

ANÁLISIS DE TIPOLOGÍAS / CLUSTER / CONGLOMERADOS

Parte 3

CURSO: Estadística IV

CARRERA: Sociología

UNIVERSIDAD ALBERTO HURTADO

PROFESORA: CAROLINA AGUILERA

AYUDANTES: Miguel Tognarelli y

Vicente Díaz



HOY

MIRADA GENERAL ANÁLISIS DE
CLUSTER

DETERMINACIÓN DE NÚMERO DE
CONGLOMERADOS
ALGORITMOS NO JERÁRQUICOS
EJEMPLO



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

ANÁLISIS DE TIPOLOGÍAS, CONGLOMERADOS , CLUSTER

VARIAS TÉCNICAS: 2 GRANDES TIPOS

Técnicas analíticas multivariadas 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.

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

*"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.

TIPOS DE ALGORITMOS

- MODELOS JERÁRQUICOS
- MODELOS NO JERÁRQUICOS



NO CONFUNDIR CON LAS MEDIDAS DE DISTANCIA Y SIMILITUD

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 **teóricamente 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.



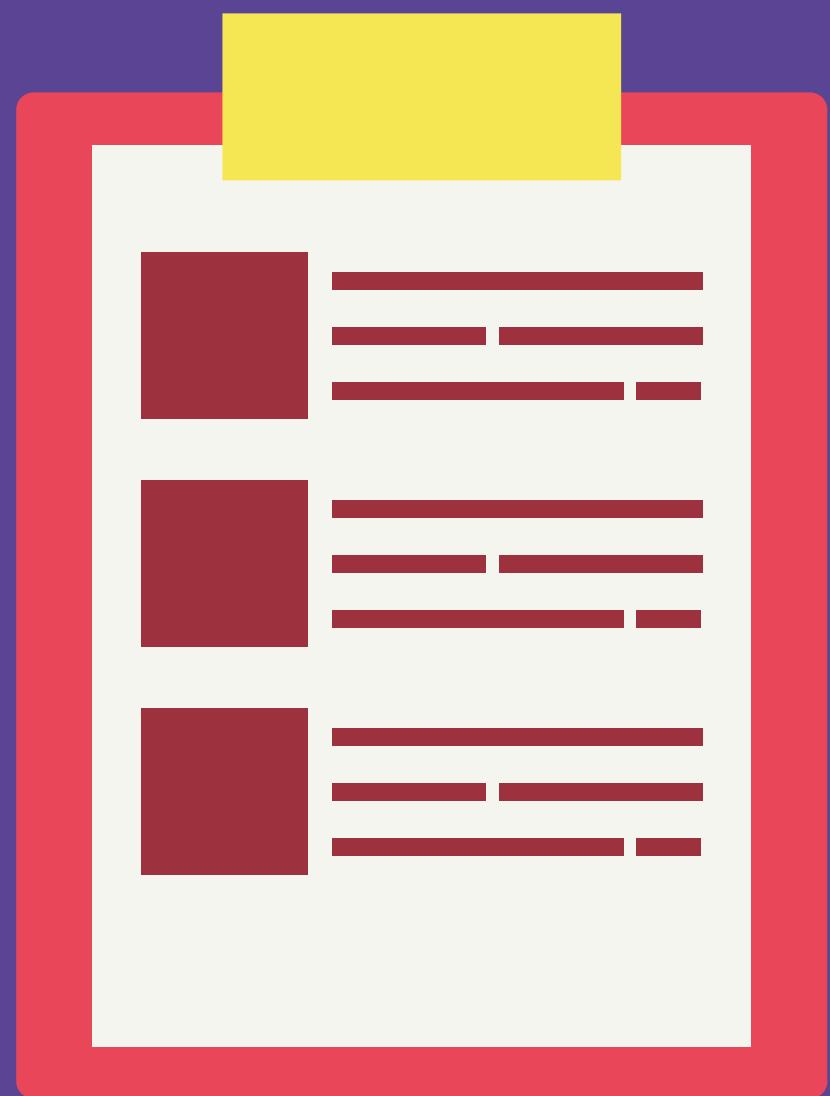
CUADRO 3.3. *Medidas de distancia o de similaridad, según el nivel de medición de las variables*

A. VARIABLES CONTINUAS	A.1. <i>Medidas de distancia</i>	<ul style="list-style-type: none"> • 1.1. Euclídea • 1.2. Euclídea al cuadrado • 1.3. D^2 de Mahalanobis • 1.4. De Manhattan o “city-block” • 1.5. De Chebychev • 1.6. De Minkowski • 1.7. De un poder métrico absoluto
	A.2. <i>Medidas de similaridad</i>	<ul style="list-style-type: none"> • 2.1. Correlación de Pearson • 2.2. Cosenos de vectores de valores
B. VARIABLES BINARIAS	B.1. <i>Medidas de similaridad</i>	<ul style="list-style-type: none"> • 1.1. De Jaccard • 1.2. De casación o parejas simples • 1.3. De Russel y Rao • 1.4. De Dice • 1.5. De Rogers y Tanimoto • 1.6. De Kulczynski 1 • 1.7. De Sokal y Sneath • 1.8. De correlación punto 4 phi (ϕ) • 1.9. De Ochiai • 1.10. De dispersión
	B.2. <i>Medidas de similaridad de probabilidades condicionales</i>	<ul style="list-style-type: none"> • 2.1. De Kulczynski 2 • 2.2. De Sokal y Sneath 4 • 2.3. De Hamann
	B.3. <i>Medidas de similaridad de predicción</i>	<ul style="list-style-type: none"> • 3.1. Lambda de Goodman y Kruskal • 3.2. D de Anderberg • 3.3. Y de Yule • 3.4. Q de Yule
	B.4. <i>Medidas de disimilaridad o distancia</i>	<ul style="list-style-type: none"> • 4.1. Euclídea binaria • 4.2. Diferencia de tamaño • 4.3. Diferencia de patrón • 4.4. Diferencia binaria de forma • 4.5. Varianza disimilar • 4.6. De Lance y Williams
C. VARIABLES CUALITATIVAS NO BINARIAS	C.1. <i>Medidas de similaridad</i>	<ul style="list-style-type: none"> • 1.1. Chi-cuadrado • 1.2. Phi-cuadrado
D. VARIABLES EN DIFERENTES NIVELES DE MEDICIÓN	D.1. <i>Medidas de similaridad</i>	1.1. Coeficiente de similaridad de Gower

SELECCIÓN DE LAS VARIABLES

El tipo de variable condiciona el tipo de métrica o forma de medir “similitud”

DETERMINACIÓN DE NÚMERO DE CLUSTERS

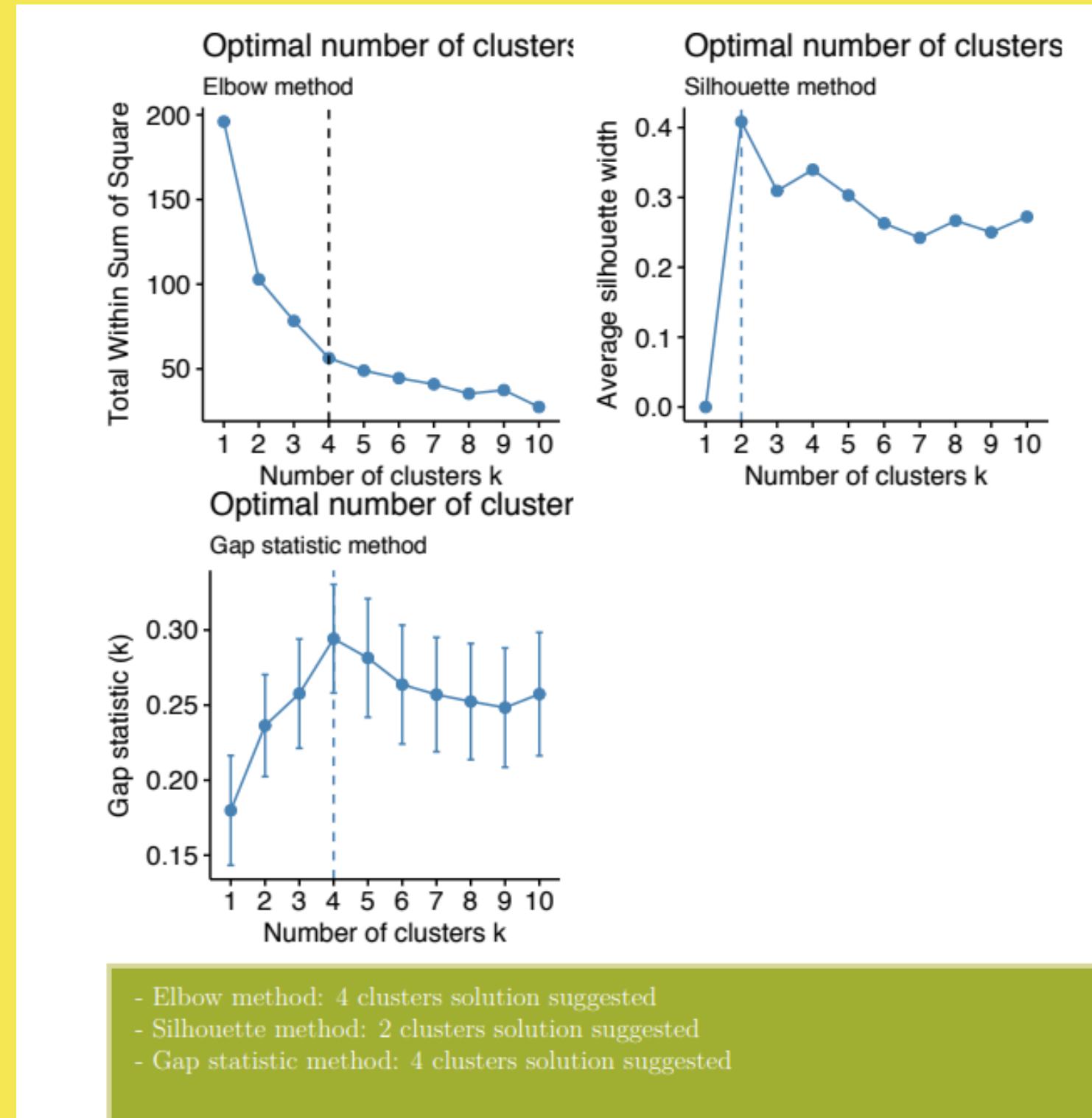


¿Cómo saber el número de clusters?

Existen diferentes métodos, uno es el dendrograma para los algoritmos jerárquicos

Sirven para diferentes métodos, pero no siempre para variables categóricas

Veremos tres alternativas



According to these observations, it's possible to define $k = 4$ as the optimal number of clusters in the data.

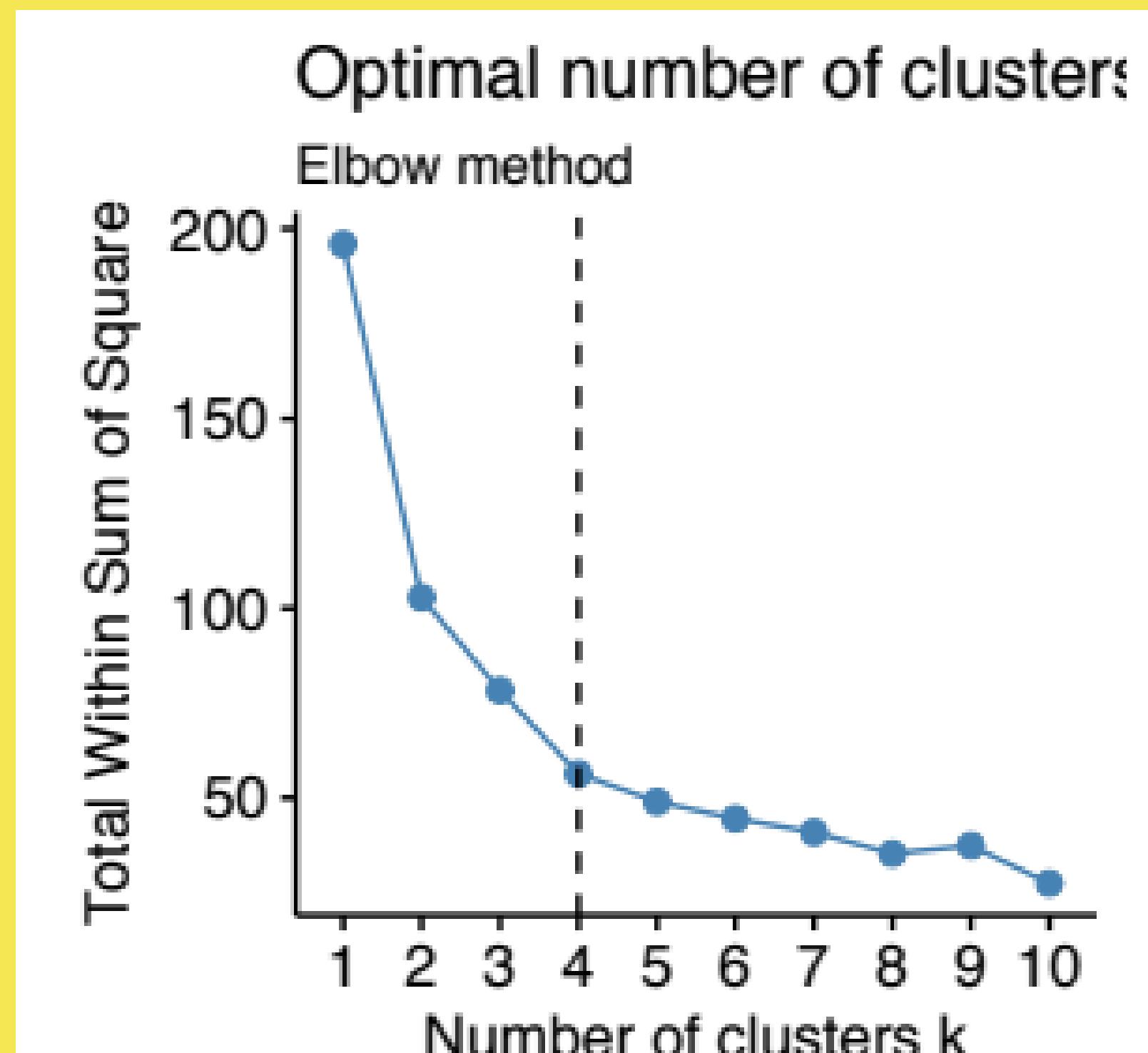
¿Cómo saber el número de clusters? (sirve para los diferentes métodos)

Métodos directos

Otimizan un criterio numérico, es decir, buscan el número de clústeres que maximiza o minimiza una medida de calidad del agrupamiento.

