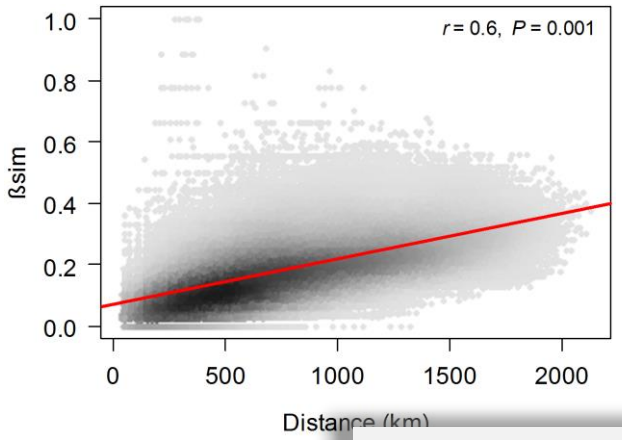
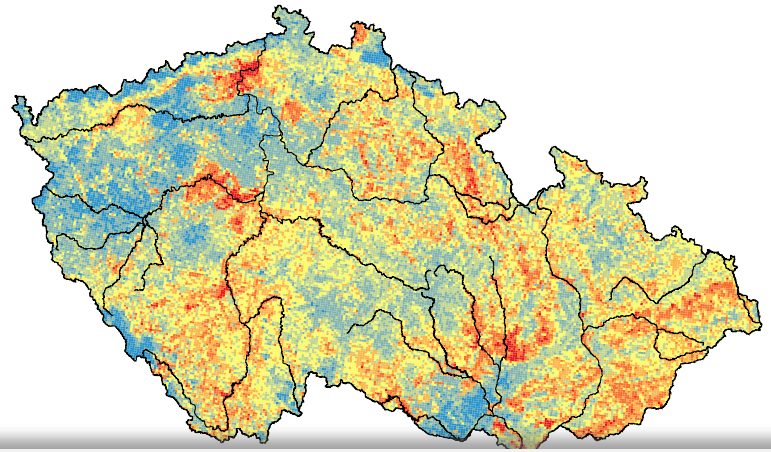
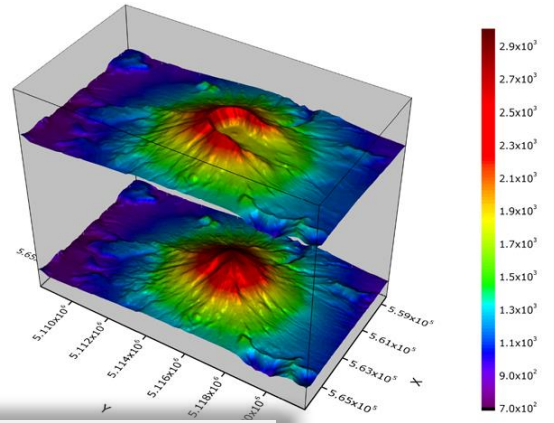
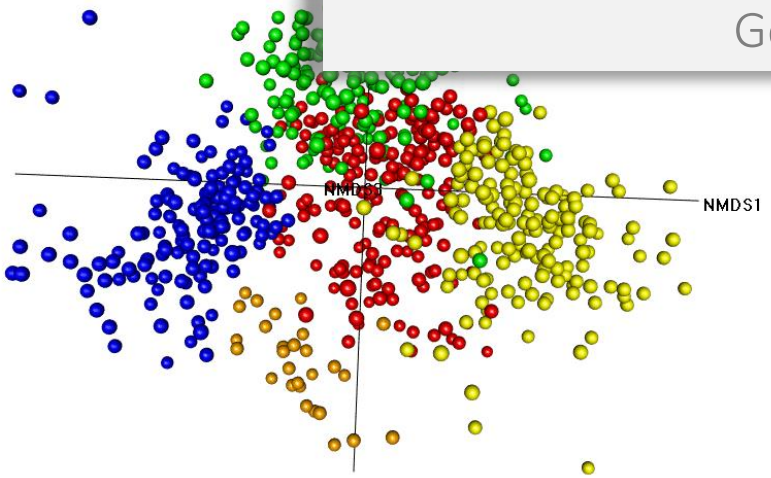


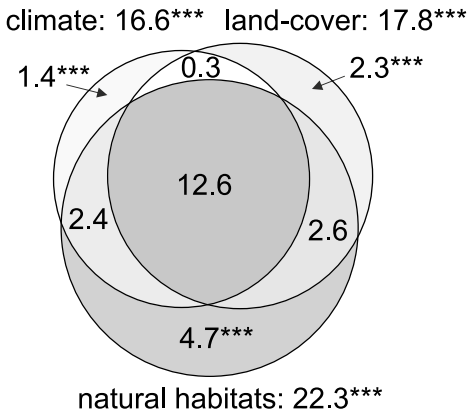




Metody fyzické geografie 3:
Biogeografie & ekologie

Jan Divíšek
Geografický ústav & Ústav botaniky a zoologie





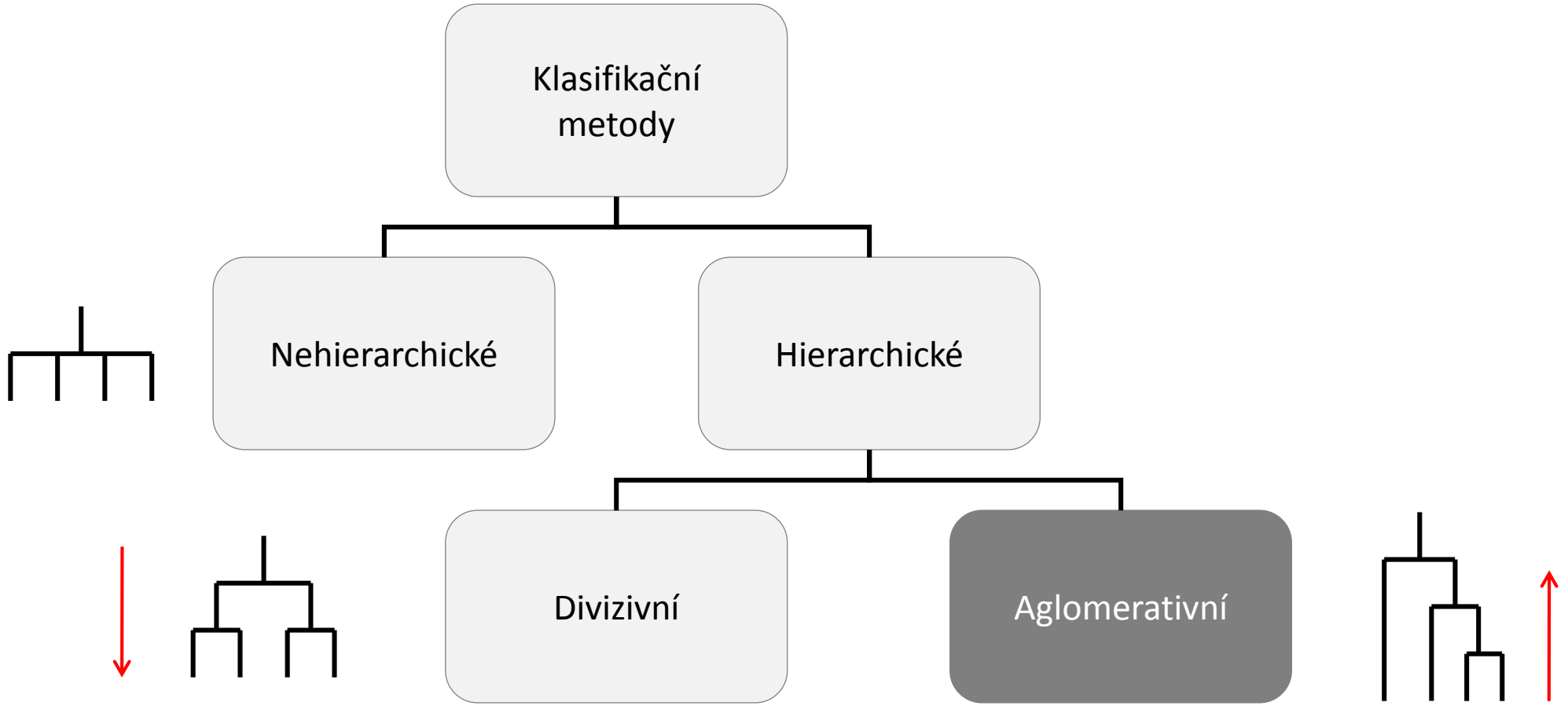
Metody fyzické geografie 3 – 14. 11. 2017

- Teoretická část
  - Hierarchické klasifikační metody
  - Výběr optimálního počtu klastrů

