# Capítulo 1

# Introducción al Análisis Cluster. Consideraciones generales.

## 1.1. El problema de la clasificación.

Una de las actividades más primitivas, comunes y básicas del hombre consiste en clasificar objetos en categorías. Las personas, objetos y sucesos encontrados en un día son demasiado numerosos para procesarlos mentalmente como entidades aisladas.

Clasificación o identificación es el proceso o acto de asignar un nuevo objeto u observación en su lugar correspondiente dentro de un conjunto de categorías establecido. Los atributos esenciales de cada categoría son conocidos, aunque haya algunas incertidumbres a la hora de asignar alguna observación dada. Como ejemplo claro, la clasificación se necesita para el desarrollo del lenguaje, el cual consiste en palabras que nos ayudan a reconocer y discutir los diferentes tipos de sucesos, objetos y gentes que nos encontramos. Cada nombre es una etiqueta usada para describir una clase de objetos que poseen notables características en común. Nombrar es clasificar.

Al igual que es una actividad humana conceptual básica, la clasificación es también fundamental en la mayoría de las ramas de la ciencia. En Biología, por ejemplo, la clasificación de organismos ha sido una preocupación desde las primeras investigaciones biológicas. Aristóteles construyó un elaborado sistema de clasificación de especies del reino animal; él empezó dividiendo los animales en dos grupos principales: los que tenían sangre roja (correspondiente a los vertebrados) y los que no la tienen (invertebrados). Además subdividió esos dos grupos de acuerdo a la forma en la que los descendientes venían al mundo (ovíparos y vivíparos). Tras Aristóteles, Teófrates redactó el primer informe fundamental sobre la estructura y clasificación de las plantas. El resultado fue unos libros ampliamente documentados y profundos, abarcando tantos conceptos en sus temas que nos han provisto de la base de las investigaciones biológicas durante muchos siglos. Fueron sustituidos en los siglos XVII y XVIII cuando los grandes exploradores europeos dieron lugar a la segunda búsqueda y colección, bajo la dirección del naturalista sueco Linnaens. Dicho naturalista publicó su trabajo *Genera Plantarum*, en el cual podemos leer:

> *...Todo el conocimiento real que nosotros poseemos depende de los métodos por los cuales distinguimos lo similar de lo no similar. El gran número de diferencias naturales que este método comprende llega a darnos una idea más clara de las cosas...*

En Biología, la teoría y práctica de la clasificación de los organismos es conocida generamente como Taxonomía. Inicialmente, la taxonomía en un sentido más amplio fue, quizás, más un arte que un método científico, pero, eventualmente, fueron desarrolladas técnicas menos subjetivas por Adanson (1727-1806), quien es avalado por Sokal y Sneath (1963) con la introducción del *polithetic*, tipo de sistemas dentro de la Biología en los cuales las clasificaciones se basan en muchas características de los objetos, siendo estudiados por oposición a los sitemas *monothetic*, los cuales usan una única característica para producir una clasificación.

La clasificación de animales y plantas ha jugado un papel importante en el campo de la Biología y de la Zoología, particularmente como una base para la teoría de la evolución de Darwin. Pero la clasificación ha jugado también un papel central en el desarrollo de teorías en otros campos de la ciencia. La clasificación de los elementos en la tabla periódica, por ejemplo, producida por Mendelejev en 1860 causó un profundo impacto en el entendimiento de la estructura del átomo. En Astronomía, la clasificación de las estrellas en *enanas*

y *gigantes* usando el campo Herstsprung-Russell de temperatura frente a luminosidad, afectó fuertemente a las teorías de la evolución de las estrellas.

Las técnicas numéricas para obtener clasificaciones se originaron en las ciencias naturales como la Biología y la Zoología, en un esfuerzo por librar a la Taxonomía de su subjetivismo tradicional y proporcionar clasificaciones objetivas y estables, objetivas en el sentido de que el análisis del mismo conjunto de organismos por diferentes métodos numéricos proporcionen la misma clasificación y estables en el sentido de que la clasificación permanezca igual bajo la inclusión de una gran variedad de organismos o de nuevos caracteres.

