

ANÁLISIS DE TIPOLOGÍAS / CLUSTER / CONGLOMERADOS

Parte 1

CURSO: Estadística IV

CARRERA: Sociología

UNIVERSIDAD ALBERTO HURTADO

PROFESORA: CAROLINA AGUILERA

AYUDANTES: Miguel Tognarelli y

Vicente Díaz



CONTENIDOS DE LA CLASE

PROCEDIMIENTOS JERÁRQUICOS

APLICACIÓN EN R STUDIO DE UN MODELO



"el arte de encontrar grupos en los datos"
(Kaufman y Rousseeuw, 1990: 1)

ANÁLISIS DE TIPOLOGÍAS, CONGLOMERADOS, CLUSTER

pero.... la disminución del número de conglomerados suele ir acompañada de una pérdida no deseada de homogeneidad dentro de los conglomerados.

VARIAS TÉCNICAS: 2 GRANDES TIPOS

Técnicas analíticas multivariadas de clasificación de interdependencia. Lógica exploratoria de análisis

OBJETIVO

Agrupar datos (individuos, objetos o variables) en un número reducido de grupos, llamados "conglomerados".

QUE SE BUSCA CON LA AGRUPACIÓN

Los casos o variables que constituyen un conglomerado deber ser lo más similar posible entre sí (con respecto a un criterio de selección determinado previamente) y diferente respecto a los integrantes de los otros conglomerados.

PARSIMONIA

Obtención de aquella estructura de los datos más simple posible que represente agrupaciones homogéneas.

*"el arte de encontrar
grupos en los datos"*

(Kaufman y Rousseeuw,
1990: 1)

ANÁLISIS DE TIPOLOGÍAS, CONGLOMERADOS , CLUSTER

LOGICA DEL MODELO

Los casos se agrupan según su grado de proximidad mutua, lo que en la literatura se denomina distancia/similitud.

Existen diferentes formas de estimar cuán lejanas o cercanas están las observaciones entre sí.

Se busca lograr la máxima homogeneidad dentro de cada clúster, mientras se maximiza la heterogeneidad entre los grupos.

Britto et. al (2014)

"el arte de encontrar
grupos en los datos"

(Kaufman y Rousseeuw,
1990: 1)

ANÁLISIS DE TIPOLOGÍAS, CONGLOMERADOS , CLUSTER

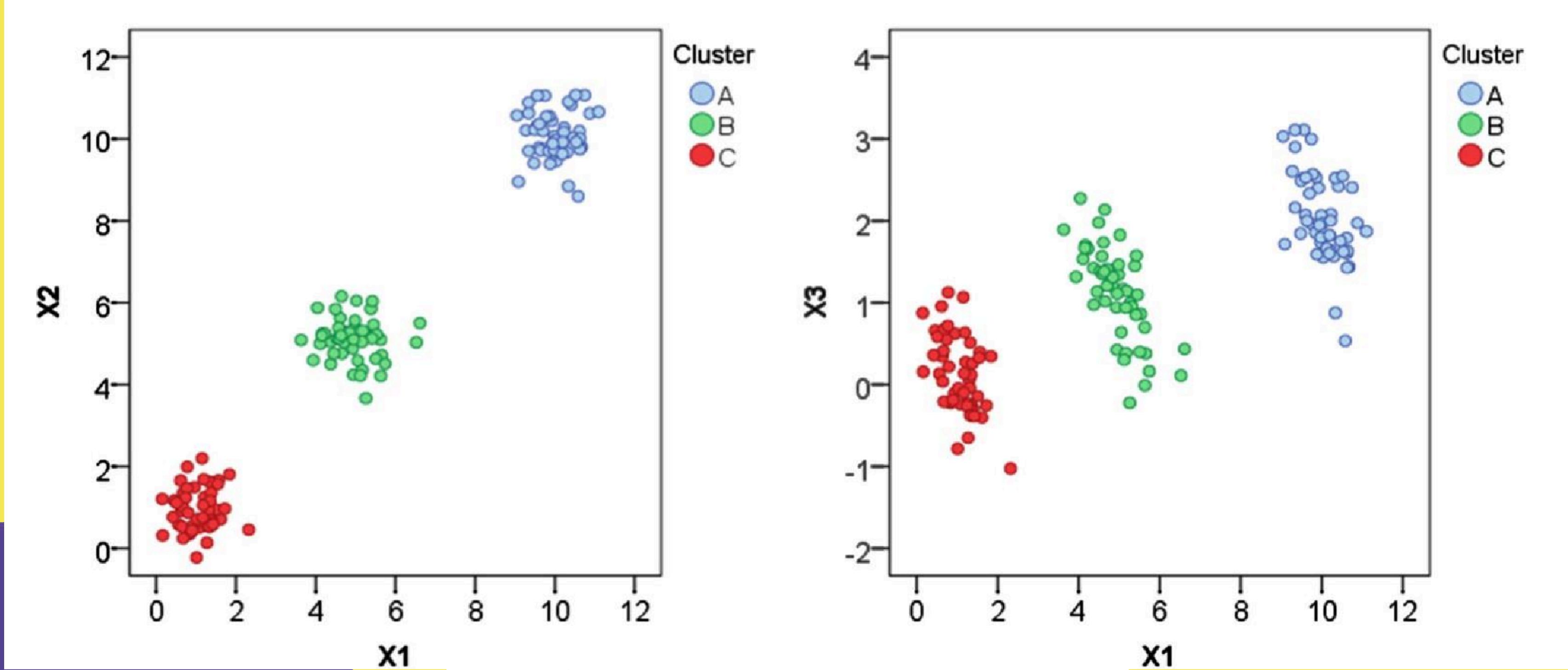
USOS

Cuatro los usos principales (Aldenderfer y Blashfield, 1984)

es lo más usado!

1. Desarrollar tipologías o clasificaciones de datos.
2. Buscar *esquemas conceptuales* útiles para agrupar entidades (o casos).
3. *Generalización de hipótesis* explorando datos.
4. La comprobación de hipótesis o el intento de *determinar si los tipos definidos a través de otros procedimientos* están de hecho presentes en una serie de datos.

Ejemplos de conglomerados “perfectos” (Britto et. al, 2014)



No hay variable dependiente e independiente,

2. Medidas de similitud entre los objetos

Todos los métodos asumen una manera específica de medir la distancia o similitud entre los casos.

