
minha cara de
felicidade de
Memedroid
voltas as aulas

Aula 3: Agrupamento

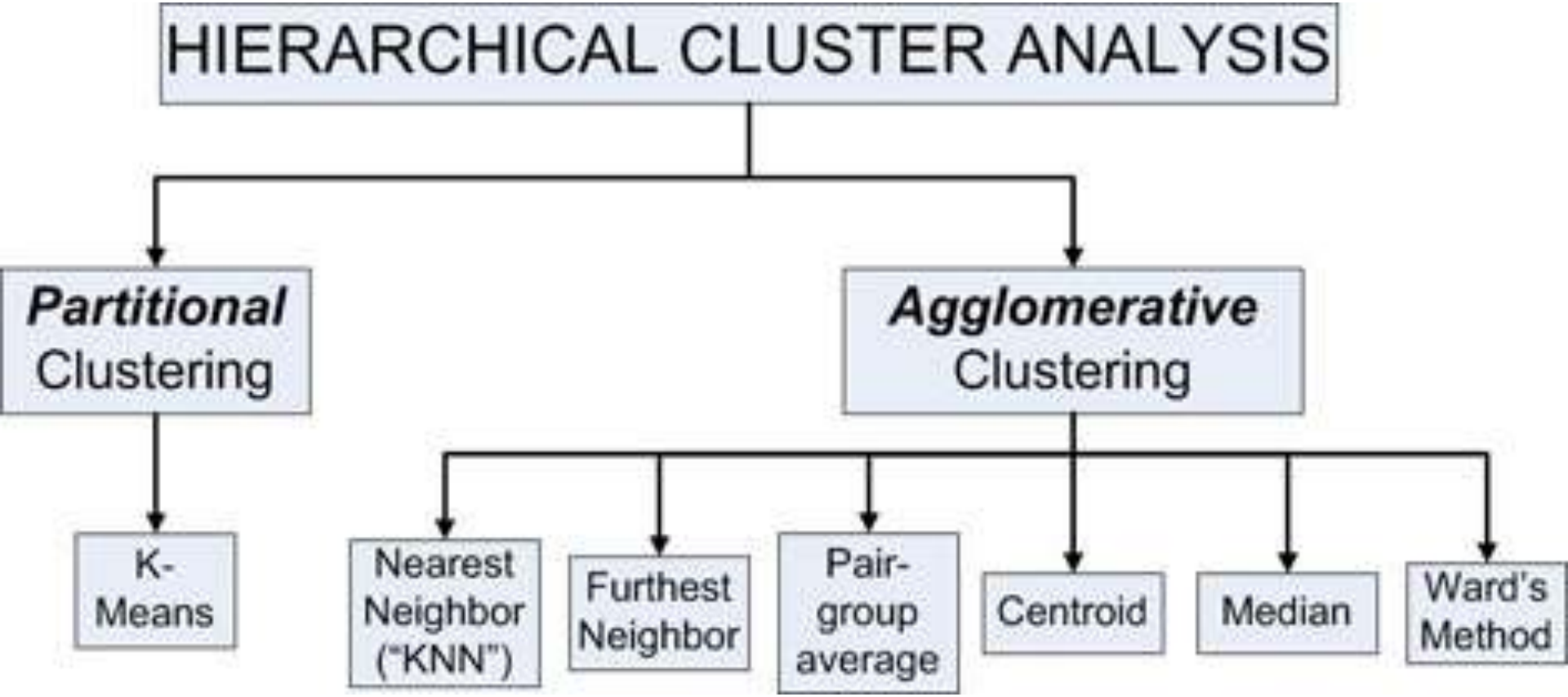
Cap. 8 Legendre & Legendre

Ao final da aula você deverá ter comprendido

- Agrupamento hierárquico: dendrograma
  - Single linkage
  - UPGMA
  - Como escolher o nível de corte para encontrar grupos
  - Como validar o agrupamento
- Agrupamento não-hierárquico: k-means
- Espécies indicadoras
  - IndVal

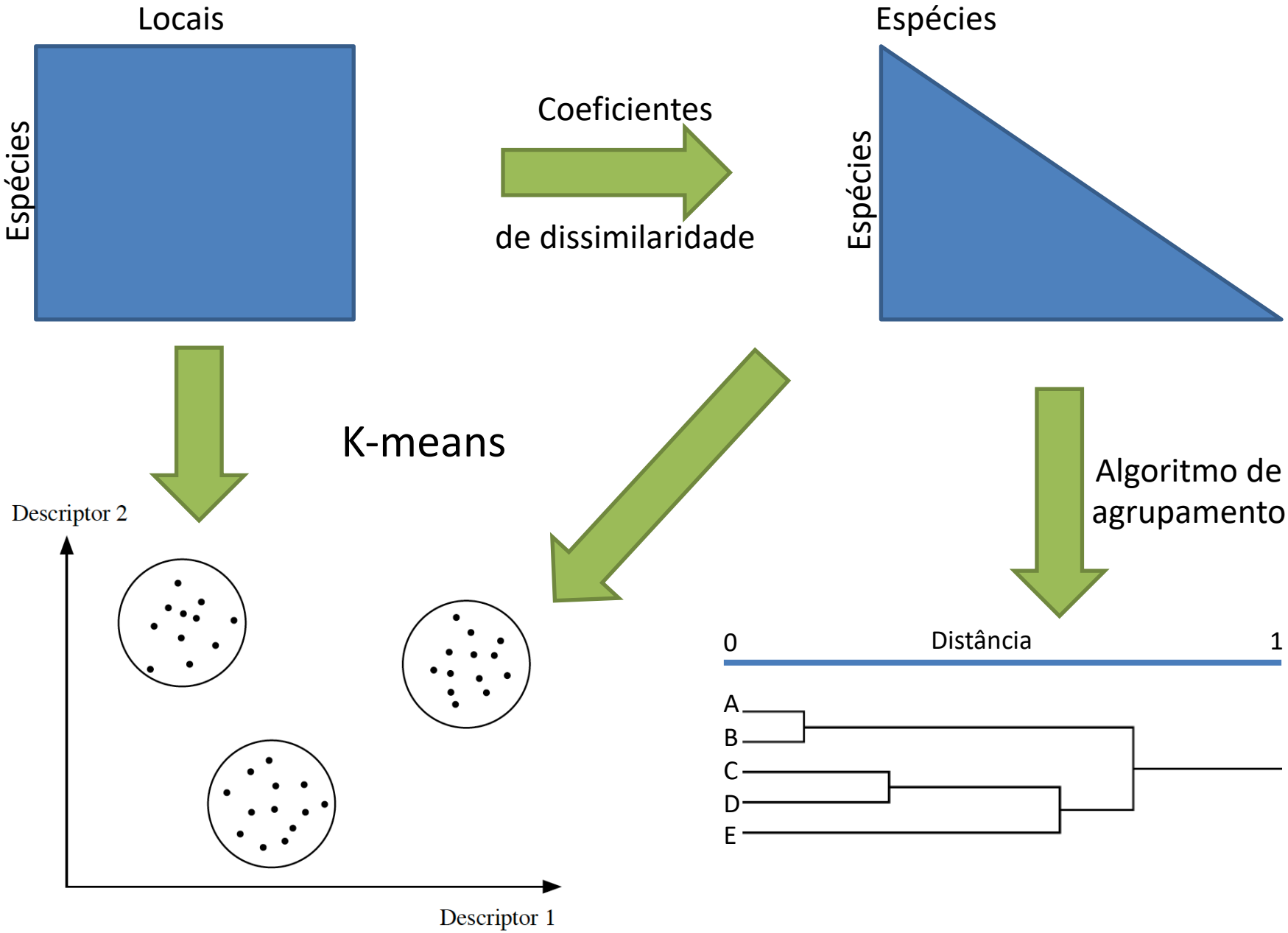
Análise de cluster

- Resumir um grande volume de informação
- Agrupa objetos por grau de similaridade
- **Busca por descontinuidades no conjunto de dados para formar grupos**
- Descontinuidade X Gradiente



http://wiki.eigenvector.com/index.php?title=Cluster

Como os métodos funcionam?


Locais
Espécies
Coeficientes
de dissimilaridade
Espécies
Espécies
K-means
Algoritmo de
agrupamento
Descriptor 2
Descriptor 1
0
Distância
1
A
B
C
D
E

Índices de Similaridade/Distância

- Já falamos sobre eles na aula passada
- Devem ser escolhidos de acordo com o tipo de dado e o objetivo da análise

Algoritmos de agrupamento

- Aglomerativos
  - Single linkage
  - Complete linkage
  - UPGMA
  - Ward

Passo-a-passo de como funciona o single linkage

Comecemos pela matriz de composição de espécies

Species	Ponds				
	212	214	233	431	432
1	3	3	0	0	0
2	0	0	2	2	0
3	0	2	3	0	2
4	0	0	4	3	3
5	4	4	0	0	0
6	0	2	0	3	3
7	0	0	0	1	2
8	3	3	0	0	0

Baseado no Item 8.2 L&L

Passo-a-passo de como funciona o single linkage

Cálculo da matriz de similaridade usando S_{20} , aqui já convertido pra distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

Passo-a-passo de como funciona o single linkage

Ranquear os pares de objetos em ordem crescente de distância

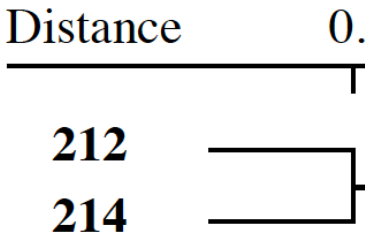
Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.400	—			
233	1.000	0.929	—		
431	1.000	0.937	0.700	—	
432	1.000	0.786	0.800	0.500	—

