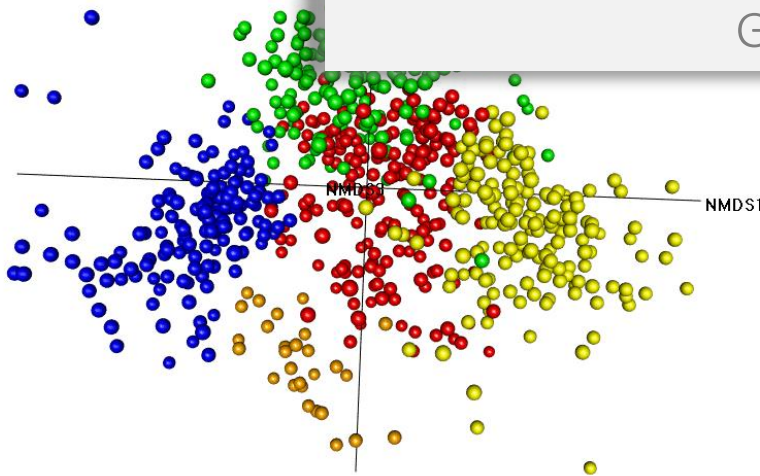


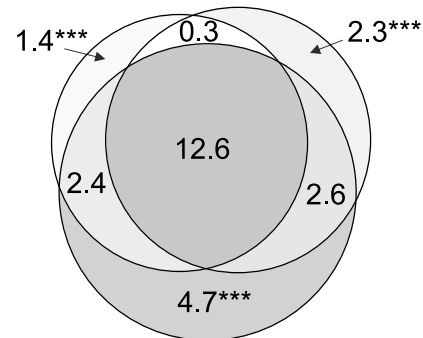
# Metody fyzické geografie 3: Biogeografie & ekologie

Jan Divíšek

Geografický ústav & Ústav botaniky a zoologie



climate: 16.6\*\*\* land-cover: 17.8\*\*\*



natural habitats: 22.3\*\*\*

I TÝ SI ZAPIŠ NOVÝ  
PŘEDMĚT Z 8055  
METODY FYZICKÉ  
GEOGRAFIE 3!

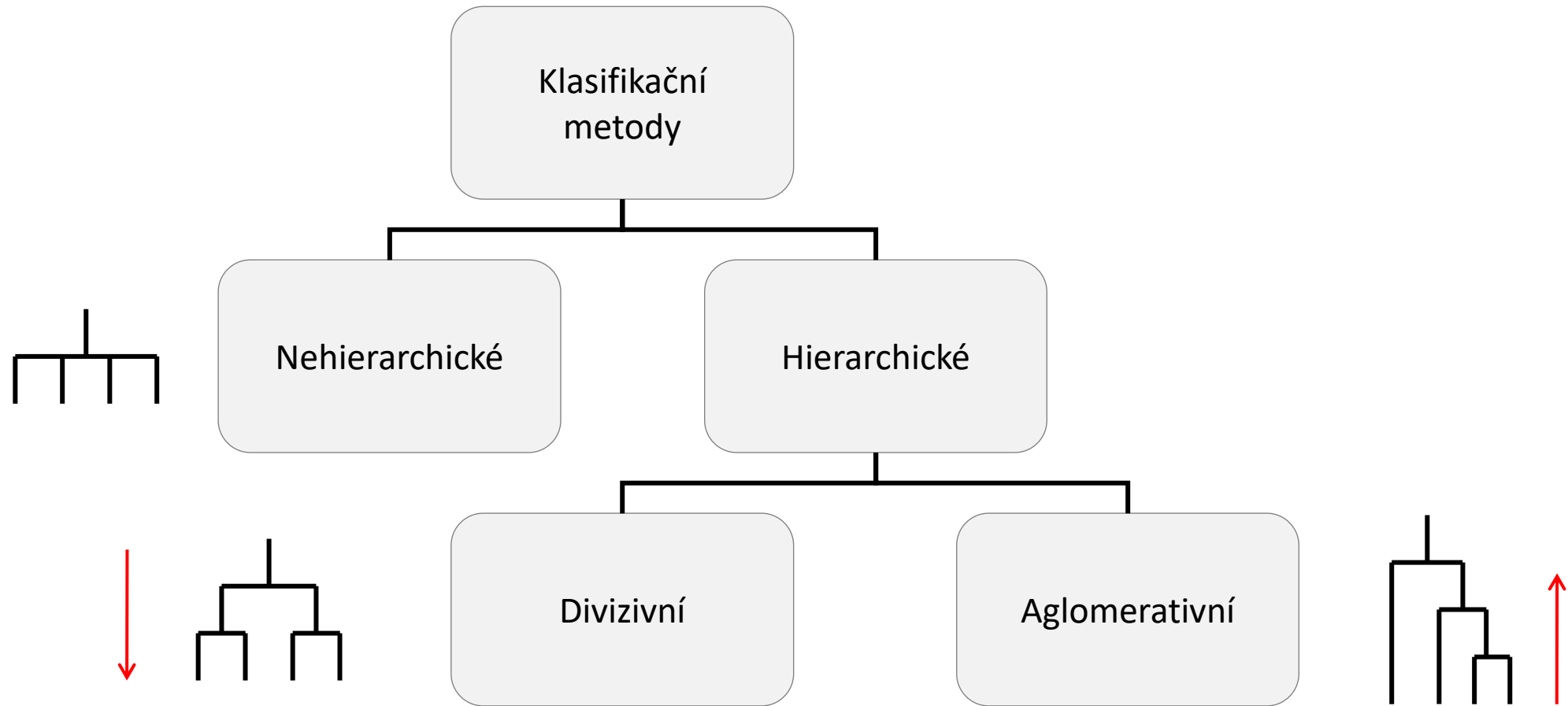


# Klasifikační metody

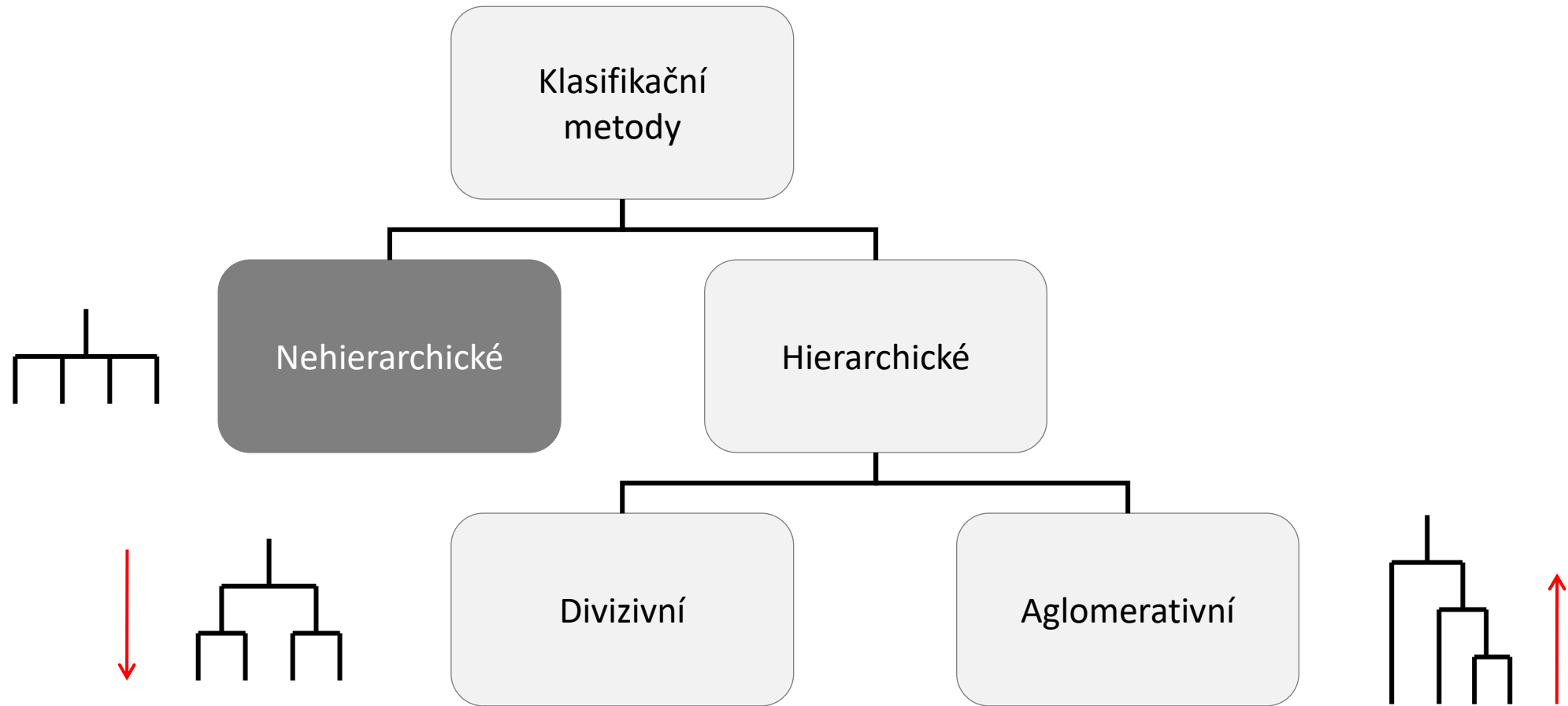
# Numerická klasifikace

- Cílem je seskupit podobné objekty (vzorky, druhy) do skupin, které jsou vnitřně homogenní, dobře popsatelné a zároveň dobře odlišitelné od ostatních skupin
- Nutná kvantifikace podobnosti vzorků × klasifikační metody pracují se vzdálenostmi (nepodobnostmi)
- Klasifikace v biogeografii a ekologii
  - pokud analyzují vzorky – daná skupina obsahuje vzorky s podobným druhovým složením (např. podobná stanoviště; výsledkem mohou být např. biogeografické regiony)
  - pokud analyzují druhy – daná skupina obsahuje druhy s podobnou distribucí a ekologickým chováním (výsledkem mohou být např. chorotypy)

# Základní typy klasifikačních metod



# Základní typy klasifikačních metod



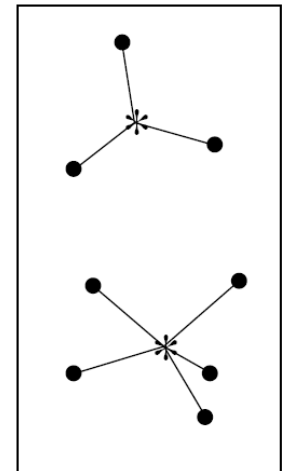
# Nehierarchická klasifikace

kmeans ()

## k-means clustering (shlukování metodou k-průměrů)

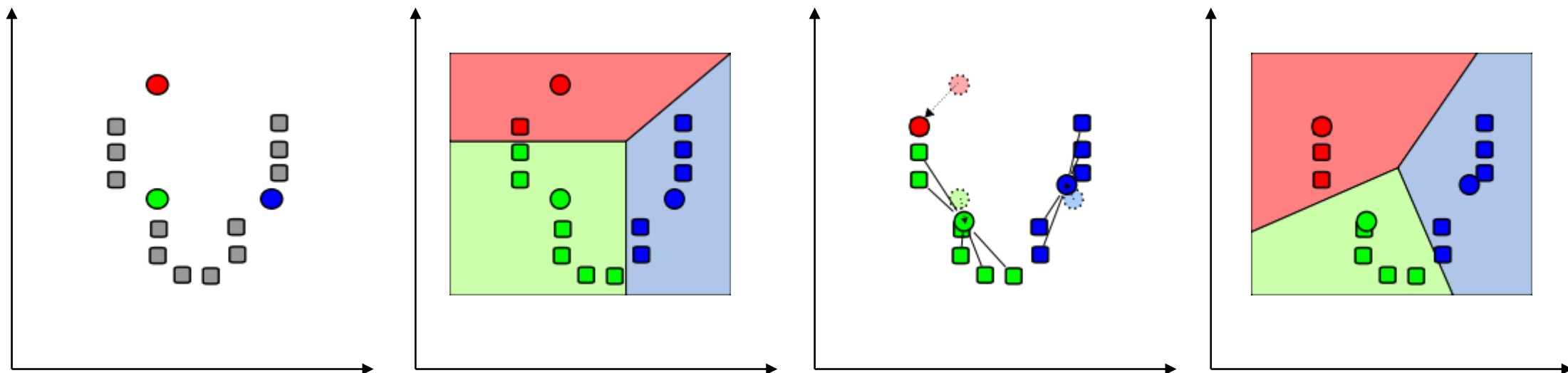
- Nehierarchická metoda – všechny shluky jsou si rovny
- Vstupem je tabulka dat, nikoliv matice vzdáleností
- Pracuje v euklidovském prostoru, tj. s euklidovskými vzdálenostmi
- Na začátku je nutné zvolit počet shluků
- Homogenní shluky jsou vytvářeny tak, aby sumy čtverců vzdáleností vzorků od centroidů shluků byly minimální
- Iterativní metoda, začne od náhodného přiřazení vzorků do shluků, postupně přehazuje vzorky mezi shluky a hledá optimální řešení
- Výsledek do určité míry záleží na počátečním rozmístění shluků do vzorků a je proto dobré proces mnohokrát zopakovat (najít stabilní řešení), protože metoda má tendenci nacházet lokální minima

$$ED = \sqrt{\sum (x_i - y_i)^2}$$



Legendre & Legendre (2012)

