Metody klasifikace a shlukování

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie,
Chemickotechnologická fakulta, Univerzita Pardubice,
Studentská 573, 532 10 Pardubice,

http: //meloun.upce.cz,

milan.meloun@upce.cz

http://meloun.upce.cz

Domů - Internet Explorer

http://meloun.upce.cz/

Soubor Úpravy Zobrazit Oblíbené položky

Převést Vybrat

Search Web Upgrade Your Toolbar Now Mail Bookmarks

Navrhované weby

Milan Meloun

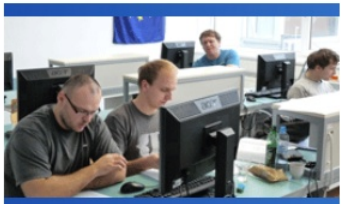
Profesor analytické chemie and chemometrie na Univerzitě Pardubice.

Vítejte na osobních stránkách

Prof. RNDr. Milana Melouna, DrSc., profesora analytické chemie a chemometrie na Univerzitě Pardubice

Menu

- Domů
- Aktuality
- Životopis
- Výuka
- Výzkum
- Publikace
- Ke stažení
- Fotogalerie
- Užitečné odkazy



Výuka

Blok obsahuje studijní materiály jako slidy přednášek, sylaby ke zkoušce, zkušební otázky, vzorové semestrální práce studentů a doporučenou literaturu k

Výzkum

Blok obsahuje oddíl **Projekty** s názvy grantů a projektů, dále tituly PhD dizertací a názvy diplomových prací ke studiu rovnováh v roztocích. Oddíl **Rovnováhy**

Publikace

Blok obsahuje 6 oddílů publikační aktivity, jako jsou **Kompendium, Původní práce, Doporučené knihy, Konference, Patenty a Citační index**. Oddíl

Start Adresa Odkazy Plocha

a pak kliknout na Výuka

Výuka - Internet Explorer

http://meloun.upce.cz/vyuka

Výuka

Soubor Úpravy Zobrazit Oblíbené položky Nástroje Nápověda

Převést Vybrat

Search Web Upgrade Your Toolbar Now Mail Bookmarks

Navrhované weby

Milan Meloun

Profesor analytické chemie and chemometrie na Univerzitě Pardubice.

Menu

- Domů
- Aktuality
- Životopis
- Výuka
  - Analytická chemie
  - Chemometrie I.
  - Chemometrie II.
  - Chemometrie III.
  - Metody a vzorové úlohy
  - Licenční studium - ARISTOTELES,GALILEO
  - Kurzy
- Výzkum
- Publikace
- Ke stažení
- Fotogalerie
- Užitečné odkazy
- Návštěvnost stránek
- Hledat
- Kontakt

Výuka (Teaching)

> *Když lidé vyučují, sami se učí.*
> *HOMINES, DUM DOCENT, DISCUNT.*
>
> — Lucius Annaeus Seneca (4 př. n. l. – 65 n. l.) Epistulae morales ad Lucilium 7, 8.

Kliknutím se fotografie zvětší na celou obrazovku. Click on this photo to enlarge it.

Analytická chemie (Analytical chemistry)

Nabízený učební text **ANALYTICKÁ CHEMIE I** podává uceleně náplň přednášek o teoretických základech chemických metod analytické chemie, kvalitativní a kvantitativní

a pak kliknout na Chemometrie II

Chemometrie I. - Internet Explorer

http://meloun.upce.cz/vyuka/chemometrie-i

Chemometrie I.

Soubor Úpravy Zobrazit Oblíbené položky Nástroje Nápověda

Převést Vybrat

Search Web Upgrade Your Toolbar Now Mail Bookmarks

Navrhované weby

Milan Meloun

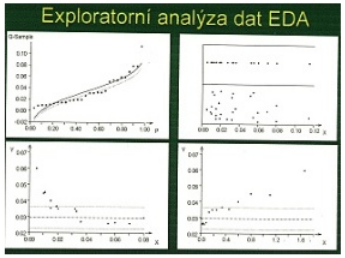
Profesor analytické chemie and chemometrie na Univerzitě Pardubice.

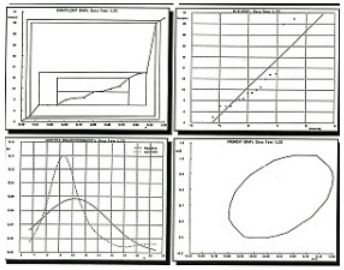
Menu

- Domů
- Aktuality
- Životopis
- Výuka
  - Analytická chemie
  - Chemometrie I.
  - Chemometrie II.
  - Chemometrie III.
  - Metody a vzorové úlohy
  - Licenční studium - ARISTOTELES,GALILEO
  - Kurzy
- Výzkum
- Publikace
- Ke stažení
- Fotogalerie
- Užitečné odkazy
- Návštěvnost stránek
- Hledat
- Kontakt

Chemometrie I - Jednorozměrná data

> *Učený člověk má své bohatství vždy v sobě.*
> *HOMO DOCTUS IN SE SEMPER DIVITIAS HABET.*
> — Phaedrus (kol. 15 př.n.l. - 50 n.l.) Fabulae Aesopiae IV, 22, 1.





A. Potřebné učebnice a vzory semestrálních prací:

1. M. Meloun, J. Militký: **STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT**, Plus, Praha 1994 (1. vydání) nebo East Publishing Praha 1998 (2. vydání), Academia 2004, **INTERAKTIVNÍ STATISTICKÁ ANALÝZA DAT**, Karolinum Praha 2012.
2. M. Meloun, J. Militký: **Kompendium statistického zpracování dat**, Učebnice s CD, Academia Praha 2002 (1. vydání), Academia Praha 2006 (2. vydání), Karolinum Praha 2013 (3. vydání)
3. M. Meloun, J. Militký: **Sbírka úloh - STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT**, Univerzita Pardubice 1997.
4. M. Meloun, J. Militký: **Počítačová analýza vícerozměrných dat v příkladech**, Učebnice s CD, Academia Praha 2002 (1. vydání), **Statistická analýza vícerozměrných dat v příkladech**, Učebnice s CD, Academia Praha 2012 (2. vydání)

Start Adresa Odkazy Plocha

A. Potřebná literatura a vzory semestrálních prací:

1. M. Meloun, J. Militký: **STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT**, Plus, Praha 1994 (1. vydání) nebo East Publishing Praha 1998 (2. vydání), Academia 2004, **INTERAKTIVNÍ STATISTICKÁ ANALÝZA DAT**, Karolinum Praha 2012.
2. M. Meloun, J. Militký: **Kompendium statistického zpracování dat**, Učebnice s CD, Academia Praha 2002 (1. vydání), Academia Praha 2006 (2. vydání), Karolinum Praha 2013 (3. vydání)
3. M. Meloun, J. Militký: **Sbírka úloh - STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT**, Univerzita Pardubice 1997.
4. M. Meloun, J. Militký: **Počítačová analýza vícerozměrných dat v příkladech**, Učebnice s CD, Academia Praha 2002 (1. vydání), **Statistická analýza vícerozměrných dat v příkladech**, Učebnice s CD, Academia Praha 2012 (2. vydání). Data k úlohám.
5. Vzory semestrálních prací studentů v řádném studiu.
6. Vzory semestrálních prací studentů v licenčním studiu.

B. Slidy přednášek v Power-Pointu

1. knihy a zahraniční prameny dat.pdf
2. 1. Charakter zdrojové matice vícerozměrných dat
3. 2. Předúprava a EDA
4. 3. Struktury ve znacích a datech
5. 4. PCA
6. 5. FA
7. 6. CCA
8. 7. DA
9. 8. LR
10. 9.CLU
11. 10. MSD
12. 11. CA

Programy v NCSS 2007:

1. 13. scatter-ncss
2. 14. pca-ncss
3. 15. clu-ncss
4. 16. da-ncss
5. 17. lr-ncss

C. Sylabus a okruhy otázek ke zkoušce:

…… a pak dolu až na Slidy a klik na přednášku 4. PCA

Klasifikace objektů

hledání struktury a vzájemných vazeb v objektech

1) *Diskriminační analýza DA:* nový objekt se zařadí do již existující třídy.

2) *Analýza shluků CLU*: neuspořádanou skupinu objektů lze uspořádat do několika vnitřně sourodých tříd či shluků

3) *Vícerozměrné škálování MDS*: hledá strukturu a vazby mezi objekty na základě jejich podobnosti.

Analyzovaný výběr jsou trénovací data, která obsahují pro každý objekt jak výstup y, tak i hodnoty všech znaků $\boldsymbol{x}$, $\{y; x_1, x_2, \ldots x_m\}$

Klasifikovaný výběr: na základě analyzovaného výběru sestavit predikční model, který umožní zařazení nových objektů do tříd.

KLASIFIKACE OBJEKTŮ

$\boxed{y}$ $\boxed{x_1\ x_2\ x_3\ x_4 \ldots\ldots\ x_m}$

→ PCA

→ FA

→ SHLUKY

→ MDS

DISKRIMINACE DA

(NORMALITA)

$\boxed{y}$ $\boxed{x_1\ x_2\ x_3\ x_4 \ldots\ldots\ x_m}$

TŘÍDA PROMĚNNÉ

LOGISTICKÁ REGRESE, Logit

(NENORMALITA)

Zdrojová matice dat X (n x m)

○ *Data*: x_1 značí délku těla, x_2 značí šířku těla, x_3 je délka předního křídla, x_4 je délka zadního křídla, x_5 je počet průduchů, x_6 je délka tykadla I, x_7 je délka tykadla II, x_8 je délka tykadla III, x_9 je délka tykadla IV, x_{10} je délka tykadla V, x_{11} je počet tykadlových ostnů, x_{12} je délka posledního článku nohy, x_{13} je délka holeně, tibia, x_{14} je délka stehna, x_{15} je délka sosáku, x_{16} je délka kladélka, x_{17} je počet kladélkových trnů, x_{18} je řitní otvor, x_{19} je počet háčků zadních křídel.

Znaky x_1 až x_{19} (sloupce, m = 19)

Objekty (řádky zde indexované od n = 1 do 23)

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
1	21.2	11.0	7.5	4.8	5.0	2.0	2.0	2.8	2.8	3.3	3.0	4.4	4.5	3.6	7.0	4.0	8.0	0.0	3.0
2	20.2	10.0	7.5	5.0	5.0	2.3	2.1	3.0	3.0	3.2	5.0	4.2	4.5	3.5	7.6	4.2	8.0	0.0	3.0
3	20.2	10.0	7.0	4.6	5.0	1.9	2.1	3.0	2.5	3.3	1.0	4.2	4.4	3.3	7.0	4.0	6.0	0.0	3.0
4	22.5	8.8	7.4	4.7	5.0	2.4	2.1	3.0	2.7	3.5	5.0	4.2	4.4	3.6	6.8	4.1	6.0	0.0	3.0
5	20.6	11.0	8.0	4.8	5.0	2.4	2.0	2.9	2.7	3.0	4.0	4.2	4.7	3.5	6.7	4.0	6.0	0.0	3.0
6	19.1	9.2	7.0	4.5	5.0	1.8	1.9	2.8	3.0	3.2	5.0	4.1	4.3	3.3	5.7	3.8	8.0	0.0	3.5
7	20.8	11.4	7.7	4.9	5.0	2.5	2.1	3.1	3.1	3.2	4.0	4.2	4.7	3.6	6.6	4.0	8.0	0.0	3.0
8	15.5	8.2	6.3	4.9	5.0	2.0	2.0	2.9	2.4	3.0	3.0	3.7	3.8	2.9	6.7	3.5	6.0	0.0	3.5
9	16.7	8.8	6.4	4.5	5.0	2.1	1.9	2.8	2.7	3.1	3.0	3.7	3.8	2.8	6.1	3.7	8.0	0.0	3.0
10	19.7	9.9	8.2	4.7	5.0	2.2	2.0	3.0	3.0	3.1	0.0	4.1	4.3	3.3	6.0	3.8	8.0	0.0	3.0
11	10.6	5.2	3.9	2.3	4.0	1.2	1.0	2.0	2.0	2.2	6.0	2.5	2.5	2.0	4.5	2.7	4.0	1.0	2.0
12	9.2	4.5	3.7	2.2	4.0	1.3	1.2	2.0	1.6	2.1	5.0	2.4	2.3	1.8	4.1	2.4	4.0	1.0	2.0
13	9.6	4.5	3.6	2.3	4.0	1.3	1.0	1.9	1.7	2.2	4.0	2.4	2.3	1.7	4.0	2.3	4.0	1.0	2.0
14	8.5	4.0	3.8	2.2	4.0	1.3	1.1	1.9	2.0	2.1	5.0	2.4	2.4	1.9	4.4	2.3	4.0	1.0	2.0
15	11.0	4.7	4.2	2.3	4.0	1.2	1.0	1.9	2.0	2.2	4.0	2.5	2.5	2.0	4.5	2.6	4.0	1.0	2.0
16	18.1	8.2	5.9	3.5	5.0	1.9	1.9	1.9	2.7	2.8	4.0	3.5	3.8	2.9	6.0	4.5	9.0	1.0	2.0
17	17.6	8.3	6.0	3.8	5.0	2.0	1.9	2.0	2.2	2.9	3.0	3.5	3.6	2.8	5.7	4.3	10.0	1.0	2.0
18	19.2	6.6	6.2	3.4	5.0	2.0	1.8	2.2	2.3	2.8	4.0	3.5	3.4	2.5	5.3	3.8	10.0	1.0	2.0
19	15.4	7.6	7.1	3.4	5.0	2.0	1.9	2.5	2.5	2.9	4.0	3.3	3.6	2.7	6.0	4.2	8.0	1.0	3.0
20	15.1	7.3	6.2	3.8	5.0	2.0	1.8	2.1	2.4	2.5	4.0	3.7	3.7	2.8	6.4	4.3	10.0	1.0	2.5
21	16.1	7.9	5.8	3.7	5.0	2.1	1.9	2.3	2.6	2.9	5.0	3.6	3.6	2.7	6.0	4.5	0.0	1.0	2.0
22	19.1	8.8	6.4	3.9	5.0	2.2	2.0	2.3	2.4	2.9	4.0	3.8	4.0	3.0	6.5	4.5	0.0	1.0	2.5
23	15.3	6.4	5.3	3.3	5.0	1.7	1.6	2.0	2.2	2.5	5.0	3.4	3.4	2.6	5.4	4.0	0.0	1.0	2.0

PŘÍKLAD 1.1 ***Popisné statistiky jednorozměrné analýzy zdrojové matice dat*** *Hrách*

Zdrojová matice dat *Hrách obsahuje znaky smyslového posouzení charakteristik rozličných odrůd hrachu. Objekty zde představují vzorky pěti různých odrůd hrachu A až E, které byly sklízeny v pěti rozličných obdobích 1 až 5. Výsledná zdrojová matice o 12 znacích převážně smyslových charakteristik obsahuje 60 vzorků hrachu. Posouzení každého objektu hrachu bylo provedeno 10 porotci dvojím odhadem tak, že smyslové charakteristiky byly bodovány ve stupnici od 1 (nejhorší) do 9 (nejlepší). Tak bylo získáno 1200 řádků (objektů) postupem: 60 vzorků × 2 hodnocení × 10 porotců. V praxi se data obvykle průměrují, aby se kompenzovaly rozdíly v subjektivní škále přísnosti jednotlivých porotců. Výsledkem je pro každý z šedesáti objektů průměrná hodnota senzorického hodnocení. Cílem úlohy je: 1*. průměrovat data, 2. vynést původní data do grafu a 3. vypočítat popisné jednorozměrné statistiky.

Data: *Zdrojová matice dat n = 1200, m = 12 byla průměrována a výsledkem byla matice 60 × 12. Obsahovala průměrné hodnoty senzorického hodnocení pro znaky ve sloupcích: Aro je aroma, Slad je sladkost, Med je medovost, Bez je bezchuťovost, Klas je klasovost, Tvrd je tvrdost, Bel je bělost, Bar1 je barva1, Bar2 je barva2, Bar3 je barva3, Slup je slupka, Ztr je ztráta.*

Objekt	*Aro*	*Slad*	*Med*	*Bez*	*Klas*	*Tvrd*	*Bel*	*Bar1*	*Bar2*
	Bar3	*Slup*	*Ztr*						
B5	6.48	6.66	4.56	2.2	2.91	3.47	4.72	5.59	5.73
	5.99	4.26	3.25						
C4	5.75	6.09	3.81	2.32	4.03	3.77	4.17	5.73	5.75
	5.32	3.82	3.38						
B2	3.94	4.12	2.44	3.63	5.77	5.39	4.77	6.67	5.11
	4.60	3.5	3.03						

STATISTICA Cz - [Data: 11Hrach.sta (13s krát 82ř)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Okno Nápověda

Přidat do sešitu Přidat do protokolu

Arial 10 Proměnné Případy

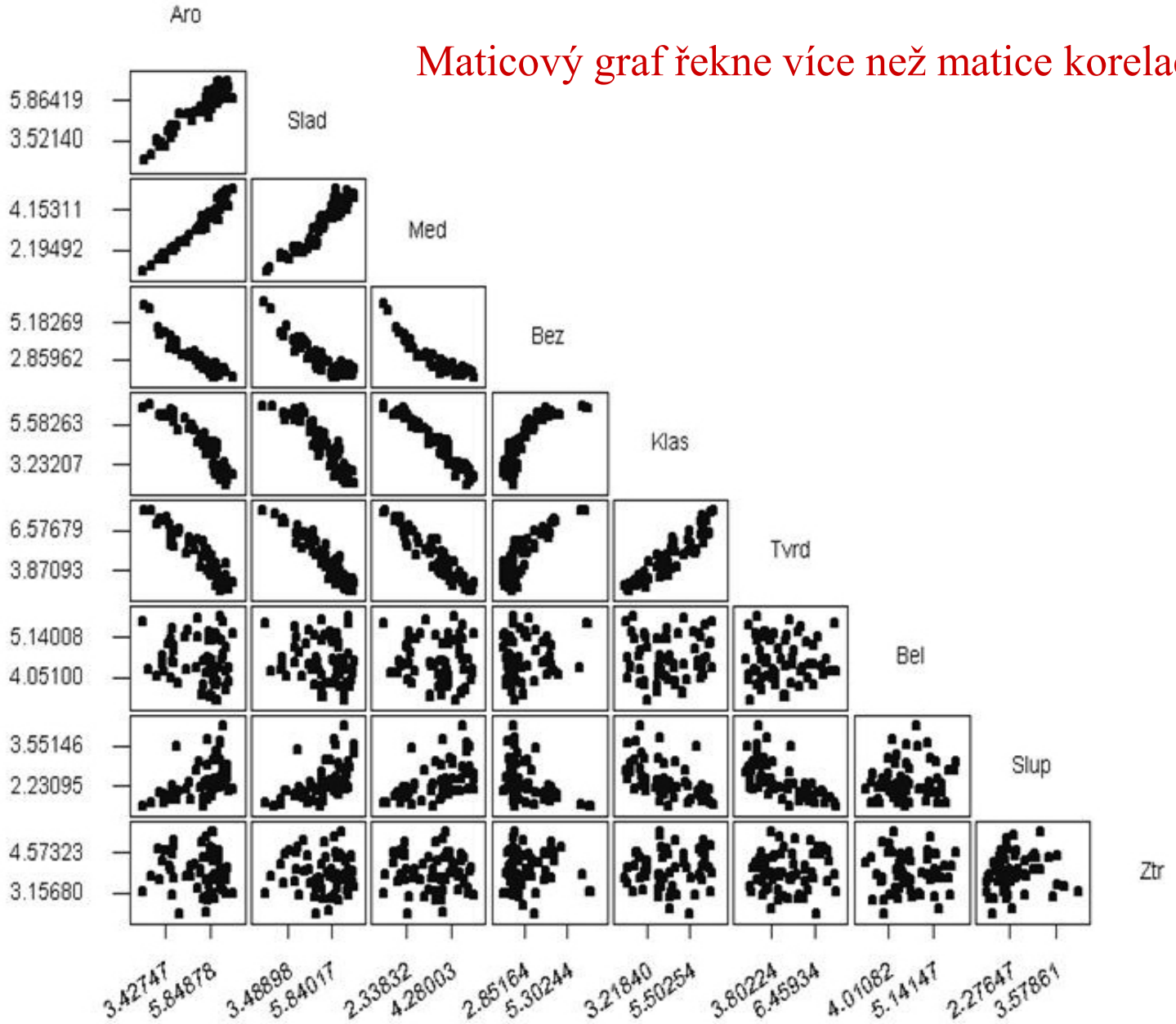
Zdrojová matice dat

	1 Objekt	2 Aro	3 Slad	4 Med	5 Bez	6 Klas	7 Tvrd	8 Bel	9 Bar1	10 Bar2	11 Bar3	12 Slup	13 Ztr
1	B5	6,480	6,660	4,560	2,200	2,910	3,470	4,720	5,585	5,735	5,985	4,260	3,250
2	C4	5,750	6,090	3,810	2,320	4,030	3,770	4,170	5,730	5,745	5,325	3,820	3,380
3	B2	3,940	4,120	2,440	3,630	5,770	5,390	4,770	6,665	5,105	4,595	3,500	3,030
4	D5	6,600	6,120	4,440	1,930	3,310	4,460	4,860	5,160	5,740	6,565	2,120	3,940
5	D4	5,680	5,980	3,800	2,120	3,850	4,140	5,030	5,635	5,220	5,480	2,380	5,160
6	E2	4,740	4,660	2,880	2,940	5,650	5,770	5,310	5,940	5,270	5,890	1,750	3,640
7	B5	6,310	6,130	4,780	1,940	2,700	3,260	5,070	5,710	5,370	6,365	3,650	4,550
8	C5	6,200	6,020	4,650	1,780	3,120	3,740	5,250	5,655	5,475	5,960	2,510	3,800
9	C2	3,790	3,880	2,310	3,520	6,240	5,730	5,390	6,300	5,135	5,230	2,010	4,110
10	A4	5,680	6,340	3,750	2,790	4,170	3,870	4,520	4,920	5,760	4,570	2,970	4,410
11	D4	6,100	6,090	3,990	2,070	4,260	4,250	4,010	5,020	6,175	5,380	2,500	4,600
12	B1	3,410	3,180	1,820	4,640	6,240	7,430	4,260	4,835	5,955	4,550	1,850	4,270
13	D4	5,890	6,090	3,990	2,290	3,900	4,590	4,530	5,890	5,630	3,815	2,200	4,710
14	E4	5,770	5,320	3,880	2,260	4,220	4,990	5,050	5,335	5,590	5,540	2,160	3,650
15	B1	3,390	3,280	1,980	4,500	6,040	7,140	4,350	5,090	5,580	4,400	2,030	4,620
16	B5	6,570	6,880	4,830	1,970	2,920	3,390	3,860	4,615	6,665	6,660	2,220	3,580
17	D4	5,860	6,180	3,940	2,200	3,800	4,910	4,350	5,180	6,255	5,760	2,270	3,610
18	C2	3,960	4,480	2,300	3,940	6,230	6,410	4,470	5,105	5,720	4,970	2,050	4,530
19	A5	6,220	6,790	4,260	2,400	2,630	3,160	5,680	7,010	4,860	3,255	3,040	4,590
20	C3	5,110	5,250	3,090	3,270	5,280	5,240	5,610	6,595	5,110	3,950	2,740	4,230
21	B2	3,770	3,970	2,180	4,370	6,470	6,550	4,950	6,055	5,310	4,395	2,210	4,760
22	B5	7,090	6,090	5,180	1,740	2,570	3,180	5,230	5,920	5,515	4,115	2,090	3,100
23	D4	5,720	5,300	3,730	2,340	3,950	4,800	3,640	4,000	6,805	6,755	1,740	2,930
24	B1	3,220	3,210	1,950	4,420	6,240	7,270	4,600	6,030	5,600	4,165	1,680	3,580
25	A5	6,110	6,620	4,290	2,580	3,200	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4,310
26	D4	6,070	6,270	3,980	2,190	3,890	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
27	A1	2,660	2,660	1,430	6,100	6,670	7,750	4,270	[illegible]	5,630	4,525	1,650	3,780
28	B3	5,260	5,490	3,460	3,030	4,850	4,170	5,220	5,410	5,415	6,120	3,080	3,950
29	C2	3,720	4,350	2,200	4,080	6,500	6,270	4,990	5,535	5,560	5,335	1,820	3,920
30	D3	5,430	5,190	3,470	2,400	4,430	5,260	4,460	4,780	5,720	5,895	1,610	3,770
31	B5	6,550	6,570	4,710	2,120	3,060	3,430	3,760	4,430	6,450	6,380	2,630	3,850
32	E3	5,530	5,410	3,680	2,470	4,720	5,780	3,880	4,335	6,470	6,790	1,800	2,950
33	C3	4,710	4,680	2,680	3,190	5,320	5,920	4,320	4,765	6,215	4,860	2,330	4,020
34	A5	6,280	7,030	4,910	2,380	2,190	2,600	4,560	5,905	5,585	4,950	3,630	3,380
35	D4	5,920	5,820	3,750	2,060	3,880	3,870	4,530	5,190	5,825	5,630	2,450	3,710
36	E4	6,090	5,720	3,800	1,940	4,440	4,450	3,940	4,625	6,510	7,175	2,180	2,990
37	B5	6,370	6,500	4,680	2,140	2,890	3,530	4,600	5,735	5,560	4,065	2,880	4,340
38	A4	5,710	5,680	3,970	2,650	4,390	3,720	5,320	6,280	5,120	6,075	2,390	2,550
39	C3	4,530	5,030	2,640	3,120	5,860	4,920	5,150	6,965	5,125	4,275	2,130	3,740
40	C4	5,950	6,280	4,040	2,190	3,930	3,610	4,120	5,395	5,815	4,505	3,090	5,340
41	D3	5,510	5,410	3,720	2,780	4,760	5,270	3,880	4,280	6,355	5,325	2,250	5,030
42	A1	3,100	3,430	1,800	4,860	6,220	7,070	4,140	5,275	5,580	3,695	2,050	4,730
43	A5	6,500	6,680	4,770	2,230	2,090	2,870	5,510	6,375	4,845	4,785	2,710	3,700
44	D3	5,460	5,410	3,270	2,970	5,150	4,980	3,610	4,305	6,600	5,435	2,370	4,440
45	A2	3,750	4,300	2,220	4,270	6,100	6,270	4,060	5,140	5,870	4,220	2,230	5,010
46	C4	5,860	5,270	3,730	2,500	3,860	4,300	4,150	4,495	6,230	6,140	2,110	3,290
47	A5	6,160	6,970	4,800	2,500	2,870	3,170	4,270	5,315	6,090	5,255	3,350	4,430
48	B2	3,870	3,880	2,230	4,060	5,990	6,310	4,450	5,530	5,785	4,980	1,930	3,620
49	B5	6,240	5,800	4,260	2,130	3,240	3,420	5,120	5,940	5,465	4,775	3,110	3,980

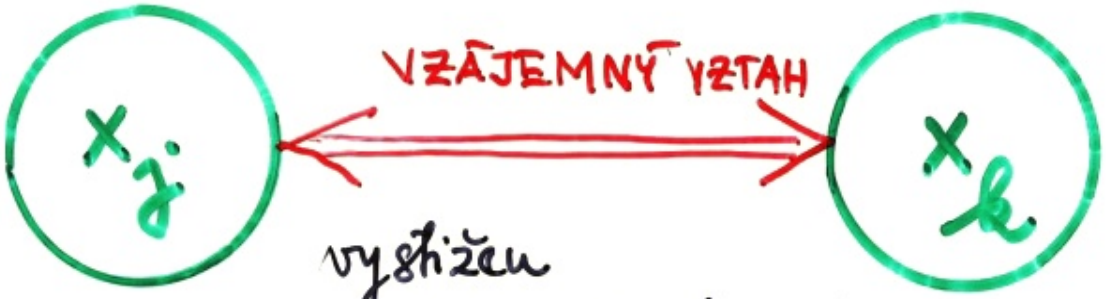
Pro nápovědu stiskněte F1 | Ř1;S11 | 5,985 | Filtr - | Váhy: VYPN | ABC | 123 | ZÁZN

Start | Total Commander 6.0 - P... | Doručená pošta - Micros... | Prezentace2 | STATISTICA Cz - [Dat... | 10:28

Maticový graf řekne více než matice korelačních koeficientů



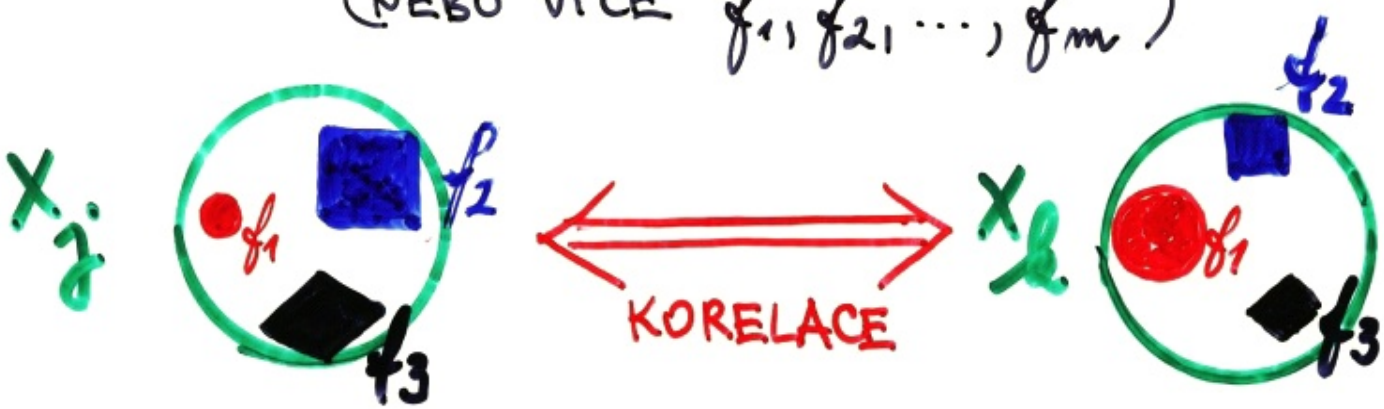
Obr. 1.3 Maticový graf korelace znaků matice *Hrách*.
Na osách jsou min. a maximální hodnoty znaku (SCAN).



vyšetřen
korelačním koeficientem
$r(x_j, x_k)$ či r_{jk}

OTÁZKA: PROČ SPOLU x_j S x_k SOUVISÍ, KORELUJE, ASOCIUJE?

ODPOVĚĎ (FA, PCA): EXISTUJE "NĚCO" SPOLEČNÉHO
EXISTUJE SPOLEČNÁ LATENTNÍ PROMĚNNÁ f
(NEBO VÍCE $f_1, f_2, \dots, f_m$)

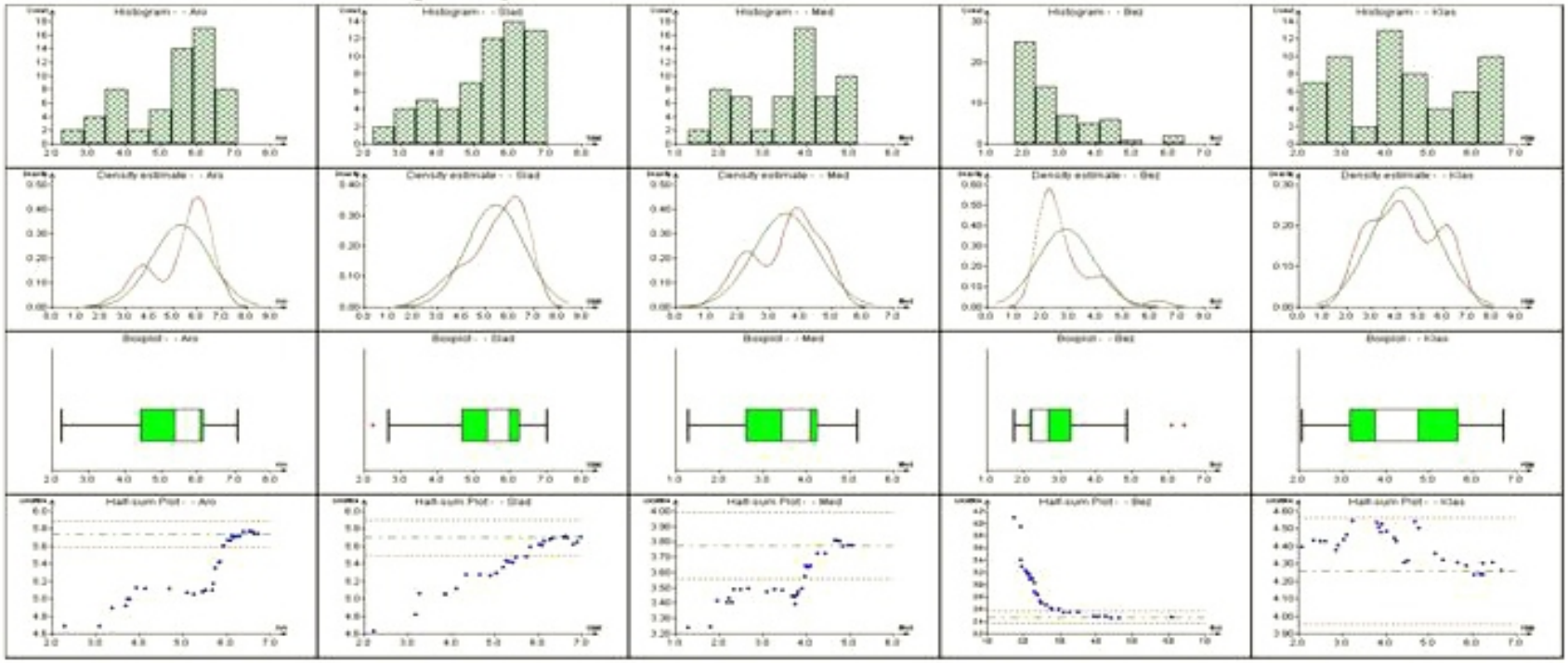


Exploratorní analýza dat

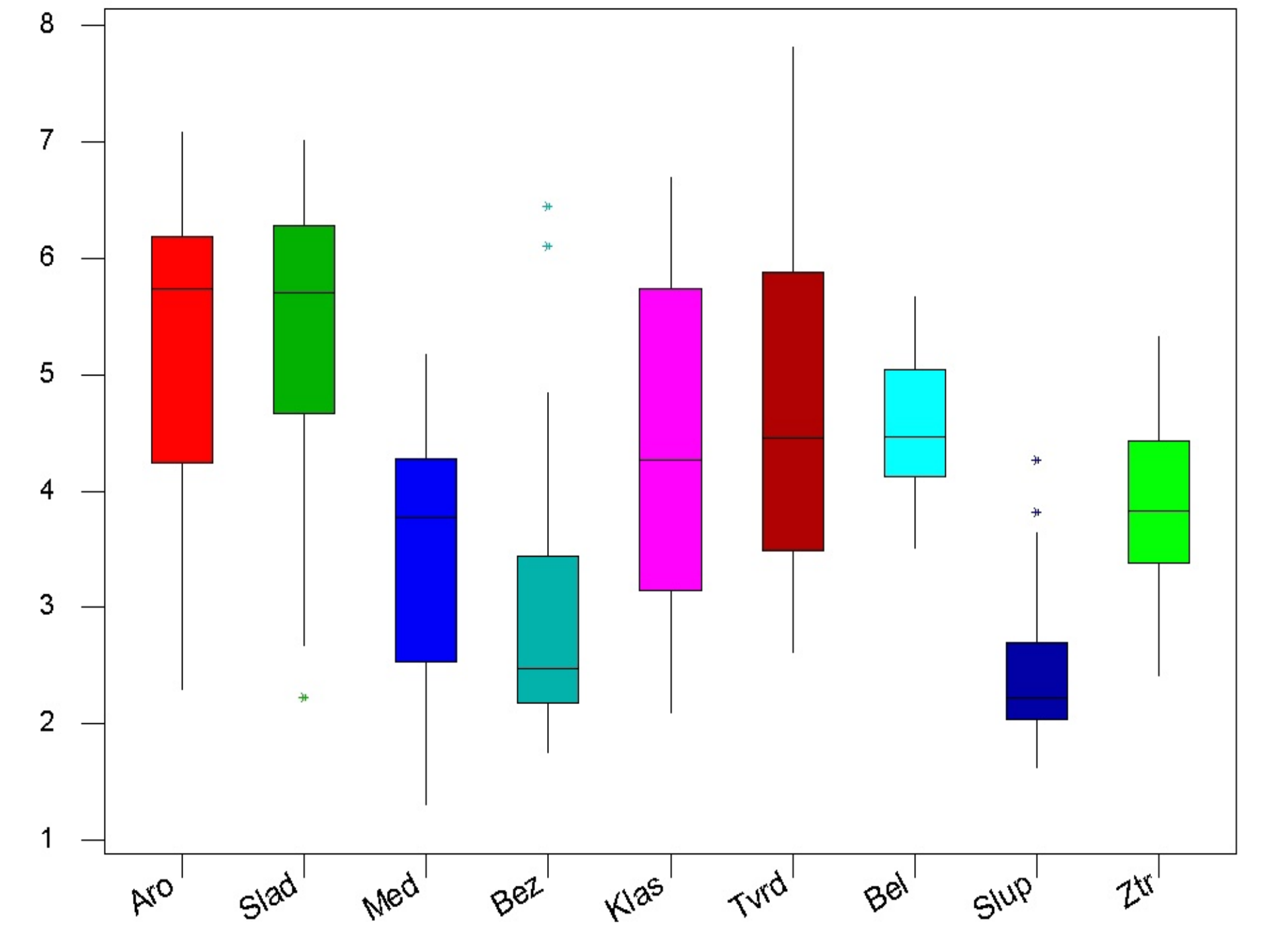
Znaky se vyšetří diagnostikami EDA:

histogram, jádrový odhad hustoty pravděpodobnosti, kvantilový graf, rankitový Q-Q graf, krabicový graf, graf polosum a symetrie, kruhový graf, atd

Zkoumá se vliv typu hrachu (A - E) a období sklizně (1 - 5) na znaky, rozdělení a odlehlé objekty,



Obr. 1.3a Histogram, graf hustoty pravděpodobnosti, krabicový graf a graf polosum pro znaky (zleva) *Aro, Slad, Med, Ber* a *Klas,* (QCEXPERT).


