobre un vector de dimensión *p* usando los *glyps y metroglyps. Un glyp* es un circulo de radio fijo con *p* rayas igualmente espaciadas emanando desde el centro del circulo, y donde cada raya corresponde a los valores de las variables. Una variación de los glyps son *las estrellas o polígonos.* En este tipo de iconos se representa el valor de cada variable a lo largo de cada radio o raya que van desde el centro hacia el exterior del círculo y el final de las rayas es conectado con líneas rectas para formar la estrella o el polígono.

Diagrama de Estrellas

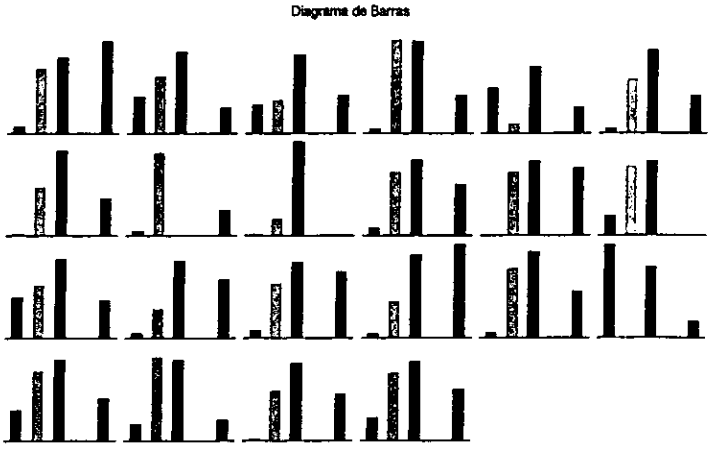


Fuente: Tabla 1.2
Diagrama de Estrellas
Gráfica 1.4

Cada estrella representa una observación multivariada con sus p variables y pueden ser agrupadas de acuerdo a sus similitudes. Otra representación gráfica son *los perfiles y los diagramas de barras* donde cada barra vertical corresponde a un valor de la variable cuya altura es proporcional al valor que toma la variable. Algunas veces el perfil es esquematizado por una línea poligonal más que barras.

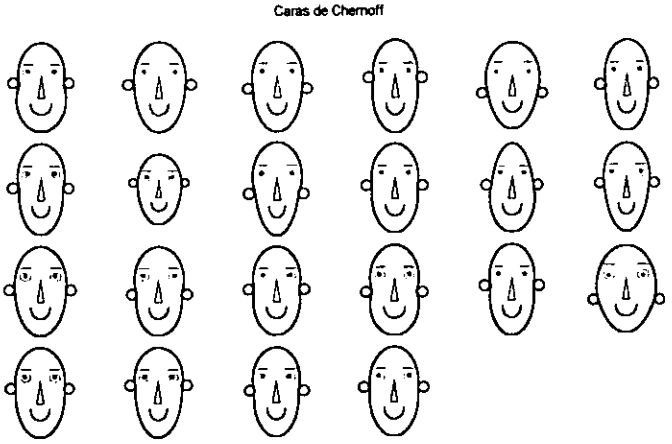


Fuente: Tabla 1.2
Diagrama de Perfiles
Gráfica 1.5



Fuente : Tabla 1.2
Diagramas de Barras
Gráfica 1.6

Las caras de Chernoff (1973) es una forma nueva para representar los datos multivariados, hoy en dia se ha convertido en una herramienta muy útil en el análisis exploratorio de los datos con aplicaciones en conglomerados y detección de datos atípicos. Chernoff, originalmente tomo en cuenta hasta 18 dimensiones en un vector, donde cada valor de la variable se asociaba con cada uno de los 18 rasgos faciales. Sin embargo, Bruckner (1978) realizó un programa que generaba caras con 6 rasgos faciales: cabeza, boca, nariz, ojos, cejas y orejas. La creación de una cara empieza asignando una variable a un rasgo facial y la asignación puede ser de manera aleatoria o designada.



Fuente: Tabla 1.2
Caras de Chernoff
Gráfica 1.7

Por último, *las series de Fourier* propuestas por Andrew (1972) representan un vector de p dimensiones con medidas $x' = (x_1, x_2, ..., x_p)$ a través de la siguiente función:

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 sent + x_3 \cos t + x_4 sen2t + x_5 \cos 2t + ... \qquad \pi < t < \pi$$

Esto significa que los valores de las variables llegan a ser los coeficientes en una expresión cuyo gráfico es una función periódica. Las gráficas que representan las series de Fourier son curvas que se pueden agrupar visualmente. Estas gráficas permiten la

identificación de datos atípicos, de conglomerados o agrupaciones y otras características interesantes de los datos.
Algunas gráficas de Andrew para el mismo conjunto de observaciones se pueden construir permutando las variables y recalculando las funciones $f(t)$. Por otro lado, Embrechts y Herzberg (1991) sugirieron que las curvas de Andrew fueran calculadas sobre los datos estandarizados (con media cero y desviación estándar de uno), esto es para evitar que los valores de las variables grandes en x disfracen de manera visual el efecto de otras variables en las funciones graficadas.

Cabe mencionar que las representaciones mencionadas anteriormente se pueden obtener con programas estadísticos. Por otro lado, también existen los árboles donde cada variable esta representada por la longitud de una rama.

1.4 VALORES FALTANTES O AUSENTES Y ALGUNAS ALTERNATIVAS

Antes de que el analista pueda encontrar una solución para la ausencia de datos o valores, debe diagnosticar su ausencia. Cuando los procesos de ausencia de datos son desconocidos, el analista se puede plantear preguntas sobre si los datos ausentes están distribuidos aleatoriamente entre las observaciones y en que medida son relevantes. El impacto de los datos ausentes puede ser perjudicial, no sólo por sus potenciales sesgos sino también por su efecto en el tamaño de la muestra. Si no se aplican soluciones para la ausencia de datos, no se debe incluir ninguna observación con valores faltantes para cualquiera de las variables. Por otro lado, el analista debe buscar observaciones adicionales o encontrar una solución para la ausencia de datos en la muestra original.

Las causas que originan la ausencia de datos son los errores en la introducción de los datos o fallas al completar el cuestionario, o bien, la acción por parte del encuestado, rehusarse a contestar. En estas situaciones, se tiene poco control en el proceso de ausencia de datos, pero si las observaciones ausentes son de carácter aleatorio se pueden aplicar ciertas soluciones. Una forma de diagnosticar si las observaciones ausentes son aleatorios, consiste en formar dos grupos para valorar una sola variable, uno con observaciones que tengan datos ausentes y el otro con valores válidos para esa variable, se aplican test para determinar si existen diferencias significativas para otras variables de interés, si existen patrones de diferencias significativas entonces la ausencia de datos no es aleatoria.

Un primer tratamiento consiste en utilizar solo las observaciones con datos completos, pero es recomendable que se aplique solo si los datos ausentes son completamente aleatorios, en caso contrario se podrían sesgar los resultados. Por otro lado, el utilizar datos completos se ajusta mejor cuando la muestra es suficientemente grande para permitir la supresión de los casos con los datos ausentes.

Un segundo tratamiento consiste en eliminar las observaciones y / o variables que peor se comportan respecto a los datos ausentes, cualquier decisión ésta basada en las consideraciones empíricas y teóricas. Si una variable que no sea la dependiente tiene valores ausentes y es una candidata a la eliminación se debe de asegurar que se tengan

variables alternativas que se espera estén altamente correlacionadas, para representar la intención de la variable original.

El tercero es el de imputación, que consiste en estimar los valores ausentes basándose en los valores que son válidos de otras variables u observaciones, es recomendable para variables métricas (cuantitativas). Los métodos de imputación se pueden clasificar en dos tipos: como el uso de toda la información disponible a partir de un subconjunto de observaciones para generalizar sobre el total, o como métodos para estimar valores para las observaciones ausentes. El primer tipo no reemplaza las observaciones ausentes sino que imputa las características de distribución (desviación estándar, media) y las correlaciones de otras observaciones. El segundo tipo consiste en la sustitución de los datos ausentes por valores estimados con información existente en la muestra, este procedimiento se puede llevar cabo mediante una sustitución directa, eligiendo observaciones que no estén en la muestra; o con estimaciones basadas en relaciones entre variables calculando el valor medio para la variable que esta ausente con todas las observaciones que tienen valores completos o bien, sustituirlo por un valor constante que se puede obtener de fuentes externas. Estos procedimientos tienen desventajas como el modificar la correlación observada puesto que todos las observaciones ausentes tendrán un valor de correlación constante, y el distorsionar la distribución de los valores.

Los Métodos para estimar valores ausentes o faltantes en el contexto de Análisis de Conglomerados son descritos por Dixon(1979) y Little y Rubin(1987).

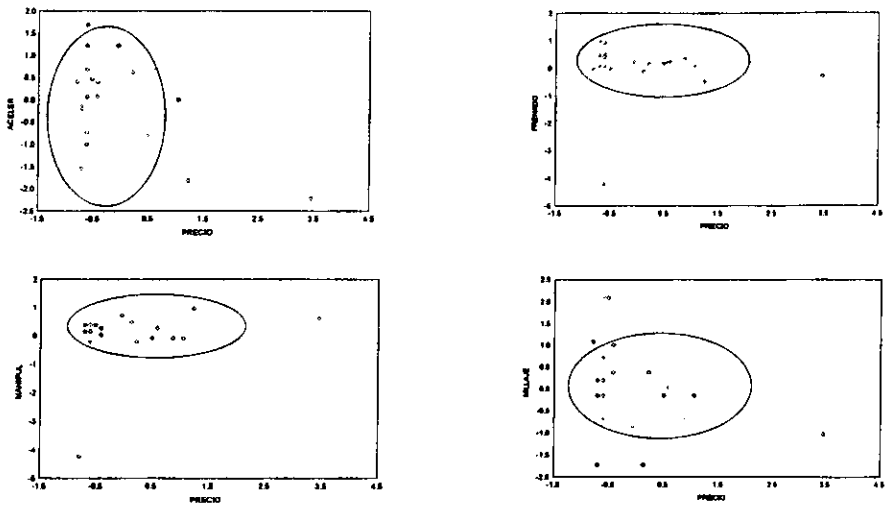
1. 5 OBSERVACIONES ATÍPICAS Y TRATAMIENTO

Las observaciones atípicas (outliers) son aquellas que no forman parte del conjunto de datos, gráficamente se verían alejadas de un conjunto de puntos, este tipo de observaciones pueden influir en el análisis de los datos, ya que por poseer características diferentes a las demás observaciones no son representativos de la población. Como se menciono anteriormente, una de las causas que origina a las observaciones atípicas son los errores en la entrada de los datos y su detección gráfica se puede hacer con el diagrama de caja y brazos.

Para datos univariados se pueden estandarizar los valores y considerar como atípicos las observaciones con valores estándar de 2.5 o mayores para una muestra de 80 o menos observaciones. Si los tamaños de muestra son mayores, los valores estándar se deben de situar entre 3 y 4.

Para datos bivariados se puede utilizar el diagrama de dispersión para analizar dos variables y las observaciones que caigan fuera del rango de observaciones pueden ser consideradas como atípicos; para establecer dicho rango se puede trazar una elipse para representar una región de confianza para una distribución normal bivariada. Ver gráfica 1.8. Para observaciones multivariados, es decir observaciones con un conjunto de variables, se puede utilizar una medida de distancia como la de Mahalanobis (D2).

Una vez que se han detectado los casos atípicos, el analista debe evaluar si realmente estas observaciones son candidatas para ser atípicas y no caer en el error de catalogarlas de esta manera solo porque no son consistentes con el resto de las observaciones. Por otro lado se pueden utilizar técnicas multivariadas como el análisis discriminante o el análisis de regresión para identificar las diferencias entre las observaciones atípicas y las restantes. Cuando ya se identificaron las observaciones atípicas, se tiene la opción de eliminarlas o mantenerlas, algunos estadísticos las mantienen a menos de que sean realmente aberrantes y no sean representativos de la población. Si se opta por eliminarlas se corre el riesgo de mejorar el análisis pero limitar su generalidad.



Fuente: Tabla 1.2
Gráficas de Dispersión para detectar datos atípicos
Gráfica 1.8

1.6 VERIFICACIÓN DE SUPUESTOS EN EL ANÁLISIS MULTIVARIADO

La verificación de los supuestos en el análisis multivariado es la última etapa del análisis de los datos, se considera que el análisis multivariado requiere que los supuestos a las técnicas estadísticas sean contrastados dos veces: una para las variables aisladas y otro para el valor teórico del modelo multivariado que actúa sobre las variables a analizar, y por tanto debe cumplir los mismos supuestos que las variables individuales. Dentro de los

supuestos que se deben analizar en las técnicas multivariadas se encuentra el supuesto de normalidad que se relaciona a la forma de la distribución de los datos para una sola variable cuantitativa y su correspondencia con una distribución normal; el supuesto de homocedasticidad que se basa en la dispersión de la varianza de la variable dependiente a lo largo del rango de los valores de la variable independiente, lo mismo sucede cuando las variables independientes no son cuantitativas; el de la linealidad que representa el grado de cambio en la variable dependiente asociado con la variable independiente y la multicolinealidad que muestra la relación tan estrecha que existe entre las variables. Este último supuesto es el único que se verifica en el Análisis de Conglomerados porque no es una técnica de inferencia estadística en la que se utilizan los parámetros de una muestra, más bien una metodología de cuantificación de las características de un conjunto de objetos.

La multicolinealidad se puede ver como si alguna variable independiente es combinación lineal de otras (ver apéndice A, A4) o bien, cuando alguno de los coeficientes de correlación simple o múltiple entre algunas variables independientes es uno, esto es cuando algunas variables independientes están correlacionadas entre sí. En el Análisis de Conglomerados, las variables que son multicolineales están implícitamente ponderadas con más fuerza. Por ejemplo, supongamos que sé esta agrupando a los encuestados sobre 10 variables, todas son afirmaciones de actitud hacia un servicio. Cuando se examina la multicolinealidad, se ve que existen dos conjuntos de variables, el primero constituido de 8 afirmaciones y el segundo de las otras dos; si el objetivo es agrupar a los encuestados a partir de las número de variables (dimensión) del producto(representadas por dos grupos de variables), entonces será un error utilizar las 10 variables iniciales. Dado que cada variable se pondera de igual manera en el Análisis de Conglomerados, la primera dimensión tendrá como mucho 4 veces más posibilidad, 8 objetos frente a dos, para afectar a la medida de proximidad, y lo mismo ocurrirá con la segunda dimensión. La multicolinealidad actúa como un proceso de ponderación, que es no aparente para el que lo observa, pero que sin embargo puede afectar el análisis.

CAPÍTULO 2

ANÁLISIS DE CONGLOMERADOS

PROXIMIDADES: MEDIDAS DE SIMILARIDAD, DISIMILARIDAD y DISTANCIA

INTRODUCCIÓN

En este capítulo se aborda el tema de las medidas de proximidad, las cuales nos miden la cercanía de los objetos que van a ser agrupados. Inicialmente se toma la distancia entre todos los objetos y los más cercanos son los que entran a un primer conglomerado o grupo, después se recalculan las distancias de éste primer grupo con los objetos restantes y se continúan formando los grupos. Estas medidas se clasifican en similaridades y disimilares, aunque dentro de estás últimas figuran las distancias. Sin embargo antes de hablar de lo que son las medidas de proximidad, es necesario introducir en que forma deben de estar las variables y cuales hay que elegir, ya que de esto depende los resultados que se obtengan en el Análisis de Conglomerados.

2.1 ELECCIÓN DE VARIABLES

La elección inicial de un conjunto particular de medidas que se utilizan para describir a cada individuo que va a ser agrupado, constituye un marco de referencia dentro del cual se establecen los conglomerados. La primera pregunta acerca de la elección de las variables, es si son o no son relevantes para el tipo de clasificación que se está haciendo, por ejemplo, si el objetivo es la clasificación de los enfermos mentales que podría ser útil para evaluar los efectos de los diferentes tratamientos, tal vez no sería necesario incluir variables como peso, altura y otras estadísticas, ya que esto podría originar que los conglomerados resultantes fueran simplemente hombres y mujeres.

La siguiente pregunta que puede ser considerada es sobre cuántas variables deberían ser medidas en cada individuo. En la mayoría de los casos, teóricamente, se considera un número ilimitado de variables para ser utilizadas en la clasificación. En la práctica muchas serán consideradas irrelevantes para el propósito que se tiene y una restricción que surgirá serán las consideraciones de tiempo o económicas. Al igual que en la pregunta sobre cuáles variables hay que medir, no hay en general una base teórica para determinar el número de variables a usar; sin embargo, el problema debe ser abordado empíricamente. Tomando en cuenta consideraciones teóricas, conceptuales y prácticas; el investigador debe incluir aquellas que caracterizan a los objetos que se están agrupando. La inclusión de variables irrelevantes pueden crear atipicidades que pueden tener un efecto significativo en los resultados por eso muy importante que solo se eligan las que esten en cocordancia con el objetivo planteado.

2.2 ESTANDARIZACIÓN

En muchas aplicaciones las variables que describen a los objetos que serán agrupados no están medidas en las mismas unidades; pueden ser variables completamente diferentes, algunas categóricas, otras ordinales y otras en una escala de intervalo. No es adecuado decir peso medido en libras, altura medida en pulgadas y ansiedad estimada en una escala de cuatro puntos. Para variables en escala de intervalo, la solución sugerida es simplemente estandarizar cada una de ellas para unificar la varianza anterior para algún análisis, usando las desviaciones estándar calculadas del conjunto de individuos a ser agrupados . Fleiss y Zubin (1969) muestran que esto puede tener una seria desventaja como diluir diferencias entre grupos sobre las variables que son los mejores discriminadores. Otra desventaja, es que se ignoran posibles correlaciones entre las variables.

Algunas de las alternativas cuando las variables son de distintos tipos, es convertir todas las variables en forma binaria antes de calcular similaridades, o aplicar análisis separados del mismo conjunto de objetos, donde cada análisis comprende variables de un solo tipo e intenta sintetizar los resultados de los distintos estudios. Otra posibilidad es un coeficiente de similaridad que pueda incorporar información de los distintos tipos de variables el cual se verá con más detalle en lo sucesivo.

2.3 PONDERACIÓN DE VARIABLES

Ponderar una variable significa darle mayor o menor importancia respecto a las otras. Varios autores cuestionan la validez de este procedimiento, por ejemplo Sokal y Sneath (1973) dicen que los pesos están basados en juicios intuitivos de lo que es importante, y que estos pueden simplemente reflejar la existencia de clasificaciones de los datos. En general esto no es lo que se requiere en una aplicación de conglomerados. Los métodos del Análisis de Conglomerados se aplican a un conjunto de datos de donde surgirán grupos aunque esto no siempre se puede dar porque los datos no apoyan la formación de grupos.

Gordon(1980) argumenta que si la información sobre la relevancia de las variables no está disponible, lo más apropiado es darles la misma ponderación. Sin embargo existen variables que pueden ser consideradas más importantes para discriminar grupos de objetos . De Sarbo et al (1984) describen un método que además de proporcionar clasificaciones al conjunto de datos también da los pesos de las variables indicando su importancia relativa para el conglomerado. El problema real de una ponderación apriori es que es un tanto difícil decidir como ponderar las variables en la práctica.

2.4 INTRODUCCIÓN A LAS PROXIMIDADES

El concepto de proximidad es fundamental para el Análisis de Conglomerados. La proximidad entre objetos es una medida de correspondencia o parecido entre objetos en relación a un cierto número de características cualitativas o cuantitativas y es calculada para todos los pares de objetos.

El punto de partida dentro del Análisis de Conglomerados es una matriz de proximidades X de n x p, donde hay n individuos u objetos y cada uno de los cuales tiene valores para p variables. Los valores para el individuo i se denotan como: $x_{i1}, x_{i2}, \ldots, x_{ip}$ y para el individuo j $x_{j1}, x_{j2}, \ldots, x_{jp}$. Esta matriz de proximidades mide la similaridad o disimilaridad entre los individuos u objetos, entonces existe una matriz de similaridad (S) y otra de disimilaridad (D).

Las medidas de similaridad también son conocidas como coeficientes de similaridad o de correlación y las de disimilaridad como coeficientes de disimilaridad, distancias o métricas. Entre más parecidos sean los individuos, la medida de similaridad aumentará mientras que la medida de disimilaridad disminuirá y viceversa.

El investigador debe tomar en cuenta ciertas caracteristicas que tienen las medidas de proximidad cuando trata de elgir alguna de ellas. Diferentes medidas de distancia o un cambio en la escala de las variables puede llevar a diferentes soluciones . Por tanto, es aconsejable utilizar varias medidas y comparar los resultados con pautas teóricas. Cuando las variables están correlacionadas, la medida más adecuada puede ser la distancia de Mahalanobis, dado que se ajusta para las correlaciones y ponderaciones de todas las variables igualmente.

2.5 MEDIDAS DE SIMILARIDAD

Las medidas de similaridad conocidas como coeficientes de asociación, indican la intensidad de la relación entre dos objetos i y j . Se han propuesto coeficientes de similaridad, dependiendo del tipo de variables: cuantitativas, categoricas, binarias, ordinales.

Cada $i-$*ésimo* individuo u objeto será representado por un vector de observaciones $x'_i = (x_1, x_2, \ldots, x_p)$ sobre las p *variables*. Si $\wp$ es una población de objetos, se puede definir una similaridad entre dos objetos i y j, como una función que mapea $\wp x \wp \rightarrow \Re^1$ si satisface los siguientes axiomas:

i) $0 \leq r(i,j) \leq 1$ para todo $i,j \in \wp$
ii) $r(i,i) = 1$
iii) $r(i,j) = 1$ si y solo si $i = j$
iv) $r(i,j) = r(j,i)$

se denotará $r(i,j) = r_{ij}$

Aunque $r(i,j)$ es la notación general, para variables cualitativas se utilizará como notación las iniciales del autor que propusó la medida de similaridad.

2.5.1 Medidas de similaridad para variables nominales o cualitativas

Los coeficientes de similaridad más simples y comunes se utilizan en variables dicotómicas o binarias. Las variables dicotómicas pueden tomar solamente dos valores como: 0 y 1, blanco o negro, hombre o mujer, verdadero o falso, etc.; aunque también se pueden tratar como la presencia o ausencia de alguna característica o atributo en dos individuos i,j. Con este tipo de variables se puede formar una tabla de contingencia $2x2$ para cada par de individuos i,j, donde cada entrada suma el número de atributos que son o no son comunes en ambos individuos (Tabla 2.1).

		Objeto j		
		Presencia (+)	Ausencia (-)	Suma
Objeto i	Presencia (+)	α	β	$\alpha+\beta$
	Ausencia (-)	γ	δ	$\gamma+\delta$
	Suma	$\alpha+\gamma$	$\beta+\delta$	$\alpha+\gamma+\beta+\delta$

Número de características presentes o ausentes para dos individuos

Tabla 2.1

Donde:

α = es el número de características presentes tanto en el individuo i como en el individuo j,

β = es el número de características presentes en el individuo i pero ausentes en j

γ = es el número de características ausentes en i pero presentes en j

δ = es el número de características ausentes en ambos individuos

$p = \alpha+\beta+\gamma+\delta$

Muchas medidas de similaridad o coeficientes de asociación han sido propuestas para combinar las cantidades α,β,γ y δ. Entre las medidas más usuales se encuentran: el coeficiente de apareamiento simple, el coeficiente de Jaccard (1908) y el coeficiente de Czekanowski (1913). En la tabla 2.2 se presentan estas y otras combinaciones.

i) Apareamiento simple

$$PS_{ij} = \frac{\alpha+\delta}{\alpha+\beta+\gamma+\delta}$$

ii) Jaccard

$$J_{ij} = \frac{\alpha}{\alpha+\beta+\gamma}$$

iii) Dice, Czekanowski y Sorenson

$$D_{ij} = \frac{2\alpha}{2\alpha + \beta + \gamma}$$

iv) Sokal y Sneath

$$SS_{ij} = \frac{2(\alpha + \delta)}{2(\alpha + \delta) + \beta + \gamma}$$

v) Rogers y Tanimoto

$$RT_{ij} = \frac{\alpha + \delta}{\alpha + \delta + 2(\beta + \gamma)}$$

vi) Sokal y Sneath (medida 2)

$$SS2_{ij} = \frac{\alpha}{\alpha + 2(\beta + \gamma)}$$

vii) Kulczynski

$$K_{ij} = \frac{\alpha}{\beta + \gamma}$$

viii) Sokal y Sneath (medida 3)

$$SS3_{ij} = \frac{\alpha + \delta}{\beta + \gamma}$$

ix) Russel y Rao

$$RR_{ij} = \frac{\alpha}{\alpha + \beta + \gamma + \delta}$$

Otras medidas dicotomicas o binarias
x) Ochiai

$$O_{ij} = \sqrt{\frac{\alpha}{\alpha + \beta} \cdot \frac{\alpha}{\alpha + \gamma}}$$

xi) Sokal y sneath (medida 5)

$$SS5_{ij} = \frac{\alpha\delta}{\sqrt{(\alpha+\beta)+(\alpha+\gamma)+(\beta+\delta)+(\gamma+\delta)}}$$

xii) Coeficiente de Correlación Phi o de Pearson

$$\varphi_{ij} = \frac{\alpha\delta - \beta\gamma}{(\alpha+\beta)+(\alpha+\gamma)+(\beta+\delta)+(\gamma+\delta)}$$

Medidas o coeficientes de similaridad para datos dicotomicos

Tabla 2.2

La elección de un coeficiente depende de los pesos relativos que se le den a las combinaciones positivas o presencias (α) y combinaciones negativas o ausencias (δ); por ejemplo el coeficiente de apareamiento simple no podría ser el adecuado cuando existan ausencias de alguna característica que tengan poco peso, o que ni lo tengan, con respecto a las presencias; sin embargo, podría ser útil cuando todas las variables son nominales y sus dos alternativas tienen el mismo peso.

Por otro lado, las distintas medidas de similaridad pueden tener diferentes valores para el mismo conjunto de datos. Estas medidas no son conjuntamente monótonicas, lo que significa que si todos los valores para distintos pares de individuos sobre una medida fueran ordenadas de menor a mayor, los valores para los mismos pares tomados por otra medida no serían concordantes con la ordenación anterior.

Ejemplo:

Se tienen los siguientes valores para dos individuos:

	1	2	3	4	5	6	7	8	9	10
individuo 1	1	0	0	0	1	1	0	0	1	0
individuo 2	0	0	0	0	1	0	0	1	1	0

y para un tercero:

	1	2	3	4	5	6	7	8	9	10
individuo 3	0	0	0	0	0	0	0	1	0	0

Los resultados obtenidos para dos de los coeficientes son:

Apareamiento simple	Jacard
$PS_{12} = 0.70$	$J_{12} = 0.40$
$PS_{13} = 0.50$	$J_{13} = 0.00$
$PS_{23} = 0.80$	$J_{23} = 0.33$

Al ordenar los valores anteriores de manera creciente para cada coeficiente, se tienen las siguientes series:

0.50, 0.70, 0.80 para el primer coeficiente
0.00, 0.40, 0.33 para el segundo coeficiente

De acuerdo a lo anterior, la serie del segundo coeficiente no cumple la condición de ser creciente, por tanto los coeficientes no son conjuntamente monotónicos.

No existe un criterio absoluto que permita decidir el coeficiente de similaridad más adecuado. En la elección de un determinado coeficiente intervienen el tipo de datos que se desea representar y el peso que se le quiera dar a las frecuencias α, β, γ y δ. Esto es un problema que debe ser resuelto para cada situación experimental concreta.

Para las *variables cualitativas* que toman *más de dos valores*, como puede ocurrir cuando la variable es el color de ojos, se puede manejar de la misma manera que para valores dicotomicos, considerando el nivel de cada variable como una variable dicotomica sola.