# Jak pracuje k-means



[https://en.wikipedia.org/wiki/K-means\\_clustering](https://en.wikipedia.org/wiki/K-means_clustering)



# Aplikace k-means v biogeografii a ekologii

- Aplikují k-means na druhová data
  - Klasifikují vzorky s podobným druhovým složením
  - Druhová data by měla být předem **transformovaná Hellingerovou transformací** → k-means bude počítat s Hellingerovými vzdálenostmi
- Aplikují k-means na environmentální data
  - Vytvářím environmentální stratifikaci
  - Environmentální data by měla být předem **standardizována na stejnou škálu jednotek** (průměr = 0, směrodatná odchylka = 1, nebo mezi 0 a 1)



# Nehierarchická klasifikace

## **k-means clustering**

- R, STATISTICA, S.A.M., ArcGIS (Grouping Analysis)

## **Další nehierarchická klasifikace**

- PAM (Partitioning Around Medoids)

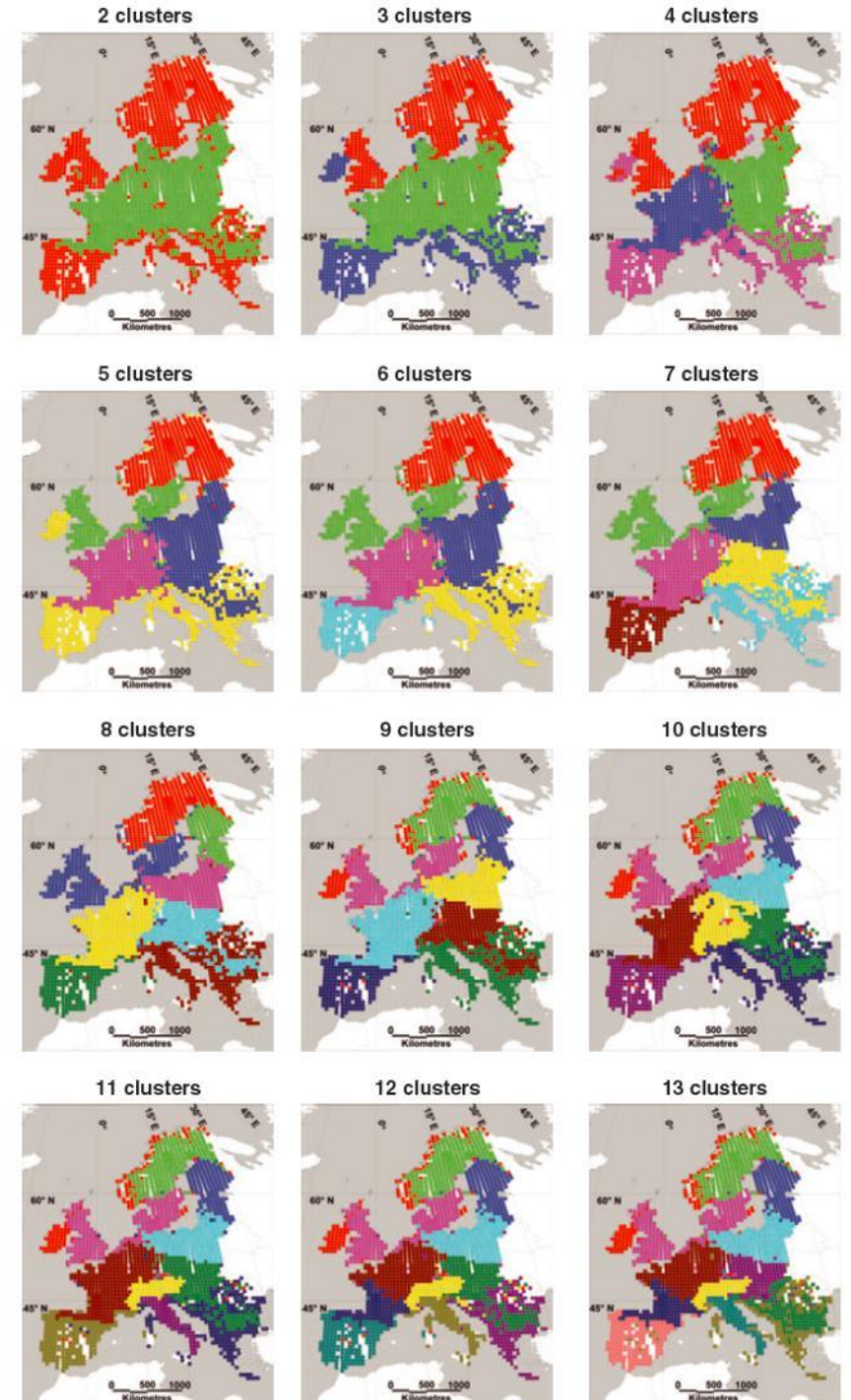
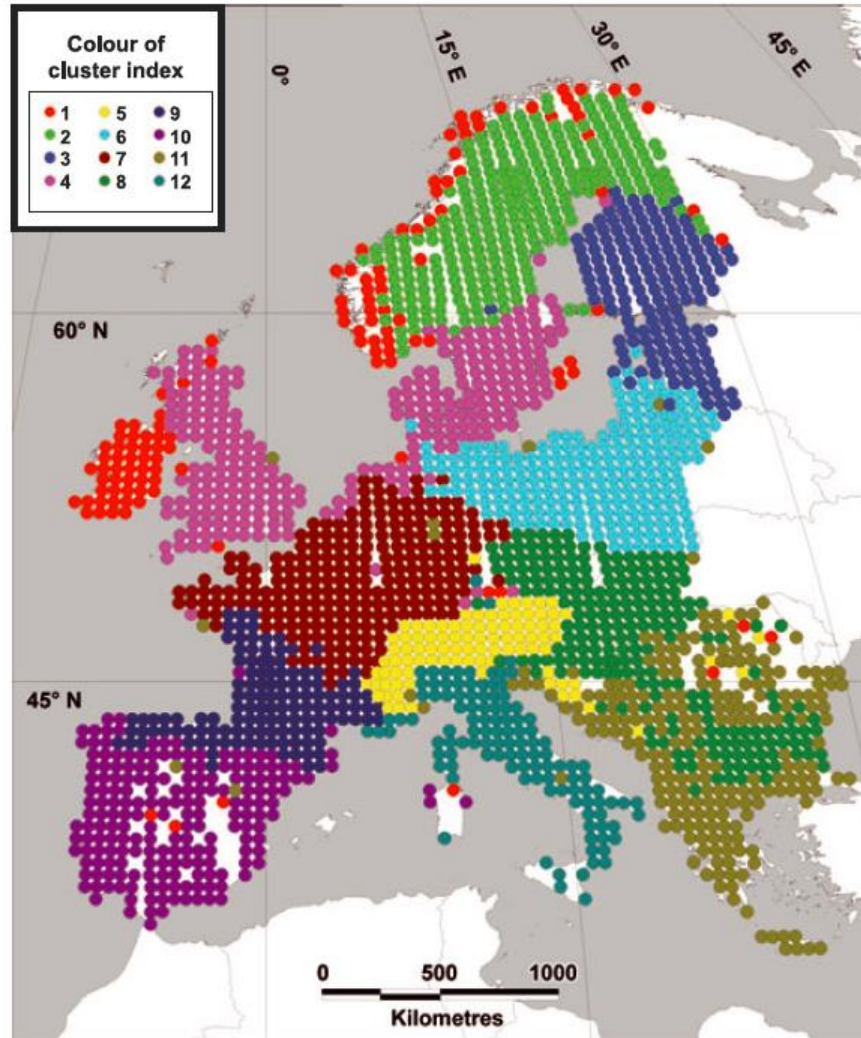
```
pam {cluster}
```

ORIGINAL  
ARTICLE

# Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters

H. Heikinheimo<sup>1\*</sup>, M. Fortelius<sup>2</sup>, J. Eronen<sup>2</sup> and H. Mannila<sup>1,3</sup>

**Figure 3** The *k*-means clustering of the mammal data cells in 12 clusters with the 'all species' set. The clustering is **the best out of 100 clustering runs in terms of squared error**. The cells are projected on to the map with the Mollweide (equal-area) NAD27 projection.



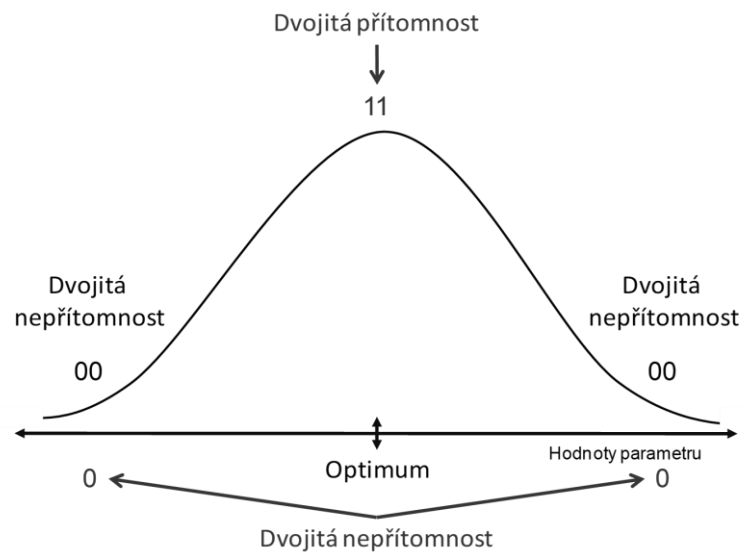
## CORRESPONDENCE

Accurate delineation of biogeographical regions depends on the use of an appropriate distance measure



SARA A. GAGNÉ\* AND RAPHAËL PROULX

### Double-zero problem

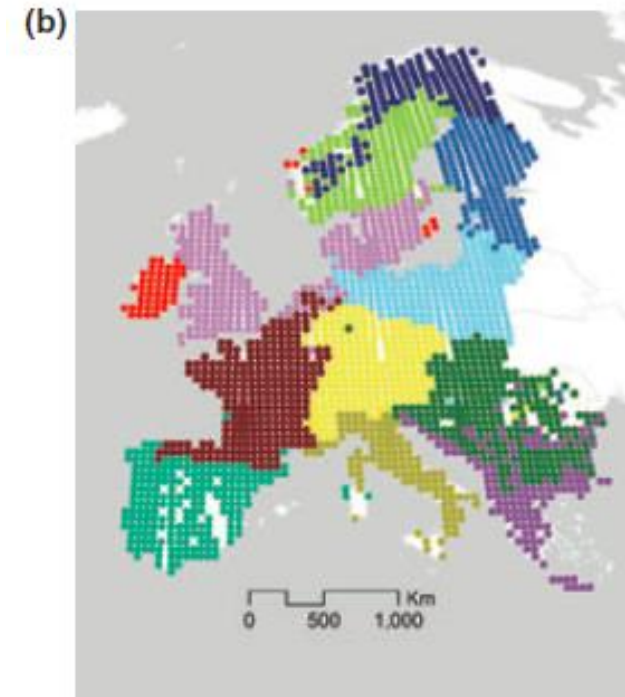
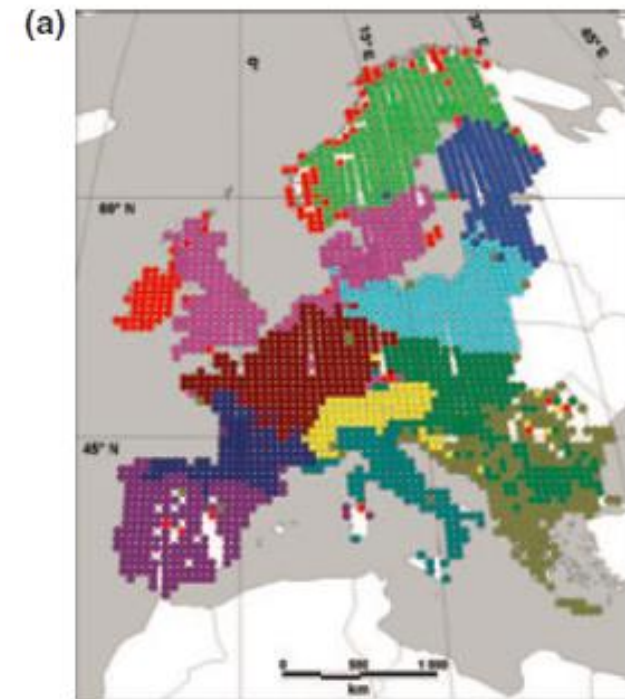