8
7
6
5
4
3
2
1
Aro
Slad
Med
Bez
Klas
Tvrd
Bel
Slup
Ztr

Druhy standardizace (škálování):

- **Sloupcové centrování.** Novým centrem stupnice j-tého znaku je nula. Původním centrem byl průměr prvků $\bar{x}_j$. Sloupcově centrovaná data y_{ij} vzniknou dle $y_{ij} = x_{ij} - \bar{x}_j$, kde

$$\bar{x}_j = \sum_{i=1}^{n} \frac{x_{ij}}{n} .$$

- **Sloupcová standardizace.** Prvky původních dat v j-tého znaku x_{ij} jsou děleny svou směrodatnou odchylkou tedy $y_{ij} = x_{ij} / s_j$, kde s_j je sloupcová směrodatná odchylka j-tého znaku

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n - 1}} .$$

- **Autoškálování.** Kombinace sloupcového centrování a sloupcové standardizace. Jde o studentizaci $y_{ij} = (x_{ij} - \bar{x}_j)/s_j$, která je analogická Z-transformaci pro velké výběry, kdy předpokládáme, že známe μ_j a σ_j

$$y_{ij} = (x_{ij} - \mu_j) / \sigma_j .$$

2.3 Průzkumová analýza vícerozměrných dat

2.3.1 Zobrazení vícerozměrných dat

EDA zobrazení umožňuje:

a) identifikovat vektory $\boldsymbol{x}_i$ nebo jejich složek, které se jeví jako vybočující,

b) indikovat struktury v datech jako jsou shluky, ukazující na heterogenitu výběru nebo přítomnost dílčích výběrů s odlišným chováním.

Zobrazení vícerozměrných dat se dá zařadit:

1. **Zobecněné rozptylové diagramy,**
2. **Symbolové grafy.**

a) Pro případ dvojice náhodných znaků ($m = 2$) lze konstruovat rozptylové grafy.

b) Problémy jsou u vícerozměrných dat pro $m > 2$: je třeba buď volit několik grafů nebo vhodně provést transformace na dvoudimenzionální data.

2. Symbolové grafy

Jednotlivé znaky jsou „kódovány“ s ohledem na jejich hodnoty do geometrických symbolů.

Každému objektu $\boldsymbol{x}_i$ pak odpovídá jistý obrazec složený z těchto symbolů (=znaků) a tak lze v jednom grafu rozlišit více znaků x_j, $j = 1, \ldots, m$.

Základní typy symbolů:

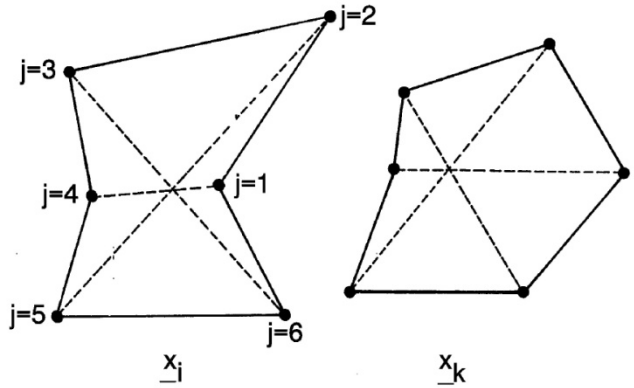
profily,

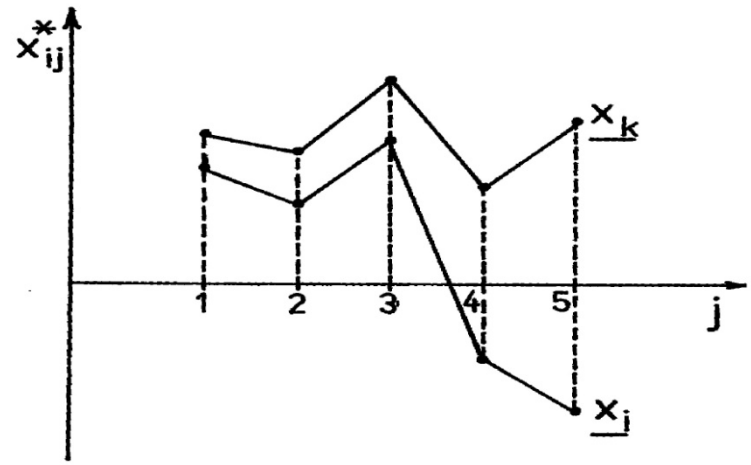
polygony,

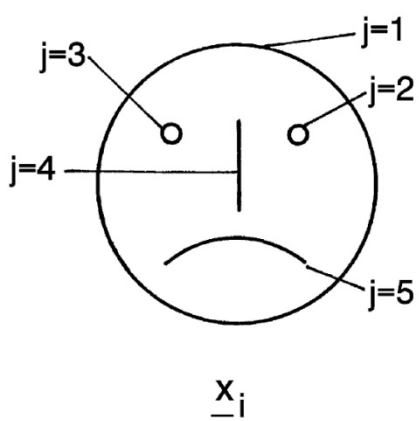
tváře,

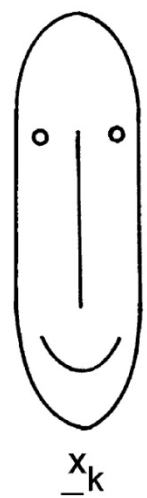
křivky,

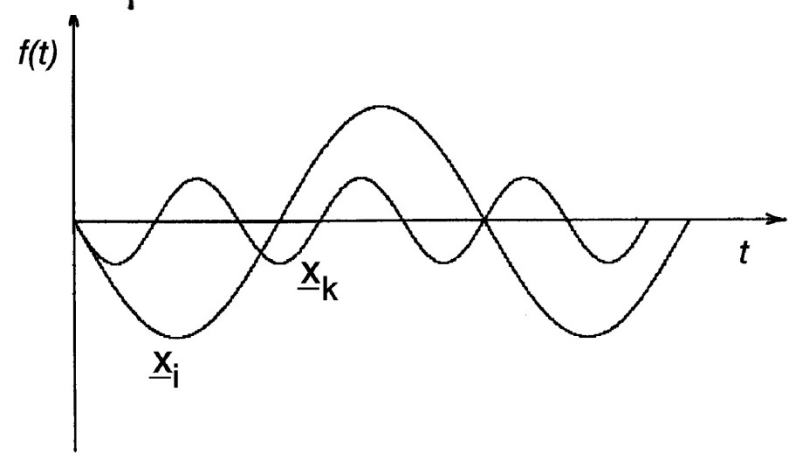
stromy.









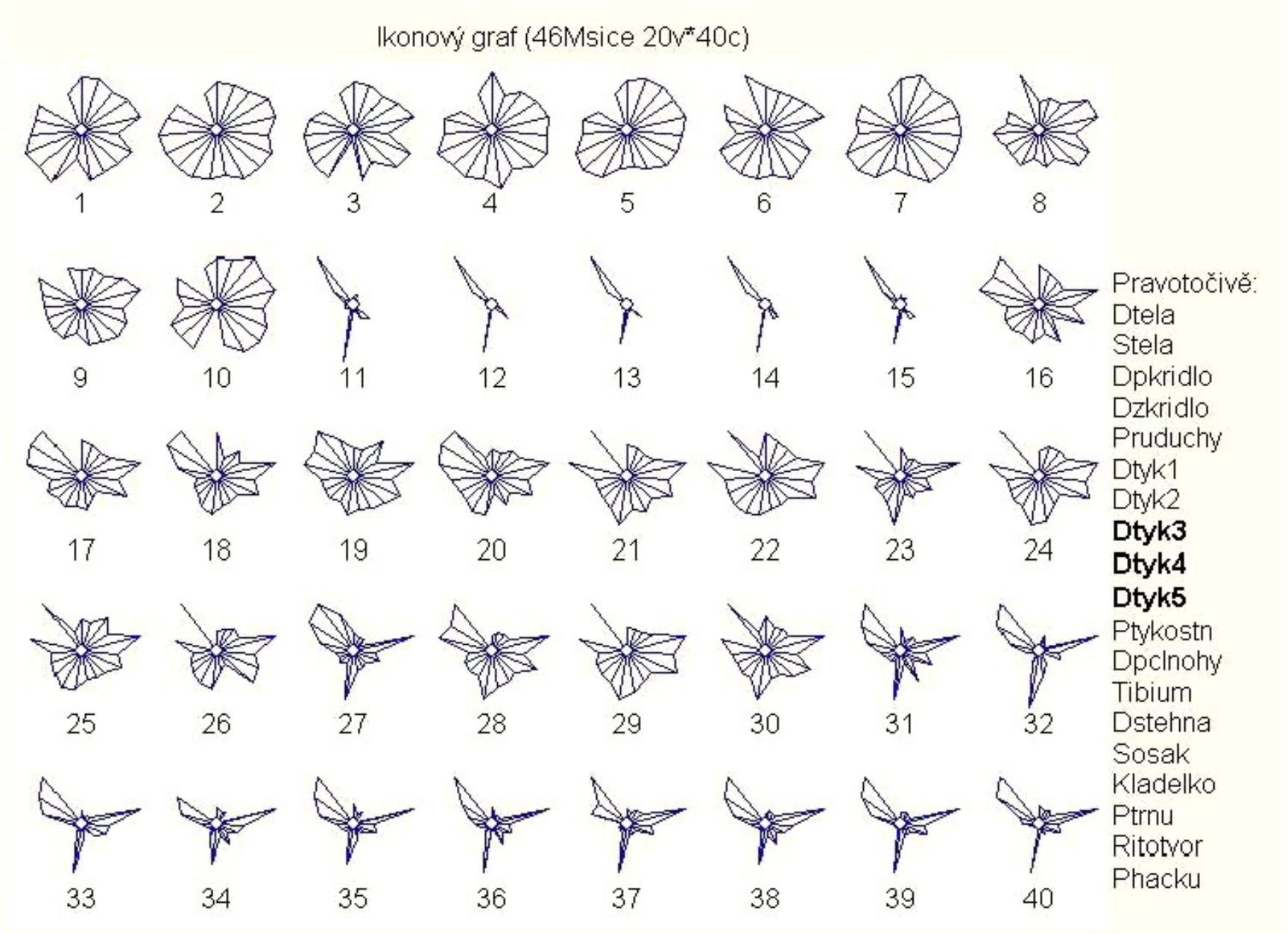


PŘÍKLAD 2.4 *Popis a třídění polétavých mšic*

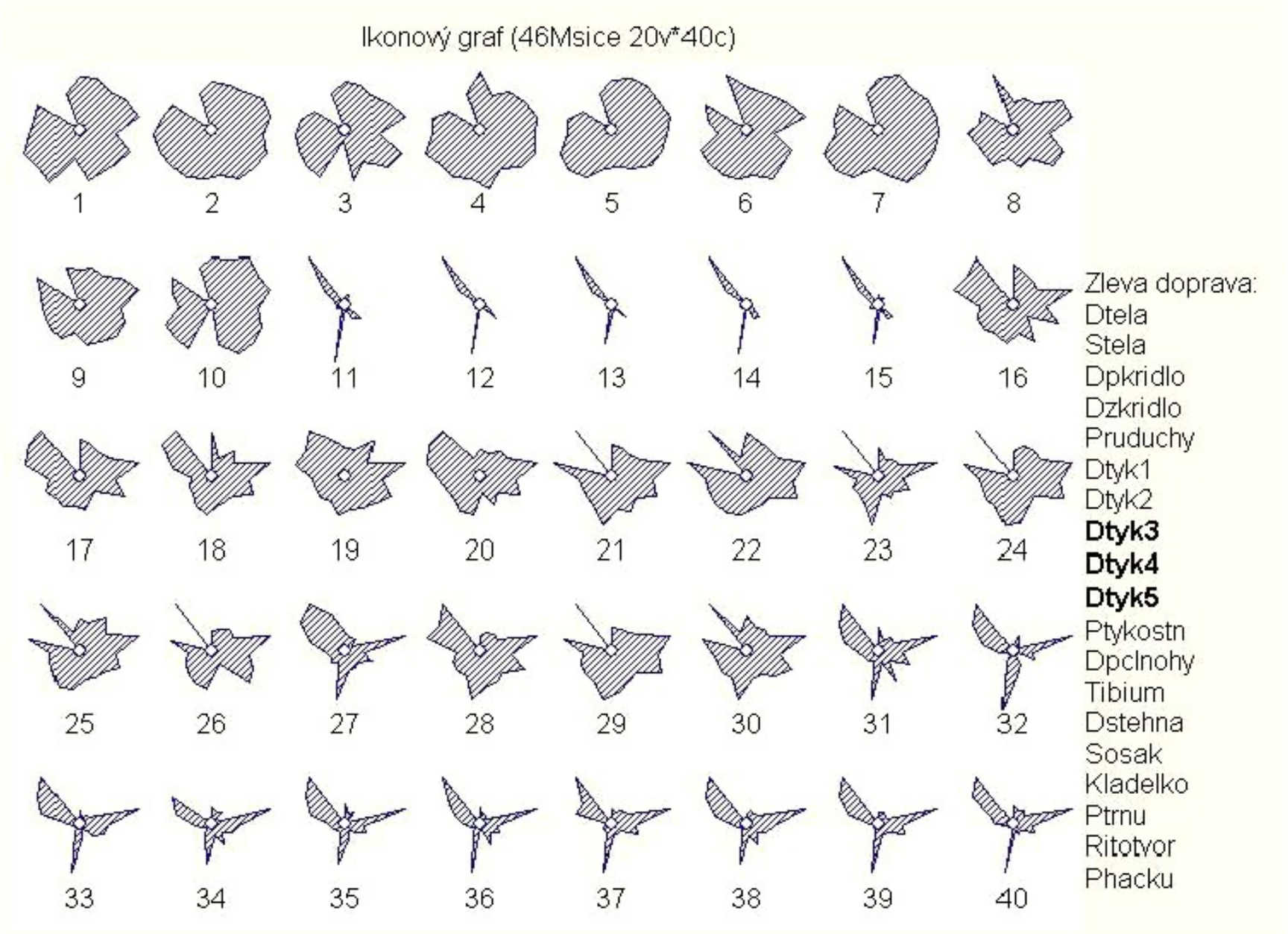
Jeffers (1967) studoval 40 polétavých mšic *(Alate adelges)* pomoci světelné pasti, změřeno 19 znaků: 14 znaků délky a šířky, 4 znaky o počtu, 1 znak binární, přítomnost či absenci. Mšice se obtížně rozlišují dle taxonometrických klíčů. Před PCA je třeba standardizaci dat, protože znaky představují směs délek a počtů.

○ *Data*: x_1 značí délku těla, x_2 značí šířku těla, x_3 je délka předního křídla, x_4 je délka zadního křídla, x_5 je počet průduchů, x_6 je délka tykadla I, x_7 je délka tykadla II, x_8 je délka tykadla III, x_9 je délka tykadla IV, x_{10} je délka tykadla V, x_{11} je počet tykadlových ostnů, x_{12} je délka posledního článku nohy, x_{13} je délka holeně, tibia, x_{14} je délka stehna, x_{15} je délka sosáku, x_{16} je délka kladélka, x_{17} je počet kladélkových trnů, x_{18} je řitní otvor, x_{19} je počet háčků zadních křídel.

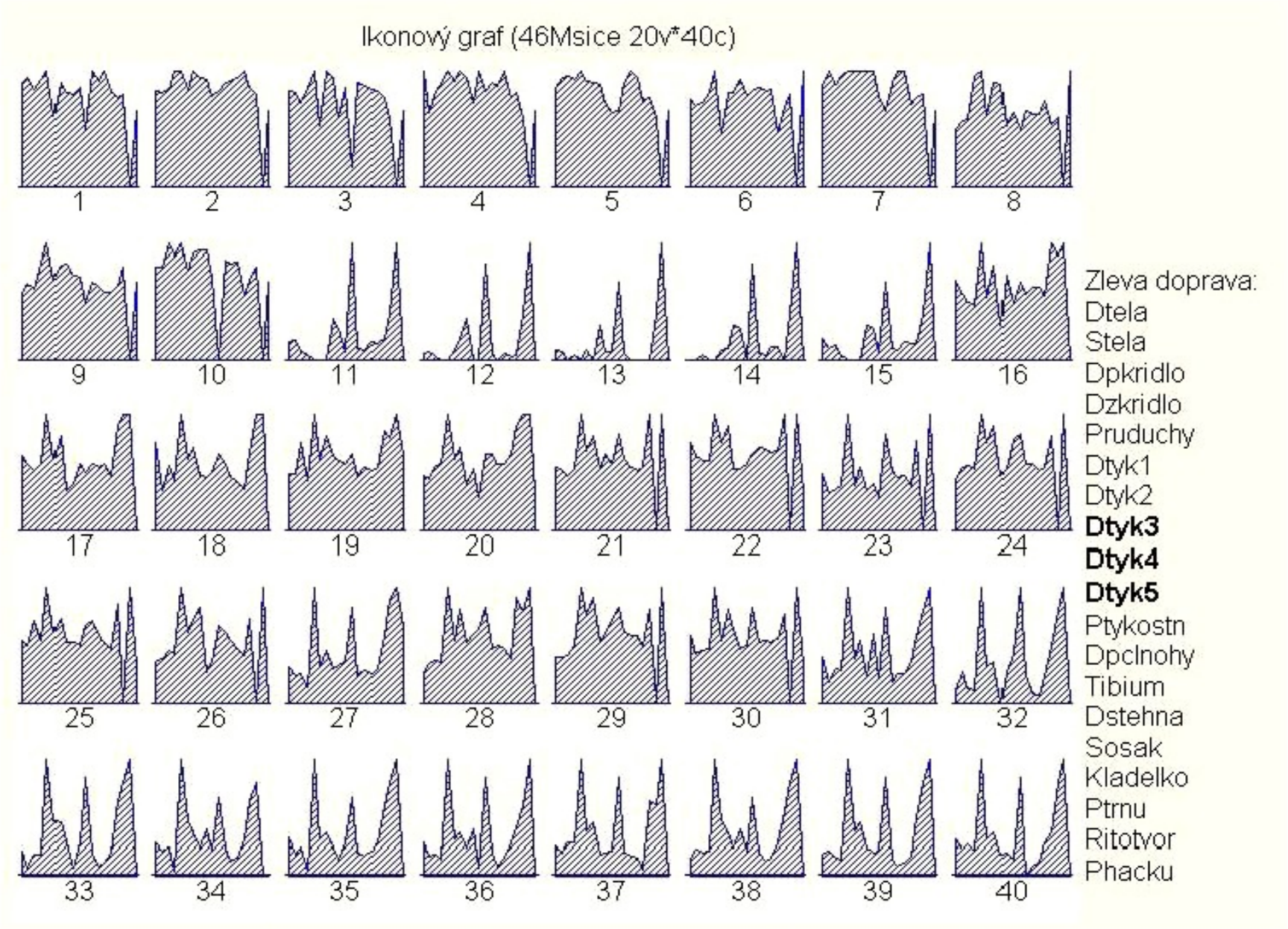
i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
1	21.2	11.0	7.5	4.8	5.0	2.0	2.0	2.8	2.8	3.3	3.0	4.4	4.5	3.6	7.0	4.0	8.0	0.0	3.0
2	20.2	10.0	7.5	5.0	5.0	2.3	2.1	3.0	3.0	3.2	5.0	4.2	4.5	3.5	7.6	4.2	8.0	0.0	3.0
3	20.2	10.0	7.0	4.6	5.0	1.9	2.1	3.0	2.5	3.3	1.0	4.2	4.4	3.3	7.0	4.0	6.0	0.0	3.0
4	22.5	8.8	7.4	4.7	5.0	2.4	2.1	3.0	2.7	3.5	5.0	4.2	4.4	3.6	6.8	4.1	6.0	0.0	3.0
5	20.6	11.0	8.0	4.8	5.0	2.4	2.0	2.9	2.7	3.0	4.0	4.2	4.7	3.5	6.7	4.0	6.0	0.0	3.0
6	19.1	9.2	7.0	4.5	5.0	1.8	1.9	2.8	3.0	3.2	5.0	4.1	4.3	3.3	5.7	3.8	8.0	0.0	3.5
7	20.8	11.4	7.7	4.9	5.0	2.5	2.1	3.1	3.1	3.2	4.0	4.2	4.7	3.6	6.6	4.0	8.0	0.0	3.0
8	15.5	8.2	6.3	4.9	5.0	2.0	2.0	2.9	2.4	3.0	3.0	3.7	3.8	2.9	6.7	3.5	6.0	0.0	3.5
9	16.7	8.8	6.4	4.5	5.0	2.1	1.9	2.8	2.7	3.1	3.0	3.7	3.8	2.8	6.1	3.7	8.0	0.0	3.0
10	19.7	9.9	8.2	4.7	5.0	2.2	2.0	3.0	3.0	3.1	0.0	4.1	4.3	3.3	6.0	3.8	8.0	0.0	3.0
11	10.6	5.2	3.9	2.3	4.0	1.2	1.0	2.0	2.0	2.2	6.0	2.5	2.5	2.0	4.5	2.7	4.0	1.0	2.0
12	9.2	4.5	3.7	2.2	4.0	1.3	1.2	2.0	1.6	2.1	5.0	2.4	2.3	1.8	4.1	2.4	4.0	1.0	2.0
13	9.6	4.5	3.6	2.3	4.0	1.3	1.0	1.9	1.7	2.2	4.0	2.4	2.3	1.7	4.0	2.3	4.0	1.0	2.0
14	8.5	4.0	3.8	2.2	4.0	1.3	1.1	1.9	2.0	2.1	5.0	2.4	2.4	1.9	4.4	2.3	4.0	1.0	2.0
15	11.0	4.7	4.2	2.3	4.0	1.2	1.0	1.9	2.0	2.2	4.0	2.5	2.5	2.0	4.5	2.6	4.0	1.0	2.0
16	18.1	8.2	5.9	3.5	5.0	1.9	1.9	1.9	2.7	2.8	4.0	3.5	3.8	2.9	6.0	4.5	9.0	1.0	2.0
17	17.6	8.3	6.0	3.8	5.0	2.0	1.9	2.0	2.2	2.9	3.0	3.5	3.6	2.8	5.7	4.3	10.0	1.0	2.0
18	19.2	6.6	6.2	3.4	5.0	2.0	1.8	2.2	2.3	2.8	4.0	3.5	3.4	2.5	5.3	3.8	10.0	1.0	2.0
19	15.4	7.6	7.1	3.4	5.0	2.0	1.9	2.5	2.5	2.9	4.0	3.3	3.6	2.7	6.0	4.2	8.0	1.0	3.0
20	15.1	7.3	6.2	3.8	5.0	2.0	1.8	2.1	2.4	2.5	4.0	3.7	3.7	2.8	6.4	4.3	10.0	1.0	2.5
21	16.1	7.9	5.8	3.7	5.0	2.1	1.9	2.3	2.6	2.9	5.0	3.6	3.6	2.7	6.0	4.5	0.0	1.0	2.0
22	19.1	8.8	6.4	3.9	5.0	2.2	2.0	2.3	2.4	2.9	4.0	3.8	4.0	3.0	6.5	4.5	0.0	1.0	2.5
23	15.3	6.4	5.3	3.3	5.0	1.7	1.6	2.0	2.2	2.5	5.0	3.4	3.4	2.6	5.4	4.0	0.0	1.0	2.0



Hvězdy znaků pro každý objekt v EDA dat *Mšice,* (STATISTICA).



Mnohoúhelníky znaků pro každý objekt v EDA dat *Mšice*, (STATISTICA).



Profily znaků pro každý objekt v průzkumové analýze dat *Mšice,* (STATISTICA).

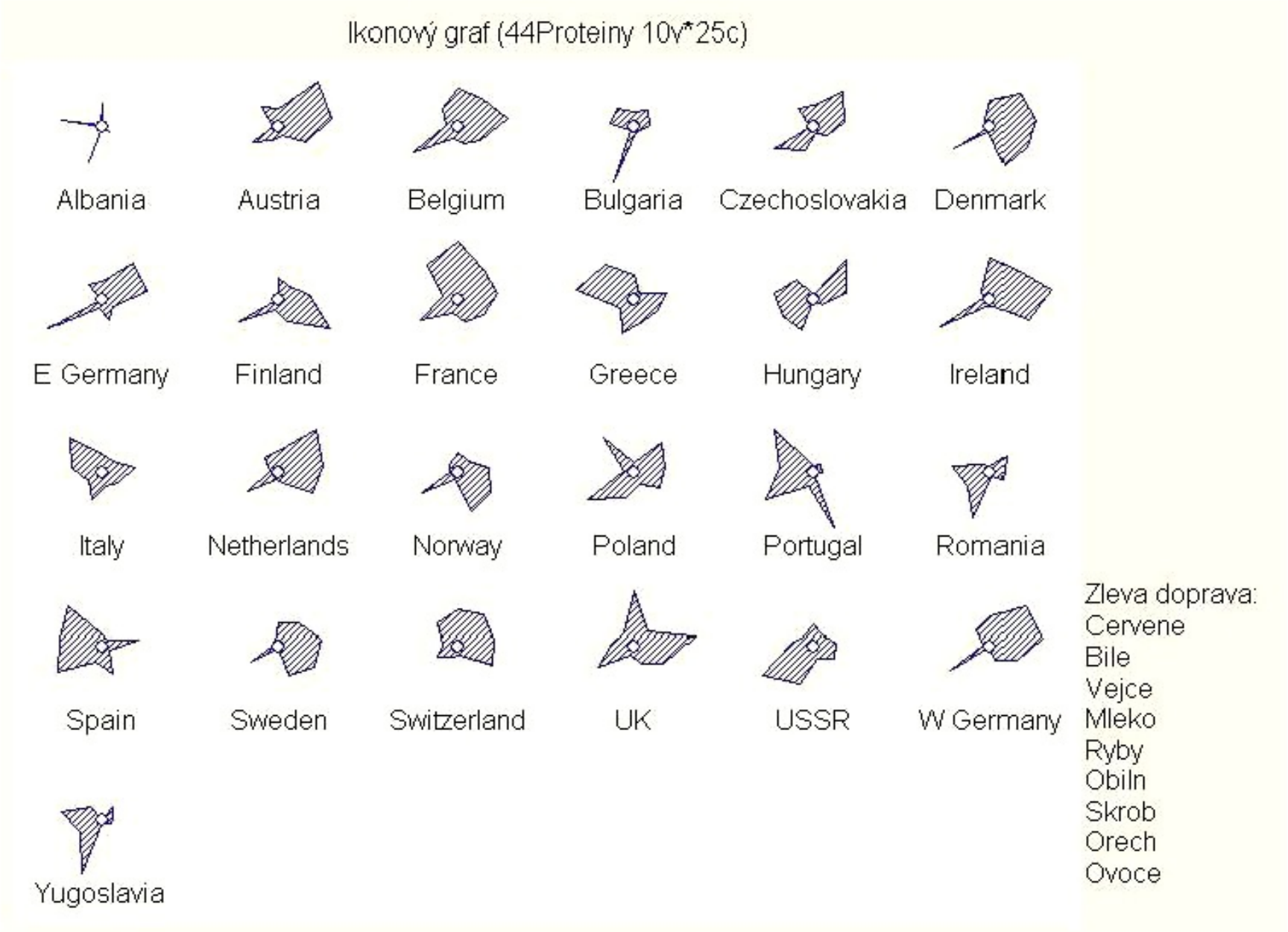
PŘÍKLAD 2.6 *Sledování spotřeby proteinů v zemích Evropy*

Sledována spotřeba proteinů v 25 zemích Evropy formou spotřeby 9 druhů potravin. Cílem je odhalit, zda existuje korelace mezi znaky, tj druhy potravin? Lze odhalit nějaké interakce mezi druhy potravin a zeměmi?

○ *Data:* v datech *Proteiny* jsou uvedeny znaky: *Cervene* značí spotřebu červeného masa, *Bile* značí spotřebu bílého masa, *Vejce* značí spotřebu vajec, *Mleko* se týká spotřeby mléka, *Ryby* značí spotřebu ryb, *Obiln* značí spotřebu obilnin, *Skrob* značí spotřebu škrobu, *Orech* značí spotřebu ořechů, *Ovoce* značí spotřebu ovoce a zeleniny.

Země	*Cervene*	*Bile*	*Vejce*	*Mleko*	*Ryby*	*Obiln*	*Skrob*	*Orech*	*Ovoce*
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
East Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
West Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Hvězdy znaků pro každý objekt v EDA dat *Proteiny,* (STATISTICA).



Mnohoúhelníky znaků pro každý objekt v EDA dat *Proteiny*, (STATISTICA).

