3

ANÁLISIS DE CONGLOMERADOS

El análisis de *conglomerados* (o "cluster analysis") se ubica dentro de las técnicas analíticas multivariables de clasificación o de *interdependencia,* al tener como objetivo principal la agrupación de datos. Concretamente, la clasificación de una serie de individuos, objetos o variables en un número reducido de grupos, llamados "conglomerados". La mínima condición que se impone es que los distintos conglomerados creados sean mutuamente excluyentes; es decir, que los casos o variables que constituyan un conglomerado han de ser lo más similar posible entre sí (con respecto a un criterio de selección determinado previamente) y diferente respecto a los integrantes de los otros conglomerados.

Kaufman y Rousseeuw (1990: 1) definen esta técnica analítica como "el arte de encontrar grupos en los datos". Los datos pueden hacer referencia a casos (individuos, objetos) y a variables. En ambas situaciones el proceso de análisis no difiere. El fin último es la consecución del principio de *parsimonia*: la obtención de aquella estructura de los datos más simple posible que represente agrupaciones homogéneas. Si bien se reconoce (Hair *et al.,* 1992, 1999), que ha de primar el equilibrio entre la definición de las estructuras más básicas (pocos conglomerados, en conformidad con el principio de *parsimonia*) y el nivel necesario de similitud dentro de los conglomerados. Se observa que la disminución del número de conglomerados suele ir acompañada de una pérdida no deseada de homogeneidad dentro de los conglomerados.

Aldenderfer y Blashfield (1984) resumen en cuatro los usos principales del análisis de conglomerados:

1. El desarrollo de tipologías o clasificaciones de datos.
2. La búsqueda de esquemas conceptuales útiles para agrupar entidades (o casos).
3. La generalización de hipótesis a través de la exploración de los datos.

4. La comprobación de hipótesis o el intento de determinar si los tipos definidos a través de otros procedimientos están de hecho presentes en una serie de datos.

De estos cuatro usos principales, el primero (la "clasificación de datos") es, sin duda, el que tradicionalmente más ha caracterizado la aplicación del análisis de conglomerados en la investigación aplicada, tanto en el campo de la investigación social, como en las otras áreas de conocimiento.

3.1. Orígenes del análisis de conglomerados y su relación con otras técnicas multivariables

Los orígenes del análisis de *conglomerados* suelen situarse a principios del siglo XX. Más concretamente, en los años treinta, en aplicaciones realizadas en el área de la biología y la botánica, con una finalidad eminentemente clasificatoria o taxonómica: la obtención de *taxonomías* de especies animales y vegetales. Se buscaba la agrupación de distintas especies de animales y vegetales en familias, en función de su similaridad: los pertenecientes a un mismo grupo debían caracterizarse por ser muy semejantes entre sí y diferentes a las otras agrupaciones.

Pero no es hasta los años cincuenta cuando el análisis de *conglomerados* adquiere un mayor desarrollo. Este avance corre paralelo al auge y creciente protagonismo de la informática como instrumento básico para el análisis de los datos, además de las aportaciones de distintos estadísticos que ayudaron a la configuración de esta técnica de análisis.

En 1963 se publica un libro escrito por dos biólogos, Robert Sokal y Peter Sneath, con el título *Principles of Numerical Taxonomy*. Esta publicación se destaca, en la literatura especializada, por ser el "estímulo principal" para el desarrollo del análisis de conglomerados. La necesidad de partir de *taxonomías* claras de seres vivos para comprender el proceso de evolución.

Desde dicha fecha, 1963, hasta 1975, Aldenderfer y Blashfield (1984) contabilizaron que el número de aplicaciones publicadas del análisis de *conglomerados* en todos los campos de conocimiento aproximadamente se doblaban una vez cada tres años, durante el período de tiempo observado (doce años). En la búsqueda de explicación de tan llamativo crecimiento de la literatura sobre el análisis de *conglomerados*, dos son las razones principales que encontraron:

1. La aparición de ordenadores cada vez más potentes y de mayor velocidad.
2. La importancia fundamental de la clasificación como procedimiento científico.

No obstante, hay que puntualizar que el análisis de *conglomerados* no es la única técnica analítica dirigida a la clasificación. Este mismo propósito básico lo cubren otras técnicas analíticas de interdependencia. Dos son las técnicas analíticas multivariables que tradicionalmente más se han relacionado con el análisis de *conglomerados*. Una

pertenece a las técnicas llamadas de "interdependencia": el análisis *factorial exploratorio*. La otra, en cambio, se incluye en las técnicas multivariables de "dependencia": el análisis *discriminante.*

Del análisis *discriminante* le separa un hecho básico: el análisis de *conglomerados* se presenta como una técnica "exploratoria" en la clasificación de datos, mientras que el análisis *discriminante* se define como "confirmatoria". Precisa de la creación previa, mediante análisis exploratorio (sea *factorial exploratorio*, *conglomerados* u otra técnica de interdependencia), de grupos para, posteriormente, derivar las reglas de clasificación.

Además, el análisis *discriminante* (capítulo 4), como técnica analítica de *dependencia*, diferencia entre variables dependientes e independientes (el análisis de *conglomerados* no). Su objetivo básico es estimar la relación de dependencia existente entre una única variable dependiente (categórica) y una serie de variables independientes. El número de categorías de la variable dependiente se refiere a los grupos creados gracias a la aplicación previa de otras técnicas analíticas. De lo que se trata es de comprobar en qué medida los grupos diferenciados quedan bien caracterizados por las variables que los definen. Asimismo, se quiere averiguar cuál es la combinación de variables (llamadas *funciones discriminantes*) que hace máxima la diferencia entre los grupos. El conocimiento de estas variables se considera clave (en el análisis *discriminante*) porque contribuye a la predicción final de la probabilidad de pertenencia de un caso concreto a uno de los grupos formados. Depende de qué valores presente en las variables independientes que forman la *función discriminante.*

En suma, mientras en el análisis de *conglomerados* el investigador desconoce, a priori, la pertenencia grupal de los casos observados, en el análisis *discriminante* se precisa conocer previamente dicha pertenencia para poder derivar la regla de clasificación. Por esta razón, ambas técnicas de análisis se presentan como complementarias. De hecho, es una práctica habitual comenzar realizando un análisis de *conglomerados* para la clasificación inicial de los datos en grupos. Después, dicha clasificación puede validarse mediante un análisis *discriminante*, que profundice en las variables que más caracterizan a los integrantes de cada grupo, diferenciándoles del resto. Sirva como ilustración la aplicación que de ambas técnicas de análisis realicé en mi tesis doctoral (Cea, 1992). Para un mismo objetivo, la descripción del menor de reforma, se aplicaron tres técnicas multivariables diferentes. Primero, un análisis de *conglomerados*, para comprobar si son uno o varios los tipos de menores de reforma; su peso e importancia en el conjunto de la población de reforma. Segundo, un análisis *discriminante*, con la finalidad de comprobar si los tres grupos de menores diferenciados, gracias al análisis de conglomerados, quedaban bien definidos por las variables referidas. Y, por último, un análisis *factorial confirmatorio* dirigido a corroborar los resultados alcanzados con las técnicas multivariables precedentes. Tras su realización se dedujeron las variables que más significativamente identifican a cada grupo de menores.

Pero, si las diferencias con el análisis *discriminante* son obvias, con el análisis *factorial exploratorio* (capítulo 5) disminuyen. Ambas técnicas multivariables son de *interdependencia* y se adecuan a un mismo objetivo básico: la identificación de grupos

internamente homogéneos (y heterogéneos entre sí), a partir de una serie de datos. No obstante, difieren en un rasgo básico: el análisis de *conglomerados* suele restringirse a la búsqueda de relaciones positivas entre las variables. De hecho, se observa –Nourisis (1986; 1994)– que si no se toman los valores absolutos de los coeficientes de correlación, las variables que correlacionan negativamente con un factor no aparecen en el mismo conglomerado con las variables que correlacionan positivamente. Por el contrario, el análisis *factorial* permite que las variables se hallen tanto positiva como negativamente relacionadas con un factor.

El análisis de *conglomerados* puede, al igual que el análisis *discriminante*, utilizarse en complementariedad con el análisis *factorial*. El proceso de análisis puede ser el siguiente: primero, realizar un análisis *factorial exploratorio* dirigido a la formación inicial de los grupos, especialmente, cuando existe un número elevado de variables originales que quiere reducirse a un número bastante inferior de *factores comunes* o *componentes principales*; después, los *coeficientes factoriales* (o "factor loadings") se convierten en los datos a analizar mediante el análisis de *conglomerados*.

Pese a esta ventaja notoria, autores como Manly (1990) recomiendan evitar esta práctica de aplicar previamente un análisis *factorial exploratorio*. Observan que los resultados del análisis de *conglomerados* difieren bastante cuando antes se ha llevado a cabo un análisis *factorial* que cuando éste no se ha realizado.

Por el contrario, la práctica de efectuar un análisis *factorial confirmatorio* con posterioridad a uno de *conglomerados* provoca mayor aceptación, incluso llega a recomendarse su práctica con la finalidad de validar los resultados del análisis de *conglomerados*.

Las similitudes entre el análisis de *conglomerados* y el análisis *factorial* pueden extenderse, finalmente, a otras técnicas analíticas de *interdependencia*. Sea el caso, por ejemplo, del *escalamiento multidimensional*, que se configura como una variedad analítica multivariable análoga al análisis *factorial*, aunque de más reciente implantación. Con él comparte un mismo objetivo: la obtención de un número reducido de dimensiones que permitan caracterizar a determinados objetos o sujetos. Sin embargo, difiere (entre otros aspectos) en el número de dimensiones a obtener. Mientras el análisis *factorial* no impone ninguna restricción al respecto, el *escalamiento multidimensional* aconseja su reducción al menor número posible de dimensiones. Ello responde a condicionamientos impuestos para la representación gráfica de los resultados de la investigación.

Salvo esta diferencia básica, el *escalamiento multidimensional* se presenta, al igual que el análisis *factorial*, como un procedimiento alternativo, además de complementario, para la formación de grupos, a partir de una matriz de datos. Asimismo, el *escalamiento multidimensional* ofrece, igualmente, la opción *exploratoria* y *confirmatoria*.

Por último, destacar las dos críticas más habituales al análisis de *conglomerados* en relación a otras técnicas multivariables:

1. El excesivo protagonismo dado al investigador en su aplicación. En él recae no sólo la decisión clave de qué variables escoger, o la medida de *distancia/similitud*

a emplear, sino también la solución de conglomerados final (referida al número de conglomerados a aceptar). Esto lleva a distintos autores (Manly, 1990; Hair *et al.,* 1992 y 1999) a otorgar al análisis de *conglomerados* el calificativo de ser una técnica "muy subjetiva".

2. A lo anterior, se añade la crítica de que el análisis de *conglomerados* no ofrece –a diferencia del análisis *discriminante*, por ejemplo– un contraste estadístico que ayude a la corroboración o refutación de las hipótesis de investigación. No avanza hacia la inferencia estadística, sino que se queda en un plano meramente descriptivo. Si se quiere proceder a la inferencia habrá, en consecuencia, que acudir a análisis como el *discriminante* o el *factorial confirmatorio*, por ejemplo, en la validación de los hallazgos del análisis de *conglomerados*.

Estas críticas habituales al análisis de *conglomerados* contribuyen a su caracterización como técnica analítica "descriptiva, ateórica y no inferencial" (Hair *et al.,* 1999: 493). De ella se afirma incluso que carece de bases estadísticas que permitan la inferencia estadística (de las estimaciones muestrales a los parámetros poblacionales), quedándose en un plano meramente exploratorio. Las soluciones tampoco son únicas. La pertenencia al conglomerado depende de muchos elementos del procedimiento, pudiéndose obtener muchas soluciones diferentes sólo variando uno o más de estos elementos. Además, la aplicación del análisis de *conglomerados* siempre concluye con la formación de varios conglomerados, aunque los datos carezcan de una estructura "auténtica".

A continuación se ofrece al lector la oportunidad de comprobar el acierto o desacierto de las críticas expuestas. Para ello se recomienda la lectura pormenorizada de las páginas que siguen, junto al ejercicio de comparar el procedimiento seguido en su realización con el desarrollado en otras técnicas analíticas que persiguen los mismos objetivos.

3.2. Fases principales en su aplicación

En la materialización de un análisis de *conglomerados* coinciden una serie de fases que pueden resumirse en las siguientes:

1. *Selección de las variables* que favorezcan la agrupación de los datos. Ésta es una decisión clave y previa a cualquier análisis de conglomerados. Las variables finalmente elegidas son las que determinan las características de "clasificación" (aquellas que identifican a cada conglomerado).
2. *Elección del procedimiento de conglomeración* a seguir (jerárquico y/o no jerárquico), junto al *algoritmo* de clasificación para la creación de los conglomerados.
3. *Elección de medidas de distancia y proximidad* para proceder a la formación de los conglomerados. Esta elección está determinada, en gran medida, por la na-

turaleza de las variables incluidas en el análisis. Si se trata de variables en su mayoría *no métricas* (*nominales* u *ordinales*), la elección se limita a las llamadas "medidas de co-ocurrencia" (subapartado 3.3.4). En cambio, para las variables *métricas* el abanico de posibilidades se amplía.

Además de las variables, incide el *algoritmo* que se haya escogido para la formación de los conglomerados y el procedimiento de conglomeración a seguir.

4. *Decisión sobre el número de conglomerados* a constituir.
5. *Presentación e interpretación de los resultados*, tanto en su forma numérica (la *tabla de aglomeración*) como gráfica (habitualmente el *dendograma* y el *gráfico de carámbanos* o de *témpanos*).
6. *Validación de los resultados* del análisis. Si éstos no logran alcanzar la calificación de "válidos", habrá que introducir modificaciones que ayuden a su mejora. La consecuencia inmediata será la repetición de todo el proceso, comenzando con el replanteamiento de las decisiones adoptadas con anterioridad a la ejecución del análisis. Los análisis se dan por concluidos cuando sus resultados logren satisfacer unos criterios mínimos de *validez* (apartado 3.6).

Estas fases esenciales en la materialización de un análisis de conglomerados pueden resumirse a modo de gráfico, en cuatro amplios bloques, como ilustra la figura 3.1.

3.3. Los preliminares del análisis: decisiones clave

Para la correcta materialización del análisis de conglomerados son decisivas una serie de decisiones "clave", que el investigador ha de adoptar, en busca de la consecución satisfactoria de sus objetivos. Estas decisiones conciernen no sólo a actuaciones compartidas con otros procedimientos analíticos, como es el tratamiento de los casos "sin respuesta" o "missing values" (subapartado 1.3.1), sino que también incluye decisiones exclusivas al análisis de conglomerados, que afectan directamente a su realización, y que se detallan a continuación.

3.3.1. Elección de variables

Una actuación tan rutinaria (en cualquier análisis), como la elección de variables, en el análisis de conglomerados se convierte en decisiva. Dependiendo de qué variables se escojan para la ejecución del análisis, varía no sólo el número, sino también la composición de los conglomerados.

La recomendación continuamente reiterada es limitar las variables sólo a aquellas que sean "relevantes" a los objetivos del estudio, de acuerdo con el *marco teórico* de la investigación. Éste es, no obstante, el ideal. Como Aldenderfer y Blashfield (1984: 20)

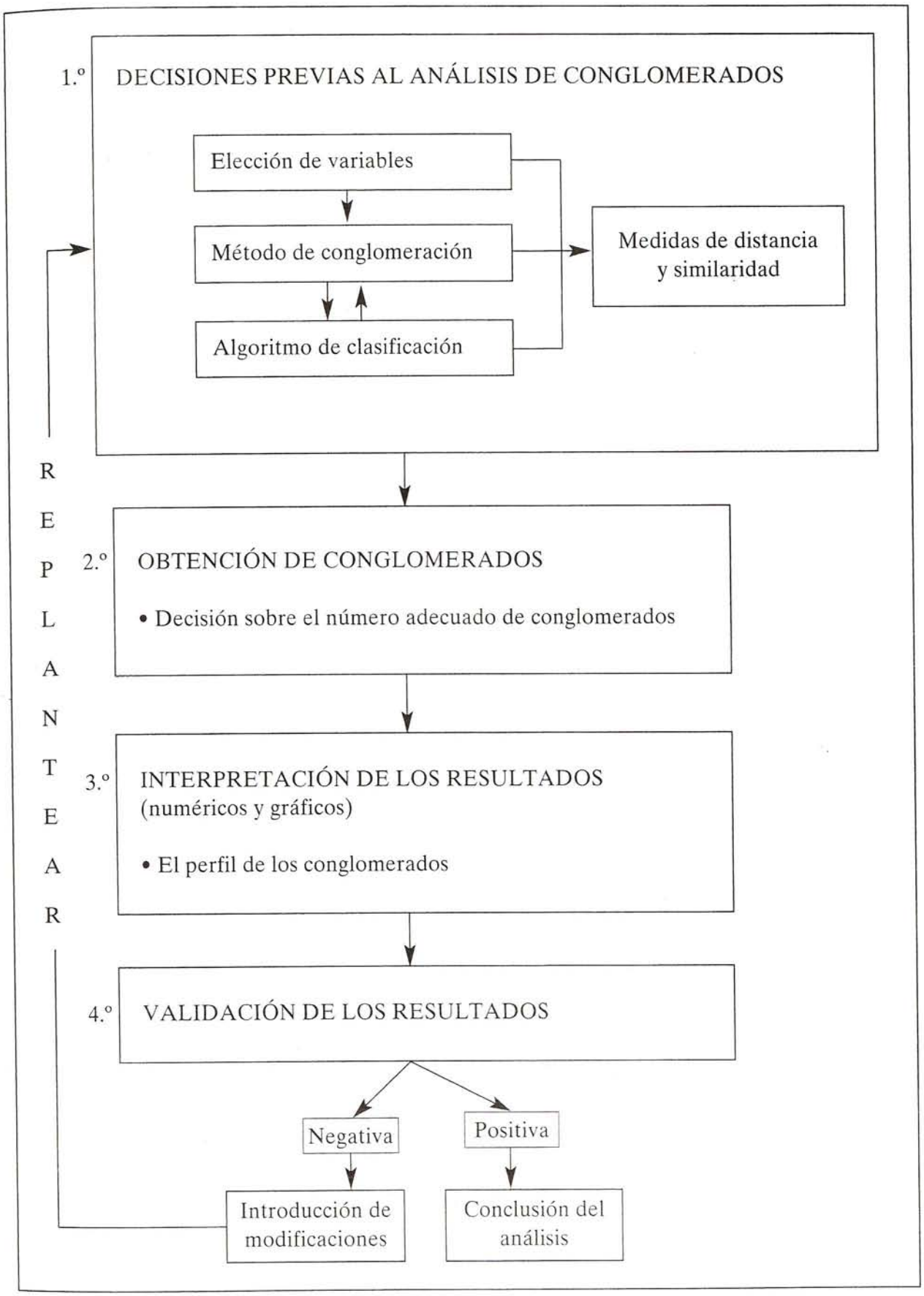


Figura 3.1. Fases principales de un análisis de conglomerados.

argumentan, "las variables deberían elegirse dentro del contexto de una teoría establecida explícitamente que se utiliza para apoyar la clasificación". Y ello porque "la teoría es la base para la elección racional de las variables a utilizar en el estudio". Aunque reconocen que, en la práctica, "la teoría que apoya la clasificación con frecuencia está implícita, y en esta situación es difícil asegurar la relevancia de las variables al problema". Esto no resta, sin embargo, que se insista en la importancia de la teoría como guía fundamental en la elección de variables.

El escoger variables que sean "relevantes" a la clasificación tiene, asimismo, una consecuencia inmediata: el restringir la presencia de "atípicos" ("outliers"). Como observan Hair *et al.* (1992 y 1999), la inclusión de variables irrelevantes aumenta la oportunidad de que aparezcan "atípicos". Por esta razón se insiste en que el investigador procure escoger variables que realmente logren diferenciar entre conglomerados. El criterio principal de selección es proceder conforme a los objetivos de la investigación.

Después, una vez que los análisis se han realizado, si alguna de las variables, previamente considerada "relevante", se torna "irrelevante", al no conseguir diferenciar significativamente a los conglomerados, lo mejor será considerar su eliminación del análisis. Si se procede a la eliminación de variables que muestren ser "irrelevantes", habrá que repetir el proceso analítico completo sólo con aquellas variables que muestren "relevancia" en la clasificación.

La relevancia no sólo ha de entenderse en sentido estadístico: el lograr caracterizar a un conglomerado concreto, diferenciándolo del resto. También es prioritario que la variable tenga un significado lógico-sustantivo. Este aspecto es inclusive preferencial al primero (el estadístico) y ha de primar en la selección de variables.

En la eliminación de variables que muestren, tras la realización del análisis, "irrelevancia" en la formación de conglomerados, es preferible actuar de forma "pautada". Quiere esto decir, que si hubiese más de una variable "irrelevante", lo mejor es proceder como es común en cualquier modalidad de análisis: la eliminación de las variables de forma secuencial. Esto supone que después de cada eliminación se han de repetir los análisis para comprobar, a continuación, qué variables muestran ser "irrelevantes" tras cada eliminación.

Además, ha de advertirse que la "irrelevancia" de una variable puede verse afectada por la relación *colineal* que dicha variable tenga con las demás incluidas en el análisis. Si la *colinealidad* es elevada, la eliminación de una de las variables muy colineales puede provocar el efecto positivo de convertir, a la variable que previamente mostró ser "irrelevante" en la configuración de los conglomerados, en "relevante". Lo que haría innecesaria su eliminación.

- Una segunda decisión clave respecto a las variables es si analizarlas en su métrica original o proceder, por el contrario, a su *estandarización*. Esta segunda decisión genera mayor controversia que la anterior.

Como en otros procedimientos analíticos, las variables que incluyen un rango elevado de valores (por ejemplo, la variable "ingresos", inclusive en relación con otras va-

riables *métricas*, como "años de estudio") ejercen mayor influencia en los resultados del análisis. Concretamente, contribuyen más en el cálculo de las medidas de *distancia* o de *proximidad*, que las variables de pequeño *recorrido* (o *rango*), indistintamente de su "relevancia" real en la diferenciación de los conglomerados.

Para evitar este problema, la solución habitual es transformar todas las variables a una escala común; es decir, *estandarizarlas*. De hecho, la correcta realización del procedimiento "K-means" (una de las modalidades más aplicadas del análisis de conglomerados), por ejemplo, precisa que todas las variables hayan sido previamente estandarizadas.

El procedimiento de *estandarización* más común consiste en transformar cada variable a puntuaciones Z, convirtiéndose su media aritmética en "0" y su desviación típica en "1". Para ello se divide la diferencia entre cada valor y la media de la variable entre la desviación típica ($Z = (X - \overline{X})/S$). Pero también puede optarse por un procedimiento alternativo: poner cada variable en un rango de "0" a "1", donde el valor más pequeño es "0" y el mayor el "1" (SPSS Inc., 1997).

No obstante, existe controversia respecto a si la *estandarización* debe o no aplicarse de forma rutinaria en el análisis de conglomerados. Everitt (1980), por ejemplo, advierte que la *estandarización* puede provocar la reducción de las diferencias entre grupos en aquellas variables que bien pueden ser las que más discriminan entre los grupos. La única *estandarización* que admite es de las variables "intragrupales", las que caractericen a cada grupo. Pero, obviamente, esta *estandarización intragrupal* no puede llevarse a cabo hasta que no se haya procedido a la clasificación en conglomerados. Lo que complica la práctica de esta propuesta de *estandarización* alternativa.

Cuando las variables se encuentran medidas en una escala común, o no existen muchas divergencias en la cuantía de sus valores (o amplitud de sus *rangos*), no es necesario proceder a la *estandarización* de las variables. En este caso, los análisis se realizan con las variables en su métrica original. Con ello se evita los perniciosos efectos observados en la *estandarización*.

3.3.2. Métodos de conglomeración

Otra decisión clave previa al análisis de conglomerados concierne al procedimiento que va a seguirse en la formación de los conglomerados. Esta decisión es igualmente importante porque la composición de los conglomerados puede verse afectada por el método seguido en su formación. Los resultados pueden no coincidir, cuando se aplican métodos de conglomeración diferentes.

A este respecto, son varias las opciones de conglomeración posibles, incluso puede elegirse aplicar dos o más de ellas de forma combinada. De nuevo, los objetivos de investigación actúan como criterio básico en la decisión. A ellos se suma la peculiaridad de la matriz de datos, principalmente, el tamaño de la muestra y las características de las variables, en especial su métrica; de ella depende la medida de *distancia* o de *similaridad* que se escoja para la formación de los conglomerados.

La decisión también puede verse afectada por las opciones de análisis que permita el programa informático que vaya a utilizarse. El cuadro 3.1 resume el amplio abanico de procedimientos existentes para la formación de conglomerados. Incluye tanto los métodos genéricos de conglomeración como los *algoritmos* concretos aplicados en la clasificación de las unidades (casos y/o variables).

CUADRO 3.1. Métodos de conglomeración y sus algoritmos de clasificación

A. MÉTODOS DE CONGLOMERACIÓN JERÁRQUICOS

A.1. *Aglomerativos*
- Distancias mínimas
- Distancias máximas
- Promedio entre grupos
- Promedio intragrupos
- Método Ward
- Método del centroide
- Método de la mediana

A.2. *Divisivos*
- Método de partición binaria de Howard-Harris
- Aplicación de algunos de los procedimientos aglomerativos. En especial, el método de Ward y el promedio entre grupos.

B. MÉTODOS DE CONGLOMERACIÓN NO JERÁRQUICOS

B.1. *De "reasignación" o de "partición iterativo"*
- Método K-means
- Quick cluster analysis
- Nubes dinámicas

B.2. *Búsqueda de densidad*
- Aproximación tipológica (análisis modal de Wishart, método de Taxmap y método de Fortín).
- Aproximación probabilística (método de combinaciones de Wolf).

B.3. *Métodos directos*
- "Block clustering" de Hartingan

En este subapartado se describen únicamente las características de los métodos genéricos de conglomeración. El detalle de los *algoritmos* concretos para la constitución de los conglomerados se encuentra en el subapartado 3.3.3.

3.3.2.1. Métodos jerárquicos

Tradicionalmente éstos han sido los procedimientos más aplicados para la formación de conglomerados (fundamentalmente los "aglomerativos"), cuando el tamaño de la muestra no es elevado (≤ 200 unidades). Si la muestra supera las 200 unidades, la "simplicidad" que caracteriza a los métodos *jerárquicos* se torna en "dificultad" de cálculo y de interpretación. Los análisis se realizan a partir de una matriz de distancias, con entradas para cada par de objetos (casos o variables). Su volumen aumenta con el tamaño de la muestra. Lo mismo sucede con la lectura e interpretación de los resultados gráficos (el *dendograma* y el *gráfico de carámbanos*).

El *dendograma* (o diagrama "en árbol") es –como se verá en el subapartado 3.5.2– un gráfico típico de los métodos *jerárquicos* de conglomeración, convirtiéndose inclusive en su expresión más familiar. Su visualización ayuda bastante a comprender cómo se han ido formando los distintos conglomerados. Pero cuando la muestra es mayor de 200 unidades, la magnitud del *dendograma* crece, llegando a ocupar varias páginas, lo que, en vez de ayudar, dificulta la lectura e interpretación de la solución de conglomerados. Lo mismo acontece con la otra expresión gráfica: el *gráfico de carámbanos* (o de *témpanos*), que también es muy aplicado en los métodos *jerárquicos* (subapartado 3.5.2).

Sin duda uno de los rasgos que más distinguen a los métodos *jerárquicos* de conglomeración es el procedimiento seguido para la agrupación de los objetos. En los "aglomerativos", los conglomerados se forman, primero, a partir de objetos individuales concretos y, después, de la conjunción de conglomerados. En los "divisivos" (o de "partición"), el proceso escalonado es el inverso: del conglomerado global se pasa, también pausadamente, mediante descomposición, a conglomerados varios hasta concluir en los objetos concretos a clasificar. Los conglomerados así creados se caracterizan por ser anidados, esto quiere decir, que cada uno de ellos puede, a su vez, ser subsumido por otro conglomerado más grande, en un nivel de similaridad superior. Éste es uno de sus rasgos distintivos frente a los métodos *no jerárquicos*. En estos últimos, la unión inicial de dos o más objetos puede variar en la solución final. Ésta no queda tan determinada por la partición inicial de los datos como sucede cuando se aplican métodos *jerárquicos*.

Cuando se conglomeran casos, una especificación mínima para proceder a un análisis de conglomerados *jerárquico* es "una o más variables numéricas". En cambio, para una conglomeración de variables, "tres o más variables numéricas" (Nourisis, 1994: 100). El proceder también varía según el método de conglomeración *jerárquico* seguido: *aglomerativo* o *divisivo*.

A) Métodos jerárquicos aglomerativos

Constituyen la variedad más popular de los métodos *jerárquicos*, por su mayor aplicación y desarrollo. También se les conoce como "métodos jerárquicos ascendentes"

(Manly, 1990). En ellos la agrupación de objetos procede de forma "ascendente" o "aglomerativa". Primero, de objetos singulares concretos y, después, de conglomerados simples a conglomerados cada vez más complejos, hasta concluir con un único conglomerado.

