Análisis Multivariante:

Análisis de Cluster

Análisis Multivariado Enfocado a la Gestión de Riesgos

Edith Johana Medina Hernández – edithjoh@gmail.com

Qué es el análisis de cluster

Objetivo: Buscar similitudes entre individuos para clasificarlos en grupos que sean lo más homogéneos posible.

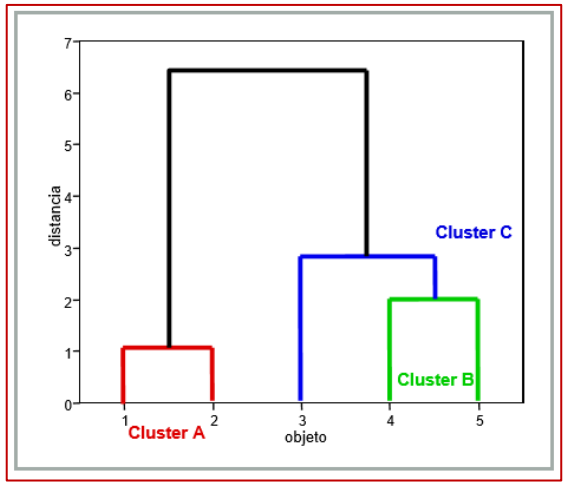
Estos métodos también se conocen como: Métodos de clasificación, Análisis de conglomerados, Análisis de segmentación (no supervisada) y Análisis clúster.

Es una técnica para combinar observaciones en grupos o cluster de forma que:

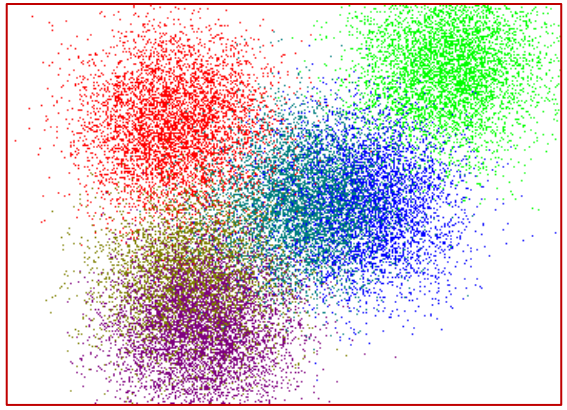
1.- Cada grupo o cluster sea homogéneo o compacto con respecto a ciertas características. Es decir las observaciones dentro de cada grupo han de ser similares entre sí.

2.- Cada grupo debe diferenciarse de los otros grupos respecto a las mismas características, es decir las observaciones de un grupo deben diferenciarse de las observaciones de los otros grupos.

Jerárquicos



No Jerárquicos



Premisas del Análisis Cluster

No existe ningún tipo de clasificación a priori (se desconoce la pertenencia de casos a grupos)

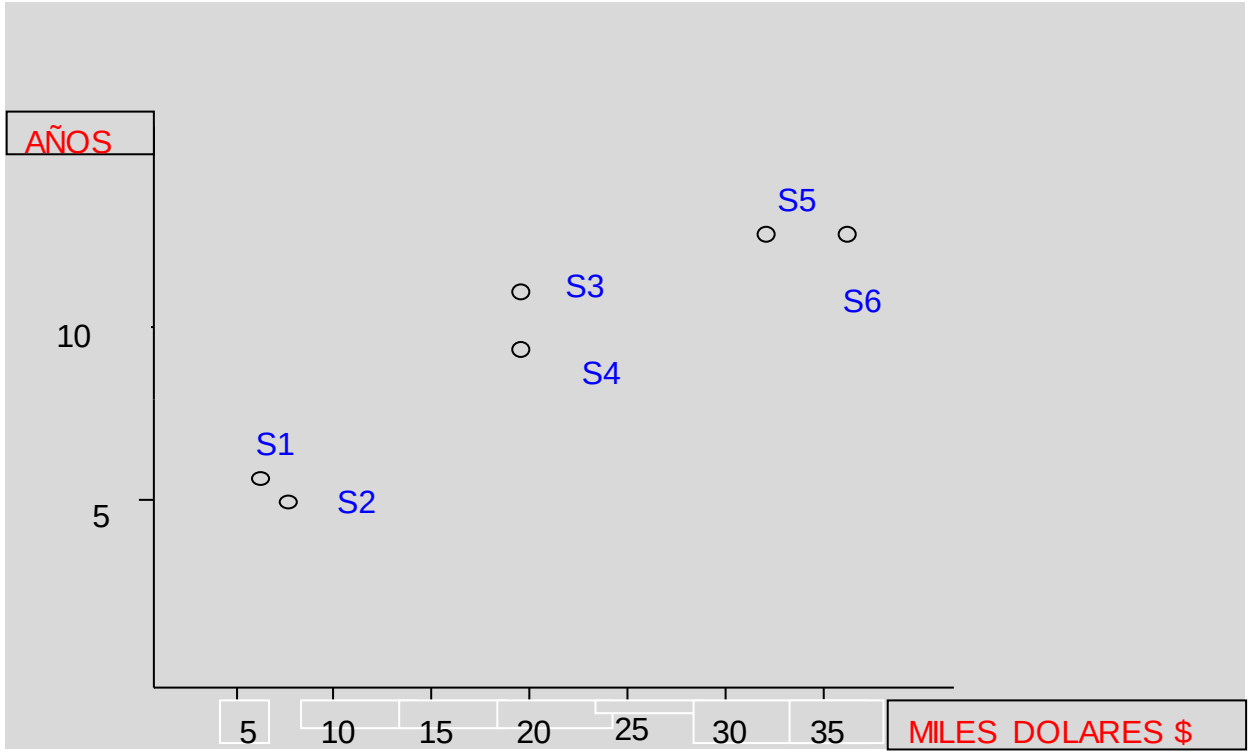
No se hipotetizan relaciones de dependencia entre las variables usadas (*Kinnear* y *Taylor*, 1996).

No hay ninguna restricción o condición que deban cumplir los datos para proceder a aplicar la técnica, por lo que con **cualquier tipo de datos** y sin preparación previa de éstos se puede realizar un análisis cluster.

Visión Geométrica del Análisis de Cluster

Consideremos los datos hipotéticos de la siguiente tabla:

INDIVIDUOS	GASTOS (miles$)	AÑOS EDUCACION
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19



Fases del Análisis de Cluster

PRIMERA FASE

- A.- Establecimiento objetivos
- B.- Selección de variables
- C.- Estandarización de variables
- D.- Detección de outliers (atípicos)
- E.- Colinealidad
- F.- **Selección medidas similitud**

SEGUNDA FASE

Selección del algoritmo

A.- Cluster jerárquicos

B.- Cluster no jerárquicos

- **TERCERA FASE**

Interpretación de los conglomerados

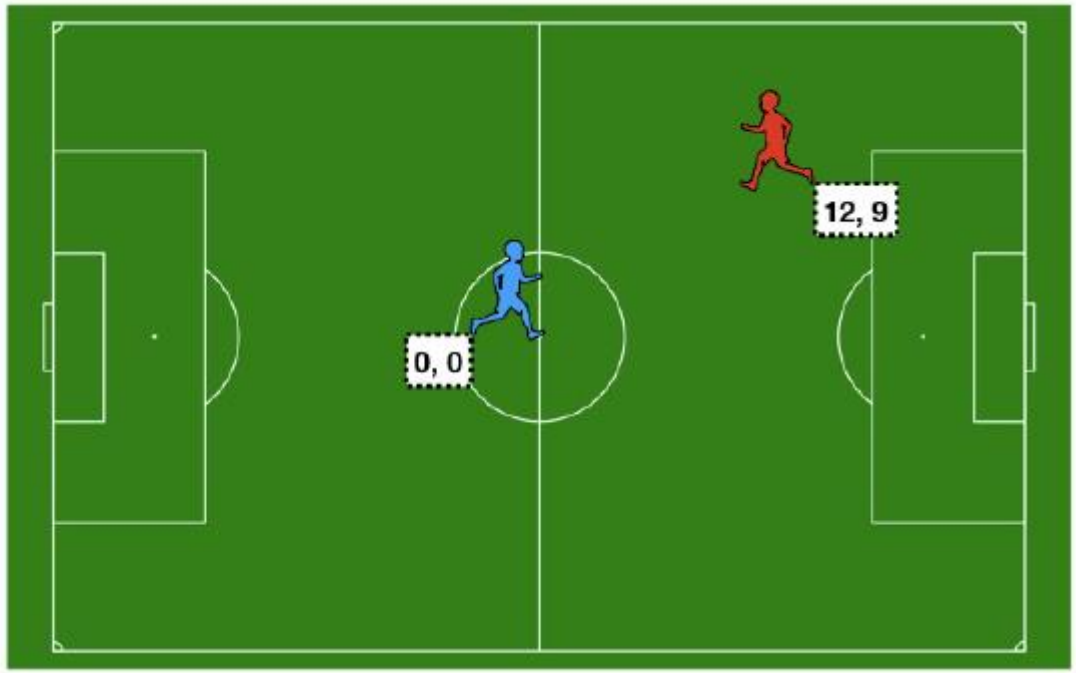
- **CUARTA FASE**
- Validación

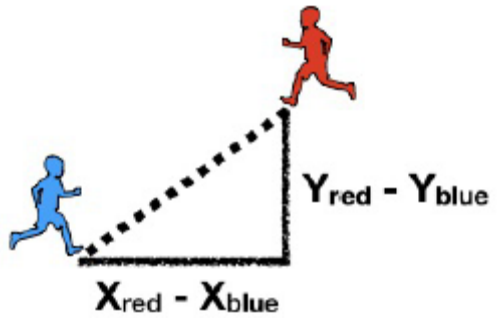
Medidas de Similitud y Disimilaridad o Distancia

