UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA

ÁREA ACADÉMICA DE INGENIERÍA

TÉCNICAS ESTADÍSTICAS DE CLASIFICACIÓN, UN EJEMPLO DE ANÁLISIS CLUSTER

M O N O G R A F Í A

QUE PARA OBTENER EL TÍTULO DE INGENIERO INDUSTRIAL

P R E S E N T A:

P.D.I.I. Ricardo Enrique Prieto Guerra

Director: Dra. Miriam M. Álvarez Suárez

PACHUCA, HGO. MARZO 2006.

DEDICATORIAS

El presente trabajo está dedicado a todas aquellas personas que han intervenido durante mi formación personal.

A Dios

Qué nos da la sabiduría para descubrir lo correcto, la voluntad para elegirlo y la fuerza para hacer que perdure.

A mi Padre

José Luís Prieto Rivero por ser el ejemplo a seguir en mi vida, admiración y respeto te tengo. Porque sin ti nada hubiera podido hacer.

A mi Madre

Maria Eugenia Guerra Gómez por el cariño y confianza que siempre me da, por ser mi apoyo y guía en lo tiempos difíciles gracias mamá.

A mí Asesora

Dra. Miriam M. Álvarez Suárez por brindarme sus consejos y comentarios, por escuchar con atención mis observaciones y preguntas. Gracias maestra por su infinita paciencia y por contestar a todas esas preguntas.

A mi Universidad

En la cual he pasado los mejores años de mi vida, Gracias por todo lo que me haz dado. Sabes que siempre pondré en alto el nombre de mi escuela.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Los métodos multivariados son extraordinariamente útiles para ayudar a los investigadores en el análisis de grandes conjuntos de datos que constan de una gran cantidad de variables medidas en gran cantidad de unidades experimentales.

A menudo el objetivo principal de los análisis multivariados es el de resumir grandes cantidades de datos por medio de pocos parámetros. En otras ocasiones, el objetivo es el de encontrar relaciones entre 1) las variables respuesta, 2) las unidades experimentales, y 3) tanto las variables respuesta como las unidades experimentales.

Algunas técnicas multivariadas tienden a ser de naturaleza exploratoria en lugar de confirmatoria. Es decir, algunos métodos multivariados tienden a motivar hipótesis en lugar de probarlas. Los métodos estadísticos tradicionales suelen exigir que un investigador establezca algunas hipótesis, reúna algunos datos y a continuación, utilice esos datos para comprobar o rechazar esas hipótesis. Una situación alternativa que se da frecuentemente es el caso en el cual un investigador dispone de una gran cantidad de datos y se pregunta si pudiera haber una información valiosa en ellos. Para resolver este último tipo de situación es que son útiles las técnicas multivariantes, ya que permiten examinar los datos en un intento por saber si hay información que vale la pena y es valiosa en dichos datos.

Una distinción fundamental entre los métodos multivariados es que se clasifican como:

- Técnicas dirigidas a las variables
- Técnicas dirigidas a los individuos

Entre estas últimas técnicas se encuentran: el análisis multivariante de la varianza (MANOVA), los modelos discriminantes y los modelos de agrupamiento o de conglomerados (análisis cluster).

Estos últimos comprenden técnicas que producen clasificaciones a partir de datos que, inicialmente, no están clasificados y no deben confundirse con los modelos discriminantes, en los cuales desde un principio se sabe cuántos grupos existen y se tienen datos que provienen de cada uno de estos grupos (Johnson, 2000).

En esta monografía se enfocará en modelos o técnicas de agrupamiento, también conocidas como tipologías, agrupamientos, clasificación y taxonomía numérica, dependiendo de las disciplinas de aplicación.

El análisis de conglomerados (cluster analysis) es la denominación de un grupo de técnicas multivariantes cuyo principal propósito es agrupar individuos u objetos basándose en las características o descriptores que poseen. Este análisis clasifica objetos; es decir, encuestados, productos, maquinarias, unidades u otras entidades, de tal forma que cada objeto es muy parecido a los que hay en el conglomerado con respecto a algún criterio de selección predeterminado. Los conglomerados de objetos resultantes deben mostrar un alto grado de homogeneidad interna (dentro del conglomerado) y un alto grado de heterogeneidad externa (entre conglomerados).

Por tanto, si la clasificación es acertada, los objetos dentro de los conglomerados estarán muy próximos cuando se representen gráficamente, y los grupos que son diferentes estarán muy alejados.

Este análisis es denominado como análisis Q, construcción de tipologías, análisis de clasificación y taxonomía numérica. Esta variedad de nombres se debe en parte al uso de los métodos de agrupación en disciplinas tan diversas como psicología, biología, sociología, economía, ingeniería y negocios. Aunque los nombres difieren entre disciplinas, todos los métodos tienen una dimensión común: clasificación de acuerdo a una relación natural (1, 2, 3, 6, 12,16 de Hair). Esta dimensión común representa la esencia de todas las aproximaciones del análisis Cluster.

Como tal, el valor fundamental de este análisis descansa en la clasificación de los datos, tal y como sugiere la agrupación "natural" de los datos en sí misma.

Este conjunto de técnicas constituyen una herramienta de análisis muy útil para diferentes situaciones. Por ejemplo, un investigador que haya recogido datos mediante un cuestionario se encuentra frente a un número elevado de observaciones que no tienen sentido a menos que se clasifiquen en grupos manejables. El análisis de conglomerados puede llevar a cabo este procedimiento mediante la reducción de la información de una población completa o de una muestra de subgrupos pequeños y específicos.

Se pueden citar ejemplos de diferentes tipos de aplicaciones del análisis cluster como la derivación de taxonomías en biología para la agrupación de todos los organismos vivientes, clasificaciones Psicológicas basadas en la personalidad y otros rasgos personales, o análisis de segmentación de mercados entre otros.

Esta tradición se ha extendido a la clasificación de objetos, incluyendo la estructura de mercado, análisis de similitudes y diferencias entre productos nuevos y evaluación del rendimiento de empresas para identificar agrupaciones basadas en las estrategias de dichas empresas u orientaciones estratégicas.

El resultado ha generado una profusión de aplicaciones en casi todas las áreas de investigación, creando no sólo una riqueza de conocimiento en el uso del análisis de conglomerados, sino también la necesidad de una mejor comprensión de la técnica para minimizar su mala utilización.

Sin embargo, junto con los beneficios del análisis cluster existen algunos inconvenientes. El análisis cluster puede caracterizarse como descriptivo, teórico y no inferencial. Esta técnica no tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una muestra y se utiliza como una técnica exploratoria. Las soluciones no son únicas en la medida en que la pertenencia al conglomerado para cualquier número de soluciones depende de muchos elementos del procedimiento y se pueden obtener muchas soluciones diferentes variando uno o más de estos elementos.

Además el análisis cluster siempre creará conglomerados, a pesar de la existencia o no de una auténtica estructura en los datos. Finalmente la solución cluster es totalmente dependiente de las variables utilizadas como base para seleccionar la medida de similitud o semejanza. La adición o eliminación de variables relevantes puede tener un impacto sustancial sobre la solución resultante. Por tanto, el investigador debe tener particular cuidado en evaluar el impacto de cada decisión implicada en el desarrollo de un análisis cluster.

El objetivo principal del análisis es definir la estructura de los datos colocando las observaciones más parecidas en grupos. Para llevar a cabo esta tarea se deben considerar 3 cuestiones básicas:

- ¿Cómo se mide la similitud o semejanza? Para ello se necesita un método de observaciones simultáneamente comparadas sobre 2 variables de aglomeración. Son posibles varios métodos, incluyendo la correlación entre objetos, medidas de asociación o midiendo su proximidad de tal forma que la distancia entre las observaciones indica similitud.

- ¿Cómo se forman los conglomerados? El procedimiento debe agrupar aquellas observaciones que son más semejantes dentro de un conglomerado. Este procedimiento debe determinar la pertenencia al grupo de cada observación.

- ¿Cuántos grupos se forman? Puede utilizarse cualquier número de reglas, pero la tarea fundamental es evaluar la similitud media dentro de los conglomerados de tal forma que a medida que la media aumenta, el conglomerado se hace menos similar. El investigador se enfrenta a una disyuntiva: pocos conglomerados frente a menos homogeneidad. A medida que el número de conglomerados disminuye, la homogeneidad dentro de los conglomerados disminuye también. Por tanto se debe buscar un equilibrio entre la definición de las estructuras mas básicas (pocos conglomerados) que todavía mantienen el necesario nivel de similitud dentro de los conglomerados. Una vez que se tengan seleccionados los procedimientos adecuados para cada una de las preguntas anteriores, se puede realizar el análisis cluster.

Esta monografía consta de 5 capítulos, donde se expone una breve Introducción, Justificación del trabajo y los Objetivos General y Específicos. En los Capítulos 1, 2 ,3 y 4 se incluye de manera resumida, algunos conceptos básicos generales de las técnicas multivariadas, las escalas de medida y la clasificación de los mismos de acuerdo con la cantidad y la naturaleza de las variables, así como el objetivo perseguido en el estudio. En el Capítulo 5, se utilizan ejemplos de clasificación de un grupo de países y empresas del estado de Hidalgo en los cuales se utilizan diferentes alternativas de solución, para, finalmente arribar a las conclusiones. A continuación aparecen las Referencias bibliográficas utilizadas y los Anexos.

En este trabajo se recopilaron algunas técnicas estadísticas multivariantes que en la Licenciatura en Ingeniería Industrial no se imparten actualmente, y que pueden ser útiles en ciertas materias de ingeniería, así como el manejo de programas estadísticos como el SPSS (Statistical Package for the Social Sciences) el cual es un programa computacional que se utiliza mayormente para cálculos estadísticos, aunque incluye un sin número de utilidades más. Actualmente, la estadística ha adquirido, de manera progresiva, una mayor relevancia en todos los sectores universitarios y en general en la sociedad, y las técnicas que presentamos a continuación pueden ser herramientas muy útiles en el manejo de una empresa, entre otras aplicaciones. Por eso se sugiere el uso de esta herramienta dado que los continuos cambios a los que están sometidas las empresas y organismos públicos demandan profesionales que sean capaces de adaptarse con éxito a las nuevas tecnologías y los nuevos avances de la ciencia.

OBJETIVO GENERAL Y ESPECIFICOS

El objetivo general de la presente Monografía es el siguiente:

"Realizar un análisis de los aspectos fundamentales del modelo de análisis de conglomerados (cluster analysis) y aplicarlos al caso de la clasificación de empresas."

Los objetivos específicos son los siguientes:

1. Desarrollar las técnicas de similitud más importantes en el análisis de cluster.

2. Analizar los métodos de formación de conglomerados más importantes para la clasificación.

3. Utilizar programas computacionales de estadística para la clasificación de empresas utilizando diferentes variantes de medidas de semejanza y algoritmos de agrupamiento.

CAPÍTULO 1. MÉTODOS DE CLASIFICACIÓN

1.1 Técnicas estadísticas multivariantes

El análisis cluster puede verse como una aproximación a la construcción de modelos en seis pasos, de los cuales los tres primeros se corresponden con los objetivos, el cuarto con la selección de un algoritmo de cluster, el quinto con la interpretación de los mismos y el sexto con la validación y perfiles de los clusters. Los pasos son los siguientes:

1. **Descripción de una taxonomía.**
2. **Simplificación de los datos.**
3. **Identificación de relaciones.**
4. **Selección de un algoritmo de cluster.**
5. **Interpretación de los clusters.**
6. **Validación y perfiles de los clusters.**

En la siguiente figura 1.1 se puede apreciar la secuencia lógica para la aplicación de un análisis cluster.

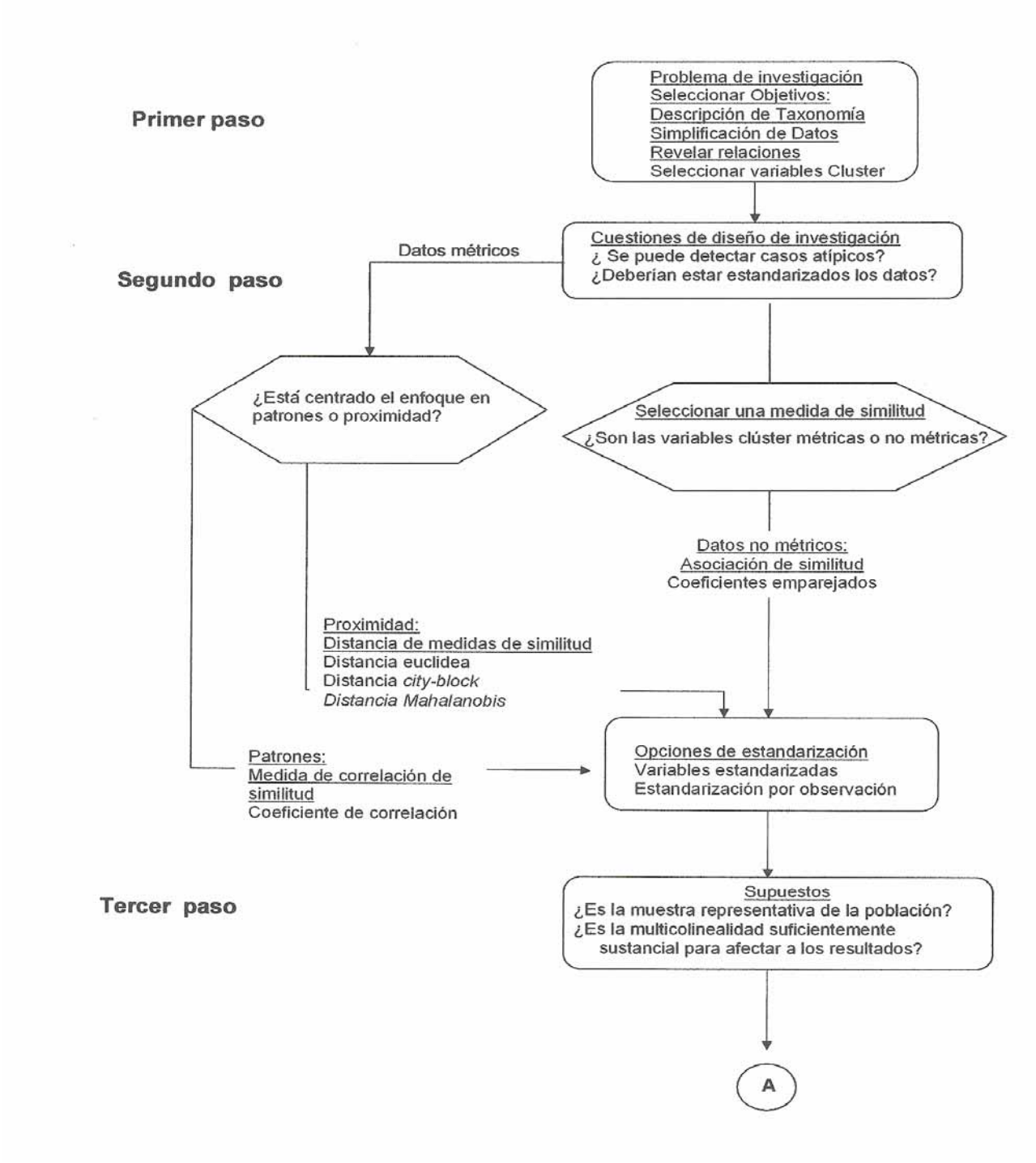


Figura 1.1 Pasos de 1 a 3 para la elaboración de un análisis de Conglomerados (cluster).

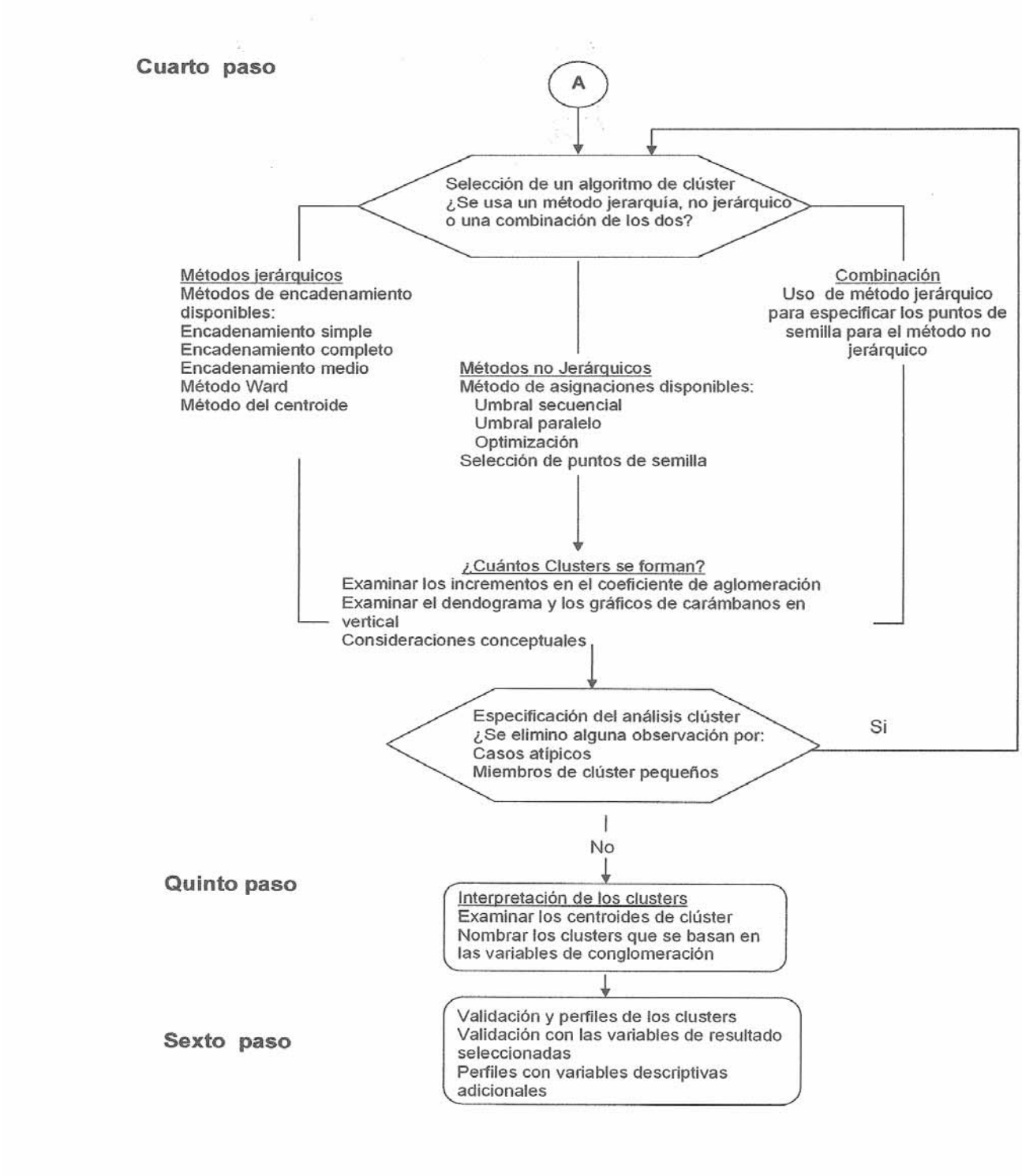


Figura 1.2 Pasos de 4 a 6 para la elaboración de un análisis de Conglomerados (cluster).

El análisis de varias variables, o análisis multidimensional, o análisis multivariado no es más que el conjunto de métodos estadísticos que tienen por objeto el estudio de las relaciones existentes entre varias variables dependientes o interdependientes, que han sido medidas sobre los mismos individuos (Dagnelie, 1977).

Las técnicas del análisis multivariante están siendo ampliamente aplicadas a la industria, a la administración y a las investigaciones científicas. Precisamente es en este último aspecto en el que se han intentado todas estas técnicas con mayor éxito. Para atender este creciente interés, se han publicado numerosos libros y artículos sobre los aspectos teóricos y matemáticos de estas herramientas. Sin embargo, se han escrito pocos libros para el investigador que no es especialista en matemática o en estadística, y que lo que necesita conocer es sus características generales, su forma de utilización y la interpretación de los resultados fundamentalmente.

En la mayor parte de los problemas actuales, los directivos no pueden fiarse de las antiguas aproximaciones donde se consideraban grupos de individuos homogéneos y caracterizados por un reducido número de variables demográficas. En su lugar, deben desarrollar estrategias para atraer a numerosos segmentos de la población con características demográficas y psicográficas en un mercado con múltiples restricciones legales, económicas, competitivas, tecnológicas, etc. Sólo a través de las técnicas de análisis multivariado se pueden examinar adecuadamente las relaciones múltiples de este tipo para llegar a una comprensión de la toma de decisiones más completa y realista (Hair, Anderson, Tatham y Black, 2000).

Es imposible discutir la aplicación de las técnicas estadísticas multivariantes sin una mención al impacto de la informática en las últimas décadas, permitiendo procesar grandes y complejas bases de datos.
Toda la estadística teórica de las técnicas multivariantes fue desarrollada a principios del siglo XX, pero sólo pudieron utilizarse ampliamente a partir del desarrollo de la computación. Existen y están a disposición de estudiantes e investigadores en todas partes del mundo programas completos de estadística diseñados para computadoras personales que contienen todo el tratamiento de datos multivariantes. Entre ellos se encuentran SPSS, SAS, BMDP Y S-PLUS, que incluyen técnicas de escala multidimensional, modelos de ecuaciones simultáneas o estructurales, y análisis conjunto. Además, más recientemente, se están desarrollando sistemas expertos dirigidos incluso a temas tales como la selección de una técnica estadística o diseñar un plan de muestreo que asegure los objetivos prácticos y estadísticos deseados.

1.2 Conceptos básicos del análisis multivariante

Cuando se han observado *"p"* características numéricas (i = 1,..., p) sobre "n" individuos (j = 1,..., n), los resultados obtenidos pueden escribirse en una ***MATRIZ DE DATOS*** de dimensión *p x n*:

$$X = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{vmatrix}$$

Cada columna de esta matriz se refiere a un individuo y constituye un vector que nombraremos x_j

$$x_j = \begin{vmatrix} x_{1j} \\ x_{2j} \\ . \\ . \\ x_{pj} \end{vmatrix} \quad .$$

Esta matriz de datos se reduce bajo la forma de ***parámetros***:

- la media,
- las varianzas y co-varianzas,
- las desviaciones típicas y
- los coeficientes de correlación

$$x_i = \begin{vmatrix} x_1 \\ x_2 \\ . \\ . \\ x_p \end{vmatrix} \quad \text{las “medias”}$$

$$S = \begin{vmatrix} s_{11} & s_{12} & . . . & s_{1p} \\ s_{21} & s_{22} & . . . & s_{2p} \\ . & . & & . \\ . & . & & . \\ s_{p1} & s_{p2} & . . . & s_{pp} \end{vmatrix} \quad \text{las “varianzas y covarianzas”}$$

$S_1 = \sqrt{S_{11}}$, $S_2 = \sqrt{S_{22}}$,. . ., $S_{p1} = \sqrt{S_{pp}}$ “las desviaciones típicas”

$$R = \begin{vmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ . & . & & . \\ . & . & & . \\ r_{p1} & r_{p2} & \dots & 1 \end{vmatrix}$$

los “coeficientes de correlación”

Es justamente a partir de estas matrices S y R que se realizarán casi todos los análisis multivariados.

El elemento esencial del análisis multivariante es el “valor teórico”, una combinación lineal de variables con ponderaciones determinadas empíricamente. El investigador especifica las variables, mientras que las ponderaciones son objeto específico de determinación por parte de la técnica multivariante. Un valor teórico de “n” variables ponderadas (X_1, …, X_n) puede expresarse matemáticamente de la siguiente forma:

Valor teórico = $w_1X_1 + w_2X_2 + \dots + w_nX_n$

donde X_n es la variable observada y w_n es la ponderación determinada por la técnica multivariante.

El resultado es un valor único que representa una combinación de *“todo el conjunto”* de variables que mejor se adaptan al objeto del análisis multivariante específico. En regresiones múltiples. El valor teórico se determina de tal forma que represente la mejor correlación con la variable que se está prediciendo. En el análisis discriminante, el valor teórico se forma de tal manera que produzca resultados para cada observación que diferencien de forma máxima entre grupos de observaciones.

Y en el análisis factorial, los valores teóricos se forman para representar mejor las estructuras subyacentes o la dimensionalidad de las variables tal y como se representan en sus ínter correlaciones.

En cada caso, el valor teórico capta el carácter multivariante del análisis, por lo que se debe entender no sólo su impacto conjunto para lograr el cumplimiento de cada técnica, sino también la contribución de cada variable separada al efecto del valor teórico en su conjunto.

1.3 Escalas de medida

El análisis de los datos implica la separación, identificación y medida de la variación en un conjunto de variables, tanto entre ellas mismas como entre una variable dependiente y una ó más variables independientes. El término clave aquí es “medida”, dado que el investigador no puede separar o identificar una variación a menos que pueda ser mesurable. La medida es importante para representar con precisión el concepto de nuestro interés y es crucial en la selección del método de análisis multivariante más apropiado.

Existen dos tipos básicos de datos: **no métricos** (cualitativos) y **métricos** (cuantitativos). Los datos no métricos son atributos, características o propiedades categóricas que identifican o describen a un individuo. Describen diferencias en tipo o clase indicando la presencia o ausencia de una característica o propiedad. Las medidas no métricas pueden tener escalas nominales u ordinales. La medida con una escala nominal asigna números que se usan para etiquetar o identificar sujetos u objetos, sin ningún significado cuantitativo ya que sólo indican la presencia o ausencia del atributo o característica bajo investigación.

Las escalas ordinales representan un nivel superior de precisión de la medida. Estas variables pueden ser ordenadas o clasificadas con relación a la cantidad del atributo poseído.

Por el contrario, las medidas de datos métricos están constituidas de tal forma que los sujetos pueden ser identificados por diferencias entre grado o cantidad.

Las variables medidas métricamente reflejan cantidades relativas o grado, y proporcionan el nivel más alto de medida de precisión, permitiendo realizar con ellas, todas las operaciones matemáticas.

1.4 Clasificación de los análisis multivariados

Los métodos estadísticos multivariados se deben seleccionar en cuanto a:

- la ***estructura*** de la ***matriz de datos,***
- el ***objetivo*** perseguido, y
- la ***naturaleza*** de esos ***datos.***

1.4.1 Clasificación según su estructura de datos

Según la estructura de la matriz de datos, los métodos pueden clasificarse en:

- ***sin*** ninguna ***estructura*** en particular, (1, 1)
  (Análisis de componentes principales y análisis factorial; conglomerados)

- una ***estructura entre variables***, (k_1, 1)

(Métodos de regresión múltiple), o (análisis de correlación canónica)

- una ***estructura entre individuos***, (1, k_2) o

(Análisis discriminante)

- ***ambas estructuras*** (k_1, k_2)

(Análisis de correspondencias múltiples)

1.4.2 Clasificación según el objetivo perseguido

Según el objetivo perseguido, los métodos son muy difíciles de clasificar, pues puede haber muchos y muy diferentes, pero se agruparan en dos grandes grupos:

- los ***descriptivos***, y
- los ***inferenciales***.

1.4.2.1 Métodos descriptivos o exploratorios

En el caso de los *métodos descriptivos o exploratorios:*

- "p" var. cuantitativas ------ Análisis Factorial (Análisis de Componentes Principales y Análisis Factorial común)

- "p" var. cualitativas ------ Análisis de Correspondencias y

y/o cuantitativas **Métodos de Conglomerados (clusters)**

1.4.2.2 Métodos inferenciales o confirmatorios

En este caso, siempre hay dos grupos de variables y casi siempre se reconocen como variables independientes y variables dependientes. Por esto, se tendrá que tener en cuenta la naturaleza y la cantidad de variables de cada uno de los grupos:

Número y naturaleza de las variables de estudio

Tabla 1.1 Comparativa del número y naturaleza de las variables y métodos que se aplican en cada caso.

Var. Dependientes	Var. Independientes	Métodos
1 var. Cuantitativa	1 ó n var. Cuantitativas	Regresión Múltiple
1 var. Cualitativa	n var. Cuantitativas	Discriminante
p var. Cuantitativas	p var. Cuantitativas	Correlación Canónica
p var. Cuantitativas	1 ó n var. Cualitativas	MANOVA
p var. Cuantitativas	n var. Cuantitativas y/o n var. Cualitativas	Análisis de Corresp. (Simple o Múltiple)

Fuente: (Hair, Anderson, Tatham y Black, 2000)

CAPÍTULO 2. MARCO TEÓRICO DEL ANÁLISIS CLUSTER

2.1 Generalidades

El análisis cluster se basa en intentar responder como es que ciertos objetos (casos) pertenecen o "caen" naturalmente en cierto número de clases o grupos, de tal manera que estos objetos comparten ciertas características.

Estas técnicas o también llamados "análisis Q", "construcción de tipología", "análisis de clasificación" y "taxonomía numérica", son procedimientos multivariados que nos permiten agrupar las observaciones de forma que los datos sean muy homogéneos dentro de los grupos (mínima varianza) y que estos grupos sean lo más heterogéneos posible entre ellos (máxima varianza). De este modo se obtiene una clasificación multivariante de los datos con la que se puede comprender mejor los mismos y la población de la que proceden. Podemos realizar conglomerados por casos, por variables o por bloques, si se agrupan variables y casos. El análisis clúster se puede utilizar para:

- La taxonomía: agrupar especies naturales.
- El marketing: clasificar consumidores tipo.
- La medicina: clasificar seres vivos con los mismos síntomas y características patológicas.
- El reconocimiento de patrones.
- Formar grupos de píxeles en imágenes digitalizadas enviadas por un satélite desde un planeta para identificar los terrenos.

Es un análisis descriptivo y no inferencial, por lo cual es una técnica exploratoria que parte de una matriz no estructurada y que tiene como objetivo la obtención de un conjunto de individuos en dos ó más grupos basándose en su similitud para un conjunto de variables o características especificadas.

Al formar grupos homogéneos, el investigador puede conseguir los siguientes objetivos:

- **Descripción de una taxonomía**, permitiendo obtener una clasificación de los individuos que a su vez puede ser comparada con una **tipología** propuesta (clasificación basada en la teoría).

- **Simplificación de los datos**, ya que las observaciones pueden agruparse para análisis posteriores; es decir, permite ver las observaciones como miembros de un conglomerado y perfiladas por sus características generales.

- **Identificación de relaciones**, ya que al estar los conglomerados definidos y la estructura subyacente de los datos representada en dichos conglomerados, el investigador tiene un medio de revelar las relaciones entre los individuos que quizá seria muy difícil de detectar a partir de las observaciones individuales.

Las soluciones pueden ser diferentes variando uno ó más de sus elementos. Es decir, la solución es totalmente dependiente de las variables utilizadas como base para la medida de similitud.

