ANÁLISIS POR CONGLOMERADOS

Material docente universitario

Asignatura: Análisis de Datos Multivariante

Máster en Salud Pública

ISABEL MARÍA TIMÓN PÉREZ

Departamento de Ciencias Sociosanitarias

Área de Medicina Preventiva y Salud Pública

Universidad de Murcia

Curso académico 2025/2026

AUTORA: ISABEL MARÍA TIMÓN PÉREZ

INSTITUCIÓN: UNIVERSIDAD DE MURCIA

DEPARTAMENTO: CIENCIAS SOCIOSANITARIAS

AÑO: 2026

TIPO DE RECURSO: MATERIAL DOCENTE UNIVERSITARIO

Descripción:
Este documento ha sido elaborado como material docente para la enseñanza de técnicas de análisis multivariante, concretamente análisis por conglomerados, aplicado al ámbito de las Ciencias de la Salud.

Este material ha sido utilizado como recurso docente en la asignatura de Análisis de Datos Multivariante para el Máster en Salud Pública.

ÍNDICE DE CONTENIDOS

1 INTRODUCCIÓN

1.1 EL ANÁLISIS DE DATOS EN EL CONTEXTO DE LA EPIDEMIOLOGÍA Y LA SALUD PÚBLICA

La epidemiología y la salud pública contemporáneas se enfrentan a fenómenos complejos que raramente pueden describirse mediante una única variable. La salud de una población, la organización de los servicios sanitarios o la distribución territorial de la mortalidad dependen simultáneamente de múltiples determinantes biológicos, sociales, ambientales y organizativos. Esta naturaleza multifactorial exige herramientas analíticas capaces de considerar de forma conjunta un conjunto amplio de variables.

En este contexto, el análisis multivariante adquiere un papel central. Frente al análisis univariante o bivariante, que examina variables de forma aislada o en pares, el análisis multivariante permite estudiar simultáneamente la estructura interna de un sistema de variables y las relaciones existentes entre múltiples dimensiones de un fenómeno sanitario.

Dentro del conjunto de técnicas multivariantes, la clasificación ocupa un lugar destacado. Clasificar significa organizar objetos —individuos, territorios, hospitales, áreas sanitarias o cualquier otra unidad de análisis— en grupos de acuerdo con determinados criterios de semejanza. En salud pública, esta tarea no es meramente descriptiva: permite identificar perfiles epidemiológicos diferenciados, detectar desigualdades estructurales o generar hipótesis sobre determinantes subyacentes.

1.2 CLASIFICACIÓN SUPERVISADA Y NO SUPERVISADA

Desde un punto de vista metodológico, las técnicas de clasificación pueden dividirse en dos grandes categorías: supervisadas y no supervisadas.

En la **clasificación supervisada**, los grupos de pertenencia son conocidos previamente. El objetivo del análisis consiste en construir una regla o función que permita asignar nuevos casos a uno de los grupos ya definidos. Técnicas como el análisis discriminante o determinados modelos de regresión logística se encuadran en este ámbito. El investigador dispone de una variable dependiente categórica que define las clases y utiliza un conjunto de variables explicativas para modelizar la separación entre ellas.

Por el contrario, en la **clasificación no supervisada**, no existe una variable que determine previamente los grupos. El objetivo no es predecir una pertenencia conocida, sino descubrir estructuras latentes en los datos. El análisis por conglomerados —o análisis cluster— pertenece a esta segunda categoría. Aquí, los grupos no están definidos a priori; emergen del propio patrón de similitudes y disimilitudes entre los objetos analizados.

La diferencia no es meramente técnica, sino epistemológica. Mientras que en el enfoque supervisado se parte de una estructura conocida para modelizarla, en el enfoque no supervisado se explora la

estructura interna del conjunto de datos sin imponer categorías previas. En salud pública, este enfoque resulta especialmente valioso cuando se pretende identificar perfiles poblacionales o territoriales cuya delimitación no está previamente establecida.

1.3 El problema de la segmentación en salud pública

La segmentación constituye un problema recurrente en epidemiología y gestión sanitaria. Algunos ejemplos ilustrativos incluyen:

- Agrupar áreas sanitarias según indicadores de mortalidad y morbilidad.
- Clasificar hospitales según patrones de utilización de recursos.
- Identificar perfiles de riesgo en poblaciones según factores sociodemográficos y clínicos.
- Segmentar territorios según esperanza de vida y otros indicadores de salud.

En todos estos casos, no se dispone de una clasificación predefinida que permita dividir la población en categorías claras. La pregunta fundamental es si existen agrupaciones naturales que reflejen patrones epidemiológicos diferenciados.

El análisis por conglomerados permite abordar este problema desde una perspectiva estructural. En lugar de imponer categorías arbitrarias, el método identifica grupos de objetos que presentan una elevada similitud interna y una clara diferenciación externa respecto a otros grupos. Este procedimiento puede revelar desigualdades territoriales, perfiles epidemiológicos diferenciados o patrones organizativos relevantes para la planificación sanitaria.

1.4 Definición formal del análisis por conglomerados

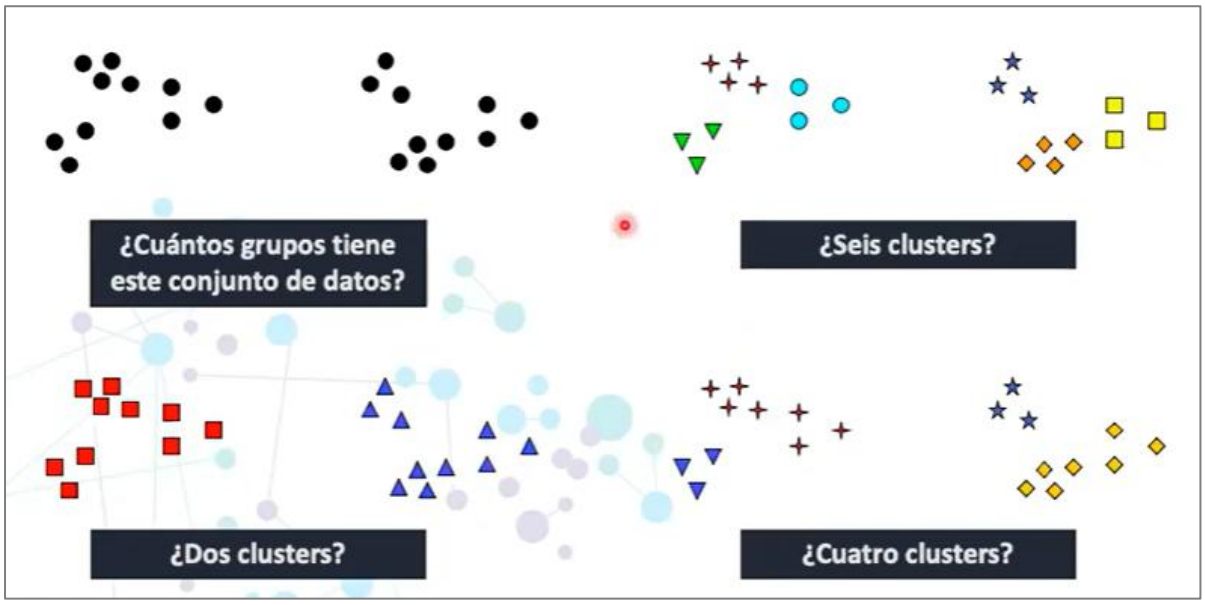
El análisis por conglomerados puede definirse como un conjunto de técnicas destinadas a clasificar un conjunto de n objetos en k grupos (o conglomerados) de modo que:

1. Los objetos pertenecientes a un mismo grupo sean lo más similares posible entre sí (homogeneidad interna).
2. Los objetos pertenecientes a grupos distintos sean lo más diferentes posible entre sí (heterogeneidad externa).

Formalmente, si se dispone de un conjunto de objetos descritos por p variables, cada objeto puede representarse como un punto en un espacio $\mathbb{R}^p$. La similitud entre dos objetos se cuantifica mediante una medida de distancia o disimilitud. El proceso de clustering consiste en particionar el conjunto de puntos de modo que se optimice algún criterio que equilibre cohesión interna y separación entre grupos.

Es importante subrayar el carácter **exploratorio** del método. A diferencia de los modelos inferenciales clásicos, el análisis por conglomerados no parte de una hipótesis paramétrica específica ni produce directamente contrastes de significación estadística. Su objetivo es descubrir estructuras potenciales en los datos que posteriormente puedan ser analizadas o contrastadas mediante otros procedimientos.

Ilustración 1. ¿Qué es un cluster? ¿Cuál es el número óptimo de clusters?



1.5 OBJETIVOS DEL ANÁLISIS POR CONGLOMERADOS

Los objetivos fundamentales del análisis por conglomerados pueden sintetizarse en tres grandes líneas:

1. **Reducción de la complejidad:** Cuando se trabaja con un gran número de objetos y múltiples variables, la estructura del conjunto de datos puede resultar difícil de interpretar. Agrupar los objetos en un número reducido de clusters facilita la comprensión global del fenómeno.
2. **Identificación de estructuras latentes:** El método puede revelar patrones subyacentes que no son evidentes mediante análisis descriptivos simples. Por ejemplo, puede identificar territorios con perfiles epidemiológicos similares, aunque no estén geográficamente próximos.
3. **Generación de hipótesis:** Al identificar conglomerados diferenciados, el investigador puede formular nuevas hipótesis sobre los determinantes que explican dichas agrupaciones, lo que constituye un punto de partida para análisis posteriores de carácter explicativo o causal.

En el ámbito de la Medicina Preventiva, estas funciones adquieren una relevancia particular, ya que la planificación de intervenciones, la asignación de recursos o la evaluación de desigualdades requieren una comprensión estructurada de la heterogeneidad poblacional.

1.6 Papel del análisis por conglomerados en el marco del análisis multivariante

Dentro del conjunto de técnicas multivariantes, el análisis por conglomerados ocupa una posición específica. No se trata de una técnica de reducción de dimensión, como el análisis de componentes principales, ni de un modelo explicativo, como la regresión múltiple. Tampoco es una técnica supervisada de clasificación, como el análisis discriminante.

Su naturaleza es esencialmente descriptiva y estructural. Permite organizar los datos, identificar patrones y segmentar poblaciones sin imponer supuestos fuertes sobre la distribución de las variables o la forma funcional de las relaciones entre ellas.

Desde la perspectiva de la Salud Pública, el estudio del análisis por conglomerados es fundamental por varias razones:

- Proporciona herramientas para la segmentación poblacional.
- Permite interpretar heterogeneidad territorial o institucional.
- Introduce en el razonamiento estructural y exploratorio.
- Complementa técnicas supervisadas como el análisis discriminante.

En definitiva, el análisis por conglomerados constituye una herramienta estratégica en el arsenal metodológico del epidemiólogo, especialmente en fases descriptivas y exploratorias del análisis de datos poblacionales.

Una vez establecido el papel del análisis por conglomerados dentro del análisis multivariante y su relevancia en epidemiología, resulta necesario formalizar sus fundamentos matemáticos, ya que la lógica geométrica subyacente determina completamente el comportamiento de los algoritmos.

2 Fundamento conceptual del análisis por conglomerados

2.1 Representación matemática de los objetos

Sea un conjunto de n objetos (individuos, territorios, hospitales, etc.), cada uno descrito por p variables cuantitativas. Podemos representar el conjunto de datos mediante una matriz:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

donde:

- x_{ij} representa el valor de la variable j para el objeto i,
- $i = 1, \dots, n$,
- $j = 1, \dots, p$.

Cada objeto puede interpretarse como un punto en el espacio euclídeo $\mathbb{R}^p$:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$$

El análisis por conglomerados parte de esta representación geométrica. El problema fundamental consiste en evaluar qué puntos están "cerca" unos de otros en este espacio multidimensional.

2.2 Concepto de similitud y disimilitud

La noción central del análisis por conglomerados es la **similitud** entre objetos. En la práctica, la similitud suele cuantificarse mediante una medida de **disimilitud** o distancia.

Una medida de distancia $d(i, j)$ debe satisfacer, al menos, las propiedades básicas de una métrica:

1. No negatividad:

$$d(i, j) \geq 0$$

2. Identidad del indiscernible:

$$d(i, j) = 0 \iff \mathbf{x}_i = \mathbf{x}_j$$

3. Simetría:

$$d(i, j) = d(j, i)$$

4. Desigualdad triangular:

$$d(i, k) \leq d(i, j) + d(j, k)$$

La elección de la medida de distancia no es un detalle técnico menor: determina la geometría del problema y condiciona la estructura final de los conglomerados.

