Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes

Trabajo Fin de Máster

"Two-step Cluster" en SPSS y técnicas relacionadas

Autor: Carlos Santos Mangudo

Tutora: Dra. Mª Purificación Galindo Villardón

2015

Dpto. de Estadística

Universidad de Salamanca

DRA. Mª PURIFICACIÓN GALINDO VILLARDÓN

Profesora Titular del Departamento de Estadística de la Universidad de Salamanca

---

CERTIFICA que **D. Carlos Santos Mangudo** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que para optar título de Máster en Análisis Avanzado de Datos Multivariantes, presenta con el título ***"Two-step Cluster" en SPSS y técnicas relacionadas*** autorizando expresamente su lectura y defensa.

Y para que conste, firman el presente certificado en Salamanca a 7 de julio de 2015.

Mª Purificacion Galindo Villardón

“Two-step Cluster” en SPSS y técnicas relacionadas

Dpto. de Estadística

Universidad de Salamanca

Trabajo para optar al título de Máster en

Análisis Avanzado de Datos Multivariantes

por la Universidad de Salamanca.

Presenta

:

Carlos Santos Mangudo

Salamanca

2015

Si no escalas la montaña,

Jamás podrás disfrutar del paisaje

(Pablo Neruda)

Los sueños parecen al principio imposibles,

Luego improbables

Y luego, cuando nos comprometemos

Se vuelven inevitables

(Mahatma Gandhi)

AGRADECIMIENTOS

Quiero expresar en primer lugar mi agradecimiento a mi directora la Dra. Mª Purificación Galindo Villardón por el gran apoyo y estímulo que me ha brindado durante todo el proceso de formación en este Master, y posteriormente en la elaboración del presente Trabajo Fin de Master, por sus enseñanzas, por su confianza y por su amistad, y sobre todo por ser parte del tren de mi vida.

Mi especial gratitud a todos los profesores del Departamento de Estadística de la Universidad de Salamanca, especialmente al Dr. D. José Luis Vicente Villardón director del Departamento, por su generosidad, por sus valiosos conocimientos y por estar siempre dispuestos a ayudarme en todo lo que he necesitado.

A mi esposa Amadora, por comprenderme y darme ese apoyo vital cuando todo se vuelve oscuro.

Por último, a mis hijas Lorena y Andrea por la fuerza que me dan en todos los momentos difíciles por los que he pasado.

Contenido

Índice General

Contenido *11*

Resumen *16*

Summary *16*

Introducción y Objetivos *17*

Introducción **19**

Objetivos **24**

Capítulo I - Análisis de Clúster *25*

1.1. Introducción **27**

1.2. Etapas del Análisis de Clúster **31**

1.3. Selección de las Variables **32**

1.4. Selección de la Medida de Distancia o Similaridad **33**

1.4.1. Medidas de distancia de tipo Cuantitativo 35

1.4.2. Medidas de Similaridad de tipo Cualitativo 36

1.4.3. Medidas de Similaridad de tipo Mixto 39

1.4.4. Correlación entre individuos 40

1.4.5. Distancias derivada de la Distancia Chi-cuadrado ($\chi 2$) 40

1.5. Selección de la Técnica Clúster de Clasificación **41**

1.6. Validación de los Resultados **41**

1.7. Métodos de Clasificación de tipo Jerárquico 45
1.7.1. Método Single-Linkage (vecino más próximo) 47
1.7.2. Método Complete-Linkage (vecino más lejano) 48
1.7.3. Método Average-Linkage (de la media) 49
1.7.4. Método del Centroide 50
1.7.5. Método de la Mediana 51
1.7.6. Método de Ward 52
1.8. Método de Clasificación de tipo No Jerárquico 53
1.8.1. Método de K-Medias 54
1.8.2. Método de Quick-Clúster 56
1.8.3. Método de Forgy 56
1.8.4. Método de las Nubes Dinámicas 58
1.8.5. Método de Densidad 59
1.8.6. Método de Block-Clustering 61
1.8.7. Método de Reducción de Dimensiones 62
1.9. Método de Clasificación de tipo Two-Step 63
Capítulo II - Método TWO-STEP 65
2.1. Introducción 67
2.2. Método BIRCH 70
2.2.1. Algoritmo del Método BIRCH 71
2.3. Algoritmo del Método Two-Step 73
2.3.1. Formación del Pre-Clúster (FASE 1) 74
2.3.2. Agrupamiento del Árbol de Características (FASE 2) 77
2.4. Cálculo para la Medida de la Distancia 77
2.4.1. Distancia Euclidea 78
2.4.2. Distancia Máxima Verosimilitud 78
2.5. Criterios de Agrupamiento 80
2.5.1. Criterio de AKAIKE (AIC) 80
2.5.2. Criterio de SCHWARTZ (BIC) 82
2.6. Selección del Número de Clusters 83
2.7. Tratamiento de Valores Atípicos 84
2.8. Importancia del Predictor en el Agrupamiento 85
2.9. Ajuste del Árbol de Características (CF) 87

2.10. Representación Gráfica 88

2.11. Aplicaciones del Método TWO-STEP 94

Capítulo III - Clúster HJ-BIPLOT *vs* TWO-STEP 97

3.1. Introducción 99

3.2. Tipo de Variables 101

3.3. Algoritmo del Clúster HJ-BIPLOT 102

3.4. Criterio de Agrupamiento 103

3.5. Representación Gráfica 105

3.6. Comparativa TWO-STEP vs Clúster HJ-BIPLOT 106

3.6.1. Comparación relativa a la información de partida 107

3.6.2. Comparación relativa al Algoritmo 108

3.6.3. Comparación relativa a la Representación Gráfica 109

3.6.4. Limitaciones HJ *vs* TWO-STEP 112

Capítulo IV CLUSPLOT *vs* TWO-STEP & Clúster HJ-BIPLOT 115

4.1. Introducción 117

4.2. Algoritmo del Método CLUSPLOT 119

4.2.1. Cálculo de la Matriz de Disimilaridades 122

4.2.2. Asignación del número de Clúster 123

4.3. Criterio de Agrupamiento 124

4.4. Comparación CLUSPLOT *vs* TWO-STEP & Clúster HJ-BIPLOT 128

4.4.1. Comparación relativa a la Información de partida 129

4.4.2. Comparación relativa al Algoritmo 130

4.4.3. Comparación relativa a la Representación Gráfica 133

4.4.4. Limitaciones CLUSPLOT 136

4.4.5. Resumen Comparativo de los 3 Métodos 137

Conclusiones 139

Bibliografía 143

Resumen

El presente Trabajo Fin de Master consiste en realizar un estudio sobre el funcionamiento y características del análisis de Clúster Bietápico o en Dos Fases del Método TWO-STEP así como un análisis comparativo de este método con respecto al Análisis de Clúster HJ-BIPLOT y CLUSPLOT, utilizados como técnicas de agrupamiento y clasificación de volúmenes de datos, que nos permitan agrupar entidades homogéneas dependiendo de una serie de variables, analizando las características de cada uno de los métodos desde un punto de vista teórico y práctico.

Summary

The present work end of Master's degree consists to carry out a study on the operation and characteristics of Two-Step Clúster Analysis Method, as well as a comparative analysis of this method with repect to the Cluster Analysis HJ-Biplot and Clusplot Clúster Analysis, used as clustering techniques and classification of Data Volumes, allow us to form homogeneous entities depending of several variables, analyzing the characterisics of each of the methods from a theroretical and practical perspective.

Palabras Clave:

Analisis de Cluster, Bietápico, Two-Step, HJ-Biplot, ClusPlot

Key Words:

Cluster Analysis, Two-Step, HJ-Biplot, ClusPlot

Introducción y Objetivos

Introducción

A lo largo del tiempo, la humanidad siempre ha querido dividir y clasificar todo lo que nos rodea, desde los animales, la religión, las sociedades, el universo, etc., todo era susceptible de ser dividido y a su vez clasificado. Aristóteles construyó un sistema para clasificar especies de animales, los que tenían sangre roja y los que no la tenían, que era en definitiva la clasificación de los animales en vertebrados e invertebrados, posteriormente Teófrates escribió el primer informe sobre estructura y clasificación de plantas, fruto de ello fueron unos libros perfectamente documentados que han servido como base en investigaciones biológicas durante muchos siglos, hasta que se sustituyeron en los siglos XVII y XVIII por investigadores europeos que ampliaron dicha información (Gallardo et al., 1994)

No parece extraño entonces ver como esta técnica sobre el Análisis de Clúster surgiera en el campo de la Antropología de la mano de los antropólogos Czekanowski (1911) en un estudio sobre unas tribus africanas, Clements, Schenck y Brown (1926) en un estudio sobre tribus en la Polinesia o Driver y Kroeber sobre unas tribus indias en California (1932), pero fue en 1935 cuando Stanislaw Klimek (1935) montó en una sola matriz 2000 coeficientes de correlación sobre 95 variables, de forma que como el número de variables y el número de correlaciones estaba aumentando, se hacía necesario el uso de métodos que fuesen capaces de reducir esas tablas tan grandes en un número más reducido (Driver y Schuessler, 1957).

Todas estas ideas fueron recogidas en el área de la Psicología por Stephenson (1936) que sugirió el uso invertido de análisis factorial para encontrar grupos de personas y Zubin (1938) quien propuso un método para la clasificación de una matriz de correlación que produciría una distribución en clúster.

Sin embargo, el primer trabajo en agrupación de clúster lo realizó R.C. Tryon (1939), debido por un lado a su interés en las diferencias de los individuos y a la influencia de Thurstone por la importante labor que estaba llevando a cabo en el desarrollo de la metodología en el análisis factorial, y aunque a Tryon no le gustaba el análisis factorial porque decía que se trataba de un complicado análisis matemático, decidió proponer un método más simple y más directo para el agrupamiento de variables en clúster, cabe señalar que la mayoría de los métodos desarrollados por

Tryon son variantes de lo que actualmente se conoce como Análisis factorial múltiple (Blashfield y Aldenderfer, 1988).

En 1944 Cattell, discutió cuatro métodos de clúster, el método “Ramyfing Linkage” que es una variación del método actual “Single Linkage”, el método “Matrix Diagonal Method” que sería actualmente un procedimiento gráfico (Hartigan, 1975), el método de Tryon, el cual está relacionado en la actualidad con el método “Average Linkage” y por último el método de Delimitación Aproximada, que era una extensión del método primero de ramificación, pero la obra más completa sobre los puntos de vista de Catell sobre análisis de clúster está recogida en su obra “El Reconocimiento Taxonometrico de tipos emergentes funcionales” (Catell et al., 1966)

A pesar de los primeros trabajos que se estaban llevando a cabo sobre el análisis de clúster dentro del campo de la psicología, el tema no suscitó mucha atención en la literatura científica hasta la década de 1960 cuando se produjo una dramática explosión de interés, disciplinas como la informática, ciencias de la información, la estadística deben mucho a la teoría y práctica del análisis clúster, Roger Needham y Karen Sparck Jones fueron dos de las figuras más influyentes en la informática a nivel mundial, fruto de esto lo avalan diferentes publicaciones sobre el análisis de clúster entre 1964 y 1968.

Pero dos fueron las razones principales para el crecimiento repentino que se estaba llevando a cabo en el análisis de clúster, por un lado la disponibilidad de ordenadores de alta velocidad, en primer lugar porque sin ellos era imposible pensar en realizar dichos análisis y en segundo porque la velocidad de proceso de los ordenadores era indispensable y fundamental para poder realizar con éxito en un tiempo razonable los análisis de clúster, ya que con anterioridad al auge de los ordenadores, los métodos de análisis de clúster resultaban tediosos y difícil desde el punto de vista del cálculo, ya que si tomamos como ejemplo el clasificar un conjunto de datos con 200 individuos, necesitaríamos trabajar con una matriz de (200x200) lo que nos daría unos 19.900 valores únicos, con lo que el número de investigadores y el tiempo necesario para hacer todo el proceso, hacía inviable este tipo de análisis.

La segunda razón que ocasionó el incremento en investigaciones sobre Análisis de Clúster fue la publicación de los "Principios de la Taxonomía Numérica" por Sokal y Sneath (1963). Sokal y Sneath propusieron reunir tantos datos como fuera posible sobre los organismos en los que un biólogo podría estar interesado en

clasificar, estimar el grado de similaridad existente entre dichos organismos y posteriormente realizar un análisis de clúster para de forma empírica formar categorías de organismos similares, formando grupos de organismos similares que representaban los taxones naturales o grupos de organismos emparentados de una clasificación biológica, y una vez que han sido agrupados analizar los miembros de cada grupo y determinar si representan especies biológicas diferentes.

El análisis de clúster o de conglomerados es una técnica capaz de dividir un conjunto de datos en grupos que sean significativos y que puedan ser útiles a posteriori, capturando por tanto la estructura natural de los mismos. Este tipo de análisis en muchas ocasiones es solo un punto de inicio para conseguir el objetivo último que se pretende con los datos que se están analizando, el análisis de clusters ha desempeñado un papel importante en multitud de áreas de la ciencia, como pueden ser la psicología, la minería de datos, la biología, las ciencias sociales e incluso el aprendizaje automático.

Scoltock en 1982, argumentó que la mayor parte de la literatura sobre análisis de clúster se producía en las áreas de Biología, Medicina (Psiquiatría), Ciencias Sociales, Geografía/Geología y Reconocimiento de patrones en Ingeniería.

Hoy en día, podemos afirmar que Scoltock no se equivocaba en 1982 al definir esas cinco áreas de la ciencia, aunque el avance de la tecnología y de la investigación en muchas áreas ha demostrado que esa división se ha quedado corta, ya que son numerosas las disciplinas o áreas donde el análisis de clúster tiene una función primordial e importante y juega un papel decisivo para la continua evolución en todos los campos, como por ejemplo:

- La Astronomía, analizando los clúster que forman las galaxias o súper galaxias y en la clasificación y evolución de las estrellas ha sido notorio la influencia, por ejemplo al clasificar las estrellas en enanas y gigantes.

- El clima, la atmósfera y los océanos para analizar los patrones atmosféricos que puedan predecir cambios en el clima y en la tierra, como terremotos y huracanes

- Las Ciencias Ambientales para clasificar ríos y poder establecer tipologías en función de la calidad de sus aguas

- La Empresa para obtener información sobre clientes actuales y futuros, y poder segmentar mercados en función del marketing mix de cada empresa

- La Sociología, Psicología y Medicina en todas sus ramas, y no solamente dentro de la Psiquiatría como decía Scoltock, detectando posibles patrones en la distribución temporal de una enfermedad.

- La Biología con su taxonomía y Microarrays en el campo de la genética, para proporcionar clasificaciones objetivas y estables de forma que la inclusión de más elementos u organismos no alteren la clasificación, ya Mendelejev en 1860 produjo un gran impacto en la comprensión de la estructura del átomo al poder clasificar los elementos en la tabla periódica.

- El área de Internet o la World Wide Web en donde la agrupación en clusters se hace indispensable para poder satisfacer los millones de peticiones y requerimientos que realizan millones de usuarios.

- El Big Data o almacenamiento masivo, estos grandes volúmenes de conjuntos de datos que son producidos por millones de personas y se almacenan en multitud de centros de datos, en donde uno de los mayores problemas asociados que tienen actualmente son su almacenamiento, gestión, recuperación y análisis, y en donde el Análisis de Clúster juega un papel clave para el análisis exploratorio de esos datos, permitiendo analizar grandes volúmenes de datos y por lo tanto ayudando en la toma de decisiones.

Es por ello que el análisis de clúster tiene una gran importancia en cualquier rama o área de investigación científica, ya que la clasificación y agrupación de individuos sean del tipo que sean y en función de cada rama de la ciencia en donde nos encontremos, es uno de los objetivos principales que se buscan a la hora de afrontar una investigación, Jain y Dubes (1980, 1988) definían ya el Análisis de Clúster como una herramienta de exploración de datos que se podía complementar con técnicas de visualización.

El análisis de clúster es por tanto un conjunto de técnicas multivariantes que nos permiten clasificar un conjunto de datos o individuos en grupos similares u homogéneos partiendo de grupos desconocidos, y de esta forma describir

asociaciones o estructuras naturales dentro de un conjunto de datos, que a priori no parecen evidentes pero que una vez encontradas pueden ser muy útiles para poder tener un conocimiento más profundo acerca de los elementos o individuos que existen en cada grupo, y de esta forma tener información relevante para la toma de decisiones, estableciendo planes de acción que permitan plantear y alcanzar los objetivos fijados.

Este Trabajo de Fin de Máster no pretende analizar en profundidad las características de todos y cada uno de los métodos existentes en el Análisis de Clúster sobre un conjunto de datos, pero si vamos a profundizar en el Método "Two-Step" o "Bietápico" o también llamado "en dos fases", propuesto por Chiu, Fang, Chen, Wang and Jeris (2001), para a continuación compararlo con los métodos HJ-Biplot de Galindo (1986) y con el ClusPlot de Pison et al. (1999).

Para el desarrollo de todo lo mencionado, se ha estructurado este trabajo de Fin de Master en distintos capítulos, definiendo las características fundamentales de cada tema tratado, con una introducción y los objetivos que se quieren alcanzar, un capítulo referente al material y métodos empleados, en donde se expondrán las características de cada uno de los métodos analizados, los resultados de las comparaciones entre los distintos métodos y terminaremos con las conclusiones más relevantes.

Capítulo I. Análisis de Clúster: donde se describe los Análisis de Clúster con un planteamiento inicial, las etapas y procesos que nos encontramos en dichos análisis y la clasificación de los diferentes métodos.

Capítulo II. Análisis del Método "Two-Step" propuesto por Chiu, Fang, Chen, Wang and Jeris (2001), donde se define el artificio matemático empleado para la reducción de la dimensionalidad, las propiedades algebraicas y un pequeño ejemplo de aplicación.

Capítulo III. Comparación del método de Clúster propuesto por Vicente-Tavera sobre "HJ-Biplot" (Galindo, 1985), con el Método "Two-Step".

Capítulo IV. Comparación del método "ClusPlot" de Pison et al. (1999), con el Método "Two-Step" y el Método Clúster HJ-Biplot.

.

Objetivos

El ***objetivo general*** de este trabajo es comparar diferentes métodos de Análisis de Clúster, en concreto el Método Two-Step propuesto por Chiu, Fang, Chen, Wang and Jeris (2001), el método de Inercia sobre HJ-Biplot Garcia-Talegón, et al. (1999) y el ClusPlot de Pison et al. (1999).

Comenzará con una pequeña revisión bibliográfica e histórica desde sus inicios hasta los métodos propuestos, y que son una parte fundamental en los estudios y en la evolución de muchas de las ciencias de investigación actuales.

Los ***objetivos específicos*** son por tanto los siguientes, y lo exploraremos mediante ejemplos prácticos sobre las diferentes tecnologías:

1. Estudiar el Método Two-Step, analizando el álgebra que lo sustenta, su algoritmo y su funcionamiento.

2. Comparar el Método de Inercia sobre HJ-Biplot con el Método Two-Step.

3. Comparar el Método Clusplot con el Método Two-Step y con el Método de Inercia sobre HJ-Biplot.

Capítulo I

AGRUPAMIENTO EN CLUSTER

Análisis de Clúster

1.1. Introducción

El término "*Análisis de Clúster*" define una gran variedad de técnicas y métodos, todas ellas con un único fin: clasificar los individuos que intervienen en el estudio y agruparlos en un número finito de clúster y, todo ello, dependiendo del comportamiento que tengan los individuos sobre las variables.

Los métodos clúster han sido desarrollados desde mediados del siglo XX, pero la mayor parte de su literatura ha sido escrita durante las pasadas 3 décadas, constituyendo su mayor exponente de desarrollo los denominados "*Principios de Taxonomía Numérica*" de Sokal y Sneath (1963), asumiendo que el proceso de reconocimiento de patrones debe ser usado como base para comprender su proceso evolutivo.

El análisis de clúster es, por tanto, un procedimiento estadístico multivariante que parte de un conjunto de datos que contienen información sobre individuos y pretende reorganizarlos en grupos homogéneos, de forma que las entidades que se encuentren dentro del mismo grupo tengan una asociación o similitud lo más fuerte posible y, al mismo tiempo, las diferencias entre los distintos grupos sea la mayor posible y, de esta forma, el grado de asociación dentro del grupo sea mayor que el grado de asociación entre distintos grupos o clúster. Jain y Dubes (1988) en su obra "*Algorithms for Clustering Data*", definen el análisis de clúster como una herramienta de exploración de datos que se complementa con técnicas de visualización de los mismos.

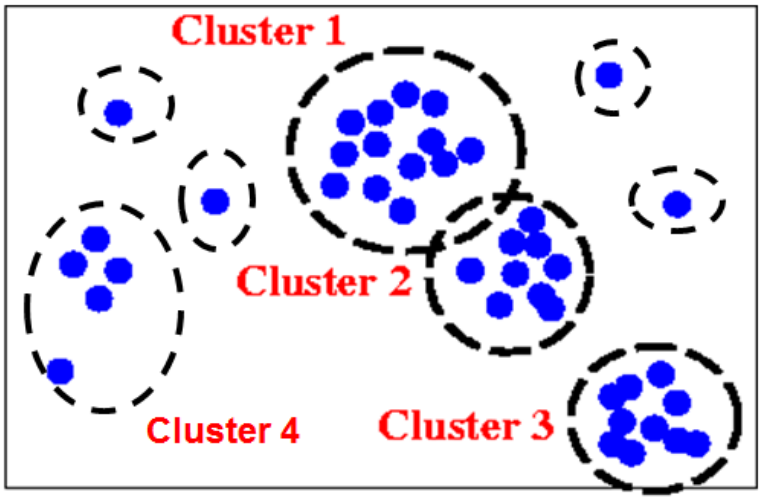


Fig. 1.1 Esquema gráfico del objetivo de un Análisis de Clúster

Por lo tanto, los procesos de **agrupación** y **clasificación** (que se llevan a cabo en un análisis de este tipo) se constituyen como las 2 tareas fundamentales dentro del mismo. De esta manera, por un lado la **clasificación** se utiliza sobre todo como un método supervisado de aprendizaje; y por otro lado la **agrupación** es un método de aprendizaje no supervisado, es decir el objetivo de la agrupación en clúster es descriptivo, mientras que el de la clasificación es predictivo (Veyssieres and Plant, 1998).

En tareas de clasificación, sin embargo, una parte importante de la evaluación es extrínseca (Rokach y Maimon, 2005), ya que los grupos deben reflejar un sistema de referencia de las clases, según Tyron y Bailey (1970) en su obra "Clúster Analysis".

> *"" La comprensión de nuestro mundo requiere la conceptualización de las similitudes y diferencias entre todas las entidades que lo componen""*

De esta forma, los individuos se organizan y se agrupan de manera eficiente para que su representación pueda caracterizar al grupo que lo conforman.

Según Gallardo (1994), las 2 principales razones para el gran desarrollo creciente que han tenido estas técnicas, son:

1) ***El desarrollo de los ordenadores***, ya que los métodos manuales de agrupamiento eran muy laboriosos y costosos, ya que trabajar con matrices de gran tamaño requería una gran cantidad de investigadores y mucho tiempo de desarrollo. En la actualidad, con la capacidad de procesamiento que tienen los ordenadores, todos esos inconvenientes se ven reducidos.

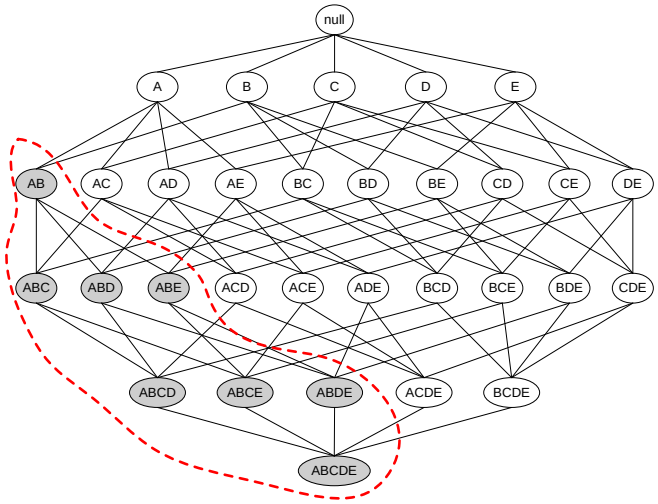


Fig. 1.2 Esquema gráfico de variaciones de Matriz de datos de gran tamaño

2) ***La importancia de la clasificación*** para el procedimiento científico (que contiene y conforma los principales conceptos usados en su ámbito), y es que todas las ciencias están construidas sobre clasificaciones que permiten estructurar su dominio de investigación.

Según Peña (2002), el análisis de clúster es útil en 3 situaciones distintas:

1º. **La Partición de los Individuos**. Cuando se desea dividir un conjunto de elementos heterogéneos de manera que:

- ❖ Cada individuo pertenezca a un sólo grupo.
- ❖ Todo individuo quede clasificado.
- ❖ Cada grupo creado sea internamente homogéneo.

2º. **La Construcción de Jerarquías**. Cuando se desea estructurar los individuos o elementos de un grupo según su similitud o jerarquía, ordenando ese conjunto en niveles, de manera que el grupo superior contiene a los grupos inferiores. Estrictamente esta práctica no define agrupaciones, sino la estructura de asociación en cadena que pueda existir.

3º. **La Clasificación de las Variables**: En investigaciones con muchas variables hay que hacer un estudio previo exploratorio para dividir el conjunto de variables iniciales en grupos o estructurarlos en una jerarquía, y de esta forma este análisis previo sirve al investigador para orientarse a la hora de plantear modelos formales y para reducir la dimensión del problema que va a estudiar.

Si consideramos una muestra formada por “n” individuos sobre los que medimos “p” variables y lo representamos en una matriz, de tal manera que situamos en sus filas los individuos y en sus columnas las variables o características de cada individuo, la i-ésima fila de la matriz “**M**” contendrá los valores de las características que diferencian al i-ésimo individuo, mientras que la j-ésima columna nos mostrará los valores correspondientes a la característica j-ésima para todos y cada uno de los individuos de la muestra.

$$\mathbf{M} = \begin{pmatrix} x_{11} & x_{12} & \dots\dots\dots & x_{1j} & \dots\dots & x_{1p} \\ x_{21} & x_{22} & \dots\dots\dots & x_{2j} & \dots\dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{i1} & x_{i2} & \dots\dots\dots & x_{ij} & \dots\dots & x_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \dots\dots\dots & x_{nj} & \dots\dots & x_{np} \end{pmatrix}$$

Para calcular el número “k” de grupos en los que se pueden clasificar las “n” observaciones o individuos, nos apoyamos en el número de Stirling de segunda especie (Abramowitz y Stegun, 1972) y (Torres, 2008).

$$S_n^{(k)} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^n$$

Como a priori no se conoce el número de grupos o clúster en los que tenemos que realizar la división y agrupamiento de los “n” individuos, el problema se complica ya que el número de posibilidades existentes sería igual a la suma de los números de Stirling, que para nuestro caso:

$$\sum_{j=1}^{n} S_n^{(j)}$$

Por ejemplo, si tomásemos 25 individuos distintos, tendríamos como resultado las siguientes posibilidades:

$$\sum_{j=1}^{25} S_{25}^{(j)} > 4 * 10^{18}$$

Tal y como podemos observar resulta un número excesivamente grande, por lo que los posibles grupos en los cuales se dividirían serian también muy numeroso, así que es necesario encontrar una solución adecuada, teniendo en cuenta únicamente un número pequeño de alternativas, y es por ello que resulta necesario utilizar las Técnicas de agrupamiento en Clúster.

1.2. Etapas del Análisis de Clúster

Aldenderfer y Blashfield (1988) establecen 4 etapas básicas para la realización de un análisis de clúster:

1. **Selección de las Variables** que se van a incluir en el análisis, puesto que la elección inicial de las características que describen a los individuos constituye un marco de referencia para poder establecer posteriormente la relación entre individuos y de esa manera establecer las agrupaciones o clústers.

2. **Selección de la Medida de Distancia o Similaridad** entre los individuos, pues la mayoría de los métodos de agrupamiento en clúster necesitan establecer una medida de asociación entre estos elementos que permita medir la proximidad de éstos, bien sea con una medida de distancia, o bien con una medida de disimilaridad o similaridad.

   Cuando se lleva a cabo un *Análisis de Clúster para individuos*, la proximidad entre ellos viene expresada como *medidas de distancias;* mientras que si se trata de *Análisis de Clúster para variables*, esta proximidad viene expresada como *coeficientes de correlación*.

3. **Selección de la Técnica Clúster** para crear grupos similares. Es conveniente en este punto no realizar la clasificación con un único algoritmo o procedimiento, sino utilizar varios de ellos para, de esa forma, poder contrastar sus resultados, lo cual nos da la posibilidad de obtener unas conclusiones mucho más válidas y consistentes sobre la estructura natural de los datos.

   Existen una gran cantidad de técnicas diferentes para realizar análisis de clasificación de clusters, la razón principal para tener diferentes métodos de clúster es la no definición exacta del significado de la palabra “clúster”, Estivill-Castro y Yang (2000). Por esta razón, se han venido desarrollando diferentes métodos y cada uno de ellos usa diferentes principios.

4. **Validación de los resultados obtenidos**, última etapa en esta secuencia lógica (aunque la más importante), pues es en ella donde se tendrá que realizar la interpretación y la validación de todas las anteriores y, por tanto, donde se van a obtener las conclusiones definitivas del estudio.

En los puntos siguientes se estudiaran en detalle cada una de las etapas enumeradas, así como las diferentes posibilidades y métodos existentes.

1.3. Selección de las Variables

En primer lugar, hay que señalar que al seleccionar las variables relevantes es muy importante la opinión del investigador acerca del propósito/objetivo que se pretende sobre la clasificación.

La elección de las variables al comienzo del proceso constituye ya de por si una categorización de los datos, para lo cual seguiremos las directrices matemáticas y estadísticas que sean necesarias. Si el número de variables es muy grande, se podría realizar antes un Análisis de Componentes Principales para, de esta forma, poder resumir el conjunto de variables óptimas y reducir así la dimensionalidad del problema.

3 son los aspectos fundamentales que se deberían considerar en la selección de variables (Martínez Ramos, 1984):

1. Que las variables no estén correlacionadas.
2. Que la unidad de medida sea la misma para todas las variables en el estudio.
3. Que el número de variables no sea demasiado grande.

1.4. Selección de la Medida de Distancia o Similaridad

Seleccionar la forma de medida que se va a utilizar entre los individuos, nos permitirá medir la proximidad o distancia entre ellos, de forma que se pueda cuantificar el grado de similaridad que existe entre cada par de elementos, construyendo una matriz de similaridades que haga posible relacionar su semejanza.

Cuanto mayor sea su valor, mayor será el grado de similaridad entre cada par de individuos y, por tanto, mayor será la probabilidad de que el método los declare como elementos de un mismo grupo; y de igual forma si es menor el valor también será menor su grado de similaridad y, por tanto, menor su probabilidad de pertenecer al mismo grupo. Análogo razonamiento para las medidas de disimilaridad.

Cuando se realiza un análisis de clúster de individuos, la proximidad viene expresada en términos de distancias, mientras que en los análisis de clúster de variables, esta proximidad suele ser del coeficiente de correlación entre ellas, pero las medidas para calcular la relación entre pares de individuos, suelen ser medidas de distancia o medidas de similitud.

El tipo de distancias o disimilaridades que se pueden emplear en un análisis de clúster es muy variado y depende fundamentalmente del tipo de variables existentes en la muestra estadística a analizar.

Los criterios que se deben seguir para la elección de la distancia más apropiada para cada estudio están definidos en los siguientes autores: Cuadras (1989), Legendre L. y Legendre P. (1979), Gower y Legendre (1986), Legendre, Dallot y Legendre (1985).

Así mismo, los individuos de cada población están caracterizados por un vector aleatorio que sigue una distribución de probabilidad y, análogamente, la distancia entre 2 poblaciones será una medida de divergencia entre los parámetros que la caracterizan.

Si disponemos de una matriz de datos $M_{n\,x\,p}$, donde “p” es el número de variables estadísticas (cuantitativas, cualitativas, binarias o categóricas) y “n” es el número de individuos de la muestra de la población, llamaremos $\delta(x_i, x_j)$ la distancia entre 2 pares de individuos "x_i" e "x_j", será una medida de distancia simétrica no

negativa que cuantificará la diferencia entre ambos individuos en relación con las variables y, por tanto, esta distancia se puede sumatorizar utilizando la matriz de distancias:

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \dots\dots\delta_{1p} \\ \delta_{21} & \delta_{22} & \dots\dots\delta_{1p} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \delta_{n1} & \delta_{n2} & \dots\dots\delta_{np} \end{pmatrix}$$

Según Cuadras (1989), algunas de las propiedades generales que se deben de cumplir, son:

1. $\delta\,(x_i, x_j) \geq 0$
2. $\delta\,(x_i, x_i) = 0$
3. $\delta\,(x_i, x_j) = \delta\,(x_j, x_i)$
4. $\delta\,(x_i, x_j) = 0 \quad si\ y\ solo\ si \quad i = j$
5. $(x_i, x_j) \leq \delta(x_i, x_k) + \delta(x_j, x_k)$
6. $\delta\,(x_i, x_j) \leq máx\{\delta(x_i, x_k)\ ,\ \delta(x_j, x_k)\}\ \ desigualdad\ ultramétrica$
7. $\delta\,(x_i, x_j) + \delta(x_k, x_l) \leq máx\{\delta\,(x_i, x_k) + \delta(x_j, x_l),\ \delta\,(x_i, x_l) + \delta(x_j, x_k)\}$
8. $\delta\,(x_i, x_j)\ \ es\ \ euclídea$
9. $\delta\,(x_i, x_j)\ \ es\ \ riemanniana$
10. $\delta\,(x_i, x_j)\ \ es\ \ una\ divergencia$

Toda distancia debe de cumplir como mínimo las 3 primeras propiedades, pero si cumple únicamente esas 3 primeras propiedades, entonces se dice que es una ***disimilaridad.***

1.4.1. Medidas de distancia de tipo Cuantitativo

Existe una lista muy extensa de coeficientes, que puede ser consultada en Gower (1985). A continuación, mostramos de todas ellas las más utilizadas:

- ***Distancia Euclídea***

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

- ***Distancia Manhattan o de ciudad***

$$\delta_{(ij)} = \sum_{k=1}^{p} \left|x_{ik} - x_{jk}\right|$$

- ***Distancia de Minkowski,*** **(Han y Kamer - 2001)**

$$\delta_{(ij)} = \sqrt[g]{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^g} \qquad 1 < g < \infty$$

- ***Distancia de Mahalanobis,*** **(Mahalanobis - 1936)**

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

- ***Distancia de Bray-Curtis,*** **(Bray y Curtis - 1957)**

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

- ***Métrica de Camberra,*** **(Lance y Williams - 1966)**

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

- ***Distancia de Pearson,*** **(Pearson - 1926)**

$$\delta_{(ij)} = \left(\sum_{k=1}^{p} \left(\frac{x_{ik} + x_{jk}}{s_k}\right)^2\right)^{1/2}$$

- ***Coeficiente de divergencia de Clark,*** **(Clark - 1952)**

$$\delta_{(ij)} = \sqrt{\frac{1}{p} \sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}}\right)^2}$$

1.4.2. Medidas de Similaridad de tipo Cualitativo

Cuando los atributos son **Nominales**, existen dos enfoques de aproximación Rokach y Maimon (2005):

1. *Por correspondencia simple*, mediante $\delta_{(ij)} = \frac{p-m}{p}$, donde p es el número total de atributos y m es el número de coincidencias
2. *Creando un atributo binario por cada estado* que contenga cada atributo nominal y calcular su disimilaridad

Cuando los atributos son **Ordinales**, la secuencia de los valores es significativa, y en tales casos, los atributos pueden ser tratados como numéricos, después de asignarles valores entre '0' y '1', donde el '0' señala que dicha característica está ausente y '1' que está presente. De esta manera, la información sobre el grado de asociación entre cualquier par de individuos puede representarse como una tabla de contingencia 2x2.

Dicha asignación se lleva a cabo de la siguiente forma:

$$z_{(i,n)} = \frac{r_{i,n} - 1}{M_n - 1}$$

Siendo:

$z_{(i,n)}$ → Valor normalizado del atributo a_n del objeto i.

$r_{i,n}$ → Valor antes de la estandarización.

M_n → Límite superior del dominio del atributo.

La mayoría de las distancias para este tipo de variables se forman partiendo del coeficiente de similaridad, que se pueden desarrollar a partir de la tabla de contingencia.

Por tanto, para pasar de una distancia a una disimilaridad (o similaridad), bastaría con transformar dicha distancia de la forma:

$$\delta_{(ij)} = 1 - S_{(ij)}$$

Aunque sugieren utilizar, las siguientes expresiones, ya que al aplicarlas sobre matrices de similitud dan lugar a una distancia métrica.

$$\delta_{(ij)} = \sqrt{S_{(ii)} - 2S_{(ij)} + S_{(jj)}} \qquad \text{Gower (1966)}$$

$$\delta_{(ij)} = \sqrt{1 - S_{(ij)}} \qquad \text{Cuadras (1986)}$$

Sobre los criterios a seguir para elegir el coeficiente de similaridad más adecuado, depende obviamente del tipo de datos que se tenga y del peso que se quiera dar a las frecuencias incluidas en la tabla de contingencia, Legendre L. y Legendre P. (1979), Gower y Legendre (1986).

Cuando los atributos son **Binarios,** la distancia se calcula basándose en la siguiente tabla de contingencia según el coeficiente elegido, cuya lista se puede consultar en Gower (1985), aunque a continuación enumeramos los más utilizados.