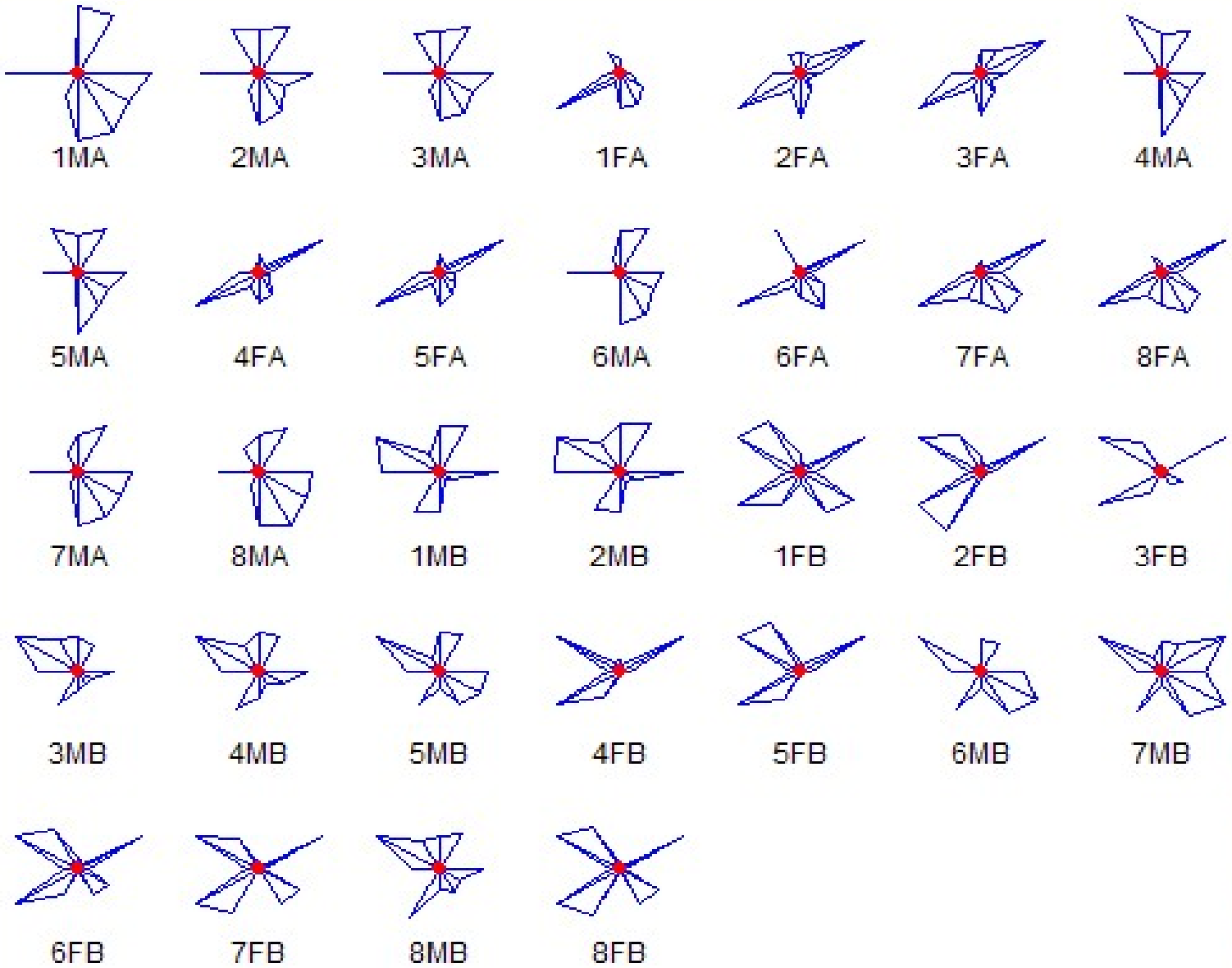
PŘÍKLAD 2.2 ***Průzkumová analýza zdrojové matice dat demografického souboru Lidé***

Vyšetřete grafickými diagnostikami průzkumové analýzy vícerozměrných dat, které ze 12 znaků demografického souboru dat *Lidé jsou nejvýhodnější k charakterizaci osob a které znaky mají největší míru rozptýlení. Matice obsahuje data pro n = 32 osob a m = 12 znaků, kde 16 osob bylo vybráno ze Skandinávie (kód A) a 16 osob ze Středomoří (kód B), 16 osob jsou muži (kód M) a 16 osob jsou ženy (kód F).*

Data: *Znaky obsahují u každé osoby výšku [cm], hmotnost [kg], délku vlasů [krátká: −1, dlouhá: +1], velikost boty [evropský standard], věk [roky], příjem [Euro], spotřeba piva [litry na rok], spotřeba vína [litry na rok], pohlaví [muž: −1, žena: +1], schopnost plavat [naměřený čas na uplavání 500 m], původ [A: −1 Skandinávie, B: +1 Středomoří], inteligenční kvocient IQ [evropský standardizovaný test IQ]. Mezi znaky jsou tři dichotomické, binární proměnné, a to pohlaví, délka vlasů a původ a ostatních 9 znaků nabývá kvantitativních hodnot.*

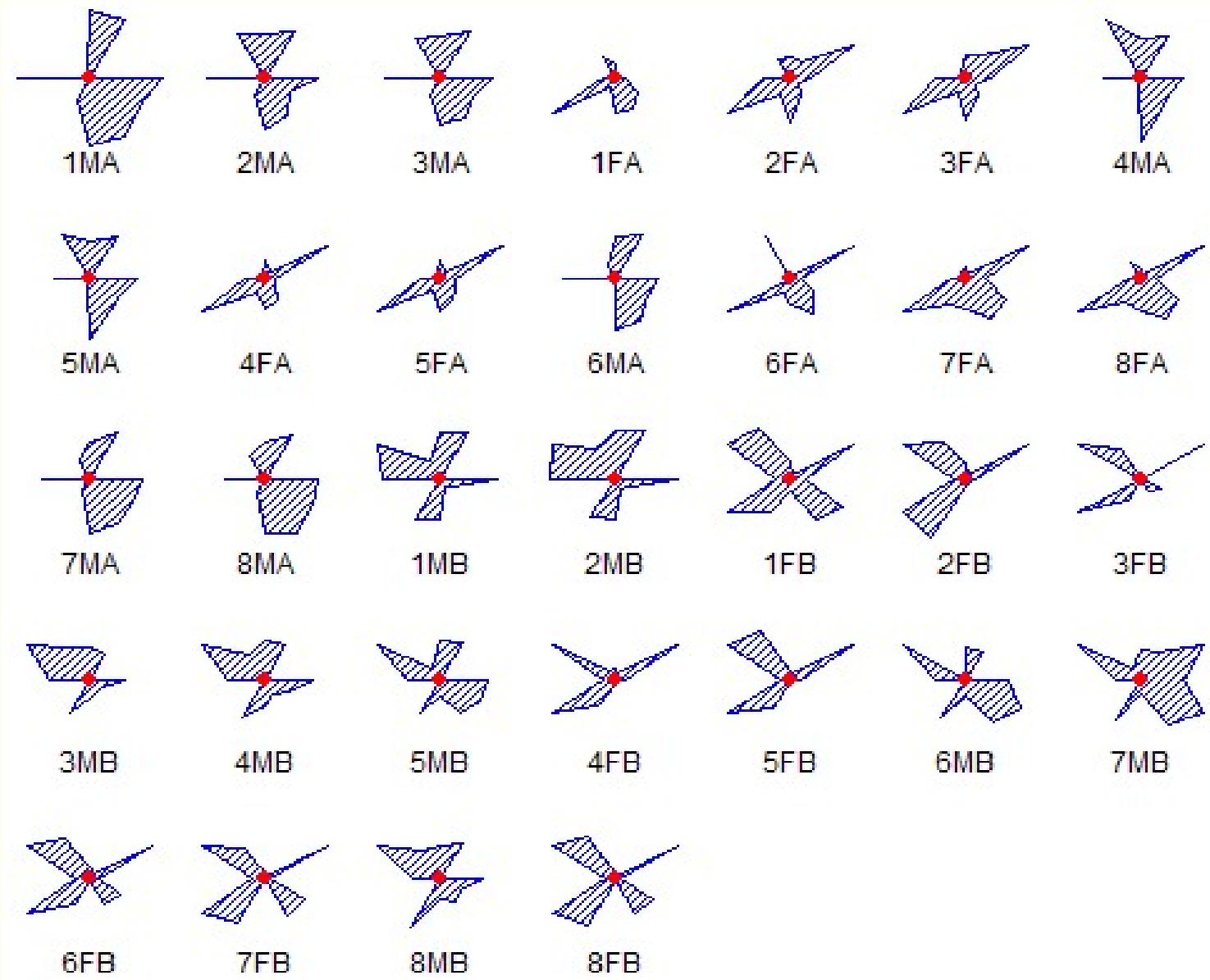
Osoba	Výška	Hmotnost	Vlasy	Boty	Věk	Příjem	Pivo	Víno	Sex
	Plavání	Původ	IQ						
MA	198	92	−1	48	48	45000	420	115	−1
	98	−1	100						
MA	184	84	−1	44	33	33000	350	102	−1
	92	−1	130						
MA	183	83	−1	44	37	34000	320	98	−1
	91	−1	127						
FA	166	47	−1	36	32	28000	270	78	1
	75	−1	112						

EDA: Symbolové grafy hvězdiček



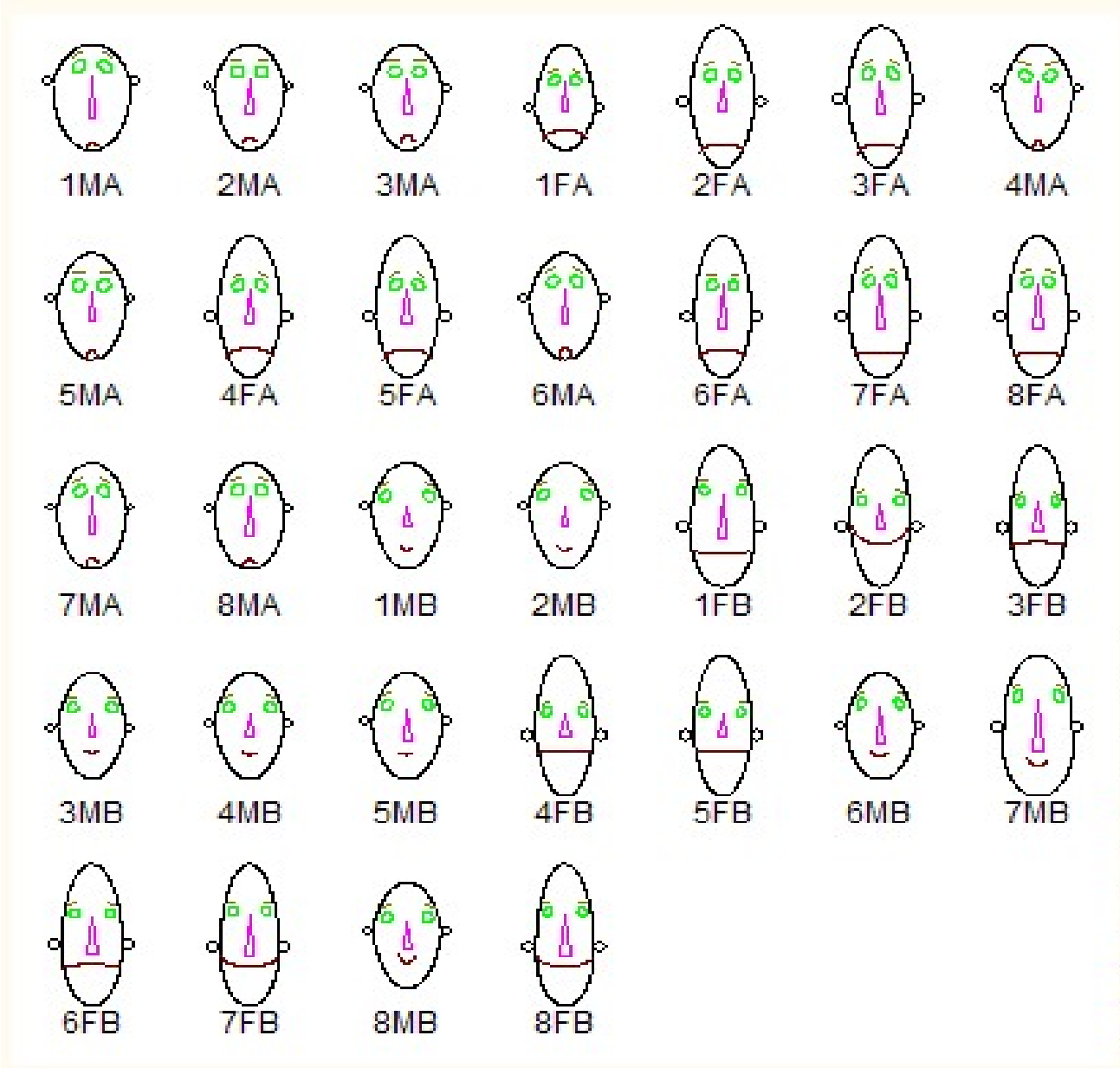
Pravotocive:
Vyska
Hmotnost
Vlasy
Boty
Vek
Prijem
Pivo
Vino
Sex
Plavani
Puvod
IQ

EDA: Symbolové grafy mnohoúhelníků



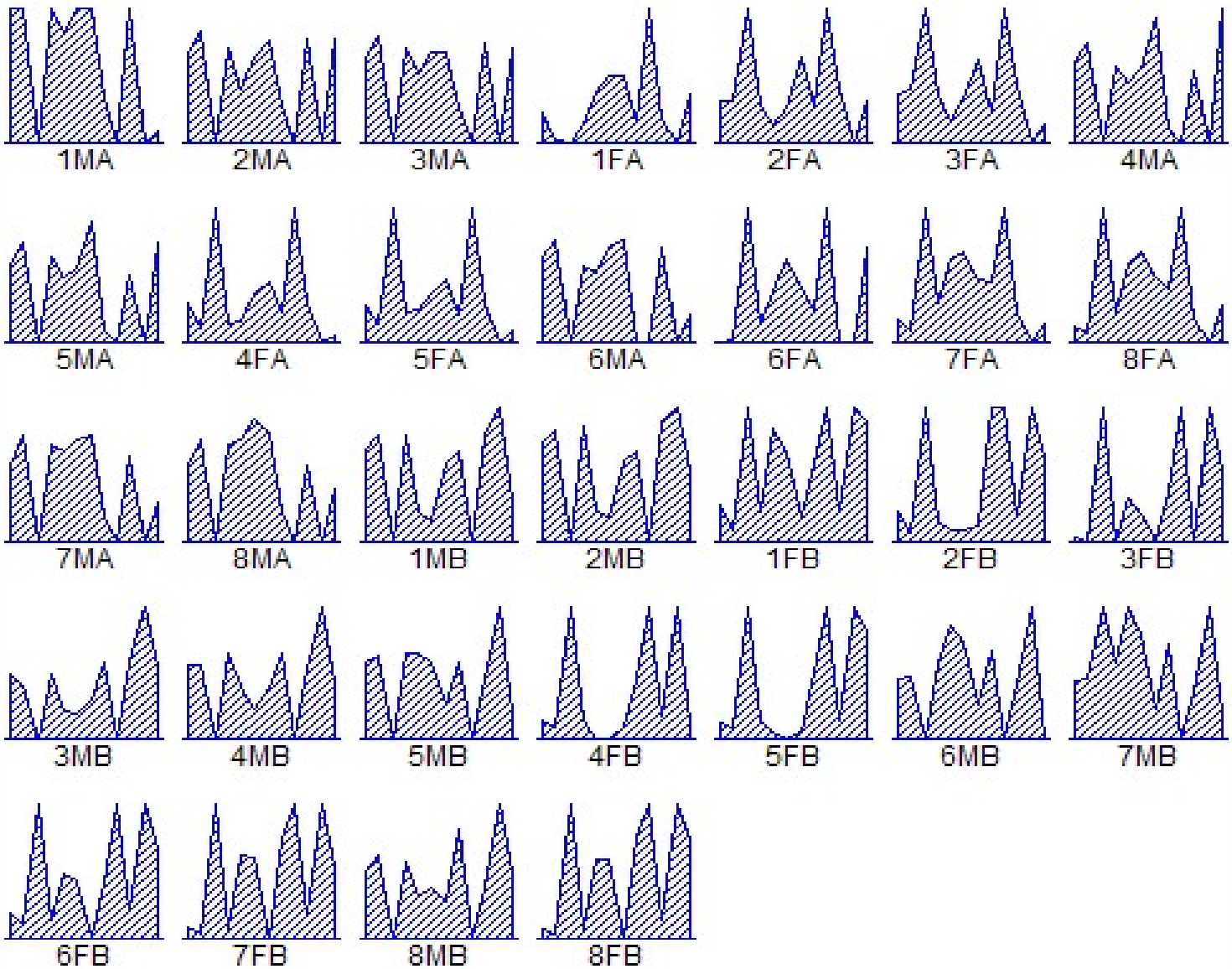
Zleva doprava:
Vyska
Hmotnost
Vlasy
Boty
Vek
Prijem
Pivo
Vino
Sex
Plavani
Puvod
IQ

EDA: Symbolové grafy obličejů



- tvár/šír = Vyska
- ucho/úrov = Hmotnost
- polovina tváre/výš = Vlasy
- horní tvár/exc = Boty
- dolní tvár/exc = Vek
- nos/dél = Prijem
- ústa/str = Pivo
- ústa/zakr = Vino
- ústa/dél = Sex
- oci/výš = Plavani
- oci/odd = Puvod
- oci/šikm = IQ

EDA: Symbolové grafy profilů



Zleva doprava:
Vyska
Hmotnost
Vlasy
Boty
Vek
Prijem
Pivo
Vino
Sex
Plavani
Puvod
IQ

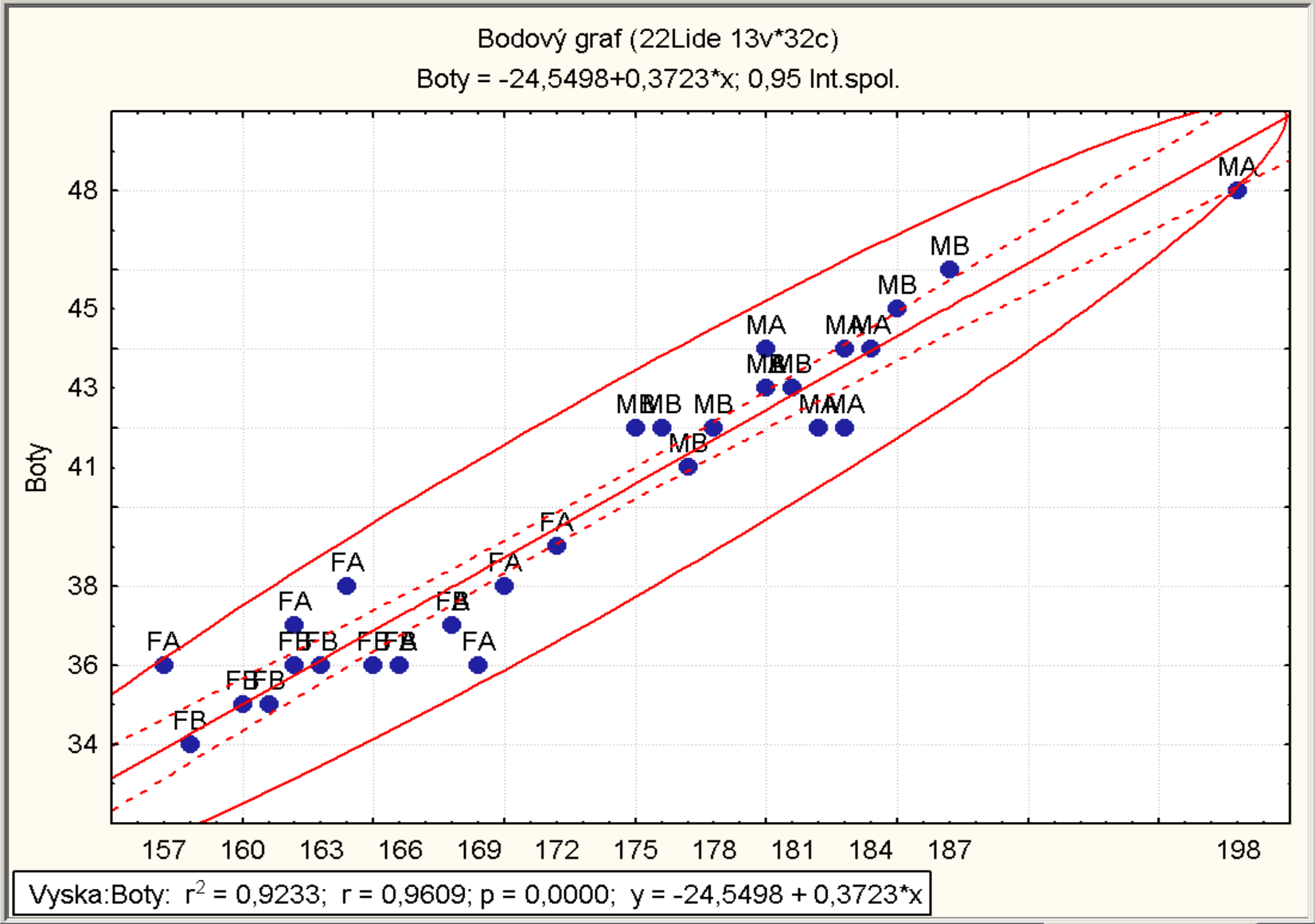
STATISTICA Cz - [PS 1* - Bodový graf (22Lide 13v*32c)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Pracovní sešit Okno Nápověda

Přidat do sešitu Přidat do protokolu

Normální graf [změ...

PS 1*
- 2D bodové grafy (22Lide)
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v



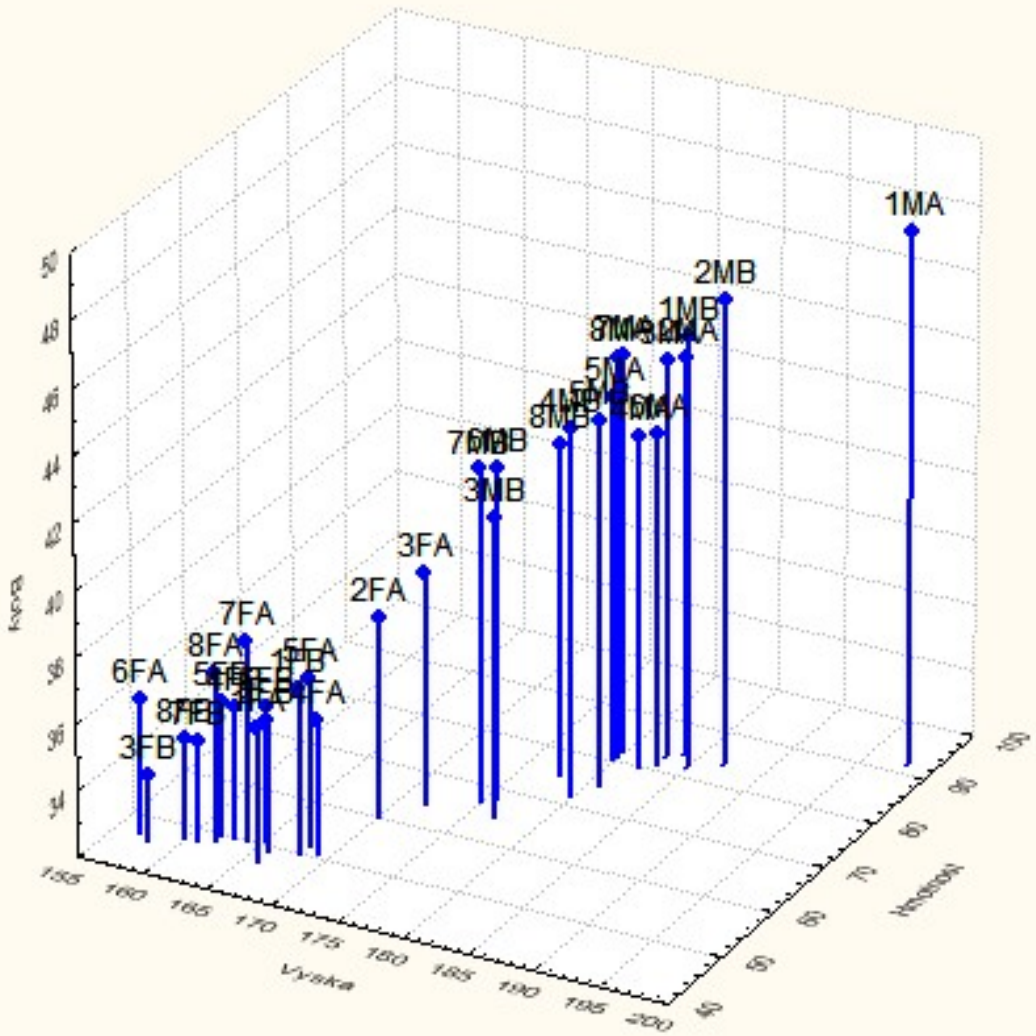
Bodový graf (22Lide 13v*32c) | Bodový graf (22Lide 13v*32c) | Bodový graf (22Lide 13v*32c) | Bodový graf (22Lide 13v*32c) | Bodový graf (22Lide 13v*32c) | Bodový graf (22Lide 13v*32c)

2D bodové grafy

Připravena

ABC 123 ZÁZN

Start | Total Commander 6.0 - P... | Zoner Photo Studio 9 - E... | Microsoft PowerPoint - [... | STATISTICA Cz - [PS ... | 18:40


1MA
2MB
1MB
5MA
3MB
3FA
2FA
7FA
8FA
6FA
3FB
Vyska
155
160
165
170
175
180
185
190
195
200

STATISTICA Cz - [PS 1* - 3D bodový graf (22Lide 13v*32c)]

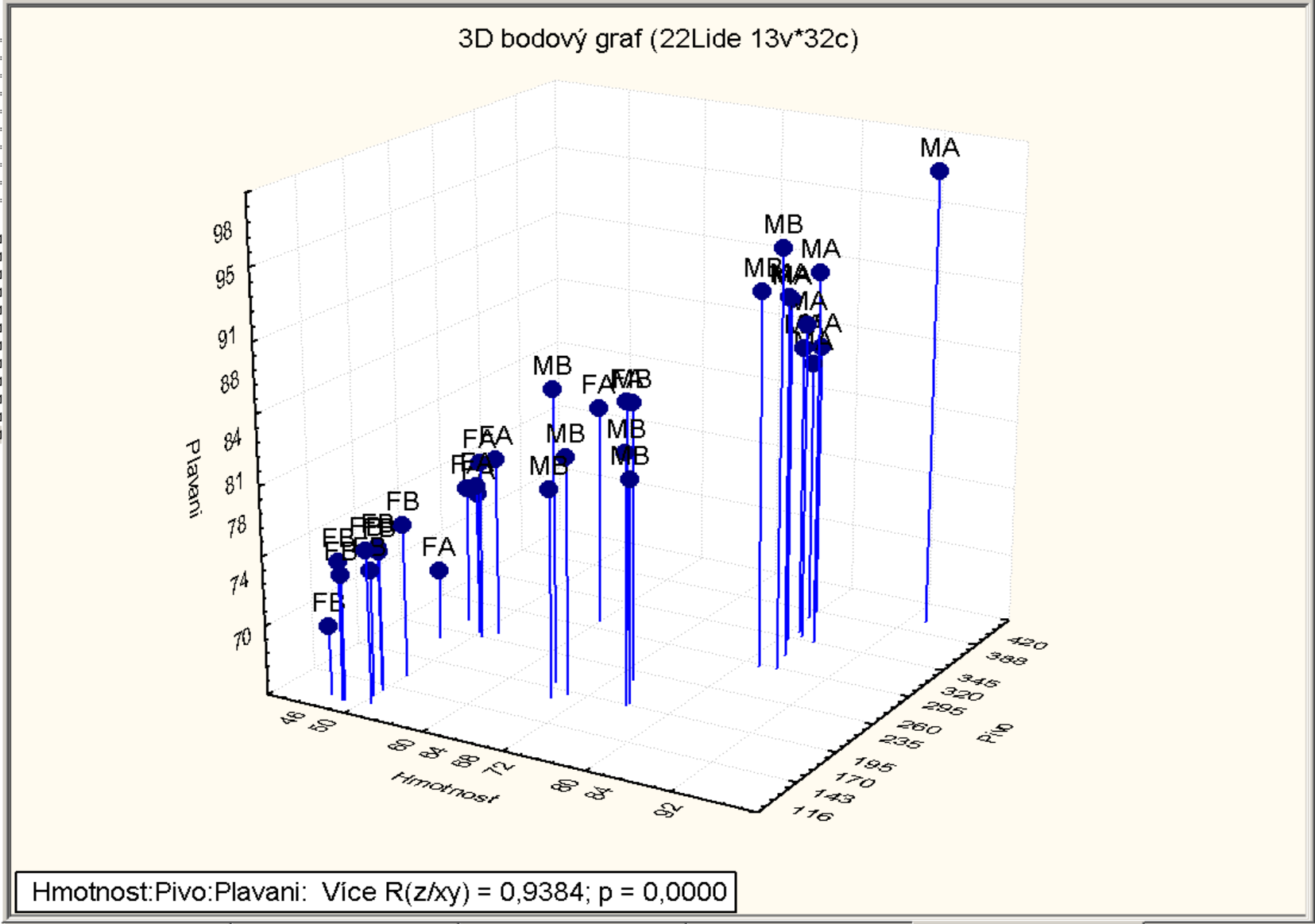
Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Pracovní sešit Okno Nápověda

Přidat do sešitu Přidat do protokolu

Normální graf [změ...

PS 1*

- 2D bodové grafy (22Lide)
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
  - Bodový graf (22Lide 13v
- 3D bodové grafy (22Lide)
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1
  - 3D bodový graf (22Lide 1



Hmotnost:Pivo:Plavani: Více R(z/xy) = 0,9384; p = 0,0000

3D bodový graf (22Lide 13v*32c) | 3D bodový graf (22Lide 13v*32c) | 3D bodový graf (22Lide 13v*32c) | 3D bodový graf (22Lide 13v*32c) | 3D bodový graf (22Lide 13v*32c)

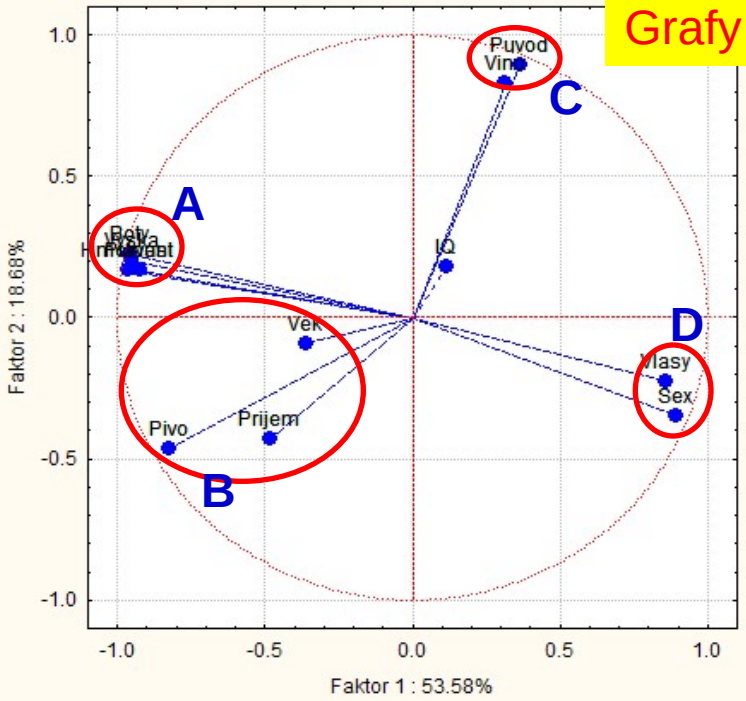
3D bodové grafy

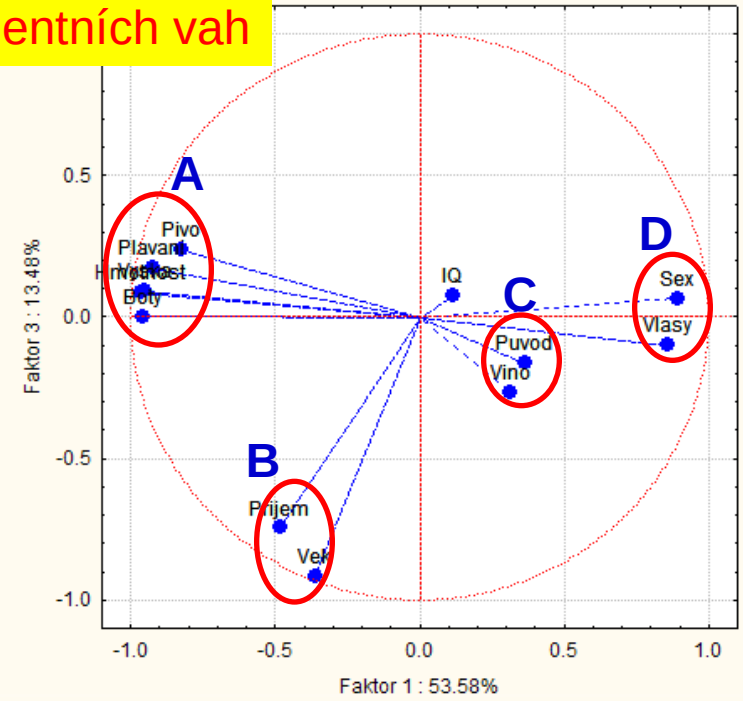
Pro nápovědu stiskněte F1

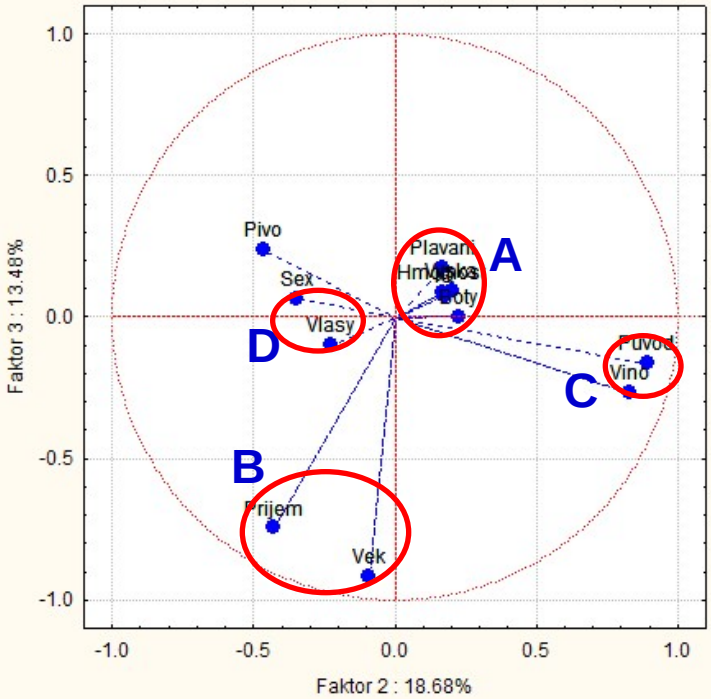
ABC 123 ZÁZN

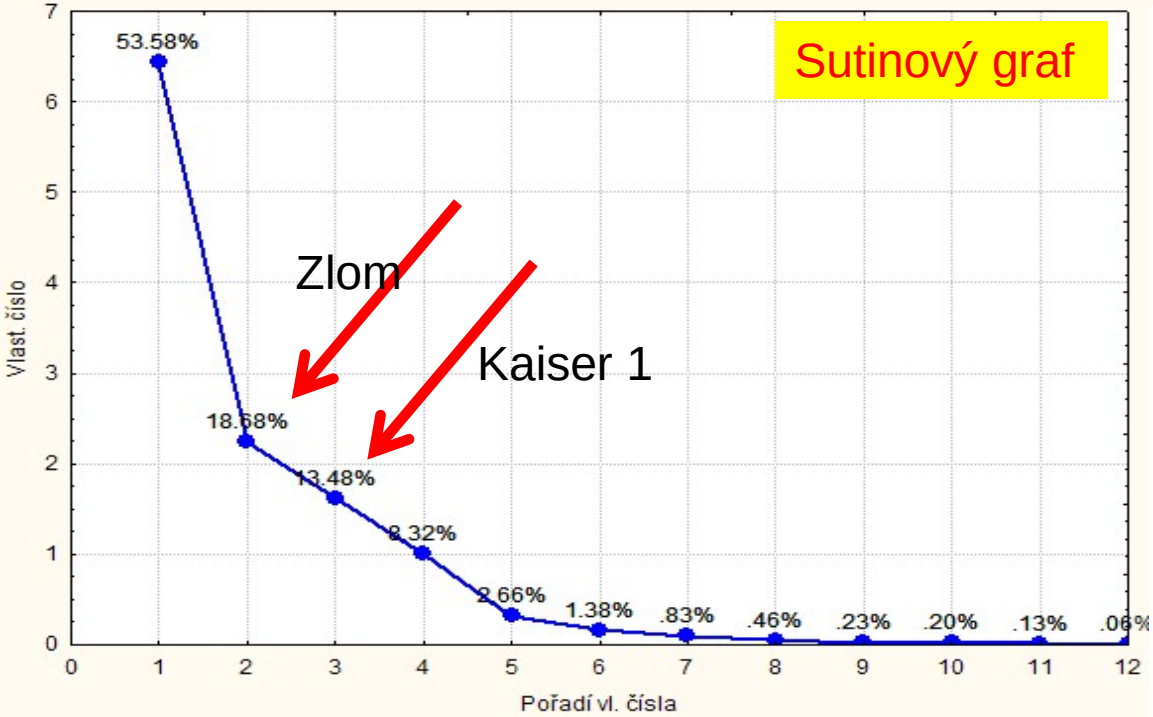
Start | Total Commander 6.0 - P... | Zoner Photo Studio 9 - E... | Microsoft PowerPoint - [... | STATISTICA Cz - [PS ... | 19:09

Grafy komponentních vah

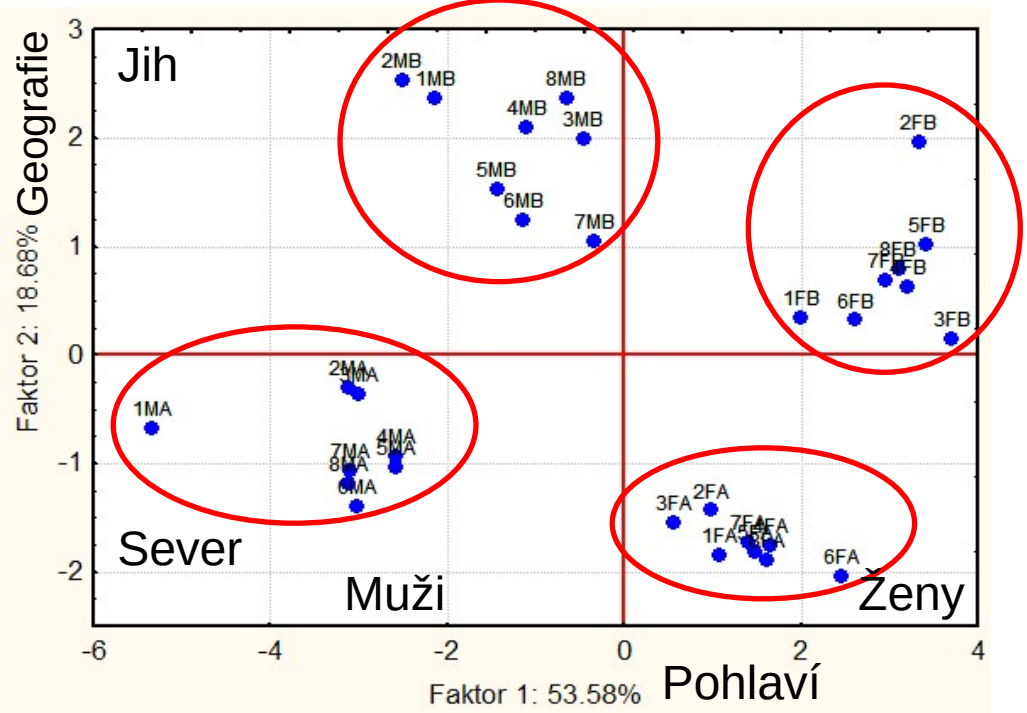


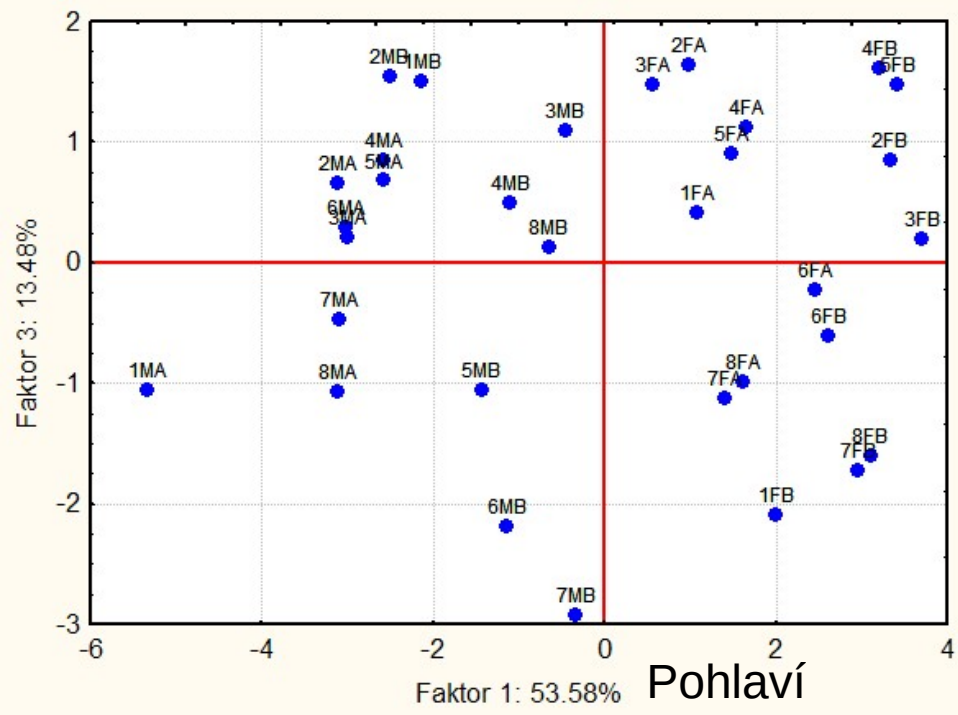


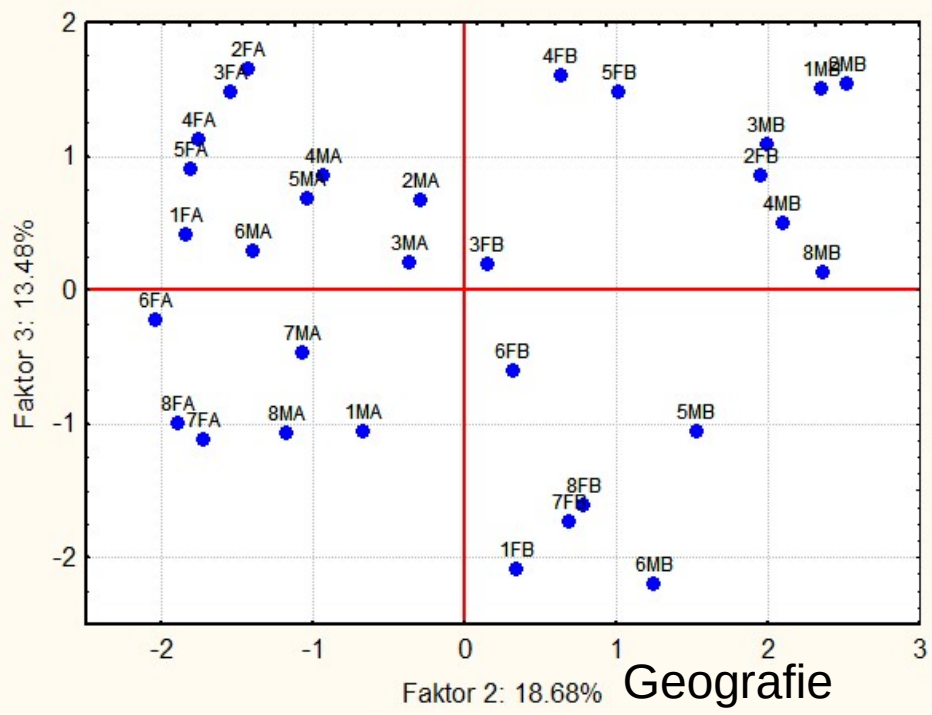




3 rozptylové diagramy komponentního skóre , tj. grafy 1-2, 1-3, 2-3







Závěr:

Určení struktury a vazeb

Zdrojová matice má rozměr $n \times m$.

