Ejemplo: Datos de Arrestos en los E.U.A. (1973)
Estadísticas (por cada 100,000 habitantes)
de asaltos, asesinatos y violaciones.
La cuarta columna es el porcentaje de la
población que vive en áreas urbanas.

La función "USarrest.R" estandariza los datos, calcula una matriz de distancias $\mathbb{D}$ y con esta obtiene un agrupamiento el cual depende del método jerarquico aglomerativo que se le indique a la función de R "hclust()", posteriormente se le puede aplicar la función "fviz_dend()"(1), al objeto que contiene el agrupamiento, con el fin de producir un dendograma. La figura en el archivo "Figura D1.pdf" contiene dendogramas producidos con los métodos jerárquicos aglomerativos: "average Linkage" (1), "complete Linkage" (2) y "single Linkage" (3).

En algunos Libros de Análisis Multivariado se

(1) Paquete "factoextra" de R

propone juzgar la "calidad" de un agrupamiento calculando la correlación entre las distancias D y las "distancias cofenéticas" en el agrupamiento. Para dos individuos x_{G_i} y x_{G_j} su distancia cofenética se define como la altura (en el dendograma) del nodo en donde x_{G_i} y x_{G_j} fueron "clasificados en el mismo grupo". Si $\mathbf{T}$ denota la matriz de distancias cofeneticas para un agrupamiento, se define el "coeficiente de correlación cofenética"[1] c como

$$c = \frac{\sum_{i<j} (d_{ij} - \bar{d})(T_{ij} - \bar{T})}{\sqrt{\sum_{i<j} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i<j} (T_{ij} - \bar{T})^2}}$$

donde $\bar{d}$ es el promedio $\bar{d} = \frac{\sum_{i<j} d_{ij}}{\frac{n(n+1)}{2}}$

y $\bar{T}$ es el promedio $\bar{T} = \frac{\sum_{i<j} T_{ij}}{\frac{n(n+1)}{2}}$

Si c tiene valores "grandes" (cercanos a 1)

(1) Sokal y Rohlf (1962) The comparison of dendograms by objective methods. Taxon, 11, 33-40

entonces se considera que el agrupamiento propuesto por el dendograma es una "buena descripción de los datos". En la práctica un valor arriba de 0.75 se considera "grande" La función "copheneD.R" calcula los coeficientes de correlación cofenética para los distintos métodos de agrupamiento jerárquico aglomerativo que tiene la función "hclust( )". Al parecer el agrupamiento "más adecuado" para los datos de arrestos el el que se obtiene usando average linkage.

Se puede obtener un agrupamiento al cortar el dendograma a una altura determinada, o bien cuando se identifica la altura en que se forman un número de grupos K pre-determinado. Si se usa una altura ó un valor de K predeterminados, eso depende de la aplicación con la que se esta trabajando.

No obstante se puede encontrar un valor "optimo" de K (cuando no se tiene idea de este), usando una medida de "qué tan adecuado" resulta considerar a K como el número de grupos.

Esta medida se conoce como el "ancho promedio de la silueta"(2) del agropamiento

(2) Alan Julian Izenman "Modern Multivariate Statistical Techniques, Regression Classification and Manifold Learning". Springer Verlag pagnes 426-429.