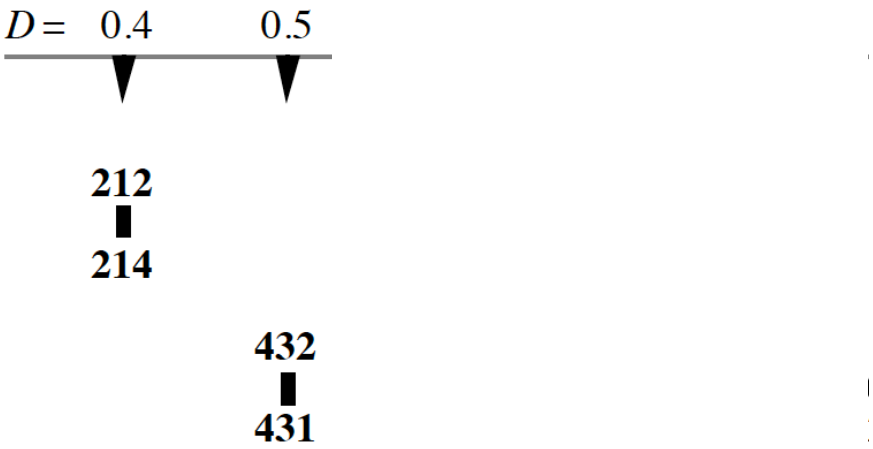
Passo-a-passo de como funciona o single linkage

The first clustering step consists in rewriting the distance values in increasing order:

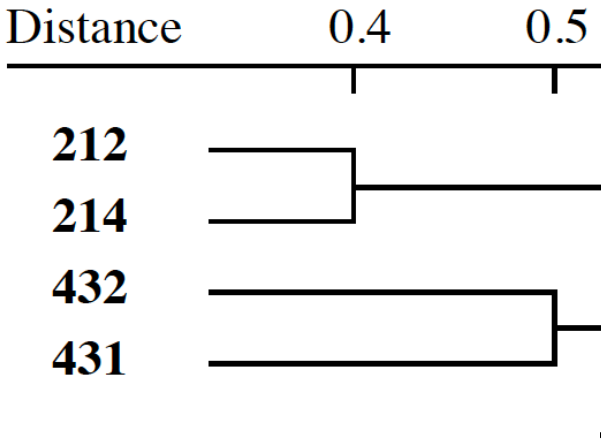
$D = 1 - S_{20}$	Pairs formed
0.400	212-214
0.500	431-432
0.700	233-431
0.786	214-432
0.800	233-432
0.929	214-233
0.937	214-431
1.000	212-233
1.000	212-431
1.000	212-432

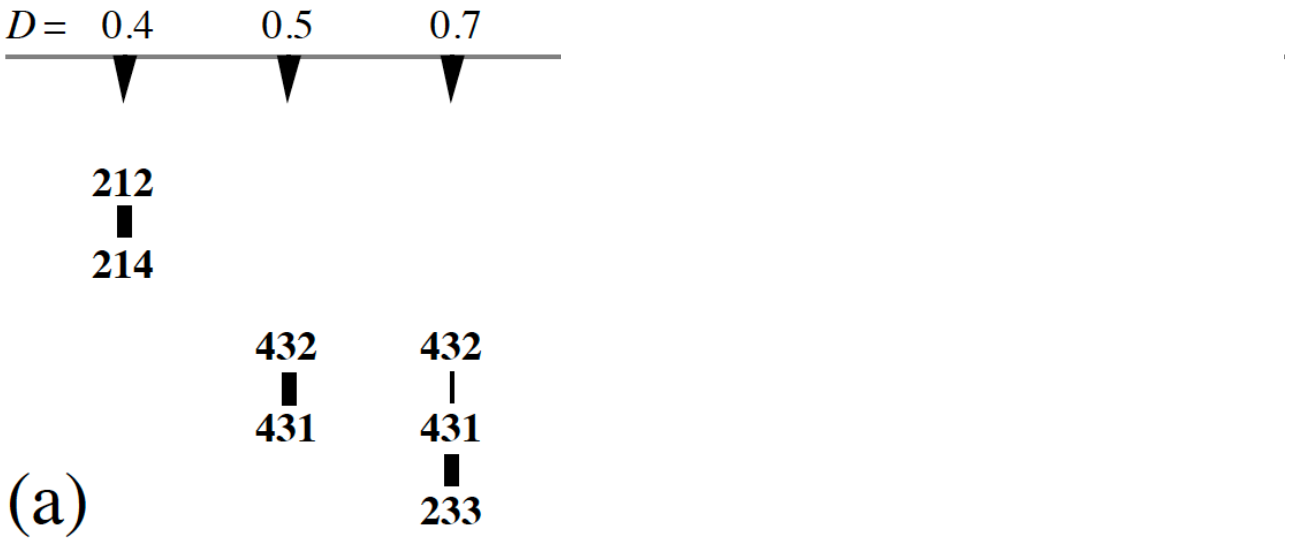
(a)



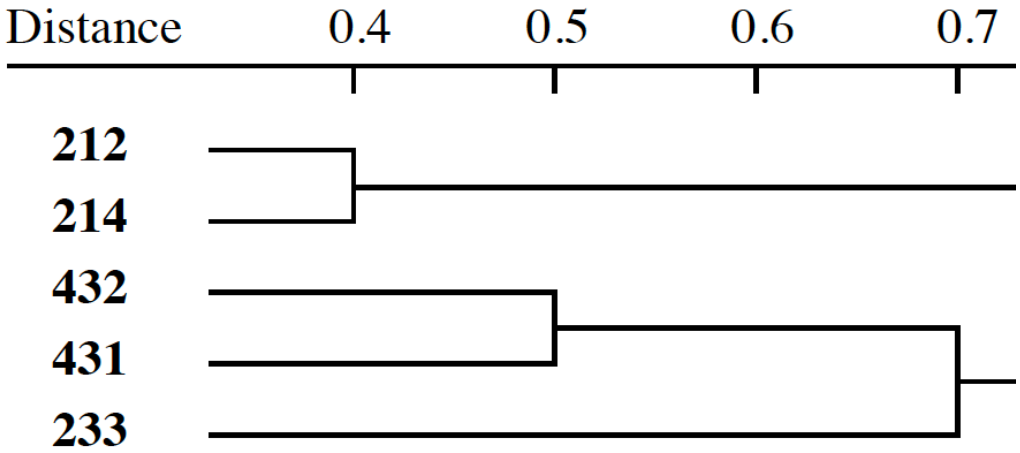


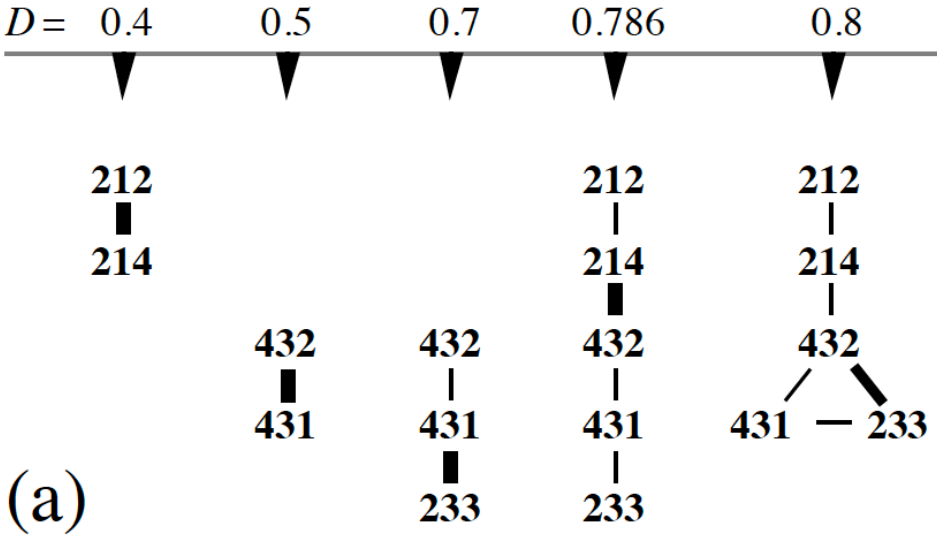
(a)



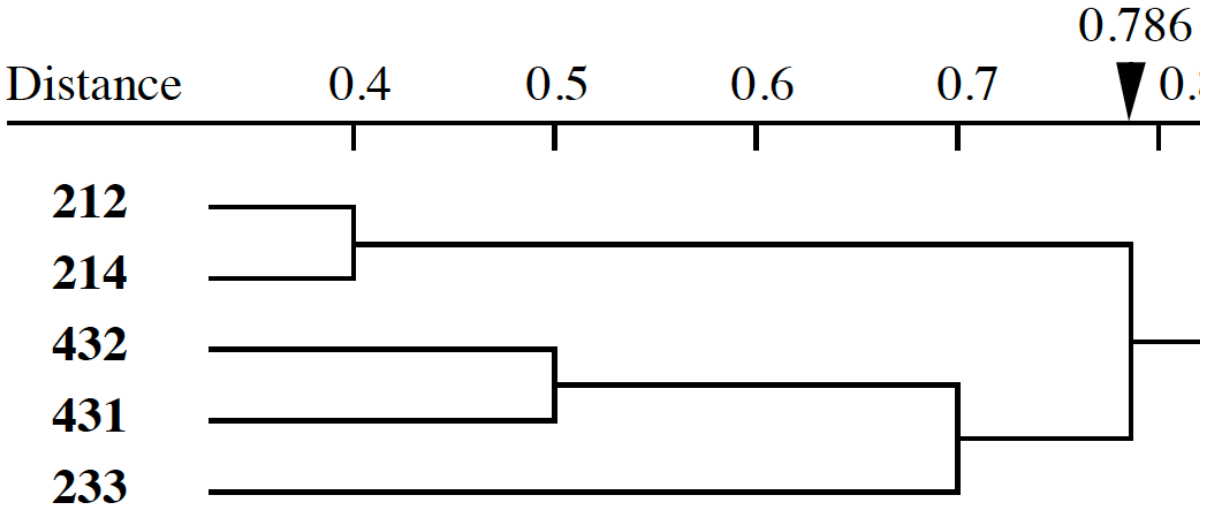

D = 0.4
0.5
0.7
212
214
432
431
432
431
233

(a)