### Species-abundance paradox

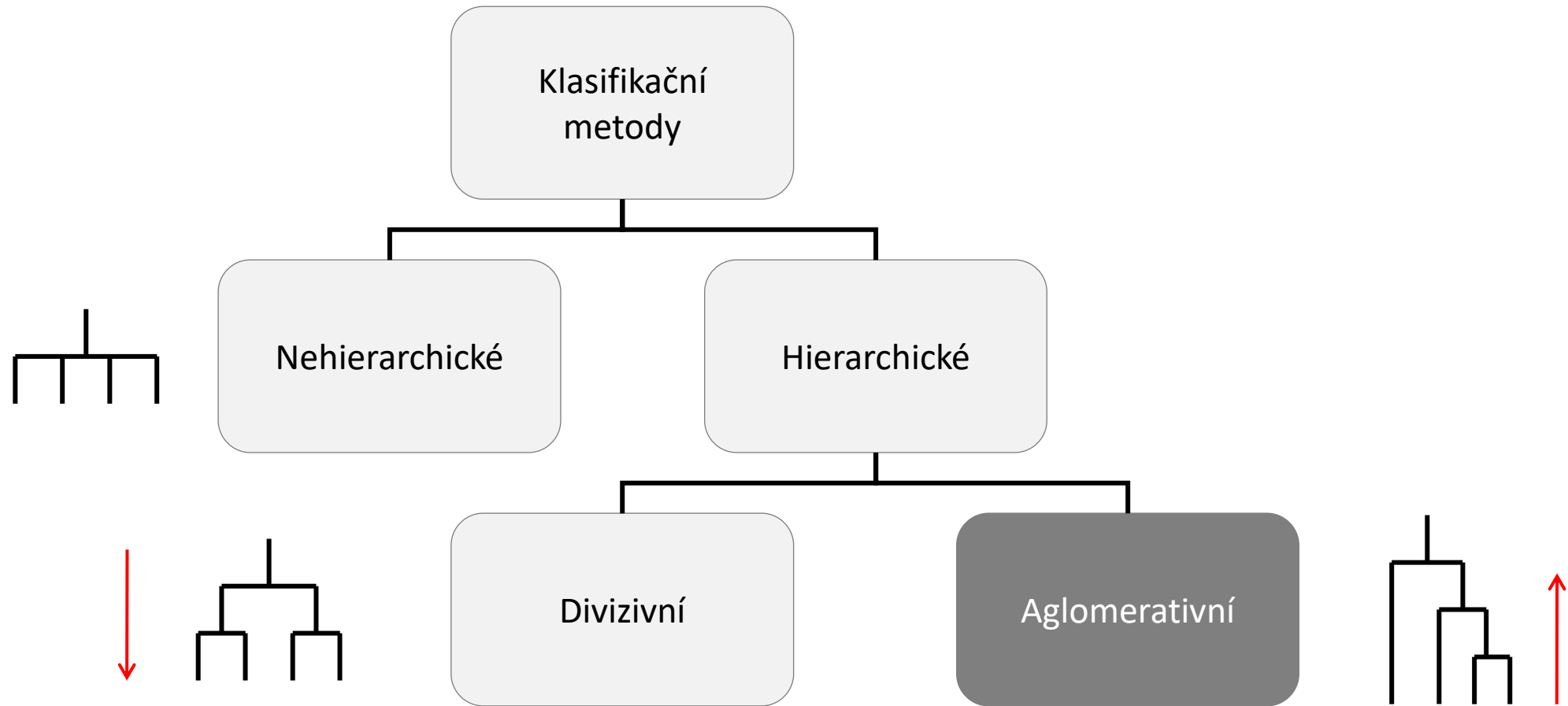
**Table 1** A hypothetical example to illustrate the species-abundance paradox (see text): occurrence (1 = presence, 0 = absence) of four species ( $j, k, l, m$ ) recorded at three sites ( $x_1, x_2, x_3$ ).

Sites	Species			
	$j$	$k$	$l$	$m$
$x_1$	0	0	1	0
$x_2$	1	0	1	0
$x_3$	1	1	1	1

$D = 1$   
 $D = 1.4$



# Základní typy klasifikačních metod

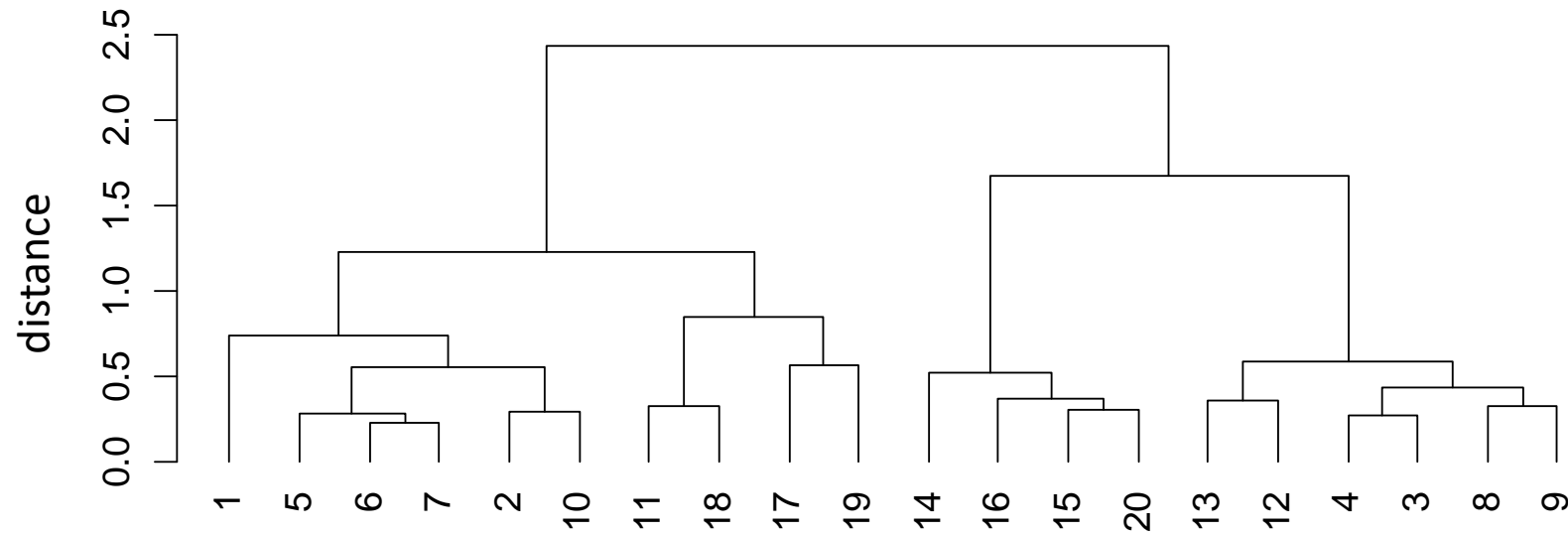


# Hierarchická aglomerativní shluková analýza

- Shluky jsou hierarchicky uspořádány (výsledkem je klasifikační strom = **dendrogram**)
- Shluky jsou tvořeny „odspodu“, tzn. postupným shlukováním jednotlivých vzorků do větších skupin
- Obecný postup shlukové analýzy:
  1. **Volba vhodného indexu vzdálenosti** (ekologické nepodobnosti)
  2. Výpočet matice vzdáleností
  3. **Volba shlukovacího algoritmu** (jakým způsobem budou vzorky shlukovány)
  4. Aplikace shlukovacího algoritmu (*clustering algorithm*) na matici vzdáleností
  5. **Volba výsledného počtu shluků**, který budu interpretovat

# Dendrogram

- záleží na tom, které vzorky jsou spojeny na které úrovni
- nezáleží na tom, který vzorek (skupina) je vpravo a který vlevo



# Nejpoužívanější algoritmy

- Single linkage
- Complete linkage
- Flexible clustering
- Average linkage (UPGMA, WPGMA, UPGMC, WPGMC)
- Wardova metoda

# Single & complete linkage

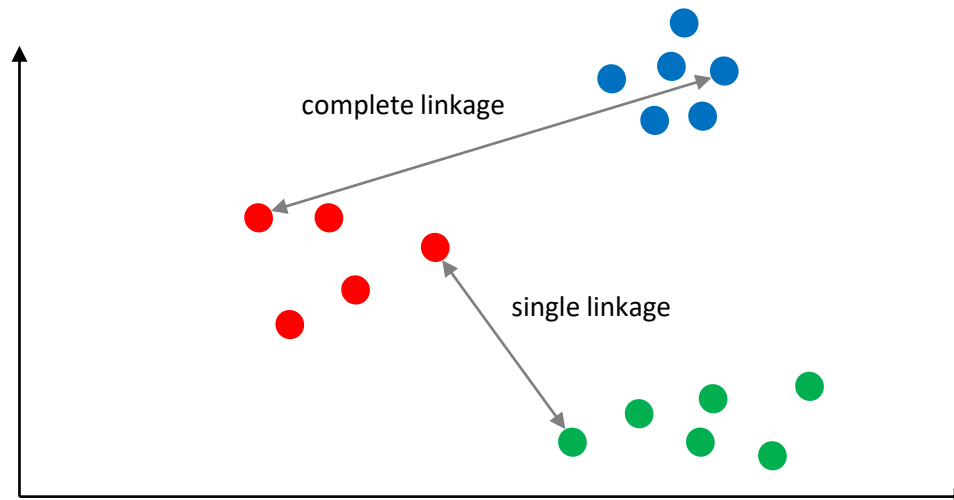
`hclust()`

## Single linkage (metoda jednospojná)

- Uvažuje **nejpodobnější** vzorky → vzorek se připojí ke shluku, ve kterém ne nejpodobnější snímek

## Complete linkage (metoda všespojná)

- Uvažuje **nejméně podobné** vzorky → vzorek se připojí ke shluku, ve kterém je nejpodobnější





# Jak funguje Single linkage?

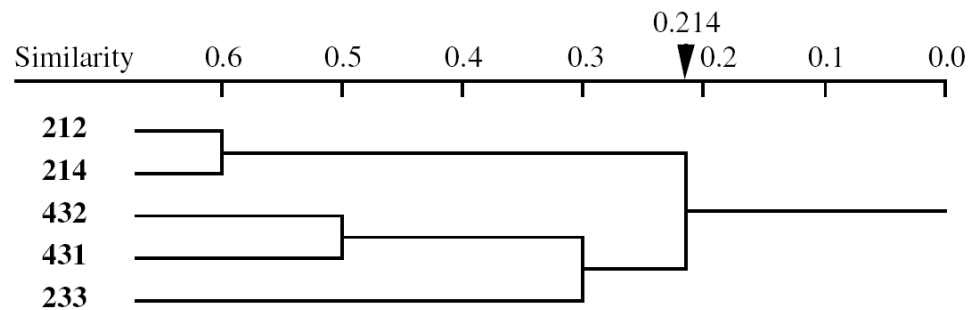
Matice podobností

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.600	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—

Páry vzorků seřazené podle podobnosti

$S_{20}$	Pairs formed
0.600	212-214
0.500	431-432
0.300	233-431
0.214	214-432
0.200	233-432
0.071	214-233
0.063	214-431
0.000	212-233
0.000	212-431
0.000	212-432

