ANÁLISIS DE CONGLOMERADOS

Prof. Esperanza Ayuga Téllez

Tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes entre ellos. Detecta grupos internamente homogéneos (y heterogéneos entre sí)

También conocido como: clasificación automática, clasificación no supervisada, reconocimiento de patrones sin supervisión

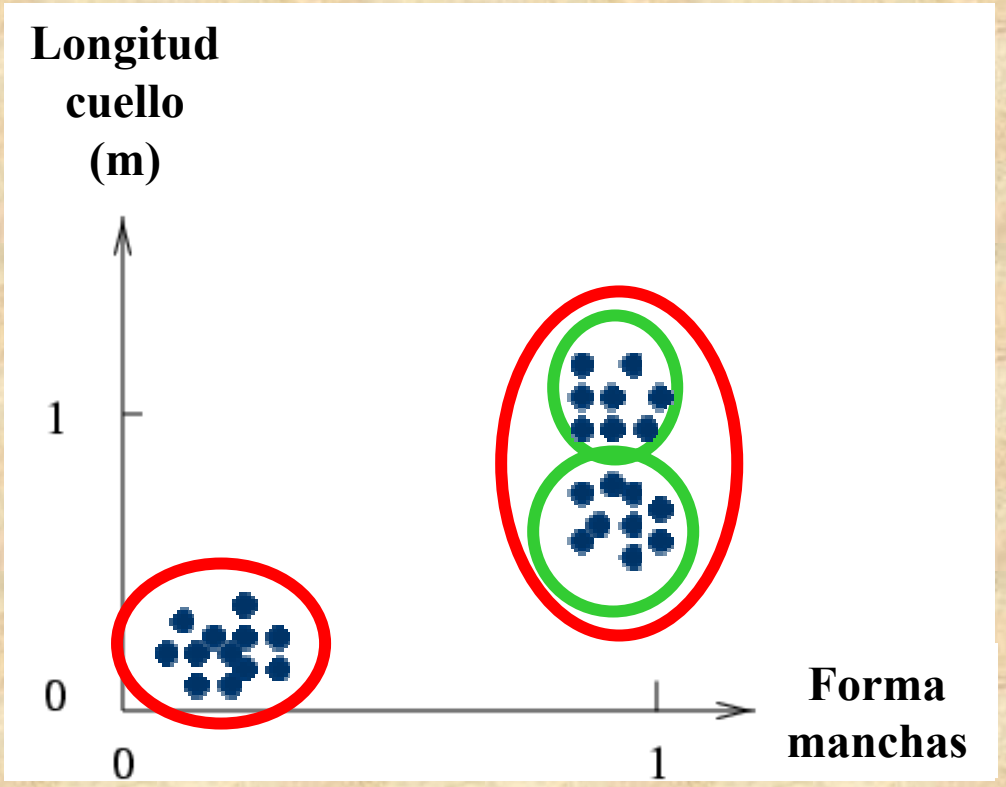
p.e. Agrupar tipos de semillas por efectos de germinación o agrupar clientes por pautas de consumo.

Estudia tres tipos de problemas:

1. Partición de los datos: disponemos de observaciones que pensamos son heterogéneas y deseamos dividirlos en un nº prefijado de grupos, de tal manera que todo elemento quede clasificado y pertenezca a un solo grupo y los grupos sean internamente homogéneos.
2. Construcción de jerarquías: deseamos estructurar los elementos de un conjunto de forma jerárquica por su similitud $\Rightarrow$ ordenar en niveles.
3. Clasificación de variables: en problemas con muchas variables es interesante hacer una división en grupos para luego reducir la dimensión.

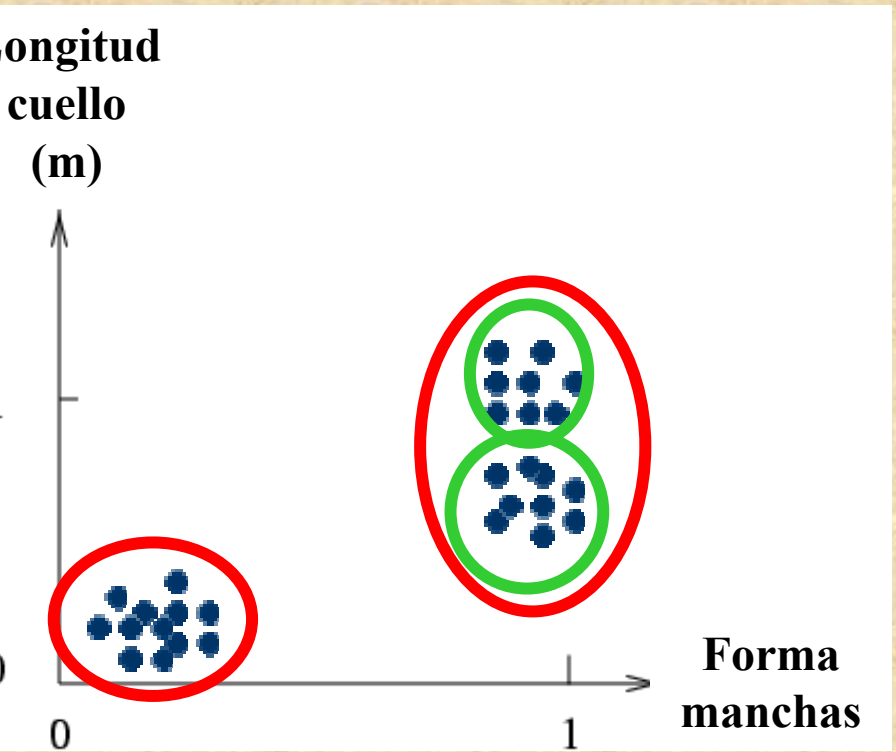
Definición del problema de cluste

Girafas



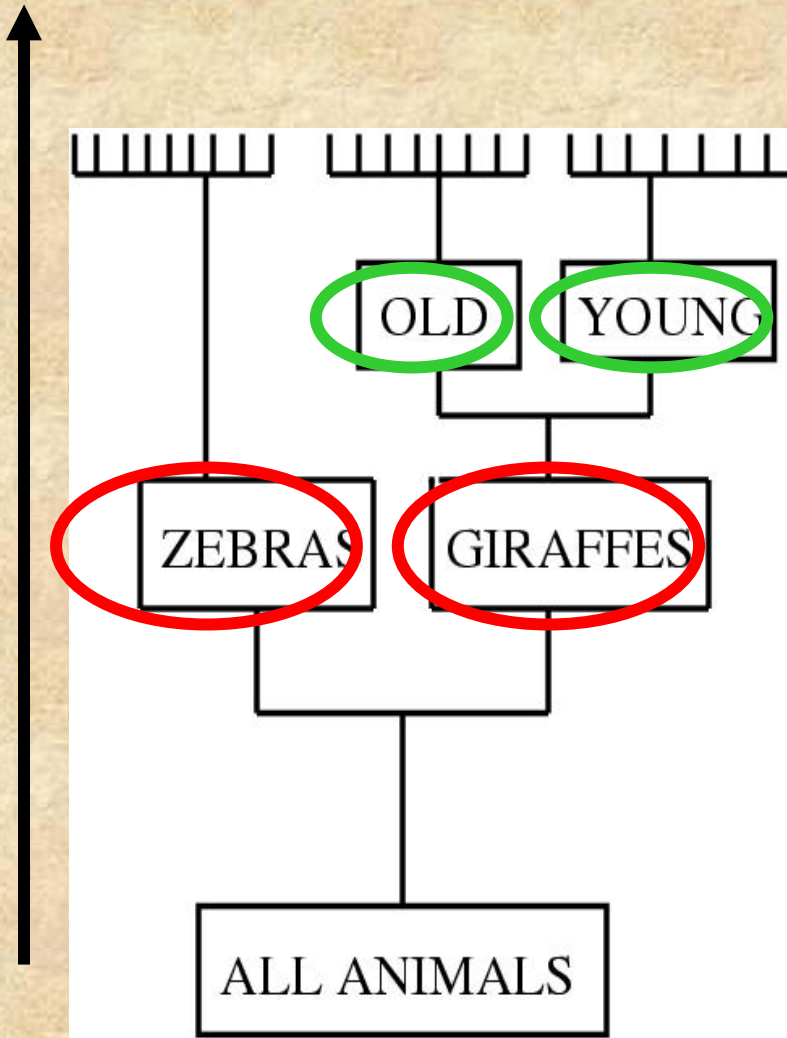
Cebras

Definición del problema

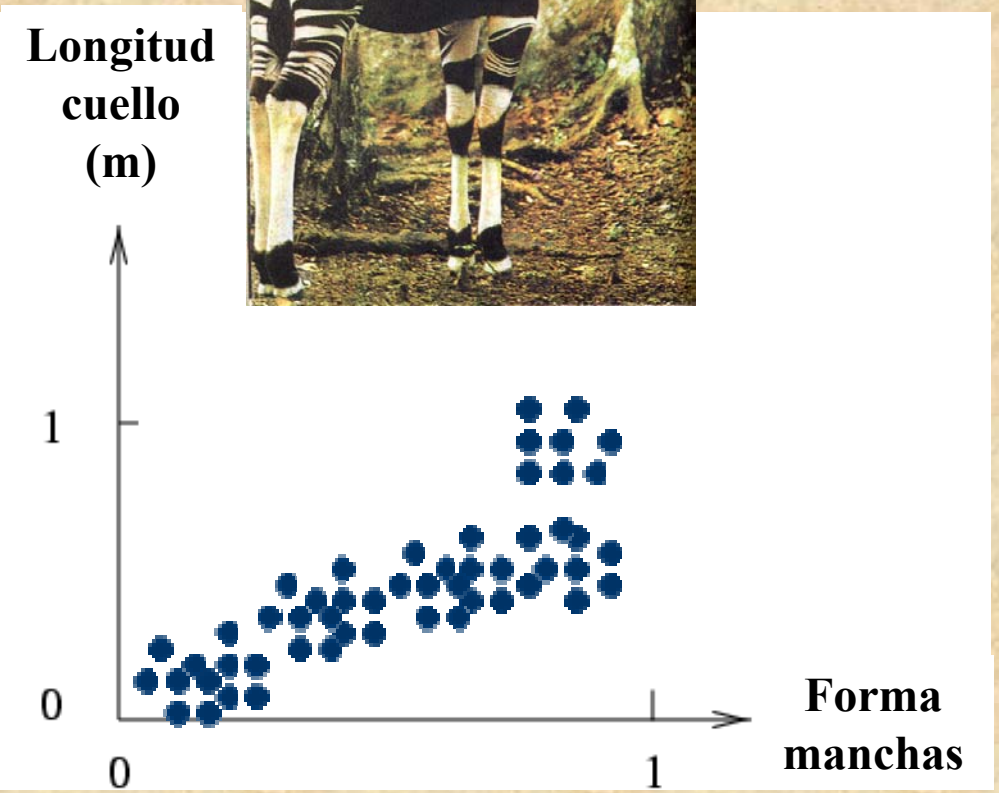


Jerarquización lineal de datos

T (resolución)



¿Pero donde está el okapi?



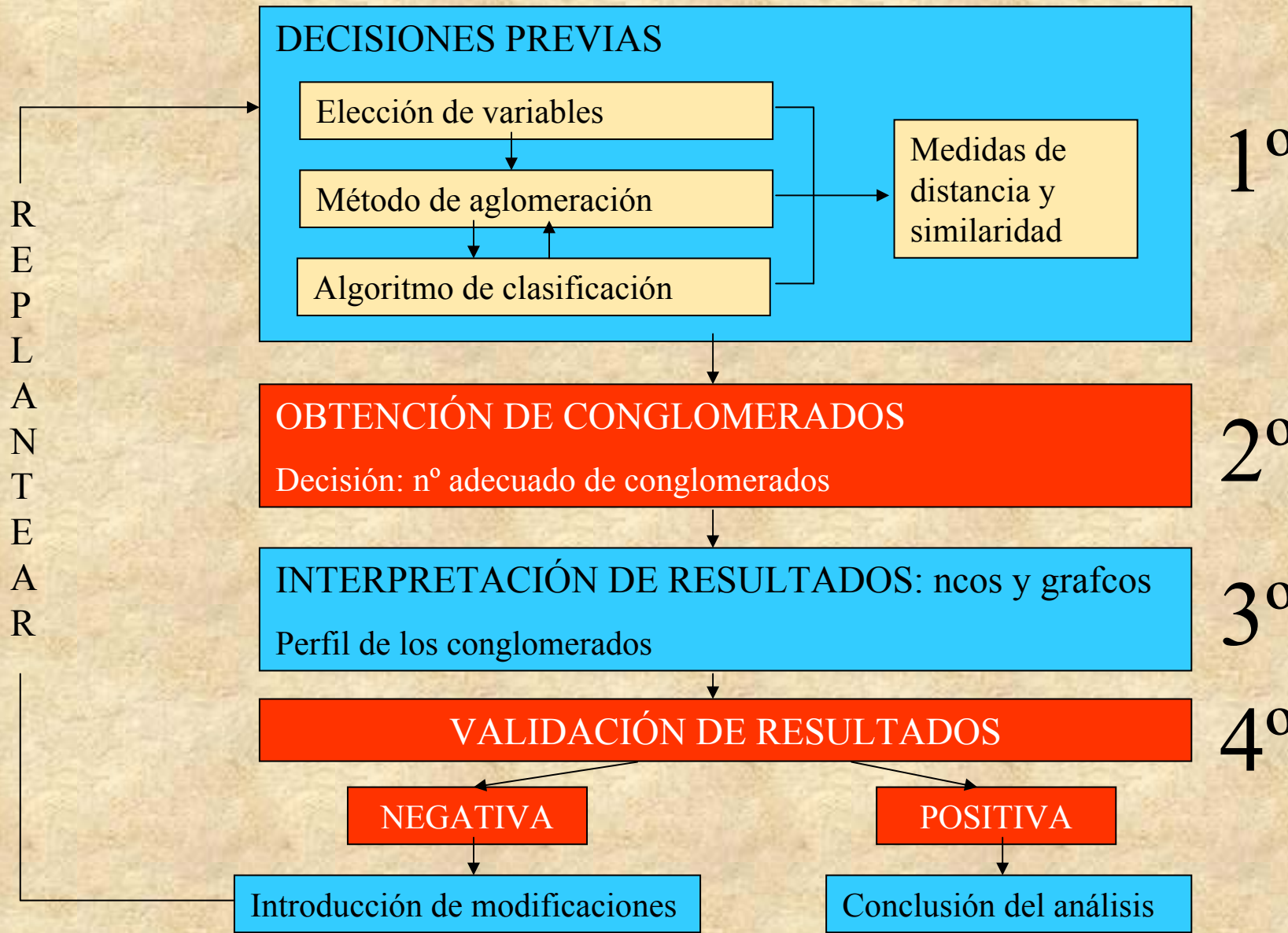
Relación con otras técnicas multivariantes:

1. Análisis discriminante: se trata de otra técnica destinada a la clasificación. Se trata de una técnica confirmatoria que parte de una clasificación previa de los individuos. Suele realizarse posteriormente al A. de conglomerados

2. Análisis factorial exploratorio: puede utilizarse para clasificar mediante los coeficientes de los factores. Puede realizarse para confirmar la agrupación por conglomerados.