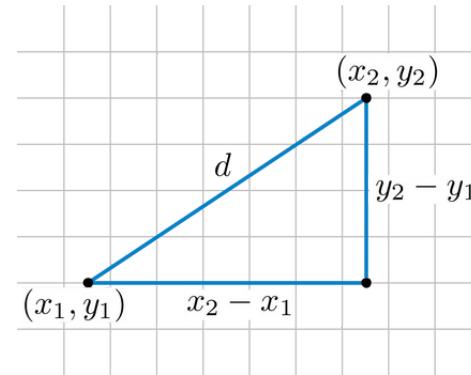
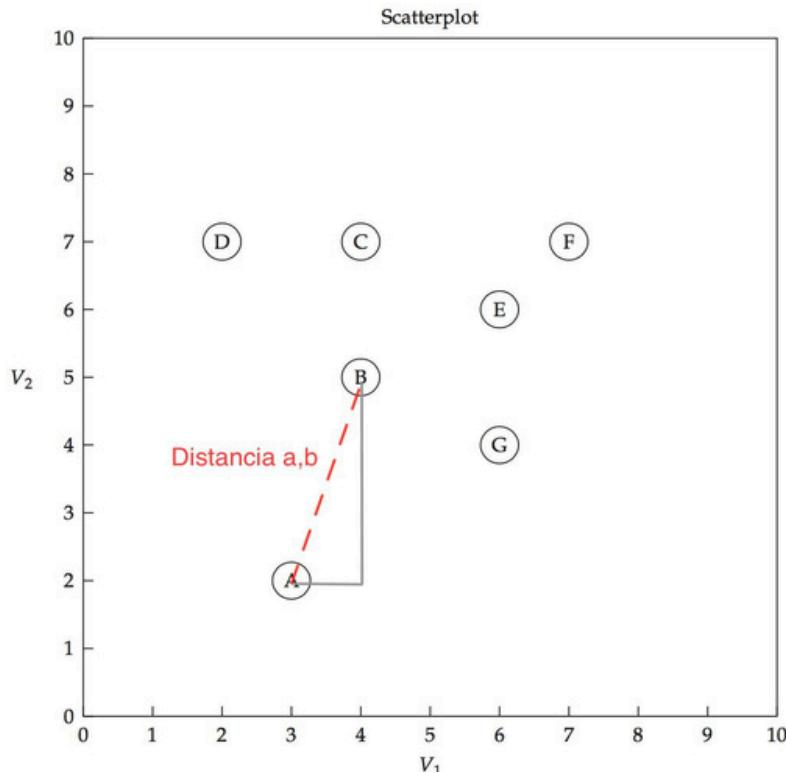
Hay diferentes formas de medir ello, según sea el tipo de variable y método -Variables numéricas: se usa algún tipo de distancia basada en la distancia “física” entre los puntos (distancia euclídea, distancia euclídea², otras

Variables ordinales: se usan medidas de similitud

2. Medidas de distancia entre los objetos

- **Medidas de distancia (similitud)**

1. **Distancia euclídea:** una de las más usadas en la investigación empírica, define similitud a través de una línea recta (cercanía en un plano). Variables métricas



$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$\text{Distancia } (a, b) = \sqrt{\sum (a_i - b_i)^2}$$

2. Medidas de distancia entre los objetos

- Medidas de distancia (similitud)**

- 1.**

2. Medidas de similitud entre los objetos

- **Medidas de distancia (similitud)**

...

2. **Distancia euclídea al cuadrado:** se desarrolló para hacer las cosas más rápidas; estadísticamente es lo mismo que la anterior
3. **Distancia de Manhattan:** se calcula como la suma de todas las diferencias absolutas entre las puntuaciones de los casos (la suma de los lados del triángulo, no sólo de la hipotenusa)
4. **Distancia de Mahalanobis:** mide distancia estandarizando las variables y considerando la existencia de correlación entre ellas (ajusta la varianza de cada grupo por la correlación entre sus miembros).

2. Medidas de similitud entre los objetos

- **Medidas de similitud (variables binarias u ordinales recodificadas).**

1. Existen diversos tipos y se basan en la idea de contar las coincidencias y las diferencias entre pares de casos. Por ejemplo

	V1	V2	V3	V4
Caso 1	1	1	0	0
Caso 2	0	1	1	1
Caso 3	1	1	0	1
Caso 4	0	0	0	1
Caso 5	1	1	1	0

2. Medidas de similitud entre los objetos

- **Medidas de similitud (variables binarias).**

1. Contar las coincidencias y las diferencias entre pares de casos. Por ejemplo

	X1	X2	X3	X4
Caso 1	1	1	0	0
Caso 2	0	1	1	1
Caso 3	1	1	0	1
Caso 4	0	0	0	1
Caso 5	1	1	1	0



	Caso 1		
	1	0	
Caso 2	1	1	2
	0	1	0



	Caso 1		
	1	0	
Caso 2	1	a	b
	0	c	d

$$\begin{aligned}a &= 1 \\b &= 2 \\c &= 1 \\d &= 0\end{aligned}$$

- Índice de Jaccard (1901):

$$\sqrt{1 - [a/(a + b + c)]}$$

- Coeficiente *simple matching* de Sokal y Michener (1958):

$$\sqrt{1 - [(a + d)/(a + b + c + d)]}$$

- Sokal y Sneath (1963):

$$\sqrt{1 - [a/(a + 2(b + c))]}$$

- Rogers y Tanimoto (1960):

$$\sqrt{1 - [(a + d)/(a + 2(b + c) + d)]}$$

- Dice (1945) o Sorenson (1948):

$$\sqrt{1 - [2a/(2a + b + c)]}$$

- Coeficiente de Hamann (Gower y Legendre, 1986)

$$\sqrt{1 - [(a - (b + c) + d)/(a + b + c + d)]}$$

- Ochiai (1957):

$$\sqrt{1 - \left[a / \sqrt{(a + b)(a + c)} \right]}$$

- Sokal y Sneath (1963):

$$\sqrt{1 - \left[ad / \sqrt{(a + b)(a + c)(d + b)(d + c)} \right]}$$

- Phi de Pearson g (Gower y Legendre, 1986):

$$\sqrt{1 - \left[(ad - bc) / \sqrt{(a + b)(a + c)(d + b)(d + c)} \right]}$$

- Coeficiente S2 de Gower y Legendre (1986):

$$\sqrt{1 - [a/(a + b + c + d)]}$$

		Caso 1	
		1	0
Caso 2	1	a	b
	0	c	d

$a = 1$

$b = 2$

$c = 1$

$d = 0$

- Índice de Jaccard (1901):
 $\sqrt{1 - [a/(a + b + c)]}$

	Caso 1		
	1		0
Caso 2	1	a	b
	0	c	d

$$\begin{aligned} a &= 1 \\ b &= 2 \\ c &= 1 \\ d &= 0 \end{aligned}$$

$$\sqrt{1 - \frac{a}{a + b + c}} = \sqrt{1 - \frac{1}{1 + 2 + 1}} = \sqrt{1 - \frac{1}{4}} = \sqrt{\frac{3}{4}} = 0,866$$

