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 UNIDAD

PRINCIPALES ELEMENTOS CONCEPTUALES Y DEFINICIONES

TIPOS DE ANÁLISIS

- MIRADA GENERAL A PROCEDIMIENTOS JERÁRQUICOS
- MIRADA GENERAL PROCEDIMIENTO NO 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 multivariables de clasificación o 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

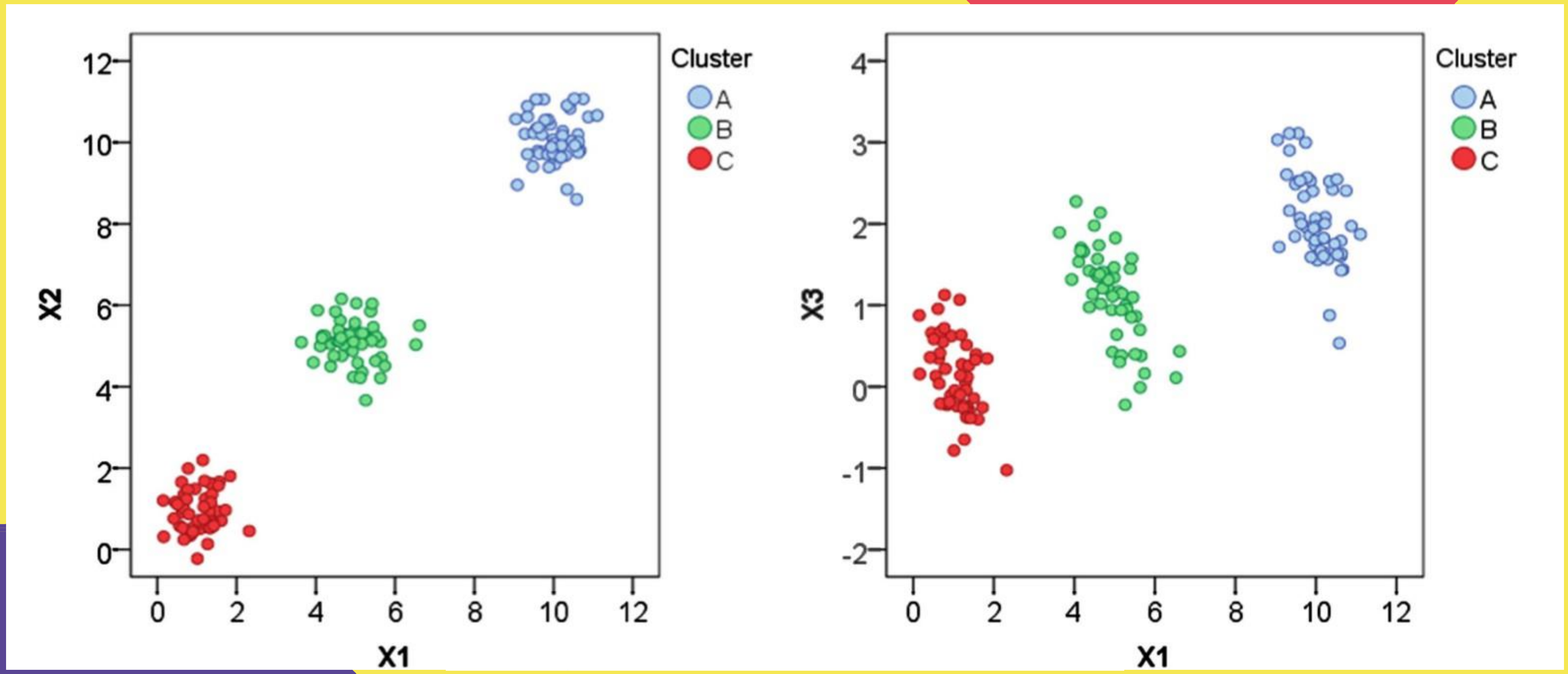
USOS

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 hlpotesls 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,

Concepto de distancia

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 euclidiana, distancia euclidiana^2, otras