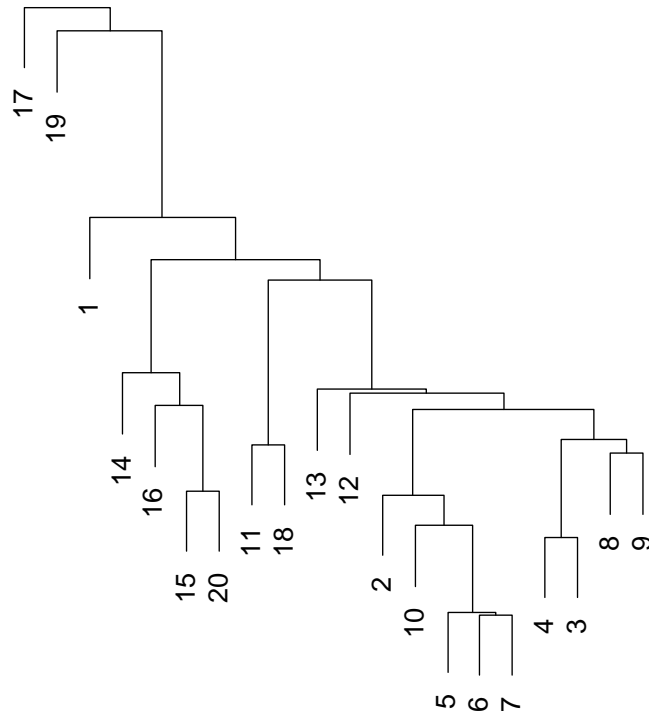
Výsledný dendrogram



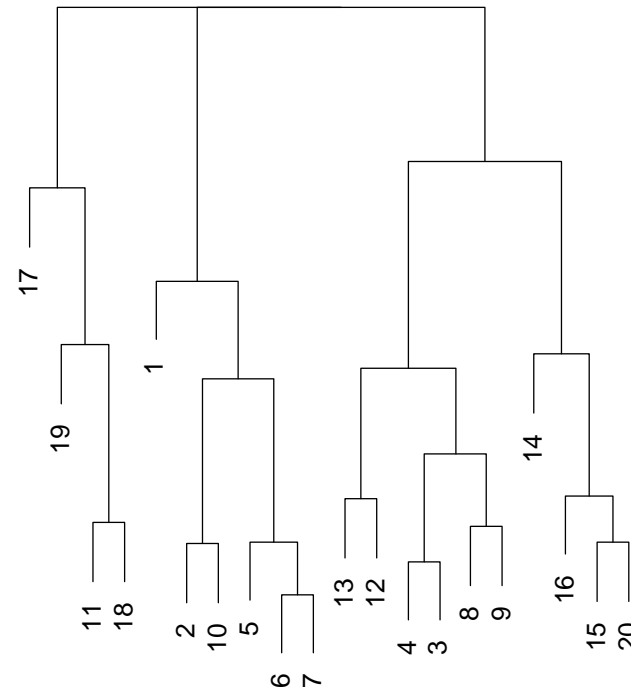
# Single & complete linkage

- Single linkage se výrazně řetězí

Bray-Curtis distance / Single linkage



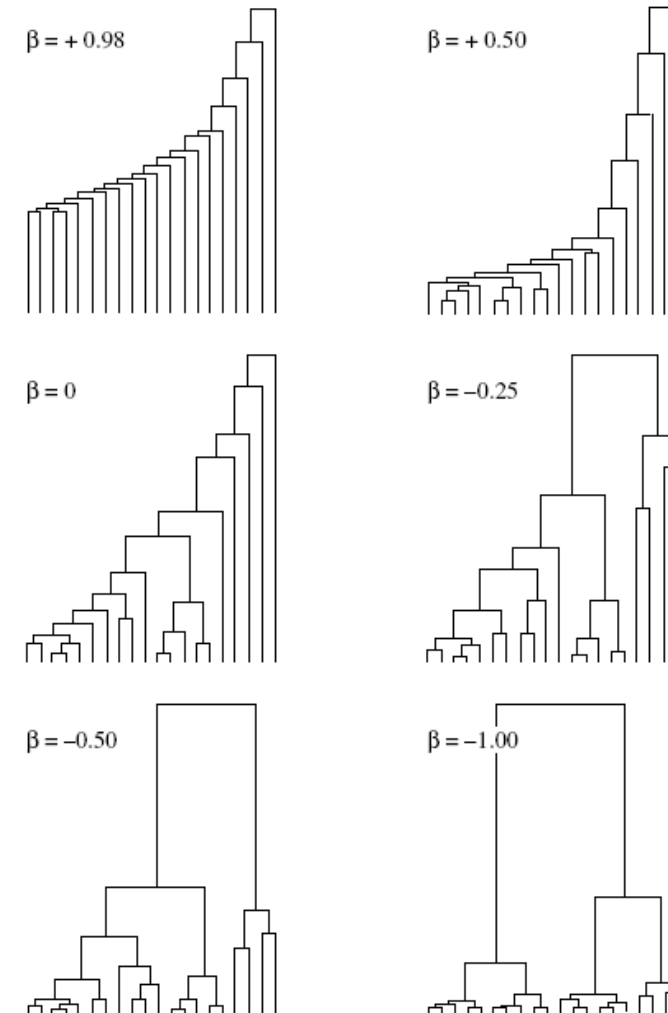
Bray-Curtis distance / Complete linkage



# Flexible clustering

## Beta flexible

- Nastavení parametru  $\beta$  ovlivňuje řetězení dendrogramu
- Nejvíc se řetězí pro  $\beta \sim 1$ , nejméně pro  $\beta = -1$
- optimální reprezentace vzdáleností mezi vzorky je při  $\beta = -0,25$



Legendre & Legendre (2012)



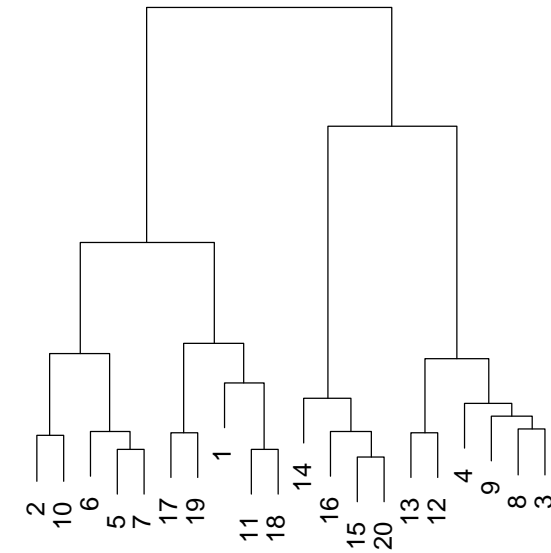
# Wardova metoda

`hclust()`

## Ward's minimum variance method (Ward, 1963)

- Ke shluku se připojí vzorek, jehož vzdálenost od centroidu shluku je nejmenší
- Počítáno přes čtverce vzdáleností mezi vzorky a centroidy shluků (stejně jako v k-means clustering)
- Vzdálenosti od centroidu počítá **euklidovskými vzdálenostmi** → podmínkou je, aby bylo možné s maticí vzdáleností pracovat v euklidovském prostoru

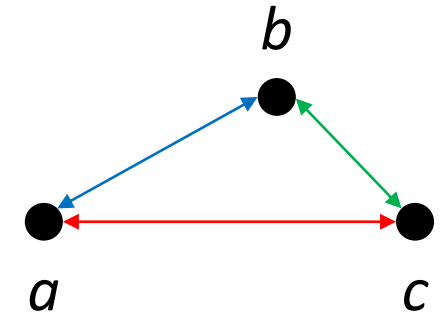
Euclidean distance / Ward's method



# Indexy vzdálenosti v euklidovském prostoru

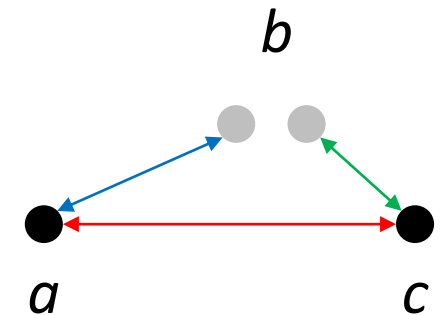
## 1. Metrické indexy

- Minimum = 0: jestliže  $a = b$ , potom  $D(a, b) = 0$
- Všechny hodnoty jsou pozitivní: jestliže  $a \neq b$ , potom  $D(a, b) > 0$
- Jsou symetrické:  $D(a, b) = D(b, a)$
- Splňují podmínku trojúhelníkové nerovnosti:  $D(a, b) + D(b, c) \geq D(a, c)$ 
  - vzdálenosti mohou být vynášeny v euklidovském prostoru
  - „euklidovský index“



## 2. Semimetrické indexy

- Nesplňují podmínku trojúhelníkové nerovnosti:  $D(a, b) + D(b, c) < D(a, c)$ 
  - vzdálenosti NEmohou být vynášeny v euklidovském prostoru
  - „neeuklidovský index“



## 3. Nemetrické indexy

- Mají negativní vzdálenosti

# Semimetrické („neeuclidovské“) indexy

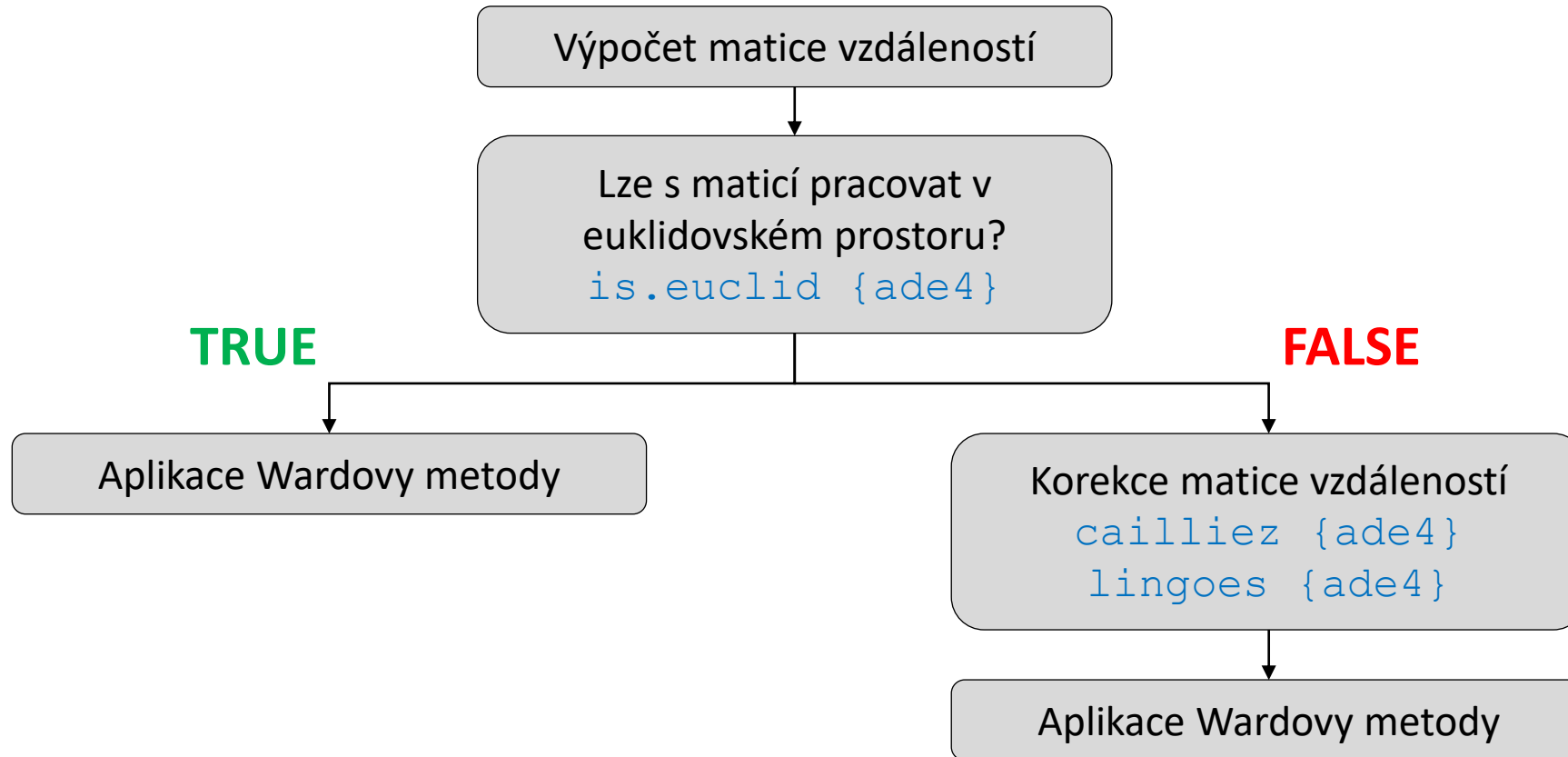
- V případě, že vzdálenosti (ekologické nepodobnosti) nemohou být vynášeny euklidovskou geometrií, lze použít korekci
- Cailliez (1983) cailliez {ade4}
  - korekce spočívá ve vyhledání nejmenší možné hodnoty (konstanty), kterou následně přičte ke všem členům matice vzdáleností
- Lingoes (1971) lingoes {ade4}
  - vyhledá hodnotu (konstantu), kterou transformuje matici vzdáleností podle rovnice

$$D'_{ij} = \sqrt{D_{ij}^2 + 2 \times k}$$

Cailliez, F. (1983) The analytical solution of the additive constant problem. *Psychometrika*, 48: 305–310.

Lingoes, J.C. (1971) Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, 36: 195–203.

# Aplikace Wardovy metody



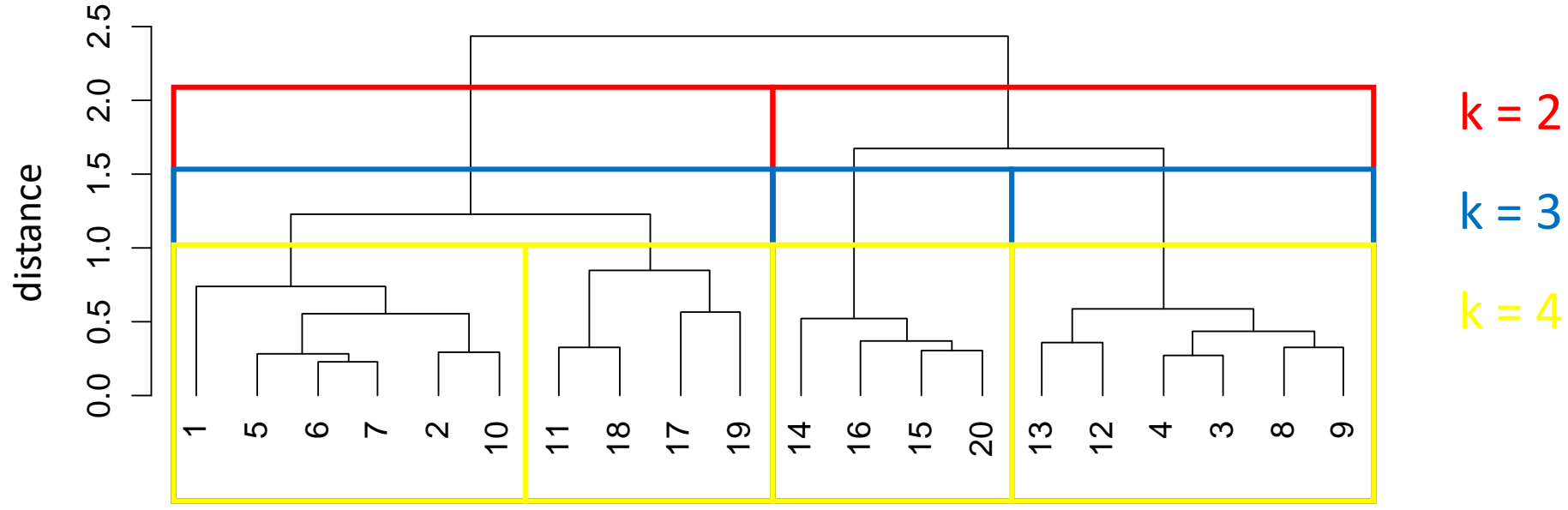


# Aplikace Wardovy metody

`hclust()`

- V literatuře se objevují dva různé způsoby výpočtu Wardovy metody
  1. Při tvorbě klastrů jsou používány prosté vzdálenosti → neodpovídá původnímu Wardovu (1963) algoritmu
    - ve funkci `hclust` atribut `ward.D`
    - ve verzích R  $\leq$  3.0.3 je pouze tento algoritmus (označen jako `ward`)
  2. Při tvorbě klastrů jsou používány čtverce vzdáleností → původní Wardův (1963) algoritmus
    - ve funkci `hclust` atribut `ward.D2`