Idea Intuitiva del cálculo de Distancias

Distancia entre 2 jugadores

	X	Y
Blue	0	0
Red	12	9





$$= \sqrt{(X_{red} - X_{blue})^2 + (Y_{red} - Y_{blue})^2} = \sqrt{(12 - 0)^2 + (9 - 0)^2} = 15$$

Distancias

Existen distintos tipos de distancias cada una con sus propiedades particulares pero las más habitualmente utilizadas son:

Distancia Euclídea

$$D_{i,j} = \left(\sum_{i=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}}$$

Donde $\mathbf{D_{ij}}$ es la distancia entre la observación **i** y la **j** y **p** es el número de variables

Ejemplo:

INDIVIDUOS	GASTOS (miles$)	AÑOS EDUCACION
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

$$D_{1,3} = \left(\sum_{i=1}^{2} (5-15)^2 + (5-14)^2 \right)^{\frac{1}{2}}$$

La distancia euclídea es la medida de similaridad más frecuentemente utilizada sin embargo tiene el problema de que no es invariante por cambios de escala.
Esto es, la distancia entre observaciones podría cambiar si cambiamos de escala.

Distancias

Supongamos que los gastos están medidos en dólares en lugar de en miles de dólares. La distancia euclídea entre las observaciones S1 y S2 que obtendríamos sería:

$$D_{1,2} = (5000 - 6000)^2 + (5 - 6)^2 = 100000 + 1 = 100001$$

Claramente es importante tener variables que estén medidas en una escala comparable!

Distancia euclídea estandarizada.-

Viene dada por:

$$D_{i,j}^2 = \sum_{i=1}^{K} \frac{\left(x_{ik} - x_{jk}\right)^2}{S_k}$$

Donde S_k es la desviación típica de la variable k-ésima

Tiene la ventaja de ser invariable por cambios de escala

Distancia de Mahalanobis.-

$$D_{i,j} = (x_i - x_j)' S^{-1} (x_i - x_j)$$

Donde S es la matriz de covarianzas dentro del grupo, en el caso del cluster la matriz de covarianzas total.

SEGUNDA FASE: A.- Selección del algoritmo

A.- CLUSTER JERARQUICOS

Puede subdividirse en **Aglometrativos y Divisivos**.

Los aglomerativos proporcionan sucesivas fusiones de los n objetos individuales en grupos.

Los divisivos particionan los n individuos (agrupados en un solo grupo) sucesivamente en grupos más finos.

Las clasificaciones jerárquicas se representan en un diagrama bidimensional conocido como **dendrograma** que indica las fusiones o divisiones hechas en las sucesivas fases del análisis.

B.- CLUSTER NO JERARQUICOS

Clusters No Jerárquicos

Clusters No jerárquicos

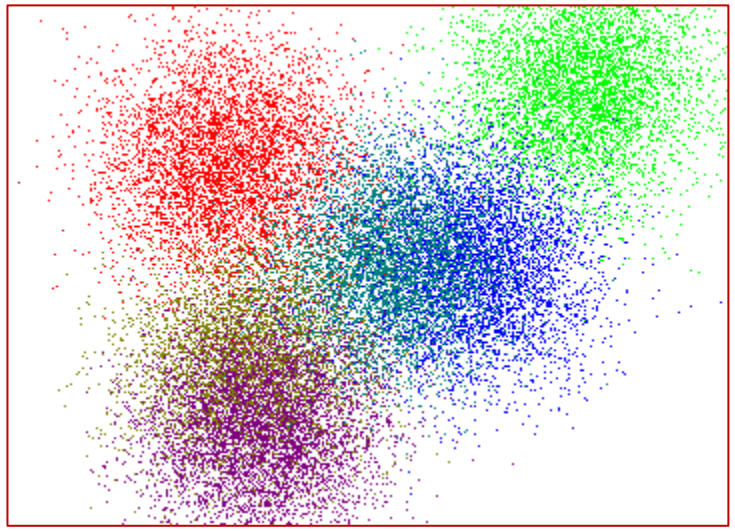
El número de cluster debe conocerse a priori.

Básicamente siguen los siguientes pasos:

1.- Seleccionar K centroides iniciales, siendo K el número de clusters deseados.
2.- Asignar cada observación al cluster que le sea más cercano.

3.- Reasignar o relocalizar cada observación a uno de los K cluster de acuerdo con alguna regla de parada.

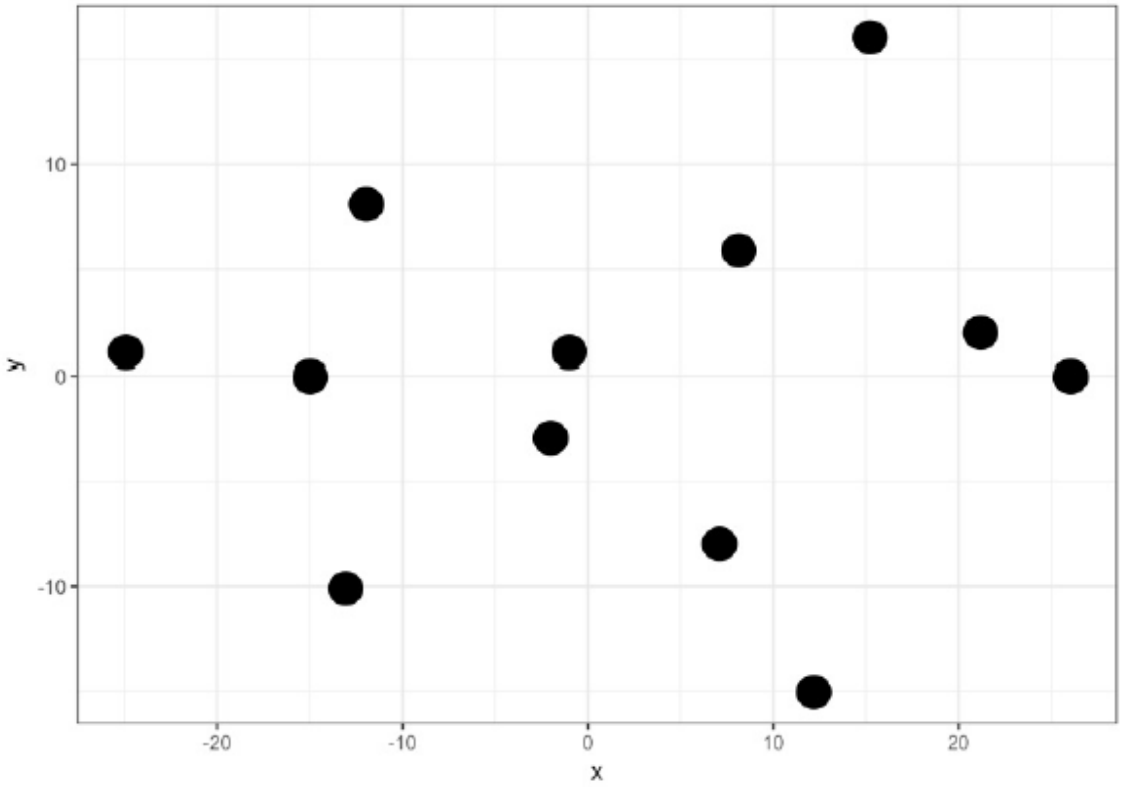
4.- Parar si no hay reasignación de los puntos o si la reasignación satisface la regla de parada. En otro caso se vuelve al paso dos.