		Individuo "j"		
		Presencia **(1)**	Ausencia **(0)**	
Individuo "i"	Presencia **(1)**	a	b	a + b
	Ausencia **(0)**	c	d	c + d
		a + c	b + d	

Fig. 1.3 Tabla de representación binaria de presencia y ausencia

Para variables de tipo **Intervalo**, la solución consiste en tipificar antes del análisis, calculando las desviaciones típicas a partir de todo el conjunto de datos a analizar. Por ejemplo Fleiss y Zubin (1969), consideran que esta técnica puede tener serias desventajas frente otras, ya que diluyen las diferencias entre grupos sobre las variables que más discriminan y, por tanto, sugieren usar la desviación estándar entre grupos para tipificar. Por otro lado, cuando las variables son de tipos diferentes se suelen convertir todas ellas a formato binario antes de calcular esas similaridades; Gower (1971) propuso usar un coeficiente de similaridad que pudiera incorporar información de diferentes tipos de variables de una forma más sensible.

Existe un amplio número de coeficientes de similaridad que se pueden consultar en Sneath y Sokal (1973), Hubálek (1982) y Gower (1985). A continuación, se muestran los coeficientes de similaridad más usados:

- **Kulczynski (1927) ……** $\frac{a}{b+c}$
- **Russel y Rao (1940)……** $\frac{a}{a+b+c+d}$
- **Jaccard (1908)……** $\frac{a}{a+b+c}$
- **Sorensen (1948)……** $\frac{a}{a+\frac{1}{2}(b+c)+d}$
- **Sokal y Michener (1958)……** $\frac{a+d}{a+b+c+d}$
- **Sokal y Sneath (1963)……** $\frac{a}{a+2(b+c)}$
- **Dice (1945)……** $\frac{2a}{2a+b+c}$
- **Ochiai (1957)……** $\frac{a}{\sqrt{(a+b)(a+c)}}$
- **Pearson (1926)……** $\frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$
- **Yule (1912)……** $\frac{ad-bc}{ad+bc}$
- **Rogers y Tanimoto (1960)……** $\frac{a+d}{a+2b+2c+d}$
- **Hamman (1961)……** $\frac{(a+d)-(b+c)}{a+b+c+d}$

1.4.3. Medidas de Similaridad de tipo Mixto

El mayor problema que podemos encontrarnos en un análisis de clúster, es cuando tenemos variables de tipo mixto, es decir variables cuantitativas y variables cualitativas (independientemente de que estas últimas se encuentren tipificadas en formato binario, ordinal o nominal o se encuentren estas en formato alfanumérico). Algunos autores sugieren transformar todas las variables a formato binario (Gordon, 1990) y, de esta forma, poder tratar a todas por igual. Otros autores, en cambio, sugieren hacer una combinación de similitudes empleando una medida de ponderación común (Wojciechowski, 1987) y (Miranda et al., 1998).

El coeficiente de similaridad de Gower (1971) está adaptado a trabajar de forma simultánea con variables de tipo cuantitativo y cualitativo, ya sean estas de tipo nominal, ordinal o binario, y por tanto con la aplicación de este coeficiente se puede determinar el grado de similaridad que tienen los individuos de la matriz de datos, estando definido por:

$$d_{ij}^2 = 1 - S_{ij}$$

$$S_{ij} \frac{\sum_{h=1}^{p1}\left[1 - \frac{|x_{ih} - x_{jh}|}{G_h}\right] + a + \propto}{p1 + (p2 - d) + p3}$$

Donde los parámetros se definen por:

p1 = Número de variables cuantitativas.

p2 = Número de variables binarias.

p3 = Numero de variables cualitativas.

d = Número de coincidencias en ausencia “0” (Binarias).

a = Número de coincidencias en presencia “1” (Binarias).

$\propto$ = Número de coincidencias para p3 (Cualitativas).

G_h = Rango de la h-ésima variable cuantitativa.

Rango = $X_{máx.} - X_{min.}$

Cuando todos los caracteres son binarios, el índice de similitud de Gower coincide con el índice de Jaccard. Cuando los caracteres son cualitativos y tienen más de 2 posibilidades o estados, entonces el índice de Gower es equivalente al coeficiente de coincidencias simple, que viene definido como la relación del número total de coincidencias y el número total de caracteres. Por último, cuando todos los caracteres son cuantitativos, el índice de Gower se asemeja a la medida absoluta de las distancias (Chávez, 2010).

1.4.4. Correlación entre individuos

El coeficiente de correlación entre vectores de individuos puede usarse como una medida de asociación entre ellos

$$r_{ij} = \frac{\sum_{l=1}^{n}(x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{s_i\, s_j}$$

En donde tenemos las medias de cada individuo como $\bar{x}_i$ y $\bar{x}_j$, y las desviaciones típicas de cada uno de ellos como s_i y s_j.

Pero el mayor problema de este coeficiente, es que un vector de datos que corresponda a un individuo tiene muchas unidades diferentes, por lo que hace muy difícil poder comparar las medias y las varianzas de los individuos. No obstante Cronbach y Gleser (1953), demostraron que este coeficiente tenía un carácter métrico.

1.4.5. Distancias derivada de la Distancia Chi-cuadrado (χ^2)

Dentro de las medidas de asociación entre individuos, la más usual y familiar en el análisis de contingencia es la Chi-cuadrado.

Sabiendo que:

o_{ij} = Valor observado en la celda i,j.

e_{ij} = Valor esperado bajo la hipótesis de independencia

$$\boldsymbol{\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

1.5. Selección de la Técnica Clúster de Clasificación

El paso siguiente, es elegir la técnica con la que realizar la agrupación, es decir el método clúster que se va a usar en función de la naturaleza de los datos que se disponen.

Los métodos clúster que se han propuesto y desarrollado en los últimos años son muy numerosos, clasificándose fundamentalmente en jerárquicos y no jerárquicos, donde su diferencia principal es la forma de asignar los individuos a cada clúster, de manera que en los métodos jerárquicos las asignaciones que se van creando permanecen estables durante todo el proceso y no se permiten reasignaciones posteriores a clusters distintos (si es que eso fuera necesario). Por el contrario, en los métodos no jerárquicos dicha reasignación es posible a posteriori.

Además, un factor a tener en cuenta para el investigador en las conclusiones, es que, por ejemplo, en los métodos no jerárquicos el número final de clusters está predefinido de antemano y el enfoque cambia totalmente que cuando el número de clusters no se conoce (esa decisión 'subjetiva', de alguna forma, impone y sesga el resultado final y por consiguiente la opinión del propio investigador).

La elección del método a emplear dependerá de la naturaleza de los datos que se vayan a usar y sobre todo del objetivo final que se pretenda; por lo tanto (tal y como señalábamos anteriormente) es conveniente no elegir un solo procedimiento y analizar sus resultados.

1.6. Validación de los Resultados

La validación e interpretación de los resultados obtenidos sería la última etapa lógica en la que se desarrolla una investigación sobre unos datos al aplicar un método de clúster; siendo es la etapa más importante, ya que es en cuando se obtienen las conclusiones definitivas del estudio.

Cuando se aplica un método jerárquico, se plantean dos problemas, Gallardo (1994):

a) ¿En qué medida representa la estructura final obtenida las similitudes o diferencias entre los objetos de estudio?

El argumento más utilizado para responder a esta pregunta es el empleo del *"Coeficiente de Correlación Cofenético"*, propuesto por Sokal y Rohlf (1962) que mide

la correlación existente entre las distancias iniciales tomadas a partir de los datos originales, y las distancias finales con las cuales los individuos se han unido durante el desarrollo del método.

Si tras el empleo de varios procedimientos clúster distintos, éstos conducen a soluciones parecidas, surge la pregunta de qué método elegiremos como definitivo, y la respuesta precisamente la aporta este coeficiente, ya que aquel método que presente un valor más alto será aquel que presente una menor distorsión en las relaciones originales existentes entre los elementos en estudio (Gallardo, 1994).

El coeficiente cofenético es la correlación entre los $n(n-1)/2$ elementos de la parte superior de la matriz de proximidades observada y los correspondientes a la llamada matriz cofenética, cuyos elementos se definen como aquellos que determinan la proximidad entre los elementos "i" y "j" cuando éstos se unen en un mismo clúster.

A continuación, representamos de forma esquemática, el proceso de comprobación del coeficiente cofenético. Una vez obtenida la matriz de distancias o disimilaridades, tendríamos el siguiente dendograma de agrupación de clúster de nuestra matriz de datos original:

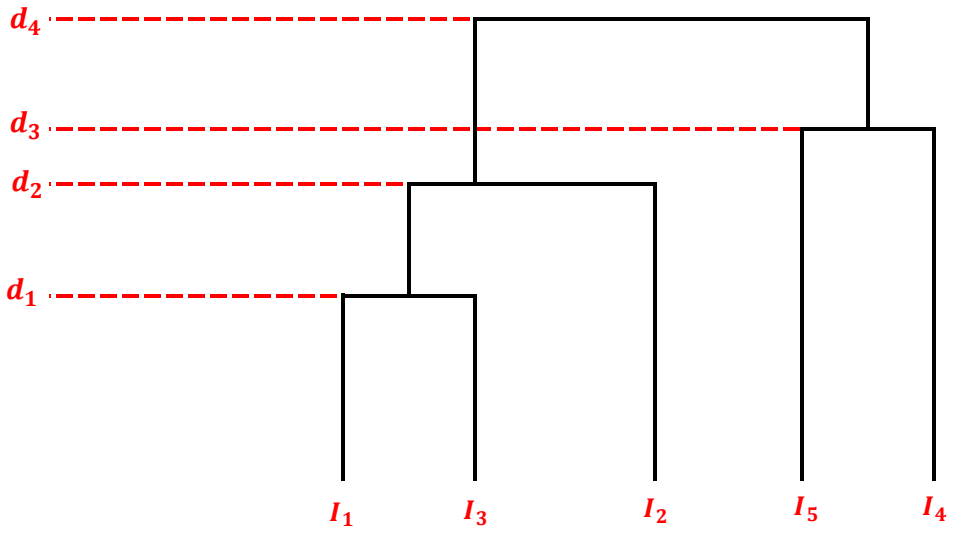


Fig. 1.5 Dendograma de la agrupación en clúster sobre 5 individuos

Dispondríamos, por tanto, las siguientes matrices: una de ellas con las distancias entre individuos, y la otra seria la matriz cofenética:

Matriz Inicial

	I_1	I_2	I_3	I_4	I_5
I_1	0	d_{12}	d_{13}	d_{14}	d_{15}
I_2		0	d_{23}	d_{24}	d_{25}
I_3			0	d_{34}	d_{35}
I_4				0	d_{45}
I_5					0

Matriz Cofenética

	I_1	I_2	I_3	I_4	I_5
I_1	0	d_{2}	d_{1}	d_{4}	d_{4}
I_2		0	d_{2}	d_{4}	d_{4}
I_3			0	d_{4}	d_{4}
I_4				0	d_{3}
I_5					0

Fig. 1.6 Ejemplo gráfico de las dos matrices en estudio

$$c = \frac{\sum (X_{(i,j)} - X^*)(Y_{(i,j)} - Y^*)}{\sqrt{(X_{(i,j)} - X^*)^2 (Y_{(i,j)} - Y^*)^2}}$$

X = valores de la matriz de distancias iniciales

X* = valor medio de los elementos de la matriz de distancias iniciales

Y = valores de la matriz cofenética

Y* = valor medio de los elementos de la matriz cofenética

b) Cuál es el número ideal de clusters que mejor representa la estructura natural de los datos?

Fisher (1968) encuentra que los valores medios del coeficiente correlación cofenético tienden a decrecer con 'n' y son casi independientes del número de variables, sugiriendo que los valores del coeficiente que sean superiores a 0,8 serán suficientes para poder rechazar la hipótesis nula, partiendo de la base que la hipótesis nula es "*no existencia de estructura natural de los datos*"

Rohlf (1970) asegura que aquellas correlaciones cofenéticas cercanas al 0,9 no garantizan que el dendograma sirva para poder definir las relaciones cofenéticas.

Jain y Dubes (1980) comentan lo siguiente:

> "*El rechazo de la hipótesis nula no es significativo, porque no han sido desarrolladas hipótesis alternativas significativas; no existe todavía una*

definición que pueda ser útil y a la vez práctica sobre una estructura de clúster, matemáticamente hablando"

Algunas variantes para la evaluación del número de clúster hallado, son:

1. **Raiz Cuadrática Media – RMSSTD (Root Mean Square Standard Deviation):**

   Consiste en una media de la desviación estándar de todas las variables y sólo se puede interpretar de forma relativa: buscando el menor RMSSTD, lo que es lo mismo un punto de equilibrio entre 'n' y 1 clúster.

   El RMSSTD de un clúster es la desviación estándar promediada de todas las variables para los individuos que forman el clúster, según Barrera (2014) cuanto más pequeño sea el valor, más homogéneas serán las observaciones con respecto a las variables y viceversa.

$$RMSSTD = \sqrt{\frac{(n-1)\sum_{j=1}^{p} S_j^2}{p(n-1)}}$$

2. **R cuadrado – RS (R Square):**

   Es el ratio entre SSE y SST, es decir $SST = SSE + SSD$. Por tanto, para un conjunto dado de datos, las mayores diferencias entre los grupos producen mayor homogeneidad dentro de los grupos y viceversa. Este valor oscila entre '0' y '1', donde '0' indica que no hay diferencias y '1' que existe el máximo de diferencia entre ellos.

3. **R cuadrado Semiparcial – SPR (Semi Partial R Square):**

   La diferencia entre la SSd acumulada del nuevo clúster y la suma de cuadrados acumulada de SSd de los clúster unidos para formar nuevos clúster se llama perdida de homogeneidad.

   Si la pérdida de homogeneidad es '0', significa que el nuevo clúster se ha obtenido mezclando 2 clusters homogéneos, pero si la pérdida de homogeneidad es grande, entonces significará que el nuevo clúster se ha obtenido combinando 2 clúster heterogéneos.

$$SPR = \frac{SSd(nuevo) - SSd(fusionado)}{SSd(total)}$$

4. **Distancia entre dos clúster – CD (Clúster Distance):**

   Esta distancia no debe de ser grande para que los dos clusters puedan mezclarse, un valor grande nos indicaría que se están fusionando dos clusters disimilares.

Respecto a los métodos no jerárquicos, las preguntas anteriormente expuestas pierden algo de sentido, ya que los procedimientos que se emplean para validar los resultados van dirigidos al estudio de la homogeneidad de los clúster encontrados durante el desarrollo del método elegido.

1.7. Métodos de Clasificación de tipo Jerárquico

Estos métodos tienen por objetivo agrupar clusters para formar un nuevo clúster o bien para separar alguno existente y de esta forma dar origen a otros dos, de manera que se minimice la distancia entre ellos o bien se maximice la medida de similitud existente, construyendo los clusters de forma recursiva, ya sea de arriba hacia abajo (Disociativo) o bien de abajo hacia arriba (Aglomerativo).

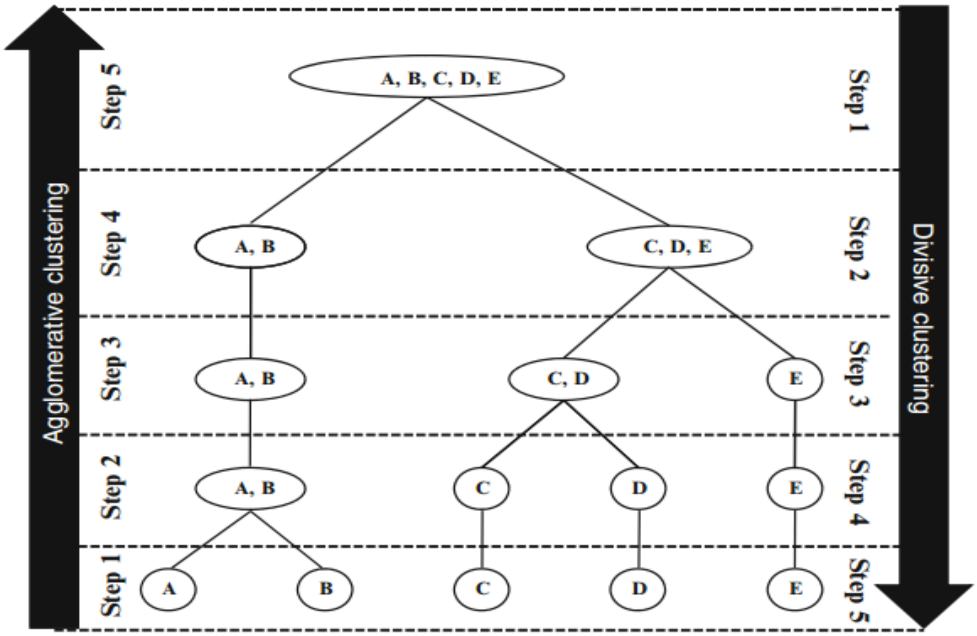


Fig. 1.7 Esquema de los distintos tipos de Método Jerárquico

El ***Método Jerárquico Aglomerativo***, comienza el análisis con tantos grupos como individuos exista en el estudio (es decir en el fichero o en la base de datos), a partir de ese punto se van formando grupos de forma ascendente, hasta que al final del proceso, todos los individuos estén englobados en un mismo clúster.

El ***Método Jerárquico Disociativo*** realiza el proceso inverso al método aglomerativo: empieza con un clúster que contiene todos los individuos y, a partir de ese punto, se van realizando sucesivas divisiones formando grupos más pequeños, para obtener al final del proceso tantos clúster como individuos existan en la muestra que se está estudiando.

Los métodos jerárquicos permiten construir un ***Dendograma*** o árbol de clasificación, donde una agrupación de datos se obtiene cortando el dendograma en el nivel de similitud deseado

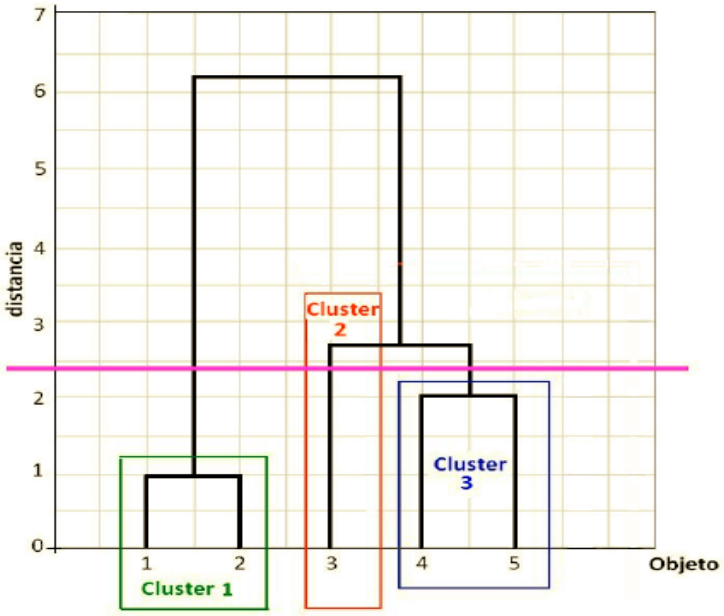


Fig. 1.8 Ejemplo gráfico de un árbol Jerárquico Aglomerativo, y la línea de corte para formar tres clusters

Independientemente del proceso de agrupamiento que se lleve a cabo, existen diversos criterios para formar los clústers y todos ellos se basan en una matriz de distancias o similitud; y, por lo tanto, se podrían dividir de acuerdo a la forma de calcular dicha medida de similitud, (Jain et al, 1999).

El análisis clúster de tipo jerárquico es una herramienta exploratoria diseñada para revelar los clústers dentro de un conjunto de datos que, de otra manera, no sería

evidente y la decisión a tomar, por tanto, del número óptimo de clusters es subjetiva, especialmente cuando se incrementa el número de objetos ya que si se seleccionan demasiado pocos, los clusters resultantes serán heterogéneos y artificiales, mientras que si se seleccionan demasiados, la interpretación de los mismos suele ser complicada.

1.7.1. Método Single-Linkage (vecino más próximo)

Este método considera que la distancia entre 2 clúster es la distancia más corta desde un miembro de un clúster a otro miembro de otro clúster, si los datos consisten en similaridades entre 2 clúster, entonces se considera la mayor similaridad desde cualquier miembro de un clúster a otro miembro de otro clúster (Sneath y Sokal, 1973).

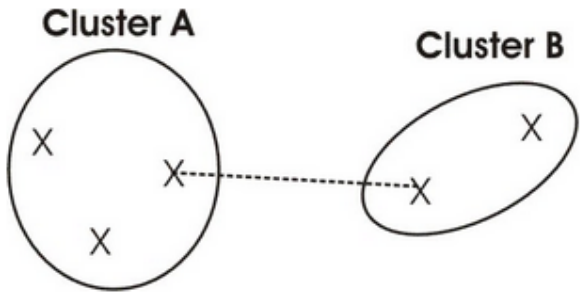


Fig. 1.9 Gráfico ejemplo del Método de Mínima Distancia

Por lo tanto, si después de haber realizado la etapa k-ésima, tenemos formados (n-k) clusters, la distancia entre el clúster C_i que tiene n_i elementos, con el clúster C_j que tiene n_j elementos, y sabiendo que x_i pertenece al conjunto C_i y que x_j pertenece al conjunto C_j sería la siguiente:

$$d(C_i, C_j) = Min\{d(x_l, x_m)\} \quad (l = 1,2, \ldots., n_i)\ y\ (m = 1,2, \ldots., n_j)$$

Y si tuviésemos similaridades, como consecuencia de haber empleado alguna medida de ese tipo en base al tipo de variables de la matriz de datos, entonces tendríamos:

$$s(C_i, C_j) = Max\{s(x_l, x_m)\} \quad (l = 1,2, \ldots., n_i)\ y\ (m = 1,2, \ldots., n_j)$$

Por lo tanto, el proceso para la unión de los dos clusters en el proceso, seria o bien minimizar las distancias o bien maximizar las similaridades, tal y como se demuestra a continuación:

a) En el caso de emplear **distancias**:

$$d(C_i, C_j) = Min_{(i_1,j_1)=1,2,\ldots.(n-k)|i_1\neq j_1}\{d(C_{i1}, C_{j1})\} =$$

$$Min_{(i_1,j_1)=1,2,\ldots.(n-k)|i_1\neq j_1}\left\{Min_{x_l\in C_{i1}\,y\,x_m\in C_{j1}\,y} d(x_l, x_m)\right\}$$

$$(l = 1,2,\ldots.,n_{i1})\ y\ \ (m = 1,2,\ldots.,n_{j1})$$

b) En el caso de emplear **similaridades**:

$$s(C_i, C_j) = Max_{(i_1,j_1)=1,2,\ldots.(n-k)|i_1\neq j_1}\{s(C_{i1}, C_{j1})\} =$$

$$Max_{(i_1,j_1)=1,2,\ldots.(n-k)|i_1\neq j_1}\left\{Max_{x_l\in C_{i1}\,y\,x_m\in C_{j1}\,y} s(x_l, x_m)\right\}$$

$$(l = 1,2,\ldots.,n_{i1})\ y\ \ (m = 1,2,\ldots.,n_{j1})$$

1.7.2. Método Complete-Linkage (vecino más lejano)

Este método considera que la distancia más grande desde cualquier miembro de un clúster a otro miembro de otro clúster, (King, 1967), si la medida es la distancia, se toma la distancia máxima de los individuos del grupo al nuevo individuo. Si se tomara la similitud o similaridad entre el grupo formado y el nuevo individuo, entonces se recoge la mínima de los individuos del grupo al nuevo individuo.

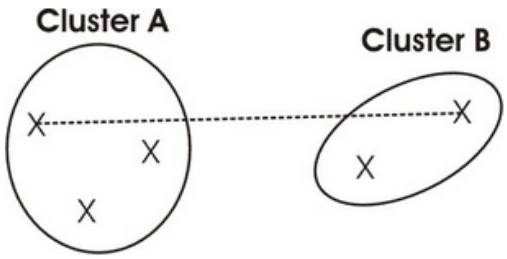


Fig. 1.10 Gráfico ejemplo del Método de Máxima Distancia

Por lo tanto, si después de haber realizado la etapa k-ésima, tenemos formados (n-k) clusters, la distancia entre el clúster C_i que tiene n_i elementos, con el clúster C_j que tiene n_j elementos, y sabiendo que x_i pertenece al conjunto C_i y que x_j pertenece al conjunto C_j sería la siguiente:

$$d(C_i, C_j) = Max\{d(x_l, x_m)\} \quad (l = 1,2,\ldots.,n_i)\ y\ \ (m = 1,2,\ldots.,n_j)$$

Y si tuviésemos similaridades, como consecuencia de haber empleado alguna medida de ese tipo en base al tipo de variables de la matriz de datos, entonces tendríamos:

$$s(C_i, C_j) = Min\{s(x_l, x_m)\} \quad (l = 1,2, \ldots., n_i)\ y\ (m = 1,2, \ldots., n_j)$$

Por lo tanto, el proceso para la unión de los 2 clusters en el proceso seria, o bien minimizar las distancias, o bien maximizar las similaridades tal y como se demuestra a continuación:

a) En el caso de emplear **distancias**:

$$d(C_i, C_j) = Min_{(i_1, j_1)=1,2,\ldots.(n-k)|i_1 \neq j_1}\{d(C_{i1}, C_{j1})\} =$$

$$Min_{(i_1, j_1)=1,2,\ldots.(n-k)|i_1 \neq j_1}\left\{Max_{x_l \in C_{i1}\ y x_m \in C_{j1}\ y} d(x_l, x_m)\right\}$$

$$(l = 1,2, \ldots., n_{i1})\ y\ (m = 1,2, \ldots., n_{j1})$$

b) En el caso de emplear **similaridades**:

$$s(C_i, C_j) = Max_{(i_1, j_1)=1,2,\ldots.(n-k)|i_1 \neq j_1}\{s(C_{i1}, C_{j1})\} =$$

$$Max_{(i_1, j_1)=1,2,\ldots.(n-k)|i_1 \neq j_1}\left\{Min_{x_l \in C_{i1}\ y x_m \in C_{j1}\ y} s(x_l, x_m)\right\}$$

$$(l = 1,2, \ldots., n_{i1})\ y\ (m = 1,2, \ldots., n_{j1})$$

1.7.3. Método Average-Linkage (de la media)

Estos métodos tienen en cuenta la distancia entre dos clúster, como la distancia media de cualquier miembro de un clúster a otro miembro de otro clúster, (Ward, 1963) y (Murtagh, 1984)

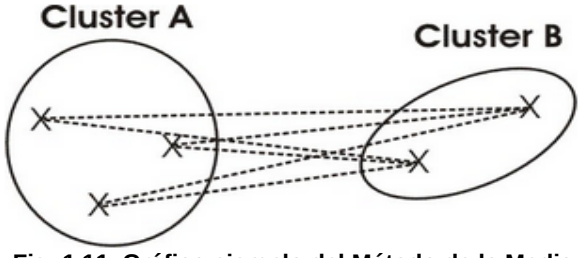


Fig. 1.11 Gráfico ejemplo del Método de la Media

Por lo tanto, si el clúster C_i que tiene n_i elementos está compuesto por 2 clusters, a saber $C_{i1}\ y\ C_{i2}$ con $n_{i1}\ y\ n_{i2}$ elementos respectivamente cada uno, y el clúster C_j que tiene n_j elementos, por lo que la distancia o similitud será la siguiente,

$$d(C_i, C_j) = \frac{d(C_{i1}, C_j) + d(C_{i2}, C_j)}{2}$$

1.7.4. Método del Centroide

Este método calcula la distancia entre cada centroide de cada clúster, pero este centro se puede mover cuando los centroides de diferentes clúster están cerca uno del otro. En consecuencia, la distancia entre los grupos combinados se puede reducir entre distintos pasos y, por tanto, puede hacer que existan problemas en los resultados del análisis (Wolfson et al, 2004), por lo que la semejanza entre 2 clusters viene dada por la semejanza entre sus centroides:

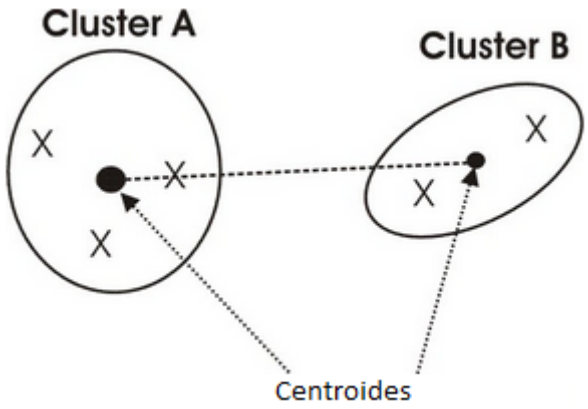


Fig. 1.12 Gráfico ejemplo del Método del Centroide

En este método, se tiene en cuenta el tamaño de los clusters a la hora de realizar los cálculos. Así, suponiendo que pretendemos medir la distancia entre el clúster C_i que tiene n_i elementos y está compuesto por 2 clusters, a saber $C_{i1}\ y\ C_{i2}$ con $n_{i1}\ y\ n_{i2}$ elementos respectivamente cada uno, y el clúster C_j que tiene n_j elementos, sabiendo que m^j , $m^{i1}\ y\ m^{i2}$ son los centroides de los clusters anteriormente descritos y siendo vectores n dimensionales (Gallardo, 1994).

Siendo las componentes del centroide del clúster C_i en notación vectorial:

$$m_l^{i1} = \frac{n_{i1}m_l^{i1} + n_{i2}m_l^{i2}}{n_{i1} + n_{i2}} \qquad (l = 1,2,\ldots.,n)$$

Por tanto la distancia euclídea al cuadrado entre los 2 clusters, viene dada por::

$$d^2(C_i, C_j) = \sum_{l=1}^{n} \left(m_l^j - m_l^i\right)^2$$

$$d^2(C_i, C_j) = \frac{n_{i1}}{n_{i1} + n_{i2}}\, d^2(C_{i1}, C_j) + \frac{n_{i2}}{n_{i1} + n_{i2}}\, d^2(C_{i2}, C_j) - \frac{n_{i1} n_{i2}}{(n_{i1} + n_{i2})^2}\, d^2(C_{i1}, C_{i2})$$

1.7.5. Método de la Mediana

Este método es similar al método del centroide. Si el tamaño de los clústers es muy distinto, entonces el centroide del nuevo clúster estará situado muy cerca del más grande, e incluso pudiera estar dentro del mismo, por esta razón Gower (1967) sugiere una estrategia alternativa, llamada "*método de la mediana*", y puede ser adecuado tanto para medidas de distancia, como para medidas de similitud (Mucha and Sofyan, 2000).

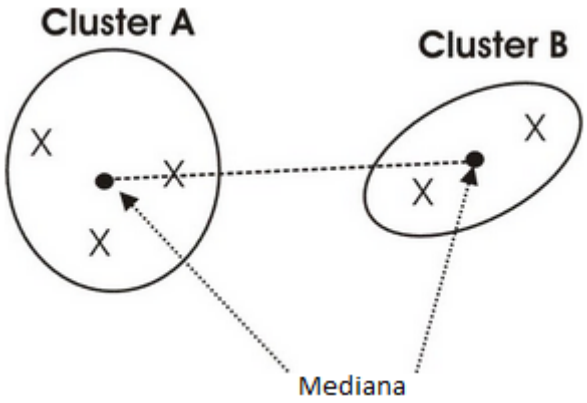


Fig. 1.13 Gráfico ejemplo del Método de la Mediana

En este método, no se tiene en cuenta el tamaño de los clústers a la hora de realizar los cálculos, ya que si los tamaños de las dos agrupaciones n_{i1} y n_{i2} del clúster C_i son muy diferentes entre sí, puede ocurrir que el centroide de dicho clúster m^i este influenciado excesivamente por el componente de mayor tamaño y, por tanto, las cualidades del más pequeño no se tengan en cuenta.

En definitiva, la estrategia a seguir en este método para calcular la distancia mediana, al considerar que $n_{i1} = n_{i2}$, provoca que el centroide del clúster C_i esté situado entre los clusters C_{i1} y C_{i2} y por lo tanto el centroide del clúster en estudio

(C_i, C_j) esté localizado en el punto central o mediana del triángulo formado por los clusters C_{i1}, C_{i2}, C_j , (Gallardo, 1994).

a) En el caso de emplear **distancias**:

$$d(C_i, C_j) = \frac{1}{2}\left[d(C_{i1}, C_j) + d(C_{i2}, C_j)\right] - \frac{1}{4}\, d(C_{i1}, C_{i2})$$

b) En el caso de emplear **similaridades**:

$$s(C_i, C_j) = \frac{1}{2}\left[s(C_{i1}, C_j) + s(C_{i2}, C_j)\right] + \frac{1}{4}\left[1 - s(C_{i1}, C_{i2})\right]$$

1.7.6. Método de Ward

El principal objetivo de este método es unir clúster en los cuales la variación dentro de ellos no haya aumentado de manera significativa, con lo que los nuevos clúster formados son más homogéneos (Mucha y Sofyan, 2000). La técnica se basa en la suma de cuadrados y tiende a crear grupos de similar tamaño, es por ello un buen método para el análisis de la varianza ya que tiende a producir clúster claramente definidos (Ward, 1963).

Supongamos que tenemos dos clúster, con los valores observados de sus "p" variables de la siguiente forma:

	Unidad	X_1	X_2			X_P	
A	1	x_{A11}	x_{A12}	.	. .	x_{A1p}	
	2	x_{A21}	x_{A22}	.	. .	x_{A2p}	
	.	.	.	.	. .	.	
	.	.	.	.	. .	.	
	.	.	.	.	. .	.	
	.	.	.	.	. .	.	
	n_A	x_{AnA1}	x_{AnA2}	.	. .	x_{AnAp}	
		$\bar{x}_{A1}$	$\bar{x}_{A2}$	.	. .	$\bar{x}_{Ap}$	**(vector de medias del clúster A)**
B	1	x_{B11}	x_{B12}	.	. .	x_{B1p}	
	2	x_{B21}	x_{B22}	.	. .	x_{B2p}	
	.	.	.	.	. .	.	
	.	.	.	.	. .	.	
	.	.	.	.	. .	.	
	.	.	.	.	. .	.	
	n_B	x_{BnB1}	x_{BnB2}	.	. .	x_{BnBp}	
		$\bar{x}_{B1}$	$\bar{x}_{B2}$	.	. .	$\bar{x}_{Bp}$	**(vector de medias del clúster B)**

Fig. 1.14 Tabla ejemplo de los dos clúster formados

Definimos por tanto la suma de cuadrados dentro de los clúster **"A"** y **"B**" de la siguiente forma:

$$SCD_A = \sum_{j=1}^{p} \sum_{m \in A} (x_{Aj_m} - \bar{x}_{Aj})^2$$

$$SCD_B = \sum_{j=1}^{p} \sum_{m \in B} (x_{Bj_m} - \bar{x}_{Bj})^2$$

Si los 2 grupos formasen un nuevo grupo, que vamos a llamar **"C",** con un vector de medias que fuese de igual forma que los anteriores, la suma de cuadrados seria de la forma:

$$SCD_C = \sum_{j=1}^{p} \sum_{m \in (A \cup B)} (x_{Cj_m} - \bar{x}_{Cj})^2$$

Por lo tanto, la disimilaridad entre los clusters **"A"** y **"B"** es el incremento que se produce en la suma de cuadrados al unir ambos grupos (es decir, que el método de Ward consiste en la unión de los grupos que tengan menor incremento en suma de cuadrados), uniéndose de esta forma los clusters más homogéneos (Molina, 2008).

$$d(AUB) = SCD_C - (SCD_A + SCD_B)$$

1.8. Método de Clasificación de tipo No Jerárquico

Los métodos no jerárquicos o particionados, tienen como objetivo principal realizar una sola partición de los individuos en **"K"** grupos o clústers, lo que significa que hay que especificar a priori el número de agrupaciones que se quieren formar, siendo ésta la diferencia fundamental con los métodos jerárquicos (aunque también este tipo de métodos trabajan con la matriz original, y por tanto no es necesario realizar una conversión a una matriz de distancias o similitudes).

Pedret (1986) agrupa los métodos no jerárquicos en 4 familias: "**Método de Reasignación o K-Medias"**, "**Análisis Quick-Clúster"**, "**Método Forgy"** y "**Método de Nubes Dinámicas"**. A continuación, realizaremos una breve descripción de cada una de estas familias ya que aunque no son el objeto central de este trabajo, presentaremos los "seminal papers" de cada caso para aquellos lectores interesados en ampliar detalles al respecto.

1.8.1. Método de K-Medias

Permite que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior, si con ello se puede optimizar el criterio de selección, acabando el proceso cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido anteriormente.

Existen 2 métodos que son denominados de igual forma, en un principio Forgy (1965) propuso un primer método de reasignación, que consiste en la iteración sucesiva hasta que se obtenga convergencia, mediante la representación de una clúster por su centro de gravedad y posteriormente asignar los individuos al clúster cuyo centro de gravedad esté más cercano.

McQueen (1967), propuso otro método muy similar, siendo este en la actualidad el que lleva el nombre de "K-Medias", donde también se representan los clúster por su centro de gravedad y se examina cada individuo para asignarlo al clúster más cercano, pero la diferencia con el método de Forgy, es que en el método de McQueen inmediatamente después de haber asignado un individuo a un clúster, el centro de gravedad es recalculado nuevamente, mientras que en el método de Forgy primero se hacen todas las asignaciones y luego se recalculan los centros.