# Jaký je „optimální“ počet shluků?



# Silhouette

- Hodnotí stupeň podobnosti daného vzorku ke klastru, do kterého byl zařazen, a srovnává ho s jeho podobností k nejbližšímu jinému klastru
- Negativní hodnoty – tyto vzorky byly s velkou pravděpodobností špatně klasifikovány (ve skutečnosti patří jinam)

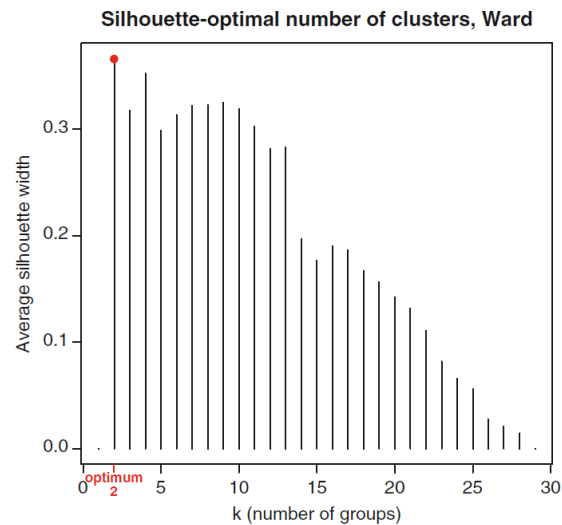


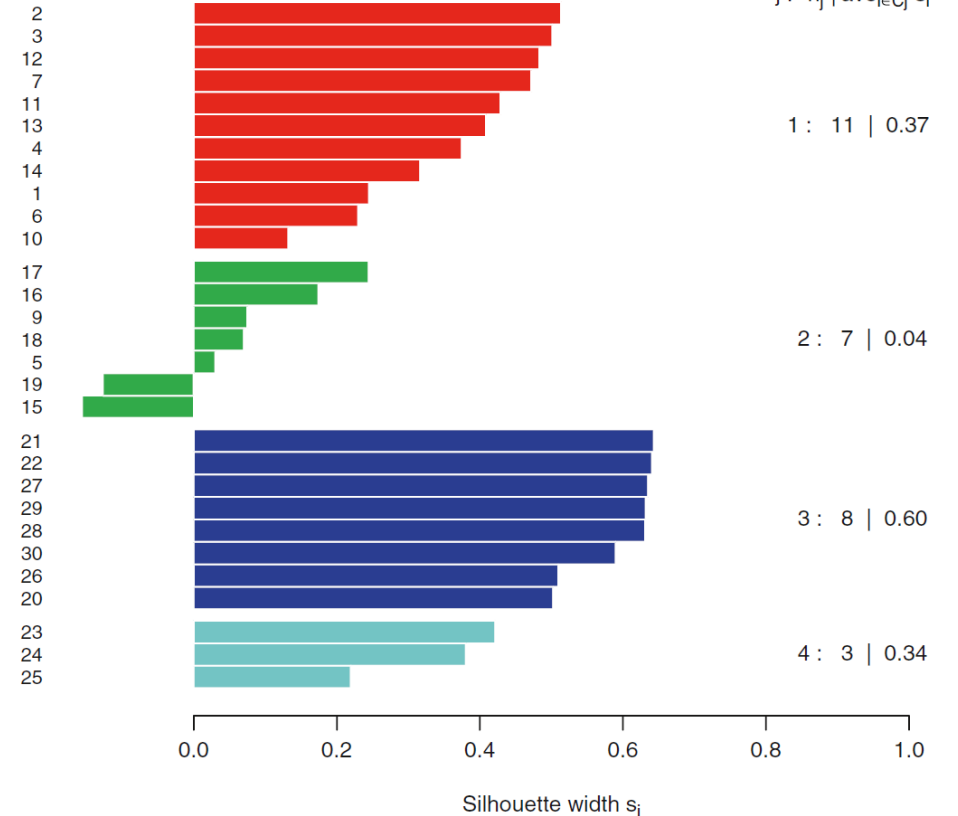
Fig. 4.8 Bar plot showing the average silhouette widths for  $k=2-29$  groups. The best partition by this criterion is the one with the largest average silhouette width

Silhouette plot - Chord - Ward

n = 29

4 clusters  $C_j$

$j : n_j \mid \text{ave}_{i \in C_j} s_i$



Average silhouette width : 0.35

Borcard et al. (2011)

# Mantelova korelace

- Porovnává matici vzdáleností vzorků s binární maticí zařazení vzorků do jednotlivých shluků
  - 1 – pokud jsou dva vzorky ve stejném shluku
  - 0 – pokud je každý v jiném shluku

Matice vzdáleností

	Samp1	Samp2	Samp3	Samp4	Samp5
Samp1	0.50	0.39	0.50	0.42	
Samp2		0.57	0.68	0.67	
Samp3			0.49	0.57	
Samp4				0.67	
Samp5					

Matice zařazení snímků

	Samp1	Samp2	Samp3	Samp4	Samp5
Samp1	0	1	1	1	
Samp2		0	0	0	
Samp3			1	0	
Samp4				0	
Samp5					

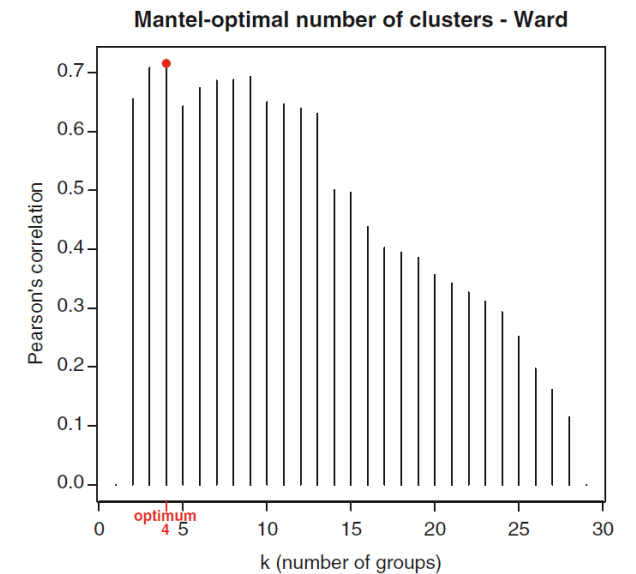


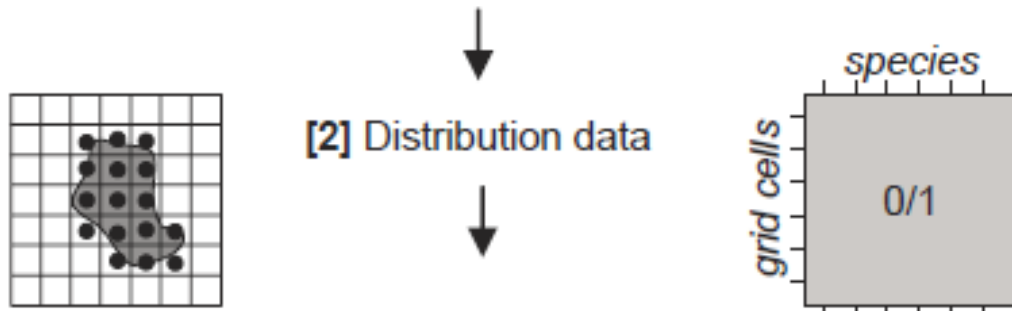
Fig. 4.9 Bar plot showing the correlations between the original distance matrix and binary matrices computed from the dendrogram cut at various levels

# A framework for delineating biogeographical regions based on species distributions

Holger Kreft<sup>1,2\*</sup> and Walter Jetz<sup>2,3</sup>



[1] Defining purpose and objectives

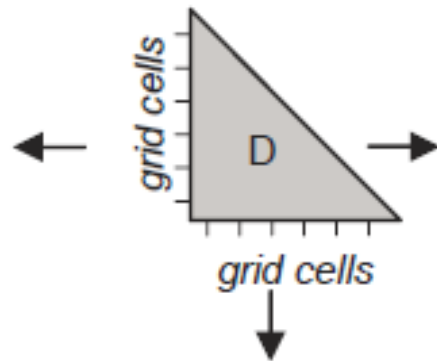


[3] Selection of distance metric (Jaccard, Sørensen,  $\beta$ -sim, Hellinger...)

