FACULTAD DE CIENCIAS ECONÓMICAS

FAMAF
Facultad de Matemática, Astronomía, Física y Computación

UNC Universidad Nacional de Córdoba

DIPLOMATURA

CIENCIA DE DATOS, INTELIGENCIA ARTIFICIAL Y SUS APLICACIONES EN ECONOMÍA Y NEGOCIOS

Análisis de conglomerados (Clustering)

Introducción

Técnica diseñada para clasificar observaciones en grupos (conglomerados o clusters)

- Cada grupo sea homogéneo respecto a las variables utilizadas para caracterizarlo (homogeneidad dentro)
- Que los grupos sean lo más distintos posible unos de otros respecto a las variables consideradas (heterogeneidad entre).

Ejemplo:
El gerente de marketing de una empresa desea dividir a sus clientes en subgrupos que tuvieran características demográficas similares (edad, nivel educativo, ingresos, género, estado civil, ocupación, etc) pero que cada subgrupo fueran lo más diferente posible de otros. Si esto podría diseñar campañas diferentes para cada grupo.

Técnica de clasificación no supervisada: se tiene un conjunto de elementos de los cuales no se conoce su pertenencia a un grupo preestablecido.

Introducción

Proceso de creación de los grupos

1.- Inicialmente se dispone de *n* observaciones (clientes) de las que se tiene información sobre *k* variables (edad, nivel educativo, ingresos, género, estado civil, ocupación, etc)

❶

	X_1	X_2	X_3	...	X_k
O_1					
O_2					
O_3					
...					
O_n					

❷

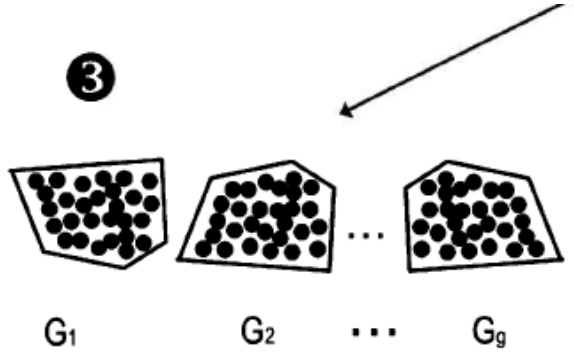
	O_1	O_2	O_3	...	O_n
O_1					
O_2					
O_3					
...					
O_n					

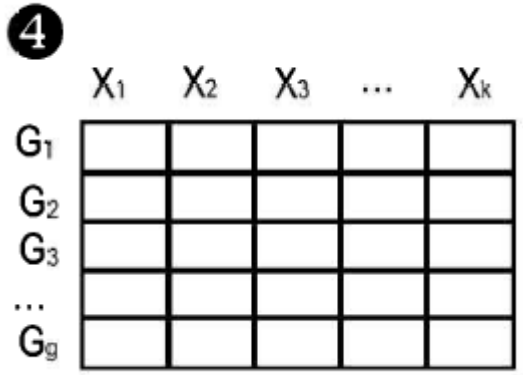
2.-Establecer un indicador que mida la similitud entre dos observaciones: distancia o similaridad.

Introducción

Proceso de creación de los grupos

3.- Crear grupos con aquellas observaciones que más se parezcan, según la distancia establecida, utilizando algún método de agrupamiento: jerárquico o no jerárquico.





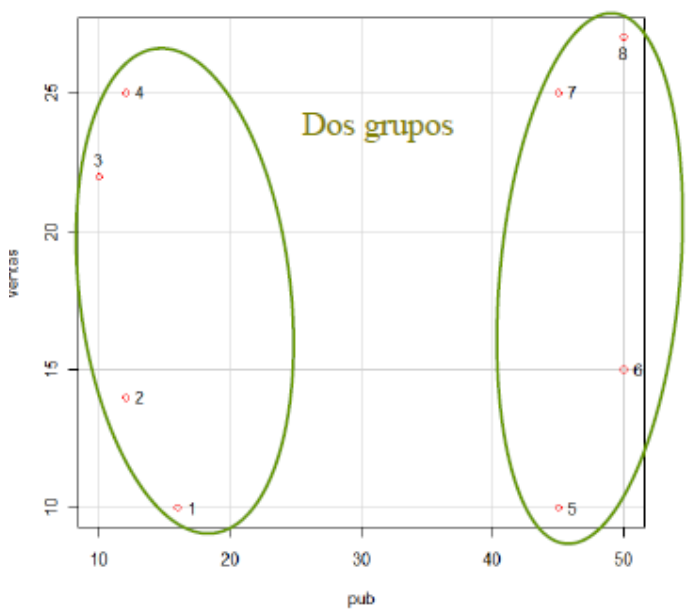
4.-Describir cada grupo y comparar unos con otros.

