Tema 3
Técnicas de análisis multivariante para agrupación

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

Medidas de distancia

Especie	SL	SW	PL	PW
Iris setosa	5.8	4.0	1.2	0.2
Iris versicolor	5.1	2.5	3.0	1.1
Iris virginica	6.3	2.8	5.1	1.5

	Distancia euclídea		
	1:setosa	51:versicolor	101:virginica
1:setosa	,000	2,606	4,312
51:versicolor	2,606	,000	2,470
101:virginica	4,312	2,470	,000

$$\begin{aligned} d_{1,51} &= \sqrt{(5.8-5.1)^2+(4.0-2.5)^2+(1.2-3.0)^2+(0.2-1.1)^2} \\ &= \sqrt{6.79} = 2.6057. \end{aligned}$$

Medidas de similitud

i / j	1	0	Total
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	$p=a+b+c+d$

- Coeficiente de concordancia simple: $m_{ij} = \frac{a+d}{a+d+b+c}$
- Coeficiente de Jaccard: $m_{ij} = \frac{a}{a+b+c}$

Resultados de la observación de la presencia (1) o ausencia (0) de tres especies de prado en 15 parcelas del experimento de Park Grass.

	d3	a3	d8	a8	d7	a7	d17	a17	d16	a16	d14	s14	d1	c1	d18
A	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1
B	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0

Tabla de contingencia de *Agrostis tenuis* frente a *Briza media*:

A / B	1	0	Total
1	$a = 6$	$b = 6$	$a + b = 12$
0	$c = 0$	$d = 3$	$c + d = 3$
Total	$a + c = 6$	$b + d = 9$	$p = 15$

- Coeficiente de concordancia simple: Supongamos que estamos interesados en contar cuantas veces las dos especies conviven y cuantas veces las dos especies no se desarrollan. En el primer caso, serán parcelas que tienen condiciones favorables a las dos especies y en el segundo caso serán parcelas que tienen condiciones desfavorables a las dos especies. En el ejemplo:

$$m_{A,B} = \frac{9}{15} = 0.600.$$

- Coeficiente de Jaccard: Supongamos que sólo queremos tener en cuenta las parcelas que tienen condiciones favorables a alguna de las dos especies. En el ejemplo:

$$m_{A,B} = \frac{6}{12} = 0.500.$$

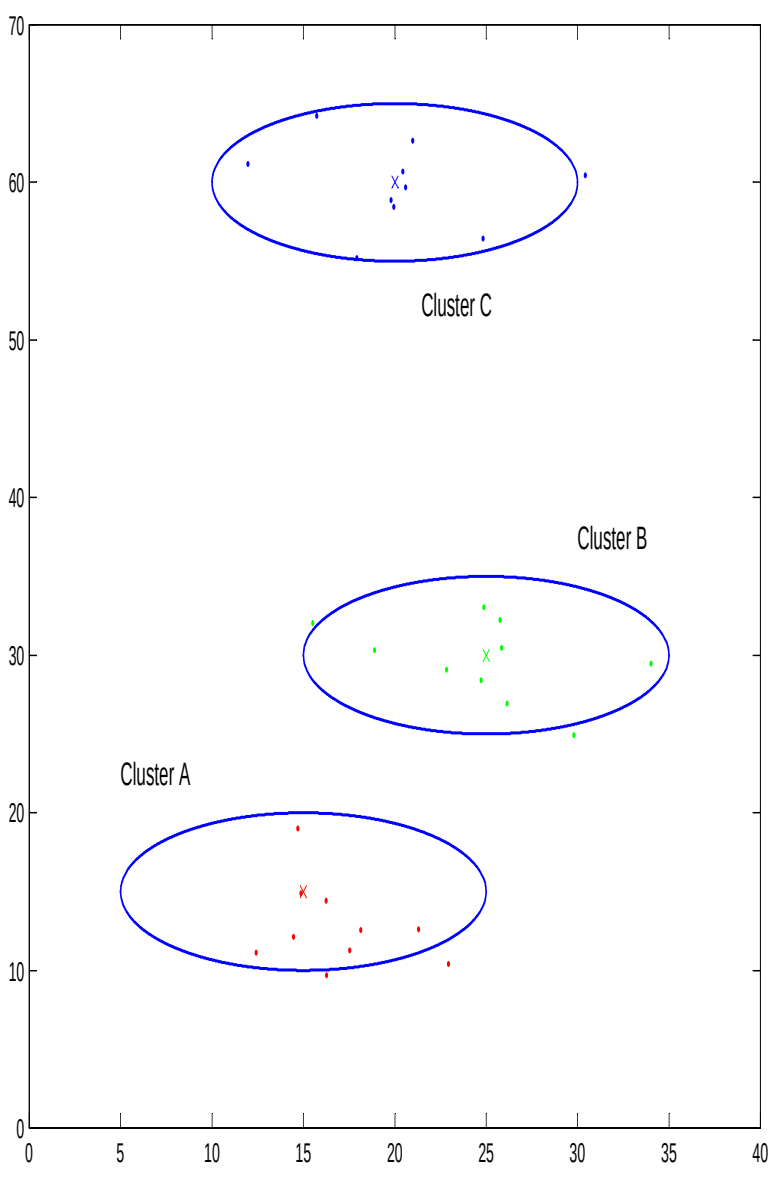
La especie *Briza media* es más parecida a *Cynosurus cristatus* que a *Agrostis tenuis* utilizando:

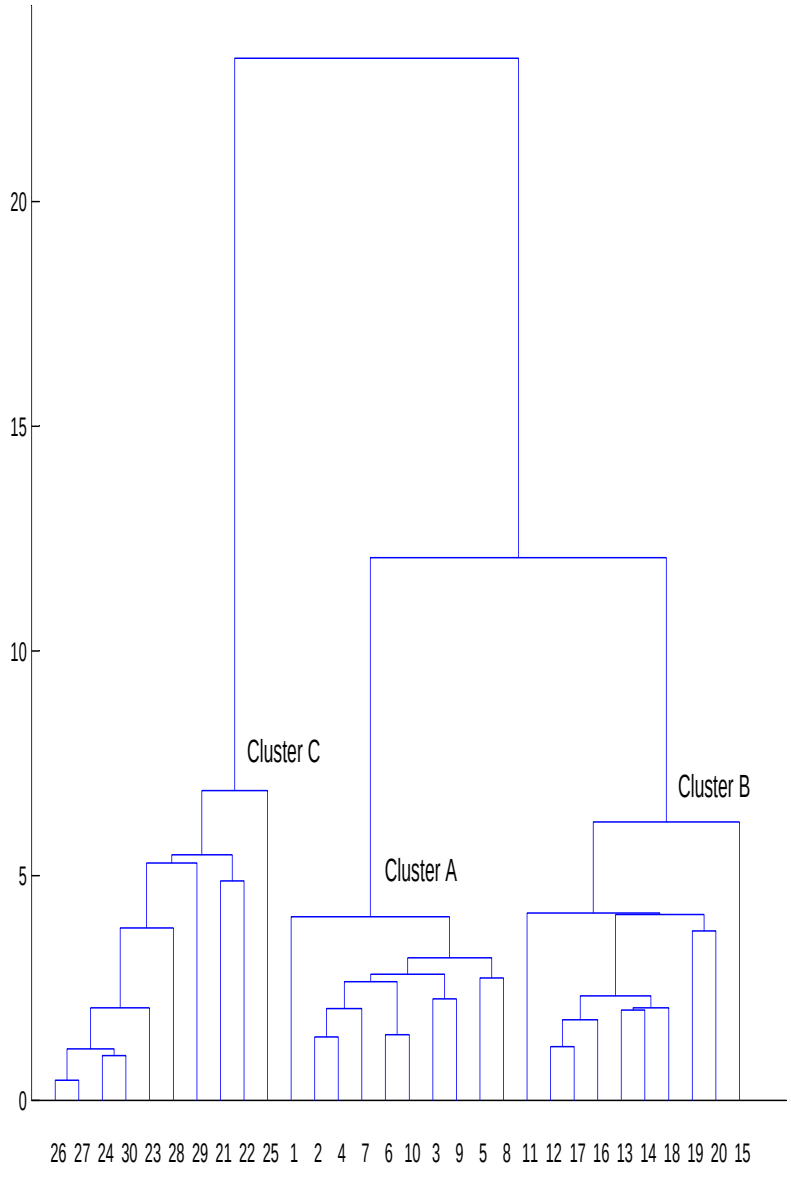
	Medida de emparejamiento simple		
Caso	1	2	3
1:Agrostis tenuis	1,000	,600	,333
2:Briza media	,600	1,000	,733
3:Cynosurus cristatus	,333	,733	1,000

La especie *Agrostis tenuis* es más parecida a *Briza media* que a *Cynosurus cristatus* utilizando:

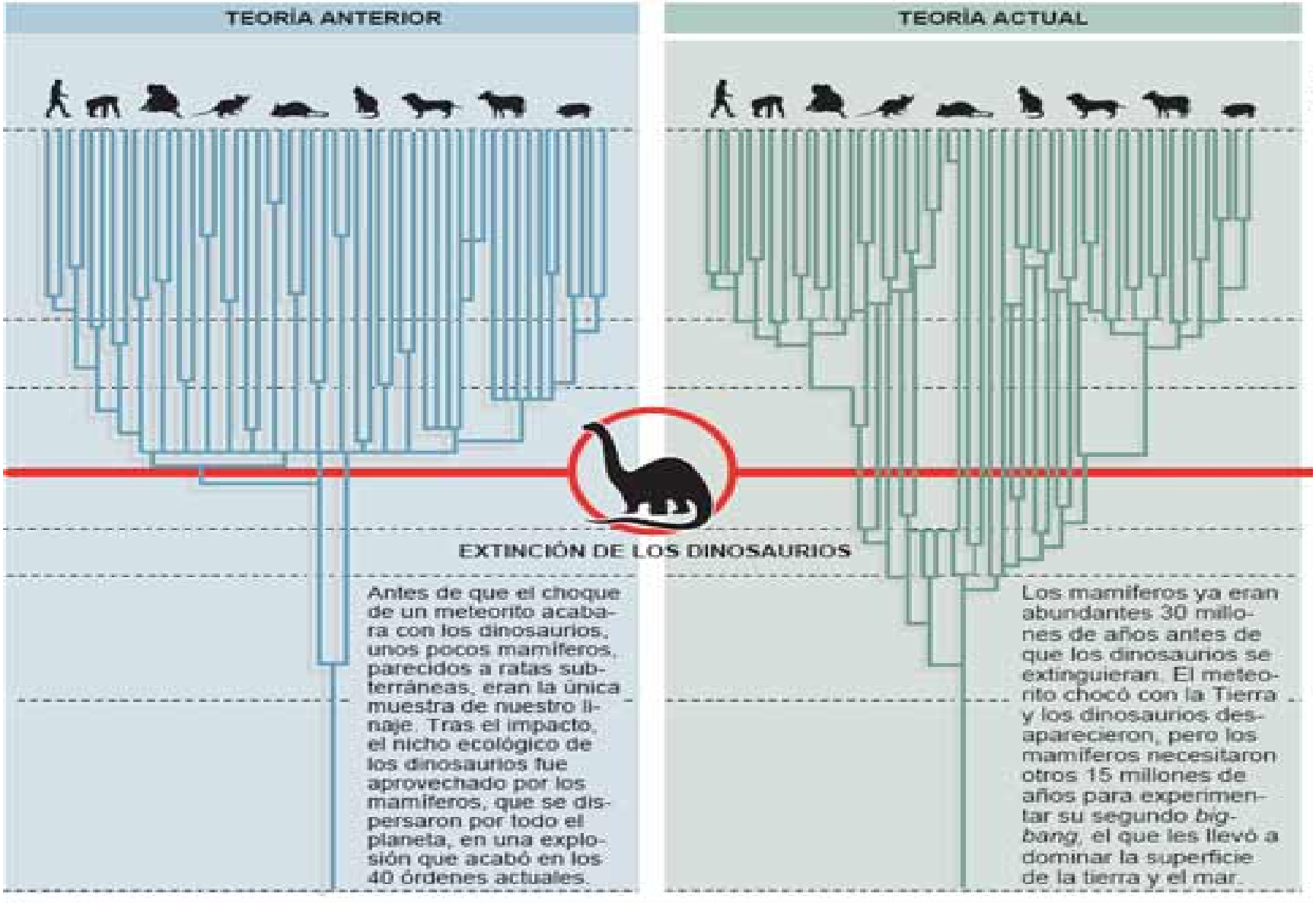
	Medida de Jaccard		
Caso	1	2	3
1:Agrostis tenuis	1,000	,500	,167
2:Briza media	,500	1,000	,333
3:Cynosurus cristatus	,167	,333	1,000

Agrupación jerárquica y no jerárquica





Arbol de evolución de los mamíferos



Fuente: Nature.

