Análise de agrupamento
CLUSTER

Marília M. Favalesso

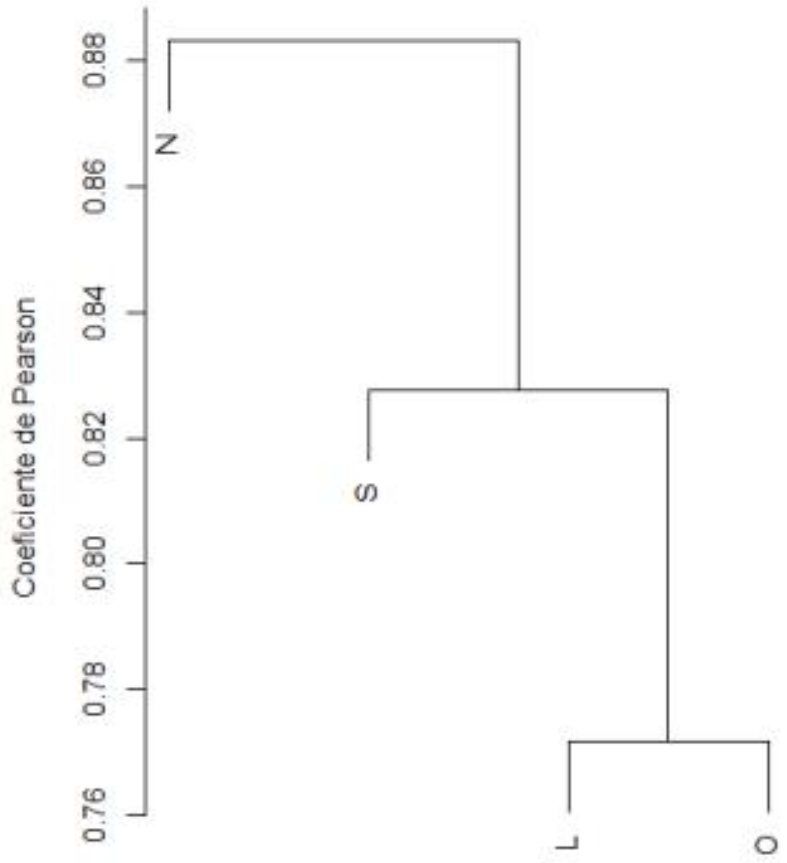
Thaís Maylin Sobjak

Mestranda do Programa de Conservação e Manejo de Recursos Naturais

Universidade Estadual do Oeste do Paraná

Análise de agrupamento

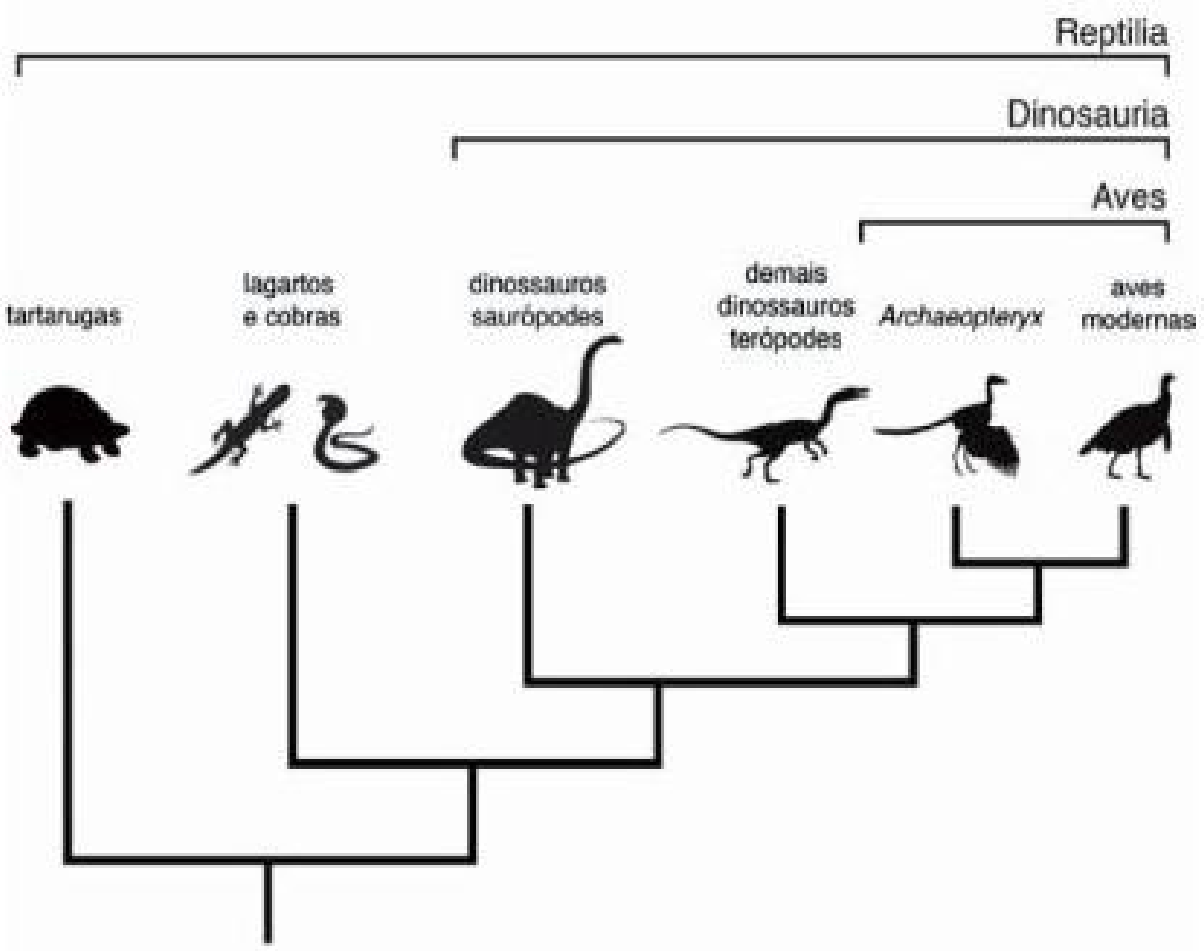
- Envolve **categorização** - dividir um grande número de observações em grupos menores
- As categorias são criadas tendo como base **medidas de similaridades/dissimilaridade**
- Resultado - variáveis em **escala nominal** - pertence/não pertence
- Var-quanti e var-quali
- Abordagem: Métodos hierarquicos