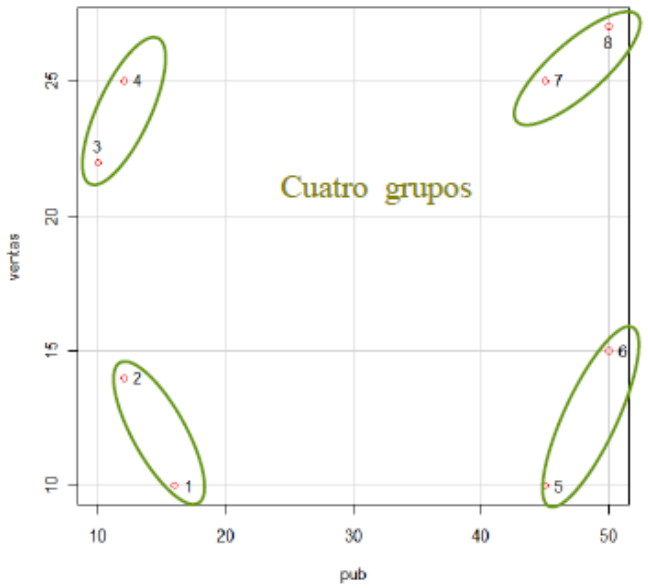
Ejemplo

¿Qué tipología de empresas puede establecerse en función de la rentabilidad obtenida de la inversión publicitaria?

Empresa	Inversión publicitaria	Ventas
1	16	10
2	12	14
3	10	22
4	12	25
5	45	10
6	50	15
7	45	25
8	50	27

Intuitivamente, los grupos se han armado por cercanía.





Medidas de distancia

Variables métricas

Empresa	Inversión publicitaria	Ventas
1	16	10
2	12	14
3	10	22
4	12	25
5	45	10
6	50	15
7	45	25
8	50	27

Euclídea

$$D_{ij} = \sqrt{\sum_{p=1}^{k} (x_{ip} - x_{jp})^2}$$

i y j son observaciones
p variables

Se mide la distancia entre dos individuos

E1 está más cerca de E2 que de cualquier otra empresa. Esa distancia es una medida de proximidad o similitud.

Entre la empresa 1 y 2, la distancia euclídea es 5,66, mientras que la distancia entre la 1 y la 3 es de 13,42.

$$D_{12} = \sqrt{(16-12)^2+(10-14)^2} = 5{,}66$$

$$D_{13} = \sqrt{(16-10)^2+(10-22)^2} = 13{,}42$$

Medidas de distancia

Variables métricas

Empresa	Inversión publicitaria	Ventas
1	16	10
2	12	14
3	10	22
4	12	25
5	45	10
6	50	15
7	45	25
8	50	27

Matriz de distancias euclídeas para los datos del ejemplo

	1	2	3	4	5	6	7	8
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	5.656854	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	13.416408	8.246211	NaN	NaN	NaN	NaN	NaN	NaN
4	15.524175	11.000000	3.605551	NaN	NaN	NaN	NaN	NaN
5	29.000000	33.241540	37.000000	36.249138	NaN	NaN	NaN	NaN
6	34.365681	38.013156	40.607881	39.293765	7.071068	NaN	NaN	NaN
7	32.649655	34.785054	35.128336	33.000000	15.000000	11.18034	NaN	NaN
8	38.013156	40.162171	40.311289	38.052595	17.720045	12.00000	5.385165	NaN

Las menores distancias entre cada par de observaciones indican la proximidad de los individuos

Medidas de distancia

Variables métricas

Distancia Euclídea al cuadrado

$$D_{ij} = \sum_{p=1}^{k} (x_{ip} - x_{jp})^2$$

Distancia de Minkowski

$$D_{ij} = \left[\sum_{p=1}^{k} |x_{ip} - x_{jp}|^n \right]^{1/n}$$

Si n = 2 estamos ante la distancias euclídea

Si n = 1 estamos ante la distancias de Manhattan

City block o Manhattan

$$D_{ij} = \sum_{p=1}^{k} |x_{ip} - x_{jp}|$$

Euclídea estandarizada

$$D_{ij} = \sqrt{\sum_{p=1}^{k} \frac{(x_{ip} - x_{jp})^2}{S_{ij}}}$$

Estandarización de valores

Para evitar el efecto de la unidad de medida en cada una de estas distancias, se estandarizan los datos.