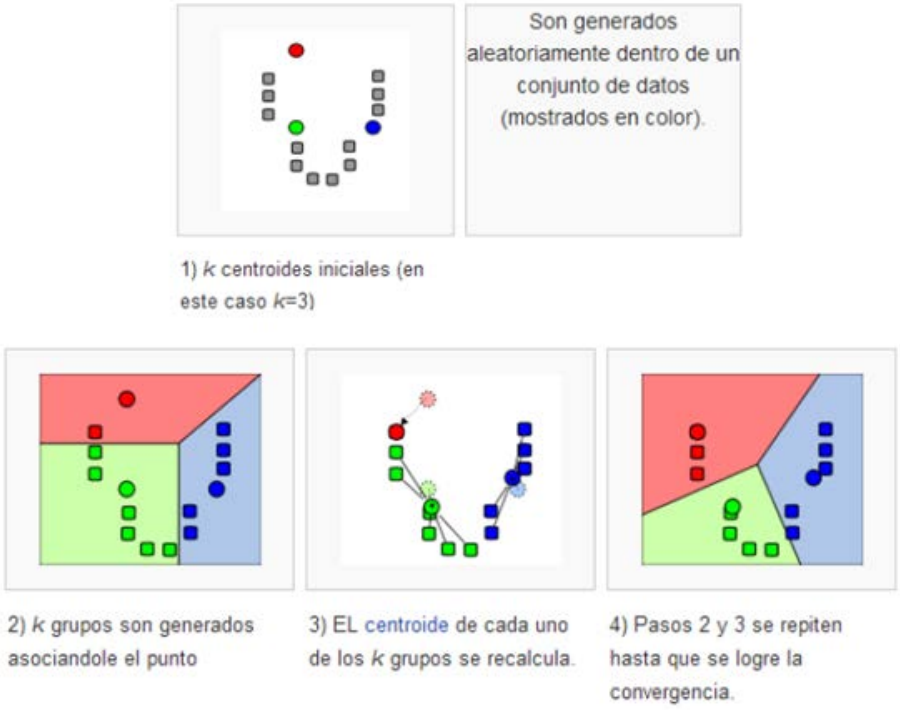


Fig. 1.15 Esquema de los pasos que sigue el algoritmo K-Medias

Los pasos del algoritmo de McQueen, son los siguientes:

1º. Se toman los primeros “k” individuos del conjunto y los considera como clusters de un único individuo.

2º. Se asigna a cada uno de los individuos restantes (según el orden de entrada de la matriz de datos) al centroide más próximo, para volver a recalcular el nuevo centroide.

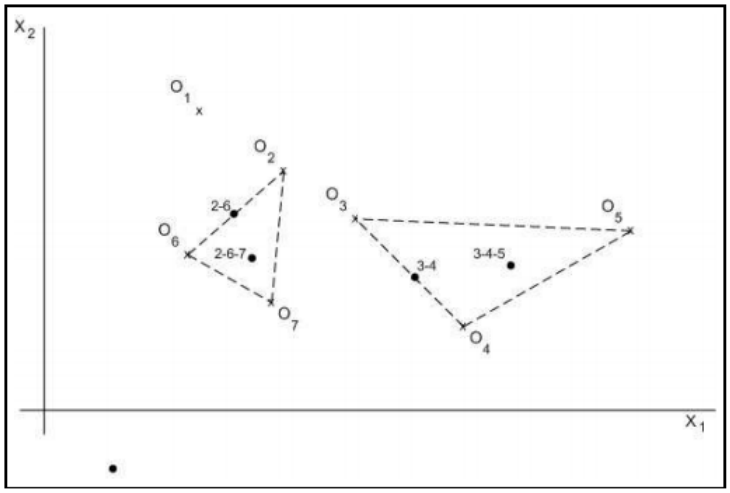


Fig. 1.16 Esquema de la asignación a los centroides próximos al realizar el paso 2.

3º. Con los centros que se han obtenido en el paso anterior, se calculan los centroides de los clusters obtenidos y se asigna cada individuo al centroide más cercano.

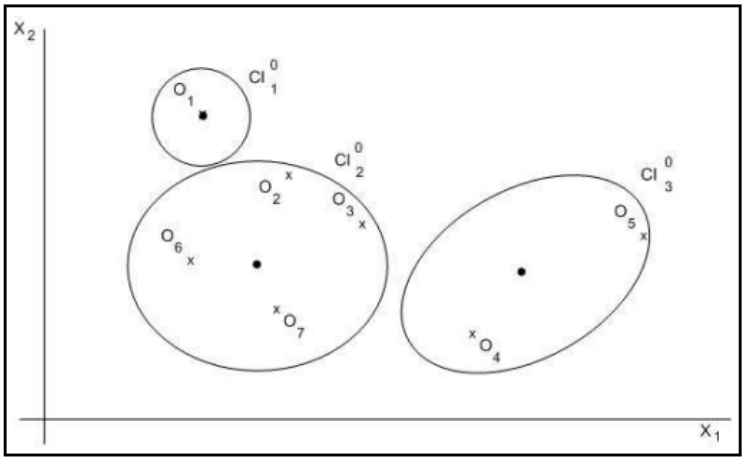


Fig. 1.17 Esquema de la agrupación en clúster al realizar el paso 3.

4º. Repetir el paso 2º y 3º hasta que se alcance un determinado criterio de parada.

La principal ventaja de este método, es que necesita menos operaciones que los otros métodos, pero requiere que los datos estén ordenados previamente dentro del conjunto inicial. Otra razón de la popularidad de este algoritmo es su facilidad de interpretación, así como su facilidad de aplicación (Dhillon y Modha, 2001); pero su uso está limitado a variables numéricas, aunque Huang (1998) presenta un prototipo sin limitaciones sobre el tipo de las variables, es decir con atributos numéricos y categóricos, siendo la medida de similaridad de las variables numéricas la distancia euclídea al cuadrado y la medida de similaridad para las variables categóricas el número de desajustes entre los individuos y los clúster.

Kaufmann y Rousseeuw (1987) proponen una alternativa al método K-Medias, denominado PAM (‘*Partition Around Medoids*’) muy similar al algoritmo de McQueen, pero que, fundamentalmente, se diferencia de éste en la representación de los clúster, donde cada agrupación está representada por el individuo más centrado en el clúster. También es más robusto que el algoritmo de McQueen en presencia de ruido y de outliers, ya que está menos influenciado por los valores atípicos o por valores extremos. Sin embargo, su procesamiento es más costoso, aunque tienen en común que ambos métodos necesitan que el número de clúster se especifique previamente, este método se verá más en detalle en el Capítulo IV del presente Trabajo Fin de Master.

1.8.2. Método de Quick-Clúster

Este método permite que un individuo asignado a un clúster en un determinado paso del proceso, sea también reasignado a otro clúster en otro paso posterior, si puede optimizar el criterio de selección. Es el procedimiento más adecuado para establecer perfiles en una muestra amplia de individuos, (Hair et al., 1998) y (Bisquerra, 1989).

1.8.3. Método de Forgy

Este método fue propuesto por Forgy (1965), y constituye la aproximación más simple al clustering particionado, y al igual que el método de McQueen o de K-medias, utiliza el concepto de centroide.

Los pasos del algoritmo de Forgy, son los siguientes:

1º. Comenzar con cualquier configuración inicial e ir al paso 2 si se comienza por un conjunto de k centroides o, por el contrario, ir al paso 3 si se comienza por una partición del conjunto de individuos en k clústers

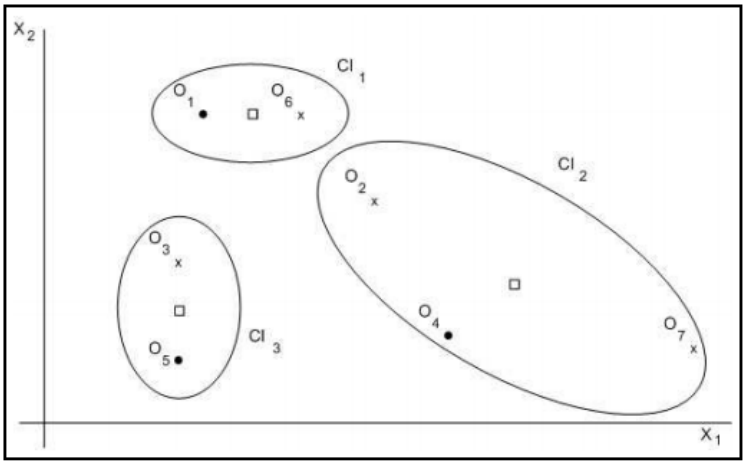


Fig. 1.18 Agrupación inicial en clúster en el paso 1

2º. Asignar a cada individuo a clasificar al centroide más próximo, permaneciendo fijos los centroides en este paso.

3º. Calcular los nuevos k centroides como los baricentros de los "k" clúster obtenidos.

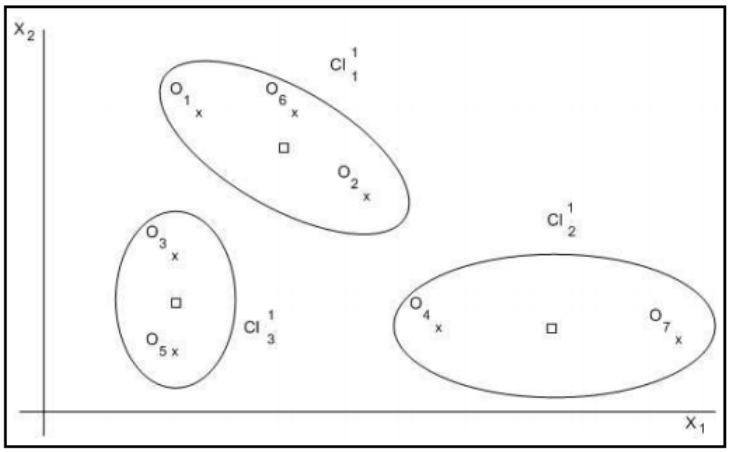


Fig. 1.19 Agrupación nueva en el paso 3

4º. Alternar los pasos 2 y 3, hasta alcanzar un determinado criterio de convergencia.

La variante de Jancey (1966) sólo modifica el paso 3, obteniendo el nuevo clúster mediante una reflexión con centro en el nuevo centro obtenido del antiguo clúster.

1.8.4. Método de las Nubes Dinámicas

Este método fue introducido por Diday (1972), generalizando el método de K-Medias de Forgy. Se basa en que cada clúster debe tener una representación llamada núcleo o centroide de clúster para, posteriormente, hacer una búsqueda iterada de centroides y de clúster, por lo que cada clase estará representada por un núcleo, que será un elemento representativo de todos los que integran la misma (Trejos, 1998).

Es decir, que en lugar de elegir como referencia de clase un sólo punto que constituye su centro y reasignar los puntos por proximidad a ese centro, se eligen varios puntos "h" para representar cada clase, siendo estos puntos el "núcleo" de la clase.

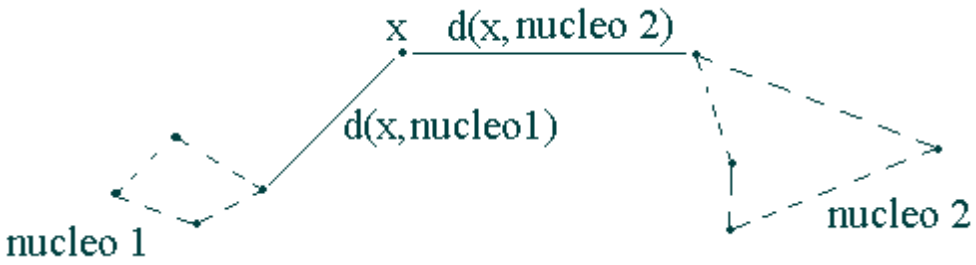


Fig. 1.20 Esquema del cálculo de distancia de x a los núcleos 1 y 2, como la mínima distancia a los "h" puntos del núcleo

Los pasos del algoritmo son los siguientes:

1º. Se establece un conjunto de "k" núcleos iniciales

2º. ***Fase de asignación***: Se forma un clúster, asignando cada individuo al clúster más próximo mediante el cálculo de la distancia al centroide.

3º. ***Fase de representación***: Se calculan los centroides de los clúster y se toman como nuevos centroides.

4º. Se repiten los pasos 2 y 3 hasta alcanzar un determinado criterio de convergencia, cuando las clases se estabilizan.

1.8.5. Método de Densidad

Los métodos de densidad, parten de la base de suponer que los puntos que pertenecen a cada clúster se extraen de una distribución de probabilidad particular (Banfield y Raftery, 1993).

La idea principal de este tipo de métodos, es identificar los clústers y sus parámetros de distribución, haciendo crecer una determinada agrupación hasta que la densidad del clúster vecino o más próximo sobrepase un cierto umbral (es decir, que contenga un determinado número de individuos dentro del clúster), los trabajos en este campo se han basado en la hipótesis de que la densidad de los clústers siguen distribuciones gaussianas para los casos en que las variables son numéricas, y distribuciones de tipo multinomial para los casos en que las variables son nominales. Por ello, una solución aceptable para este tipo de casos es tomar el principio de máxima verosimilitud.

Entre los algoritmos utilizados para este tipo de métodos podemos citar los siguientes:

1. El **algoritmo de la esperanza matemática "EM"** (Dempster et al., 1977), se aplica en este tipo de estimación de los parámetros. Comienza con una estimación inicial del vector parámetro y lo alterna entre dos pasos (Fraley y Raftery, 1998), de forma que en el primer **"paso-E"** se calcula la esperanza de la máxima verosimilitud entre los datos observados y el parámetro estimado, y en el segundo **"paso-M"** se maximiza la función de verosimilitud calculada en el paso primero.

2. El **algoritmo DBSCAN** donde sus siglas en ingles significan ('***D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise'*), el cual es eficaz para volúmenes de datos grandes. Realiza búsquedas sobre cada individuo existente en la Base de Datos y verifica si contiene más de un número mínimo de objetos (Ester et al., 1996).

3. El **algoritmo AUTOCLASS** que cubre una variedad de distribuciones del tipo, Gauss, Bernouili, Poisson, Log-normal (Cheeseman y Stutz, 1996)

4. El **algoritmo SNOB** es una aplicación MML ("***M**inimum **M**essage **L**ength")* al problema de la taxonomía numérica, cada variable continua se supone

que proviene de una distribución Normal y cada variable categórica proviene de una distribución Multinomial, este algoritmo utiliza la aplicación MML tanto para la selección del modelo como para la estimación de los parámetros (media, desviación estándar, número de grupos, etc.) (Wallace y Dowe, 1994).

5. El **algoritmo MCLUST** es una interface software para el análisis de clúster escrito en lenguaje Fortran e incluido en el software comercial S-Plus. La entrada es la matriz de datos, el máximo y mínimo número de clúster que se quiere formar. MCLUST compara los valores del Criterio de Información Bayesiano, *(Bayesian Information Criterion, BIC),* para la optimización de parámetros, aplicando distintas restricciones a la matriz de varianzas y covarianzas, (Fraley y Raftery, 1998)

Algunos de estos métodos constituyen una variante del método de las distancias mínimas (Vallejo, 1992), pero imponen reglas para evitar el problema de obtener un solo clúster cuando existen puntos intermedios, dentro de estos métodos se distinguen, aquellos que proporcionan una ***aproximación tipológica*** y los que proporcionan una ***aproximación probabilística***. En el primer tipo, los clúster se forman buscando las zonas en las cuales se da una mayor concentración de individuos, encontrándose entre ellos, el *Análisis modal de Wishart* (1969), que parte del supuesto de que los clusters son esféricos, el *Método Taxmap* (Charmichael et al., 1968), en donde una de las características esenciales de este método es que tiene en cuenta el problema del chaining (es decir cuando los clúster no forman grupos claramente aislados sino que forman uno continuo, el usuario debe de introducir el valor de corte, lo que proporciona una subjetividad a los resultados obtenidos) y el *Método de Fortin* (Fortin et al., 2002).

En el segundo tipo, se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata por tanto, de encontrar los individuos que pertenecen a la misma distribución. Dentro de este tipo se encuentra el *Método de las Combinaciones de Wolf* (1971).

1.8.6. Método de Block-Clustering

La agrupación simultánea, generalmente llamado '*Bi-Clustering'*, '*Co-Clustering*' o '*Block-Clustering*' ('agrupamiento en bloques'), es una técnica importante en el análisis de datos bidireccional. El término fue introducido primeramente por Mirkin (1996), aunque la técnica originaria fue introducida por Hartigan (1975).

El objetivo de este método es encontrar sub-matrices, con filas y columnas con una alta correlación, pero uno de los problemas que plantean es que el número de clúster a calcular debe de ser suministrado previamente al cálculo del algoritmo. Sin embargo, estas estrategias sólo pueden ser realizadas utilizando algoritmos de una vía y existe una falta de enfoque claro para conseguir el mejor número de clústers en algoritmos de '*Block-Clustering*' (Charrad et al., 2010).

El algoritmo CROK12 propuesto por Govaert (1983), Govaert (1995) y Nadif y Govaert (2005) se basa en una versión adaptada del algoritmo K-medias, visto en el punto 1.8.1 del Capítulo I del presente Trabajo Fin de Master, basado en la distancia Chi-Cuadrado, de manera que se aplica a la tabla de contingencia para identificar una partición fila "P" y una partición columna "Q" que maximicen el valor Chi-cuadrado $\boldsymbol{\chi^2}$**.**

Este algoritmo consiste en aplicar el algoritmo K-medias inicialmente, tanto en filas, como en columnas, de forma alternativa para construir una serie de pares de particiones (P^n, Q^n) que optimicen el valor Chi-Cuadrado $\boldsymbol{\chi^2}$ de la nueva matriz de datos.

El objetivo, por tanto, es encontrar una partición fila $P = (P_1, P_2, P_3, \dots\dots, P_K)$ compuesta por "**K"** clusters, y una partición columna $Q = (Q_1, Q_2, Q_3, \dots\dots, Q_L)$ compuesta por "**L"** clusters, que maximicen el valor de Chi-cuadrado $\boldsymbol{\chi^2}$ de la nueva tabla de contingencia (P, Q) obtenida mediante la agrupación de las filas **K** y las columnas **L**.

$$\boldsymbol{\chi^2(P,Q) = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(f_{kl} - f_{k\cdot}\, f_{\cdot l})^2}{f_{k\cdot}\, f_{\cdot l}}}$$

$$\boldsymbol{f_{kl} = \sum_{i\in P_k}\sum_{j\in Q_l} f_{ij}}$$

$$f_{k\cdot} = \sum_{l=1,L} f_{kl} = \sum_{i \in P_k} f_{i\cdot}$$

$$f_{\cdot l} = \sum_{k=1,K} f_{kl} = \sum_{j \in Q_l} f_{\cdot j}$$

$$T_1(k,l) = \sum_{i \in P_k} \sum_{j \in Q_l} a_{ij} \qquad (nueva\ tabla\ de\ contingencia)$$

Siendo los pasos de ejecución:

1º. Empezar en la posición inicial (P^0, Q^0).

2º. Calcular (P^{n+1}, Q^{n+1}) comenzando en (P^n, Q^n), mediante dos pasos, a saber:

   a. Calcular (P^{n+1}, Q^n) comenzando en (P^n, Q^n), aplicando el algoritmo K-medias en la partición P^n.

   b. Calcular (P^{n+1}, Q^{n+1}) comenzando en (P^{n+1}, Q^n), aplicando el algoritmo K-medias en la partición Q^n.

3º. Repetir el paso 2º hasta que exista convergencia.

1.8.7. Método de Reducción de Dimensiones

Estos métodos consisten en la búsqueda de unos factores en el espacio de los individuos que contiene la matriz de datos, donde cada factor corresponde a un clúster y se les conoce como "*Análisis Factorial tipo Q*" (Overall y Klett, 1972).

El objetivo consiste en encontrar grupos de individuos con valores semejantes en las variables, con la finalidad de determinar un número reducido de clúster, de los que se espera que los individuos contenidos en cada uno de ellos tengan algún tipo de propiedad común. El método parte de la matriz de correlaciones entre individuos y somete los factores encontrados a una rotación ortogonal (Vallejo, 1992), el problema es que los individuos pueden pertenecer a varios y, por tanto, los clusters pueden presentar solapamiento, resultando (por lo tanto) compleja su interpretación.

1.9. Método de Clasificación de tipo Two-Step

Al método Two-Step (también llamado Bietápico o método de agrupamiento en dos fases), se le dedica en exclusiva el Capítulo II para su análisis en detalle, por ser el objetivo principal del presente Trabajo Fin de Master. Desarrollado por Chiu, Fang, Chen, Wang y Jeris (2001), permite trabajar con grandes volúmenes de datos, y por tanto gestionar grandes matrices de datos.

El nombre Two-Step, es una clara indicación de que el algoritmo está basado en dos fases, de forma que el algoritmo en la primera ejecuta un algoritmo muy similar al K-medias y, posteriormente y basado en los resultados obtenidos de la primera fase, ejecuta una modificación de un procedimiento jerárquico aglomerativo tradicional, (visto en el apartado 1.7 del presente Capítulo I), y que combina los individuos secuencialmente para formar clusters homogéneos, siendo capaz de trabajar y combinar variables de tipo continuo con variables de tipo categórico.

Dedicaremos el próximo capítulo a estudiar este método en profundidad.

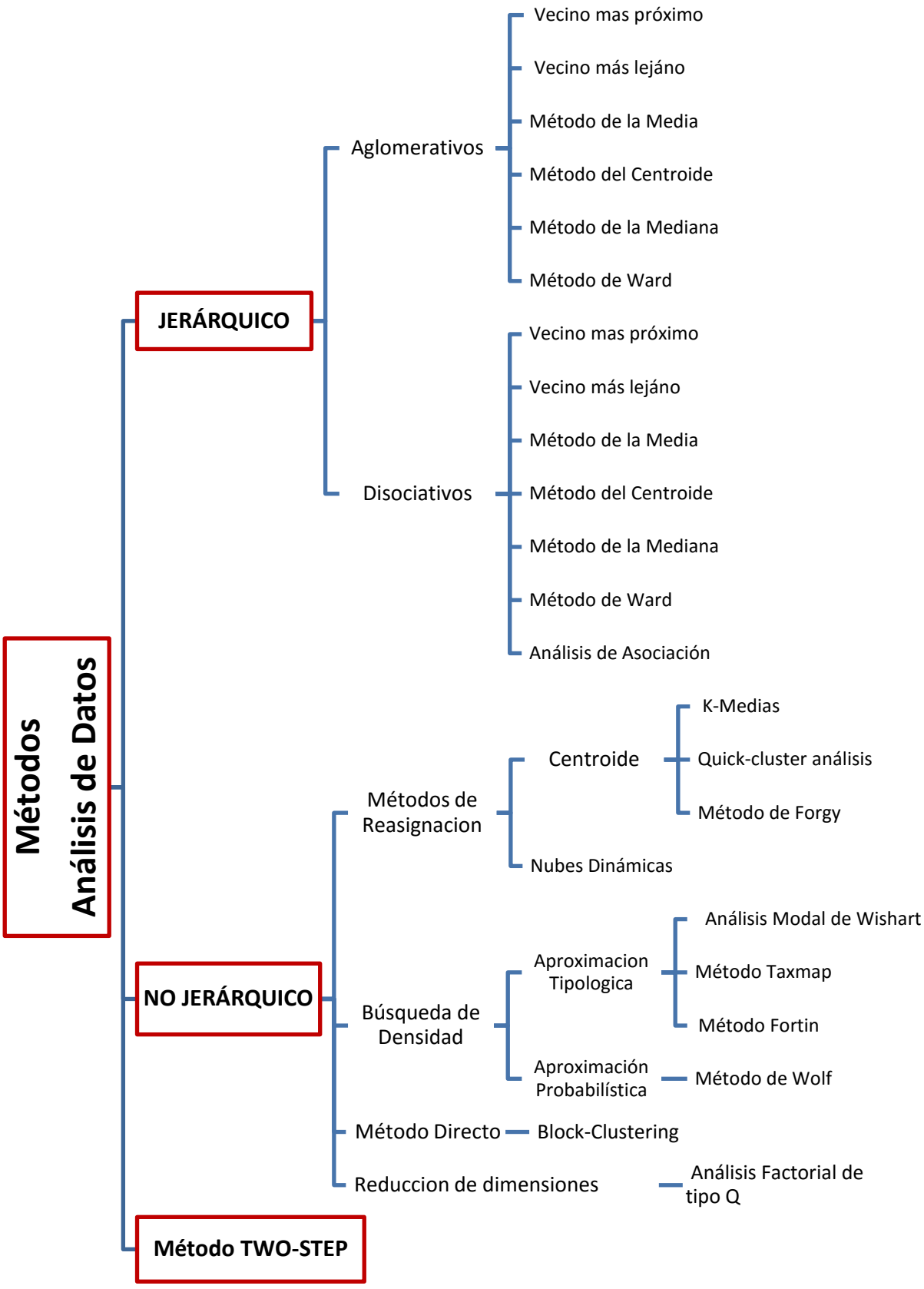


Fig. 1.20 Diagrama con los diferentes Tipos de Métodos en Análisis de Clúster

Capítulo II

AGRUPAMIENTO EN CLUSTER

Método TWO-STEP

2.1. Introducción

Los métodos anteriormente descritos presentan problemas cuando hay un número de individuos muy grande para ser clasificados (en minería de datos, por ejemplo). Como alternativa surge en la literatura el Método “Two-Step” o “Bietápico” (“en dos fases”), propuesto por Chiu et al. (2001), que ha conseguido hacerse muy popular porque es único y ademas está implementado en el SPSS[1]. SPSS de IBM (IBM Inc., 2001).

Esto motiva que casi cada día podamos encontrar nuevos trabajos de investigación publicados donde la técnica central de Análisis es el Método Two-Step, poniéndolo incluso en el título del trabajo.

Los tres aspectos fundamentales y diferenciadores del resto de técnicas de clúster son: permite trabajar con información mixta (variables categóricas y continuas), realiza una selección automática del número de clusteres y permite el análisis de grandes volúmenes de datos.

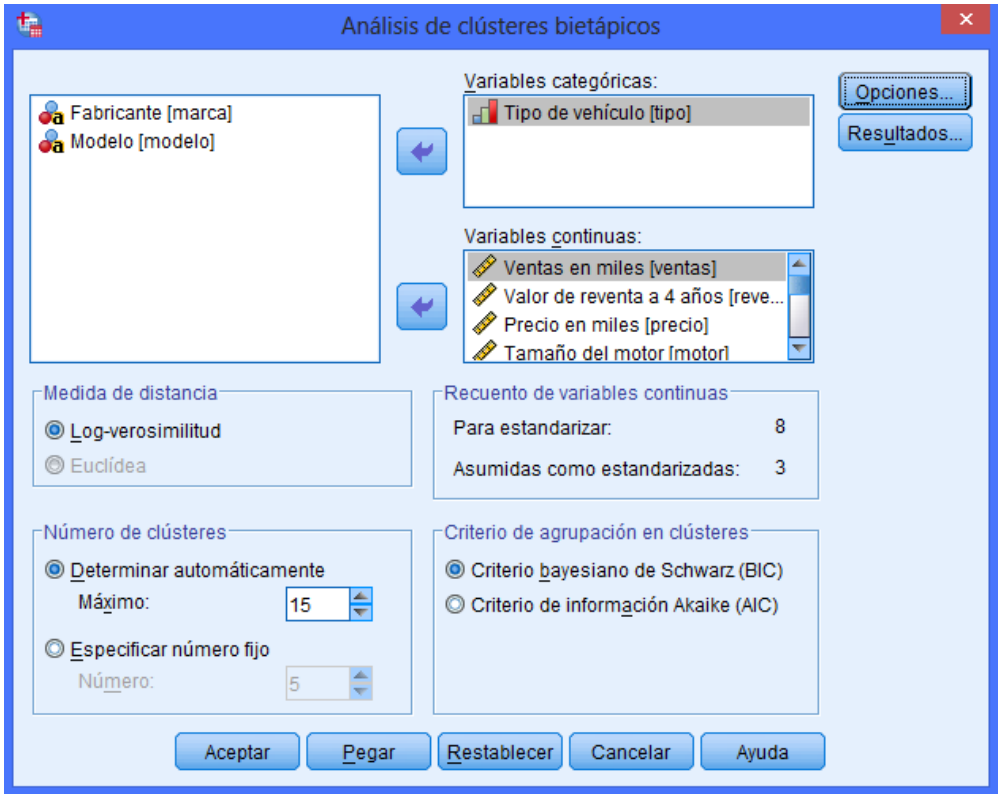


Fig. 2.1 Pantalla petición sobre la matriz de Datos del Software SPSS

[1] Los autores del artículo son precisamente desarrolladores de SPSS.

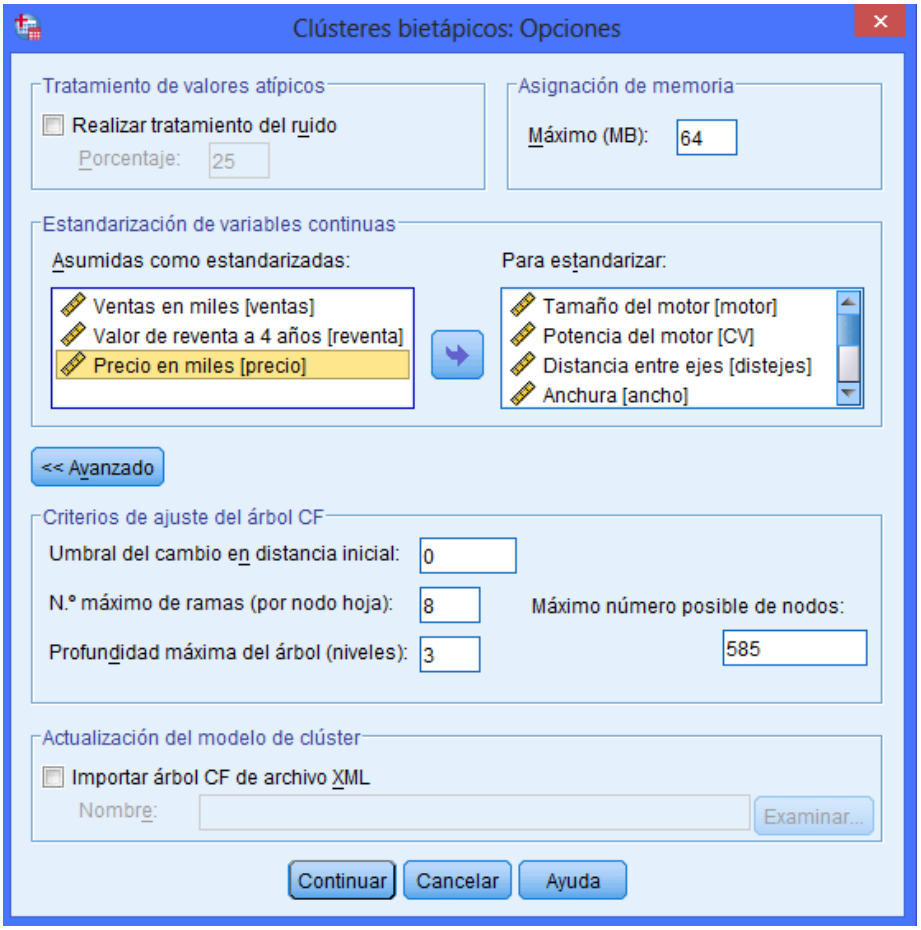


Fig. 2.2 Pantalla opciones sobre la matriz de Datos del Software SPSS

Un trabajo en el que se inspiran los creadores del método es el de Guha et al. (1999), donde los autores propusieron un método de análisis de clúster que presentaron como ROCK (*ROCK: A robust clustering algorithm for categorical attributes*), el cual está basado en describir links entre relaciones booleanas creadas a partir de lo que ellos llaman transaciones y que no es más que trabajar en vez de con los individuos/items directamente, trabajar con categorías de elementos y definir variables que tomarán el valor 1 si el elemento pertenece a la transación concreta o cero si no pertenece[2]. Usa el concepto que llama "link" para medir la similaridad entre pares de individuos o registros con variables categóricas, analizando la similaridad entre las variables booleanas creadas a partir de las transaciones estudiadas.

Banfield y Raftery (1993) habían introducido un modelo basado en la medida de distancias para individuos con variables continuas. Chiu et al. (2001) extendieron

[2] Los autores ponen un bonito ejemplo: productos de venta para bebes (leche, cereales, potitos, galletas, etc)
Para afirmar que la cesta de la compra de dos familias es similar no hace falta que las dos cestas tengan todo, ni exactamente los mismos productos, pero si una (transación) tienen potitos y leche de bebes y otra (transación) cereales y galletas para bebes, ambas transaciones nos permiten afirmar que esos dos perfiles de compradores tienen una alta similaridad; existe, pues, un link entre ellas). Observese que trabajar comparando transaciones requiere unos tiempos de computo mucho mas pequeños y nos llevará a resultados tan buenos como si analizaramos ítem a ítem.

los modelos basados en la distancia, para modelos con variables mixtas, es decir modelos con variables categóricas y continuas al mismo tiempo, basándose en las propuestas de estos dos artículos.

El algoritmo determina de forma automática el número óptimo de clusters. Esta característica no figura en los métodos de agrupamiento tradicionales, sino que se suele realizar de forma separada y realizando pruebas sobre diferentes modelos cambiando manualmente el número de grupos que se quieren realizar en cada momento.

Permite el análisis de grandes volúmenes de datos, mediante la construcción de un árbol de características, cuyas siglas en inglés son “Characterictics Features (CF)”, lo que los otros autores llamaron transaciones por el contexto en el que lo presentaron.

Esta característica tiene una importancia vital en el análisis de clúster de grandes volúmenes de datos, ya que los algoritmos de clúster tradicionales mantienen en la memoria del procesador del ordenador, el conjunto de datos que están analizando para así poder realizar su división en grupos homogéneos, sin embargo cuando el conjunto de datos es muy grande, este no puede ser cargado en la memoria principal de una sola vez, lo cual ocasiona una paginación de la memoria del procesador del ordenador, ralentizando de forma dramática su proceso, lo que hace que el tiempo de ejecución del análisis aumente exponencialmente en función del número de registros o de individuos que tenga la base de datos que se está analizando.

El árbol de características (CF) y su solución final pueden depender del orden de los individuos o casos que se traten, por lo que es necesario que estén ordenados de forma aleatoria para minimizar el efecto del orden inicial. Si se dispone de archivos de datos demasiado grandes, es probable que sea más difícil o que no sea factible realizar este tipo de acciones, por lo que se podrán sustituir varias ejecuciones por una muestra de casos ordenados con diferentes órdenes aleatorios.

2.2. Método BIRCH

Otro de los trabajos sobre los que se asienta la propuesta del TwoStep, es el propuesto por Zhang, et al. (1996) en el cual desarrollan el método **BIRCH** (*"Balanced Iterative Reducing and Clustering using Hierarchies"*) el cual, según los autores, es el primer algoritmo de agrupamiento vinculado al ámbito de las bases de datos.

Es un método de clusterización basado en distancias (utiliza la Euclídea y la Manhattan) que no exige independencia entre las variables. Con respecto a otros métodos que trabajan sobre distancias presenta la ventaja de que en lugar de chequear cada registro, analiza la densidad de la estructura. De esta forma no necesita inspeccionar todos los registros en cada agrupación, sino los cercanos.

Una contribución importante del algoritmo BIRCH, es la formulación del problema de agrupamiento en clúster de una manera apropiada para manejar grandes volúmenes de bases de datos lo cual supone una gran ventaja en términos de tiempo de cómputo y de memoria de procesador. (Zhang et al., 1996)

1.- Es un algoritmo local, en que cada decisión de agrupamiento se realiza sin escanear todos los datos, utiliza mediciones que reflejan la cercanía natural de los puntos de la matriz, y al mismo tiempo puede ser mantenido incrementalmente durante el proceso de agrupamiento.

2.- Explota al máximo la observación de que el espacio de datos no suele ser ocupado de manera uniforme, y por lo tanto no todos los puntos de la matriz de datos es igualmente importante para los propósitos de agrupamiento, por tanto una región densa de puntos se trata conjuntamente como si fuese un solo grupo y los puntos de la matriz de datos que están en regiones dispersas se tratan como valores atípicos y por lo tanto se eliminan opcionalmente.

3.- Hace un uso completo de la memoria disponible en el procesador para obtener los mejores clusters posibles, reduciendo al mínimo las instrucciones de entrada y salida del procesador, y se caracteriza por el uso de una estructura de árbol en la memoria del procesador, de forma equilibrada, y debido a estas características, el tiempo de ejecución es linealmente escalable, ya que escanea y carga los datos de la matriz una sola vez en memoria.

El método BIRCH introduce los conceptos Clustering Feature y Clustering Feature tree (CF) o árbol de Características del Clúster, en donde este árbol se utiliza como forma de representación de manera más resumida de los clusters creados, para lograr mayor velocidad de proceso y mayor escalabilidad a la hora de aplicar el análisis de clúster a grandes volúmenes de datos.

Este árbol CF de características de clúster, es en definitiva un resumen estadístico del clúster, que tiene fundamentalmente tres parámetros (Zhang, 1996)

$$\boldsymbol{CF} = (\boldsymbol{N}, \boldsymbol{LS}, \boldsymbol{SS})$$

Donde los parámetros incluidos se definen como sigue:

- N: Número de individuos incluidos en un clúster.
- LS: Suma de los atributos de los N individuos $\left[\sum_{i=1}^{N} \boldsymbol{X_i}\right]$
- SS: Suma al cuadrado de los N individuos $\left[\sum_{i=1}^{N} \boldsymbol{X_i^2}\right]$ de un clúster.

El Clustering Feature (un vector conformado por la anterior tripleta) y el teorema de aditividad que permite combinar los subclusters formados, posibilita recoger la información esencial contenida en un cluster y procesar únicamente esa información para realizar la agrupación.

2.2.1. Algoritmo del Método BIRCH

El procedimiento del algoritmo BIRCH, se basa en cuatro fases:

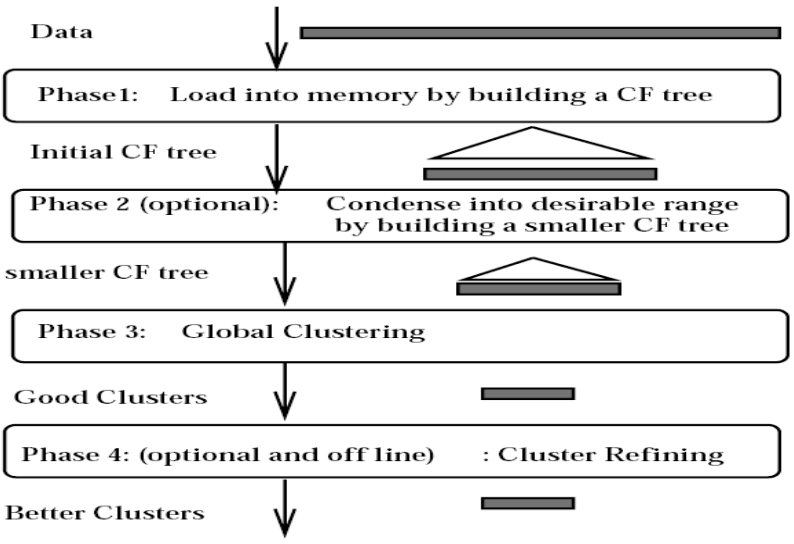


Fig. 2.3 Gráfica de proceso de las etapas del algoritmo BIRCH. Fuente: Zhang et al. (1996).