- Índice de Jaccard (1901):

$$\sqrt{1 - [a/(a + b + c)]}$$

		Caso 1	
	1	1	0
Caso 2	1	a	b
	0	c	d

$$\begin{aligned} a &= 1 \\ b &= 2 \\ c &= 1 \\ d &= 0 \end{aligned}$$

Cuadro 3.5.: Matriz de distancias euclídeas para los datos del ejemplo

```
> dist.binary(DatosCuadro3.5[,c("X1","X2","X3","X4")], method = 1, diag = TRUE, upper = FALSE)
```

	1	2	3	4	5
1	0.0000000				
2	0.8660254	0.0000000			
3	0.5773503	0.7071068	0.0000000		
4	1.0000000	0.8164966	0.8164966	0.0000000	
5	0.5773503	0.7071068	0.7071068	1.0000000	0.0000000

```
> dist.binary(DatosCuadro3.5[,c("X1","X2","X3","X4")], method = 2, diag = TRUE, upper = FALSE)
```

	1	2	3	4	5
1	0.0000000				
2	0.8660254	0.0000000			
3	0.5000000	0.7071068	0.0000000		
4	0.8660254	0.7071068	0.7071068	0.0000000	
5	0.5000000	0.7071068	0.7071068	1.0000000	0.0000000

SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos jerárquicos aglomerativos (“ascendentes”)

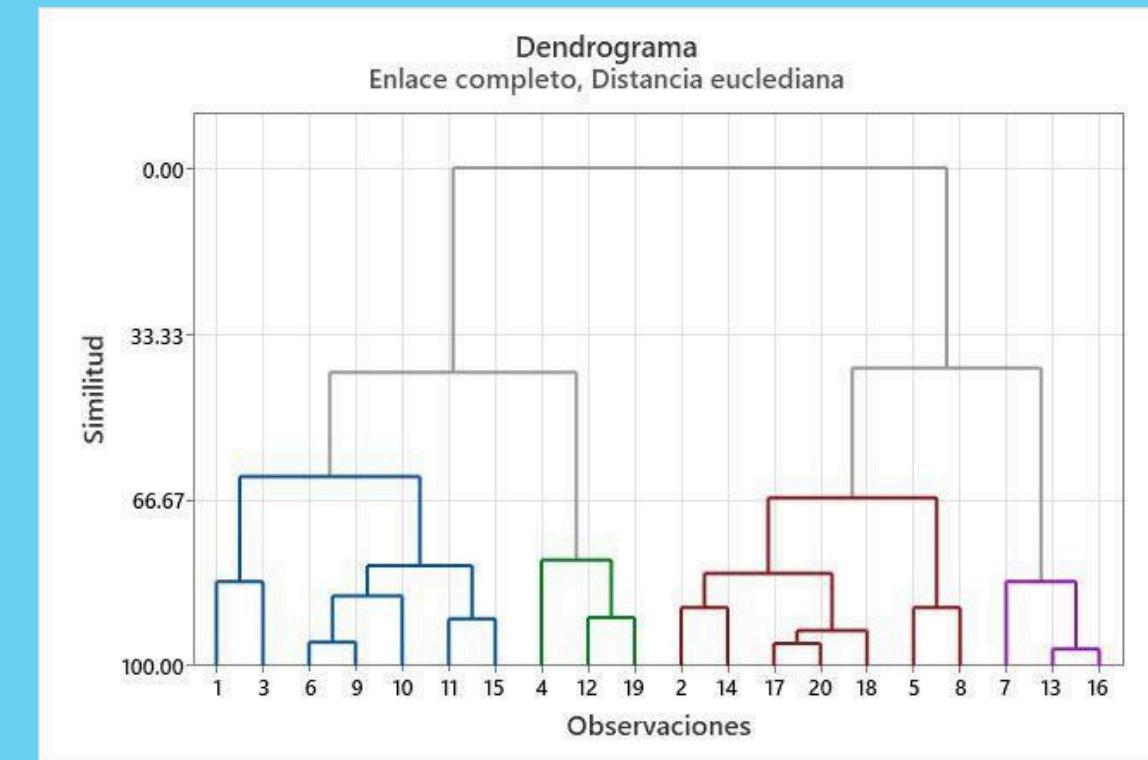
Se comienza con tantos conglomerados como objetos a clasificar (ya sean variables o casos).

2. Dos de los objetos se combinan en un único conglomerado.

3. Surge un nuevo conglomerado de la fusión de otros dos objetos adicionales, o de un tercer objeto que se une al conglomerado previamente formado por los dos objetos.

En cada paso se constituye un nuevo conglomerado, bien como resultado de la unión de dos objetos que permanecían todavía aislados (sin pertenecer a ningún conglomerado), o bien por la anexión de un objeto a un conglomerado ya constituido, o por la conjunción de dos conglomerados ya existentes.

El proceso de conglomeración concluye cuando se llega a un único conglomerado que reúne a todos los objetos.