Puntuación Z

$$z_{ip} = \frac{x_{ip} - \bar{x}_p}{S_p}$$

Rango 1

$$z_{ip} = \frac{x_{ip}}{max_p - min_p}$$

Rango 0 a 1

$$z_{ip} = \frac{x_{ip} - min_p}{max_p - min_p}$$

Matriz de distancias euclídeas para los datos estandarizados

	1	2	3	4	5	6	7	8
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	0.649050	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	1.856060	1.221720	NaN	NaN	NaN	NaN	NaN	NaN
4	2.292081	1.672629	0.470025	NaN	NaN	NaN	NaN	NaN
5	1.642460	1.965484	2.694237	2.948813	NaN	NaN	NaN	NaN
6	2.070298	2.157554	2.503051	2.635158	0.811312	NaN	NaN	NaN
7	2.810691	2.508161	2.034090	1.869006	2.280858	1.546716	NaN	NaN
8	3.223380	2.922230	2.389634	2.173569	2.600437	1.824686	0.415545	NaN

Medidas de similaridad

Variables binarias

Si las variables son binarias (posee o no una determinada característica) o dicotómicas

	Obs. j		
Obs. i	**Presencia**	**Ausencia**	**Suma**
Presencia	a	b	a+b
Ausencia	c	d	c+d
Suma	a+c	b+d	a+b+c+d

Ejemplo:

Obs	**Variables**			
	X1	**X2**	**X3**	**X4**
1	1	1	0	0
2	0	1	1	1

	Obs. 1	
Obs. 2	**Presencia**	**Ausencia**
Presencia	1	2
Ausencia	1	0

Distancia euclídea

$$D_{ij} = \sqrt{b+c}$$

Proporción de coincidencias (Sokal y Michener)

$$D_{ij} = \frac{a+d}{a+b+c+d}$$

Proporción de apariciones (Jaccard)

$$D_{ij} = \frac{a}{a+b+c}$$

Formación de los grupos

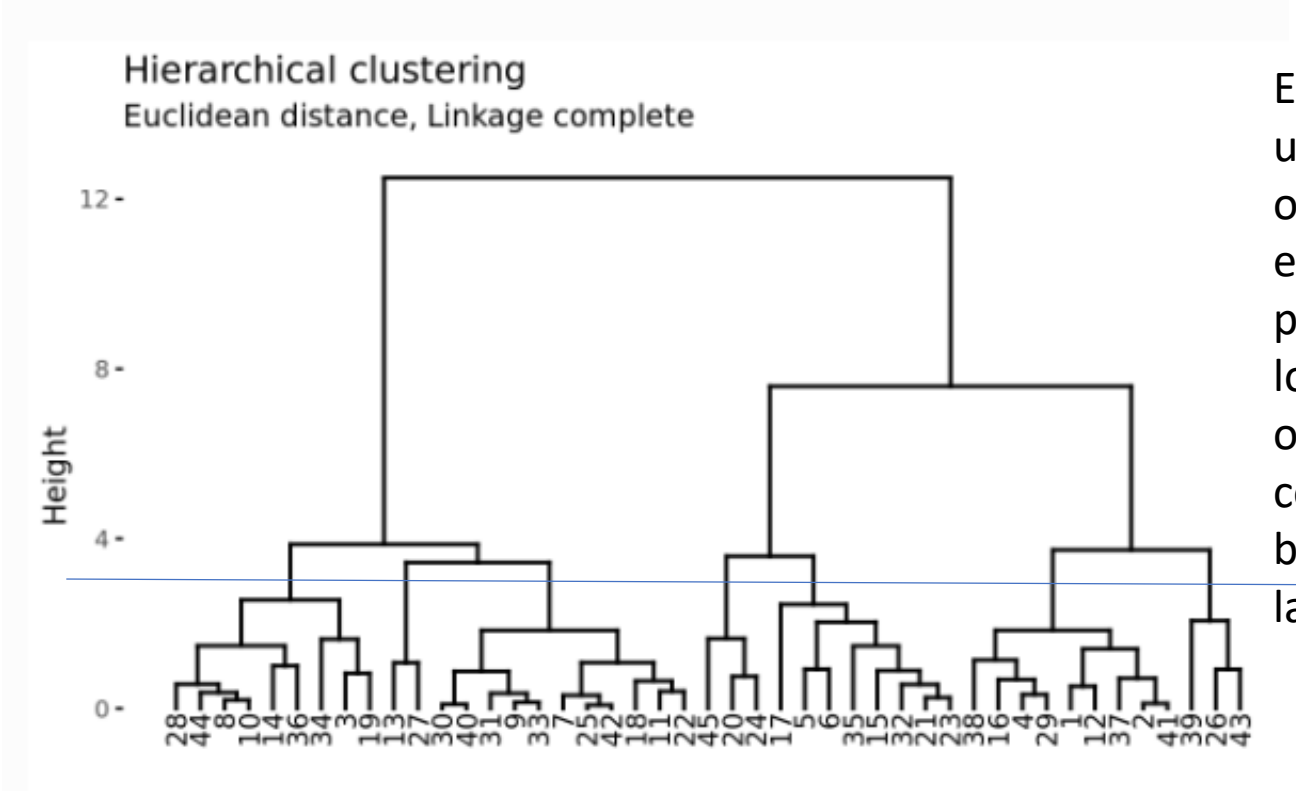
Formación de grupos: selección del algoritmo de decisión y determinación del numero de grupos en función de los datos.