Métodos de agrupación jerárquicos

	d3	a3	d8	a8	d7	a7	d17	a17	d16	a16	d14	s14	d1	c1	d18
A	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1
B	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
D	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
F	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1

La matriz de similitudes (con el coeficiente de Jaccard) es:

	Medida de Jaccard				
Caso	1	2	3	4	5
1:Agrostis tenuis	1,000	,500	,167	,600	,857
2:Briza media	,500	1,000	,333	,500	,429
3:Cynosurus cristatus	,167	,333	1,000	,167	,143
4:Dactylis glomerata	,600	,500	,167	1,000	,733
5:Festuca rubra	,857	,429	,143	,733	1,000

Un método jerárquico general tiene los siguientes pasos:

1. Comenzamos con n clusters (cada uno contiene una observación), y con una matriz de similitudes. En el ejemplo, comenzamos con $n = 5$ clusters y la matriz de similitudes

$$\mathbf{M} = \begin{bmatrix} 1.000 & 0.500 & 0.167 & 0.600 & 0.857 \\ 0.500 & 1.000 & 0.333 & 0.500 & 0.429 \\ 0.167 & 0.333 & 1.000 & 0.167 & 0.143 \\ 0.600 & 0.500 & 0.167 & 1.000 & 0.733 \\ 0.857 & 0.429 & 0.143 & 0.733 & 1.000 \end{bmatrix}.$$

2. En la matriz de similitudes buscamos la pareja de clusters más parecidos. En el ejemplo, es la pareja formada por las especies: *Agrostis tenuis* y *Festuca rubra* que denotaremos por A y F, respectivamente.

3. Unimos la pareja de clusters del paso anterior y actualizamos la matriz de similitudes del siguiente modo:
   - 3.1 Eliminamos las filas y las columnas que corresponden a la pareja de clusters.
   - 3.2 Añadimos una fila y una columna que contendrá las similitudes del nuevo cluster con los restantes clusters.

   En el ejemplo, unimos A y F en un nuevo cluster $\overline{AF}$ y actualizamos $\mathbf{M}$:

   3.1 $$\mathbf{M}_{-A,-F} = \begin{bmatrix} 1.000 & 0.333 & 0.500 \\ 0.333 & 1.000 & 0.167 \\ 0.500 & 0.167 & 1.000 \end{bmatrix}.$$

   3.2 $$\mathbf{M} = \mathbf{M}_{+\overline{AF}} = \begin{bmatrix} 1.000 & m_{\overline{AF},B} & m_{\overline{AF},C} & m_{\overline{AF},D} \\ m_{\overline{AF},B} & 1.000 & 0.333 & 0.500 \\ m_{\overline{AF},C} & 0.333 & 1.000 & 0.167 \\ m_{\overline{AF},D} & 0.500 & 0.167 & 1.000 \end{bmatrix}.$$

4. Los pasos 2 y 3 se repiten $n - 1$ veces.

Método del encadenamiento simple o vecino más próximo: La similitud entre el nuevo cluster y los restantes clusters se calcula por la máxima similitud de los miembros del nuevo grupo con los miembros de los restantes clusters. En el ejemplo:

$$\mathbf{M} = \begin{bmatrix} 1.000 & 0.500 & 0.167 & 0.600 & 0.857 \\ 0.500 & 1.000 & 0.333 & 0.500 & 0.429 \\ 0.167 & 0.333 & 1.000 & 0.167 & 0.143 \\ 0.600 & 0.500 & 0.167 & 1.000 & 0.733 \\ 0.857 & 0.429 & 0.143 & 0.733 & 1.000 \end{bmatrix}$$

(i) $m_{\overline{AF},B} = \max\{m_{A,B},\, m_{F,B}\} = \max\{0.500, 0.429\} = 0.500.$

(ii) $m_{\overline{AF},C} = \max\{m_{A,C},\, m_{F,C}\} = \max\{0.167, 0.143\} = 0.167.$

(iii) $m_{\overline{AF},D} = \max\{m_{A,D},\, m_{F,D}\} = \max\{0.600, 0.733\} = 0.733.$

La nueva matriz de similitudes es

$$\mathbf{M} = \begin{bmatrix} 1.000 & 0.500 & 0.167 & 0.733 \\ 0.500 & 1.000 & 0.333 & 0.500 \\ 0.167 & 0.333 & 1.000 & 0.167 \\ 0.733 & 0.500 & 0.167 & 1.000 \end{bmatrix}.$$

Ahora debemos repetir el paso 2 del algoritmo aglomerativo, es decir buscar la pareja de clusters más parecidos, que resulta ser la formada por el cluster $\overline{AF}$ y D. Unimos los dos clusters en un nuevo $\overline{AFD}$ y actualizamos la matriz

$$\mathbf{M} = \begin{bmatrix} 1.000 & 0.500 & 0.167 & 0.733 \\ 0.500 & 1.000 & 0.333 & 0.500 \\ 0.167 & 0.333 & 1.000 & 0.167 \\ 0.733 & 0.500 & 0.167 & 1.000 \end{bmatrix}$$

(i) $m_{\overline{AFD},B} = \max\{m_{\overline{AF},B},\, m_{D,B}\} = \max\{0.500, 0.500\} = 0.500.$

(ii) $m_{\overline{AFD},C} = \max\{m_{\overline{AF},C},\, m_{D,C}\} = \max\{0.167, 0.167\} = 0.167.$

La nueva matriz de similitudes es $\mathbf{M} = \begin{bmatrix} 1.000 & 0.500 & 0.167 \\ 0.500 & 1.000 & 0.333 \\ 0.167 & 0.333 & 1.000 \end{bmatrix}$.

