
Estatística: Aplicação ao Sensoriamento Remoto

SER 204

Análise de Agrupamento

Camilo Daleles Rennó

camilo.renno@inpe.br

acesso do conteúdo do curso em [Bibdigital do INPE](#) ou [GitHub](#)

Análise de Agrupamento (*Cluster Analysis*)

A Análise de Agrupamento inclui técnicas ou métodos que visam o particionamento de uma população heterogênea em grupos mais homogêneos.

Este processo é feito sem a pré-determinação de classes, ou seja, os elementos da população são agrupados de acordo com alguma métrica que descreve o grau de similaridade ou dissimilaridade entre estes elementos. Por isso, muitas vezes são chamados de **classificadores não supervisionados**.

Os grupos (ou *clusters*) são formados de modo a obter-se a maior homogeneidade dentro dos grupos e a maior heterogeneidade entre eles.

Basicamente há 2 tipos de métodos:

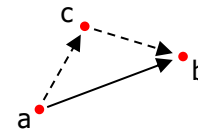
- a) Hierárquicos
- b) Não-Hierárquicos ou por Particionamento

Medidas de Distância

Grande parte dos métodos de agrupamento se baseia em medidas de similaridade ou de dissimilaridade entre os elementos a serem agrupados. Em geral, a dissimilaridade é traduzida por uma métrica que representa uma distância.

Sendo a , b e c elementos de uma população, uma métrica será uma distância desde que:

- a) $d(a, b) \geq 0$
- b) $d(a, a) = 0$
- c) $d(a, b) = d(b, a)$ (simetria)
- d) $d(a, b) \leq d(a, c) + d(c, b)$ (desigualdade triangular)

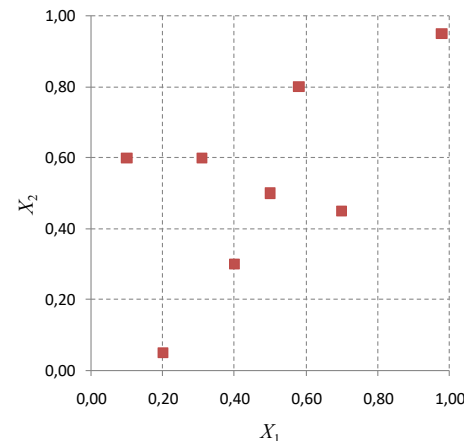


Supondo que n elementos sejam avaliadas segundo m variáveis diferentes (atributos), podemos representar o conjunto de valores observados por um matriz, onde cada elemento x_{ij} representa o valor da i -ésima amostra para a j -ésima variável.

$$\begin{bmatrix} X_1 & X_2 & \dots & X_m \\ x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

Exemplo 2D:

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

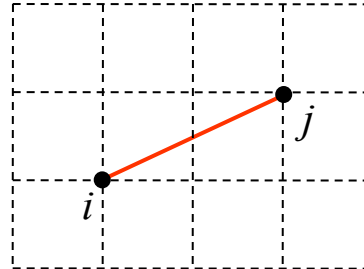


Medidas de Distância

Distância entre pares de pontos

Distância Euclidiana

$$d(i, j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

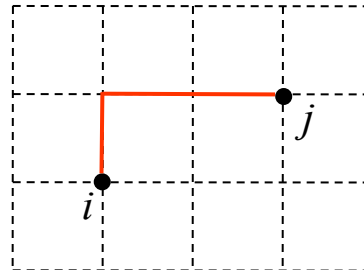


No R:

```
d <- dist(data, method="euclidian")
```

Distância de Manhattan (Quarteirão)

$$d(i, j) = \sum_{k=1}^m |x_{ik} - x_{jk}|$$



```
d <- dist(data, method="manhattan")
```

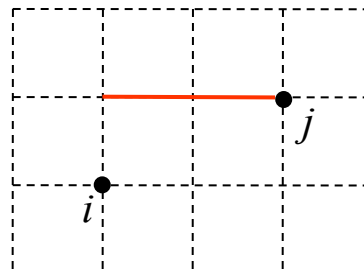
Distância de Minkowski

$$d(i, j) = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} \quad \text{sendo } p \geq 1$$

```
d <- dist(data, method="minkowski", p=p)
```

Distância de Chebychev

$$d(i, j) = \max_k |x_{ik} - x_{jk}|$$



```
d <- dist(data, method="maximum")
```

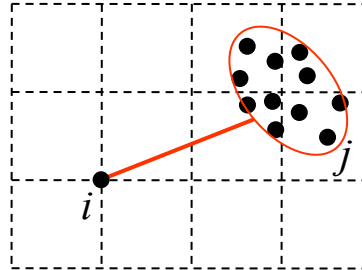
Calcula a distância entre todos os pares de pontos

Gera matrizes $n \times n$

Medidas de Distância

Distância de Mahalanobis
(distância de um ponto a uma nuvem)

$$d(i, j) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j)}$$

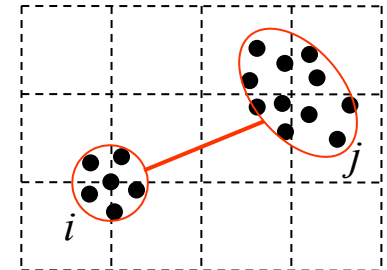


No R:

```
d <- mahalanobis(x, center, cov)
```

Distância de Bhattacharyya (considerando uma distribuição multigaussiana)
(distância entre nuvens)

$$d_B(i, j) = \frac{1}{8} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \left(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2} \right)^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2} \ln \left(\frac{\left| \frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2} \right|}{\sqrt{|\boldsymbol{\Sigma}_i| |\boldsymbol{\Sigma}_j|}} \right)$$



Distância de Jeffrey-Matusita

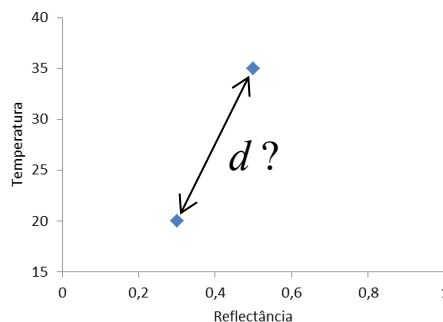
$$d_{JM}(i, j) = 2 \left(1 - e^{-d_B(i, j)} \right) \quad 0 \leq d_{JM}(i, j) \leq 2$$

Normalização das Variáveis

A medida de distância é afetada pela escala das variáveis.

Há duas situações em que a normalização das variáveis é imprescindível:

a) quando as variáveis apresentam diferentes unidades



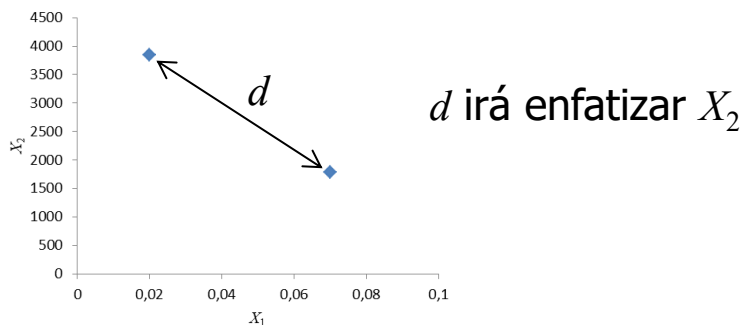
$$d(i, j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

Normalização das Variáveis

A medida de distância é afetada pela escala das variáveis.

Há duas situações em que a normalização das variáveis é imprescindível:

- a) quando as variáveis apresentam diferentes unidades
- b) quando não se deseja considerar o impacto das amplitudes de cada variável



Normalização das Variáveis

A medida de distância é afetada pela escala das variáveis.

Há duas situações em que a normalização das variáveis é imprescindível:

- a) quando as variáveis apresentam diferentes unidades
- b) quando não se deseja considerar o impacto das amplitudes de cada variável

Em geral, a normalização é um reescalonamento dos valores de cada variável

Este reescalonamento pode ser feito com base na informação contida na própria amostra ou utilizando o conhecimento do comportamento da variável.

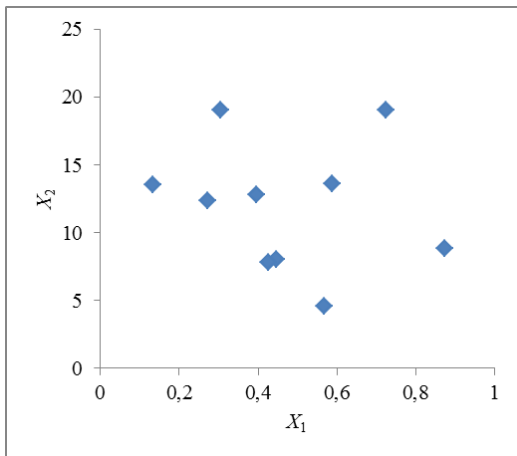
Muita atenção quando a normalização usar características amostrais e a metodologia proposta for ser aplicada em outras regiões.

Normalização das Variáveis

A normalização pode ser feita de diversas maneiras:

a) normalizar entre 0 e 1 $\rightarrow (x - x_{\min}) / (x_{\max} - x_{\min})$

X_1	X_2
0,725	19,06
0,874	8,88
0,567	4,58
0,271	12,38
0,306	19,10
0,446	8,08
0,397	12,82
0,133	13,56
0,587	13,64
0,426	7,82



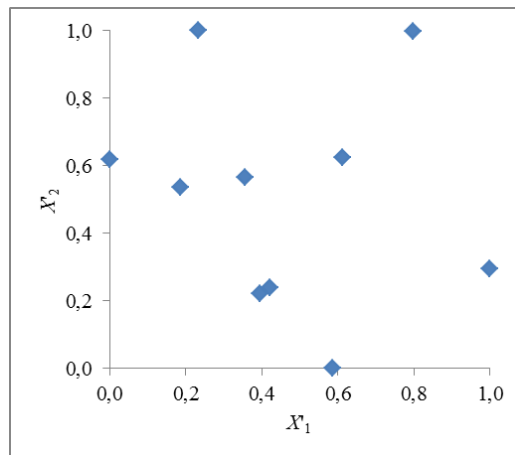
No R:

```
maxs <- apply(data, 2, max)
```

```
mins <- apply(data, 2, min)
```

```
datanew <- scale(data, center = mins, scale = maxs - mins)
```

X'_1	X'_2
0,799	0,997
1,000	0,296
0,586	0,000
0,186	0,537
0,233	1,000
0,422	0,241
0,356	0,567
0,000	0,618
0,613	0,624
0,395	0,223



Cuidado com a presença de *outliers*!

os valores mínimos e máximos podem ser determinados de outras formas:

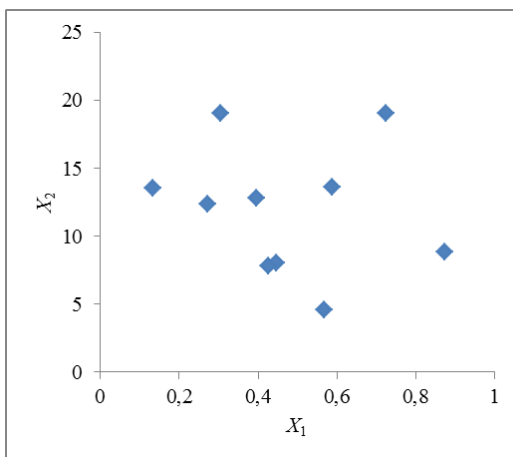
- valores esperados ou encontrados na literatura
- valores "aparados": por exemplo, desprezando-se 2,5% dos valores extremos (*outliers*)
nesses casos, os valores fora do escopo serão saturados em zero e um?

Normalização das Variáveis

A normalização pode ser feita de diversas maneiras:

b) dividir pelo valor máximo

X_1	X_2
0,725	19,06
0,874	8,88
0,567	4,58
0,271	12,38
0,306	19,10
0,446	8,08
0,397	12,82
0,133	13,56
0,587	13,64
0,426	7,82

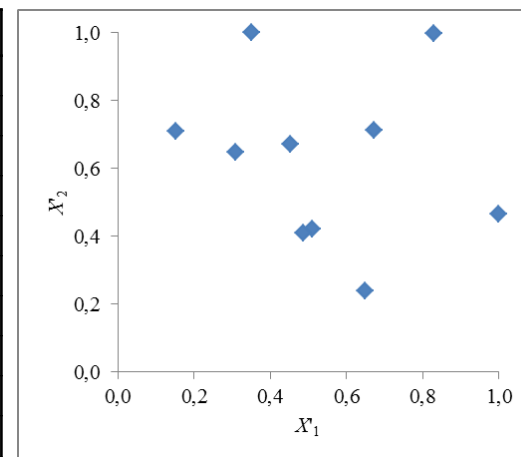


No R:

```
maxs <- apply(data, 2, max)
```

```
datanew <- scale(data, center=rep(0,2), scale = maxs)
```

X'_1	X'_2
0,830	0,998
1,000	0,465
0,649	0,240
0,310	0,648
0,350	1,000
0,510	0,423
0,454	0,671
0,152	0,710
0,672	0,714
0,487	0,409



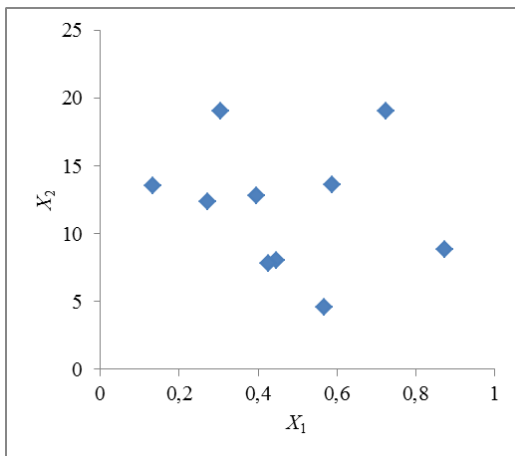
Cuidado com a presença de *outliers*!