La segunda mitad de este siglo ha visto un gran aumento en el número de técnicas numéricas de clasificación disponibles. Este crecimiento ha ido paralelo con el desarrollo de los ordenadores, que son necesarios para poder realizar el gran número de operaciones aritméticas que se precisan. Asimismo, un desarrollo similar ha tenido lugar en las áreas de aplicación. Actualmente tales técnicas son usadas en campos como la arqueología, psiquiatría, astronomía e investigación de mercados.

Una variedad de nombres han sido aplicados a estos métodos, dependiendo del área de aplicación. *Taxonomía Numérica* se usa en Biología. En Psicología se emplea el término *Q-análisis*. En inteligencia artificial se usa el nombre de *Reconocimiento de Patrones*. En otras áreas se emplea *Agrupación* y *agrupamiento*. Actualmente, no obstante, el término más genérico es *Análisis Cluster*. El problema con el que estas técnicas se encuentran puede ser establecido en general como sigue:

> *Dado un conjunto de m objetos individuales (animales, plantas, etc.), cada uno de los cuales viene descrito por un conjunto de n características o variables, deducir una división útil en un número de clases. Tanto el número de clases como las propiedades de dichas clases deben ser determinadas.*

La solución generalmente buscada es una partición de los $m$ objetos, o sea, un conjunto de grupos donde un objeto pertenezca a un grupo sólo y el conjunto de dichos grupos contenga a todos los objetos. Formalmente hablando, se parte de una muestra $\Xi$ de $m$ individuos, $X_1, \ldots, X_m$, cada uno de los cuales está representado por un vector $n$-dimensional, $X_j = (x_{j1}, x_{j2}, \ldots, x_{jn})'$, $j = 1, \ldots, m$ y debemos encontrar una partición de la muestra en regiones $\omega_1, \ldots, \omega_c$ de forma que

$$\bigcup_{i=1}^{c} \omega_i = \Xi$$

$$\omega_i \cap \omega_j = \emptyset \quad ; \quad i \neq j$$

El problema de la clasificación puede ser complicado debido a varios factores, como la presencia de clases definidas de forma imperfecta, la existencia de categorías solapadas y posibles variaciones aleatorias en las observaciones. Una forma de tratar estos problemas, desde el punto de vista estadístico, sería encontrar la probabilidad que tiene cada nueva observación de pertenecer a cada categoría. En este sentido, el criterio de clasificación más simple sería elegir la categoría más probable, mientras que pueden necesitarse reglas más sofisticadas si las categorías no son igualmente probables o si los costos de mala clasificación varían entre las categorías.

## 1.2. El Análisis Cluster.

Análisis Cluster es el nombre genérico de una amplia variedad de procedimientos que pueden ser usados para crear una clasificación. Más concretamente, un método cluster es un procedimiento estadístico multivariante que comienza con un conjunto de datos conteniendo información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que llamaremos *clusters*.

En Análisis Cluster poca o ninguna información es conocida sobre la estructura de las categorías, lo cual lo diferencia de los métodos multivariantes de asignación y discriminación. De todo lo que se dispone es de una colección de observaciones, siendo el objetivo operacional en este caso, descubrir la estructura de las categorías en la que se encajan las observaciones. Más concretamente, el objetivo es ordenar las observaciones en grupos tales que el grado de asociación natural es alto entre los miembros del mismo grupo y bajo entre miembros de grupos diferentes.

Aunque poco o nada se conoce sobre la estructura de las categorías a priori, se tiene con frecuencia algunas nociones sobre características deseables e inaceptables a la hora de establecer un determinado esquema de clasificación. En términos operacionales, el analista es informado suficientemente sobre el problema, de tal forma que puede distinguir entre buenas y malas estructuras de categorías cuando se encuentra con ellas. Entonces, ¿por qué no enumerar todas las posibilidades y elegir la más atrayente?

