Multivariační morfometrika, geometrická morfometrika, rekonstrukce evoluce, tvorba fylogenetických stromů

Fenetický přístup (multivariační metody; “pattern”; shluková analýza, ordinační metody, diskriminační analýza)

Kladistický přístup (parsimonická analýza)

Alternativní přístupy k rekonstrukci fylogenézy
(metoda spájení sousedních objektů – *neighbour joining method*;
metody největší věrohodnosti – *maximum likelihood*;
Bayesovské metody – *Bayesian statistical methods*)

Geometrická morfometrika (Booksteinovy souřadnice tvaru,
Prokrustova analýza, metoda ohybných pásků - *thin plate spline*)

Fenetický přístup

Department of Entomology, University of Kansas, Lawrence, U.S.A.

Michener, Ch.D. & Sokal, R.R. 1957. A quantitative approach to a problem in classification. *Evolution* 11: 130-162.

Department of Microbiology, University of Leicester, U.K.

Sneath, P.H.A. 1957. Some thoughts on bacterial classification. *J. Gen. Microbiol*. 17: 184-200.

Sokal, R.R. & Sneath, P.H.A. 1963. *Principles of numerical taxonomy*. W. H. Freeman and comp., San Francisco & London.

Sneath, P.H.A. & Sokal, R.R. 1973. *Numerical taxonomy, the principles and practice of numerical classification*. W. H. Freeman and comp., San Francisco.

Neo-adansonovské princípy

Čím větší je obsah informace v taxonech a na čím větším počtu znaků je klasifikace založena, tím je tato klasifikace lepší.

Každý znak má při tvorbě taxonů stejnou váhu.

Celková podobnost mezi taxony je funkcí podobností v jednotlivých znacích.

Taxony se rozeznávají na základě toho, že se korelace mezi znaky v různých skupinách líší.

Taxonomie se považuje za praktickou a empirickou vědu.

Klasifikace se zakládají na empirické podobnosti.

Klasifikace znaků

(1) **kvalitativní (*qualitative*):**

binární (*binary*, dvoustavové, dvouhodnotové, alternativní)
vícestavové (*multistate*, vícehodnotové)

(2) **semikvantitativní (*semiquantitative*)**

(3) **kvantitativní (*quantitative*)**

nespojité, diskrétní (*discontinuous, discrete, meristic*)
spojité, kontinuální (*continuous*)

Koeficienty vyjadřující vztahy mezi objekty nebo znaky (*resemblance coefficients*)

(1) koeficienty vzdálenosti pro kvantitativní a binární znaky (*metric distances*)

(2) koeficienty podobnosti pro binární znaky (*binary similarity coefficients*)

(3) koeficienty pro smíšená data (*coefficients for mixed data*)

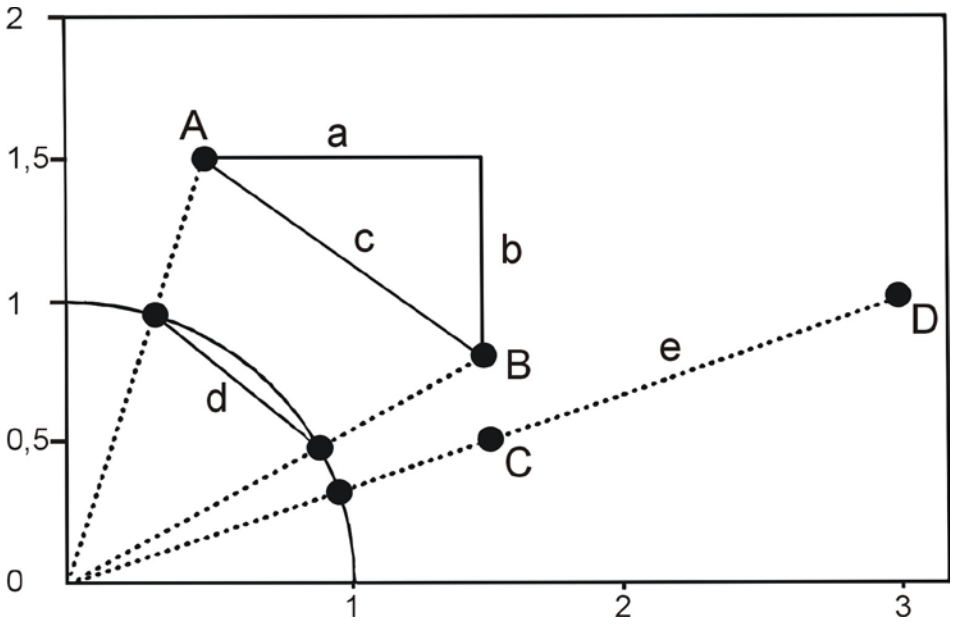
(4) korelační koeficienty (*correlation coefficients*)

Metriky (vzdálenosti)

Euklidovská vzdálenost (Euclidean distance):

$$EU=c \qquad EU_{jk} = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - x_{ik}\right)^2}$$

kde x_{ij} je hodnota znaku *i* pro objekt *j*, x_{ik} je hodnota znaku *i* pro objekt *k*, n je celkový počet znaků



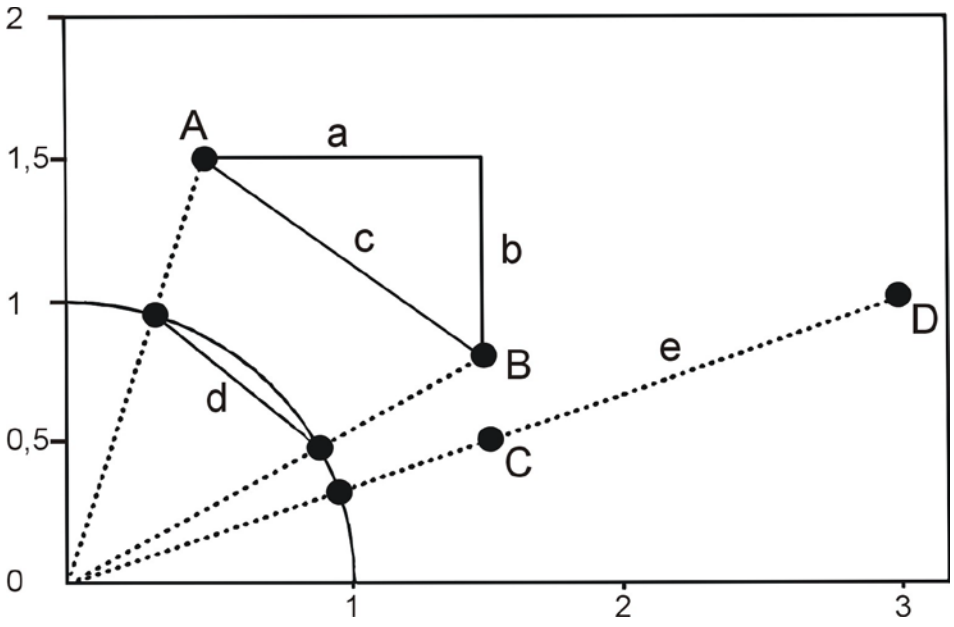
Metriky (vzdálenosti)

Manhattanská (city block) metrika:

$CB = a + b$

$$CB_{jk} = \sum_{i=1}^{n} \left| x_{ij} - x_{ik} \right|$$

Připomíná severoamerická města s kolmými ulicemi, kde se musí chodit kolem bloků



Minkowského metrika:

$$MNK_{jk} = \sqrt[r]{\sum_{i=1}^{n} \left(x_{ij} - x_{ik} \right)^{r}}$$

kde $r \geq 1$;
pro $r=1$ …. CB
***Pro* $r=2$ … EU**

Koeficienty podobnosti pro binární data

Výběr koeficientu podobnosti

	objekt 2		
objekt 1		1	0
	1	*a*	*b*
	0	*c*	*d*

a – počet znaků, ve kterých mají oba objekty hodnotu + (resp. 1) (pozitivní shoda)

b – počet znaků, ve kterých má objekt i hodnotu – (resp. 0) a objekt j hodnotu + (resp. 1)

c – počet znaků, ve kterých má objekt i hodnotu + (resp. 1) a objekt j hodnotu – (resp. 0)

d – počet znaků, ve kterých mají oba objekty hodnotu – (resp. 0) (negativní shoda)

Volba mezi koeficienty závisí především na tom, jestli pro dané znaky má nebo nemá smysl **negativní shoda**, tj. zdali má nebo nemá smysl uvažovat, že nulová hodnota znaku má u porovnávaných objektů stejnou příčinu

Koeficienty podobnosti pro binární data

Koeficienty hodnotící *a* a *d* symetricky:

	objekt 2		
		1	0
objekt 1	1	*a*	*b*
	0	*c*	*d*

Koeficient jednoduché shody (*simple matching*):

koeficient je blízký ED:
$ED^2 = n(1\text{-}SM)$
$n=a+b+c+d$
$ED = \sqrt{b+c}$

$$SM = \frac{a+d}{a+b+c+d}$$

Koeficient Rogerse a Tanimota:
neshody jsou vážené dva krát;
hodnoty vždy nižší než u SM, s výjimkou b+c=0

$$RT = \frac{a+d}{a+2b+2c+d}$$

Hamannův index:
rozpětí [-1,1]

$$HAM = \frac{a+d-b-c}{a+d+b+c}$$

$$SM = \frac{HAM+1}{2}$$

Koeficienty podobnosti pro binární data

Koeficienty, které neberou do úvahy negativní shodu:

		objekt 2	
		1	0
objekt 1	1	*a*	*b*
	0	*c*	*d*

Jaccardův koeficient: $JAC = \frac{a}{a+b+c}$

rozpětí [0,1]

konverze $d_{jk} = \sqrt{1 - s_{jk}}$

má za výsledek Euklidovskou vzdálenost

Sorensenův koeficient: $SOR = \frac{2a}{2a+b+c}$

pozitivní shoda se váží dva krát

genetické vzdálenosti podle Nei & Li (1979), Link et al. (1995) využívané při NJ, PCoA odpovídají také tomuto typu koeficientů

Nei & Li (1979):

$$NL = 1 - \frac{2a}{2a+b+c}$$

Link et al. (1995):

$$L = \frac{b+c}{b+c+a}$$

Koeficienty pro smíšená data

Do této kategorie patří Gowerův koeficient a vzdálenost pro smíšená data. Používají se v případech, kdy jsou v matici současně zastoupeny kvalitativní znaky a znaky kvantitativní nebo binární (případně všechny tři druhy znaků).

Gowerův koeficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j – objekty charakterizované znakem k,
n – celkový počet znaků,
s_{ijk} – skóre znaku k

w_{ijk} je váha, která může nabývat hodnot 1 nebo 0 podle toho, jestli je nebo není možné srovnání hodnot znaku k u objektů i a j (krome binárních znaku může mít nulovou hodnotu jenom tehdy, pokud hodnota znaku k není u jednoho nebo obou objektů známá); s_{ijk} je skóre (hodnota) pro príslušný znak k.

a) pro binární znaky:

$w_{ijk} = 1$ a $s_{ijk} = 0$ pokud $x_{ik} \neq x_{jk}$ (hodnoty znaku k pro objekty i a j)

$w_{ijk} = s_{ijk} = 1$ pokud $x_{ik} = x_{jk} = 1$ nebo pokud $x_{ik} = x_{jk} = 0$ a negativní shoda se bere do úvahy (odpovídá koeficientu jednoduché shody)

$w_{ijk} = s_{ijk} = 0$ pokud $x_{ik} = x_{jk} = 0$ a negativní shoda se nebere do úvahy (odpovídá Jaccardovu koeficientu)

Koeficienty pro smíšená data

Gowerův koeficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j – objekty charakterizované znakem k,
n – celkový počet znaků,
s_{ijk} – skóre znaku k

b) pro nominální znaky:
$w_{ijk} = 1$ pokud x_{ik} a x_{jk} jsou známé; pak
$s_{ijk} = 0$ pokud $x_{ik} \neq x_{jk}$; $s_{ijk} = 1$ pokud $x_{ik} = x_{jk}$ (počet stavů se nebere do úvahy)

Koeficienty pro smíšená data

Gowerův koeficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j – objekty charakterizované znakem k,
n – celkový počet znaků,
s_{ijk} – skóre znaku k

c) pro kvantitativní znaky:
w_{ijk} = 1 pokud x_{ik} a x_{jk} jsou oba známé, a $s_{ijk} = 1 - \{|x_{ik} - x_{jk}| / (\text{rozpětí znaku } i)\}$
(odpovídá Manhattanské metrice s daty stadardizovanými na rozpětí)

Korelační koeficienty

Pearsonův korelační koeficient

$$r_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}}$$

n počet objektů,
x_{i1} hodnota znaku 1 pro objekt i

lineární korelace, predpokládá normální rozdělení dat

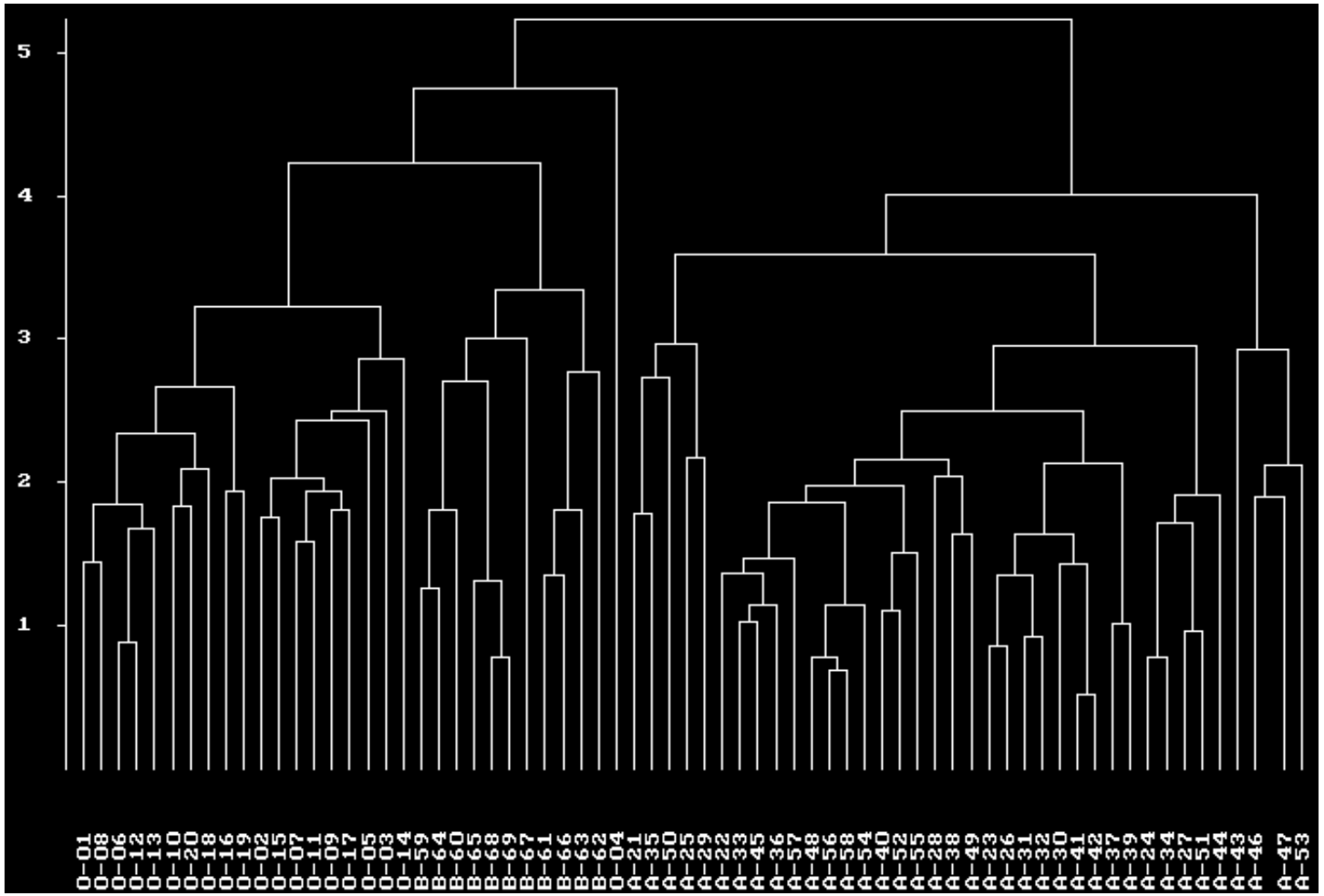
Spearmanův korelační koeficient (rank koeficient, koeficient pořadí):

$$r_{12} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

do úvahy se neberou konkrétní hodnoty znaků ale pořadí objektů,
kde d_i je rozdíl v pořadí mezi objekty;

Pearsonův korelační koeficient a Spearmanův korelační koeficient:
rozpětí [-1, +1], +1 ´přímá závislost, -1 nepřímá závislost, 0 absence vztahu

Shluková analýza



Shluková analýza

způsob tvorby shluků: **aglomerativní metody – divizivní metody**

uspořádání shluků: **hierarchické metody – nehierarchické metody**

překryv shluků: **nepřekrývající** nebo **překrývající se shluky** (*fuzzy clustering*)

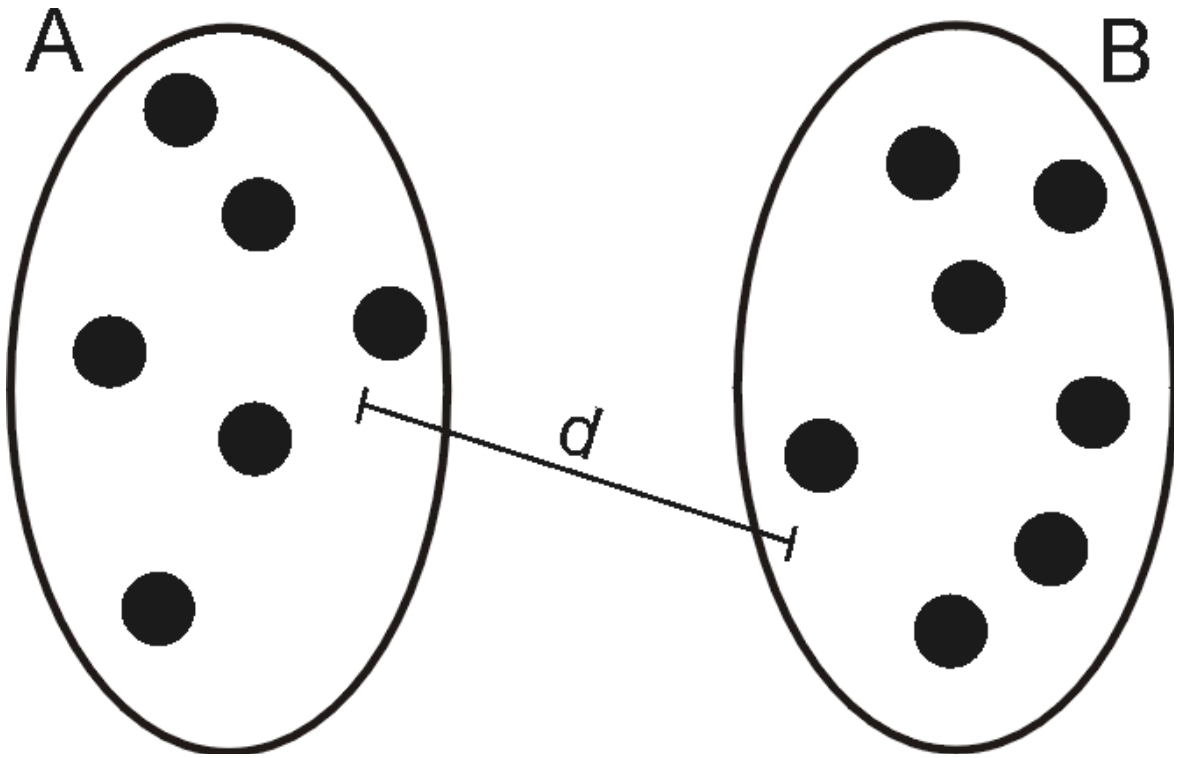
postup shlukování: **sekvenční metody – simultánní metody**

Shlukovací metody kategorie SAHN:

(a) metody založené na minimalizaci vzdálenosti mezi shluky

(b) metody založené na optimalizaci homogenity shluků podle určitého kritéria

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



$D_1 =$

	1	**2**	**3**	**4**	**5**
1	0,0	1,0	7,0	4,0	12,0
2	1,0	0,0	2,0	5,0	9,0
3	7,0	2,0	0,0	8,0	3,0
4	4,0	5,0	8,0	0,0	6,0
5	12,0	9,0	3,0	6,0	0,0