Normalização das Variáveis

A normalização pode ser feita de diversas maneiras:

c) padronizar (subtrair a média e dividir pelo desvio padrão)

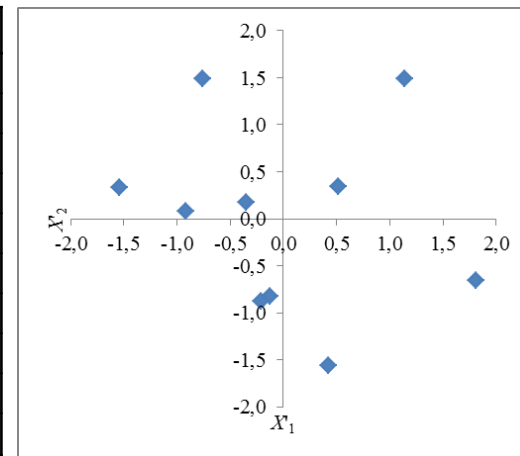
X_1	X_2
0,725	19,06
0,874	8,88
0,567	4,58
0,271	12,38
0,306	19,10
0,446	8,08
0,397	12,82
0,133	13,56
0,587	13,64
0,426	7,82



No R:

```
datanew <- scale(data, center=TRUE, scale = TRUE)
```

X'_1	X'_2
1,140	1,485
1,814	-0,654
0,425	-1,558
-0,915	0,082
-0,757	1,494
-0,123	-0,822
-0,345	0,174
-1,540	0,329
0,515	0,346
-0,214	-0,877



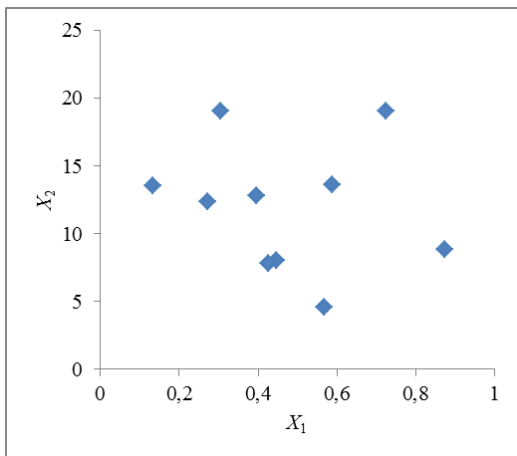
Cuidado com amostras muito pequenas

Normalização das Variáveis

A normalização pode ser feita de diversas maneiras:

d) transformar os valores em ordem (*rank*)

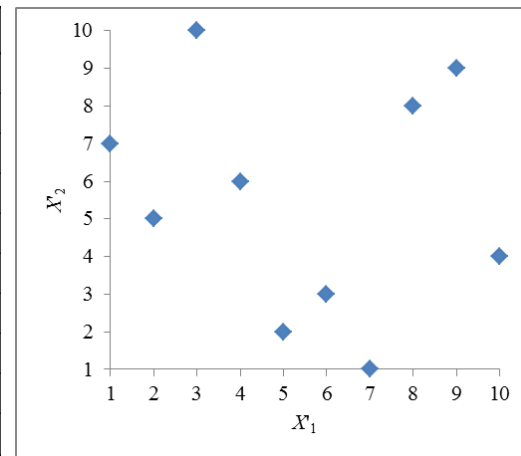
X_1	X_2
0,725	19,06
0,874	8,88
0,567	4,58
0,271	12,38
0,306	19,10
0,446	8,08
0,397	12,82
0,133	13,56
0,587	13,64
0,426	7,82



No R:

```
datanew <- apply(data, 2, rank)
```

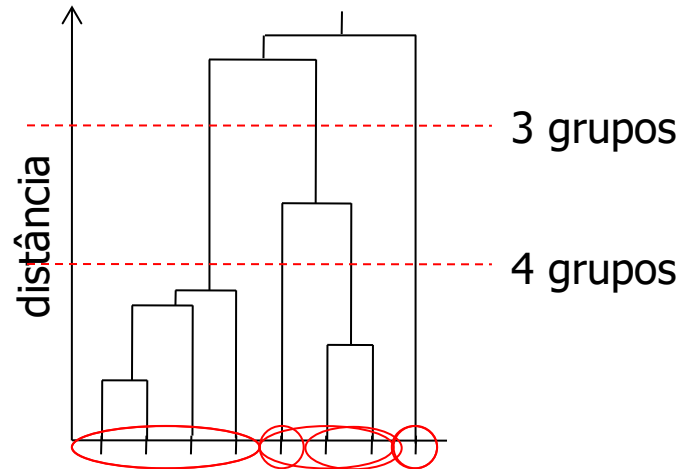
X'_1	X'_2
9	9
10	4
7	1
2	5
3	10
6	3
4	6
1	7
8	8
5	2



“conserta” distribuições muito assimétricas
minimiza problemas com *outliers*

Agrupamento Hierárquico

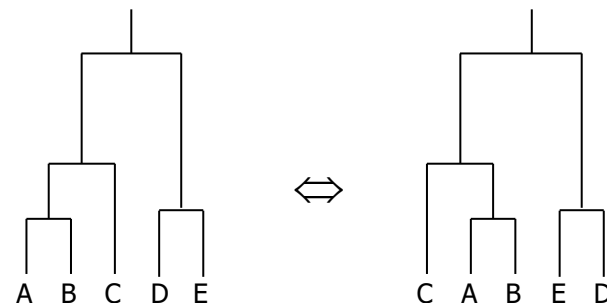
Este método produz um gráfico de saída chamado dendrograma ou diagrama de árvore que representa a estrutura hierárquica do agrupamento.



O dendrograma pode ser construído por aglomeração (mais comum) ou por divisão

Para definir o número de grupos, escolhe-se um valor de corte (subjetivo), em função do objetivo da análise

Cuidado ao analisar um dendrograma:
seu comportamento se assemelha a
um móbile!



Método Hierárquico por Aglomeração

No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

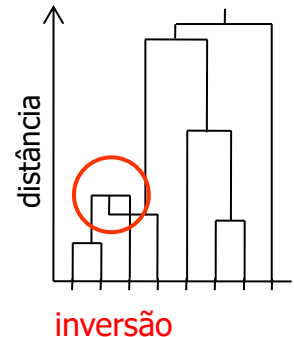
Há 3 métodos principais de aglomeração:

- a) Métodos de ligação (simples, completo, média, mediana)
- b) Método de centróide (pode ter problema de inversão no dendrograma)
- c) Método de Ward (baseia-se na minimização dos erros quadráticos)

Se A , B e C são 3 grupos, então A será unido a B se:

$$\begin{cases} SQE_{AB} - (SQE_A + SQE_B) < SQE_{AC} - (SQE_A + SQE_C) \\ SQE_{AB} - (SQE_A + SQE_B) < SQE_{BC} - (SQE_B + SQE_C) \end{cases}$$

$$SQE_G = \sum_{k=1}^m \sum_{i=1}^{n_G} (x_{ik} - \bar{x}_{ik})^2 \quad \forall x_{ik} \in G$$



Método Hierárquico por Aglomeração

No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

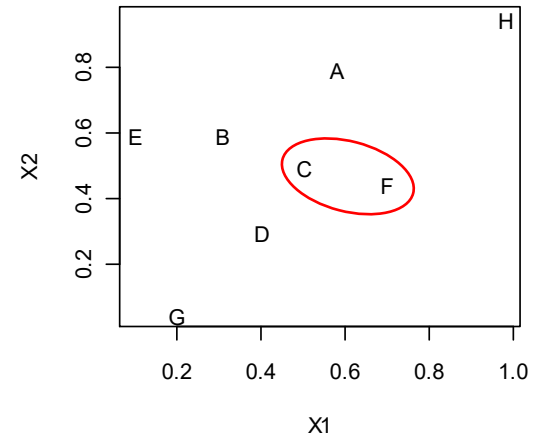
Método de ligação **simples - vizinho mais próximo**

possuem a mesma unidade e portanto não precisam ser normalizadas!

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	B	C	D	E	F	G
B	0,336						
C	0,310	0,215					
D	0,531	0,313	0,224				
E	0,520	0,210	0,412	0,424			
F	0,370	0,418	0,206	0,335	0,618		
G	0,841	0,561	0,541	0,320	0,559	0,640	
H	0,427	0,756	0,658	0,871	0,947	0,573	1,191



$$\text{dist}(A, CF) = \min\{\text{dist}(A, C), \text{dist}(A, F)\}$$

$$\text{dist}(B, CF) = \min\{\text{dist}(B, C), \text{dist}(B, F)\} \dots$$

Método Hierárquico por Aglomeração

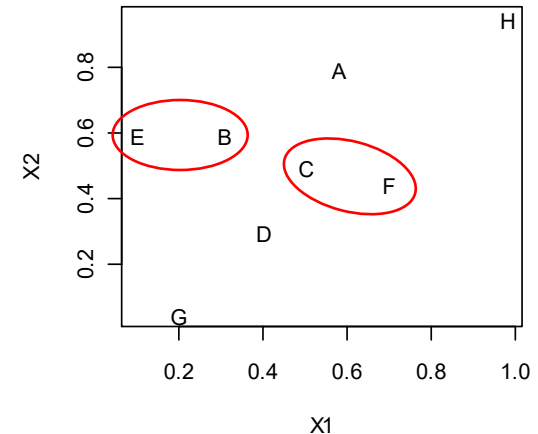
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **simples - vizinho mais próximo**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	B	CF	D	E	G
B	0,336					
CF	0,310	0,215				
D	0,531	0,313	0,224			
E	0,520	0,210	0,412	0,424		
G	0,841	0,561	0,541	0,320	0,559	
H	0,427	0,756	0,573	0,871	0,947	1,191



Método Hierárquico por Aglomeração

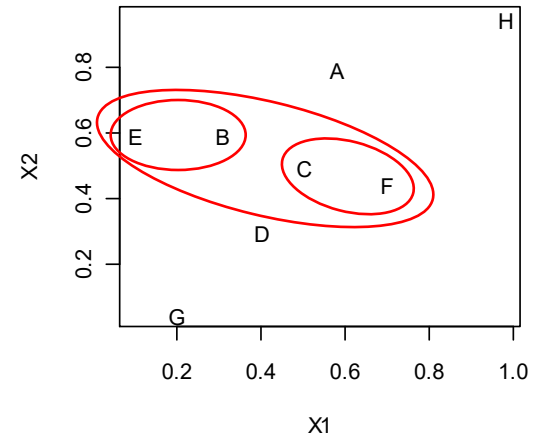
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **simples - vizinho mais próximo**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	BE	CF	D	G
BE	0,336				
CF	0,310	0,215			
D	0,531	0,313	0,224		
G	0,841	0,559	0,541	0,320	
H	0,427	0,756	0,573	0,871	1,191



Método Hierárquico por Aglomeração

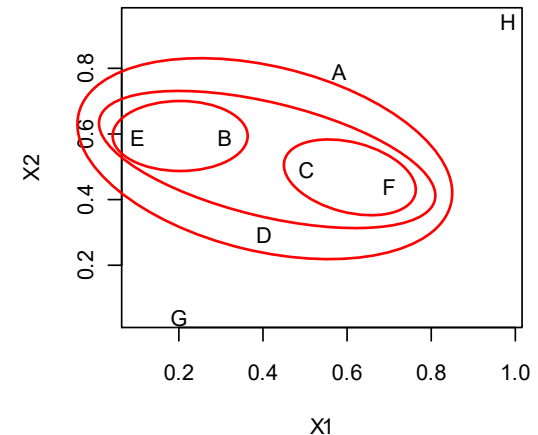
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **simples - vizinho mais próximo**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	BCEF	D	G
BCEF	0,310			
D	0,531	0,224		
G	0,841	0,541	0,320	
H	0,427	0,573	0,871	1,191



Método Hierárquico por Aglomeração

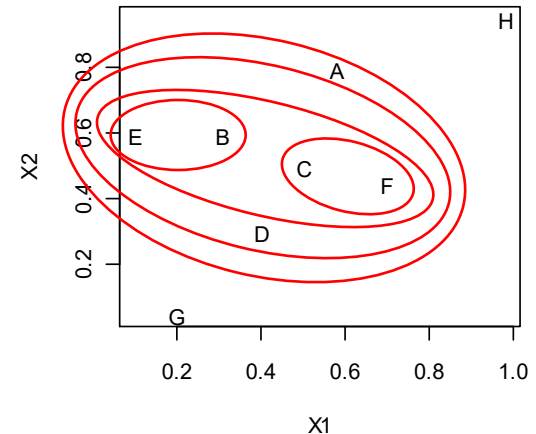
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **simples - vizinho mais próximo**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	BCDEF	G
BCDEF	0,310		
G	0,841	0,320	
H	0,427	0,573	1,191



Método Hierárquico por Aglomeração

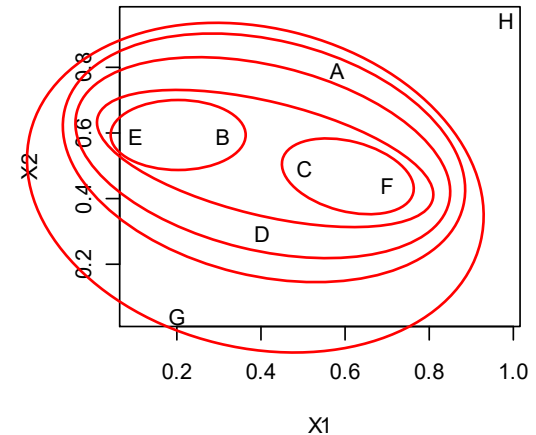
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **simples - vizinho mais próximo**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	ABCDEF	G
G	0,320	
H	0,427	1,191