1) **FASE 1 (CF-Tree Insertion)**

Analiza todos los datos y construye un *"CF-Tree"* inicial en la memoria principal, usando para ello la cantidad de memoria y de espacio en disco disponible. Este "*CF-Tree*" intenta reflejar en una estructura clúster toda la información del conjunto de datos, tan bien como sea posible, bajo el límite impuesto por la memoria de manera que se asigna cada individuo leído de la matriz de datos a una rama o a otra, teniendo en cuenta que si la distancia de un individuo nuevo y los anteriores se hace mayor que un parámetro definido, entonces se creará una nueva rama. En esta fase inicial se crea en memoria un *"CF Tree"* inicial.

2) **FASE 2 (Opcional)**

Esta segunda fase es opcional, pues el objetivo es revisar nuevamente el árbol creado en la fase 1, y comprobar si es demasiado grande, para modificar el valor del parámetro asignado en la primera fase, de forma que si este parámetro aumenta entonces las ramas del árbol CF creado se juntarán, con lo que se generará un árbol más pequeño, y de esta forma la fase tres podrá ejecutarse con mayor rapidez para un gran volumen de datos. Su resultado es la creación de un *"CF-Tree"* más pequeño.

3) **FASE 3 (CF-Tree Rebuilding)**

En esta fase, se ejecuta un algoritmo de agrupación de tipo jerárquico aglomerativo sobre la información obtenida, bien de la fase primera o de la fase segunda, para obtener el conjunto de clusters que capture el principal patrón de distribución de los datos de la matriz original.

3) **FASE 4 (Opcional)**

En esta última fase, se utiliza los centroides de los clusters generados en la fase tercera y redistribuye los datos en base a un algoritmo para volver a agrupar los datos en base a su cercanía.

De lo anterior se deduce que: el "***CF-Tree***" es una estructura en forma de árbol definido a partir de 2 parámetros: el '*factor de ramificación*' (**B**) y el '*umbral*' (**T**), donde su tamaño depende directamente de este último factor. Esta estructura está conformada por 3 elementos tipo: el "*root node*" (o "nodo raíz") el "*non-leaf node*" (o

"nodo no-hoja") y el "l*eaf node*" (o "nodo hoja") y su construcción se basa en las siguientes premisas:

- El tamaño de los nodos está determinado por la dimensionalidad de los datos y por la de un tercer parámetro general: el '*tamaño de la página*' (**P**) que se deriva de la memoria principal disponible.
- El "*root node*" y cada "*non-leaf node*" tienen a lo sumo **B** 'CF-Entradas' (las marcadas a priori en este parámetro denominado, tal y como señalábamos antes, 'factor de ramificación'). Este último tipo de nodo representa un clúster conformado por cada una de las entradas a todos y cada uno de sus 'subclusters hijos' (o secundarios) que conforman los "leaf node". Cada "*leaf node*" tiene a lo sumo **L** 'CF-Entradas' y cada una de ellas debe satisfacer el umbral **T.**

Pero el algoritmo BIRCH presenta un inconveniente importante: funciona sólo cuando las variables son continuas y, en la práctica, muchas observaciones contienen también (o sólo) variables categóricas. Por lo tanto y aprovechando sus ventajas, era necesaria una mejora y un nuevo enfoque que abarcara la existencia de una serie estadística que contenga datos mixtos.

En este contexto aparece la investigación de Chiu et al. (2001) que permite manejar registros continuos y categóricos para un mismo conjunto de datos y que se conoce como método "TwoStep" o 'Clúster Bietápico'.

2.3. Algoritmo del Método Two-Step

El algoritmo Two-Step propuesto por Chiu et al. (2001) está implementado en el software SPSS (IBM® SPSS® Statistics),

El algoritmo TWO-STEP se realiza en dos etapas, en el primer paso, todos los registros son escaneados y almacenados como regiones densas, guardando un resumen estadístico de ellas, a continuación en el segundo paso cada región densa almacenada es tratada como un punto individual mediante un algoritmo jerárquico, y como el número de regiones densas es bastante menor que el numero inicial de individuos, el método jerárquico es el más eficiente para ello.

2.3.1. Formación del Pre-Clúster (FASE 1)

Se forma un clúster inicial, llamado pre-clúster, que corresponde a los datos originales y que se van a utilizar en lugar de las filas de los individuos de los datos originales, para posteriormente poder utilizarlos en la realización de los clusters jerárquicos, de forma que todos los individuos que pertenezcan al mismo pre-clúster se trataran como una entidad.

Siguiendo la notación del método BIRCH, propuesto por Zhang et al. (1997), donde el resumen de las estadísticas almacenadas de las regiones densas de los individuos eran las funciones características de dichas regiones

Se define CF como las Características del Clúster (Clúster Features en inglés), como:

$$CF = \{N, SA, SA^2, N_B\}$$

Donde los parámetros incluidos se definen como sigue:

- N: Tamaño del Clúster, es decir número de individuos que están incluidos en un clúster concreto.
- SA : Suma de los atributos de las variables continuas correspondientes al clúster.
- SA^2: Suma al cuadrado de los atributos de las variables continúas correspondientes al clúster.
- N_B: $\sum(L_k - 1)$, que es la suma de categorías de cada variable categórica

Cuando dos clusters se quieren fusionar, significa que las características de cada uno de los clúster almacenadas se unen para formar una nueva (CF), de forma que:

$$CF_{(j,s)} = \{N_j + N_s, SA_j + SA_s, SA_j^2 + SA_s^2, N_{Bj} + N_{Bs}\}$$

Esta nueva característica de clúster es un camino eficiente de representación de los datos, ya que guarda bastante menos información que todos los datos incluidos

en la región densa creada al inicio y ello es suficiente para calcular las medidas de similaridad que necesita el algoritmo.

El árbol CF de características se construye siguiendo la notación propuesto por Zhang et al. (1996), que consiste en una serie de nodos organizados por niveles, a su vez de cada nodo salen ramas que pueden contener también más nodos o acabar en una hoja, y una hoja finalmente representa el sub-clúster que vamos buscando.

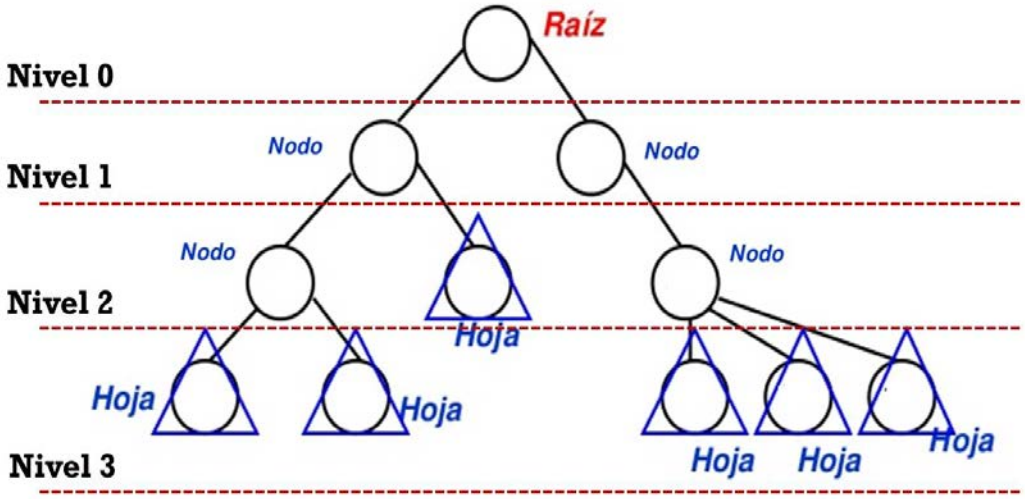


Fig. 2.4 Gráfica de formación del Árbol de características CF

Un árbol es por tanto una estructura de datos ramificada que puede representarse como un conjunto de nodos que están enlazados entre sí por medio de ramas, y estas a su vez tienen otros nodos de los que vuelven a salir más ramas, o bien acaban en ese punto en un nodo hoja.

Al nodo base, se le denomina *"**Nodo Raiz**"* y se identifica como el nodo que no tiene padre y se encuentra en el "***Nivel***" más bajo del árbol, desde este nivel y en orden ascendente se crearan el resto de niveles que tendrá el árbol en base al número de niveles especificado al algoritmo, todos estos niveles nos proporcionará la "***profundidad***" del árbol, desde este nodo raíz saldrán los "***Nodos Ramas*** que son aquellos que no pertenecen a ninguna de las otras categorías, Raíz u Hoja, finalmente el arbola acabará en "***Nodos Hoja***" que son aquellos que quedan identificados como los que no tienen hijos.

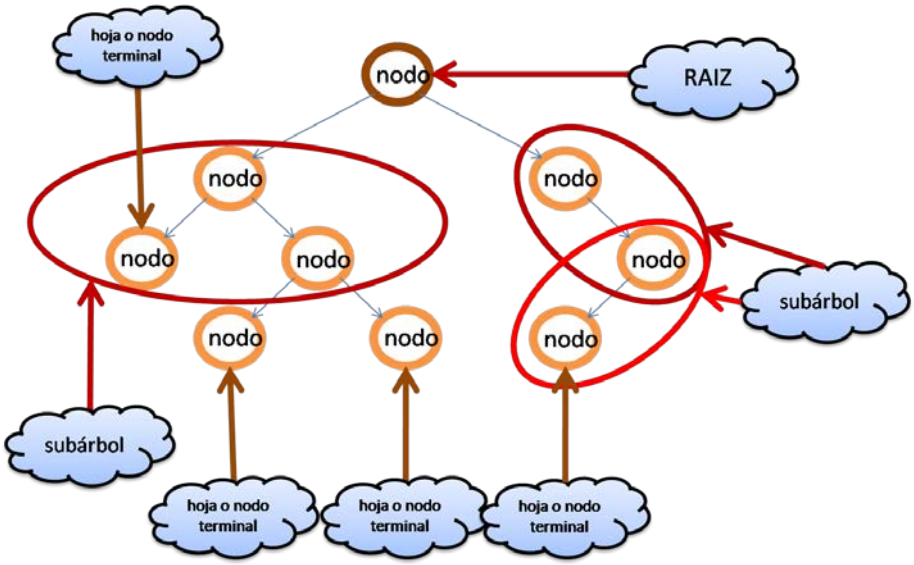


Fig. 2.5 Ejemplo de Descripción de las partes del Árbol de características CF

Cada nodo del árbol CF se caracteriza por los parámetros que incorpora en el proceso secuencial que sigue de escaneado de todos los registros, y que consiste en el número de registros que contiene, la media y la varianza de cada variable continua y el recuento de cada categoría de cada variable categórica.

El proceso empieza escaneando todos los registros usando el enfoque secuencial propuesto por Theodoridis y Koutroumbas (1999), para la construcción del árbol de características (CF), de forma que uno a uno se decide si el registro leído debe de fusionarse con el registro anterior o por el contrario se debe de formar un nuevo clúster basándose en el criterio de distancia que se haya seleccionado.

El árbol empieza colocando el primer individuo en la raíz del árbol y en un nodo de hoja que contiene toda la información de la variable sobre ese individuo, de forma que cada individuo se añade a continuación a un nodo existente o forma uno nuevo, tomando como base la similaridad o disimilaridad que tenga con los nodos existentes y utilizando la medida de distancia definida entre ellos.

Un nodo que contiene varios individuos, contiene un resumen sobre la información de los individuos que lo integran, y por consiguiente el árbol CF proporcionará un resumen del archivo de datos utilizado.

Cuando un registro de datos pasa a través de un nodo hoja, busca la entrada más cercana y pasa hasta el nodo hijo, el proceso continúa de forma recursiva y los registros descienden a través del árbol (CF) para alcanzar un nodo hoja, cada registro

es absorbido dentro de un nodo hoja si la distancia del registro a la entrada más cercana se encuentran dentro del umbral definido, en caso contrario se crea una nueva entrada en el nodo hoja.

Si el árbol de características (CF) crece por encima del tamaño máximo permitido, se vuelve a reconstruir con un criterio más apropiado, el nuevo árbol de características (CF) será más pequeño y por tanto con más espacio para los registros entrantes, el proceso continuará hasta que todos los registros correspondientes a todos los individuos de la matriz inicial hayan pasado por este proceso.

2.3.2. Agrupamiento del Árbol de Características (FASE 2)

Después de que se ha construido el árbol de características CF en el paso uno o fase uno descrita anteriormente, todas las regiones densas que se han ido formando en el proceso de pre-clúster se identifican y se almacenan en los nodos hojas del árbol CF.

Como el número de regiones densas es generalmente menor que el número de individuos que tiene la matriz de datos, y el resumen de los estadísticos almacenados en la fase de pre-clúster son suficientes para calcular la distancia mediante el criterio de agrupamiento seleccionado, se utiliza un algoritmo de agrupamiento aglomerativo jerárquico basado en la medida de la distancia del logaritmo de la función máximo verosímil.

Para determinar el número de clusters optimo, cada una de las soluciones del clúster se compara utilizando el criterio de agrupamiento seleccionado, Criterio Bayesiano de Schwarz (BIC) o el Criterio de Información de Akaike (AIC).

2.4. Cálculo para la Medida de la Distancia

La medida de distancia, nos permite poder calcular la similaridad o disimilaridad entre los miembros de un mismo clúster y también entre distintos clúster, para ello el Método Two-Step nos permite dos tipos de medida, en función del tipo de variables que tengamos en nuestra matriz de datos.

2.4.1. Distancia Euclidea

La distancia euclidea, es la distancia definida según una línea recta entre dos clusters, pero únicamente se puede utilizar cuando las variables son del tipo continuo.

$$d(x,y) = \sqrt{\sum_{j=1}^{J}(x_j - y_j)^2}$$

- $d(x,y)$ es el valor de la distancia entre las unidades de análisis “x” e “y”
- "x_j" e "y_j" son los valores que toman las variables desde 1 hasta J.

A los efectos de realizar el cálculo de esta métrica de distancia, las variables suelen estandarizarse previamente para que la escala de medición no distorsione el valor de las distancias. Con los valores de distancia de cada par de unidades de análisis se construye la llamada matriz de distancia, sobre la cual se aplican diversos algoritmos matemáticos para hallar grupos compuestos por unidades que, para ponerlo en términos sencillos, tengan mayor similitud entre sí.

2.4.2. Distancia Máxima Verosimilitud

La medida de verosimilitud, realiza una distribución de probabilidad entre variables, sabiendo que las variables continuas siguen una distribución Normal y las variables categóricas siguen una distribución Multinomial, y partiendo de la base de que se toman todas las variables como independientes, por lo que este tipo de distancia se debe de tomar con datos de tipo mixto (Bacher et al., 2001)

La distancia del logaritmo de máxima verosimilitud entre dos clusters se define como:

$$\boldsymbol{d(i,s) = \xi_i + \xi_s - \xi_{(i,s)}}$$

De forma que obtenemos la medida de distancia entre dos grupos o clusters como la disminución en el logaritmo de la función de máxima verosimilitud formada al juntarse ambos clúster.

El modelo asume que las variables continuas x_j $(j = 1,2,3,\ldots p)$ están dentro del cluster “i”, con distribución normal de media μ_{ij} y varianza σ_{ij}^2, y las variables

categóricas a_j dentro del clúster "i" con distribución multinomial con probabilidad π_{ijl}, donde (jl) es el índice para la categoría que ocupa la posición "l-esima" $(l = 1,2,3,\dots q)$ correspondiente a la variable a_j $(j = 1,2,3,\dots q)$

Siendo sus componentes:

$$\xi_i = -n_i\left(\sum_{j=1}^{p}\frac{1}{2}\log\left(\hat{\sigma}_{ij}^2 + \hat{\sigma}_j^2\right) - \sum_{j=1}^{q}\sum_{l=1}^{m_j}\hat{\pi}_{ijl}\log\left(\hat{\pi}_{ijl}\right)\right)$$

$$\xi_s = -n_s\left(\sum_{j=1}^{p}\frac{1}{2}\log\left(\hat{\sigma}_{sj}^2 + \hat{\sigma}_j^2\right) - \sum_{j=1}^{q}\sum_{l=1}^{m_j}\hat{\pi}_{sjl}\log\left(\hat{\pi}_{sjl}\right)\right)$$

$$\xi_{(i,s)} = -n_{(i,s)}\left(\sum_{j=1}^{p}\frac{1}{2}\log\left(\hat{\sigma}_{(i,s)j}^2 + \hat{\sigma}_j^2\right) - \sum_{j=1}^{q}\sum_{l=1}^{m_j}\hat{\pi}_{(i,s)jl}\log\left(\hat{\pi}_{(i,s)jl}\right)\right)$$

En donde ξ_v se puede interpretar como el tipo de dispersión que hay dentro del clúster, o lo que es lo mismo su varianza.

El parámetro ξ_v, se compone de dos partes:

- La primera parte está definida por "$-n_v\left(\sum_{j=1}^{p}\frac{1}{2}\log\left(\hat{\sigma}_{vj}^2 + \hat{\sigma}_j^2\right)\right)$", y mide la dispersión que tienen las variables continuas x_j dentro del clúster definido por "v", si se usa únicamente $\hat{\sigma}_{vj}^2$ entonces la distancia entre clusters $d(i,s)$ seria exactamente el decrecimiento del logaritmo de la función de verosimilitud después de haberse fusionado los clúster "i" con "s", y el termino $\hat{\sigma}_j^2$ se añade para evitar la posible situación cuando $(\hat{\sigma}_{vj}^2 = 0)$.

- La segunda parte está definida por "$n_v\left(\sum_{j=1}^{p}\sum_{l=1}^{m_j}\hat{\pi}_{vjl}\log\left(\hat{\pi}_{vjl}\right)\right)$", y mide la dispersión que tienen las variables categóricas dentro del clúster.

Siguiendo el proceso aglomerativo jerárquico en la fase 2, los clúster que tengan menor distancia entre ellos se fusionaran para formar un nuevo clúster.

2.5. Criterios de Agrupamiento

Los criterios estadísticos están basados en el principio de parsimonia, este principio conocido como “la navaja de Ockham”, dice que cuando dos supuestos en igualdad de condiciones tienen las mismas consecuencias, la teoría más simple tiene más probabilidad de ser la correcta que la más compleja, este principio debe su nombre al filósofo y teólogo franciscano de origen inglés Guillermo de Ockham quien vivió entre los años 1.280 y 1.349, y aunque no fue el primero en usar el principio si es verdad que fue el primero que lo puso por escrito.

Estos criterios pueden ser divididos en tres clases, Criterios de predicción, Criterios de información o verosimilitud y Criterios de maximización Bayesiana con distribución a posteriori de Probabilidad, siendo estos dos últimos los que se encuentran implementados en mayor medida en productos software en el mercado al permitir ajustar modelos mixtos de datos, en concreto se encuentra implementado en el producto SPSS.

La máxima verosimilitud nos permite poder seleccionar el modelo que sea capaz de realizar el mejor ajuste a los datos que estemos analizando y además no penalizan la complejidad que tengan estos, pero en mayor o menor medida, todos penalizan el logaritmo de la función de verosimilitud por el número de parámetros.

El modelo de agrupamiento en clúster Two-Step que estamos analizando en el presente Trabajo de Fin de Máster, utiliza los criterios de agrupamiento Bayesiano y el criterio de Akaike.

2.5.1. Criterio de AKAIKE (AIC)

Hirotugu Akaike (1974, 1978) es uno de los pioneros en el campo de la evaluación de modelos estadísticos, desarrollo un método para la comparación de modelos, llamado Criterio de Información de Akaike (AIC) en honor a su nombre, desarrollado en un principio para series temporales y que posteriormente fue propuesto para el análisis factorial.

El criterio combina la teoría de la máxima verosimilitud, información teórica y la entropía de información (Motulsky y Christopoulos, 2003), y está definido por la siguiente ecuación:

$$\mathrm{AIC} = -2 * \ln \mathrm{L}(\hat{\theta}) + 2\mathrm{K}$$

La estructura del AIC está compuesta por dos términos:

a) El primer término es la maximización del logaritmo de la función de verosimilitud $(-2 * \ln L(\hat{\theta}))$ como componente de la falta de ajuste del modelo y por consiguiente como medida de bondad de ajuste del modelo.

b) El segundo término, el número de parámetros estimados dentro del modelo $(2\mathrm{K})$ como componente de penalidad, ya que esta medida de penalidad es la compensación por el sesgo debido a la falta de ajuste del modelo cuando empleamos estimadores de máxima verosimilitud, creciente por tanto conforme aumenta el número de parámetros, de acuerdo al Principio de Parsimonia.

Se puede ver como el termino de penalización del criterio de Akaike $(2\mathrm{K})$**,** no depende del tamaño de la muestra y por consiguiente de la población considerada, de lo que se puede deducir que el mismo parámetro de penalización, será para tamaños muestrales pequeños o grandes, sin distinción alguna, lo que nos lleva a pensar que el Criterio de Akaike no es un estimador consistente del número adecuado de factores comunes.

Este criterio tiene en cuenta los cambios en la bondad de ajuste del modelo y también las diferencias en el número de parámetros entre dos modelos, siendo los mejores modelos los que tengan un valor de AIC más bajo, y el modelo que mejor explica los datos con el mínimo número de parámetros es el que presenta más bajo valor de AIC (Molinero 2003 y Balzarini et al. 2005).

Cuando dos valores de AIC están muy próximos, para poder escoger el mejor modelo se realiza en función del cálculo de probabilidad, también llamado pesos de Akaike, y la probabilidad relativa o relación de evidencia entre los dos modelos, según las siguientes ecuaciones (Posada et al., 2007):

$$PROBABILIDAD = \frac{\boldsymbol{e^{-0.5^{\Delta}}}}{\mathbf{1} + \boldsymbol{e^{-0.5^{\Delta}}}}$$

$$PROBABILIDAD\ \ RELATIVA = \frac{P\ (Modelo\ \ "1"\ \ sea\ correcto)}{P\ (Modelo\ \ "2"\ \ sea\ correcto)} = \frac{1}{e^{-0.5^{\Delta}}}$$

Donde "Δ" es la diferencia entre valores de AIC ((Motulsky y Christopoulos, 2003)

2.5.2. Criterio de SCHWARTZ (BIC)

La estadística bayesiana surge del famoso teorema de Bayes, el cual permite en el caso de conocer la probabilidad de que ocurra un cierto suceso, poder modificar su valor cuando se dispone de información nueva (Molinero 2002).

Para poder mejorar la inconsistencia del criterio de Akaike (AIC), tanto Akaike (1978) como Schwartz (1978) presentaron un criterio de selección de modelos desde un enfoque bayesiano.

En este sentido Schwartz estableció que la solución de Bayes consistía en seleccionar el modelo con una alta probabilidad a posteriori, por lo que está probabilidad a posterior para grandes volúmenes de datos podía aproximarse mediante el desarrollo de Taylor.

El criterio está definido por la siguiente ecuación:

$$\mathrm{BIC} = (-2 * \ln \mathrm{L}(\hat{\theta})) + (\ln(\mathrm{n}) * \mathrm{K})$$

La estructura del BIC está compuesta por dos términos:

a) El primer término de la estructura del BIC viene dato por la maximización del logaritmo de la función de verosimilitud, igual que en el criterio de Akaike $(-2 * \ln \mathrm{L}(\hat{\theta}))$

b) El segundo término como componente de penalidad, está definido por el número de parámetros estimados dentro del modelo y multiplicado por el logaritmo neperiano del tamaño de la muestra $(\ln(\mathrm{n}) * \mathrm{K})$

El criterio para elegir el mejor modelo sigue el mismo procedimiento que en el criterio de Akaike, es decir el modelo con el valor BIC más bajo se considera el mejor en explicar los datos del análisis con el mínimo número de parámetros.

2.6. Selección del Número de Clusters

Cuando no se especifica el número de clúster que se quiere y por tanto se le deja al algoritmo que decida de forma automática, es como añadir un paso adicional más al algoritmo para que pueda determinar el número de clúster más adecuado según los datos que tiene que analizar, la estrategia más básica seria calcular criterios estadísticos para modelos con 1 clúster, con 2 clúster, con 3 y así sucesivamente, eligiendo el más óptimo según los resultados obtenidos.

El método en dos etapas, que estamos analizando, utiliza un criterio diferente en cada una de ellas, mientras que en la primera fase o etapa consiste en detectar una estimación del número de clúster optimo en función del Criterio Bayesiano (BIC) o del Criterio de Akaike (AIC) previamente seleccionado.

En la fase o etapa dos se usa el ratio de cambio en la distancia de los clúster, que viene definido por:

$$R_{(k)} = d_{k-1}/d_k$$

En donde d_{k-1} es la distancia entre el clúster "k" y el clúster "k-1", así mismo la distancia d_k se calcula como $d_k = l_{k-1} - l_k$

Y según el criterio que se use, Criterio Bayesiano o Criterio de Akaike, tendremos:

$$l_v = (r_v * \log_n - BIC_v)/2$$

o bien

$$l_v = (2r_v - AIC_v)/2$$

El número de clúster se obtiene como la solución donde exista un gran salto en el ratio, que viene calculado como el cociente entre los dos valores mayores obtenidos en la primera fase:

$$R_{(k1)} / R_{(k2)}$$

Si el ratio de cambio es mayor que el umbral definido, entonces el número de clusters será igual al valor de distancia $k1$, en caso contrario el número de clusters

será igual al mayor valor existente entre $k1$ y $k2$, y por tanto el número de clúster se obtiene.

Las dos formas de selección para el número de clusters que tiene el algoritmo, son las siguientes:

1.- Selección Automática

El procedimiento determina automáticamente el número óptimo de clúster necesario para realizar el agrupamiento de datos propuesto, utilizando para ello el Criterio de Agrupamiento especificado para tal fin, especificando a su vez el número máximo de clúster que debe de tener en cuenta el algoritmo.

2.- Selección Manual

El algoritmo es capaz también de permitir que sea el propio usuario el que de forma manual introduzca el número de clúster que desee para la realización del agrupamiento de los datos, impidiendo de esta manera que el algoritmo lo calcule de forma automática.

2.7. Tratamiento de Valores Atípicos

El método Two-Step permite tratar los valores atípicos de forma especial durante el proceso de agrupamiento de los datos, de los individuos existentes, si se llena el clúster de características (CF), de forma que el Clúster no aceptará ningún individuo más dentro del mismo si se considera lleno y si considera que tampoco hay ninguna hoja que se pueda dividir.

Se considera un valor atípico o outlier, a aquellos valores que no encajan bien en ningún clúster, en este sentido, podemos seleccionar si se quiere tratamiento del ruido o no, en el caso que se seleccione el tratamiento del ruido y el árbol se llenara, se hará volver a crecer después de colocar los individuos existentes en hojas que estén poco llenas en una hoja denominada de ruido.

Se considera que una hoja esta poco llena si contiene un número de individuos inferior a un determinado porcentaje de individuos del tamaño máximo existente en un nodo hoja del árbol de características CF, por lo que después de volver a realizar el proceso y hacer que crezca de nuevo el árbol, los individuos atípicos se colocarán en

el árbol CF en el caso de que fuese posible ubicarlos una última vez, pero si no fuese posible entonces los valores atípicos se descartarán.

En el supuesto de que no se quiera tratamiento de ruido, y el árbol CF se llenara, entonces se volverá hacer crecer el árbol CF utilizando un umbral del cambio en distancia mayor, y después de este nuevo agrupamiento, los valores correspondientes a los individuos que no se hayan asignado a un clúster se considerarán valores atípicos, a los cuales se les asignará el valor de "-1" y no se incluirán en el recuento del número de clusters.

2.8. Importancia del Predictor en el Agrupamiento

En base a la importancia que cada variable tiene en la formación de un clúster, tanto dentro como entre clusters, se distinguen dos situaciones, estando definido la importancia asignada por:

$$VI_i = \frac{-\log_{10}(p\_valor_i)}{\max_{j\in\Omega}(-\log_{10}(p\_valor_j))}$$

Donde Ω denota la matriz de datos, el p-valor se calcula tal como se describe a continuación, y si el $(p\_valor)_i = 0$ entonces $(p\_valor)_i$ es el valor mínimo doble.

a) <u>ENTRE Clusters</u>

Para variables categóricas: El p-valor está basado en una Chi-Cuadrado

$$p\_valor = Prob(\chi_d^2 > \chi^2)$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(N_{ij} - \hat{N}_{ij})^2}{\hat{N}_{ij}}$$

$$\hat{N}_{ij} = \frac{N_{i.}\ N_{.j}}{N}$$

Sabiendo que los grados de libertad serán:$(I' - 1, J' - 1)$:

$N = 0$	→	La importancia será indefinida o desconocida
$N_{i.} = 0$	→	I' se calcula como (I' = I -1)

$N_{.j} = 0$ → J' se calcula como (J' = J – 1)

si $J' \leq 1$ o $I' \leq 1$ → La importancia será indefinida o desconocida

Para variables continuas: El p-valor está basado en una F de Snedecor:

$$p\_valor = Prob\{F(J-1, N-J) > F)\}$$

$$F = \frac{\sum_{j=1}^{J}\left(\bar{x}_j - \bar{\bar{x}}\right)^2/(J-1)}{\sum_{j=1}^{J}(N_j - 1)s_j/(N-J)}$$

Sabiendo que los grados de libertad son $(J' - 1, N - J')$:

$N = 0$ → La importancia será indefinida o desconocida

$N_j = 0$ → J' se calcula como (J' = J – 1)

si $J' \leq 1$ o $N \leq J'$ → La importancia será indefinida o desconocida

Si el denominador de la formula para la F de Snedecor es cero, entonces el p-valor se tomará como igual a 1

b) <u>DENTRO del Clúster</u>

Para variables categóricas El p-valor está basado de forma que la proporción de individuos en el clúster <j> es la misma que la proporción global, mediante la Chi-cuadrado.

$$\chi^2 = \sum_{i=1}^{I}\frac{(N_{ij} - N_j p_i)^2}{N_j p_i}$$

Sabiendo que los grados de libertad serán: $d = (I' - 1)$:

$N_j = 0$ → La importancia será indefinida o desconocida

$p_i = 0$ → I' se calcula como (I' = I – 1)

si $I' \leq 1$ → La importancia será indefinida o desconocida

Para variables continuas: El p-valor está basada en que la media del clúster <j> es la misma que la media global, mediante una T-Student:

$$t = \frac{(\overline{x}_j - \overline{\overline{x}})}{s_j/\sqrt{N_j}}$$

Sabiendo que: los grados de libertad serán: $d = (N_j - 1)$

$si\ N_j \leq 1\ o\ s_j = 0$ → La importancia será indefinida o desconocida

Si el denominador de la formula para la t-Student es cero, entonces el p-valor se tomará como igual a 1, y se calculará como:

$$p\_valor = 1 - Prob\{|T(d)| \leq |t|\}$$

2.9. Ajuste del Árbol de Características (CF)

Dentro de los criterios u opciones para ajustar las características del clúster (CF), se encuentra fundamentalmente la definición del árbol que se tiene que construir mediante la primera etapa del algoritmo, y que viene especificado mediante los siguientes valores por defecto, pudiéndose modificar manualmente.

a) <u>Umbral del cambio en la distancia inicial</u>

Se refiere al que se utiliza para hacer crecer al árbol CF, de forma que si se ha insertado una hoja concreta en el árbol CF que va a producir una densidad inferior al umbral marcado, entonces no se producirá la división de la hoja, de igual forma si la densidad supera el umbral marcado, entonces se realizará la división de la hoja.

b) <u>Umbral del cambio en la distancia inicial</u>

Es el número máximo de nodos hijo que puede tener un nodo hoja

c) <u>Máxima profundidad del árbol</u>

Se refiere al número máximo de niveles que puede tener un árbol CF

d) Máximo número posible de nodos

Como su nombre indica, se refiere al número máximo de nodos del árbol CF que puede generar el algoritmo, ya que un árbol CF demasiado grande, puede consumir los recursos del sistema y afectar de manera negativa al rendimiento del algoritmo, por lo que se necesitan 16 bytes como mínimo para cada nodo.

El algoritmo sigue la siguiente función:

$$\frac{(b^{d+1}-1)}{(b-1)}$$

En donde:

"b" es el número máximo de ramas

"d" es la profundidad máxima del árbol

2.10. Representación Gráfica

El procedimiento muestra un gráfico resumen, en donde figura el tipo de algoritmo utilizado, el número de variables categóricas y continuas introducidas en el análisis, el número de clusters que ha definido como solución a los datos introducidos.

También muestra la calidad que tienen estos clusters, permitiendo mediante esta gráfica comprobar si la calidad de agrupamiento realizada es insuficiente o no, en cuyo caso se podrían realizar los ajustes necesarios en el modelo para de esta forma producir mejores resultados que los obtenidos.

El gráfico de calidad, nos muestra mediante diferentes tipos de colores, y una barra del color alcanzado, la cohesión y separación de los clúster, indicándonos si los resultados son pobres, suficientes o buenos (IBM® SPSS® Statistics, 2015).

Fig. 2.6 Resumen gráfico del programa SPSS

La interpretación y validación gráfica de este tipo de análisis de clúster, está basada en el trabajo desarrollado por Kaufman y Rousseeuw (1987) sobre la interpretación de estructuras de grupos, según esta valoración, un resultado "bueno" nos indica que los datos reflejan una sólida evidencia de que existe una estructura de clúster, un resultado "suficiente" significa que esa evidencia es débil, y un resultado "pobre" nos indica que no hay evidencias claras de agrupamiento.

Para poder construir el gráfico Silueta (Silhouette), se necesitan dos cosas, por un lado la partición obtenida mediante el algoritmo y la colección de las proximidades o similiaridades entre los individuos u objetos, en donde para cada objeto "i" se introduce un valor llamado "$s(i)$", y con todos ellos se forma el gráfico.

Si llamamos "i" al objeto que ha sido asignado al cluster "A", podemos calcular la media de las disimilaridades entre este objeto y el resto de objetos dentro del cluster, de forma que tendremos:

- $a(i)$ = Media de las disimilaridades de (i) con el resto de objetos incluidos en el clúster A
- $d(i, C)$ = Media de las disimilaridades de (i) con todos los objetos incluidos en el clúster C, sabiendo que ambos clúster A y B son distintos entre sí.
- $b(i) = minimo\ \ d(i, C)$

Una vez que se calculan todas las distancias entre (i) y todos los clúster formados, se toma el menor de ellos, si estimamos por ejemplo que en nuestro gráfico el menor es el constituido con el clúster B, tendremos:

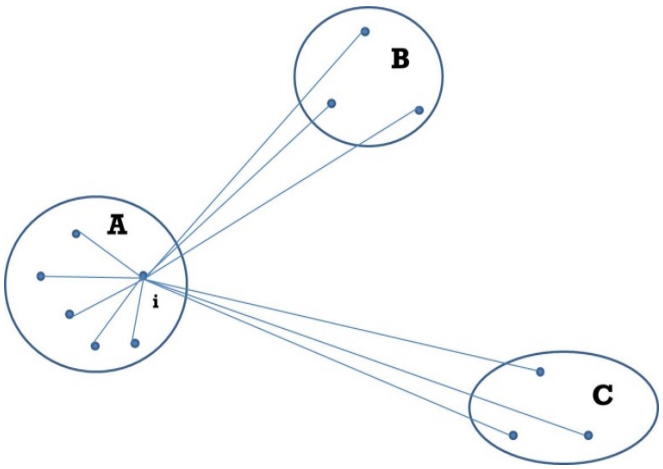


Fig. 2.7 Ejemplo gráfico del cálculo de las similaridades entre clusters

Por lo tanto el valor $s(i)$ se obtiene como combinación entre los valores estimados anteriormente, y tendremos lo siguiente:

$$s(i) = \begin{cases} 1 - {}^{a_i}/_{b_i} & si \quad a_i < b_i \\ 0 & si \quad a_i = b_i \\ {}^{b_i}/_{a_i} - 1 & si \quad a_i > b_i \end{cases}$$

$$s(i) = \frac{b_1 - a_i}{\max\ \{a(i),\ b(i)\}} \qquad (-1 \leq s(i) \leq 1)$$

Un valor de "1" podría implicar que todos los individuos están situados directamente en los centros de sus grupos, por el contrario un valor de "-1" significaría que todos los individuos se encuentran en los centros de los grupos de otro clúster, esta distribución se muestra por defecto como un diagrama de sectores, en donde se especifica la frecuencia que tiene cada clúster.

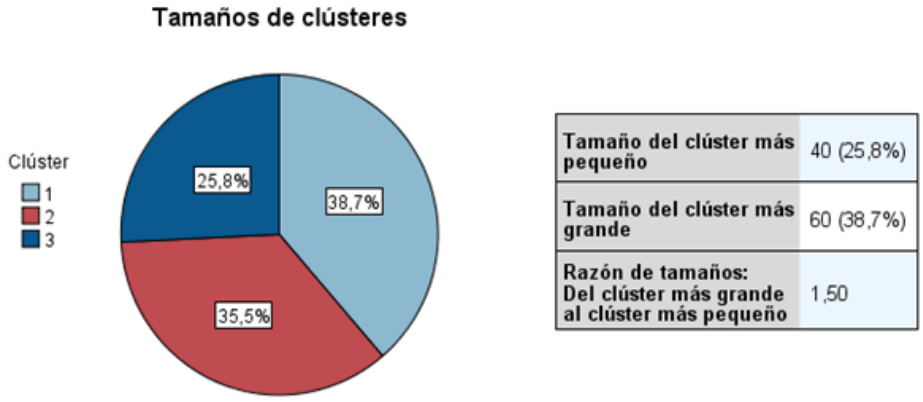


Fig. 2.8 Gráfica del programa SPSS del número de clusters creado

La importancia del predictor o del conjunto de variables que se han agrupado, indica el orden que han tomado las variables de la matriz de datos para diferenciar los diferentes clusers que se han formado, tanto para variables categóricas como para variables numéricas, de manera que cuanto mayor es la medida de importancia que tiene la variable, menos probable será la variación para una variable entre clusters debido al azar y mas probable será debido a alguna diferencia subyacente de las propias variables.