La tabla de contingencia se puede formar como sigue:

		Objeto j				
		1	2	...	u	suma
Objeto i	1	n_{11}	n_{12}	...	n_{1u}	$n_{1.}$
	2	n_{21}	n_{22}	...	n_{2u}	$n_{2.}$
	⋮					⋮
	v	n_{v1}	n_{v2}	...	n_{vu}	$n_{v.}$
	suma	$n_{.1}$	$n_{.2}$	...	$n_{.u}$	$n_{..}$

donde u y v son el número de categorías para los objetos i y j respectivamente. La entrada n_{ij} es el número de características que caen en el objeto i y j. Se puede colocar un valor de cero o uno a cada variable k dependiendo de si los objetos i y j son los mismos sobre la variable. Los valores para todas las variables son promediados:

$$n_{ij=} \frac{\sum_{k=1}^{p} n_{ijk}}{p}$$

2.5.2 Medidas de disimilaridad para variables cuantitativas

Para obtener el valor del coeficiente de correlación entre ambos objetos i,j, se tiene que tomar el valor de cada una de las variables de cada objeto que se esté comparando; es decir: x_{i1} con x_{j1}, x_{i2} con x_{j2}, y así sucesivamente. Esto es:

$$r_{ij} = \frac{\sum_{k}(x_{ik} - \bar{x}_{i\bullet})(x_{jk} - \bar{x}_{j\bullet})}{\left\{\sum_{k}(x_{ik} - \bar{x}_{i\bullet})^2 \sum_{k}(x_{jk} - \bar{x}_{j\bullet})^2\right\}^{1/2}}$$

donde $k = 1,2,\ldots,p$ y $-1 \leq r_{ij} \leq 1$

Además de que este coeficiente de correlación no cumple con el axioma (1), ha sido criticado por varios autores por las desventajas que posee, por ejemplo, sobre el significado que se le da a $\bar{x}_{i\bullet}$ (media sobre todas las variables del objeto i), o cuando $r_{ij} = 1$ no significa que $x_i = x_j$, a menos que los elementos de x_i estén linealmente relacionados a los de x_j.

Aunque ha habido ciertos autores que difieren de esta opinion, la evidencia sugiere que las disimilaridades basadas en métricas son mejores medidas de proximidad que las correlaciones. Cormark (1971) estableció que el uso del coeficiente de correlación debe ser restringido en situaciones donde las variables no son codificadas y no son medidas comparables o numerados, este no es invariante bajo el escalamiento de las variables, o incluso bajo alteraciones en la dirección de codificación de algunas variables.

2.5.3 Medidas de disimilaridad para mezcla de variables

Gower(1971) propuso un coeficiente para conjuntos de datos donde hubiera diferentes tipos de variables,

$$r_{ij} = \frac{\sum_{k=1}^{p} r_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

Aqui $w_{ijk} = 1$, excepto cuando una comparación no es posible, como con observaciones faltantes o combinaciones negativas (ausencias) de variables dicotómicas, en cuyo caso $r_{ijk} = w_{ijk} = 0$.

En la tabla 2.4, se presentan los valores de los coeficientes para variables dicotomicas o variables cualitativas de dos niveles:

presencia / ausencia de variables dicótomicas

objeto i	+	+	-	-
objeto j	+	-	+	-
r_{ijk}	1	0	0	0
w_{ijk}	1	1	1	0

Variable cualitativa de dos niveles o estados

objeto i	1	1	2	2
objeto j	1	2	1	2
r_{ijk}	1	0	0	1
w_{ijk}	1	1	1	1

Coeficientes de Similaridad de Gower

Tabla 2.3

Con variables cualitativas con más de dos niveles (variables multiestados), se hace $r_{ijk} = 1$ si los objetos *i* y *j* concuerdan en la variable k, y $r_{ijk} = 0$ de otra manera. En ambos casos $w_{ijk} = 1$

Para una variable cuantitativa $w_{ijk} = 1$ y :

$$r_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$$

$$= 1 - |x'_{ik} - x'_{jk}|$$

donde R_k es el rango de la variable k y $x'_{ik} = \frac{x_{ik}}{R_k}$. De esta manera, si se tienen d_1 variables cuantitativas, d_2 variables dicotómicas, y d_3 variables multiestados, se obtiene:

$$r_{ijk} = \frac{\sum_{k=1}^{d_1}(1 - |x'_{ik} - x'_{jk}|) + a_2 + m_3}{(d_1 + (d_2 - \delta_2) + d_3)}$$

donde a_2 y δ_2 son el número de presencias y ausencias respectivamente para las variables dicotomicas, y m_3 es el número de combinaciones para las variables con varias categorías (multiestados).

Si todas las variables son dicotomicas, entonces r_{ij} se reduce al coeficiente de Jaccard. Pero si todas las variables son de dos niveles, entonces r_{ij} se reduce al coeficiente de apareamiento simple.

2.6 MEDIDAS DE DISIMILARIDAD Y DISTANCIAS

Una medida de disimilaridad que mide la cercanía de dos objetos x_i y x_j, es una función d que mapea $\Re^p \times \Re^p \rightarrow \Re^1$ y satisface los siguientes axiomas:

i) $d(i,j) \geq 0$ para todo $i,j \in \Re^p$
ii) $d(i,i) = 0$
iii) $d(i,j) = d(j,i)$ para todo i,j en $\Re^p$

Los supuestos *i* y *v* significan que d es definida positiva (ver apéndice A, A.2), *ii* quiere decir que la distancia es cero si el par está formado por elementos iguales y el supuesto *iii* implica que d es simétrica.

Si una medida de disimilaridad además cumple con las condiciones:

iv) $d(i,j) \leq d(i,k) + d(k,j)$ para todo i,j,k *en* $\Re^p$
v) $d(i,j) = 0$ si y solo si $i = j$

se le llama métrica, el supuesto *iv* es la desigualdad del triángulo y el *v* significa que siempre que la distancia sea cero, los dos elementos son iguales.

Toda proximidad cumple al menos con los tres primeros supuestos. Algunas medidas de disimilaridad no satisfacen el supuesto *iv* pero esto no es un requerimiento necesario. Por otro lado, no existe alguna regla que permita decidir que medida de disimilaridad es la óptima, esto depende de la naturaleza de los objetos, de las variables y de la finalidad del análisis. A continuación se muestran algunas medidas de disimilaridad.

2.6.1 Métrica de Minkowski

A partir de esta métrica se pueden obtener distintas métricas según sea el valor de λ

$$d_{ij} = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right\}^{1/\lambda} \qquad para\ \lambda \geq 1$$

2.6.2 Distancia Euclideana

Cuando $\lambda = 2$ en la métrica de Minkowski se obtiene la distancia euclideana.

$$d_{ij} = \left\{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right\}^{1/2}$$

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$

Esta medida de disimilaridad es la más común y la más usada, cumple con los 4 axiomas y es calculada sobre los datos originales. La distancia euclideana entre dos puntos es la longitud de la hipotenusa de un triángulo rectángulo, puede ser generalizada a p dimensiones aunque su representación gráfica no es posible.

Esta distancia tiene ciertas ventajas, como por ejemplo cuando se calcula la distancia entre dos objetos, ésta no queda afectada por la introducción de nuevos objetos al análisis, los cuales pueden ser observaciones aberrantes. Sin embargo las distancias pueden ser afectadas por diferencias en escala entre las variables, es decir si una de las variables tiene mayor variabilidad que las otras, ésta influirá en el cálculo de las distancias, puede tenr un efecto considerable sobre la clasificación de las distancias.

Ejemplo:

A tres niños se les tomaron medidas sobre dos variables: altura (en pies) y peso (en libras) con los siguientes resultados:

	peso	*altura*
niño 1	60	3.0
niño 2	65	3.5
niño 3	63	4.0

Las distancias euclideanas son las siguientes: $d_{12} = 5.02$, $d_{13} = 3.16$ y $d_{23} = 2.06$; sin embargo, si la altura se mide en pulgadas, entonces las distancias son: $d_{12} = 7.81$, $d_{13} = 12.37$ y $d_{23} = 6.32$, comparando ambos conjuntos de distancias se puede ver que el niño 1 es más cercano al niño 2 que al niño 3, caso contrario al primer conjunto de distancias, el niño 1 es más cercano al niño 3 que al 2; de esta manera se puede ver que un cambio de escala en alguna de las variables tiene un efecto sobre la posición de las distancias.

Este problema se puede solucionar dividiendo cada variable por su rango o por su desviación estándard muestral. Este método removerá la dependencia de las variables; pero tiene otros problemas como diluir las diferencias entre conglomerados con respecto a las variables que son los mejores discriminadores. La distancia entre dos puntos dentro de los conglomerados se incrementa con relación a la distancia entre conglomerados, por tanto los conglomerados son menos claros. El cálculo de la distancia euclideana supone que las variables son no correlacionadas por lo que esta medida puede ser muy pobre; por otro lado, los efectos de escalamiento dependen mucho de los sesgos de los datos.

Existe una variante de la distancia euclideana que es la distancia euclideana al cuadrado, su expresión es la siguiente:

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

2.6.3 Distancia Euclideana en forma matricial

Dada una matriz X de (nxp) con vectores renglón $(1xp)$: $x'_1, x'_2, \ldots, x'_n$, la distancia euclideana al cuadrado d_{ij}^2 entre los objetos i y j puede escribirse:

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j) \qquad i,j = 1,2,\ldots,n$$

2.6.4 Distancia Euclideana Estandarizada

Como ya se mencionó anteriormente, la distancia euclideana es sensible a las escalas de medición, por lo que si ésta no es común a las p variables, es mejor utilizar una distancia ponderada en donde los pesos reflejen la importancia de cada una de las variables; una alternativa a esto es la distancia euclideana estandarizada:

$$d_{ij}^2 = \sum_{k=1}^{p} \frac{1}{s_k^2}(x_{ik} - x_{jk})^2 = (x_i - x_j)'D^{-1}(x_i - x_j)$$

s_k^2 con $k = 1,2,\ldots,p$ es la varianza de la variable x_j sobre los n objetos y D es la matriz diagonal cuyos elementos son s_k^2.

2.6.5 Métrica City Block o de Manhattan

Se obtiene cuando $\lambda = 1$ en la métrica de Minkowski:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

El nombre de city block se debe a que las ciudades americanas son construidas en un arreglo rectangular y la distancia que se tiene que recorrer entre dos puntos es como la expresión de arriba con $p = 2$. Esta métrica cumple con las 4 propiedades y tienen las mismas carencias que la distancia euclideana.

2.6.6 Distancia de Mahalanobis

Esta métrica D^2 es una extensión del sistema de pesos, se utiliza para resolver no solo problemas de escalamiento sino también los efectos de correlación entre las variables.

$$d_{ij} = \left\{(x_i - x_j)' S^{-1} (x_i - x_j)\right\}^{1/2}$$

donde $S = \dfrac{\sum_m (x_m - \bar{x})(x_m - \bar{x})'}{n-1}$ ha sido propuesta como una medida de distancia.

La distancia de Mahalanobis es invariante bajo transformaciones de la forma $y_m = Ax_m + b$, donde A es una matriz no singular (ver definición en el apéndice A); sin embargo, Hartigan comenta que la invarianza bajo transformaciones lineales parece menos forzado que la invarianza bajo un cambio de unidades por cada variable; pero además comenta que la distancia de Mahalanobis reduce la claridad de los conglomerados incluso más que el escalamiento de las variables para que tengan varianza unitaria.

2.6.7 Métrica de Gower

En esta métrica el escalamiento a cada variable se da a través de su rango. Considera un promedio de los rangos estandarizados de las variables.

$$d_{ij} = \frac{1}{p} \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{R_k}$$

donde Rj es el rango de la variable k.

2.6.8 Métrica de Canberra

Fue introducida por Lance y Williams. Es utilizada para variables positivas y generalmente insensible a sesgos y a valores distantes.

$$d_{ij} = \frac{1}{p} \sum_{k=1}^{p} \left\{ \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}} \right\}$$

2.6.9 Distancia de Chebyschev

Esta distancia puede ser apropiada en casos donde se desea definir 2 objetos como "diferentes" si ellos son diferentes en alguna de las variables.

$$d_{ij} = \max|x_{ik} - x_{jk}|$$

2.7 MEDIDA ENTRE GRUPOS

También se han propuesto medidas de distancia entre grupos. Esto se puede hacer sustituyendo medias de grupos de las p variables en la fórmula de la distancia euclideana o la de cityt block entre individuos. Por ejemplo el grupo A tiene un vector medio $\bar{x}'_A = (\bar{x}_{A1}, \bar{x}_{A2}, \ldots, \bar{x}_{Ap})$ y el grupo B un vector medio o centroide $\bar{x}'_B = (\bar{x}_{B1}, \bar{x}_{B2}, \ldots, \bar{x}_{Bp})$, entonces la medida de distancia entre los dos grupos es:

$$d_{AB} = \sqrt{\sum_{k=1}^{p} (\bar{x}_{Ai} - \bar{x}_{Bi})^2}$$

Las medidas entre grupos deben de incorporar de una manera u otra, conocimiento de la variación dentro de grupos. Una posibilidad es la distancia de Mahalanobis D^2 la cual toma en cuenta las correlaciones entre las variables

$$d_{ij} = (\bar{x}_A - \bar{x}_B)' W^{-1} (\bar{x}_A - \bar{x}_B)$$

donde W es una matriz de pxp de dispersiones dentro de grupos para los 2 grupos. Cuando las correlaciones entre las variables son pequeñas, D^2 será similar a la distancia euclideana al cuadrado calculada sobre los datos estandarizados. Utilizar D^2 implica que el investigador está dispuesto a suponer que las dispersiones de las variables son casi lo mismo en ambos grupos. Cuando esto no sucede, D^2 es una medida inapropiada.

2.8 SIMILARIDADES A PARTIR DE DISTANCIAS

Una disimilaridad se puede obtener de una similaridad colocando $d_{ij} = 1 - r_{ij}$, aunque d_{ij} no será una métrica a menos que r_{ij} satisfaciera el axioma *iii* de las medidas de similaridad y d_{ij} satisfaciera la desigualdad del triángulo.

Por ejemplo si se aplica la distancia euclideana estandarizada o la métrica de Canberra a datos binarios 0-1, se obtiene $(\beta + \gamma)/d$, el complemeto del coeficiente de apareamiento r_{ij}, así que $1 - r_{ij}$ es una métrica. De manera similar, el complemento de los coefieicntes de Jacard y Czekanowski satisfacen la desigualdad del triángulo para que también sean métricas.

Es posible construir similaridades a partir de distancias, por ejemplo, con la expresión:

$$r_{ij} = \frac{1}{1 + d_{ij}} \qquad \text{donde } 0 < r_{ij} \leq 1$$

donde r_{ij} es la similaridad entre i y j, y d_{ij} es la distancia correspondiente.

Sin embargo, las distancias no siempre pueden ser construidas a partir de similaridades. Gower (1971) demostró que esto se puede hacer solo si la matriz de similaridad es semi definida positiva, $r_{ij} = 1$ con la condición de semi definida positiva (ver apéndice A,A2) y con la similaridad estandarizada máxima, además,

$$d_{ij} = \sqrt{2(1 - r_{ij})}$$

tiene las propiedades de distancia.

Por otro lado, si se agrupan variables más que objetos, entonces, la distancia euclideana entre dos variables estandarizadas kl se puede definir como:

$$d_{kl}^2 = 2(1 - r_{kl})$$

donde $r_{kl} = \sum_i (x_{ik} - \bar{x}_{.k})(x_{il} - \bar{x}_{.l}) / \left\{ \sum_i (x_{ik} - \bar{x}_{.k})^2 \sum_i (x_{il} - \bar{x}_{.l})^2 \right\}^{1/2}$,

Entonces se puede utilizar, $d_{kl} = \sqrt{2(1 - r_{kl})}$ para transformar la medida de similaridad r_{kl} en una distancia. Sin embargo, hay algunos problemas cuando se usa el coeficiente de correlación si una o ambas variables k y l son nominales. Lance y Williams (1968) han propuesto una solución para esto. En las variables binarias se pueden utilizar los valores 0 y1.

CAPÍTULO 3

MÉTODOS JERÁRQUICOS Y NO JERÁRQUICOS

INTRODUCCIÓN

En este capítulo se hablará sobre los métodos que permiten agrupar un conjunto de observaciones, los cuales se dividen en jérarquicos y no jeráquicos, y donde los primeros a su vez, se subdividen en aglomerativos y divisivos. También se hablará sobre algunas de las consideraciones que se deben de tomar en cuenta cuando se decide aplicar alguna de estas técnicas, es decir, sobre el número óptimo de conglomerados o grupos que se tienen que elegir y sobre el método que se debe de utilizar, para lo cual mencionaremos algunas propiedades y problemas que presentan algunos de estos algoritmos. Y por último, se menciona como aplicar el Análisis de Componentes principales dentro del Análisis de Conglomerados.

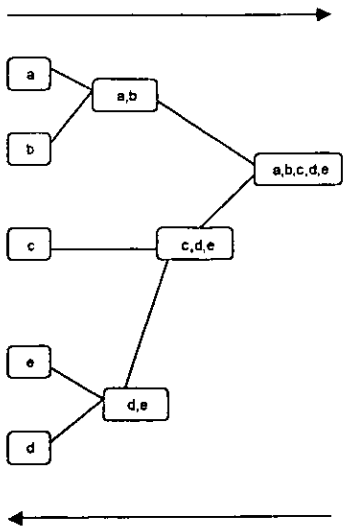
3.1 MÉTODOS JERÁRQUICOS

Los métodos jerárquicos consisten en la construcción de una estructura en forma de árbol. Los datos no son particionados en un número particular de clases en un solo paso. La clasificación consiste en una serie de particiones las cuales pueden correrse desde un solo conglomerado conteniendo a todos los individuos, a n conglomerados conteniendo cada uno, un solo individuo. Existen dos tipos de procedimientos jerárquicos para obtener conglomerados: aglomerativos y divisivos.

Una característica de los procedimientos jerárquicos es que los resultados obtenidos en un paso previo siempre necesitan encajarse dentro de los resultados del siguiente paso, creando un árbol.. Dado que los conglomerados se forman solo por unión de los conglomerados existentes, cualquier miembro de un conglomerado se puede rastrear hasta su origen por simple observación. La representación de este proceso se denomina dendrograma o gráfica en forma de árbol. En la gráfica 3.1 los métodos aglomerativos van de izquierda a derecha y los métodos divisivos van de derecha a izquierda.

Con tales métodos una vez hechas las divisiones o fusiones son irrevocables de modo que que cuando un algoritmo aglomerativo ha unido dos individuos ellos no pueden subsequentemente ser separados y cuando un algoritmo divisivo ha hecho una división, ésta no puede ser reunida nuevamente. Como dicen Kaufman y Rousseuw (1990): un método jerárquico tiene el defecto de que nunca puede reparar lo que fué hecho en pasos anteriores. Todas las técnicas jerárquicas aglomerativas reducen los datos a un solo conglomerado conteniendo todos los individuos y las técnicas divisivas dividirán al conjunto de datos entero en n grupos conteniendo cada uno un solo individuo.

Aglomerativos



Divisivos

Ejemplo de un dendrograma

Gráfica 3.1

3.1.1 MÉTODOS AGLOMERATIVOS

Los *métodos aglomerativos* producen una serie de particiones de los datos, P_n, P_{n-1}, ... ,P_1. La primera P_n consiste de n conglomerados con un solo objeto, la última P_1, consiste de un solo grupo conteniendo a los n individuos. En estos métodos, cada objeto u observación empieza dentro de su propio conglomerado, en etapas posteriores, los dos conglomerados más cercanos (o individuos) se combinan en un nuevo conglomerado, reduciendo así el número de conglomerados paso a paso. En algunos casos un tercer individuo se une a los dos primeros en un conglomerado. En otros, dos grupos de individuos formados en un paso anterior pueden unirse en un nuevo conglomerado. Eventualmente, todos los individuos se agrupan en un único conglomerado. Por esta razón, los procedimientos de aglomeración son denominados a veces como métodos de construcción.

A continuación, se presentan los algoritmos utilizados para desarrollar conglomerados, los cuales por conveniencia serán descritos en términos de distancia. Además se tiene un apéndice A donde se presenta el algebra de matrices como material de apoyo.

3.1.1.1 Liga o encadenamiento simple o vecino más cercano

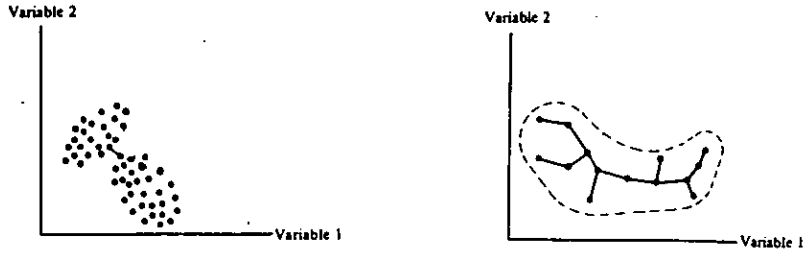
La liga simple es uno de los métodos jerárquicos aglomerativos más simples, la característica de este procedimiento esta basada en la distancia mínima.

Si C_1 y C_2 son dos conglomerados, entonces la distancia entre ellos es la disimilaridad más pequeña entre un miembro de C_1 y C_2 :

$d_{(C_1)(C_2)} = \min\{d_{ij} : i \in C_1, j \in C_2\}$ donde i y j denotan individuos u objetos

Este procedimiento consiste en encontrar los dos objetos separados por la distancia más corta y colocarlos en un primer conglomerado. Después se encuentra la distancia más corta de este conglomerado con los demás objetos, y o bien un tercer objeto se une a los dos primeros para formar un conglomerado o se forma un nuevo conglomerado de dos miembros. El proceso continúa hasta que todos los objetos se encuentran en un conglomerado. La distancia entre dos conglomerados cualquiera es la distancia más corta desde cualquier punto en un conglomerado a cualquier punto en el otro. Dos conglomerados se fusionan en cualquier nivel por el vínculo más corto o más fuerte entre ellos.

Dado que este método junta conglomerados por la distancia más corta entre ellos, la técnica no puede distinguir conglomerados separados (ver gráfica 3.2a). Este método, es de los pocos que delinea conglomerados no elipsoidales. Por otro lado, el método de la liga simple presenta el efecto de encadenamiento, que surge cuando dos conglomerados distintos se unen por unos pocos puntos intermedios (gráfica 3.2b). El efecto de encadenamiento puede ser engañoso si los objetos finales opuestos de la cadena son muy parecidos.



a b

Conglomerados del Método de la Liga Simple

Gráfica 3.2

Ejemplo:

La siguiente matriz es una matriz de distancia donde el elemento en el $i-ésimo$ renglón y la $j-ésima$ columna da la distancia dij entre los individuos i y j. En esta matriz se utiliza la distancia euclideana y se denotará como D_1 :

$$D_1 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0.0 & 1.0 & 5.0 & 6.0 & 8.0 \\ 1.0 & 0.0 & 3.0 & 8.0 & 7.0 \\ 5.0 & 3.0 & 0.0 & 4.0 & 6.0 \\ 6.0 & 8.0 & 4.0 & 0.0 & 2.0 \\ 8.0 & 7.0 & 6.0 & 2.0 & 0.0 \end{pmatrix} \end{array}$$

En la primera etapa se unen los individuos A y B cuya distancia es la más cercana para formar un conglomerado; $d_{AB} = 1.0$ es el elemento más pequeño en D_1. Una vez que se forma el conglomerado se calculan las distancias con este conglomerado y con los otros individuos que no están agrupados C, D, E. Estas son obtenidas como sigue:

$$d_{(AB)C} = \min\{d_{AC}, d_{BC}\} = d_{BC} = 3.0$$

$$d_{(AB)D} = \min\{d_{AD}, d_{BD}\} = d_{BD} = 6.0$$

$$d_{(AB)E} = \min\{d_{AE}, d_{BE}\} = d_{BE} = 7.0$$

Ahora se formará una nueva matriz con estas distancias, D_2 cuyas elementos son distancias entre individuos y entre grupos:

$$D_2 = \begin{array}{c} \\ AB \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} AB & C & D & E \end{array} \\ \begin{pmatrix} 0.0 & 3.0 & 6.0 & 7.0 \\ 3.0 & 0.0 & 4.0 & 6.0 \\ 6.0 & 4.0 & 0.0 & 2.0 \\ 7.0 & 6.0 & 2.0 & 0.0 \end{pmatrix} \end{array}$$

En la matriz D_2, la entrada más pequeña es $d_{DE} = 2.0$, en lugar de incluir un nuevo individuo en el conglomerado que ya se había formado, se formará un segundo conglomerado, con los individuos *D y E*. Las distancias son:

$$d_{(AB)C} = 3.0$$

$$d_{(AB)(DE)} = \min\{d_{AD}, d_{AE}, d_{BD}, d_{BE}\} = d_{AD} = 6.0$$

$$d_{(DE)C} = \min\{d_{CD}, d_{CE}\} = d_{CD} = 4.0$$

De aquí se puede formar una nueva matriz de distancias, D_3

$$D_3 = \begin{array}{c} \\ \text{AB} \\ \text{C} \\ \text{DE} \end{array} \begin{array}{c} \begin{array}{ccc} \text{AB} & \text{C} & \text{DE} \end{array} \\ \begin{pmatrix} 0.0 & 3.0 & 6.0 \\ 3.0 & 0.0 & 4.0 \\ 6.0 & 4.0 & 0.0 \end{pmatrix} \end{array}$$

El elemento más pequeño en D_3 es $d_{(AB)C}$ lo cual indica que el individuo *C* debería unirse al primer conglomerado, que contienen a los individuos *A y B*.

$$d_{(ABC)(DE)} = \min\{d_{AD}, d_{AE}, d_{BD}, d_{BE}, d_{CD}, d_{CE}\} = 4.0$$

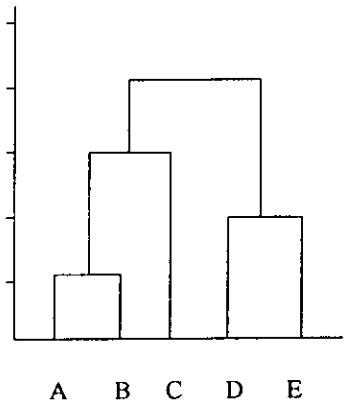
Entonces la matriz D_4 es la siguiente:

$$D_4 = \begin{array}{c} \\ \text{ABC} \\ \text{DE} \end{array} \begin{array}{c} \begin{array}{cc} \text{ABC} & \text{DE} \end{array} \\ \begin{pmatrix} 0 & 4 \\ 4 & 0 \end{pmatrix} \end{array}$$

En la última etapa, los dos conglomerados son unidos para formar uno solo que contenga a los 5 individuos. Las particiones que se produjeron en cada estapa son las siguientes:

etapa	Grupos
P_5	$(A), (B), (C), (D), (E)$
P_4	$(A\ B), (C), (D), (E)$
P_3	$(A\ B), (C), (D\ E)$
P_2	$(A\ B\ C), (D\ E)$
P_1	$(A\ B\ C\ D\ E)$

El dendrograma que resume las etapas donde se realizan las fusiones se ilustra la figura 3.3.



Dendrograma del Método de la Liga Simple

Gráfica 3.3

3.1.1.2 Liga o encadenamiento completo o Vecino más lejano

El método de la liga completa o vecino más lejano es el opuesto al de la liga simple, su métodología es de la misma manera que este último, con excepción de que el método de la liga completa se basa en la distancia máxima.

Sean C_1 y C_2 dos conglomerados, entonces la distancia entre ellos está definida para ser la disimilaridad más grande entre un miembro de C_1 y C_2 :

$d_{(C_1)(C_2)} = \max\{d_{ij} : i \in C_1, j \in C_2\}$ donde i y j denotan individuos u objetos

La distancia entre grupos es ahora definida como la distancia entre los pares de individuos más distantes , es decir en cada etapa, la distancia entre conglomerados es determinada por la distancia entre dos elementos, uno de cada conglomerado, que son los más distantes. La liga completa vincula todos los objetos de un conglomerado con el resto, a alguna distancia máxima o por la mínima similaridad.

El algoritmo aglomerativo general, nuevamente inicia encontrando la distancia más pequeña en la matriz y uniendo los individuos u objetos correspondientes .