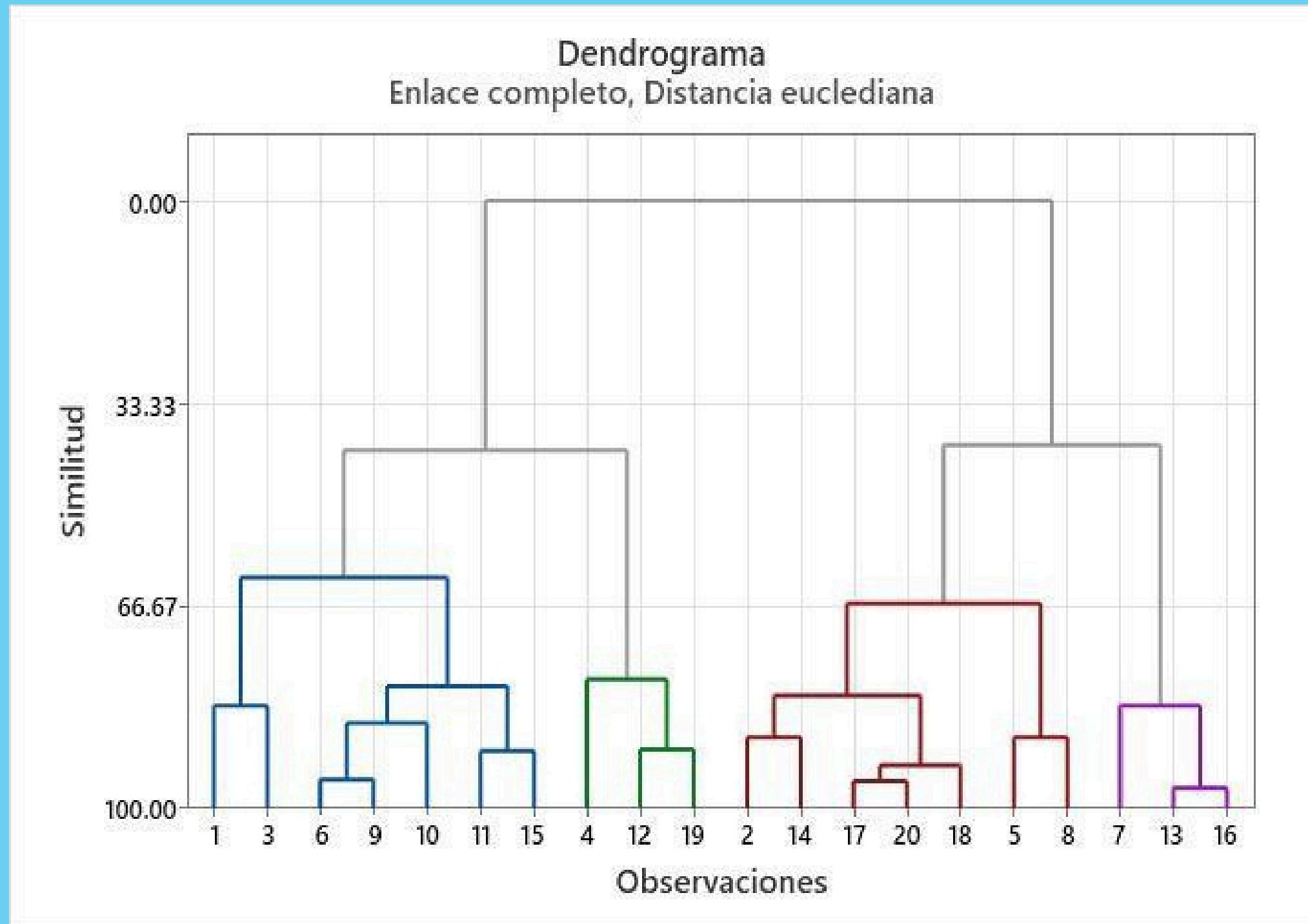
Ejemplo

SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos jerárquicos aglomerativos - dendrograma

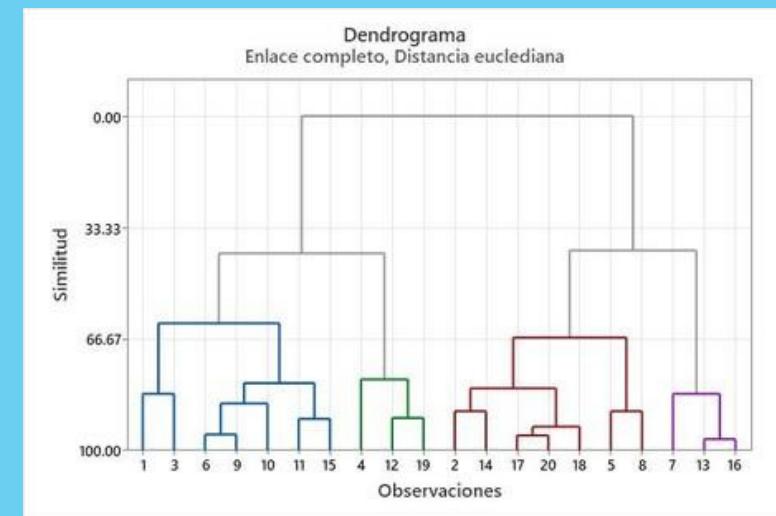
4 conglomerados (colores)

Si se cortara el dendrograma más arriba, entonces habría menos conglomerados finales, pero su nivel de similitud sería menor. Si se cortara el dendrograma más abajo, entonces el nivel de similitud sería mayor, pero habría más conglomerados finales.



SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos jerárquicos aglomerativos (“ascendentes”)



- Característica distintiva de este método: una vez que el conglomerado se ha constituido(dos objetos se han vinculado) no puede dividirse en etapas posteriores
- Tras cada nueva agrupación, se recalculan las distancias, de acuerdo con el algoritmo de clasificación y la medida de distancia/similaridad escogida para la formación de conglomerados.
- Cuando el análisis de conglomerados es de casos, el criterio que decide la pertenencia a los conglomerados se basa en la matriz de distancias o, en su caso, de similaridad, entre pares de casos.
- Si, por el contrario, se quiere agrupar variables, las medidas de distancia/similaridad se calculan entre pares de variables.

SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos jerárquicos divisivos (“descendentes”)

- Mucho menos usado.

Se comienza con un único conglomerado (con todos los objetos a clasificar (ya sean variables o casos)).



De forma gradual, se va disgregando ese gran conglomerado inicial, con la excepción de aquel objeto (caso o variable) que se halle más distante del promedio de los otros objetos en el conglomerado.



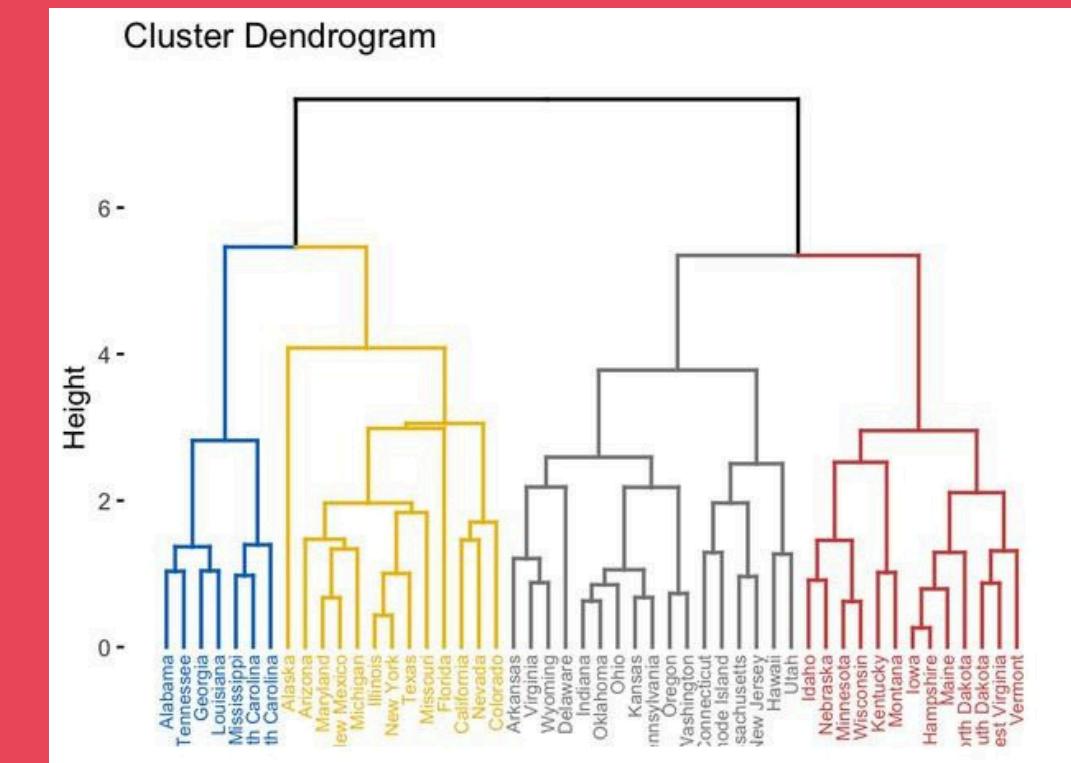
El conglomerado inicial se divide en dos conglomerados, entre los que se distribuyen los casos o variables. Éstos quedan ubicados en el conglomerado hacia el que estén más próximos.



