UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES

"ACATLAN"

ESTUDIO CONJUNTIVO DE ANALISIS ESTADISTICO MULTIVARIADO: ANALISIS DISCRIMINANTE Y ANALISIS DE CONGLOMERADOS CON APLICACION AL PROCESO DE CALIDAD DE LA INDUSTRIA HULERA

T E S I S

QUE PARA OBTENER EL TITULO DE

LICENCIADO EN MATEMATICAS APLICADAS Y COMPUTACION

P R E S E N T A

BIDHAULL ALVA HERNANDEZ

ASESOR: ING. ELVIRA BEATRIZ CLAVEL DIAZ

NAUCALPAN, EDO. DE MEX. FEBRERO, 1999

AGRADECIMIENTOS

Le doy gracias a Dios por haberme permitido tener una gran familia, que me fortaleció lo suficiente para poderme desempeñar como lo debo hacer.

Muchas gracias papá y mamá, por motivarme y animarme con cada palabra que recibo de ustedes, sin las cuales no hubiera podido desarrollar este trabajo. En especial les doy las gracias por haberme dado todas las bases necesarias para ser como soy.

Gracias Mauricio y gracias Hilda, por su apoyo y confianza, que me daban la fuerza para la realización de este trabajo.

Agradezco a la fuente de información que me permitió obtener las herramientas fundamentales para la creación de este trabajo, y a todas las personas que me brindaron su ayuda, para poder concluirlo, GRACIAS.

En especial le doy las gracias a una gran persona (M.) la cual tiene lo que muchos no tenemos, calidad humana.

ÍNDICE

INTRODUCCIÓN

INTRODUCCIÓN

Fenómenos multivariados[1] son todos aquellos que aparecen en el tiempo y en el espacio, en los que se manifiestan relaciones determinadas por las categorías que resultan de la medición de múltiples atributos[2] que poseen los objetos, elementos o unidades sujetos a estudio. Este tipo de fenómenos aparecen en todas las ramas de la ciencia, desde psicología hasta biología, y desde la matemática hasta la política. Los métodos para analizar fenómenos multivariados constituyen un área joven, importante y creciente de la estadística.

La particularidad del análisis multivariado es la consideración de un grupo de N objetos, de los cuales se tienen valores observados sobre m variables. El grupo de objetos puede ser completo o una muestra de una población. A su vez los objetos poseen varios atributos o variables de medición, los cuales tienen su propio espacio muestral finito o infinito, continuo o discreto, y pueden ser también un subconjunto de otro conjunto más grande de variables. Estos tipos de fenómenos por tener, en general, un número grande de objetos y éstos a su vez poder ser medidos en una gran cantidad de variables, resultan ser fenómenos de un comportamiento muy complejo y difícil de manejar así como de obtener. Es por eso que se requieren estudios con distintos propósitos, de los cuales los más importantes son:

1) *Simplificación estructural.* El objetivo aquí es "*ver el bosque por los árboles*", es decir, examinar los datos mediante varias transformaciones. Hay formas simples de representar la complejidad del fenómeno, por ejemplo, transformar un conjunto de variables interdependientes en variables independientes o reducir la dimensionalidad de lo complejo.

---

[1] Definición tomada del *DICCIONARIO FILOSOFICO*, Agustín Ezcurdia y Pedro Chávez Calderón, 1996, pág. 93.

[2] Atributo: Cada una de las cualidades o propiedades de un objeto. Enciclopedia *SALVAT DICCIONARIO*, Tomo 2, pág. 341.

2) *Clasificación.* En este tipo de estudio las preguntas a las que se desea dar respuesta son, ¿los objetos forman o pertenecen a grupos (conglomerados) mediante los cuales pueden ser distinguidos y clasificados?, o ¿cómo están esparcidos casualmente sobre el territorio de variación?.

3) *Agrupación de variables.* Ya que la clasificación es tomada como la agrupación de objetos, también en algunos estudios interesa saber si las variables caen de la misma forma dentro de grupos semejantes.

4) *Análisis de interdependencia lineal.* El objetivo es examinar la interdependencia (dependencia recíproca) entre variables o sea, las posibilidades de que las variables sean linealmente independientes o tengan una colinearidad, es decir, una situación en la cual una variable es una función lineal de otras o dado el caso no lo sea.

5) *Análisis de dependencia.* En el análisis de interdependencia lineal las variables son consideradas todas fijas, en bases semejantes, de esta manera son afectadas sobre sus relaciones mutuas. Por el contrario, en el *análisis de dependencia* una o más variables son tomadas para realizar el examen de sus dependencias sobre las demás como en análisis de regresión y en análisis de correlación.

6) *Construcción de hipótesis y pruebas de hipótesis.* En este caso el propósito es hacer una afirmación o suposición en base a las evidencias, señales o características que presentan los datos multivariados.

El método de *Análisis Discriminante* (AD) pertenece al tipo de propósito del *análisis de interdependencia lineal* y sus raíces surgen por los años 1920's. El estadístico inglés Karl Pearson propuso un coeficiente que interpretaba la distancia entre grupos indexados (*Coeficiente de semejanza racial*), es decir, un coeficiente que indica qué tanto está un grupo separado de otro, o se puede considerar como la distancia entre dos grupos. Este tipo de coeficiente fue estudiado extensivamente por G. M. Morant en los años 1920's. Por esos mismos años en la India hubo otros estudios sobre distancias

entre grupos de datos que fueron formalizados por P. C. Mahalanobis en los 1930's, y en esa misma década la idea de la distancia intergrupal multivariada fue traducida a una combinación lineal de variables derivadas para el propósito de discriminación entre grupos por R. A. Fisher. Los escritos acerca del AD aparecieron en las primeras tres o cuatro décadas y fueron enfocados en la predicción de miembros de grupos (*Análisis Discriminante Predictivo*, ADP), pero hasta la década de los 1960's la combinación lineal de variables se consideró para propósitos de interpretar efectos relevados por un análisis multivariado de varianza (MANOVA), y este aspecto del AD es llamado *Análisis Discriminante Descriptivo* (ADD). El ADD tiene como principal interés describir los diferentes efectos que se producen en los objetos mediante diferentes agrupaciones de variables. En cambio en el ADP el interés básico es predecir los miembros de grupos de acuerdo a la agrupación de las variables que puede existir.

El AD es una técnica mediante la cual se obtiene un resultado por el que se puede describir el comportamiento de un fenómeno *objeto* (variable dependiente) por medio de *varios supuestos* (variable independiente). Si la variable dependiente que se desea describir es *dicótoma* (por ejemplo, macho - hembra) o *multicótoma* (por ejemplo, alta - media - baja) y por lo tanto no es un dato métrico[3], entonces la técnica multivariada AD es la apropiada para analizar a esta variable. El AD es útil en situaciones donde la muestra total puede ser dividida en grupos basados sobre una variable dependiente que tiene distintas clases conocidas. Los objetivos primarios de este análisis de interdependencia lineal son el entender las diferencias de grupos y predecir la probabilidad de que una entidad (individuo u objeto) pertenecerá a una clase particular o grupo basado sobre distintos datos métricos o variables independientes. Este método es de aplicación universal, incluso en muchas situaciones, el objetivo primario es identificar el grupo al cual un objeto pertenece. En cada instancia, los fenómenos se encuentran en grupos y pueden, esto es lo esperado, ser predichos, descritos o explicados, por un conjunto de variables independientes seleccionadas mediante el análisis señalado.

---

[3] Dato Métrico: Son datos cuantitativos, mediciones usadas para identificar o describir objetos, no solamente en las bases del tipo, sino también por el valor o grado para el cual el objeto puede ser caracterizado por un atributo particular.
Dato No Métrico: Son datos cualitativos, atributos, características o propiedades categóricas que pueden ser usadas para identificar o describir un objeto. Ejemplo, ocupación (Físico, Doctor, Profesor, etc.). También conocidos como datos nominales u ordinales. Hair, Joseph F., *MULTIVARIATE DATA ANALYSIS with Readings*, 1987, pág. 2.

Por otro lado, el *Análisis de Conglomerados* (AC) es un método analítico que sintetiza la información de la población de objetos, por medio de la clasificación de los mismos en un cierto número de grupos que se pueden considerar como poblaciones en miniatura, puesto que cada grupo conserva las mismas propiedades de la población total pero con un menor número de objetos, los más representativos y significantes.

Específicamente, el objetivo es el clasificar una muestra de objetos (unidades o individuos) en un número pequeño de grupos mutuamente exclusivos, basados sobre las similaridades entre las entidades. En el AC a diferencia del AD, los grupos no están predefinidos. En el AD se supone que los objetos o unidades a analizar pertenecen ya a un grupo predefinido, pero en el AC los grupos todavía no están preasignados. De esa forma el AC toma la preferencia sobre el AD de ser el primer paso de un análisis con tal de definir los grupos de objetos. El AC usualmente inicia con la etapa en que se obtiene un tipo de medición en base a las variables independientes, que de alguna forma nos indica la *proximidad* o asociación entre las entidades en orden para determinar si varios grupos realmente existen en la muestra, y después perfila los objetos en orden para determinar la agrupación que puedan tener. Este paso podría ser perfeccionado al aplicarse el AD a los grupos identificados por el AC.

El AC es un método que clasifica los objetos en grupos que son homogéneos en sí mismos y heterogéneos entre ellos y esta clasificación es realizada con el conjunto completo de variables, pero *¿es necesario usar todo el conjunto de variables, para clasificar a los objetos correctamente?* o *¿podría ser menor el número de variables que mejor clasifican a los objetos?*.

Los grupos de datos están compuestos de mediciones de diferentes variables que describen a cada objeto del grupo, es por eso que al realizar un análisis a estos datos es apropiado considerar de manera simultánea todas las variables (ADP) sin olvidar a los objetos (AC).

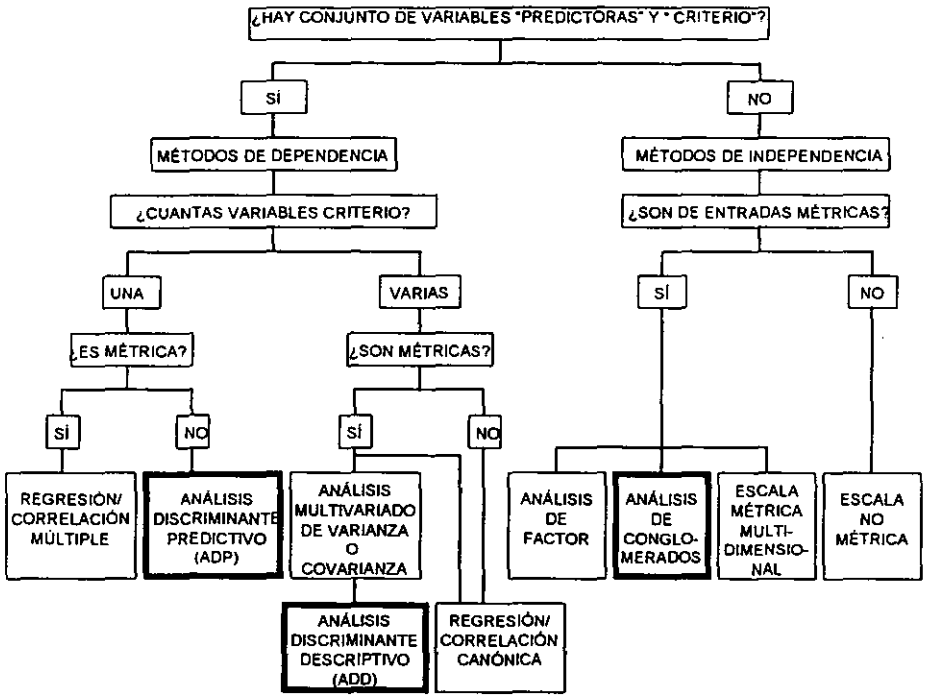


Figura 1. Clasificación de Métodos Multivariados

Para poder tener una mejor idea de qué lugar ocupan los métodos de AD y de AC en la *Estadística Multivariada*, y qué tan relacionados están ambos métodos, se muestra un organigrama (Figura 1) basado en los casos en que es apropiado aplicar un análisis multivariado. En este esquema se presentan claramente las condiciones para aplicar los distintos métodos de la estadística multivariada.

El AD y el AC son dos tipos de análisis completamente diferentes y con distintos fines, pero en estos dos métodos sus objetivos básicos son respecto a las dos partes fundamentales de la estadística multivariada (*objetos y variables predictoras*), entonces se puede pensar *que los dos métodos podrían formar un sólo método que defina a los mejores grupos en los que se pueden clasificar a los objetos basándose en las variables predictoras que mejor los describen*, o como un método que encuentra el grupo de variables predictoras que mejor describen a cada grupo predefinido de objetos.

El AC obtiene varios grupos de objetos, los cuales son homogéneos en su interior, pero muy heterogéneos en su exterior, estos grupos (*conglomerados*) puede ser que no estén compuestos por las variables que de mejor manera predicen su comportamiento con respecto al fenómeno del que surgen y pertenecen, entonces *¿qué variables predictoras son las que mejor influyen en el objetivo del análisis de conglomerados?*.

Como se nota en la figura 1, entre los métodos de análisis estadístico multivariado AD y AC no se aprecian claramente las relaciones que interactúan entre el funcionamiento de ambos métodos. Estas relaciones son difíciles de identificar, por no encontrarse de una manera superficial, sino por estar en el interior del desarrollo de ambos métodos. Al considerar estos tipos de relaciones que unen a los dos métodos de análisis, surgen ciertas preguntas que generan el objetivo del trabajo:

- Si se tienen variables para un grupo predefinido de objetos,
  ¿es aceptable aplicar primero el AD para asegurar que ese es el grupo adecuado de variables que describen al fenómeno?
- ¿Qué son el AD y AC de sí mismos
  - comprobantes de validez de sus resultados?
  - refuerzos de sus resultados?
- ¿Existe una relación conceptual entre ambos métodos de análisis? o ¿El AD y el AC son complementarios de sí mismos?

Las respuestas a cada una de las preguntas puede ser que estén basadas en los supuestos de cada uno de los dos métodos o en los conceptos de cualquiera de sus procedimientos, o tal vez en los efectos de los resultados, es por eso que en el documento se desarrolla un estudio de manera conjunta de los métodos de análisis de conglomerados y de análisis discriminante predictivo examinando los principios teóricos matemáticos en los que subyace el principio de proximidad, para establecer a través de éste ¿cuáles serían las k agrupaciones, si las hay, que se podrían dar entre los objetos de un proceso de producción de calidad?, así como para predecir ¿cuál sería la agrupación de variables que podría darse o existir en dicho proceso, que proporcionara una mejor clasificación de objetos y en consecuencia un mejor control y descripción de

la calidad del producto?

Para dar respuesta a estas preguntas se requiere de una síntesis de los fundamentos teóricos y metodológicos del ADP y del AC, que facilitará la realización de un estudio de la pertinencia, agrupación y selección que se puede hacer con un conjunto de datos sobre las variables de los objetos en términos de semejanza o distancia, que puede conducir al mejoramiento del proceso de un producto para maximizar su calidad.

Para poder desarrollar el estudio de una manera adecuada, en el **Capítulo 1: Análisis de Conglomerados**, se muestran los principales conceptos básicos para conformar el entendimiento teórico y práctico del AC, esto con el propósito de poder distinguir, los conceptos de sus procedimientos y etapas, así como sus supuestos y efectos, que se desarrollan en este método de análisis estadístico multivariado. Se muestran las etapas del método teniendo un subcapítulo de las principales medidas de proximidad que pueden ser usadas por el método, así como las mejores formas de transformar las variables no-métricas a métricas. Después se muestran los distintos procesos de conglomeración (jerárquicos y no-jerárquicos) a usarse, dependiendo del caso a analizar, y sus maneras de interpretar los resultados obtenidos.

Al igual que en el capítulo anterior, mostrando los principales conceptos básicos para obtener un entendimiento teórico y práctico pero ahora con respecto al **Análisis Discriminante** en el **Capítulo 2**, se muestran las etapas del ADP, presentando como subcapítulos, cada uno de los procesos para perfilar las variables y así poder desarrollar y validar la *función discriminante*, y obtener los resultados que caracterizan al método.

El **Capítulo 3: Clasificación y Discriminación**, es la parte práctica del texto, en el cual se muestra la aplicación de los métodos de ADP y AC a un mismo caso, que tiene las condiciones necesarias para poderse aplicar de una manera conjunta los procedimientos de los métodos de clasificación y discriminación; y así poderse comprobar los resultados supuestos y deseados, por los que se podrá conjuntar la parte teórica con la práctica y de esta forma responder las preguntas antes hechas y así corroborar los supuestos que se han planteado.

Lo anterior es posible aplicando el análisis de conglomerados al grupo de datos de un periodo mensual de producción de llantas, para establecer las k agrupaciones que se pueden dar entre los objetos (llantas). Y como siguiente paso aplicar el análisis discriminante predictivo con las k agrupaciones definidas anteriormente para predecir cuál es el subconjunto de variables que puede existir y que proporcione la mejor descripción de los objetos.

El procedimiento de la fase práctica se hace en 3 etapas, mediante el programa estadístico *SPSS* (*Statistical Package for Social Sciences*), con el fin de mostrar que no es exclusivo para ciencias sociales:

1) Se hace un AC con las m variables originales, obteniendose k conglomerados, de los N objetos.
2) Se realiza un ADP a los N objetos clasificados en los k conglomerados en base a las m variables originales, y se obtendrá una función lineal discriminante compuesta por las h variables ($h \leq m$) que mejor describen al fenómeno.
3) Se aplica el AC a los objetos medidos con las h variables que se obtuvieron en la etapa anterior.
   - *¿ Se obtendrán los mismos k conglomerados que se obtuvieron en la etapa 1?*
   - *¿ Se obtendrán otros conglomerados mejor definidos para el objetivo del análisis ?*

- Si se obtienen los mismos conglomerados que se obtuvieron en la primera etapa, entonces se puede decir que, ***el ADP sirvió como una prueba de validez para el AC.***

- Si se obtienen otros conglomerados distintos a los de la primera etapa, pero mejor apegados al objetivo del análisis, entonces, ***el ADP es como un complemento para mejorar los resultados del AC.***

- Si se obtienen otros conglomerados distintos a los de la primera etapa, sin apegarse al objetivo del análisis, entonces, ***el ADP no es un complemento y no se relaciona***

de una forma conjunta con el AC.

Con el propósito de que el lector cuente con un esquema para poderse auxiliar y conocer el lugar en que se ubican cada uno de los componentes que forman a los datos multivariados, y así entender mejor la manera en que operan cada una de las funciones de proximidad en el AC y las reglas de clasificación en el ADP, el que sustenta este trabajo preparó la siguiente matriz objetos-variables.

VARIABLES

OBJETOS		X_1	X_2	.	.	X_j	.	.	X_m
	Y_{g1}	x_{g11}	x_{g12}	.	.	x_{g1j}	.	.	x_{g1m}
	Y_{g2}	x_{g21}	x_{g22}	.	.	x_{g2j}	.	.	x_{g2m}
	.	.	.			.			.
	.	.	.			.			.
	.	.	.			.			.
	Y_{gi}	x_{gi1}	x_{gi2}	.	.	x_{gij}	.	.	x_{gim}
	.	.	.			.			.
	.	.	.			.			.
	.	.	.			.			.
	Y_{gn_g}	x_{gn_g1}	x_{gn_g2}	.	.	x_{gn_gj}	.	.	x_{gn_gm}

donde x_{gij} es el dato (crudo o estandarizado) del objeto i del conglomerado g medido en la variable j

$g = 1, 2, \ldots, k$ (número de conglomerados)

$i = 1, 2, \ldots, n_g$ (cantidad de objetos en el conglomerado g)

$j = 1, 2, \ldots, m$ (número de variables independientes "predictoras")

CAPÍTULO 1

ANÁLISIS DE CONGLOMERADOS

1.1 ESTANDARIZACIÓN

En muchos análisis de datos, las mediciones que se tienen de los objetos (que son las variables que los describen desde distintos puntos de vista) que serán analizados, no son medidas con las mismas unidades. En efecto, pueden ser variables de diferentes tipos, como categóricas, ordinales, nominales, cuantitativas, binarias, etc. Por ejemplo, hay ocasiones en que las variables pueden describir al objeto con varias unidades métricas tales como, libras, pulgadas, gramos, etc., pero a veces en distintas categorías (multicótoma) como bajo, medio, alto, macho, hembra, etc. De aquí surge el problema de cómo considerar los distintos tipos de mediciones de un objeto, cuando se pretende obtener un coeficiente que indique la cercanía que hay de un objeto con otro, sin ignorar las relaciones que haya entre las variables que los describen. La solución que más se sugiere para este tipo de problema, es la estandarización de los datos (al ser estandarizados tendrán una distribución normal con media cero y varianza de unidad, N(0,1)), esto se hace utilizando la desviación estándar y la media, calculadas del grupo completo de observaciones de una variable que describe a los objetos a ser conglomerados, haciéndose de la siguiente manera,

$$X_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

donde

X_{ij} es el dato estandarizado del i-ésimo objeto medido con la j-ésima variable

x_{ij} es el dato crudo[1] a estandarizar del i-ésimo objeto medido con la j-ésima variable

$\bar{x}_j$ es la media de la muestra de la j-ésima variable

σ_j es la desviación estándar de la j-ésima variable

[1] Se le llama dato crudo, a los datos originales, que son aquellos que permanecen de la manera como se obtuvieron inicialmente. (n. a.)

Pero esto no es una solución completa y definitiva para este problema, ya que se ha mostrado que la estandarización de datos puede traer serias desventajas, y diluir diferencias entre grupos sobre las variables, las cuales son las mejores que distinguen y diferencian los objetos a tratar (Fleiss y Zubin (1969)). Una de las desventajas es que al estandarizar cada objeto separadamente, se ignora a las posibles correlaciones entre las variables.

Cuando las variables medidas son de diferentes tipos, varias sugestiones se han hecho para saber cómo pueden ser usadas en forma conjunta, y así no ignorar las relaciones que pueden existir entre ellas. Una de esas sugestiones es el de obtener un coeficiente que indique el nivel de similitud de un objeto con otro, utilizando todas las mediciones de los dos objetos; a este coeficiente se le conoce como coeficiente de *proximidad*[2]. Una manera sencilla es el hacer binarios a los objetos, esta conversión tiene la ventaja de ser muy sencilla, pero tiene la desventaja de sacrificar mucha información útil. Una mejor alternativa es el de usar un coeficiente de proximidad, el cual puede incorporar información de diferentes tipos de variables en un modo sensible.

1.2 MEDIDAS DE PROXIMIDAD

Los coeficientes de proximidad se usan para indicar dos tipos de semejanzas entre los objetos, los cuales son:

similaridad s_{ij} indica la similitud que hay del objeto i con el objeto j, donde un valor pequeño indica que dos objetos no son tan similares y un valor grande indica que dos objetos son muy similares.

disimilaridad d_{ij} es el concepto de la distancia entre el objeto i con el objeto j, donde un valor pequeño indica que dos objetos son muy cercanos y un valor grande indica que dos objetos están muy lejanos el uno con el otro.

En el contexto se hablará de ambas medidas de proximidad, pero esto no es ningún

[2] Semejanza o analogía de una variable con otra. *Everitt, B.S.,CLUSTER ANALYSIS,* pág. 37.

problema, porque las medidas de *disimilaridad* pueden ser transformadas en medidas de *similaridad* mediante la siguiente fórmula

$$s_{ij} = 1/(1 + d_{ij})$$

pero es recomendable establecer desde un principio el tipo de proximidad con el que se llevará a cabo el análisis.

Un coeficiente de proximidad indica la fuerza de la relación entre dos objetos, dados los valores de un conjunto de m variables comunes para ambos. La *similaridad* entre los objetos i y j, es una función de sus valores observados, es decir,

$$s_{ij} = f(\mathbf{x}_i, \mathbf{x}_j)$$

donde $\mathbf{x}'_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ y $\mathbf{x}'_j = [x_{j1}, x_{j2}, \ldots, x_{jm}]$ son los valores de las variables observadas para los objetos i y j. Muchas funciones han sido propuestas dependiendo del tipo de las variables tratadas (cuantitativas, categóricas, binarias, ordinales, etc.) y particularmente del tipo de objeto.

Hay distintos tipos de medidas de *similaridad*, usualmente se consideran como una relación simétrica $s_{ij} = s_{ji}$, por eso la mayoría de los coeficientes de *similaridad* no son negativos y tienen un límite superior que es la unidad, aunque algunos se comportan con una naturaleza correlacional, así que en ocasiones ciertas medidas de *similaridad* tienen un rango de un coeficiente de correlación

$$-1 \leq s_{ij} \leq 1$$

Asociada con cada medida de *similaridad* y limitada por cero y uno, es una *disimilaridad* $d_{ij} = 1 - s_{ij}$, la cual es simétrica y no negativa. Con esta medida, el grado de *similaridad* entre dos objetos se incrementa con s_{ij} y se decrementa con el incremento de d_{ij}, y de esta manera un objeto tiene su máximo grado de *similaridad* con sí mismo, respetando ambas medidas de proximidad, así $s_{ii} = 1$ y $d_{ii} = 0$.

Si ζ es el conjunto de objetos a clasificar, entonces un coeficiente de *disimilaridad* es una función definida de la siguiente manera:

$d : \zeta \times \zeta \Rightarrow \mathbf{R}$ tal que

$$0 \leq d_{ij} \leq 1 \qquad \forall\, i,j \in \zeta$$
$$d_{ii} = 0 \qquad \forall\, i \in \zeta$$
$$d_{ij} = d_{ji} \qquad \forall\, i,j \in \zeta$$

Así la *disimilaridad* (distancia) del objeto *i* y el objeto *j* toma como máximo valor el uno y como mínimo valor el cero. La *disimilaridad* de un objeto con sí mismo es nula, y la *disimilaridad* del objeto *i* al objeto *j* es la misma que la del objeto *j* al objeto *i*.

1.2.1 DISIMILARIDADES PARA DATOS CUANTITATIVOS

Existe un gran número de tipos de disimilaridades, dependiendo del tipo de variables con las que se esté tratando, por ejemplo, si todas las variables son cuantitativas, hay medidas de proximidad exclusivas para este caso, pero sólo se pueden usar si todas las variables son de este tipo. Para este tipo de variables una disimilaridad tiene la idea de una distancia que hay entre un objeto de otro, esa es la razón por la que muchas medidas de disimilaridad tienen como base el concepto fundamental sobre las distancias entre dos puntos, que es la distancia mediante el *Teorema de Pitágoras*, que comúnmente se conoce como la Distancia Euclidiana[3].

A continuación se mostrarán las disimilaridades más comunes en los análisis de conglomerados. Supongamos que existe una muestra de N objetos, y cada uno es medido en m variables; se usarán los subíndices i, j para indicar objetos, y el subíndice k para indicar variables en las fórmulas siguientes. El valor numérico x_{ik} es una observación de la k-ésima variable sobre el i-ésimo objeto en la muestra y d_{ij} es la

[3] Se le conoce como Distancia Euclidiana, por el hecho de tener como clave de su comprobación las 47 proposiciones de Euclides. W. M. Jackson. *ENCICLOPEDIA PRACTICA JACKSON Conjunto de Conocimientos para la Formación Autodidáctica*, Tomo VIII, pág. 191.

disimilaridad entre el objeto *i* y el objeto *j*.

