Shluková analýza

Shluk (klastr, *cluster*) je skupina objektů, které uvnitř nějaké větší skupiny nemají ani nahodilý ani rovnoměrný výskyt a jejich vzájemná vzdálenost resp. nepodobnost je menší než vzdálenost resp. nepodobnost s objekty, které patří do jiných shluků.

Těžiště (*centroid*) shluku je hypotetický (nikoliv nutně existující) prvek, jehož souřadnice ve znakovém prostoru jsou dány průměrnými hodnotami souřadnic jednotlivých objektů.

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

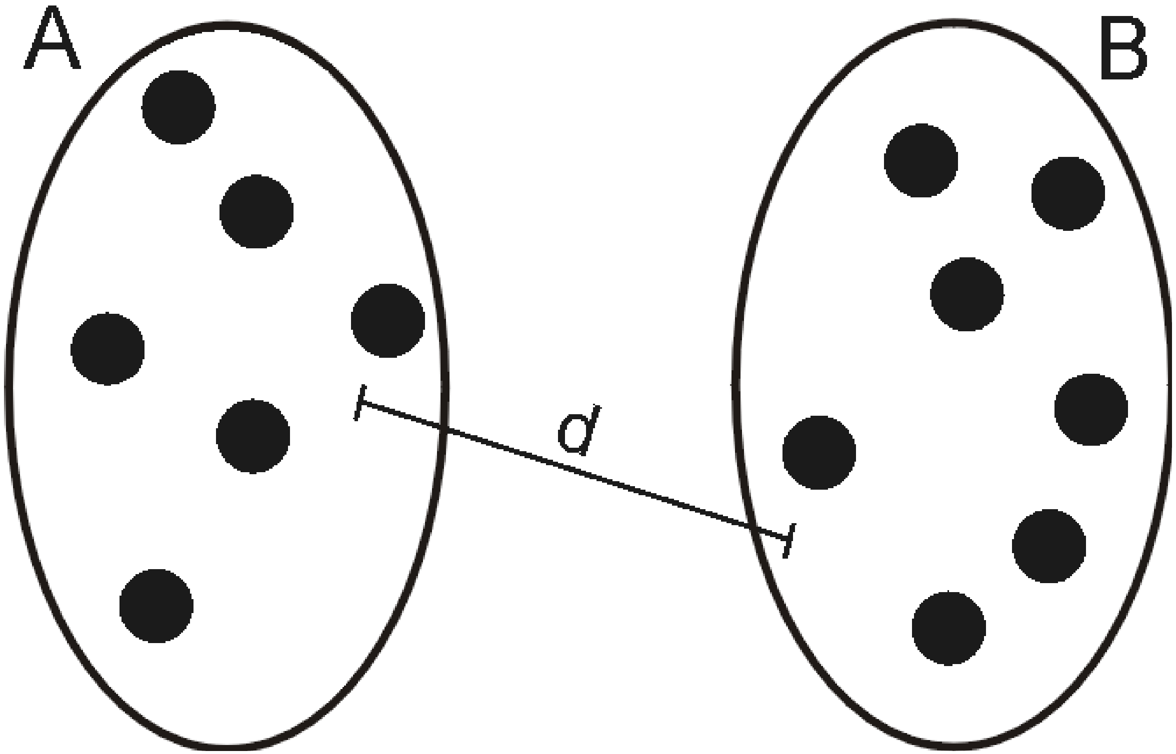
Příklady koeficientů používaných ve shlukových analýzách:

(a) pro binární data – Jaccardův koeficient, Sörensenův koeficient, koeficient jednoduché shody, euklidovská vzdálenost, tětivová vzdálenost;

(b) pro smíšená data – Gowerův koeficient, vzdálenost pro smíšená data;

(c) pro kvantitativní data – euklidovská vzdálenost, manhattanská metrika, tětivová vzdálenost.

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



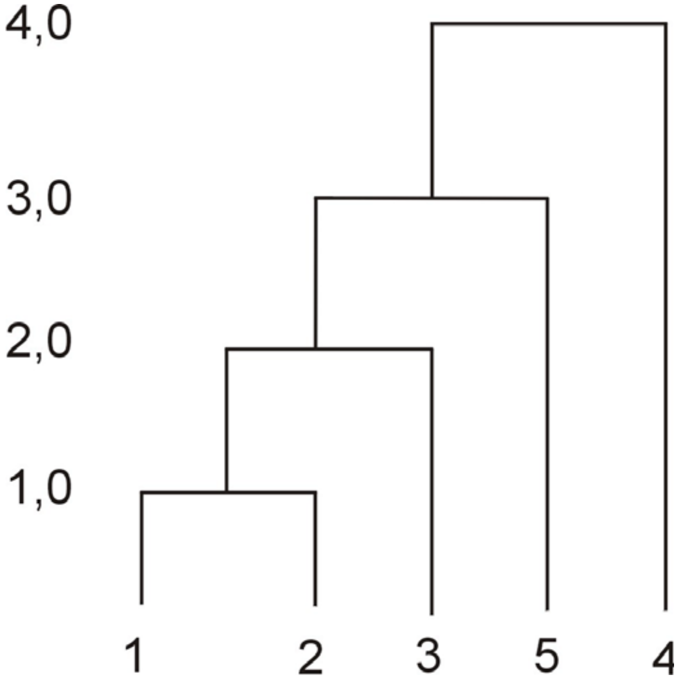
$$D_1 = \begin{array}{cccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{1} & 0,0 & 1,0 & 7,0 & 4,0 & 12,0 \\ \mathbf{2} & 1,0 & 0,0 & 2,0 & 5,0 & 9,0 \\ \mathbf{3} & 7,0 & 2,0 & 0,0 & 8,0 & 3,0 \\ \mathbf{4} & 4,0 & 5,0 & 8,0 & 0,0 & 6,0 \\ \mathbf{5} & 12,0 & 9,0 & 3,0 & 6,0 & 0,0 \end{array}$$

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

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

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

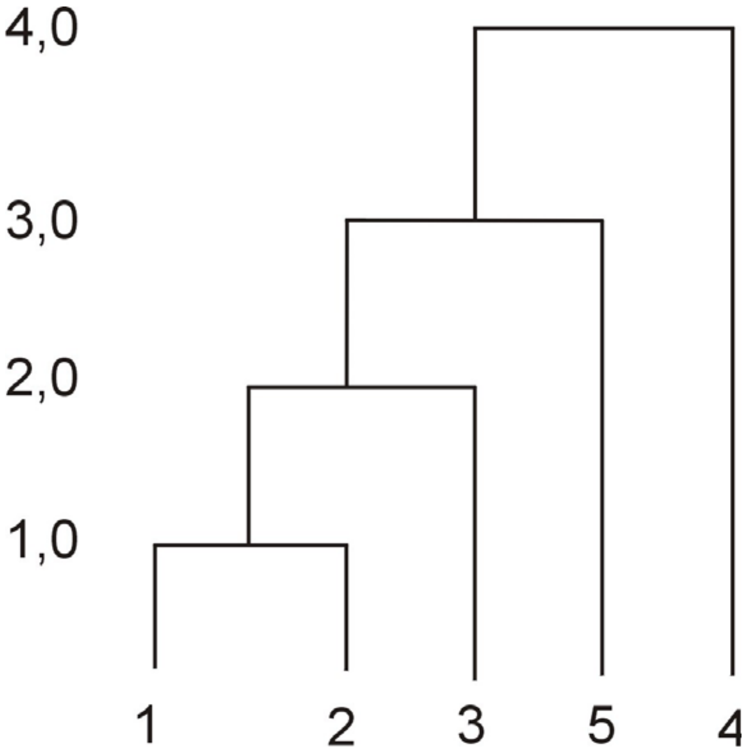
$$D_2 = \begin{array}{ccccc} & \mathbf{(1,\,2)} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{(1,\,2)} & 0,0 & 2,0 & 4,0 & 9,0 \\ \mathbf{3} & 2,0 & 0,0 & 8,0 & 3,0 \\ \mathbf{4} & 4,0 & 8,0 & 0,0 & 6,0 \\ \mathbf{5} & 9,0 & 3,0 & 6,0 & 0,0 \end{array}$$



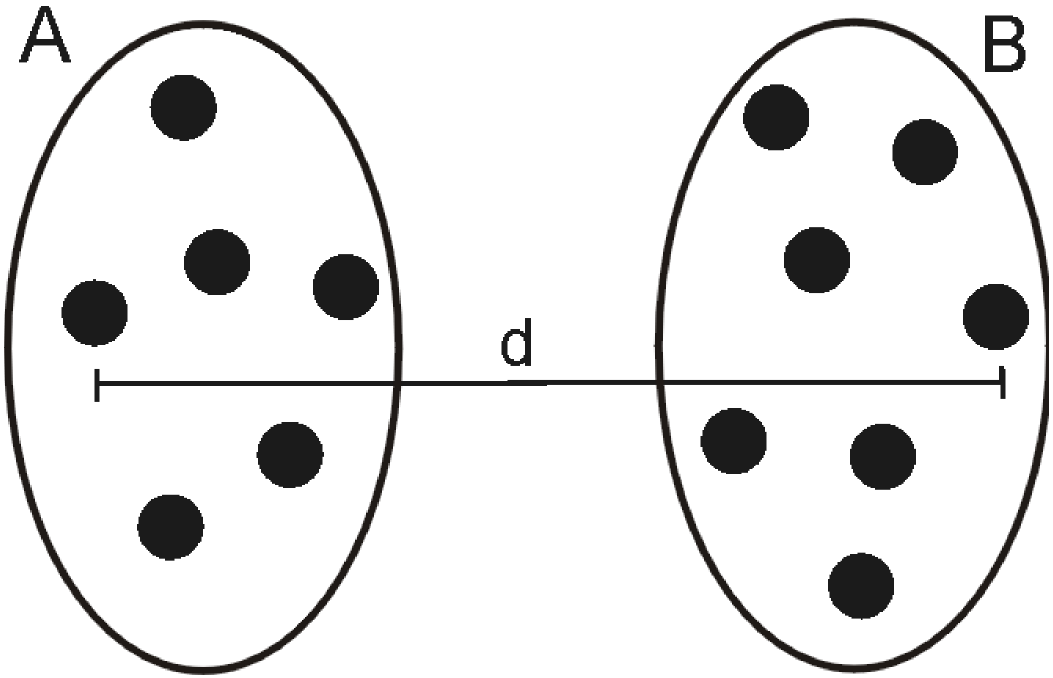
$$d_{(1,\,2,\,3)4} = \min\{d_{(1,\,2)\,4}, d_{3,\,4}\} = d_{(1,\,2)\,4} = 4{,}0$$

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

$$D_3 = \begin{array}{llll} & \mathbf{(1,\,2,\,3)} & \mathbf{4} & \mathbf{5} \\ \mathbf{(1,\,2,\,3)} & 0{,}0 & 4{,}0 & 3{,}0 \\ \mathbf{4} & 4{,}0 & 0{,}0 & 6{,}0 \\ \mathbf{5} & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$



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



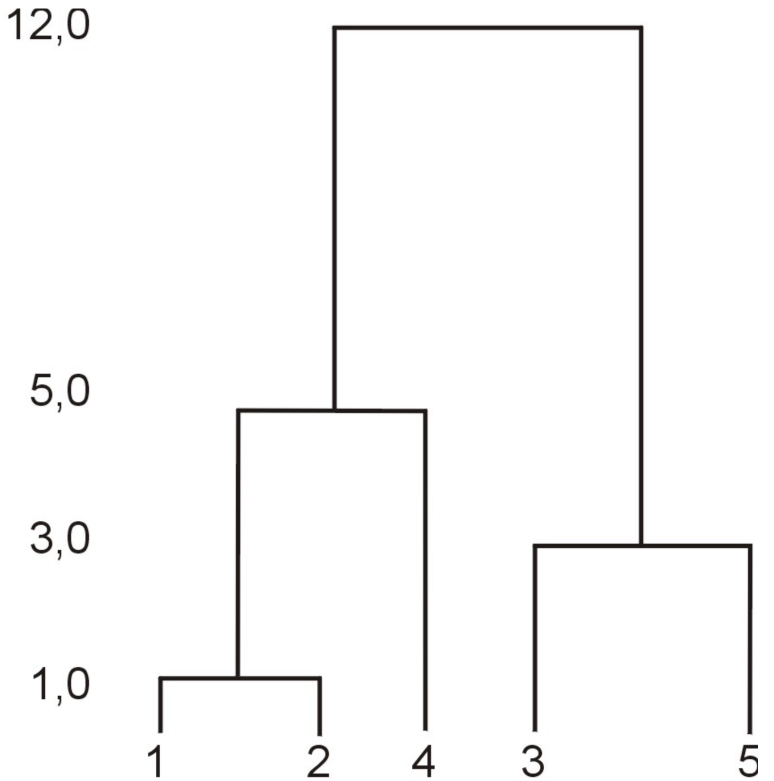
$$D_1 = \begin{array}{c|ccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{1} & 0{,}0 & 1{,}0 & 7{,}0 & 4{,}0 & 12{,}0 \\ \mathbf{2} & 1{,}0 & 0{,}0 & 2{,}0 & 5{,}0 & 9{,}0 \\ \mathbf{3} & 7{,}0 & 2{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ \mathbf{4} & 4{,}0 & 5{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ \mathbf{5} & 12{,}0 & 9{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