El análisis comienza con tantos conglomerados como objetos a clasificar (ya sean variables o casos). En un segundo paso, dos de los objetos se combinan en un único conglomerado. En el tercer paso, surge un nuevo conglomerado de la fusión, bien de otros dos objetos adicionales, bien de un tercer objeto que se une al conglomerado previamente formado por los dos objetos. La formación de conglomerados es gradual y ascendente. En cada paso se constituye un nuevo conglomerado, bien como resultado de la unión de dos objetos que permanecían todavía aislados (sin pertenecer a ningún conglomerado), o bien por la anexión de un objeto a un conglomerado ya constituido, o por la conjunción de dos conglomerados ya existentes. El proceso de conglomeración concluye cuando se llega a un único conglomerado que reúne a todos los objetos. El *dendograma* permite la visualización de cómo se han ido formando los conglomerados en las distintas etapas del análisis.

La característica distintiva de este método de conglomeración es que una vez que el conglomerado se ha constituido (dos objetos se han vinculado) no puede dividirse en etapas posteriores. Al contrario, sólo puede ampliarse por la anexión de nuevos miembros, algunos de ellos pertenecientes a conglomerados ya existentes.

Tras cada nueva agrupación, se recalculan las *distancias*, de acuerdo con el *algoritmo* de clasificación y la medida de *distancia/similaridad* escogida para la formación de conglomerados. Cuando el análisis de conglomerados es de casos, el criterio que decide la pertenencia a los conglomerados se basa en la *matriz de distancias* o, en su caso, de *similaridad*, entre pares de casos. Si, por el contrario, se quiere agrupar variables, las medidas de *distancia/similaridad* se calculan entre pares de variables.

B) Métodos jerárquicos divisivos o de partición

También se les conoce como métodos "descendentes" o "disociativos", en contraposición a los aglomerativos. Su aplicación en la investigación social ha sido menor, en parte debido a su escasa presencia en los paquetes estadísticos iniciales.

En la formación de conglomerados ahora se procede de forma inversa a la anterior. El análisis comienza con un único conglomerado que incluye a todos los casos o variables observados. Después, y de forma gradual, se procede a la disgregación de ese gran conglomerado inicial, con la excepción de aquel objeto (caso o variable) que se halle más distante del promedio de los otros objetos en el conglomerado. De este modo, el conglomerado inicial se divide en dos conglomerados, entre los que se distribuyen los casos o variables. Éstos quedan ubicados en el conglomerado hacia el que estén más próximos.

Tras cada escisión o división de conglomerados se vuelven a calcular las distancias entre sus integrantes. Los objetos situados a mayor distancia del promedio del con-

glomerado se separan del mismo, ya sea constituyendo un nuevo conglomerado, ya añadiéndose al conglomerado hacia el que ahora se sitúen más "próximos".

El proceso de división de conglomerados continúa iterativamente hasta que existan tantos conglomerados como objetos a clasificar. La distancia que se permite entre los integrantes de un mismo conglomerado es reducida, lo que favorece la disgregación en cada vez mayor número de conglomerados y de menor tamaño. Cuando esto sucede se alcanza la disgregación máxima, que supone el fin del proceso.

Los conglomerados creados mediante métodos *jerárquicos divisivos* pueden ser de dos clases diferentes:

a) *Monotéticos*: si todos los objetos incluidos en el conglomerado tienen el mismo valor en una variable concreta. Esta variable es la que define al conglomerado, al determinar la pertenencia al conglomerado según el valor que los objetos presenten en la misma.
b) *Politéticos*: cuando el protagonismo en la definición del conglomerado lo comparten dos o más variables. La conjunción de estas variables determina la pertenencia al conglomerado.

Los conglomerados *politéticos* suelen ser los más habituales. Las estrategias divisivas *monotéticas* se restringen, preferentemente, a datos *binarios;* cuando la división de conglomerados se basa en la identificación de una variable que hace máxima la diferencia (o disimilaridad) entre los conglomerados. Uno de los criterios divisivos más comúnmente empleado es el estadístico *chi-cuadrado*, como se verá en el subapartado 3.3.4.

3.3.2.2. Métodos no jerárquicos

Otra alternativa a la formación de conglomerados la ofrecen los métodos *no jerárquicos*, también llamados de "optimización". Este último nombre responde a cómo se produce la asignación de los objetos a los conglomerados. La finalidad es "optimizar" el criterio de selección.

En cambio, la primera denominación ("no jerárquicos") se debe al procedimiento seguido en la constitución de conglomerados: procedimientos de partición "no jerárquicos".

Pero, como sucede con los métodos *jerárquicos*, la categoría genérica de métodos *no jerárquicos* engloba una amplia variedad de procedimientos en la constitución de los conglomerados. Si bien, en el establecimiento de una tipología básica de métodos *no jerárquicos* existe menos consenso, entre los autores, que en los métodos *jerárquicos*.

Una de las clasificaciones de métodos *no jerárquicos* más amplia es la resumida por Bisquerra (1989) en tres categorías extensas: cada una de ellas se acompaña de los *algoritmos* de clasificación principales.

A) Métodos de reasignación

Permiten que los objetos asignados a un conglomerado en una fase del proceso sean reasignados a otro conglomerado en otra fase posterior. La condición es que la "reasignación" consiga "optimizar" el criterio de selección.

La formación de conglomerados concluye cuando no queda ningún objeto cuya reasignación logre optimizar el resultado.

Aldenderfer y Blashfield (1984) llaman a estos métodos de *reasignación* métodos de "partición iterativos". Parten de una partición inicial de los datos, que puede verse modificada por el desplazamiento (o reasignación) de los objetos a otros conglomerados, a cuyo *centroide* se hallen más próximos. Esta alteración en la composición de los conglomerados acontece mediante procedimientos iterativos.

Algunos de los *algoritmos* más conocidos dentro de estos métodos son:

1. Método "K-means" de McQueen.
2. El "quick cluster analysis" y el método de "Forgy". Ambos se agrupan bajo el nombre genérico de "métodos de centroide" (o "centros de gravedad").
3. El método de nubes dinámicas de Diday.

B) Métodos de búsqueda de densidad

Aldenderfer y Blashfield (1984: 51) los definen como "desarrollos naturales del concepto que concibe al conglomerado como una región de una 'elevada' densidad de puntos en un espacio relacionado con aquellas regiones que los rodean". Estos métodos "buscan", esencialmente, el espacio para "modas" naturales en los datos que representan estas áreas de elevada densidad.

Los métodos de *búsqueda de densidad* incluyen dos aproximaciones básicas: la aproximación tipológica y la probabilística.

a) La aproximación tipológica puede considerarse una variante del método de *distancias mínimas* ("single link"). De él difiere en que ofrece reglas para iniciar nuevos conglomerados, más que unir las entidades encontradas recientemente a los conglomerados ya existentes. Los conglomerados se forman a partir de la búsqueda de aquellas zonas en las que se dé una mayor concentración de objetos. Los *algoritmos* más conocidos son:

1. El análisis modal de Wishart.
2. El método de Taxmap de Carmichael y Sneath.
3. El método de Fortín.

b) La aproximación probabilística. Parte del supuesto de que las variables siguen una ley de probabilidad, de acuerdo con la cual, los parámetros varían de

un conglomerado a otro. De lo que se trata es de encontrar los objetos que pertenecen a la misma distribución.

Entre los *algoritmos* aplicados destaca el método de las combinaciones de Wolf.

C) Métodos directos

Permiten la clasificación simultánea de los individuos y de las variables. Las entidades que se agrupan ya no son los casos o las variables, por separado. Por el contrario, se procede a su análisis conjunto, es decir, al cruce de ambas (casos por variables), tal y como figura en la matriz de datos.

El *algoritmo* de mayor aplicación en este tercer grupo de métodos *no jerárquicos* es el llamado "block clustering" de Hartingan.

• Aldenderfer y Blashfield (1984) añaden otra categoría aparte que denominan método "clumping". Este método ha alcanzado una amplia aplicación en el campo de la lingüística, donde interesa la representación de palabras que incluyan múltiples significados, pero, en otras áreas de conocimiento, apenas es conocido.

El método de "clumping" se incluye en los métodos *no jerárquicos* porque no produce clasificaciones "jerárquicas". Aunque difiere de los otros métodos (*no jerárquicos*) porque permite la creación de conglomerados que se superponen: un mismo objeto puede pertenecer a más de un conglomerado. Éste es el rasgo que más caracteriza al método "clumping" y le distingue del resto de métodos de conglomeración, tanto *jerárquicos* como *no jerárquicos.*

Asimismo, se distingue por requerir el cálculo de una *matriz de similaridad* entre los casos. Los datos suelen partirse mediante métodos aleatorios en un número de configuraciones de partida diferentes, de modo que cada vez se crea sólo dos grupos. Los objetos entonces se vuelven a localizar iterativamente hasta conseguir que la función a optimizar sea estable. La finalidad es intentar "optimizar" el valor de un criterio estadístico que técnicamente se refiere como una "función de cohesión".

El problema más importante que se observa en este procedimiento de formación de conglomerados es que los mismos conglomerados se descubren, con frecuencia, y de forma reiterada. Por lo que no se proporciona información nueva (Aldenderfer y Blashfield, 1984).

• Hair *et al.* (1992 y 1999) se distancian de estos autores, ofreciendo otra clasificación de los métodos *no jerárquicos*. Aunque su clasificación es menos completa que la anteriormente enunciada. Diferencia tres modalidades alternativas para la obtención de conglomerados mediante procedimientos *no jerárquicos:*

1. Método de umbral secuencial.
2. Método de umbral paralelo.
3. Método de optimización.

1. *Método de umbral secuencial.* Una variedad del procedimiento de conglomeración no jerárquica que se ajusta a grandes series de datos. Se encuentra en programas como FASTCLUS, en SAS.

   El análisis comienza, como en cualquier procedimiento *no jerárquico*, con la indicación por parte del analista del número máximo de conglomerados permitido. A partir de esta especificación, el programa comienza seleccionando "semillas" de conglomerado, que se emplean como conjeturas iniciales de las *medias* de los conglomerados. La primera *semilla* es la primera observación en la serie de datos con ningún valor "sin respuesta" ("missing value"). La segunda *semilla* es la siguiente observación completa (es decir, sin ningún valor *sin respuesta*), que se separa de la primera "semilla" por una distancia mínima específica. La opción que el programa aplica por defecto es una distancia mínima de cero.

   A la selección de un conglomerado "semilla" le sigue la asignación de todos los objetos que se hallen dentro de la distancia especificada previamente. Después, se selecciona otro conglomerado "semilla", con la consiguiente asignación de objetos que estén en la distancia especificada. El proceso continúa hasta que no quede ningún objeto por clasificar.

   Tras cada atribución de un objeto a un conglomerado se actualizan las *semillas* de conglomerado, si se quiere. Ello supone el cálculo de las *medias* de los conglomerados después de cada asignación de un objeto a un conglomerado.
2. *Método de umbral paralelo.* Difiere del anterior en que la selección de "semillas" se hace de forma simultánea y al principio del proceso. Los objetos dentro de la *distancia umbral* se asignan, igualmente, al conglomerado "semilla" más próximo. En algunas variantes del método, cabe la opción de que algún objeto quede fuera de los conglomerados, si se halla fuera de la distancia previamente especificada desde cualquiera de las "semillas" de conglomerados.

   Tras las atribuciones de los objetos a los conglomerados, las distancias umbrales pueden ajustarse para incluir "más" o "menos" objetos en los conglomerados.

   Como ejemplo de este procedimiento *no jerárquico* se cita el procedimiento "Quick Cluster" del programa SPSS. En él se establecen los puntos de "semilla" bien como puntos proporcionados por el usuario, o bien como puntos seleccionados aleatoriamente de todas las observaciones.
3. *Método de optimización.* El procedimiento de formación de conglomerados se asemeja a los dos precedentes, salvo en un aspecto importante: se permite la reasignación de objetos a otros conglomerados, desde el original, si con ello se satisface algún criterio de *optimización* global.

• De la comparación de las distintas clasificaciones propuestas de los métodos de conglomeración *no jerárquicos*, puede concluirse el predominio de los denominados

"métodos de reasignación" o de "partición iterativos". Dentro de ellos puede también incluirse la tipología asumida por Hair *et al.* (1992 y 1999).

Estos "métodos de reasignación" han sido, dentro de los procedimientos *no jerárquicos*, los más aplicados en la investigación social, incluso se han convertido en su referente. De modo que, cuando se comparan los métodos *jerárquicos* con los *no jerárquicos*, las características a las que se hace mención corresponden, en su generalidad, al proceder de los "métodos de reasignación". De las otras especificidades apenas se hace mención.

• Como resumen de lo expuesto, cabe destacar tres diferencias básicas que separan a los métodos *jerárquicos* de los *no jerárquicos*:

a) En los métodos *no jerárquicos* el procedimiento de formación de conglomerados comienza a partir de una partición inicial de los datos. El investigador especifica, previamente, el número máximo de conglomerados que debe haber en la matriz de datos. Quiere esto decir que parte de una clasificación inicial de los objetos, de acuerdo con algún criterio de investigación.

A partir de esta clasificación inicial (determinada por un número concreto de conglomerados) se produce la asignación de objetos a los conglomerados a cuyo centro (o *centroide*, que corresponde al valor medio de las variables que configuran el conglomerado) se hallen más próximos. A la constitución de los conglomerados le puede seguir un nuevo cálculo de los *centroides*. Para ello se consideran los objetos que finalmente se asignaron a los conglomerados. Los nuevos *centroides* pueden provocar el desplazamiento de objetos a otro conglomerado, si ahora el objeto se sitúa más próximo al *centroide* de ese nuevo conglomerado.

Los *centroides* se recalculan cada vez que se produce una alteración en la composición de los conglomerados. El proceso concluye cuando no se produce ninguna nueva modificación en los conglomerados.

Este proceder, descrito de forma genérica, varía en función del *algoritmo* de clasificación escogido (*K-means, quick cluster...*). Además, es más característico de los métodos de *partición iterativos*, también llamados de "reasignación". Las otras variedades ya han sido expuestas.

b) Si en los métodos *jerárquicos* la asignación de un objeto a un conglomerado suele considerarse definitiva, en los métodos *no jerárquicos* puede ser accidental. Continuamente se valora la pertenencia de los objetos a los conglomerados a los que se les ha asignado inicialmente. Esto es posible gracias a la aplicación de procedimientos *iterativos*, de los que pueden derivarse modificaciones sustantivas en la composición de los conglomerados.

Este proceder resuelve uno de los inconvenientes principales tradicionalmente atribuidos a los métodos *jerárquicos*: una mala asignación inicial de los objetos a los conglomerados puede determinar una desacertada solución final. Recuérdese que en los métodos *jerárquicos*, en especial en los *aglomera-*

tivos, una vez que un objeto es asignado a un conglomerado, queda definitivamente en él. Tras la formación de nuevos conglomerados no se evalúa la pertenencia del objeto al conglomerado inicialmente asignado. En cambio, los métodos *no jerárquicos* se caracterizan por lo contrario: la valoración continua de la pertenencia de los objetos a los conglomerados, tras cada alteración en la composición de los mismos.

Pero esta valoración "continua" de la composición de los conglomerados no logra solventar una de las deficiencias importantes generalmente atribuidas a los métodos *no jerárquicos*: la derivada de una mala decisión inicial sobre el número de conglomerados "real" que existe en la matriz de datos. Esto puede ocasionar una errónea clasificación de los objetos (ya sean casos o variables).

Por esta razón, se recomienda repetir los análisis variando, cada vez, la especificación inicial del número máximo de conglomerados que quiere formarse. De las distintas posibilidades de clasificación existentes se escogerá aquella que ofrezca una mejor interpretación, desde el punto de vista estadístico y lógico-sustantivo. Ante todo, ha de tener sentido lógico, relacionado con el marco teórico de la investigación.

c) Los métodos *jerárquicos* operan a partir de una matriz de *similaridades*, ya sea de casos ($N \times N$) o de variables ($p \times p$). En los métodos *no jerárquicos* se trabaja, en cambio, con los datos brutos originales. Esto proporciona una ventaja inicial importante: el facilitar el análisis de muestras grandes (mayores de 200 unidades). Estas muestras, por el contrario, son difíciles de analizar mediante métodos *jerárquicos*, como ya se mencionó.

3.3.2.3. La combinación de métodos de conglomeración

De la lectura de los subapartados anteriores puede concluirse que cada método de conglomeración ofrece unas ventajas, pero también presentan unos inconvenientes o límites importantes. El cuadro 3.2 resume los inconvenientes principales observados en la práctica de ambos métodos de conglomeración.

Estos y otros inconvenientes pueden solventarse, si se opta por combinar métodos *jerárquicos* de constitución de conglomerados con métodos *no jerárquicos*, para cubrir un mismo objetivo de investigación. Una estrategia analítica posible es la siguiente:

1. Aplicar inicialmente un método *jerárquico*, con la finalidad de conocer el número de conglomerados que se pueden formar en la matriz de datos concreta que se analiza.

   La información no se limitará a la identificación del número y la composición de los distintos conglomerados. También abarca otros aspectos de gran interés para un análisis posterior, como es el conocimiento de los *centroides* de los conglomerados y los casos *atípicos*. Recuérdese que la valoración de estos últimos se hace en función de su número.

2. La solución que resulte del método *jerárquico* se toma como punto de partida del método *no jerárquico*, lo que ayuda a ajustar o precisar más la constitución de los conglomerados obtenidos con la aplicación del método jerárquico.

CUADRO 3.2. *Inconvenientes principales de los métodos jerárquicos y no jerárquicos*

MÉTODOS JERÁRQUICOS	*MÉTODOS NO JERÁRQUICOS*
Dificultad de determinar *a priori* el mejor *algoritmo* de clasificación, cuando el investigador desconoce la estructura de la muestra.	Dificultad de conocer *a priori* el número de conglomerados "real" existente en los datos observados.
A menos que se empleen *algoritmos* especiales, es difícil operar con muestras superiores a 200 unidades porque se parte de una *matriz de similaridad.* Al confeccionarse ésta con cada par de objetos (casos o variables) adquiere un tamaño desorbitado, conforme aumenta el tamaño de la muestra. En especial, cuando se clasifican casos. La lectura de los resultados gráficos (mediante el *dendograma* o el gráfico de *carámbanos*) también es difícil de realizar en muestras grandes.	Formar todas las particiones posibles de la serie de datos (que se presenta como la forma más directa de descubrir la partición óptima de una serie de datos), iterativamente, supone la realización de cálculos muy complejos para un número elevado de casos y de conglomerados. Ello dificulta su puesta en práctica.
Una mala partición inicial de los datos no puede modificarse en fases posteriores del proceso de conglomeración.	Una mala decisión inicial sobre el número de conglomerados "real" puede resultar en una errónea clasificación de los datos.
Mayor predisposición a la presencia de "atípicos" (o *outliers*).	Mayor complejidad de los análisis que le hace muy dependiente de la capacidad del ordenador que se utilice.

Por ejemplo, puede aplicarse el procedimiento *jerárquico* de *Ward* en una fase inicial del análisis. De él se obtiene información sobre el número de conglomerados, su composición, sus *centroides* y los casos *atípicos*. Toda esta información es de gran utilidad para la aplicación idónea de un método *no jerárquico*. Concretamente, el procedimiento *K-means* ("K-medias"), uno de los métodos *no jerárquicos* más populares, precisa de la especificación previa no sólo del número de conglomerados, sino también de sus *centroides*. Si esta información no es aleatoria, sino que se basa en un análisis exhaustivo precedente, es más factible que logre una mayor aproximación a la "realidad". Asimismo, la detección de *atípicos* y su tratamiento posterior también ayuda a la mejora de la clasificación final de los datos.

En los últimos años, la aplicación conjunta de los procedimientos de *Ward* y *K-means* ha llegado a convertirse incluso en "la combinación perfecta" en la investigación aplicada (Gómez Suárez, 1999: 542).

3.3.3. Algoritmos de clasificación

Una tercera decisión clave en los preliminares del análisis de conglomerados concierne al *algoritmo* de clasificación a aplicar. Esta decisión también es importante porque dependiendo del *algoritmo* elegido, varía el número y la composición de los conglomerados.

El cuadro 3.1 incluye algunos de los principales *algoritmos* de clasificación de los métodos *jerárquicos* y de los *no jerárquicos*. Todos ellos persiguen el mismo objetivo básico: crear conglomerados "homogéneos" pero, a su vez, muy diferentes unos de otros. En términos de varianza, se trata de formar conglomerados de una elevada homogeneidad intragrupal y, por el contrario, una elevada heterogeneidad entre los grupos. Pero la forma como se alcanza este objetivo varía en función del *algoritmo* escogido.

Por *algoritmo* se entiende –siguiendo la definición dada por Moliner en *Diccionario de uso del Español* (1984)– "notación propia de una forma particular de cálculo". En el análisis de conglomerados, en concreto, con el término *algoritmo* se hace referencia al procedimiento a seguir en la disposición de objetos similares en conglomerados. Los procedimientos posibles son varios y, como afirma Manly (1990: 105), "no existe ninguno generalmente aceptado como 'mejor'". Es el investigador quien tiene que decidir qué *algoritmo* aplicar ante unos objetivos específicos de investigación. La elección del *algoritmo* de clasificación se ve afectada por los siguientes aspectos:

a) Los objetivos del estudio.
b) Las características de los datos a analizar: métrica de las variables y tamaño muestral, principalmente.
c) El método de conglomeración elegido: *jerárquico* y *no jerárquico*.
d) Los límites operativos impuestos por la capacidad del ordenador y, en especial, del programa estadístico utilizado al efecto.

De la conjunción de estos factores puede suceder que se tenga que elegir entre *algoritmos* igualmente aplicables a los datos concretos que se quiere clasificar. En tal circunstancia, la mejor decisión puede ser probar varios *algoritmos* y, a la vista de los resultados, elegir. Como bien apuntan Kaufman y Rousseeuw (1990: 37), "es permisible *probar* varios algoritmos en los mismos datos, porque el análisis de conglomerados principalmente se utiliza como una herramienta descriptiva o exploratoria, en contraste con las pruebas estadísticas que se llevan a cabo para propósitos confirmatorios o inferenciales".

A continuación se describen algunos de los *algoritmos* más aplicados en la investigación social. Ante la extensión que supondría informar de "todos" los *algoritmos*

propuestos en el análisis de conglomerados, se ha optado por reseñar aquéllos de uso más común. Y, dentro de ellos, los que proporcionan criterios distintos para la formación de los conglomerados. Así, por ejemplo, en los *algoritmos* pertenecientes a la conglomeración *no jerárquica* sólo se ha escogido uno de ellos en cada clasificación, como representación de los otros que comparten sus mismas características: "K-means" en representación de los métodos de *reasignación* (o de *partición iterativos*), el "análisis modal de Wishart" de los métodos de *búsqueda de densidad* y "block clustering" de los llamados métodos *directos*. En concreto, se han seleccionado los siguientes 13 *algoritmos* de clasificación:

a) Distancias mínimas.
b) Distancias máximas.
c) Promedio entre grupos.
d) Promedio intragrupos.
e) Método Ward.
f) Método del centroide.
g) Método de la mediana.
h) Partición binaria.
i) Método de Howard-Harris.
j) "K-means".
k) Análisis modal de Wishart.
l) "Block clustering".

En el cuadro 3.1 aparecen ubicados cada uno de estos *algoritmos* de clasificación. Los siete primeros (hasta el *método de la mediana*, inclusive) pertenecen a la conglomeración *jerárquica*, en especial, a los procedimientos *aglomerativos*, si bien, también pueden aplicarse cuando se realiza un análisis de conglomerados *jerárquico divisivo*. Dos de los más habituales son el *promedio entre grupos* y el *método de Ward*. Los dos *algoritmos* siguientes (*partición binaria* y el *método de Howard-Harris*) son dos *algoritmos* específicos a la *jerarquización divisiva*. En cambio, los tres últimos se incluyen en la conglomeración *no jerárquica*. Los *algoritmos* comúnmente llamados "K-means" y "Quick cluster", pertenecen a los métodos de *reasignación* o de *partición iterativos*. El *análisis modal de Wishart* constituye, por el contrario, una aproximación tipológica en los métodos de *búsqueda de densidad*. Mientras que "block clustering", propuesto por Hartingan, se incluye como *algoritmo* característico de los *métodos directos* de formación de conglomerados mediante procedimientos *no jerárquicos*.

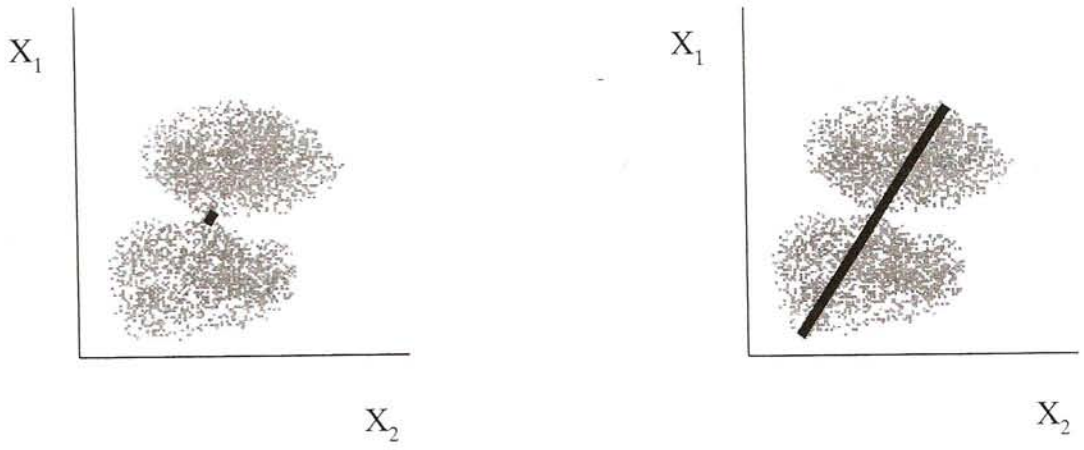
A) Distancias mínimas

Del inglés "single-link" (eslabón único), también conocido como del "vecino más próximo" ("nearest neighbour"). Fue propuesto por Sneath en 1957 (en "The application of computers to taxonomy", *Journal of General Microbiology*, 17: 201-226).

Constituye uno de los procedimientos más sencillos para formar conglomerados de manera jerárquica. Como su nombre expresa, los conglomerados se constituyen siguiendo el criterio de "distancia mínima". De acuerdo con este criterio, los objetos que se agrupan son aquellos que presentan la menor distancia entre ellos o, dicho en otros términos, los más semejantes.

Los dos primeros objetos que se combinan son los más próximos entre sí. Los otros objetos van, uno a uno, combinándose en un nuevo conglomerado, o uniéndose a un conglomerado ya existente, depende del conglomerado hacia el que se sitúe a menor distancia.

La distancia existente entre el nuevo objeto y el conglomerado es respecto al objeto en el conglomerado con quien el nuevo objeto tenga una menor distancia y, por consiguiente, una mayor similitud. La distancia entre dos conglomerados cualesquiera se calcula desde sus dos puntos (objetos) más próximos, como puede verse en la figura 3.2. En ella se comparan los *algoritmos* de *distancia mínima* y de *distancia máxima.*



a) Distancia mínima *b)* Distancia máxima

Figura 3.2. Comparación de los *algoritmos* de *distancia mínima* y *distancia máxima.*

Este proceder en la constitución de conglomerados presenta varias ventajas importantes:

1. Su mayor atractivo matemático. "Es el único método de agrupación jerárquica que satisface todas las condiciones" (Chatfield y Collins, 1980: 227).

   "Es invariante de transformaciones monotónicas de la matriz de similaridad y no está afectada por ataduras en los datos" (Aldenderfer y Blashfield, 1984: 38). La primera de estas propiedades se considera bastante importante. Significa que el procedimiento de creación de conglomerados no se verá afectado por cualquier transformación que se haga en los datos, siempre que ésta retenga el

mismo orden relativo de los valores en la *matriz de similaridad*. Esta propiedad no la presentan todos los *algoritmos aglomerativos*.

2. Su mayor facilidad de cálculo, que se materializa en dos aspectos importantes: uno, la mayor rapidez en la obtención de los conglomerados; dos, la posibilidad de llevarse a cabo con tamaños muestrales elevados.

Chatfield y Collins (1980: 227) matizan que "el método es bueno para los datos que tienen significación ordinal sólo".

Pero, como en todo proceder, también se observan desventajas. Su principal inconveniente es su tendencia a "encadenar" conglomerados, aparentemente distintos, por unos cuantos puntos intermedios que unen a ambos conglomerados. De esta forma se crean grandes conglomerados alargados, cuyos puntos extremos mantienen una gran distancia entre ellos. O, dicho con otras palabras, que son bastante disimilares entre sí, por lo que se pierde la homogeneidad en el conglomerado. Este inconveniente se da, sobre todo, en conglomerados que están escasamente delimitados.

B) Distancias máximas

Es una alternativa opuesta a la anterior, como expresa su propia denominación: "eslabón completo" ("complete-link"), el "vecino más alejado" ("furthest-neighbour") o, simplemente, la "distancia máxima". El criterio fundamental que rige la agrupación de los objetos es el opuesto al anterior: la distancia entre los dos objetos más alejados (y no entre los más próximos, como sucede en el *algoritmo* de *distancias mínimas*). Para que un nuevo objeto se incluya en un conglomerado ya existente es preciso que tenga un nivel elevado de *similaridad* con todos los miembros de ese conglomerado, y no sólo con aquél hacia el que tenga una menor *distancia*.

