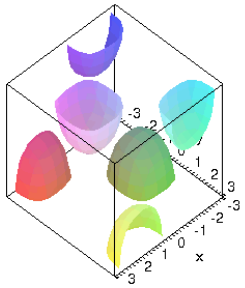
Vícerozměrná analýza biodiversity

Metody analýzy biodiverzity

Species abundance modely

Vícerozměrná analýza



Indexy diverzity

X

Vícerozměrná analýza společenstev: výhody a nevýhody

- Na data biodiverzity může být aplikována řada shlukovacích, ordinačních, regresních a klasifikačních vícerozměrných technik.

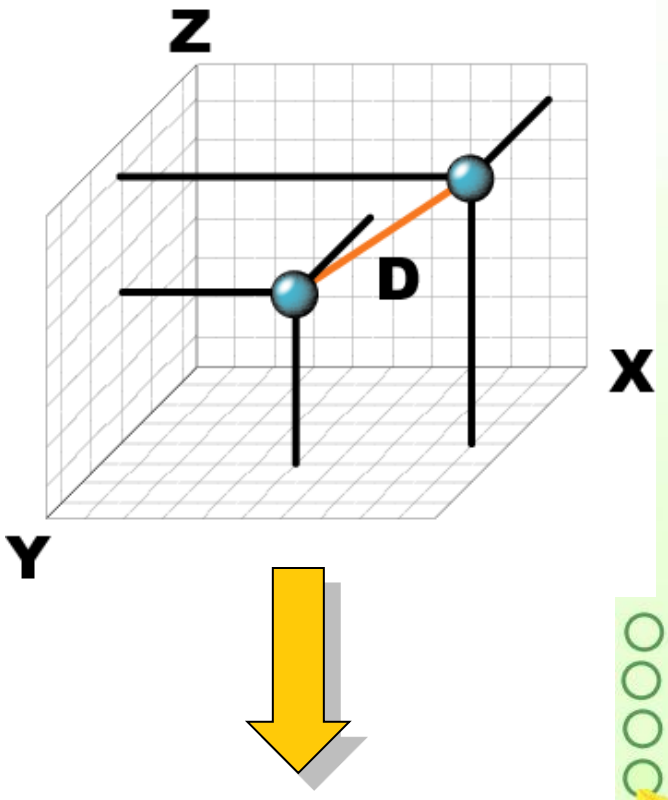
- Tyto metody hledají v rozsáhlých datech vícerozměrné vzory společenstev umožňující odpovědět na následující otázky:
  - Vztah druhů k prostředí
  - Prostorové vztahy
  - Interakce taxonů

- Výhody:
  - Shrnující výsledky postihující všechny aspekty dat
  - Identifikace skrytých interakcí a vztahů mezi proměnnými

- Nevýhody:
  - Náročné na data a metodiku
  - Vyžadují expertní znalosti jak v oblasti statistické metodiky, tak biologických společenstev, v opačném případě mohou vést k nesprávným závěrům a interpretacím

Cíle vícerozměrné analýzy dat

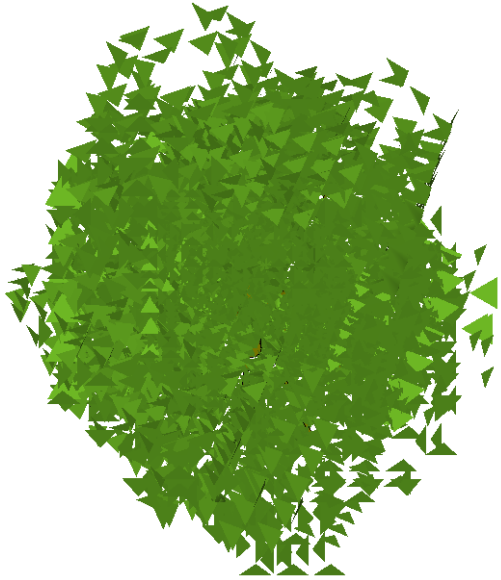
- Každý objekt reálného světa můžeme popsat jeho pozicí v mnohorozměrném prostoru
- Více než 3D prostor je pro nás vizuálně neuchopitelný a hledání vztahů ve více než 3 dimenzích je problematické
- Vícerozměrná analýza se tento problém snaží řešit různými přístupy:
  - Redukce dimenzionality dat „sloučením“ korelovaných proměnných do menšího počtu „faktorových“ proměnných
  - Identifikace shluků objektů ve vícerozměrném prostoru a následná redukce vícedimenzionálního problému kategorizací objektů do zjištěných shluků



Zjednodušení
Interpretace

Vícerozměrná analýza dat = pohled ze správného úhlu

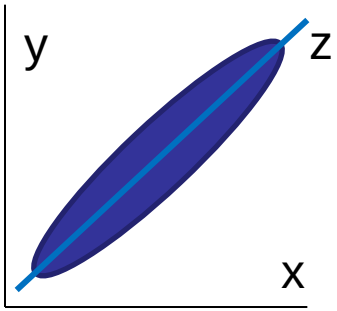
- Vícerozměrná analýza nám pomáhá nalézt v x-dimenzionálním prostoru nejvhodnější pohled na data poskytující maximum informací o analyzovaných objektech



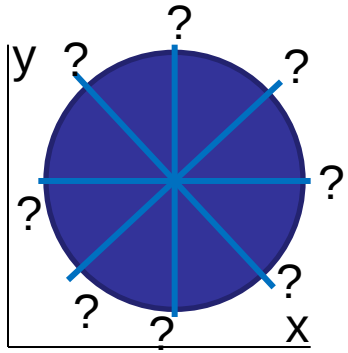
Všechny obrázky ukazují stejný objekt z různých úhlů v 3D prostoru.

Obecný princip redukce dimenzionality dat

- V převážné většině případů existují mezi dimenzemi **korelační vztahy**, tedy dimenze se navzájem vysvětlují a pro popis kompletní informace v datech není třeba všech dimenzí vstupního souboru
- Všechny tzv. ordinační metody využívají principu identifikace korelovaných dimenzí a jejich sloučení do souhrnných nových dimenzí zastupujících několik dimenzí vstupního souboru
- Pokud mezi dimenzemi vstupního souboru neexistují korelace, nemá smysl hledat zjednodušení vícerozměrné struktury takovéhoto souboru !!!



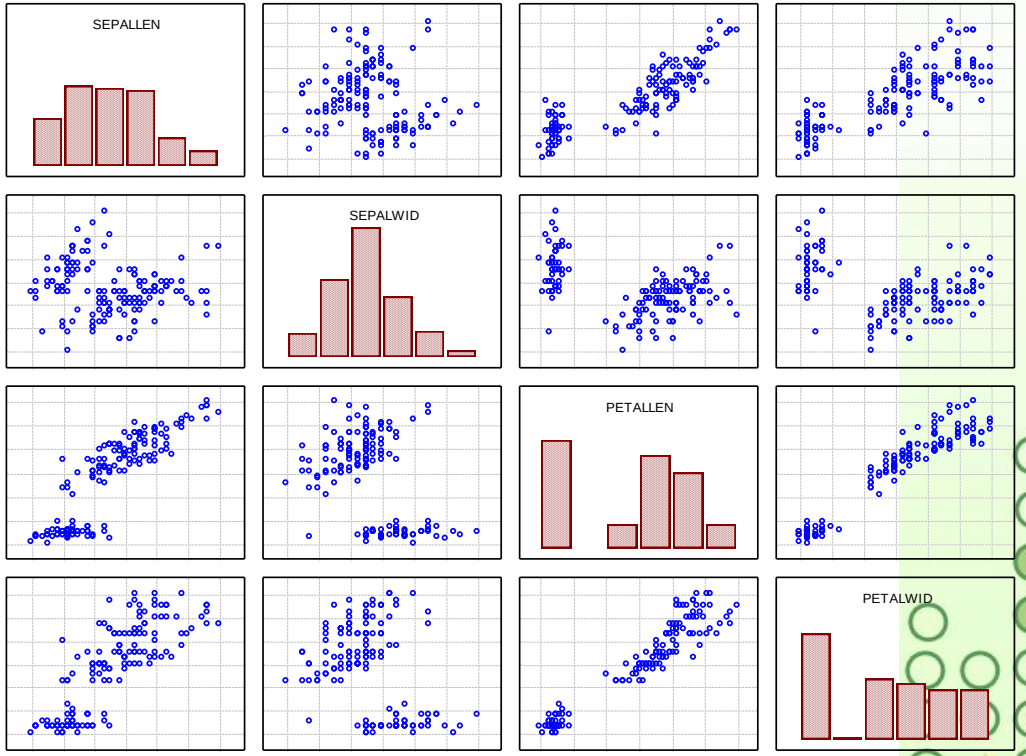
Jednoznačný vztah dimenzí x a y umožňuje jejich nahrazení jedinou novou dimenzí z



V případě neexistence vztahu mezi x a y nemá smysl definovat nové dimenze – nepřináší žádnou novou informaci oproti x a y

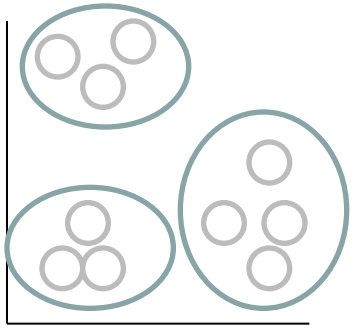
Příklad vícerozměrného popisu objektů

	Dimenze 1	Dimenze 2	Dimenze 3	Dimenze 4
ID objektu	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SETOSA	5.0	3.3	1.4	0.2
VIRGINIC	6.4	2.8	5.6	2.2
VERSICOL	6.5	2.8	4.6	1.5
VIRGINIC	6.7	3.1	5.6	2.4
VIRGINIC	6.3	2.8	5.1	1.5
SETOSA	4.6	3.4	1.4	0.3
VIRGINIC	6.9	3.1	5.1	2.3
VERSICOL	6.2	2.2	4.5	1.5
VERSICOL	5.9	3.2	4.8	1.8
SETOSA	4.6	3.6	1.0	0.2
	...	...	...	...

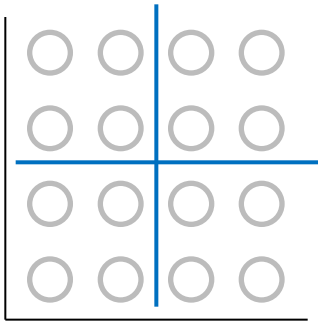


Obecný princip hledání shluků v datech

- Vzájemnou pozici objektů ve vícerozměrném prostoru lze popsat jejich **vzdáleností**
- Dle vzdálenosti objektů je můžeme slučovat do shluků a přiřazení objektů ke shlukům ve vícerozměrném prostoru následně využít pro zjednodušení jejich x-dimenzionálního popisu
- Smysluplnost výsledků shlukování závisí jednak na objektivní existenci shluků v datech, jednak na arbitrárně nastavených kritériích definice shluků



Jednoznačné odlišení existujících shluků v datech (obdoba multimodálního rozložení)



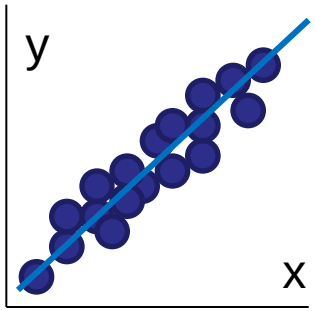
Shluková analýza je možná i v tomto případě, nicméně hranice shluků jsou dány pouze naším rozhodnutím.

Omezení vícerozměrné analýzy dat

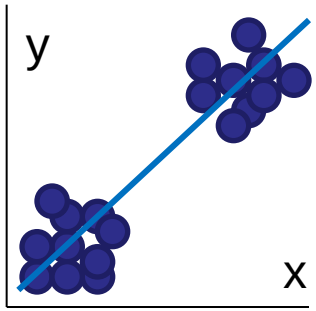
- Vícerozměrná analýza může přinést zjednodušení dimenzionality dat pouze v případě, kdy **data** skrývají nějakou identifikovatelnou **vícerozměrnou strukturu**
  - Mezi dimenzemi existují **vztahy (korelace)** umožňující nahrazení korelovaných dimenzí zástupnou souhrnnou dimenzí
  - Objekty vytváří v x-dimenzionálním prostoru **shluky** nebo jiné nenáhodné struktury
- Pro náhodně rozmístěné objekty bez korelací mezi dimenzemi jejich x-dimenzionálního prostoru nepřináší vícerozměrná analýza žádné nové informace oproti původním dimenzím
- Důležitý je **poměr počtu objektů** (řádky tabulky) **a dimenzí** (sloupce tabulky). Čím je tento poměr menší tím větší je šance, že výsledky analýzy jsou ovlivněny náhodnými procesy. Za minimální poměr pro získání validních výsledků je považováno 10 objektů na 1 dimenzi.
- Pro vícerozměrné analýzy platí obdobné **předpoklady** jako pro jednorozměrnou statistickou analýzu; vzhledem k jejich možnému porušení na úrovni kombinace několika dimenzí je tyto předpoklady třeba kontrolovat ještě pečlivěji než u jednorozměrné analýzy
- Kromě klasických statistických předpokladů je při vícerozměrných analýzách třeba věnovat pozornost **výběru metrik vzdáleností mezi objekty** (klíčové ovlivnění interpretace výsledků) a jejich předpokladům
- Pokud **výsledky** vícerozměrné analýzy nejsou **interpretovatelné** je třeba zvážit, zda použití vícerozměrné analýzy přináší oproti sadě jednorozměrných analýz nějakou přidanou hodnotou
- Využitelná vícerozměrná analýza by měla být:
  - Vybrána vhodná metoda pro řešení daného problému
  - korektně spočítána za dodržení všech předpokladů
  - Interpretovatelná a přinášející novou informaci oproti analýze původních dimenzí

Korelace jako princip výpočtu vícerozměrných analýz

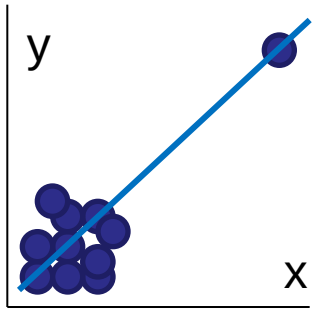
- Kovariance a Pearsonova korelace je základem analýzy hlavních komponent, faktorové analýzy jakož i dalších vícerozměrných analýz pracujících s lineární závislostí proměnných
- Předpokladem výpočtu kovariance a Pearsonovy korelace je:
  - Normalita dat v obou dimenzích
  - Linearita vztahu proměnných
- Pro vícerozměrné analýzy je nejzávažnějším problémem přítomnost odlehlých hodnot



Lineární vztah – bezproblémové použití Pearsonovy korelace



Korelace je dána dvěma skupinami hodnot – vede k identifikaci skupin objektů v datech



Korelace je dána odlehlou hodnotu – analýza popisuje pouze vliv odlehlé hodnoty

Analýza kontingenčních tabulek jako princip výpočtu vícerozměrných analýz

- Abundance taxonů (nebo počet jakýchkoliv objektů) na lokalitách lze brát jako kontingenční tabulku a mírou vztahu mezi řádky (lokality) a sloupci (taxony) je velikost chi-kvadrátu

$$\chi^2_{(1)} = \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$

Počítáno pro každou buňku tabulky

	☠	☺
A	10	0
B	0	10

Pozorovaná tabulka

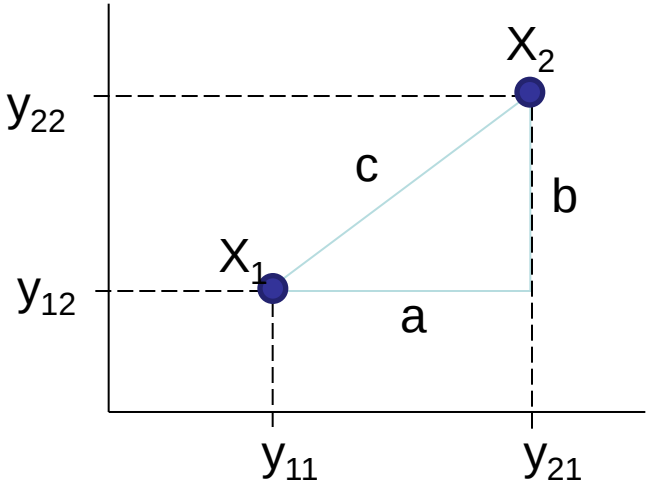
	☠	☺
A	5	5
B	5	5

Očekávaná tabulka

Hodnota chi-kvadrátu definuje míru odchylky dané buňky (v našem kontextu vztahu taxon-lokalita) od situace, kdy mezi řádky a sloupci (taxon-lokalita) není žádný vztah

Euklidovská vzdálenost jako princip výpočtu vícerozměrných analýz

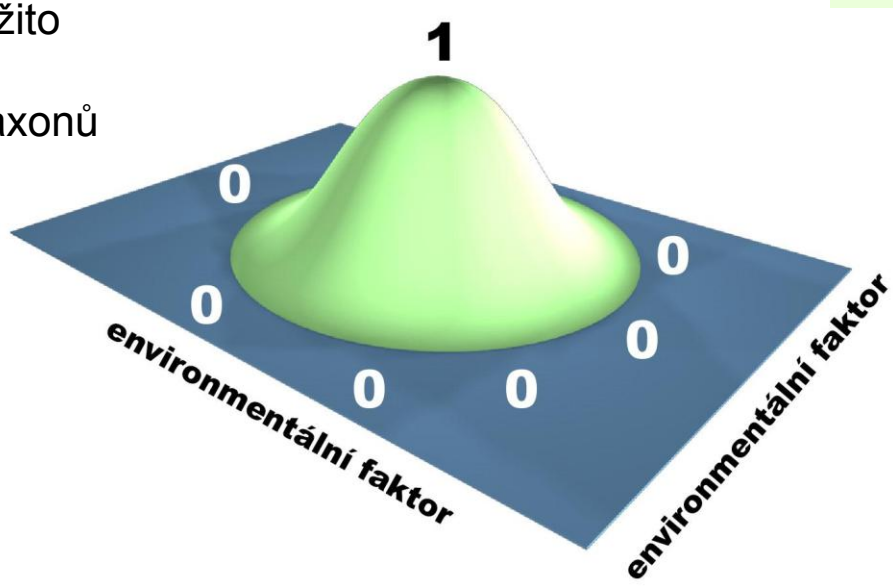
- Nejsnáze představitelným měřítkem vztahu dvou objektů ve vícerozměrném prostoru je jejich vzdálenost
- Nejjednodušším typem této vzdálenosti (bohužel s omezeným použitím na data společenstev) je Euklidovská vzdálenost vycházející z Pythagorovy věty



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

Double zero problém

- V případě binárních metrik (druh se vyskytuje/nevyskytuje) není možné uvažovat stejnou váhu pro souhlas přítomnosti (11) a nepřítomnosti (00) taxonů (symetrický koeficient)
- Problémem využití všech typů metrik pro data abundancí spočívá v odlišném významu přítomnosti a nepřítomnosti taxonů
- Pokud se taxon nachází v obou srovnávaných společenstvech – znamená to že společenstva si budou v tomto ohledu podobná, protože mají podmínky umožňující přítomnost taxonu
- Pokud se taxon nenachází ani v jednom ze dvou srovnávaných společenstev – příčina může být nejrůznější – double zero problem
- Pro odstranění tohoto problému je použito asymetrické hodnocení souhlasné přítomnosti (11) a nepřítomnosti (00) taxonů (asymetrické koeficienty)



Pojmy vícerozměrných analýz

- **Vícerozměrné metody**: Název vícerozměrné vychází z typu vstupních dat, tato data jsou tvořena jednotlivými objekty (i.e. klienti) a každý z nich je charakterizován svými parametry (věk, příjem atd.) a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.

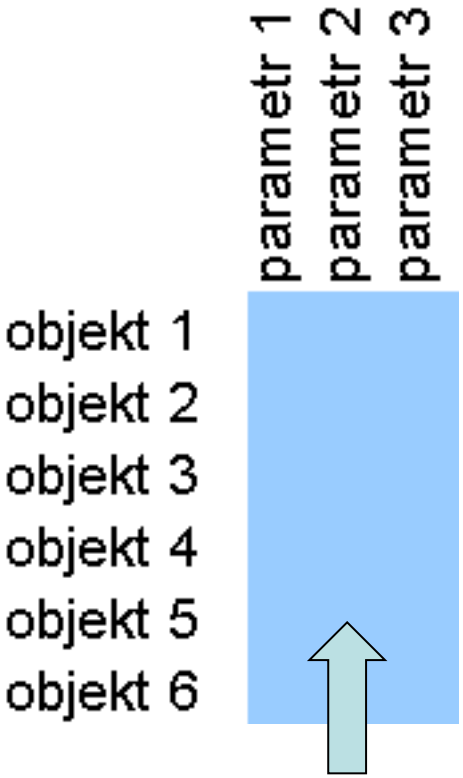
- **Maticová algebra**: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.