3. Escalamiento multidimensional: por su analogía con el A. factorial puede relacionarse con la técnica de conglomerados.

Fases de su aplicación



ELECCIÓN DE VARIABLES

Elegir variables relevantes: dependiendo del objetivo de la investigación $\Rightarrow$ afecta a los conglomerados, al número óptimo de éstos y a la presencia de datos atípicos.

Estandarizar las variables: dependiendo de su rango $\Rightarrow$ si existen pocas diferencias en la magnitud y rango de las variables usadas no es preciso estandarizar.

MÉTODOS DE CONGLOMERACIÓN

Métodos jerárquicos:

- Aglomerativos
  distancias mínimas
  distancias máximas
  promedio entre grupos
  promedio intra grupos
  método Ward
  método del centroide
  método de la mediana

- Divisivos
  método de partición binaria

Métodos no jerárquicos:

- Reasignación
  método K-medias
  nubes dinámicas

- Búsqueda de densidad
  aproximación tipológica
  aproximación probabilística

- Métodos directos
  Grupos por bloque de Hartingan

MÉTODOS DE CONGLOMERACIÓN

Métodos jerárquicos:

Se basan en el cálculo de una matriz de distancias y se aplican con n<200, ya que los cálculos y resultados se complican al aumentar el tamaño de la muestra.

Se pueden aplicar a variables o a observaciones:

- Si se agrupan variables se precisan 3 o más v. numéricas.
- Si se agrupan observaciones se necesita al menos una v. numérica.

MÉTODOS JERÁRQUICOS

Dada la matriz, los algoritmos son de dos tipos:

1. De aglomeración los elementos se van agregando. Requieren menor tiempo y son los más usados.

2. De división parten del conjunto de datos y se van dividiendo. los elementos que se incluyen en un grupo no se pueden reasignar.

MÉTODOS JERÁRQUICOS

Métodos jerárquicos aglomerativos:

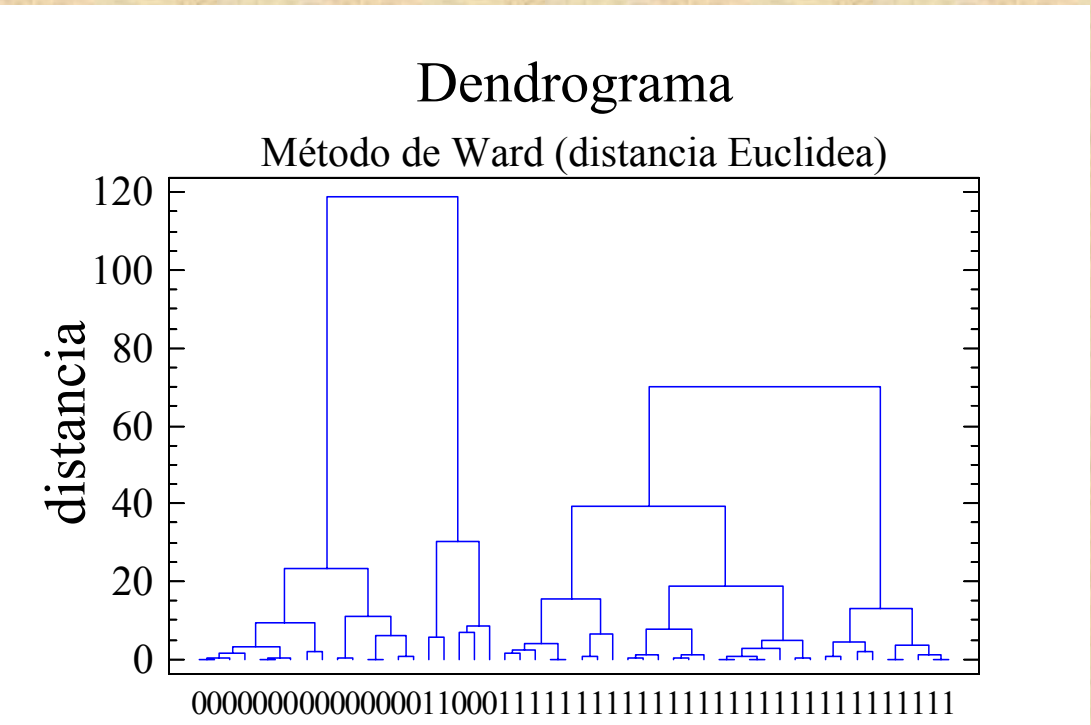
También se denominan ascendentes (Manly, 1990). Parten de objetos singulares (cada objeto un conglomerado) para ir construyendo conglomerados cada vez más complejos hasta concluir en uno sólo.

Se pueden aplicar a variables o a observaciones:

- Si se agrupan variables se precisan 3 o más v. numéricas.
- Si se agrupan observaciones se necesita al menos una v. numérica.

MÉTODOS JERÁRQUICOS

La representación gráfica del resultado de la agrupación jerárquica es el dendrograma, útil si efectivamente los puntos tienen una estructura jerárquica y engañoso en otro caso.

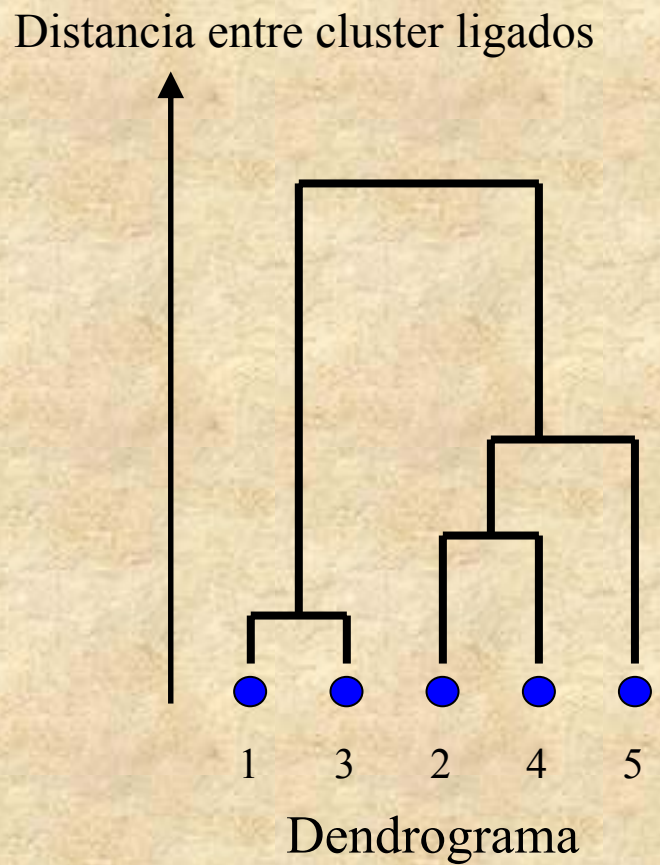


Se recomienda elegir el criterio más adecuado para los datos a tratar y, en caso de duda, probar con varios y comparar los resultados.

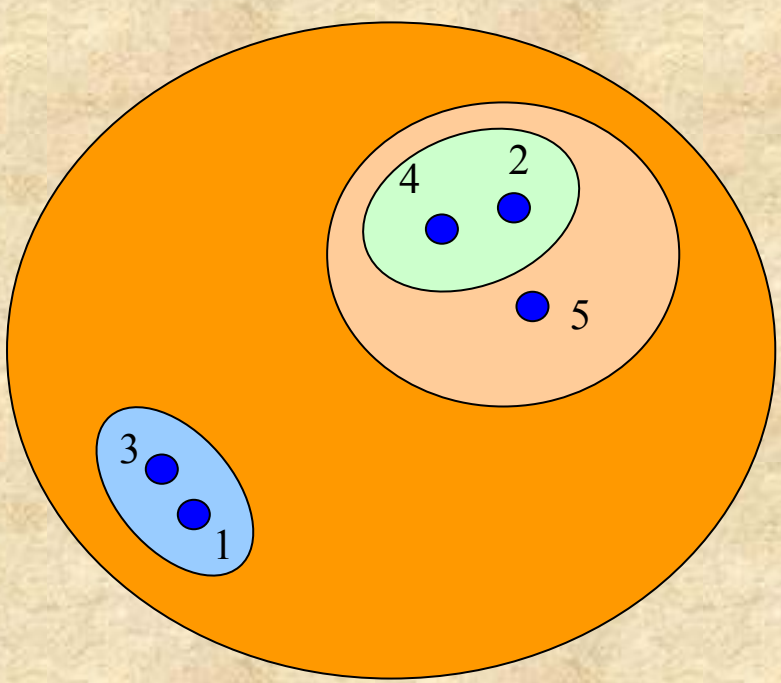
Cluster jerárquico aglomerativo

A cada paso se junta el par de GRUPOS más próximos

Inicialmente ⇒ cada punto = cluster



Dendrograma



erativo

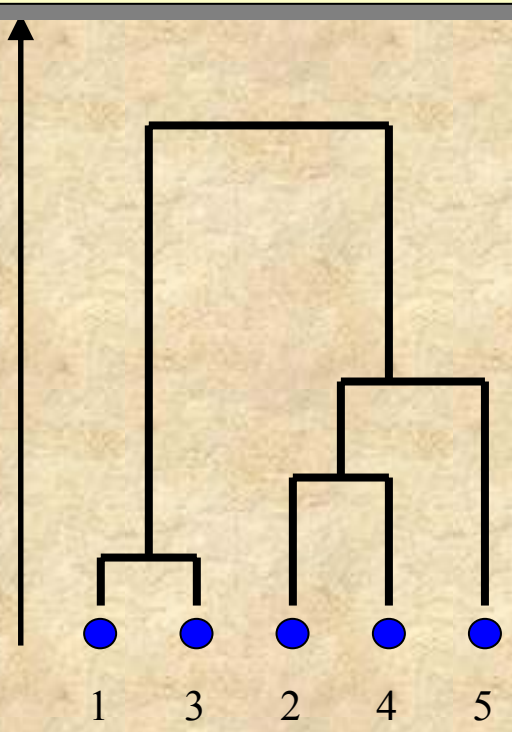
próximos

Algoritmos definidos por la distancia entre el nuevo grupo y el grupo anterior.

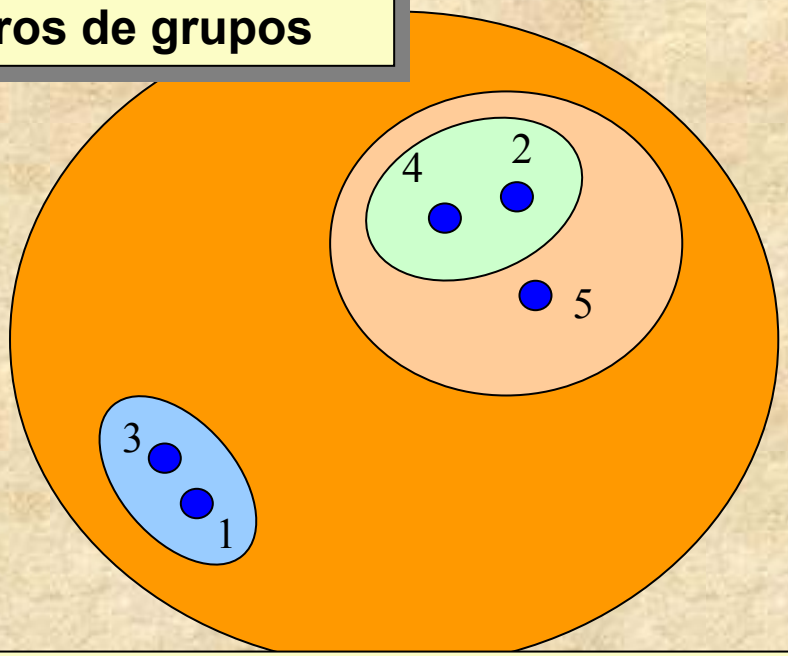
Enlace simple: distancia mínima entre grupos.

Enlace completo : distancia máxima entre grupos.

Enlace medio: distancia media entre todos los pares o distancia entre centros de grupos



Dendrograma



El dendrograma muestra un orden lineal de los datos.

MÉTODOS DE AGLOMERACIÓN

Métodos no jerárquicos:

También conocidos como métodos de “optimización”.

- Los métodos de reasignación permiten reasignar objetos a distintos conglomerados en cada fase.
- Los métodos de búsqueda de densidad agrupan mediante la búsqueda de altas densidades (modas).
- Los métodos directos permiten clasificar de forma simultánea individuos y variables

MÉTODOS DE CONGLOMERACIÓN

Diferencias básicas:

Métodos jerárquicos:	Métodos no jerárquicos:
Comienza con las observaciones y no precisa determinar a priori el número de conglomerados.	Comienza con una partición inicial de conglomerados. A priori se determina el número y composición de los conglomerados
La asignación de objetos es definitiva.	El procedimiento es iterativo y permite la reasignación de objetos.
Operan con una matriz de similaridades.	Operan con la matriz de datos originales.

MÉTODOS DE CONGLOMERACIÓN

Inconvenientes principales

Métodos jerárquicos:	Métodos no jerárquicos:
Si la estructura de la muestra es desconocida resulta difícil escoger el algoritmo.	Dificultad en conocer a priori el número real de los conglomerados existentes en la muestra.
Es difícil operar e interpretar los gráficos con más de 200 datos.	Formar todas las particiones posibles para escoger la óptima es muy complejo.
Mayor cantidad de atípicos en esta partición.	Mayor complejidad en los análisis.
Una mala partición no puede modificarse.	Una mala decisión inicial sobre el nº y composición de los grupos ocasiona una errónea clasificación.

ALGORITMOS DE CONGLOMERACIÓN

De la elección del algoritmo de clasificación dependen el número y composición de los conglomerados obtenidos.

