Zhluková analýza

Danka Haruštiaková

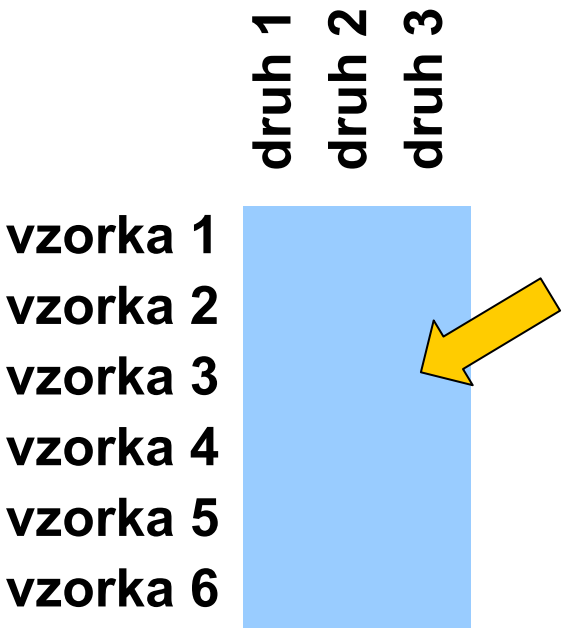
Podzim 2009

MU IBA Inštitút bioštatistiky a analýz, Masarykova univerzita

Úvod

- **Mnohorozmerné metódy**:
  názov „mnohorozmerné“ – dáta sú tvorené objektami (vzorky, lokality), každý z nich je charakterizovaný viacerými parametrami (druhmi)
  každý z týchto parametrov môžme považovať za jeden rozmer objektu (vzorky)

DÁTOVA MATICA



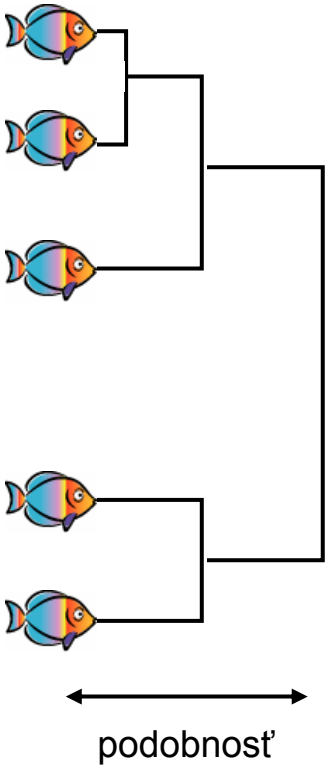
Hodnoty pre druhy (presencia/absencia; abundancia; dominancia) pre každú vzorku

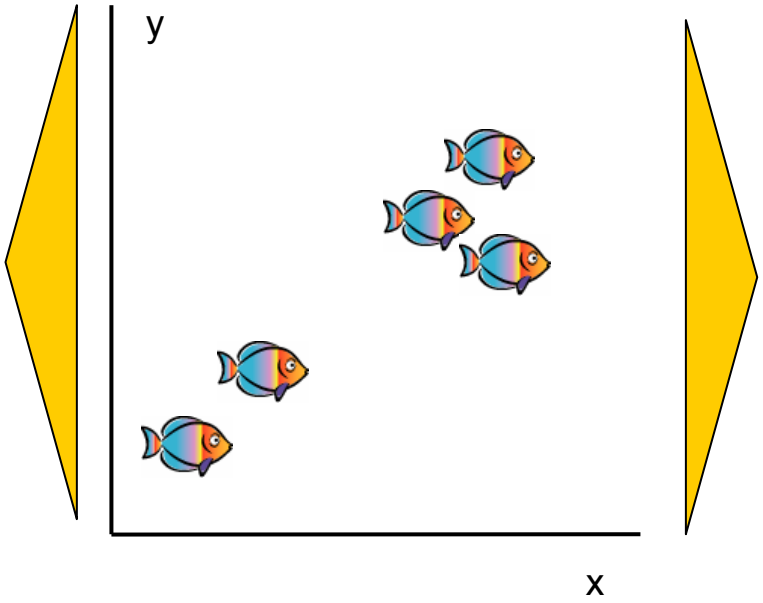
Ordinácia a zhluková analýza sú jediné možné techniky, ktoré môžeme použiť bez nameraných environmentálnych dát.

Úvod

ZHLUKOVÁ ANALÝZA

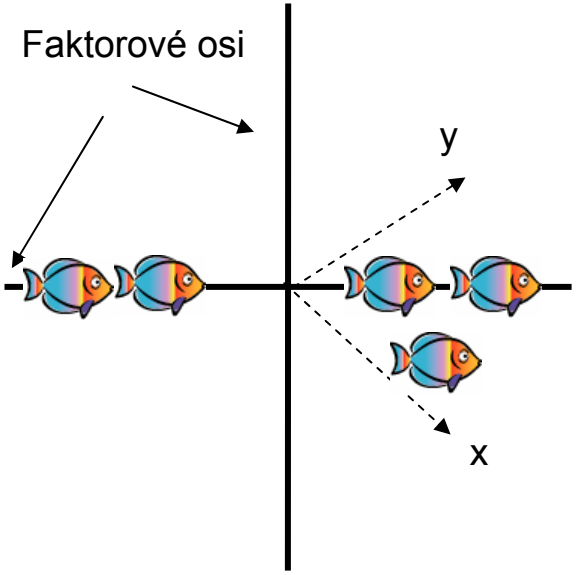
- Klasifikuje vzorky (lokality), druhy alebo premenné
- Nachádza skupiny v dátach





ORDINÁCIA

- Usporadúva vzorky pozdĺž trendu v dátach

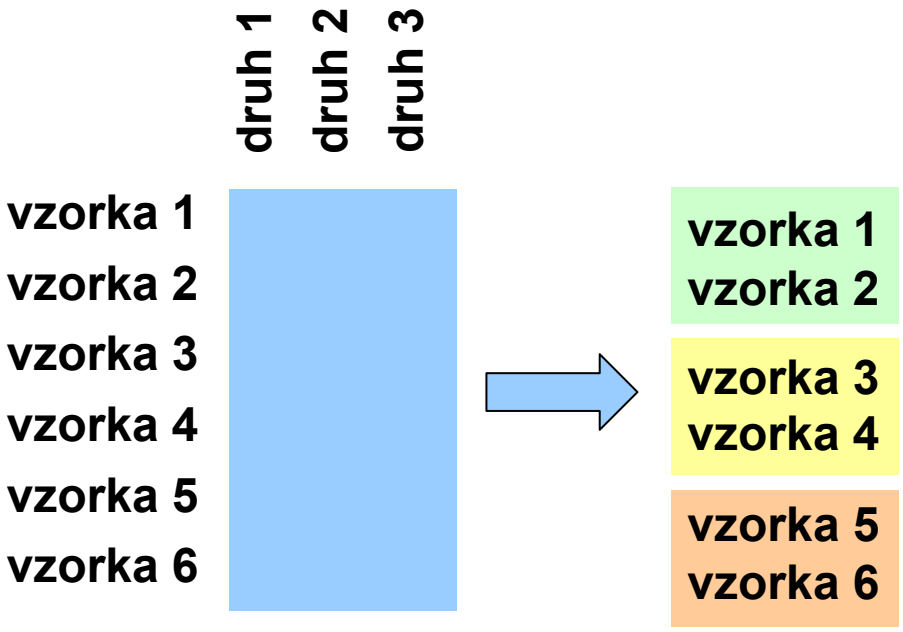


Zhluková analýza

Zhluková analýza:

- Roztriedenie objektov do niekoľkých pomerne homogénnych zhlukov
- Zníženie počtu dimenzií objektov tak, že radu uvažovaných premenných (druhy) zastúpi jediná premenná, vyjadrujúca príslušnosť objektu k definovanej skupine

Na základe druhov (premenných) → **Klasifikácia objektov do skupín**



- zhluky sú disjunktné
- objekty vnútri zhluku si sú čo najviac podobné a s objektami z rôznych zhlukov čo najmenej

Zhluková analýza

Ciele klasifikácie sú hlavne:

- poskytnúť informáciu o konkurencii druhov (vnútorná štruktúra dát),
- stanoviť typy spoločenstiev pre deskriptívne štúdie (syntaxonómia alebo mapovanie),
- odhaliť vzťahy medzi spoločenstvami a prostredím analyzovaním skupín vytvorených zhlukovou analýzou s ohľadom na environmentálne premenné (externá analýza).