(1) *Distancia Euclidiana cuadrada*:

Es la distancia euclidiana al cuadrado entre dos objetos, que se obtiene al sumar las diferencias cuadradas entre los valores de las variables.

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

(2) *Distancia Euclidiana*:

Es la disimilaridad mediante el teorema de Pitágoras, que es la raíz cuadrada de la suma de las diferencias al cuadrado, entre los valores de las variables de los objetos a conglomerar.

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

(3) *Distancia Euclidiana Escalada*:

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

para algún peso conveniente w_k, el cual puede ser cualquiera de

w_k = (desviación estandar de la *k*-ésima variable)$^{-1}$

o w_k = (rango de la *k*-ésima variable)$^{-1}$

el efecto de ambas selecciones es el igualar la importancia de cada variable en la muestra.

(4) *City-Block*:

De esta aproximación es aconsejable su uso cuando los objetos no están correlacionados y sus unidades de medición son compatibles.

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

(5) *Métrico de Minkowski*:

Esta disimilaridad es la generalización de la distancia Euclidiana y la de City-Block

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

Donde λ es la cantidad que se considera para aumentar la relatividad de los objetos; para el caso de λ=1 se da la medida de similaridad *City-Block*, mientras que en el caso de λ=2 da la *distancia euclidiana*. Mientras λ se incremente, se está aumentando la consideración de la mayoría de los objetos relativos a las que son similares.

(6) *Métrico de Camberra*: $$d_{ij} = \sum_{k=1}^{m} \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})}$$

(7) *Coeficiente de Czekanowski*: $$d_{ij} = 1 - \frac{2\sum_{k=1}^{m} min(x_{ik}, x_{jk})}{\sum_{k=1}^{m} (x_{ik} + x_{jk})}$$

(8) *Distancia de Mahalanobis*

Todas las disimilaridades anteriores tienen el inconveniente de resultar sustancialmente afectadas por cambios de unidad de medición en las variables, lo que lógicamente conduce a agrupaciones diferentes de los objetos de la muestra. En cambio la *distancia de Mahalanobis* elimina los problemas de dimensionalidad al ser invariante respecto a cambios de origen y escala, y se define como:

$$d_{ij} = [(X_i - X_j)' S^{-1} (X_i - X_j)]^{1/2}$$

donde X_i y X_j son los vectores de observaciones de los objetos i y j respectivamente, y S^{-1} es la inversa de la matriz de covarianza de las m variables

que caracterizan a cada uno de los N objetos que se desean clasificar (ver secciones 2.2 y 2.3).

1.2.2 MEDIDAS DE PROXIMIDAD PARA DATOS BINARIOS

Puede haber ocasiones en que todas las variables tengan solamente dos valores (dicótomas), ya sean valores de categorías o clases, o que representen la presencia o ausencia de algún atributo. Como se comentó anteriormente, existen medidas de proximidad para este tipo de variables siempre y cuando todas sean del mismo tipo. Los coeficientes de similaridad que se usan para este tipo de variables son los más simples y más comúnmente usados. En algunos casos los valores de las variables pueden relacionarse a (1) como la presencia o (0) ausencia de alguna cualidad, en otros simplemente a las dos alternativas equivalentes, alta / baja, áspero / liso, etc.

Tales datos para los dos objetos, i y j pueden ser acomodados en una tabla de 2x2 (como se muestra a continuación), donde a, b, c y d son el número de veces en que se presentan las diferentes formas de combinacion de los valores de los objetos y $p = a + b + c + d$. Esta tabla de 2x2 indica la suma de los valores para cada combinación de valores de dos objetos, pero no muestra el porcentaje con respecto a los mismos valores como indica una tabla de contingencia, es por eso por lo que no deben ser confundidas estos dos tipos de tablas.

Contadores de variables binarias para dos objetos.

		objeto i 1	0	
objeto j	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	p

Muchos coeficientes de proximidad han sido propuestos al combinar las cantidades de a, b, c y d. Las tres medidas más comunes son las siguientes:

(8) s_{ij} = (*simple coeficiente de ser igual*)= $\frac{a+d}{p}$ y su disimilaridad relativa es

$d_{ij} = 1 - s_{ij} = 1 - \frac{a+d}{p} = \frac{b+c}{p}$. Así la disimilaridad entre las dos unidades es medida como la proporción de variables que muestra desacuerdo en sus valores registrados entre las dos unidades, y ésta es una medida muy aceptable en la mayoría de los estudios.

(9) *Coeficiente de Jaccard*: $s_{ij} = \frac{a}{a+b+c}$ y su disimilaridad relativa es

$d_{ij} = \frac{b+c}{a+b+c}$. Este coeficiente es particularmente apropiado cuando los dos valores 1 y 0 de la dicotomía representan respectivamente presencia y ausencia de un atributo. En este caso, uno puede frecuentemente sostener que la ausencia de un atributo en ambas unidades no debería contribuir en sus grados de *igualdad*. En tales casos, el divisor $a + b + c$ es más apropiado que el divisor $p = a+b+c+d$ dado en (8).

(10) *Coeficiente de Czekanowski*: $s_{ij} = \frac{2a}{2a+b+c}$ y su respectiva disimilaridad es

$d_{ij} = \frac{b+c}{2a+b+c}$. Este está en el mismo sentido que (9), pero cada marca de 1 1 es dada como un peso extra para compensar todas las marcas de 0 0 dejadas.

Los diferentes coeficientes de similaridad pueden tener valores muy diferentes para el mismo grupo de datos. Supongamos por ejemplo, a dos objetos que tienen las siguientes marcas sobre diez variables binarias.

		Variable									
		1	2	3	4	5	6	7	8	9	10
Objeto	**1**	1	0	0	0	1	1	0	0	1	0
Objeto	**2**	0	0	0	0	1	0	0	1	1	0

La correspondiente tabla 2x2 es

		Objeto 1		
		1	**0**	
Objeto 2	1	2	1	3
	0	2	5	7
		4	6	10

Los valores tomados por los diferentes coeficientes de similaridad mencionados son (8) 0.70, (9) 0.40, (10) 0.57. Estos coeficientes toman diferentes valores para el mismo par de objetos, sería relativamente sin importancia si los coeficientes fueran conjuntamente *monotómicos*, en el sentido que si todos los valores para diferentes pares de objetos sobre un coeficiente fueran ordenados, así que ellos formaran una serie monotómica (que es una serie por la cual cualquier incremento o decremento esta por toda su longitud), y los correspondientes valores para otros coeficientes fueran similarmente ordenados. Esto es fácilmente demostrado al introducir otro objeto a la tabla de objetos

		Variable									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Objetos	**3**	0	0	0	0	0	0	0	1	0	0

Los valores por los primeros dos coeficientes, para los tres pares de objetos son

Simple coeficiente de ser igual	***Coeficiente de Jaccard***
s_{12} = 0.70	s_{12} = 0.40
s_{13} = 0.50	s_{13} = 0.00
s_{23} = 0.80	s_{23} = 0.33

Los coeficientes no son conjuntamente monotómicos, así los datos categóricos donde las variables tienen más de dos niveles podrían ser repartidos con un camino similar al de datos binarios, con cada nivel de una variable siendo considerada como simple

variable binaria. Esta no es una aproximación atractiva, simplemente porque un número grande de marcas *negativas* (sin el atributo) inevitablemente serían confusas.

Un método que soluciona lo anterior es el asignar una marca s_{ijk} de cero o uno, a cada variable k, dependiendo si los dos objetos i y j son iguales en la variable. Las marcas para todas las variables son simplemente promediadas para dar el coeficiente de similaridad requerido.

$$s_{ij} = \frac{\sum_{k=1}^{m} s_{ijk}}{m}$$

1.2.3 MEDIDAS DE SIMILARIDAD PARA VARIABLES DE TIPO MIXTO

Como los objetos pueden ser tomados desde varios puntos de vista, es decir, pueden ser medidos en varios tipos de variables ya sean binarias, nominales, cuantitativas, etc., no se puede evitar enfrentarse al problema de tener varios tipos de variables, y el tener que buscar un coeficiente de proximidad para objetos que son medidos con distintas variables.

Un coeficiente de proximidad para estos casos, debe ser una fórmula que tome en cuenta las propiedades de cada uno de los diferentes tipos de variables, para poder así tener un coeficiente que mantenga todas las relaciones que haya entre las distintas mediciones de los dos objetos que se toman.

Un coeficiente de similaridad para el problema de los conjuntos de datos que contienen una variedad de tipos de variables como las mencionadas anteriormente, fue sugerido por Gower (1971), el cual es definido como

$$s_{ij} = \sum_{k=1}^{m} w_{ijk} s_{ijk} \Big/ \sum_{k=1}^{m} w_{ijk} \qquad (1.1)$$

En este coeficiente, s_{ijk} es la similaridad entre los individuos i-ésimo y j-ésimo medidos

por la k-ésima variable. y w_{ijk} es 1 o 0 dependiendo sobre "si" o "no" la comparación es considerada válida para la k-ésima variable. Las marcas de 0 son asignadas cuando la variable k es desconocida para uno o ambos objetos, o para variables binarias, donde éste es requerido para excluir marcas negativas (i.e., falta del atributo). Para datos categóricos los componentes de similaridades, s_{ijk} toman el valor 1 cuando los dos objetos tienen el mismo valor y 0 en otro caso. Para variables cuantitativas los componentes de similaridad son dados por

$$s_{ijk} = 1 - | x_{ik} - x_{jk} | / R_k$$

donde x_{ik} y x_{jk} son los valores de los dos objetos para la variable k, y R_k es el rango de la variable k, usualmente en el conjunto de objetos a ser conglomerados.

Para ilustrar el uso del coeficiente de Gower se toma como un ejemplo los siguientes datos para pacientes malos psiquiátricamente.

	Peso (libras)	Nivel de ansiedad	¿Presenta Depresión?	¿Presenta Alucinación?	Edad
Paciente 1	120	Pacífico	No	No	Joven
Paciente 2	150	Moderado	Si	No	Medio
Paciente 3	110	Severo	Si	Si	Viejo
Paciente 4	145	Pacífico	No	Si	Viejo
Paciente 5	120	Pacífico	No	Si	Joven

En este caso supongamos que en el estudio se desea excluir marcas negativas sobre la Depresión y Alucinaciones del cálculo de la similaridad entre pacientes. Los valores de las medidas de similaridad de Gower para cada par de pacientes es calculada de (1.1), por ejemplo para el par de pacientes 1 y 2

$$s_{12} = \frac{1 \times (1 - (30/40)) + 1 \times 0 + 1 \times 0 + 0 \times 1 + 1 \times 0}{1 + 1 + 1 + 0 + 1} = 0.0625$$

Los valores para cada par de pacientes son desplegados en la siguiente matriz de similaridades S.

$$S = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1.000 & & & & \\ 0.062 & 1.000 & & & \\ 0.150 & 0.200 & 1.000 & & \\ 0.344 & 0.175 & 0.425 & 1.000 & \\ 0.750 & 0.005 & 0.350 & 0.475 & 1.000 \end{pmatrix} \end{array}$$

En tratar de seleccionar una medida de distancia particular, el analista debería recordar los siguientes conceptos. En la mayoría de las situaciones, diferentes medidas de distancias conducen a diferentes soluciones de conglomerados. De este modo es conveniente usar distintas medidas y comparar los resultados a lo teórico o conocer modelos. También cuando las variables son medidas en diferentes unidades, uno debería estandarizar los datos antes de seguir con el análisis de conglomerados. La estandarización es particularmente recomendable cuando el rango de una variable es mucho más grande que las otras.

1.3 ALGORITMOS DE CONGLOMERACIÓN

Las medidas de proximidad para saber la distancia o la semejanza de un objeto con otro es el primer paso del análisis de conglomerados, y al tener concluido este paso dá el permiso para poder comenzar la siguiente etapa del método, que es el clasificar los objetos de acuerdo a sus proximidades. De esta manera los objetos más *homogéneos* entre ellos (que son los objetos con una similaridad más grande o una disimilaridad más pequeña) formaran grupos o conjuntos, los cuales son lo suficientemente *heterogéneos*[4], con cualquier otro grupo que sea formado.

A estos grupos o conjuntos de objetos se les conoce como ***conglomerados*** y su principal característica que los distingue de cualquier tipo de grupos o conjuntos de

[4] Heterogéneo: Término que indica que los objetos o cosas son diferentes entre si. *ENCICLOPEDIA SALVAT Diccionario*. Tomo 6. pág. 1674

elementos, es que son muy homogéneos entre si, pero muy heterogéneos entre ellos, es decir, todos los objetos que los forman son muy semejantes entre ellos pero muy diferentes con los objetos de los otros grupos, y asi cada conglomerado tiene una definición muy particular con los objetos que los forman.

Pero de ¿qué manera o proceso debe ser seguido para formar conglomerados y clasificar los objetos dentro de éstos?

Los algoritmos de conglomeración más comúnmente usados pueden ser clasificados en dos categorías generales: (1) jerárquicos, y (2) no jerárquicos.

1.3.1 ALGORITMOS DE CONGLOMERACIÓN JERÁRQUICOS

A estos tipos de procedimientos se les considera jerárquicos por el hecho de crear una estructura arbólica que puede ser tomada como una estructura con diferentes jerarquías. Hay básicamente dos tipos de procedimientos de conglomeración jerárquicos que son los aglomerativos y los divisivos.

En los métodos de conglomeración aglomerativos, cada objeto inicia siendo como su propio conglomerado. En subsecuentes pasos, los dos conglomerados más estrechos son combinados en un nuevo conglomerado agregado, y de este modo reduciéndose el número de conglomerados, hasta que todos los individuos están agrupados dentro de un conglomerado más grande y significante.

Los métodos de conglomeración divisivos son procedimientos que proceden en la dirección opuesta de los métodos aglomerativos, es decir, en vez de iniciar con N conglomerados (el número de objetos) se inicia con un sólo conglomerado formado por todos los objetos. En consecutivos pasos los objetos que son más diferentes son divididos y cambiados en conglomerados más pequeños. Este proceso continúa hasta que hay N (número de objetos) conglomerados de un sólo miembro.

Los métodos de conglomeración aglomerativos son mucho más populares que los métodos divisivos, es por eso que los métodos divisivos no son muy usados y conocidos.

Con tales métodos las divisiones o uniones una vez hechas son irrevocables, así que cuando un algoritmo aglomerativo ha juntado dos individuos, éstos no pueden ser subsecuentemente separados, y cuando un algoritmo divisivo ha hecho una partitura ésta no puede ser reunida. Kaufman y Rousseeuw (1990) comentaron que *un método aglomerativo sufre del defecto de que éste nunca puede reparar lo que fue hecho en pasos anteriores.*

Las clasificaciones jerárquicas pueden ser representadas por un diagrama de dimensión dos llamado ***dendograma***[5], el cual ilustra sobre un eje a todos los objetos y los conglomerados que se van formando de acuerdo a uniones o divisiones hechas en las estrategias sucesivas del análisis, y sobre el otro eje la distancia entre cada conglomeración combinada.

Los métodos de conglomeración jerárquicos más comúnmente usados para desarrollar conglomerados en un análisis son: (1) *Método de Ligaje Simple*, (2) *Método de Ligaje Completo*, (3) *Método de Ligaje en Promedio*, (4) *Conglomeración con Centroides*, (5) *Método de la Mediana*, y (6) *Método de Ward.*

1.3.1.1 MÉTODO DE LIGAJE SIMPLE

El método de Ligaje Simple es una de las técnicas más sencillas para desarrollar conglomerados, también es conocido como la técnica del *vecino más cercano*. El método fue descrito por Florek (1951) y después por Sneath (1957) y Johnson (1967). La principal característica del método es que la distancia entre los grupos es definida como el par más estrecho de objetos, donde los pares solamente consisten de un objeto de cada grupo que es considerado.

El procedimiento de ligaje simple está basado sobre la distancia mínima. Este encuentra los dos objetos que estén separados por la distancia más corta y los coloca en el primer conglomerado. Entonces la próxima distancia más corta es encontrada y puede ser un tercer objeto junto a los primeros dos para formar un nuevo conglomerado de tres elementos o un nuevo conglomerado de dos objetos es formado. Este proceso continúa

[5] Dendrogram. (n. a.)

hasta que todos los objetos forman un conglomerado. La distancia entre dos conglomerados es la distancia más corta de algún punto de un conglomerado a un punto en el segundo conglomerado. Dos conglomerados son fusionados por alguna estrategia en la simple liga más corta entre ellos. Este método tiene problemas, los cuales suelen ocurrir cuando los conglomerados son pobremente delineados. En tales casos, los procedimientos del ligaje simple forman largas cadenas serpentinas (figura 1.1), y eventualmente todos los objetos (ya sean objetos o conglomerados pequeños) son colocados en una cadena, y de ese modo los objetos que se encuentran en los extremos finales de la cadena pueden ser muy diferentes.

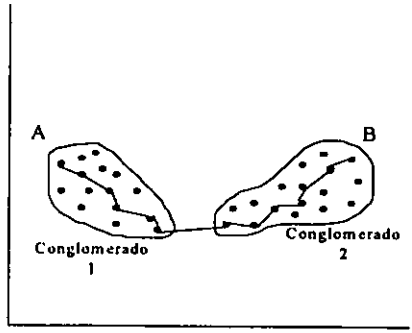


Figura 1.1. Puntos A y B no muy similares

Como un ejemplo de este método y para conocer sus operaciones de técnicas aglomerativas jerárquicas, el método será aplicado a la siguiente matriz de disimilaridades, la cual es calculada por la medida de disimilaridad más adecuada, dependiendo del tipo de variables con que son medidos los objetos, la matriz es de disimilaridades ya que la diagonal principal de ésta es 0, lo cual es una propiedad de las disimilaridades, donde la disimilaridad entre el mismo objeto es 0 ($d_{ii} = 0$).

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}$$

La entrada más pequeña en la matriz, es de los objetos 1 y 2, consecuentemente éstos son reunidos para formar un conglomerado de dos miembros. Las distancias entre este conglomerado y los otros tres objetos se obtuvieron como

$$\begin{aligned} d_{(12)3} &= \min [d_{13}, d_{23}] = d_{23} = 5.0 \\ d_{(12)4} &= \min [d_{14}, d_{24}] = d_{24} = 9.0 \\ d_{(12)5} &= \min [d_{15}, d_{25}] = d_{25} = 8.0 \end{aligned} \qquad (1.2)$$

Una nueva matriz es ahora construida donde las entradas son las distancias entre objetos y los valores del conglomerado-objeto.

$$\mathbf{D}_2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & \\ 5.0 & 0.0 & & \\ 9.0 & 4.0 & 0.0 & \\ 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}$$

La entrada más pequeña en $\mathbf{D}_2$ es de los objetos 4 y 5, así de esta manera éstos forman un segundo conglomerado de dos miembros, y un nuevo conjunto de distancias es encontrado

$$\begin{aligned} d_{(12)3} &= 5.0 \text{ como antes} \\ d_{(12)(45)} &= \min [d_{14}, d_{15}, d_{24}, d_{25}] = d_{25} = 8.0 \\ d_{(45)3} &= \min [d_{34}, d_{35}] = d_{34} = 4.0 \end{aligned} \qquad (1.3)$$

Y todas estas distancias pueden ser acomodadas para formar la matriz $\mathbf{D}_3$

$$\mathbf{D}_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{pmatrix} 0.0 & & \\ 5.0 & 0.0 & \\ 8.0 & 4.0 & 0.0 \end{pmatrix} \end{array}$$

La entrada más pequeña es ahora $d_{(45)3}$ y así el objeto 3 es añadido al conglomerado que contiene los objetos 4 y 5. Finalmente los conglomerados que contienen los objetos 1, 2 y 3, 4, 5 son combinados en un solo conglomerado. Las particiones que se hicieron en cada estrategia fueron las siguientes:

Estrategia	Conglomerados
P_5	[1], [2], [3], [4], [5]
P_4	[1 2], [3], [4], [5]
P_3	[1 2], [3], [4 5]
P_2	[1 2], [3 4 5]
P_1	[1 2 3 4 5]

El dendograma correspondiente es el que se muestra en la figura 1.2.

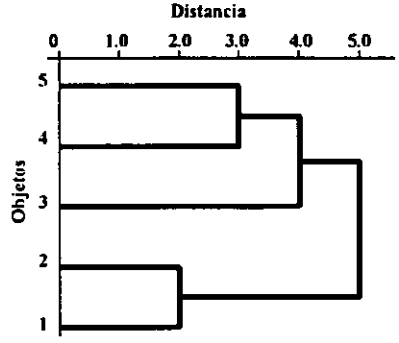


Figura 1.2. Dendograma del Ligaje Simple

Un punto importante acerca de la conglomeración hecha jerárquicamente, es que un conglomerado es obtenido cada vez mediante las combinaciones de conglomerados de previas estrategias.

1.3.1.2 MÉTODO DE LIGAJE COMPLETO

El método de conglomeración de *Ligaje Completo* o *Vecino más lejano* es la oposición

del método de ligaje simple, en el sentido de que la distancia entre los grupos (conglomerados) esta ahora definida como el par de individuos más distante, siendo uno de cada conglomerado. Esta medida en comparación con la del ligaje simple es mostrada en la figura 1.3. El procedimiento del ligaje completo es similar al ligaje simple excepto que el criterio de conglomerado esta basado en distancias máximas. Por esta razón también es conocido como el método del *vecino más lejano*. El método es llamado ligaje completo por que todos los objetos (individuos) en un conglomerado están ligados con cada otro por alguna distancia máxima o por la mínima similaridad. Se puede decir que la similaridad dentro del grupo es igual al diámetro del grupo, por el hecho de ser la distancia más grande que hay dentro del grupo, así la máxima distancia entre dos puntos es igual a la máxima distancia entre dos puntos del perímetro de un circulo, lo cual es el diámetro. Así esta técnica elimina el problema serpentino identificado con el método del ligaje simple.

El problema de medir la distancia entre grupos sin embargo, todavía se presenta. La figura 1.3 muestra como las distancias más cortas y más largas no pueden representar la verdadera similaridad entre grupos. El uso de la distancia más corta indica que los dos grupos son muy similares, mientras que el uso de la distancia más larga indica que son muy diferentes.

Usando este método sobre la matriz $\mathbf{D}_1$ de la sección anterior, la primera estrategia es de nuevo la fusión de los objetos 1 y 2. Las distancias entre este grupo y los tres objetos restantes son

$$\begin{aligned} d_{(12)3} &= \max [d_{13}, d_{23}] = d_{13} = 6.0 \\ d_{(12)4} &= \max [d_{14}, d_{24}] = d_{14} = 10.0 \\ d_{(12)5} &= \max [d_{15}, d_{25}] = d_{15} = 9.0 \end{aligned} \qquad (1.4)$$

Los resultados de las siguientes estrategias son

$$D_2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & \\ 6.0 & 0.0 & & \\ 10.0 & 4.0 & 0.0 & \\ 9.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}$$

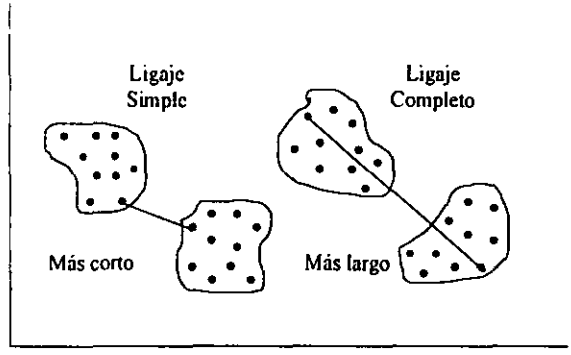


Figura 1.3. Comparación de las medidas de distancia del Ligaje Simple y el Ligaje Completo.

La entrada más pequeña es la de los objetos 4 y 5, y sus correspondientes distancias con los demás objetos son

$$\begin{aligned} d_{(12)3} &= 6.0 \text{ como antes} \\ d_{(45)(12)} &= \max [d_{14}, d_{15}, d_{24}, d_{25}] = d_{14} = 10.0 \\ d_{(45)3} &= \max [d_{34}, d_{35}] = d_{35} = 5.0 \end{aligned} \qquad (1.5)$$

Y se forma la matriz D_3

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{pmatrix} 0.0 & & \\ 6.0 & 0.0 & \\ 10.0 & 5.0 & 0.0 \end{pmatrix} \end{array}$$

Y se junta el objeto 3 al conglomerado que contiene los objetos 4 y 5, por ser la entrada más pequeña de la matriz. El dendograma que se obtiene de la aplicación del método de

ligaje completo a la matriz D_1 se muestra en la figura 1.4.

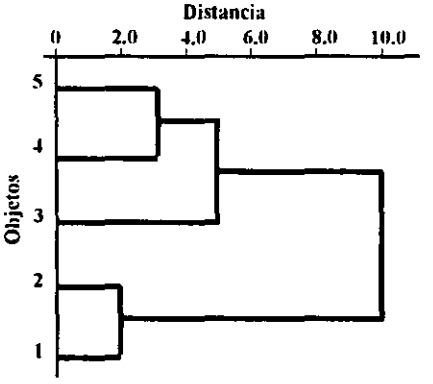


Figura 1.4. Dendograma del Ligaje Completo

1.3.1.3 MÉTODO DE LIGAJE EN PROMEDIO

El método de *Ligaje en Promedio* se inicia de la misma manera que el ligaje simple y el ligaje completo, pero el criterio de conglomerar es el promedio de las distancias de objetos de un conglomerado a los objetos de otro conglomerado.

Tal técnica no usa los valores extremos, como el ligaje simple o el ligaje completo, y su partición está basada sobre todos los miembros de los conglomerados y no sobre un simple par de miembros extremos.

La aproximación del ligaje en promedio tiende a combinar conglomerados con varianzas pequeñas; y estos conglomerados tienden a estar sesgados hacia la producción de conglomerados con la misma varianza aproximadamente.

Como se mencionó la distancia entre dos conglomerados está definida como el promedio de las distancias entre todos los pares de objetos, que son compuestos de un objeto de cada grupo. Tal medida es ilustrada en la figura 1.5.

