MAT-269: Sesión práctica de Análisis de Conglomerados

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Datos de mediciones corporales

```
> measure
    chest  waist  hips  gender
1      34     30    32    Male
2      37     32    37    Male
3      38     30    36    Male
4      36     33    39    Male
5      38     29    33    Male
6      43     32    38    Male
7      40     33    42    Male
8      38     30    40    Male
9      40     30    37    Male
10     41     32    39    Male
11     36     24    35  Female
12     36     25    37  Female
13     34     24    37  Female
14     33     22    34  Female
15     36     26    38  Female
16     37     26    37  Female
17     34     25    38  Female
18     36     26    37  Female
19     38     28    40  Female
20     35     23    35  Female
```

Datos de mediciones corporales

```
> dm <- dist(measure[,c("chest","waist","hips")]) # distancia Euclidiana
> dm
       1     2     3     4     5     6     7     8     9    10
2   6.16
3   5.66  2.45
4   7.87  2.45  4.69
5   4.24  5.10  3.16  7.48
6  11.00  6.08  5.74  7.14  7.68
7  12.04  5.92  7.00  5.00 10.05  5.10
8   8.94  3.74  4.00  3.74  7.07  5.74  4.12
9   7.81  3.61  2.24  5.39  4.58  3.74  5.83  3.61
10 10.10  4.47  4.69  5.10  7.35  2.24  3.32  3.74  3.00
11  7.00  8.31  6.40  9.85  5.74 11.05 12.08  8.06  7.48 10.25

...

20  7.68  9.43  7.68 10.82  7.00 12.41 13.19  9.11  8.83 11.53  ...
```

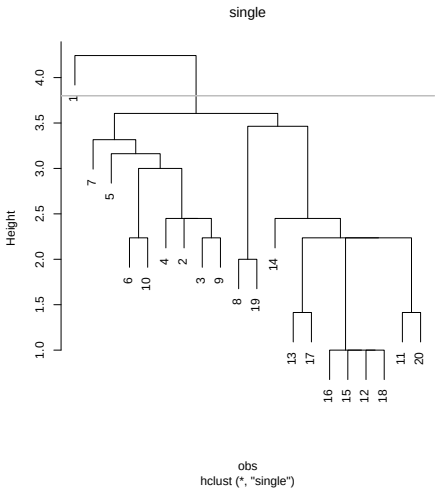
Opciones de `dist`: parámetro `method` con euclidean, maximum, manhattan, canberra, binary, minkowski.

Datos de mediciones corporales

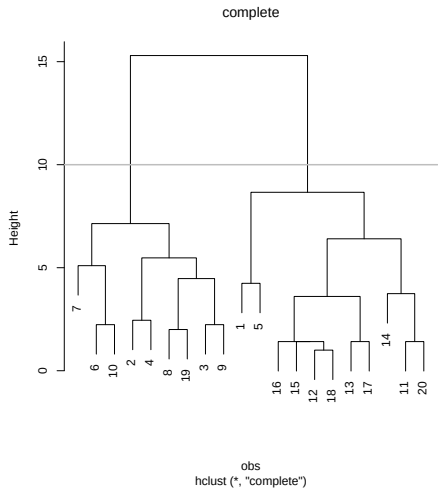
```
# cluster aglomerativo usando distintos metodos
> cs <- hclust(dm, method = "single")
> cc <- hclust(dm, method = "complete")
> ca <- hclust(dm, method = "average")

# dendogramas para cada metodo de enlace
> plot(cs, main = "single", font.main = 1, xlab = "obs")
> plot(cc, main = "complete", font.main = 1, xlab = "obs")
> plot(ca, main = "average", font.main = 1, xlab = "obs")
```

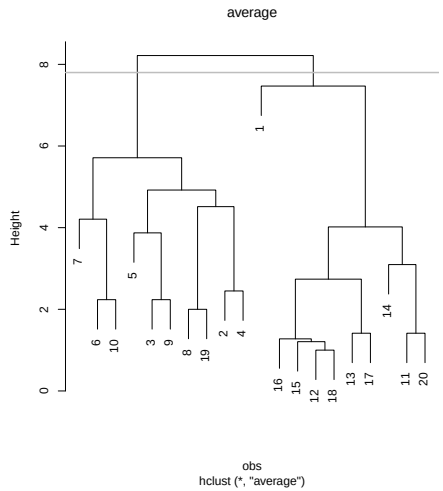
Datos de mediciones corporales



(a)