Algoritmos de agrupación

- Jerárquicos
  - Aglomerativos
    - Centroide
    - Vecino más cercano
    - Vecino más lejano
    - Ward
    - Vinculación promedio
  - Desagregativos o divisivos
- No Jerárquicos
  - K-Means

Algoritmos jerárquicos

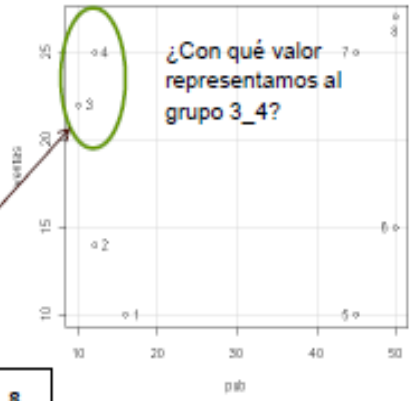
Los métodos jerárquicos realizan sucesivas fusiones de los objetos dando por resultados una serie de particiones encerradas las unas en las otras, que son resumidas en un diagrama, denominado dendograma que ilustra las agrupaciones que han sido efectuadas en las sucesivas etapas.



En la base del dendrograma, cada observación forma una terminación individual conocida como hoja o leaf del árbol. A medida que se asciende por la estructura, pares de hojas se fusionan formando las primeras ramas. Estas uniones se corresponden con los pares de observaciones más similares. También ocurre que las ramas se fusionan con otras ramas o con hojas. Cuanto más temprana (más próxima a la base del dendrograma) ocurre una fusión, mayor es la similitud.

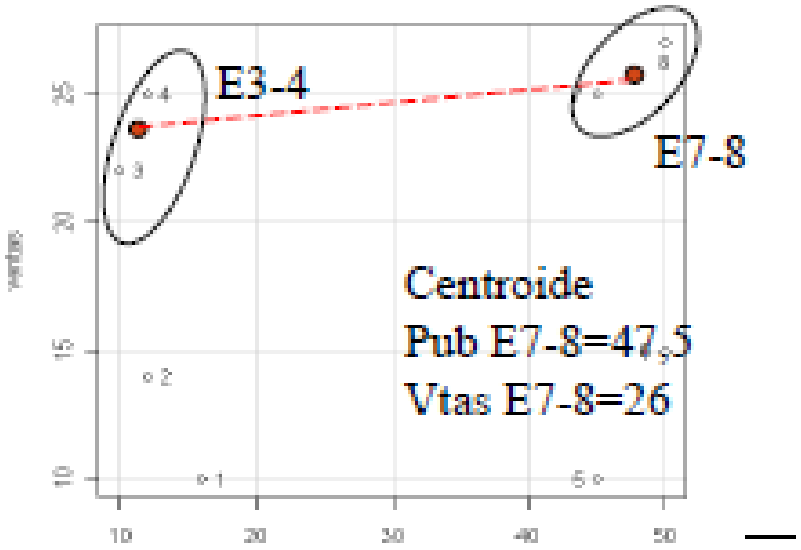
Método del centroide

- Se unen las dos observaciones más cercanas.
- El grupo formado es sustituido por una observación que lo representa (su promedio o centroide) y así las variables toman los valores medios de todas las observaciones que constituyen el grupo representado.
- Se recalcula la matriz de distancias, cada vez que hay un agrupamiento, se unen aquellas dos observaciones que están de nuevo mas cerca y se repetirá el proceso. Esto termina cuando todas las observaciones están en un solo grupo.