Aplicando este método a la matriz D_1, la primera etapa es nuevamente la unión de los individuos A *y* B como en el método de la liga simple, la distancia mínima es $d_{AB} = 1$. Las distancias entre este grupo y los tres individuos restantes son las siguientes :

$$d_{(AB)C} = \max\{d_{AC}, d_{BC}\} = d_{AC} = 5.0$$

$$d_{(AB)D} = \max\{d_{AD}, d_{BD}\} = d_{BD} = 8.0$$

$$d_{(AB)E} = \max\{d_{AE}, d_{BE}\} = d_{AE} = 8.0$$

De aquí se puede obtener la siguiente matriz , D_2 :

$$D_2 = \begin{array}{c} \\ AB \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} AB & C & D & E \end{array} \\ \begin{pmatrix} 0.0 & 5.0 & 8.0 & 8.0 \\ 5.0 & 0.0 & 4.0 & 6.0 \\ 8.0 & 4.0 & 0.0 & 2.0 \\ 8.0 & 6.0 & 2.0 & 0.0 \end{pmatrix} \end{array}$$

En la matriz D_2, la entrada más pequeña es $d_{DE} = 2.0$, entonces un segundo conglomerado se forma con los individuos D y E. Las distancias son:

$$d_{(AB)C} = 5.0$$

$$d_{(AB)(DE)} = \max\{d_{AD}, d_{AE}, d_{BD}, d_{BE}\} = 8.0$$

$$d_{(DE)C} = \max\{d_{DC}, d_{EC}\} = d_{EC} = 6.0$$

La matriz de distancias, D_3 es:

$$D_3 = \begin{array}{c} \\ \text{AB} \\ \text{C} \\ \text{DE} \end{array} \begin{array}{c} \begin{array}{ccc} \text{AB} & \text{C} & \text{DE} \end{array} \\ \begin{pmatrix} 0.0 & 5.0 & 8.0 \\ 5.0 & 0.0 & 6.0 \\ 8.0 & 6.0 & 0.0 \end{pmatrix} \end{array}$$

El elemento más pequeño en D_3 es $d_{(AB)C} = 5.0$, lo cual indica que el individuo C debería unirse al primer conglomerado, que contienen a los individuos A y B.

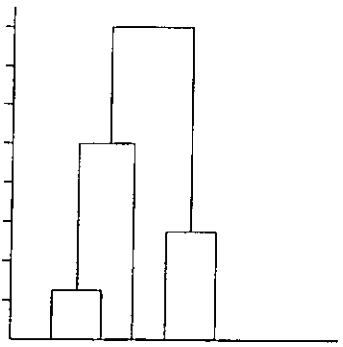
$$d_{(ABC)(DE)} = \max\{d_{AD}, d_{AE}, d_{BD}, d_{BE}, d_{CD}, d_{CE}\} = 8.0$$

Entonces, la matriz D_4 es la siguiente:

$$D_4 = \begin{array}{c} \\ \text{ABC} \\ \text{DE} \end{array} \begin{array}{c} \begin{array}{cc} \text{ABC} & \text{DE} \end{array} \\ \begin{pmatrix} 0 & 8 \\ 8 & 0 \end{pmatrix} \end{array}$$

En la última etapa, los dos conglomerados son unidos para formar uno solo que contenga a los 5 individuos al igual que en el método de la liga simple.

El dendrograma correpondiente a este método se muestra en la gráfica 3.4.



Dendrograma del Método de la Liga Completa

Gráfica 3.4

El método de la liga simple y el de la liga completa fueron discutidos inicialmente por Sneath (1957) posteriormente por Sokal y Sneath (1963). Estos métodos se parecen al método del mínimo y máximo discutidos por Jonhson (1967).

3.1.1.3 Promedio grupal o encadenamiento promedio

El criterio de este método es la distancia media. La distancia entre dos conglomerados se define como el promedio de las distancias entre todos los pares de individuos, uno de un conglomerado y uno del otro. Esta técnica no depende de los valores extremos, como se hace en los métodos de la liga simple y completa, y la partición se basa en todos los miembros de los conglomerados en lugar de un par único de miembros extremos. El enfoque de este método tiende a combinar los conglomerados con variaciones reducidas dentro del conglomerado.

Sean C_1 *y* C_2 2 conglomerados, entonces la distancia entre ellos esta definida como el promedio de las $n_1 n_2$ disimilaridades entre todos los pares, esto es:

$$d_{(C_1)(C_2)} = \frac{1}{n_1 n_2} \sum_{i \in C_1} \sum_{j \in C_2} d_{ij}$$

Aplicando este método a la matriz D_1, la primera etapa como en la liga simple y completa es la formación de un conglomerado conteniendo a los individuos A y B. Las distancias promedio entre el conglomerado formado por AB y los otros individuos se muestran en la matriz D_2.

$$D_2 = \begin{array}{c} \\ AB \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} AB & C & D & E \end{array} \\ \begin{pmatrix} 0 & 4 & 7 & \frac{15}{2} \\ 4 & 0 & 4 & 6 \\ 7 & 4 & 0 & 2 \\ \frac{15}{2} & 6 & 2 & 0 \end{pmatrix} \end{array}$$

Los individuos D y E deben ser unidos para formar un nuevo conglomerado. Las nuevas distancias son las siguientes y están dadas en la matriz D_3.

$$d_{(AB)C} = \frac{1}{2}(d_{AC} + d_{BC}) = \frac{1}{2}(5+3) = 4$$

$$d_{(AB)(DE)} = \frac{1}{4}(d_{AD} + d_{AE} + d_{BD} + d_{BE}) = \frac{1}{4}(6+8+8+7) = \frac{29}{4}$$

$$d_{(DE)C} = \frac{1}{2}(d_{DC} + d_{EC}) = \frac{1}{2}(4+6) = 5$$

$$D_3 = \begin{array}{c} \\ AB \\ C \\ DE \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & DE \end{array} \\ \begin{pmatrix} 0 & 4 & \frac{29}{4} \\ 4 & 0 & 5 \\ \frac{29}{4} & 5 & 0 \end{pmatrix} \end{array}$$

La distancia más pequeña es $d_{(AB)C} = 4$, entonces se forma un nuevo conglomerado con estos individuos. las distancias son:

$$d_{(ABC)DE} = \frac{1}{5}(d_{AD} + d_{AE} + d_{BD} + d_{BE} + d_{CD} + d_{CE}) = \frac{1}{5}(6+8+8+7+4+6) = \frac{39}{5}$$

La matriz D_4 es :

$$D_4 = \begin{array}{c} \\ \text{ABC} \\ \text{DE} \end{array} \begin{array}{c} \begin{array}{cc} \text{ABC} & \text{DE} \end{array} \\ \begin{pmatrix} 0 & \frac{39}{5} \\ \frac{39}{5} & 0 \end{pmatrix} \end{array}$$

En al última etapa se unen los 5 individuos al igual que en los métodos anteriores.

Como se ha visto, los tres métodos descritos anteriormente operan con una matriz de proximidades y no necesitan accesar a los valores de las variables originales de los individuos.

3.1.1.4 Método del Centroide

En este método la distancia entre dos conglomerados es la distancia entre sus centroides. Los centroides de los grupos son los valores medios de las observaciones del conglomerado. Los centroides de los conglomerados cambian, se calcula un nuevo centroide a medida que las uniones van teniendo lugar; es decir, existe un cambio en un centroide de un grupo, cada vez que un nuevo individuo o grupos de individuos se añaden al conglomerado existente.

Con este método, una vez que los conglomerados están formados son representados por sus valores medios para cada variable, esto es su vector de medias o centroide, y la distancia entre grupos es definida en términos de la distancia entre los dos vectores medios.

Sea:

$$x_1 = \sum_{i \in C_1} \frac{x_i}{n_1} \qquad y \qquad x_2 = \sum_{j \in C_2} \frac{x_i}{n_2}$$

los centroides de los n_1 y n_2 miembros de C_1 y C_2 respectivamente, entonces:

$$d_{(C_1)(C_2)} = P(\bar{x}_1, \bar{x}_2)$$

donde P es una medida de proximidad como la correlación, la distancia euclideana u otra medida de disimilaridad. El centroide de $C_1 \cup C_2$, la fusión de C_1 y C_2 está dado por el promedio ponderado:

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Ilustraremos la operación de este método al siguiente conjunto de datos bivariados:

Individuo	*variable*1	*variable*2
A	1.0	1.0
B	1.0	2.0
C	6.0	3.0
D	8.0	2.0
E	8.0	0.0

En la matriz D_1 se encuentran las distancias euclideanas entre los individuos:

$$D_1 = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{A} & \text{B} & \text{C} & \text{D} & \text{E} \end{array} \\ \begin{pmatrix} 0.00 & 1.00 & 5.39 & 7.07 & 7.07 \\ 1.00 & 0.00 & 5.10 & 7.00 & 7.28 \\ 5.39 & 5.10 & 0.00 & 2.24 & 3.61 \\ 7.07 & 7.00 & 2.24 & 0.00 & 2.00 \\ 7.07 & 7.28 & 3.61 & 2.00 & 0.00 \end{pmatrix} \end{array}$$

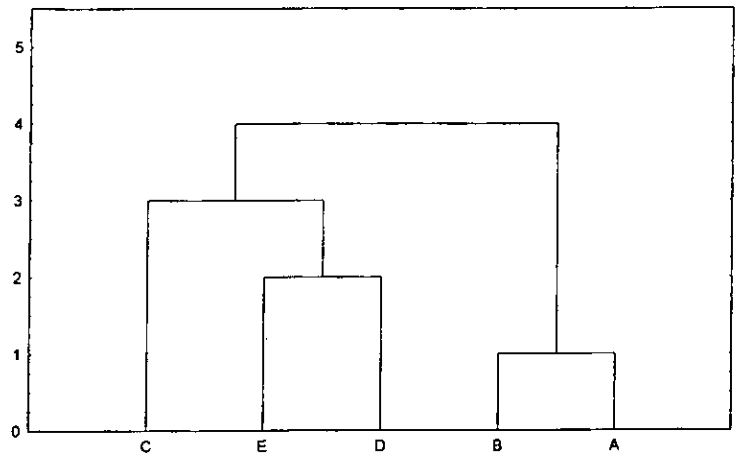
La distancia d_{12} es la más pequeña en la matriz D_1, entonces se unen los individuos A y B para formar un grupo. Se calcula el vector medio de este grupo, (1.0, 1.5), entonces se deduce una nueva matriz de distancia euclideana de este grupo (1,2) con los individuos restantes.

$$D_2 = \begin{array}{c} \\ AB \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} AB & C & D & E \end{array} \\ \begin{pmatrix} 0.00 & 5.22 & 7.02 & 7.16 \\ 5.22 & 0.00 & 2.24 & 3.61 \\ 7.02 & 2.24 & 0.00 & 2.00 \\ 7.16 & 3.61 & 2.00 & 0.00 \end{pmatrix} \end{array}$$

En esta matriz el elemento más pequeño es $d_{DE} = 2.0$ y los individuos A y B se unen para formar un segundo conglomerado y su vector medio es: (8.0,1.0). Entonces se obtiene una nueva matriz D_3 :

$$D_3 = \begin{array}{c} \\ AB \\ C \\ DE \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & DE \end{array} \\ \begin{pmatrix} 0.00 & 5.22 & 7.02 \\ 5.22 & 0.00 & 2.83 \\ 7.02 & 2.83 & 0.00 \end{pmatrix} \end{array}$$

En D_3 la entrada más pequeña es $d_{(DE)C}$ y los individuos C, D y E se unen en un conglomerado de tres miembros. La etapa final consiste de la unión de los dos grupos restantes para formar uno solo. El dendrograma se encuentra en la figura 3.5.



Dendrograma del Método del Centroide

Gráfica 3.5

3.1.1.5 Método de la Mediana

Este método es como el método del centroide, excepto que un nuevo conglomerado es reemplazado por el promedio no ponderado, $x = \frac{x_1 + x_2}{2}$. Fué introducido para superar una deficiencia del método del centroide, esto es, que si un grupo pequeño se une con uno mas grande, o sea si los tamaños de los 2 grupos son diferentes, el grupo más pequeño pierde su identidad, sus características se pierden virtualmente, y entonces el centroide del nuevo grupo será muy cercano al del grupo más grande y puede quedarse dentro de ese grupo. La estrategia se puede hacer independientemente del tamaño del grupo suponiendo que los grupos que van a ser unidos son de igual tamaño, la posición aparente del nuevo grupo será entonces entre los 2 grupos que van a ser unidos. Por otra parte, si los centroides de los grupos que van hacer unidos son representados por (i) y (j), entonces la distancia del centroide de un tercer grupo (h) del grupo formado por la union de (i) y (j) se situa a lo largo de la mediana del triángulo formado por (i), (j) y (h). Por esta razón Gower (1967), quien primero sugirió la estrategia, propuso el nombre mediana.

3.1.1.6 Método de Ward

Ward (1963) propuso un método para formar conglomerados que esta basado en la "pérdida de información" que resulta de la agrupación de individuos en conglomerados. Esta perdida de información esta definida en términos de un criterio de error de las sumas de cuadrados (ESC). En cada paso dentro del análisis, se considera la unión de todo posible par de conglomerados, y los dos conglomerados cuya fusión da como resultado un mínimo incremento en "perdida de información" son combinados.

Este método busca formar las particiones $P_n, P_{n-1}, \ldots, P_1$ de tal manera que se minimice la perdida asociada con cada agrupamiento. El proceso inicia considerando g grupos de un solo objeto cada uno. El primer grupo se forma por la selección de 2 de esos g grupos que cuando se unen producirán el mínimo daño en el valor de ESC. Este $g - 1$ conjunto de grupos se reexamina para determinar los siguientes 2 de esos $g - 1$ grupos para unirlos mientras se minimiza el incremento en ESC. Los g grupos iniciales son de esta manera sistemáticamente reducidos de g a $g - 1$ a $g - 2$ hasta 1 grupo. Los cambios en el valor de ESC en cada estapa proporcionan una pista importante para la determinación de un número natural de agrupaciones para los g objetos. Los procedimientos jerárquicos no pueden conducir a un valor óptimo de ESC para un número especifico de grupos. Sin embargo, una solución no óptima puede ocurrir solo en circunstancias donde el agrupamiento natural de las características de los objetos sea muy débil.

La pérdida de información esta definida por la siguiente expresión:

$$ESC = \sum_{m=1}^{g} \left(\sum_{i=1}^{n_m} (x_{im} - \frac{1}{n} \left(\sum_{i=1}^{n_m} x_{im} \right)^2 \right)$$

donde x_{ij} denota el valor la característica o variable para el i – *ésimo* individuo en el m – *ésimo* conglomerado, g es el número total de conglomerados en cada etapa y n_m es el número de individuos en el m – *ésimo* conglomerado.

Ejemplo:

Se ha recolectado información de 5 individuos para una sola variable, los datos se encuentran en la siguiente tabla.

Individuo	*Valor de la variable*
A	2
B	5
C	9
D	10
E	15

La *ESC* es cero en la primera etapa, puesto que cada individuo constituye un conglomerado. Después se consideran todos los posibles conglomerados de tamaño 2 y se combinan esos individuos que producen el *ESC* más pequeño. Para este conjunto de individuos *C* y *D* son fusionados. La siguiente etapa calcula *ESC* por agregar cada uno de los individuos restantes al primer conglomerado y por formar todos los posibles pares de los 3 individuos no agrupados. Puesto que el incremento en el *ESC* asociado por agregar cada individuo restante al primer conglomerado es más grande que el asociado con la fusión de los individuos *A* y *B*, se forma un nuevo conglomerado, el cual contiene esos 2 individuos. Después el individuo *E* se une al primer coanglomerado que contiene a los individuos *C* y *D* ya que este produjó el incremento más pequeño en el *ESC*. Finalmente, los 2 conglomerados restantes son unidos. La propuesta de Ward se puede mostrar considerando datos univariados.

En la tabla 3.1 se muestran los cálculos de estas etapas y el dendrograma correspondiente en la gráfica 3.6:

Primera etapa

A=2	B=5	C=9	D=10	E=15

Segunda etapa

AB = 4.5	BD = 12.5	
AC = 24.5	BE = 60.5	
AD = 32.0	CD = 0.5	
AE = 98.0	CE = 24.5	
BC = 8.0	DE = 18.0	

Tercera etapa

CDA=28.0	CDB=14	CDE=28.66	AB=4.5	
AE=98.0	BE=60.5			

Cuarta etapa

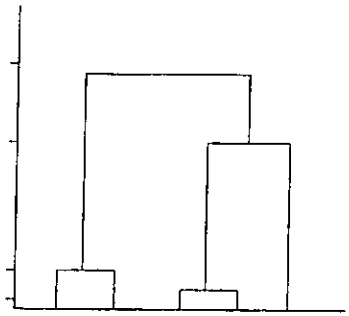
ABCD=41.0	ABE=108.66	CDE=28.66	

Quinta etapa

ABCDE=113.2

Cálculo de ESC

Tabla 3.1



Dendrograma del Método de Ward

Gráfica 3.6

3.1.1.7 Método de incremento en la suma de cuadrados

Wishart (1969), basado en la idea de Ward (1963) para el caso de datos univariados, sugirió fusionar los 2 conglomerados que minimizaran $I_{(C_1)(C_2)}$, que es el incremento en la suma de cuadrados totales de las distancias de los respectivos centroides a fusionar dentro delconglomerado.

$$I_{(C_1)(C_2)} = \sum_{i \in C_1 \cup C_2} \|x_i - \bar{x}\|^2 - \left\{ \sum_{i \in C_1} \|x_i - \bar{x}\|^2 + \sum_{i \in C_2} \|x_i - \bar{x}\|^2 \right\}$$

$$= \sum_{a=1}^{2} n_a \|\bar{x}_a - \bar{x}\|^2$$

$$= \frac{n_1 n_2}{n_1 + n_2} \|\bar{x}_1 - \bar{x}_2\|^2$$

En particular para los objetos i y j,

$$I_{(C_1)(C_2)} = \tfrac{1}{2} \|x_i - x_j\|^2 = \tfrac{1}{2} d_{ij}^2$$

Si iniciamos con $D = [(d_{ij}^2)]$, se puede definir la "distancia" entre 2 conglomerados por,

$$d_{(C_1)(C_2)} = 2l_{(C_1)(C_2)}$$

3.1.1.8 Método Flexible de Lance y William

Lance y William (1967a) mostraron que los métodos como el de la liga simple, liga completa, y el de la liga promedio, método del centroide y el de la mediana son casos especiales de la formula para la distancia entre conglomerados $C_1, C_2 \cup C_3$:

$$d_{(C_3)(C_1 \cup C_2)} = \alpha_1 d_{(C_3)(C_1)} + \alpha_2 d_{(C_3)(C_2)} + \beta d_{(C_1)(C_2)} + \gamma |d_{(C_3)(C_1)} - d_{(C_3)(C_2)}|$$

Wishart demostró que el método de incremento en la suma de cuadrados, satisfacía también la fórmula anterior, los valores de los párametros se encuentran en la tabla 3.2, donde n_g es el número de objetos en el conglomerado C_g $(g = 1,2,3)$.

Por otro lado, Lance y William propusieron un esquema, que satisfaciera las condiciones: $\alpha_1 + \alpha_2 + \beta = 1, \beta < 1$, y $\gamma = 0$ y recomendaron un valor para β negativo y pequeño, tal que $\beta = 0.25$

Lo apropiado es iniciar con $D = \left[(d_{(i)(j)})\right]$, para que la definición de $d_{(C_1)(C_2)}$ se considere para conglomerados de un solo objeto. Esto significa que para métodos como el de la mediana, el centroide y el método de incremento en la suma de cuadrados se utiliza $d_{(i)(j)} = d_{ij}^2$, y en otros como el de la liga simple, liga completa y liga promedio, $d_{(i)(j)} = d_{ij}$. La expresión 4.1 se satisface si se utiliza d_{ij}^2 o d_{ij}, aunque en casos posteriores, la expresión 4.1 se consideraría si se trabajara con d_{ij}^2 en lugar de d_{ij}.

En la mayoría de los métodos se puede aplicar las similaridades en lugar de disimilaridades, proporcionandolo necesario para que las mediciones tengan sentido. Por ejemplo, una correlación promedio en el método de la liga promedio es generalmente inaceptable , ya que las correlaciones pueden ser negativas. Por otro lado, si se utilizan las correlaciones, las propiedades geométricas del método del centroide y de la mediana se pierden.

En general , se puede interpretar cercanía en términos de una disimilaridad pequeña o una similaridad grande, y todo esto es requerido intercambiando las palabras máximo y mínimo en los lugares apropiados. Por ejemplo, en el método de la liga simple, la similaridad se puede expresan como:

$C_{(C_1)(C_2)} = \max\{r_{ij} : i \in C_1, j \in C_2\}$ y se unen los coglomerados con la similaridad máxima.

	Párametro		
	α_i	β	γ
Vecino más cercano	$\frac{1}{2}$	0	$-\frac{1}{2}$
Vecino más lejano	$\frac{1}{2}$	0	$\frac{1}{2}$
Centroide	$\frac{n_i}{n_1+n_2}$	$\frac{-n_1n_2}{(n_1+n_2)^2}$	0
Incremental	$\frac{n_i+n_3}{n_1+n_2+n_3}$	$\frac{-n_3}{n_1+n_2+n_3}$	0
Mediana	$\frac{1}{2}$	$-\frac{1}{4}$	0
Promedio grupal	$\frac{n_i}{n_1+n_2}$	0	0
Flexible	$\frac{1}{2}(1-\beta)$	$\beta < 1$	0

Párametros para Lance y Williams (1967)

Tabla 3.2

3.1.2 MÉTODOS DIVISIVOS

Los *métodos divisivos* separan a los *n* individuos sucesivamente en agrupaciones más finas, el proceso de obtención de conglomerados procede en dirección opuesta al método de aglomeración. En los métodos divisivos se empieza con un gran conglomerado que contiene todas las observaciones, En los pasos sucesivos, los objetos que son más diferentes se dividen y se construyen conglomerados más pequeños. Este proceso continúa hasta que cada objeto es un conglomerado en sí mismo.

En estos métodos, el primer paso es dividir los n objetos en 2 grupos, esto se puede hacer de $2^{n-1} - 1$maneras; pero incluso en las computadoras con grandes capacidades de almacenamiento se tiene que restringir el número de subconjuntos considerados. Una vez que se hace la división inicial, los objetos son movidos de un conglomerado a otro o se hacen subdivisiones muy finas de los conglomerados que ya están formados. Lo que distingue a los métodos divisivos es: 1) como se efectua la división inicial, 2) como los conglomerados que ya se formaron son subdivididos.

Los métodos divisivos son de 2 tipos, monotéticos que dividen a los datos basándose en una sola variable, y los politéticos donde las divisiones están basadas en todas las variables. El método divisivo politético más factible, es el que describieron MacNaughton–Smith et al (1964).

3.1.2.1 Métodos Politéticos: Método de la distancia promedio "splinter "

MacNaughton–Smith et al. propusieron un método basado en un grupo principal y un grupo *splinter*.El grupo *splinter* se inicia dividiendo aquel objeto con la distancia promedio más grande con respecto a los otros $n - 1$ objetos, entonces se forman dos grupos. Se calcula la distancia promedio de cada objeto en el grupo principal con objetos del grupo *splinter*, y la distancia promedio de los objetos del grupo principal con objetos de este mismo grupo. Si esos números son todos negativos, el proceso para o continuaría con cada grupo separadamente; de otra manera, el objeto con el valor más grande positivo es trasladado al grupo splinter y el proceso repetido; en otras palabras, significa que si la distancia del objeto al grupo splinter es menor que su distancia al conglomerado principal, este debe ser removido y fusionarse con el grupo *splinter*.

Ejemplo:

Consideremos la siguiente matriz:

	A	B	C	D	E
A	0	12	9	32	31
B	12	0	9	25	27
C	9	9	0	23	24
D	32	25	23	0	9
E	31	27	24	9	0

El individuo con la distancia promedio más grande de los individuos restantes es el individuo *E*. De esta manera, la división inicial es $\{E\}$ y $\{A,B,C,D\}$. En el siguiente paso se calculan las distancias promedio para el grupo splinter y para el grupo principal. La siguiente tabla muestran los cálculos realizados:

Individuo	*Distancia promedio para el grupo splinter*	*Distancia promedio para el grupo principal*	*Diferencia*
A	31	17.67	-13.33
B	27	15.33	-11.67
C	24	13.67	-10.33
D	9	26.67	17.67

Puesto que la distancia promedio para el grupo splinter para el individuo *D*, es menor que la distancia promedio para el grupo principal, el individuos *D* es separado y fusionado con el individuo E. Entonces la composición de los 2 conglomerados es: $\{D,E\}$, $\{A,B,C\}$. En el próximo paso nuevamente se consideran las distancias promedio de los individuos *A*, *B* y *C* para el grupo splinter y para el grupo principal. Se muestran los cálculos a continuación:

Individuo	*Distancia promedio para el grupo splinter*	*Distancia promedio para el grupo principal*	*Diferencia*
A	31.5	10.5	- 21.0
B	26.0	10.5	-15.5
C	23.5	9.0	-14.5

Dado que todas las diferencias son negativas, no es necesario hacer más movimientos. El proceso iniciaría nuevamente en el sentido de que cada uno de los 2 conglomerados sería dividido separadamente. Por ejemplo, si agrupamos $\{A,B,C\}$, los individuos *A* y *B* podrían ser unidos en un grupo y el individuo *C* en otro.

Este método tiene la ventaja de que los cálculos que se requieren son considerablemente menos que los que se utilizan para los métodos donde se hacen " todas las subdivisiones posibles". Por otro lado, una partición ineficiente no puede ser corregida en la última etapa, pero esto también se presenta en los métodos aglomerativos.

3.1.2.2 Métodos Monotéticos

Los métodos monoteticos son generalmente utilizados cuando las variables son binarias. La división esta dada por los objetos que poseen una característica o atributo especifico y por los que no lo tienen. Si las divisiones de este tipo son consideradas solamente, entonces para un conjunto de datos con p variables binarias, hay p divisiones del conjunto inicial, $p-1$ divisiones en cada uno de los 2 subconjuntos formados, y así sucesivamente. Así para cada par de atributos, k y l se puede construir una tabla de $2x2$ de presencias y ausencias como la tabla 2.1 del capitulo 2, pero basada en 2 variables más que en 2 objetos y calcular la estadistica ji -cuadrada para una tabla de contingencia de $2x2$.

$$x_{kl} = \frac{n(\alpha\delta - \beta\gamma)^2}{(\alpha+\beta)(\alpha+\gamma)(\beta+\delta)(\gamma+\delta)}$$

Por ejemplo la división podría ser sobre la variable k, la cual hace $\sum_{k=l} x_{kl}^2$ un máximo. Cada división se puede localizar en un nivel jerarquico preciso sobre un árbol de división de acuerdo al valor max $\sum x^2$. Otros criterios propuestos son:

$$\max \sum \sqrt{x_{kl}^2}$$

$$\max \sum |\alpha\delta - \beta\gamma|$$

$$\max \sum (\alpha\delta - \beta\gamma)^2$$

Ejemplo:

Consideremos el siguiente conjunto de datos para 5 individuos con 3 variables binarias:

	variable		
individuo	1	2	3
A	0	1	1
B	1	1	0
C	1	1	1
D	1	1	0
E	0	0	1

Los resultados de estadisticas *ji – cuadrada* para cada par de variables son:

$$x_{12}^2 = 1.87, \quad x_{13}^2 = 2.22, \quad x_{23}^2 = 0.83$$

$$x_{12}^2 + x_{13}^2 = 4.09$$

$$x_{12}^2 + x_{23}^2 = 2.70$$

$$x_{13}^2 + x_{23}^2 = 3.05$$

Utilizando el criterio max $\sum x^2$, la primera división de los datos en 2 subconjuntos, en individuos que poseen la variable 1 y los que no la tienen es:

$$(B,C,D) \text{ y } (A,E)$$

Estos métodos se han utilizado amplamente en estudios de ecología (Lambert y William (1962,1966) y Pielou (1969))

Las técnicas jerárquicas son utilizadas en muchos campos en los cuales las estructuras jerárquicas no son las más apropiadas, y la lógica de su uso en esas áreas necesita un sumo cuidado. El peligro de imponer un esquema jerárquico sobre datos que no son esencialmente jerárquicos es claro.

Dentro de los métodos jerárquicos, los métodos aglomerativos son los más utilizados. Entre estos se encuentra el método de Ward, el de la liga promedio, y el de la liga completa, aunque los resultados no son claros, y muchas iteraciones del método y el tipo de datos parece existir.

3.2 MÉTODOS NO JERÁRQUICOS O PARTITIVOS

En los procedimientos no jerárquicos en lugar de implicar procesos de construcciones de árboles (dendrogramas), se asignan los objetos a conglomerados una vez que el número de conglomerados a formar es especificado y conocido. En este tipo de técnicas se produce una partición de los objetos para un número particular de grupos a través de la maximización o minimización de algún criterio.

Las técnicas no jerárquicas están basadas en :1) como se inician los conglomerados y como deben ser colocados los objetos en cada uno de ellos, 2) como son reasignados algunos o todos los objetos que ya estaban agrupados en otros conglomerados porque tal vez su asignación inicial fue realmente inadecuada. .