Esta consideración de las distancias hacia los miembros más distantes del conglomerado (que supone la valoración de todos sus integrantes) supone, inevitablemente, la aplicación de un criterio para la formación de conglomerados más riguroso que el aplicado en el *algoritmo* de *distancias mínimas*. Elimina la posibilidad de encontrar conglomerados encadenados, al no considerarse las distancias entre sus dos integrantes más próximos, sino entre los dos más alejados. Este proceder, sin embargo, suele provocar la creación de conglomerados "hiperesféricos, relativamente compactos y compuestos de casos bastante similares" (Aldenderfer y Blashfield, 1984: 39).

C) Promedio entre grupos

Los *algoritmos* que (para la vinculación de objetos) siguen el criterio de "promedio", ya sea entre grupos o intragrupal, fueron propuestos por Sokal y Michener en 1958 (en "A statistical method of evaluating systematic relationship", *University of Kansas Scientific Bulletin*, 38: 1409-1438). De ellos la variante más comúnmente aplicada

es, precisamente, la llamada "*promedio entre grupos*" ("average linkage between groups"), a veces también denominada UPGMA ("Unweighted Pair-Group Method using Arithmetic Averages").

Difiere de los dos *algoritmos* anteriores en que impone, como criterio básico de agrupación, la distancia *promedio* de los integrantes de un conglomerado respecto a los pertenecientes a otro conglomerado. En el cálculo de la *distancia* participan todos los integrantes del conglomerado y no sólo un único par de miembros extremos (ya sean los más próximos, ya los más alejados). Esta consideración de todos los integrantes del conglomerado convierte a este *algoritmo* de clasificación en uno de los más aplicados. De hecho, es el *algoritmo* que se oferta por defecto en programas tan populares como el SPSS.

Una vez que el investigador decide qué medida de distancia aplicar, se procede a calcular la distancia de cada objeto de un conglomerado con todos los objetos de los demás conglomerados. Después, se calcula el *promedio* de todos ellos. De esta manera, la distancia entre dos conglomerados queda definida como el *promedio* de las distancias entre todos los pares de objetos. Un miembro del par pertenece a cada uno de los conglomerados formados. Nourisis (1986) y Bisquerra (1989) lo ilustran con el siguiente ejemplo: si los individuos 1 y 2 componen el conglomerado A, y los individuos 3, 4 y 5, el conglomerado B, la distancia entre los conglomerados A y B será el *promedio* de las distancias entre los siguientes pares de casos: (1, 3), (1, 4), (1,5), (2, 3), (2, 4) y (2, 5).

Dos son los inconvenientes principales que se observan en la aplicación de este *algoritmo* de clasificación:

1. Tiende a combinar conglomerados con varianzas pequeñas.
2. Sesgo en la creación de conglomerados con aproximadamente la misma varianza (Hair *et al.*, 1992 y 1999).

D) Promedio intragrupal

Constituye la variante del *algoritmo* anterior. Los conglomerados que se agrupan son aquéllos cuya unión presenta la menor distancia promedio. De dos en dos se agrupan los objetos en conglomerados. Después, se calcula el *promedio* de las distancias de todos los integrantes del conglomerado, de acuerdo con la medida de distancia elegida. La combinación o agrupación de conglomerados se produce entre aquellos cuya distancia *promedio* entre todos los integrantes del conglomerado que resulte de la unión sea la menor posible.

E) Método Ward

También conocido como "momento central de orden dos" o "pérdida de inercia mínima". Fue diseñado por Ward en 1963 (en "Hierarchical grouping to optimize an ob-

jective function", *Journal of the American Statistical Association*, 58: 236-244). Su objetivo principal es "optimizar" la *varianza mínima intragrupal* (la suma de cuadrados intragrupal). Para este propósito, la distancia entre dos conglomerados pasa a definirse como la suma de cuadrados entre los dos conglomerados, sumados en todas las variables.

Calcula la *media* de todas las variables de cada conglomerado. Luego se procede al cálculo de la distancia (normalmente la distancia *euclídea al cuadrado*) entre cada objeto y la media del conglomerado en el que está incluido. A continuación, se suman las distancias de todos los objetos. En cada paso del proceso de agrupación se trata de combinar aquellos dos conglomerados que provoquen el menor incremento en la suma total de las distancias al cuadrado dentro de los conglomerados. Esto significa que se unen aquellos objetos o conglomerados que ocasionan un menor incremento de la *varianza intragrupal*. Ésta se trata de minimizar en todo el proceso de agrupación.

Este *algoritmo* provoca los mismos inconvenientes que el llamado de *distancias máximas*. Es decir, tiende a generar conglomerados de forma hiperesférica y de tamaño relativamente igual. Ello se debe a que están integrados, aproximadamente, con los mismos objetos. A esto se añade otro problema comúnmente observado: la tendencia a combinar conglomerados con un número pequeño de observaciones (Aldenderfer y Blashfield, 1984; Hair *et al.*, 1992 y 1999).

F) Método del centroide

La distancia se define como la habida entre los *centroides* grupales (o vectores de la media grupal). Éstos se obtienen de la *media* de las variables en el conglomerado, de manera que, el valor del *centroide* se ve afectado por los cambios que acontezcan en la composición de los conglomerados. Su valor cambia con cada variación en la configuración del conglomerado.

Si un nuevo conglomerado surge de la combinación de dos conglomerados ya existentes, el nuevo *centroide* será la combinación ponderada de los *centroides* correspondientes a los dos conglomerados individuales. Su peso será, asimismo, proporcional al tamaño de los conglomerados respectivos.

Frente a los otros *algoritmos* de clasificación, el método del *centroide* ofrece la ventaja de ser, de los procedimientos *jerárquicos*, el menos afectado por la presencia de *atípicos*. Pese a ello se observan dos inconvenientes principales en su aplicación:

1. Los últimos conglomerados que se forman suelen ser menos homogéneos que los creados en las fases iniciales del proceso. Esto se debe a la disminución que se produce en el valor de la distancia que permite la unión de dos conglomerados. Esta distancia disminuye paulatinamente de un paso al siguiente.
2. Como sucede con el *método Ward* y el método de la *mediana*, el método del *centroide* precisa, igualmente, que los datos a clasificar sean *métricos*. Este requisito limita bastante su aplicación en la investigación social, donde es habitual

la presencia de variables *cualitativas*. Cuando esto sucede, habrá que elegir preferentemente alguno de los otros *algoritmos* de clasificación adecuados a este tipo de variables. Además, estos últimos *algoritmos* permiten la aplicación de cualquier medida de *proximidad*. En los *algoritmos* del *centroide*, de la *mediana* y de *Ward*, sin embargo, la medida de *distancia* que se aplica normalmente es la *euclidea al cuadrado* (subapartado 3.3.4).

G) Método de la mediana

A las características mencionadas en el párrafo anterior, se añade una peculiaridad que le distingue de los otros *algoritmos* de clasificación (en especial, del método del *centroide*): cuando se aplica el método de la *mediana*, el tamaño del conglomerado no afecta al cálculo del *centroide*.

Tras la creación de un nuevo conglomerado, a partir de la combinación de dos ya existentes, el nuevo *centroide* (del conglomerado recién creado) es, igualmente, la combinación ponderada de los *centroides* correspondientes a los dos conglomerados individuales iniciales. Pero, a diferencia del método del *centroide*, en el método de la *mediana* el peso atribuido a cada conglomerado no es proporcional a su tamaño. Es independiente del número de integrantes que exista en cada conglomerado precedente.

Este proceder en la constitución de conglomerados favorece, indudablemente, a los conglomerados de menor tamaño. Éstos pueden tener el mismo efecto en la caracterización de nuevos conglomerados que los conglomerados de mayor número de componentes. A esta ventaja principal se suma la dicha al respecto en el método del *centroide*, al igual que sus inconvenientes.

H) Partición binaria

Este octavo algoritmo de clasificación se ubica en los métodos *jerárquicos divisivos* (o de *partición*). Por medio de él se pasa, de forma secuencial, de conglomerados más genéricos o globales a un pequeño número de conglomerados de menor tamaño, gracias a la "partición" o "división" de los conglomerados originales.

Su uso se restringe, preferentemente, a variables *binarias*, como indica su nombre: "de partición binaria". Esta peculiaridad favorece su aplicación cuando se analizan variables en su mayoría *binarias*. "Aunque los algoritmos divisivos generalmente son menos eficientes que los algoritmos aglomerativos, lo opuesto puede sostenerse con este tipo de datos" (Chatfield y Collins, 1980: 224).

I) Método de Howard-Harris

Otro de los algoritmos aplicados en la conglomeración *jerárquica divisiva* es el de *Howard-Harris*. Este último comparte con el anterior (el método de *partición binaria*)

el hecho de que los conglomerados se crean mediante la división de conglomerados de mayor tamaño. El procedimiento seguido es, igualmente, secuencial, aunque difiere en que su aplicación no se limita a variables *binarias,* extendiéndose a otros tipos de variables. A ello se suma la particularidad de que el proceso de formación de conglomerados sigue el criterio fundamental de que con cada división se logre hacer mínima la *varianza intragrupal.* Esto significa que se quiere crear conglomerados cuyos integrantes sean muy homogéneos entre sí y diferentes de aquellos que componen otros conglomerados. En consecuencia, la división de conglomerados en dos (o posteriormente más) conglomerados de menor tamaño se produce sólo cuando dicha división favorece la homogeneidad de los conglomerados.

J) K-means

Éste es el *algoritmo* más característico y de mayor aplicación en los métodos de conglomeración *no jerárquicos.* Fue diseñado por McQueen en 1967 (en "Some method for classification and analysis of multivariate observations", *Proceedings* 5^{th} *Berkeley Symposium,* 1: 281-296), como un procedimiento para la clasificación totalmente opuesto a la conglomeración *jerárquica.* Este procedimiento puede resumirse en cuatro pasos básicos:

1. El investigador especifica el número de conglomerados que deben formarse con los datos. El valor "K" expresa dicho número (por ejemplo, K = 4 conglomerados, 3, o los que se decida).
2. Se calculan los *centroides* iniciales de los conglomerados. En caso de no disponer de esta información previa (cuando no se parte, por ejemplo, de conglomerados ya constituidos mediante algún procedimiento *jerárquico* u otro *algoritmo* de clasificación), el programa informático que se use para su realización los estima iterativamente, utilizando los valores de los "K" primeros casos en el fichero de datos como estimaciones "provisionales" de los *centroides* (de las "K-medias" de los conglomerados; donde "K", recuérdese, expresa el número de conglomerados especificado por el investigador).
3. Mediante un proceso iterativo se asignan los objetos a los conglomerados a cuyo centro se sitúen más próximos. Para ello se calcula la distancia entre todos los objetos (casos o variables) y los *centroides.* La medida de distancia más utilizada en este *algoritmo* es la *distancia euclídea.*
4. Tras cada reasignación de los objetos a los conglomerados se vuelven a calcular los *centroides* de los conglomerados. Esto supone el cálculo de los valores promedio para las variables que caracterizan al conglomerado, tomando en consideración los objetos ahora asignados a los conglomerados.

   Los nuevos *centroides* pueden provocar una nueva reasignación de objetos a conglomerados a cuyo *centroide* se encuentren más próximos. A cada modificación en la composición de los conglomerados le sigue un nuevo cálculo de

los *centroides*, lo que puede provocar un nuevo desplazamiento de objetos a otros conglomerados. Y, así sucesivamente, hasta que un nuevo recálculo de los *centroides* no provoque ninguna alteración en la composición (volumen y características) de los conglomerados.

También puede suceder que se haya llegado al número máximo de iteraciones posible. En el programa SPSS, por ejemplo, el número máximo de iteraciones, aplicado por defecto, para actualizar los *centroides* mediante un procedimiento iterativo es 10. En cada una de dichas iteraciones los objetos se asignan por turnos al *centroide* más cercano. Cada iteración provoca un nuevo cálculo de los *centroides*.

Al final del proceso iterativo se obtiene los *centroides* finales. Es factible que éstos no coincidan con los "iniciales", sobre todo cuando se ha producido un número elevado de iteraciones y, en consecuencia, de modificaciones en la composición de los conglomerados.

A diferencia de los procedimientos *jerárquicos*, que proporcionan varias clasificaciones alternativas de los datos (al considerarse diversos números de conglomerados), *K-means* sólo proporciona una solución. Ésta se atiene al número de conglomerados previamente especificado por el analista. Este proceder es común a la conglomeración *no jerárquica* y tiene el peligro de no ajustarse a la realidad. Por esta razón, se recomienda que, a menos que se disponga de una clasificación previa (que resulte de un análisis de conglomerados inicial, principalmente *jerárquico*), se prueben varias clasificaciones alternativas. Esto se puede hacer de forma automática, dejando que el ordenador pruebe distintos valores de "K" y, finalmente, escoja el modelo más relacionado con algún criterio numérico. El investigador también puede realizar esta comprobación (de forma manual), probando varias soluciones, con diferentes valores de "K". Para ello realizará el análisis de conglomerados varias veces. De las distintas agrupaciones posibles escogerá aquella que proporcione una interpretación más significativa, tanto desde la vertiente estadística como de la lógico-sustantiva.

K-means (o *K-medias*) se presenta como una opción ideal cuando se manejan tamaños muestrales elevados (superiores a 200 unidades). Además, ayuda a la detección de casos *atípicos*, al proporcionar la distancia de cada caso al centro del conglomerado al que ha sido asignado: si se halla muy próximo o, por el contrario, está muy alejado, lo que le convierte en *atípico*. Éstas son dos de sus ventajas principales. A ellas se une otra que distingue a este *algoritmo* de clasificación: la posibilidad de identificar, con cierta precisión, aquellas variables que más contribuyen a la caracterización del grupo. Esta identificación de variables "relevantes" es posible porque este procedimiento de conglomeración incluye el cálculo de estadísticos F univariados para cada variable que compone el conglomerado.

En contra del procedimiento *K-means* está el no ser un *algoritmo* de aplicación universal. Al utilizar los *centroides* como criterio básico que determina la pertenencia al conglomerado, y aplicar la distancia *euclídea* para medir la distancia que separa al caso del *centroide*, su uso se limita a variables *métricas*. Asimismo, se recomienda que és-

tas hayan sido estandarizadas antes de comenzar el análisis. Éste es un requisito imprescindible cuando se analizan variables que presentan unidades de medición muy variadas, que provocan varianzas muy dispares.

Características similares a las enunciadas son compartidas por otros *algoritmos* de clasificación incluidos en los métodos de *reasignación*, también llamados de *partición iterativos*. Sea el caso, por ejemplo, del procedimiento conocido popularmente como "quick cluster", de uso también frecuente en la investigación social.

K) Análisis modal de Wishart

Pertenece a los *algoritmos* de clasificación que proporcionan una aproximación tipológica en los métodos de *búsqueda de densidad*, dentro de la conglomeración *no jerárquica*.

Como puede deducirse de su nombre, este *algoritmo* se distingue por buscar "puntos densos"; es decir, por localizar zonas donde exista una mayor concentración de objetos. Por esta razón se llama "análisis de la moda". Estos puntos están contenidos dentro de una hiperesfera de R radios. Partiendo de un valor pequeño de R, este procedimiento busca una hiperesfera de R radios alrededor de cada punto. Además, cuenta el número de otros puntos dentro de esta hiperesfera.

L) Block clustering

Este último *algoritmo* de clasificación fue propuesto por Hartingan en 1975 (en *Clustering algorithms*, Nueva York, John Wiley). A diferencia de los demás *algoritmos,* busca la conglomeración conjunta de individuos y de variables. Asimismo se distingue por adecuarse más a variables *cualitativas* (o *no métricas)*, preferiblemente, de menos de 10 categorías. Si la variable fuese, por el contrario, *continua*, habría que proceder previamente a su transformación en variable *categórica*.

Los conglomerados se forman, igualmente, mediante procedimientos iterativos. Mediante ellos se identifican aquellos "bloques" (casos por variable) que presentan un modelo similar sobre un conjunto de variables para cada caso. Los casos para cada variable pueden considerarse como conglomerados de variables (Bisquerra, 1989).

3.3.4. Medidas de distancia y de similaridad

De la lectura del subapartado anterior puede concluirse que los *algoritmos* de clasificación operan a partir de dos matrices de datos básicas:

a) Una *matriz* $N \times p$ (de casos por variables), donde las filas se corresponden a los casos y las columnas a las variables. Esto sucede cuando los casos se representan por sus atributos en las variables.

b) Una *matriz de proximidad* para todos los pares de objetos, ya sean casos ($N \times N$) o variables ($p \times p$). La matriz de *proximidad* puede ser, a su vez, de *distancia* o de *similaridad*:

1. De *distancia*, si mide lo alejado que se hallan dos objetos, uno respecto al otro.
2. De *similaridad*, cuando se mide la similitud o semejanza existente entre los objetos a ser clasificados. De modo que los valores elevados indican mayor "similitud" entre los objetos que se comparan, a diferencia de la matriz de *distancia*, en la que son los valores bajos los que expresan "similitud" entre los objetos.

Los criterios a seguir para determinar qué objetos se combinan para formar un conglomerado se basan, fundamentalmente, en alguna de estas dos matrices (de *distancia* o de *similaridad,* entre pares de objetos). Si bien, las opciones posibles son igualmente variadas.

Se puede, por ejemplo, calcular *coeficientes de correlación*, a partir de una matriz de *correlación*, como forma alternativa de comprobar la "similaridad" de los objetos. Como los valores próximos a 1,0 expresan semejanza y los situados en 0,0 disimilaridad, aquellos objetos que muestren intercorrelaciones elevadas (ya sean positivas o negativas) –en la matriz de *correlación*– comparten patrones similares.

No obstante, se advierte que "las medidas de correlación se utilizan rara vez porque el interés de la mayoría de las aplicaciones del análisis de conglomerados está en las magnitudes de los objetos, no en los patrones de los valores" (Hair *et al.,* 1999: 502-503).

Los conglomerados que se basan en medidas de *correlación* suelen caracterizarse por tener patrones similares. Los conglomerados basados en medidas de *distancia* tienen, en cambio, valores más parecidos para el conjunto de variables, aunque sus patrones sean bastante diferentes. Y, como el interés se sitúa más en las "magnitudes" que en los "patrones", el empleo de medidas de *distancia* o, en su caso, de *similaridad*, es más predominante en el análisis de conglomerados.

Ambas medidas (de *correlación* y de *distancia*) se adecuan más a variables *métricas* (o cuantitativas). Para variables *cualitativas* (o no métricas) existen unas medidas de *similaridad* alternativas. Para datos *binarios* se han propuesto "medidas de similaridad probabilística", como fueron denominadas por Sneath y Sokal (1973). Estas últimas se caracterizan por calcularse a partir de datos brutos, y no de matrices de *distancia-proximidad*. Mediante ellas se trata, igualmente, de comprobar cuál es la combinación de objetos que proporciona la mayor ganancia de información para proceder a su fusión y consiguiente constitución de conglomerados.

En resumen, desde la propuesta de Sneath y Sokal de 1973 (en *Numerical Taxonomy*, San Francisco, W. H. Freeman) son cuatro los tipos genéricos de *coeficientes de similaridad* que pueden aplicarse en el análisis de conglomerados:

a) Coeficientes de correlación.
b) Medidas de distancia.
c) Coeficientes de asociación.
d) Medidas de similaridad probabilística.

En la exposición de los distintos *coeficientes de similaridad* aquí se va a seguir la clasificación resumida en el cuadro 3.3. Esta clasificación responde al nivel de medición de las variables. Se ha elegido este criterio por ser el que más incide en la decisión de qué medida de *similaridad* o *distancia* escoger. En cada grupo, las opciones que se ofertan son varias. Lo más factible es que cada una provoque una solución de conglomeración diferente. Esto lleva a la recomendación reiterada (como sucede con los métodos de conglomeración y los *algoritmos* de clasificación) de aplicar, en los mismos datos, diversos *coeficientes de similaridad*. De las distintas soluciones se escogerá aquella que siendo relevante, desde la vertiente estadística, proporcione una interpretación más acorde en relación con el marco teórico de la investigación.

El cuadro 3.3 incluye una selección de medidas de *distancia* o *similaridad*. Aunque el cuadro sea extenso, hay que indicar que no logra abarcar el amplio abanico de medidas propuestas hasta la fecha.

A) Variables continuas

Cuando las variables son *continuas*, pueden aplicarse tanto medidas de *distancia* como de *similaridad*. Aunque son más habituales las primeras, en especial, la distancia *euclídea al cuadrado*.

A.1. Medidas de distancia

Los coeficientes que a continuación se exponen miden la "distancia" o "disimilaridad" entre los objetos a clasificar. Su valor siempre será positivo ($d_{ij} \geq 0$), no habiendo ningún límite superior al mismo. Si bien, cuanto mayor es su valor (más se distancie de 0), mayor es la disparidad entre los dos objetos medidos (i y j). Un $d_{ij} = 0$ expresa inexistencia de distancia entre los dos objetos. Lo que significa que son "idénticos" o muy similares, pudiéndose describir cada uno de ellos mediante las variables referidas al otro.

La *matriz de distancias* es una matriz cuadrada, que se confecciona a modo de la siguiente:

"p" variables

"N" objetos

$$\begin{array}{l} X_{11}X_{12}X_{13} \ldots\ldots\ldots\ldots X_{1P} \\ X_{21}X_{22}X_{23} \ldots\ldots\ldots\ldots X_{2P} \\ X_{31}X_{32}X_{33} \ldots\ldots\ldots\ldots X_{3P} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X_{N1}X_{N2}X_{N3} \ldots\ldots\ldots\ldots X_{NP} \end{array}$$

CUADRO 3.3. *Medidas de distancia o de similaridad, según el nivel de medición de las variables*

A. VARIABLES CONTINUAS	A.1. *Medidas de distancia*	• 1.1. Euclídea • 1.2. Euclídea al cuadrado 1.3. D^2 de Mahalanobis 1.4. De Manhattan o "city-block" • 1.5. De Chebychev • 1.6. De Minkowski 1.7. De un poder métrico absoluto
	A.2. *Medidas de similaridad*	• 2.1. Correlación de Pearson • 2.2. Cosenos de vectores de valores
B. VARIABLES BINARIAS	B.1. *Medidas de similaridad*	• 1.1. De Jaccard 1.2. De casación o parejas simples • 1.3. De Russel y Rao • 1.4. De Dice • 1.5. De Rogers y Tanimoto • 1.6. De Kulczynski 1 • 1.7. De Sokal y Sneath • 1.8. De correlación punto 4 phi (ϕ) • 1.9. De Ochiai • 1.10. De dispersión
	B.2. *Medidas de similaridad de probabilidades condicionales*	• 2.1. De Kulczynski 2 • 2.2. De Sokal y Sneath 4 • 2.3. De Hamann
	B.3. *Medidas de similaridad de predicción*	• 3.1. Lambda de Goodman y Kruskal • 3.2. D de Anderberg • 3.3. Y de Yule • 3.4. Q de Yule
	B.4. *Medidas de disimilaridad o distancia*	• 4.1. Euclídea binaria • 4.2. Diferencia de tamaño • 4.3. Diferencia de patrón 4.4. Diferencia binaria de forma 4.5. Varianza disimilar 4.6. De Lance y Williams
C. VARIABLES CUALITATIVAS NO BINARIAS	C.1. *Medidas de similaridad*	• 1.1. Chi-cuadrado • 1.2. Phi-cuadrado
D. VARIABLES EN DIFERENTES NIVELES DE MEDICIÓN	D.1. *Medidas de similaridad*	1.1. Coeficiente de similaridad de Gower

Las filas corresponden a los objetos (o casos) mientras que en las columnas se posicionan las "p" variables analizadas. En tamaños muestrales elevados, esta matriz adquiere una gran magnitud, lo que dificulta su correcta lectura e interpretación.

A.1.1. Distancia euclídea

Una de las medidas de distancia más populares, cuando se analizan variables *continuas,* es la *distancia euclídea.* Esta medida de distancia se obtiene de la aplicación del teorema de Pitágoras. Este teorema dice que la hipotenusa al cuadrado es igual a la suma de los cuadrados de los catetos, como ilustra la figura 3.4.

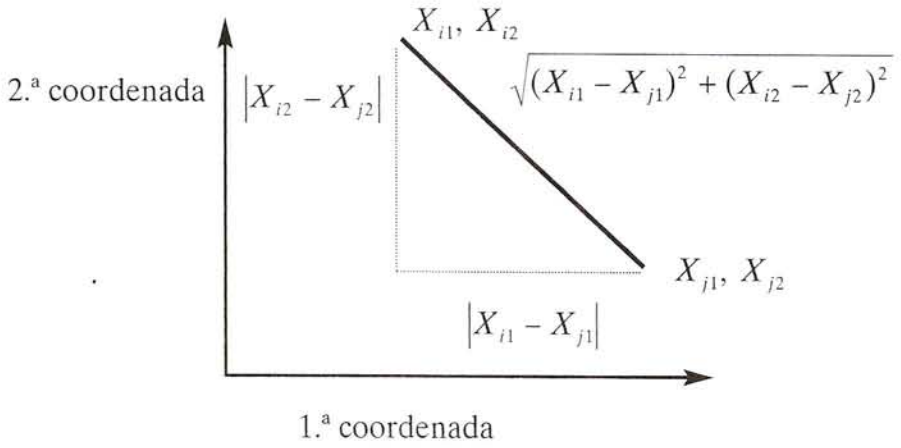


Figura 3.4. Representación gráfica del teorema de Pitágoras

En el gráfico puede verse que si "X_{i1}, X_{i2}" es un punto en el plano de coordenadas (X_1, X_2), correspondiente al objeto i, y "X_{j1}, X_{j2}" es otro punto del plano, perteneciente al objeto j, la distancia entre ambos puntos viene dada por:

$$d_{ij} = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2}$$

Siguiendo este teorema, la *distancia euclídea* se define como la raíz cuadrada de la suma de las diferencias cuadradas entre los valores de la variable K (X_K), para el objeto i y el objeto j. Y esto para todas las p variables que se analizan.

$$d_{ij} = \sqrt{\sum_{K=1}^{p} (X_{iK} - X_{jK})^2}$$

Donde: "d_{ij}" representa la distancia entre los casos i y j.
"X_{iK}" es el valor de la variable X_K para el caso i.
"X_{jK}" es el valor de la variable X_K para el caso j.

Cuando las variables están estandarizadas, las variables X_{iK} y X_{jK} se expresarían respectivamente como Z_{iK} y Z_{jK}.

A.1.2. Distancia euclídea al cuadrado

Es la medida de distancia empleada por defecto para datos de *intervalo* (Nourisis, 1994), en especial, cuando se agrupan casos. De hecho, es la medida recomendada en los *algoritmos* del *centroide* y de *Ward*.

Se define como la suma de las distancias (o diferencias cuadradas) entre los valores de la variable K, para los objetos i y j, en todas las variables analizadas.

$$d_{ij}^2 = \sum_{K=1}^{p} \left(X_{iK} - X_{jK} \right)^2$$

Un problema importante que se detecta en la aplicación de esta medida de *distancia* (aunque es igualmente extensible a otras medidas) es la desigual influencia de las variables, cuando éstas se hallan en diferentes unidades de medida. Las variables que incluyen valores más elevados (de mayor variabilidad) –sea el caso de la variable ingresos, por ejemplo– contribuyen más a la medida de distancia que aquellas que incluyen un menor rango de valores (como las variables edad, calificación académica u horas de descanso, por ejemplo).

La manera más popular de resolver este problema es *estandarizar* las variables. La *estandarización* tiene el efecto beneficioso de reducir la influencia del tamaño relativo de las variables. Pese a ello, diversos autores (como Aldenderfer y Blashfield, 1984; Nourisis, 1986) advierten que la variabilidad de una medida particular puede proporcionar información útil. La *estandarización*, en cambio, puede provocar el efecto adverso de minimizar las diferencias grupales. Su aplicación se haría antes de proceder al cálculo de la distancia y, preferentemente, cuando el rango de una variable sea bastante superior al de otras variables, cuya influencia conjunta en la clasificación de objetos trata de medirse.

EJEMPLO DEL EFECTO DE LA ESTANDARIZACIÓN EN EL CÁLCULO DE LA DISTANCIA EUCLÍDEA

Como ilustración del efecto de la *estandarización* en el cálculo de la distancia *euclídea* se toma el ejemplo dado en el manual del SPSS (1997) por su claridad expositiva. Dice lo siguiente:

	Unidades originales		*Unidades estandarizadas*	
	Edad	*Renta*	*Edad*	*Renta*
Juan	45	7	1,1	1,9
David	30	2	0,1	0,1

La variable "edad" se halla medida en años y la variable "renta mensual", en miles de dólares. Aplicando la fórmula de la distancia *euclídea al cuadrado* para los dos casos (Juan y David), respecto a las dos variables analizadas, se obtienen los valores siguientes:

• Tomando las unidades de medición originales de las variables:

$$(45 - 30)^2 + (7 - 2)^2 = 15^2 + (-5)^2 = 225 + 25 = 250$$

• En unidades estandarizadas:

$$(1{,}1 - 0{,}1)^2 + (1{,}9 - 0{,}1)^2 = 1^2 + 1{,}8^2 = 1 + 3{,}24 = 4{,}24$$

Si se comparan ambos resultados, puede observarse que la variable "edad", en su nivel de medida original (en años), supone el 90% de la medida de *distancia*. Al transformarse en unidades *estandarizadas*, su influencia se reduce al 23,6%, lo que muestra el efecto de la *estandarización*, de minimizar las diferencias grupales. Esto debería considerarse cuando se esté ante la decisión de tomar a las variables en sus unidades de medición originales o proceder a su estandarización

---

A.1.3. Distancia D^2 de Mahalanobis

Propuesta por Mahalanobis en 1927, aunque su divulgación se posterga a 1936 (en "On the generalized distance in statistics", India: *Proceedings of the National Institute of Science*, 12: 49-55).