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

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

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

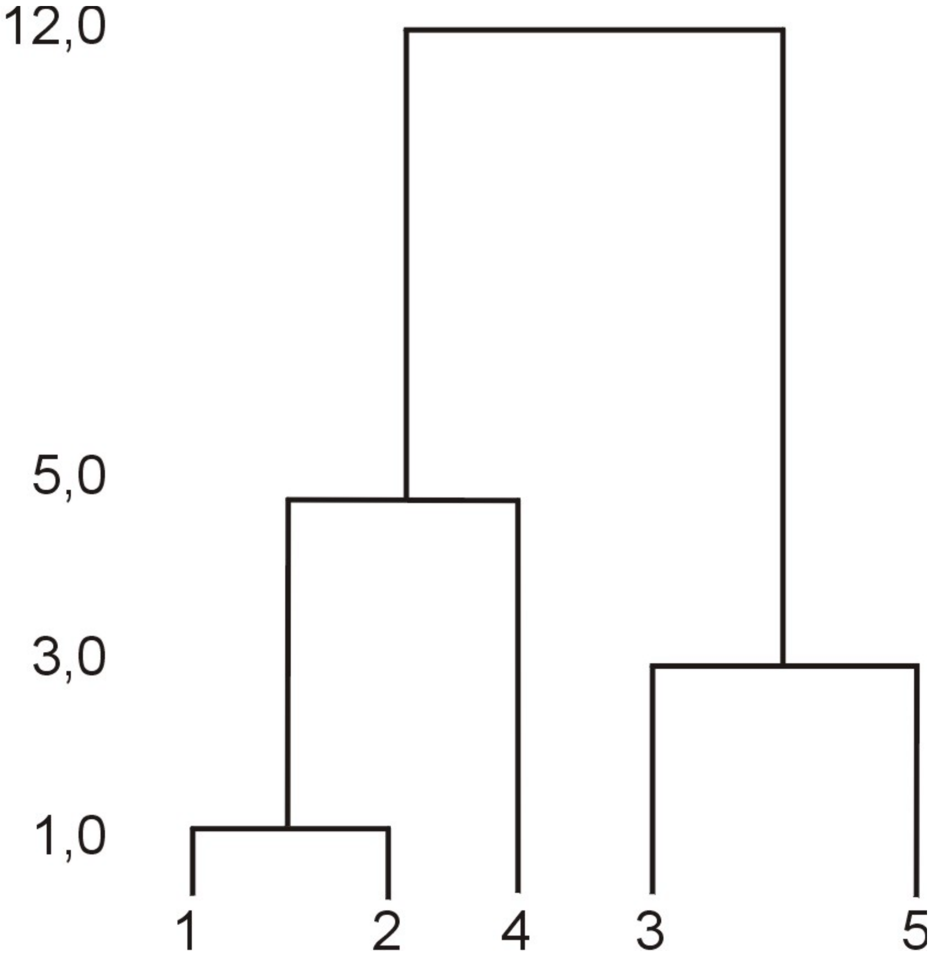
$$D_2 = \begin{array}{c|cccc} & \mathbf{(1,\,2)} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{(1,\,2)} & 0{,}0 & 7{,}0 & 5{,}0 & 12{,}0 \\ \mathbf{3} & 7{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ \mathbf{4} & 5{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ \mathbf{5} & 12{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$



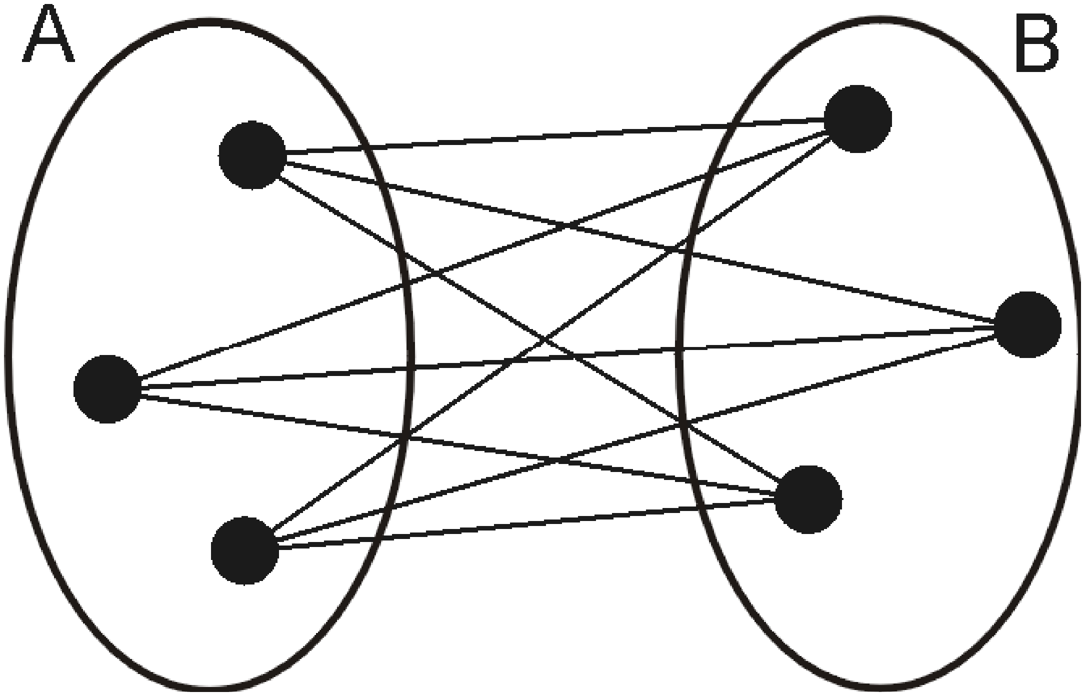
$$d_{(1,\ 2)(3,\ 5)} = \max\{d_{(1,\ 2)\ 3}, d_{(1,\ 2)\ 5}\} = d_{(1,2),\ 5} = 12{,}0$$

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

$$D_3 = \begin{array}{llll} & \mathbf{(1,\ 2)} & \mathbf{(3,\ 5)} & \mathbf{4} \\ \mathbf{(1,\ 2)} & 0{,}0 & 12{,}0 & 5{,}0 \\ \mathbf{(3,\ 5)} & 12{,}0 & 0{,}0 & 8{,}0 \\ \mathbf{4} & 5{,}0 & 8{,}0 & 0{,}0 \end{array}$$



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



$$D_1 = \begin{array}{cccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{1} & 0{,}0 & 1{,}0 & 7{,}0 & 4{,}0 & 12{,}0 \\ \mathbf{2} & 1{,}0 & 0{,}0 & 2{,}0 & 5{,}0 & 9{,}0 \\ \mathbf{3} & 7{,}0 & 2{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ \mathbf{4} & 4{,}0 & 5{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ \mathbf{5} & 12{,}0 & 9{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

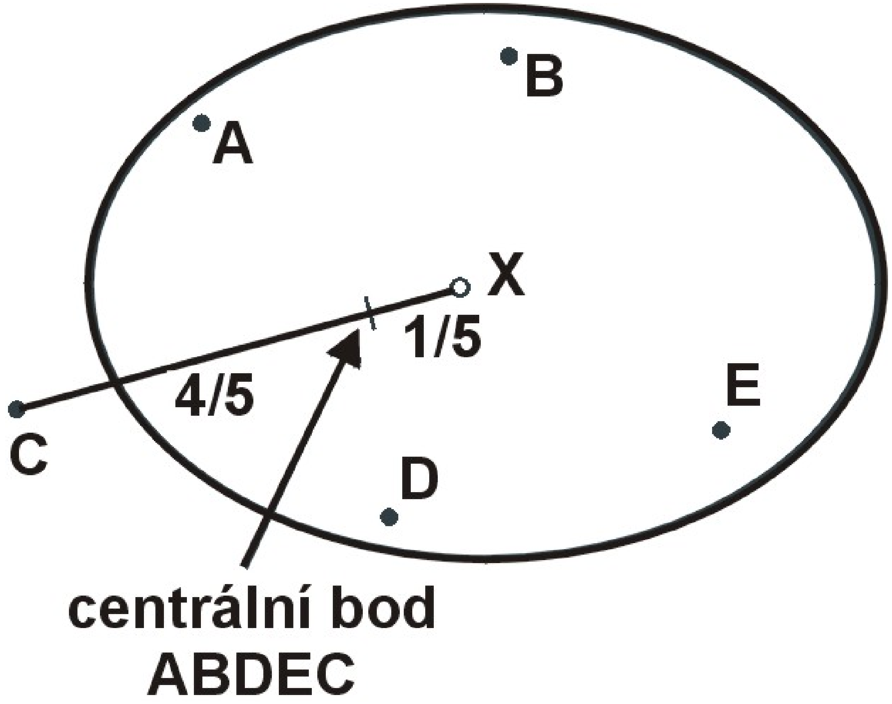
$d_{(1,\ 2)3} = 1/2\ (d_{1,\ 3} + d_{2,\ 3}) = 4{,}5$

$d_{(1,\ 2)4} = 1/2\ (d_{1,\ 4} + d_{2,\ 4}) = 4{,}5$

$d_{(1,\ 2)5} = 1/2\ (d_{1,\ 5} + d_{2,\ 5}) = 10{,}5$

$$D_2 = \begin{array}{ccccc} & \mathbf{(1,\ 2)} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{(1,\ 2)} & 0{,}0 & 4{,}5 & 4{,}5 & 10{,}5 \\ \mathbf{3} & 4{,}5 & 0{,}0 & 8{,}0 & 3{,}0 \\ \mathbf{4} & 4{,}5 & 8{,}0 & 0{,}0 & 6{,}0 \\ \mathbf{5} & 10{,}5 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

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

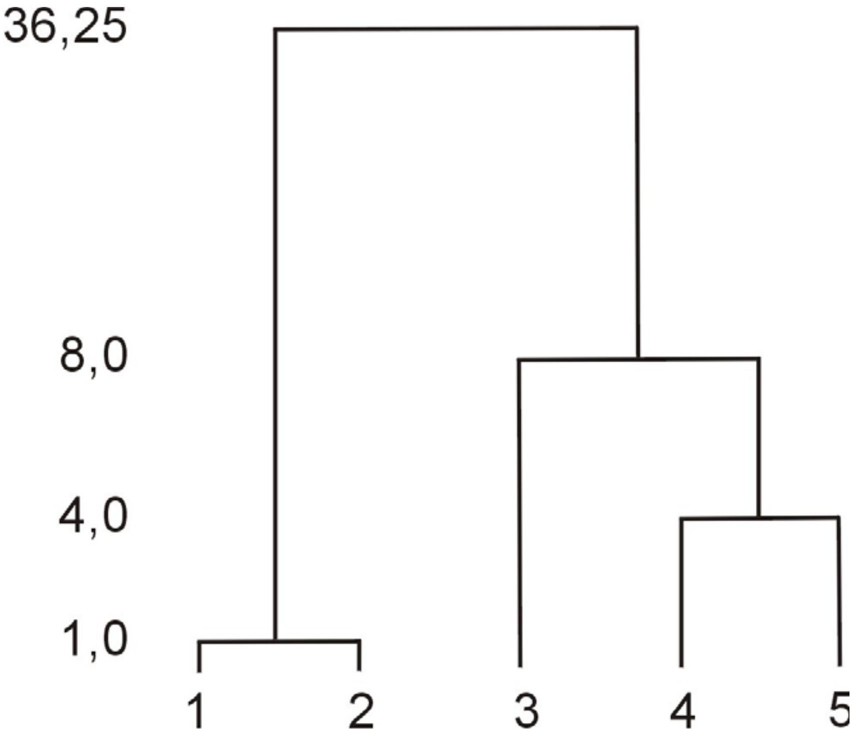


		znak	
		1	**2**
	1	1,0	1,0
	2	1,0	2,0
OTU	**3**	6,0	3,0
	4	8,0	2,0
	5	8,0	0,0

koeficient - druhá mocnina euklidovské vzdálenosti

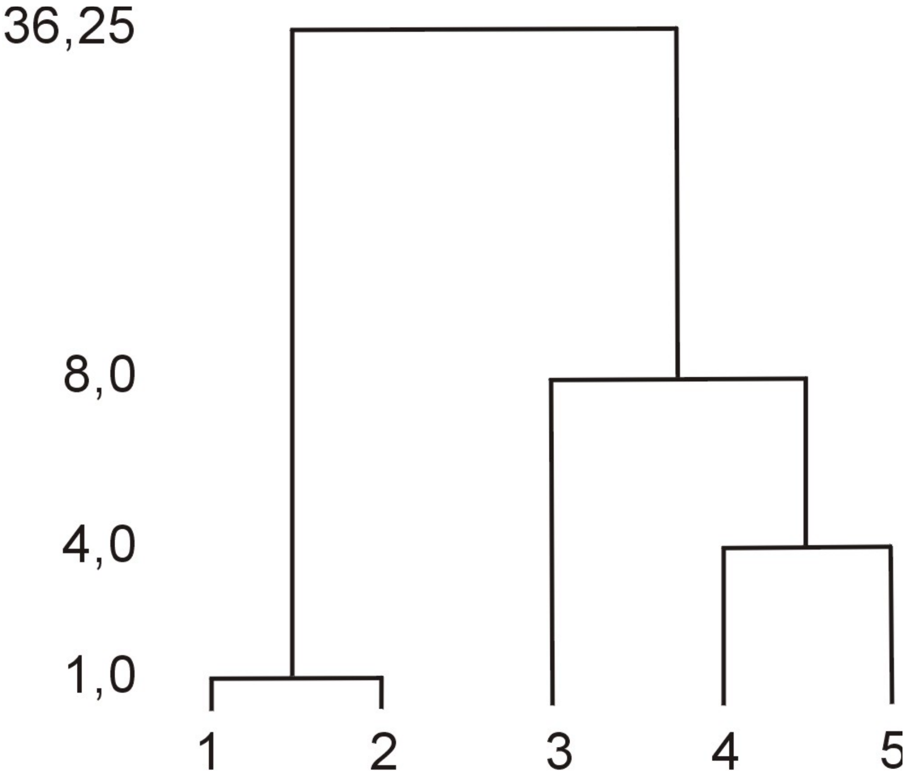
		1	**2**	**3**	**4**	**5**
	1	0,0	1,0	29,0	50,0	50,0
	2	1,0	0,0	26,0	49,0	53,0
D_1 =	**3**	29,0	26,0	0,0	5,0	13,0
	4	50,0	49,0	5,0	0,0	4,0
	5	50,0	53,0	13,0	4,0	0,0