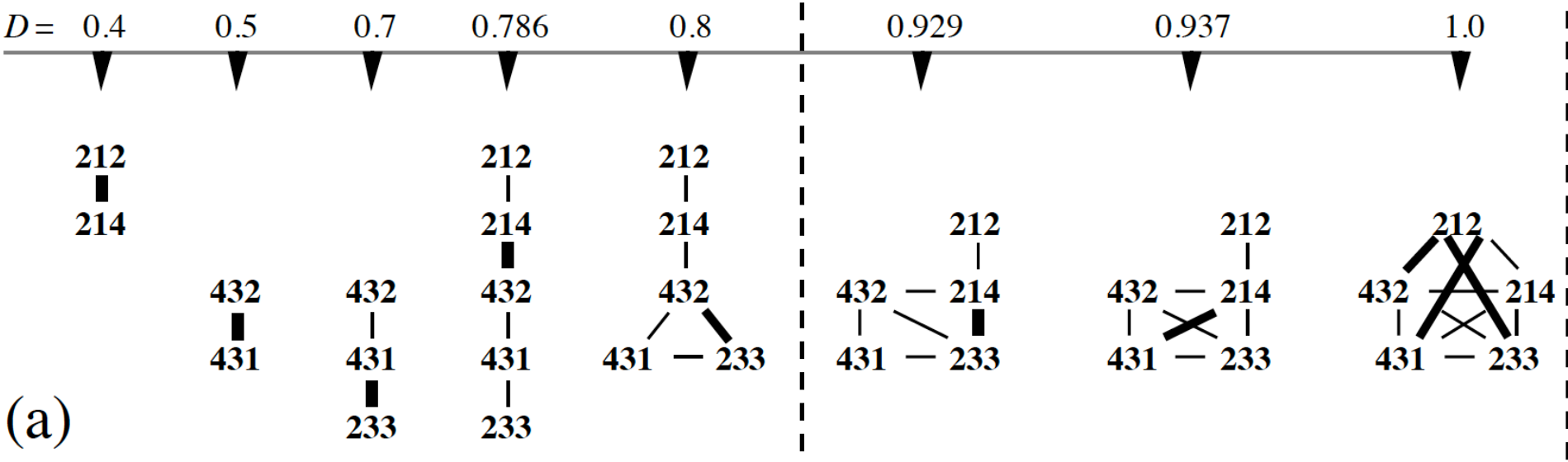

Distance
0.4
0.5
0.6
0.7
212
214
432
431
233

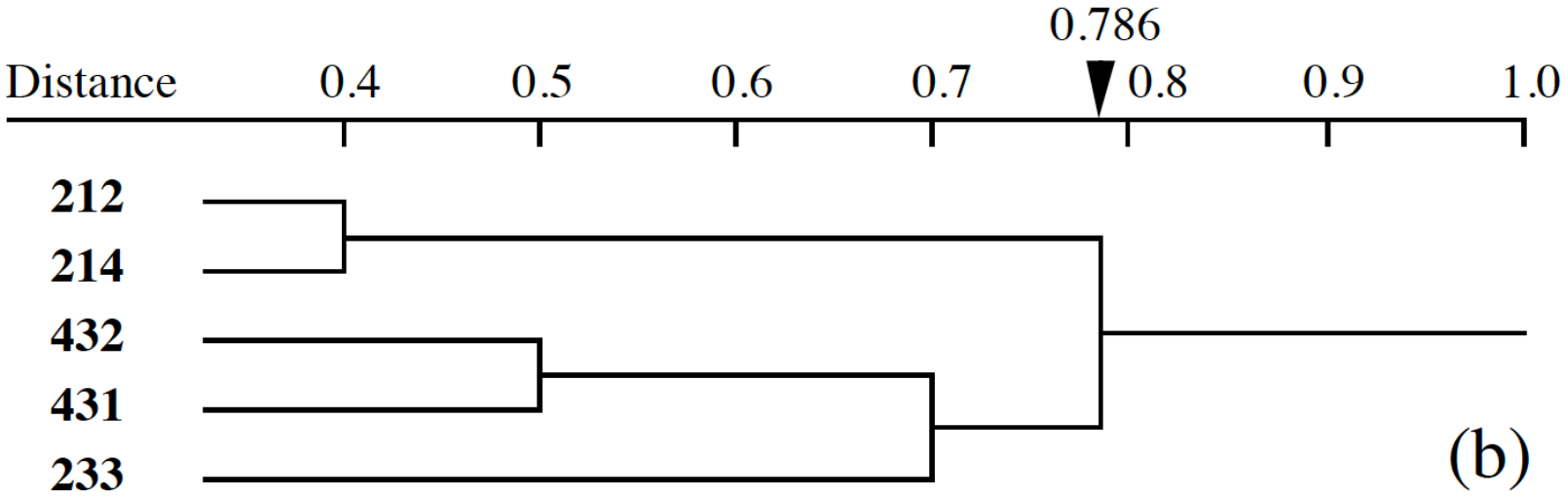

D = 0.4 0.5 0.7 0.786 0.8
212
214
212
214
212
214
432
432
431
432
431
233
432
431
431
233
233

(a)


0.786
Distance 0.4 0.5 0.6 0.7
212
214
432
431
233

Opcional

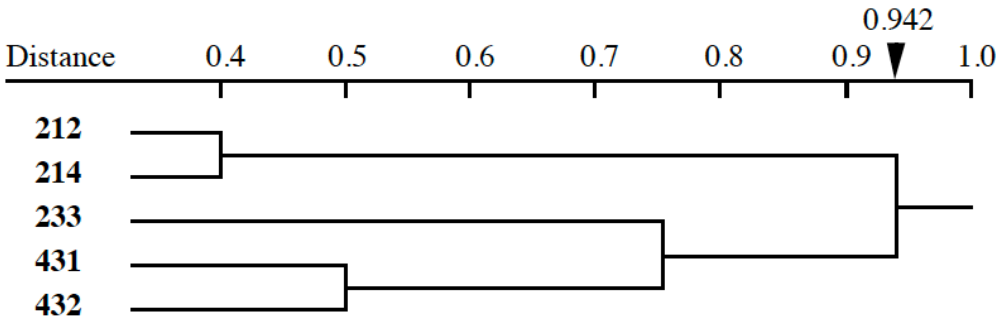

D = 0.4
0.5
0.7
0.786
0.8
0.929
0.937
1.0
212
214
432
431
233
(a)


Distance
0.4
0.5
0.6
0.7
0.786
0.8
0.9
1.0
212
214
432
431
233
(b)

UPGMA

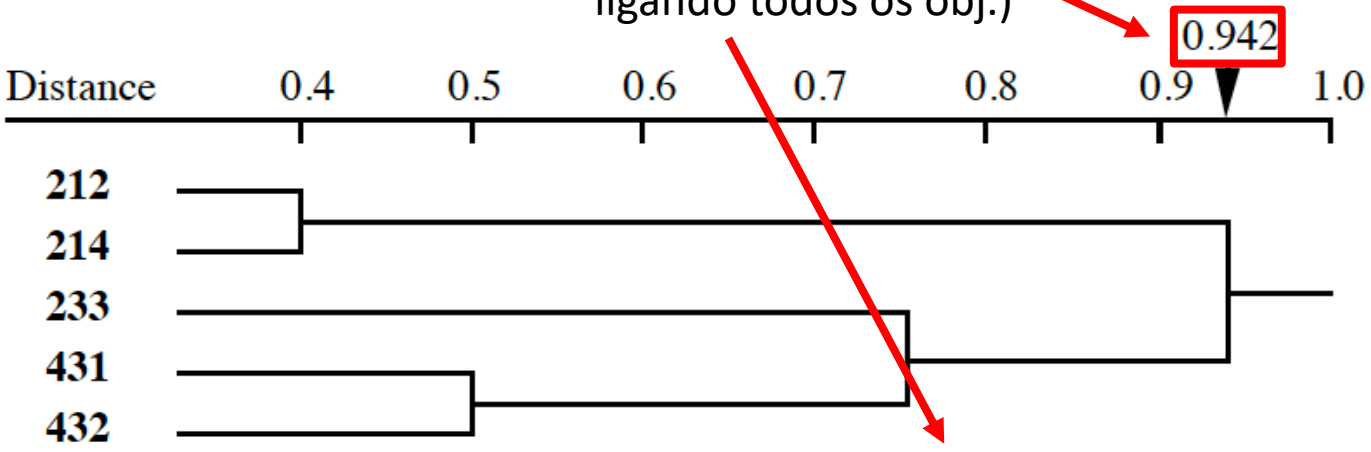
Table 8.3 Unweighted arithmetic average clustering (UPGMA) of the pond data. At each step, the lowest distance value is identified (italicized boldface value) and the two corresponding objects or groups are fused by averaging their distances as described in the text (boxes).