Constituye una extensión de la distancia *euclídea*, a la que incorpora la *estandarización*. Ello permite medir las respuestas en unidades de desviación típica, además de realizar los ajustes mediante intercorrelaciones entre las variables.

En el cálculo de la distancia ahora interviene la *matriz de varianza-covarianza* (Σ), que ajusta para las intercorrelaciones entre las variables. Esto es importante porque en la aplicación del análisis de conglomerados se ha observado (Hair *et al.*, 1992 y 1999) que series bastante intercorrelacionadas de variables pueden, implícitamente, sobreponderar una serie de variables en la formación de los conglomerados. Al incluirse la *matriz de varianza-covarianza* se trata, precisamente, de evitar esto. La distancia de *Mahalanobis* ajusta para las intercorrelaciones y pondera todas las variables de igual manera, lo que la convierte en la medida de distancia más apropiada cuando las variables están bastante intercorrelacionadas positiva y/o negativamente.

La distancia entre dos objetos se obtiene, mediante la medida de *Mahalanobis*, del producto siguiente: $d_{i,j} = (x_i - x_j)' \sum^{-1} (x_i - x_j)$

Donde: "$\sum^{-1}$" es la inversa de la matriz de varianza-covarianza intragrupos.
"x_i" y "x_j" son vectores de los valores de las variables para los objetos i y j.
La prima (´) indica la matriz transpuesta.

Asimismo, la distancia de un objeto al centro del grupo (*centroide*) viene dada por:

$$D^2 = (x - \overline{X})' \sum^{-1} (x - \overline{X})$$

Su valor expresa la distancia del objeto hacia el centro (o *centroide*). Éste se define en consonancia con el conjunto de variables que configuran el conglomerado y le diferencian del resto. Valores D^2 muy elevados indican que el objeto al que corresponde el valor se halla muy distanciado del *centroide* del conglomerado donde se le ha clasificado. De esta forma, D^2 ayuda a la detección de *atípicos* ("outliers"): los objetos situados fuera del rango de valores esperados en una variable.

La D^2 de *Mahalanobis* es comparable a la aplicación de R^2 en el análisis de regresión *lineal* en la medición de la distancia entre objetos. Pero no todos los programas estadísticos incluyen esta medida de distancia. Cuando esto sucede, las preferencias se decantan por la distancia *euclídea al cuadrado*. Además, ambas medidas de distancia son equivalentes, cuando las variables no están correlacionadas.

El cálculo de la D^2 de *Mahalanobis* ofrece, como ventaja, la posibilidad de estimar la F de *Fisher* y utilizarla como prueba de contraste (Bisquerra, 1989).

$$F = D^2 \frac{n_p n_q (n_p + n_q - v - 1)}{(n_p + n_q)(n_p + n_q - 2)v}$$

La F de *Fisher* se distribuye según la distribución F de *Snedecor*, siendo sus grados de libertad "v" y "$n_p + n_q - v - 1$", respectivamente. "n_p" y "n_q" indican el número de objetos de las poblaciones correspondientes.

A.1.4. Distancia de Manhattan o "city-block"

A diferencia de las medidas anteriores, la distancia de *Manhattan* considera la suma de las diferencias absolutas de los valores de las variables y no su cuadrado. Ello incide en la menor ponderación de las diferencias grandes.

$$d_{ij} = \sum_{K=1}^{P} |X_{ik} - X_{jk}|$$

Los requisitos matemáticos son los mismos de la función de distancia *euclídea*. A decir:

a) La distancia siempre es un número positivo: $d_{ij} \geq 0$.
b) La distancia de un objeto consigo mismo es cero: $d_{ij} = 0$.
c) Simetría de la función de distancia: $d_{ij} = d_{ji}$.
d) Desigualdad de triángulo: ir directamente de i a j es más corto que hacer una desviación sobre el objeto h: $d_{ij} \leq d_{ih} + d_{hj}$.

Su uso se recomienda en "aquellas situaciones donde, por ejemplo, una diferencia de 1 en la primera variable y de 3 en la segunda variable es la misma que una diferencia de 2 en la primera variable y de 2 en la segunda variable" (Kaufman y Rousseew, 1990: 13).

A.1.5. Distancia de Chebychev

Difiere de la distancia de *Manhattan* en que sólo considera la diferencia absoluta máxima de los valores de las variables. Lo que supone "ignorar mucha de la información disponible" (Nourisis, 1986: B-82). Su definición es la siguiente:

$$d_{ij} = Máx|X_{iK} - X_{jK}|$$

A.1.6. Distancia de Minkowski

Se considera una generalización de la distancia *euclídea* y la de *Manhattan*. Se define como la raíz "q" de la suma de las diferencias absolutas a la potencia "q" entre los valores de la variable K para el caso "i" y el caso "j".

$$d_{ij} = \sqrt[q]{\sum_{K=1}^{P} |X_{iK} - X_{jK}|^q}$$

Donde "q" es un número real ≥ 1. Cuando q = 2, se está ante la distancia *euclídea*. Si q = 1, es la distancia de *Manhattan*.

A.1.7. Distancia de un poder métrico absoluto

Se distingue de la distancia de *Minkowski* en que la raíz y la potencia a la que se eleva la diferencia entre los valores de la variable difieren.

$$d_{ij} = \sqrt[r]{\sum_{K=1}^{P} |X_{iK} - X_{jK}|^q}$$

A.2. Medidas de similaridad

En vez de emplear un coeficiente de distancia (d_{ij}) para medir lo alejados que están dos objetos ("i" y "j"), alternativamente se puede aplicar un *coeficiente de similaridad* (s_{ij}). A diferencia de los *coeficientes de distancia*, los de *similaridad* miden la "semejanza"o la "proximidad" existente entre dos objetos. Por esta razón, ahora interesa la obtención de *coeficientes de similaridad* elevados porque expresan "similitud" entre los objetos a clasificar.

Los *coeficientes de similaridad* típicamente tienen un rango de valores de 0 a 1. El 1 expresa *similaridad* máxima, mientras que el 0, inexistencia de *similaridad.*

Estos coeficientes cumplen las mismas condiciones que los *coeficientes de distancia*. Éstas se resumen en tres principales:

a) $0 \leq s_{ij} \leq 1$
b) $s_{ij} = s_{ji}$
c) $s_{ii} = 1$

Sus valores se obtienen de una matriz de *similaridad* ($N \times N$). Adviértase que sus valores son los adversos a los de *distancia*, ya que: $s_{ij} = 1 - d_{ij}$.

A.2.1. Correlación de Pearson

El coeficiente de correlación más popular es el de *Pearson*, aunque igualmente pueden aplicarse otros, como el de Spearman o el de Kendall.

El *coeficiente de correlación de Pearson* originariamente se define como "un método para correlacionar variables". También se ha utilizado "en la clasificación cuantitativa para determinar la correlación entre casos" (Aldenderfer y Blashfield, 1984: 22). En este contexto, este coeficiente se define del modo siguiente:

$$r_{ij} = \frac{\sum_{K=1}^{P} (X_{ij} - \overline{X}_j)\left(X_{iK} - \overline{X}_K\right)}{\sqrt{\sum_{K=1}^{P} (X_{ij} - \overline{X}_j)^2 (X_{ik} - \overline{X}_K)^2}}$$

Donde: "X_{ij}" es el valor de la variable i para el objeto j.
"$\overline{X}_j$" es la media de todos los valores de las variables para el objeto j.

Cuando las variables están estandarizadas, el símbolo "Z" sustituye a "X" en la fórmula.

A diferencia de otras medidas de *similaridad*, el rango de valores del *coeficiente de correlación de Pearson* va de –1,00 a +1,00. Si $r_{ij} = 0$, la relación entre los objetos es inexistente. La relación entre ellos aumenta conforme más se aproxime r_{ij} a 1.

Pero, a diferencia del análisis de regresión lineal, en el análisis de conglomerados el signo que acompaña al *coeficiente de correlación* (r) no siempre es interpretable. Se toma el valor absoluto del coeficiente como medida de *similaridad*, al proporcionar el grado de relación entre las variables. El signo sólo indica la dirección de la relación. De ahí que se recomiende (Nourisis, 1994) mantener el signo sólo cuando se quieren conglomerados únicamente para variables correlacionadas positivamente.

La aplicación de la *correlación de Pearson* es más habitual cuando se desea conglomerar variables, más que casos. Para estos últimos, la distancia *euclídea al cuadrado* suele presentarse como la mejor opción.

A.2.2. Cosenos de vectores de valores

Una medida de *similaridad* de uso menos generalizado que el coeficiente de correlación, que se obtiene del siguiente cociente:

$$Cos_{ij} = \frac{\sum_{K=1}^{P} X_{iK} X_{jK}}{\sqrt{\sum_{K=1}^{P} X_{iK}^2 \sum_{K=1}^{P} X_{jK}^2}}$$

B) Variables binarias

Las variables *binarias* son variables que incluyen sólo dos opciones de respuesta, a modo de las siguientes: "a favor"-"en contra"; "sí"-"no"; "varón"-"mujer"; "aprobado"-"suspenso". En la matriz de datos estas variables normalmente aparecen con los códigos numéricos 1 o 0. El código 1 suele aplicarse para denotar la presencia del atributo que se mide (por ejemplo, "a favor", "sí", "aprobado"). El código numérico 0 se atribuye, por el contrario, a la inexistencia del atributo en cuestión ("en contra", "no", "suspenso"). Pero, aunque ésta sea la codificación predominante, existen otros procedimientos de codificación alternativos. La mayoría de los paquetes estadísticos permiten la opción de emplear otros valores *integer* para indicar la presencia o la ausencia del atributo de una variable. Recuérdese lo dicho al respecto en la creación de variables *ficticias*, que adquieren la codificación *binaria*.

Las variables *nominales* pueden, de hecho, transformarse en *binarias*. Si la variable incluye más de dos categorías (por ejemplo, las variables religión, estado civil, nacionalidad), cabe la opción de convertir cada categoría de la variable en *binaria*. De esta forma, la variable "estado civil", por ejemplo, se transforma en 5 variables *binarias*. A decir,

	Categorías u opciones de respuesta	
Variables	*Sí*	*No*
X_1 Soltero	1	0
X_2 Casado	1	0
X_3 Viudo	1	0
X_4 Divorciado/separado	1	0
X_5 En pareja	1	0

Asimismo, podría decidirse dejar la variable "estado civil" en una única variable *binaria*. Ésta podría ser, por ejemplo, soltero 1, no soltero 0. Las otras categorías diferentes a "soltero" quedarían, de esta forma, agrupadas en la opción "no soltero".

Esta última alternativa de codificación *binaria* presenta, no obstante, el gran inconveniente de suponer una pérdida importante de información. Esto incide en la aplicación más generalizada de la primera opción de codificación *binaria*, aunque suponga una mayor complejidad en los análisis. Consiste en la transformación de toda variable *nominal* en varias variables *binarias*. Su número lo determina el número de categorías que incluya la variable original.

Respecto a las variables *ordinales*, el proceder más habitual es darles el tratamiento de variables de *intervalo*. Su tratamiento a modo de variables *nominales* supone una pérdida de información relevante.

Por último, si se quiere analizar, de forma conjunta, variables en diferentes niveles de medición (*métricas* y *no métricas*), como es práctica habitual en la investigación social, se puede optar por transformar todas las variables de interés en variables *binarias*. Ello facilitaría su tratamiento conjunto. En este caso, las variables de *intervalo* se transformarían en *binarias*. Un procedimiento a seguir sería seleccionar un valor "central" y atribuir el código numérico 0 a todo valor que se sitúe por debajo de dicho referente; y el código 1, para los que se hallan por encima del valor "central" elegido. Por ejemplo, en la variable "edad" podría elegirse como valor central (a la vista de las frecuencias de respuesta) la edad de "30 años". Todo sujeto con edad ≤ 30 años se codificaría 0, mientras que los de edad > 30 años recibirían el código 1.

El principal inconveniente que se observa en la aplicación de la codificación *binaria* en variables de *intervalo* es, obviamente, la pérdida de información que supone su transformación.

Otra opción alternativa para el análisis conjunto de variables en distintos niveles de medición es proporcionar a todas las variables el tratamiento de variables *continuas*. Esta actuación se adecua bastante en variables *binarias simétricas*, para los rangos que se originan de variables *ordinales* (de 0 a 1) y para los logaritmos de las variables *de razón*, pero no para las variables *nominales* con más de dos categorías. La razón está en que algunos códigos pueden estar bastante alejados unos de otros, sin que reflejen una "lejanía" intrínseca de los estados correspondientes (Kaufman y Rousseeuw, 1990).

Dada la complejidad del tratamiento conjunto de variables mezcladas, el investigador deberá elegir el tratamiento conjunto que proporcionará a dichas variables: *continuo* o *binario*. Para ello deberá sopesar las ventajas e inconvenientes principales de cada alternativa de actuación.

En caso de indecisión, la mejor opción puede ser ejecutar, en análisis separados, las distintas alternativas de tratamiento conjunto expuestas. A la vista de los resultados, se escogerá aquella opción que proporcione resultados más "significativos", analítica y conceptualmente.

- Respecto a las *medidas de distancia*, para las variables *binarias* se han propuesto varias *medidas de similaridad* y de *disimilaridad* específicas. Estas medidas se adecuan más a las características de estas variables que las medidas aplicadas en variables *continuas*. Para facilitar la comprensión de las diversas medidas (de *similaridad*

y de *disimilaridad*), a continuación se muestra cómo se configura la matriz de *similaridad* (o en su caso de *distancia*) para variables *binarias*:

		Objeto j 1	Objeto j 0	
Objeto i	1	a	b	a + b
	0	c	d	c + d
		a + c	a + d	p

Los coeficientes propuestos para medir la *similaridad* en variables *binarias* superan la treintena (Aldenderfer y Blashfield, 1984). Ante su elevado número, se descarta una exposición detallada de cada uno de los coeficientes existentes. Se prefiere resaltar los más aplicados en la investigación empírica, por su disponibilidad en los paquetes estadísticos estándares, como el SPSS.

B.1. Medidas de similaridad

También se las conoce como "medidas de asociación de similitud". A excepción de los coeficientes 1.° de Kulczynski y el 3.° de Sokal y Sneath, la generalidad de las *medidas de similaridad* alcanza un valor en el rango de 0 a 1. La plena correspondencia entre los objetos a clasificar se logra cuando el coeficiente de *similaridad* se aproxima al valor 1. Por el contrario, un valor próximo a 0 significa la plena divergencia de los objetos.

B.1.1. Coeficiente de Jaccard

Igualmente conocido como "razón de similaridad" (Nourisis, 1994). Constituye una de las medidas de similaridad más aplicadas en variables *binarias*, especialmente, en el campo de la biología. Su rasgo más característico es que no considera las "ausencias" conjuntas de la variable en los dos objetos observados. Estas ausencias se expresan con la letra "d" en la matriz de asociación correspondiente, como puede verse en la ilustración de la tabla anterior. El coeficiente queda entonces definido de la manera siguiente:

$$S_{ij} = \frac{a}{(a + b + c)}$$

Las co-ocurrencias positivas de la presencia del atributo de la variable que se mide en dos objetos (i y j), representada por la letra "a", se divide por la suma de las situaciones en las que la presencia del atributo coincide en ambos casos ("a") o, al menos, en uno de ellos ("b" y "c"). No se considera, por tanto, la situación de "no presencia" del atributo en ambos objetos ("d").

Como ya se ha indicado, este coeficiente puede tener un valor comprendido en el rango de 0 a 1. El valor 1 expresa la plena correspondencia o similitud entre los objetos a clasificar, mientras que 0, su divergencia.

B.1.2. Coeficiente de casación o de parejas simples

Un coeficiente asimismo muy aplicado en la medición de la *similaridad* en variables binarias. Del coeficiente anterior le distingue la inclusión de la "ausencia conjunta" de una variable (como se indica en la celdilla con la letra "d" de la matriz de *similaridad*). Esta *ausencia conjunta* se incluye tanto en el numerador como en el denominador de la ecuación.

$$S_{ij} = \frac{a + d}{a + b + c + d}$$

El rango de valores posibles es, igualmente, de 0 a 1.

B.1.3. Coeficiente de Russel y Rao

Como el coeficiente de *casación*, incluye la situación de *ausencia conjunta* del atributo de la variable en los dos objetos. Pero, a diferencia de dicho coeficiente, sólo en el denominador:

$$S_{ij} = \frac{a}{a + b + c + d}$$

B.1.4. Coeficiente de Dice

También llamado "medida de Czekanowski y Sorensen". Se asemeja al coeficiente de *Jaccard*, al no considerar las *ausencias conjuntas*. Pero, difiere de él, al conceder un peso doble a la *coincidencia positiva* (celdilla "a" en la matriz de *similaridad*). Su definición es la siguiente:

$$S_{ij} = \frac{2a}{2a + b + c}$$

B.1.5. Coeficiente de Rogers y Tanimoto

Como los coeficientes de *casación* y de *Russel y Rao*, este quinto coeficiente incluye las *ausencias conjuntas* del atributo de la variable. Pero, su particularidad consiste en conceder un doble peso a las *no coincidencias*, representadas por las celdillas "c" y "b".

$$S_{ij} = \frac{a+d}{a+d+2(b+c)}$$

B.1.6. Coeficiente de Kulczynski 1

Al igual que los coeficientes de *Jaccard* y de *Dice*, el de *Kulczynski* 1 no incluye la *ausencia conjunta* (también llamada "coincidencias o co-ocurrencias negativas"), indicada en la celdilla "d". Se limita a dividir las *coincidencias positivas* (celdilla "a") entre la suma de las *no coincidencias* (celdillas "b" y "c"). Pero, a diferencia del coeficiente de *Dice*, no concede más importancia (doble peso) a las *coincidencias positivas* (celdilla "a"). Queda definido de la manera siguiente:

$$S_{ij} = \frac{a}{b+c}$$

A diferencia de los cinco coeficientes precedentes, el coeficiente de *Kulczynski* 1 puede presentar un valor superior a 1. El valor mínimo sigue siendo, no obstante, 0.

La indefinición se alcanza cuando existe pleno acuerdo. Dicho en otros términos, cuando no existe ninguna *no-coincidencia*. En consecuencia, b = 0 y c = 0.

El programa SPSS, por ejemplo, asigna un límite superior artificial de "9999.999", cuando existe indefinición o el valor del coeficiente es mayor de "1".

B.1.7. Coeficientes de Sokal y Sneath

Estos autores proponen cuatro coeficientes de *similaridad:*

- Primero: $S_{ij} = \dfrac{2(a+d)}{2(a+d)+b+c}$

  Se otorga doble peso a la suma de las *coincidencias negativas* (celdilla "d") y *positivas* (celdilla "a").

- Segundo: $S_{ij} = \dfrac{a}{a+2(b+c)}$

  En este segundo coeficiente, el doble peso se concede a las *no coincidencias* (celdillas "b" y "c").

- Tercero: $S_{ij} = \dfrac{a+d}{b+c}$

  La suma de las *coincidencias positivas* y *negativas* se divide por la suma de las *no coincidencias*.

  Al igual que el primer coeficiente de *similaridad* de *Kulcznski*, este tercer coeficiente de *similaridad* propuesto por Sokal y Sneath también puede presentar un valor superior a 1. El programa SPSS, por ejemplo, asigna igualmente un límite superior artificial de 999,999, cuando está indefinido o excede el valor de 1. La indefinición ocurre cuando no existe ninguna *no coincidencia*. El valor mínimo, sin embargo, continúa siendo 0.

- Cuarto: $S_{ij} = \dfrac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$

  En realidad ésta es la quinta medida de *similaridad* de Sokal y Sneath. En ella se conjugan las *coincidencias* con las *no coincidencias*. El rango de valores posibles va de 0 a 1.

B.1.8. Coeficiente de correlación punto 4 phi (ϕ)

Es la configuración, en forma *binaria*, del coeficiente de correlación *phi de Pearson*. Si este coeficiente se compara con la quinta medida de *similaridad de Sokal y Sneath* precedente, podrá comprobarse que ambos guardan bastante similitud. Tan sólo les diferencia que en el numerador, al producto de las *coincidencias positivas* y *negativas*, se le resta el producto de las *no coincidencias* en el *coeficiente de correlación punto 4 phi*.

$$S_{ij} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

El rango de valores posibles es, asimismo, de 0 a 1. El valor 0 expresa la inexistencia de *similitud* entre los objetos.

B.1.9. Coeficiente de Ochiai

La alternativa binaria del *coseno* (para variables *continuas*). Su rango va de 0 a 1, con igual interpretación que en los demás coeficientes de *similitud*.

$$S_{ij} = \sqrt{\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)}$$

B.1.10. Coeficiente de dispersión

Coincide con el *coeficiente de correlación punto 4 phi de Pearson* en el numerador, pero no en el denominador. Éste está integrado, en el *coeficiente de dispersión*, por la suma cuadrada de todas las *coincidencias* y las *no coincidencias*. También difiere en el rango de valores, ahora va de –1 a +1. Es decir, incluye valores positivos y negativos, pero con interpretación similar al resto de coeficientes.

$$S_{ij} = \frac{ad - bc}{(a + b + c + d)^2}$$

B.2. Medidas de similaridad de probabilidades condicionales

La probabilidad de que se presenten *coincidencias* o *divergencias* se condiciona a la existencia o de *coincidencias positivas* (celdilla "a") o de *negativas* (celdilla "d"). Tres son las medidas más comúnmente clasificadas en este segundo bloque de *medidas de similaridad* para datos *binarios* (Nourisis, 1986; 1994; Bisquerra, 1989).

B.2.1. Coeficiente de similaridad de Kulczynski 2

$$S_{ij} = \frac{a/(a + b) + a/(a + c)}{2}$$

Su valor oscila entre 0 (objetos dispares) y 1 (objetos plenamente similares).

B.2.2. Coeficiente de similaridad de Sokal y Sneath 4

$$S_{ij} = \frac{a/(a + b) + a/(a + c) + d/(b + d) + d/(c + d)}{4}$$

La probabilidad condicional de una característica de un objeto está en la misma situación (presencia "a" o ausencia "d") que la característica del otro objeto. La medida es, igualmente, un promedio de ambos objetos actuando como predictores. El rango de valores va de 0 a 1.

B.2.3. Coeficiente de Hamann

A diferencia de los anteriores, en el *coeficiente de Hamann* la probabilidad de que una característica tenga el mismo estado en ambos objetos –quiere esto decir, *presente*

(celdilla "a") o *ausente* (celdilla "d") en ambos– se resta la probabilidad de que una característica tenga estados distintos en los dos objetos: *presente* en uno pero, en cambio, *ausente* en el otro. Su valor va de –1 a +1:

$$S_{ij} = \frac{(a+d)-(b+c)}{a+b+c+d}$$

B.3. Medidas de similaridad de predicción

B.3.1. Coeficiente lambda (λ) de Goodman y Kruskal

Este coeficiente de predictibilidad se define, cuando se aplica al análisis de conglomerados, de la siguiente manera:

$$S_{ij} = \frac{\text{máx}(a,b)+\text{máx}(c,d)+\text{máx}(a,c)+\text{máx}(b,d)}{2(a+b+c+d)} - \frac{\text{máx}(a+c,b+d)+\text{máx}(a+b,c+d)}{2(a+b+c+d)}$$

Mide la reducción proporcional en el error de predicción del valor de la característica de un objeto (*presencia* o *ausencia*), a partir del valor de otro objeto.

Se basa, igualmente, en los valores máximos de la conjunción de las celdillas. El rango de valores va de 0 a 1. El valor 1 expresa la total predictibilidad de la *presencia* o *ausencia* de dicha característica en el objeto, a partir del conocimiento del otro objeto. Ambos son bastante similares. Por el contrario, el valor 0 denota inexistencia de *similaridad* entre los objetos, no pudiéndose predecir la situación de uno de los objetos desde el conocimiento del otro objeto.

B.3.2. Coeficiente D de Anderberg

Es otro coeficiente de predictibilidad de la situación de un objeto respecto a una variable concreta (*presencia* o *ausencia*), a partir de la situación de otro objeto. Dicha predictibilidad es igualmente factible, si ambos objetos son "similares". Pero, a diferencia del *coeficiente lambda*, en el *coeficiente D* los dos cocientes de la ecuación no se restan, sino que se suman. El coeficiente queda definido en los términos siguientes:

$$S_{ij} = \frac{\text{máx}(a,b)+\text{máx}(c,d)+\text{máx}(a,c)+\text{máx}(b,d)}{2(a+b+c+d)} + \frac{\text{máx}(a+c,b+d)+\text{máx}(a+b,c+d)}{2(a+b+c+d)}$$

El rango de valores va, asimismo, de 0 a 1.

B.3.3. Coeficiente Y de Yule

Se define como la razón entre las diferencias de las raíces cuadradas de las *coincidencias* (celdillas "a" y "d") y las *no coincidencias* (celdillas "b" y "c") y la suma de ambas raíces. Puede presentar un valor de –1 a +1.

$$S_{ij} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

B.3.4. Coeficiente Q de Yule

Difiere del *coeficiente Y* en que los productos de *coincidencias* y *no coincidencias* no están en raíz cuadrada. Su rango también va de –1 a +1.

$$S_{ij} = \frac{ad - bc}{ad + bc}$$

B.4. Medidas de disimilaridad o distancia

Se presentan como la alternativa para variables *binarias* del hacer en las variables *continuas*. Algunas de las medidas, como la distancia *euclídea*, comparten nominación con las medidas de *distancia* para variables *continuas*. Pero no su definición, que cambia para adecuarse a la métrica correspondiente a variables *binarias*. Los coeficientes de *disimilaridad* o *distancia* más comúnmente aplicados son los siguientes (Nourisis, 1986, 1994; Bisquerra, 1989):

B.4.1. Distancia euclídea binaria

$$d_{ij} = \sqrt{b + c}$$

Se obtiene de la raíz cuadrada de la suma de *no coincidencias*. Su valor mínimo es 0. No tiene límite superior.

En programas como el SPSS la distancia *euclídea al cuadrado* se aplica también por defecto en datos *binarios*. Su rango va de 0 a infinito, dada la inexistencia de límite superior.

$$d_{ij}^2 = b + c$$

B.4.2. Diferencia de tamaño

Guarda bastante similitud con el *coeficiente de dispersión* (B.1.10), aunque difiere en el numerador. Éste se halla integrado por el cuadrado de las diferencias entre las

no coincidencias. También difiere en el rango de valores que va de 0 hasta infinito al no haber ningún límite superior.

$$d_{ij} = \frac{(b-c)^2}{(a+b+c+d)^2}$$

B.4.3. Diferencia de patrón

A diferencia del coeficiente anterior, en éste el numerador se halla integrado exclusivamente por el producto de las *no coincidencias*. El denominador, en cambio, es el mismo. El rango de valores va de 0 a 1, que es el límite superior.

$$d_{ij} = \frac{bc}{(a+b+c+d)^2}$$

B.4.4. Diferencia binaria de forma

$$d_{ij} = \frac{(a+b+c+d)(b+c)-(b-c)^2}{(a+b+c+d)^2}$$

No tiene ni límite superior ni inferior. Los valores posibles son infinitos.

B.4.5. Varianza disimilar

$$d_{ij} = \frac{bc}{4(a+b+c+d)}$$

Su formulación se asemeja bastante al coeficiente de *diferencia de patrón*. Difiere en el denominador. El producto de *no coincidencias* se divide por cuatro veces la suma de *coincidencias* y *no coincidencias*. Tampoco coincide en el rango de valores posibles. Como la distancia *euclídea binaria* y el coeficiente de *diferencia de tamaño*, la *varianza disimilar* no tiene ningún límite superior. Su valor mínimo es, igualmente, 0.

B.4.6. Medida de disimilaridad no métrica binaria de Lance y Williams

También se la conoce como coeficiente *no métrico de Bray-Curtis*. Difiere de las cuatro últimas medidas de *disimilaridad* en que no considera las *coincidencias* negativas (celdilla "d"). El rango de valores va de 0 a 1.

$$d_{ij} = \frac{b + c}{2a + b + c}$$

C) Variables cualitativas, no binarias

Cuando las variables que se analizan para formar conglomerados son *cualitativas*, pero no *binarias*, la *similaridad* se mide con la ayuda de dos coeficientes básicos: *chi-cuadrado* y *phi-cuadrado*.

C.1. Chi-cuadrado

Uno de los estadísticos de mayor aplicación en el análisis de variables *cualitativas* es también de gran utilidad en el análisis de conglomerados. Aunque se introducen modificaciones importantes en su formulación respecto a su uso en otras técnicas analíticas. Mide la distancia entre dos variables cualitativas (X_A y X_B) como dependiente de las frecuencias totales que presentan las distintas categorías de las variables. Se consideran tanto las frecuencias observadas como las esperadas. Su cálculo es el siguiente:

$$d\chi^2_{AB} = \sqrt{\sum_{i=1}^{n} \frac{\left[X_{Ai} - E(X_{Ai})\right]^2}{E(X_{Ai})} + \sum_{i=1}^{n} \frac{\left[X_{Bi} - E(X_{Bi})\right]^2}{E(X_{Bi})}}$$

Donde: "X_A" y "X_B" representan las frecuencias observadas en las i (i = 1, 2, 3... n) categorías de las variables X_A y X_B, respectivamente.
"$E(X_{Ai})$" y "$E(X_{Bi})$" son las frecuencias esperadas de las i categorías de las variables X_A y X_B. Téngase presente que los valores esperados son los del modelo de *independencia* de las variables X_A y X_B, el habido cuando ambas variables no se hallan relacionadas.