3.2.1 Iniciación de los métodos no jerárquicos

El primer paso consiste en seleccionar g puntos en un espacio de p dimensiones como centros para iniciar la formación de los conglomerados. Existen varias técnicas para elegir los centros de los conglomerados, es decir, elegir los g primeros puntos. Esto puede ser seleccionando g puntos aleatoriamente (Ball y Hall [1965]); regularmente espaciados (Beale [1969a,b]); mutuamente lejanos (Thorndige [1953]); o elegir g puntos aleatoriamente en $\Re^p$, asignar los objetos para formar g conglomerados, y utilizar los g centroides para una partición posterior (Jancey [1966], Forgy [1965]). Ball y Hall (1967) usaron la media total como el primer centro, entonces se seleccionan los subsiguientes centroides aceptando cualquier dato que se encuentre al menos a alguna distancia especificada, δ.

Una vez que son elegidos los g centroides, los objetos restantes son asignados a los núcleos más cercanos. El método más utilizado es el Método $k-medias$, cuya explicación se da a continuación.

3.2.2 Método de las k - Medias

Mac Queen propuso el término $k-medias$ para describir su algoritmo. Aunque le llamaremos virtualmente algoritmo de $g-medias$, porque en lugar de utilizar el término k, de k grupos, utilizamos g, de g grupos. Una vez que de los g centros de los conglomerados son elegidos, los objetos restantes son asignados al conglomerado cuyo centroide es el más cercano, utilizando alguna medida de proximidad, generalmente la distancia euclideana cuadrada. El centroide se recalcula cada vez que el conglomerado que recibe un nuevo objeto Esto se repite hasta que todos los objetos han sido colocados.

La distancia entre el $i-ésimo$ individuo y el $m-ésimo$ conglomerado esta dado por la siguiente expresión:

$$d_{i,m} = \sum_{k=1}^{p} ([x_{i,k} - \bar{x}_{m,k}]^2)^{1/2}$$

donde $x_{i,k}$ es el valor del $i-ésimo$ individuo sobre la $k-ésima$ variable; $i = 1,2,...,n$, $k = 1,2,...,p$ y $\bar{x}_{m,k}$ es la media de la $k-ésima$ variable en el $m-ésimo$ conglomerado. $P(n,g)$ es la partición que resulta dfe cada uno de los n individuos para asignarse a uno de los conglomerados $1,2,...,g.$,y el número de individuos que pertenecen al $m-ésimo$ conglomerado por n_m.

El error de la partición se define como:

$$E(P_{n,g}) = \sum_{i=1}^{n} d^2_{i,m(i)}$$

donde $m(i)$ es el conglomerado que contiene el $i-ésimo$ individuo, y $d_{i,m(i)}$ es la distancia euclideana entre el individuo i y la media del conglomerado donde esta contenido el individuo. En el proceso de agrupamiento se busca una partición con un error pequeño moviendo individuos de un conglomerado a otro hasta que no se transfiera un individuo que resulte en una reducción en el error.

Ejemplo:

Consideremos 3 nutrientes contenidos en 6 tipos de pescados. En la siguiente tabla se muestran los datos:

tipo de pescado	*energía*	*grasa*	*calcio*	*suma(i)*
mackerel(MC)	5	9	20	34
perca(PR)	6	11	2	19
salmon(SL)	4	5	20	29
sardina(SD)	6	9	46	61
atun(AT)	5	7	1	13
camaron(CA)	3	1	12	16

Tres nutrientes en seis tipos de pescados

Tabla 3.3

Los datos se pueden representar por el elemento $a_{i,k}$ que se refiere al $i-ésimo$ individuo con la $k-ésima$ variable donde $1 \leq i \leq 6$ y $1 \leq k \leq 3$.

Los siguientes pasos muestran la formación de los conglomerados basados en el algoritmo de las k medias

Paso1. Se puede formar un conglomerado inicial considerado el individuo i como parte del $m-ésimo$ conglomerado donde k es la parte integral de $g[(sum(i)-min)]/(max-min)+1$, donde max y min son los valores máximos y mínimos respectivamente de sum(i).

El número de conglomerados es $g = 3$, con $max = 61$ y $min = 13$. Si se aplican las reglas propuestas anteriormente, se tienen los siguientes conglomerados:

Conglomerado	*Elementos*
1	(PR, AT,CA)
2	(MC, SL)
3	(SD)

Paso 2. Ahora calculamos $\bar{x}_{m,k}$, que es la media de la $k-ésima$ variable sobre todos los individuos en el $m-ésimo$ conglomerado. Los valores son los siguientes:

Conglomerado	*energía*	*grasa*	calcio
(PR, AT, CA)	$\frac{14}{3}$	$\frac{19}{3}$	5
(MC, SL)	$\frac{9}{2}$	7	20
(SD)	6	9	46

Paso 3. Se calculan las distancias del $i-ésimo\ individuo$ al $m-ésimo$ conglomerado según la expresión que corresponde a $d_{i,m}$.

Entonces los errores en la partición son:

$$E(P(6,3)) = d_{1,2}^2 + d_{2,1}^2 + d_{3,2}^2 + d_{4,3}^2 + d_{5,1}^2 + d_{6,1}^2$$

$$= (5-\tfrac{9}{2})^2 + (9-7)^2 + (20-20)^2 + (6-\tfrac{14}{3})^2 + (11-\tfrac{19}{3})^2 + (2-5)^2 + (4-\tfrac{9}{2})^2 + (5$$

$$+ (20-20)^2 + (6-6)^2 + (9-9)^2 + (46-46)^2 + (5-\tfrac{14}{3})^2 + (7-\tfrac{19}{3})^2 + (1-5)^2 + (3$$

$$+ (1-\tfrac{19}{3})^2 + (12-5)^2 = 137.805$$

Paso 4. En este paso se verifica si algún movimiento de un individuo de un conglomerado a otro produce una reducción en E (error). Se tiene que calcular el siguiente valor para cada individuo

$$R_{m(i),m} = \frac{n_m d^2_{i,m}}{n_m + 1} - \frac{n_{m(i)} d^2_{i,m(i)}}{n_{m(I)} + 1}$$

n_m = número de individuos en el m – *ésimo* conglomerado
$m(i)$ = i – *ésimo* individuo contenido en el conglomerado

Para el primer individuo tenemos,

$$d^2_{1,1} = (5 - \tfrac{14}{3})^2 + (9 - \tfrac{19}{3})^2 + (20 - 5)^2 = 232.22$$

$$d^2_{1,2} = (5 - \tfrac{9}{2})^2 + (9 - 7)^2 + (20 - 20)^2 = 4.25$$

$$d^2_{1,3} = (5 - 6)^2 + (9 - 9)^2 + (20 - 46)^2 = 677$$

$$R_{2(1),1} = \tfrac{3}{4}(233.22) - 2(4.25) = 166.66 > 0$$

$$R_{2(1),3} = \frac{677}{2} - 2(4.25) = 330.25 > 0$$

Se puede observar que hay un cambio del primer individuo en el segundo conglomerado al primer conglomerado o al tercero y produce un incremento en E por lo que el primer individuo permanecerá en el segundo conglomerado.

Para los individuos 2, 3, 4 y 5, se hacen cálculos similares y el resultado que se obtiene es que deben permanecer en sus conglomerados iniciales. Para el individuo 6, los resultados muestran que hay una reducción en el error, si este individuo se mueve del primer conglomerado al segundo; la reducción es de 52.15. Sin embargo, ahora se encuentra que

$$E(P'_{n,g}) = 137.805 - 52.15 = 85.655$$

donde la partición está conformada por: (PR,AT)(MC,SL,CA)(SD).

Paso 5. Se recalculan los x_{mk}, cuyos valores se encuentran en la tabla 3.5.

Conglomerado	*energía*	*grasa*	*calcio*
(PR, AT)	$\frac{11}{2}$	9	$\frac{3}{2}$
(MC, SL, CA)	4	5	$\frac{52}{3}$
(SD)	6	9	46

Medias de los conglomerados del método de las k – medias

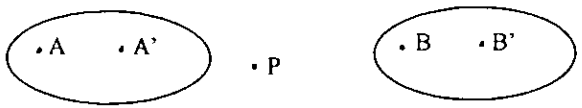
Tabla 3.4

Se hacen las operaciones del paso 4, cuyos valores de los $R^{m(i),m'}$s resultan ser positivos, por lo tanto si no hay más cambios se producirán errores más pequeños. Por tanto, los conglomerados finales quedan como:

Conglomerado	*Elementos*
1	(PR, AT)
2	(MC, SL, CA)
3	(SD)

3.2.3 Criterio de agrupamiento: Reasignación de los objetos

El siguiente paso consiste en la búsqueda de objetos que deberían de ser reasignados a otros conglomerados. Por ejemplo, en la figura 4.7, el punto P es más cercano al núcleo B que al núcleo A; pero una vez que los conglomerados son formados, P es más cercano al nuevo núcleo A' que al nuevo núcleo B'. Cada individuo es obsevado cuidadosamente y reasignado si causa un incremento, o decremento en el caso de la minimización, en el valor de un criterio de agrupación dado. Se repite este proceso a todos los individuos hasta que no se obtiene mejoría al mover un solo objeto, hasta que se alcanza un óptimo local del criterio .



Reasignación de objetos

Gráfica 3.7

Forgy (1965) sugiere un procedimiento donde cada objeto es considerado y reasignado si este es más cercano al centroide de otro conglomerado. Despúes de considerar todos los objetos, los centroides son actualizados y el proceso se repite hasta que ningún objeto cambia su conglomerado. Colecciones de métodos y sus variantes se describen usualmente como algortimos *k – medias* (*g – medias*).

Otros criterios de agrupamiento has sido propuestos, pero los más usados comunmente son los que están basados en la minimización y maximización dentro de grupos y entre grupos respectivamente (métodos basados en la traza).

3.2.3.1 Métodos basados en la traza

Estos se basan en las siguientes matrices de p x p donde p es el número de variables:

$$T = \frac{1}{n}\sum_{m=1}^{g}\sum_{i=1}^{n_m}(x_{im} - \bar{x})(x_{im} - \bar{x})'$$

$$W = \frac{1}{n-g}\sum_{m=1}^{g}\sum_{i=1}^{n_m}(x_{im} - \bar{x}_l)(x_{im} - \bar{x}_m)'$$

$$B = \sum_{m=1}^{g} n_m(\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})'$$

donde T es la matriz de dispersión total, W es la matriz de dispersión dentro de grupos y B es la matriz de dispersión entre grupos y satisfacen la siguiente ecuación:

$$T = W + B$$

Para $p = 1$ esta ecuación representa una relación entre escalares; simplemente la división de la suma de cuadrados totales para una variable, la suma de cuadrados entre y dentro de grupos. Para $p > 1$ el criterio de agrupamiento no esta definido, por lo que se dan algunas otras alternativas.

3.2.3.2 Minimización de la traza W

Una extensión del criterio de minimización de la suma de cuadrados dentro de grupos, mencionado anteriormente para el caso $p = 1$, cuando los datos no son univariados, es minimizar la traza de la matriz de suma de cuadrados y productos cruzados dentro de grupos. Se puede demostrar que esto es equivalente a minimizar la suma de distancias euclideanas al cuadrado entre individuos y la media del conglomerado, esto es:

$$E = \sum_{i=1}^{N} d^2_{i,m(i)}$$

Por oto lado, minimizar W es equivalente a maximizar B, ya que

$$tr(T) = tr(W) + tr(B)$$

3.2.3.3 Minimización del determinante de W

Dentro del análisis de varianza multivariada, una de las pruebas para diferencias de vectores medios dentro de grupos, está basado en la razón de los determinantes de las matrices de dispersión total y dentro de grupos.Valores grandes de $|T|/|W|$ indican que los vectores medios de grupos difieren, por lo que Friedman y Rubin (1967) sugirieron como criterio de agrupamiento la maximización de este cociente. Puesto que para todas las particiones de los n individuos en los g grupos, T permanece igual, la maximización de $|T|/|W|$ es equivalente a la minimización de $|W|$.

3.2.3.4 Maximización de la traza BW^{-1}

Este criterio propuesto por Friedmman y Rubin (1967) maximiza la traza de la matriz que se obtiene del producto de la matriz de dispersión entre grupos y la inversa de la matriz dentro de grupos. Este criterio se puede expresar en terminos de los valores propios $\lambda_1, \lambda_2, ..., \lambda_p$, de BW^{-1}, esto es:

$$tr(BW^{-1}) = \sum_{i=1}^{p} \lambda_i$$

El uso de un algortimo partitivo requiere de la elección de un criterio de agrupamiento; el uso de criterios propuestos restringiran la forma de los conglomerados formados, por ejemplo, utilizar el criterio de $tr(W)$ significa que los conglomerados revelados serán de una variedad esférica. En contraste, el criterio W no restringe a los conglomerados a ser esfericos, pero supone que todos los conglomerado tienen la misma forma. Por esta razón, Scott y Symon (1971) sugieren que sea utilizado $\Pi_{i=1}^{k}|W_i|^{n_i}$ (donde n_i es el número de individuos en el grupo i) en lugar de $|W|$.

3.2.4 Optimización del Criterio de Agrupamiento: Número de particiones

Una vez que un criterio de agrupamiento se ha elegido y suponiendo que el número de conglomerados g ya esta determinado, el procedimiento natural es particionar los n objetos en g conglomerados para optimizar algún criterio. El número de posibles particiones de los objetos es:

$$P_{n,g} = \frac{1}{g!}\sum_{i=1}^{g}\binom{g}{i}(-1)^{g-i}\,i^n \sim \frac{g^n}{g!} \qquad cuando\ n \to \infty$$

Esta aproximación es impractica ya que $P_{n,g}$ es demasiado grande, incluso para valores pequeños de g. incluso con la capacidad de las computadoras hoy en día, los números que se obtienen son muy grandes, que la enumeración completa de cada posible partición de los n objetos en los g grupos no es posible. Estos ejemplos ilustran el grado del problema,

$$P_{15,3} = 2,375,101$$

$$P_{20,4} = 45,232,115,901$$

$$P_{25,8} = 690,223,721,118,368,580$$

la situación es peor si g no está especificado, ya que el número total de posibles particiones es: $\sum_{g=1}^{n} P_{n,g}$.

Lo imprácticо de analizar cada una de las particiones ha llevado al desarrollo de algoritmos diseñados para buscar el valor óptimo de un criterio de agrupamiento re-ordenando las particiones que ya existen y manteniendo las nuevos solo si proporcionan una mejoría; estos algoritmos se llaman "hill climbing", aunque en el caso de criterios que requieren minimización se deberían de llama "hill descending".

Las técnicas jerárquicas se utilizan para producir una partición con un número g de conglomerados pero una debilidad de esos métodos es que una fusión o división incorrecta en una etapa no se puede corregir posteriormente. Sin embargo los métodos no jerárquicos permiten que los objetos sean reasignados. Estas técnicas de reasignación se pueden aplicar a los metodos divisivos pero no a los aglomerativos.

3.3 ELECCIÓN DEL NÚMERO DE CONGLOMERADOS

Un problema del análisis de conglomerados es la elección del número de conglomerados, g. El investigador o analista tendrá que "estimar el número de conglomerado en los datos". Existe una variedad de métodos que son útiles en situaciones particulares. La mayoría son relativamente informales y consisten en analizar las diferencias entre los niveles de fusión en el dendograma. Los cambios grandes son generalmente tomados para indicar un número particular de grupos. Los procedimientos para analizar dendrogramas pueder ser muy subjetivos.

Se han descrito técnicas que tratan de resolver el problema de la subjetividad. Beale (1969), propone una prueba F que se usa para probar si una subdivisión en g_2 conglomerados es significativamente mejor que una subdivisión en algún número más pequeño de conglomerados g_1. La estadistica se define de la siguiente manera:

$$F(g_1,g_2) = \frac{\dfrac{R_{g_1} - R_{g_2}}{R_{g_2}}}{\left[\left\{\dfrac{n-g_1}{n-g_2}\right\}\left(\dfrac{g_1}{g_2}\right)^{2/p} - 1\right]}$$

donde $R_g = (n-g)S_g^2$ y S_g^2 es la desviación media cuadrada de los centros de los conglomerados en la muestra. La estadistica es comparada con F, con $p(g_2 - g_1)$ y $p(n-g_2)$ grados de libertad. Si se obtiene un resultado significativo indica que una subdivisión en g_2 conglomerados es una mejoría sobre una subdivisión en un número más pequeño g_1. De acuerdo a experiencias anteriores se ha visto, que este procedimiento es útil solo si los conglomerados están moderadamente separados y con una forma esférica aproximadamente.

Calinsgi y Harabasz (1974) proponen un indice C que esta basado en la suma de cuadrados.

$$C = \frac{tr(B)/(g-1)}{tr(W)/(n-g)}$$

donde B y W es la suma de cuadrados dentro y entre conglomerados y las matrices de los productos cruzados, y g es el número de grupos. El máximo valor de C en la jerarquía se toma como indicador del número correcto de grupos.

Mojena (1977) propone un procedimiento basado en los tamaños relativos de los diferentes niveles de fusión en el dendrograma. El próposito es seleccionar un numero de grupos correspondiente a la primera etapa en el dendograma satisfaciendo,

$$\alpha_{j+1} > \bar{\alpha} + ks_{\alpha}$$

donde $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, son los niveles de fusión que corresponden a las etapas con $n, n-1, \ldots, 1$ conglomerados. El término $\bar{\alpha}$ y s_{α} son la media y la desviación estándar de los valores de α respectivamente, y k es una constante. Este autor sugiere que los valores de k que se encuentran en el rango de 2.75 a 3.50 dan los mejores resultados. Sin embargo, Milligan y Cooper (1985) sugieren que el valor de k como regla de Mojena debería ser 1.25.

Cualquier conceptualización que sugiera un número de conglomerados debe complementarse con el juicio empírico del analista. Se puede empezar este proceso especificando algún criterio basándose en consideraciones prácticas, como decir, "los resultados serán más manejables y más fáciles de manejar si se tengo de 3 a 6 conglomerados, y a continuación resolver para este número de conglomerados y seleccionar la mejor alternativa después de evaluar todas ellas. Las soluciones que se obtengan tendrán una mejoría mediante la restricción de la solución de acuerdo con los aspectos conceptuales del problema.

3.4 PROPIEDADES Y PROBLEMAS DE LAS TÉCNICAS JERÁRQUICAS AGLOMERATIVAS

Algunos de los métodos que se describieron anteriormente son propensos a formar conglomerados esfericos incluso cuando los datos contienen conglomerados de otras formas. Jardine y Sibson (1971) consideraron que los métodos de agrupamiento jerárquicos deberían de satisfacer algunas condiciones matemáticas. Lo que esos autores muestran es que el método de agrupamiento, el cual transforma un coeficiente de disimilaridad en un dendrograma jerárquico se puede ver como un método con lo cual la desigualdad ultramétrica puede ser impuesta sobre el coeficiente que originalmente puede haber satisfacido solo la desigualdad triángulo. Esta transformación tiene que satisfacer ciertas condiciones. Una condición fundamental es la de continuidad , la cual dice que " pequeños cambios en los datos deberían producir pequeños cambios en el árbol resultante y pequeños cambios en las disimilaridades se pueden producir si hay errores en los datos. También en algunos métodos de agrupamiento, pequeños cambios en los datos pueden producir grandes cambios en los árboles que se obtengan. Otro conjunto de condiciones importantes son las llamada "fitting together", esto es, si se sustrae o se agrega un individuo al conjunto original, se espera que la estructura del árbol cambie relativamente poco, aunque la clasificación algunas veces cambie de una manera no trivial. Jardine y Sibson recomiendan también como el mejor método el que recurra más a lo matemático. El método de la liga simple es el que cumple con esto y con las condiciones antes propuestas, además de que es el más rápido para los cálculos en computadoras y da soluciones que son invariantes en estructura bajo una transformación de las medidas de disimilaridad. Este método es bueno para datos que tienen significancia solamente ordinal.

El método de la liga simple no ha sido aceptado universalmente porque en muchas aplicaciones no produce soluciones útiles. Tiene como desventaja el efecto de encadenamiento, que consiste en la existencia de puntos intermedios entre las agrupaciones, es decir puntos que conectan a los conglomerados. Esta propiedad puede causar que el método no logre una distinción de los conglomerados cuando hay un número pequeño de individuos dentro de ellos. El encadenamiento es visto generalmente como un defecto de la liga simple, aunque llamarlo así, es engañador (Jardine y Sibson), más bien, es una simple descripción de lo que es método hace. Se han propuesto alternativas para resolver el problema del encadenamiento pero se producen otros problemas como la falta de continuidad. Por otro lado, autores como Gower (1988) siente que las condiciones de Jardine y Sibson son demasiadas estrictas.

El método de la mediana también tienen la propiedad del encadenamiento. Los métodos de la liga completa y liga promedio son poco atrayentes por la falta de continuidad, aunque el primero tiene la ventaja de que es invariante bajo trnsformaciones monotónicas de la matríz de proximidad, lo cual significa que el método dará el mismo resultado sobre otras matrices de proximidad, cuyos elementos están en el mismo orden que los originales, solamente las propiedades ordinales de las medidas de disimilaridad o similaridad son consecuencia. También el método de la liga simple presenta esta ventaja.

Por otro lado, se han hecho numerosas investigaciones sobre las técnicas de agrupamiento jerárquicas y han mostrado cuales métodos son los más útiles en la práctica. Por ejemplo Baker (1974) y Hubert (1974) muestran que el método de la liga completa es menos sensible a errores observacionales que el de la liga simple. Cuningham y Ogilvie (1972) comparan siete técnicas jerárquicas y encuentran que el método del grupo promedio se desarrolla más satisfactoriamente en el conjunto de datos, además de que existe una fuerte interacción en los resultados entre los datos de entrada y el método de agrupamiento utilizado. Kiper y Fisher (1975) investigan seis técnicas jerárquicas y encuentran que para tamaños de mustra iguales de distribuciones normal multivariada, el método de Ward es el más adecuado, y para tamaños de muestras distintos, el método del centroide, el del grupo promedio y el de la liga completa son los más exitosos. El estudio de Milligan (1980) muestra que el método de la liga simple, el método del centroide y el de la mediana no son afectados virtualmente; el método de Ward y el grupo promedio se desarrollan probremente. Además cuando los datos contienen una estructura de agrupamiento verdadera, el método del centroide, liga simple y el de la mediana dan resultados pobres; y el método de Ward y el del grupo promedio tienen más exito. En un estudio hecho por Everritt y Hands (1987) el método de Ward presenta una buena ejecución cuando los datos son binarios y cuando los datos contienen conglomerados de tamaños casi iguales, pero pobremente cuando los conglomerados son de diferentes tamaños en donde el método del centroide fue el que dio resultados más satisfactorios. Como se ha visto, el método preferido para Jardine-Sibson es el de la liga simple, sin embargo otros autores como Marriot recomienda el método de Wishart como un método efectivo y poco propenso para dar resultados engañosos .

El éxito o fracaso de un método sobre datos particulares puede radicar en el tipo de datos que se tenga más que en el criterio teórico. Como se ha visto, se han realizado comparaciones de los métodos con datos generados artificialmente, pero los resultados obtenidos son inconsistentes. Algunos métodos trabajan mejor en ciertos tipos de datos que en otros. Estudios empíricos enfatizan el hecho de que la mayoría de los métodos hacen supuestos implícitos en la muestra de datos, si estos supuestos no se sostienen, entonces la técnica de agrupamiento impone una estructura en los datos más que encontrarla; de esta manera se pueden encontrar soluciones espurias.

Es difícil hacer recomendaciones sobre el método más adecuado, sin embargo, el de la liga simple podría ser el más recomendable, aunque también se sugiere tratar más de un metodo. Algunas veces es recomendable utilizar varios de ellos, y si todos revelan los mismos agrupamientos, entonces se puede tener confianza en que las agrupaciones naturales realmente existen. Sin embargo, estudios empíricos han demostrado que es difícil que suceda, a menos que los grupos sean de forma esférica y bien separados.

Una alternativa es utilizar los métodos jerárquicos y no jerárquicos para obtener beneficios de cada uno. La técnica jerárquica puede establecer el número de conglomerados, los perfiles de los centros de los conglomerado y los datos atípicos. Una vez que se han eliminado los atípicos, las observaciones restantes se pueden agrupar con un método no jerárquico con los centros de conglomerados desde los resultados jerárquicos como los puntos iniciales. De esta manera, las ventajas de los métodos jerárquicos se complementan con la capacidad de los no jerárquicos para ajustar los resultados permitiendo el cambio de pertenencia a un conglomerado.

3.5 ANÁLISIS DE CONGLOMERADOS CON ANÁLISIS DE COMPONENTES PRINCIPALES

Existen métodos que se utilizan para obtener puntajes que resuman la información de los datos u observaciones de interés, como pueden ser el análisis factorial, el escalamiento multidimensional y el análisis canónico, sin embargo el más usual es el análisis de componentes principales.

El Análisis de Conglomerados algunas veces inicia con un análisis de componentes principales para reducir un número grande de variables originales a un número pequeño de variables llamado componentes principales. Esto puede reducir drásticamente el tiempo de cálculo para el análisis de conglomerados, además de que la interpretación de cada conglomerado es mucho más fácil. Sin embargo, se sabe que los resultados de un Análisis de Conglomerados puede ser muy diferentes con o sin el Análisis de Componentes Principales inicial. Por lo que éste análisis es mejor evitarlo. Por otro lado, cuando los 2 primeros componentes cuentan con un alto porcentaje de variación en los datos, una forma útil para buscar agrupaciones, es una gráfica de los objetos contra esos 2 componentes.

La característica principal del Análisis de Componentes Principales es describir la variación de un conjunto de datos multivariados en términos de un conjunto de variables no correlacionadas llamadas componentes principales que son combinaciones lineales de las variables originales, es decir es la transformación de las variables originales $x_1, x_2, \ldots\ldots x_p$ que son correlacionadas en un nuevo conjunto de variables $z_1, z_2, \ldots\ldots, z_p$ que son combinaciones lineales de las originales y son no correlacionadas. Las nuevas variables son derivadas en orden de importancia decreciente, de modo que la primera componente principal explica la mayor cantidad de variación de las variables originales.

El objetivo de este método es ver si unos cuantos de las primeras componentes principales explican la mayor parte de la variabilidad de los datos originales, si esto se logra, las componentes principales se usan para resumir los datos con poca pérdida de información proporcionando una reducción en la dimensión de los datos. Es de esperarse que unos cuantos de las primeros componentes principales serán intuitivamente significativos, ayudarán a entender los datos de una mejor manera y serán útiles en análisis posteriores donde se pueda trabajar con un número menor de variables. El análisis de componentes principales no siempre puede lograr su objetivo de reducción de dimensionalidad en el número inicial de variables, ya que puede suceder que las variables originales no se encuentren correlacionadas entre ellas. Así los mejores resultados se obtienen cuando las variables iniciales se encuentran altamente correlacionadas entre ellas, ya sea negativa o positivamente.

El Análisis de Componentes Principales está basado en un conjunto de datos de n individuos en p variables.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Suponemos que $x' = [x_1, x_2, \ldots\ldots, x_p]$ es una variable aleatoria con media μ y matriz de varianzas y covarianzas Σ.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

donde los elementos de la diagonal σ_{ii} son las varianzas de x_i y σ_{ij} son las covarianzas entre los pares de variables x_i y x_j.

El objetivo es encontrar un conjunto de nuevas variables $z_1, z_2, \ldots, z_p$, las cuales son no correlacionadas y cuyas varianzas decrecen desde la primera hasta la última, cada z_i es una combinación lineal (ver apéndice A, A4) de las x s tal que:

$$z_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ip}x_p = a_i'x$$

donde $a_i' = [a_{i1}, a_{i2}, \ldots, a_{ip}]$ es un vector de constantes.

La primera componente principal es la combinación lineal

$z_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$ tal que su varianza es la más grande posible sujeta a la restricción de normalización:

$$\sum_{k=1}^{p} a_{1k}^2 = a_{11}^2 + a_{12}^2 + \ldots + a^2{}_{1p} = 1$$

El objetivo consiste en encontrar el vector de pesos o coeficientes $a_1 = (a_{11}, a_{12}, \ldots, a_{1p})$ que maximiza la varianza de z_1.