Vstupné dáta

- Tabuľka spojitých alebo kategoriálnych dat popisujúca objekty

	druh 1	druh 2	druh 3
vzorka 1			
vzorka 2			
vzorka 3			
vzorka 4			
vzorka 5			
vzorka 6			

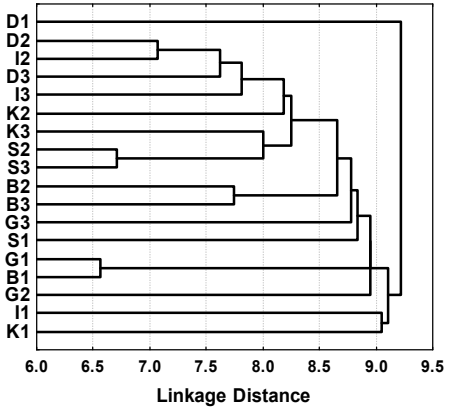
Výstupy analýzy

- Tzv. dendrogram popisujíci väzby medzi objektami alebo parametrami
- Rozdelenie objektov alebo parametrov do daného počtu skupín

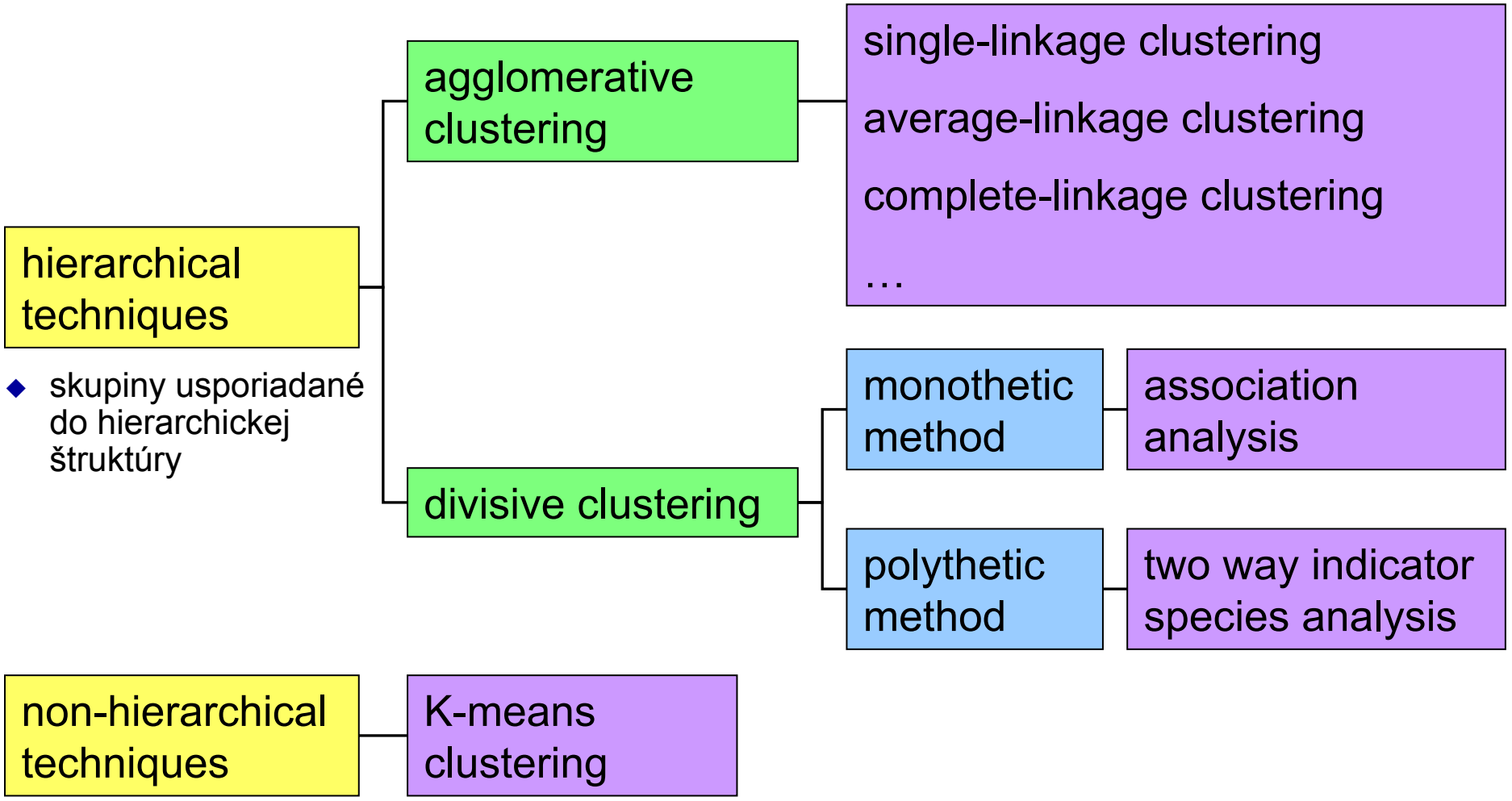
vzorka 1
vzorka 2

vzorka 3
vzorka 4

vzorka 5
vzorka 6



Zhluková analýza



- skupiny usporiadané do hierarchickej štruktúry

Hierarchické aglomeratívne zhlukovanie

Koeficienty podobnosti

Sorensenov koeficient

$$S_8(x_1, x_2) = \frac{2a}{2a+b+c} \qquad S_9(x_1, x_2) = \frac{3a}{3a+b+c}$$

Jaccardov koeficient

$$S_7(x_1, x_2) = \frac{a}{a+b+c}$$

Sorensenov kvantitatívny koef.

$$C_N = \frac{2\,jN}{(aN+bN)}$$

Morisota-Horn index

$$C_{mH} = \frac{2\sum(an_i bn_i)}{(da+db).aN.bN} \qquad da = \frac{\sum an_i^2}{aN^2}$$

Hierarchické aglomeratívne zhlukovanie

Metriky vzdialenosti

Euklidovská vzdialenosť

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{j1} - y_{j2})^2}$$

Vážená euklidovská vzdialenosť

$$D_9(x_1, x_2) = \sqrt{\sum_{j=1}^{p} w_j^2 (y_{j1} - y_{j2})^2}$$

Manhattanská vzdialenosť

$$D_7(x_1, x_2) = \sum_{j=1}^{p} |y_{j1} - y_{j2}|$$

Minkowski (power distance)

$$D_6(x_1, x_2) = \left[\sum_{j=1}^{p} |y_{j1} - y_{j2}|^{\lambda} \right]^{\frac{1}{\lambda}}$$

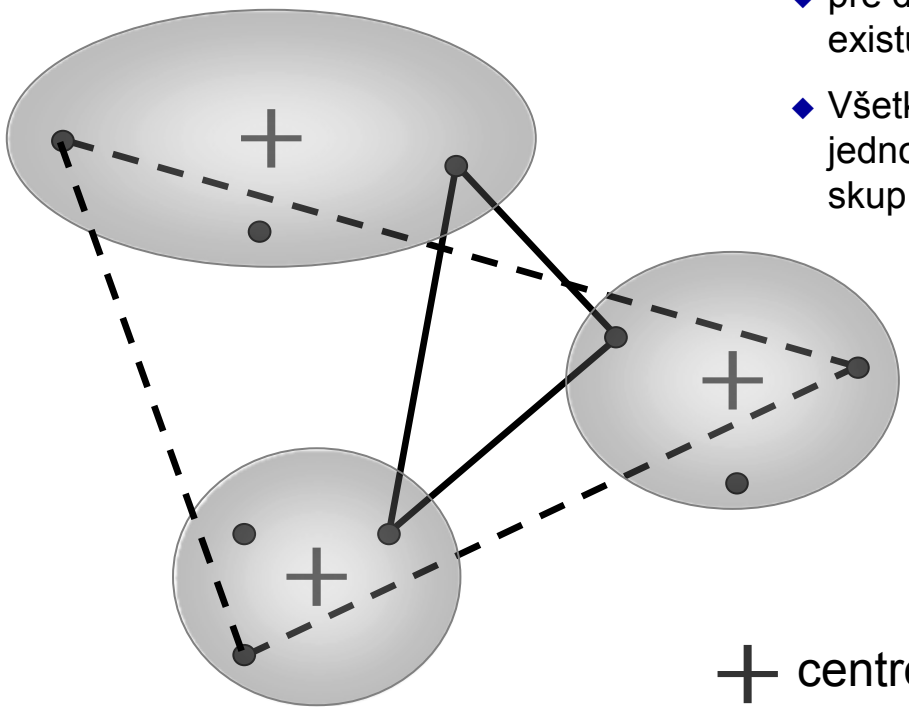
λ - celé číslo
λ =1 Manhattan (city block)
λ = 2 Euklidovská vzdialenosť

Výsledok zhlukovej analýzy je silne ovplyvnený výberom metriky vzdialenosti, resp. indexu podobnosti

Hierarchické aglomeratívne zhlukovanie

hierarchical techniques — agglomerative clustering

- začína jednotlivými objektami, ktoré sú spájané do väčších zhlukov
- vyžaduje maticu podobností alebo nepodobností (site by site), ktorou začína
- pre dáta presencie/absencie aj pre kvantitatívne dáta existuje mnoho indexov podobnosti
- Všetky aglomeratívne metódy sú založené na spájaní jednotlivých objektov (vzoriek) alebo zhlukov do väčších skupín



centroid

Definícia podobnosti medzi skupinami sa u jednotlivých metód líši. Metódy sa navzájom líšia chápaním vzdialenosti medzi zhlukmi.

vzdialenosť pri **single linkage**

vzdialenosť pri **complete linkage**

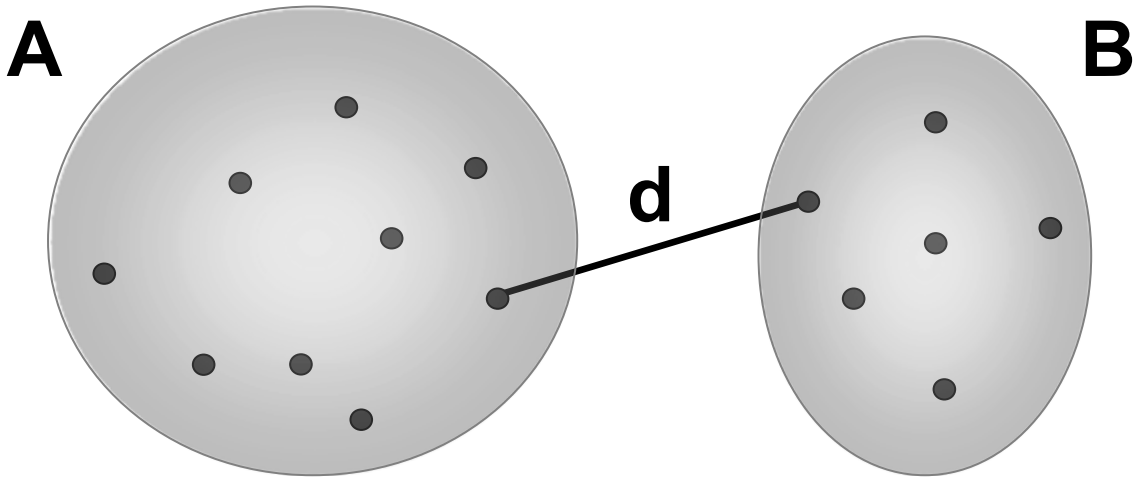
Iné metódy:

- vzdialenosť medzi centroidmi
- average linkage
- …

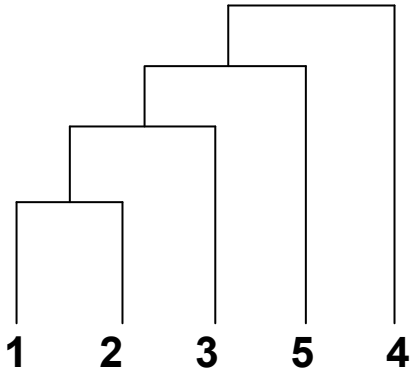
Hierarchické aglomeratívne zhlukovanie

Metóda najbližšieho suseda

(jednospojná metóda, metoda jedinej väzby, *single linkage, the nearest neighbor method*)



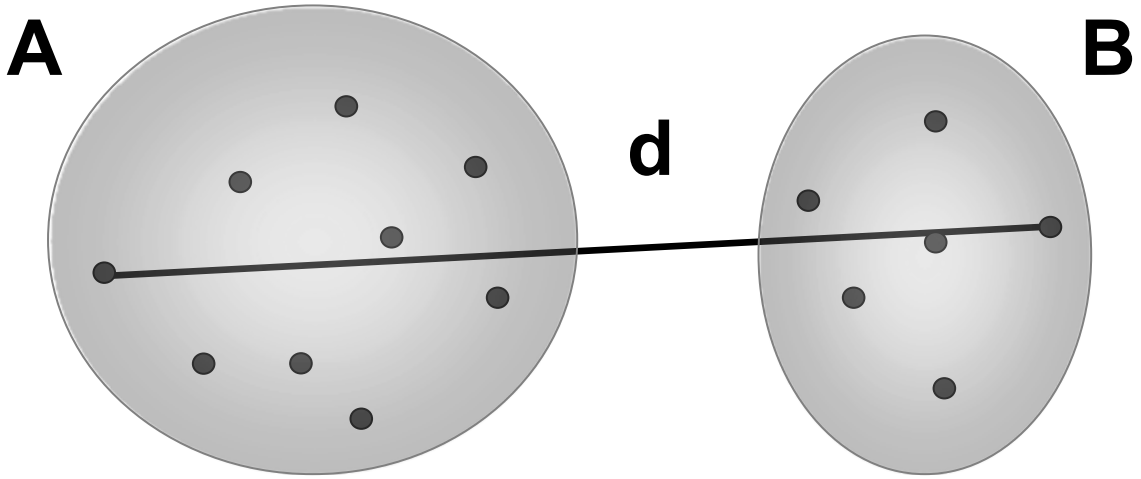
Vzdialenosť medzi dvoma zhlukmi je daná ako minimálna vzdialenosť medzi všetkými možnými zástupcami zhluku. Často sa i veľmi vzdialené objekty môžu zísť v rovnakom zhluku, ak väčší počet ďalších objektov medzi nimi vytvorí akýsi most.



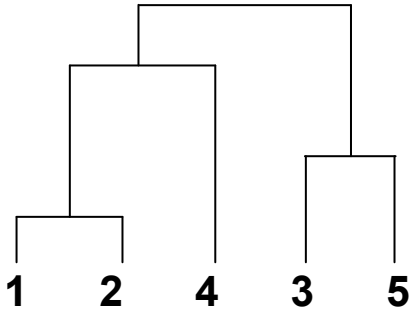
Hierarchické aglomeratívne zhlukovanie

Metóda najvzdialenejšieho suseda

(všespojná metóda, metóda úplnej väzby, *complete linkage, the furthest neighbor method*)

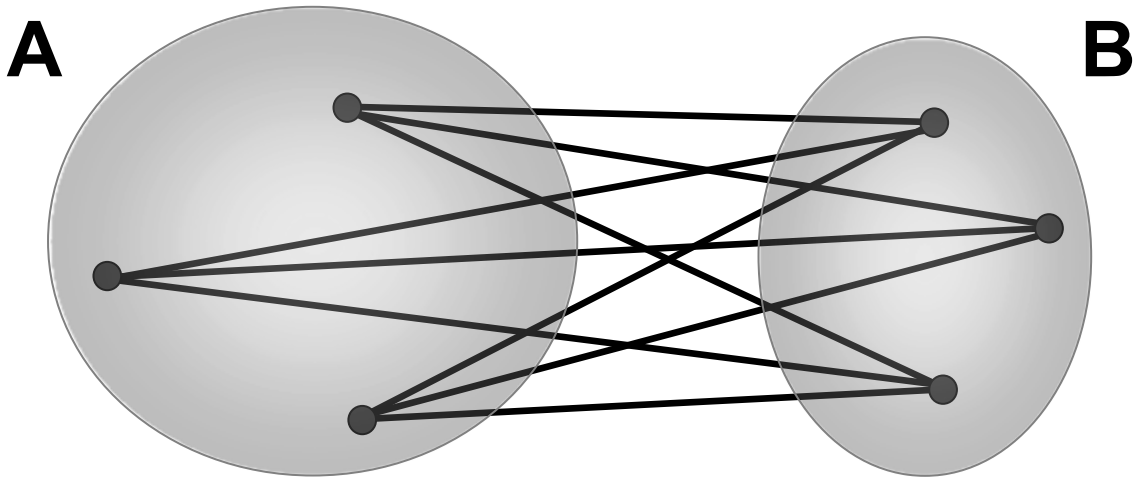


Vzdialenosť medzi dvoma zhlukmi je daná maximálnou vzdialenosťou medzi všetkými možnými zástupcami oboch zhlukov.
Zhluky sú medzi sebou dobre oddelené.
Tendencia k tvorbe kompaktých zhlukov, nie však veľmi veľkých.



Hierarchické aglomeratívne zhlukovanie

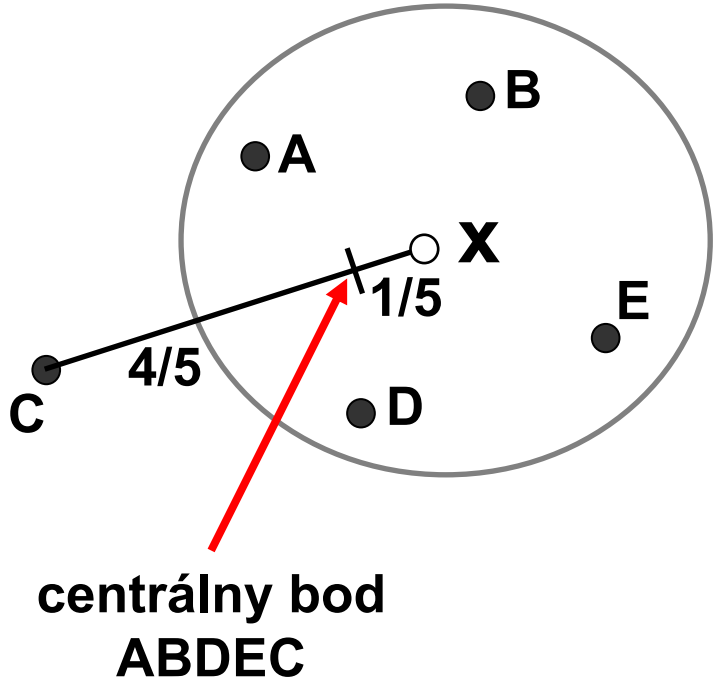
Metóda priemernej vzdialenosti (stredospojná metóda, metóda priemernej väzby, *average linkage*, *UPGMA – unweighted pair-group method using arithmetic averages*)

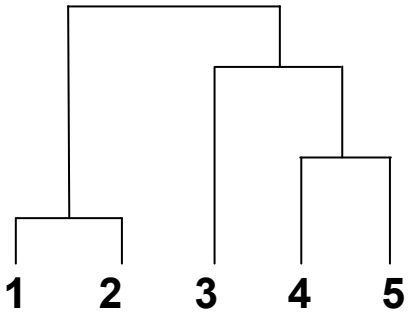


Medziskupinová (ne)podobnosť je definovaná ako priemerná (ne)podobnosť medzi všetkými možnými pármi členov.
Metóda vedie často k podobným výsledkom ako metóda najvzdialenejšieho suseda.

Hierarchické aglomeratívne zhlukovanie

Centroidová metóda (Gowerova metóda, *centroid method, UPGMC – unweighted pair-group method using centroids*)

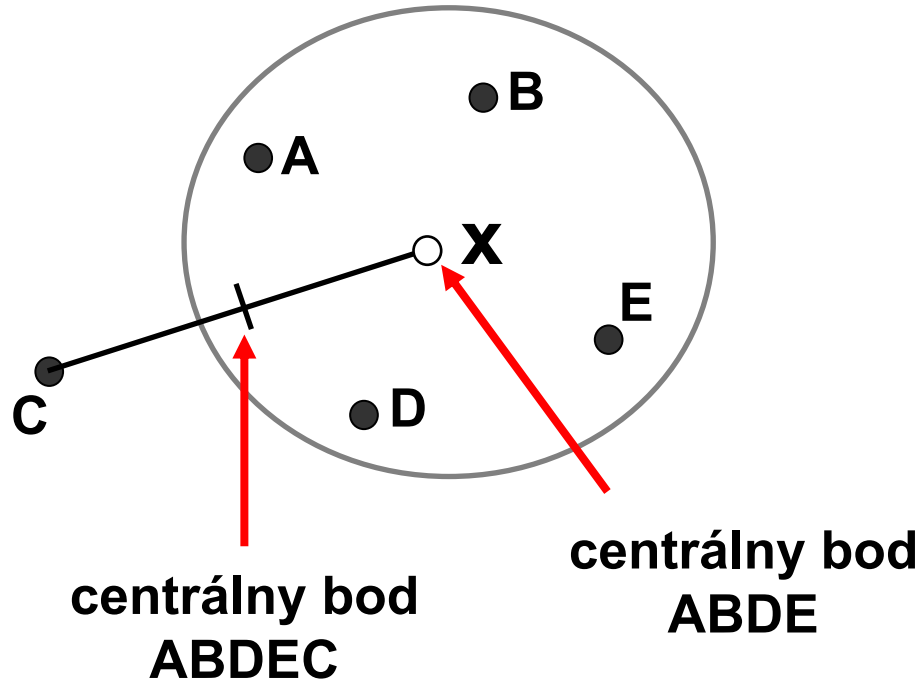




Táto metóda nevychádza už z agregácie informácií o medzizhlukových vzdialenostiach objektov. Kritérium je euklidovská vzdialenosť centroidov. Pri tejto metóde je vzdialenosť medzi zhlukmi počítaná ako vzdialenosť medzi centroidmi týchto zhlukov.