Como en otros análisis estadísticos, en el análisis de conglomerados χ^2 puede alcanzar un valor de 0 a infinito. No existe ningún límite superior. El valor 0 expresa *similitud*. Cuanto más se aleje el valor de 0, mayor es la *distancia* existente entre las variables consideradas.

C.2. Phi-cuadrado

El coeficiente *phi-cuadrado* se define a partir del estadístico *chi-cuadrado* de la manera siguiente:

$$d\phi^2 = \frac{\chi^2}{N} \text{ siendo } d\phi = \sqrt{\frac{\chi^2}{N}}$$

Donde: "N" expresa el tamaño muestral (o número de casos analizados).

Este coeficiente permite graduar mejor la distancia entre las variables, al estar el rango de valores posibles comprendido en el intervalo de 0 a 1. El valor 0 expresa *similitud* plena, mientras que el valor 1, la plena *disparidad* o *disimilaridad* entre las variables medidas.

D) Variables en diferentes niveles de medición

En el análisis simultáneo de variables en distintos niveles de medición cabe la opción de proceder a su transformación a un mismo nivel de medición. Por ejemplo, operar con variables *binarias* (con la pérdida consiguiente de información, ya aludida) y escoger algunas de las medidas de *similaridad* adecuadas a este tipo de variables. Pero también cabe la opción de analizarlas, de forma conjunta, en sus distintos niveles de medición originales. Para dicho propósito puede aplicarse el *coeficiente de Gower* como medida de *similitud.*

D.1. Coeficiente de similaridad de Gower

Este coeficiente fue propuesto por Gower en 1971 (en "A general coefficient of similarity and some of its properties", *Biometrics*, 27: 857-872). Tiene la particularidad de permitir el uso simultáneo de variables en diferentes niveles de medición, en la estimación de la *similaridad*. Su definición es la siguiente:

$$S_{ij} = \frac{\sum_{K=1}^{p} S_{ijk}}{\sum_{K=1}^{p} W_{ijk}}$$

Donde: "W_{ijk}" es una variable de ponderación de valor 1, si se considera válida una comparación de la variable K; y 0, si no se considera. En variables *binarias*, "W_{ijk}" es cero, cuando la variable K no se conoce para uno o ambos objetos que se comparan. "W_{ijk}" también se fija en 0 cuando se trata de conjunciones negativas.
"S_{ijk}" es una puntuación de *similaridad* basada en el resultado de la comparación de la variable K en los casos "i" y "j".

Cuando las variables son todas *binarias*, el coeficiente de *similaridad de Gower* es idéntico al coeficiente de *Jaccard* (Aldenderfer y Blashfield, 1984: 31). El rango de valores va de 0 a 1.

EJEMPLO DE ELECCIÓN DEL PROCEDIMIENTO DE CONGLOMERACIÓN: MÉTODO, ALGORITMO Y MEDIDA DE DISTANCIA O DE SIMILARIDAD

El análisis de conglomerados permite tanto la clasificación de "casos" como de "variables", en función de su similitud. Con los mismos datos de la encuesta del CIS se han realizado distintos análisis de conglomerados de "casos" y de "variables".

A) Para la *conglomeración de variables*, se han aplicado distintos métodos de conglomeración *jerárquica*, variando el *algoritmo* de clasificación e incluso las variables a clasificar, de acuerdo con su "relevancia". El modelo de conglomeración finalmente elegido fue el obtenido aplicando el *algoritmo* de clasificación de "distancias mínimas" (o "vinculación simple") y la *correlación de Pearson* como medida de *similaridad*.

El método de *distancias mínimas* (se agrupan las variables más próximas) se eligió por ser, de los métodos *jerárquicos*, el que muestra más adecuación a tamaños muestrales elevados. Recuérdese que el tamaño de la muestra total es 2.493 casos. Si bien, se decide realizar los análisis de conglomeración sólo con la mitad de la muestra, seleccionada de forma aleatoria, para posteriormente "validar" los resultados con la otra mitad. Pese a esta reducción considerable en el tamaño de la muestra y excluyendo, además, los "casos sin respuesta" en algunas de las variables incluidas en el análisis, la *muestra de análisis* se halla integrada por 809 casos, cantidad que supera bastante a la comúnmente referida como cantidad máxima "ideal" para la realización de un análisis de conglomerados *jerárquico* (200 unidades muestrales). Aunque hay que precisar que el efecto del tamaño muestral en la conglomeración jerárquica de "variables" es inferior al obtenido en la clasificación de "casos". Las dimensiones de la *matriz de distancias*, de las tablas del *historial de conglomeración* y de los *conglomerados de pertenencia*, al igual que los gráficos (de *témpanos* y el *dendograma*) se ajusta al número de variables a clasificar y no al de casos analizados, lo que facilita considerablemente su lectura e interpretación.

En este ejemplo, 14 fueron las variables inicialmente analizadas: "simpatía marroquí" (X_{15}), "leyes inmigración" (X_1), "ideología política" (X_2), "sexo" (X_3), "edad" (X_4), "número de inmigrantes" (X_6), "regularizar inmigrantes" (X_7), "entrada inmigrantes" (X_8), "partido racista" (X_9), "casar con marroquí" (X_{10}), "estudios" (X_{11}), "ingresos" (X_{12}), "vecino marroquí" (X_{13}) e "inmigrante delincuente" (X_{14}). La variable "simpatía hacia latinoamericanos" (X_5) se decide excluirla de los análisis por tener una *colinealidad* elevada con la variable X_{15} ("simpatía marroquí"). Se quiere evitar el efecto distorsionador de la inclusión de variables muy *colineales* en el proceso de conglomeración, no tanto de "variables", sino de "casos". En la conglomeración de "casos", el incluir varias variables que midan una misma dimensión aumenta la probabilidad de que éstas ejerzan un mayor efecto en el proceso de agrupación.

Una segunda razón principal que fundamenta la elección del *algoritmo* de *distancias mínimas* es su mayor adecuación a variables "ordinales". Recuérdese que varias de las variables

que participan del análisis son *ordinales*. Además, este *algoritmo* de clasificación se ve menos afectado por las transformaciones que se realicen en la matriz de datos.

Pese a estas ventajas esenciales, el inconveniente principal del *algoritmo de distancias mínimas* es su tendencia a "encadenar" conglomerados aparentemente distintos, como se constató en su aplicación a la clasificación de "casos".

Mediante la *conglomeración de variables* se quiere identificar variables "relacionadas". El objetivo básico es analizar qué características comparten las variables para, a partir de ellas, observar las diferencias entre los sujetos en sus actitudes ante la inmigración.

Para eliminar la influencia negativa de analizar variables en distintas escalas de medición, se utiliza como medida de "distancia" (en este caso de "similitud") la *correlación de Pearson*, que mide la semejanza o *similitud* entre dos variables. Ésta es la medida de distancia más utilizada en la conglomeración de variables y se calcula para todos los pares de variables.

Como se quiere medir la "fuerza" de las relaciones entre las variables, se utiliza el *valor absoluto* de cada *correlación*, y no el "signo" del coeficiente (que sólo indica la "dirección" de la relación entre las variables). A diferencia del *análisis factorial* (capítulo 5), que permite que las variables estén positiva o negativamente relacionadas con un factor, en el *análisis de conglomerados* el "signo" del coeficiente de *correlación de Pearson* afecta a la creación de conglomerados: las variables que correlacionan negativamente con un conglomerado no aparecen en el mismo conglomerado que las variables que correlacionan positivamente. Ésta es la razón por la que se recomienda emplear sólo los valores absolutos de los coeficientes de *correlación*, salvo que se quiera conglomerados compuestos por variables positivamente correlacionadas.

El empleo de la *correlación de Pearson* exige que las variables sean *continuas*. Ello necesariamente lleva a la transformación de las variables *nominales* en *ficticias*, como se expone en el capítulo 1. De los distintos análisis realizados, finalmente se escoge el obtenido cuando se excluye la variable "ideología política" (X_2) del análisis. En este último modelo la agrupación de variables resulta más "lógica" y coincide con la obtenida mediante el análisis factorial de *componentes principales* (capítulo 5). Como se verá en el subapartado 3.5.1, las 13 variables a clasificar se agrupan en 4 conglomerados:

- El primer conglomerado agrupa variables que expresan "simpatía" hacia los inmigrantes, preferentemente marroquíes: "simpatía marroquí", "partido racista", "casar con marroquí" y "vecino marroquí".
- El segundo conglomerado reúne variables principalmente relacionadas con la política inmigratoria: "leyes inmigración", "número de inmigrantes", "regularizar inmigrantes", "entrada inmigrantes" e "inmigrante delincuente".
- El tercer conglomerado sólo incluye la variable "sexo".
- El cuarto conglomerado agrupa variables sociodemográficas comúnmente vinculadas con la "posición social": "ingresos", "estudios" y "edad".

A esta agrupación de variables se llega a partir de la siguiente *matriz de distancias* (tabla A), aplicando la *correlación de Pearson* como *medida de similaridad*:

De acuerdo con esta matriz, las variables más "próximas" son las que presentan *correlaciones* elevadas (próximas a 1,0). Las correlaciones mayores se dan entre variables agrupadas en el primer conglomerado. A decir, las variables X_{10} ("casar con marroquí") y X_{13} ("vecino marroquí"): r = ,585; y X_{10} con X_{15} ("simpatía marroquí"): r = ,510. Correlaciones pró-

ximas a 0,0 expresan, por el contrario, "disimilaridad" entre las variables. En el subapartado 3.5.1 figuran las tablas del *historial de conglomeración* y de *conglomerado de pertenencia* de este modelo de conglomeración; en el subapartado 3.5.2, los gráficos (*dendograma* y de *témpanos*).

Tabla A. Matriz de distancias

Caso	Archivo matricial de entrada												
	simpatía marroquí	*leyes inmigr.*	*sexo*	*edad*	*n.º inmigr.*	*regular. inmigr.*	*entrada inmigr.*	*partido racista*	*casar con marroquí*	*estudios*	*ingresos*	*vecino marroquí*	*inmigr. delinc.*
Simpatía marroquí		,296	,048	,174	,316	,276	,284	,275	,510	,213	,108	,395	,304
Leyes inmigración	,296		,014	,138	,351	,263	,381	,100	,216	,230	,189	,210	,281
Sexo	,048	,014		,003	,000	,041	,026	,007	,016	,001	,065	,014	,088
Edad	,174	,138	,003		,169	,081	,118	,154	,208	,433	,197	,142	,188
Núm. inmigrantes	,316	,351	,000	,169		,254	,296	,222	,212	,320	,230	,208	,345
Regularizar inmig.	,276	,263	,041	,081	,254		,377	,248	,286	,198	,126	,280	,222
Entrada inmigran.	,284	,381	,026	,118	,296	,377		,309	,317	,174	,170	,261	,331
Partido racista	,275	,100	,007	,154	,222	,248	,309		,322	,110	,051	,353	,182
Casar con marroq.	,510	,216	,016	,208	,212	,286	,317	,322		,162	,068	,585	,241
Estudios	,213	,230	,001	,433	,320	,198	,174	,110	,162		,436	,127	,172
Ingresos	,108	,189	,065	,197	,230	,126	,170	,051	,068	,436		,046	,087
Vecino marroquí	,395	,210	,014	,142	,208	,280	,261	,353	,585	,127	,046		,206
Inmigrante delin.	,304	,281	,088	,188	,345	,222	,331	,182	,241	,172	,087	,206	

Aunque la consideración de la variable "ideología política" (X_2) también ocasionaba la agrupación de las variables en cuatro conglomerados, este segundo modelo de conglomeración se descartó por ser más "ilógica" la composición de los conglomerados, aun teniendo significado "estadístico". El primer conglomerado agrupaba variables relacionadas tanto con "política inmigratoria" como con "simpatía" hacia los inmigrantes: "simpatía marroquí", "leyes inmigración", "n.° inmigrantes", "regularizar inmigrantes", "entrada inmigrantes", "partido racista", "casar con marroquí", "vecino marroquí" e "inmigrante delincuente". El segundo conglomerado sólo la variable "ideología política". El tercer conglomerado sólo la variable "sexo". Y, el cuarto conglomerado las variables "edad", "estudios" e "ingresos". Éste es un ejemplo de cómo la elección del modelo de conglomerados no sólo ha de fundamentarse en criterios "estadísticos", sino también "lógico-sustantivos". La solución final ha de tener, sobre todo, interpretación "lógica".

La *matriz de distancias*, al igual que el *historial de conglomeración*, la tabla de *pertenencia grupal* y los gráficos también difieren, al haberse añadido una variable más al modelo de clasificación. Como muestra, compárese la *matriz de distancias* correspondiente al modelo que incluye la variable "ideología política" (tabla B) con el que la excluye (tabla A).

B) Respecto a la *conglomeración de casos*, el elevado tamaño muestral, pese a su seccionamiento aleatorio a la mitad, lleva a la aplicación preferente de un método de conglomeración *no jerárquica*, como "K-means" (o *K-medias*), que es el más popular y se incluye en las últimas versiones del SPSS.

La elección del algoritmo de clasificación *K-medias* se decide no sólo porque está incluido en la mayoría de los programas estadísticos y su adecuación a tamaños muestrales elevados (lo habitual en datos de encuesta), sino también porque:

Tabla B. Matriz de distancias

Caso	Archivo matricial de entrada													
	simpatía marroquí	leyes inmigr.	ideología política	sexo	edad	n.º inmigr.	regular. inmigr.	entrada inmigr.	partido racista	casar con marroquí	estudios	ingresos	vecino marroquí	inmigr. delinc.
Simpatía marroquí		,306	,120	,066	,179	,333	,299	,292	,273	,510	,246	,121	,388	,327
Leyes inmigración	,306		,215	,014	,156	,369	,272	,392	,121	,237	,259	,212	,220	,292
Ideología política	,120	,215		,014	,145	,135	,109	,135	,202	,137	,089	,015	,134	,152
Sexo	,066	,014	,014		,003	,023	,053	,024	,003	,027	,017	,052	,009	,103
Edad	,179	,156	,145	,003		,164	,069	,119	,156	,196	,410	,178	,150	,188
Núm. inmigrantes	,333	,369	,135	,023	,164		,262	,306	,218	,224	,320	,227	,228	,345
Regularizar inmig.	,299	,272	,109	,053	,069	,262		,371	,268	,293	,211	,117	,278	,214
Entrada inmigran.	,292	,392	,135	,024	,119	,306	,371		,335	,326	,174	,156	,273	,318
Partido racista	,273	,121	,202	,003	,156	,218	,268	,335		,323	,114	,056	,358	,203
Casar con marroq.	,510	,237	,137	,027	,196	,224	,293	,326	,323		,166	,064	,599	,241
Estudios	,246	,259	,089	,017	,410	,320	,211	,174	,114	,166		,439	,135	,170
Ingresos	,121	,212	,015	,052	,178	,227	,117	,156	,056	,064	,439		,044	,060
Vecino marroquí	,388	,220	,134	,009	,150	,228	,278	,273	,358	,599	,135	,044		,212
Inmigrante delin.	,327	,292	,152	,103	,188	,345	,214	,318	,203	,241	,170	,060	,212	

- Comprueba la "relevancia" de las variables en la diferenciación de los grupos (mediante la prueba de significación *F*).
- Ayuda a la identificación de "atípicos", al proporcionar la distancia de cada caso al centro del conglomerado en el que se incluye.

Pero, a diferencia de los métodos *jerárquicos*, precisa de la especificación previa del número de conglomerados a formar con los datos. Esta información puede obtenerse de análisis precedentes (como la realización previa de un análisis de conglomerados *no jerárquico*, aunque sólo sea con una parte de la muestra, que señale un número de conglomerados "razonable" a constituir) o a partir de razonamientos teóricos. Si no se dispone de ninguna información al respecto, siempre cabe la opción de probar varias soluciones de conglomeración *K-medias,* variando el número de conglomerados a formar. En nuestro ejemplo, la solución de 3 conglomerados se presenta como más "lógica" desde todas las vertientes. Los análisis de conglomeración de casos *jerárquicos* realizados previamente indicaban como clasificación más "idónea" aquella en la que los casos se clasifican en 3 conglomerados diferentes de acuerdo con su actitud "declarada" ante la inmigración. Posteriormente, el análisis *discriminante* "valida" el modelo de clasificación obtenido mediante el procedimiento *K-medias*, como se verá en el capítulo 4.

Los resultados del análisis de conglomerados *K-medias* se exponen y comentan en el subapartado 3.5, dedicado a la "presentación" e "interpretación" de los resultados del análisis de conglomerados. A continuación se informa de los diversos análisis de conglomerados *jerárquicos aglomerativos* de "casos", realizados para alcanzar un mismo objetivo básico: diferenciar distintos grupos homogéneos de personas de acuerdo con sus actitudes manifiestas ante la inmigración. De manera que, una vez conocidas sus características, pueda preverse qué actitud tendrán ante la inmigración.

Como las variables consideradas se hallan en distinta métrica, se prueban las dos opciones de tratamiento conjunto posibles: su tratamiento como variables "continuas" y como variables "ficticias".

1. El tratamiento como variable *continua* exige la necesaria transformación de las variables *nominales* en *ficticias*. Una vez hecha dicha conversión, se procede a la elección del *algoritmo* de clasificación. Los tres elegidos fueron los siguientes:

   - *Distancias mínimas* (o *vinculación simple*), por las razones anteriormente expuestas.
   - *Promedio entre grupos*, por ser la opción aplicada por defecto en el programa SPSS, y porque considera (en el cálculo de la *distancia*) todos los objetos y no sólo los extremos. Aunque presenta el gran inconveniente de crear conglomerados con aproximadamente la misma varianza.
   - *Método Ward* porque vincula objetos a conglomerados que provocan un menor incremento de la varianza intragrupal. Minimiza la varianza intragrupal evitando, al mismo tiempo, el "encadenamiento" de conglomerados habitual en los métodos de *vinculación* (en especial, en el *simple*). Su principal inconveniente es la creación de conglomerados de tamaño similar y de forma hiperesférica, así como la combinación de conglomerados con un número reducido de observaciones.

   En los tres *algoritmos* de clasificación se aplica la misma medida de *distancia*: la *distancia euclídea al cuadrado*. Ésta es la medida de *distancia* aplicada por defecto en el *algoritmo de Ward* y la más utilizada en variables *continuas*, cuando se agrupan "casos". Su uso correcto exige que las variables estén *estandarizadas* para evitar que las variables de mayor variabilidad contribuyan más al cálculo de la distancia entre los casos. Si bien, la *estandarización* presenta el gran inconveniente de minimizar las diferencias grupales.

   La *estandarización* se recomienda cuando el rango de alguna de las variables a analizar supera bastante al de otras variables que también participan del análisis. En este caso se optó por el procedimiento de *estandarización* habitual, que consiste en la transformación de las variables a "puntuaciones Z", de manera que su media sea 0 y su desviación típica 1. El rango de dos variables concretas, las variables "edad" e "ingresos" (en especial, la segunda), supera bastante al de las otras variables incorporadas al análisis. Lo que necesariamente lleva a la *estandarización* para que todas las variables se encuentren expresadas en una escala comparable. Además, la aplicación posterior del *algoritmo K-medias* también obliga a la *estandarización* de las variables.

   Como era de prever, las soluciones difieren dependiendo del *algoritmo* de clasificación que se aplique. El de *vinculación simple* (o *distancia mínima*) cumple el inconveniente principal generalmente atribuido a este *algoritmo* de clasificación: su tendencia a "encadenar" conglomerados aparentemente distintos. A excepción de 3 casos –el caso 129 (conglomerado 2), 282 (conglomerado 3) y el caso 352 (conglomerado 4)–, todos los casos analizados son clasificados en el conglomerado 1. Este hecho lleva a descartar dicha solución de conglomeración, por no satisfacer el objetivo principal del análisis.

   Por su parte, el *algoritmo de promedio entre grupos* engloba algo más de 3 de cada 4 casos en un solo conglomerado: el conglomerado 2. Por esta razón, se descarta este segundo modelo de conglomeración por no adecuarse tampoco al objetivo del análisis.

   La solución finalmente elegida es la obtenida mediante el método *Ward*, aplicando asimismo la distancia *euclídea al cuadrado*. De acuerdo con esta medida de distancia, coeficientes pequeños (próximos a "0") indican la presencia de conglomerados claramente

homogéneos (a diferencia de la *correlación de Pearson*), mientras que coeficientes elevados informan que los casos o conglomerados que se agrupan son heterogéneos. Cuanto mayor sea su valor, peor. Esta medida de distancia carece de límite superior, a diferencia de la *correlación de Pearson* y otras medidas de distancia (subapartado 3.3.4).

Mediante este tercer *algoritmo* de clasificación se obtienen conglomerados de similar tamaño y con menor varianza intragrupal, lo cual lleva a su elección como solución de conglomeración *jerárquica*. En el subapartado 3.5.1 se incluye un extracto de las tablas del *historial de conglomeración* y del *conglomerado de pertenencia* de este modelo de clasificación de casos.

2. Por último, se prueba el tratamiento conjunto *binario*, aun sabiendo que presenta como inconveniente principal la pérdida de información consecuente con la traducción de las variables de *intervalo*, e inclusive las variables *ordinales*, a *binarias* (con los códigos 1 y 0). Para dicha transformación se ha seguido el procedimiento usual descrito en el subapartado 3.3.4. Para la variable *continua* "edad", por ejemplo, se ha seguido el procedimiento usual de buscar un valor central (en este caso la edad de 45 años), de manera que todos los casos situados por debajo de este valor de referencia (≤ 45 años) se codifican 0, mientras que los situados por encima (> 45 años) se codifican 1.

   Se prueban, asimismo, diferentes *algoritmos* de clasificación, no sólo porque ello permite la elección de la solución más "significativa" analítica y conceptualmente, sino también porque se quiere validar los resultados anteriores.

   Dos han sido los *algoritmos* de clasificación utilizados: *distancias mínimas* y *distancias máximas* (o *vinculación completa*). Este último difiere del anterior en que calcula la distancia entre dos conglomerados como la distancia entre sus dos casos más distantes. Normalmente crea conglomerados más compactos que la *vinculación simple*.

   El número de conglomerados posibles se fija, igualmente, entre 2 y 4. También se prueban distintas soluciones de conglomerados variando además las medidas de distancia. Tres fueron las medidas de *distancia* elegidas: *correlación punto 4 phi*, distancia *euclídea binaria* y la distancia *Jaccard*, que es una de las más aplicadas en datos binarios. En todas ellas el valor 0 expresa inexistencia de *similitud* entre los casos, mientras el valor 1 la plena correspondencia o *similitud* entre ellos, a excepción de la *distancia euclídea binaria* que carece de límite superior. El modelo aplicando la medida de distancia *euclídea binaria* se descartó por diferir bastante de las anteriores clasificaciones. En cambio, los obtenidos aplicando las medidas de *correlación punto 4* y la distancia *Jaccard* guardaban bastante similitud. No obstante, se prima la solución de conglomeración obtenida aplicando el *algoritmo de Ward* y la distancia *euclídea cuadrada*, por suponer menor pérdida de información y proporcionar una clasificación más "lógica" de los casos. A ella se hará referencia en páginas posteriores, cuando se interpreten los resultados de los análisis realizados.

3.4. La obtención de conglomerados

A la elección de la medida de *similaridad* o de *distancia* le sigue la obtención de la solución de conglomerados, en conformidad con las diversas decisiones adoptadas. A decir, el *método* de conglomeración, el *algoritmo* de clasificación y la medida de *similaridad* o *distancia*. Pero antes de proceder a la interpretación de los resultados, hay

que dirimir una cuestión crucial: la referente al número de conglomerados a retener, entre las distintas alternativas posibles de clasificación de los objetos de interés.

Como ya se dijo en la presentación de los diferentes métodos de conglomeración, la decisión sobre el número de conglomerados a retener es previa a la ejecución de cualquier análisis de conglomerados *no jerárquico*. En la conglomeración *jerárquica*, por el contrario, es ésta una cuestión a debatir en las postrimerías del análisis, una vez que éstos han concluido. De ahí su inclusión en este apartado posterior a la exposición de decisiones clave previo al análisis de conglomerados.

3.4.1. Elección del número de conglomerados

La finalidad de todo análisis de conglomerados es la clasificación de una serie de objetos en conglomerados (o grupos) homogéneos. Pero, ¿cuántos conglomerados se requieren para describir, de forma precisa, la similitud y la diversidad en una población?

Para resolver esta cuestión trascendental no existe ninguna respuesta que sea comúnmente aceptada. Sin embargo, existen distintos procedimientos alternativos que, como en el análisis *factorial exploratorio*, se aplican para determinar el número de conglomerados idóneo. De éstos destacan los siguientes:

a) Seguir algún *criterio teórico* que fundamente la elección de un número de conglomerados específico, aunque hay que tener presente que la clasificación propuesta *a priori* no siempre es coincidente con la sugerida tras la realización de los análisis. Por esta razón, se recomienda no ceñirse a un número determinado de conglomerados, sino probar diferentes soluciones de clasificación con números de conglomerados varios. Y, después, elegir aquella solución que tenga un mayor significado teórico y estadístico. Este proceder es muy habitual en la conglomeración *no jerárquica*, como se dijo en el subapartado 3.3.2.

b) En la conglomeración *jerárquica* además se puede aplicar criterios similares a los utilizados en el análisis *factorial exploratorio*, en la decisión del número de factores a retener. Si en el análisis *factorial* (capítulo 5) los *autovalores* son los protagonistas, en el análisis de *conglomerados jerárquico* lo son los *coeficientes de conglomeración*. Llámense "coeficientes de fusión o amalgamiento" (Aldenderfer y Blashfield (1984) o "coeficientes de aglomeración" (Hair *et al.,* 1992; 1999), indican el valor numérico (medida de *distancia* o *similitud*) que propicia la unión de objetos (casos o variables) para formar conglomerados.

Se trata de observar grandes "variaciones" en los valores de los coeficientes, como indicativo del número de conglomerados a retener. Cuando se aplican medidas de *distancia* (como la distancia *euclídea al cuadrado*), las variaciones han de ser un fuerte "aumento" en la cuantía de los coeficientes, al pasar de un número concreto de conglomerados al inmediato superior (de 3 a 4 conglomerados, por ejemplo). En cambio, si se utilizan medidas de *similitud* (como, por ejemplo, la *correlación de Pearson*), las variaciones han de ser una fuerte "disminución" en la magnitud del coeficiente. La unión de dos conglomerados

muy diferentes coincide con coeficientes elevados (medida de *distancia*) o bajos (medida de *similitud*), concretamente, con un incremento porcentual considerable en el coeficiente respecto al nivel siguiente.

En consecuencia, la solución "idónea" del número de conglomerados es, de acuerdo con este segundo criterio de decisión, la correspondiente al número de conglomerados previo al "salto" (o variación) apreciable en el valor del coeficiente de conglomeración. Ello se debe a que los "saltos" acontecen cuando dos conglomerados con relativa disimilaridad (o heterogeneidad) se unen. La dificultad está en decidir cuál de los "saltos" se considera relevante como indicador de que se ha alcanzado el número correcto de conglomerados, de manera especial, cuando se observan varios "saltos". Esto puede introducir nuevamente "subjetividad" en la decisión del número de conglomerados a formar. Crítica que acompaña, en general, a los métodos de conglomeración *jerárquica*, como se expuso en el subapartado 3.3.2.

EJEMPLO DE USO DE LOS COEFICIENTES DE CONGLOMERACIÓN EN LA ELECCIÓN DEL NÚMERO DE CONGLOMERADOS

Para ilustrar la aplicación de los coeficientes de *conglomeración* en la decisión del número de conglomerados a formar, se toman los coeficientes obtenidos mediante el *algoritmo de Ward* y la medida de distancia *euclídea al cuadrado* correspondiente al análisis de conglomerados *jerárquico aglomerativo* que fue finalmente aceptado como el más adecuado para la clasificación de los "casos", de acuerdo con su actitud ante la inmigración. Los valores que figuran a continuación (tabla A) se han extractado de la tabla del *historial de conglomeración* respectiva. Corresponden a las 10 últimas etapas del proceso de conglomeración, cuando el número de conglomerados se reduce hasta uno en la última etapa.

A la vista de las variaciones en el valor de los coeficientes puede concluirse que la solución de 3 conglomerados puede ser la correcta, al haberse observado sólo un "salto" apreciable en el valor del coeficiente. El reducir el número de conglomerados a 2 supone el mayor aumento en la magnitud del coeficiente (19,3%), que pasa de ser 2.743,817 a 2.978,380. Dicho incremento significa que se unen dos conglomerados muy diferentes, si se pasa de 3 a 2 conglomerados. Por esta razón, se escoge la solución de 3 conglomerados como la más "idónea". Se puede proceder a la clasificación de los encuestados en 3 grupos diferentes, de acuerdo con sus actitudes manifiestas ante la inmigración.

Aunque se descartó la solución de conglomeración obtenida mediante el *algoritmo promedio inter-grupos* por producir, al igual que la *vinculación simple,* conglomerados "encadenados", la tabla B extracta los coeficientes de conglomeración correspondientes a las diez últimas etapas para compararlos con los incluidos en la tabla A. Estos coeficientes se han obtenido, igualmente, aplicando la medida de distancia *euclídea al cuadrado.* Como puede apreciarse en la comparación de las variaciones porcentuales, la solución de la clasificación de los encuestados en 3 grupos, en consonancia con su actitud ante la inmigración, muestra ser, igualmente, la más idónea.