Los más comunes son:

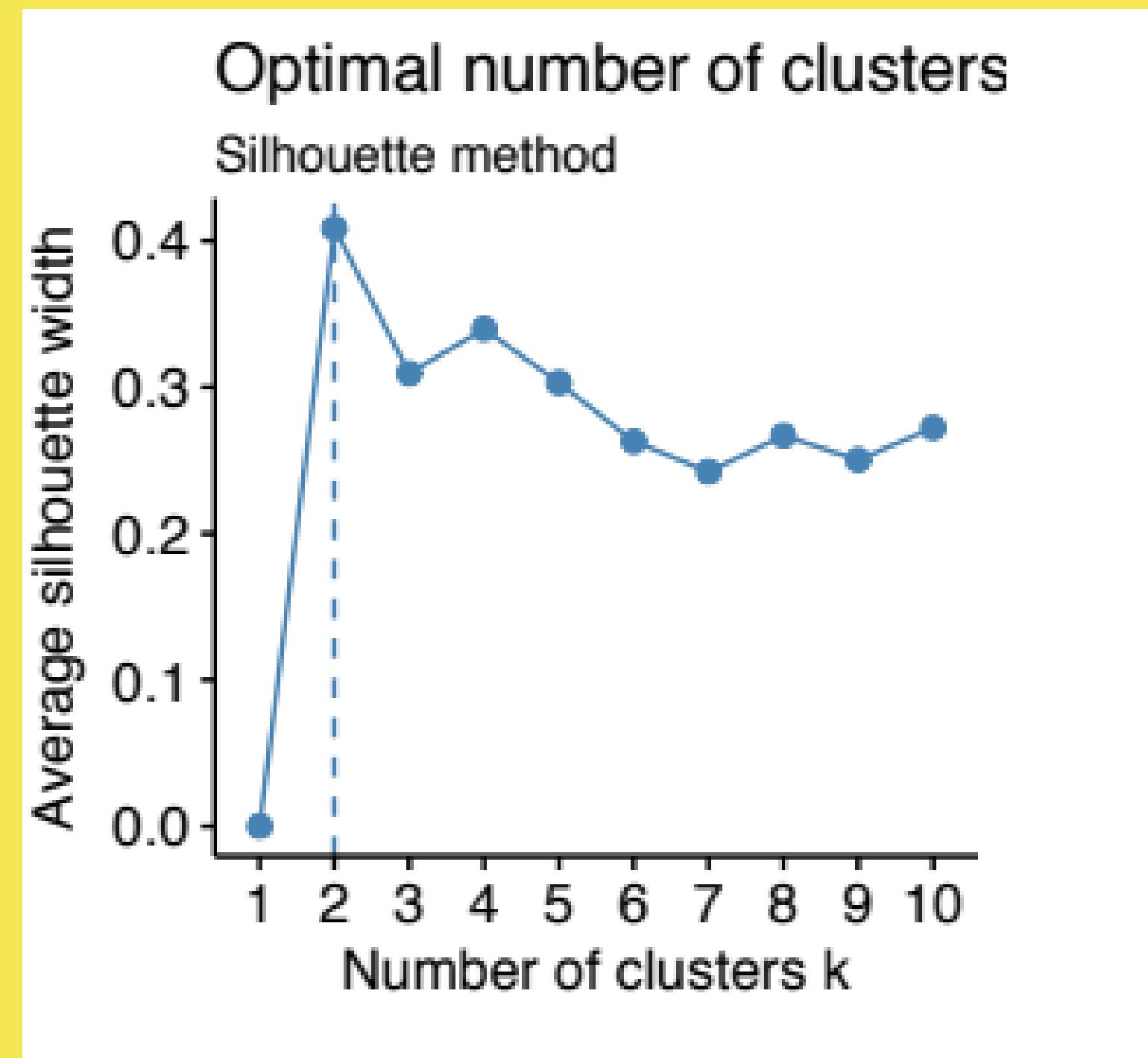
- **Método del codo** (elbow method): busca el punto donde dejar de aumentar el número de clústeres ya no reduce significativamente la variación interna (within-cluster sum of squares). (variables numéricas)
- Visualmente, es el “codo” en la curva.



¿Cómo saber el número de clusters? (sirve para los diferentes métodos)

1. Métodos directos

- **Método del silhouette:** elige el número de clústeres que maximiza el promedio del índice de silhouette, que mide qué tan bien cada observación se ajusta a su clúster frente a otros.
- Sirve para variables categóricas



(ver código de R en Kassambara, 2017: 133)

Kassambara, 2017

¿Cómo saber el número de clusters? (sirve para los diferentes métodos)

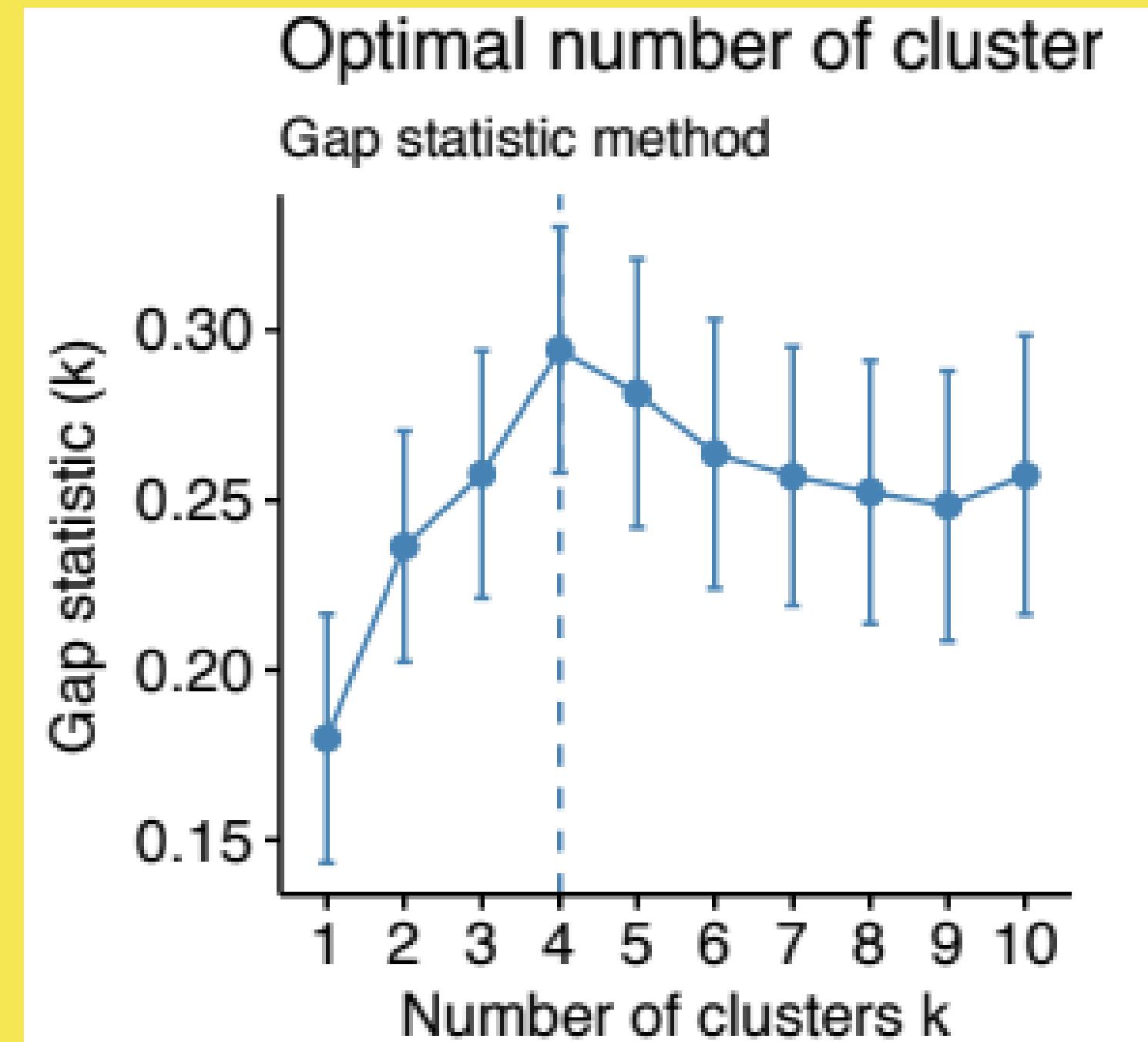
2. Métodos basados en pruebas estadísticas

Estos consisten en comparar la evidencia observada con una hipótesis nula, que generalmente dice que “los datos no tienen estructura de clúster”.

Gap (gap statistic):

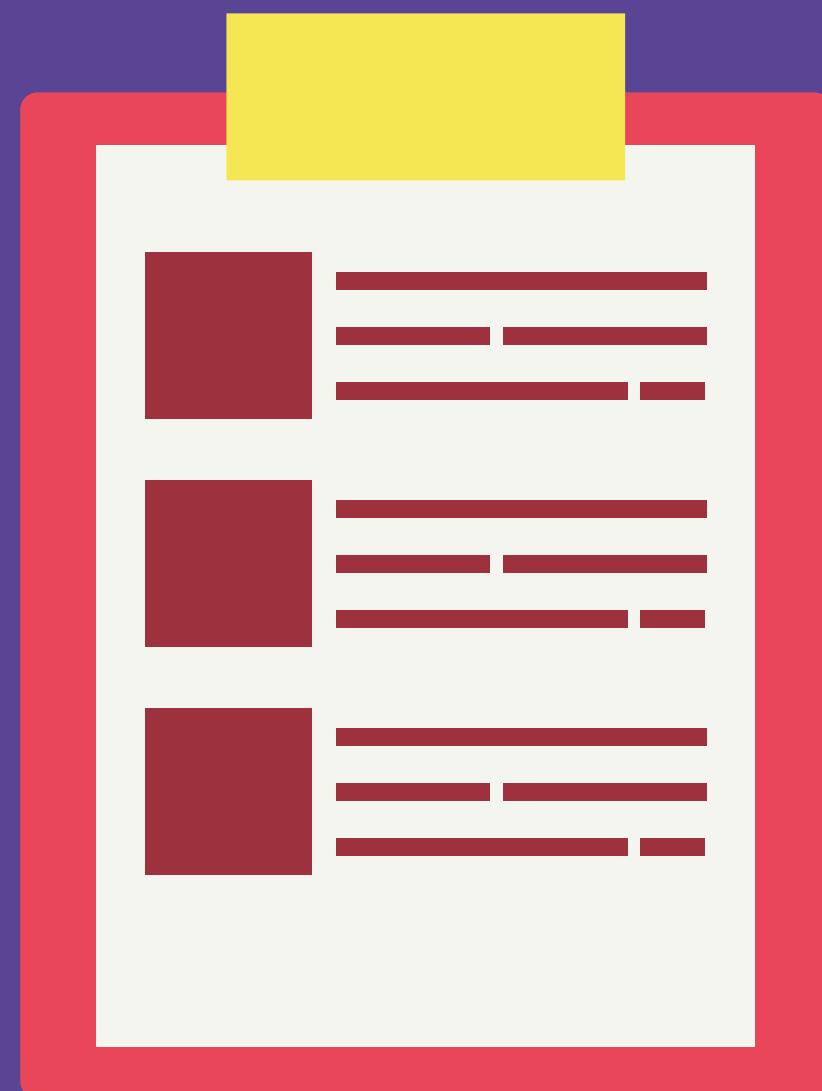
- Compara la dispersión dentro de los clústeres reales con la esperada bajo una distribución aleatoria (sin estructura).
- El número óptimo de clústeres es aquel donde la diferencia (“gap”) entre ambas dispersiones es máxima.