- **NxP matice**: N objektů s p parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.

- **Asociační matice**: Na základě těchto matic jsou počítány matice asociační na nichž pak probíhají další výpočty, jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. metriky) buď objektů (Q mode analýza) nebo parametrů (R mode analýza). Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

Vstupní matice vícerozměrných analýz

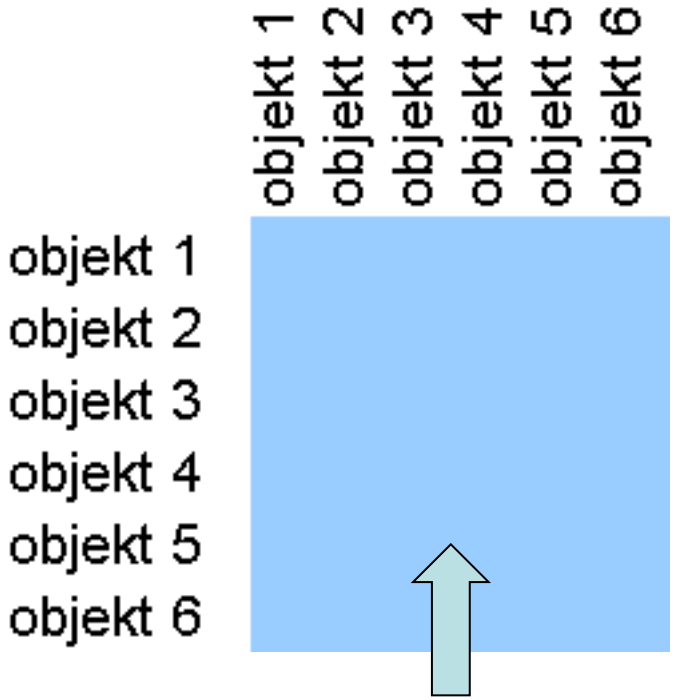
NxP MATICE



Hodnoty parametrů pro jednotlivé objekty

Výpočet metriky podobností/ vzdáleností

ASOCIAČNÍ MATICE



Korelace, kovariance, vzdálenost, podobnost

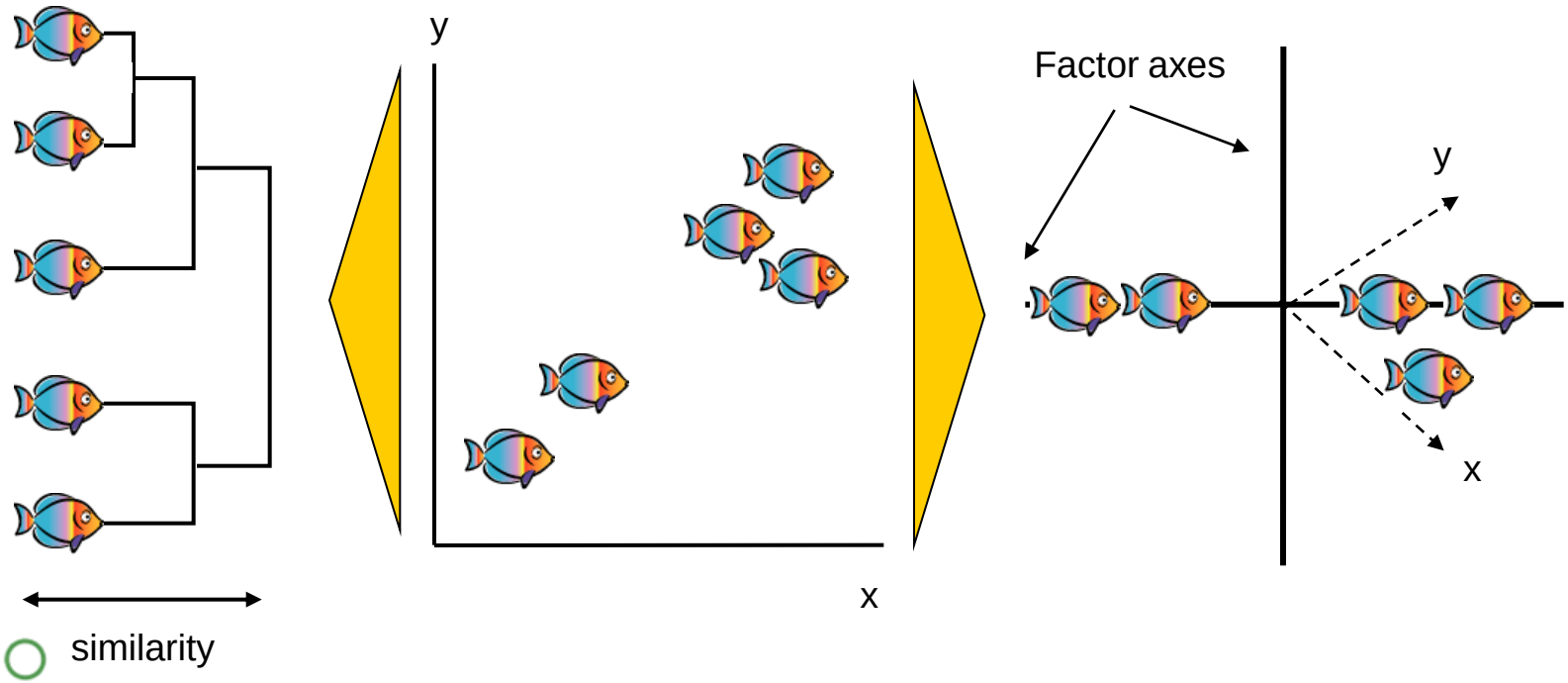
Základní typy vícerozměrných analýz

Shluková analýza

- Vytváření shluků objektů na základě jejich podobnosti
- Identifikace typů objektů

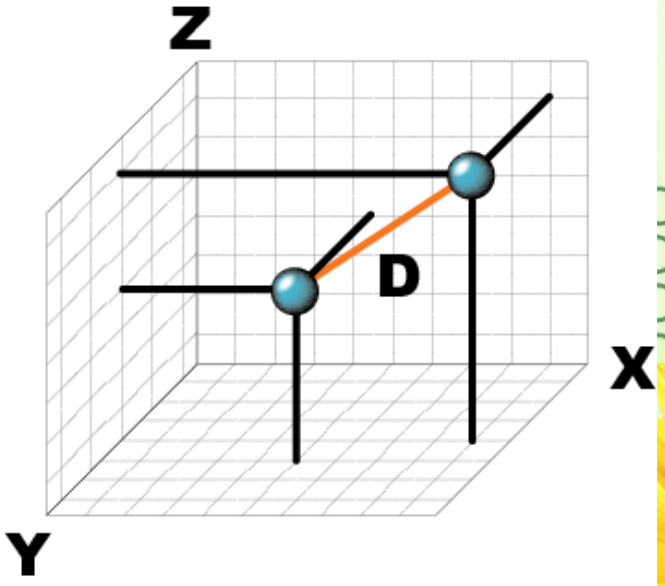
Ordinace

- Zjednodušení vícerozměrného problému do menšího počtu rozměrů
- Principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat



Seznam taxonů – vícerozměrný popis společenstva

- Na seznam taxonů lze pohlížet také jako seznam rozměrů společenstva
- Záznam o nalezených taxonech tak vlastně tvoří vícerozměrný popis daného společenstva
- Společenstva můžeme srovnávat podle jejich vzájemné pozice v n-rozměrném prostoru
- Pro srovnání společenstev lze teoreticky využít libovolnou metriku vícerozměrné podobnosti nebo vzdálenosti



Koeficienty podobnosti

Koeficienty podobosti (indexy podobnosti)

- V ekologii se využívá řada indexů podobnosti založených buď na přítomnosti/nepřítomnosti taxonů nebo na abundancích

Binární koeficienty podobnosti

Společenstvo 1			
Společenstvo 2		1	0
	1	a	b
	0	c	d

a, b, c, d = počet případů, kdy souhlasí binární charakteristika společenstev 1 a 2
a+b+c+d=p

Symetrické binární koeficienty - není rozdíl mezi případem 1-1 a 0-0
Asymetrické binární koeficienty - rozdíl mezi případem 1-1 a 0-0

Více informací a další měření vzdáleností a podobností najdete v knize **LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elseviere Science BV, Amsterodam.**

Symetrické binární koeficienty

Jednoduchý srovnávací koeficient (Sokal & Michener, 1958)

- Obvyklou metodou pro výpočet podobnosti mezi dvěma objekty je podíl počtu deskriptorů, které kódují objekt stejně, a celkového počtu deskriptorů. Při použití tohoto koeficientu předpokládáme, že není rozdíl mezi nastáním 0 a 1 u deskriptorů.

$$S_1(x_1, x_2) = \frac{a+d}{p}$$

Rogers & Tanimoto koeficient (1960)

- Dává větší váhu rozdílům než podobnostem.

$$S_2(x_1, x_2) = \frac{a + d}{a + 2b + 2c + d}$$

Sokal & Sneath (1963)

- Další čtyři navržené koeficienty obsahují double-zero, ale jsou navrženy tak, aby se snížil vliv double-zero:

$$S_3(x_1, x_2) = \frac{2a + 2d}{2a + b + c + 2d}$$

- tento koeficient dává dvakrát větší váhu shodným deskriptorům než rozdílným;

$$S_4(x_1, x_2) = \frac{a + d}{b + c}$$

- porovnává shody a rozdíly prostým podílem v měřítku jdoucím od 0 do nekonečna;

$$S_5(x_1, x_2) = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$$

- porovnává shodné deskriptory se součty okrajů tabulky;

$$S_6(x_1, x_2) = \frac{a}{\sqrt{(a+b)(a+c)}} \frac{d}{\sqrt{(b+d)(c+d)}}$$

- je vytvořen z geometrických průměrů členů vztahujících se k *a* a *d,* podle koeficientu S5.

Asymetrické binární koeficienty

Jaccardův koeficient (1900, 1901, 1908)

- Všechny členy mají stejnou váhu

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

Sørensenův koeficient (1948) (Coincidence index, Dice(1945))

- varianta předchozího koeficientu dává dvojnásobnou váhu dvojitým prezencím, protože se může zdát, že přítomnost druhů je více informativní než jejich absence, která může být způsobena různými faktory a nemusí nutně odrážet rozdílnost prostředí. Prezence druhu na obou lokalitách je silným ukazatelem jejich podobnosti. S7 je monotónní k S8, proto podobnost pro dvě dvojice objektů vypočítaná podle S7 bude podobná stejnému výpočtu S8. Oba koeficienty se liší pouze v měřítku. Tento index byl poprvé použit Dicem v R-mode studii asociací druhů. Jiná varianta tohoto koeficientu dává duplicitním prezencím trojnásobnou váhu.

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

Sokal & Sneath (1963)

- navržen jako doplněk Rogers & Tanimotova koeficientu (S2), dává dvojnásobnou váhu rozdílům ve jmenovateli.

$$S_{10}(x_1, x_2) = \frac{a + d}{a + 2b + 2c}$$

Russel & Rao (1940)

- navržená míra umožňuje porovnání počtu duplicitních prezencí (v čitateli) proti celkovému počtu druhů, nalezených na všech lokalitách, zahrnujícím druhy, které chybějí (*d*) na obou uvažovaných lokalitách.

$$S_{11}(x_1, x_2) = \frac{a}{p}$$

Kulczynski (1928)

- koeficient porovnávající duplicitní prezence s diferencemi

$$S_{12}(x_1, x_2) = \frac{a}{b + c}$$

Binární verze asymetrického kvantitativního Kulczynski koeficientu (1928)

- Mezi svými koeficienty pro presence/absence data zmiňují Sokal & Sneath (1963) tuto verzi kvantitativního koeficientu, kde jsou duplicitní prezence srovnávány se součty okrajů tabulky (*a*+*b*) a (*a*+*c*).

$$S_{13}(x_1, x_2) = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right]$$

Ochiachi (1957)

- použil jako míru podobnosti geometrický průměr poměrů *a* k počtu druhů na každé lokalitě, tj. se součty okrajů tabulky (*a*+*b*) a (*a*+*c*), tento koeficient je obdobou S6, bez části, týkající se double-zero (d).

$$S_{14}(x_1, x_2) = \sqrt{\frac{a}{(a+b)}\frac{a}{(a+c)}} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Faith (1983)

- V tomto koeficientu je neshoda (přítomnost na jedné a absence na druhé lokalitě) vážena proti duplicitní prezenci. Hodnota S26 klesá s růstem double-zero

$$S_{26}(x_1, x_2) = \frac{a + d/2}{p}$$

Kvantitativní koeficienty

„Klasické“ indexy podobnosti

- **Sørensenův kvantitativní koeficient**, kde aN a bN jsou celkové počty jedinců v společenstvech A a B, jN je pak suma abundancí pokud se druh nachází v obou společenstvech, je počítána vždy z nižší abundance daného druhu ve společenstvu

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

- **Morisita-Horn index**, kde aN je celkový počet jedinců ve společenstvu A a an_i počet jedinců druhu i ve společenstvu A (obdobně platí pro společenstvo B)

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

Jednoduchý srovnávací koeficient (Sokal & Michener, 1958)

- modifikovaný jednoduchý srvonávací koeficient může být použit pro multistavové deskriptory - čitatel obsahuje počet deskriptorů, pro které jsou dva objekty ve stejném stavu – např. je-li dvojice objektů popsána následujícími deseti multistavovými deskriptory: hodnota S1,vypočítaná pro 10 multistavových deskriptorů bude S1,(x1,x2) = 4 agreements/ 10 descriptors = 0.4
- Podobným způsobem je možné rozšířit všechny binární koeficienty pro multistavové deskriptory.

$$S_1(x_1, x_2) = \frac{agreements}{p}$$

	Deskriptors										**Σ**
Object x_1	9	3	7	3	4	9	5	4	0	6	
Object x_2	2	3	2	1	2	9	3	2	0	6	
Agreements	0	+1	+0	+0	+0	+1	+0	+0	+1	+1	4

Gowerův obecný koeficient podobnosti (1971) I.

- Obecný koeficient podobnosti může kombinovat různé typy deskriptorů. Podobnost mezi dvěma objekty je vypočítána jako průměr podobností, vypočítaných pro všechny deskriptory. Pro každý deskriptor j je hodnota parciální podobnosti s_{12j} mezi objekty x1 a x2 vypočítána následovně:

$$S_{15}(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} s_{12j}$$

- ✓ Pro binární deskriptory sj=1 (shoda) nebo 0 (neshoda). Gower navrhl dvě formy tohoto koeficientu (symetrická, dává sj=1 double-zero; asymetrická dává pro double-zero sj=0)
- ✓ Kvalitativní a semikvantitivní deskriptory jsou upraveny podle jednoduchého zaměňovacího pravidla, sj=1 při souhlasu a sj = 0 při nesouhlasu deskriptorů. Double zero jsou ošetřeny stejně jako v předchozím odstavci.
- ✓ Kvantitativní deskriptory (reálná čísla) jsou zpracovány následovně: pro každý deskriptor se nejprve vypočte rozdíl mezi stavy obou objektů který je poté vydělen největším rozdílem (Rj), nalezeným pro daný deskriptor mezi všemi objekty ve studii (nebo v referenční populaci – doporučuje se vypočítat největší diferenci Rj každého deskriptoru j pro celou populaci, aby byla zajištěna konzistence výsledků pro všechny parciální studie).

Gowerův obecný koeficient podobnosti (1971) II.

- normalizovaná vzdálenost může být odečtena od 1 aby byla transformována na podobnost:

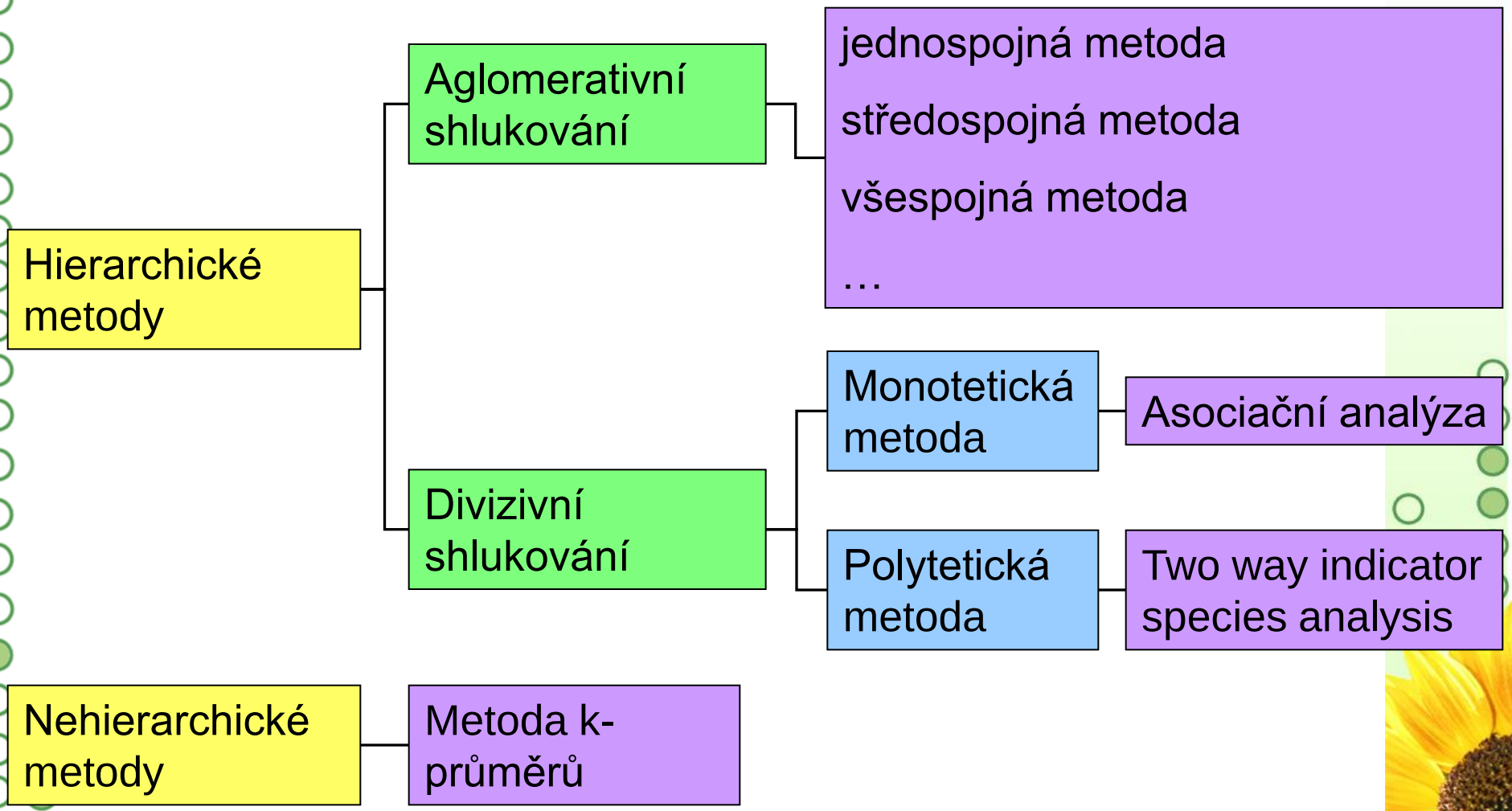
$$s_{12j} = 1 - \left[\frac{|y_{1j} - y_{2j}|}{R_j} \right]$$

- Gowerův koeficent může být nastaven tak, aby zahrnoval přídavný flexibilní prvek: žádné porovnání není vypočítáno u deskriptorů, u nichž chybí informace buď u jednoho, nebo u druhého objektu. Toto zajišťuje člen *wj*, nazývaný Kroneckerovo delta, popisující přítomnost/nepřítomnost informace v obou objektech: je-li informace o deskriptoru yj přítomna u obou objektů (*wj*=1), jinak (*wj*=0), tento koeficient nabývá hodnot podobnosti mezi 0 a 1 (největší podobnost objektů). Další možností je vážení různých deskriptorů prostým přiřazením čísla v rozsahu 0-1 *wj*.