El algoritmo es la forma particular de cálculo empleado en los métodos descritos.

La elección del algoritmo de clasificación depende de:

a) Los objetivos del estudio

b) Las características de los datos: métrica de las variables y tamaño muestral

c) El método elegido: jerárquico o no jerárquico

d) Los límites del programa y ordenador que usemos.

ALGORITMOS DE MÉTODOS JERÁRQUICOS

Los algoritmos tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos.

1. Empezar con tantas clases como elementos, n.
2. Seleccionar los dos datos más próximos y formar con ellos una clase.
3. Sustituir los dos elementos anteriores por uno sólo que representa a la clase. Se calculan las distancias entre éste nuevo elemento y los anteriores.
4. Repetir 2 y 3 hasta agrupar todos los datos en una sola clase

ALGORITMOS DE MÉTODOS JERÁRQUICOS

Parte de una matriz de distancias o similaridades a partir de la que se construye una jerarquía.

La distancia más usada es la euclídea (entre v. estand. univ.)

No es conveniente la distancia de Mahalanobis.

Para v. binarias se trabaja con similaridades.

$$d_{ij} = \sqrt{2(1 - s_{ij})}; \text{simil} = \frac{a}{a + b + c}$$

a=nº de coincidencias;

b=nº de (0,1) y c=nº de (1,0)

ALGORITMOS DE MÉTODOS JERÁRQUICOS

Supongamos dos grupos: A y B con n_A y n_B elementos, la distancia del nuevo grupo (AB) a otro C de n_C se calcula con:

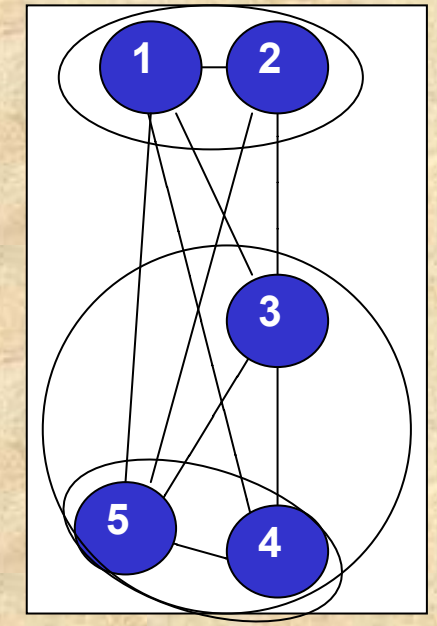
1. Encadenamiento simple o vecino más próximo: tiende a producir grupos alargados que pueden incluir puntos muy distintos en los extremos:

$$d(C;AB)=\min(d_{CA};d_{CB})$$

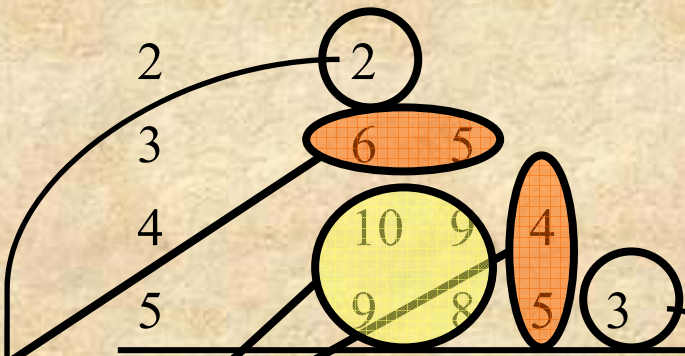
2. Encadenamiento completo o vecino más alejado: tiende a producir grupos esféricos:

$$d(C;AB)=\max(d_{CA};d_{CB})$$

Enlace simple o vecino más próximo o distancia mínima



Objeto	1	2	3	4	5
1					
2					
3					
4					
5					



Matriz de distancias

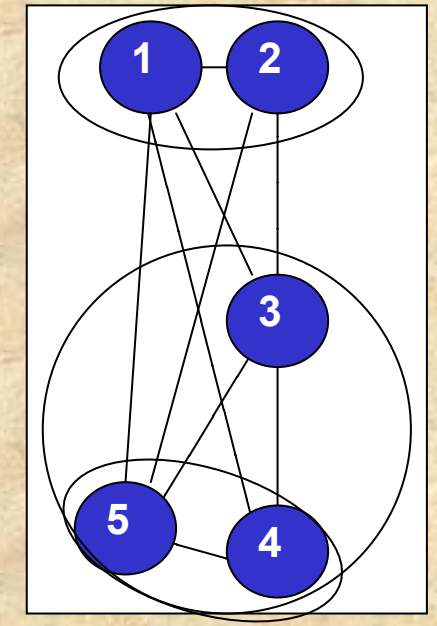
$$d_{(12)3} = \min[d_{13}, d_{23}] = d_{23} = 5$$

$$d_{(45)3} = \min[d_{43}, d_{53}] = d_{43} = 4$$

$$d_{(12)(45)} = \min[d_{14}, d_{24}, d_{15}, d_{25}] = d_{25} = 8$$

Distancia	Cluster
0	1,2,3,4,5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4	(1, 2), (3, 4, 5)
5	(1, 2, 3, 4, 5)

Enlace completo o vecino más lejano o distancia máxima



Objeto	1	2	3	4	5
1					
2					
3					
4					
5					

Matriz de distancias

Distancia	Cluster
0	1,2,3,4,5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
5	(1, 2), (3, 4, 5)
10	(1, 2, 3, 4, 5)

$$d_{(12)3} = \max[d_{13}, d_{23}] = d_{13} = 6$$

$$d_{(45)3} = \max[d_{43}, d_{53}] = d_{53} = 5$$

$$d_{(12)(45)} = \max[d_{14}, d_{24}, d_{15}, d_{25}] = d_{14} = 10$$

ALGORITMOS DE MÉTODOS JERÁRQUICOS

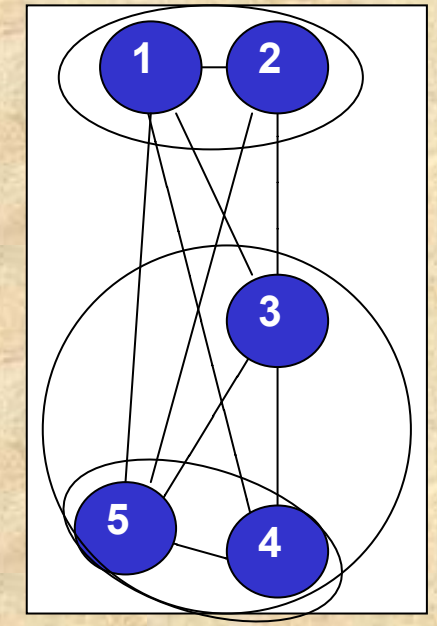
3. Media de grupos: media ponderada de las distancias. Como el primer método, es invariante ante transformaciones monótonas:

$$d(C;AB) = \frac{n_A}{n_A + n_B} d_{CA} + \frac{n_B}{n_A + n_B} d_{CB}$$

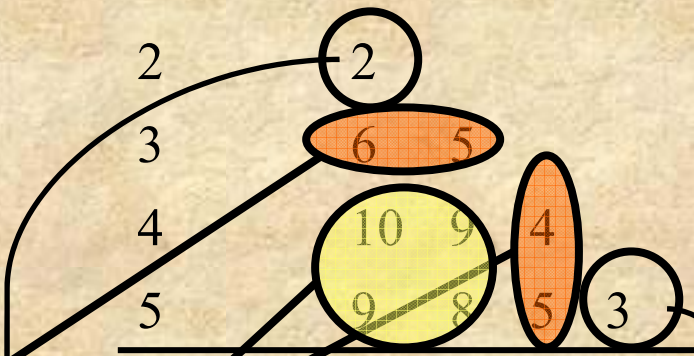
4. Método del centroide: sólo válido para variables continuas. Equivale a la distancia euclídea entre sus centros:

$$d^2(C;AB) = \frac{n_A}{n_A + n_B} d^2_{CA} + \frac{n_B}{n_A + n_B} d^2_{CB} - \frac{n_A n_B}{(n_A + n_B)^2} d^2_{AB}$$

Media de grupos



Objeto	1	2	3	4	5
1					
2					
3					
4					
5					



Matriz de distancias

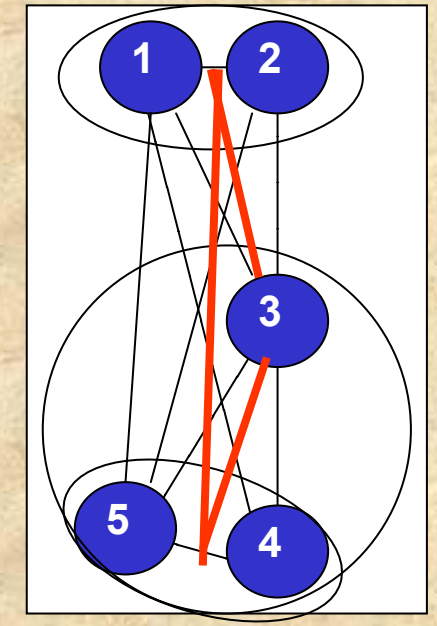
Distancia	Cluster
0	1,2,3,4,5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4,5	(1, 2), (3, 4, 5)
7,8	(1, 2, 3, 4, 5)

$$d_{(12)3} = \text{med}[d_{13}, d_{23}] = (6+5)/2 = 5{,}5$$

$$d_{(45)3} = \text{media}[d_{43}, d_{53}] = 4{,}5$$

$$d_{(12)(45)} = \text{med}[d_{14}, d_{24}, d_{15}, d_{25}] = 36/4 = 9$$

Método del centroide



Objeto	1	2	3	4	5
1					
2	2				
3	6	5			
4	10	9	4		
5	9	8	5	3	

Matriz de distancias

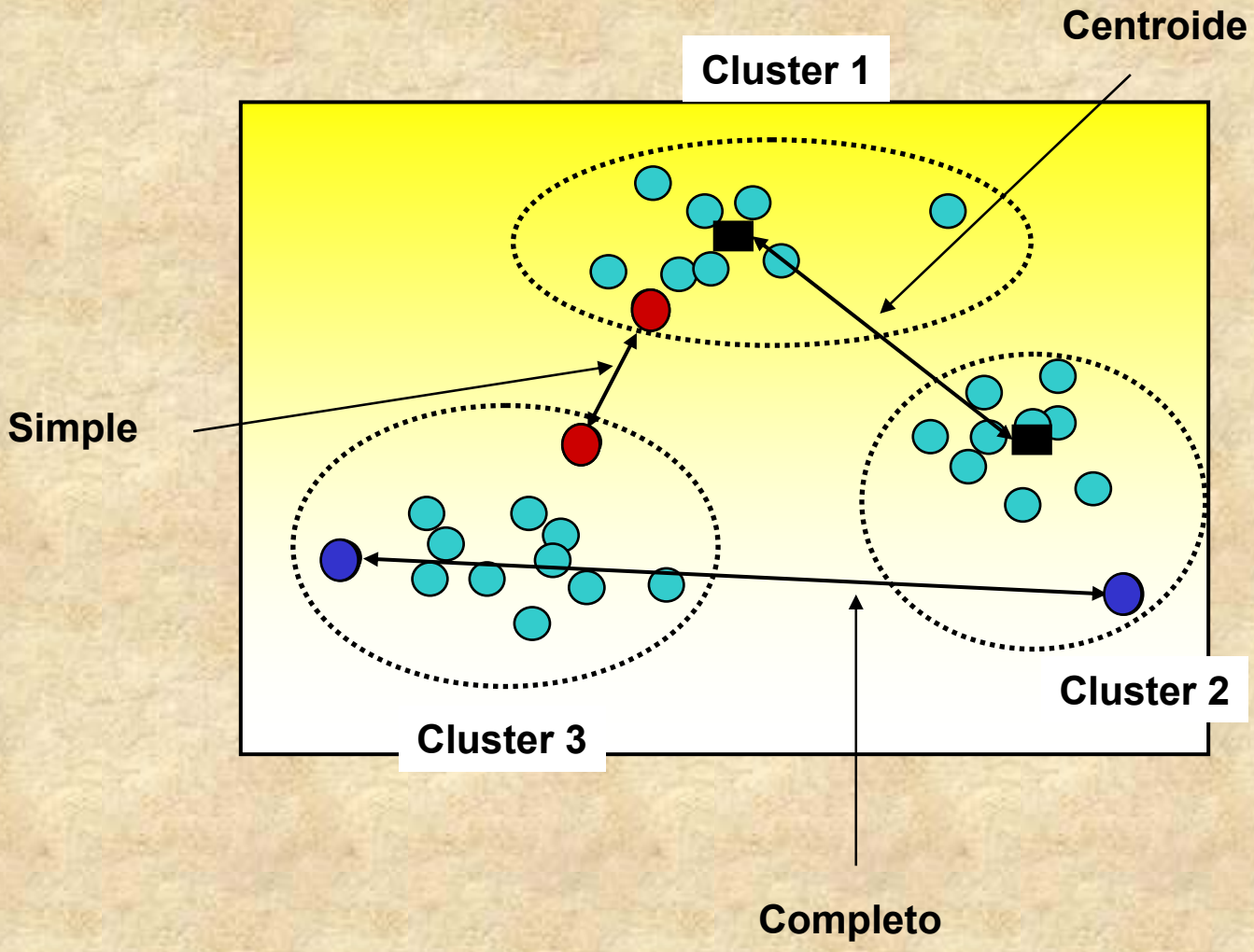
$$d_{(12)3} = d[c_{12},3] = 4$$

$$d_{(45)3} = d[c_{45},3] = 3{,}75$$

$$d_{(12)(45)} = d[c_{12},c_{45}] = 9$$

Distancia	Cluster
0	1,2,3,4,5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
3,75	(1, 2), (3, 4, 5)
6	(1, 2, 3, 4, 5)

ALGORITMOS DE MÉTODOS JERÁRQUICOS



ALGORITMOS DE MÉTODOS JERÁRQUICOS

5. Método de Ward: sigue un proceso algo diferente de los anteriores, se parte de una medida global de la heterogeneidad de una agrupación: **W**.

$$\mathbf{W} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(\mathbf{x_{ig}} - \bar{\mathbf{x}}_{\mathbf{g}}\right)\left(\mathbf{x_{ig}} - \bar{\mathbf{x}}_{\mathbf{g}}\right)'$$

- Se comienza suponiendo que cada dato forma un grupo (G=n y **W**=0). A continuación se unen los elementos que produzcan un incremento mínimo de **W** (los 2 objetos con mínima distancia), obteniendo n-1 grupos (1 con 2 objetos y el resto con 1).Repetimos hasta unir todos los puntos.
- El método equivale a unir, en cada etapa, los grupos tales que:

$$\min \frac{n_a n_b}{n_a + n_b} (\bar{x}_a - \bar{x}_b)'(\bar{x}_a - \bar{x}_b)$$

Método de Ward

Ejemplo: Se parte de 3 grupos con 2 variables medidas en cada objeto

Grupo	Tamaño	$\overline{X}_1$	$\overline{X}_2$
A	4	2	1
B	10	6	4
C	7	4	2

Método de Ward

Se calcula la expresión anterior entre cada par de grupos

A-B	[(4.10)/(4+10)]25=71,43
A-C	[(4.7)/(4+7)]5=12,73
B-C	[(10.7)/(10+7)]8=32,94

Se agrupan A y C, obteniendo un cluster de 11 objetos (AC) y otro de 10 (B)

$$(\overline{X}_A - \overline{X}_B) = \begin{bmatrix} 2-6 \\ 1-4 \end{bmatrix} = \begin{bmatrix} -4 \\ -3 \end{bmatrix} \Rightarrow \begin{bmatrix} -4 & -3 \end{bmatrix} \begin{bmatrix} -4 \\ -3 \end{bmatrix} = 16+9=25$$

$$(\overline{X}_A - \overline{X}_C) = \begin{bmatrix} 2-4 \\ 1-2 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \end{bmatrix} \Rightarrow \begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} -2 \\ -1 \end{bmatrix} = 4+1=5$$

$$(\overline{X}_B - \overline{X}_C) = \begin{bmatrix} 6-4 \\ 4-2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = 4+4=8$$

ALGORITMOS JERÁRQUICOS CON VARIABLES

Se construye una matriz de similitudes entre variables y se aplica un algoritmo jerárquico de clasificación.