(ver código de R en Kassambara, 2017: 133)



Kassambara, 2017

EJEMPLO CON VARIABLE CATEGÓRICA



● Método jerárquico

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

VARIABLES CATEGÓRICA:

Se usa una medida de similitud para ese tipo de variables

Sirve para 2, 3 y 4

Ejemplo de método aglomerativo con variables categóricas

Se usará la distancia de Jackard y el método de “vínculo promedio”

- La distancia de Jackard es una de las más usadas para variables categóricas. Convierte las variables en variables dummy
- Vínculo promedio: al fusionar dos grupos, para el paso siguiente calcula la distancia entre ellos como el promedio de las distancias entre todos los pares de observaciones (una de cada grupo).
- Ver el script en documento adjunto



Ejemplo de método aglomerativo con variables categóricas

Considere una análisis exploratorio de una base de datos de 500 personas, con las variables categóricas

-
- nivel educacional
- nivel de ingresos
- edad
- posición política

Paso 1. Determinación de número de clusters

EJEMPLO DE MÉTODO AGLOMERATIVO CON VARIABLES CATEGÓRICAS

Con la función “model.matrix” se crean las variables dummy y la matriz de distancias entre estas, para luego hacer el análisis de cluster

```
X <- model.matrix(~ ingresos + educacion + edad + eje_politico - 1, data =  
datos_cluster)
```

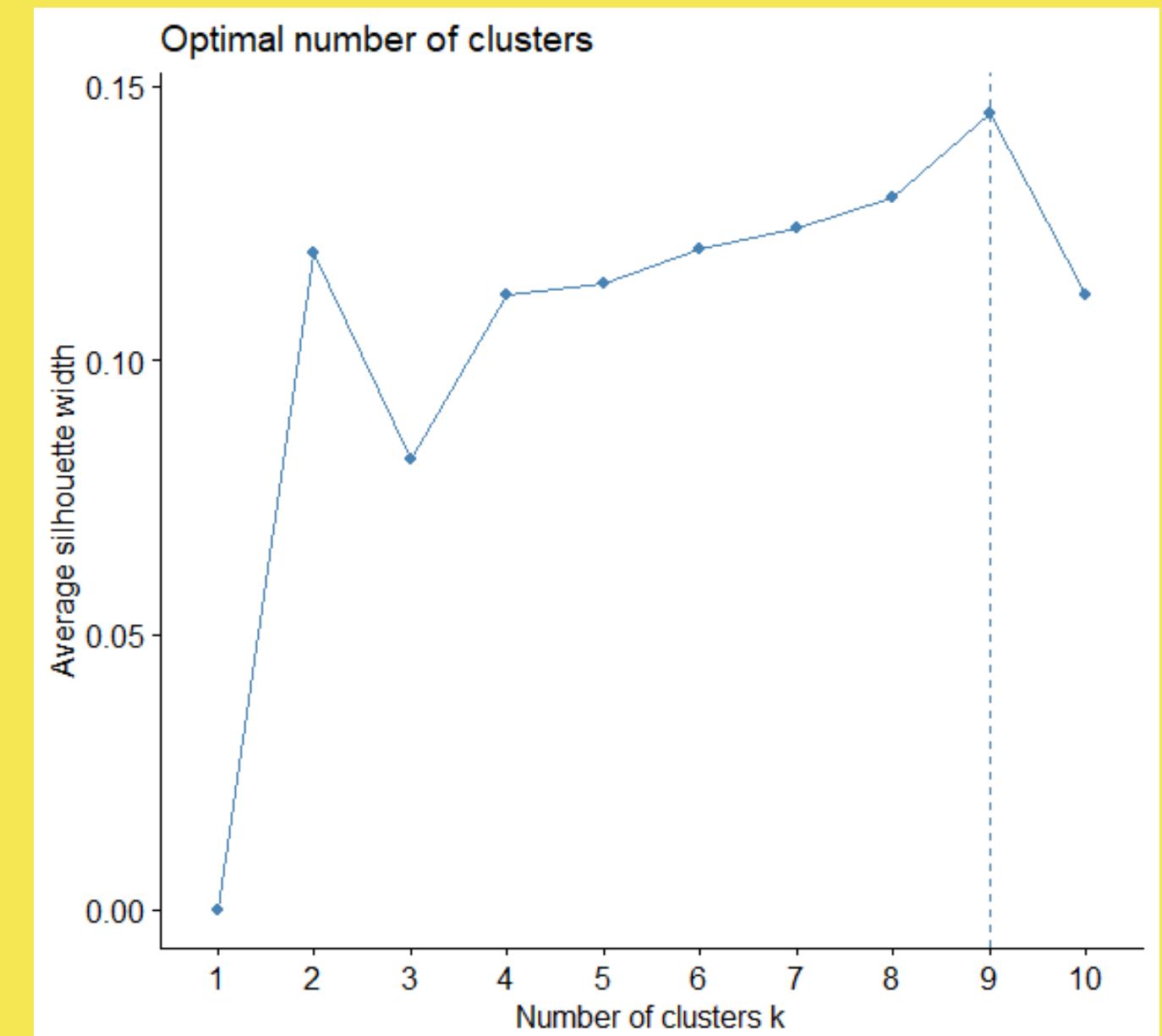
X es matriz de variables dummy (0 y 1), que representa las variables categóricas, y que se usarán para hacer el cluster

¿Cómo saber el número de clusters? (sirve para los diferentes métodos)

Método del silhouette

```
fviz_nbclust(X, FUN = hcut,  
method = "silhouette", diss =  
dist_jaccard)
```

```
gr9  
 1   2   3   4   5   6   7   8   9  
62 131 112 41  62  2  36  49  5  
|> |
```



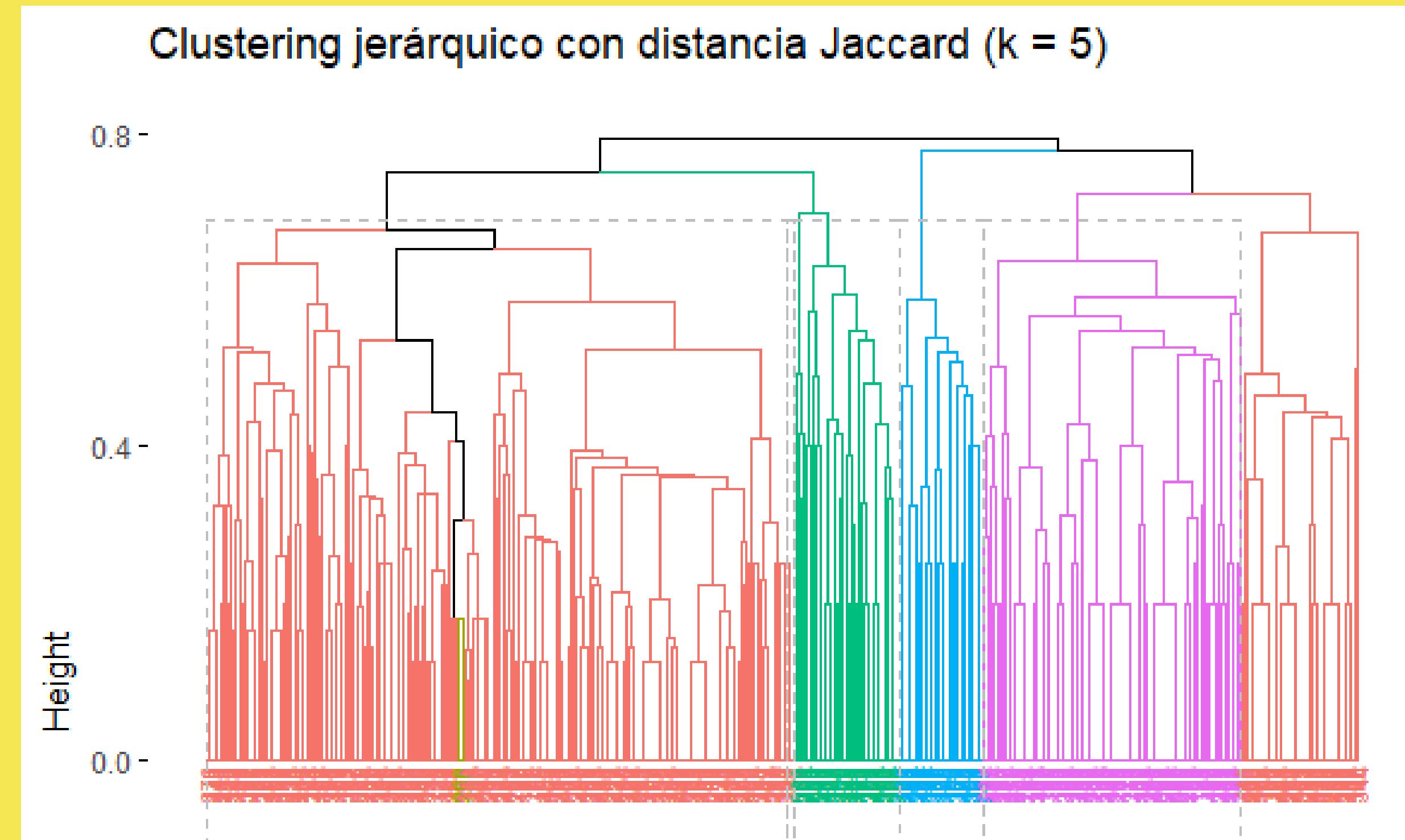
¿Cómo saber el número de clusters? (sirve para los diferentes métodos)

```
gr7
 1 2 3 4 5 6 7
62 193 112 41 51 36 5
> # Cortar el dendrograma en k = 9
> gr6 <- cutree(hc, k = 6)
> # Tamaños de clúster
> table(gr6)
gr6
 1 2 3 4 5 6
255 112 41 51 36 5
> # Cortar el dendrograma en k = 9
> gr6 <- cutree(hc, k = 6)
> # Tamaños de clúster
> table(gr5)
gr5
 1 2 3 4 5
255 112 46 51 36
> # Cortar el dendrograma en k = 5
> gr5 <- cutree(hc, k = 5)
> # Tamaños de clúster
> table(gr5)
gr5
 1 2 3 4 5
255 112 46 51 36
> |
```

despues de probar con 9, hasta 5, nos quedamos con 5, porque los $k > 5$ arrojan algunos cluster muy pequeños

Haremos un dendrograma para visualizar la conformación de los 5 cluster

EJEMPLO DE MÉTODO AGLOMERATIVO CON VARIABLES CATEGÓRICAS



EJEMPLO DE MÉTODO AGLOMERATIVO CON VARIABLES CATEGÓRICAS

Podemos visualizar estas cluster de otra manera?

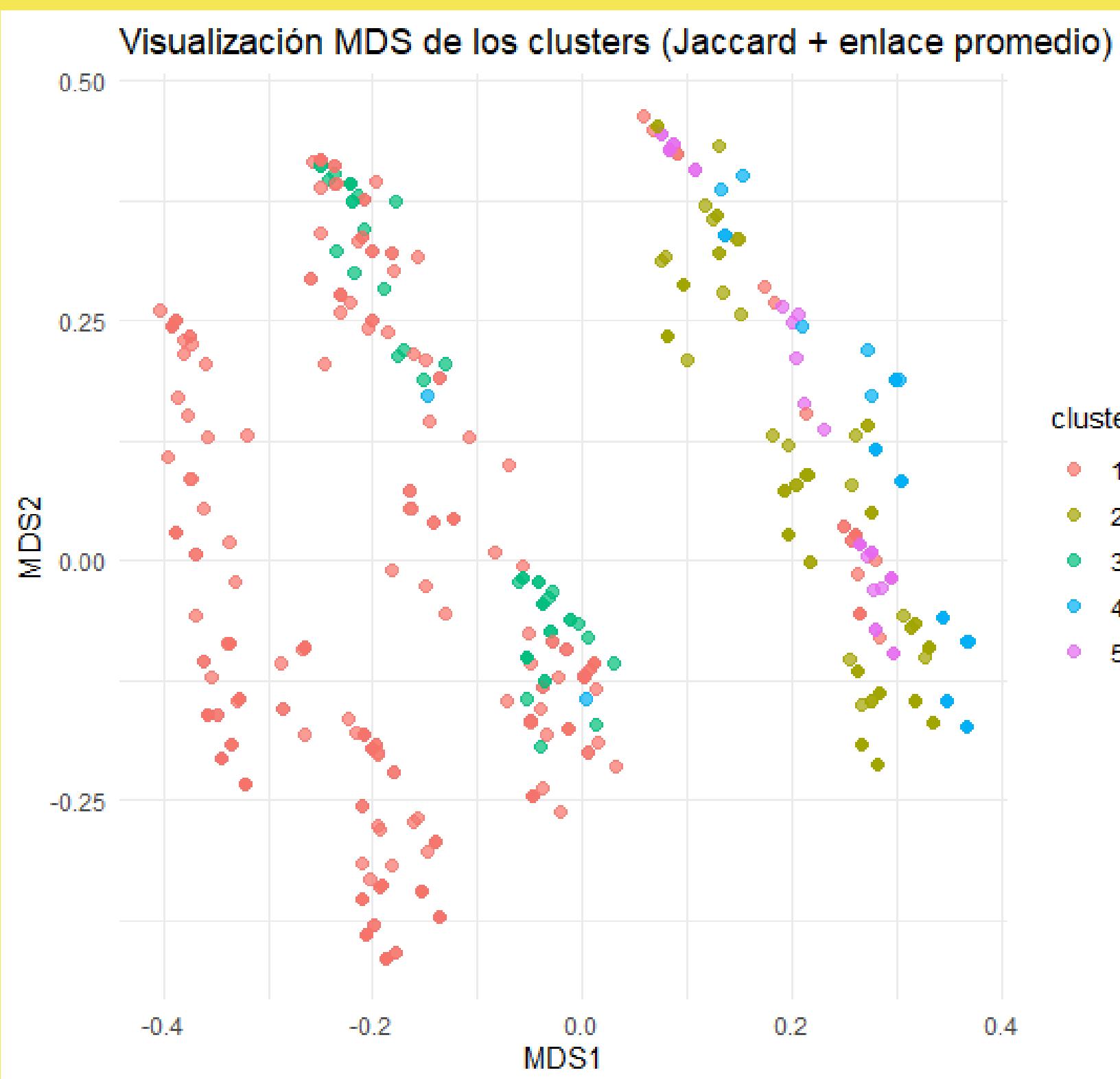
Cómo caracterizar cada cluster?

- Dismunir la dimensiónalidad, con una método que identifique dos variables que condensen la mayor información. Para este tipo de casos de usa.