Tras cada división de conglomerados se vuelven a calcular las distancias entre sus integrantes. Los objetos situados a mayor distancia del promedio del conglomerado se separan del mismo, ya sea constituyendo un nuevo conglomerado, ya añadiéndose al conglomerado hacia el que ahora se sitúen más “próximos”.



El proceso de división de conglomerados continúa iterativamente hasta que existan tantos conglomerados como objetos a clasificar.



MÉTODOS JERARQUICOS - OPTIMIZACIÓN



1. Método del centroide

2. Método del vecino más cercano

3. Método del vecino más lejano

4. Método de vinculación promedio

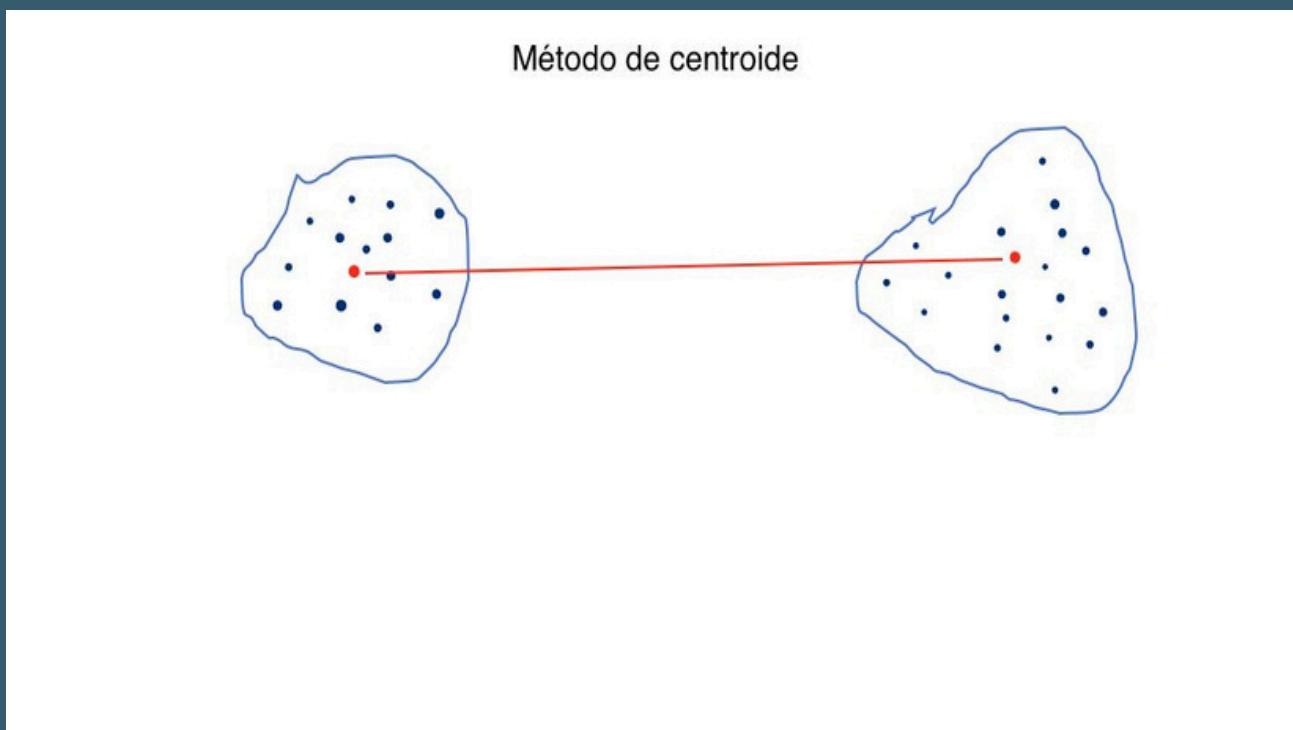
5. Método de Ward



3. Métodos de aglomeración jerárquicos: algoritmos

1. Método del centroide

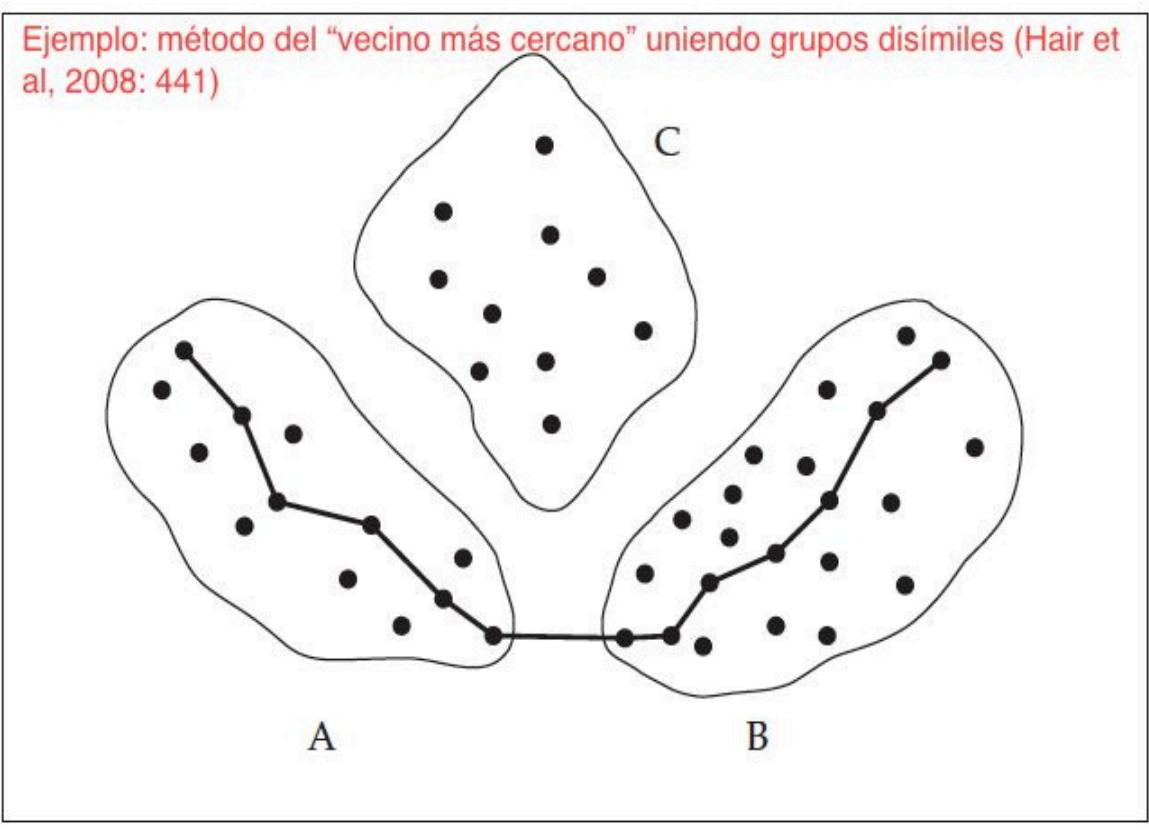
- Similitud = distancia entre centroides (valores promedios) de los grupos Cálculo:
- cada vez que los casos son reagrupados, un nuevo centroide se reclacula



