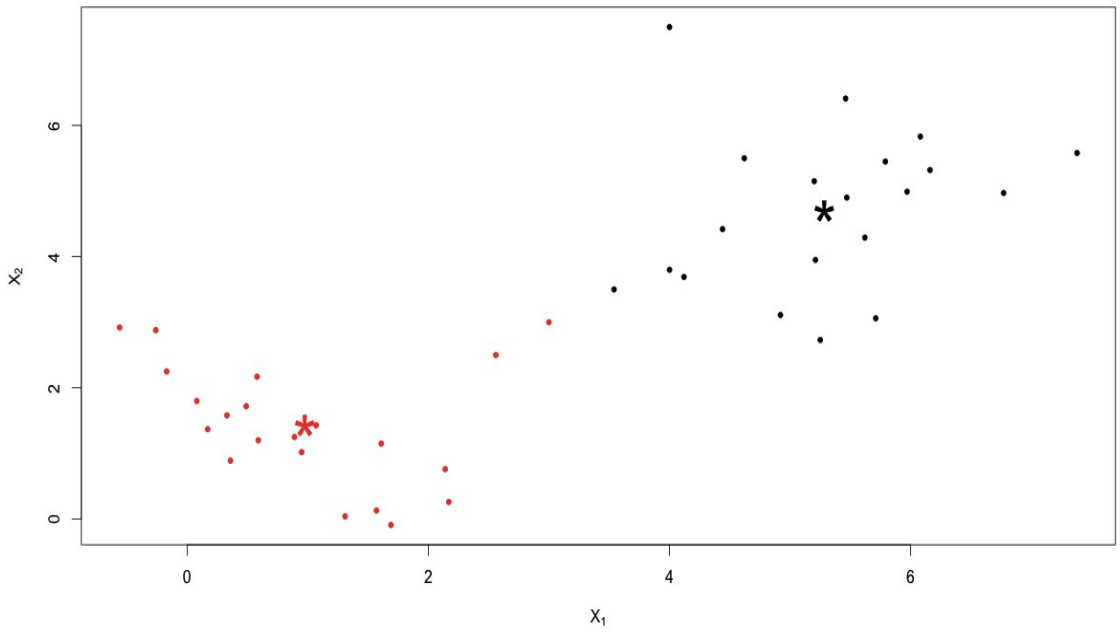
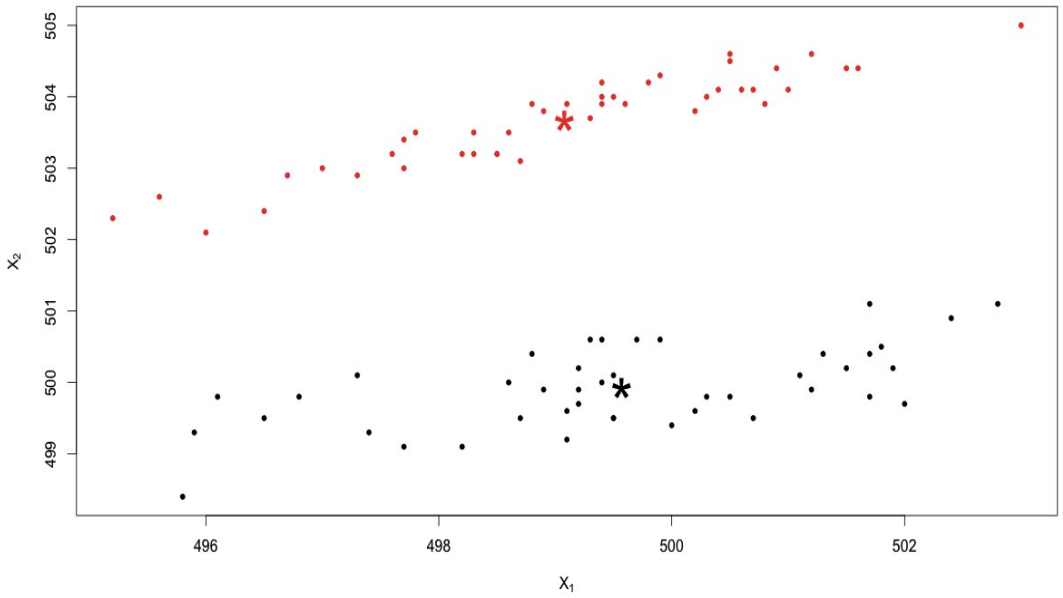
Análise de agrupamentos por métodos não hierárquicos