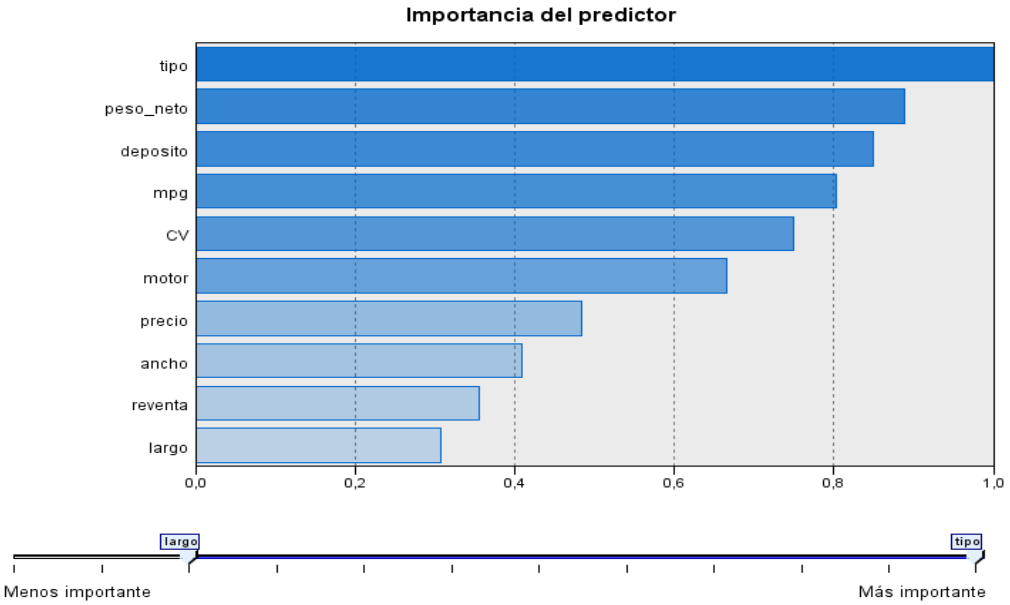
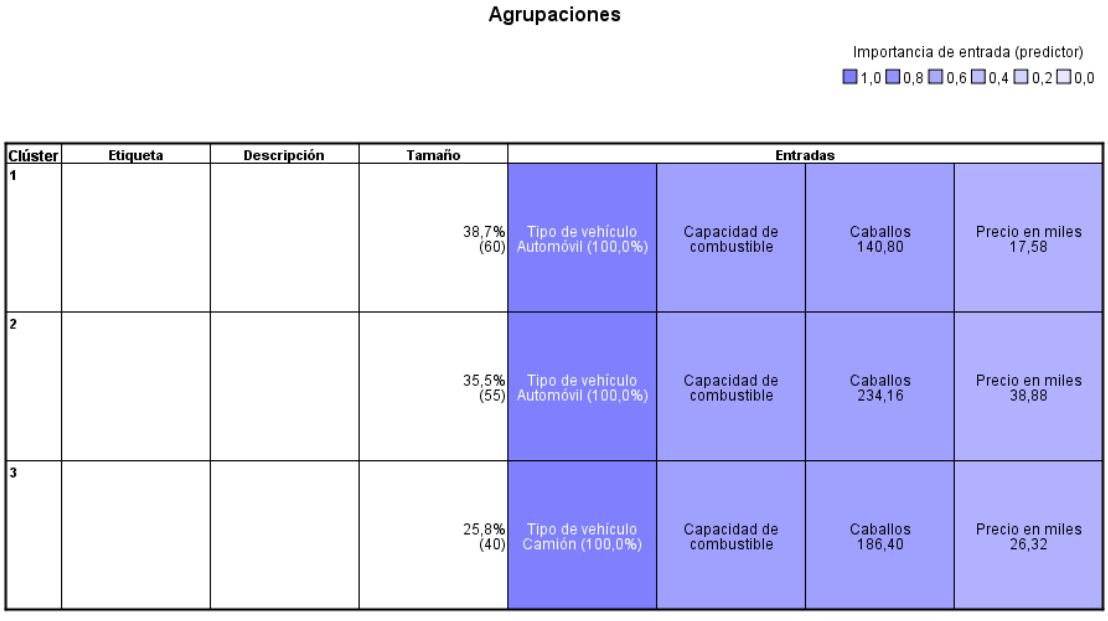


Fig. 2.9 Gráfica del programa SPSS de la importancia de las variables en estudio

En el siguiente gráfico se puede observar el agrupamiento que ha tenido lugar y la importancia según el predictor o variable seleccionada para el análisis de clúster, tanto continua como categórica, así como el orden de importancia por clúster.



Clúster	Etiqueta	Descripción	Tamaño	Entradas			
1			38,7% (60)	Tipo de vehículo Automóvil (100,0%)	Capacidad de combustible	Caballos 140,80	Precio en miles 17,58
2			35,5% (55)	Tipo de vehículo Automóvil (100,0%)	Capacidad de combustible	Caballos 234,16	Precio en miles 38,88
3			25,8% (40)	Tipo de vehículo Camión (100,0%)	Capacidad de combustible	Caballos 186,40	Precio en miles 26,32

Fig. 2.10 Gráfica del programa SPSS de la agrupación de las variables en estudio

En donde se pueden ver los diferentes campos incluidos en el gráfico, a saber:

- ***Clúster***: Indica el número de clúster creado por el algoritmo
- ***Etiqueta***: Indica el nombre que queremos dar al clúster, por defecto está en blanco, pero se puede introducir la denominación manualmente.
- ***Descripción***: Indica lo que contiene el clúster, y también por defecto está en blanco, aunque igualmente se puede introducir el contenido manualmente.
- ***Tamaño***: Indica el recuento de registros que están incluidos dentro del clúster y el porcentaje que representa respecto del total del clúster.
- ***Entradas***: Indica los predictores, o variables incluidas en el análisis de clúster, ordenadas por orden de importancia global entre ellas, esta importancia se indica mediante el color sombreado de la casilla, siendo más oscuro cuanto más importante sea la característica. La clasificación también se puede realizar por la importancia de cada variable dentro del clúster en el que se encuentra, por tamaño e incluso por nombre.

También se puede observar la distribución de los clúster en función de las distribuciones absolutas o relativas.

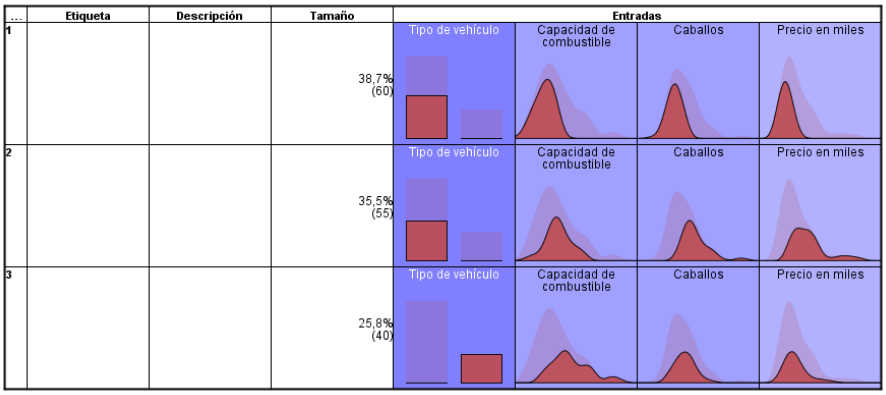


Fig. 2.11 Gráfica del programa SPSS de las distribuciones de las variables

En donde se puede observar que entre los clusters 1 y 2 existe un cierto solapamiento en las características de tipo de vehículo, que es la variable categórica, pero como se diferencian respecto al resto de variables, que en este caso son las variables numéricas, mientras que el clúster 3 es totalmente diferente en cuanto a la

variable categórica, pero existiendo un cierto parecido con el clúster 2 en cuanto a la variable “potencia el vehículo”

Otra forma de comparar los clúster, es mediante el siguiente gráfico, seleccionando aquellos clúster a comparar, donde las variables categóricas se representan cada una de ellos con un color diferente y tamaño diferente, indicando el orden global que tiene cada una de ellas, y las variables continuas se representan mediante un gráfico de caja donde se puede ver la situación de las medianas de cada clúster así como los cuantiles.

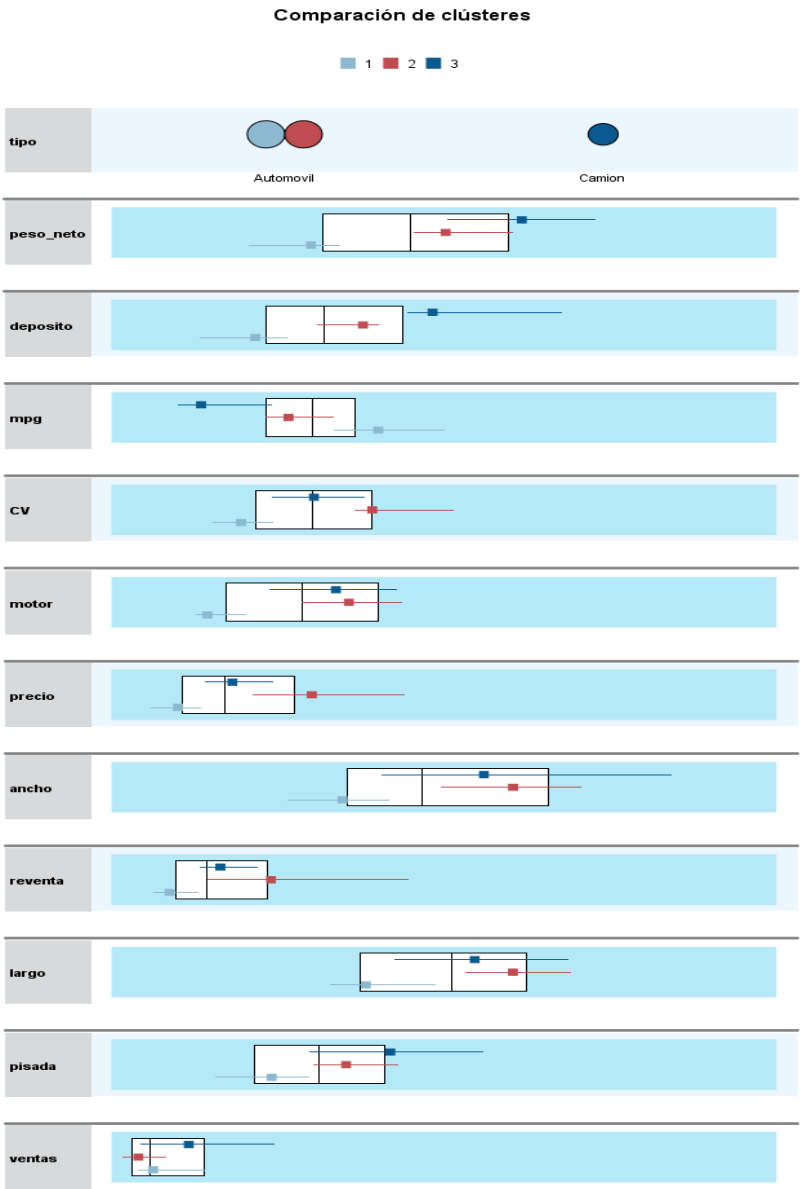


Fig. 2.12 Gráfica del programa SPSS de la comparación de los clusters

2.11. Aplicaciones del Método TWO-STEP

A continuación se incluyen un conjunto de referencias, sin pretender ser exhaustivo, en donde el método Two-Step se está aplicando para realizar análisis de clúster, en diferentes campos las cuales ponen de manifiesto su vigente actualidad.

REFERENCIA	CAMPO DE APLICACIÓN
Alho y Abreu (2015)	Análisis Urbanístico
Akter y Ahmed (2015)	Potencial sobre la recolección del agua de lluvia en regiones de Bangladesh
Altas, Kubas y Sezen (2013)	Análisis de la 'Sensibilidad Ambiental' de las Empresas
Andersen, Silva y Levy (2013)	Análisis de la importancia en la colaboración universidad/empresa (I+D+i)
Arrighetti y Ninni (2014)	Economía Industrial Aplicada
Budeva y Mullen (2014)	Segmentación del Mercado Internacional
Cardona (2014)	Grado de innovación del sector financiero español
Chin y Choi (2015)	Análisis de los factores de suicidio en adolescentes coreanos
Dhillon y Godfrey (2013)	Monitorización de protocolos de salud pública
Dutt et al. (2015)	Data Mining sobre datos Educacionales
DuBois et al. (2015)	Análisis sobre las decisiones en el ámbito de la Investigación Profesional
Francisco (2012)	Salud Mental
Giuliano et al. (2014)	Análisis de rutas de transporte
Griffin et al. (2014)	Medicina Preventiva
Gunten et al. (2014)	Análisis de datos administrativos
Harvey (2014)	Urbanismo
Hauser et al. (2015)	Análisis sobre el Impacto del clima sobre la biología marina
Nadotti y Constantin (2014)	Análisis de Riesgos
Nikolaj et al. (2015)	Análisis sobre la Dirección de Negocio
Opitz y Hofmann (2015)	Aprendizaje artificial
Satish & Bharadhwaj (2010)	Marketing en el sector automovilístico
Sigmund et al. (2014)	Diagnóstico obesidad en el ámbito escolar
Stanton (2015)	Pautas de los profesores, aplicación en la intervención y sus efectos sobre los resultados de los estudiantes

Stranak et al. (2014)	Diagnóstico hipotensión en recién nacidos
Xiong et al. (2014)	Análisis de Patrones de Desarrollo Global
Yan-yan et al. (2011)	Evaluación factores de riesgo propagación de incendios en grandes edificios
Yu y Chan (2012)	Evaluación del desempeño operativo de un sistema de refrigeración en un edificio institucional

Fig. 2.13 Tabla de Referencias en distintos campos de la Ciencia. Método TWO-STEP

Capítulo III

AGRUPAMIENTO EN CLUSTER

HJ-BIPLOT *vs* TWO-STEP

3.1. Introducción

Los métodos Biplot (Gabriel, 1971) son una representación gráfica de datos multivariantes de una matriz de datos formada por “n” individuos con “p” variables, de forma que en las filas tengamos a los individuos y en las columnas las variables o características de cada individuo, que ordenados en una matriz, podría ser como la siguiente:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots\dots\dots & x_{1j} & \dots\dots & x_{1p} \\ x_{21} & x_{22} & \dots\dots\dots & x_{2j} & \dots\dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{i1} & x_{i2} & \dots\dots\dots & x_{ij} & \dots\dots & x_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \dots\dots\dots & x_{nj} & \dots\dots & x_{np} \end{pmatrix}$$

El Biplot aproxima la distribución de una muestra mutivariante en un espacio de dimensión reducida, normalmente de dimensión dos, y superpone sobre la misma representaciones de las variables sobre las que se mide la muestra (Gower, 1966).

Con un enfoque diferente podemos encontrar a (Gower, 1992), (Gower y Harding, 1988) y (Gower y Hand, 1996), en donde se propone primero obtener una ordenación de los individuos de la matriz de datos utilizando métodos de escalamiento multidimensional, y posteriormente superponer en los biplots las variables, de manera que pueda ser interpretada de forma conjunta, este tipo de enfoque se puede relacionar con la forma factorial clásica de la escuela francesa de análisis de datos (Frutos, 2014).

El método Biplot está basado en la descomposición en valores y vectores singulares de la matriz **X** (Greenacre, 1984) es decir:

$$X = U_{nxr}\Sigma_r V_{rxp}^t$$

U_{nxr} : Es la matriz cuyas columnas contienen los vectores propios de XX^t

V_{rxp} : Es la matriz cuyas columnas contienen los vectores propios de X^tX

Σ_r : Es la matriz diagonal de los valores propios de X, positivos o cero.

Verificando estas matrices, lo siguiente:

$$U^t\,U = V^t\,V = I\ (matriz\ identidad)$$

Gabriel introduce el término Biplot usando la siguiente definición (Martins, 2003):

> *"Toda matriz de rango dos puede ser representada gráficamente como un biplot que consiste en un vector para cada fila y en un vector para cada columna, elegidos de modo que cada elemento de la matriz, sea exactamente el producto interno de los vectores que corresponden a esa fila y a esa columna. Si una matriz tiene rango superior a 2, se puede representar esa matriz, de modo aproximado, por un biplot de una matriz de rango 2 – que es una aproximación de la matriz original".*

Esto significa por tanto que el termino biplot, está asociado al concepto de descomposición de una matriz de datos $\mathbf{X_{nxp}}$ con "m" filas de los individuos y "p" variables de cada individuo

$$X_{nxp} = [x_{ij}] = [a_i^t \; b_j] = [< a_i \; b_j >] = [|a_i||b_j| \; cos\theta_{ij}]$$

En donde los vectores a_i son los marcadores de los individuos o filas de la matriz y los vectores b_j son los marcadores designados de las variables o columnas de la matriz $\mathbf{X_{nxp}}$ y θ_{ij} es el ángulo formado por el marcador de la fila i con el marcador de la columna j.

Las dos factorizaciones Biplot mas importantes propuestas por Gabriel (Gabriel, 1971) fueron denominadas: GH-Biplot y JK-Biplot, aunque posteriormente Greenacre (Grenacre, 1984) introdujo la terminología CMP (Column Metric Preserving) y RMP (Row Metric Preserving) respectivamente.

Galindo propone una alternativa para obtener simultáneamente altas calidades de representación tanto para filas como para columnas, proponiendo el método HJ-BIPLOT o también llamado RCMP (Row Column Metric Preserving), siguiendo la terminología de Greenacre (Galindo, 1986).

El HJ-BIPLOT es una representación gráfica multivariante de marcadores fila y marcadores columna, de forma que se pueden superponer en el mismo sistema de referencia y con la máxima calidad de representación (Galindo, 1986), (Galindo y Cuadras 1986), y partiendo de la Descomposición de Valores Singulares (SVD), ya que al tener tanto filas como columnas idéntica bondad de ajuste, es posible el

interpretar además de las filas y las columnas, las relaciones entre ellas, es decir las relaciones entre individuos y variables.

$$X = U_{nxr}\,\Sigma_r\,V^t_{rxp}$$

Lo que permite representar las coordenadas de las filas y columnas con referencia a los mismos ejes cartesianos, por lo tanto los espacios para representar las filas y las columnas, pueden ser superpuestos para obtener una representación sobre un mismo sistema cartesiano, y que será el sistema de ejes factoriales, cuyo origen coincide con el punto de equilibrio de las nubes, proporcionando también el HJ-Biplot las mejores representaciones β-Baricentricas en el sentido propuesto por Lebart (Lebart et al., 1983).

Algunas de las aplicaciones de este método se pueden encontrar en, (Pedraz y Galindo, 1986), (Galindo et al., 1987), (Galindo et al., 1993), (Perez-Mellado y Galindo, 1986).

3.2. Tipo de Variables

El algoritmo trata únicamente variables de tipo numérico, pudiendo tomar cualquier valor dentro de la escala de los números reales, para las variables de tipo categórico, tienen que ser previamente cambiadas, es decir es necesario realizar un paso previo para cambiar el tipo de codificación de este tipo de variables a formato nominal, ordinal o binario.

Por tanto las variables de tipo Nominal y Ordinal, deben de ser codificadas como números enteros, empezando en el número 1 y en orden ascendente, sin espacios vacíos entre un valor y otro, por otro lado las variables de tipo Binario deben de ser codificadas con el valor 0 para ausencia y el valor 1 para presencia, posteriormente y dentro del software para el método de agrupamiento del HJ-BIPLOT "***MULTBIPLOT***" (Vicente-Villardon, J.L., 2010). *MULTBIPLOT: A package for Multivariate Analysis using Biplots*. Departamento de Estadística. Universidad de Salamanca; **http://biplot.usal.es/ClassicalBiplot/index.html**), se les puede asignar diferentes colores para su representación gráfica.

Aquellas variables que no tienen valor asignado, el algoritmo lo marca como no numérico (NaN) y es remplazado por la media o la mediana, en función del valor que se asigne manualmente en el proceso.

Si los datos no tienen variables nulas, entonces elegiremos el método que viene por defecto “Descomposición en Valores y vectores singulares”, pero si existen valores que no están incluidos en nuestra matriz de datos, entonces tendremos que elegir “Regresión Lineal Alternada”

3.3. Algoritmo del Clúster HJ-BIPLOT

El algoritmo que se utiliza para la realización del agrupamiento o clustering en el modelo HJ-BIPLOT, es una extensión del Criterio de Inercia o de la varianza propuesto por Benzecri (Benzecri, 1973), el cual está basado en la representación del HJ-BIPLOT en vez de un análisis de correspondencias (Vicente-Tavera et al., 1998), y se encuentra ampliado en la Tesis doctoral del Dr. Vicente-Tavera (Vicente-Tavera, 1992).

Es necesario ejecutar como pasos previos al algoritmo, en primer lugar modificar nuestra matriz de datos originales de tipo mixto, por una matriz de datos numéricos, y posteriormente al cargar la matriz modificada dentro del software MULTBIPLOT (Vicente-Villardon, J.L., 2010). “*MULTBIPLOT: A package for Multivariate Analysis using Biplots*. Departamento de Estadística. Universidad de Salamanca; **http://biplot.usal.es/ClassicalBiplot/index.html)**”, asignar manualmente aquellas variables que sean de tipo nominal, ordinal o binario, y por último ejecutar el HJ-Biplot mediante la estimación de Valores y Vectores singulares, para obtener la matriz de coordenadas del método HJ-Biplot, y estandarizando por columnas.

Tomando como base la matriz obtenida del paso anterior, de las coordenadas del HJ- BIPLOT de los individuos sobre los primeros “q” factores, se realizan los siguientes pasos (Vicente-Tavera et al., 1994):

1) Se construye una matriz inicial de distancias entre los individuos, de la forma siguiente:

Inercia Dentro: $I(S_i^k) = \sum_{j=1}^{n_i} \|S_{ij}^k - S_{i\bullet}^k\|^2 = \sum_{j=1}^{n_i} \sum_{\alpha} [F_\alpha(S_{ij}^k) - F_\alpha(S_{i\bullet}^k)]^2$

$$\mathbf{d^2(i_1, i_2) = \sum_{\alpha=1}^{q} [F_\alpha(S_{i_1}) - F_\alpha(O_{(i_1,i_2)})]^2 + \sum_{\alpha=1}^{q} [F_\alpha(S_{i_2}) - F_\alpha(O_{(i_1,i_2)})]^2}$$

2) Se agrupan aquellos individuos entre los cuales existe una distancia mínima.

3) Se vuelven a calcular en cada etapa, las distancias existentes entre el clúster formado y el resto, por medio de:

Inercia Entre: $\mathrm{I_E(P_k)} = \sum_{i=1}^{k} \mathrm{n_i} \sum_{\alpha} \left[\mathrm{F_\alpha}\left(\mathrm{S_{i\bullet}^k}\right) - \mathrm{F_\alpha(O)}\right]^2$

$$\mathbf{d^2(S_t, S_v) = I_E(\{S_t, S_v\})}$$

4) Se repite el procedimiento desde el paso 2 hasta que se hayan agrupado todos los individuos en los clúster correspondientes.

Sabiendo que:

q	→	Numero de factores retenidos en la representación HJ-Biplot
P^k	→	Partición del conjunto a clasificar, formada por "k" clusters
n_i	→	Número de clases elementales "s_j" que hay en la clase "S_{ij}^k"
S_i^k	→	Clase "i-esima" de la partición P^k con (i = 1,2,3…, k)
S_{ij}^k	→	Clase "j-esima" perteneciente a la clase S_i^k
$S_{i\bullet}^k$	→	Centro de gravedad de la clase S_i^k
$F_\alpha(S_{ij}^k)$	→	Coordenadas para la clase S_{ij}^k en la representación HJ-Biplot
$F_\alpha(S_{i\bullet}^k)$	→	Coordenadas para $S_{i\bullet}^k$ en la representación HJ-Biplot con "α" factores
O	→	Centro de gravedad donde hemos retenido "q" factores
$F_\alpha(O)$	→	Coordenadas para "α" factores de O

3.4. Criterio de Agrupamiento

El criterio de clasificación que incorpora este método, es el criterio de clasificación ascendente jerárquico y está basado en la representación del HJ-BIPLOT, (Vicente-Tavera et al., 1994).

Dentro de este criterio de agrupamiento jerárquico aglomerativo, se aplica la distancia euclidea como medida de distancia, teniendo que seleccionar de forma manual tanto el número de clusters que se quieren conseguir como cualquiera de los métodos incluidos en la agrupación jerárquica aglomerativa, desarrollada ampliamente en el Capítulo I del presente Trabajo Fin de Master.

El paso de una partición $\boldsymbol{P_k}$ formada por "k" clases a otra partición $\boldsymbol{P_{k-1}}$ en la que hemos agrupado las clases $\boldsymbol{S_t^k}$ y $\boldsymbol{S_v^k}$ y formada a clase $\left(\boldsymbol{S_t^k} \cup \boldsymbol{S_v^k}\right)$, se hace de forma que:

$$\boldsymbol{P_{k-1}} = \left[\boldsymbol{P_k} - \left\{\boldsymbol{S_t^k}, \boldsymbol{S_v^k}\right\}\right] \cup \left\{\boldsymbol{S_t^k} \cup \boldsymbol{S_v^k}\right\}$$

Si aplicamos la definición de Inercia Dentro sobre la partición "k-1", comprobamos que la Inercia Dentro de la partición "k-1" depende únicamente de la Inercia Entre las clases que hemos añadido y no de la Inercia del resto de clases que existen en el conjunto definido, por lo que el criterio que se va a seguir será el de minimizar esta inercia entre las clases que estamos agrupando:

$$\boldsymbol{I_D(P_{k-1})} = \boldsymbol{I_D(P_k)} + \boldsymbol{I_E}\left(\left\{\boldsymbol{S_t^k}, \boldsymbol{S_v^k}\right\}\right)$$

Minimizar la inercia entre las clases $\left\{\boldsymbol{S_t^k}, \boldsymbol{S_v^k}\right\}$ es lo mismo que minimizar la siguiente relación, tomada de (Vicente-Tavera et al., 1994).

$$\boldsymbol{I_E}\left(\left\{\boldsymbol{S_t^k}, \boldsymbol{S_v^k}\right\}\right) = \boldsymbol{n_t} \sum_{\boldsymbol{\alpha=1}}^{\boldsymbol{q}} \left[\boldsymbol{F_\alpha}\left(\boldsymbol{S_{t\bullet}^k}\right) - \boldsymbol{F_\alpha(O_{tv})}\right]^2 + \boldsymbol{n_v} \sum_{\boldsymbol{\alpha=1}}^{\boldsymbol{q}} \left[\boldsymbol{F_\alpha}\left(\boldsymbol{S_{v\bullet}^k}\right) - \boldsymbol{F_\alpha(O_{tv})}\right]^2$$

Donde $\boldsymbol{F_\alpha(O_{tv})}$ son las coordenadas del centro de gravedad de la clase agrupada, y $\boldsymbol{n_t}$ y $\boldsymbol{n_v}$ son el número de elementos de las clases y "**q**" el número de factores retenidos para la clasificación en la representación del HJ-Biplot.

Este criterio tiene dos puntos a favor importantes (Vicente-Tavera et al., 1994), a saber:

1. Formar particiones que contienen ***clases homogéneas***, ya que la agregación o inclusión se hace de forma que la inercia alrededor del centro de la clase sea lo más pequeña posible, siendo de esta manera para todas las clases de la partición, y por tanto la Inercia Dentro también será baja.

2. La ***separación de*** unas ***clases*** con otras se encuentran bien separadas, puesto que las nubes de los centros de gravedad de las clases de una partición está dispersa por ser grande la Inercia Entre.

Cada nudo formado a través de este criterio en el árbol, incorpora un nivel equivalente a la inercia acumulada hasta el mismo, de manera que si sumamos los nudos de todos los niveles del árbol, tendremos la inercia total de la nube de puntos en estudio, y por tanto esta inercia representará la contribución relativa del nudo a la inercia total de la nube.

3.5. Representación Gráfica

El procedimiento permite elegir representaciones gráficas respecto a las variables nominales y respecto a las variables numéricas, ejecuta el método aglomerativo jerárquico descrito en el Capítulo I del presente Trabajo Fin de Master, por lo que habrá que marcar también manualmente el número de clúster y la técnica que se quiere.

En el gráfico siguiente se ha marcado el método de Ward, y una agrupación en 3 clusters, y como se puede observar las variables precio y potencia (CV) tienen una mayor incidencia sobre el cluster 3, mientras que el clúster 2 se encuentran los que poseen una mayor capacidad del depósito de combustible.

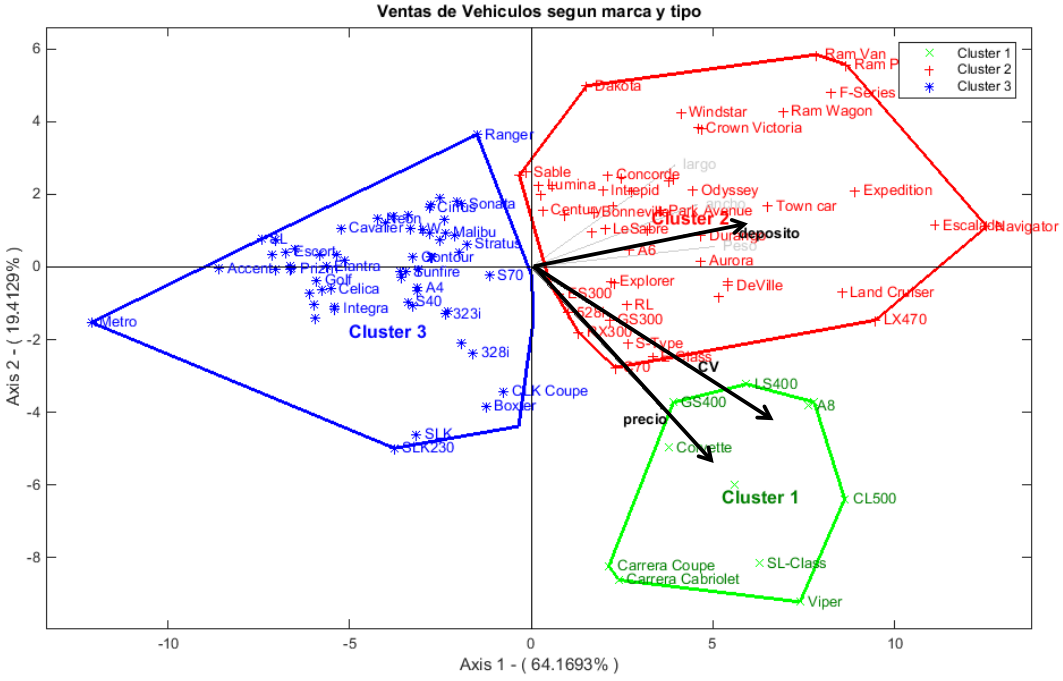


Fig. 3.1 Gráfica de la formación de 3 clúster

3.6. Comparativa TWO-STEP vs Clúster HJ-BIPLOT

El método TWO-STEP y el método Clúster del HJ-BIPLOT son dos procedimientos que tienen como objetivo común la agrupación en diferentes Clúster de la matriz de partida o inicial, sin embargo presentan algunas diferencias que se manifiestan en los resultados que de ellos se derivan. Por ello consideramos que, a pesar de la similitud que aparentemente presentan ambos métodos, es importante realizar una comparación detallada de los mismos.

Para las pruebas de contenido práctico se ha usado el software SPSS (IBM® SPSS® Statistics v22) para el método TWO-STEP y el software "***MULTBIPLOT***" (Vicente-Villardon, J.L., 2010). *MULTBIPLOT: A package for Multivariate Analysis using Biplots*. Departamento de Estadística. Universidad de Salamanca; **http://biplot.usal.es/ClassicalBiplot/index.html**), para el método de agrupamiento sobre el HJ-BIPLOT.

Se ha tomado de la web de IBM Knowledge Center (*IBM,* 2013), como base para la comparación práctica entre ambos métodos, un fichero de datos de coches que contiene estimaciones de ventas, precios de lista y especificaciones físicas hipotéticas de varias marcas y modelos de vehículos, y al cual se le ha cambiado las etiquetas

Es por tanto una matriz de 157 filas con 14 variables cada fila (157x14), con la siguiente composición:

Nombre	Tipo	Anchura	Decimales	Etiqueta	Medida
marca	Cadena	39	0	Fabricante	Nominal
modelo	Cadena	51	0	Modelo	Nominal
ventas	Numérico	11	3	Ventas en miles	Escala
reventa	Numérico	11	3	Valor de reventa a 4 años	Escala
tipo	Numérico	11	0	Tipo de vehículo	Ordinal
precio	Numérico	11	3	Precio en miles	Escala
motor	Numérico	11	1	Tamaño del motor	Escala
CV	Numérico	11	0	Potencia del motor	Escala
distejes	Numérico	11	1	Distancia entre ejes	Escala
ancho	Numérico	11	1	Anchura	Escala
largo	Numérico	11	1	Longitud	Escala
peso_neto	Numérico	11	3	Peso neto	Escala
deposito	Numérico	11	1	Capacidad del deposito	Escala
mpg	Numérico	11	0	Eficiencia en el consumo de combustible	Escala

Fig. 3.2 Tabla de variables del fichero de coches (*source IBM, 2013*)

3.6.1. Comparación relativa a la información de partida

El método ***TWO-STEP*** está pensado para trabajar con todo tipo de variables, ya sean estas de tipo alfanumérico o numérico, es decir de tipo categórico o de tipo continuo, independientemente de que puedan ser clasificadas las variables categóricas como nominales, ordinales o binarias.

marca	modelo	ventas	reventa	tipo	precio	motor	CV	distejes	ancho	largo	peso_neto	deposito	mpg
Lexus	GS400	3,334	.	Automóvil	46,305	4,0	300	110,2	70,9	189,2	3,693	19,8	21
Lexus	LS400	6,375	40,375	Automóvil	54,005	4,0	290	112,2	72,0	196,7	3,890	22,5	22
Lexus	LX470	9,126	.	Camión	60,105	4,7	230	112,2	76,4	192,5	5,401	25,4	15
Lexus	RX300	51,238	.	Camión	34,605	3,0	220	103,0	71,5	180,1	3,900	17,2	21
Lincoln	Continental	13,798	20,525	Automóvil	39,080	4,6	275	109,0	73,6	208,5	3,868	20,0	22
Lincoln	Town car	48,911	21,725	Automóvil	43,330	4,6	215	117,7	78,2	215,3	4,121	19,0	21
Lincoln	Navigator	22,925	.	Camión	42,660	5,4	300	119,0	79,9	204,8	5,393	30,0	15
Mitsubishi	Mirage	26,232	8,325	Automóvil	13,987	1,8	113	98,4	66,5	173,6	2,250	13,2	30
Mitsubishi	Eclipse	42,541	10,395	Automóvil	19,047	2,4	154	100,8	68,9	175,4	2,910	15,9	24
Mitsubishi	Galant	55,616	10,595	Automóvil	17,357	2,4	145	103,7	68,5	187,8	2,945	16,3	25
Mitsubishi	Diamante	5,711	16,575	Automóvil	24,997	3,5	210	107,1	70,3	194,1	3,443	19,0	22
Mitsubishi	3000GT	,110	20,940	Automóvil	25,450	3,0	161	97,2	72,4	180,3	3,131	19,8	21
Mitsubishi	Montero	11,337	19,125	Camión	31,807	3,5	200	107,3	69,9	186,6	4,520	24,3	18
Mitsubishi	Montero Sport	39,348	13,880	Camión	22,527	3,0	173	107,3	66,7	178,3	3,510	19,5	20
Mercury	Mystique	14,351	8,800	Automóvil	16,240	2,0	125	106,5	69,1	184,8	2,769	15,0	28
Mercury	Cougar	26,529	13,890	Automóvil	16,540	2,0	125	106,4	69,6	185,0	2,892	16,0	30
Mercury	Sable	67,956	11,030	Automóvil	19,035	3,0	153	108,5	73,0	199,7	3,379	16,0	24
Mercury	Grand Marquis	81,174	14,875	Automóvil	22,605	4,6	200	114,7	78,2	212,0	3,958	19,0	21
Mercury	Mountaineer	27,609	20,430	Camión	27,560	4,0	210	111,6	70,2	190,1	3,876	21,0	18
Mercury	Villager	20,380	14,795	Camión	22,510	3,3	170	112,2	74,9	194,7	3,944	20,0	21
Mercedes-Benz	C-Class	18,392	26,050	Automóvil	31,750	2,3	185	105,9	67,7	177,4	3,250	16,4	26

Fig. 3.3 Tabla de entrada de Datos del Método Two-Step

El método de ***Clúster del HJ-BIPLOT*** las variables alfanuméricas no pueden ser usadas de forma directa, por tanto las variables categóricas tienen que ser previamente codificadas a formato numérico, para posteriormente y mediante el programa de software "***MULTBIPLOT***" definido al comienzo del punto 3.6, definir cada variable como nominal, ordinal o binaria.

	ventas	reventa	tipo	precio	motor	CV	distejes	ancho	largo
Escort	70.2270	7.4250	0	12.0700	2	110	98.4000	67	174.7000
Mustang	113.3690	12.7600	0	21.5600	3.8000	190	101.3000	73.1000	183.2000
Contour	35.0680	8.8350	0	17.0350	2.5000	170	106.5000	69.1000	184.6000
Taurus	245.8150	10.0550	0	17.8850	3	155	108.5000	73	197.6000
Focus	175.6700	NaN	0	12.3150	2	107	103	66.9000	174.8000
Crown Victoria	63.4030	14.2100	0	22.1950	4.6000	200	114.7000	78.2000	212
Explorer	276.7470	16.6400	1	31.9300	4	210	111.6000	70.2000	190.7000
Windstar	155.7870	13.1750	1	21.4100	3	150	120.7000	76.6000	200.9000
Expedition	125.3380	23.5750	1	36.1350	4.6000	240	119	78.7000	204.6000
Ranger	220.6500	7.8500	1	12.0500	2.5000	119	117.5000	69.4000	200.7000
F-Series	540.5610	15.0750	1	26.9350	4.6000	220	138.5000	79.1000	224.5000
Civic	199.6850	9.8500	0	12.8850	1.6000	106	103.2000	67.1000	175.1000
Accord	230.9020	13.2100	0	15.3500	2.3000	135	106.9000	70.3000	188.8000
CR-V	73.2030	17.7100	1	20.5500	2	146	103.2000	68.9000	177.6000
Passport	12.8550	17.5250	1	26.6000	3.2000	205	106.4000	70.4000	178.2000

Fig. 3.4 Tabla de entrada de Datos del Método Clúster HJ-Biplot

3.6.2. Comparación relativa al Algoritmo

Respecto de criterio de clasificación y agrupamiento, ambos métodos se componen de dos fases, perfectamente diferenciadas una de otra, aunque la forma y manera de realizar cada fase los hace totalmente distintos entre sí.