3. Métodos de aglomeración jerárquicos: algoritmos

2. Método del “Vecino más cercano” (single-linkage)

- Similitud entre grupos = distancia más corta entre el objeto de uno y otro grupo
- Problema: riesgo de encadenamiento



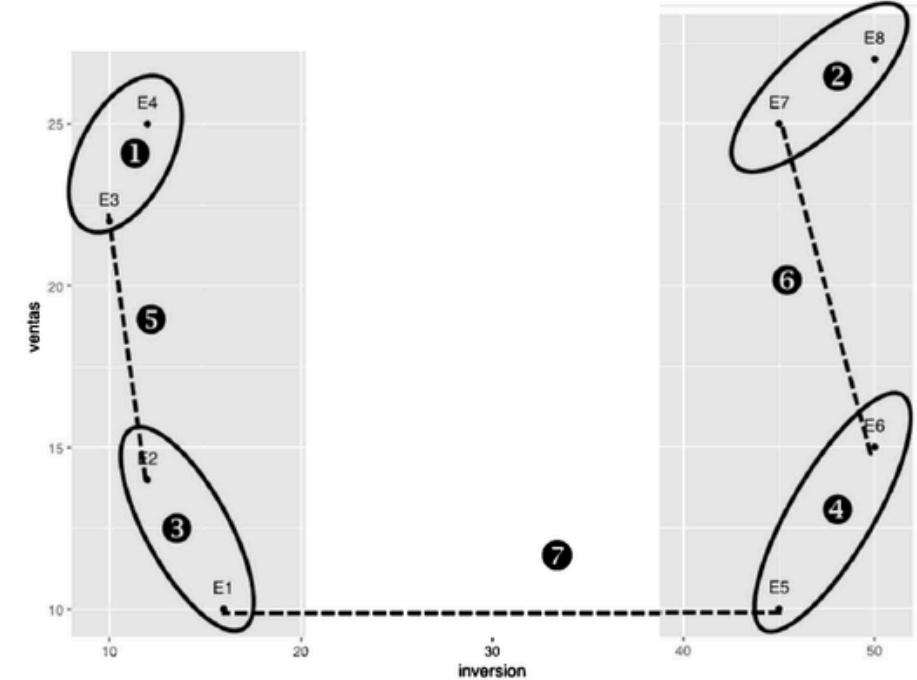
3. Métodos de aglomeración jerárquicos: algoritmos

2. Método del “Vecino más cercano” (single-linkage)

- Similitud entre grupos = distancia más corta entre el objeto de uno y otro grupo

- Problema: riesgo de encadenamiento

Figura 3.5.: Historial de conglomeración “vecino más cercano”



	height	merge.1	merge.2
1	13	-3	-4
2	29	-7	-8
3	32	-1	-2
4	50	-5	-6
5	68	1	3
6	125	2	4
7	841	5	6

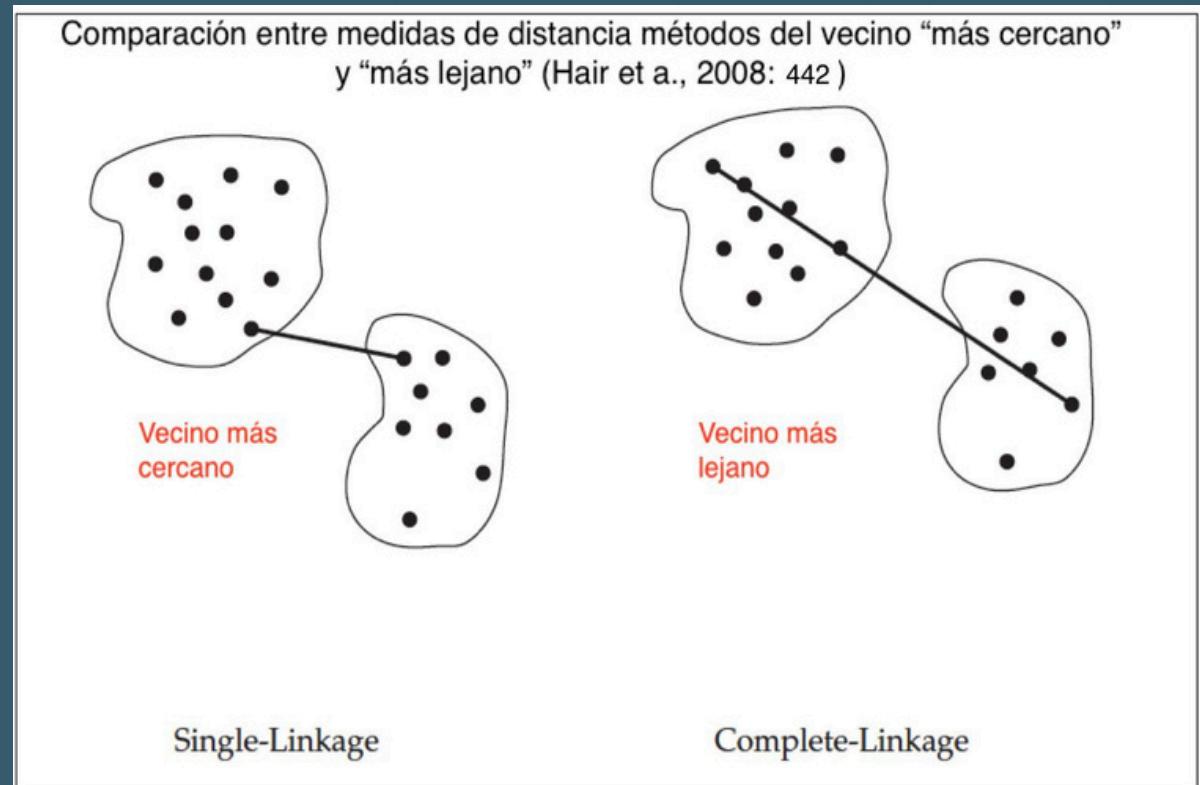
3. Métodos de aglomeración jerárquicos: algoritmos