Variables ordinales: se usan medidas de similitud

Lo veremos en las próximas clases

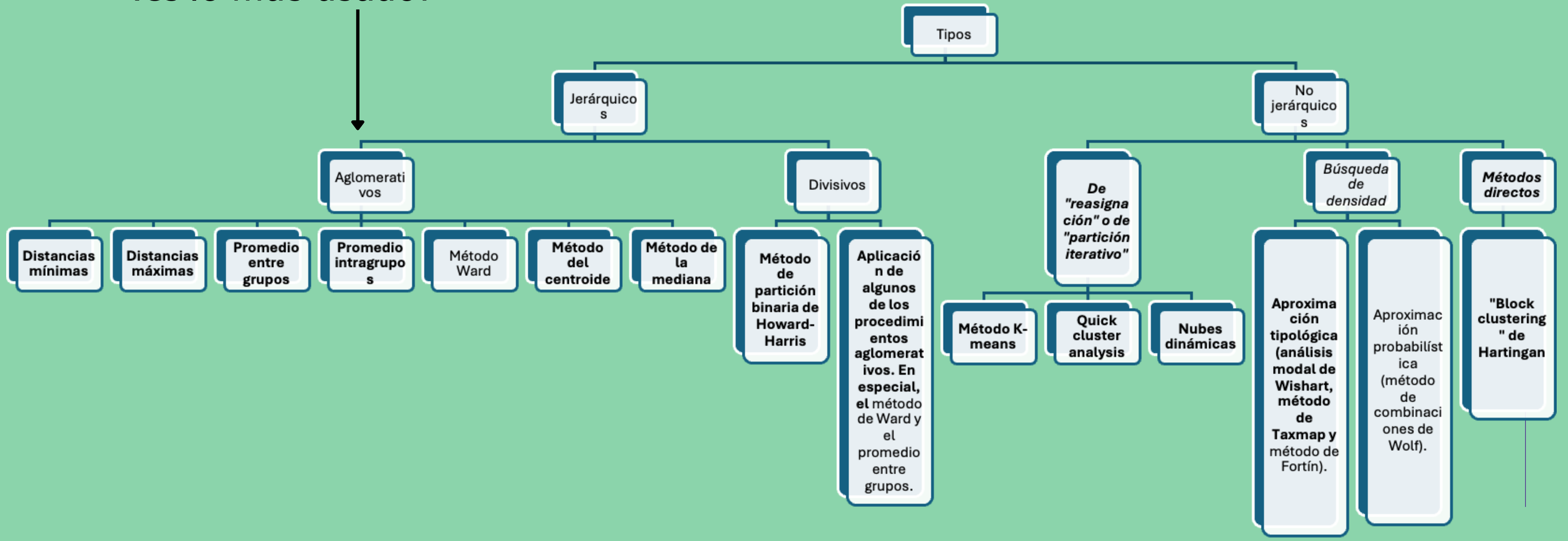
TIPOS

- MODELOS JERÁRQUICOS
- MODELOS NO JERÁRQUICOS

LOS RESULTADOS PUEDEN NO COINCIDIR, CUANDO SE APLICAN MÉTODOS DE CONGLOMERACIÓN DIFERENTES.

TIPOS

¡es lo más usado!



muestras de hasta 200 casos

muestras de más de 200 casos

LOS RESULTADOS PUEDEN NO COINCIDIR, CUANDO SE APLICAN MÉTODOS DE CONGLOMERACIÓN DIFERENTES.

Cea de Anacona, 2004

PASOS PARA REALIZAR UN ANÁLISIS DE CONGLOMERADOS

Pasos previos a realización del cálculo

Selección y tratamiento de los datos

Selección de variables

Selección de método de conglomeración

Obtención de conglomerados

Decisión sobre el número de conglomerados

Interpretación de los resultados

Validación de los resultados

Positiva / Negativa (volver al inicio)

Britt et. al, 2014; Cea de Anacona, 2020

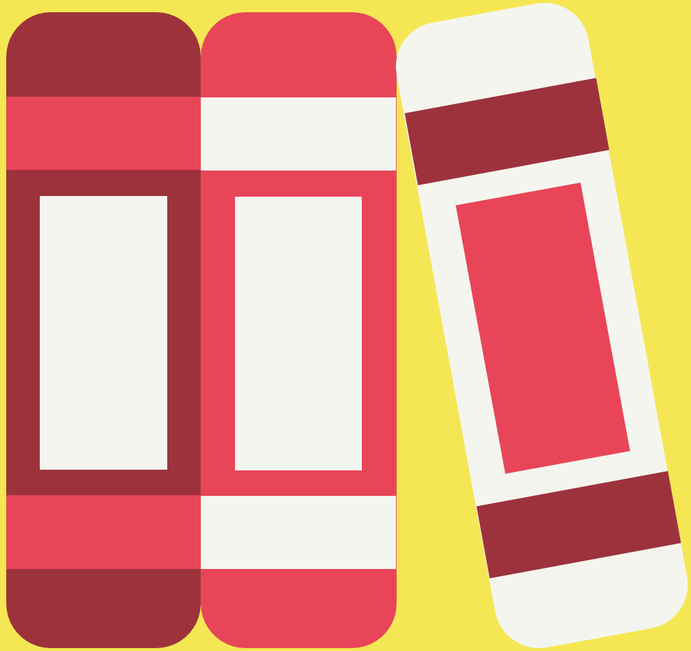
SELECCIÓN DE LAS VARIABLES

ASPECTO CRÍTICO:

Deben incluirse únicamente **variables que caractericen** a los objetos que se desean agrupar y que estén específicamente relacionadas con los objetivos del análisis de clúster.

Incluir únicamente variables t**eóricamente relevantes** para la clasificación de los casos.

De no ser así, existe un serio riesgo de caer en un **empirismo ingenuo**, produciendo resultados conceptualmente vacíos y que no contribuyen a la acumulación del conocimiento.



Britt et. al, 2014; Cea de Anacona, 2020

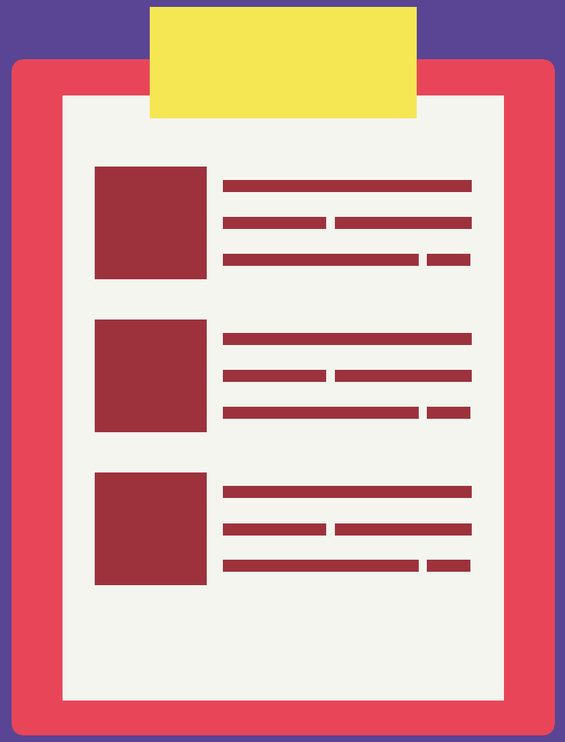
SELECCIÓN DE LAS VARIABLES

Segunda decisión clave:

- **Usar la variable con su métrica original o estandarizar.**

No hay consenso sobre una regla general

Britt et. al, 2014; Cea de Anacona, 2020



DOS TIPOS DE LÓGICAS DE CONGLOMERACIÓN

- **Métodos jerárquicos**
- **Métodos no jerárquico**

Cea de Anacona, 2004

SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos jerárquicos

Son los procedimientos más aplicados (especialmente los "aglomerativos"), cuando el tamaño de la muestra no es elevado (< 200 unidades).

Los análisis se realizan a partir de una matriz de distancias, con entradas para cada par de objetos (casos o variables).

El volúmen de los conglomerados aumenta con el tamaño de la muestra, al igua que con la lectura e interpretación de los resultados gráficos (el dendograma y el gráfico de carámbanos).

Cea de Anacona, 2004

SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos jerárquicos

Los conglomerados se caracterizan por ser anidados: cada conglomerado puede ser subsurrúdo por otro conglomerado más grande, en un nivel de similaridad superior.

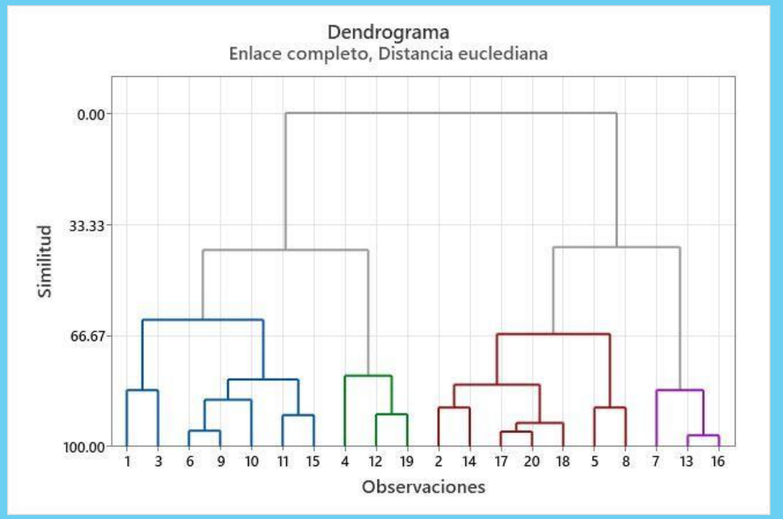
Se puede partir desde las unidades hacía “arriba” (método aglomerativo o ascendente) o desde el total hacia las unidades (método divisivo o “descendente”)

Existen diferentes formas de calcular esas distancias de similitud dando lugar a diferentes modelos dentro de cada tipo

Cea de Anacona, 2020

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.

Cea de Anacona, 2004

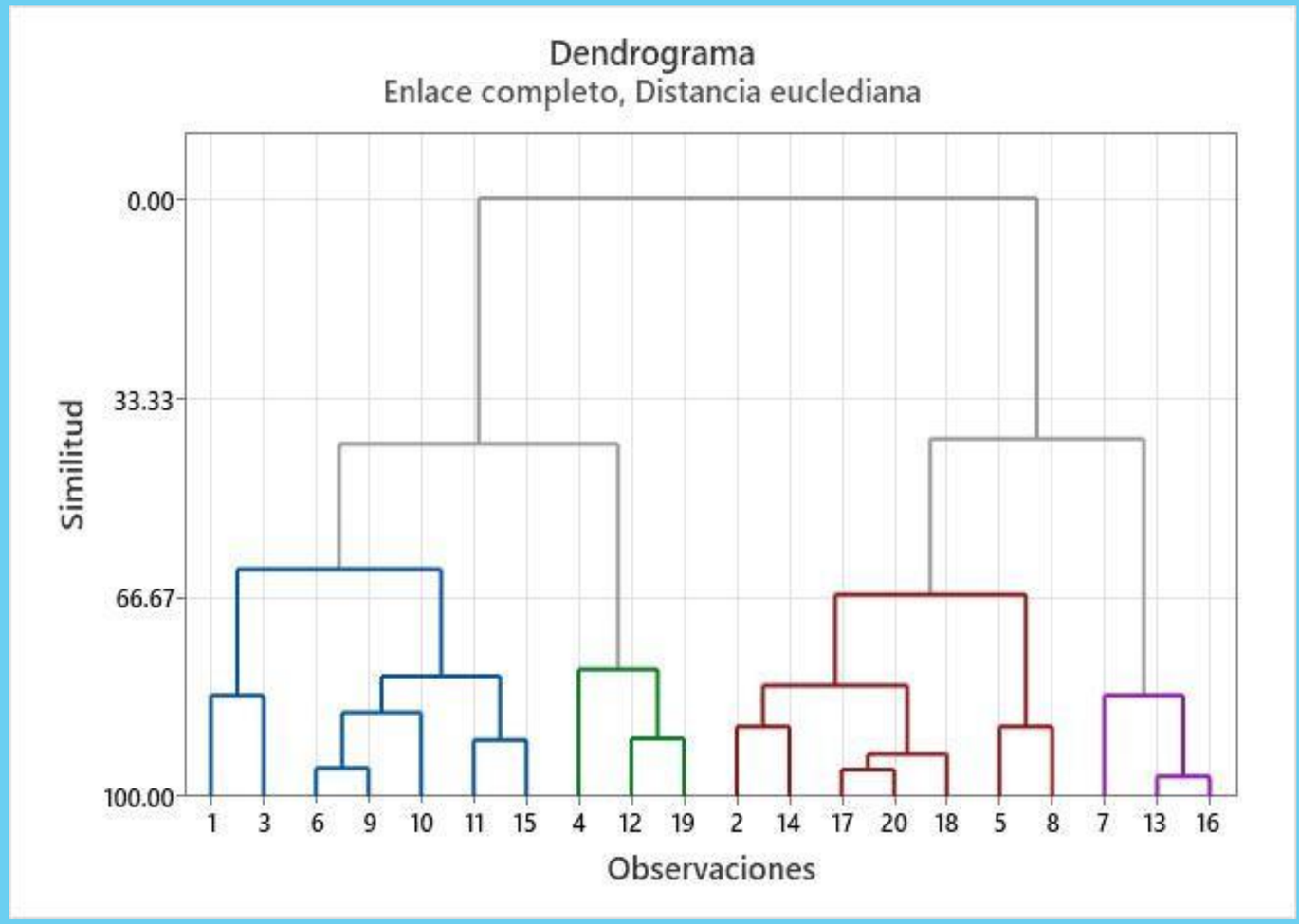
SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Ejemplo