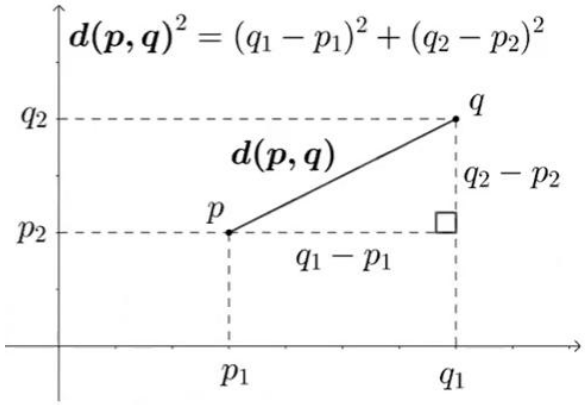
2.3 DISTANCIA EUCLÍDEA

La medida de distancia más utilizada en variables cuantitativas es la **distancia euclídea**, definida como:

$$d_E(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Esta fórmula puede interpretarse como la extensión en p dimensiones del teorema de Pitágoras.

Ilustración 2. Representación gráfica de la Distancia Euclídea con el Teorema de Pitágoras para 2 dimensiones



INTERPRETACIÓN GEOMÉTRICA

La distancia euclídea mide la longitud del segmento que une los puntos $\mathbf{x}_i$ y $\mathbf{x}_j$ en el espacio multidimensional.

En términos epidemiológicos:

- Si dos territorios presentan valores muy similares en todos los indicadores de salud considerados, su distancia será pequeña.
- Si difieren sustancialmente en varias dimensiones, la distancia será grande.

2.4 DISTANCIA EUCLÍDEA AL CUADRADO

En muchos algoritmos de clustering, especialmente en el método de Ward, se utiliza la **distancia euclídea al cuadrado**:

$$d_E^2(i,j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

Desde el punto de vista matemático, elevar al cuadrado elimina la raíz y simplifica cálculos relacionados con varianzas y sumas de cuadrados.

Desde el punto de vista conceptual:

- Se penalizan más intensamente las diferencias grandes.
- Se mantiene la misma ordenación relativa de distancias.

Esta forma es coherente con criterios basados en la minimización de la varianza intracluster.

2.5 La matriz de distancias

El primer paso práctico en muchos métodos de clustering consiste en construir la **matriz de distancias**:

$$\mathbf{D} = \begin{pmatrix} 0 & d(1,2) & \dots & d(1,n) \\ d(2,1) & 0 & \dots & d(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

Se trata de una matriz simétrica de dimensión $n \times n$ que contiene todas las distancias pareadas entre objetos.

En términos prácticos:

- Esta matriz resume toda la información necesaria para los métodos jerárquicos.
- La estructura de los conglomerados depende directamente de esta matriz.

2.6 El problema de la escala: necesidad de estandarización

Un aspecto crítico del análisis por conglomerados es la **sensibilidad a la escala de las variables**.

Supongamos que analizamos:

- Esperanza de vida (rango aproximado 60–85 años),
- Producto Nacional Bruto per cápita (rango miles o decenas de miles),
- Tasa de mortalidad (valores pequeños).

En la distancia euclídea:

$$(x_{ik} - x_{jk})^2$$

las variables con mayor varianza o mayor escala numérica contribuirán de forma desproporcionada al valor total de la distancia.

ESTANDARIZACIÓN MEDIANTE TIPIFICACIÓN

La transformación más habitual consiste en convertir cada variable en una variable tipificada:

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$$

donde:

- $\bar{x}_k$es la media de la variable k,
- $s_k$es su desviación típica.

Tras esta transformación:

- Cada variable tiene media 0.
- Cada variable tiene desviación típica 1.
- Todas las variables contribuyen de forma comparable a la distancia.

Desde un punto de vista conceptual:

La estandarización evita que la estructura de los conglomerados esté determinada artificialmente por la escala de medición.

2.7 VARIABLES BINARIAS Y OTRAS MEDIDAS DE DISTANCIA

En el caso de variables binarias (por ejemplo, presencia/ausencia de enfermedad), la distancia euclídea puede no ser la más apropiada.

Si $x_{ik} \in \{0,1\}$, la varianza está dada por:

$$\mathrm{Var}(X_k) = p_k(1 - p_k)$$

donde p_k es la proporción de unos.

Sin embargo, la coincidencia en ceros (ausencia conjunta) puede no tener el mismo significado epidemiológico que la coincidencia en unos (presencia conjunta).

Por ello se emplean medidas específicas como el índice de Jaccard:

$$J(i,j) = \frac{a}{a + b + c}$$

donde:

- a: número de coincidencias en presencia (1–1),

- b: número de (1–0),
- c: número de (0–1).

Esta medida ignora las coincidencias en ausencia (0–0), lo cual resulta relevante en contextos clínicos donde la presencia de un factor tiene mayor significado que su ausencia.

Para datos mixtos (continuos y categóricos), puede utilizarse la distancia de Gower, que combina contribuciones estandarizadas de distintas naturalezas de variables.

2.8 Criterio general de partición

En términos generales, el análisis por conglomerados busca una partición del conjunto de objetos en k grupos:

$$\{C_1, C_2, \dots, C_k\}$$

tal que se optimice algún criterio de cohesión interna y separación externa.

En formulaciones basadas en varianza, el objetivo puede expresarse como la minimización de la suma de cuadrados dentro de los grupos:

$$W(k) = \sum_{r=1}^{k} \sum_{i \in C_r} \| \mathbf{x}_i - \bar{\mathbf{x}}_r \|^2$$

donde:

- $\bar{\mathbf{x}}_r$ es el centroide del cluster C_r.

Esta expresión será clave al estudiar el método de Ward y el método de K-medias.

A partir de la definición de distancia y del criterio general de partición, podemos abordar los distintos procedimientos algorítmicos que permiten construir los conglomerados.

3 Métodos jerárquicos: Dendograma

Los métodos jerárquicos constituyen una de las aproximaciones más clásicas y conceptualmente transparentes del análisis por conglomerados. Su principal característica es que construyen una estructura anidada de agrupaciones, en la que los conglomerados formados en etapas iniciales quedan contenidos dentro de conglomerados de mayor tamaño en etapas posteriores.

A diferencia de los métodos particionales, en los que el número de grupos debe fijarse previamente, los métodos jerárquicos permiten explorar simultáneamente todas las posibles soluciones, desde n conglomerados (cada objeto constituye su propio grupo) hasta un único conglomerado que incluye a todos los objetos.

3.1 Estructura general del algoritmo aglomerativo

Existen dos grandes familias de métodos jerárquicos:

- Métodos **aglomerativos**, que parten de n grupos individuales y los van fusionando progresivamente.
- Métodos **divisivos**, que parten de un único grupo y lo subdividen sucesivamente.

En la práctica aplicada en salud pública, los métodos aglomerativos son los más utilizados, por su implementación sencilla y su disponibilidad en software estadístico.

El procedimiento aglomerativo puede describirse formalmente del siguiente modo:

1. Se parte de n clusters iniciales, cada uno formado por un único objeto.
2. Se calcula la matriz de distancias $\mathbf{D}$.
3. Se identifican los dos clusters cuya distancia sea mínima.
4. Se fusionan dichos clusters en uno nuevo.
5. Se recalculan las distancias entre el nuevo cluster y los restantes.
6. Se repiten los pasos 3–5 hasta que todos los objetos queden agrupados en un único cluster.

Tras $n - 1$ fusiones, el proceso concluye.

Lo que distingue a los distintos métodos jerárquicos no es la lógica general del algoritmo, sino la manera en que se define la distancia entre clusters.

3.2 Distancia entre clusters: criterios de vinculación

Cuando los clusters contienen más de un objeto, surge la pregunta clave:

¿Cómo se define la distancia entre dos grupos de objetos?

Existen diversos criterios de vinculación (linkage criteria), cada uno con propiedades geométricas distintas.

3.2.1 Vinculación por vecino más próximo (*Single Linkage - MIN*)

La distancia entre dos clusters A y B se define como:

$$d(A,B) = \min_{i \in A,\, j \in B} d(i,j)$$

Es decir, la menor distancia entre cualquier par de objetos pertenecientes a ambos grupos.

Interpretación:

- Los clusters se fusionan si existe al menos un par de puntos muy próximos.
- Tiende a producir estructuras alargadas (efecto "cadena").

Limitación:

- Puede generar conglomerados poco compactos.
- Sensible a puntos intermedios que actúan como "puentes".

3.2.2 Vinculación por vecino más lejano (*Complete Linkage - MAX*)

La distancia entre clusters se define como:

$$d(A,B) = \max_{i \in A,\, j \in B} d(i,j)$$

Se considera la mayor distancia entre cualquier par de objetos de ambos grupos.

Interpretación:

- Garantiza que todos los puntos dentro del cluster estén relativamente próximos.
- Tiende a generar grupos más compactos que single linkage.

Limitación:

- Puede fragmentar grupos naturales si existen valores extremos.

3.2.3 Vinculación intergrupo (*Average Linkage – Group Average*)

La distancia entre clusters es la media de todas las distancias pareadas:

$$d(A,B) = \frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} d(i,j)$$

Interpretación:

- Compromiso entre single y complete linkage.
- Produce conglomerados relativamente equilibrados.

3.2.4 Método de centroides

Se define la distancia entre clusters como la distancia entre sus centroides:

$$d(A,B) = \| \bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B \|$$

donde:

$$\bar{\mathbf{x}}_A = \frac{1}{|A|}\sum_{i \in A} \mathbf{x}_i$$

Limitación:

- Puede producir inversiones en el dendrograma (problema de reversión).

3.3 Método de Ward

El método de Ward ocupa una posición especial dentro de los métodos jerárquicos, ya que no se basa directamente en una distancia geométrica entre puntos extremos, sino en un criterio de minimización de la varianza intracluster.

3.3.1 Fundamento matemático

Recordemos que en el apartado anterior definimos la suma de cuadrados dentro de los grupos como:

$$W(k) = \sum_{r=1}^{k} \sum_{i \in C_r} \| \mathbf{x}_i - \bar{\mathbf{x}}_r \|^2$$

El método de Ward selecciona en cada paso la fusión de dos clusters que produzca el menor incremento en $W(k)$.

Es decir, entre todas las posibles fusiones, se elige aquella que minimiza:

$$\Delta W = W_{\text{nuevo}} - W_{\text{actual}}$$

Interpretación conceptual:

- Se fusionan los dos grupos cuya unión altera menos la homogeneidad interna.
- Se busca mantener clusters compactos y de varianza reducida.

3.3.2 Relación con la distancia euclídea al cuadrado

En la práctica, el método de Ward se implementa utilizando la distancia euclídea al cuadrado, ya que la suma de cuadrados se basa en diferencias cuadráticas.

Si A y B son dos clusters con tamaños n_A y n_B, el incremento de varianza al fusionarlos puede expresarse como:

$$\Delta W = \frac{n_A n_B}{n_A + n_B} \parallel \bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B \parallel^2$$

Esta expresión muestra que:

- El incremento depende de la distancia entre centroides.
- También depende del tamaño de los clusters.

Esta característica introduce un componente estructural importante: el método penaliza la unión de grupos grandes muy distantes.

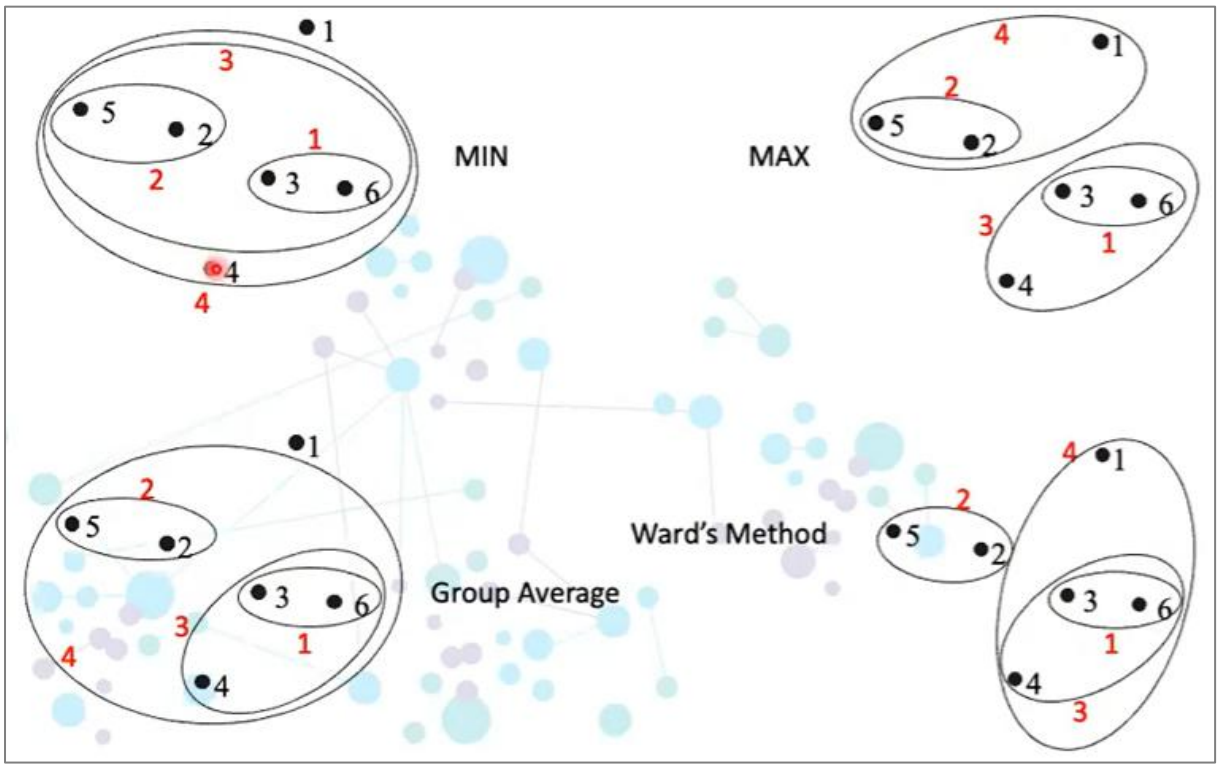
3.3.3 Interpretación epidemiológica

En estudios de salud pública, el método de Ward resulta especialmente útil porque:

- Genera conglomerados internamente homogéneos.
- Tiende a producir grupos relativamente equilibrados en tamaño.
- Se fundamenta en un criterio de minimización de variabilidad.