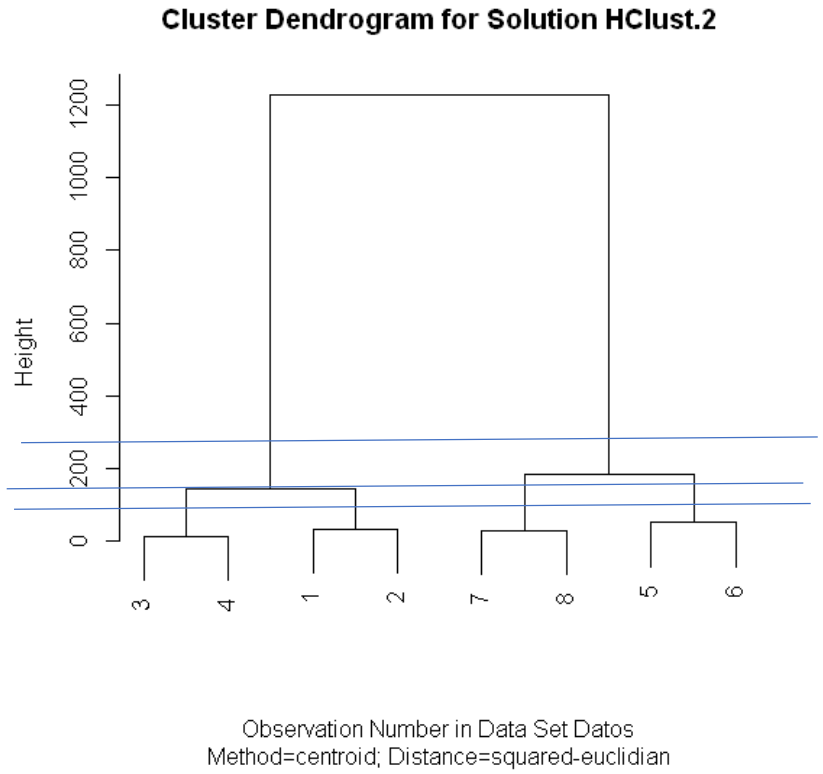
Matriz de distancias euclideas al cuadrado

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



Método del centroide

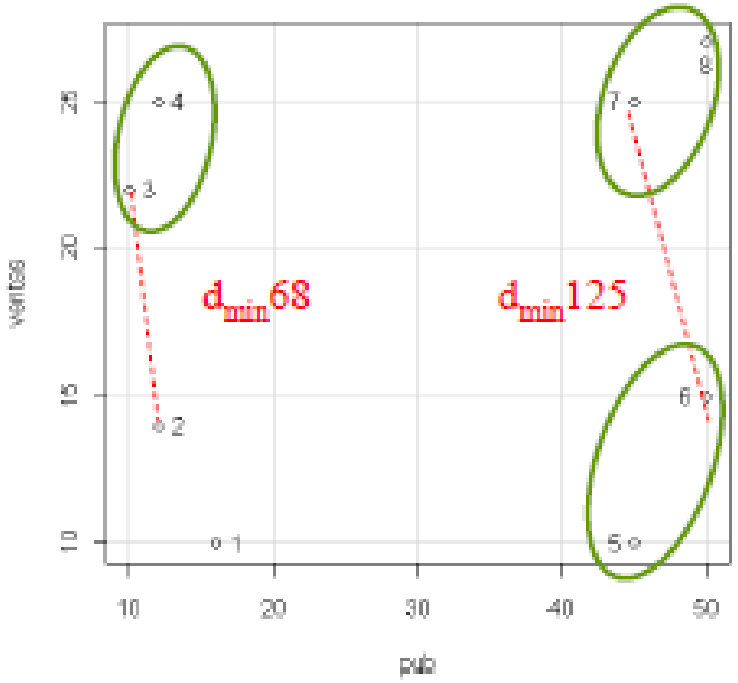
- El historial de conglomeración muestra que primero se unieron las empresas 3 y 4 a una distancia de 13. Estas empresas son reemplazadas por su centroide y se recalcula la matriz de distancias.
- En el paso 5 dejan de fusionarse empresas individuales y comienzas a unirse grupos de empresas.
- Termina cuando todas las empresas están en un solo grupo (paso 7).



Etapa	Agrupamiento		Distancia	Nro grupos
[1,]	3	4	13	7
[2,]	7	8	29	6
[3,]	1	2	32	5
[4,]	5	6	50	4
[5,]	[3-4]	[1-2]	141,25	3
[6,]	[7-8]	[5-6]	182,25	2
[7,]	[3-4-1-2]	[7-8-5-6]	1227,3	1

Método del vecino más cercano (vinculación simple)