Método Hierárquico por Aglomeração

No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

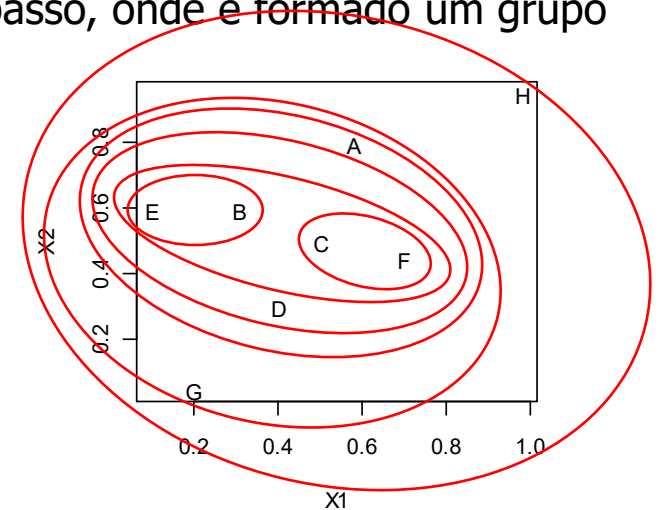
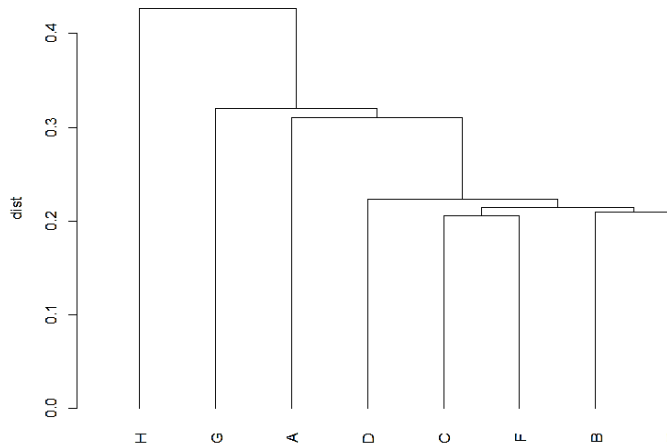
Método de ligação **simples - vizinho mais próximo**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	B	C	D	E	F	G
H							

0,427



Método Hierárquico por Aglomeração

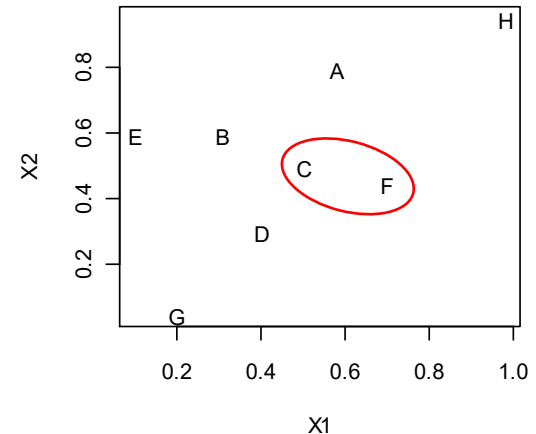
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **completo - vizinho mais distante**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	B	C	D	E	F	G
B	0,336						
C	0,310	0,215					
D	0,531	0,313	0,224				
E	0,520	0,210	0,412	0,424			
F	0,370	0,418	0,206	0,335	0,618		
G	0,841	0,561	0,541	0,320	0,559	0,640	
H	0,427	0,756	0,658	0,871	0,947	0,573	1,191



$$\text{dist}(A, CF) = \max \{ \text{dist}(A, C), \text{dist}(A, F) \}$$

$$\text{dist}(B, CF) = \max \{ \text{dist}(B, C), \text{dist}(B, F) \} \dots$$

Método Hierárquico por Aglomeração

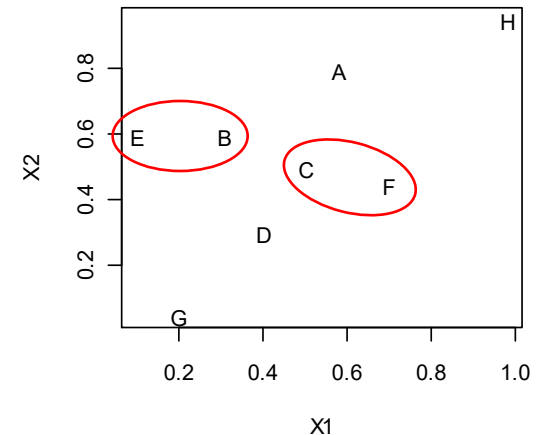
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **completo - vizinho mais distante**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	B	CF	D	E	G
B	0,336					
CF	0,370	0,418				
D	0,531	0,313	0,335			
E	0,520	0,210	0,618	0,424		
G	0,841	0,561	0,640	0,320	0,559	
H	0,427	0,756	0,658	0,871	0,947	1,191



Método Hierárquico por Aglomeração

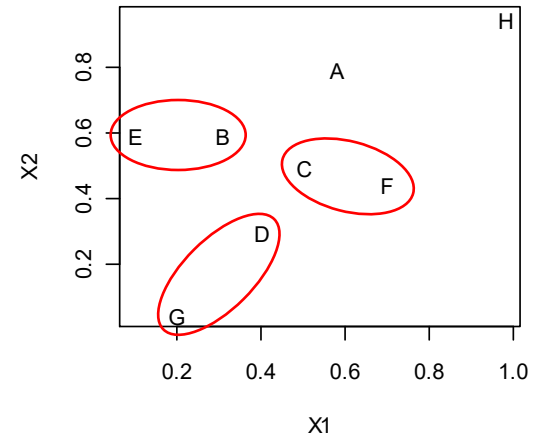
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **completo - vizinho mais distante**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	BE	CF	D	G
BE	0,520				
CF	0,370	0,618			
D	0,531	0,424	0,335		
G	0,841	0,561	0,640	0,320	
H	0,427	0,947	0,658	0,871	1,191



Método Hierárquico por Aglomeração

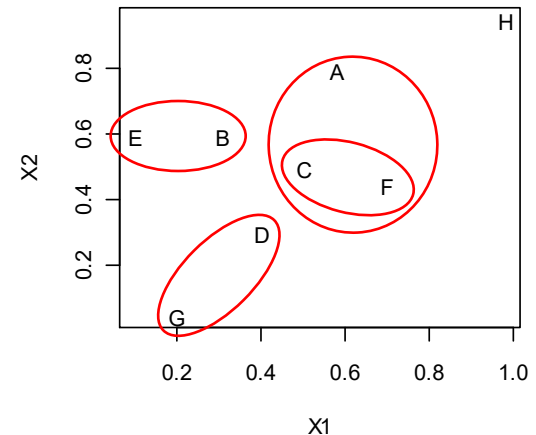
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **completo - vizinho mais distante**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	BE	CF	DG
BE	0,520			
CF	0,370	0,618		
DG	0,841	0,561	0,640	
H	0,427	0,947	0,658	1,191



Método Hierárquico por Aglomeração

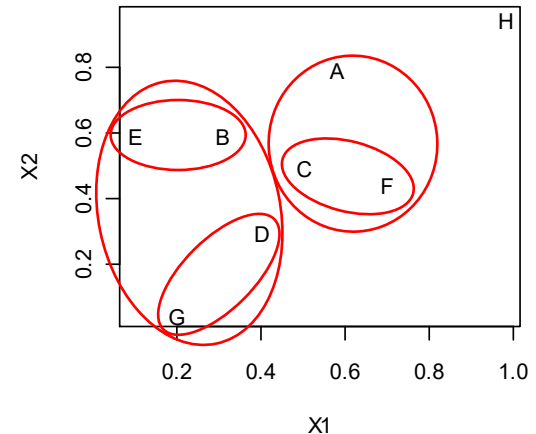
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **completo - vizinho mais distante**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	ACF	BE	DG
BE	0,618		
DG	0,841	0,561	
H	0,658	0,947	1,191



Método Hierárquico por Aglomeração

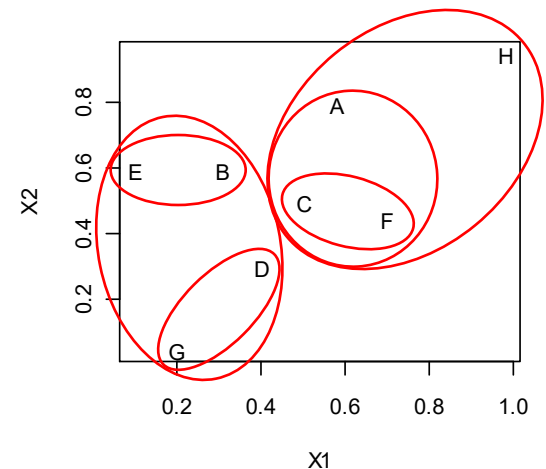
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **completo - vizinho mais distante**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	ACF	BDEG
BDEG	0,841	
H	0,658	1,191



Método Hierárquico por Aglomeração

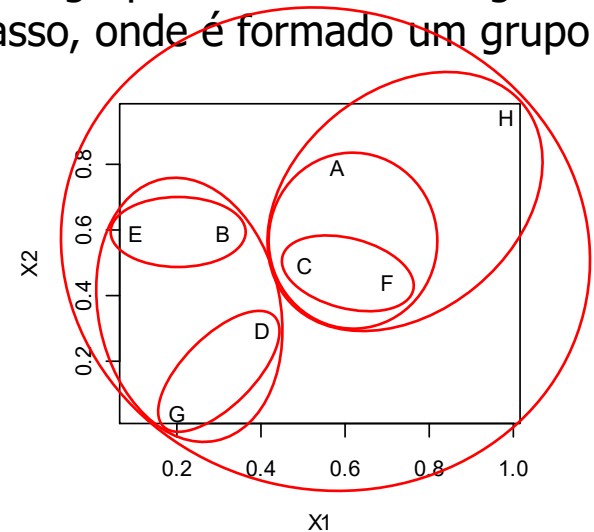
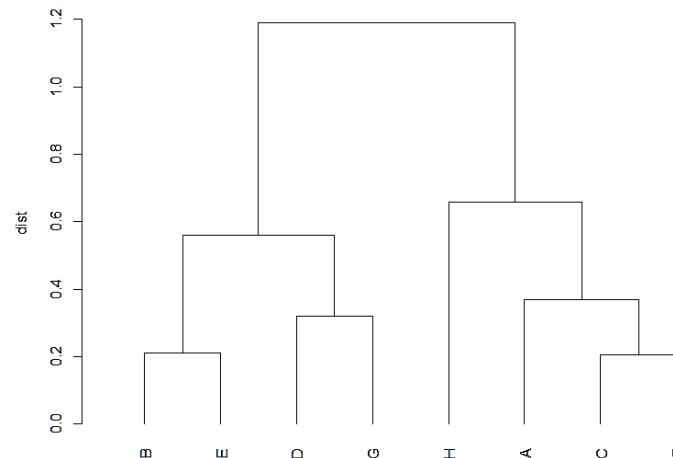
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **completo - vizinho mais distante**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	ACFH
BDEG	1,191



Método Hierárquico por Aglomeração

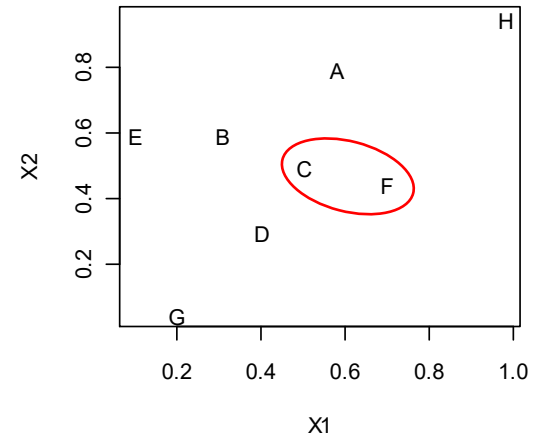
No método hierárquico por aglomeração, definem-se tantos grupos quantos elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **média**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	B	C	D	E	F	G
B	0,336						
C	0,310	0,215					
D	0,531	0,313	0,224				
E	0,520	0,210	0,412	0,424			
F	0,370	0,418	0,206	0,335	0,618		
G	0,841	0,561	0,541	0,320	0,559	0,640	
H	0,427	0,756	0,658	0,871	0,947	0,573	1,191



$$\text{dist}(A, CF) = \text{média}\{\text{dist}(A, C), \text{dist}(A, F)\}$$

$$\text{dist}(B, CF) = \text{média}\{\text{dist}(B, C), \text{dist}(B, F)\} \dots$$

Método Hierárquico por Aglomeração

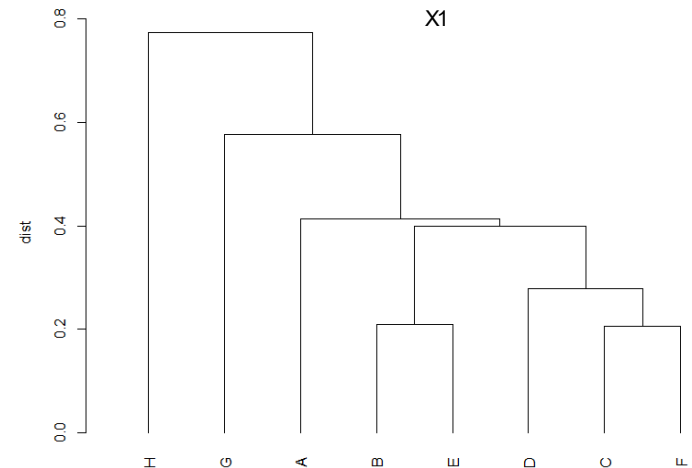
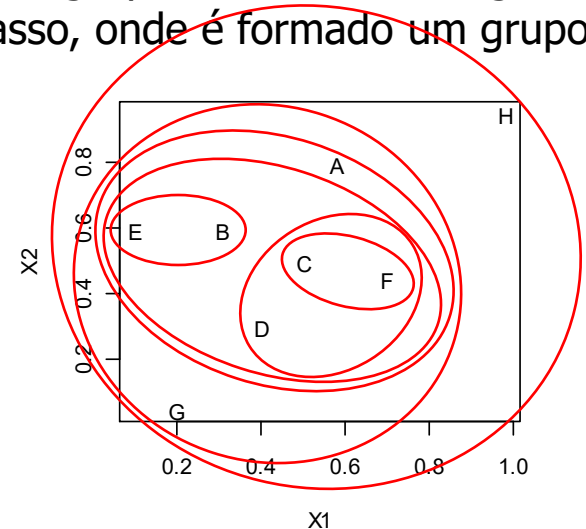
No método hierárquico por aglomeração, definem-se tantos grupos quanto elementos (cada elemento representa um grupo) e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Método de ligação **média**

