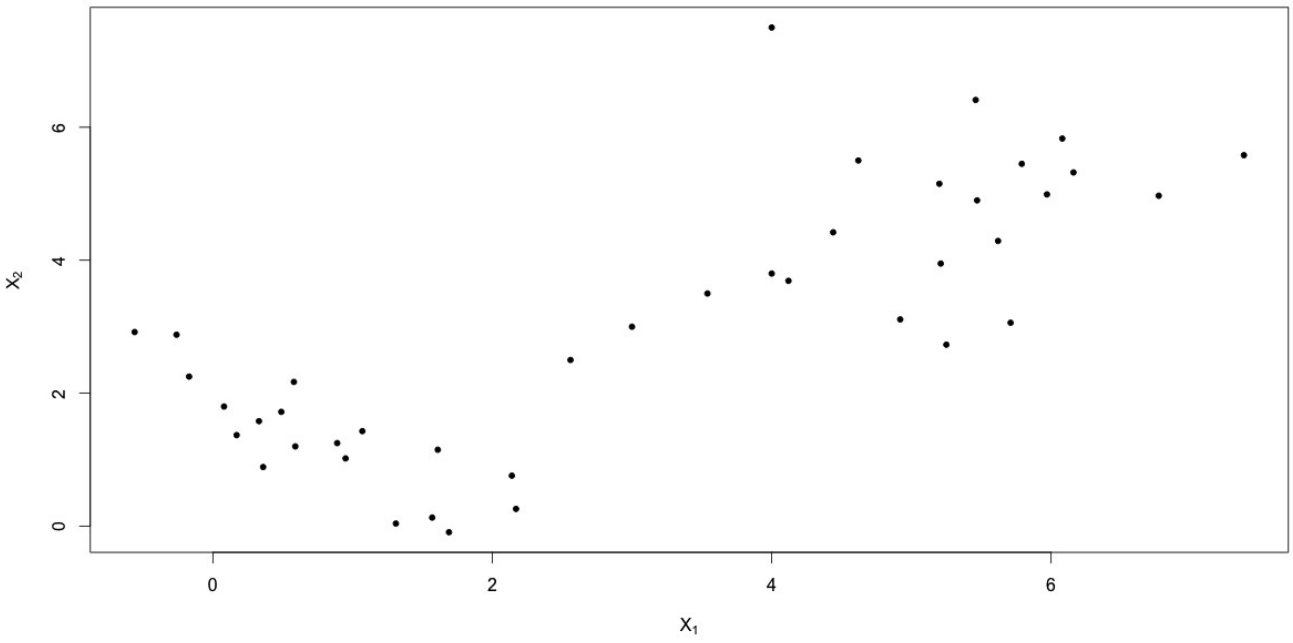
Análise de agrupamentos por métodos hierárquicos aglomerativos

```
## Exemplo 1 (p = 2)

dados <- read.table("dadosex1.txt")
cat("\n n =", n <- nrow(dados))
```

```
n = 41
```

```
plot(dados, pch = 20, xlab = expression(X[1]), ylab = expression(X[2]))
```



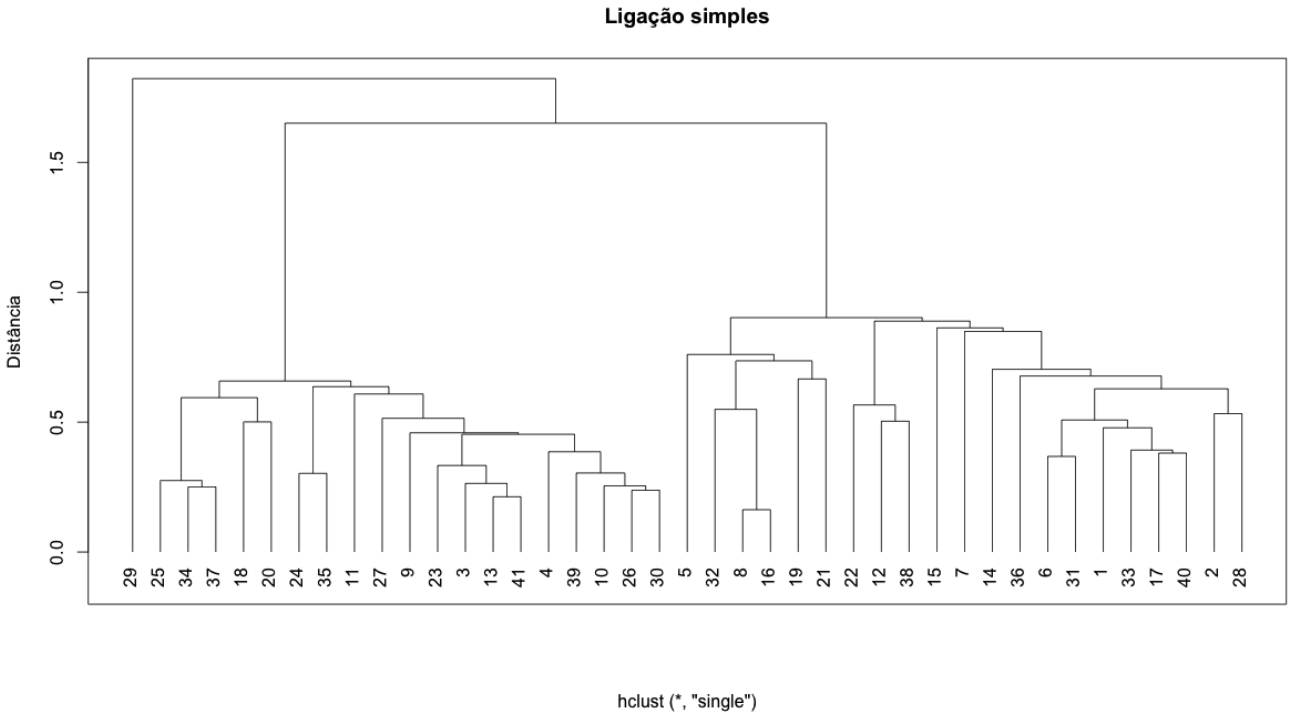

```
## Distância euclidiana (default: method = "euclidean")
distancia <- dist(dados)

## Ligação simples
mls <- hclust(distancia, method = "single")
cat("\n Correlação cofenética = ", cor(distancia, cophenetic(mls)))
```

```
Correlação cofenética =  0.8349257
```