Como el objetivo principal del análisis de conglomerados es definir la estructura de los datos colocando las observaciones más parecidas en grupos, se deben abordar tres cuestiones básicas:

1- ¿Cómo medir la similitud?

Existen varias formas, pero hay tres métodos que dominan las aplicaciones del análisis de clúster:

- medidas de correlación,
- medidas de asociación
- medidas de semejanza o desemejanza (distancias).

Cada uno de los métodos representa una perspectiva particular de similitud, dependiendo tanto de sus objetivos como del tipo de datos. Tanto las medidas de distancia como la correlación exigen datos métricos, mientras que las medidas de asociación son utilizadas para datos no métricos.

2- ¿Cómo se forman los conglomerados?

No importa cómo se mida la similitud, el procedimiento debe agrupar aquellas observaciones que son más similares dentro de un conglomerado. Este procedimiento debe determinar la pertenencia al grupo de cada observación.

3- ¿Cuántos grupos se forman?

Para esto puede utilizarse cualquier número de reglas, pero la tarea fundamental es evaluar la similitud "media" dentro de los conglomerados, de tal forma que a medida que la media aumenta, el conglomerado se hace menos similar. Una estructura simple, al tender hacia la parsimonia, se refleja en el menor número de conglomerados posible.

Pero a medida que el número de conglomerados disminuye, la homogeneidad dentro de los conglomerados también disminuye, luego se debe buscar un equilibrio entre las definición de las estructuras más básicas (pocos conglomerados) que todavía mantienen el nivel necesario de similitud dentro de los conglomerados.

2.2 Estudio previo al análisis de conglomerados

Después de definir los objetivos y haber seleccionado las variables para el estudio, el investigador debe tratar tres cuestiones antes de iniciar el proceso de partición. Estas cuestiones son:

- Selección de la muestra de datos. Detectar valores atípicos
- Selección y transformación de variables a utilizar
- Selección del concepto de distancia o similitud y medición de las mismas.

Para resolver estas cuestiones no hay reglas generales, y desafortunadamente, muchas de las aproximaciones ofrecen diferentes resultados para el mismo conjunto de datos. Es por esto que se proponen algunas opciones, sin que sea una regla general y dejando claro, que el conjunto de datos debe ser analizado por el investigador antes de ofrecer una solución final.

2.2.1 Detección de valores atípicos

Como se sabe, los valores atípicos pueden ser producto de observaciones verdaderamente aberrantes que no son representativos de la población en general, o también una muestra reducida del grupo pero que sí pertenece a la población y que puede provocar una mala representación del grupo. Hay muchas formas de detectar datos atípicos, sin embargo, cuando el número de variables e individuos no es demasiado grande, es aconsejable realizar un “diagrama de perfil gráfico”, que consiste en situar en el eje horizontal las variables y los valores correspondientes de cada variable en el eje vertical. Así, se obtendrá una línea quebrada para cada individuo.

Los valores atípicos serán aquellos individuos con perfiles muy diferentes, caracterizados por tener valores extremos para una ó más variables. El investigador debe decidir si elimina o no dichos atípicos ya que al hacerlo pudiera distorsionar la estructura efectiva de los datos.

2.2.2 Estandarización de los datos

Un problema al que se enfrentan todas las medidas de distancia es que el uso de datos no estandarizados implica inconsistencias entre las soluciones clúster cuando cambia la escala de las variables. El orden de las similitudes puede cambiar profundamente con sólo un cambio de escala en una de las variables. Debería emplearse, por tanto, la estandarización de las variables de aglomeración, siempre que sea conceptualmente posible, para evitar diferentes soluciones por el solo hecho de contar, por ejemplo, con una variable medida en metros y cambiarla para el análisis en centímetros.

Se recomienda incorporar el procedimiento de estandarización que aporta la distancia de Mahalanobis (D^2) y que además evalúa la varianza-covarianza dentro del grupo, que ajusta las intercorrelaciones entre las variables. Conjuntos de variables altamente intercorrelacionadas del análisis clúster pueden ponderar implícitamente un conjunto de variables en los procedimientos de aglomeración.

En resumen, esta distancia calcula una medida de distancia entre objetos comparable al R^2 del análisis de regresión. En caso de no contar con esta medida de similitud, los investigadores pueden utilizar la distancia euclidiana al cuadrado como alternativa.

2.3 Medidas de semejanza

El concepto de semejanza es fundamental para el análisis de clúster. La semejanza entre individuos es una medida de correspondencia, o del parecido entre individuos que van a ser agrupados. Aquí, las características que definen la semejanza, se especifican en primer lugar, y a continuación, se combinan las características en una medida de semejanza calculada para todos los pares de individuos. El procedimiento del análisis de conglomerados procede a continuación a agrupar individuos similares en el mismo conglomerado.

2.3.1 Análisis Q y R

Según lo observado por Cattell (1952), la matriz de los datos se puede estudiar a partir de dos puntos de vista fundamentales: si se desean las relaciones entre los objetos o las relaciones entre los descriptores o variables. El aspecto importante es que ambos modos de análisis están basados en diferentes medidas de asociación.

La medida de la dependencia entre los descriptores se realiza utilizando el coeficiente de correlación r de Pearson por lo que el estudio de la matriz de base con tales coeficientes se llama análisis *R*. Por el contrario, el estudio de la matriz para analizar las relaciones entre objetos es llamado el análisis *Q*.

Cattell (1966) también describió que la caja de datos (formada por descriptores **x** objetos **x** tiempo) se puede analizar desde otros puntos de vista además del Q y R, definiendo finalmente, seis modos de análisis:

O: tiempos x descriptores (un solo objeto):
P: descriptores x tiempos (un solo objeto);
Q: objetos x descriptores (un solo tiempo);
R: descriptores x objetos (un solo tiempo);
S: objetos x tiempos (un solo descriptor);
T: tiempos x objetos (un solo descriptor).

A continuación, la discusión de los coeficientes de asociación se centrará solamente en los dos modos básicos, es decir las medidas Q (entre objetos) y las medidas R (entre descriptores).

2.3.2 Coeficientes de asociación

El enfoque más usual para determinar la semejanza entre objetos o descriptores es, en primer lugar, condensar toda la (o la parte más relevante de) información disponible de la matriz de los datos en una matriz cuadrada de asociación entre los objetos o los descriptores. En la mayoría de los casos, la matriz de asociación es simétrica. Las matrices no-simétricas se pueden descomponer en componentes simétricos y componentes asimétricos y entonces, los componentes se pueden analizar por separado.

Los objetos o los descriptores podrán ser agrupados en conglomerados o representados en un espacio reducido después de analizarse la matriz de asociación.

Por lo tanto, la estructura que resulta del análisis numérico es la de la matriz de asociación; aunque los resultados del análisis no reflejan necesariamente toda la información contenida originalmente en la matriz inicial de los datos.

Esto pone de relieve la importancia de elegir una medida apropiada de asociación. Esta opción determina la aplicación del análisis. Por lo tanto, debe tenerse en cuenta las consideraciones siguientes:

- La naturaleza del estudio (i.e. la pregunta inicial y la hipótesis) determina la clase de estructura que se evidenciará a través de una matriz de asociación, y por lo tanto el tipo de medida de semejanza que debe ser utilizado.

- Las diferentes medidas disponibles están sujetas a diversas restricciones matemáticas. Los métodos de análisis a los cuales la matriz de asociación será aplicada (clasificación, ordenación) requiere a menudo medidas de semejanza con características matemáticas específicas.

- También debe considerarse el aspecto computacional, y preferiblemente, elegir una medida que esté disponible en un programa computacional o puede ser programado fácilmente.

Los investigadores son, en principio, libres de definir y utilizar cualquier medida de asociación conveniente al fenómeno de estudio; las matemáticas imponen pocas restricciones a esta elección. Esta es la razón por la cual se encuentran tantos coeficientes de asociación en la literatura.

Algunos de ellos son de amplia aplicabilidad, mientras que otros se han creado para necesidades específicas. Varios coeficientes han sido vueltos a descubrir por diferentes autores a través del tiempo y se pueden conocer bajo varios nombres.

Las medidas de similitud de asociación se utilizan para comparar objetos cuyas características se miden sólo en términos no métricos (nominales y ordinales).

Si los datos están divididos en clases, el estadístico Chi-cuadrado es el más utilizado. Si los datos son binarios, existen una diversidad de distancias que van desde la distancia euclidiana hasta las medidas de SOKAL y Sneath, Jaccard, Lambda, Ochiai y otras y si se trata de variables ordinales o nominales, la distancia de Gower es la más conocida.

2.3.3 Medidas de distancia

Las medidas de similitud de distancia, que representan la similitud como la proximidad de las observaciones respecto a las otras, para las variables del valor teórico del análisis de clúster, son las medidas de similitud más utilizadas. Los conglomerados basados en la distancia, tienen valores más parecidos para el conjunto de variables. Las medidas de distancia utilizadas para el agrupamiento pueden ser también muy diversas; entre ellas se encuentran: la distancia euclidiana, la distancia euclidiana al cuadrado, la distancia Coseno, la distancia de Tchebychev, la de Minkowski y otras que el investigador pueda concebir para datos métricos.

Al intentar seleccionar una medida de distancia particular, el investigador debe recordar que diferentes medidas de distancia o un cambio en la escala de las variables, pueden llevar a diferentes soluciones de clúster.

Por tanto, es aconsejable utilizar varias medidas y comparar los resultados con pautas teóricas o conocidas por trabajos anteriores.

2.3.4 Medidas de proximidad y de distancia

Una vez establecidas las variables y los objetos a clasificar el siguiente paso consiste en establecer una medida de proximidad o de distancia entre ellos que cuantifique el grado de similaridad entre cada par de objetos.

Las **medidas de proximidad, similitud o semejanza** miden el grado de semejanza entre dos objetos de forma que, cuanto mayor (respecto al menor) es su valor, mayor (respecto la menor) es el grado de similaridad existente entre ellos y con más (respectivamente menos) probabilidad los métodos de clasificación tenderán a ponerlos en el mismo grupo.

Las **medidas de disimilitud, de semejanza o distancia** miden la distancia entre dos objetos de forma que, cuanto mayor(respecto al menor) sea su valor, más diferentes son los objetos y menor (respecto al mayor) la probabilidad de que los métodos de clasificación los pongan en el mismo grupo.

En la literatura existen multitud de medidas de semejanza y de distancia dependiendo del tipo de variables y datos considerados. En esta monografía solamente se verán algunas de las más utilizadas. Para otros ejemplos ver Anderberg (1973) o el manual de SPSS. Siguiendo el manual de SPSS se puede distinguir los siguientes tipos de datos, los cuales son presentados a continuación.

2.3.4.1 Tipos de datos

1) **De intervalo**: se trata de una matriz objetos x variables en donde todas las variables son cuantitativas, medidas en escala intervalo o razón

2) **Frecuencias:** las variables analizadas son categóricas de forma que, por filas, tenemos objetos o categorías de objetos y, por columnas, las variables con sus diferentes categorías. En el interior de la tabla aparecen frecuencias.

3) **Datos binarios:** se trata de una matriz objetos *x* variables pero en la que las variables analizadas son binarias de forma que 0 indica la ausencia de una característica y 1 su presencia.

2.3.4.2 Medidas de proximidad

a) Medidas para variables cuantitativas

1) Coeficiente de congruencia

$$C_{rs} = \frac{\sum_{j=1}^{P} X_{rj} X_{sj}}{\sqrt{\sum^{P} X_{rj}^{2}} \sqrt{\sum^{P} X_{Sj}^{2}}} \qquad \textbf{(2.1)}$$

que es el coseno del ángulo que forman los vectores $(x_{r1},\ldots,x_{rp})'$ y $(x_{s1},\ldots,x_{sp})'$.

2) Coeficiente de correlación

$$r_{rs} = \frac{\sum_{J=1}^{p} (x_{rj} - \bar{x}_r)(x_{sj} - \bar{x}_s)}{\sqrt{\sum_{j=1}^{p} (x_{rj} - \bar{x}_R)^2} \sqrt{\sum_{j=1}^{p} (x_{sj} - \bar{x}_s)^2}} \qquad \textbf{(2.2)}$$

$$\textit{donde} \qquad \overline{X}_r \frac{\sum_{j=1}^{P} x_{rj}}{p} \qquad y \qquad \overline{X}_s = \frac{\sum_{j=1}^{p} x_{sj}}{p} \qquad \textbf{(2.3)}$$

Si los objetos r y s son variables, r_{rs} mide el grado de asociación lineal existente entre ambas.

Estas dos medidas se utilizan, preferentemente para clasificar variables siendo, en este caso, invariantes por cambios de escala y, en el caso del coeficiente de correlación, invariante por cambio de origen. Por esta razón es más conveniente utilizar el coeficiente de congruencia con variables tipo razón en las cuales el origen está claramente definido.

Conviene observar, además, que tanto c_{rs} como r_{rs} toman valores comprendidos entre -1 y 1 pudiendo tomar, por lo tanto, valores negativos. Dado que, en algunos casos, (por ejemplo, si los objetos a clasificar son variables), los valores negativos cercanos a -1 pueden implicar fuerte semejanza entre los objetos clasificados. Conviene, en estas situaciones, utilizar como medida de semejanza sus valores absolutos.

b) Medidas para datos binarios

En este caso se construye una tabla de contingencia, para cada par de objetos **r** y **s**, de la forma:

*Tabla 2.1. Tabla de Contingencia para objetos **r** y **s***

Objeto s \ Objeto r	0	1
0	a	b
1	c	d

Fuente: (Hair, Anderson, Tatham y Black, 2000)

donde a = número de variables en las que los objetos r y s toman el valor 0, etc. y p = a+b+c+d. Utilizando dichas tablas algunas de las medidas de semejanza más utilizadas son:

Coeficiente de Jaccard: $\frac{d}{b + c + d}$

Coeficiente de acuerdo simple: $\frac{a + d}{p}$

Ambas toman valores entre 0 y 1 y miden, en tanto por uno, el porcentaje de acuerdo en los valores tomados en las p variables, existente entre los dos objetos. Difieren en el papel dado a los acuerdos en 0. El coeficiente de Jaccard no los tiene en cuenta y el de acuerdo simple. Ello es debido a que, en algunas situaciones, las variables binarias consideradas son asimétricas en el sentido de que es más informativo el valor 1 que el valor 0. Así, por ejemplo, si el color de los ojos de una persona se codifica como 1 si tiene los ojos azules y 0 en caso contrario. En éste tipo de situaciones es más conveniente utilizar coeficientes tipo Jaccard.

c) Medidas para datos nominales y ordinales

Una generalización de las medidas anteriores viene dada por la expresión:

$$S_{rs} = \sum_{k=1}^{p} S_{rsk} \qquad \textbf{(2.4)}$$

donde s_{rsk} es la contribución de la variable k-ésima a la semejanza total. Dicha contribución suele ser de la forma $1-d_{rsk}$ donde d_{rsk} es una distancia que suele tener la forma $\delta_{k\ell m}$ siendo ℓ el valor del estado de la variable X_k en el r-ésimo objeto y m el del s-ésimo objeto.

En variables nominales suele utilizarse $\delta_{k\ell m}$ = 1 si ℓ =m y 0 en caso contrario. En variables ordinales suele utilizarse medidas de la forma $|\ell-m|^r$ con r>0.

2.3.4.3 Medidas de distancia para variables cuantitativas, tablas de frecuencias, datos binarios y tipo mixto

a) Medidas para variables cuantitativas

1) Distancia euclidiana

$$\sqrt{\sum_{j=1}^{p}\left(x_{rj}-x_{sj}\right)^2} \qquad \textbf{(2.5)}$$

2) Distancia euclidiana al cuadrado

$$\sum_{j=1}^{p} \left(x_{rj} - x_{sj} \right)^2 \qquad \textbf{(2.6)}$$

3) Distancia métrica de Tchebychev

$$\max_{i} \left| x_{ri} - x_{si} \right| \qquad \textbf{(2.7)}$$

4) Distancia de Manhattan

$$\sum_{i=1}^{p} \left| x_{ri} - x_{si} \right| \qquad \textbf{(2.8)}$$

5) Distancia de Minkowski

$$\sqrt[q]{\sum_{i=1}^{p} (x_{ri} - x_{si})^q} \quad \text{con } q \in \mathbf{N}. \qquad \textbf{(2.9)}$$

Las tres primeras medidas son variantes de la distancia de Minkowski con q=2, ∞ y 1, respectivamente. Cuanto mayor es ***q*** más énfasis se le da a las diferencias en cada variable.

Todas estas distancias no son invariantes a cambios de escala por lo que se aconseja estandarizar los datos si las unidades de medida de las variables no son comparables. Además, no tienen en cuenta las relaciones existentes entre las variables.

Si se quieren tener en cuenta se aconseja utilizar la **distancia de Mahalanobis** que viene dada por la forma cuadrática:

$$(x_r - x_s)' S^{-1} (x_r - x_s)$$

donde $x_r = (x_{r1},\ldots,x_{rp})'$ y $x_s = (x_{s1},\ldots,x_{sp})'$

b) Medidas para tablas de frecuencias

Suelen estar basadas en la χ^2 de Pearson. Algunas de las más utilizadas son:

$$\chi^2 = \sqrt{\sum_{i=1}^{p} \frac{(X_{ri} - E(X_{ri}))^2}{E(X_{ri})} + \sum_{i=1}^{p} \frac{(X_{si} - E(X_{si}))^2}{E(X_{si})}} \qquad \textbf{(2.10)}$$

$$\varphi^2 = \sqrt{\sum_{i=!}^{p} \frac{(X_{ri} - E(X_{ri}))^2}{E(X_{ri})} + \sum_{i=!}^{p} \frac{(X_{si} - E(X_{si}))^2}{E(X_{si})}} \qquad \textbf{(2.11)}$$

donde $E(X_{ri}) = \frac{X_r X_i}{N}$ *con* $X_r = \sum_{i=1}^{p} X_{ri}$ *y* $X_i = X_{ri} + X_{si}$ es el valor esperado de la frecuencia x_{ri} si hay independencia entre los individuos r y s y las categorías 1,...,p de las variables y $N = x_{r.} + x_{s.}$ es el total de observaciones. La diferencia entre ambas medidas radica en la división por N en el caso de φ^2 para encubrir la dependencia que tiene la χ^2 de Pearson respecto a N.

c) Medidas para datos binarios

Distancia euclidiana al cuadrado: $(b+c)^2$

Lance y Williams: $\frac{b + c}{2d + b + c}$

Esta última ignora las concordancias en 0.

d) Medidas para datos de tipo mixto

Si en la base de datos existen diferentes tipos de variables: binarias, categóricas, ordinales, cuantitativas no existe una solución universal al problema de cómo combinarlas para construir una medida de distancia. Anderberg (1973) o Gordón (1990) sugieren las siguientes soluciones:

- Expresar todas las variables en una escala común, habitualmente binaria, transformando el problema en uno de los ya contemplados anteriormente. Esto tiene sus costes, sin embargo, en términos de pérdida de información si se utilizan escalas menos informativas como las nominales u ordinales o la necesidad de incorporar información extra si se utilizan escalas más informativas como son los intervalo o razón.

- Combinar medidas con pesos de ponderación mediante expresiones de la forma:

$$d_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} d_{ijk}}{\sum_{k=1}^{p} w_{ijk}} \qquad \textbf{(2.12)}$$

donde d_{ijk} es la distancia entre los objetos i y j en la k-ésima variable y w_{ijk} = 0 ó 1 dependiendo de si la comparación entre i y j es válida en la k-ésima variable

2.3.4.4 Medidas de correlación

Las medidas de correlación representan la similitud mediante la correspondencia de patrones entre las características (variables). Es decir, que las correlaciones representan patrones para todas las variables más que las magnitudes. Las medidas de correlación, sin embargo, se utilizan en raras ocasiones porque el interés de la mayoría de las aplicaciones del análisis de clúster está en las magnitudes de los individuos, y no en los patrones de los valores.

CAPÍTULO 3. COEFICIENTES DE DEPENDENCIA, DE SEMEJANZA Y DISTANCIA

En este capítulo se describirá un grupo de medidas de semejanza y distancias, indicando en cada caso la naturaleza de las variables para la cual deben ser utilizadas.

En las secciones siguientes, la asociación será utilizada como término general para describir la medida o el coeficiente usado para cuantificar la semejanza o la diferencia entre los objetos o los descriptores, según lo propuesto por Orlóci (1975). Con los ***coeficientes de dependencia***, se usa en el **modo R**, al cero le corresponde la no asociación. En estudios del **modo Q**, los ***coeficientes de semejanza o similaridad*** entre objetos será diferenciado de los ***coeficientes de distancia (o desemejanza).*** Las similaridades o semejanzas son ***máximas*** cuando los dos objetos son idénticos y ***mínimas*** cuando dos objetos son totalmente diferentes; mientras que las distancias siguen el sentido opuesto.

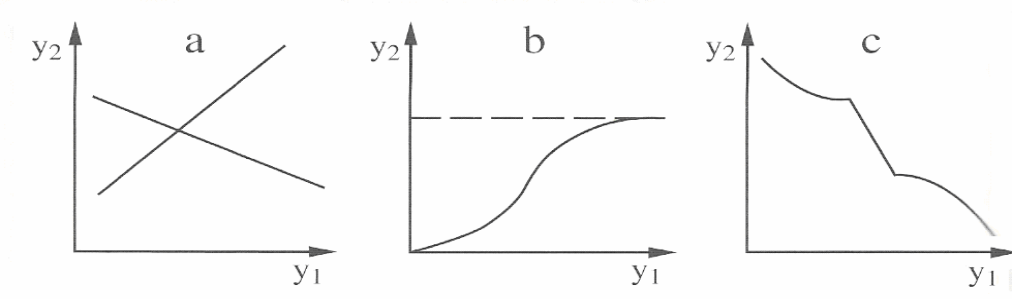


Figura 3.1 Relaciones monotónicas. En esta figura se muestran tres tipos de relaciones monotónicas entre dos descriptores: (a) lineal (aumentando y disminuyendo): logístico (aumento monotónico): (c) anormal (disminución monotónica).

En la figura 3.1 se muestra claramente la diferencia entre los dos tipos de medidas: la longitud de la línea entre dos objetos es una medida de su distancia, mientras que su grosor, que disminuye cuando los dos objetos se alejan, es proporcional a su similaridad. Si es necesario, una semejanza se puede transformar en una distancia, por ejemplo, calculando su complemento. Para una medida de semejanza que varía entre 0 y 1, como generalmente es el caso, la distancia correspondiente se puede calcular como:

$$D = 1 - S \qquad D = \sqrt{1 - S}\ , \qquad D = \sqrt{1 - S^{\,2}} \tag{3.1}$$

Las distancias, que en algunos casos no están acotadas por un valor superior predeterminado, se pueden normalizar, utilizando las ecuaciones.

$$y_i' = \frac{y_i}{y_{\max}} \quad , y_i' = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \tag{3.2}$$

$$D_{norm} = \frac{D}{D_{\max}} \quad o \quad D_{norm} = \frac{D - D_{\min}}{D_{\max} - D_{\min}} \tag{3.3}$$

Donde D_{norm} es la distancia normalizada entre [0, 1] mientras que D_{max} y D_{min} son los valores máximos y mínimos tomados por el coeficiente de la distancia, respectivamente. Las distancias normalizadas se pueden utilizar para calcular semejanzas, invirtiendo las transformaciones dadas anteriormente:

$$S = 1 - D^{\,2}{}_{norm'} \quad S = \sqrt{1 - D^{2}{}_{norm}} \ \ ó \ \ S = \sqrt{1 - D^{\,2}{}_{norm}}$$

3.1 Modo Q: Coeficientes de semejanza

El grupo más amplio de coeficientes en la literatura es el de las semejanzas. Estos coeficientes se utilizan para medir la asociación entre los objetos. En contraste con la mayoría de los coeficientes de distancia, las medidas de semejanza no son nunca métricas, puesto que siempre es posible encontrar dos objetos, A y B, que son más similares que la suma de sus semejanzas con otro más distante: el objeto C. De esto, se concluye que las semejanzas no se pueden utilizar directamente para situar objetos en un espacio métrico; sino que deben ser convertidas en distancias. Los métodos de clasificación, por otra parte, se pueden realizar tanto con una matriz de distancias como con una de semejanzas.

Los coeficientes de semejanza primero fueron desarrollados para datos binarios, es decir, datos de presencia-ausencia, o respuestas a las preguntas sí-no. Fueron generalizados más tarde a los descriptores multiestado, cuando las computadoras lo hicieron posible. Otra dicotomía importante entre los coeficientes de semejanza está relacionada con el tratamiento de los **doble ceros** o parejas de valores negativos.

3.1.1 Problema del doble-cero

El problema del doble-cero tiene gran importancia debido a la naturaleza especial de muchos descriptores. Muchas veces las casillas de la matriz contienen eventos que son escasos. Si una casilla está presente para dos sitios diferentes, ésta es una indicación de la semejanza de estos sitios; pero si una casilla aparece como ausente para dos sitios diferentes, esto puede ser porque los dos sitios están ambos sobre el valor óptimo para esas especies, o porque ambos están por debajo del óptimo;

más aún porque un sitio está por encima y el otro por debajo del valor óptimo para esa especie, y por tanto, no se puede decir cuál de estas circunstancias es la correcta.

Es así preferible abstenerse de sacar cualquier conclusión de la ausencia de datos en una casilla para dos sitios diferentes. En términos numéricos, esto significa no considerar los doble-cero al calcular coeficientes de semejanza o distancias usando datos de presencia-ausencia.

Los coeficientes de este tipo se llaman ***asimétricos*** porque tratan los ceros de manera diferente que el resto de los valores. Por otra parte, la presencia de un individuo en uno de los dos sitios y su ausencia en el otro sitio es considera como la diferencia entre estos sitios.

En coeficientes ***simétricos***, el estado cero para dos objetos se trata exactamente de la misma manera que cualquier otro par de valores, al calcular una semejanza. Estos coeficientes se pueden utilizar en los casos donde el estado cero es una base válida para comparar dos objetos o individuos y representa la misma clase de información que cualquier otro valor. Esto excluye, obviamente, el caso especial donde el cero significa "carencia de información".

Pueden existir varias razones para que un sujeto se encuentre ausente en un sitio. Entre otras, la ausencia puede también ser el resultado del patrón de la dispersión de un individuo, de acontecimientos históricos, o, más simplemente, de una variación estocástica del muestreo.

Ocurre a menudo que un gran número de individuos están presentes en la matriz de datos, y en cambio, hay varios sitios donde solamente se encuentra un pequeño número de especies, no existe una regularidad en la presencia de individuos de lugar a lugar. El considerar los doble-cero en la comparación entre lugares, daría altos valores de similaridad o semejanza para las parejas de lugares que tienen pocos individuos, lo que no reflejaría la situación adecuadamente. Es por esto que en muchas ocasiones es preferible utilizar coeficientes asimétricos para los cuales las dobles ausencias no se cuentan como semejanzas.

El resto de esta sección distingue entre los coeficientes binarios y cuantitativos de la semejanza y, para cada tipo, los que utilizan doble-ceros o los excluyen. Termina con una descripción de coeficientes probabilísticas

3.1.2 Coeficientes binarios simétricos

En los casos más simples, la semejanza entre dos sitios se basa en datos de presencia-ausencia. Los descriptores binarios pueden describir la presencia o la ausencia de condiciones ambientales o de individuos. Las observaciones pueden ser resumidas en la siguiente tabla de frecuencias 2 x 2:

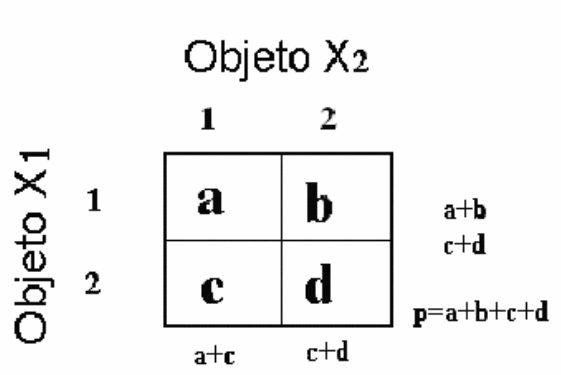


Figura 3.2 Tabla de frecuencia 2 x 2

donde “a” es el número de descriptores para los cuales los dos objetos se han codificado como 1, d es el número de los descriptores de dos objetos codificados como 0; mientras que b y c son los números de los descriptores para los cuales los dos objetos se codifican de manera diferente; y p es el número total de descriptores.

Una manera obvia de calcular la semejanza entre dos objetos es contar el número de descriptores que codifican los objetos de la misma manera y dividir esta cantidad por el número total de descriptores:

$$S_1^*(x_1, x_2) = \frac{a + d}{p} \quad \textbf{(3.4)}$$

El coeficiente S_1* se llama el coeficiente de las parejas simple (Sokal y Michener, 1958). Al usar este coeficiente, se asume que no hay diferencia entre el doble-0 y el doble-1. Este es el caso, por ejemplo, cuando cualquiera de los dos estados de cada descriptor se podría codificar 0 o 1 indistintamente. Una variante de esta medida es el coeficiente de Rogers y Tanimoto (1960) en el cual a las diferencias se les da más peso que a las semejanzas:

$$S_2(x_1, x_2) = \frac{a + d}{a + 2b + 2c + d} \quad \textbf{(3.5)}$$

Sokal y Sneath (1963) propusieron cuatro medidas más que incluyen los doble-cero, que además, tienen sus contrapartes en cuatro coeficientes que excluyen los doble-ceros:

$$S_3(x_1, x_2) = \frac{2a + 2d}{2a + b + c + 2d} \quad \textbf{(3.6)}$$

que cuenta las semejanzas con el doble de la importancia de las diferencias;

$$S_4(x_1, x_2) = \frac{a + d}{b + c} \tag{3.7}$$

que compara las semejanzas con las diferencias, en una medida que va desde 0 al infinito:

$$S_5(x_1, x_2) = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right] \tag{3.8}$$

para la comparación de las semejanzas con los totales marginales;

$$S_6(x_1, x_2) = \frac{a}{\sqrt{(a+b)*(a+c)}} \frac{d}{\sqrt{(b+d)(c+d)}} \tag{3.9}$$

que es el producto de las medias geométricas de los términos concernientes **a** y **d**, respectivamente, en el coeficiente S_5.

Entre los coeficientes anteriores, S_1 hasta S_3 son los de interés más general, pero los otros pueden ser útiles para el tratamiento de descriptores especiales.

Existen tres medidas adicionales que están disponibles en diferentes programas computacionales como de NT-SYS:el coeficiente de Hamman:

$$S = \frac{a + d - b - c}{p} \tag{3.10}$$

el coeficiente de Yule:

$$S = \frac{ad - bc}{ad + bc} \tag{3.11}$$

y el coeficiente de Pearson's Φ (phi):

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{3.12}$$

donde el numerador es el determinante de la tabla de frecuencias 2 x 2. Como ϕ es realmente la raíz cuadrada del estadístico $\chi 2$ (Chi-cuadrado) para tablas de 2 x 2. Los coeficientes de este tipo se utilizan sobre todo en análisis del **modo R**. Estos últimos índices se describen detalladamente en Sokal y Sneath (1963).