La segunda componente principal:

$z_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$. Se desea encontrar a_2 tal que que maximize la varianza de z_2, es decir tal que $var(z_2)$ sea lo más grande posible (de la variabilidad restante de los datos) no mayor a la de z_1 sujeta a la restricción $\sum_{k=1}^{p} a_{2k}^2 = 1$ y con la condición de que z_1 y z_2 sean no correlacionadas.

$$\begin{aligned} Cov(z_2, z_1) &= cov(a_2'x, a_1'x) \\ &= E[(a_2'(x-\mu)(x-\mu)'a_i] \\ &= a_2'\Sigma a_1 \end{aligned}$$

Las demás componentes $z_3, z_4, \ldots, z_p$ son definidas de la misma manera, son no correlacionadas y tienen varianza decreciente.

$$var(z_1) \geq var(z_2) \geq \ldots \geq var(z_p)$$

La restricción $\sum_{k=1}^{p} a_{ik}^2 = 1$ es introducida porque de otro modo la varianza de z_i podría ser aumentada simplemente al aumentar los valores de los coeficientes a_{ik}. En otras palabras se debe a que no se desea incrementar de forma arbitraria la varianza original de los datos.

Lo que se desea encontrar es a_i tal que maximice la varianza de z_i sujeto a la condición $a_i'a_i = 1$, es decir

$$Var(z_i) = Var(a_i'x) = a_i' \Sigma a_i$$

donde $a_i' \Sigma a_i$ es la función objetivo. El procedimiento para maximizar una función con variables, sujeto a una o más condiciones es el método de multiplicadores de Lagrange, aplicando este método se llega a que el valor de a_i que maximiza a la $Var(z_i)$ debe satisfacer $\Sigma a_i = \lambda_i a_i$, es decir:

$$(\Sigma - \lambda_i I)a_i = 0$$

este es un conjunto homogéneo de p ecuaciones con p incógnitas y para una solución no trivial $a_i \neq 0$ se requiere:

$$| \Sigma - \lambda_i I | = 0$$

donde $(\Sigma - \lambda I)$ es una matriz singular de pxp. Por lo tanto λ_i es un valor propio de Σ (ver apéndice A,A5) y la solución a_i es su correspondiente vector propio. Σ tendrá p valores propios (ver apéndice A,A5), como Σ es una matriz semidefinida positiva (ver apéndice A, A2), los valores propios no pueden ser negativos.

Las varianzas de las componentes principales son los valores propios λ_i de la matriz Σ los cuales pueden ordenarse como $\lambda_1 \geq \lambda_2 \geq \ldots\ldots\ldots \geq \lambda_P \geq 0$ y asociarse a cada uno de los componentes, esto es λ_i corresponde a la i-ésimo componente principal.

Tenemos:

$$\begin{aligned} Var(z_i) = Var(a_i'x) &= a_i' \Sigma a_i \\ &= a_i' \lambda_i I a_i \\ &= \lambda_i \end{aligned}$$

a_i es el vector propio de Σ asociado al valor propio λ_i. Algunos de los valores propios de Σ pueden ser iguales pero los vectores propios asociados a las raices (valores propios) deben de ser ortogonales.

A es la matriz de (pxp) de vectores propios donde:

$$A = [a_1, a_2, \ldots, a_p]$$

y z es un vector de $px1$ de componentes principales. Entonces:

$$z = A'x$$

La matriz de varianzas y covarianzas *pxp* de *z* es denotada de la siguiente manera:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Esta matriz es diagonal ya que las componentes han sido elegidos para ser no correlacionados.

Utilizando la siguiente expresión:

$$\Lambda = A' \Sigma A$$

se muestra la importancia de la relación entre la matriz de covarianza de x y los correspondientes componentes principales. La ecuación anterior se puede reescribir como sigue:

$$\Sigma = A \Lambda A'$$

donde A es una matriz ortogonal $AA' = I$

Dado que los valores propios son interpretados como las varianzas respectivas de las distintas componentes principales. La suma de las varianzas está dada como:

$$\sum_{i=1}^{p} Var(zi) = \sum_{i=1}^{p} \lambda_i = traza(\Lambda)$$

pero

$$\begin{aligned} traza(\Lambda) &= traza(A' \Sigma A) \\ &= traza(\Sigma A' A) \\ &= traza(\Sigma) \\ &= \sum_{i=1}^{p} Var(x_i) \end{aligned}$$

es decir:
$Traza(\Sigma) = \sigma_{11} + \sigma_{22} + \dots\dots\dots + \sigma_{pp} = \lambda_1 + \lambda_2 + \dots\dots\dots + \lambda_P$

esto significa que la suma de las varianzas de las variables originales es igual a la suma de las varianzas de las componentes principales por lo que las componentes se quedan con toda la variación de los datos originales.

La $i-ésima$ componente principal cuenta con una proporción de $\lambda_i / \sum_{i=1}^{p} \lambda_i$ sobre la variación total de los datos originales. La proporción de varianza explicada de las primeros m componentes sobre la variación total es:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Para evitar problemas de influencia por escala natural de las variables, algunos autores recomiendan antes de llevar a cabo el análisis de componentes principales, estandarizar las variables $x_1, x_2, \ldots, x_p$, de manera que tengan media cero y varianza unitaria.

$$x_i = \frac{x - \bar{x}}{\sigma}$$

De esta manera las componentes principales se calculan sobre una matriz de correlación P de las $x's$.

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

donde $\rho_{ij} = \rho_{ji}$ es la correlación entre x_i y x_j y

$Traza(P) = \rho_{11} + \rho_{22} + \ldots\ldots\ldots + \rho_{pp} = 1 + 1 + \ldots\ldots\ldots + 1 = p$ donde p es el número de variables.

Una de la decisiones que se deben de tomar cada vez que se realiza un análisis de componentes principales es saber cuantas componentes principales se tienen que elegir de tal manera que los datos se resuman efectivamente. Los siguientes pasos sirven de guía para elegir el número óptimo de estos:

- Elegir a las componentes que acumulen un 80% de varianza, es decir eliminar a los componentes que contengan solo un pequeña porción de la variabilidad de las $x's$. Por ejemplo si se tiene un análisis donde $p = 20$ y las tres primeras componentes suman un 80% de la variabilidad de los datos, las 17 restantes se pueden eliminar, ya que contendrán muy poca información de los datos originales, solo 20% de la variabilidad.

- Excluir a las componentes cuyos valores propios sean menor que el promedio de todos los valores propios,

$\sum_{i=1}^{p} \lambda_i/p$. Para una matriz de correlación este promedio es uno.

- Utilizar una grafica de λ_i contra i. Aquí se puede observar un codo que divide los valores propios grandes y los valores propios pequeños.

- Probar la significancia de los componentes más grandes los cuales corresponden a vectores propios más grandes (cuando se ha incorporado la hipótesis de normalidad).

CAPITULO 4

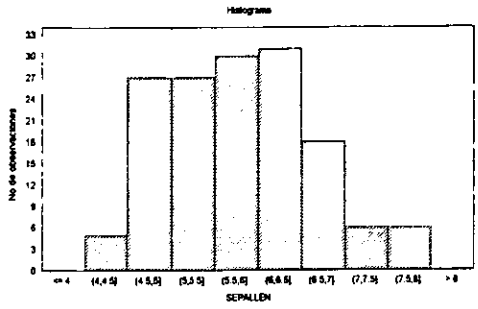
APLICACIÓN

4.1 INTRODUCCIÓN

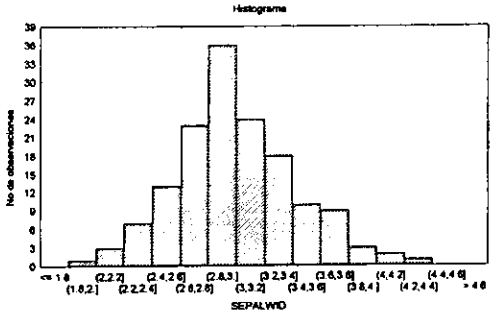
Este capítulo tiene como finalidad aplicar algunos de los métodos jerárquicos aglomerativos y los métodos no jerárquicos. Primero se realizará un análisis gráfico exploratorio para ver el comportamiento de las variables y la posible presencia de conglomerados como un análisis previo de los datos. También se aplicará el Análisis de Componentes Principales. Los datos a los que se les realizará el análisis, son las observaciones de Iris de Fisher que contienen medidas sobre la longitud del sépalo (sepalen) y del pétalo (petalen), y del ancho del sépalo (sepalwid) y del pétalo (petalwid), de 150 especies de plantas iris de las cuales se desprenden tres tipos: iris setosa, iris versicolor e iris virginica. Estos datos están medidos en centímetros y se encuentran en el Apéndice B. El paquete que se ocupa es el Statistica y el Statgraphics.

4.2 ANÁLISIS EXPLORATORIO

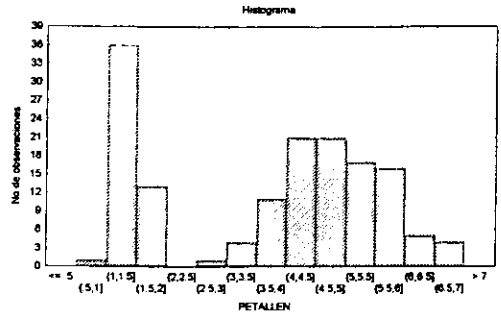
La finalidad de este análisis gráfico, es mostrar la presencia de posibles grupos sin haber aplicado ninguna de las técnicas que ofrece el Análisis de Conglomerados. Este examen es a través de representaciones univariadas, bivariadas y multivariadas. Primero se obtendrán las distribuciones de cada una de las variables a través de los histogramas. En este ejercicio no es necesario estandarizarlas ni ponderarlas, ya que las variables están medidas en las mismas unidades.



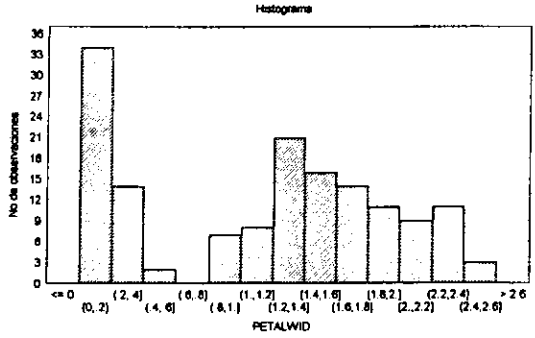
Fuente: Datos iris
Histograma de la variable sepallen
Gráfica 4.1



Fuente: Datos iris
Histograma de la variable sepalwid
Gráfica 4.2



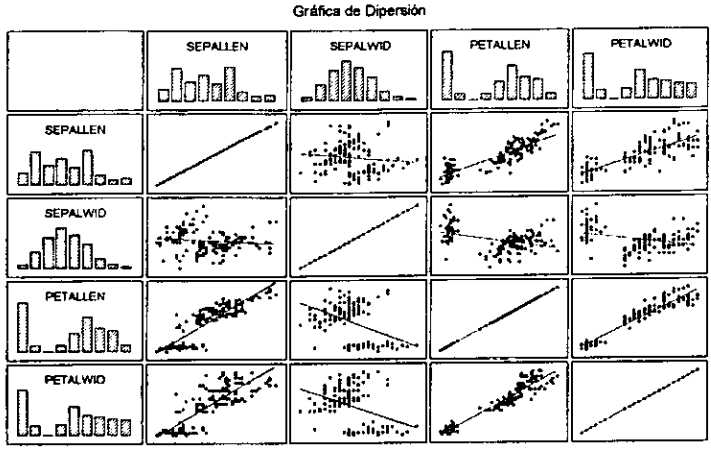
Fuente: Datos iris
Histograma de la variable petallen
Gráfica 4.3



Fuente: Datos iris
Histograma de la variable petalwid
Gráfica 4.4

De acuerdo a estas gráficas se observa que el ancho (petalwid) y longitud del pétalo (petallen) presentan una distribución multimodal, lo cual se podrían tomarse como una evidencia preeliminar de que esta variable es la que proporciona más información para distinguir a los grupos.

Otra gráfica que puede resultar también muy útil para ver la relación que existe entre las variables y encontrar la presencia de conglomerados es la gráfica de dispersión.



Fuente: Datos iris
Gráfica de dispersión
Gráfica 4.5

En esta gráfica, visualmente se nota que las variables petallen y petalwid presentan una forma lineal, esto indica que ambas variables tienen una correlación alta, y que flores con pétalos largos están relacionados con flores de pétalos anchos. Este mismo comportamiento se nota también con las variables sepallen y petallen, pero en menor magnitud. En el siguiente cuadro se muestran los valores de las correlaciones que corroboran lo anterior.

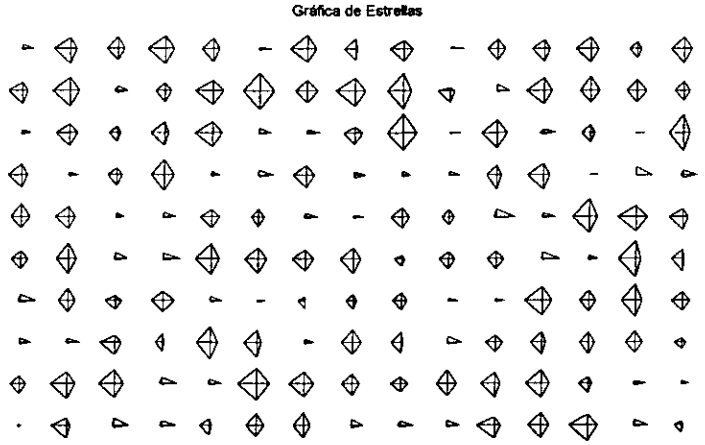
	Sepallen	Sepalwid	Petallen	Petalwid
Sepallen	1.00	-.12	.87	.82
Sepalwid	-.12	1.00	-.43	-.37
Petallen	.87	-.43	1.00	.96
Petalwid	.82	-.37	.96	1.00

Fuente: Datos iris
Correlaciones de las variables
Tabla 4.1

La gráfica de dispersión también muestra que hay concentraciones de datos y de forma separada, las variables petallen y petalwid; y petallen y sepallen presentan

claramente éste comportamiento, y como habíamos visto en las gráficas univariadas, éstas variables pueden ser relevantes en la formación de grupos.

Hemos visto que los histogramas y la gráfica de dispersión mostraron la presencia de conglomerados, sin embargo esto se realizó de manera univariada y bivariada. Para finalizar con el análisis exploratorio, se presenta la siguiente grafica multivariada que permite observar de manera conjunta todas las observaciones con sus respectivas variables.



Fuente: Datos iris
Gráfica de Estrellas
Gráfica 4.6

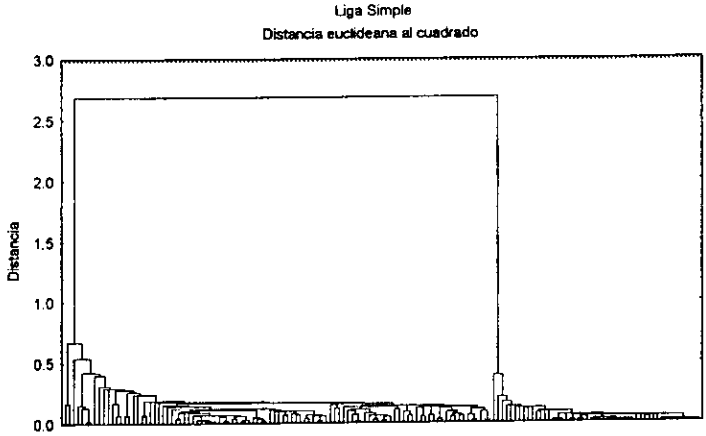
En esta representación se puede ver también la presencia de grupos por la forma que tiene cada una de las estrellas. Con todas las gráficas anteriores ya tenemos suficiente evidencia de que existen grupos de observaciones similares.

4.3 ANÁLISIS DE CONGLOMERADOS

Se ha mencionado que no hay una regla para la elección de la medida de proximidad y del algoritmo que agrupe a los objetos, aunque algunos de ellos tienen ventajas que otros no tienen. Aplicaremos el método de la liga simple porque es considerado el más rápido para los cálculos, es invariante bajo transformaciones monotónicas en la matriz de proximidad y fue el más recomendado por Jardine y Sibson (capítulo 3, sección 3.4); sin embargo, tiene como desventaja el encadenamiento y en algunas aplicaciones no produce soluciones útiles de conglomerados. A pesar de las ventajas y desventajas que presenta este

método, es de gran utilidad hacer una comparación de esta técnica con otra para confrontar los resultados. El siguiente algoritmo de agrupación a utilizar es el método Ward por ser un método que tiende a formar grupos pequeños. La medida de proximidad será la distancia euclideana al cuadrado por ser la más usual.

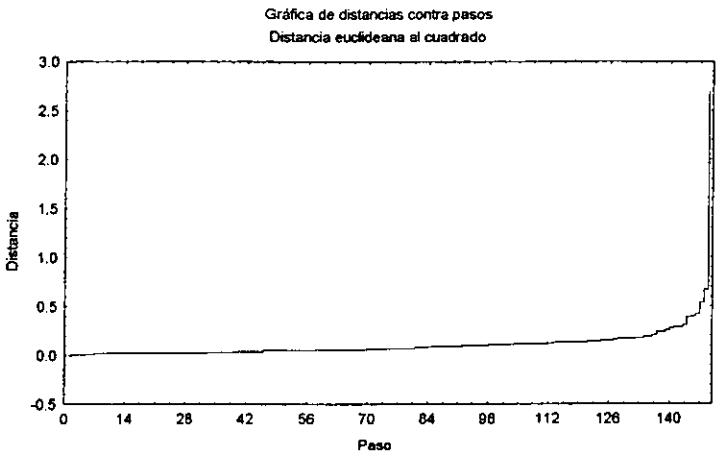
En el siguiente dendrograma que pertenece a la liga simple, se visualizan dos grandes grupos de los tres existentes (iris versicolor, iris setosa e iris virginica), lo cual se puede deber a que existen dos de ellos muy similares. Si se decidiera elegir tres grupos, según la gráfica 4.7, uno de los conglomerados solo tendría dos observaciones, los objetos 3 y 64; sin embargo, esto no sería útil para fines prácticos. Por lo que es mejor quedarse con dos.



Fuente: Datos iris
Dendrograma del Método de la Liga Simple
Gráfica 4.7

El dendrograma y la gráfica de la tabla de amalgamación (gráfica 4.8), son dos formas de ver la formación de conglomerados, pero el segundo muestra que tan natural se unen los objetos, además señala paso a paso a que distancia se van formando los conglomerados. Cada uno de los objetos inicialmente esta formando un conglomerado y de acuerdo a los cálculos de la tabla de amalgamación que se encuentra en el apéndice B, tabla B2, se observa que en el primer paso se unen los conglomerados 18, 67 y 42 a una distancia de .05; en el segundo paso, también a una distancia de .05, se unen los conglomerados 13 y 27, aquí se nota que los elementos son demasiados parecidos porque durante los nueve primeros pasos se unen elementos a esa distancia, después se fusionan a una distancia de 0.059, 0.060, realmente aquí no hay una variación tan grande en las distancias, esto nos sigue indicando que los objetos son muy parecidos. Posteriormente los

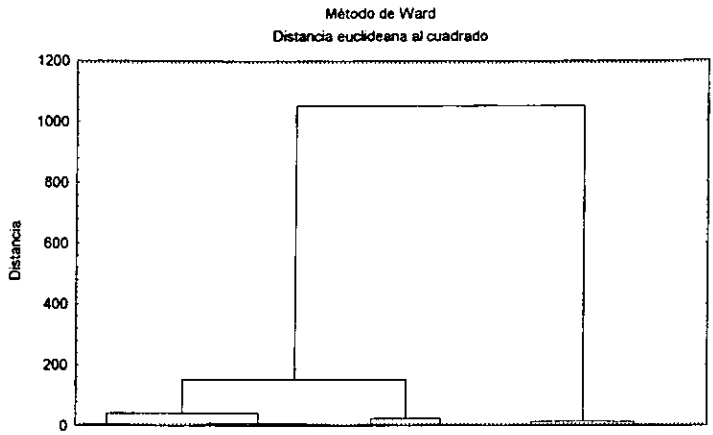
elementos se fusionan a una distancia de 0.069, aquí ya se ve mayor amplitud en la distancia respecto a las dos anteriores. Las siguientes distancias de unión son 0.070, 0.079, .080, .090, 099, .10. hasta 2.6 donde se unen todos los objetos. Esto gráficamente se corrobora con la gráfica 4.8, se observa que hasta el paso 140 aproximadamente hay cambios grandes en las distancias.



Fuente: Datos iris
Gráfica de la Tabla de amalgamación del Método de la Liga Simple
Gráfica 4.8

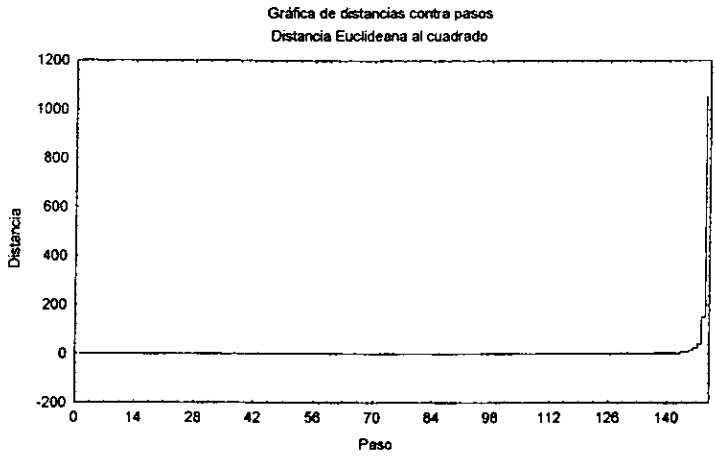
Dado que es mejor quedarse con dos conglomerados, en el apéndice B, tabla B4 se muestran que elementos conforman cada uno de ellos, en el primero hay 50 objetos y en el segundo 100.

Ahora se aplicará el método de Ward. Primero se obtendrá el dendrograma y posteriormente la gráfica de la tabla de amalgamación.



Fuente: Datos iris
Dendrograma del Método de la Liga Simple
Gráfica 4.9

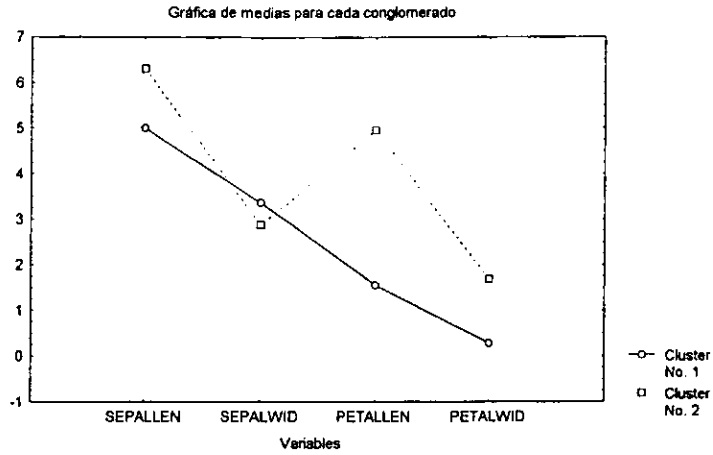
En este dendrograma se notan más conglomerados a simple vista, y las distancias a las que se unen resultan ser más pequeñas que en el método anterior. Según la tabla de amalgamación B.3, del apéndice B, en el primer paso, se unieron los conglomerados 16 y 75 a una distancia de .00, en el siguiente paso a una distancia de. 010 con los conglomerados 95 y 106. Así sucesivamente hasta que todos los objetos queden en un solo conglomerado.



Fuente: Datos iris
Gráfica de la Tabla de amalgamación del Método de Ward
Gráfica 4.10

Lo que tiene éste método es que los primeros conglomerados se unen a una distancia más pequeñas pero los últimos conglomerados se fusionan a distancias muy grandes. Esto se puede notar en la escala, en el de Ward los conglomerados se unen hasta una distancia de 1052 donde se encuentran todos los elementos, y en el de la liga simple la máxima es de 2.6. En este método como se mencionó anteriormente, se forman más grupos, si se tomarán tres grupos, el primero estaría conformado por 50 elementos, el segundo por 36 y el tercero por 64. Los resultados son muy diferentes a lo obtenido por la liga simple. Sin embargo, si se tomarán dos conglomerados, tendría el mismo número de elementos que el método de la liga simple, 50 y 100. En la tabla B5 se encuentran los elementos pertenecientes a cada conglomerado para Ward.

El número de conglomerados finales lo decide el analista; sin embargo, se puede aplicar el método de las k-medias para hacer una comparación con los métodos anteriores. Cabe mencionar que en este método se debe de conocer el número de grupos, en este caso es 2.



Fuente: Datos iris
Gráfica de las k-medias
Gráfica 4.11

En la gráfica 4.11 se nota que los centros de cada conglomerado son muy distantes para las variables petallen y petalwid, y que como lo habíamos notado en el análisis exploratorio eran las que más discriminaban a los grupos, aunque también eran las que más se correlacionaban. La variable que menos lo hace es sepalwid. Claramente se nota la formación de dos conglomerados. Además, si se observa el siguiente cuadro, los centros de los conglomerados finales son parecidos en sepallen y sepalwid, y los que marcan la diferencia son las variables petallen y petalwid. De aquí se confirma que es mejor utilizar dos conglomerados.

Variables	Conglomerado 1	Conglomerado2
Sepallen	5.0	6.3
sepalwid	3.4	2.9
Petallen	1.6	5.0
Petalwid	.3	1.7

Fuente: Datos iris
Centros de los conglomerados finales
Tabla 4.2

El primer conglomerado aplicando el método de las k-medias esta formado por 53 objetos y el segundo por 97. El número de objetos en cada conglomerado casi es el mismo que en los otros dos métodos. Dado que ambos métodos jerárquicos tienen el mismo número de

elementos y los mismos objetos al tomar dos grupos, los conglomerados finales son los que tienen 50 y 100 objetos. Lo que caracteriza al primer conglomerado es que las flores tienen pétalos cortos y angostos, es decir son flores pequeñas; y el segundo pétalos y sépalos muy largos aunque sus pétalos son más anchos que los del primer grupo, se les puede considerar como flores grandes.

4.4 ANÁLISIS DE CONGLOMERADOS CON ANÁLISIS DE COMPONENTES PRINCIPALES

Como se menciono en el capítulo 3, el Análisis de Componentes Principales (ACP) reduce el número de variables originales cuando estas son demasiadas y garantiza que las nuevas variables (componentes) no están correlacionadas. También se mencionó que se pueden obtener resultados muy diferentes con o sin el Análisis de Componentes Principales. Los datos de la aplicación tienen solo cuatro variables, por tanto no se puede decir que es para reducir dimensionalidad; sin embargo, las variables petallen y petalwid estaban correlacionadas, pero en la sección 4.3 se vio que eran las que diferenciaban los centros de cada conglomerado por el método de las k-medias, por tanto no es necesario aplicarle un ACP. Por otro lado, si se aplicara un ACP a las flores iris; según la tabla 4.3, solo bastaría el primer componente para tener un porcentaje alto en la variabilidad de los datos. Realmente no tiene mucho sentido aplicar esta técnica para tener una sola variable que forme grupos.

Componente	Valor propio	Porcentaje de varianza	Porcentaje acumulado
1	4.22824	92.462	92.462
2	0.242671	5.307	97.769
3	0.0782095	1.710	99.479
4	0.0238351	0.521	100.00

Fuente: Datos iris
Componentes principales
Tabla 4.3

CONCLUSIONES Y RECOMENDACIONES

A pesar de que en el Análisis de Conglomerados existen diversos métodos que permiten la formación de los grupos, no existe un método óptimo que defina mejor las agrupaciones, esto depende de la naturaleza de los datos y del conocimiento que tenga el analista sobre el comportamiento de las observaciones y de sus objetivos planteados previamente. Por lo que es necesario evaluar dos o tres métodos para ver cual es el que mejor describe los grupos. Por otro lado tampoco se tiene un algoritmo que diga cual es el número óptimo de conglomerados, existen métodos auxiliares que dan una idea pero el analista es el que decide finalmente con cuantos grupos se queda. Es de gran utilidad comparar los métodos jerárquicos y los no jerárquicos (método de las k medias), ya que en estas técnicas se tienen los centroides de cada conglomerado y de esta manera se puede saber que tan heterogéneos son los grupos.