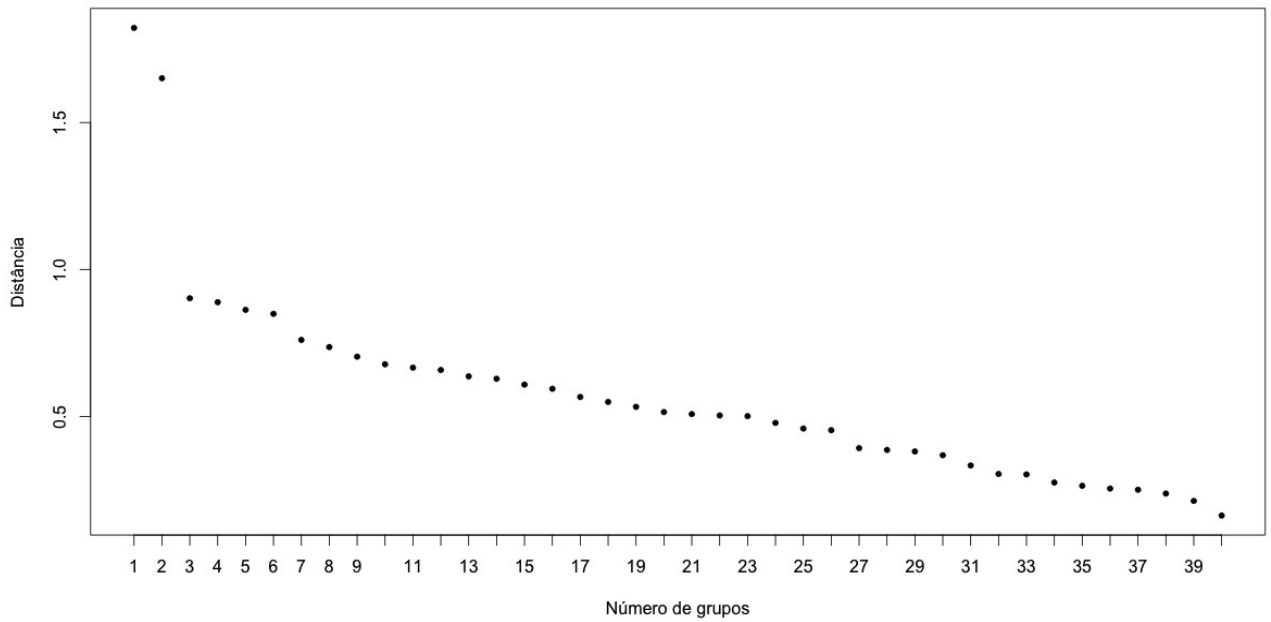
```
# Componentes de mls
names(mls)
```

```
[1] "merge"       "height"      "order"       "labels"
[5] "method"      "call"        "dist.method"
```

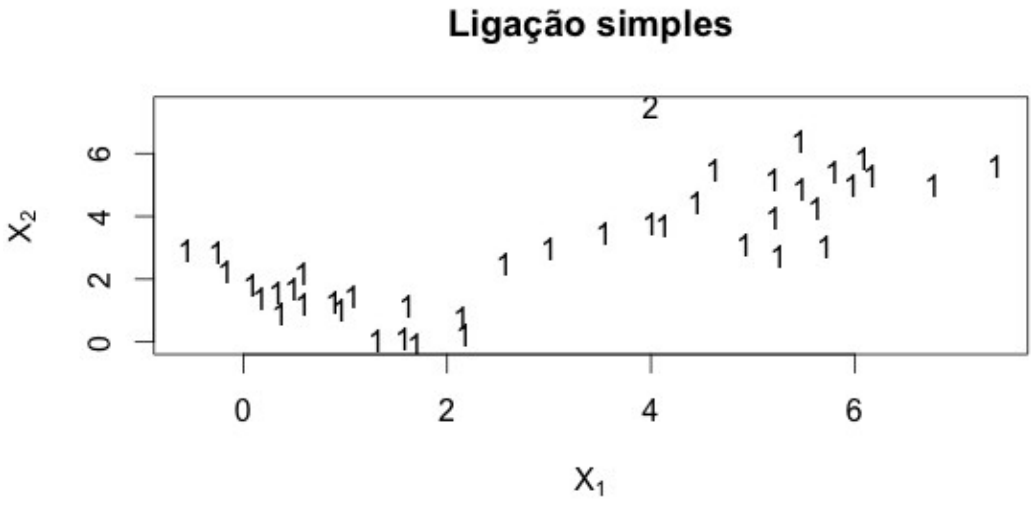
```
# Dendrograma
plot(mls, xlab ="", ylab = "Distância", main = "Ligação simples", hang = -1)
box()
```



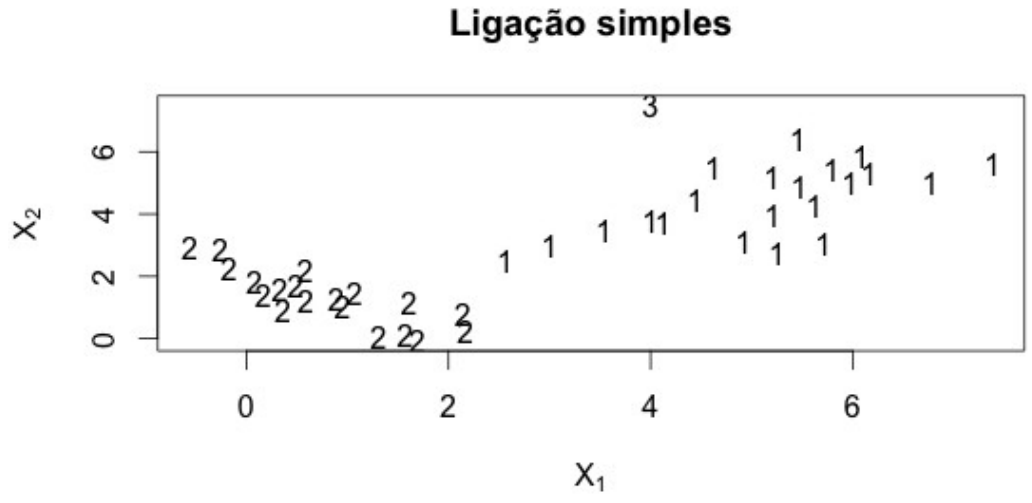

```
# Distâncias
plot((n - 1):1, mls$height, pch = 20, xlab = "Número de grupos",
     ylab = "Distância", main = "", axes = FALSE)
axis(1, 1:(n-1))
axis(2)
box()
```