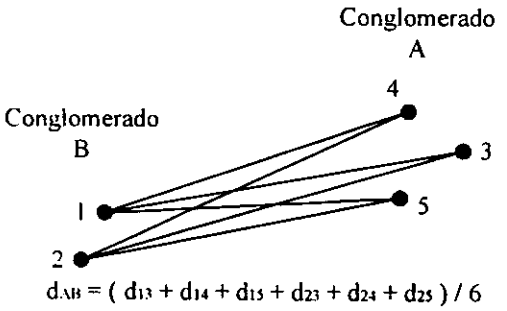


Figura 1.5. Distancia del Método de la Mediana

Aplicando el método a la misma matriz $\mathbf{D}_1$ que se utilizó en las secciones anteriores, la primera estrategia, como lo es con el ligaje simple y completo, es la formación de un conglomerado que contiene a los objetos 1 y 2 por tener la entrada más pequeña de la matriz. Así un nuevo conjunto de distancias es encontrado de

$$\begin{aligned} d_{(12)3} &= (d_{13} + d_{23})(\tfrac{1}{2}) = 5.5 \\ d_{(12)4} &= (d_{14} + d_{24})(\tfrac{1}{2}) = 9.5 \\ d_{(12)5} &= (d_{15} + d_{25})(\tfrac{1}{2}) = 8.5 \end{aligned} \qquad (1.6)$$

Y acomodando las distancias en una matriz $\mathbf{D}_2$ se obtiene

$$\mathbf{D}_2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & \\ 5.5 & 0.0 & & \\ 9.5 & 4.0 & 0.0 & \\ 8.5 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}$$

La entrada más pequeña es d_{45} y así un segundo conglomerado es formado por los objetos 4 y 5. La distancia promedio entre los dos conglomerados (formados por dos-miembros) está dada por

$$d_{(12)(45)} = (d_{14} + d_{15} + d_{24} + d_{25})(1/4) = 9.0$$
$$d_{(12)3} = (d_{13} + d_{23})(1/2) = 5.5 \qquad (1.7)$$
$$d_{(45)3} = (d_{43} + d_{53})(1/2) = 4.5$$

Y se forma la matriz D_3

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{pmatrix} 0.0 & & \\ 5.5 & 0.0 & \\ 9.0 & 4.5 & 0.0 \end{pmatrix} \end{array}$$

Y se junta el objeto 3 al conglomerado formado por los objetos 4 y 5, por ser la entrada más pequeña de la matriz. Y el procedimiento continúa como se ha descrito en las secciones anteriores.

1.3.1.4 CONGLOMERACIÓN CON CENTROIDES

Los tres métodos descritos hasta ahora, operan directamente con los datos de la matriz de proximidades, y no les es necesario manejar los valores originales de las variables de los objetos (individuos). Un método el cual requiere los datos originales es el de *Conglomeración con Centroides*.

En el método de conglomeración con centroides la distancia entre dos conglomerados es la distancia (típicamente Euclidiana) entre sus ***centroides***[6]. En este método, cada vez que unos objetos son agrupados y forman un conglomerado o son agrupados a un conglomerado ya creado, un nuevo centroide es calculado del conglomerado que se haya formado. Los centroides de conglomerados toman lugar como conglomerados fusionados. En otras palabras, hay un cambio en el centroide de un conglomerado cada vez que un nuevo objeto o grupo de objetos es añadido a un conglomerado existente.

[6] Centroide: Es el centro de gravedad del conjunto de variables con que se tienen los valores de todos los objetos que forman la población. Este vector es una medida que representa el centro de todas las mediciones que se tienen de la población que representa. Mood, Alexander *M., INTRODUCTION TO THE THEORY OF STATISTICS*. 1983, pág. 58, 65.

Este método frecuentemente provoca confusiones, esto ocurre por reversiones, cuando la distancia entre los centroides de un par puede ser menor que la distancia entre los centroides de otro par fusionado en un tiempo anterior. La ventaja de este método es que se afecta menos por las salidas (resultados de estrategias) como en otros métodos jerárquicos.

Los métodos de centroides son métodos que requieren de datos métricos, los cuales pueden severamente limitar su aplicación a ciencias sociales. Y otros métodos de ligaje no tienen este requerimiento.

Con este método, los grupos una vez formados son representados por sus valores medios para cada variable, que es su *vector de medias*, y la distancia entre grupos esta ahora definida en términos de la distancia entre dos vectores de medias. El uso de una media implica estrictamente, que las variables están sobre un *intervalo escala*[7].

Para ilustrar cómo opera la conglomeración con centroides, será aplicado al siguiente conjunto de datos bivariados.

Objetos	Variable 1	Variable 2
1	1.0	1.0
2	1.0	2.0
3	6.0	3.0
4	8.0	2.0
5	8.0	0.0

Mediante la distancia Euclidiana se obtuvo la siguiente matriz de distancias

[7] Intervalo escala: Por ejemplo, la altitud de una montaña, la cual es determinada solamente de una posición estándar (nivel del mar) y en términos de una unidad estándar (ft, pies)
Hartigan, John A., *CLUSTERING ALGORITHMS*, 1975, pág. 9

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.00 & & & & \\ 1.00 & 0.00 & & & \\ 5.39 & 5.10 & 0.00 & & \\ 7.07 & 7.00 & 2.24 & 0.00 & \\ 7.07 & 7.28 & 3.61 & 2.00 & 0.00 \end{pmatrix} \end{array}$$

La matriz $\mathbf{D}_1$ nos muestra que d_{12} es la entrada más pequeña y los objetos 1 y 2 son fusionados para formar un grupo. El vector media del grupo es calculado, (1.0, 1.5) y se obtendrá una nueva matriz de los datos siguientes

Objetos	Variable 1	Variable 2
(12)	1.0	1.5
3	6.0	3.0
4	8.0	2.0
5	8.0	0.0

$$\mathbf{D}_2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.00 & & & \\ 5.22 & 0.00 & & \\ 7.02 & 2.24 & 0.00 & \\ 7.16 & 3.61 & 2.00 & 0.00 \end{pmatrix} \end{array}$$

La entrada más pequeña en $\mathbf{D}_2$ es d_{45} y los objetos 4 y 5 por lo tanto son fusionados para formar un segundo grupo, y siendo calculado el vector media de este grupo dá (8.0, 1.0), y se obtiene ahora la matriz $\mathbf{D}_3$ con los datos siguientes

Objeto	Variable 1	Variable 2
(1 2)	1.0	1.5
3	6.0	3.0
(4 5)	8.0	1.0

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{pmatrix} 0.00 & & \\ 5.22 & 0.00 & \\ 7.02 & 2.83 & 0.00 \end{pmatrix} \end{array}$$

En la matriz D_3 la entrada más pequeña es $d_{(45)3}$ y así los objetos 3, 4 y 5 son fusionados en un conglomerado de tres miembros. La estrategia final consiste de fusionar los dos grupos restantes en un solo grupo. El dendograma resultante se muestra en la figura 1.6.

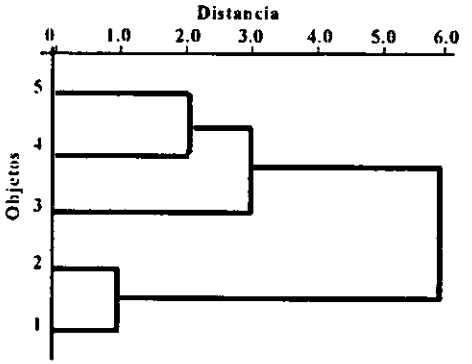


Figura 1.6. Dendograma de Conglomeración con Centroides

1.3.1.5 MÉTODO DE LA MEDIANA

Una desventaja del método de centroides es que si los tamaños de los dos grupos que son fusionados son muy diferentes, entonces el centroide del nuevo grupo será muy estrecho (cercano) al grupo más grande y puede permanecer dentro del grupo y de ésta manera las propiedades de las características de el grupo más pequeño virtualmente se pierden. La estrategia puede ser hecha independiente del tamaño del grupo, asumiendo que los grupos a fusionar son de igual tamaño, la posición aparente del nuevo grupo siempre estará entre los dos grupos a ser fusionados. Además si los centroides de los grupos a ser fusionados son representados por (*i*) y (*j*), entonces la distancia del centroide de un tercer grupo (*h*) del grupo formado por la fusión de (*i*) y (*j*) está situada a lo largo de la media del triángulo definido por (*i*), (*j*) y (*h*). Es por esta razón

que Gower (1967), fue quien sugirió primero la estrategia, y le propuso el nombre de la *Mediana*.

Aunque este método podria ser hecho conveniente para medidas de similaridad y distancia, Lance y Williams (1967) sugirieron que el método deberia ser considerado inapropiado para medidas tales como coeficientes de correlación, ya que la interpretación en un sentido geométrico no es posible.

1.3.1.6 MÉTODO DE WARD

Ward (1963) propuso un procedimiento de conglomeración buscando el formar particiones P_N , P_{N-1} , . . . , P_1 en una manera que minimice las pérdidas de información asociadas con cada grupo, y el cuantificar pérdidas en una forma que es fácilmente interpretable. En cada paso del análisis, la unión de cada posible par de conglomerados es considerada y los dos conglomerados de quien la fusión resulta con el mínimo incremento en *información perdida*, son combinados. La información perdida es definida por Ward en términos de un criterio de suma de errores al cuadrado, ESS[8].

En este método la distancia entre dos conglomerados es la suma de ESS entre los dos conglomerados sumados sobre todas las variables, que es la suma de errores al cuadrado de cada variable con que son medidos los objetos que componen a los conglomerados que se combinan. En cada etapa del procedimiento de conglomeración, la suma de errores al cuadrado (ESS) en un conglomerado es minimizada sobre todas las particiones (el conjunto completo de divisiones o conglomerados separados) que se pueden obtener al combinar dos conglomerados en la etapa previa. El procedimiento tiende a combinar conglomerados con un pequeño número de observaciones, ocasionando que se obtengan conglomerados con varianzas iguales y aproximadamente con el mismo número de observaciones.

El criterio del método de Ward puede ser ilustrado simplemente con considerar los siguientes datos univariados, supongamos por ejemplo, 10 individuos (objetos) que

[8] Error Sum of Squares. (n. a.)

tienen los valores de 2, 6, 5, 6, 2, 2, 2, 0, 0, 0, sobre alguna variable particular. La pérdida de información que resultaría de tratar los diez valores como un grupo con una media de 2.5 es representada por ESS dado por,

$$ESS = \sum_{i=1}^{N} (x_i - \bar{x})^2 \qquad (1.8)$$

Para este ejemplo

$$ESS_{\text{un grupo}} = (2 - 2.5)^2 + (6 - 2.5)^2 + \ldots + (0 - 2.5)^2 = 50.5 \qquad (1.9)$$

Similarmente si los diez individuos son clasificados acotando sus valores dentro de cuatro conjuntos,

$$\{0, 0, 0\}, \{2, 2, 2, 2\}, \{5\}, \{6, 6\}$$

la ESS puede ser evaluada como la suma de cuatro separadas sumas de errores al cuadrado

$$ESS_{\text{cuatro grupos}} = ESS_{\text{grupo 1}} + ESS_{\text{grupo 2}} + ESS_{\text{grupo 3}} + ESS_{\text{grupo 4}} = 0.0 \qquad (1.10)$$

Finalmente el método de Ward combina los conglomerados que resulten con el mínimo incremento en las sumas de los errores al cuadrado dentro de las distancias de los conglomerados.

El método de conglomeración jerárquico de Ward está basado sobre las sumas ESS dentro de los grupos en vez de ligar grupos y así un algoritmo aglomerativo es usado. Para cada etapa el número de grupos es reducido por uno, al combinar los dos grupos con el más pequeño incremento posible en el total de la suma de errores al cuadrado dentro de los grupos; de hecho, cuando se inicia con N grupos de un individuo, el total de la suma de cuadrados dentro de los grupos es cero (es decir, no hay pérdida de información).

1.3.2 ALGORITMOS DE CONGLOMERACIÓN NO JERÁRQUICOS

En contraste a los métodos jerárquicos, los procedimientos de conglomeración no jerárquicos no involucran el proceso de construcción arbólico. En vez de eso, el primer paso es el seleccionar un conglomerado centro o semilla, y todos los objetos (individuos) dentro de una distancia comienzo preespecificada son incluidos en el conglomerado resultante.

Las técnicas de conglomeración no jerárquicas son más conocidas como los métodos de partición y típicamente usan una de las siguientes tres aproximaciones. El **Procedimiento de Comienzo Secuencial**, el cual inicia por seleccionar un conglomerado semilla, y todos los objetos dentro de una distancia preespecificada son incluidos. Cuando todos los objetos están incluidos dentro de la distancia, un segundo conglomerado semilla es seleccionado, y todos los objetos dentro de la distancia preespecificada son incluidos. Entonces una tercera semilla es seleccionada, y el proceso continúa como se mencionó. Cuando un objeto es conglomerado con una semilla, ya no es considerado por subsecuentes semillas. En contraste, el **Procedimiento de Comienzo Paralelo**, selecciona varias semillas simultáneamente en el principio, y los objetos dentro de la distancia *comienzo* son asignados a la semilla más cercana. Como el proceso involucra distancias iniciales pueden ser ajustados a incluir menos o más objetos en los conglomerados. También en algunos métodos, los objetos pueden quedar fuera de los conglomerados si ellos están afuera de la distancia inicial preespecificada de algún conglomerado semilla. El tercer procedimiento, es el de **Optimización**, es similar a los otros dos, excepto que éste permite el reasignamiento de objetos a otro conglomerado del original sobre las bases de algún criterio global de optimización.

El principal problema visto en todos los procedimientos de conglomeración no jerárquicos, es el cómo seleccionar el conglomerado semilla o semillas. Por ejemplo, con un *Comienzo Paralelo*, el conglomerado inicial y probablemente el final resulta dependiendo de los ordenes de las observaciones en el conjunto de datos, y al arrastrar el orden es posible el afectar los resultados. Específicamente los conglomerados semillas iniciales con el procedimiento de *comienzo paralelo* pueden reducir este

problema. Pero al seleccionar el proceso para semillas (es decir, diferentes puntos semillas aleatorios) pueden todavia afectar los resultados y de este modo permanece el problema.

1.3.2.1 MÉTODOS DIVISIVOS

Las técnicas de conglomeración divisivas son esencialmente de dos tipos, *monotéticas* las cuales dividen los datos sobre las bases de la posesión o de otra manera de un simple atributo especificado, y *politéticas* donde las divisiones están basadas sobre los valores tomados por todos los atributos.

El más factible de los métodos politéticos divisivos es el descrito por McNaughton-Smith (1964). En este método un grupo de *sobras* es acumulado por una adición secuencial del objeto de quien la total disimilaridad con el resto menos su disimilaridad total con el grupo de *sobras*, es un máximo.

Cuando esta diferencia llega a ser negativa el proceso es repetido sobre los dos sub-grupos. La medida usual de disimilaridad a usarse es el promedio de la distancia Euclidiana entre cada objeto y los otros objetos en el grupo.

1.3.2.2 MÉTODOS DE PARTICIÓN

La mayoría de los métodos de partición comienzan con el procedimiento de *Comienzo Paralelo*, es decir, inician tomando k semillas que serán los k conglomerados finales.

Un método de partición construye k conglomerados, en el que se clasifican los datos, los k conglomerados resultantes juntos satisfacen los requerimientos de una partición:

- Cada grupo debe contener por lo menos un objeto.
- Cada objeto debe pertenecer exactamente a un grupo.

Estas condiciones implican que puede haber tantos conglomerados como el número de objetos,

$$k \leq N$$

La segunda condición indica que dos conglomerados diferentes no pueden tener algún objeto en común, y que los k conglomerados componen a todo el grupo de datos.

El valor de k es un número dado por el analista, y así el algoritmo construirá una partición con la cantidad de conglomerados que se desee.

Es conveniente correr el algoritmo varias veces con diferentes valores de k, ya que hay ocasiones en que el valor de k no lleva a una conglomeración *natural*. Y la mejor forma de seleccionar el valor correcto de k es de acuerdo a ciertas características o señales de las gráficas, o retener la conglomeración que dé la interpretación más significante de acuerdo a algún criterio numérico. La idea de los métodos de partición es el obtener un número fijo de k conglomerados.

En el proceso de conglomeración para obtener k conglomerados, se seleccionan k objetos (los cuales son llamados *objetos representativos*) en el conjunto de datos.

Los conglomerados correspondientes son entonces obtenidos al asignar cada objeto al *objeto representativo* más cercano. Los *objetos representativos* deben ser seleccionados de manera que estén centralmente localizados en los conglomerados que definen. Para ser exactos, la distancia promedio (o disimilaridad promedio) del *objeto representativo* a todos los otros objetos del mismo conglomerado es minimizada.

A tal *objeto representativo* óptimo se le conoce como **centro** de su conglomerado y el método de partición alrededor de centros es conocido como el método de ***k*-centros**.

Hay otro método con una gran semejanza con el método anterior, el cuál minimiza el promedio de las distancias cuadradas, dando así los llamados centroides, y el método es llamado el método de ***k*-medias**.

El procedimiento de conglomeración del método de *k*-medias comienza por usar los

valores de los primeros k objetos del conjunto de datos, como estimaciones temporales de los puntos medios de los k conglomerados, donde k es el número de conglomerados especificado por el analista. Los puntos medios iniciales de los conglomerados están formados al asignar a cada objeto en turno al conglomerado con el punto medio más cercano a éste, y entonces se vuelve a calcular el centroide. Y así mediante un proceso iterativo se encuentran los puntos medios de los conglomerados finales. Este proceso continúa hasta que no ocurran cambios más escandalosos en los centroides o hasta que un número máximo de iteraciones es alcanzado.

El método de *k-centros* es de mayor aplicación, debido a que es más robusto con respecto a los *outliers* (ver sección 2.10), y por que este método no solamente trata con coeficientes de disimilaridad sino también con mediciones de *intervalo escala*. Pero el procedimiento del análisis de conglomerados de *k-medias* es útil cuando se tiene un número grande de objetos.

1.4 MÉTODOS DE OPTIMIZACIÓN

Estas técnicas de conglomeración producen una partición de los objetos para un número particular de grupos, para minimizar o maximizar algún criterio numérico. En estos métodos se asume que el número de grupos ha sido fijado por el investigador.

La idea básica de estos métodos es asociar un índice *f(N,k)* con cada partición de los N objetos dentro del número requerido de k grupos, su valor es indicador de la *calidad* de esta particular conglomeración.

El criterio de conglomeración más comúnmente usado surge de considerar las siguientes tres matrices, las cuales son calculadas de una partición de los datos.

$$\mathbf{T} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})' \quad , \quad \mathbf{W} = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)' \qquad (1.11)$$

$$\mathbf{B} = \sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})'$$

Estas matrices son del orden de *mxm* (donde *m* es el número de variables), y la matriz T representa la total dispersión, W la dispersión dentro de grupos, y B la dispersión entre grupos, y satisfacen la siguiente ecuación:

$$T = W + B \qquad (1.12)$$

1.4.1 CRITERIO DE MINIMIZACIÓN DE LA TRAZA DE W

Si los datos que se tienen son multivariados, un buen criterio de conglomeración es el criterio de minimizar la suma de cuadrados dentro de los grupos. De esta manera el criterio toma en cuenta todas las variables para llevar a minimizar la traza(W)[9]. Como se menciona en la otra sección W es la matriz de dispersión al interior de los grupos, y al minimizar su traza se obtiene una de las ideas básicas de la conglomeración, que es el de obtener grupos que sean muy homogéneos en el interior de ellos, pero al ser minimizada la traza(W) se maximiza la traza(B) y así se llega a la otra idea básica de la conglomeración, que los grupos sean más heterogéneos entre ellos, formando estas dos ideas lo fundamental de la conglomeración, es decir, obtener grupos o conglomerados muy homogéneos en su interior pero muy heterogéneos entre ellos.

1.4.2 CRITERIO DE MINIMIZACIÓN DEL DETERMINANTE DE W

En el análisis multivariado de varianza, una de las formas para saber si hay diferencias entre vectores de medias se basa en la razón de los determinantes de las matrices de dispersión dentro (W) y el total (T) de los grupos. En el análisis de conglomerados, los vectores de medias son los vectores compuestos por las medias de cada una de las *m* variables con que son formados los conglomerados. Valores grandes del cociente

$$\det(T) / \det(W) \qquad (1.13)$$

indican que los vectores de medias de conglomerados se diferencian. Estas

[9] Traza(A) es la suma de los elementos de la diagonal principal de la matriz A.
Chatfield. C., Collins. A. J., *INTRODUCTION TO MULTIVARIATE ANALYSIS*, pág. 11

consideraciones dichas por Friedman y Rubin (1967) fueron tomadas como un criterio de conglomeración, que es el minimizar el det(**W**), y de esta manera maximizar (1.13).

1.4.3 CRITERIO DE MAXIMIZACIÓN DE LA TRAZA DE ($\mathbf{BW}^{-1}$)

Otro criterio que sugirieron Friedman y Rubin (1967) es el de maximizar la traza de la matriz que se obtiene del producto de la inversa de la matriz de dispersión dentro de los grupos ($\mathbf{W}^{-1}$) y la matriz de dispersión entre los grupos (**B**).

$$\text{traza}(\mathbf{BW}^{-1}) \tag{1.14}$$

Los criterios de la traza($\mathbf{BW}^{-1}$) y det(**T**) / det(**W**) pueden ser expresados en términos de los eigenvalores, λ_i de la matriz $\mathbf{BW}^{-1}$,de la siguiente forma,

$$\text{traza}(\mathbf{BW}^{-1}) = \sum_{i=1}^{m} \lambda_i \tag{1.15}$$

$$[\det(\mathbf{T}) / \det(\mathbf{W})] = \prod_{i=1}^{m} (1 + \lambda_i) \tag{1.16}$$

Este criterio de conglomeración es muy conveniente para datos con variables continuas.

1.5 PARTICIONES DE UNA JERARQUÍA

Frecuentemente en algunos estudios no se está interesado en la jerarquía completa dentro del análisis, sino en una o dos particiones obtenidas de ésta misma. En la conglomeración jerárquica, las particiones son logradas por *cortar* un dendograma o una selección de la solución de la secuencia de conglomerados que comprenden las jerarquías. En aplicaciones particulares se tiene el interés de saber y determinar cuál de todas las posibles particiones produce el mejor ajuste de los datos; esencialmente ésto significa conocer el apropiado número de conglomerados para los datos. Un método informal el cual es frecuentemente usado para este propósito es el de examinar las

diferencias entre niveles de fusiones en el dendograma. Grandes cambios o saltos en el dendograma, son tomados para indicar un número particular de conglomerados. Como ejemplo, el dendograma mostrado en la figura 1.6, muestra una grande diferencia en el nivel entre dos grupos y la estrategia final en el cual todos los objetos están en un sólo grupo (conglomerado). Esto indica tomar como evidencia el considerar la solución de dos conglomerados.

Duda y Hart (1973), propusieron una razón criterio, E(2) / E(1), donde E(2) es la suma de errores al cuadrado (ver sección 1.3.1.6) en los conglomerados cuando los datos son particionados en dos conglomerados, y E(1) da los errores al cuadrado cuando solamente un conglomerado se presenta. La hipótesis de un simple conglomerado es rechazada si la razón es más pequeña que un valor crítico especificado.

Calinski y Harabasz (1974) también sugirieron un indicador para el número de grupos basado en términos de la suma de cuadrados, que son la *traza* de la matrices que representan en forma resumida los conglomerados, como se muestra

$$\frac{\text{traza}(\mathbf{B}) / (g - 1)}{\text{traza}(\mathbf{W}) / (N - g)} \qquad (1.17)$$

donde **B** y **W** son el producto cruzado de matrices de la suma de cuadrados entre y dentro de conglomerados, y g es el número de grupos. El máximo valor del indicador en la jerarquía es tomado para indicar el correcto número de grupos.

1.6 ESTIMACIÓN DEL NÚMERO DE GRUPOS

En la mayoría de las aplicaciones de los métodos de optimización del análisis de conglomerados, se debe estimar el número de conglomerados que es más adecuado en los datos. La mayoría de los métodos de estimación son relativamente informales e involucran esencialmente trazar el valor del criterio de conglomeración contra el número de grupos.

Beale (1969) sugirió un método formal, el cual es por medio de una *prueba F*, y se usa

para probar si una subdivisión g_2 dentro de los conglomerados es significativamente mejor que una subdivisión g_1 dentro de algún número más pequeño de conglomerados.

El estadístico de prueba está definido como sigue:

$$F(g_1, g_2) = \frac{R_{g1} - R_{g2}}{R_{g2}} \Bigg/ \left[\left\{ \frac{N - g_1}{N - g_2} \right\} \left(\frac{g_2}{g_1} \right)^{2/p} - 1 \right] \qquad (1.18)$$

donde

$R_{g2} = (n\text{-}g) S_g^2$

S_g^2 es la varianza de los conglomerados centros en la muestra

$\nu_1 = p(g_2 - g_1)$

$\nu_2 = p(N - g_2)$

Un resultado significante indica que una subdivisión de g_2 dentro de los conglomerados es mejor que una subdivisión g_1 en un número más pequeño. Este procedimiento es recomendado sólo si los conglomerados están completamente separados.

Calinski y Harabasz (1974) sugirieron un método en el que se toma el valor de la g que dé el máximo valor de C, donde C es:

$$C = [\text{ traza}(\mathbf{B}) / (g - 1)\ ] / [\text{ traza}(\mathbf{W}) / (N - g)\ ] \qquad (1.19)$$

1.7 COMPARACIÓN DE RESULTADOS

Frecuentemente cuando se lleva a cabo un análisis de conglomerados de un conjunto de datos multivariados, puede ser necesario que se comparen dos o más conglomeraciones del mismo conjunto de objetos. Las soluciones a ser comparadas pueden ser hechas de diferentes métodos de conglomeración sobre el mismo conjunto de datos, o del mismo metodo de conglomeración aplicado a diferentes matrices de similaridad o distancia de

los datos crudos (sin estandarizar). Otra forma de obtener diferentes soluciones es por medio de aplicar el mismo procedimiento de conglomeración a la misma matriz de proximidad derivada de grupos de datos de diferentes fuentes. También se pueden hacer comparaciones informalmente, ésto puede ser simplemente con inspeccionar las conglomeraciones para determinar los conglomerados más importantes y/o examinar los dendogramas y señalar donde los conglomerados son más parecidos y donde son más diferentes.

CAPÍTULO 2

ANÁLISIS DISCRIMINANTE

I.- UNA PEQUEÑA HISTORIA

Los primeros escritos sobre el Análisis Discriminante en las primeras cuatro décadas se enfocaron en la predicción de miembros de grupos (***Análisis Discriminante Predictivo***[1] , ADP). R. A. Fisher (1930's), fue quien consideró a la discriminación de grupos de variables, como la idea de la distancia entre grupos multivariados de una combinación lineal de variables (Funciones Lineales Discriminantes, LDF's[2]), con el propósito de distinguir a los objetos que más se apegan al conjunto de variables; no fue hasta la década de los 1960's en que las LDF's se consideraron seriamente para propósitos de interpretar efectos reveladores de un análisis multivariado de varianza (MANOVA). Este aspecto del análisis discriminante es conocido como ***Análisis Discriminante Descriptivo***[3] (ADD).

Las ideas de la distancia y la combinación de variables fueron presentadas a priori en un documento de Fisher en 1936 ("*The use of multiple measurements in taxonomic problems*", el cual apareció en *Annals of Eugenics*). Muchas ampliaciones y refinamientos de las ideas de Fisher aparecieron en los 1940´s desarrollados por metodologistas de la academia de la Universidad de Harvard.

Este interés ocasionó que se escribieran textos que abarcaban el análisis discriminante en varias perspectivas. Algunos de esos libros escritos antes de 1970 con una perspectiva aplicada fueron escritos por: Rao (1952), Tatsuoka-Tiedeman (1954-1969), Kendall (1957), Cooley and Lohnes (1962), y Rulon et al. (1967).

Cuando grupos de objetos son conocidos y el propósito del investigador es el describir diferencias de grupos o el predecir los miembros de grupos sobre las bases de las mediciones de variables respuestas, en general las técnicas del análisis discriminante son las apropiadas.