Exemplo Cluster

Exemplos de aplicações potênciais

- Taxonomia numérica
  - Evolução
  - Ecologia



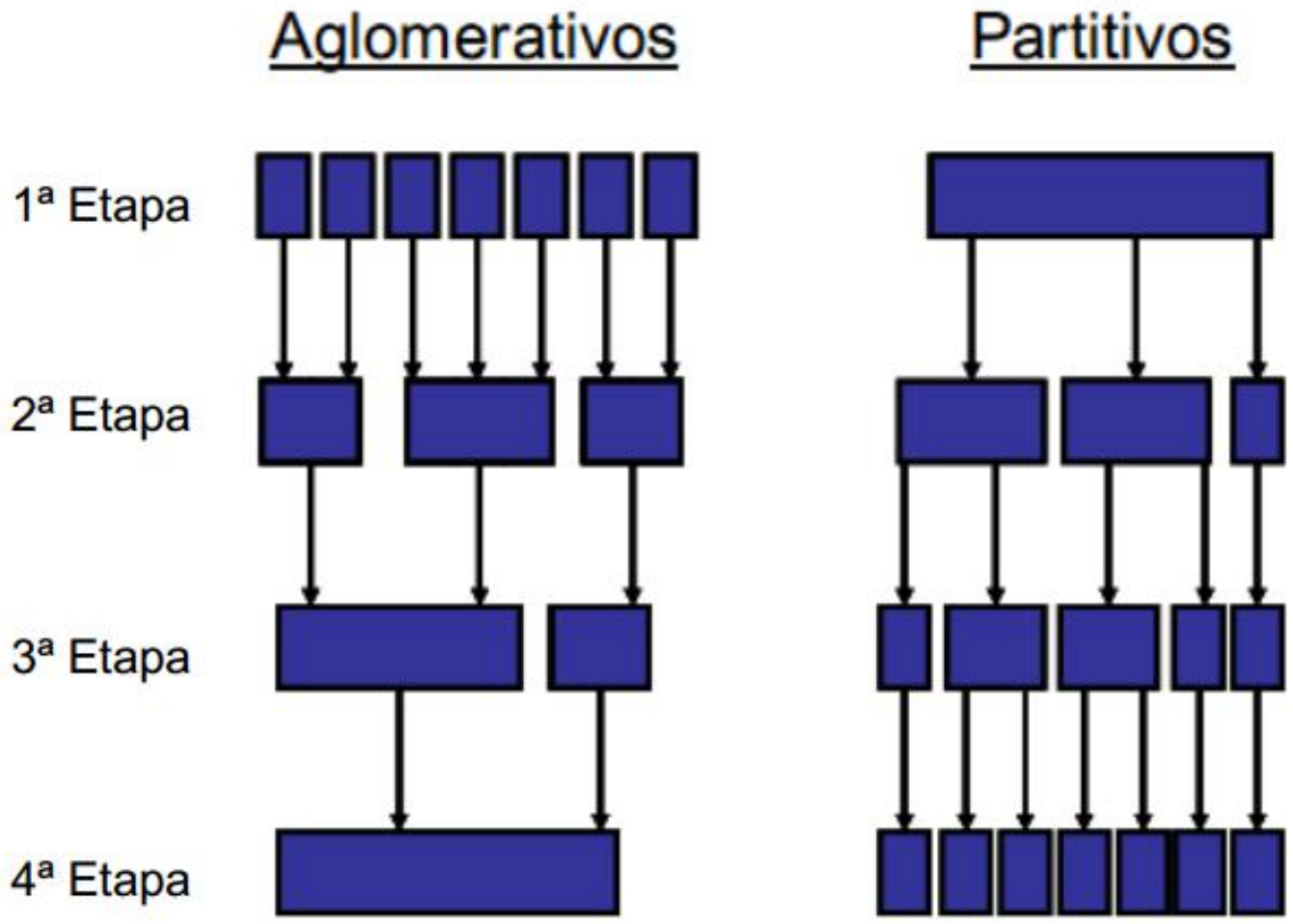
Medidas de distância

↑ Similaridade ↑ igualdade entre os grupos

↑ Dissimilaridade ↓ igualdade entre os grupos

Quais são as medidas?

Métodos hierárquicos



Quando parar?

↓

Os algoritmos de agrupamento não apresentam solução para determinação do número ideal de grupos

↓

Avaliar o número de grupos (mais grupos, maior homogeneidade)

↓

Coeficiente cofenétic (depois)

Método hierárquico aglomerativo

Método de ligação simples ou do vizinho mais próximo (*Single linkage*)

Método da ligação completa ou do vizinho mais distante (*Complete linkage*)

Método da ligação média (*Average linkage*)

Exemplo

Espécies	A	B	C	D	E
Local 1	0	0	10	8	0
Local 2	0	0	12	9	0
Local 3	0	0	13	5	10
Local 4	2	3	0	4	12
Local 5	5	10	0	0	16
Local 6	15	20	0	0	0

Abundância de cinco espécies amostradas em seis diferentes localidades

Exemplo

Espécies	A	B	C	D	E
Local 1	0	0	10	8	0
Local 2	0	0	12	9	0
Local 3	0	0	13	5	10
Local 4	2	3	0	4	12
Local 5	5	10	0	0	16
Local 6	15	20	0	0	0