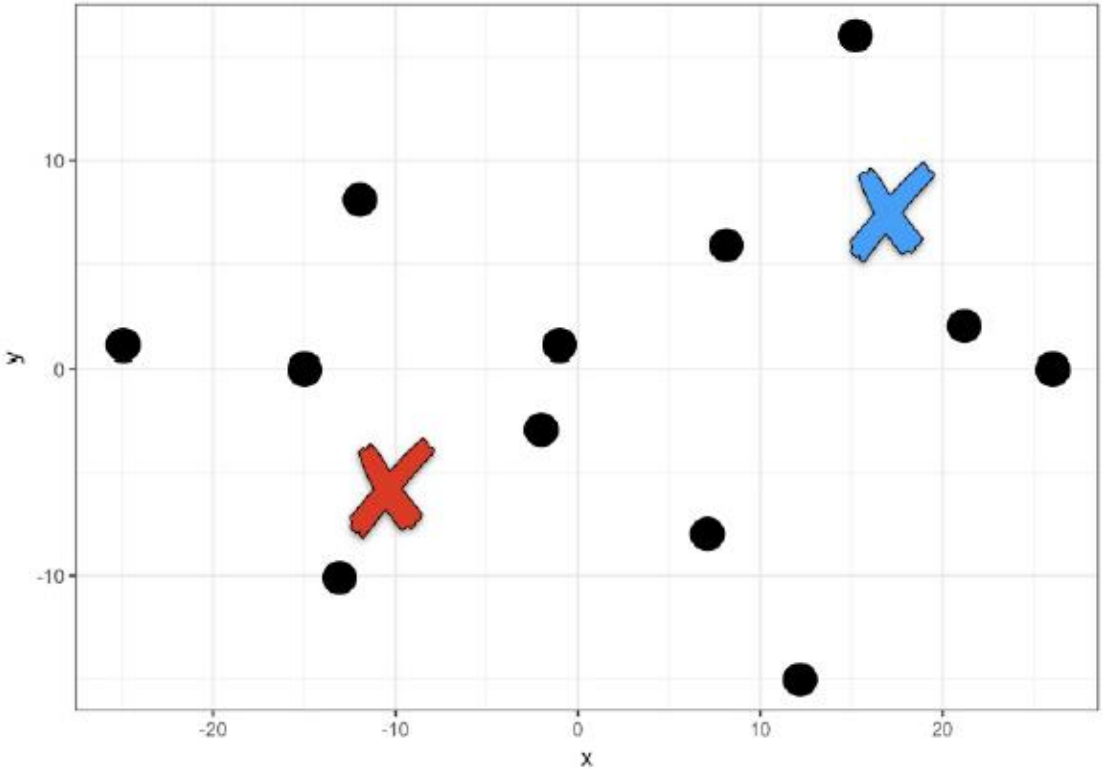


La mayoría de los algoritmos no jerárquicos difieren con respecto a:

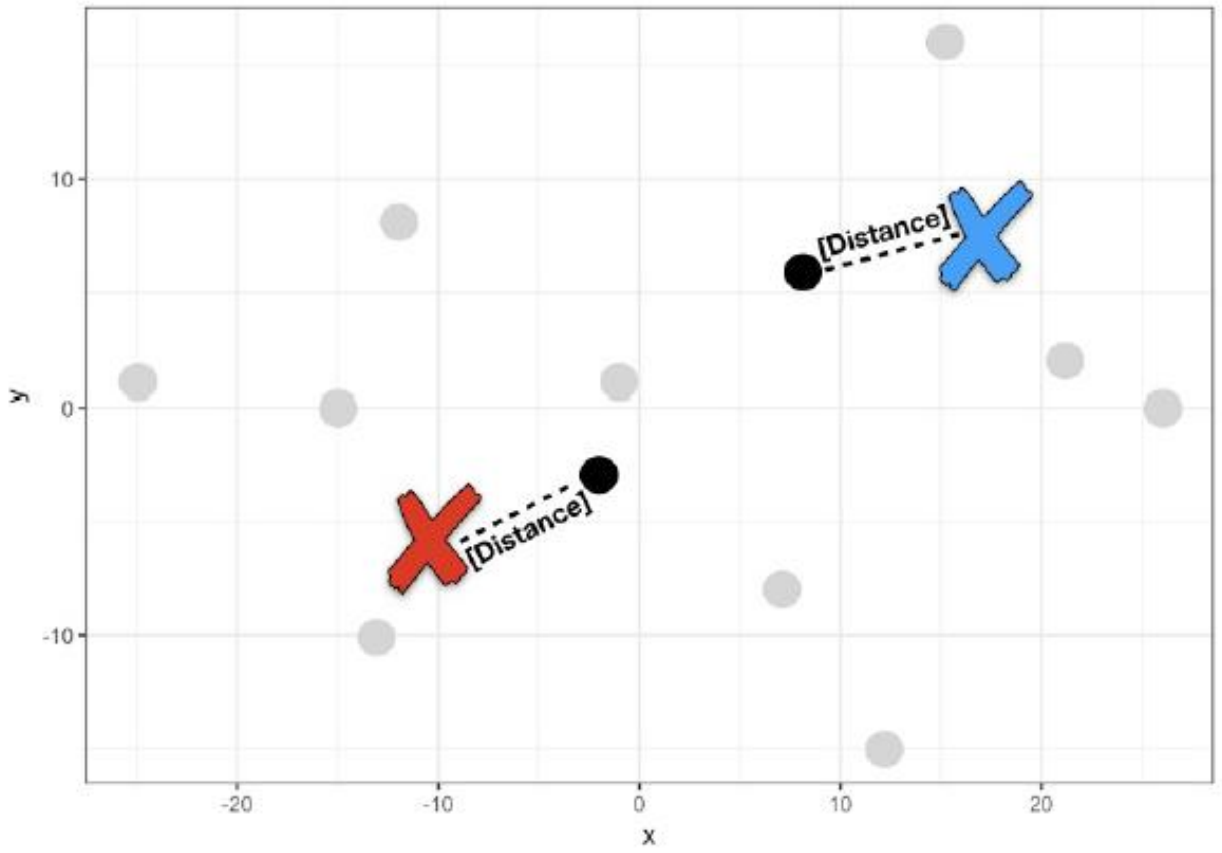
- El procedimiento para obtener los centroides iniciales
- La regla que se usa para reasignar las observaciones