1. **Fase de Construcción**

   En el método ***TWO-STEP***, la construcción del árbol de características, se realiza siguiendo la notación del método BIRCH (Zhang et al., 1996) y posteriormente implementado en al programa software SPSS (IBM® SPSS® Statistics) por Chiu, Fang, Chen, Wang y Jeris (Chiu et al., 2001), ampliamente expuesto en el Capítulo II del Presente Trabajo Fin de Máster.

   Se basa en leer los registros de la matriz de partida de uno en uno y en base al criterio de la distancia seleccionado, fusionar un registro con el anterior o formar un nuevo clúster, y para ello dispone de dos tipos de medida de distancia posibles que se pueden seleccionar indistintamente, aunque por defecto se usa la distancia de máxima verosimilitud para permitir la entrada de diferentes tipos de variables, también dispone de la opción para la distancia euclidea, además la selección del número de clusters que hay que formar se efectúa de forma automática, sin necesidad de definir previamente que número se quiere, aunque permite la posibilidad de forzar de forma manual el número de clúster que se desea.

   En el caso del método ***Clúster HJ-BIPLOT***, y una vez modificadas las variables de tipo mixto a tipo numérico, se ejecuta el HJ-Biplot para obtener las coordenadas mediante el método SVD de valores y vectores singulares, y una vez normalizadas todas las variables, se construye la matriz de distancias al cuadrado entre individuos en base a los criterios de inercia, definidos en el apartado 3.3 del Capítulo III del presente Trabajo Fin de Master, tanto para el cálculo de los individuos situados dentro del mismo clúster como para individuos en diferentes clusters, de forma que los nudos principales del árbol generado, va a contener la inercia total de los elementos que se encuentran dependiendo de dicho nudo.

2. **Fase de Agrupación**

   En el método ***TWO-STEP***, la agrupación se lleva a cabo mediante dos diferentes criterios que se tienen que seleccionar de forma manual, el Criterio Bayesiano de Schwarz o BIC (Schwartz, 1978) y el Criterio de Akaike o AIC (Akaike, 1974),

descritos en el Capítulo II del presente Trabajo Fin de Máster, y que sirven para ajustar modelos con variables de tipo mixto, aunque ambos criterios buscan minimizar el valor del criterio calculado maximizando el logaritmo de la función máximo verosímil.

La estandarización de las variables se realiza de forma automática al inicio del proceso de agrupamiento, aunque también es posible realizarlo de forma manual antes de su ejecución.

En el caso del método ***Clúster HJ-BIPLOT***, la agrupación está totalmente definida mediante el criterio jerárquico aglomerativo, y dentro de este cabe la posibilidad de elegir cualquiera de los métodos definidos para los modelos jerárquicos, y que han sido descritos en el Capítulo I del presente Trabajo Fin de Master, pues el objetivo es conseguir clusters totalmente homogéneos en base a la Inercia Dentro calculada según lo descrito en el apartado 3.3 del Capítulo III del presente Trabajo Fin de Master, y además conseguir que la separación entre clusters sea la mayor, en base a la Inercia Entre calculada según lo descrito en el apartado 3.3 del Capítulo III del presente Trabajo Fin de Master.

La estandarización se hace por defecto en columnas de la matriz, es decir sobre las variables, ya que aunque las variables tienen que ser numéricas, pueden existir variables nominales, ordinales, binarias y numéricas, y por tanto es necesario que todas estén en la misma métrica y que hablen el mismo lenguaje, aunque siempre se puede cambiar manualmente y elegir uno de los métodos incluidos dependiendo del análisis que se quiera realizar.

3.6.3. Comparación relativa a la Representación Gráfica

El método ***TWO-STEP***, usa el programa SPSS (IBM® SPSS® Statistics), y está basado principalmente en la interpretación de estructuras llevada a cabo por Kaufman y Rousseeuw, donde se puede observar la distribución realizada de los clúster mediante un gráfico de sectores y sus frecuencias, la importancia que han tenido las variables en la realización de dichos clusters, la distribución de cada clúster en base a las variables fijadas para el agrupamiento, tanto las categóricas como las numéricas, y una comparación de todos los cluster y las variables seleccionadas para su agrupación, mediante diagramas de caja para las variables continuas y color y tamaño diferente para las variables categóricas (Kaufman y Rousseeuw, 1987).

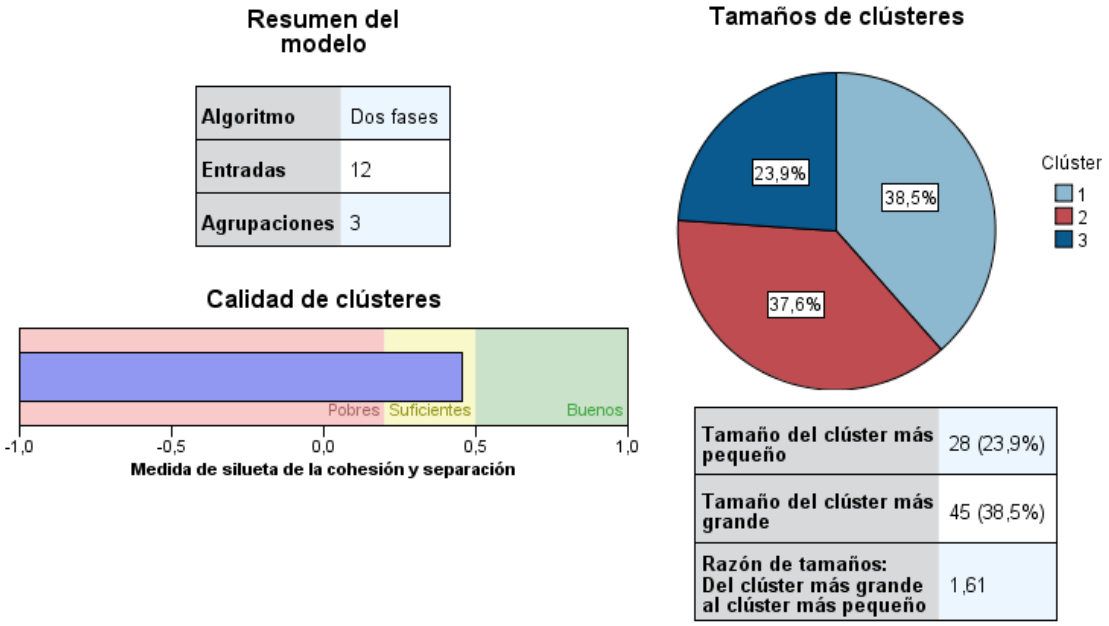


Fig. 3.5 Gráficos Two-Step, Número de clústers, tamaño y calidad de la agrupación

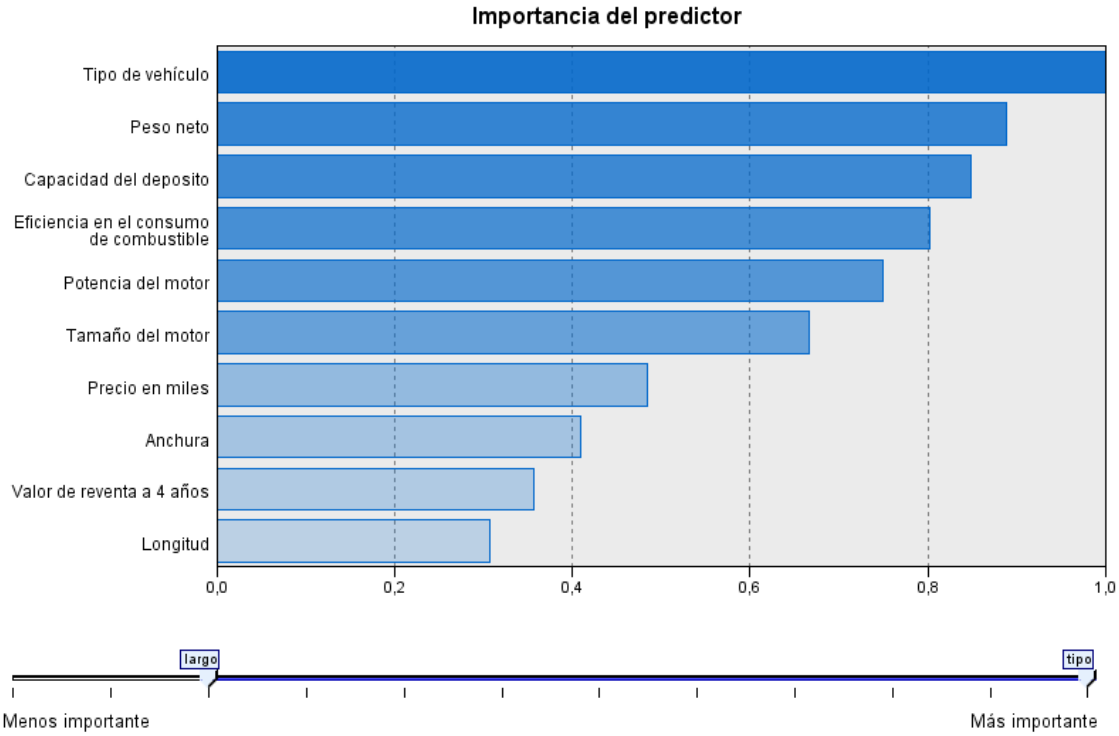


Fig. 3.6 Gráficos Two-Step, Importancia de las variables en la formación de los clusters

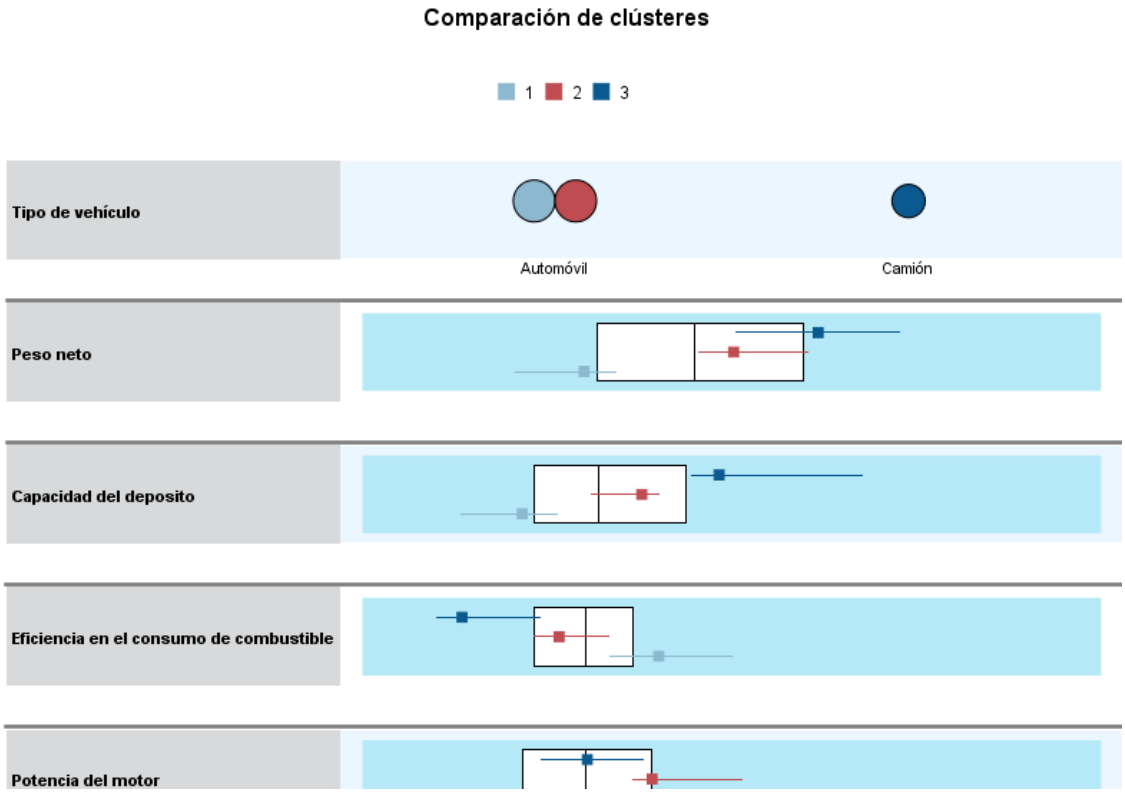


Fig. 3.7 Gráficos Two-Step, Comparación de los clusters formados

El método de ***Clúster HJ-BIPLOT***, se puede realizar o bien para las variables continuas o bien para cada variable nominal, por separado, el siguiente grafico se ha marcado la Distancia de Ward y el Método Jerárquico.

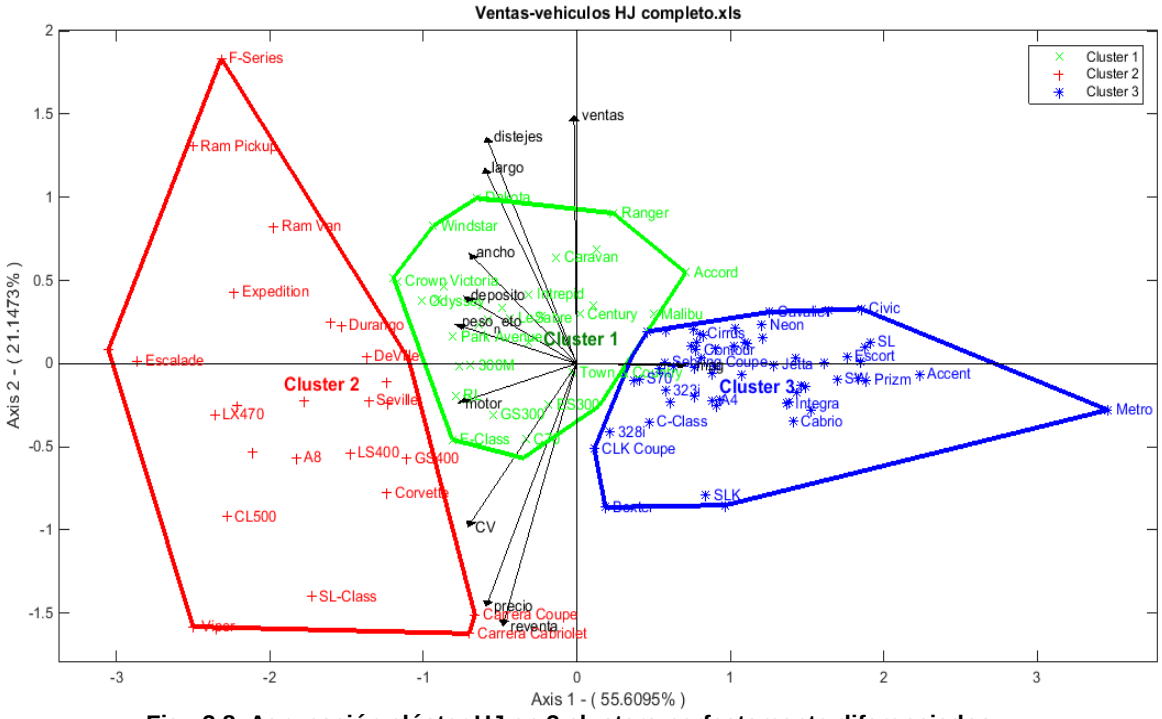


Fig. 3.8 Agrupación clúster HJ en 3 clusters perfectamente diferenciados

3.6.4. Limitaciones HJ *vs* TWO-STEP

Finalmente se exponen las limitaciones encontradas en la comparación de ambos algoritmos:

1) En primer lugar hay que señalar el uso de los tipos de variables que puede existir en una matriz de Datos, el TWO-STEP permite la entrada de cualquier tipo de variable, incluso permite la tipificación de variables alfanuméricas como variables nominales, ordinales o binarias, mientras que el Clúster HJ-BIPLOT solo permite variables numéricas, y por tanto es necesario un paso previo al algoritmo para poder tipificar cualquier variable de la matriz de datos de entrada, y aun cuando no es un hándicap en sí mismo, si plantea cierta limitación en el tratamiento de variables no numéricas, aspecto importante en el tratamiento de variables en diferentes campos de la ciencia, que en muchas ocasiones no están tipificadas e incluso puede resultar complicado su tipificación..

2) El método TWO-STEP usa la distancia de máxima verosimilitud con datos de tipo categórico, mientras que el algoritmo Clúster del HJ-BIPLOT tiene que modificar en primer lugar todas las variables categóricas a tipo numérico y a continuación crear la matriz de coordenadas del HJ-Biplot, y finalmente usar el criterio de Inercia maximizando la Inercia Entre clusters y al mismo tiempo minimizando la Inercia Dentro de cada clúster, partiendo de variables tipificadas todas como numéricas y teniendo en cuenta que como en el HJ-Biplot las masas son unitarias, la inercia coincide entonces con la distancia al cuadrado.

Parece en principio parecido el procedimiento para crear los nudos del árbol con ambos algoritmos, dado que en el Two-Step los nudos van a contener las caracterisiticas del clúster (CF) y los nudos en el método de inercia van a contener la inercia total del nudo, pero mientras que en el Two-Step la agrupación se consigue con el método de la máxima verosimilitud, y ese procedimiento se sabe que toma como hipótesis que las variables son todas independientes y que las variables numéricas siguen una distribución normal, y las variables categóricas siguen una distribución multinomial, lo cual no siempre se puede asegurar matemáticamente, y nos llevaría a errores de apreciación al tener dicha hipótesis para construir los

clúster, en cambio en el algoritmo del criterio de inercia es mucho mas racional y optimo, ya que se construye la inercia total de los individuos que conforman cada clúster, con el único inconveniente que parte de cambiar todas las variables categóricas como nominales o binarias con la valoración que el investigador estime oportuna.

Lo más razonable según las investigaciones llevadas a cabo, es utilizar el coeficiente de Gower (Gower, 1971) para variables de tipo mixto, ya que incluso este coeficiente nos permitiría tratar variables con datos perdidos, pero seria necesario crear y comprobar con diferentes software la interface necesaria para ejecutar diferentes soluciones matemáticas.

3) En cuanto a la representación gráfica ambos métodos lo realizan de diferentes maneras, pero es evidente la gran representación gráfica que proporciona el método Clúster del HJ-BIPLOT frente al método TWO-STEP, consiguiendo la mayor calidad de representación para filas y columnas, ya que en el primero se sabe perfectamente la variabilidad que se puede ver representada en los planos que se eligen, además de ver de qué manera afectan las variables del modelo a los individuos de la matriz.

Capítulo IV

AGRUPAMIENTO EN CLUSTER

CLUSPLOT

vs

TWO-STEP & Clúster HJ-BIPLOT

4.1. Introducción

CLUSPLOT es una nueva forma de representar Clusters en el cual los objetos son representados como puntos en un gráfico bidimensional y los cluster como elipses de varios tamaños y formas. (Pison et al., 1999).

Entre los algoritmos utilizados en la formación del CLUSPLOT, siguiendo a Kaufman y Rousseeuw (1990) en su obra *"Finding Groups in Data*", se incluyen PAM, CLARA, FANNY, AGNES, DIANA y MONA. Los tres primeros están considerados como algoritmos de tipo particionado, mientras que los tres últimos están considerados como algoritmos de tipo jerárquico. En la misma obra Kaufman y Rousseeuw desarrollan el algoritmo CLUSPLOT que estaba escrito en lenguaje Pascal.

Struyf y col. (1997), implementan estos algoritmos dentro del programa software S-PLUS, por lo que fue necesario transformar las rutinas del lenguaje Fortran originario en lenguaje que pudiese entender el nuevo paquete software. También están implementadas dichas rutinas dentro del software "R", ya que como software libre permite llamadas a rutinas escritas en Fortran, C, C++,

Debido a que CLUSPLOT solo representa en el plano bidimensional los datos de nuestra matriz original, es necesario utilizar alguno de los algoritmos antes mencionados si las variables a utilizar son de tipo numérico o bien siguiendo a Kaufman y Rousseeuw (1990) utilizar el algoritmo DAISY, para transformar nuestra matriz de variables categóricas en una matriz de disimilaridades, utilizando el coeficiente de Gower, descrito en el Capítulo I del presente trabajo.

En el gráfico siguiente, se detalla el diagrama de proceso para la ejecución del algoritmo CLUSPLOT, partiendo de la matriz de datos original, y teniendo en cuenta el tipo de variables que puede tener esta matriz de datos original, creando en cada paso con el algoritmo correspondiente la matriz necesaria para que CLUSPLOT pueda representar gráficamente la matriz original.

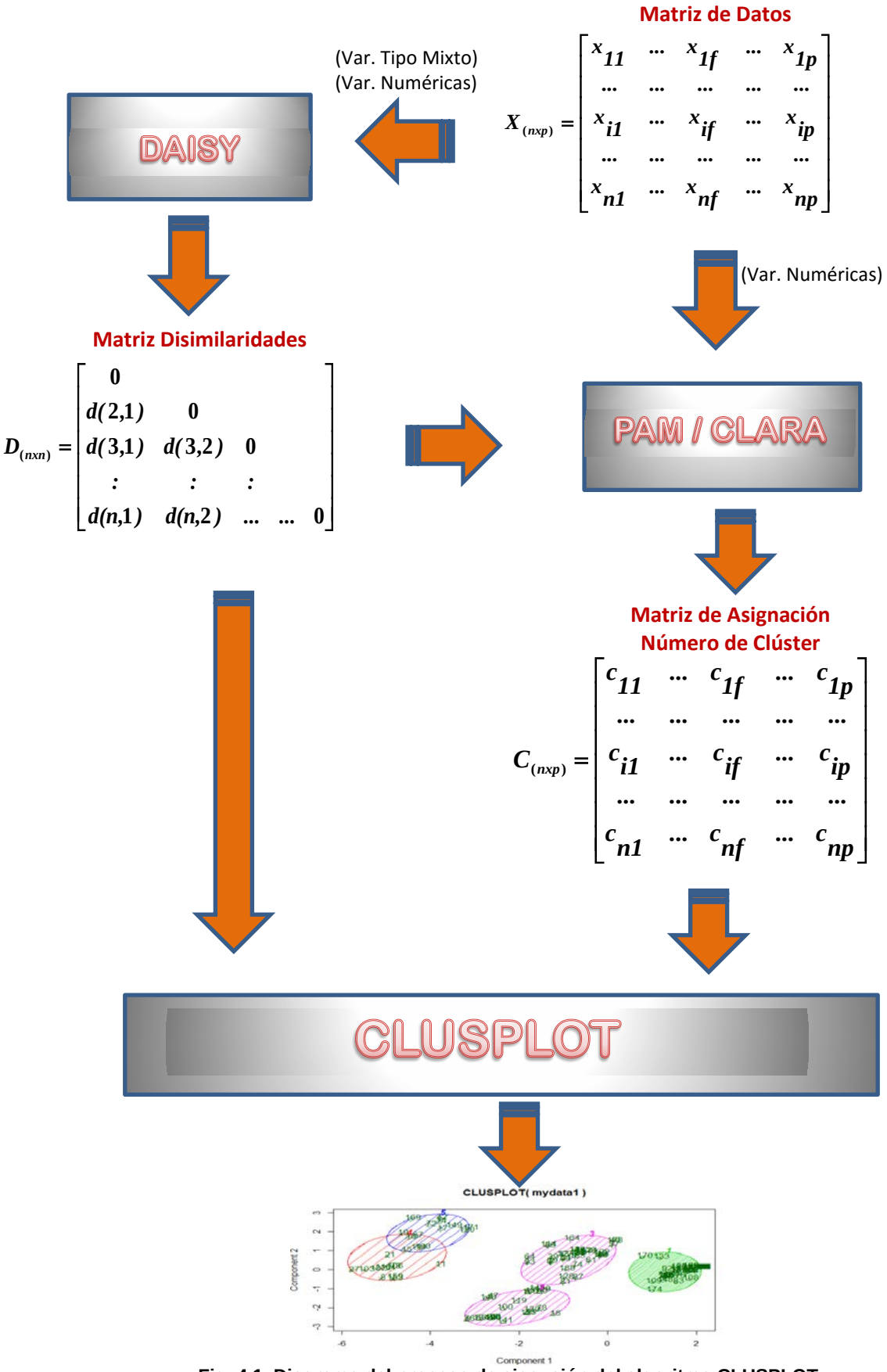


Fig. 4.1 Diagrama del proceso de ejecución del algoritmo CLUSPLOT

4.2. Algoritmo del Método CLUSPLOT

CLUSPLOT es por tanto, según Kaufman y Rousseeuw (1990), una representación gráfica que usa la salida que le proporciona un algoritmo de datos particionados para poder visualizar los individuos y los clúster formados por dichos individuos según sus variables sobre un gráfico en dos dimensiones

La función CLUSPLOT desarrollada dentro del software S-PLUS (Pison et al., 1999), así como dentro del software R, proporciona un gráfico en representación bidimensional del agrupamiento en clúster, anteriormente las herramientas gráficas representaban gráficos de distancia (Chen et al., 1974) o gráficos de silueta (silhouette) (Rousseeuw, 1987), el algoritmo CLUSPLOT puede también ser visto como una versión generalizada y automatizada de mapas taxometricos (Carmichael y Sneath, 1969).

La gráfica que proporciona CLUSPLOT puede tomar diferentes formas, dependiendo del algoritmo que se use para ello y que se encuentran implementados en el software S-PLUS, por ejemplo remplazando cada elipse por una forma convexa de todos los puntos que están en el clúster usando el algoritmo de Eddy (1977), por el determinante de mínima covarianza "MCD", construido por (Rousseeuw y Van Driessen, 1997), o bien con el algoritmo de Titterington (1976), basado en la media y la matriz de covarianzas de cada clúster, donde su tamaño se corresponde con el total de puntos que lo contiene, por lo que siempre aparecen puntos en el borde de las elipses que forman cada clúster.

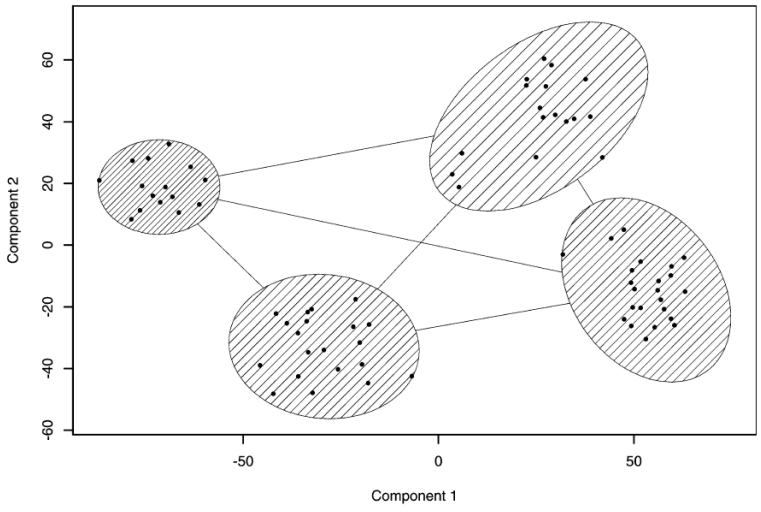


Fig. 4.2 Clusplot de la tabla de Rusipini de 75 elementos y 4 clusters (Source: Pison et al., 1999)

La sintaxis de esta rutina, tiene la forma siguiente (Pison et al., 1999), y al menos es necesario especificar los tres primeros argumentos, siendo opcionales el resto de ellos, en cuyo caso el algoritmo tomará los valores que tienen por defecto, algunos de ellos se muestran a continuación:

clusplot	
(X,	Matriz de datos o matriz de disimilaridad, dependiendo si se ha usado el algoritmo "Daisy" para crear la matriz de disimilaridades previamente. No hace falta especificar el número de individuos que tiene la matriz, ya que el algoritmo lo calcula automáticamente
clus,	Un vector de longitud "n" que representa el agrupamiento de la matriz "X", en donde para cada individuo el vector indica el número o el nombre del clúster al que se le ha asignado, por lo general es la salida del algoritmo PAM, CLARA o FANNY
diss = TRUE,	Si la matriz inicial de datos es una matriz de disimilaridades, entonces el argumento tiene que ser TRUE (valor por defecto), en caso contrario tiene que ser FALSE
Lines = 2,	Este argumento puede tener los valores de 0, 1, 2, y nos indica la distancia entre dos elipses, dibujada a través de las líneas que las unen. 0 = No aparece la línea de conexión entre clusters 1 = La línea que une los centros de las elipses 2 = La línea que une los bordes de las elipses
Shade = FALSE,	Es un argumento lógico, siendo TRUE para sombrear las elipses en base a su densidad, siendo la densidad el número de puntos que hay en el clúster dividido por el área de la elipse, el valor por defecto es FALSE
Labels = 0,	Este argumento puede tener los valores de 0, 1, 2, 3 y 4 0 = No se ponen etiquetas en el gráfico a los individuos 1 = Individuos y elipses, pueden ser identificados 2 = Individuos y elipses estarán identificados, (valor por defecto) 3 = Solo los individuos estarán identificados 4 = Solo las elipses estarán identificadas
Stand = FALSE	Valor por defecto, cuando todos los puntos de la matriz original están estandarizados, en caso contrario hay que poner el valor TRUE
Color = FALSE	Es un argumento lógico, si es TRUE las elipses serán coloreadas con respecto a sus densidades, lo que significa que a medida que aumenta la densidad en las elipses los colores son azul claro, verde claro, rojo y púrpura. Cuando se tienen 4 o menos clusters, entonces el valor de TRUE le da a cada clúster un color diferente, y cuando hay más de 4, el algoritmo CLUSPLOT usa la función PAM para dividir las densidades en 4 grupos, de forma que las elipses con igual densidad tengan el mismo color.
Col.p = "black"	Es el código de color usado para los puntos que representan los individuos. "black"; "red", "dark Green", etc…

Existen algunos casos especiales dentro del algoritmo CLUSPLOT:

a) ***Cuando un clúster solo contiene a un individuo***, entonces se dibuja un circulo alrededor de dicho individuo

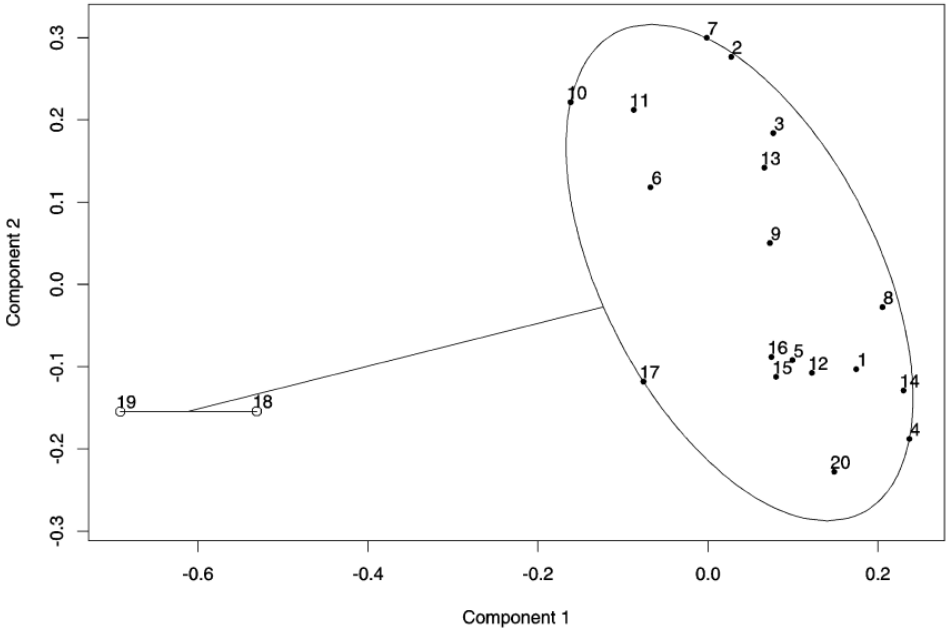


Fig. 4.3 Clusplot of the dissimilarity data, (Abbot y Perkins, 1978)

b) ***Cuando los individuos de un clúster caen sobre una line recta***, con el argumento "span = FALSE", entonces dibuja una elipse alrededor de esta, y si la opción fuese "span = TRUE", entonces dibuja una línea recta exacta.

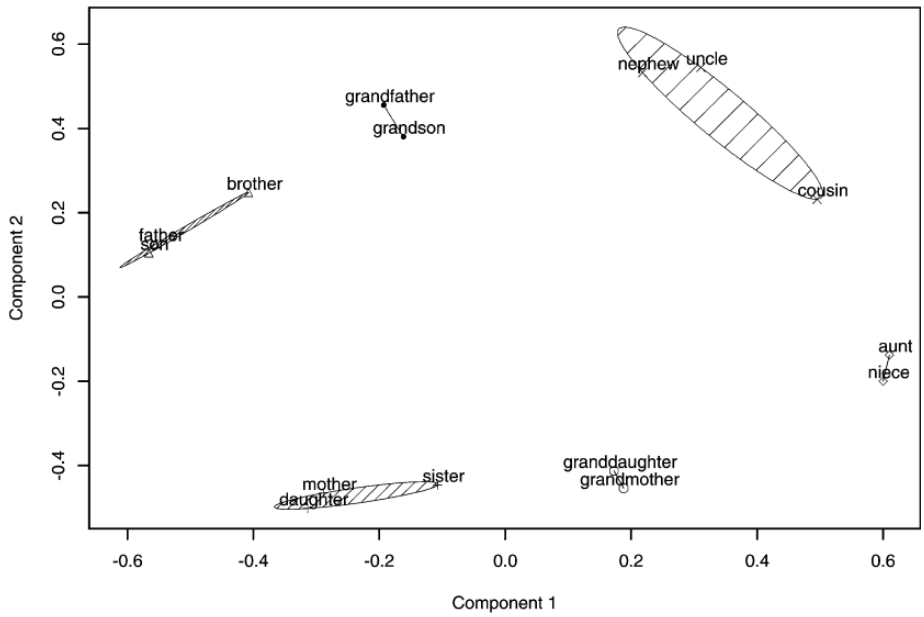


Fig. 4.4 (Source: Clusplot of the Rosenberg dissimilarity data, 1982)

4.2.1. Cálculo de la Matriz de Disimilaridades

Si la matriz de partida tiene variables de tipo mixto se dispone de un algoritmo denominado DAISY que se incluye dentro de la obra titulada "Finding Groups in Data" de los autores Kaufman y Rousseeuw (1990), que aporta la posibilidad de transformar dicha matriz en una nueva matriz de disimilaridades necesaria para que el algoritmo CLUSPLOT pueda a partir de ella elaborar la representación gráfica en dos dimensiones de la matriz original, y para ello aplicará la métrica de Gower (Gower, 1971), este algoritmo, también se suele usar aunque los datos de partida sean numéricos, para transformar la matriz de partida en una matriz de disimilaridades.

Para ahorrar espacio de almacenamiento en el proceso de ejecución del algoritmo, la matriz de disimilaridades se representa como un vector y no como una matriz completa, únicamente se considera la parte triangular superior de la matriz ya que es simétrica respecto de la diagonal principal

En el siguiente gráfico correspondiente a la matriz de datos de Harman (1967), contiene disimilaridades entre 13 individuos acerca de resultados sobre test de psicología, indicándole que realice 3 clúster, y el algoritmo CLUSPLOT nos muestra el siguiente gráfico, indicando la variabilidad de los dos componentes elegidos, asi como los correspondientes clusters.

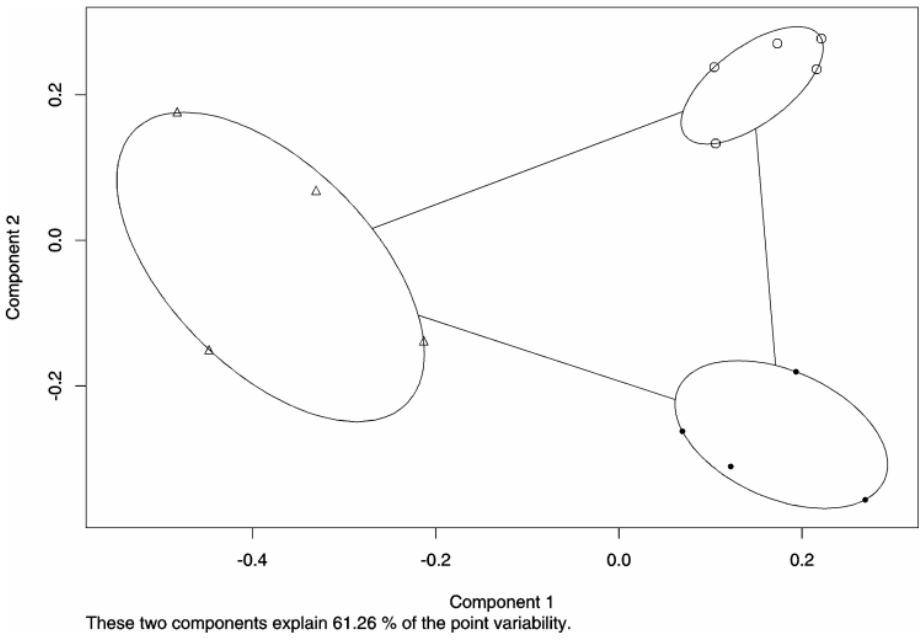


Fig. 4.5 Clusplot of the Harman dissimilarity data

4.2.2. Asignación del número de Clúster

Para realizar el cálculo de la matriz que contiene el número de clúster asignado a cada elemento de la matriz de datos original, necesaria para que el algoritmo CLUSPLOT pueda interpretar y representar en un plano de dos dimensiones nuestra matriz de datos, es necesario utilizar cualquiera de los algoritmos particionados que se han descrito en el apartado 4.1 del presente Capítulo IV, los cuales a partir de la matriz original si son datos numéricos o bien con la matriz de disimilaridades calculada previamente por el algoritmo DAISY si son datos de tipo mixto pueda generar dicha matriz de vectores.