El Análisis Exploratorio de los datos es una herramienta importante que muchas veces es omitida, pero puede ser de gran utilidad para evitar cálculos innecesarios y tener una mejor idea del comportamiento de los datos. En particular, los métodos gráficos son los más sencillos de manejar y permiten anticiparse a problemas futuros. Por otro lado, el Análisis de Conglomerados por si solo ha sido una técnica muy utilizada en distintas ramas, ya que ha permitido describir comportamientos y características especificas de los objetos, y corroborar grupos que se habían definido previamente por la experiencia; sin embargo, se puede conjugar con otras técnicas multivariadas para tener una mejor evaluación de los resultados, en este trabajo solo se menciono como análisis complementario el Análisis de componentes principales, aunque no tuvo sentido aplicárselo a los datos porque eran muy pocas variables; sin embargo, sería de gran utilidad usarlo cuando se tuviera un número considerable de variables; y aplicar el escalamiento multidimensional que es una técnica que esta basada en distancias y permite ver como se encuentran distribuidas las observaciones en los cuatro cuadrantes.

Es recomendable hacer una evaluación minuciosa de los diferentes métodos de agrupación con distintos tipos de datos para tener una mejor idea de cuál método es el que otorga óptimos resultados. Aunque algunos autores realizaron estudios comparativos sobre diferentes técnicas hace algunos años, hoy en día se pueden aprovechar los diversos programas estadísticos con los que se cuentan ya estos permiten realizar un mayor número de cálculos y de manera más rápida.

En este trabajo se dio un panorama del Análisis de Conglomerados, dentro del cual se explicaron sus técnicas de agrupación y sus medidas de proximidad, así como algunos problemas con los que se enfrenta el analista al utilizar esta técnica, pero también existen otras técnicas de agrupamiento como son las *técnicas clumping y la agrupación simultanea de individuos y variables* que pueden ser abordados en otro trabajo; también es de interés abordar con mayor profundidad los criterios de agrupamiento que se mencionaron dentro de los métodos no jerárquicos como son los métodos basados en la traza.

APÉNDICE A

ALGEBRA DE MATRICES

A.1 Matrices y Vectores

Una *matriz* es un arreglo cuadrado o rectangular de números o variables arreglados en renglones y columnas. Se utilizan letras en mayúsculas para representar a las matrices. Las entradas son números reales o variables las cuales representan números reales. Los elementos de una matriz se despliegan en paréntesis y la posición de cada elemento se denota de acuerdo al número de renglón o columna donde se encuentre la variable. La dimensión de una matriz se determina mediante el número de renglones y columnas que tiene. Se dice que las matrices son del mismo tamaño cuando cada una de ellas tiene el mismo número de renglones y de columnas que las otras.

En la siguiente matriz el elemento 5 se encuentra en la posición a_{11}, es decir en el renglón 1 y columna 1, el elemento 10 se encuentra en la posición a_{32}, es decir en el renglón 3 y columna 2.

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

Una *matriz cuadrada* es una matriz que tiene el mismo número de columnas y renglones, una matriz de *nxn*.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \quad \text{o bien} \quad A = (a_{ij}) \quad \text{donde } a_{ij} \text{ es un elemento de la matriz}$$

Una *matriz rectangular* es una matriz que tiene distinto número de columnas y renglones, esto es, *n* renglones y *p* columnas o viceversa, una matriz de *nxp o pxn*.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

Un *vector* es un arreglo con una sola columna o un solo renglón. Se utilizan letras minúsculas para representarlos y sus elementos se pueden identificar por un índice suscrito. Geométricamente un vector con p elementos representa un punto en un espacio p-dimensional. Además se dice que los vectores son del mismo tamaño cuando cada uno de ellos tiene el mismo número de renglones o de columnas que los restantes.

$$x' = \begin{pmatrix} x_1 & x_2 & \dots & x_p \end{pmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

vector renglón *vector columna*

La *suma de matrices* se determina sumando cada uno de los elementos correpondientes de cada matriz si éstas son del mismo tamaño. Esto es, si dos matrices A es nxp y B es nxp, entonces $C = A + B$, también de nxp, se desea encontrar

$$c_{ij} = (a_{ij} + b_{ij}).$$

Ejemplo:

$$A = \begin{bmatrix} -2 & 4 \\ 9 & 3 \\ -6 & 5 \end{bmatrix} \qquad B = \begin{bmatrix} 8 & -3 \\ 10 & 7 \\ 3 & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} -2 & 4 \\ 9 & 3 \\ -6 & 5 \end{bmatrix} + \begin{bmatrix} 8 & -3 \\ 10 & 7 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 1 \\ 19 & 10 \\ -3 & 6 \end{bmatrix}$$

La *resta de matrices* se obtiene restando cada uno de los elementos de cada matriz.

Para el *producto de matrices* AB se requiere que el número de columnas de A sea igual al número de renglones de B, en este caso se dice que A y B son conformables. Para obtener $C = AB$ se multiplica cada renglón de A por cada columna de B.

El $i-j-ésimo$ elemento de AB esta dado por:

$$c_{ij} = \sum_k a_{ik} b_{kj}$$

es decir c_{ij} es la suma de productos del i-ésimo renglón de A y la $j-ésima$ columna de B.

El tamaño de la matriz AB es el número de renglones de A y el número de columnas de B.

Ejemplo:

$$A = \begin{bmatrix} 2 & 7 & -3 \\ 8 & 9 & 5 \end{bmatrix} \qquad B = \begin{bmatrix} 7 & 6 & 2 \\ 5 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix}$$

$2x3 \qquad 3x3$

$$AB = \begin{bmatrix} 2\cdot 7+7\cdot 5+(-3)\cdot 1 & 2\cdot 6+7\cdot 2+(-3)\cdot 4 & 2\cdot 2+7\cdot 3+(-3)\cdot 9 \\ 8\cdot 7+9\cdot 5+5\cdot 1 & 8\cdot 6+9\cdot 2+5\cdot 4 & 8\cdot 2+9\cdot 3+5\cdot 9 \end{bmatrix}$$

$$AB = \begin{bmatrix} 46 & 14 & -2 \\ 106 & 86 & 88 \end{bmatrix}$$

$3x3$

La *multiplicación de vectores por matrices* tiene la misma regla descrita que para matrices, el número de renglones de una matriz debe ser igual al número de columnas del vector o viceversa.

Ejemplo:

$$A = \begin{bmatrix} 3 & -2 & 4 \\ 1 & 3 & 5 \end{bmatrix} \qquad a = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \qquad b = \begin{bmatrix} 2 \\ -5 \end{bmatrix}$$

$$Aa = \begin{bmatrix} 3 & -2 & 4 \\ 1 & 3 & 5 \end{bmatrix}\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 16 \\ 31 \end{bmatrix}$$

$$b'A = \begin{bmatrix} 2 & -5 \end{bmatrix}\begin{bmatrix} 3 & -2 & 4 \\ 1 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & -19 & -17 \end{bmatrix}$$

El producto de dos vectores es de forma similar. Si a y b son de $nx1$ entonces:

$a'b = a_1b_1 + a_2b_2 + \ldots + a_nb_n$, cuyo resultado es un escalar

Por otro lado, ab' es definida como una matriz cuadrada:

$$ab' = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \begin{pmatrix} b_1 & b_2 & \cdots & b_p \end{pmatrix} = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_p \\ a_2b_1 & a_2b_2 & \cdots & a_2b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_p \end{bmatrix}$$

El producto $a'a$ es llamado producto punto,

$$a'a = a_1a_1 + a_2a_2 + \ldots + a_na_n = a_1^2 + a_2^2 + \ldots + a_n^2$$

y la raíz cuadrada de esta expresión es la distancia desde el origen a un punto a y también se refiere como la longitud de a,

$$\textit{longitud de } a = \sqrt{a'a} = \sqrt{\sum_{i=1}^{n} a_i^2}$$

A.2 Matrices Transpuestas, Definidas Positivas y Traza de una Matriz

La *transpuesta* de una matriz A es denotada por A' o A^T. Se obtiene intercambiando los renglones y columnas. Los renglones de A' son las columnas A y las columnas de A' son los renglones de A'.

Ejemplo:

$$A = \begin{bmatrix} 3 & -5 \\ 8 & 2 \end{bmatrix} \qquad A' = \begin{bmatrix} 3 & 8 \\ -5 & 2 \end{bmatrix}$$

Una matriz simétrica A se dice que es *definida positiva* si $x'Ax > 0$ para todos los posibles vectores x (excepto $x = 0$). Pero A es *semidefinida positiva* si $x'Ax \geq 0$ para todo $x \neq 0$

Ejemplo: Los elementos de la diagonal a_{ii} de una matriz definida positiva son positivos.

Sea $x' = (0,\ldots 0,1,0,\ldots,0)$ con 1 en la $i-$*ésima* posición. $x'Ax = a_{ii} > 0$.

de manera similar para una matriz A semidefinida positivas, $a_{ii} \geq 0$.

La *traza* se define como la suma de los elementos de la diagonal de una matriz A y se denota como $tr(A) = \sum_{i=1}^{n} a_{ii}$

Ejemplo:

$$A = \begin{bmatrix} 2 & 8 & -5 \\ 1 & 3 & -3 \\ 6 & 4 & 9 \end{bmatrix} \qquad tr(A) = 2 + 3 + 9 = 14$$

A.3 Matrices y Vectores ortogonales

Dos vectores a y b del mismo tamaño, son *ortogonales* si:

$$a'b = a_1b_1 + a_2b_2 + \ldots + a_nb_n = 0$$

si $a'a = 1$ entonces el vector a es normalizado. Este vector puede ser normalizado si es dividido entre su longitud, $\sqrt{a'a}$. Esto es:

$$c = \frac{a}{\sqrt{a'a}}, \qquad c'c = 1$$

Una matriz $C = (c_1, c_2, \ldots, c_p)$cuyas columnas están normalizadas y mutuamente ortogonales se denomina *matriz ortogonal*. Los elementos de $C'C$ son producto de las columnas de C, las cuales tienen la propiedad de que $c_i'c_i = 1$ para todo i y $c_i'c_j = 0$ para todo $i \neq j$. $C'C = I$ por lo que los renglones también son normalizados y ortogonales entre sí.

A.4 Combinación Lineal y dependencia e independencia lineal

Dados dos vectores x_1 y x_2 distintos de cero y no paralelos en $\Re^2$, el conjunto de todas las *combinaciones lineales* de x_1 y x_2 esta dado por:

$$\{a_1x_1 + a_2x_2 \mid a_1, a_2 \in \Re\} \text{ donde } a_1 \text{ y } a_2 \text{ son constantes}$$

Esta expresión es extensiva para un espacio de *p dimensiones*, $\Re^p$.

Un conjunto de vectores $\{x_1, x_2, \ldots, x_p\}$ es *linealmente dependiente* si existe una relación de dependencia, es decir si:

$$a_1x_1 + a_2x_2 + \ldots + a_px_p = 0, \text{ para alguna } a_j \neq 0$$

Un conjunto de vectores $\{x_1, x_2, \ldots, x_p\}$ es *linealmente independiente* si no existe ninguna relación de dependencia, es decir si:

$$a_1x_1 + a_2x_2 + \ldots + a_px_p = 0, \text{ para todo } a_j = 0$$

Una combinación lineal de las $x's$ puede se escrita como:

$$z = a_1x_1 + a_2x_2 + \ldots + a_px_p = a'x$$

donde $a' = (a_1, a_2, \ldots, a_p)$. Si el vector a es aplicado a cada x_i, se tiene:

$$z_i = a_1x_{i1} + a_2x_{i2} + \ldots + a_px_{ip} = a'x_i \qquad i = 1, 2, \ldots, n$$

n es el número de observaciones, individuos u objetos y p son las variables correspondientes a cada observación.

A.5 Determinantes, Valores propios y Vectores propios

El símbolo para representar el *determinante* de una matriz cuadrada A es $|A|$ o bien $\det A$. El determinante una matriz de $2x2$ se define de la siguiente forma:

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

y para una matriz de $3x3$ como sigue:

$$|A| = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} - a_{31}a_{22}a_{13} - a_{32}a_{23}a_{11} - a_{33}a_{12}a_{21}$$

Para dar una definición generalizada del determinante se necesita conocer lo siguiente:
La matriz menor A_{ij} de $A = (a_{ij})$, es una matriz de $(n-1)$ x $(n-1)$ obtenida al eliminar la $i-ésima$ fila y la $j-ésima$ de A. Entonces el cofactor de a_{ij} en A se define como:

$$a'_{ij} = (-1)^{i+j}|A_{ij}|$$

- Así entonces el determinante de una matriz de nxn con $n > 1$ se define como:

$$|A| = a_{11}a'_{11} + a_{12}a'_{12} + \ldots + a_{1n}a'_{1n}$$

Ejemplo:

$$A = \begin{bmatrix} 5 & -2 & 4 & -1 \\ 0 & 1 & 5 & 2 \\ 1 & 2 & 0 & 1 \\ -3 & 1 & -1 & 1 \end{bmatrix}$$

$$|A| = 5(-1)^2\begin{vmatrix} 1 & 5 & 2 \\ 2 & 0 & 1 \\ 1 & -1 & 1 \end{vmatrix} + (-2)(-1)^3\begin{vmatrix} 0 & 5 & 2 \\ 1 & 0 & 1 \\ -3 & -1 & 1 \end{vmatrix} + 4(-1)^4\begin{vmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ -3 & 1 & 1 \end{vmatrix}$$

$$+ (-1)(-1)^5\begin{vmatrix} 0 & 1 & 5 \\ 1 & 2 & 0 \\ -3 & 1 & -1 \end{vmatrix} = 5(-8) + 2(-22) + 4(10) + 1(36) = -8$$

Por tanto:
$|A| = -8$

Para una matriz cuadrada A, un escalar λ y un vector distinto de cero x pueden ser encontrados tal que

$$Ax = \lambda x$$

donde λ es llamado *valor propio* de A y x es un *vector propio*

- Para encontrar λ y x la expresión anterior se escribe como

$$(A - \lambda I)x = 0$$

- Si $|A - \lambda I| \neq 0$ entonces $(A - \lambda I)$ tiene una inversa y $x = 0$ es la única solución. pero para encontrar soluciones no triviales ($x \neq 0$), $|A - \lambda I| = 0$. De este modo:

$$(A - \lambda I)x = 0 \text{ .....(1)}$$

y la matriz $A - \lambda I$ debe ser una matriz singular para encontrar una solución de $x \neq 0$

- $|A - \lambda I| = 0$ se llama ecuación característica o polinimio característico. Si A es nxn, A tendrá n raíces, es decir tendrá n valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$. Las λ ´ s no son necesariamente distintas o todas cero, pero si se procesan en computadoras sobre datos reales, las λ ´ s serán distintas.

- Después de encontrar los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$, se calculan los vectores propios $x_1, x_2, \ldots, x_n$ utilizando la ecuación 1.

Ejemplo:

$$A = \begin{bmatrix} 1 & 2 \\ -1 & 4 \end{bmatrix}$$

$$|A - \lambda I| = \begin{vmatrix} 1-\lambda & 2 \\ -1 & 4-\lambda \end{vmatrix} = (1-\lambda)(4-\lambda) + 2 = 0$$

$$= \lambda^2 - 5\lambda + 6 = (\lambda - 3)(\lambda - 2) = 0$$

los valores propios son: $\lambda_1 = 3$ y $\lambda_2 = 2$

Pra encontrar los vectores propios a partir de$\lambda_1 = 3$ y $\lambda_2 = 2$ se utilizará la ecuación 1.

$$(A - \lambda I)x = 0$$

$$\begin{bmatrix} 1-3 & 2 \\ -1 & 4-3 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$-2x_1 + 2x_2 = 0$$
$$-x_1 + x_2 = 0$$

de este sistema de ecuaciones se obtiene $x_1 = x_2$, ambas variables son desconocidas entonces un vector solución puede ser obtenido asignando una constante arbitraria como:

$$x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

si se normalizan los elementos del vector x_1, se tiene:

$$x_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Similarmente para $\lambda_2 = 2$, se obtiene:

$$x_2 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}$$

Si se multiplica por un escalar k ambos lados de la ecuación 1, se obtiene

$$(A - \lambda I)kx = k0 = 0$$

de esta manera si x es un vector propio de A, kx también es un vector propio

Para una matriz cuadrada A con valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$tr(A) = \sum_{i=1}^{n} \lambda_i \qquad |A| = \prod_{i=1}^{n} \lambda_i$$

Los valores propios de una matriz definida positiva son todos positivos.

Los valores propios de una matriz semidefinida positiva son positivos o cero con el número de valores propios positivos igual al rango de la matriz.

Los valores propios de una matriz definida positiva se listan en forma descendente: $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. los vectores propios $x_1, x_2, \ldots, x_n$ se listan de la misma manera, x_1 corresponde a λ_1, x_2 a λ_2 y así sucesivamente.

A.6 Matrices de Covarianza y Correlación poblacional y muestral

La *matriz de varianza y covarianza para una muestra* para p variables se denota como: $S = (s_{ij})$

$$S = (s_{ij}) = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix} \qquad (1)$$

en la diagonal de esta matriz se encuentran la varianzas de las p variables y los elementos fuera de la diagonal son las covarianzas de todos los posibles pares de variables. El $i-ésimo$ renglón (columna) contiene las covarianzas de la variable x_i con las $p-1$ variables.

La matriz (1) se le llama también matriz de varianza, matriz de covarianza y matriz de dispersión.

La varianza de la i – *ésima* variable $s_{ii} = s_i^2$ se calcula de la siguiente manera:

$$s_{ii} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \qquad (2)$$

$$= \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}^2 - n\bar{x}_i^2)$$

donde $\bar{x}$ es la media de la i – *ésima* variable

La covarianza de la i – *ésima* y j – *ésima* variable, s_{ij} se calcula como:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2(x_{kj} - \bar{x}_j) \qquad (3)$$

$$= \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}x_{kj} - n\bar{x}_i\bar{x}_j)$$

Los subindices i y j son para las variables y k para observaciones. La matriz S es simétrica porque $s_{ij} = s_{ji}$ de acuerdo a (3).

La *matriz de covarianza* S se puede expresar en términos de los vectores de las observaciones:

$$S = \frac{1}{n-1}\sum_{k=1}^{n}(x_i - \bar{x})(x_i - \bar{x})' \qquad (4)$$

$$= \frac{1}{n-1}\left(\sum_{k=1}^{n}x_i x'_i - n\bar{x}\bar{x}'\right)$$

El elemento en la posición $(1,1)$ en $(x_i - \bar{x})(x_i - \bar{x})'$ es $(x_{i1} - \bar{x}_1)^2$ y el elemento $(1,2)$ es $(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$. Así la ecuación (4) es equivalente a la ecuación (2) y (3).

La *matriz de varianza y covarianza para una población* se denota como sigue:

$$\Sigma = cov(y) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \qquad (5)$$

Los elementos de la diagonal $\sigma_{ii} = \sigma_i^2$ son las varianzas de la población de las variables x´ s y los elementos fuera de la diagonal son las covarianzas de todos los posibles pares de las x´ s de la población.

La matriz (5) se puede expresar como.

$$\Sigma = E[(y-\mu)(y-\mu)']$$

$$= \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_p - \mu_p \end{bmatrix} \begin{bmatrix} y_1 - \mu_1, & y_2 - \mu_2, & \cdots & y_p - \mu_p \end{bmatrix}$$

$$= E\begin{bmatrix} (y_1-\mu_1)^2 & (y_1-\mu_1)(y_2-\mu_2) & \cdots & (y_1-\mu_1)(y_p-\mu_p) \\ (y_2-\mu_2)(y_1-\mu_1) & (y_2-\mu_2)^2 & \cdots & (y_2-\mu_2)(y_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (y_p-\mu_p)(y_1-\mu_1) & (y_p-\mu_p)(y_2-\mu_2) & \cdots & (y_p-\mu_p)^2 \end{bmatrix}$$

$$= E\begin{bmatrix} E(y_1-\mu_1)^2 & E(y_1-\mu_1)(y_2-\mu_2) & \cdots & E(y_1-\mu_1)(y_p-\mu_p) \\ E(y_2-\mu_2)(y_1-\mu_1) & E(y_2-\mu_2)^2 & \cdots & E(y_2-\mu_2)(y_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(y_p-\mu_p)(y_1-\mu_1) & E(y_p-\mu_p)(y_2-\mu_2) & \cdots & E(y_p-\mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Además es análoga a (4).

La correlación muestral entre las variables i – *ésima* y j – *ésima* esta definida como:

$$\rho_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} = \frac{s_{ij}}{s_i s_j}$$

La *matriz de correlación muestral* es análoga a la matriz de varianza y covarianza con correlaciones en lugar de covarianzas.

$$R = (r_{ij}) = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix} \qquad (6)$$

por ejemplo en el renglón 3 se tiene la correlación de x_3 con cada una de las $x's$, incluyendo la correlación con si misma la cual es 1. la matriz (6) es simétrica ya que $r_{ij} = r_{ji}$.

La matriz correlación se puede obtener a partir de la matriz de varianza y covarianza y viceversa, definiendo:

$$D_S = diag\left(\sqrt{s_{11}}, \sqrt{s_{22}}, \dots, \sqrt{s_{pp}}\right) = diag(s_1, s_2, \dots, s_p)$$

$$= \begin{bmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_1 \end{bmatrix}$$

La *matriz* de *correlación poblacional* se obtiene similarmente que (6) y se define:

$$P = (\rho_{ij}) = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{bmatrix} \qquad (7)$$

APÉNDICE B

TABLA B1 : TABLA DE DATOS IRIS

Iristype	número de elemento	Sepallen	Sepalwid	Petallen	Petalwid
SETOSA	1	5	3,3	1,4	0,2
VIRGINIC	2	6,4	2,8	5,6	2,2
VERSICOL	3	6,5	2,8	4,6	1,5
VIRGINIC	4	6,7	3,1	5,6	2,4
VIRGINIC	5	6,3	2,8	5,1	1,5
SETOSA	6	4,6	3,4	1,4	0,3
VIRGINIC	7	6,9	3,1	5,1	2,3
VERSICOL	8	6,2	2,2	4,5	1,5
VERSICOL	9	5,9	3,2	4,8	1,8
SETOSA	10	4,6	3,6	1	0,2
VERSICOL	11	6,1	3	4,6	1,4
VERSICOL	12	6	2,7	5,1	1,6
VIRGINIC	13	6,5	3	5,2	2
VERSICOL	14	5,8	2,5	3,9	1,1
VIRGINIC	15	6,5	3	5,5	1,8
VIRGINIC	16	5,8	2,7	5,1	1,9
VIRGINIC	17	6,8	3,2	5,9	2,3
SETOSA	18	5,1	3,3	1,7	0,5
VERSICOL	19	5,7	2,8	4,5	1,3
VIRGINIC	20	6,2	3,4	5,4	2,3
VIRGINIC	21	7,7	3,8	6,7	2,2
VERSICOL	22	6,3	3,3	4,7	1,6
VIRGINIC	23	6,7	3,3	5,7	2,5
VIRGINIC	24	7,6	3	6,6	2,1
VIRGINIC	25	4,9	2,5	4,5	1,7
SETOSA	26	5,5	3,5	1,3	0,2
VIRGINIC	27	6,7	3	5,2	2,3
VERSICOL	28	7	3,2	4,7	1,4
VERSICOL	29	6,4	3,2	4,5	1,5
VERSICOL	30	6,1	2,8	4	1,3
SETOSA	31	4,8	3,1	1,6	0,2
VIRGINIC	32	5,9	3	5,1	1,8
VERSICOL	33	5,5	2,4	3,8	1,1
VIRGINIC	34	6,3	2,5	5	1,9
VIRGINIC	35	6,4	3,2	5,3	2,3
SETOSA	36	5,2	3,4	1,4	0,2
SETOSA	37	4,9	3,6	1,4	0,1
VERSICOL	38	5,4	3	4,5	1,5
VIRGINIC	39	7,9	3,8	6,4	2
SETOSA	40	4,4	3,2	1,3	0,2
VIRGINIC	41	6,7	3,3	5,7	2,1
SETOSA	42	5	3,5	1,6	0,6
VERSICOL	43	5,6	2,6	4	1,2
SETOSA	44	4,4	3	1,3	0,2
VIRGINIC	45	7,7	2,8	6,7	2
VIRGINIC	46	6,3	2,7	4,9	1,8
SETOSA	47	4,7	3,2	1,6	0,2
VERSICOL	48	5,5	2,6	4,4	1,2
VIRGINIC	49	7,2	3,2	6	1,8
SETOSA	50	4,8	3	1,4	0,3
SETOSA	51	5,1	3,8	1,6	0,2
VIRGINIC	52	6,1	3	4,9	1,8
SETOSA	53	4,6	3,4	1,9	0,2
SETOSA	54	5	3	1,6	0,2
SETOSA	55	5	3,2	1,2	0,2
VIRGINIC	56	6,1	2,6	5,6	1,4
VIRGINIC	57	6,4	2,8	5,6	2,1
SETOSA	58	4,3	3	1,1	0,1
SETOSA	59	5,8	4	1,2	0,2
SETOSA	60	5,1	3,5	1,9	0,4
VERSICOL	61	6,7	3,1	4,4	1,4
VIRGINIC	62	6,2	2,8	4,8	1,8
SETOSA	63	4,9	3	1,4	0,2
SETOSA	64	5,1	3,5	1,4	0,2
VERSICOL	65	5,6	3	4,5	1,5
VERSICOL	66	5,8	2,7	4,1	1
SETOSA	67	5	3,4	1,6	0,4
SETOSA	68	4,6	3,2	1,4	0,2
VERSICOL	69	6	2,9	4,5	1,5
VERSICOL	70	5,7	2,6	3,5	1
SETOSA	71	5,7	4,4	1,5	0,4
SETOSA	72	5	3,6	1,4	0,2
VIRGINIC	73	7,7	3	6,1	2,3
VIRGINIC	74	6,3	3,4	5,6	2,4
VIRGINIC	75	5,8	2,7	5,1	1,9
VERSICOL	76	5,7	2,9	4,2	1,3

Iristype	número de elemento	Sepallen	Sepalwid	Petallen	Petalwid
VIRGINIC	77	7,2	3	5,8	1,6
SETOSA	78	5,4	3,4	1,5	0,4
SETOSA	79	5,2	4,1	1,5	0,1
VIRGINIC	80	7,1	3	5,9	2,1
VIRGINIC	81	6,4	3,1	5,5	1,8
VIRGINIC	82	6	3	4,8	1,8
VIRGINIC	83	6,3	2,9	5,6	1,8
VERSICOL	84	4,9	2,4	3,3	1
VERSICOL	85	5,8	2,7	4,2	1,3
VERSICOL	86	5,7	3	4,2	1,2
SETOSA	87	5,5	4,2	1,4	0,2
SETOSA	88	4,9	3,1	1,5	0,2
VIRGINIC	89	7,7	2,6	6,9	2,3
VIRGINIC	90	6	2,2	5	1,5
SETOSA	91	5,4	3,9	1,7	0,4
VERSICOL	92	6,6	2,9	4,6	1,3
VERSICOL	93	5,2	2,7	3,9	1,4
VERSICOL	94	6	3,4	4,5	1,6
SETOSA	95	5	3,4	1,5	0,2
SETOSA	96	4,4	2,9	1,4	0,2
VERSICOL	97	5	2	3,5	1
VERSICOL	98	5,5	2,4	3,7	1
VERSICOL	99	5,8	2,7	3,9	1,2
SETOSA	100	4,7	3,2	1,3	0,2
SETOSA	101	4,6	3,1	1,5	0,2
VIRGINIC	102	6,9	3,2	5,7	2,3
VERSICOL	103	6,2	2,9	4,3	1,3
VIRGINIC	104	7,4	2,8	6,1	1,9
VERSICOL	105	5,9	3	4,2	1,5
SETOSA	106	5,1	3,4	1,5	0,2
SETOSA	107	5	3,5	1,3	0,3
VIRGINIC	108	6,8	2,8	4,9	2
VERSICOL	109	6	2,2	4	1
VIRGINIC	110	7,3	2,9	6,3	1,8
VIRGINIC	111	6,7	2,5	5,8	1,8
SETOSA	112	4,8	3,1	1,5	0,1
VERSICOL	113	6,7	3,1	4,7	1,5
VERSICOL	114	6,3	2,3	4,4	1,3
SETOSA	115	5,4	3,7	1,5	0,2
VERSICOL	116	5,8	3	4,1	1,3
VERSICOL	117	6,3	2,5	4,9	1,5
VERSICOL	118	6,1	2,8	4,7	1,2
VERSICOL	119	6,4	2,9	4,3	1,3
VERSICOL	120	5,1	2,5	3	1,1
VERSICOL	121	5,7	2,8	4,1	1,3
VIRGINIC	122	6,5	3	5,8	2,2
VIRGINIC	123	6,9	3,1	5,4	2,1
SETOSA	124	5,4	3,9	1,3	0,4
SETOSA	125	5,1	3,5	1,4	0,3
VIRGINIC	126	7,2	3,6	6,1	2,5
VIRGINIC	127	6,5	3,2	5,1	2
VERSICOL	128	6,1	2,9	4,7	1,4
VERSICOL	129	5,6	2,9	3,6	1,3
VERSICOL	130	6,9	3,1	4,9	1,5
VIRGINIC	131	6,4	2,7	5,3	1,9
VIRGINIC	132	6,8	3	5,5	2,1
VERSICOL	133	5,5	2,5	4	1,3
SETOSA	134	4,8	3,4	1,6	0,2
SETOSA	135	4,8	3	1,4	0,1
SETOSA	136	4,5	2,3	1,3	0,3
VIRGINIC	137	5,7	2,5	5	2
SETOSA	138	5,7	3,8	1,7	0,3
SETOSA	139	5,1	3,8	1,5	0,3
VERSICOL	140	5,5	2,3	4	1,3
VERSICOL	141	6,6	3	4,4	1,4
VERSICOL	142	6,8	2,8	4,8	1,4
SETOSA	143	5,4	3,4	1,7	0,2
SETOSA	144	5,1	3,7	1,5	0,4
SETOSA	145	5,2	3,5	1,5	0,2
VIRGINIC	146	5,8	2,8	5,1	2,4
VERSICOL	147	6,7	3	5	1,7
VIRGINIC	148	6,3	3,3	6	2,5
SETOSA	149	5,3	3,7	1,5	0,2
VERSICOL	150	5	2,3	3,3	1