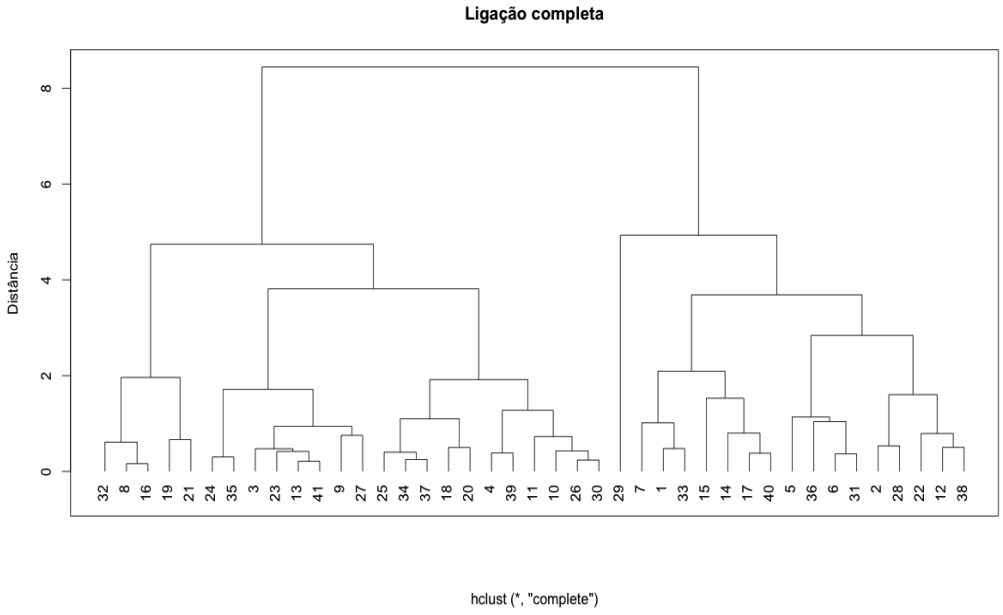
```
# Solução com dois grupos
k <- 2
grupls <- cutree(mls, k = k)
plot(dados, pch = as.character(1:k)[grupls], xlab = expression(X[1]),
     ylab = expression(X[2]), main = "Ligação simples")
```



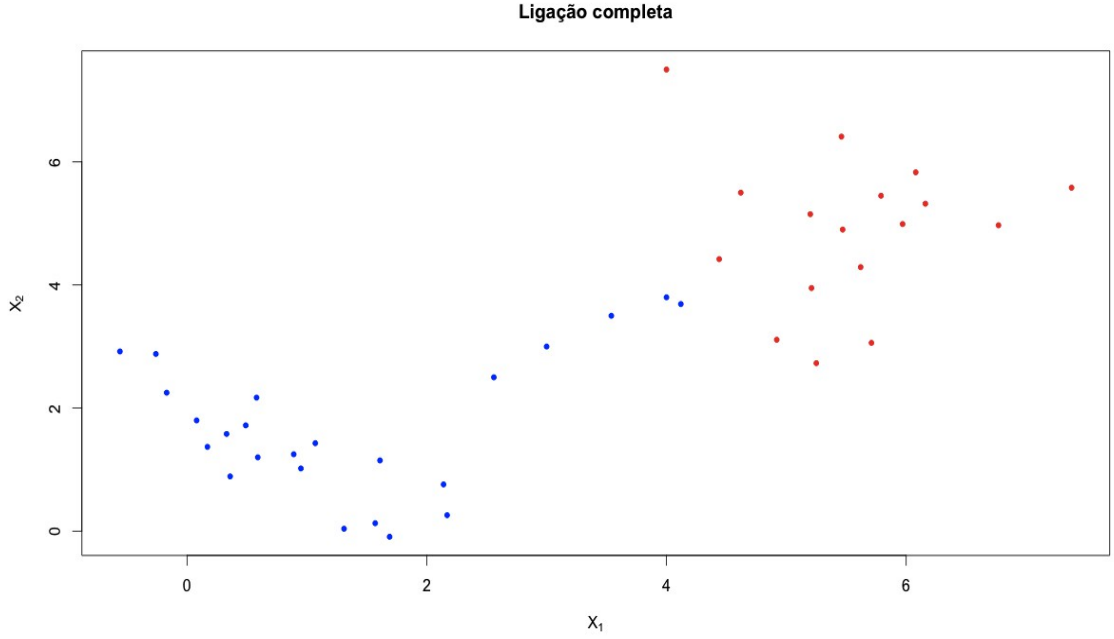

```
# Solução com  três grupos
k <- 3
grupls <- cutree(mls, k = k)
plot(dados, pch = as.character(1:k)[grupls], xlab = expression(X[1]),
     ylab = expression(X[2]),  main = "Ligação simples")
```




```
## Ligação completa
mlc <- hclust(distancia, method = "complete")
cat("\n Correlação cofenética = ", cor(distancia,cophenetic(mlc)),"\n")

Correlação cofenética =  0.8118873

plot(mlc, xlab ="", ylab = "Distância", main = "Ligação completa", hang = -1)
box()
```