Tabla A

Número de conglomerados	*Coeficiente de conglomeración*	*Diferencia de coeficientes*	*Cambio porcentual en el coeficiente del nivel siguiente*
10	1.975,104	66,810	3,4
9	2.041,914	68,644	3,4
8	2.110,558	80,191	3,8
7	2.190,749	93,524	4,3
6	2.284,273	109,087	4,8
5	2.393,360	169,517	7,1
4	2.562,877	180,940	7,1
3	2.743,817	234,563	8,5
2	2.978,380	573,620	19,3
1	3.552,000	–	–

Tabla B

Número de conglomerados	*Coeficiente de conglomeración*	*Diferencia de coeficientes*	*Cambio porcentual en el coeficiente del nivel siguiente*
10	26,152	,003	,01
9	26,155	1,250	4,8
8	27,405	2,490	9,1
7	29,895	1,390	4,6
6	31,285	3,097	9,9
5	34,382	1,319	3,8
4	35,701	,263	,7
3	35,964	5,350	14,9
2	41,314	8,499	20,6
1	49,813	–	–

En la conglomeración de "variables", al emplearse una medida de *similitud* (la *correlación de Pearson*), las variaciones en los coeficientes han de ser contrarias a las producidas cuando se aplican medidas de *distancia*: coeficientes de correlación "elevados" expresan agrupación de variables "similares" o muy próximas, mientras que coeficientes próximos a cero significan que se vinculan variables muy heterogéneas.

En el subapartado 3.5.1 figura la tabla del *historial de conglomeración* correspondiente al modelo de conglomerados de variables, aplicando el *algoritmo* de *distancias mínimas* y la *correlación de Pearson*, como medida de *similitud* de las variables. Extráctese de dicha tabla los coeficientes de conglomeración e indíquese cuántos conglomerados deberían seleccionarse como solución más "idónea" en la clasificación de las variables. Téngase presente que la lectura de dichos coeficientes y sus variaciones porcentuales ha de hacerse de forma inversa a la anteriormente expuesta, al expresar "similitud" y no "distancia". Aquí se avanza que el "salto" más apreciable se produce cuando se pasa de 2 (coeficiente de correlación igual a ,317) a 1 conglomerado (siendo su coeficiente igual a ,088), lo que lleva a conceder más protagonismo a criterios lógico-sustantivos en la decisión del número de conglomerados a constituir. Como se verá en la tabla de *conglo-*

merado de pertenencia (subapartado 3.5.1), la agrupación de las 13 variables en 4 conglomerados es la más "lógica", considerando el "significado" de las variables.

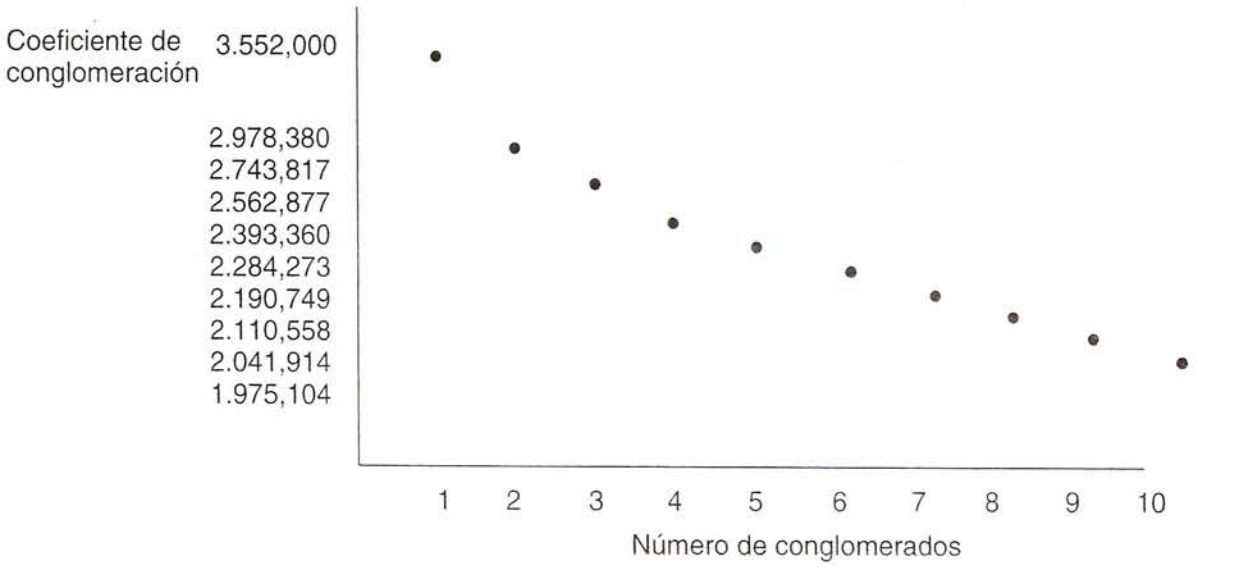
---

c) Como en el análisis *factorial exploratorio,* en el análisis de *conglomerados* se puede elaborar un *gráfico de sedimentación* que ayude a la decisión del número de conglomerados a retener. En él figuran los coeficientes de conglomeración con el número de conglomerado correspondiente. La interpretación del gráfico es la misma que en el análisis *factorial* (subapartado 5.5.2): un marcado "allanamiento" de la pendiente se toma como indicativo de número de conglomerados "idóneo". Éste viene determinado por el punto de inflexión de la trayectoria de caída de la pendiente del gráfico: cuando la pendiente descendente comienza a nivelarse. Adviértase que la disposición de los coeficientes obtenidos aplicando medidas de *similitud* ha de ser opuesta a la descrita, si se quiere que los puntos formen una pendiente descendente. En caso contrario, la pendiente será ascendente, aunque se trata igualmente de observar un punto de inflexión en la trayectoria de la pendiente.

Este criterio de selección del número de conglomerados comparte las mismas críticas pronunciadas contra el *gráfico de sedimentación* en el análisis *factorial*. Principalmente, la arbitrariedad y el subjetivismo que introduce en la decisión del número de conglomerados a retener.

EJEMPLO DE GRÁFICO DE SEDIMENTACIÓN

El siguiente gráfico ha sido elaborado a partir de los coeficientes que figuran en la tabla A del ejemplo anterior. Aunque el programa SPSS no lo ofrece como opción gráfica, a diferencia del análisis factorial que sí lo incluye, he decidido realizarlo manualmente para ilustrar su uso como criterio alternativo en la selección del número de conglomerados.



En él puede observarse cómo la pendiente decreciente comienza a "allanarse" a partir de 3 conglomerados, y más a partir de 5. Nuevamente se apunta a la solución de 3 conglomerados como una clasificación "idónea" de los casos analizados.

---

d) En programas específicos del análisis de conglomerados, como el CLUSTAN, se ofrece además un procedimiento alternativo de selección de conglomerados propuesto por Wishart en 1982 (en *Supplement, CLUSTAN user manual*, 3.º edition: Program Library Unit, Edinburgh University) y que Aldenderfer y Blashfield (1984) consideran "óptimo". Parte del procedimiento llamado "stopping rule ≠ 1", propuesto por Mojena en 1977 (en "Hierarchical grouping methods and stopping rules: an evaluation", *Computer Journal*, 20: 359-363), que establece como partición óptima del número de conglomerados aquella que satisface la desigualdad siguiente: $Z_{j+1} > Z + Ks_Z$ (siendo "Z" el valor del *coeficiente de fusión* y "Z_{j+1}" el correspondiente a la etapa "j + 1" del proceso de conglomeración; y "Ks_Z" la desviación típica respecto al *coeficiente de fusión*). Lo que Wishart añade es la comprobación de la *significatividad* estadística de la aplicación de la regla de Mojena, mediante el estadístico "t" con "n – 2" grados de libertad (donde "n" es el número total de *coeficientes de fusión*). Pero, desgraciadamente, esta comprobación no está incluida en todos los paquetes estadísticos, lo que dificulta su aplicación.

e) A estos criterios hay que añadir la información dada por uno de los gráficos más característicos de la conglomeración *jerárquica*: el *dendograma* (que se describe en el subapartado 3.5.2).

3.5. Presentación de los resultados y su interpretación

En el análisis de conglomerados los resultados se presentan de forma gráfica y mediante tablas de resultados. En el análisis de conglomerados *jerárquico* las *tablas de resultados* comunes son el *historial de conglomeración* y el *conglomerado de pertenencia*, mientras que las tablas de *centros de conglomerados* (iniciales y finales) y *ANOVA* caracterizan a los métodos de conglomeración *no jerárquica*. Los gráficos típicos de la conglomeración *jerárquica* son el *dendograma* y el gráfico de *témpanos*. En los métodos de conglomeración *no jerárquica* destacan, en cambio, los que representan la ubicación de cada conglomerado y sus *centros* respectivos. A continuación se exponen las presentaciones numéricas y gráficas más características de cada método de conglomeración.

3.5.1. Las tablas de resultados

Como procedimientos de conglomeración diferentes, las tablas de resultados difieren cuando se aplica un método *jerárquico* o uno *no jerárquico*. Para facilitar su com-

prensión, se ha decidido su presentación aparte, diferenciando las tablas de resultado según el método de conglomeración aplicado.

3.5.1.1. Métodos de conglomeración jerárquicos

A) Historial de conglomeración

El *historial de conglomeración* es una tabla de resultados básica en la conglomeración *jerárquica*, dirigida a la descripción del proceso de convergencia de los objetos (casos o variables) y a la selección del número de conglomerados idóneo para su clasificación. Adopta la forma de una tabla o cuadro que resume el proceso de constitución de los conglomerados e incluye la información siguiente:

- En la primera columna ("etapa") figuran numeradas las distintas etapas del análisis, desde la primera hasta la última. En general, habrá tantas etapas como casos menos uno (si se clasifican casos) o variables menos uno (cuando la conglomeración es de variables). Por esta razón, su exposición e interpretación sólo es viable cuando los objetos (casos o variables) a clasificar son de tamaño reducido. La recomendación usual es que no superen las 200 unidades. Por encima de este referente, las descripciones del análisis, tanto numéricas como gráficas, adquieren elevadas dimensiones que dificultan su lectura e interpretación.
- A continuación figuran dos columnas que indican los dos objetos o, en su caso, conglomerados que se combinan en cada etapa.
- La columna con el cabecero de "coeficiente" informa del valor de la medida de *distancia* o *similitud* escogida en la clasificación de los objetos. Mediante este valor se cuantifica la homogeneidad o heterogeneidad de los conglomerados que se combinan en cada fase del análisis. Cuando se aplican medidas de *distancia,* valores pequeños (próximos a "0") expresan que los objetos (o conglomerados) que se combinan son bastante homogéneos. En cambio, coeficientes con valores elevados informan de lo contrario: de la agrupación de objetos o conglomerados bastante disimilares.

  Cuando se utilizan medidas de *similitud*, la interpretación de los coeficientes es contraria a la anterior: valores elevados expresan "homogeneidad", mientras que valores bajos indican "heterogeneidad" de los conglomerados.

  El valor de los coeficientes depende no sólo de la medida de *distancia* o *similitud* empleada, también afecta el *algoritmo* de clasificación elegido, como se demostró en el ejemplo del subapartado 3.4.1. En dicho ejemplo se expuso el uso de los coeficientes como guía principal en la decisión del número de conglomerados a formar para representar adecuadamente los datos.
- Después se añaden dos columnas con información referente al "paso anterior" en el que cada uno de los dos conglomerados, combinados en dicho paso, aparecen por primera vez. A ello se debe que estas dos columnas aparezcan ba-

jo el cabecero genérico de "etapa en la que el conglomerado aparece por primera vez". Una observación que nunca se ha unido antes a un conglomerado tendrá un valor 0 en estas columnas. Esta información es de utilidad en la identificación de observaciones "únicas" que se unen tarde al proceso de conglomeración (en los últimos pasos del análisis). Estas observaciones "únicas" pueden ser, a su vez, potenciales *atípicos* (o "outliers"), como se verá en el subapartado 3.5.3.

- La última columna ("próxima etapa") señala la etapa siguiente en la que el conglomerado que acaba de formarse se agrupará con otro conglomerado u objeto. Recuérdese que en la conglomeración *jerárquica* (*aglomerativa*) paulatinamente se va de más a menos conglomerados, mediante la integración de nuevos objetos o la fusión de conglomerados ya existentes.

EJEMPLO DE HISTORIAL DE CONGLOMERACIÓN

La tabla A corresponde al *historial de conglomeración* de la clasificación de variables realizada. Aunque son 809 los casos finalmente analizados, la dimensionalidad de la tabla es pequeña porque hace referencia a "variables" y no a "casos". Como son 13 las variables a clasificar, el número de etapas se reduce a 12. Recuérdese que siempre es el número total de objetos que quiere agruparse menos 1. La pequeña dimensión de la tabla facilita bastante su lectura e interpretación. Por el contrario, en la clasificación de "casos" (tabla B) la tabla adquiere una dimensión tan desorbitada que, aun restringiendo los análisis a la mitad de la *muestra de análisis* (297 casos válidos), ocupa siete páginas de la salida de ordenador, al incluir 296 "etapas". Por esta razón, la tabla B sólo es un extracto de la salida original que incluye sólo los datos correspondientes a las 10 primeras etapas y las 10 últimas, para que puedan compararse las *historias de conglomeración* correspondientes a "casos" y a "variables".

Los números que figuran en la tabla A corresponden al número de la variable que es asignado por el programa, en consonancia con su disposición en el *archivo matricial de entrada* (la matriz de distancia). Si se observa la *matriz de distancia* (ejemplo del subapartado 3.3.4), puede constatarse que el número 9 designa a la variable "casar con marroquí" (X_{10}) y el número 12 a la variable "vecino marroquí" (X_{13}). Ambas variables hacen referencia explícita a un colectivo concreto de inmigrantes: los "marroquíes". De las 13 variables analizadas, éstas son precisamente las dos más relacionadas, al ser su coeficiente de *correlación de Pearson* el más elevado de los incluidos en la *matriz de distancias*: r = ,585 (subapartado 3.3.4).

Como éstas son las dos variables primeramente agrupadas, en las columnas quinta y sexta, que informan de la "etapa en la que el conglomerado aparece por primera vez", figuran ceros. La "próxima etapa" en la que se unirá una nueva variable a este primer conglomerado constituido es la etapa 2, a decir por la columna séptima en la etapa 1. La variable en cuestión es la que figura en primer lugar (1) en el *archivo matricial de entrada*: la variable "simpatía marroquí" (X_{15}), la tercera y última variable que hace referencia expresa al colectivo específico de inmigrantes marroquíes. En la *matriz de distancias* puede igualmente observarse que la correlación entre X_{15} y X_{10} es igual a r = ,510. Este valor es, asimismo, el que aparece en la columna "coeficientes" en la etapa 2.

Tabla A. Clasificación de variables

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglom. 1	Conglom. 2		Conglom. 1	Conglom. 2	
1	9	12	,585	0	0	2
2	1	9	,510	0	1	7
3	10	11	,436	0	0	4
4	4	10	,433	0	3	10
5	2	7	,381	0	0	6
6	2	6	,377	5	0	8
7	1	8	,353	2	0	11
8	2	5	,351	6	0	9
9	2	13	,345	8	0	10
10	2	4	,320	9	4	11
11	1	2	,317	7	10	12
12	1	3	,088	11	0	0

Hay que advertir que el programa SPSS utiliza el número asignado al primer objeto con el que se forma el conglomerado para nombrar al conglomerado en cuestión. Por esta razón, este primer conglomerado, integrado en las dos primeras etapas por tres variables, que indican "simpatía" hacia los marroquíes (ya sea directamente –grado de "simpatía"– o indirectamente –preocupación por un futuro "matrimonio" de un hijo con un marroquí o por tener como "vecinos" a una familia de marroquíes–) se registra inicialmente con el número 9, la primera variable que se agrupa. Pero, posteriormente, el conglomerado constituido pasa a designarse con el número 1. Éste es el número que corresponde a la primera variable ("simpatía marroquí") que se añade al primer conglomerado fruto de la unión de las variables X_{10} y X_{13}.

La siguiente etapa en la que se sumará una nueva variable a dicho conglomerado es la etapa 7. En ella se incorpora la variable número 8: "partido racista" (X_9). Obsérvese que en esta etapa (la séptima) no coincide el coeficiente ,353 con el registrado en la *matriz de distancias*. Ello se debe a que se ha sumado una cuarta variable a un conglomerado ya compuesto por tres variables. El coeficiente de conglomeración es el promedio de las distancias (o correlaciones) de todas las variables que conforman el componente.

En la fila correspondiente a la etapa 7 se informa, además, que la variable número 1 se combinó por primera vez en la etapa 2 y que la "próxima etapa" en la que se sumará una nueva variable es la etapa 11. Pero, adviértase que ahora no se trata de una variable que se suma a un conglomerado, sino de la fusión de dos conglomerados en la etapa 11: el llamado 1 (anteriormente descrito) y el designado con el número 2. Éste es el número que nombra a otro conglomerado que fue primeramente constituido en la etapa 5, tras la unión de las variables número 2 ("leyes de inmigración", X_1) y 7 ("entrada inmigrantes", X_8). La correlación entre estas dos variables (r = ,381) es inferior a la de las variables que forman el primer conglomerado. En la etapa "6" se suma una nueva variable a dicho conglomerado: la variable nú-

mero 6 ("regularizar inmigrantes", X_7). Esta variable hace también referencia a aspectos relacionados con la "política inmigratoria", lo que hace que su vinculación con las dos variables precedentes (la número 2 y 7) sea "lógica".

Obsérvese que en la etapa 6, el coeficiente que figura no es el mismo que el registrado entre las variables 2 y 6 en la *matriz de distancias*. Ahora el número 2 designa un conglomerado y no una variable concreta. Por lo que, el "coeficiente" es el promedio de las *correlaciones de Pearson* de las tres variables que componen este segundo conglomerado (que hace referencia explícita a cuestiones relacionadas con la "política inmigratoria") en la etapa 6.

En la etapa 8 se suma una nueva variable a este segundo conglomerado. Se trata de la variable número 5 ("número de inmigrantes", X_6). En la etapa 9 se añade a este mismo conglomerado una nueva variable, la número 13 ("inmigrante delincuente", X_{14}).

En cambio, la etapa 10 informa de la unión de dos conglomerados: el número 2 (anteriormente descrito y que vamos a llamar "política inmigratoria") y el número 4. Este tercer conglomerado fue inicialmente constituido en la etapa 3, tras la unión de las variables número 10 ("estudios", X_{11}) y 11 ("ingresos", X_{12}). En la etapa 4 se suma una tercera variable, la número 4 ("edad", X_4). El número de esta tercera variable (el 4) es el que pasa a designar a este tercer conglomerado.

La única variable que no queda agrupada con otras variables en un conglomerado concreto es la variable número 3 ("sexo", X_3). Esta variable forma un conglomerado aparte. Por esta razón, esta variable aparece por primera vez en la última etapa (la número 12), cuando se forma un único conglomerado tras la combinación de los anteriores, a los que se añade la variable número 3 (que constituye ella sola un conglomerado).

En las etapas 10, 11 y 12 se combinan conglomerados, mientras que las etapas precedentes describen la constitución de los conglomerados mediante la agrupación de variables "similares". En primer lugar, en la etapa 10, se combinan los conglomerados designados con el número 2 (aquí llamado "política inmigratoria") y el 4 (que incluye variables relacionadas con la "posición social": "estudios", "ingresos" y "edad"). De los 4 conglomerados se pasa a 3, tras la combinación de los conglomerados 2 y 4. En la etapa 11 se agrupa el conglomerado 1 (llamado genéricamente "simpatía hacia inmigrantes") con el 2 ("política inmigratoria"). En consecuencia, sólo quedan dos conglomerados en dicha etapa. En la etapa 12 y última sólo queda un conglomerado, al fusionarse el gran conglomerado anteriormente descrito con el conglomerado número 3 (que sólo está integrado por la variable número 3: "sexo"). En la tabla *conglomerado de pertenencia* se constata, nuevamente, la clasificación "clara" de las 13 variables en 4 conglomerados diferentes, como después se verá.

La interpretación de la tabla B es similar a la realizada respecto a la tabla A. Si bien, adviértase que los números que figuran en cada etapa designando a los conglomerados ahora corresponden a "casos" y no a "variables". Salvo esta puntualización, la interpretación del *historial de conglomeración* es igual a la anterior, aunque los coeficientes que incluye miden *distancia* y no *similitud* entre los casos. En concreto, corresponden a la medida de distancia *euclídea al cuadrado*, aplicada con el método *Ward*. Valores bajos (próximos a "0") expresan homogeneidad entre los casos o conglomerados, mientras que valores elevados indican heterogeneidad (los casos o conglomerados que se combinan son "disimilares" o heterogéneos). Aunque esta tabla sea un extracto de la original, inténtese su interpretación, a modo de la efectuada respecto a la tabla A.

Tabla B. Clasificación de casos

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglom. 1	Conglom. 2		Conglom. 1	Conglom. 2	
1	268	291	2,007E-03	0	0	53
2	60	95	3,412E-02	0	0	16
3	18	217	,106	0	0	21
4	114	248	,283	0	0	113
5	6	289	,463	0	0	112
6	46	169	,686	0	0	16
7	152	166	1,040	0	0	39
8	188	260	1,414	0	0	49
9	48	51	1,818	0	0	79
10	58	286	2,234	0	0	114
.....	.....	.....	.....	.....	.....	.....
287	2	29	1.975,104	270	278	293
288	13	19	2.041,914	273	283	291
289	11	17	2.110,558	282	271	293
290	3	8	2.190,749	285	281	291
291	3	13	2.284,273	290	288	294
292	1	4	2.393,360	280	286	295
293	2	11	2.562,877	287	289	294
294	2	3	2.743,817	293	291	296
295	1	5	2.978,380	292	284	296
296	1	2	3.552,000	295	294	0

B) Conglomerado de pertenencia

Una segunda tabla de resultados de interés en el análisis de conglomerados *jerárquico* es la llamada *conglomerado de pertenencia*. En ella se muestra la asignación de los objetos (variables o casos) a los conglomerados, una vez concluido el proceso de clasificación. Esta información se puede obtener para una única solución de conglomerados o para un rango de soluciones que el investigador determina previo a la materialización del análisis.

EJEMPLO DE CONGLOMERADO DE PERTENENCIA

En consonancia con lo reflejado en el *historial de conglomeración*, la clasificación de las "variables" en cuatro conglomerados es la siguiente, según recoge la tabla de *conglomerado de pertenencia*:

- Conglomerado 1: "simpatía marroquí", "partido racista", "casar con marroquí" y "vecino marroquí".
- Conglomerado 2: "leyes inmigración", "n.° inmigrantes", "regularizar inmigrante", "entrada inmigrante" e "inmigrante delincuente".
- Conglomerado 3: "sexo".
- Conglomerado 4: "edad", "estudios" e "ingresos".

El conglomerado 1 agrupa variables habitualmente utilizadas en la medición de la "simpatía" hacia los inmigrantes (en este caso, hacia un colectivo concreto de inmigrantes: los marroquíes). El conglomerado 2 reúne variables más relacionadas con "política inmigratoria", aunque actúan igualmente de indicadores de la actitud ante la inmigración. El conglomerado 3 incluye una única variable: "sexo". Y el conglomerado 4 a tres variables sociodemográficas ("edad", "estudios" e "ingresos") más vinculadas con la "posición social", en especial, las variables "estudios" e "ingresos".

Como en el análisis *factorial exploratorio*, al investigador incumbe asignar una etiqueta al conglomerado que identifique y sintetice, lo mejor posible, la información contenida en las variables que agrupa. Pero, a diferencia del análisis *factorial exploratorio*, en el análisis de conglomerados *jerárquico* no se cuantifica la aportación de cada variable al conglomerado y la significatividad de su contribución. La obtención de esta información lleva necesariamente a complementar estos resultados con la realización de un análisis *factorial exploratorio* con las mismas variables (capítulo 5). Lo cual además ayudará a comprobar la *validez* de los resultados.

Por último, adviértase que la tabla A informa de la composición de los conglomerados para tres soluciones de conglomerados diferentes: de 2, 3 y de 4. De la comparación de las tres opciones barajadas puede concluirse que la solución que considera la clasificación de las 13 variables en 4 conglomerados es la finalmente aceptada no sólo siguiendo criterios estadísticos, sino también lógico-sustantivos, a decir por el significado de las variables que componen los conglomerados.

Tabla A. Conglomerado de pertenencia

Caso	*4 conglomerados*	*3 conglomerados*	*2 conglomerados*
Simpatía marroquí	1	1	1
Leyes inmigración	2	2	1
Sexo	3	3	2
Edad	4	2	1
Núm. inmigrantes	2	2	1
Regularizar inmigrante	2	2	1
Entrada inmigrante	2	2	1
Partido racista	1	1	1
Casar con marroquí	1	1	1
Estudios	4	2	1
Ingresos	4	2	1
Vecino marroquí	1	1	1
Inmigrante delincuente	2	2	1

Para la clasificación de "casos", también se solicitó el rango de soluciones de 2 a 4 conglomerados. Al ser los "casos" los objetos a clasificar, la tabla adquiere una gran dimensión. Por lo que sólo se expone un extracto de la misma (tabla B). Para que el lector pueda nuevamente observar cómo la solución de conglomerados se ve afectada por el *algoritmo* de clasificación y la medida de *distancia* o *similitud*, la tabla B incluye extractos de las tablas de *conglomerados de pertenencia* para el rango de 2 a 4 conglomerados en tres de los procedimientos probados.

Primero figura la solución finalmente aceptada: la obtenida aplicando el método *Ward*. Segundo, la correspondiente al *algoritmo* de *distancias mínimas* (o *vinculación simple*). Y, tercero, la obtenida con el método de *promedio inter-grupos*. En las tres modalidades se aplicó la misma medida de *distancia*: la *euclídea al cuadrado*. De esta forma puede comprobarse, igualmente, el efecto del *algoritmo* de clasificación. Obsérvese que se cumple la crítica comúnmente pronunciada contra los *algoritmos* de *distancias* ("mínimas"), e incluso de *promedio intergrupos*, de formar conglomerados "encadenados". Además, recuérdese que la solución de 3 conglomerados era la más idónea para la clasificación de "casos", a decir por los coeficientes de conglomeración.

Tabla B. Conglomerados de pertenencia en distintas soluciones posibles de conglomerados

Casos	*Método Ward*			*Distancias mínimas*			*Promedio intergrupos*		
	4 congl.	*3 congl.*	*2 congl.*	*4 congl.*	*3 congl.*	*2 congl.*	*4 congl.*	*3 congl.*	*2 congl.*
1: Caso 16	1	1	1	1	1	1	1	1	1
2: Caso 48	2	2	2	1	1	1	2	2	2
3: Caso 72	3	2	2	1	1	1	2	2	2
4: Caso 87	1	1	1	1	1	1	2	2	2
5: Caso 94	4	3	1	1	1	1	2	2	2
6: Caso 104	2	2	2	1	1	1	2	2	2
7: Caso 109	1	1	1	1	1	1	1	1	1
8: Caso 116	3	2	2	1	1	1	2	2	2
9: Caso 129	1	1	1	2	2	1	1	1	1
10: Caso 141	4	3	1	1	1	1	2	2	2
11: Caso 164	2	2	2	1	1	1	2	2	2
12: Caso 171	4	3	1	1	1	1	2	2	2
13: Caso 188	3	2	2	1	1	1	2	2	2
14: Caso 191	2	2	2	1	1	1	2	2	2
15: Caso 234	3	2	2	1	1	1	2	2	2
16: Caso 242	4	3	1	1	1	1	2	2	2
17: Caso 246	2	2	2	1	1	1	2	2	2
18: Caso 256	2	2	2	1	1	1	2	2	2
19: Caso 257	3	2	2	1	1	1	2	2	2
20: Caso 265	3	2	2	1	1	1	2	2	2
.....	...	...	...	...	...	...	...	...	...

* Esta tabla extracta la clasificación de los primeros 20 casos de la muestra de análisis.

** La medida de distancia aplicada en los tres *algoritmos* de clasificación es la *euclídea al cuadrado*.

3.5.1.2. Métodos de conglomeración no jerárquicos

En la conglomeración *no jerárquica* se desestima toda información relativa al número de conglomerados creados en cada etapa, al no seguirse un proceso de formación de conglomerados gradual. El interés no está en comprobar cuántos conglomerados pueden constituirse, sino en analizar, pormenorizadamente, la composición de los conglomerados previamente definidos. En la variedad analítica más popular dentro de los métodos *no jerárquicos*, la llamada "K-means" (o *K-medias*), las tablas de resultados fundamentales son las siguientes:

A) Tablas de los centros de los conglomerados

En el procedimiento *K-medias* se obtiene información, por separado, de los centros de los conglomerados (o *centroides*) "iniciales" y los "finales" y de la distancia habida entre ellos. Lo más probable es que los *centroides iniciales* y *finales* no coincidan, sobre todo, cuantas más iteraciones se hayan realizado hasta la solución final. Recuérdese que los *centroides* o *centros* de los conglomerados son, simplemente, la media de las variables para los casos que forman el conglomerado. Su valor depende de la composición de los mismos.