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



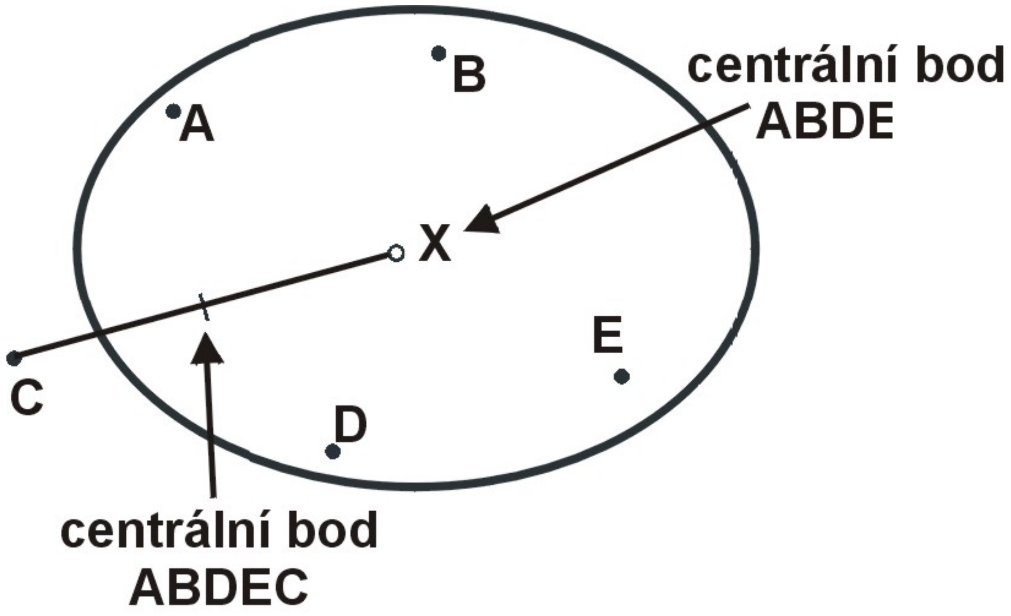
$$D_2 = \begin{array}{ccccc} & \mathbf{(1,\,2)} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{(1,\,2)} & 0{,}0 & 27{,}25 & 49{,}25 & 51{,}25 \\ \mathbf{3} & 27{,}25 & 0{,}0 & 5{,}0 & 13{,}0 \\ \mathbf{4} & 49{,}25 & 5{,}0 & 0{,}0 & 4{,}0 \\ \mathbf{5} & 51{,}25 & 13{,}0 & 4{,}0 & 0{,}0 \end{array}$$

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



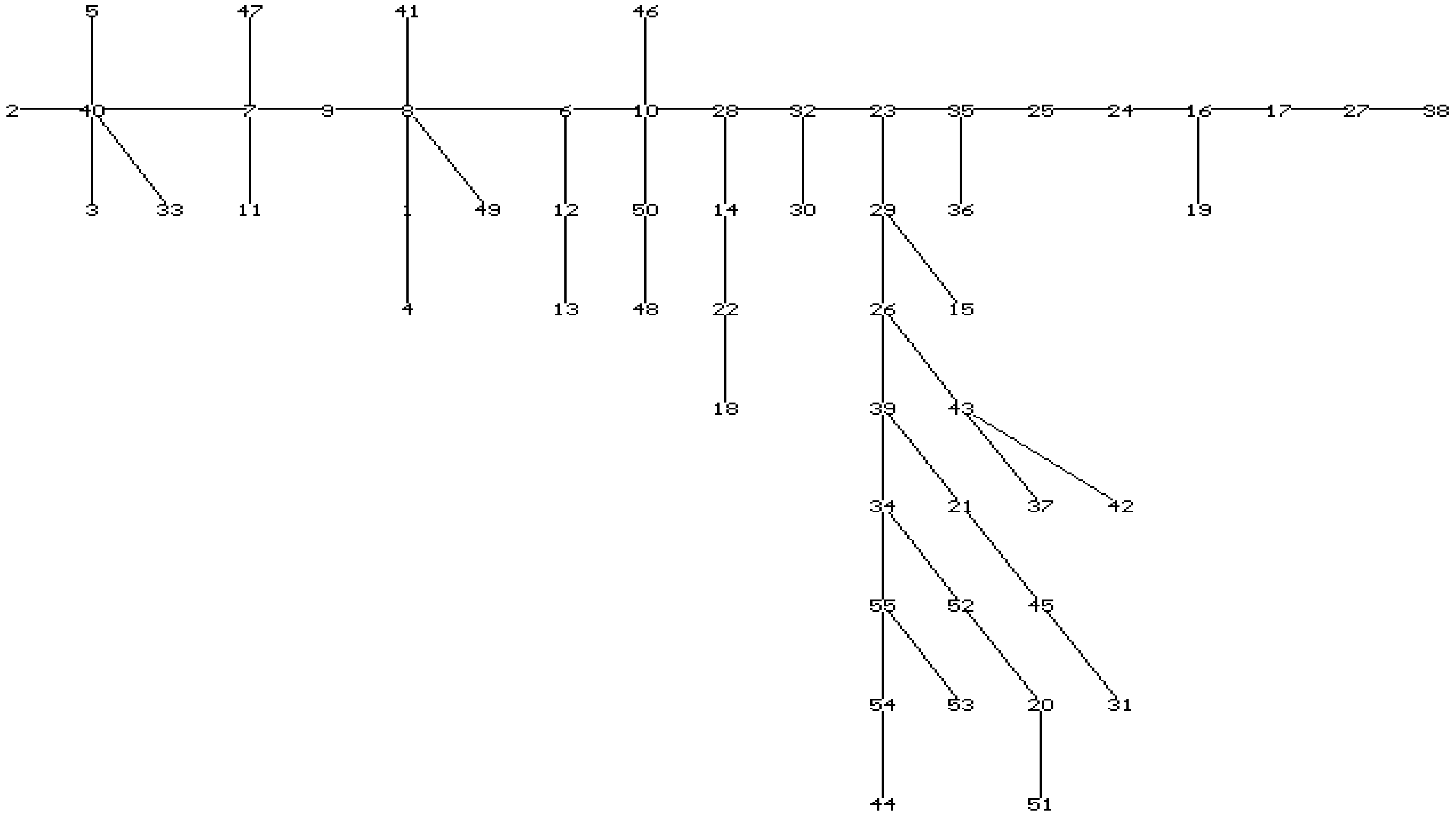
$$D_3 = \begin{array}{cccc} & \mathbf{(1,\ 2)} & \mathbf{3} & \mathbf{(4,\ 5)} \\ \mathbf{(1,\ 2)} & 0{,}0 & 27{,}25 & 49{,}25 \\ \mathbf{3} & 27{,}25 & 0{,}0 & 8{,}0 \\ \mathbf{(4,\ 5)} & 49{,}25 & 8{,}0 & 0{,}0 \end{array}$$

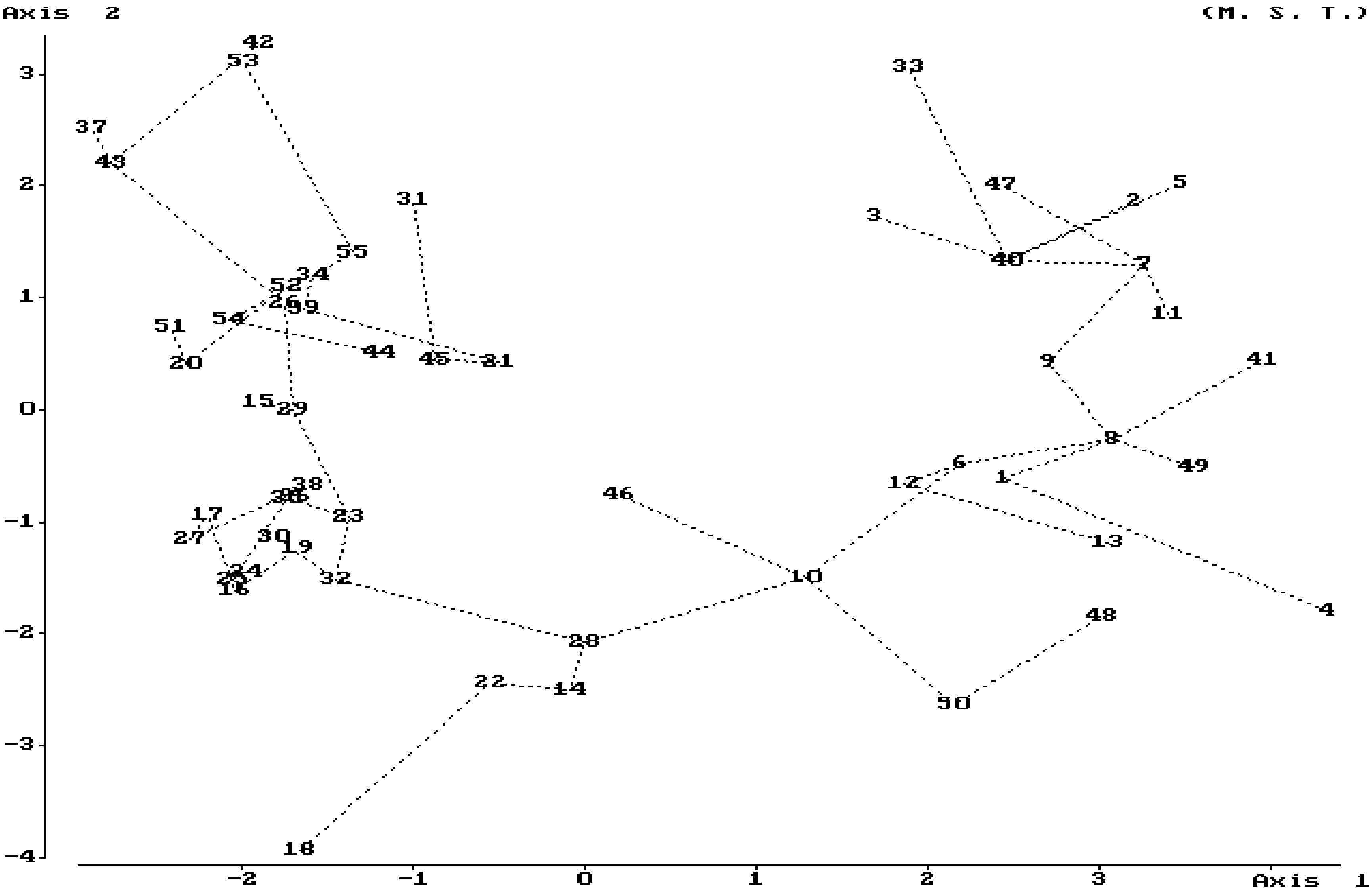
Mediánová metoda (*median method, WPGMC – weighted pair-group method using centroids, weighted centroid clustering*)



Minimální kostra (*minimum spanning tree*)

graf, který spojuje všechny objekty tak, že se zde nevyskytují žádné smyčky nebo kružnice a zároveň součet délky spojnic mezi uzly (objekty) je minimální




Axis 2
(M. S. T.)
Axis 1
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
42
53
37
43
55
34
52
26
39
51
54
20
44
45
21
31
15
29
38
36
35
17
23
27
30
19
24
25
16
32
28
22
14
18
33
47
5
2
3
40
7
11
9
41
8
6
49
12
1
46
13
10
4
48
50

Obecné poznámky ke shlukovacím metodám

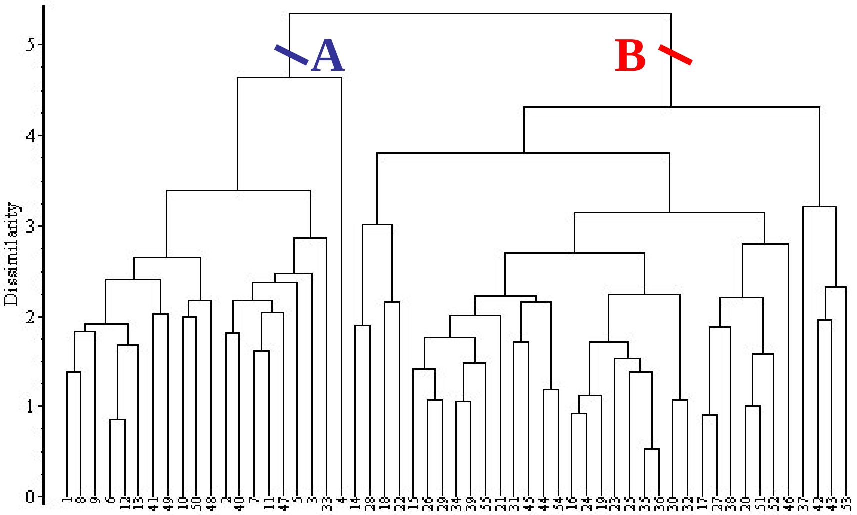
Pokud data nemají zcela jednoznačnou a zřetelnou strukturu (jedná se víceméne o náhodně rozptýlené objekty), je pravděpodobné, že použití různých shlukovacích technik přinese odlišné výsledky.