Cuando se agrupan territorios según indicadores sanitarios, este enfoque permite identificar perfiles claramente diferenciados sin que un único punto extremo determine la estructura.

Ilustración 3. Comparación de clustering según la medida utilizada



3.4 REPRESENTACIÓN DEL PROCESO: EL DENDROGRAMA

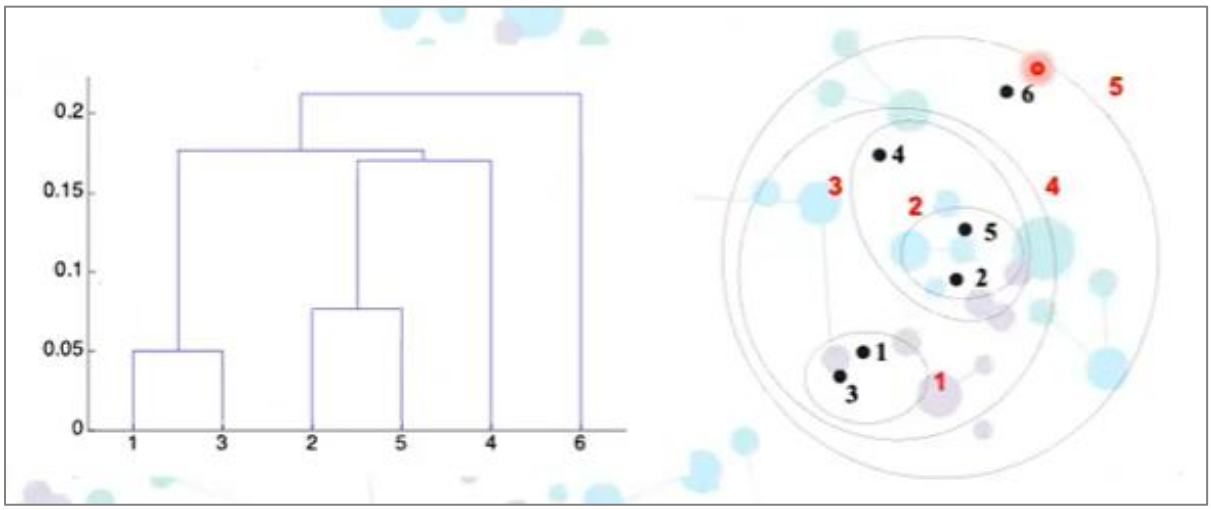
El resultado del proceso jerárquico se representa mediante un **dendrograma**, que constituye una representación gráfica del árbol de fusiones.

En el dendrograma:

- El eje vertical representa los objetos.
- El eje horizontal representa la distancia o el incremento de varianza en el momento de la fusión.

Cada unión se representa mediante una estructura en forma de "U" cuya altura indica la disimilitud entre los grupos fusionados.

Ilustración 4. Correspondencia entre la estructura del dendrograma y la proximidad geométrica de los objetos.



3.5 PROPIEDADES Y LIMITACIONES DE LOS MÉTODOS JERÁRQUICOS

VENTAJAS

- No requieren fijar previamente el número de clusters.
- Permiten explorar múltiples soluciones.
- Proporcionan una representación gráfica intuitiva.
- Son adecuados para análisis exploratorios.

LIMITACIONES

- Una vez realizada una fusión, no puede deshacerse (irreversibilidad).
- Sensibles a la elección de la medida de distancia.
- Complejidad computacional elevada cuando n es grande.
- Sensibles a valores atípicos.

4 MÉTODOS PARTICIONALES: EL ALGORITMO K-MEDIAS

Mientras que los métodos jerárquicos construyen una estructura de agrupaciones anidadas a partir de la matriz de distancias, los **métodos particionales** persiguen directamente una partición del conjunto de objetos en un número fijo de grupos. El algoritmo más representativo de esta familia es el método de **K-medias (K-means)**.

A diferencia del enfoque jerárquico, el número de clusters k debe fijarse previamente. El problema consiste entonces en encontrar la partición que optimice un criterio de homogeneidad interna.

4.1 Formulación matemática

Sea un conjunto de n objetos representados en $\mathbb{R}^p$, y sea k el número de clusters prefijado.

El objetivo del método K-medias es minimizar la suma de cuadrados dentro de los grupos:

$$W(k) = \sum_{r=1}^{k} \sum_{i \in C_r} \| \mathbf{x}_i - \bar{\mathbf{x}}_r \|^2$$

donde:

- C_r representa el conjunto de objetos del cluster r,
- $\bar{\mathbf{x}}_r$ es el centroide del cluster r,
- $\|\cdot\|$ es la norma euclídea.

Este criterio coincide formalmente con el utilizado por el método de Ward, con la diferencia de que en K-medias el número de clusters está fijado desde el inicio y el procedimiento no es jerárquico.

Interpretación conceptual:

K-medias busca la partición que minimice la variabilidad interna de los grupos.

4.2 Descripción del algoritmo

El algoritmo K-medias es iterativo y puede describirse en cuatro etapas fundamentales:

PASO 1. INICIALIZACIÓN

Se seleccionan k puntos iniciales que actuarán como centroides provisionales.

Estos pueden elegirse aleatoriamente entre los objetos o mediante procedimientos más sofisticados (por ejemplo, K-means++).

PASO 2. ASIGNACIÓN

Cada objeto $\mathbf{x}_i$ se asigna al cluster cuyo centroide esté más próximo, generalmente según la distancia euclídea:

Asignar $\mathbf{x}_i$ a C_r si $\| \mathbf{x}_i - \bar{\mathbf{x}}_r \|$ es mínima.

PASO 3. RECALCULADO DE CENTROIDES

Para cada cluster C_r, se recalcula el centroide como la media de los objetos asignados:

$$\bar{\mathbf{x}}_r = \frac{1}{|\ C_r\ |} \sum_{i \in C_r} \mathbf{x}_i$$

PASO 4. ITERACIÓN

Se repiten los pasos 2 y 3 hasta que no haya cambios en la asignación o la reducción de $W(k)$ sea insignificante.

El algoritmo converge en un número finito de pasos, ya que en cada iteración la función objetivo $W(k)$ disminuye o permanece constante.

4.3 PROPIEDADES DEL MÉTODO

El método K-medias presenta varias propiedades relevantes:

- **Convergencia a mínimos locales:** El algoritmo no garantiza encontrar el mínimo global de $W(k)$. La solución depende de la inicialización.
- **Sensibilidad a outliers:** Al basarse en medias, valores extremos pueden desplazar significativamente los centroides.
- **Necesidad de variables métricas:** El método requiere variables cuantitativas y es coherente con la distancia euclídea.
- **Escalabilidad:** Es computacionalmente más eficiente que los métodos jerárquicos cuando n es grande.

4.4 COMPARACIÓN CON MÉTODOS JERÁRQUICOS

Desde una perspectiva conceptual:

- El método de Ward puede entenderse como una aproximación jerárquica al mismo criterio de minimización de la suma de cuadrados que utiliza K-medias.
- Ambos comparten una base en la variabilidad intracluster.

Sin embargo, presentan diferencias estructurales:

Tabla 1. Comparación entre métodos jerárquicos y K-Means

Métodos jerárquicos	K-medias
No requieren fijar $\boldsymbol{k}$ previamente	Requieren fijar k
Generan dendrograma	No generan estructura jerárquica
Irreversibles	Permiten reasignaciones iterativas
Más costosos computacionalmente	Más eficientes para grandes muestras

En contextos de salud pública exploratoria, los métodos jerárquicos resultan especialmente útiles cuando se desea visualizar la estructura de los datos y explorar distintas soluciones. Por esta razón, en el presente desarrollo se prioriza el enfoque jerárquico como herramienta principal, utilizando K-medias como referencia conceptual complementaria.

5 DETERMINACIÓN DEL NÚMERO ÓPTIMO DE CONGLOMERADOS

Con independencia del procedimiento utilizado, jerárquico o particional, surge una cuestión metodológica central: determinar cuántos conglomerados deben conservarse.

Uno de los problemas centrales del análisis por conglomerados no es únicamente cómo construir los grupos, sino decidir cuántos deben conservarse. A diferencia de técnicas inferenciales clásicas, no existe una regla universal que determine el número "correcto" de clusters. La decisión requiere combinar criterios estadísticos, interpretación sustantiva y coherencia metodológica.

Desde un punto de vista formal, el problema puede enunciarse del siguiente modo: dado un conjunto de n objetos y un procedimiento de agrupación, determinar el valor de k que proporciona la mejor partición del conjunto.

5.1 CRITERIO VISUAL EN MÉTODOS JERÁRQUICOS

En los métodos jerárquicos, el dendrograma constituye la herramienta inicial para evaluar el número de conglomerados.

Recordemos que en el dendrograma:

- Cada fusión se representa mediante una estructura en forma de "U".
- La altura de la unión indica la distancia o el incremento de varianza asociado a la fusión.

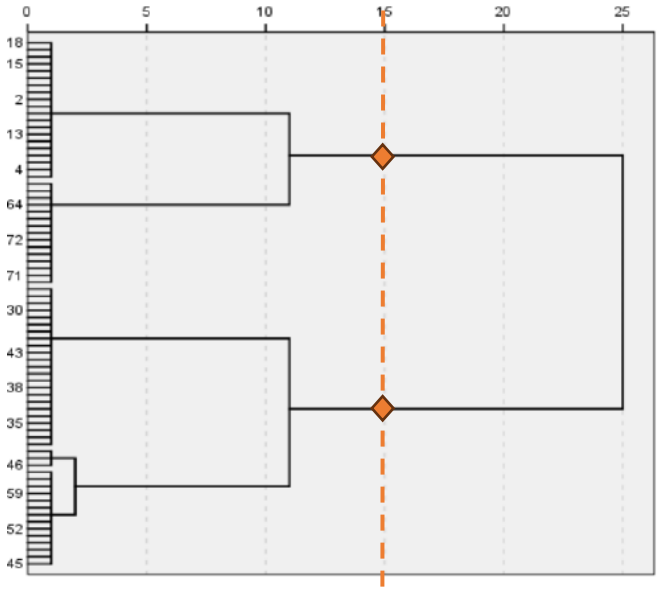
PRINCIPIO DEL "SALTO GRANDE"

La idea central consiste en identificar un incremento notable en la distancia de fusión. Cuando dos grupos se fusionan a una distancia considerablemente mayor que las anteriores, puede interpretarse que se están uniendo estructuras ya claramente diferenciadas.

Formalmente, si $d_1 \leq d_2 \leq \cdots \leq d_{n-1}$ representan las distancias de fusión ordenadas, un aumento abrupto entre d_r y d_{r+1} puede sugerir que la solución adecuada se sitúa en $k = n - r$ clusters.

No obstante, este criterio es inherentemente subjetivo y depende de la escala de las distancias.

Ilustración 5. Dendrograma con punto de corte basado en incremento abrupto de la distancia de fusión.



5.2 CRITERIO DE LA SUMA DE CUADRADOS: EL MÉTODO DEL "CODO"

En métodos particionales como K-medias, puede representarse la función:

$$W(k) = \sum_{r=1}^{k} \sum_{i \in C_r} \| \mathbf{x}_i - \bar{\mathbf{x}}_r \|^2$$

Dado que al aumentar k la suma de cuadrados intracluster disminuye monótonamente, se busca un punto a partir del cual la reducción adicional es marginal.