Escalamiento multidimensional (no lo hemos visto en el curso).

permite representar visualmente relaciones de similitud o disimilitud entre un conjunto de objetos (ya sean distancias métricas o no métricas)

Ejemplo de método aglomerativo con variables categóricas



Usando el método de “escalamiento multidimensional” se pueden visualizar los tres conglomerados

```
mds <- cmdscale(dist_jaccard, k = 2)
df_plot <- data.frame(MDS1 = mds[, 1],
                      MDS2 = mds[, 2],
                      cluster =
factor(cutree(hc, k = 3)))
library(ggplot2)
ggplot(df_plot, aes(MDS1, MDS2, color =
cluster)) +
  geom_point(size = 2, alpha = 0.7) +
  labs(title = "Visualización MDS de los
clusters (Jaccard + enlace promedio)") +
  theme_minimal()
```

Ejemplo de método aglomerativo con variables categóricas. Caracterizar a partir de los valores de las variables en cada cluster.

```
> # Proporciones por clúster para una variable categórica (ej.: educación)  
> prop.table(table(datos_cluster$cluster, datos_cluster$educacion), 1)
```

	Básico	Medio	Técnico	Universitario	Postgrado
1	0.01568627	0.12156863	0.23529412	0.08235294	0.54509804
2	0.00000000	0.00000000	0.00000000	1.00000000	0.00000000
3	0.13043478	0.34782609	0.52173913	0.00000000	0.00000000
4	0.00000000	0.03921569	0.00000000	0.96078431	0.00000000
5	0.00000000	0.00000000	0.00000000	1.00000000	0.00000000

Caraceterizar según variable educación
(5 niveles)

El Clúster 1 representa a personas de educación más alta (más posgrados que el promedio).
El Clúster 3 reúne a quienes tienen niveles medios o bajos (básico, medio, técnico).
Los Clústeres 2, 4 y 5 son grupos casi exclusivamente universitarios, aunque pueden diferenciarse por otras variables (edad, ingresos, ideología, etc.).

Ejemplo de método aglomerativo con variables categóricas

```
> tabs_cat$ingresos
```

cluster	Decil 1	Decil 2	Decil 3	Decil 4	Decil 5	Decil 6	Decil 7	Decil 8	Decil 9	Decil 10
1	0.4%	0.4%	5.1%	17.3%	20.4%	14.1%	9.4%	11.0%	11.8%	10.2%
2	3.6%	1.8%	5.4%	19.6%	17.9%	17.9%	16.1%	6.2%	3.6%	8.0%
3	0.0%	0.0%	8.7%	8.7%	15.2%	6.5%	15.2%	13.0%	21.7%	10.9%
4	0.0%	2.0%	3.9%	21.6%	29.4%	23.5%	9.8%	7.8%	2.0%	0.0%
5	0.0%	2.8%	5.6%	8.3%	8.3%	13.9%	8.3%	16.7%	19.4%	16.7%
Total	1.0%	1.0%	5.4%	16.8%	19.4%	15.2%	11.4%	10.2%	10.4%	9.2%

Caracterizar según variable ingresos(10 niveles)

- Clúster 1, 2 y 4: Perfil medio
- Clúster 3 y 5: Perfil medio-alto

Ejemplo de método aglomerativo con variables categóricas

```
> tabs_cat$edad
```

cluster	15-24	25-39	40-54	55-69	70-99
1	13.7%	30.2% 22.7%		8.2%	25.1%
2	0.9%	55.4% 42.0%		0.0%	1.8%
3	0.0%	4.3%	0.0%	95.7%	0.0%
4	98.0%	0.0%	0.0%	0.0%	2.0%
5	0.0%	0.0%	0.0%	100.0%	0.0%
Total	17.2%	28.2%	21.0%	20.2%	13.4%

Caracterizar según variable edad (5 niveles)

- Clúster 1 y 2: Perfil edad adulto-joven/adulto
- Clúster 3 y 5: Perfil adulto entre 55 y 69 años
- Clúster 4: Perfil jóvenes

Ejemplo de método aglomerativo con variables categóricas

```
> tabs_cat$eje_politico
```

cluster	Izquierda	Centro	Derecha
1	27.5%	47.5%	25.1%
2	23.2%	53.6%	23.2%
3	8.7%	58.7%	32.6%
4	45.1%	47.1%	7.8%
5	16.7%	55.6%	27.8%
Total	25.8%	50.4%	23.8%

Caracterizar según variable
posicionamiento político (3
niveles)

- Clúster 1, 2, 3 y 5: Perfil personas de centro
- Clúster 4: Perfil personas de centro izquierda

Ejemplo de método aglomerativo con variables categóricas. Lectura conjunta

Clúster 1:

educación más alta

ingresos medios

adulto-joven/adulto

centro político

Clúster 2:

casi exclusivamente universitarios

ingresos medios

adulto-joven/adulto

centro político

Clúster 3

niveles educacionales medios o bajos

ingresos medio-alto

adulto entre 55 y 69 años

centro político

Clúster 4

casi exclusivamente universitarios

ingresos medios

jóvenes

centro izquierda

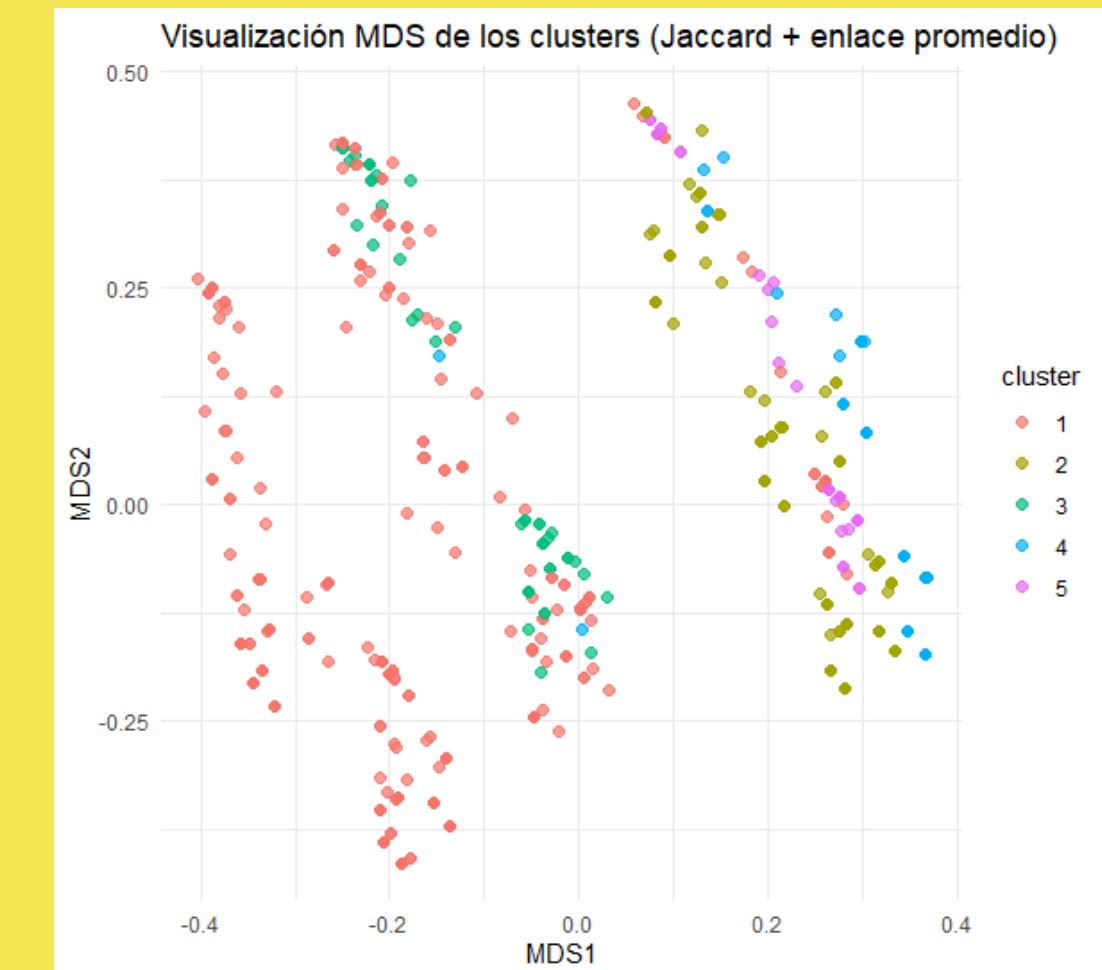
Clúster 5:

casi exclusivamente universitarios

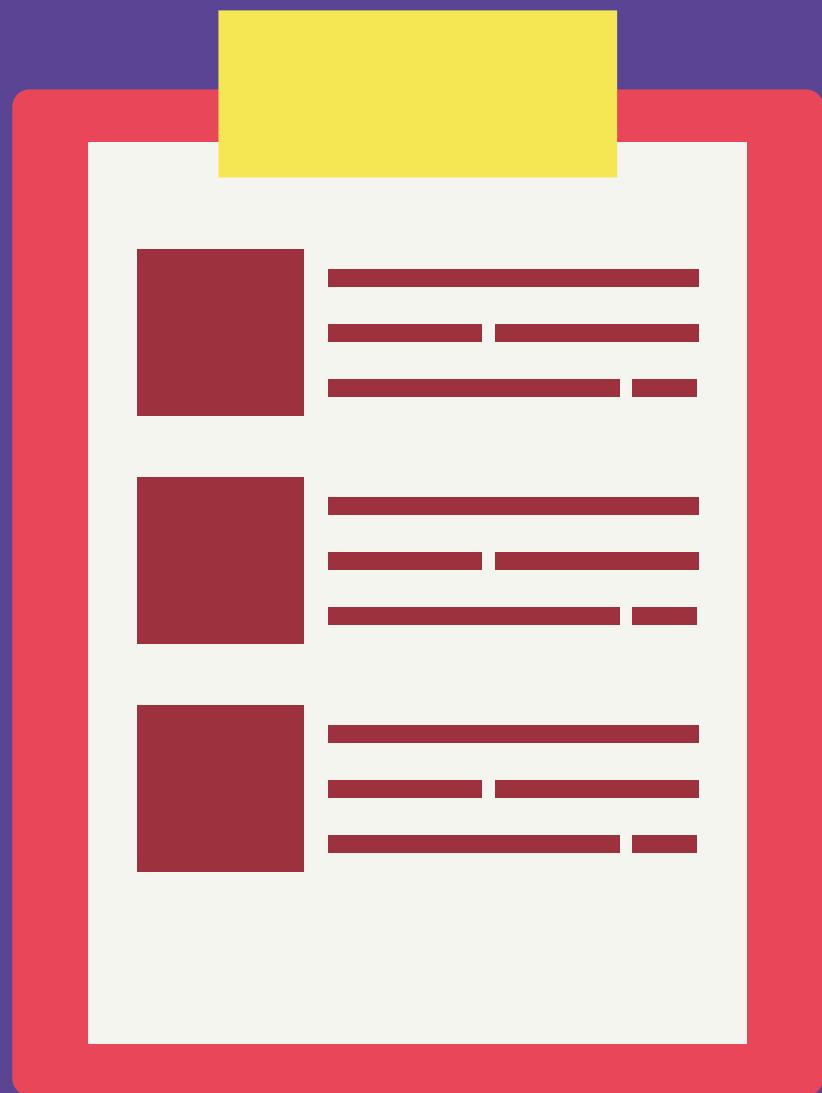
ingresos medio-alto

adulto entre 55 y 69 años

centro político



DOS TIPOS DE LÓGICAS DE CONGLOMERACIÓN



Métodos no
jerárquico

MÉTODOS NO JERÁRQUICOS O DE OPTIMIZACIÓN



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

MÉTODOS NO JERARQUICOS - OPTIMIZACIÓN

METODO DE REASIGNACIÓN (O PARTICIÓN ITERATIVOS)

Permiten que los objetos asignados a un conglomerado en una fase del proceso sean reasignados a otro conglomerado en otra fase posterior.

La "reasignación" debe "optimizar" el criterio de selección (a un conglomerado al que estén más cerca de su centroide)

Algunos de los algoritmos más conocidos dentro de estos métodos son:

K-means clustering (MacQueen, 1967), in which, each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to anomalous data points and outliers. • K-medoids clustering or PAM (Partitioning Around Medoids, Kaufman & Rousseeuw, 1990), in which, each cluster is represented by one of the objects in the cluster. PAM is less sensitive to outliers compared to k-means. • CLARA algorithm (Clustering Large Applications), which is an extension to PAM adapted for large data sets

- **Método "K-means"** de McQueen.
- Métodos de centroide ("quick cluster analysis" y método de "Forgy")
- Método de nubes dinámicas de Diday.
- K-medoids clustering
- PAM (Partitioning Around Medoids)
- CLARA algorithm (Clustering Large Applications), extension de PAM
- Fuzzy clustering
- Density-based clustering
- Model-based clustering



MÉTODOS NO JERARQUIOS - OPTIMIZACIÓN

Método de k - means

Algoritmo más característico y de mayor aplicación en los métodos de conglomeración no jerárquicos.

Diseñado por McQueen en 1967

Adecuado para variables continuas.

Se puede usar para variables ordinales, pero se debe asumir que la distancia entre las categorías es contante.

Existen varios tipos de algoritmos, siendo el here are several k-means algorithms available. The standard algorithm is the Hartigan-Wong algorithm (1979) el más común:

Define el total de la variación interna de los cluster como la suma de cuadrados de distancias Euclídeas entre los ítems y el centroide correspondiente. Se pueden usar otras medidas.