Métodos jerárquicos aglomerativos - dendrograma

4 conglomerados (colores)

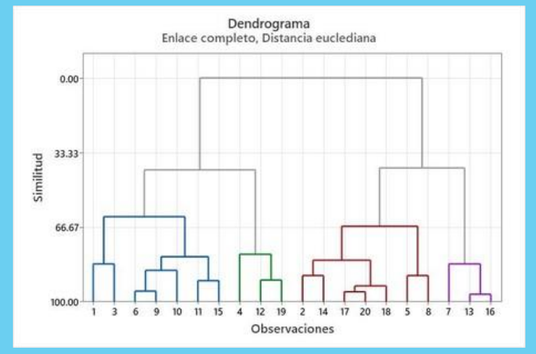
Si se cortara el dendrograma más arriba, entonces habría menos conglomerados finales, pero su nivel de similitud sería menor. Si se cortara el dendrograma más abajo, entonces el nivel de similitud sería mayor, pero habría más conglomerados finales.



Cea de Anacona, 2004

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 posterio

- 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.

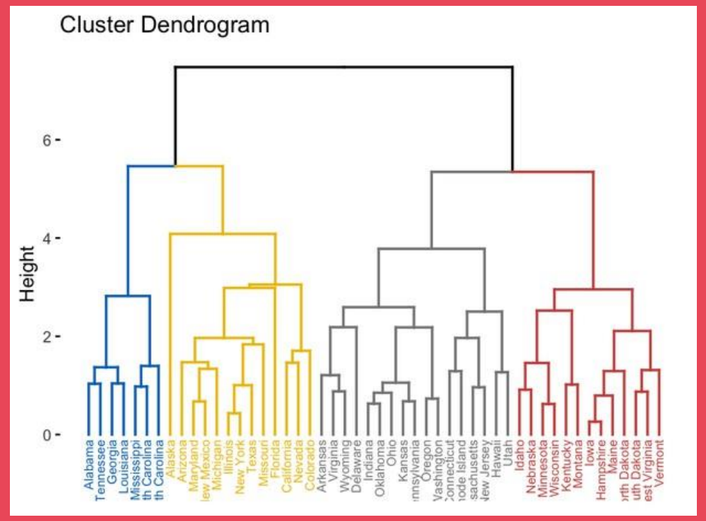
- Si, por el contrario, se quiere agrupar variables, las medidas de distancia/similaridad se calculan entre pares de variables.

Cea de Anacona, 2004

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.

Cea de Anacona, 2004

SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos no jerárquicos o de optimización

Comienza a partir de una partición inicial de los datos de acuerdo con algún criterio de investigación (el investigador especifica, previamente el número máximo de conglomerados que debe haber en la matriz de datos)

A partir de esta clasificación inicial se produce la asignación de objetos a los conglomerados a cuyo centro (o centroide, que corresponde al valor medio de las variables que configuran el conglomerado) se hallen más próximos.

Los centroides se recalculan cada vez que se produce una alteración en la composición de los conglomerados.

Este proceder, descrito de forma genérica, varía en función del algoritmo de clasificación escogido

El proceso concluye cuando no se produce ninguna nueva modificación en los conglomerados..

Continuamente se valora la pertenencia de los objetos a los conglomerados a los que se les ha asignado inicialmente (por iteración).

Este proceder resuelve uno de los inconvenientes principales tradicionalmente atribuidos a los métodos jerárquicos: una mala asignación inicial de los objetos a los conglomerados puede determinar una desacertada solución final

Cea de Anacona, 2004

SELECCIÓN DE MÉTODO DE CONGLOMERACIÓN

Métodos no jerárquicos o de optimización

Se pueden clasificar en tres tipos según los algoritmos de agrupación:

- Métodos de reasignación
- Métodos de búsqueda de densidad
- Métodos directos

Cea de Anacona, 2004

CUADRO 3.2. *Inconvenientes principales de los métodos jerárquicos y no jerárquicos*

