**Departamento de Estadística e Investigación Operativa. Universidad de Sevilla**

# Análisis de Datos Multivariantes.

## Grado en Matemáticas. Doble Grado en Matemáticas y Estadística

## Hoja 5. Prácticas de Análisis de Conglomerados con R

**1.** Se considera el fichero leche.sav que contiene la composición de la leche materna para un conjunto de mamíferos.

i) Leer los datos y resumirlos numérica y gráficamente.

ii) Identificar pares de variables que presenten una correlación lineal positiva.

iii) Usando la función hclust, aplicar un análisis de conglomerados jerárquico aglomerativo a partir de la matriz de distancias euclídeas (Método completo).

iv) A partir de los resultados del apartado anterior, establecer una partición en 3 conglomerados. Calcular los centroides (vectores de medias de los conglomerados respectivos).

v) ¿Qué variables presentan mayor variabilidad entre los conglomerados? Representan gráficamente la partición efectuada sobre dichas variables.

vi) Repetir el apartado iii) con los métodos "Single" y "Average". ¿Cómo cambian las particiones de tamaño 3?

vii) Realizar la partición en 3 conglomerados con la función agnes.

**2.** Trabajando con el *data frame* flowers de la librería cluster:

i) Calcular el coeficiente general de disimilaridad de Gower como aparece en las transparencias.

ii) Realizar y comparar agrupaciones mediante las técnicas jerárquicas "Single" y "Complete".

iii) Calcular el coeficiente de aglomeración para ambas soluciones.

**3.** Acceder al conjunto de datos swiss.

i) Realizar un análisis de conglomerados de centros móviles (k=3) utilizando como centroides iniciales los resultantes de un análisis jerárquico (método "complete").

ii) Representar gráficamente la solución sobre las dos variables con mayor variabilidad entre los conglomerados.

iii) Repetir el análisis de centros móviles utilizando 15 soluciones iniciales diferentes.

**4.** Acceder a la imagen "*mandril.jpg*" mediante las librerías jpeg y pixmap.

i) Leer los datos en R. Generar una matriz de datos a partir de las componentes de los tres colores básicos RGB.

ii) Para valores de k en el conjunto{2, 5, 10} aplicar k-medias y reemplazar la imagen original por los centroides de las particiones obtenidas.

**5.** Acceder al *data frame* agriculture de la librería cluster.

i) Leer los datos, y representarlos gráficamente.

ii) Realizar una partición mediante el método de los k-medioides (k=2). Representar gráficamente la partición.

iii) ¿Se puede mejorar la solución para algún valor de k entre 3 y 7?.

**6.** Acceder al *data frame* ruspini de la librería cluster. Comprobar que para estos datos simulados la mejor partición con el método de los k-medoides corresponde a k=4.

**7.** Realizar un estudio similar al del problema anterior con los datos xclara usando el método *clara*.

**8.** Aplicar el método *fanny* sobre los datos agriculture.

**9.** Ajustar un modelo de mixturas de normales bivariantes sobre los datos faithful.

**10.** Ajustar un modelo de mixturas sobre los datos wreath de la librería mclust.