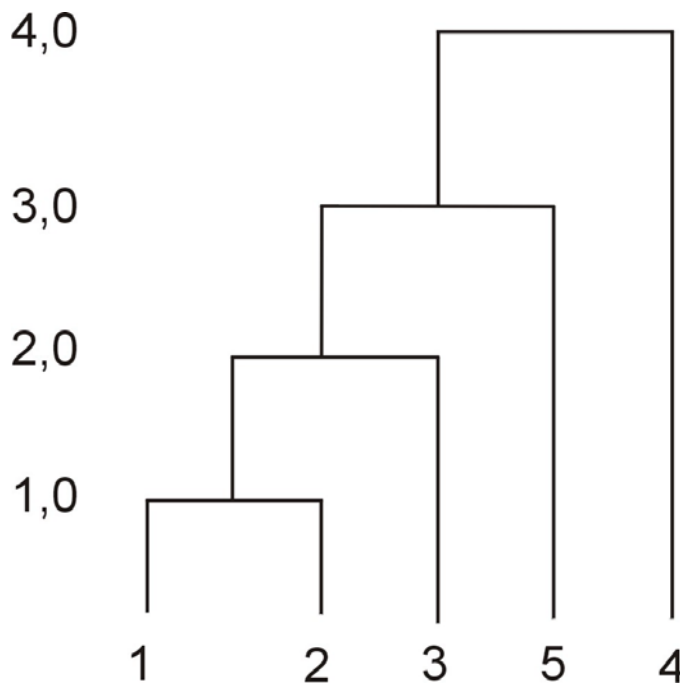
$d_{(1,2)3} = \min \{d_{1,3}, d_{2,3}\} = d_{2,3} = 2,0$

$d_{(1,2)4} = \min \{d_{1,4}, d_{2,4}\} = d_{1,4} = 4,0$

$d_{(1,2)5} = \min \{d_{1,5}, d_{2,5}\} = d_{2,5} = 9,0$

$D_2 =$

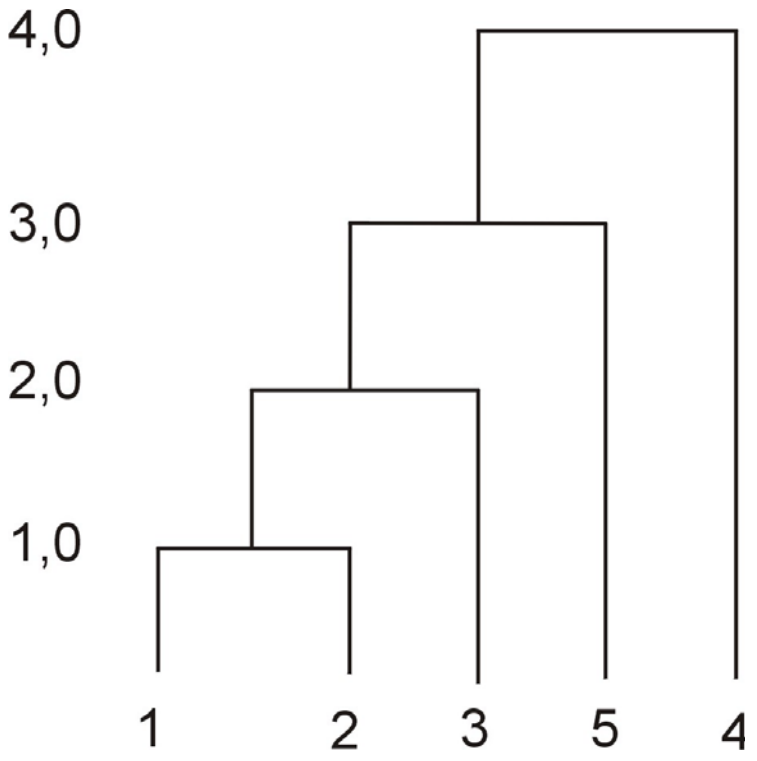
	(1, 2)	**3**	**4**	**5**
(1, 2)	0,0	2,0	4,0	9,0
3	2,0	0,0	8,0	3,0
4	4,0	8,0	0,0	6,0
5	9,0	3,0	6,0	0,0



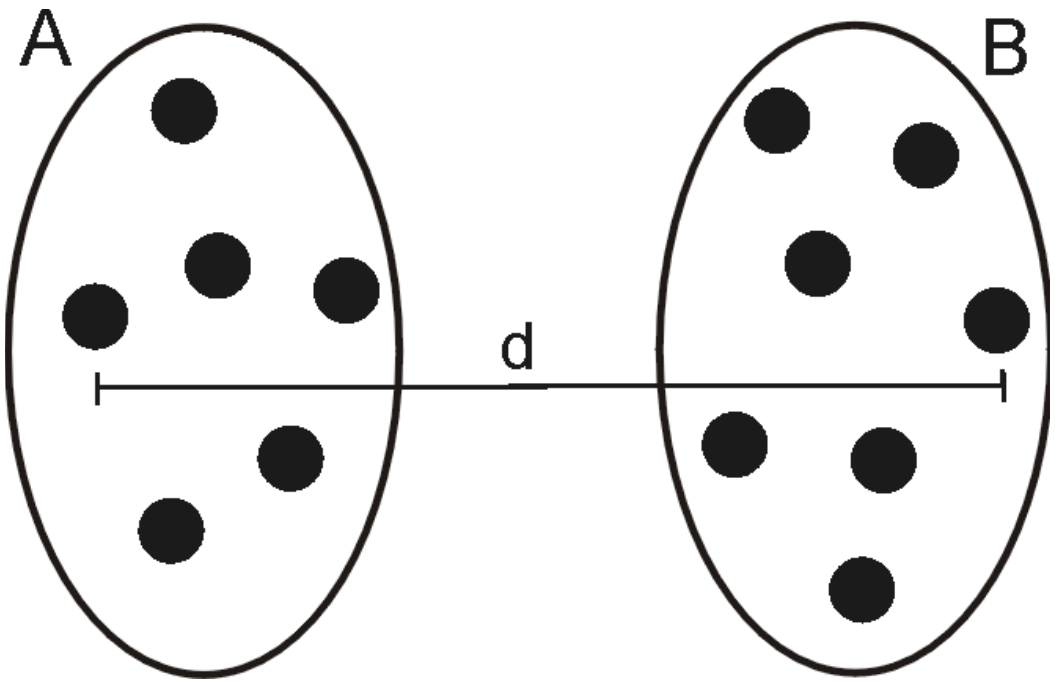
$d_{(1, 2, 3)4} = \min \{d_{(1, 2)\,4}, d_{3, 4}\} = d_{(1, 2)\,4} = 4{,}0$

$d_{(1, 2, 3)5} = \min \{d_{(1, 2)\,5}, d_{3, 5}\} = d_{3, 5} = 3{,}0$

		(1, 2, 3)	**4**	**5**
	(1, 2, 3)	0,0	4,0	3,0
$D_3 =$	**4**	4,0	0,0	6,0
	5	3,0	6,0	0,0



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



$D_1 =$

	1	**2**	**3**	**4**	**5**
1	0,0	1,0	7,0	4,0	12,0
2	1,0	0,0	2,0	5,0	9,0
3	7,0	2,0	0,0	8,0	3,0
4	4,0	5,0	8,0	0,0	6,0
5	12,0	9,0	3,0	6,0	0,0

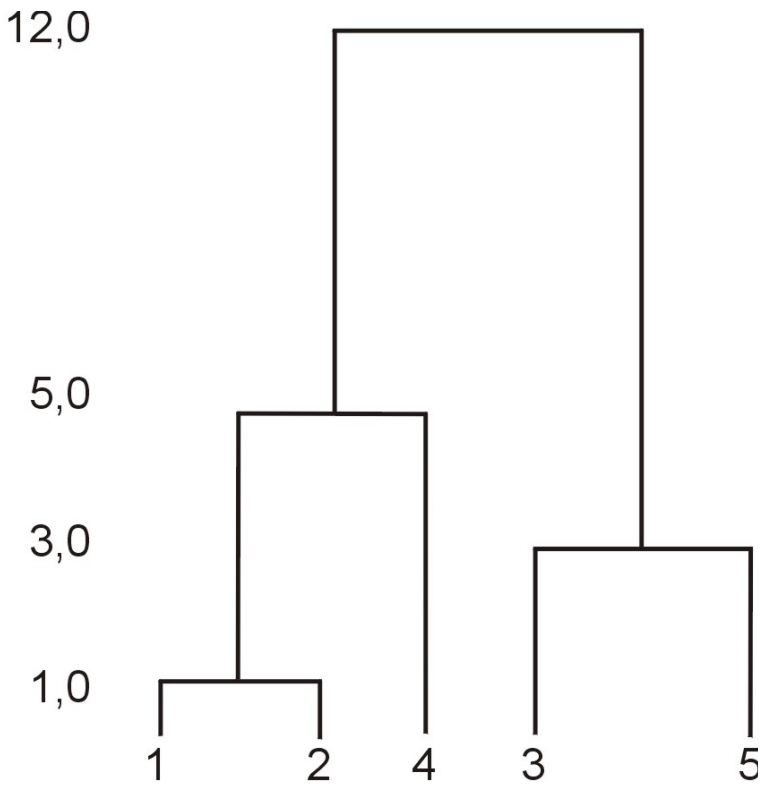
$d_{(1,\,2)3} = \max\{d_{1,\,3}, d_{2,\,3}\} = d_{1,\,3} = 7{,}0$

$d_{(1,\,2)4} = \max\{d_{1,\,4}, d_{2,\,4}\} = d_{2,\,4} = 5{,}0$

$d_{(1,\,2)5} = \max\{d_{1,\,5}, d_{2,\,5}\} = d_{1,\,5} = 12{,}0$

$D_2 =$

	(1, 2)	**3**	**4**	**5**
(1, 2)	0,0	7,0	5,0	12,0
3	7,0	0,0	8,0	3,0
4	5,0	8,0	0,0	6,0
5	12,0	3,0	6,0	0,0

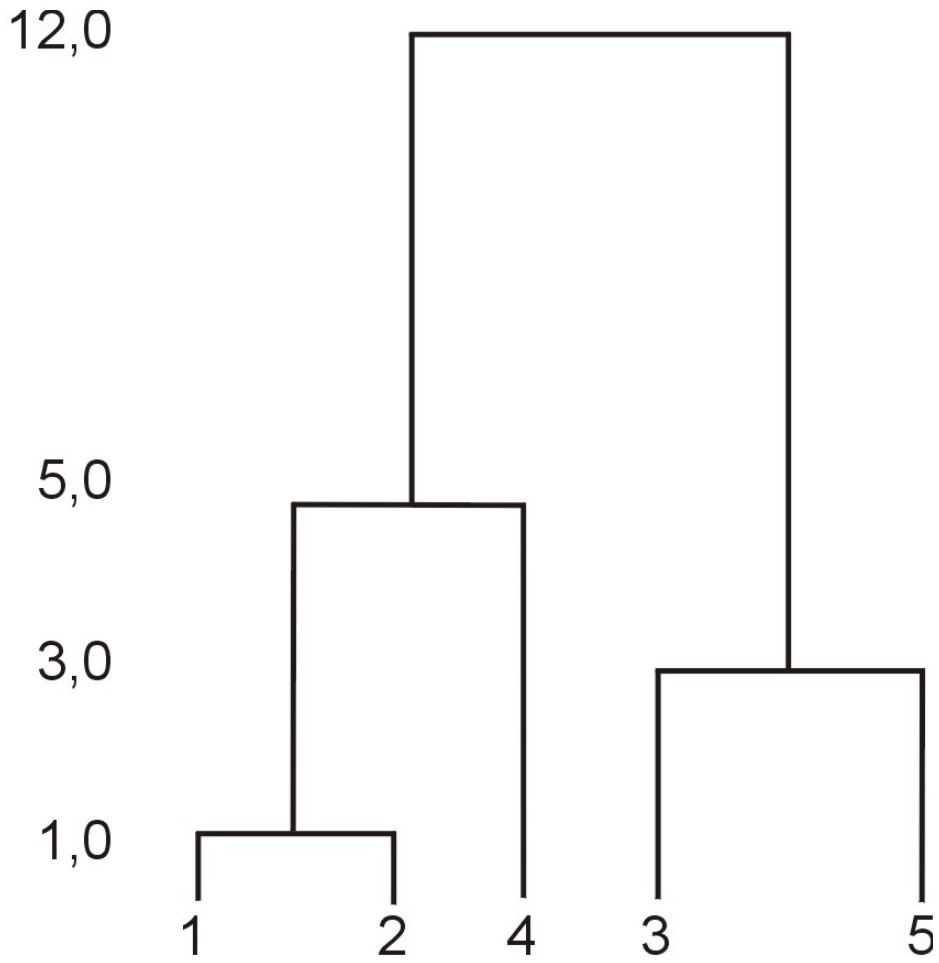


$d_{(1,\ 2)(3,\ 5)} = \max\{d_{(1,\ 2)\ 3}, d_{(1,\ 2)\ 5}\} = d_{(1,2),\ 5} = 12{,}0$

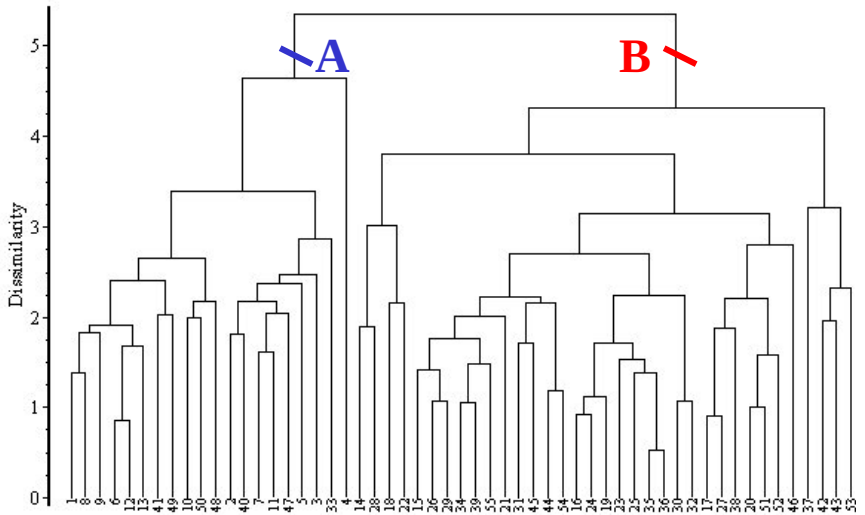
$d_{(3,\ 5)4} = \max\{d_{3,\ 4}, d_{3,\ 5}\} = d_{3,\ 4} = 8{,}0$

$D_3 =$

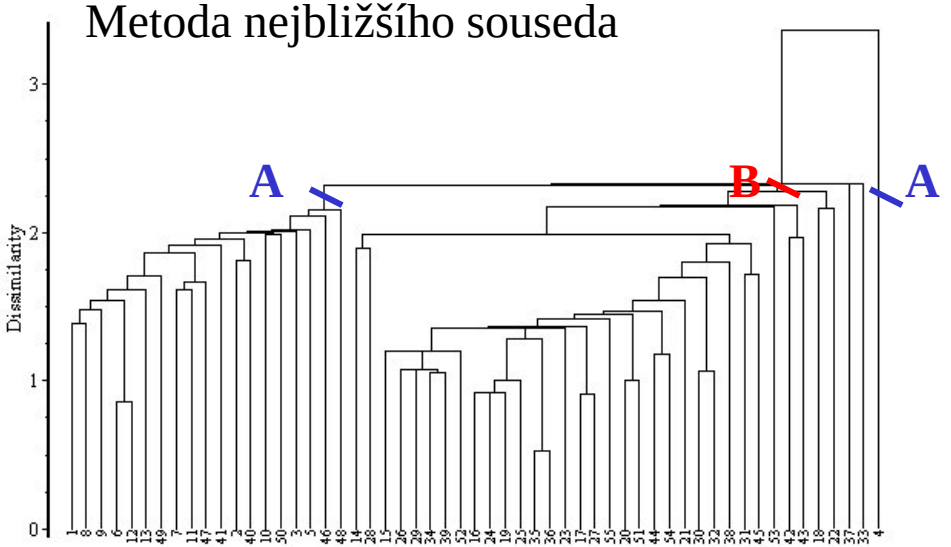
	(1, 2)	**(3, 5)**	**4**
(1, 2)	0,0	12,0	5,0
(3, 5)	12,0	0,0	8,0
4	5,0	8,0	0,0



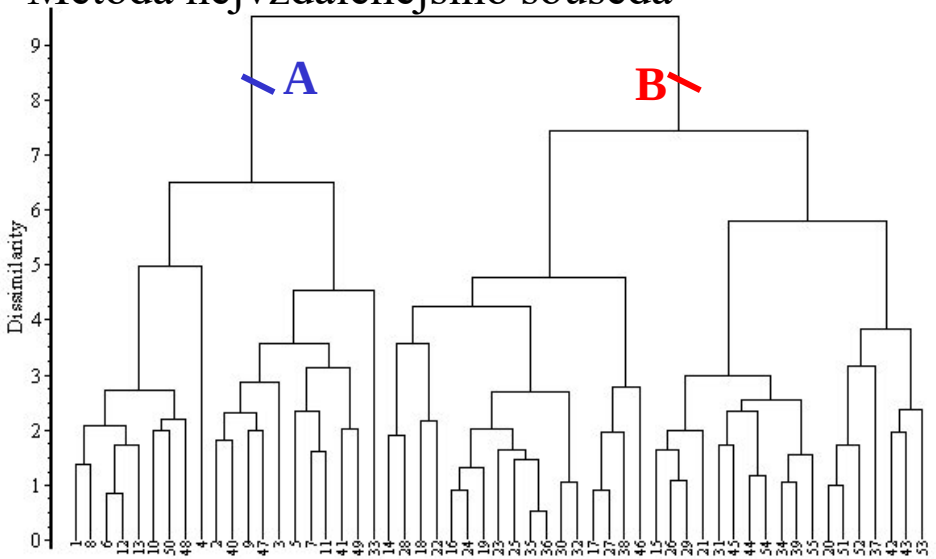
Metoda průměrné vzdálenosti



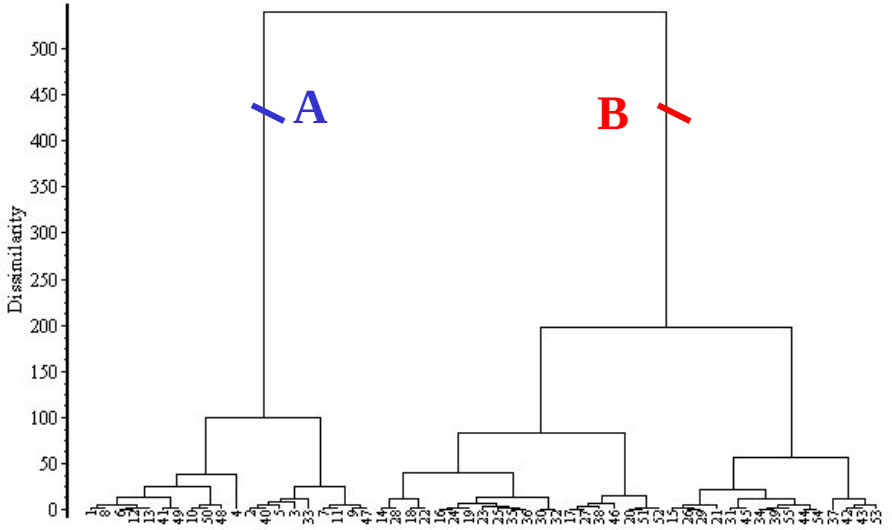
Metoda nejbližšího souseda



Metoda nejvzdálenějšího souseda



Wardova metoda



jednoznačná podpora je pro dva taxony, další seskupení reflektují rozdíly ve shlukovacích algoritmech

Obecné poznámky ke shlukovacím metodám

Pokud data nemají zcela jednoznačnou a zřetelnou strukturu (jedná se víceméne o náhodně rozptýlené objekty), je pravděpodobné, že použití různých shlukovacích technik přinese odlišné výsledky.

Pokud různé shlukovací techniky přinášejí z téhož souboru dat shodné, resp. podobné výsledky, je to do jisté míry potvrzení struktury obsažené 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 atypických případů). Před samotnou shlukovou analýzou je proto vhodné použít některou z metod na jejich detekci, např. PCA. Výrazně odlehlé objekty se zpravidla z dalších analýz vylučují.

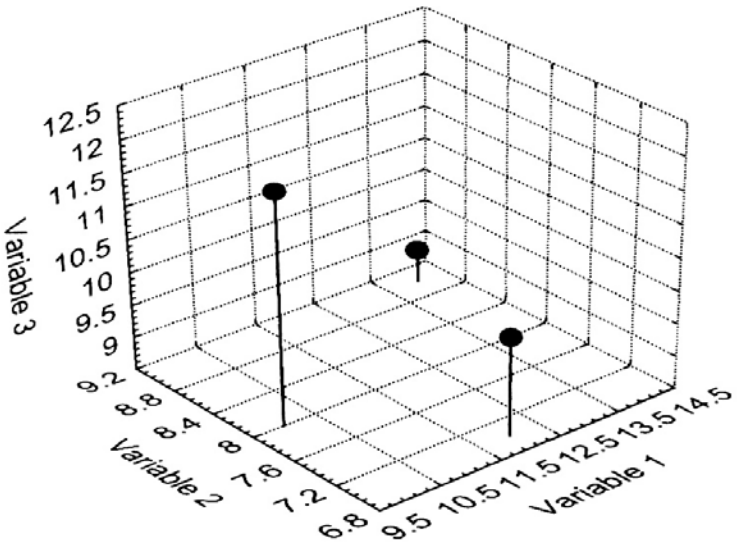
Shlukové analýzy obecně nejsou vhodné pro data, která popisují klinální variabilitu znaků (*cline* = variabilita znaku závislá na gradientu prostředí).