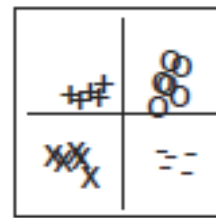
[4] Visualizing turnover



Distance matrix



[5] Ordination

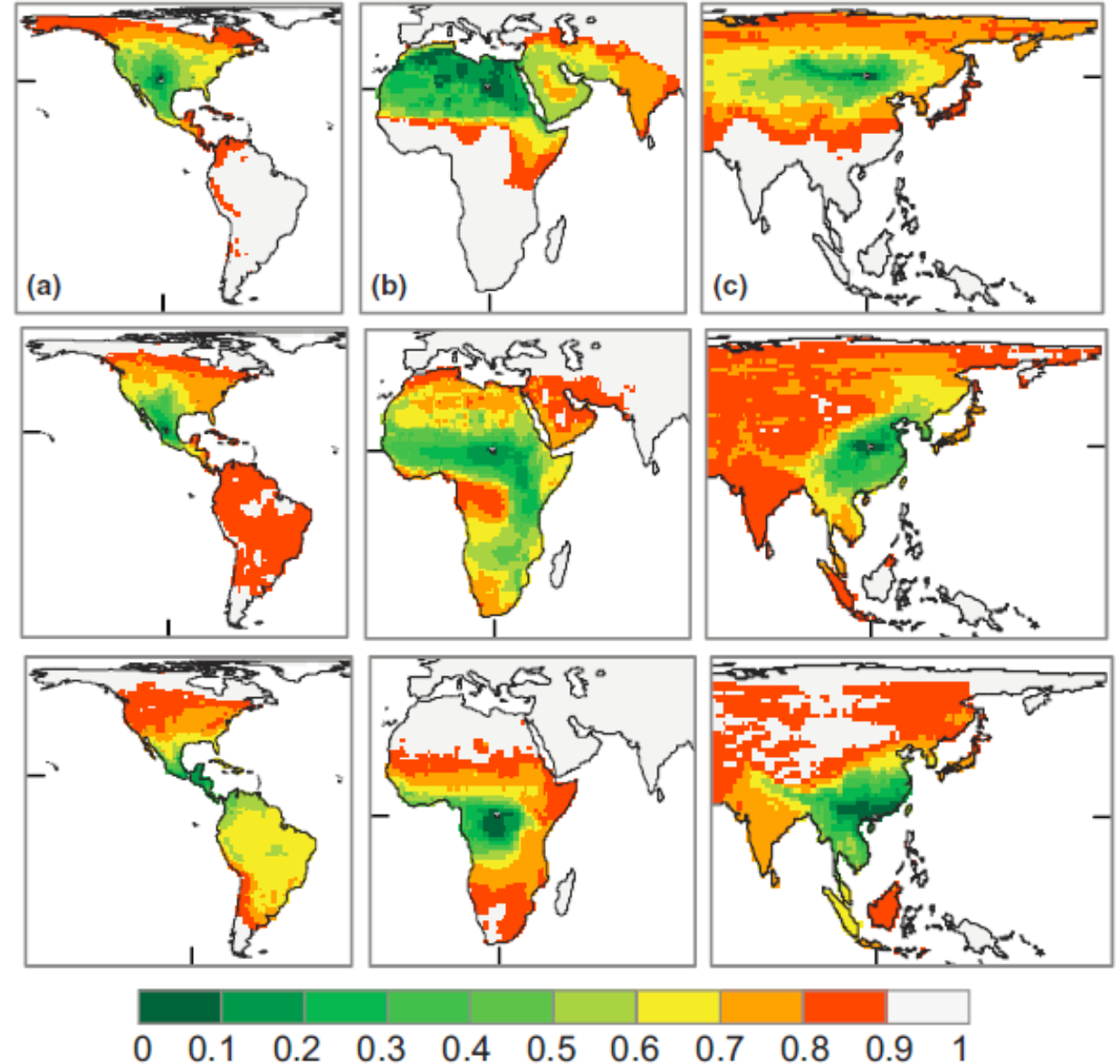


Map of results



# A framework for delineating biogeographical regions based on species distributions

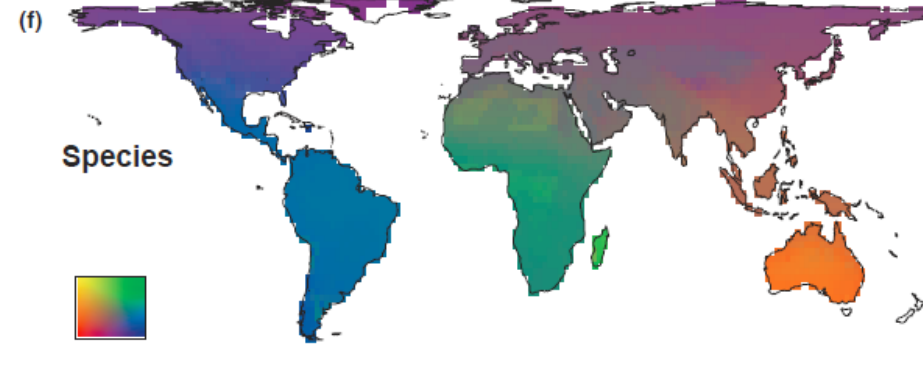
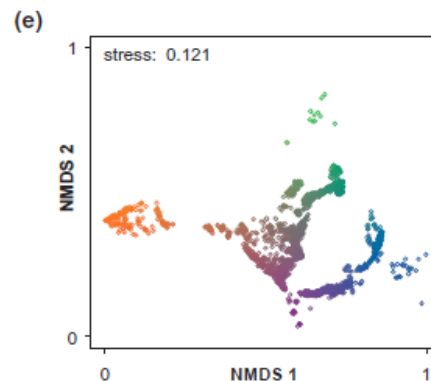
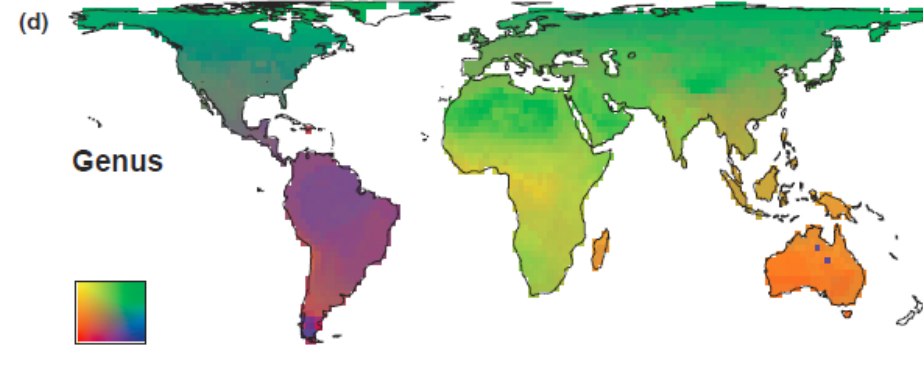
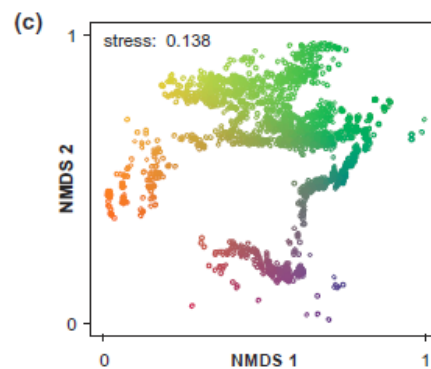
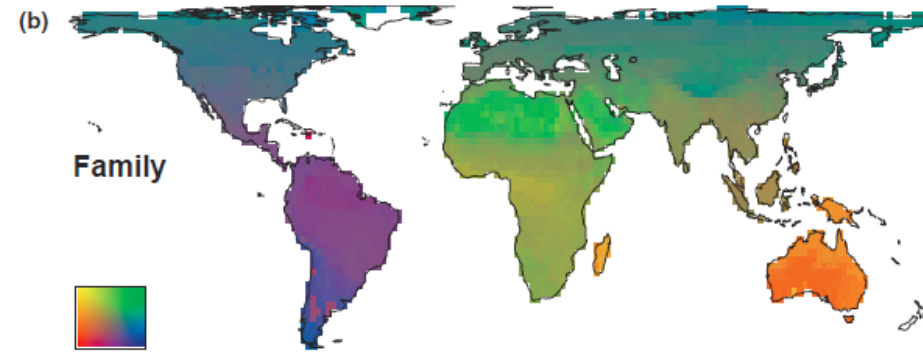
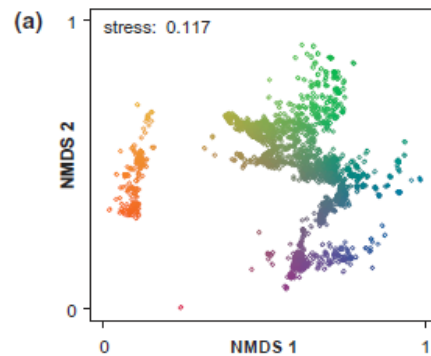
Holger Kreft<sup>1,2\*</sup> and Walter Jetz<sup>2,3</sup>





# A framework for delineating biogeographical regions based on species distributions

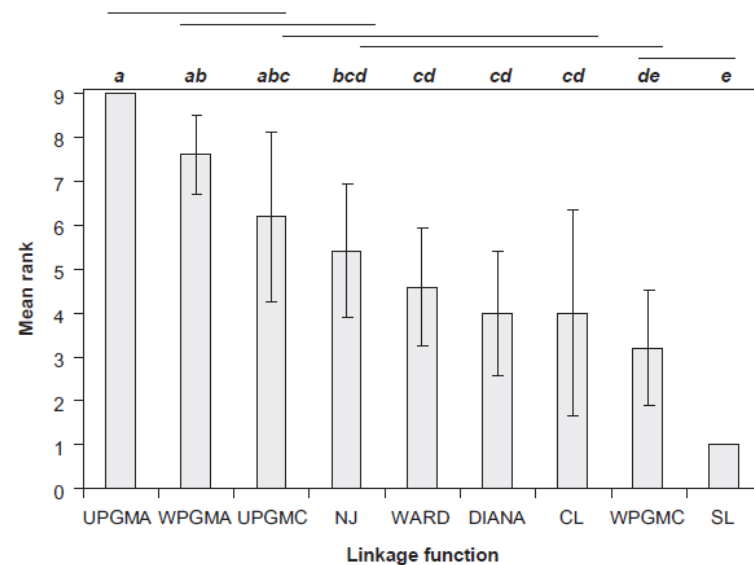
Holger Kreft<sup>1,2\*</sup> and Walter Jetz<sup>2,3</sup>



SPECIAL  
PAPER

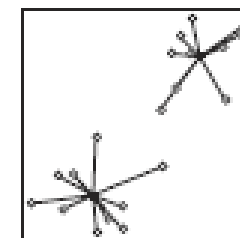
# A framework for delineating biogeographical regions based on species distributions

Holger Kreft<sup>1,2\*</sup> and Walter Jetz<sup>2,3</sup>

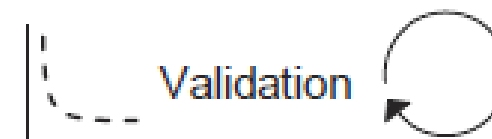
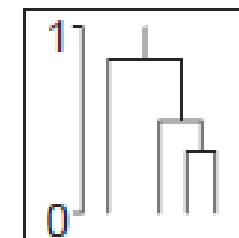


## [6] Cluster analysis

Non-hierarchical



Hierarchical



Map of results

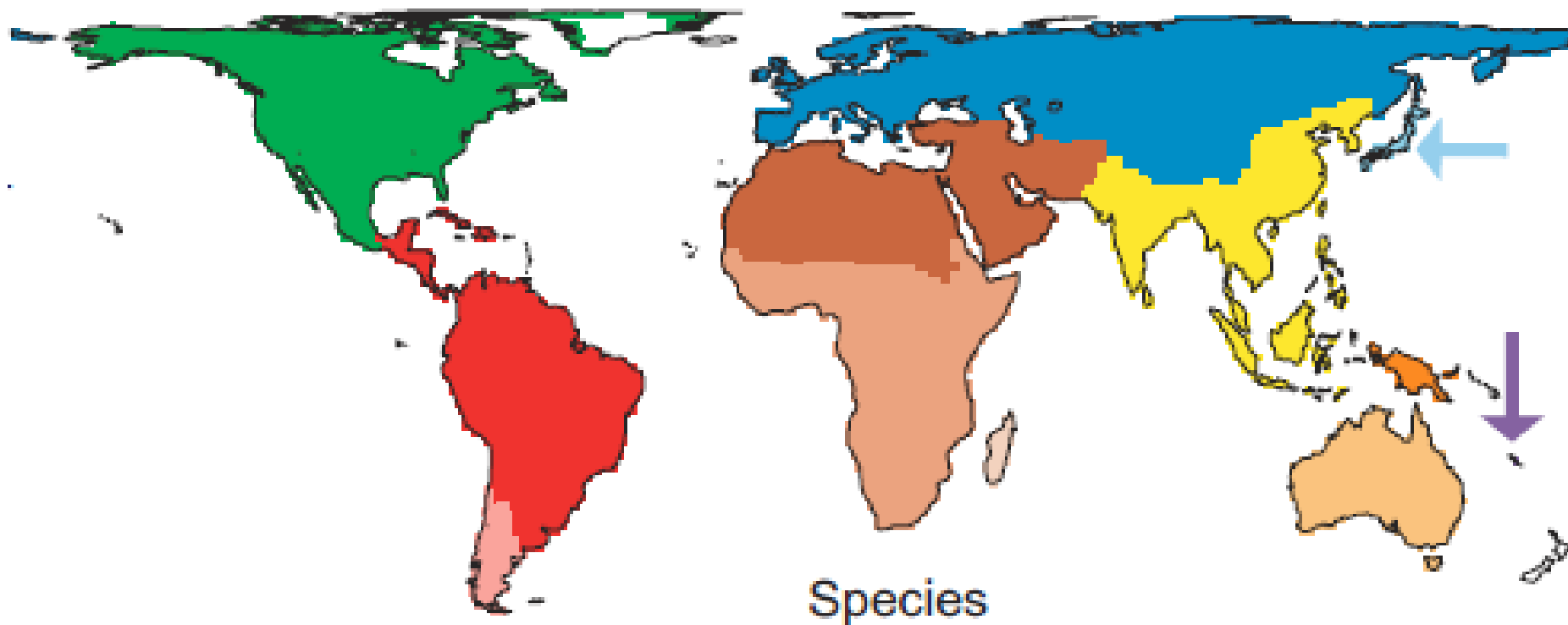
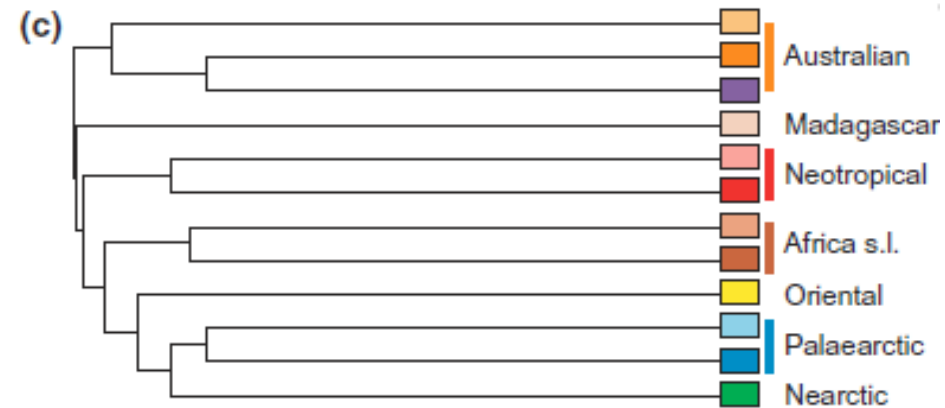
Spatial framework for further historical, ecological, and conservation biogeographical studies





