

4. Análisis de Conglomerados (clusters)

El análisis consiste en agrupar un conjunto de datos multidimensionales (filas de la matriz de datos X) en un conjunto de grupos homogéneos. Para ello se utilizan funciones de similitud o similaridad entre ellos.

El análisis de conglomerados también puede utilizarse para agrupar variables (Columnas de la matriz de datos X).

Estos métodos también se conocen como métodos de clasificación automática o no supervisada.

Hay básicamente dos tipos de análisis de conglomerados:

- Métodos jerárquicos que no asumen ningún modelo estadístico para los datos.
- Métodos que asumen un modelo definido para los datos.

En el caso de los métodos jerárquicos los datos se ordenan en niveles de manera que los niveles superiores contienen a los inferiores. La jerarquía construida permite obtener también una partición de los datos en grupos. En este caso se utiliza la matriz de distancias o similitudes entre elementos de la matriz de datos.

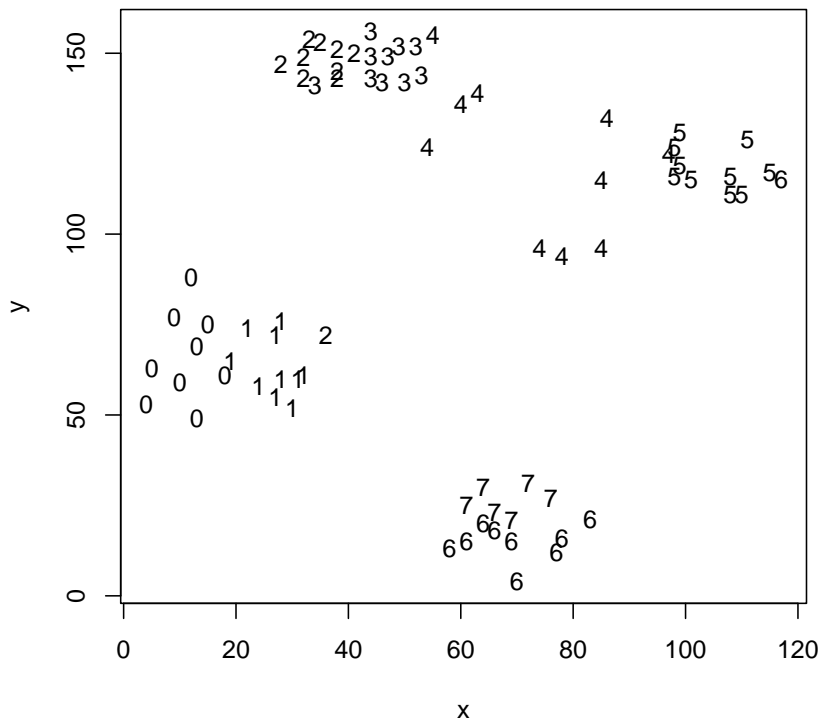
4.1 Conglomerados jerárquicos

Se analizarán los datos de Ruspini (1970)

```
> library(cluster)
> data(ruspini)
> help (ruspini)
```

Estos datos consisten de 75 observaciones de dos variables.

El gráfico de los datos es el siguiente:



Los números representan el primer dígito de la i -ésima observación. Los valores 0 son las primeras nueve observaciones.

Tipos de algoritmos jerárquicos

Los algoritmos jerárquicos pueden ser de dos tipos: De División y de Aglomeración.

- El algoritmo de división asume que en un primer paso todos los datos conforman un sólo conglomerado. Este cluster se va dividiendo sucesivamente en conglomerados más pequeños de acuerdo a algún criterio seleccionado previamente. El resultado de este procedimiento se representa por el *dendograma*.
- En el algoritmo de aglomeración cada observación inicialmente es un conglomerado y en cada paso se asocian los conglomerados más similares hasta llegar a un sólo cluster.