Ordinační metody

Objekty charakterizované *p* znaky je možné si představit jako body v *p* rozměrném prostoru, kde každý z rozměrů představuje hodnoty jednoho znaku.

Pokud pracujeme pouze se dvěma nebo třemi znaky je možné bez problémů sledovat na dvoj- případně trojrozměrném grafu vztahy mezi objekty, jejich vzdálenosti a seskupení.

Větší počet znaků **=>** nutnost redukce jejich počtu s co nejmenší ztrátou informace



Ordinační metody

analýza hlavních komponent (PCA)

analýza hlavních koordinát (PCoA)

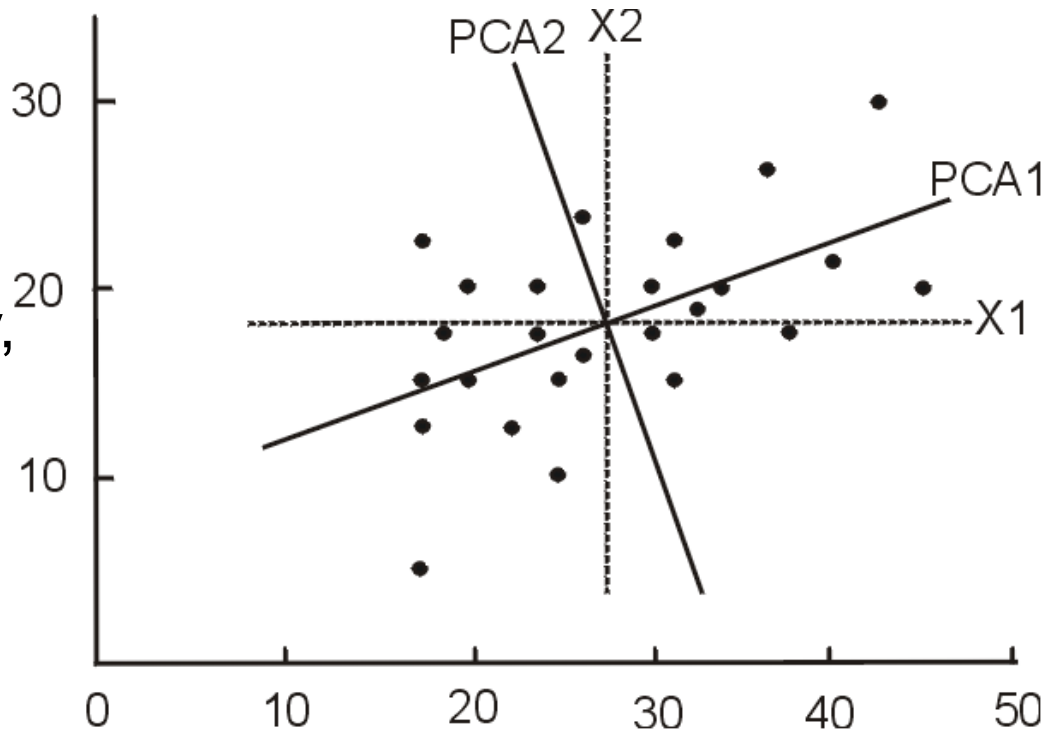
nemetrické mnohorozměrné škálování (NMDS)

Užitečné informace o ordinačních metodách se nacházejí na WWW stránce

http://ordination.okstate.edu/

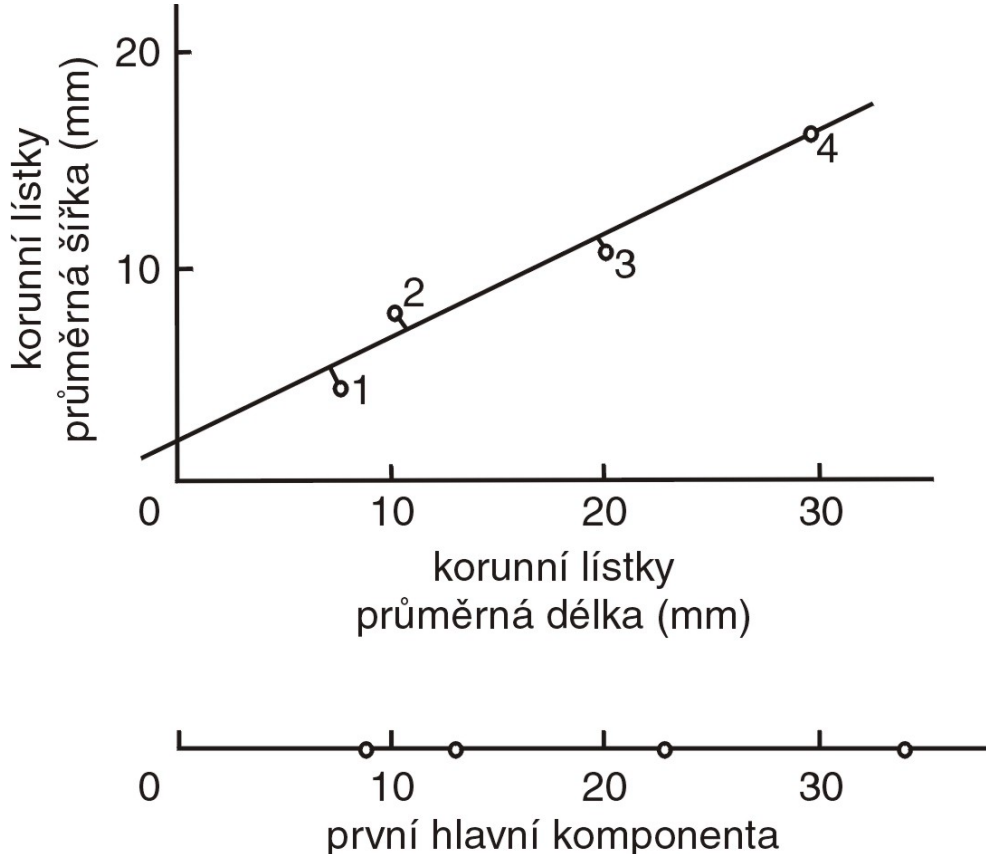
Analýza hlavních komponent (PCA – *principal component(s) analysis*)

nahrazuje původní soubor pozorovaných znaků souborem nových (hypotetických), vzájemně nekorelovaných znaků tak, že první nová osa (první hlavní komponenta, PC1, první nový znak) je vedena ve směru největší variability mezi objekty, druhá osa (druhá hlavní komponenta, PC2, druhý nový znak) je vedena ve směru největší variability, který je kolmý na směr první komponenty, atd.



Geometrická interpretace PCA (podle Dunn & Everitt 1982):

OTU	**1**	**2**	**3**	**4**
průměrná délka korunních lístků (mm)	8	10	20	30
průměrná šířka korunních lístků (mm)	4	9	11	18



Původní soubor p pozorovaných znaků $x_1, x_2, ..., x_p$

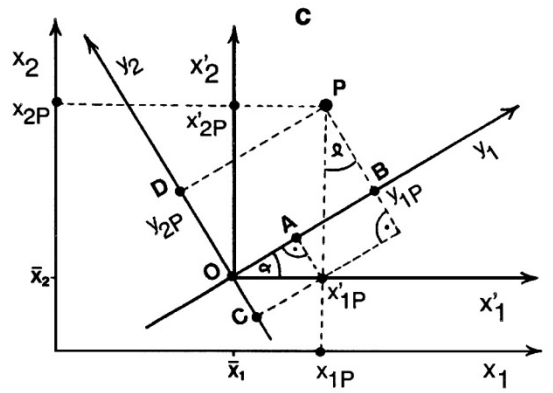
se transformuje do nového souboru znaků $y_1, y_2, ..., y_p$

$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$

.

.

$y_p = a_{p1}x_1 + a_{p2}x_2 + ... + a_{pp}x_p$



Koeficienty první hlavní komponenty - vektor $\boldsymbol{a}_1$

první hlavní komponenta $y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$ vyjádřena

vektorově $\boldsymbol{a}_1'\boldsymbol{x}$

Podobně $y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p$ můžeme zapsat jako $\boldsymbol{a}_2'\boldsymbol{x}$ atd.

Komponenty nejsou vzájemně korelované
čiže platí: $\boldsymbol{a}_2'\boldsymbol{a}_1 = 0$

Suma čtverců koeficientů každé z lineárních kombinací
se rovná jedné $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$, $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ atd.

Obecně pro *j*-tou hlavní komponentu platí
$y_j = \boldsymbol{a}_j'\boldsymbol{x}$
a tato má největší rozptyl za podmínek,
že $\boldsymbol{a}_j'\boldsymbol{a}_j = 1$ a $\boldsymbol{a}_j'\boldsymbol{a}_i = 0$, $i \neq j$.

K symetrické matici S_{pp} (jakou je kovarianční anebo korelační matice), je možné přiřadit *p* reálných vlastních čísel (charakteristických čísel, *eigenvalue*, *characteristic root*, *latent root*) λ_1 ... λ_p a *p* sloupcových *p*-složkových vlastních vektorů (charakteristických vektorů, *eigenvector*, *characteristic vector*, *latent vector*) $\boldsymbol{a}_1$, ..., $\boldsymbol{a}_p$,

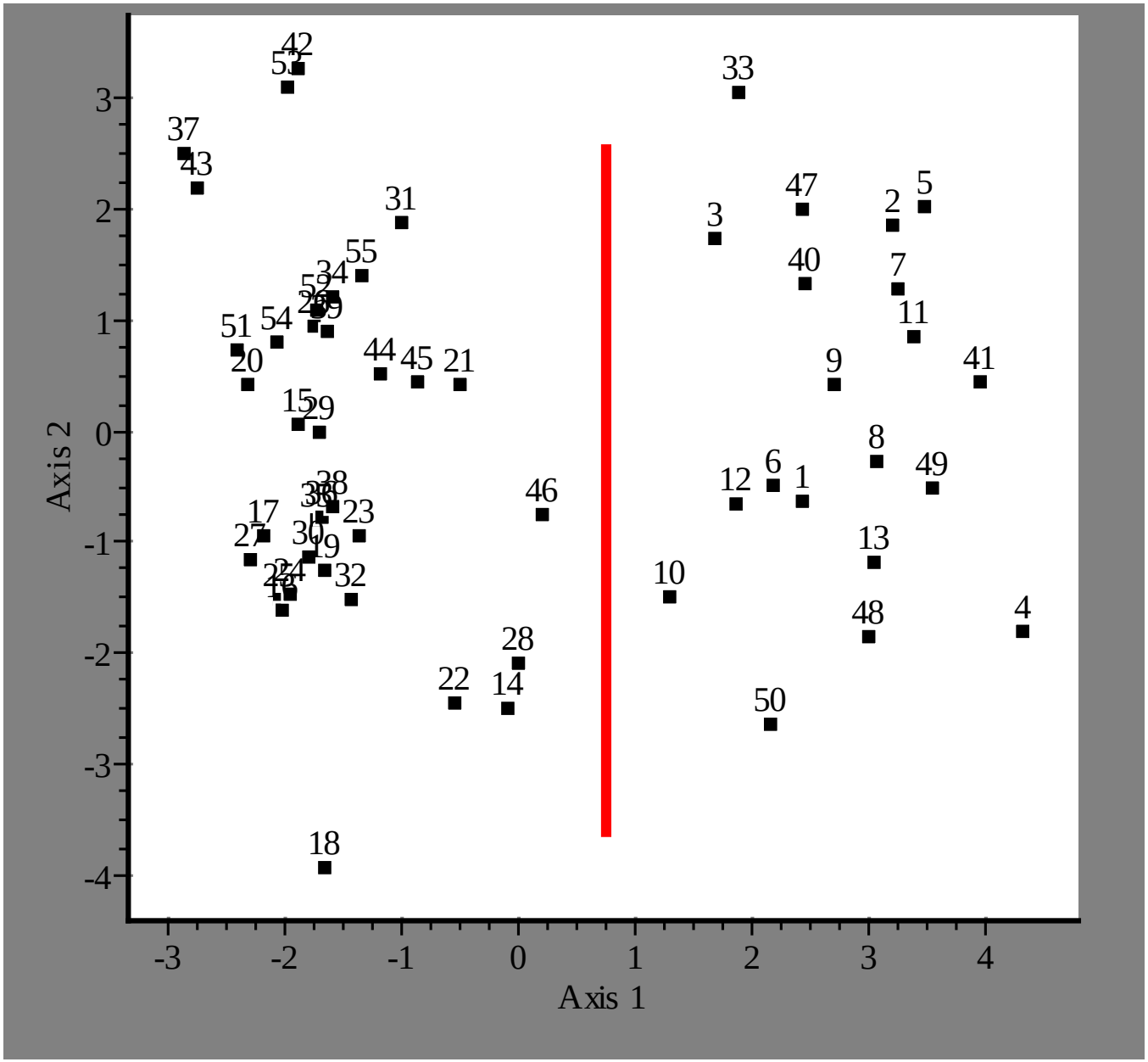
přičemž platí $S_{pp} = A_{pp} \Lambda_{pp} A_{pp}'$.

Je možné dokázat, že vektory koeficientů $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, ... $\boldsymbol{a}_p$ jsou vlastní vektory kovarianční nebo korelační matice; v případě, že suma jejich čtverců je 1 (viz výše $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$),
jsou vlastní čísla této matice λ_1, λ_2, ... λ_p interpretovatelné jako míry rozptylu zachycené komponentami y_1, ... , y_p.

Cardamine amara (Brassicaceae)

subsp. *amara*

subsp. *opicii*



Ordinace objektů

Vlastní čísla

(1) NUMBER OF POSITIVE EIGENVALUES = 10

(2) SUM OF POSITIVE EIGENVALUES = 0.10000000E+02

(3) EIGENVALUES

0.5030E+01	0.2590E+01	0.1127E+01	0.3886E+00	0.3164E+00
0.1992E+00	0.1353E+00	0.1054E+00	0.6441E-01	0.4339E-01

(4) EIGENVALUES AS PERCENT

50.30	25.90	11.27	3.89	3.16
1.99	1.35	1.05	.64	.43

(5) CUMULATIVE PERCENTAGE OF EIGENVALUES

50.30	76.20	87.47	91.36	94.52
96.52	97.87	98.92	99.57	100.00

(6) SQUARE ROOTS OF EIGENVALUES

2.242671	1.609441	1.061811	.623400	.562464
.446362	.367773	.324655	.253790	.208292

Procento variability znaků vyjádřené příslušnou komponentou

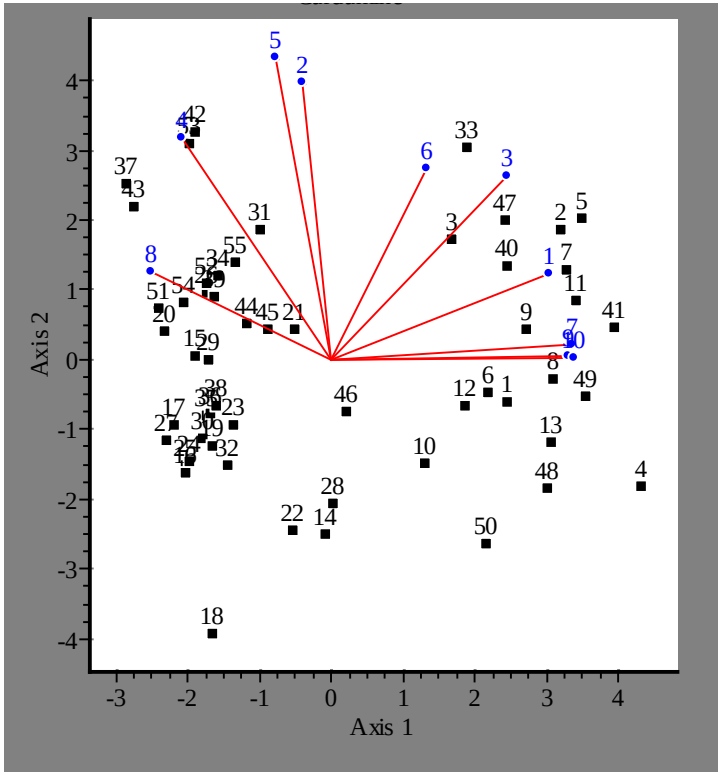
PERCENTAGE OF VARIANCE OF VARIABLES ACCOUNTED FOR BY EACH COMPONENT

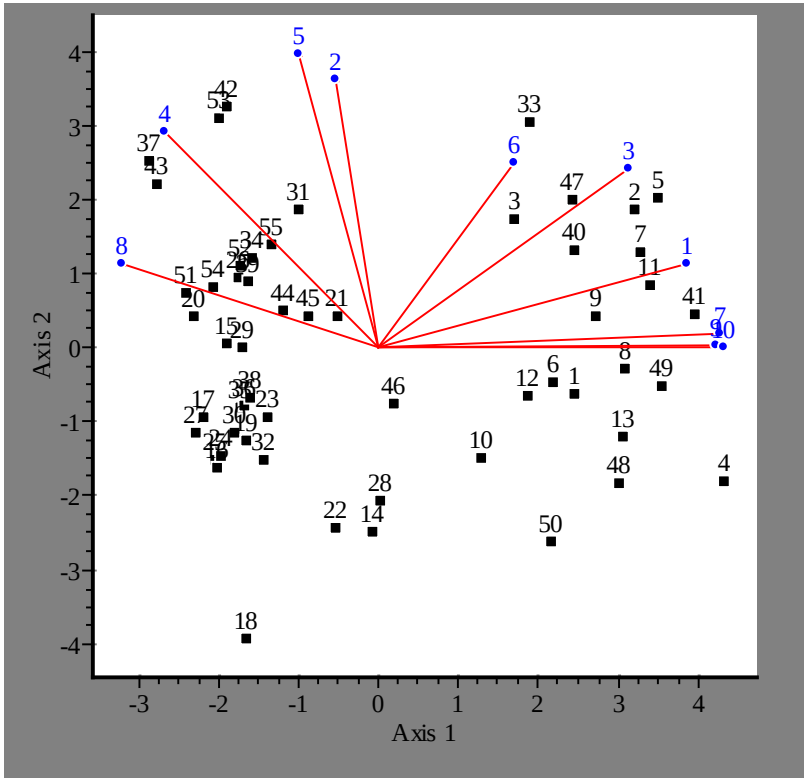
VARIABLE 1	*(šířka báze lodyhy)*		
74.352	6.290	7.935	
VARIABLE 2	*(délka nitek delších tyčinek)*		
1.395	65.706	16.284	
VARIABLE 3	*(délka kališních lístků)*		
48.986	28.849	2.002	
VARIABLE 4	*(šířka korunních lístků)*		
35.274	42.023	.322	
VARIABLE 5	*(délka korunních lístků)*		
4.887	78.187	6.603	
VARIABLE 6	*(počet květů v hlavním květenství)*		
14.201	31.260	44.392	
VARIABLE 7	*(počet lístků na lodyžních listech)*		
90.539	.204	.002	
VARIABLE 8	*(větvení lodyhy)*		
51.548	6.499	34.909	
VARIABLE 9	*(počet lodyžních listů)*		
88.628	.010	.019	
VARIABLE 10	*(nahloučení listů pod květenstvím)*		
93.148	.002	.275	

Ordinace objektů a znaků (biplot)

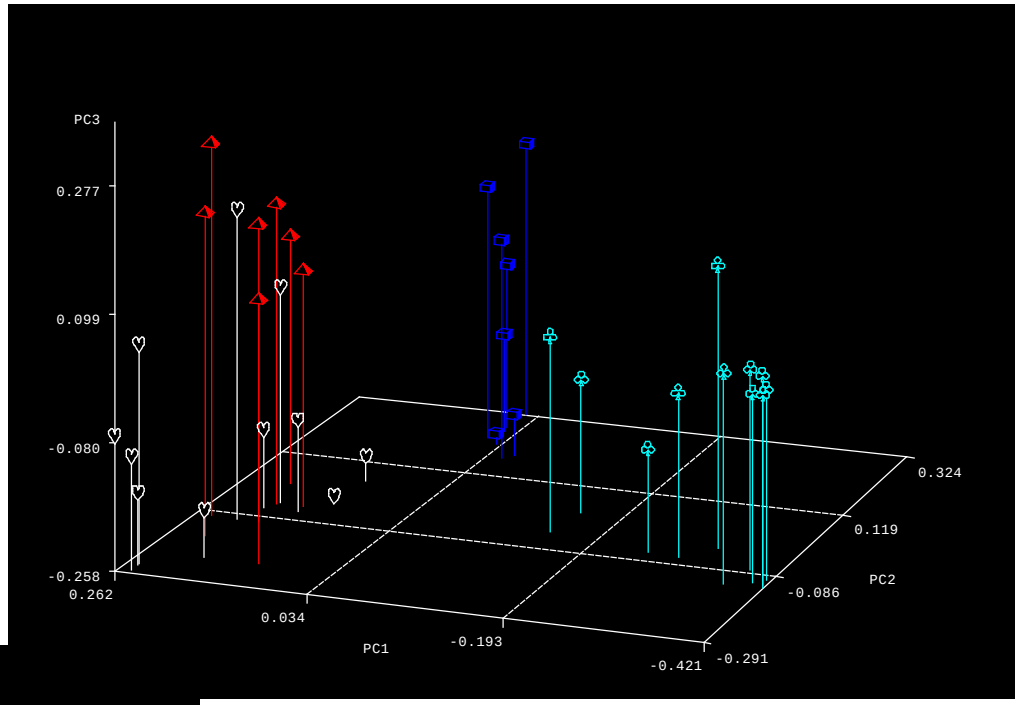
Euclidean biplot - poloha znaků vyjadřuje polohy vektorů příslušných znaků

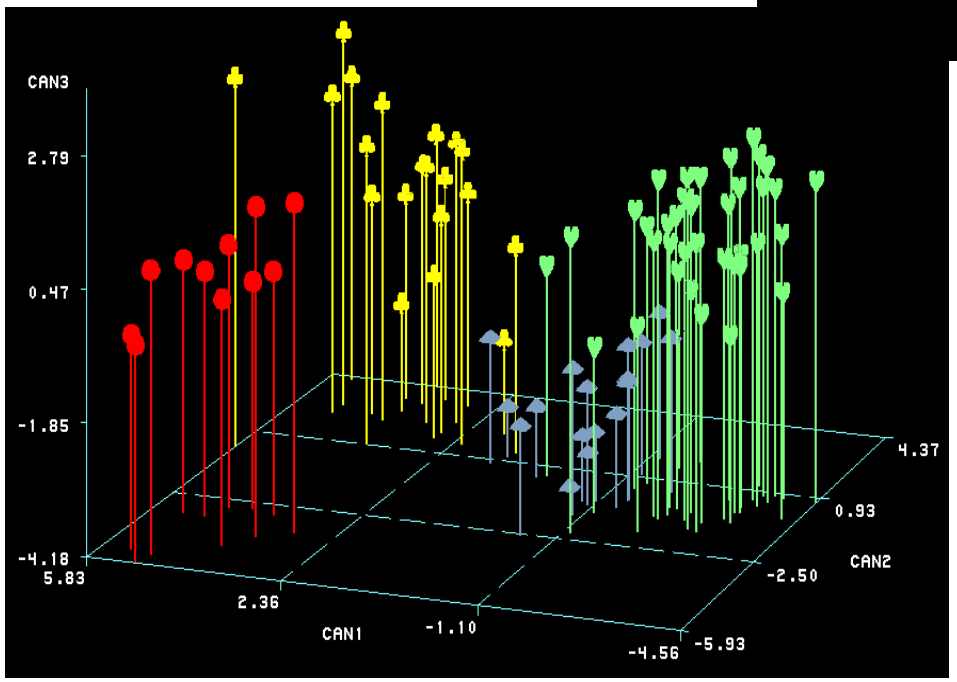
Rohlf mixed option“ - poloha znaků vyjadřuje hodnoty korelace (případně kovariance) znaků s příslušnými komponentami





Analýza hlavných komponentov





Kanonická diskriminačná analýza

Diskriminační analýza (DA)

testování hypotéz

(a) interpretace rozdílů - kanonická diskriminační analýza
(aa) zda a do jaké míry je možné odlišit stanovené skupiny objektů na základě znaků, které máme k dispozici,
(ab) které ze znaků k tomuto odlišení přispívají největší mírou.