El número de formas en las que se pueden clasificar $m$ observaciones en $k$ grupos es un número de Stirling de segunda especie (Abramowitz y Stegun, 1968).

$$\mathbb{S}_m^{(k)} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^m$$

El problema se complica aún más por el hecho de que usualmente el número de grupos es desconocido, por lo que el número de posibilidades es suma de números de Stirling; así, por ejemplo, en el caso de $m$ observaciones tendríamos que el número total de posibles clasificaciones sería

$$\sum_{j=1}^{m}\mathbb{S}_m^{(j)}$$

que es un número excesivamente grande, por lo que el número de posibles clasificaciones puede ser enorme (por ejemplo, en el caso de 25 observaciones, se tiene que $\sum_{j=1}^{25}\mathbb{S}_{25}^{(j)} > 4\times 10^{18}$). Así es necesario encontrar una solución aceptable considerando sólo un pequeño número de alternativas.

Los métodos cluster han sido desarrollados a lo largo de este siglo, pero la mayor parte de la literatura sobre Análisis Cluster ha sido escrita durante las pasadas tres décadas. El principal estímulo para el desarrollo de estos métodos fue el libro *Principios de Taxonomía Numérica*, publicado en 1963 por dos biólogos, Sokal y Sneath. Dichos autores argumentan que un procedimiento eficiente para la generación de clasificaciones biológicas debe recoger todos los posibles datos sobre un conjunto de organismos de interés, estimar el grado de similaridad entre esos organismos y usar un método cluster para colocar los organismos similares en un mismo grupo. Una vez que los grupos de organismos similares han sido encontrados, los miembros de cada uno de ellos deben ser analizados para determinar si representan especies biológicas diferentes. En efecto, Sokal y Sneath asumen que el proceso de reconocimiento de patrones debe ser usado como base para comprender el proceso evolutivo.

A partir de ese momento, la literatura sobre Análisis Cluster se desarrolla de forma considerable. Hay dos razones para el rápido crecimiento y desarrollo de este tipo de técnicas:

1. **El desarrollo de los ordenadores.**

   Antes del auge de los ordenadores, los métodos clusters resultaban molestos y dificultosos desde el punto de vista computacional cuando eran aplicados a conjuntos grandes de datos. Por ejemplo, clasificar un conjunto de datos con 200 entidades requiere buscar una matriz de similaridad con 19.900 valores, y trabajar con una matriz de ese tamaño es una tarea costosa en tiempo que muchos investigadores debían emprender. Obviamente, con la difusión de los ordenadores, el proceso de manejo de grandes matrices se vuelve mucho más factible.

2. **La importancia fundamental de la clasificación como un procedimiento científico.**

   Todas las ciencias están construidas sobre clasificaciones que estructuran sus dominios de investigación. Una clasificación contiene los mejores conceptos usados en una ciencia. La clasificación de los elementos, por ejemplo, es la base para comprender la química inorgánica y la teoría atómica de la materia; la clasificación de las enfermedades proporciona la base estructural para la medicina.

A pesar de su popularidad, los métodos cluster están todavía poco comprendidos y desarrollados en comparación con otros procedimientos estadísticos multivariantes como el análisis factorial, análisis discriminante o multidimensional scaling. La literatura en las ciencias sociales sobre cluster refleja una serie desconcertante y con frecuencia contradictoria de terminología, métodos y aproximaciones, lo cual ha creado un complejo mundo que es virtualmente impenetrable.

Como hemos notado, los métodos cluster se han diseñado para crear grupos homogéneos de casos o entidades. La mayor parte de los usos del Análisis Cluster pueden ser resumidos bajo cuatro objetivos principales:

1. Desarrollar una tipología o clasificación.

2. Investigar esquemas conceptuales útiles para agrupar entidades.

3. Generar hipótesis a través de la exploración de los datos.

4. Contrastar hipótesis o intentar determinar si tipos definidos por otros procedimientos están de hecho presentes en un conjunto de datos.