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).

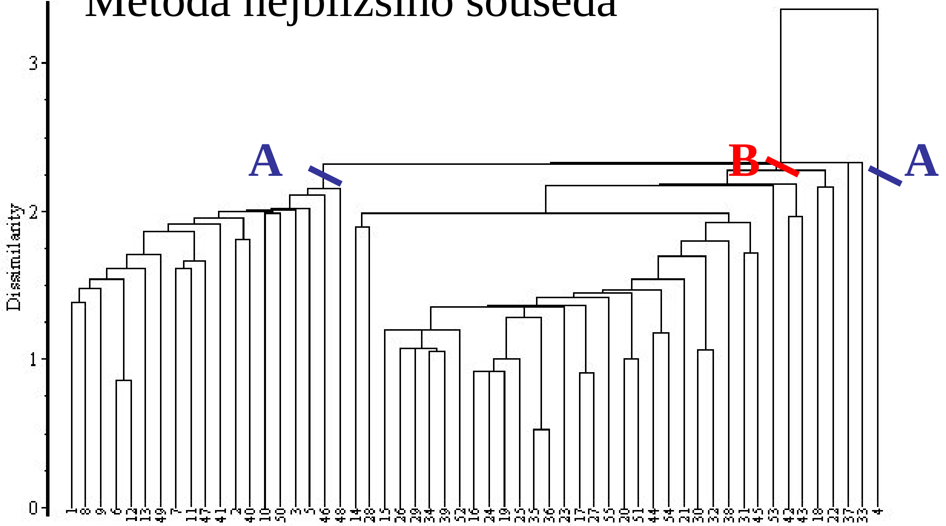
Mnohé shlukovací techniky jsou citlivé na přítomnost odlehlých objektů (*outliers*, výrazně atypických případů). Před samotnou shlukovou analýzou je proto vhodné použít některou z metod na jejich detekci, např. PCA. Výrazně odlehlé objekty se zpravidla z dalších analýz vylučují.

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

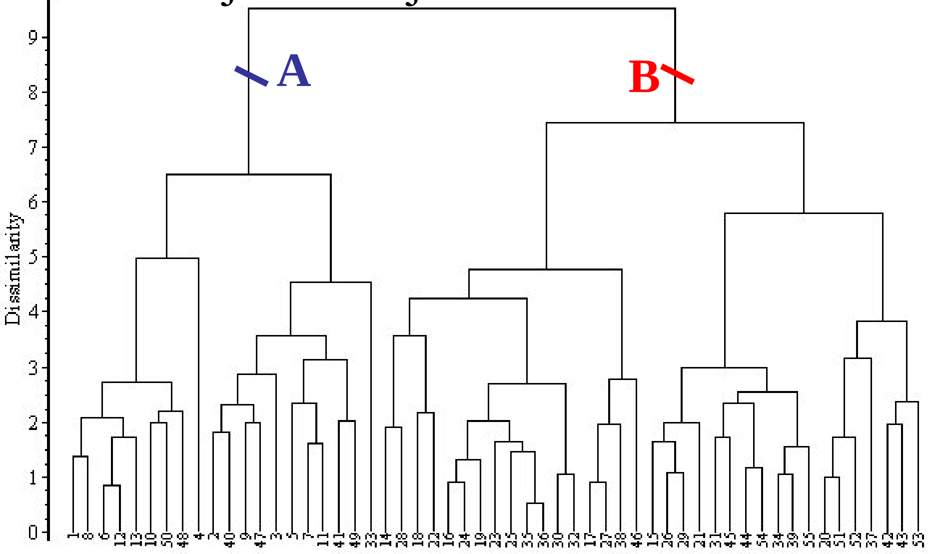
Metoda průměrné vzdálenosti



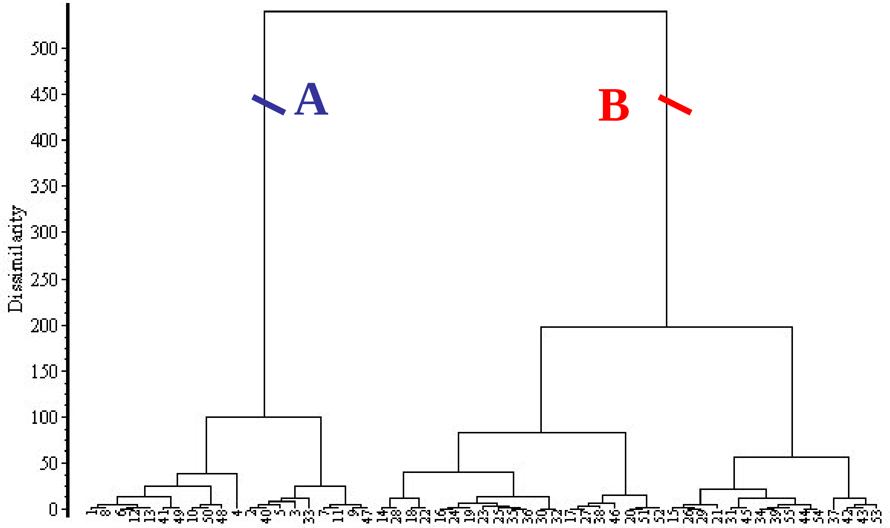
Metoda nejbližšího souseda



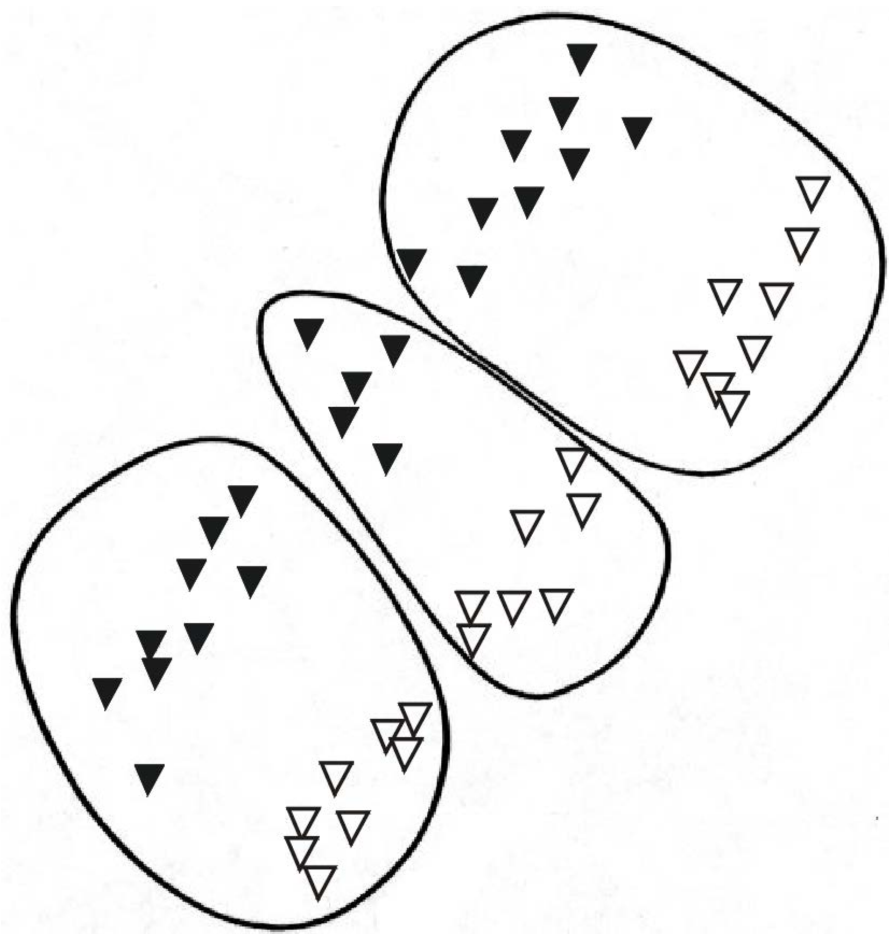
Metoda nejvzdálenějšího souseda



Wardova metoda

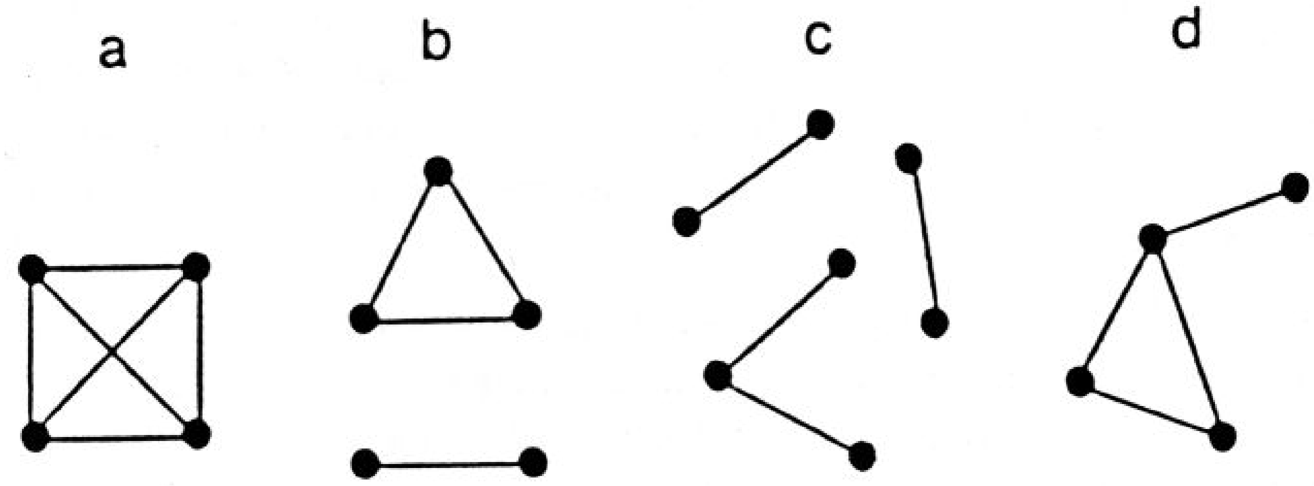


jednoznačná podpora je pro dva taxony, další seskupení reflektují rozdíly ve shlukovacích algoritmech



Metoda nejbližšího souseda by v důsledku řetězového efektu spojila do jednoho shluku plné trojúhelníky a do druhého prázdné trojúhelníky, zatímco Wardova a metoda průměrné vzdálenosti by přinesly skupiny ohraničené čárami (podle Everitt & Dunn 1983)

Shody (*ties*)

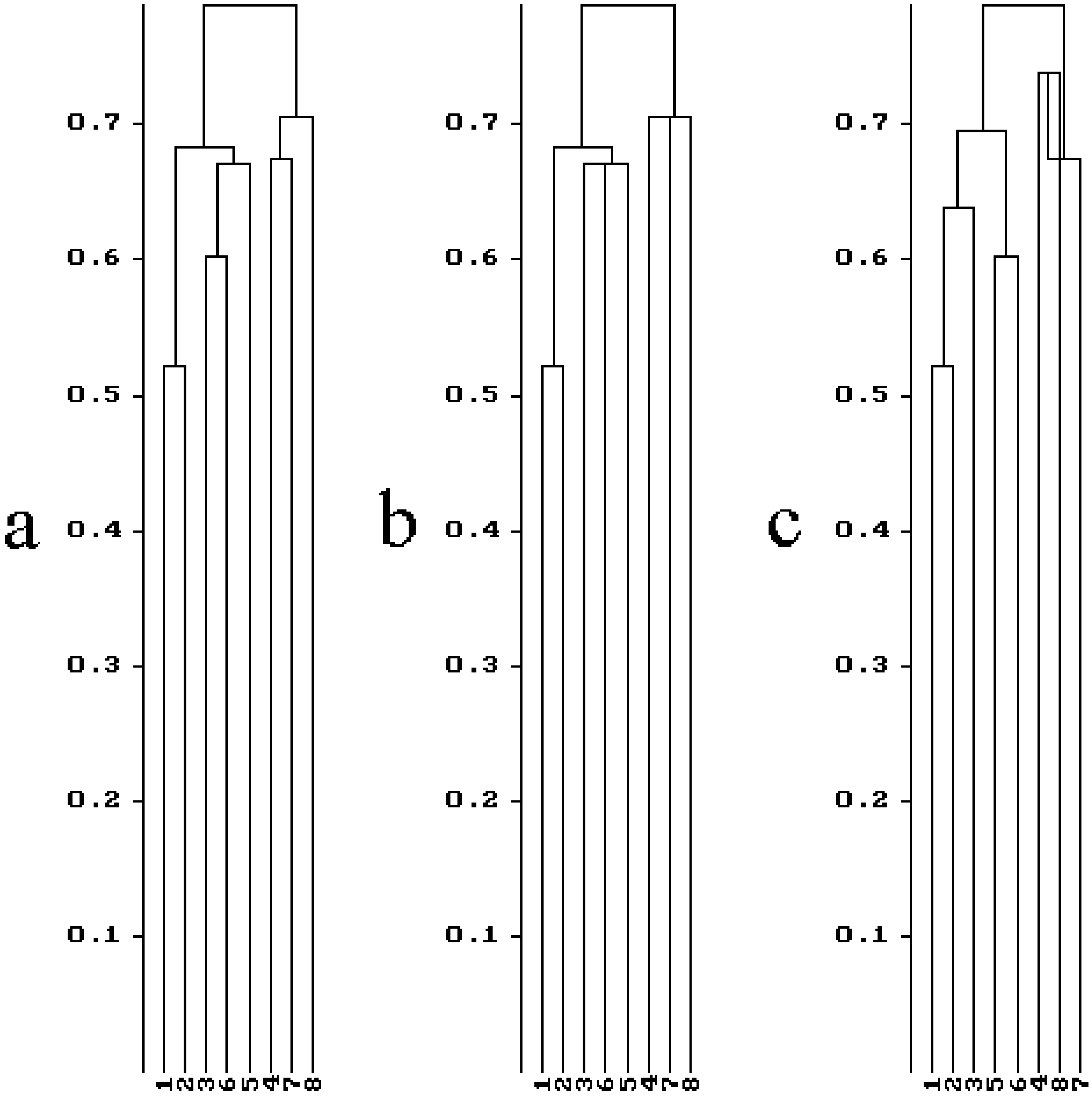


a – graf je úplný, b – graf je nesouvislý a všechny izolované komponenty jsou úplné, c – graf je nesouvislý a alespoň jedna komponenta není úplná, d – graf je souvislý, ale není úplný