El análisis se inicia a partir de una composición de los conglomerados concreta: los *centros iniciales* de los conglomerados. Ésta se obtiene bien tras la realización de un análisis previo (normalmente un análisis de conglomerados *jerárquico*) o bien ha sido estimada por el programa una vez especificado el número de conglomerados que se desea formar. En este último caso, es el propio programa quien estima iterativamente los *centros iniciales* de los conglomerados. Para ello utiliza los "K" primeros casos del fichero de datos como estimaciones "provisionales" de los *centroides iniciales* (siendo "K" el número de conglomerados indicado).

Los objetos se asignan, siguiendo un proceso iterativo, a los conglomerados hacia cuyo *centroide* se sitúen más próximos. Tras cada reasignación, se recalculan los *centros* de los conglomerados, considerando las características de los objetos ahora asignados a cada conglomerado. Éstos no tienen por qué coincidir, necesariamente, con los objetos inicialmente asignados al conglomerado.

Cada nueva reasignación provoca un nuevo cálculo de la distancia entre los objetos y los nuevos *centroides*. La medida de *distancia* utilizada es la *euclídea*. El proceso concluye cuando se ha llegado al criterio de *convergencia*, que hace referencia a la proporción de modificaciones que se han producido. El investigador suele determinarlo al principio del análisis. La condición que se impone es que el criterio de *convergencia* fijado sea superior a 0 e inferior a 1.

Por último, indicar que la interpretación de los conglomerados, atendiendo a sus *centroides*, es más sencilla cuando las variables no están *estandarizadas* (expresadas en su unidad de medida original) que cuando lo están (en unidades de desviación típica). No obstante, ha de insistirse en que la correcta realización de *K-medias* exige que las

variables hayan sido previamente estandarizadas. Excepto cuando las variables compartan una misma escala de medida o no exista mucha disparidad entre sus rangos; es decir, que incluyan un número similar de valores.

EJEMPLO DE CENTROS DE LOS CONGLOMERADOS

Para la realización del análisis de conglomerados *K-medias* se decide que los centros *iniciales* de los conglomerados sean estimados iterativamente por el programa (SPSS). En su estimación se emplean los "K" primeros casos en el fichero de datos como estimaciones "provisionales" de los *centroides*. Como los análisis de conglomerados *jerárquicos* hechos con anterioridad concluían que la clasificación de los casos, en relación a su actitud declarada ante la inmigración, en tres grupos es la más adecuada, se especifica que el número de conglomerados a formar sea 3 (K = 3). A continuación se exponen juntas las tablas de los centros de conglomerados *iniciales* (tabla A) y *finales* (tabla B), para facilitar la comparación de ambas configuraciones de los conglomerados. Recuérdese que los centros de los conglomerados se hallan definidos por los valores promedios de las variables en los casos que componen el conglomerado. Sus valores obviamente se alteran tras cada modificación en la composición de los conglomerados. Los siguientes datos corresponden a la *muestra de análisis*. Las variables están *tipificadas* para evitar el efecto distorsionador que supone la inclusión de las variables "ingresos" y "edad", fundamentalmente, al tener un rango de valores muy superior al del resto de variables.

Tabla A. Centros iniciales de los conglomerados

	Conglomerado		
	1	2	3
Puntuac.: simpatía marroquí	– 2,18611	– 2,18611	– ,35302
Puntuac.: leyes inmigración	– 1,63676	2,22651	1,26070
Puntuac.: sexo	– ,96277	1,03825	1,03825
Puntuac.: edad	1,05182	– 1,09959	– ,60311
Puntuac.: núm. inmigrantes	1,23176	1,23176	– ,35958
Puntuac.: regularizar inmigrante	,56676	– 1,76361	,56676
Puntuac.: entrada inmigrante	1,73368	3,31970	– 1,43838
Puntuac.: partido racista	1,01157	4,52535	– ,74532
Puntuac.: casar con marroquí	2,30546	2,30546	– ,66661
Puntuac.: estudios	– ,95759	– ,12358	2,37843
Puntuac.: vecino marroquí	– ,39717	4,06515	– ,39717
Puntuac.: ideología política	,16690	,16690	– ,85375
Puntuac.: ingresos	– ,65560	6,94657	6,94657
Puntuac.: inmigrante delincuente	,86559	,86559	– 1,15475

Tabla B. Centros de los conglomerados finales

	Conglomerado		
	1	2	3
Puntuac.: simpatía marroquí	– ,07100	– 1,07887	,39915
Puntuac.: leyes inmigración	– ,24637	– ,75394	,54050
Puntuac.: sexo	– ,13882	,23784	,05351
Puntuac.: edad	,62919	,23539	– ,67026
Puntuac.: núm. inmigrantes	,33460	,60549	– ,52554
Puntuac.: regularizar inmigrante	– ,06827	– ,84117	,41497
Puntuac.: entrada inmigrante	,04485	,93063	– ,34354
Puntuac.: partido racista	,06268	,77813	– ,31330
Puntuac.: casar con marroquí	– ,08016	1,56400	– ,40302
Puntuac.: estudios	– ,64826	– ,23629	,60673
Puntuac.: vecino marroquí	– ,26942	2,01062	– ,31847
Puntuac.: ideología política	,11977	,39017	– ,17369
Puntuac.: ingresos	– ,42680	– ,12941	,54231
Puntuac.: inmigrante delincuente	,39976	,66489	– ,52084

Como todas las variables se hallan tipificadas (expresadas en la misma unidad de medida: puntuaciones Z), su interpretación se hace en términos de unidades de desviación típica por encima (signo positivo) o por debajo (signo negativo) de la media. La *media* de una variable tipificada o *normalizada* es siempre 0 y su *desviación típica* 1.

De la comparación de las tablas A y B puede concluirse que los centros de los conglomerados *finales* difieren bastante de los *iniciales*. Esta divergencia es normal porque se han producido varias modificaciones en los casos que componen cada conglomerado. La tabla C detalla el *historial de las iteraciones*.

Tabla C. Historial de iteraciones[a]

	Cambio en los centros de los conglomerados		
Iteración	*1*	*2*	*3*
1	4,872	9,032	4,037
2	,702	,519	2,862
3	,426	,274	,504
4	9,277 E – 02	5,317 E – 02	7,420 E – 02

[a] Covergencia alcanzada debido a un cambio en la distancia nulo o pequeño. La distancia máxima en la que ha cambiado cada centro es 4,655 E – 02. La iteración actual es 4. La distancia mínima entre los centros inicales es 10,151.

Aunque el programa SPSS aplica por defecto un número máximo de iteraciones de 10, su número se ha reducido a 4, al especificarse como criterio de *convergencia* el valor ,02. Significa que el proceso de iteración concluye cuando no se obtiene ninguna modificación en los *centroides* superior al 2% de la distancia más pequeña entre cualquiera de los centros de los conglomerados *iniciales*.

De la interpretación de los centros de los conglomerados *finales* (tabla B) se concluye la existencia de 3 grupos diferentes de españoles de acuerdo a su actitud declarada ante la inmigración. Los grupos más polares son los pertenecientes al conglomerado 2 (los más "reacios" a la inmigración) y al conglomerado 3 (los más "favorables"). Los incluidos en el conglomerado 1 se sitúan entre ambos extremos. El conglomerado 1 reúne al 40,1% de los 1.229 casos válidos que componen la *muestra de análisis* (exactamente, 493 casos); el conglomerado 2, a 165 casos (el 13,4% de la *muestra de análisis*); y el conglomerado 3 agrupa casi a la mitad de la muestra (571 casos que suponen el 46,5% de la muestra).

La configuración de los conglomerados obtenida en la *muestra de validación* (1.256 casos) coincide con la extraída en la *muestra de análisis*, como puede verse en la tabla D, salvo alguna divergencia en los valores "exactos" de los *centroides*. El conglomerado 1 reúne a 516 casos (41,1%), el conglomerado 2 a 116 casos (9,2%) y el conglomerado 3 a 624 casos (49,7%) de los 1.256 casos de la *muestra de validación*.

Tabla D. Centros de los conglomerados finales

	Conglomerado		
	1	2	3
Puntuac.: simpatía marroquí	– ,21947	– 1,26956	,41504
Puntuac.: leyes inmigración	– ,38826	– ,76972	,41341
Puntuac.: sexo	– ,12126	,10674	,02171
Puntuac.: edad	,48722	,32279	– ,47192
Puntuac.: núm. inmigrantes	,51249	,72610	– ,53871
Puntuac.: regularizar inmigrante	– ,31324	– ,81762	,41469
Puntuac.: entrada inmigrante	,27232	,82738	– ,42245
Puntuac.: partido racista	,21763	,93986	– ,29126
Puntuac.: casar con marroquí	,07441	1,74436	– ,40331
Puntuac.: estudios	– ,55937	– ,25991	,50262
Puntuac.: vecino marroquí	– ,22834	2,81465	– ,32508
Puntuac.: ideología política	,22377	,22662	– ,19137
Puntuac.: ingresos	– ,36695	,01537	,35722
Puntuac.: inmigrante delincuente	,53849	,68024	– ,58816

En suma, casi la mitad de los encuestados pueden clasificarse como "favorables" a la inmigración (conglomerado 3), de acuerdo con sus respuestas a los ítems analizados. Atendiendo a la descripción de las 14 variables (descritas en el capítulo 1), puede observarse que el conglomerado 3 agrupa a los sujetos que más "simpatía" manifiestan ante los marroquíes (se sitúan por encima de la media) y, en consecuencia, a los que menos preocupa un futuro "matrimonio" de un hijo con un marroquí o tener como "vecinos" a una familia de marroquíes. Son contrarios al auge de los "partidos" de ideología "racista" (valor negativo) y, por el contrario, favorables a la "regularización" de los inmigrantes ilegales (signo positivo) y a facilitarles la "entrada". El signo negativo de esta segunda variable ("entrada inmigrante") es consonante con cómo está medida: el valor más bajo, 1, corresponde a facilitar la entrada a trabajadores inmigrantes; el 2 sólo a aquellos que tengan un contrato de trabajo; 3, dificultad su entrada y 4 prohibirla.

Los integrantes del componente 3 también se caracterizan por considerar "duras" las "leyes de inmigración" y "pocos" los inmigrantes existentes en nuestro país. Además, no pien-

san que su presencia incremente la "delincuencia" en España (signo negativo). Entre sus características sociodemográficas destaca el tener un nivel de "ingresos" superior a la media, al igual que su nivel de "estudios". Son preferentemente jóvenes, en proporciones similares de varones que de mujeres. Respecto a su "ideología política" se ubican más a la izquierda. Recuérdese que la escala aplicada iba desde 01 (extrema izquierda) a 10 (extrema derecha). Para seguir la interpretación de la composición de cada conglomerado, se recomienda releer la descripción de las variables expuesta en el capítulo 1.

Por el contrario, el grupo menos populoso es el que forma el conglomerado 2: los "contrarios" a la inmigración. Se distinguen por ser los que menos manifiestan "simpatía" hacia los marroquíes (se sitúan a 1,08 y 1,27 unidades de desviación típica por debajo de la media en la muestra de *análisis* y de *validación*, respectivamente) y, consecuentemente, son a los que más preocupa un futuro "matrimonio" o tener como "vecinos" a inmigrantes marroquíes. Consideran que el aumento de los inmigrantes favorece el incremento de la "delincuencia" en España y califican muy positivo el auge de los "partidos de ideología racista". Su consideración del "número de inmigrantes" se sitúa por encima de la media. Son contrarios a la "regularización" de los inmigrantes ilegales y favorables a prohibirles la "entrada" en nuestro país. Califican las "leyes de inmigración" de muy "tolerantes". Respecto a sus características sociodemográficas, se distinguen por tener un nivel de "estudios" y de "ingresos" por debajo de la media (inferior al de los componentes del conglomerado 3). Su "edad" media supera a la de los integrantes del conglomerado 3, aunque es inferior a la de los clasificados en el conglomerado 1. Predominan los "varones" y, en general, las personas de "ideología política" de "derechas".

Por último, los integrantes del conglomerado 1, que agrupa a cuatro de cada diez encuestados, se sitúan entre ambos extremos, en cuanto a su actitud ante la inmigración. Respecto a su perfil sociodemográfico, se distinguen por ser los que menor nivel de "ingresos" y de "estudios" presentan y, en consonancia, los de más "edad". Es un colectivo integrado sobre todo por "mujeres" y por personas de "derechas".

De los *centroides* también se ofrece información relativa a las distancias entre los *centroides* finales. Esta información permite conocer qué conglomerados se sitúan más alejados y cuáles más próximos.

Ejemplo de tabla de distancias entre los centros de conglomerados finales

Una vez constituidos los conglomerados, se comprueba las "distancias" (*euclídeas*) entre los pares de los centros de los conglomerados *finales*. Ello permite conocer lo "separado" que están los distintos conglomerados: cuáles se sitúan más "próximos" y cuáles más "distantes". Se quieren conglomerados muy "alejados" unos respecto de otros, e integrados por casos bastante "próximos" al centro del conglomerado. La siguiente tabla incluye las distancias entre los centros de los conglomerados *finales*. En ella puede observarse que los dos conglomerados más "distantes" son, obviamente, los conglomerados 2 y 3 (4,750). Recuérdese que éstos eran los dos conglomerados "polares": el conglomerado 2 agrupa a los más "reacios" a la inmigración, mientras los integrantes del conglomerado 3 se presentan co-

mo los más "favorables" a la misma. El conglomerado 1 se sitúa entre ambos, al estar integrado por personas con una actitud "intermedia" ante la inmigración. De acuerdo con las distancias entre los centros de los conglomerados *finales*, el conglomerado 1 se halla más próximo al conglomerado 3 (2,718), los "favorables" a la inmigración, que respecto al conglomerado 2 (3,441), los más "reacios o contrarios" a la misma. Quiere esto decir, que sus características medias coinciden más con las que definen al conglomerado 3 que las que perfilan al conglomerado 2. Se insiste en que los centros de los conglomerados hacen referencia a las *medias* de las variables de los casos que forman el conglomerado.

Distancias entre los centros de los conglomerados finales

Conglomerado	*1*	*2*	*3*
1		3,441	2,718
2	3,441		4,750
3	2,718	4,750	

B) Tabla ANOVA

En el análisis de conglomerados *no jerárquico K-medias*, la adecuación de las variables en la configuración y diferenciación de los grupos se comprueba realizando un análisis de la varianza univariable (ANOVA). Mediante éste se comprueba la significatividad de la contribución de cada variable a la diferenciación entre los grupos.

Como en cualquier análisis univariable de la varianza, la tabla *ANOVA* incluye información relativa a las "medias cuadráticas" (que indican variabilidad de las variables). Pero, difiere en que las "medias cuadráticas" que se comparan son las "entre conglomerados" (en la columna etiquetada "conglomerado" o "cluster") y la "intra conglomerado" o dentro del conglomerado (en la columna etiquetada "error"). Ambas "medias cuadráticas" figuran acompañadas por sus grados de libertad. En el caso de la "entre conglomerados", los grados de libertad son iguales al número de conglomerados menos uno (gl = K – 1); en la "media cuadrática intra conglomerado", al número de casos válidos menos el número de conglomerados (gl = N – K).

Del cociente entre ambas "medias cuadráticas" se obtiene el estadístico *F*, que aparece acompañado de su significatividad. Si bien, se insiste en que el uso de este estadístico de comprobación en el análisis de conglomerados se adecua a fines descriptivos y no inferenciales. Los conglomerados se forman siguiendo el criterio principal de que sean máximas las diferencias entre los objetos de conglomerados distintos. Los niveles de significatividad no se hallan corregidos por esto, lo que limita su aplicación usual como prueba de hipótesis. A decir, la contrastación de la hipótesis *nula* de que las *medias* de los conglomerados (o *centroides*) sean iguales.

El estadístico *F* se utiliza para conocer qué medias de las variables son las que más difieren entre los conglomerados: aquéllas a las que correspondan un valor *F* más elevado y un nivel de significación bajo (usualmente, ≤ ,05). Un valor *F* bajo indica, por

el contrario, que las *medias* de dicha variable apenas difieren entre los conglomerados. Esto significa que no consigue diferenciar a los conglomerados. Hecho que la convierte en candidata a ser descartada en la descripción de las características que definen a los conglomerados. En cambio, las variables que presenten un valor *F* elevado participan en la interpretación y posterior etiquetamiento de los conglomerados creados, al diferir bastante sus *medias* entre los conglomerados.

EJEMPLO DE TABLA ANOVA EN SU APLICACIÓN EN UN ANÁLISIS DE CONGLOMERADOS

La tabla ANOVA obtenida en la muestra de *análisis* (prácticamente coincidente con la extraída en la muestra de *validación*) se expone a continuación. Obsérvese que todas las 14 variables analizadas muestran ser "relevantes" en la diferenciación de los conglomerados, a decir por las pruebas *F* realizadas. Las razones de sus *medias cuadráticas entre conglomerados* ("conglomerados") y las *intra conglomerados* ("error") resultan en valores elevados de *F*, que llevan a corroborar la "significatividad" de la contribución de todas las variables a la diferenciación de los tres grupos de personas en cuanto a su actitud ante la inmigración.

Asimismo, obsérvese que los grados de libertad entre los conglomerados son, en todas las variables, 2 (gl = K – 1 = 3 – 1), mientras que los grados de libertad dentro de los conglomerados ("error") difieren en cada variable (gl = N – K). Ello se debe a que se ha activado la opción "excluir casos (valores perdidos) según pareja" de variables para aprovechar al máximo los datos posibles.

Tabla ANOVA

	Conglomerado		*Error*		*F*	*Sig.*
	Media cuadrática	*gl*	*Media cuadrática*	*gl*		
Puntuac.: simpatía marroquí	129,218	2	,761	1.072	169,727	,000
Puntuac.: leyes inmigración	103,139	2	,816	826	126,404	,000
Puntuac.: sexo	10,234	2	,985	1.226	10,385	,000
Puntuac.: edad	229,954	2	,605	1.226	380,236	,000
Puntuac.: núm. inmigrantes	117,394	2	,763	1.033	153,939	,000
Puntuac.: regularizar inmigrante	94,641	2	,743	1.063	127,304	,000
Puntuac.: entrada inmigrante	100,344	2	,771	1.123	130,123	,000
Puntuac.: partido racista	70,547	2	,853	1.113	82,700	,000
Puntuac.: casar con marroquí	241,370	2	,569	1.189	424,059	,000
Puntuac.: estudios	196,861	2	,675	1.134	291,773	,000
Puntuac.: vecino marroquí	369,874	2	,395	1.213	936,435	,000
Puntuac.: ideología política	18,851	2	,915	896	20,594	,000
Puntuac.: ingresos	93,611	2	1,033	898	90,620	,000
Puntuac.: inmigrante delincuente	131,220	2	,748	1.028	175,457	,000

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

C) Tabla resumen

Por último, puede confeccionarse una tabla resumen que incluya, para cada caso, el conglomerado al que fue finalmente asignado. A ello se añade la distancia *euclídea* (preferente en este *algoritmo* de clasificación) entre el caso y el centro del conglomerado empleado en la clasificación del caso. Esta información es de utilidad en la identificación de los casos que se encuentran alejados de sus respectivos *centroides* (los *atípicos*). Su lejanía les convierte en casos no representativos de los conglomerados a los que han sido asignados. En cambio, la información relativa a la pertenencia del caso al conglomerado resulta relevante para la "validación" posterior de la clasificación obtenida mediante el análisis de conglomerados, acudiendo a otros procedimientos analíticos. En el ejemplo aquí ilustrado, los análisis se validaron además realizando *ex profeso* un análisis *discriminante* (capítulo 4), que partió de la configuración de los tres conglomerados obtenida con la aplicación del *algoritmo* de clasificación *K-medias*. El procedimiento seguido se detalla en el susodicho capítulo.

EJEMPLO DE TABLA DE PERTENENCIA GRUPAL

Debido al elevado tamaño muestral, aun considerando sólo la *muestra de análisis* (1.229 casos), a continuación figura un extracto de la tabla de *pertenencia a los conglomerados*. En ella figura la ubicación de los veinte primeros casos que componen la muestra de *análisis* (elegidos aleatoriamente) y los cinco últimos. Los situados a mayor distancia del *centroide* del conglomerado en el que han sido ubicados son posibles "atípicos".

Número de caso	*Conglomerado*	*Distancia*
2	3	3,029
3	1	2,594
4	1	1,574
5	1	2,610
6	1	2,229
7	1	2,516
8	3	3,133
9	1	2,249
12	3	3,792
15	1	2,214
16	2	4,211
18	2	2,284
19	3	1,842
20	1	2,351
21	1	2,663
22	3	2,422
26	2	2,883
29	2	2,384
30	3	2,980
33	1	3,250
2485	3	2,078
2486	3	2,327
2488	3	3,814
2489	3	1,870
2493	3	2,837

3.5.2. Las representaciones gráficas

Como sucede con los resultados numéricos, las representaciones gráficas también difieren atendiendo al método de conglomeración aplicado para la clasificación de los objetos. En el análisis de conglomerados *jerárquico* dos son los gráficos estrella: el *dendograma* y el gráfico de *témpanos*. En la conglomeración *no jerárquica* destacan, en cambio, los gráficos de pertenencia al conglomerado por distancia desde los *centroides* y los que representan los casos por variables representativas desde los conglomerados de variables. Para estos últimos se escogen variables que muestren ser representativas de los conglomerados formados: aquéllas a las que correspondan valores *F* elevados. Cualquiera de estos dos últimos gráficos cumple la función de ayudar a visualizar la homogeneidad de los conglomerados.

A) El dendograma

El *dendograma* (o diagrama en árbol) es la expresión gráfica que mejor representa la estructura jerárquica implícita en los procedimientos de conglomeración *jerárquica*. Muestra qué objetos (casos o variables) componen cada conglomerado, cómo se van uniendo los diversos conglomerados y la distancia a la que se unen. Es decir, ofrece, en forma gráfica, la información contenida en el *historial de conglomeración*.

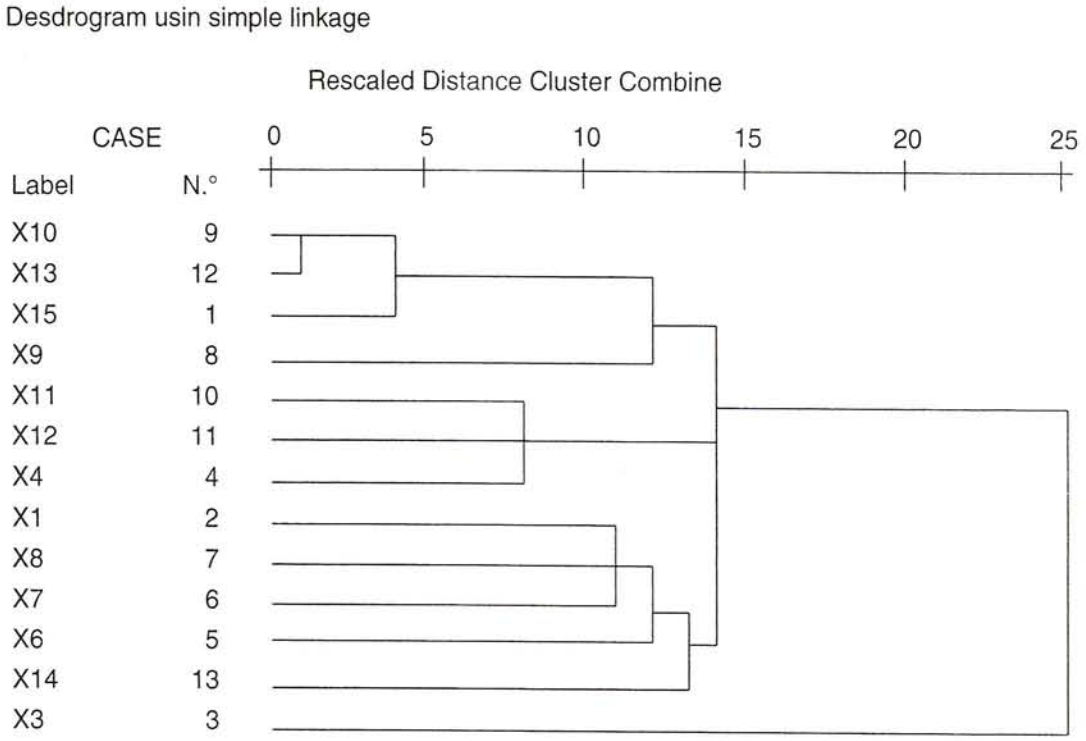
No obstante, hay que precisar que los valores de *distancia* dispuestos en el gráfico (normalmente en la parte superior, si la disposición es horizontal) no se corresponden con los valores de *distancia* reales. Están reescalados a números comprendidos en el rango de 0 a 25. Para conocer la distancia "real" a la que se combinan los objetos y los conglomerados hay que observar la tabla del *historial de conglomeración*, que contiene dicha información.

Asimismo, hay que destacar que, cuando se aplica el *algoritmo* de clasificación de *distancias máximas* (o *vinculación completa*), el coeficiente de *distancia* que se muestra para la última etapa es la mayor distancia entre un miembro de un conglomerado y un integrante de otro, al menos en el programa SPSS.

La disposición del *dendograma* puede ser horizontal o vertical. Es vertical, cuando las "ramas" del árbol se sitúan en el fondo y las "raíces" en la parte de arriba, lo que determina que su lectura sea ascendente. Los valores de *distancia* se ubican en el margen izquierdo o derecho del gráfico. En las salidas de ordenador estándares es muy habitual, sin embargo, la impresión horizontal del *dendograma*. Las "ramas" del árbol se disponen a la izquierda y los valores de distancia en la parte superior del gráfico. Su lectura es horizontal, de izquierda a derecha. Para evitar que las ramas del árbol se crucen, los objetos a clasificar suelen posicionarse juntos. Su orden de aparición no coincide, por tanto, con el otorgado antes de proceder a los análisis.

EJEMPLO DE DENDOGRAMA

Dada la gran dimensionalidad del *dendograma* correspondiente a la clasificación de "casos" (mediante el procedimiento de conglomeración *no jerárquico aglomerativo* utilizando el método *Ward* y la distancia *euclídea al cuadrado*), se ha optado por exponer sólo el *dendograma* del modelo de clasificación de "variables" (mediante el *algoritmo* de *vinculación simple*, utilizando la *correlación de Pearson* como medida de similitud entre las variables). Para facilitar la interpretación del *dendograma*, se recomienda releer la tabla del *historial de conglomeración* de dicho modelo de clasificación (subapartado 3.5.1).



La disposición del *dendograma* ilustrado es "horizontal", siendo su lectura transversal, de izquierda a derecha. Las variables están dispuestas no por orden de entrada para la realización del análisis, sino por "similitud". Las variables que acaban agrupándose en un mismo conglomerado aparecen juntas para evitar que las ramas del *dendograma* se crucen.

Como se ha dicho, la medida de distancia que figura en la parte de arriba del gráfico se encuentra reescalada al rango de valores de 0 a 25. Las líneas verticales indican la formación de un conglomerado y la posición de la línea en la escala la distancia a la que los conglomerados se unieron: la distancia más pequeña ha sido la 1 y la más elevada 25. Pero, aunque las distancias estén reescaladas, adviértase que la razón de dichas distancias es la misma que la razón de las distancias originales.

Las variables que primero se unen son las más correlacionadas entre sí: X_{10} ("casar con marroquí") y X_{13} ("vecino marroquí"), como ya se vio en el *historial de conglomeración*. Lo que añade el *dendograma* es la visualización gráfica del proceso de formación de los conglomerados; cómo se van uniendo, primero, variables y después conglomerados hasta la solución final de

cuatro conglomerados. Obsérvese que la variable X_3 no se halla vinculada a ninguna otra variable, sino que ella sola constituye un conglomerado. En el valor de distancia máximo (reescalado al valor 25) se produce la solución de un único conglomerado que resulta de la unión de dicha variable (X_3, "sexo"), que difiere bastante de las demás (al haberse unido al resto al final del proceso de conglomeración, en el valor de distancia máximo), al gran conglomerado obtenido de la agrupación de los otros tres conglomerados. Nuevamente se insiste en la conveniencia de leer dicho gráfico en conjunción con el *historial de conglomeración.*

Por último, señalar que la diferencia entre los cuatro conglomerados en los que pueden agruparse las 13 variables es amplia según indica el *dendograma*: su unión se produce a niveles de distancia (reescalados) elevados. Además, la unión final de todos los conglomerados en un único conglomerado se da en el último valor posible: el 25. Todo lo cual indica diferencias considerables entre los conglomerados formados.

B) Diagrama de témpanos

El nombre de este gráfico, *témpanos* o *carámbanos* (del inglés "icicle"), le viene de la forma que adopta. Ésta se asemeja a una fila de *carámbanos*, o estalactitas de hielo, que cuelgan de los aleros de los tejados cuando ha nevado.

Como el *dendograma*, el gráfico de *témpanos* puede disponerse de manera horizontal y vertical. La opción horizontal suele preferirse cuando los objetos a clasificar son muchos y existe dificultad para su representación en una única página. Salvo en esta situación, la disposición habitual del gráfico es la vertical.

En las columnas se representan los objetos a clasificar. Éstos se identifican bien por una etiqueta o bien por el número secuencial que se les asignó en el fichero de datos. En las filas se localizan los distintos pasos habidos en la conglomeración *jerárquica* realizada.

Para seguir la secuencia de pasos correctamente, el gráfico se lee, como sucede con el *dendograma* vertical, de abajo a arriba. La primera fila (en la parte superior del gráfico) incluye un único conglomerado. Éste se halla integrado por todos los objetos que se quiere clasificar. En cambio, en la última fila habrá tantos conglomerados como objetos. Dada su irrelevancia para el análisis, esta última fila, que corresponde al paso 0 del análisis, no siempre aparece dibujada en el gráfico.