(b)



(c)

Datos de mediciones corporales

```
# corta el arbol generado por el dendograma
> lab.cs <- cutree(cs, h = 3.8)
> lab.cc <- cutree(cc, h = 10)
> lab.ca <- cutree(ca, h = 7.8)

# asignacion a cada cluster
> lab.cs
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2

> lab.cc
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1  2  2  2  1  2  2  2  2  2  1  1  1  1  1  1  1  2  1

> lab.ca
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1  2  2  2  2  2  2  2  2  2  1  1  1  1  1  1  1  2  1
```

Datos de mediciones corporales

```
# una variacion de los grupos
> lab <- cutree(cs, k = 1:4)
> lab
   1 2 3 4
1  1 1 1 1
2  1 2 2 2
3  1 2 2 2
4  1 2 2 2
5  1 2 2 2
6  1 2 2 2
7  1 2 2 2
8  1 2 3 3
9  1 2 2 2
10 1 2 2 2
11 1 2 3 4
12 1 2 3 4
13 1 2 3 4
14 1 2 3 4
15 1 2 3 4
16 1 2 3 4
17 1 2 3 4
18 1 2 3 4
19 1 2 3 3
20 1 2 3 4
```

Datos de esperanza de vida

```
> source("lifeexp.R") # descritos en Slide 17.
> life
                    m0  m25  m50  m75  w0  w25  w50  w75
Algeria             63   51   30   13  67   54   34   15
Cameroon            34   29   13    5  38   32   17    6
Madagascar          38   30   17    7  38   34   20    7
Mauritius           59   42   20    6  64   46   25    8
Reunion             56   38   18    7  62   46   25   10
Seychelles          62   44   24    7  69   50   28   14
South Africa(C)     50   39   20    7  55   43   23    8
South Africa(W)     65   44   22    7  72   50   27    9
Tunisia             56   46   24   11  63   54   33   19
Canada              69   47   24    8  75   53   29   10

...

Argentina           65   46   24    9  71   51   28   10
Chile               59   43   23   10  66   49   27   12
Columbia            58   44   24    9  62   47   25   10
Ecuador             57   46   28    9  60   49   28   11
```