II.- ANÁLISIS DISCRIMINANTE PREDICTIVO

[1] Predictive Discriminant Analysis. (n. a.)
[2] Linear Discriminant Functions. (n. a.)
[3] Descriptive Discriminant Analysis. (n. a.)

En la mayoría de las actividades de hoy en día, se tiene el interés de realizar la predicción de los posibles miembros de las subpoblaciones que conforman los fenómenos socioeconómicos administrativos de la actualidad. Por ejemplo, en la mayoría de los sectores académicos se tiene el interés de realizar la predicción de los alumnos de nuevo ingreso en el instituto, en predecir el dominio de la inteligencia, en identificar a aquellos estudiantes que llegarán a la cima de los estudios en la academia, o en identificar a aquellos contribuyentes que llevan en forma correcta sus impuestos. En cada uno de los objetos (alumnos de nuevo ingreso) hay una o más variables predictoras (explicatorias) con una variable criterio.

En algunas instancias la característica principal que distingue a cada uno de los objetos (variable criterio) es medida en escalas ordinales, pero en otras instancias el criterio es categórico y medido en una escala nominal. Este tipo de situaciones con un criterio categórico, es en el que el ***Análisis Discriminante Predictivo*** (ADP) es aplicable.

En el análisis discriminante predictivo la regla de predicción es desarrollada con una combinación lineal de las variables predictoras, sin embargo, la regla consiste de tantas combinaciones lineales como el número de categorías (grupos) que se consideran en el estudio. Tal regla hace capaz al investigador de predecir los miembros (objetos) de las categorías, y determinar el grupo de variables con el cual las categorías son mejor identificadas.

III.- ANÁLISIS DISCRIMINANTE DESCRIPTIVO

Como se notó en el *análisis discriminante predictivo*, las variables de respuestas múltiples juegan un papel de variables predictoras, en cambio, en el *Análisis Discriminante Descriptivo* (ADD) son vistas como simples resultados de mediciones sin tener un papel significante, es decir, variables resultados que cuando son agrupadas toman un especial papel en el método; y cada agrupación de éstas son las variables explicatorias. Por lo tanto, en el ADD la pregunta básica de interés pertenece a la agrupación de los efectos de variables explicatorias en base a las variables de respuesta múltiples, o más específicamente a la diferencia de grupos con respecto a las variables

de respuesta múltiple.

PREDICCIÓN

2.1 ANALOGÍA DE LA REGRESIÓN Y EL ANÁLISIS DISCRIMINANTE
(Ideas básicas de clasificación)

Una solución comúnmente usada para hacer predicciones empíricas o estadísticas es la *Regresión Múltiple*. Las técnicas de regresión múltiple son apropiadas en una situación en la que se tiene de un lado un conjunto de m variables predictoras (aleatorias o fijas) $X_1, X_2, \ldots, X_m$, y del otro lado una simple variable criterio (aleatoria) Y. Hay que tomar en cuenta que en estas técnicas se trata con un solo grupo de N objetos (unidades) experimentales, para cada uno de los cuales se tienen m+1 *medidas*[4]. Una meta del análisis de regresión múltiple es el crear una regla basada en una matriz de datos del orden Nx(m+1), para usarse en predecir (o estimar) una medida de variable criterio que dará valores a los predictores. Esto determina un conjunto de cargas de regresión $b_1, b_2, \ldots, b_m$ correspondientes a un conjunto dado de medidas de m variables predictoras, para obtener el valor de un compuesto lineal (combinación lineal), el cual es un valor predictivo para el criterio. La medida del criterio predictivo para el objeto u (Y_u) puede ser representada como:

$$Y_u = b_0 + b_1 X_{1,u} + b_2 X_{2,u} + \ldots + b_m X_{m,u}$$

donde b_0 es la constante de regresión. El compuesto puede ser también expresado como

$$Y_u = b_0 + b' X_u$$

donde b' es el vector renglón (1xm) de cargas de regresión y X_u es el vector columna (mx1) de medidas de variables predictoras para el objeto u.

[4] Medida: Expresión de una cantidad o dimensión con relación a una unidad determinada. *ENCICLOPEDIA SALVAT Diccionario*, Tomo 8, pág. 2178.

En cambio el ADD involucra la combinación lineal de las dos (o más) variables independientes que se distinguen mejor entre los grupos definidos con anterioridad.

La combinación lineal para un ADD son derivados de una ecuación que toma la siguiente forma:

$$Z = W_1 X_1 + W_2 X_2 + W_3 X_3 + \cdots + W_n X_n$$

donde

Z = La marca discriminante

W = Peso discriminante

X = Variable independiente

Otra aproximación usada en hacer predicciones empíricas, que involucra el aspecto del análisis discriminante del maximizar la varianza relativa entre los grupos para la varianza dentro de un grupo, pero sin utilizar una combinación lineal, es el tipo de análisis conocido como ***Análisis Discriminante Predictivo*** (ADP).

Las técnicas del ADP son apropiadas en un conjunto de grupos múltiples en los cuales se tienen m medidas para cada objeto que pertenece a uno de los k grupos. Los k grupos de n_g (g=1,...,k) objetos representan k poblaciones significantes. Y la variable criterio es una agrupación de variables dicótomas o multicótomas. Unas de las metas del ADP es el crear una regla basada en k matrices de datos de orden n_g x m, que predeciran la población a la que pertenece un objeto en la que sus miembros son desconocidos.

Una regla de clasificación en el ADP puede tomar 3 diferentes formas:

1) Un compuesto de las variables predictoras.
2) Una probabilidad de los miembros de la población.
3) Una distancia entre dos puntos.

2.2 LA NOCIÓN DE LA DISTANCIA

En un espacio bivariado (X_1 , X_2) se representa la distancia d_{AB} entre dos puntos como A: $(X_{1,A} , X_{2,A})$ y B: $(X_{1,B} , X_{2,B})$ por el teorema de Pitágoras (usualmente conocido como la distancia euclidiana) como se muestra

$$d^2_{AB} = (X_{1,A} - X_{1,B})^2 + (X_{2,A} - X_{2,B})^2$$

Note que d^2_{AB} puede ser expresado como

$$[X_A - X_B]' [X_A - X_B]$$

donde X_A y X_B son vectores columnas del orden (2x1) de cantidades y $X_A - X_B$ es un vector columna (2x1) de diferencias. Esto es apropiado sólo si se cumplen dos condiciones:

1) Las medidas de X_A y X_B están incorrelacionadas (es decir, $\rho_{AB} = .00$[5]); y
2) Las medidas de X_A y X_B tienen varianza de unidad.

Extendiendo la idea de la distancia euclidiana a un espacio general m-variado, se puede escribir como

$$d^2_{AB} = [X_A - X_B]' [X_A - X_B]$$

donde X_A y X_B son vectores (m x 1). La última expresión de d^2 es un vector renglón de m diferencias multiplicadas por un vector columna de las mismas m diferencias. En el mismo caso que el bivariado, ésto se basa en las suposiciones de que se tienen variables incorrelacionadas y de varianza de unidad, así la matriz de covarianza Σ de orden (m x m), para las m variables, es una matriz identidad.

---

[5] Se dice que se tiene un grado de interdependencia alto entre dos conjuntos de números si el valor de su coeficiente de correlación ρ es cercano a 1 o -1. Y es bajo o incorrelacionado si es cero.
Walpole. Ronald E., *PROBABILIDAD Y ESTADISTICA*, pág. 410.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1m}\sigma_1\sigma_m \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2m}\sigma_2\sigma_m \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}\sigma_m\sigma_1 & \rho_{m2}\sigma_m\sigma_2 & \cdots & \sigma_m^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Se toma como un caso especial cuando las varianzas de las medidas (variables) de los objetos son iguales. Si éste no es el caso, las varianzas distintas deben ser tomadas en consideración.

Un indicador de la distancia (cuadrada) entre dos puntos (*A* y *B*) que tome en consideración el caso en que las varianzas de las variables son distintas de uno y sus intercorrelaciones diferentes de cero, es

$$\Delta^2_{AB} = [X_A - X_B]' \Sigma^{-1} [X_A - X_B] \qquad (2.1)$$

en el que los puntos *A* y *B* son definidos por los vectores columnas X_A y X_B respectivamente. En la expresión (2.1) Σ es la matriz de covarianza de la población, y Δ^2_{AB} es una distancia generalizada que es atribuida al investigador estadístico ***P. C. Mahalanobis*** (1893 - 1972).

2.3 DISTANCIA Y CLASIFICACIÓN

El indicador de distancia Δ^2 entre dos puntos representados por dos vectores de *m* observaciones cada uno (que es el perfil de dos unidades experimentales) es de un particular interés en el *análisis de conglomerados*, por que los conglomerados de los objetos experimentales se determinan mediante valores que muestran la misma función, que es el de indicar la distancia que hay entre un objeto y un conglomerado, en el que valores de distancias entre dos puntos, como Δ^2, son vistos en el análisis de conglomerados como medidas de *proximidad* (disimilaridad y similaridad).

Hay dos tipos de distancias *Mahalanobis*, que son muy importantes en el análisis discriminante. El primero es el indicador de la distancia entre dos puntos, donde cada

punto es un vector de medias de las m variables. El vector de medias es llamado ***centroide***, por ejemplo, el centroide de la población g es denotado por

$$\mu'_g = [\mu_{1g}, \mu_{2g}, \ldots, \mu_{mg}]$$

donde μ_{ig} es la media de la variable i en la población g. La distancia entre dos poblaciones, 1 y 2, es la distancia entre sus centroides, y es

$$\Delta_{12} = [(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)]^{1/2}$$

El segundo tipo de *distancia Mahalanobis*, es la distancia entre un objeto (unidad) y una población, ésto es la distancia entre dos puntos, donde un punto es un vector de m observaciones de un objeto experimental, y el otro punto es el centroide de una población. Por ejemplo, supongamos que se tienen k poblaciones, y se quiere tener la distancia entre el objeto u y la población g, esto es la distancia entre X_u (vector de observaciones del objeto u) y μ_g (centroide de la población g).

Si tenemos un vector de observaciones para un objeto experimental u y una población g, entonces

$$\Delta^2_{ug} = (X_u - \mu_g)' \Sigma_g^{-1} (X_u - \mu_g) \qquad (2.2)$$

donde Σ_g es la matriz de covarianza de la población g, y así se indica el cuadrado de la distancia del punto X_u a el punto que representa el centroide de la población g.

Esta última distancia es de gran interés en los análisis de clasificación (Análisis de Conglomerados), por que su principal objetivo es el clasificar un objeto en la población a la que se encuentre más cercana.

2.4 REGLAS DE CLASIFICACIÓN EN GENERAL

El propósito básico del ADP es el crear una regla de clasificación, mediante la cual se

puedan clasificar objetos en una de las k poblaciones que se tienen. Esto basándose en la matriz de datos del orden N x m, donde $N = \Sigma n_g$, y n_g (g = 1, ..., k) es el número de objetos que hay en la población g, y cada objeto debe tener m mediciones.

Las reglas de clasificación o decisión comúnmente usadas, se basan en el principio de *máxima probabilidad* que es: *Asignar un objeto a la población en la cual su vector de observaciones tenga la más grande probabilidad de ocurrencia.*

Supongamos que f_1 y f_2 son las funciones de densidad para las poblaciones 1 y 2; la probabilidad de una observación $X=a$ en la población 1 es $f_1(a)$. Aplicando el principio de *máxima probabilidad*, se llega a la regla de asignar un objeto $X=a$ a la población 1 si $f_1(a) > f_2(a)$, que es la probabilidad de una observación $X=a$ en la población 1 más grande que en la población 2, de otro modo se asigna el objeto a la población 2.

En términos múltiples se tiene que la regla de máxima probabilidad es, asignar el objeto u a la población g, si la probabilidad del vector de observaciones X_u , es más grande para el grupo g que para algún otro grupo; pero en el caso de que la probabilidad más grande del vector de observaciones X_u , sea igual en dos o más grupos, se asigna el objeto u al grupo que tenga el mayor número de elementos en ese instante.

Al ser f la función de densidad, entonces la regla de máxima probabilidad es: asignar la unidad u a la población g si la probabilidad del vector de observaciones X_u es de más valor para el grupo g que para algún otro grupo, y la regla queda establecida como

Asignar u a la población g si
$f(X_u \mid g) > f(X_u \mid g')$
para $g' \neq g$

donde $f(X_u,g)$ es el valor de la función de densidad de X_u dado que se tiene una población g. En términos probabilísticos una probabilidad es denotada como $P(X_u|g)$, que es la probabilidad del vector de observaciones X_u dado que se tiene la población g.

La regla puede ser interpretada de mejor forma en términos de probabilidad inversa al intercambiar los parámetros, y de esa manera establecer como el evento del que se tiene evidencia al vector de observaciones de la unidad u, X_u , y así, una probabilidad es denotada como $P(g\ X_u)$ que es la probabilidad del objeto u dado que está en la población g; considerando ahora la probabilidad de que la unidad u pertenezca a la población g. $P(g\ X_u)$ es llamada la *probabilidad posterior* [6] de los miembros en la población g, *posterior* en el sentido de que es la probabilidad de los miembros de la población g con la condición de que el vector X_u es conocido.

La probabilidad de que el objeto u pertenezca a la población g, es igual a la razón de la probabilidad del vector de observaciones de u (X_u) en la población g, respecto a la suma de las probabilidades del mismo vector asociadas con todas las k poblaciones.

$$P(g|X_u) = \frac{P(X_u|g)}{\sum_{g'=1}^{k} P(X_u|g')} \qquad (2.3)$$

Y la regla es interpretada como

Asignar u a la población g si
$P(g \mid X_u) > P(g' \mid X_u)$
para $g' \neq g$, donde $P(g\|X_u)$ es definida de (2.3)

Los valores estimados de las probabilidades $P(X_u \mid g)$, $g = 1, 2, \ldots, k$, están basados en las k muestras, y por lo tanto dependen del tamaño de las mismas. Es por eso por lo que deben ser tomados en cuenta los tamaños de las k muestras.

Sea π_g la proporción de unidades en el universo total (es decir, el conjunto de las k

[6] Probabilidad Posterior, es la probabilidad condicional P(A|B) de un evento A dado que ya se tiene el grado de creencia del evento B que puede suceder. (n. a.)

distribución son conocidos). Los parámetros que usualmente son desconocidos son Σ y μ, entonces para poder usar las reglas deben ser estimados.

Hay varias formas de estimar los parámetros, y así poder construir una regla de clasificación que los use. La más comúnmente usada es cuando se especifica un modelo de distribución de probabilidad teórica, suponiendo que los datos se ajustan al modelo, y así estimar los parámetros del modelo tomando los datos para construir una regla usando las estimaciones. Otra manera de estimar es, usando la anterior forma de estimación y además estimar los valores de densidad directamente de los datos con un modelo de especificación no priori, y así construir una regla de especificación usando estas estimaciones.

Con este tipo de aproximaciones, en el que se tienen valores adicionales conocidos (las estimaciones de los parámetros de la función de densidad) se obtendrá una regla de clasificación del tipo bayesiano, por el simple hecho de usar valores adicionales conocidos y así obtener la regla de clasificación óptima.

La familia de las funciones de densidad probabilística normal univariada está definida por

$$f(X|g) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_g^2}} \exp\left[-\frac{1}{2}\frac{\left(X-\mu_g\right)^2}{\sigma_g^2}\right] \qquad (2.6)$$

donde μ_g y σ^2_g son la media y varianza, de la población g.

Pero como casi nunca se conocen los valores de μ_g y σ^2_g, entonces es necesario usar estimaciones basadas en la muestra de datos y así obtener una estimación de $f(X|g)$

$$\hat{f}(X|g) = \frac{1}{\sqrt{2\pi}\sqrt{s_g^2}} \exp\left[-\frac{1}{2}\frac{\left(X-\bar{X}_g\right)^2}{s_g^2}\right]$$

donde $\bar{X}_g = \frac{1}{n_g}\sum_u X_u$ y $s_g^2 = \frac{1}{n_g - 1}\sum_u (X_u - \bar{X}_g)^2$

La generalización de (2.6) a el caso multivariado puede ser hecha por analogía para llegar a las funciones de densidad de probabilidad normal *m*-variada, la cual es definida por

$$f(X|g) = \frac{1}{\sqrt{(2\pi)^m}\sqrt{|\Sigma_g|}} \exp[-\tfrac{1}{2}(X - \mu_g)' \Sigma_g^{-1} (X - \mu_g)] \qquad (2.7)$$

La generalización de σ^2_g es la matriz de covarianza de la población g, Σ_g de orden (*mxm*), y el vector de medias de orden (*m*x1) μ_g es la generalización de μ_g; el determinante de Σ_g , que es $|\Sigma_g|$, es llamada la varianza generalizada del conjunto de *m* variables. Análogos al caso univariado, la especificación de μ_g y Σ_g complétamente determinan la función de densidad normal para la población g.

Pero en el análisis de datos rara vez son conocidos los valores de estos parámetros, entonces es necesario determinar estimadores para $f(X \mid g)$.

Esto se hace insertando estimaciones de μ_g y Σ_g en la expresión (2.7), el vector μ_g es estimado por $\bar{x}_g$ el cual es el vector de medias (*m*x1) de la muestra g, y Σ_g puede ser estimado por S_g que es la matriz de covarianza (*mxm*) para la muestra g. Sustituyendo en la expresión (2.7) queda como

$$\hat{f}(X|g) = \frac{1}{\sqrt{(2\pi)^m}\sqrt{|S_g|}} \exp[-\tfrac{1}{2}(X - \bar{x}_g)' S_g^{-1} (X - \bar{x}_g)] \qquad (2.8)$$

donde $\bar{x}_g$ es el vector de medias *m*x1 de la muestra g, y S_g es la matriz de covarianza *mxm* para la muestra g. Los elementos de la diagonal principal de S_g son las *m* varianzas definidas por

$$S_g^{(i,i)} = \frac{1}{n_g - 1}\sum_u \left(X_u^{(i)} - \bar{X}_g^{(i)}\right)^2$$

y los elementos que no están en la diagonal principal son covarianzas definidas por

$$S_g^{(i,i')} = \frac{1}{n_g - 1}\sum_u (X_u^{(i)} - \bar{X}_g^{(i)})(X_u^{(i')} - \bar{X}_g^{(i')})$$

Así la *distancia Mahalanobis* (2.2), que indica el cuadrado de la distancia entre un vector de observaciones del objeto u y el centroide para la muestra g, puede ser escrita como:

$$D^2_{ug} = (X_u - \bar{x}_g)' S_g^{-1} (X_u - \bar{x}_g)$$

y la expresión (2.8) puede ser expresada como

$$\hat{f}(X_u|g) = (2\pi)^{-n/2} |S_g|^{-1/2} \exp(-\tfrac{1}{2} D^2_{ug}) \qquad (2.9)$$

2.6 REGLAS DE CLASIFICACIÓN BASADAS EN NORMALIDAD

Como la regla de clasificación se basa en probabilidades posteriores, éstas serán dadas por (2.4), pero como se trabaja con estimadores dados por (2.9) entonces queda de la forma

$$\hat{P}(g|X_u) = \frac{q_g \cdot \hat{f}(X_u|g)}{\sum_{g'=1}^{k} q_{g'} \cdot \hat{f}(X_u|g')} \qquad (2.10)$$

donde $q_g = \hat{\pi}_g$

Sustituyendo (2.9) en (2.10) y simplificando queda como

$$\hat{P}(g|X_u) = \frac{q_g \cdot |S_g|^{-1/2} \quad \exp(-\frac{1}{2} D^2_{ug})}{\sum_{g'=1}^{k} q_{g'} \cdot |S_{g'}|^{-1/2} \quad \exp(-\frac{1}{2} D^2_{ug'})} \qquad (2.11)$$

de esta manera se llega a la siguiente regla de clasificación para un caso normal m-variado

> Asignar u a la población g si
>
> $\hat{P}(g|X_u) > \hat{P}(g'|X_u)$ (2.12)
>
> para $g \neq g'$, donde $\hat{P}(g|X_u)$ está definido por (2.11)

Considerando el caso especial donde las k matrices de covarianzas de las poblaciones son iguales

$$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k = \Sigma \qquad (2.13)$$

sólo es necesario estimar una matriz de covarianza (mxm) Σ por S.

La distancia cuadrada del objeto u al centroide de la población g es expresada por

$$D^{*2}_{ug} = (X_u - \bar{x}_g)' S^{-1} (X_u - \bar{x}_g)$$

y asi la probabilidad estimada de (2.9) es

$$\hat{f}(X_u|g) = (2\pi)^{-m/2} \cdot |S|^{-1/2} \exp(-\tfrac{1}{2} D^{*2}_{ug})$$

Sustituyendo en (2.10) y simplificando queda como

$$\hat{P}(g|X_u) = \frac{q_g \cdot \exp(-\frac{1}{2} D^{*2}_{ug})}{\sum_{g'=1}^{k} q_{g'} \cdot \exp(-\frac{1}{2} D^{*2}_{ug'})} \qquad (2.14)$$

Y se obtiene una regla de máxima probabilidad para el caso normal m-variado, bajo la condición (2.13)

Asignar u a la población g si $\hat{P}(g\|X_u) > \hat{P}(g'\|X_u)$ para $g \neq g'$, donde $\hat{P}(g\|X_u)$ está definido por (2.14)	(2.15)

2.7 FUNCIONES DE CLASIFICACIÓN

Las reglas (2.12) y (2.15) pueden ser establecidas en términos de los numeradores, y de esta forma maximizar el numerador.

Así para (2.11) la regla queda en maximizar

$$q_g \cdot |S_g|^{-1/2} \exp(-\tfrac{1}{2} D^2_{ug}) \qquad (2.16)$$

El numerador de la expresión (2.11) siempre será positivo, pero el máximo valor que puede alcanzar es 1. Aunque ese valor nunca será alcanzado, ya que si se obtuviera se estaría calculando la distancia de un punto con sí mismo, que es un caso sin sentido en este tipo de análisis. Las probabilidades posteriores (2.11) y (2.14) son creadas en base al criterio de maximizar la varianza relativa entre los grupos con respecto a la varianza del grupo al cual es asignado el objeto que se pretende clasificar; una expresión más sencilla pero con estas mismas propiedades, las cuales siguen siendo respetadas por (2.16), es el logaritmo natural de éste mismo, que es

$$Q_{ug} = \ln q_g - \tfrac{1}{2} \ln |S_g| - \tfrac{1}{2} D^2_{ug} \qquad (2.17)$$

(2.17) es una expresión en la que intervienen los dos puntos específicos de los que se pretende encontrar una probabilidad posterior, es por eso por lo que el maximizar (2.16) es equivalente a maximizar su logaritmo natural. De esta manera la regla para el caso normal m-variado puede ser expresada como

Asignar u a la población g si
$Q_{ug} > Q_{ug'}$
para $g \neq g'$, donde Q_{ug} es definida por (2.17)

(2.18)

Es el mismo caso, para cuando se trata con la condición (2.13), que sería el obtener el máximo valor del logaritmo natural del numerador de (2.14) que es

$$L_{ug} = \ln q_g - \tfrac{1}{2} D^{*2}_{ug} \qquad (2.19)$$

y la regla queda como

Asignar u a la población g si
$L_{ug} > L_{ug'}$
para $g \neq g'$, donde L_{ug} es definida por (2.19)

(2.20)

(2.17) en términos de X_u es una expresión cuadrática (ver anexo 1) y es llamada *Función de Clasificación Cuadrática* (QCF, Quadratic Classification Function) y la regla (2.18) es conocida como ***Regla de Clasificación Cuadrática***; (2.19) viene siendo una expresión lineal en términos de X_u (ver anexo 1) y es por eso por lo que se le llama *Función de Clasificación Lineal* (LCF, Linear Classification Function) y la regla (2.20) es conocida como ***Regla de Clasificación Lineal***.

Refiriéndonos a la función QCF (2.17), el maximizar Q_{ug} es equivalente a minimizar

$$d_{ug} = -2Q_{ug} = \ln |S_g| + D^2_{ug} - 2 \ln q_g \qquad (2.21)$$

Así la máxima probabilidad es igual a encontrar la distancia mínima, y la regla puede ser expresada como

Asignar u a la población g si
$d_{ug} < d_{ug'}$
para $g \neq g'$, donde d_{ug} es definida por (2.21)

(2.22)

Cuando todas las matrices de covarianza son iguales, y esta condición es considerada, entonces el maximizar L_{ug} es equivalente a minimizar

$$d^*_{ug} = -2\, L_{ug} = D^{*2}_{ug} - 2\, ln\, q_g \qquad (2.23)$$

Y la regla de máxima probabilidad (o mínima distancia) para el caso de matrices de covarianza iguales, puede ser expresada como

Asignar u a la población g si
$d^*_{ug} < d^*_{ug'}$
para $g \neq g'$, donde d^*_{ug} es definida por (2.23)

(2.24)

El considerar o saber que las matrices de covarianza de las k poblaciones sean iguales

$$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$$

es tomado como un caso especial para poder establecer con mayor precisión la regla de clasificación a usar. Pero también debe ser vista con la misma importancia el caso de cuando las k probabilidades a priori π_g , $g = 1, 2, \ldots, k$ son iguales, ésto es generado sólo cuando el tamaño de las k poblaciones es el mismo

$$q_1 = q_2 = \ldots = q_k$$

al considerar esta condición, la expresión (2.11) que define la regla (2.12) queda como

$$\hat{P}(g|X_u) = \frac{|S_g|^{\frac{1}{2}} \exp(-\frac{1}{2} D^2_{ug})}{\sum_{g'=1}^{k} |S_{g'}|^{\frac{1}{2}} \exp(-\frac{1}{2} D^2_{ug'})} \qquad (2.25)$$

y para el caso en el que las k matrices de covarianza son iguales la expresión (2.14) que define la regla (2.15) se expresa como

$$\hat{P}(g|X_u) = \frac{\exp(-\frac{1}{2} D^{*2}_{ug})}{\sum_{g'=1}^{k} \exp(-\frac{1}{2} D^{*2}_{ug'})} \qquad (2.26)$$

Al ser igual el tamaño de las k poblaciones ocasiona que las probabilidades a priori sean también iguales, así la *función de clasificación cuadratica* (2.17) queda simplificada como

$$Q_{ug} = -\frac{1}{2} \ln |S_g| - \frac{1}{2} D^2_{ug} \qquad (2.27)$$

y en términos de distancia es

$$d_{ug} = -2Q_{ug} = \ln |S_g| + D^2_{ug} \qquad (2.28)$$

Y para la función de *clasificación lineal* (2.19) queda como

$$L_{ug} = -\frac{1}{2} D^{*2}_{ug} \qquad (2.29)$$

y en términos de distancia es

$$d^{*}_{ug} = -2 L_{ug} = D^{*2}_{ug} \qquad (2.30)$$

En la siguiente tabla se muestran todas las reglas de predicción, junto con las condiciones que las caracterizan para su aplicación correcta:

PROBABILIDES A PRIORI	MATRICES DE COVARIANZA	
	DIFERENTES	IGUALES
DIFERENTES *Las k poblaciones son de distinto tamaño*	$\hat{P}(g\|X_u) = \dfrac{q_g \cdot \|S_g\|^{-1/2} \exp(-\frac{1}{2} D^2_{ug})}{\sum_{g'=1}^{k} q_{g'} \cdot \|S_{g'}\|^{-1/2} \exp(-\frac{1}{2} D^2_{ug'})}$	$\hat{P}(g\|X_u) = \dfrac{q_g \cdot \exp(-\frac{1}{2} D^{*2}_{ug})}{\sum_{g'=1}^{k} q_{g'} \cdot \exp(-\frac{1}{2} D^{*2}_{ug'})}$
	Reglas Cuadráticas	Reglas Lineales
	$Q_{ug} = \ln q_g - \frac{1}{2} \ln\|S_g\| - \frac{1}{2} D^2_{ug}$ $d_{ug} = \ln\|S_g\| + D^2_{ug} - 2 \ln q_g$ [a]	$L_{ug} = \ln q_g - \frac{1}{2} D^{*2}_{ug}$ $d^*_{ug} = D^{*2}_{ug} - 2 \ln q_g$ [a]
IGUALES *El tamaño de las k poblaciones es el mismo*	$\hat{P}(g\|X_u) = \dfrac{\|S_g\|^{-1/2} \exp(-\frac{1}{2} D^2_{ug})}{\sum_{g'=1}^{k} \|S_{g'}\|^{-1/2} \exp(-\frac{1}{2} D^2_{ug'})}$	$\hat{P}(g\|X_u) = \dfrac{\exp(-\frac{1}{2} D^{*2}_{ug})}{\sum_{g'=1}^{k} \exp(-\frac{1}{2} D^{*2}_{ug'})}$
	Reglas Cuadráticas	Reglas Lineales
	$Q_{ug} = -\frac{1}{2} \ln\|S_g\| - \frac{1}{2} D^2_{ug}$ $d_{ug} = -2Q_{ug} = \ln\|S_g\| + D^2_{ug}$ [a]	$L_{ug} = -\frac{1}{2} D^{*2}_{ug}$ $d^*_{ug} = D^{*2}_{ug}$ [a]

Nota:

$$D^2_{ug} = (X_u - \bar{X}_g)' S_g^{-1} (X_u - \bar{X}_g)$$

$$D^{*2}_{ug} = (X_u - \bar{X}_g)' S^{-1} (X_u - \bar{X}_g)$$

[a] Por propósitos de clasificación, el máximo de k valores es considerado, excepto para éstos cuatro, donde el mínimo es considerado, por estar en términos de distancias.