Las distancias entre variables se miden con:

$$d_{jh} = \sqrt{\frac{1 - r_{jh}}{2}} \Leftrightarrow \text{usa correlaciones} \Rightarrow \text{sólo para v. continuas}$$

$$d_{jh} = 1 - \sqrt{\frac{\chi^2}{n}} \Leftrightarrow \text{usa coef. de contingencia} \Rightarrow \text{para v. binarias}$$

Para que las v. no dependan de las unidades deben estandarizarse.

Permite disminuir la dimensión del problema desde el punto de vista descriptivo.

ALGORITMOS DE MÉTODOS NO JERÁRQUICOS

Método clásico de las k-medias: queremos clasificar n elementos de p-variables en G grupos prefijados.

1. Seleccionamos G puntos como centros iniciales (aleatoria)
2. Calculamos las distancias de cada punto al centro. Asignamos el punto al grupo de centro más cercano. Al introducir un nuevo elemento se recalcula el centro.
3. Definir un criterio de optimalidad y comprobar si reasignando alguno de los puntos mejora el criterio.
4. Si no es posible mejorar el criterio de optimalidad, terminar.

ALGORITMOS DE MÉTODOS NO JERÁRQUICOS

Criterio de optimalidad

Minimizar la suma de cuadrados dentro de los grupos para todas las variables.

$$SCDG = \sum_{g=1}^{G}\sum_{j=1}^{p}\sum_{i=1}^{n_g}\left(x_{ijg} - \bar{x}_{jg}\right)^2$$

$$\min(SCDG) = \min \operatorname{tr}(\mathbf{W});\ \mathbf{W} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(\mathbf{x_{ig}} - \bar{\mathbf{x}}_{\mathbf{g}}\right)\left(\mathbf{x_{ig}} - \bar{\mathbf{x}}_{\mathbf{g}}\right)'$$

Equivale a minimizar la suma ponderada de las varianzas de las variables en los grupos o minimizar las distancias al cuadrado entre los puntos y sus centros de grupos.

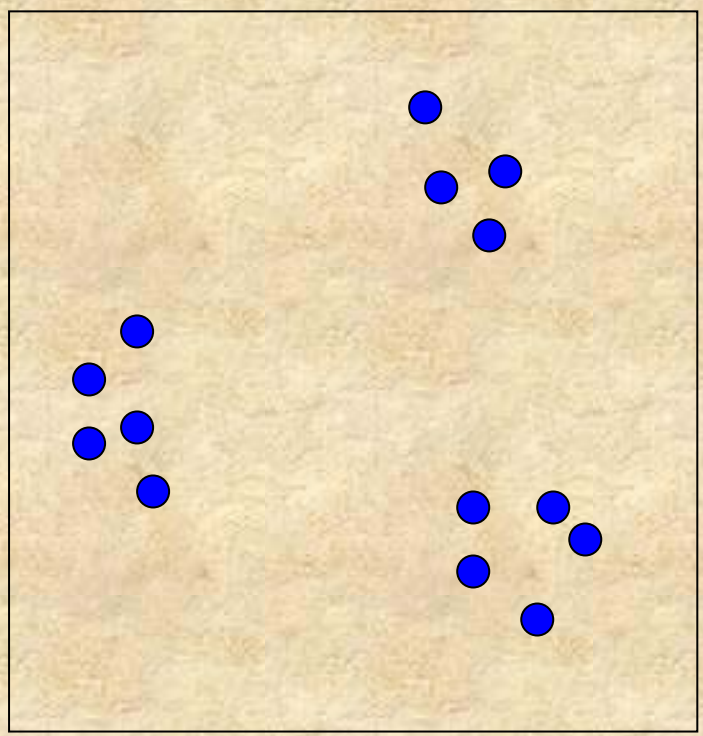
ALGORITMOS DE MÉTODOS NO JERÁRQUICOS

El algoritmo de k-medias

- Parte de una asignación inicial, permite mover sólo un elemento en cada iteración y termina al no poder reducir más la tr(W).
- Conviene repetir el algoritmo con asignaciones iniciales diferentes.
- Estandarizar las variables si están en distintas unidades (ya que el criterio varía con los cambios de escala).
- Minimizar la distancia euclídea conduce a grupos esféricos y supone que las v. son cuantitativas.
- Si existen muchas v. atributos $\Rightarrow$ métodos jerárquicos.

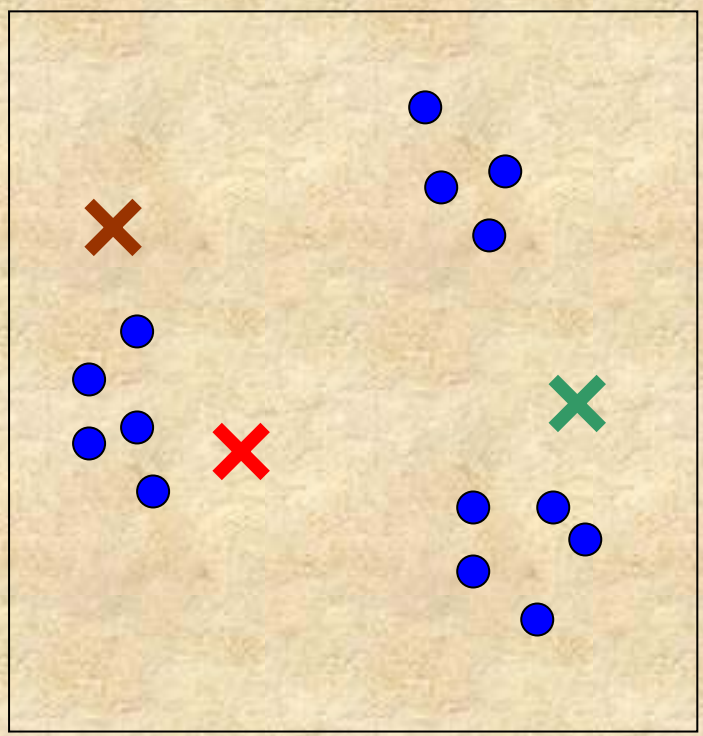
K-medias

supongamos que K=3



K-medias

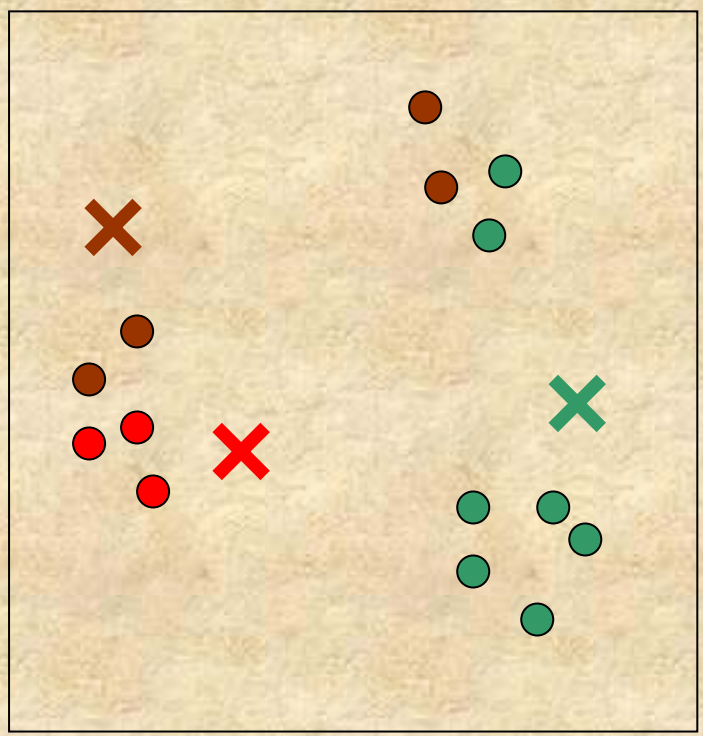
- Comenzamos con una posición aleatoria del centroide.



Iteración = 0

K-medias

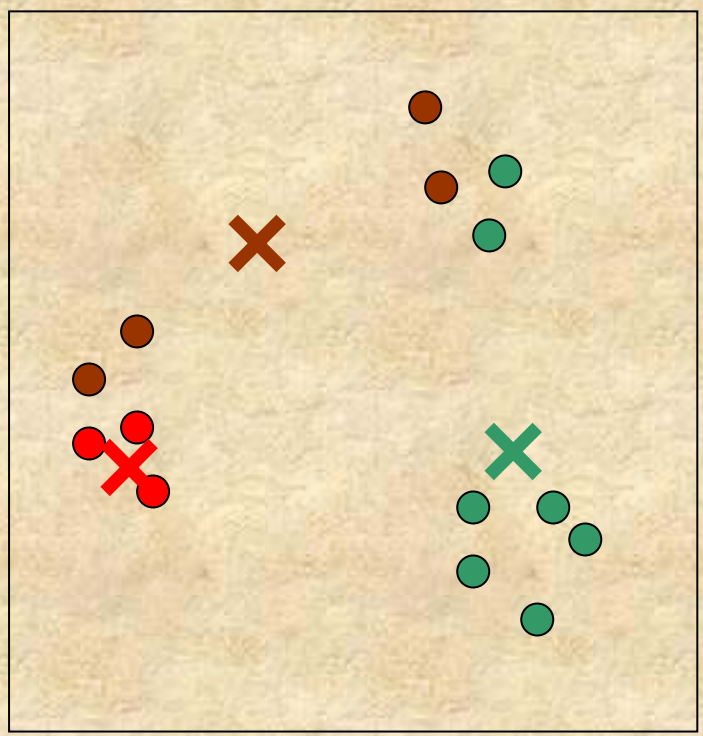
- Comenzamos con una posición aleatoria del centroide.
- Asignamos cada observación al centroide más próximo.



Iteración = 1

K-medias

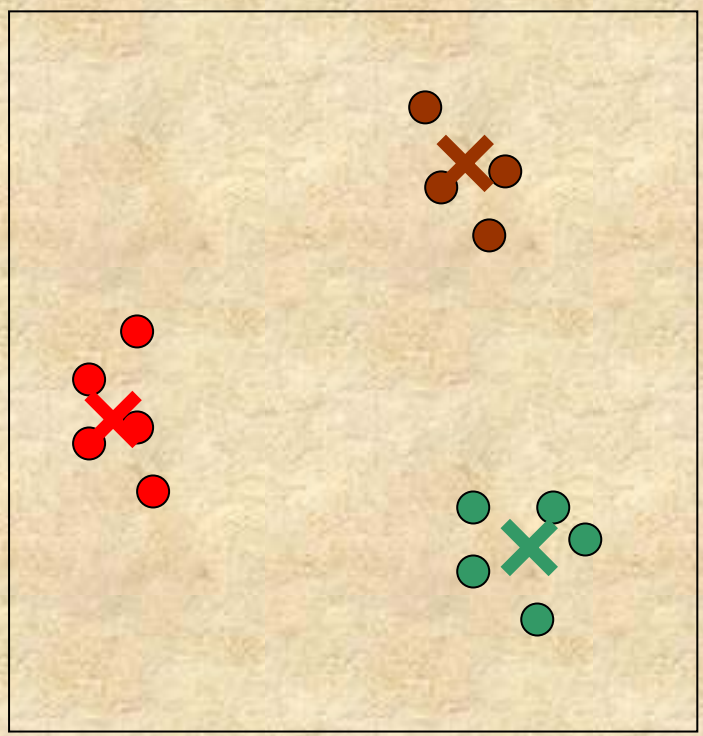
- Comenzamos con una posición aleatoria del centroide.
- Asignamos cada observación al centroide más próximo.
- Mover el centroide al centro de los puntos asignados



Iteración = 2

K-medias

- Comenzamos con una posición aleatoria del centroide.
- Asignamos cada observación al centroide más próximo.
- Mover el centroide al centro de los puntos asignados
- Iterar hasta terminar con mínima distancia



Iteración = 3

ELECCIÓN DE MEDIDAS DE DISTANCIA Y SIMILARIDAD

Los criterios para decidir qué objeto se incluye o no en un conglomerado se utilizan matrices de distancias o similaridades entre los pares de objetos.