(a) „*silent mode (arbitrary)*“

(b) „*single linkage*“

(c)„*suboptimal fusions*“

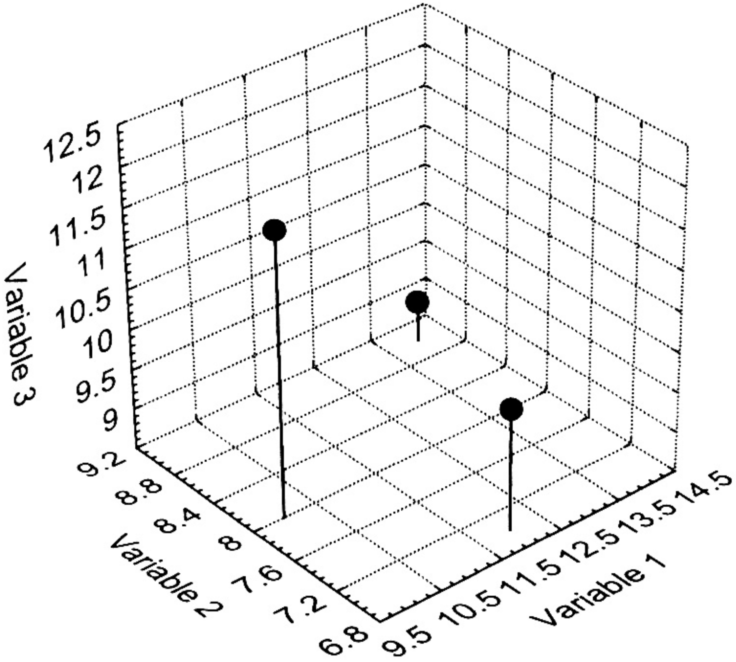


Ordinační metody

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 jejich vzdálenosti a seskupení.

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



Ordinační metody

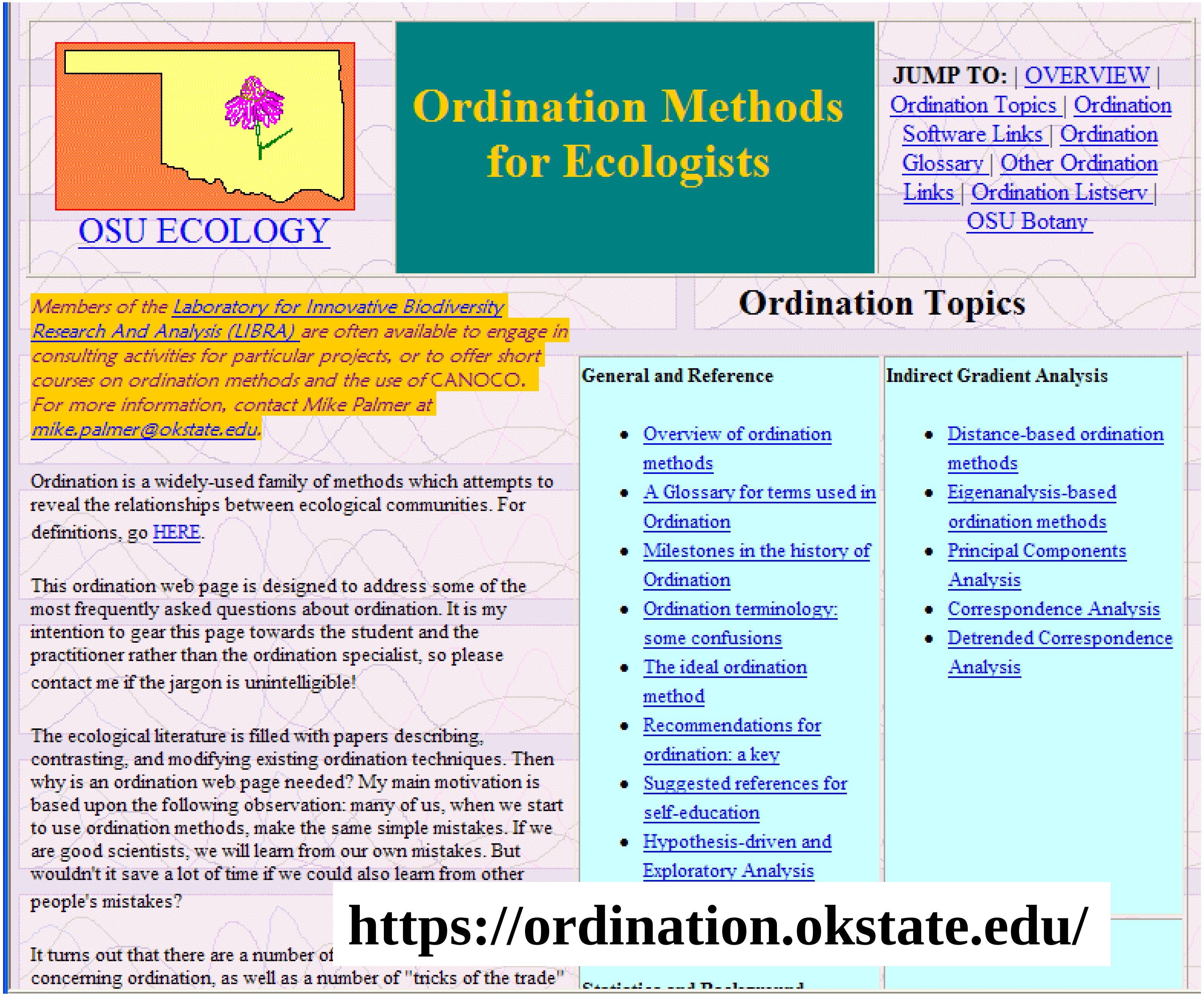
analýza hlavních komponent (PCA)

analýza hlavních koordinát (PCoA)

nemetrické mnohorozměrné škálování (NMDS)

Užitečné informace o ordinačních metodách se nacházejí na WWW stránce

http://ordination.okstate.edu/



OSU ECOLOGY

Ordination Methods for Ecologists

JUMP TO: | OVERVIEW | Ordination Topics | Ordination Software Links | Ordination Glossary | Other Ordination Links | Ordination Listserv | OSU Botany

Members of the Laboratory for Innovative Biodiversity Research And Analysis (LIBRA) are often available to engage in consulting activities for particular projects, or to offer short courses on ordination methods and the use of CANOCO. For more information, contact Mike Palmer at mike.palmer@okstate.edu.

Ordination is a widely-used family of methods which attempts to reveal the relationships between ecological communities. For definitions, go HERE.

This ordination web page is designed to address some of the most frequently asked questions about ordination. It is my intention to gear this page towards the student and the practitioner rather than the ordination specialist, so please contact me if the jargon is unintelligible!

The ecological literature is filled with papers describing, contrasting, and modifying existing ordination techniques. Then why is an ordination web page needed? My main motivation is based upon the following observation: many of us, when we start to use ordination methods, make the same simple mistakes. If we are good scientists, we will learn from our own mistakes. But wouldn't it save a lot of time if we could also learn from other people's mistakes?

It turns out that there are a number of

concerning ordination, as well as a number of "tricks of the trade"

Ordination Topics

General and Reference

- Overview of ordination methods
- A Glossary for terms used in Ordination
- Milestones in the history of Ordination
- Ordination terminology: some confusions
- The ideal ordination method
- Recommendations for ordination: a key
- Suggested references for self-education
- Hypothesis-driven and Exploratory Analysis

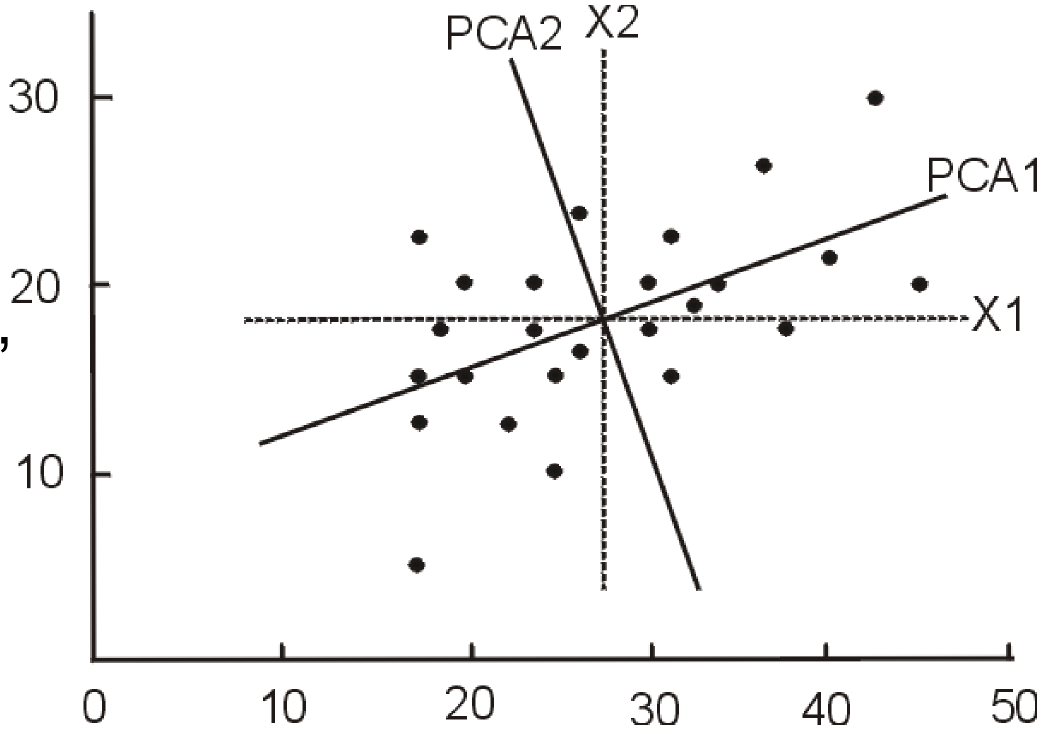
Indirect Gradient Analysis

- Distance-based ordination methods
- Eigenanalysis-based ordination methods
- Principal Components Analysis
- Correspondence Analysis
- Detrended Correspondence Analysis

https://ordination.okstate.edu/

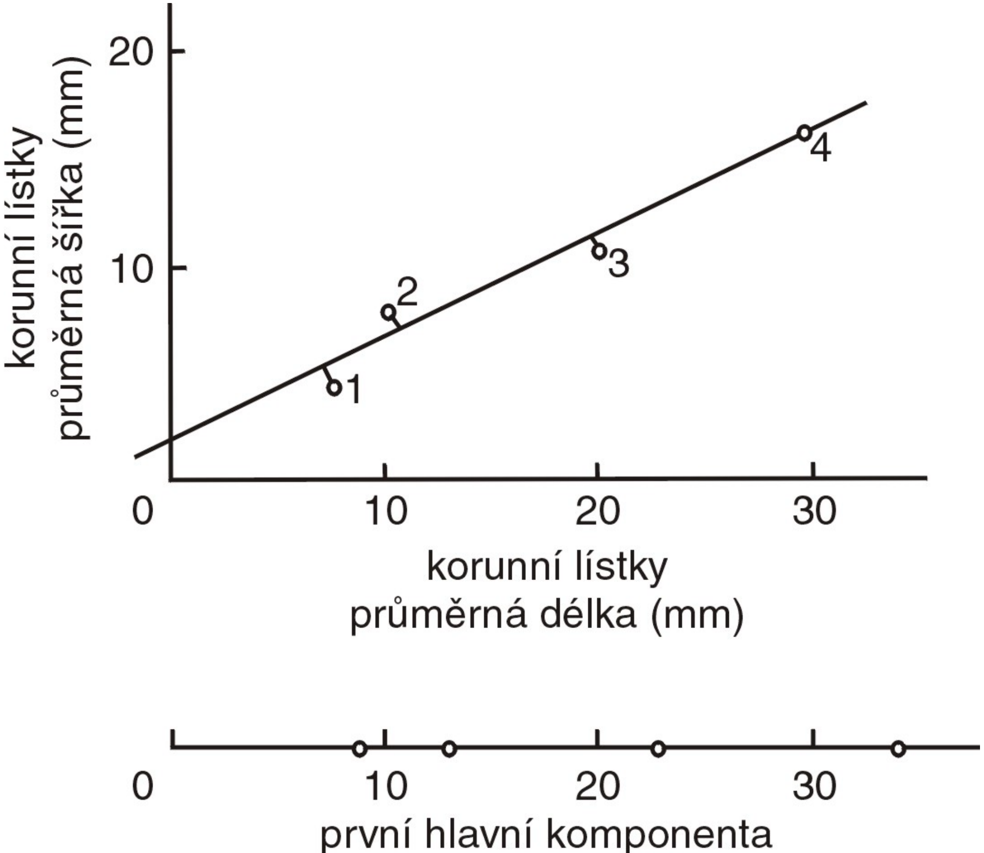
Analýza hlavních komponent (PCA – *principal component(s) analysis*)

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.