- Praktická část
  - UPGMA
  - Wardova metoda + transformace matice vzdáleností

Hierarchické klasifikační metody

Základní typy klasifikačních metod

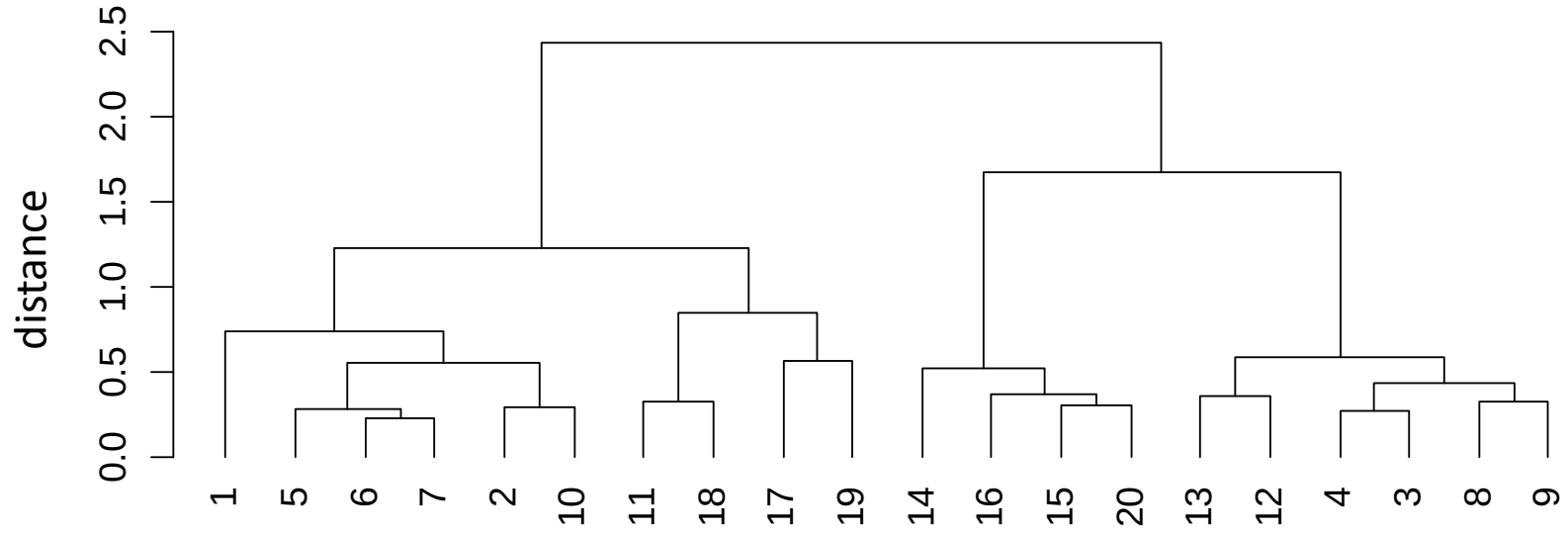


Hierarchická aglomerativní shluková analýza

- 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

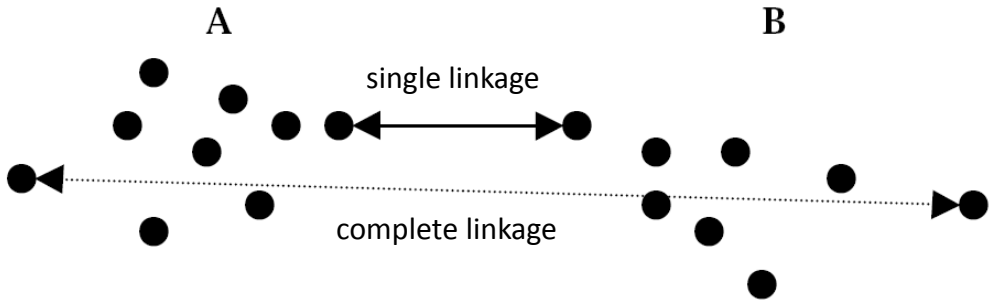
`hclust()`

Sigle 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 snímek



Jak funguje Single linkage?

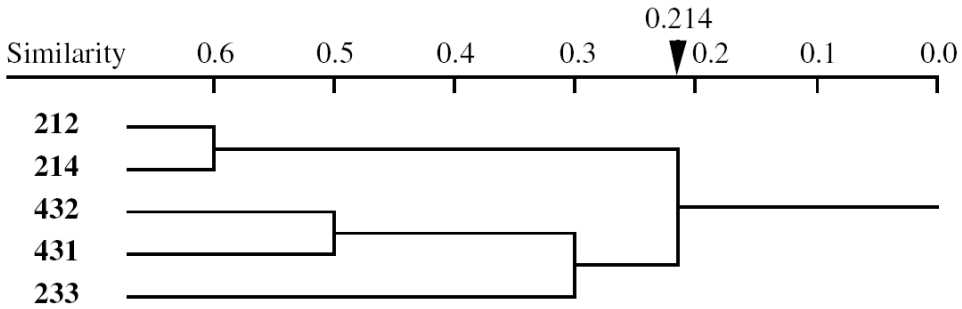
Matice podobností

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

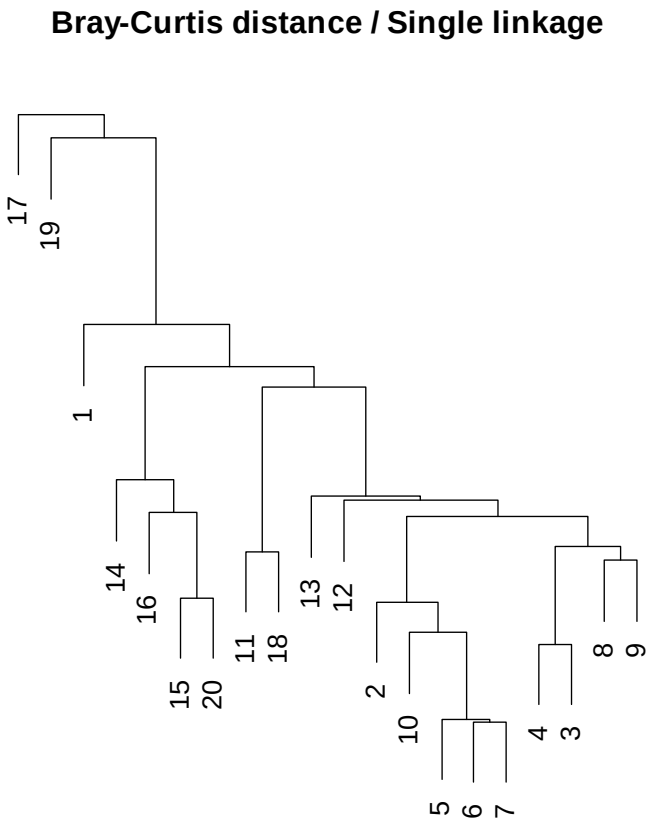
Výsledný dendrogram

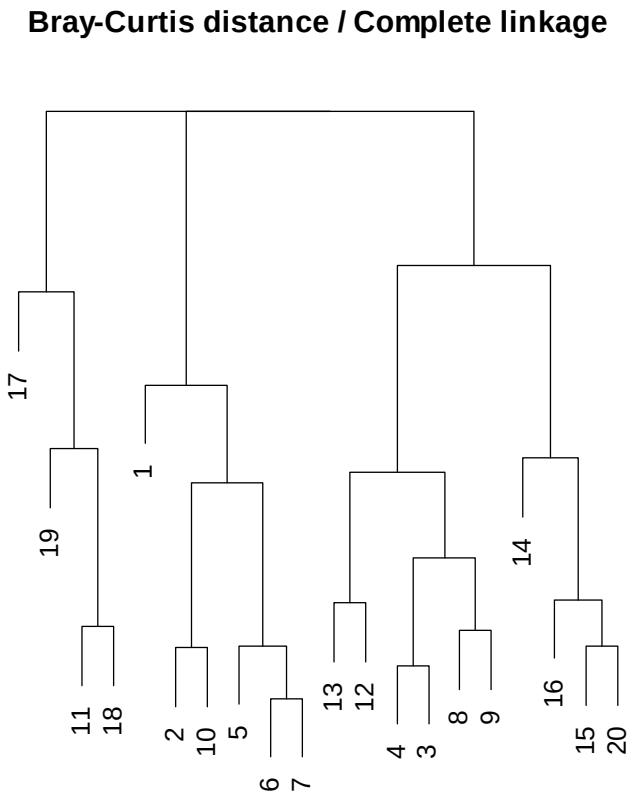


Legendre & Legendre (2012)

Single & complete linkage

- Single linkage se výrazně řetězí

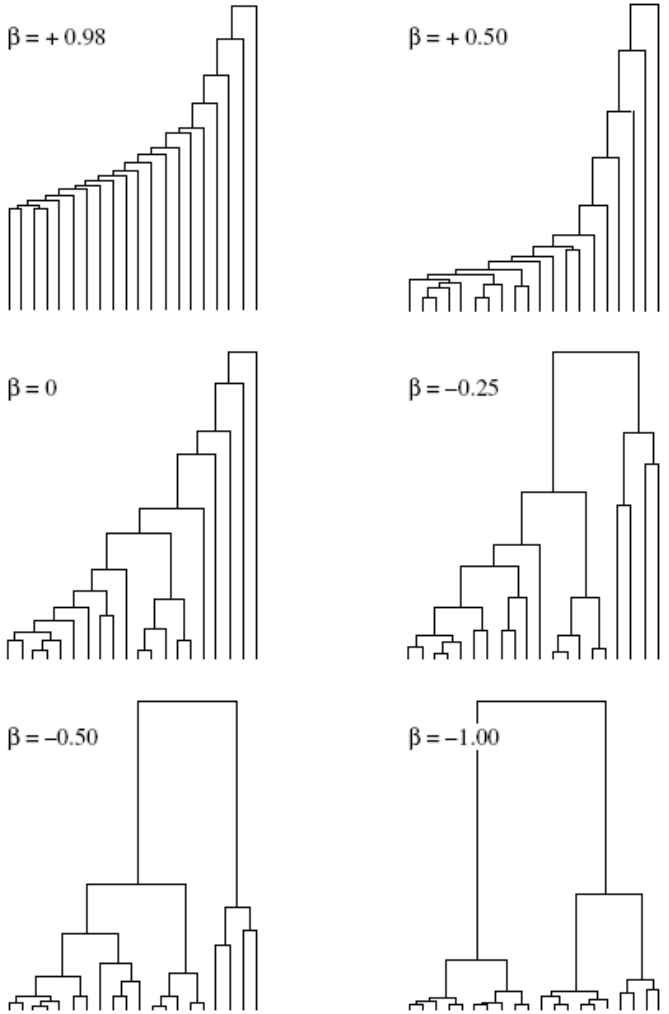




Flexible clustering

Beta flexible

- Nastavení parametru β ovlivňuje řetězení dendrogramu
- Nejvíc se řetězí pro β ~ 1, nejméně pro β = -1
- optimální reprezentace vzdáleností mezi vzorky je při β = -0,25

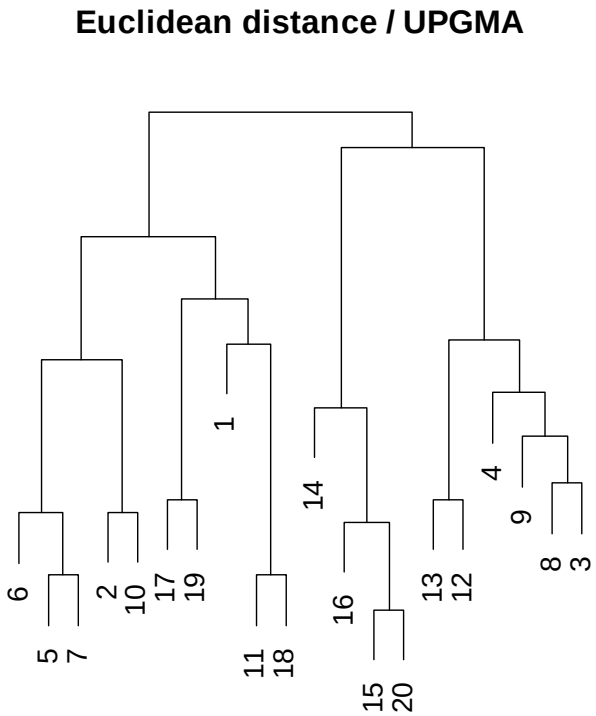


Legendre & Legendre (2012)

Average linkage

`hclust()`

- Zahrnuje řadu metod, které stojí mezi *single* a *complete linkage* a v ekologii jsou smysluplnější
- **Unweighted pair-group method using arithmetic averages** (UPGMA)
  - Vzorek se připojí ke shluku, ke kterému má největší (neváženou) průměrnou podobnost se všemi jeho vzorky
- Další metody
  - Weighted arithmetic average clustering (WPGMA)
  - Unweighted centroid clustering (UPGMC)
  - Weighted centroid clustering (WPGMC)