$$S_{15}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}$$

Shlukování dat biodiverzity

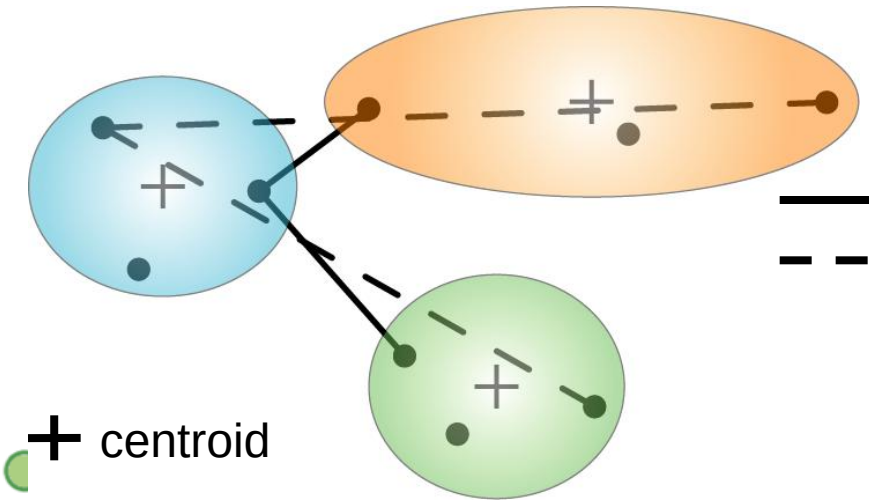
Shluková analýza



Hierarchické aglomerativní shlukování

Hierarchické metody — Aglomerativní shlukování

- začíná jednotlivými objekty, které jsou spojeny do větších shluků
- vyžaduje matici podobností nebo nepodobností (site by site), kterou začíná
- pro data prezence/absence i pro kvantitativní data existuje mnoho indexů podobnosti
- všechny aglomerativní metody jsou založeny na spájení jednotlivých objektů (vzorek) nebo shluků do větších skupin



\+ centroid

vzdálenost u **jednospojné metody**

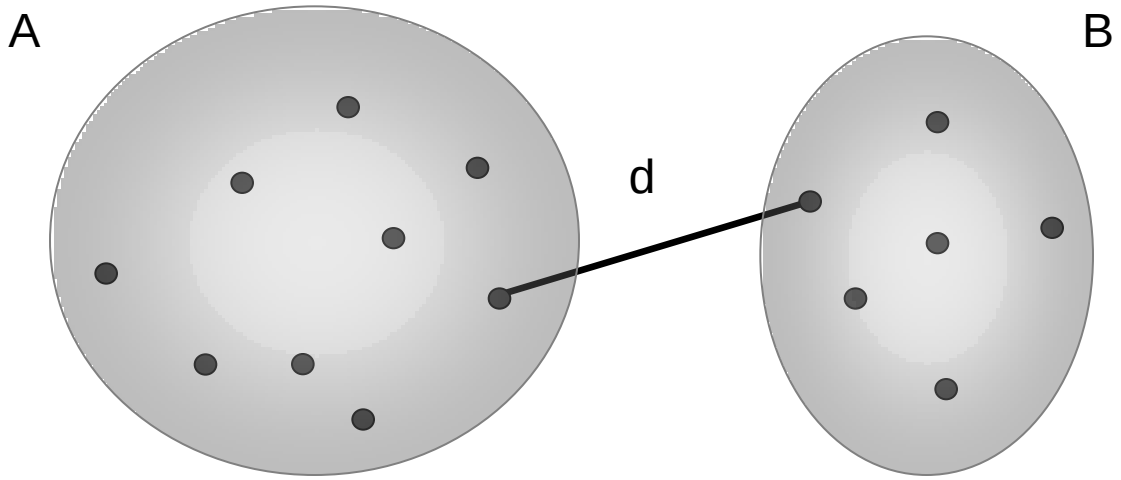
vzdálenost u **všespojné metody**

jiné metody:
Vzdálenost mezi centroidy
Průměrná vzdálenost
...

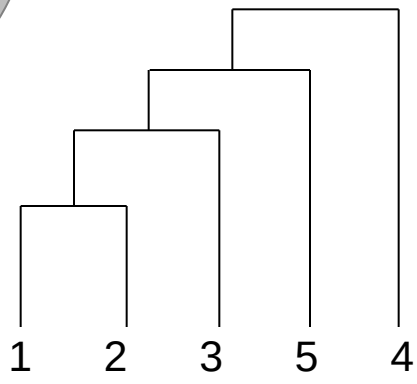
Hierarchické aglomerativní shlukování

Metoda nejbližšího souseda

(jednospojná metoda, metoda jediné vazby, *single linkage, the nearest neighbor method*)



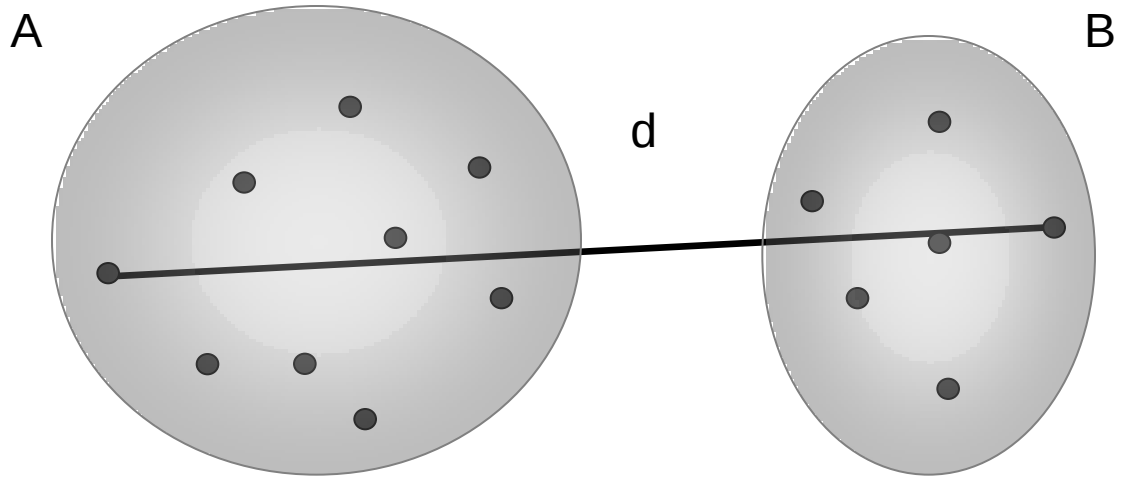
Vzdálenost mezi dvěma shluky je daná jako minimální vzdálenost mezi všemi možnými zástupci shluků.
Často se i velmi vzdálené objekty můžou sejít ve stejném shluku, když větší počet dalších objektů mezi nimi tvoří jakýsi most.



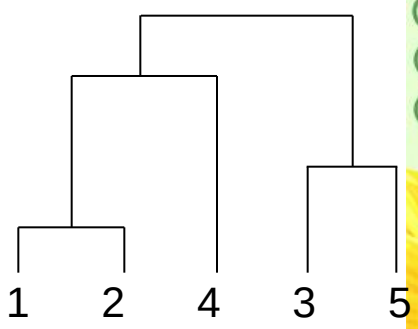
Hierarchické aglomerativní shlukování

Metoda nejvzdálenějšího souseda

(všespojná metoda, metoda úplné vazby, *complete linkage, the furthest neighbor method*)

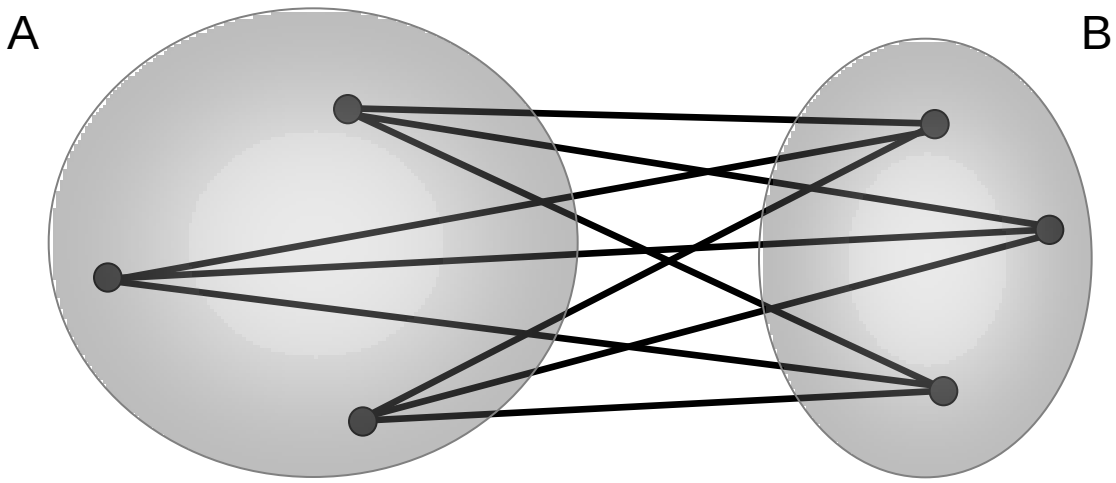


Vzdálenost mezi dvěma shluky je daná maximální vzdáleností mezi všemi možnými zástupci obou shluků.
Shluky jsou mezi sebou dobře oddělené.
Tendence ke tvorbě kompaktních shluků, ne ovšem velmi velkých.



Hierarchické aglomerativní shlukování

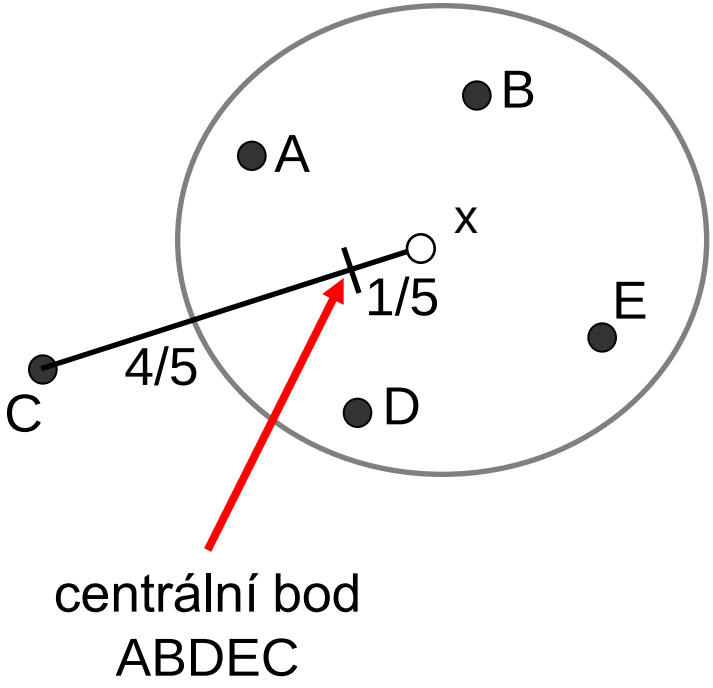
Metoda průměrné vzdálenosti (středospojná metoda, metoda průměrné vazby, *average linkage*, *UPGMA – unweighted pair-group method using arithmetic averages*)

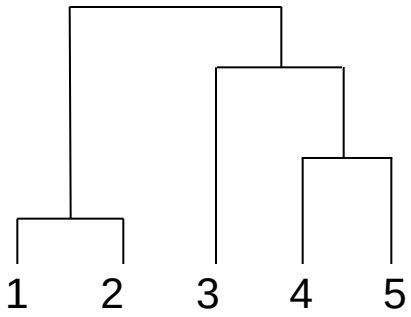


Meziskupinová (ne)podobnost je definována jako průměrná (ne)podobnost mezi všemi možnými páry členů.
Metoda vede často k podobným výsledkům jako metoda najvzdálenějšího souseda.

Hierarchické aglomerativní shlukování

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

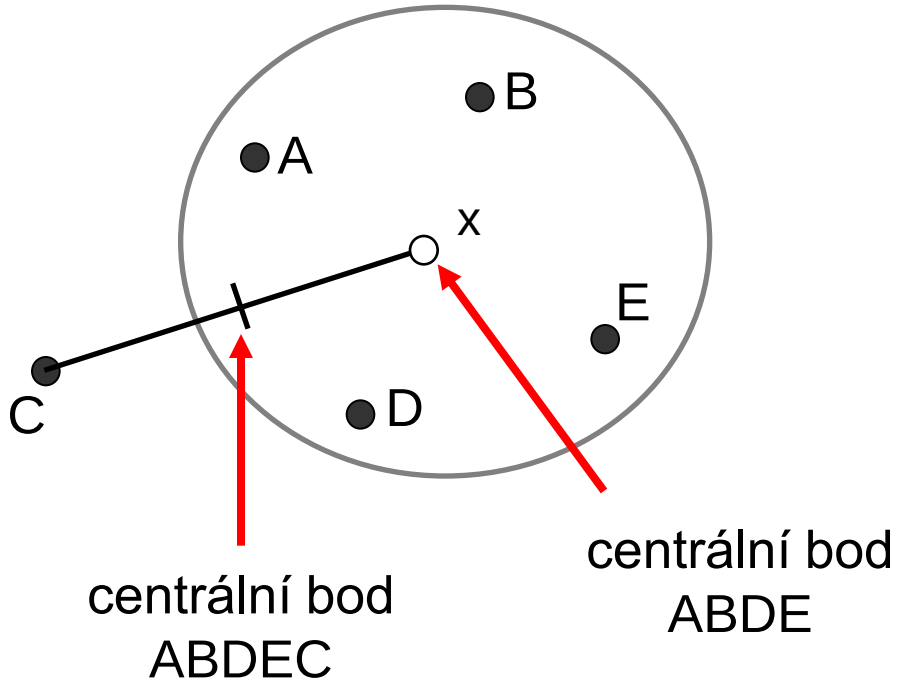




Tato metoda nevychází již z agregace informací o mezishlukových vzdálenostech objektů. Kritérium je euklidovská vzdálenost centroidů. Při této metodě je vzdálenost mezi shluky počítána jako vzdálenost mezi centroidy těchto shluků.

Hierarchické aglomerativní shlukování

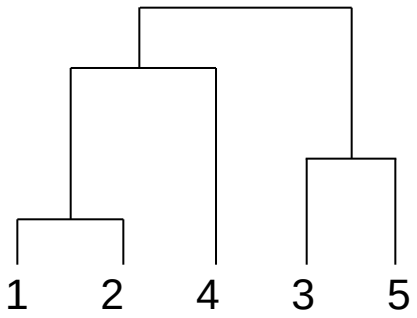
Mediánová metoda (*median method, WPGMC – weighted pair-group method using centroids, weighted centroid clustering*)



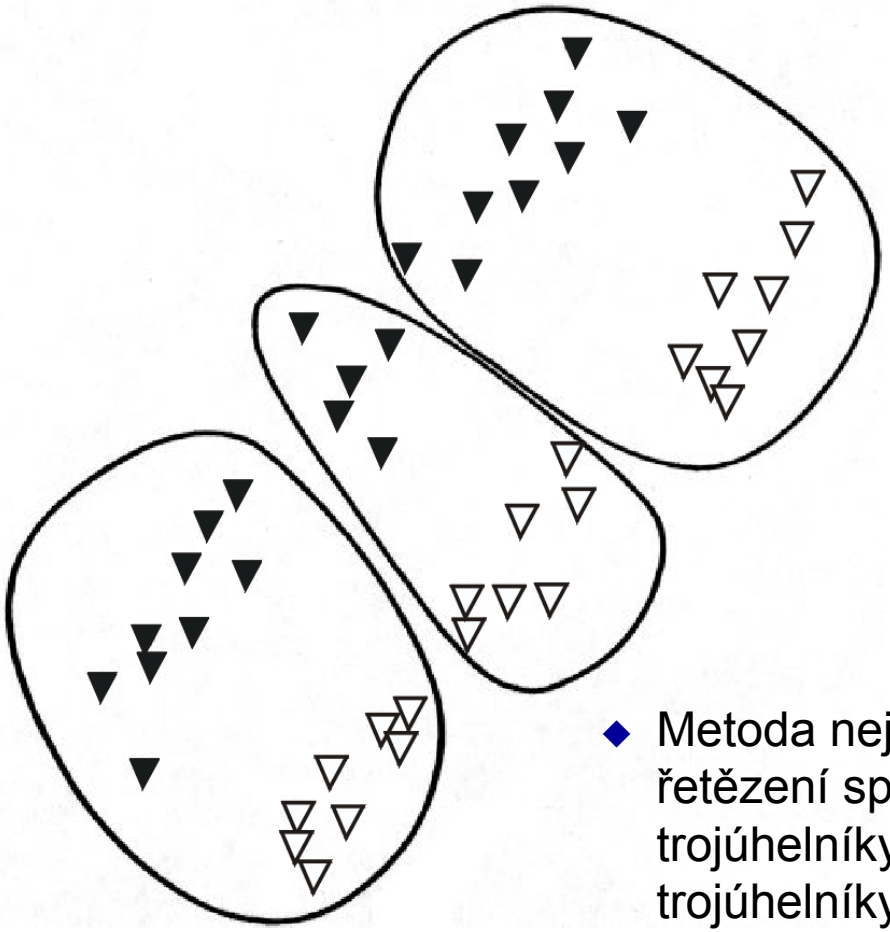
Hierarchické aglomerativní shlukování

Wardova metoda (*Minimum variance clustering*)

Wardova metoda je podobná středospojné a centroidové metodě. Kritérium pro spojování shluků je příspěvek celkového vnitroskupinového součtu čtverců odchylek pozorování od shlukového průměru. Příspěvek je vyjádřen jako součet čtverců v novo vznikajícím shluku, zmenšený o součty čtverců v obou zanikajících shlucích. Wardova metoda má tendenci odstraňovat malé shluky, teda tvořit shluky zhruba shodné velikosti.



Hierarchické aglomerativní shlukování



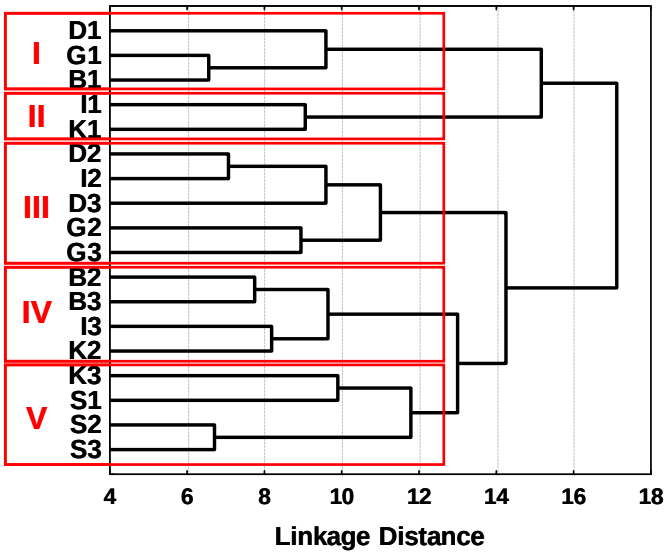
- Metoda nejbližšího souseda by v důsledku řetězení spojila do jednoho shluku plné trojúhelníky a do druhého prázdné trojúhelníky, zatím co Wardova metoda a metoda průměrné vzdálenosti by vytvořili skupiny ohraničené čárami (podle Everitt & Dunn 1983).

Hierarchické aglomerativní shlukování

Výsledkem hierarchického aglomeratívního shlukování je dendrogram (strom).

V tomto případě jsme použily:

- všespojnou shlukovací metodu
- míru vzdálenosti: Euklidovskou vzdálenost



Dendrogram znázorňuje podobnost společenstev korýšů šesti lokalit v záplavové oblasti Dunaje ve třech obdobích

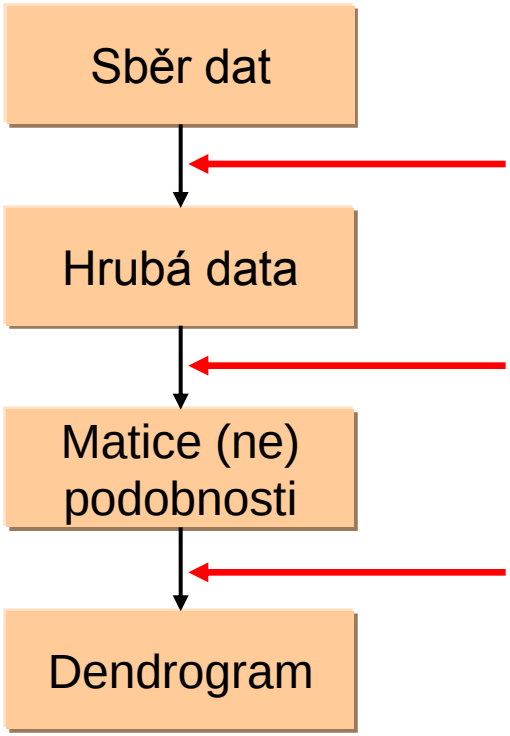
- 1: 1991-1992 před přehrazením Dunaje
- 2: 1993-1997 prvních 5 let po přehrazení
- 3: 1999-2004 dalších 6 let po přehrazení

Sledované lokality:

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

Hierarchické aglomerativní shlukování

Výsledek klasifikace je ovlivněn rozhodnutím na několika úrovních



důležitostná hodnota (pokryvnost, početnost)

transformace, standardizace, měření podobnosti

shlukovací algoritmus

Podle Kováře a Lepše (1986) mají transformace větší vliv na výsledek shlukování než metody shlukování.