3. Método del “Vecino más lejano” (complete-linkage)

- Similitud entre grupos:

Análisis de las distancias máximas entre las observaciones de cada grupo

Se unen grupos cuyas distancias “más lejanas” son las menores



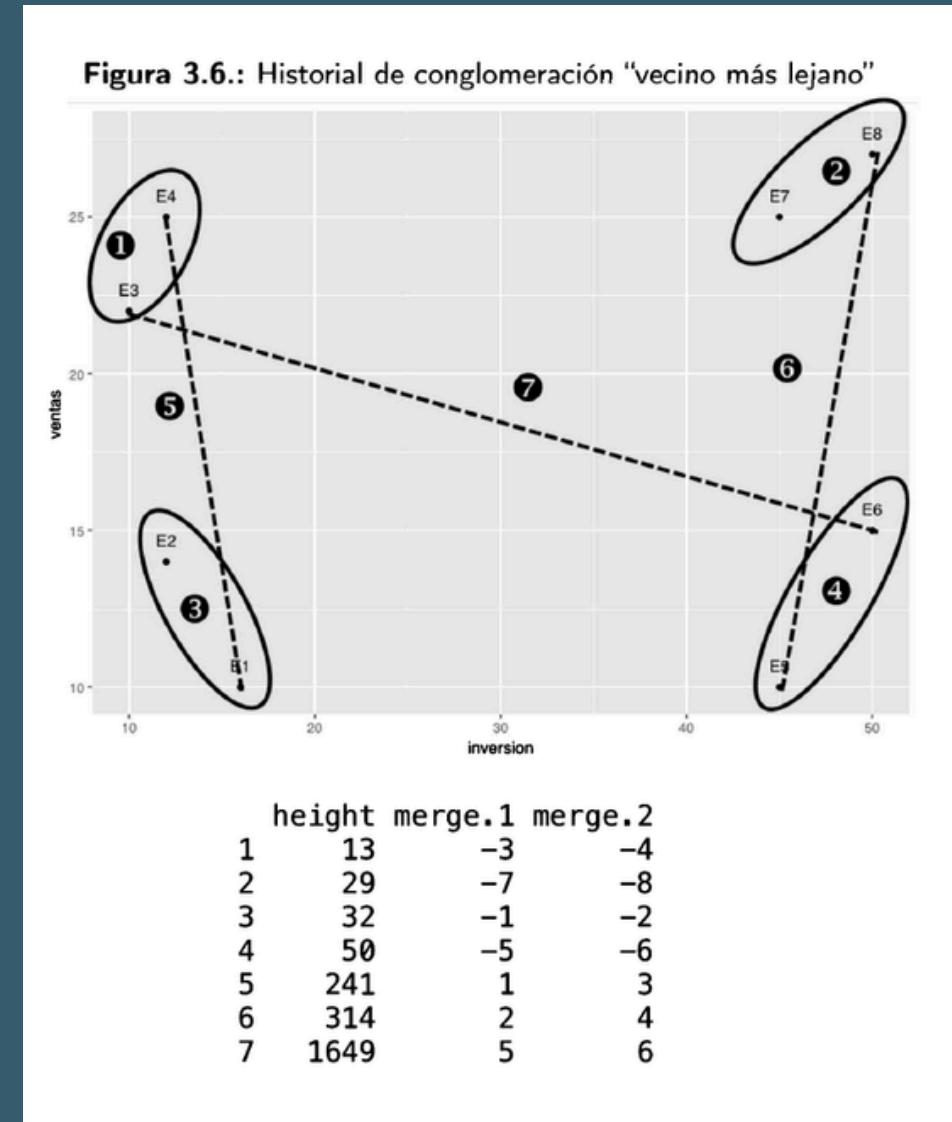
3. Métodos de aglomeración jerárquicos: algoritmos

3. Método del “Vecino más lejano” (complete-linkage)

- Similitud entre grupos:

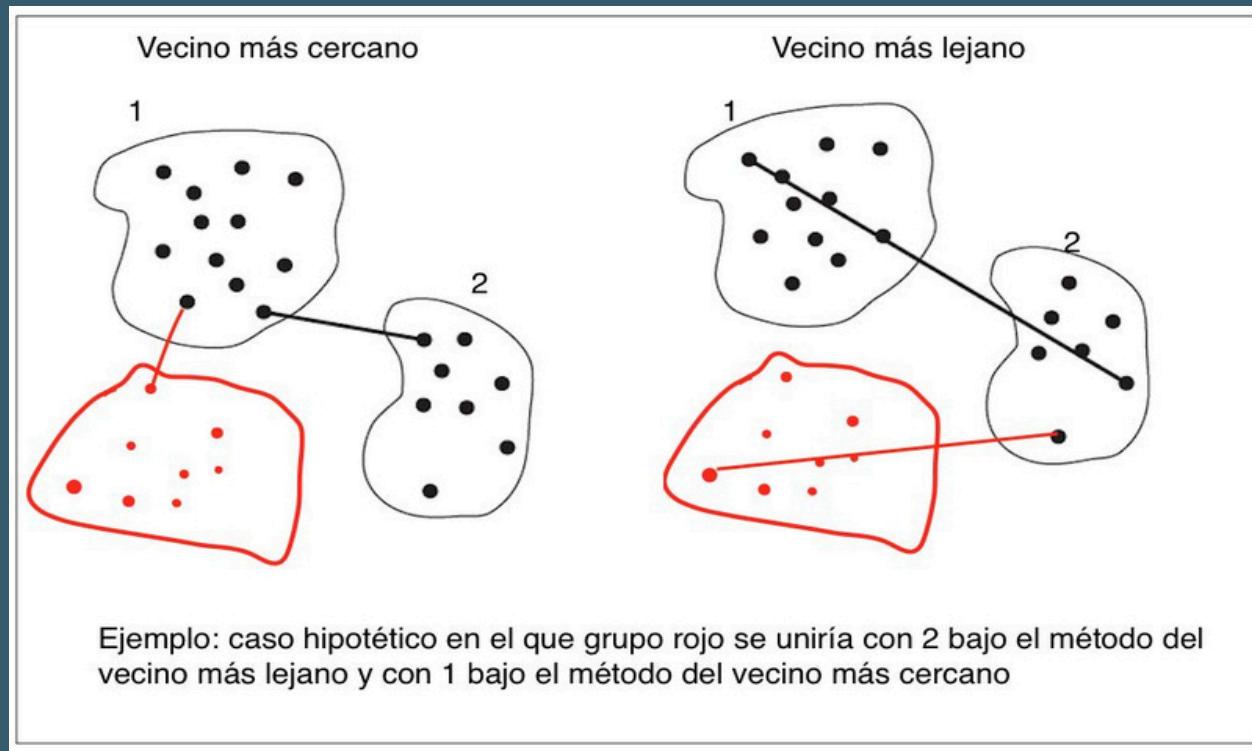
Análisis de las distancias máximas entre las observaciones de cada grupo