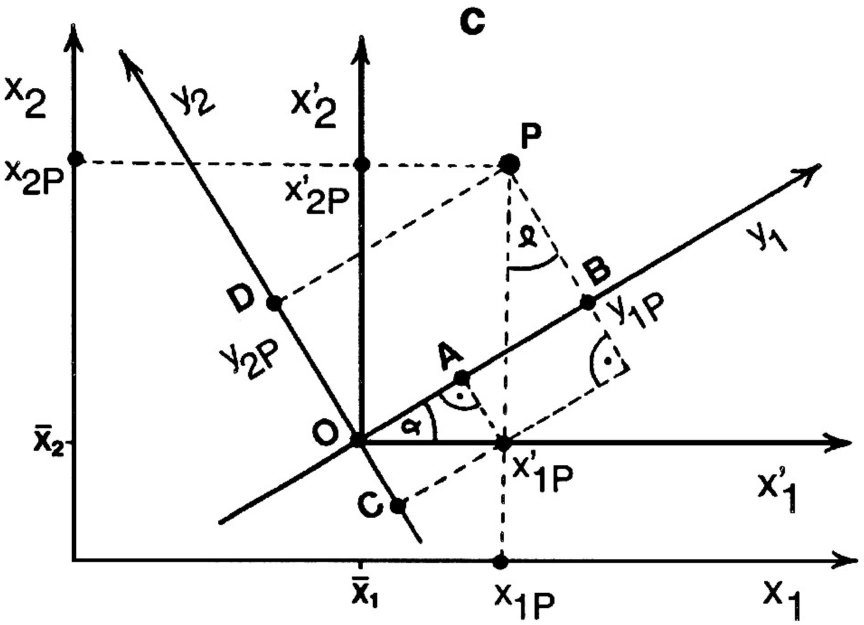
Geometrická interpretace PCA (podle Dunn & Everitt 1982):

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



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

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



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

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

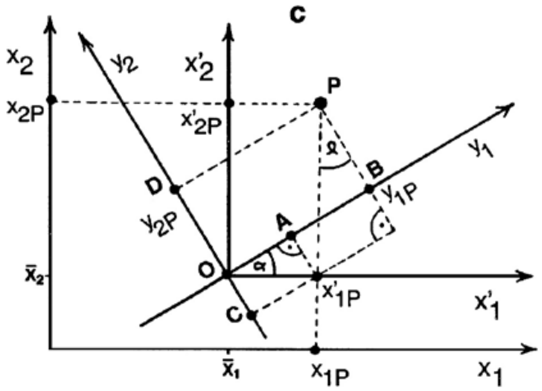
se transformuje do nového souboru znaků $y_1, y_2, ..., y_p$

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

.

.

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



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.

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

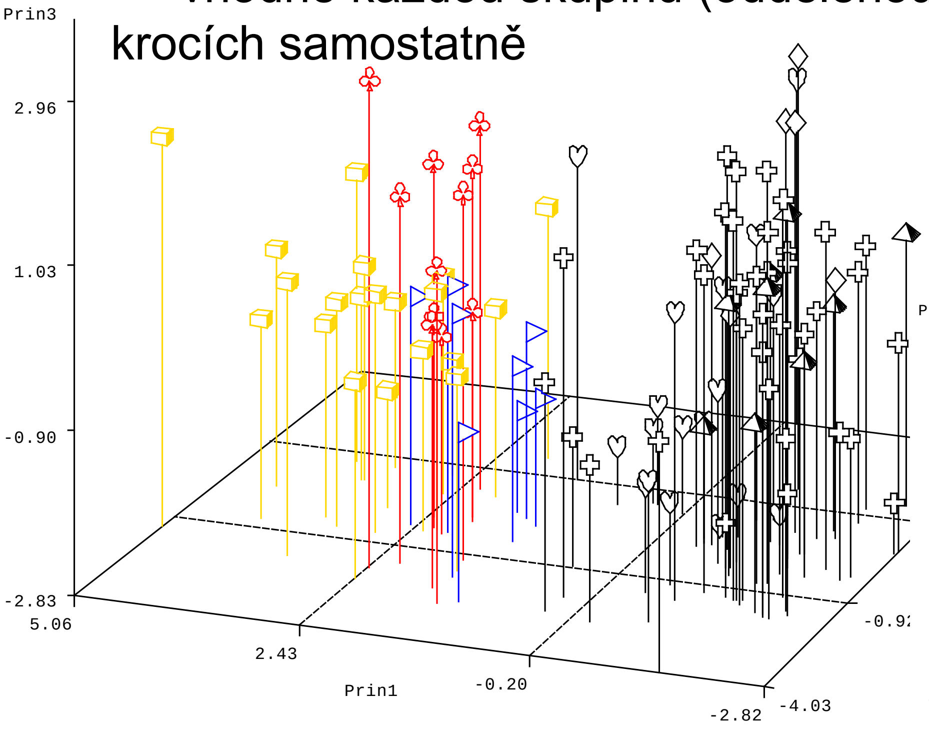
Počet objektů při PCA musí být alespoň o jeden větší než je počet analyzovaných znaků.

Obvykle se však doporučuje, aby se počet objektů blížil druhé mocnině počtu znaků (souvisí s počtem stupňů volnosti).

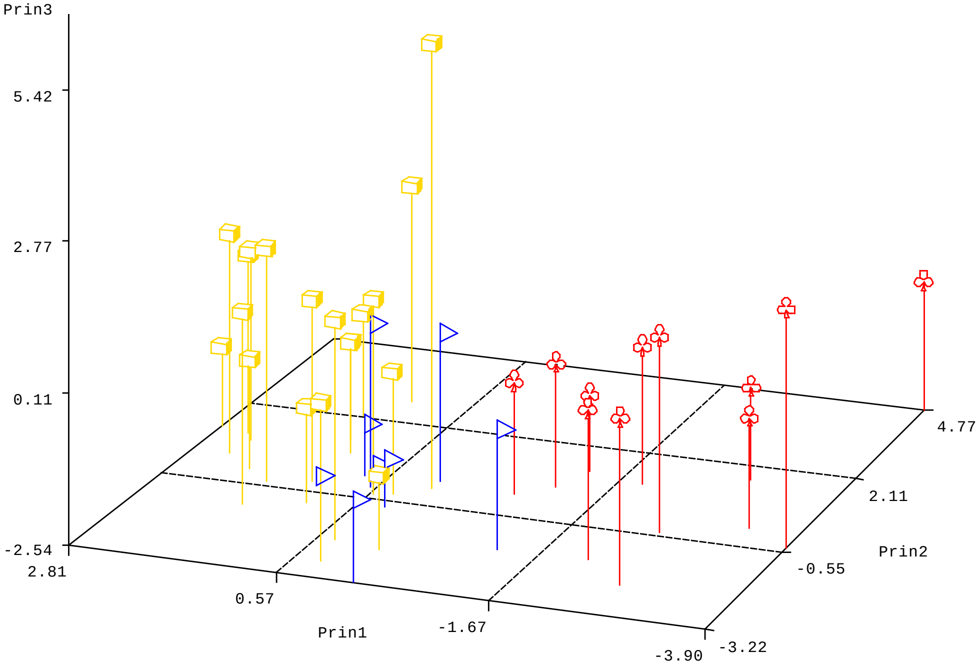
V případě, že $n \leq p$ (kde n je počet objektů a p počet znaků), výsledná matice (korelační nebo kovarianční) řádu p má jen $n - 1$ nezávislých řádků nebo sloupců. V takovém případě příslušná matice má $p - (n - 1)$ nulových vlastních čísel (na umístnění n objektů podle jejich vzájemných vzdáleností je zapotřebí jen $n - 1$ rozměrů)

Interpretace ordinace objektů může být složitá v případě příliš složité struktury obsažené v datech. Pokud jsou např. dvě základní skupiny (oddělené podél první osy) uvnitř členěné komplikovanějším způsobem, bývají druhá, třetí a další osy jistým kompromisem mezi strukturou v obou základních skupinách

=> vhodné každou skupinu (oddělenou podle první osy) analyzovat v dalších krocích samostatně

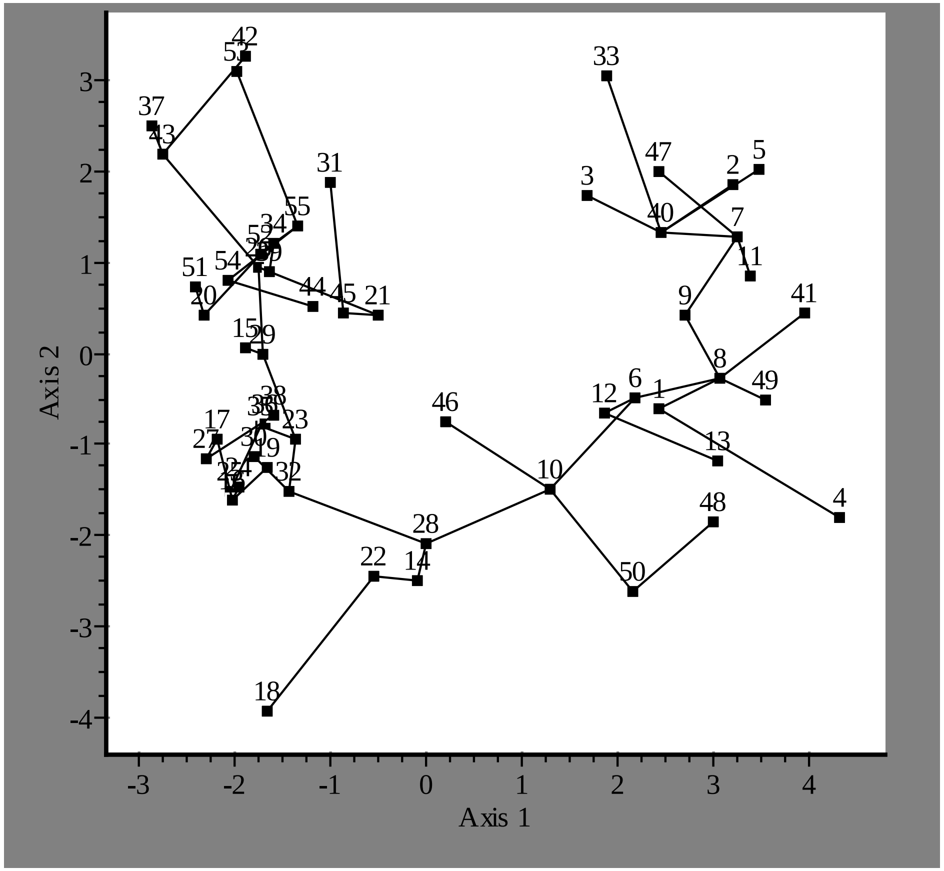


7 poddruhů (definovaných stupněm ploidie a areálem)



Ačkoliv byla technika PCA byla původně navržena pro kvantitativní znaky, může se použít také pro znaky binární a semikvantitativní. Binární data však ve zvýšené míře způsobují tzv. „podkovový efekt“ (*horseshoe effect*), kdy jsou objekty v ploše definované prvními dvěma komponentami uspořádané do tvaru podkovy.

Toto zakřivení odstraňují tzv. detrendované techniky, např. v programu DECORANA (Hill 1979), v taxonomických aplikacích se však podobná „narovnání“ zpravidla nepoužívají.