(b) identifikace objektů - klasifikační diskriminační analýza
odvození jedné nebo více rovnic za účelem identifikace objektů

Požadavky na data:

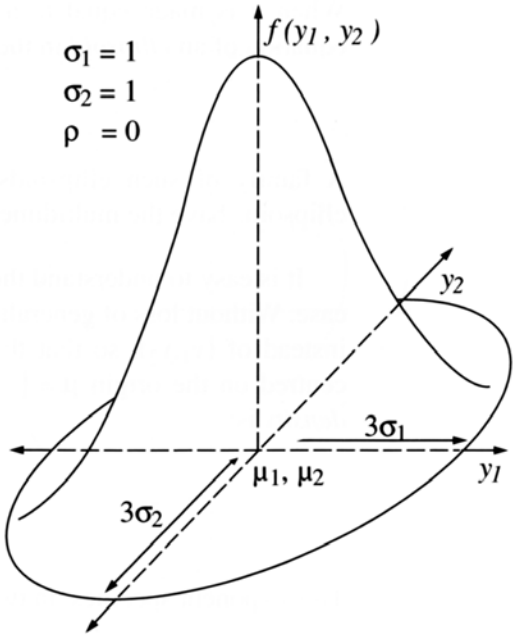
(a) kvantitativní anebo binárními znaky

(b) žádný ze znaků nesmí být lineární kombinací jiného znaku nebo jiných znaků

(c) nelze současně používat dva nebo více velmi silně korelovaných znaků

(d) kovarianční matice pro jednotlivé skupiny musí být přibližně shodné

(e) znaky charakterizující každou skupinu by měly splňovat požadavek mnohorozměrného normálního rozdělení



Pro počty skupin (g), počty znaků (p), počty objektů v skupinách a celkové počty objektů v analýze (n) v diskriminačních analýzách musí platit:

(a) musí být alespoň dvě skupiny objektů: $g \geq 2$;

(b) v každé ze skupin musí být nejméně 2 objekty;

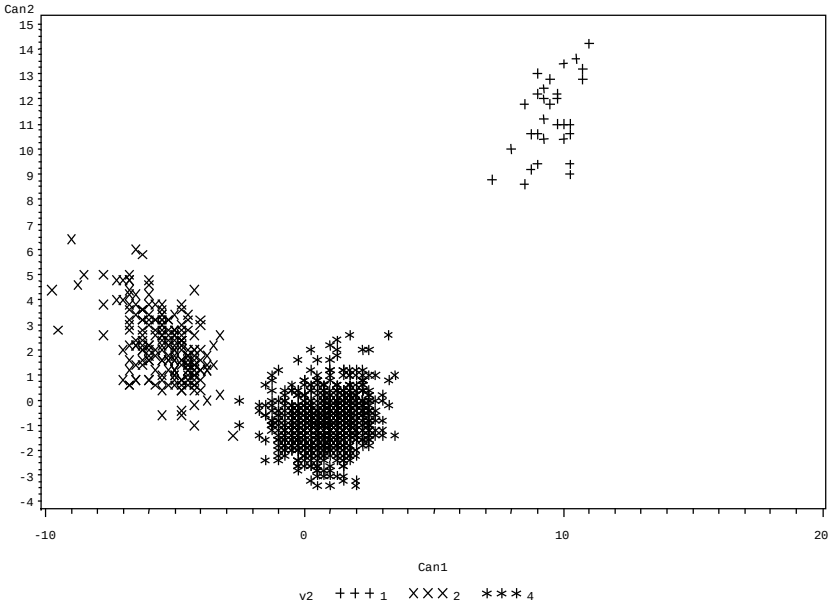
(c) počet znaků použitých v analýze musí být menší než počet objektů zmenšený o počet skupin: $0 < p < (n - g)$;

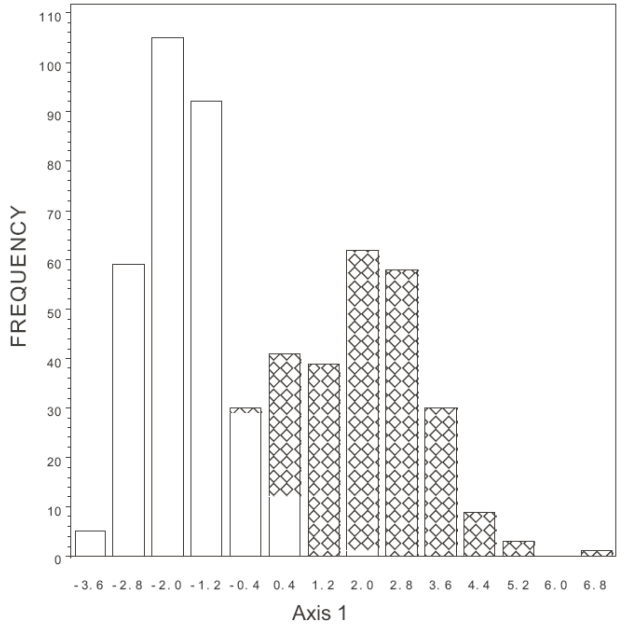
(d) žádný znak by neměl být v některé skupině konstantní

Kanonická diskriminační analýza (CDA – *canonical discriminant analysis*, *canonical variates analysis*)

umožňuje sledovat vztahy mezi objekty v prostoru definovaném kanonickými osami

ordinační procedura, která maximalizuje rozdíly mezi skupinami





Kanonická diskriminační analýza (CDA – *canonical discriminant analysis*, *canonical variates analysis*)

kanonická diskriminační funkce

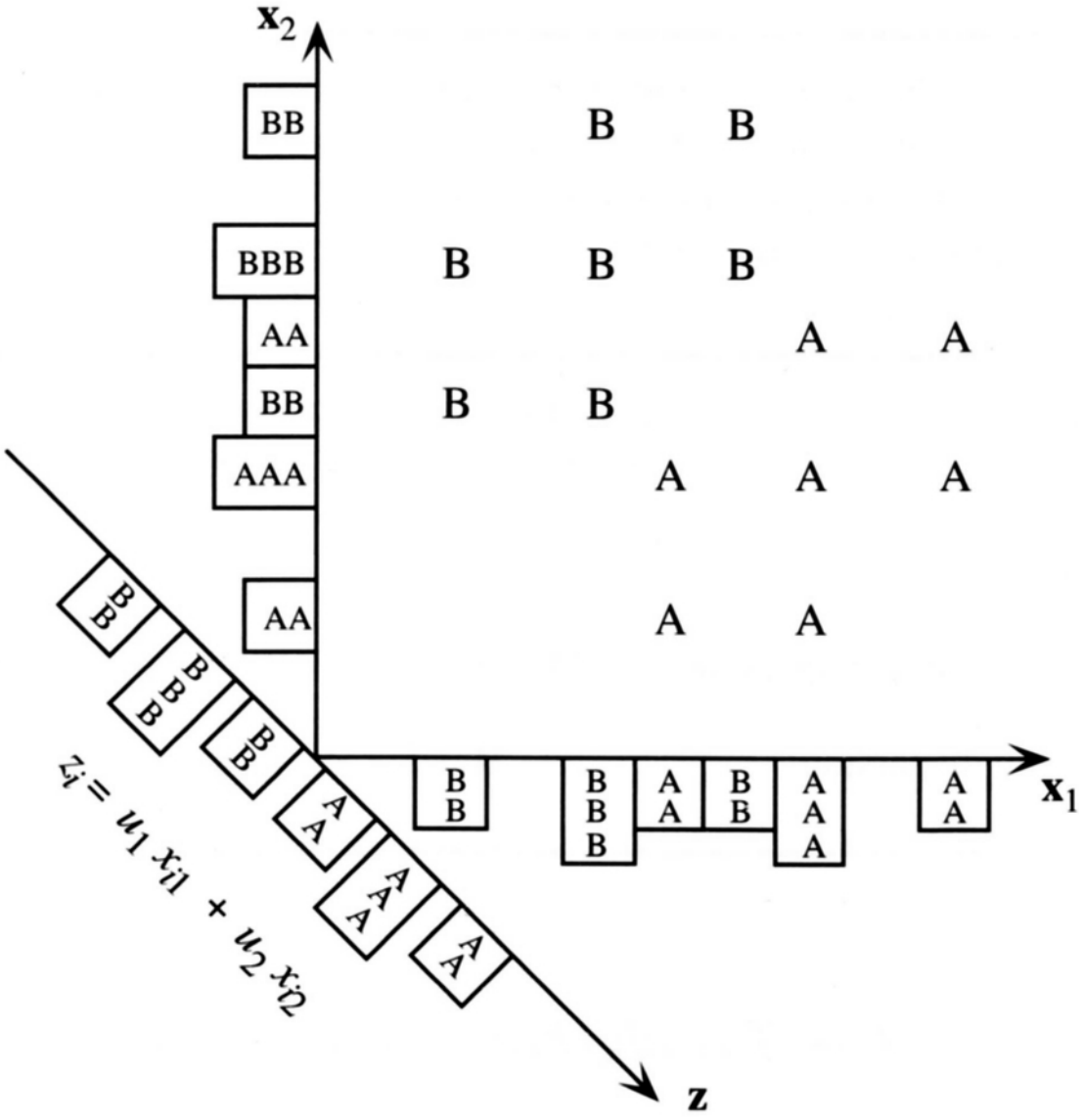
$f_{km} = a_0 + a_1x_{1km} + a_2x_{2km} + ... + a_px_{pkm}$,

f_{km} = hodnota (skóre) kanonické diskriminační funkce pro případ *m* v skupině *k*;

x_{ikm} = hodnota diskriminačního znaku x_i pro případ *m* v skupině *k*

a_i = koeficienty diskriminační funkce (i = 0, 1 ..., p);

Koeficienty (*a*) pro první funkci se odvodí tak, aby skupinové těžiště (centroidy, průměry) byly maximálně vzdálené (ve smyslu Mahalanobisovy vzdálenosti). Koeficienty vypočtené pro druhou funkci musí dále maximalizovat rozdíly mezi skupinovými centroidy a současně hodnoty obou funkcí nesmí být korelovány.


x2
BB
BBB
AA
BB
AAA
AA
B B
B B B
A A
A A A
A A
x1
z
z_i = u_1 x_i1 + u_2 x_i2

	PCA, PCoA, NMDS	DA
Předem stanovené skupiny	**ne**	**ano**
Vysvětlení maximální variability	**celkové**	**meziskupinové**
Vážení znaků	**ne**	**ano**

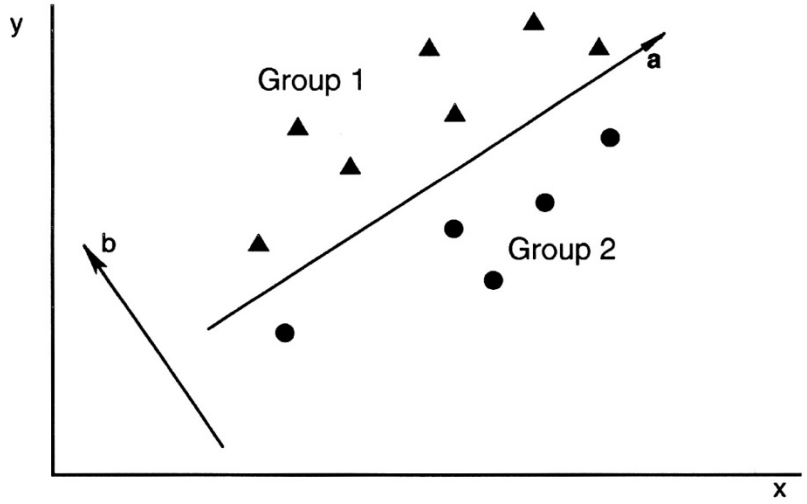
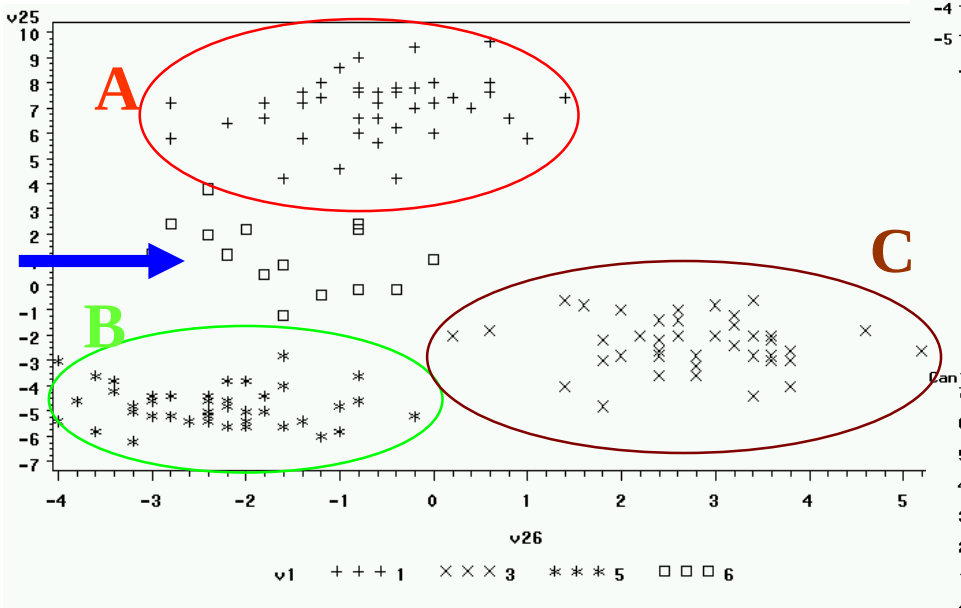
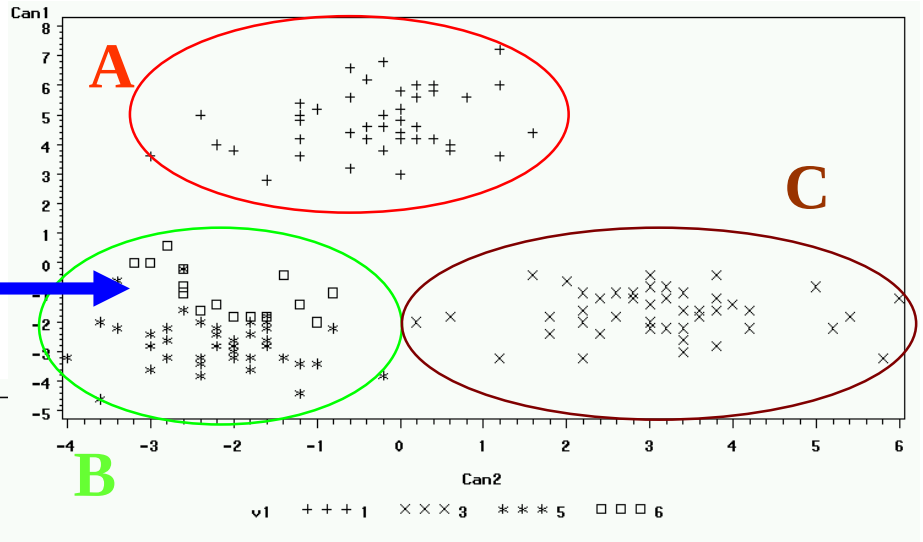


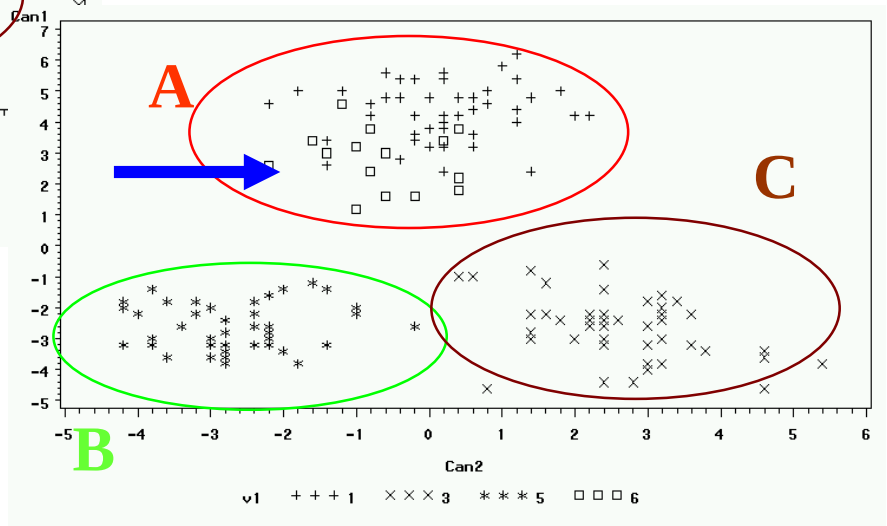
Figure 7.22. Comparison of the underlying ideas in PCA and CVA by an artificial example with two original dimensions. Component 1 (**a**) coincides with the main trend of variation in the entire sample, whereas canonical variate 1 (**b**, there is only one in this case) explains the optimum separation of the two groups.

nezařazené





zařazené do B



zařazené do A

Pozor: zařazení přechodných objektů do různých skupin může přinést různé diskr. funkce a různé výsledky

Klasifikační diskriminační analýza

(a) hledání identifikačního (klasifikačního) kritéria

skupiny objektů známého zařazení
skupinu objektů neurčitého postavení

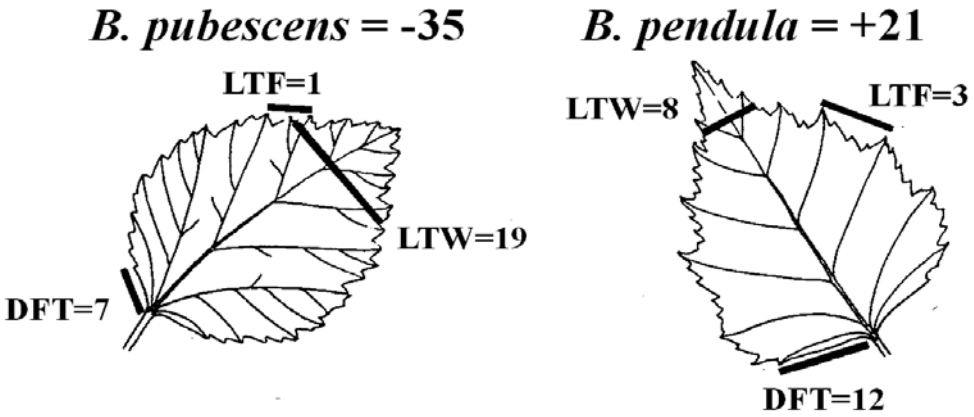
(b) zjištění účinnosti klasifikačního kritéria

resubstituce (*resubstitution*)
křížové ověření (*cross-validation*)

Účinnost klasifikačního kriteria testujeme na stejném souboru dat, z něhož se toto klasifikační pravidlo odvozuje (tento způsob testu se nazývá **resubstituce**, *resubstitution*). Pokud máme menší počet objektů, je vhodné použít tzv. **křížové ověření** (*cross-validation*): Ze souboru *n* objektů vybereme *n* – 1 objektů, které použijeme jako tréninkový soubor. Na základě tohoto tréninkového souboru odvodíme klasifikační kritérium, které potom aplikujeme na jeden vypuštěný případ. Celý postup opakujeme *n*-krát.