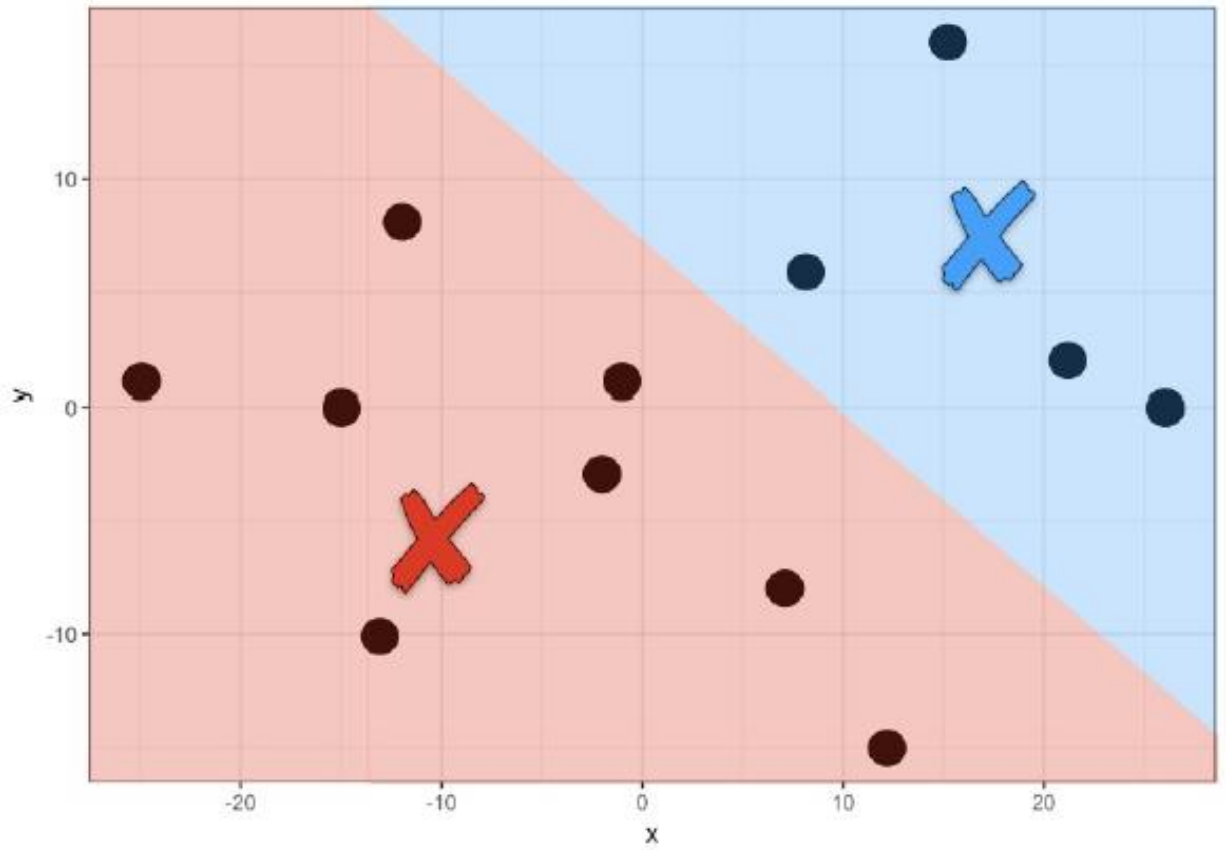
K-means



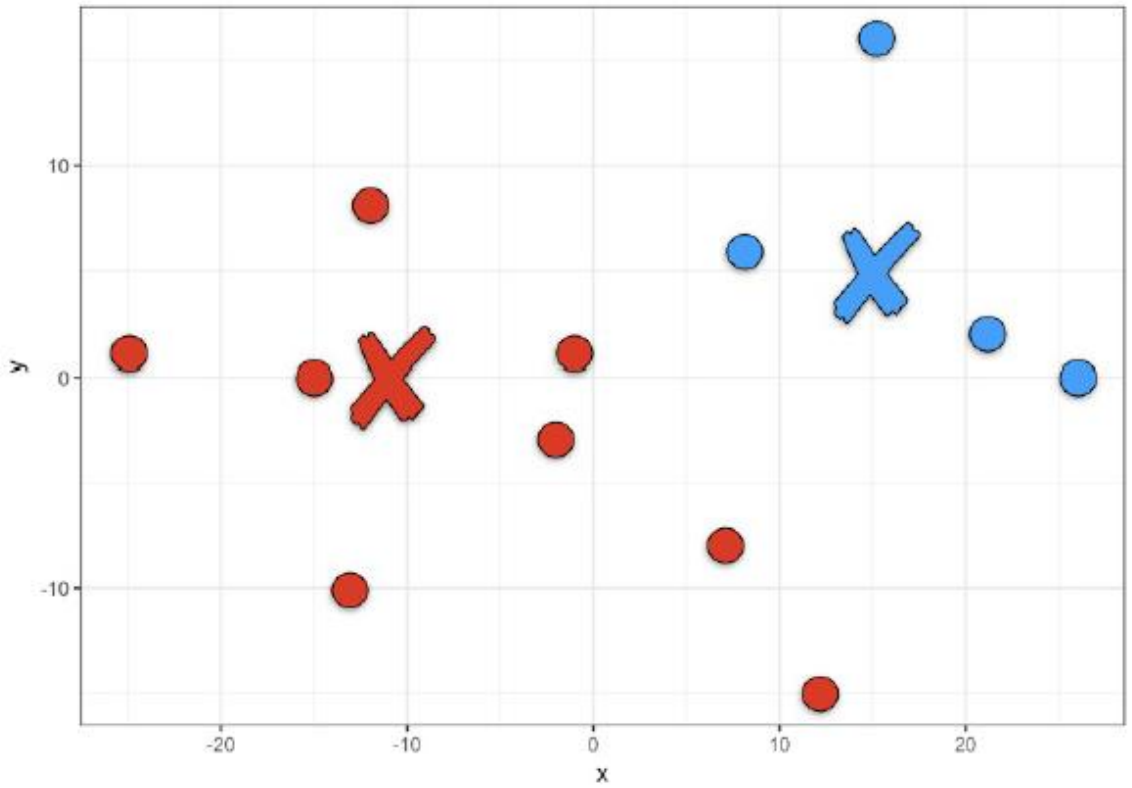


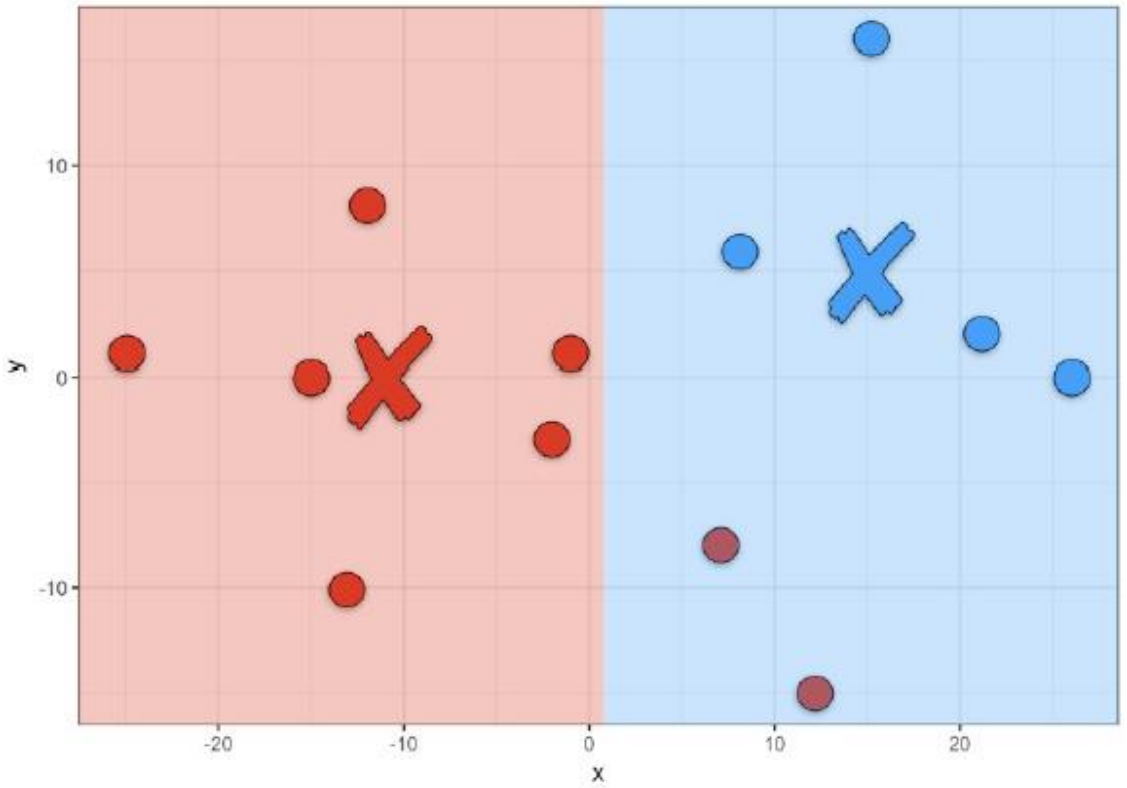
K-means





K-means



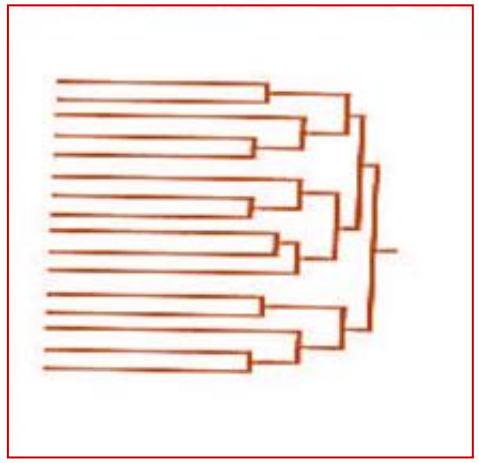


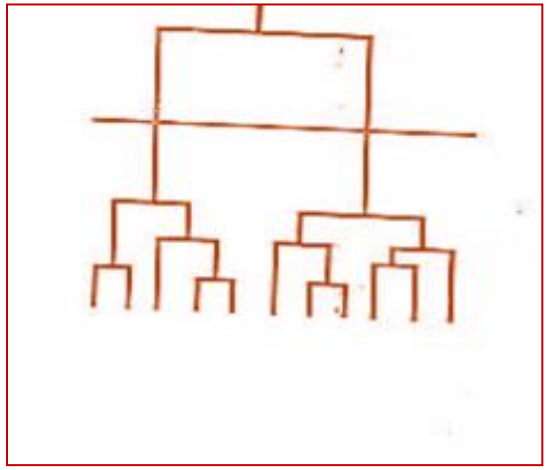
Clusters Jerárquicos

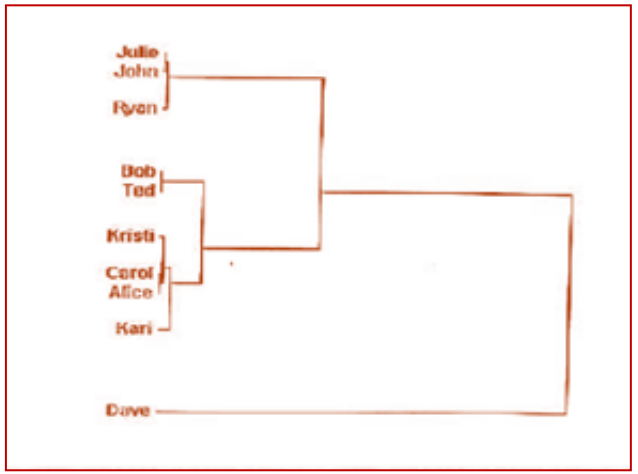
A.- CLUSTER JERARQUICOS

Dendrograma.- Los objetos se representan como nodos y las ramas del árbol indican los sujetos que se han fusionado en un cluster, la longitud de las ramas indican la distancia de la fusión**.**

Un dendrograma que diferencie grupos de objetos claramente tendrá pequeñas distancias en las ramas lejanas del árbol y grandes diferencias en las ramas cercanas.
Cuando las distancias de las ramas lejanas son relativamente grandes con respecto a las cercanas el agrupamiento no será tan efectivo.