De estos objetivos, la creación de clasificaciones, probablemente, resulta el objetivo más frecuente de los métodos cluster, pero en la mayor parte de los casos muchos de estos objetivos se combinan para formar la base de estudio.

No obstante, hay que tener algunas precauciones sobre los métodos cluster:

1. La mayor parte de los métodos de Análisis Cluster son procedimientos que, en la mayor parte de los casos, no están soportados por un cuerpo de doctrina estadística teórica. En otras palabras, la mayor parte de los métodos son heurísticos. Esto contrasta con otros procedimientos como el Análisis Factorial, por ejemplo, que está basado sobre una extensa teoría estadística.

2. La mayor parte de los métodos clusters han nacido al amparo de ciertas ramas de la ciencia, por lo que, inevitablemente, están impregnados de un cierto sesgo procedente de esas disciplinas. Esta cuestión es importante puesto que cada disciplina tiene sus propias preferencias tales como los tipos de datos a emplear en la construcción de la clasificación. Así puede haber, por ejemplo, métodos que sean útiles en psicología pero no en biología o viceversa.

3. Distintos procedimientos clusters pueden generar soluciones diferentes sobre el mismo conjunto de datos. Una razón para ello radica en el hecho ya comentado de que los métodos clusters se han desarrollado a partir de fuentes dispares que han dado origen a reglas diferentes de formación de grupos. De esta manera, lógicamente, es necesaria la existencia de técnicas que puedan ser usadas para determinar qué método produce los grupos naturalmente más homogéneos en los datos.

## 1.3. Cluster por individuos y por variables.

El punto de partida para el Análisis Cluster es, en general, una matriz $X$ que proporciona los valores de las variables para cada uno de los individuos objeto de estudio, o sea

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

La $i$-ésima fila de la matriz $X$ contiene los valores de cada variable para el $i$-ésimo individuo, mientras que la $j$-ésima columna muestra los valores pertenecientes a la $j$-ésima variable a lo largo de todos los individuos de la muestra.

El objetivo de clasificar los datos, como ya se ha comentado, es agrupar individuos u objetos representados por las filas de $X$. Aparentemente no hay razón para que estos procedimientos no se apliquen a $X'$, obteniéndose así una clasificación de las variables que describen cada individuo. De hecho, muchas de las técnicas cluster existentes (no todas) pueden ser aplicadas para clasificar variables; incluso algunos paquetes estadísticos, como es el caso de BMDP, incluyen implementaciones por separado que permiten realizar análisis cluster por variables (1M) y análisis cluster por individuos (2M).

## 1.4. Clasificación de las técnicas clusters.

La clasificación que vamos a dar está referida a algunas de las distintas técnicas clusters existentes. Como se podrá comprobar, es bastante extensa, ya que múltiples son los métodos existentes. Asimismo hay que hacer notar que no todos los procedimientos mencionados van a ser tratados con posterioridad, sino que trataremos solamente los más usuales en las aplicaciones prácticas, y por ende sobre los que se posee un mayor grado de experiencia, y que suelen ser los normalmente implementados en los paquetes estadísticos existentes, ya que no se debe perder de vista que sin un potente ordenador y programa informático no es factible el desarrollo práctico de ninguna técnica cluster.

A grandes rasgos se distinguen dos grandes categorías de métodos clusters: métodos jerárquicos y métodos no jerárquicos.

### 1.4.1. Métodos Jerárquicos.

Estos métodos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen a su vez en aglomerativos y disociativos. Los aglomerativos comienzan el análisis con tantos grupos como individuos haya en el estudio. A partir de ahí se van formando grupos de forma ascendente, hasta que, al final del proceso, todos los casos están englobados en un mismo conglomerado. Los métodos disociativos o divisivos realizan el proceso inverso al anterior. Empiezan con un conglomerado que engloba a todos los individuos. A partir de este grupo inicial se van formando, a través de sucesivas divisiones, grupos cada vez más pequeños. Al final del proceso se tienen tantos grupos como individuos en la muestra estudiada.