Objects	212	214	233	431	432	
212	—					**Step 1**
214	***0.400***	—				
233	1.000	0.929	—			
431	1.000	0.937	0.700	—		
432	1.000	0.786	0.800	0.500	—	
212-214		—				**Step 2**
233		0.9645	—			
431		0.9685	0.700	—		
432		0.8930	0.800	***0.500***	—	
212-214		—				**Step 3**
233		0.9645	—			
431-432		0.93075	***0.750***	—		
212-214		—				**Step 4**
233-431-432		***0.942***	—			

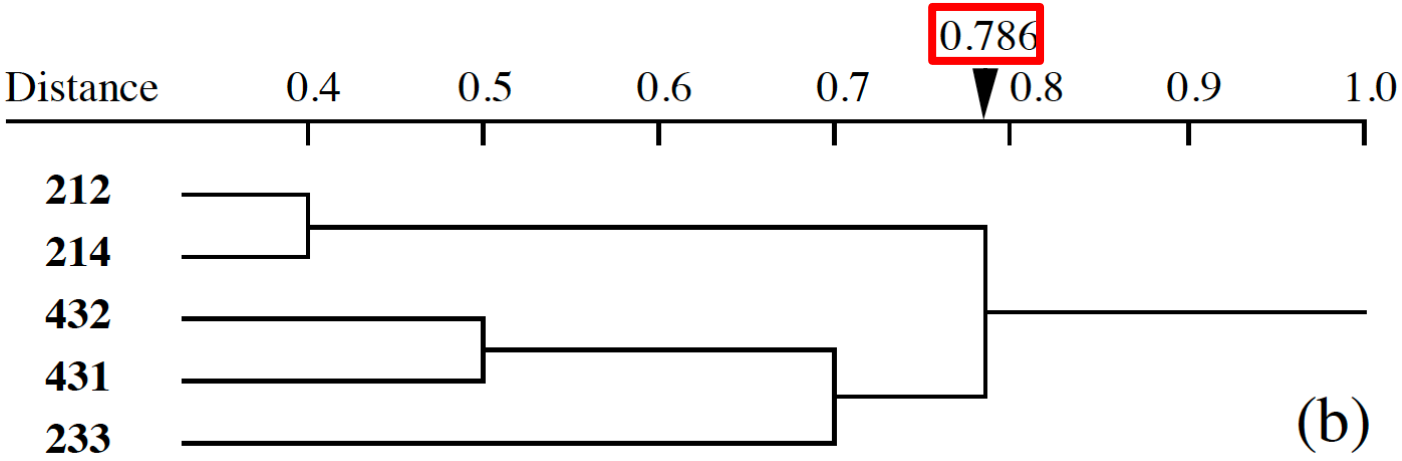


Ponto de parada do agrupamento (distância máxima ligando todos os obj.)

UPGMA



Single linkage



Variações da UPGMA

Table 4.1 The four methods of average clustering. The names in quotes are the corresponding arguments of function **hclust()**

	Arithmetic average	Centroid clustering
Equal weights	Unweighted pair-group method using arithmetic averages (UPGMA) "`average`"	Unweighted pair-group method using centroids (UPGMC) "`centroid`"
Unequal weights	Weighted pair-group method using arithmetic averages (WPGMA) "`mcquitty`"	Weighted pair-group method using centroids (WPGMC) "`median`"

Método de Ward

Minimiza o quadrado dos resíduos (ANOVA)

Table 8.7 Ward's minimum variance clustering of the pond data. Step 1 of the table contains *squared distances* computed as D^2 from the distance values in the upper panels of Tables 8.3 to 8.6. At each step, the lowest squared distance is identified (italicized boldface value) and the two corresponding objects or groups are fused using eq. 8.10.

Objects	212	214	233	431	432
212	—				**Step 1**
214	***0.16000***	—			
233	1.00000	0.86304	—		
431	1.00000	0.87797	0.49000	—	
432	1.00000	0.61780	0.64000	0.25000	—
212-214		—			**Step 2**
233		1.18869	—		
431		1.19865	0.49000	—	
432		1.02520	0.64000	***0.25000***	—
212-214		—			**Step 3**
233		1.18869	—		
431-432		1.54288	***0.67000***	—	
212-214		—			**Step 4**
233-431-432		***1.6795***	—		

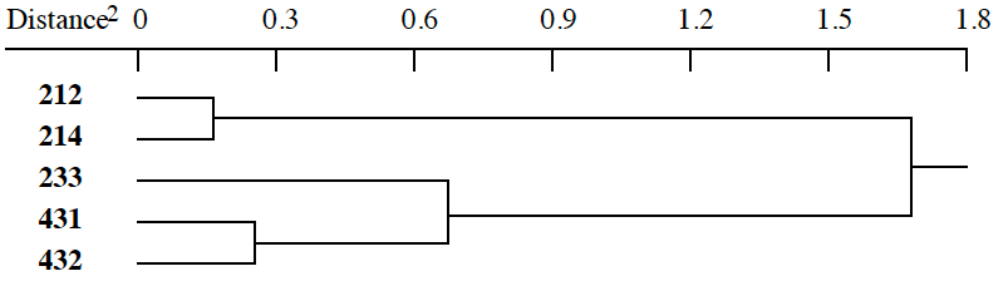


Table 8.8 Clustering steps in Ward's minimum variance clustering for the pond data. The objects are renamed 1 to 5 for shortness. K is the number of clusters, represented by underscored groups of objects. The total sum of squares (SS_{Total}) of the 5 objects is 1.37976 (eq. 8.6). SS_{Within} is also computed using eq. 8.6; $SS_{Among} = SS_{Total} - SS_{Within}$. Between clustering levels, $\Delta E^2_{\mathbf{hi}}$ is computed using eq. 8.8 or by difference between the successive values of SS_{Within} or SS_{Among}. D^2_{min} was computed in Table 8.7; D_{min} is the square root of D^2_{min}.

K	Objects	SS_{Within}	SS_{Among}	$\Delta E^2_{\mathbf{hi}}$	D^2_{min}	D_{min}
5	1 2 3 4 5	0	1.37976			
				0.08000	0.16000	0.40000
4	1 2 3 4 5	0.08000	1.29976			
				0.12500	0.25000	0.50000
3	1 2 3 4 5	0.20500	1.17476			
				0.33500	0.67000	0.81854
2	1 2 3 4 5	0.54000	0.83976			
				0.83976	1.67952	1.29596
1	1 2 3 4 5	1.37976	0			

Existe uma maneira de combinar vários algoritmos?

- *Flexible clustering*

4 parêmtros que

permitem alterar

o grau de agrupamento

- Implementado em

  cluster::agnes

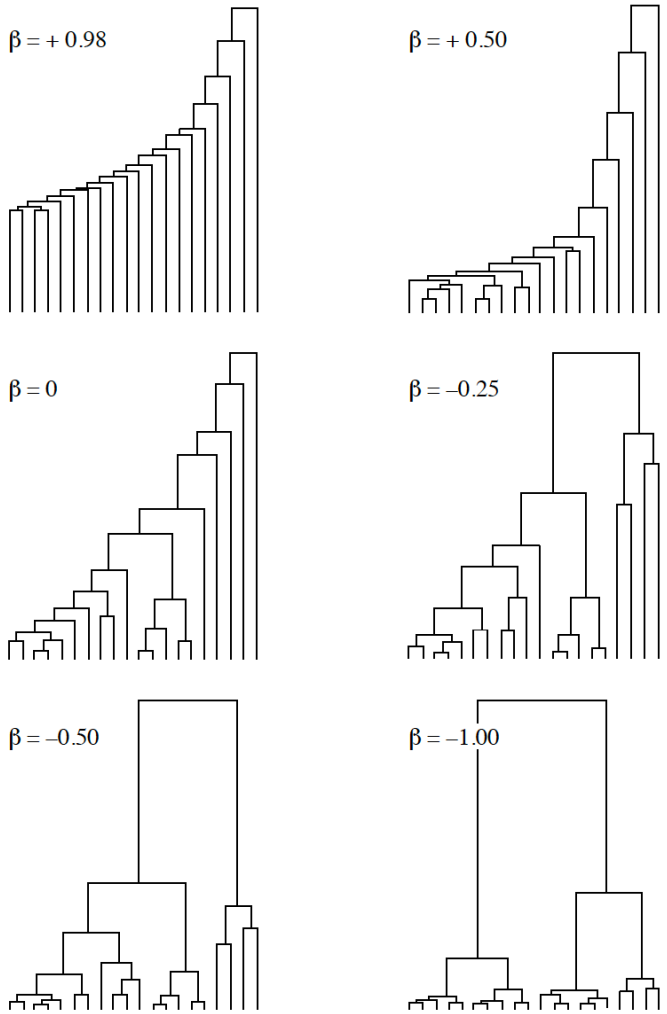


Table 8.9 Values of parameters α_h, α_i, β, and γ in Lance and Williams' general model for combinatorial agglomerative clustering. Modified from Sneath & Sokal (1973) and Jain & Dubes (1988).

Clustering method	α_h	α_i	β	γ	Effect on space A
Single linkage	1/2	1/2	0	–1/2	Contracting*
Complete linkage	1/2	1/2	0	1/2	Dilating*
UPGMA	$\frac{n_h}{n_h + n_i}$	$\frac{n_i}{n_h + n_i}$	0	0	Conserving*
WPGMA	1/2	1/2	0	0	Conserving
UPGMC	$\frac{n_h}{n_h + n_i}$	$\frac{n_i}{n_h + n_i}$	$\frac{-n_h n_i}{(n_h + n_i)^2}$	0	Conserving
WPGMC	1/2	1/2	–1/4	0	Conserving
Ward's	$\frac{n_h + n_g}{n_h + n_i + n_g}$	$\frac{n_i + n_g}{n_h + n_i + n_g}$	$\frac{-n_g}{n_h + n_i + n_g}$	0	Conserving
Flexible	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$-1 \le \beta < 1$	0	Contracting if $\beta \approx 1$ Conserving if $\beta \approx -.25$ Dilating if $\beta \approx -1$

* Terms used by Sneath & Sokal (1973).

L&L 2012

Como determinar grupos?