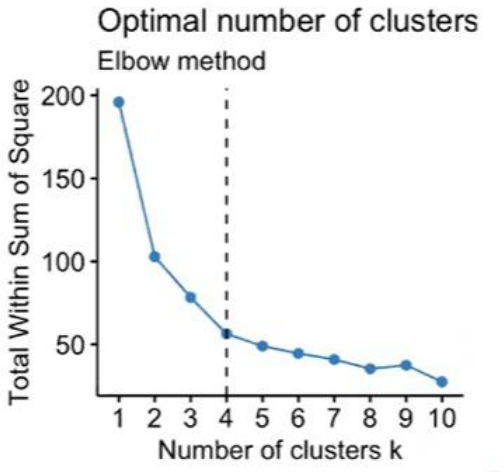
Si se representa $W(k)$ frente a k, suele observarse una curva decreciente. El punto en el que la pendiente cambia de forma notable se denomina "codo".

Interpretación conceptual:

Se busca un equilibrio entre simplicidad (menos clusters) y homogeneidad interna (menor variabilidad).

Este criterio, aunque útil, también tiene un componente interpretativo y no siempre produce un punto claramente definido.

Ilustración 6. Determinación del número óptimo de conglomerados mediante el método del "codo".



5.3 Criterio de Calinski–Harabasz

Para introducir mayor objetividad en la elección de k, pueden utilizarse índices basados en la comparación entre variabilidad entre clusters y variabilidad dentro de clusters.

El índice de Calinski–Harabasz (CH) se define como:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

donde:

- $B(k)$ es la suma de cuadrados entre clusters,
- $W(k)$ es la suma de cuadrados dentro de clusters,
- n es el número total de objetos,
- k es el número de clusters.

DEFINICIONES FORMALES

La suma de cuadrados total puede descomponerse como:

$$T = B(k) + W(k)$$

donde:

$$B(k) = \sum_{r=1}^{k} n_r \parallel \bar{\mathbf{x}}_r - \bar{\mathbf{x}} \parallel^2$$

y

$$W(k) = \sum_{r=1}^{k} \sum_{i \in C_r} \parallel \mathbf{x}_i - \bar{\mathbf{x}}_r \parallel^2$$

siendo:

- $\bar{\mathbf{x}}$ el vector media global,
- n_r el tamaño del cluster r.

Obsérvese que el índice aumenta cuando la variabilidad entre grupos crece y la variabilidad intragrupo disminuye.

INTERPRETACIÓN ESTADÍSTICA

El índice CH es análogo a un estadístico F en un análisis de la varianza:

- El numerador mide la variabilidad entre grupos.
- El denominador mide la variabilidad dentro de los grupos.

Un valor elevado de $CH(k)$ indica:

- Alta separación entre clusters.
- Alta cohesión interna.

En la práctica, se calcula el índice para distintos valores de k y se selecciona aquel que maximiza $CH(k)$.

Ilustración 7. Selección del número óptimo de clusters mediante el índice de Calinski–Harabasz (máximo).

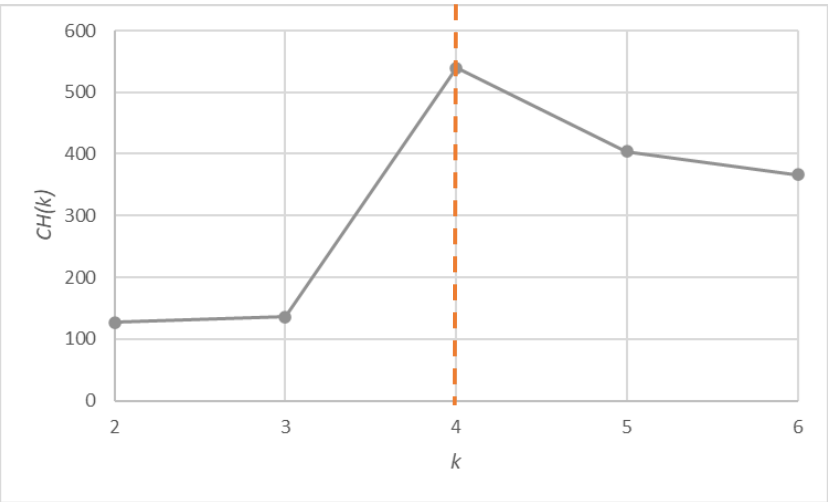
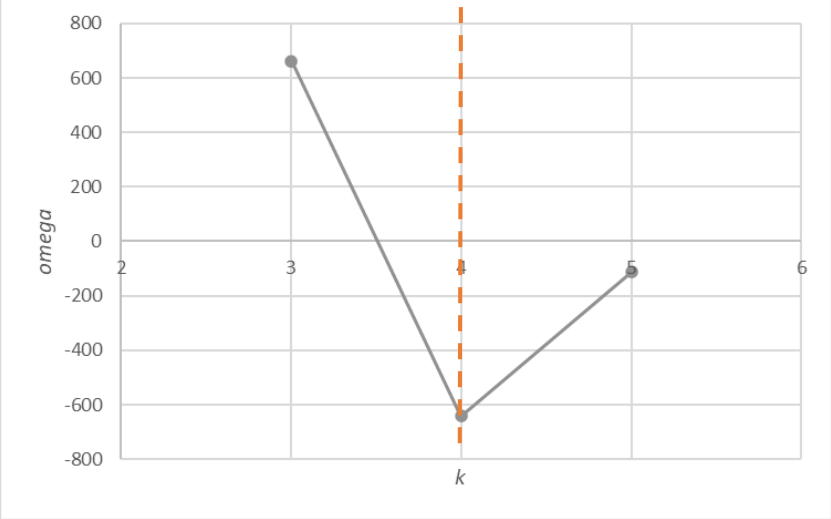


Ilustración 8. Selección del número óptimo de clusters mediante el índice omega (mínimo).



5.4 Interpretación práctica en salud pública

En el análisis de territorios según indicadores sanitarios, la elección de k no debe basarse exclusivamente en un índice numérico. Es necesario considerar:

- Coherencia epidemiológica de los grupos.
- Tamaño razonable de los clusters.
- Interpretabilidad sustantiva.

- Estabilidad de la solución ante pequeñas modificaciones.

En contextos aplicados, la combinación de:

1. Inspección del dendrograma,
2. Criterios cuantitativos como Calinski–Harabasz,
3. Juicio epidemiológico,

proporciona una base sólida para justificar la elección final.

5.5 Limitaciones en la determinación de k

- No existe un número "verdadero" universal de clusters.
- Diferentes índices pueden sugerir valores distintos.
- La solución puede depender de la escala y de la medida de distancia.
- En datos con estructura continua, cualquier partición puede ser artificial.

En consecuencia, el análisis por conglomerados debe entenderse como una herramienta exploratoria cuyo resultado requiere interpretación crítica.

6 Interpretación y perfilado de los conglomerados

Una vez determinada la partición del conjunto de objetos en k clusters, el análisis no concluye con la mera asignación de etiquetas. El verdadero valor del análisis por conglomerados reside en la interpretación sustantiva de los grupos obtenidos.

Desde un punto de vista metodológico, el clustering es una técnica estructural; desde un punto de vista epidemiológico, debe convertirse en una herramienta explicativa y generadora de hipótesis.

6.1 Variable de pertenencia al cluster

El resultado operativo del análisis es una nueva variable categórica:

$$G_i \in \{1,2,\dots,k\}$$

que indica la pertenencia del objeto i al cluster correspondiente.

Esta variable permite:

- Comparar sistemáticamente los grupos.
- Realizar análisis descriptivos estratificados.
- Representar gráficamente diferencias entre clusters.

En términos prácticos, esta fase marca la transición entre el análisis estructural y la caracterización sustantiva.

6.2 DESCRIPCIÓN ESTADÍSTICA DE LOS CONGLOMERADOS

El primer paso en el perfilado consiste en describir cada cluster mediante estadísticas resumen de las variables originales.

Para cada cluster C_r, pueden calcularse:

- Media:

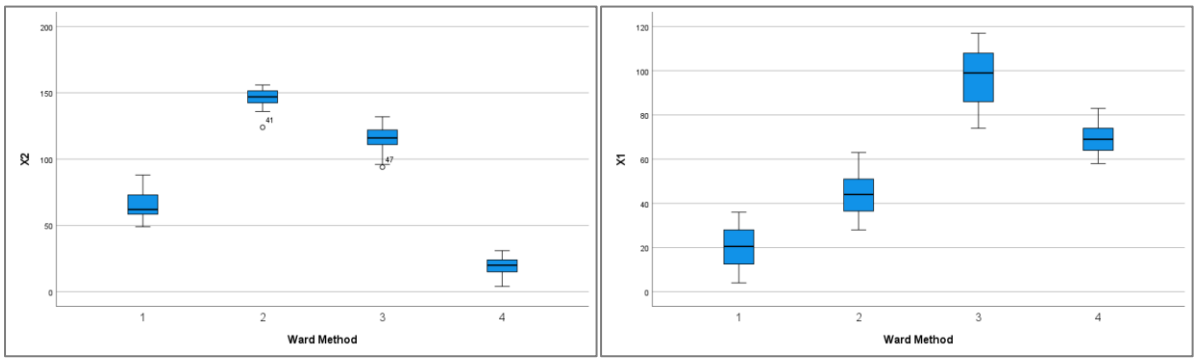
$$\bar{x}_{rk} = \frac{1}{n_r} \sum_{i \in C_r} x_{ik}$$

- Mediana (especialmente útil en presencia de asimetrías).
- Desviación típica o rango intercuartílico.
- Distribución completa mediante gráficos (boxplots).

El objetivo es identificar qué variables diferencian claramente los clusters.

Ilustración 9. Perfil descriptivo de los conglomerados (Ruspini, k = 4): medias y distribución de X1 y X2.

Cluster	Media X_1	Media X_2
1	20,15	64,95
2	43,91	146,04
3	98,18	114,88
4	68,93	19,40



6.3 Interpretación geométrica

Desde la perspectiva matemática, cada cluster puede interpretarse como una región del espacio $\mathbb{R}^p$ centrada en su centroide $\bar{\mathbf{x}}_r$.

La distancia media al centroide refleja la cohesión interna:

$$\frac{1}{n_r}\sum_{i\in C_r} \| \mathbf{x}_i - \bar{\mathbf{x}}_r \|$$

Clusters con baja dispersión interna indican alta homogeneidad; clusters con elevada dispersión pueden sugerir subestructuras no captadas o una partición insuficiente.

6.4 Interpretación epidemiológica

La interpretación epidemiológica implica traducir diferencias estadísticas en perfiles sustantivos.

Por ejemplo, en un análisis territorial de esperanza de vida, podrían identificarse:

- Un cluster con valores sistemáticamente altos en ambos sexos.
- Un cluster con valores intermedios pero elevada desigualdad entre sexos.
- Un cluster con valores sistemáticamente bajos.

La interpretación no debe limitarse a describir diferencias numéricas; debe buscar:

- Patrones regionales.
- Posibles determinantes estructurales.
- Implicaciones para políticas públicas.

En este sentido, el análisis por conglomerados no proporciona causalidad, pero sí segmentación estructural.