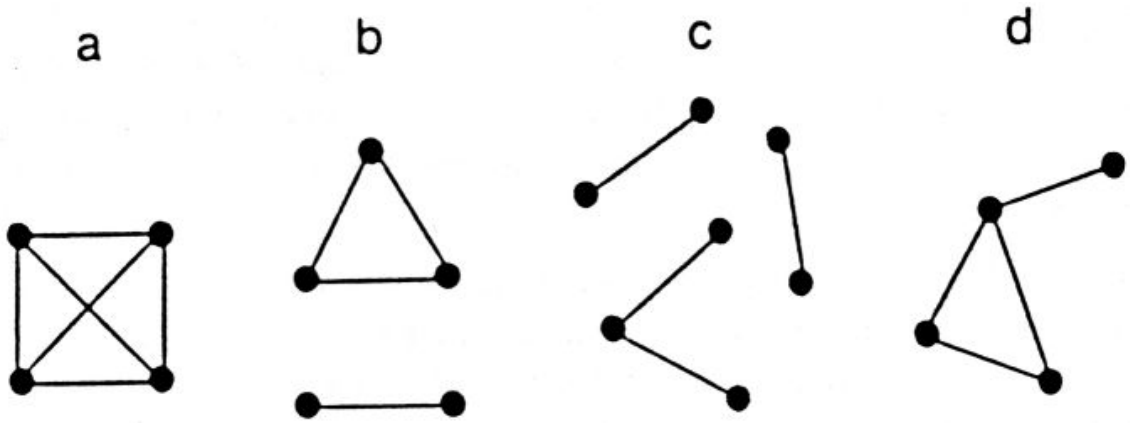
Kritické problémy analýzy

- Velké množství proměnných nebo objektů v dendrogramu je obtížné interpretovat
- Analýza je silně závislá na zvolení vhodné metriky vzdálenosti
- Analýza je silně závislá na shlukovacím algoritmu

Hierarchické aglomerativní shlukování

Shody (*ties*)

- Při použití aglomerativních shlukovacích metod může nastat situace, kdy se v matici podobnosti vyskytnou tzv. shody (*ties*)
- Nejčastěji dochází ke shodám při analýze binárních dat, je tu veliká pravděpodobnost stejné vzdálenosti mezi objekty
- Náhodné řešení takové situace může ovlivnit výslednou klasifikaci (dendrogram)



a – graf je úplný, b – graf je nesouvislý a všechny izolované komponenty jsou úplné , c – graf je nesouvislý a alespoň jedna komponenta není úplná, d – graf je souvislý, ale není úplný

Hierarchické aglomerativní shlukování

Řešení situací

a) spojí se všechny objekty naráz

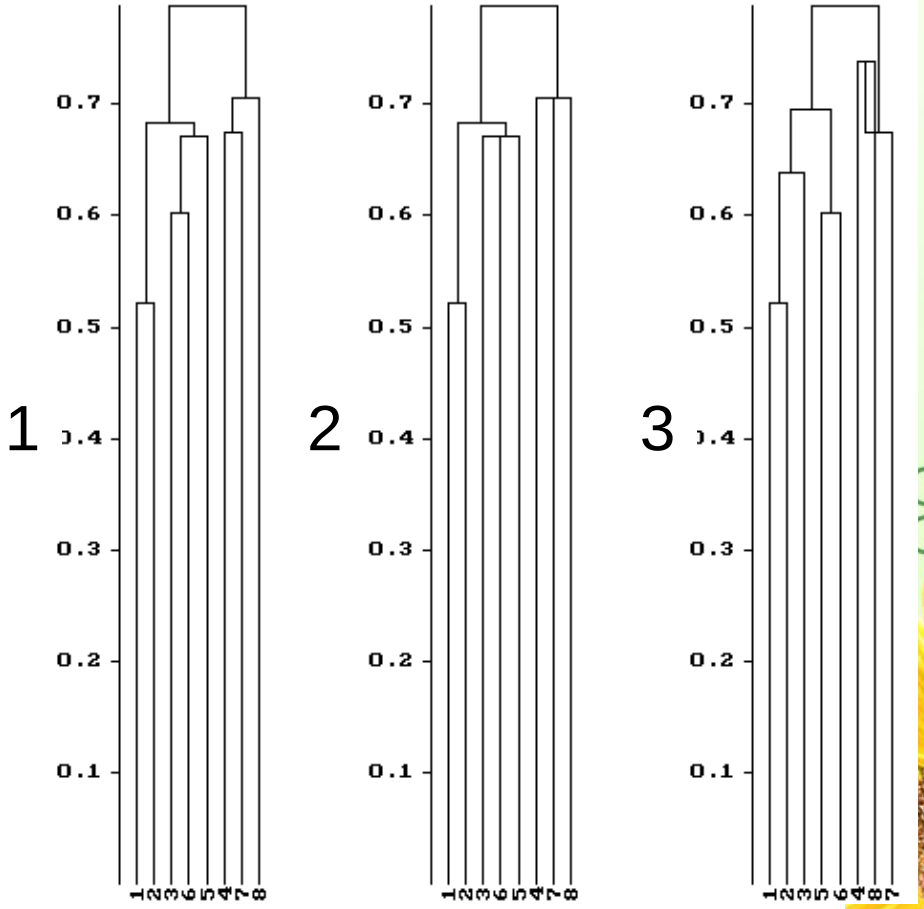
b) paralelně se vytvoří více skupin (tzv. *multiple fusion*)

c) a d) tři možnosti řešení:

1 „*silent mode (arbitrary)*“
Vazby se řeší náhodně, spojí se jenom poslední nalezená dvojice (je tu vliv pořadí objektů v primární matici)

2 „*single linkage*“
Všechny objekty, které jsou spojené vazbou, se spojí do jednoho shluku

3 „*suboptimal fusions*“
Nekompletní komponenty se ignorují a hledání nejmenších vzdáleností v matici pokračuje pokým se už žádné nekompletní komponenty nevyskytují



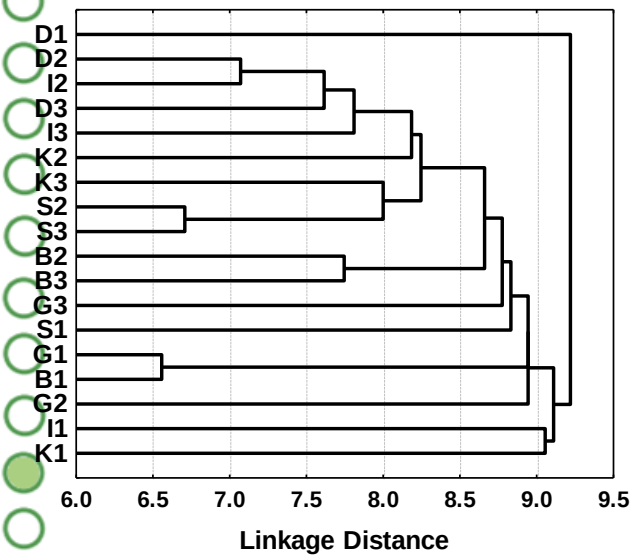
Hierarchické aglomerativní shlukování

Hierarchické metody — Aglomerativní shlukování

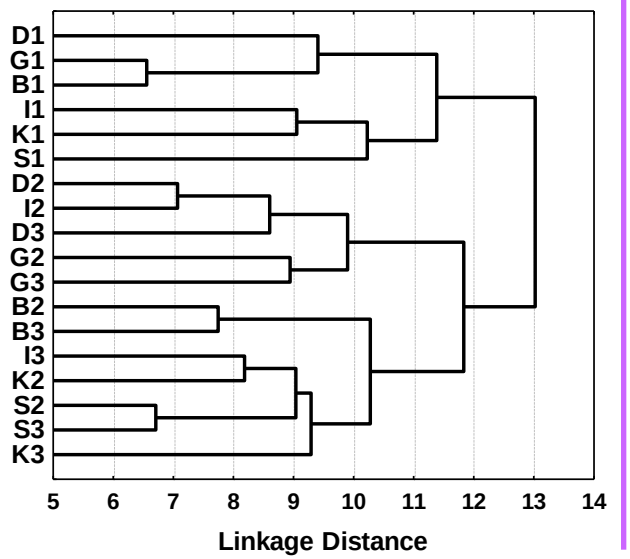
REÁLNA DATA

- 6 lokalit, každá lokalita monitorována ve 3 obdobích
- datová matice: 18 vzorek x 63 planktonních druhů korýšů; hodnoty = stupeň dominance

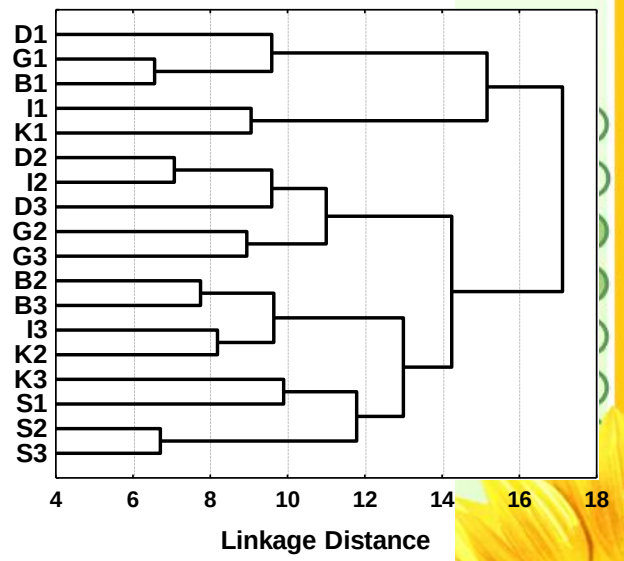
jednospojná metoda



středospojná metoda



všespojná metoda

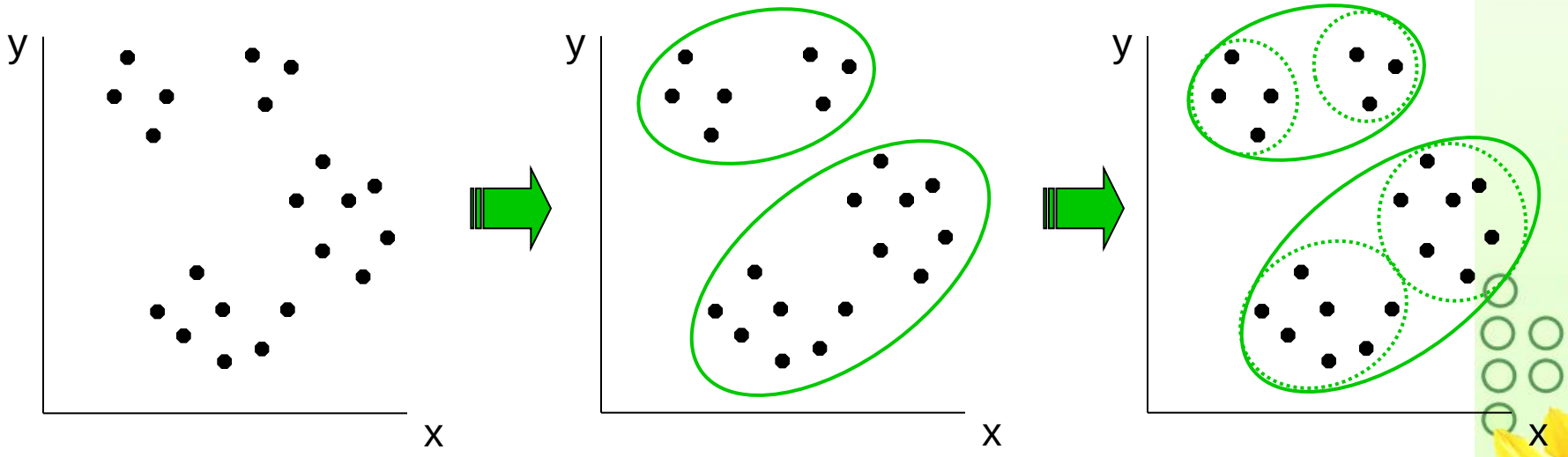


Dendrogramy vytvořeny pomocí tří různých shlukovacích algoritmů: jednospojná, středospojná a všespojná metoda.
V prvním případě je zjevné silné řetězení objektů.

Hierarchické divizivní shlukování

Hierarchické metody — Divizivní shlukování

- dělení probíhá „shora“; začíná všemi objekty jako s jednou skupinou
- rozdělení souboru na dvě části - podskupiny
- další dělení podskupin



Časté použití ke klasifikaci biologických společenstev

Hierarchické divizivní shlukování

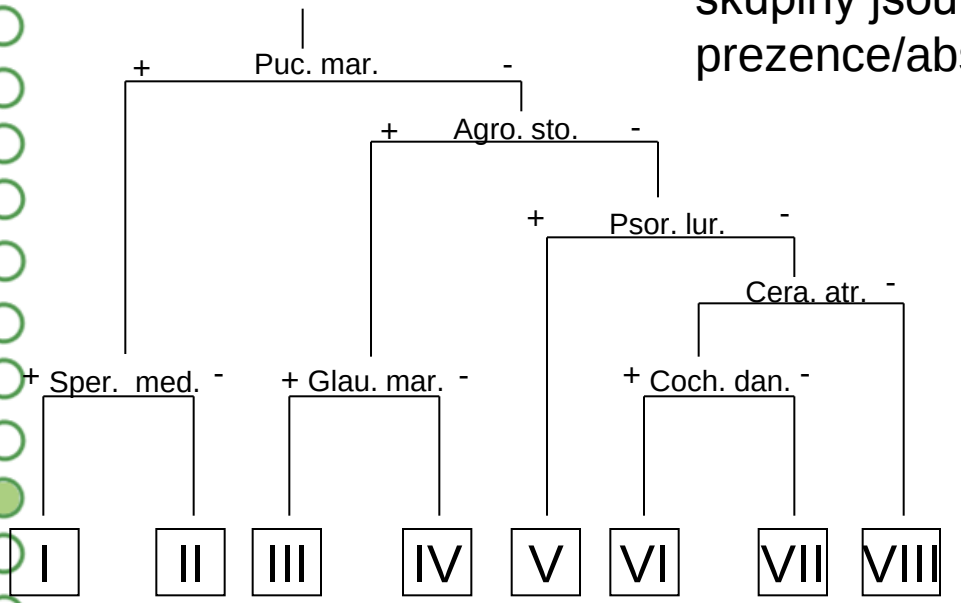
Hierarchické metody

Divizivní shlukování

Monotetická metoda

Asociační analýza

- dělení na základě jednoho parametru
- nejdříve je nalezen druh, který je nejvíce asociovaný s ostatními druhy; skupiny jsou rozděleny na základě prezence/absence tohoto druhu



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

Polytetická metoda

Two way indicator species analysis

TWINSPAN

Hierarchické divizivní shlukování

Polytetická metoda

Two way indicator species analysis

TWINSPAN

- dělení skupiny je založeno na všech druzích podle jejich skóre na první ose vytvořené ordinací (v TWINSPAN-e korespondenční analýza)
- dichotomie vzniká ordinací lokalit na základě diferenciálních druhů
- bere do úvahy aj abundanci druhů vo formě tzv. pseudo-druhů => potřeba určit hraniční hodnoty (cut levels)

Původní tabulka

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

cut levels
1, 5 a 20

Tabulka s pseudodruhy použitými v TWINSPAN

Species	A	B
Cirsoler1	0	1
Glechede1	1	0
Glechede2	1	0
Junctenu1	1	1
Junctenu2	1	1
Junctenu3	1	1
Junctenu4	0	1

Hierarchické divizivní shlukování

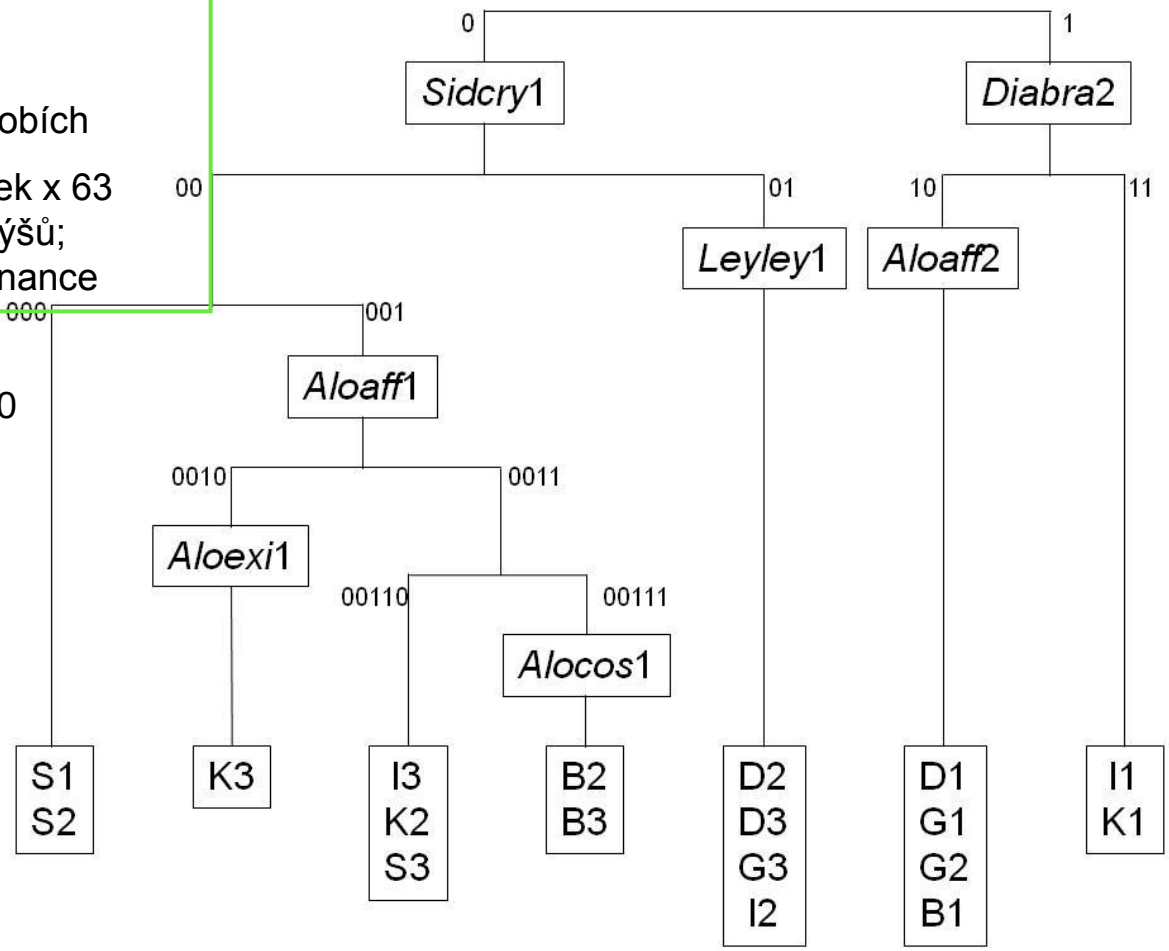
Two way indicator species analysis

TWINSPAN for Windows, WinTWINS,
http://www.canodraw.com/wintwins.htm

REÁLNA DATA

- 6 lokalit, každá lokalita monitorována ve 3 obdobích
- datová matice: 18 vzorek x 63 planktonních druhů korýšů; hodnoty = stupeň dominance

- Cut levels 0, 2, 5, 10, 20



Hierarchické divizivní shlukování

Hierarchické metody

Divizivní shlukování

- začíná se všemi objekty jako s jednou skupinou
- skupina je rozdělena na dvě menší skupiny, …

Monotetická metoda

Asociační analýza

+ poskytuje jednoduchý binární klíč, který sa dá použít ke klasifikaci dalšího vzorku

− jenom pro data prezence/absence

− vytvořené skupiny – méně homogenní jako skupiny vytvořeny polytetickou metodou

− konečni klasifikace – není robustní

Polytetická metoda

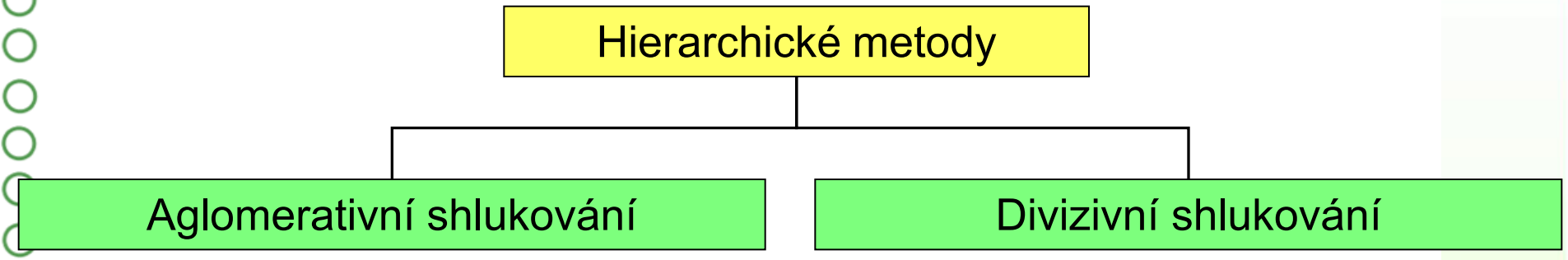
Two way indicator species analysis

+ vytvořené skupiny jsou více homogenní jako skupiny vytvořeny monotetickou metodou