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

Distância Euclidiana

	A	B	C	D	E	F	G
B	0,336						
C	0,310	0,215					
D	0,531	0,313	0,224				
E	0,520	0,210	0,412	0,424			
F	0,370	0,418	0,206	0,335	0,618		
G	0,841	0,561	0,541	0,320	0,559	0,640	
H	0,427	0,756	0,658	0,871	0,947	0,573	1,191



Método Hierárquico por Aglomeração no R

	X_1	X_2
A	0,58	0,80
B	0,31	0,60
C	0,50	0,50
D	0,40	0,30
E	0,10	0,60
F	0,70	0,45
G	0,20	0,05
H	0,98	0,95

```
x1<-c(0.58,0.31,0.5,0.4,0.1,0.7,0.2,0.98)
x2<-c(0.8,0.6,0.5,0.3,0.6,0.45,0.05,0.95)
data<-cbind(x1,x2)
rownames(data)<-c("A","B","C","D","E","F","G","H")
d<-dist(data,method="euclidian")
d
```

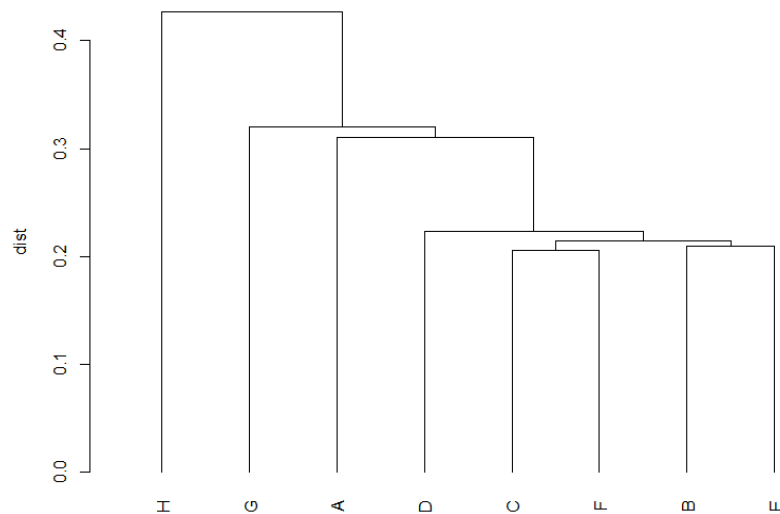
```

      A          B          C          D          E          F          G
B 0.3360060
C 0.3104835 0.2147091
D 0.5314132 0.3132092 0.2236068
E 0.5200000 0.2100000 0.4123106 0.4242641
F 0.3700000 0.4178516 0.2061553 0.3354102 0.6184658
G 0.8407735 0.5608921 0.5408327 0.3201562 0.5590170 0.6403124
H 0.4272002 0.7559100 0.6579514 0.8711487 0.9470480 0.5730620 1.1909660

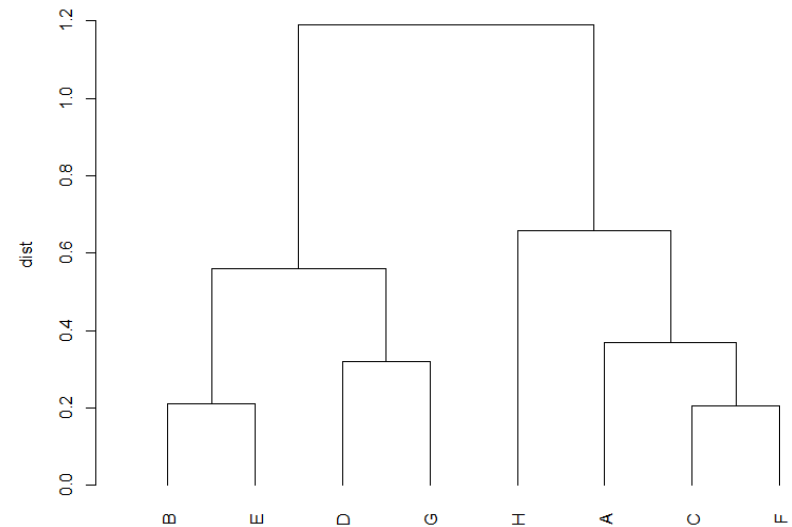
```

Método Hierárquico por Aglomeração no R

```
fit<-hclust(d,method="single")  
plot(as.dendrogram(fit),ylab="dist")
```



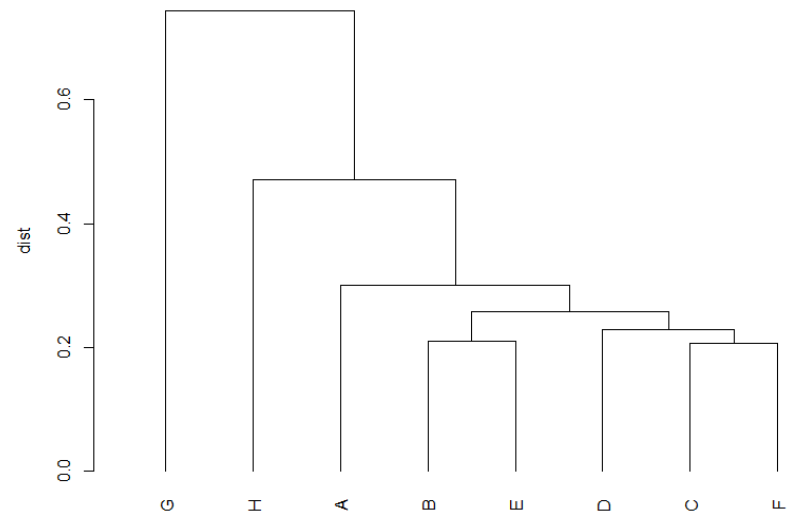
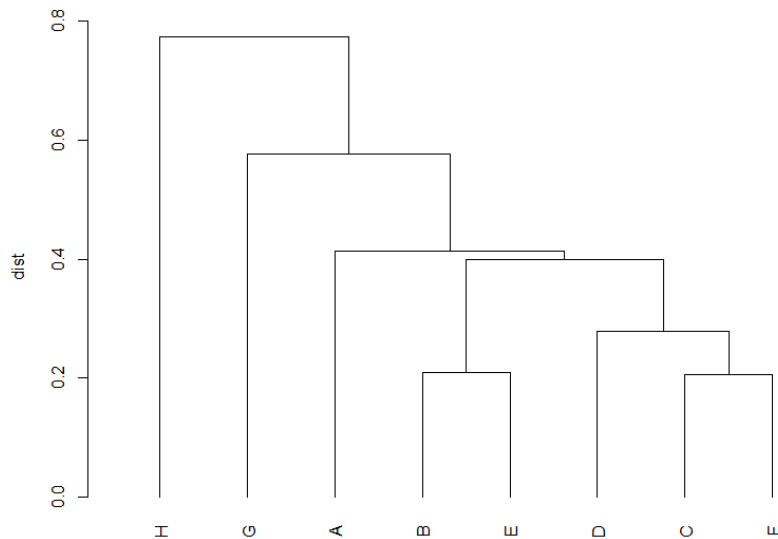
```
fit<-hclust(d,method="complete")  
plot(as.dendrogram(fit),ylab="dist")
```



Método Hierárquico por Aglomeração no R

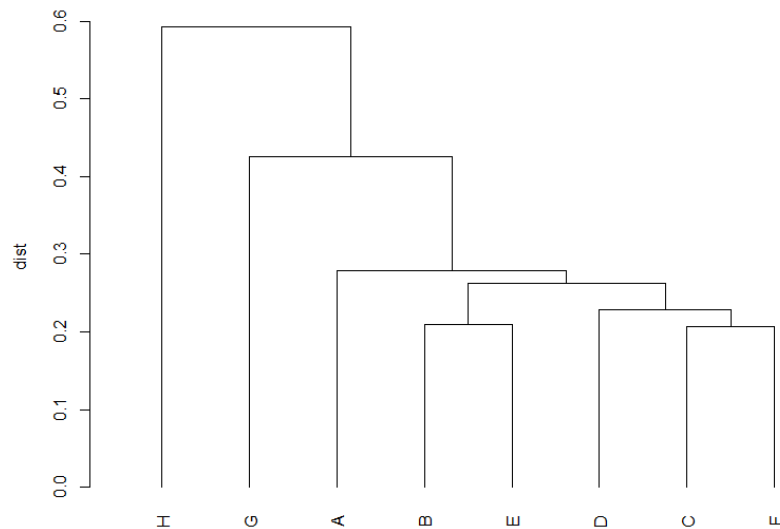
```
fit<-hclust(d,method="average")  
plot(as.dendrogram(fit),ylab="dist")
```

```
fit<-hclust(d,method="median")  
plot(as.dendrogram(fit),ylab="dist")
```

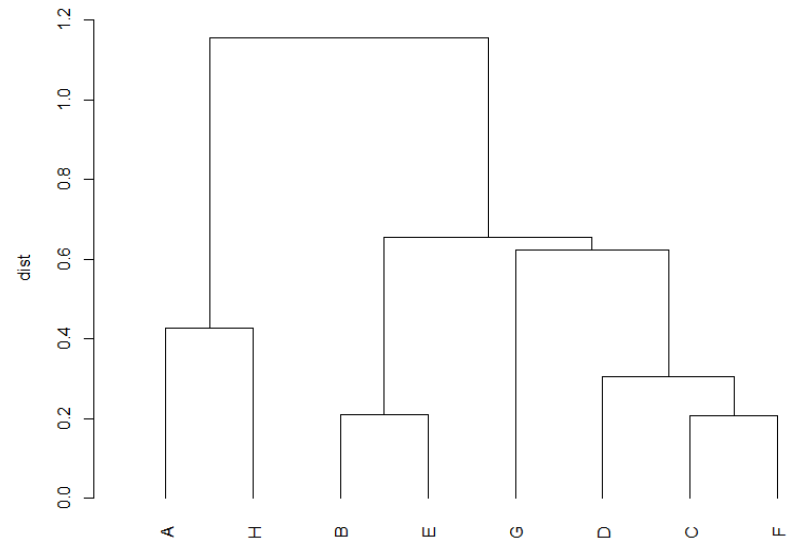


Método Hierárquico por Aglomeração no R

```
fit<-hclust(d,method="centroid")  
plot(as.dendrogram(fit),ylab="dist")
```



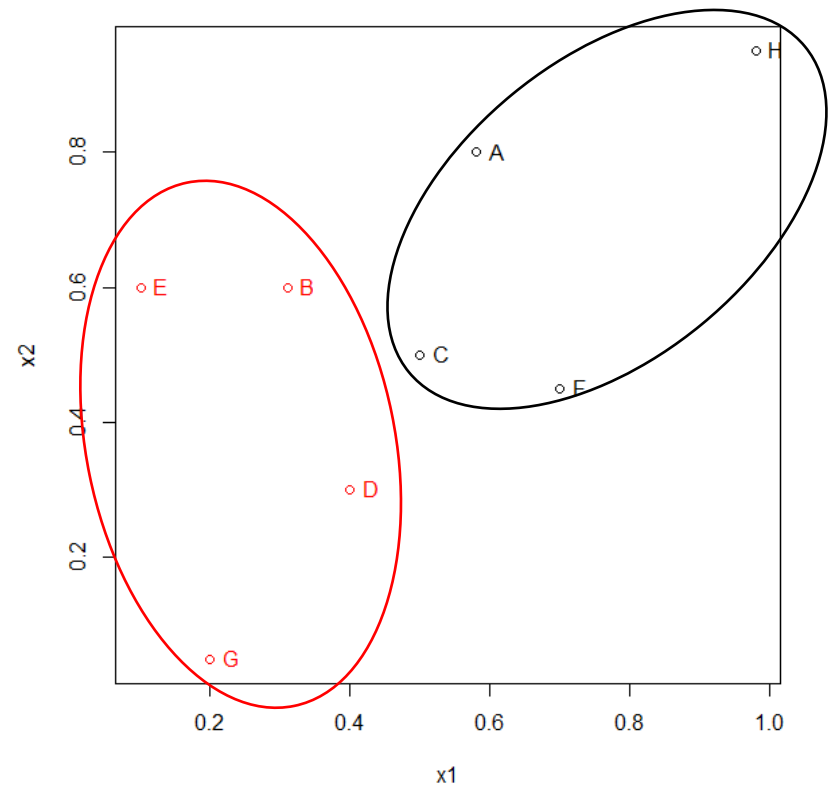
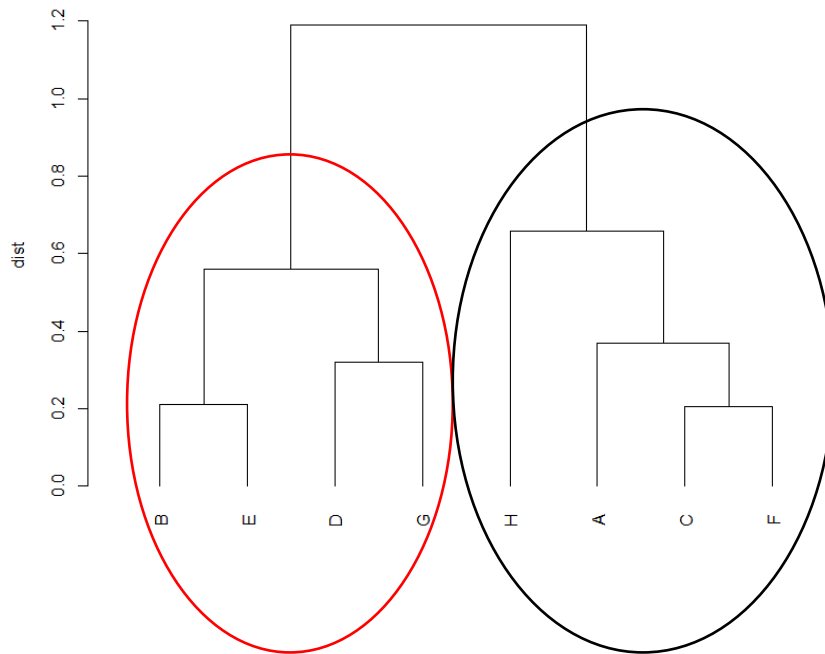
```
fit<-hclust(d,method="ward.D")  
plot(as.dendrogram(fit),ylab="dist")
```



Método Hierárquico por Aglomeração no R

```
fit<-hclust(d,method="complete")  
plot(as.dendrogram(fit),ylab="dist")
```

```
groups <- cutree(fit, k=2)  
plot(data,col=groups)  
text(data,rownames(data),col=groups,pos=4)
```



Método por Particionamento

A ideia básica deste método é agrupar-se os elementos em K grupos, onde K é previamente definido.