1. **Hledání struktury *v proměnných* v metrické škále:**
   faktorová analýza a *analýza hlavních komponent*.

2. **Hledání struktury *v objektech* v metrické škále**: *shluková analýza*.

3. **Hledání struktury *v objektech* v metrické i v nemetrické škále:**
   vícerozměrné škálování.

4. **Hledání struktury *v objektech* v nemetrické škále:**
   korespondenční analýza.

Analýza hlavních komponent (PCA)

Zaměření metody PCA: PCA jedna z nejstarších a nejvíce používaných metod vícerozměrné analýzy.

Zavedena: Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933.

Cílem PCA: zjednodušení popisu lineárně závislých tj. korelovaných znaků, a to rozkladem matice dat do ***matice strukturní*** (= využité hlavní komponenty) a do ***matice šumové*** (= nevyužité hlavní komponenty).

Popis metody PCA:

1) Lineární transformace původních znaků na nové, nekorelované proměnné, zvané *hlavní komponenty*.

2) Základní charakteristikou každé hlavní komponenty je její míra variability čili *rozptyl.*

3) Hlavní komponenty jsou *seřazeny dle důležitosti*, tj. dle klesajícího rozptylu, od největšího k nejmenšímu.

4) *Většina informace* o variabilitě dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě.

5) ***Pravidlo:*** má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.

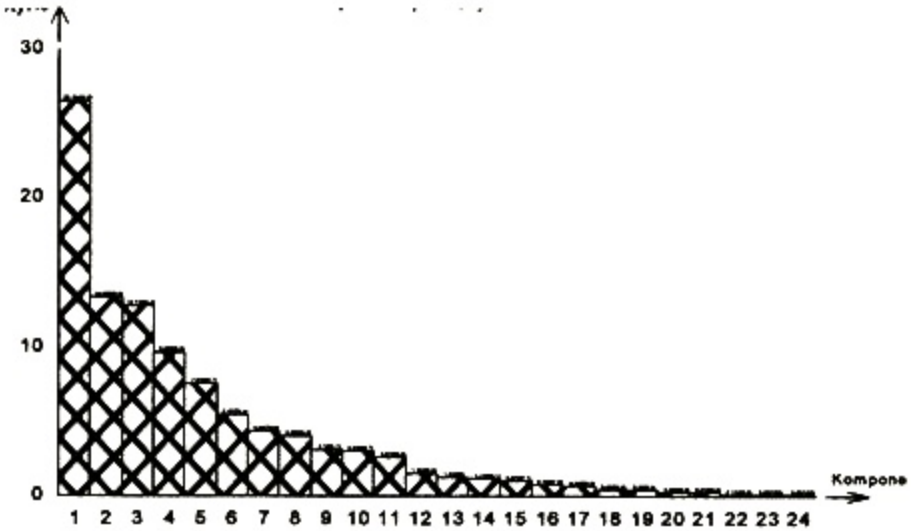
6) Využitím PCA je *snížení dimenze úlohy* čili redukce počtu znaků bez velké ztráty informace, užitím pouze prvních několika hlavních komponent.

7) *Nevyužité hlavní komponenty* obsahují malé množství informace, protože jejich rozptyl je příliš malý.

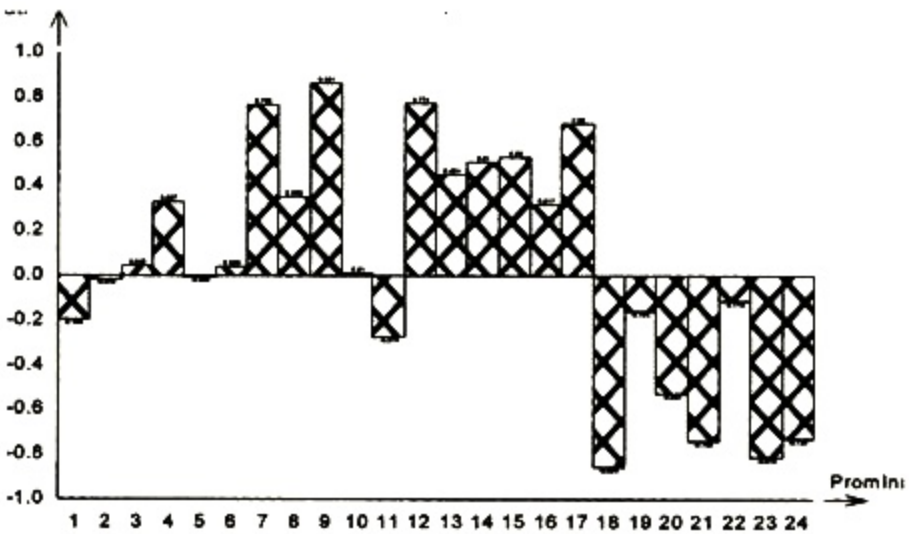
8) Hlavní komponenty jsou nekorelované.

9) První hlavní komponenta je například vhodným ukazatelem jakosti.

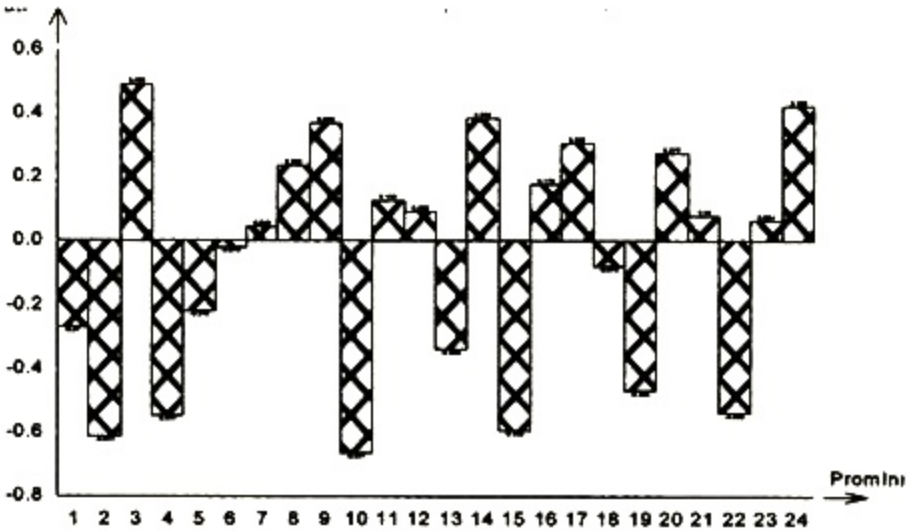
10) První dvě resp. první tři hlavní komponenty se využívají především jako techniky zobrazení vícerozměrných dat v projekci do roviny (nebo do prostoru).



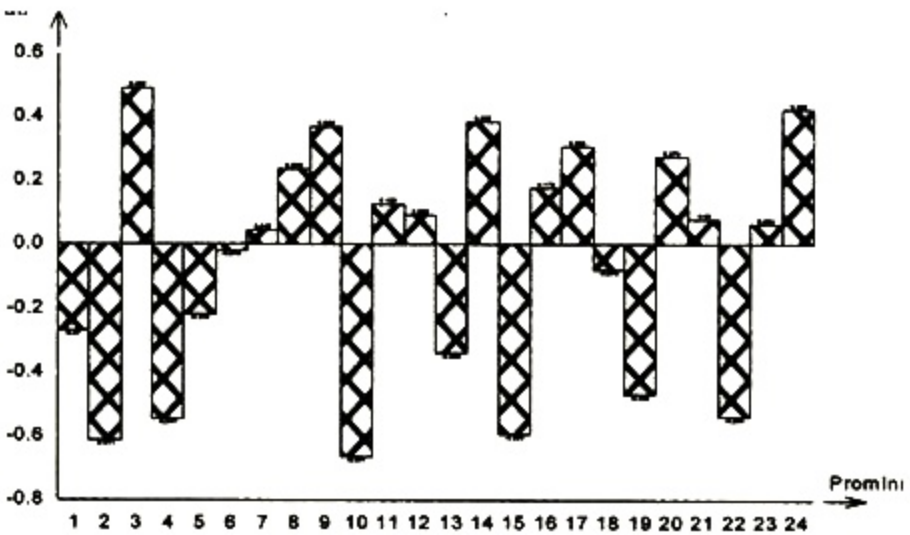
Obr. 1. Sloupcový diagram indexového grafu úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 2. Složení 1. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 3. Složení 2. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 4. Složení 3. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.

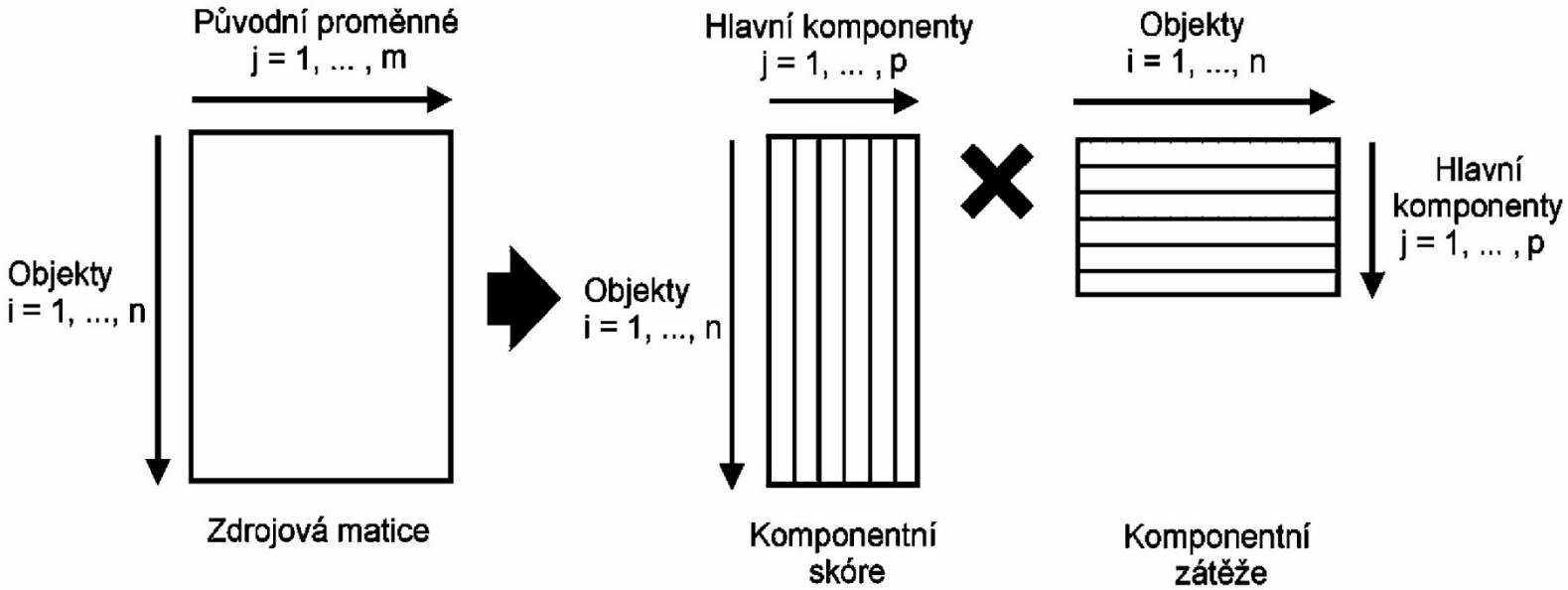
Analýza hlavních komponent (PCA)

Cílem PCA: transformace dat z původních proměnných x_i, j=1, ..., m, do menšího počtu latentních proměnných y_j.

PCA vystihují téměř *proměnlivost* původních proměnných a jsou vzájemně nekorelované.

První hlavní komponenta y_1 popisuje největší část proměnlivosti čili rozptylu původních dat,

Druhá hlavní komponenta y_2 popisuje zase největší část rozptylu neobsaženého v y_1 atd.



Obr. 4.1 Schéma maticových výpočtů v PCA.

● **Maximální počet hlavních komponent:** počet efektivních hlavních komponent se rovná *hodnosti zdrojové matice* $\boldsymbol{X}$.

● $\boldsymbol{X}$ = **Struktura + šum:** všechny hlavní komponenty jsou vzájemně ortogonální a souvisí postupně se snižující hodnotu rozptylu objektů.

Model hlavních komponent má tvar

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} = \textit{struktura dat} + \textit{šum}.$$

● **Střed modelu:** hlavní komponenty mají společný počátek, který odpovídá *průměrnému objektu* čili těžišti celého shluku objektů (centrování).

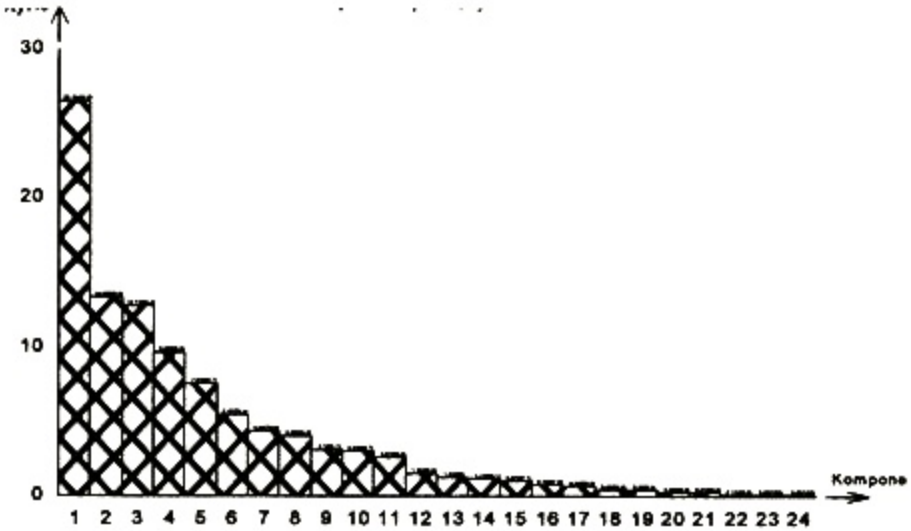
• Komponentní váhy,

zátěže - vztah mezi *X* a *PC*:

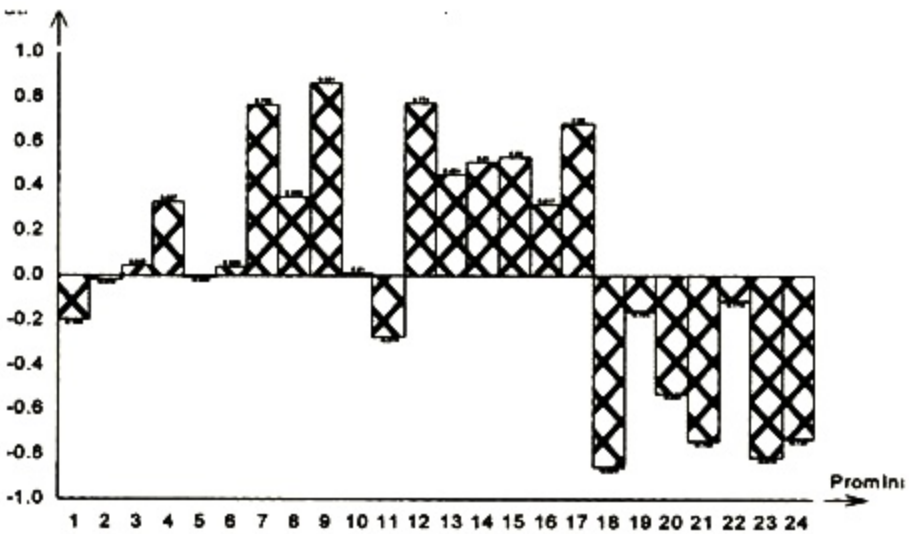
Hlavní komponenta představuje *lineární kombinaci všech m* vektorů v prostoru znaků v *m* rozměrném prostoru a jejich koeficienty se nazývají *komponentní váhy*.

Komponentní váhy informují o vztahu mezi původními *m* znaky a hlavními komponentami.

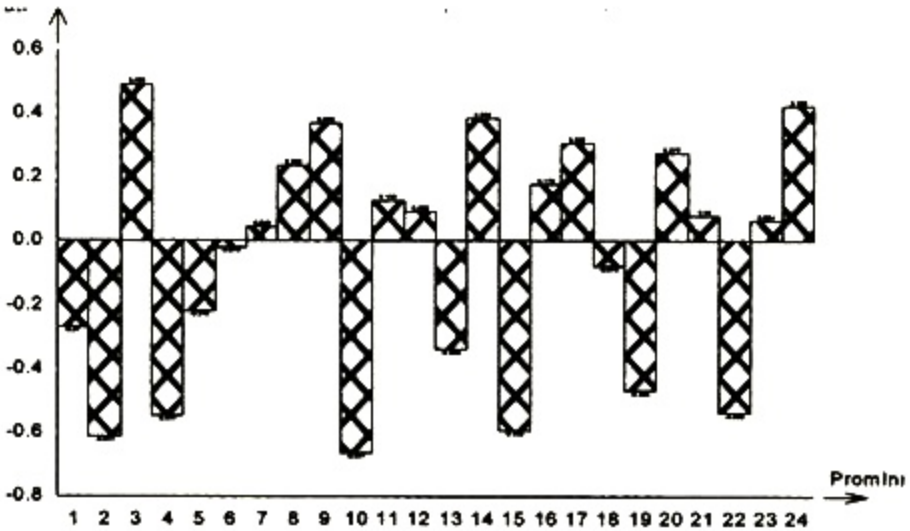
Na *grafu komponentních vah p* pro *PC*1 a *PC*2 jsou místo objektů jejich znaky a lze tak vyšetřovat závislosti a podobnosti mezi znaky.



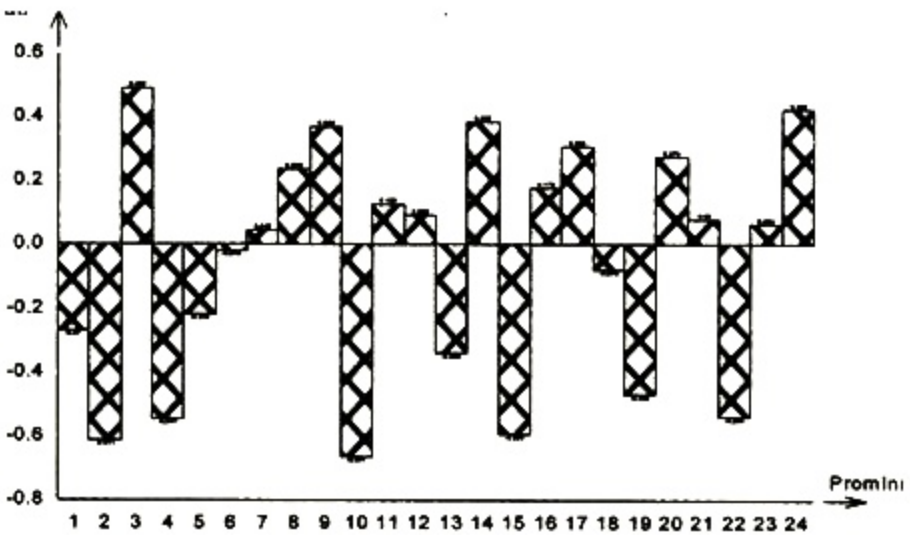
Obr. 1. Sloupcový diagram indexového grafu úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 2. Složení 1. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 3. Složení 2. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 4. Složení 3. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.

Výklad:

Indexový graf úpatí vlastních čísel

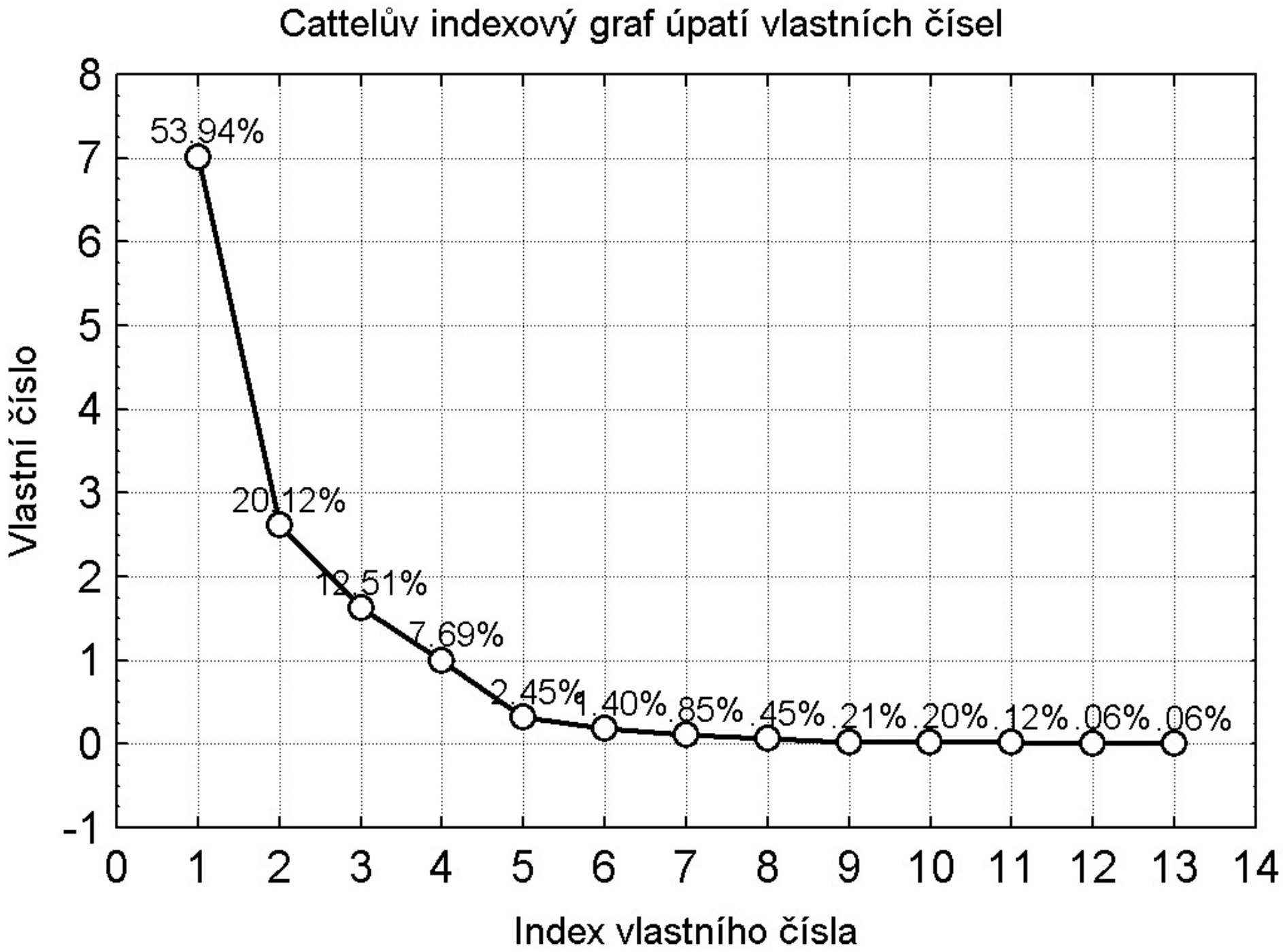
(Scree Plot)

Je to sloupcový diagram vlastních čísel proti indexu *A*.

Zobrazuje: relativní velikost jednotlivých vlastních čísel.

Využití: k určení počtu *A* "užitečných" hlavních komponent. Graf úpatí se jeví nejobjektivnějším kritériem.

Kritérium “1": hrubším kritériem PC, jejichž vlastní číslo je větší než jedna. Graf úpatí se však jeví objektivnějším.


Cattelův indexový graf úpatí vlastních čísel
Vlastní číslo
Index vlastního čísla
53.94%
20.12%
12.51%
7.69%
2.45%
1.40%
.85%
.45%
.21%
.20%
.12%
.06%
.06%
8
7
6
5
4
3
2
1
0
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Výklad:

Graf komponentních vah, zátěží

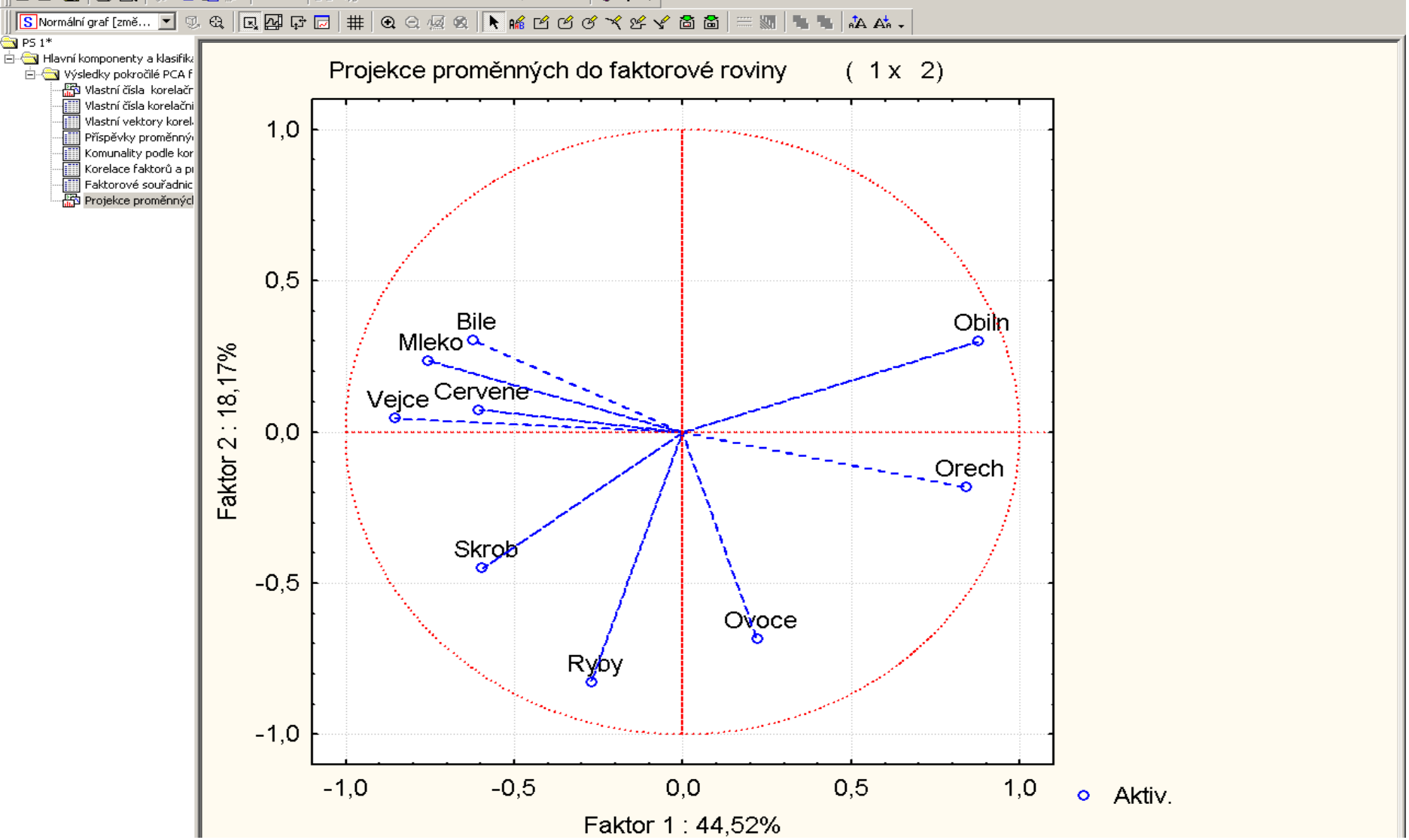
(Plot Components Weights)

Zobrazí: komponentní váhy

Porovnávají se: vzdálenosti mezi proměnnými.
Krátká znamená silnou korelaci.

Nalezneme: shluk podobných proměnných, jež spolu korelují.

Představuje: most mezi původními proměnnými a hlavními komponentami.



152. Otevřeme první *Graf komponentních vah 1. a 2. hlavní komponenty* a kliknutím na kolečka znaků nastavíme jejich grafiku (velikost, tvar, barvu).

Rozptylový diagram komponentního skóre
(Scatterplot)

1. **Umístění objektů:** daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.

2. **Podobnost objektů:** objekty blízko sebe si jsou podobné, daleko od sebe jsou si nepodobné.

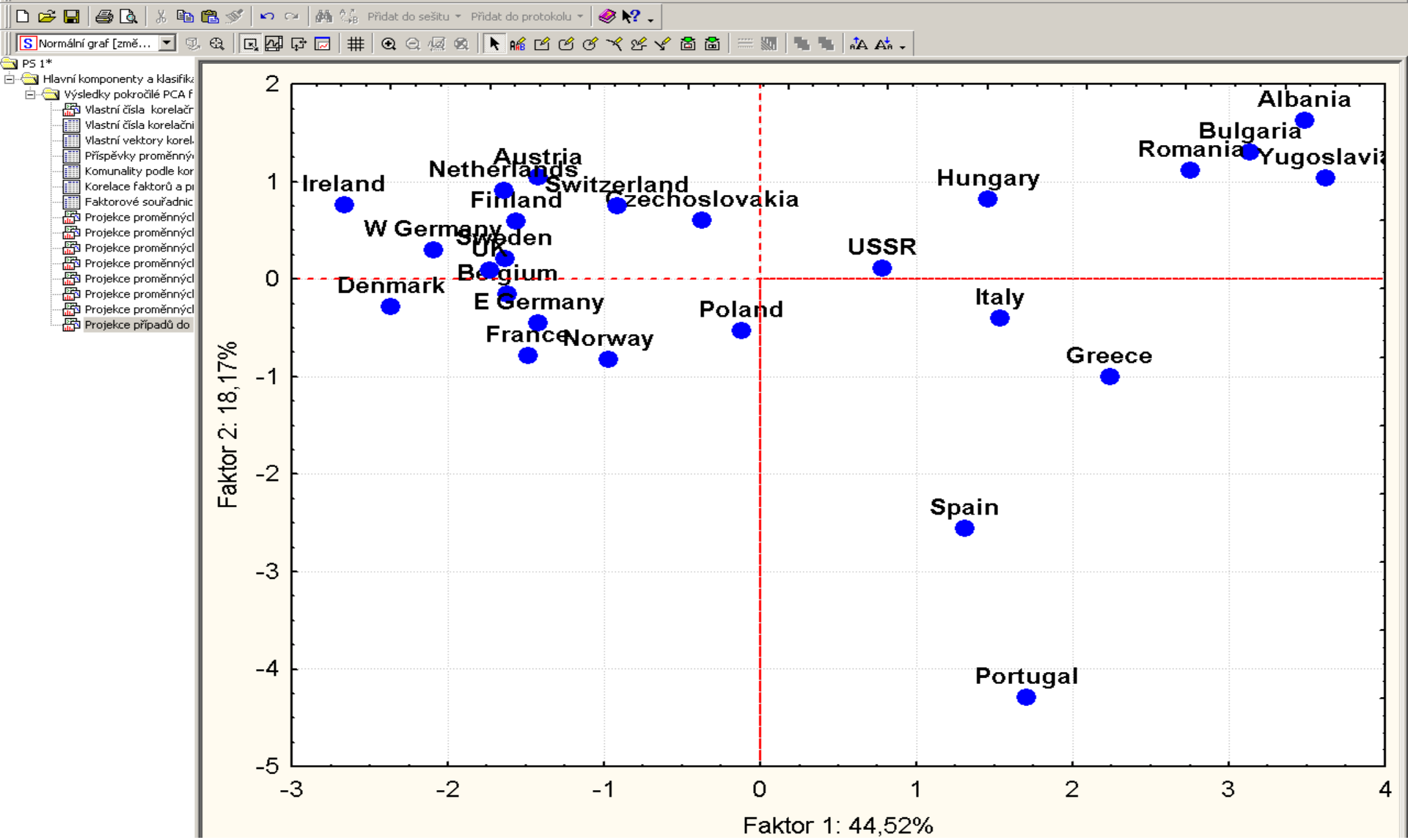
3. **Objekty v shluku:** umístěné zřetelně v jednom shluku jsou si podobné a nepodobné objektům v ostatních shlucích. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.