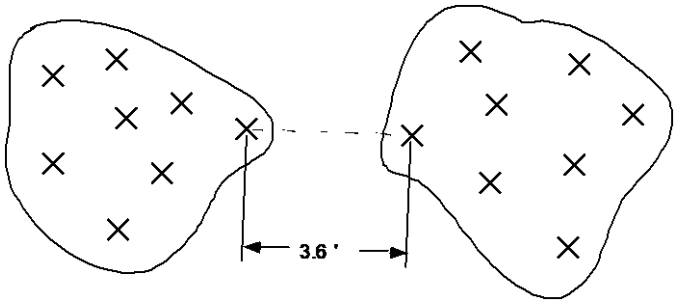
A.- Selección del algoritmo

La diferencia entre los distintos métodos está en la diferente forma de definir la distancia entre un individuo y un grupo que contenga varios individuos, o entre dos grupos de individuos.

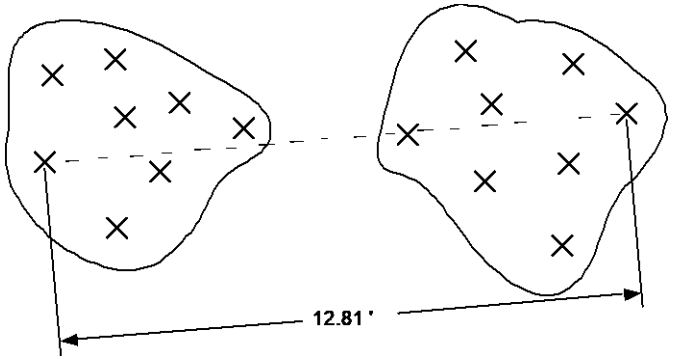
Posibles métodos aglomerativos:

1.- Vecino más próximo (Single linkage)
2.- Vecino más lejano (completo linkage)
3.- Grupo mediano (Group Average)
4.- Método del centroide
5.- Cluster mediano
6.- Método de Ward.

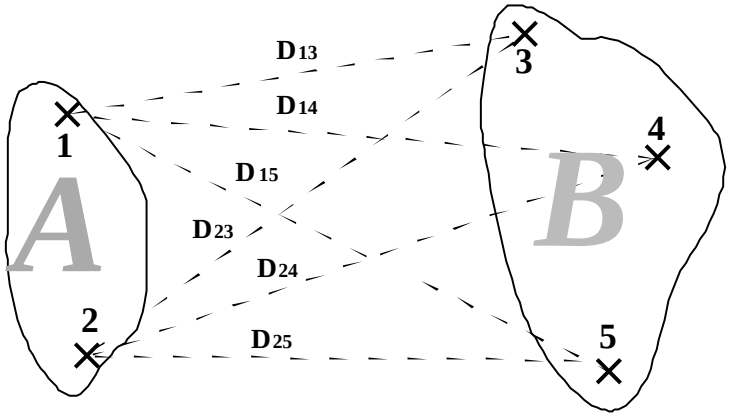
Vecino más próximo



Vecino mas lejano



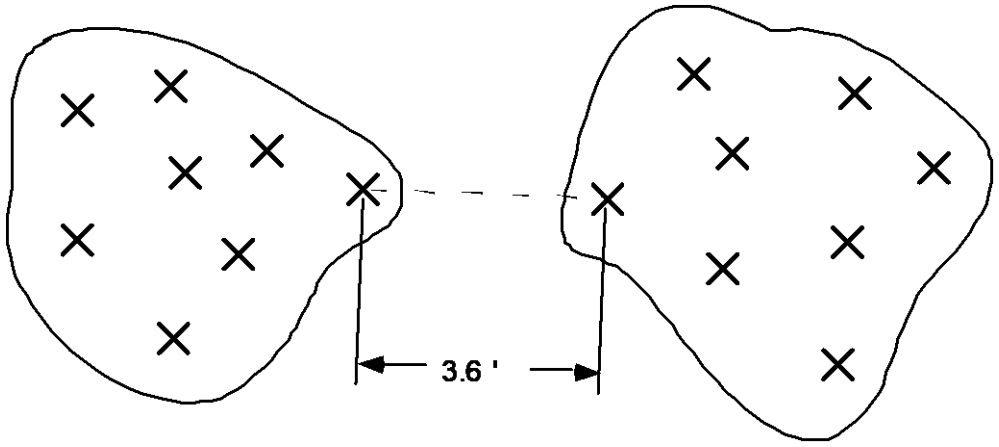
Grupo mediano
Distancia media entre todos los pares de puntos



1.- Método del vecino más próximo.

Es uno de los métodos jerárquicos aglomerativos más sencillos. Fue descrito por Florek et al. (1951) y posteriormente por Sneath (1957) y Jonson (1967).

La característica que define a este método es que la distancia entre grupos se define como la del par de individuos que está más cercano.



Ejemplo del Cálculo de Distancias

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0,0 & & & & \\ 2,0 & 0,0 & & & \\ 6,0 & 5,0 & 0,0 & & \\ 12,0 & 9,0 & 4,0 & 0,0 & \\ 9,0 & 8,0 & 5,0 & 3,0 & 0,0 \end{bmatrix} \end{array}$$

El valor más pequeño se da entre los individuos 1 y 2, consecuentemente los juntaremos en un cluster con dos miembros.

Las distancias entre este cluster y los otros clusters individuales se obtienen:

$$D_{(1-2),3} = \min\{D_{1,3}; D_{2,3}\} = D_{2,3} = 5,0$$

$$D_{(1-2),4} = \min\{D_{1,4}; D_{2,4}\} = D_{2,4} = 9,0$$

$$D_{(1-2),5} = \min\{D_{1,5}; D_{2,5}\} = D_{2,5} = 8,0$$

1.- Método del vecino más próximo.

Se construye una nueva matriz de distancias:

$$D_2 = \begin{array}{c} \\ 1-2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1-2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0,0 & & & \\ 5,0 & 0,0 & & \\ 9,0 & 4,0 & 0,0 & \\ 8,0 & 5,0 & 3,0 & 0,0 \end{bmatrix} \end{array}$$

El valor más pequeño se da entre los individuos 4 y 5 por lo que ellos serán los que formen el siguiente cluster de dos miembros. De nuevo hemos de recalcular las distancias:

$$D_{(1-2),3} = \min\{D_{1,3}; D_{2,3}\} = D_{2,3} = 5,0$$

$$D_{(1-2),(4-5)} = \min\{D_{1,4}; D_{1,5}; D_{2,4}; D_{2,5}\} = D_{2,5} = 8,0$$

$$D_{(4-5),3} = \min\{D_{3,4}; D_{3,5}\} = D_{3,4} = 4,0$$

Formamos la matriz de distancias D_3:

$$D_3 = \begin{array}{c} \\ 1-2 \\ 3 \\ 4-5 \end{array} \begin{array}{c} \begin{array}{ccc} 1-2 & 3 & 4-5 \end{array} \\ \begin{bmatrix} 0,0 & & \\ 5,0 & 0,0 & \\ 8,0 & 4,0 & 0,0 \end{bmatrix} \end{array}$$

El valor más pequeño se da entre 3 el cluster 4-5, por lo que el individuo 3 se añade al cluster que contiene a los individuos 4 y 5.