Repetimos el paso 2 del algoritmo aglomerativo: al cluster $\overline{AFD}$ se le une B y actualizamos la matriz $\mathbf{M} = \begin{bmatrix} 1.000 & 0.500 & 0.167 \\ 0.500 & 1.000 & 0.333 \\ 0.167 & 0.333 & 1.000 \end{bmatrix}$

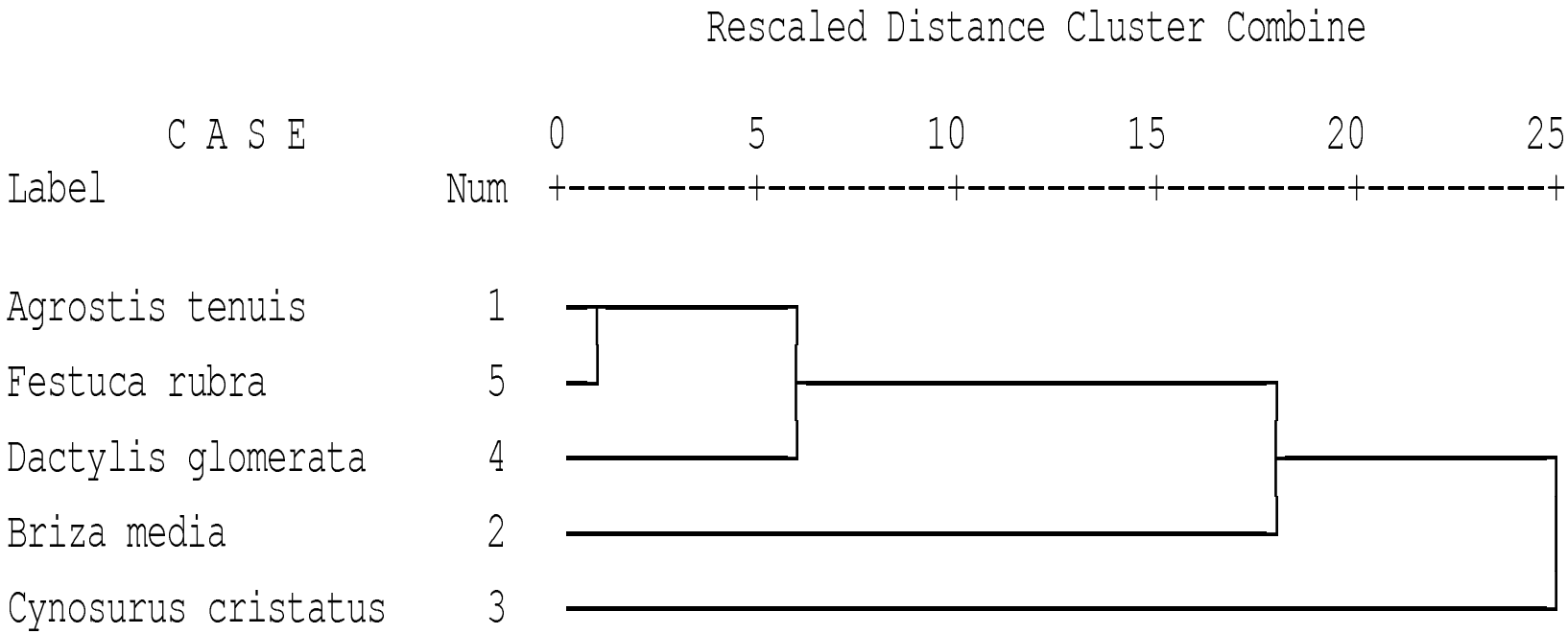
(i) $m_{\overline{AFDB},C} = \max\{m_{\overline{AFD},C}, m_{B,C}\} = \max\{0.167, 0.333\} = 0.333.$

La nueva matriz de similitudes es $\mathbf{M} = \begin{bmatrix} 1.000 & 0.333 \\ 0.333 & 1.000 \end{bmatrix}$.

Historial de conglomeración

Etapa	Cluster que se combina		Coeficientes	Etapa en la que el cluster aparece por primera vez		Próxima etapa
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	5	,857	0	0	2
2	1	4	,733	1	0	3
3	1	2	,500	2	0	4
4	1	3	,333	3	0	0

Otra manera de presentar el historial de agrupamiento es un árbol jerárquico o dendograma:



Método del encadenamiento completo o vecino más lejano: La similitud entre el nuevo cluster y los restantes clusters se calcula por la mínima similitud de los miembros del nuevo grupo con los miembros de los restantes clusters. En el ejemplo:

$$\mathbf{M} = \begin{bmatrix} 1.000 & 0.500 & 0.167 & 0.600 & 0.857 \\ 0.500 & 1.000 & 0.333 & 0.500 & 0.429 \\ 0.167 & 0.333 & 1.000 & 0.167 & 0.143 \\ 0.600 & 0.500 & 0.167 & 1.000 & 0.733 \\ 0.857 & 0.429 & 0.143 & 0.733 & 1.000 \end{bmatrix}$$

(i) $m_{\overline{AF},B} = \min\{m_{A,B}, m_{F,B}\} = \min\{0.500, 0.429\} = 0.429.$

(ii) $m_{\overline{AF},C} = \min\{m_{A,C}, m_{F,C}\} = \min\{0.167, 0.143\} = 0.143.$