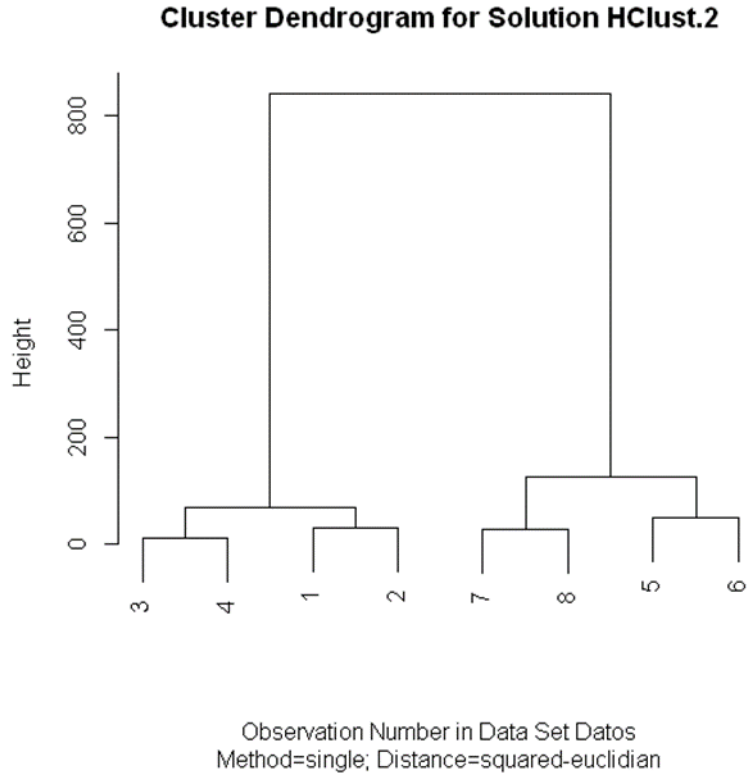
Este método no calcula la distancia entre los centroides, sino que calcula la distancia entre dos grupos, como aquella que se da entre los individuos mas cercanos de esos grupos.



Matriz de distancias euclídeas al cuadrado

	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

Método del vecino más cercano (vinculación simple)



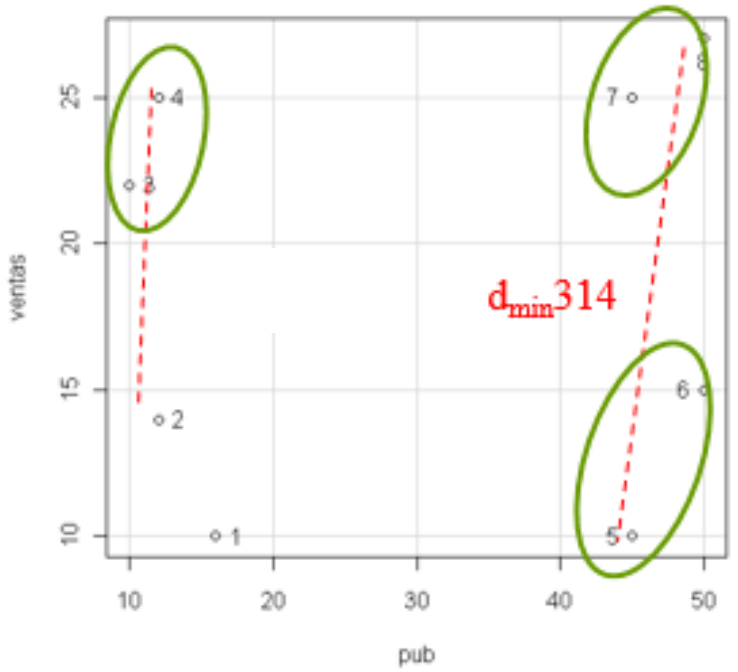
Etapa	Agrupamiento		Distancia	Nro grupos
[1,]	3	4	13	7
[2,]	7	8	29	6
[3,]	1	2	32	5
[4,]	5	6	50	4
[5,]	[3-4]	[1-2]	68	3
[6,]	[7-8]	[5-6]	125	2
[7,]	[3-4-1-2]	[7-8-5-6]	841	1

Método del vecino más lejano (vinculación compleja)