Hierarchické aglomeratívne zhlukovanie

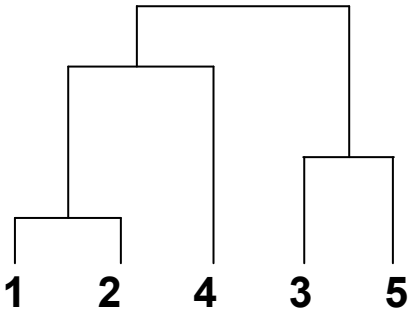
Mediánová metoda (*median method, WPGMC – weighted pair-group method using centroids, weighted centroid clustering*)



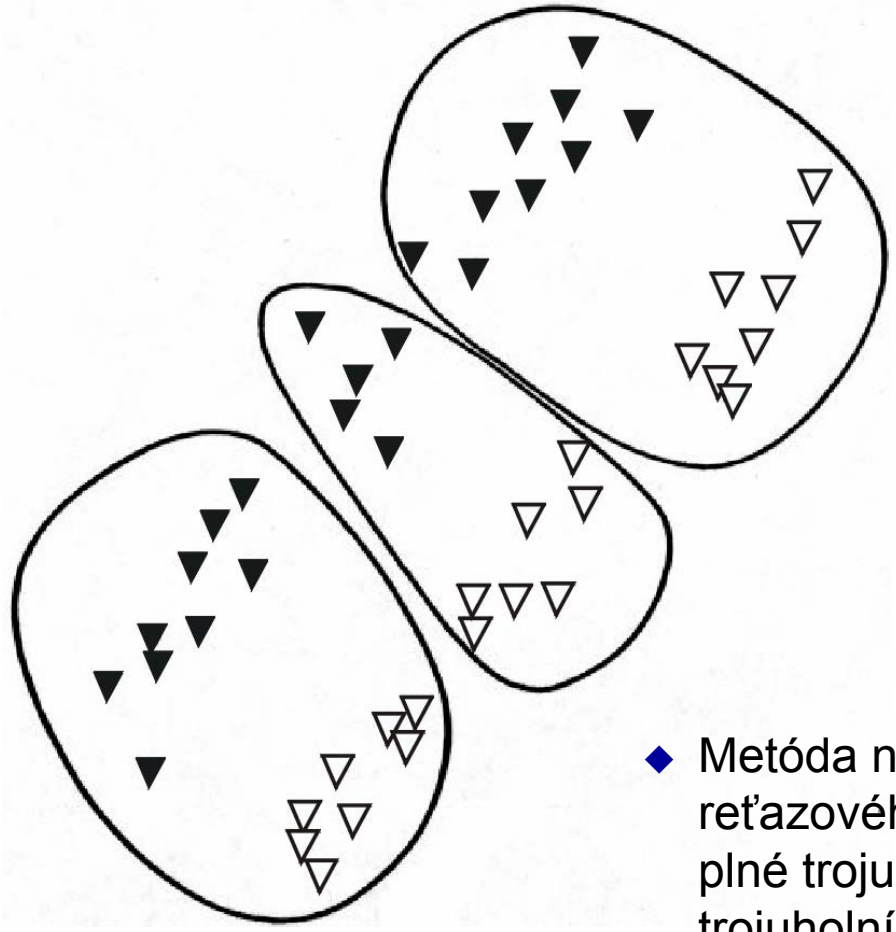
Hierarchické aglomeratívne zhlukovanie

Wardova metóda (*Minimum variance clustering*)

Wardova metóda je podobná stredospojnej a centroidnej metóde. Kritérium pre spojovanie zhlukov je prírastok celkového vnútroskupinového súčtu štvorcov odchýlok pozorovaní od zhlukového priemeru. Prírastok je vyjadrený ako súčet štvorcov v novo vznikajúcom zhluku, zmenšený o súčty štvorcov v oboch zanikajúcich zhlukoch. Wardova metóda má tendenciu odstraňovať malé zhluky, teda tvoriť zhluky zhruba zhodnej veľkosti.



Hierarchické aglomeratívne zhlukovanie



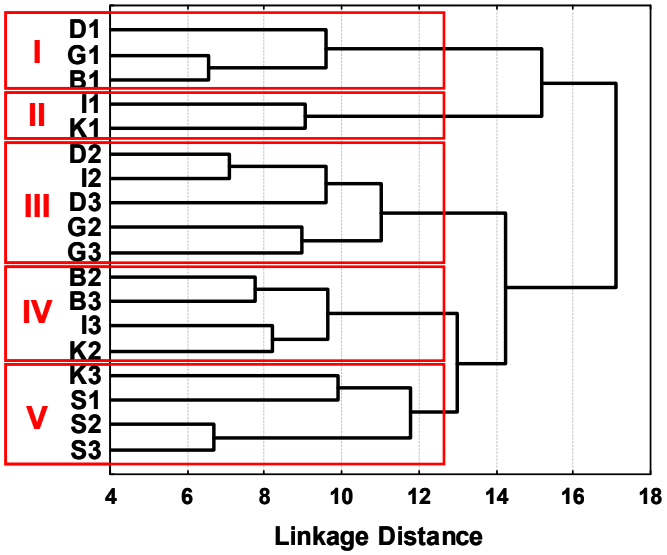
- Metóda najbližšieho suseda by v dôsledku reťazového efektu spojila do jedného zhluku plné trojuholníky a do druhého prázdne trojuholníky, zatiaľ čo Wardova metóda a metóda priemernej vzdialenosti by priniesli skupiny ohraničené čiarami (podľa Everitt & Dunn 1983).

Hierarchické aglomeratívne zhlukovanie

Výsledkom hierarchického aglomeratívneho zhlukovania je **dendrogram (strom).**

V tomto prípade boli použité:

- všespojná zhlukovacia metóda (complete linkage)
- miera vzdialenosti: Euklidovské vzdialenosti



Dendrogram znázorňuje podobnosť spoločenstiev **kôrovcov** šiestich lokalít v záplavovej oblasti **Dunaja v troch obdobiach**

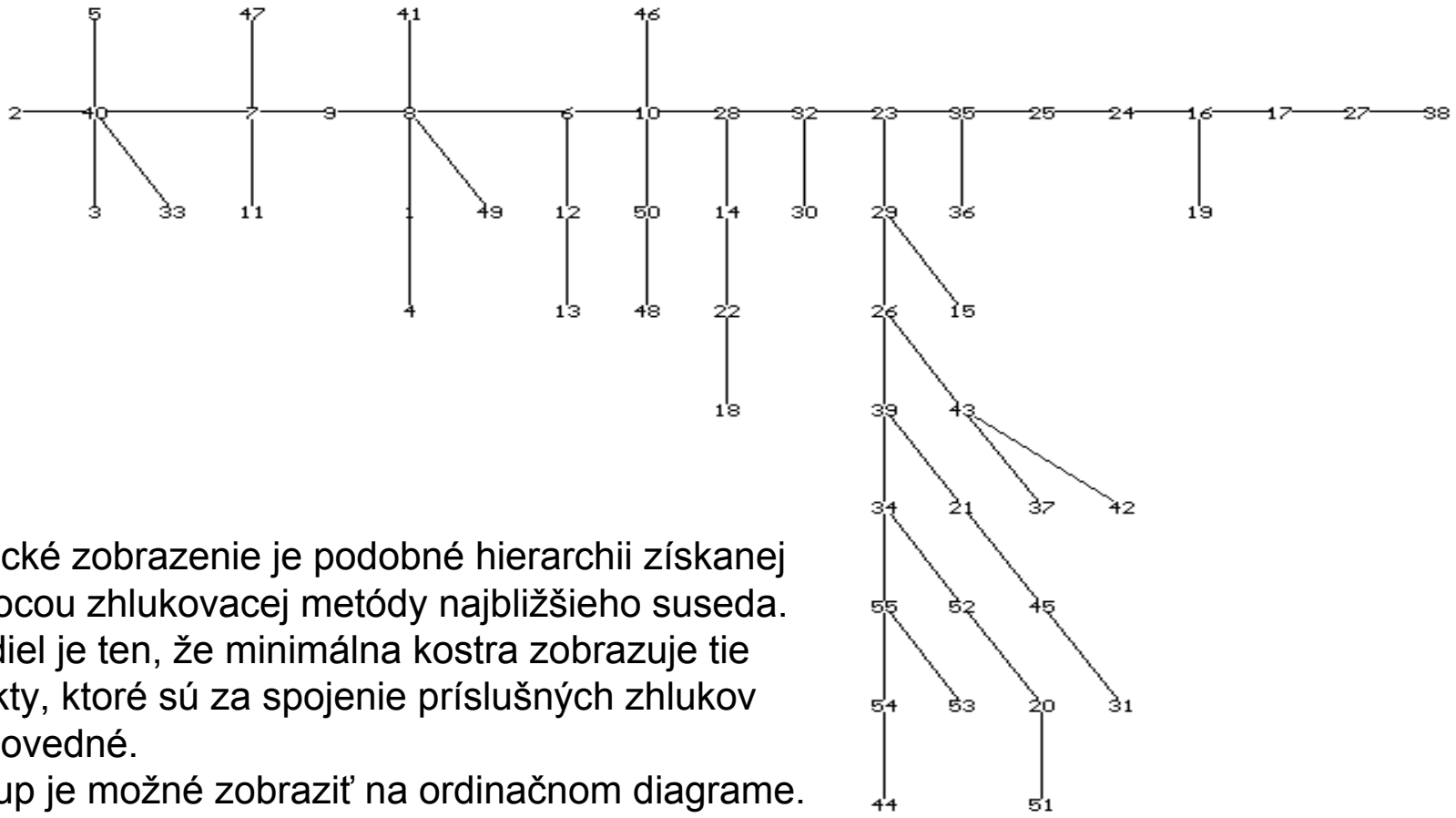
- 1: 1991-1992 pred prehradením Dunaja
- 2: 1993-1997 prvých 5 rokov po prehradení
- 3: 1999-2004 ďalších 6 rokov po prehradení

Sledované lokality:

- D: Dobrohošť
- G: Gabčíkovo
- B: Bodíky
- I: Istragov
- K: Kráľovská lúka
- S: Sporná sihoť

Hierarchické aglomeratívne zhlukovanie

Minimálna kostra (*minimum spanning tree*)
graf, ktorý spojuje všetky objekty tak, že sa tu nevyskytujú žiadne smyčky alebo kružnice a zároveň súčty dĺžok spojníc medzi uzlami (objektami) je minimálny.