De modo geral, o método inicia-se escolhendo-se uma partição inicial dos elementos e, em seguida, por um processo iterativo, os elementos são redistribuídos nos grupos observando-se a máxima similaridade (ou mínima distância) até obter-se a melhor partição possível.

Como a escolha do número K é arbitrária, nem todos K grupos são necessariamente representativos e, por isso, aplica-se o método várias vezes para diferentes valores de K , escolhendo os resultados que apresentem melhor a interpretação dos grupos

Quando comparado com o método hierárquico, o método por particionamento é mais rápido e simples pois não é necessário recalcular a cada iteração a matriz de distâncias entre os elementos.

É indicado quando o número de elementos é muito grande.

Os métodos por particionamento mais conhecidos são o k -médias (*k-mean*) e o k -medóides* (*k-medoid*)

*Medóide: é o elemento representativo de um grupo cuja dissimilaridade média para todos os demais elementos é mínima

K-médias

Algoritmo Básico do Método K-Médias:

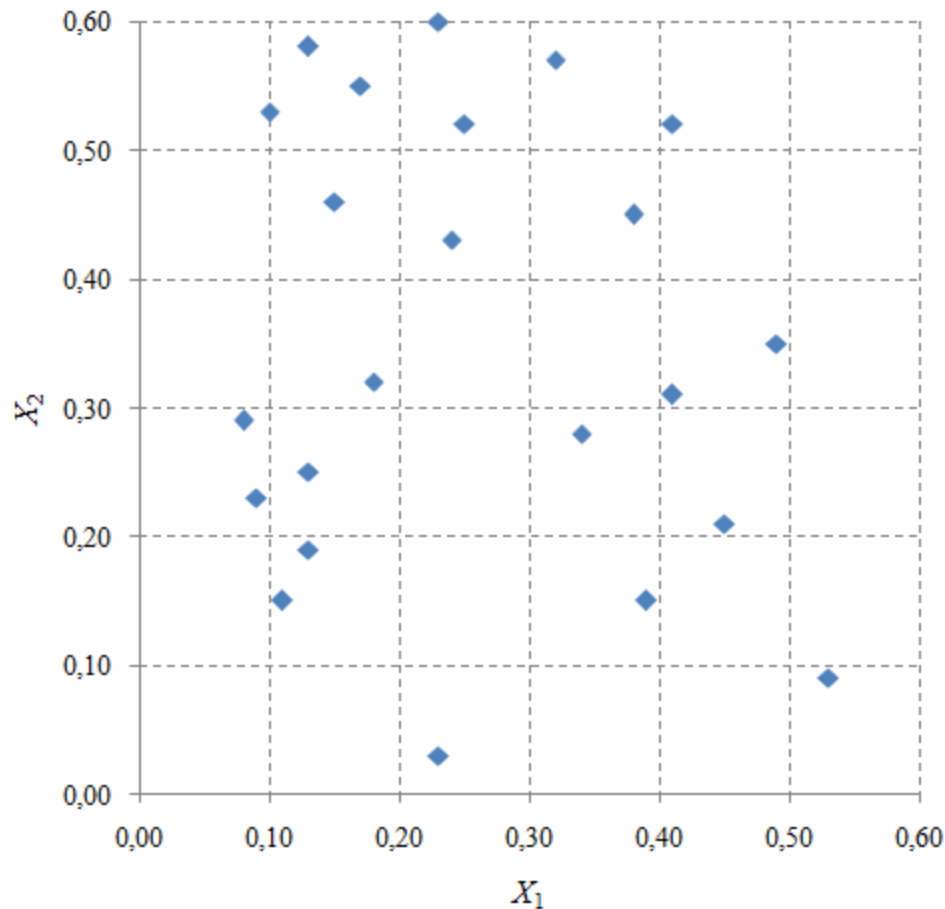
- a) Normalize cada variável (se necessário)
- b) Determine o número de agrupamentos desejado k
- c) Divida os elementos **aleatoriamente** em k grupos
- d) Calcule os centróides de cada grupo (em geral através da posição média)
- e) Para cada elemento, calcule a **distância euclidiana** em relação ao centróide de cada grupo
- f) Redefina os grupos de acordo com a distância mínima
- g) Repita os procedimentos de d a f **até que os centróides não mudem de posição**

OBS:

- no item c, outros critérios podem ser utilizados no momento de definir os k grupos iniciais
- no item e, outras distâncias podem ser utilizadas
- no item g, quando o número de elementos é muito grande, pode-se definir um outro critério de parada (por exemplo, menos de 1% dos elementos mudaram de grupo entre duas iterações consecutivas)
- em geral, testa-se vários valores de k , escolhendo-se aquele que minimiza a variância intra-grupos e maximiza a variância entre grupos

K-médias

Exemplo 2D: $k = 4$, distância euclidiana

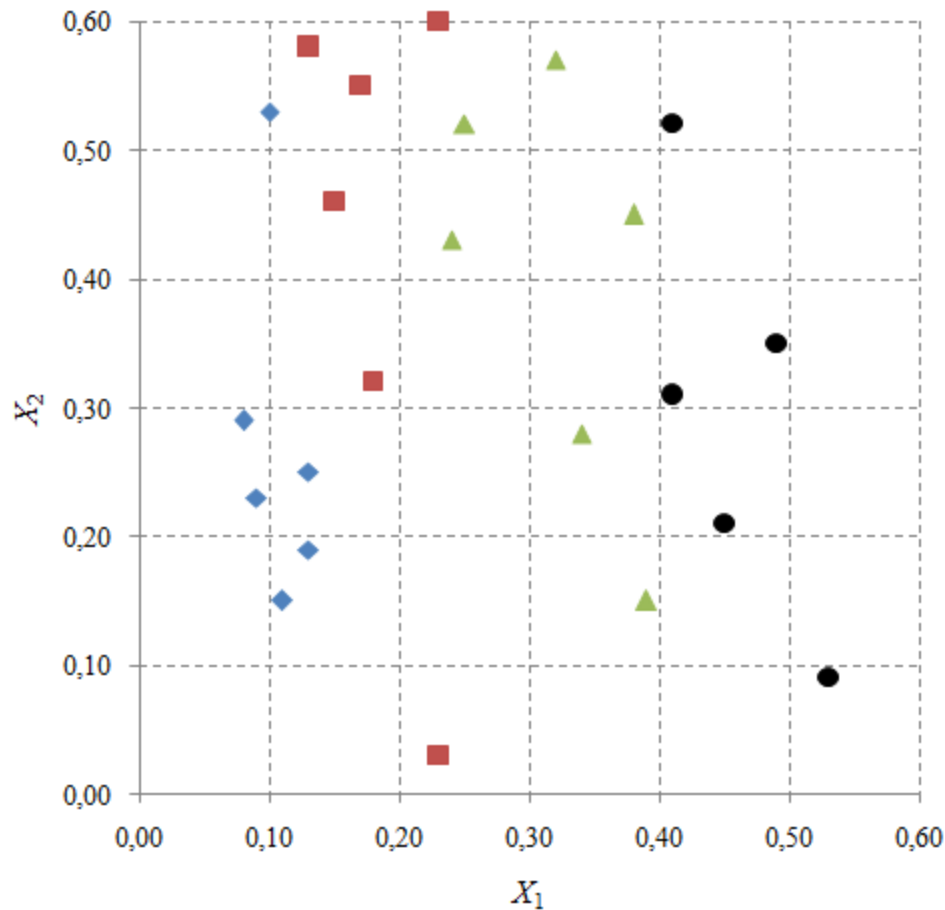


	X_1	X_2
1	0,08	0,29
2	0,09	0,23
3	0,11	0,15
4	0,10	0,53
5	0,13	0,19
6	0,13	0,25
7	0,13	0,58
8	0,15	0,46
9	0,17	0,55
10	0,18	0,32
11	0,23	0,03
12	0,23	0,60
13	0,24	0,43
14	0,25	0,52
15	0,32	0,57
16	0,34	0,28
17	0,38	0,45
18	0,39	0,15
19	0,41	0,52
20	0,41	0,31
21	0,45	0,21
22	0,49	0,35
23	0,53	0,09

Classificam-se aleatoriamente os elementos...

K-médias

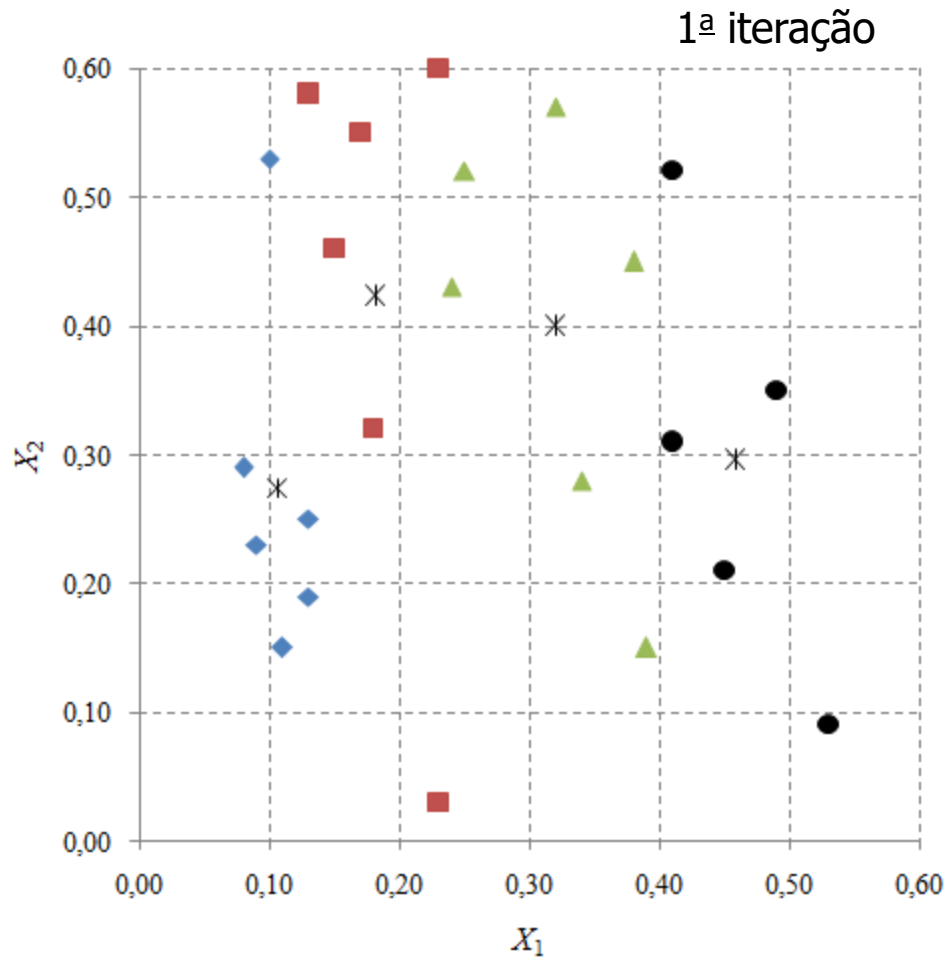
Exemplo 2D: $k = 4$, distância euclidiana



	X_1	X_2
1	0,08	0,29
2	0,09	0,23
3	0,11	0,15
4	0,10	0,53
5	0,13	0,19
6	0,13	0,25
7	0,13	0,58
8	0,15	0,46
9	0,17	0,55
10	0,18	0,32
11	0,23	0,03
12	0,23	0,60
13	0,24	0,43
14	0,25	0,52
15	0,32	0,57
16	0,34	0,28
17	0,38	0,45
18	0,39	0,15
19	0,41	0,52
20	0,41	0,31
21	0,45	0,21
22	0,49	0,35
23	0,53	0,09

K-médias

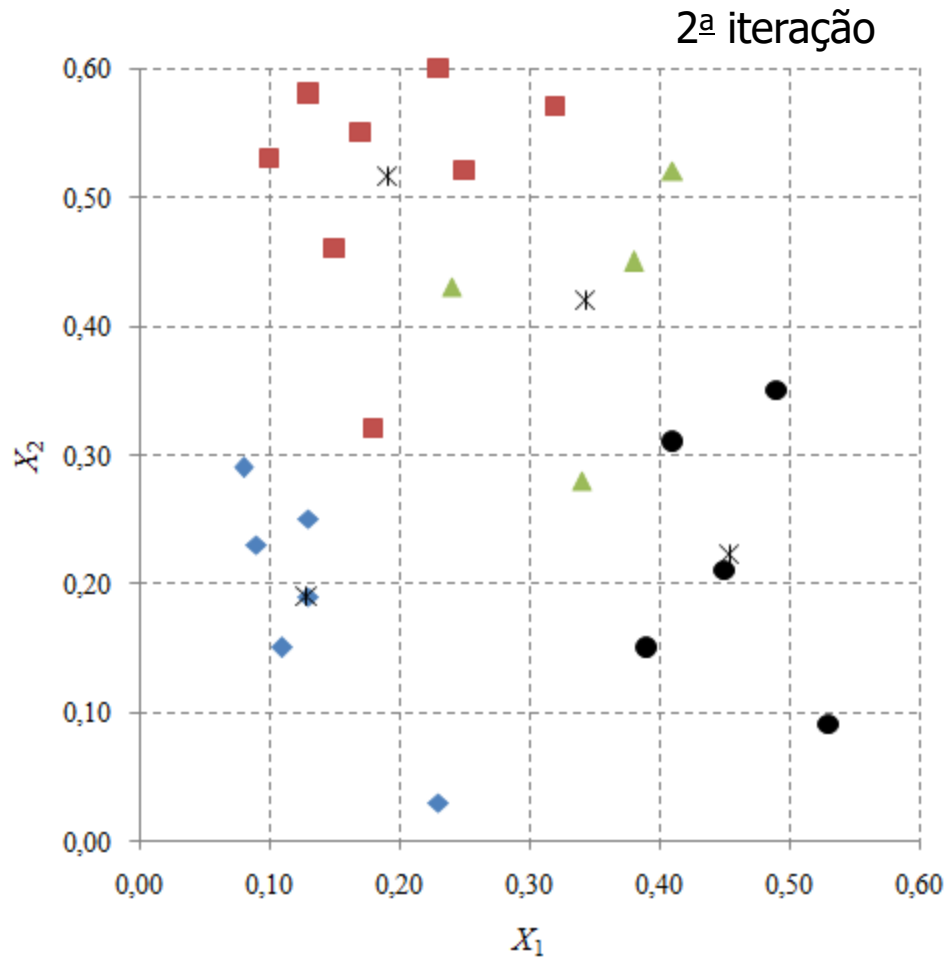
Exemplo 2D: $k = 4$, distância euclidiana