Abundância de seis espécies amostradas em cinco diferentes localidades

Exemplo

Espécies	A	B	C	D	E
Local 1	0	0	10	8	0
Local 2	0	0	12	9	0
Local 3	0	0	13	5	10

Distância euclidiana

$$di,j = \sqrt{(yi,1 - yj,1)^2 + (yi,2 - yj,2)^2}$$

$$d1,2 = \sqrt{(0-0)^2 + (0-0)^2} + (10-12)^2 + (8-9)^2 + (0-0)^2$$

$$d1,2 = 2,24$$

$$d1,3 = \sqrt{(0-0)^2 + (0-0)^2} + (10-13)^2 + (8-5)^2 + (0-10)^2$$

$$d1,3 = 10,86$$

Exemplo

Espécies	A	B	C	D	E
Local 1	0	0	10	8	0
Local 2	0	0	12	9	0
Local 3	0	0	13	5	10

Distância euclidiana

$$di,j = \sqrt{(yi,1 - yj,1)^2 + (yi,2 - yj,2)^2}$$

$$d2,3 = \sqrt{(0-0)^2 + (0-0)^2} + (12-13)^2 + (9-5)^2 + (0-10)^2$$

$$d1,2 = 10,82$$

Exemplo

	Local 2	**Local 3**
Local 1	2,24	10,86
Local 2		10,82

Resultado do exemplo de distância euclidiana

As vezes é melhor padronizar!

↓

$$Z = \frac{(Yi - \bar{Y})}{S}$$

Transformação Z

Exemplo

	Local 1	**Local 2**	**Local 3**	**Local 4**	**Local 5**	**Local 6**
Local 1	0,00	2,24	10,86	16,52	23,35	28,09
Local 2	2,24	0,00	10,82	18,06	24,62	29,15
Local 3	10,86	10,82	0,00	13,67	18,84	30,32
Local 4	16,52	18,06	13,67	0,00	9,49	24,86
Local 5	23,35	24,62	18,84	9,49	0,00	21,35
Local 6	28,09	29,15	30,32	24,86	21,35	0,00

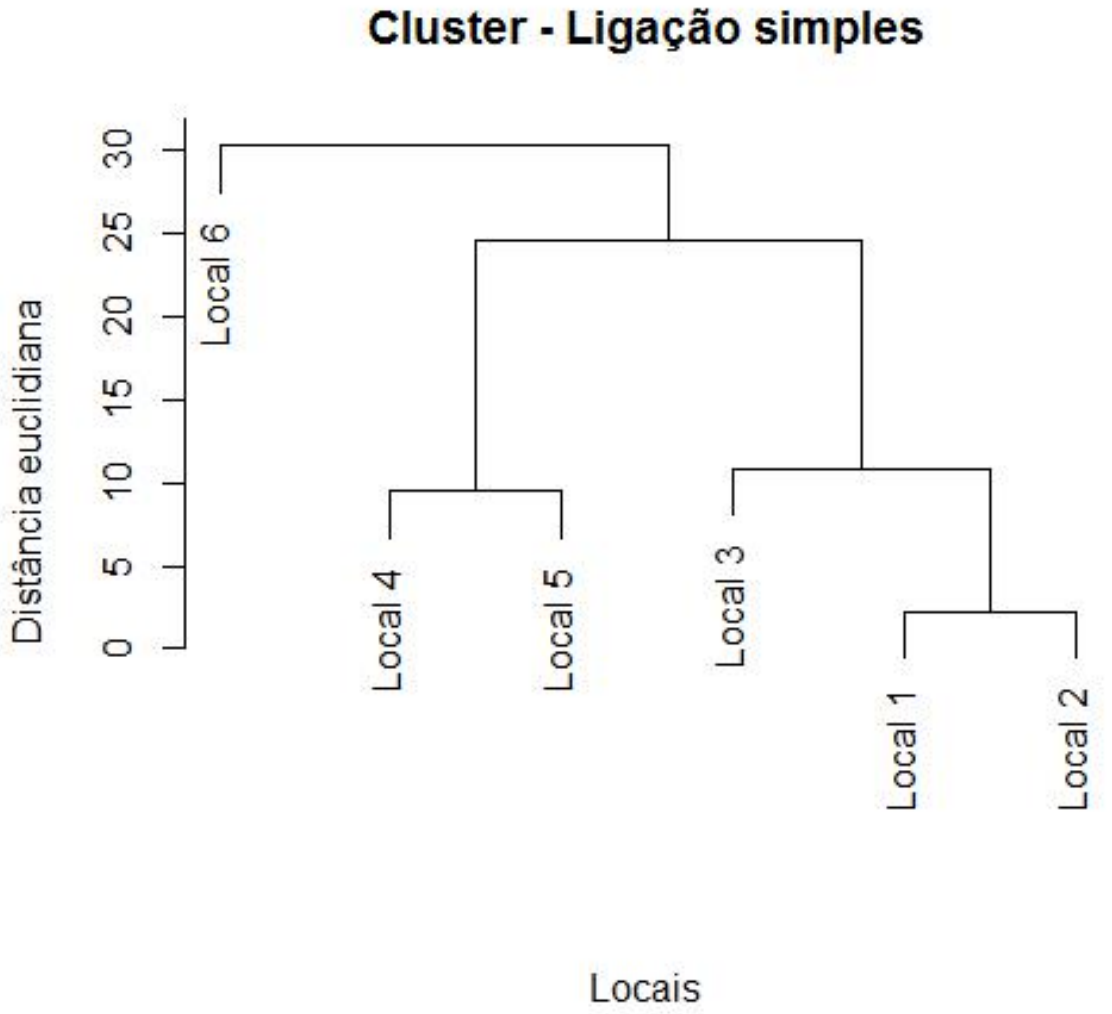
Matriz de distância euclidiana (com variáveis padronizadas)