Independientemente del proceso de agrupamiento, hay diversos criterios para ir formando los conglomerados; todos estos criterios se basan en una matriz de distancias o similitudes. Por ejemplo, dentro de los métodos aglomerativos destacan:

1. Método del amalgamamiento simple.
2. Método del amalgamamiento completo.
3. Método del promedio entre grupos.
4. Método del centroide.
5. Método de la mediana.
6. Método de Ward.

Dentro de los métodos disociativos, destacan, además de los anteriores, que siguen siendo válidos:

1. El análisis de asociación.
2. El detector automático de interacción.

### 1.4.2. Métodos no Jerárquicos.

En cuanto a los métodos no jerárquicos, también conocidos como partitivos o de optimización, tienen por objetivo realizar una sola partición de los individuos en $K$ grupos. Ello implica que el investigador debe especificar a priori los grupos que deben ser formados, siendo ésta, posiblemente, la principal diferencia respecto de los métodos jerárquicos, (no obstante hay que señalar que hay diversas versiones de estos procedimientos que flexibilizan un tanto el número final de clusters a obtener). La asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia de estos métodos respecto a los jerárquicos reside en que trabajan con la matriz de datos original y no precisan su conversión en una matriz de distancias o similitudes. Pedret en 1986 agrupa los métodos no jerárquicos en cuatro familias:

1. **Métodos de Reasignación.**

   Permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior, si ello optimiza el criterio de selección. El proceso acaba cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Dentro de estos métodos están:

   *a*) El método $K$-Medias.
   *b*) El Quick-Cluster análisis.
   *c*) El método de Forgy.
   *d*) El método de las nubes dinámicas.

2. **Métodos de búsqueda de la densidad.**

   Dentro de estos métodos están los que proporcionan una aproximación tipológica y una aproximación probabilística.

   En el primer tipo, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre ellos destacan:

*a*) El análisis modal de Wishart.

*b*) El método Taxmap.

*c*) El método de Fortin.

En el segundo tipo se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Entre los métodos de este tipo destaca el método de las combinaciones de Wolf.

3. **Métodos directos.**

   Permiten clasificar simultáneamente a los individuos y a las variables. El algoritmo más conocido dentro de este grupo es el Block-Clustering.

4. **Métodos de reducción de dimensiones.**

   Estos métodos consisten en la búsqueda de unos factores en el espacio de los individuos; cada factor corresponde a un grupo. Se les conoce como Análisis Factorial tipo Q.

- Métodos de Análisis Cluster
  - Jerárquicos
    - Aglomerativos
      - Linkage Simple
      - Linkage Completo
      - Promedio entre grupos
      - Método del Centroide
      - Método de la Mediana
      - Método de Ward
    - Disociativos
      - Linkage Simple
      - Linkage Completo
      - Promedio entre grupos
      - Método del Centroide
      - Método de la Mediana
      - Método de Ward
      - Análisis de Asociación
      - Detector Automático de Interacción
  - No Jerárquicos
    - Métodos de Reasignación
      - Centroides
        - K-Medias
        - Quick-Cluster
        - Método de Forgy
      - Nubes Dinámicas
    - Búsqueda de densidad
      - Aproximación Tipológica
        - Análisis Modal
        - Método Taxmap
        - Método de Fortin
      - Aproximación Probabilística: Método de Wolf
    - Métodos Directos: Block-Clustering
    - Metodos Reductivos: Análisis Factorial tipo Q

## 1.5. Etapas en Análisis Cluster.

Las etapas a seguir en el empleo de una técnica cluster pueden ser resumidas en los siguientes puntos:

1. **Elección de las variables.**

   La elección inicial del conjunto concreto de características usadas para describir a cada individuo constituye un marco de referencia para establecer las agrupaciones o clusters; dicha elección, posiblemente, refleje la opinión del investigador acerca de su propósito de clasificación. Consecuentemente, la primera cuestión a responder sobre la elección de variables es si son relevantes para el tipo de clasificación que

se va buscando. Es importante tener en cuenta que la elección inicial de variables es, en sí misma, una categorización de los datos, para lo cual sólo hay limitadas directrices matemáticas y estadísticas.