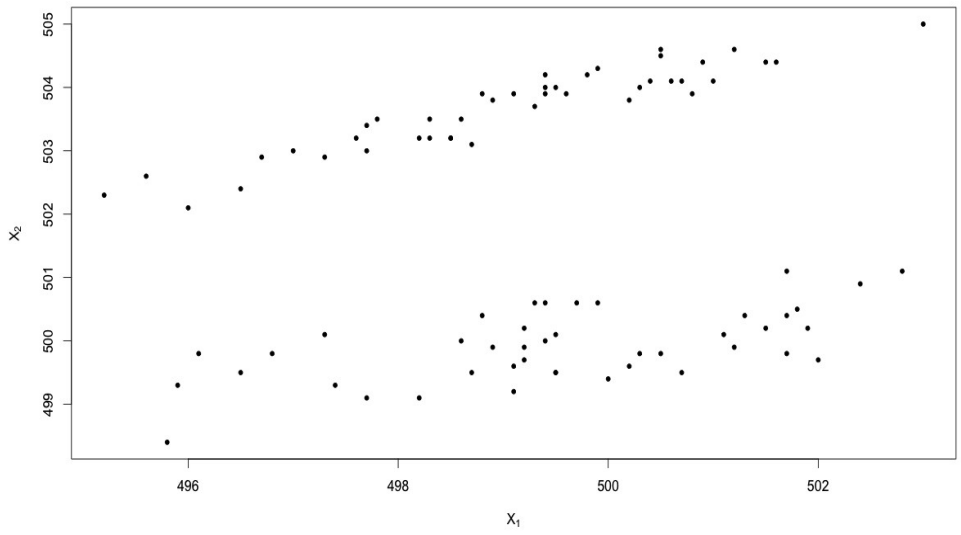

```
# Solução com dois grupos
gruplc <- cutree(mlc, k = 2)
plot(dados, pch = 20, xlab = expression(X[1]), ylab = expression(X[2]),
     col = c("red","blue")[gruplc], main = "Ligação completa")
```



Nota 1. Apresente as soluções obtidas com os métodos de ligação média e de Ward.

```
## Exemplo 2 (p = 2)
dados <- read.table("dadosex2.txt")
cat("\n n =", n <- nrow(dados))
n = 86
```

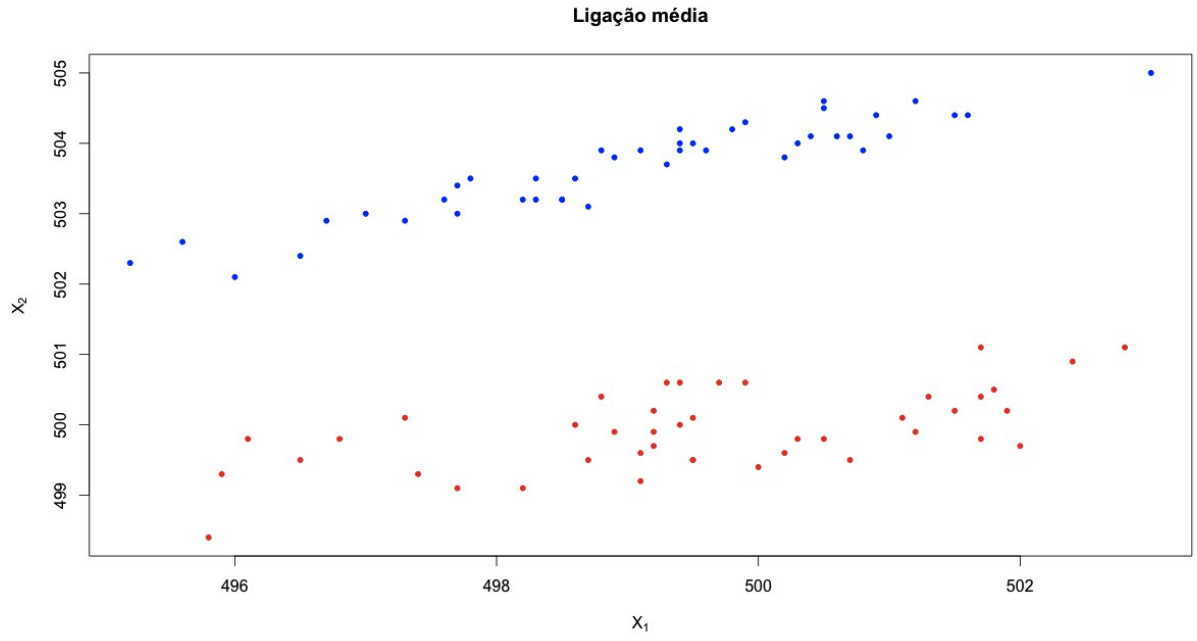
```
plot(dados, pch = 20, xlab = expression(X[1]), ylab = expression(X[2]))
```




```
## Distância euclidiana (default: method = "euclidean")
distancia <- dist(dados)

## Ligação média e solução com dois grupos
mlm <- hclust(distancia, method = "average")

cat("\n Correlação cofenética = ", cor(distancia, cophenetic(mlm)))
```

```
Correlação cofenética =  0.8173515
```

```
gruplm <- cutree(mlm, k = 2)
plot(dados, pch = 20, xlab = expression(X[1]), ylab = expression(X[2]),
     col = c("red","blue")[gruplm], main = "Ligação média")
```



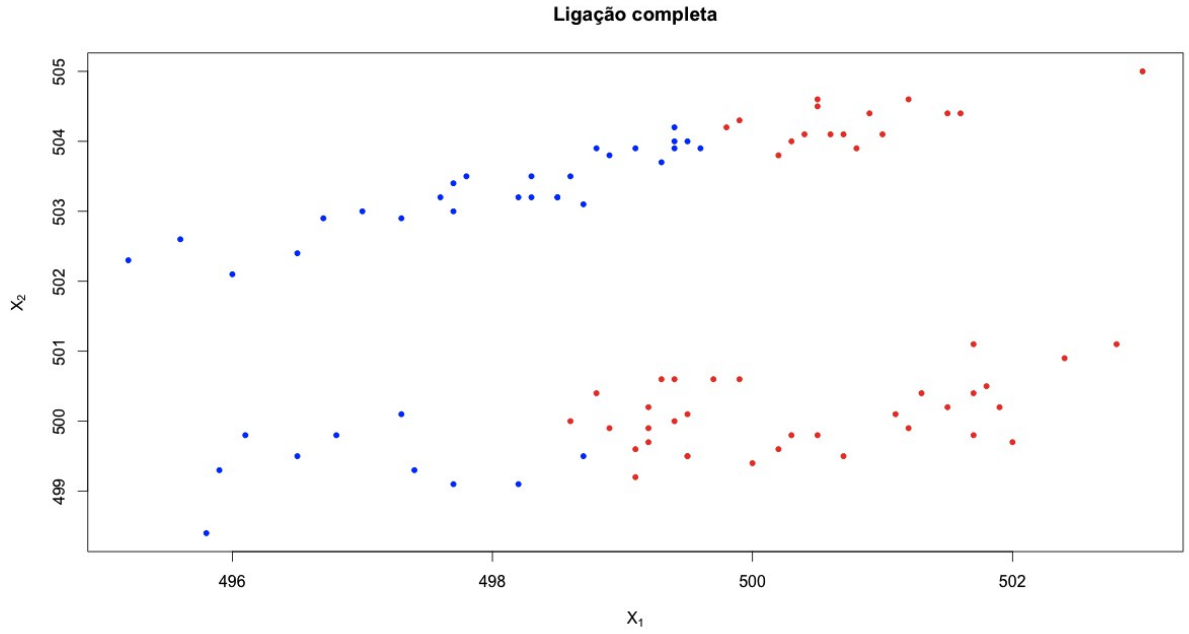
```
## Ligação completa e solução com dois grupos
mlc <- hclust(distancia, method = "complete")
cat("\n Correlação cofenética = ", cor(distancia, cophenetic(mlc)))
```

```
Correlação cofenética =  0.651931
```

```
gruplc <- cutree(mlc, k = 2)
plot(dados, pch = 20, xlab = expression(X[1]), ylab = expression(X[2]),
     col = c("red","blue")[gruplc], main = "Ligação completa")
```