− neposkytuje jednoduchý klíč vhodný pro zařazení nové vzorky do dané třídy (skupiny)

− předpokládá jenom jeden základní trend v datech

Hierarchické shlukování



+ Shlukování je intuitivní => je to nejpopulárnější klasifikační metoda

Výsledek je sumarizovaný v dendro-gramu – jednoduchá interpretace

− Neexistuje „správný“ shlukovací algoritmus
Výsledky se dramaticky mění s
- různým shlukovacím algoritmem
- různým indexem podobnosti

Aglomerativní shlukování není efektivní pro velmi velká data

+ jednoduchá interpretace výsledků

divizivní techniky jsou pro velmi objemná data vhodnější jako aglomerativní techniky

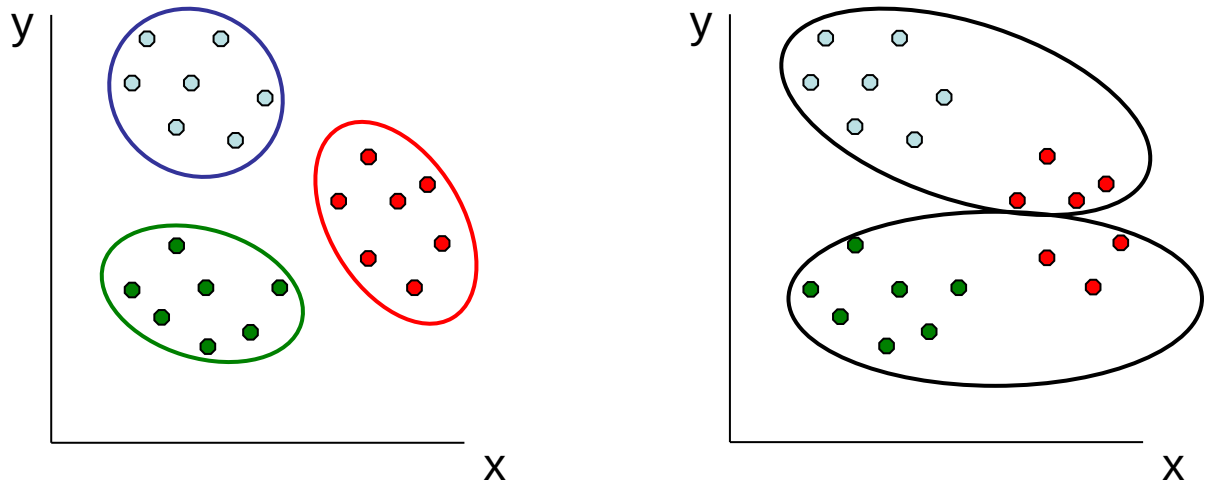
− monoteticka metoda není robustní

polytetická metoda neposkytuje jednoduchý klíč vhodný poe zařazení nového vzorku do dané skupiny

Nehierarchické shlukování

Nehierarchické metody

Objekty jsou na základě zadaného počtu shluků rozděleny podle kritéria maximální homogenity shluků.



Ukázka rozdělení objektů do shluků nehierarchickou metodou *k-průměrů*. Výsledek je ovlivněn volbou počtu shluků.

Vlevo: počet shluků k = 3 je dobrá volba; vpravo: počet shluků k = 2 je špatná volba.

Nehierarchické shlukování

Princip nehierarchického shlukování

- Pro výpočet se používá opakovaná relokační procedura. Začíná s *k* skupinami a pak přesouvá objekty tak, aby minimalizovala variabilitu uvnitř skupin a maximalizovala variabilitu mezi skupinami.
- Relokační procedura se ukončí, když žádný další přesun už kriteria nezlepší.
- Takto získáváme ovšem pouze lokální extrém, nemáme jistotu, že je taky globálním extrémem.
- Doporučuje se začít s různými počátečními skupinami a sledovat, zda jsou výsledky těchto analýz stejné.

Rizika analýzy

- při chybném odhadu počtu shluků dává metoda chybné výsledky
- výpočet je možný pouze na Euklidovských vzdálenostech se všemi jejími omezeními

Nehierarchické shlukování

Nehierarchické metody — metoda *k*-průměrů

- skupiny nejsou zahrnuty do nadskupin, ani neobsahují podskupiny
- rozděluje objekty do určitého počtu skupin
- metoda *k*-průměrů pracuje s euklidovskými vzdálenostmi

Nehierarchické metody můžou být vhodnější jako hierarchické techniky
- v případě většího objemu dat
- v případě, že v datech neexistuje hierarchická struktura

počet skupin *k* je třeba specifikovat předem uživatelem

metoda *k*-průměrů pracuje s euklidovskými vzdálenostmi
=> to může být problémem v případě, když euklidovská vzdálenost není „nejlepší“ metrikou

Shluková analýza – souhrn

Shluková analýza obecně

Když data nemají úplně jednoznačnou a zřetelnou strukturu (jedná se spíše o náhodně rozptýleny objekty), je pravděpodobné, že použití různých shlukovacích technik přinese odlišné výsledky.

Když různé shlukovací techniky dávají ze stejného datového souboru shodné, reps. podobné výsledky, je to do jisté míry potvrzení struktury obsáhlé v datech (ačkoliv shlukovací metody patří k postupům produkujícím hypotézy a nejsou určeny k jejich testování)

Mnohé shlukovací techniky jsou citlivé na přítomnost odlehlých objektů (*outliers*, výrazně atypické případy). Před samotnou shlukovou analýzou je vhodné použít některou z metod na jejich odhalení, např. PCA. Výrazně odlehlé objekty zpravidla z ďalších analýz vyloučíme.

Shlukové analýzy obecně nejsou vhodná na data, ktorá popisujú variabilitu znaku závislém na gradientu prostředí.

Shluková analýza souhrn

Vstup shlukové analýzy:

- Matice podobnosti anebo vzdálenosti objektů
- Tabulka objektů charakterizovaných několika parametry

Výstup shlukové analýzy:

- Strom (dendrogram) při hierarchické shlukové analýze
- Zařazení objektů do předem definovaného počtu shluků při nehierarchické analýze

Při použití shlukové analýzy je nutné pamatovat na omezení:

- Aglomerativní shlukování není efektivní pro velmi velká data
- Při hierarchické aglomerativní analýze je výsledek silně ovlivněn výběrem indexu podobnosti, resp. metrikou vzdálenosti a shlukovacím algoritmem
- *! neexistuje správný shlukovací algoritmus !!!*
- Při hierarchické divizivní analýze: Twinspan předpokládá jeden hlavní trend v datech a je ovlivněn nastavením hranic pseudo-druhů
- Při nehierarchickém shlukování je nutné určit počet skupin předem

Ordinace dat biodiverzity a definice environmentálního gradientu

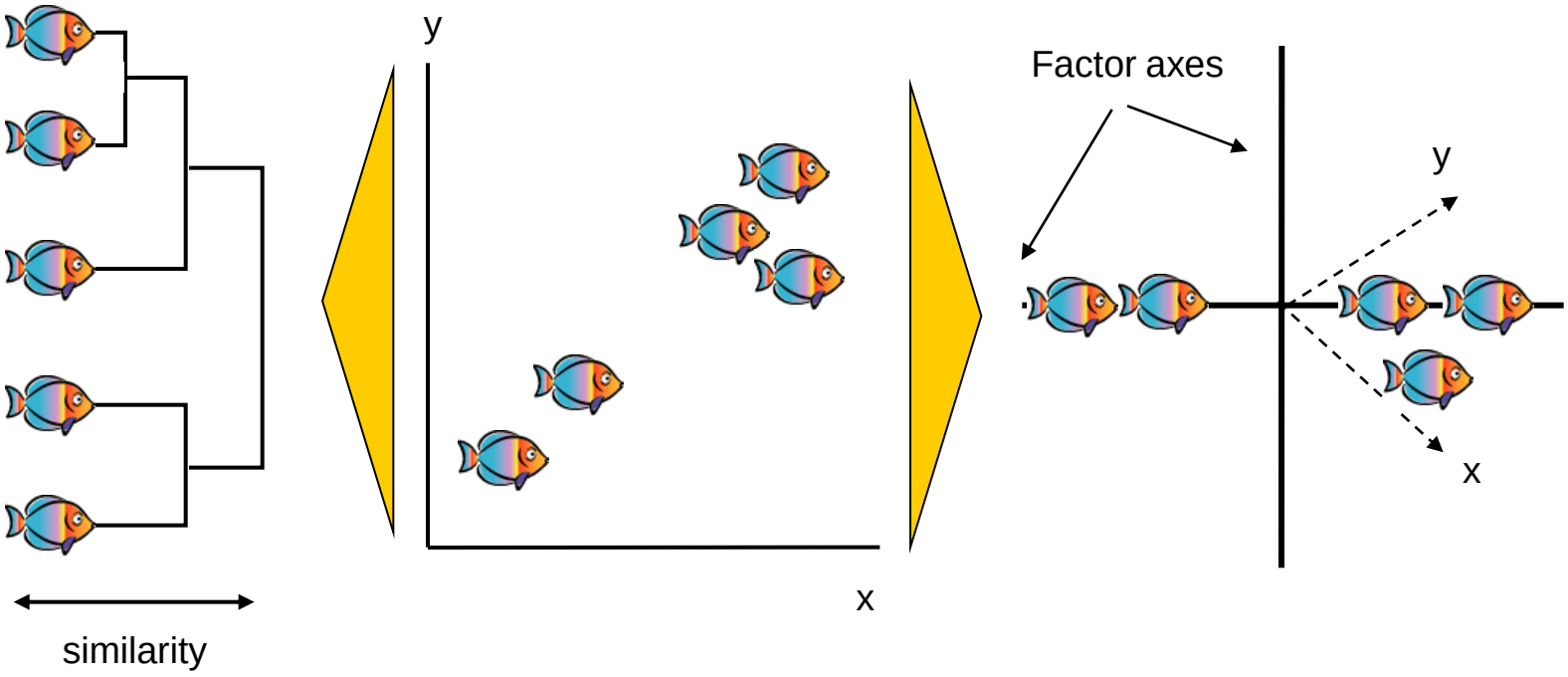
Základní typy vícerozměrných analýz

Shluková analýza

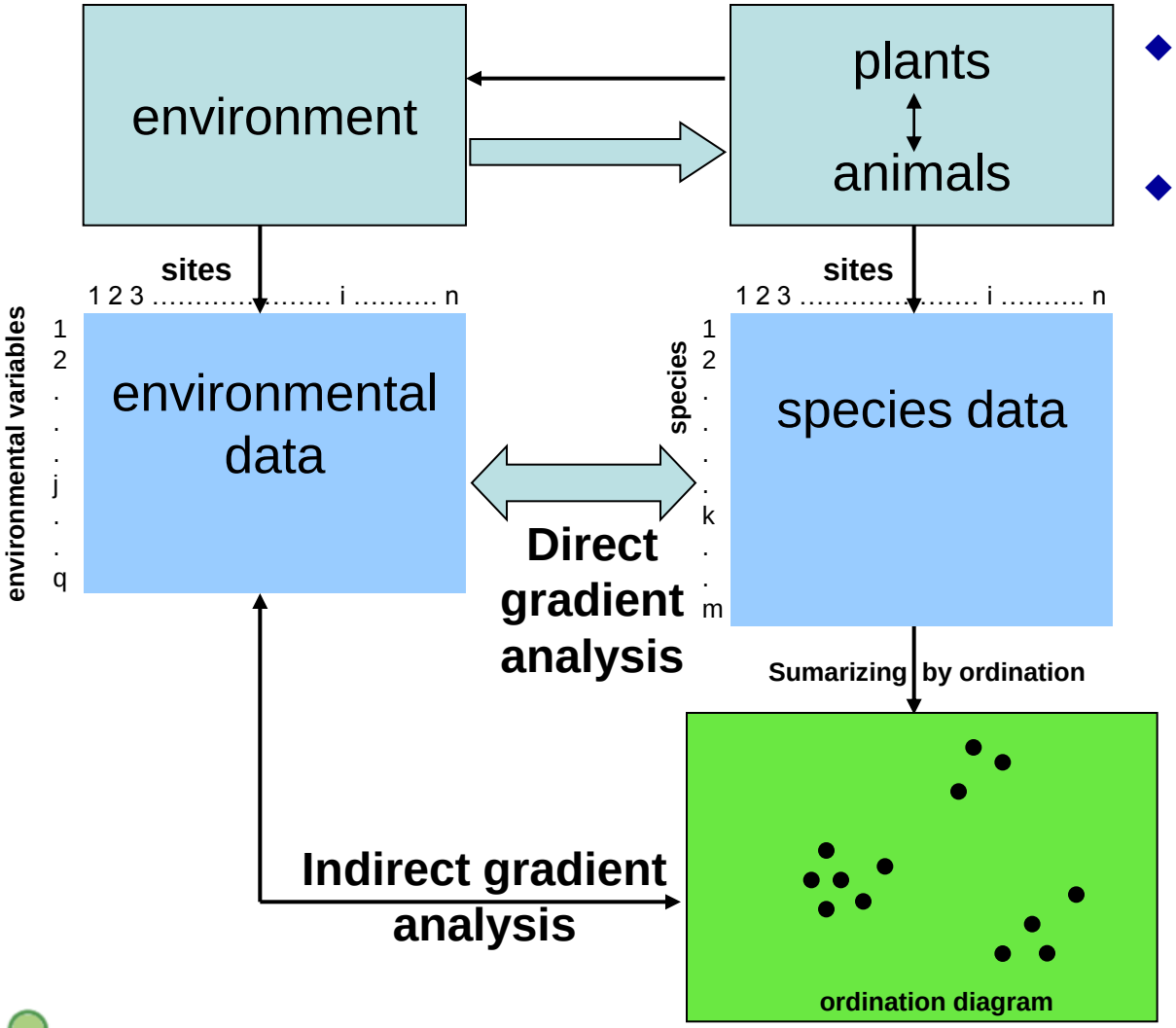
- Klasifikuje vzorky (lokality), druhy nebo proměnné
- Nachází skupiny v datech

Ordinace

- Uspořádá vzorky podél trendu v datech



Ordinační metody a data diverzity

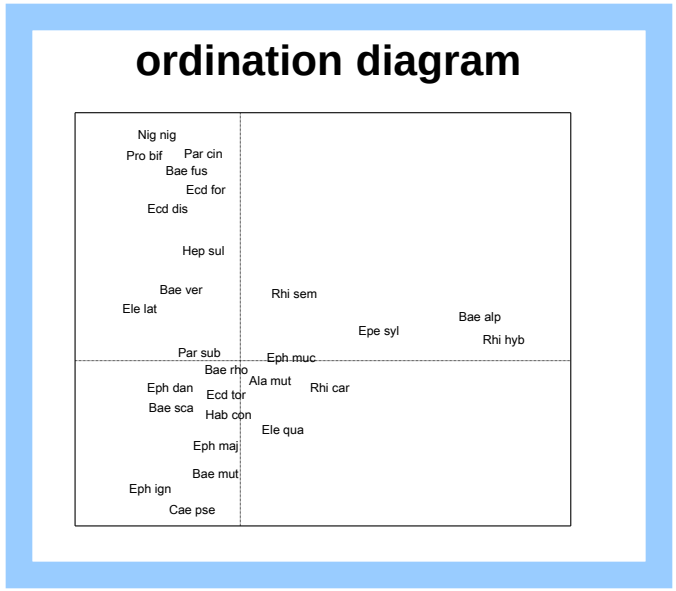


- Seřadí objekty podél environmentálního gradientu
- Cílem ordinace je sformulovat hypotézy o vztahu mezi druhovým složením společenstva na lokalitách a základními environmentálními faktory

- Ordinační metody nepředpokládají žádné apriorní seskupení objektů.
- Ordinační metody používáme zejména ke tvorbě hypotéz.

Ordinace

- Ordinace a shluková analýza jsou jediné možné techniky, které můžeme použít bez naměřených environmentálních dat
- Vysvětlující (explanatory) proměnné v ordinaci jsou teoretické proměnné = environmentální gradienty

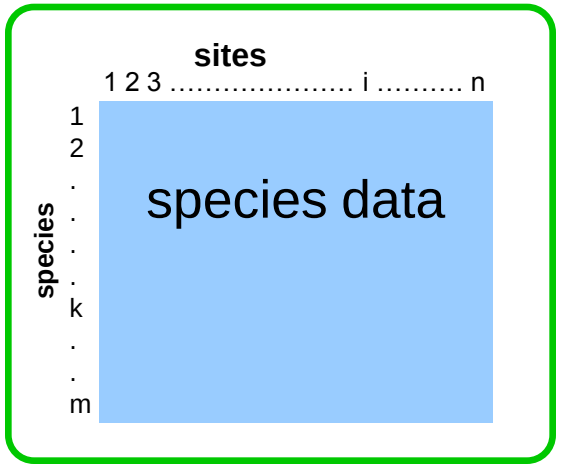


Každý vzorek zahrnuje hodnoty mnoho druhů.

vysvětlované = závislé proměnné
druhová data

Ordinační analýza: typy dat

Biodiverzitní data:



kvantitativní data
- počet jedinců jednotlivých druhů
- procentická pokryvnost
- odhad biomasy

semikvantitativní data
- Braun-Blanquetová stupnice

kvalitativní data
- přítomnost / nepřítomnost

Ordinační metody, gradientová analýza

- Termín **gradientová analýza** používáme pro metody, které dávají do vztahu druhová data a gradienty prostředí (měřeny nebo hypotetické).
- Gradientová analýza se zabývá vztahem složení společenstva k (známým nebo neznámým) gradientům prostředí.

Nepřímá gradientová analýza

(*indirect gradient analysis*)

- Osi vytvořeny na základě druhových dat

Přímá gradientová analýza

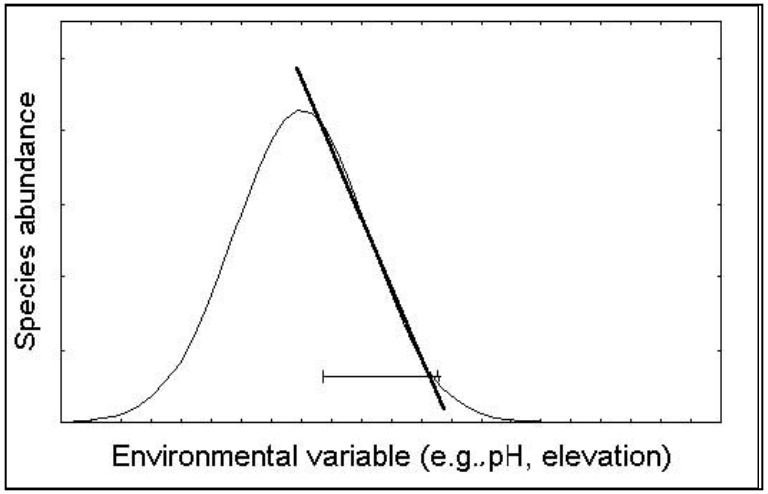
(*direct gradient analysis*)

- Ordinace kombinovaná s regresí – ordinační osy jsou omezeny (*constrained*) nebo kanonické (*canonical*) – jsou lineárně závislé na měřených vysvětlujících proměnných.