La siguiente cuestión que debe considerarse es el número de variables a emplear. En muchas aplicaciones es probable que el investigador se equivoque tomando demasiadas medidas, lo cual puede dar origen a diversos problemas, bien sea a nivel computacional o bien porque dichas variables adicionales oscurezcan la estructura de los grupos.

En muchas aplicaciones las variables que describen los objetos a clasificar no están medidas en las mismas unidades. En efecto, puede haber variables de tipos completamente diferentes, algunas categóricas, otras ordinales e incluso otras que tengan una escala de tipo intervalo.

Es claro que no sería correcto tratar como equivalentes en algún sentido, por ejemplo, el peso medido en kilos, la altura en milímetros y valorar la ansiedad en una escala de cuatro puntos.

Para variables de tipo intervalo, la solución general consiste en tipificar las variables antes del análisis, calculando las desviaciones típicas a partir de todos los individuos. Algunos autores, por ejemplo Fleiss y Zubin (1969), consideran que esta técnica puede tener serias desventajas al diluir las diferencias entre grupos sobre las variables que más discriminen; como alternativa sugieren emplear la desviación estándar entre grupos para tipificar.

Cuando las variables son de tipos diferentes se suele convertir todas las variables en binarias antes de calcular las similaridades. Esta técnica tiene la ventaja de ser muy clarificadora, pero la desventaja de sacrificar información. Una alternativa más atractiva es usar un coeficiente de similaridad que pueda incorporar información de diferentes tipos de variables de una forma sensible, como el propuesto por Gower en 1971 y que después trataremos. Asimismo, para variables mixtas existe la posibilidad de hacer un análisis por separado e intentar sintetizar los resultados a partir de los diferentes estudios.

2. **Elección de la medida de asociación.**

   La mayor parte de los métodos cluster requieren establecer una medida de asociación que permita medir la proximidad de los objetos en estudio. Cuando se realiza un Análisis Cluster de individuos, la proximidad suele venir expresada en términos de distancias, mientras que el Análisis Cluster por variables involucra generalmente medidas del tipo coeficiente de correlación, algunas de las cuales tienen interpretaciones en distintos sentidos mientras que otras son difíciles de describir, dado el carácter subjetivo de las mismas.

   En el capítulo 2 se hace un breve repaso a las medidas de asociación más usuales que suelen emplearse. Destacamos el hecho de estar clasificadas en medidas para variables y para individuos, si bien algunas de ellas pueden considerarse de uso común. La clasificación se ha establecido sobre todo atendiendo a que las prácticas en ordenador se realizarán con el paquete estadístico BMDP, donde existen dos capítulos específicos, uno para Análisis Cluster por variables y otro por individuos, cada uno de los cuales proporciona un conjunto de medidas a poder usar.

   Hay que tener en cuenta, asimismo, la importancia que tienen los tipos de datos a emplear, bien sean éstos categóricos o no. En el capítulo 2 se muestra toda una serie de posibles medidas que abarcan diversas posibilidades según el tipo de datos a utilizar.