EJEMPLO DE DIAGRAMA DE TÉMPANOS

En la representación del proceso de formación de los conglomerados de "variables", de cómo se combinan las variables en conglomerados en cada iteración del proceso analítico, también se realizó un diagrama de *témpanos*. La opción escogida fue la "vertical" porque el número de objetos a clasificar (variables) no era elevado y favorecía esta disposición. El gráfico obtenido fue el siguiente:

Diagrama de témpanos vertical

Número de conglomer.	*Caso*																								
	Sexo		Ingresos		Estudios		Edad		Inmigrante delincuente		Núm. inmigrantes		Regularizar inmigrante		Entrada inmigrante		Leyes inmigración		Partido racista		Vecino marroquí		Casar con marroquí		Simpatía marroquí
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
4	X		X	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
5	X		X	X	X	X	X		X		X	X	X	X	X	X	X		X	X	X	X	X	X	X
6	X		X	X	X	X	X		X		X		X	X	X	X	X		X	X	X	X	X	X	X
7	X		X	X	X	X	X		X		X		X	X	X	X	X		X		X	X	X	X	X
8	X		X	X	X	X	X		X		X		X		X	X	X		X		X	X	X	X	X
9	X		X	X	X	X	X		X		X		X		X		X		X		X	X	X	X	X
10	X		X	X	X		X		X		X		X		X		X		X		X	X	X	X	X
11	X		X		X		X		X		X		X		X		X		X		X	X	X	X	X
12	X		X		X		X		X		X		X		X		X		X		X	X	X		X

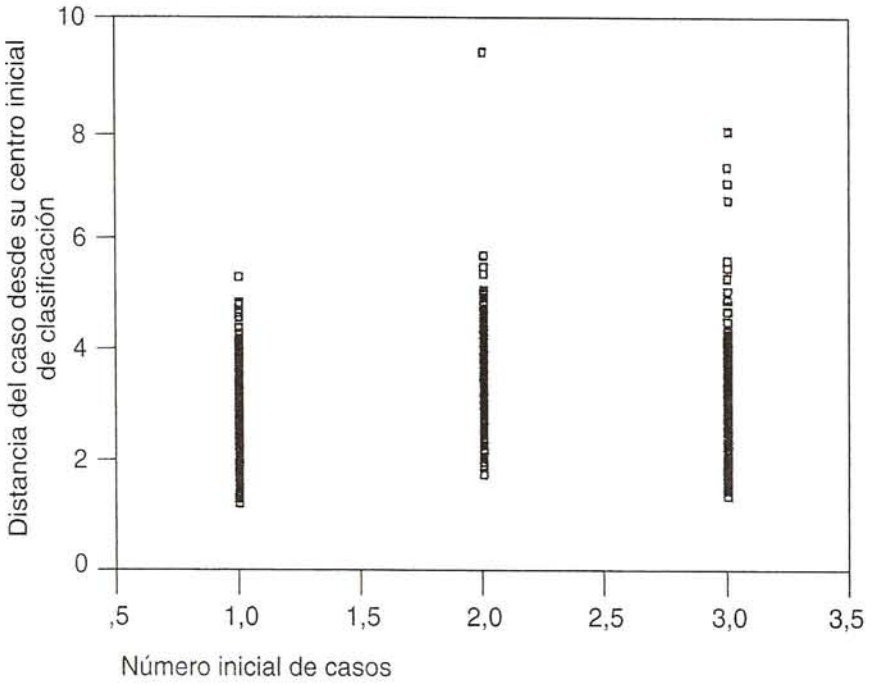
Este gráfico resume los pasos habidos en la formación de los conglomerados: 12, al ser 13 las variables que quiere agruparse. Obsérvese que la disposición del gráfico es contraria a la que presenta la tabla del *historial de conglomeración*. En el diagrama de *témpanos* las filas representan el número de conglomerados y no la etapa de conglomeración. La primera fila incluye todas las variables formando un único conglomerado, mientras en la última (la 12) son 12 los conglomerados existentes: uno agrupa a las dos variables más correlacionadas ("casar con marroquí" y "vecino marroquí") y los 11 restantes están integrados sólo por una variable.

El *diagrama de témpanos* puede considerarse un diagrama de "síntesis" que, sin embargo, no logra ofrecer toda la información contenida en la tabla del *historial de conglomeración*. Para su interpretación se recomienda, igualmente, revisar el *historial de conglomeración*.

C) Gráficos de pertenencia al conglomerado

En el análisis de conglomerados *no jerárquico*, los gráficos de interés no atienden a la descripción de los conglomerados que se constituyen en cada etapa, sino a la composición de dichos conglomerados. Las opciones gráficas más habituales son los gráficos de pertenencia al conglomerado por distancias desde los *centroides* y aquellos que representan los casos por variables representativas desde los conglomerados de variables. Para estos últimos se escogen variables que muestren ser representativas de los conglomerados surgidos; aquellas a las que correspondan valores *F* elevados. Cualquiera de estos gráficos cumple la función de ayudar a visualizar la homogeneidad de los conglomerados.

EJEMPLO DE GRÁFICO DE PERTENENCIA AL CONGLOMERADO



Este gráfico de dispersión representa los casos teniendo en cuenta su pertenencia al conglomerado y su distancia del centro del conglomerado al que ha sido asignado. Muestra gran utilidad en la identificación de *atípicos*, de casos que se sitúan bastante distanciados de los centros de sus conglomerados, no siendo representativos de los mismos. En él puede apreciarse que en el conglomerado 3 es donde más casos no se ajustan al perfil que define dicho conglomerado, al no coincidir con sus características medias. En el conglomerado 2 la presencia de *atípicos* es más evidente (su distancia respecto al *centroide* es mayor), si bien su cuantía es menor. Aunque existan casos bastante distanciados del centro del conglomerado, su número es inferior al habido en el conglomerado 3. En cambio, en el conglomerado 1 las disimilaridades entre sus integrantes es menos apreciable.

3.5.3. La detección de atípicos

A lo largo del presente capítulo se ha hecho referencia al posible efecto distorcionador de los *atípicos* y cómo puede detectarse su presencia. Véase, por ejemplo, lo referente a *atípicos* potenciales en el *historial de conglomeración* y en el subapartado anterior. En la conglomeración *jerárquica* los *atípicos* suelen coincidir con aquellos objetos que se unen tarde al proceso de conglomeración (en la tabla del *historial de conglomeración* y en sus expresiones gráficas, como el *dendograma*, por ejemplo, cuando se visualice una "rama" que se une al resto casi al final del proceso de agrupación). En la conglomeración *no jerárquica*, los *atípicos* coinciden con objetos (casos o variables)

situados a bastante distancia de los centros de los conglomerados a los que han sido asignados.

A la detección común de *atípicos* (todo aquel que exceda el valor estandarizado de +3,0 o se halle por debajo de –3,0) en el análisis de conglomerados se añade, para el mismo objetivo, la utilización del estadístico D^2 *de Mahalanobis*. Cualquier observación que presente una puntuación extrema en este estadístico se convierte en *atípico* potencial, que puede distorsionar la verdadera estructura de los datos. De ahí la conveniencia de evaluar su incidencia en los resultados del análisis, antes de proseguir con su interpretación.

Para no redundar en aspectos ya tratados con anterioridad, se remite a la lectura de los capítulos anteriores. En especial, del subapartado 1.5.4, que trata de los procedimientos habituales para la detección de *atípicos*, así como de los remedios más comunes.

3.5.4. El perfil de los conglomerados

A la obtención de los conglomerados le sigue el análisis de los perfiles de los conglomerados. Trata de la descripción de las características que más diferencian o discriminan a los diversos conglomerados surgidos del análisis. Este conocimiento (el perfil del conglomerado) es de gran utilidad para la posterior predicción de la pertenencia de un determinado objeto a un conglomerado concreto.

El *perfil* de los conglomerados puede lograrse mediante la aplicación combinada del análisis de *conglomerados* con el análisis *discriminante*. Primero, se lleva a cabo un análisis de *conglomerados* para la clasificación de los objetos en conglomerados. Los conglomerados que resultan del análisis pasan, en una fase posterior, a ser analizados con detenimiento mediante la realización de un análisis *discriminante* en la misma serie de datos. Los conglomerados actúan de variables dependientes y las variables que ayudarán a perfilar las características de los conglomerados (variables sociodemográficas, económicas u otras analizadas) como variables independientes. Lo usual es utilizar variables no incluidas en la solución de conglomerados. El procedimiento seguido se expone en el capítulo 4, dedicado al análisis *discriminante*. En él se ejemplifica el uso combinado de ambas técnicas analíticas multivariables, a partir de la clasificación de "casos" obtenida mediante le procedimiento de conglomeración *no jerárquica K-medias* (o "K-means").

3.6. Validación de los resultados

La *validación* de los resultados es una tarea que no debe descuidarse en cualquier procedimiento analítico. Se sitúa en las postrimerías del análisis y previo a su conclusión. A la obtención de los conglomerados y de sus perfiles, le sigue la comprobación de su *significatividad* y posibilidades de inferencia a la población a la que pertenece la muestra analizada. Pero, dado el carácter "exploratorio" y, a veces, ateórico, del

análisis de conglomerados, la utilidad de esta última fase analítica se halla, al igual que la configuración de los perfiles, sujeta a debate.

Autores como Hair *et al.* (1999: 528) reivindican la necesidad de su realización. En su opinión "es esencial que el investigador lleve a cabo todos los tests para confirmar la validez de la solución *cluster* a la vez que asegura que la solución cluster tiene significación práctica. Los investigadores que minimizan o se saltan este paso se exponen al riesgo de aceptar una solución que se especifica sólo para la muestra y tiene una generalización limitada o incluso reducida cuando se utiliza más allá de la mera descripción de los datos sobre las variables de obtención de conglomerados".

Aldenderfer y Blashfield (1984) destacan cinco técnicas posibles para *validar* una solución de análisis de conglomerados:

A) La correlación "cophenetic"

Fue originariamente propuesta por Sokal y Rohfl en 1962 (en "The comparison of dendograms by objective methods", *Taxon*, 11: 33-40) y ha llegado a convertirse en el principal procedimiento de *validación* defendido por los taxonomistas numéricos.

Se define como la correlación entre los valores de la matriz de similaridad original y los valores de la matriz de similaridad resultante del análisis. Ambos tipos de valores no siempre coinciden en número. Lo habitual es que el número de valores únicos en la matriz de similaridad final sea inferior al número de valores únicos en la matriz de similaridad original. Ello revierte negativamente en que la cantidad de información contenida en ambas matrices sea bastante diferente.

Su uso, no obstante, se limita a la conglomeración *jerárquica*. La finalidad principal es comprobar si la solución de conglomerados *jerárquica* (representada, por ejemplo, mediante un *dendograma*) logra plasmar el modelo de similaridades entre los objetos.

B) Tests de significatividad en las variables utilizadas en la creación de conglomerados

Consiste en aplicar un análisis multivariable de la varianza (MANOVA) en el conjunto de variables empleadas en la creación de los conglomerados. También puede realizarse un análisis de la varianza por separado, en cada una de las variables analizadas (ANOVA). Cualquiera de estas dos actuaciones ayuda a conocer la significatividad de las variables que caracterizan a los conglomerados. El mismo objetivo puede, sin embargo, alcanzarse recurriendo a otras técnicas analíticas, como el análisis discriminante, por ejemplo.

A diferencia de la *correlación "cophenetic"*, esta segunda técnica de *validación* es de uso generalizado. Puede llevarse a efecto en toda la variedad de procedimientos de conglomeración existente.

C) Tests de significatividad en variables no empleadas para la formación de los conglomerados

Esta tercera técnica, también llamada de "validación externa", se propone como uno de los mejores procedimientos para validar la solución del análisis de conglomerados. Consiste en la realización de pruebas de significatividad que comparen los conglomerados en variables que no se han utilizado en la generación de la solución de conglomerados. Este proceder coincide con la modalidad llamada "validez de criterio", propuesta por Carmines y Zeller (1979) para todo procedimiento de medición. Su poder descansa en la comprobación de la generabilidad de una solución de conglomerados contra "criterios relevantes". La dificultad está, precisamente, en cómo definir la serie de criterios externos "relevantes" para que sirvan de referente en la contrastación de la clasificación obtenida mediante el análisis de conglomerados.

D) Replicación

La *replicación* hace referencia a lo que comúnmente se entiende por *fiabilidad*. Es decir, "la capacidad de obtener resultados consistentes en mediciones sucesivas del mismo fenómeno" (Jacob, 1994: 363).

En el análisis de conglomerados la *replicación* consiste en comprobar la consistencia interna de la solución de conglomerados. Para ello se aplica el mismo procedimiento de análisis a diferentes muestras extraídas de la misma población. La finalidad es demostrar la generabilidad de la clasificación obtenida mediante el análisis de conglomerados. Se quiere mostrar que aparecen los mismos conglomerados en distintas subseries de datos, cuando se aplica el mismo método de conglomeración. Si las soluciones no coinciden, la clasificación se considera no estable. Esto significa su pérdida de *validez* y posterior utilidad, lo que puede llevar a rechazar la solución del análisis de conglomerados.

Por el contrario, si se obtiene la misma solución de conglomerados en análisis repetidos, puede significar que tiene generabilidad. Si bien, se advierte que una replicación exitosa no garantiza la validez de la solución. Se recomienda que, antes de llegar a dicha conclusión, la solución de conglomerados se compruebe, asimismo, mediante otros procedimientos alternativos.

E) Procedimientos de Monte Carlo

Esta última técnica para comprobar la *validez* de la solución de conglomerados ha tenido menor aplicación que las precedentes. Consiste en la aplicación de procedimientos de *Monte Carlo*. Primero, se generan números aleatorios con el objetivo de crear una serie de datos con características generales que casen con las globales de los datos originales. A continuación, se emplea el mismo método de conglomeración tanto en los datos "reales" como en los "artificiales" creados al efecto. Después se comparan ambas soluciones.

Para facilitar la comparación, pueden aplicarse procedimientos de *partición iterativos*, como por ejemplo *K-medias*. Este *algoritmo* de clasificación posibilita la comparación de los valores *F* en todas las variables analizadas.

Los resultados también pueden compararse con la ayuda de gráficos que retraten la propuesta de clasificación que resulta de los análisis.

• Entroncando con esta quinta técnica de *validación* propuesta por Aldenderfer y Blashfield (1984) en el análisis de conglomerados, está la aproximación de uso generalizado en cualquier modalidad analítica (y es uno de los seguidos en la ejemplificación de la realización de esta técnica analítica). Consiste en dividir la muestra en dos submuestras. Ambas se analizan por separado para, después, comparar sus resultados. Este procedimiento se ve limitado, no obstante, por el tamaño de la muestra. Exige que el tamaño de la muestra original sea elevado para que su división en dos submuestras (ya sea a la mitad o al 60 y 40%, por ejemplo) no repercuta negativamente en la pérdida de significatividad estadística de los resultados del análisis.

• En resumen, existen varios procedimientos para comprobar la *validez* y consistencia (o *fiabilidad*) de los resultados de un análisis de conglomerados. A veces no serán todos ellos factibles, como sucede con la *correlación "cophenetic"* (cuando la conglomeración es *no jerárquica*) o la división de la muestra en dos submuestras (si el tamaño muestral original es pequeño), por ejemplo. Pero para la generabilidad de las situaciones el investigador puede elegir entre varias de las opciones posibles. Además, hay que insistir en la conveniencia de aplicar distintas técnicas de *validación*; ya sea repitiendo los mismos análisis en otras muestras, ya analizando la misma muestra mediante otro método de conglomeración (utilizar primero uno *jerárquico* para, posteriormente, validarlo mediante uno *no jerárquico*), o ya acudiendo a otra técnica analítica multivariable con la que analizar la misma serie de datos. En especial, el análisis *discriminante* y el análisis *factorial confirmatorio*. Ambas técnicas analíticas se presentan como "confirmatorias" y "explicativas", y no como meramente "exploratorias", como sucede con el análisis de *conglomerados*, favoreciendo la "inferencia" y posibilidades de generalización de los resultados del estudio. Ambas técnicas se explican en capítulos posteriores, a cuya lectura se remite. Asimismo, para obtener mayor información sobre distintos procedimientos de comprobación de la *validez* y *fiabilidad*, se aconseja leer Carmines y Zeller (1973), Cea D´Ancona (1996) o De Vaus (1990).

LECTURAS COMPLEMENTARIAS

Aldenderfer, M. S. y Blashfield, R. K. (1984). *Cluster analysis*, Beverly Hills, Sage.

Bailey, K. D. (1994). *Typologies and taxonomies: an introduction to classification techniques*, Thousand Oaks, California, Sage.

Everitt, B. S. (1980). *Cluster analysis*, Nueva York, Halster.
Fernández Santana, O. (1991). "El análisis de cluster: aplicación, interpretación y validación", *Papers*, 37: 65-76.
Hair, J. F., Anderson, R. E., Tathan, R. L. y Black, W. C. (1999). *Análisis multivariante*, 5.ª edición, Madrid, Prentice Hall, pp. 491-546.
Martínez Ramos, E. (1984). "Aspectos teóricos del análisis de cluster y aplicación a la caracterización del electorado potencial de un partido", en Sánchez Carrión, J. J (ed.), *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*, Madrid, CIS (Centro de Investigaciones Sociológicas), pp. 165-208.

EJERCICIOS PROPUESTOS

1. La exclusión de la variable "ingresos" puede llevar a desconsiderar la necesidad de tipificar las variables para realizar un análisis de conglomerados *K-medias.* Para comprobarlo, con la misma base de datos se repiten los análisis, con la mitad de la muestra total, excluyendo la variable "ingresos" y analizando, primero, variables *sin estandarizar* (A). Compárense los resultados siguientes con los expuestos en el subapartado 3.5.1.2. Asimismo, compárense con la clasificación obtenida utilizando variables *estandarizadas* (B).

   A) Clasificación de casos con variables no estandarizadas

Centros de los conglomerados finales

	Conglomerado		
	1	2	3
Simpatía marroquí	6,35	5,77	5,63
Leyes inmigración	2,82	2,58	2,47
Ideología política	4,43	4,65	5,09
Sexo	,53	,47	,44
Edad	27,53	49,16	69,29
Núm. inmigrantes	2,09	2,27	2,36
Regularizar inmigrante	,79	,73	,69
Entrada inmigrante	1,85	1,93	2,04
Partido racista	1,38	1,41	1,51
Casar con marroquí	1,49	1,72	1,92
Estudios	2,76	1,82	1,37
Vecino marroquí	1,20	1,27	1,41
Inmigrante delincuente	,48	,59	,71

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3
1		21,668	41,802
2	21,668		20,139
3	41,802	20,139	

Tabla ANOVA

	Conglomerado		*Error*		*F*	*Sig.*
	Media cuadrática	*gl*	*Media cuadrática*	*gl*		
Simpatía marroquí	56,214	2	7,412	1.102	7,585	,001
Leyes inmigración	9,201	2	1,078	860	8,535	,000
Ideología política	29,537	2	3,723	920	7,934	,000
Sexo	,996	2	,249	1.255	4,004	,018
Edad	178.398,600	2	38,352	1.255	4.651,641	,000
Núm. inmigrantes	6,694	2	,386	1.062	17,324	,000
Regularizar inmigrante	,943	2	,188	1.097	5,015	,007
Entrada inmigrante	3,381	2	,405	1.164	8,354	,000
Partido racista	1,625	2	,329	1.115	4,940	,007
Casar con marroquí	18,209	2	,995	1.211	18,304	,000
Estudios	190,468	2	1,131	1.141	168,449	,000
Vecino marroquí	4,321	2	,458	1.242	9,425	,000
Inmigrante delincuente	4,316	2	,238	1.092	18,143	,000

Número de casos en cada conglomerado

Conglomerado	1	536,000
	2	407,000
	3	316,000
Válidos		1.259,000
Perdidos		,000

B) Clasificación de casos con variables estandarizadas

Centros de los conglomerados finales

	Conglomerado		
	1	2	3
Puntuac.: simpatía marroquí	– 1,08625	– ,41673	,42250
Puntuac.: leyes inmigración	– ,86411	– ,59638	,44676
Puntuac.: sexo	,11241	– ,09666	,05160
Puntuac.: edad	,39479	,42831	– ,30472
Puntuac.: núm. inmigrantes	,61075	,60196	– ,47938
Puntuac.: regularizar inmigrante	– ,72789	– ,62936	,42202
Puntuac.: entrada inmigrante	,77207	,51332	– ,37481
Puntuac.: partido racista	,51404	,40280	– ,29830
Puntuac.: casar con marroquí	1,71412	,17625	– ,41013
Puntuac.: estudios	– ,35882	– ,50574	,33880
Puntuac.: vecino marroquí	2,63627	– ,21542	– ,32664
Puntuac.: ideología política	,46186	,25763	– ,24680
Puntuac.: inmigrante delincuente	,53973	,53283	– ,39912

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3
1		3,349	4,913
2	3,349		2,846
3	4,913	2,846	

Anova

	Conglomerado		*Error*		*F*	*Sig.*
	Media cuadrática	*gl*	*Media cuadrática*	*gl*		
Puntuac.: simpatía marroquí	156,833	2	,725	1.102	216,242	,000
Puntuac.: leyes inmigración	133,038	2	,716	860	185,752	,000
Puntuac.: sexo	3,613	2	,997	1.255	3,625	,027
Puntuac.: edad	80,868	2	,853	1.255	94,805	,000
Puntuac.: núm. inmigrantes	152,955	2	,722	1.062	211,728	,000
Puntuac.: regularizar inmigrante	155,154	2	,748	1.097	207,431	,000
Puntuac.: entrada inmigrante	132,708	2	,805	1.164	164,904	,000
Puntuac.: partido racista	71,641	2	,896	1.115	79,954	,000
Puntuac.: casar con marroquí	254,573	2	,585	1.221	434,803	,000
Puntuac.: estudios	90,861	2	,859	1.141	105,722	,000
Puntuac.: vecino marroquí	488,793	2	,243	1.242	2.014,551	,000
Puntuac.: ideología política	37,042	2	,906	920	40,897	,000
Puntuac.: inmigrante delincuente	116,577	2	,790	1.092	147,612	,000

Número de casos en cada conglomerado

Conglomerado	1	134,000
	2	403,000
	3	722,000
Válidos		1.259,000
Perdidos		,000

2. Siguiendo la recomendación de Fernández Santana (1991) de no incluir datos sociodemográficos juntamente con opiniones u otros ítems de naturaleza psicológica (para realizar un correcto análisis de conglomerados), se repite la clasificación de "variables" excluyendo las variables sociodemográficas. Interprétense los resultados siguientes, obtenidos de la aplicación del procedimiento de conglomeración *jerárquica* de *vinculación simple* (o *distancia mínima*). La medida de distancia utilizada es la *correlación de Pearson*. Compárense los resultados con los incluidos en el subapartado 3.5.1. Además, señálese el número de conglomerados que debería formarse. Justifíquese la respuesta.

Matriz de distancia

Caso	*Archivo matricial de entrada*									
	simpatía marroquí	*leyes inmigr.*	*latino-americano*	*n.º inmigr.*	*regular. inmigr.*	*entrada inmigr.*	*partido racista*	*casar marroquí*	*vecino marroquí*	*inmigr. delinc.*
Simpatía marroquí		,337	,525	,311	,302	,295	,240	,470	,329	,323
Leyes inmigración	,337		,201	,380	,341	,413	,107	,311	,322	,322
Simpatía latinoamer.	,525	,201		,169	,188	,259	,185	,221	,213	,204
Núm. inmigrantes	,311	,380	,169		,315	,299	,183	,290	,223	,352
Regularizar inmigrante	,302	,341	,188	,315		,421	,235	,286	,280	,251
Entrada inmigrante	,295	,413	,259	,299	,421		,289	,354	,265	,309
Partido racista	,240	,107	,185	,183	,235	,289		,302	,291	,138
Casar con marroquí	,470	,311	,221	,290	,286	,354	,302		,606	,292
Vecino marroquí	,329	,322	,213	,223	,280	,265	,291	,606		,206
Inmigrante delincuente	,323	,322	,204	,352	,251	,309	,138	,292	,206	

Vinculación simple

Historial de conglomeración

Etapa	*Conglomerado que se combina*		*Coeficientes*	*Etapa en la que el conglomerado aparece por primera vez*		*Próxima etapa*
	Conglom. 1	*Conglom. 2*		*Conglom. 1*	*Conglom. 2*	
1	8	9	,606	0	0	3
2	1	3	,525	0	0	3
3	1	8	,470	2	1	7
4	5	6	,421	0	0	5
5	2	5	,413	0	4	6
6	2	4	,380	5	0	7
7	1	2	,354	3	6	8
8	1	10	,352	7	0	9
9	1	7	,302	8	0	0

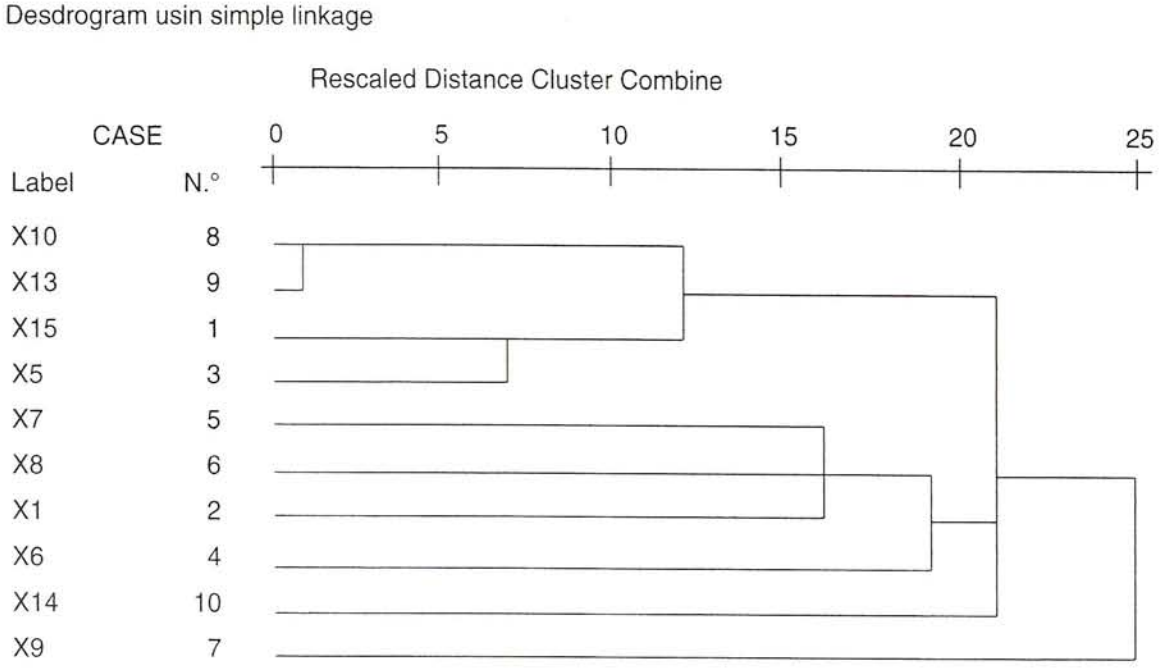
Conglomerado de pertenencia

Caso	4 *conglomerados*	3 *conglomerados*	2 *conglomerados*
Simpatía marroquí	1	1	1
Leyes inmigración	2	1	1
Simpatía latinoamericano	1	1	1
Núm. inmigrantes	2	1	1
Regularizar inmigrante	2	1	1
Entrada inmigrante	2	1	1
Partido racista	3	2	2
Casar con marroquí	1	1	1
Vecino marroquí	1	1	1
Inmigrante delincuente	4	3	1

Diagrama de témpanos vertical

Número de conglomerados	*Caso*																	
	Partido racista		Inmigrante delincuente		Núm. inmigrantes		Entrada inmigrante		Regularizar inmigrante		Leyes inmigración		Vecino marroquí		Casar con marroquí		Simpatía latinoamericano		Simpatía marroquí
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X		X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X		X		X	X	X	X	X	X	X		X	X	X	X	X	X	X
5	X		X		X		X	X	X	X	X		X	X	X	X	X	X	X
6	X		X		X		X	X	X		X		X	X	X	X	X	X	X
7	X		X		X		X		X		X		X	X	X	X	X	X	X
8	X		X		X		X		X		X		X	X	X		X	X	X
9	X		X		X		X		X		X		X	X	X		X		X

Dendograma



3. En la investigación de Baró Llinas, J. *et al.* (1999) "Agrupaciones de las comunidades autónomas según distribución de la ocupación para el año 1997" (*Revista del Ministerio de Trabajo y Asuntos Sociales*, 16: 117-130), se obtuvo una

clasificación de las distintas comunidades autónomas, de acuerdo con las variables edad, sector económico, nivel de estudios, categoría profesional, sexo y tipo de contrato. Analícense las distancias *euclídeas* de cada comunidad autónoma respecto a la media de España.

Comunidades Autónomas	*Distancia euclídea*
Andalucía	7,930265
Aragón	4,745556
Asturias	6,822812
Baleares	5,195898
Canarias	4,718522
Cantabria	5,327134
Castilla-La Mancha	6,655783
Castilla y León	3,667722
Cataluña	5,754900
C. Valenciana	4,053136
Extremadura	10,815473
Galicia	4,911625
Madrid	7,923963
Murcia	4,274389
Navarra	6,958548
País Vasco	7,089011
La Rioja	5,085058