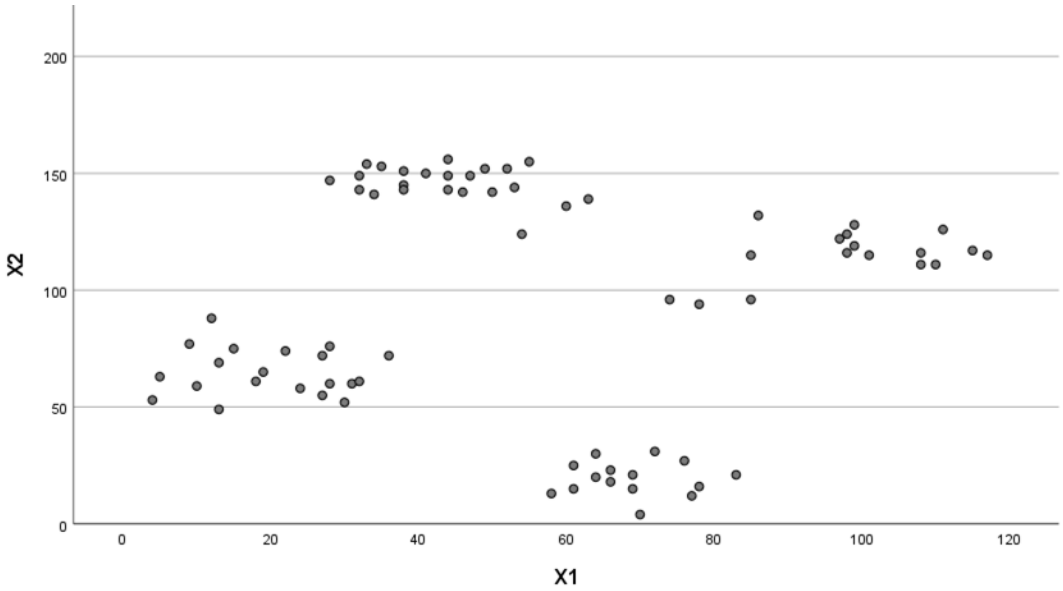
6.5 Representación gráfica y espacial

La representación gráfica es esencial para facilitar la interpretación:

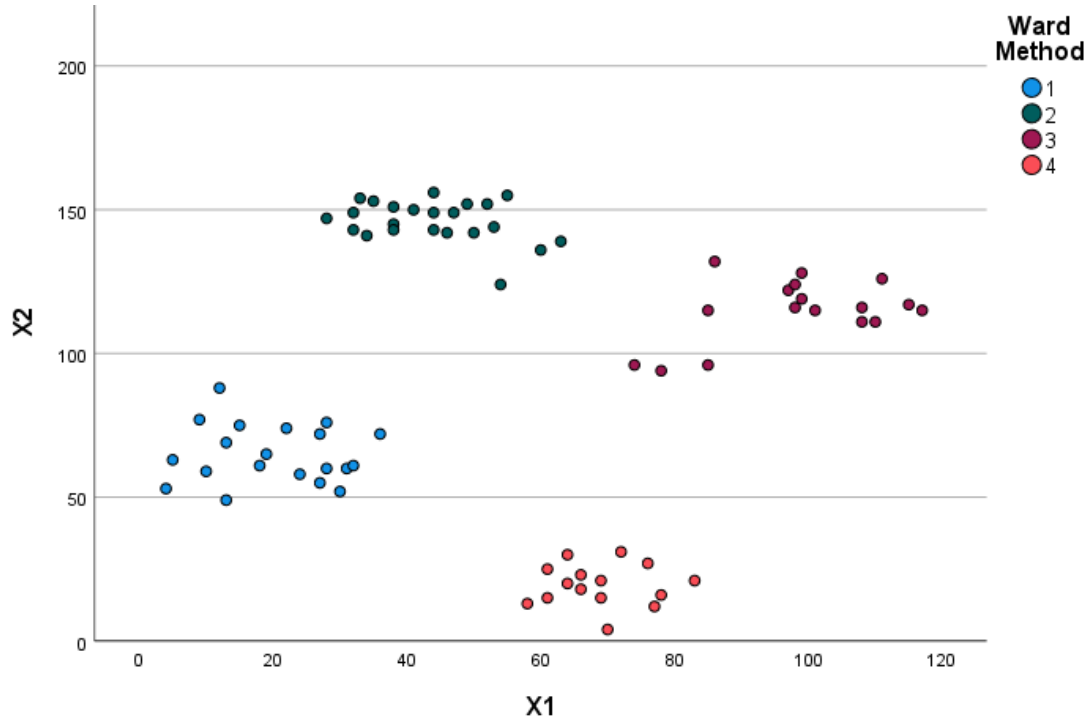
1. **Gráficos comparativos por cluster:** Permiten visualizar diferencias sistemáticas entre grupos.
2. **Mapas temáticos (cuando los objetos son territorios):** La representación espacial puede revelar patrones geográficos coherentes.

Ilustración 10. Distribución original y segmentación por conglomerados del conjunto Ruspini (método de Ward, k = 4).

(a) Distribución original



(b) Clasificación por conglomerado



Este paso es especialmente relevante en Medicina Preventiva, donde la dimensión territorial es central.

6.6 Evaluación crítica de la solución

Una interpretación rigurosa exige plantearse:

- ¿Los clusters son estables?
- ¿Existen clusters de tamaño muy reducido?
- ¿La solución depende excesivamente de un valor extremo?
- ¿La diferenciación es clínicamente relevante o meramente estadística?

Además, debe evitarse la sobreinterpretación:

- El clustering identifica estructuras, pero no prueba que los grupos sean entidades naturales discretas.
- En datos con gradientes continuos, cualquier partición es una simplificación.

6.7 Generación de hipótesis

El perfilado de clusters puede constituir el punto de partida para análisis posteriores:

- Modelos explicativos (regresión).
- Comparaciones formales entre grupos.
- Estudios de desigualdad.
- Evaluación de políticas públicas diferenciadas por perfil territorial.

Así, el análisis por conglomerados se integra dentro de una estrategia analítica más amplia.

Tras la construcción e interpretación de la solución, resulta imprescindible adoptar una perspectiva crítica que permita evaluar el alcance y las limitaciones del procedimiento.

7 Limitaciones generales del análisis por conglomerados

El análisis por conglomerados constituye una herramienta potente de exploración estructural; sin embargo, su utilización exige una comprensión clara de sus limitaciones metodológicas y conceptuales. Una interpretación acrítica de los resultados puede conducir a conclusiones simplificadoras o incluso erróneas.

7.1 Carácter exploratorio y ausencia de inferencia formal

El análisis por conglomerados no es, en sentido estricto, un procedimiento inferencial clásico. No se basa en un modelo probabilístico explícito ni proporciona contrastes de hipótesis con niveles de significación asociados.

Aunque pueden utilizarse índices como Calinski–Harabasz para comparar soluciones, estos no constituyen pruebas formales de existencia de grupos "reales" en la población. Los clusters identificados describen una estructura en los datos observados, pero no garantizan que dicha estructura represente categorías discretas subyacentes.

Desde la perspectiva epidemiológica, esto implica que los conglomerados deben interpretarse como segmentaciones descriptivas y no como entidades ontológicamente definidas.

7.2 Sensibilidad a la escala y a la elección de distancia

Como se ha señalado en el fundamento conceptual, la distancia utilizada determina la geometría del espacio de análisis.

Cambios en:

- La estandarización,
- La inclusión o exclusión de variables,
- La medida de distancia empleada,

pueden modificar sustancialmente la estructura de los clusters obtenidos.

Esta dependencia metodológica obliga a justificar de manera explícita:

- La selección de variables,
- La necesidad o no de tipificación,
- La elección del método de vinculación.

7.3 Sensibilidad a valores atípicos

Los valores extremos pueden influir de manera desproporcionada en la estructura de los conglomerados, especialmente en métodos basados en medias y varianzas (como Ward o K-medias).

Un único territorio con un perfil excepcional puede:

- Alterar la matriz de distancias.
- Generar un cluster aislado.

- Distorsionar la interpretación global.

En análisis aplicados es recomendable:

- Realizar una exploración previa de outliers.
- Evaluar la estabilidad de la solución ante su exclusión.

7.4 Irreversibilidad en métodos jerárquicos

En los métodos jerárquicos aglomerativos, las fusiones son irreversibles. Una vez que dos objetos o grupos se han unido, no pueden separarse en etapas posteriores.

Esto implica que:

- Errores iniciales de agrupación pueden propagarse.
- La solución final depende del orden secuencial de fusiones.

En contraste, los métodos particionales permiten reasignaciones iterativas.

7.5 Determinación no única del número de clusters

La elección de k no está determinada por una regla universal. Diferentes criterios pueden sugerir valores distintos. Además, en datos con estructura continua sin fronteras claras, cualquier partición introduce una segmentación artificial.

En contextos epidemiológicos, debe primar la coherencia sustantiva sobre la mera optimización matemática.

7.6 Dependencia del conjunto de variables

El análisis por conglomerados no identifica "tipos naturales" universales; identifica estructuras relativas al conjunto de variables consideradas.

Si se modifican las variables incluidas en el análisis:

- Puede cambiar la estructura de los clusters.
- Puede alterarse la interpretación epidemiológica.

Por tanto, la selección de variables constituye una decisión sustantiva, no únicamente técnica.

7.7 Riesgo de sobreinterpretación

Existe una tendencia frecuente a:

- Etiquetar clusters con nombres sugerentes.
- Atribuirles significado causal.
- Asumir que representan categorías reales discretas.

Sin embargo, en muchos casos los conglomerados son simplificaciones de gradientes continuos.

El investigador debe mantener una actitud crítica y evitar considerar los conglomerados como categorías naturales o definitivas.

7.8 Consideraciones finales sobre las limitaciones

El análisis por conglomerados es especialmente útil en fases exploratorias y descriptivas del análisis epidemiológico. No obstante:

- No sustituye a modelos explicativos.
- No establece relaciones causales.
- No determina automáticamente la estructura óptima del fenómeno.

Su valor reside en la organización estructural de la información y en la generación de nuevas preguntas analíticas.

8 Aplicaciones ilustrativas del análisis por conglomerados

La comprensión profunda del análisis por conglomerados exige observar su comportamiento en conjuntos de datos concretos. Los ejemplos clásicos no sólo tienen valor pedagógico, sino que permiten verificar cómo los principios matemáticos previamente desarrollados se traducen en estructuras empíricas reconocibles.

En este apartado se presentan dos conjuntos de datos paradigmáticos: el dataset Iris de Anderson y el dataset bidimensional de Ruspini. Ambos permiten ilustrar con claridad la lógica del método jerárquico y del dendrograma, así como sus implicaciones interpretativas.

Los pasos detallados para la reproducción del análisis en SPSS se recogen en los Anexos I y II.

8.1 El dataset Iris de Anderson

8.1.1 Descripción del conjunto de datos

El dataset Iris, introducido por Ronald A. Fisher en 1936 a partir de mediciones realizadas por Edgar Anderson, constituye uno de los ejemplos más conocidos en estadística multivariante.

Contiene:

- 150 observaciones.
- Tres especies de iris: *Setosa*, *Versicolor* y *Virginica*.
- Cuatro variables cuantitativas:
    - Longitud del sépalo.
    - Anchura del sépalo.
    - Longitud del pétalo.
    - Anchura del pétalo.

Desde el punto de vista geométrico, cada flor puede representarse como un punto en el espacio $\mathbb{R}^4$:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$$

Aunque el conjunto incluye la variable "Species", en el contexto del análisis por conglomerados esta variable no se utiliza en la construcción de los clusters. La clasificación real sólo se emplea posteriormente como referencia para evaluar la coherencia del resultado.

8.1.2 Aplicación del método jerárquico

Se aplica un método jerárquico aglomerativo utilizando:

- Distancia euclídea al cuadrado.
- Método de Ward.

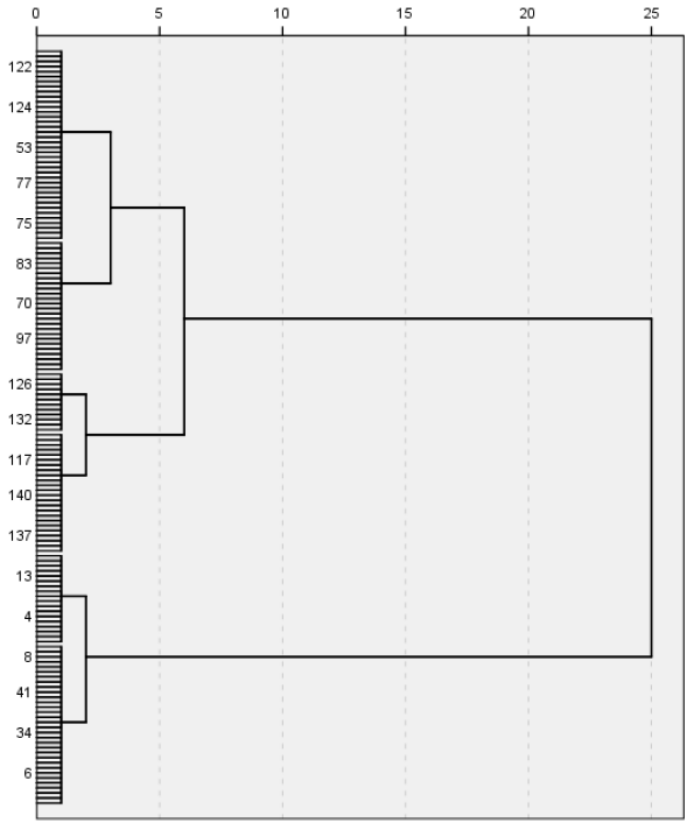
Esta elección se justifica por:

1. La naturaleza cuantitativa de las variables.
2. El fundamento en la minimización de la varianza intracluster.
3. La coherencia con la formulación:

$$W(k) = \sum_{r=1}^{k} \sum_{i \in C_r} \| \mathbf{x}_i - \bar{\mathbf{x}}_r \|^2$$

La construcción del dendrograma permite explorar distintas soluciones para k.

Ilustración 11. Dendrograma obtenido mediante el método de Ward para el conjunto Iris de Anderson.



8.1.3 Interpretación del dendrograma

El análisis visual del dendrograma suele mostrar:

- Una separación clara de un grupo bien definido (correspondiente a *Setosa*).
- Una mayor proximidad estructural entre *Versicolor* y *Virginica*.

Este resultado es coherente con la estructura biológica del conjunto:

- *Setosa* presenta características morfológicas claramente diferenciadas.
- *Versicolor* y *Virginica* presentan solapamiento parcial en el espacio multidimensional.