Úkol

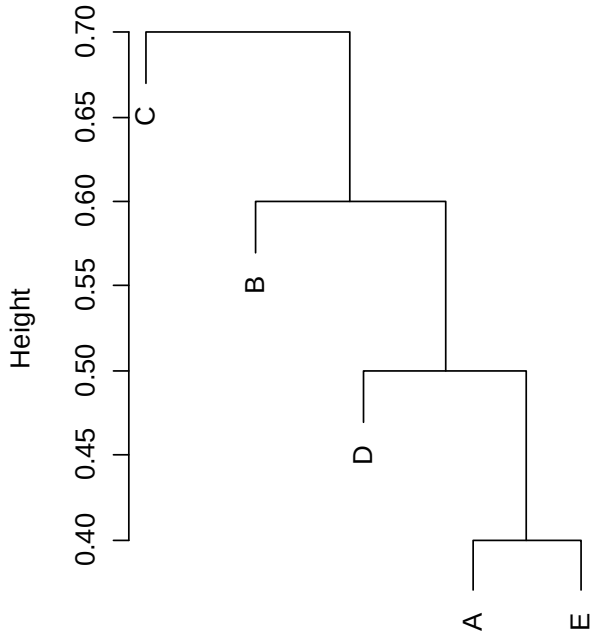
- Zkonstruujte ručně dendrogram metodami single linkage, complete linkage a average linkage

Matice podobností pěti vzorků A - E

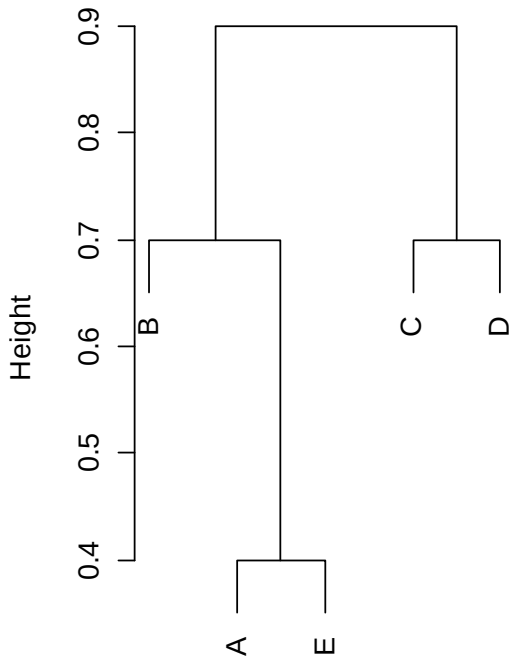
	A	B	C	D	E
A	100				
B	30	100			
C	20	10	100		
D	50	10	30	100	
E	60	40	20	10	100

Řešení

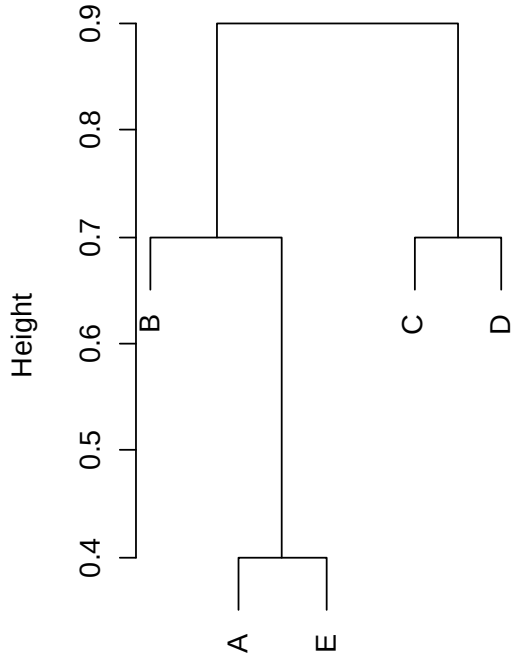
Single linkage



Complete linkage



UPGMA

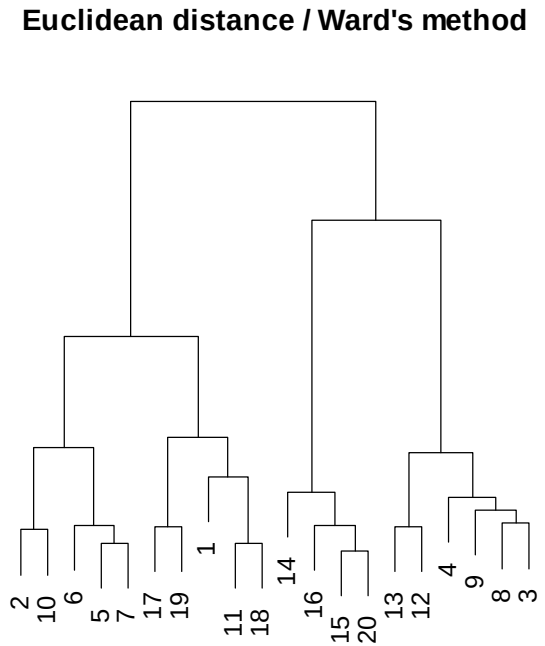


Wardova metoda

`hclust()`

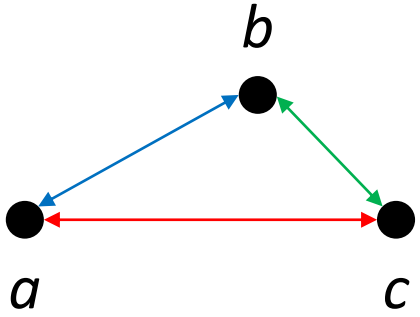
Ward's minimum variance method (Ward, 1963)

- 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

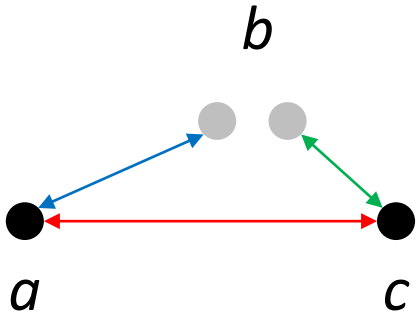


Indexy vzdálenosti v euklidovském prostoru

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

Legendre & Legendre (2012)

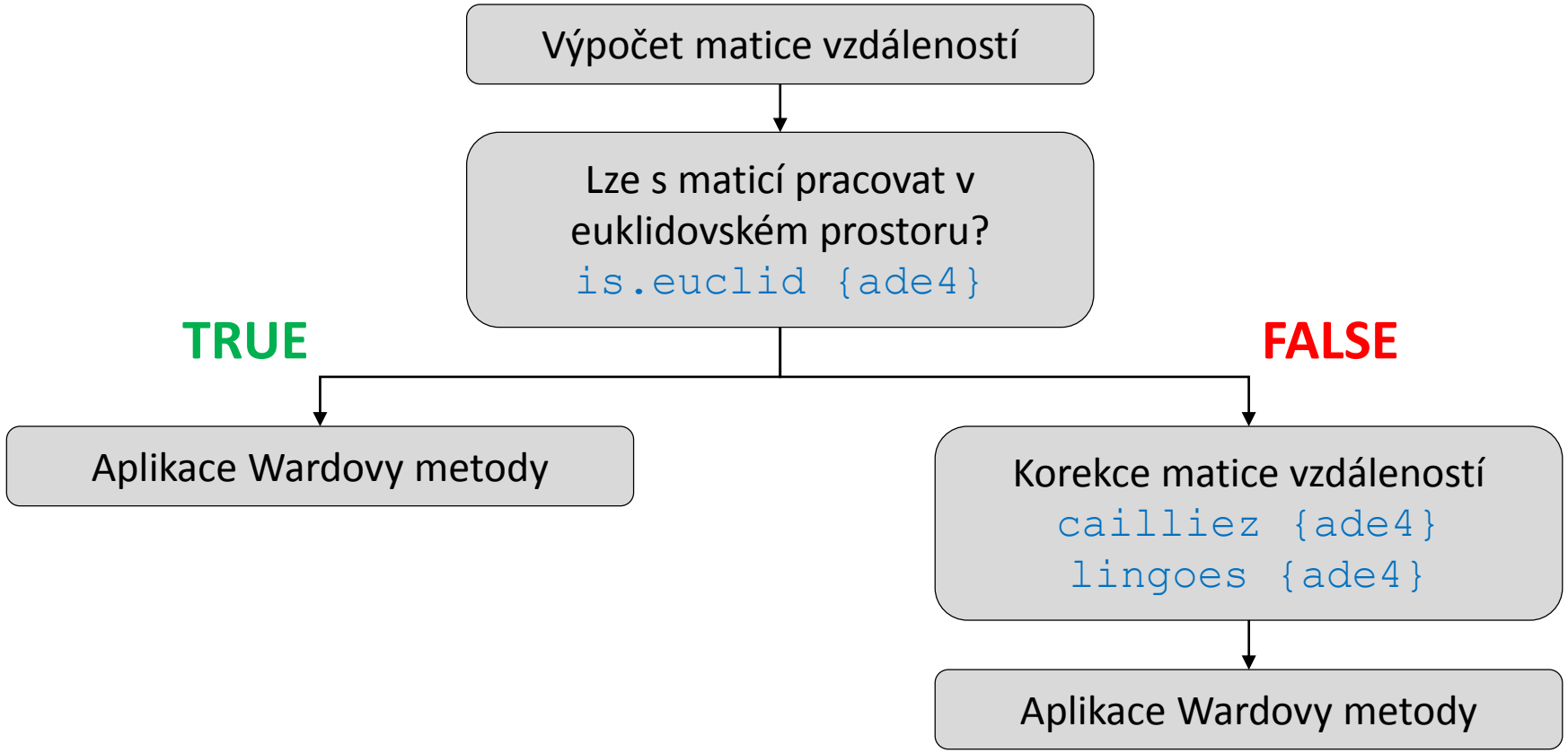
Semimetrické („neeuklidovské“) indexy

- V případě, že vzdálenosti (ekologické nepodobnosti) nemohou být vynášeny euklidovskou geometrií, lze použít korekci
- Cailiez (1983) `cailliez {ade4}`
  - korekce spočívá ve vyhledání nejmenší možné hodnoty (konstanty), kterou následně přičte ke všem členům matice vzdáleností
- Lingoez (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