Datos de esperanza de vida

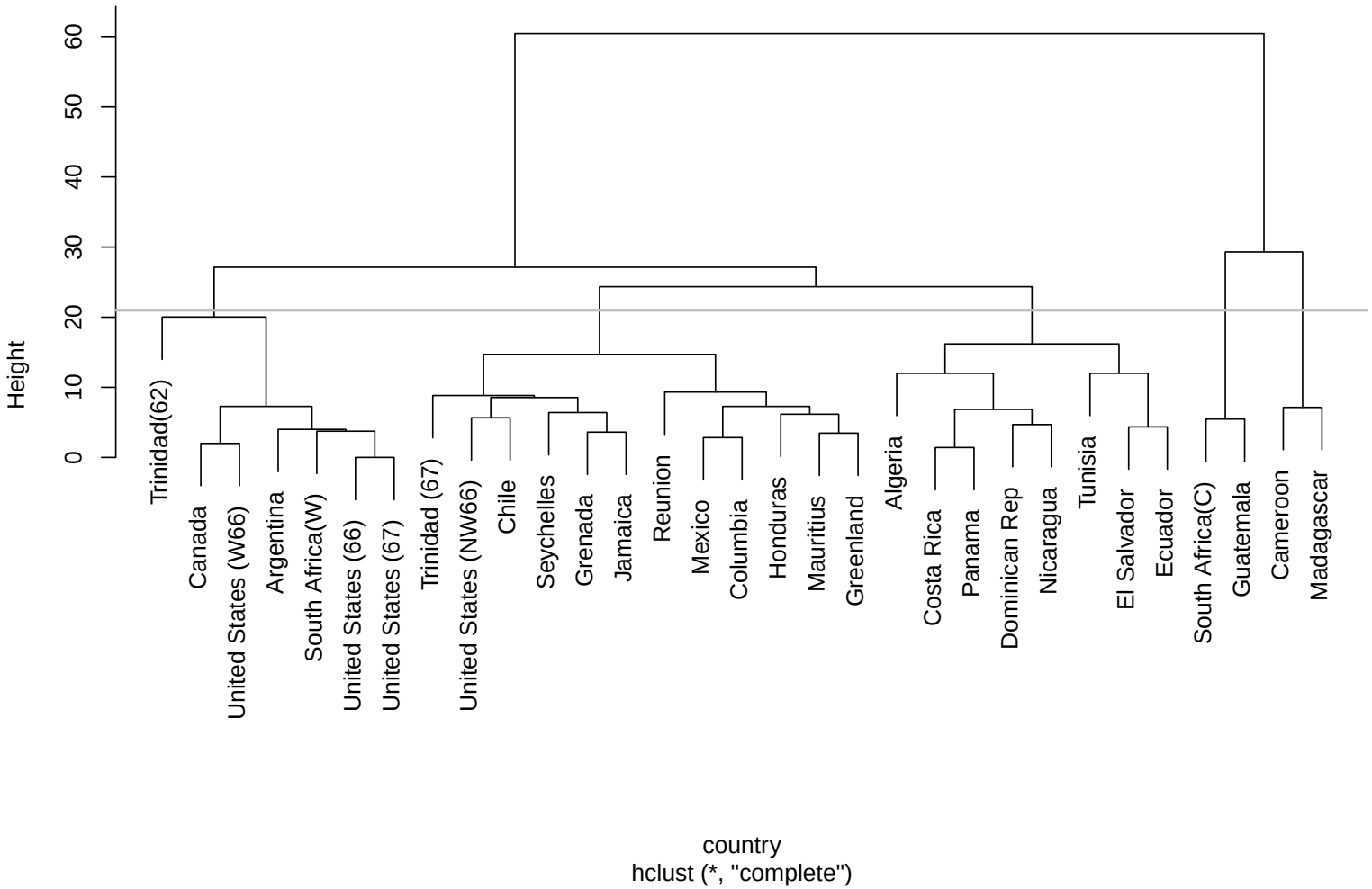
```
# analisis de agrupamientos
> clust.life <- hclust(dist(life), method = "complete")
> groups <- cutree(clust.life, h = 21)
> country.clus <- lapply(1:5, function(nc) row.names(life)[groups==nc])
> country.mean <- lapply(1:5, function(nc) apply(life[groups==nc,],
+                          2, mean))

# output
> clust.life

Call:
hclust(d = dist(life), method = "complete")

Cluster method   : complete
Distance         : euclidean
Number of objects: 31
```

Datos de esperanza de vida



Datos de esperanza de vida

```
# informacion de grupos
> groups
             Algeria              Cameroon            Madagascar
                   1                     2                     2
            Mauritius               Reunion            Seychelles
                   3                     3                     3
     South Africa(C)      South Africa(W)               Tunisia
                   4                     5                     1
              Canada           Costa Rica        Dominican Rep
                   5                     1                     1
         El Salvador             Greenland               Grenada
                   1                     3                     3
           Guatemala              Honduras               Jamaica
                   4                     3                     3
              Mexico             Nicaragua                Panama
                   3                     1                     1
        Trinidad(62)         Trinidad (67)  United States (66)
                   5                     3                     5
United States (NW66)  United States (W66)  United States (67)
                   3                     5                     5
           Argentina                 Chile              Columbia
                   5                     3                     3
             Ecuador
                   1
```

Datos de esperanza de vida

```
> country.clus
[[1]]
[1] "Algeria"        "Tunisia"        "Costa Rica"     "Dominican Rep"
[5] "El Salvador"    "Nicaragua"      "Panama"         "Ecuador"

[[2]]
[1] "Cameroon"    "Madagascar"

[[3]]
 [1] "Mauritius"             "Reunion"               "Seychelles"
 [4] "Greenland"             "Grenada"               "Honduras"
 [7] "Jamaica"               "Mexico"                "Trinidad (67)"
[10] "United States (NW66)" "Chile"                 "Columbia"

[[4]]
[1] "South Africa(C)" "Guatemala"

[[5]]
[1] "South Africa(W)"     "Canada"                "Trinidad(62)"
[4] "United States (66)"  "United States (W66)" "United States (67)"
[7] "Argentina"
```