3.1.3 Coeficientes binarios asimétricos

Los coeficientes que se corresponden con los anteriores están disponibles para comparar sitios usando matrices de datos de presencia-ausencia de individuos, cuando la comparación debe excluir los doble-cero.

La mejor medida conocida es el coeficiente de Jaccard (1900, 1901, 1908), conocido también como el coeficiente de comunidad, o simplemente, coeficiente de Jaccard:

$$S_7(x_1, x_2) = \frac{a}{a + b + c} \tag{3.13}$$

en el cual todos los términos tienen igual peso. Como una variante, el coeficiente de Sorensen (1948) nos ofrece una doble ponderación para presencias dobles:

$$S_8(x_1, x_2) = \frac{2a}{2a + b + c} \qquad \textbf{(3.14)}$$

ya que es posible considerar que la presencia de una especie es más informativa que su ausencia. La ausencia puede ser debida a varios factores, según lo discutido arriba y no refleja necesariamente diferencias en el ambiente. La doble-presencia, por el contrario, es una indicación fuerte de la semejanza. Se puede notar, sin embargo, que S_8 es monotónica a S_7.

Esta propiedad significa que si la semejanza para un par de objetos calculado con S_7 es más alto que para otro par de objetos, se obtiene el mismo resultado utilizando S_8. Es decir, S_7 y S_8 solamente se diferencian por sus escalas. Anteriormente, Sorensen y Dice (1945) habían utilizado S_8 bajo el nombre de **"índice de coincidencia"** en un estudio del **modo R** para asociaciones de especies.

La versión de la distancia para este coeficiente, $D_{l3} = 1 - S_8$ es una semimétrica. Una consecuencia es que el análisis de coordenadas principales de una matriz de semejanzas S_8 ó D_{l3} es probable que produzca valores negativos La manera más fácil de resolver este problema es basar el análisis de coordenadas principales en distancias de raíces cuadradas transformadas:

$$D \sqrt{1 - S_8} \qquad en \ \ lugar \ \ de \ : \ \ D = 1 - S_8$$

Otra variante de S_7 ofrece una triple ponderación a las presencias dobles:

$$S_9(x_1, x_2) = \frac{3a}{3a + b + c} \quad \textbf{(3.15)}$$

La contraparte del coeficiente de Rogers y Tanimoto (S_2), fue propuesto por Sokal y Sneath (1963). Este coeficiente ofrece una doble ponderación a las diferencias en el denominador:

$$S_{10}(x_1, x_2) = \frac{a}{a + 2b + 2c} \quad \textbf{(3.16)}$$

Russell y Rao (1940) sugirieron una medida que permite la comparación del número de presencias dobles, en el numerador, con el número total del objeto encontrado en todos los sitios, incluyendo las especies que están ausentes (d) de las parejas de sitios considerados:

$$S_{11}(x_1, x_2) = \frac{a}{p} \quad \textbf{(3.17)}$$

Kulczynski (1928) propuso un coeficiente opuesto para las doble-presencias de las diferencias:

$$S_{12}(x_1, x_2) = \frac{a}{b + c} \quad \textbf{(3.18)}$$

Entre los coeficientes para los datos de presencia-ausencia, Sokal y de Sneath (1963) mencionan la versión binaria del coeficiente S_{18} de Kulczynski para datos cuantitativos:

$$S_{13}(x_1, x_2) = \frac{1}{2}\left[\frac{a}{a + b} + \frac{a}{a + c}\right] \quad \textbf{(3.19)}$$

Donde las doble - presencias se comparan a los totales marginales (a + b) y (a + c).

Ochiai (1957) utilizó, como medida de semejanza, la media geométrica de las razones de **a** con respecto al número de individuos en cada sitio; es decir, los totales marginales (a + b) y (a + c):

$$S_{14}(x_1, x_2) = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} = \frac{a}{\sqrt{(a+b)(a+c)}} \qquad \textbf{(3.20)}$$

Esta medida es igual que S_6 excepto en lo que concierne a los doble-ceros (d)

Faith (1983) sugirió el coeficiente siguiente, en el cual las des-concordancia (presencia en un sitio y ausencia en el otro) se consideran con una ponderación opuesta al doble de las presencias. El valor de S_{26} disminuye cuando aumenta el número de doble-ceros:

$$S_{26}(x_x, x_2) = \frac{a + d/2}{p} \qquad \textbf{(3.21)}$$

3.1.4 Coeficientes cuantitativos simétricos

Los descriptores tienen a menudo más de dos estados. Los coeficientes binarios descritos anteriormente, se pueden ampliar a descriptores multi-estado. Por ejemplo, el coeficiente de las parejas simple puede ser utilizado con descriptores multi-estado de la siguiente forma:

$$S_1(x_1, x_2) = \frac{Concordancias}{p} \qquad \textbf{(3.22)}$$

Donde el numerador contiene el número de descriptores para los cuales los dos objetos están en el mismo estado. Por ejemplo, si un par de objetos fue descrito por los 10 descriptores multi-estado siguientes:

Tabla 3.1 Ejemplo de concordancias.

	Descriptores
Objetos X_1	9 3 7 3 4 9 5 4 0 6
Objetos X_2	2 3 2 1 2 9 3 2 0 6
Concordancia	0+1+0+0+0+1+0+0+1+1

= 4

Fuente:(Legendre 1986)

El valor del S_1 calculado para los 10 descriptores multi-estado sería:

$$S_1(x_1, x_2) = 4\ Concordancias/10\ descriptores = 0.4$$

Es posible extender de la misma manera el uso de todos los coeficientes binarios a descriptores multi-estado. Sin embargo, los coeficientes de este tipo a veces dan lugar a una pérdida de información valiosa, especialmente en el caso de descriptores ordenados para los cuales dos objetos pueden compararse basándose en la ***cantidad que expresa la diferencia*** entre los estados.

Gower (1971ª) propuso un coeficiente general de semejanza que combina diferentes tipos de descriptores y procesa cada uno de acuerdo con su propio tipo matemático. Aunque la descripción de este coeficiente puede parecer algo complicada, puede ser fácilmente resuelta con un pequeño programa de computación. El coeficiente inicialmente se expresa de la siguiente forma:

$$S_{15}(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} S_{12j} \qquad \textbf{(3.23)}$$

La semejanza entre dos objetos es el promedio, sobre los "p" descriptores de las semejanzas calculadas para todos los descriptores. Para cada descriptor j, el ***valor de la semejanza parcial*** $S_{12\,j}$ entre los objetos x_1 y x_2 se calcula de la siguiente forma:

- Para los descriptores binarios, S_j = 1 (acuerdo) u O (desacuerdo). Gower propuso dos formas para este coeficiente. La forma utilizada aquí es simétrica, dando S_j = 1 a los ceros dobles. La otra forma, utilizada en el coeficiente asimétrico de Gower S_{I9}. da S_j = 0 para los doble-ceros.
- Los descriptores cualitativos y semi-cuantitativos son tratados siguiendo la siguiente regla de parejas establecida anteriormente: S_j = 1 cuando hay acuerdo y S_j = 0 cuando hay desacuerdo. Los doble-cero se tratan como en el párrafo anterior.
- Los descriptores cuantitativos (números reales) se tratan de una manera interesante. Para cada descriptor, primero se calcula la diferencia entre los estados de los dos objetos $\left| Y_{1j} - Y_{2j} \right|$ como en el caso de los coeficientes de distancia que pertenecen al grupo métrico de Minkowski.

Este valor entonces es dividido por la diferencia más grande (R_j) encontrada para este descriptor a través de todos los sitios en el estudio (o, si uno prefiere, utilizando una población de referencia). Puesto que este cociente es realmente una distancia normalizada, se resta de 1 para transformarla en una semejanza:

(3.24)

$$S_{12} = 1 - \left[\left| Y_{1j} - Y_{2j} \right| / R_j \right]$$

El coeficiente de Gower se puede programar incluyendo un elemento adicional de flexibilidad: No se calcula ninguna comparación para los descriptores donde falte información para uno o para el otro objeto. Esto es obtenido mediante un valor $W_{j,}$ conocido como la delta de Kronecker, que describe la presencia o la ausencia de la información: Wj = 0 cuando la información sobre Y_j falta para un u otro objeto, o ambos; W_j =1cuando la información está presente para ambos objetos. La forma final del coeficiente de Gower es la siguiente:

$$S_{15}(x_1, x_2) = \frac{\sum_{j=1}^{p} W_{12\ j} S_{12\ j}}{\sum_{j=1}^{p} W_{12\ j}} \qquad \textbf{(3.25)}$$

El coeficiente S_{15} produce valores de semejanza entre 0 y 1 (semejanza máxima)

Un último aspecto de la complejidad, que no fue sugerida en el artículo de Gower pero se agrega aquí, consiste en ponderar a varios descriptores. En vez de 0 o de 1, uno puede asignar a W_j un valor entre O y 1 que corresponde al peso que uno desea que cada descriptor tenga en el análisis. Los descriptores con pesos cercanos a O contribuyen poco al valor final de la semejanza, mientras que los descriptores con pesos más altos (más cercanos a 1) contribuyen más. Dar un peso de 0 a un descriptor es equivalente a quitarlo del análisis. Un valor que falta automáticamente cambia el peso W_j a 0.El ejemplo numérico siguiente ilustra el cálculo del coeficiente S_{15}.

En el ejemplo, se describen dos sitios por ocho valores ambientales cuantitativos de los descriptores. Los valores R_j (rango de valores entre todos los objetos, porque cada descriptor $y_{j)}$ dados en la tabla se han calculado para la base de datos entera antes de calcular el coeficiente S_{15}. Los pesos W_{12j} se utilizan solamente en este ejemplo para eliminar descriptores con valores faltantes (función delta de Kronecker):

Tabla 3.2 Función delta de Kronecker calculo del coeficiente S_{15}.

Descriptores j									**Suma**
Objeto X_1	2	2	-	2	2	4	2	6	
Objeto X_2	1	3	3	1	2	2	2	5	
W_{12j}	1	1	0	1	1	1	1	1	=7
R_J	1	4	2	4	1	3	2	5	
$\mathbf{I\ Y_{Ij} - Y_{2j}\ I}$	1	1	-	1	0	2	0	1	
W_{12j}	1	0.25	-	0.25	0	0.67	0	0.2	
S_{12j}	0	0.75	0	0.75	1	0.33	1	0.8	= 4.63
Así, S_{15} (X1 , X2) = 4.63 / 7 = 0.66									

Fuente:(Legendre 1983:260)

Cuando se calcula S_{15}, se puede decidir el manejo de descriptores semi cuantitativos como si fueran cuantitativos, para utilizar diferencias entre los estados en el gravamen final de la semejanza. Es importante en tales casos cerciorarse de que las distancias entre los estados adyacentes son comparables en magnitud. Por ejemplo, con descriptores ordenados (semi cuantitativos) codificados del 1 al 3, $\mathbf{I\ Y_{1\ j} - Y_{2j}\ I}$ puede ser utilizado solamente si la diferencia entre los estados 1 y 2 se puede pensar que es casi igual a la que hay entre los estados 2 y 3. Si hay demasiada diferencia, los valores $\mathbf{I\ Y_{1\ j} - Y_{2j}\ I}$ no son comparables y los descriptores semi-cuantitativos no se deben utilizar de esa manera en el coeficiente S_{15}.

Otro coeficiente general de semejanza fue propuesto por Estabrook y Rogers (1966). La semejanza entre dos objetos es, como en S_{15}, la suma de las semejanzas parciales por descriptores, dividida por el número de los descriptores para los cuales hay información para los dos objetos. En la publicación original, los autores utilizaron el estado 0 para significar "ninguna información disponible", pero cualquier otra convención sería aceptable. La forma general de este coeficiente es por lo tanto igual que el coeficiente de Gower:

$$S_{16}(x_1, x_2) = \frac{\sum_{J=1}^{P} w_{12j} s_{12J}}{\sum_{j=1}^{p} w_{12j}} \qquad \textbf{(3.26)}$$

Como en S_{15}, los parámetros de W_j se pueden utilizar como pesos (entre 0 y 1) en vez solamente de desempeñar las funciones de las deltas de Kronecker. El coeficiente de Estabrook y de Rogers se diferencia de S_{15} en el cálculo de las semejanzas parciales S_j.

En el artículo de Estabrook y de Rogers (1966), los valores del estado eran enteros positivos y los descriptores eran u ordenados o no ordenados. La semejanza parcial entre dos objetos para un descriptor dado *j* se calcula usando una función monotónicamente decreciente de la semejanza parcial. Sobre una base empírica, y entre todas las funciones de este tipo, los autores propusieron utilizar la función siguiente de dos números ***d*** y ***k***:

$$S_{12j} = f(d_{12j,}, k_j) = 0 \qquad cuando \qquad d > k$$
$$S_{12j} = f(d_{12j}, k_j) = 2(k + 1 - d) \; cuando \quad d \leq k$$

donde ***d*** es la distancia entre los estados de los dos objetos X_1 y X_2 para el descriptor *j*, es decir, el mismo valor $|Y_{1j} - Y_{2j}|$ que en el coeficiente de Gower, y ***k*** es un parámetro determinado a priori por los usuarios para cada descriptor, describiendo hasta donde se permite que lleguen las semejanzas parciales no nulas. El parámetro ***k*** es igual a la diferencia más grande d para la cual la semejanza parcial S_{12j} (para el descriptor *j*) se permite ser diferente de 0. Los valores de ***k*** para diferentes descriptores pueden ser bastante diferentes unos de los otros. Por ejemplo, para un descriptor codificado de 1 a 4, podemos decidir utilizar ***k*** = 1, y para otro descriptor con valores codificados del 1 al 50, se puede utilizar ***k*** = 10.

Para entender completamente la función de semejanza parcial S_{12J}, se puede calcular a mano S_{12j} para algunos descriptores en el ejemplo numérico siguiente. Los valores de k, que son generalmente números pequeños, se dan para cada descriptor en la siguiente tabla:

Tabla 3.3 Valores tomados por la función parcial de la semejanza para los primeros valores de k que se dan en la tabla 3.1

		Descriptores j					**S_{16} (X_1, X_2)**
Objeto X_1	2	1	3	4	2	1	
Objeto x_2	2	2	4	3	2	3	
K_j	1	2	1	2	1	1	
	▼	▼	▼	▼	▼	▼	
$S_{12j} = f(d_{12j}, k_j)$	1	+0	+0.04	+0.5	+1	+0	=2.9 / 6=0.483

Fuente: (Legendre 1983: 260)

Los valores en la tabla demuestran que, si k = 0 para todos los descriptores, S_{16} es idéntica al coeficiente de las parejas simple para los descriptores multi-estado.

Tabla 3.4 Valores de la función de similaridad parcial f(d,k) para los coeficientes S_{16} y S_{20} para algunos valores de k

k		**d**						
	0	1	2	3	4	5	6	7
0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	1	0.40	0.00	0.00	0.00	0.00	0.00	0.00
2	1	0.50	0.20	0.00	0.00	0.00	0.00	0.00
3	1	0.55	0.28	0.12	0.00	0.00	0.00	0.00
4	1	0.57	0.33	0.18	0.08	0.00	0.00	0.00
5	1	0.59	0.36	0.22	0.13	0.05	0.00	0.00

Fuente: (Legendre & Rogers, 1972: 594)

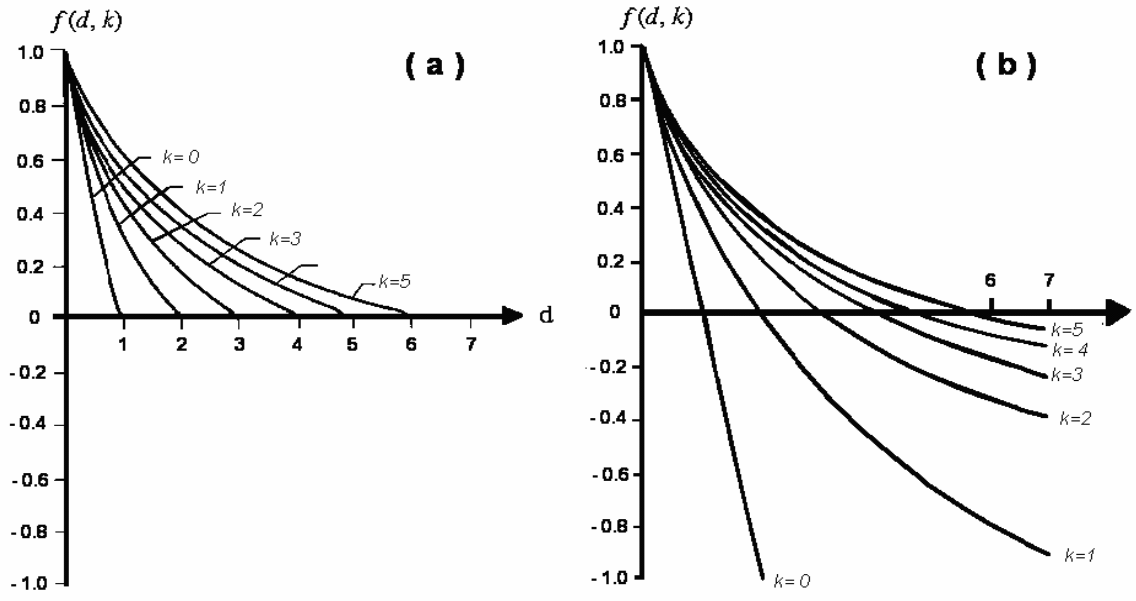


Figura 3.3 Coeficientes S_{16} y S_{20}: cambio en f (d, k) en función de d, para seis valores de k, (a) bajo f (d de la condición, k) = 0 cuando k; (b) sin esta condición. Adaptaciones de Legendre y de Chodorowski (1977).

3.1.5 Coeficientes cuantitativos asimétricos

En las subdivisiones anteriores se analizaron los descriptores multi-estado del coeficiente S_1; de la misma manera, los coeficientes binarios se pueden extender para acomodar descriptores multi-estado. Por ejemplo, el coeficiente de Jaccard se convierte en:

$$S_7 = (x_1, x_2) = \frac{Concordancias}{p - doble\ ceros} \qquad \textbf{(3.27)}$$

donde en el numerador se encuentran los individuos con la misma categoría en los dos sitios. Esta forma puede ser utilizada cuando las categorías de los individuos se codifican en un número pequeño de clases y se desea poner en contraste fuertemente las diferencias en dichas categorías. En otros casos, usar tal coeficiente causaría obviamente la pérdida de parte de la información ocasionada por las categorías de los individuos.

Otras medidas son más eficientes que la eq. 3.27 al usar la información de las categorías de los individuos. Se dividen en dos categorías: ***los coeficientes para los datos escasos (raw data) y las medidas para datos normalizados***.

- Como es conocido, la distribución de las categorías de un individuo a través de un gradiente a menudo se sesga fuertemente, de modo que la normalización de las categorías de los individuos exige, a menudo, transformaciones como la raíz cuadrada, la doble raíz cuadrada o las logarítmicas. Otra manera de obtener datos aproximadamente normales es utilizar una escala de categorías relativos con los límites que forman una progresión geométrica, por ejemplo una escala de 0 (ausente) a 7 (muy abundante).

- Las categorías así normalizadas reflejan el papel de cada individuo en el sistema mejor que los datos crudos de cada categoría, puesto que un descriptor representado por 100 individuos en un sitio no tiene un papel 10 veces tan importante en el equilibrio del sistema como otro descriptor representado por 10 individuos. Lo primero es quizás dos veces más importante que lo último ya que este es el cociente obtenido después de aplicar una transformación logarítmica base-10 (si se asume que los números 100 y 10 en un sitio son representativos de la verdadera categoría relativa en la población).

Algunos coeficientes disminuyen el efecto de las diferencias más grandes y se pueden por lo tanto utilizar con categorías escasas (raw data) de un descriptor, mientras que otros comparan los diversos valores de la categoría de una manera más lineal y son así mejores adaptados a los datos normalizados.

Entre el grupo de coeficientes que se utilizarán con categorías escasas de un descriptor, el más conocido es un coeficiente atribuido al matemático polaco H. Steinhaus descrito por Motyka (1947). Esta medida ha sido redescubierta muchas veces; su uno-complemento se conoce como el coeficiente de Odum Bray -Curtis (D_{14}). Se atribuye a veces incorrectamente al antropólogo Czekanowski (1909 y 1913). El coeficiente de la distancia de Czekanowski se describe en la (D_8) Este coeficiente compara dos sitios (x_1, x_2) en función de la categoría mínima de cada descriptor:

$$S_{17} = (x_1, x_2) = \frac{W}{(A+B)/2} = \frac{2W}{(A+B)} \qquad \textbf{(3.28)}$$

dónde W es la suma de las categorías mínimas de los diferentes descriptores. Este mínimo es definido como la categoría en el sitio donde los descriptores apenas se encuentran. A y B son las sumas de las categorías de todos los descriptores en cada uno de los dos sitios o, en otras palabras, el número total de los individuos observados en cada sitio, respectivamente. Considere el ejemplo numérico siguiente:

Tabla 3.5 Comparativa de 2 sitios (X_1, X_2) en función de la categoría mínima de cada descriptor

	Especies de abundancia						*A*	*B*	*W*
Sitio X1	7	3	0	5	0	1	16		
Sitio X2	2	4	7	6	0	3		22	
Mínimo	2	3	0	5	0	1			11

Fuente: (Legendre, 1983: 265)

$$S_{17} = (x_1, x_2) = \frac{2\,x11}{16 + 22} = .579$$

Esta medida se relaciona estrechamente con el coeficiente de Sorensen (S_8): si se utilizan datos de presencia-ausencia en vez de conteos o frecuencias del descriptor, S_{17} se convierte en S_8.

La versión de la distancia de este coeficiente, $D_{14} = 1 - S_{17}$ es una semimétrica, como se demuestra en el ejemplo que sigue a la ecuación (D_{14}). Una consecuencia es que el análisis de coordenadas principales de la matriz de semejanzas S_{17} o de las distancias D_{l4} es probable que produzca valores negativos. Una de las soluciones posibles es utilizar el análisis de coordenadas principales con las distancias transformadas a raíz cuadrada $D=\sqrt{1-S_{17}}$ en lugar de $D=1-S_{17}$

El coeficiente de Kulczynki también pertenece al grupo de las medidas que cuentan con categorías de datos escasos. La suma de los mínimos se compara primero con el gran total en cada sitio; entonces los dos valores se promedian:

$$S_{18}(x_1, x_2) = \frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right) \qquad \textbf{(3.29)}$$

Para los datos de presencia-ausencia, S_{18} se convierte en S_{13}. Utilizando el ejemplo numérico anterior, el coeficiente S_{18} se calcula como sigue:

$$S_{18}(x_1, x_2) = \frac{1}{2}\left(\frac{11}{16} + \frac{11}{22}\right) = 0.594$$

Los coeficientes S_{17} y S_{18} producen siempre valores entre 0 y 1, aunque Kulczynki (1928) multiplicó el valor final por 100 para obtener un porcentaje. El enfoque de Kulczynki, que consiste en calcular el promedio de dos comparaciones, parece más arbitrario que el método del Steinhaus, en el cual la suma de los mínimos se compara con la media de la suma de dos sitios. En la práctica, los valores de estos dos coeficientes son casi monotónicos.

Los coeficientes siguientes pertenecen al grupo adaptado a datos de categorías normalizadas; esto significa que tienen distribuciones de frecuencia insesgadas. Estos coeficientes son paralelos a los coeficientes S_{15} y S_{16} descritos anteriormente. Con respecto al coeficiente S_{19}, Gower (1971a) había propuesto inicialmente que su coeficiente general S_{15} debe excluir los doble-cero de la comparación.Lo que hace que este coeficiente sea muy útil para categorías de descriptores con datos cuantitativos, puesto que las diferencias entre los estados se calculan como $\mathbf{| Y_{1j} - Y_{2j} |}$ y están linealmente relacionados con la escala de la medida.

Este coeficiente se debe utilizar con datos previamente normalizados. La forma general es:

$$S_{19}(x_1, x_2) = \frac{\sum_{j=1}^{p} W_{12j} S_{12j}}{\sum_{j=1}^{p} W_{12j}} \qquad \textbf{(3.30)}$$

- $S_{12j} = \left[\left| \frac{y_{1j} - y_{2j}}{R_j} \right|\right]$ como en S_{15}

- $W_{12j} = 0$ cuando y_{1j} o y_{2j} = ausencia de información

- Mientras $W_{12j} = 1$ en todos los otros casos.

Con datos de las categorías de los descriptores, los valores de W_j se podrían hacer variar entre 0 y 1, como en el coeficiente S_{15}, para compensar los efectos selectivos del proceso de muestreo.

Legendre y Chodorowski (1977) propusieron un coeficiente general de semejanza que es paralelo a S_{16}. Esta medida utiliza una versión levemente modificada de la función parcial de semejanza f (d, k), o bien una matriz de semejanzas parciales. Como S_{20} calcula todas las diferencias d de la misma forma, independientemente de si corresponden a valores altos o bajos en la escala de categorías, es mejor utilizar esta medida con datos de categorías insesgados. La única diferencia entre S_{16} y S_{20} estriba en la manera en la cual se manejan los doble-cero.

La forma general del coeficiente es la suma de valores de semejanza parciales para todos los descriptores, dividido por el número total de descriptores en los dos sitios combinados:

$$S_{20}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{12\,j} s_{12\,j}}{\sum_{j=1}^{p} w_{12\,j}} \qquad \textbf{(3.31)}$$

- $S_{12\,j} = f(d_{j12}, k_{1\,j})$ $\begin{cases} = \dfrac{2(k+1-d)}{2k+2+dk} \text{ cuando } d \le k \\ = 0 \text{ cuando } d > k \\ = 0 \text{ cuando } y_{j1} \text{ o } y_{j2} = 0 \end{cases}$
- bien $S_{12\,j} = f(y_{1\,j}, y_{2\,j})$ dado por una matriz de semejanza parcial
- Y $w_{12j} = 0$ cuando. $S_{12j} = f(y_{1j}, y_{2j})$ o y_{2j} = ausencia de información, o cuando y_{12j} = ausencia de las categorías ($y_{1j} + y_{2j} = 0$)
- Mientras que $w_{12j} = 1$ en el resto de los casos. Además, W_{12j} puede recibir un valor entre 0 y 1, según lo explicado anteriormente para S_{19}.

En resumen, las propiedades de coeficiente S_{20} son las siguientes:

1. Cuando el dj es mayor que kj la semejanza parcial entre los sitios es $S_{12j} = 0$ para las categorías j

2. Cuando dj = 0, entonces $S_{12j} = 1$

3. f (d, k) disminuye cuando d aumenta, para una k dada

4. f (d, k) aumenta con el aumento de k, para una d dada;

5. Cuando es $Y_{1j} = 0$ ó $Y_{2j} = 0$, la semejanza parcial entre los sitios es $S_{12j} = 0$ para las categorías j, aun cuando d_{12j} no es mayor que el k_j

6. cuando $k_j = O$ para toda las categorías j, S_{20} es igual que el coeficiente de Jaccard (S_7) para los descriptores multi-estado.

El último coeficiente cuantitativo que excluye los doble-cero es llamado el coeficiente de similaridad χ^2; que no es más que el complemento de la métrica χ^2 (D_{15}).

$$S_{21}(x_1, x_2) = 1 - D_{15}(x_1, x_2) \qquad \textbf{(3.32)}$$

3.1.6 Coeficientes probabilísticos

Las medidas probabilísticas forman una categoría especial. Estos coeficientes están basados en la estimación estadística de la significación de la relación entre objetos.

Cuando los datos son conteos (frecuencias), una primera medida probabilística podría obtenerse calculando el estadístico Chi-cuadrado de Pearson $\left(X_p^2\right)$ entre parejas de sitios, en lugar de coeficiente de similaridad χ^2 (S_{21}) descrito anteriormente. El complemento de la probabilidad asociada con X_p^2 proporciona una medida probabilística de semejanza entre los sitios. El número de grados de libertad (ν) debe excluir el número de doble-ceros:

$$\nu = \left[\left(\sum_{j=1}^{p} w_{12\,j}\right) - 1\right] \qquad \textbf{(3.33)}$$

donde W_{12j} es la delta de Kronecker para cada descriptor j, como el coeficiente S_{19} de Gower. La semejanza probabilística de X_p^2 se define como:

$$S_{22} = 1 - prob\left[X_p^2\,(x_1, x_2)\right] \qquad \textbf{(3.34)}$$

Siguiendo un enfoque diferente. el coeficiente probabilístico de Goodall (1964, 1966a) considera la distribución de frecuencia de diferentes estados de cada descriptor en el conjunto total de objetos.

De hecho, es menos probable que dos sitios tengan los mismos descriptores raros que descriptores más frecuentes. En este caso, se le debe dar más importancia a los descriptores raros que a los descriptores frecuentes al estimar la semejanza entre los sitios.

El índice probabilístico de Goodall, que había sido desarrollado originalmente para la taxonomía, es especialmente útil en clasificaciones ecológicas, porque las categorías de las especies en diferentes sitios son funciones estocásticas (Sneath y Sokal, 1963: 141). Orlóci (1978) sugieren para utilizarlo para los agrupar sitios (**modo Q**). El índice también se ha utilizado para el **modo R**, agrupando especies e identificando asociaciones.

El coeficiente probabilística de Goodall se basa en las probabilidades de diferentes estados de cada descriptor. La medida de semejanza que resulta es en sí misma una probabilidad, nombrada el complemento de la probabilidad del parecido entre dos sitios es debido al azar.

El índice probabilístico, según lo formulado por Goodall (1966a), es una medida taxonómica general en la cual los descriptores binarios y cuantitativos se pueden utilizar juntos. El coeficiente según lo presentado aquí sigue las modificaciones de Orlóci (1978) y se limita a agrupar sitios basados en categorías abundantes.

También considera las observaciones hechas referente a los doble-cero. La medida que resulta es por lo tanto una simplificación del coeficiente original de Goodall, orientada al agrupamiento de sitios.

Los pasos computacionales son como sigue:

(a) Se calcula una medida de similaridad parcial S_j para todos los pares de sitios y para cada descriptor j. Como hay n sitios, el número de las semejanzas parciales Sj a calcular, para cada descriptor, es: n(n - 1)/2. Si se han normalizado las categorías de las especie, se puede elegir la medida de semejanza $S_{12} = 1 - \left[\left| y_1 j - y_{2j} \right| / R_j \right]$ del coeficiente S_{19} de Gower o la función S_{12j} del coeficiente S_{20}, que fueron descritos anteriormente. En todos los casos, los doble-cero deben ser excluidos. Esto se puede hacer multiplicando las semejanzas parciales Sj por la delta de Kronecker W_{12j} cuyo valor es 0 cuando ocurre un doble-cero. Para datos cuyas categorías son escasas debe utilizarse el coeficiente de semejanza parcial de Steinhaus (S_{17}), calculado para descriptores que aparecen solos una vez. El resultado de este primer paso es una matriz de semejanza parcial, que contiene tantas filas como objetos en la matriz de datos (p) y, $n(n-1)/2$ columnas de n, es decir, una columna para cada par de sitios.