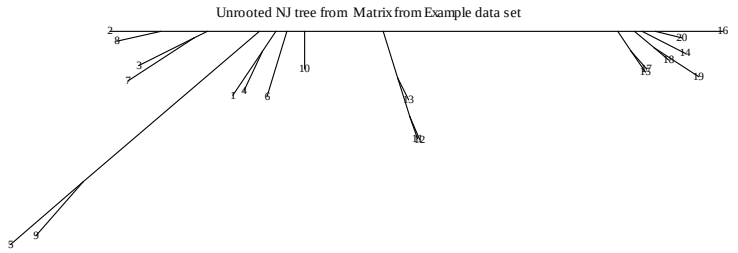


Grafické zobrazenie je podobné hierarchii získanej pomocou zhlukovacej metódy najbližšieho suseda. Rozdiel je ten, že minimálna kostra zobrazuje tie objekty, ktoré sú za spojenie príslušných zhlukov zodpovedné.
Výstup je možné zobraziť na ordinačnom diagrame.

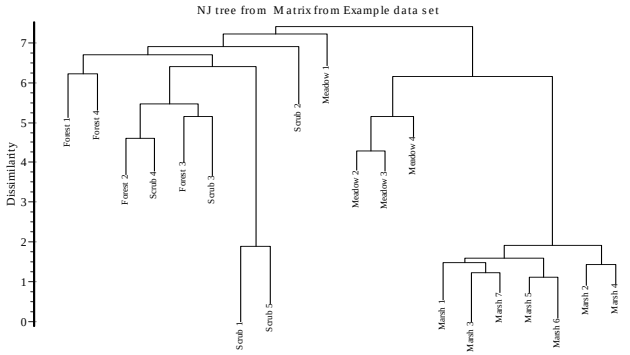
Hierarchické aglomeratívne zhlukovanie

Metóda spojovania susedných objektov (*neighbor-joining method*)
Metóda je podobná zhlukovacím metódam.
Používa sa napr. k hodnoteniu dát získaných pri analýze dĺžkového polymorfizmu DNA, tj. v situáciách, kedy sú výsledkom analýz matice binárnych dat (prítomnost alebo neprítomnosť prúžkov v odpovedajúcich pozíciách na elektroforetickom gély).
Je založená na genetickej vzdialenosti, ktorá závisí na počte zhodujúcich sa prúžkov v príslušných vzorkách.
Pri výpočte vzdialenosti vytvorených zhlukov od zostávajúcich objektov sa postupuje podobne ako pri metóde priemernej vzdialenosti. Ale „susedné objekty“ sa nespájajú tie, ktoré ležia najbližšie, ale tak, aby bol výsledkom čo najkratší strom (dendrogram).
Dendrogram sa skladá z uzlov (*node*) spojených medziuzlami (*internode*) a vetví (*branch*).

nezakorenený dendrogram (*unrooted*)

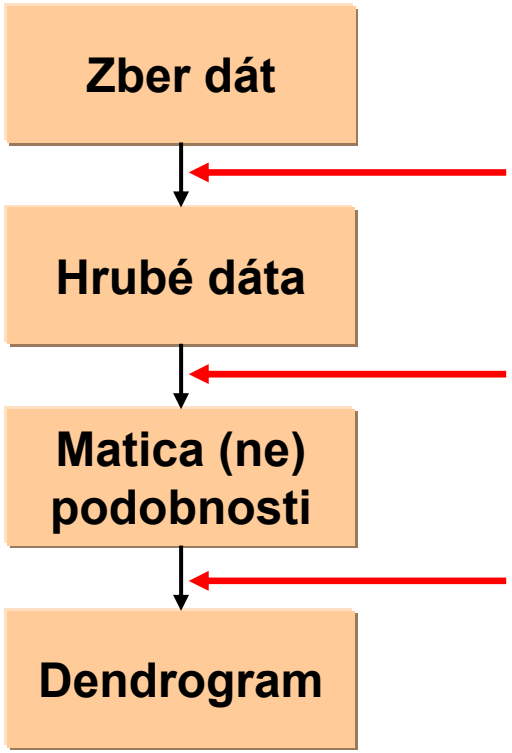


zakorenený dendrogram (*rooted*)



Hierarchické aglomeratívne zhlukovanie

Výsledok klasifikácie je ovplyvnený rozhodnutím na niekoľkých úrovniach



dôležitostná hodnota (pokryvnosť, početnosť)

transformácia, štandardizácia, meranie podobnosti

zhlukovací algoritmus

Podľa Kováča a Lepša (1986) majú transformácie väčší vplyv na výsledok zhlukovania než metódy zhlukovania.

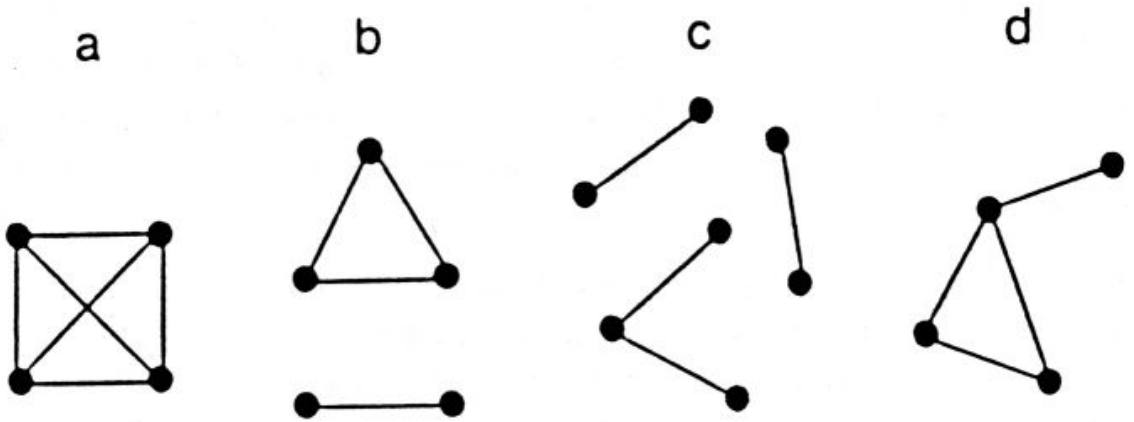
Kritické problémy analýzy

- Veľké množstvo parametrov alebo objektov v dendrograme je obtiažne interpretovať
- Analýza je silne závislá na zvolení vhodnej **metriky vzdialenosti**
- Analýza je silne závislá na **zhlukovacom algoritme**

Hierarchické aglomeratívne zhlukovanie

Zhody (*ties*)

- Při použití aglomeratívnych zhlukových metód môže nastať situácia, kedy sa v matici podobností vyskytujú tzv. zhody (*ties*)
- Najčastejšie dochádza k zhodám pri analýze binárnych dát, je tu veľká pravdepodobnosť rovnakej vzdialenosti medzi objektami
- Náhodné riešenie takejto situácie môže ovplyvniť výslednú klasifikáciu (dendrogram)



a – graf je úplný, b – graf je nesúvislý a všetky izolované komponenty sú úplné, c – graf je nesúvislý a aspoň jedna komponenta nie je úplná, d – graf je súvislý, ale nie je úplný

Hierarchické aglomeratívne zhlukovanie

Riešenie situácií

a) spoja sa všetky objekty naraz

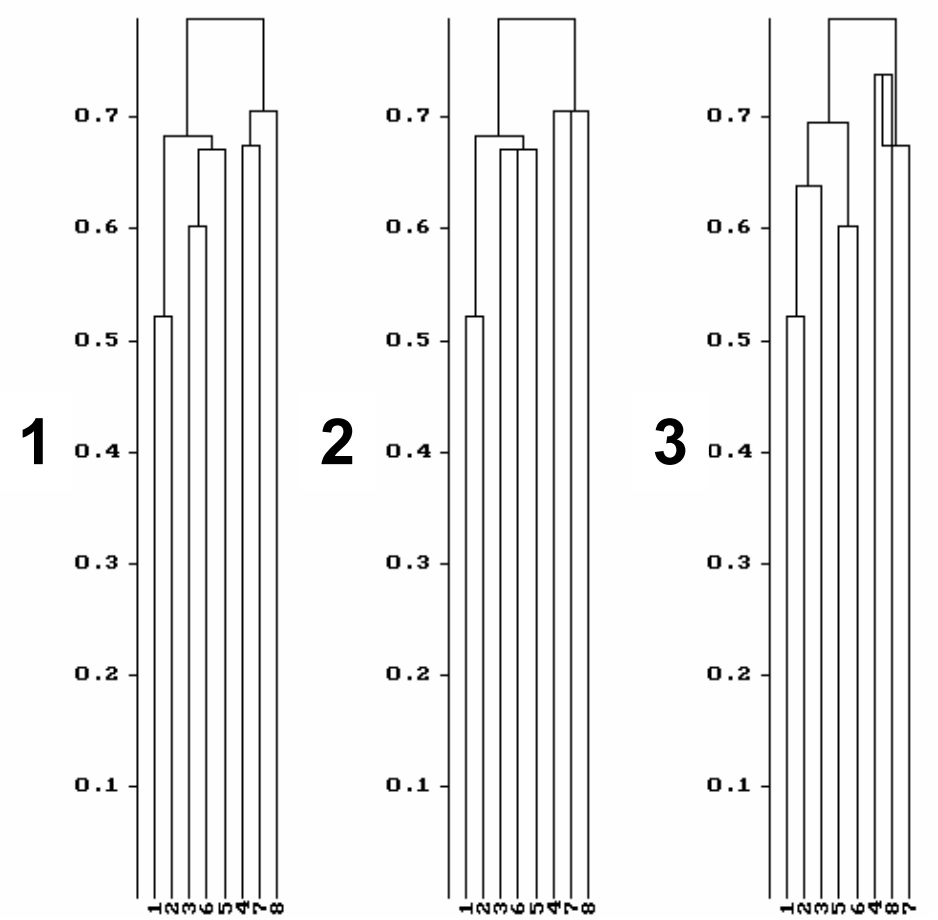
b) paralelne sa vytvorí viac skupín (tzv. *multiple fusion*)

c) a d) tri možnosti riešenia:

1 „*silent mode (arbitrary)*“
Väzby sa riešia náhodne, spojí sa len posledná nájdená dvojica (je tu vplyv poradia objektov v primárnej matici)

2 „*single linkage*“
Všetky objekty, ktoré sú spojené väzbou, sa spoja do jedného zhluku

3 „*suboptimal fusions*“
Nekompletné komponenty sa ignorujú a hľadanie najmenších vzdialeností v matici pokračuje kým sa už žiadne nekompletné komponenty nevyskytujú



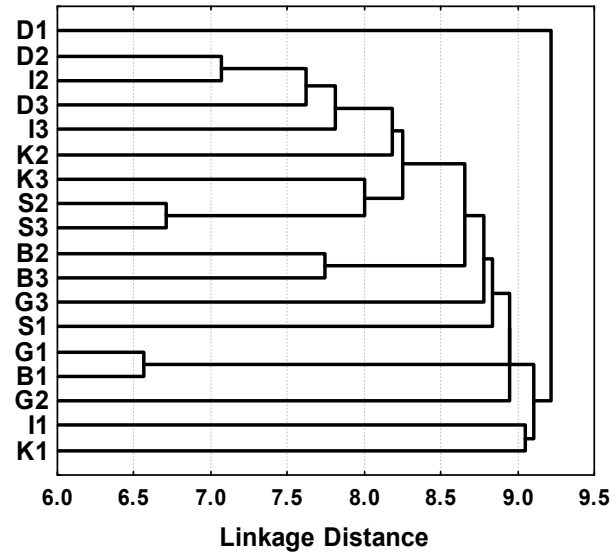
Hierarchické aglomeratívne zhlukovanie

hierarchical techniques — agglomerative clustering

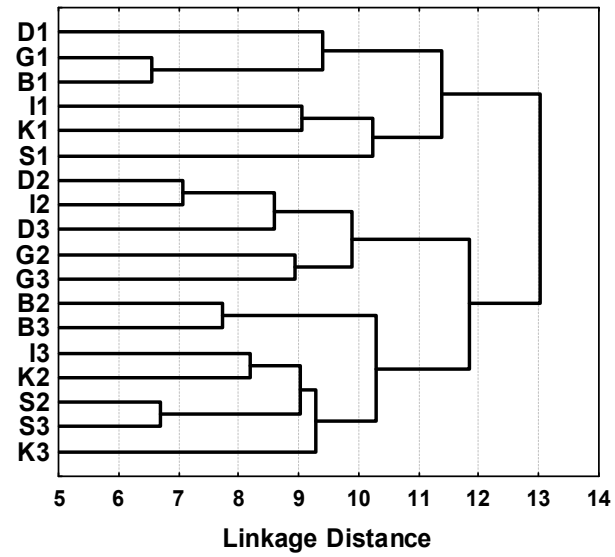
REÁLNE DÁTA

- 6 lokalít, každá lokalita monitorovaná v 3 obdobiach
- dátová matica: 18 vzoriek x 63 planktónnych druhov; hodnoty = stupeň dominancie

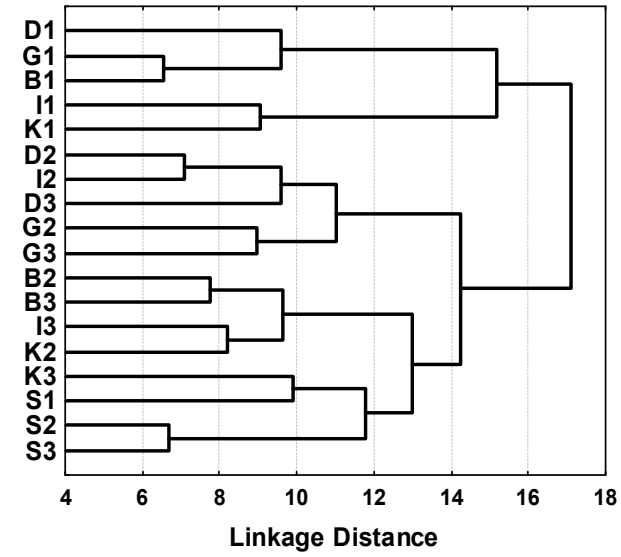
single-linkage



average-linkage



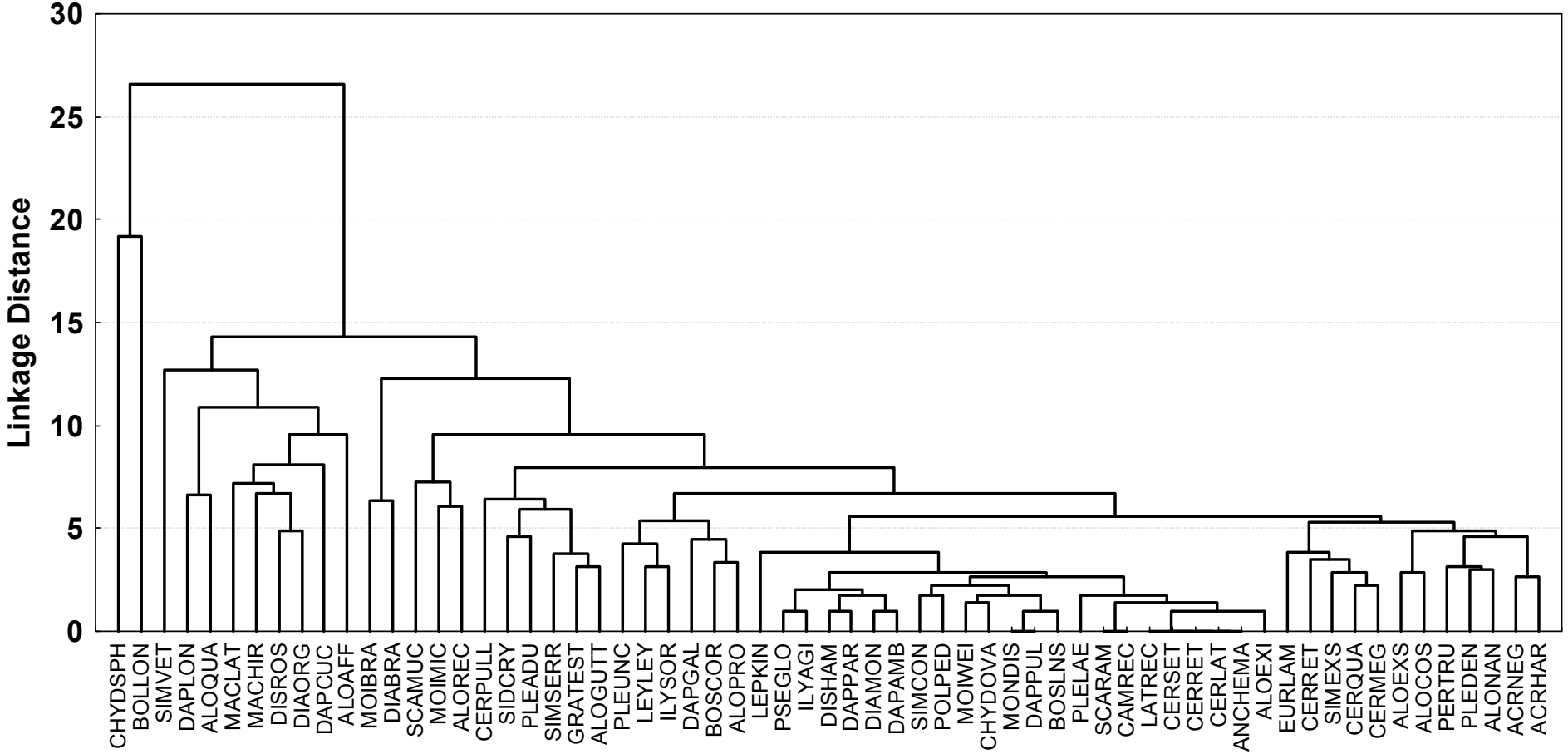
complete-linkage



Dendrogramy vytvorené pomocou troch rôznych zhlukovacích algoritmov: single-linkage, average, and complete-linkage. V prvom prípade (single-linkage) je zjavné silné zreťazenie objektov.