3. **Elección de la técnica cluster a emplear en el estudio.**

   Los métodos cluster que se han propuesto y desarrollado en los últimos años son bastante numerosos y muy diversos en cuanto a su concepción, clasificándose, en un primer estado, en jerárquicos y no jerárquicos, distinguiéndose los primeros de los segundos en que las asignaciones de los individuos, hechas por los métodos jerárquicos a los clusters que se van creando permanecen estables durante todo el proceso, no permitiendo reasignaciones posteriores a clusters distintos si hubiera lugar a ello, cuestión que sí es factible en los métodos no jerárquicos. Además, en los métodos jerárquicos, el investigador deberá sacar sus propias conclusiones mientras que en los procedimientos no jerárquicos el número final de clusters está, por lo general, impuesto de antemano, si bien se han desarrollado, dentro de este tipo de métodos, técnicas que permiten una cierta flexibilidad en el número final de clusters, con el fin de evitar posibles perturbaciones en los resultados definitivos.

   Así pues, en algunos problemas prácticos, la elección del método a emplear será relativamente natural, dependiendo, sobre todo, de la naturaleza de los datos usados y de los objetivos finales perseguidos, si bien en otros la elección no será tan clara. Lo que sí es conveniente siempre, a la hora de las aplicaciones prácticas, es no elegir un sólo procedimiento, sino abarcar un amplio abanico de posibilidades y contrastar los resultados obtenidos con cada una de ellas. De este modo, si los resultados finales son parecidos,

podremos obtener unas conclusiones mucho más válidas sobre la estructura natural de los datos. En caso contrario no obtendremos mucha información, si bien grandes diferencias en los resultados obtenidos pueden llevar a plantearnos el hecho de que tal vez los datos con los que se está trabajando no obedezcan a una estructura bien definida.

En los capítulos 3 y 4 desarrollamos los principales métodos cluster existentes, tanto jerárquicos como no jerárquicos.

4. **Validación de los resultados e interpretación de los mismos.**

   Ésta es la última etapa en la secuencia lógica en la que se desarrolla una investigación a través de un método cluster. Sin duda alguna es la más importante, ya que es en ella donde se van a obtener las conclusiones definitivas del estudio.

   Son diversos los métodos propuestos para validar un procedimiento cluster. Por ejemplo, cuando se está trabajando con métodos jerárquicos se plantean dos problemas:

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

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

   El argumento más empleado para responder a la primera pregunta es el empleo del coeficiente de correlación cofenético, propuesto por Sokal y Rohlf en 1962. Dicho coeficiente mide la correlación entre las distancias iniciales, tomadas a partir de los datos originales, y las distancias finales con las cuales los individuos se han unido durante el desarrollo del método. Altos valores de tal coeficiente mostrarán que durante el proceso no ha ocurrido una gran perturbación en lo que concierne a la estructura original de los datos. En cuanto a la segunda pregunta, muchas son las técnicas existentes, algunas de las cuales, las más empleadas a nivel práctico, están recogidas en el capítulo 3.

   En cuanto a los métodos no jerárquicos, las cuestiones anteriores van perdiendo sentido, mientras que los procedimientos empleados para validar los resultados van encaminados al estudio de la homogeneidad de los grupos encontrados durante el desarrollo del método. Algunos autores han propuesto el empleo de técnicas multivariantes como el análisis multivariante de la varianza (MANOVA), o bien (como BMDP incluye) desarrollar múltiples análisis de la varianza (ANOVA) sobre cada variable en cada cluster. Estos procedimientos, evidentemente, plantean serios problemas y no deben ser considerados como definitivos. Una técnica usualmente empleada, de tipo remuestreo, es la de tomar varias submuestras de la muestra original y repetir el análisis sobre cada una. Si tras repetir el análisis sobre ellas se consiguen soluciones aproximadamente iguales, y parecidas a la obtenida con la muestra principal, se puede *intuir* que la solución obtenida puede ser válida, si bien ésto no sería argumento suficiente para adoptar tal decisión. No obstante, este método es más útil empleado de forma inversa, en el sentido de que si las soluciones obtenidas en las diversas submuestras no guardan una cierta similitud, entonces parece evidente que se debiera dudar de la estructura obtenida con la totalidad de la muestra.