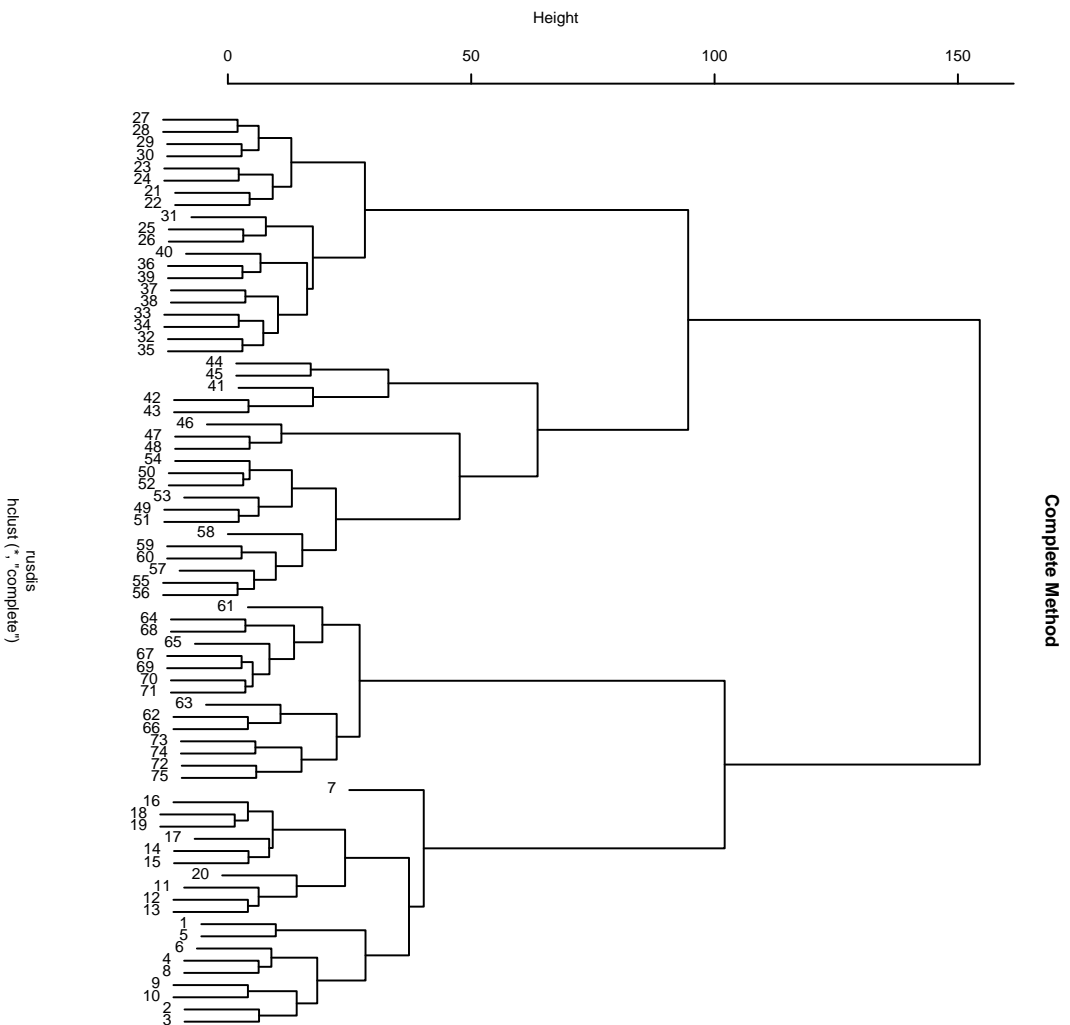
En el dendograma la escala vertical representa la distancia. La distancia entre dos conglomerados que se calcula según un algoritmo predeterminado. La implementación *hclust* de **R** utiliza el método Lance-Williams que calcula y actualiza en cada paso la disimilaridad entre clusters.

Si cortamos el dendograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

Hay que ser cuidadoso ya que el dendograma dos puntos pueden parecer próximos cuando no lo están y pueden parecer alejados cuando están próximos.

A continuación se calcula la matriz de distancias, se aplica un cluster jerárquico y se grafica el dendograma resultante:

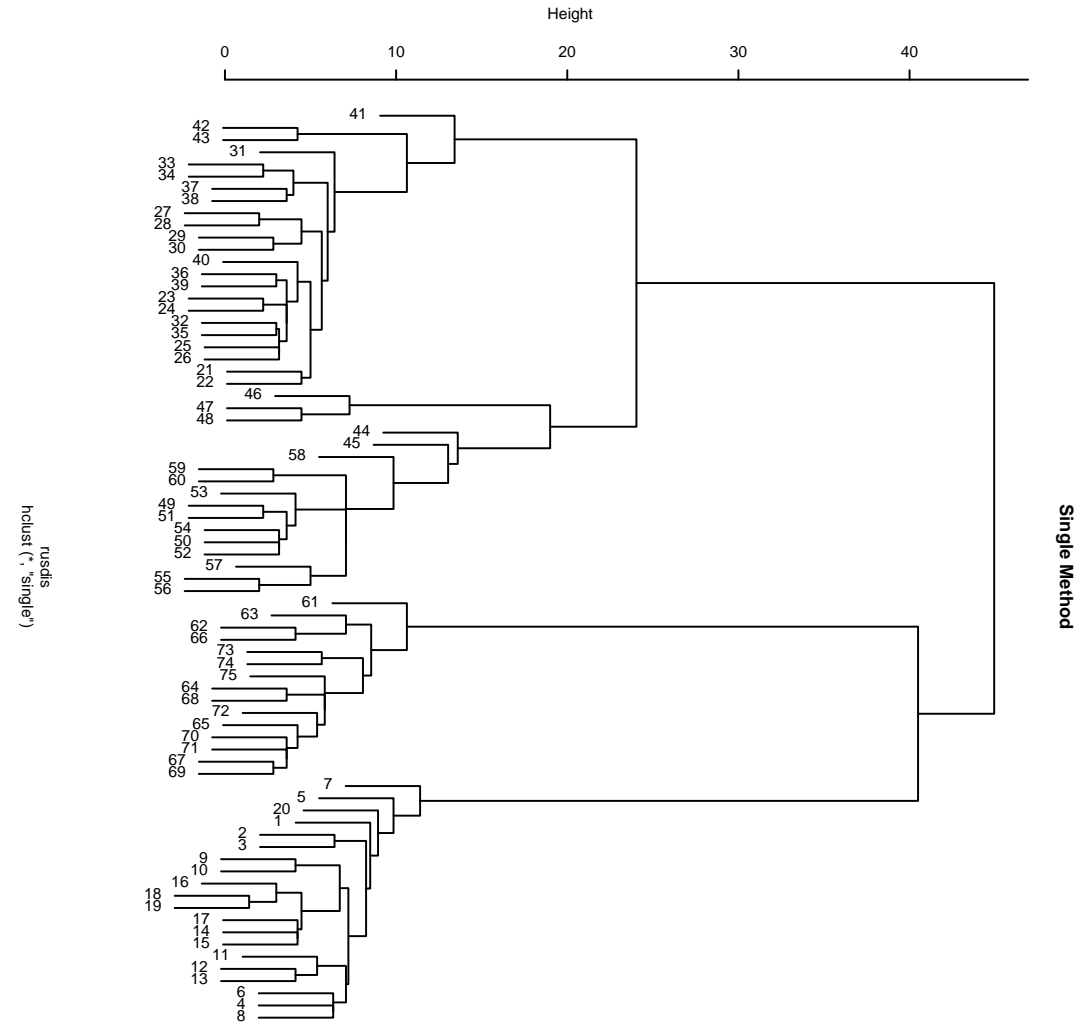
```
> rusdis<-dist(ruspini,method="euclidean")
> rushclus<-hclust(rusdis,method="complete")
> plclust(rushclus)
```