4. **Osamělé objekty:** izolované objekty mohou být odlehlé.

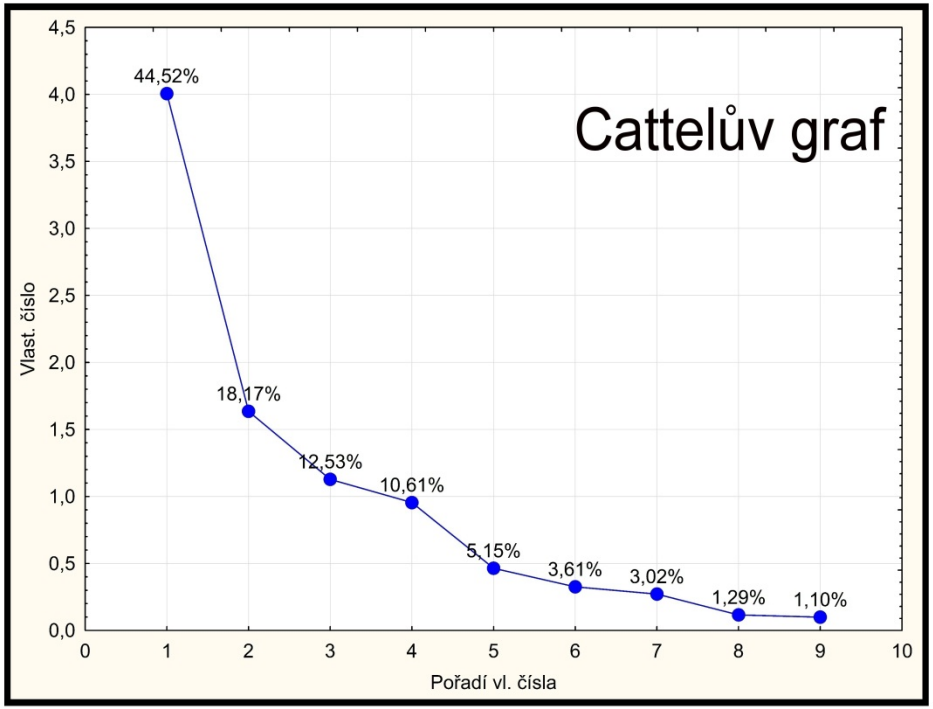
5. **Odlehlé objekty:** ideálně bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je špatný model.

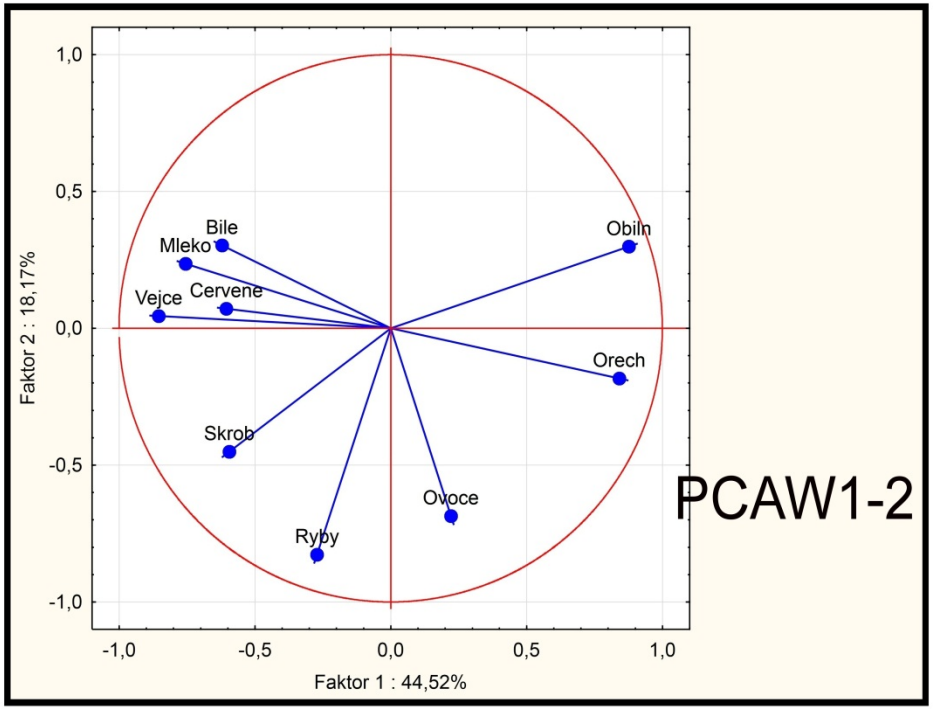
6. **Pojmenování objektů:** výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a vystihneme tak jejich fyzikální či biologický vztah.

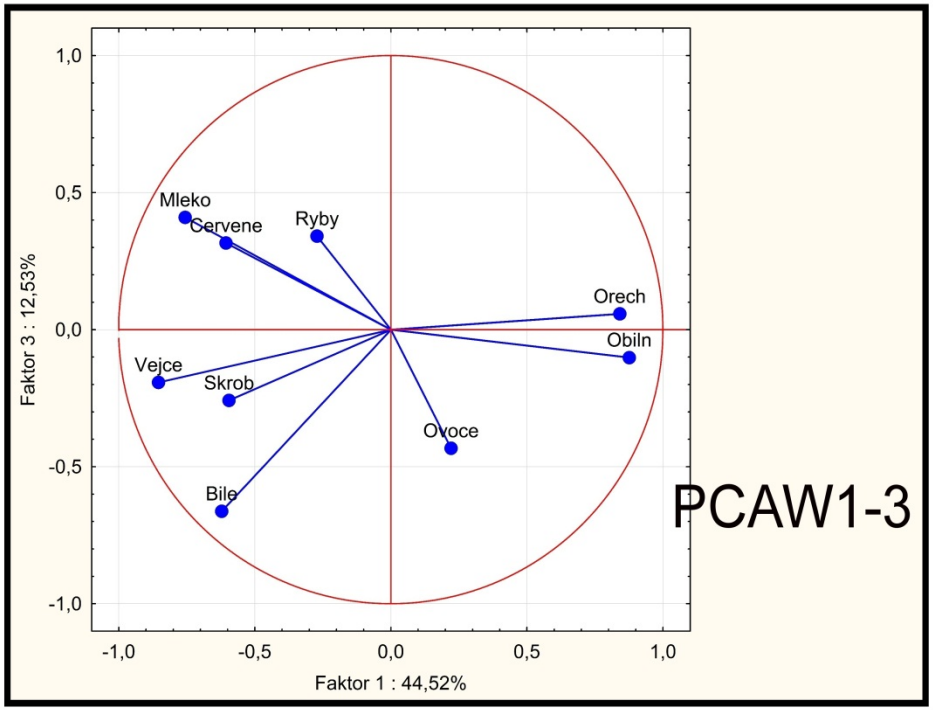
7. **Vysvětlení místa objektu:** umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami původních proměnných ve dvojném grafu.

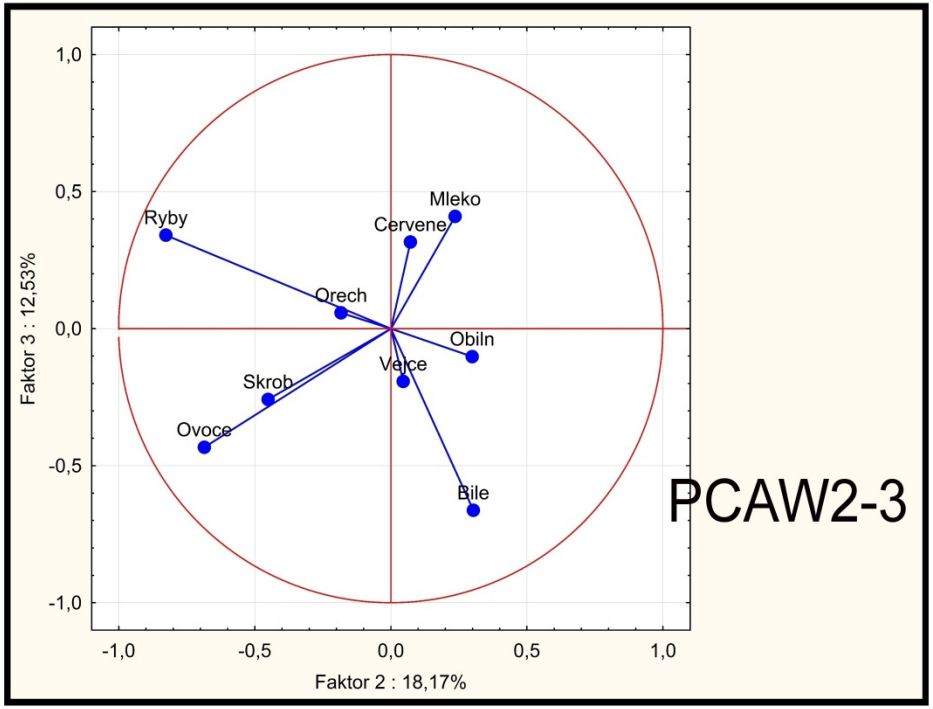


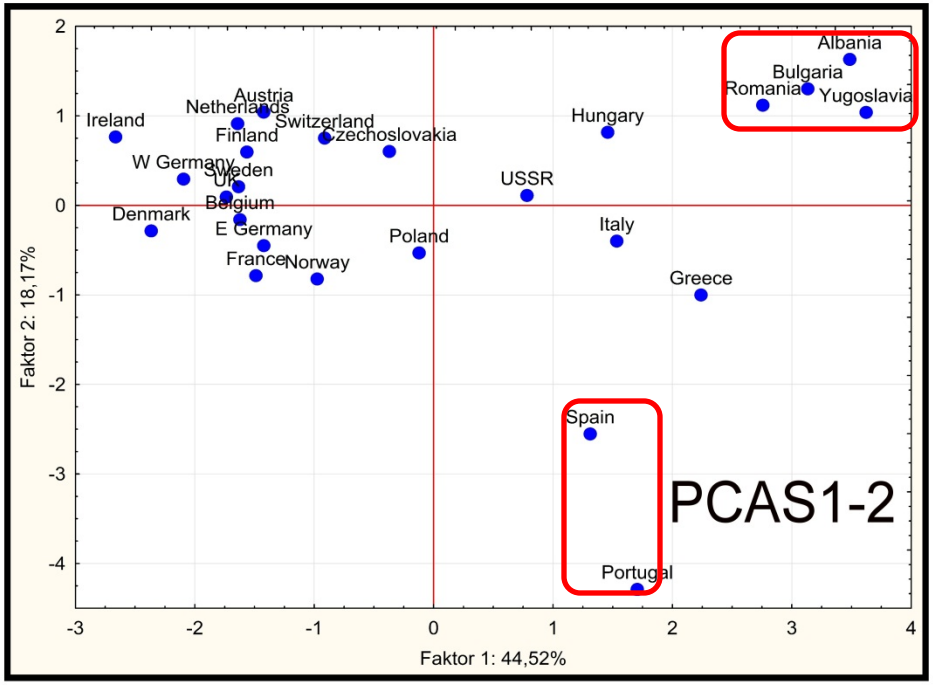
159. Kliknutím na čísla na osách se otevře okno *Rozložení osy*, ve kterém nastavíme měřítko *Minimum* a *Maximum*, dále pak *Upravit krok* obou os *x* a *y* a obdržíme tento výsledek.

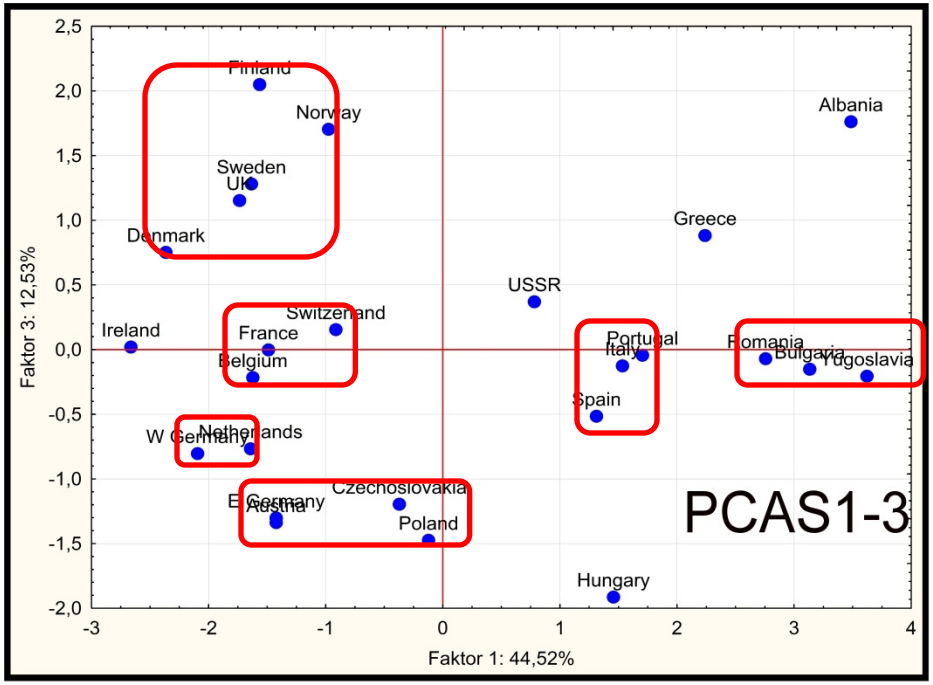

Cattelův graf
44,52%
18,17%
12,53%
10,61%
5,15%
3,61%
3,02%
1,29%
1,10%
Vlast. číslo
Pořadí vl. čísla

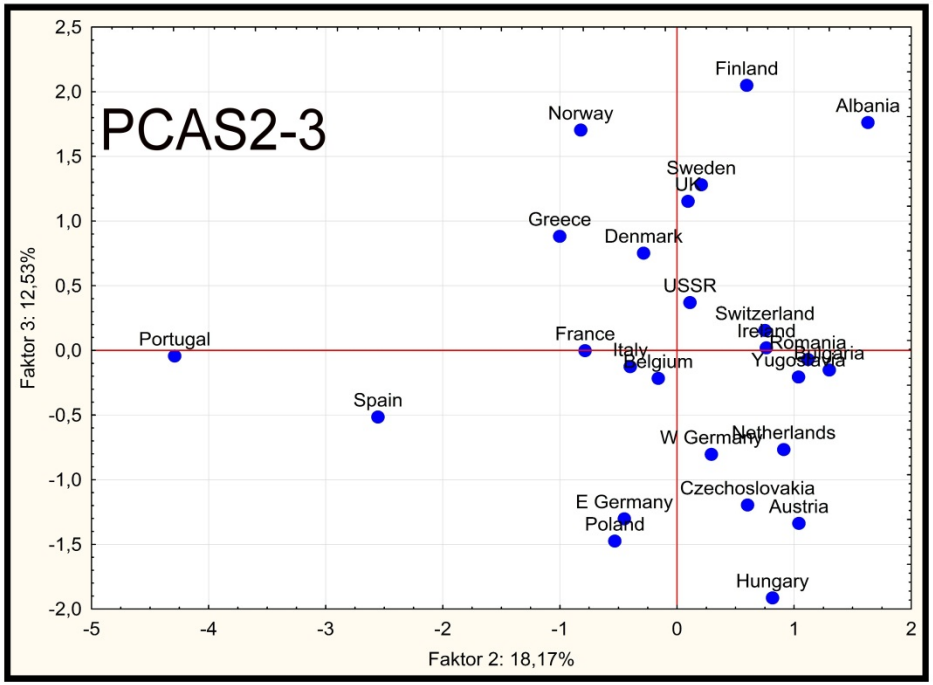

PCAW1-2
Bile
Mleko
Cervene
Vejce
Obiln
Orech
Skrob
Ovoce
Ryby
Faktor 2 : 18,17%
Faktor 1 : 44,52%

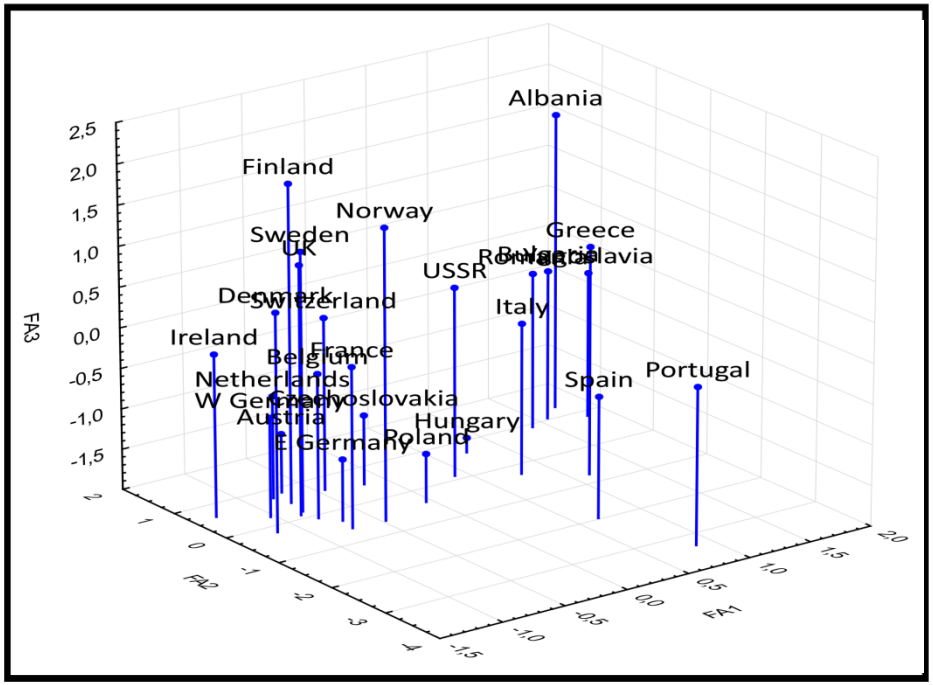

PCAW1-3
Mleko
Cervene
Ryby
Orech
Obiln
Vejce
Skrob
Ovoce
Bile
Faktor 3 : 12,53%
Faktor 1 : 44,52%


PCAW2-3
Ryby
Mleko
Cervene
Orech
Obiln
Vejce
Skrob
Ovoce
Bile
Faktor 3 : 12,53%
Faktor 2 : 18,17%


PCAS1-2
Faktor 1: 44,52%
Faktor 2: 18,17%
Albania
Bulgaria
Romania
Yugoslavia
Austria
Netherlands
Switzerland
Ireland
Finland
Czechoslovakia
Hungary
W Germany
Sweden
UK
USSR
Belgium
Denmark
E Germany
Poland
Italy
France
Norway
Greece
Spain
Portugal


PCAS1-3
Faktor 1: 44,52%
Faktor 3: 12,53%
Finland
Norway
Albania
Sweden
UK
Denmark
Greece
USSR
Switzerland
Ireland
France
Portugal
Italy
Romania
Bulgaria
Yugoslavia
Belgium
Spain
W Germany
Netherlands
Czechoslovakia
E Germany
Austria
Poland
Hungary


PCAS2-3
Faktor 2: 18,17%
Faktor 3: 12,53%
Finland
Albania
Norway
Sweden
UK
Greece
Denmark
USSR
Switzerland
Ireland
Portugal
France
Romania
Italy
Bulgaria
Belgium
Yugoslavia
Spain
W Germany
Netherlands
Czechoslovakia
E Germany
Austria
Poland
Hungary


FA3
FA2
FA1
Albania
Finland
Norway
Sweden
UK
Greece
USSR
Italy
Ireland
Belgium
France
Netherlands
Spain
Portugal
Austria
Hungary
E Germany
Poland

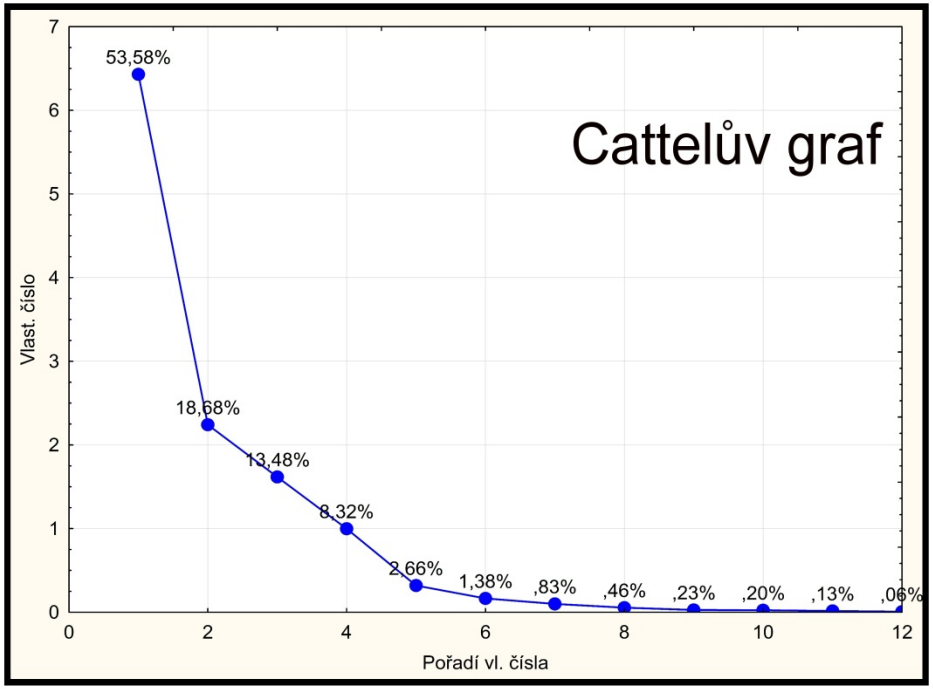
PŘÍKLAD 4.1

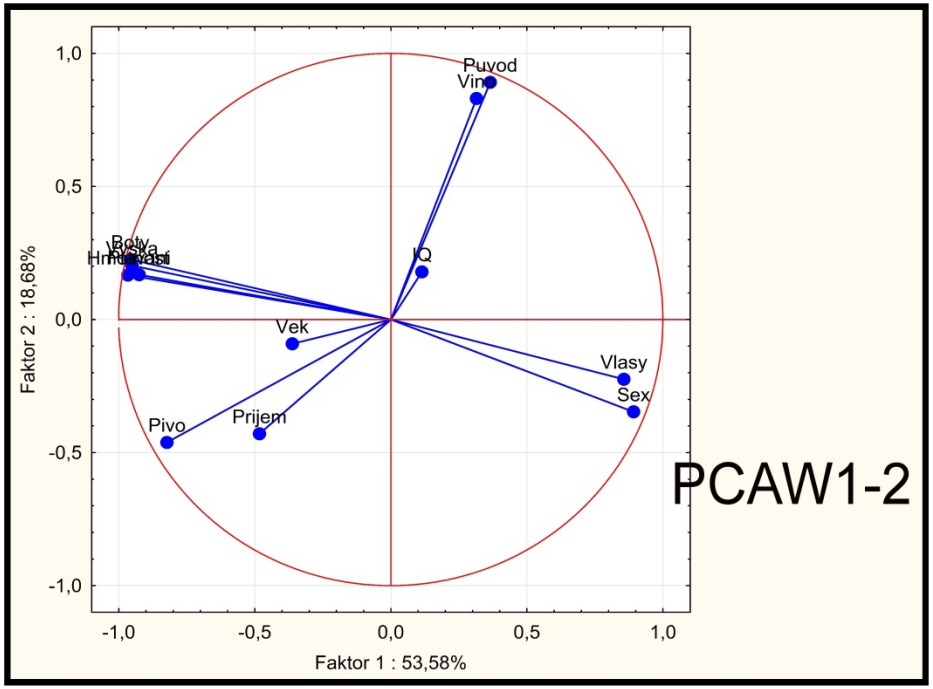
Interpretace grafů PCA demografického souboru Lidé

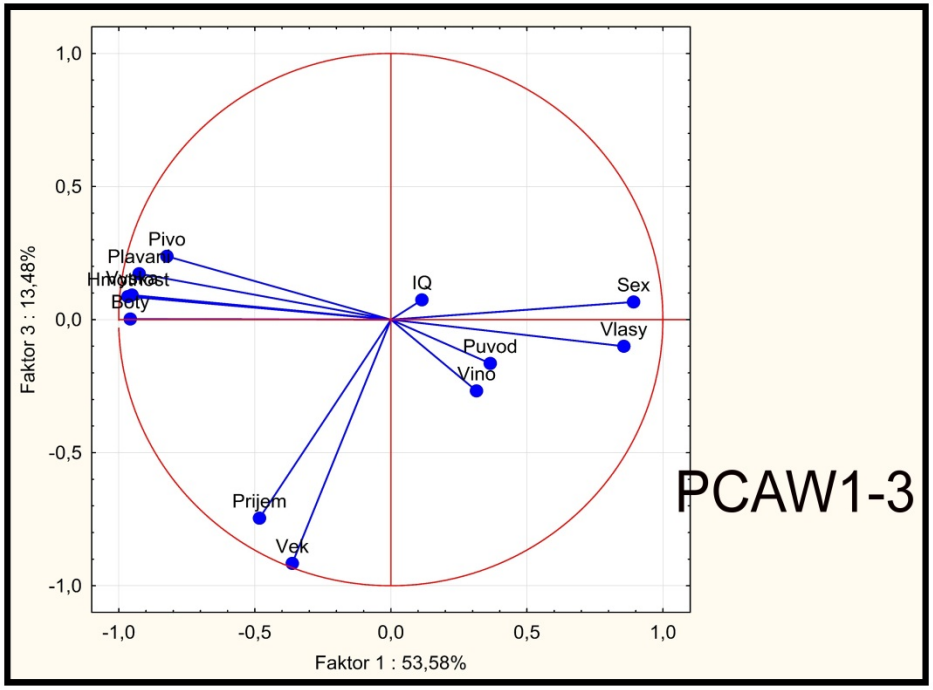
Cílem je interpretace grafu komponentního skóre a komponentních vah pro soubor *Lidé*

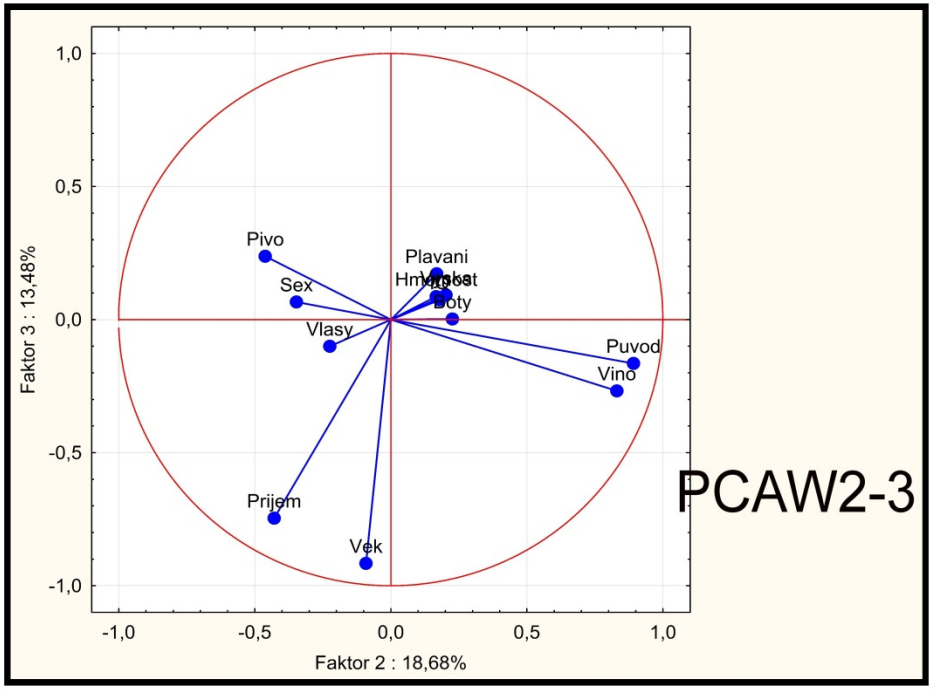
- *Data:* použita data *Lidé*
- *Řešení:* Na základě maticového grafu korelace byla získána počáteční informace o korelaci znaků. Grafy hlavních komponent zdrojové matice dat *Lidé* se týkají 32 osob (objektů) a 12 jejich vlastností (znaků).

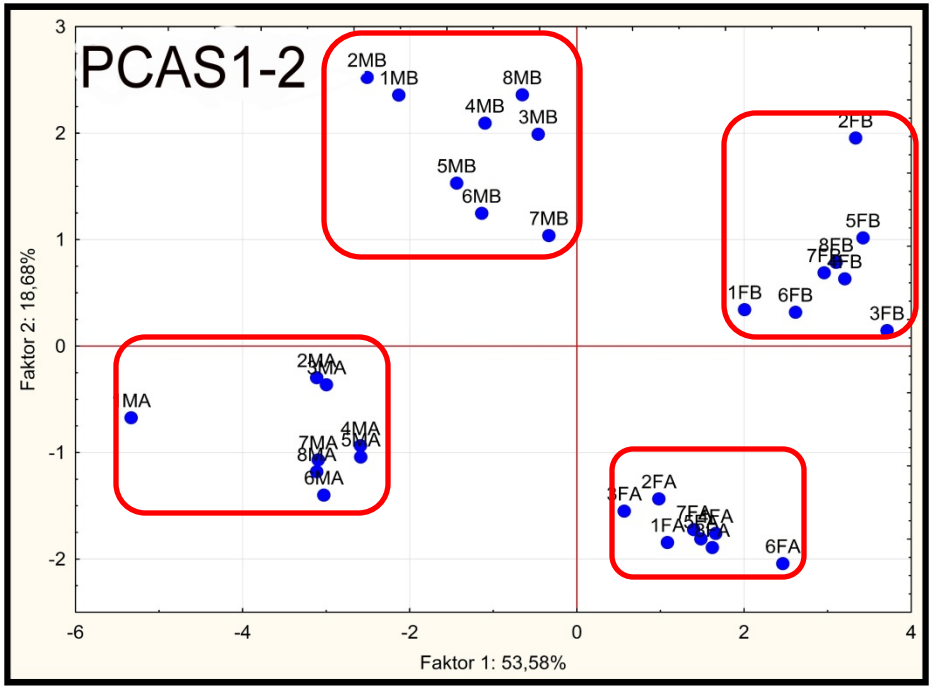
1. Počet potřebných hlavních komponent: indexový graf úpatí objasněného rozptylu ukazuje, že první dvě hlavní komponenty objasňují okolo 74% z celkového rozptylu.

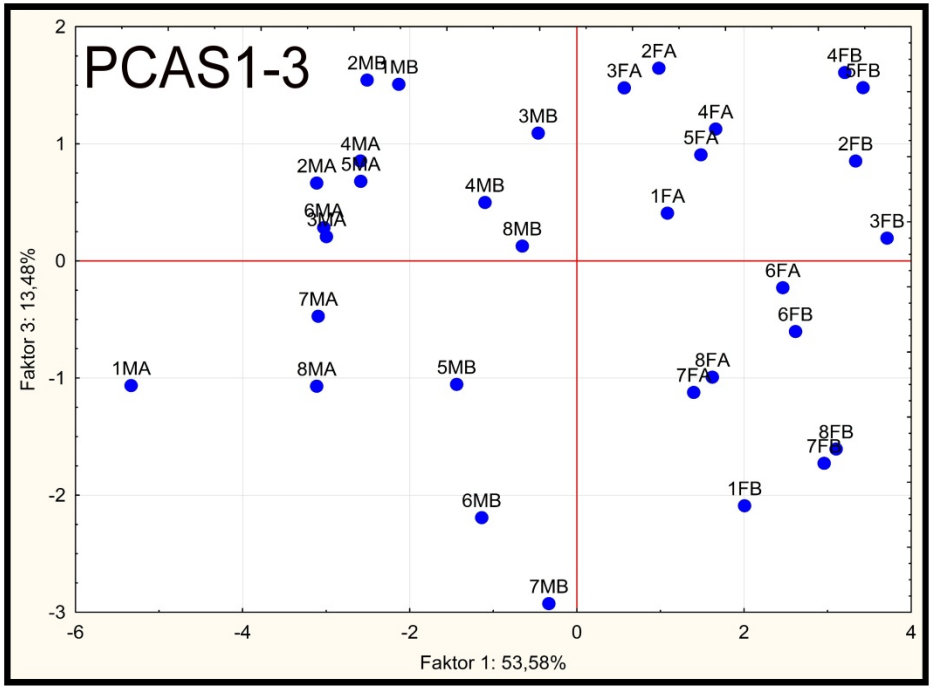

Cattelův graf
Vlast. číslo
Pořadí vl. čísla
53,58%
18,68%
13,48%
8,32%
2,66%
1,38%
,83%
,46%
,23%
,20%
,13%
,06%

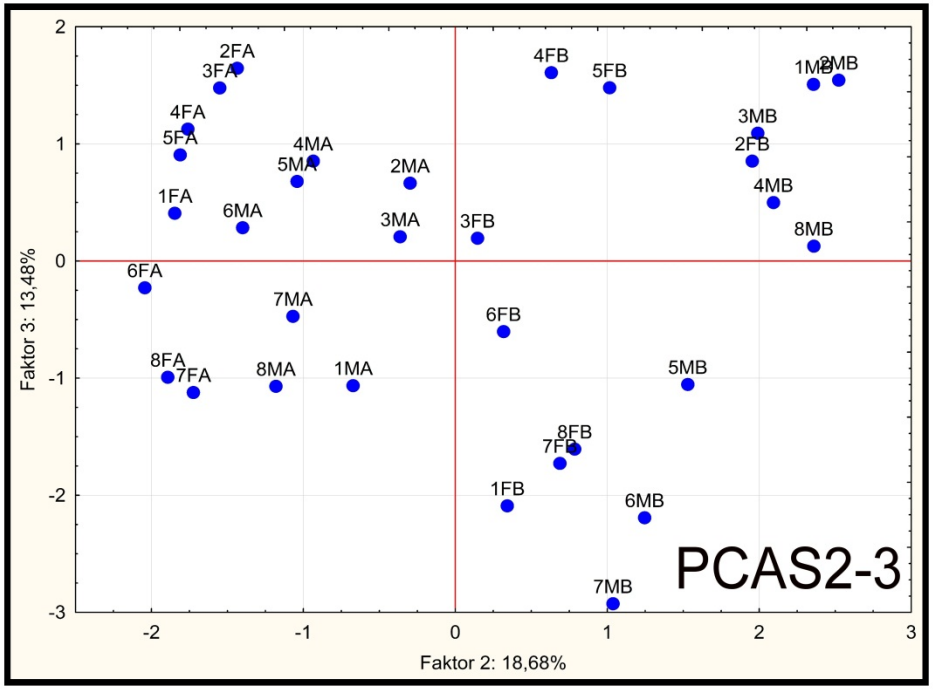

PCAW1-2
Faktor 1 : 53,58%
Faktor 2 : 18,68%
Puvod
Vino
IQ
Vek
Vlasy
Sex
Pivo
Prijem


PCAW1-3
Faktor 1 : 53,58%
Faktor 3 : 13,48%
Pivo
Plavani
Boty
IQ
Sex
Vlasy
Puvod
Vino
Prijem
Vek


PCAW2-3
Faktor 2 : 18,68%
Faktor 3 : 13,48%
Pivo
Plavani
Sex
Boty
Vlasy
Puvod
Vino
Prijem
Vek


PCAS1-2
Faktor 1: 53,58%
Faktor 2: 18,68%


PCAS1-3
Faktor 1: 53,58%
Faktor 3: 13,48%

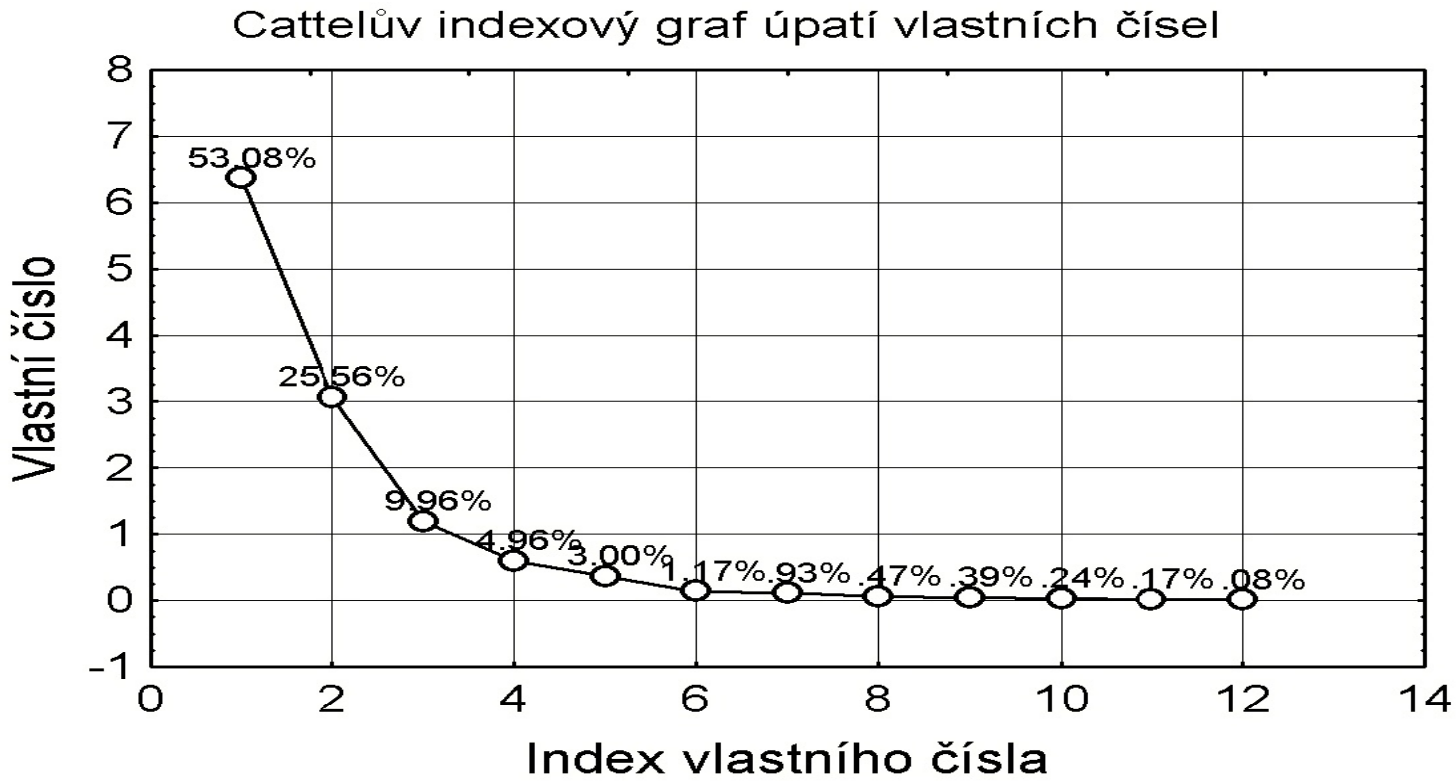

PCAS2-3
Faktor 2: 18,68%
Faktor 3: 13,48%

PŘÍKLAD 4.2 ***Posouzení hrachu diagramem komponentního skóre***

Je třeba roztřídit druhy vyšetřovaného hrachu dle smyslového posouzení hrachu člověkem, které znaky subjektivního posouzení se nejlépe hodí k popisu. Které znaky se nejlépe podílejí na popisu proměnlivosti hrachu?