Nota 2. Apresente as soluções obtidas com os métodos de ligação simples e de Ward.

Nota 3. Apresente as soluções obtidas com a distância de quarteirão (`method = "manhattan"` na função `dist`).




```
## Exemplo 3
dados <- read.table("dadosex3.txt", header = TRUE)
```

Dados do conteúdo de nove compostos químicos em 45 peças cerâmicas.

```
cat("\n n =", n <- nrow(dados), ", p =", ncol(dados))
```

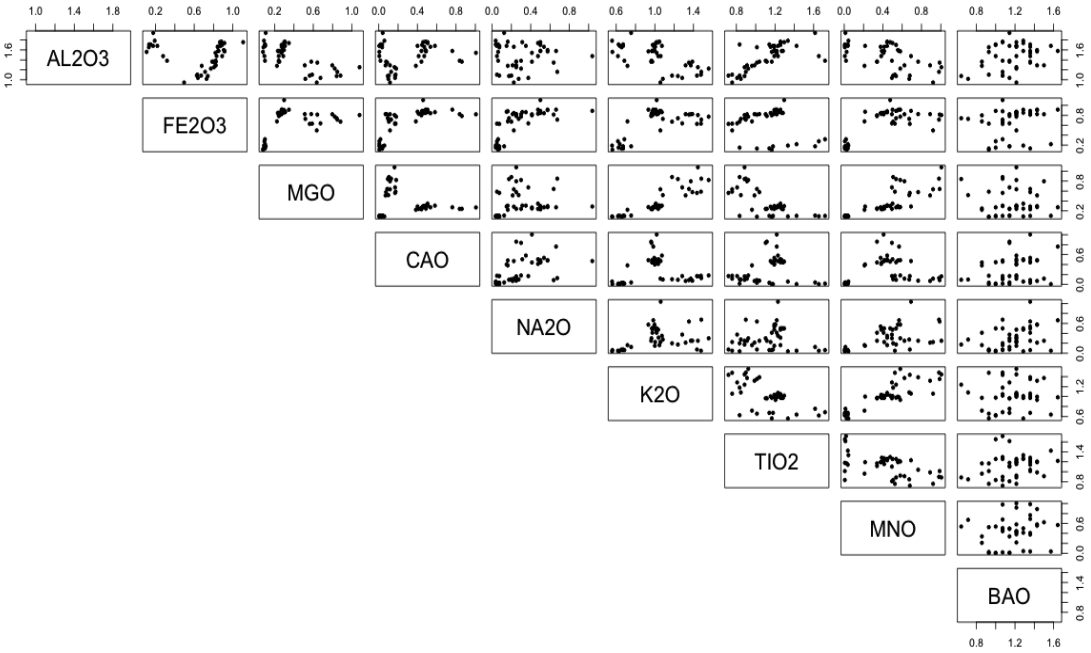
```
n = 45 , p = 9
```

```
summary(dados)
```