Exemplo

	Local 2	Local 3	Local 4	Local 5	Local 6
Local 1	**2,24**	10,86	16,52	23,35	28,09
Local 2		10,82	18,06	24,62	29,15
Local 3			13,67	18,84	30,32
Local 4				9,49	24,86
Local 5					21,35

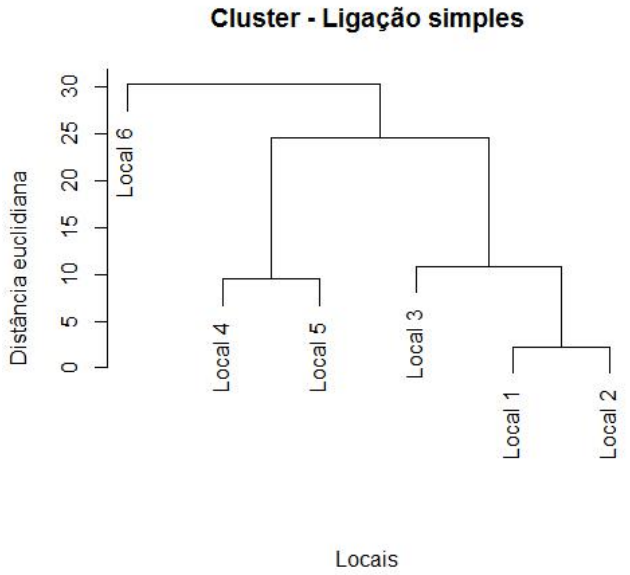
+ **Método do vizinho mais próximo (“single linkage”)**