Cardamine amara **(Brassicaceae)**


subsp. *amara*


subsp. *opicii*

Korelační matice

```
RESEMBLANCE MATRIX

  ROW    1
  0.1000E+01 -0.3509E-01  0.6541E+00 -0.2760E+00 -0.2020E-01
  0.5767E+00  0.7991E+00 -0.3839E+00  0.8776E+00  0.7924E+00
  ROW    2
 -0.3509E-01  0.1000E+01  0.3224E+00  0.4878E+00  0.7738E+00
  0.2010E+00 -0.7949E-01  0.6593E-01 -0.1057E+00 -0.6612E-01
  ROW    3
  0.6541E+00  0.3224E+00  0.1000E+01 -0.9458E-01  0.3456E+00
  0.4514E+00  0.6788E+00 -0.4203E+00  0.5752E+00  0.6789E+00
  ROW    4
 -0.2760E+00  0.4878E+00 -0.9458E-01  0.1000E+01  0.6564E+00
  0.1018E+00 -0.4963E+00  0.5808E+00 -0.4945E+00 -0.5906E+00
  ROW    5
 -0.2020E-01  0.7738E+00  0.3456E+00  0.6564E+00  0.1000E+01
  0.2194E+00 -0.1818E+00  0.2400E+00 -0.1856E+00 -0.2039E+00
```

Vlastní čísla

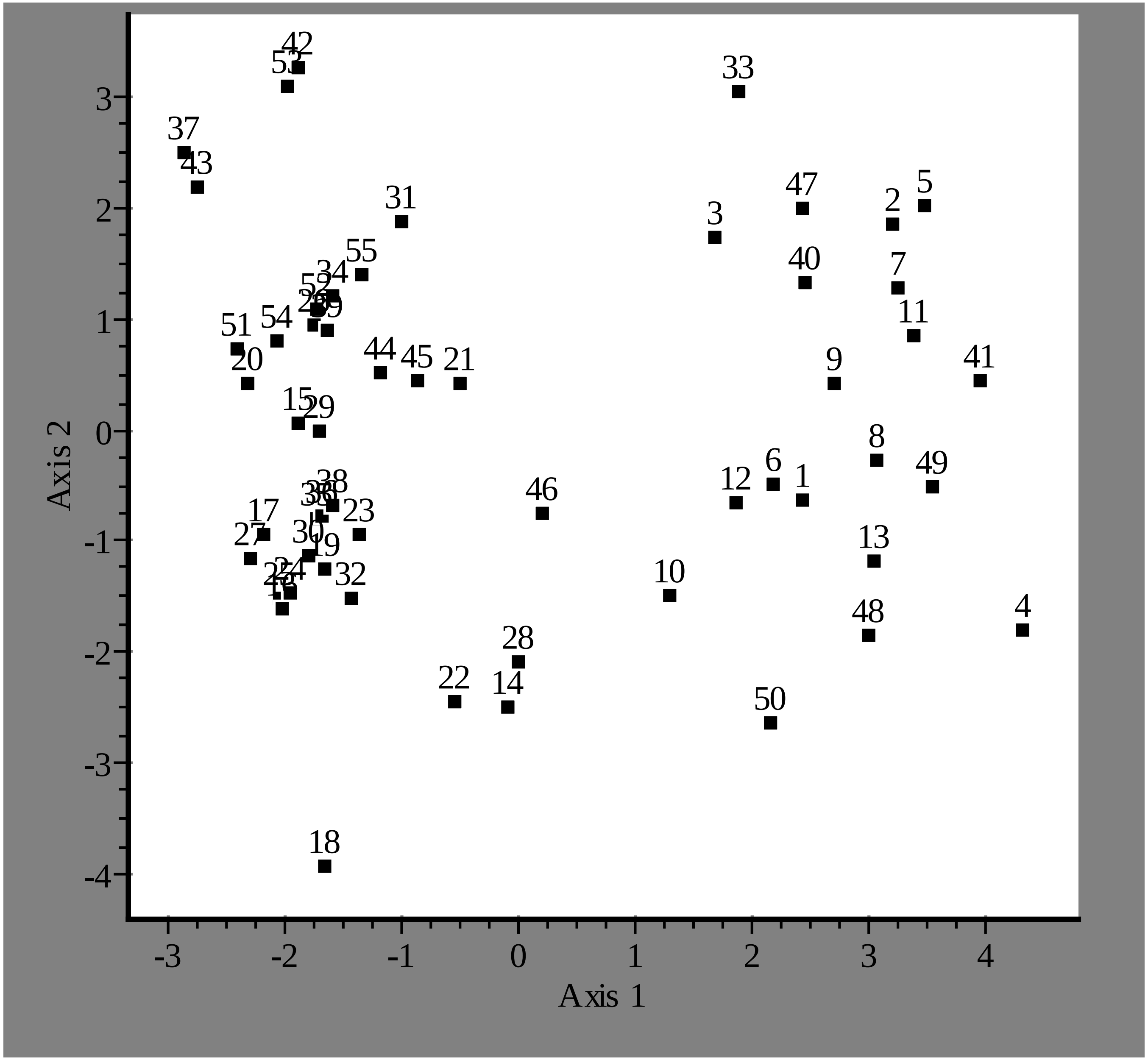
```
(1) NUMBER OF POSITIVE EIGENVALUES =    10
(2) SUM OF POSITIVE EIGENVALUES =       0.10000000E+02
(3) EIGENVALUES
 0.5030E+01  0.2590E+01  0.1127E+01  0.3886E+00  0.3164E+00
 0.1992E+00  0.1353E+00  0.1054E+00  0.6441E-01  0.4339E-01
(4) EIGENVALUES AS PERCENT
      50.30       25.90       11.27        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
```

Vlastní vektory (vektory směrových kosinů)

```
EIGENVECTORS (DIRECTION COSINES)
  VECTOR    1
      .38449      -.05267       .31208      -.26483      -.09857
      .16803       .42428      -.32014       .41978       .43035
  VECTOR    2
      .15583       .50365       .33372       .40278       .54941
      .34739       .02809       .15840       .00622       .00253
  VECTOR    3
      .26530      -.38005      -.13325       .05345      -.24200
      .62749      -.00397       .55644       .01315      -.04942
```

Komponentní skóre

```
COMPONENT SCORES


   1         2.440         -.617         -.296
   2         3.203         1.866         -.442
   3         1.689         1.730         -.520
   4         4.332        -1.803         1.150
   5         3.485         2.018          .820
   6         2.192         -.477         -.485
   7         3.268         1.295         -.575
   8         3.077         -.271         -.586
```



Ordinace objektů

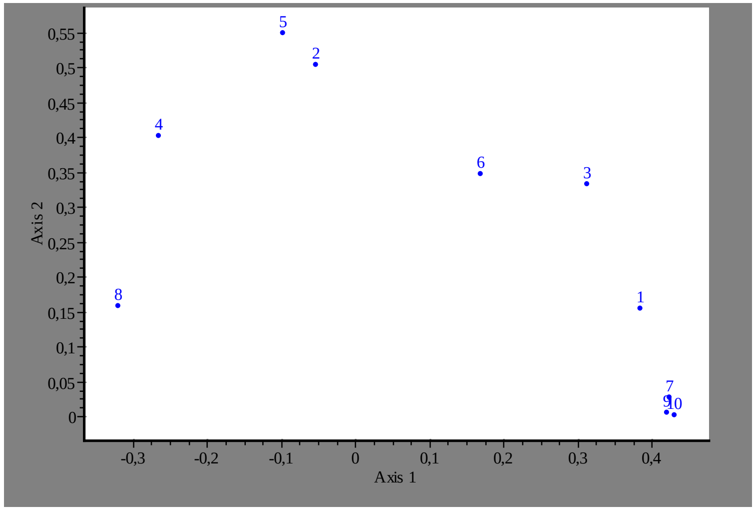
Procento variability znaků vyjádřené příslušnou komponentou

```
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 znaků

„*Euclidean option*“ - poloha znaků vyjadřuje polohy vektorů příslušných znaků

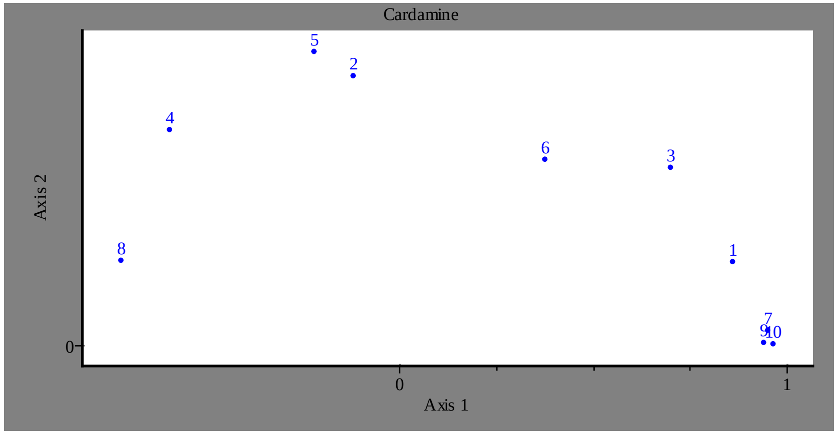
```
EIGENVECTORS AS COORDINATES OF VAR.

SCORES FOR VARIABLES
  VARIABLE   1         .384        .156        .265
  VARIABLE   2        -.053        .504       -.380
  VARIABLE   3         .312        .334       -.133
  VARIABLE   4        -.265        .403        .053
  VARIABLE   5        -.099        .549       -.242
  VARIABLE   6         .168        .347        .627
  VARIABLE   7         .424        .028       -.004
  VARIABLE   8        -.320        .158        .556
  VARIABLE   9         .420        .006        .013
  VARIABLE  10         .430        .003       -.049
```



Ordinace znaků

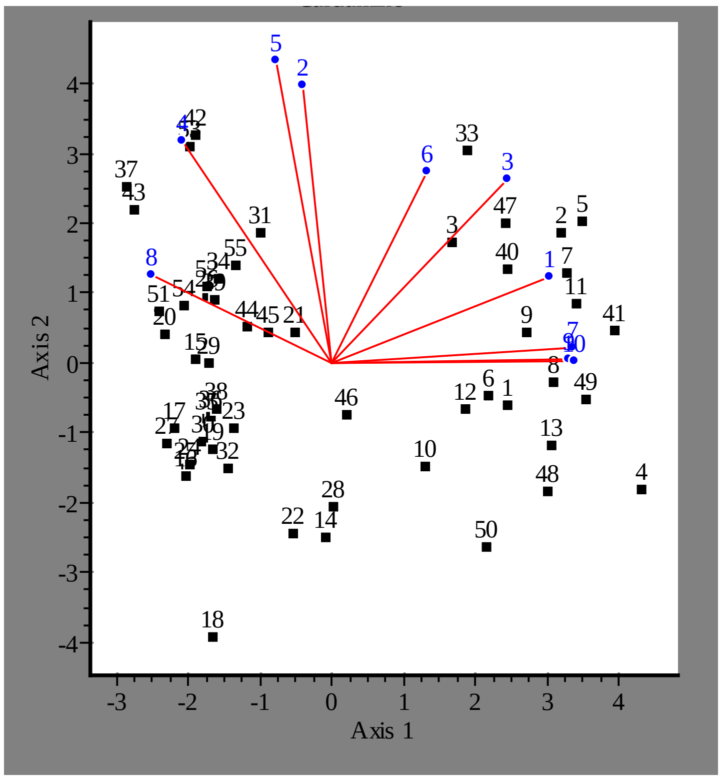
„*Rohlf mixed option*“ - poloha znaků vyjadřuje hodnoty korelace (případně kovariance) znaků s příslušnými komponentami