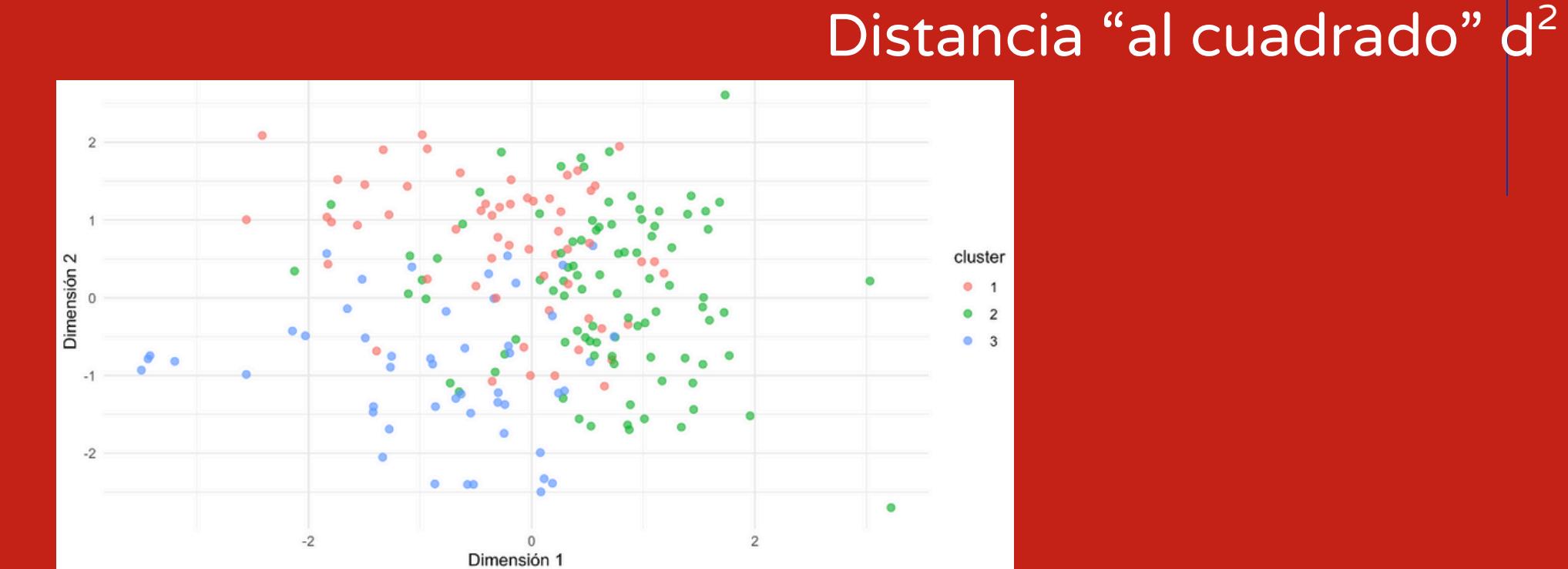
MÉTODOS NO JERARQUIOS - OPTIMIZACIÓN

Métdo de k - means



4 pasos

1. Especificar (a priori) el número de conglomerados que deben formarse con los datos (se puede obtener por criterio teórico o por análisis jerárquico previo).
 - 1.1 En general se recomienda estandarizar las variables.
2. Se calculan los centroides iniciales de los conglomerados. En caso de no disponer de esta información previa (cuando no se parte, por ejemplo, de conglomerados ya constituidos mediante algún procedimiento jerárquico u otro algoritmo de clasificación), el programa informático que se use para su realización los estima iterativamente, utilizando los valores de los "K" primeros casos en el fichero de datos como estimaciones "provisionales" de los centroides (de las "K-medias" de los conglomerados ("K" = el número de conglomerados especificado por el investigador).

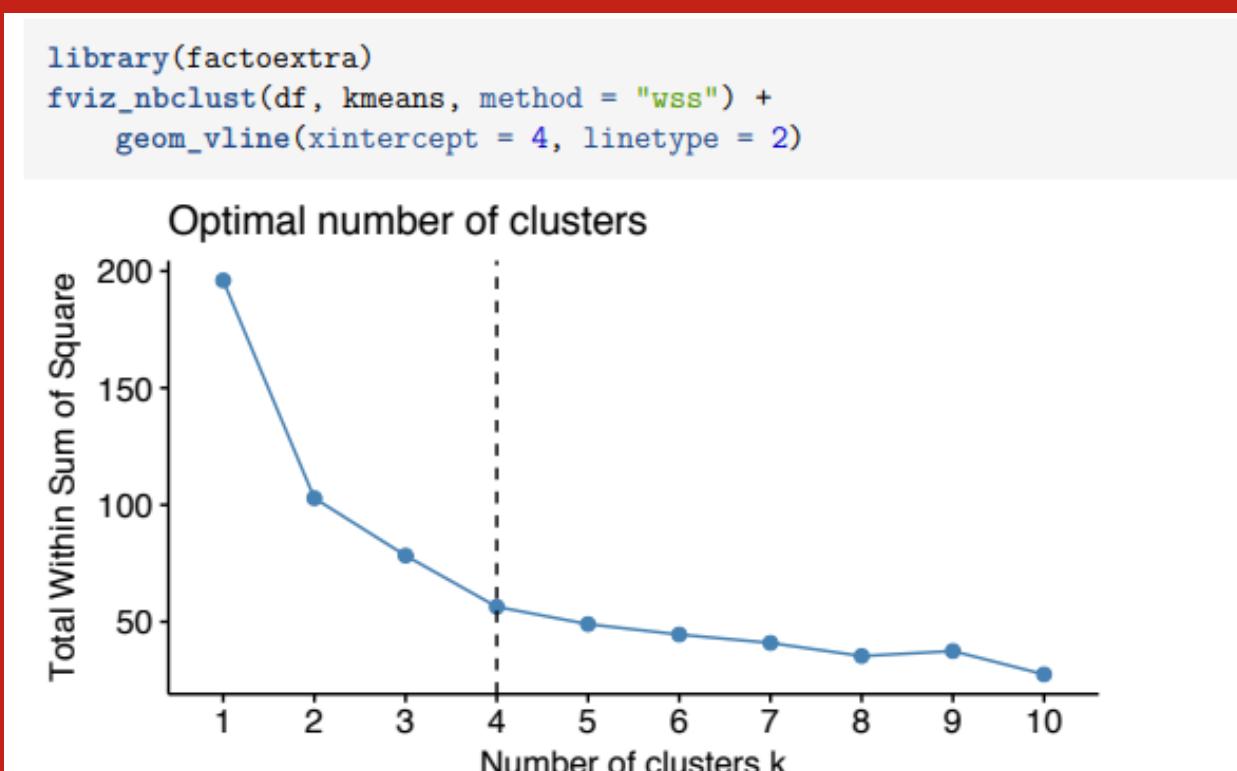


SELECCIÓN DE NÚMERO DE CONGLOMERADOS - ALTERNATIVA

Para k-means

Correr el programa con diferentes número de conglomerados y luego aplicar el método de sedimentación (calculando la suma de cuadrados que representa la varianza al interior de los cluster) y buscar el codo (donde ya no se disminute significativamente la varianza interior, sumando más particiones).

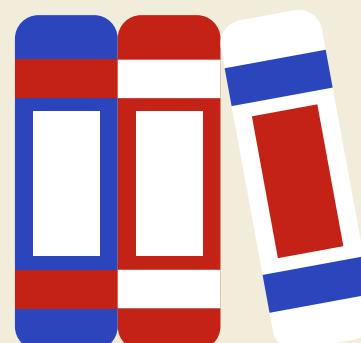
Función fviz_nbclust() de R (incluida en el paquete factoextra)



MÉTODOS NO JERARQUICOS - OPTIMIZACIÓN

Método de k - means

4 pasos



En el paso 1, el algoritmo selecciona k objetos del total de objetos de la base de datos, los que toma como “centros” (centroides).

Paso 2. Para cada uno de los objetos restantes se asigna a su centroide más cercano, donde “más cercano” se define utilizando la distancia euclídea entre el objeto y la media del grupo (cluster): “etapa de asignación de conglomerados” (cluster assignment step).

Paso 3. El algoritmo calcula el nuevo valor medio de cada conglomerado: “actualización del centroide del conglomerado” (cluster centroid update)

Paso 4: Una vez que los centros han sido recalculados, cada observación se vuelve a comprobar para determinar si podría estar más cerca de un conglomerado diferente.

Proceso iterativo de re-asignación, utilizando las medias actualizadas de los conglomerados.

La medida de distancia más utilizada en este algoritmo es la distancia euclídea (puede ser otra)

MÉTODOS NO JERARQUIOS - OPTIMIZACIÓN

Métdo de k - means

4 pasos



Las etapas de asignación de conglomerados y actualización de centroides se repiten de forma iterativa hasta que las asignaciones de los conglomerados dejan de cambiar (es decir, hasta que se alcanza la convergencia).

En ese punto, los conglomerados formados en la iteración actual son los mismos que los obtenidos en la iteración anterior..

También puede suceder que se haya llegado al número máximo de iteraciones posible.

Al final del proceso iterativo se obtiene los centroides finales. Es factible que éstos no coincidan con los "iniciales", sobre todo cuando se ha producido un número elevado de iteraciones y, en consecuencia, de modificaciones en la composición de los conglomerados.

MÉTODOS NO JERARQUICOS - OPTIMIZACIÓN

Método de k - means

Aplicación R

Use un data set de variables continuas o escala Likert

```
data("base_de_datos") # cargar base de datos  
df <- scale(base_de_datos) # Escalar las variables  
head(df, n = 3) # ver las primeras tres filas
```

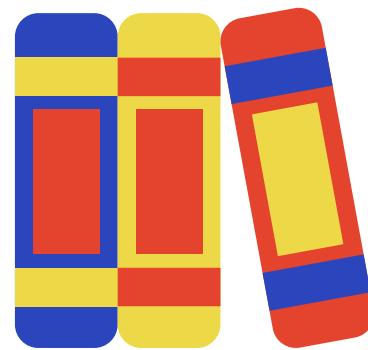
La función estandart para k-means clustering en R es

kmeans() [stats package]

```
kmeans(x, centers, iter.max = 10, nstart = n)
```

```
set.seed(123) # fijar la selección random para futuras reproducciones  
km.res <- kmeans(df, 4, nstart = 25) # realizar k-means con 4 cluster
```

- x: matriz numérica de distancias
- centers: número de clusters (k)
- iter.max: indica el número máximo de iteraciones permitidas. el valor por defecto es 10.
 - nstart: el número de particiones iniciales, cuando “centers” es un numero. Se recomienda 25, que dice que escoja 25 veces puntos iniciales y escoje la solución más estable y óptima
 - Así se reduce el riesgo de caer en un mínimo local (una mala partición).



MÉTODOS NO JERARQUICOS - OPTIMIZACIÓN

Método de k - means

Ejemplo aplicación R

Como el método k-means clustering comienza con una selección random de centroids, se recomienda usar la función `set.seed()` de modo que la próxima vez que se corra el mismo script se seleccionen los mismos casos y el método sea reproducible.

```
print(km.res) # para ver el resultado
```

```
aggregate(base_de_datos, by=list(cluster=km.res$cluster), mean) # para  
ver el resultado de cada cluster
```

```
aggregate(USArrests, by=list(cluster=km.res$cluster), mean)
```

```
##   cluster   Murder   Assault UrbanPop      Rape
## 1       1  3.60000  78.53846 52.07692 12.17692
## 2       2  5.65625 138.87500 73.87500 18.78125
## 3       3 10.81538 257.38462 76.00000 33.19231
## 4       4 13.93750 243.62500 53.75000 21.41250
```

If you want to add the point classifications to the original data, use this:



MÉTODOS NO JERARQUICOS - OPTIMIZACIÓN

Método de k - means

Ejemplo aplicación R

Como el método k-means clustering comienza con una selección random de centroids, se recomienda usar la función `set.seed()` de modo que la próxima vez que se corra el mismo script se seleccionen los mismos casos y el método sea reproducible.

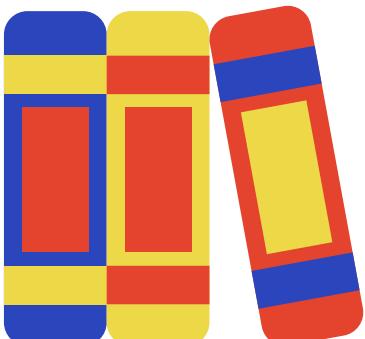
```
print(km.res) # para ver el resultado
```

```
aggregate(base_de_datos, by=list(cluster=km.res$cluster), mean) # para  
ver el resultado de cada cluster
```

```
aggregate(USArrests, by=list(cluster=km.res$cluster), mean)
```

```
##   cluster   Murder   Assault UrbanPop      Rape
## 1       1  3.60000  78.53846 52.07692 12.17692
## 2       2  5.65625 138.87500 73.87500 18.78125
## 3       3 10.81538 257.38462 76.00000 33.19231
## 4       4 13.93750 243.62500 53.75000 21.41250
```

If you want to add the point classifications to the original data, use this:



MÉTODOS NO JERARQUICOS - OPTIMIZACIÓN

Métdo de k - means

Ejemplo aplicación R

Para generar gráficos se recomienda

factoextra package y hacer un análisis de componentes principales, de modo de graficar los puntos con colores en ese plano.

Con la función fviz_cluster() del paquete factoextra, si tus datos (data = df) tienen más de dos variables, la función automáticamente aplica un Análisis de Componentes Principales (PCA) internamente para poder mostrar el gráfico en dos dimensiones (2D).