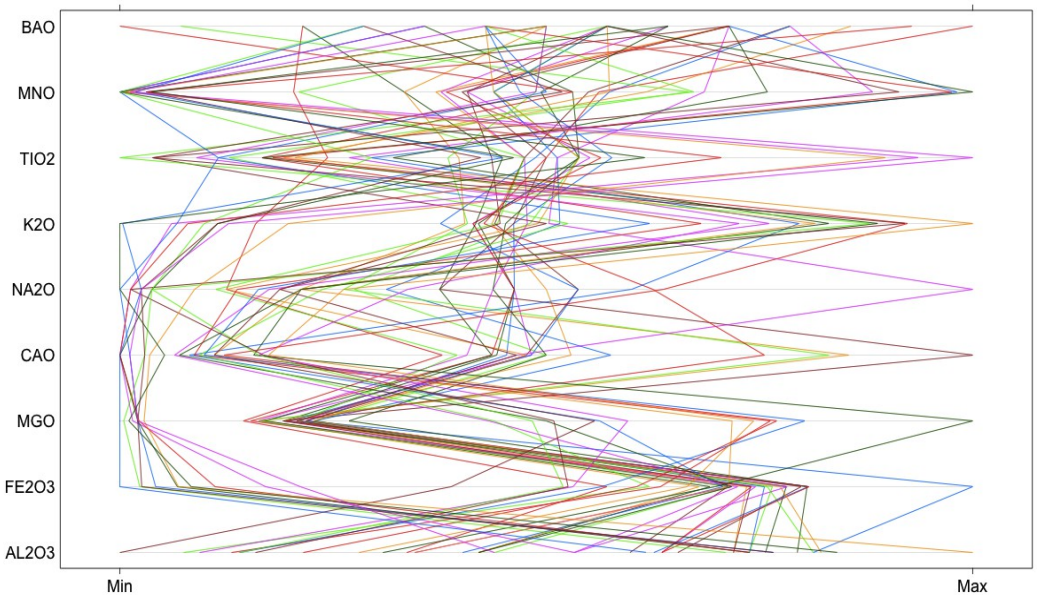
```
     AL2O3           FE2O3             MGO              CAO         
 Min.   :0.9439   Min.   :0.1070   Min.   :0.0791   Min.   :0.005814  
 1st Qu.:1.2897   1st Qu.:0.6267   1st Qu.:0.2328   1st Qu.:0.075581  
 Median :1.5421   Median :0.8047   Median :0.2866   Median :0.174419  
 Mean   :1.4681   Mean   :0.6693   Mean   :0.3714   Mean   :0.298579  
 3rd Qu.:1.6822   3rd Qu.:0.8523   3rd Qu.:0.5627   3rd Qu.:0.482558  
 Max.   :1.9439   Max.   :1.1070   Max.   :1.0791   Max.   :1.005814  
```

```
      NA2O              K2O              TIO2              MNO               BAO        
 Min.   :0.0375   Min.   :0.5573   Min.   :0.7179   Min.   :0.006173   Min.   :0.6429  
 1st Qu.:0.1250   1st Qu.:0.9331   1st Qu.:0.9231   1st Qu.:0.216049   1st Qu.:1.0714  
 Median :0.2500   Median :1.0032   Median :1.1667   Median :0.444444   Median :1.2143  
 Mean   :0.3036   Mean   :1.0191   Mean   :1.1239   Mean   :0.435254   Mean   :1.1794  
 3rd Qu.:0.4750   3rd Qu.:1.1783   3rd Qu.:1.2564   3rd Qu.:0.580247   3rd Qu.:1.3571  
 Max.   :1.0375   Max.   :1.5573   Max.   :1.7179   Max.   :1.006173   Max.   :1.6429  
```

```
# Gráficos de dispersão
pairs(dados, pch = 20, lower.panel = NULL, cex.labels = 2)
```




```
# Gráfico de coordenadas paralelas
library(lattice)
parallelplot(dados)
```



```
# Distância euclidiana
distancia <- dist(dados)

## Distância entre grupos
mls <- hclust(distancia, method = "single")
cat("\n Correlação cofenética = ", cor(distancia, cophenetic(mls)))
```

```
Correlação cofenética =  0.8717956
```

```
mlc <- hclust(distancia, method = "complete")
cat("\n Correlação cofenética = ", cor(distancia,cophenetic(mlc)))
```

```
Correlação cofenética =  0.8894245
```

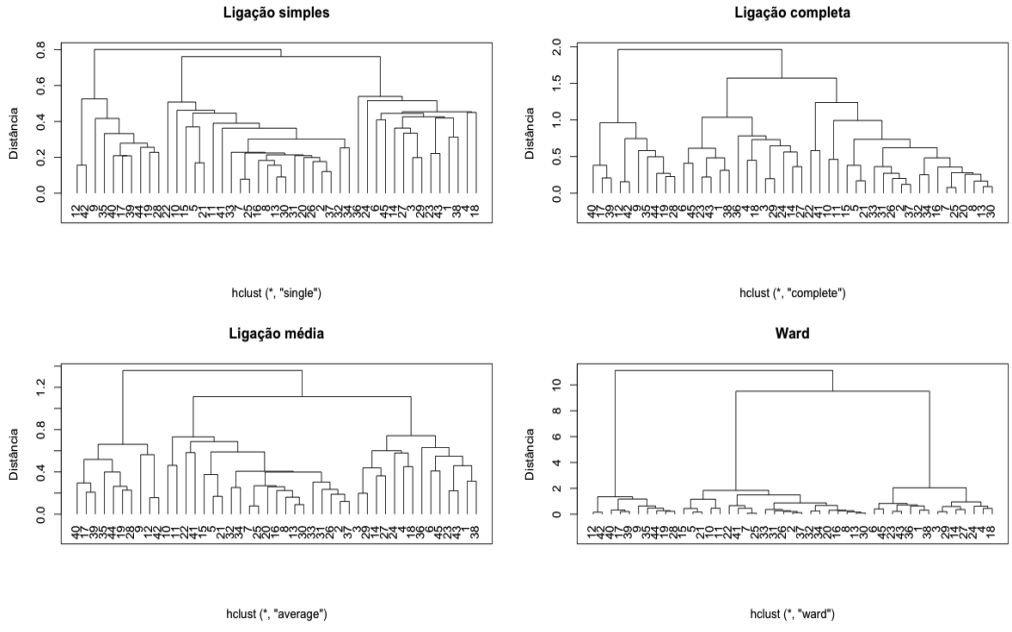
```
mla <- hclust(distancia, method = "average")
cat("\n Correlação cofenética = ", cor(distancia, cophenetic(mla)))
```

```
Correlação cofenética =  0.8968238
```

```
mlw <- hclust(distancia, method = "ward")
cat("\n Correlação cofenética = ", cor(distancia, cophenetic(mlw)))
```