MÉTODOS JERÁRQUICOS	*MÉTODOS NO JERÁRQUICOS*
Dificultad de determinar *a priori* el mejor *algoritmo* de clasificación, cuando el investigador desconoce la estructura de la muestra.	Dificultad de conocer *a priori* el número de conglomerados "real" existente en los datos observados.
A menos que se empleen *algoritmos* especiales, es difícil operar con muestras superiores a 200 unidades porque se parte de una *matriz de similaridad.* Al confeccionarse ésta con cada par de objetos (casos o variables) adquiere un tamaño desorbitado, conforme aumenta el tamaño de la muestra. En especial, cuando se clasifican casos. La lectura de los resultados gráficos (mediante el *dendograma* o el gráfico de *carámbanos*) también es difícil de realizar en muestras grandes.	Formar todas las particiones posibles de la serie de datos (que se presenta como la forma más directa de descubrir la partición óptima de una serie de datos), iterativamente, supone la realización de cálculos muy complejos para un número elevado de casos y de conglomerados. Ello dificulta su puesta en práctica.
Una mala partición inicial de los datos no puede modificarse en fases posteriores del proceso de conglomeración.	Una mala decisión inicial sobre el número de conglomerados "real" puede resultar en una errónea clasificación de los datos.
Mayor predisposición a la presencia de "atípicos" (o *outliers*).	Mayor complejidad de los análisis que le hace muy dependiente de la capacidad del ordenador que se utilice.

Cea de Anacona, 2004

CUADRO 3.2. *Inconvenientes principales de los métodos jerárquicos y no jerárquicos*

MÉTODOS JERÁRQUICOS	*MÉTODOS NO JERÁRQUICOS*
Dificultad de determinar *a priori* el mejor *algoritmo* de clasificación, cuando el investigador desconoce la estructura de la muestra.	Dificultad de conocer *a priori* el número de conglomerados "real" existente en los datos observados.
A menos que se empleen *algoritmos* especiales, es difícil operar con muestras superiores a 200 unidades porque se parte de una *matriz de similaridad.* Al confeccionarse ésta con cada par de objetos (casos o variables) adquiere un tamaño desorbitado, conforme aumenta el tamaño de la muestra. En especial, cuando se clasifican casos. La lectura de los resultados gráficos (mediante el *dendograma* o el gráfico de *carámbanos*) también es difícil de realizar en muestras grandes.	Formar todas las particiones posibles de la serie de datos (que se presenta como la forma más directa de descubrir la partición óptima de una serie de datos), iterativamente, supone la realización de cálculos muy complejos para un número elevado de casos y de conglomerados. Ello dificulta su puesta en práctiea.
Una mala partición inicial de los datos no puede modificarse en fases posteriores del proceso de conglomeración.	Una mala decisión inicial sobre el número de conglomerados "real" puede resultar en una errónea clasificación de los datos.
Mayor predisposición a la presencia de "atípicos" (o *outliers*).	Mayor complejidad de los análisis que le hace muy dependiente de la capacidad del ordenador que se utilice.

Cada método de conglomeración ofrece ventajas e inconvenientes o límites importantes

Estos y otros inconvenientes pueden solventarse, si se opta por **combinar métodos** jerárquicos con no jerárquicos, para cubrir un mismo objetivo de investigación

Cea de Anacona, 2004

CUADRO 3.2. *Inconvenientes principales de los métodos jerárquicos y no jerárquicos*

MÉTODOS JERÁRQUICOS	*MÉTODOS NO JERÁRQUICOS*
Dificultad de determinar *a priori* el mejor *algoritmo* de clasificación, cuando el investigador desconoce la estructura de la muestra.	Dificultad de conocer *a priori* el número de conglomerados "real" existente en los datos observados.
A menos que se empleen *algoritmos* especiales, es difícil operar con muestras superiores a 200 unidades porque se parte de una *matriz de similaridad.* Al confeccionarse ésta con cada par de objetos (casos o variables) adquiere un tamaño desorbitado, conforme aumenta el tamaño de la muestra. En especial, cuando se clasifican casos. La lectura de los resultados gráficos (mediante el *dendograma* o el gráfico de *carámbanos*) también es difícil de realizar en muestras grandes.	Formar todas las particiones posibles de la serie de datos (que se presenta como la forma más directa de descubrir la partición óptima de una serie de datos), iterativamente, supone la realización de cálculos muy complejos para un número elevado de casos y de conglomerados. Ello dificulta su puesta en práctiea.
Una mala partición inicial de los datos no puede modificarse en fases posteriores del proceso de conglomeración.	Una mala decisión inicial sobre el número de conglomerados "real" puede resultar en una errónea clasificación de los datos.
Mayor predisposición a la presencia de "atípicos" (o *outliers*).	Mayor complejidad de los análisis que le hace muy dependiente de la capacidad del ordenador que se utilice.

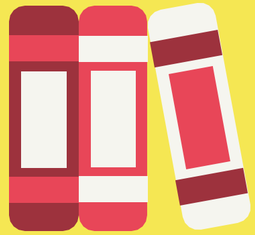
Combinar métodos jerárquicos con no jerárquicos

1. Comenzar con un método jerárquico (identificar número de conglomerados que se pueden formar en la matriz de datos concreta que se analiza, conocimiento de los centroides de Ios conglomerados y los casos atípicos.
2. La solución que resulte se toma como punto de partida del método no Jerárquico, lo que ayuda a ajustar o precisar más la constitución de los conglomerados obtenidos con la aplicación del método jerárquico.

Cea de Anacona, 2004

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**



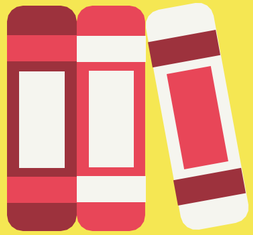
MÉTODOS JERARQUICOS - OPTIMIZACIÓN

Método de centroide ("quick cluster analysis" y método de "Forgy")

Variables numéricas (incluye escalas Likert)
En este método, la distancia entre dos conglomerados se calcula como la **distancia entre los centroides** (promedios de las variables) de los grupos.