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

fviz_cluster(km.res, data = df,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800",
                         "#FC4E07"),
             ellipse.type = "euclid", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
           )
```



Ejemplo con base latinobarómetro k-means, distancia euclídea al cuadrado

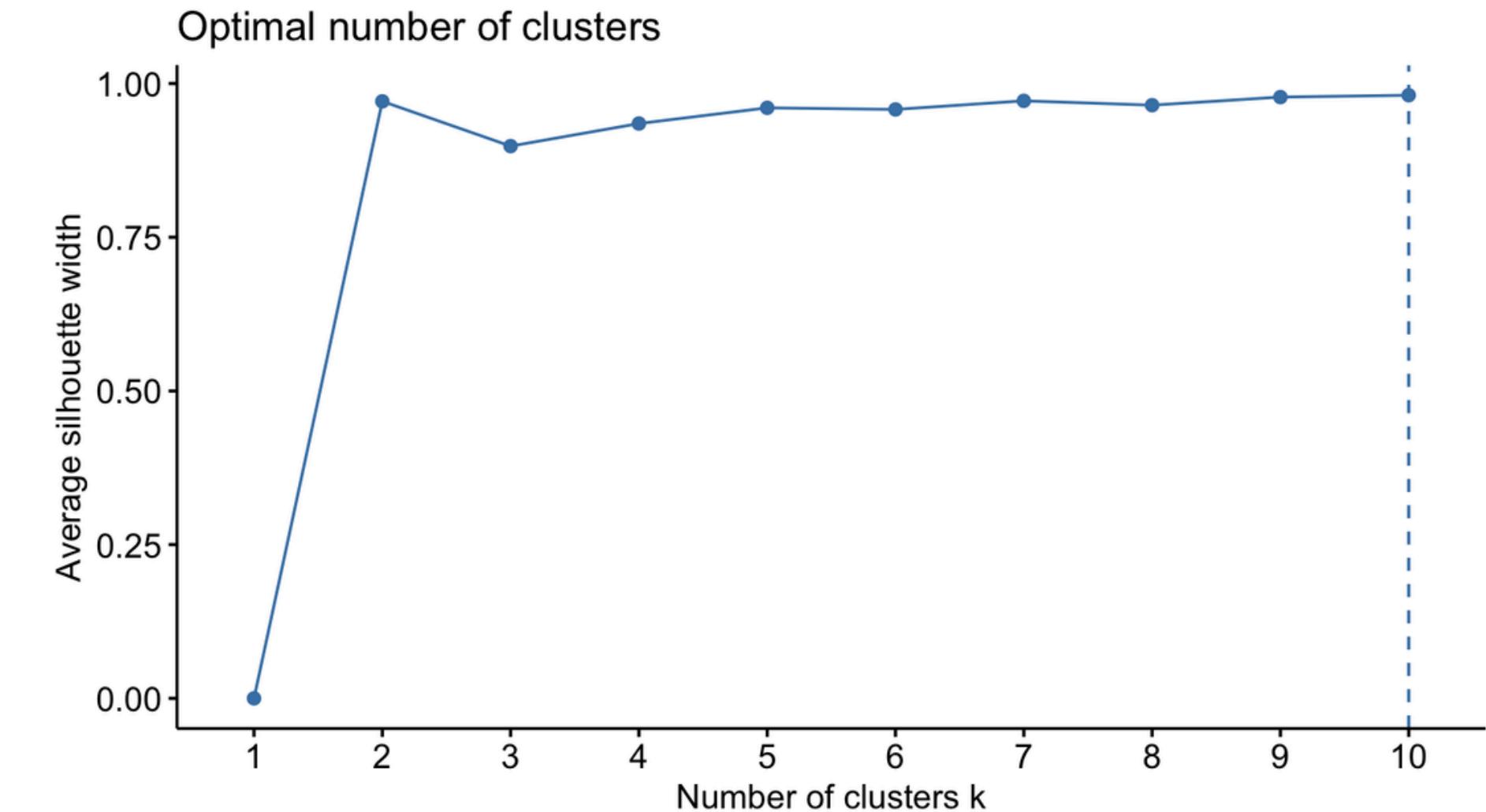
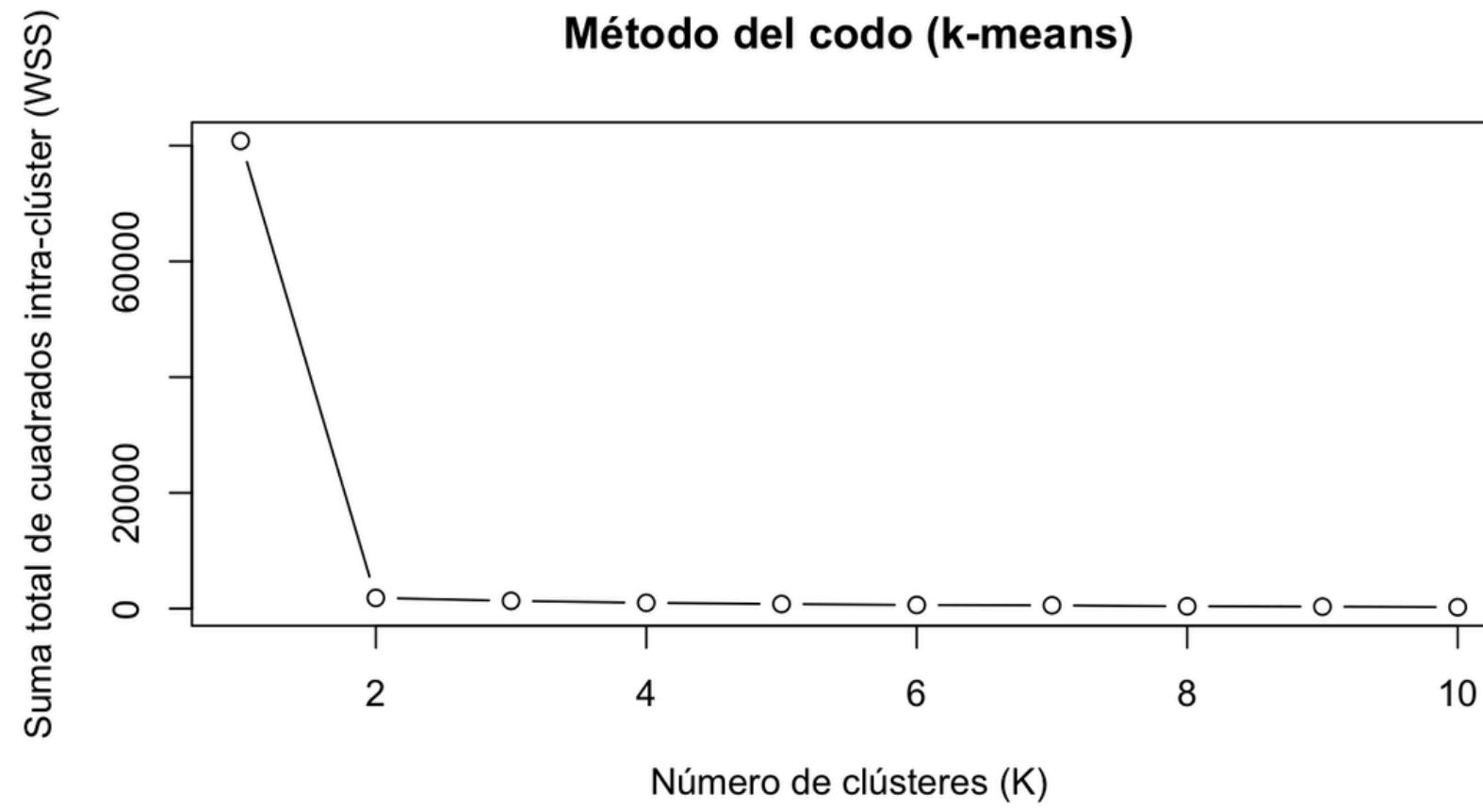
4 variables latinobarómetro
Escala Likert

```
# "P34NI.BB" = "Confianza en los empresarios",
# "P34NI.BE" = "Confianza en los partidos políticos",
# "P34NI.BD" = "Confianza en el Congreso Nacional",
# "P34NI.BC" = "Confianza en el Poder Judicial"
```

Ejemplo con base latinobarómetro k-means, distancia euclídea al cuadrado

determinar n de cluster

Tests me indica $k = 2$



Ejemplo con el script clase anterior

4 variables latinobarómetro

```
> print(resumen_km)
# A tibble: 2 × 5
  Cluster `Confianza en los empresarios` `Confianza en los partidos políticos` `Confianza en el Congreso Nacional` `Confianza en el Poder Judicial`
  <dbl> <dbl> <dbl> <dbl> <dbl>
1     1   -1.98  -1.97  -1.97  -1.96
2     2    1.97   1.99   1.96   1.95
```