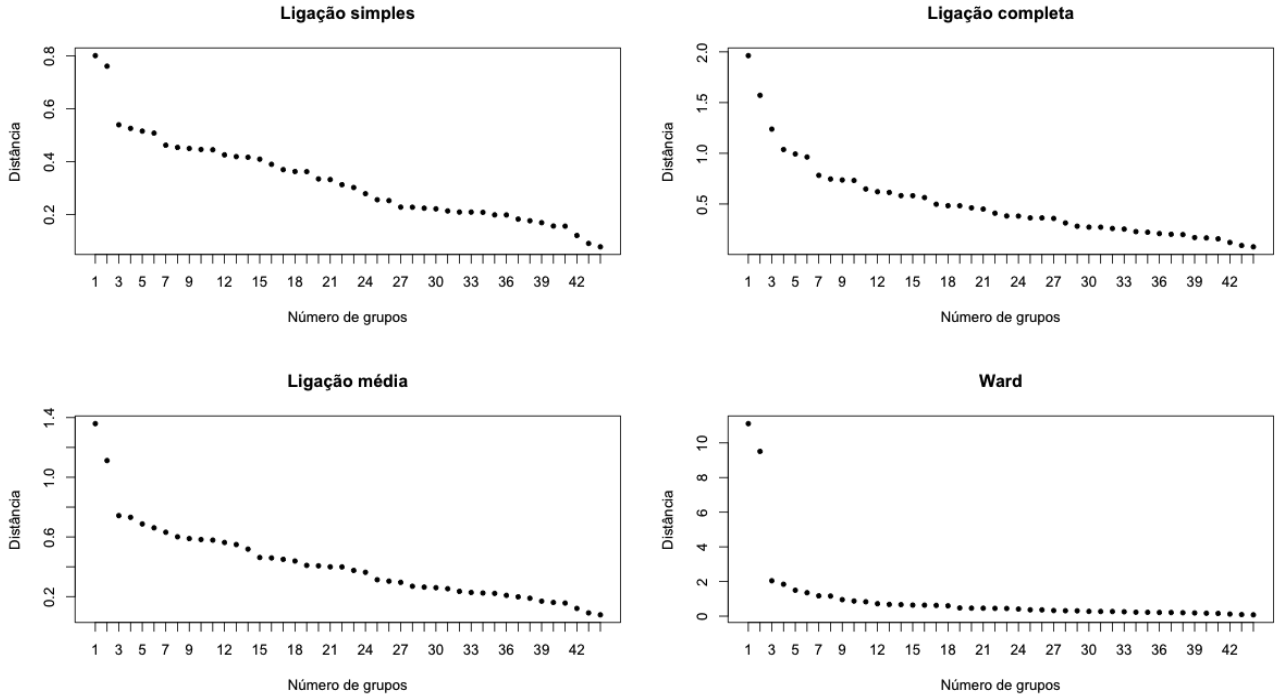
```
Correlação cofenética =  0.8654336
```

```
# Dendrogramas
par(mfrow = c(2, 2))
plot(mls, xlab ="", ylab = "Distância", main = "Ligação simples", hang = -1)
box()
plot(mlc, xlab ="", ylab = "Distância", main = "Ligação completa", hang = -1)
box()
plot(mla, xlab ="", ylab = "Distância", main = "Ligação média", hang = -1)
box()
plot(mlw, xlab ="", ylab = "Distância", main = "Ward", hang = -1)
box()
```



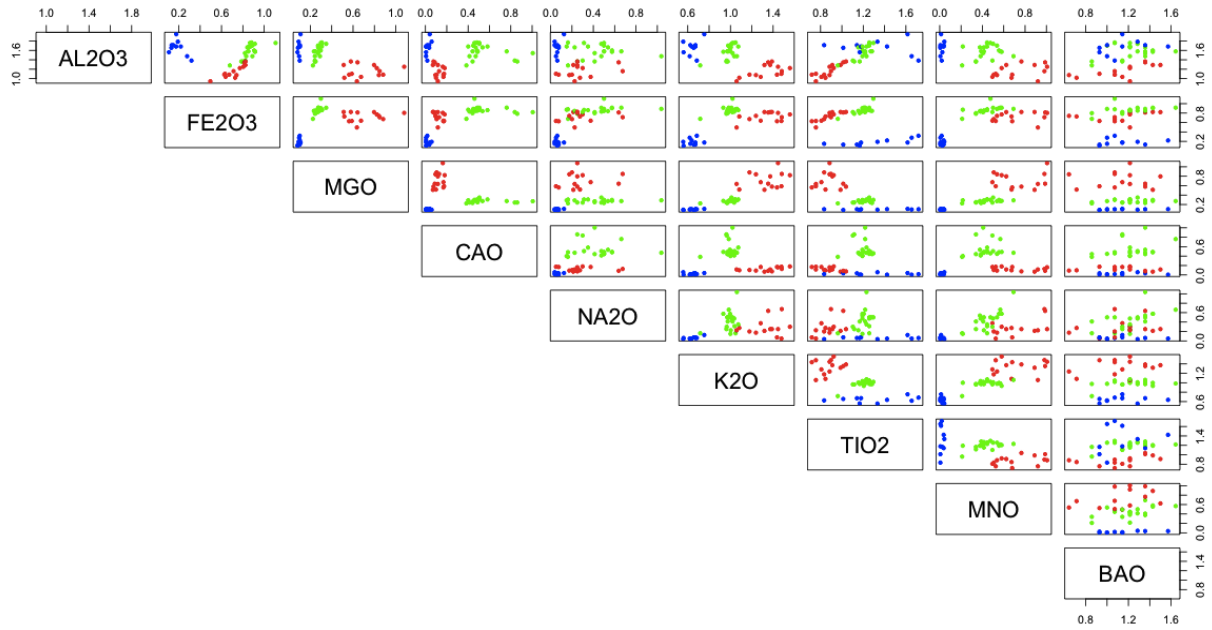
```
# Distâncias
par(mfrow = c(2, 2))
plot((n - 1):1, mls$height, pch = 20,
xlab = "Número de grupos",
     ylab = "Distância", main =
"Ligação simples", axes = FALSE)
axis(1, 1:(n - 1))
axis(2)
box()

plot((n - 1):1, mlc$height, pch = 20,
xlab = "Número de grupos",
     ylab = "Distância", main =
"Ligação completa", axes = FALSE)
axis(1, 1:(n - 1))
axis(2)
box()
```