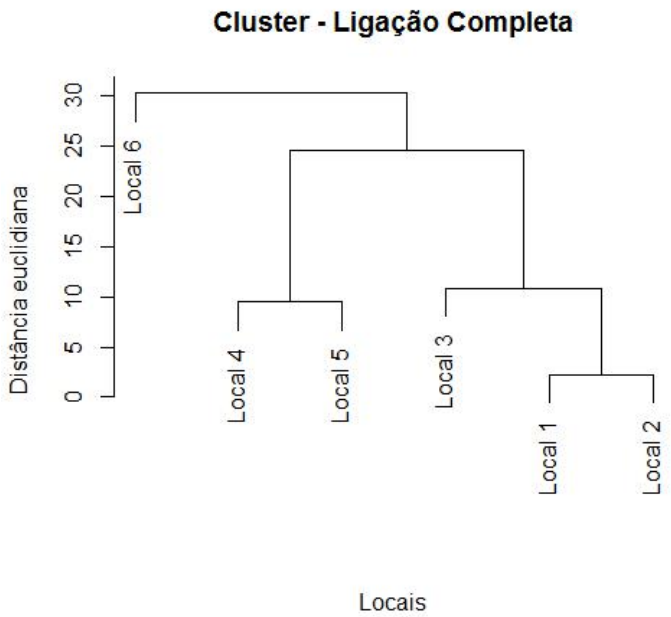
Exemplo

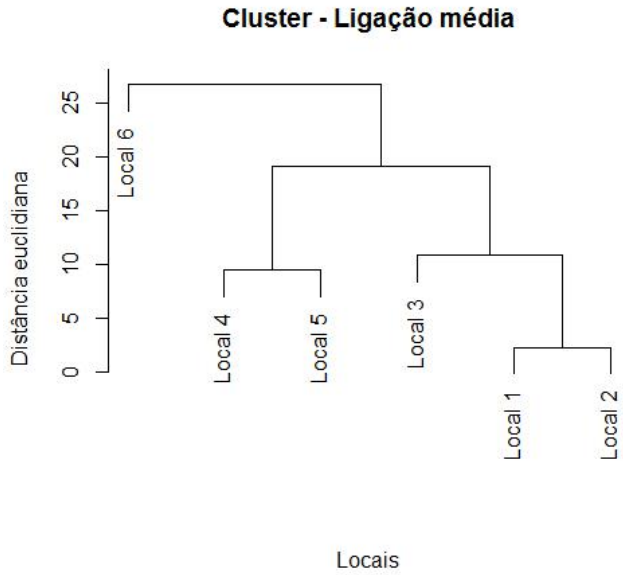


E os outros métodos de ligação?

Exemplo


Cluster - Ligação simples
Distância euclidiana
0
5
10
15
20
25
30
Local 6
Local 4
Local 5
Local 3
Local 1
Local 2
Locais


Cluster - Ligação Completa
Distância euclidiana
0
5
10
15
20
25
30
Local 6
Local 4
Local 5
Local 3
Local 1
Local 2
Locais


Cluster - Ligação média
Distância euclidiana
0
5
10
15
20
25
Local 6
Local 4
Local 5
Local 3
Local 1
Local 2
Locais

Validação do agrupamento

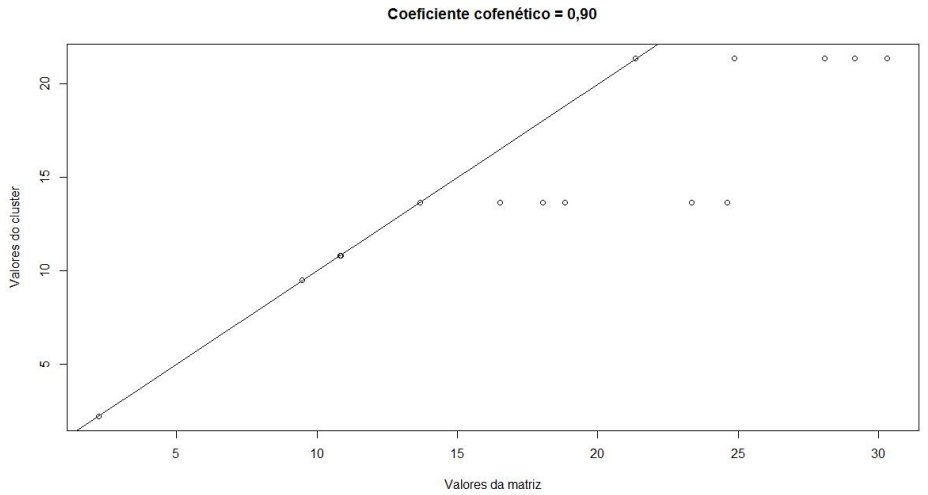
Coeficiente de correlação cofenética (ccc)