Datos de esperanza de vida

```
> country.mean
[[1]]
    m0     m25     m50     m75      w0     w25     w50     w75
61.375 47.625 26.875 10.750 65.000 50.750 29.250 12.625

[[2]]
  m0   m25   m50   m75    w0   w25   w50   w75
36.0 29.5 15.0   6.0 38.0 33.0 18.5   6.5

[[3]]
    m0     m25     m50     m75      w0     w25     w50     w75
60.083 42.750 22.000   7.583 64.916 46.833 25.333   9.666

[[4]]
  m0   m25   m50   m75    w0   w25   w50   w75
49.5 39.5 21.0   8.0 53.0 42.0 23.0   8.0

[[5]]
    m0     m25     m50     m75      w0     w25     w50     w75
66.429 48.000 22.857   7.857 72.714 50.714 27.714   9.714
```

Alfarería Romano-Británica

Composición química de 45 especímenes de alfarería Romano-Británica, determinada por espectrofotometría por absorción atómica, para nueve óxidos (Tubb et al., 1980).[1]

```
> source("pottery.R")
> pottery
      AL2O3 FE2O3   MGO     CAO   NA2O  K2O TIO2     MNO  BAO
 [1,]  1.76  1.11 0.299 0.4593 0.500 1.02 1.29 0.4753 1.07
 [2,]  1.58  0.85 0.246 0.4884 0.500 0.97 1.27 0.4136 1.29
 [3,]  1.70  0.89 0.272 0.4477 0.500 0.98 1.26 0.5370 1.00
 [4,]  1.58  0.85 0.233 0.4419 0.500 0.97 1.28 0.3889 1.36
 [5,]  1.66  0.84 0.273 0.5349 0.538 0.99 1.19 0.3765 1.36
 [6,]  1.76  0.87 0.307 0.5058 0.312 1.04 1.26 0.4444 1.21
 [7,]  1.54  0.82 0.270 1.0058 0.413 1.02 1.22 0.4074 1.36
 [8,]  1.68  0.86 0.307 0.5814 0.350 1.07 1.23 0.4444 1.21
 [9,]  1.48  0.83 0.242 0.4128 0.475 1.04 1.19 0.3827 1.21
[10,]  1.36  0.80 0.249 0.4419 0.413 0.97 1.17 0.3395 0.86
[11,]  1.28  0.68 0.224 0.3837 0.163 0.72 0.96 0.2099 0.86

...

[43,]  1.56  0.11 0.079 0.0058 0.063 0.56 1.17 0.0247 0.93
[44,]  1.38  0.32 0.100 0.0174 0.063 0.68 1.72 0.0185 1.07
[45,]  1.79  0.19 0.090 0.0581 0.037 0.56 1.33 0.0432 1.29
```