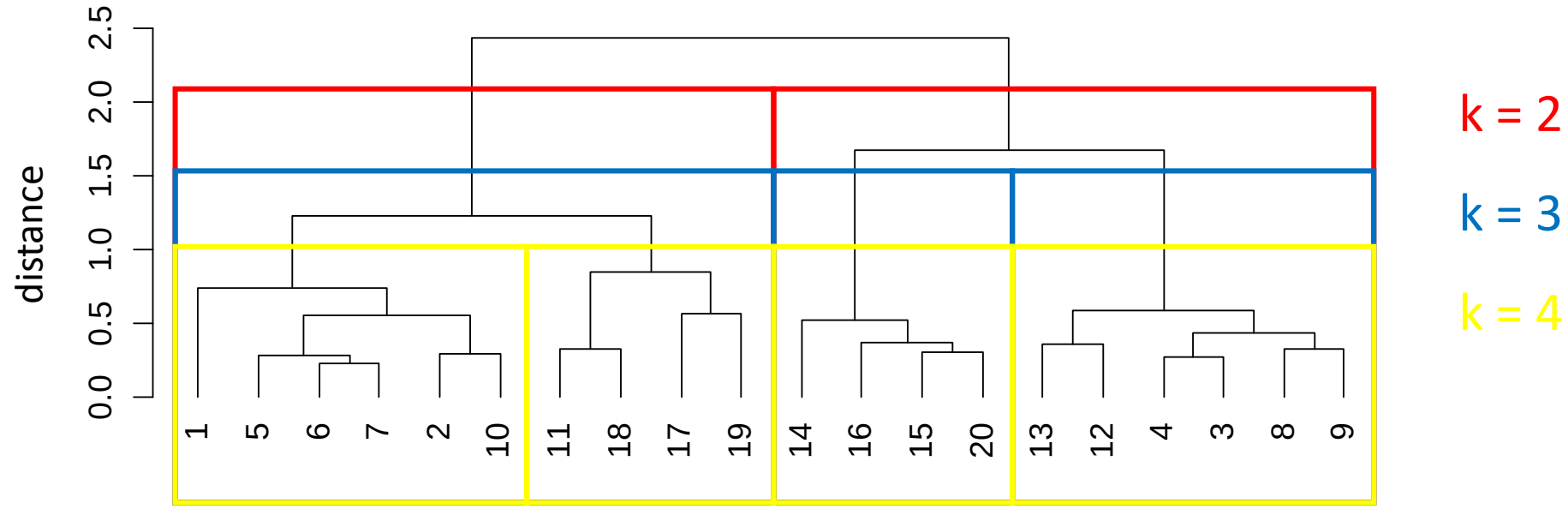


Aplikace Wardovy metody

`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 <= 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)

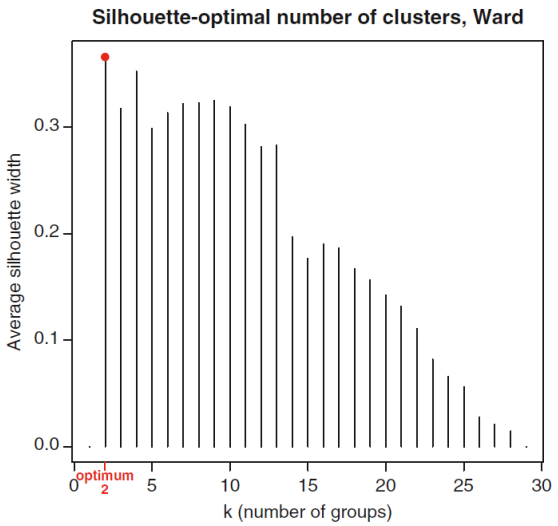
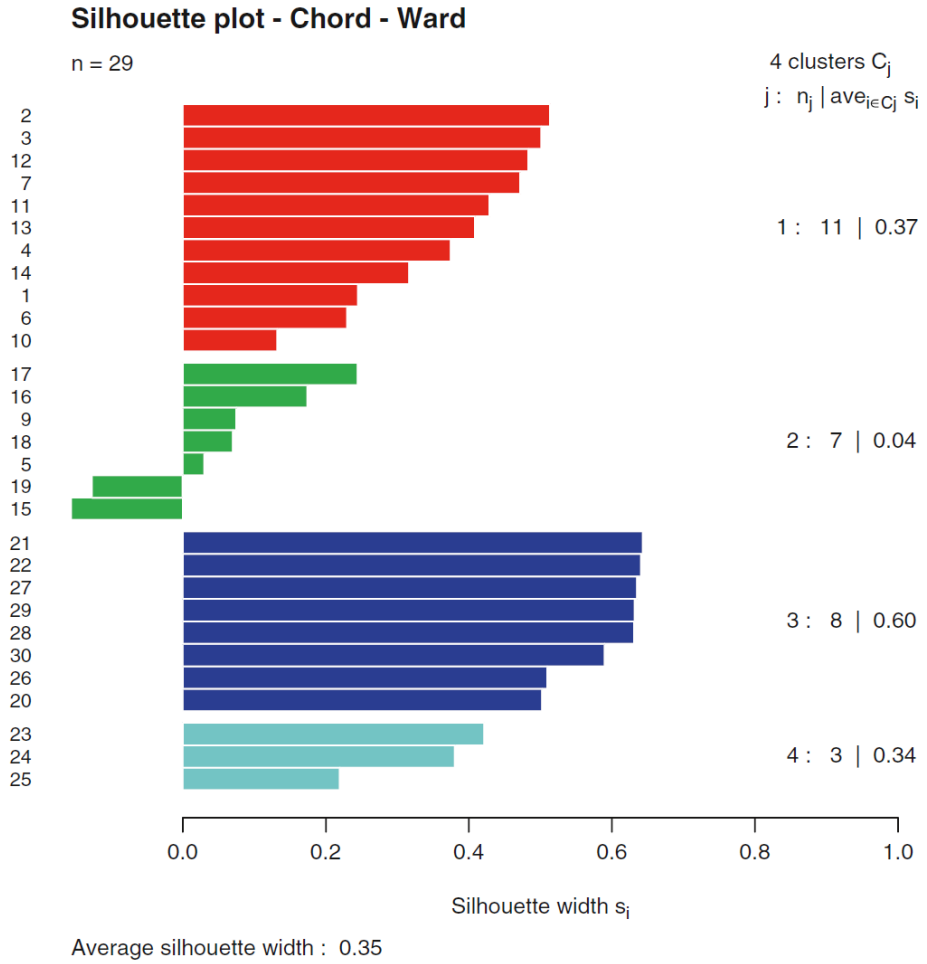


Fig. 4.8 Bar plot showing the average silhouette widths for $k = 2–29$ groups. The best partition by this criterion is the one with the largest average silhouette width



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					

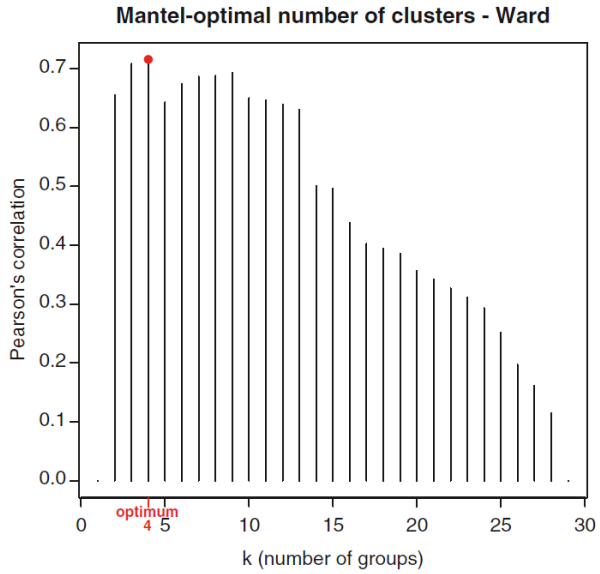


Fig. 4.9 Bar plot showing the correlations between the original distance matrix and binary matrices computed from the dendrogram cut at various levels

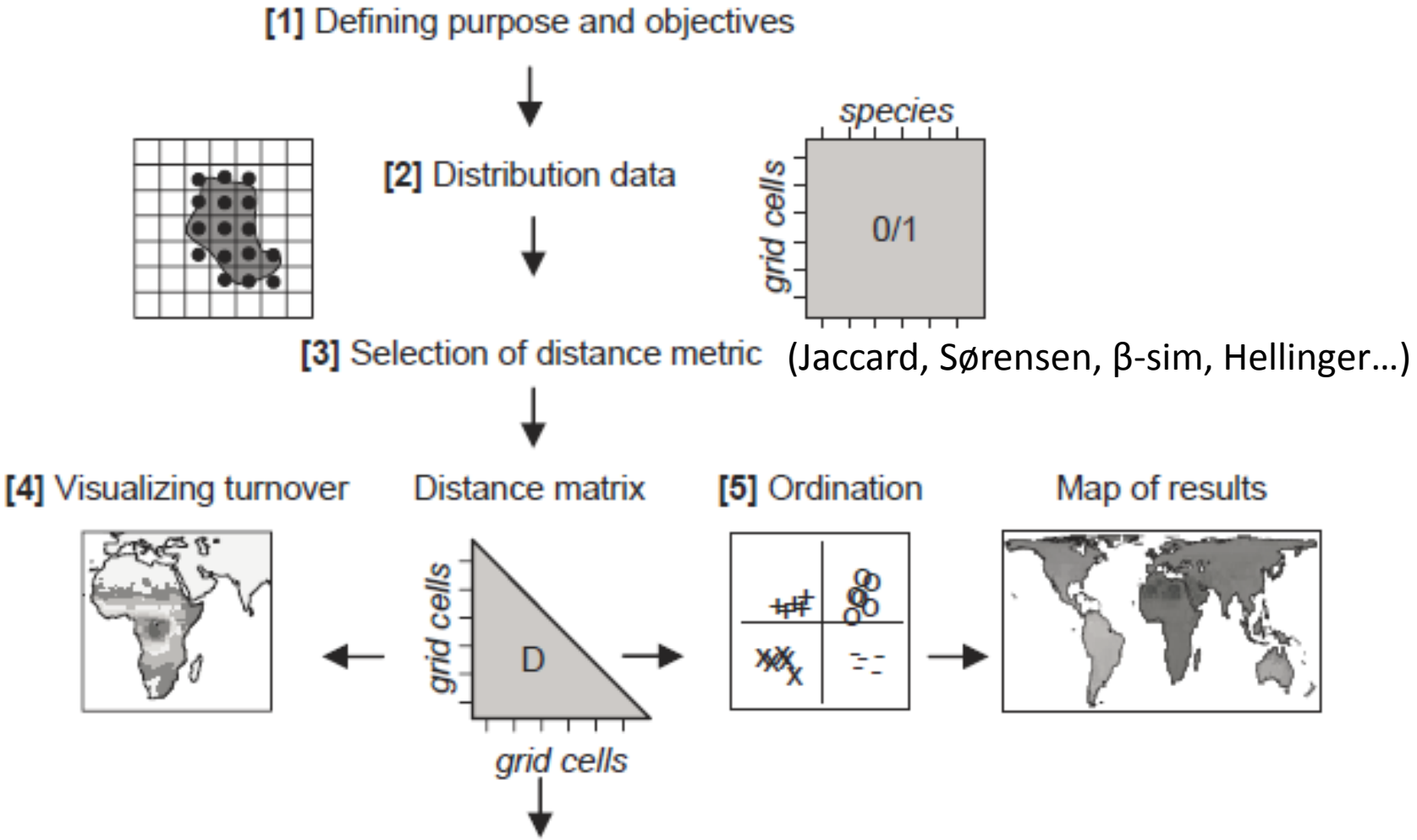
Borcard et al. (2011)