- Escolher nível de corte
  - Valor de distância que estabelece grupos
- Vários critérios (50% similaridade, inspeção visual etc.
- Reamostragem por Bootstrap multiescala
  - Calcula um valor de *P* não-enviesado para cada agrupamento
  - Pacote pvclut


O QUÊ?
ME EXPLICA ISSO, PORQUE EU NÃO ENTENDI NADA...

The Annals of Statistics
2004, Vol. 32, No. 6, 2616–2641
DOI 10.1214/009053604000000823

APPROXIMATELY UNBIASED TESTS OF REGIONS USING MULTISTEP-MULTISCALE BOOTSTRAP RESAMPLING[1]

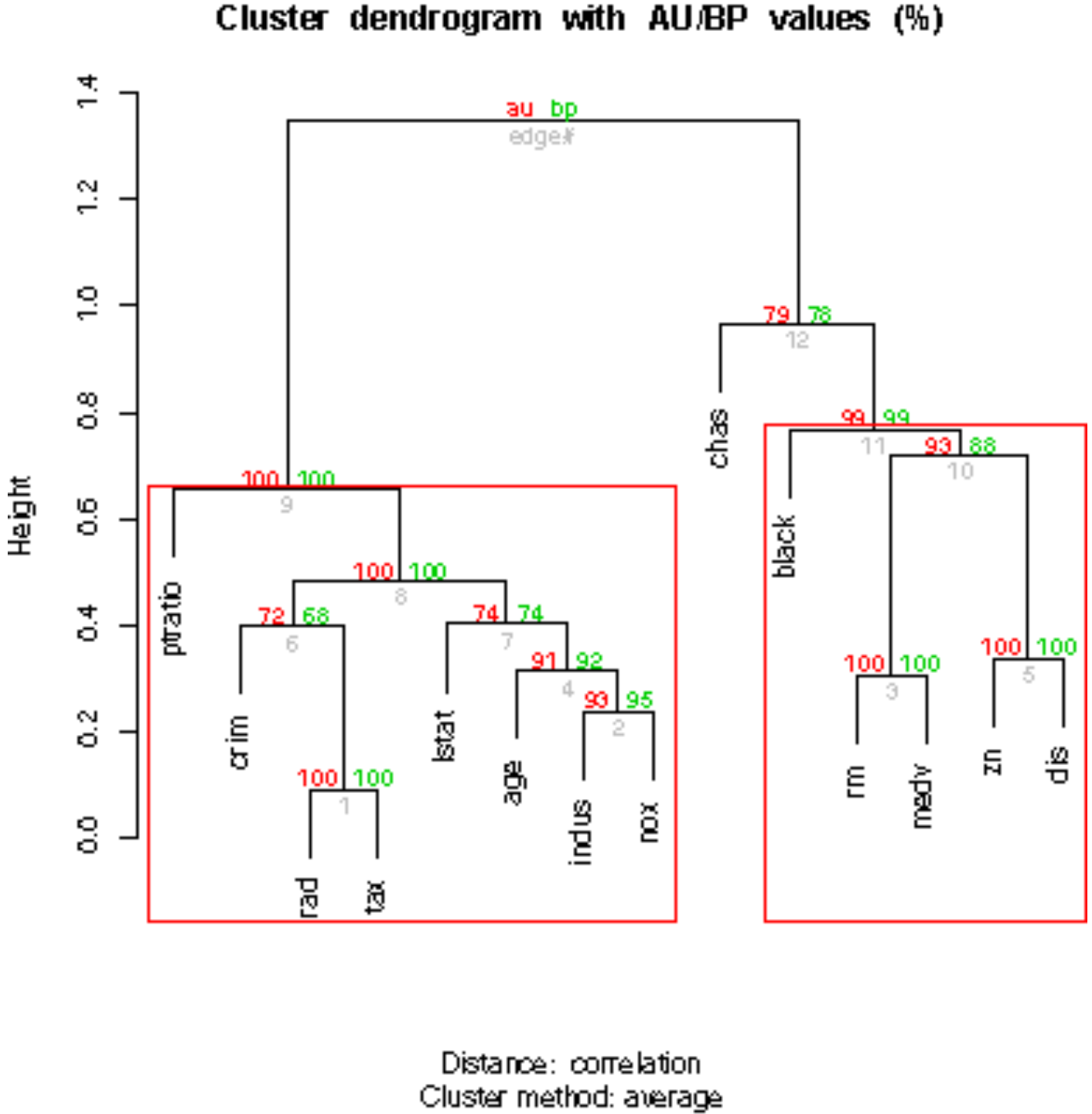
BY HIDETOSHI SHIMODAIRA

PVCLUST

- Reamostragem utilizando bootstrap em multiescala.
  - Utiliza 2 estimadores para calcular o valor de p: **AU** e **BP**
  - Verifica se os agrupamentos (clusters) se manteríam se o tamanho amostral fosse aumentado.
  - função pvrect circula grupos com **AU** maiores que 95%
  - AU values > 0.95 => for suporte para agrupamento

```
> result <- pvclust(lung, method.dist="cor", method.hclust="average", nboot=1000)
Bootstrap (r = 0.5)... Done.
Bootstrap (r = 0.6)... Done.
Bootstrap (r = 0.7)... Done.
Bootstrap (r = 0.8)... Done.
Bootstrap (r = 0.9)... Done.
Bootstrap (r = 1.0)... Done.
Bootstrap (r = 1.1)... Done.
Bootstrap (r = 1.2)... Done.
Bootstrap (r = 1.3)... Done.
Bootstrap (r = 1.4)... Done.
```

?pvclust


Cluster dendrogram with AU/BP values (%)
au bp
edge#
Height
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
100 100
9
100 100
8
72 68
6
74 74
7
91 92
4
99 95
2
100 100
1
79 78
12
99 99
11
93 88
10
100 100
3
100 100
5
ptratio
crim
rad
tax
lstat
age
indus
nox
chas
black
rm
medv
zn
dis
Distance: correlation
Cluster method: average

PVCLUST

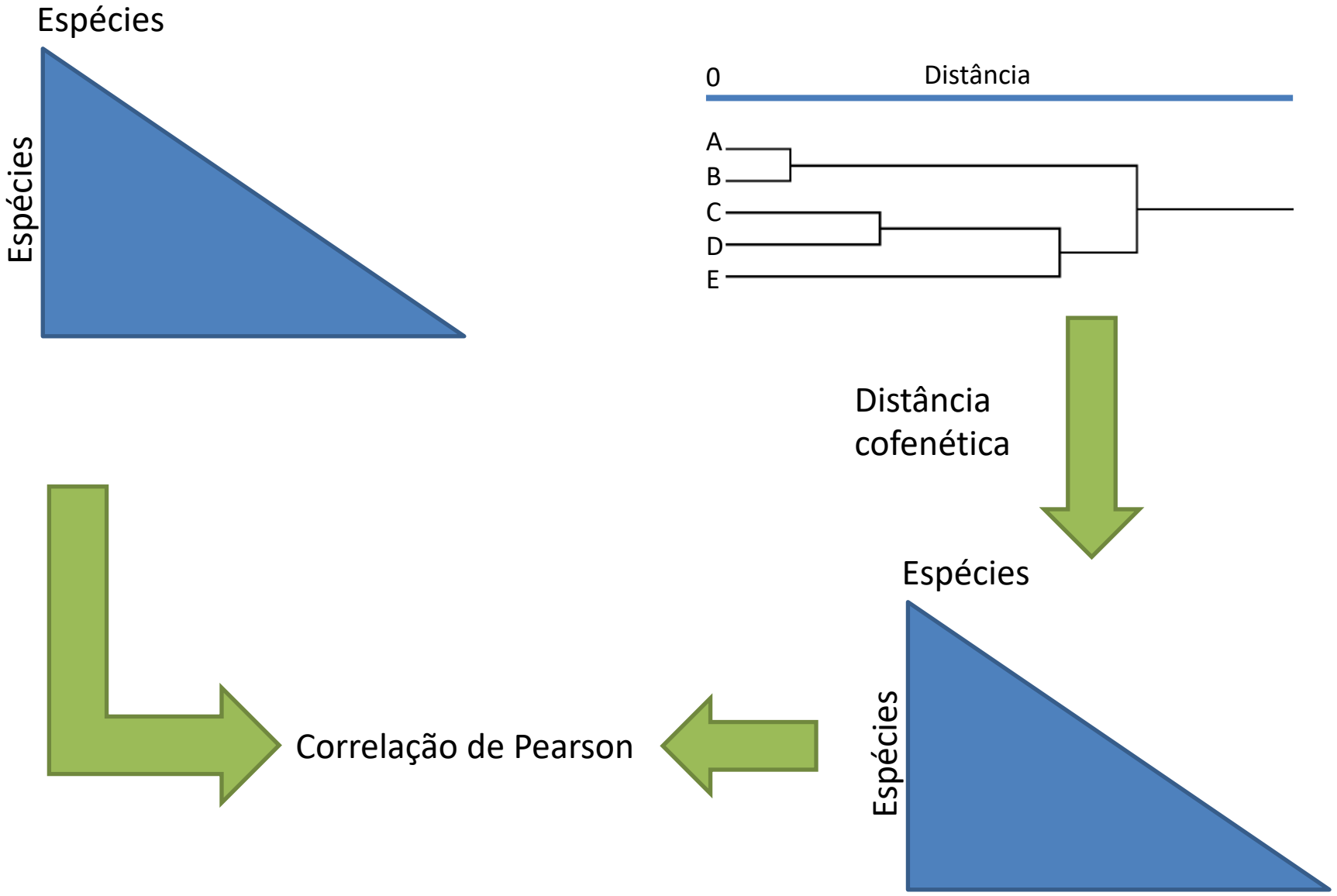
- Desvantagem: só aceita alguns coeficientes de distância (tem de usar uma função que permite usar coeficientes do vegan::vegdist)