	X_1	X_2
1	0,08	0,29
2	0,09	0,23
3	0,11	0,15
4	0,10	0,53
5	0,13	0,19
6	0,13	0,25
7	0,13	0,58
8	0,15	0,46
9	0,17	0,55
10	0,18	0,32
11	0,23	0,03
12	0,23	0,60
13	0,24	0,43
14	0,25	0,52
15	0,32	0,57
16	0,34	0,28
17	0,38	0,45
18	0,39	0,15
19	0,41	0,52
20	0,41	0,31
21	0,45	0,21
22	0,49	0,35
23	0,53	0,09

K-médias

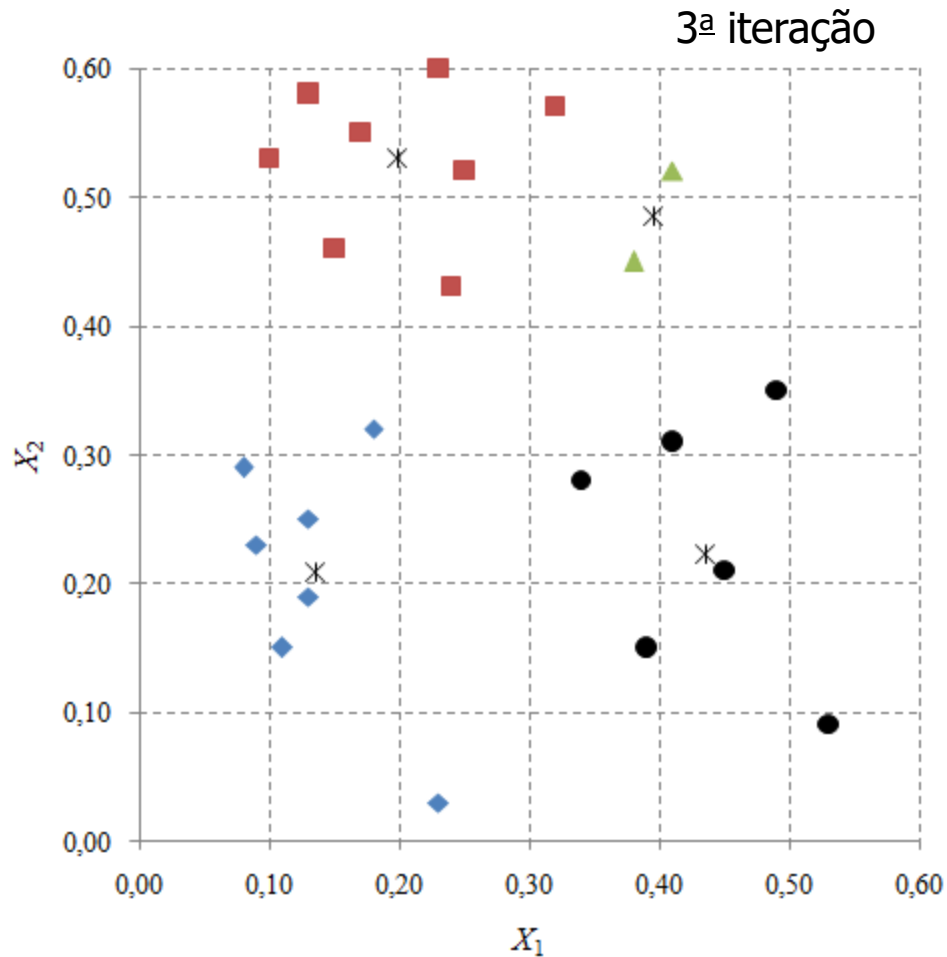
Exemplo 2D: $k = 4$, distância euclidiana



	X_1	X_2
1	0,08	0,29
2	0,09	0,23
3	0,11	0,15
4	0,10	0,53
5	0,13	0,19
6	0,13	0,25
7	0,13	0,58
8	0,15	0,46
9	0,17	0,55
10	0,18	0,32
11	0,23	0,03
12	0,23	0,60
13	0,24	0,43
14	0,25	0,52
15	0,32	0,57
16	0,34	0,28
17	0,38	0,45
18	0,39	0,15
19	0,41	0,52
20	0,41	0,31
21	0,45	0,21
22	0,49	0,35
23	0,53	0,09

K-médias

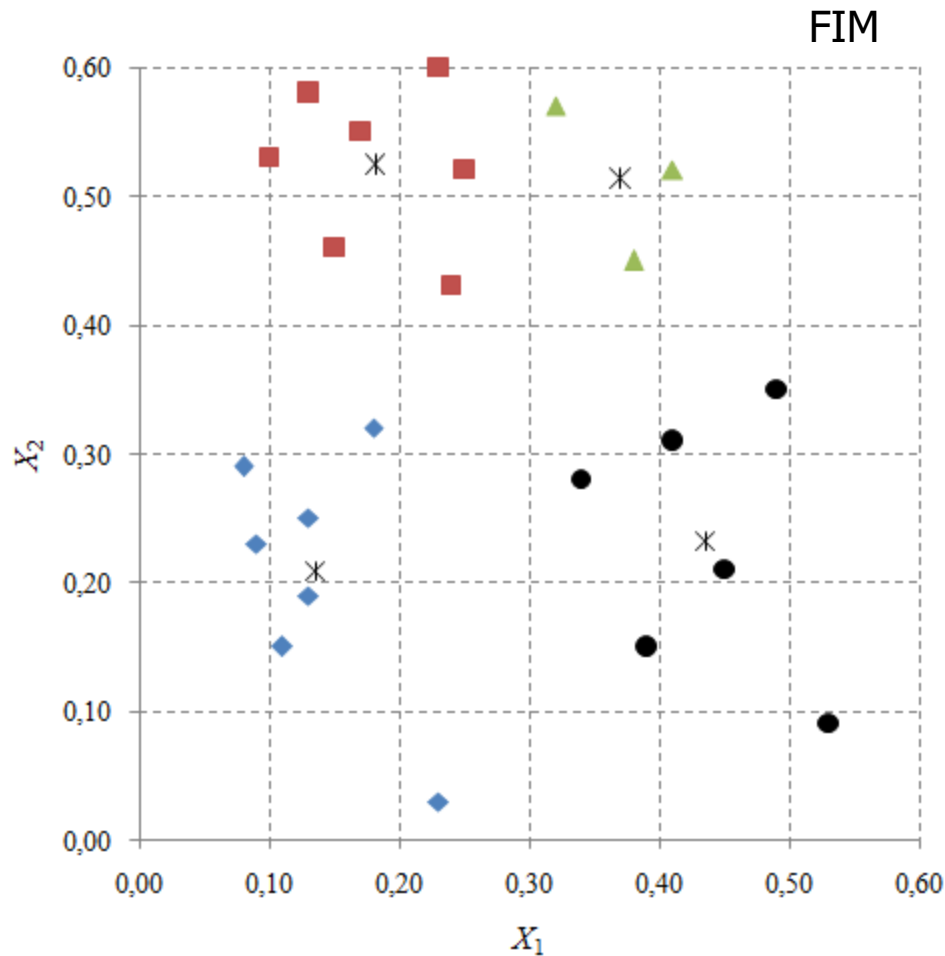
Exemplo 2D: $k = 4$, distância euclidiana



	X_1	X_2
1	0,08	0,29
2	0,09	0,23
3	0,11	0,15
4	0,10	0,53
5	0,13	0,19
6	0,13	0,25
7	0,13	0,58
8	0,15	0,46
9	0,17	0,55
10	0,18	0,32
11	0,23	0,03
12	0,23	0,60
13	0,24	0,43
14	0,25	0,52
15	0,32	0,57
16	0,34	0,28
17	0,38	0,45
18	0,39	0,15
19	0,41	0,52
20	0,41	0,31
21	0,45	0,21
22	0,49	0,35
23	0,53	0,09

K-médias

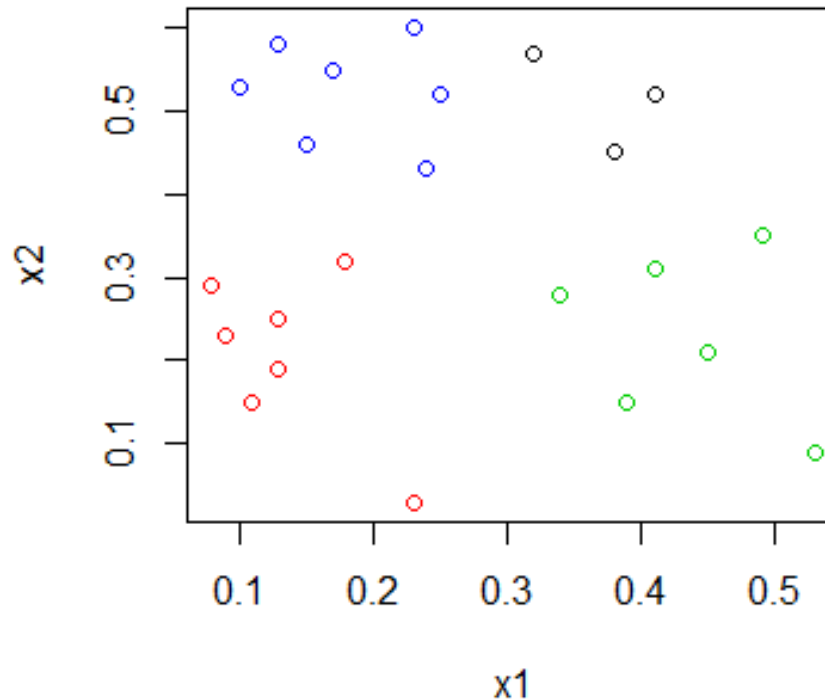
Exemplo 2D: $k = 4$, distância euclidiana



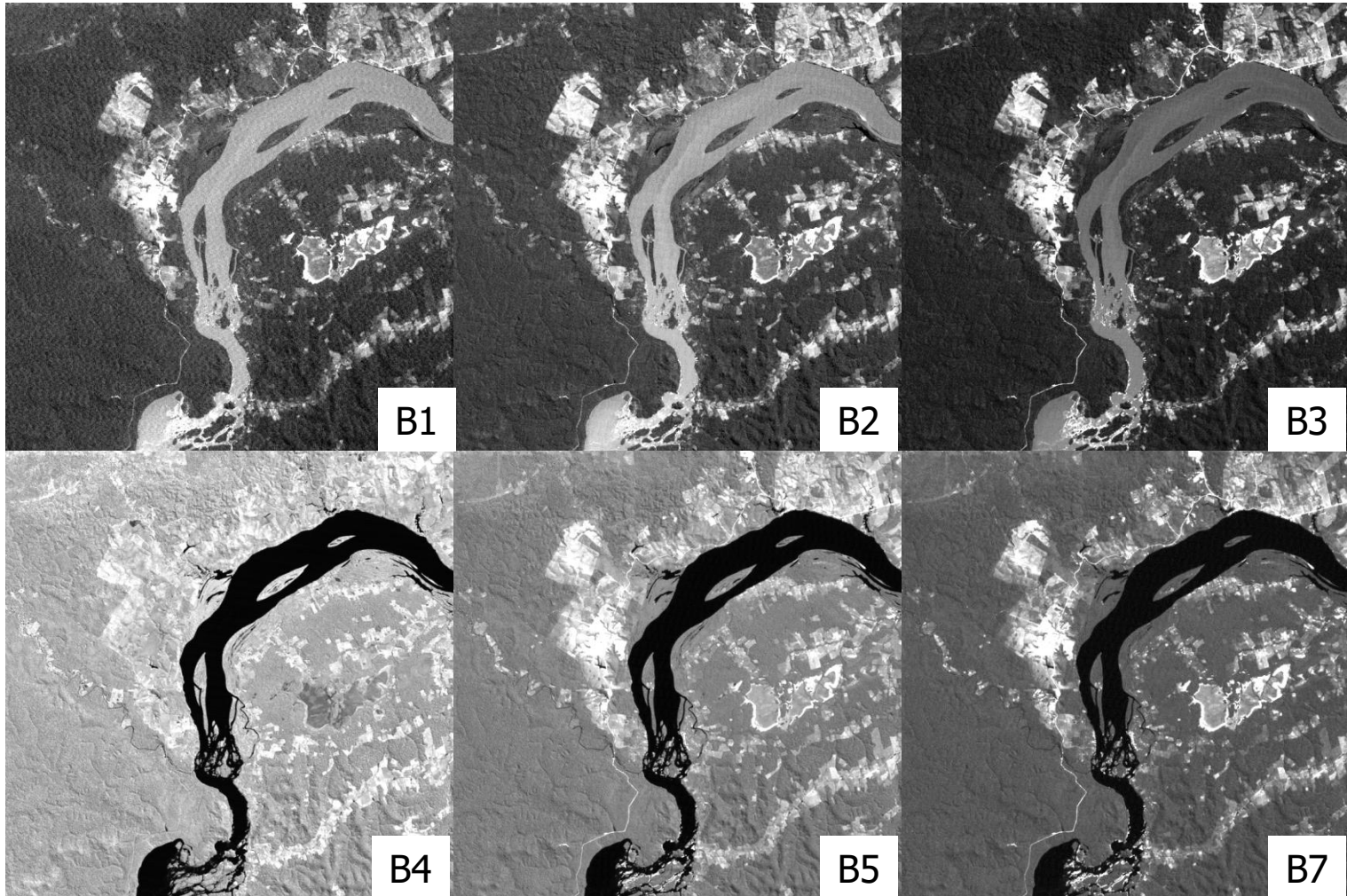
	X_1	X_2
1	0,08	0,29
2	0,09	0,23
3	0,11	0,15
4	0,10	0,53
5	0,13	0,19
6	0,13	0,25
7	0,13	0,58
8	0,15	0,46
9	0,17	0,55
10	0,18	0,32
11	0,23	0,03
12	0,23	0,60
13	0,24	0,43
14	0,25	0,52
15	0,32	0,57
16	0,34	0,28
17	0,38	0,45
18	0,39	0,15
19	0,41	0,52
20	0,41	0,31
21	0,45	0,21
22	0,49	0,35
23	0,53	0,09