Způsoby odvození klasifikačního pravidla:

(1) Kanonická diskriminační funkce - objekty se klasifikují na základě jejich skóre na kanonické diskriminační funkci anebo na základě jejich projekce do kanonického prostoru

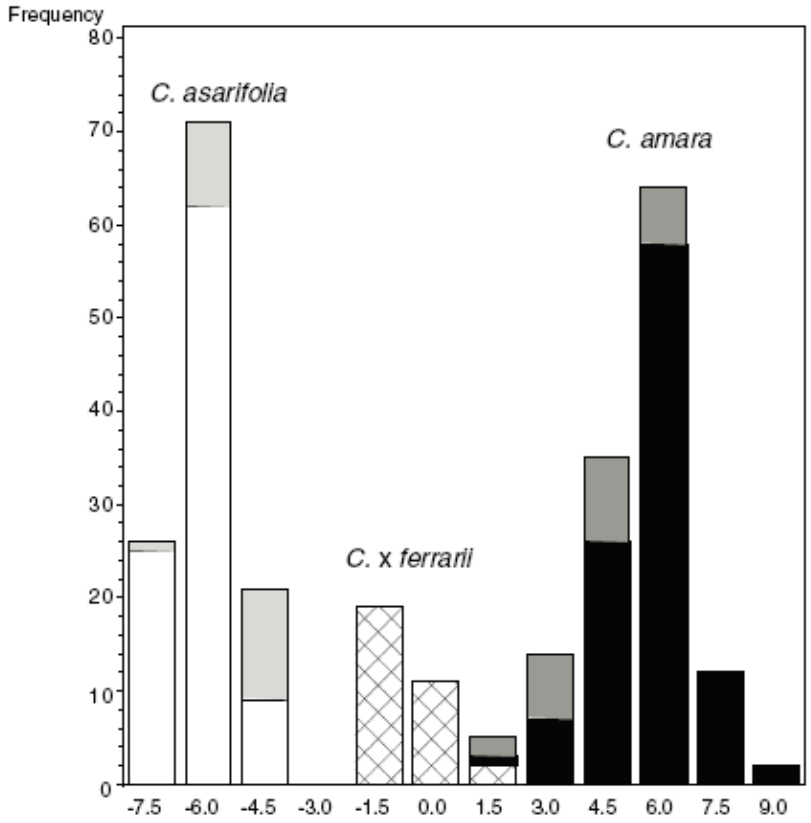


diskriminační funkce na určení druhů *Betula pubescens* a *B. pendula*
12LTF + 2DFT – 2LTW – 23
kladné hodnoty *B. pendula*, záporné hodnoty *B. pubescens*
pravděpodobnost správného určení 93%
(Stace, C. A., 1991, New Flora of the British Isles)

Způsoby odvození klasifikačního pravidla:

(1) Kanonická diskriminační funkce - objekty se klasifikují na základě jejich skóre na kanonické diskriminační funkci anebo na základě jejich projekce do kanonického prostoru

Klasifikovaný objekt se zobrazí v kanonickém prostoru spolu se souborem známých objektů (jejichž příslušnost ke skupinám je známá). Podle vzájemné pozice klasifikovaného objektu a souboru známých objektů se usuzuje na příslušnost tohoto prvku k některé skupině.



(2) výpočet lineární klasifikační funkce pro každou skupinu

Pro každou skupinu objektů se vypočítá samostatná lineární klasifikační funkce. Dále se vypočítá klasifikační skóre neznámého (klasifikovaného) objektu pro každou z těchto funkcí. Objekt bude zařazen do skupiny, pro kterou klasifikační skóre dosáhne nejvyšší hodnoty.

(3) klasifikační pravidla založená na pravděpodobnostních modelech

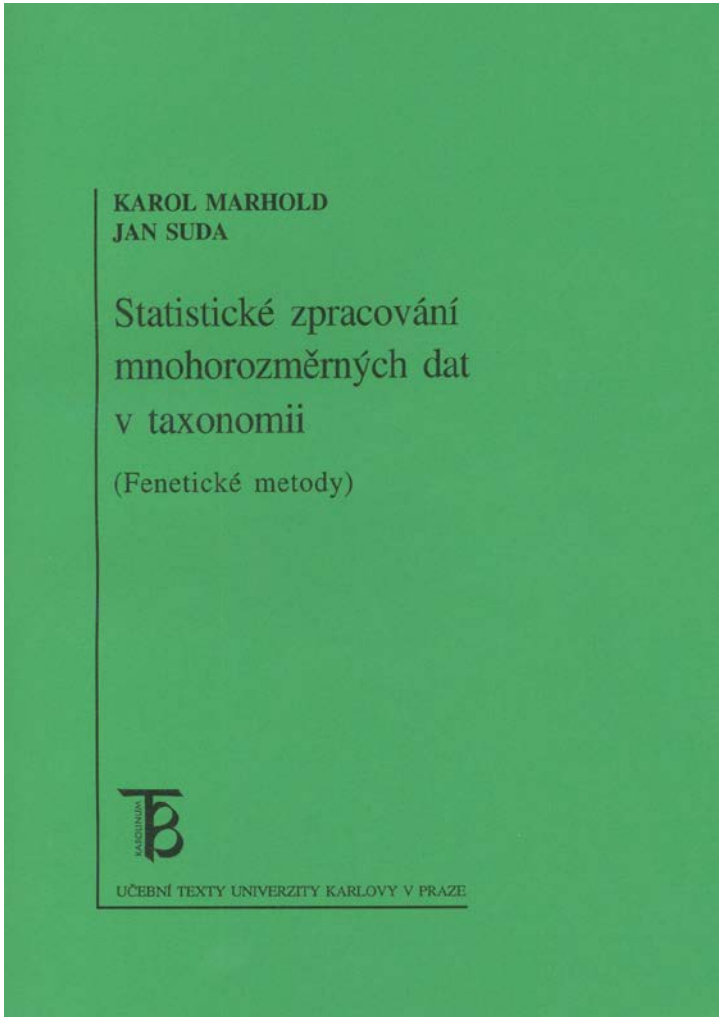
(i) lineární diskriminační funkce
(ii) kvadratické diskriminační funkce
(iii) neparametrické metody, např. *k*-nejbližších sousedů (*k-nearest neighbors*)

Klasifikační diskriminační analýza

skupina — příslušnost rostlin k stanoveným skupinám na základě klasifikačního kriteria (absolutní počet a procento rostlin klasifikovaných do jednotlivých skupin)

	amara	**austr.**	**olot.**	**opicii**	**pyren.**	Celkom
amara	**349**	20	3	1	7	380
	91.84	5.26	0.79	0.26	1.84	100.00%
austriaca	51	**302**	1	6	8	368
	13.86	**82.07**	0.27	1.63	2.17	100.00%
olotensis	2	0	**99**	0	0	101
	1.98	0.00	**98.02**	0.00	0.00	100.00%
opicii	1	9	0	**326**	42	378
	0.26	2.38	0.00	**86.24**	11.11	100.00%
pyrenaea	1	11	0	19	**207**	238
	0.42	4.62	0.00	7.98	**86.97**	

Marhold, K. & Suda, J. 2002: *Statistické zpracování mnohorozměrných dat v taxonomii*. Karolinum, Praha.

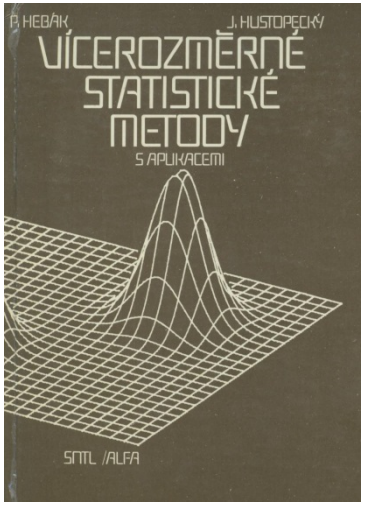


Hebák, P. & Hustopecký, J. 1987: *Vícerozměrné statistické metody s aplikacemi*. SNTL – nakladatelství technické literatury, Alfa, vydavateľstvo technickej a ekonomickej literatúry, Praha.

Hebák, P., Hustopecký, J., Jarošová, E. & Pecáková, I. 2007. *Vícerozměrné statistické metody (1)*. Ed. 2. Informatorium, Praha.

Hebák, P., Hustopecký, J. & Malá, I. 2005. *Vícerozměrné statistické metody (2)*. Informatorium, Praha.

Hebák, P., Hustopecký, J., Pecáková, I., Průša, M., Řezanková, H., Svobodová, A. & Vlach, P. 2007. *Vícerozměrné statistické metody (3)*. Ed. 2. Informatorium, Praha.

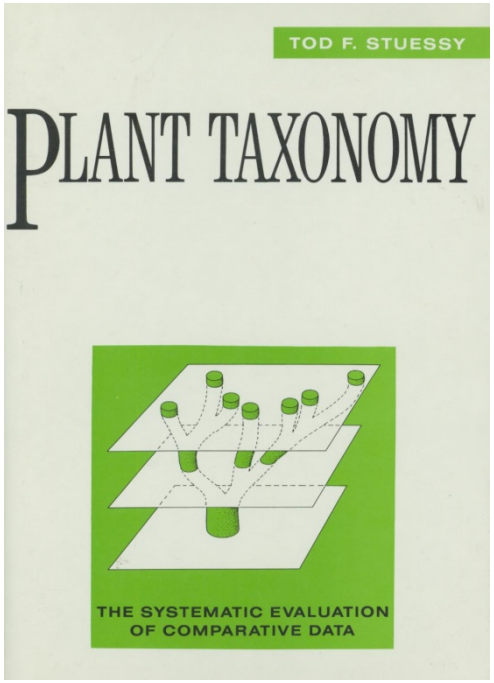


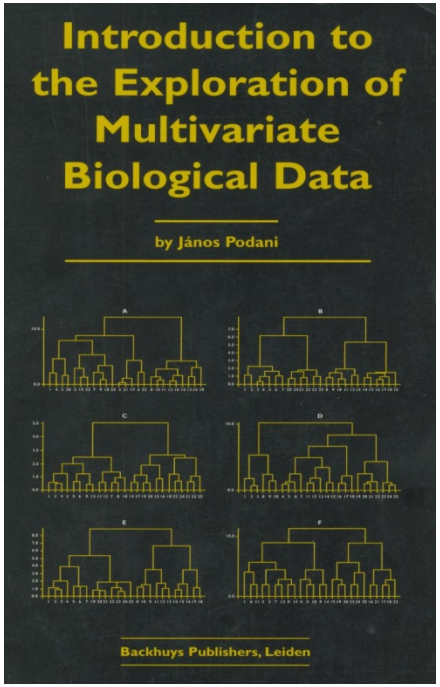
Legendre, P. & Legendre, L. 1998. *Numerical ecology*. Second English edition. Elsevier, Amsterdam.

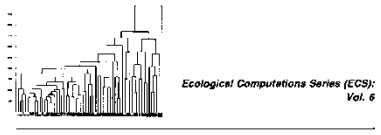
Podani, J. 1994. *Multivariate data analysis in ecology and systematics*. SPB Academic Publishing bv, The Hague.

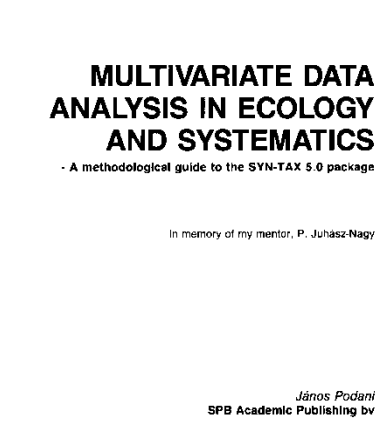
Podani, J. 2000. *Introduction to the exploration of multivariate biological data*. Backhuys Publishers, Leiden.

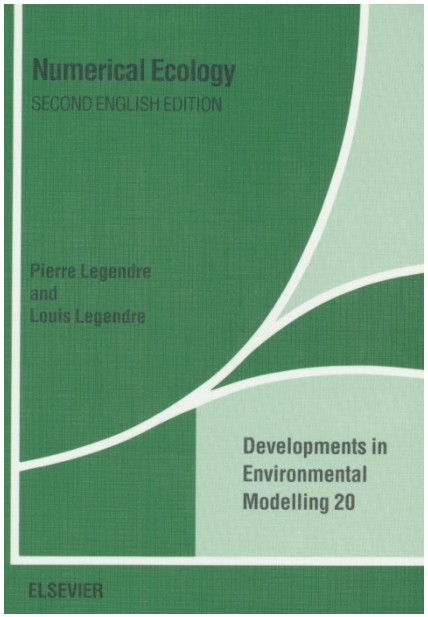
Stuessy, T. F. 1990. *Plant taxonomy: the systematic evaluation of comparative data*. Columbia University Press, New York.











Kladistický prístup

Hennig, W.

1950: *Grundzüge einer Theorie der phylogenetischen Systematik*. Deutsche Zentralverlag, Berlin.

1965: Phylogenetic systematics. *Annual Review of Entomology* 10: 97-116.

1966: *Phylogenetic systematics*. University of Illinois Press, Urbana.

Botanika:

Koponen, T., 1968: Generic revision of Mniaceae Mitt. (Bryophyta). *Ann. Bot. Fenn.* 5: 117-151.

Funk, V. & Stuessy, T. F. 1978: Cladistics for practicing plant taxonomist. *Syst. Bot.* 3: 159-178.

Bremer, K. & Wantorp, H.- E. 1978: Phylogenetic systematics in botany. *Taxon* 27: 317-329.

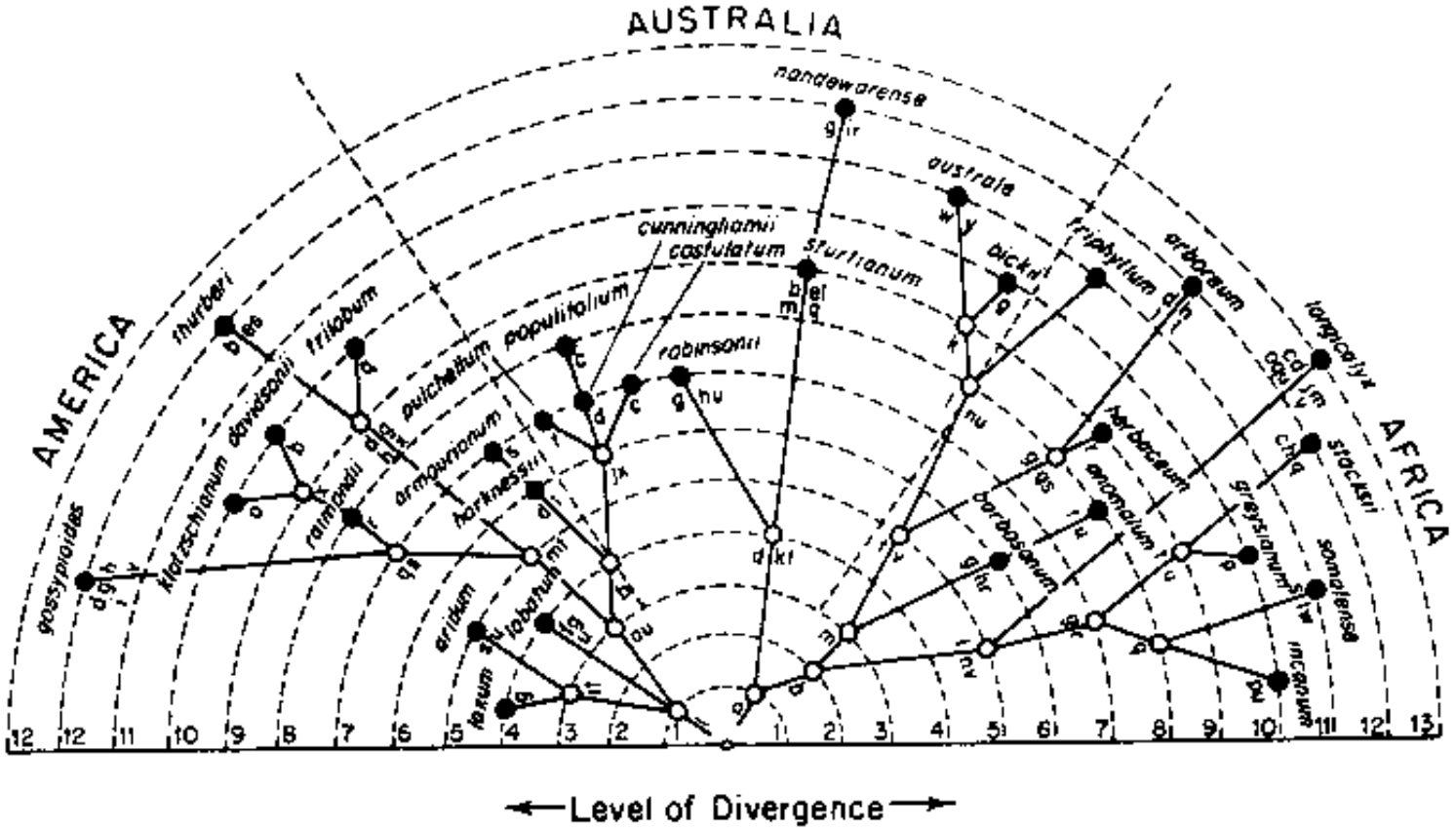


Fig. 2.17 Cladogram (Wagner tree) of 30 species of *Gossypium* (Malvaceae), modified from Frywell[142].

W.H. Wagner, University of Michigan - Groundplan/divergence method

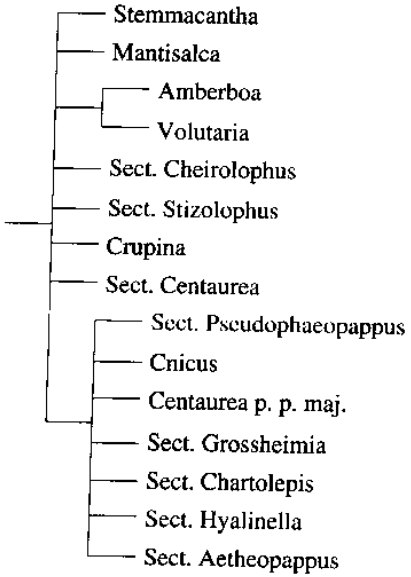


FIGURE 8-4. Strict consensus tree of six equally parsimonious cladograms of *Centaurea* sections and related genera based on cypsela characters from Dittrich (1966, pp. 138–139). The data matrix is given in Table 8-4.

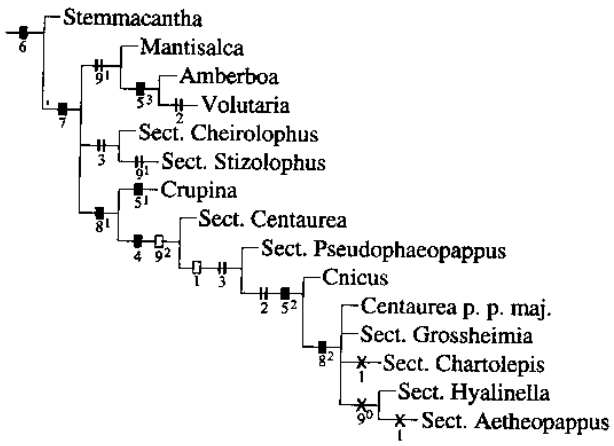


FIGURE 8-5. One of six equally parsimonious cladograms of *Centaurea* sections and related genera based on cypsela characters from Dittrich (1966, pp. 138–139). The characters are given in Table 8-3 and the data matrix in Table 8-4. Solid bars indicate nonhomoplastic synapomorphies; open bars indicate homoplastic synapomorphies with reversals; double bars indicate parallelisms; crosses indicate reversals.