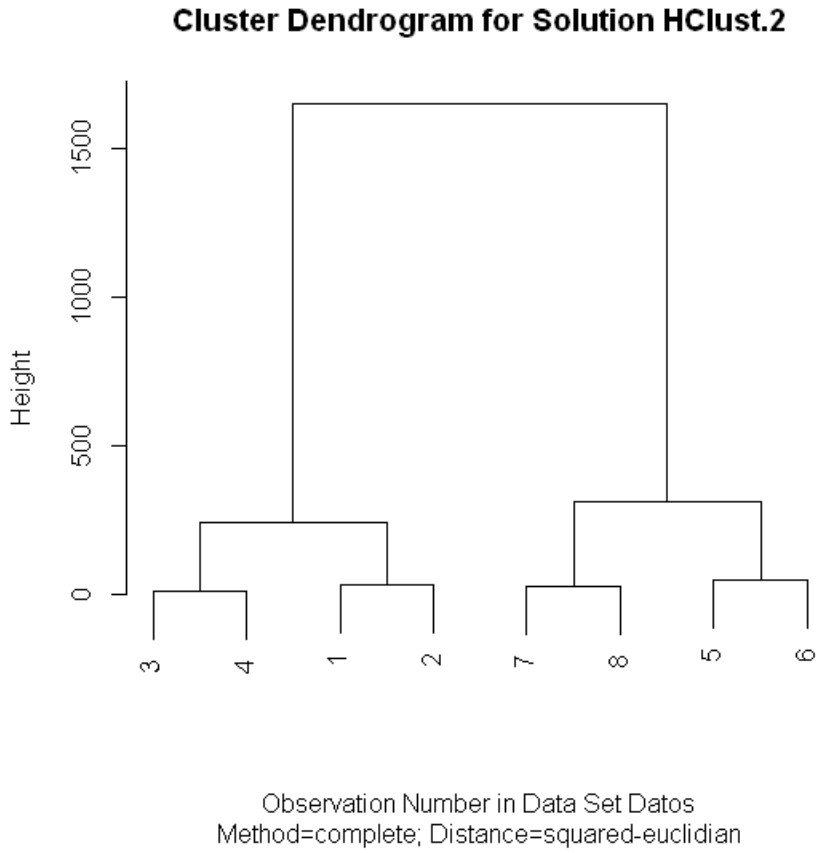
Este método calcula la distancia entre dos grupos, como aquella que se da entre los individuos mas lejanos de esos grupos.

Matriz de distancias euclideas al cuadrado

	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



Método del vecino más lejano (vinculación compleja)



Etapa	Agrupamiento		Distancia	Nro grupos
[1,]	3	4	13	7
[2,]	7	8	29	6
[3,]	1	2	32	5
[4,]	5	6	50	4
[5,]	[3-4]	[1-2]	241	3
[6,]	[7-8]	[5-6]	314	2
[7,]	[3-4-1-2]	[7-8-5-6]	1649	1

Método de la vinculación promedio

Este método calcula la distancia entre dos grupos, como la distancia promedio entre todos los pares de observaciones que pueden formarse tomando un individuo de un grupo y otro individuo de otro grupo.

Matriz de distancias euclideas al cuadrado

	1	**2**	**3**	**4**
1	0	32	180	241
2	32	0	68	121
3	180	68	0	13
4	241	121	13	0

Pares de observaciones	**Distancia**	**Promedio**
E1, E3	180	152,5
E1, E4	241	
E2, E3	68	
E2, E4	121	

Método de Ward (mínima varianza)

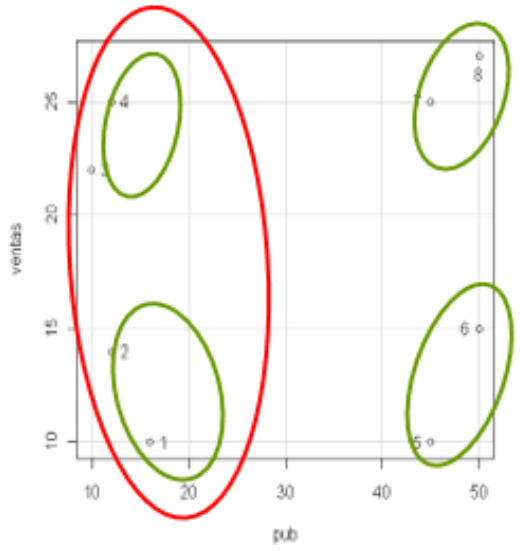
Este método no calcula distancia entre los distintos conglomerados para decidir cuales se deben fusionar, ya que su objetivo es maximizar la homogeneidad dentro de cada conglomerado. Para ello plantea todas las posibles combinaciones de observaciones para el número de grupos que se esté considerando en cada etapa concreta.

- Calcula los centroides de las posibles fusiones.
- Calcula la distancia euclídea al centroide de todas las observaciones del grupo (suma de cuadrados total).
- La solución con menor variabilidad dentro de los grupos garantiza mayor homogeneidad.

En cada etapa se combinan aquellos dos clusters que producen el mínimo incremento en la suma de cuadrados total (SCT).

Matriz de incrementos en la SCT

	1-2	**3-4**	**5-6**	**7-8**
1-2	0			
3-4	282,5	0		
5-6	2245	2907	0	
7-8	2637	2677	364,5	0



Algoritmo no jerárquico

En este caso se conoce a priori el número *k* de grupos que se desea y las observaciones son asignadas a cada uno de esos *k* conglomerados.

- Se determinan los centroides iniciales (medias) de las variables que caracterizan a las observaciones en cada uno de los *k* grupos (semilla).