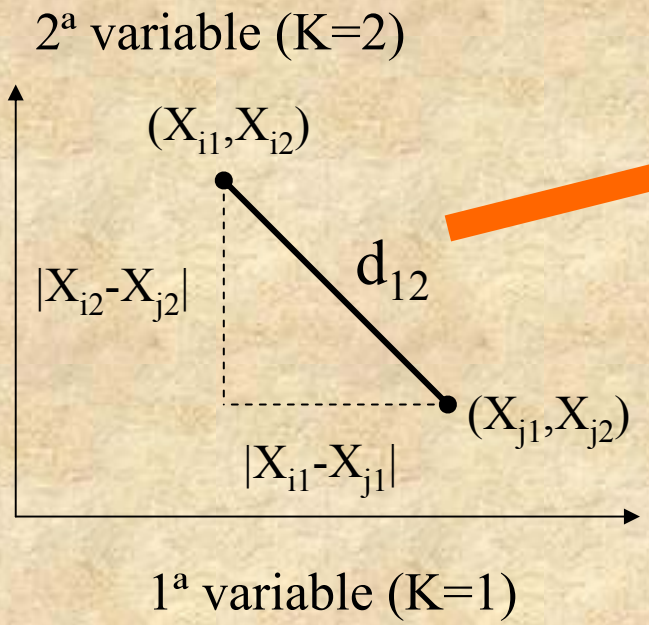
Las más empleadas para variables cuantitativas son las distancias euclídea, euclídea al cuadrado, “city block” y la correlación.

Las más empleadas para variables binarias son la distancia euclídea junto con el coeficiente de Jaccard.

La más empleada para variables cualitativas es la chi-cuadrado.

ELECCIÓN DE MEDIDAS DE DISTANCIA Y SIMILARIDAD

Euclídea



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

Donde:

d_{ij} representa la distancia entre los casos i y j

X_{iK} es el valor de la variable X_K para el caso i

X_{jK} es el valor de la variable X_K para el caso j

ELECCIÓN DE MEDIDAS DE DISTANCIA Y SIMILARIDAD

Euclídea al cuadrado

Empleada por defecto para datos de intervalo en especial cuando se agrupan casos

Medida recomendada en el algoritmo del centroide y de Ward

En la que más influyen las diferencias en las medidas.

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

ELECCIÓN DE MEDIDAS DE DISTANCIA Y SIMILARIDAD

De Manhattan o "city-block"

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

Correlación

Se aplica a v. continuas, y usa correlaciones (Pearson, Spearman o Kendall). También se emplea en métodos para jerárquizar variables.

$$d_{jh} = \sqrt{\frac{1 - r_{jh}}{2}}$$

ELECCIÓN DE MEDIDAS DE DISTANCIA Y SIMILARIDAD

Coeficiente de Jaccard

Conocido como razón de similaridad, se aplica a v. binarias.

		Objeto j: 1	Objeto j: 0
Objeto i	1	a	b
	0	c	d

$$s_{ij} = \frac{a}{a+b+c}; \qquad d_{ij} = \sqrt{2(1-s_{ij})}$$

Chi-cuadrado

$$d_{jh} = \sqrt{\chi^2} \Leftrightarrow \text{usa coef. de contingencia} \Rightarrow \text{para v. binarias}$$

OBTENCIÓN DE CONGLOMERADOS

A la elección de métodos algoritmos y distancias le sigue la obtención de los conglomerados.

Antes de interpretar los resultados hay que decidir el **número** adecuado de éstos

Si el método es **jerárquico**: se puede elegir el número de conglomerados adecuado **posteriormente** al análisis.

Si el método es **no jerárquico**: la elección del número de conglomerados adecuado es **previa** a la ejecución del análisis.

OBTENCIÓN ALGEBRAICA

Número de grupos

Para seleccionar el número de grupos aconsejable se usa el cociente:

$$F = \frac{SCDG(G) - SCDG(G+1)}{SCDG(G+1)/(n-G-1)}$$

El valor obtenido se compara con el de una $F_{p\ ;\ p(n-G-1)}$ para ver si el cociente es significativo. Si lo es se introduce un grupo más.

Una regla empírica es introducir un grupo más si F>10.

OBTENCIÓN ALGEBRAICA

Número de grupos

Para seleccionar el núme
cociente:

Suma de Diferencias al Cuadrado entre G+1 grupos

$$F = \frac{SCDG(G)}{SCDG(G+1)/(n-G-1)}$$

El valor obtenido se compara con el de una $F_{p\ ;\ p(n-G-1)}$ para ver si el cociente es significativo. Si lo es se introduce un grupo más.

Una regla empírica es introducir un grupo más si F>10.

OBTENCIÓN ALGEBRAICA

Número de grupos

$$\sum_{j=1}^{2}\sum_{i=1}^{60}\left(Z_{ij}-\bar{Z}_j\right)^2 = 60+60 = 120$$

Nº de	SCDG(G)	F	p-valor
1	120		
2	34,0656	143,789	1,44428E-7
3	14,0512	79,766	4,56299E-8
4	12,0345	9,216	0,00019934
5	3,5424	129,452	1,220450E-7
6	2,9582	10,467	0,00007106

OBTENCIÓN ALGEBRAICA

Número de grupos

Para valores estandarizados:

$$\sum_{j=1}^{2}\sum_{i=1}^{60}\left(Z_{ij}-\bar{Z}_j\right)^2 = 60+60 = 120$$

$$\sum_{G=1}^{2}\sum_{j=1}^{2}\sum_{i=1}^{30}\left(Z_{ijg}-\bar{Z}_{jg}\right)^2 =$$

$$=((12{,}26694+10{,}75995)+(0{,}0887619+10{,}94991))$$

30 datos de la v1 G1

30 datos de la v2 G1

30 datos de la v1 G2

30 datos de la v2 G2

Nº de	SCDG(
1	120
2	34,0656
3	14,0512
4	12,0345
5	3,5424
6	2,9582

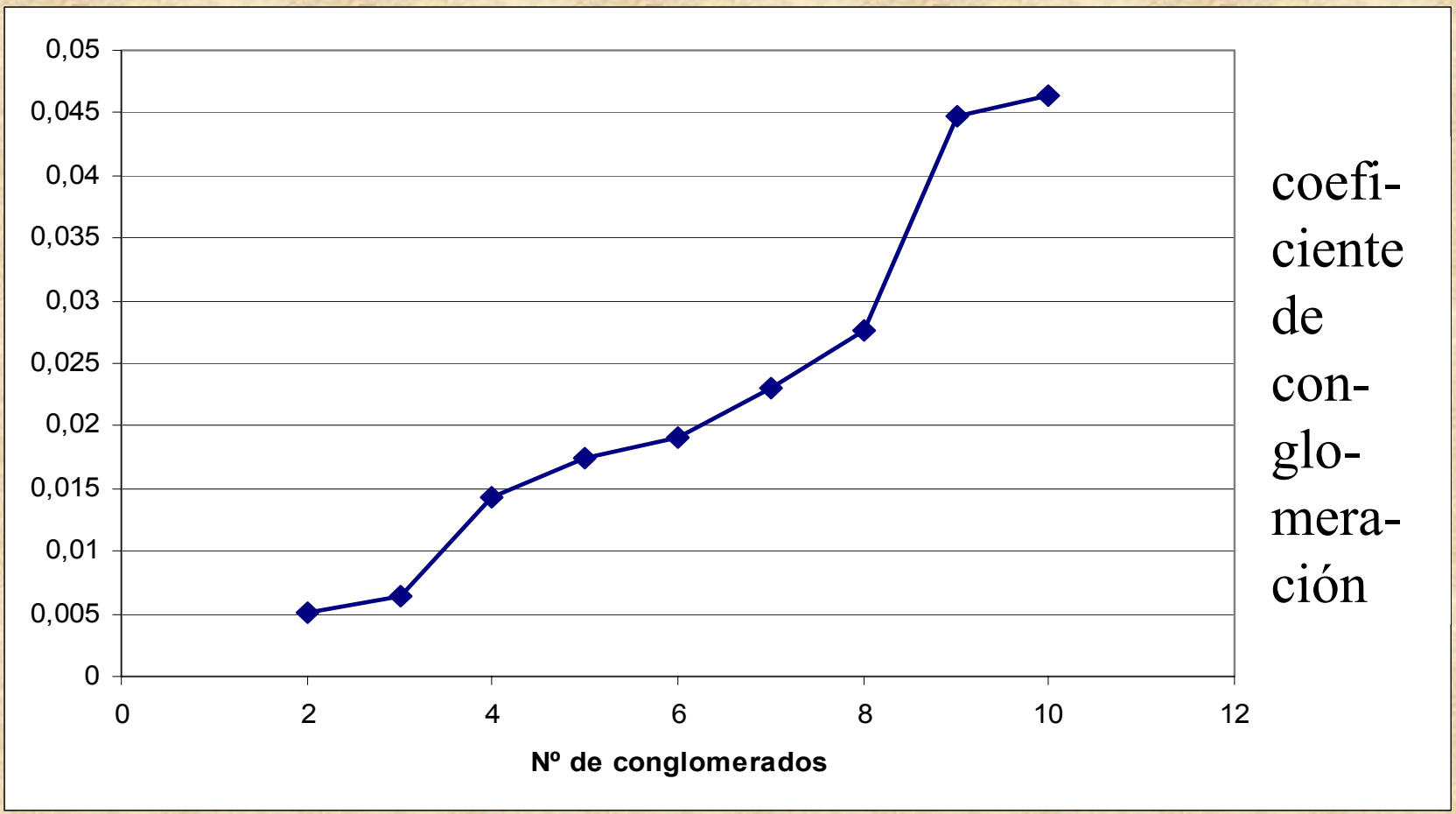
OBTENCIÓN ALGEBRAICA

Número de grupos

Se pueden utilizar también las variaciones entre los coeficientes de aglomeración (valor nco. que propicia la unión de objetos)

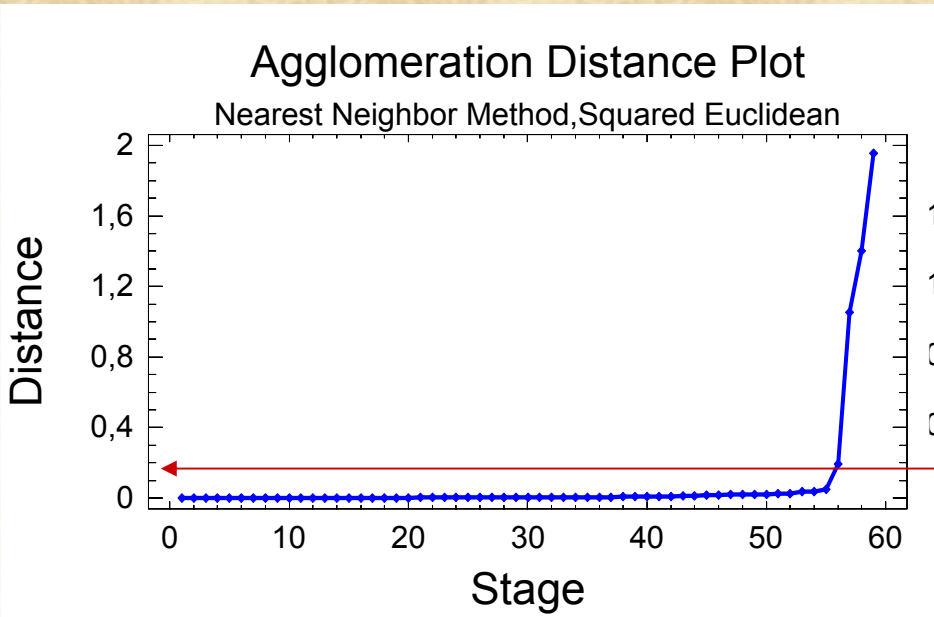
Nº de G	Coeficiente de aglomeración	Diferencia de coeficientes	Cambio porcentual
1	4,9937E-05		
2	5,0261E-05	3,244E-07	0,00502614
3	6,4276E-05	1,4015E-05	0,00642759
4	0,00014356	7,9288E-05	0,0143564
5	0,00017413	3,0562E-05	0,0174126
6	0,00019057	1,6444E-05	0,019057
7	0,00023068	4,0112E-05	0,0230682
8	0,00027572	0,00004504	0,0275722
9	0,00044768	0,00017196	0,0447681
10	0,00046428	1,6596E-05	0,0464277

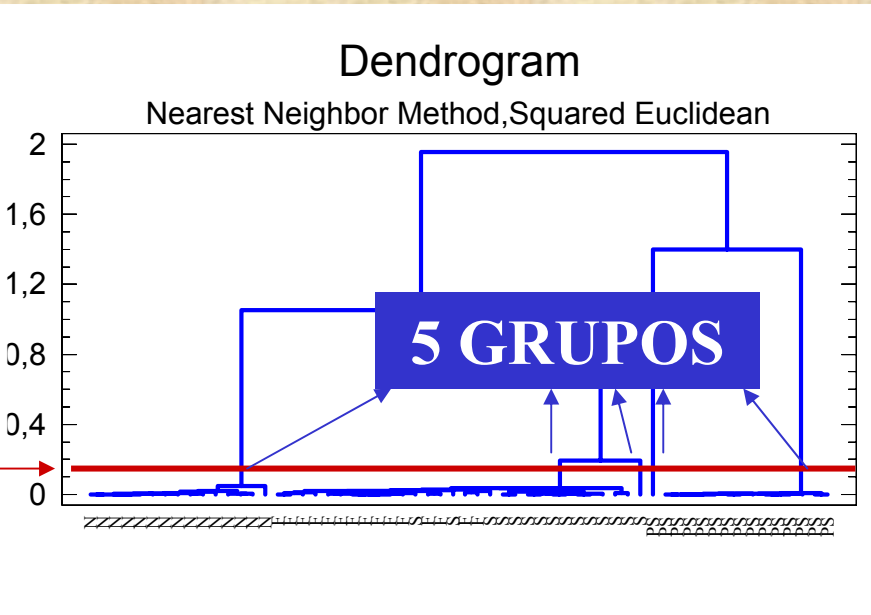
OBTENCIÓN ALGEBRAICA
Número de grupos



OBTENCIÓN GRÁFICA
Número de grupos

En el gráfico de distancias de aglomeración se elige la distancia para la cual el número de conglomerados es el más adecuado: se determina con el cambio brusco de pendiente





PRESENTACIÓN DE RESULTADOS

Métodos jerárquicos → Historial de agrupación

Métodos jerárquicos → Grupo de pertenencia

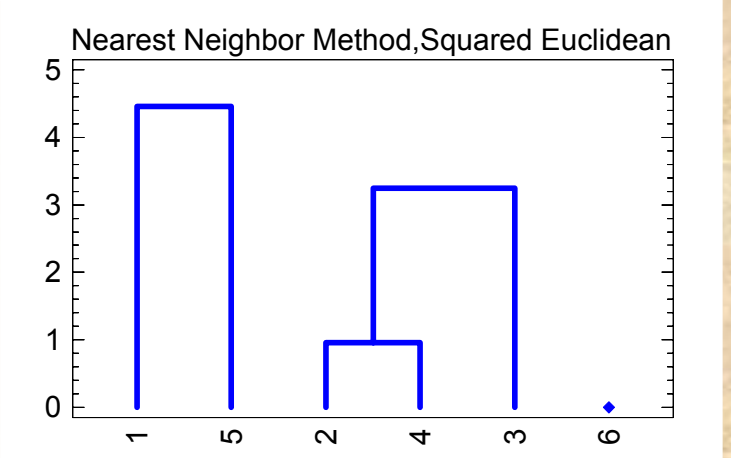
Métodos no jerárquicos → Centros de grupo