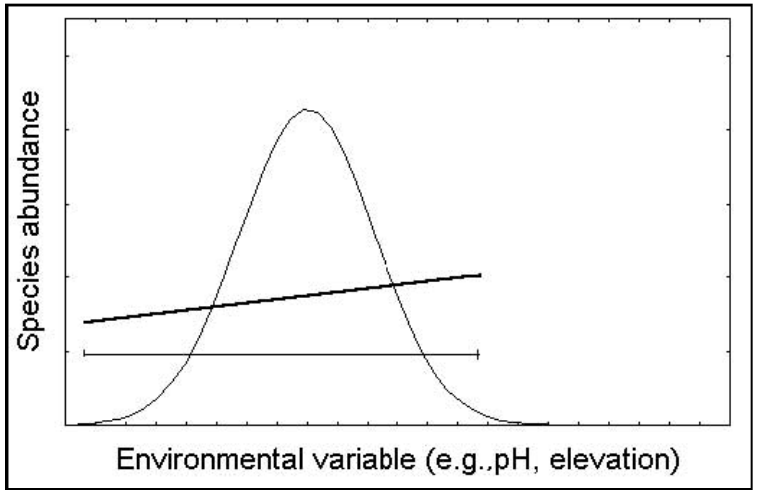
Odpověď druhů na gradient prostředí

Dva typy modelu odpovědi druhu na (známý nebo teoretický) gradient

- **lineární** (*linear*) – nejjednodušší odhad (na krátkém gradientu dobře funguje lineární aproximace jakékoliv funkce)
- **unimodální** (*unimodal*) – druh má na gradientu své optimum (na dlouhém gradientu není aproximace lineární funkcí vhodná)

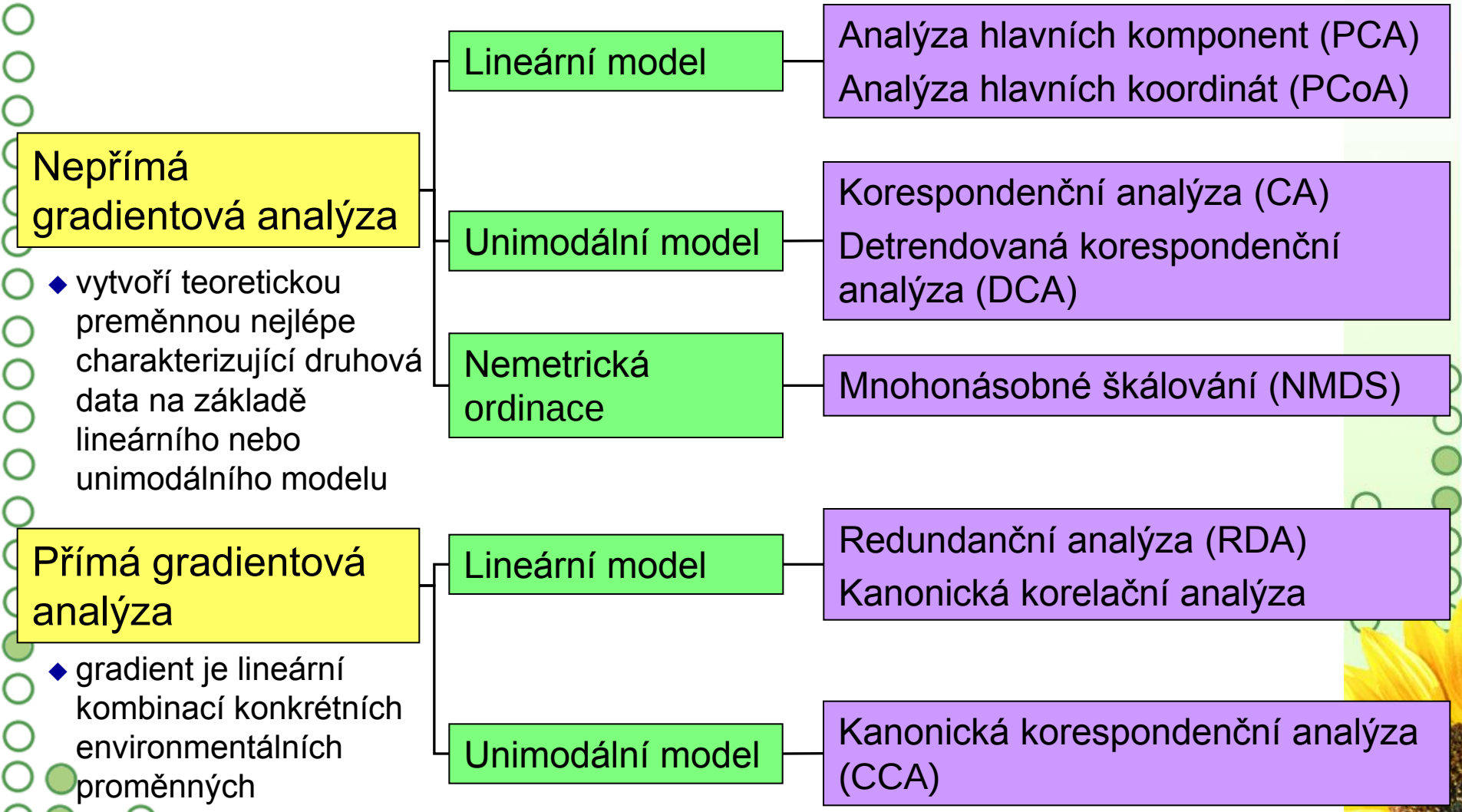


Lineární aproximace unimodální odpovědi na krátké části gradientu



Lineární aproximace unimodální odpovědi na dlouhé části gradientu

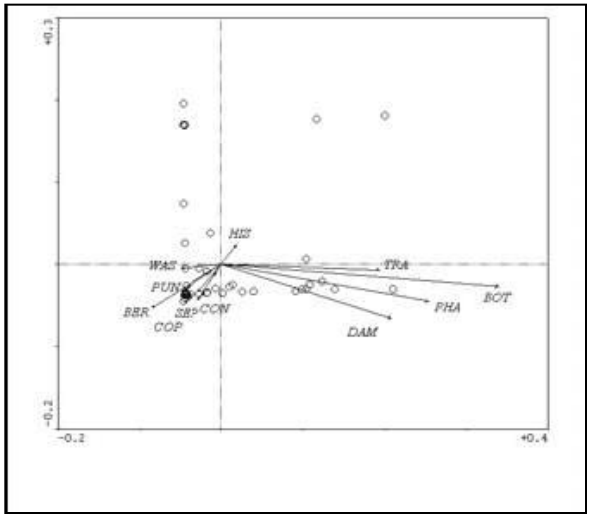
Základní techniky ordinačných metod



Příklady ordinačních diagramů

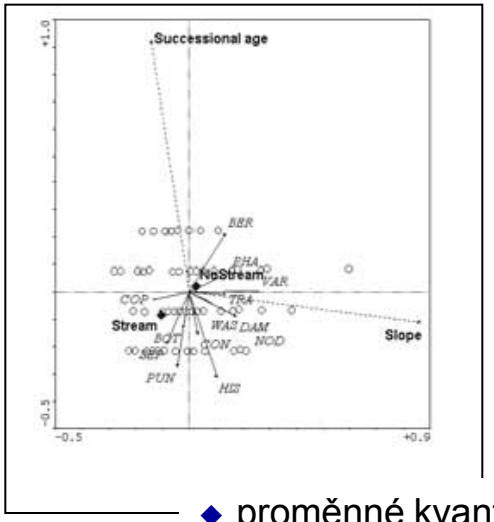
Výsledky ordinací se obvykle prezentují jako **ordinační diagramy**.

PCA



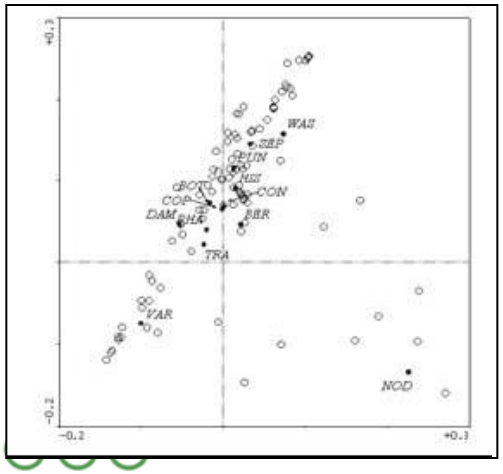
- vzorky: body
- druhy: šipky

RDA



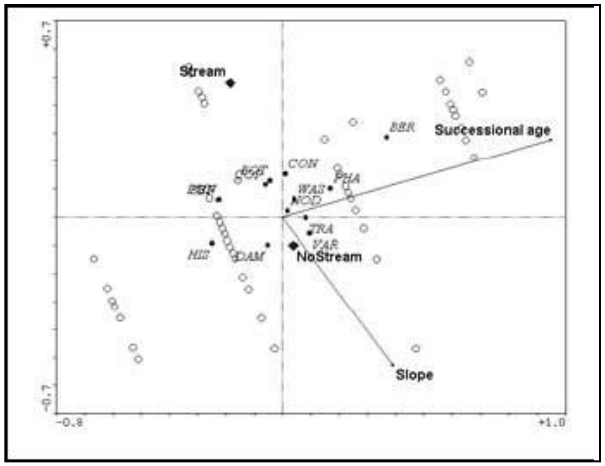
- proměnné kvantitativní: šipky
- proměnné kvalitativní: body

CA



- vzorky: body
- druhy: body

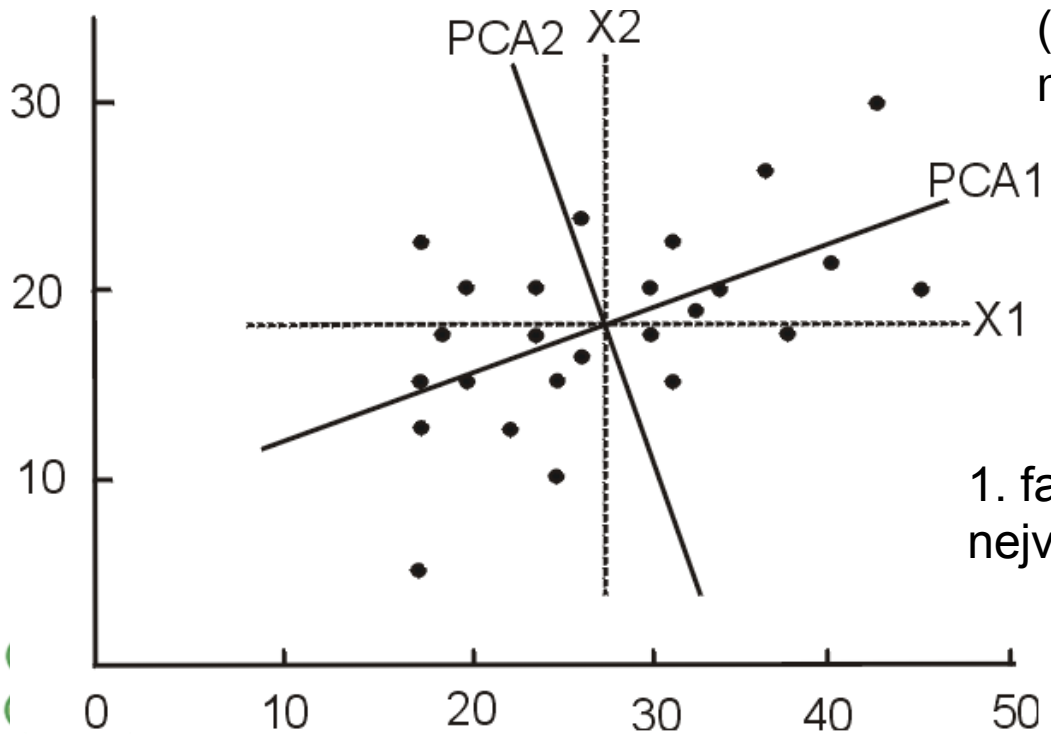
CCA



Analýza hlavních komponent

Analýza hlavních komponent (PCA)

- Proměnné jsou vzájemně korelované, tedy část informace v souboru je duplicitní
- Analýza odstraní duplicitu z dat a zobrazí pouze unikátní informaci – tj. nahradí původní soubor proměnných souborem nových proměnných vzájemně nekorelovaných.



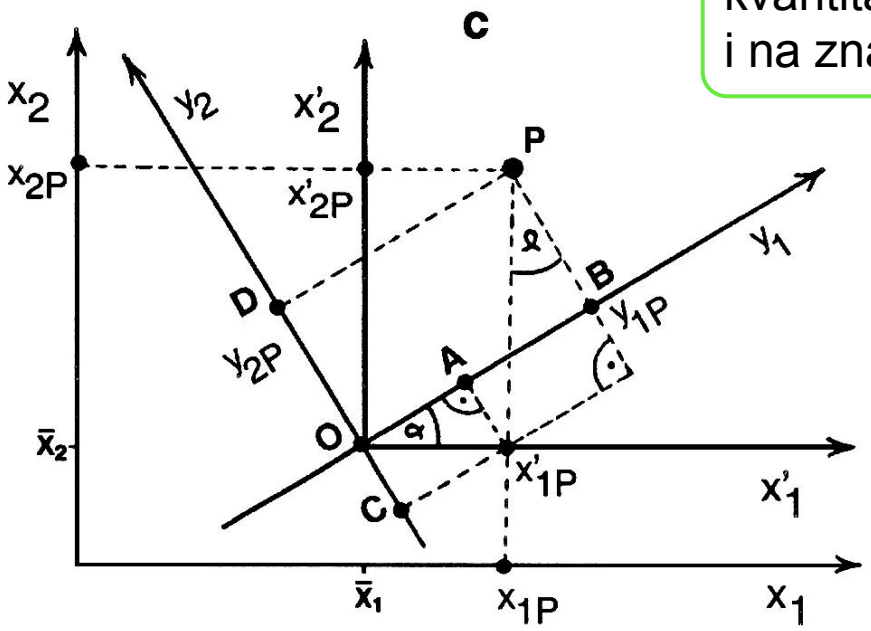
Je založena na **vlastní analýze** (eigenanalysis) symetrických matic (**korelační, kovarianční**)

1. faktorová osa vyčerpá nejvíc celkové variability

Analýza hlavních komponent (PCA)

Cíl PCA: určení uhlů mezi původními a novými osami souřadnicové soustavy, souřadnice objektů v novém systému souřadnic.

Původně byla PCA navrhnuta pro kvantitativní znaky, může sa ovšem použít i na znaky binární a semikvantitativní.



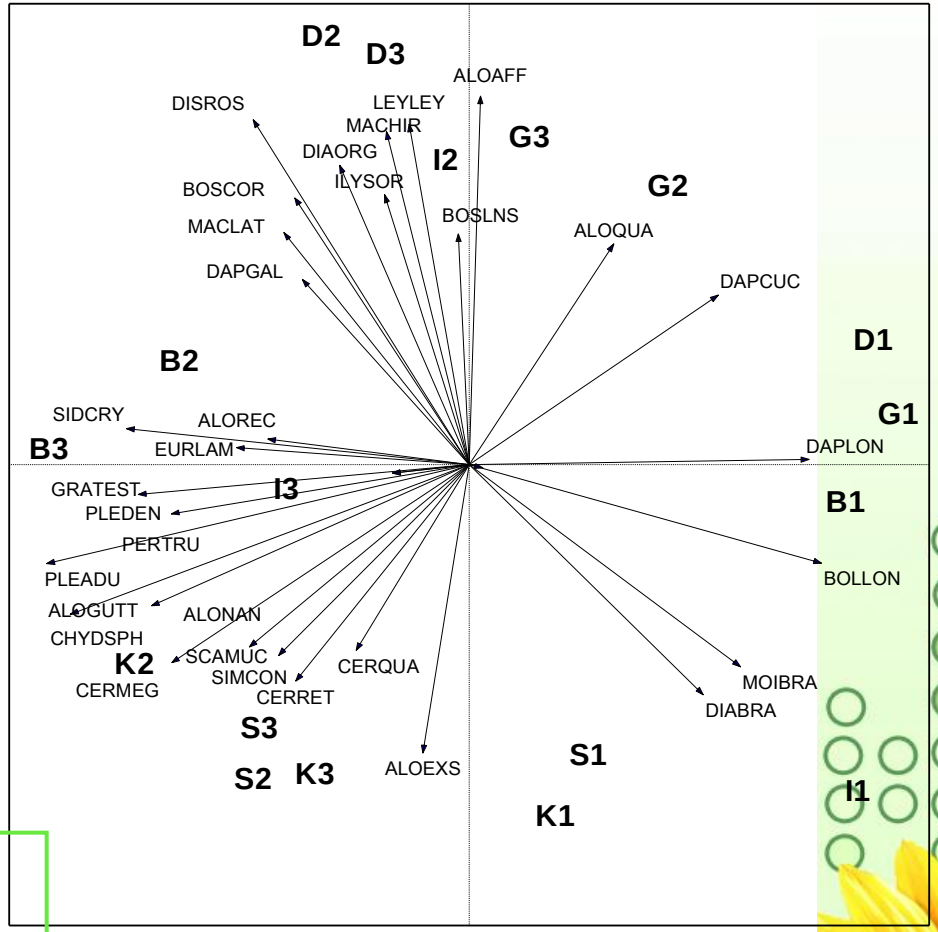
Vlastní čísla matice $\lambda_1, \lambda_2, \ldots \lambda_p$ jsou interpretovatelné jako míry rozptylu zachycené komponenty $y_1, \ldots, y_p$.

Analýza hlavních komponent (PCA)

Indirect gradient analysis

Principal component analysis

- PCA je postavena na lineárním modelu; abundance každého druhu roste ve směru šipky
- PCA je definováná pro kovarianční a pro korelační matici
- PCA není vhodna pro datovou matici s hodně nulami



REÁLNA DATA

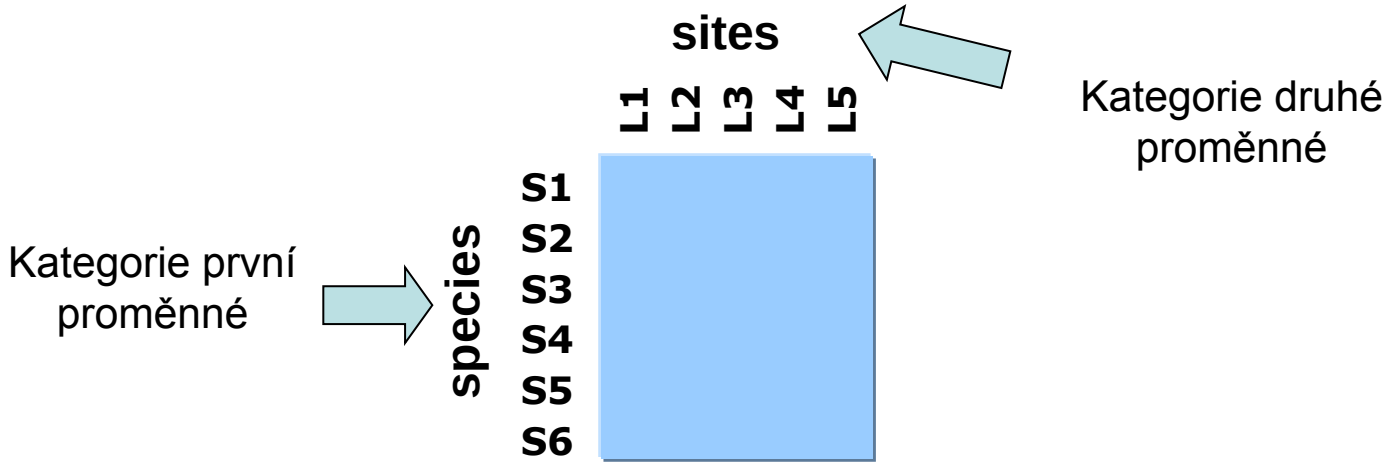
- 6 lokalit, každá lokalita sledována ve 3 obdobích
- datová matice: 18 vzorek x 63 plankt. dr. korýšů; hodnoty = stupeň dominance

Korespondenční analýza, Detrendovaná korešpondenční analýza

Korespondenční analýza

Korespondenční analýza – nástroj pro analýzu vztahů mezi řádky a sloupci kontingenční tabulky => dvě kategoriální proměnné.