○ *Řešení:*

1. Počet potřebných hlavních komponent: První hlavní komponenta popisuje 53% celkového rozptylu, druhá hlavní komponenta 25.6% a třetí hlavní komponenta 9.9%.



Obr. 4.7a Cattelův indexový graf úpatí vlastních čísel Scree Plot zdrojové matice dat *Hrách* (STATISTICA).

2. Graf komponentních vah: vysvětlení grafu

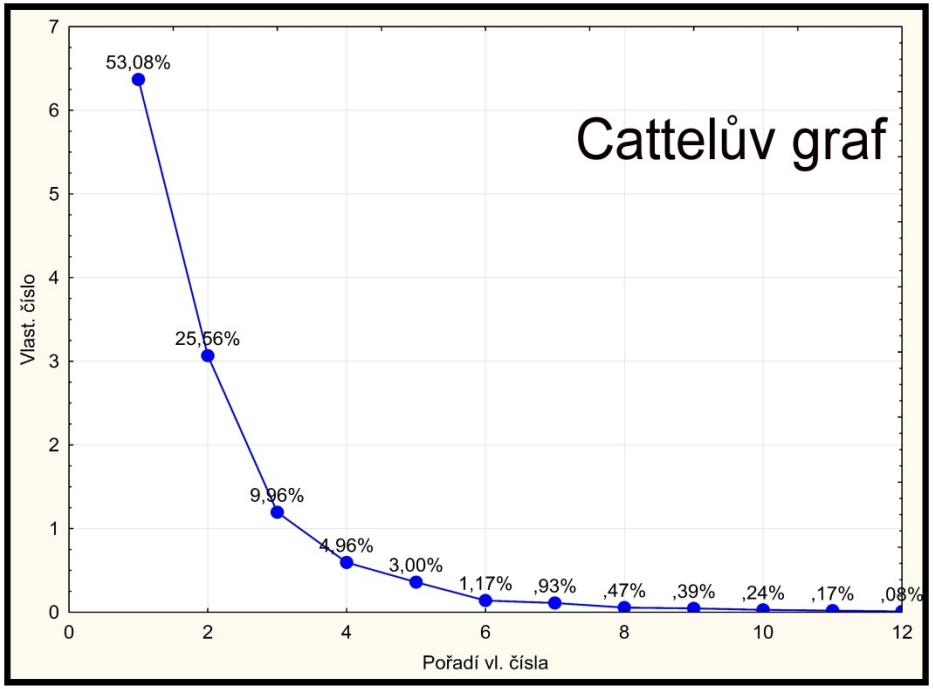
1) Vůně hrachu (znak *Aro*), sladkost (*Slad*) a medovost (*Med*) pozitivně korelují,

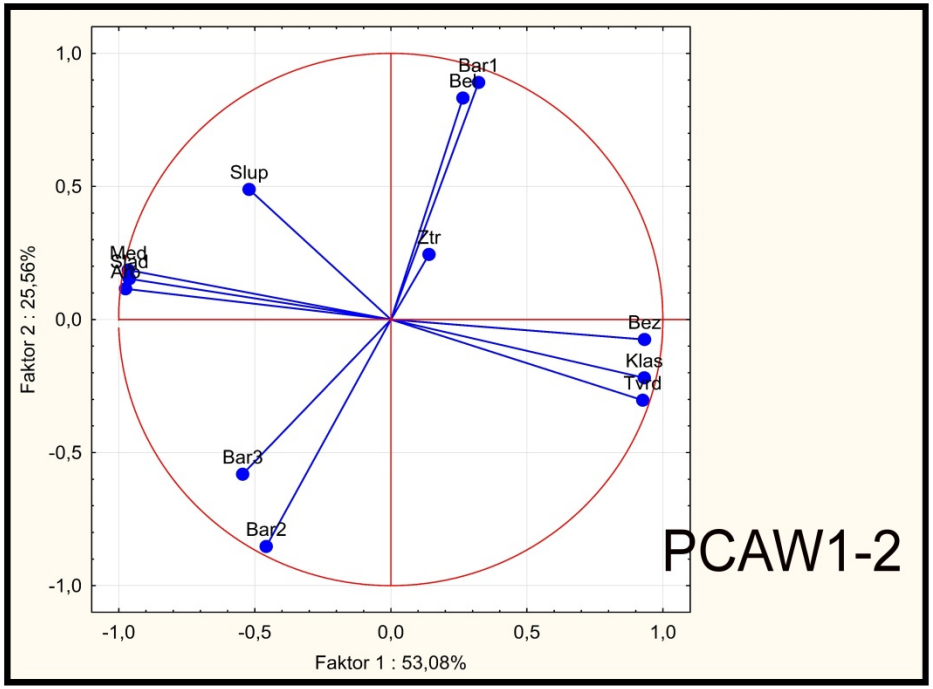
2) Tvrdost hrachu (*Tvrd)*, klasovost (*Klas*) a bezchuťovost (*Bez*) jsou rovněž pozitivně korelovány ale jsou negativně korelovány se znaky vůně hrachu (*Aro*), sladkost (*Slad*) a medovost (*Med*), protože oba shluky znaků leží na opačných stranách vůči počátku.

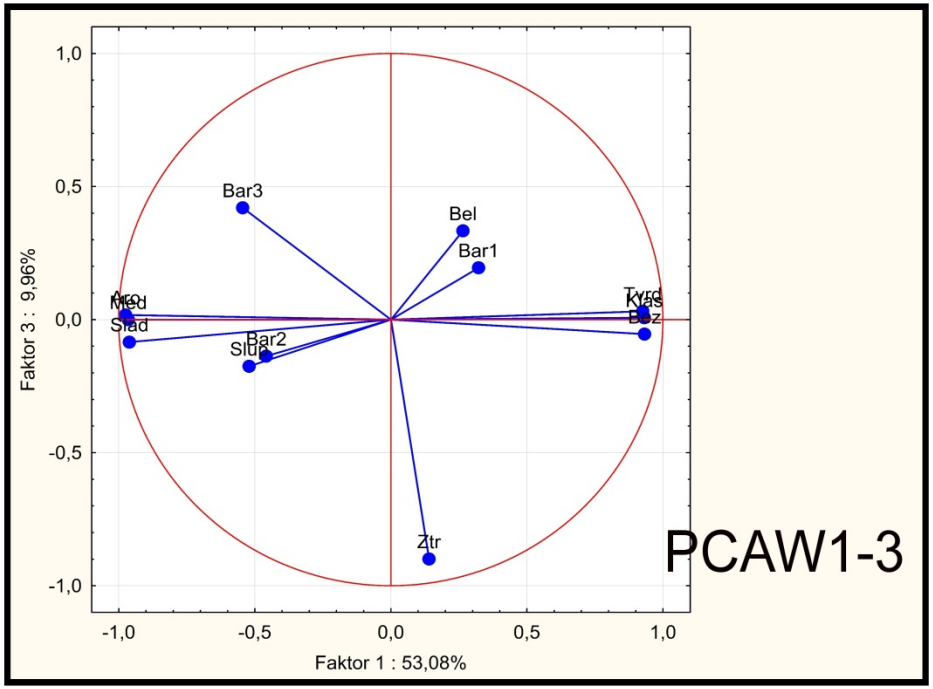
3) Druhá hlavní komponenta *PC*2 ukazuje, že barva 1 (*Bar1*), bělost (*Bel*) a ztráta (*Ztr*) jsou v horní části diagramu a obě jsou negativně korelovány s barvou 2 (*Bar2*) a barvou 3 (*Bar3*), které jsou umístěny v dolní části diagramu.

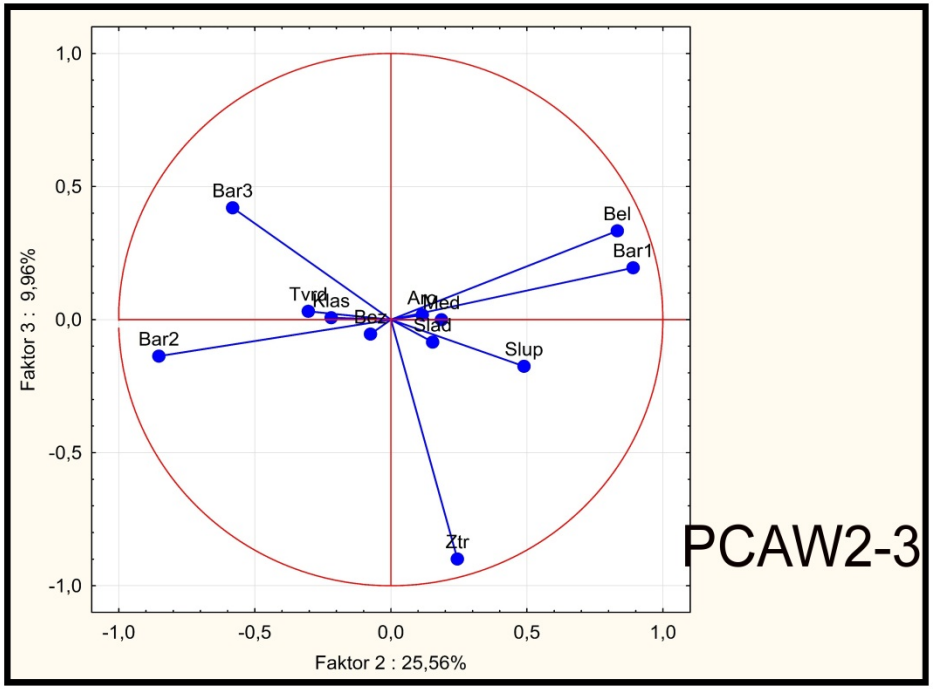
4) Vzorky hrachu nahoře diagramu jsou bělejší a vzorky v dolní části budou barevnější.

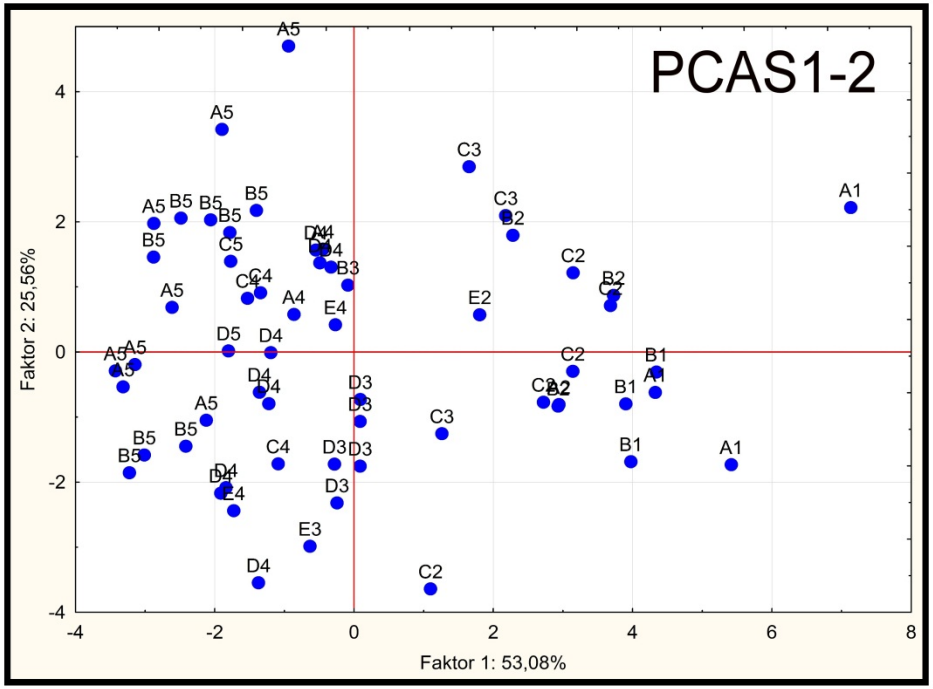
5) Slupka zrn *Slup* hrachu nekoreluje ani s bělostí (*Bel*) ani s chuťovými vlastnostmi hrachu vůně (*Aro*), sladkost (*Slad*) a medovost (*Med*).

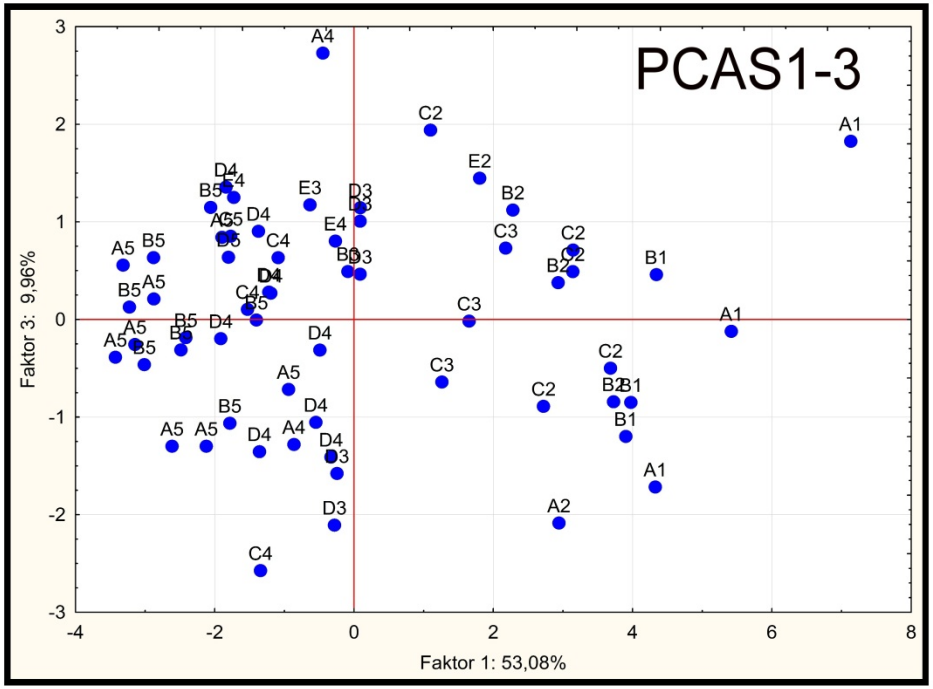

Cattelův graf
Vlast. číslo
Pořadí vl. čísla
53,08%
25,56%
9,96%
4,96%
3,00%
1,17%
,93%
,47%
,39%
,24%
,17%
,08%

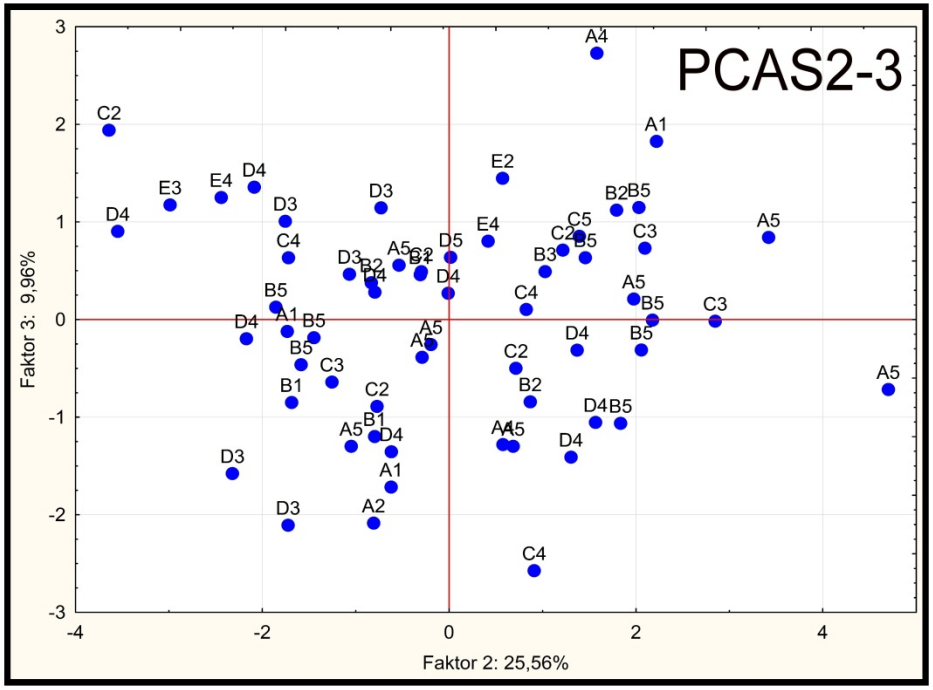

PCAW1-2
Faktor 1 : 53,08%
Faktor 2 : 25,56%
Bar1
Bel
Slup
Ztr
Med
Slad
Aro
Bez
Klas
Tvrd
Bar3
Bar2


PCAW1-3
Faktor 1 : 53,08%
Faktor 3 : 9,96%
Bar3
Bel
Bar1
Aro
Med
Slad
Tvrd
Klas
Bez
Bar2
Slup
Ztr


PCAW2-3
Faktor 2 : 25,56%
Faktor 3 : 9,96%
Bar3
Bel
Bar1
Tvrd
Klas
Aro
Med
Bez
Slad
Bar2
Slup
Ztr


PCAS1-2
Faktor 2: 25,56%
Faktor 1: 53,08%


PCAS1-3
Faktor 3: 9,96%
Faktor 1: 53,08%


PCAS2-3
Faktor 3: 9,96%
Faktor 2: 25,56%

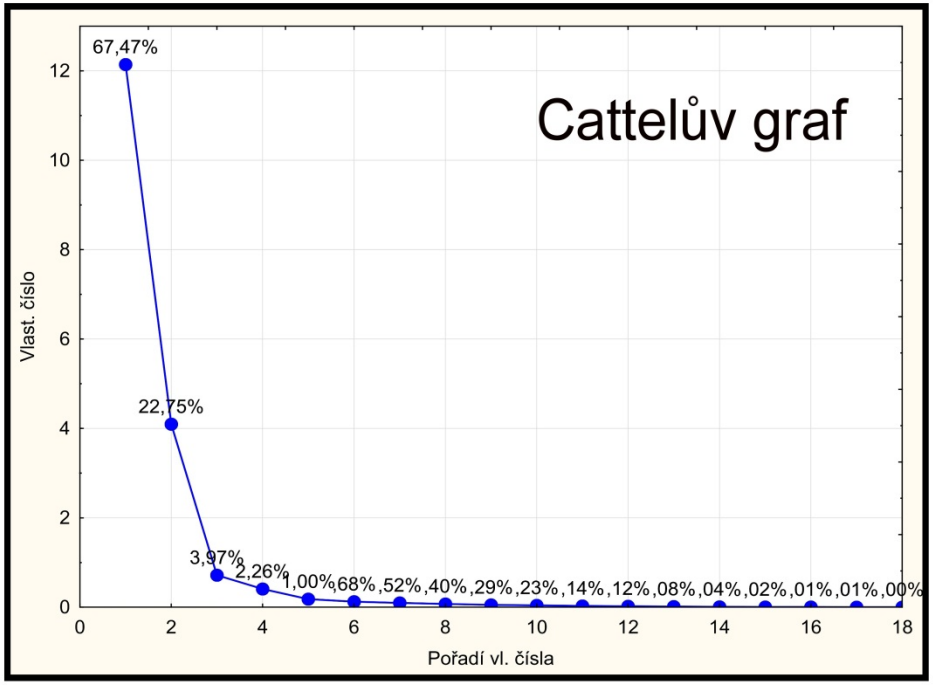
PŘÍKLAD 4.5

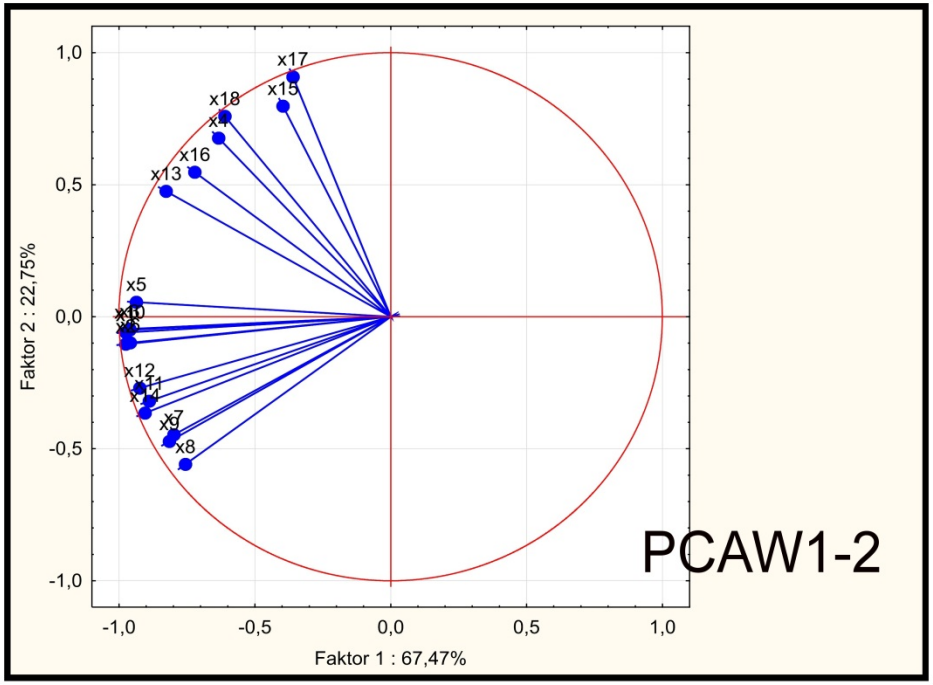
Chromatografická analýza farmakologických sloučenin

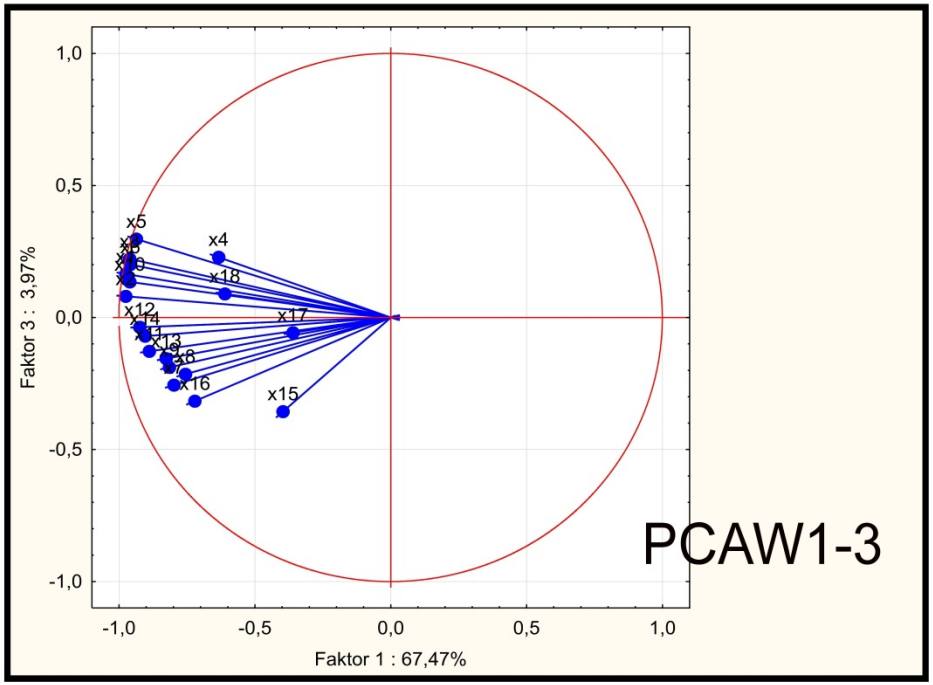
Byly měřeny hodnoty R_F pro 20 sloučenin s 18 eluenty. Žádné eluční činidlo však neprovedlo úplné rozdělení. Cílem je nalézt minimální výběr elučních činidel, které by daly dostatek informace pro kvalitativní analýzu.

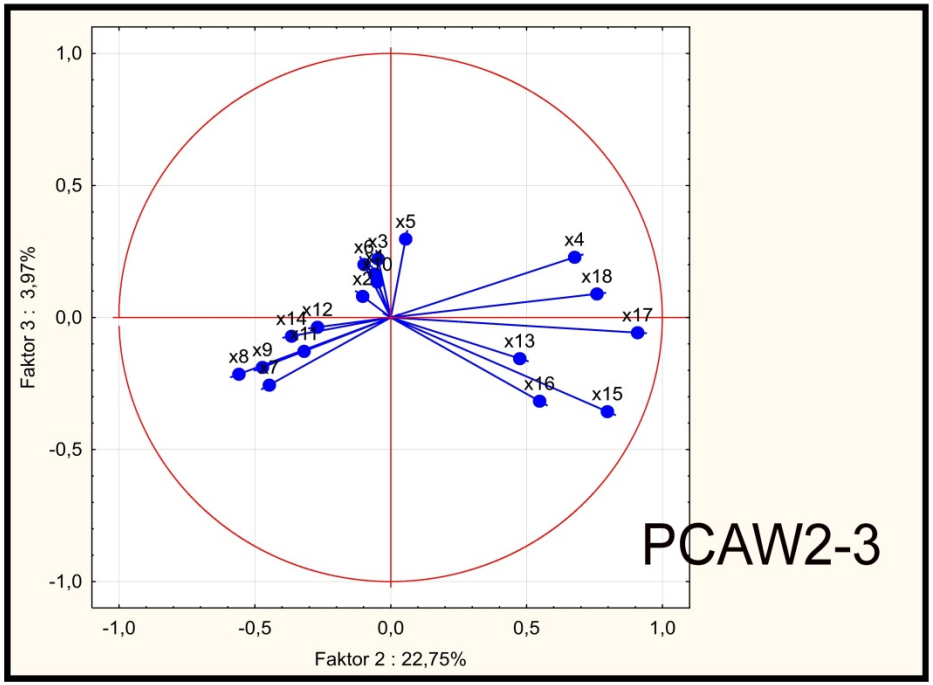
○ ***Data***: Datový soubor GIUSEPPE obsahuje 100 × R_F pro 20 sloučenin (v řádcích byla jména zkrácena na maximálně 8 písmen) a ve sloupcích je 18 elučních činidel představujících zde znaky: i vzorek, x_1 směs toluen : aceton : ethanol : 30 % amoniak = 45 : 45 : 7 : 3, x_2 směs ethylacetát : benzen : methanol : 30 % amoniak = 60 : 35 : 6.5 : 2.5, x_3 směs benzen : dioxan : ethanol : 30 % amoniak = 50 : 40 : 7.5 : 2.5, x_4 směs methanol : 30 % amoniak = 100 : 1.5, x_5 směs benzen : 2-propanol : methanol : 30 % amoniak = 70 : 30 : 20 : 5, x_6 směs ethylacetát : methanol : 30 % amoniak = 85 : 10 : 5, x_7 směs cyklohexan : toluen : diethylamin = 65 : 25 : 10, x_8 směs cyklohexan : toluen : diethylamin = 75 : 15 : 10, x_9 směs cyklohexan : benzen : metanol : diethylamin = 70 : 20 : 10 : 5, x_{10} směs chloroform : aceton : diethylamin = 50 : 40 : 10, x_{11} směs cyklohexan : chloroform : diethylamin = 50 : 40 : 10, x_{12} směs benzen : ethylacetát : diethylamin = 50 : 40 : 10, x_{13} směs xylen : methylethylketon : methanol : diethylamin = 40 : 40 : 6 : 2, x_{14} směs diethylether : diethylamin = 95 : 5, x_{15} směs ethylacetát : chloroform = 50 : 50, x_{16} směs ethylacetát : chloroform [A] = 50 : 50, x_{17} směs butanol : methanol = 40 : 60, x_{18} směs butanol : methanol [A] = 40 : 60, kde [A] značí, že byl užit 0.1M methanolát draselný.