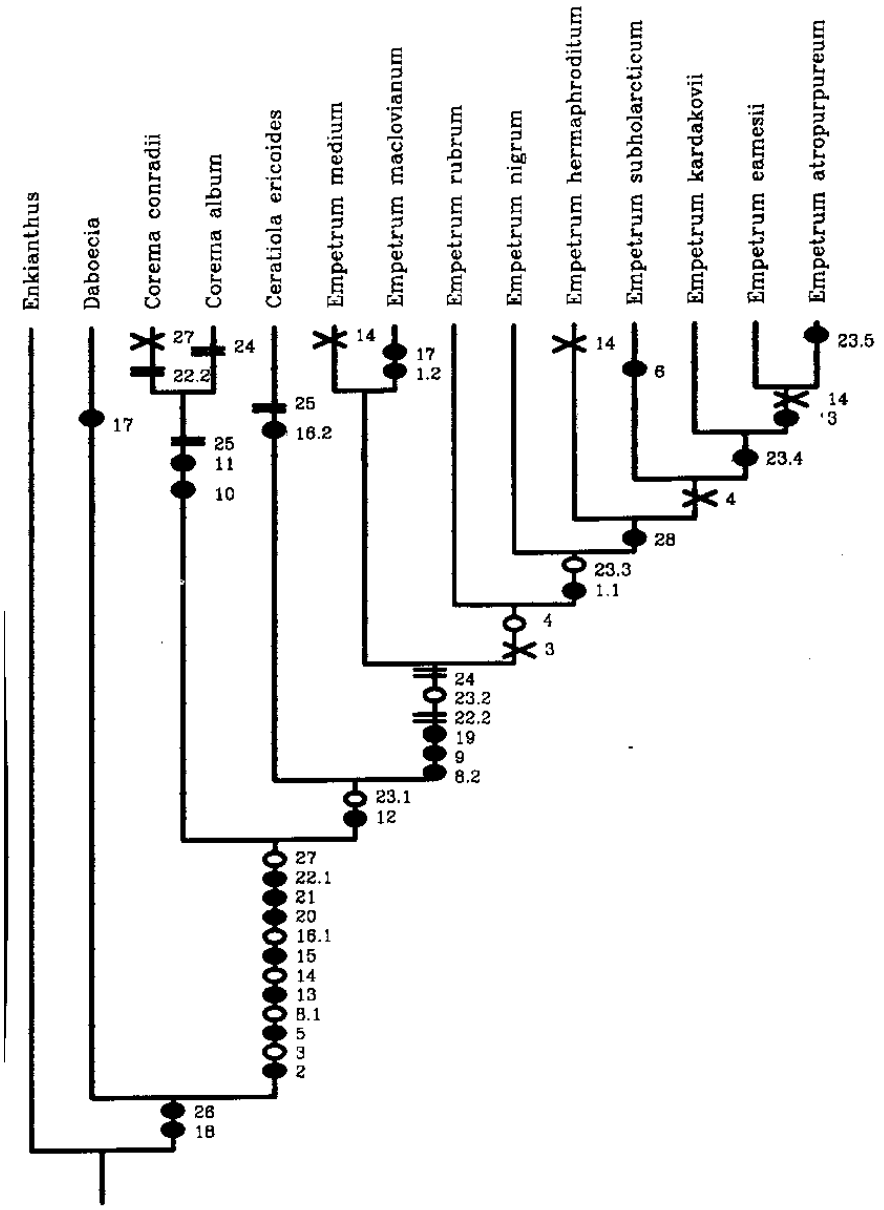


FIG. 2. One of five equally parsimonious cladograms of the Empetraceae. *Enkianthus* and *Daboecia* are outgroup taxa. Characters are numbered in accordance with the text, Appendix 1, and with Table 1. Black dots = synapomorphies (ci = 1), white dots = synapomorphies (ci < 1), parallel lines = parallelisms, crosses = reversals.

A Y-Z, X-Y-Z monofyletické skupiny

B X-Y parafyletická skupina

C X-Y polyfyletická skupina, paralelizmus

D X-Y polyfyletická skupina, konvergencia

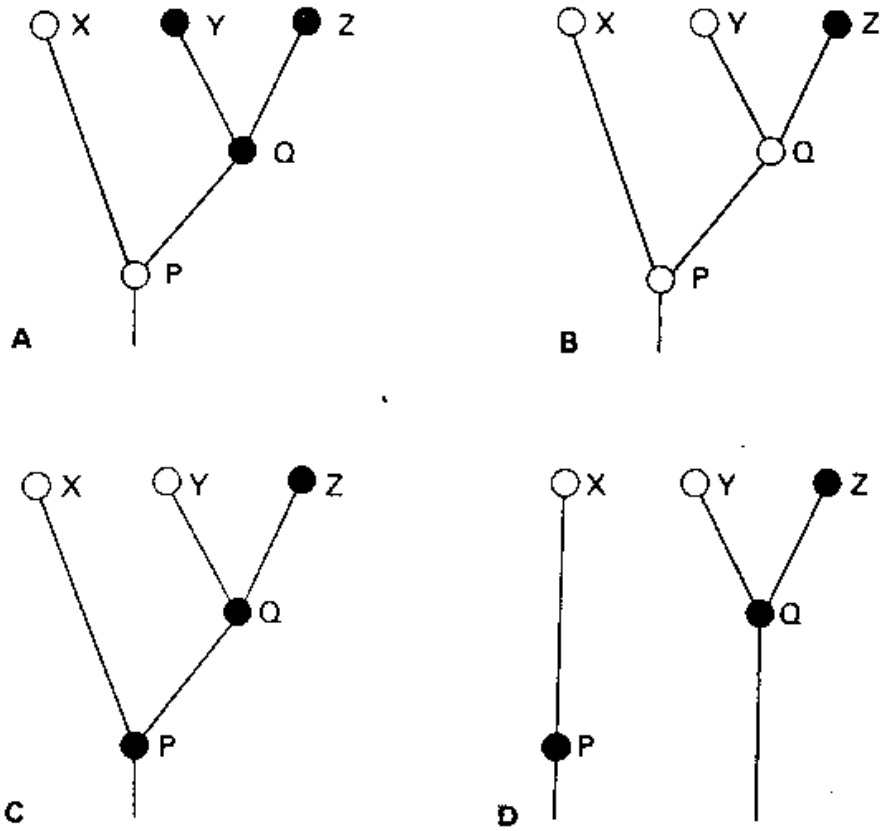
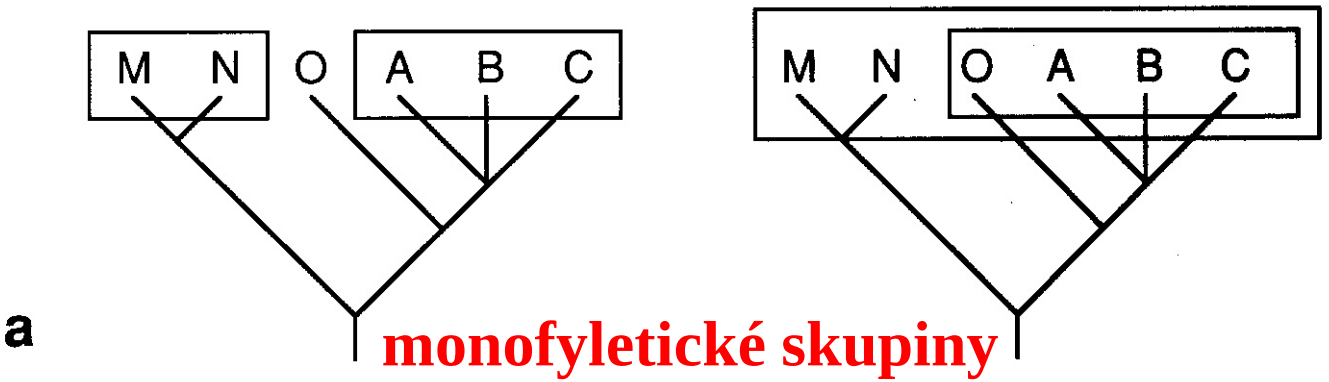
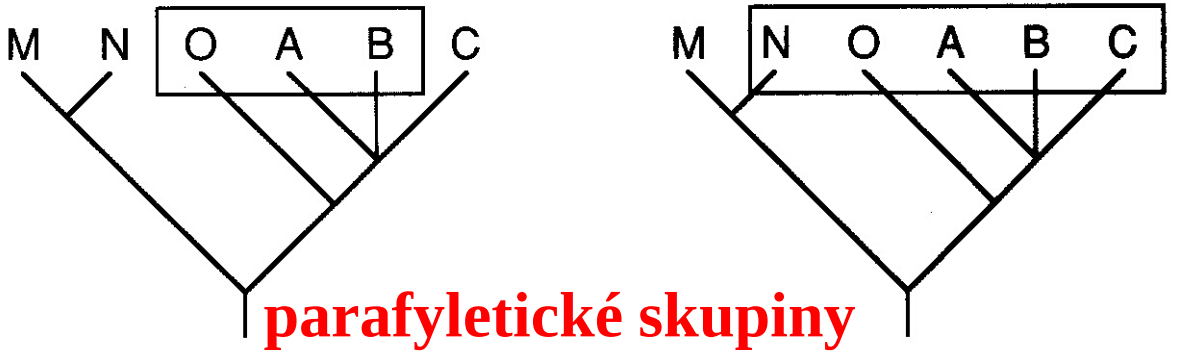


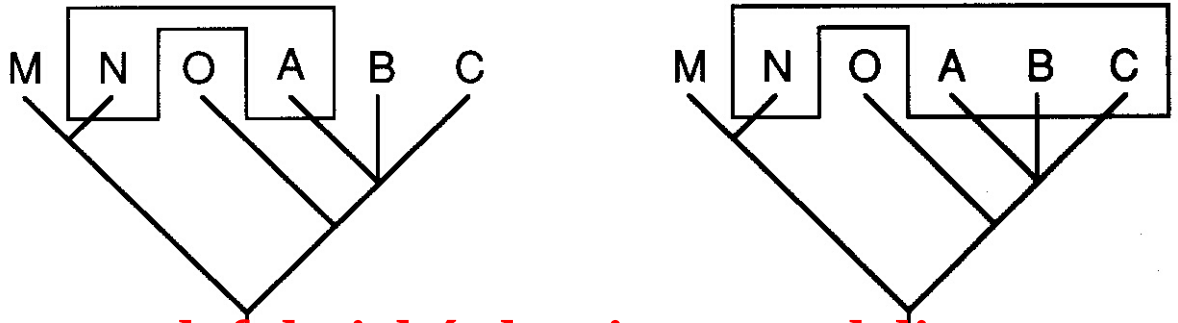
Fig. 2.6 Four diagrams showing different origins of three species (X, Y, Z) from the ancestral taxa P and Q in order to illustrate the concepts of monophyly, paraphyly, polyphyly, parallelism and convergence. The possession of one or other of two contrasting character-states by each of the five taxa is indicated by an open or closed circle respectively. **A.** Groups YZ and XYZ are both monophyletic; the similarity between Y and Z is a synapomorphy; the difference between X and YZ is due to divergence. **B.** Group XY is paraphyletic; group XYZ is monophyletic; the similarity between X and Y is a symplesiomorphy; the difference between Y and Z is due to divergence. **C.** Group XY is polyphyletic; group XYZ is monophyletic; the similarity between X and Y is a false synapomorphy caused by parallelism. **D.** Groups XY and XYZ are both polyphyletic; group YZ is monophyletic; the similarity between X and Y is a false synapomorphy caused by convergence.



a **monofyletické skupiny**



b **parafyletické skupiny**



c **polyfyletické skupiny, paralelizmus**

Primitívny stav znaku

Pleziomorfia
Sympleziomorfia

Odvodený stav znaku

Apomorfia
Autapomorfia
Synapomorfia

Homoplázia = konvergencia + paralelizmus

Mimoskupinové porovnanie (outgroup comparison)

Ingroup - študovaná skupina

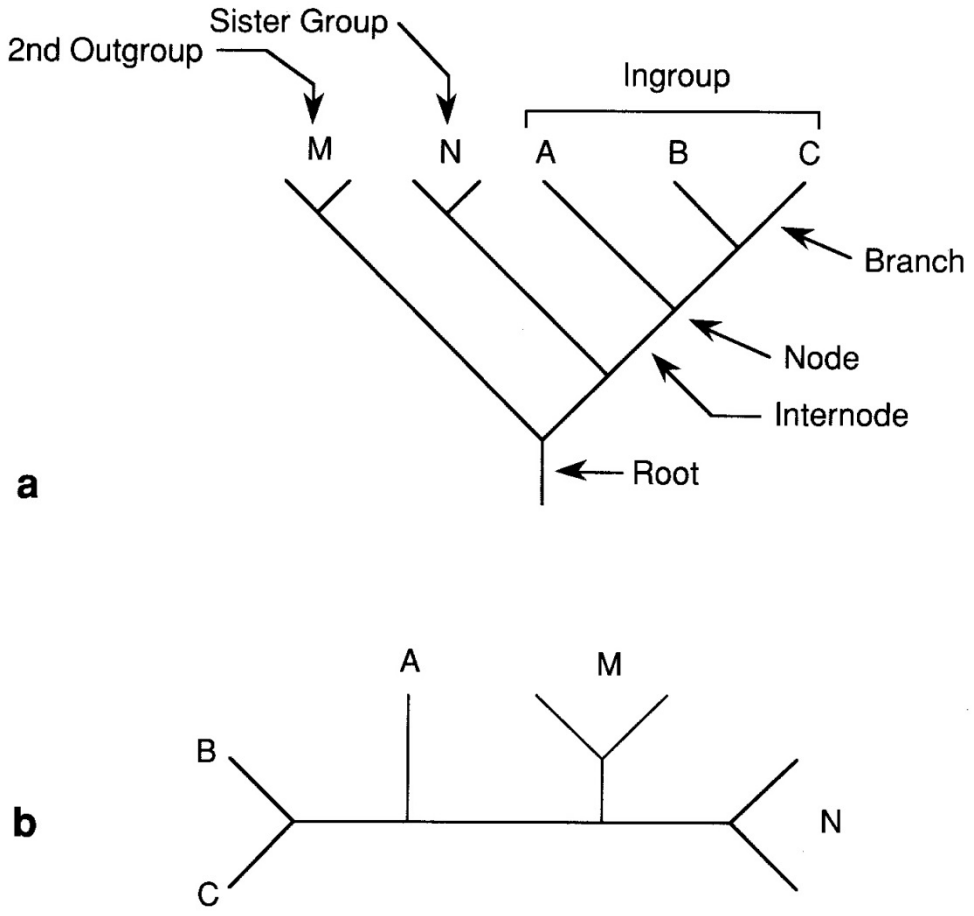
Sesterská skupina (sister group)

Mimoskupina (outgroup)

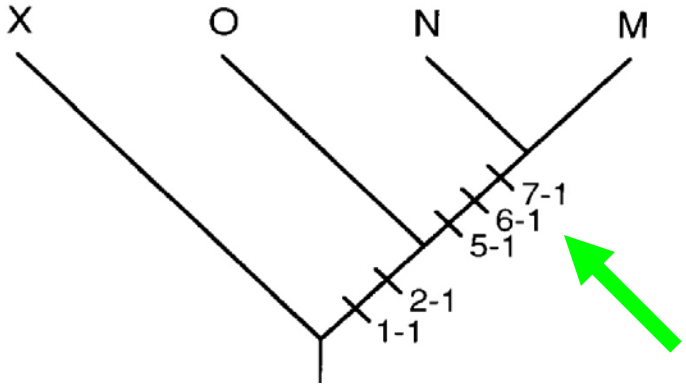
Polarizácia znakov

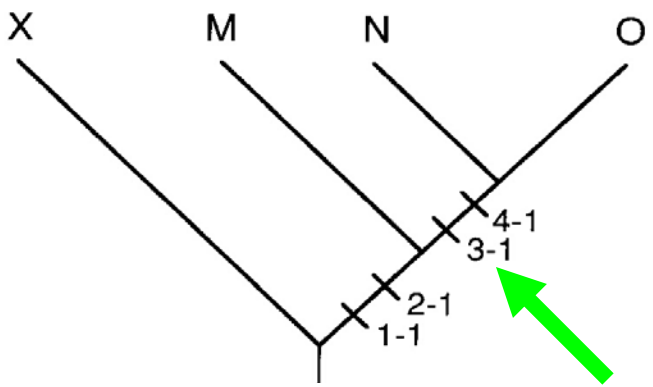
Mimoskupinové porovnania (outgroup comparison)

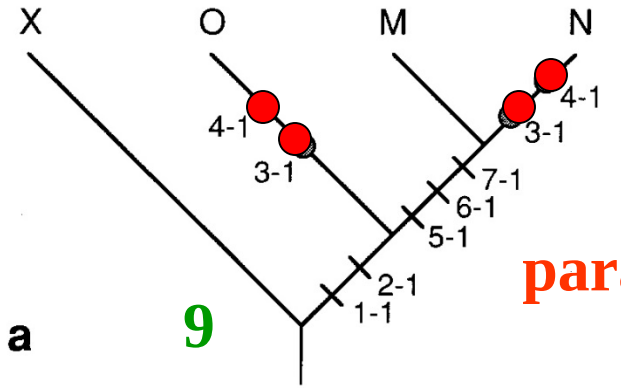
Uzol (node) - speciačná udalosť, vznik druhu



Konár (branch)
Medziuzly (internode)
Koreň (root)
Strom zakorenený - nezakorenený



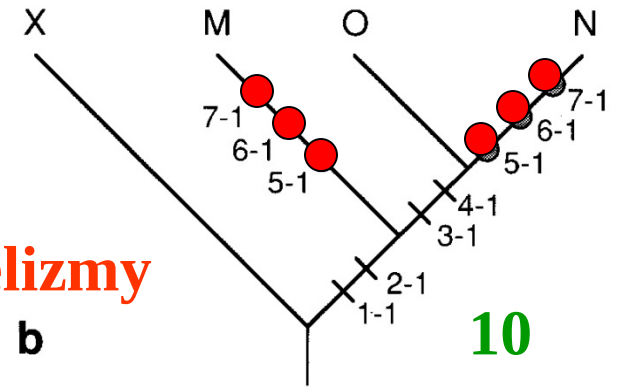




a

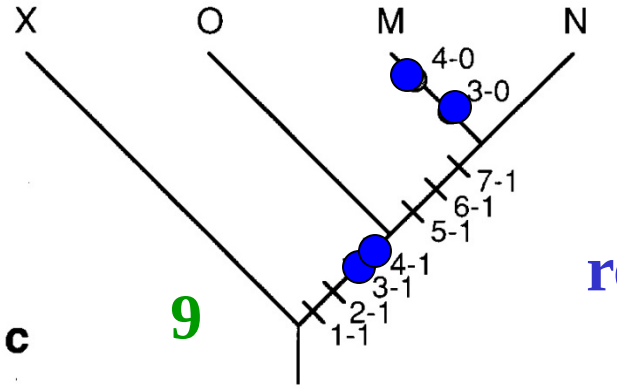
9

paralelizmy



b

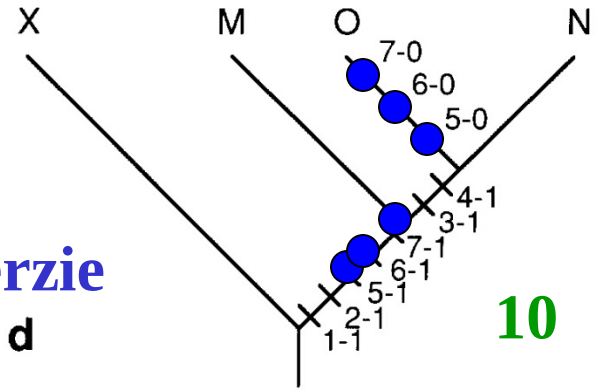
10



c

9

reverzie



d

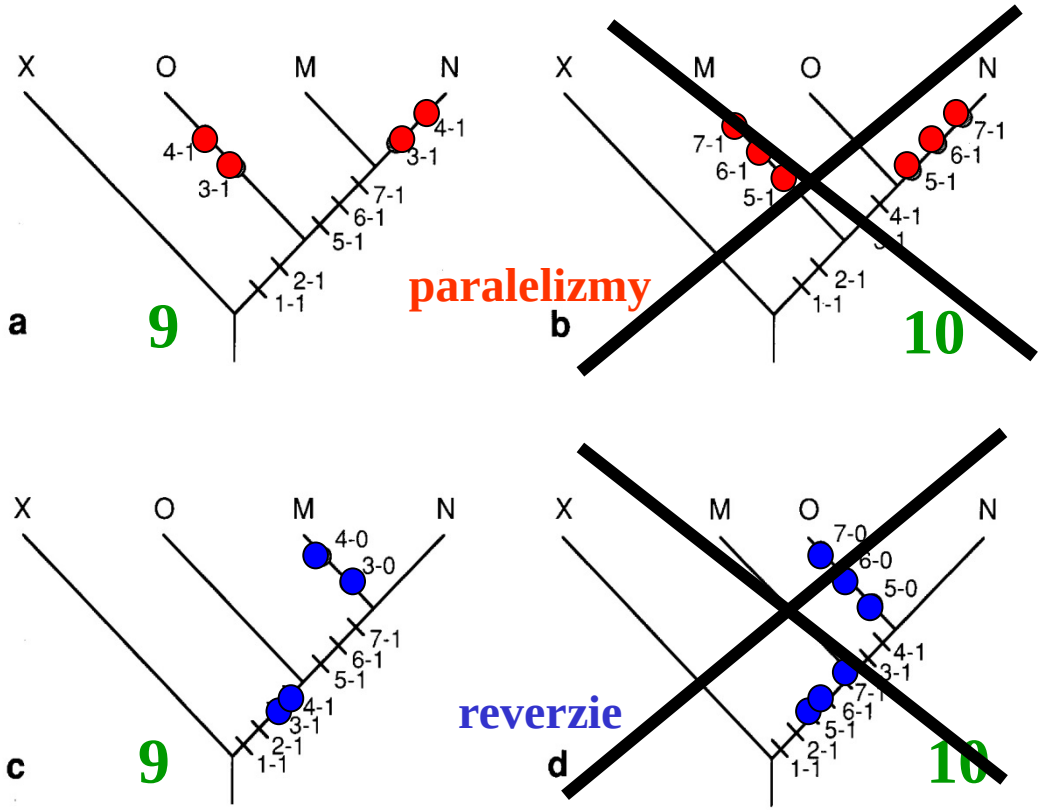
10

Parsimónia (maximum parsimony, MP)