Hierarchické aglomeratívne zhlukovanie

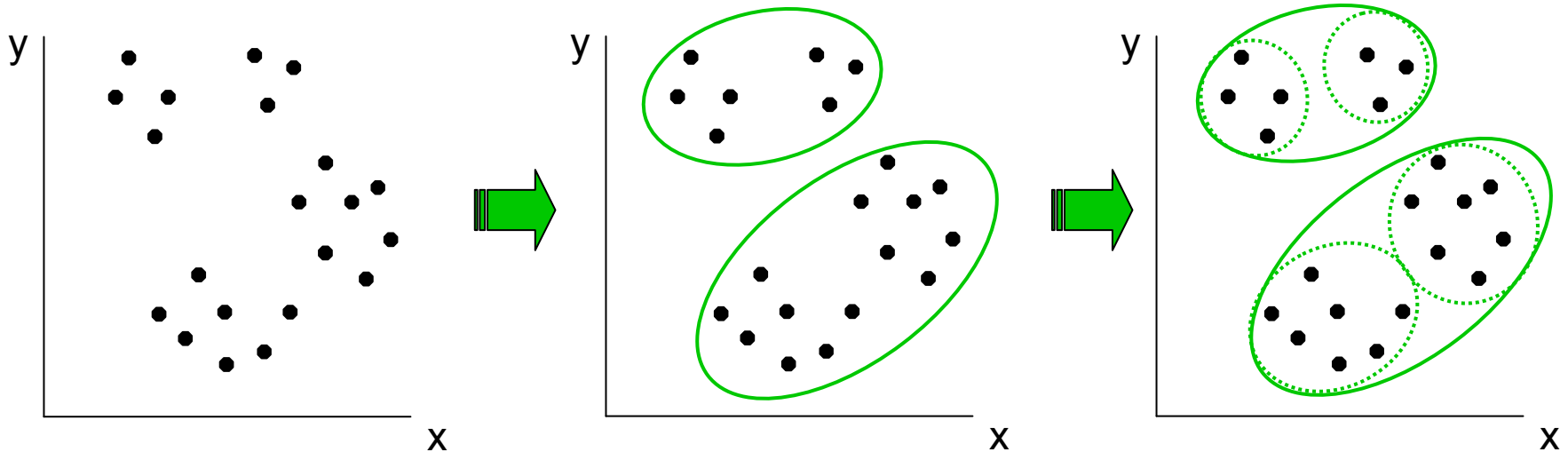
Podobne môžeme počítať aglomeratívnu hierarchickú klasifikáciu (*cluster analysis*) **pre premenné** (napr. pre **druhy**). V tomto prípade bude zrejme rozumným merítkom distribučnej podobnosti druhu **korelačný koeficient** (merítko rozumnej podobnosti sa líši podľa toho, či porovnávame vzorky alebo druhy).



Hierarchické divizívne zhlukovanie

hierarchical techniques — divisive clustering

- delenie prebieha „zhora“; začína všetkými objektami ako s jednou skupinou
- rozdelenie súboru na 2 časti
- ďalšie delenie častí

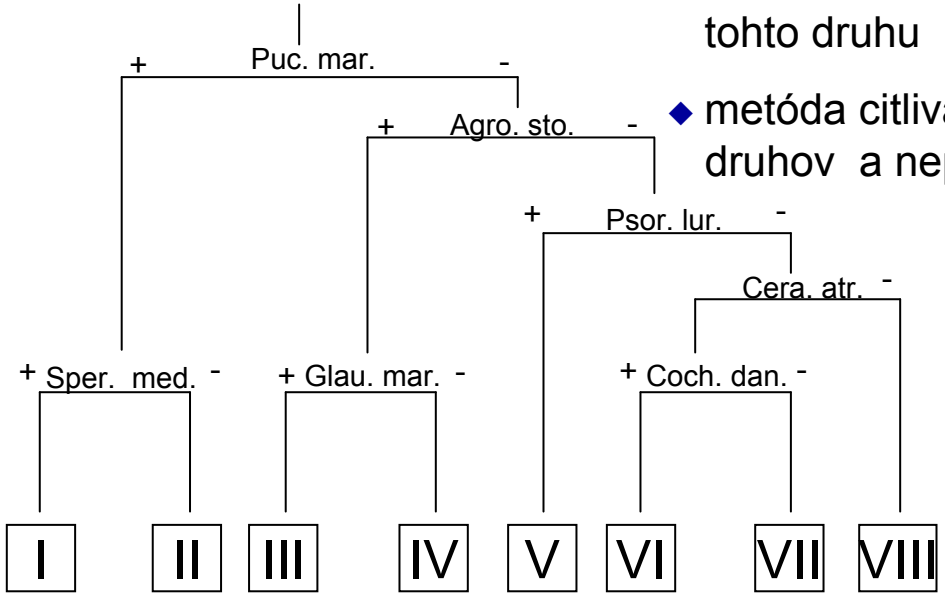


Časté použitie ku klasifikácii biologických spoločenstiev

Hierarchické divizívne zhlukovanie

hierarchical techniques — divisive clustering

- monothetic method
  - association analysis
    - delenie na základe jedného parametra
    - najprv je nájdený druh, ktorý je najviac asociovaný s ostatnými druhmi; skupiny sú rozdelené na základe prezencie/absencie tohto druhu
    - metóda citlivá na prítomnosť vzácnych druhov a neprítomnosť bežnejších druhov



A binary key for identifying types of salt-marsh habitat (Ivemey-Cook, Proctor 1966)

- polythetic method
  - two way indicator species analysis

TWINSPAN

Hierarchické divizívne zhlukovanie

polythetic method

two way indicator species analysis

TWINSPAN

- delenie skupiny je založené na všetkých druhoch podľa ich skóre na prvej osi vytvorenej ordináciou (v TWINSPAN-e korešpondenčná analýza)
- dichotómia vzniká ordináciou lokalít na základe diferenciálnych druhov
- berie do úvahy aj abundancie druhov vo forme tzv. pseudo-druhov => potrebné určiť hraničné hodnoty (cut levels)

Pôvodná tabuľka

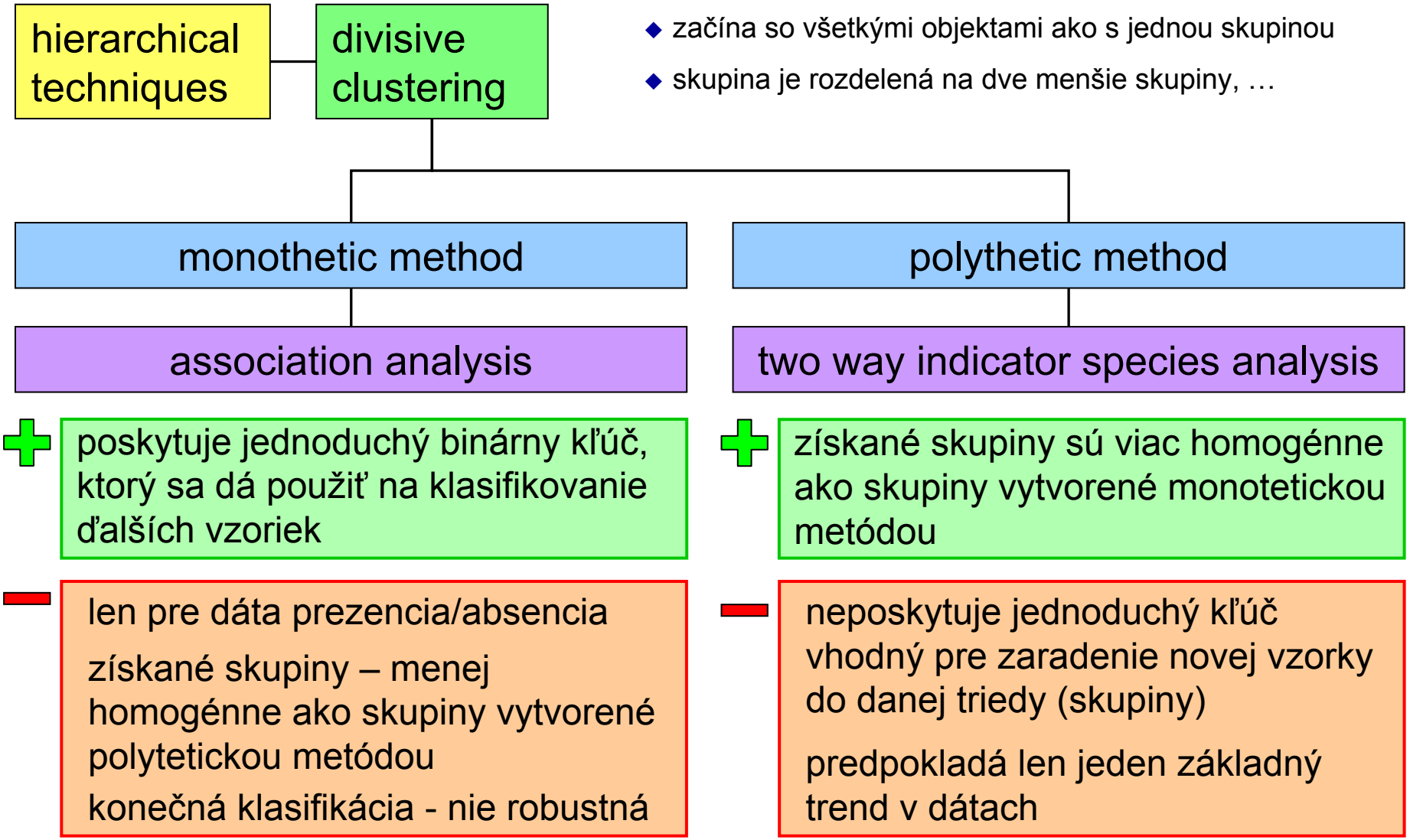
Species	A	B
Cirsium oleraceum	0	1
Glechoma hederacea	6	0
Juncus tenuis	15	25