Journal of Biogeography (J. Biogeogr.) (2010)

SPECIAL PAPER

A framework for delineating biogeographical regions based on species distributions

Holger Kreft[1,2]* and Walter Jetz[2,3]

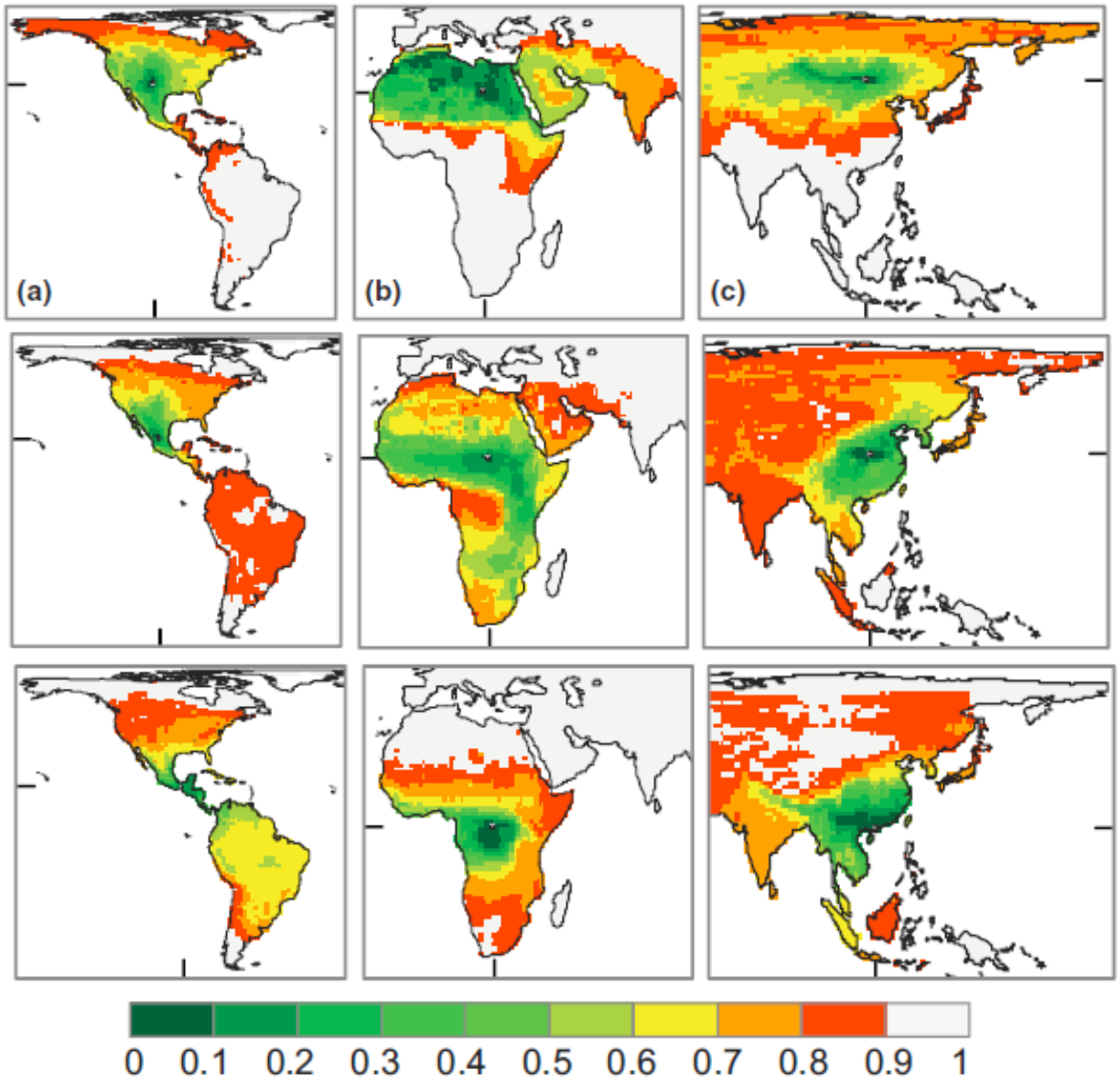


Journal of Biogeography (J. Biogeogr.) (2010)

SPECIAL PAPER

A framework for delineating biogeographical regions based on species distributions

Holger Kreft[1,2]* and Walter Jetz[2,3]

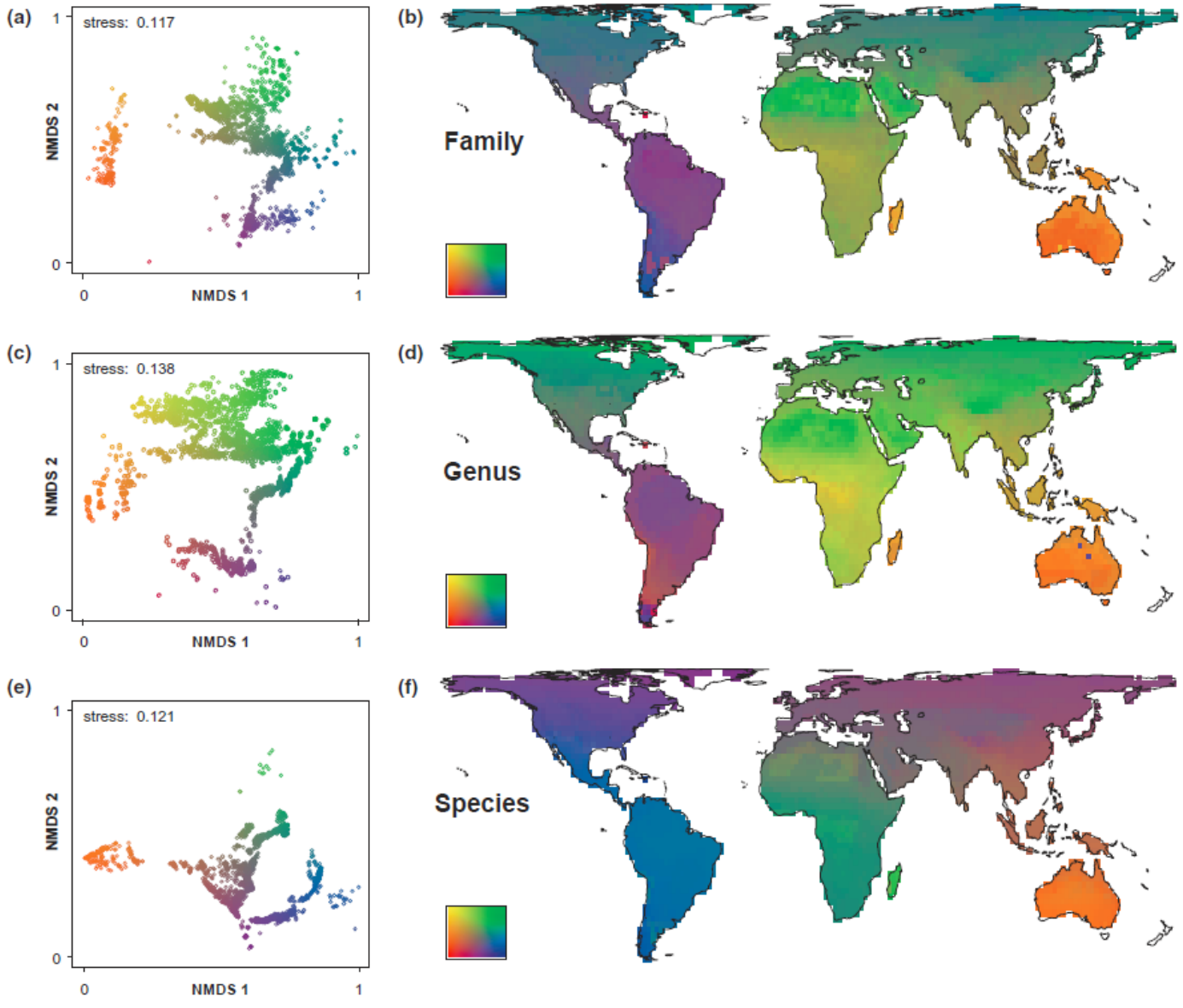


Journal of Biogeography (J. Biogeogr.) (2010)

SPECIAL PAPER

A framework for delineating biogeographical regions based on species distributions

Holger Kreft[1,2]* and Walter Jetz[2,3]