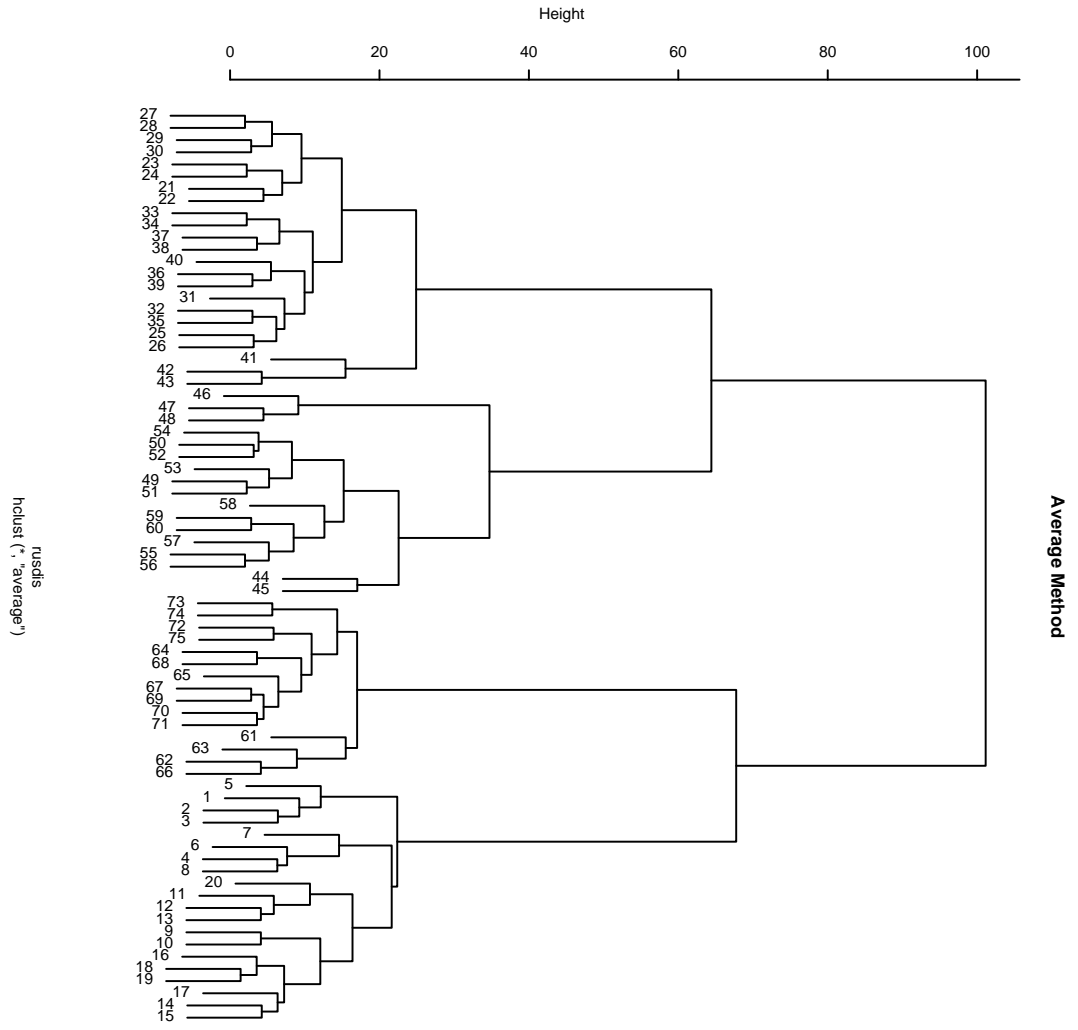


Del gráfico anterior se observa que el método ha identificado los conglomerados que visualmente se observaron en el gráfico.

Métodos utilizados para determinar el número de clusters

- *Método Completo o Vecino más alejado (Complete Linkage Method):*
La distancia entre dos conglomerados es la máxima de las distancias individuales entre puntos del cluster. Tiende a producir grupos alargados. El criterio es invariante ante transformaciones monótonas.
- *Método Simple o Vecino más cercano (Single Linkage Clustering):*
La distancia entre dos conglomerados es la mínima de las distancias individuales entre un punto de un cluster y un punto del otro cluster. Tiende a producir grupos esféricos. El criterio es invariante ante transformaciones monótonas.
- *Método del promedio (Average Method):*
La distancia entre los conglomerados es el promedio de las distancias entre los puntos de un cluster y los puntos del otro cluster. Es un método intermedio entre los dos anteriores. El método no es invariante ante transformaciones monótonas.
- *Método del centroide (Centroid Method):*
La distancia entre dos conglomerados es la distancia entre sus centroides.
- *Método de Ward ó Método de la Suma de Cuadrados:*
Los nuevos conglomerados se crean de tal manera de que se minimice la suma de cuadrados total de las distancias dentro de cada cluster.





La función *agnes* de la librería *cluster* de **R**

La función *agnes* de la librería *cluster* de **R** es una más reciente implementación del conglomerado jerárquico. Esta implementación sigue el método descrito por Kaufman y Rousseeuw (1990).

agnes y *hclust* utilizan la fórmula de Lance-Williams para calcular las distancias entre conglomerados.

Si C_1 y C_2 son dos conglomerados que se van a unir en un nuevo conglomerado, la distancia entre su unión y un nuevo conglomerado Q es:

$$D(C_1 \cup C_2, Q) = \alpha_1 * D(C_1, Q) + \alpha_2 * D(C_2, Q) + \beta * D(C_1, C_2) + \gamma * |D(C_1, Q) - D(C_2, Q)|$$

donde $\alpha_1, \alpha_2, \beta, \gamma$ son parámetros que deben especificarse.