cut levels 1, 5 a 20

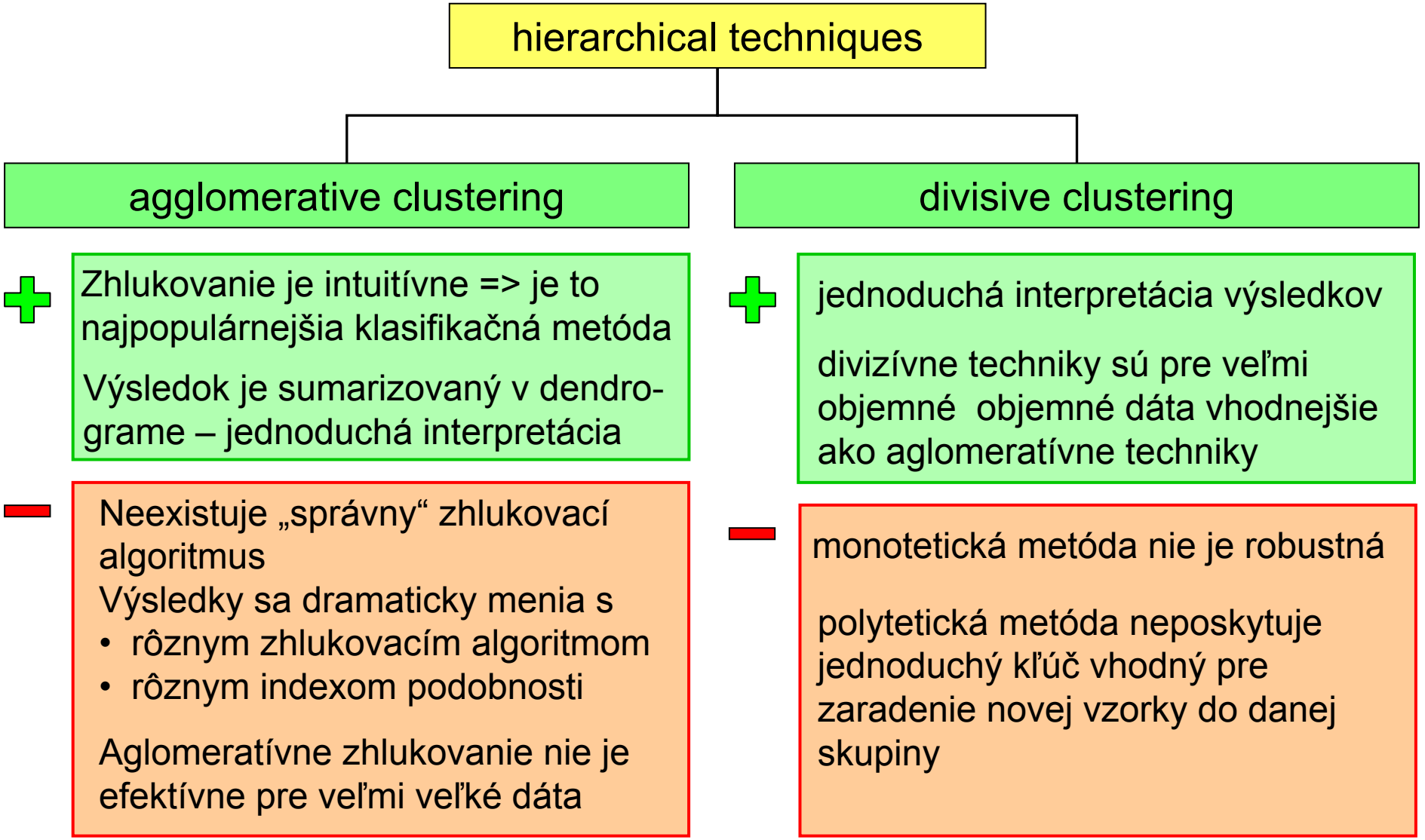
Tabuľka s pseudodruhmi použitými v TWINSPAN

Species	A	B
Cirsoler1	0	1
Glechede1	1	0
Glechede2	1	0
Junctenu1	1	1
Junctenu2	1	1
Junctenu3	0	1

Hierarchické divizívne zhlukovanie



Hierarchické zhlukovanie

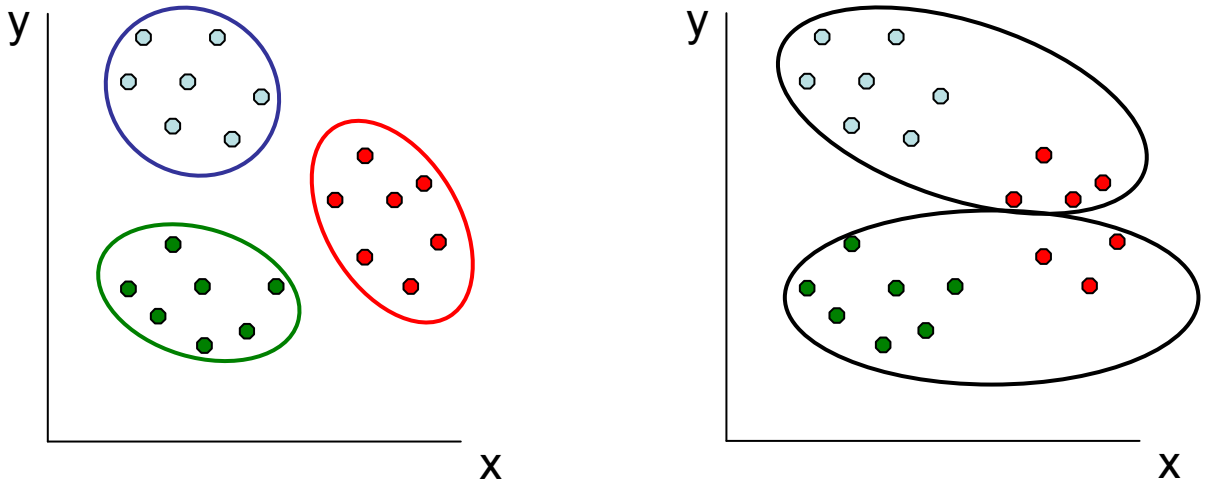


Nehierarchické zhlukovanie

non-hierarchical techniques

Nehierarchické zhlukovanie

Objekty sú na základe zadaného počtu zhlukov rozdelené podľa kritéria maximálnej homogenity zhlukov.



Ukážka rozdelenia objektov do zhlukov nehierarchickou metódou *k-means clustering*. Výsledok je ovplyvnený voľbou počtu zhlukov.

Vľavo: počet zhlukov 3 je dobrá voľba; vpravo: počet zhlukov 2 je zlá voľba.

Nehierarchické zhlukovanie

Princíp nehierarchického zhlukovania

- Pre výpočet sa používa opakovaná relokačná procedúra. Začína s *k* skupinami a potom presúva objekty tak, aby minimalizovala variablitu vnútri skupín a maximalizovala variabilitu medzi skupinami.
- Relokačná procedúra sa ukončí, keď žiadny ďalší presun už kritéria nezlepší.
- Takto získavame však len lokálny extrém, nemáme istotu, že je zároveň globálnym extrémom
- Odporúča sa začať s rôznymi počiatočnými skupinami a sledovať, či sú výsledky týchto analýz rovnaké.

Rizika analýzy

- pri chybnom odhade počtu zhlukov dáva metóda chybné výsledky
- výpočet je možný len na Euklidovských vzdialenostiach so všetkými jej obmedzeniami

Nehierarchické zhlukovanie

non-hierarchical techniques — K-means clustering

- skupiny nie sú zahrnuté do väčších skupín, ani neobsahujú menšie skupiny
- rozdeľuje objekty do určitého počtu skupín
- *K*-means clustering pracuje s euklidovskými vzdialenosťami

Nehierarchické metódy môžu byť vhodnejšie ako hierarchické techniky
- v prípade väčšieho objemu dát
- v prípade, že v dátach neexistuje hierarchická štruktúra

počet skupín *K* je potrebné špecifikovať vopred užívateľom

K-means clustering pracuje s euklidovskými vzdialenosťami
=> to môže byť problémom v prípade, keď euklidovská vzdialenosť nie je „najlepšiou“ metrikou

Zhluková analýza všeobecne

Keď dáta nemajú úplne jednoznačnú a zreteľnú štruktúru (jedná sa viacmenej o náhodne rozptýlené objekty), je pravdepodobné, že použitie rôznych zhlukovacích techník prinesie odlišné výsledky.

Pokiaľ rôzne zhlukovacie techniky prinášajú z toho istého súboru dát zhodné, resp. podobné výsledky, je to do istej miery potvrdenie štruktúry obsiahnutej v dátach (hoci zhlukovacie metódy patria k postupom produkujúcim hypotézy a nie sú určené k ich testovaniu).

Mnohé zhlukovacie techniky sú citlivé na prítomnosť odľahlých objektov (*outliers*, výrazne atypické prípady). Pred samotnou zhlukovou analýzou je preto vhodné použiť niektorú z metód na ich detekciu, napr. PCA. Výrazne odľahlé objekty spravidla z ďalších analýz vylúčime.

Zhlukové analýzy všeobecne nie sú vhodné na dáta, ktoré popisujú variabilitu znaku závislom na gradiente prostredia.

Zhluková analýza súhrn

Vstup zhlukovej analýzy:

- Matica podobnosti alebo vzdialenosti objektov
- Tabuľka objektov charakterizovaných niekoľkými parametrami

Výstup zhlukovej analýzy:

- Strom (dendrogram) pri hierarchickej zhlukovej analýze
- Zaradenie objektov do vopred definovaného počtu zhlukov pri nehierarchickej analýze

Pri použití zhlukovej analýzy je nutné pamätať na obmedzenia:

- aglomeratívne zhlukovanie nie je efektné pre veľmi veľké dáta
- pri hierarchickej aglomeratívnej analýze je výsledok silne ovplyvnený výberom indexu podobnosti, resp. metrikou vzdialenosti a zhlukovacím algoritmom
- *! neexistuje správny zhlukovací algoritmus !!!*
- pri hierarchickej divizívnej analýze: monotetická metóda nie je robustná; polytetická metóda predpokladá jeden hlavný trend v dátach a je ovplyvnená nastavením hraníc pseudo-druhov
- pri nehierarchickom zhlukovaní je nutné určiť počet skupín vopred