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

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

n = 41

# k-médias com dois grupos
mk1 <- kmeans(dados, centers = 2)

plot(dados, pch = 20, xlab = expression(X[1]), ylab = expression(X[2]),
     col = mk1$cluster)
points(mk1$centers, pch = "*", cex = 4, col = 1:2)
```




```
cat("\n Coordenadas dos centróides:")
mk1$centers

Coordenadas dos centróides:
         V1       V2
1 5.2855000 4.707500
2 0.9795238 1.439524

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

```
# k-médias com dois grupos
mk2 <- kmeans(dados, centers = 2)

plot(dados, pch = 20, xlab = expression(X[1]), ylab = expression(X[2]),
     col = mk2$cluster)
points(mk2$centers, pch = "*", cex = 4, col = 1:2)
```




```
cat("\n Coordenadas dos centróides:")
mk2$centers
```

```
 Coordenadas dos centróides:
        V1       V2
1 499.5698 499.9233
2 499.0814 503.6744
```

```
## 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

# k-médias com k = 2,3,4,5 grupos
mk32 <- kmeans(dados, centers = 2, nstart = 10)
mk33 <- kmeans(dados, centers = 3, nstart = 10)
mk34 <- kmeans(dados, centers = 4, nstart = 10)
mk35 <- kmeans(dados, centers = 5, nstart = 10)
```

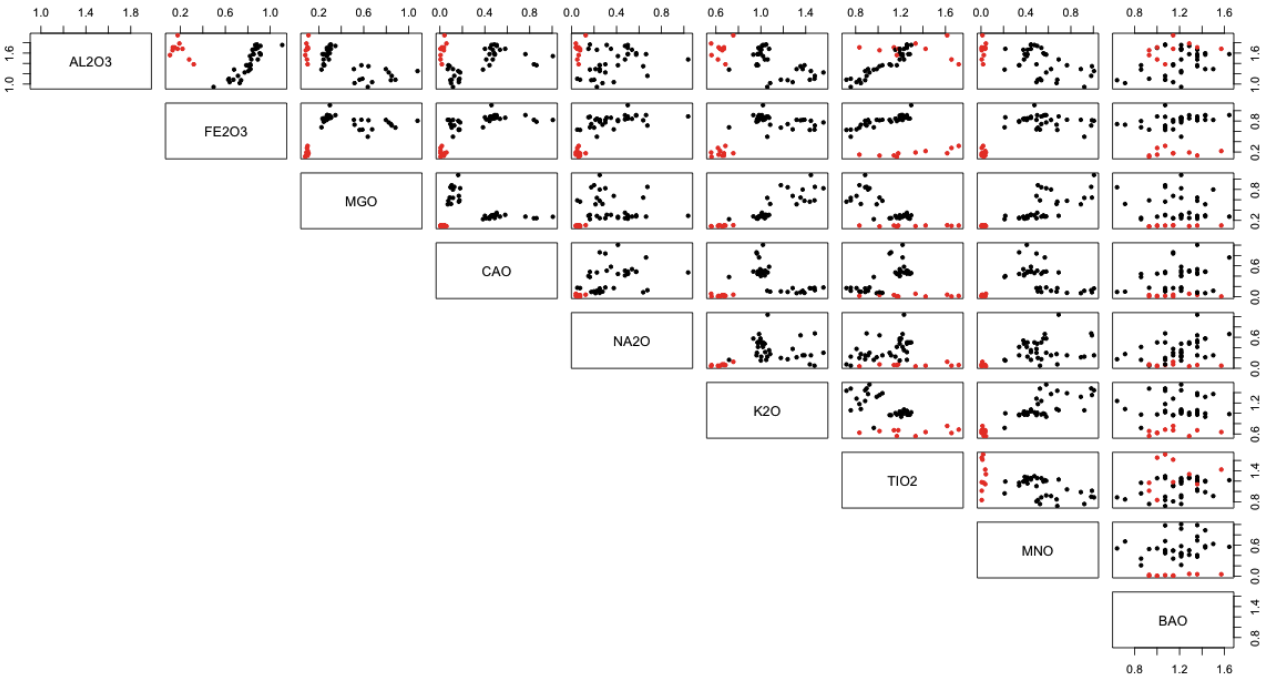
A semente é obtida calculando o centróide de `nstart` = 10 observações selecionadas aleatoriamente do conjunto de dados.