Abundance taxonů (nebo počet jakýchkoliv objektů) na lokalitách lze brát jako kontingenční tabulku a mírou vztahu mezi řádky (lokality) a sloupci (taxony) je velikost chi-kvadrátu



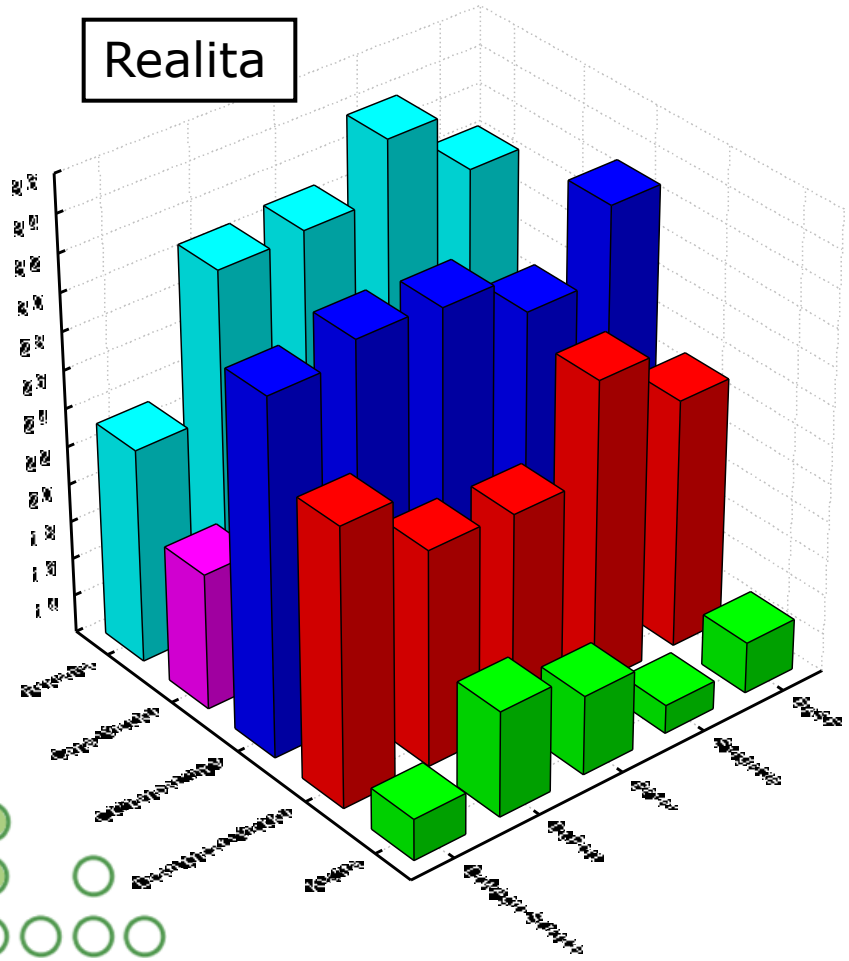
Korespondenční analýza a data biodiverzity

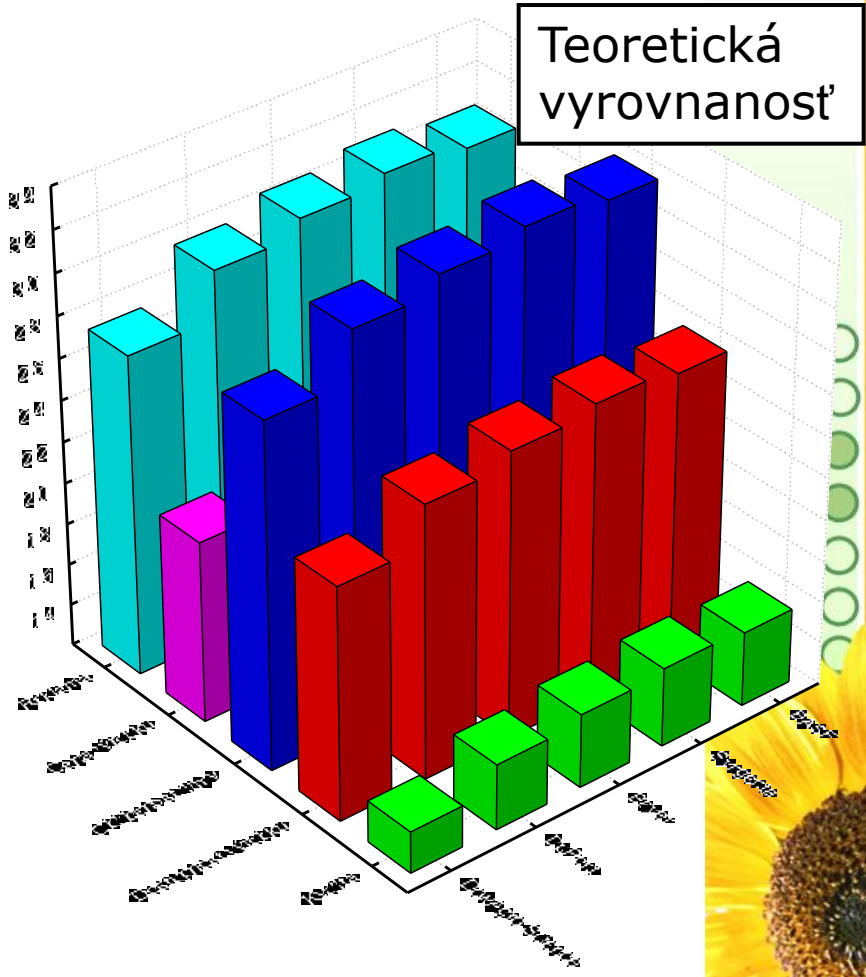
- Nepřímá gradientová analýza
- Založená na unimodální odpovědi – odhaduje optimum druhu na teoretickém gradientu

Korespondenční analýza

Princip

Korešpondenční analýza hledá, které kombinace řádků a sloupců hodnocené tabulky nejvíce přispívají k její variabilitě





Korespondenční analýza

Korespondenční analýza obecně

- Základní myšlenkou metody korespondenční analýzy je odvodit indexy (osy), které budou kvantifikovat vztahy mezi řádkovými a sloupcovými kategoriemi. Z těchto indexů můžeme odvodit, která sloupcová kategorie má větší či menší váhu v daném řádku a opačně.
- V grafu interpretujeme relativní pozice bodů řádků a sloupců jako váhy přislouchající danému sloupci a řádku.

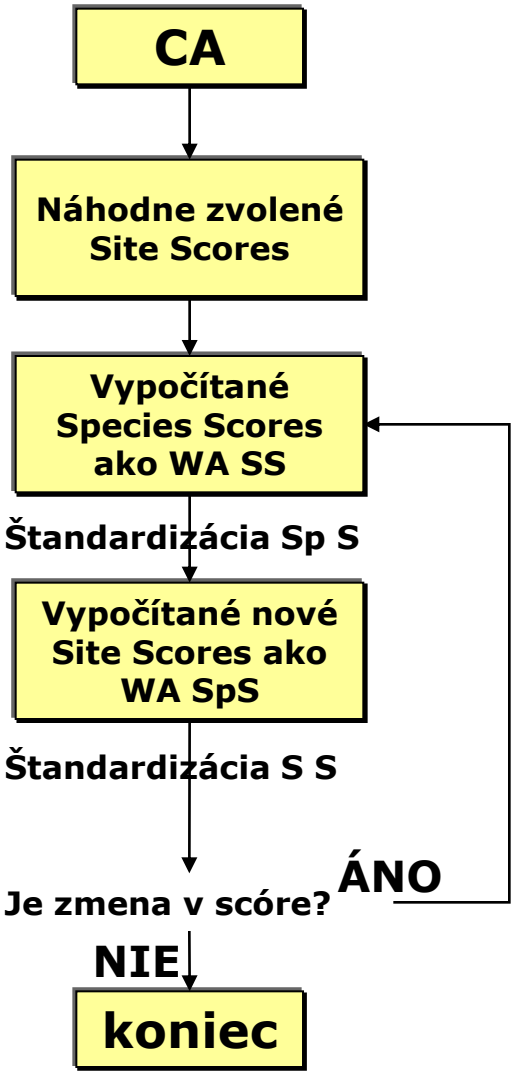
Korespondenční analýza a data diverzity

- Nejjednodušší cestou jak odhadnou optimum druhu pro unimodální model je spočítat vážený průměr těch hodnot charakteristik prostředí, při kterých se druh vyskytuje.
- Jako váha při výpočtu se používá početnost či jiná důležitostní hodnota druhu.
- Při váženém průměrování je implicitně zahrnuta standardizace po vzorcích i po druzích.

Korespondenční analýza

Korespondenční analýza, Correspondence analysis (CA)
Reciproční průměřování (reciprocal averaging)
nebo vlastní analýza (eigenanalysis)

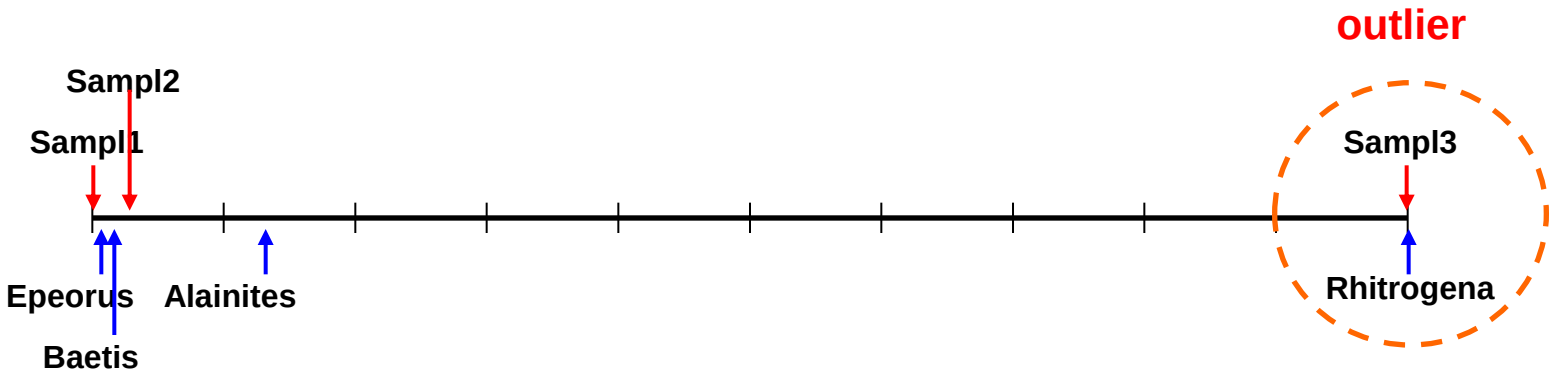
	Site1	Site2	Site3	WA1	WA2	WA3	WA4
Rhitrogena	**0**	**0**	**3**	13.000	10.000	10.000	10.000
Alainites	**5**	**2**	**1**	4.625	1.363	1.312	1.310
Baetis	**6**	**2**	**0**	3.250	0.113	0.062	0.060
Epeorus	**8**	**1**	**0**	2.556	0.050	0.028	0.027
initial value	*2*	*7*	*13*				
WA1	3.319	3.661	10.906				
WA1resc.	0.000	0.450	10.000				
WA2	0.415	0.600	7.841				
WA2resc.	0.000	0.249	10.000				
WA3	0.377	0.555	7.828				
WA3resc.	0.000	0.240	10.000				
WA4	0.375	0.553	7.827				
WA4resc.	0.000	0.239	10.000				



Korespondenční analýza

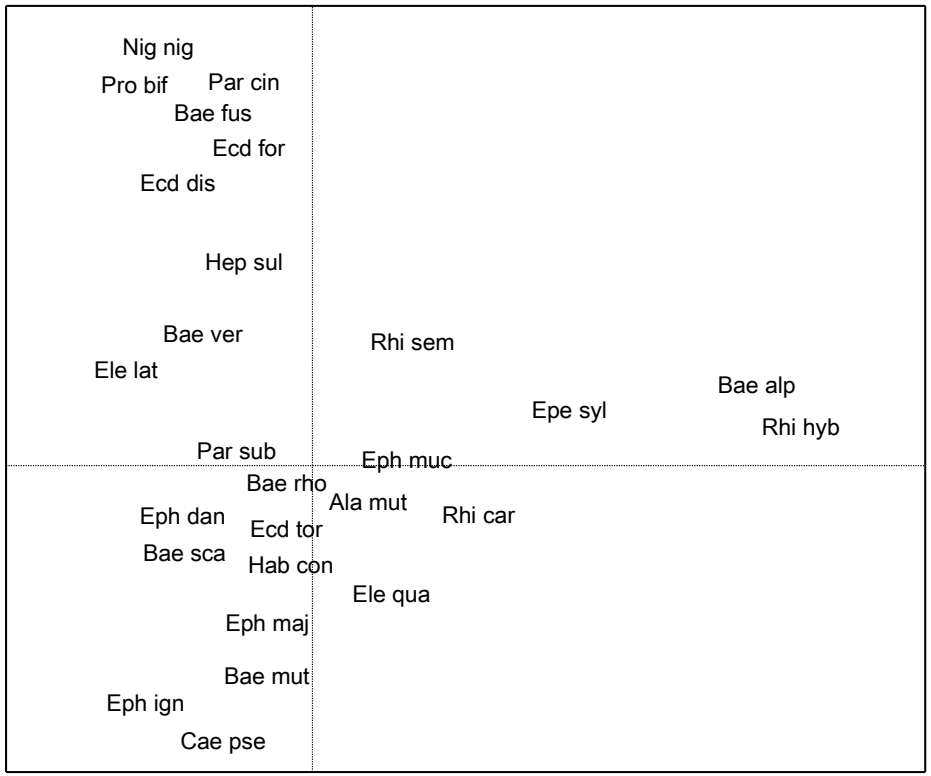
Korespondenční analýza, Correspondence analysis (CA)
Reciproční průměřování (reciprocal averaging) nebo vlastní analýza (eigenanalysis)

	Site1	Site2	Site3	WA4
Rhitrogena	**0**	**0**	**3**	10.000
Alainites	**5**	**2**	**1**	1.310
Baetis	**6**	**2**	**0**	0.060
Epeorus	**8**	**1**	**0**	0.027
WA4resc.	0.000	0.239	10.000	



Korespondenční analýza: výsledky

- **Ordinační diagram:** ordinační osy jsou ortogonální, tj. na sobě lineárně nezávislé
- Skóre druhů a vzorků (řádky a sloupce původní kontingenční tabulky)
- Vlastní hodnoty, vlastní vektory (eigenvalues, eigenvectors)



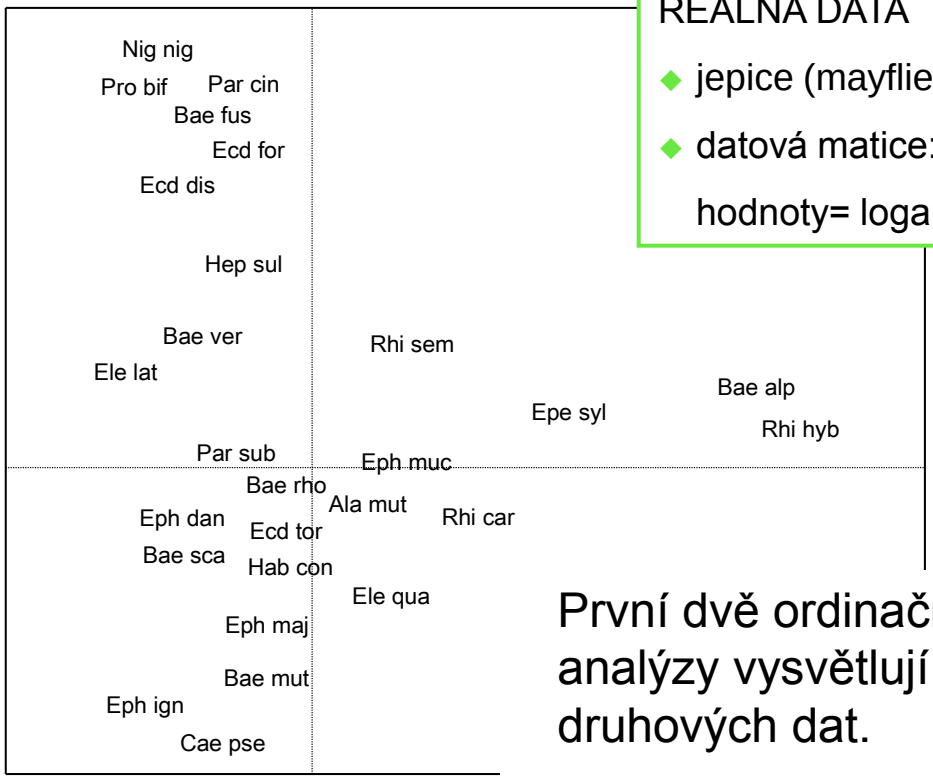
Vysoké skóre: druh s nízkou frekvenci

Vlastní hodnota (eigenvalue) představuje informaci vysvětlenou danou osou.

Většinou interpretujeme pouze 2-3 **ordinační osy**.

Korespondenční analýza (CA)

- CA počítá s unimodální odpovědí druhů na gradient prostředí; každý druh se vyskytuje v určitém rozpětí hodnot hypotetického gradientu
- CA se doporučuje pro data obsahující hodně nul



REÁLNA DATA

- jepice (mayflies)
- datová matice: 30 lokalit x 40 druhů

  hodnoty= logaratimicky transformované abundance

První dvě ordinační osy korespondenční analýzy vysvětlují 24.7% variability druhových dat.

V diagramu jsou znázorněny pouze druhy s nejlepším fitem, lokality nejsou znázorněny.

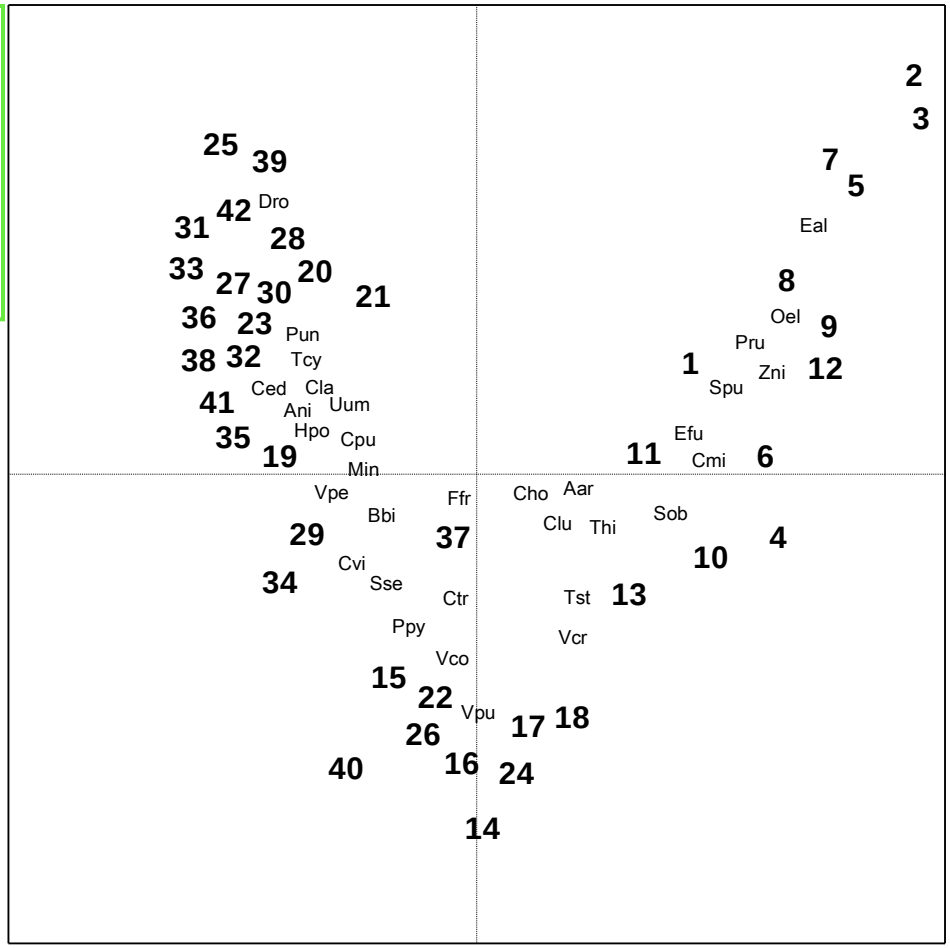
Korespondenční analýza: „arch effect“

- CA počítá s unimodální odpovědí druhů na gradient prostředí
- Silná unimodální odpověď může vést k tzv. podkovitému efektu „arch effect“ v ordinačním diagramu; jde o artefakt metody
- Detrendovaná forma CA odstraňuje „arch effect“