Jednoduchá, intuitívna a logická metóda (odvodená od stredovekej logiky – uprednostňujeme najjednoduchšie riešenie), žiadna štatistika

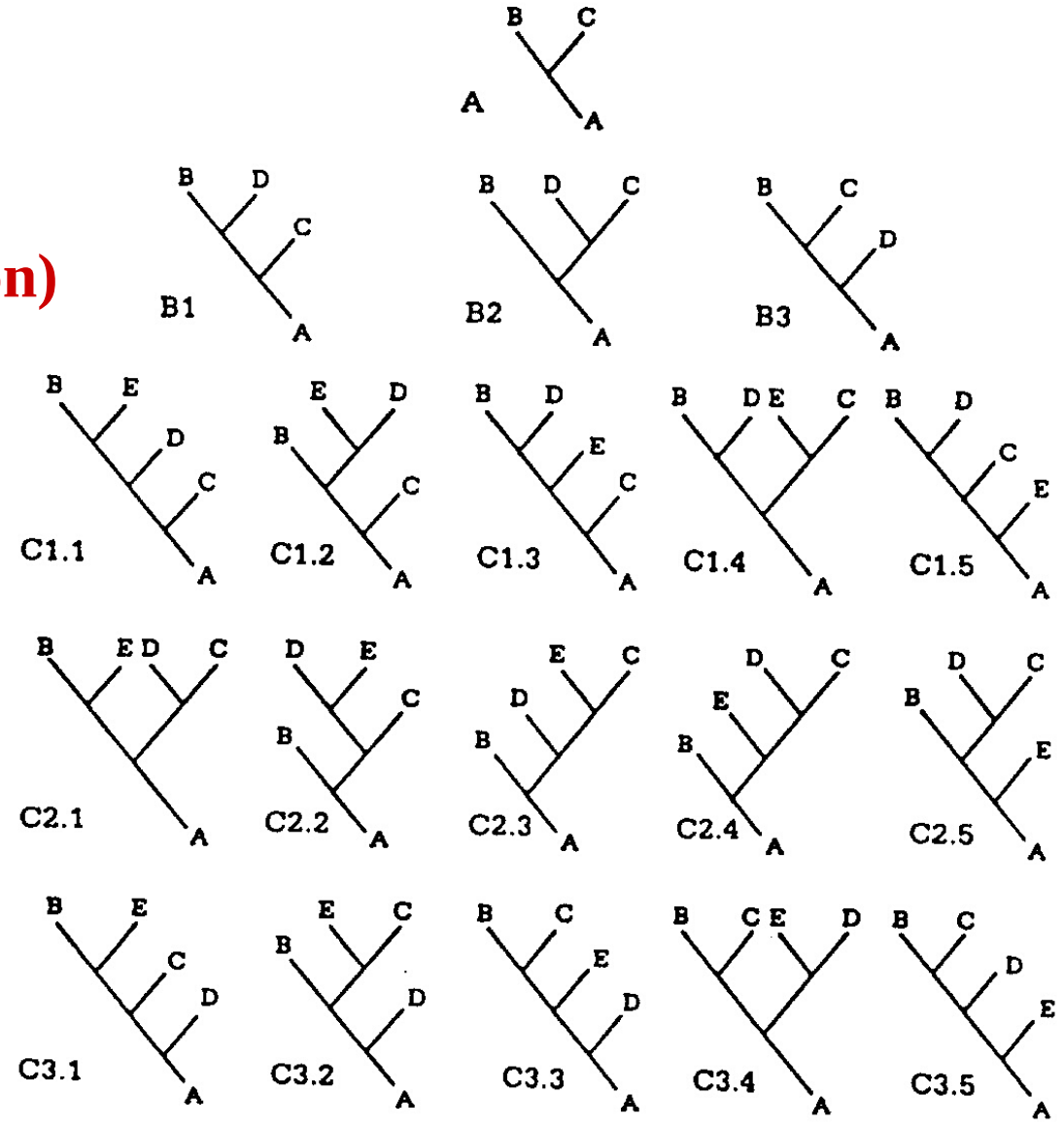
Minimalizuje ad hoc vysvetlenia – homoplázie

Maximalizuje výpovednú hodnotu

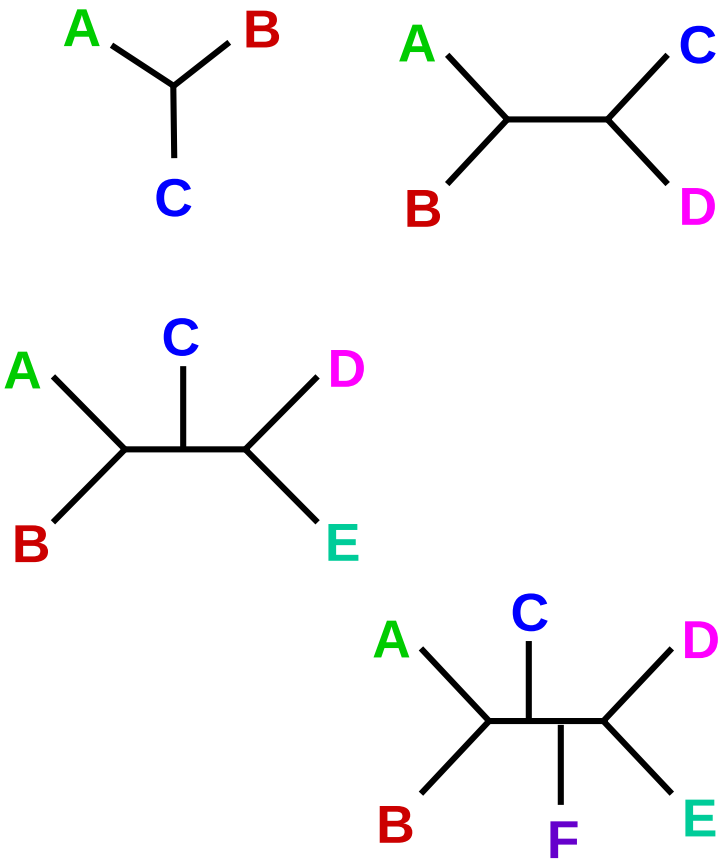


Metódy tvorby stromov

Vyčerpávajúce hľadanie (exhaustive search, implicit enumeration)

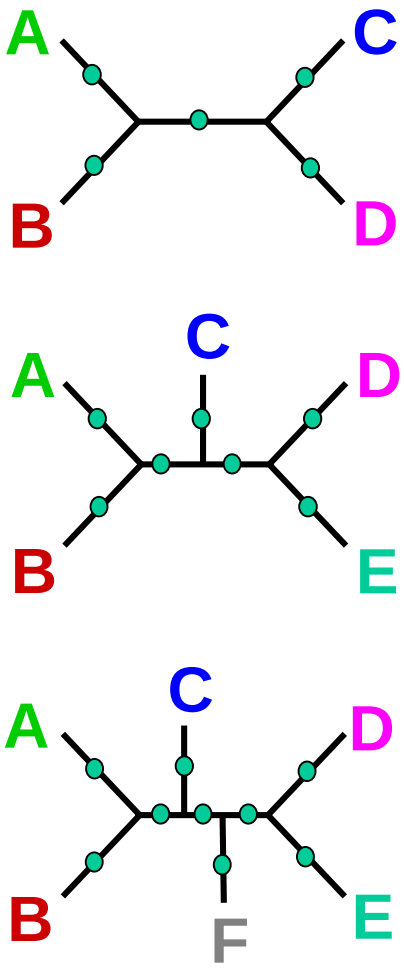


Vyčerpávajúce hľadanie má zmysel ca. do 11 taxónov



Taxóny (N)	Nezakorenené stromy
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.
.	.
30	3.58×10^{36}

Každý nezakorenený strom môže byť (teoreticky) zakorenený pozdĺž ktoréhokoľvek konára al. medziuzla



Taxóny	Nezakorenené stromy	x	Korene	=	Zakorenené stromy
3	1		3		3
4	3		5		15
5	15		7		105
6	105		9		945
7	945		11		10,395
8	10,935		13		135,135
9	135,135		15		2,027,025
.	.		.		.
.	.		.		.
30	~3.58x10^{36}		57		~2.04x10^{38}
.	.		.		.
135					2.11x10^{267}

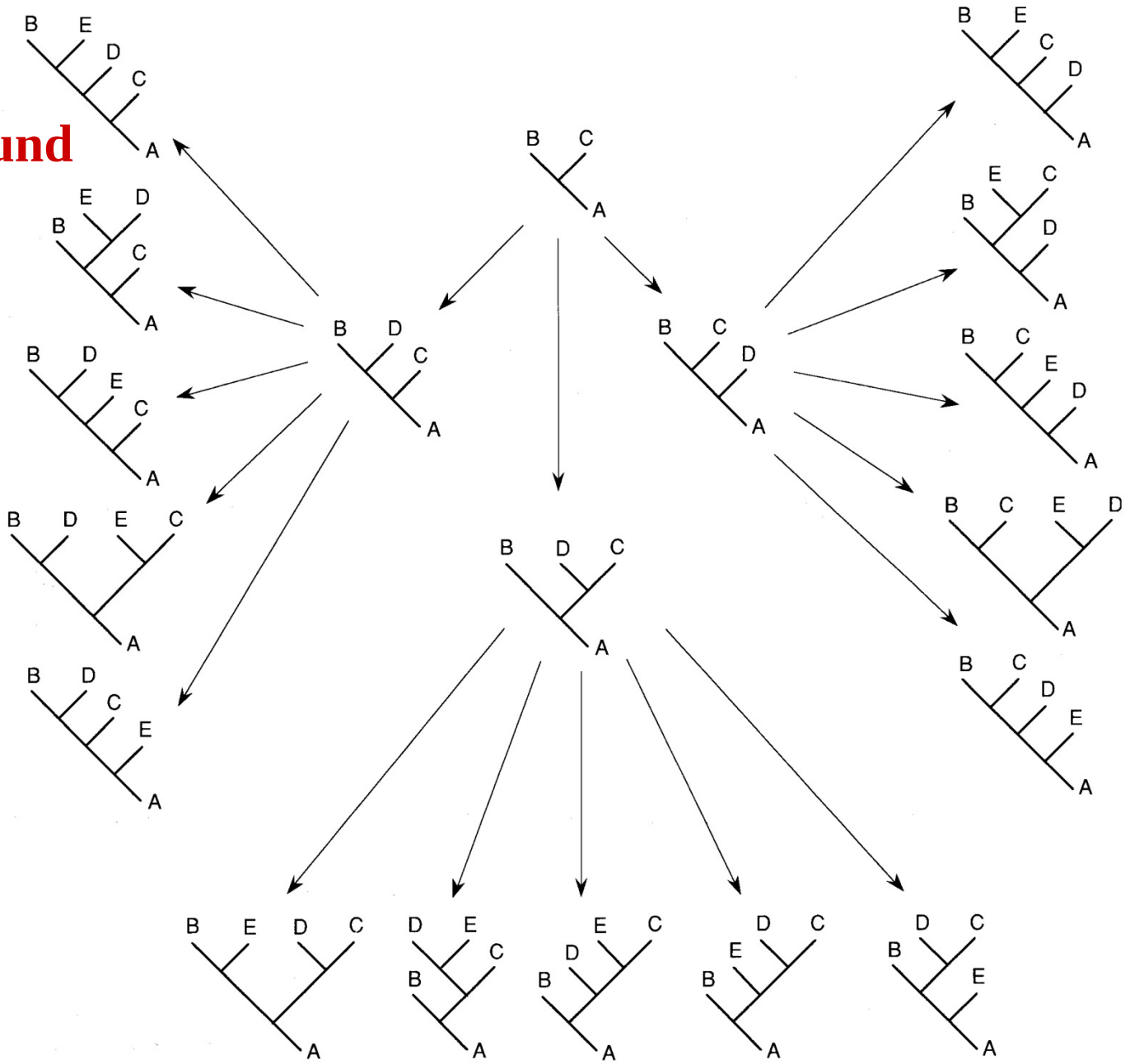
presahuje počet častíc v celom známom vesmíre!!!

Metódy tvorby stromov

Branch-and-bound (ohraničovanie vetiev)

heuristickou metódou sa nájde suboptimálny strom, ktorý slúži ako východiskové kritérium

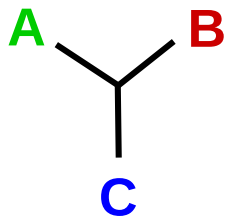
pri počte taxónov do 25



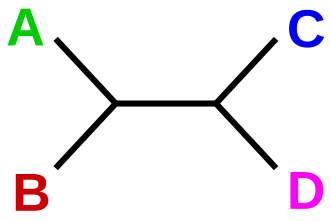
Heuristické algoritmy

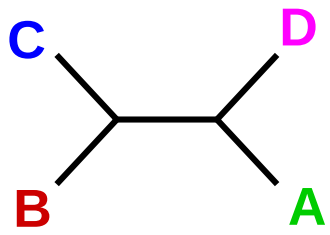
Pridávanie po krokoch (stepwise addition)

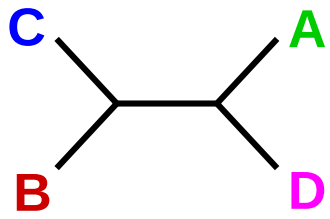
Najprv sa spoja tri objekty



Potom sa náhodne vyberie štvrtý a postupne sa pridáva k trom existujúcim vetvám (konárom, branch)

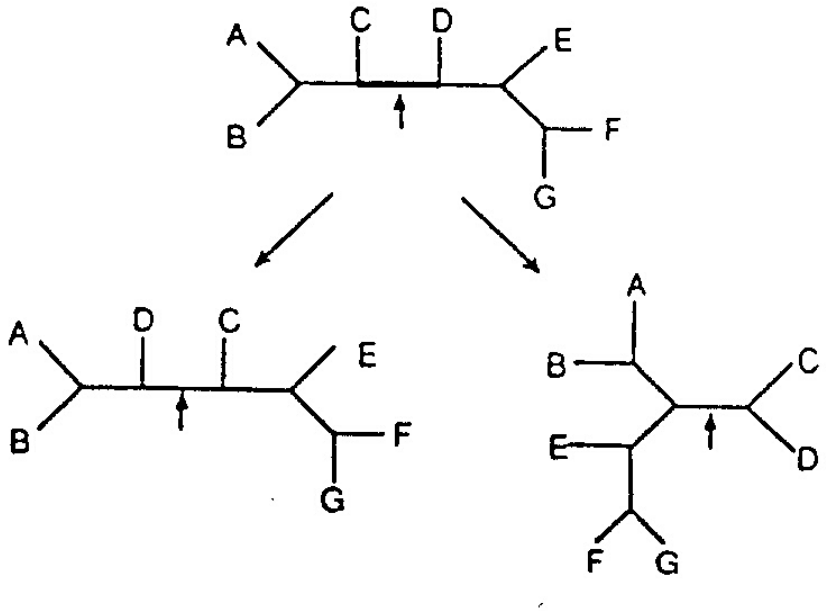






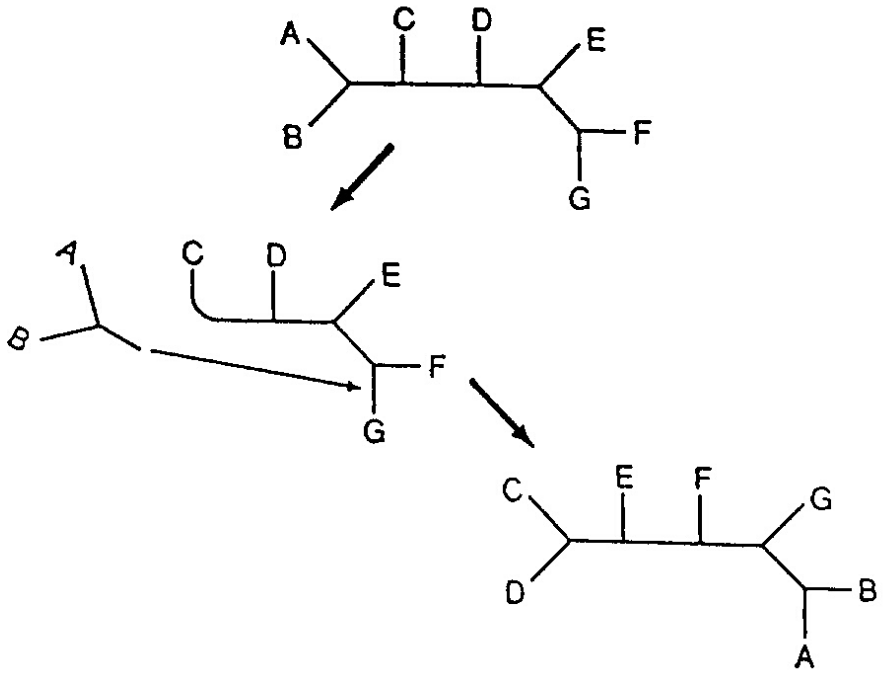
Jednotlivé stromy sa posudzujú podľa optimalizačného kritéria a jeden alebo viaceré najkratšie sa ponechajú do ďalšieho kola, kde sa pridáva piaty objekt, atď.

Výmena vetiev (branch swapping)



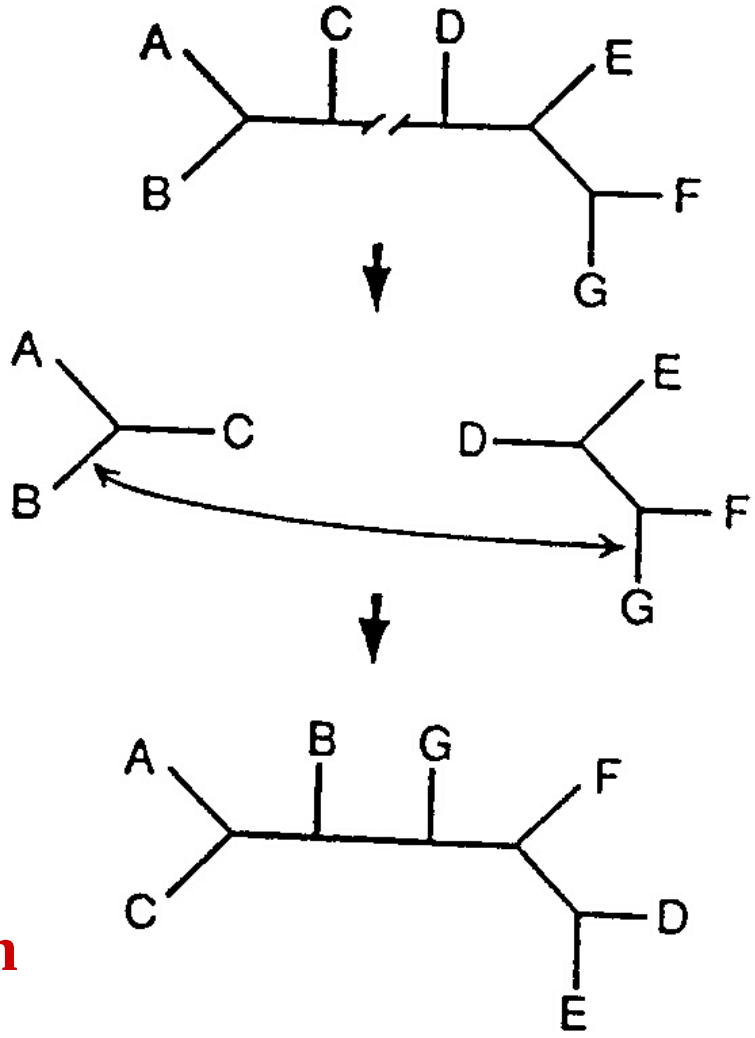
Výmena susedných objektov – nearest neighbor interchange (NNI)

Výmena vetiev (branch swapping)



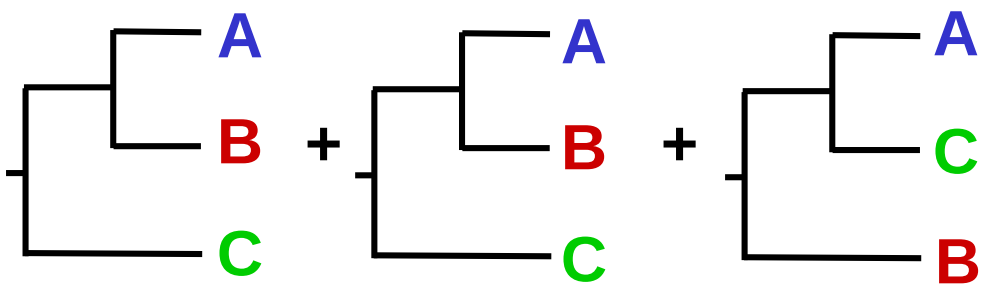
Prerezávanie vetiev (podstromov) a vrúbľovanie (roubování) – subtree pruning and regrafting (SPR)