Se unen grupos cuyas distancias “más lejanas” son las menores



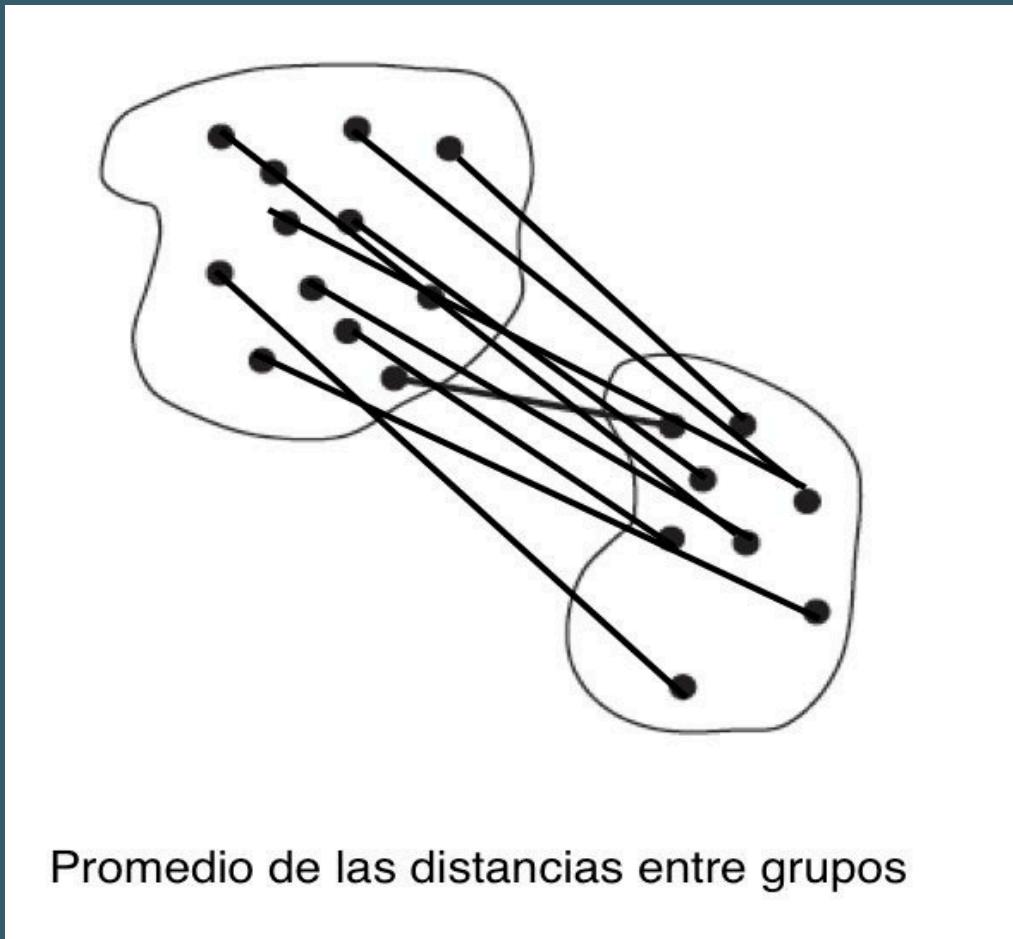
3. Métodos de aglomeración jerárquicos: algoritmos

Diferencias entre métodos del "Vecino más cercano" y "Vecino más lejano"



3. Métodos de aglomeración jerárquicos: algoritmos

4. Promedio de las distancias entre grupos (average linkage between groups)



- Se calcula la distancia promedio de todos los integrantes de un grupo respecto de los integrantes de otro grupo; dos grupos se combinan cuando el promedio distancia es la menor posible

3. Métodos de aglomeración jerárquicos: algoritmos

5. Promedio intragrupal

- Variante del algoritmo anterior
-
- Se parte combinando de dos en dos los grupos, de manera sucesiva.
-
- Luego se calcula la distancia promedio entre los miembros de esos grupos.

Se mantienen sólo los grupos cuya distancia promedio intragrupal es la menor

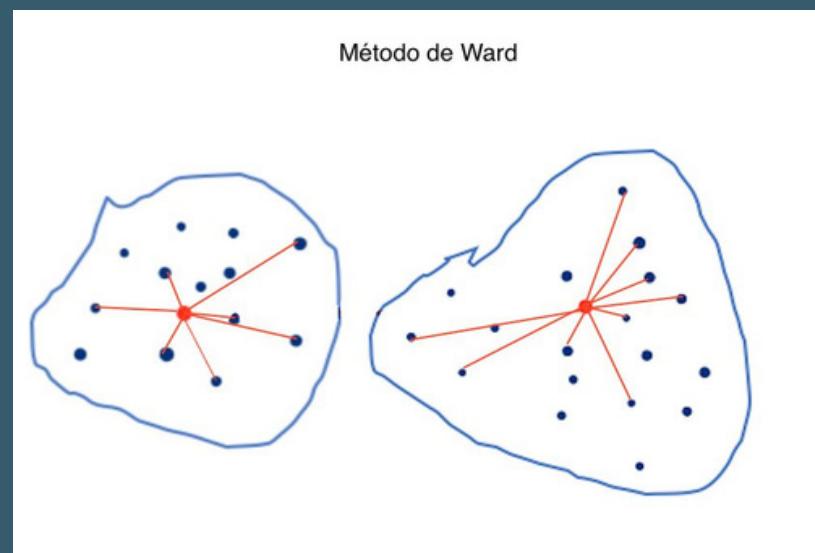
3. Métodos de aglomeración jerárquicos: algoritmos

6. Método de Ward

Se busca minimizar la varianza el interior de cada grupo en cada paso, calculando las distancias al cuadrado de cada caso a la media de su grupo al cuadrado

Luego se calcula la suma total de dichas distancias

Dos grupos se unen cuando esa suma total sea la menor posible, es decir, cuando se cumpla el criterio de minimización de “varianza” (suma de distancias al cuadrado), asumiendo que siempre el valor de las distancias al cuadrado va a aumentar



Ejemplo

- Medida de distancia: distancia euclídea al cuadrado
- Algoritmo de agrupamiento: Ward
- Especificación rango de soluciones: 2 a 4



Módulo 2 Ejercicio. 0,2 puntos para Control 2

Elija algún grupo de variables continuas o de Likert de la encuesta ELSOC para las que usted considere adecuado hacer un análisis de conglomerado, **es decir, que usted considere generen grupos diferenciados de personas.** Estandarice las variables. Considere hacer un muestreo de pocos casos para facilitar el trabajo del computador y la interpretación (100 casos por ejemplo)

Aplique el algoritmo de Ward según script adjunto.

Interprete