```
> # Tamaños de clúster
> table(clusters_km)
clusters_km
  1    2
  2377 17827
```

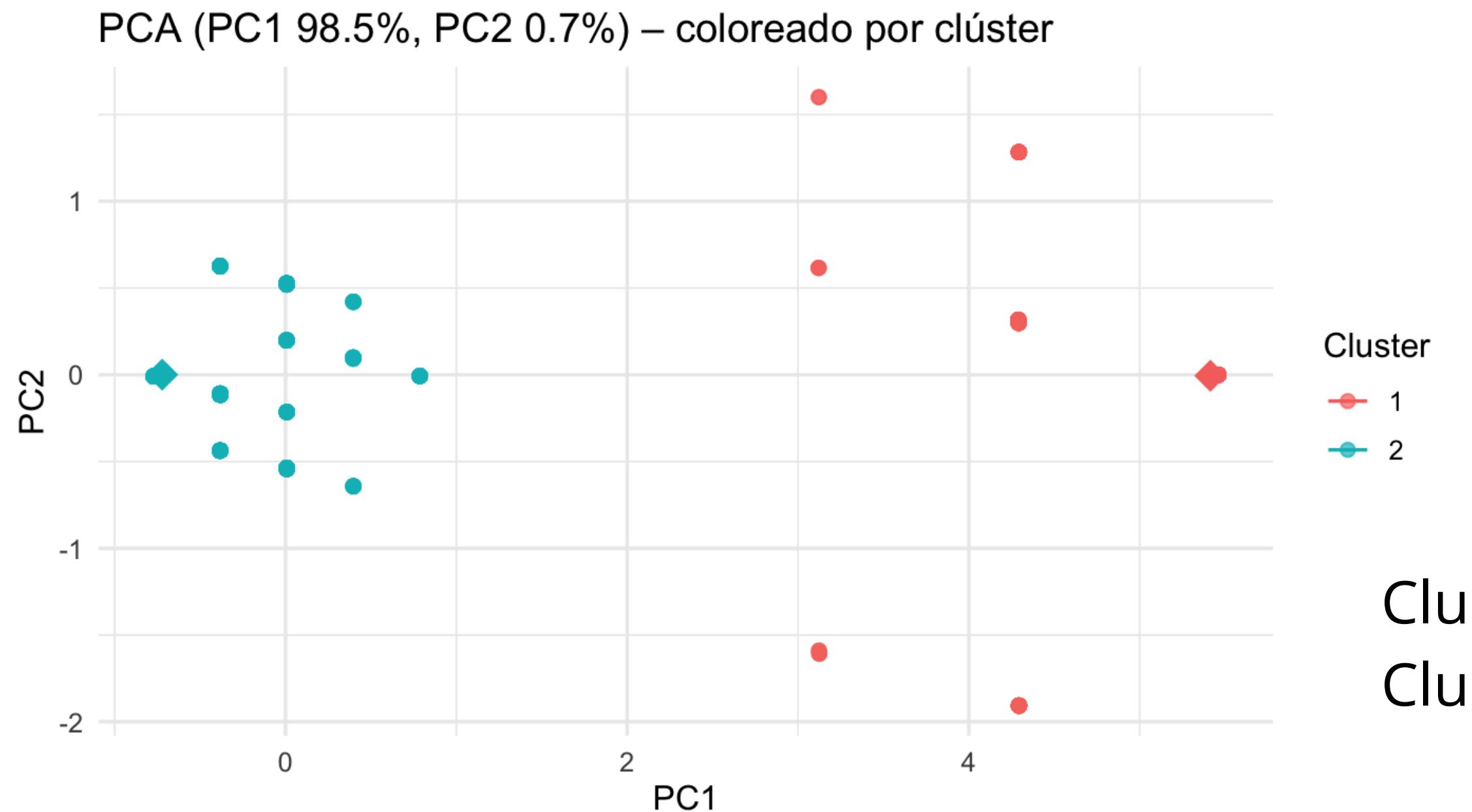
Se observan ambos grupos con claridad:

Quienes tienen confianza y quienes no tienen confianza. Con esto se podría construir una variable cluster, por ejemplo

Ejemplo con el script clase anterior

4 variables latinobarómetro

Se observan ambos grupos con claridad. Se pueden visualizar en un plano de 2 coordenadas mediante Componentes Principales



Cluster 1: no tienen confianza
Cluster 2: tienen confianza

EJEMPLO

ALDAS Y URIELA



Diseño de un plan de incentivos
para vendedores

EJEMPLO

Aldas y urielas



Diseño de un plan de incentivos para vendedores

El director de ventas de una cadena de tiendas de electrodomésticos con implantación nacional está estudiando el plan de incentivos de sus vendedores.



Considera que **los incentivos deben estar ajustados a las dificultades de las distintas zonas de ventas**, siendo necesario fijar incentivos más altos en aquellas zonas geográficas en que las condiciones de vida de sus habitantes hacen más difícil las ventas. Por este motivo quiere **determinar si las comunidades autónomas se pueden segmentar en grupos homogéneos respecto al equipamiento de los hogares**.

Datos cuadro 3.22

Objetivo: **establecer cuántos grupos de comunidades autónomas con niveles de equipamiento similar pueden establecerse y en qué radican las diferencias entre esos grupos.**

EJEMPLO

Diseño de un plan de incentivos para vendedores

Aldas y urielia



Objetivo: establecer cuántos grupos de comunidades autónomas con niveles de equipamiento similar pueden establecerse y en qué radican las diferencias entre esos grupos.



Cuadro 3.22.: Equipamiento de los hogares en distintas comunidades autónomas

CC.AA.	Porcentaje de hogares que poseen					
	Auto-móvil	TV color	Vídeo	Micro-ondas	Lava-vajillas	Teléfono
España	69,0	97,6	62,4	32,3	17,0	85,2
Andalucía	66,7	98,0	82,7	24,1	12,7	74,7
Aragón	67,2	97,5	56,8	43,4	20,6	88,4
Asturias	63,7	95,2	52,1	24,4	13,3	88,1
Baleares	71,9	98,8	62,4	29,8	10,1	87,9
Canarias	72,7	96,8	68,4	27,9	5,80	75,4
Cantabria	63,4	94,9	48,9	36,5	11,2	80,5
Castilla y León	65,8	97,1	47,7	28,1	14,0	85,0
Cast.-La Mancha	61,5	97,3	53,6	21,7	7,10	72,9
Cataluña	70,4	98,1	71,1	36,8	19,8	92,2
Com. Valenciana	72,7	98,4	68,2	26,6	12,1	84,4
Extremadura	60,5	97,7	43,7	20,7	11,7	67,1
Galicia	65,5	91,3	42,7	13,5	14,6	85,9
Madrid	74,0	99,4	76,3	53,9	32,3	95,7
Murcia	69,0	98,7	59,3	19,5	12,1	81,4
Navarra	76,4	99,3	60,6	44,0	20,6	87,4
País Vasco	71,3	98,3	61,6	45,7	23,7	94,3
La Rioja	64,9	98,6	54,4	44,4	17,6	83,4

Fuente: Panel de Hogares de la Unión Europea. INE.

EJEMPLO

Aldas y uriel



Diseño de un plan de incentivos para vendedores. Estrategia de análisis



1. Análisis de la existencia de *outliers* en la medida en que pueden generar importantes distorsiones en la detección del número de grupos.
2. Realización de un análisis de conglomerados jerárquicos, evaluando la solución de distintos métodos de conglomeración, aplicando los criterios presentados para identificar el número adecuado de grupos y obtención de los centroides que han de servir de partida para el paso siguiente.
3. Realización de un análisis de conglomerados no jerárquico mediante el método de k -medias para la obtención de una solución óptima en términos de homogeneidad intrasegmentos y heterogeneidad intersegmentos.

EJEMPLO

Diseño de un plan de incentivos para vendedores.

Aldas y uriel

-
1. Análisis de la existencia de *outliers* en la medida en que pueden generar importantes distorsiones en la detección del número de grupos.

Cuadro 3.23.: Resultados de la detección de *outliers*

CC.AA.	D^2	p-value $\chi^2(df = 6)$
España	0,40	0,99
Andalucía	3,93	0,68
Aragón	1,94	0,92
Asturias	4,46	9,61
Baleares	6,02	0,42
Canarias	10,47	0,10
Cantabria	7,27	0,29
Castilla y León	3,25	0,77
Castilla-La Mancha	4,12	0,66
Cataluña	4,21	0,64
Com. Valenciana	2,85	0,82
Extremadura	0,29	0,15
Galicia	13,30	0,03
Madrid	9,49	0,14
Murcia	4,61	0,59
Navarra	9,58	0,14
País Vasco	2,55	0,86
La Rioja	4,25	0,64

Detectar los outliers mediante la distancia de Mahalanobis.

A la luz de esta información no cabría considerar a ninguna comunidad autónoma como un valor atípico.

(se comparan los valores con la distribución y valores críticos de chi-cuadrado.

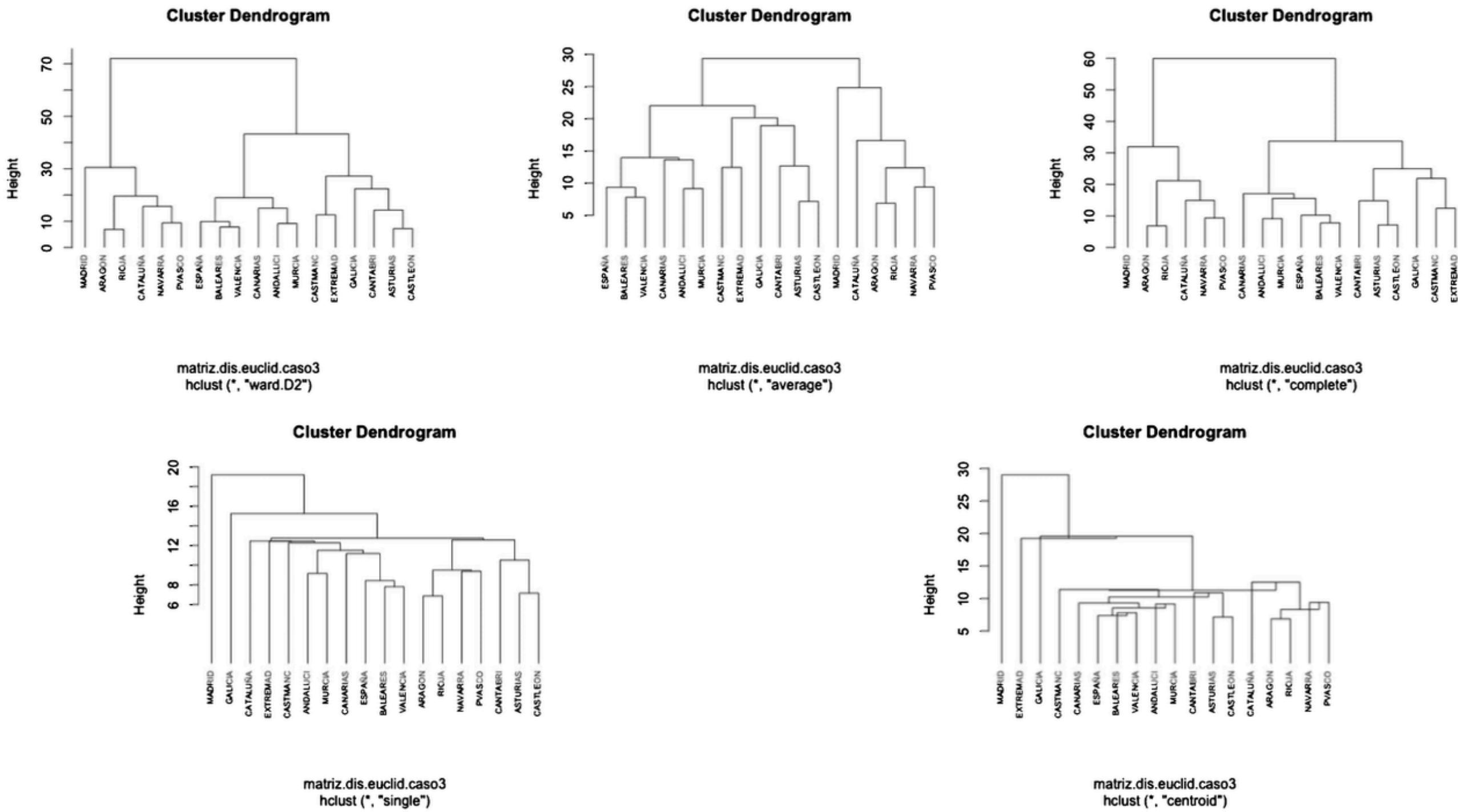


EJEMPLO

Diseño de un plan de incentivos para vendedores.

Aldas y uriel

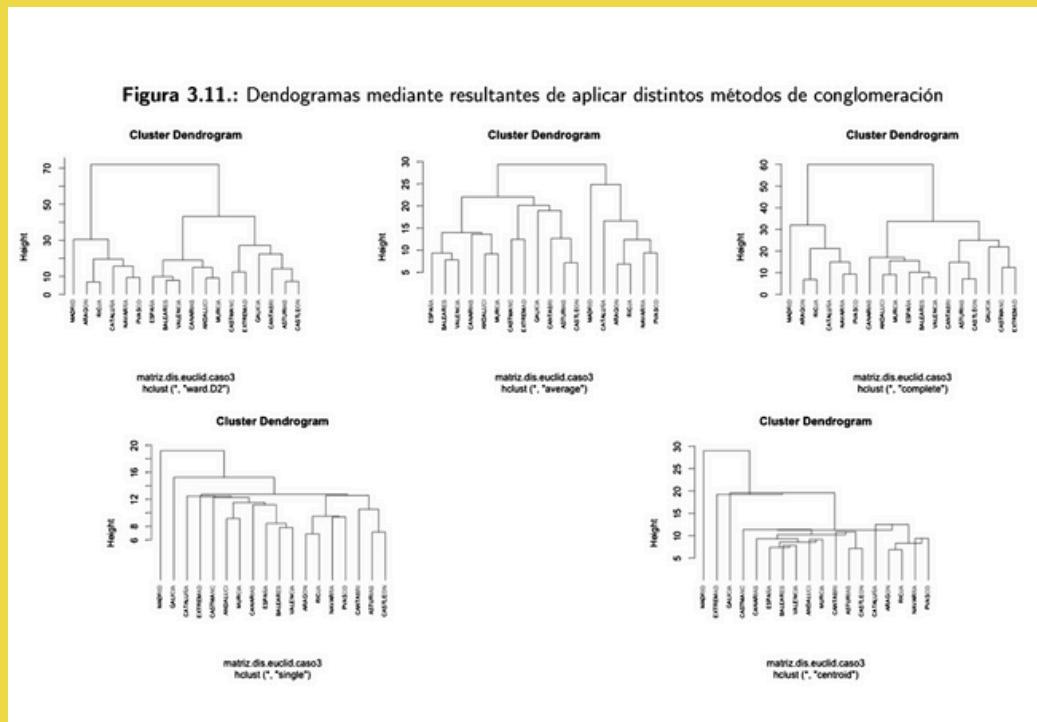
Figura 3.11.: Dendogramas mediante resultantes de aplicar distintos métodos de conglomeración



EJEMPLO

Diseño de un plan de incentivos para vendedores.

Aldas y urielia



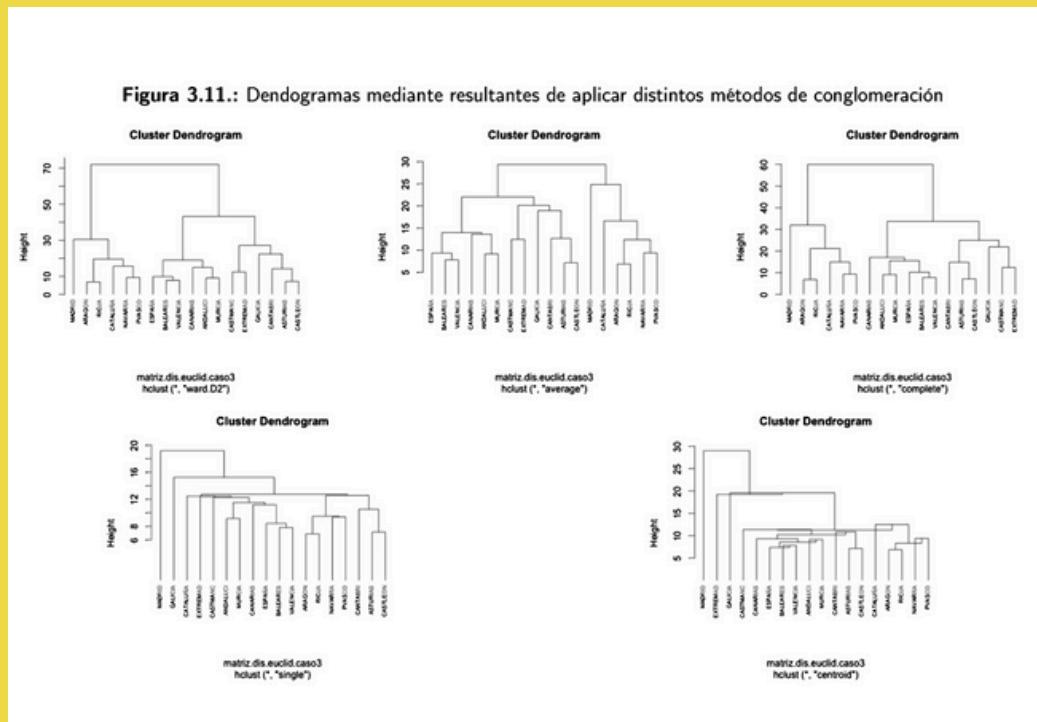
Dos patrones: el correspondiente a los métodos centroid y single, que agregan a todas las comunidades -salvo a Madrid— en un mismo grupo pero con claras distorsiones en el dendograma,

Los otros tres procedimientos –Ward, complete y average— que generan una solución muy parecida. Hemos de establecer cuántos grupos, pero los dendogramas nos permiten intuir que las comunidades que acabarán en cada grupo van a ser las mismas independientemente del método de conglomeración empleado.

EJEMPLO

Diseño de un plan de incentivos para vendedores.

Aldas y uriel

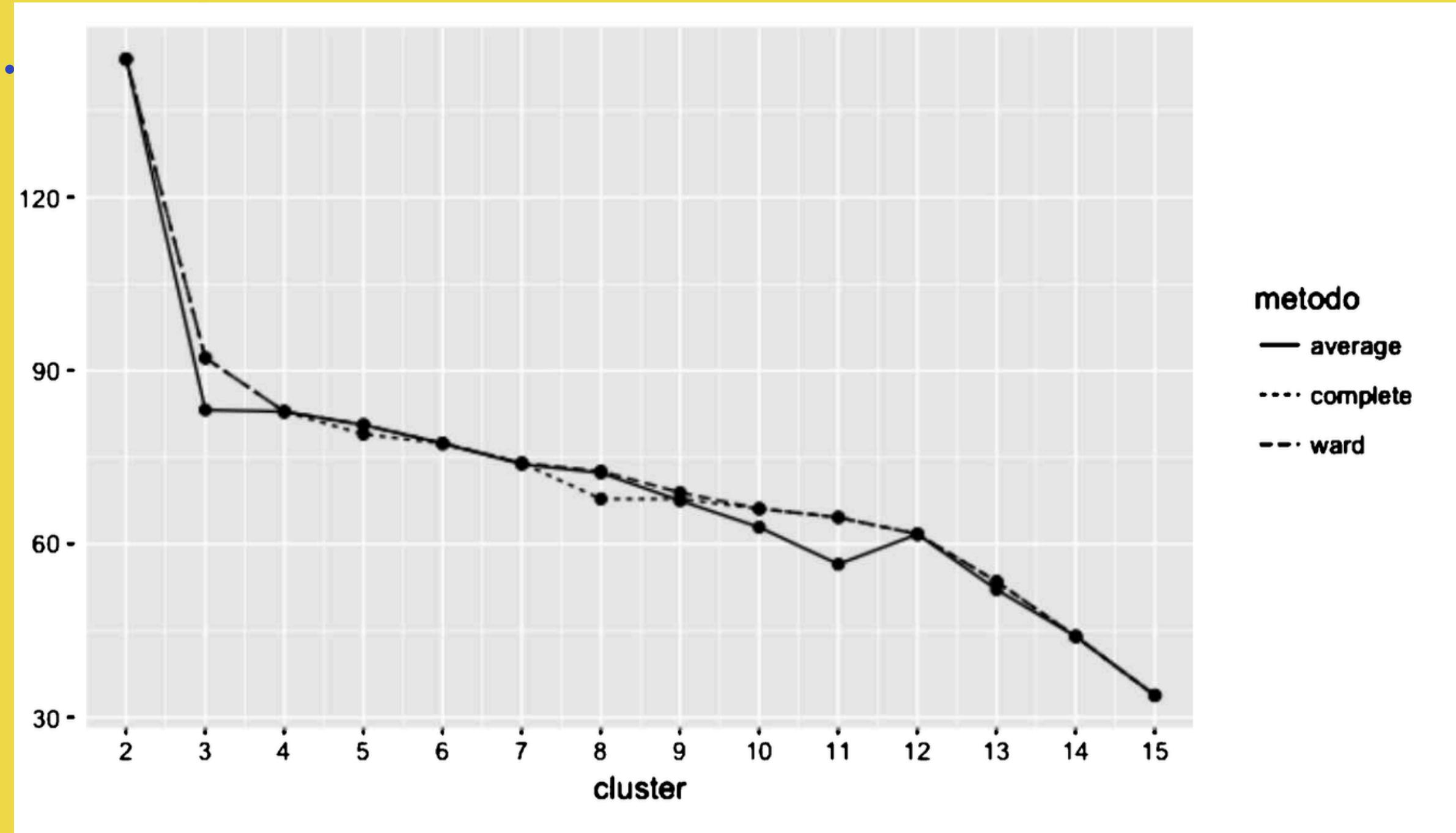


Siguiente fase -determinación del número de conglomerados— solo para las tres técnicas que no nos generan dudas respecto a la claridad de los dendogramas: Ward, complete y average.

EJEMPLO

Diseño de un plan de incentivos para vendedores.

Aldas y uriel



Determinación del número de conglomerados— método del codo:
k=2

EJEMPLO

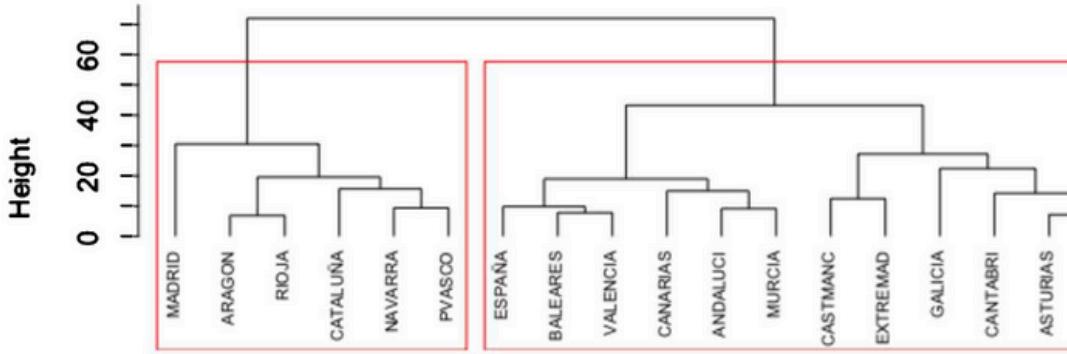
Diseño de un plan de incentivos para vendedores.

Aldas y urielia



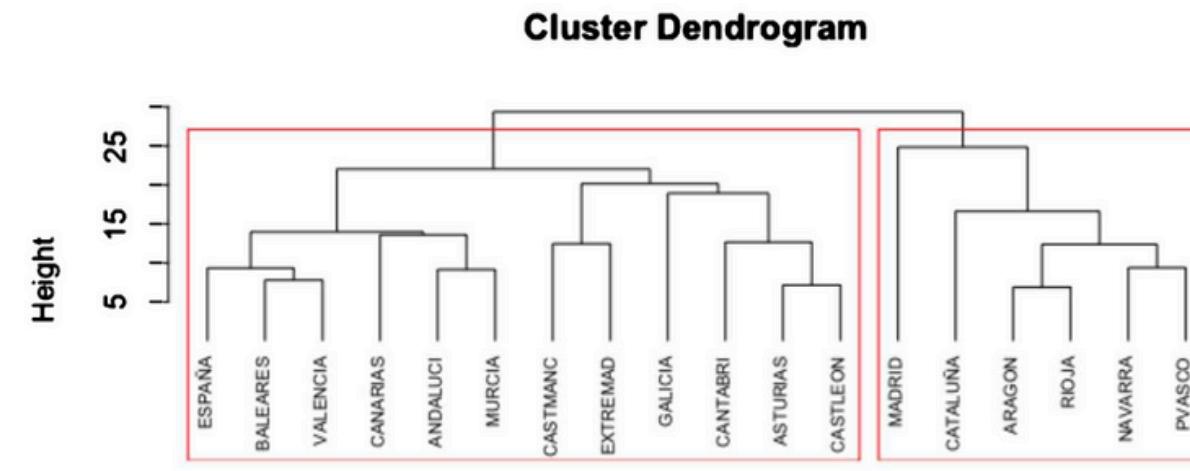
Figura 3.13.: Dendogramas mediante resultantes de aplicar distintos métodos de conglomeración

Cluster Dendrogram



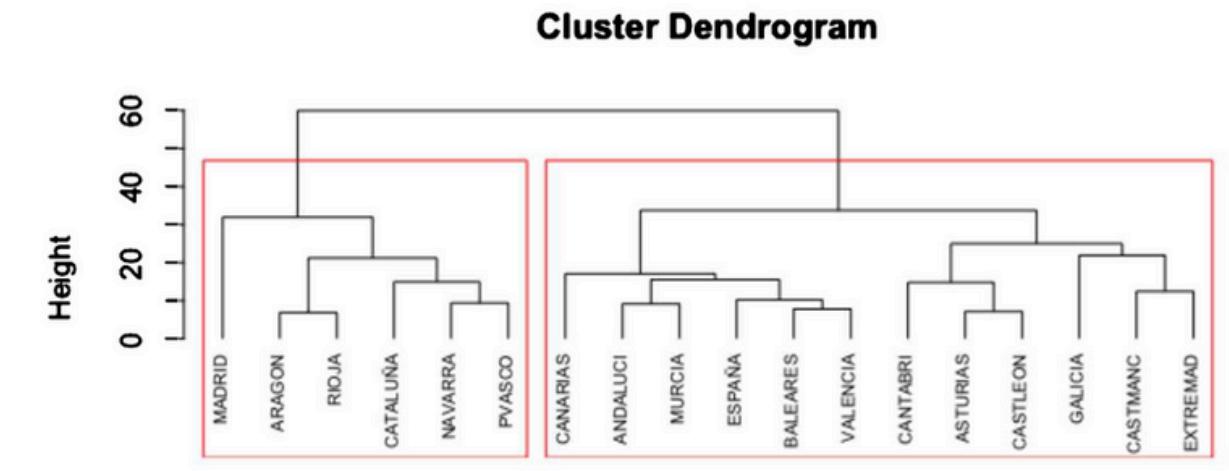
matriz.dis.euclid.caso3
hclust (*, "ward.D2")

Cluster Dendrogram



matriz.dis.euclid.caso3
hclust (*, "average")

Cluster Dendrogram



Determinación del número de conglomerados— método del codo:
 $k=2$

EJEMPLO

Aldas y urielas

Diseño de un plan de incentivos para vendedores.

A continuación obtenemos los centroides:

la media de las seis variables analizadas en cada uno de los dos grupos obtenidos.

Cuadro 3.25.: Centroides resultantes del método no jerárquico K-means clustering with 2 clusters of sizes 12, 6

Cluster means:

	automovi	tvcolor	video	microond	lavavaji	telefono
1	66.86667	96.81667	56.00833	25.425	11.80833	80.70833
2	70.70000	98.53333	63.46667	44.700	22.43333	90.23333

Clustering vector:

```
[1] 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 2 2 2
```

Within cluster sum of squares by cluster:

```
[1] 2176.3133 848.3533  
(between_SS / total_SS = 46.2 %)
```

EJEMPLO

Diseño de un plan de incentivos para vendedores.

Aldas y uriel

.....

Se puede hacer una prueba de test de students (o ANOVA si tuvieramos más de dos conglomerados, para saber si son estadísticamente significativas las diferencias

Cuadro 3.26.: Significatividad de las diferencias entre los perfiles de los conglomerados

Variable	Grupo de CC.AA.		Prueba <i>t</i>
	Grupo 1	Grupo 2	
Automóvil	66,86	70,70	1,81
TV color	96,82	98,53	2,51*
Vídeo	56,01	63,46	1,71
Microondas	25,43	44,70	6,73**
Lavavajillas	11,81	22,43	4,61**
Teléfono	80,71	90,23	3,50**

** $p < 0,01$; * $p < 0,05$

EJEMPLO

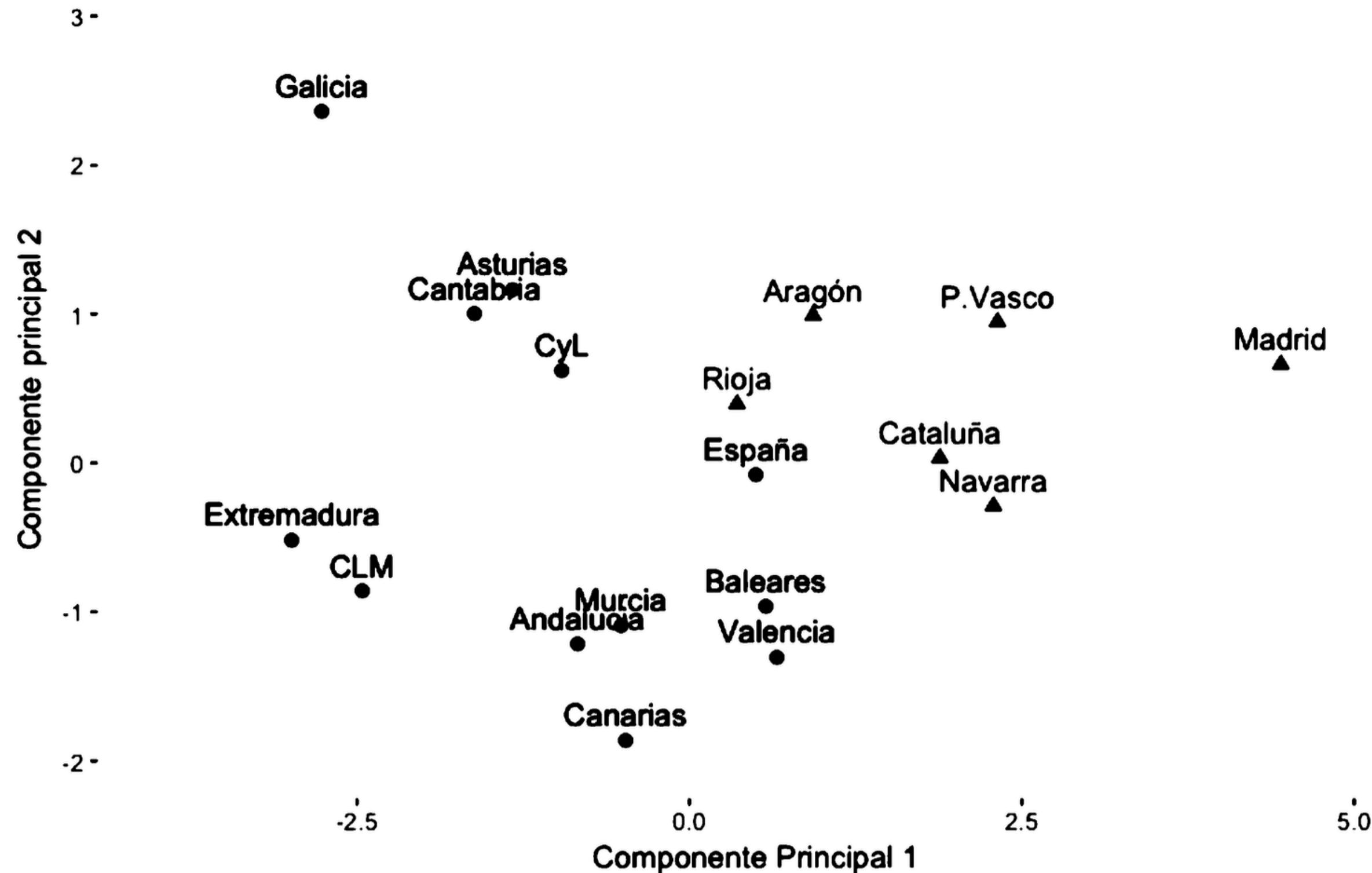
Aldas y uriel

.....

Visualización en un
plano de dos
dimensiones a
partir de la
reducción de
dimensionalidad
con componentes
principales

Diseño de un plan de incentivos para vendedores.

Figura 3.14.: Visualización de los resultados de un análisis de conglomerados



RECURSOS DE ILUSTRACIÓN GRATIS

PUEDES CAMBIAR EL COLOR DE ESTOS
ÍCONOS E ILUSTRACIONES GRATUITOS Y
USARLOS EN TU DISEÑO DE CANVA.