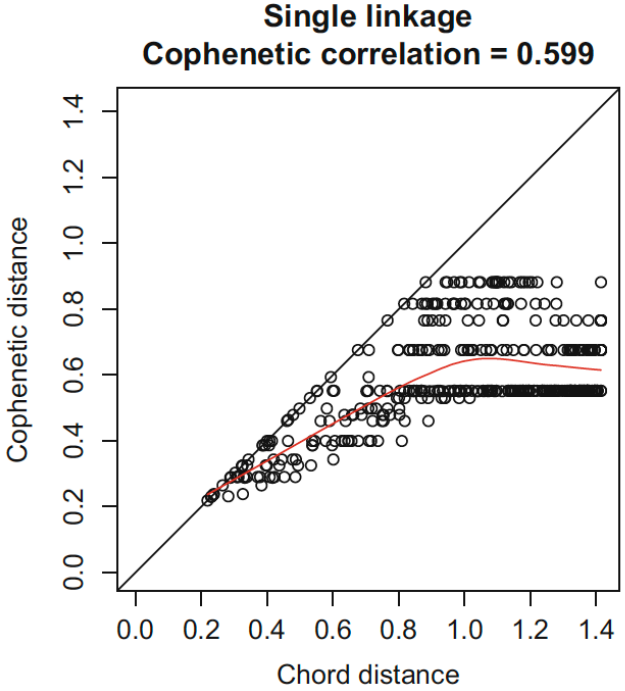
Fusion level values

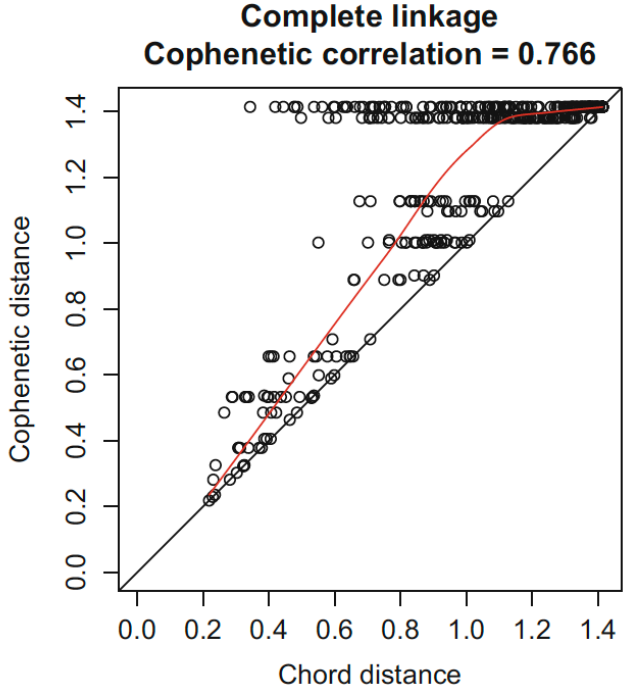
- Outra alternativa para encontrar o melhor nível de corte de dissimilaridade
- Ver item 4.7.3.1 de Borcard et al. (2018)

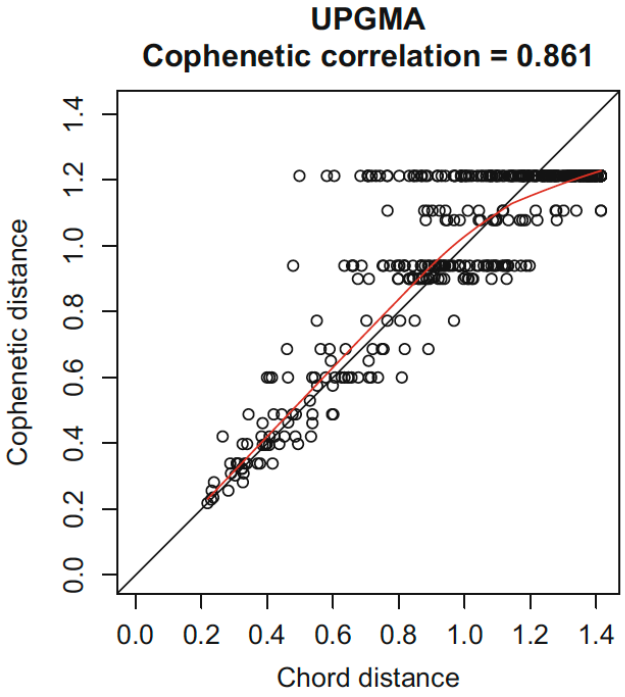
Como avaliar a representatividade do dendrograma?

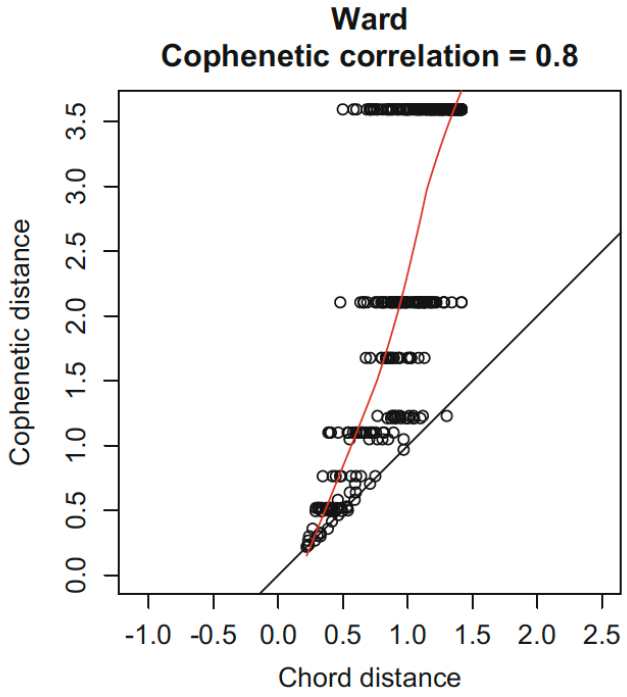
- Coeficiente de Correlação Cofenética
  - Correlação de Pearson entre a matriz de similaridade e a matriz cofenética (distância entre objetos dada pelo dendrograma)
  - Escolha métodos que tenham a maior correlação possível
  - Mostra que o dendrograma não distorceu muito as distâncias originais
  - Regra do 0,8


Espécies
Espécies
0
Distância
A
B
C
D
E
Distância
cofenética
Espécies
Espécies
Correlação de Pearson









Borcard et al. 2018

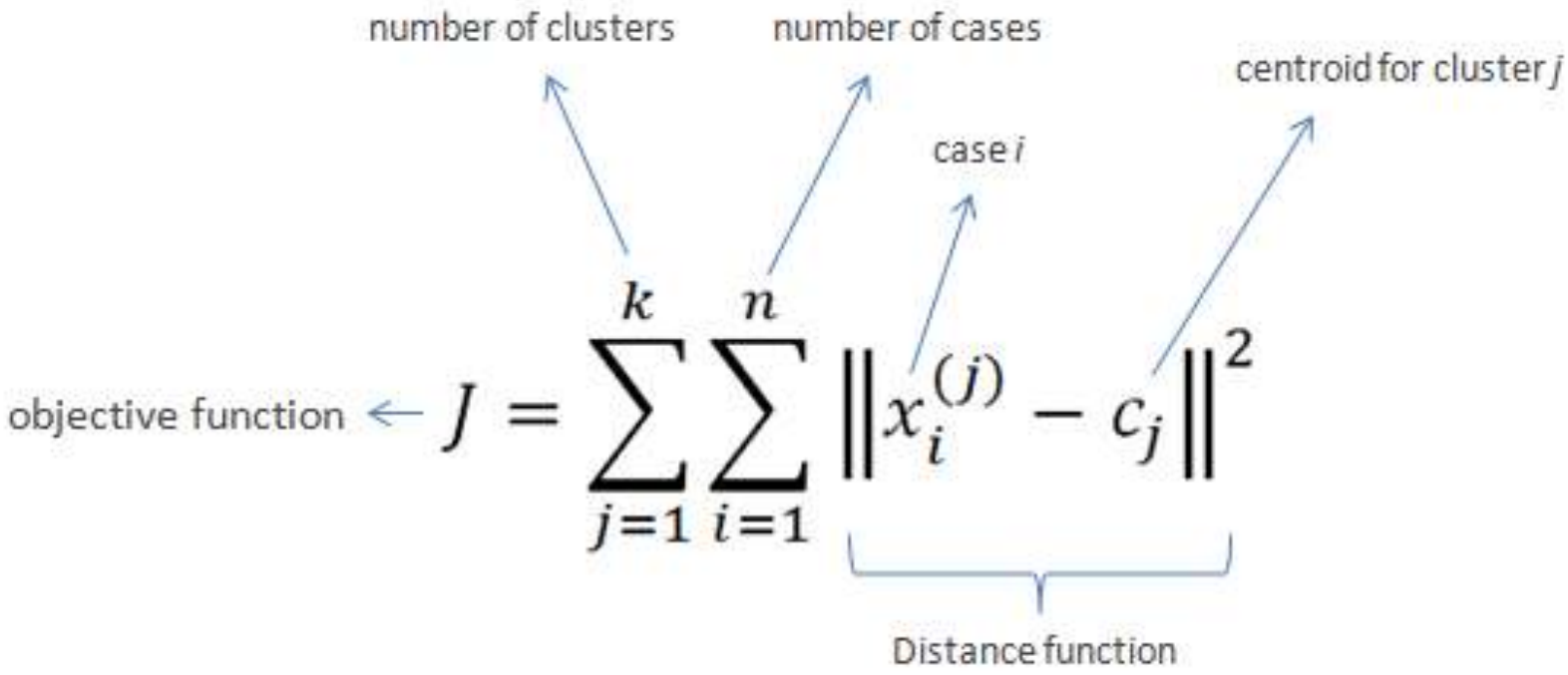
K-means

K-means

- Técnica de agrupamento não-hierárquica
- Útil para encontrar grupos que não são subgrupos de outros
- Agrupa objetos ao redor de centroides (objeto representante do grupo)
- Número de grupos (k) é pré-definido pelo usuário