Cada vez que se fusionan dos grupos, se recalcula un nuevo centroide para el grupo resultante.

El proceso continúa de forma **aglomerativa** (de abajo hacia arriba), construyendo un **dendrograma jerárquico**.



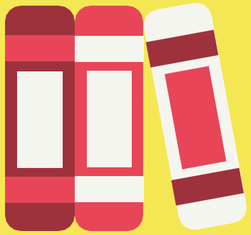
Aldas y Uriel, 2017

Método del centroide

Siguiendo a Aldas y Uriel (2017) se sigue el ejemplo con los datos de ocho empresas, donde se busca saber su la inversión en publicidad afecta o no las ventas (si esta variable implica generación de dos grupos diferenciados)

Cuadro 3.1.: Inversión en publicidad y ventas de 8 empresas hipotéticas

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3	10	22
E4	12	25
E5	45	10
E6	50	15
E7	45	25
E8	50	27



Aldas y Uriel, 2017

Método del centroide

El método del centroide (Sokal y Michener, 1958) está implementado en la función de R, hclust{**stats**}.
Paso 1. Calcular la distancia entre los datos. Se usa la distancia euclídea al cuadrado

```
#calculo de la distancia euclidea al cuadrado

(primero hacemos la sin exponente)

matriz.dis.euclid<-dist (DatosCaso3.1[,c("inversion", "ventas")], method="euclidean",

,diag=TRUE)

matriz.dis.euclid2<-(matriz.dis.euclid)^2
```

Cuadro 3.1.: Inversión en publicidad y ventas de 8 empresas hipotéticas

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3	10	22
E4	12	25
E5	45	10
E6	50	15
E7	45	25
E8	50	27

Cuadro 3.10.: Matriz de distancias euclídeas al cuadrado para los datos del caso 3.1

	1	2	3	4	5	6	7	8
1	0							
2	32	0						
3	180	68	0					
4	241	121	13	0				
5	841	1105	1369	1314	0			
6	1181	1445	1649	1544	50	0		
7	1066	1210	1234	1089	225	125	0	
8	1445	1613	1625	1448	314	144	29	0

Aldas y Uriel, 2017

¿Qué es la distancia euclidiana?

¿Qué es la distancia euclidiana?

Cuadro 3.1.: Inversión en publicidad y ventas de 8 empresas hipotéticas

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3	10	22
E4	12	25
E5	45	10
E6	50	15
E7	45	25
E8	50	27

Grafiquemos y veamos la distancia geométrica entre los puntos

$d = (x_2 - x_1)^2 + (y_2 - y_1)^2$

¿Qué es x_i e y_j ?

$d(P1,P2) = (x_{21} - x_{11})^2 + (x_{22} - x_{12})^2 + (x_{23} - x_{13})^2 + \cdots + (x_{2n} - x_{1n})^2$

donde

$P1 = (x_{11}, x_{12}, x_{13}, \ldots, x_{1n})$
$P2 = (x_{21}, x_{22}, x_{23}, \ldots, x_{2n})$

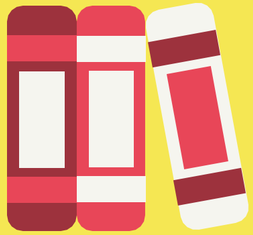
Método del centroide

Paso 2. Análisis de cluster con método centroide

```
hclust.centroide<-hclust(matriz.dis.euclid2,method="centroid")
plot (hclust. centroide, labels-Datos Caso3.1$nombre.
empresa)
```

Cuadro 3.1.: Inversión en publicidad y ventas de 8 empresas hipotéticas

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3	10	22
E4	12	25
E5	45	10
E6	50	15
E7	45	25
E8	50	27



Aldas y Uriel, 2017

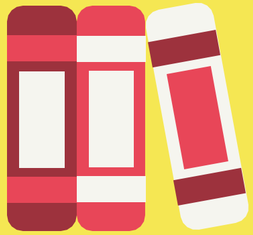
Método del centroide

Paso 3. Sacar el historial de aglomeración del objeto

```
# hclust.centroide data.frame(hclust.centroide [2:1])
```

Cuadro 3.1.: Inversión en publicidad y ventas de 8 empresas hipotéticas

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3	10	22
E4	12	25
E5	45	10
E6	50	15
E7	45	25
E8	50	27



Aldas y Uriel, 2017

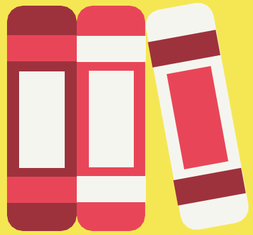
Método del centroide

Este método comienza uniendo aquellas dos observaciones que están más cercanas, en este caso las empresas E3 y E4 (la distancia es 13).