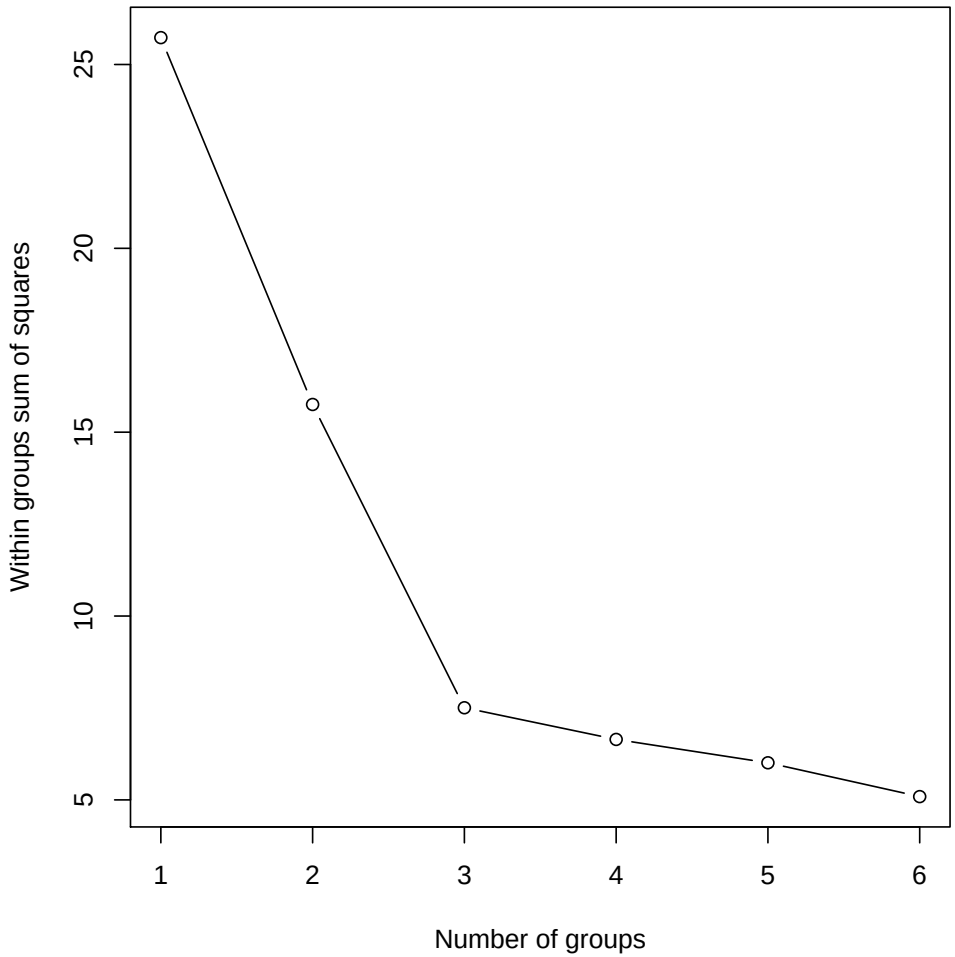
[1] Archaeometry **22**, 153-171.

Alfarería Romano-Británica

```
# permite identificar el numero de grupos
> n <- nrow(pottery)
> wss <- rep(0, 6)
> wss[1] <- (n - 1) * sum(apply(pottery, 2, var))
> for (i in 2:6)
+   wss[i] <- sum(kmeans(pottery, centers = i)$withinss)
> plot(1:6, wss, type = "b", xlab = "Number of groups",
+   ylab = "Within groups sum of squares")

# output
> wss
[1] 25.731238 15.754439  7.505893  6.644948  6.009360  5.089980
```

Alfarería Romano-Británica



Alfarería Romano-Británica

```
# K-means con 3 grupos
> pottery.kmean <- kmeans(pottery, centers = 3)
> pottery.kmean
K-means clustering with 3 clusters of sizes 5, 5, 35

Cluster means:
     AL2O3     FE2O3         MGO         CAO       NA2O        K2O      TIO2
1 1.654206 0.2362791 0.09850746 0.02558140 0.0650000 0.6503185 1.548718
2 1.663551 0.1386047 0.09253731 0.01976744 0.0625000 0.6369427 1.066667
3 1.413618 0.8070100 0.45023454 0.37740864 0.3721429 1.1263876 1.071429
         MNO      BAO
1 0.02345679 1.214286
2 0.01604938 1.071429
3 0.55396825 1.189796

Clustering vector:
 [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[34] 3 3 2 1 2 1 1 2 2 2 1 1

Within cluster sum of squares by cluster:
[1]  0.5619463  0.2479702 13.6304007
 (between_SS / total_SS =  43.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Alfarería Romano-Británica

```
# informacion desde el horno en que la ceramica fue hallada
> kiln <- c(rep(1,21),rep(2,12),rep(3,2),rep(4,5),rep(5,5))
> kiln
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
[31] 2 2 2 3 3 4 4 4 4 4 5 5 5 5 5

# tabla de contingencia
> table(kiln, pottery.kmean$cluster)

kiln  1  2  3
   1  0  0 21
   2  0  0 12
   3  0  0  2
   4  3  2  0
   5  2  3  0

# validacion?
> dp <- dist(scale(pottery, center = FALSE))
> library("lattice")
> trellis.par.set(canonical.theme(color = FALSE))
> levelplot(as.matrix(dp), xlab = "Pot number", ylab = "Pot number")
```

Alfarería Romano-Británica