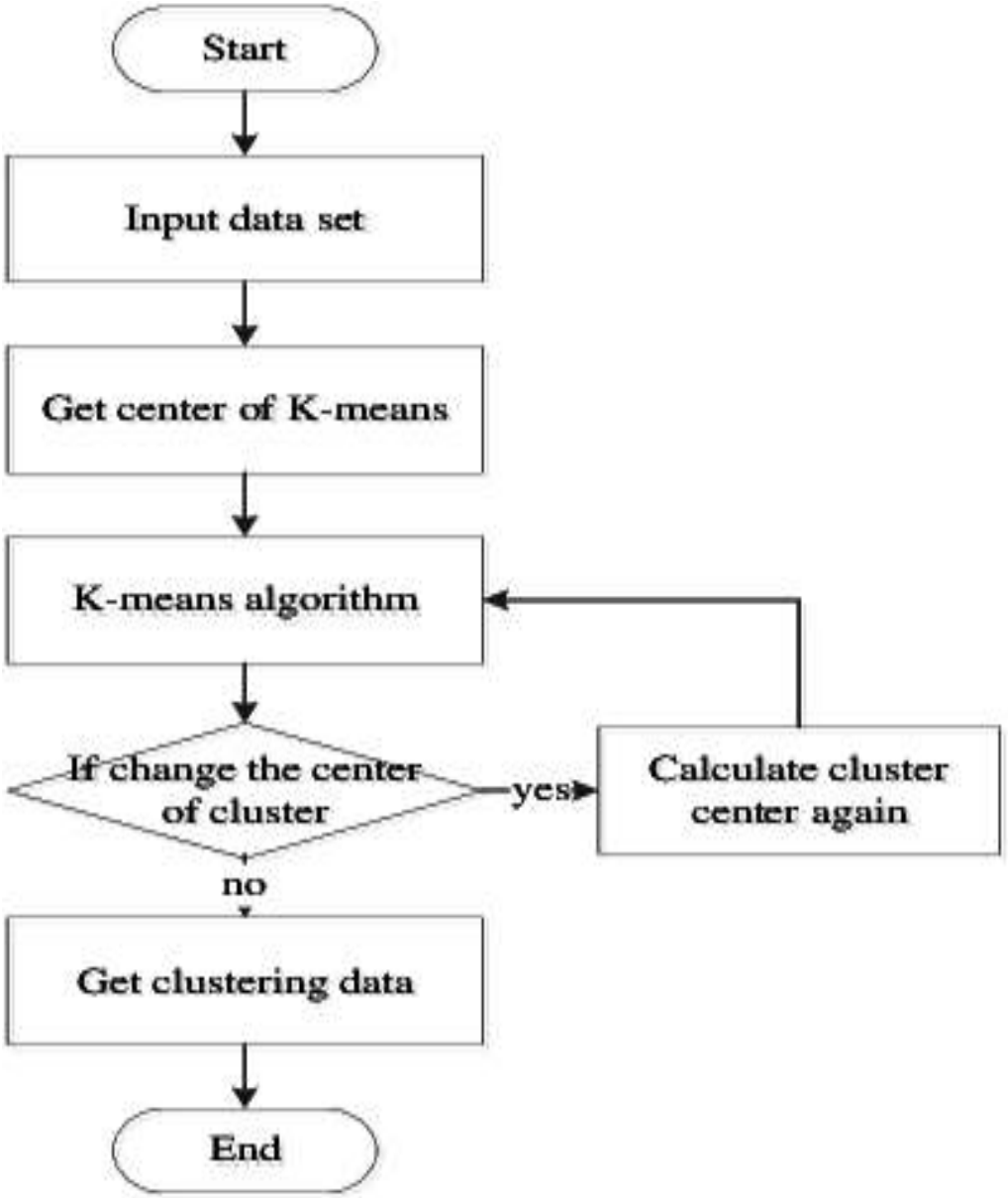
Como funciona o algoritmo?

- Grupos podem ser calculados a partir da matriz bruta ou de distância
  - TESS é a mesma usando D^2 ou dados brutos
  - Se o conjunto de dados é muito grande, é melhor calcular direto na matriz bruta
    - Usa distância Euclidiana nesse caso
- Minimiza os Soma de quadrados total dos resíduos (TESS) da mesma maneira que método de Ward



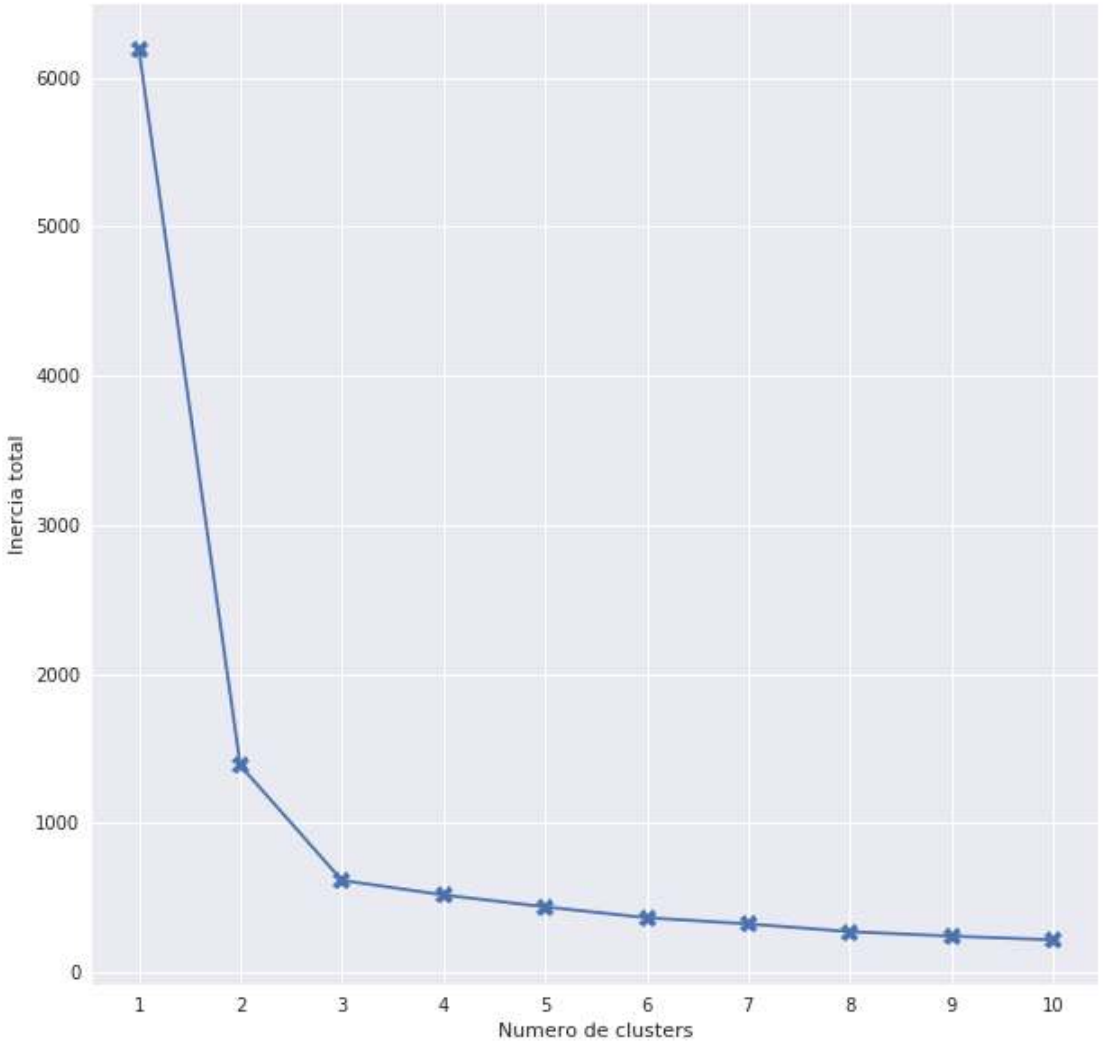
https://www.saedsayad.com/clustering_kmeans.htm

Como funciona o algoritmo?