REÁLNA DATA

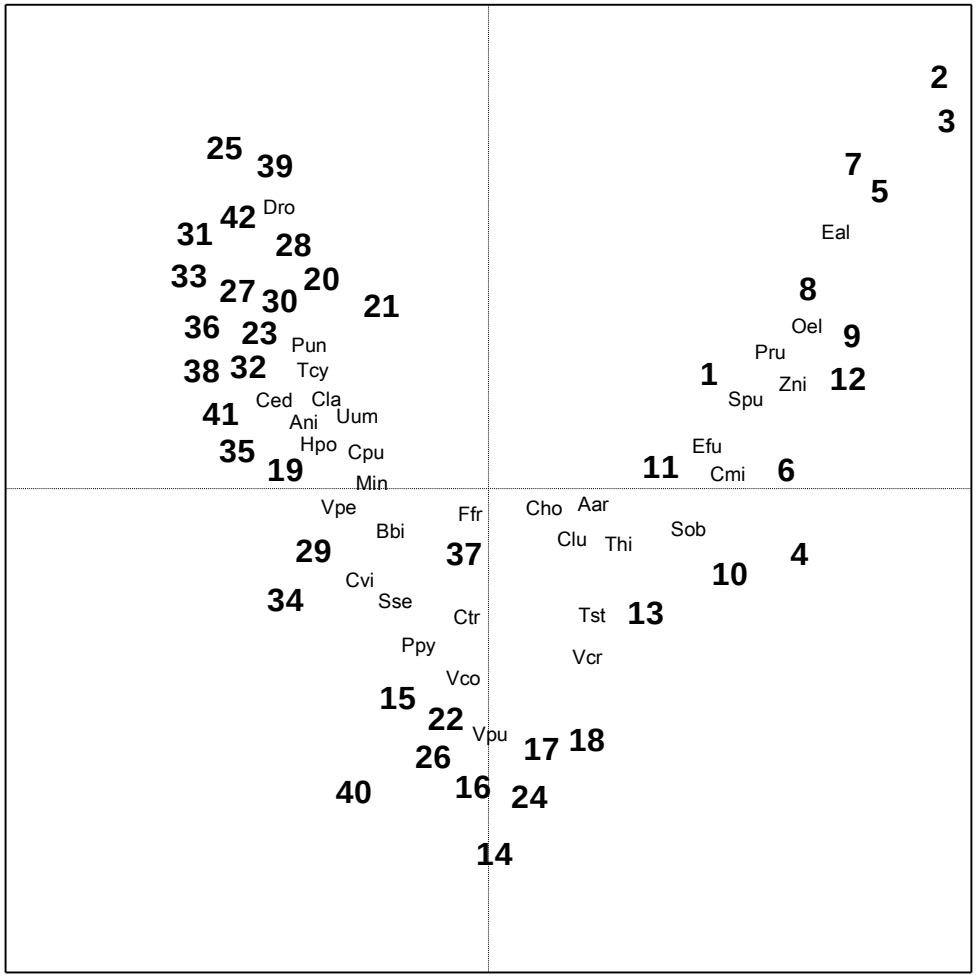
- suchozemské slimáky
- datová matice: 42 lokalit x 33 druhů slimáku

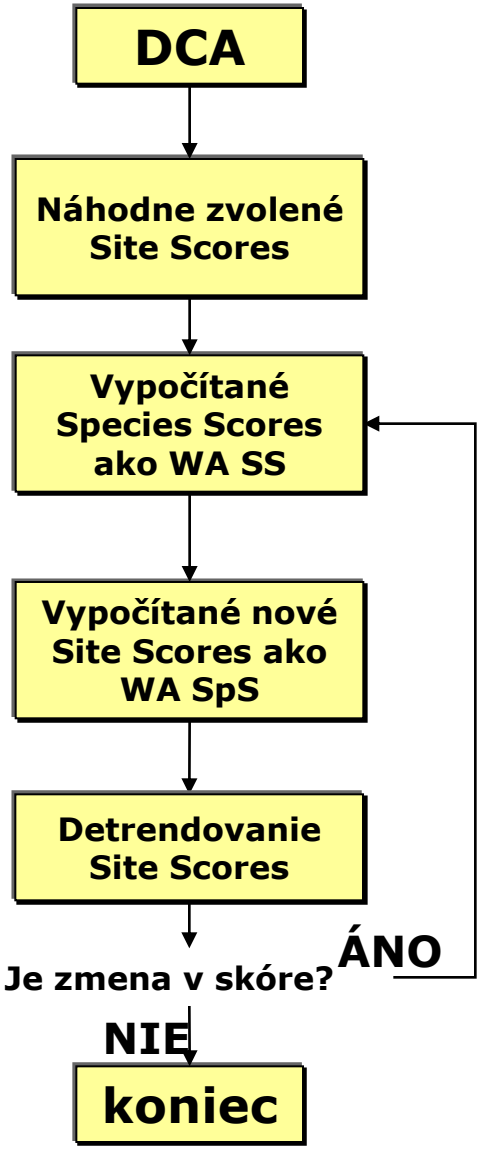
hodnoty = stupnice dominance



Korespondenčná analýza: „arch effect“

„arch effect“, „horse shoe effect“





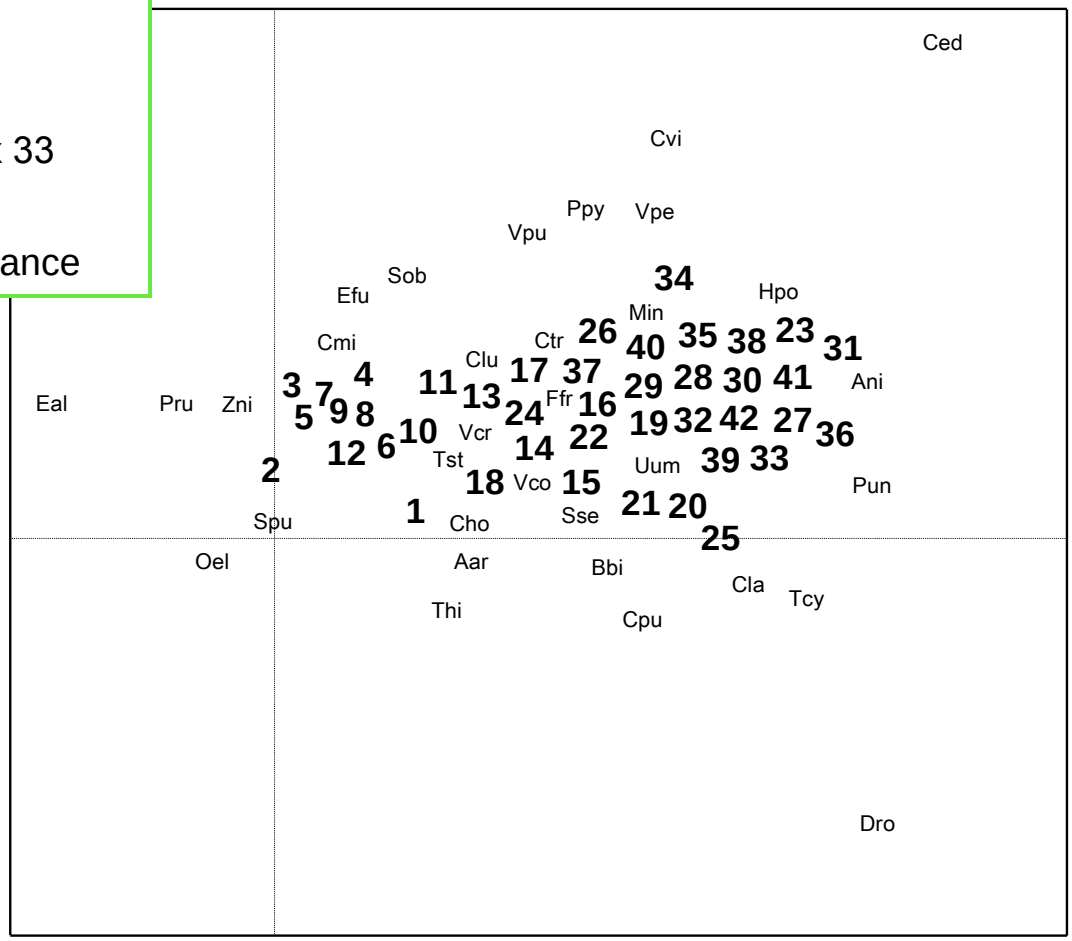
Detrendovaná korespondenční analýza (DCA)

- DCA je založena na unimodální odpovědi druhu na gradient prostředí
- DCA odstraňuje „arch effect“ různými metodami

REÁLNA DATA

- suchozemské slimáky
- datová matice: 42 lokalit x 33 druhů slimáku

  hodnoty = stupnice dominance



Mnohorozměrné škálování

Nemetrické mnohorozměrné škálování

- Mnohorozměrné škálování se používá jako průzkumná metoda
- Cílem analýzy je zobrazit pozorované podobnosti nebo nepodobnosti (vzdálenosti) mezi zkoumanými objekty v euklidovském prostoru
- Pomocí NMDS můžeme analyzovat nejenom korelační matice (tak jako je tomu v PCA) ale i jakoukoliv jinou matici podobnosti/nepodobnosti

\+

neparametrická ordinace je robustnější k vychýleným hodnotám (např. druh s výjimečně vysokou abundancí na lokalitě v jednom roku)

dá sa použít před použitím nehierarchického shlukování *k*-průměrů (v případech kdy není možné použít euklidovské vzdálenosti)

−

počet dimenzí musí byt určen předem

těžko interpretovatelné výsledky

Kanonická ordinační analýza

Kanonické ordinační metody

Přímé (kanonické) ordinační metody:
Hledání nejlepších vysvětlujících proměnných.

V kanonických ordinacích jsou ordinační osy vážené charakteristiky prostředí.
Čím méně těchto proměnných máme, tím přísnější bude omezení.

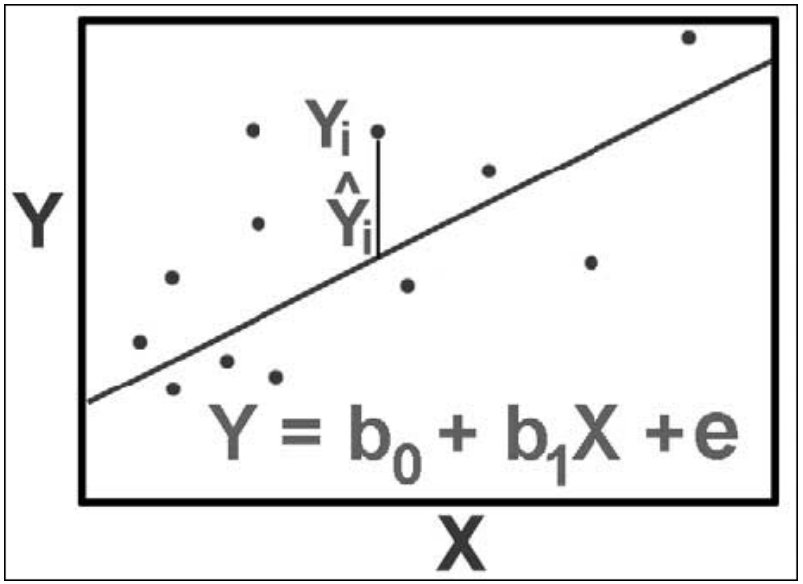
Když je jejich počet větší než počet vzorků snížený o jednu, tak se ordinace stává nepřímou.

Neomezené (*unconstrained*) ordinační osy odpovídají směru největší variability v souboru dat. **Omezené** (constrained) **ordinační osy** odpovídají směru největší variability v datovém souboru, ktorá může být vysvětlena charakteristikami prostředí.

Počet omezených os nemůže být větší než počet charakteristik prostředí.

Kanonické ordinační metody

Grafické znázornění jednoduchého lineárního regresného modelu



Y závislá proměnná (vysvětlovaná)
X nezávislá proměnná (vysvětlující)

regresní reziduál, označený jako **e**: rozdíl mezi pozorovanými hodnotami vysvětlované proměnné Y a hodnotami predikovanými modelem (predikované hodnoty, Y se stříškou).

Všechny statistické modely mají dvě důležité složky:

1. **systematická** – část variability vysvětlovaných proměnných, kterou můžeme vysvětlit vysvětlujícími proměnnými (prediktormi) pomocí zvolené parametrické funkce.
2. **stochastická** – zbývající část variability hodnot vysvětlované proměnné, kterou nemůžeme předpovědět systematickou částí modelu.

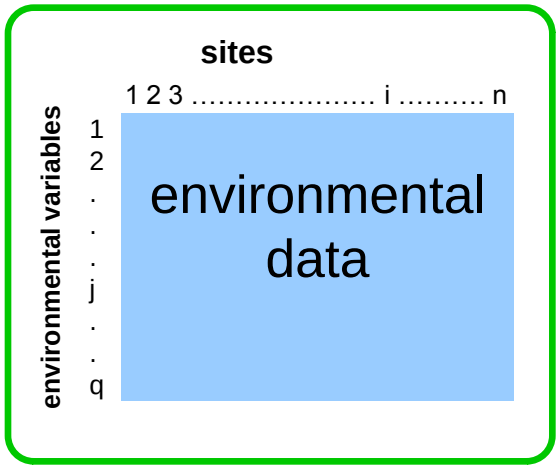
Kanonické ordinační metody

Přímá gradientova analýza (*direct gradient analysis; constrained, canonical ordination methods*) – kombinace ordinace a regrese

- Nepřímé gradientové analýzy hledaly teoretické gradienty, které byly „optimálními“ prediktory v regresních modelech lineární či unimodální odpovědi druhů.
- Metody přímé gradientové analýzy se snaží o to samé, ale gradienty, které je těmto metodam „dovoleno najít“, jsou více omezené. Tyto gradienty jsou lineární kombinací vysvětlujících proměnných (charakteristik prostředí). Abundance jednotlivých druhů se snažíme vysvětlit pomocí složených proměnných, definovaných hodnotami pozorovaných charakteristik prostředí.
- Metody přímé gradientové analýzy se podobají mnohorozměrné násobné regresi.
- Existuje tolik kanonických os, kolik je nezávislých vysvětlujících proměnných.

Kanonická korespondenční analýza (CCA)

Kromě druhových dat máme k dispozici i **vysvětlující proměnné**



Můžou být použity k předpovídání hodnot vysvětlovaných proměnných

Vysvětlující proměnné (charakteristiky prostředí)

- kvantitativní proměnné
- semikvantitativní proměnné
- faktoriální (kategoriální) proměnné - překódování do 0,1

- Kategoriální proměnné – potřeba překódovat do tzv. **indikátorových proměnných** (*dummy variables*)

vzorek	**Geo**
Vz 1	akal
Vz 2	akal
Vz 3	psamal
Vz 4	pelal

vzorek	**akal**	**psamal**	**pelal**
Vz 1	1	0	0
Vz 2	1	0	0
Vz 3	0	1	0
Vz 4	0	0	1

Kanonická korespondenční analýza (CCA)

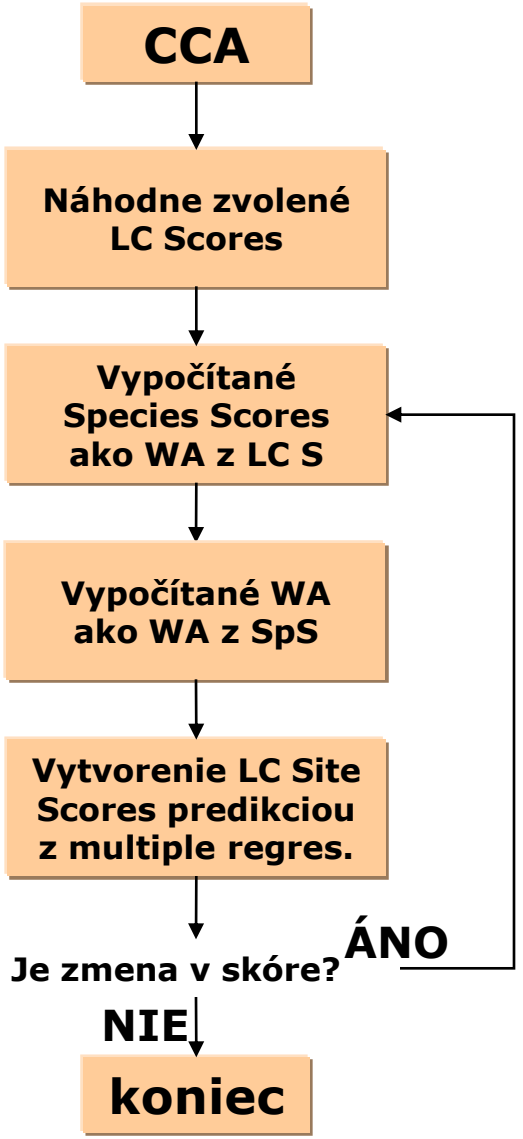
CCA je omezená ordinace

- druhová data + vysvětlující proměnné
- pouze „smysluplné“ vysvětlující preměnné

- Forward selection:

  Permutační test H_0:

  Vysvětlovací síla skupiny environmentálních proměnných se po přidání dané proměnné nezvýší víc, než kdybychom přidali takovou proměnnou, která má stejné distribuční vlastnosti jako uvažovaná proměnná, ale nemá vztah k druhovým datům.



Kanonická korespondenční analýza (CCA)

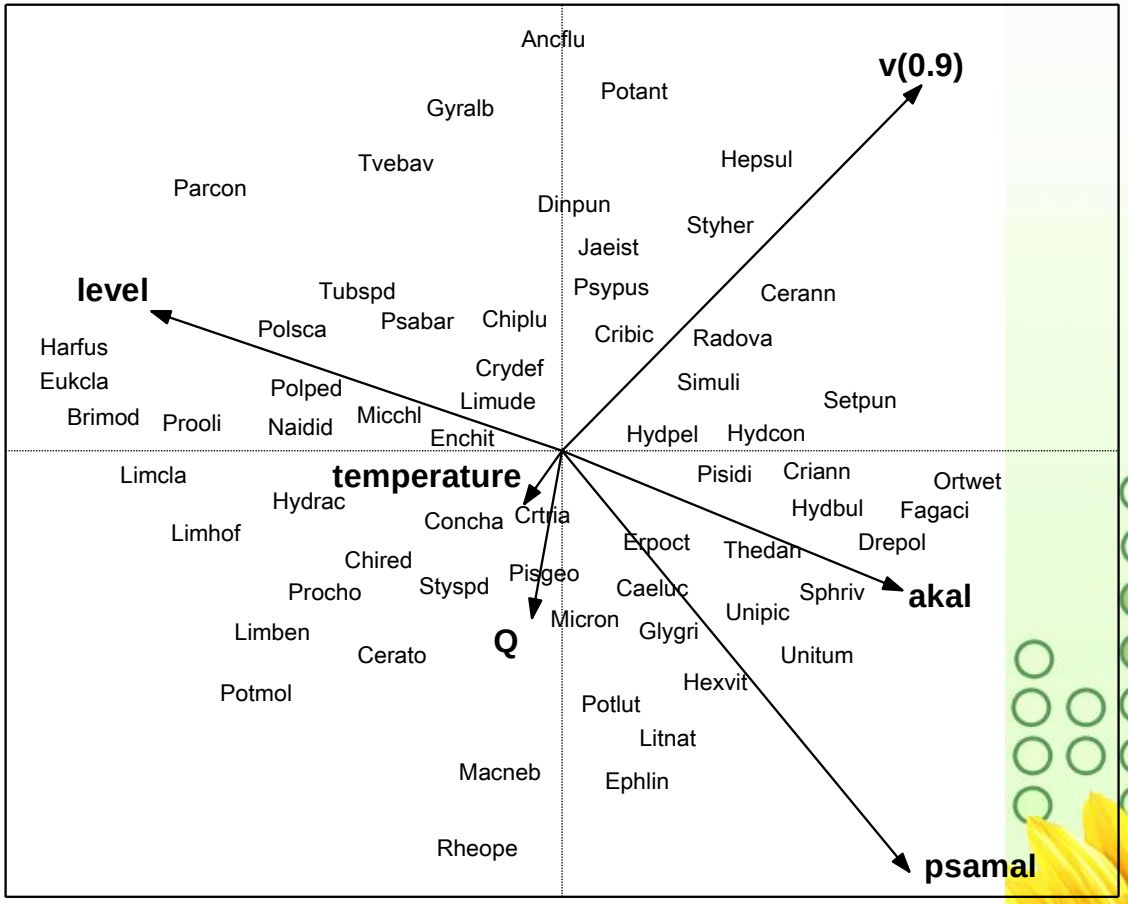
Direct gradient analysis

Canonical correspondence analysis

- CCA je kanonická forma CA
- CCA se doporučuje pro druhová data s velkým výskytem nulových hodnot

REÁLNA DATA

- společenstva makrozoobentosu
- datové matice:
  60 lok. x 63 tax. (stupeň dominance)
  60 lok. x 13 environm. faktorů (fs)



Závěrem

Využití základních typů vícerozměrných analýz

- Shluková analýza
  - Možnost využití libovolných asociačních koeficientů netrpících problémem double zero
  - Poskytuje rozdělení společenstev do shluků
  - Problematické použití při velkém počtu shluků
- Korespondenční analýza
  - Na rozdíl od PCA apod. netrpí problémem double zero
  - Poskytuje pozici lokalit v xy grafu
  - Omezena pouze na chi-square vzdálenost
  - Arch effect
- Multidimensional scaling
  - Poskytuje pozici lokalit v xy grafu
  - Možnost využití libovolných asociačních koeficientů netrpících problémem double zero
  - V řadě případů problematická interpretace os

Software

- Canoco for Windows (ter Braak & Šmilauer 2004)
- SYN-TAX 2000 (Podani 1997)
- Statistica (StatSoft, Inc. 2005)
- PAST (Hammer, Harper, Ryan 2001)

  http://folk.uio.no/ohammer/past/

SYN-TAX 2000

STATISTICA

StatSoft®

Děkuji za pozornost