K-médias no R

```
x1<-c(0.08,0.09,0.11,0.10,0.13,0.13,0.13,0.15,0.17,0.18,0.23,0.23,0.24,0.25,0.32,0.34,0.38,0.39,0.41,0.41,0.45,0.49,0.53)
x2<-c(0.29,0.23,0.15,0.53,0.19,0.25,0.58,0.46,0.55,0.32,0.03,0.6,0.43,0.52,0.57,0.28,0.45,0.15,0.52,0.31,0.21,0.35,0.09)
data<-cbind(x1,x2)
fit<-kmeans(data,4)
plot(data,col=fit$cluster)
```

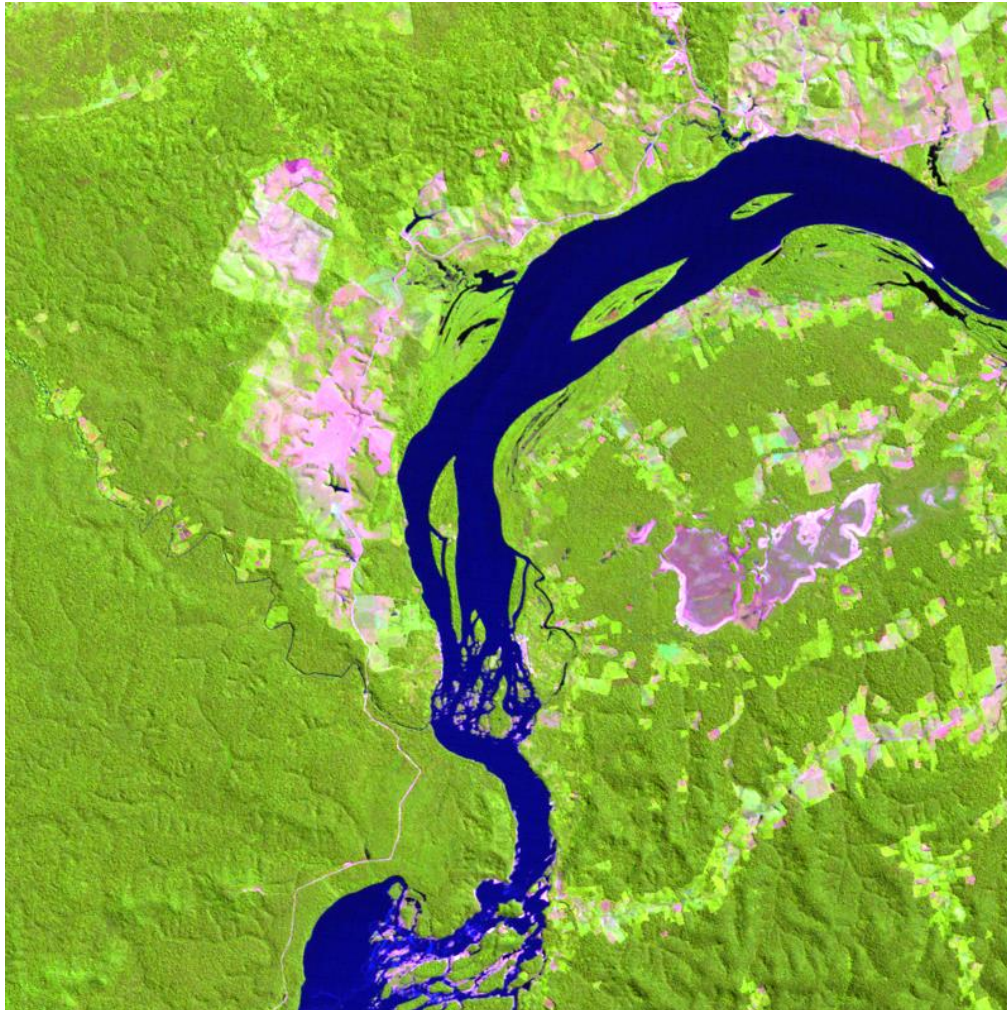


Exemplo K-médias em imagens



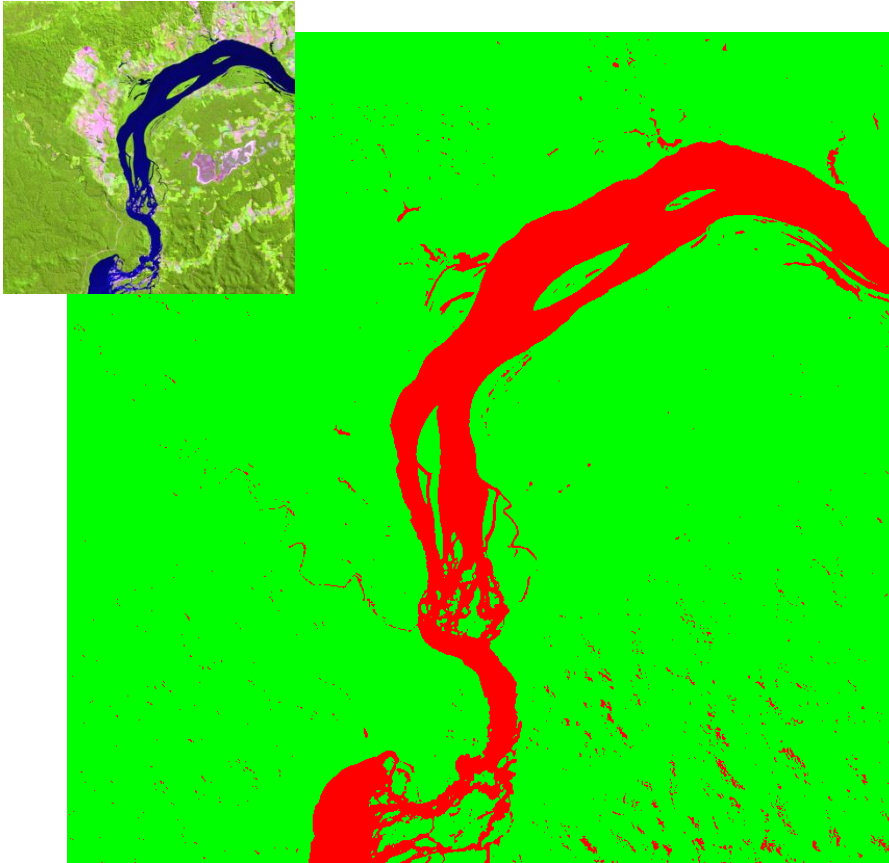
TM/Landsat

Exemplo K-médias em imagens

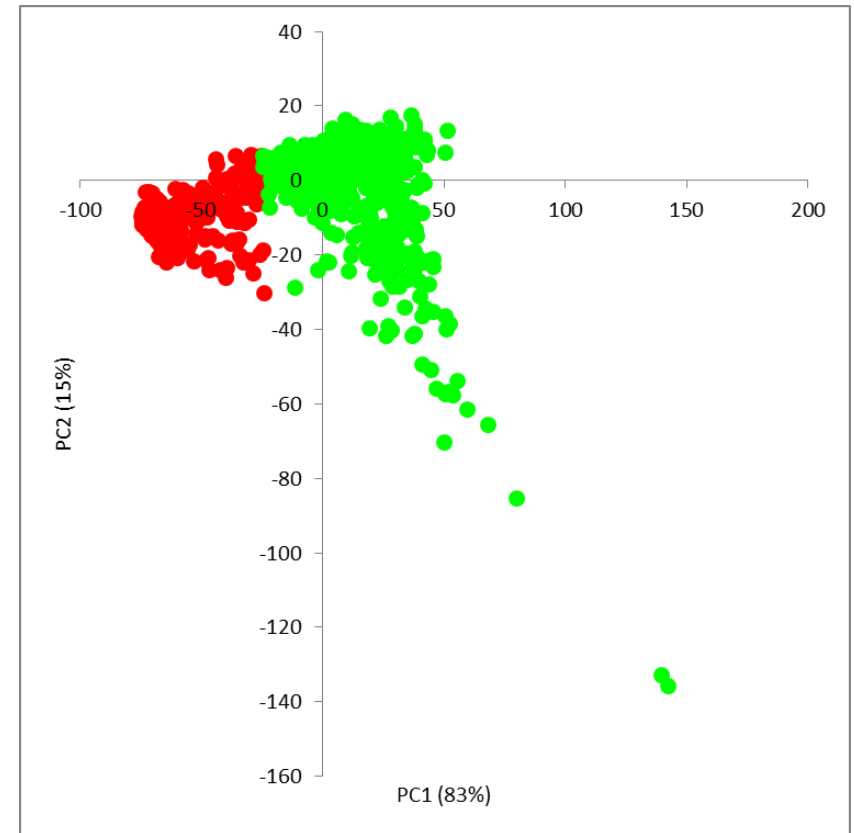


TM/Landsat 5R4G3B

Exemplo K-médias em imagens

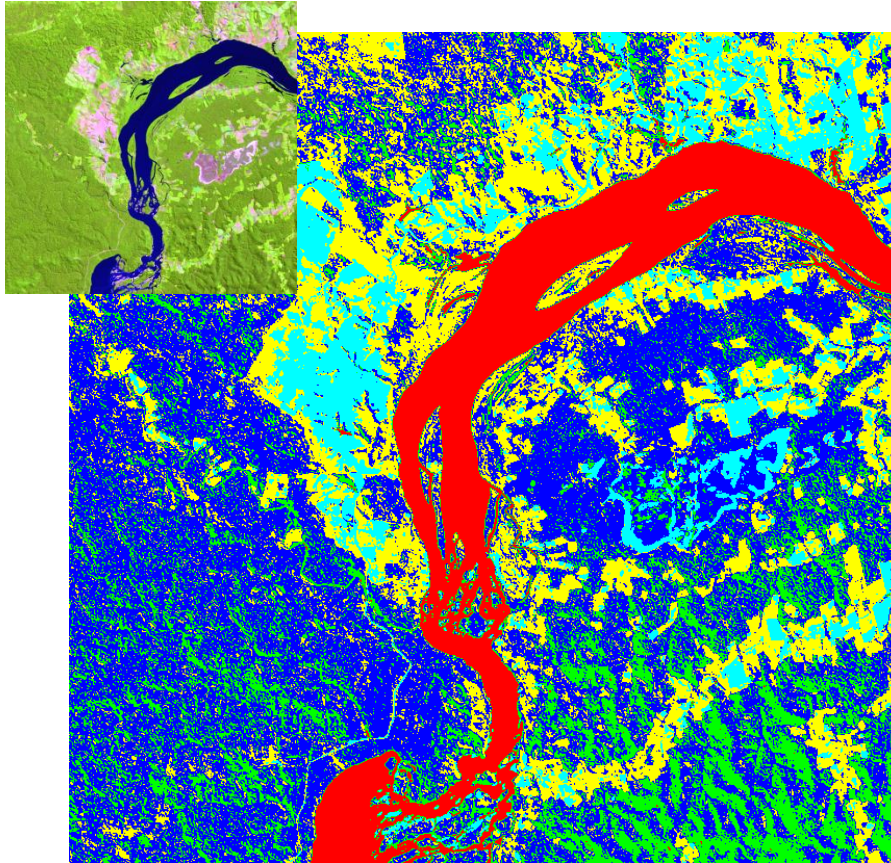


K-médias (6 bandas)
2 classes
10 iterações

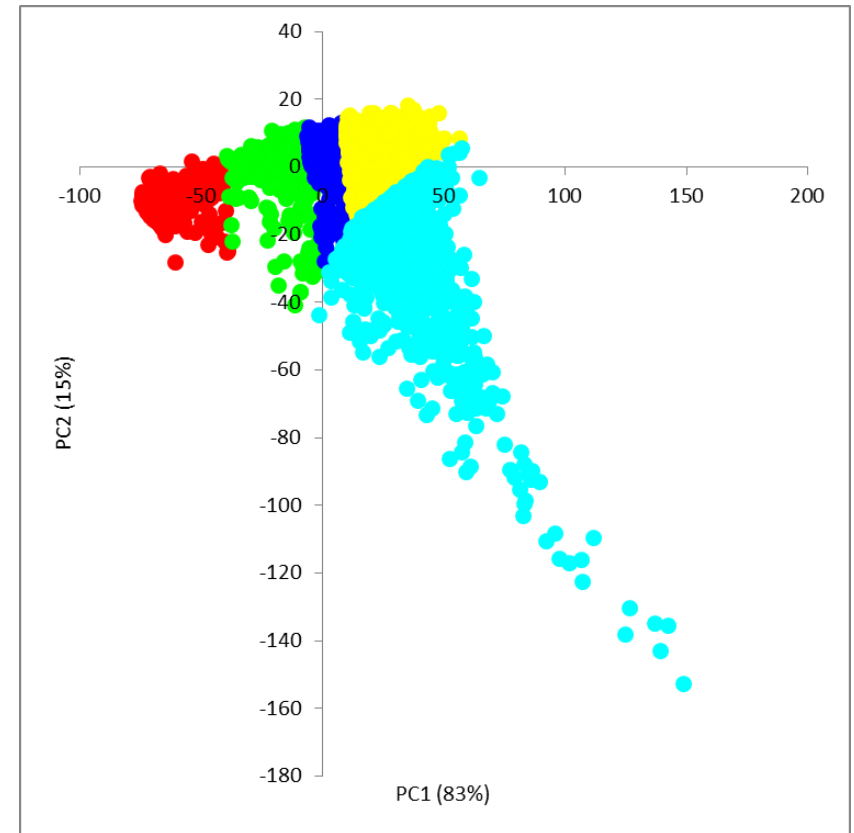


1000 amostras por classe

Exemplo K-médias em imagens

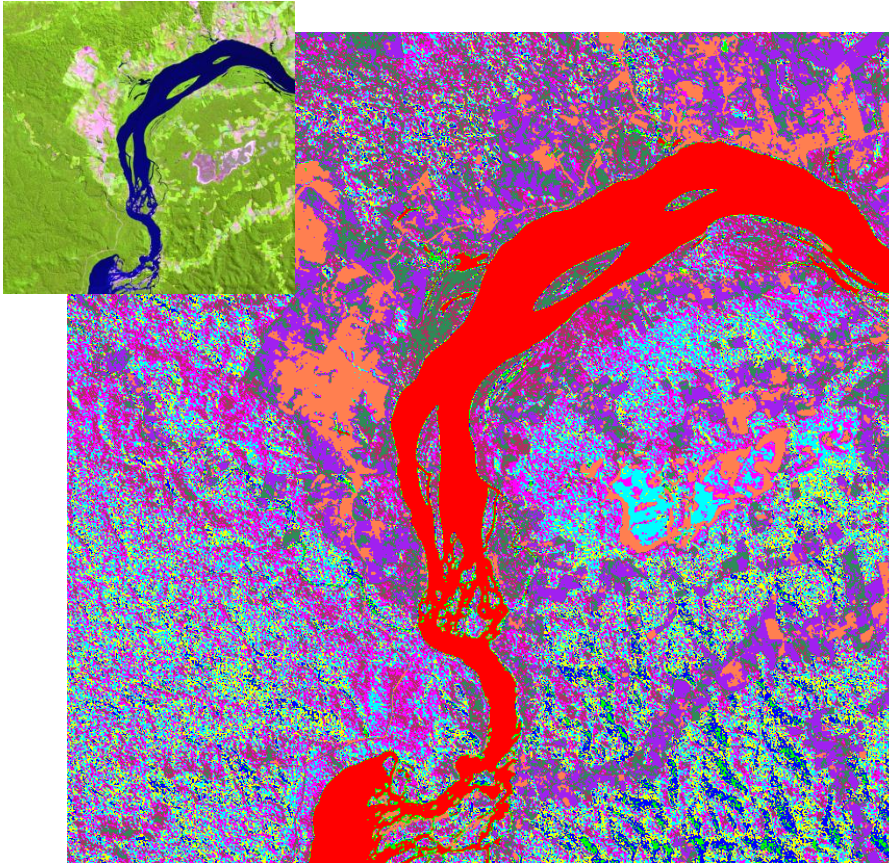


K-médias (6 bandas)
5 classes
10 iterações

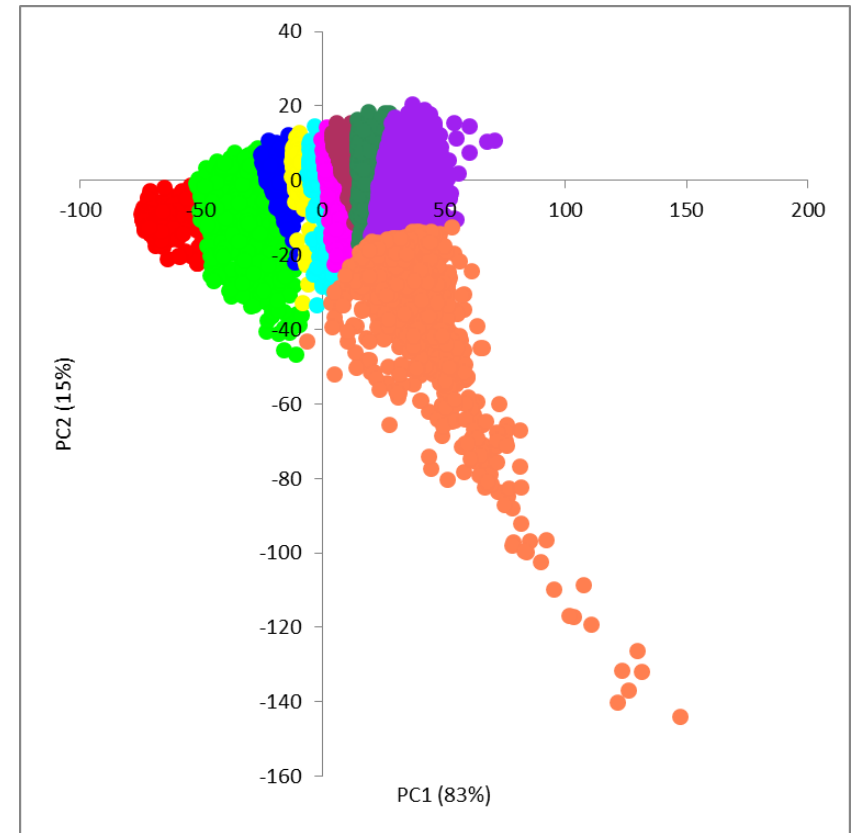


1000 amostras por classe

Exemplo K-médias em imagens



K-médias (6 bandas)
10 classes
10 iterações



1000 amostras por classe

Considerações finais

Normalização das variáveis

- indicado quando as variáveis têm diferentes unidades ou variações muito diferentes
- cuidado ao utilizar dados amostrais para se fazer a normalização
é aconselhável sempre apresentar uma tabela com os valores utilizados!

Método hierárquico

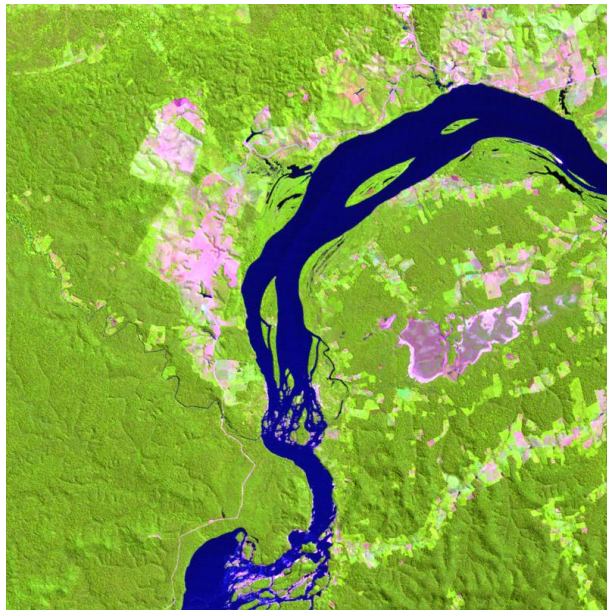
- adequado quando o número de elementos a agrupar é relativamente pequeno
- a análise do dendrograma pode indicar o número de grupos ideal

Método por Particionamento

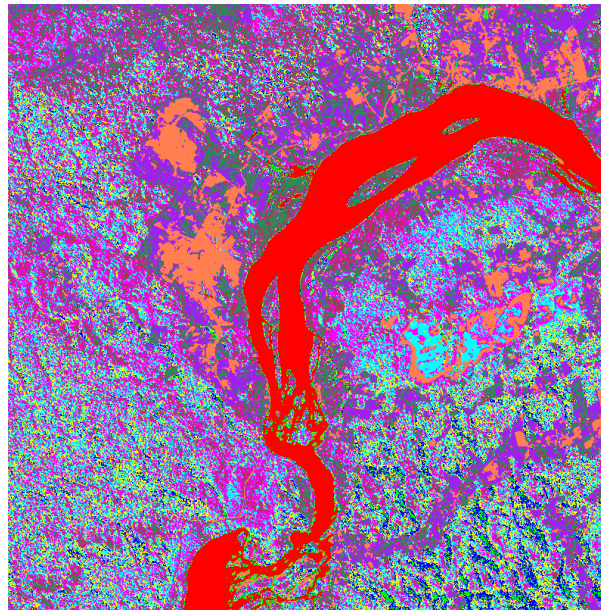
- indicado quando o número de elementos a agrupar é muito grande (imagens, p. ex.)