(b) En una segunda tabla del mismo tamaño, para cada descriptor j y cada uno de los $n(n-1)/2$ pares de sitios, se calcula la proporción de los valores de la semejanza parcial que se corresponden con los descriptores j que son mayores o iguales a las semejanzas parciales de las parejas de sitios que han sido considerados. El valor Sj considerado se encuentra incluido en le calculo de la proporción cuanto mayor es la proporción, los dos sitios son menos similares con respecto a los descriptores dados.

(c) Las proporciones o probabilidades anteriores son combinadas en una matriz de semejanza sitio x sitio, utilizando el método de Fisher; es decir, calculando el producto π de las probabilidades relativas a la diferentes descriptores. Dado que ningunas de las probabilidades son iguales a 0, no hay problema para combinar estos valores, pero se debe asumir que las probabilidades de los diferentes descriptores son vectores independientes. Si existen correlaciones entre los descriptores, se puede utilizar, en lugar de los descriptores originales de las categorías (Orlóci, 1978: 62), matriz de puntuación de componentes extraída de un análisis de coordenadas principales o de un análisis de correspondencia: con los datos originales.

(d) Hay dos formas para definir el índice de semejanza de Goodall. En el primer enfoque, los productos π se ponen en orden creciente. A continuación se calcula de diferencia entre dos sitios como la proporción de los productos que son mayores o iguales al producto para el par de sitios considerados:

$$S_{23}(x_1, x_2) = \frac{\sum_{Pares} d}{n(n-1)/2}, \quad donde \quad \begin{cases} d = 1 \; si \;\; \Pi \geq \Pi_{12} \\ d = 0 \; si \;\; \Pi < \Pi_{12} \end{cases} \qquad \textbf{(3.35)}$$

(e) En el segundo enfoque el valor χ^2 que le corresponde a cada producto se calcula bajo la hipótesis de que las probabilidades de los diferentes descriptores son vectores independientes:

$$\chi_{12}{}^2 = -2 \ln \prod_{12} \qquad \textbf{(3.36)}$$

la cuál tiene 2p grados de libertad (p es el número de descriptores).

El índice de semejanza es el complemento de la probabilidad asociada con este χ^2; es decir el complemento de que un valor de χ^2 tomado al azar exceda al valor observado χ^2:

$$S_{23}(x_1, x_2) = 1 - prob\left(\chi^2_{12}\right) \quad \textbf{(3.37)}$$

Debe quedar claro que el valor del índice de Goodall para un par dado de sitios puede variar dependiendo de los sitios incluidos en el cálculo, puesto que está basado en el rango de las semejanzas parciales para ese par de sitios dentro de todos los pares de sitios. Esto hace diferente a la medida de Goodall de todos los otros coeficientes discutidos anteriormente.

3.2. Modo Q: coeficientes de distancia

Los coeficientes de distancia son las funciones que toman sus valores máximos (a menudo 1) para dos objetos que son completamente diferentes, y 0 para dos objetos que idénticos para todos los descriptores excesivos. Las distancias, como las semejanzas, son utilizadas para medir la asociación entre objetos. Los coeficientes de distancia se pueden subdividir en tres grupos. El primer grupo consiste en las métricas que comparten las siguientes cuatro propiedades:

1. Mínimo 0: si a = b, entonces D (a,b) = 0
2. Positividad: si a ≠ b, entonces D (a,b) > 0
3. Simetría: D (a,b) = D (b,a)
4. Desigualdad triangular: D(a, b) +D (b, c) ≥ D (a, c). En este caso la suma de ambos lados de un triángulo en el espacio euclidiano es necesariamente igual o mayor que el tercer lado.

Tabla 3.6 Propiedades de los coeficientes de distancia calculados para los coeficientes de semejanza presentados anteriormente.

Similaridad	$D = 1 - S$ *Métrica , etc .*	$D = 1 - S$ *Euclideana*	$D = \sqrt{1 - S}$ *Métrica*	$D = \sqrt{1 - S}$ *Euclideana*
$S_1 = \frac{a + d}{a + b + c + d}$ Emparejamiento Simple	Métrica	No	Si	Si
$S_2 = \frac{a + d}{a + 2b + 2c + d}$ Roger & Tanimoto	Métrica	No	Si	Si
$S_3 = \frac{2a + 2d}{2a + b + c + 2d}$	Semimétrica	No	Si	No
$S_4 = \frac{a + d}{b + c}$	No métrica	No	No	No
$S_5 = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$	Semimétrica	No	No	No
$S_6 = \frac{a}{\sqrt{(a+b)(a+c)}} \frac{d}{\sqrt{(b+d)(c+d)}}$	Semimétrica	No	Si	Si
$S_7 = \frac{a}{a + b + c}$ Jaccard	Métrica	No	Si	Si
$S_8 = \frac{2a}{2a + b + c}$	Semimétrica	No	Si	Si
$S_9 = \frac{3a}{3a + b + c}$	Semimétrica	No	No	No
$S_{10} = \frac{a}{a + 2b + 2c}$	Métrica	No	Si	Si
$S_{11} = \frac{a}{a + b + c + d}$ Russell & Rao	Métrica	No	Si	Si
$S_{12} = \frac{a}{b + c}$	No Métrica	No	No	No

Fuente:(Gower & Legendre, 1986)

Todos los coeficientes de semejanza expuestos anteriormente, pueden ser transformados en distancias, utilizando las transformaciones de la Tabla anterior, asumiendo que una distancia es no métrica o Euclidiana cuando realmente se puede encontrar un ejemplo; Esto no significa que el coeficiente nunca sería métrico o Euclidiano. Un coeficiente es probable que sea métrico o euclidiano cuando la forma binaria del coeficiente se conoce que es métrico o euclidiano, y las pruebas nunca han demostrado lo contrario.

Se dice que un coeficiente es Euclidiano si las distancias están completamente definidas en un espacio Euclidiano. El análisis de coordenadas principales de tal matriz de distancias no produce valores propios negativos.

Tabla 3.7 Propiedades de los coeficientes de distancia calculados cuando no hay datos faltantes

Similaridad	$D = 1 - S$ *Métrica, etc.*	$D = 1 - S$ *Euclidiana*	$D = \sqrt{1 - S}$ *Métrica*	$D = \sqrt{1 - S}$ *Euclidiana*
$S_{13} = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right]$	Semimétrica	No	No	No
$S_{14} = \frac{a}{\sqrt{(a+b)(a+c)}}$ Ochiai	Semimétrica	No	Si	Si
$S_{15} = \sum w_j s_j / \sum w_j$ Gower	Métrica	No	Si	Como*(S_1)
$S_{16} = \sum w_j s_j / \sum w_j$ Estabrook & Rogers	Métrica	No	Si	Como*(S_1)

$S_{17} = \frac{2W}{A + B}$ Steinhaus	Semimetrica	No	Como*(S_8)	Como*(S_1)
$S_{18} = \frac{1}{2}\left[\frac{W}{A} + \frac{W}{B}\right]$ Kulczynski	Semimétrica	No	No*(S_{13})	No*(S_{13})
$S_{19} = \sum w_j s_s / \sum w_j$ Gower	Métrica	No	Si	Como
$S_{20} = \sum w_j s_j / \sum w_j$ Legendre & Chodorowski	Métrica	No	Si	Como*(S_7)
$S_{21} = 1 - \chi^2$ Métrica	Métrica	Si	Si	Si
$S_{22} = 1 - p(\chi^2)$	Semimétrica	No	-	-
$S_{23} = 2\ (\sum d) / n(n-1)$ Goodall	Semimétrica	No	-	-
$S_{23} = 1 - p(\chi^2)$ Goodall	Semimétrica	No	-	-
$S_{26} = (a + d/2)/p$ Faith	Métrica	-	Si	-

Fuente: (Gower & Legendre, 1986)

3.2.1 Distancias métricas

Las distancias métricas se han desarrollado para descriptores cuantitativos, pero en ocasiones se han utilizado con descriptores semi cuantitativos. Algunas de estas medidas (D_1, D_2, D_5 hasta D_8 y D_{12}) procesan los doble-cero de la misma forma que cualquier otro valor de los descriptores. Estos coeficientes no se deben utilizar, en general, con variables categóricas, como será visto en la paradoja descrita más adelante. Los coeficientes D_3, D_4, D_9, hasta D_{11}, y D_{15} hasta D_{17}, por el contrario, se adaptan bien a datos categorizados. La métrica más común es la distancia euclidiana, que se calcula utilizando la fórmula de Pitágoras para puntos situados en un espacio p-dimensional llamado un espacio métrico o euclidiano:

$$D_1 = (x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad \textbf{(3.38)}$$

Cuando hay solamente dos descriptores, esta expresión se convierte en la medida de la hipotenusa de un triángulo rectángulo.

$$D_1(x_1, x_2) = \sqrt{(y_{11} - y_{21})^2 + (y_{12} - y_{22})^2} \qquad \textbf{(3.39)}$$

El cuadrado de D_1 se puede también utilizar para los análisis de conglomerados. Debemos destacar, sin embargo, que D_1 es un semimétrica y que es menos apropiada que D_1 para los métodos de clasificación:

$$D_1 (x_1, x_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2 \qquad \textbf{(3.40)}$$

La distancia euclidiana no tiene un límite superior y su valor aumenta indefinidamente con el número de descriptores. El valor también depende de la escala de cada descriptor, hasta tal punto que cambiar la escala de algunos descriptores puede dar lugar a las medidas que no son monotónicas con respecto a cada una de ellas.

El problema puede ser evitado utilizando variables estandardizadas en vez de los datos originales, o restringiendo el uso de D_1 y otras distancias del mismo tipo (D_2, D_6 D_7 y D_8) a matrices de datos dimensionalmente homogéneas.

La distancia euclidiana, utilizada como medida de semejanza entre sitios con base conteos de variables categorizadas, puede conducir a la paradoja siguiente: dos sitios sin una especie común pueden estar en una distancia más pequeña que otro par de sitios que comparten especies. Esta paradoja es ilustrada por un ejemplo numérico de Pierre & Louis Legendre (1986:278):

Tabla 3.8 Ejemplo numérico de dos sitios sin una especie

Sitios	**Especies**		
	Y_1	**Y_2**	**Y_3**
X_1	0	1	1
X_2	1	0	0
X_3	0	4	4

Fuente: (Legendre, 1986:278)

De estos datos, las distancias siguientes se calculan entre los sitos:

Tabla 3.9 Ejemplo numérico de calculo de distancias D_1

Sitios	**Sitios**		
	X_1	**X_2**	**X_3**
X_1	0	1. 732	4.243
X_2	1.732	0	5.745
X_3	4.243	5.745	0

Fuente: (Legendre, 1986:278)

Así, la distancia euclidiana entre los sitios X_1 y X_2, que no tienen ninguna especie en común, es más pequeña que la distancia entre X_1 y X_3 que compartan las especies Y_2 y Y_3. En general, los doble-cero conducen a una reducción de las distancias. Esto se debe evitar con datos de conteos categorizados.

Para los descriptores ambientales, por el contrario, los doble-cero pueden considerarse como válidos para comparar sitios. Por tanto, la distancia euclidiana no se debe utilizar para comparar sitios donde los datos sean conteos categorizados.

Se han propuesto varias modificaciones para tratar las desventajas de la distancia euclidiana aplicada a conteos categorizados. Primero, el efecto del número de descriptores puede ser tratado calculando una distancia promedio:

$$D_2^2(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p}(y_{1j} - y_{2j})^2 \quad o \quad D_2(x_1, x_2) = \sqrt{D_2^2} \qquad \textbf{(3.41)}$$

Orlóci (1967b) propuso utilizar la distancia de la cuerda, que tiene un valor máximo igual a $\sqrt{2}$ para sitios sin especies en común y un mínimo de 0 cuando dos sitios comparten la misma especie en las mismas proporciones, sin que sea necesario, para estas especies, estar representadas por las mismas cantidades de individuos en los dos sitios.

Esta medida es la distancia Euclidiana calculada después de llevar los vectores de cada sitio a vectores de longitud 1. (normalizarlos). Después de esta normalización, la distancia euclidiana calculada entre dos objetos (sitios) es equivalente a la longitud de una cuerda que une dos puntos mediante un segmento de esfera o de una hiperesfera de radio 1.

Si hay solamente dos especies involucradas, la normalización coloca los sitios en una circunferencia de un sector de 90° de un círculo de radio 1. La distancia de la cuerda se puede también calcular directamente de datos no-normalizados con la siguiente fórmula:

$$D_3(x_1, x_2) = \sqrt{2\left[1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right]} \qquad \textbf{(3.42)}$$

La parte interna de esta ecuación es realmente el coseno del ángulo (***θ***) entre los dos vectores del sitio, normalizado o no. La fórmula de la distancia de la cuerda puede escribirse como:

$$D_3 = \sqrt{2(1 - \cos\theta)} \qquad \textbf{(3.43)}$$

La distancia de la cuerda es máxima cuando las especies en dos sitios son totalmente diferentes. En tal caso, los vectores normalizados de los sitios están a 90° uno del otro en la circunferencia de un sector de 90° de un circulo (cuando hay solamente dos especies), o en la superficie de un segmento de una hiperesfera (para p especies), y la distancia entre los dos sitios es $\sqrt{2}$. Esta medida soluciona el problema causado por sitios que tienen conteos totales de categorías diferentes, así como la paradoja explicada anteriormente para D_1. En realidad, con D_3 las distancias entre pares de sitios para el ejemplo numérico son:

Tabla 3.10 Calculo de distancias con D_3

Sitios	**Sitios**		
	X_1	**X_2**	**X_3**
X_1	0	1.414	0
X_2	1.414	0	1.414
X3	0	1.414	0

Fuente:(Legendre, 1986:280)

La distancia de la cuerda es una métrica. Puesto que los doble-cero no influyen en la distancia de la cuerda, puede ser utilizada para comparar los sitios descritos por conteos categorizados.

Una transformación de la medida anterior, conocida como la ***métrica geodésica***, mide la longitud del arco en la superficie de la hiperesfera de radio unidad:

$$D_4(x_1, x_2) = arc\ \cos\left[1 \frac{D_3^2(x_1, x_2)}{2}\right] \qquad \textbf{(3.44)}$$

En el ejemplo numérico, los pares de sitios (X_1, X_2) y (X_2, X_3) sin especie en común, están en un ángulo de 90°, mientras que el par (X_1,X_2), que comparte dos de las tres especies, está en un ángulo más pequeño (88°).

Mahalanobis (1936) desarrolló una distancia generalizada que considera las correlaciones entre descriptores y es independiente de las escalas de los diferentes descriptores. Esta medida calcula la distancia entre dos puntos en un espacio cuyos ejes no sean necesariamente ortogonales. Para considerar las correlaciones entre descriptores. Orlóci (1978: 48) ofrece una fórmula para calcular las distancias entre sitios individuales, pero, en la práctica, la ***distancia generalizada de Mahalanobis*** se utiliza solamente para comparar grupos de sitios. Para dos grupos de sitios, W_1 y W_2 que contienen n_1 y n_2 sitios respectivamente, y descrito por las mismas p variables, el cuadrado de la distancia generalizada está dado por la siguiente fórmula matricial:

$$D_5^2(w_1, w_2) = \overline{d_{12}} V^{-1} \overline{d_{12}'} \qquad \textbf{(3.45)}$$

En esta ecuación, $\overline{d_{12}}$ es el vector (longitud = p) de las diferencias entre las medias de las **p** variables en los dos grupos de sitios. **V** es la matriz de dispersión dentro de grupos mancomunada para los dos grupos de sitios, estimados de las matrices de las sumas de cuadrados y de los productos cruzados entre los descriptores de grupos centrados para cada uno de los dos grupos, sumando término a término y dividiendo por (n_1 + n_2 - 2), como en el análisis discriminante y en el análisis de varianza multivariante:

$$V = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1 + (n_2 - 1)S_2\right] \qquad \textbf{(3.46)}$$

donde S_1 y S_2 son las matrices de dispersión para cada uno de los dos grupos. Mientras que el vector $\overline{d_{12}}$ mide la diferencia entre las medias p-dimensionales de dos grupos (descriptores de p), **V** considera las covariaciones entre descriptores. Esta fórmula se puede adaptar para calcular la distancia entre un solo objeto y un grupo.

Si desea probar la significación para D_5, las matrices de dispersión dentro del grupo deben ser homogéneas. La homocedasticidad de las matrices S_1 y S_2 se puede probar utilizando la prueba de Kullback. La prueba de significación también asume la multinormalidad de las distribuciones dentro de grupo, aunque la distancia generalizada tolera un cierto grado de desviación de esta condición.

La distancia generalizada de Mahalanobis conserva las medias entre grupos en un espacio canónico de funciones discriminantes. Para realizar la prueba de significación, la distancia generalizada es transformada en el estadístico T^2 de Hotelling (1931) utilizando la ecuación siguiente:

$$T^2 = \frac{n_1 + n_2}{(n_1 + n_2)} D_5^2 \qquad \textbf{(3.47)}$$

y el estadístico F se calcula de la siguiente forma:

$$F = \frac{n_1 + n_2 - (p + 1)}{(n_1 + n_2 - 2)\, p} T^2 \qquad \textbf{(3.48)}$$

con p y $[n_1 + n_2 - (p + 1)]$ grados de libertad.

El estadístico T^2 es una generalización para el caso multidimensional del estadístico t -Student, que permite probar la hipótesis entre dos grupos originarios de poblaciones con centroides similares. La generalización final para varios grupos, es llamada la Δ (lambda) de Wilks.

Los coeficientes D_2 y D_5 están relacionados con la distancia euclidiana D_1, que es el segundo grado (r = 2) de la métrica de Minkowski:

$$D_6(x_1, x_2) = \left[\sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|^r \right]^{1/r} \quad \textbf{(3.49)}$$

Las formas de esta métrica con r > 2 se utilizan raramente en ecología porque las potencias mayores que 2 dan demasiada importancia a las diferencias más grandes $\left| y_{1j} - y_{2j} \right|$.

Por la razón contraria, el exponente r = 1 se utiliza en muchos casos. La forma básica:

$$D_7(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right| \quad \textbf{(3.50)}$$

se conoce como la métrica de Manhattan, métrica de taxis, o métrica de ciudad-cuadras (city-block). Esto se refiere al hecho de que, para dos descriptores, la distancia entre dos sitios es la abscisa de la distancia (descriptor Y_1) más la distancia en la ordenada (el descriptor Y_2), similar a la distancia recorrida por un taxi alrededor de las cuadras en una ciudad con un plano ortogonal como Manhattan. Esta métrica presenta el mismo problema para los doble-cero que la distancia euclidiana y nos conduce a la misma paradoja.

La ***diferencia de carácter promedio*** propuesta por el antropólogo Czekanowski (1909),

$$D_8(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p}\left|y_{1j} - y_{2j}\right| \qquad \textbf{(3.51)}$$

y tiene como ventaja sobre D_7 que no aumenta con el número p de descriptores. Puede utilizarse con variables de conteo si se modifica la ecuación anterior eliminando los doble-cero del cálculo de $\left|y_{1j} - y_{2j}\right|$ y p se sustituye por: p – la cantidad de doble-ceros.

El ***índice de asociación de Whittaker*** (1952) se adapta bien a conteos de categorías porque cada descriptor primero se transforma en una fracción de la cantidad total de individuos en el sitio, antes de la substracción. El complemento de este índice es la distancia siguiente:

$$D_9(x_1, x_2) = \frac{1}{2}\sum_{j=1}^{p}\left[\frac{y_{1j}}{\sum_{j=1}^{p} y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{21j}}\right] \qquad \textbf{(3.52)}$$

La diferencia es cero para aquellos objetos cuyas proporciones son idénticas en ambos sitios. Un resultado idéntico es obtenido calculando, para todos los objetos, la suma de las fracciones más pequeñas calculadas para los dos sitios:

$$D_9(x_1, x_2) = \left[1 - \sum_{i=1}^{p} \min\left(\frac{y_1}{\sum y}\right)\right] \qquad \textbf{(3.53)}$$

Los australianos Lance y Williams (1967ª) desarrollaron variantes de la métrica de Manhanttan, incluyendo su ***métrica de Canberra*** (Lance y Williams, 1966c) es la siguiente formula:

$$D_{10}(x_1, x_2) = \sum_{j=1}^{p} \left[\frac{|y_{1j} - y_{2j}|}{y_{1j} - y_{2j}} \right] \qquad \textbf{(3.54)}$$

que debe excluir los doble-ceros para evitar indeterminaciones. Esta medida no tiene límite superior. Puede ser demostrado que, en D_{1O}, una diferencia entre objetos abundantes contribuye menos a la distancia que la misma diferencia encontrada entre especies más raras.

Como medida de semejanza ecológica, Stephenson et al. (1972) y Moreau y Legendre (1979) utilizaron el complemento de la métrica de Canberra, después de recodificarla a valores entre 0 y 1:

$$S(x_1, x_2) = 1 - \frac{1}{p} D_{10} \qquad \textbf{(3.55)}$$

Otra versión del coeficiente D_{10} codificado y basado en la distancia euclidiana, ha sido utilizada para propósitos taxonómicos por Clark (1952) bajo el nombre de ***coeficiente de divergencia***:

$$D_{11}(x_1, x_2) = \sqrt{\frac{1}{p} \sum_{j=1}^{p} \left(\frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}} \right)^2} \qquad \textbf{(3.56)}$$

La distancia D_{11} es a D_{10} como D_2 es a D_7 porque, en D_{11} la diferencia para cada descriptor primero se expresa como una fracción, antes de elevar al cuadrado y sumarlos. Este coeficiente se puede utilizar con datos de conteos categorizados. Como en el coeficiente D_8, sin embargo, los doble-cero deben ser excluidos del cálculo y su cantidad restada de p. A menos que uno se preponga utilizar este coeficiente como base para la clasificación, es mejor, para datos de conteos, utilizar la semimétrica D_{14} descrita más adelante.

Otro coeficiente, que se relaciona con el D_{11}, fue desarrollado por Pearson (1926) para estudios antropológicos bajo el nombre de ***coeficiente de la semejanza racial***.

Usando este coeficiente, es posible medir una distancia entre grupos de objetos, como con la distancia generalizada Mahalanobis D_5, pero sin eliminar el efecto de las correlaciones entre descriptores:

$$D_{12}(w_1, w_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left[\frac{(\bar{y}_{1j} - \bar{y}_{2j})^2}{(s_{1j}^2 / n_1) + (s_{2j}^2 / n_2)}\right]} - \frac{2}{p} \qquad \textbf{(3.57)}$$

para dos grupos de los sitios W_1 y W_2 que contienen n1 y n2 sitios respectivamente; $\bar{y}_{1j}$ es la media del descriptor j en el grupo i y el s_{ij}^2 es su varianza correspondiente. Otras medidas, relacionadas con χ^2, están disponibles para calcular la distancia entre objetos usando conteos u otros datos de frecuencias; no se permite ningún valor negativo en los datos. El primero de estos coeficientes se llama la ***métrica*** χ^2.

La suma de los cuadrados de las diferencias es calculada entre perfiles de probabilidades condicionales en dos filas (o columnas) de una tabla de frecuencias, ponderando cada término de la suma de cuadrados por el inverso de la frecuencia de la columna (o de la fila) en la tabla general. Esta medida ha sido utilizada por Roux y Reyssac (1975) para calcular las distancias entre objetos descritos por conteos de especies.

Para calcular la métrica χ^2, la matriz de los datos se debe transformar en una matriz de probabilidades condicionales. Si la métrica se calcula entre las filas de la matriz, las probabilidades condicionales son calculadas por filas. Los elementos de la matriz se convierten en nuevos términos Y_{ij} / Y_{i+} donde Y_{i+} es la suma de las frecuencias en la fila i.

En el ejemplo numérico que se vera a continuación, las filas de la matriz de datos, en la parte izquierda, son los objetos y las columnas son los descriptores. La matriz de probabilidades condicionales (por filas) en la derecha, es utilizada posteriormente para calcular la asociación entre filas (objetos):

$$y = \begin{bmatrix} 45 & 10 & 15 & 0 & 10 \\ 25 & 8 & 10 & 0 & 3 \\ 7 & 15 & 20 & 14 & 12 \end{bmatrix} \begin{matrix} [y_{i+}] \\ \begin{bmatrix} 80 \\ 46 \\ 68 \end{bmatrix} \end{matrix} \rightarrow \left[\frac{y_{ij}}{y_{i+}} \right] = \begin{bmatrix} 0.563 & 0.125 & 0.188 & 0.000 & 0.125 \\ .543 & 0.174 & 0.217 & 0.000 & 0.065 \\ 0.103 & 0.221 & 0.294 & 0.206 & 0.176 \end{bmatrix}$$

$$[y_{+j}] = [77 \quad 33 \quad 45 \quad 14 \quad 25 \;] \quad 194$$

La distancia entre las primeras dos filas de la parte de la derecha de la matriz se podría calcular utilizando la fórmula de la distancia euclidiana D_1.

A partir de la ecuación de la distancia entre los dos objetos entonces sería:

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{y_1 +} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad \textbf{(3.58)}$$

Con esta ecuación, sin embargo, los objetos más abundantes contribuirían de una manera predominante en la suma de cuadrados. En su lugar, la métrica χ^2 se calcula utilizando la expresión ponderada:

$$D_{15}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2j}} \right)^2} \qquad \textbf{(3.59)}$$

donde Y_{+j} es la suma de frecuencias en la columna j. Aunque esta medida no tiene límite superior, proporciona distancias más pequeñas que 1 en la mayoría de los casos. Para el ejemplo numérico, el cálculo de D_{15} entre los primeros dos sitios (filas) da como resultado:

$$D_{15}(x_1, x_2) =$$

$$\left[\frac{(0.536 - 0.543)^2}{77} + \frac{(0.125 - 0.174)^2}{33} + \frac{(0.188 - 0.217)^2}{45} + \frac{(0 - 0)^2}{14} + \frac{(0.125 - 0.065)^2}{25} \right]^{1/2}$$

La cuarta especie, que está ausente en los primeros dos sitios, queda eliminada. Así es como la métrica χ^2 excluye los doble-cero del cálculo.

Los dos sitios fueron comparados utilizando los perfiles de las probabilidades condicionales de los objetos ponderados. D_{15} puede también utilizarse para medir la distancia entre objetos a partir de los perfiles de la distribución ponderada entre varios sitios. La matriz de probabilidad condicional entonces se puede calcular por columnas $\left|y_{ij}/y_{.j}\right|$ antes de aplicar la fórmula anterior, intercambiando las columnas por las filas.

Una medida relacionada con la métrica se llama la ***distancia*** χ^2 (Lebart y Fénelon, 1971), que difiere de la métrica χ^2 en que los términos de la suma de cuadrados son divididos por la probabilidad (frecuencia relativa) de cada fila en la tabla general, en vez de utilizar su frecuencia absoluta. Es decir, es idéntica a la métrica χ^2 multiplicada por $\sqrt{y_{++}}$ donde y_{++} es la suma de todas las frecuencias en la tabla de datos:

$$D_{16}(x_1,x_2)=\sqrt{\sum_{j=1}^{p}\frac{1}{y_{+j}/y_{++}}\left(\frac{y_{1j}}{y_{1+}}-\frac{y_{1j}}{y_{2+}}\right)^2}=\sqrt{y_{++}}\sqrt{\sum_{j=1}^{p}\frac{1}{y_{+j}}\left(\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{21+}}\right)^2} \quad \textbf{(3.60)}$$

La distancia χ^2 es la distancia utilizada en el análisis de correspondencias, al calcular la asociación entre objetos o entre descriptores. Más generalmente, se utiliza para calcular la asociación entre filas o columnas de una tabla de contingencia. Esta medida no tiene límite superior.

El pequeño ejemplo numérico utilizado para ilustrar la paradoja asociada con la distancia euclidiana (D_1) calculada para los descriptores categorizados, se utiliza de nuevo aquí para contrastar D_{16} con D_1.

$$\begin{bmatrix} y_{i+} \end{bmatrix}$$

$$y = \begin{bmatrix} 0\,1\,1 \\ 1\,0\,0 \\ 0\,4\,4 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 8 \end{bmatrix} \rightarrow \begin{bmatrix} \frac{y_{ij}}{y_{i+}} \end{bmatrix} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \end{bmatrix}$$

$$\begin{bmatrix} y_{+j} \end{bmatrix} = \begin{bmatrix} 1 & 5 & 5 \end{bmatrix} \quad 11$$

calculando D_{16} entre las primeras dos filas (objetos) resulta:

$$D_{16}(x_1, x_2) = \left[\frac{(0-1)^2}{1/11} + \frac{(0.5-0)^2}{5/11} + \frac{(0.5-0)^2}{5/11} \right]^{1/2} = 3.477$$

Las distancias entre todos los pares de sitios son:

Tabla 3.11 Ejemplo numérico de la distancia calculada con D_{16}

Sitios	**Sitios**		
	X_1	**X_2**	**X_3**
X_1	0	3.477	0
X_2	3.477	0	3.477
X_3	0	3.477	0

Fuente: (Legendre, 1986:286)

La comparación con los resultados obtenidos por D_1, anteriormente realizado, demuestra que no existe el problema causado por la presencia de doble-ceros. La distancia D_{16} puede, por lo tanto, ser utilizada directamente con los sitios (objetos) descritos por descriptores categorizados, contrariamente a los D_1.

Un coeficiente relacionado con D_{15} y D_{16} es la distancia de Hellinger, descrita por Rao (1995). Cuya fórmula para esta distancia:

$$D_{17}=(x_1,x_2)=\sqrt{\sum_{j=1}^{p}\left[\sqrt{\frac{y_{1j}}{y_{1+}}}-\sqrt{\frac{y_{2j}}{y_{2+}}}\right]^2} \qquad \textbf{(3.61)}$$

Igual que D_{16}, esta distancia no tiene límite superior. Rao (1995) recomienda esta medida como base para utilizarla en un nuevo método de clasificación. Se puede obtener una clasificación similar calculando la función de distancia entre los objetos y realizando un análisis de coordenadas principales.

3.2.2 Distancias semimétricas

Algunas medidas de distancia no siguen la cuarta propiedad de las métricas, es decir el axioma de la desigualdad triangular. Por consiguiente, no permiten una clasificación apropiada de los objetos en un espacio Euclidiano completo. Pueden, sin embargo, ser utilizadas para la clasificación mediante un análisis de coordenadas principales después de realizada la corrección para los valores propios negativos o mediante un escalamiento multidimensional no métrico. Estas medidas se llaman ***semimétricas o pseudométricas***.