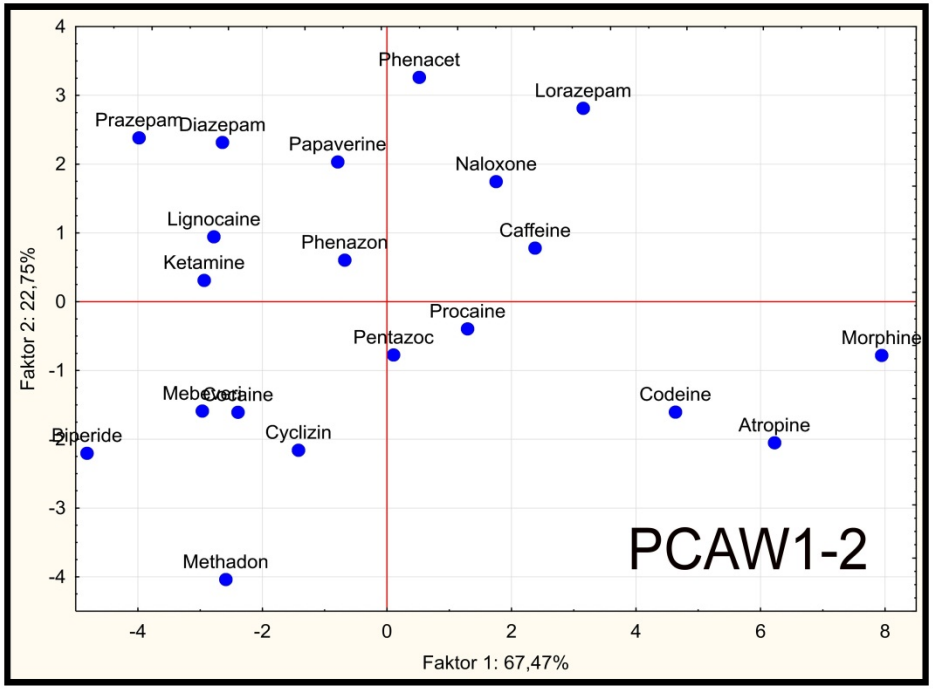
i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
Atropine	20	16	29	23	62	33	4	2	13	47	25	42	18	12	0	0	5	8
Biperide	91	90	87	68	92	87	73	72	64	85	81	86	68	94	11	40	40	65
Caffeine	55	42	52	68	77	54	8	5	13	60	30	51	41	20	13	12	54	57
Cocaine	81	82	81	71	87	82	46	41	38	81	72	80	52	72	6	24	30	57
Codeine	38	31	44	39	71	43	12	9	16	49	29	36	22	14	0	0	15	21
Cyclizin	71	72	80	64	85	80	49	47	40	75	71	73	39	59	2	9	34	54
Diazepam	76	79	80	78	85	80	28	21	29	80	61	75	72	54	54	50	85	87
Ketamine	77	79	80	76	86	79	71	33	32	81	66	76	67	66	27	37	66	79
Lignocaine	77	79	80	73	86	80	35	30	28	84	73	77	66	64	25	54	68	84
Lorazepam	47	34	53	77	73	46	2	0	12	52	7	22	47	8	28	15	85	79
Mebeveri	85	90	90	65	90	85	43	33	38	88	70	87	62	76	5	29	29	53
Methadon	85	84	88	48	89	83	63	64	48	85	73	86	38	86	1	10	13	29
Morphine	18	9	15	39	56	20	2	0	5	20	2	8	13	3	0	1	16	18
Naloxone	48	40	62	75	79	48	15	11	22	52	26	40	60	21	18	21	67	77
Papaverine	68	66	76	79	88	71	12	7	18	78	56	62	54	30	28	41	76	80
Pentazoc	72	66	81	65	87	76	22	18	26	69	41	54	39	44	2	11	32	59
Phenacet	64	58	62	79	87	66	4	1	13	68	18	41	58	24	41	40	86	84
Phenazon	66	53	70	83	86	65	30	22	24	77	60	66	45	54	15	21	68	74
Prazepam	81	83	86	83	88	81	41	31	35	83	66	82	75	74	65	67	86	88
Procaine	64	60	70	65	82	73	8	5	16	66	24	54	37	50	1	11	29	53

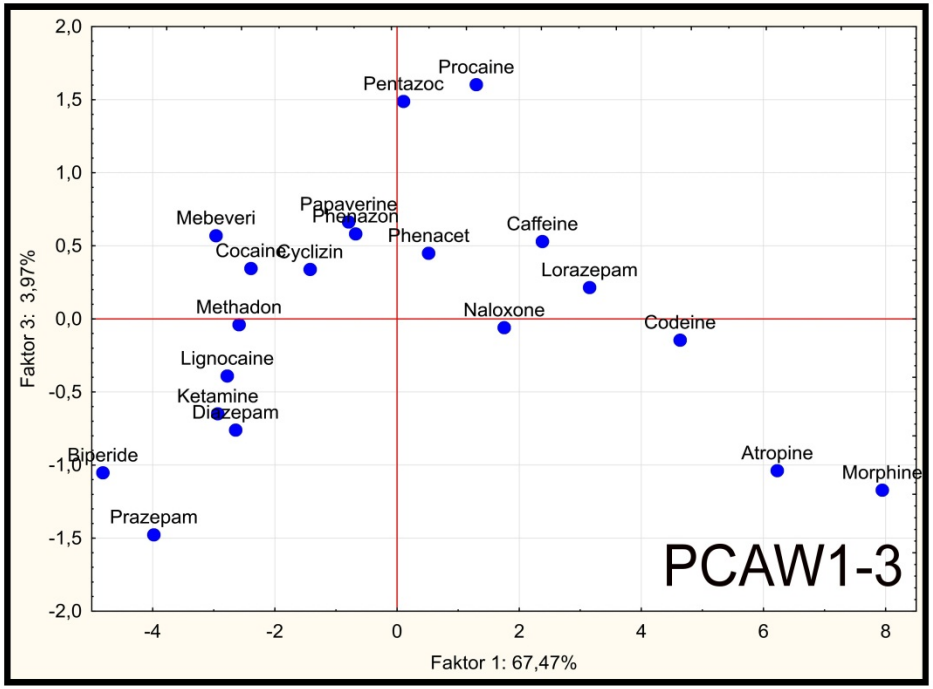

Cattelův graf
67,47%
22,75%
3,97%
2,26%
1,00%
68%
52%
40%
29%
23%
14%
12%
08%
04%
02%
01%
01%
00%
Vlast. číslo
Pořadí vl. čísla

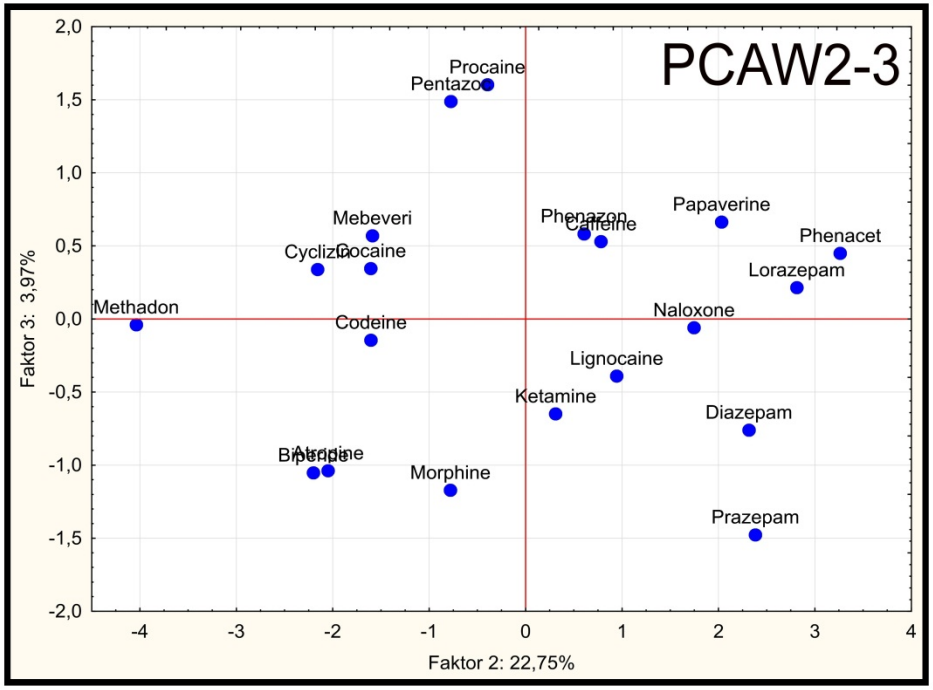

PCAW1-2
Faktor 2 : 22,75%
Faktor 1 : 67,47%


PCAW1-3
Faktor 3 : 3,97%
Faktor 1 : 67,47%


PCAW2-3
Faktor 3 : 3,97%
Faktor 2 : 22,75%


PCAW1-2
Faktor 1: 67,47%
Faktor 2: 22,75%
Phenacet
Lorazepam
Prazepam
Diazepam
Papaverine
Naloxone
Lignocaine
Caffeine
Phenazon
Ketamine
Procaine
Pentazoc
Morphine
Mebeveri
Cocaine
Codeine
Atropine
Biperide
Cyclizin
Methadon


PCAW1-3
Faktor 1: 67,47%
Faktor 3: 3,97%
Procaine
Pentazoc
Papaverine
Mebeveri
Phenazon
Caffeine
Phenacet
Cocaine
Cyclizin
Lorazepam
Methadon
Naloxone
Codeine
Lignocaine
Ketamine
Diazepam
Biperide
Atropine
Morphine
Prazepam

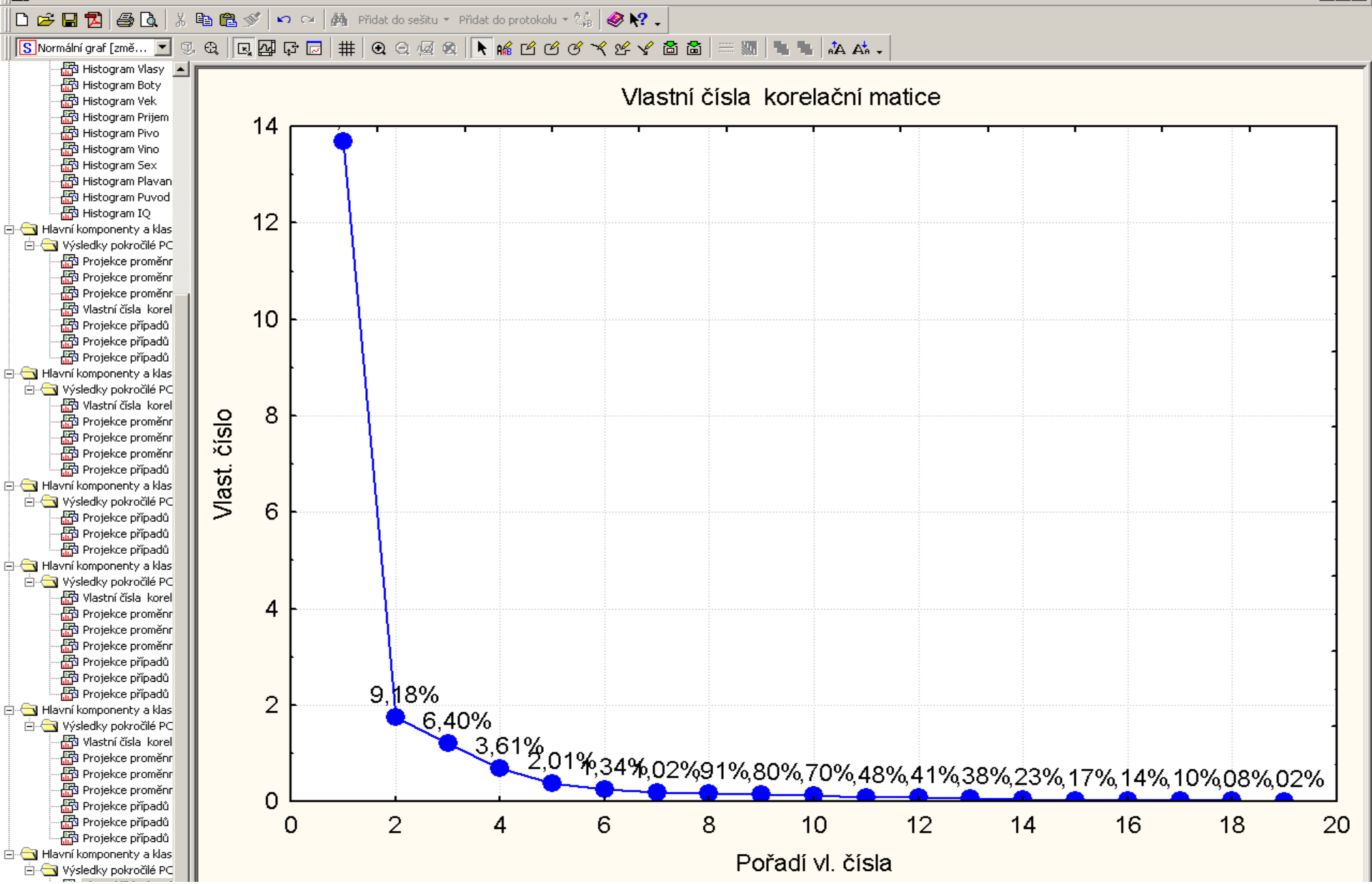

PCAW2-3
Faktor 2: 22,75%
Faktor 3: 3,97%
Procaine
Pentazoc
Mebeveri
Phenazon
Papaverine
Caffeine
Phenacet
Cyclizin
Cocaine
Lorazepam
Methadon
Naloxone
Codeine
Lignocaine
Ketamine
Diazepam
Morphine
Prazepam

PŘÍKLAD 4.6 *Popis a třídění polétavých mšic*

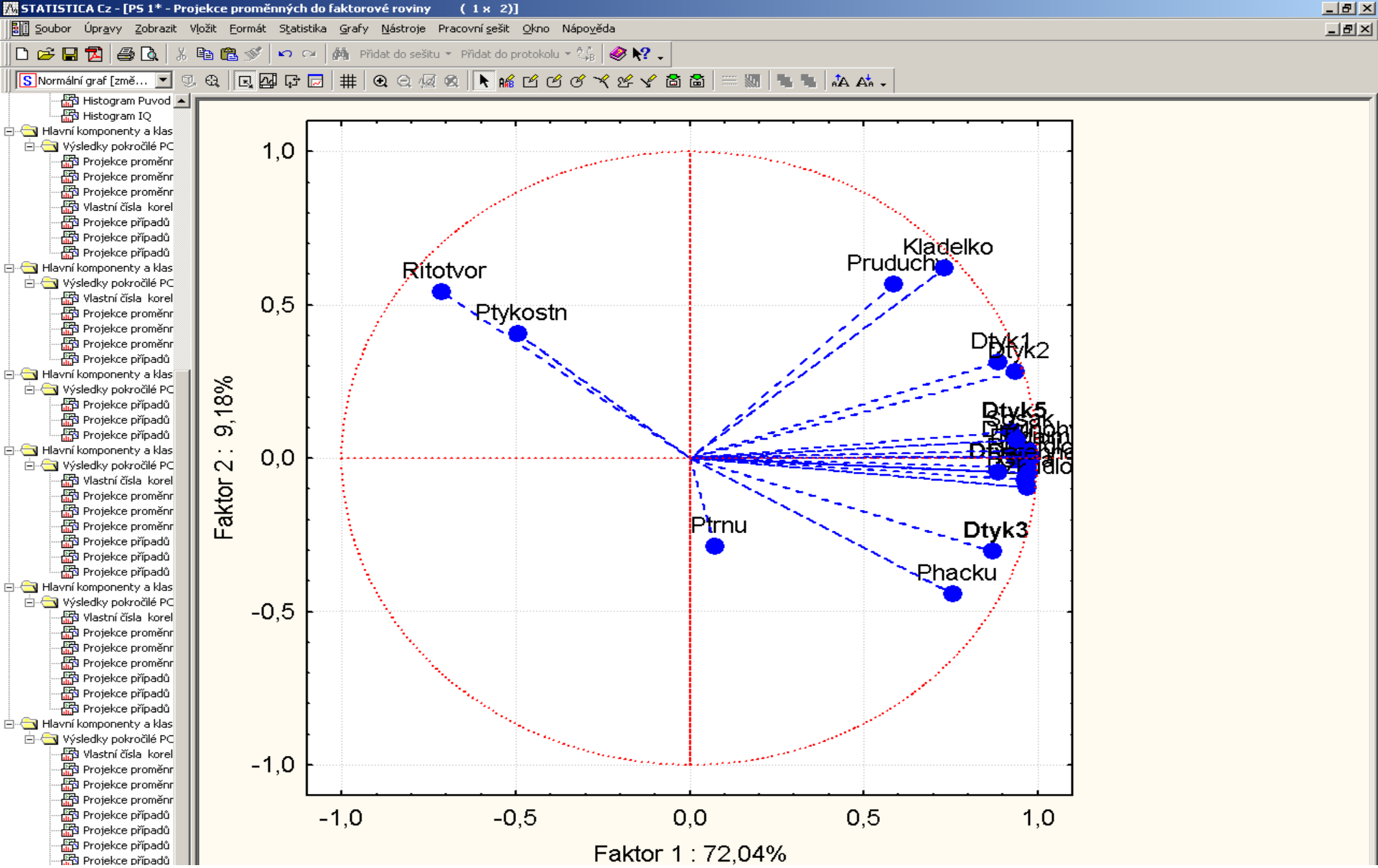
Jeffers (1967) studoval 40 polétavých mšic *(Alate adelges)* pomoci světelné pasti, změřeno 19 znaků: 14 znaků délky a šířky, 4 znaky o počtu, 1 znak binární, přítomnost či absenci. Mšice se obtížně rozlišují dle taxonometrických klíčů. Před PCA je třeba standardizaci dat, protože znaky představují směs délek a počtů.

○ *Data*: x_1 značí délku těla, x_2 značí šířku těla, x_3 je délka předního křídla, x_4 je délka zadního křídla, x_5 je počet průduchů, x_6 je délka tykadla I, x_7 je délka tykadla II, x_8 je délka tykadla III, x_9 je délka tykadla IV, x_{10} je délka tykadla V, x_{11} je počet tykadlových ostnů, x_{12} je délka posledního článku nohy, x_{13} je délka holeně, tibia, x_{14} je délka stehna, x_{15} je délka sosáku, x_{16} je délka kladélka, x_{17} je počet kladélkových trnů, x_{18} je řitní otvor, x_{19} je počet háčků zadních křídel.

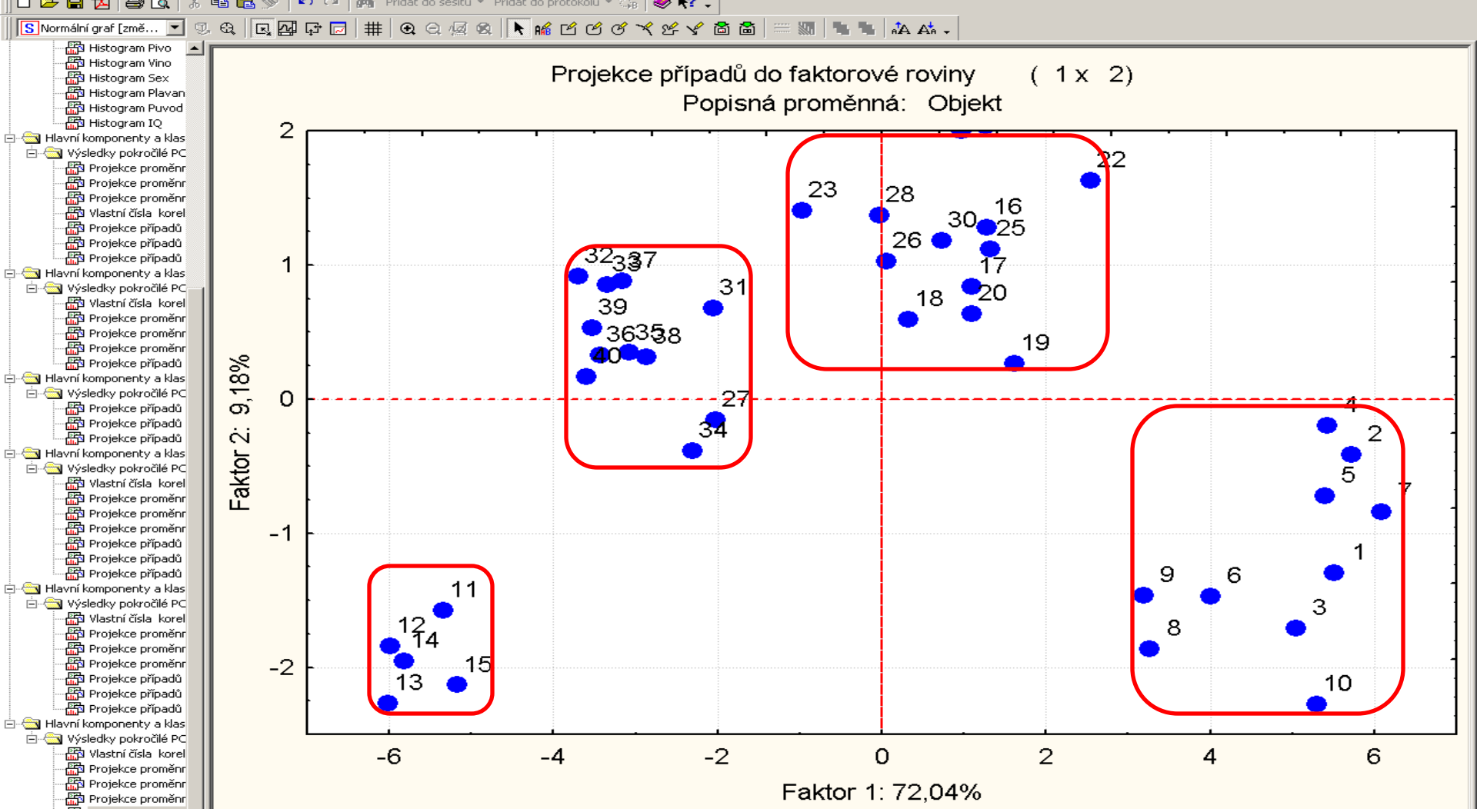
i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
1	21.2	11.0	7.5	4.8	5.0	2.0	2.0	2.8	2.8	3.3	3.0	4.4	4.5	3.6	7.0	4.0	8.0	0.0	3.0
2	20.2	10.0	7.5	5.0	5.0	2.3	2.1	3.0	3.0	3.2	5.0	4.2	4.5	3.5	7.6	4.2	8.0	0.0	3.0
3	20.2	10.0	7.0	4.6	5.0	1.9	2.1	3.0	2.5	3.3	1.0	4.2	4.4	3.3	7.0	4.0	6.0	0.0	3.0
4	22.5	8.8	7.4	4.7	5.0	2.4	2.1	3.0	2.7	3.5	5.0	4.2	4.4	3.6	6.8	4.1	6.0	0.0	3.0
5	20.6	11.0	8.0	4.8	5.0	2.4	2.0	2.9	2.7	3.0	4.0	4.2	4.7	3.5	6.7	4.0	6.0	0.0	3.0
6	19.1	9.2	7.0	4.5	5.0	1.8	1.9	2.8	3.0	3.2	5.0	4.1	4.3	3.3	5.7	3.8	8.0	0.0	3.5
7	20.8	11.4	7.7	4.9	5.0	2.5	2.1	3.1	3.1	3.2	4.0	4.2	4.7	3.6	6.6	4.0	8.0	0.0	3.0
8	15.5	8.2	6.3	4.9	5.0	2.0	2.0	2.9	2.4	3.0	3.0	3.7	3.8	2.9	6.7	3.5	6.0	0.0	3.5
9	16.7	8.8	6.4	4.5	5.0	2.1	1.9	2.8	2.7	3.1	3.0	3.7	3.8	2.8	6.1	3.7	8.0	0.0	3.0
10	19.7	9.9	8.2	4.7	5.0	2.2	2.0	3.0	3.0	3.1	0.0	4.1	4.3	3.3	6.0	3.8	8.0	0.0	3.0
11	10.6	5.2	3.9	2.3	4.0	1.2	1.0	2.0	2.0	2.2	6.0	2.5	2.5	2.0	4.5	2.7	4.0	1.0	2.0
12	9.2	4.5	3.7	2.2	4.0	1.3	1.2	2.0	1.6	2.1	5.0	2.4	2.3	1.8	4.1	2.4	4.0	1.0	2.0
13	9.6	4.5	3.6	2.3	4.0	1.3	1.0	1.9	1.7	2.2	4.0	2.4	2.3	1.7	4.0	2.3	4.0	1.0	2.0
14	8.5	4.0	3.8	2.2	4.0	1.3	1.1	1.9	2.0	2.1	5.0	2.4	2.4	1.9	4.4	2.3	4.0	1.0	2.0
15	11.0	4.7	4.2	2.3	4.0	1.2	1.0	1.9	2.0	2.2	4.0	2.5	2.5	2.0	4.5	2.6	4.0	1.0	2.0
16	18.1	8.2	5.9	3.5	5.0	1.9	1.9	1.9	2.7	2.8	4.0	3.5	3.8	2.9	6.0	4.5	9.0	1.0	2.0
17	17.6	8.3	6.0	3.8	5.0	2.0	1.9	2.0	2.2	2.9	3.0	3.5	3.6	2.8	5.7	4.3	10.0	1.0	2.0
18	19.2	6.6	6.2	3.4	5.0	2.0	1.8	2.2	2.3	2.8	4.0	3.5	3.4	2.5	5.3	3.8	10.0	1.0	2.0
19	15.4	7.6	7.1	3.4	5.0	2.0	1.9	2.5	2.5	2.9	4.0	3.3	3.6	2.7	6.0	4.2	8.0	1.0	3.0
20	15.1	7.3	6.2	3.8	5.0	2.0	1.8	2.1	2.4	2.5	4.0	3.7	3.7	2.8	6.4	4.3	10.0	1.0	2.5
21	16.1	7.9	5.8	3.7	5.0	2.1	1.9	2.3	2.6	2.9	5.0	3.6	3.6	2.7	6.0	4.5	0.0	1.0	2.0
22	19.1	8.8	6.4	3.9	5.0	2.2	2.0	2.3	2.4	2.9	4.0	3.8	4.0	3.0	6.5	4.5	0.0	1.0	2.5
23	15.3	6.4	5.3	3.3	5.0	1.7	1.6	2.0	2.2	2.5	5.0	3.4	3.4	2.6	5.4	4.0	0.0	1.0	2.0
24	14.8	8.1	6.2	3.7	5.0	2.2	2.0	2.2	2.4	3.2	5.0	3.5	3.7	2.7	6.0	4.1	0.0	1.0	2.0
25	16.2	7.7	6.9	3.7	5.0	2.0	1.8	2.3	2.4	2.8	4.0	3.8	3.7	2.7	5.7	4.2	0.0	1.0	2.5
26	13.4	6.9	5.7	3.4	5.0	2.0	1.8	2.8	2.0	2.6	4.0	3.6	3.6	2.6	5.5	3.9	0.0	1.0	2.0
27	12.9	5.8	4.8	2.6	5.0	1.6	1.5	1.9	2.1	2.6	5.0	2.8	3.0	2.2	5.1	3.6	9.0	1.0	3.0
28	12.0	6.5	5.3	3.2	5.0	1.9	1.9	2.3	2.5	3.0	5.0	3.3	3.5	2.6	5.4	4.3	8.0	1.0	2.0
29	14.1	7.0	5.5	3.6	5.0	2.2	2.0	2.3	2.5	3.1	5.0	3.6	3.7	2.8	5.8	4.1	0.0	1.0	2.0
30	16.7	7.2	5.7	3.5	5.0	1.9	1.9	2.5	2.3	2.8	5.0	3.4	3.6	2.7	6.0	4.0	0.0	1.0	2.5
31	14.1	5.4	5.0	3.0	5.0	1.7	1.6	1.8	2.5	2.4	5.0	2.7	2.9	2.2	5.3	3.6	8.0	1.0	2.0
32	10.0	6.0	4.2	2.5	5.0	1.6	1.4	1.4	2.0	2.7	6.0	2.8	2.5	1.8	4.8	3.4	8.0	1.0	2.0
33	11.4	4.5	4.4	2.7	5.0	1.8	1.5	1.9	1.7	2.5	5.0	2.7	2.5	1.9	4.7	3.7	8.0	1.0	2.0
34	12.5	5.5	4.7	2.3	5.0	1.8	1.4	1.8	2.2	2.4	4.0	2.8	2.6	2.0	5.1	3.7	8.0	0.0	2.0
35	13.0	5.3	4.7	2.3	5.0	1.6	1.4	1.8	1.8	2.5	4.0	2.7	2.7	2.1	5.0	3.6	8.0	1.0	2.0
36	12.4	5.2	4.4	2.6	5.0	1.6	1.4	1.8	2.2	2.2	5.0	2.7	2.5	2.0	5.0	3.2	6.0	1.0	2.0
37	12.0	5.4	4.9	3.0	5.0	1.7	1.5	1.7	1.9	2.4	5.0	2.7	2.7	2.0	4.2	3.7	6.0	1.0	2.0
38	10.7	5.6	4.5	2.8	5.0	1.8	1.4	1.8	2.2	2.4	4.0	2.7	2.6	2.0	5.0	3.5	8.0	1.0	2.0
39	11.1	5.5	4.3	2.6	5.0	1.7	1.5	1.8	1.9	2.4	5.0	2.6	2.5	1.9	4.6	3.4	8.0	1.0	2.0
40	12.8	5.7	4.8	2.8	5.0	1.6	1.4	1.7	1.9	2.3	5.0	2.3	2.5	1.9	5.0	3.1	8.0	1.0	2.0



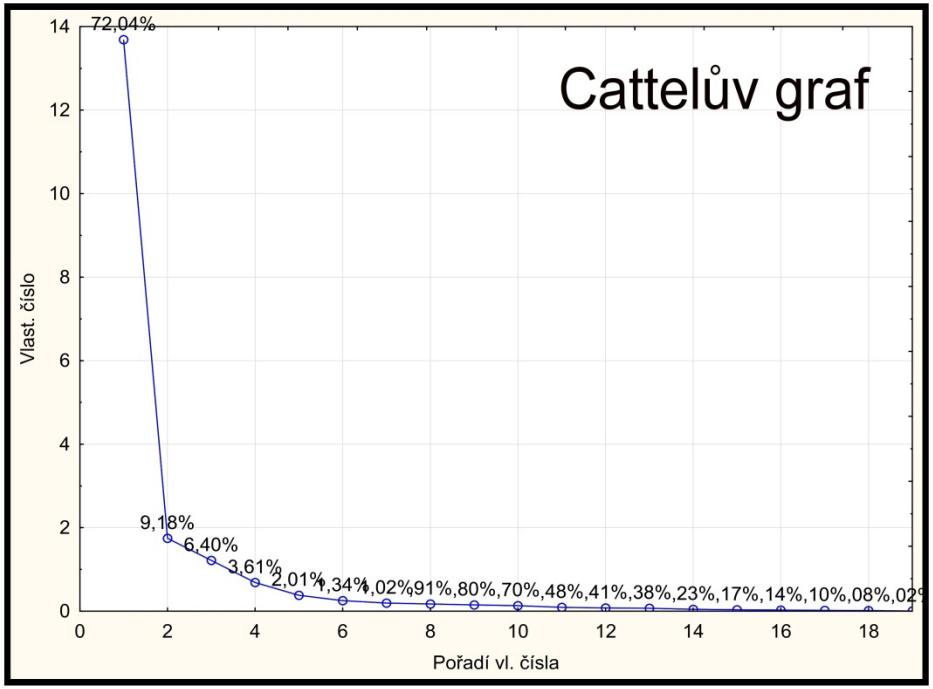
203. Cattelův indexový graf vlastních čísel odkrývá 2 první hlavní komponenty.

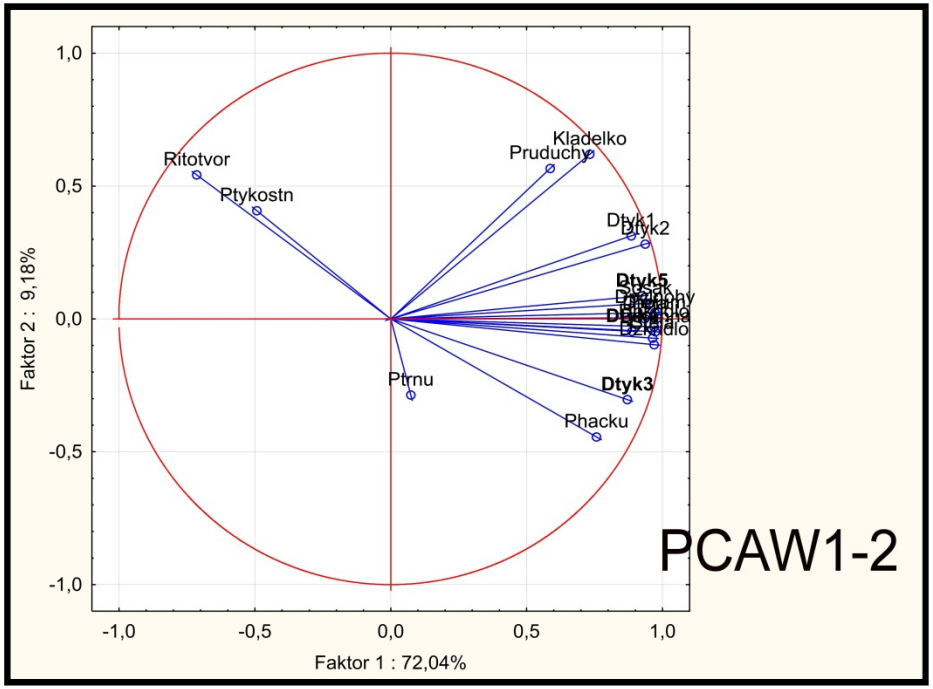


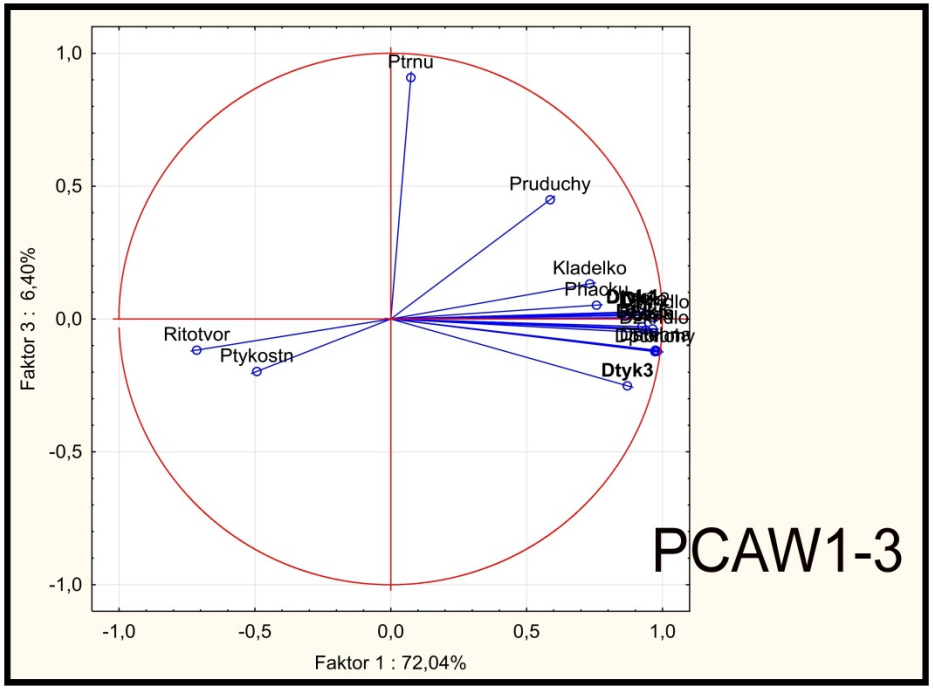
204. Otevřeme první graf komponentních vah 1. a 2. hlavní komponenty a kliknutím na kolečka znaků nastavíme jejich grafiku (velikost, tvar, barvu).

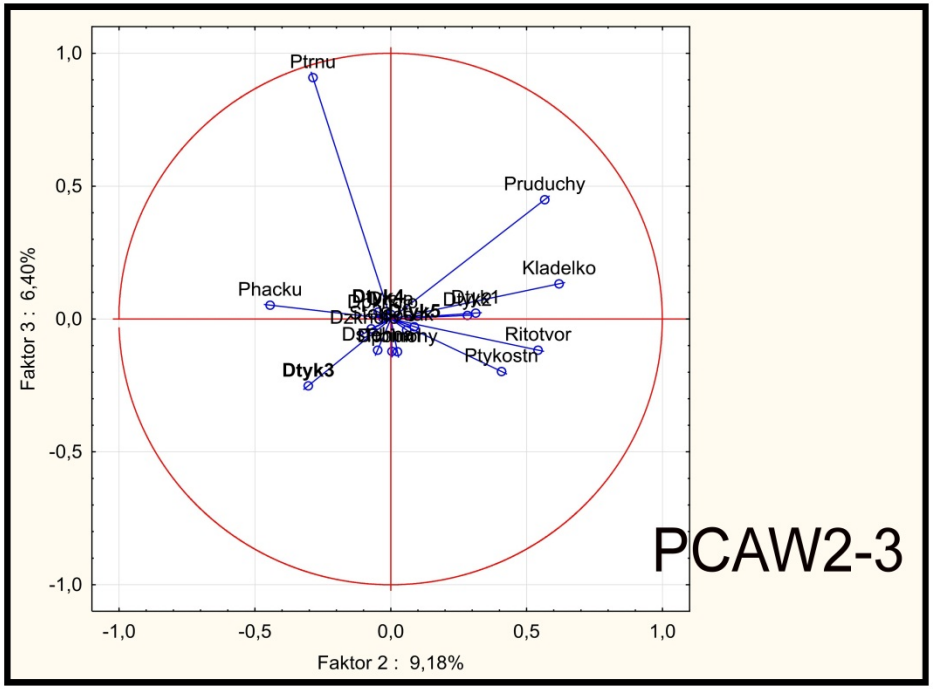


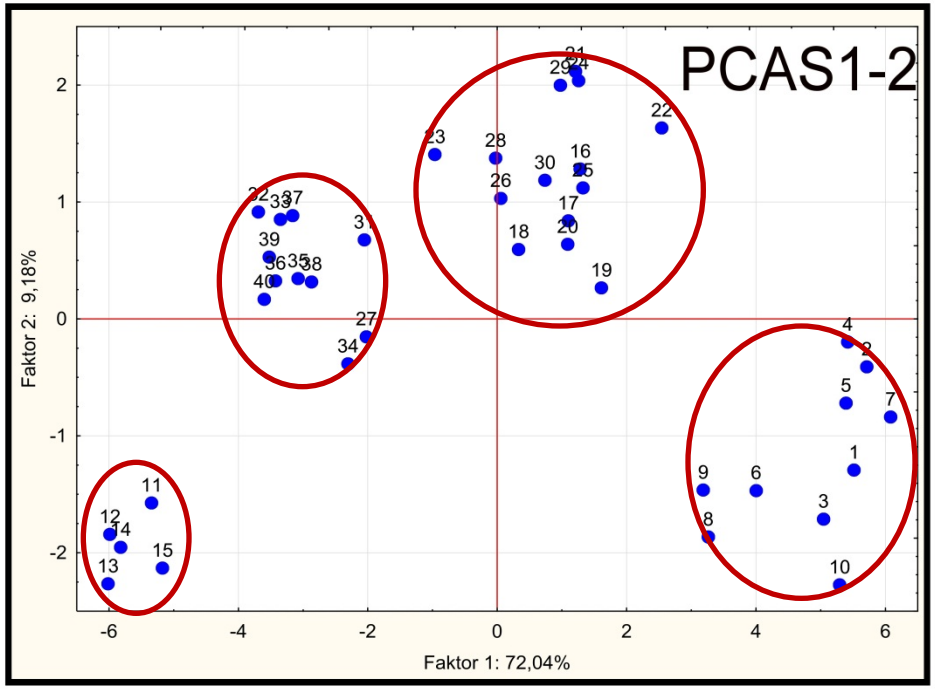
207. Otevřeme rozptylový diagram komponentního skóre pro 1. a 2. hlavní komponentu. Kliknutím na čísla na osách se otevře okno *Rozložení osy*, ve kterém nastavíme měřítko *Minimum* a *Maximum*, dále pak *Upravit krok* obou os *x* a *y* a obdržíme tento výsledek.