# A framework for delineating biogeographical regions based on species distributions

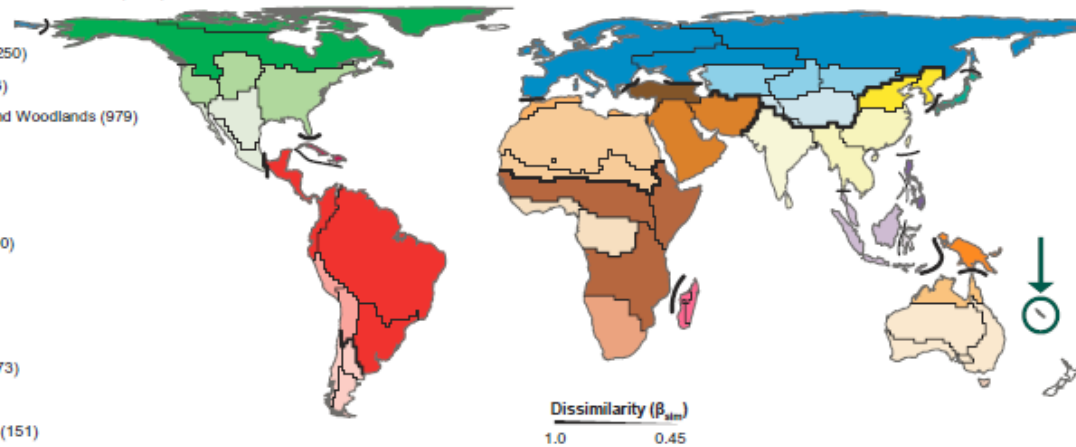
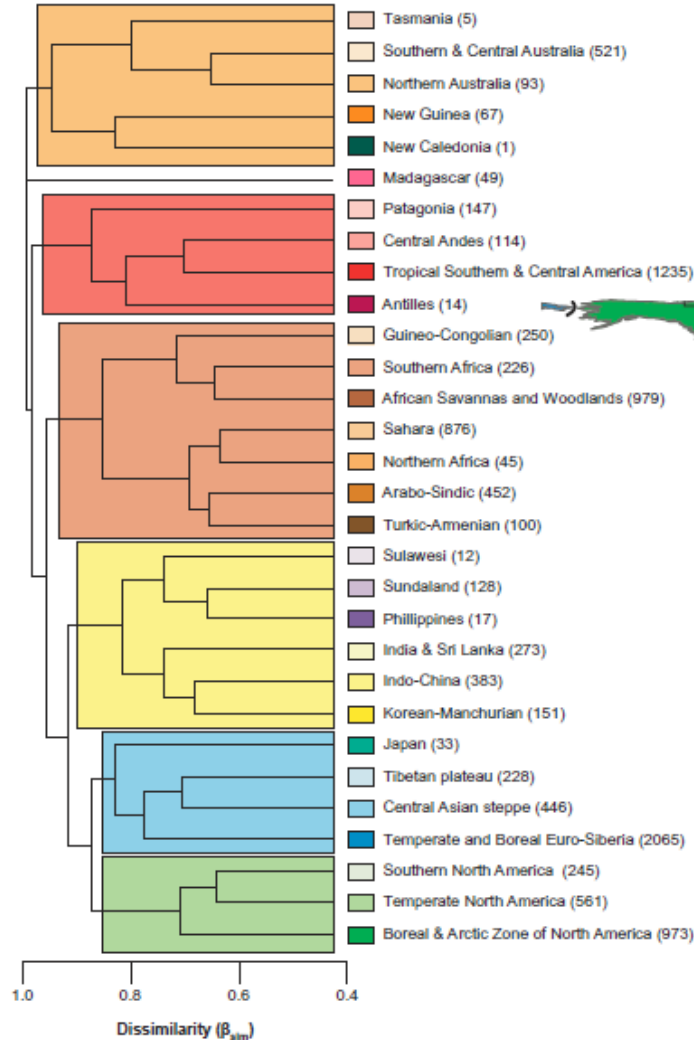
Holger Kreft<sup>1,2\*</sup> and Walter Jetz<sup>2,3</sup>



SPECIAL  
PAPER

# A framework for delineating biogeographical regions based on species distributions

Holger Kreft<sup>1,2\*</sup> and Walter Jetz<sup>2,3</sup>



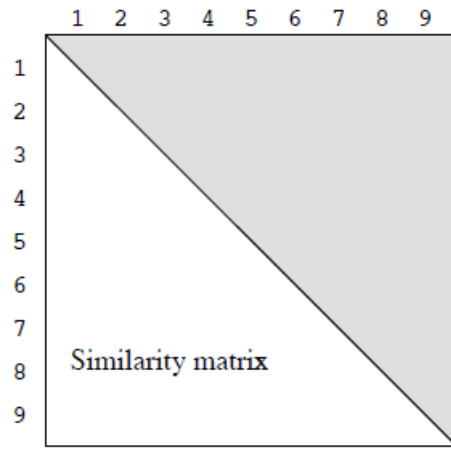
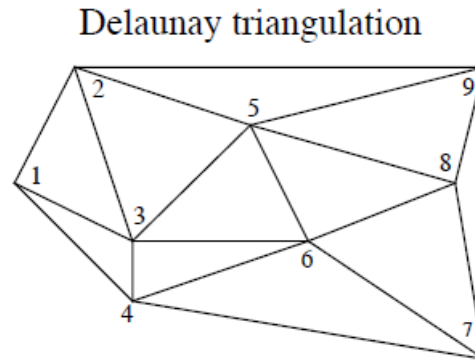
# Prostorově omezené shlukování

- Modifikace tradičních shlukovacích metod
- Shlukuje vzorky podle jejich podobnosti (vzdálenosti), ale pouze ty, které spolu sousedí podle zvoleného schématu propojenosti vzorků
- Do procesu shlukování vstupuje:
  1. Matice vzdáleností (nepodobností) vzorků (*distance/dissimilarity matrix*)
  2. Matice propojenosti vzorků (*connectivity matrix*)
    - 1 = propojené (sousedící) vzorky
    - 0 = nepropojené (nesousedící) vzorky
- Výhody
  - Vytváří prostorově kompaktní shluky, které jsou často lépe interpretovatelné
  - Díky prostorovým omezením dávají různé klasifikační algoritmy podobné výsledky
- Nevýhody
  - Shluky jsou vnitřně více heterogenní než v případě neomezené klasifikace
  - Výsledek značně závisí na volbě schématu propojenosti vzorků (další arbitrární rozhodnutí)

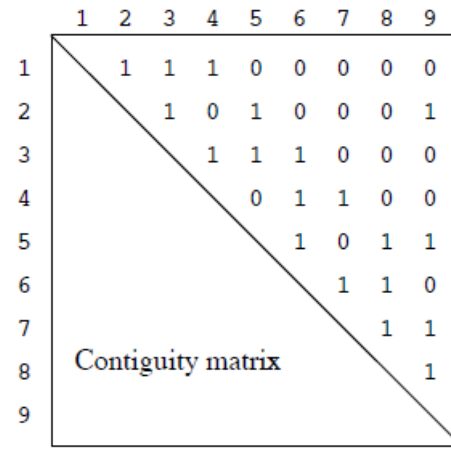
# Prostorově omezené shlukování

Data file

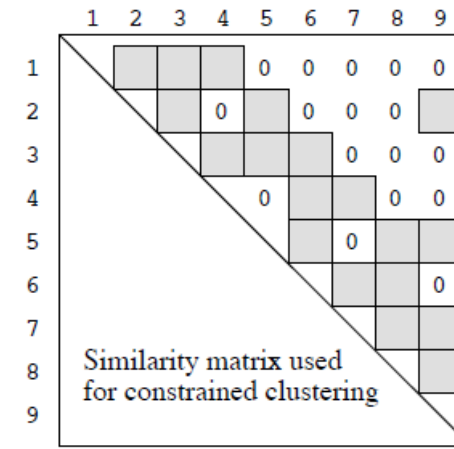
Sites	Descriptors
1	---
2	---
3	---
4	---
5	---
6	---
7	---
8	---
9	---



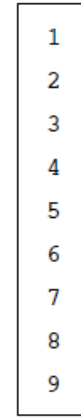
Hadamard product  
\*



=



Group membership

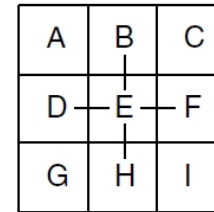


Legendre & Legendre (2012)

# Schémata propojenosti

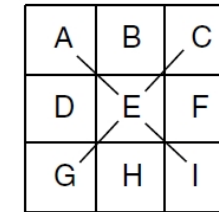
- Vzorky rozmístěné v pravidelné síti
  - Rook connection
  - Bishop connection
  - Queen connection
  - Kritérium vzdálenosti
- Vzorky rozmístěné nepravidelně
  - Delaunay triangulation
  - Gabriel criterion
  - Relative neighbourhood
  - Maximum distance
  - Minimum spanning tree (MST)

Rook



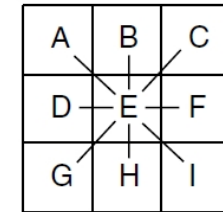
	A	B	C	D	E	F	G	H	I
A	-								
B	0	-							
C	0	0	-						
D	0	0	0	-					
E	0	1	0	1	-				
F	0	0	0	1	1	-			
G	0	0	0	0	0	0	-		
H	0	0	0	1	1	0	0	-	
I	0	0	0	0	0	0	0	0	-

Bishop



	A	B	C	D	E	F	G	H	I
A	-								
B	0	-							
C	0	0	-						
D	0	0	0	-					
E	1	0	1	0	-				
F	0	0	0	0	0	-			
G	0	0	0	0	1	0	-		
H	0	0	0	0	0	0	0	-	
I	0	0	0	0	1	0	0	0	-

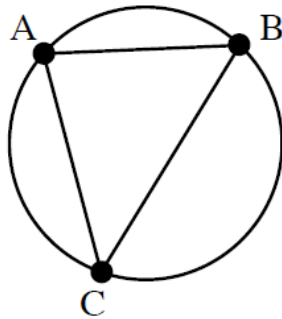
Queen



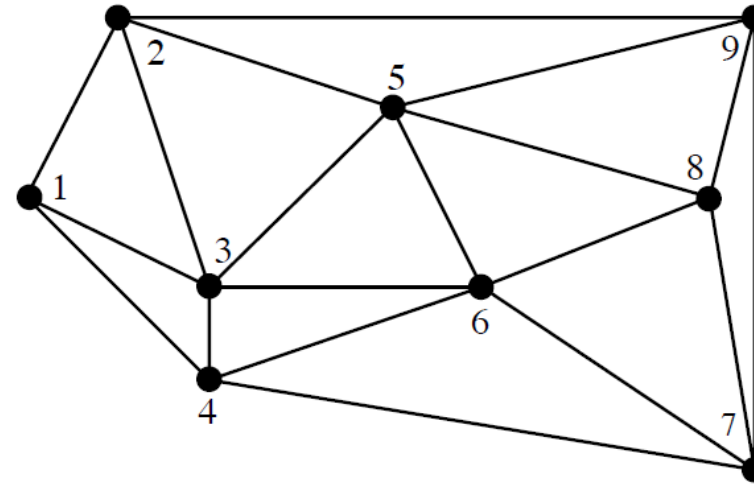
	A	B	C	D	E	F	G	H	I
A	-								
B	0	-							
C	0	0	-						
D	0	0	0	-					
E	1	1	1	1	-				
F	0	0	0	0	1	-			
G	0	0	0	0	1	0	-		
H	0	0	0	0	1	0	0	-	
I	0	0	0	0	1	0	0	0	-

# Delaunay triangulation

- Trojice vzorků je propojena pouze pokud kružnice, na které tyto vzorky leží neobsahuje žádný další bod



Point identifiers	Coordinates	
	X	Y
1	0	3
2	1	5
3	2	2
4	2	1
5	4	4
6	5	2
7	8	0
8	7.5	3
9	8	5



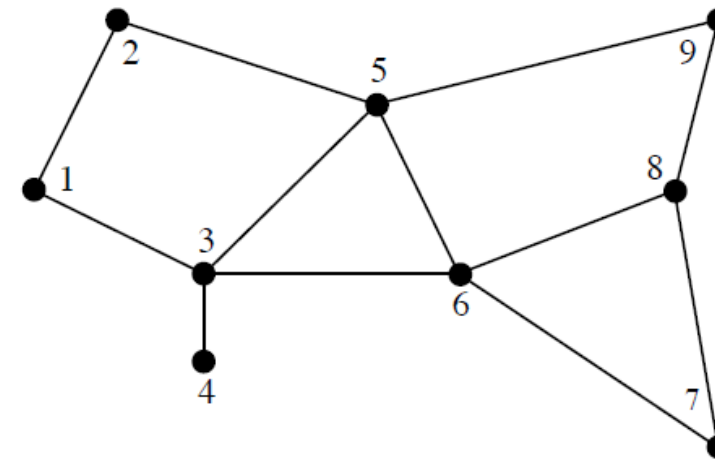
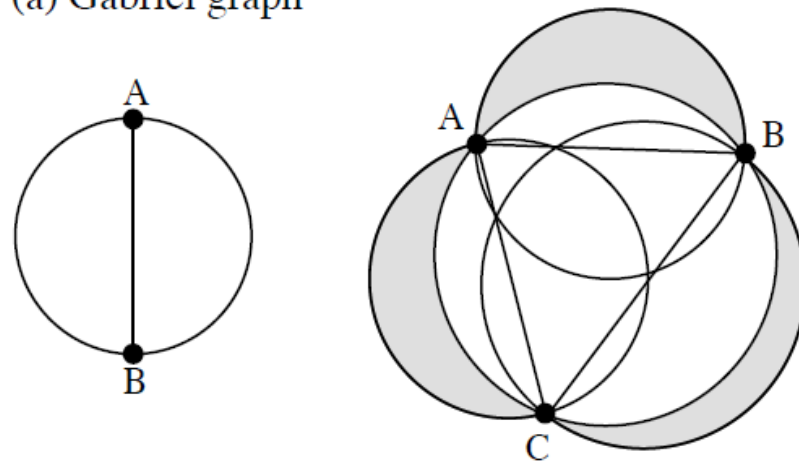
19 edges form the Delaunay triangulation:

1-2   1-3   1-4   2-3   2-5   2-9   3-4  
 3-5   3-6   4-6   4-7   5-6   5-8   5-9  
 6-7   6-8   7-8   7-9   8-9

# Gabriel criterion

- Dvojice vzorků je propojena pouze pokud kružnice procházející těmito vzorky neobsahuje žádný další vzorek