La distancia correspondiente al coeficiente de Sorensen S_8 fue descrito por Watson bajo el nombre de ***coeficiente no métrico***.

$$D_{13}(x_1,x_2)=1-\frac{2a}{2a+b+c}=\frac{b+c}{2a+b+c} \qquad \textbf{(3.62)}$$

El ejemplo numérico siguiente demuestra que D_{13} no obedece el axioma de la desigualdad triangular:

Tabla 3.12 Ejemplo numérico donde D_{13} no respeta axioma de desigualdad triangular

Sitios	**Especies**				
	Y_1	**Y_2**	**Y_3**	**Y_4**	**Y_5**
X_1	1	1	1	0	0
X_2	0	0	0	1	I
X_3	1	1	1	1	1

Fuente:(Legendre, 1986:287)

Las distancias entre los tres pares de sitios son:

$$D_{13}(x_1, x_2) = \frac{3 + 2}{0 + 3 + 2} = 1.00$$

$$D_{13}(x_1, x_3) = \frac{0 + 2}{(2 \times 3) + 0 + 2} = 0.25$$

$$D_{13}(x_2, x_3) = \frac{0 + 3}{(2 \times 2) + 0 + 3} = 0.43$$

Por lo tanto, 1.00 > 0.25 + 0.43, contrariamente al axioma de la desigualdad triangular.

Entre las medidas para datos de conteos entre objetos, los coeficientes S_{17} de Steinhaus y S_{18} de Kulczynski son semimétricas cuando están transformados en distancias. En particular, $D_{14} = 1 - S_{17}$ fue descrita previamente por Odum (1950), que la llamó ***la diferencia de porcentaje***, y luego por Bray y Curtis (1957):

$$D_{14} = (x_1, x_2) = \frac{\sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} = 1 - \frac{2W}{A + B} \qquad \textbf{(3.63)}$$

Contrario a la métrica D_{10} de Canberra, diferencias entre objetos abundantes contribuyen igual en D_{14} que las diferencias entre objetos raros. Esto se puede ver como una propiedad deseable, por ejemplo al usar datos de frecuencia de objetos normalizados. Bloom (1981) comparó la métrica de Canberra, la diferencia de porcentaje y otros índices con un estándar teórico.

Él demostró que solamente D_{I4} o S_{17} reflejaban con precisión la verdadera semejanza a lo largo de su escala completa entre 0 a 1, mientras que D_{10} por ejemplo, subestimaba demasiado la semejanza en el rango de 0 a 1.

El ejemplo numérico siguiente, de Orlóci (1978: 59) demuestra que D_{14} no obedece el axioma de la desigualdad triangular:

Tabla 3.13 Ejemplo numérico donde D_{14} no obedece desigualdad triangular

Cuadrantes	**Especies**				
	Y_1	**Y_2**	**Y_3**	**Y_4**	**Y_5**
X_1	2	5	2	5	3
X_2	3	5	2	4	3
X_3	9	1	1	1	1

Fuente: (Orlóci (1978:59)

Las distancias entre los tres pares de sitios son:

$$D_{14}(x_1, x_2) = \frac{1+0+0+1+0}{17+17} = 0.059$$

$$D_{14}(x_1, x_3) = \frac{7+4+1+4+2}{17+13} = 0.600$$

$$D_{14}(x_2, x_3) = \frac{6+4+1+3+2}{17+13} = .0533$$

Por lo tanto, 0.600 > 0.059 + 0.533, contrariamente al axioma de la desigualdad triangular. Si las cantidades de individuos son las mismas en todos los sitios (sumas de filas), entonces D_{14} es una métrica: por lo que, cuando las cantidades de individuos son bastante diferentes de sitio a sitio, la clasificación mediante un análisis de coordenadas principales basado en matrices de D_{14} (o S_{17} y S_{18}), es probable que se obtengan valores propios negativos.

3.3 Modo R: Coeficientes de Dependencia

El propósito principal del análisis del modo- R es investigar las relaciones entre descriptores; las matrices R se pueden también utilizar, en algunos casos, para la clasificación de objetos (en el análisis de componentes principales o en el análisis discriminante). Después de la clasificación de los descriptores en la matriz, los coeficientes de dependencia se seleccionarán para descriptores cuantitativos, semi-cuantitativos, y cualitativos.

La mayoría de los coeficientes de dependencia permiten la realización de una prueba estadística. Para tales coeficientes, es posible asociar una matriz de probabilidades con la matriz **R**. A pesar de que no siempre es posible aplicar pruebas estadísticas de significación, no es incorrecto calcular un coeficiente de dependencia. Por ejemplo, no hay ninguna objeción para calcular un coeficiente de correlación de Pearson para cualquier par de variables métricas, pero estas mismas variables deben estar distribuidas normalmente. Además, una prueba de significación permite solamente que se acepte o se rechace una hipótesis específica referente al valor del estadístico (aquí, el coeficiente de semejanza) mientras que el coeficiente en sí mismo, mide la intensidad de la relación entre los descriptores.

3.3.1 Descriptores de conteo

El por qué la semejanza entre los descriptores de conteo (frecuencias) se debe medir usando coeficientes especiales se ha explicado anteriormente. Las medidas de semejanza se utilizan aquí para comparar los descriptores para los cuales los doble-cero proporcionan información inequívoca.

La semejanza entre ***descriptores cuantitativos*** se puede calcular usando las medidas paramétricas de dependencia; es decir, medidas basadas en los parámetros de las distribuciones de frecuencia de los descriptores. Estas medidas son la covarianza y el coeficiente de correlación de Pearson, aunque ellos se adaptan solamente a descriptores cuyas relaciones son lineales.

La ***covarianza*** S_{jk} entre los descriptores j y k se calcula para variables centradas $(y_{ij} - \overline{y}_j)$ y $(y_{ik} - \overline{y}_k)$. El rango de valores de la covarianza no tiene ningún límite superior o inferior a priori. Las varianzas y covarianzas entre un grupo de descriptores forman su matriz de dispersión S.

$$S_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \overline{y_j}) (y_{ik} - \overline{y_k}) \qquad \textbf{(3.64)}$$

El ***coeficiente de correlación de Pearson*** r_{jk} es la covarianza de los descriptores j y k calculados para las variables estandarizadas.

$$r_{jk} = \frac{s_{jk}}{s_j\, s_k} = \frac{\sum_{i=1}^{n} (\, y_{ij} - \overline{y_j}\,)\;(\, y_{ik} - \overline{y_k}\,)}{\sqrt{\sum_{i=1}^{n} (\, y_{ij} - \overline{y_j}\,)^2 \sum^{n} (\, y_{ik} - \overline{y_k}\,)^2}} \qquad \textbf{(3.65)}$$

Los coeficientes de correlaciones entre un grupo de descriptores forman la matriz de correlación **R.**

$$R = \begin{bmatrix} 1 & \rho_{12} & \bullet\bullet\bullet & \rho_{1p} \\ \rho_{21} & 1 & \bullet\bullet\bullet & \rho_{2p} \\ \bullet & \bullet & \bullet\bullet\bullet & \bullet \\ \bullet & \bullet & \bullet\bullet\bullet & \bullet \\ \bullet & \bullet & \bullet\bullet\bullet & \bullet \\ \rho_{p1} & \rho_{p2} & \bullet\bullet\bullet & 1 \end{bmatrix}$$

El rango de los coeficientes de correlación se encuentra entre -1 y + 1. La significación de los coeficientes individuales (la hipótesis nula generalmente es H_0: r = 0) usando la ecuación siguiente ($F = \nu \frac{r_{jk}^2}{1 - r_{jk}^2}$), mientras que esta ecuación ($\chi^2 = -\left[n - (2p+11)/6\right] \ln |R|$) se utiliza para probar la independencia completa entre todos los descriptores.

Algunos autores han utilizado el coeficiente r de Pearson para análisis del **modo Q** después de intercambiar las posiciones de objetos y de descriptores en la matriz de los datos. Lefebvre (1980) llama a esta medida de Q el ***coeficiente de parecido***. Hay por lo menos cinco objeciones a esto:

- En el **modo R**, el coeficiente r de Pearson es un coeficiente sin dimensiones. Cuando los descriptores no son dimensionalmente homogéneos, el coeficiente de correlación del **modo Q**, que combina todos los descriptores, tiene dimensiones complejas que no puedan ser interpretadas.

- En la mayoría de los casos, uno puede arbitrariamente recodificar los descriptores cuantitativos (por ej., multiplicando uno por 100 y dividiendo otro por 10). En el **modo R**, el valor de r no se altera después de la recodificación, mientras que si se hace esto en el **modo Q**, puede cambiar el valor de la semejanza entre objetos de manera imprevisible y no monotónica.

- Para evitar los dos problemas anteriores, se ha sugerido estandardizar los descriptores antes de calcular las correlaciones en el **modo Q**. Consideremos dos objetos x_l y x_2: su semejanza debe ser independiente de los otros objetos del estudio; quitar objetos del estudio no debe cambiarlo. Cualquier cambio en la composición de los objeto del estudio, cambia las variables estandardizadas; y afecta el valor de la correlación calculada entre x_1 y x_2. Por lo tanto, la estandarización no soluciona los problemas.

- Incluso con datos dimensionalmente homogéneos (por ej. Con datos de conteos), el segundo problema todavía se mantiene. Además, en el **modo R**, el teorema del límite central predice que, cuando el número de objetos aumenta, las medias, las varianzas y las covarianzas (o las correlaciones) convergen hacia los valores de sus parámetros poblacionales.

En el **modo Q**, por el contrario, la adición de nuevos descriptores (ya que sus posiciones se han intercambiado con la de los objetos en la matriz de datos) causa variaciones importantes en el coeficiente de parecido si estos descriptores adicionales no se correlacionan perfectamente con los anteriores.

- Si los coeficientes de correlación se podrían utilizar como medida general de parecido en el **modo Q**, deben ser aplicables, en particular, al caso simple de la descripción de las proximidades entre sitios, calculado de sus coordenadas geográficas X y Y en un mapa; las correlaciones obtenidas de este cálculo deben reflejar de una cierta manera las distancias entre los sitios. Éste no es el caso: los coeficientes de correlación calculados entre sitios a partir de sus coordenadas geográficos son todos + 1 ó - 1.

Por lo cual, las medidas diseñadas para el análisis del **modo R** no se deben utilizar para estudios del **modo Q**. La semejanza entre ***descriptores semiquantitativos*** y, más generalmente, entre cualquier par de descriptores ordenados cuya relación es monotónica, puede ser determinada utilizando medidas de dependencia no paramétricas.

Como los descriptores cuantitativos están ordenados, los coeficientes no paramétricos pueden utilizarse para medir su dependencia, ya que están monotónicamente relacionados.

Los dos coeficientes de correlación no parametricos son: el coeficiente **r** de Spearman y el coeficiente τ(tau) de kendall. En el coeficiente **r** de Spearman,

$$r_{jk} = \frac{\frac{1}{2}\left[\frac{n^3 - n}{12} + \frac{n^3 - n}{12} - \sum_{i=1}^{n} d_i^2\right]}{\frac{n^3 - n}{12}} = 1 - \frac{6\sum_{i=1}^{n} d_1^2}{n^3 - n} \qquad \textbf{(3.66)}$$

Los valores cuantitativos se sustituyen por rangos antes de aplicar la formula r de Pearson. El coeficiente τ de Kendall

$$\tau_a = \frac{S}{n\ (n-1)/2} = \frac{2\ S}{n\ (n-1)} \qquad \textbf{(3.67)}$$

$$\tau_c = \frac{S}{\frac{1}{2} n^2 \left(\frac{\min - 1}{\min}\right)} \qquad \textbf{(3.68)}$$

mide la semejanza de una forma bastante diferente del coeficiente r de Pearson. Los valores de r de Spearman y τ de Kendall oscilan entre -1 y +1. Igual que para el coeficiente r de Pearson, el **modo Q** no debe utilizarse para coeficientes de correlación de rango. En realidad, aunque los descriptores cuantitativos estén estandarizados surge e mismo problema que para el coeficiente r de Pearson; es decir, la medida Q para un par de objetos es una función de todos los objetos del conjunto de datos.

además en muchos estudios de muestreo algunos objetos están representados por frecuencias muy bajas y como estas frecuencias están sujetas a variaciones estocásticas grandes, los rangos de los descriptores correspondientes son inciertos en el estudio general. Como consecuencia, las correlaciones de rango entre sitios están sujetas ha variaciones aleatorias importantes porque sus valores están basados en un gran número de rangos inciertos. Esto es equivalente ha darle un peso preponderante a muestras de objetos que están pobremente muestreados.

La medición de la semejanza entre los pares de ***descriptores cualitativos*** se basa en las tablas de contingencia de doble entrada, cuyo análisis se realiza generalmente usando el estadístico χ^2 (chi-cuadrado). El análisis de la tabla de contingencia es también el enfoque principal disponible para medir la dependencia entre los descriptores ordenados cuantitativos o semiquantitativos que no están monotónicamente relacionados. El valor mínimo de χ^2 es cero, pero no tiene ningún límite superior a priori. Cuyas formulas son:

$$\chi_p^2 = \sum_{todas\ las\ celdas} \frac{(0 - E)^2}{E} \qquad \textbf{(3.69)}$$

$$\chi_W^2 = \sum_{todas\ las\ celdas} 0 \ln\left(\frac{0}{E}\right) \qquad \textbf{(3.70)}$$

3.3.2 Coeficientes del tipo 1

Considera dos objetos, cada uno representado por un vector de frecuencias, para ser comparado usando una medida de **modo Q**. Con los coeficientes del tipo 1, si hay una diferencia entre sitios para frecuencias abundantes y la misma diferencia para objetos raros, las dos especies contribuyen igualmente a la semejanza o a la distancia entre los sitios. Un ejemplo numérico pequeño ilustra esta propiedad para la diferencia del porcentaje (D_{14}) que es el complemento de la semejanza de Steinhaus (S_{17}).

Tabla 3.14 Propiedad para la diferencia del porcentaje (D_{14}), complemento de la semejanza de Steinhaus

Especies:	**Y_1**	**Y_2**	**Y_3**
Sitio X_1	100	40	20
Sitio X_2	90	30	10
$\lvert y_{1j-}y_{2j}\rvert$	10	10	10
$(y_{1j}+y_{2j})$	190	70	30

Fuente: (Legendre, 1986:296)

Si se usa D_{16} se puede demostrar que cada uno de las tres especies contribuye 10/290 al total de la distancia entre los dos sitios. Con algunos coeficientes (D_3, D_4, D_9), la estandardización de los sitio-vectores, que se hace automáticamente antes de el cálculo del coeficiente, puede hacer que el resultado sea debido ha la importancia dada a cada especies.Con estos coeficientes, la propiedad de "contribución igual" se encuentra solamente cuando los dos sitio-vectores son igualmente importantes, siendo esta importancia medida de diferentes formas dependiendo del coeficiente.

3.3.3 Coeficientes del Tipo 2a

Con coeficientes de este tipo, Una diferencia entre valores para frecuencias abundantes contribuyen menos a la distancia (y, más a la semejanza) que la misma diferencia para frecuencias escasas. La métrica de Canberra (D_{10}) pertenece a este tipo, Para el ejemplo numérico anterior, el cálculo de D_{10} demuestra que la especie y1, que es la más abundante, contribuye 10/190 a la distancia, y2 contribuye 10/70, mientras que la contribución de y1, que es una especie rara, es la más grande de los tres (10/30). La distancia total es D_{10} = 0.529. El coeficiente de la divergencia (D_{11}) también pertenece a este tipo.

3.3.4 Coeficientes de Tipo 2b

Los coeficientes de este tipo se comportan de semejante a los anteriores excepto que la importancia de cada especie se calcula con respecto al conjunto de datos total, en vez de para las comparaciones dos sitio-vectores. La métrica χ^2 (D_{15}) es representativo de esto.

En la distancia D_{15} y el ejemplo subsiguiente del acompañamiento, el cuadrado de la diferencia entre las probabilidades condicionales, para una especie dada, es dividida por Y_{+j} que es el número total de individuos que pertenecen a esta especie en todos los sitios. Si este número es grande, se reduce la contribución de la especie a la distancia total entre dos filas (sitios) más que lo que ocurriría en el caso de especies escasas. El coeficiente de Gower (S_{19}) tiene el mismo comportamiento (excepto cuando se utilizan los pesos especiales W_{12j} para algunas especies), ya que la importancia de cada especie se determina a partir de su rango de variación en todos los sitios.

El coeficiente de Legendre y Chodorowski (S_{20}) también pertenece a este tipo cuando el parámetro k en la función de semejanza parcial S_{12j} para cada especie se hace proporcional a su rango de variación en todos los sitios.

Legendre (1985) sugirió que es más informativo comparar especies dominantes o bien representadas que especies escasas, ya que estas últimas no están generalmente bien muestreadas. Esto da un enfoque para elegir un coeficiente. En comunidades no establecidas, la mayoría de las especies están representadas por pequeñas cantidades de individuos, de manera que sólo pueden ser muestreadas pocas especies, mientras que, en comunidades establecidas, varias especies presentan frecuencias de ocurrencia intermedias o altas. cuando se calcula semejanzas entre especies de comunidades no establecidas, un enfoque razonable puede ser darle mayor peso a las pocas especies bien muestreadas (coeficientes tipo 2), mientras que, para sitios de comunidades establecidas, los coeficientes del tipo 1 pueden ser más apropiados.

Otra manera de elegir un coeficiente de semejanza es construir datos artificiales que representen situaciones contrastantes de modo que la medida de la semejanza o de la distancia puedan diferenciarlas. Si calculamos varios coeficientes posibles para probarlos con nuestros datos indicarán cuál coeficiente es el más apropiado para datos del mismo tipo. En este sentido, Hajdu (1981) construyó una serie de pruebas para diferentes casos, que llamó comparaciones ordenadas de series casos (OCCAS), correspondiente a cambios lineales en las abundancias de dos especies a lo largo de diferentes tipos de gradientes ambientales simulados. Los resultados son distancias entre sitios, calculados usando diferentes coeficientes, para composiciones de especies que cambian linealmente.

CAPÍTULO 4. FORMACIÓN DE LOS CONGLOMERADOS

Con las variables seleccionadas y la matriz de similitud calculada, comienza el proceso de seleccionar el algoritmo de aglomeración para la formación de conglomerados (clústers).

Esto no es algo sencillo ya que existen muchos algoritmos y siempre se están desarrollando más. No obstante, el criterio esencial de todos los algoritmos es que intentan maximizar las diferencias entre los conglomerados relativa a la variación dentro de los conglomerados, La razón entre la variación dentro de los conglomerados y la media de la variación dentro de los conglomerados es semejante (aunque no igual) a la razón F del análisis de varianza.

Existen dos grandes categorías de algoritmos de obtención de conglomerados: los jerarquizados y los no jerarquizados.

4.1 Métodos de clasificación jerárquicos

Los procedimientos jerárquicos consisten en la construcción de una estructura en forma de árbol. La agrupación se realiza mediante un proceso que conlleva un conjunto de fases de agrupación o desagrupación sucesiva.

El resultado final es una jerarquía de unión completa en la que cada grupo se une o separa en una determinada fase. Existen dos tipos de procedimientos de obtención de conglomerados jerárquicos: **aglomerativos y divisivos.**

- En los **métodos aglomerativos**, cada individuo empieza dentro de su propio conglomerado. En etapas posteriores, los dos conglomerados más cercanos (o individuos), se combinan en un nuevo conglomerado, reduciendo así el número de conglomerados paso a paso. En cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo, creando algo parecido a un árbol. Esta representación se denomina **DENDOGRAMA o gráfico en forma de árbol**.

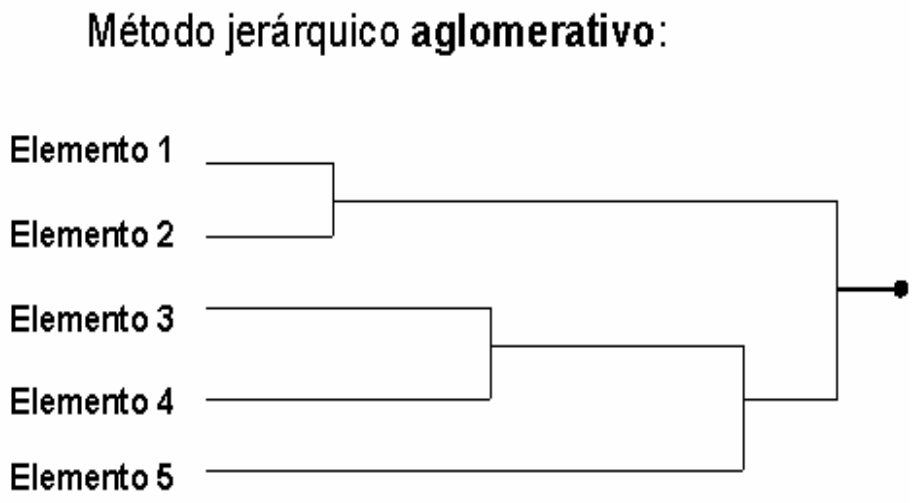


Figura 4.1 Método jerárquico aglomerativo

- En los **métodos divisivos** el proceso de obtención de conglomerados procede en dirección opuesta al anterior; es decir, empezamos con un gran conglomerado que contiene todos los individuos y en los pasos sucesivos, las observaciones que son más diferentes se dividen y se construyen conglomerados más pequeños. Este procedimiento continúa hasta que cada individuo es un conglomerado en sí mismo.

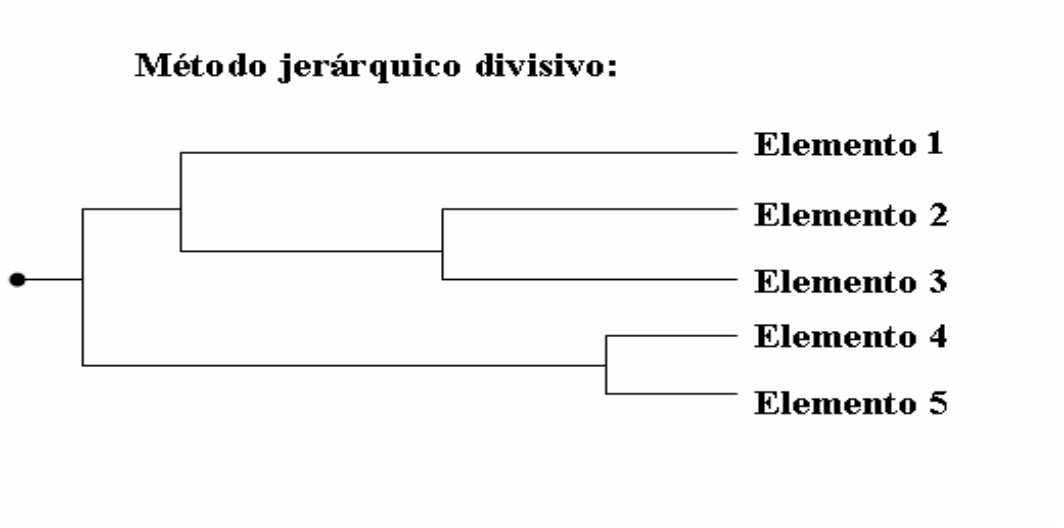


Figura 4.2 Método jerárquico divisivo

Entre los algoritmos más utilizados para desarrollar conglomerados se encuentran:

El método de ligamiento simple, se basa en la distancia mínima entre dos individuos y con ellos forma el primer conglomerado y así sucesivamente. También se conoce como el método del vecino más cercano. La distancia entre dos conglomerados cualesquiera es la distancia más corta desde cualquier punto de un conglomerado hasta cualquier punto en el otro.

Este método tiene el problema de que puede producir largas cadenas llegando a conformar una sola cadena cuando los conglomerados no están bien definidos.

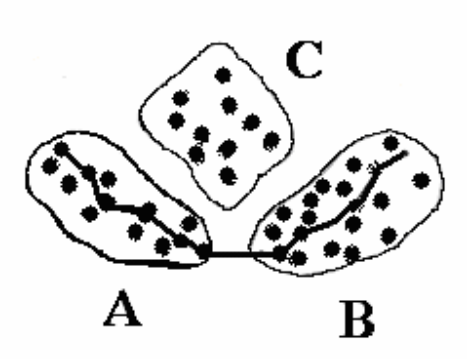


Figura 4.3 Ligamiento simple que une a los conglomerados diferentes A y B

El método de ligamiento completo, se basa en la distancia máxima, ya que todos los individuos de un conglomerado se vinculan con el resto a alguna distancia máxima o por la mínima similitud. Esta técnica elimina el problema de las largas cadenas del método anterior..También es conocido como el método del vecino más lejano.

Figura 4.4 Ligamiento completo

Para aplicar este método se aplican los siguientes pasos:

1. Se inicia con *N* agrupamientos, en donde cada uno de ellos contiene exactamente un punto dato.
2. Enlazar los dos puntos mas cercanos según una de las tres medidas seleccionadas de la distancia
3. Definir la desemejanza entre este nuevo agrupamiento y cualquier otro punto como la distancia mínima entre los dos puntos del agrupamiento y este punto.
4. Combinar los agrupamiento que sean los mas cercanos entre si de modo que en cada etapa, la cantidad de agrupamientos se reduzca en uno y la desemejanza entre cualesquiera dos de éstos siempre se defina como la distancia entre sus miembros más cercanos

Para ilustrar cómo funciona el método del vecino más cercano, se consideró una muestra que contiene seis puntos y supóngase que las distancias entre éstos se expresan por la siguiente matriz de desemejanza:

Tabla 4.1 Matriz de desemejanza

	1	**2**	**3**	**4**	**5**	**6**
1		0.31	0.23	0.32	0.26	0.25
2			0.34	0.21	0.36	0.28
3				0.31	0.04	0.07
4					0.31	0.28
5						0.09
6						

Fuente: (Hair, Anderson, Tatham y Black, 2000)

La agrupación inicial se denota por C_0 y tiene a cada punto en un agrupamiento por sí mismo. Por tanto, la agrupación inicial es:

$$C_0 = \{[1], [2], [3], [4], [5], [6]\}$$

Buscando por toda la matriz de desemejanza, se puede ver que los dos puntos más cercanos entre sí son el 3 y el 5. Por consiguiente, el primer paso del proceso de agrupación sería producir el agrupamiento:

$$C_1 = \{[1], [2], [3,5], [4], [6]\}$$

Enseguida, debe calcularse una nueva matriz de distancias entre los agrupamientos que se encuentran en C_1. El método de vecino más cercano toma la distancia entre [1] y [3, 5] como el mínimo de 0.23 y 0.26, de modo que la distancia entre [1] y [3, 5] es 0.23; de manera semejante, se pueden determinar las distancias entre todos los demás agrupamientos.

Una nueva matriz de distancias, para la agrupación definida por C_1, es:

Tabla 4.2 Matriz de distancias para la agrupación definida por C_1

	1	**2**	**3**	**4**	**6**
1		0.31	0.23	0.32	0.25
2			0.34	0.21	0.28
[3,5]				0.31	0.07
4					0.28
6					

Fuente: (Hair, Anderson, Tatham y Black, 2000)

Aquí, los dos agrupamientos más próximos son [6] y [3, 5] y, de donde, se combinaría estos dos agrupamientos. Los resultados del segundo paso producen esta agrupación:

$$C_2 = \{[1], [2], [3,5,6], [4]\}$$

Entonces, debe calcularse una nueva matriz de distancias. La ventaja de los métodos de un solo enlace es que la nueva matriz de distancias se puede calcular a partir de la del paso anterior. De este modo, al aplicar los métodos de un solo enlace no se necesita regresar a la matriz original de distancias. La matriz de distancias, para la agrupación definida por C_2, es

Tabla 4.3 Matriz de distancias para la agrupación definida por C_2

	1	**2**	**[3,5,6]**	**4**
1		0.31	0.23	0.32
2			0.28	0.21
[3,5,6]				0.28
4				

Fuente: (Hair, Anderson, Tatham y Black, 2000)

Esta matriz de distancias produce la siguiente agrupación:

$$C_3 = \{[1], [2,4], [3,5,6]\}$$

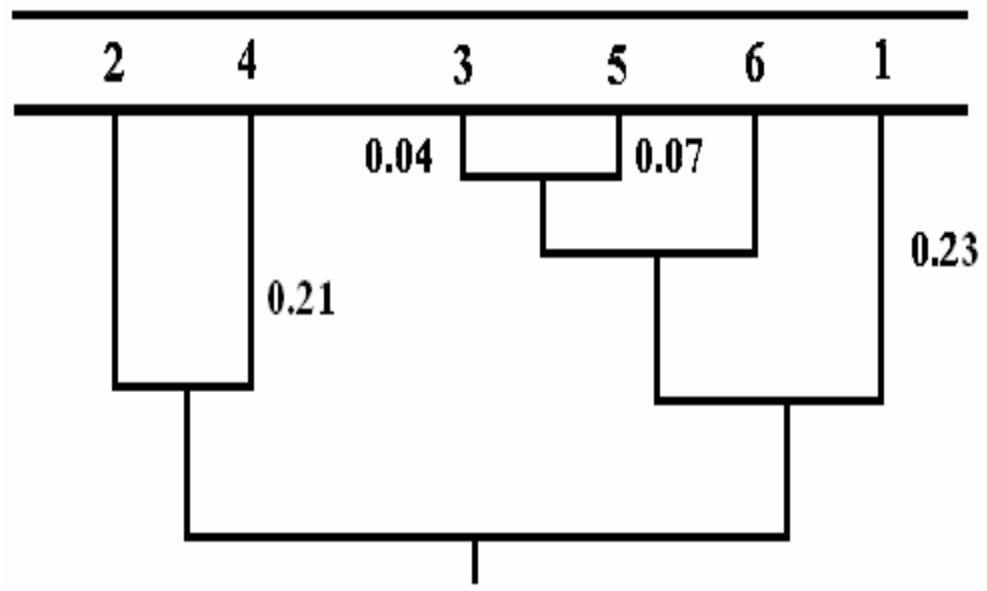
La nueva matriz de distancias, para la agrupación definida por C_3, es

Tabla 4.4 Matriz de distancias para la agrupación definida por C_3

	1	**[2,4]**	**[3,5,6]**
1			
[2,4]			
[3,5,6]			

Fuente: (Hair, Anderson, Tatham y Black, 2000)

la cual produce la agrupación siguiente:



$$C_4 = \{[1,3,5,6],[2,4]\}$$

Figura 4.5 Dendograma resultado de la agrupación C_4

La matriz de distancia para esta agrupación es

Tabla 4.5 Matriz de distancias para la agrupación definida por C_4

	[1,3,5,6]	**[2,4]**
[1,3,5,6]		0.28
[2,4]		

Fuente: (Hair, Anderson, Tatham y Black, 2000)

Y la agrupación final es

$$C_5 = \{[1,2,3,4,5,6]\}$$

Método de ligamiento medio, comienza igual que los métodos anteriores, pero el criterio de aglomeración es la distancia media de todos los individuos de un conglomerado con todos. Tales técnicas no dependen de los valores extremos, como en los métodos anteriores, y la partición se basa en todos los miembros de los conglomerados en lugar de un par único de miembros extremos. Este enfoque tiende a combinar los conglomerados con variaciones reducidas dentro del conglomerado, aunque tiende a estar sesgado hacia la producción de conglomerados con aproximadamente la misma varianza. Mide la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos o la media de las similitudes entre objetos de ambos grupos. Así, por ejemplo, si se utilizan distancias, la distancia entre los grupos r y s vendría dada por:

$$\frac{1}{n_r n_s} \sum_{j \in r} \sum_{k \in s} d(j, k) \qquad \textbf{(4.1)}$$

donde d (j, k) = distancia entre los objetos j y k y n_r, n_s son los tamaños de los grupos r y s, respectivamente.