- o número de grupos ideal depende de uma avaliação posterior
- numa classificação de imagens, em geral, cada pixel é classificado independentemente
- há implementações que consideram as relações espaciais!

Considerações finais

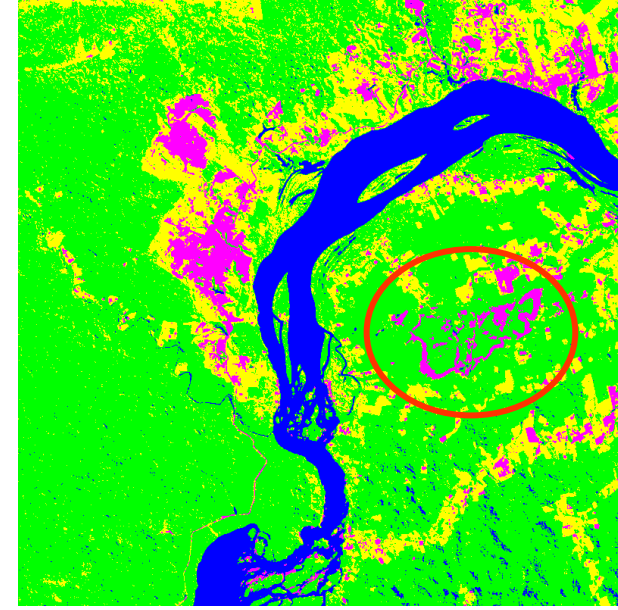
Independente do método, para gerar uma classificação final, há a necessidade de se atribuir a cada grupo (classe) uma classe temática de interesse.



TM/Landsat 5R4G3B



K-médias (6 bandas)
10 classes
10 iterações



Classificação Final

- floresta
- regeneração/pastagem
- solo exposto
- água

Considerações finais

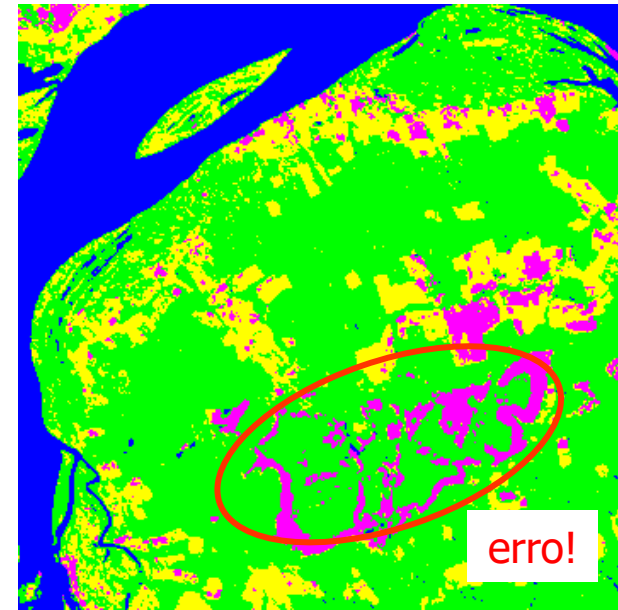
Independente do método, para gerar uma classificação final, há a necessidade de se atribuir a cada grupo (classe) uma classe temática de interesse.



TM/Landsat 5R4G3B



K-médias (6 bandas)
10 classes
10 iterações



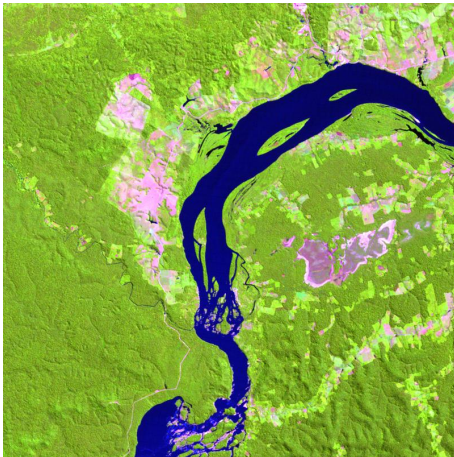
Classificação Final

- floresta
- regeneração/pastagem
- solo exposto
- água

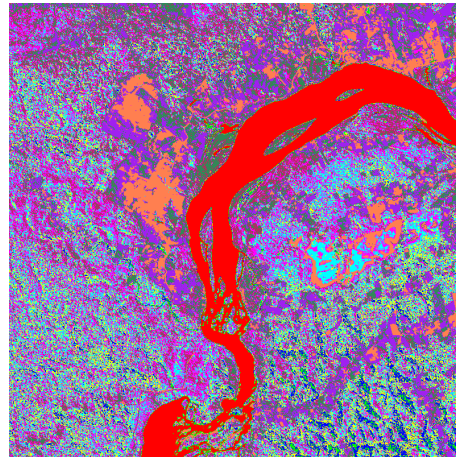
Solução: aumentar K?

Considerações finais

A classificação K-Médias pode ser usada para avaliar previamente a diversidade espectral da área de estudo e para orientar a definição das classes e escolha das amostras de treinamento que serão utilizadas no classificador supervisionado



TM/Landsat 5R4G3B



K-médias (6 bandas)
10 classes
10 iterações

Idealmente escolhe-se um valor para K muito superior ao número de classes de interesse

É interessante avaliar as propriedades estatísticas de cada grupo de modo a verificar a necessidade da definição de subclasses que ajudarão no desempenho do classificador supervisionado