Tabla 1. Tabla de clasificación de las reglas de clasificación.

Para hacer una selección exacta de la regla de predicción a usar, se deben considerar dos conceptos básicos, que son

1) La igualdad de las matrices de covarianza de las k poblaciones.
2) La igualdad del tamaño de las k poblaciones (probabilidades a priori).

Pero hay una condición fundamental que no debe ser descartada para poder usar las reglas de predicción de manera correcta. Esta condición es el saber si los vectores de observaciones de las m variables predictoras en los k grupos poseen una distribución normal, es decir, si cada población de la que surgen las distintas mediciones para los objetos tienen distribución normal. Esta condición debe ser tomada en cuenta por el hecho de que todas las reglas son creadas con el supuesto de que los datos tienen una distribución normal multivariada.

Si no se conociera la distribución de las poblaciones de las m variables predictoras, hay varias técnicas gráficas y empíricas, para poder saber si los datos poseen una distribución normal multivariada.

Pero, ¿es necesario saber esta condición para poder usar las reglas de clasificación? Si se desea hacer un análisis desde su inicio de la forma más estricta, la respuesta es "sí", pero si se crearan las reglas con datos no-normales y se aplicaran, se estaría haciendo un "análisis a ciegas", y no sería extraño el obtener resultados ilógicos o erróneos y no apegados a la realidad, sin descartar el caso de obtener resultados correctos, que expliquen el fenómeno que se analizó.

2.8 HIT RATE (Porcentaje Correcto)

Los resultados finales de un ADP son frecuentemente presentados de manera resumida en una tabla de clasificación, en esta tabla se muestran los números de objetos de cada grupo que son asignados a cada uno de los otros grupos.

		Grupos predecidos			
		1	2	3	Total
Grupo actual	1	n_{11}	n_{12}	n_{13}	n_1
	2	n_{21}	n_{22}	n_{23}	n_2
	3	n_{31}	n_{32}	n_{33}	n_3
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

Tabla 2. Tabla de los resultados de clasificación, para el caso de $k = 3$ (3 conglomerados).

Por ejemplo en la Tabla 2, la celda $n_{gg'}$ $(g \neq g')$ indica el número de objetos del grupo g que fueron asignados (predecidos) a estar en el grupo g'. El porcentaje de los objetos que son clasificados correctamente por las reglas de clasificación, se le conoce como *hit rate*. El *hit rate* para el grupo g está dado por n_{gg} / n_g y el *hit rate* con respecto a la cantidad total de objetos es $\Sigma n_{gg} / N$.

Hay tres tipos de probabilidades, las cuales pueden ser consideradas como buenos estimadores de la proporción de clasificación correcta (*hit rate*). Una es el ***hit rate óptimo*** denotado como $P^{(o)}$. Este es el *hit rate* obtenido cuando se es aplicada a la población una regla de clasificación basada con parámetros conocidos (es decir, los k vectores de medias de las subpoblaciones y la matriz común de covarianza). La segunda es el ***hit rate actual*** (a veces llamado el *hit rate incondicional*) denotado como $P^{(a)}$. Este *hit rate* es obtenido cuando se aplica a futuras muestras (subpoblaciones, tomadas de la misma población) una regla de clasificación basada en una muestra (subpoblación) particular. $P^{(a)}$ puede ser vista como la proporción esperada de clasificaciones correctas sobre muestras futuras dadas por una regla basada en estadísticos de una muestra particular. El tercer *hit rate* a considerar es el ***hit rate actual esperado*** denotado como $P^{(e)}$, que es la proporción de clasificaciones correctas que surge al aplicar la regla de clasificación, basada en una muestra particular en la población de tamaño $N = \Sigma n_g$, $g = 1, \ldots, k$ (es decir, sobre las k subpoblaciones), como el $P^{(e)}$ es aplicado en la población completa, entonces, $P^{(e)} = E(P^{(a)})$. Este *hit rate* es de interés especial antes de que alguna muestra de objetos sea clasificada.

Los resultados de clasificación así como sus respectivos *hit rates* pueden ser obtenidos

mediante dos tipos de procesos:

Análisis Interno o *Análisis Externo*.

2.8.1 ANÁLISIS INTERNO

Un *análisis de clasificación interno* es cuando las muestras son clasificadas en base a la regla de clasificación con los parámetros estimados a partir de ellas mismas (los k vectores de medias de las subpoblaciones, μ_g, la matriz común de covarianza, Σ, la proporción de objetos en la población total, π_g, etc.).

2.8.2 ANÁLISIS EXTERNO

Mientras que en el análisis interno los objetos clasificados son los mismos de los que se obtuvieron los parámetros para la regla de clasificación, en un *análisis de clasificación externo* la regla de clasificación es determinada de un conjunto particular de objetos, y entonces usada para clasificar otros conjuntos de objetos.

Hay tres métodos para poder llevar a cabo un *análisis externo*:

2.8.2.1 MÉTODO EXTENDIDO

Una forma de llevar a cabo una clasificación externa es por medio de una partición de la muestra en dos submuestras: (1) una muestra diseñadora, y (2) una muestra de prueba. La regla de clasificación es determinada usando los datos de la muestra diseñadora y entonces es aplicada a los datos de la muestra de prueba. Este método es llamado el *método extendido*. Su *hit rate* estimado es la proporción de los objetos de la muestra de prueba que son correctamente clasificados (usando la regla desarrollada sobre la muestra diseñadora).

2.8.2.2 MÉTODO Dejar - uno - fuera

Una segunda manera de llevar a cabo un análisis externo es mediante el método llamado *método dejar-uno-fuera* (método que P. A. Lachenbruch dió a conocer en 1967). El método involucra un proceso de dos pasos. En el primero, se borra un objeto y (por considerar que las matrices de covarianza son iguales) las funciones de clasificación lineal son determinadas en base a los N-1 objetos sobrantes. Entonces las LCF's (funciones lineales de clasificación) son usadas para clasificar el objeto borrado en uno de los k grupos criterio. Este proceso se lleva a cabo N veces borrando en cada iteración un objeto hasta tener un solo elemento en la muestra diseñadora, y las proporciones de clasificación correcta de los objetos borrados son usadas como un estimador del *hit rate* de la clasificación correcta.

Para cada clasificación, se puede considerar a la muestra de los objetos sobrantes como la muestra diseñadora, y a la muestra de los objetos borrados como la muestra de prueba.

2.8.2.3 MÉTODO DE LA PROBABILIDAD POSTERIOR MÁXIMA (P. P. M.)

La razón del uso del término *Máxima* es implicada por la definición de un estimador para $P_g^{(a)}$ y es explícita en la definición de un estimador para $P^{(a)}$. El estimador de P.P.M. para $P_g^{(a)}$, es simplemente una "media" de las probabilidades posteriores estimadas para objetos de todos los grupos asignados a la población g por la regla de clasificación usada. La suma de estas estimaciones es dividida por $N \cdot q_g$ y de ese modo se obtiene el *hit rate*.

$$\hat{P}_g^{(a)} = \frac{1}{N \cdot q_g} \sum_{g'=1}^{k} \left\{ \sum_{u=1}^{n_{g'}} \left[\begin{array}{c} prob.\ post.\ para\ todos\ X_u\ en \\ el\ grupo\ g'\ asignado\ al\ grupo\ g \end{array} \right] \right\} \qquad (2.31)$$

El valor de $P^{(a)}$ puede ser estimado usando

$$\hat{P}^{(a)} = \sum_{g=1}^{k} q_g \hat{P}_g^{(a)} = \frac{1}{N} \sum_{u=1}^{N} max \left[\begin{array}{cccc} \hat{P}(1|X_u), & \hat{P}(2|X_u), & ..., \\ \hat{P}(g|X_u), & ..., & \hat{P}(k|X_u) \end{array} \right] \qquad (2.32)$$

El estimado de $I^{**'}$ se calcula de la media de las máximas probabilidades posteriores para cada objeto.

Las probabilidades posteriores $\hat{P}(g|X_u)$ de (2.32) pueden ser calculadas mediante cualquiera de los análisis internos y externos.

2.9 SELECCIÓN DE VARIABLES PREDICTORAS

En el análisis de clasificación, un incremento en el número de variables predictoras afecta los resultados. A diferencia de regresión cuando m (número de variables predictoras) es aumentada, los *hit rates* (de grupos separados y/o grupo total) se decrementan; ésto sucede si las variables a ser añadidas no contribuyen substancialmente a las diferencias entre grupos, y similar a regresión como m aumente, el *sesgo positivo* del *hit rate* interno se incrementa.

Una meta del análisis de clasificación es el crear una regla con alto grado de precisión para ser usada con muestras subsecuentes. Si esta propuesta no descarta mejorar el grado de exactitud para predecir de un subconjunto de variables predictoras, entonces es apropiado borrar las variables predictoras que menos influyan en el objetivo de la regla, y con ésto se reduce la complejidad de la clasificación. Como esta regla será usada con muestras subsecuentes, el número de predictoras debe ser relativamente menor al tamaño de la muestra (Huberty, C. J.), y las variables predictoras deben ser ordenadas con respecto a su contribución de la clasificación corréctamente obtenida.

Modelos de un *análisis discriminante predictivo* con pocas variables predictoras, relativas a N ($N = \Sigma n_g$, g = 1,..., k) dejan relativamente más exactitud (es decir, menos sesgo) y estimadores más precisos (ver, Hora y Wilcox, 1982).

Los dos métodos más populares y conocidos para seleccionar y ordenar a las variables predictoras más significativas en el proceso de predicción son:

2.9.1 MÉTODO SIMULTÁNEO

Este método obtiene la *función discriminante* considerando a todas las variables predictoras concurrentemente. Así la función o funciones discriminantes a ser calculadas se basan en el conjunto de todas las variables predictoras considerando a éstas con el mismo ***poder discriminante*** [8]. El método es apropiado cuando por razones teóricas del fenómeno, el analista necesita incluir todas las variables predictoras en el análisis, y no se interesa por ver los resultados intermedios basados solamente en la mayoría de las variables predictoras.

2.9.2 MÉTODO *Stepwise*

Este método involucra las variables predictoras en la función discriminante en base a los poderes discriminantes de cada variable. Este método fue sugerido por Smith (1984) y comienza por escoger a la mejor variable predictora mediante su *poder discriminante*, utilizando el método de análisis externo *Dejar-uno-fuera* como se describe a continuación:

1.- Hacer m ADP's univariados, con cada variable predictora y obtener el *hit rate* de interés por medio del análisis externo *Dejar-uno-fuera*, entonces el mejor subconjunto de tamaño 1 será el de aquella variable predictora que dé el más alto de los *hit rates*. Supongamos que la variable V1 es la que dio el mayor *hit rate*.

2.- Hacer m-1 ADP's bivariados con todos los subcojuntos de 2 variables (incluida V1), que son V1 y V2, V1 y V3, . . . , V1 y Vm, y obtener el *hit rate* de interés para cada par (usando *Dejar-uno-fuera*). El mejor subconjunto de tamaño 2 (dado que V1 está incluida) consiste del par, supongamos que es V1 y V2, que da el más alto de los m-1 *hit rates*.

3.- Hacer m-2 ADP's de tres variables, usando todos los subconjuntos de

[8] La taza real "hit rate", es considerada como el poder discriminante. (n. a.)

CAPÍTULO 3

CLASIFICACIÓN Y DISCRIMINACIÓN

3.1 INTRODUCCIÓN AL CASO DE APLICACIÓN

Hoy en día la calidad[1] es el factor más importante para el crecimiento de una empresa. Todos los procedimientos en el desarrollo y producción de un producto son partes fundamentales de la calidad que finalmente definirá el valor del producto. Todos los procedimientos por los que pasa un producto, desde los procesos de llegada de la materia prima con que se producen, hasta el proceso de almacenamiento del producto terminado, forman parte de la calidad con que éste termina.

La planta de ***México*** de una ***Compañía Hulera Transnacional***, la cual puede producir más de 100 tipos de ***llantas radiales***[2] para carro y camioneta (de distinto tamaño y diseño), y de ese modo considerada como la planta de América Latina con la más amplia variedad en producción de llantas radiales, enfrenta el problema de tener una gran cantidad de procedimientos en la producción de llantas; es por eso que la *Compañía Hulera* desea tener un mayor control de los procedimientos de desarrollo y producción de las llantas para poder aumentar su calidad.

Por tener la planta información multivariada se consideró a la *Estadística Multivariada* como la herramienta adecuada para poder encontrar la solución a este tipo de problema.

El *análisis de conglomerados* es una buena técnica de análisis exploratorio para usarse cuando se desea saber si existen agrupaciones de objetos (en este caso las llantas), con un gran parecido o semejanza entre ellos (*homogéneos*), pero con una gran diferencia (*heterogéneos*) con las llantas de las demás agrupaciones, y así obtener grupos de llantas (*conglomerados*) con una gran semejanza en sus procedimientos de desarrollo y producción que darán facilidad al manejo del control de calidad de las mismas cuando esten en producción.

Para esto se requieren de los dos factores fundamentales de la estadística multivariada,

---

[1] Calidad: Estado de una cosa, su naturaleza, sus circunstancias y condiciones que se requieren para tener una gran importancia y gravedad.
ENCICLOPEDIA SALVAT Diccionario, Tomo 3, pág. 618

[2] Llantas radiales: Llantas en que las cuerdas de sus capas tienen la misma dirección del radio de la llanta. (n.a.) (ver anexo 2)

que son los *objetos* a analizar y sus atributos o *variables* en que serán medidos. A cada tipo de llanta se le considera un objeto (caso) el cual tiene asignado un número como etiqueta para su identificación.

3.2 ESTABLECIMIENTO DE LAS VARIABLES

Hoy en dia los clientes están solicitando cada vez más llantas con alto rendimiento y desempeño, y de esa manera más sensibles al manejo, es por eso por lo que se desea clasificar a las llantas en distintos grupos mediante los cuales se puedan establecer procedimientos especiales para el mejoramiento de su calidad; para eso se especificó como principal factor de los grupos (conglomerados), el desempeño de las llantas, pero como en la planta hay una gran variedad de llantas en medidas y diseños, también es necesario que los grupos tengan un factor que distinga a las llantas y otro que especifique su volumen de producción.

La *Uniformidad* es el principal indicador de calidad sobre el cual se mide el desempeño de las llantas. Uniformidad significa obtener y mantener un proceso estable, es decir, sin variaciones, ésto es fabricar un producto dentro de especificaciones de calidad, con las mismas características de operación y materiales con que fue producido ayer y se producirá mañana.

La *uniformidad* en una llanta es un conjunto de características las cuales provocan un rodaje agradable, sin vibraciones, trepidaciones[3] o inestabilidad en un vehículo, dando una sensación de confort y seguridad para el conductor y sus ocupantes.

Los parámetros de uniformidad que solicitan los clientes en una llanta son:

- El que se considera como una condición en la cual un auto está rodando en un pavimento plano, pero al recibir carga y presión, el auto brinca hacia arriba y hacia abajo como si estuviera en un camino de terracería; a esta condición se le llama ***Fuerza Radial***, el cuál es causado por las llantas que no son perfectamente redondas (ver Figura 3.1).

[3] Trepidar: temblar, titubear. (n. a.)

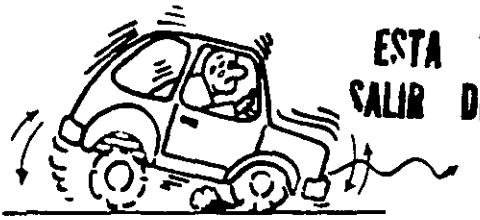


Figura 3.1. Condición de la Fuerza Radial.

- El que se considera como una condición en la cual el auto se va de un lado hacia el otro al mismo tiempo que vibra, a esta condición se le llama ***Fuerza Lateral***, la cual puede definirse como una condición que hace que el vehículo viaje bamboleándose de lado a lado (ver Figura 3.2).

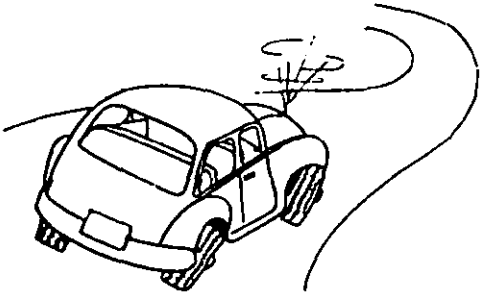


Figura 3.2. Condición de la Fuerza Lateral.

- Hay un parámetro muy especial en el desempeño de las llantas, al cual se le llama ***Primera Armónica***, este parámetro es obtenido mediante una descomposición (matemática) de la variación de *fuerza radial* en una serie infinita de ondas periódicas senoidales, de las cuales la *1ra. armónica* es aquella que tiene una periodicidad de 360° y normalmente tiene una magnitud del 60% al 80% del total de variación de la *fuerza radial*, y es medida como la variación de pico a pico en la curva sinusoidal[4] que mejor se apega a la curva total de la variación de la *fuerza radial* (ver Figura 3.3).

[4] Gráfica de la función de seno que se apega a una gráfica de variación de manera que la suaviza. (n. a.)

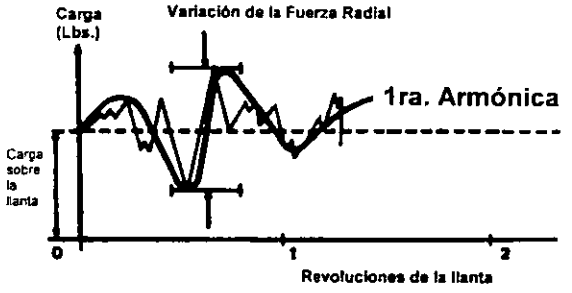


Figura 3.3. Gráfica de la Primera Armónica.

- Uno que es la condición que hace que la llanta desvíe al auto hacia un lado, el cual es llamado ***Conicidad***. Una sección cónica se comporta de la misma manera que una llanta con *conicidad* alta, de aquí el término *conicidad*. Si el auto se va hacia la izquierda se dice que su *conicidad* es negativa, y si se va hacia la derecha es positiva (ver Figura 3.4).

Figura 3.4. Condición del Parámetro 4.

Los cuatro parámetros de uniformidad mencionados, son mediciones de carga y presión (fuerza) de las llantas, pero en el estudio serán tomadas las mediciones con que son dados a conocer a los clientes, que es el porcentaje de llantas de una producción mensual que respetan los límites (establecidos por la compañía) de cada parámetro.

La *uniformidad total* es medida con un indicador llamado ***Desempeño***, que son los límites establecidos por la Compañía Hulera para cada parámetro de uniformidad. El indicador ***Desempeño*** es el porcentaje de llantas que respetan los parámetros de calidad en forma simultánea, por ejemplo, se puede decir que un tipo de llanta tiene 65%

de desempeño, cuando el 65% de su producción cumple con los cuatro parámetros de uniformidad simultáneamente (***Fuerza radial***, ***Fuerza lateral***, ***1ra. Armónica*** y ***Conicidad***).

De esa manera se establecen como las variables predictoras con el factor de calidad, a los 5 parámetros de uniformidad.

El otro factor de los grupos es la distinción de las llantas, todas las llantas son diferentes en tamaño, compuestos, peso, y diseño, entonces cualquier variable de las anteriores se pudo haber tomado como la variable que diferenciará a las llantas, pero por ser el ***Peso*** de la llanta (en Kilogramos), una variable sencilla y clara para distinguir los tipos de llantas, se tomó como la variable predictora que especifica la diferencia (distinción) entre las llantas.

Como el volumen de producción de las llantas es muy grande, la uniformidad se obtiene mediante muestras relacionadas con su cantidad de producción, es por eso que se agregó una variable predictora que considere el volumen de producción de cada llanta, esta variable es la ***Cantidad Muestral*** con que se obtuvo la uniformidad mensual de la llanta.

De esta forma quedan establecidas las siguientes m=7 variables predictoras con que se realizará el análisis:

X_1	**Cantidad Muestral**
X_2	**Desempeño**
X_3	**Fuerza Radial**
X_4	**Fuerza Lateral**
X_5	**Primera Armónica**
X_6	**Conicidad**
X_7	**Peso**

Este conjunto de variables considera, la calidad, especificación y volumen de producción de las llantas, que son las características con que se desean obtener los conglomerados.

3.3 ANÁLISIS

El *análisis de conglomerados* es un análisis que inicia con una sola población, de la cual se trata de encontrar todas las agrupaciones definidas de objetos (conglomerados) que puedan existir en la población inicial. Las agrupaciones son definidas por el hecho de tener cada una características de distinción muy particulares. En cambio el *análisis discriminante* comienza con los objetos ya clasificados en varios grupos (poblaciones) definidos.

De este modo, la población con la que se inicia el análisis exploratorio es el grupo de las N=41 principales llantas más solicitadas en el mercado, de la categoría carros y camionetas radiales, de la cual se obtendrán los grupos donde puedan ser clasificadas las llantas en base a las características particulares que distingan a cada grupo a partir de las m=7 variables predictoras establecidas en la sección anterior.

Para hacer posible esto, el análisis se divide en 3 etapas, donde la Etapa 1, es la primera aplicación del análisis de conglomerados tomando en cuenta las m=7 variables predictoras establecidas (ver sección anterior), con el fin de tener los conglomerados que de mejor manera distinguen a las llantas, los cuales son los resultados base para llevar a cabo el análisis.

La Etapa 2, es la aplicación del *análisis discriminante predictivo*, que es la parte del análisis donde se obtienen las bases para aumentar el grado de precisión de la regla de clasificación del *análisis de conglomerados*. Esto mediante la selección de las variables predictoras que más y mejor influyen en el objetivo de la regla de clasificación, por medio del *análisis discriminante predictivo*.

En la Etapa 3, se aplica por segunda vez el *análisis de conglomerados*, a las mismas N=41 llantas (objetos), pero en esta ocasión con una regla de clasificación con alto grado de precisión; al considerar el subconjunto de variables que se obtuvo en la etapa anterior.

3.3.1 ETAPA 1: PRIMER ANÁLISIS DE CONGLOMERADOS

Este primer *análisis de conglomerados*, en el que se toman a las m=7 variables predictoras establecidas, se hace con el fin de tener unos resultados base (ver sección anterior), los cuales serán perfeccionados en la tercera etapa (considerando los resultados de la etapa 2).

Se tienen N=41 objetos (principales llantas productivas) que son medidos con las m=7 variables predictoras. Las distintas mediciones que se tienen de los N=41 objetos se tienen hechas en distintas unidades, la variable X_1 *Cantidad Muestral* es medida en unidades (cantidades de llantas), las variables X_2 *Desempeño*, X_3 *Fuerza Radial*, X_4 *Fuerza Lateral*, X_5 *Primera Armónica* y X_6 *Conicidad*, son medidas en porcentajes de llantas que respetan los límites de cada variable, en cambio la variable X_7 *Peso* tiene como unidad el peso en kilogramos de cada llanta, ocasionando ésto que los datos estén a diferentes escalas. Con el fin de tener un mejor manejo de los datos, se llevó a cabo la **estandarización** de los datos, el cual es un cálculo para transformar o re-expresar todas las variables en una misma escala.

Todos los datos son del mismo tipo, que es el cuantitativo, así como también hay variables muy correlacionadas entre si, las cuales son X_3 *Fuerza Radial* y X_5 *Primera Armónica* por ser X_5 consecuencia de X_3 (ver sección 3.2). Por estas razones se seleccionó una medida de disimilaridad exclusiva para datos cuantitativos que tome en forma simultánea a todas las variables (es decir, si los dos objetos son muy similares lo son con respecto a todas las variables), esta medida es la **Distancia euclidiana al cuadrado**, que es la más adecuada para datos cuantitativos correlacionados; de la cual su expresión matemática es la siguiente

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

donde d_{ij} es la disimilaridad o distancia que hay entre el objeto i y el objeto j

m es el número de variables predictoras

x_{ik} es la medición de la k-ésima variable predictora en el objeto i.

Por ser este tipo de análisis del tipo exploratorio, es decir, un análisis con el que se inicia en blanco con respecto a conocimientos del tipo de comportamiento que influye entre los datos, y de esta manera se pretende descubrir el comportamiento de la información, es necesario hacer la comparación de varias conglomeraciones con el mismo grupo de datos, ya que no existe un método de conglomeración óptimo con el que se puedan obtener los conglomerados mejor definidos para el grupo de objetos que se tiene.

De esa manera se realizaron cinco análisis de conglomerados por medio del paquete estadístico *SPSS* (Statistical Package for Social Sciences) *Base 7.5 for Windows*, con los mismos datos estandarizados y con la misma medida de disimilaridad (Distancia Euclidiana al cuadrado), pero con distintos métodos de conglomeración jerárquicos, los cuales fueron:

- Método de Ligaje Simple
- Método de Ligaje Completo
- Método de Centroides
- Método de Mediana
- Método de Ward

Se seleccionaron métodos de conglomeración jerárquicos por no tener una solución base, es decir, no tener un número específico de conglomerados adecuado con que se desee contar.