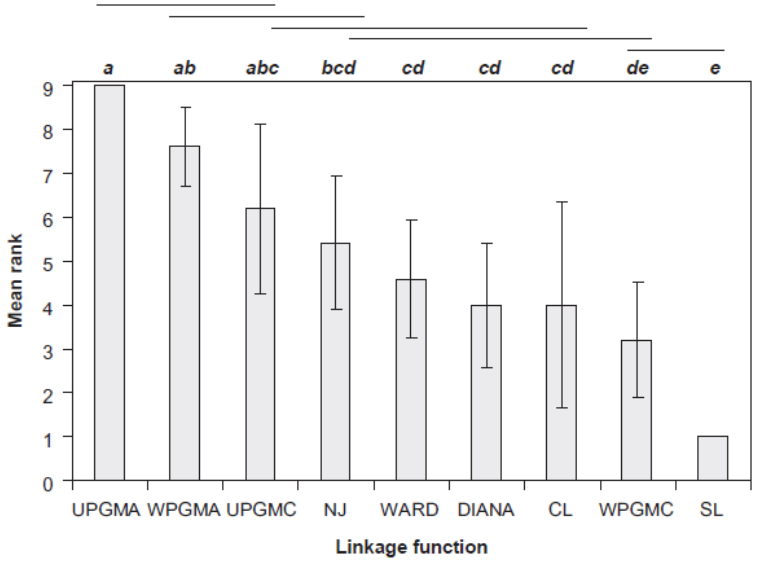


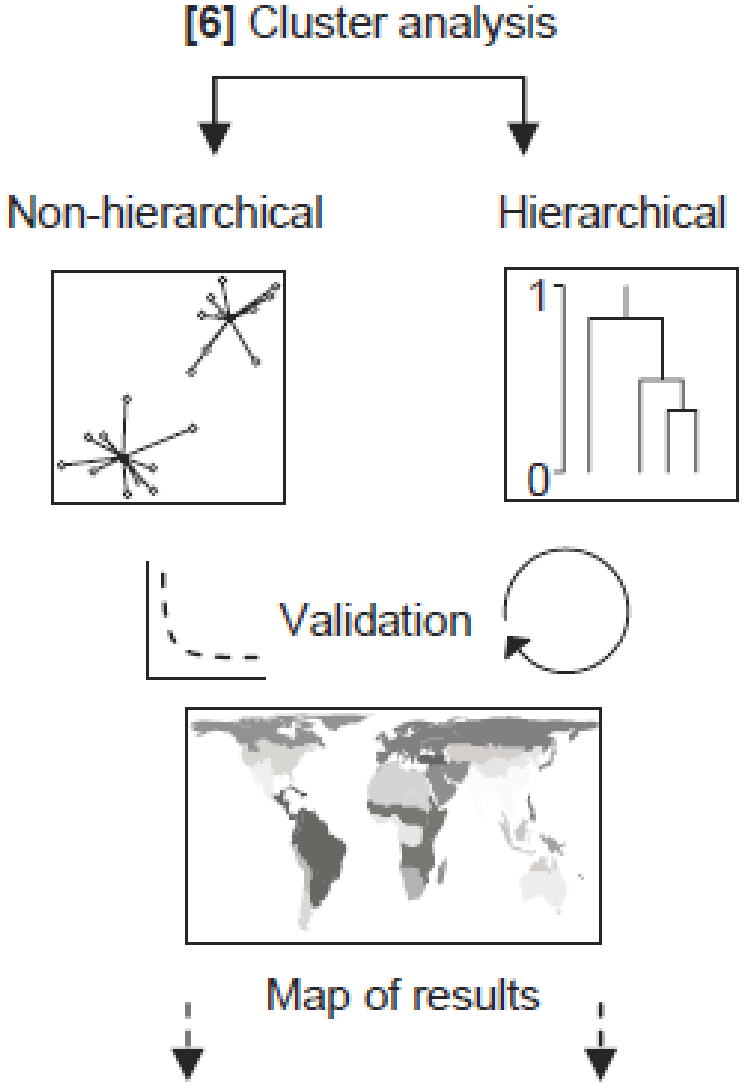
Journal of Biogeography (J. Biogeogr.) (2010)

SPECIAL PAPER

A framework for delineating biogeographical regions based on species distributions

Holger Kreft[1,2]* and Walter Jetz[2,3]



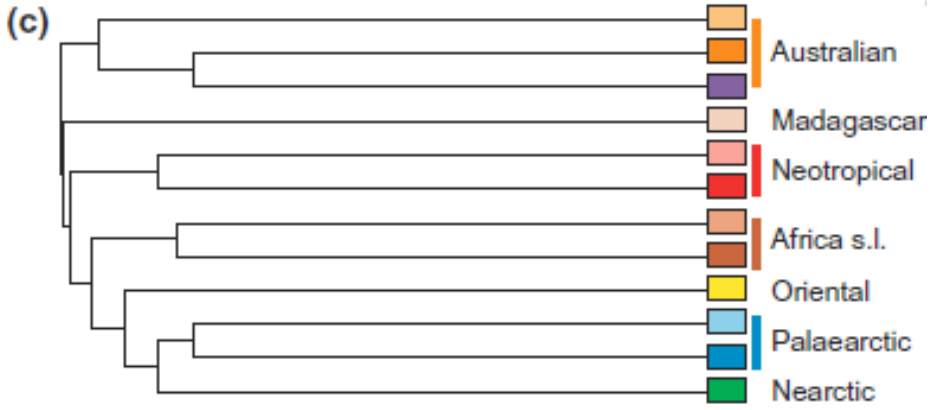


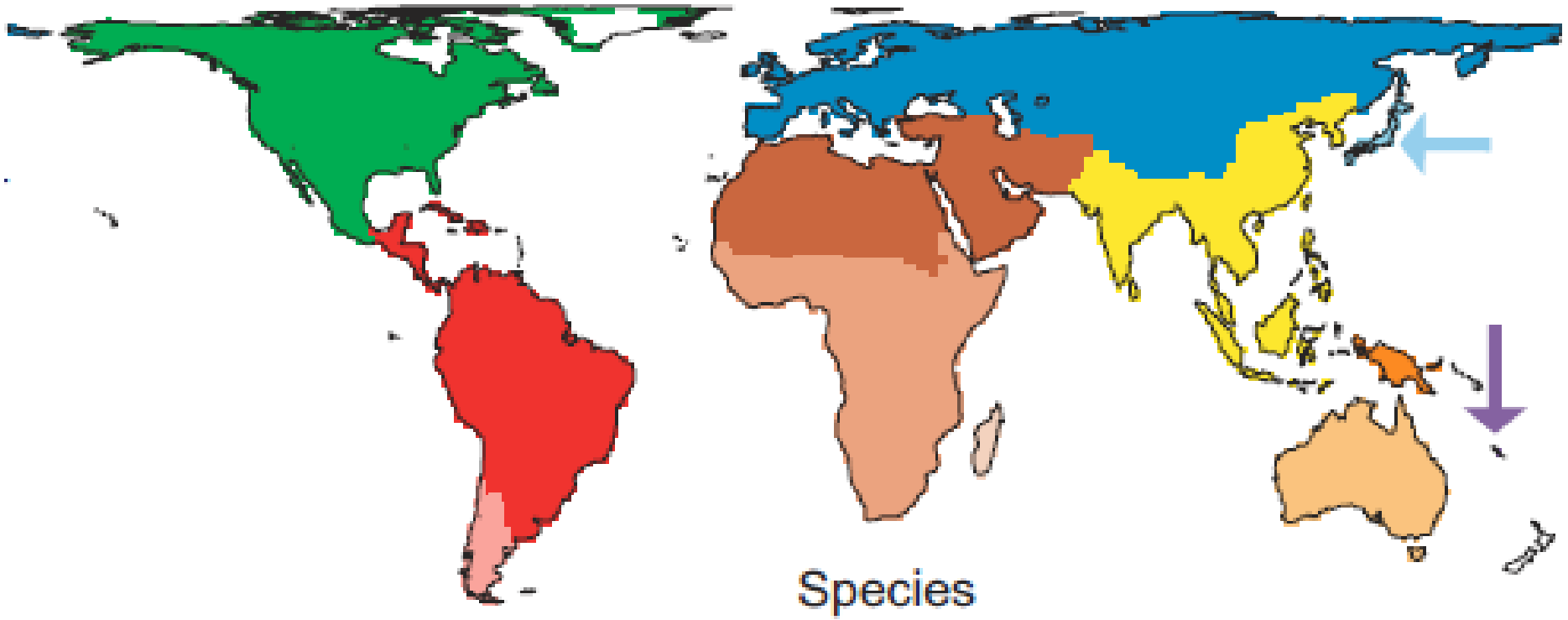
Journal of Biogeography (J. Biogeogr.) (2010)

SPECIAL PAPER

A framework for delineating biogeographical regions based on species distributions

Holger Kreft[1,2]* and Walter Jetz[2,3]



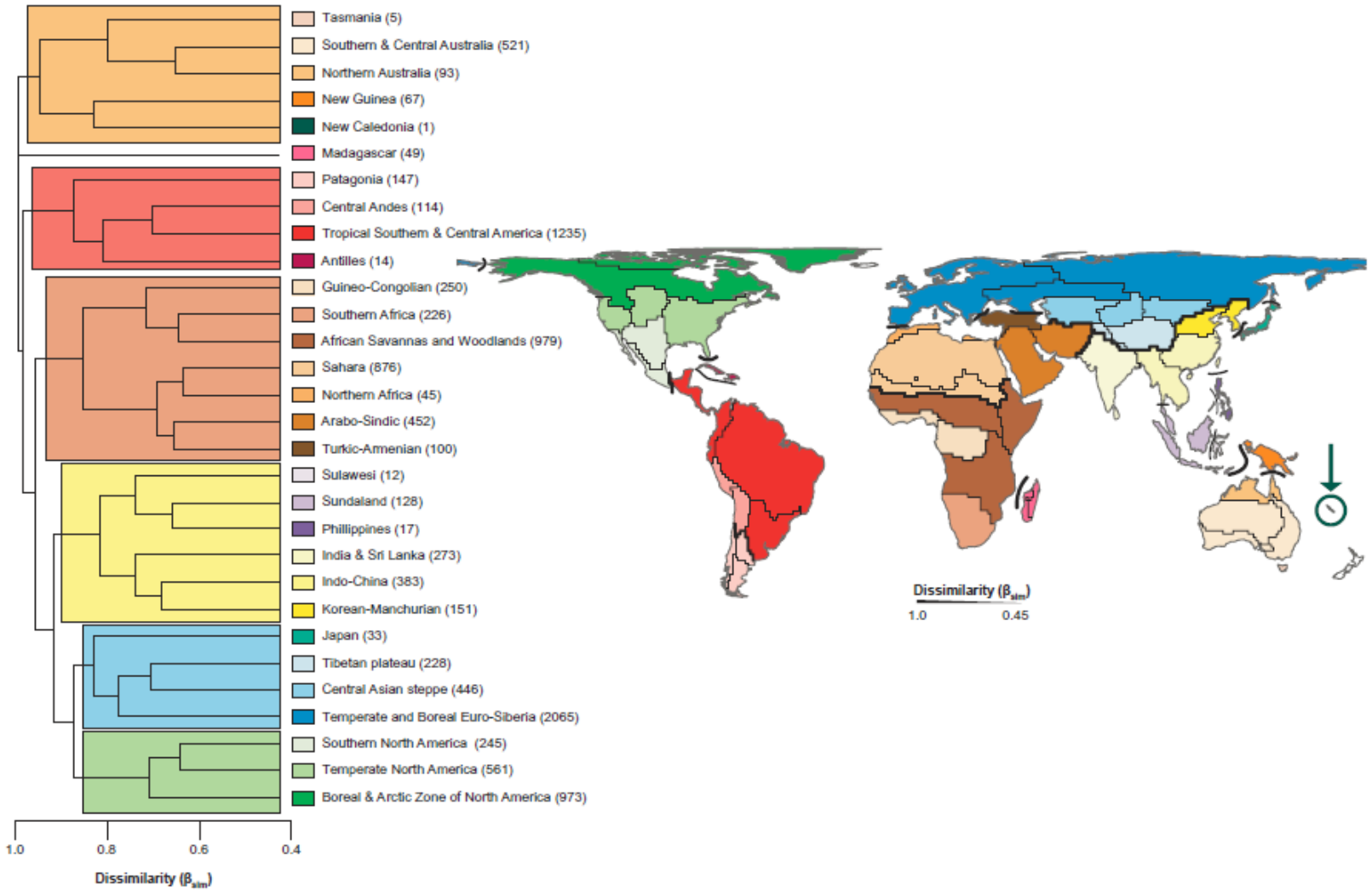


Journal of Biogeography (J. Biogeogr.) (2010)

SPECIAL PAPER

A framework for delineating biogeographical regions based on species distributions