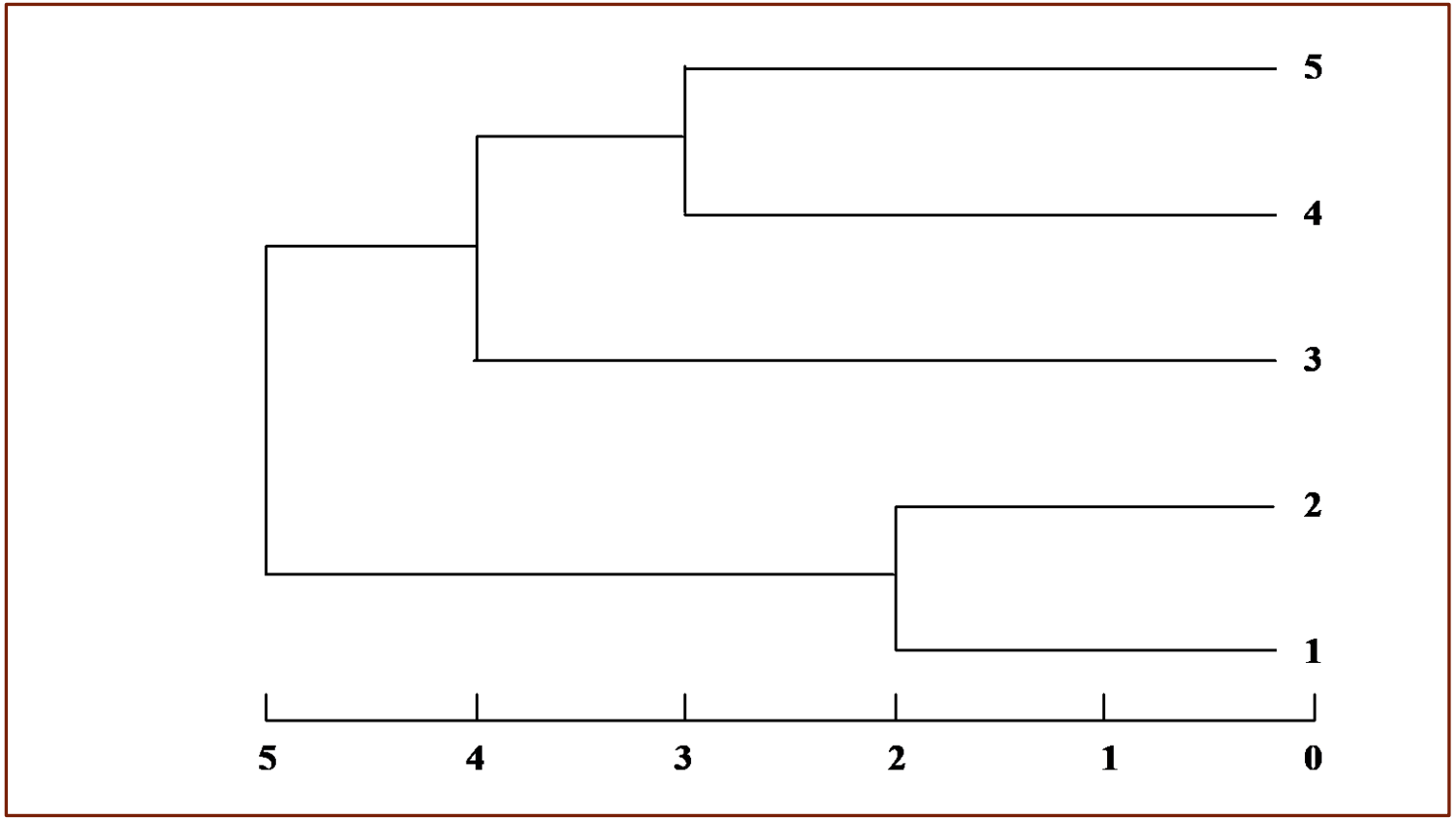
Finalmente el grupo que contienen los elementos 1 y 2 y el que contiene los individuos 3, 4 y 5 se combinan en un cluster único.

Las particiones producidas en cada paso son:

P_5 [1] [2] [3] [4] [5]
P_4 [1 – 2] [3] [4] [5]
P_3 [1 – 2] [3] [4 – 5]
P_2 [1 – 2] [3 – 4 – 5]
P_1 [1 – 2 – 3 – 4 – 5]

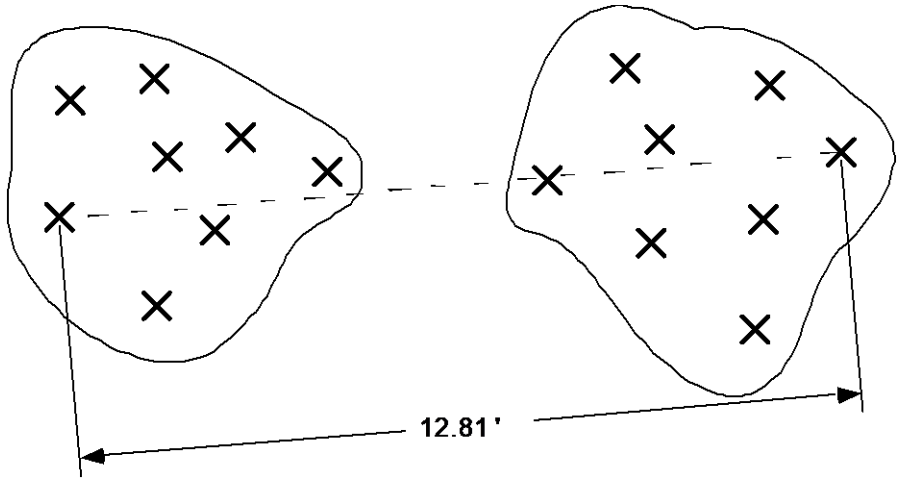
1.- Método del vecino más próximo.

El correspondiente dendograma sería:



2.- Método del vecino mas lejano (COMPLETE LINKAGE).

Este método es el opuesto al anterior, en el sentido que la distancia entre grupos la definimos ahora como la mayor distancia entre pares de individuos, uno de cada grupo.



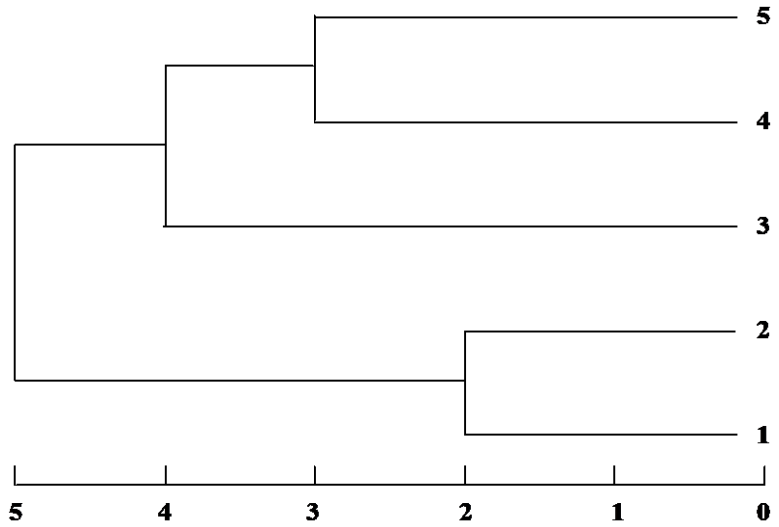
Usando este método sobre la matriz D_1, el primer paso será, de nuevo, fusionar los individuos 1 y 2. Las distancias entre este grupo y los tres individuos restantes será:

$$D_{(1-2),3} = \max\left\{D_{1,3}; D_{2,3}\right\} = D_{2,3} = 6,0$$

$$D_{(1-2),4} = \max\left\{D_{1,4}; D_{2,4}\right\} = D_{2,4} = 10,0$$

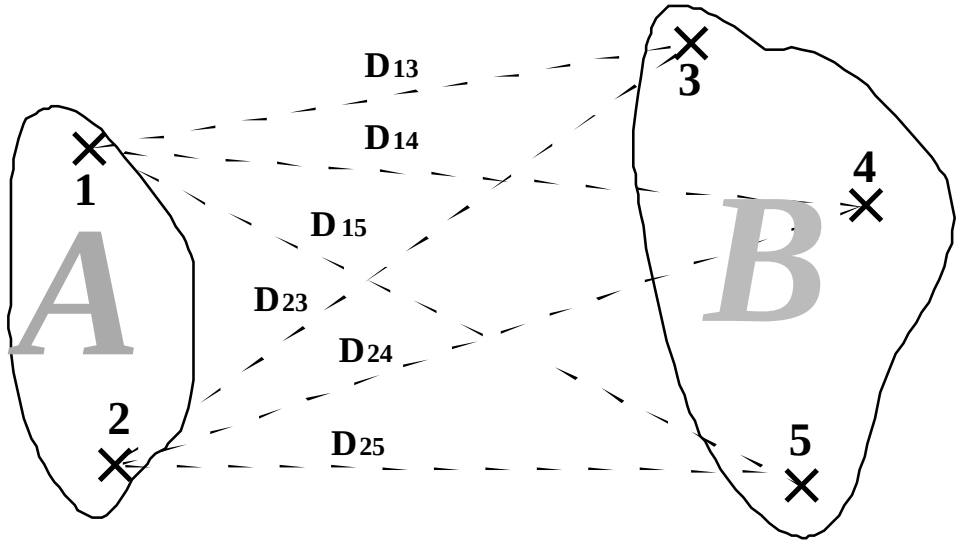
$$D_{(1-2),5} = \max\left\{D_{1,5}; D_{2,5}\right\} = D_{2,5} = 9,0$$

El proceso se repetiría y el dendograma resultante sería:



3.- (Group Average).

En el la distancia entre dos cluster se define como la distancia media entre todos los pares de individuos de cada grupo.



$$D_{AB} = \frac{D_{13} + D_{14} + D_{15} + D_{23} + D_{24} + D_{25}}{6}$$

Aplicando el método a la matriz D_1, el primer paso lo mismo que en los métodos anteriores es la formación del cluster que contiene a los elementos 1 y 2.

Recalculando:

$$D_{(1-2),\,3} = \tfrac{1}{2}\left(D_{1,3} + D_{2,3}\right) = \tfrac{(6+5)}{2} = 5{,}5$$

$$D_{(1-2),\,4} = \tfrac{1}{2}\left(D_{1,4} + D_{2,4}\right) = \tfrac{(10+9)}{2} = 9{,}5$$

$$D_{(1-2),\,5} = \tfrac{1}{2}\left(D_{1,5} + D_{2,5}\right) = \tfrac{(9+8)}{2} = 8{,}5$$

$$D_2 = \begin{array}{c} \\ 1-2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1-2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0 & & & \\ 5{,}5 & 0{,}0 & & \\ 9{,}5 & 4{,}0 & 0{,}0 & \\ 8{,}5 & 5{,}0 & 3{,}0 & 0{,}0 \end{bmatrix} \end{array}$$

3.- (Group Average).