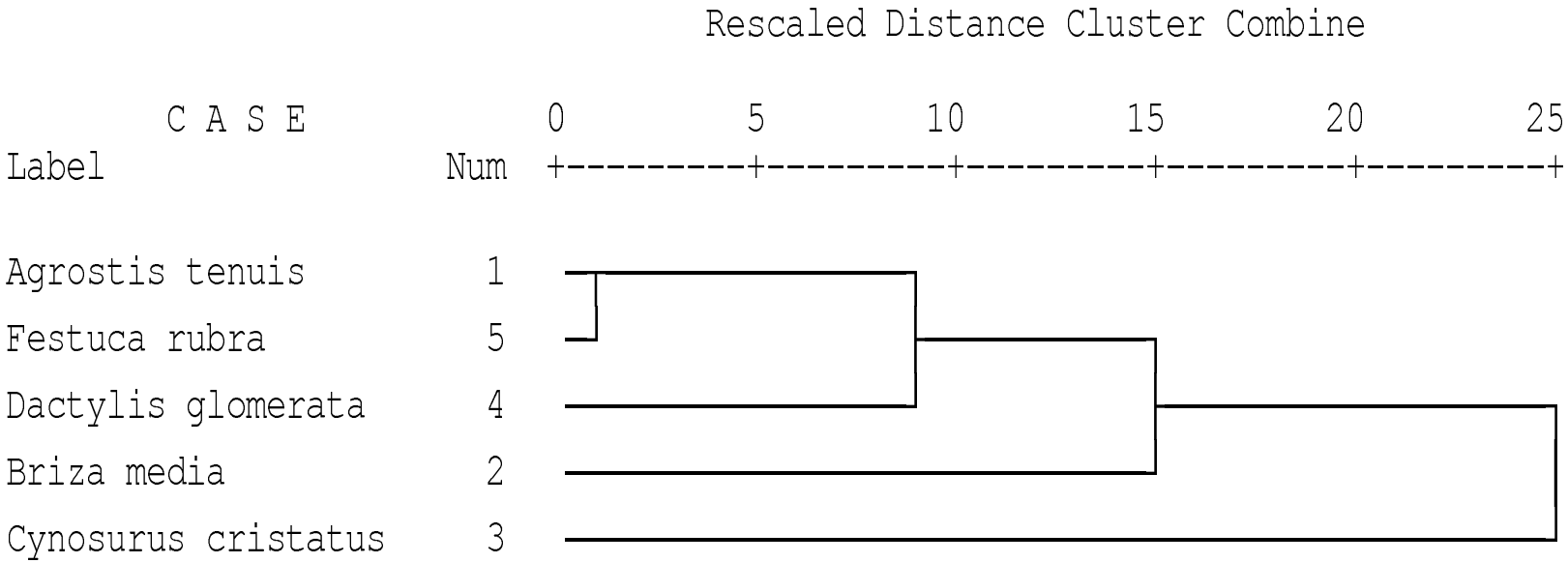
(iii) $m_{\overline{AF},D} = \min\{m_{A,D}, m_{F,D}\} = \min\{0.600, 0.733\} = 0.600.$

La nueva matriz de similitudes es $\mathbf{M} = \begin{bmatrix} 1.000 & 0.429 & 0.143 & 0.600 \\ 0.429 & 1.000 & 0.333 & 0.500 \\ 0.143 & 0.333 & 1.000 & 0.167 \\ 0.600 & 0.500 & 0.167 & 1.000 \end{bmatrix}$.

Historial de aglomeración y Dendograma basado en el método del vecino más lejano:

Historial de conglomeración

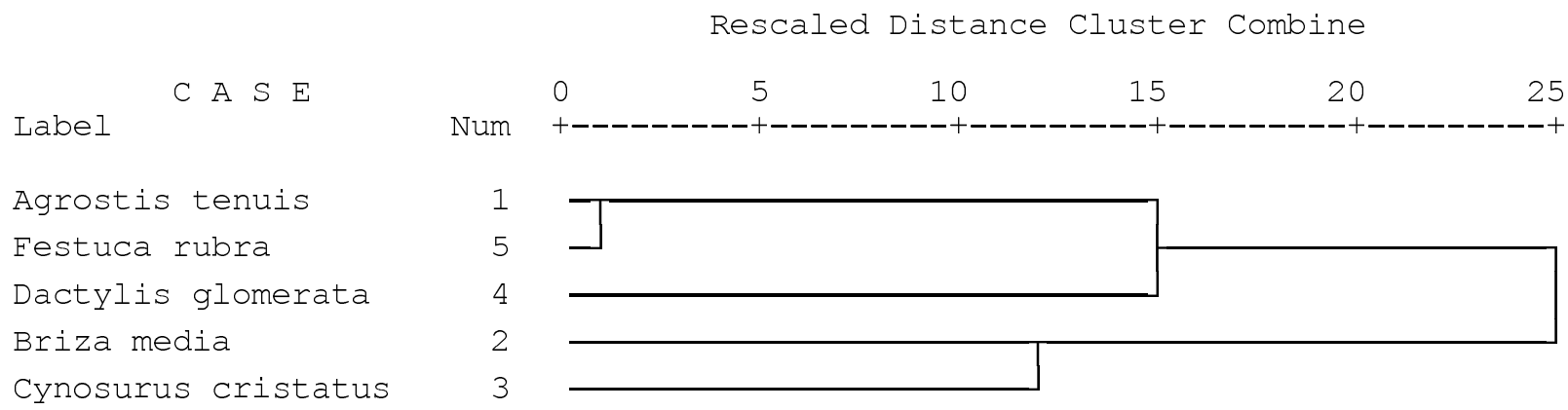
Etapa	Cluster que se combina		Coeficientes	Etapa en la que el cluster aparece por primera vez		Próxima etapa
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	5	,857	0	0	2
2	1	4	,600	1	0	3
3	1	2	,429	2	0	4
4	1	3	,143	3	0	0



Método de agrupación de centroides: La distancia entre dos clusters se mide como la distancia euclídea al cuadrado entre los centros de los clusters. En el ejemplo:

Historial de conglomeración

Etapa	Cluster que se combina		Coeficientes	Etapa en la que el cluster aparece por primera vez		Próxima etapa
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	5	2,000	0	0	3
2	2	3	4,000	0	0	4
3	1	4	4,500	1	0	4
4	1	2	6,333	3	2	0



Cuestiones sobre métodos jerárquicos

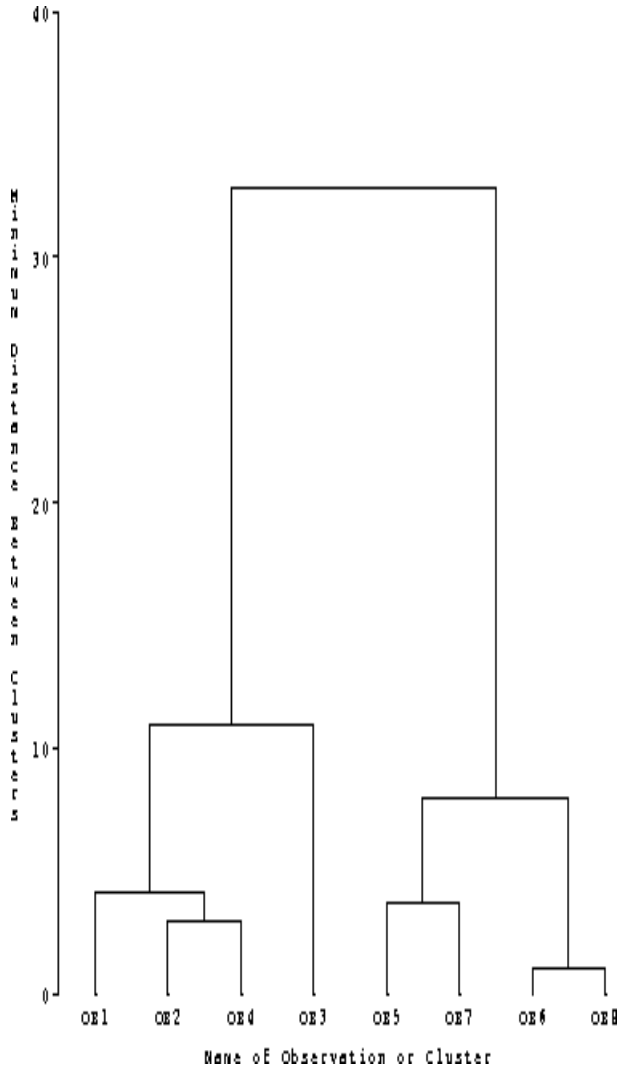
Datos de temperaturas y cosechas en 8 localizaciones geográficas y la correspondiente matriz de distancias euclídeas:

$$\begin{bmatrix} 24.0 & 113 \\ 25.1 & 109 \\ 24.7 & 124 \\ 25.5 & 106 \\ 31.9 & 156 \\ 30.3 & 167 \\ 29.6 & 159 \\ 30.8 & 168 \end{bmatrix}$$

Group	1	2	3	4	5	6	7	8
1	0							
2	4.15	0						
3	11.02	15.01	0					
4	7.16	3.03	18.02	0				
5	43.72	47.49	32.80	50.41	0			
6	54.37	58.23	43.36	61.19	11.12	0		
7	46.34	50.20	35.34	53.16	3.78	8.03	0	
8	55.42	59.27	44.42	62.23	12.05	1.12	9.08	0

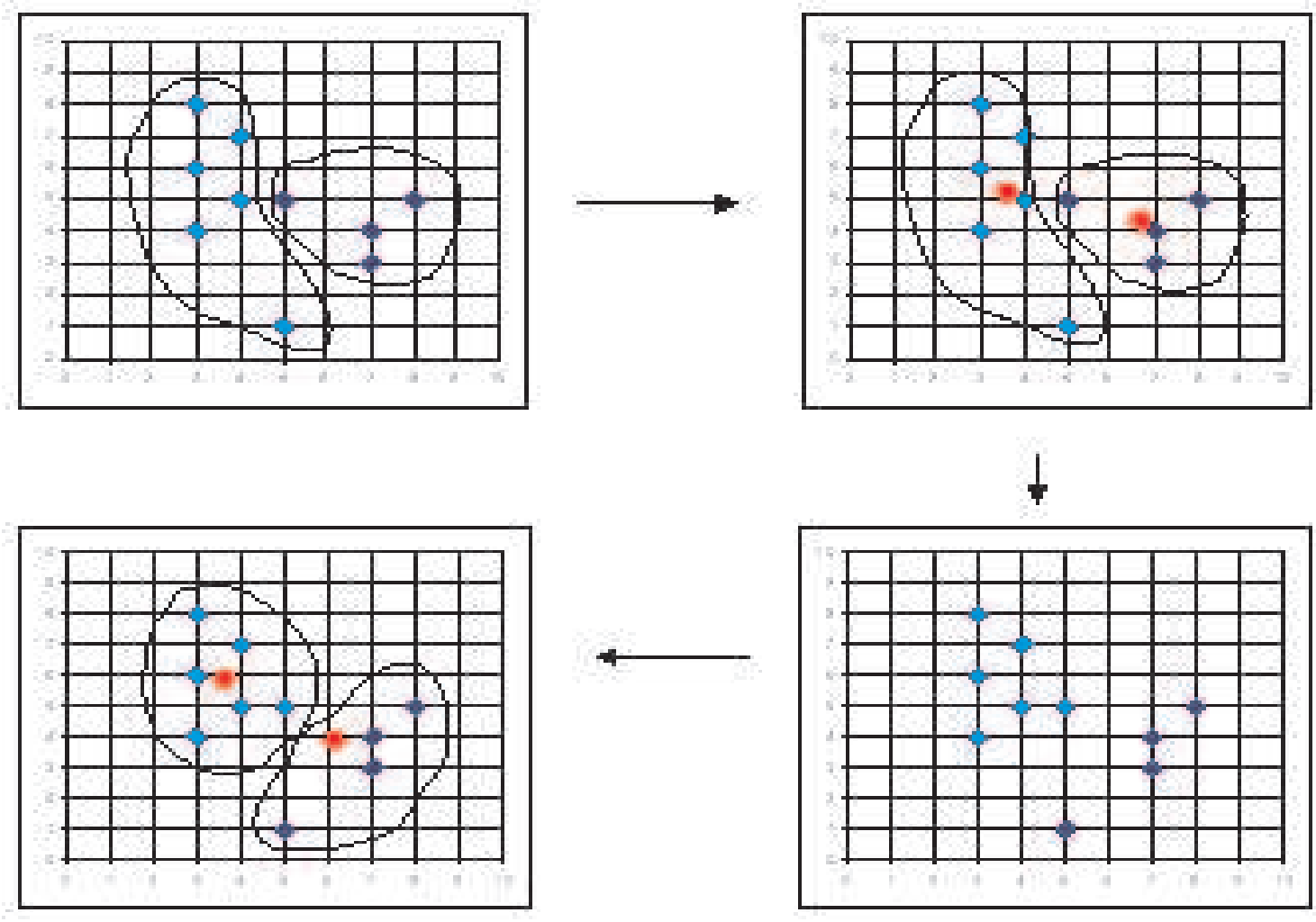
1. ¿Cuál es el valor de la distancia rectangular entre $\mathbf{x}_1$ y $\mathbf{x}_2$?
2. ¿Qué dos observaciones son las que primero se unen en un método de agrupación jerárquico usando encadenamiento simple? ¿Y usando encadenamiento completo?
3. Calcula las distancias entre el grupo formado en la primera iteración y el resto de grupos si se utiliza encadenamiento simple.

Una vez concluye el algoritmo el resultado se representa mediante un dendograma



1. ¿Cuántos grupos relativamente homogéneos hay en estos datos?
2. ¿Qué grupos se unieron en la segunda iteración?
3. ¿Qué grupos se unieron en la última?
4. ¿Cuál es la distancia aproximada entre los grupos $\{5, 7\}$ y $\{6, 8\}$?