Desde un punto de vista geométrico, esto significa que las distancias entre observaciones de *Setosa* y las otras especies son sistemáticamente mayores que las distancias internas entre *Versicolor* y *Virginica*.

8.1.4 Comparación con la clasificación real

Una vez fijado $k = 3$, puede compararse la variable de cluster obtenida con la variable “Species”.

El análisis suele mostrar:

- Clasificación prácticamente perfecta de *Setosa*.
- Algunas confusiones entre *Versicolor* y *Virginica*.

Este resultado ilustra dos aspectos fundamentales del análisis por conglomerados:

1. El método es capaz de recuperar estructuras reales cuando la separación geométrica es clara.
2. La existencia de solapamiento en el espacio de variables se traduce en ambigüedad estructural.

Desde una perspectiva metodológica, el ejemplo demuestra que el clustering refleja la geometría del espacio de datos y no impone artificialmente separaciones perfectas.

8.2 El dataset Ruspini

8.2.1 Descripción del conjunto de datos

El dataset de Ruspini (1970) consiste en un conjunto de puntos en $\mathbb{R}^2$, diseñado específicamente para ilustrar métodos de clustering.

Presenta:

- Distribución claramente agrupada.
- Cuatro conglomerados visualmente distinguibles.
- Separación geométrica evidente.

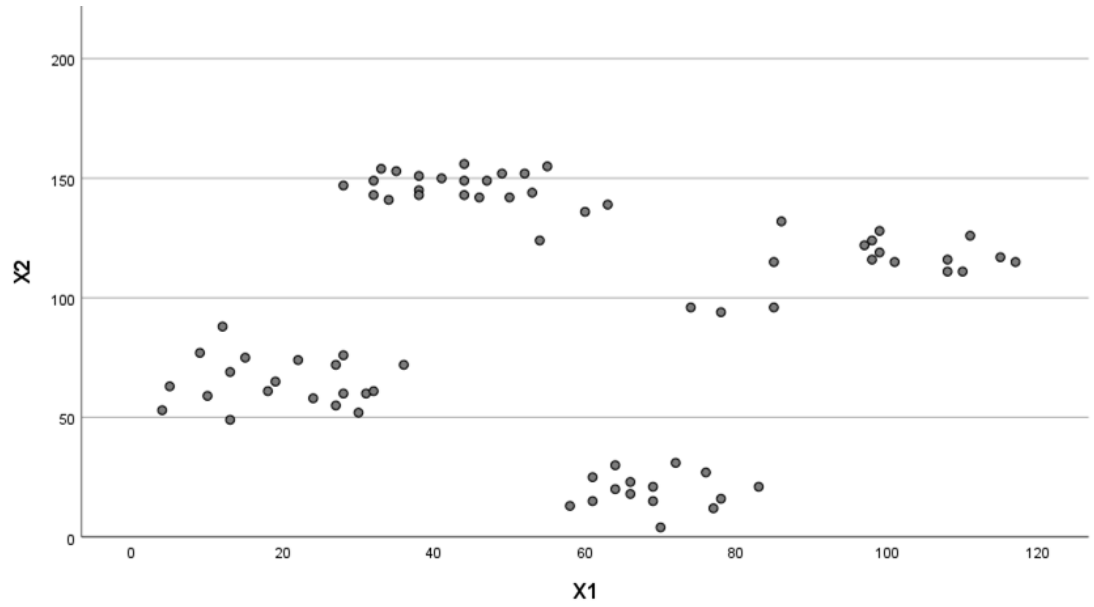
Cada objeto se representa como:

$$\mathbf{x}_i = (x_{i1}, x_{i2})$$

8.2.2 Visualización geométrica

En este caso, la representación gráfica directa permite observar cuatro nubes de puntos relativamente compactas y separadas.

Ilustración 12. Distribución original del conjunto Ruspini.



Es evidente la estructura de los grupos, sin necesidad de algoritmos.

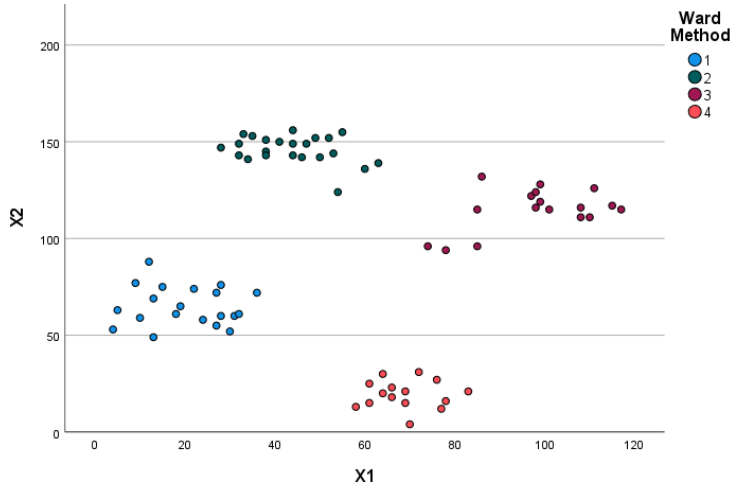
8.2.3 Aplicación del método jerárquico

Aplicando Ward con distancia euclídea al cuadrado, el dendrograma suele mostrar:

- Fusiones tempranas dentro de cada nube de puntos.
- Saltos marcados al intentar unir conglomerados geométricamente separados.

El análisis del dendrograma permite identificar claramente una solución con $k = 4$.

Ilustración 13. Segmentación por conglomerados del conjunto Ruspini (método de Ward, k = 4).



8.2.4 Interpretación

En este caso, la coincidencia entre:

- Estructura geométrica observable.
- Resultado del análisis jerárquico.

confirma la coherencia del método.

La separación clara en el espacio bidimensional se traduce en:

- Distancias pequeñas dentro de cada grupo.
- Distancias grandes entre grupos.
- Incrementos significativos de ΔW al fusionar clusters distintos.

Este ejemplo ilustra de manera especialmente clara el significado geométrico de la minimización de la varianza intracluster.

8.3 Comparación metodológica

La comparación entre Iris y Ruspini permite extraer conclusiones fundamentales:

1. Cuando la separación geométrica es clara (Ruspini), el clustering reproduce fielmente la estructura.
2. Cuando existe solapamiento parcial (Iris), el clustering refleja dicha ambigüedad.
3. El método no “inventa” grupos; organiza el espacio de datos según la métrica elegida.
4. La interpretación debe considerar la estructura geométrica subyacente.

Desde el punto de vista formativo en Salud Pública, estos ejemplos muestran que el análisis por conglomerados:

- Es una herramienta estructural basada en distancia.
- Depende de la naturaleza de las variables.
- Requiere interpretación crítica.
- Puede utilizarse para segmentar poblaciones o territorios con fundamento matemático.

9 BIBLIOGRAFÍA

- Ward JH Jr. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58(301):236–244.
- Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat.* 1974;3(1):1–27.
- Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen.* 1936;7(2):179–188.
- Ruspini EH. Numerical methods for fuzzy clustering. *Inf Sci.* 1970;2(3):319–350.
- Everitt BS, Landau S, Leese M, Stahl D. *Cluster Analysis.* 5th ed. Chichester: Wiley; 2011.
- Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis.* 6th ed. Upper Saddle River: Pearson; 2007.
- Hair JF, Black WC, Babin BJ, Anderson RE. *Multivariate Data Analysis.* 7th ed. Harlow: Pearson; 2014.
- Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley; 1990.

ANEXO I. PROCEDIMIENTO EN SPSS: DATASET IRIS

DESCRIPCIÓN DEL CONJUNTO DE DATOS

El dataset Iris, introducido por Fisher (1936) a partir de mediciones realizadas por Anderson, contiene 150 observaciones correspondientes a tres especies de iris (*setosa*, *versicolor* y *virginica*).

Para cada planta se registran cuatro variables cuantitativas:

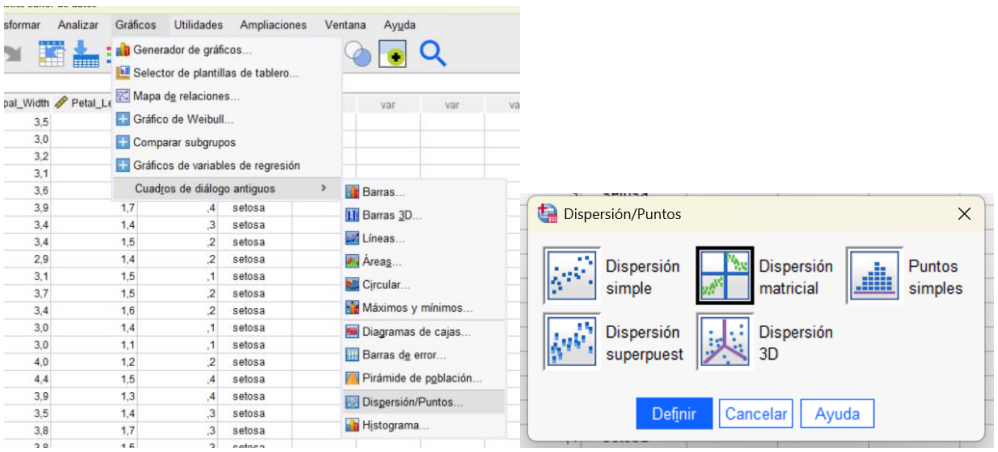
- Longitud del sépalo (Sepal Length)
- Anchura del sépalo (Sepal Width)
- Longitud del pétalo (Petal Length)
- Anchura del pétalo (Petal Width)

El objetivo del análisis por conglomerados en este ejemplo es identificar agrupaciones naturales en el espacio de estas cuatro variables sin utilizar la variable *Species*, que se reservará únicamente para evaluar la concordancia posterior.

EXPLORACIÓN GRÁFICA INICIAL

Antes de realizar el análisis jerárquico, se recomienda explorar la estructura geométrica de los datos mediante una matriz de diagramas de dispersión.

En SPSS: **Gráficos → Constructor de gráficos → Matriz de dispersión**



Se seleccionan las cuatro variables cuantitativas.

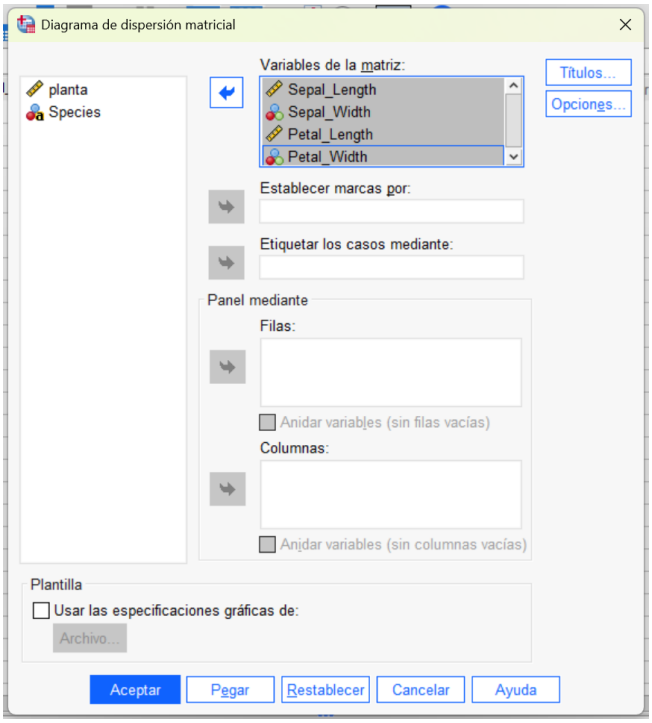
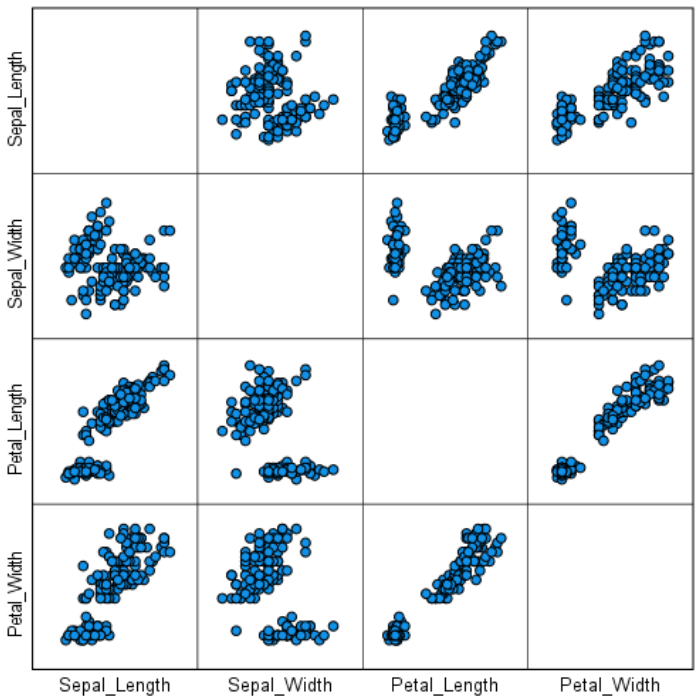


Ilustración 14. Matriz de dispersión de las variables del dataset Iris.