A continuación el grupo formado es sustituido por una observación que lo representa y en la que las variables toman los valores medios de todas las observaciones que constituyen el grupo representado (centroide

Cuadro 3.10.: Matriz de distancias euclídeas al cuadrado para los datos del caso 3.1

	1	2	3	4	5	6	7	8
1	0							
2	32	0						
3	180	68	0					
4	241	121	13	0				
5	841	1105	1369	1314	0			
6	1181	1445	1649	1544	50	0		
7	1066	1210	1234	1089	225	125	0	
8	1445	1613	1625	1448	314	144	29	0



Aldas y Uriel, 2017

Método del centroide

Las empresas E3 y E4 son sustituidas por un conglomerado (E3-4), formado por ambas empresas, que tiene un valor promedio en productividad y en ventas.

$$Publicidad \text{ de E3-4} = \frac{10 + 12}{2} = 11$$

$$Ventas \text{ de E3-4} = \frac{22 + 25}{2} = 23,5$$

Cuadro 3.10.: Matriz de distancias euclídeas al cuadrado para los datos del caso 3.1

	1	2	3	4	5	6	7	8
1	0							
2	32	0						
3	180	68	0					
4	241	121	13	0				
5	841	1105	1369	1314	0			
6	1181	1445	1649	1544	50	0		
7	1066	1210	1234	1089	225	125	0	
8	1445	1613	1625	1448	314	144	29	0

Aldas y Uriel, 2017

Método del centroide

Con esto, se recalcula la matriz de distancias, ahora con el nuevo conglomerado

hclust muestra esas distancias sucesivas en lo que denominamos el historial de conglomeración (Cuadro 3.11).

En el primer paso se fusionaron los casos E3 y E4 y lo hicieron a una distancia *(height*) de 13.
En el segundo paso, se fusionan las empresas E7 y E8, que están a una distancia de 29

Ahora las empresas E7 y E8 serán sustituidas por su centroide E7-8, se recalculará la matriz de distancias y se repetirá el proceso.

El proceso termina cuando todas las empresas están en un solo grupo.

Cuadro 3.11.: Datos en el paso 2 del proceso de conglomeración e historial de conglomeración

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3-4	11	23,5
E5	45	10
E6	50	15
E7	45	25
E8	50	27

Cuadro 3.10.: Matriz de distancias euclídeas al cuadrado para los datos del caso 3.1

	1	2	3	4	5	6	7	8
1	0							
2	32	0						
3	180	68	0					
4	241	121	13	0				
5	841	1105	1369	1314	0			
6	1181	1445	1649	1544	50	0		
7	1066	1210	1234	1089	225	125	0	
8	1445	1613	1625	1448	314	144	29	0

Aldas y Uriel, 2017

Método del centroide

En el historial de conglomeración se observa que en la etapa siguiente (etapa 3) se juntan las empresas E1 y E2 y, en la etapa 4, las E5 y E6.

En la etapa 5 dejan de fusionarse empresas individuales para fusionarse dos grupos (lo que hclust identifica en el cese en el uso del signo −).

Se fusionan las empresas que lo hicieron en el paso 1 (E3-4) con las que lo hicieron en el paso 3 (E1-2), y ese es el indicador (el número del paso) que se muestra en el cuadro

	height	merge.1	merge.2
1	13.00	−3	−4
2	29.00	−7	−8
3	32.00	−1	−2
4	50.00	−5	−6
5	141.25	1	3
6	182.25	2	4
7	1227.25	5	6

Paso 1: Se forma el conglomerado {3,4}
Paso 2: Se forma el conglomerado {7,8}
Paso 3: Se forma el conglomerado {1,2}
Paso 4: Se forma el conglomerado {5,6}
Paso 5: Une el conglomerado nuevo **1** con el nuevo 3
("1" = clúster del **paso 1** = {3,4})
("3" = clúster del **paso 3** = {1,2})
Dando lugar a un nuevo conglomerado **{1,2,3,4}**
Paso 6. fusiona los pares **{5,6}** (paso 4) con **{7,8}** (paso 2) dando lugar a un nuevo conglomerado **{5,6,7,8}**
Paso 7: fusión final: **{1,2,3,4}** con **{5,6,7,8}**

•Primero se arman **cuatro pares naturales**: (3–4), (7–8), (1–2), (5–6).
•Luego se juntan en **dos bloques grandes**:
- **A = {1,2,3,4}** (a height ≈ 141)
- **B = {5,6,7,8}** (a height ≈ 182)

• La **fusión final** A↔B ocurre recién a **1227.25**, un salto grandísimo respecto de 141–182.

•**Solución con 2 conglomerados** (A y B) es muy razonable.
- En dendrograma: "corta" el árbol en cualquier altura **entre ~182 y ~1227** para obtener **k = 2**.