Se compararon los resultados de cada método, inspeccionando sus procedimientos de conglomeración mediante los dendogramas (ver anexos 3-6) y con criterios de expertos en calidad de llantas; de todos los resultados que se obtuvieron de los cinco análisis, los resultados del **Método de Ward** son los que mostraron, con mayor claridad y sentido, la definición de cada conglomerado final en que quedaron clasificadas las llantas, ésto por ser el *método de Ward* un método que minimiza las pérdidas de información de las

mediciones de las variables de calidad y producción de cada llanta. Y así obtener los conglomerados con la mayor información posible sobre la calidad y producción de las llantas.

La Tabla 3 es el **Historial de Conglomeración**, que muestra en forma resumida los resultados de cada etapa del procedimiento de conglomeración del *método de Ward.* mostrando para cada etapa del procedimiento el tipo de combinación (conjunción) que sucedió en la misma, la cuál puede ser la unión de un objeto con otro objeto, un objeto con un conglomerado o un conglomerado con otro conglomerado. La tabla muestra el coeficiente de cada etapa con el que se hizo la combinación y de esa manera se sabe qué tan homogéneos son los conglomerados que se combinan. Los coeficientes pequeños indican que los conglomerados que se juntaron son completamente homogéneos, mientras que los coeficientes grandes indican que los miembros de los conglomerados son muy distantes (disimilares).

Como se ve en el historial de conglomeración (Tabla 3), en la etapa 1, el objeto 20 (llanta "20") es combinado con el objeto 26 (llanta "26") con un coeficiente de 0.027 y de esa manera el conglomerado resultante es (20, 26), se señala que la próxima etapa en la que vuelve a tomar parte de una combinación es en la etapa 12, donde el conglomerado (20, 26) se combina con el objeto 21 resultando el conglomerado (20, 21, 26) el cual en la etapa 25 se combina con el conglomerado (13, 14, 18, 23, 24) con un coeficiente de 12.831.

En la etapa final del historial de conglomeración, al ver la diferencia de los coeficientes entre dos etapas adyacentes, si hay un salto brusco en el tamaño de su diferencia, se debe considerar que una solución es alcanzada. Así el coeficiente de la etapa 40 indica una solución de 2 conglomerados; se puede ver que la diferencia de los coeficientes de las etapas 37 y 36 es grande, y esto hace considerar una solución de 5 o 6 conglomerados respectivamente.

De acuerdo a lo indicado por la tabla 3, se consideró un rango de soluciones de 3 a 6 conglomerados, esto por tener los coeficientes de las etapas 36, 37, 38 y 39, las diferencias más grandes entre las etapas adyacentes de todos los coeficientes del

procedimiento de conglomeración.

La Tabla 4 muestra los miembros resultantes para las soluciones tomadas, que son de la solución de 3 hasta la solución de 6 conglomerados. Esta tabla es llamada **Lista de Miembros** donde la primera columna muestra los objetos (llantas) junto con el número que son identificados, en la siguiente columna se muestra la lista de miembros para la solución de 6 conglomerados y después es mostrada la lista de miembros para las soluciones de 5, 4 y 3 conglomerados. En cada una de las listas de las soluciones se tiene identificado el conglomerado al que cada objeto pertenece.

Una gran ayuda visual para saber qué tan homogéneos son los conglomerados resultantes, es el **dendograma** (Figura 3.5) o *Arbol de Conglomerados*, este diagrama es una buena imagen del historial de conglomeración, pero con la desventaja de que no muestra los coeficientes con que se combinan los objetos y los conglomerados, ya que su escala es diferente a la que tienen los coeficientes que muestra el historial de conglomeración. El dendograma que se muestra (Figura 3.5) es el resultante del procedimiento de conglomeracion, método de Ward.

El *dendograma* muestra claramente la formación de los conglomerados, y su principal ventaja es el saber qué tan homogéneos y heterogéneos son los conglomerados que forman la solución con el método de conglomeración seleccionado. Pero las desventajas del dendograma son el no poder saber con que coeficientes se combinaron los conglomerados, ni saber su orden de formación.

Historial de conglomeración

Etapa	Conglomerado que se combina 1	Conglomerado que se combina 2	Coeficientes	Etapa en la que el conglomerado aparece por primera vez 1	Etapa en la que el conglomerado aparece por primera vez 2	Próxima etapa
1	20	26	0.027	0	0	12
2	35	36	0.120	0	0	22
3	13	14	0.235	0	0	20
4	32	33	0.369	0	0	7
5	22	28	0.511	0	0	14
6	18	24	0.673	0	0	10
7	31	32	0.839	0	4	17
8	38	40	1.031	0	0	30
9	12	16	1.254	0	0	19
10	18	23	1.502	6	0	20
11	37	39	1.752	0	0	35
12	20	21	2.074	1	0	25
13	29	30	2.423	0	0	31
14	22	27	2.934	5	0	27
15	2	4	3.465	0	0	26
16	1	7	4.021	0	0	23
17	31	34	4.582	7	0	22
18	11	19	5.246	0	0	29
19	12	15	5.912	9	0	29
20	13	18	6.602	3	10	25
21	17	25	7.622	0	0	27
22	31	35	8.805	17	2	30
23	1	10	10.004	16	0	37
24	3	9	11.274	0	0	28
25	13	20	12.831	20	12	36
26	2	5	14.609	15	0	32
27	17	22	16.400	21	14	34
28	3	6	18.232	24	0	32
29	11	12	20.105	18	19	34
30	31	38	22.040	22	8	31
31	29	31	25.373	13	30	35
32	2	3	29.131	26	28	33
33	2	8	34.383	32	0	37
34	11	17	39.907	29	27	36
35	29	37	47.403	31	11	38
36	11	13	56.817	34	25	39
37	1	2	76.215	23	33	39
38	29	41	100.629	35	0	40
39	1	11	159.007	37	36	40
40	1	29	280.000	39	38	0

Tabla 3. Historial de Conglomeración del Método de Ward.

Conglomerado de pertenencia

		Número de soluciones			
		6	5	4	3
Objeto	" 1 "	1	1	1	1
	" 2 "	2	2	1	1
	" 3 "	2	2	1	1
	" 4 "	2	2	1	1
	" 5 "	2	2	1	1
	" 6 "	2	2	1	1
	" 7 "	1	1	1	1
	" 8 "	2	2	1	1
	" 9 "	2	2	1	1
	"10"	1	1	1	1
	"11"	3	3	2	2
	"12"	3	3	2	2
	"13"	4	3	2	2
	"14"	4	3	2	2
	"15"	3	3	2	2
	"16"	3	3	2	2
	"17"	3	3	2	2
	"18"	4	3	2	2
	"19"	3	3	2	2
	"20"	4	3	2	2
	"21"	4	3	2	2
	"22"	3	3	2	2
	"23"	4	3	2	2
	"24"	4	3	2	2
	"25"	3	3	2	2
	"26"	4	3	2	2
	"27"	3	3	2	2
	"28"	3	3	2	2
	"29"	5	4	3	3
	"30"	5	4	3	3
	"31"	5	4	3	3
	"32"	5	4	3	3
	"33"	5	4	3	3
	"34"	5	4	3	3
	"35"	5	4	3	3
	"36"	5	4	3	3
	"37"	5	4	3	3
	"38"	5	4	3	3
	"39"	5	4	3	3
	"40"	5	4	3	3
	"41"	6	5	4	3

Tabla 4. Lista de Miembros.

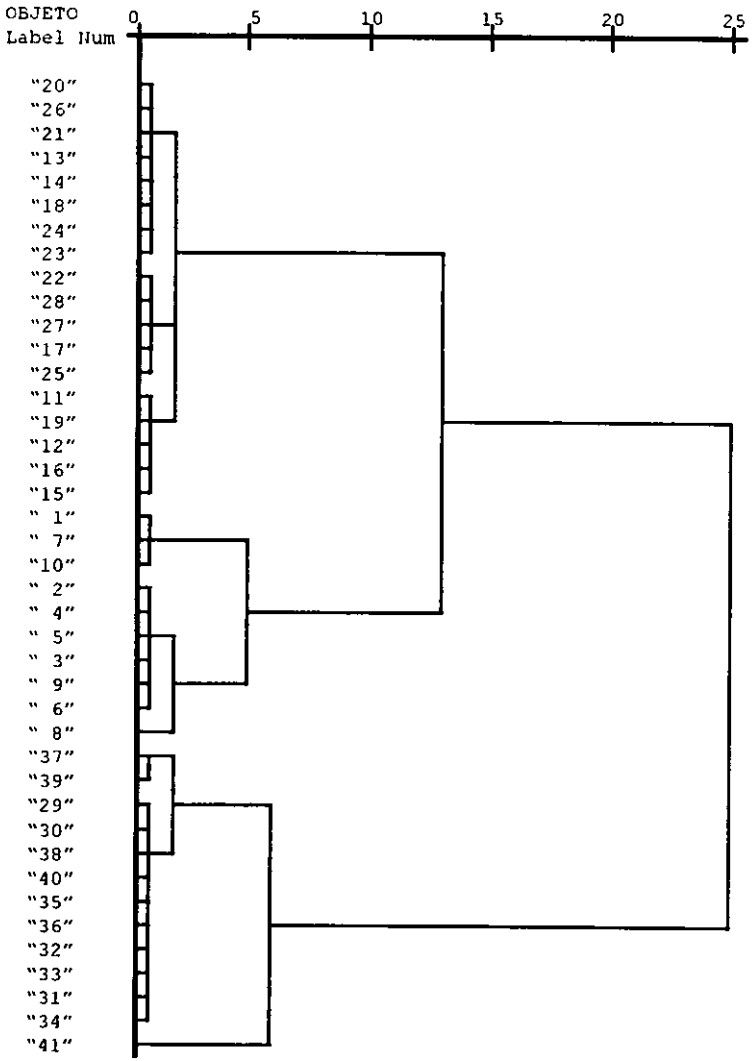


Figura 3.5. Dendograma usando el Método de Ward.

El análisis proporcionó las soluciones de 3 hasta 6 conglomerados, y de acuerdo a opiniones y puntos de vistas de personal experto en llantas, se tomó la solución de k=6 conglomerados por haber sido la que mejor distingue a cada conglomerado de acuerdo a las llantas clasificadas en ellos, y los 6 conglomerados resultantes son presentados en la

Tabla 5.

GRUPO 1	GRUPO 2	GRUPO 3	GRUPO 4	GRUPO 5	GRUPO 6
"1"	"2"	"11"	"13"	"29"	"41"
"7"	"3"	"12"	"14"	"30"	
"10"	"4"	"15"	"18"	"31"	
	"5"	"16"	"20"	"32"	
	"6"	"17"	"21"	"33"	
	"8"	"19"	"23"	"34"	
	"9"	"22"	"24"	"35"	
		"25"	"26"	"36"	
		"27"		"37"	
		"28"		"38"	
				"39"	
				"40"	

Tabla 5. Lista de miembros de los conglomerados resultantes de la etapa 1.

Y los estadisticos descriptivos de cada grupo (conglomerado) son presentados en la Tabla 6.

Estadísticos del grupo (Ward Method)

		Media	Desv. Est.	No. de Miembros
1	X1: C. Muestral	5023.3	4380.1	3
	X2: Desempeño	2.31	1.47	3
	X3: Fuerza Radial	3.34	1.82	3
	X4: Fuerza Lateral	82.62	9.44	3
	X5: 1ra. Armónica	19.44	10.30	3
	X6: Conicidad	92.71	5.36	3
	X7: Peso	18.93	0.54	3
2	X1: C. Muestral	3782.7	4960.0	7
	X2: Desempeño	3.30	2.34	7
	X3: Fuerza Radial	9.45	5.56	7
	X4: Fuerza Lateral	49.36	14.80	7
	X5: 1ra. Armónica	26.50	7.93	7
	X6: Conicidad	64.18	12.38	7
	X7: Peso	15.35	2.58	7
3	X1: C. Muestral	2792.0	2427.1	10
	X2: Desempeño	14.28	4.23	10
	X3: Fuerza Radial	21.19	8.50	10
	X4: Fuerza Lateral	88.04	7.89	10
	X5: 1ra. Armónica	37.97	11.55	10
	X6: Conicidad	83.78	8.38	10
	X7: Peso	10.91	1.95	10
4	X1: C. Muestral	3508.3	2475.0	8
	X2: Desempeño	26.85	4.75	8
	X3: Fuerza Radial	32.05	6.14	8
	X4: Fuerza Lateral	96.97	3.40	8
	X5: 1ra. Armónica	54.51	6.95	8
	X6: Conicidad	93.37	2.45	8
	X7: Peso	9.69	1.27	8
5	X1: C. Muestral	8915.42	9247.2	12
	X2: Desempeño	54.05	6.44	12
	X3: Fuerza Radial	71.53	8.97	12
	X4: Fuerza Lateral	95.02	4.39	12
	X5: 1ra. Armónica	69.31	7.61	12
	X6: Conicidad	90.25	7.31	12
	X7: Peso	6.87	0.80	12
6	X1: C. Muestral	67735.0	.	1
	X2: Desempeño	54.20	.	1
	X3: Fuerza Radial	63.40	.	1
	X4: Fuerza Lateral	98.20	.	1
	X5: 1ra. Armónica	69.40	.	1
	X6: Conicidad	94.50	.	1
	X7: Peso	8.41	.	1

Tabla 6. Estadísticos de los grupos de la etapa 1.

La Tabla 6, nos muestra los principales indicadores estadísticos, (media y desviación

estándar) para saber qué comportamiento es el que presentan las variables en cada uno de los 6 diferentes grupos. Y de esa forma saber las diferencias que existen en los grupos. Los cuales son identificados de la siguiente manera:

GRUPO 1: Grupo de *llantas grandes*, por que contiene a las llantas de mayor tamaño y peso.

GRUPO 2: Grupo de *llantas complejas*, por mostrar un bajo porcentaje en uniformidad. ya que su procedimiento de desarrollo es complejo.

GRUPO 3: Grupo de *llantas indecisas*, ya que muestran porcentajes bajos y altos en los parámetros de uniformidad.

GRUPO 4: Grupo *de llantas sencillas*, por contener llantas que son muy simples en su procedimiento de construcción y de esa manera tener alto porcentaje en uniformidad.

GRUPO 5: Grupo de *llantas productivas*, ya que la mayoría de las llantas que contiene son de un volumen grande en producción y de un porcentaje alto en uniformidad.

GRUPO 6: Grupo de *llanta especial*, por representar esta llanta el 40% de la producción actual de la planta, y ser la más productiva y experimentada.

Pero ¿es buena la clasificación de las llantas en los distintos grupos? y ¿son necesarias las 7 variables predictoras para una clasificación correcta de las llantas?

3.3.2 ETAPA 2: ANÁLISIS DISCRIMINANTE PREDICTIVO

En esta etapa se obtendrán las bases que elevaran el grado de precisión de la regla de clasificación del *análisis de conglomerados*, con el fin de perfeccionar la clasificación de las llantas en los grupos que muestren una mejor definición de los diferentes procedimientos de desarrollo y producción de las llantas. En la etapa anterior se realizó el *análisis de conglomerados* considerando al conjunto completo de las m=7 variables predictoras establecidas, pero debido al tiempo y costo en que se lleva tomar cada variable surge la pregunta, ¿es necesario considerar a todas las variables?. Si la regla de clasificación es hecha con las variables que realmente tienen sentido para el objetivo del análisis, entonces su grado de precisión es alto, es por eso que en ocasiones no es necesario realizar el análisis con todas las variables predictoras que se disponen, por que puede ser que no todas influyan de la manera adecuada en la regla de clasificación, y así habrá variables innecesarias para la clasificación de los objetos, y por lo tanto afectan el grado de precisión de la regla de clasificación. Entonces, la regla de clasificación tendrá una mayor exactitud en su función si es creada con las variables predictoras que en verdad tienen peso en la clasificación de los objetos, y de esa manera lo correcto será no tomar en cuenta las variables innecesarias para el objetivo.

Pero, ¿cuáles son las variables que mejor influyen en la regla de clasificación del *análisis de conglomerados*?, la mejor manera de obtener la respuesta es mediante el *análisis discriminante predictivo* (ADP), con el cual se obtendrá (si existe) el subconjunto de variables que mejor influyen y definen a los conglomerados en los que pueden ser diferenciados los procedimientos de desarrollo y producción de las llantas.

Por medio de la etapa anterior (primer análisis de conglomerados) se obtuvo la variable de agrupación, que divide a la población en k=6 conglomerados (subpoblaciones), y de esta manera se tienen las condiciones necesarias y suficientes para ser aplicado el análisis discriminante predictivo (ADP).

El ADP tiene varios supuestos para los datos con que se realizará el análisis, esto por el hecho de que sus reglas de clasificación son creadas en base a principios de la distribución normal multivariada.

Los supuestos son:

- Las observaciones de las variables predictoras provienen de una población con una distribución normal multivariada.
- Las matrices de covarianza intra-grupos deben ser iguales.

Se asume que la pertenencia de los objetos a los grupos sea:

- Exclusiva (es decir, ningún objeto pertenece a más de un grupo).
- Exhaustiva de modo colectivo (es decir, todos los objetos son miembros de un grupo).

Es necesario saber si para el análisis se respetan las suposiciones anteriormente señaladas, por ser puntos importantes para su diseño.

Con respecto al supuesto de que todas las variables predictoras deben tener una distribución normal multivariada en el análisis, en este estudio no todas las m=7 variables predictoras pueden tener una distribución normal multivariada, ya que las variables predictoras, X_1 *Cantidad Muestral* y X_7 *Peso*, tienen la función de diferenciar y especificar a cada llanta de las 40 llantas restantes, es decir, funcionan como variables categóricas por el hecho de diferenciar a las llantas por medio de su nivel de producción y tamaño respectivamente.

Pero las 5 variables restantes (X_2 *Desempeño*, X_3 *Fuerza Radial*, X_4 *Fuerza Lateral*, X_5 *Primera Armónica* y X_6 *Conicidad*) que son las variables que indican calidad, son las variables predictoras que puede ser que respeten el supuesto de la distribución normal multivariada, para saberlo se realizaron varias pruebas de hipótesis, y así saber si se tiene la suficiente evidencia de que las variables de calidad provienen de una población con distribución normal multivariada.

Se utilizó el paquete estadístico *SPSS* (*Statistical Package for Social Sciences*) Base *7.5 for Windows*, para elaborar las pruebas de normalidad, y el paquete estadístico

STATGRAPHICS (*Statistical Graphics System*) *versión 7 DOS*, para la obtención de la ayuda visual de los histogramas de frecuencia con la curva normal.

En la tabla 7 se muestran los resultados de las pruebas de hipótesis de normalidad, que fueron las pruebas de *Kolmogorov-Smirnov* y la de *Shapiro-Wilk* (para las muestras con 50 o menos observaciones) como alternativas más poderosas a la prueba de bondad de ajuste ji-cuadrada, con un nivel de significancia de 0.05. La hipótesis nula a probar es

H_0 : Los datos no constituyen una muestra aleatoria de una población con distribución normal multivariada.

Pruebas de normalidad

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
X_2 Desempeño	0.151	41	0.019	0.877	41	0.010
X_3 Fuerza Radial	0.177	41	0.002	0.886	41	0.010
X_4 Fuerza Lateral	0 242	41	0.000	0.750	41	0.010
X_5 1ra. Armónica	0 129	41	0 086	0.946	41	0.075
X_6 Conicidad	0.191	41	0.001	0.864	41	0.010

** Este es un límite superior de la significación verdadera.
a Corrección de la significación de Lilliefors

Tabla 7. Pruebas de Normalidad de Kolmogorov-Smirnov y Shapiro-Wilk.

De acuerdo al criterio de rechazo[5] de H_0, la variable X_5 *1ra. Armónica* es la única que tiene la suficiente evidencia para decir que posee una distribución normal multivariada. Lo cual puede confirmarse al ver los histogramas de frecuencia de la figura 3.6, en los que se nota claramente que la variable predictora que más se apega a la campana de la distribución normal, es la variable predictora X_5 *1ra. Armónica*.

El otro supuesto a probar, es si las matrices de covarianza intra-grupos son iguales en los k=6 grupos. Esta comprobación se realizó con la prueba de ***M-Box***[6], el cual es un contraste sobre la igualdad de las matrices de covarianza de los grupos. La prueba se

[5] Si la significancia de Kolmogorov-Smirnov y Shapiro-Wilk es > 0.05 se rechaza H_0.
SPSS® Base 7.5 Applications Guide, pág. 53, 200.

[6] El estadístico de M-Box puede ser afectado por el tamaño de los 6 grupos (si difieren mucho) o cuando la suposición de la normalidad multivariada es violada.
SPSS® Base 7.5 Applications Guide, pág. 231-232

realizó con el *SPSS Base 7.5 for Windows*.

En la tabla 8, se muestran los resultados de la prueba de *M-Box*, en donde los valores de los *Logaritmos de los determinantes* proveen una indicación de qué matrices de covarianza son más diferentes. De acuerdo a los estadísticos descriptivos presentados en la tabla 6, las dispersiones (desv. estándar) de los grupos 2, 3, 4 y 5 son muy similares, por esa razon los logaritmos de los determinantes de sus respectivas matrices de covarianza tienen unos valores similares.

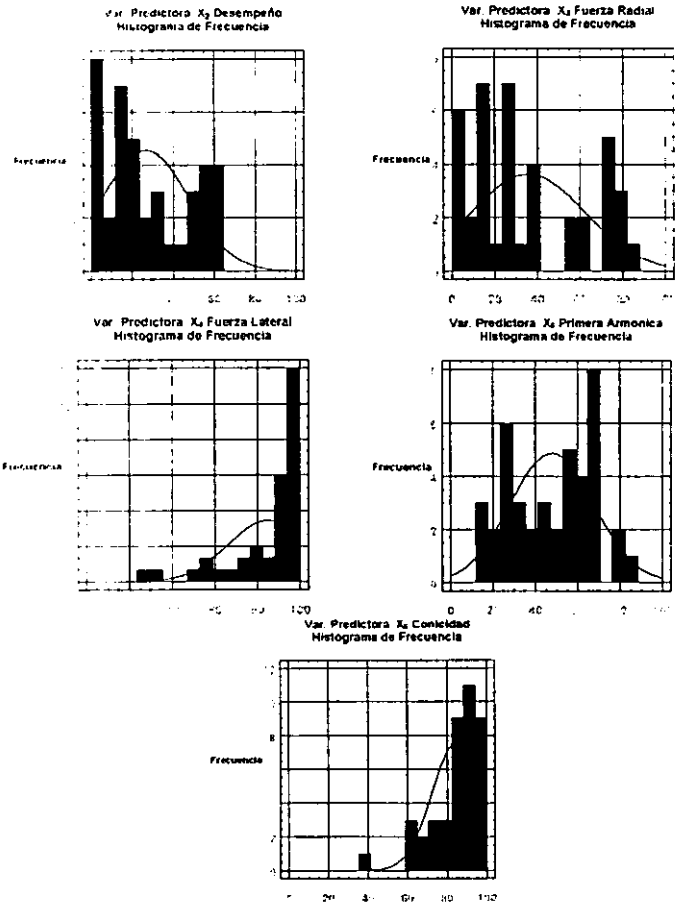


Figura 3.6. Histogramas de Frecuencia de las variables predictoras.

Logaritmo de los determinantes

		Rango	Logaritmo del determinante
Ward Method	1	a	.b
	2	4	24.142021
	3	4	23.517314
	4	4	19.470656
	5	4	23.769656
	6	.c	.b
	Intra-grupos combinada	4	25.520329

Los rangos y logaritmos naturales de los determinantes impresos son los de las matrices de covarianza de los grupos.

a Rango < 3

b Muy pocos casos para no ser singular

c Rango < 1

Resultados de la prueba

M de Box		95.66905
F	Aprox.	2.410861
	gl1	30
	gl2	2131.666
	Sig.	2.97E-05

Contrasta la hipótesis nula de que las matrices de covarianza poblacionales son iguales.

a. Algunas matrices de covarianza son singulares y el procedimiento ordinario no es válido.

Los grupos no singulares se probarán sobre sus propias matrices de covarianza intra-grupo combinadas

El logaritmo de su determinante es 25.756.

Tabla 8. Prueba de M-Box.

Si el nivel de significancia del estadístico **M - Box** es muy cercano o igual a cero entonces, se rechaza la hipótesis nula de que las matrices de covarianza son iguales. En este caso es rechazada la hipótesis nula, por ser el nivel de significancia menor a 0.0005, el cual es un valor muy cercano a cero, indicando que no se tiene la suficiente evidencia de que las k=6 matrices covarianza intra-grupos sean iguales.

De esa manera con el ADP se obtuvieron las reglas de clasificación que respetan las condiciones de que las matrices de covarianza son diferentes y que los tamaños de los k=6 grupos no son iguales.

El método de selección de variables *Stepwise* (ver sección 2.9.2) y la *Regla Cuadrática* (ver Tabla 1) para matrices de covarianza y tamaños de grupos diferentes, fueron los controladores de entradas y salidas de variables del paquete estadístico *SPSS Base 7.5 for Windows*, y así se obtuvo como resultado el subconjunto de las variables predictoras

que más y mejor influyen en la *regla de clasificación* del análisis de conglomerados.

El método Stepwise se basa en el poder discriminante de las variables predictoras, combinandolas hasta formar el grupo con el mayor poder discriminante (ver anexos 7 y 8), es así como se obtiene el subconjunto de variables predictoras que más y mejor influyen en la regla de clasificación del análisis de conglomerados, y de esa manera aumentando su grado de precisión. El subconjunto es formado por las siguientes h=4 variables predictoras:

X_1: Cantidad Muestral

X_2: Desempeño

X_4: Fuerza Lateral

X_7: Peso

El subconjunto de variables predictoras indica que las variables de calidad X_2: Desempeño y X_4: Fuerza Lateral son las necesarias para mostrar el factor de calidad en los conglomerados, ya que la variable X_2 es la variable global que toma en cuenta en forma simultánea los cuatro parámetros de uniformidad, y la variable X_4 es parte del subconjunto por ser una condición muy particular en las llantas (ver sección 3.2). En cambio las variables X_1 y X_7 forman parte del subconjunto por ser las variables predictoras que dan indicación sobre el volumen de producción y especificación de las llantas respectivamente. Obviamente estas dos últimas variables son parte del subconjunto por tener la función de diferenciar las llantas que componen a cada conglomerado mediante su volumen de producción y tamaño.

Así el número de variables predictoras que mejor influyen en la regla de clasificación es menor (h=4) al número de variables predictoras con el que se inició (m=7) el análisis. De esa manera estas variables predictoras, disminuirán el sesgo del *hit rate* de la regla de clasificación del análisis de conglomerados.

3.3.3 ETAPA 3: SEGUNDO ANÁLISIS DE CONGLOMERADOS

Esta etapa es el último paso del análisis y es donde se acompleta el objetivo principal del trabajo. En esta sección es donde se sabrá el tipo de relación que existe entre el análisis de conglomerados y el análisis discriminante predictivo, y de esta manera saber si los dos métodos son complementarios uno del otro o si no tienen nada que ver entre si.