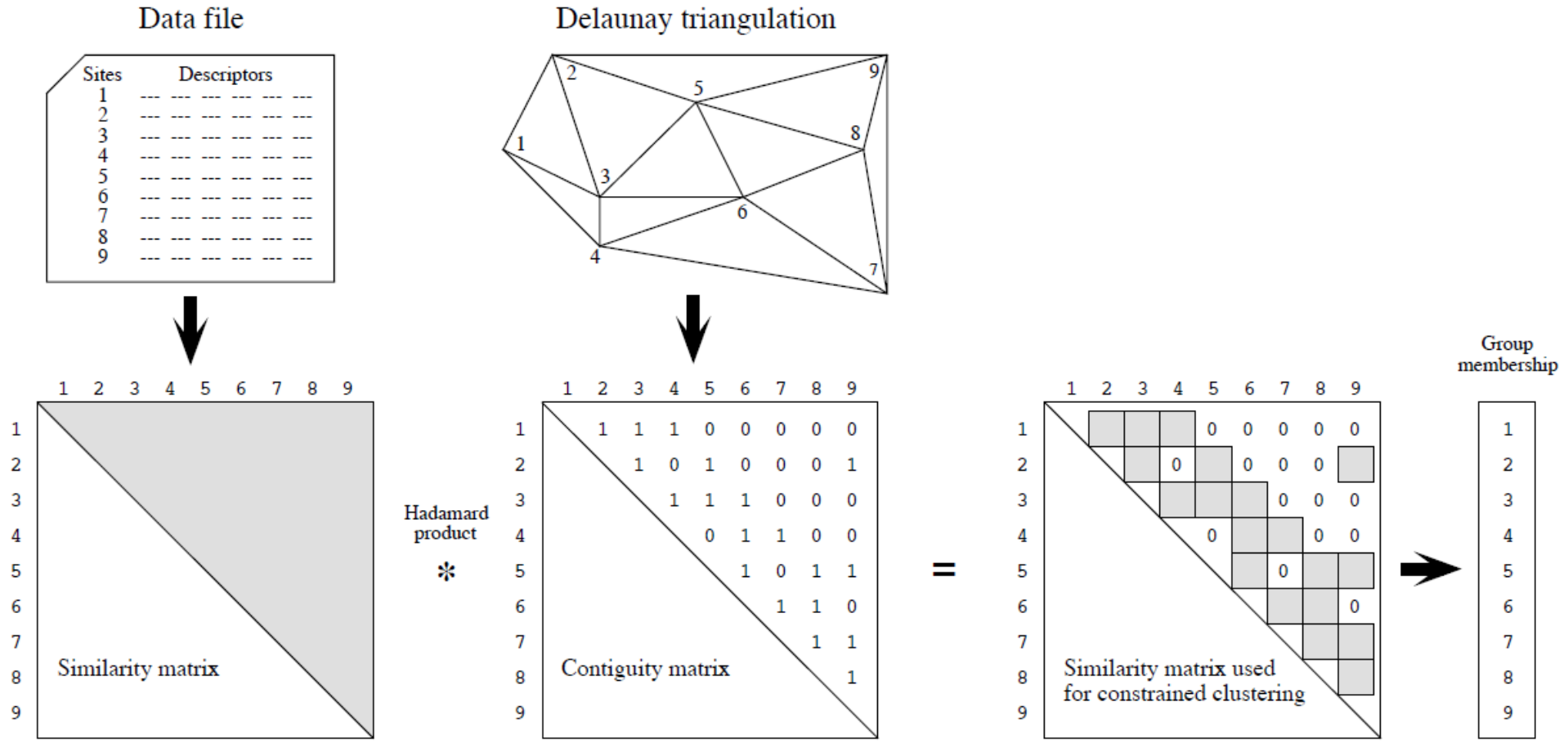
Holger Kreft[1,2]* and Walter Jetz[2,3]



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

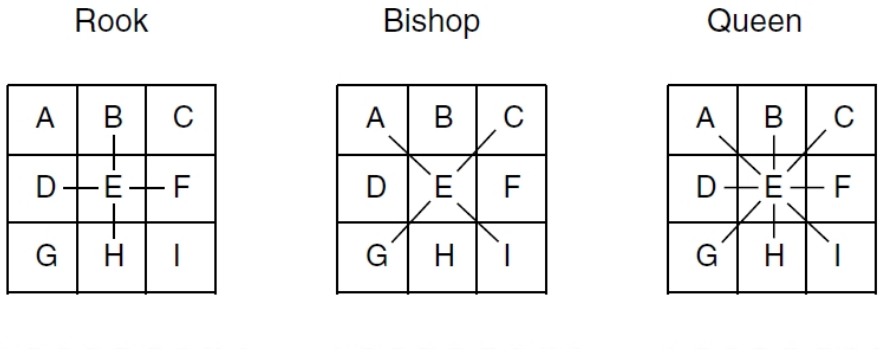


Legendre & Legendre (2012)

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)



	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	-

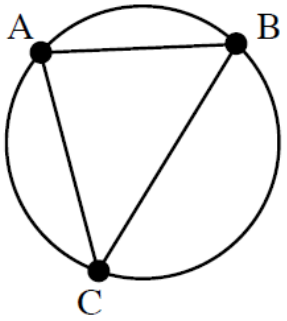
	A	B	C	D	E	F	G	H	I
A	-								
B	0	-							
C	0	0	-						
D	0	0	0	-					
E	1	0	1	0	-				
F	0	0	0	0	0	-			
G	0	0	0	0	1	0	-		
H	0	0	0	0	0	0	0	-	
I	0	0	0	0	1	0	0	0	-

	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	-

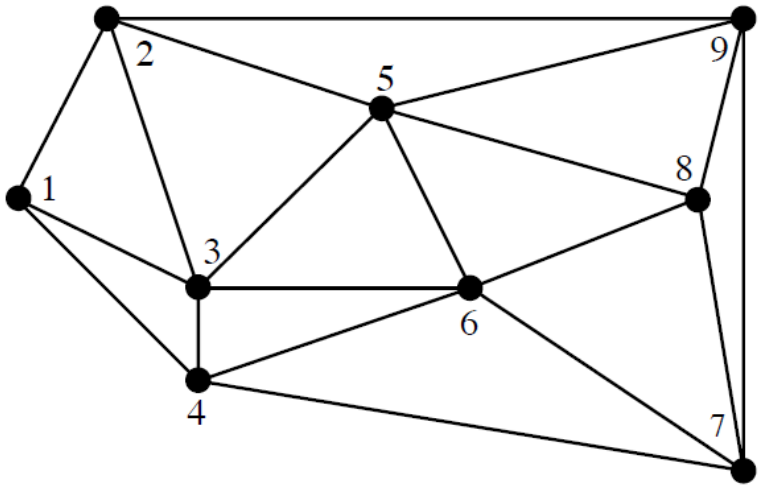
Fortin & Dale (2005)

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



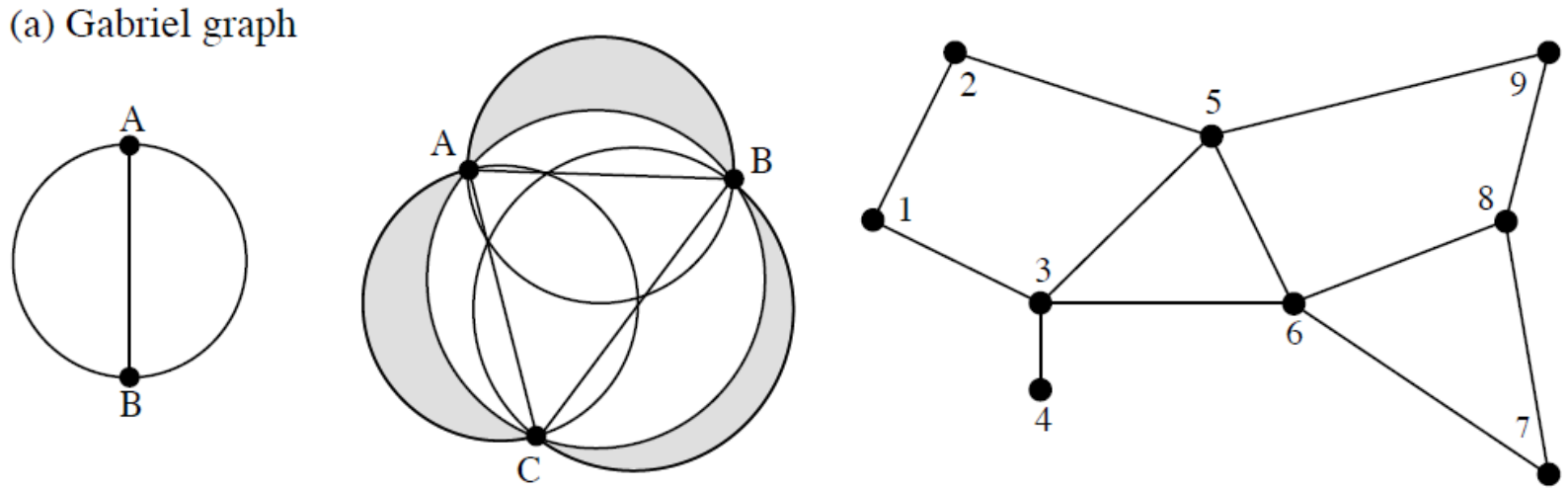
19 edges form the Delaunay triangulation:

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		

Legendre & Legendre (2012)

Gabriel criterion

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



12 edges form the Gabriel graph:

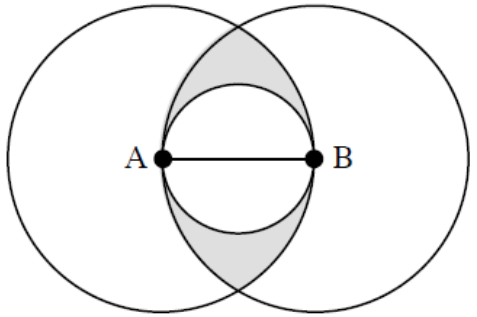
1–2	1–3	2–5	3–4	3–5	3–6
5–6	5–9	6–7	6–8	7–8	8–9

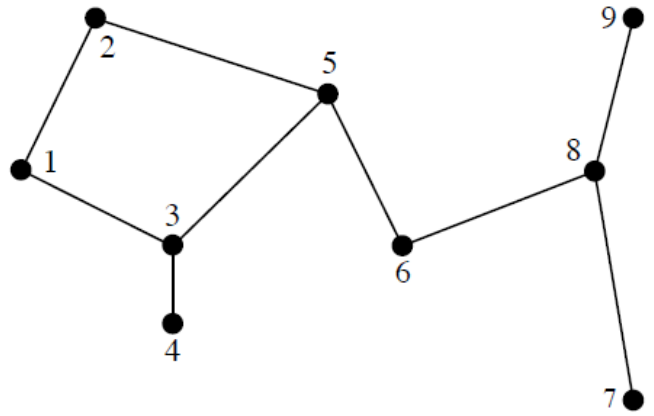
Legendre & Legendre (2012)