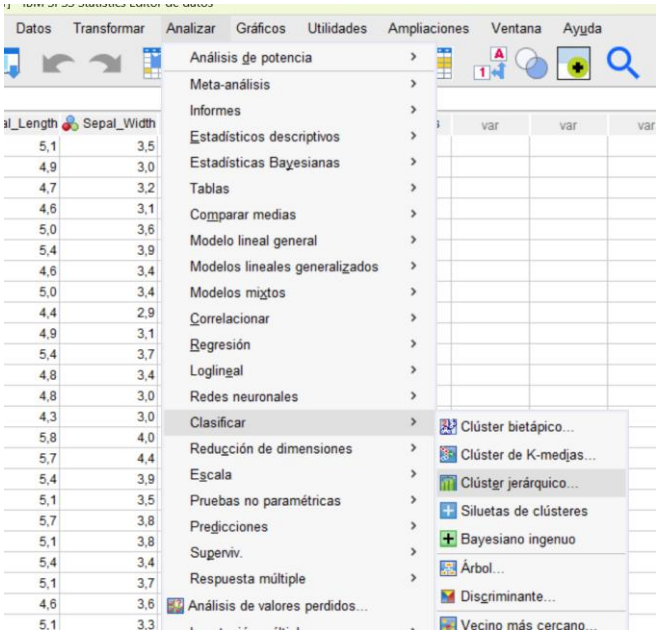
Interpretación:

En varias combinaciones de variables (especialmente Petal Length vs Petal Width) se observan al menos dos agrupaciones claramente diferenciadas. No obstante, el solapamiento parcial entre algunas observaciones sugiere que la estructura puede ser más compleja cuando se consideran simultáneamente las cuatro dimensiones.

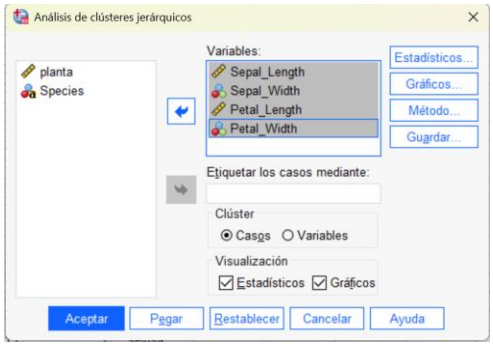
Esta exploración preliminar justifica la aplicación de un procedimiento jerárquico.

ANÁLISIS JERÁRQUICO (MÉTODO DE WARD)

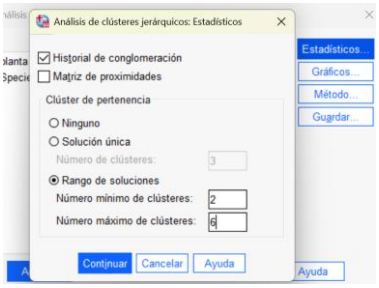
En SPSS: **Analizar → Clasificar → Conglomerados jerárquicos**



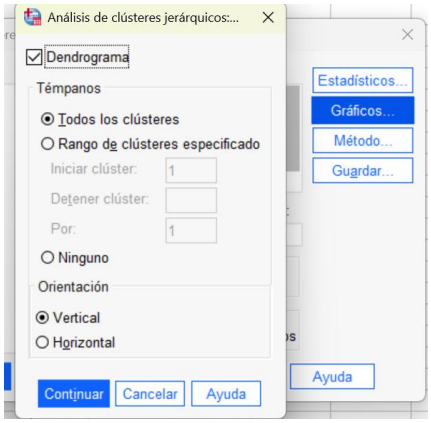
- Variables: Sepal Length, Sepal Width, Petal Length, Petal Width



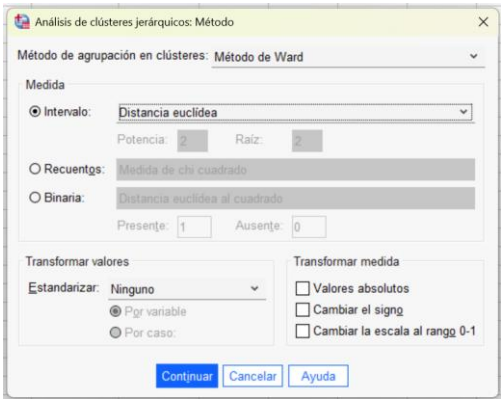
- Estadísticos: Rango de soluciones (mín. 2; máx. 6)



- Gráficos: Dendrograma



- Método:
  - Método de agrupación en clusteres: Método Ward
  - Medida: Intervalo -> Distancia Euclídea
  - **Posibilidad de estandarizar las variables si no estamos seguros de que el rango sea similar.* Transformar variables -> Estandarizar: Puntuaciones Z



- Guardar: Rango de soluciones “Número mínimo 2; Número máximo 6”.

  SPSS crea cinco nuevas variables (CLU2_1; CLU3_1; CLU4_1; CLU5_1; CLU6_1) que asigna a cada observación su conglomerado correspondiente

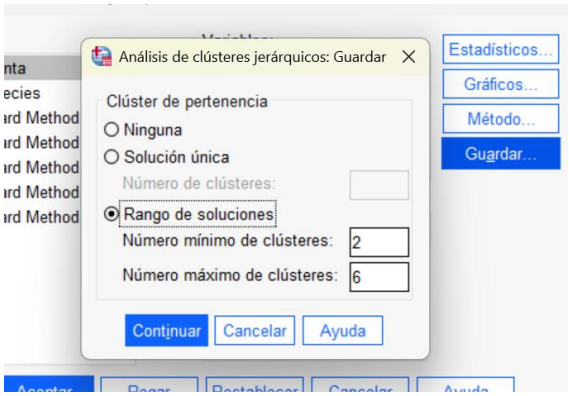
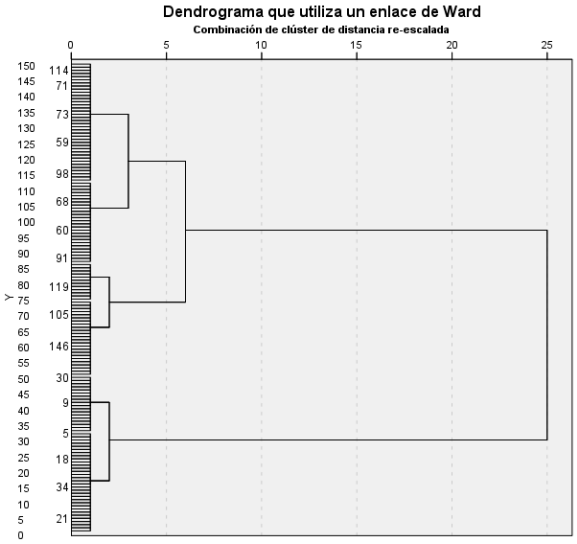


Ilustración 15. Dendrograma obtenido mediante el método de Ward (Iris).



Interpretación del dendrograma:

El árbol jerárquico muestra una separación muy marcada de un grupo respecto a los demás en una etapa temprana del proceso de fusión. Posteriormente, se observa una subdivisión adicional dentro del conjunto restante.

Visualmente, el dendrograma sugiere una solución de tres conglomerados como estructura razonable, en coherencia con la información taxonómica conocida.

VALIDACIÓN DEL NÚMERO DE CLUSTERS

Para evaluar formalmente la solución, se aplica el criterio de Calinski y Harabasz, concretamente, vamos a obtener los valores omega para encontrar el número de clusters como el valor de k que minimiza omega.

Procedimiento (EXCEL):

1. Para cada solución cluster (k = 2, 3, 4…), realizar un ANOVA de un factor:
    - Factor: variable de pertenencia al cluster.
    - Variables dependientes: las cuatro variables originales.
2. Obtener el valor F para cada variable.
3. Calcular el **VRC (Variance Ratio Criterion)** como la suma de los valores F.
4. Calcular el valor omega:

$$\omega_i = VRC_{i+1} - 2VRC_i + VRC_{i-1}$$

Tabla 2. Valores VRC y omega para distintas soluciones cluster.

k	VRC	omega
2	1638,352	
3	2189,004	-531,301
4	2208,355	-59,438
5	2168,268	-310,378
6	1817,803	

Interpretación:

La solución con tres conglomerados presenta el valor óptimo según el criterio empleado y coincide con la estructura sugerida por el dendrograma.

**Comprueba el número de conglomerados que se obtiene con el método de maximizar el valor de CH(k).*

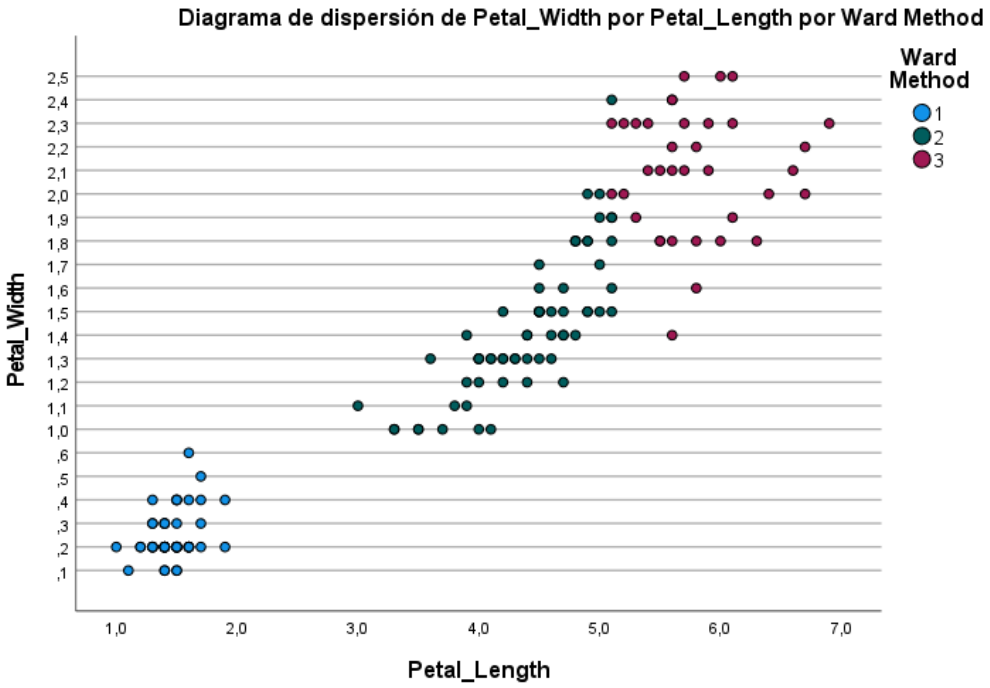
REPRESENTACIÓN GRÁFICA DE LA SOLUCIÓN SELECCIONADA

Para visualizar la solución seleccionada:

En SPSS: **Gráficos → Constructor de gráficos → Dispersión simple**

- Eje X: Petal Length
- Eje Y: Petal Width
- Color: variable de cluster (CLU3_1)

Ilustración 16. Scatterplot de Iris (Petal Length vs Petal Width) coloreado por conglomerado (k = 3).



Interpretación:

El gráfico evidencia una separación clara de uno de los conglomerados, correspondiente a la especie *setosa*. Los otros dos grupos muestran cierto grado de solapamiento, especialmente en la zona intermedia del plano.

La representación gráfica confirma la coherencia entre:

- El dendrograma,
- La estructura geométrica bidimensional,
- Y la clasificación final obtenida.

En conjunto, la convergencia entre:

- Evidencia gráfica,

- Estructura jerárquica,
- Concordancia con la variable taxonómica,
- Y criterio cuantitativo de validación,

apoya la elección de k = 3 como solución final.

EVALUACIÓN DE LA CLASIFICACIÓN "EXTRA"

Para evaluar la coherencia de la clasificación obtenida, se realiza una tabla de contingencia entre la variable de cluster y la variable *Species*:

Analizar → Estadísticos descriptivos → Tablas cruzadas

Tabla 3. Tabla de contingencia Species × Cluster (Ward, k = 3).

Tabla cruzada Species*Ward Method

Recuento

		Ward Method			
		1	2	3	Total
Species	setosa	50	0	0	50
	virginica	0	14	36	50
	versicolor	0	50	0	50
Total		50	64	36	150

Interpretación:

La clasificación muestra una correspondencia casi perfecta para la especie *setosa*, que aparece agrupada en un único cluster. Las especies *versicolor* y *virginica* presentan mayor solapamiento, lo cual es consistente con la proximidad geométrica observada en el espacio de variables.

Este resultado confirma que el análisis por conglomerados recupera adecuadamente la estructura biológica subyacente cuando existe separación clara.