En la etapa 1, se realizó el análisis de conglomerados considerando las 7 variables predictoras establecidas para el análisis,

X_1: Cantidad Muestral
X_2: Desempeño
X_3: Fuerza Radial
X_4: Fuerza Lateral
X_5: Primera Armónica
X_6: Conicidad
X_7: Peso

En esta sección es realizada una segunda aplicación del análisis de conglomerados, considerando como variables predictoras el subconjunto de variables que el análisis discriminante predictivo (etapa 2) propone como las variables que aumentan el grado de precisión de la regla de clasificación del análisis de conglomerados.

Las variables predictoras que componen el subgrupo resultante del análisis discriminante predictivo son:

X_1: Cantidad Muestral
X_2: Desempeño
X_4: Fuerza Lateral
X_7: Peso

Esta segunda aplicación del análisis de conglomerados, es una comprobación para saber si en realidad el subconjunto de las variables predictoras propuestas por el análisis

discriminante predictivo, mejoran la clasificación de las llantas en los distintos grupos que mejor diferencien los procedimientos de producción y desarrollo de éstas mismas, y de esa manera tener un mejor control de su calidad.

Para que se pueda saber si en realidad el subconjunto de las h=4 variables predictoras aumenta el grado de precisión de la regla de clasificación del análisis de conglomerados, se debe contar con las mismas condiciones con que se aplicó por primera vez, es decir, es necesario que el diseño del análisis de conglomerados sea el mismo con el que se realizó el de la etapa 1. El diseño del análisis debe ser el mismo en las dos etapas (etapa 1 y 3), el cual tiene como principales factores de diseño los que a continuación se muestran:

- Los datos son estandarizados con respecto a las variables.
- Método de Ward, como procedimiento de conglomeración.
- Distancia euclidiana cuadrada, como medida de disimilaridad.
- Rango de soluciones de 3 a 6 conglomerados.

La tabla 9, es el *Historial de Conglomeración* desarrollado mediante el método de Ward, al considerar las h=4 variables predictoras mencionadas anteriormente. Como se explicó en la sección 3.3.1, el historial de conglomeración indica que debe ser considerada una solución de 6 conglomerados, por tener la etapa 36 del historial el último salto brusco en las diferencias entre los coeficientes de etapas adyacentes.

En el dendograma (Figura 3.7) se ve claramente una mejor distinción de los k=6 conglomerados propuestos como solución, distinguiéndose con una mejor claridad la cantidad de conglomerados señalada mediante la tabla 9, en cambio en el dendograma de la etapa 1 (Figura 3.5) no se tiene una clara distinción del número de conglomerados que el historial de conglomeración respectivo ofrece como solución.

Así para poder saber si el subconjunto de las h=4 variables predictoras, aumenta el grado de precisión de la regla de clasificación del análisis de conglomerados, es

necesario obtener el mismo número de conglomerados (k=6) que en la etapa 1 se obtuvieron.

En la tabla 10 se muestran los k=6 conglomerados que se obtienen al considerar como variables predictoras el subconjunto de variables predictoras que aumentan el grado de precisión de la regla de clasificación del análisis de conglomerados.

Historial de conglomeración

Etapa	Conglomerado que se combina 1	Conglomerado que se combina 2	Coeficientes	Etapa en la que el conglomerado aparece por primera vez 1	Etapa en la que el conglomerado aparece por primera vez 2	Próxim etapa
1	20	26	0.005	0	0	10
2	22	28	0.032	0	0	16
3	33	36	0.087	0	0	8
4	30	31	0.155	0	0	15
5	38	40	0.225	0	0	25
6	16	27	0.300	0	0	20
7	1	7	0.375	0	0	29
8	33	34	0.458	3	0	13
9	13	14	0.544	0	0	20
10	20	21	0.631	1	0	27
11	18	23	0.724	0	0	14
12	12	15	0.839	0	0	22
13	33	35	0.954	8	0	28
14	18	24	1.076	11	0	25
15	30	32	1.216	4	0	24
16	22	25	1.383	2	0	26
17	37	39	1.552	0	0	36
18	6	9	1.780	0	0	32
19	2	4	2.026	0	0	30
20	13	16	2.298	9	6	22
21	11	19	2.601	0	0	31
22	12	13	2.907	12	20	27
23	3	8	3.222	0	0	32
24	29	30	3.582	0	15	28
25	18	38	4.065	14	5	34
26	17	22	4.561	0	16	31
27	12	20	5.207	22	10	33
28	29	33	5.879	24	13	34
29	1	10	6.652	7	0	37
30	2	5	7.577	19	0	35
31	11	17	9.169	21	26	33
32	3	6	11.277	23	18	35
33	11	12	13.637	31	27	39
34	18	29	16.980	25	28	36
35	2	3	20.651	30	32	37
36	18	37	27.709	34	17	38
37	1	2	36.124	29	35	40
38	18	41	61.263	36	0	39
39	11	18	90.972	33	38	40
40	1	11	160.000	37	39	0

Tabla 9. Historial de Conglomeración, de la Etapa 3

(considerando el subconjunto de h=4 variables predictoras, de la etapa 2).

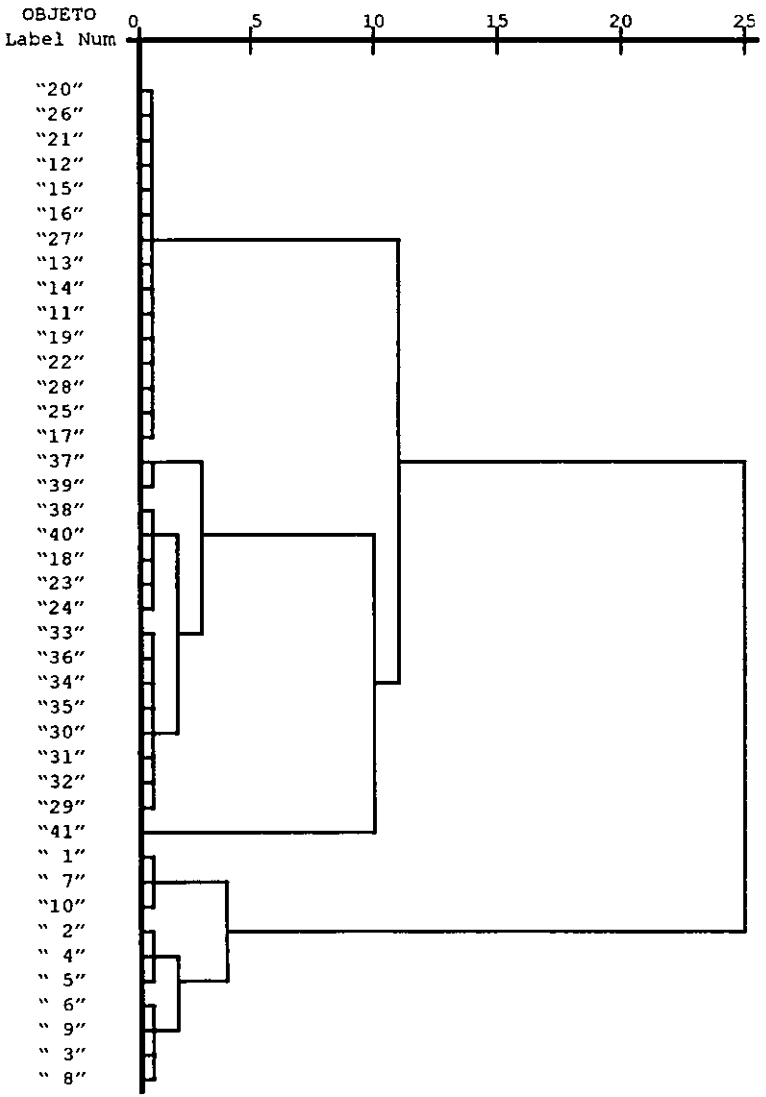


Figura 3.7. Dendograma de la etapa 3

(considerando el subconjunto de h=4 variables, resultante de la etapa 2).

GRUPO 1	GRUPO 2	GRUPO 3	GRUPO 4	GRUPO 5	GRUPO 6
"1"	"2"	"11"	"18"	"37"	"41"
"7"	"3"	"12"	"23"	"39"	
"10"	"4"	"13"	"24"		
	"5"	"14"	"29"		
	"6"	"15"	"30"		
	"8"	"16"	"31"		
	"9"	"17"	"32"		
		"19"	"33"		
		"20"	"34"		
		"21"	"35"		
		"22"	"36"		
		"25"	"38"		
		"26"	"40"		
		"27"			
		"28"			

Tabla 10. Lista de los miembros clasificados en los conglomerados resultantes de la etapa 3.

La tabla 11 muestra la descripción, mediante los estadísticos descriptivos, de los grupos que cambiaron en comparación con los de la etapa 1, los cuales son los grupos 3, 4 y 5:

Estadísticos del grupo (Ward Method)

		Media	Desv. Est.	No. de Miembros
3	X1: C. Muestral	2629.7	2104.2	15
	X2: Desempeño	17.35	5.65	15
	X3: Fuerza Radial	23.40	7.60	15
	X4: Fuerza Lateral	90.63	7.67	15
	X5: 1ra. Armónica	43.91	13.38	15
	X6: Conicidad	87.01	8.39	15
	X7: Peso	10.66	1.75	15

		Media	Desv. Est.	No. de Miembros
4	X1: C. Muestral	5235.7	2704.7	13
	X2: Desempeño	48.34	10.78	13
	X3: Fuerza Radial	64.73	16.76	13
	X4: Fuerza Lateral	95.49	4.42	13
	X5: 1ra. Armónica	65.27	10.62	13
	X6: Conicidad	89.79	6.44	13
	X7: Peso	7.27	1.22	13

5	X1: C. Muestral	27730.5	4349.4	2
	X2: Desempeño	58.75	1.48	2
	X3: Fuerza Radial	67.10	6.51	2
	X4: Fuerza Lateral	97.75	2.62	2
	X5: 1ra. Armónica	70.20	0.42	2
	X6: Conicidad	97.70	1.84	2
	X7: Peso	7.29	0.17	2

Tabla 11. Estadísticos descriptivos de los grupos resultantes de la etapa 3.

De acuerdo a los datos descriptivos de la tabla 11, claramente se nota que las llantas del grupo 5 son productivas, las del grupo 4 son sencillas en el procedimiento de producción y las del grupo 3 muestran altos y bajos porcentajes en los parámetros de uniformidad, y asi los grupos 3, 4 y 5 quedan identificados con la misma nomenclatura que se les dió en la etapa 1 (ver sección 3.3.1).

3.4 CONJUNCIÓN DE RESULTADOS

En la tabla 12, se muestra la clasificación resultante de los miembros en cada uno de los 6 conglomerados que se obtuvieron al considerar las m=7 variables predictoras iniciales (etapa 1), y al considerar el subconjunto de h=4 variables predictoras (etapa 3).

GRUPO 1	
7 var	4 var
"1"	"1"
"7"	"7"
"10"	"10"

GRUPO 2	
7 var	4 var
"2"	"2"
"3"	"3"
"4"	"4"
"5"	"5"
"6"	"6"
"8"	"8"
"9"	"9"

GRUPO 3	
7 var	4 var
"11"	"11"
"12"	"12"
"15"	"13"
"16"	"14"
"17"	"15"
"19"	"16"
"22"	"17"
"25"	"19"
"27"	"20"
"28"	"21"
	"22"
	"25"
	"26"
	"27"
	"28"

GRUPO 4	
7 var	4 var
"13"	"18"
"14"	"23"
"18"	"24"
"20"	"29"
"21"	"30"
"23"	"31"
"24"	"32"
"26"	"33"
	"34"
	"35"
	"36"
	"38"
	"40"

GRUPO 5	
7 var	4 var
"29"	"37"
"30"	"39"
"31"	
"32"	
"33"	
"34"	
"35"	
"36"	
"37"	
"38"	
"39"	
"40"	

GRUPO 6	
7 var	4 var
"41"	"41"

Tabla 12. Comparación de la clasificación de los objetos en los conglomerados resultantes al considerar las m=7 variables (etapa 1) y al considerar el subconjunto de las h=4 variables (etapa 3).

Al comparar los conglomerados resultantes de la etapa 1 y de la etapa 3, los conglomerados 1, 2 y 6 resultan con los mismos objetos en las dos aplicaciones del análisis de conglomerados, mientras que los conglomerados 3, 4 y 5, tienen diferentes objetos clasificados en las dos etapas (1 y 3).

Un caso muy especial es el conglomerado 5, el cual en la etapa 1 contiene 12 objetos, y en la etapa 3 resulta con sólo 2 objetos, estos dos objetos son las llantas "37" y "39". El conglomerado 4 de la etapa 3 contiene los 10 objetos restantes que contenía el conglomerado 5 de la etapa 1. Y el conglomerado 3 de la etapa 3 contiene 5 objetos que el conglomerado 4 de la etapa 1 contenía.

Una gran ayuda visual para saber cuál puede ser el motivo por el que en la etapa 3 se obtiene este cambio en la clasificación de los objetos, es mediante la *gráfica de dispersión* bidimensional de las variables predictoras más significantes (ver anexos 9-

CONCLUSIONES

CONCLUSIONES

Al examinar los principios teóricos matemáticos del análisis de conglomerados se constató que el principio de la distancia, la cual es establecida como la proximidad (similaridad o disimilaridad) entre dos puntos, es la base para determinar las diferentes agrupaciones en las que pueden ser clasificados los objetos. En cambio los principios teórico matemáticos del análisis discriminante predictivo muestran que el principio de la distancia es el factor fundamental de la función de clasificación, la cual decreta la regla de clasificación que predice la población (conglomerado) a la que un objeto podría pertenecer o estar más apegado. Y de acuerdo a la exactitud de lo predicho se determina el subconjunto de variables predictoras que más influyen para alcanzar el objetivo de la regla de clasificación del análisis de conglomerados.

Al detallar cada una de las etapas de los procesos de desarrollo del análisis de conglomerados y del análisis discriminante predictivo, se obtuvo una síntesis de los fundamentos teóricos y metodológicos de ambos métodos, mediante la cual se determinó que el concepto de distancia es el factor que se comparte en los dos métodos, y que explica la razón de obtener una regla de clasificación con alto grado de precisión, sin descartar el grado de exactitud en la proyección de las variables predictoras.

De esta manera, el análisis de conglomerados y el análisis discriminante predictivo, mantienen a la distancia como el factor clave para la clasificación de los objetos en los conglomerados que son logrados a partir de la combinación de las variables predictoras que ofrecen el máximo grado de precisión de la regla de clasificación.

Todo lo anterior se confirmó, al haber obtenido la mejor clasificación y distinción de las N=41 llantas en los k=6 grupos (conglomerados) resultantes del análisis de conglomerados, al considerar el subconjunto de las h=4 variables predictoras proporcionadas por el análisis discriminante predictivo, las cuales mostraron una mejor definición de los conglomerados, en comparación con los que se obtuvieron a partir del

grupo de las m=7 variables predictoras establecidas al inicio del estudio. Lo cual hace concluir con certeza que el análisis discriminante predictivo enriquece el resultado del análisis de conglomerados; determinando así que el análisis discriminante predictivo puede ser tomado como una etapa más para el desarrollo de un análisis de conglomerados de mayor precisión.

Cuando una llanta especifica esté en producción, y se requiera hacer un cambio en su proceso de desarrollo, que afecte de manera significativa a uno o varios de los parámetros de uniformidad, se puede tener la seguridad de que el cambio afectará de la misma manera al proceso de desarrollo de todas y cada una de las llantas que pertenezcan al mismo conglomerado en el que se encuentra clasificada la llanta seleccionada. De esta manera, los conglomerados pueden ser una gran herramienta, para tomar decisiones con el fin de aumentar la calidad en forma simultánea, de las llantas clasificadas en un mismo grupo.

Los objetivos y resultados fueron logrados por medio del software estadístico *SPSS*, que como su nombre lo indica (*Statistical Package for Social Sciences*) es un programa estadístico diseñado para el manejo y análisis de información de aspectos sociales, pero de acuerdo al estudio realizado sobre la calidad de las llantas, se manejó información técnica de la rama industrial, con la que se obtuvieron resultados lo suficientemente apegados a la realidad y lógica de la calidad de las llantas, demostrando así que el *SPSS* no es un programa para un uso exclusivo en ciencias sociales, sino que puede ser muy bien aplicado en la ingeniería y otras ramas de la ciencia, por poseer un gran dominio para el control y manejo de todo tipo de datos, así como un gran número de alternativas para analizarlos.

El estudio conjuntivo fue realizado de manera que el análisis de conglomerados es la base para mostrar la relación existente entre éste y el análisis discriminante predictivo, ya que proporciona los resultados por medio de los cuales, se comprueba que el análisis discriminante predictivo aumenta el grado de precisión de la regla de clasificación del análisis de conglomerados, pero por el hecho de no haber reglas que indiquen la forma en que debe ser realizado el estudio conjuntivo, existen varias maneras con las que se puede demostrar que el análisis discriminante predictivo junto con el análisis de

conglomerados forman un método de clasificación con un alto grado de precisión. Por ejemplo, puede ser tomado como primer paso la aplicación del análisis discriminante predictivo al grupo establecido de variables predictoras, y de esa manera asegurarse si el grupo de variables predictoras es el adecuado para obtener una correcta clasificación de los objetos a tratar, y por lo tanto satisfacer el objetivo del estudio.

Con el resultado del trabajo se puede afirmar, hoy en día, que la estadistica multivariada es una gran herramienta para coadyuvar a encontrar alternativas de solución a los problemas de manufactura, cuya importancia es sustantiva en los procesos de planeación y control de la producción de un determinado bien.

ANEXOS

ANEXO 1

Regla de clasificación lineal

$$L_{ug} = \ln q_g - \tfrac{1}{2} D_{ug}^{*2} = \ln q_g - \tfrac{1}{2}\left(X_u - \bar{X}_g\right)' S^{-1}\left(X_u - \bar{X}_g\right)$$

$$= \ln q_g - \tfrac{1}{2} X_u' S^{-1} X_u + \tfrac{1}{2} X_u' S^{-1} \bar{X}_g + \tfrac{1}{2} \bar{X}_g' S^{-1} X_u - \tfrac{1}{2} \bar{X}_g' S^{-1} \bar{X}_g$$

donde $X_u' S^{-1} X_u$ para un objeto u sería comun en todas las g y puede ser ignorado por propósitos de clasificación. Por ser S simétrica $\Rightarrow$ $X_u' S^{-1} \bar{X}_g = \bar{X}_g' S^{-1} X_u$

$$\Rightarrow \quad L_{ug} = \bar{X}_g' S^{-1} X_u - \tfrac{1}{2} \bar{X}_g' S^{-1} \bar{X}_g + \ln q_g$$

donde $b_g' = \bar{X}_g' S^{-1}$

$$c_g = -\tfrac{1}{2} \bar{X}_g' S^{-1} \bar{X}_g + \ln q_g$$

$$L_{ug} = b_g' X_u + c_g$$

Regla de clasificación cuadrática

$$Q_{ug} = \ln q_g - \tfrac{1}{2} \ln |S_g| - \tfrac{1}{2} D_{ug}^2 = \ln q_g - \tfrac{1}{2} \ln |S_g| - \tfrac{1}{2}\left(X_u - \bar{X}_g\right)' S_g^{-1}\left(X_u - \bar{X}_g\right)$$

$$= \ln q_g - \tfrac{1}{2} \ln |S_g| - \tfrac{1}{2} X_u' S_g^{-1} X_u + \tfrac{1}{2} X_u' S_g^{-1} \bar{X}_g + \tfrac{1}{2} \bar{X}_g' S_g^{-1} X_u - \tfrac{1}{2} \bar{X}_g' S_g^{-1} \bar{X}_g$$

donde $X_u' S_g^{-1} X_u$ es la forma cuadrática[1] de X_u

Por ser S_g simétrica $\Rightarrow$ $X_u' S_g^{-1} \bar{X}_g = \bar{X}_g' S_g^{-1} X_u$

$$= \ln q_g - \tfrac{1}{2} \ln |S_g| - \tfrac{1}{2} X_u' S_g^{-1} X_u + \bar{X}_g' S_g^{-1} X_u - \tfrac{1}{2} \bar{X}_g' S_g^{-1} \bar{X}_g$$

Tomando a la forma cuadrática como X_u^2, entonces Q_{ug} en términos de X_u puede ser escrita como una ecuación cuadrática

donde $a = -\tfrac{1}{2}$

$$b_g' = \bar{X}_g' S_g^{-1}$$

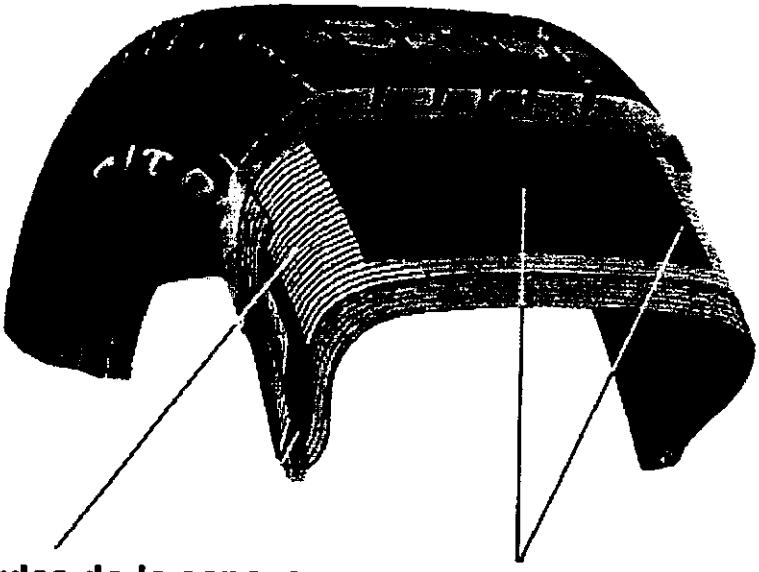
$$c_g = \ln q_g - \tfrac{1}{2} \ln |S_g| - \tfrac{1}{2} \bar{X}_g' S_g^{-1} \bar{X}_g$$

$$Q_{ug} = a X_u' S_g^{-1} X_u + b_g' X_u + c_g = a X_u^2 + b_g' X_u + c_g$$

[1] MATRICES. Frank Ayres Jr. p.131

ANEXO 2

Corte representativo de una *Llanta Radial*

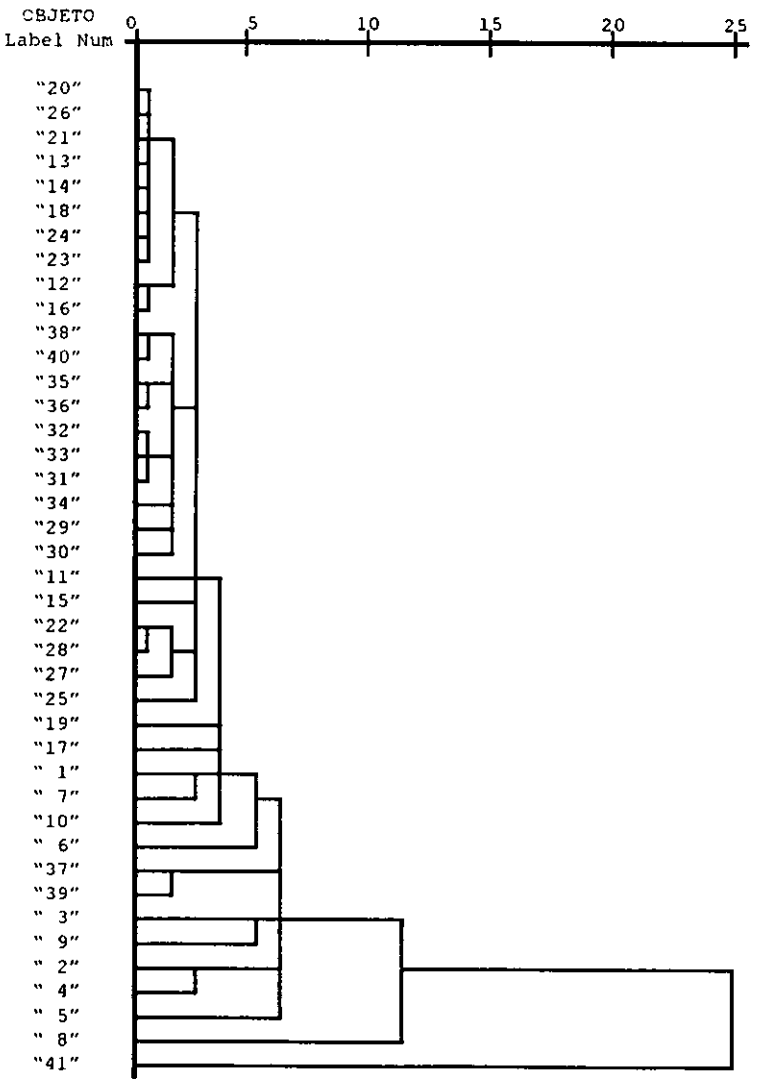


Cuerdas de la capa, con la misma dirección del radio.