Enlace medio dentro de los grupos, mide la proximidad entre dos grupos con la distancia media existente entre los miembros del grupo unión de los dos grupos. Así, por ejemplo, si se trata de distancias, la distancia entre los grupos r y s vendría dada por:

$$\frac{1}{C_{n_r + n_s}^{2}} \sum_{(j,k) \in r \cup s} d(j,k) \qquad \textbf{(4.2)}$$

Método de Ward, considera que la distancia entre dos conglomerados es la suma de los cuadrados entre dos conglomerados sumados para todas las variables. En cada paso del procedimiento, se minimiza la suma de los cuadrados dentro del conglomerado para todas las particiones (el conjunto completo de conglomerados disjuntos o separados) obtenida mediante la combinación de dos conglomerados en un paso previo; es decir, se minimizará la variación intra grupal de la estructura formada. Este procedimiento tiende a combinar los conglomerados con un número reducido de observaciones y también está sesgado hacia la producción de conglomerados con aproximadamente el mismo número de observaciones. *(Tiende a generar conglomerados demasiado pequeños y demasiado equilibrados en tamaño).*

El método busca minimizar $\sum_r SSW$ donde SSW_r es, para cada grupo r, las sumas de cuadrados intragrupo que viene dada por:

$$SSW_r = \sum_{m=1}^{nr} \sum_{j=1}^{p} \left(x_{rjm} - \overline{x}_{rj} \right)^2 \qquad \textbf{(4.3)}$$

donde x_{rjm} denota el valor de la variable X_j en el m-ésimo elemento del grupo r.

En cada paso del algoritmo une los grupos r y s que minimizan:

$$SSW_t - SSW_r - SSW_s = \frac{n_r n_s}{n_r + n_s} d_{rs}^2 \qquad \textbf{(4.4)}$$

con t = r ∪s y d_{rs}^2 la distancia entre los centroides de r y s.

Los métodos del centroide y de la mediana, plantean que la distancia entre dos conglomerados es la distancia (normalmente euclidiana simple o cuadrada) entre sus centroides. Los centroides de los grupos son los valores medios de las observaciones de las variables en el valor teórico del conglomerado. Cada vez que se agrupa a los individuos, se calcula un nuevo centroide; es decir, los centroides de un grupo cambian a medida que se fusionan conglomerados. Este método a veces produce resultados desordenados y a menudo confusos; aunque tiene la ventaja de que se ve menos afectado por los individuos atípicos que otros métodos jerárquicos.

Ambos métodos miden la proximidad entre dos grupos calculando la distancia entre sus centroides

$$d_{rs}^2 = \sum_{j=1}^{p} \left(\overline{x}_{rj} - \overline{x}_{sj} \right)^2 \qquad \textbf{(4.5)}$$

donde $\overline{x}_{rj}$ y $\overline{x}_{sj}$ son las medias de la variable X_j en los grupos r y s, respectivamente. Los dos métodos difieren en la forma de calcular los centroides: el método del centroide utiliza las medias de todas las variables de forma que las coordenadas del centroide del grupo r = s ∪ t vendrán dadas por:

$$\overline{x}_{rj} = \frac{1}{n_r} \sum_{m=1}^{n_r} x_{rjm} = \frac{n_s}{n_s + n_t} \overline{x}_{sj} + \frac{n_t}{n_s + n_t} \overline{x}_{tj} \qquad j = 1 \ldots p \qquad \textbf{(4.6)}$$

En el método de la mediana el nuevo centroide es la media de los centroides de los grupos que se unen

$$\bar{x}_{rj} = \frac{1}{2}\bar{x}_{sj} + \frac{1}{2}\bar{x}_{tj} \qquad \textbf{(4.7)}$$

4.1.1 Comparación de los diversos métodos aglomerativos

1) El enlace simple conduce a clusters encadenados
2) El enlace completo conduce a clusters compactos
3) El enlace completo es menos sensible a outliers que el enlace simple
4) El método de Ward y el método del enlace medio son los menos sensibles a outliers
5) El método de Ward tiene tendencia a formar clusters más compactos y de igual tamaño y forma en comparación con el enlace medio

4.2 Métodos de clasificación no jerárquicos o de k medias

Este tipo de método es conveniente utilizarlo cuando los datos a clasificar son muchos y/o para refinar una clasificación obtenida utilizando un método jerárquico. Supone que el número de grupos es conocido a priori.

En los procedimientos no jerárquicos no se construyen árboles. En su lugar, se asignan los objetos a conglomerados una vez que el número de conglomerados a formar esta especificado. Por tanto, la solución de 6 conglomerados no es sólo una combinación de dos conglomerados a partir de una solución de 7 conglomerados, sino que se basa en la búsqueda de la mejor solución de esos 6 conglomerados.

El proceso opera seleccionando una **"semilla de conglomerado"** como centro de conglomerado inicial, y todos los individuos que se encuentran dentro de una "**distancia umbral"** previamente especificada se incluyen dentro del conglomerado resultante. Entonces se selecciona otra "semilla de conglomerado" y el proceso de asignación continúa hasta que todos los individuos están asignados.

4.2.1. Pasos para implementar el Método de K- medias

En general se consideran 4 pasos para implementar un método no jerárquico, que son:

1) Se seleccionan k centroides o semillas donde k es el número de grupos deseado
2) Se asigna cada observación al grupo cuya semilla es la más cercana
3) Se calculan los puntos semillas o centroides de cada grupo
4) Se iteran los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, los puntos semillas apenas cambian o los grupos obtenidos en dos iteraciones consecutivas son los mismos.

Este método suele ser muy sensible a la solución inicial dada, por lo que es conveniente utilizar una que sea buena. Una forma de construirla es mediante una clasificación obtenida por un algoritmo jerárquico. Los procedimientos de aglomeración no jerarquizados también se llaman métodos de aglomeración de K-medias y normalmente utilizan una de las siguientes aproximaciones para asignar las observaciones individuales de uno de los conglomerados:

- **Umbral secuencial**. Este método selecciona una semilla de conglomerado e incluye todos los individuos que caen dentro de una distancia previamente especificada. Cuando todos los objetos dentro de la distancia están incluidos, se selecciona una segunda semilla de conglomerado y se incluyen todos los individuos dentro de la distancia previamente especificada. Cuando un individuo se incluye en un conglomerado con una semilla, no se considera a efectos de posteriores semillas. En este método, la primera semilla es la primera observación del conjunto de datos sin valores perdidos. La segunda semilla es la siguiente observación del conjunto sin valores perdidos, que se separa de la primera semilla mediante una distancia mínima especificada. Por esto, los resultados del conglomerado inicial y probablemente del final dependerán del orden de las observaciones en el conjunto de datos y arrastrar el orden de los datos es como afectar a los resultados; aunque la opción de especificar las semillas de conglomerado iniciales puede reducir este problema. Cada objeto ya asignado no se considera para posteriores asignaciones. En general, los programas de computadora ofrecen la opción por defecto que considera una distancia mínima igual a cero.

- **Umbral paralelo**. En contraste, este método selecciona varias semillas de conglomerado simultáneamente al principio y asigna individuos dentro de la distancia umbral hasta la semilla más cercana. A medida que el proceso avanza, se pueden ajustar las distancias umbral para incluir más o menos individuos en los conglomerados. También, en algunas variantes de este método, los objetos permanecen fuera de los conglomerados si están fuera de la distancia previamente especificada desde cualquiera de las semillas de conglomerado.

  En los programas de computación, para este método se establece que los puntos de semilla pueden ser aportados por el usuario o seleccionados aleatoriamente de las observaciones. No obstante, la selección aleatoria de las semillas de conglomerado producirá diferentes resultados para cada conjunto de puntos de semilla aleatorios; luego el investigador deberá estar consciente del impacto del procedimiento de selección de las semillas de conglomerado en los resultados finales.

- **Optimización**. Este método es parecido a los otros dos, excepto que permite la reubicación de los individuos. O sea, si en el curso de la asignación de los individuos, uno de ellos se acerca más a otro conglomerado que no es el que tiene asignado en este momento, entonces un procedimiento de optimización cambia el individuo hacia el conglomerado más parecido (cercano).

4.2.2. Selección de puntos de semilla

Los procedimientos no jerárquicos se encuentran disponibles en varios programas informáticos, incluyendo los principales programas estadísticos. El procedimiento del umbral secuencial (por ejemplo, el programa FASTCLUS en SAS) es un ejemplo de programa de formación de conglomerados no jerarquizado diseñado para conjuntos con gran cantidad de datos. Una vez que el investigador especifica el número máximo de conglomerados permitidos, el procedimiento comienza con la selección de semillas de conglomerados, que se utilizan como conjeturas iniciales de las medias de los conglomerados.

La primera semilla es la primera observación del conjunto de datos sin valores perdidos. La segunda semilla es la siguiente observación completa (sin datos perdidos) que se separa de la primera semilla mediante una distancia mínima especificada. La opción por defecto es una distancia mínima de cero. Una vez que se han seleccionado todas las semillas, el programa asigna cada observación al conglomerado con las semillas más próxima.

El investigador puede especificar que los conglomerados de semillas se revisen (actualicen) mediante el cálculo de medias de los conglomerados de semillas cada vez que se asigna una observación. Como contraste, los métodos del umbral paralelo (por ejemplo, QUICK-CLUSTER en SPSS) establecen los puntos de semilla como puntos aportados por el usuario o seleccionado aleatoriamente de las observaciones.

El principal problema a que se enfrentan todos los métodos de formación de conglomerados no jerárquicos es cómo seleccionar las semillas de conglomerado. Por ejemplo, con una opción de umbral secuencial, los resultados del conglomerado inicial y probablemente del final dependerán del orden de las observaciones en el conjunto de datos y arrastrar el orden de los datos es cómo afectar a los resultados. La especificación de las semillas de conglomerado iniciales, como se hace en el procedimiento de umbral secuencial, puede reducir este problema.

Pero incluso la selección aleatoria de las semillas de conglomerado producirá diferentes resultados para cada conjunto de puntos de semilla aleatorios. Por tanto, el investigador debe ser consciente del impacto del proceso de selección de las semillas de conglomerados en los resultados finales.

4.3 Análisis de conglomerados en 2 pasos

En tiempo más reciente se ha desarrollado un método de análisis *cluster* que permite evitar ciertas limitaciones de los métodos tradicionales. El *TwoStep* análisis *cluster* o método en dos etapas desarrollado por Zhang, Ramakrishnan y Livny (1996) se basa en una metodología denominada BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*), cuyas ventajas principales son las siguientes:

1. Es especialmente eficiente cuando se analiza un gran número de observaciones.
2. Es un método iterativo que permite un aprovechamiento mayor de la información a medida que se ejecuta el algoritmo en el cual se basa.

3. Es el primer algoritmo de análisis *cluster* que permite el análisis de los *outliers*, por lo que es posible el estudio más adecuado de los grupos estimados.
4. Es un método local (frente a los tradicionales métodos globales), en el cual, la decisión de agrupación refleja la cercanía entre los puntos sin recurrir al análisis de toda la nube.
5. No asigna la misma importancia a cada uno de los puntos, ya que, en el espacio, éstos no están uniformemente distribuidos.
6. Permite la selección automática del número de clusters. Este modelo compara los valores de un criterio de modelo escogido contra diferentes soluciones de conglomerados y automáticamente determina el número de clusters óptimo.
7. Calcula la escalabilidad a partir de un Árbol de Características de los Conglomerados que resume los pasos dados, permitiendo analizar ficheros muy grandes.
8. Finalmente, permite tratar de manera diferenciada las variables continuas y las categóricas, de manera que las distancias para el primer tipo de variables se calculan con base en su media y su varianza, mientras que para las variables no continuas permite incluir las frecuencias.

Se recomienda el cálculo de correlaciones bivariadas para probar la independencia entre dos variables continuas y la prueba de Chi-cuadrado para probar la independencia de dos variables categóricas.

También se debe realizar un análisis de varianza para probar la independencia entre una variable continua y una categórica, así como probar la normalidad de las variables continuas y en caso de no encontrar normalidad, realizar la estandarización de dichas variables.

Debe utilizarse la prueba de bondad de ajuste con el estadístico Chi-cuadrado para corroborar si las variables categóricas siguen una distribución multinomial.

Así, en el primer paso cada una de las observaciones es pre-agrupada a través de distancias cuantificadas por el logaritmo de la verosimilitud o por la distancia euclidiana (si solamente contempla variables continuas), generándose un árbol de características (CF). Los subclusters resultantes se agregan posteriormente, en el segundo paso, comparando sus distancias con un umbral específico. De esta manera si la distancia es mayor que el umbral, los dos clusters se fusionan. La distancia entre dos *clusters j* y *s* se define como la reducción en el logaritmo de la verosimilitud debida a la fusión de dos *clusters*, es decir:

$$d(j,s) = \xi_j + \xi_s - \xi_{<j,s>} \qquad \textbf{(4.7)}$$

Donde :

$$\xi_v = N_v \left(\sum_{k=1}^{K^A} 1/2 \log \left(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2 \right) + \sum_{k=1}^{K^B} \hat{E}_{vk}^2 \right)$$

y de aquí :

$$\hat{E}_{vk}^2 = -\sum_{I=1}^{Lk} \frac{N_{vkl}}{N_v} \log \frac{N_{vkI}}{N_v}$$

siendo: K^A el número total de variables continuas, K^B el número total de variables categóricas, L_k el número de categorías de cada una de las *k*-ésimas variables categóricas, N_j el número de observaciones del cluster $j, \hat{\sigma}^2{}_k$ la varianza de la *k*-ésima variable continúa en la base original y, finalmente $j, \hat{\sigma}^2{}_k$ la varianza de la *k*-ésima variable continua en el *cluster j*, N_{jkl} es el número de observaciones en el cluster *j* cuya *k*-ésima variable categórica toma la *l*-ésima categoría y < *j,s* > representa

el cluster formado por la unión de los clusters *j* y *s*.

Para el cálculo del logaritmo de la verosimilitud se asume que las variables continuas están normalmente distribuidas y las categóricas siguen una distribución multinomial. Chiu, Fang, Chen, Wang, y Jeris (2001) desde una perspectiva teórica y Ma y Kockleman (2005) desde una perspectiva aplicada, adoptan el método BIRCH siendo el árbol de características típico CF_j para un cluster *Cj* el siguiente:

$$CF_j = \left\{N_j, s_{AJ}, s_{Aj}^2, N_{bj}\right\} \tag{4.8}$$

donde s_{Aj} es la suma de las variables continuas del *cluster* C_j, s_{Aj}^2 es la suma del cuadrado de las variables continuas del cluster C_j, *y* $N_{Bj} = \left(N_{Bj1}, N_{Bj2}, \Lambda, N_{Bjk}^B\right)$ el vector $\sum_{K=1}^{K^B} (L_K - 1)$ dimensional cuyo *k*-ésimo subvector es de dimensión $\left(L_K - 1\right)$

Cuando dos *clusters* C_j y C_s se fusionan, el árbol de características del cluster resultante $CF_{<j,s>}$ puede obtenerse a partir de:

$$CF_{<j,s>} = \left\{N_j + N_s, s_{Aj} + s_{As}, s_{Aj}^2 + s_{As}^2, N_{Bj} + N_{Bs}\right\} \tag{4.9}$$

El número óptimo de clusters puede determinarse utilizando, bien el Criterio de Información Bayesiano o, el de Akaike. Así, para el caso de *J* clusters, pueden obtenerse de la siguiente manera:

$$BIC(J) = -2\sum_{j=1}^{j} \xi_j + m_j \log(N)$$

$$AIC(J) = -2\sum_{j=1}^{j} \xi_j + 2m_j$$

$$Donde: \qquad m_j = J\left(2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right)$$

Por lo tanto, la información puede ser finalmente agrupada en función de sus características o atributos.

4.4 Métodos jerárquicos vs. no jerárquicos

No puede darse una respuesta definitiva a esta cuestión por dos razones. En primer lugar, el problema a investigar en ese momento puede sugerir un método u otro. En segundo lugar, lo que aprendemos con la continua aplicación de estos métodos a un contexto particular puede sugerir un método u otro como el más aconsejable para ese contexto. Las ventajas y desventajas de los métodos jerárquicos son las siguientes:

En el pasado, las técnicas jerárquicas de formación de conglomerados eran las más populares, siendo el método de Ward y el encadenamiento medio probablemente los mejores disponibles.Los procedimientos jerárquicos tienen la ventaja de ser más rápidos y llevar menos tiempo de cálculo. No obstante, con el poder de cálculo de hoy en día, incluso los computadores personales pueden manejar grandes conjuntos de datos fácilmente. Los métodos jerárquicos pueden dar una idea equivocada, sin embargo, porque combinaciones iniciales indeseables pueden persistir a lo largo del análisis y llevar a resultados

artificiales.

De interés específico es el impacto substancial de los valores atípicos sobre los métodos jerárquicos, particularmente con el método del encadenamiento completo. Para reducir esta posibilidad, el investigador puede querer realizar el análisis de *clúster* de los datos repetidas veces, eliminando los atípicos o las observaciones problemáticas.

La destrucción de casos, sin embargo, incluso aquellos que no sean atípicos, puede muchas veces distorsionar la solución. Por tanto, el investigador debe tener un cuidado extremo en la destrucción de las observaciones por la razón que sea.

También, aunque los cálculos de los procesos de formación de conglomerados son relativamente rápidos, los métodos jerárquicos no son susceptibles de analizar muestras muy grandes. A medida que aumenta el tamaño de la muestra, los requisitos de almacenamiento de datos aumenta enormemente por ejemplo, una muestra de 400 casos exige el almacenamiento de aproximadamente 80,000 similitudes que se incrementan a 125,000 para una muestra de 500.

Incluso con los avances tecnológicos actuales, problemas de este calibre exceden la capacidad de la mayoría de las computadoras actuales, limitando por tanto la aplicación en muchos casos. Se puede considerar una muestra aleatoria de las observaciones originales para reducir el tamaño de la muestra pero debe cuestionarse ahora la representatividad de la muestra tomada de la muestra original.

4.5 Elección del número de grupos o conglomerados

La determinación del número final de conglomerados a formar es también conocida como la **"regla de paro"**, y no existe un procedimiento objetivo o estándar para su determinación.

Existen diversos métodos de determinación del número de grupos: algunos están basados en intentar reconstruir la matriz de distancias original, otros, en los coeficientes de concordancia de Kendall y otros realizan análisis de la varianza entre los grupos obtenidos. No existe un criterio universalmente aceptado.

Dado que la mayor parte de los programas estadísticos proporciona las distancias de aglomeración, es decir, las distancias a las que se forma cada grupo, una forma de determinar el número de grupos consiste en localizar en qué iteraciones del método utilizado dichas distancias pegan grandes saltos. Con dichas distancias se pueden utilizar criterios como el *criterio de Mojena* que determina el primer **S** $\in$ **N** tal que $\alpha_{s+1} > \overline{\alpha} + ks_{\alpha}$ si se utilizan distancias y si son similitudes donde $\{\alpha_j ; j=1,...,n-1\}$ son las distancias de aglomeración, $\overline{\alpha}$, s_{α} su media y su desviación típica respectivamente y k una constante entre 2.5 y 3.5.

Al no existir un criterio estadístico para decidir, los investigadores han desarrollado varios criterios y líneas a seguir para aproximarse a la solución del problema.

Una regla de paro que es relativamente simple examina alguna medida de similitud entre los conglomerados a cada paso sucesivo, considerando como solución cuando la medida de similitud excede a un valor especificado o cuando los valores sucesivos entre los pasos dan un salto súbito.

Entonces se selecciona la solución clúster previa a dicho salto ya que esa combinación provocó la sustancial reducción en su similitud. Existen otros criterios más sofisticados, como el “criterio cúbico de elaboración de conglomerados” (CCC), aunque no se ha encontrado ninguno que sea mejor en todas las situaciones.

A veces es necesario complementar estas reglas con un juicio meramente empírico con cualquier conceptualización de las relaciones teóricas que pueda sugerir un número natural de conglomerados.

Es decir, puede ser interesante para un trabajo específico el contar con 5 conglomerados, o con 3 y a continuación, y después de repetir el análisis para ambos, seleccionar la mejor alternativa utilizando criterios a priori, juicios prácticos, sentido común o fundamentos teóricos. Las soluciones se verán mejoradas mediante la restricción de la solución de acuerdo con los aspectos conceptuales del problema.

4.6 Interpretación de los conglomerados

Interpretar la clasificación obtenida por un Análisis Cluster requiere, en primer lugar, un conocimiento suficiente del problema analizado. Hay que estar abierto a la posibilidad de que no todos los grupos obtenidos tienen por qué ser significativos. Algunas ideas que pueden ser útiles en la interpretación de los resultados son las siguientes:

- Realizar ANOVAS y MANOVAS para ver qué grupos son significativamente distintos y en qué variables lo son,
- Realizar Análisis Discriminantes,
- Realizar un Análisis Factorial o de Componentes Principales para representar, gráficamente los grupos obtenidos y observar las diferencias existentes entre ellos, y
- Calcular perfiles medios por grupos y compararlos

El paso de la interpretación implica el examen de cada conglomerado en términos del valor teórico del conglomerado o asignar una etiqueta precisa que describa la naturaleza de los conglomerados.

Determinar qué caracteriza cada conglomerado, es decir, cuáles son sus perfiles y su interpretación, es la parte más importante del análisis, ya que proporcionan un medio de evaluar la correspondencia de los conglomerados de aquellos propuestos por una teoría a priori o por la experiencia práctica.

Si se utiliza de forma confirmatoria, los perfiles del análisis clúster ofrecen un medio directo de evaluación de la correspondencia, comparando los conglomerados derivados con una tipología preconcebida.

La agrupación exacta no es una tarea sencilla y es difícil hacer recomendaciones generales. Siempre es aconsejable intentar con más de un método. Si varios métodos dan resultados semejantes, entonces se puede suponer que en realidad existen agrupaciones naturales.

CAPÍTULO 5. EJEMPLOS DE APLICACIÓN

En lo que sigue, se analizarán los pasos a seguir para llevar a cabo un Análisis Cluster, ilustrándolos con aplicaciones al Análisis Económico Internacional.

5.1 Ejemplo (Clasificación de países de la UE)

En este ejemplo los datos corresponden a la situación de 6 países europeos en 1996 con respecto a los 4 criterios exigidos por la UE para entrar en la Unión Monetaria: Inflación, Interés, Déficit Público y Deuda Pública y vienen dados en la tabla siguiente:

Tabla 5.1 Tabla de datos ejemplo de la UE.

País	Inflación	Interés	Déficit	Deuda
Alemania	1	1	1	0
España	1	1	1	0
Francia	1	1	1	1
Grecia	0	0	0	0
Italia	1	1	0	0
Reino Unido	1	1	0	1

Fuente: Salvador Figueras, M (s.f.). "Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

Este es un ejemplo en el que todas las variables son binarias de forma que, este caso 1 significa que el país sí satisfacía el criterio exigido y 0 que no lo satisfacía. En este caso todas las variables son binarias simétricas y se puede utilizar como medida de distancia la distancia euclidiana al cuadrado. La matriz de distancias obtenida viene dada por:

Tabla 5.2 Matriz de distancias obtenidas con la distancia Euclidiana al cuadrado

	Al	Es	Fr	Gr	It	RU
Al	0	0	1	3	1	2
Es		0	1	3	1	2
Fr			0	4	2	1
Gr				0	2	3
It					0	1
RU						0

Fuente: Salvador Figueras, M (s.f.). "Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

Así, por ejemplo, la distancia entre España y Francia es 1 puesto que solamente difieren en un criterio: el de la deuda pública que Francia satisfacía y España no. Los resultados de aplicar un método jerárquico aglomerativo con enlace completo utilizando el programa estadístico SPSS 12.0 se muestran a continuación:

Tabla 5.3 Historial de conglomeración

Historial de conglomeración

	Conglomerado que se combina			Etapa en la que el conglomerado aparece por primera vez		
Etapa	Conglomerado 1	Conglomerado 2	Coeficientes	Conglomerado 1	Conglomerado 2	Próxima etapa
1	1	2	0	0	0	3
2	5	6	1	0	0	4
3	1	3	1	1	0	4
4	1	5	2	3	2	5
5	1	4	4	4	0	0

Fuente: Salvador Figueras, M (s.f.). "Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

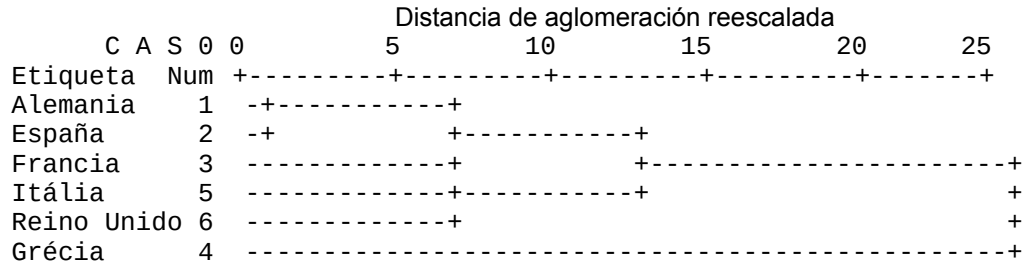


Figura 5.1 Diagrama de arbol (Dendograma)

El historial de aglomeración muestra las distancias de aglomeración y los grupos que se han ido formando al aplicar el algoritmo. El diagrama de témpanos y el dendograma dan dicha información de forma gráfica. Así, en el primer paso del algoritmo se unieron Alemania y España a una distancia de aglomeración igual a 0. Posteriormente, a dicho grupo, se unió Francia e Italia y Reino Unido formaron otro grupo, todo ello a una distancia de aglomeración igual a 1.

Estos dos grupos se unieron formando un único grupo a una distancia de aglomeración igual a 2. Finalmente Grecia se unió a todos los demás países a una distancia de aglomeración igual a 4, la máxima posible. Si tomamos como punto de corte 1 nos quedaríamos con 3 grupos: {España, Alemania y Francia}, {Italia, Reino Unido} y {Grecia}. Estos grupos están formados por países que difieren entre sí en a lo más un criterio.

5.2 Ejemplo (Clasificación de países de la UE)

Este ejemplo corresponde a datos sobre diversas variables económicas, sanitarias y demográficas correspondientes a 102 países del mundo en el año 1995. Dichas variables vienen detalladas en la siguiente tabla:

Tabla 5.4 Variables utilizadas (económicas, sanitarias y demográficas correspondientes a 102 países)

Variable	Significado
POB	Logaritmo de la Población
DENS	Logaritmo de la Densidad
ESPF	Logaritmo de 83-Esperanza de vida Femenina
ESPM	Logaritmo de 78 - Esperanza de vida masculina
ALF	Logaritmo de 101-Tasa de Alfabetización
MINF	Logaritmo de la Tasa de Mortalidad Infantil
PIBCA	Logaritmo del PIB per cápita
NACDEF	Logaritmo de Nacimientos/Defunciones
FERT	Logaritmo del número medio de hijos por mujer

Fuente: Salvador Figueras, M (s.f.). "Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

En los dos ejemplos el objetivo es el mismo: encontrar grupos de países que muestren un comportamiento similar con respecto a las variables analizadas.

En este caso todas las variables son cuantitativas pero medidas en diferentes unidades. Por esta razón utilizaremos la distancia euclidiana pero con los datos estandarizados previamente.

En la figura 5.2 se muestran las distancias de aglomeración del algoritmo jerárquico aglomerativo tomando como función de enlace, el enlace intergrupos y utilizando el programa estadístico SPSS 12.0

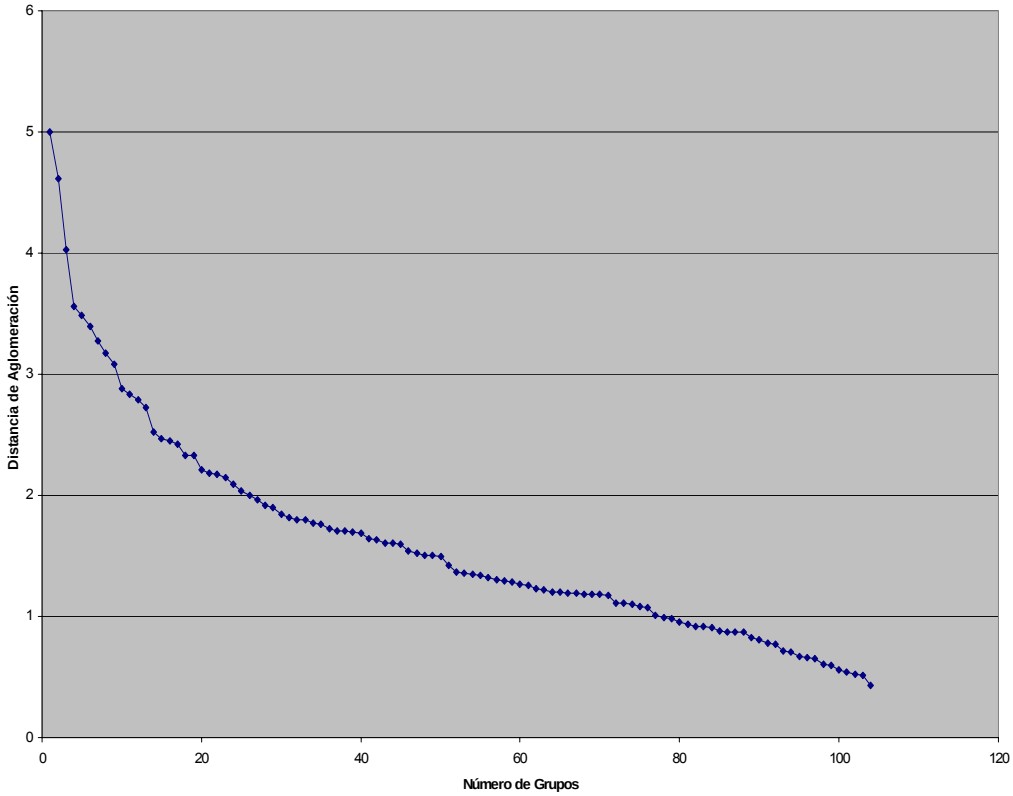


Figura 5.2 Distancias de aglomeración

Se observa que los mayores saltos se dan cuando el algoritmo pasa de 4 a 3, 3 a 2 y 2 a 1 grupo. El criterio de Mojena aplicado con k=2.5 da una distancia de corte igual a 3.83 y selecciona un número de grupos igual a 4. Por todas estas razones toma como número de grupos 4.