TABLA B2 : TABLA DE AMALGAMACIÓN DEL MÉTODO DE LA LIGA SIMPLE

Distancia	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10	Obj. No. 11	Obj. No. 12	Obj. No. 13	Obj. No. 14	Obj. No. 15	Obj. No. 16	Obj. No. 17	Obj. No. 18	Obj. No. 19	Obj. No. 20	Obj. No. 21
0500000	C_18	C_67	C_42																		
0500000	C_13	C_127																			
0500000	C_8	C_52	C_62																		
0500000	C_11	C_129	C_69	C_118																	
0500000	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42						
0500000	C_6	C_31	C_66	C_112	C_90	C_63	C_135	C_47	C_54	C_88	C_100	C_101	C_40	C_44	C_96						
0500001	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
0500001	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
0500001	C_17	C_102																			
0500000	C_15	C_81	C_63																		
0500000	C_61	C_141	C_92																		
0500000	C_14	C_33	C_98	C_133	C_140																
0800000	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
0800000	C_4	C_23																			
0800000	C_6	C_52	C_62	C_46	C_42																
0800000	C_8	C_52	C_62	C_46	C_42	C_34															
0800000	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
0800000	C_3	C_61	C_141	C_92																	
0800000	C_20	C_74																			
0800000	C_43	C_69	C_66																		
0800001	C_7	C_27																			
0800000	C_3	C_61	C_141	C_92	C_103	C_119															
0800000	C_22	C_29																			
0700000	C_24	C_45																			
0700000	C_43	C_69	C_66	C_78	C_66	C_121	C_65	C_118													
0700001	C_8	C_114																			
0700001	C_14	C_33	C_66	C_133	C_140	C_43	C_69	C_66	C_78	C_66	C_121	C_65	C_118								
0700001	C_18	C_75	C_137																		
0700001	C_14	C_33	C_66	C_133	C_140	C_43	C_69	C_66	C_78	C_66	C_121	C_65	C_118	C_46							
0700001	C_29	C_130																			
0700001	C_104	C_119																			
0700001	C_4	C_23	C_17	C_102																	
0700000	C_8	C_52	C_62	C_46	C_42	C_34	C_32														
0800000	C_78	C_143																			
0800002	C_29	C_130	C_113																		
0800000	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
0800000	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
0800000	C_19	C_36	C_65																		
0800001	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
0800001	C_2	C_57	C_122																		
0800001	C_4	C_23	C_17	C_102	C_41																
0800001	C_14	C_33	C_66	C_133	C_140	C_43	C_69	C_66	C_78	C_66	C_121	C_65	C_118	C_46	C_105						
0800001	C_20	C_74	C_35																		
0800000	C_3	C_61	C_141	C_92	C_103	C_119	C_29	C_130	C_113												
0800000	C_14	C_33	C_66	C_133	C_140	C_43	C_69	C_66	C_78	C_66	C_121	C_65	C_118	C_46	C_105	C_19	C_36	C_65			
0800000	C_3	C_61	C_141	C_92	C_103	C_119	C_29	C_130	C_113	C_22	C_29										
1000000	C_18	C_75	C_137	C_108																	
1000001	C_3	C_61	C_141	C_92	C_103	C_119	C_29	C_130	C_113	C_22	C_29	C_147									
1700000	C_21	C_36																			
1700000	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
1700001	C_24	C_45	C_69																		
1800001	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
1800002	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
1900001	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
1800002	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
2000000	C_1	C_36	C_64	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_67	C_42	C_6	C_31	C_66	C_112	C_90	C_63
2300000	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
2400001	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
2600001	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
2700000	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
2600000	C_2	C_57	C_122	C_15	C_61	C_63	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35

2900002	C_2	C_87	C_122	C_15	C_81	C_83	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
3000000	C_2	C_87	C_122	C_15	C_81	C_83	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
3900001	C_1	C_36	C_84	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_87	C_42	C_8	C_31	C_86	C_112	C_50	C_63
3999995	C_2	C_87	C_122	C_15	C_81	C_83	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
4200001	C_2	C_87	C_122	C_15	C_81	C_83	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
5400000	C_2	C_87	C_122	C_15	C_81	C_83	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
6700000	C_2	C_87	C_122	C_15	C_81	C_83	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27	C_13	C_127	C_131	C_20	C_74	C_35
7.680000	C_1	C_36	C_84	C_125	C_72	C_37	C_95	C_108	C_145	C_107	C_134	C_55	C_18	C_87	C_42	C_8	C_31	C_86	C_112	C_50	C_63

Obj. No. 22	Obj. No. 23	Obj. No. 24	Obj. No. 25	Obj. No. 26	Obj. No. 27	Obj. No. 28	Obj. No. 29	Obj. No. 30	Obj. No. 31	Obj. No. 32	Obj. No. 33	Obj. No. 34	Obj. No. 35	Obj. No. 36	Obj. No. 37	Obj. No. 38	Obj. No. 39	Obj. No. 40	Obj. No. 41	Obj. No. 42	Obj. No. 43
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96													
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149											
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149	C_58										
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149	C_58	C_51	C_139	C_144							
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149	C_58	C_51	C_139	C_144	C_53						
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149	C_58	C_51	C_139	C_144	C_53	C_26					
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149	C_58	C_51	C_139	C_144	C_53	C_26	C_78	C_143			
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149	C_58	C_51	C_139	C_144	C_53	C_26	C_78	C_143	C_79	C_91	C_136
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_135	C_47	C_54	C_66	C_100	C_101	C_40	C_44	C_96	C_115	C_149	C_58	C_51	C_139	C_144	C_53	C_26	C_78	C_143	C_79	C_91	C_136
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_3	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103
C_5	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_16	C_75	C_137	C_108	C_117	C_49	C_77	C_80	C_3	C_81	C_141	C_92	C_103

C_3	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_18	C_75	C_137	C_106	C_117	C_49	C_77	C_80	C_3	C_61	C_141	C_92	C_103
C_8	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_18	C_75	C_137	C_106	C_117	C_49	C_77	C_80	C_3	C_61	C_141	C_92	C_103
C_138	C_47	C_54	C_88	C_100	C_101	C_40	C_44	C_98	C_115	C_149	C_58	C_81	C_139	C_144	C_53	C_28	C_76	C_143	C_79	C_91	C_138
C_5	C_12	C_8	C_52	C_82	C_46	C_82	C_34	C_32	C_18	C_75	C_137	C_106	C_117	C_49	C_77	C_80	C_3	C_61	C_141	C_92	C_103
C_5	C_12	C_8	C_52	C_82	C_46	C_82	C_34	C_32	C_18	C_75	C_137	C_106	C_117	C_49	C_77	C_80	C_3	C_61	C_141	C_92	C_103
C_8	C_12	C_8	C_52	C_82	C_46	C_82	C_34	C_32	C_18	C_75	C_137	C_106	C_117	C_49	C_77	C_80	C_3	C_61	C_141	C_92	C_103
C_8	C_12	C_9	C_52	C_82	C_46	C_82	C_34	C_32	C_18	C_75	C_137	C_106	C_117	C_49	C_77	C_80	C_3	C_61	C_141	C_92	C_103
C_138	C_47	C_54	C_88	C_100	C_101	C_40	C_44	C_98	C_115	C_149	C_58	C_51	C_139	C_144	C_53	C_28	C_76	C_143	C_79	C_91	C_138

Obj. No. 44	Obj. No. 45	Obj. No. 46	Obj. No. 47	Obj. No. 48	Obj. No. 49	Obj. No. 50	Obj. No. 51	Obj. No. 52	Obj. No. 53	Obj. No. 54	Obj. No. 55	Obj. No. 56	Obj. No. 57	Obj. No. 58	Obj. No. 59	Obj. No. 60	Obj. No. 61	Obj. No. 62	Obj. No. 63	Obj. No. 64	Obj. No. 65
C_124	C_87	C_80	C_71	C_58																	
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_128	C_89	C_118	C_14	C_33	C_98	C_133	C_140	C_43	C_98	C_86	C_76
C_119	C_28	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_128	C_89	C_118	C_14	C_33	C_96	C_133	C_140	C_43	C_98	C_86	C_78
C_119	C_29	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_128	C_89	C_118	C_14	C_33	C_96	C_133	C_140	C_43	C_99	C_88	C_76
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_128	C_89	C_118	C_14	C_33	C_96	C_133	C_140	C_43	C_98	C_86	C_76
C_124	C_87	C_80	C_71	C_58	C_19																
C_119	C_28	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	C_89	C_118	C_14	C_33	C_98	C_133	C_140	C_43	C_99	C_88	C_78
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	C_89	C_118	C_14	C_33	C_96	C_133	C_140	C_43	C_99	C_88	C_76
C_119	C_29	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_128	C_89	C_118	C_14	C_33	C_96	C_133	C_140	C_43	C_99	C_88	C_78
C_119	C_28	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	C_89	C_118	C_14	C_33	C_96	C_133	C_140	C_43	C_98	C_88	C_76
C_118	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	C_89	C_119	C_14	C_33	C_90	C_133	C_140	C_43	C_98	C_86	C_76

C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	[illegible]	C_118	C_14	C_33	[illegible]	C_133	C_140	C_43	[illegible]	[illegible]	C_76
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	[illegible]	C_118	C_14	C_33	[illegible]	C_133	C_140	C_43	[illegible]	[illegible]	C_76
C_124	C_87	C_60	C_71	C_50	C_10	C_136															
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	[illegible]	C_118	C_14	C_33	[illegible]	C_133	C_140	C_43	[illegible]	[illegible]	C_76
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	[illegible]	C_118	C_14	C_33	[illegible]	C_133	C_140	C_43	[illegible]	[illegible]	C_76
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	[illegible]	C_118	C_14	C_33	[illegible]	C_133	C_140	C_43	[illegible]	[illegible]	C_76
C_119	C_26	C_130	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	[illegible]	C_118	C_14	C_33	[illegible]	C_133	C_140	C_43	[illegible]	[illegible]	C_76
C_124	C_87	C_80	C_71	C_39	C_18	C_138	C_2	C_57	C_122	C_15	C_81	C_83	C_4	C_23	C_17	C_102	C_41	C_123	C_132	C_7	C_27

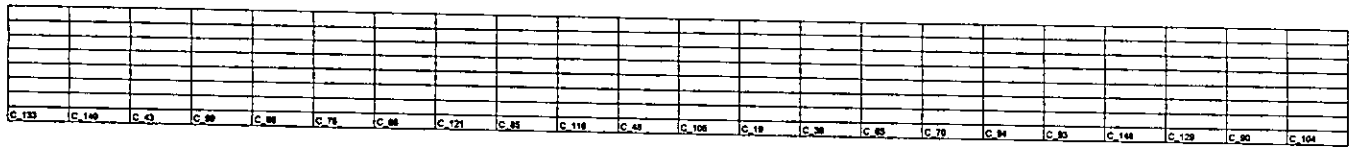
Obj. No. 66	Obj. No. 67	Obj. No. 68	Obj. No. 69	Obj. No. 70	Obj. No. 71	Obj. No. 72	Obj. No. 73	Obj. No. 74	Obj. No. 75	Obj. No. 76	Obj. No. 77	Obj. No. 78	Obj. No. 79	Obj. No. 80	Obj. No. 81	Obj. No. 82	Obj. No. 83	Obj. No. 84	Obj. No. 85	Obj. No. 86	Obj. No. 87
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148									
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148	C_129								
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148	C_129	C_80							
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148	C_129	C_80	C_104	C_110					
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148	C_129	C_80	C_104	C_110	C_109				
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148	C_129	C_80	C_104	C_110	C_109	C_146			
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148	C_129	C_80	C_104	C_110	C_109	C_146	C_8	C_114	
C_66	C_121	C_85	C_110	C_46	C_105	C_19	C_36	C_65	C_70	C_84	C_93	C_148	C_129	C_80	C_104	C_110	C_109	C_146	C_8	C_114	C_24

C_86	C_121	C_86	C_118	C_46	C_105	C_19	C_38	C_85	C_70	C_94	C_83	C_148	C_129	C_90	C_104	C_110	C_108	C_146	C_8	C_114	C_24
C_86	C_121	C_86	C_118	C_48	C_105	C_19	C_38	C_85	C_70	C_94	C_83	C_148	C_129	C_90	C_104	C_110	C_108	C_146	C_8	C_114	C_24
C_86	C_121	C_85	C_118	C_48	C_105	C_19	C_38	C_85	C_70	C_94	C_83	C_148	C_129	C_90	C_104	C_110	C_109	C_146	C_8	C_114	C_24
C_86	C_121	C_85	C_118	C_48	C_105	C_19	C_38	C_85	C_70	C_94	C_83	C_148	C_129	C_90	C_104	C_110	C_108	C_146	C_8	C_114	C_24
C_86	C_121	C_86	C_118	C_48	C_105	C_19	C_38	C_85	C_70	C_94	C_83	C_148	C_129	C_90	C_104	C_110	C_108	C_146	C_8	C_114	C_24
C_86	C_121	C_85	C_118	C_48	C_105	C_19	C_38	C_85	C_70	C_94	C_83	C_148	C_129	C_90	C_104	C_110	C_108	C_146	C_8	C_114	C_24
C_13	C_127	C_131	C_20	C_74	C_35	C_5	C_12	C_9	C_52	C_82	C_46	C_62	C_34	C_32	C_18	C_75	C_137	C_108	C_117	C_49	C_77

Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.
88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109
C_45	C_89																				
C_45	C_89	C_73																			

C_45	C_89	C_73	C_56																		
C_45	C_89	C_73	C_56	C_111																	
C_45	C_89	C_73	C_56	C_111	C_129																
C_45	C_89	C_73	C_56	C_111	C_129	C_84	C_150	C_97	C_120												
C_45	C_89	C_73	C_56	C_111	C_129	C_84	C_150	C_97	C_120	C_25											
C_45	C_89	C_73	C_56	C_111	C_126	C_84	C_150	C_97	C_120	C_25	C_21	C_39									
C_80	C_3	C_61	C_141	C_82	C_165	C_119	C_26	C_139	C_113	C_22	C_29	C_147	C_142	C_30	C_11	C_126	C_89	C_116	C_14	C_33	C_86

Obj. No. 110	Obj. No. 111	Obj. No. 112	Obj. No. 113	Obj. No. 114	Obj. No. 115	Obj. No. 116	Obj. No. 117	Obj. No. 118	Obj. No. 119	Obj. No. 120	Obj. No. 121	Obj. No. 122	Obj. No. 123	Obj. No. 124	Obj. No. 125	Obj. No. 126	Obj. No. 127	Obj. No. 128	Obj. No. 129	Obj. No. 130	Obj. No. 131


C_133
C_149
C_43
C_99
C_86
C_75
C_66
C_121
C_85
C_119
C_48
C_106
C_19
C_38
C_65
C_79
C_94
C_93
C_144
C_129
C_90
C_104

Obj. No. 132	Obj. No. 133	Obj. No. 134	Obj. No. 135	Obj. No. 136	Obj. No. 137	Obj. No. 138	Obj. No. 139	Obj. No. 140	Obj. No. 141	Obj. No. 142	Obj. No. 143	Obj. No. 144	Obj. No. 145	Obj. No. 146	Obj. No. 147	Obj. No. 148	Obj. No. 149	Obj. No. 150

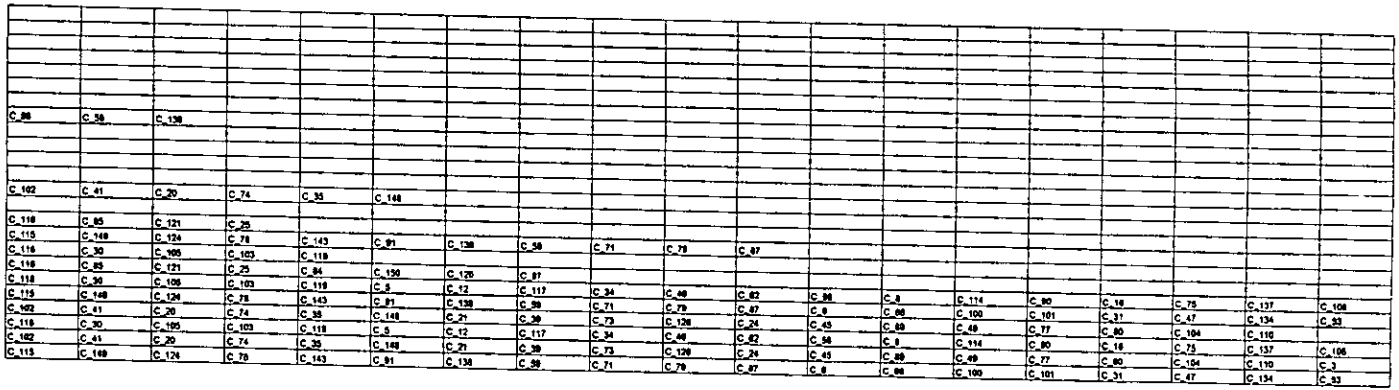
C_110	C_109	C_148	C_8	C_114	C_24	C_45	C_69	C_73	C_98	C_111	C_128	C_84	C_150	C_87	C_120	C_25	C_21	C_39

TABLA B3 : TABLA DE AMALGAMACIÓN DEL MÉTODO DE WARD

Distancia	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10	Obj. No. 11	Obj. No. 12	Obj. No. 13	Obj. No. 14	Obj. No. 15	Obj. No. 16	Obj. No. 17	Obj. No. 18
.0000000	C_16	C_75																
.0100000	C_95	C_109																
.0100000	C_115	C_149																
.0100000	C_86	C_112																
.0100000	C_64	C_123																
.0100000	C_2	C_57																
.0200000	C_11	C_128																
.0300000	C_15	C_81																
.0200000	C_52	C_82																
.0200000	C_81	C_141																
.0200000	C_78	C_86																
.0200000	C_138	C_144																
.0200000	C_57	C_72																
.0200000	C_83	C_126																
.0200000	C_33	C_96																
.0200000	C_36	C_148																
.0200000	C_44	C_88																
.0200000	C_86	C_100																
.0200000	C_43	C_89																
.0200000	C_84	C_139																
.0200001	C_51	C_47																
.0299999	C_85	C_121																
.0299999	C_123	C_132																
.0300000	C_84	C_125	C_107															
.0300000	C_1	C_85	C_106															
.0300000	C_46	C_82																
.0333333	C_78	C_86	C_118															
.0333333	C_80	C_83	C_135															
.0399999	C_35	C_65																
.0400000	C_18	C_67																
.0400000	C_133	C_140																
.0400001	C_103	C_118																
.0433333	C_54	C_88	C_112															
.0466667	C_51	C_139	C_144															
.0466667	C_88	C_100	C_101															
.0500000	C_13	C_127																
.0500001	C_17	C_102																
.0599999	C_11	C_128	C_88															
.0600000	C_4	C_23																
.0600000	C_14	C_33	C_98															
.0600000	C_3	C_92																
.0600000	C_20	C_74																
.0600001	C_7	C_27																
.0666667	C_18	C_67	C_42															
.0666666	C_22	C_29																
.0700000	C_24	C_45																
.0700001	C_9	C_114																
.0700001	C_26	C_130																
.1700000	C_21	C_39																
.1730003	C_4	C_23	C_17	C_102	C_41													
.1783333	C_31	C_47	C_134	C_83														
.1933332	C_84	C_130	C_129															
.1933333	C_89	C_63	C_133	C_54	C_88	C_112	C_65											
.1933334	C_26	C_115	C_149	C_124														
.2000000	C_70	C_129																
.2100001	C_22	C_29	C_94															
.2199999	C_13	C_127	C_147															
.2183333	C_51	C_139	C_144	C_80														
.2316666	C_15	C_81	C_63	C_131														
.2316666	C_14	C_33	C_96	C_133	C_140													
.2399999	C_49	C_77	C_85															

2600005	C_7	C_27	C_123	C_132														
2600002	C_30	C_105	C_103	C_119														
2638886	C_3	C_92	C_142	C_61	C_141													
1.300000	C_21	C_39	C_73	C_126														
1.461111	C_19	C_48	C_38	C_66	C_78	C_86	C_118	C_65	C_121	C_25								
1.672486	C_26	C_115	C_149	C_134	C_78	C_143	C_91	C_136	C_58	C_71	C_79	C_87						
1.752867	C_8	C_12	C_117	C_34	C_48	C_62	C_56	C_8	C_114	C_50								
1.880361	C_6	C_66	C_100	C_101	C_31	C_47	C_134	C_53	C_50	C_63	C_135	C_54	C_88	C_112	C_55	C_10	C_46	C_44
1.897024	C_2	C_57	C_123	C_111	C_15	C_61	C_63	C_131	C_7	C_27	C_123	C_132	C_13	C_127	C_147			
2.236453	C_9	C_57	C_62	C_32	C_22	C_78	C_94	C_11	C_126	C_89	C_116	C_30	C_105	C_103	C_119			
2.502857	C_21	C_39	C_73	C_126	C_24	C_43	C_88											
3.094887	C_3	C_12	C_117	C_34	C_48	C_62	C_56	C_8	C_114	C_50	C_16	C_75	C_137	C_108	C_148			
3.098187	C_2	C_57	C_122	C_111	C_15	C_61	C_63	C_131	C_7	C_27	C_123	C_132	C_13	C_127	C_147	C_4	C_23	C_17
3.407142	C_21	C_39	C_73	C_126	C_24	C_45	C_88	C_49	C_77	C_60	C_104	C_110						
3.671383	C_14	C_33	C_66	C_133	C_140	C_83	C_70	C_129	C_43	C_69	C_86	C_109	C_19	C_48	C_38	C_65	C_76	C_68
3.881614	C_1	C_65	C_108	C_36	C_145	C_84	C_125	C_107	C_37	C_72	C_18	C_67	C_42	C_51	C_139	C_144	C_60	C_26
4.217388	C_3	C_92	C_142	C_61	C_141	C_29	C_130	C_113	C_9	C_62	C_62	C_32	C_22	C_29	C_94	C_11	C_129	C_69
7.916232	C_14	C_33	C_66	C_133	C_140	C_83	C_70	C_129	C_43	C_69	C_86	C_109	C_19	C_48	C_38	C_65	C_76	C_68
8.233008	C_3	C_92	C_142	C_61	C_141	C_78	C_130	C_113	C_9	C_62	C_62	C_32	C_22	C_29	C_94	C_11	C_129	C_69
14.65388	C_1	C_65	C_108	C_36	C_145	C_84	C_125	C_107	C_37	C_72	C_18	C_67	C_42	C_51	C_139	C_144	C_60	C_26
23.90077	C_2	C_57	C_122	C_111	C_15	C_61	C_63	C_131	C_7	C_27	C_123	C_132	C_13	C_127	C_147	C_4	C_23	C_17
40.99241	C_3	C_92	C_142	C_61	C_141	C_29	C_130	C_113	C_9	C_62	C_62	C_32	C_22	C_29	C_94	C_11	C_129	C_69
191.2998	C_2	C_57	C_122	C_111	C_15	C_61	C_63	C_131	C_7	C_27	C_123	C_132	C_13	C_127	C_147	C_4	C_23	C_17
1002.847	C_1	C_65	C_108	C_36	C_145	C_84	C_125	C_107	C_37	C_72	C_18	C_67	C_42	C_51	C_139	C_144	C_60	C_26

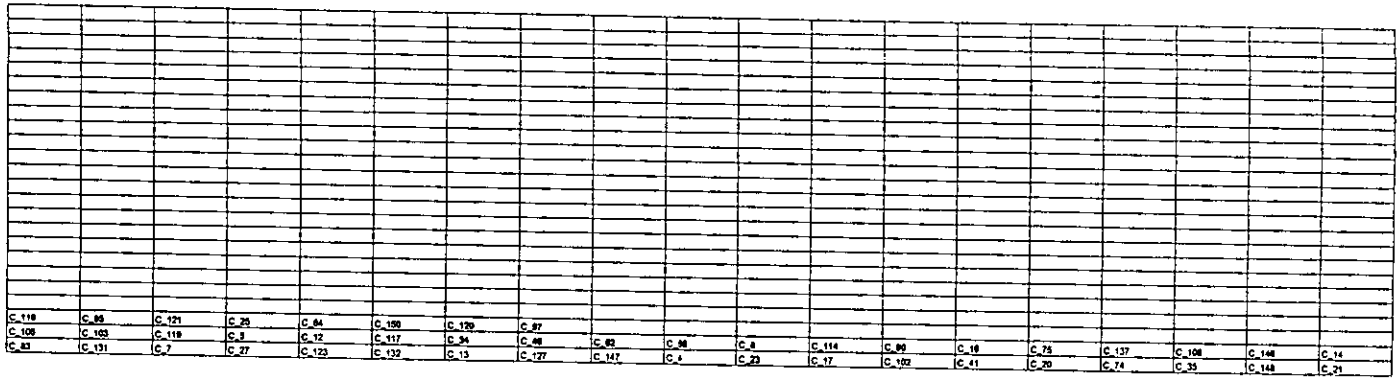
Obj. No. 19	Obj. No. 20	Obj. No. 21	Obj. No. 22	Obj. No. 23	Obj. No. 24	Obj. No. 25	Obj. No. 26	Obj. No. 27	Obj. No. 28	Obj. No. 29	Obj. No. 30	Obj. No. 31	Obj. No. 32	Obj. No. 33	Obj. No. 34	Obj. No. 35	Obj. No. 36	Obj. No. 37


C_86 C_58 C_139
C_102 C_41 C_20 C_74 C_35 C_148
C_116 C_85 C_121 C_25
C_115 C_149 C_124 C_78 C_143 C_91 C_138 C_58 C_71 C_79 C_87
C_116 C_30 C_105 C_103 C_118
C_116 C_85 C_121 C_25 C_84
C_118 C_30 C_105 C_103 C_119 C_130 C_120 C_87
C_115 C_149 C_124 C_78 C_5 C_12 C_117 C_34 C_46 C_62 C_86 C_8 C_114 C_80 C_16 C_75 C_137 C_106
C_102 C_41 C_20 C_74 C_143 C_91 C_138 C_58 C_71 C_79 C_87 C_8 C_86 C_100 C_101 C_31 C_47 C_134 C_53
C_118 C_30 C_105 C_103 C_35 C_148 C_21 C_39 C_73 C_129 C_24 C_45 C_88 C_49 C_77 C_80 C_104 C_110
C_102 C_41 C_20 C_74 C_118 C_5 C_12 C_117 C_34 C_46 C_62 C_56 C_8 C_114 C_80 C_16 C_75 C_137 C_106
C_115 C_149 C_124 C_78 C_35 C_148 C_21 C_39 C_73 C_129 C_24 C_45 C_88 C_49 C_77 C_80 C_104 C_110 C_3
C_143 C_91 C_138 C_58 C_71 C_79 C_87 C_8 C_86 C_100 C_101 C_31 C_47 C_134 C_53