Qiu-yu et al. 2015

Como determinar o número ideal de grupos? Scree plot



http://minerandodados.com.br/index.php/2017/12/12/entenda-o-algoritmo-k-means/

Análise de espécie indicadora (IndVal)

Ecological Monographs, 67(3), 1997, pp. 345–366

SPECIES ASSEMBLAGES AND INDICATOR SPECIES: THE NEED FOR A FLEXIBLE ASYMMETRICAL APPROACH

MARC DUFRÊNE[1] AND PIERRE LEGENDRE[2]

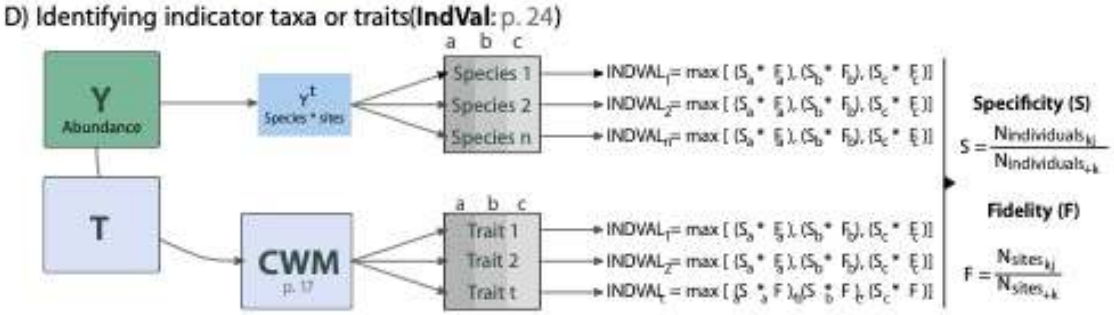
Veja também De Cáceres & Legendre 2009, De Cáceres et al. 2012

IndVal

- Técnica que permite encontrar espécies indicadoras de hábitats
  - Melhor que o TWINSPAN
- Útil no contexto de monitoramento ambiental para encontrar espécies substitutas (*surrogates*)
- Essas *surrogate species* seriam então priorizadas para amostragem, já que seriam indicadoras de uma dada condição ambiental

p. 397-403 L&L

IndVal

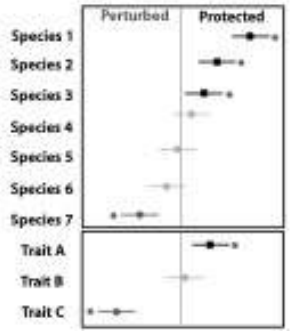


i) Ethnobiology

What are the preferred plants (and traits) used by local people under climate change?

ii) Ecology

What are the species (and traits) affected by habitat loss?



Gonçalves-Souza et al. 2019

IndVal

Specificity

$$A_{kj} = Nindividuals_{kj} / Nindividuals_{+k}$$

Fidelity

$$B_{kj} = Nsites_{kj} / Nsites_{k+}$$

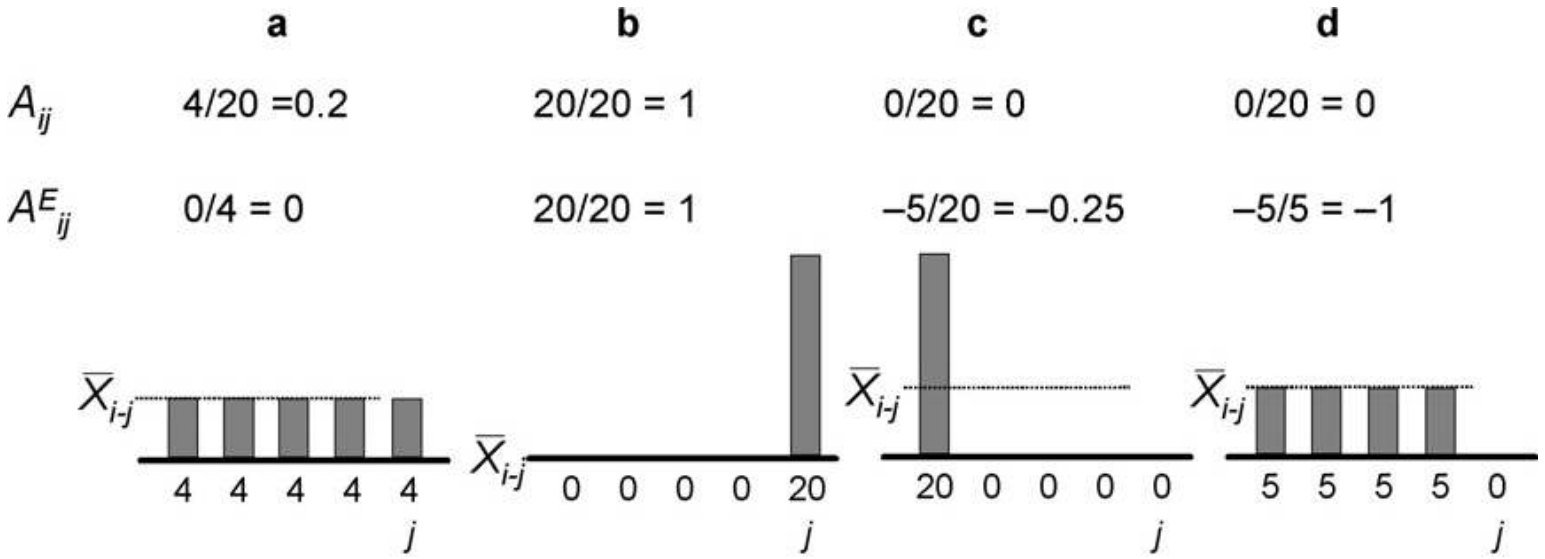
$$INDVAL_{kj} = A_{kj}\, B_{kj}$$

Indicator value

$$INDVAL_{j} = \max\left[INDVAL_{kj}\right]$$

IndVal

- Conceitos
  - **Especificidade** => todas as unidades amostrais do grupo
  - **Fidelidade** => só em 1 grupo



Podani & Csányib 2010

Table 8.10 Numerical example: abundance of three species at 25 sites divided into 5 groups. Modified from Dufrêne & Legendre (1997). Top panel: data. Bottom panel: calculation of the specificity (A_{kj}), fidelity (B_{kj}) and $INDVAL_{kj}$ index for each species (j) in each group of sites (k). The maximum value of $INDVAL_{kj}$ for each species is in bold.

Groups	Group 1					Group 2					Group 3					Group 4					Group 5				
Sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Species 1	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	3	3	3	3	3	2	2	2	2	2
Species 2	8	8	8	8	8	4	4	4	4	4	6	6	6	6	6	4	4	2	0	0	0	0	0	0	0
Species 3	18	18	18	18	18	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	Group 1	Group 2	Group 3	Group 4	Group 5
Species 1					
A_{k1}	4/20 = 0.20	5/20 = 0.25	6/20 = 0.30	3/20 = 0.15	2/20 = 0.10
B_{k1}	5/5 = 1	5/5 = 1	5/5 = 1	5/5 = 1	5/5 = 1
$INDVAL_{k1}$	0.20	0.25	**0.30**	0.15	0.10
Species 2					
A_{k2}	8/20 = 0.40	4/20 = 0.20	6/20 = 0.30	2/20 = 0.10	0/20 = 0.00
B_{k2}	5/5 = 1	5/5 = 1	5/5 = 1	3/5 = 0.6	0/5 = 0
$INDVAL_{k2}$	**0.40**	0.20	0.30	0.06	0.00
Species 3					
A_{k3}	18/20 = 0.90	2/20 = 0.10	0/20 = 0.00	0/20 = 0.00	0/20 = 0.00
B_{k3}	5/5 = 1	5/5 = 1	0/5 = 0	0/5 = 0	0/5 = 0
$INDVAL_{k3}$	**0.90**	0.10	0.00	0.00	0.00

Table 5

Indicator value (IndVal) for different Ephemeroptera species in anthropized and preserved streams. The Monte Carlo test was performed using 9.999 runs

	Anthropized	Preserved	*P*
Americabaetis alphus	0.462	0.438	0.868
Aturbina georgei	0.299	0.552	0.162
Baetodes prosculus	0.000	0.800	0.001
Camelobaetidius juparana	0.000	0.267	0.110
Paracloeodes eurybranchus	0.400	0.187	0.277
Rivudiva trichobasis	0.401	0.0188	0.045
Tupiara ibirapitanga	0.000	0.267	0.100
Caenis cuniana	0.027	0.213	0.262
Campylocia anceps	0.023	0.043	1.000
Tricorythodes hiemalis	0.230	0.236	0.994
Tricorythopsis gibbus	0.101	0.099	1.000
Farrodes carioca	0.003	0.445	0.008
Needhamella ehrhardti	0.281	0.010	0.082

Gonçalves-Souza et al. 2019

Method	Pros & cons	Use in ecology
Hierarchical agglomeration: linkage clustering	Pairwise relationships among the objects are known.	
Single linkage	Computation simple; contraction of space (chaining); combinatorial method.	Good complement to ordination.
Complete linkage (see also: species associations)	Dense nuclei of objects; space expansion; many objects cluster at high distance; arbitrary rules to resolve conflicts; combinatorial method.	To increase the contrast among clusters.
Intermediate linkage	Preservation of reference space A; non-combinatorial: not included in Lance & Williams' general model.	Preferable to the above two methods if only one clustering method is to be used.
Hierarchical agglomeration: average clustering	Preservation of reference space A; pairwise relationships between objects are lost; combinatorial method.	
Unweighted arithmetic average (UPGMA)	Fusion of clusters when the distance reaches the mean inter-cluster distance value.	For a collection of objects obtained by simple random or systematic sampling.
Weighted arithmetic average (WPGMA)	As UPGMA, with adjustment for group sizes.	Preferable to the previous method in all other sampling situations.
Unweighted centroid (UPGMC)	Fusion of clusters with closest centroids; may produce reversals.	For simple random or systematic samples of objects.
Weighted centroid (WPGMC)	As UPGMC, with adjustment for group sizes; may produce reversals.	Preferable to the previous method in all other sampling situations.
Ward's method	Minimizes the within-group sum of squares.	When looking for hyperspherical clusters in space A.
Hierarchical agglomeration: flexible clustering	Allows contraction, conservation, or dilation of space A; pairwise relationships between objects are	All combinatorial methods, including this one, are implemented using the simple
***K*-means partitioning**	Minimizes within-group sum of squares; different rules may	Produces a partition of the objects into K groups, K being determined
Indicator species		
TWINSPAN	Only for classifications of sites obtained by splitting CA axes; justification of some steps unclear.	Gives indicator values for the pseudospecies.
Indicator value index	For any hierarchical or non-hierarchical classification of sites; *IndVal* for a species is not affected by the other species in the study.	Gives indicator values for the species under study; the *IndVal* index is tested by permutation.

Minha cara depois de assistir a essa aula