La aplicación del método de K-medias es conveniente utilizarlo cuando los datos a clasificar son muchos y/o para refinar una clasificación obtenida utilizando un método jerárquico. Supone que el número de grupos es conocido a priori.

Existen varias formas de implementarlo pero todas ellas siguen, básicamente, los siguientes pasos:

1) Se seleccionan k centroides o semillas donde k es el número de grupos deseado

2) Se asigna cada observación al grupo cuya semilla es la más cercana

3) Se calculan los puntos semillas o centroides de cada grupo

4) Se iteran los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, los puntos semillas apenas cambian o los grupos obtenidos en dos iteraciones consecutivas son los mismos.

El método suele ser muy sensible a la solución inicial dada por lo que es conveniente utilizar una que sea buena. Una forma de construirla es mediante una clasificación obtenida por un algoritmo jerárquico

Los resultados de aplicar el algoritmo de las k-medias implementado en SPSS 12.0, con un número de grupos igual a 4 y tomando como punto de partida los centroides de los grupos obtenidos anteriormente vienen dados por las siguientes tablas y gráficos. El algoritmo converge en 10 iteraciones y obtiene 4 grupos de tamaños 24, 39, 1 y 41 países respectivamente.

Tabla 5.5 Historial de Iteraciones

	Cambio en los centros de los conglomerados			
Iteración	1	2	3	4
1	,592	,109	1,036E-07	,172
2	,487	6,262E-02	,000	,125
3	,214	,000	,000	4,648E-02
4	,231	,000	,000	5,287E-02
5	,225	6,193E-02	,000	3,981E-02
6	,306	5,276E-02	,000	9,411E-02
7	,235	,000	,000	9,347E-02
8	,250	6,932E-02	,000	,115
9	,227	7,083E-02	,000	,121
10	,305	,174	,000	5,141E-02

Fuente: Salvador Figueras, M (s.f.). “Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

Las iteraciones se han detenido porque se ha llevado a cabo el número máximo de iteraciones. Las iteraciones no han convergido. La distancia máxima en la que han cambiado los centros es .172. La iteración actual es 10. La distancia mínima entre los centros iniciales es 3.007.

En la tabla siguiente se muestran los países miembros de cada grupo junto con las distancias de cada país al centroide de su grupo. Así mismo se muestran las distancias entre los centroides de cada grupo.

Se observa que los grupos 1 y 4 contienen países del tercer mundo, el grupo 2 está compuesto por países del primer y segundo mundos y el grupo 3 contiene únicamente a Islandia.

Tabla 5.6 Grupos obtenidos

PAIS	GRUPO	DISTANCIA
Venezuela	1	1,10992
Ecuador	1	1,17341
Malasia	1	1,19941
Panamá	1	1,24843
Acerbaján	1	1,27096
Colombia	1	1,31659
Armenia	1	1,33676
Chile	1	1,36857
Rep. Dominicana	1	1,49939
Turquía	1	1,57329
Uzbekistán	1	1,65333
Líbano	1	1,67326
México	1	1,69396
Tailandia	1	1,81748
El Salvador	1	1,81842
Corea del Norte	1	1,82812
Paraguay	1	1,88032
Jordania	1	1,90393
Argentina	1	2,05071
Emiratos Árabes	1	2,26097
Corea del Sur	1	2,28927
Costa Rica	1	2,56727
Kuwait	1	2,5803
Bahrein	1	2,78161
Austria	2	0,84751
Irlanda	2	1,02262
Dinamarca	2	1,03776
Croacia	2	1,17118
Bélgica	2	1,25977
Finlandia	2	1,29839
Grecia	2	1,39139

Polonia	2	1,39569
España	2	1,41288
Lituania	2	1,42745
Hungía	2	1,43235
Portugal	2	1,45946
Bielorusia	2	1,47973
Gran Bretaña	2	1,53294
Bulgaria	2	1,53866
Georgia	2	1,62389
Nueva Zelanda	2	1,68732
Suecia	2	1,69381
Rumanía	2	1,69529
Italia	2	1,71363
Alemania	2	1,71408
Países Bajos	2	1,77523
Noruega	2	1,83862
Uruguay	2	1,93886
Cuba	2	1,94022
Francia	2	1,98214
Estonia	2	2,01381
Letonia	2	2,02654
Suiza	2	2,04078
Ucrania	2	2,19731
Estados Unidos	2	2,30185
Canadá	2	2,60291
Australia	2	2,69585
Israel	2	2,71955
Rusia	2	2,89912
Japón	2	3,11629
Barbados	2	3,15042
Singapur	2	3,48935
Hong Kong	2	3,75342
Islandia	3	0,0000

Camerún	4	0,57933
Senegal	4	0,72504
Kenia	4	0,81205
Egipto	4	1,01448
Guatemala	4	1,07179
Camboya	4	1,17287
Marruecos	4	1,34473
Burkina Faso	4	1,35581
Nicaragua	4	1,40744
Tanzania	4	1,44743
Irán	4	1,45366
Nigeria	4	1,47222
Iraq	4	1,50176
Sudáfrica	4	1,51414
Perú	4	1,53181
Liberia	4	1,54648
Bolivia	4	1,56759
Uganda	4	1,57074
Honduras	4	1,58019
Zambia	4	1,58128
Etiopía	4	1,68095
Pakistán	4	1,69868
Afganistán	4	1,73597
Somalia	4	1,78696
Siria	4	1,86294
Haití	4	1,86689
Burundi	4	1,99972
Filipinas	4	2,03681
Indonesia	4	2,12085
Ruanda	4	2,13195
Vietnam	4	2,14496
Gambia	4	2,31622
Brasil	4	2,31901

Rep. C. Africana	4	2,41386
Arabia Saudí	4	2,4842
Bangladesh	4	2,5958
Libia	4	2,77066
Gabón	4	2,94421
India	4	2,96665
Botswana	4	2,96857
China	4	3,63459

Fuente: Salvador Figueras, M (s.f.). "Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

Tabla 5.7 Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3	4
1		3,038	5,466	2,534
2	3,038		4,233	4,967
3	5,466	4,233		7,460
4	2,594	4,967	7,460	

Fuente: Salvador Figueras, M (s.f.). "Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

5.2.1 Interpretación de los resultados

En la tabla siguiente se muestran los resultados de aplicar un ANOVA para cada una de las variables analizadas. Se observa que existen diferencias significativas en todas las variables al 1 y al 5% con excepción de las variables POB y DENS en las que solamente existen diferencias al 5%.

Tabla 5.8 Análisis de Varianza

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Puntua(POB)	3,563	3	,941	101	3,786	,013
Puntua(DENS)	3,086	3	,929	101	3,321	,023
Puntua(ESPF)	26,744	3	,263	101	101,702	,000
Puntua(ESPM)	23,077	3	,375	101	61,560	,000
Puntua(ALF)	26,760	3	,254	101	105,301	,000
Puntua(MINF)	28,487	3	,180	101	157,954	,000
Puntua(PIBCA)	23,533	3	,355	101	66,341	,000
Puntua(NACDE)	23,794	3	,303	101	78,483	,000
Puntua(FERT)	26,351	3	,238	101	110,829	,000

Fuente: Salvador Figueras, M (s.f.). "Análisis de conglomerados o cluster".<http://www.5campus.org/leccion/cluster>

Las pruebas F se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los conglomerados son iguales.

Los dos gráficos siguientes muestran los perfiles medio de cada grupo y los diagramas de cajas de las variables analizadas para cada uno de los grupos. Se observa que los países de los grupos 1 y 4 poseen una menor renta per cápita y peores indicadores los índices de alfabetización, mortalidad y esperanza de vida así como una mayor fertilidad y natalidad que la de los países de los grupos 2 y 3. Siendo estas diferencias más acusadas en los países del grupo 4 que la de los grupo 1. También queda de manifiesto el carácter atípico de Islandia debido a su baja natalidad, mortalidad infantil, población y densidad y su alta alfabetización, esperanza de vida.

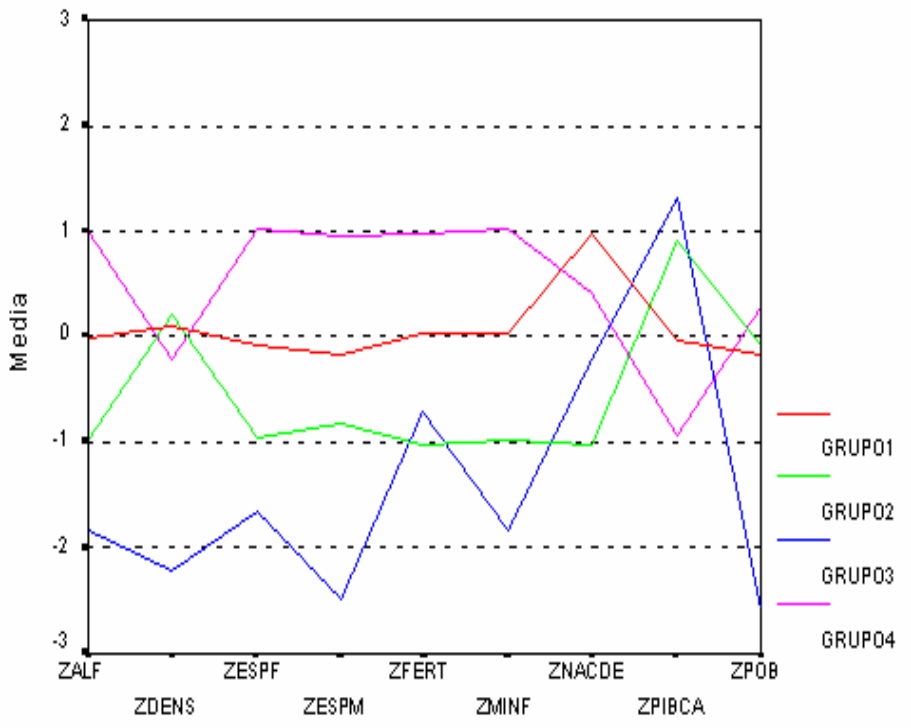


Figura 5.3 Perfiles medios de cada grupos

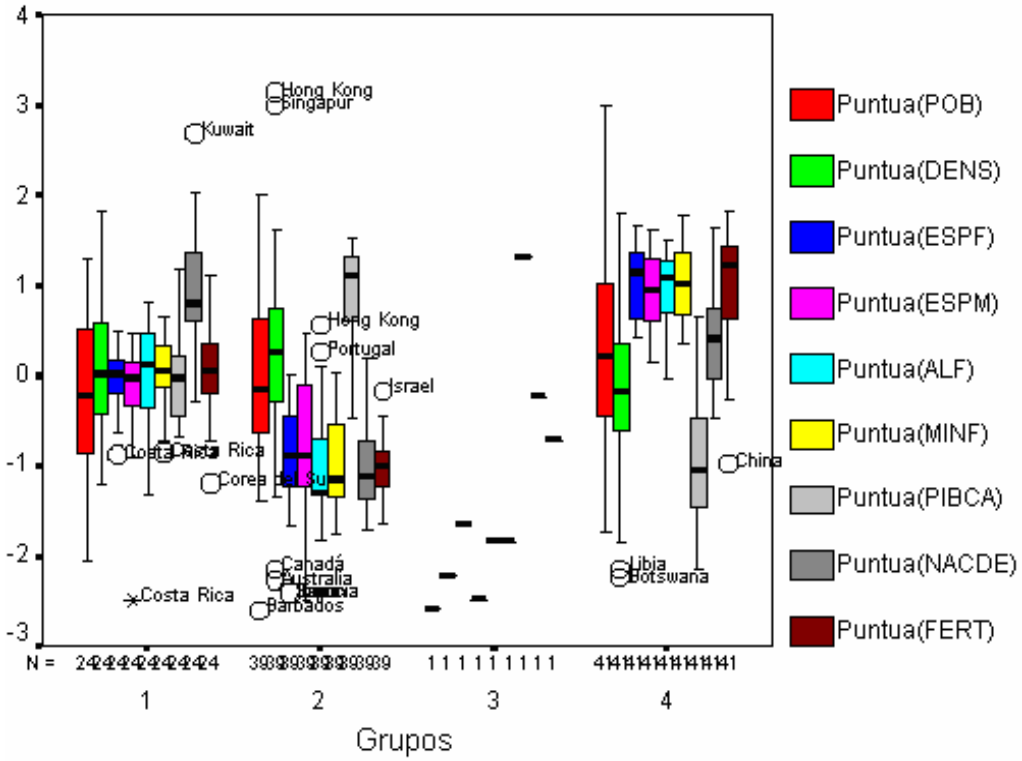


Figura 5.4 Diagrama de cajas correspondiente a cada grupo

5.2.2 Validación de la solución

Una vez obtenidos los grupos e interpretado los resultados conviene, siempre que sea posible, proceder a la validación de los mismos con el fin de averiguar, por un lado, hasta qué punto los resultados obtenidos son extrapolables a la población de la que vienen los objetos seleccionados y, por el otro, por qué han aparecido dichos grupos. Esta validación se puede realizar de forma externa o interna.

5.2.2.1 Validez interna

Se puede establecer utilizando procedimientos de validación cruzada. Para ello se dividen los datos en dos grupos y se aplica el algoritmo de clasificación a cada grupo comparando los resultados obtenidos en cada grupo. Por ejemplo, si el método utilizado es el de las k-medias se asignaría cada objeto de uno de los grupos al cluster más cercano obtenido al clasificar los datos el otro grupo y se mediría el grado de acuerdo entre las clasificaciones obtenidas utilizando los dos métodos

5.2.2.2 Validez externa

Se puede realizar comparando los resultados obtenidos con un criterio externo (por ejemplo, clasificaciones obtenidas por evaluadores independientes o analizando en los grupos obtenidos, el comportamiento de variables no utilizadas en el proceso de clasificación) o realizando un Análisis Cluster con una muestra diferente de la realizada.

En los 3 gráficos siguientes se muestra la composición de cada grupo por religión mayoritaria, región económica y clima predominante. Se observa que la mayor parte de los países cristianos pertenecen al grupo 2 siendo esta diferencia más clara en los cristianos ortodoxos y protestantes.

Por otro lado, los países musulmanes y los que practican otras religiones están en los grupos 1 y 4. Los países budistas se distribuyen equitativamente en los 3 grupos

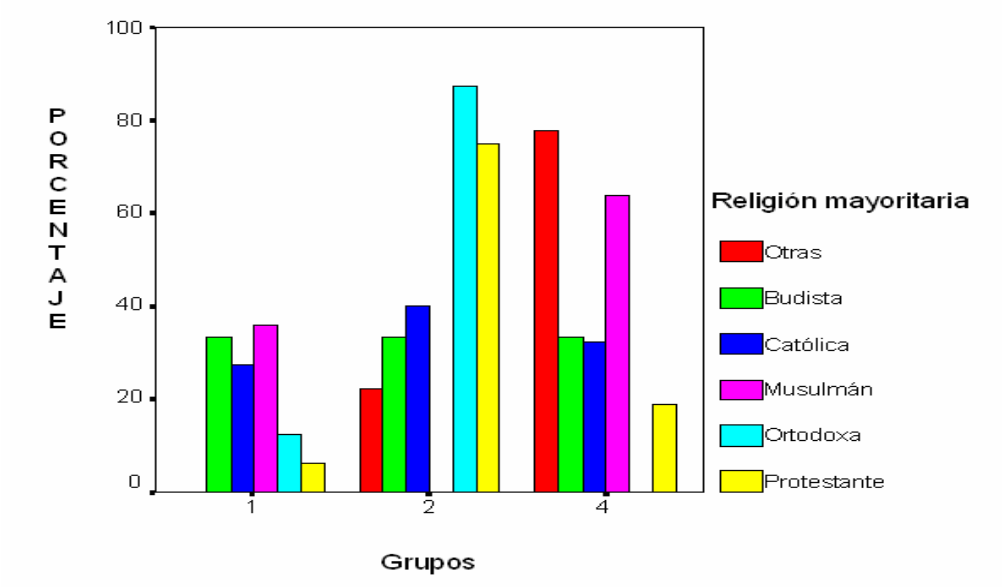


Figura 5.5 Composición de los grupos por religión

Por regiones económicas, los países del primer y segundo mundos (OCDE y Europa Oriental) pertenecen todos al segundo grupo, los países de América Latina y Oriente Medio tiende a estar en el grupo 1 mientras que todos los países africanos y la mayor parte de los países de Asia están incluidos en el grupo 4. Los grupos reflejan, por lo tanto, las diferencias existentes entre las diversas regiones económicas del mundo.

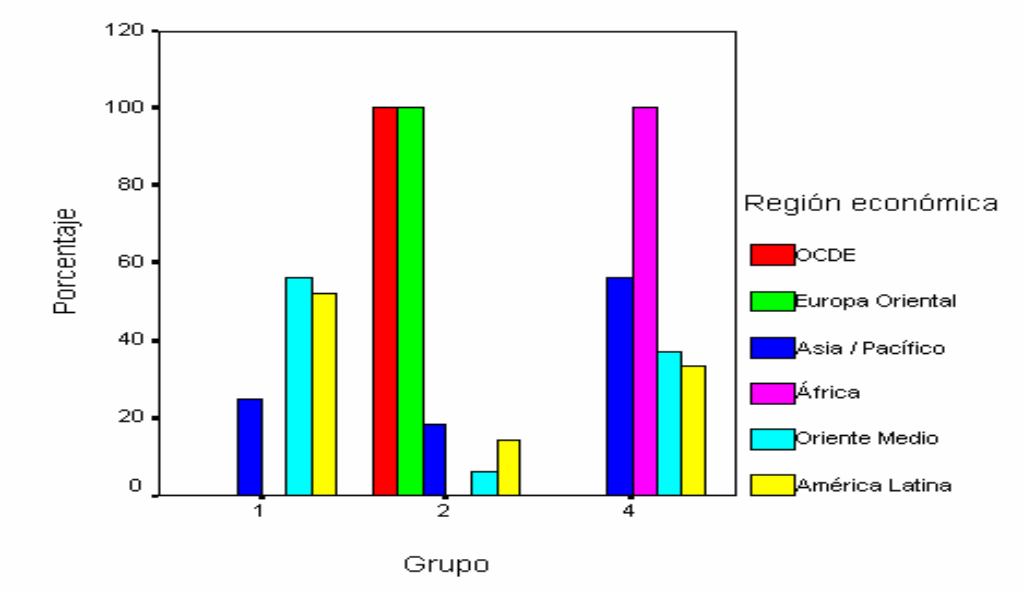


Figura 5.6 Composición de los grupos por región económica

En la figura 5.6 pone de manifiesto la influencia del clima en la composición de los grupos. La mayor parte de los países con climas templados y frío pertenecen al grupo 2 mientras que los países con clima desértico, ecuatorial y tropical tienden a estar en el grupo 4 y los de clima árido en el grupo 1.

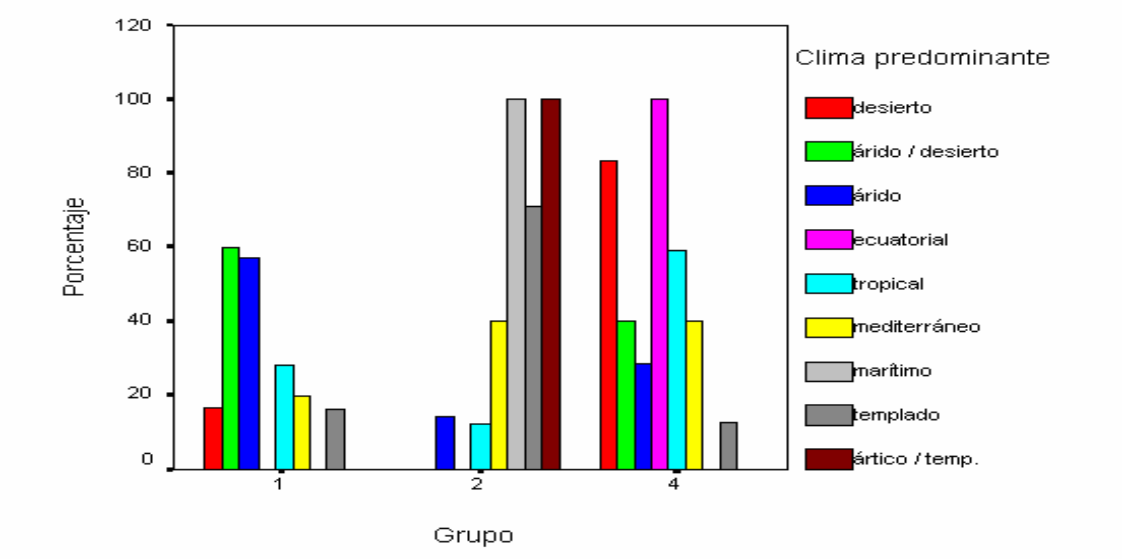


Figura 5.7 Composición de los grupos por clima predominante

5.3 Ejemplo de aplicación en industrias dentro de la región de Pachuca

A continuación se utilizará una base de datos sobre las micro, pequeñas y medianas empresas del sector textil y la confección de la región Pachuca realizada entre septiembre del 2001 y marzo del 2002, por estudiantes del Instituto Tecnológico Regional de Pachuca y dirigido por los Profesores – Investigadores de CIAII, UAEH; Mtro. Heriberto Niccolás Morales, Mtro. Jaime Garnica Gonzáles, Dr. Gilberto Pérez Lechuga y el Ing. Germán Reséndiz López.

De dicha base de datos se han tomado solamente las siguientes variables: Personal Especializado, Calidad de Materia Prima, Nivel de Exigencia con la Calidad de sus Productos y Aplicación de Sistema de Calidad. Con ellas se utilizarán diferentes medidas de semejanza o distancia y diferentes técnicas de agrupación a modo de ejemplo.

Tabla 5.9 Variables utilizadas y tipo de variable

Variables	Tipo de variable
Personal Especializado	Nominal (Si y No)
Estudios de Distribución de Planta	Nominal (Sí y No)
Estudios de Distribución de Servicios	Nominal (Si y No)
Aplicación de Sistema de Calidad	Nominal (Sí y No)

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

En este ejemplo los datos corresponden a la situación de 12 empresas en 1996 con respecto a los 4 criterios: Personal Especializado, Calidad de Materia Prima, Nivel de Exigencia con la Calidad de sus Productos y Aplicación de Sistema de Calidad y vienen dados en la siguiente tabla:

Tabla 5.10 Matriz de datos

ID	AÑO DE INICIO	Personal Especializado	Estudios de Distribución De Planta	Estudios de Distribución de Servicios	Aplicación de Sistema de Calidad
1	1945	1	1	2	2
2	1992	1	1	1	1
3	2001	2	2	2	2
4	1922	2	2	2	1
5	1996	2	2	2	2
6	1998	1	1	2	1
7	1989	2	1	2	1
8	1998	2	2	2	2
9	1994	1	2	1	2
10	1971	2	1	2	2
11	1965	2	2	2	2
12	1935	2	2	2	2

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

Se realizaron todas las combinaciones entre métodos de clasificación jerárquicos y medidas de distancia posibles, eliminando posteriormente aquellos cuyos resultados no fueron satisfactorios. No se incluyeron todos los tipos de enlace ya que la naturaleza de las variables eran de naturaleza nominal y además hay métodos de clasificación que no se recomiendan para algunas medidas de distancia.

Finalmente los mejores agrupamientos fueron los conglomerados jerárquicos utilizando la vinculación completa con la distancia euclidiana al cuadrado y la distancia de City-block, no encontrándose diferencia ninguna con las distancias euclidianas y Minkowsky.

El método de vinculación simple con cualquiera de las distancias utilizadas produjo clusters encadenados y el método de vinculación promedio también con cualquier distancia arrojó clusters demasiado compactos.

A continuación aparecen los detalles fundamentales del análisis jerárquico utilizando la vinculación completa con la distancia euclidiana al cuadrado, que resultó ser la mejor clasificación.

Tabla 5.11 Historial de conglomeración

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	8	12	.000	0	0	2
2	3	8	.000	0	1	4
3	4	7	.000	0	0	7
4	3	5	.000	2	0	7
5	1	9	1.000	0	0	9
6	2	6	1.000	0	0	10
7	3	4	1.000	4	3	9
8	10	11	3.000	0	0	11
9	1	3	3.000	5	7	10
10	1	2	7.000	9	6	11
11	1	10	10.000	10	8	0

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

En el siguiente dendograma o diagrama de arbol se muestral como se forman los conglomerados

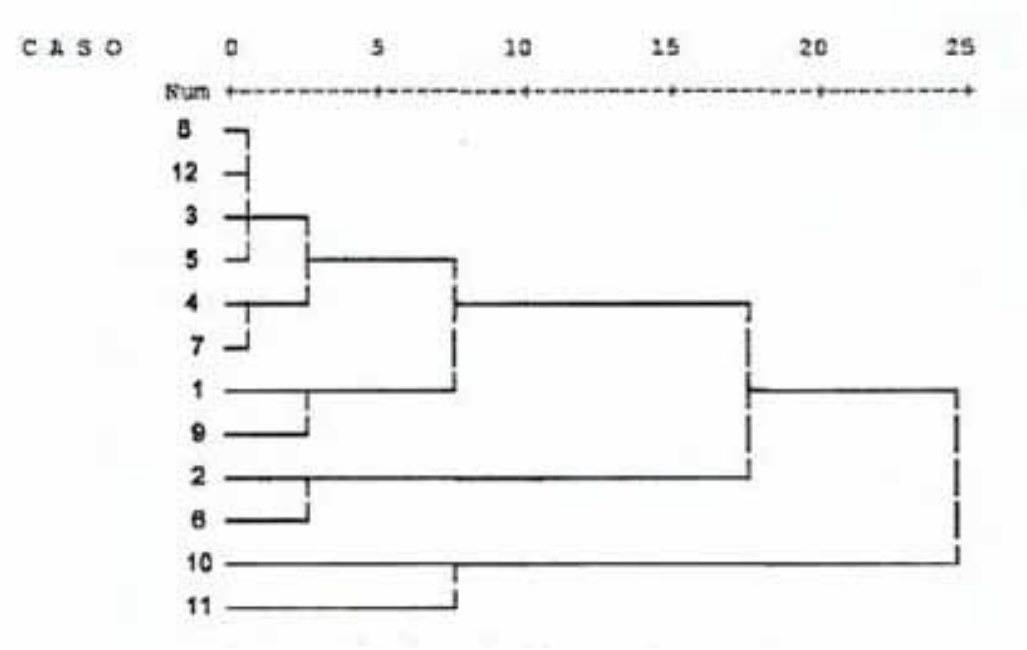


Figura 5.8 Dendograma utilizando la vinculación completa

5.4 Análisis de conglomerados de K medias

Este método es recomendable cuando los datos a clasificar son muchos o para refinar una clasificación obtenida utilizando un método jerárquico previamente. Nuestra base de datos sólo comprende 12 empresas, luego su utilización se justifica para confirmar los resultados obtenidos previamente.

Como la solución del método jerárquico aportó tres grupos bien diferenciados, se decidió utilizar $k = 3$ puntos semilla o centroides iniciales. El algoritmo convergió en 2 iteraciones y se obtuvieron 3 grupos de tamaños 2, 3 y 7 empresas respectivamente, como se puede apreciar en las tablas siguientes.

El agrupamiento final coincidió en gran medida con el cluster utilizando el método jerárquico, ya que en el grupo 3 se encontraron las mismas 7 empresas (1, 3, 4, 5, 7, 8 y 12), en el grupo 2 se encontraron las empresas 2, 6 y 9 y en el grupo 1 se encontraron las empresas 10 y 11.

La única diferencia detectada fue con la empresa 9 que anteriormente se encontraba en el grupo 1, aunque en una posición limítrofe, y con el procedimiento de K-medias se pudo concluir que pertenece al grupo 2.

Tabla 5.12 Número de casos en cada conglomerado

Conglomerado	1	2.000
	2	3.000
	3	7.000
Válidos		12.000
Perdidos		.000

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

Tabla 5.13 Pertenencia a los conglomerados

Número de caso	Identificacion de la Empresa	Conglomerado	Distancia
1	1	3	.904
2	2	2	.816
3	3	3	.319
4	4	3	.728
5	5	3	.319
6	6	2	.577
7	7	3	.728
8	8	3	.319
9	9	2	1.000
10	10	1	.866
11	11	1	.866
12	12	3	.319

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

La interpretación de la clasificación obtenida por un Análisis de Conglomerados requiere del conocimiento del problema analizado. Como es posible que los grupos obtenidos no sean realmente significativos entre sí se realizó un ANOVA para ver si existían diferencias entre dichos grupos. Como se puede observar en la siguiente tabla, los resultados arrojaron diferencias entre grupos (P< .01) para todas las variables, excepto para la variable: aplicación de sistemas de calidad.

Tabla 5.14 Análisis de la varianza (Anova)

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Tiene personal especializado en manejo de tecnología	1.446	2	.151	9	9.592	.006
Calidad de materia prima que ofrecen sus proveedores	1.875	2	.130	9	14.464	.002
Nivel de exigencia con la calidad de sus productos	1.875	2	.130	9	14.464	.002
Aplican sistema de calidad	.286	2	.233	9	1.227	.338

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

Las pruebas F sólo se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

5.5 Conglomerado en dos pasos (Two Step)

Una de las características fundamentales de este tipo de análisis es que es muy útil para la clasificación de grandes bases de datos. Como nuestra base sólo cuenta con 12 empresas, el procedimiento que seguiremos será a modo de ejemplo. La base de datos para este caso tiene que contar con variables cuantitativas y categóricas por lo que utilizamos las variables, cantidad de personal calificado, porcentajes de personal femenino y masculino y la calificación del personal, como se aprecia a continuación.