Výmena vetiev
(branch swapping)

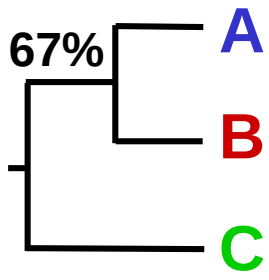


Delenie a znovuspájanie stromov –
tree bisection and reconnection (TBR)

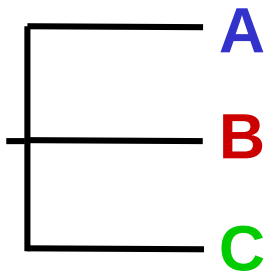
Konsenzuálne stromy (consensus trees)



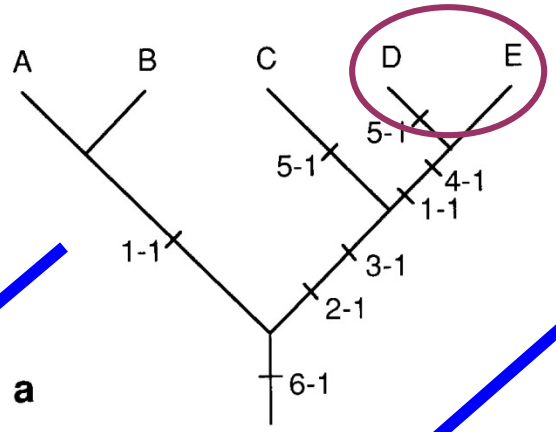
väčšinové stromy (majority-rule)



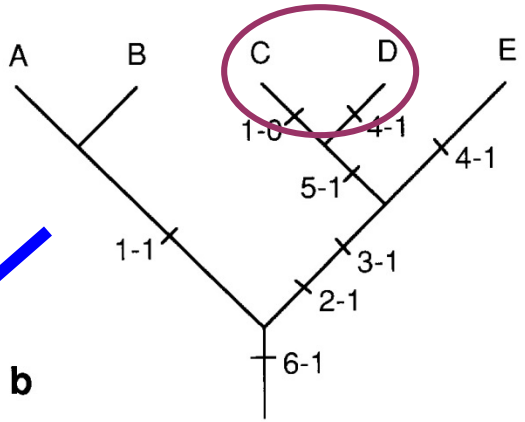
striktné konsenzuálne stromy (strict)



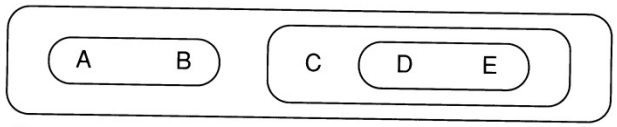
Striktné konsenzuálne stromy (strict consensus trees)



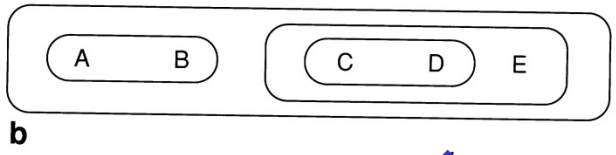
a



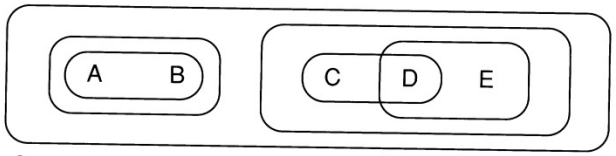
b



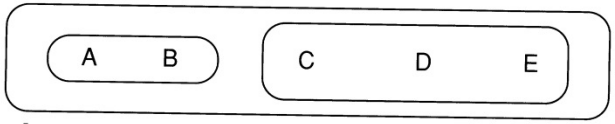
a



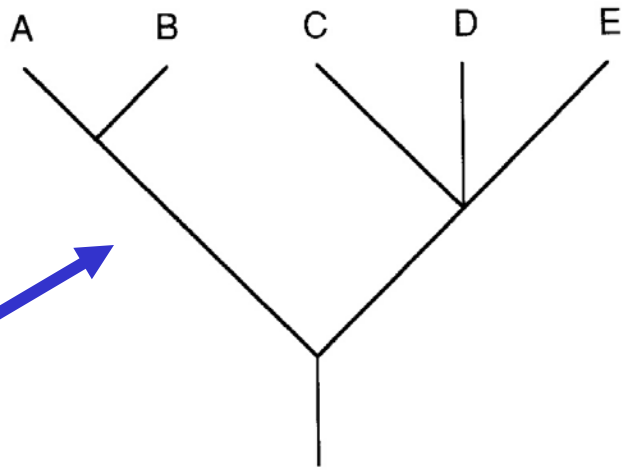
b



c



d



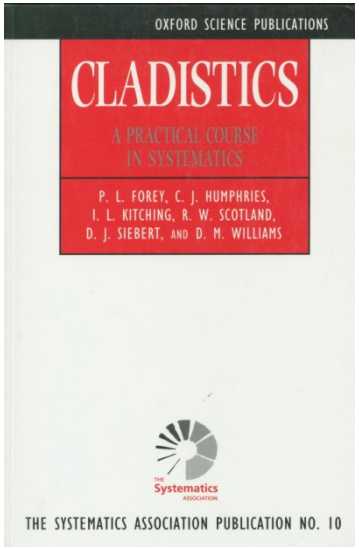
Forey, P.L., Humphries, C.J., Kitching, I.J., Scotland, R.W., Siebert, D.J. & Wiliams, D.M., 1992. *Cladistics. A practical course in systematics.* Clarendon Press, Oxford.

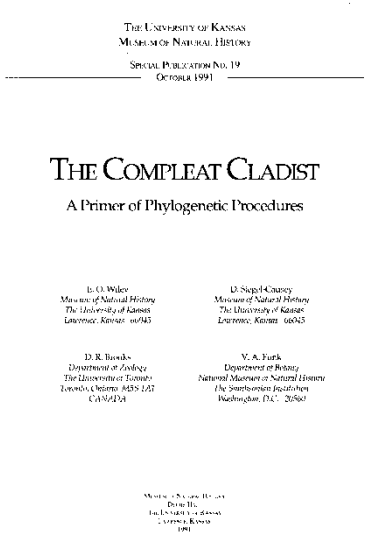
Kitching, I.J., Forey, P.L., Humphries, C.J. & Williams, D.M., 1998. *Cladistics. The theory and practice of parsimony analysis*. Ed. 2. Oxford University Press, Oxford.

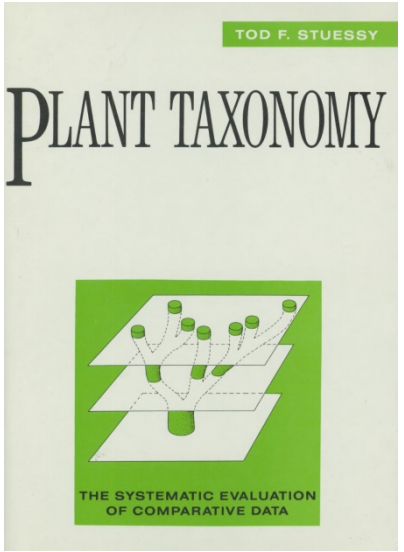
Stuessy, T. F. 1990. *Plant taxonomy: the systematic evaluation of comparative data*. Columbia University Press, New York.

Wiley, E.O., Siegel-Causey, D., Brooks, D.R. & Funk, V.A. 1991. *The compleat cladist, a primer of phylogenetic procedures*. The University of Kansas, Museum of Natural History, Lawrence.

K dispozícii na www stránke: http://nhm.ku.edu/cc.html

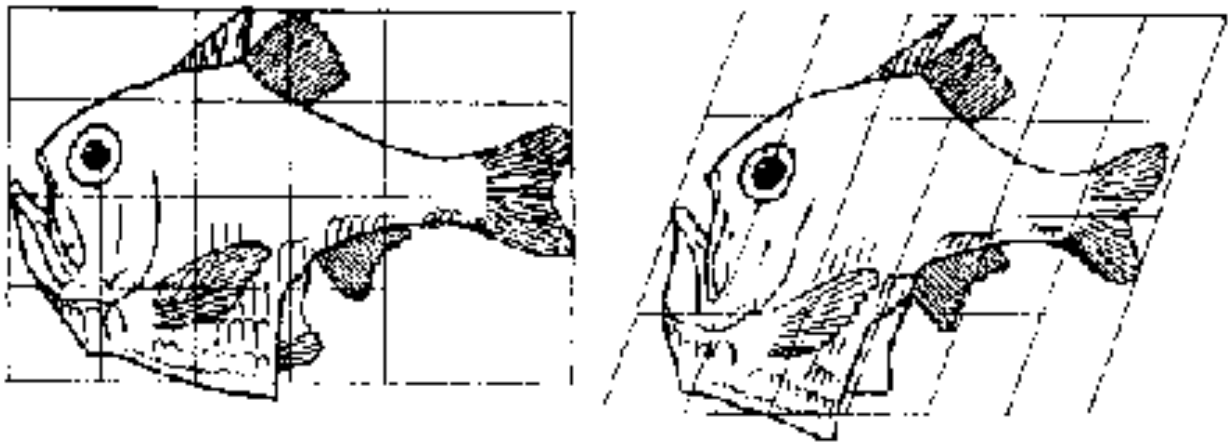






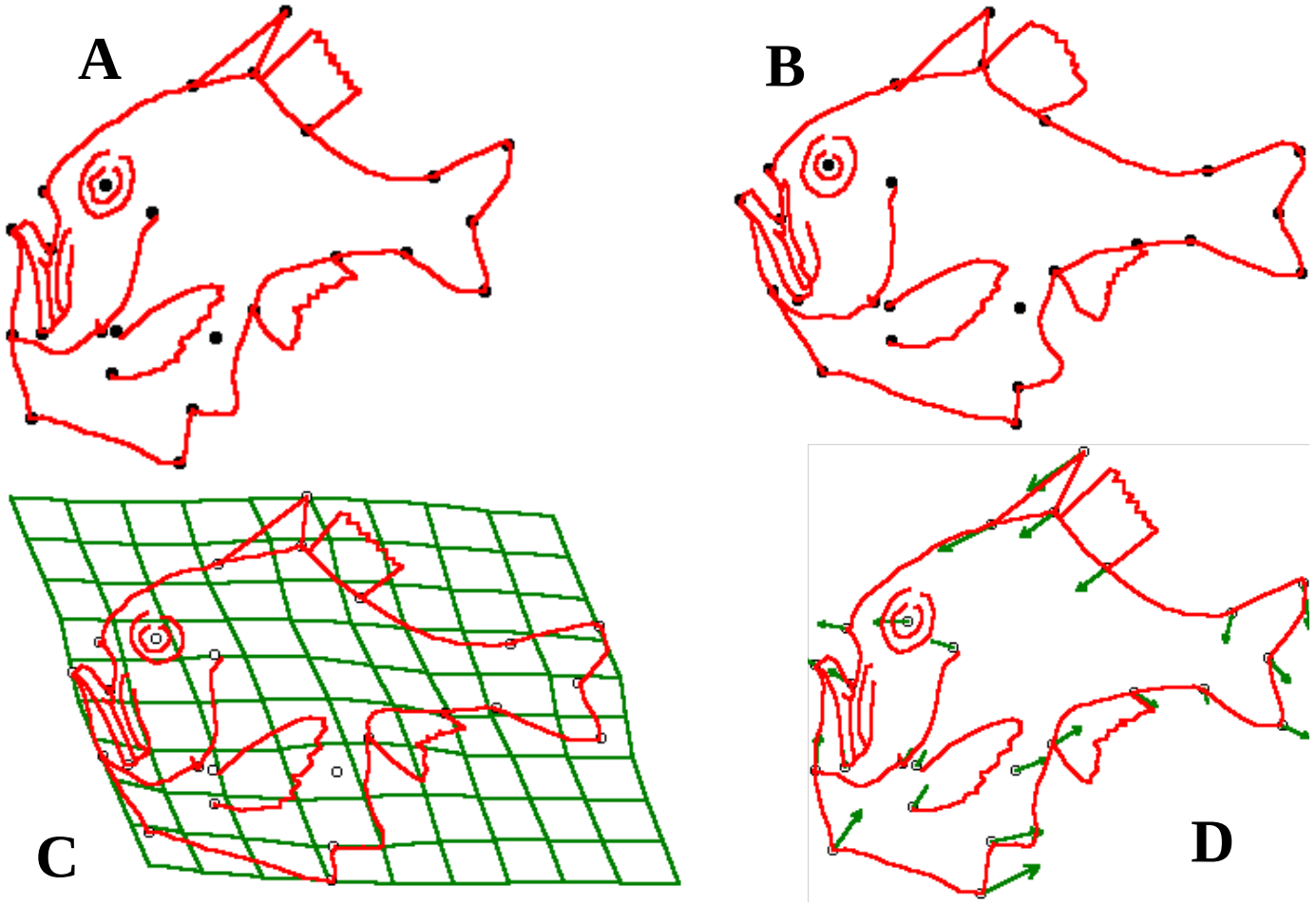
Geometrická morfometrika

Thompson, A. W. 1917. ***On growth and form.*** **Cambridge University Press, Cambridge.**



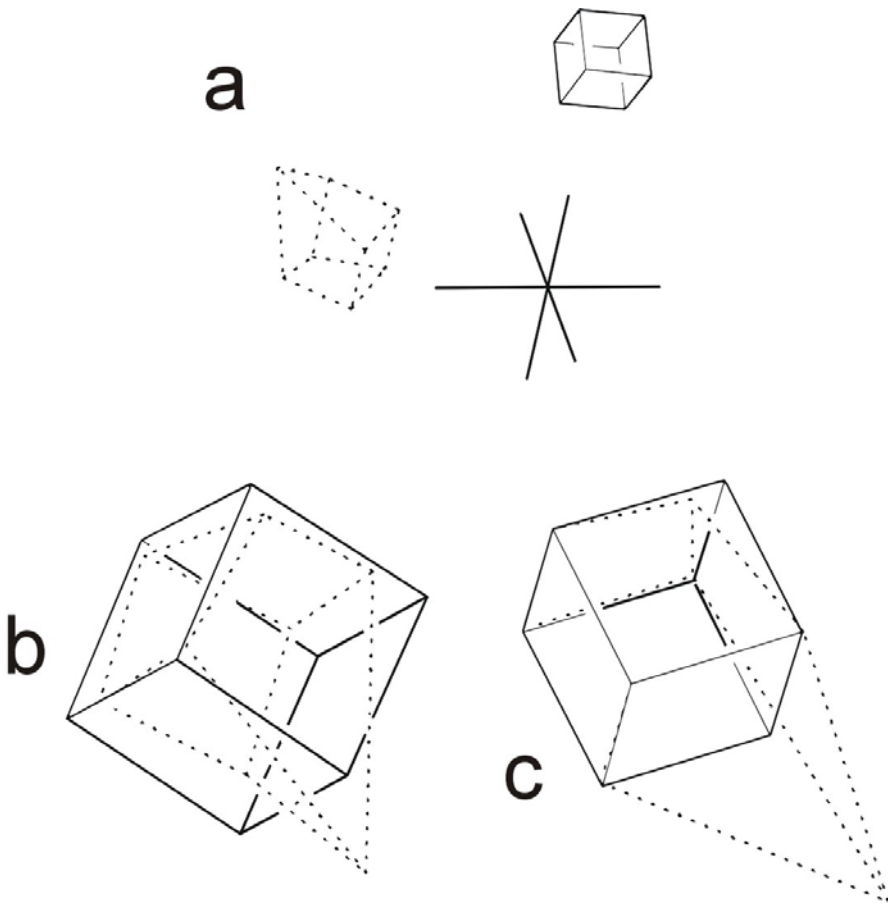
Argyropelecus olfersi. *Sternoptyx diaphana.*

Geometrická morfometrika



Vzájomné vzťahy tvarov druhov *Stenoptyx diaphana* (A) a *Argyropelecus olfersi* (B) – vzorové dáta z programu tpsSpline (http://life.bio.sunysb.edu/morph/), C – zobrazenie celkovej transformácie pomocou ohybnej pásky (*thin-plate spline*), D – to isté vyjadrené pomocou vektorov

Geometrická morfometrika



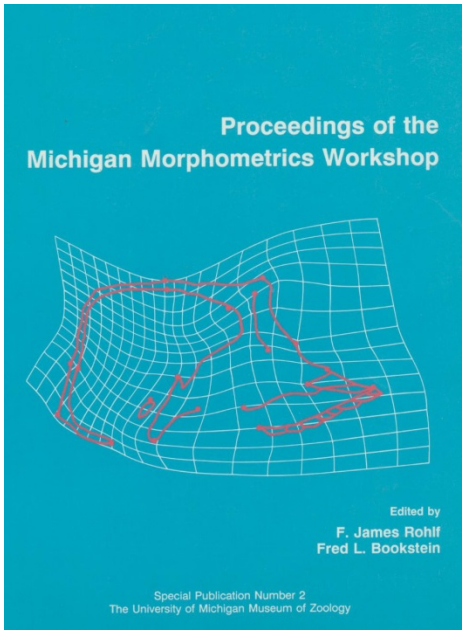
Prokrustova analýza. a – konsenzuálna konfigurácia plnou čiarou, jednotlivý objekt bodkovane; b – superpozícia metódou GLS (rozdíely v pozícii zodpovedajúcich význačných bodov sú porovnateľné); c – superpozícia metódou rezistentného prispôsobenia (objekty sa výrazne líšia v pozícii jediného bodu)

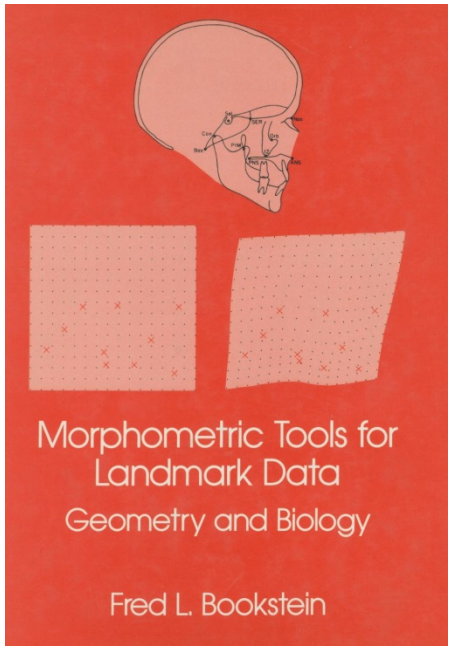
Rohlf, F.J. & Bookstein, F.L., eds., 1990. Proceedings of the Michigan morphometric workshop. *Special Publ. No. 2, The University of Michigan Museum of Zoology*. [Blue book]

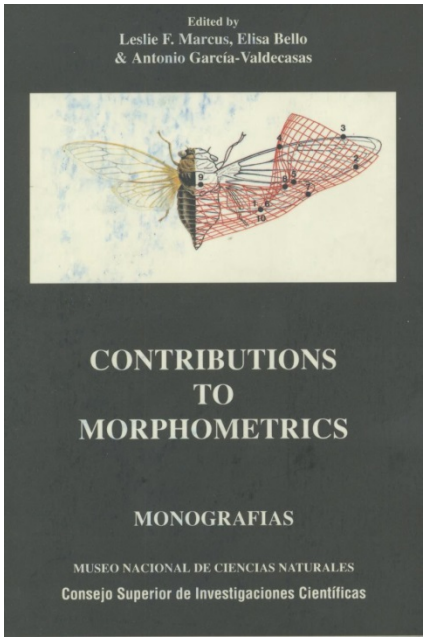
Bookstein, F.L. 1991. *Morphometric tools for landmark data: geometry and biology*. Cambridge University Press, New York. [Red book]

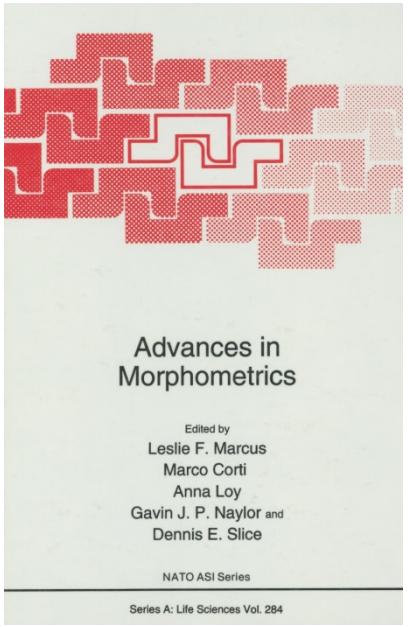
Marcus, L.F., Bello, E. & García-Valdecasas, A., eds., 1993. *Contributions to morphometrics*. Museo Nacional de Ciencias Naturales, Madrid. [Black book]

Marcus, L.F., Corti, M., Loy, A., Naylor, G.J.P. & Slice, D.E., eds., 1996. Advances in morphometrics. *NATO ASI Series A: Life Sciences* 284. [White book]









Macleod, N & Forey, P. 2002. *Morphology, shape and phylogeny*. Taylor and Francis, London, New York.

Macholán, M. 1999. Prokrustes, deformace a nová morfometrie. *Vesmír* 78: 35-39.