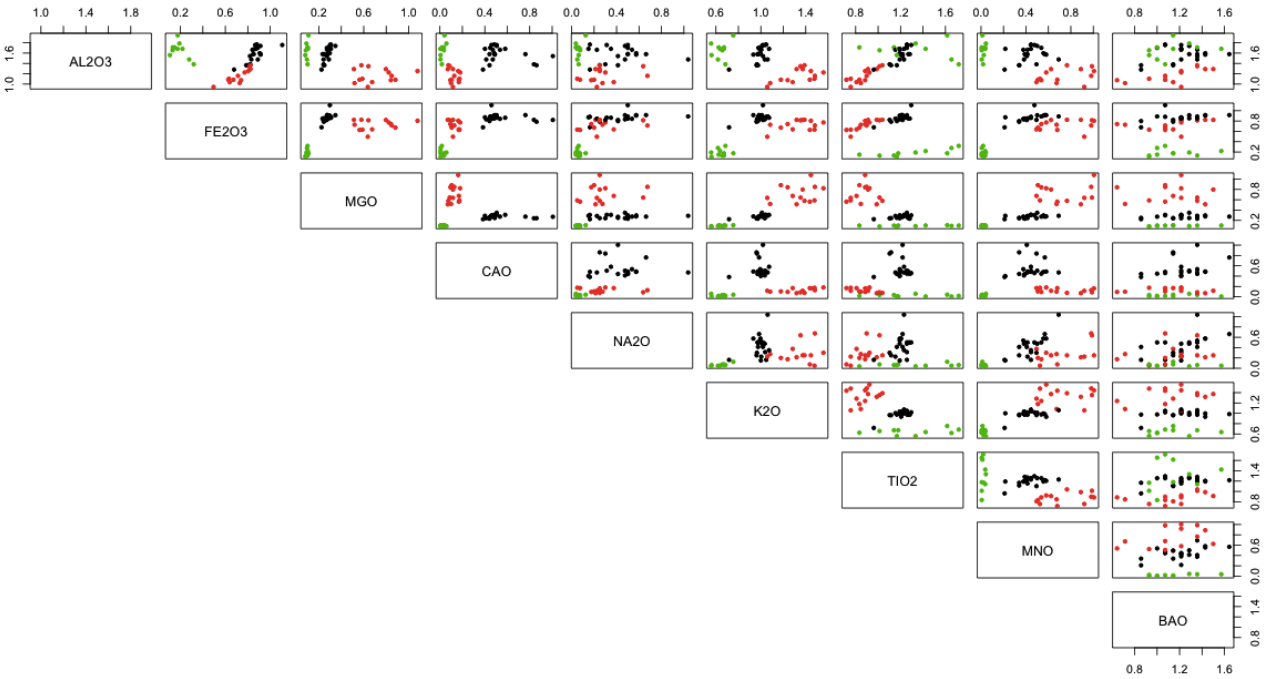
```
## Resultados da solução com k = 3 grupos
mk33

K-means clustering with 3 clusters of sizes 14, 21, 10

Cluster means:
     AL2O3     FE2O3        MGO        CAO      NA2O       K2O
1 1.162216 0.7218439 0.71311301 0.12458472 0.2821429 1.3337125
2 1.581219 0.8637874 0.27498223 0.54595792 0.4321429 0.9881711
3 1.658879 0.1874419 0.09552239 0.02267442 0.0637500 0.6436306
       TIO2        MNO      BAO
1 0.8754579 0.72619048 1.137755
2 1.2020757 0.43915344 1.224490
3 1.3076923 0.01975309 1.142857

Clustering vector:
 [1] 1 2 1 1 2 1 2 2 3 2 2 3 2 1 2 2 3 1 3 2 2 2 1 1 2 2 1 3 1
[30] 2 2 2 2 2 3 1 2 1 3 3 2 3 1 3 1

Within cluster sum of squares by cluster:
[1] 2.874794 3.164386 1.466713
 (between_SS / total_SS =  70.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"

# R2
cat("\n R2 = ", mk32$betweenss / mk32$totss)

R2 =  0.4132768

cat("\n R2 = ", mk33$betweenss / mk33$totss)

R2 =  0.7082965

cat("\n R2 = ", mk34$betweenss / mk34$totss)

R2 =  0.7516567

cat("\n R2 = ", mk35$betweenss / mk35$totss)

R2 =  0.7849956
```