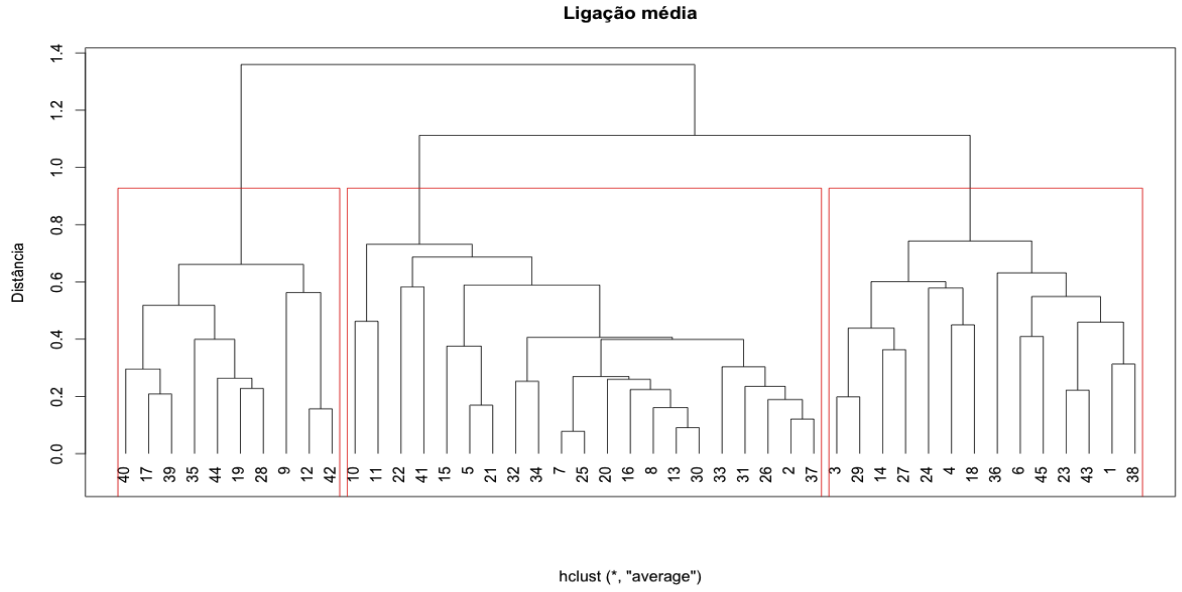
```
plot((n - 1):1, mla$height, pch = 20,
xlab = "Número de grupos",
     ylab = "Distância", main = "Ligação
média", axes = FALSE)
axis(1, 1:(n - 1))
axis(2)
box()

plot((n - 1):1, mlw$height, pch = 20,
xlab = "Número de grupos",
     ylab = "Distância", main = "Ward",
axes = FALSE)
axis(1, 1:(n - 1))
axis(2)
box()
```



Os métodos de ligação média e Ward apontam a formação de três grupos.

```
# Solução com três grupos, ligação média
grupla <- cutree(mla, k = 3)
cores <- rainbow(3)
pairs(dados, pch = 20, col = cores[grupla], main = "", lower.panel = NULL,
      cex.labels = 2)

plot(mla, xlab ="", ylab = "Distância", main = "Ligação média", hang = -1)
box()
rect.hclust(mla, k = 3)
```



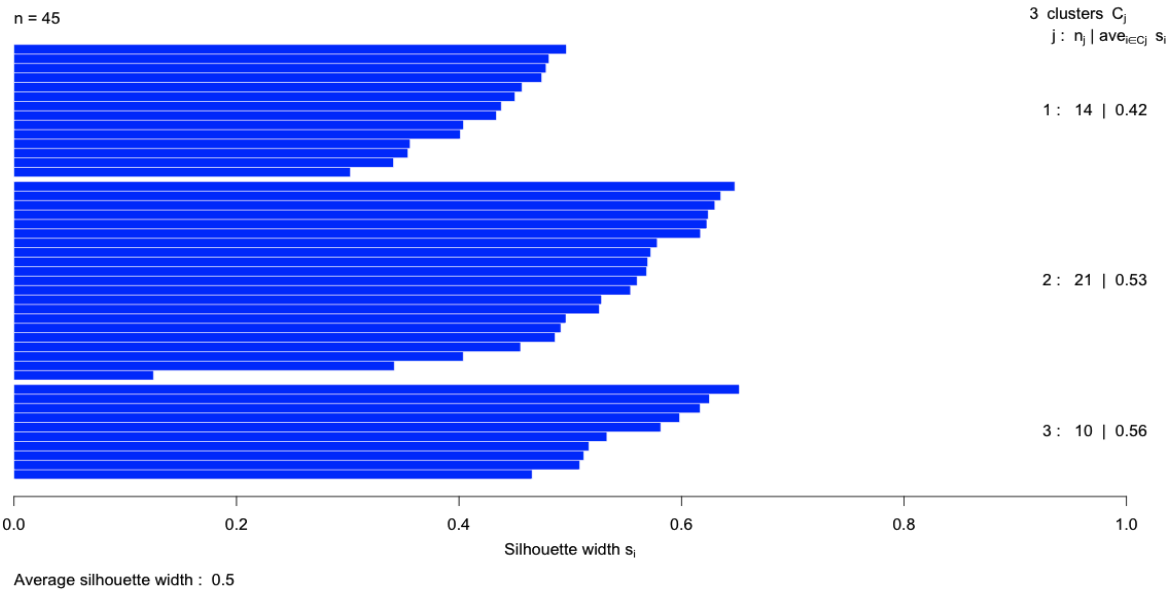



```
# Observações em cada grupo
table(grupla)

grupla
1  2  3
14 21 10

nomes <- paste("O", 1:n, sep = "")
for (j in 1:ng) {
   cat("\n Obs. no grupo ", j, ":", nomes[grupla == j])
}
```

```
 Obs. no grupo  1 : O1 O3 O4 O6 O14 O18 O23 O24 O27 O29 O36 O38 O43 O45
 Obs. no grupo  2 : O2 O5 O7 O8 O10 O11 O13 O15 O16 O20 O21 O22 O25 O26 O30
                    O31 O32 O33 O34 O37 O41
 Obs. no grupo  3 : O9 O12 O17 O19 O28 O35 O39 O40 O42 O44
```

Nota 4. Nos três exemplos obtenha o número de grupos utilizando o critério pseudo *F*.

```
# Gráfico da silhueta
library(cluster)
smla <- silhouette(cutree(mla, k = 3), dist = distancia)
summary(smla)

Silhouette of 45 units in 3 clusters from
silhouette.default(x = cutree(mla, k = 3), dist = distancia) :
 Cluster sizes and average silhouette widths:
       14        21        10
0.4190934 0.5254417 0.5609320
Individual silhouette widths:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1253  0.4503  0.5084  0.5002  0.5782  0.6521

plot(smla, nmax.lab = n, main = "", col = "blue")
```



Nota 5. O objeto `smla` contém os valores da silhueta *s(i)* para cada observação, $i = 1, ..., n$.

Nota 6. Apresente o gráfico da silhueta para os exemplos 1 e 2.

Nota 7. Refaça os exemplos com a função `agnes` (*agglomerative nesting*) do pacote `cluster` no lugar da função `hclust`.

Nota 8. Refaça os exemplos aplicando os métodos às variáveis padronizadas (função `scale` do pacote `base`).

Nota 9. Procure refazer os exemplos utilizando outros pacotes estatísticos (SAS, SPSS, Minitab e Statistica, por exemplo).