Métodos no jerárquicos → ANOVA

PRESENTACIÓN DE RESULTADOS

Historial de agrupación

Etapa	Conglomerado que se combina		Coeficientes	Etapa en que el conglomerado aparece por primera vez		Próxima etapa
	Conglom. 1	Conglom. 2		Conglom. 1	Conglom. 2	
1	2	4	0,963394	0	0	2
2	2	3	3,24342	1	0	0
3	1	5	4,45441	0	0	0



PRESENTACIÓN DE RESULTADOS

Historial de agrupación

Etapa	Conglomerado que se combina		Coeficientes	Etapa en que el conglomerado aparece por primera vez		Próxima etapa
	Conglom. 1	Conglom. 2		Conglom. 1	Conglom. 2	
1	2	4	0,963394	0	0	2
2	2	3	3,24342	1	0	
3	1	5	4,45441	0	0	

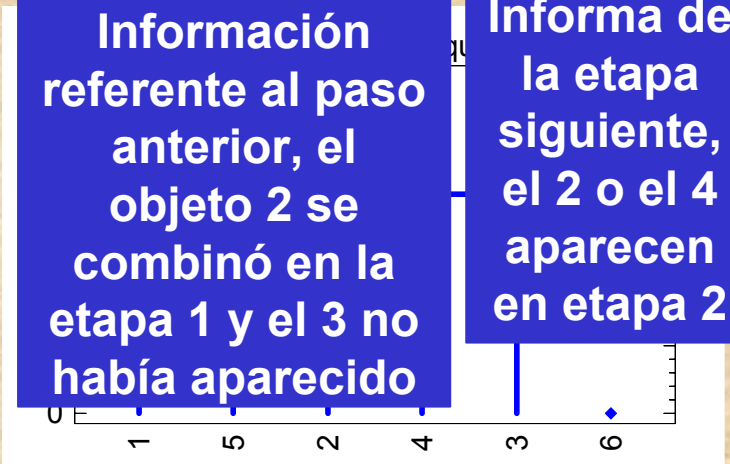
Figuran las etapas del análisis G-1

Objetos que se combinan en cada etapa

Medida de la distancia entre objetos

Información referente al paso anterior, el objeto 2 se combinó en la etapa 1 y el 3 no había aparecido

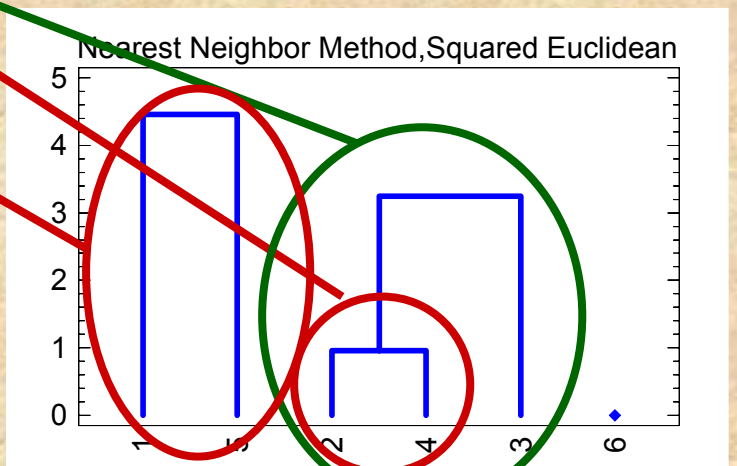
Informa de la etapa siguiente, el 2 o el 4 aparecen en etapa 2



PRESENTACIÓN DE RESULTADOS

Historial de agrupación

Etapa	Conglomerado que se combina		Coeficientes	Etapa en que el conglomerado aparece por primera vez		Próxima etapa
	Conglom. 1	Conglom. 2		Conglom. 1	Conglom. 2	
1	2	4	0,963394	0	0	2
2	2	3	3,24342	1	0	0
3	1	5	4,45441	0	0	0

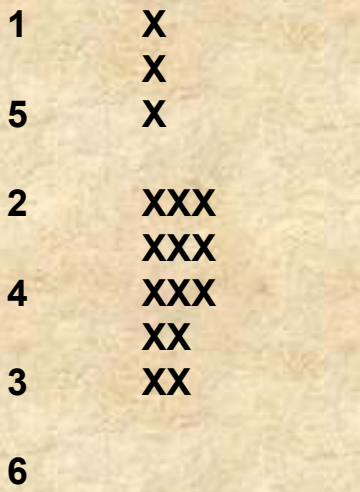


PRESENTACIÓN DE RESULTADOS

Gráfico de témpanos y Grupo de pertenencia

Número de Cluster

Distancia: Euclídea cuadrado



Objeto	Cluster
1	1
2	2
3	2
4	2
5	1
6	3

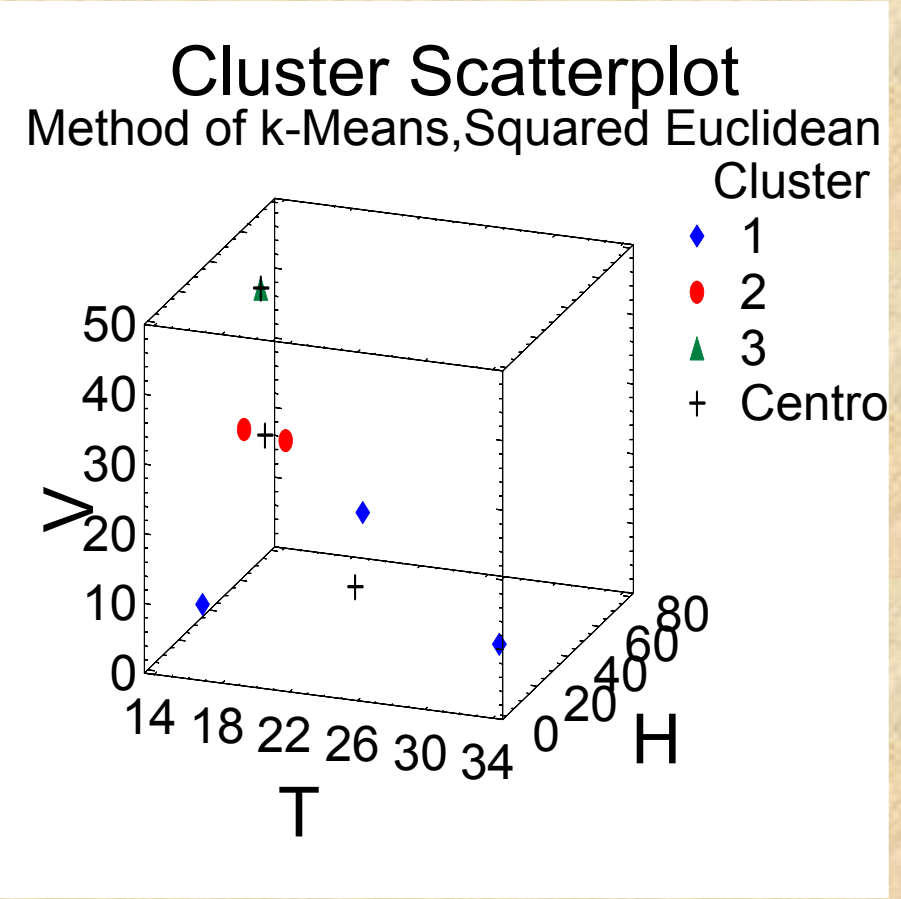
PRESENTACIÓN DE RESULTADOS

Tabla de centroides

G	nº obj.	%
1	3	50,00
2	2	33,33
3	1	16,67

Centroides

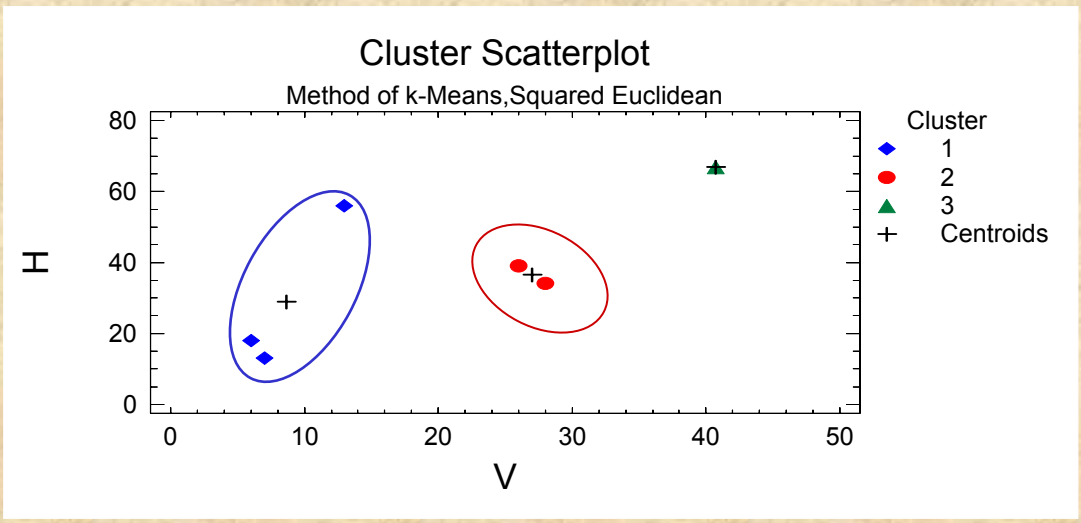
Cluster	T	H	V	P
1	23,0	29,0	8,67	1020,0
2	17,0	36,5	27,0	1012,0
3	14,0	67,0	40,7	1014,0



PRESENTACIÓN DE RESULTADOS

ANOVA

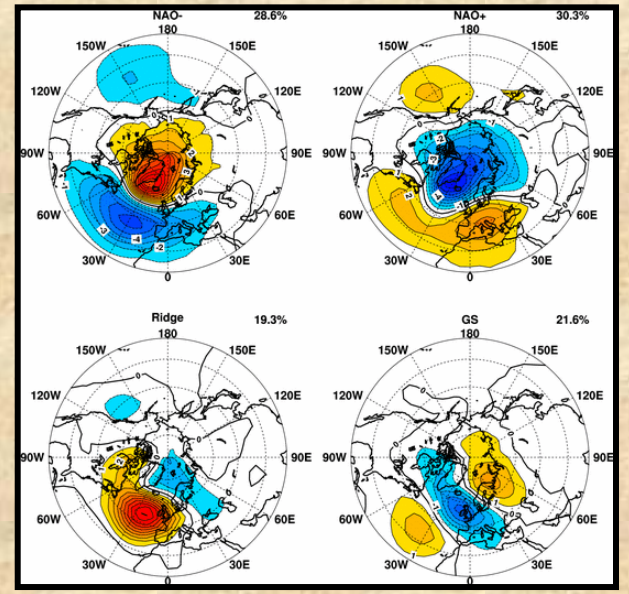
Con G	Conglomerado		Error		F	p-valor
	Media cuadrática	g.l.	Media cuadrática	g.l.		
T	39,75	2	56,6667	3	0,7	0,5624
P	41,6667	2	5,33333	3	7,81	0,0646
H	544,167	2	372,833	3	1,46	0,3608
V	455,871	2	10,2222	2	44,6	0,0059



Valores altos de F indican gran contribución de la variable a la diferenciación entre grupos

ULTIMAS APLICACIONES

Se buscan grupos de pacientes según el tiempo de estancia y coste de asistencia hospitalaria. Con más de 7200 enfermos se obtuvieron 25 grupos.



Estudios climáticos (2005):

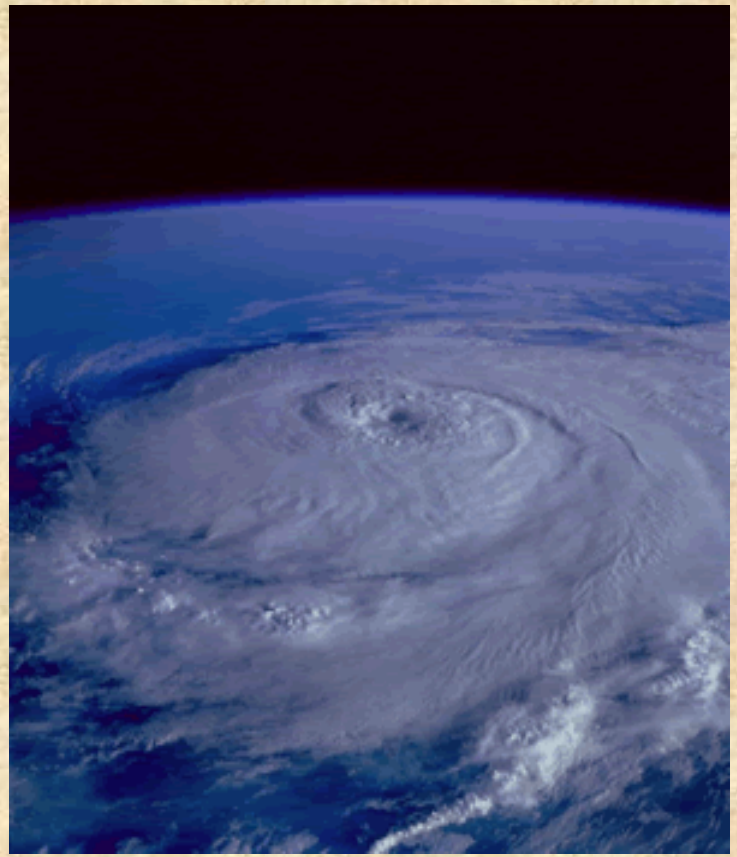
Se intentan caracterizar los inviernos en el Atlántico Norte mediante cuatro grupos de diferente régimen climático.