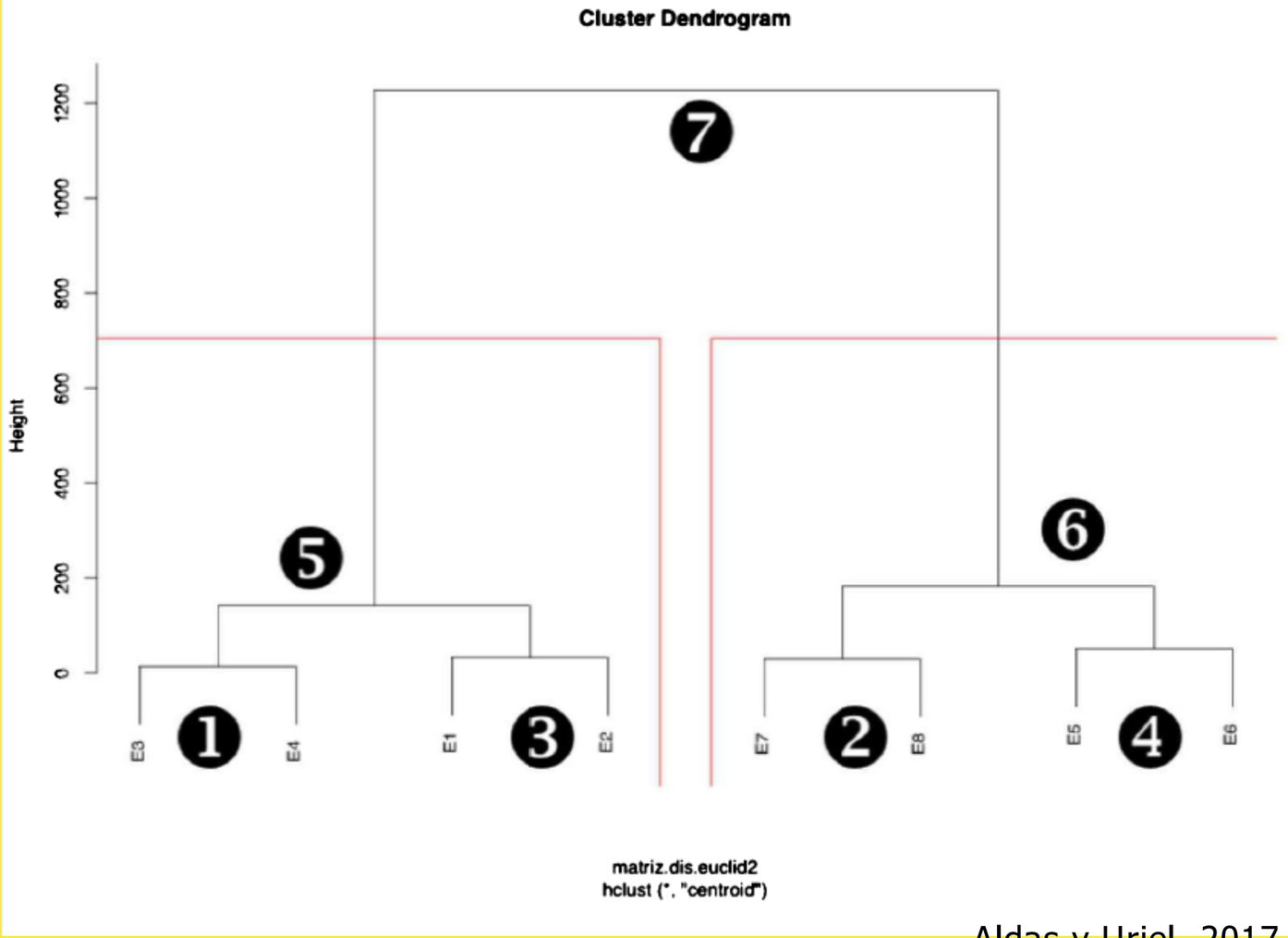
Aldas y Uriel, 2017

Método del centroide - dendrograma

Estemismo historial se puede graficar en un dendrograma

	height	merge.1	merge.2
1	13.00	-3	-4
2	29.00	-7	-8
3	32.00	-1	-2
4	50.00	-5	-6
5	141.25	1	3
6	182.25	2	4
7	1227.25	5	6

Figura 3.4.: Dendograma



Aldas y Uriel, 2017

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)

Calcule al matriz de distancias distancia euclidiana al cuadrado. Interprete

Aplique el análisis de conglomerado con centroide. Interprete.

¿Cuántos conglomerados seleccionó y porqué?

Módulo 2 Ejercicio. 0,2 puntos para Control 2

Códigos para el ejercicio

```
install.packages("amap")
library(amap)

install.packages(“tidyverse")
library(tidyverse)


Seleccionar variables:
data <- elsoc %>%
  select(v1, v2, v3, v4)

# Opción:
# Hacer una selección aleatoria de 100 casos

set.seed(123)
muestra_100 <- data %>%
  slice_sample(n = 100)

#Si lo hace cambie el nombre de los datos)

# estandarice las variables
data.scaled <- scale(data)
d <- Dist(data.scaled, method = "euclidean")
```

Módulo 2 Ejercicio. 0,2 puntos para Control 2

Códigos para el ejercicio

Calcule la distancia euclidiana

```
d <- dist(data.scaled, method = "euclidean")

# Elevar cada distancia al cuadrado
d2 <- d^2

# Ejecutar el análisis de conglomerado
hc <- hclust.centroide(d2)


# Visualizar y cortar en k clusters (ej. k = 4)

plot(hc, hang = -1, main = "Clustering por centroide (amap)")
rect.hclust(hc, k = 4, border = 2)
```

Módulo 2 Ejercicio. 0,2 puntos para Control 2

Códigos para el ejercicio

Analizar los cluster

con el dendrograma defino el número de cluster. Digamos que nos quedamos con 4.
pido que el dendrograma se corte en 4 y le llamo a eso “cluster”. Con eso se genera un asignación de un determinado conglomerado para cada caso

```
cluster <- cutree(hc, k=4)
```

#definimos cluster como una variable, ya que el análisis de conglomerados le asigna un conglomerado a cada caso, es decir un valor entre 1, 2, 3 y 4 (4 conglomerados)

```
data$hc <- factor(cluster)
```

#lo agregamos a nuestra base de datos y le pedimos que nos arroje el valor promedio en cada variable, según el cluster)

```
agreggate(cbind(v1, v2, v3) ~ hc, data = data, FUN = mean)
```

Debiera aparecer algo así (donde aparece el valor promedio de cada variable para cada cluster. Como estandarizo las variables debe considerar eso para interpretar)

	cluster4	ingreso_percapita	escolaridad_media	IDH
1	1	22846.42	10.96250	542.6529
2	2	19729.00	11.03191	632.3560
3	3	2421665.50	9.05000	398.8150
4	4	2699722.00	15.90000	528.0000