- Cada observación se asigna a aquel conglomerado, dentro de los *k* existentes, cuyo centroide esta mas cercano a esa observación en términos de distancia euclídea.

- Se recalculan los centroides con las observaciones asignadas a cada grupo y se comparan los viejos centroides con estos nuevos hasta conseguir un criterio de convergencia.

Selección de métodos

¿Cuál es el método que debo elegir?

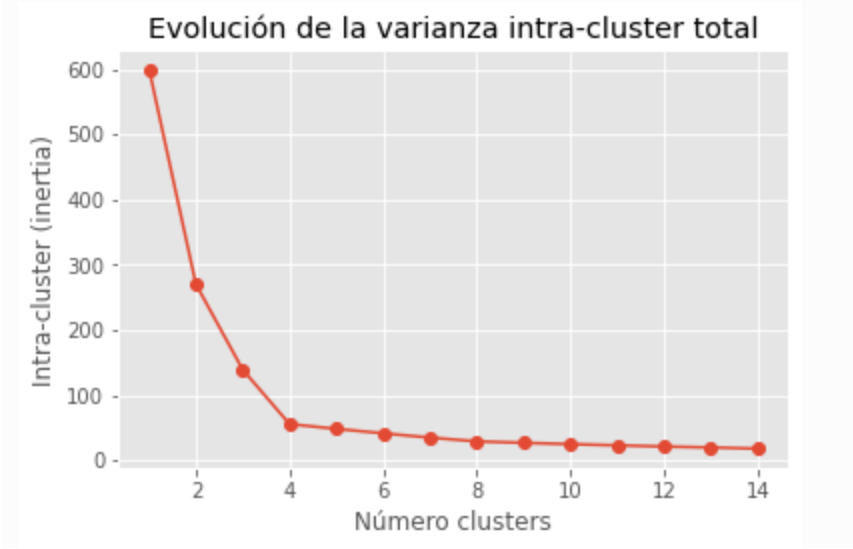
- Depende de los objetivos del estudio y de las propiedades de los distintos métodos (Hair *et al.,* 2014).
- Si el investigador no tiene ninguna idea a priori se sugiere aplicar algoritmos jerárquicos primero.

- El método del centroide es el más utilizado.
- El método del vecino más cercano es más sensible a la presencia de observaciones anómalas (atípicos) que el método del vecino más lejano.
- El método del vecino más lejano identifica habitualmente grupos muy homogéneos, en los que las observaciones son muy parecidas unas a otras.
- El método del vecino mas cercano tiene tendencia a crear menos grupos que el del vecino mas lejano.
- El método de Ward tiende a encontrar conglomerados no solo muy compactos, sino también de tamaño similar.

Selección del número de clusters

¿Cómo definir el número de clusters?

- Dendograma: detener el proceso de fusión cuando los grupos a unir están a una distancia mayor que los que previamente se han fusionado.
- El método *Elbow*, también conocido como método del codo, calcula la varianza total *intra-cluster* en función del número de clusters y escoge como óptimo aquel valor a partir del cual añadir más clusters apenas consigue mejoría.



A partir de 4 clusters la reducción en la suma total de cuadrados internos parece estabilizarse, indicando que $K = 4$ es una buena opción.

Selección del número de clusters

- Cálculo del coeficiente $S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$ que cuantifica qué tan buena es la asignación que se ha hecho de una observación comparando su similitud con el resto de observaciones de su cluster frente a las de los otros clusters.
- Su valor puede estar entre -1 y 1, siendo valores próximos a 1 un indicativo de que la observación se ha asignado al clúster correcto.

a_i= distancia promedio entre la observación i y el resto de las observaciones del mismo cluster.

b_i= distancia promedio entre la observación i y el cluster más próximo.

Algoritmo no jerárquico

"Online" es un conjunto de que contiene todas las transacciones que ocurrieron entre el 01/12/2010 y el 09/12/2011 para un comercio minorista en línea (Reino Unido). La empresa vende principalmente regalos únicos para todas las ocasiones. Muchos clientes de la empresa son mayoristas. El objetivo es realizar un agrupamiento de los clientes y elegir a qué grupos dirigirse.

Vamos a analizar los Clientes en función de los siguientes factores:

- Días: número de días desde la última compra
- Frecuencia: número de transacciones
- Gasto: gastos totales

Los clientes del Cluster 1 son los clientes con una gran cantidad de transacciones en comparación con otros clientes. Su gasto es superior al de otros clientes ya que son compradores frecuentes.

Los clientes del Cluster 2 no son compradores frecuentes y su última compra fue hace más tiempo, por lo tanto, son los menos importantes desde el punto de vista comercial.

Los clientes del Cluster 0 realizaron su compra más recientemente pero tienen poca frecuencia de compra.