Para realizar este proceso se utilizan fundamentalmente el algoritmo "PAM" o el algoritmo "CLARA"

a) <u>Algoritmo "PAM"</u>

Este método fue introducido por Kaufman y Rousseeuw (1990), el algoritmo PAM (Partitioning Around Medoids) es una versión más robusta del algoritmo de K-Medias, también se le denomina a este algoritmo como K-medoid, al usar la función del algoritmo K-medias pero con la restricción de que los centros de los clusters formados deben de pertenecer a la matriz de datos, por lo que estos medoides estarán localizados lo más cerca del centro de cada clúster y que cada individuo estará agrupado con el medoide más cercano (Leiva et al., 2010), realizando todos los cambios necesarios entre los individuos más representativos hasta que se minimiza la medida de disimilitud entre los k-medoides y los vectores de los individuos que forman los clusters (Kamer y Han, 2006).

El algoritmo PAM, por tanto está basado en buscar los ***"k"*** individuos representativos de la matriz de datos del análisis, ya que estos ***"k"*** individuos son los denominados medoides del algoritmo, por lo que estos medoides serán calculados de forma que la disimilaridad total de todos los individuos a su medoide más cercano sea mínima, es decir, que el objetivo fundamental es encontrar un subconjunto $\{m_1, m_2 m_3, \dots\dots, m_k\} \subset \{1,2, \dots \dots n\}$ que sea capaz de minimizar la función objetivo:

$$\sum_{i=1}^{n} min_{t=1,2,\dots k}\, d(i, m_t)$$

b) Algoritmo "CLARA"

Este método fue introducido por Kaufman y Rousseeuw (1990), el algoritmo CLARA (Clustering Large Applications) es un método de particionado similar al algoritmo PAM, pero es mucho más eficiente al trabajar con matrices de datos muy grandes, más de 250 individuos, debido a que el algoritmo no guarda la matriz de dismilaridades en memoria como lo hace PAM y no produce una ralentización del proceso de cálculo de los clusters sobre la matriz de datos (Struyf et al., 1997).

El algoritmo CLARA, separa cada muestra de la matriz de datos en "k" diferentes clusters aplicando el algoritmo PAM sobre cada una de dichas muestras, para poder encontrar los k-medoides de cada muestra, lo que hace necesario considerar cada muestra del mismo tamaño, de forma que las necesidades de memoria de proceso y de espacio sean lineales para todas ellas.

Según los autores (Kaufman y Rousseeuw, 1990), los resultados experimentales indican que 5 muestras con (40 + 2k) individuos cada una, produce unos resultados satisfactorios (Leiva et al., 2010), siendo la calidad de los clúster medida con la disimilitud media de todos los datos, y no únicamente con aquellos individuos que están incluidos en las muestras (Ng y Han, 1994).

4.3. Criterio de Agrupamiento

La salida de los algoritmos PAM y CLARA, se compone de los llamados silhouettes, explicado en el Capítulo 2 del presente Trabajo Fin de Master, y por tanto CLUSPLOT utiliza la salida de un algoritmo de partición para visualizar los objetos y grupos en una trama bidimensional, y CLUSPLOT utiliza como entrada la salida que genera estos algoritmos asi como la salida del algoritmo DAISY.

La estructura de datos de entrada puede ser o bien una matriz de datos que ha preparado previamente cualquiera de los algoritmos mencionados anteriormente o bien una matriz de disimilaridades que ha preparado previamente el algoritmo DAISY, si se utiliza una matriz de disimilaridades como entrada, CLUSPLOT comienza mediante la conversión de los datos a coordenadas bidimensionales por medio del método de escalamiento multidimensional, y la misma técnica se utiliza si los datos consisten en una matriz de datos original con más de dos variables (Kaufman y Rousseeuw, 1990),

El algoritmo PAM tiene dos fases (Struyf et al., 1997), es un método que intenta determinar k particiones o clusters conformados por n individuos cada clúster que sean representativos (Ng y Han, 1994).

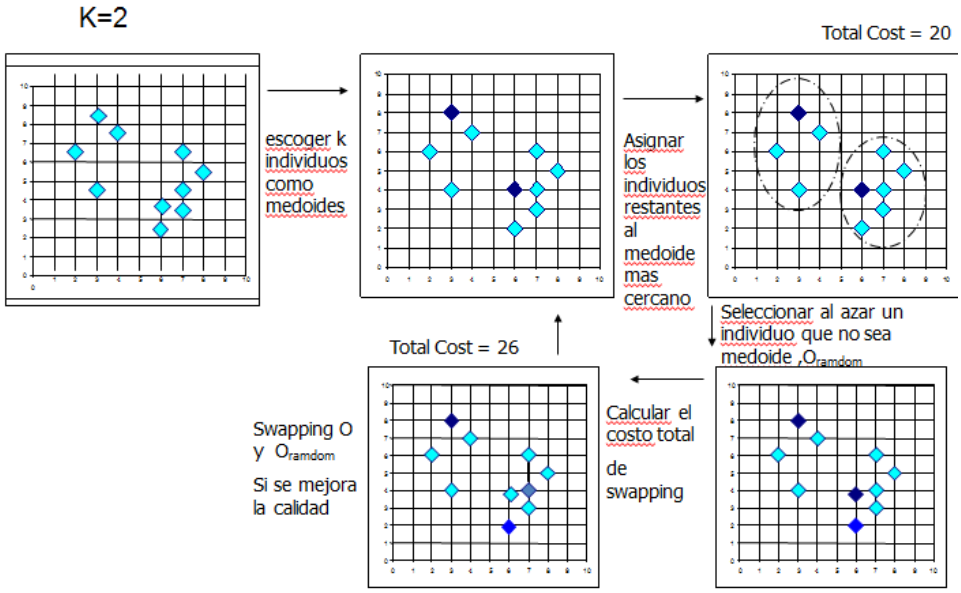


Fig. 4.6 - Diagrama de funcionamiento del algoritmo PAM

1º. FASE DE CONSTRUCCION (BUILD)

Selecciona de forma arbitraria ***k*** individuos representativos de entre la matriz de datos, bien sea esta en formato original o una vez que ha sido transformada a una nueva matriz de disimilaridades.

Haciendo mínima la disimilaridad entre todos los integrantes del clúster, es decir encontrar un subconjunto $\{m_1, m_2 m_3, \ldots\ldots, m_k\} \subset \{1,2, \ldots\ldots n\}$ que sea capaz de minimizar la función objetivo:

$$\sum_{i=1}^{n} min_{t=1,2,\ldots k}\, d(i, m_t)$$

2º. FASE DE CAMBIO (SWAP)

Para conseguir que mejore la calidad del intercambio, hay que hacer mínima la función de Costos (Struyf et al., 1997) y (Leiva et al., 2010), calculando los costos para todos los individuos no seleccionados, y si

realizando el intercambio se mejora dicha función de costos, usando por defecto la distancia euclidea,

Esta fase se realiza repetidamente hasta conseguir minimizar la función de costos.

$$\mathbf{Min\ TC_{mh}\,(O_m, O_h)} \quad \Leftrightarrow \quad \mathbf{TC_{mh}} = \sum_j \mathbf{C_{jmh}}$$

Cuando un medoide "m" tiene que cambiarse con uno que no es medoide "h", tiene que chequear cada uno de los no medoides "j" , de manera que se pueden dar las dos siguientes posibilidades:

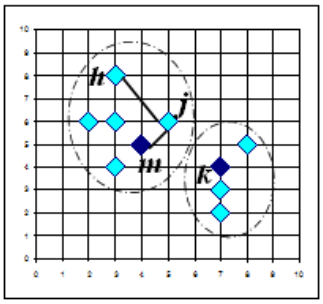
1) Que el elemento "j" esté en el clúster del elemento "m" y haya que reasignar al elemento "j"

a) *Caso 1*

El elemento "j" está cerca del elemento "k" que está dentro de su clúster, pero no del elemento "h" que está situado en otro clúster diferente, por lo que después del intercambio de "m" y de "h", el elemento "j" estará alojado en el clúster del elemento "k"

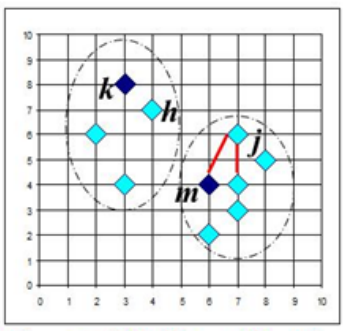
b) *Caso 2*

El elemento "j" está más cerca del elemento "h" que del elemento "k", siendo ambos de un clúster diferente a donde se encuentra "j", por lo que después del intercambio de "m" y de "h", el elemento "j" estará alojado en el clúster del elemento "h"



$C_{jmh} = d(j, k) - d(j, m) \geq 0$

Caso 1



$C_{jmh} = d(j, h) - d(j, m)$

Caso 2

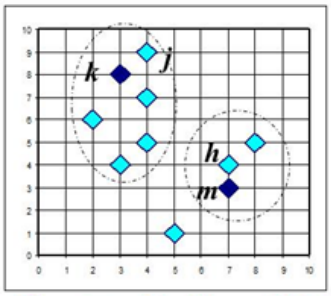
2) Que el elemento "j" esté en el clúster del elemento "k", pero no en el clúster del elemento "m", por lo que habrá que comparar los elementos "k" con "h"

a) *Caso 3*

El elemento "j" está más cerca del elemento "k" que del elemento "h" que está situado en otro clúster diferente, por lo que después del intercambio de "m" y de "h", el elemento "j" seguirá alojado en el clúster del elemento "k"

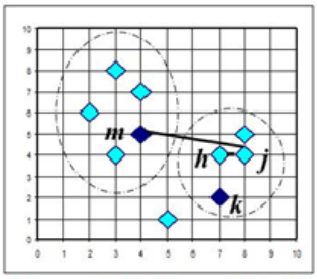
b) *Caso 4*

El elemento "j" está más cerca del elemento "h" que pertenece a otro clúster diferente que del elemento "k" que pertenece al mismo clúster de "j", por lo que después del intercambio de "m" y de "h", el elemento "j" estará alojado en el clúster del elemento "h"



$C_{jmh} = d(j, k) - d(j, k) = 0$

Caso 3



$C_{jmh} = d(j, h) - d(j, k) < 0$

Caso4

Una vez que el algoritmo ha construido el número de clusters que se le ha indicado manualmente en el parámetro correspondiente del algoritmo, la salida particionada se la entrega al algoritmo CLUSPLOT para que realice la representación gráfica en un modelo de dos dimensiones, en donde los individuos estarán representados como puntos y los clusters estarán representados por elipses de distintos tamaños y formas (Pison et al., 1999)

4.4. Comparación CLUSPLOT *vs* TWO-STEP & Clúster HJ-BIPLOT

El método ***TWO-STEP,*** el método ***CLUSPLOT*** y el método ***Clúster HJ-BIPLOT*** son procedimientos que tienen como objetivo común la agrupación en diferentes Clúster de la matriz de partida o inicial, sin embargo presentan algunas diferencias que se manifiestan en los resultados que de ellos se derivan. Por ello consideramos que, a pesar de la similitud que aparentemente presentan todos los métodos, es importante realizar una comparación detallada de los mismos.

Para las pruebas de contenido práctico se ha usado:

A. El software ***SPSS*** (IBM® SPSS® Statistics, 1989) para el método TWO-STEP

B. El software ***R*** versión 3.1.2 (copyright © R Foundation for Statistical Computing, 2014) para el método de CLUSPLOT.

C. El software "***MULTBIPLOT***" (Vicente-Villardon, J.L., 2010). *MULTBIPLOT: A package for Multivariate Analysis using Biplots*. Departamento de Estadística. Universidad de Salamanca; http://biplot.usal.es/ClassicalBiplot/index.html), para el método de agrupamiento sobre el HJ-BIPLOT.

Se ha tomado de la web de IBM Knowledge Center (*IBM,* 2013), como base para la comparación práctica entre ambos métodos, un fichero de datos de coches que contiene estimaciones de ventas, precios de lista y especificaciones físicas hipotéticas de varias marcas y modelos de vehículos, y al cual se le ha cambiado las etiquetas, es por tanto una matriz de 157 filas con 14 variables cada fila (157x14), con la siguiente composición:

Nombre	Tipo	Anchura	Decimales	Etiqueta	Medida
marca	Cadena	39	0	Fabricante	Nominal
modelo	Cadena	51	0	Modelo	Nominal
ventas	Numérico	11	3	Ventas en miles	Escala
reventa	Numérico	11	3	Valor de reventa a 4 años	Escala
tipo	Numérico	11	0	Tipo de vehículo	Ordinal
precio	Numérico	11	3	Precio en miles	Escala
motor	Numérico	11	1	Tamaño del motor	Escala
CV	Numérico	11	0	Potencia del motor	Escala
distejes	Numérico	11	1	Distancia entre ejes	Escala
ancho	Numérico	11	1	Anchura	Escala
largo	Numérico	11	1	Longitud	Escala
peso_neto	Numérico	11	3	Peso neto	Escala
deposito	Numérico	11	1	Capacidad del deposito	Escala
mpg	Numérico	11	0	Eficiencia en el consumo de combustible	Escala

Fig. 4.7 Tabla de variables del fichero de coches (*source IBM, 2013*)

4.4.1. Comparación relativa a la Información de partida

El método ***TWO-STEP*** está pensado para trabajar con todo tipo de variables, ya sean estas de tipo alfanumérico o numérico, es decir de tipo categórico o de tipo continuo, independientemente de que puedan ser clasificadas las variables categóricas como nominales, ordinales o binarias, ya que en las pruebas llevadas a cabo se han realizado indistintamente con las variables sin clasificar y clasificándolas.

marca	modelo	ventas	reventa	tipo	precio	motor	CV	distejes	ancho	largo	peso_neto	deposito	mpg
Lexus	GS400	3,334	.	Automóvil	46,305	4,0	300	110,2	70,9	189,2	3,693	19,8	21
Lexus	LS400	6,375	40,375	Automóvil	54,005	4,0	290	112,2	72,0	196,7	3,890	22,5	22
Lexus	LX470	9,126	.	Camión	60,105	4,7	230	112,2	76,4	192,5	5,401	25,4	15
Lexus	RX300	51,238	.	Camión	34,605	3,0	220	103,0	71,5	180,1	3,900	17,2	21
Lincoln	Continental	13,798	20,525	Automóvil	39,080	4,6	275	109,0	73,6	208,5	3,868	20,0	22
Lincoln	Town car	48,911	21,725	Automóvil	43,330	4,6	215	117,7	78,2	215,3	4,121	19,0	21
Lincoln	Navigator	22,925	.	Camión	42,660	5,4	300	119,0	79,9	204,8	5,393	30,0	15
Mitsubishi	Mirage	26,232	8,325	Automóvil	13,987	1,8	113	98,4	66,5	173,6	2,250	13,2	30
Mitsubishi	Eclipse	42,541	10,395	Automóvil	19,047	2,4	154	100,8	68,9	175,4	2,910	15,9	24
Mitsubishi	Galant	55,616	10,595	Automóvil	17,357	2,4	145	103,7	68,5	187,8	2,945	16,3	25
Mitsubishi	Diamante	5,711	16,575	Automóvil	24,997	3,5	210	107,1	70,3	194,1	3,443	19,0	22
Mitsubishi	3000GT	,110	20,940	Automóvil	25,450	3,0	161	97,2	72,4	180,3	3,131	19,8	21
Mitsubishi	Montero	11,337	19,125	Camión	31,807	3,5	200	107,3	69,9	186,6	4,520	24,3	18
Mitsubishi	Montero Sport	39,348	13,880	Camión	22,527	3,0	173	107,3	66,7	178,3	3,510	19,5	20
Mercury	Mystique	14,351	8,800	Automóvil	16,240	2,0	125	106,5	69,1	184,8	2,769	15,0	28
Mercury	Cougar	26,529	13,890	Automóvil	16,540	2,0	125	106,4	69,6	185,0	2,892	16,0	30
Mercury	Sable	67,956	11,030	Automóvil	19,035	3,0	153	108,5	73,0	199,7	3,379	16,0	24
Mercury	Grand Marquis	81,174	14,875	Automóvil	22,605	4,6	200	114,7	78,2	212,0	3,958	19,0	21
Mercury	Mountaineer	27,609	20,430	Camión	27,560	4,0	210	111,6	70,2	190,1	3,876	21,0	18
Mercury	Villager	20,380	14,795	Camión	22,510	3,3	170	112,2	74,9	194,7	3,944	20,0	21
Mercedes-Benz	C-Class	18,392	26,050	Automóvil	31,750	2,3	185	105,9	67,7	177,4	3,250	16,4	26

Fig. 4.8 – Tabla de Entrada de Datos del Método Two-Step

El método ***CLUSPLOT*** no define directamente el tipo de variable de la matriz de datos inicial, requiere de un paso previo a su ejecución para poder adaptar la matriz inicial de datos a una matriz que pueda leer dicho algoritmo, para lo cual se apoya en el ***algoritmo DAISY*** y en el ***algoritmo PAM***, analizado en el presente Capítulo IV en detalle, y que permite trabajar con todo tipo de variables, ya sean estas de tipo categórico o de tipo continuo.

	ventas	reventa	tipo	precio	motor	CV	pisada	ancho	largo	peso_neto
1	16,919	16,360	Automovil	21,500	1,80	140	101,20	67,30	172,40	2,64
2	39,384	19,875	Automovil	28,400	3,20	225	108,10	70,30	192,90	3,52
3	14,114	18,225	Automovil		3,20	225	106,90	70,60	192,00	3,47
4	8,588	29,725	Automovil	42,000	3,50	210	114,60	71,40	196,60	3,85
5	20,397	22,255	Automovil	23,990	1,80	150	102,60	68,20	178,00	3,00
6	18,780	23,555	Automovil	33,950	2,80	200	108,70	76,10	192,00	3,56
7	1,380	39,000	Automovil	62,000	4,20	310	113,00	74,00	198,20	3,90
8	19,747		Automovil	26,990	2,50	170	107,30	68,40	176,00	3,18
9	9,231	28,675	Automovil	33,400	2,80	193	107,30	68,50	176,00	3,20
10	17,527	36,125	Automovil	38,900	2,80	193	111,40	70,90	188,00	3,47
11	91,561	12,475	Automovil	21,975	3,10	175	109,00	72,70	194,60	3,37
12	39,350	13,740	Automovil	25,300	3,80	240	109,00	72,70	196,20	3,54
13	27,851	20,190	Automovil	31,965	3,80	205	113,80	74,70	206,80	3,78
14	83,257	13,360	Automovil	27,885	3,80	205	112,20	73,50	200,00	3,59
15	63,729	22,525	Automovil	39,895	4,60	275	115,30	74,50	207,20	3,98
16	15,943	27,100	Automovil	44,475	4,60	275	112,20	75,00	201,00	
17	6,536	25,725	Automovil	39,665	4,60	275	108,00	75,50	200,60	3,84
18	11,185	18,225	Automovil	31,010	3,00	200	107,40	70,30	194,80	3,77
19	14,785		Camion	46,225	5,70	255	117,50	77,00	201,20	5,57
20	145,519	9,250	Automovil	13,260	2,20	115	104,10	67,90	180,90	2,68
21	135,126	11,225	Automovil	16,535	3,10	170	107,00	69,40	190,40	3,05
22	24,629	10,310	Automovil	18,890	3,10	175	107,50	72,50	200,90	3,33

Fig. 4.9 - Entrada de datos de la Matriz Inicial en el programa R

En el método ***Clúster HJ-BIPLOT*** las variables alfanuméricas no pueden ser usadas de forma directa, por tanto las variables categóricas tienen que ser previamente codificadas a formato numérico, para posteriormente y mediante el programa de software "***MULTBIPLOT***".

	ventas	reventa	tipo	precio	motor	CV	distejes	ancho	largo
Escort	70.2270	7.4250	0	12.0700	2	110	98.4000	67	174.7000
Mustang	113.3690	12.7600	0	21.5600	3.8000	190	101.3000	73.1000	183.2000
Contour	35.0680	8.8350	0	17.0350	2.5000	170	106.5000	69.1000	184.6000
Taurus	245.8150	10.0550	0	17.8850	3	155	108.5000	73	197.6000
Focus	175.6700	NaN	0	12.3150	2	107	103	66.9000	174.8000
Crown Victoria	63.4030	14.2100	0	22.1950	4.6000	200	114.7000	78.2000	212
Explorer	276.7470	16.6400	1	31.9300	4	210	111.6000	70.2000	190.7000
Windstar	155.7870	13.1750	1	21.4100	3	150	120.7000	76.6000	200.9000
Expedition	125.3380	23.5750	1	36.1350	4.6000	240	119	78.7000	204.6000
Ranger	220.6500	7.8500	1	12.0500	2.5000	119	117.5000	69.4000	200.7000
F-Series	540.5610	15.0750	1	26.9350	4.6000	220	138.5000	79.1000	224.5000
Civic	199.6850	9.8500	0	12.8850	1.6000	106	103.2000	67.1000	175.1000
Accord	230.9020	13.2100	0	15.3500	2.3000	135	106.9000	70.3000	188.8000
CR-V	73.2030	17.7100	1	20.5500	2	146	103.2000	68.9000	177.6000
Passport	12.8550	17.5250	1	26.6000	3.2000	205	106.4000	70.4000	178.2000

Fig. 4.10 Tabla de entrada de Datos del Método Clúster HJ-Biplot

4.4.2. Comparación relativa al Algoritmo

Respecto de criterio de clasificación y agrupamiento, ambos métodos se componen de dos fases, perfectamente diferenciadas una de otra, aunque la forma y manera de realizar cada fase los hace totalmente distintos entre sí.

A. Fase de Construcción

En el método ***TWO-STEP***, la construcción del árbol de características, se realiza siguiendo la notación del método BIRCH propuesto por Zhang et al. (1996) y posteriormente implementado en al programa software SPSS (IBM® SPSS® Statistics) por Chiu et al. (2001), ampliamente expuesto en el Capítulo II del Presente Trabajo Fin de Máster.

Se basa en leer los registros de la matriz de partida de uno en uno y en base al criterio de la distancia seleccionado, fusionar un registro con el anterior o formar un nuevo clúster, y para ello dispone de dos tipos de medida de distancia posibles que se pueden seleccionar indistintamente, aunque por defecto se usa la distancia de máxima verosimilitud para permitir la entrada de diferentes tipos de variables, también dispone de la opción para la distancia euclidea, además la selección del número de clusters que hay que formar se

efectúa de forma automática, sin necesidad de definir previamente que número se quiere, aunque permite la posibilidad de forzar de forma manual el número de clúster que se desea.

En el caso del método ***CLUSPLOT***, la construcción no la realiza directamente dicho método, sino que se apoya principalmente, en los algoritmos DAISY y PAM, ampliamente descritos en el Capítulo IV del Presente Trabajo Fin de Máster, en un principio el algoritmo DAISY se emplea para la construcción de una matriz de disimilaridades en vez de utilizar la matriz original utilizando las métricas euclidea o Manhattan, pero dicho algoritmo es absolutamente fundamental cuando la matriz de datos de entrada tiene variables de tipo mixto, para lo cual dicho algoritmo usa la métrica de Gower (Gower, 1966).

Es por ello, que al algoritmo PAM hay que especificarle manualmente todo, desde el número de clúster que se desea para que el algoritmo pueda calcular la mejor manera de agrupar los elementos de la matriz, como si se quiere estandarizar los datos de la matriz original,

En el caso del algoritmo PAM solo puede leer matrices de datos de tipo numérico o matrices de tipo disimilaridad, ya que si la matriz de partida tiene variables de tipo mixto, se hace necesario el paso previo con el algoritmo DAISY para de esa forma cambiar a una matriz de disimilaridades y así permitir al algoritmo PAM que realice la fase de construcción de clusters, pues el objetivo es minimizar la disimilaridad entre individuos del mismo clúster, seleccionando previamente un numero de medoides o representantes de clúster que han sido especificados manualmente al inicio.

En el caso del método ***Clúster HJ-BIPLOT***, y una vez modificadas las variables de tipo mixto a tipo numérico, se ejecuta el HJ-Biplot para obtener las coordenadas mediante el método SVD de valores y vectores singulares, y una vez normalizadas todas las variables, se construye la matriz de distancias al cuadrado entre individuos en base a los criterios de inercia, definidos en el apartado 3.3 del Capítulo III del presente Trabajo Fin de Master, tanto para el cálculo de los individuos situados dentro del mismo clúster como para individuos en diferentes clusters, de forma que los nudos principales del árbol generado, va a contener la inercia total de los elementos que se encuentran dependiendo de dicho nudo

B. Fase de Agrupación

En el método ***TWO-STEP***, la agrupación se lleva a cabo mediante dos diferentes criterios que se tienen que seleccionar de forma manual, el Criterio Bayesiano de Schwarz o BIC (Schwartz, 1978) y el Criterio de Akaike o AIC (Akaike, 1974), descritos en el Capítulo II del presente Trabajo Fin de Máster, y que sirven para ajustar modelos con variables de tipo mixto, aunque ambos criterios buscan minimizar el valor del criterio calculado maximizando el logaritmo de la función máximo verosímil.

En el caso del método ***CLUSPLOT***, la agrupación no la realiza directamente dicho método, sino que se apoya principalmente en el algoritmo PAM, o bien en el algoritmo CLARA si la matriz de datos de partida tiene un volumen grande de registros, y que está ampliamente descrito en el Capítulo IV del Presente Trabajo Fin de Máster

La estandarización de las variables hay que hacerla de forma manual, es decir, especificando en el parámetro correspondiente del algoritmo si se quiere realizar estandarización o no, dependiendo de que o bien las variables hayan sido estandarizadas previamente o bien se esté utilizando la matriz de disimilaridades obtenidas con el algoritmo DAISY.

En el caso del método ***Clúster HJ-BIPLOT***, la agrupación está totalmente definida mediante el criterio jerárquico aglomerativo, y dentro de este cabe la posibilidad de elegir cualquiera de los métodos definidos para los modelos jerárquicos, y que han sido descritos en el Capítulo I del presente Trabajo Fin de Master, pues el objetivo es conseguir clusters totalmente homogéneos en base a la Inercia Dentro calculada, y además conseguir que la separación entre clusters sea la mayor, en base a la Inercia Entre calculada según lo descrito en el Capítulo III del presente Trabajo Fin de Master.

La estandarización se hace por defecto en columnas de la matriz, es decir sobre las variables, ya que aunque las variables tienen que ser numéricas, pueden existir variables nominales, ordinales, binarias y numéricas, y por tanto es necesario que todas estén en la misma métrica y que hablen el mismo lenguaje, aunque siempre se puede cambiar manualmente y elegir uno de los métodos incluidos dependiendo del análisis que se quiera realizar.

4.4.3. Comparación relativa a la Representación Gráfica

El método ***TWO-STEP***, usa el programa SPSS (IBM® SPSS® Statistics), y está basado principalmente en la interpretación de estructuras llevada a cabo por Kaufman y Rousseeuw, donde se puede observar la distribución realizada de los clúster mediante un gráfico de sectores y sus frecuencias, la importancia que han tenido las variables en la realización de dichos clusters, la distribución de cada clúster en base a las variables fijadas para el agrupamiento, tanto las categóricas como las numéricas, y una comparación de todos los cluster y las variables seleccionadas para su agrupación, mediante diagramas de caja para las variables continuas y color y tamaño diferente para las variables categóricas (Kaufman y Rousseeuw, 1987).

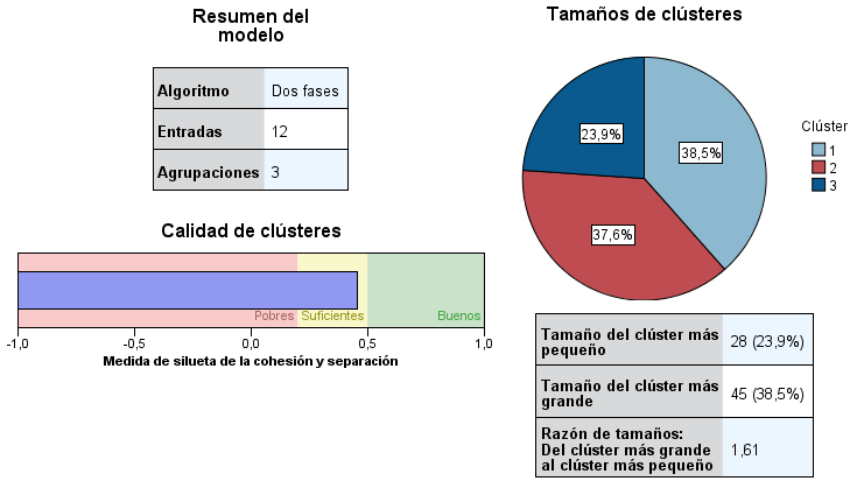


Fig. 4.11 - Gráficos Two-Step, Número de clústers, tamaño y calidad de la agrupación

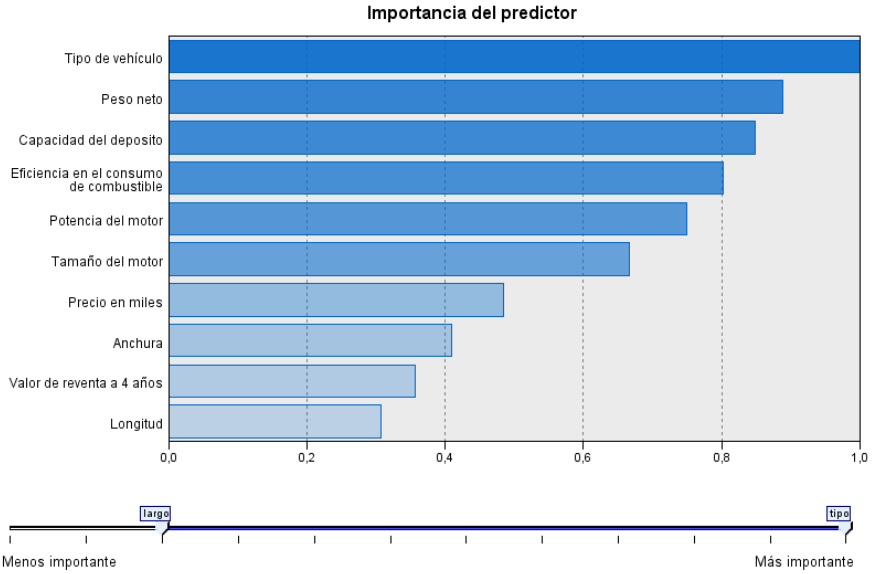


Fig. 4.12 Gráficos Two-Step, Importancia de las variables en la formación de los clusters

El método ***CLUSPLOT***, tal y como se ha indicado en el apartado 4.3 del presente Capítulo IV, se realiza con el software R versión 3.1.2 (copyright © 2014 The R Foundation for Statistical Computing),

Se carga previamente en memoria, la matriz de datos original, idéntica que para el método TWO-STEP y con variables de tipo mixto.

```
> library(cluster)
> coches = read.table(file.choose(), header=TRUE, sep="\t")
> coches
     ventas reventa      tipo precio motor  CV pisada ancho  largo peso_neto
1    16,919  16,360 Automovil 21,500  1,80 140 101,20 67,30 172,40      2,64
2    39,384  19,875 Automovil 28,400  3,20 225 108,10 70,30 192,90      3,52
3    14,114  18,225 Automovil         3,20 225 106,90 70,60 192,00      3,47
4     8,588  29,725 Automovil 42,000  3,50 210 114,60 71,40 196,60      3,85
5    20,397  22,255 Automovil 23,990  1,80 150 102,60 68,20 178,00      3,00
6    18,780  23,555 Automovil 33,950  2,80 200 108,70 76,10 192,00      3,56
7     1,380  39,000 Automovil 62,000  4,20 310 113,00 74,00 198,20      3,90
8    19,747         Automovil 26,990  2,50 170 107,30 68,40 176,00      3,18
9     9,231  28,675 Automovil 33,400  2,80 193 107,30 68,50 176,00      3,20
10   17,527  36,125 Automovil 38,900  2,80 193 111,40 70,90 188,00      3,47
11   91,561  12,475 Automovil 21,975  3,10 175 109,00 72,70 194,60      3,37
12   39,350  13,740 Automovil 25,300  3,80 240 109,00 72,70 196,20      3,54
13   27,851  20,190 Automovil 31,965  3,80 205 113,80 74,70 206,80      3,78
```

Fig. 4.13 - Salida gráfica del programa R – carga de la Matriz Inicial de empleados

A continuación se realiza la ejecución del algoritmo DAISY, para calcular la nueva matriz de disimilaridades, con variables de tipo mixto, y estandarizando las variables previamente,

>coches.diss <- daisy (coches,stand=TRUE, metric="Gower")

A continuación se ejecuta el algoritmo PAM, el cual realizará las dos fases que hemos descrito anteriormente y correspondientes a la construcción y agrupación de los datos de la nueva matriz calculada, indicando cual es la matriz entrada y que corresponde a la matriz de salida previa del algoritmo DAISY,

Además se le indica que efectivamente la matriz de datos de entrada corresponde a una matriz de disimilaridades y por último se le indica que queremos 3 clúster, ya que en pruebas realizadas sobre mas agrupaciones el gráfico resultante no era el mas indicado

>coches.clus <- pam (coches.diss,3,diss=TRUE)

Finalmente ejecutamos el algoritmo CLUSPLOT, indicándole la matriz de disimilaridades calculada previamente por el algoritmo DAISY y también la matriz que ha calculado el algoritmo PAM y que le va a permitir realizar la representación gráfica de los puntos de la matriz, pues tal y como se ha descrito en el apartado 4.3 del presente Capítulo IV, le indica a que clúster pertenece cada punto de la matriz de dismilaridades calculada, y de esta forma dibujar los clúster que contienen a los individuos de la matriz original, y en la que se puede observar que no consigue diferenciar los 3 clusters fijados, en las pruebas llevadas a cabo tanto con 2 clusters, como con 4 e incluso con 5, no se ha conseguido la división total de los clusters.

>clusplot (coches.diss,coches.clus,lines=1,diss=TRUE,shade=TRUE, col.p="black",color=TRUE,labels=2)

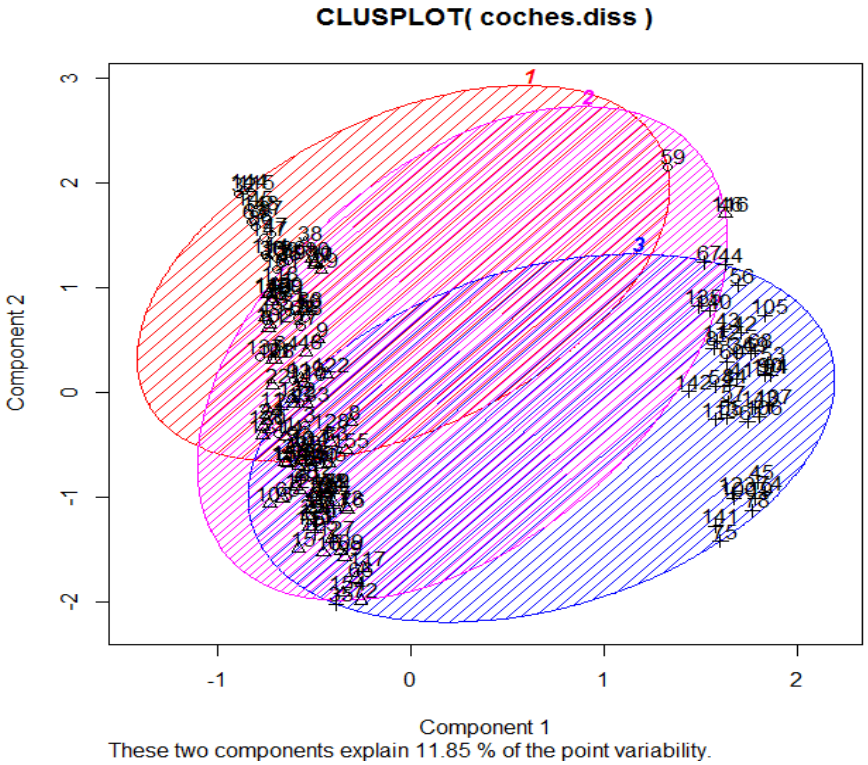


Fig. 4.14 Salida gráfica del programa R, Agrupación en 3 clúster de la matriz inicial

El método de ***Clúster HJ-BIPLOT***, se puede realizar o bien para las variables continuas o bien para cada variable nominal, por separado, el siguiente grafico se ha marcado la Distancia de Ward y el Método Jerárquico.