El valor más pequeño, y por tanto el consiguiente cluster, está formado por los individuos 4 y 5. La distancia entre ambos grupos se calcula:

$$D_{(1-2),(4-5)} = \frac{D_{14} + D_{15} + D_{24} + D_{25}}{4} = 9$$

y el proceso continuaría como antes.

Este método se denomina también :UPGMA.

Los métodos descritos operan directamente sobre la matriz de proximidades y no necesitan los valores originales de las variables en los individuos.

4- Método del centroide

Un método que requiere los datos originales es el método del centroide.

Con este método, los grupos una vez formados se representan por sus valores medios para cada variable, es decir **su vector de medias** y las distancias entre los grupos se definen en términos de las distancia entre vectores de medias.

El uso de medias implica, estrictamente hablando, que las **variables estén medidas en escala de intervalo**, el método sin embargo a menudo se usa para otro tipo de variables.

Para ilustrar el método trabajaremos con los siguientes datos:

Individuo	Variable 1	Variable 2
1	1,0	1,0
2	1,0	2,0
3	6,0	3,0
4	8,0	2,0
5	8,0	0,0

Supongamos que elegimos la distancia euclídea común como medida de distancia entre individuos, dando la siguiente matriz de distancias:

1	0				
2	1	0			
3	5,39	5,10	0		
4	7,07	7,0	2,24	0	
5	7,07	7,28	3,61	2	0

4- Método del centroide

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 5{,}39 & 5{,}10 & 0 & & \\ 7{,}07 & 7{,}0 & 2{,}24 & 0 & \\ 7{,}07 & 7{,}28 & 3{,}61 & 2 & 0 \end{bmatrix}$$

Si examinamos la matriz vemos que $D_{1,2}$ es el valor más pequeño y los individuos 1 y 2 se fusionan para formar un grupo.

Se calcula el vector de medias del grupo (1.0, 1.5) y se calcula una nueva matriz de distancias.

$$D_{(1-2),3} = \sqrt{(6-1)^2 + (3-1{,}5)^2} = 5{,}22$$

$$\begin{array}{c} (1,2) \\ 3 \\ 4 \\ 5 \end{array}\begin{bmatrix} 0 & & & \\ 5{,}22 & 0 & & \\ 7{,}02 & 2{,}24 & 0 & \\ 7{,}16 & 3{,}61 & 2 & 0 \end{bmatrix}$$

El valor más pequeño de esta matriz es entre los individuos 4 y 5 y por lo tanto se fusionan en un segundo grupo, el vector de medias se calcula y da (8.0, 1.0) y se vuelve a calcular la matriz de distancias:

$$\begin{array}{c} (1,2) \\ 3 \\ (4,5) \end{array}\begin{bmatrix} 0 & & \\ 5{,}22 & 0 & \\ 7{,}02 & 2{,}83 & 0 \end{bmatrix}$$

En esta el valor más pequeño es entre el individuo (4,5) y el 3, por lo que se fusionan en un cluster de tres individuos.
El paso final consiste en la fusión de los dos grupos restantes en uno.

5- Método del cluster mediano

Una de las desventaja el método del centroide es que si los tamaños de los dos grupos que se fusionan es muy diferente entonces el centroide del nuevo grupo estará más próximo al grupo mayor.

Las propiedades características del grupo más pequeño se pierden.

La estrategia puede hacerse independiente del tamaño del grupo asumiendo que los grupos que se fusionan tienen el mismo tamaño, la posición aparente del nuevo grupo estará siempre entre los dos grupos a fusionarse.

La distancia entre un individuo o grupo K de centroide k y el grupo formado por la fusión de los grupos I y J de centroides i y j viene dada por **la mediana del triángulo i,j, k.** Razón por la cual Gower propuso el nombre de método (distancia) de la mediana.

6- Método del Ward

Este método no calcula distancias entre cluster. Lo que hace es formar cluster de forma que se maximice la homogeneidad intra cluster.

Se usa como medida de la homogeneidad la suma de cuadrados intra grupos.

Los cluster que se forman en un paso son los que minimizan la suma de cuadrados intra grupos.

Esta suma de cuadrados se le conoce como suma de cuadrados de los errores (ESS).

Paso 1 Cada sujeto es un grupo.

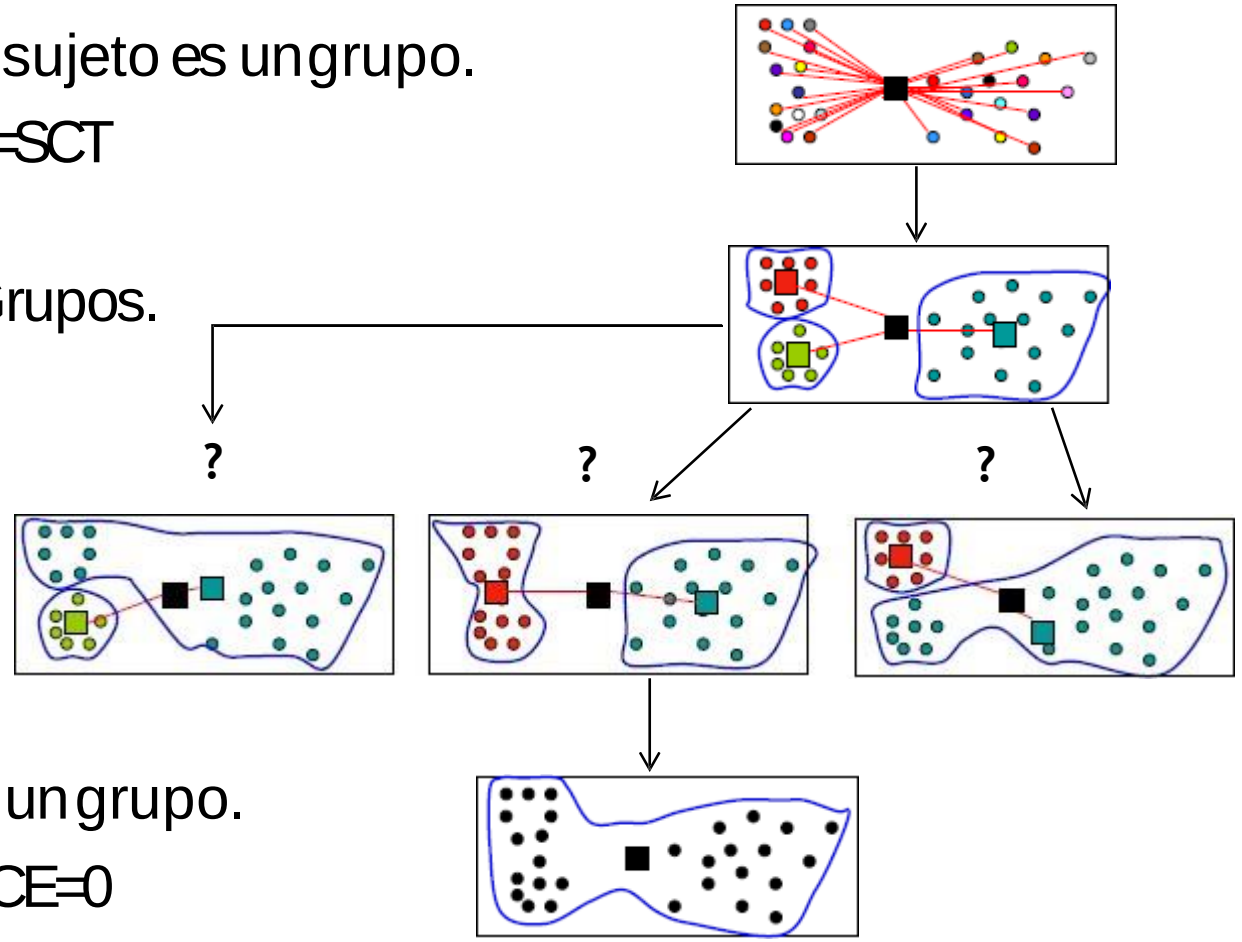
SCD=0, SCE=SCT

Paso i-2: 3 Grupos.

Paso i-1:
2 Grupos.

Paso i. Solo un grupo.

SCD=SCT, SCE=0



6- Método del Ward

Supongamos de nuevo la tabla de partida:

INDIV	GASTO	AÑOS EDUC
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

Inicialmente cada observación es un cluster y por tanto su ESS es cero.

El siguiente paso es formar cinco cluster, un cluster de tamaño dos y otros cuatro de tamaño uno.