Método de las k medias: esquema



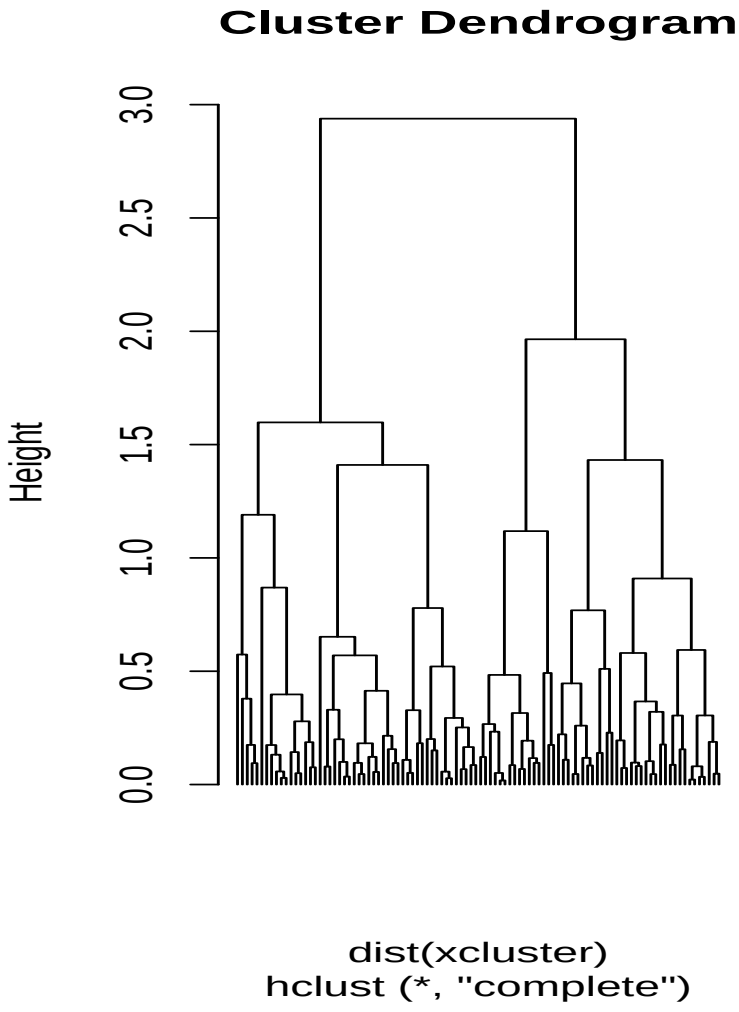
xcluster[,2]
xcluster[,1]
−1.0
0.0
1.0
−0.5
0.0
0.5
1.0
1.5

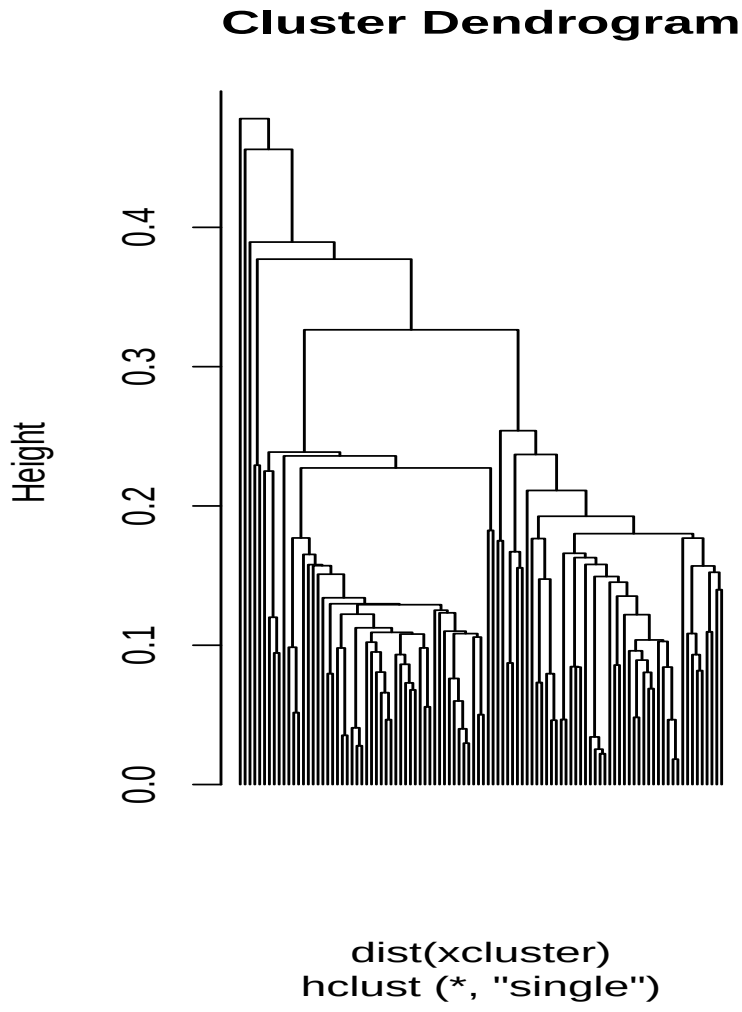
xcluster2[,2]
xcluster2[,1]
−2
0
1
2
3
−2
−1
0
1
2
3
4