EJERCICIO PROPUESTO – PARA PRACTICAR

Como ejercicio obtener el dendrograma usando el método del vecino más próximo, el método del vecino más lejano, vinculación inter-grupo e intra-grupo, y agrupación de centroides. Comentar las diferencias observadas. Para finalizar usa el método de Calinski y Harabasz para determinar el número de conglomerados.

ANEXO II. PROCEDIMIENTO EN SPSS: DATASET RUSPINI

DESCRIPCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos de Ruspini está formado por 75 observaciones definidas mediante dos variables cuantitativas, X1 y X2. Cada observación representa un punto en el plano bidimensional $\mathbb{R}^2$.

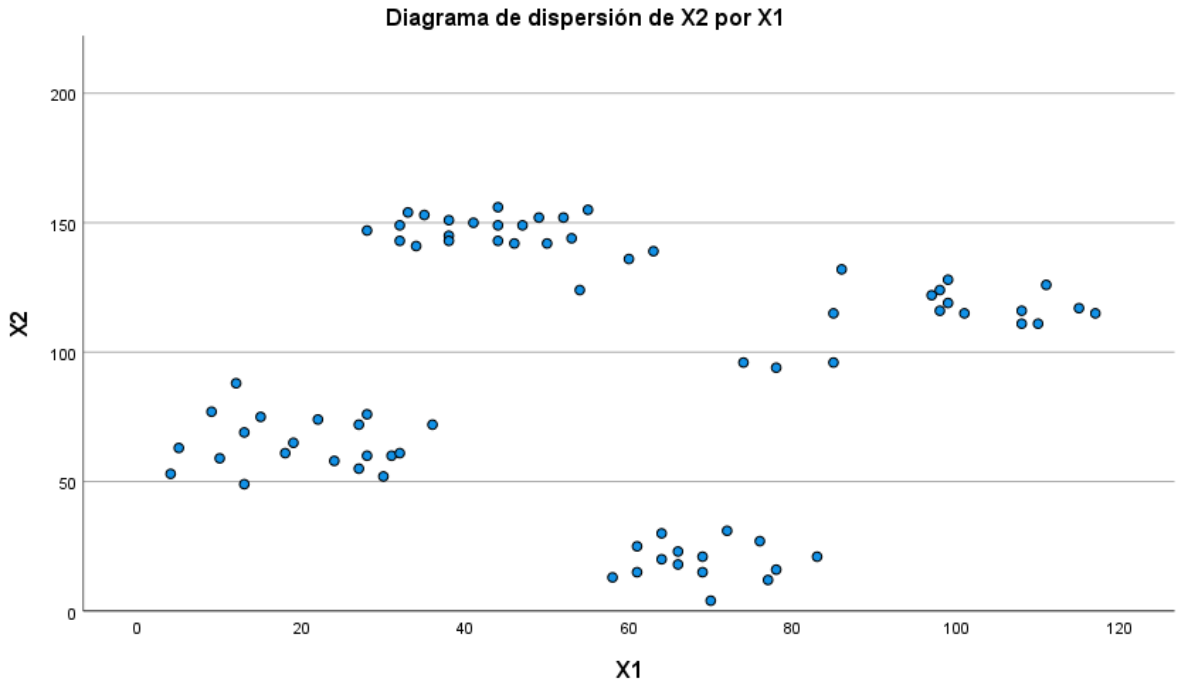
El objetivo del análisis es identificar agrupaciones naturales en función de la proximidad geométrica entre los puntos, utilizando un procedimiento jerárquico aglomerativo.

A diferencia del dataset Iris, en este caso no existe una clasificación externa de referencia; el análisis tiene carácter puramente estructural.

EXPLORACIÓN GRÁFICA INICIAL

La representación gráfica inicial consiste en un diagrama de dispersión de los puntos (X1, X2).

Ilustración 17. Distribución original del conjunto Ruspini en el plano (X1–X2).



Interpretación:

La nube de puntos sugiere visualmente la existencia de cuatro agrupaciones claramente diferenciadas:

- Un grupo en la zona inferior izquierda.
- Un grupo en la zona superior izquierda.
- Un grupo en la zona superior derecha.
- Un grupo en la zona inferior derecha.

La separación entre estas regiones es notable, lo que anticipa que un procedimiento de clustering debería identificar cuatro conglomerados bien definidos.

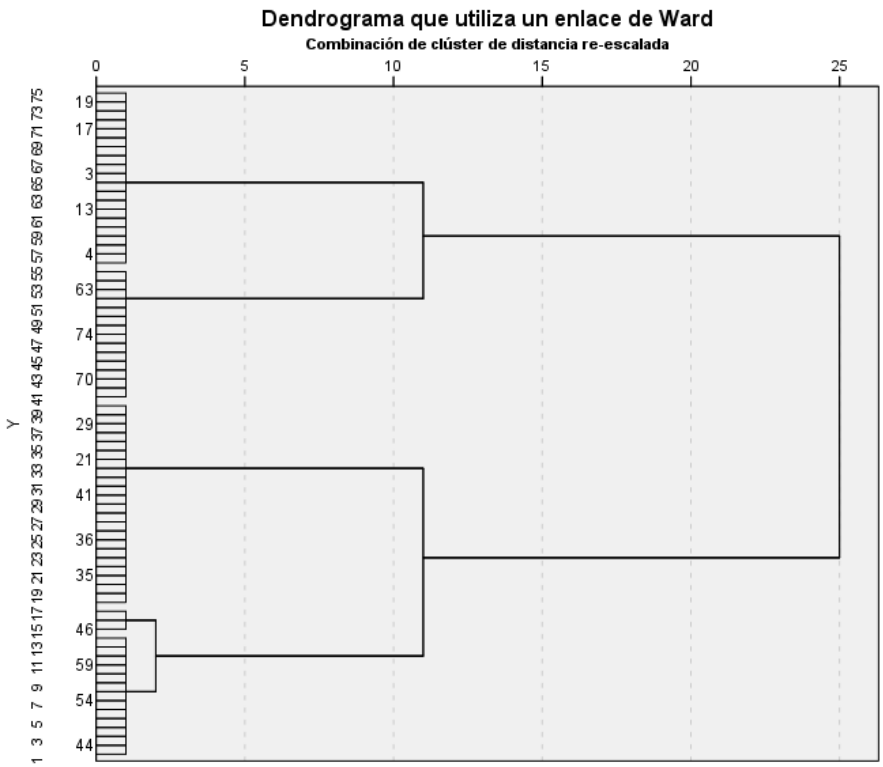
ANÁLISIS JERÁRQUICO (MÉTODO DE WARD)

Se aplica un análisis jerárquico aglomerativo utilizando:

- Método de vinculación: Ward.
- Medida de distancia: Euclídea.

El dendrograma obtenido representa la secuencia de fusiones entre observaciones y grupos.

Ilustración 18. Dendrograma del conjunto Ruspini (método de Ward).



Interpretación del dendrograma:

El dendrograma muestra:

- Fusiones tempranas entre puntos cercanos dentro de cada agrupación visible.
- Un incremento notable en la distancia de fusión al intentar unir las cuatro estructuras principales.

El análisis visual sugiere que una solución con cuatro conglomerados constituye una partición natural del conjunto.

Además, el examen de las soluciones intermedias (k = 2, 3, 4, 5 y 6) permite observar cómo los grupos grandes se subdividen progresivamente, manteniendo coherencia con la estructura geométrica del plano.

VALIDACIÓN DEL NÚMERO DE CLUSTERS

Para validar formalmente la solución, se aplica el criterio de Calinski y Harabasz. Procedimiento (EXCEL).

Para cada solución cluster (k = 2, 3, 4, 5 y 6), se realiza un análisis de varianza tomando como factor la variable de pertenencia al cluster y como variables dependientes X1 y X2.

El valor VRC (Variance Ratio Criterion) se obtiene como la suma de los valores F correspondientes a ambas variables.

Posteriormente se calcula el valor omega mediante:

$$\omega_i = VRC_{i+1} - 2VRC_i + VRC_{i-1}$$

Tabla 4. Valores VRC y omega para distintas soluciones cluster (Ruspini).

k	VRC	omega
2	331,218	
3	286,524	661,664
4	903,494	-642,718
5	877,746	-112,785
6	739,213	

Interpretación:

La solución con cuatro conglomerados presenta:

- El mayor valor de VRC.
- El menor valor de omega.

Esta convergencia entre criterios cuantitativos respalda la elección de $k = 4$ como número óptimo de clusters.

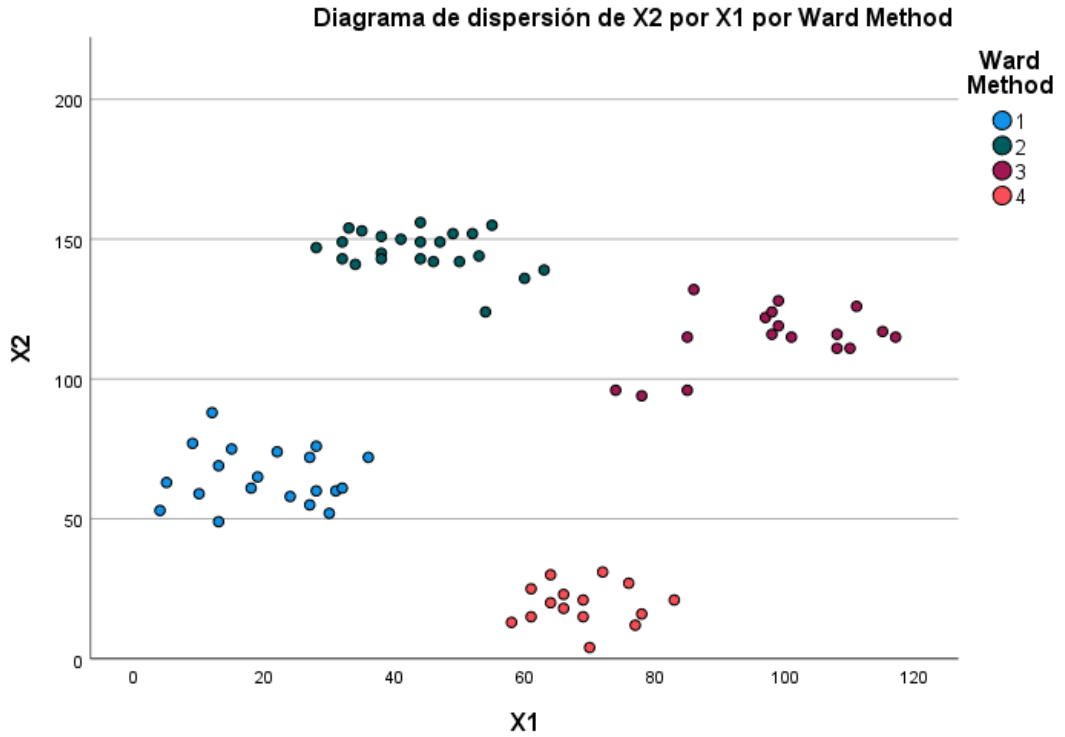
Cabe destacar que el criterio omega es un criterio local y requiere la comparación con soluciones adyacentes; en este caso, la solución de cuatro grupos destaca claramente frente a las alternativas evaluadas.

**Comprueba el número de conglomerados que se obtiene con el método de maximizar el valor de CH(k).*

REPRESENTACIÓN GRÁFICA DE LA SOLUCIÓN SELECCIONADA

Para visualizar la solución con cuatro conglomerados, se representa nuevamente el diagrama de dispersión (X1, X2), coloreando los puntos según la variable de pertenencia al cluster.

Ilustración 19. Scatterplot de Iris (Petal Length vs Petal Width) coloreado por conglomerado (k = 3).



Interpretación:

La representación coloreada confirma:

- Alta cohesión interna dentro de cada cluster.
- Clara separación entre conglomerados.
- Correspondencia directa entre estructura geométrica y clasificación obtenida.

La segmentación final reproduce fielmente las cuatro agrupaciones visibles en la exploración inicial, mostrando coherencia entre:

- Evidencia gráfica,
- Estructura jerárquica,
- Y criterios cuantitativos de validación.

CIERRE INTERPRETATIVO DEL ANEXO

En este ejemplo, el análisis por conglomerados no impone una estructura artificial, sino que formaliza una segmentación que ya era sugerida por la disposición geométrica de los datos.

La concordancia entre:

- Visualización inicial,
- Dendrograma,
- Criterios de validación,
- Y representación final,

convierte este ejemplo en un caso paradigmático de identificación de estructura clara mediante clustering jerárquico.

EJERCICIO PROPUESTO – PARA PRACTICAR

Como ejercicio obtener el dendrograma usando el método del vecino más próximo, el método del vecino más lejano, vinculación inter-grupo e intra-grupo, y agrupación de centroides. Comentar las diferencias observadas. Para finalizar usa el método de Calinski y Harabasz para determinar el número de conglomerados.