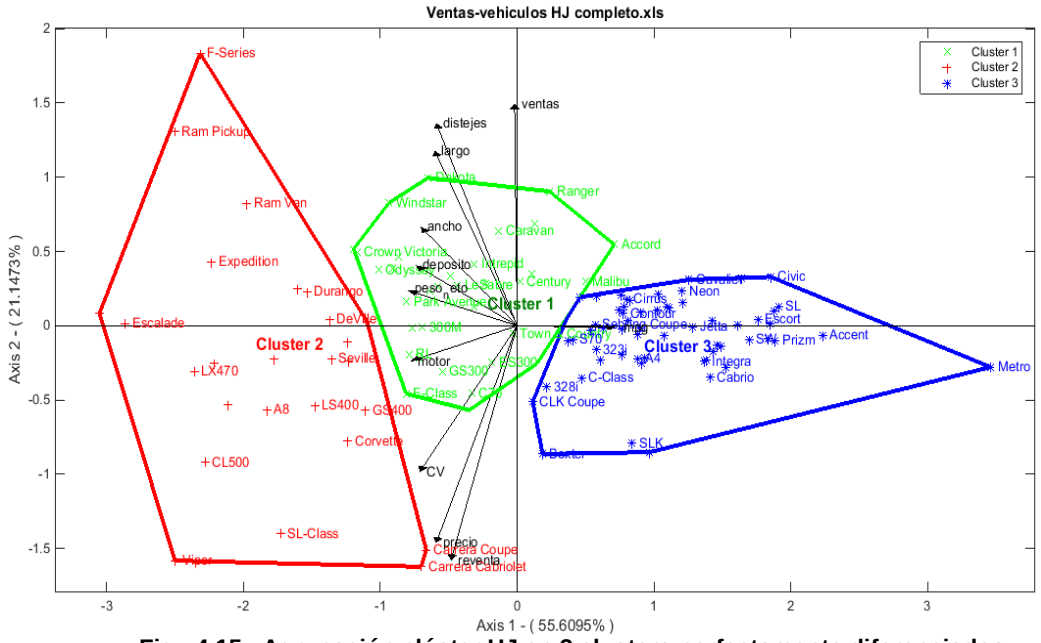


Fig. 4.15 Agrupación clúster HJ en 3 clusters perfectamente diferenciados

4.4.4. Limitaciones CLUSPLOT

Las limitaciones encontradas en la comparación de los algoritmos son las siguientes:

1) En primer lugar hay que señalar que el CLUSPLOT solo es un algoritmo de representación gráfica y por tanto carece del resto de opciones de cálculo sobre los datos de la matriz inicial que tiene el TWO-STEP y el HJ-BIPLOT.

2) La introducción de todos los valores que hay que proporcionar al algoritmo CLUSPLOT y al algoritmo Clúster HJ-BIPLOT son de forma manual, así como también a los algoritmos previos y necesarios para su ejecución, frente a la posibilidad de realizar de forma autónoma todas las operaciones en el algoritmo TWO-STEP.

3) CLUSPLOT tiene la necesidad de usar otros algoritmos previos a su ejecución, tanto para crear una matriz de disimilaridades si hay datos de tipo mixto como para crear la matriz de vectores que le pueda servir al CLUSPLOT para dibujar en el plano todos los puntos de la matriz original, mientras que el TWO-STEP o el Clúster del HJ-BIPLOT lo tiene de forma integrada.

4) CLUSPLOT no tiene la posibilidad de realizar diferentes opciones de agrupación en base a las variables, mientras que el TWO-STEP si tiene

dicha facilidad, tanto para variables categóricas como numéricas, mediante diferentes métodos, así como también el método de Clúster del HJ-BIPLOT permite realizar diferentes agrupaciones.

5) El método TWO-STEP usa la distancia de máxima verosimilitud cuando tiene datos de tipo categórico, mientras que el algoritmo PAM previo al CLUSPLOT usa la distancia de Gower y el algoritmo Clúster del HJ-BIPLOT tiene que modificar en primer lugar todas las variables categóricas a tipo numérico y a continuación crear la matriz de coordenadas del HJ-BIPLOT, y finalmente usar el criterio de Inercia maximizando la Inercia Entre clusters y al mismo tiempo minimizando la Inercia Dentro de cada clúster.

6) En cuanto a la representación gráfica lo realizan de diferentes maneras, pero es evidente la gran representación gráfica que proporciona el método Clúster del HJ-BIPLOT frente al método TWO-STEP y mucho mas respecto al método CLUSPLOT que no consigue dividir los tres clusters fijados

4.4.5. Resumen Comparativo de los 3 Métodos

	TWO-STEP	CLÚSTER HJ-BIPLOT	CLUSPLOT
Software	IBM SPSS	MultBiplot	R
Variables de Entrada	MIXTO	NUMÉRICO	MIXTO
Ejecución autónoma	SI	NO	NO
Big Data	SI	NO	NO
Medida de Distancia	COEFICIENTE VEROSIMILITUD	METODO INERCIA	COEFICIENTE GROWER
Número de Clusters	AUTOMATICO	MANUAL	MANUAL
Representación Gráfica	BUENA	EXCELENTE	NORMAL

Fig. Fuente Propia: Tabla de diferencias entre los tres métodos

Conclusiones

1) El análisis de clúster como forma de clasificación, es una técnica para encontrar patrones, relaciones o descubrir conocimiento en Bases de Datos sobre unidades taxonómicas respecto de una serie de variables.

2) La literatura sobre el análisis de clúster refleja terminologías, métodos y aproximaciones contradictorias, debido a que se han venido creando al amparo de distintas ramas de la ciencia, por lo que están impregnadas de ciertos sesgos que proceden de esas disciplinas, ya que cada disciplina de la ciencia tiene sus preferencias para la construcción de grupos.

3) Distintos procedimientos generan soluciones diferentes sobre la misma matriz de datos, lo que hace necesario la existencia de distintas técnicas para poder determinar que método puede generar grupos que sean homogéneos de forma natural sobre la matriz de datos de partida.

4) El método TWO-STEP y el método CLUSPLOT permiten la entrada de cualquier tipo de variable, tanto numéricas como categóricas, y estas últimas en cualquiera de sus posibilidades, mientras que el Método Clúster HJ- BIPLOT solo permite variables numéricas, teniendo que ser posteriormente tipificadas esas variables numéricas a variables nominales, ordinales o binarias.

5) El método CLUSPLOT tiene la necesidad de usar otros algoritmos previos a su ejecución, tanto para crear una matriz de disimilaridades si hay datos de tipo mixto como para crear la matriz de vectores que le pueda servir al CLUSPLOT para dibujar en el plano todos los puntos de la matriz original, así como el Clúster del HJ-BIPLOT necesita del algorimto del HJ-Biplot para obtener la matriz de coordenadas de valores singulares para que el criterio de inercia pueda ejecutarse respecto de dichas coordenadas, mientras que el TWO-STEP lo tiene de forma integrada.

6) El método TWO-STEP usa la distancia de máxima verosimilitud cuando tiene datos de tipo mixto, mientras que el Método Clúster del HJ-BIPLOT usa el criterio de Inercia maximizando la Inercia Entre clusters y al mismo tiempo minimizando la Inercia Dentro de cada clúster, mientras que método CLUSPLOT necesita del algoritmo PAM previamente para realizar dicho cálculo, que usa el coeficiente de Gower para calcular las distancias, siendo este coeficiente el más adecuado para variables de tipo mixto (Gower, 1971), bien es verdad que el método cluster del

HJ-Biplot crea inercia totales en cada nudo sobre todas las variables lo cual le hace mas optimo que respecto al criterio de máxima verosimilitud que parte de ciertas hipótesis que no siempre son correctas ni reales, seria necesario probar una interface capaz de trabajar con el coeficiente de Gower para crear las inercias dentro y entre los clúster.

7) El Método TWO-STEP gestiona de forma automática el número de clusters necesarios para el agrupamiento, mientras que los métodos de clúster del HJ-BIPLOT y el CLUSPLOT tienen que hacerlo de forma manual.

8) CLUSPLOT no tiene la posibilidad de realizar diferentes opciones de agrupación en base a las variables que se quieran, seria necesario ejecutar repetidamente el algoritmo con distintas variables cada vez, mientras que el TWO-STEP y el método de Clúster del HJ-BIPLOT si tienen dicha facilidad.

9) El método Clúster del HJ-BIPLOT proporciona una gran representación gráfica frente al método TWO-STEP y al método CLUSPLOT, consiguiendo la mayor calidad de representación tanto para los individuos como para las variables, por lo que su interpretación se hace más evidente al poder relacionar filas y columnas en el mismo gráfico.

Bibliografía

Abbott, R. D., & Perkins, D. (1978). Development and Construct Validation of a Set of Student Rating-of-Instruction Items. *Educational and Psychological Measurement*, *38*(4), 1069–1075. doi:10.1177/001316447803800427

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical funcions*. National Bureau of Standards, Applied Mathematics. Retrieved from http://people.math.sfu.ca/~cbm/aands/abramowitz_and_stegun.pdf

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, *30*(1), 9–14. doi:10.1007/BF02480194

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332.

Akter, A., & Ahmed, S. (2015). Potentiality of Rainwater Harvesting for an Urban Community in Bangladesh. *Journal of Hydrology*, *528*, 84–93. doi:10.1016/j.jhydrol.2015.06.017

Aldenderfer, M. S., & Blashfield, R. K. (1988). *Cluster Analysis Series: Quantitative Applications in the Social Sciences*. *Sage University Paper*. London: Sage.

Alho, A. R., & de Abreu e Silva, J. (2015). Utilizing urban form characteristics in urban logistics analysis: a case study in Lisbon, Portugal. *Journal of Transport Geography*, *42*, 57–71. doi:10.1016/j.jtrangeo.2014.11.002

Altas, D., Kubas, A., & Sezen, J. (2013). Analysis of environmental sensitivity in thrace region through Two Step Cluster. *Trakia Journal of Sciences*, *3*, 318–329. Retrieved from http://tru.uni-sz.bg/tsj/N3, Vol.11, 2013/D.Altas.pdf

Andersen, B., Silva, M., & Levy, C. (2013). *How business can work with universities to generate knowledge and drive innovation*. Retrieved from http://www.mbsportal.bl.uk/taster/subjareas/techinnov/bic/152858collaborate13.pdf

Arrighetti, A., & Ninni, A. (2014). *La trasformazione "silenziosa."* Retrieved from http://dspace-unipr.cineca.it/bitstream/1889/2565/1/La_trasformazione_silenziosa-A Arrighetti_A Ninni.pdf

Bacher, J., Wenzig, K., & Vogler, M. (2001). SPSS TwoStep Cluster - A First Evaluation.

Balzarini, M., Macchiavelli, R., & Casanoves, F. (2005). Aplicaciones de modelos mixtos en agricultura y foresteria. In *Curso Internacional de Aplicaciones de Modelos Mixtos* (p. 189). CATIE. Turrialba, Costa Rica.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*, 803–821.

Benzecri, J. (1973). *L'Analyse des données, Toumo I: La Taxinomie*. Paris: Dunod.

Bisquerra, R. (1989). *Introducción conceptual al análisis multivariable. Un enfoque informático con los paquetes SPSS-X, BMDP, LISREL y SPAD. Vol. 2*. Barcelona: PPU.

Bradu, D., & Gabriel, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, *20*, 47–68.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*, 325–349.

Budeva, D., & Mullen, M. (2014). International market segmentation: Economics, national culture and time. *European Journal of Marketing*, *48*(7/8), 1209–1238. doi:10.1108/EJM-07-2010-0394

Cardona, I. (2014). La innovació oberta al sector financer espanyol: una anàlisi dels intermediaris financers i asseguradors. Retrieved from https://riunet.upv.es/handle/10251/34732

Carmichael, J. W., & Sneath, P. H. A. (1969). Taxometric maps. *Systematic Zoology*, *18*, 402–415.

Carmichael, L. E., & Bruner, D. W. (1968). Characteristics of a newly-recognized species of Brucella responsible for infectious canine abortions. *The Cornell Veterinarian*, *48*(4), 579–92. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5693645

Cattell, R. (1944). A note on correlation clusters and cluster search methods. *Psychometrika*, *9*, 169–184.

Cattell, R., Coulter, M., & Tsujioka, B. (1966). *The taxonomic recognition of types and functional emergents. Handbook of Multivariate Experimental Psychology* (Catell, RB.). Chicago: Rand, McNally.

Charrad, M., Lechevallier, Y., Ben Ahmed, M., & Saporta, G. (2010). On the Number of Clusters in Block Clustering Algorithms. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference*.

Chávez, D. E., Miranda, I., Varela, M. N., & Fernandez, L. (2010). Utilización del análisis de cluster con variables mixtas. *Revista Investigacion Operacional*, *30*(3), 209–216.

Cheeseman, P., & Stutz, J. (1996). Bayesian Classification (Autoclass):Theory and Results. *Advances in Knowledge Discovery and Data Mining*, 153–180.

Chen, H., Gnanadesikan, R., & Kettering, J. R. (1974). Statistical methods for grouping corporations. *Sankhya Ser B*, *36*, 1–28.

Chin, Y. R., & Choi, K. (2015). Suicide Attempts and Associated Factors in Male and Female Korean Adolescents A Population-Based Cross-Sectional Survey. *Community Mental Health Journal*. doi:10.1007/s10597-015-9856-6

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *International conference on Knowledge discovery and data mining* (pp. 263–268). San Francisco, USA.

Clark, P. F. (1952). An extension of the coefficient of divergence for use with multiple characters (pp. 61–64). Copeia.

Clements, F. E., Schenck, S. M., & Brown, T. K. (1926). A new objective methd for showing special relationships. *American Anthropologist*, *28*, 585–604.

Correa, J. C., & Gonzalez, N. (2002). *Gráficos Estadísticos con R*. Retrieved from http://cran.r-project.org/doc/contrib/grafi3.pdf

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin*, *50*(6), 456–473.

Cuadras, C. M. (1986). *Problemas de Probabilidad y Estadistica. Vol. 2*. Barcelona: PPU.

Cuadras, C. M. (1989). Distancias Estadisticas. *Estadistica Española*, *30*, 295–378.

Cuadras, C. M. (2014). *Nuevos Métodos de Análisis Multivariante*. CMC Editions.

Czekanowski, J. (1911). Objectiv kriterien in der ethnologie. *Korrespondenzblatt Der Deutschen Gesselschaft Fur Anthropologie, Ethnologie Und Urgeschichte 47: Pp 1-5*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data using the EM algorithm. *Journal of the Royal Statistical Society*, *39*(B).

Dhillon, A., & Godfrey, A. R. (2013). Using routinely gathered data to empower locally led health improvements. *London Journal of Primary Care*, *5*, 70–3. Retrieved from http://www.radcliffehealth.com/sites/radcliffehealth.com/files/ljpc_articles/05_dhillon_ljpc5_1d2.pdf

Dhillon, I., & Modha, D. (200AD). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*(1), 143–175.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26:*, 297–302.

Diday, E. (1972). Optimisation en classification automatique et reconnaissance es formes. *Note Scientifique IRIA*, *6*.

Driver, H. E. (1941). Girl's puberty rites in western North America. *University of California Anthropological Records*, *6*, 21–90.

Driver, H. E., & Kroeber, A. L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Achaeology and Ethnology 31:211-56*.

Driver, H. E., & Schuessler, K. F. (1957). Factor Analysis of Ethnographic Data. *American Anthropologist*, *59*(4), 655–663. doi:10.1525/aa.1957.59.4.02a00080

DuBois, J. M., Chibnall, J. T., Tait, R. C., Vander Wal, J. S., Baldwin, K. A., Antes, A. L., & Mumford, M. D. (2015). Professional Decision-Making in Research (PDR): The Validity of a New Measure. *Science and Engineering Ethics*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26071940

Dutt, A., Aghabozrgi, S., Akmal, M. B. I., & Mahroeian, H. (2015). Clustering Algorithms Applied in Educational Data Mining. *International Journal of Information and Electronics Engineering*, *5*(2). Retrieved from http://www.ijiee.org/vol5/513-F1002.pdf

Eddy, W. F. (1977). A New Convex Hull Algorithm for Planar Sets. *ACM Transactions on Mathematical Software*, *3*(4), 398–403. doi:10.1145/355759.355766

Ester, M., Kriegel, H. P., Sander, S., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Miining (KDD-96)* (pp. 226–231). Menlo Park, CA: In E. Simoudis, J. Han and U. Fayyad - AAAI, AAAI Press.

Estivill-Castro, V., & Yang, J. A. (2000). A fast and robust general purpose clustering algorithm. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 208–218).

Everitt, B. s. (2001). *Cluster Analysis*. Edward Arnold.

Fisher, W. D. (1968). *Clustering and aggregation in economics*. John Hopkins University Press, Baltimore.

Fleiss, & Zubin, J. (1969). On the methods ant theory of clustering. *Multivariate Behavioral Research*, *4*, 235–250.

Forgy, E. W. (1965). Cluster Analysis of multivariate data: Efficiency vs Interpretability of classification. *Biometrics*, *21*, 768.

Fortin, M. ., Dale, M., & Hoef, J. (2002). Spatial analysis in ecology. *Encyclopedia of Environmetrics*, *4*, 2051–2058.

Fraley, C., & Raftery, A. E. (1998). *How many clusters? Which clustering method? Answers Via model-based Clúster Analysis*. *THE COMPUTER JOURNAL, Vol.*

41, No. 8. Department of Statistics University of Washington. Retrieved from http://www.stat.washington.edu/raftery/Research/PDF/fraley1998.pdf

Francisco, D. (2012). Poor mental health symptoms among Romanian employees. A Two-Step Cluster analysis. *Procedia - Social and Behavioral Sciences*, *33*, 293–297. doi:10.1016/j.sbspro.2012.01.130

Frutos, E. B. (2014). *Análisis de Datos Acoplados: Modelo T3-PCA*. Tesis Doctoral. Universidad de Salamanca.

Gabriel, K. R. (1971). The Biplot-graphic display of matrices with application to Principal Component Analysis. *Biometrika*, *58*, 453–467.

Gabriel, K. R. (1995). Biplot display of multivariate categorical data, with comments on multiple correspondence analysis. *W.J. Krzanowski Ed. Recent Advances on Descriptive Multivariate Analysis*, 190–226.

Gabriel, K. R., & Odoroff, C. L. (1990). Biplot in biomedical research. *Statistics in Medecine*, *9*(469-485).

Galindo, M. P. (1986). Una Alternativa de Representación simultánea: HJ-Biplot. *Questió*, *10*(1), 13–23.

Galindo, M. P., & Cuadras, C. M. (1986). Una extensión del método Biplot a su relación con otras técnicas. *Publicación de Bioestadística Y Biomatemática*, *17*.

Galindo, M. P., Lorente, F., Romo, A., & Martín, M. (1987). Inspección de matrices de datos multivariantes utilizando el método HJ-Biplot: Aplicación a un problema médico. *Cuadernos de Bioestadística Y Sus Aplicaciones Informáticas*, *5*(1), 88–101.

Galindo, M. P., Vicente-Villardon, J., Vicente-Tavera, S., Barrera, I., Fernandez, M. J., & Martin, A. (1993). Analisis gráfico y descripción estructural de la variabilidad de cultivos en Castilla-León. *Investigación Agraria. Economía*, *8*(3), 17–329.

Gallardo, J. A., Gutierrez, R., Gonzalez, A., & Torres, F. (1994). *Técnicas de Análisis de datos multivariable. Tratamiento Computacional*. (U. de G. (Facultad de Ciencias), Ed.). Granada, Spain. Retrieved from http://www.ugr.es/~gallardo/

Garcia-Talegón, J., Vicente, M. A., Molina-Ballesteros, E., & Vicente-Tavera, S. (1998). Determination of the origin and evolution of building stones as a function of their chemical composition using the inertia criterion based on an HJ-Biplot. *Chemical Geology*, *153*, 37–51.

Giulano, G., Rhoads, M., & Chakrabarti, S. (2014). New data, new applications: using transportation system data for regional monitoring. *Transport Research Arena*. Retrieved from http://tra2014.traconference.eu/papers/pdfs/TRA2014_Fpaper_18601.pdf

Gordon, A. D. (1990). *Cluster Classification*. New York: Wiley.

Govaert, G. (1983). *Classification croisée. Thèse d'état*. Université Paris 6, France.

Govaert, G. (1995). Simultaneous clustering of rows and columns. *Control and Cybernetics*, 437–458.

Gower, J. C. (1966). Some distance properties of latent roo and vector mehods in multivariate analysis. *Biometrika1*, *53*, 315–328.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857–874. Retrieved from http://www.jstor.org/stable/2528823?seq=1#page_scan_tab_contents

Gower, J. C. (1985). Measures of similarity, dissimilarity and distance. In *Encyclopedia of Statistics. Vol 5*. Johnson, NL, Kotz, S, Read CB (eds). Wiley. New York.

Gower, J. C. (1992). Generalized Biplots. *Biometrika*, *79*(3), 475–493.

Gower, J. C., & Hand, D. (1996). *Biplots, Chapman and Hall*. London.

Gower, J. C., & Harding, S. (1988). Nonlinear biplots. *Biometrika*, *75*, 445–455.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, *3*, 5–48. Retrieved from http://fitelson.org/coherence/gower_legendre.pdf

Greenacre, M. (1984). *Theory and applications of correspondence analysis*. *Academic press*. London.

Griffin, B., Sherman, K. A., Jones, M., & Bayl-Smith, P. (2014). The Clustering of Health Behaviours in Older Australians and its Association with Physical and Psychological Status, and Sociodemographic Indicators. *Annals of Behavioral Medicine*, *48*(2), 205–214. doi:10.1007/s12160-014-9589-8

Guha, S., Rastogi, R., & Shim, K. (1999). *ROCK: A Robust clustering Algorihm for Categorical Attributes*.

Gunten, L. von, Húmbelin, O., & Fritschi, T. prof. (2014). Administrative Data: Benefits and Challenges for Social Security Research. *Berne University of Applied Science*. Retrieved from http://ecpr.eu/filestore/paperproposal/c37024ae-ed80-4fc8-b8b2-0c22a14a648f.pdf

Hair, J. F., Anderson, R. E., Tatham, R. R., & Black, W. C. (1998). *Multivariate data analysis (5th ed.)*. Upper Saddle River, NJ: Prentice Hall. Retrieved from http://library.wur.nl/WebQuery/clc/1809603

Hamann, U. (1961). Merkmalsbestand und verwandtschaftsbeziehungen der farinosae ein beitrag zum system der monokotyledonen. *Willdenowia*, *2*, 639–768.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Harman, H. H. (1967). *Modern Factor Analysis*. Chicago: University of Chicago Press.

Hartigan, J. (1975). *Clustering Algoithms*. New York: Jhon Wiley and Sons.

Harvey, C. (2014). Measuring Streetscape Design for Livability Using Spatial Data and Methods. *Graduate College Dissertations and Theses*. Retrieved from http://scholarworks.uvm.edu/graddis/268

Hauser, D. D. W., Tobin, E. D., Feifel, K. M., Shah, V., & Pietri, D. M. (2015). Disciplinary reporting affects the interpretation of climate change impacts in global oceans. *Global Change Biology*, n/a–n/a. doi:10.1111/gcb.12978

Huang, Z. (1997). A fast custering algorithm to cluster very large categorical data sets. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.83&rep=rep1&type=pdf

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3).

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary an evaluation. *Biological Reviews*, *57*, 669–689.

IBM Corporation. (1989). SPSS - Modo Bietápico. Retrieved from http://www-01.ibm.com/support/knowledgecenter/SS3RA7_16.0.0/com.ibm.spss.modeler.help/clementine/clusternode_general.htm?lang=es

IBM Corporation. (2013). IBM Knowledge Center. Retrieved from http://www-01.ibm.com/support/knowledgecenter/SSLVMB_22.0.0/com.ibm.spss.statistics.tut/spss/tutorials/data_files.htm?lang=es

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise Sci. Nat.*, *44*, 223–270.

Jain, A., & Dubes, R. (1980). Clustering methodologies in exploratory data analysis. *Advances in Computers*, *19*, 113–228.

Jain, A., & Dubes, R. (1988). *Algorithms for clustering data*. Prentice Hall. Retrieved from http://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Survey. *ACM Computing Surveys*, *31*(3).

Jancey, R. C. (1966). Multidimensional group analysis. *Austral J. Botany*, *14*, 127–130.

Jiang, M., Tseng, S., & Su, C. (2001). Two-phase clustering process for outlier detection.

Kamber, M., & Han, J. (2006). *Data Mining. Concepts and Techniques*. San Francisco (USA): Morgan Kaufmann Publishers.

Kaufman, L., & Rousseeuw, P. (1987). Clustering by Means of Medoids. *Statistical Data Analysis Based on the L1 Norm and Related Methods*, 405–416.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data*. John Wiley and sons Inc. doi:10.1002/9780470316801

King, B. (1967). Step-wise Clustering Procedures. *J. Am. Stat. Assoc.*, *69*, pp. 86–101.

Klimek, S. (1935). The structure of California Indian culture. *University of California Publications in American Achaeology and Ethnology 1*, *37*, 1–70.

Kulczynski, S. (1927). Die Pflanzenassoziationen der Pieninem. *Bull. Int. Acad. Polonaise Sci. et Lettres. Classe Sci. Mth. et Nat. Serie B, Suppl II*, 57–203.

Lance, G. N., & Williams, W. T. (1966). Computer programs for hierarchical polythetic classification. *Computer Journal1*, *9*, 64–64.

Lebart, L., Morineau, A., & Fenelon, J. P. (1983). *Tratamiento estadístico de datos*. Marcombo.

Legendre, L., & Legendre, P. (1979). *Ecologie Numerique*. Paris: Masson.

Legendre, P., Legendre, L., & Dallot, S. (1985). Succession of species within a community chronological clustering with applications to marine and freshwath zooplankton. *American Naturalist*, *125*, 257–258.

Leiva-Valdebenito, S., & Torres-Avilés, F. J. (2010). Una revisión de los algoritmos de partición mas comunes en el análisis de conglomerados: un estudio comparativo. *Revista Colombiana de Estadística*, *33*(2), 321–339.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, *2(1)*, 49–55.

Mártinez, E. R. (1984). *Aspectos teóricos del Análisis de Cluster y Aplicación a la caracterización del electorado potencial de un partido. En J.J. Carrión (Ed), Introducción a las técnicas de Análisis Multivariable aplicadas a las ciencias sociales*. CIS.

Martins, V. V. (2003). *Desarrollo de un Sistema para Mineria de datos basado en los métodos Biplot*. Tesis Doctoral. Universidad de Salamanca.

McQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1* (pp. 281–297).

Miranda, I., & Torres, V. (1998). Coeficientes de similaridad para variables mixtas I. *Rev. Proteccion Veg*, *13*, 127–131.

Mirkin, B. (1996). *Mathematical classification and clustering*. Dordrecht: Kluwer.

Molina, I. P. (2008). *Análisis Cluster*. Universidad Carlos III, Madrid. Retrieved from http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/TecnicasInvestigacion/SlidesAClusterEstudi0809.pdf

Molinero, L. L. (2002). El metodo Bayesiano en la investigacion médica. *Liga Española Para La Lucha Contra La Hipertension arterial2*, 3–10.

Molinero, L. L. (2003). ¿Que es el metodo de estimacion de maxima verosimilitud y como se interpreta? *Liga Española Para La Lucha Contra La Hipertension Arterial*.

Motulsky, H., & Christopoulos, A. (2003). Fitting models to biological data using linear and nonlinear regresion. *Biological Sciences*, 351 p. Retrieved from www.graphpad.com

Mucha, H. J., & Sofyan, H. (2000). *Cluster Analysis, Sonderforschungsbereich 373, Discussion Paper 2000-49*. Humboldt University at Berlin.

Murtagh, F. (1984). A survey of recent advances in hierarchical clustering algorithms which use cluster centers. *Journal of Computational*, *26*, 354–359.

Nadif, M., & Govaert, G. (2005). Block clustering of contingency table and mixture model. In *Intelligent Data Analysis IDA'2005, LNCS 3646* (pp. 249–259). Heidelberg, Springer-Verlag Berlin.

Nadotti, L. L., & Constantin, L.-G. (2014). Catastrophe Bonds Structures at European Level – A Cluster Analysis Approach. *The Romanian Economic Journal*, *54*. Retrieved from http://www.rejournal.eu/sites/rejournal.versatech.ro/files/articole/2014-12-23/3177/8nadotticonstantin.pdf

Needham, E. M., & Sparck, K. J. (1964). KEYWORDS AND CLUMPS. *Journal of Documentation*, *20*(1), 5–15. doi:10.1108/eb026337

Needham, R. ., & Sparck Jones, K. (1964). Theory of Subject Analysis. *Journal of Documentation*, *20*(5-15).

Needham, R. ., & Sparck Jones, K. (1967). Automatic term classifications and retrieval. In *First Cranfield International Conference on Mechanised Information Storage and Retrieval Systems*.

Needham, R. ., & Sparck Jones, K. (1968). nformation Storage and Retrieval, *4*(91-100).

Needham, R. M., & Sparck, K. J. (1968). Automatic term classifications and retrieval. *Information Storage and Retrieval*, *4*(2), 91–100. doi:10.1016/0020-0271(68)90013-2

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th VLDB Conference* (pp. 144–155). Santiago de Chile.

Nikolaj, P. B., Klaudi, K. K., Minbaeva, D., & Peter, N. M. (2015). Prioritisation of marketing investments in different types of marketing functions. *Danish Journal of Management & Business*, *79*(1). Retrieved from https://www.djoef-forlag.dk/services/djm/full/DJM_vol79_no_1_2015.pdf#page=45

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bull Jnp Soc Sci Fish*, *22*, 526–530.

Opitz, B., & Hofmann, J. (2015). Concurrence of rule- and similarity-based mechanisms in artificial grammar learning. *Cognitive Psychology*, *77*, 77–99. doi:10.1016/j.cogpsych.2015.02.003

Osuna, Z. M. (2006). *Contribuciones al Análisis de Datos Textuales*. Tesis Doctoral. Universidad de Salamanca.

Overall, J. E., & Klett, C. J. (1972). *Applied Multivariate Analysis*. New York: McGraw-Hill.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika*, *18*, 337–343.

Pedraz, C., & Galindo, P. (1986). Study of socio-clultural factors influencing the decision to breast-feed instead of bottle-feed. *Arch. Pediat.*, *36*, 469–477.

Pedret, R. (1986). *Técnicas cuantitativas al servicio de marketing: Métodos descriptivos de análisis multivariable*. (Tesis Doctoral) - Facultad de Ciencias Económicas (Universidad de Barcelona).

Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw-Hill.

Perez-Mellado, V., & Galindo, P. (1986). Biplot Graphic display of Iberian and North African populations of podarcis. *Rocek Z (ed), Studies in Herpetology*, 197–200.

Pison, G., Rousseeuw, P., & Struyf, A. (1999). Displaying a clustering with CLUSPLOT. *Computational Statistics & Data Analysis*, *30*, 381–392.

Posada, S. L., Zoot, M. S., & Romero, R. N. (2007). Comparación de modelos matemáticos: una aplicación en la evaluación de alimentos para animales. *Revi. Col. de Ciencias Pec.*, *20*, 141–148.

R Foundation for Statistical. (2014). R: The R Project for Statistical Computing. Retrieved from http://www.r-project.org/

Rogers, D. J., & Tanimoto, T. . (1960). A computer program for classifing plants. *Science*, *132*, 1115–1118.

Rohlf, F. J. (1970). Adaptative hierarchical clustering schemes. *Systematic Zool*, *19*, 58–82.

Rokach, L., & Maimon, O. (2005). *The data mining and Knowledge discovery handbook*. (Springer, Ed.). N. Retrieved from http://www.ise.bgu.ac.il/faculty/liorr/

Rosenberg, S. (1982). *The method of sorting in multivariate research with applications selected from cognitive psychology and person perception*. Erlbaum, Hillsdale, NJ,: In: Hirschberg, N., Humphreys, L.G. (Eds.), Multivariate Applications in the Social Sciences.

Rousseeuw, P. (1987). Silhouettes, a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*.

Rousseeuw, P., & Ruts, I. (1997). The Baglot: a bivariate box-and-whiskers plot, submitted for publication.

Rousseeuw, P., & Van Driessen, K. (1997). *A fast algorithm for the minimum covariance determinant estimator, submited for publication*.

Ruspini, E. H. (1970). Numerical Methods for fuzzy clustering. *Inform Sci.*, *2*, 319–350.

Russell, P. F., & Rao, T. R. (1940). On habitat and association of species of anopheline larvae in south-eastern Madras. *J. Malar Inst. India*, *3*, 153–178.

Satish, S. M., & Bharadhwaj, S. (2010). Information search behaviour among new car buyers: A two-step cluster analysis. *IIMB Management Review*, *22*(1-2), 5–15. doi:10.1016/j.iimb.2010.03.005

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. Retrieved from http://qwone.com/~jason/trg/papers/schwarz-dimension-78.pdf

Scoltock, J. (1982). A survey of the literature of cluster analysis. *Computer Journal*, *25*, 130–134.

Sigmund, E., Sigmundová, D., Snoblová, R., & Gecková, A. M. (2014). ActiTrainer-determined segmented moderate-to-vigorous physical activity patterns among normal-weight and overweight-to-obese Czech schoolchildren. *European Journal of Pediatrics*, *173*(3), 321–9. doi:10.1007/s00431-013-2158-5

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical Taxonomy: The principles and practice of numerical classification*. San Francisco. USA: Freeman W.H. and Co. (573p.).

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendograms by objective methods. *Taxon*, *11*, 33–40.

Sokal, R., & Sneath, P. (1963). *Principles of numerical taxonomy*. San Francisco (USA): Freeman W.H. and Co. (pp. 359).

Sorensen, T. (1948). A mehod of stablishing groups of equal amplitude in plant sociology based on similarity of species content and its aplication to analyses of the vegetation on Danish commons. *Biological SKR*, *5*, 1–34.

Stanton, B. (2015). Teachers' Patterns of Implementation of an Evidence-Based Intervention and Their Impact on Student Outcomes: Results from a Nationwide Dissemination over 24-Months Follow-Up. *AIDS and Behavior*. doi:10.1007/s10461-015-1110-2

Stephenson, W. (1936). Introduction of inverted factor analysis with some applications to studies in orexia. *Journal of Educational Psychology*, *5*, 353–367.

Stranak, Z., Semberova, J., Barrington, K., O'Donnell, C., Marlow, N., Naulaers, G., & Dempsey, E. (2014). International survey on diagnosis and management of hypotension in extremely preterm babies. *European Journal of Pediatrics*, *173*(6), 793–8. doi:10.1007/s00431-013-2251-9

Strickland, J. (2015). K-Means Clustering using R. Retrieved June 27, 2015, from https://www.linkedin.com/pulse/k-means-clustering-using-r-jeffrey-strickland-ph-d-cmsp

Struyf, A., Hubert, M., & Rousseeuw, P. (1997). Integrating robust clustering techniques in S-Plus. *Comput. Statist. Data Anal.*, *26*, 17–37.

Theodoridis, S., & Koutroumbas, K. (1999). *Pattern recognition*. *Academic press*. New York.

Titterington, D. N. (1976). *Algorithms for computing D-optimal design on finite design spaces*. *Proc. 1976 Conf. on Information Science and Systems*. Johns Hopkins University. Baltimore, MD.

Torres, F. A. (2008). *Estadistica Multivariante aplicada a la Geologia*. Universidad de Granada; Falcultad de Ciencias, Granada, Spain. Retrieved from http://www.ugr.es/~fdeasis/Material/MultivarianteGeologia/MultivarianteGeologia_Tema4_Teoria.pdf

Trejos, J. (1998). Métodos de Clasificación y Discriminación. In *Notas de Curso; Maestría en Matemática Aplicada*. Universidad de Costa Rica, San José.

Tryon, R. (1939). *Cluster Analysis*. New York: McGraw-Hill.

Tyron, R. C., & Bailey, D. E. (1970). *Cluster Analysis*. McGraw-Hill.

Vallejo, G. (1992). *Técnicas multivariadas aplicadas a las ciencias del comportamiento*. Universidad de Oviedo. Retrieved from https://books.google.com.pe/books/about/T%25C3%25A9cnicas_multivariadas_aplicadas_a_las.html?id=zwvFV8etpdsC&pgis=1

Veyssieres, M. P., & Plant, R. E. (1998). Identification of vegetation state and transition domains in California's hardwood rangelands. *University of California*.

Vicente-Tavera, S. (1992). *Las técnicas de representación de datos multidimensionales en el estudio del índice de producción industrial en la C.E.E.* Tesis Doctoral. Universidad de Salamanca.

Vicente-Tavera, S., Galindo, M. P., & Ramirez, G. (1994). El HJ-BIPLOT como base para la busqueda de clusters en función de la distribución de parados según profesiones en la Comunidad de Castilla y León. *IV Congreso de Economía Regional de Castilla Y León*, *2*, 822–835.

Vicente-Villardón, J. L. (2010). MultBiplot: A package for Multivariate Analysis using Biplots. Retrieved from http://biplot.dep.usal.es/classicalbiplot/

Wallace, C. S., & Dowe, D. (1994). Intrinsic Classification by MML—the Snob Program. *Proceeding of the 7th Australian Joint Conference on Artificial Intelligence*, 37–44. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.5077

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Wishart, D. (1969). Mode Analysis: A Generalization of the nearest neighbor which reduces chaining effects. *Computer and Information Science. Numerical Taxonomy*, *2*, 282–319.

Wojciechowsaki, T. J. (1987). Nearest neighbour classification rule for mixturs of discrete and continous random variables. *Biometrics*, *29*, 953–959.

Wolf, W. . (1971). Proposed method for determining density of traps required to reduce an insect population. *Journal of Economic Entomology*, *64 (4)*, 872–877.

Wolfson, M., Zagros, M., & James, P. (2004). Identifying national types: a cluster analysis of politics, economics and conflict. *Journal of Peace Research*, *41*(5), pp. 607–623.

Xiong, J., Qureshi, S., & Najjar, L. (2014). A Cluster Analysis of Research in Information Technology for Global Development: Where to from here ? . *AIS Elecronic Library*. Retrieved from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1007&context=globdev2014

Yan-yan, C., Dong, L., & Hui, Z. (2011). Multi-factor Risk Analysis in a Building Fire by Two Step Cluster. *Procedia Engineering*, *11*, 658–665. doi:10.1016/j.proeng.2011.04.710

Yu, F. W., & Chan, K. T. (2012). Using cluster and multivariate analyses to appraise the operating performance of a chiller system serving an institutional building. *Energy and Buildings*, *44*, 104–113. doi:10.1016/j.enbuild.2011.10.026

Yule, G. U. (1912). On the methods of measuring asociation between two attributes. *Journal of the Royal Statistical Society1*, *75*, 579–642.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.2882&rep=rep1&type=pdf

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, *1*(2).

Zubin, J. (1938). A technique for measuring likemindedness. *Journal of Abnormal Psychology*, *33*, 508–516.