Obj. No. 38	Obj. No. 39	Obj. No. 40	Obj. No. 41	Obj. No. 42	Obj. No. 43	Obj. No. 44	Obj. No. 45	Obj. No. 46	Obj. No. 47	Obj. No. 48	Obj. No. 49	Obj. No. 30	Obj. No. 31	Obj. No. 32	Obj. No. 33	Obj. No. 34	Obj. No. 35	Obj. No. 36

C_146																		
C_80	C_63	C_135	C_54	C_66	C_112	C_55	C_10	C_40	C_44	C_66	C_56	C_136						
C_146	C_14	C_53	C_66	C_133	C_146	C_63	C_70	C_129	C_43	C_66	C_66	C_106	C_19	C_48	C_36	C_85	C_78	C_66
C_92	C_142	C_61	C_141	C_26	C_130	C_113	C_8	C_52	C_62	C_32	C_27	C_29	C_94	C_11	C_128	C_66	C_116	C_30
C_80	C_63	C_135	C_54	C_66	C_112	C_55	C_10	C_40	C_44	C_66	C_56	C_136	C_2	C_57	C_122	C_111	C_15	C_61

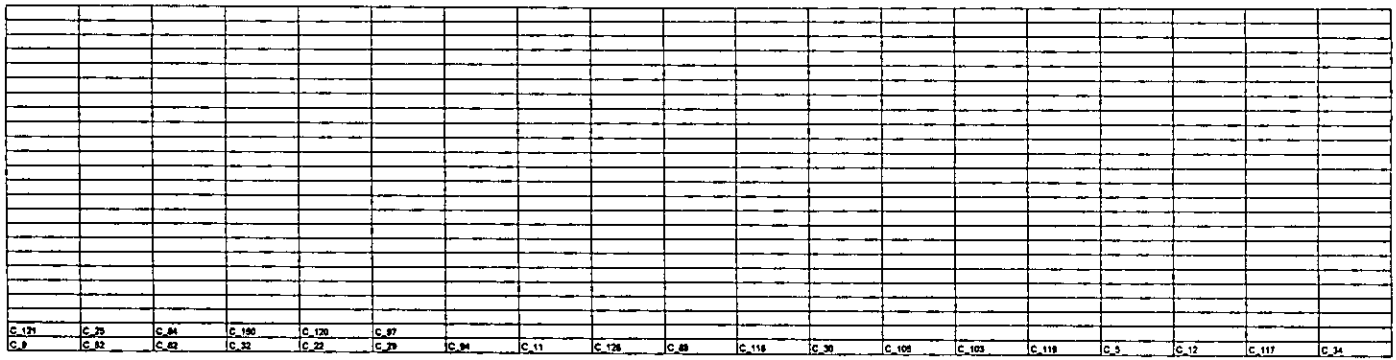
Obj. No. 57	Obj. No. 58	Obj. No. 59	Obj. No. 60	Obj. No. 61	Obj. No. 62	Obj. No. 63	Obj. No. 64	Obj. No. 65	Obj. No. 66	Obj. No. 67	Obj. No. 68	Obj. No. 69	Obj. No. 70	Obj. No. 71	Obj. No. 72	Obj. No. 73	Obj. No. 74	Obj. No. 75


C_118 C_89 C_121 C_25 C_64 C_150 C_129 C_97
C_108 C_163 C_119 C_3 C_12 C_117 C_94 C_46 C_93 C_98 C_8 C_114 C_80 C_16 C_75 C_137 C_108 C_146 C_14
C_83 C_131 C_7 C_27 C_123 C_132 C_13 C_127 C_147 C_4 C_23 C_17 C_102 C_41 C_70 C_74 C_35 C_148 C_21

Obj. No. 76	Obj. No. 77	Obj. No. 78	Obj. No. 79	Obj. No. 80	Obj. No. 81	Obj. No. 82	Obj. No. 83	Obj. No. 84	Obj. No. 85	Obj. No. 86	Obj. No. 87	Obj. No. 88	Obj. No. 89	Obj. No. 90	Obj. No. 91	Obj. No. 92	Obj. No. 93	Obj. No. 94

C_23	C_86	C_133	C_140	C_93	C_70	C_128	C_43	C_99	C_66	C_109	C_19	C_48	C_38	C_85	C_76	C_66	C_116	C_65
C_39	C_73	C_129	C_24	C_45	C_68	C_48	C_77	C_80	C_104	C_110	C_3	C_92	C_142	C_81	C_141	C_28	C_130	C_113

Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.
95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113


C_121
C_9
C_79
C_82
C_84
C_62
C_160
C_32
C_120
C_22
C_87
C_29
C_94
C_11
C_126
C_68
C_116
C_30
C_105
C_103
C_119
C_5
C_12
C_117
C_34

Obj. No. 114	Obj. No. 115	Obj. No. 116	Obj. No. 117	Obj. No. 118	Obj. No. 119	Obj. No. 120	Obj. No. 121	Obj. No. 122	Obj. No. 123	Obj. No. 124	Obj. No. 125	Obj. No. 126	Obj. No. 127	Obj. No. 128	Obj. No. 129	Obj. No. 130	Obj. No. 131	Obj. No. 132

C_46	C_62	C_58	C_9	C_114	C_90	C_16	C_75	C_137	C_109	C_146	C_14	C_33	C_96	C_133	C_140	C_83	C_70	C_129

Obj. No. 133	Obj. No. 134	Obj. No. 135	Obj. No. 136	Obj. No. 137	Obj. No. 138	Obj. No. 139	Obj. No. 140	Obj. No. 141	Obj. No. 142	Obj. No. 143	Obj. No. 144	Obj. No. 145	Obj. No. 146	Obj. No. 147	Obj. No. 148	Obj. No. 149	Obj. No. 150

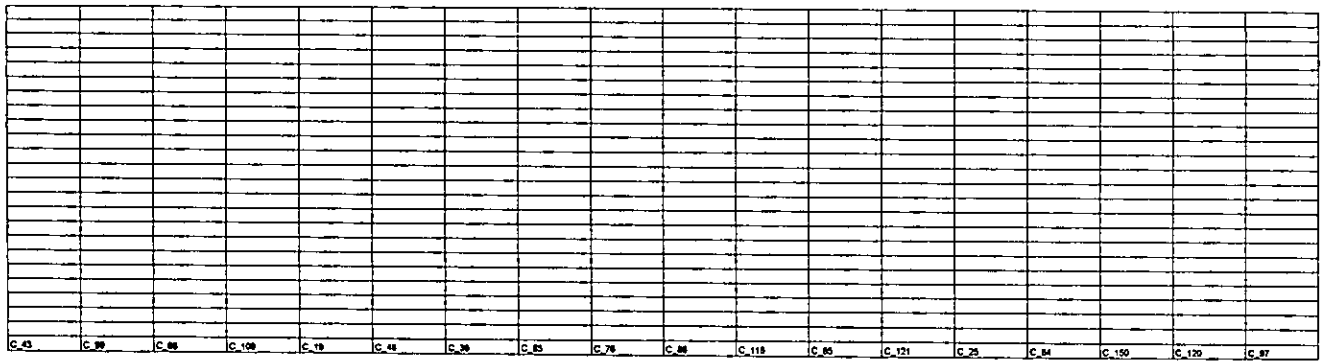

C_43
C_99
C_66
C_109
C_19
C_48
C_36
C_83
C_76
C_86
C_115
C_65
C_121
C_25
C_84
C_150
C_129
C_97

TABLA B4: TABLA DE CONGLOMERADOS POR EL MÉTODO DE LA LIGA SIMPLE

Iristype	número de elemento	Sepallen	Sepalwid	Petallen	Petalwid	número de conglomerado
SETOSA	1	5	3,3	1,4	0,2	1
SETOSA	6	4,6	3,4	1,4	0,3	1
SETOSA	10	4,6	3,6	1	0,2	1
SETOSA	18	5,1	3,3	1,7	0,5	1
SETOSA	26	5,5	3,5	1,3	0,2	1
SETOSA	31	4,8	3,1	1,6	0,2	1
SETOSA	36	5,2	3,4	1,4	0,2	1
SETOSA	37	4,9	3,6	1,4	0,1	1
SETOSA	40	4,4	3,2	1,3	0,2	1
SETOSA	42	5	3,5	1,6	0,6	1
SETOSA	44	4,4	3	1,3	0,2	1
SETOSA	47	4,7	3,2	1,6	0,2	1
SETOSA	50	4,6	3	1,4	0,3	1
SETOSA	51	5,1	3,6	1,6	0,2	1
SETOSA	53	4,8	3,4	1,9	0,2	1
SETOSA	54	5	3	1,6	0,2	1
SETOSA	55	5	3,2	1,2	0,2	1
SETOSA	58	4,3	3	1,1	0,1	1
SETOSA	59	5,8	4	1,2	0,2	1
SETOSA	60	5,1	3,8	1,9	0,4	1
SETOSA	63	4,9	3	1,4	0,2	1
SETOSA	64	5,1	3,5	1,4	0,2	1
SETOSA	67	5	3,4	1,6	0,4	1
SETOSA	68	4,6	3,2	1,4	0,2	1
SETOSA	71	5,7	4,4	1,5	0,4	1
SETOSA	72	5	3,6	1,4	0,2	1
SETOSA	78	5,4	3,4	1,5	0,4	1
SETOSA	79	5,2	4,1	1,5	0,1	1
SETOSA	87	5,5	4,2	1,4	0,2	1
SETOSA	88	4,9	3,1	1,5	0,2	1
SETOSA	91	5,4	3,9	1,7	0,4	1
SETOSA	95	5	3,4	1,5	0,2	1
SETOSA	96	4,4	2,9	1,4	0,2	1
SETOSA	100	4,7	3,2	1,3	0,2	1
SETOSA	101	4,8	3,1	1,5	0,2	1
SETOSA	108	5,1	3,4	1,5	0,2	1
SETOSA	107	5	3,5	1,3	0,3	1
SETOSA	112	4,9	3,1	1,5	0,1	1
SETOSA	115	5,4	3,7	1,5	0,2	1
SETOSA	124	5,4	3,9	1,3	0,4	1
SETOSA	125	5,1	3,5	1,4	0,3	1
SETOSA	134	4,8	3,4	1,6	0,2	1
SETOSA	135	4,8	3	1,4	0,1	1
SETOSA	136	4,5	2,3	1,3	0,3	1
SETOSA	138	5,7	3,8	1,7	0,3	1
SETOSA	139	5,1	3,8	1,5	0,3	1
SETOSA	143	5,4	3,4	1,7	0,2	1
SETOSA	144	5,1	3,7	1,5	0,4	1
SETOSA	145	5,2	3,5	1,5	0,2	1
SETOSA	149	5,3	3,7	1,5	0,2	1
VIRGINIC	2	6,4	2,8	5,6	2,2	2
VERSICOL	3	6,5	2,8	4,6	1,5	2
VIRGINIC	4	6,7	3,1	5,6	2,4	2
VIRGINIC	5	6,3	2,8	5,1	1,5	2
VIRGINIC	7	6,9	3,1	5,1	2,3	2
VERSICOL	8	6,2	2,2	4,5	1,5	2
VERSICOL	9	5,9	3,2	4,8	1,8	2
VERSICOL	11	6,1	3	4,6	1,4	2
VERSICOL	12	6	2,7	5,1	1,6	2
VIRGINIC	13	6,5	3	5,2	2	2
VERSICOL	14	5,6	2,5	3,9	1,1	2
VIRGINIC	15	6,5	3	5,5	1,8	2
VIRGINIC	16	5,8	2,7	5,1	1,9	2
VIRGINIC	17	6,8	3,2	5,9	2,3	2
VERSICOL	19	5,7	2,8	4,5	1,3	2
VIRGINIC	20	6,2	3,4	5,4	2,3	2
VIRGINIC	21	7,7	3,8	6,7	2,2	2
VERSICOL	22	6,3	3,3	4,7	1,6	2
VIRGINIC	23	6,7	3,3	5,7	2,5	2
VIRGINIC	24	7,6	3	6,6	2,1	2
VIRGINIC	25	4,9	2,5	4,5	1,7	2
VIRGINIC	27	6,7	3	5,2	2,3	2
VERSICOL	28	7	3,2	4,7	1,4	2
VERSICOL	29	6,4	3,2	4,5	1,5	2
VERSICOL	30	6,1	2,8	4	1,3	2
VIRGINIC	32	5,9	3	5,1	1,8	2

Iristype	número de elemento	Sepallen	Sepalwid	Petallen	Petalwid	número de conglomerado
VERSICOL	33	5,5	2,4	3,8	1,1	2
VIRGINIC	34	6,3	2,5	5	1,9	2
VIRGINIC	35	6,4	3,2	5,3	2,3	2
VERSICOL	38	5,4	3	4,5	1,5	2
VIRGINIC	39	7,9	3,8	6,4	2	2
VIRGINIC	41	6,7	3,3	5,7	2,1	2
VERSICOL	43	5,8	2,6	4	1,2	2
VIRGINIC	45	7,7	2,8	6,7	2	2
VIRGINIC	46	6,3	2,7	4,9	1,8	2
VERSICOL	48	5,5	2,6	4,4	1,2	2
VIRGINIC	49	7,2	3,2	6	1,8	2
VIRGINIC	52	6,1	3	4,9	1,8	2
VIRGINIC	56	6,1	2,8	5,6	1,4	2
VIRGINIC	57	6,4	2,8	5,6	2,1	2
VERSICOL	61	6,7	3,1	4,4	1,4	2
VIRGINIC	62	6,2	2,8	4,8	1,8	2
VERSICOL	65	5,6	3	4,5	1,5	2
VERSICOL	66	5,8	2,7	4,1	1	2
VERSICOL	69	6	2,9	4,5	1,5	2
VERSICOL	70	5,7	2,6	3,5	1	2
VIRGINIC	73	7,7	3	6,1	2,3	2
VIRGINIC	74	6,3	3,4	5,6	2,4	2
VIRGINIC	75	5,6	2,7	5,1	1,9	2
VERSICOL	76	5,7	2,9	4,2	1,3	2
VIRGINIC	77	7,2	3	5,8	1,6	2
VIRGINIC	80	7,1	3	5,9	2,1	2
VIRGINIC	81	6,4	3,1	5,5	1,8	2
VIRGINIC	82	6	3	4,8	1,8	2
VIRGINIC	83	6,3	2,9	5,6	1,8	2
VERSICOL	84	4,9	2,4	3,3	1	2
VERSICOL	85	5,6	2,7	4,2	1,3	2
VERSICOL	86	5,7	3	4,2	1,2	2
VIRGINIC	89	7,7	2,6	6,9	2,3	2
VIRGINIC	90	6	2,2	5	1,5	2
VERSICOL	92	6,6	2,9	4,6	1,3	2
VERSICOL	93	5,2	2,7	3,9	1,4	2
VERSICOL	94	6	3,4	4,5	1,6	2
VERSICOL	97	5	2	3,5	1	2
VERSICOL	98	5,5	2,4	3,7	1	2
VERSICOL	99	5,8	2,7	3,9	1,2	2
VIRGINIC	102	6,9	3,2	5,7	2,3	2
VERSICOL	103	6,2	2,9	4,3	1,3	2
VIRGINIC	104	7,6	2,8	6,1	1,9	2
VERSICOL	105	5,8	3	4,2	1,5	2
VIRGINIC	106	5,8	2,8	4,9	2	2
VERSICOL	109	6	2,2	4	1	2
VIRGINIC	110	7,3	2,9	6,3	1,8	2
VIRGINIC	111	6,7	2,5	5,8	1,8	2
VERSICOL	113	5,7	3,1	4,7	1,5	2
VERSICOL	114	6,3	2,3	4,4	1,3	2
VERSICOL	116	5,9	3	4,1	1,3	2
VERSICOL	117	6,3	2,5	4,9	1,5	2
VERSICOL	118	6,1	2,8	4,7	1,2	2
VERSICOL	119	6,4	2,9	4,3	1,3	2
VERSICOL	120	5,1	2,5	3	1,1	2
VERSICOL	121	5,7	2,8	4,1	1,3	2
VIRGINIC	122	6,5	3	5,8	2,2	2
VIRGINIC	123	6,9	3,1	5,4	2,1	2
VIRGINIC	126	7,2	3,6	6,1	2,5	2
VIRGINIC	127	6,5	3,2	5,1	2	2
VERSICOL	128	6,1	2,8	4,7	1,4	2
VERSICOL	129	5,8	2,8	3,8	1,3	2
VERSICOL	130	6,8	3,1	4,8	1,5	2
VIRGINIC	131	6,4	2,7	5,3	1,9	2
VIRGINIC	132	6,8	3	5,5	2,1	2
VERSICOL	133	5,5	2,5	4	1,3	2
VIRGINIC	137	5,7	2,5	5	2	2
VERSICOL	140	5,5	2,3	4	1,3	2
VERSICOL	141	6,9	3	4,4	1,4	2
VERSICOL	142	6,9	2,8	4,9	1,4	2
VIRGINIC	146	6,3	2,8	5,1	2,4	2
VERSICOL	147	6,7	3	5	1,7	2
VIRGINIC	148	6,2	3,4	5,4	2,3	2
VERSICOL	150	5,9	3	5,1	1,8	2

TABLA B5: TABLA DE CONGLOMERADOS POR EL MÉTODO DE WARD

Instype	número de elemento	Sepallen	Sepalwid	Petallen	Petalwid	número de conglomerado
SETOSA	1	5	3,3	1,4	0,2	1
SETOSA	6	4,6	3,4	1,4	0,3	1
SETOSA	10	4,6	3,6	1	0,2	1
SETOSA	18	5,1	3,3	1,7	0,5	1
SETOSA	26	5,5	3,5	1,3	0,2	1
SETOSA	31	4,8	3,1	1,6	0,2	1
SETOSA	36	5,2	3,4	1,4	0,2	1
SETOSA	37	4,9	3,6	1,4	0,1	1
SETOSA	40	4,4	3,2	1,3	0,2	1
SETOSA	42	5	3,5	1,6	0,6	1
SETOSA	44	4,4	3	1,3	0,2	1
SETOSA	47	4,7	3,2	1,6	0,2	1
SETOSA	50	4,8	3	1,4	0,3	1
SETOSA	51	5,1	3,8	1,6	0,2	1
SETOSA	53	4,8	3,4	1,9	0,2	1
SETOSA	54	5	3	1,6	0,2	1
SETOSA	55	5	3,2	1,2	0,2	1
SETOSA	58	4,3	3	1,1	0,1	1
SETOSA	59	5,8	4	1,2	0,2	1
SETOSA	60	5,1	3,8	1,9	0,4	1
SETOSA	63	4,9	3	1,4	0,2	1
SETOSA	64	5,1	3,5	1,4	0,2	1
SETOSA	67	5	3,4	1,5	0,4	1
SETOSA	68	4,6	3,2	1,4	0,2	1
SETOSA	71	5,7	4,4	1,5	0,4	1
SETOSA	72	5	3,6	1,4	0,2	1
SETOSA	75	5,4	3,4	1,5	0,4	1
SETOSA	79	5,2	4,1	1,5	0,1	1
SETOSA	87	5,5	4,2	1,4	0,2	1
SETOSA	88	4,9	3,1	1,5	0,2	1
SETOSA	91	5,4	3,9	1,7	0,4	1
SETOSA	95	5	3,4	1,5	0,2	1
SETOSA	98	4,4	2,9	1,4	0,2	1
SETOSA	100	4,7	3,2	1,3	0,2	1
SETOSA	101	4,6	3,1	1,5	0,2	1
SETOSA	106	5,1	3,4	1,5	0,2	1
SETOSA	107	5	3,5	1,3	0,3	1
SETOSA	112	4,9	3,1	1,5	0,1	1
SETOSA	115	5,4	3,7	1,5	0,2	1
SETOSA	124	5,4	3,9	1,3	0,4	1
SETOSA	125	5,1	3,5	1,4	0,3	1
SETOSA	134	4,8	3,4	1,6	0,2	1
SETOSA	135	4,8	3	1,4	0,1	1
SETOSA	136	4,5	2,3	1,3	0,3	1
SETOSA	138	5,7	3,8	1,7	0,3	1
SETOSA	139	5,1	3,8	1,5	0,3	1
SETOSA	143	5,4	3,4	1,7	0,2	1
SETOSA	144	5,1	3,7	1,5	0,4	1
SETOSA	145	5,2	3,5	1,5	0,2	1
SETOSA	149	5,3	3,7	1,5	0,2	1
VIRGINIC	2	6,4	2,8	5,6	2,2	2
VERSICOL	3	6,5	2,8	4,6	1,5	2
VIRGINIC	4	6,7	3,1	5,6	2,4	2
VIRGINIC	5	6,3	2,8	5,1	1,5	2
VIRGINIC	7	6,9	3,1	5,1	2,3	2
VERSICOL	8	6,2	2,2	4,5	1,5	2
VERSICOL	9	5,9	3,2	4,8	1,8	2
VERSICOL	11	6,1	3	4,6	1,4	2
VERSICOL	12	6	2,7	5,1	1,6	2
VIRGINIC	13	6,5	3	5,2	2	2
VERSICOL	14	5,6	2,5	3,9	1,1	2
VIRGINIC	15	6,5	3	5,5	1,8	2
VIRGINIC	16	5,8	2,7	5,1	1,9	2
VIRGINIC	17	6,8	3,2	5,9	2,3	2
VERSICOL	19	5,7	2,8	4,5	1,3	2
VIRGINIC	20	6,2	3,4	5,4	2,3	2
VIRGINIC	21	7,7	3,8	6,7	2,2	2
VERSICOL	22	6,3	3,3	4,7	1,6	2
VIRGINIC	23	6,7	3,3	5,7	2,5	2
VIRGINIC	24	7,6	3	6,6	2,1	2
VIRGINIC	25	4,9	2,5	4,5	1,7	2
VIRGINIC	27	6,7	3	5,2	2,3	2
VERSICOL	28	7	3,2	4,7	1,4	2
VERSICOL	29	6,4	3,2	4,5	1,5	2
VERSICOL	30	6,1	2,8	4	1,3	2
VIRGINIC	32	5,9	3	5,1	1,8	2

Instype	número de elemento	Sepallen	Sepalwid	Petallen	Petalwid	número de conglomerado
VERSICOL	33	5,5	2,4	3,8	1,1	2
VIRGINIC	34	6,3	2,5	5	1,9	2
VIRGINIC	35	6,4	3,2	5,3	2,3	2
VERSICOL	38	5,4	3	4,5	1,5	2
VIRGINIC	39	7,9	3,8	6,4	2	2
VIRGINIC	41	6,7	3,3	5,7	2,1	2
VERSICOL	43	5,8	2,6	4	1,2	2
VIRGINIC	45	7,7	2,6	6,7	2	2
VIRGINIC	46	6,3	2,7	4,9	1,8	2
VERSICOL	48	5,5	2,6	4,4	1,2	2
VIRGINIC	49	7,2	3,2	6	1,8	2
VIRGINIC	52	6,1	3	4,9	1,8	2
VIRGINIC	56	6,1	2,8	5,6	1,4	2
VIRGINIC	57	6,4	2,8	5,6	2,1	2
VERSICOL	61	6,7	3,1	4,4	1,4	2
VIRGINIC	62	6,2	2,8	4,8	1,8	2
VERSICOL	65	5,8	3	4,5	1,5	2
VERSICOL	66	5,6	2,7	4,1	1	2
VERSICOL	69	6	2,9	4,5	1,5	2
VERSICOL	70	5,7	2,8	4,5	1	2
VIRGINIC	73	7,7	3	6,1	2,3	2
VIRGINIC	74	6,3	3,4	5,6	2,4	2
VIRGINIC	75	5,6	2,7	5,1	1,9	2
VERSICOL	78	5,7	2,9	4,2	1,3	2
VIRGINIC	77	7,2	3	5,9	1,6	2
VIRGINIC	80	7,1	3	5,9	2,1	2
VIRGINIC	81	6,4	3,1	5,5	1,8	2
VIRGINIC	82	6	3	4,8	1,8	2
VIRGINIC	83	6,3	2,9	5,6	1,8	2
VERSICOL	84	4,9	2,4	3,3	1	2
VERSICOL	85	5,6	2,7	4,2	1,3	2
VERSICOL	86	5,7	3	4,2	1,2	2
VIRGINIC	89	7,7	2,6	6,9	2,3	2
VIRGINIC	90	6	2,2	5	1,5	2
VERSICOL	92	6,6	2,9	4,6	1,3	2
VERSICOL	93	5,2	2,7	3,9	1,4	2
VERSICOL	94	6	3,4	4,5	1,6	2
VERSICOL	97	5	2	3,5	1	2
VERSICOL	98	5,5	2,4	3,7	1	2
VERSICOL	99	5,8	2,7	3,9	1,2	2
VIRGINIC	102	6,9	3,2	5,7	2,3	2
VERSICOL	103	6,2	2,9	4,3	1,3	2
VIRGINIC	104	7,4	2,8	6,1	1,9	2
VERSICOL	105	5,9	3	4,2	1,5	2
VIRGINIC	108	6,8	2,8	4,8	2	2
VERSICOL	109	6	2,2	4	1	2
VIRGINIC	110	7,3	2,9	6,3	1,8	2
VIRGINIC	111	6,7	2,5	5,8	1,8	2
VERSICOL	113	6,7	3,1	4,7	1,5	2
VERSICOL	114	6,3	2,3	4,4	1,3	2
VERSICOL	116	5,6	3	4,1	1,3	2
VERSICOL	117	6,3	2,5	4,9	1,5	2
VERSICOL	118	6,1	2,8	4,7	1,2	2
VERSICOL	119	6,4	2,9	4,3	1,3	2
VERSICOL	120	5,1	2,5	3	1,1	2
VERSICOL	121	5,7	2,8	4,1	1,3	2
VIRGINIC	122	6,5	3	5,8	2,2	2
VIRGINIC	123	6,9	3,1	5,4	2,1	2
VIRGINIC	126	7,2	3,6	6,1	2,5	2
VIRGINIC	127	6,5	3,2	5,1	2	2
VERSICOL	128	6,1	2,9	4,7	1,4	2
VERSICOL	129	5,6	2,9	3,6	1,3	2
VERSICOL	130	6,6	3,1	4,9	1,5	2
VIRGINIC	131	6,4	2,7	5,3	1,9	2
VIRGINIC	132	6,8	3	5,5	2,1	2
VERSICOL	133	6,5	2,5	4	1,3	2
VIRGINIC	137	6,4	2,5	5	2	2
VERSICOL	140	6,5	2,3	4	1,3	2
VERSICOL	141	6,6	3	4,4	1,4	2
VERSICOL	142	6,8	2,8	4,8	1,4	2
VIRGINIC	146	6,8	2,8	5,1	2,4	2
VERSICOL	147	6,7	3	5	1,7	2
VIRGINIC	148	6,3	3,3	6	2,5	2
VERSICOL	150	5	2,3	3,3	1	2

BIBLIOGRAFÍA

Abraira, Victor S., Pérez de Vargas, Alberto L., Métodos Multivariantes en Bioestadística. Centro de Estudios Ramón Aceres, S.A.. Madrid, 1996.

Aguirre, Victor T., Artaloitia, Begoña C., Análisis Exploratorio de datos, Instituto Tecnológico Autónomo de México. México,1998.

Anderberg, Michael R., Cluster Analysis for Applications, Academic Press. New York & London, 1973.

Chatfield, Christopher, Collins, Alexander J., Introduction to Multivariate Analysis, Chapman & Hall. London, 1980.

Dillon, William R., Goldstein, Mathew., Multivariate Analysis: Methods and Applications, Wiley. New York,1984.

Esme, Diego C., Análisis Multivariante, Prentice hall, 1999.

Everitt, Brian S., Cluster Analysis, Wiley, New York, 1992.

Fraleigh, John A. Beauregard Raymond A., Algebra lineal, Addison – Wesley Iberoamericana. México, 1989.

Hair, Joseph F., Anderson, Rolph E., Tatham, Ronald L., Black, William C., Análisis Multivariante, Prentice Hall. Madrid, 1999.

Johnson, Richard A., Wichern, Dean W., Applied Multivariate Statistical Analysis, Prentice Hall, 1998.

Manly, B. F., Multivariate Statistical Methods, Chapman & Hall. London, 1986.

Rencher, Alvin C., Methods of Multivariate Analysis, Wiley. New York, 1995.

Seber, G.A.F., Multivariate Observations, Wiley. New York, 1984.