(a) Gabriel graph



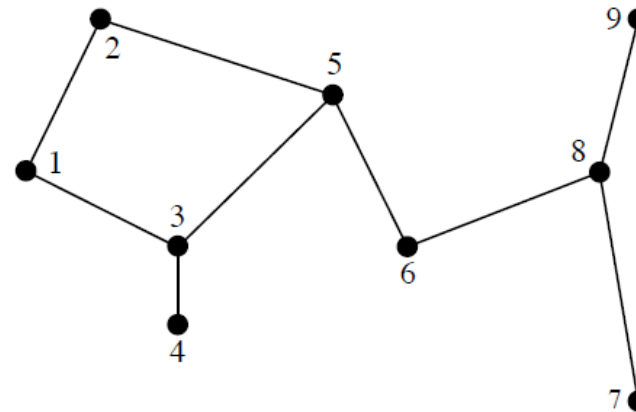
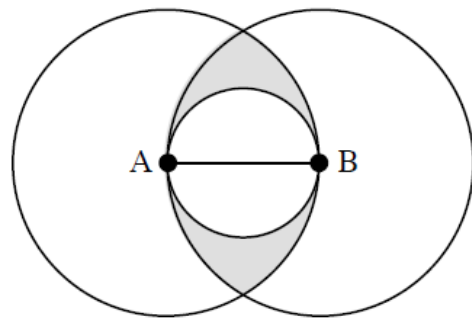
12 edges form the Gabriel graph:

1-2 1-3 2-5 3-4 3-5 3-6  
 5-6 5-9 6-7 6-8 7-8 8-9

# Relative neighbourhood

- Dvojice vzorků je propojena pouze pokud průnik kružnic, jejich poloměr se rovná vzdálenosti mezi body neobsahuje žádný další vzorek (vzorky tvoří středy kružnic)

(b) Relative neighbourhood graph



9 edges form the relative neighbourhood graph:

1-2 1-3 2-5 3-4 3-5 5-6  
6-8 7-8 8-9

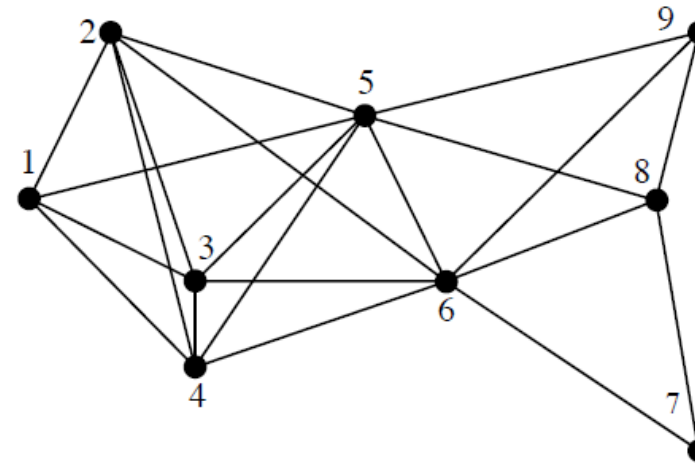


# Maximum distance

- Vzorčky jsou propojeny na základě kritéria vzdálenosti

(c) Maximum distance graph

Criterion:  $D \leq \text{threshold}$   
 In this example,  $D \leq 5$

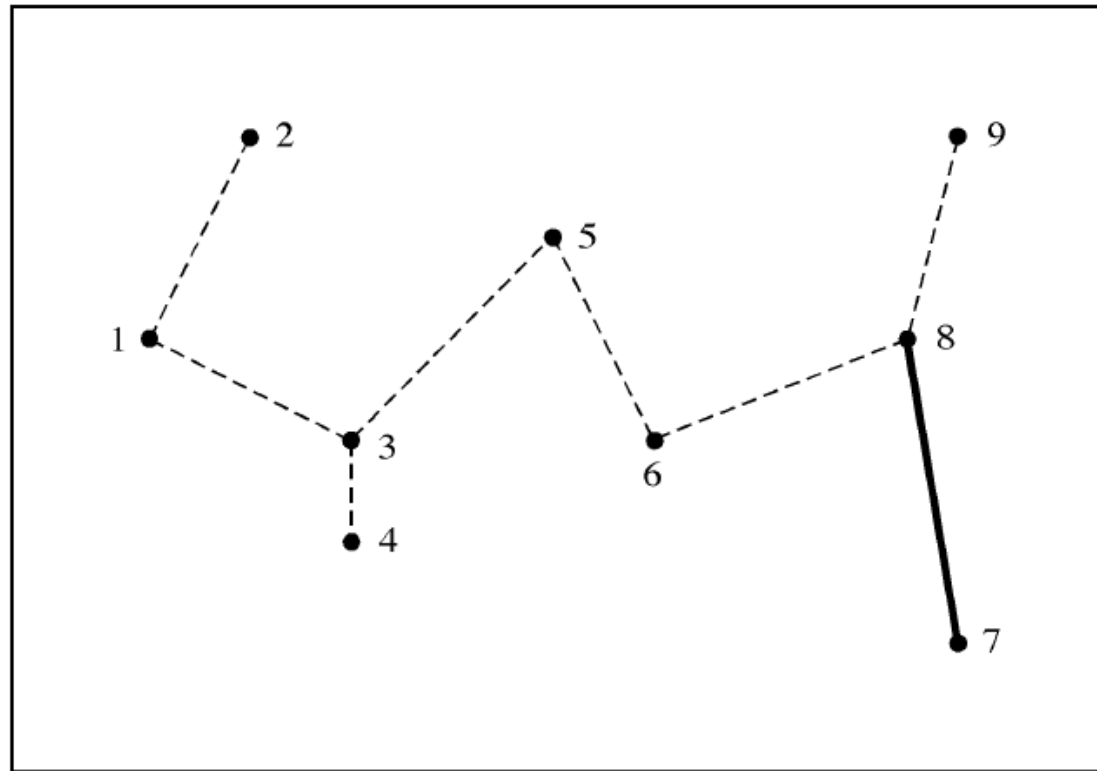


22 edges form the maximum distance graph:

1-2	1-3	1-4	1-5	2-3	2-4	2-5	2-6
3-4	3-5	3-6	4-5	4-6	5-6	5-8	5-9
6-7	6-8	6-9	7-8	7-9	8-9		

# Minimum spanning tree (MST)

- Spojuje  $n$  vzorků  $n-1$  spojnícemi tak, aby součet jejich délek byl minimální



# Software

- Pro hierarchické aglomerativní metody
  - funkce `constrained.clust` v knihovně `const.clust` (není na CRANu, ale [zde](#))
- Pro nehierarchické metody
  - spatially constrained k-means v S.A.M.
  - Grouping Analysis v ArcGIS



## Landscape classification of the Czech Republic based on the distribution of natural habitats

Klasifikace krajiny České republiky na základě rozšíření přírodních biotopů

Jan Divíšek<sup>1,2,3</sup>, Milan Chytrý<sup>3</sup>, Vít Grulich<sup>3</sup> & Lucie Poláková<sup>4</sup>

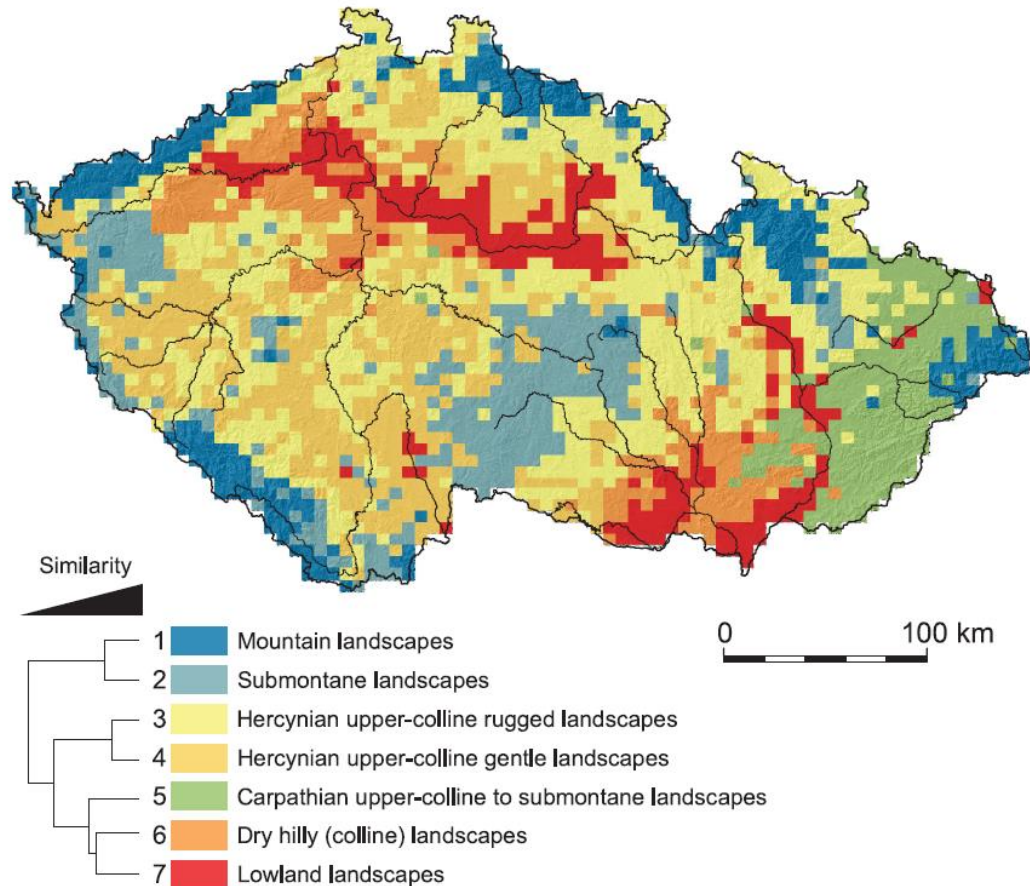


Fig. 3. – Landscape classification of the Czech Republic based on spatially unconstrained clustering with the optimal number of seven clusters according to the cross-validation procedure.

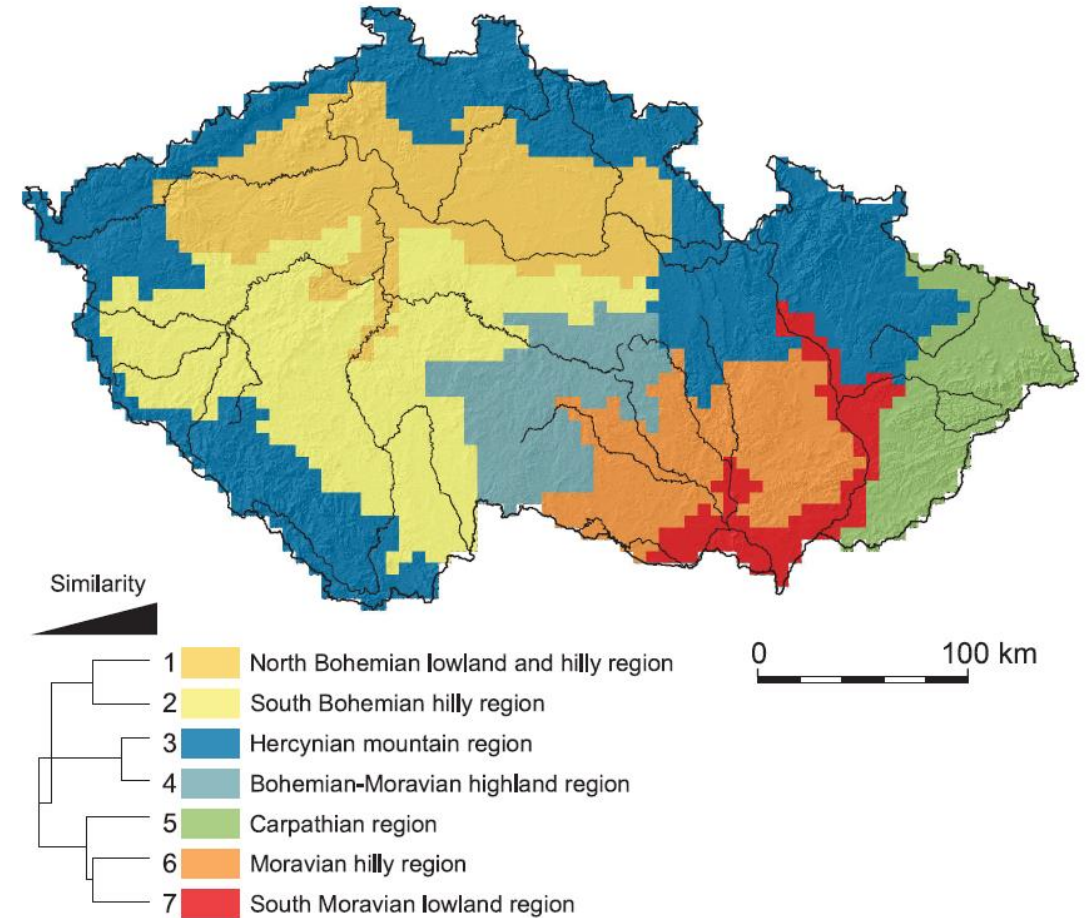


Fig. 7. – Regions of the Czech Republic based on the spatially constrained clustering with the optimal number of seven clusters according to cross-validation procedure. Reversal in the dendrogram is due to spatial constraints.

# Literatura

- Legendre, P. & Legendre, L. (2012): Numerical ecology. Third Edition. Elsevier, Amsterdam.
- Borcard, D., Gillet, F. & Legendre, P. (2011): Numerical ecology with R. Springer, New York.
- Koleff, P., Gaston, K.J. & Lennon, J.J. (2003): Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, 72(3): 367–382
- Fortin, M-J. & Dale, M.R.T. (2005): Spatial analysis: a guide for ecologists. Cambridge University Press. New York.