Por ejemplo podemos formar un cluster con los individuos 1 y 2 y otros cuatro con el resto

El ESS del cluster con dos observaciones (S1, S2) es:

$$(5-5{,}5)^2+(6-5{,}5)^2+(5-5{,}5)^2+(6-5{,}5)^2=1$$

6- Método del Ward

Si calculamos las 15 posibles soluciones de cinco cluster

	1	2	3	4	5	ESS
1	S1-S2	S3	S4	S5	S6	1
2	S1-S3	S2	S4	S5	S6	90.5
3	S1-S4	S2	S3	S5	S6	110.5*
4	S!-S5	S2	S3	S4	S6	312.5
5	S1-S6	S2	S3	S4	S5	410.5
6	S2-S3	S1	S4	S5	S6	72.5
7	S2-S4	S1	S3	S5	S6	90.5
8	S2-S5	S1	S3	S4	S6	278.5
9	S2-S6	S1	S3	S4	S5	372.5
10	S3-S4	S1	S2	S5	S6	1
11	S3-S5	S1	S2	S4	S6	68
12	S3-S6	S1	S2	S4	S5	125
13	S4-S5	S1	S2	S3	S6	53
14	S4-S6	S1	S2	S3	S5	106
15	S5-S6	S1	S2	S3	S4	13

*** Ejemplo de cálculo de ESS entre S1 y S4:**

Hacemos las medias

$$\frac{5+16}{2}=10.5 \;;\; \frac{5+15}{2}=10$$

Hacemos las sumas de cuadrados:

$$(5-10.5)^2+(16-10.5)^2+(5-10)^2+(15-10)^2=110.5$$

Basado en el criterio de minimizar ESS, podemos seleccionar el cluster 1 ó el 10 ya que ambos dan un ESS de 1. La elección es al azar. Supongamos que elegimos el cluster 1 es decir el formado por S1-S2.

6- Método del Ward

El siguiente paso es formar cuatro cluster. Hay 10 posibles soluciones

$$\frac{5x4}{2} = 10$$

Por ejemplo el ESS para el cluster formado por S1-S2-S3 es:

$$(5-8.67)^2 + (6-8.67)^2 + (15-8.67)^2 + (5-8.33)^2 +$$
$$+(6-8.33)^2 + (14-8.33)^2 = 109.33$$

	1	2	3	4	ESS
1	S1-S2-S3	S4	S5	S6	109.33
2	S1-S2-S4	S3	S5	S6	134.66
3	S1-S2-S5	S3	S4	S6	394.66
4	S1-S2-S6	S3	S4	S5	522.66
5	S1-S2	S3-S4	S5	S6	**2**
6	S1-S2	S3-S5	S4	S6	69
7	S1-S2	S3-S6	S4	S5	126
8	S1-S2	S4-S5	S3	S6	54
9	S1-S2	S4-S6	S3	S5	107
10	S1-S2	S5-S6	S3	S4	14

El cluster 5 es el que minimiza el ESS, con un valor de 2
Este procedimiento se repite en todos los restantes pasos.

¿Qué método jerárquico deberíamos de utilizar?.

1.- Algunos jerárquicos son susceptibles de *encadenamiento.* En general el **vecino más próximo** es más susceptible que el de el más lejano.

2.- Si comparamos el simple linkage con el **completo linkage** a este último *le afectan menos los outliers.*

3.- El del **completo linkage** identifica *cluster compactos* en los que las observaciones son muy similares entre sí.

4.- El método de **ward** tiende a encontrar *cluster compactos y prácticamente de igual tamaño y forma.*

En general se recomienda usar varios métodos para comparar la consistencia y usar el método que tenga solución interpretable.

¿Qué método de cluster elegir?

Cluster jerárquicos:

- No requieren un conocimiento a priori del número de cluster o de la partición de partida.
- Los jerárquicos se usan a menudo con fines exploratorios y la solución resultante se utiliza en los no jerárquicos para afinar la solución.
- Ambas técnicas podrían verse como métodos complementarios y no como competitivos.

Cluster no jerárquicos:

- Necesitan conocimiento previo del número de cluster
- Hemos de identificar los centros de los cluster antes de proceder con las observaciones.
- Los algoritmos son muy sensibes a las particiones iniciales

Interpretación de los conglomerados

La interpretación implica examinar cada cluster para **asignar una etiqueta precisa que describa la naturaleza de cada cluster**, para ello se analizan los centroides de los grupos.

Para la interpretación se suele utilizar el centroide del conglomerado, pero **si los datos se estandarizan el investigador tendría que retroceder a las puntuaciones dadas por los encuestados en las variables originales.**

La interpretación del conglomerado consigue algo más que una descripción ya que proporciona un medio de evaluar los conglomerados obtenidos con aquellos propuestos por una teoría a priori o por la experiencia práctica.

Interpretación de los conglomerados

¿ Como saber si la solución cluster obtenida es representativa de la población?

- Realizar análisis de cluster para muestras distintas.

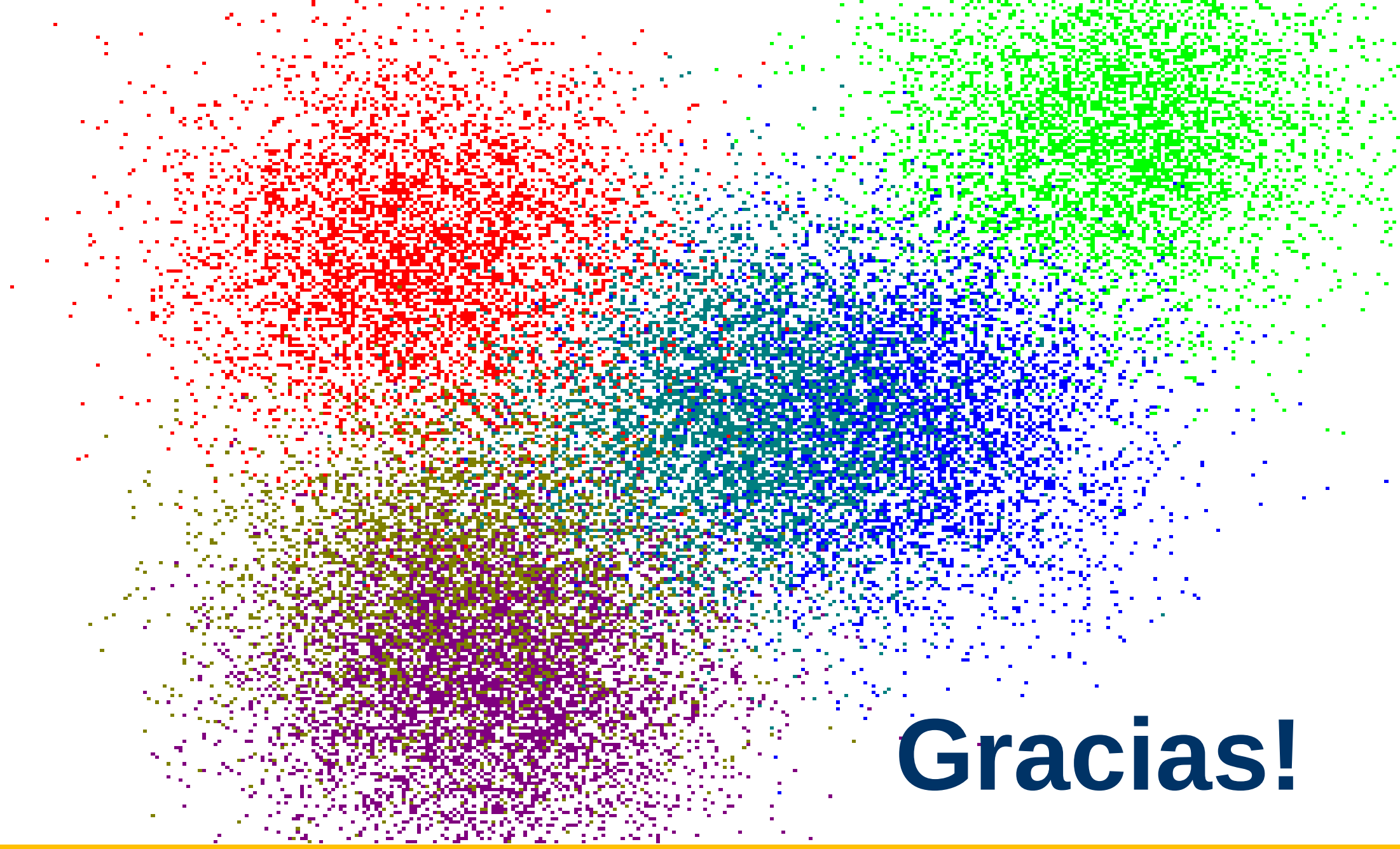
- Dividir la muestra en dos grupos.
  Cada submuestra se analiza por separado y se comparan luego los resultados.

-Criterio de validez predictiva
Selección una o más variables no utilizadas en el análisis pero que se sabe que cambian a lo largo de los conglomerados.

Revisar ejemplo en:

https://www.r-bloggers.com/k-means-clustering-in-r/

Clusters con R



Gracias!