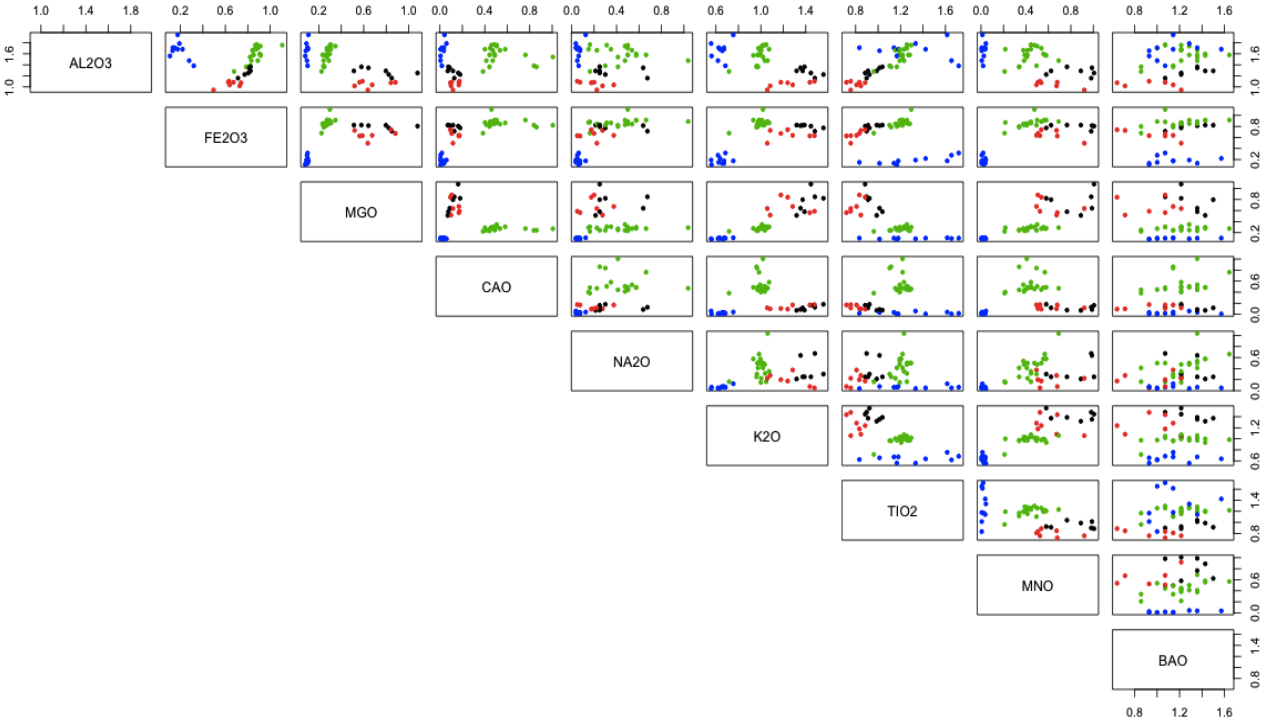
O valor de R^2 não diminui quando o número de grupos aumenta. A maior variação no valor de R^2 ocorre quando passamos de $k = 2$ para $k = 3$ grupos, indicando a formação de três grupos. Este critério baseado em R^2 também pode ser aplicado aos métodos hierárquicos aglomerativos.

Nota 1. A soma de quadrados total (`totss`) depende do número de grupos?

```
# Gráficos de dispersão
pairs(dados, pch = 20, col = mk32$cluster, lower.panel = NULL)
pairs(dados, pch = 20, col = mk33$cluster, lower.panel = NULL)
pairs(dados, pch = 20, col = mk34$cluster, lower.panel = NULL)
```


AL2O3
FE2O3
MGO
CAO
NA2O
K2O
TIO2
MNO
BAO


AL2O3
FE2O3
MGO
CAO
NA2O
K2O
TIO2
MNO
BAO




```
# Observações em cada grupo (k = 3)
mk33$size
```

```
[1] 14 21 10
```

```
nomes <- paste("O", 1:n, sep = "")
for (j in 1:3) {
   cat("\n Obs. no grupo ", j, ":", nomes[mk33$cluster == 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 2. Refaça os exemplos com as funções `pam` (*partitioning around medoids*) e `clara` (*clustering large applications*) do pacote `cluster` no lugar da função `kmeans`.

Nota 3. Procure refazer os exemplos utilizando outros pacotes estatísticos (SAS, SPSS, Minitab e Statistica, por exemplo).