```
CORRELATIONS OF VAR. WITH COMPONENTS


SCORES FOR VARIABLES
  VARIABLE   1        .862        .251        .282

  VARIABLE   2       -.118        .811       -.404

  VARIABLE   3        .700        .537       -.141

  VARIABLE   4       -.594        .648        .057

  VARIABLE   5       -.221        .884       -.257

  VARIABLE   6        .377        .559        .666

  VARIABLE   7        .952        .045       -.004

  VARIABLE   8       -.718        .255        .591

  VARIABLE   9        .941        .010        .014

  VARIABLE  10        .965        .004       -.052
```

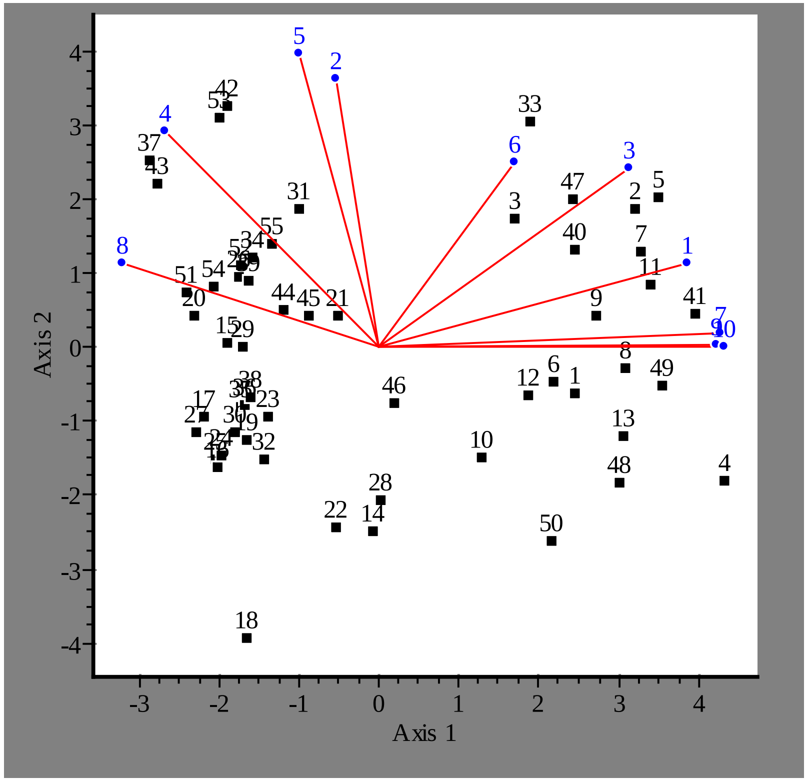


Ordinace objektů a znaků (biplot)

Euclidean biplot - poloha znaků vyjadřuje polohy vektorů příslušných znaků



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



Typy PCA

Centrovaná PCA - vychází z kovarianční matice znaků

Počáteční bod nové souřadnicové soustavy (hlavních komponent) je posunut z původního počátečního bodu souřadnicové soustavy původních znaků do centroidu „oblaku“ ordinovaných objektů (hypotetického bodu představujícího „průměrný objekt“). Vzdálenosti mezi objekty v nové souřadnicové soustavě zůstávají stejné jako v soustavě původní.

Standardizovaná PCA - vychází z korelační matice znaků

Její součástí je standardizace dat (srovnej vztah kovariance a korelace: vydělením hodnot kovariance součinem směrodatných odchylek získáme korelační koeficient). Počáteční bod nové souřadnicové soustavy se přesouvá do centroidu oblaku objektů a zároveň jsou původní znaky přeškálované tak, aby měly jednotkový rozptyl. Nové znaky (hlavní komponenty) sice nemají jednotkový rozptyl, ale jejich rozptyl odpovídá příslušným vlastním číslům (suma vlastních čísel se rovná počtu znaků).

Necentrovaná PCA - vychází z matice skalárních součinů znaků (*cross-products between variables*)

Nezahrnuje ani standardizaci ani centrování. Počáteční bod nové souřadnicové soustavy je ve stejném místě jako u původní soustavy. Tato technika se používá v některých ekologických aplikacích.

Analýza hlavních koordinát

(metrické mnohorozměrné škálování, PCoA – *principal coordinate(s) analysis*, *metric multidimensional scaling, classical scaling*)

Rozmístění souboru objektů v novém prostoru definovaném hlavními koordináty (novými osami).

Vzájemné (euklidovské) vzdálenosti objektů odrážejí vztahy mezi původními objekty měřené libovolným koeficientem podobnosti nebo vzdálenosti.

Binární znaky

Vícestavové kvalitativní znaky

Smíšená data

(1) Primární matice dat → sekundární matice vzdáleností → → symetrická matice, ekvivalentní korelační nebo kovarianční matici používané v PCA

(2) Výpočet vlastních čísel, vlastních vektorů a komponentních skóre

Souřadnice v prostoru určeném hlavními koordinátami nejsou lineárně závislé na hodnotách původních znaků

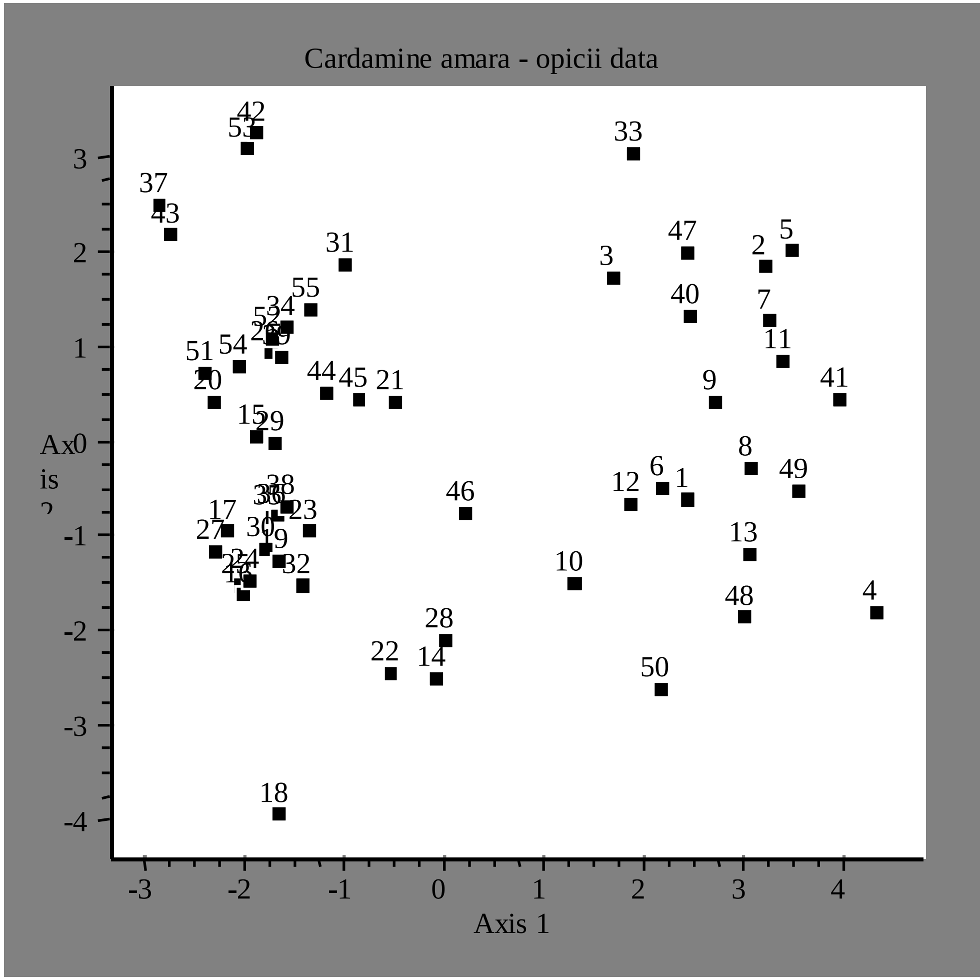
Lze vhodně použít také tehdy, pokud počet znaků převyšuje počet objektů (např. u molekulárních dat)

```
(1) EIGENVALUES
271.59700      139.87590      60.88202       20.98608       17.08320
10.7592        7.30412        5.69138        3.47811        2.34287
.00060         .00001         .00001         .00001          .00001
.00001         .00001         .00001         .00001         0.00000
0.00000        0.00000        0.00000        0.00000        0.00000
0.00000        0.00000        0.00000        0.00000        0.00000
0.00000        0.00000        0.00000        0.00000        0.00000
0.00000        0.00000        0.00000        0.00000        0.00000
0.00000        0.00000        0.00000        0.00000        0.00000
.00001         .00001         .00001         .00001         .00001
.00001         .00001         .00001         .00001         .00001

(2) SUM OF POSITIVE EIGENVALUES
        540.00054

(3) NUMBER OF POSITIVE EIGENVALUES
      33
```

```
(4) EIGENVALUES AS PERCENT
       50.30       25.90       11.27        3.89        3.16
        1.99        1.35        1.05         .64         .43
        0.00        0.00        0.00        0.00        0.00
        0.00        0.00        0.00        0.00        0.00
        0.00        0.00        0.00        0.00        0.00
        0.00        0.00        0.00        0.00        0.00
        0.00        0.00        0.00
(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
      100.00      100.00      100.00      100.00      100.00
      100.00      100.00      100.00      100.00      100.00
      100.00      100.00      100.00      100.00      100.00
      100.00      100.00      100.00      100.00      100.00
      100.00      100.00      100.00
```

```
COORDINATES

 OBJECT  1
    2.44015      -.61730       .29618
 OBJECT  2
    3.20319      1.86563       .44235
 OBJECT  3
    1.68897      1.73017       .51992
```

(zkráceno)



Nemetrické mnohorozměrné škálování
(NMDS – *non-metric multidimensional scaling*)

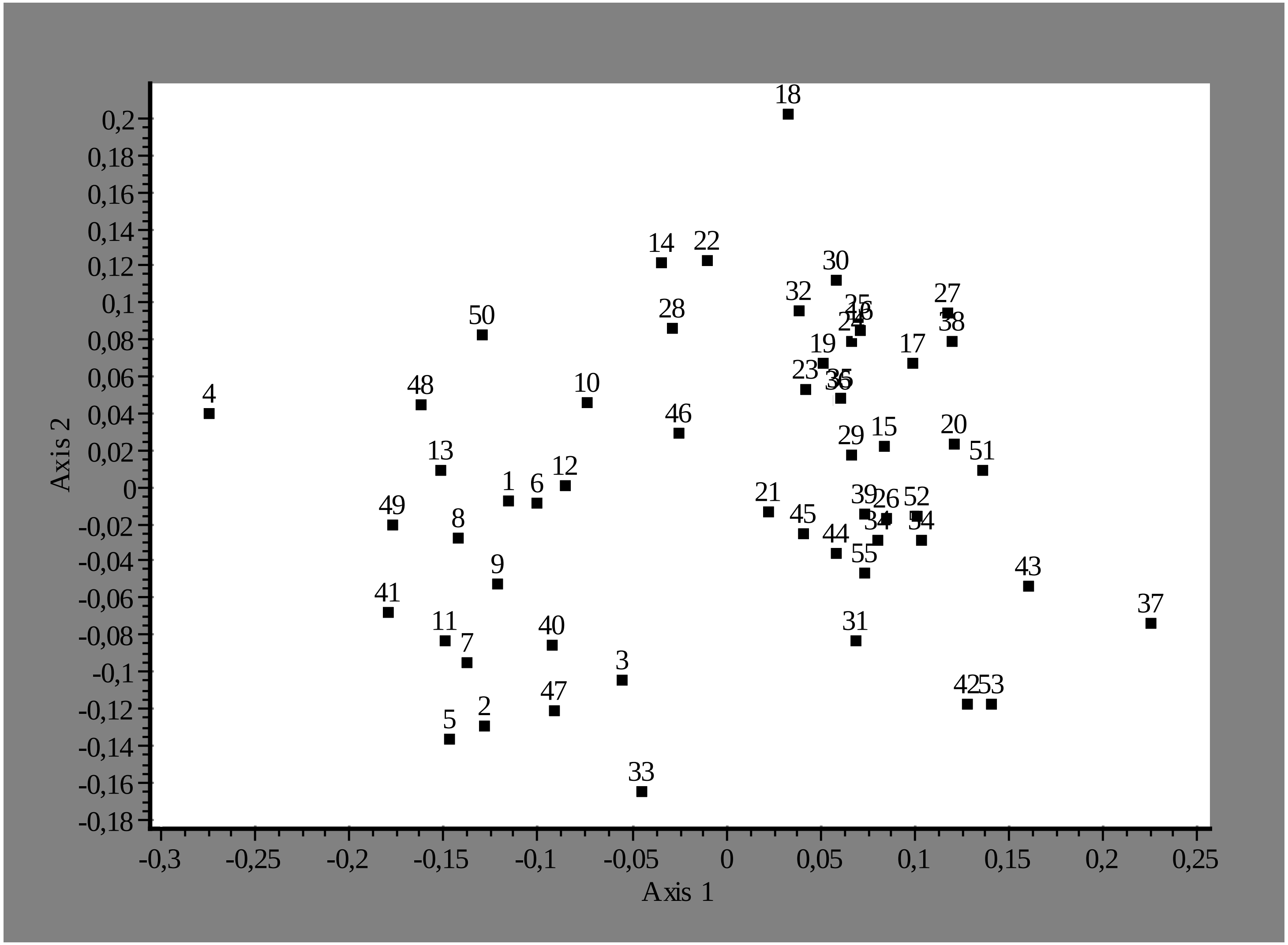
redukování dimenze původního znakového prostoru

zachování pořadí vzdáleností mezi objekty

4 OTU a 6 hodnot nepodobnosti:
$\boldsymbol{\delta_{23} < \delta_{12} < \delta_{34} < \delta_{13} < \delta_{24} < \delta_{14}}$

OTU = body v euklidovském prostoru, jejich vzájemné vzdálenosti jsou: d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}
Předpokládá se, že tyto vzdálenosti se perfektně shodují s pozorovanými nepodobnostmi, pokud:

$\mathbf{d_{23} \leq d_{12} \leq d_{34} \leq d_{13} \leq d_{24} \leq d_{14}}$


Axis 2
0,2
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
-0,02
-0,04
-0,06
-0,08
-0,1
-0,12
-0,14
-0,16
-0,18
-0,3
-0,25
-0,2
-0,15
-0,1
-0,05
0
0,05
0,1
0,15
0,2
0,25
Axis 1
18
14
22
30
32
25
16
27
50
28
24
38
19
17
48
10
23
35
36
4
46
29
15
20
13
12
51
1
6
21
39
26
52
49
8
45
34
54
44
55
9
43
41
37
11
40
31
7
3
47
42
53
2
5
33

stres (*stress*)

míra shody vzdáleností na ordinačním diagramu a původními hodnotami nepodobnosti

pod 0,05 - shoda výborná

0,05-0,10 - shoda uspokojivá

0,10-0,15 - shoda akceptovatelná s výhradami

```
DIMENSIONALITY USED BELOW =    5
   CHANGE VERY SMALL, FINAL STRESS =         .01672606

DIMENSIONALITY USED BELOW =    4
   CHANGE VERY SMALL, FINAL STRESS =         .02733298

DIMENSIONALITY USED BELOW =    3
   CHANGE VERY SMALL, FINAL STRESS =         .04378727

DIMENSIONALITY USED BELOW =    2
   CHANGE VERY SMALL, FINAL STRESS =         .10600522
```