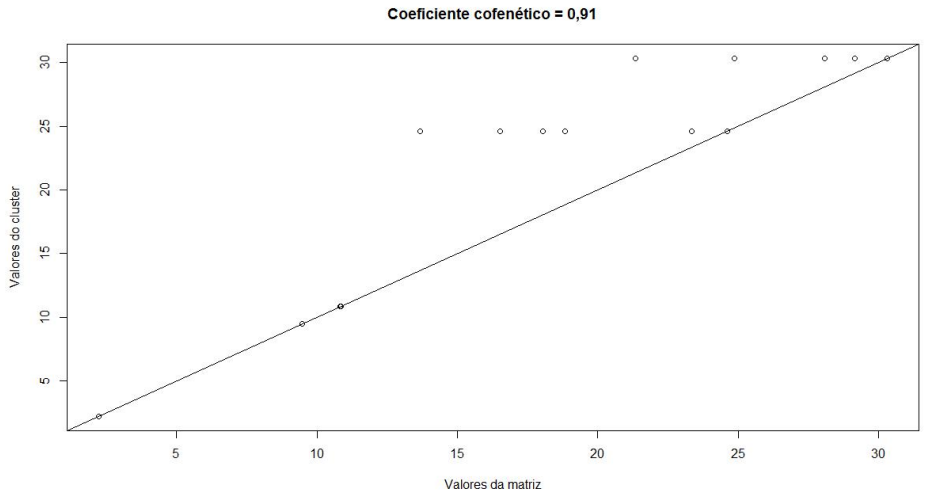
Avaliar o grau de similaridade da matriz de distâncias dos dados originais preservados

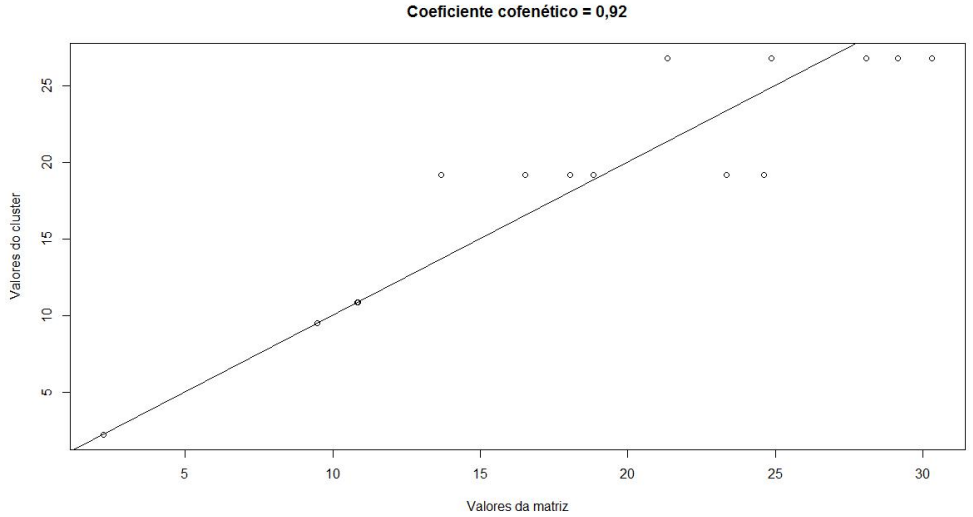
↓

Correlação linear

Validação do agrupamento







Idealmente > 0,80
(Hair, 2009)