ULTIMAS APLICACIONES

Predicción de huracanes (2005) :

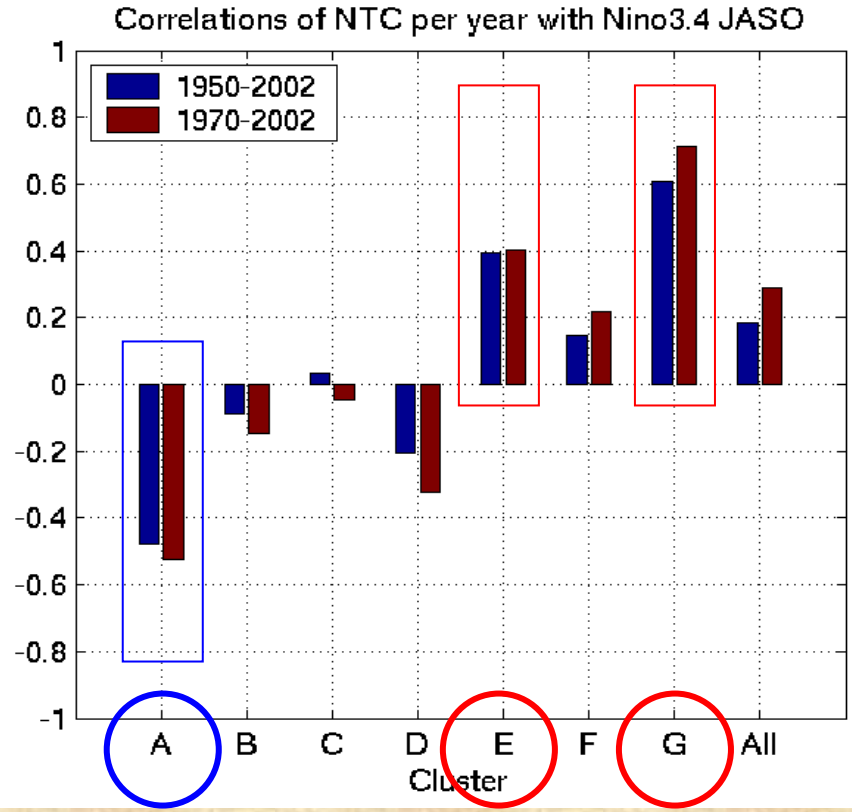
Se buscan trayectorias y comportamientos tipo de los huracanes del pacífico, empleando datos de los últimos 50 años.

Se relacionan los conglomerados encontrados con el fenómeno del Niño, según se muestra en la figura siguiente, para el NO del Pacífico.



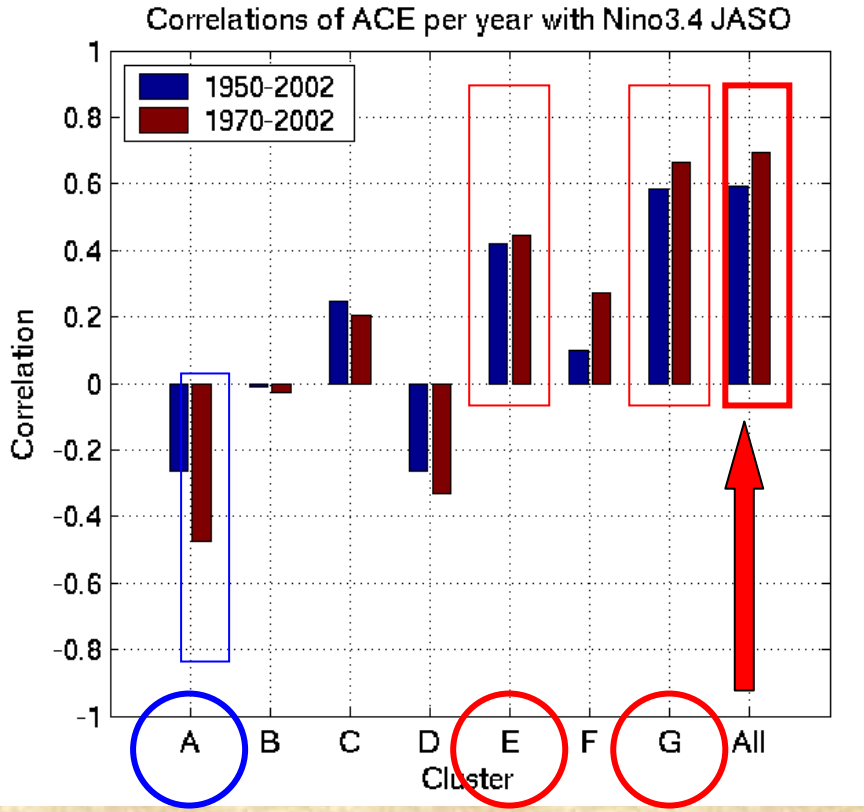
ULTIMAS APLICACIONES

NTC- Number of Tropical Cyclones



Total NTC per year is not significantly correlated with ENSO (e.g. Wang & Chan, 2002).

ACE – Accumulated Cyclone Energ



Total ACE has a well known relationship with ENSO (Camargo & Sobel, 2005).

Ejemplo:

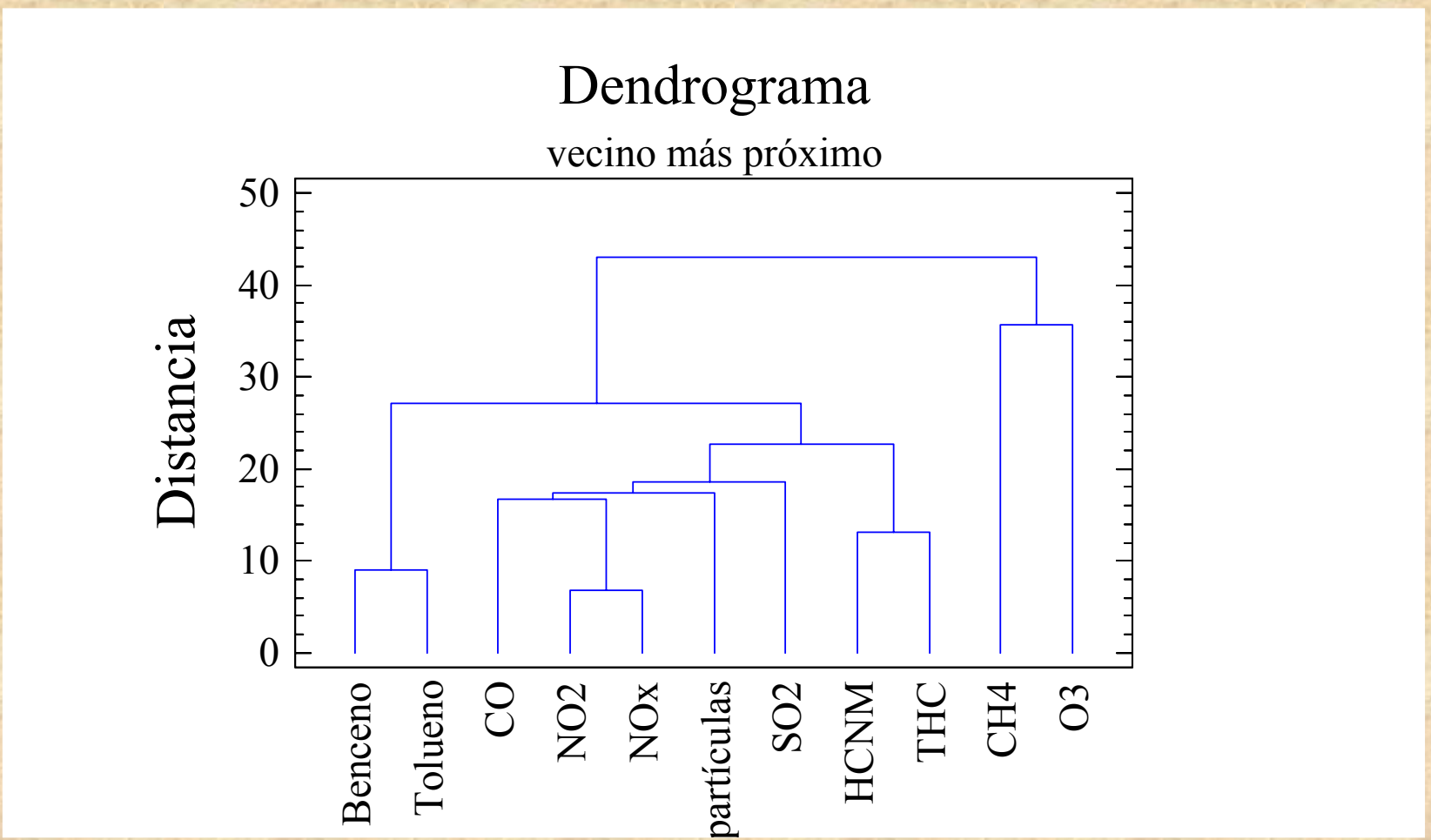
Se han obtenido 25 observaciones de diferentes variables en cinco lugares distintos.

Las variables medidas son concentraciones, por metro cúbico de aire, de diferentes agentes contaminantes:

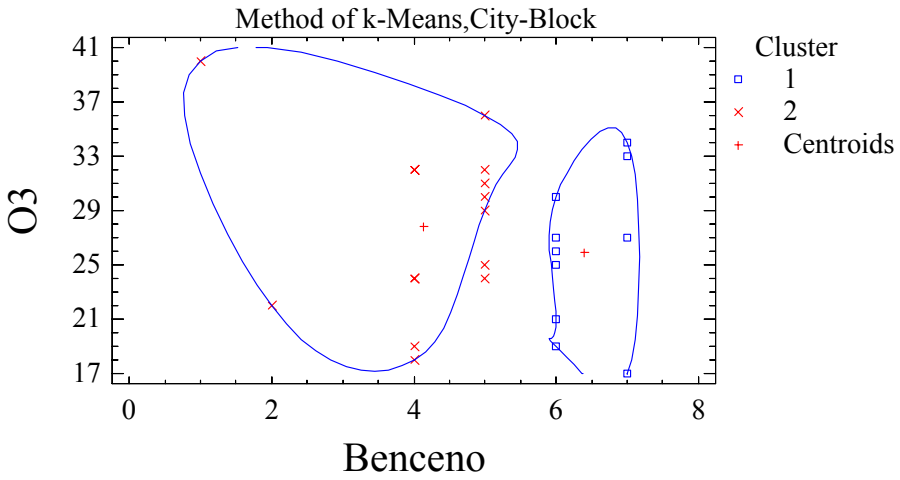
benceno, tolueno, ozono, CO, NO_2, NO_x, SO_2, CH_4, dos tipos de hidrocarburos y partículas suspendidas.

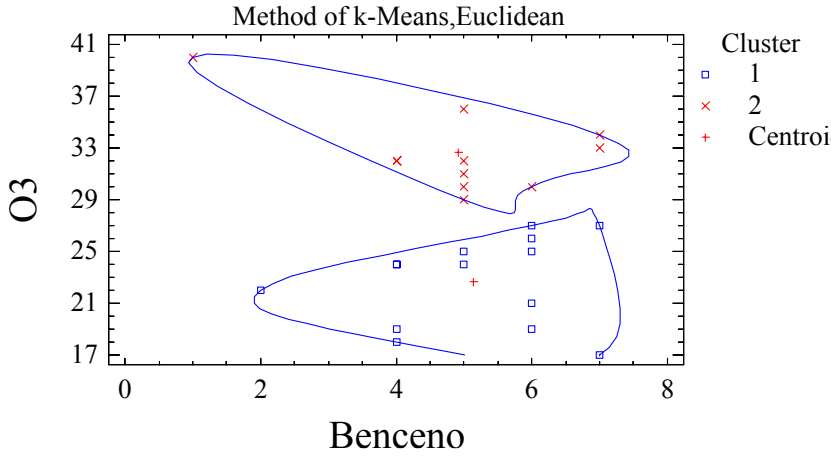
Veamos si se pueden agrupar las variables y las observaciones.

Ejemplo:

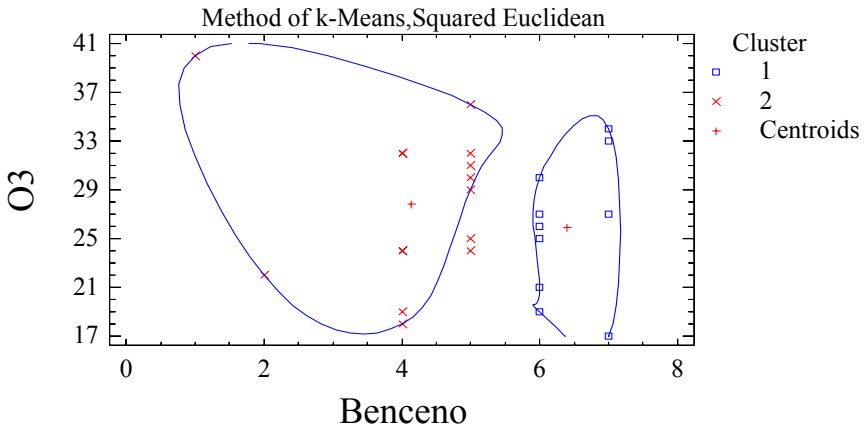

Dendrograma
vecino más próximo
Distancia
50
40
30
20
10
0
Benceno
Tolueno
CO
NO2
NOx
partículas
SO2
HCNM
THC
CH4
O3

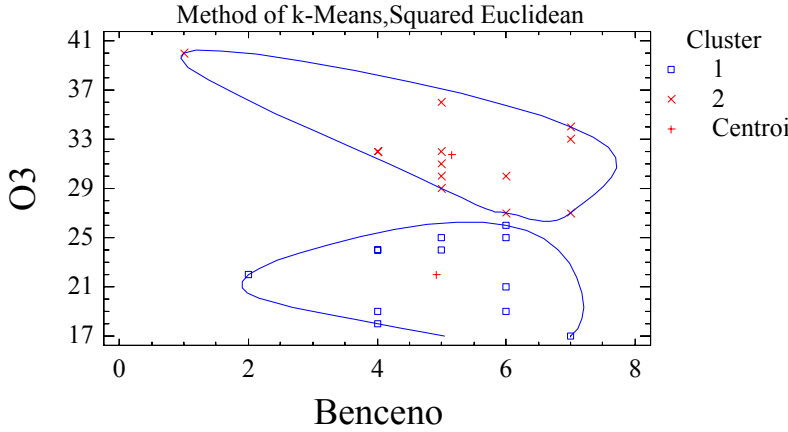
Ejemplo:



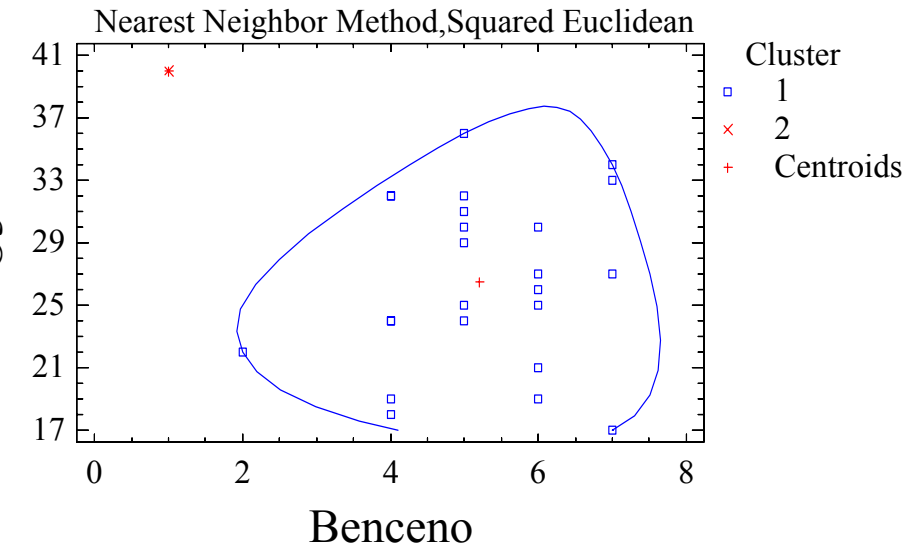


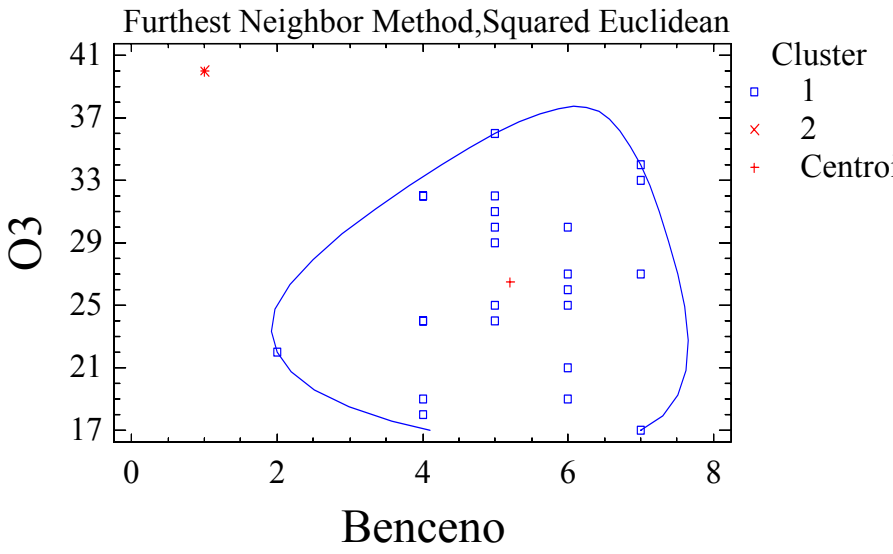
Sin estandarizar





Ejemplo:



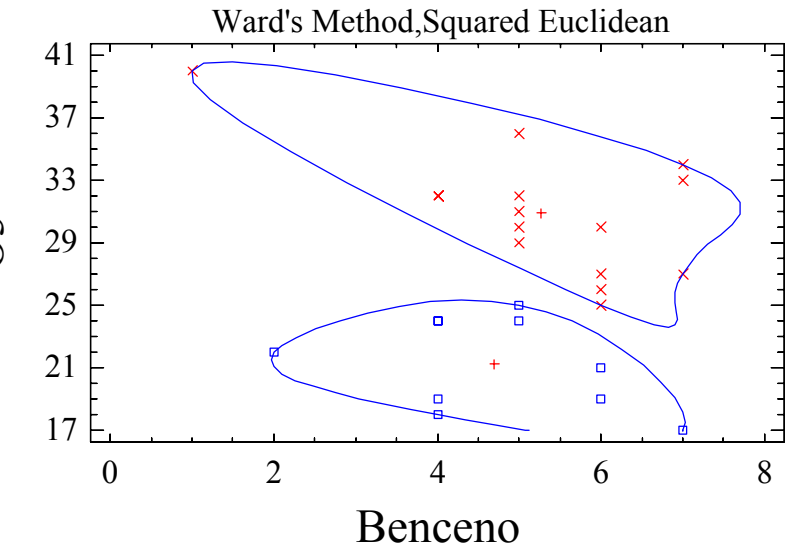


Con el resto de métodos y distancias (excepto Ward)

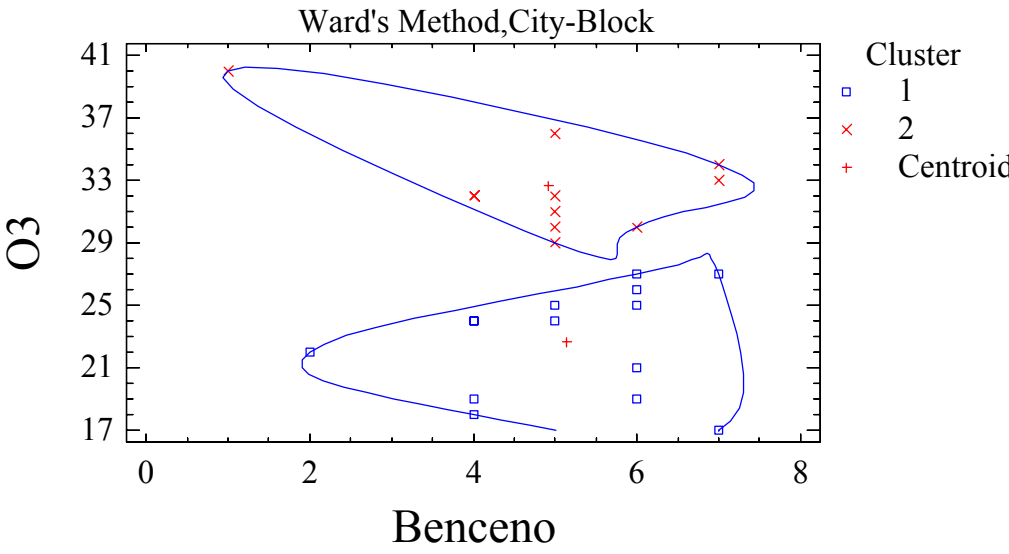
se obtienen estos mismos grupos

Ejemplo:

Estandarizados



Sin estandarizar



Ejemplo:

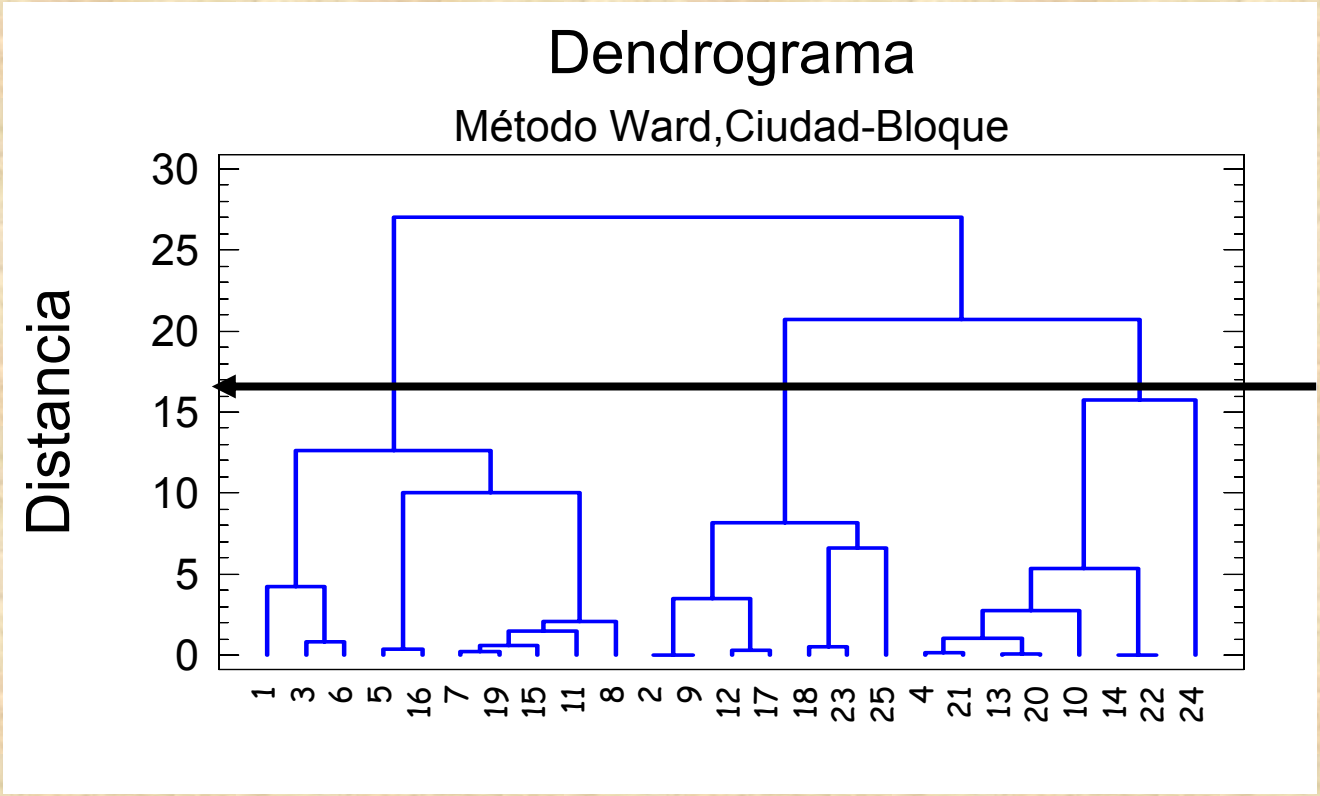
Sin embargo dos grupos pueden no ser adecuados. Para determinar el número de grupos calculamos la suma de distancias cuadradas:

Nº grupos	Tamaño	SCDG	F
2	10; 15	852,82	
3	**10; 7; 8**	**442,83**	**19,4**
4	10; 7; 7; 1	371,59	3,8
5	2; 7; 7; 8; 1	167,65	23,11

DISMINUCIÓN IMPORTANTE (F>10)

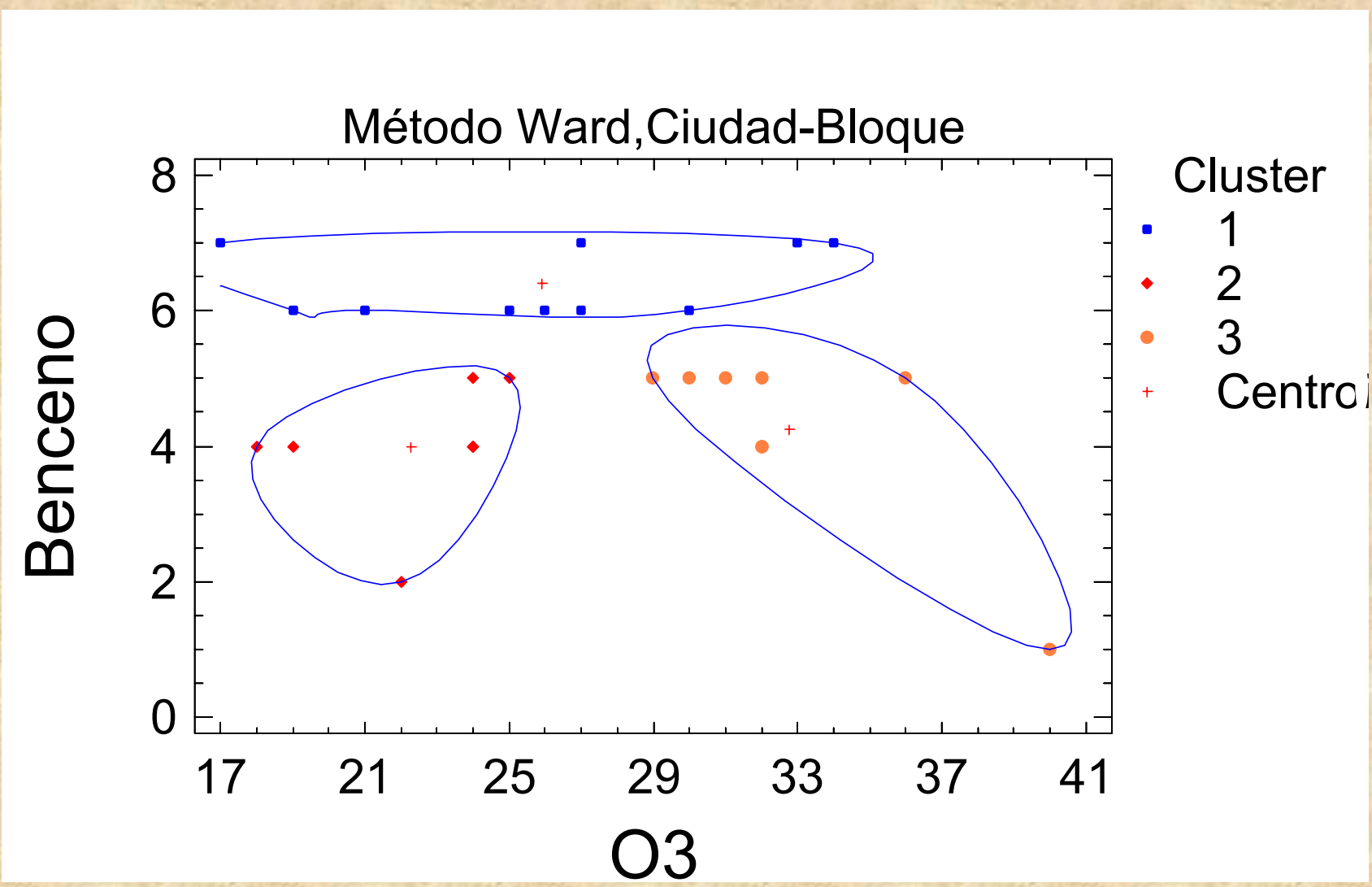
Ejemplo:

En el ordenador, nos ayudamos con el gráfico de distancias:



El instante en que la curva da un suave salto, indica el nº de grupos

Ejemplo:


Método Ward,Ciudad-Bloque
Benceno
O3
0
2
4
6
8
17
21
25
29
33
37
41
Cluster
1
2
3
Centroi