Tabla 5.15 Variables Estandarizadas

	Tipo de empresa	Personal	Per_fem	Per_masc	Calif_per
1	Zapato militar industrial	2	10	90	6
2	Camisas, pijama	2	90	10	8
3	Ropa interior	3	80	20	10
4	Material triturado e hidratado	4	1	99	10
5	Camisas de caballero	3	80	20	9
6	Escudos	1	50	50	8
7	Uniformes deportivos	3	85	15	10
8	Cordeles	1	80	20	10
9	Pantalón	3	90	10	8
10	Uniformes industriales	2	85	15	7
11	Corsetería	2	62	38	10
12	Uniformes para trabajo	3	90	10	7

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

Este tipo de conglomerados se realizó con variables diferentes, luego no se podra esperar los mismos resultados. Los conglomerados

uno y dos sólo contaron con una empresa que evidentemente presenta características en su personal bastante diferentes del resto que se encuentra concentrado en el cluster número tres. En el cluster 1 sólo se encuentra la empresa número 4 y el cluster 2 la empresa número 1.

Tabla 5.16 Distribución del Cluster

		N	% de Combinación	% del Total
Cluster	1	1	8.3%	8.3%
	2	1	8.3%	8.3%
	3	10	83.3%	83.3%
	Combinado	12	100%	100%
Total		12		100%

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

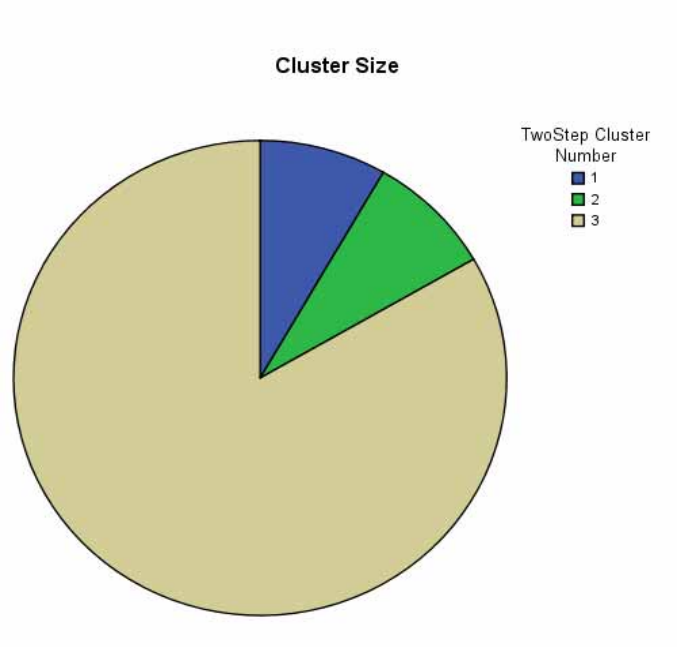


Figura 5.9 Tamaño del Cluster

Tabla 5.17 Centroides

Centroids

	numero de personal que labora en la planta		Porcentaje de mujeres (distribucion de personal)		Porcentaje de hombres		calificacion de la capacitacion de personal	
	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Cluster 1	4.00	.	1.00	.	99.00	.	10.00	.
2	2.00	.	10.00	.	90.00	.	6.00	.
3	2.30	.823	79.20	13.198	20.80	13.198	8.70	1.252
Combined	2.42	.900	66.92	31.132	33.08	31.132	8.58	1.443

Fuente: Tabla propia realizada de una base de datos sobre las micro, pequeñas y medianas empresas corrida del SPSS

Como se puede apreciar a continuación la variación dentro de cada cluster por variable solo es interesante en el cluster 3 ya que los restantes tienen solo un individuo. Las variables que más influyeron fueron la capacitación del personal y el personal calificado en las empresas.

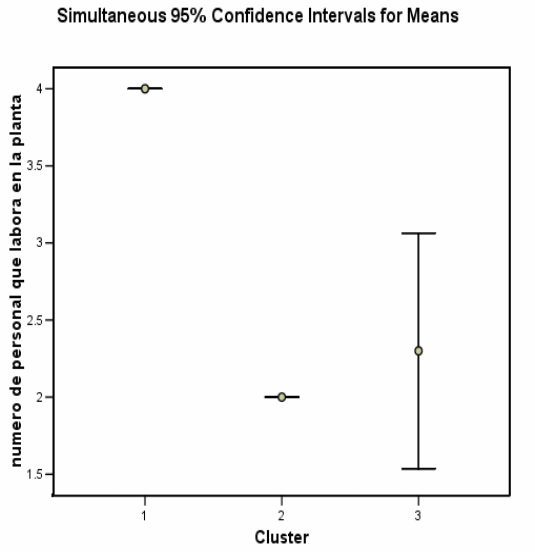


Figura 5.10 Intervalos de confianza para medias para el numero de personal que labora en la planta

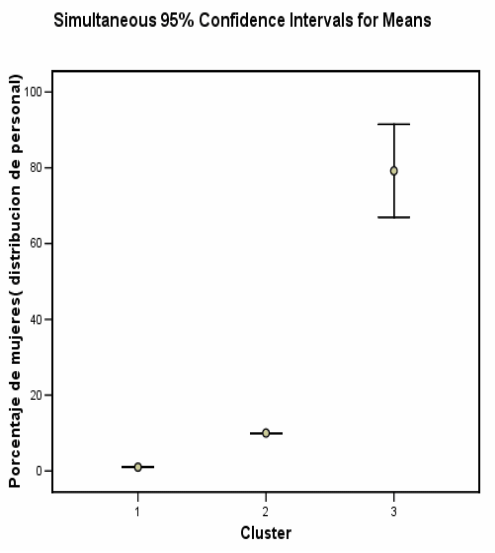


Figura 5.11 Intervalos de confianza para medias para el porcentaje de mujeres.

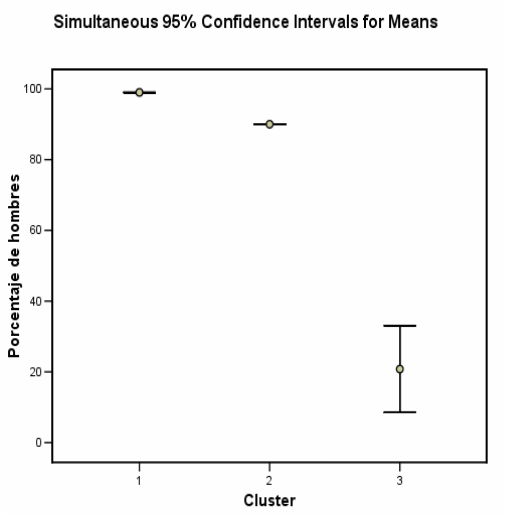


Figura 5.12 Intervalos de confianza para medias para el porcentaje de hombres

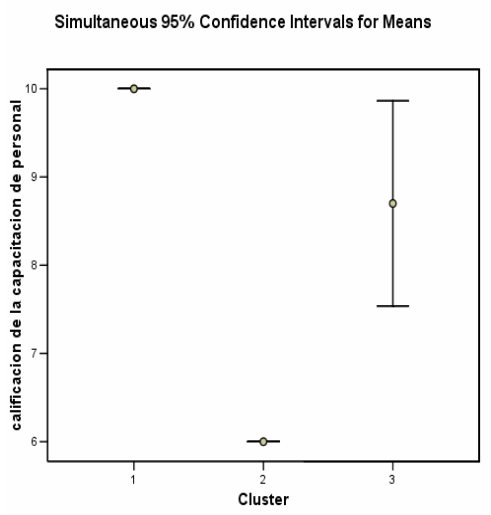


Figura 5.13 Intervalos de confianza para medias para la calificación de la capacitación de personal

El conglomerado en 2 pasos es una técnica confirmatorio de grupos anteriormente detectados utilizando un análisis de conglomerados jerárquico y apropiado cuando el número de objetos a clasificar es muy grande; no obstante lo hemos querido utilizar para incluir dos tipos de variables (que es una de sus bondades más importante) y así demostrar la importancia de su utilización.

Como se utilizaron diferentes variables que en los análisis anteriores, no se pueden comparar los resultados con los analizados utilizando el conglomerado jerárquico y el conglomerado de K-medias.

CONCLUSIONES Y RECOMENDACIONES

A continuación presentaremos las características fundamentales, las bases, el modo de empleo, las ventajas y las desventajas de cada una de las múltiples variantes de las técnicas de clasificación estadística, así como la interpretación de algunas de ellas, las cuales fueron utilizadas como ejemplo.

- El análisis de conglomerados (***cluster análisis)*** es la denominación de un grupo de técnicas multivariantes cuyo principal propósito es agrupar individuos u objetos basándose en las características o descriptores que poseen.

- Si la clasificación es acertada, los conglomerados poseen un alto grado de homogeneidad interna (dentro del conglomerado) y un alto grado de heterogeneidad externa (entre conglomerados).

- Este análisis multidimensional es denominado como análisis Q, construcción de tipologías, análisis de clasificación y taxonomía numérica según el uso de los métodos de agrupación en disciplinas tan diversas como psicología, biología, sociología, economía, ingeniería y negocios.

- Las desventajas fundamentales de este grupo de técnicas es que son técnicas exploratorias, descriptivas, y no inferenciales, por lo que deben utilizarse como paso previo a la confirmación de los conglomerados con otra técnica multivariante.

- Seleccionar adecuadamente la medida de similitud o semejanza teniendo en cuenta la naturaleza de las variables incluidas en la clasificación,

- Seleccionar el procedimiento de pertenencia al grupo de cada objeto,

- Definir cuántos grupos deseamos obtener; es decir, hallar el equilibrio entre la definición de las estructuras mas básicas (pocos conglomerados) pero que mantienen el necesario nivel de similitud dentro de los conglomerados.

- El análisis de conglomerados puede verse como un modelo en 6 pasos de los cuales los tres primeros se corresponden con los objetivos, el cuarto con la selección de un algoritmo de cluster, el quinto con la interpretación de los mismos y el sexto con la validación y perfiles de los clusters. Los pasos son los siguientes:

    1 Descripción de una taxonomía.
    2 Simplificación de los datos.
    3 Identificación de relaciones.
    4 Selección de un algoritmo de cluster
    5 Interpretación de los clusters
    6 Validación y perfiles de los clusters.

- La selección de la medida de proximidad o de distancia es uno de los problemas más complejos del análisis de conglomerados., por la diversidad de opciones que existen.

- Las **medidas de proximidad, similitud o semejanza** miden el grado de semejanza entre dos objetos de forma que, cuanto mayor (respecto al menor) es su valor, mayor (respecto al menor) es el grado de similaridad existente entre ellos y con más (respectivamente menos) probabilidad los métodos de clasificación tenderán a incluirlos en el mismo grupo.

- Las **medidas de disimilitud, desemejanza o distancia** miden la distancia entre dos objetos de forma que, cuanto mayor (respecto al menor) sea su valor, más (respectivamente menos) diferentes son los objetos y menor (respecto al mayor) la probabilidad de que los métodos de clasificación los incluyan en el mismo grupo.

- Existen dos grandes categorías de algoritmos de obtención de conglomerados: los jerarquizados (aglomerativos y divisivos) y los no jerarquizados.

- Los algoritmos más utilizados para los métodos jerárquicos son el de ligamiento simple, el de ligamiento completo, el de ligamiento medio, el enlace medio dentro de los grupos, el método de Ward, el método del centroide y el método de la mediana.

- Las características fundamentales de estos algoritmos es que el enlace simple conduce a clusters encadenados, el enlace completo conduce a clusters compactos, el enlace completo es menos sensible a valores atípicos que el enlace simple, el método de Ward y el método del enlace medio son los menos sensibles a valores atípicos y el método de Ward tiene tendencia a formar clusters más compactos y de igual tamaño y forma en comparación con el enlace medio.

- Para saber si realmente los conglomerados formados constituyen grupos homogéneos hay muchas técnicas que van desde los coeficientes de concordancia de Kendall hasta la realización de análisis de la varianza entre los grupos obtenidos, aunque no existe un criterio universalmente aceptado.

- Dado que la mayor parte de los paquetes estadísticos proporciona las distancias de aglomeración, es decir, las distancias a las que se forma cada grupo, una forma de determinar el número de grupos consiste en localizar en qué iteraciones del método utilizado dichas distancias pegan grandes saltos. Utilizando dichas distancias se pueden utilizar criterios como el *criterio de Mojena* que determina el primer $S \in \mathbb{N}$ tal que $\alpha_{s+1} > \overline{\alpha} + ks_{\alpha}$ si se utilizan distancias y < si son similitudes donde $\{\alpha_j ; j=1,...,n-1\}$ son las distancias de aglomeración, $\overline{\alpha}$, S_{α} su media y su desviación típica respectivamente y k una constante entre 2.5 y 3.5.

- A veces es necesario complementar estas reglas con un juicio meramente empírico con cualquier conceptualización de las relaciones teóricas que pueda sugerir un número natural de conglomerados. Es decir, puede ser interesante para un trabajo específico el contar con 5 conglomerados, o con 3 y a continuación, y después de repetir el análisis para ambos, seleccionar la mejor alternativa utilizando criterios a priori, juicios prácticos, sentido común o fundamentos teóricos. Las soluciones se verán mejoradas mediante la restricción de la solución de acuerdo con los aspectos conceptuales del problema.

- Los algoritmos no jerárquicos o métodos de aglomeración de K-medias son muy importantes cuando el número de objetos a clasificar es muy grande. En general, para su análisis se consideran 4 pasos, que son:

    - Se seleccionan k centroides o semillas donde k es el número de grupos deseado
    - Se asigna cada observación al grupo cuya semilla es la más cercana
    - Se calculan los puntos semillas o centroides de cada grupo, y
    - Se iteran los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, los puntos semillas apenas cambian o los grupos obtenidos en dos iteraciones consecutivas son los mismos.

- Estos métodos suelen ser muy sensibles a la solución inicial dada, por lo que es conveniente utilizar una que sea buena. Se recomienda seleccionar la solución inicial mediante una clasificación obtenida por un algoritmo jerárquico.

- Los procedimientos de aglomeración no jerarquizados normalmente utilizan una de las siguientes aproximaciones para asignar las observaciones individuales de uno de los conglomerados: Umbral secuencial, Umbral paralelo y Optimización.

- El principal problema a que se enfrentan todos los métodos de formación de conglomerados no jerárquicos es cómo seleccionar las semillas de conglomerado. Por ejemplo, con una opción de umbral secuencial, los resultados del conglomerado inicial y probablemente del final dependerán del orden de las observaciones en el conjunto de datos, y arrastrar el orden de los datos es como afectar a los resultados; aunque la opción de

especificar las semillas de conglomerado iniciales puede reducir este problema. Cada objeto ya asignado no se considera para posteriores asignaciones. En general, los programas de computadora ofrecen la opción por defecto que considera una distancia mínima igual a cero.

- El método de conglomerado en dos pasos permite el análisis conjunto de variables cuantitativas y cualitativas, cuestión que no es posible en los modelos jerárquicos o en el de K-medias, lo que lo hace un método mucho más general, además de que permite la solución óptima del número de conglomerados y la utilización de grandes bases de datos.

- Determinar la característica de cada conglomerado, es decir, cuáles son sus perfiles y su interpretación, es la parte más importante del análisis, ya que proporciona un medio de evaluar la correspondencia de los conglomerados de aquellos propuestos por una teoría a priori o por la experiencia práctica. Si se utiliza de forma confirmatoria, los perfiles del análisis cluster ofrecen un medio directo de evaluación de la correspondencia, comparando los conglomerados derivados con una tipología preconcebida.

- La agrupación exacta de un cluster no es una tarea sencilla y es difícil hacer recomendaciones generales. Siempre es aconsejable intentar con más de un método. Si varios métodos dan resultados semejantes, entonces se puede suponer que en realidad existen agrupaciones naturales.

- Recomendamos el uso de otros métodos de clasificación que permiten obtener agrupamientos más precisos entre los que se encuentran los métodos de la lógica difusa (fuzzy logic) para variables cuantitativas y los modelos de redes neuronales, que permiten múltiples consideraciones y se puede obtener una clasificación óptima. La utilización de algún método de validación (validación cruzada, jacknife o bootstrap) para conocer la repetibilidad y estabilidad de las respuestas cuando se han utilizado diversos modelos de clasificación.

- Recomendamos incluir las técnicas de clasificación en los programas de estudio de la licenciatura en Ingeniería Industrial, debido a su importancia y amplio campo de aplicaciones.

- Recomendamos también la utilización del software estadístico SPSS en la enseñanza de las asignaturas de Estadística en la carrera de Licenciatura en Ingeniería Industrial por sus amplias posibilidades de aplicación.

Aportes del trabajo de monografía:

- Estudio y aprendizaje del significado de las técnicas de clasificación, y de su importancia para la aplicación en diversos campos de la ciencia.
- Estudio de las condiciones previas de la información para seleccionar las medidas de distancia o semejanza, las técnicas de conglomerados y los métodos de enlace a aplicar según el caso.
- Entrenamiento en el uso del software estadístico SPSS para la aplicación de las técnicas de clasificación.
- Detección de diversos usos y posibles aplicaciones en el campo de la ingeniería.

BIBLIOGRAFÍA

Alderfer, Mark S., y Roger K. Blashfield .1984. *Cluster Analysis.* Thousand Oaks.: Sage Publications.

Anderberg, M. 1973. Cluster Analysis for Applications. New York: Academic Press.

Bailey, Kenneth D. 1994. Typologies and Taxonomies: An Introduction to Classification Techniques, Thousand Oaks.: Sage Publications

Bray, R. J. & J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27: 325 -349. [287, 535,569,600,633,707]

Bloom, S. A. 198. similarity indices in community studies: potencial pitfalls. *Mars. Ecol. Prog. Ser.* 5: 125 -128. [287,298]

Cattell, R.B. 1952. *Factor analysis* – An introduction and manual for the psychologist and social scientist. Harper,New York. 462 pp [248]

Cattel, R. B. 1966. The data box: its ordering of total resources in terms of possible relational systems. 67 – 128 in: R. B. Cattell [ed] Handbok of multivariate experimental psychology. Rand McNally & Co., Chicago. [248, 249]

Clark, P. J. 1952. An extension of the coefficient of divergence fr use with multiple characters. Copeia 1952: 61-64 [283]

Czekanowski, J. 1909. Zur Differentialdiagnose der Neandertalgruppe. Krrespondenz- Blatt deutsch.Ges. Anthropol. Ethnol.Urgesch. 40: 44-47. [xiii,265,282,371]

Dagnielie, P. 1975. L'analyse statistique á plusier variables. Les Presses agronomiques de Gembloux(Belgique).362 pp. [184]

Dallas E. Jonson. 1998. *Metodos Multivariados aplicados al analisis de datos.* Selecciones Empresariales.

Dallas E. Johnson. 2000. *Metodos multivariados aplicados al analisis de datos.* Internacional Thomson Editores S. A de C. V.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. Ecology 49: 297 – 302. [257,294]

Estabrook, G. F. 1966. A mathematical model in graph theory for biological classification. *J theor. Boil.* 12: 297 -310. [311,374]

Faith, D. P 1983. Asymmetric binary similarity maesures. Oecologia (Berl.) 57: 287- 290. [258]

Goodall, D. W. 1964. A probabilisticsimilarity index . *Nature (lond.)* 203: 1098. [269]

Goodall, D. W. 1966. A new Similarity index based on probability. Biometrics 22: 882-907. [269,269]

Gower, J. C. 1971a. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857- 871. [258, 266]

Green, P. E., and J. Douglas Carroll. 1978. Mathematical. Tools for Applied Multivariate Analysis. New York: Academic Press.

Hair, F. Jr. Anderson, Tatham, E. R. Black W. C. 1999. Analsis Multivariante. 5° ed. Pretince Hall Iberia, Madrid.

Hajdu, L. J. 1981. Geographical comparison of resemblance measures in phytosociology. *vegetatio* 48: 47 -59. [297,298]

Hotelling, H. 1931 The generalization of Student's ratio. *Ann. Math. Statist .2*: 360- 378. [281]

Jaccard, P. 1900 Contribution au probleme de l'immigration post- glaciaire de la flore alpine. *Bull. Soc. Vaudoise Sci. Nat.* 36 87 – 130. [xiii,256]

Jaccard, P. 1901. Etude comparative de la distribution florale dans portion des alpes et du Jura. *Bull. Soc, Vaudoise Sci. nat.* 37: 547 -579. [256]

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale.*bull. Soc. Vaudoise Sci. Nat.* 44: 223- 270.[256]

Kulczynski, S. 1928. Die Pflanzenassoziationen der Pieninen. *Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat. Ser. B,* Suppl. II (1927): 57 – 203. [257, 266,306, 371, 372, 373]

Lance, G. N. & W. T. y Williams. 1966c. Computer programs for classification. *Proc. ANCCAC Conference,* Canberra, May 1966, Paper 12/3. [282]

Lance, G. N. & W. T. Williams. 1967a. Mixed- data classificatory

programs. I. Agglomerative systems. Aust. Comput. *J.*1: 15-20.[282]

Lebart, L. & J. P. Fenelon. 1971. Statisque et informatique appliqueés. Dunod, Paris. 426 pp. [285,455]

Legrenge, P. & Chodorowski. 1977. A generalization of jaccard's association coefficient for Q analysis of multi-state ecologival data matrices. *Ekol. Pol.* 25*:* [297- 308. [262,263,267,271,308, 1977,484,485]

Legendre. P., S. Dallot & L. Legendre. 1985. Succession of species within a community: Chroonological clustering, with applications to marine and freshwater zooplankton. *Am. Nat.* 125: 257- 288. [297, 692, 696, 697, 698, 699, 700, 701, 7591]

Legendre, P. & D.J. Rogers. 1972. Characters and clustering in taxonomy: A synthesis of two taximetric procedures. Taxon. 21: 567 – 606. [262, 305, 312, 344]

Mahalanobis, P. C. 1936. On the generalizad distence in statics. *Proc. Natl. Inst. Sci. India* 2: 49-55. [280]

Motyka, J. 1947. O zadaniach i metodach badan geobotanicznych. Sur les buts et les methodes des recherches geobotanique. Annales Uniersitatis Mariae curie- Sklodowska (lublin polonia). Sectio C, Supplementum *I* viii + 168 pp. [265]

Moreau, G & L. Legendre. 1979. Relation entre habitat et peuplements de poissons: essai de definition d'une method numerique pour de rivieres mordiques. *Hydrobiologia* 67 : 81 – 87. [283]

Morales, N. H., Garnica G. J., Pérez L. G., Resendiz L. G. 2002. *Textos.*

57: 58 – [59].

Ochiai, A. 1957 Zoogeographic studies on the soleoid fishes foubd in japan and its neighbouring regions. *Bull. Jpn. Soc. Sci. Fish.* 22:526-530. [257]

Orlóci, L. 1967 b. Anagglomerative method for classification of plany communities. *. ecol.* 55: 193-205.[279]

Orlici, L.1975. *multivariante analysis in vegetation research*. Dr.W.Junk B. V., The Haugue. Ix + 276 pp. [252,451]

Orloci, L.1978. *multivariante analysis in vegetation research*. 2nd edition. Dr. W. Junk B. V., The Hague. ix + 451 pp. [252, 269, 270, 278, 280, 288, 371,461]

Punj, G., and D. Stewart. 1983. «Cluster Analysis Marketing Research: Review and Suggestions for Aplication. » *Journal of Marketing Research* 20 (May): 48.

Rogers, D. J. & T.T. Tanimoto. 1960. A computer program for classifying plants. Science (Wash. D. C.) 132: 1115-1118. [255]

Russell, P. F. & T. R. Rao. 1940. On habitat and association of species of anopheline larvae insouth – eastern Madras. *J. malar. Inst. India* 3: 153 - 178. [257]

Sneath, P. H. A., and R. R. Sokal .1973. Numerical *Taxonomy.* San Francisco: Freeman Press.

Sokal, R. R. & P. H. A. Snealth. 1963. Principles of numerical taxonomy. W. H. Freeman, San Francisco. Xvi + 359 pp. [xiii,252, 255, 256, 257]

Sokal, R. R. & C. d. michener. 1958. A statistical method for evaluating systematic relationships. *Univ, Kans. Sci. Bull.* 38: 1409 -1438. [255,306,321]

Sokal, R. R. & P. H. A. Snealth. 1963. Principles of numerical taxonomy. W. H. Freeman, San Francisco. Xvi + 359 pp. [xiii,252, 255, 256, 257]

Sokal, R. R. & P. H. A. Snealth. 1963. Principles of numerical taxonomy. W. H. Freeman, San Francisco. Xvi + 359 pp. [xiii,252, 255, 256, 257]

Sokal, R. R. & P. H. A. Snealth. 1963. Principles of numerical taxonomy. W. H. Freeman, San Francisco. Xvi + 359 pp. [xiii,252, 255, 256, 257]

Sorensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biol. Skr.* 5: 1- 34. [256,316,317]

Stephenson, W., W. T. Williams & S. D. Cook. 1972. Computer analyses of Petersen's original data on bottom communities. *Ecol. Monogr.*42: 387 – 415. [283]

Rao, C. R. 1995. A review of canonical coordinates anda n alternative to correspondence analysis using Hellinger distance. *Qúestiio (Quaderns d' Estadistica I Investigacio operative)* 19: 23- 63. [286]

Roux, G. & J. Reyssac 1975. essai d'application au phytoplancton marin

de methodes satatisques utilisees en phytosociologie terrestre. *Ann. Ints. Oceaonogr.* (Paris)51 : 89 -97. [283]

Whittaker,R. H. 1952. A study of summer foliage insect communities in the Great Smoky Mountains. *Ecol. Monogr.* 22: 1- 44. [282]

Wiley, J. and Sons, N.Y. 1992.Introduction to Linear Regression Analysis, 2ª edición.

Zhang, t.; ramakrishnan, r., y livny, M. 1996. "BIRCH: An efficient data clustering method for very large databases", *Proceedings de ACM SIGMOD*

ENLACES

Salvador Figueras, M (s.f.). “Análisis de conglomerados o cluster". Recuperado el 15 de febrero del 2005, del sitio web *campus.org, Estadística* de la Universidad de Zaragoza: <http://www.5campus.org/leccion/cluster>

Práctica sobre Análisis Cluster. (s.f.). Recuperado el 20 de febrero de 2005, de www.ual.es/~freche/practicas/practica7/practica7.html

Modelos y técnicas de análisis de datos. (s.f). Recuperado el 20 de febrero de 2005, de http://home-3.tiscali.nl/~xp117079/mtad/

GLOSARIO

Concordancia: (lat. med. *-ntia*) *f.* Correspondencia o conformidad de una cosa con otra.
2 *fís.* Estado de dos fenómenos vibratorios que no presentan ninguna diferencia de fase.
3 *gram.* Relación de dos o más palabras diferentes por la conformidad de accidentes.
4 mús. Justa proporción que guardan entre sí las voces que suenan juntas.
5 *f., pl.* Índice alfabético de todas las palabras de un libro, con todas las citas de los lugares en que se hallan.

Correlación: *f.* Relación recíproca o mutua entre dos o más cosas.
2 *ling.* Conjunto de dos series de fonemas opuestas por un mismo rasgo distintivo.
3 Relación que se establece entre ellas.
4 *mat.* Existencia de mayor o menor dependencia mutua entre dos variables aleatorias.

Costes: *m.* Costa (cantidad).
2 *econ.* Medida y valoración del consumo realizado o previsto por la aplicación de los factores para la obtención de un producto, trabajo o servicio.

Discriminante *adj.* Que discrimina.
2 *adj.-s.* Función especial de las raíces de una ecuación expresada en términos de sus coeficientes.

Exploratoria: *adj.-m.* Que sirve para explorar.
2 *med.* Instrumento o medio que se emplea para explorar cavidades o heridas en el cuerpo.
Heterogeneidad: *f.* Calidad de heterogéneo.

Heterogéneo: (lat. *heterogeneu* ← gr. *heterogenés* ← *hetero-* + *génos,* género)*adj.* Compuesto de partes de diversa naturaleza.

2 Diferente.

Homocedasticidad: Término estadístico que significa igualdad de varianzas entre grupos (Contrario: **Heterocedasticidad)**.

Homogeneidad: *f.* Calidad de homogéneo.

Homogéneo, -ea: (b. lat. *homogeneus* ← gr. *homogenés* ← + *génos,* género)
adj. Relativo a un mismo género.
2 Formado por elementos de igual naturaleza.
3 fig. Muy junto o espeso.
4 *quím.* [sistema] Que consta de una sola fase.

Interdependencia: (*inter-* + *dependencia f.* Dependencia mutua entre personas, entidades, naciones, principios, etc.: *la ~ económica de los países europeos.*

Interdependientes: *adj.* Que tiene interdependencia.

Marketing: (voz inglesa) *m.* Mercadotecnia.

Paradoja: lat.-gr. *-oxa*)

. Especie opuesta a la opinión común y, esp., la que parece opuesta siendo exacta.
2 Aserción inverosímil presentada con apariencias de verdadera.

Parámetros: *para-* + *metro*)
m. Línea constante e invariable que entra en la ecuación de algunas curvas, esp. en la de la parábola.
2 Variable tal que otras variables pueden ser expresadas por funciones de ella.

3 fig. Elemento importante cuyo conocimiento es necesario para comprender un problema o un asunto.

Parsimonia: f. (lat. Parsimonia). Frugalidad, moderación, escasez:

Priori: lat. med., por lo que precede)
loc. adv. fil. [conocimiento] Independiente de la experiencia, es decir, que ésta supone pero no puede explicar, aunque sea necesario a la posibilidad de la experiencia; *a priori* no designa, pues, una anterioridad psicológica, sino una anterioridad lógica o de validez.

2 *fil.* En la filosofía escolástica, [razonamiento] que desciende de la causa al efecto, o de la esencia de una cosa a sus propiedades.

Segmentación: . Acción de segmentar o segmentarse.

2 Efecto de segmentar o segmentarse.
3 División en fragmentos.
4 Técnica de división de un programa en partes denominadas segmentos a fin de no requerir la presencia simultánea de la totalidad del programa a la memoria del ordenador.

5 *biol.* División de la célula huevo de animales y plantas, en virtud de la cual se constituye un cuerpo pluricelular, que es la primera fase del embrión.

Sesgado: de *sesgo* II) *adj.* p. us. Término estadístico que significa falto de simetría con respecto a la distribución normal

Sesgo: probl. de *sesgo* II)
adj. Torcido, cortado o situado oblicuamente: *al ~,* oblicuamente, al través.
2 fig. Grave o torcido en el semblante.
3 *m.* Oblicuidad o torcimiento de una cosa hacia un lado.

SIN. *1* **Soslayado, oblicuo.**

Taxonomía: Conjunto de principios y métodos sobre la clasificación de individuos, animales, plantas u objetos, así como los resultados obtenidos

Tipología: *tipo- + -logía)*
f. Estudio y clasificación de tipos que se practica en diversas ciencias.
2 Ciencia que estudia los distintos tipos raciales en que se divide la especie humana.
3 *med.* Ciencia que estudia los varios tipos de la morfología del hombre en relación con sus funciones vegetativas y psíquicas.
4 Tipología lingüística, disciplina que compara las lenguas para clasificarlas y establecer entre ellas relaciones, genealógicas o no, según las afinidades de sus sistemas fonológicos, morfológicos y sintácticos, etc.