Interpretação do resultado final

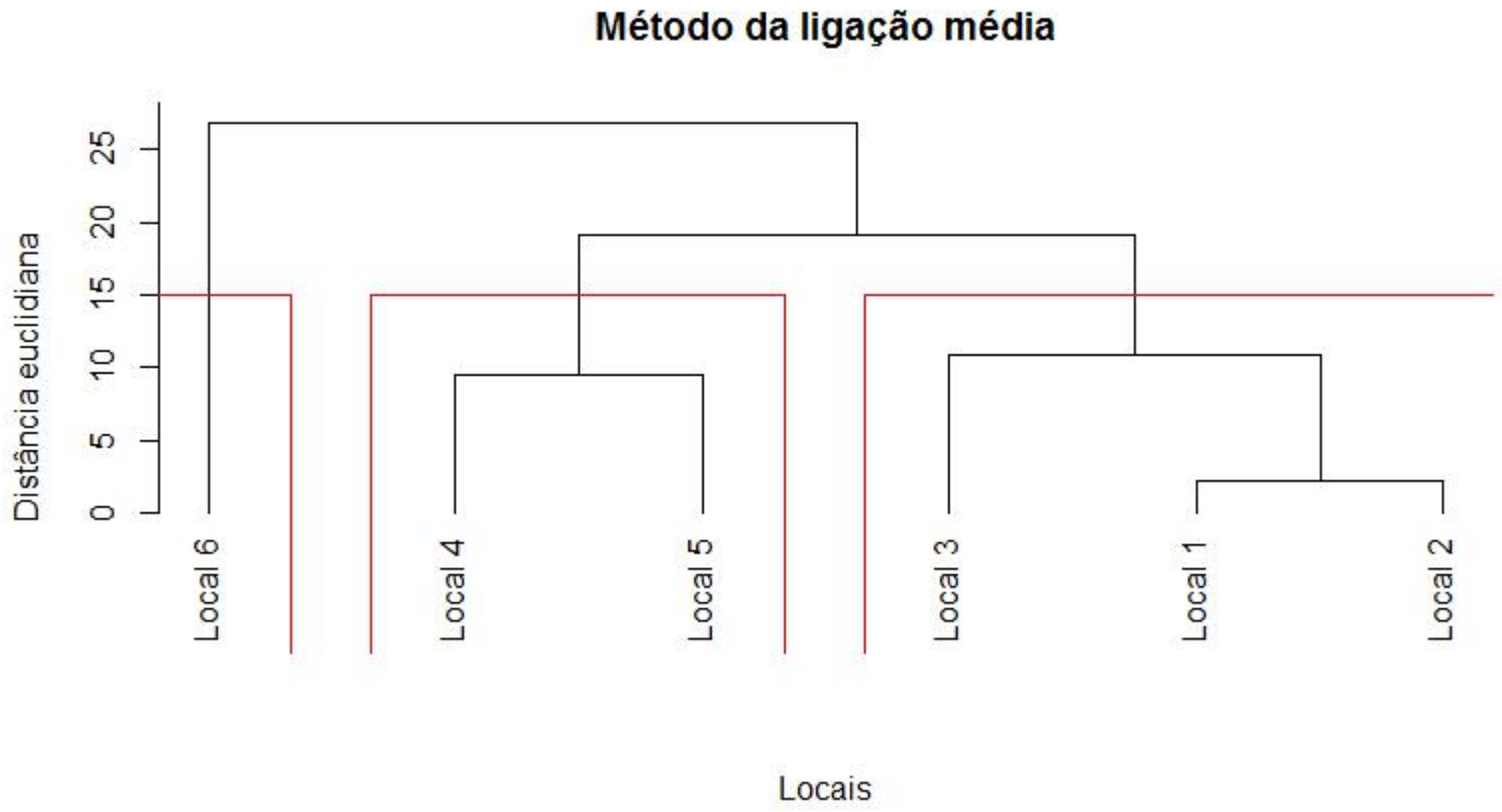


Figura 1 - Agrupamento cluster dos locais amostrados tendo como base as espécies A, B, C, D, E. Coef. Cofenético = 0,92 (P<0,02)