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

Legendre & Legendre (2012)

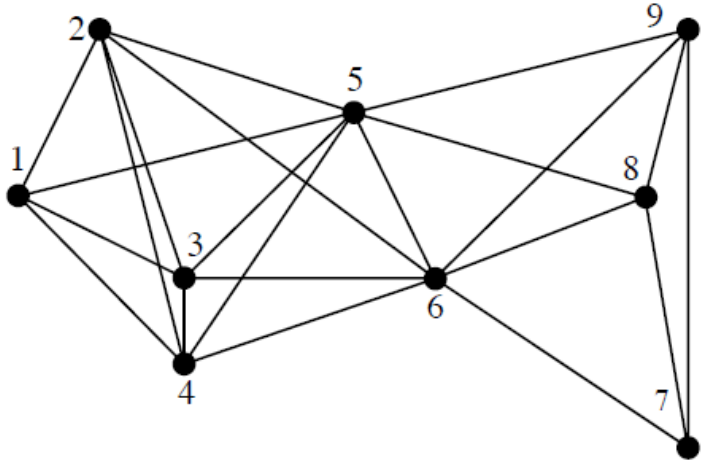
Maximum distance

- Vzorky jsou propojeny na základě kritéria vzdálenosti

(c) Maximum distance graph

Criterion: $D \leq$ threshold

In this example, $D \leq 5$



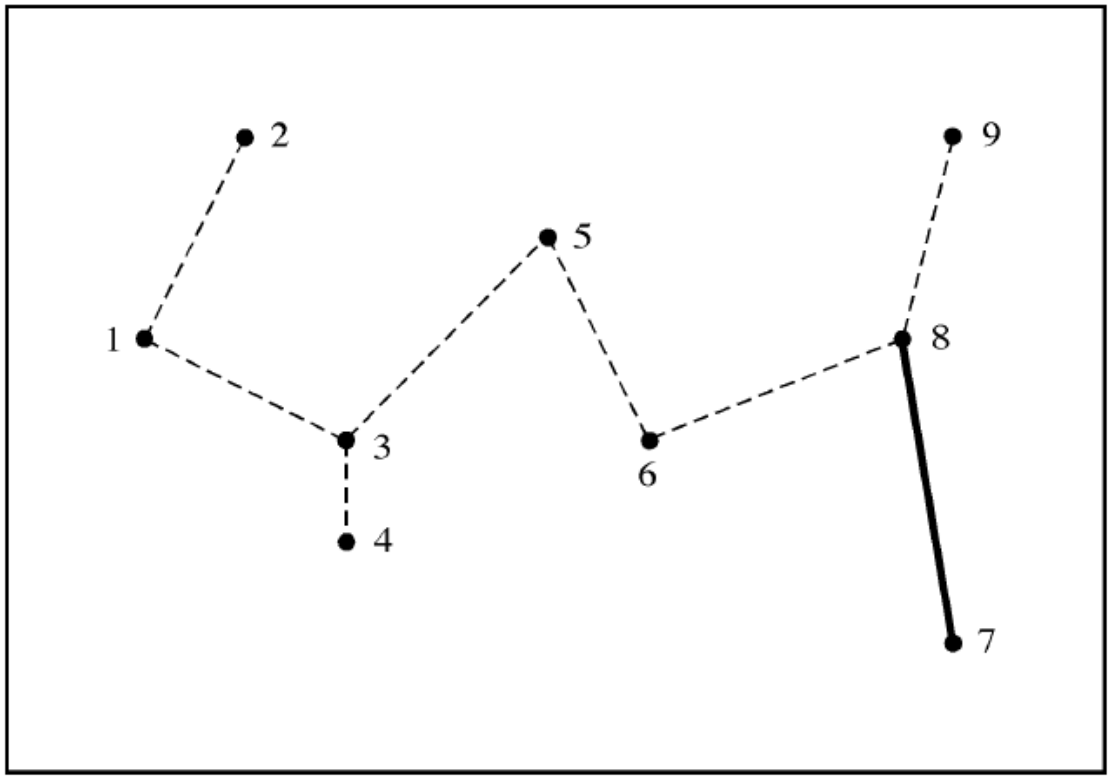
22 edges form the maximum distance graph:

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		

Legendre & Legendre (2012)

Minimum spanning tree (MST)

- Spojuje *n* vzorků *n*-1 spojnicemi tak, aby součet jejich délek byl minimální



Legendre & Legendre (2012)

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[1,2,3], Milan Chytrý[3], Vít Grulich[3] & Lucie Poláková[4]

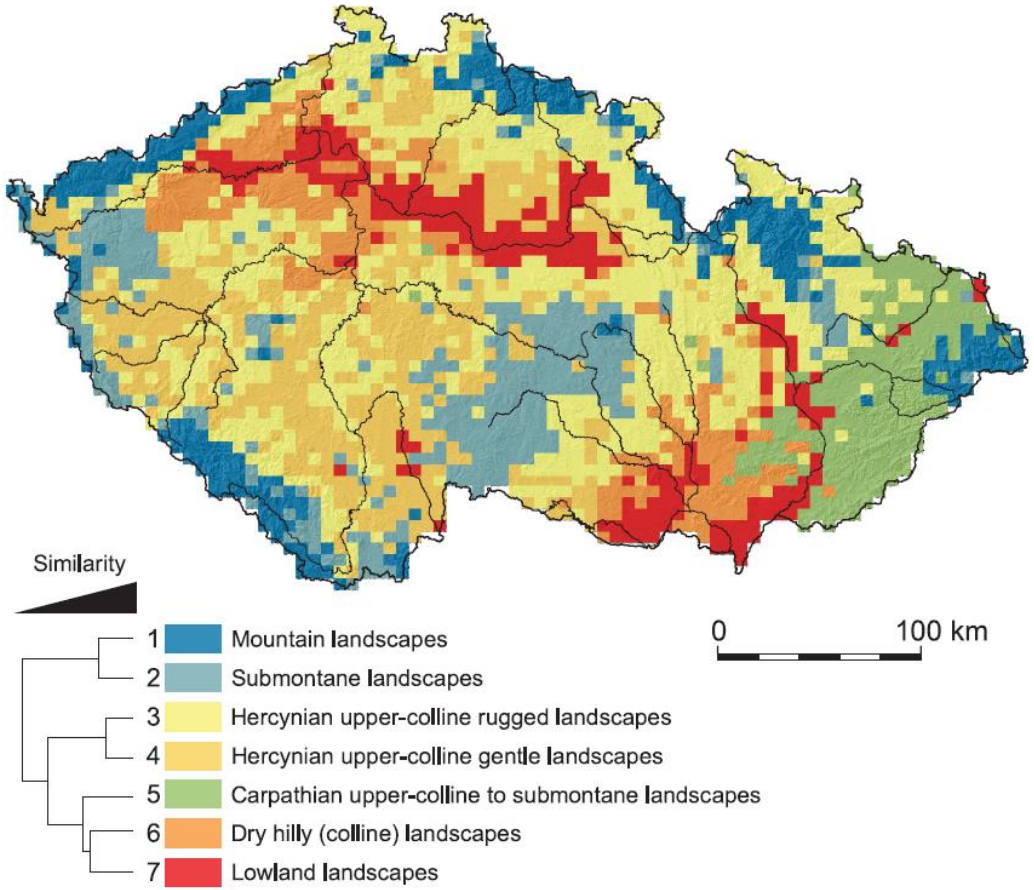


Fig. 3. – Landscape classification of the Czech Republic based on spatially unconstrained clustering with the optimal number of seven clusters according to the cross-validation procedure.

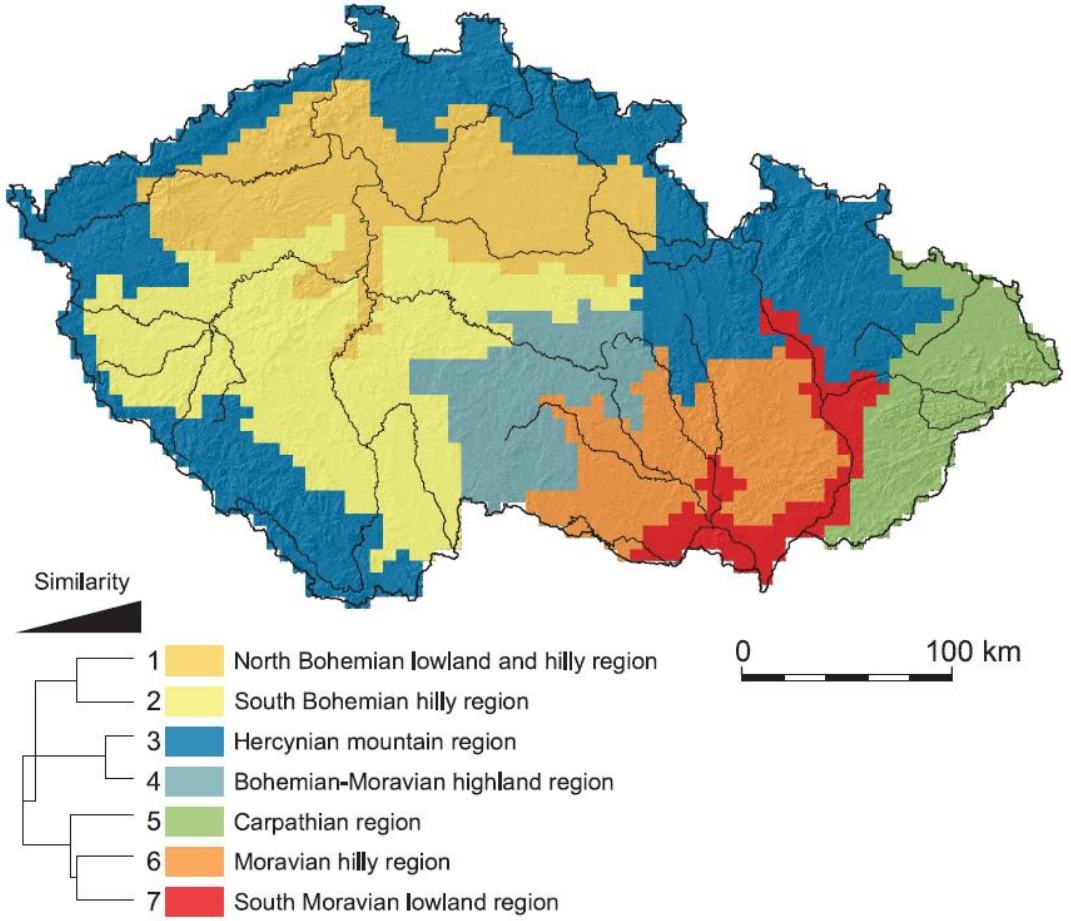


Fig. 7. – Regions of the Czech Republic based on the spatially constrained clustering with the optimal number of seven clusters according to cross-validation procedure. Reversal in the dendrogram is due to spatial constraints.

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.