BREAKERS (cinturones de acero)

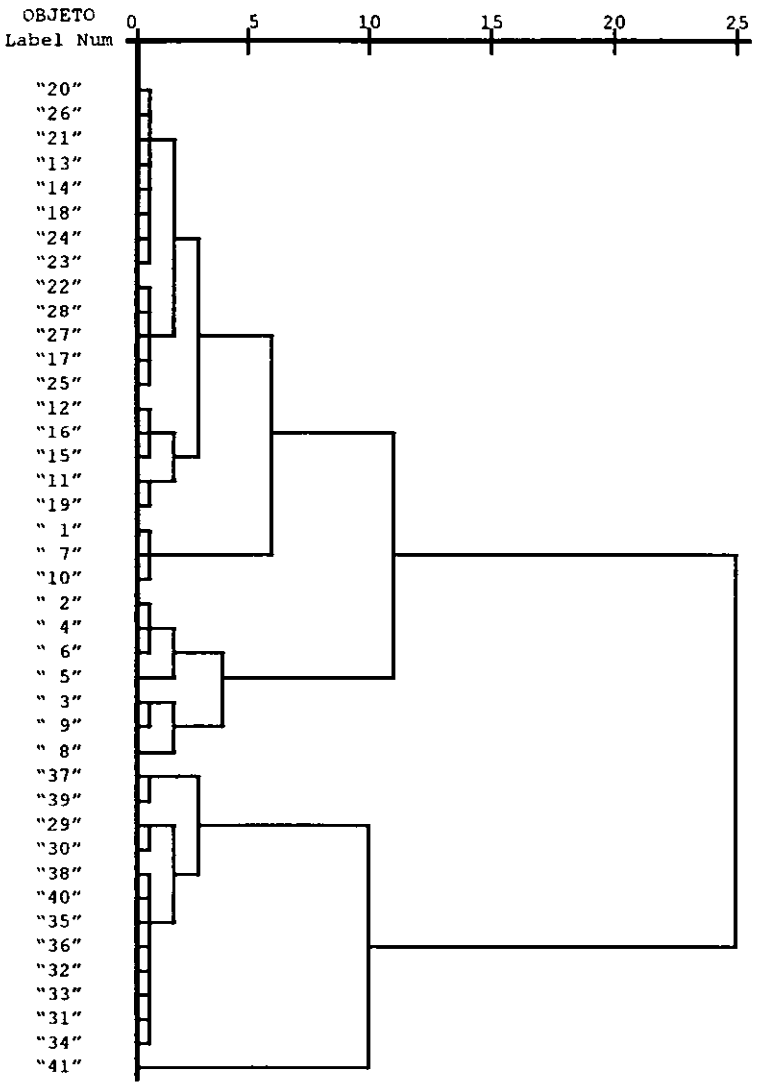
ANEXO 3

Dendograma resultante del Método de Ligaje Simple considerando las *m*=7 variables predictoras iniciales



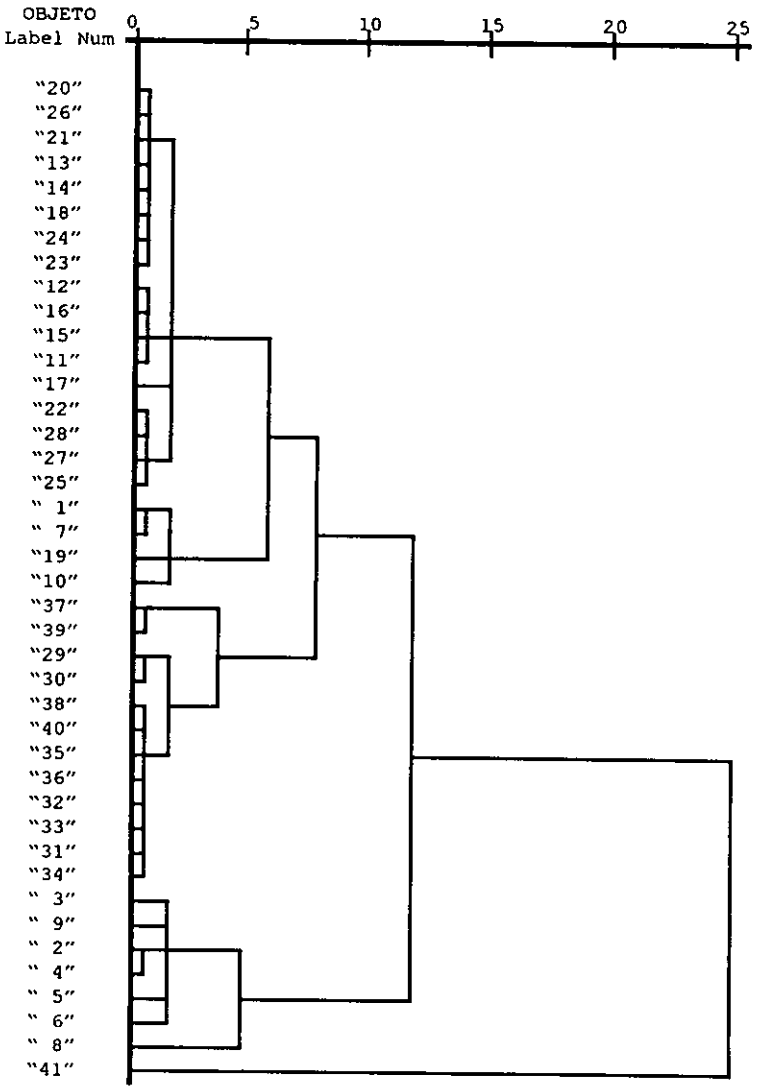
ANEXO 4

Dendograma resultante del Método de Ligaje Completo considerando las m=7 variables predictoras iniciales.



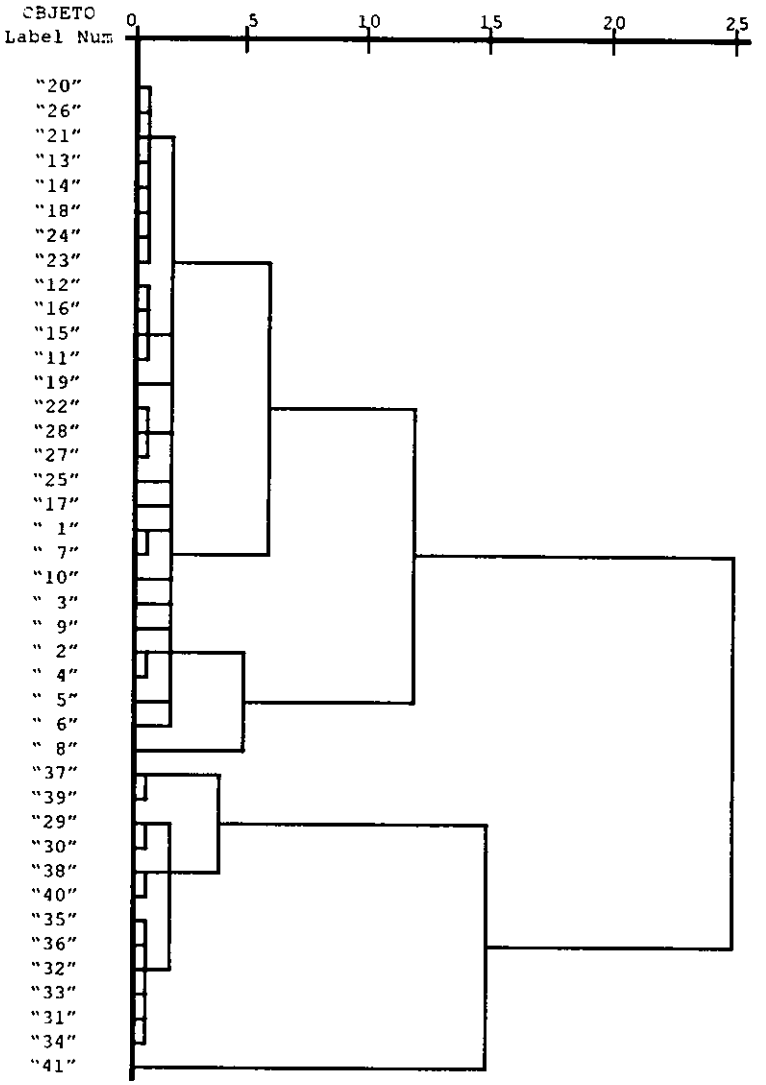
ANEXO 5

Dendograma resultante del Método de Conglomeración con Centroides considerando las m=7 variables predictoras iniciales.



ANEXO 6

Dendograma usando el Método de la Mediana considerando las m=7 variables predictoras iniciales.



ANEXO 7

La variable que tenga la distancia Mahalanobis cuadrada más grande en cada paso, será la que pase al grupo de *variables en el análisis* (anexo 8). Por ejemplo: en el paso 0, la variable X_3 *F. Radial* tiene la distancia D^2 más grande entre los conglomerados 1 y 2.

Variables no incluidas en el análisis

Paso		Tolerancia	Tolerancia mín	F que introducir	Mín. D cuadrado	Entre grupos
0	X1: C. Muestral	1	1	23.403	0.002	2 y 4
	X2: Desempeño	1	1	145.368	0.001	5 y 6
	X3: F. Radial	1	1	92.616	0.655	1 y 2
	X4: F. Lateral	1	1	34.387	0.023	4 y 6
	X5: 1ra. Armónica	1	1	32.747	0.000	5 y 6
	X6: Conicidad	1	1	13.159	0.007	1 y 4
	X7: Peso	1	1	40.899	0.557	3 y 4
1	X1: C. Muestral	0.996	0.996	19.834	0.682	1 y 2
	X2: Desempeño	0.463	0.463	6.795	0.977	1 y 2
	X4: F. Lateral	0.998	0.998	22.068	1.274	5 y 6
	X5: 1ra. Armónica	0.778	0.778	3.195	0.874	1 y 2
	X6: Conicidad	0.902	0.902	12.107	1.203	5 y 6
	X7: Peso	0.960	0.960	9.734	1.714	5 y 6
2	X1: C. Muestral	0.972	0.937	19.777	2.309	3 y 4
	X2: Desempeño	0.454	0.454	5.902	3.472	5 y 6
	X4: F. Lateral	0.976	0.939	17.607	1.918	5 y 6
	X5: 1ra. Armónica	0.768	0.737	2.599	1.970	5 y 6
	X6: Conicidad	0.883	0.852	11.092	1.813	5 y 6
3	X1: C. Muestral	0.809	0.378	19.228	5.735	1 y 2
	X4: F. Lateral	0.904	0.421	14.034	3.481	5 y 6
	X5: 1ra. Armónica	0.767	0.398	2.282	3.771	5 y 6
	X6: Conicidad	0.701	0.327	6.911	3.575	5 y 6
4	X1: C. Muestral	0.935	0.898	18.630	5.136	1 y 2
	X3: F. Radial	0.463	0.454	2.032	3.472	5 y 6
	X4: F. Lateral	0.965	0.925	16.928	1.203	5 y 6
	X5: 1ra. Armónica	0.893	0.840	1.393	0.970	5 y 6
	X6: Conicidad	0.992	0.939	8.175	1.308	5 y 6
5	X3: F. Radial	0.401	0.378	2.368	5.735	1 y 2
	X4: F. Lateral	0.950	0.888	16.764	7.664	3 y 4
	X5: 1ra. Armónica	0.890	0.799	1.327	6.022	1 y 2
	X6: Conicidad	0.903	0.850	9.208	8.615	3 y 4
6	X3: F. Radial	0.306	0.306	1.079	8.788	3 y 4
	X4: F. Lateral	0.713	0.677	4.870	8.639	3 y 4
	X5: 1ra. Armónica	0.759	0.759	2.374	11.429	3 y 4
7	X3: F. Radial	0.306	0.306	0.989	8.815	3 y 4
	X5: 1ra. Armónica	0.746	0.623	2.302	11.573	3 y 4
8	X3: F. Radial	0.381	0.357	1.151	8.327	3 y 4
	X5: 1ra. Armónica	0.811	0.767	1.999	9.521	3 y 4
	X6: Conicidad	0.677	0.677	1.259	8.639	3 y 4

ANEXO 8

Las variables que en cada paso salen de la tabla anterior (anexo 7) son las que entran en la siguiente tabla, y de ésta manera se obtiene el subgrupo de las variables predictoras que más influyen en la regla de clasificación del análisis de conglomerados, aumentando el grado de precisión de la regla de clasificación.

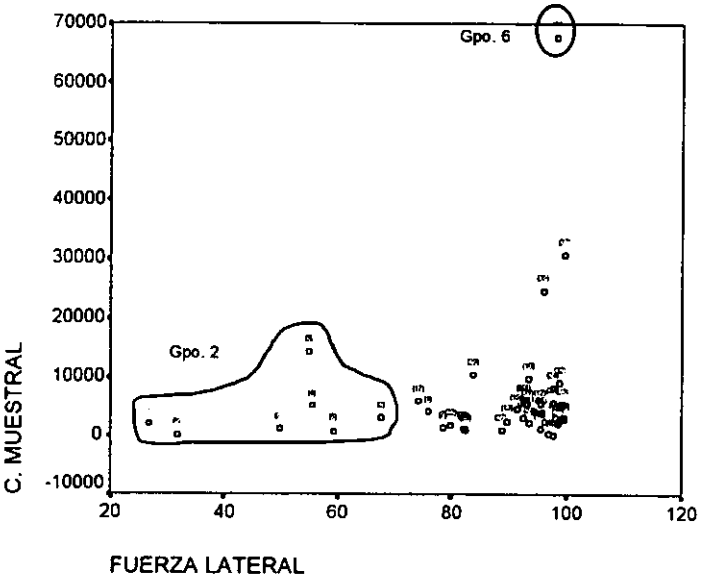
Variables en el análisis

Paso		Tolerancia	F que eliminar	Mín. D cuadrado	Entre grupos
1	X3: F. Radial	1	92.616		
2	X3: F. Radial	0.960	27.586	0.557	3 y 4
	X7: Peso	0.960	9.734	0.655	1 y 2
3	X3: F. Radial	0.463	2.032	0.963	5 y 6
	X7: Peso	0.940	8.605	0.977	1 y 2
	X2: Desempeño	0.454	5.902	1.714	5 y 6
4	X7: Peso	0.942	8.857	0.001	5 y 6
	X2: Desempeño	0.942	43.004	0.557	3 y 4
5	X7: Peso	0.934	8.185	0.116	1 y 2
	X2: Desempeño	0.898	37.265	0.558	3 y 4
	X1: C. Muestral	0.935	18.630	0.963	5 y 6
6	X7: Peso	0.934	7.875	8.260	1 y 3
	X2: Desempeño	0.897	34.388	1.998	3 y 4
	X1: C. Muestral	0.850	20.252	1.308	5 y 6
	X6: Conicidad	0.903	9.208	5.136	1 y 2
7	X7: Peso	0.927	6.031	8.626	3 y 4
	X2: Desempeño	0.884	33.848	2.223	3 y 4
	X1: C. Muestral	0.849	19.259	1.357	5 y 6
	X6: Conicidad	0.677	1.259	7.664	3 y 4
	X4: F. Lateral	0.713	4.870	8.615	3 y 4
8	X7: Peso	0.927	6.968	7.345	1 y 3
	X2: Desempeño	0.888	35.028	1.573	3 y 4
	X1: C. Muestral	0.920	18.439	1.203	5 y 6
	X4: F. Lateral	0.950	16.764	5.136	1 y 2

ANEXO 9

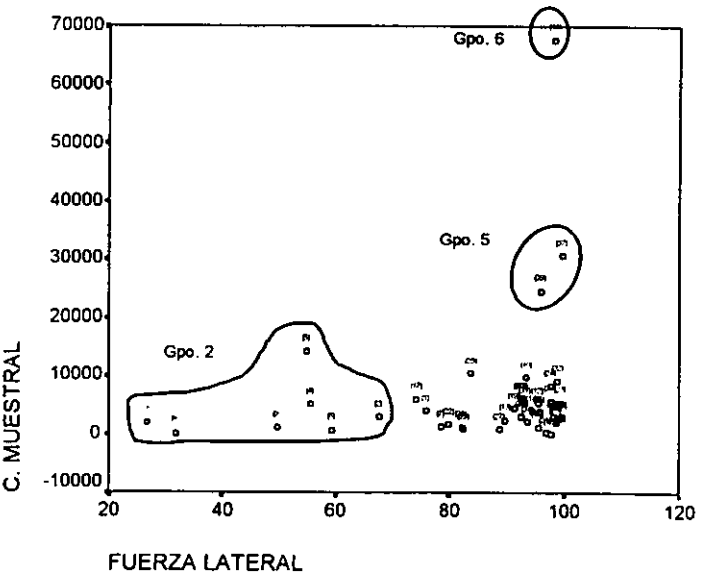
Gráficas de dispersión, de las variables predictoras

X_1: Cantidad Muestral vs. X_4: Fuerza Lateral



Conglomerados definidos mediante las $m = 7$ variables predictoras

X1 C Muestral
X2 Desempeño
X3 F Radial
X4 F Lateral
X5 1ª Harmonica
X6 Conicidad
X7 Peso



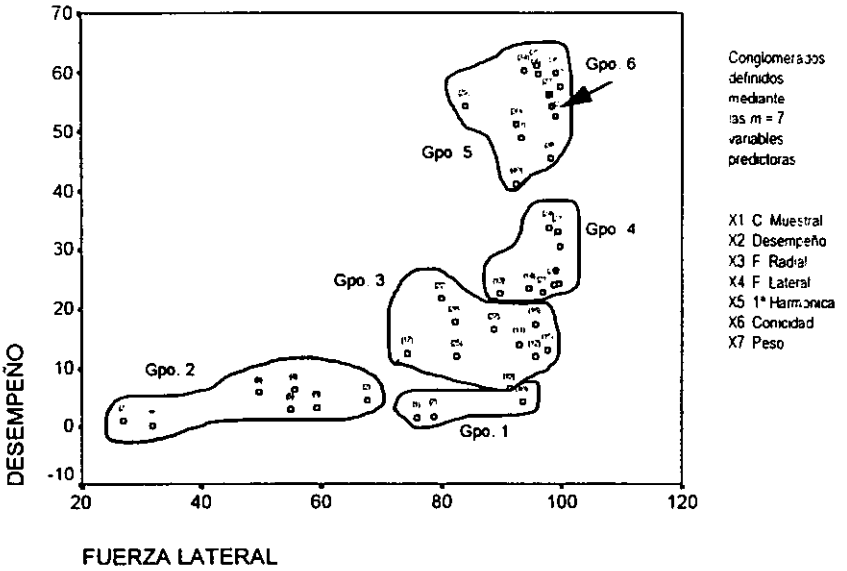
Conglomerados definidos mediante las $h = 4$ variables predictoras

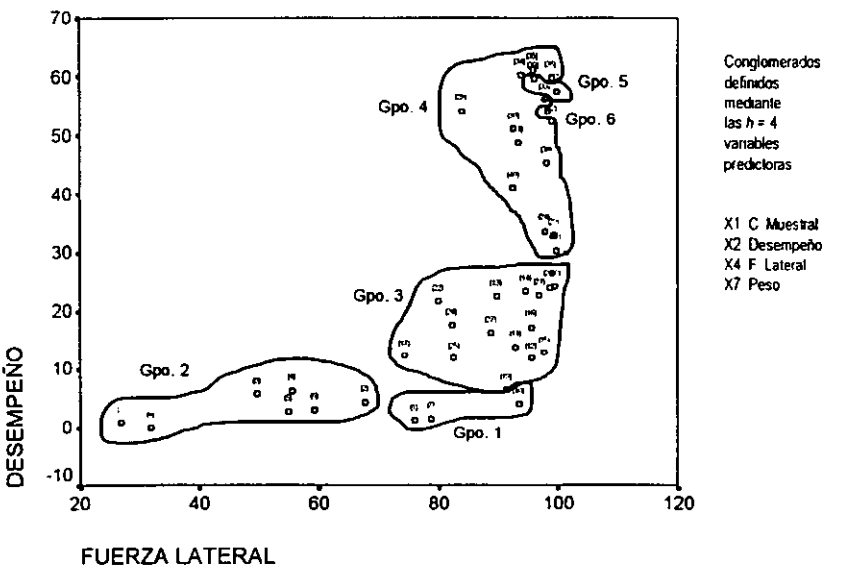
X1 C Muestral
X2 Desempeño
X4 F Lateral
X7 Peso

ANEXO 10

Gráficas de dispersión, de las variables predictoras

X_2: Desempeño vs. X_4: Fuerza Lateral

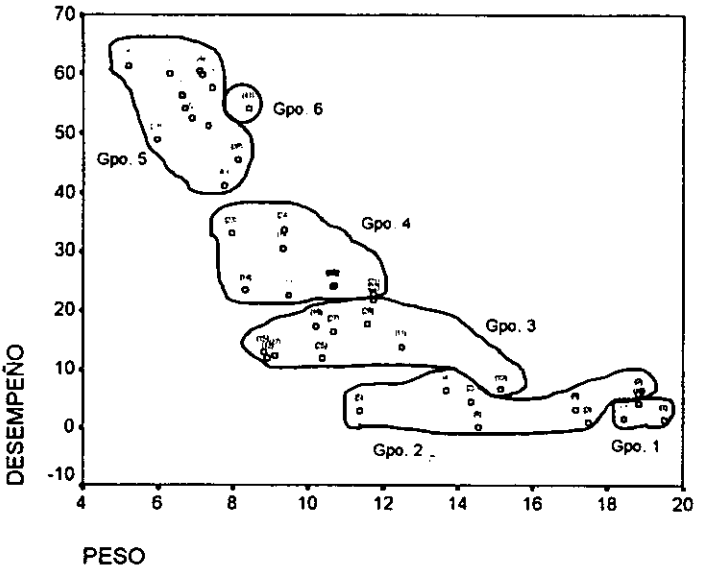




ANEXO 11

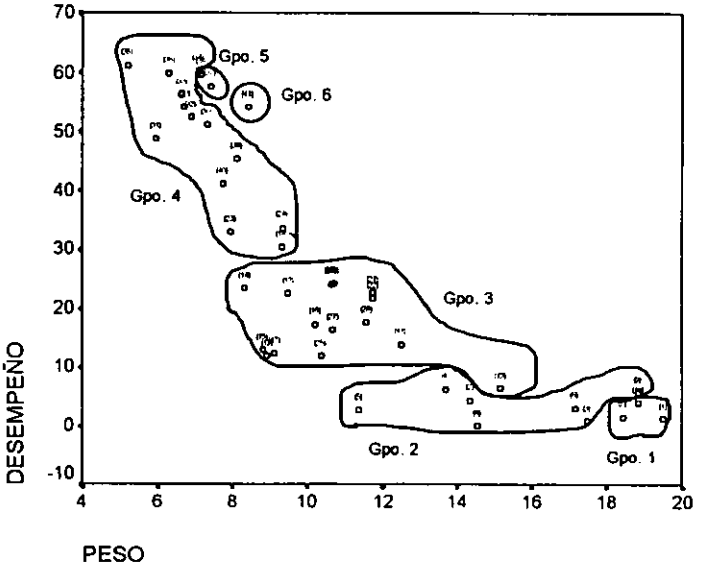
Gráficas de dispersión, de las variables predictoras

X_2: Desempeño vs. X_7: Peso



Conglomerados definidos mediante las $m = 7$ variables predictoras

X1 C Muestral
X2 Desempeño
X3 F Radial
X4 F Lateral
X5 1ª Harmonica
X6 Concidad
X7 Peso



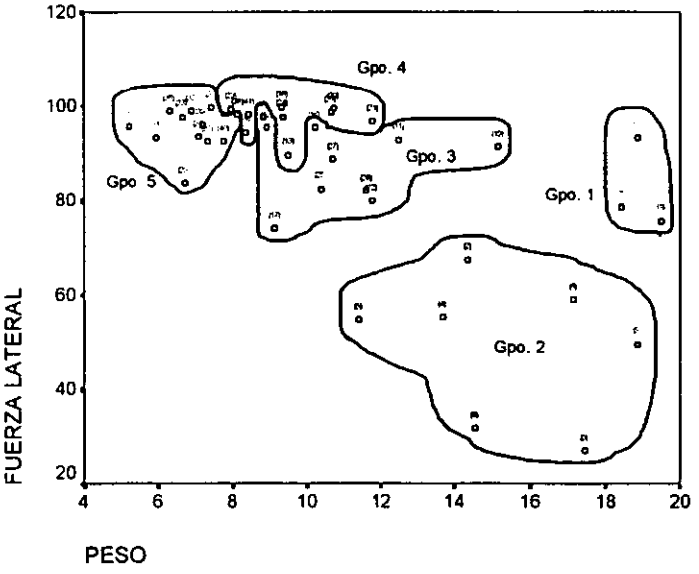
Conglomerados definidos mediante las $h = 4$ variables predictoras

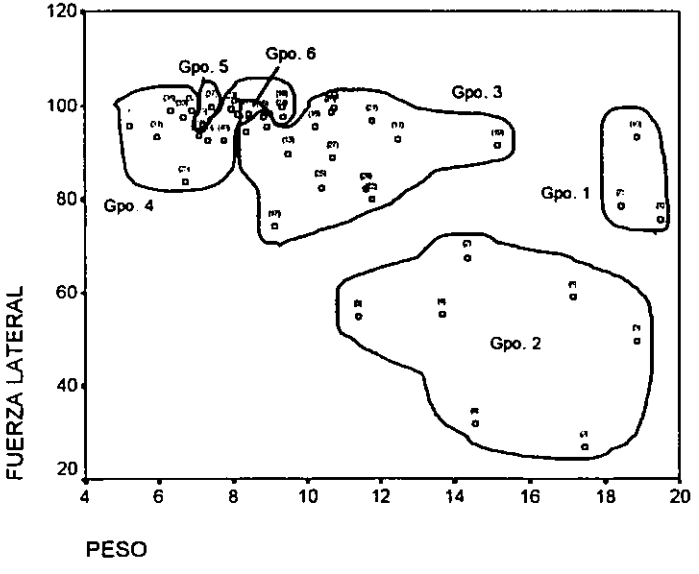
X1 C Muestral
X2 Desempeño
X4 F Lateral
X7 Peso

ANEXO 12

Gráficas de dispersión, de las variables predictoras

X_4: Fuerza Lateral vs. X_7: Peso





BIBLIOGRAFÍA

BIBLIOGRAFÍA

- **Ayres Frank, Jr.**
  Matrices
  Edit: McGraw - Hill, Primera edición, México, 1992

- **Cuadras, Carles M.**
  Métodos de Análisis Multivariante
  Edit: Promociones y Publicaciones Universitarias, Segunda edición , 1991
  Colección: Estadística y Análisis de datos. Tomo 6 , Barcelona, España

- **Everitt, Brian S.**
  Cluster Analysis
  Edit: Edward Arnold, 3ra. Edición, Great Britain, 1993

- **Ezcurdia Hijar, Agustín, y Chávez Calderón Pedro**
  Diccionario Filosófico
  Edit: LIMUSA, México, 1996.

- **Field, Christopher Chat & Collins, Alexander J.**
  Introduction to Multivariate Analysis
  Edit: Science Paperbacks, 1ra. Edición, Great Britain, 1980

♣ **Hair Joseph F., Jr., Anderson Rolph E.**
Tatham Ronald L., Black William C.
Multivariate Data Analysis (with Readings)
Edit: Macmillan, 3ra. Edición, U.S.A., 1992

♣ **Hand, D. J.**
Discrimination and Classification
Edit: John Wiley & Sons, Inc. ,
Wiley Series in Probability and Mathematical Statistics, U.S.A. 1981

♣ **Hartigan, John A.**
Clustering Algorithms
Edit: John Wiley & Sons, 1ra. Edición, U.S.A., 1975

♣ **Huberty, Carl J.**
Applied Discriminant Analysis
Edit: John Wiley & Sons, Inc. ,
Wiley Series in Probability and Mathematical Statistics, U.S.A. 1994

♣ **Kaufman, Leonard & Rousseeuw, Peter J.**
Finding Groups in Data
Edit: John Wiley & Sons, Inc., 1ra. Edición, U.S.A., 1990

♣ **Kendall, Maurice**
Multivariate Analysis
Edit: Charles Griffin & Company LTD, Second edition, Great Britain, 1980

- **Krzanowski, W. J.**
  Principles of Multivariate Analysis
  Edit: Oxford Science Publications, First edition, U.S.A., 1988

- **McLachlan, Geoffrey J.**
  Discriminant Analysis and Statistical Pattern Recognition
  Edit: John Wiley & Sons, Inc.
  Wiley Series in Probability and Mathematical Statistics, 1ra. Edición, U.S.A., 1992

- **Mood Alexander M., Franklin A. Graybill, C. Boes Duane**
  Introduction to the Theory of Statistics
  Edit: McGraw-Hill,
  International Student Edition, 3ra. edición, México, 1983

- **Pérez, López César**
  Econometria y análisis estadístico multivariable con STATGRAPHICS (Técnicas Avanzadas)
  Edit: RA-MA, España, 1996

- **SPSS® Base 7.5 Applications Guide**
  Copyright© 1997 by SPSS Inc.
  United States of America

- **Walpole, Ronald E. & Myers, Raymond H.**
  Probabilidad y Estadística
  Edit: McGraw - Hill, 4ta. Edición, México, 1994

Enciclopedias y Colecciones

- **SALVAT Editores, S. A.**
  Enciclopedia SALVAT Diccionario
  © 1976 Salvat Editores de México, S. A. México

- **W. M. Jackson, Inc., Editores**
  ENCICLOPEDIA PRÁCTICA JACKSON
  Conjunto de conocimientos para la formación autodidáctica
  11va Edición, México, 1970