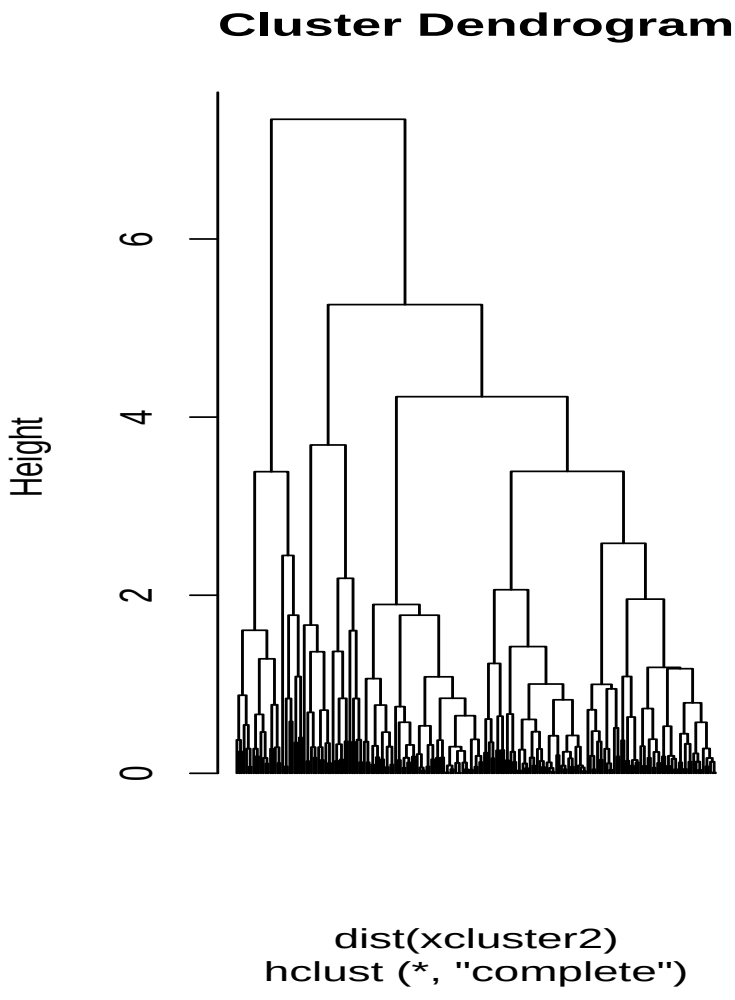
xcluster[,2]
xcluster[,1]
−1.0
0.0
1.0
−0.5
0.0
0.5
1.0
1.5

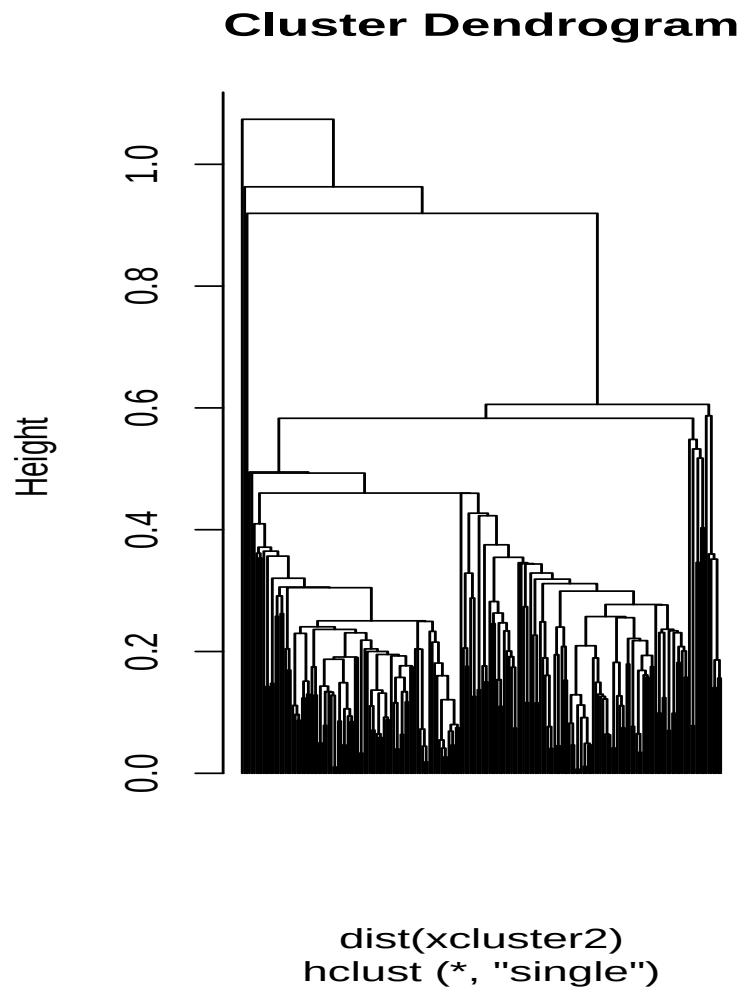
xcluster2[,2]
xcluster2[,1]
−2
0
1
2
3
−2
−1
0
1
2
3
4

Comparación con los métodos jerárquicos






Cluster Dendrogram
Height
0
2
4
6
dist(xcluster2)
hclust (*, "complete")


Cluster Dendrogram
Height
0.0
0.2
0.4
0.6
0.8
1.0
dist(xcluster2)
hclust (*, "single")

Ejemplo: tortugas

La siguiente tabla presenta tres medidas (longitud, ancho y altura) del caparazón de 4 tortugas pintadas. Aplicamos el algoritmo K-medias con $K = 2$:

Id.	Longitud	Ancho	Altura
m1	120	89	40
m2	119	93	41
f1	159	118	63
f2	155	115	63

1. Se asignan aleatoriamente $n/K = 2$ elementos a cada grupo. Supongamos que m1 y f1 se asignan al primer grupo y m2 y f2 se asignan al segundo grupo.

2. Calculamos el vector de medias de cada grupo $\bar{\mathbf{x}}_1$ y $\bar{\mathbf{x}}_2$ y efectuamos la asignación secuencial de cada observación:

Id.	Grupo Inicial	Dist. a $\bar{\mathbf{x}}_1$	Dist. a $\bar{\mathbf{x}}_2$	Grupo Final	Nueva $\bar{\mathbf{x}}_1'$	Nueva $\bar{\mathbf{x}}_2'$
m1	1	26.8	25.7	2	[159.0 118.0 63.0]	[131.3 99.0 48.0]
m2	2	52.0	15.4	2	[159.0 118.0 63.0]	[131.3 99.0 48.0]
f1	1	0.0	36.8	1	[159.0 118.0 63.0]	[131.3 99.0 48.0]
f2	2	5.0	32.3	1	[157.0 116.5 63.0]	[119.5 91.0 40.5]

3. Repetimos el paso 2 como sigue:

Id.	Grupo Inicial	Dist. a $\bar{\mathbf{x}}_1$	Dist. a $\bar{\mathbf{x}}_2$	Grupo Final	Nueva $\bar{\mathbf{x}}_1'$	Nueva $\bar{\mathbf{x}}_2'$
m1	2	51.5	2.1	2	[157.0 116.5 63.0]	[119.5 91.0 40.5]
m2	2	49.8	2.1	2	[157.0 116.5 63.0]	[119.5 91.0 40.5]
f1	1	2.5	52.9	1	[157.0 116.5 63.0]	[119.5 91.0 40.5]
f2	1	2.5	48.4	1	[157.0 116.5 63.0]	[119.5 91.0 40.5]