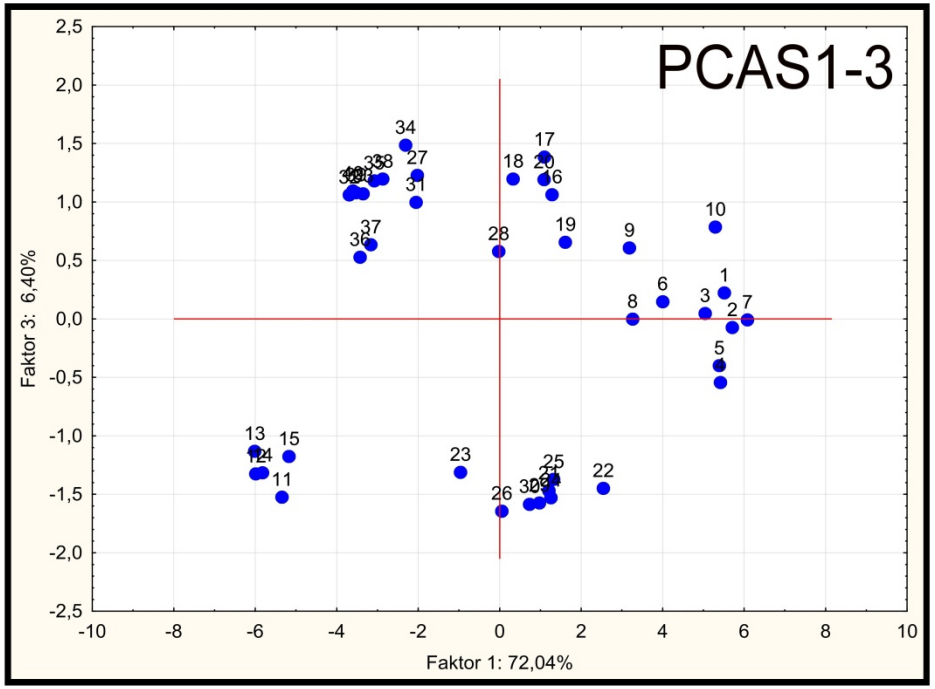

Cattelův graf
Vlast. číslo
Pořadí vl. čísla
72,04%
9,18%
6,40%
3,61%
2,01%
1,34%
1,02%
,91%
,80%
,70%
,48%
,41%
,38%
,23%
,17%
,14%
,10%
,08%

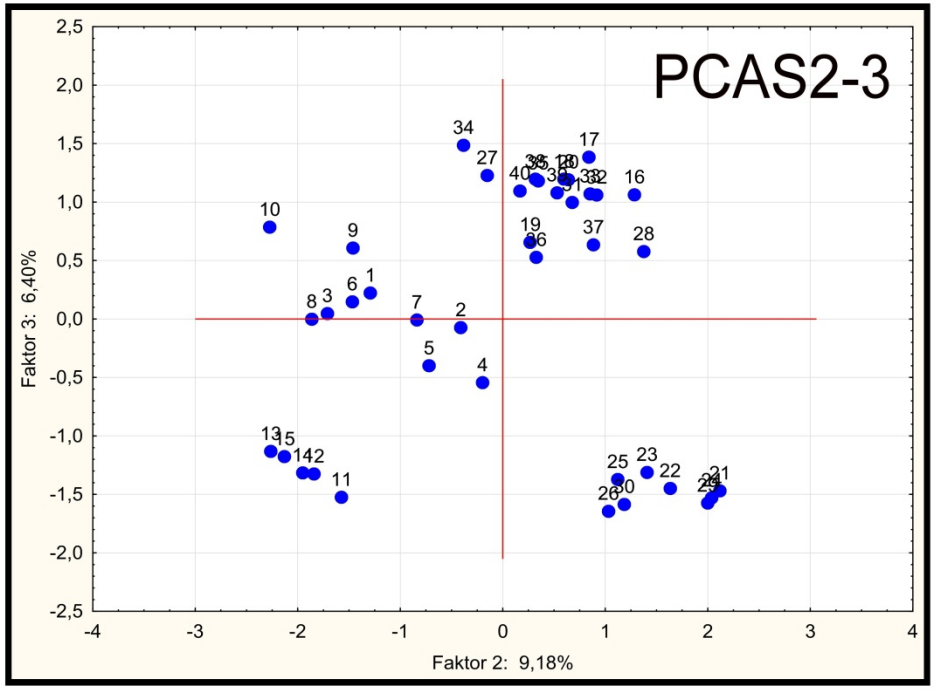

PCAW1-2
Faktor 1 : 72,04%
Faktor 2 : 9,18%
Ritotvor
Ptykostn
Kladelko
Pruduchy
Dtyk1
Dtyk2
Dtyk5
Ptrnu
Dtyk3
Phacku


PCAW1-3
Faktor 1 : 72,04%
Faktor 3 : 6,40%
Ptrnu
Pruduchy
Kladelko
Phacku
Ritotvor
Ptykostn
Dtyk3


PCAW2-3
Faktor 2 : 9,18%
Faktor 3 : 6,40%
Ptrnu
Pruduchy
Kladelko
Phacku
Dtyk1
Dtyk2
Ritotvor
Ptykostn
Dtyk3


PCAS1-2
Faktor 1: 72,04%
Faktor 2: 9,18%


PCAS1-3
Faktor 1: 72,04%
Faktor 3: 6,40%


PCAS2-3
Faktor 2: 9,18%
Faktor 3: 6,40%

Faktorová analýza (FA)

1. Zaměření faktorové analýzy:

1) K vyšetření **vnitřních souvislostí** (korelace) a **odhalení struktury** datové matice.

2) Analýza struktury vnitřních vztahů mezi velkým počtem znaků, **zavedením** několika ***faktorů***.

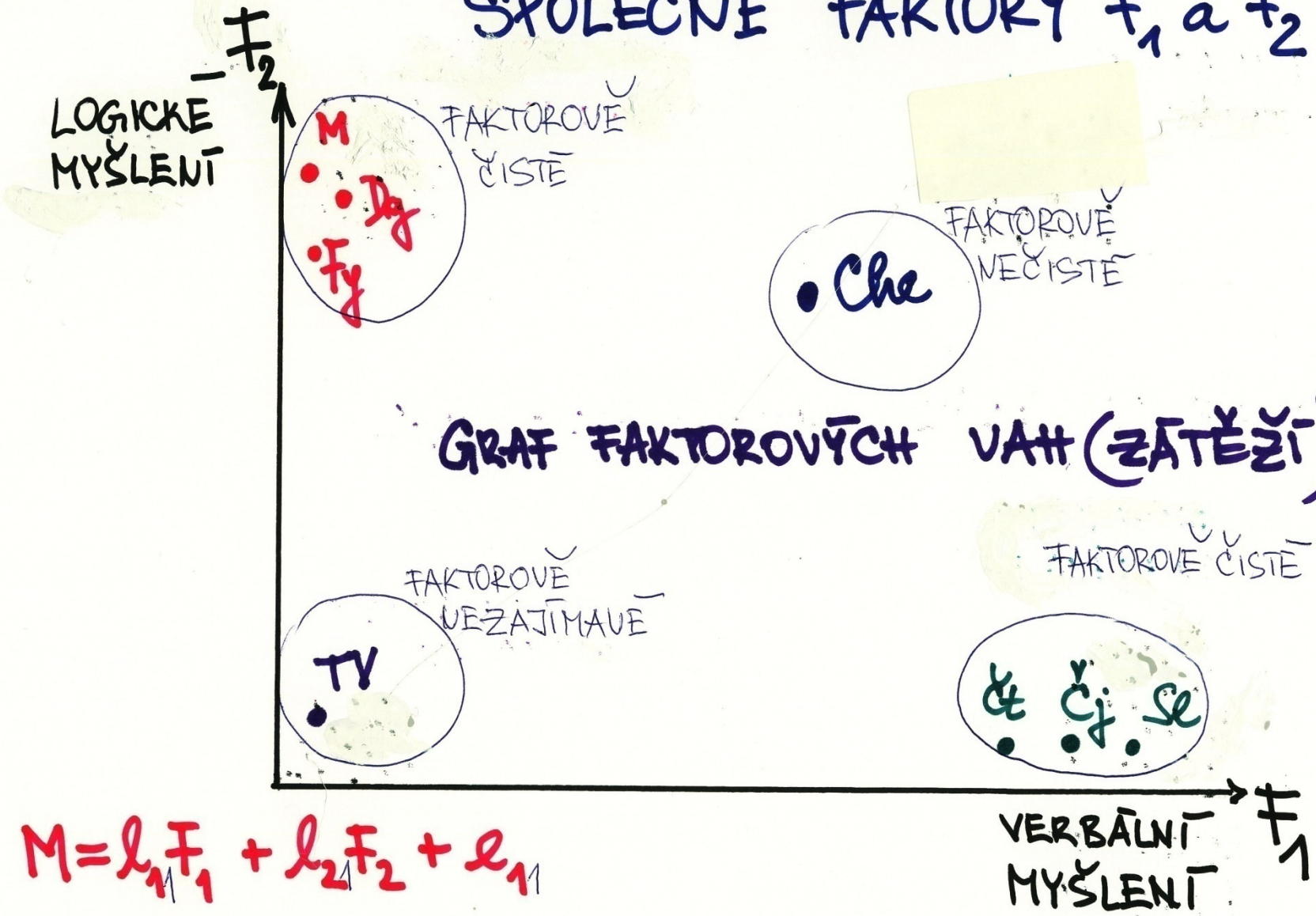
3) Po identifikaci faktorů je každému faktoru přidělen **obsahový význam,** pomocí kterého je každý znak vysvětlen.

4) **Dva primární cíle FA**:

   a) sumarizace a **zestručnění dat**

   b) **redukce dat**.

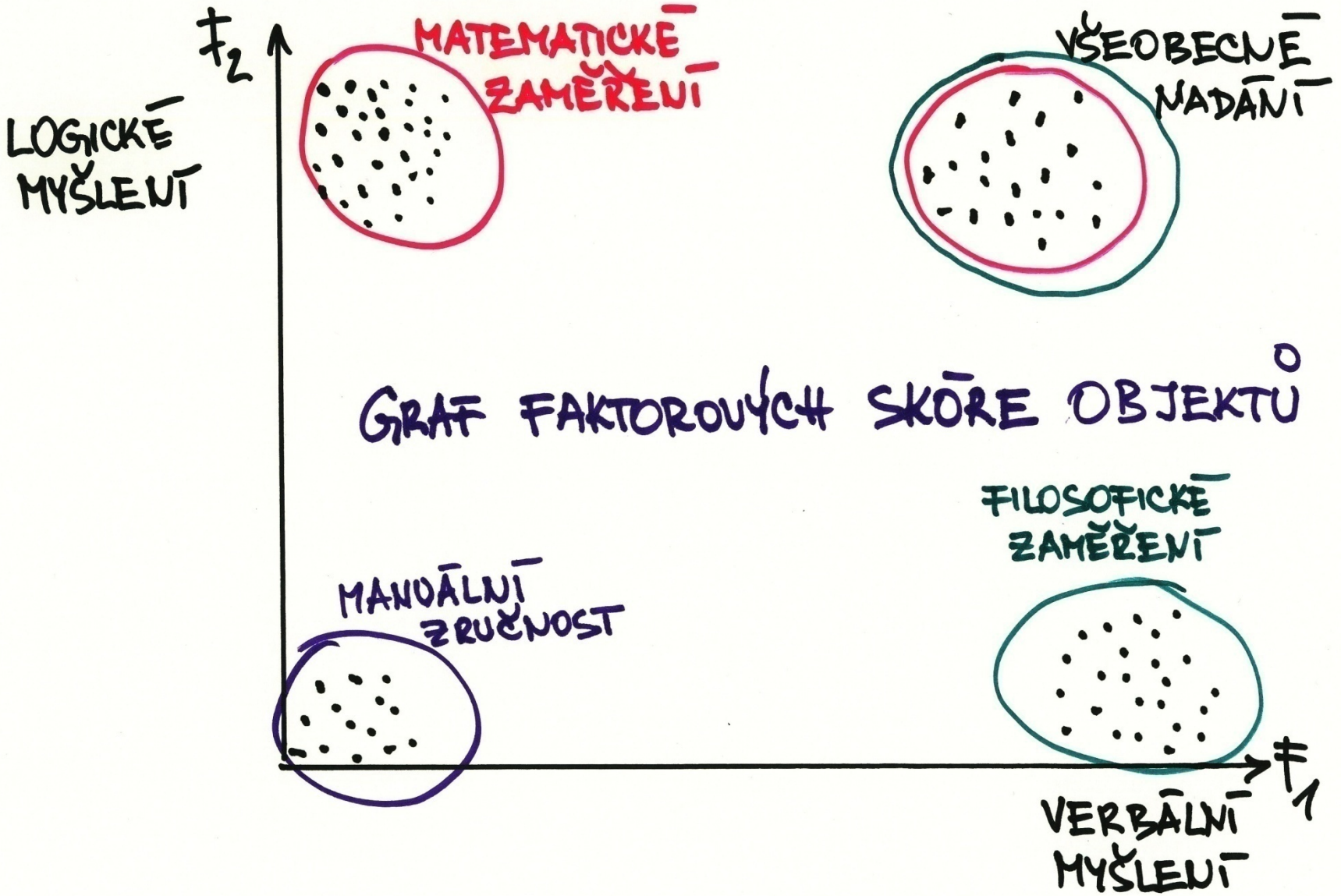
SPOLEČNÉ FAKTORY F_1 a F_2



$$M = l_{11}F_1 + l_{21}F_2 + e_{11}$$

$$Čj = l_{12}F_1 + l_{22}F_2 + e_2$$

SPECIFICKÉ FAKTORY e_1 a e_2
(CHYBOVÉ)


f_2
LOGICKÉ MYŠLENÍ
MATEMATICKÉ ZAMĚŘENÍ
VŠEOBECNÉ NADÁNÍ
GRAF FAKTOROVÝCH SKÓRE OBJEKTŮ
FILOSOFICKÉ ZAMĚŘENÍ
MANUÁLNÍ ZRUČNOST
f_1
VERBÁLNÍ MYŠLENÍ

2. Podstata faktorové analýzy

Matice standardizovaných dat $x_1 = (X_i - \bar{X}_i)/ s_i$ obsahuje P sloupců znaků $X_1, X_2, ..., X_P$ a N řádků objektů.

Standardizovaná data mají u znaků střední hodnotu 0 a rozptyl 1 a jejich kovariance jsou vlastně korelační koeficienty.

$\boldsymbol{x}^{\mathbf{T}} = (x_1, x_2, ..., x_P)^{\mathrm{T}}$ značí jeden i-tý objekt o P znacích s korelační maticí **R**.

Objekty představíme jako lineární kombinaci faktorů $F_1, F_2, ..., F_m$ plus specifický faktor e_i

Formou ***modelu faktorové analýzy (FA)*** ve tvaru

$$x_1 = l_{11}F_1 + l_{12}F_2 + \dots l_{1m}F_m + e_1 ,$$

$$x_2 = l_{21}F_1 + l_{22}F_2 + \dots l_{2m}F_m + e_2 ,$$

$$\dots\dots\dots\dots$$

$$x_P = l_{P1}F_1 + l_{P2}F_2 + \dots l_{Pm}F_m + e_P ,$$

kde ◆ F_1, F_2, ..., F_m jsou *vybrané společné faktory o počtu m*, které vyvolávají korelace mezi P původními znaky.

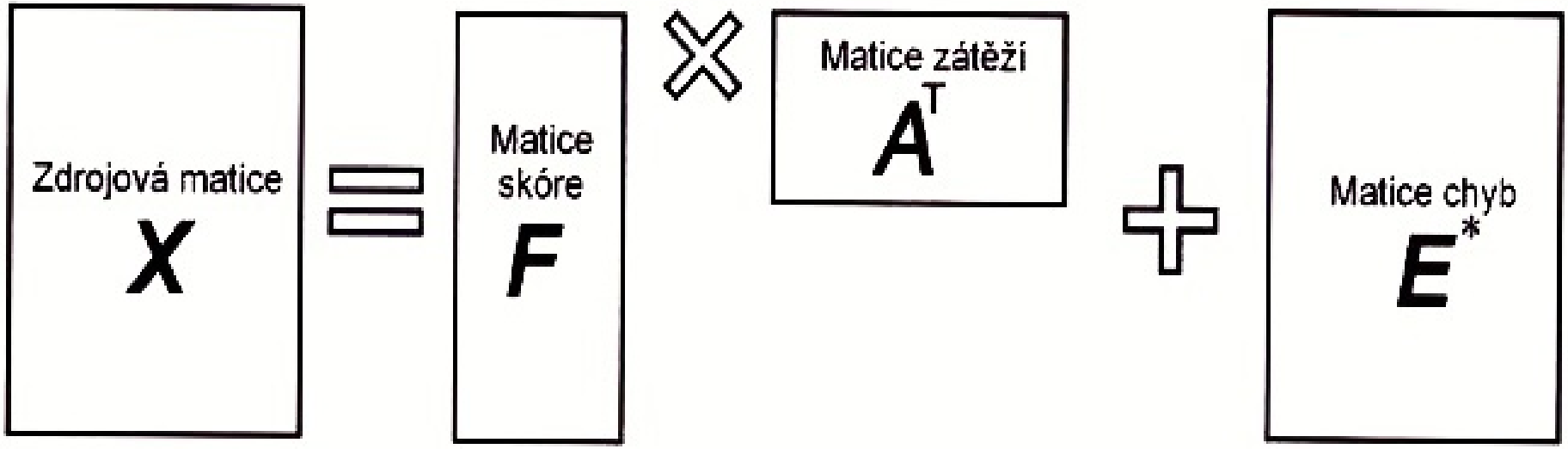
◆ Faktory mají nulovou střední hodnotu a jednotkový rozptyl.

◆ Faktory mají nulovou střední hodnotu a jednotkový rozptyl.

◆ $e_1, e_2, ..., e_P$ jsou *specifické (chybové) faktory*, které přispívají k rozptylu jednotlivých proměnných.

◆ Koeficienty l_{ij} nazýváme *faktorové zátěže* i-tého znaku na j-tém společném faktoru F_j a představují prvky matice faktorových zátěží.

◆ Výběr N objektů lze vyjádřit pouhým přidáním indexu ke každému x_i nebo F_j a e_i, a tak vystihnout každý objekt.



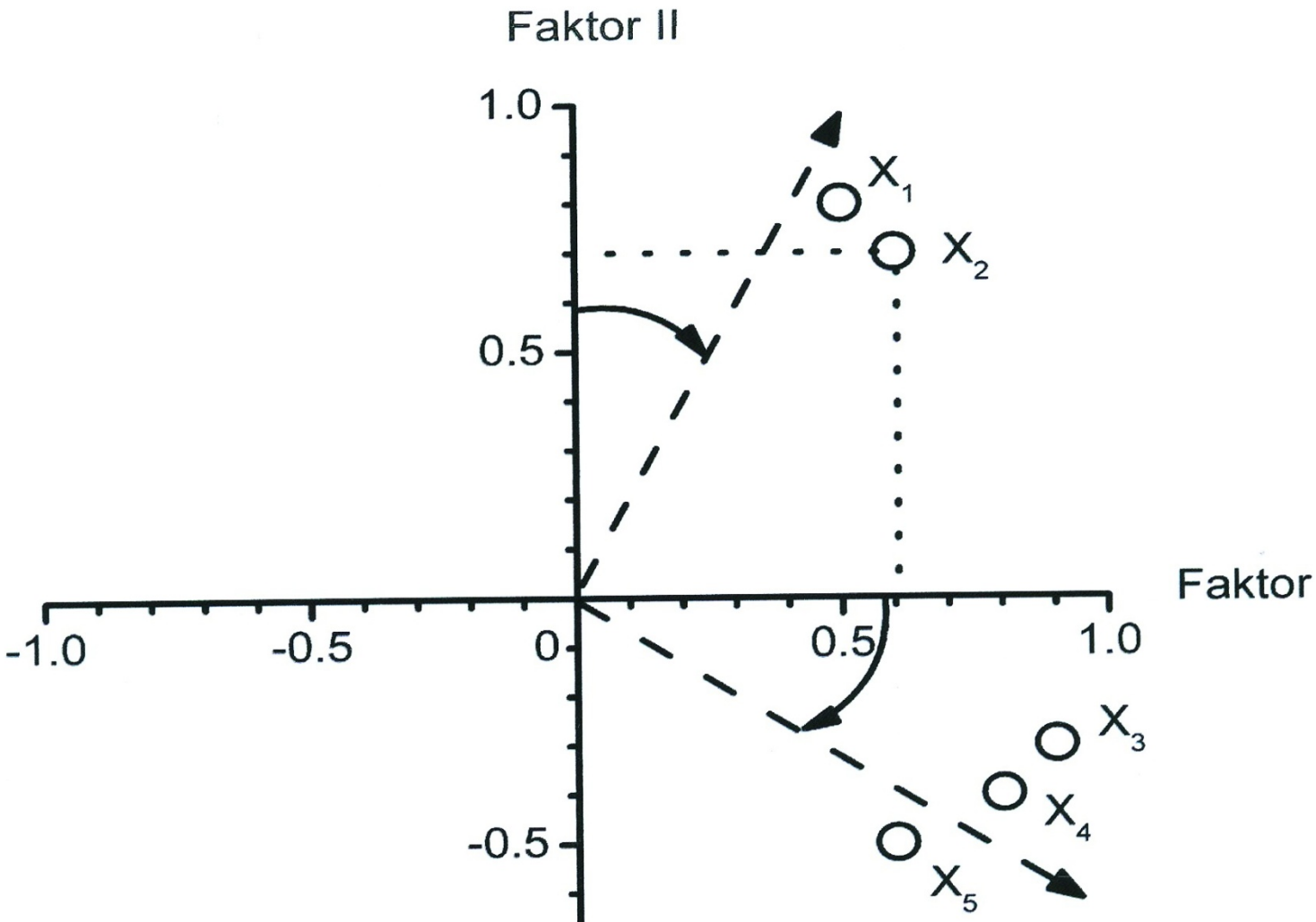
Faktorový model rozdělí:

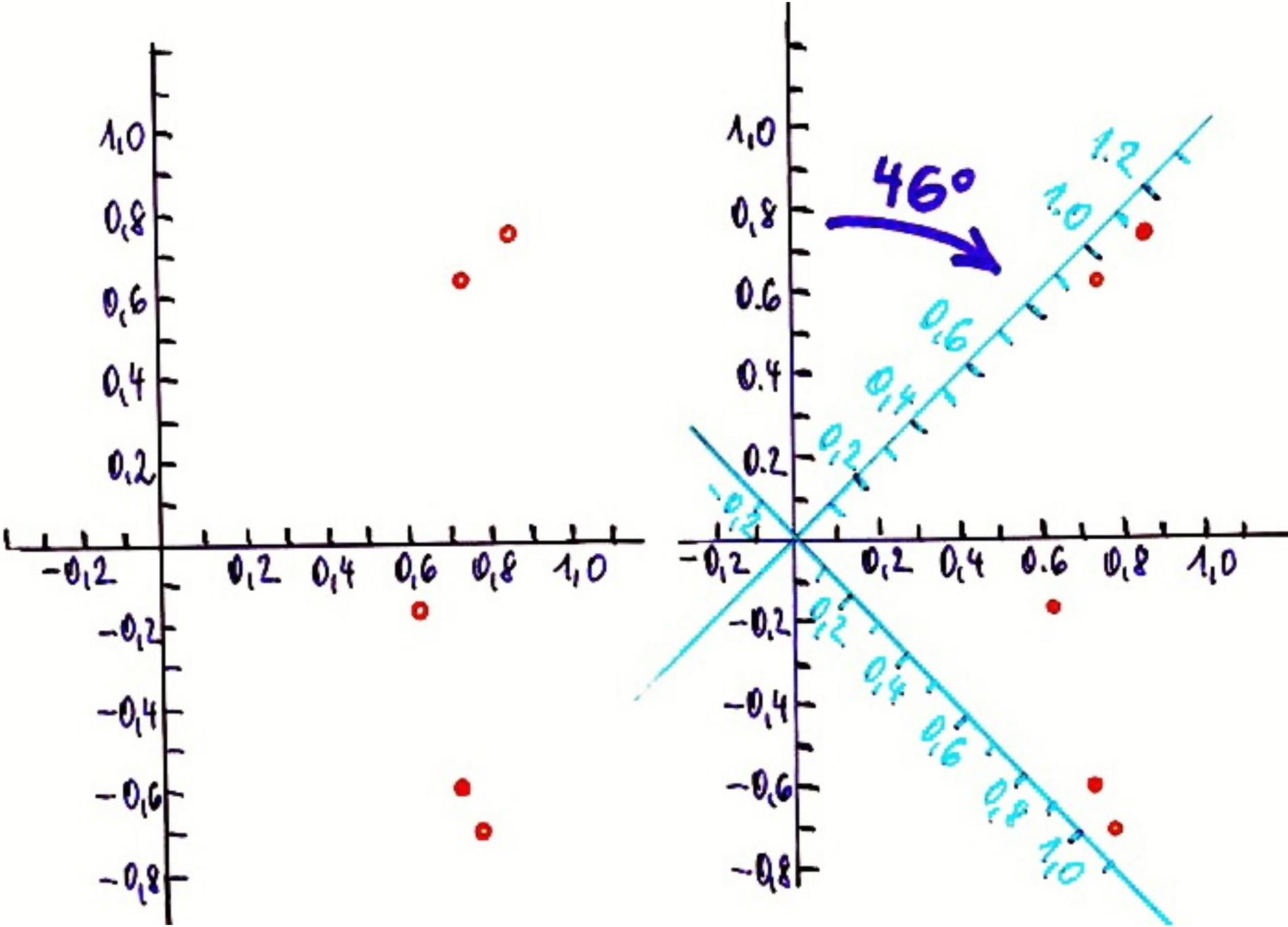
jednak každý znak x_i na dvě části,

jednak rozptyl každého znaku na dvě části.

Otočené faktory: některé zátěže znaků dosahují hodnot blízko ±1 a zbývající zátěže pak téměř nulových.

Proto si budeme přát, aby některý znak dosahoval vysoké zátěže pouze u jediného faktoru čili byl *faktorově čistý*.




1,0
0,8
0,6
0,4
0,2
-0,2
-0,4
-0,6
-0,8
-0,2
0,2
0,4
0,6
0,8
1,0
46°
1.2
1.0
0,6
0,4
0,2
-0,2
0,2
0,4
0,6
0,8
1,0

6.Grafické pomůcky faktorové analýzy

jsou daleko názornější než hodnoty faktorových zátěží v tabulce.

1. Cattelův indexový graf úpatí vlastních čísel

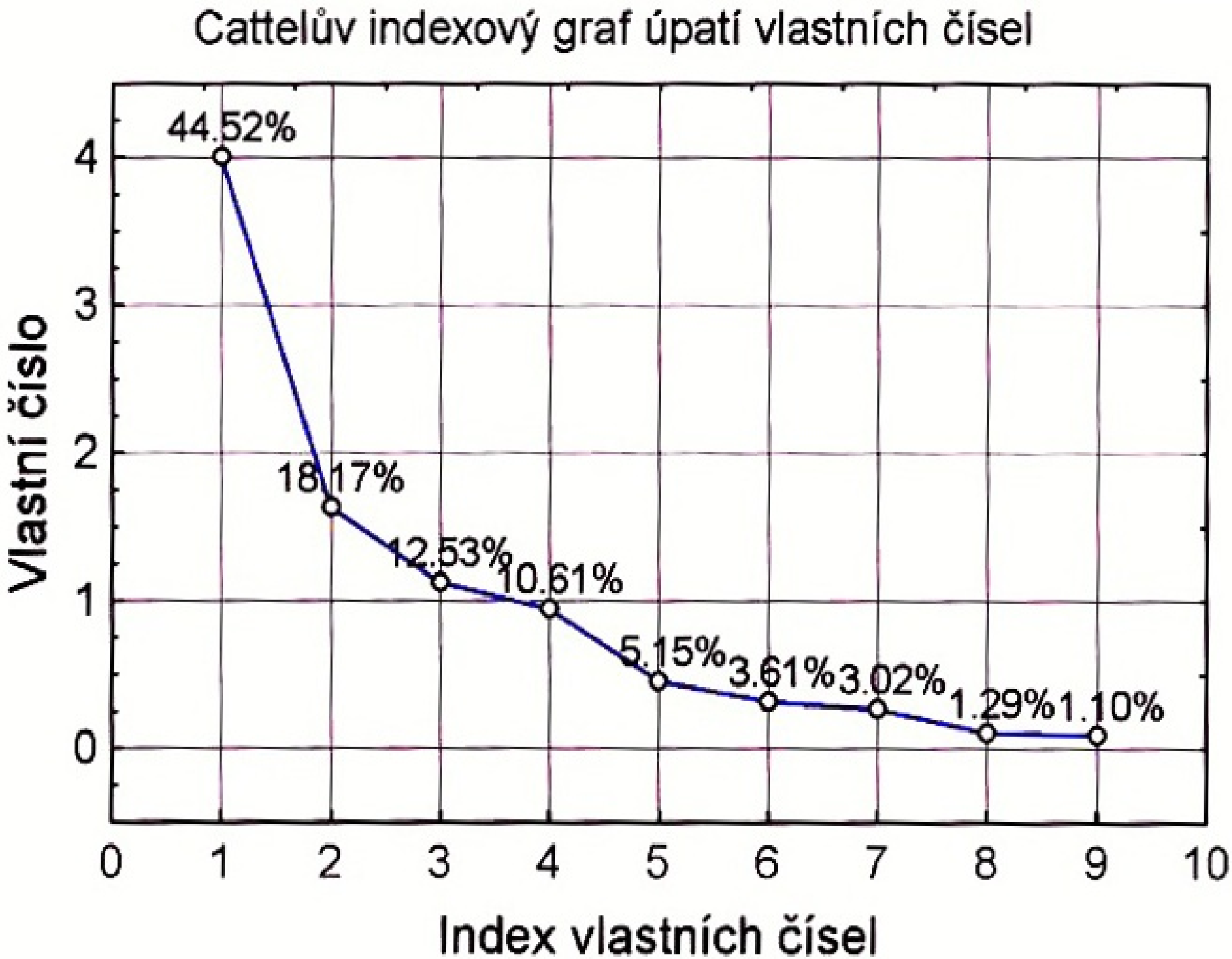
(Scree Plot): Je diagramem vlastních čísel $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ zdrojové matice dat v závislosti na indexu i a zobrazuje relativní velikost jednotlivých vlastních čísel.

Alternativou je graf úpatí pro $\ln \lambda_j$.

Pravidlo „zlomené hole“: pokud je jednotkový úsek náhodně rozdělen na m částí, je očekávaná délka k-té části rovna

$$g_k = \frac{1}{m} \sum_{i=k}^{m} \left(\frac{1}{i} \right) .$$

Zahrnují se pak ty komponenty, pro které platí, že $\lambda_k > g_k$.

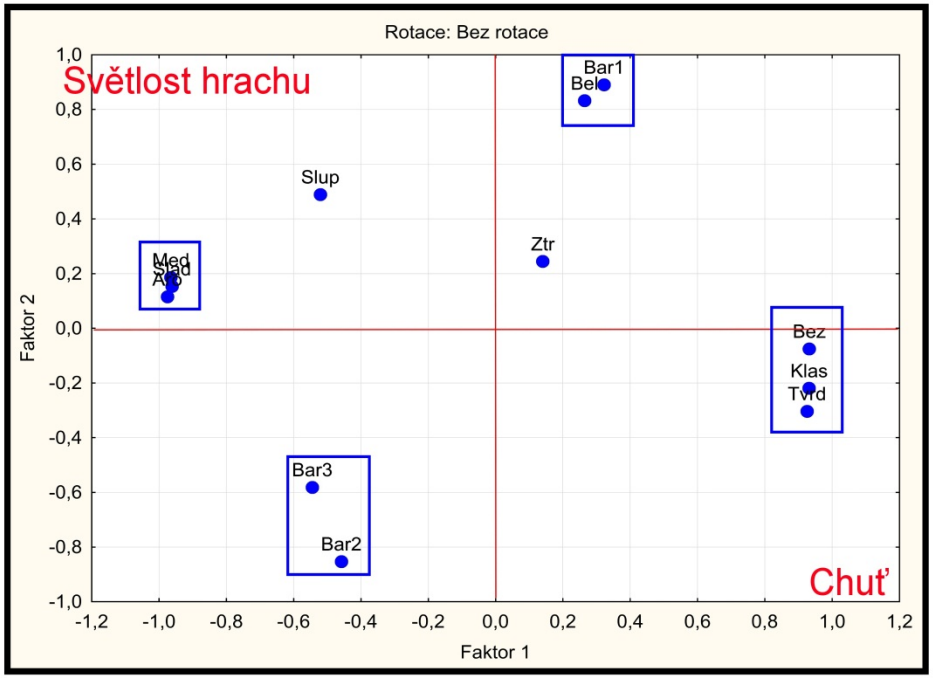

Cattelův indexový graf úpatí vlastních čísel
44.52%
18.17%
12.53%
10.61%
5.15%
3.61%
3.02%
1.29%
1.10%
Vlastní číslo
Index vlastních čísel
0
1
2
3
4
5
6
7
8
9
10

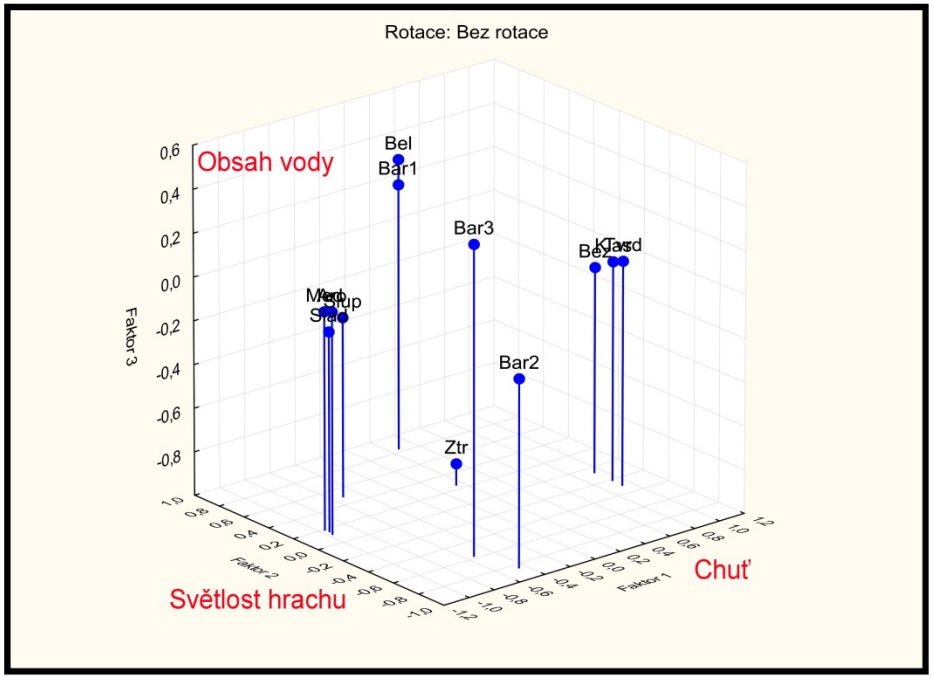
PŘÍKLAD 1.1 ***Popisné statistiky jednorozměrné analýzy zdrojové matice dat*** *Hrách*

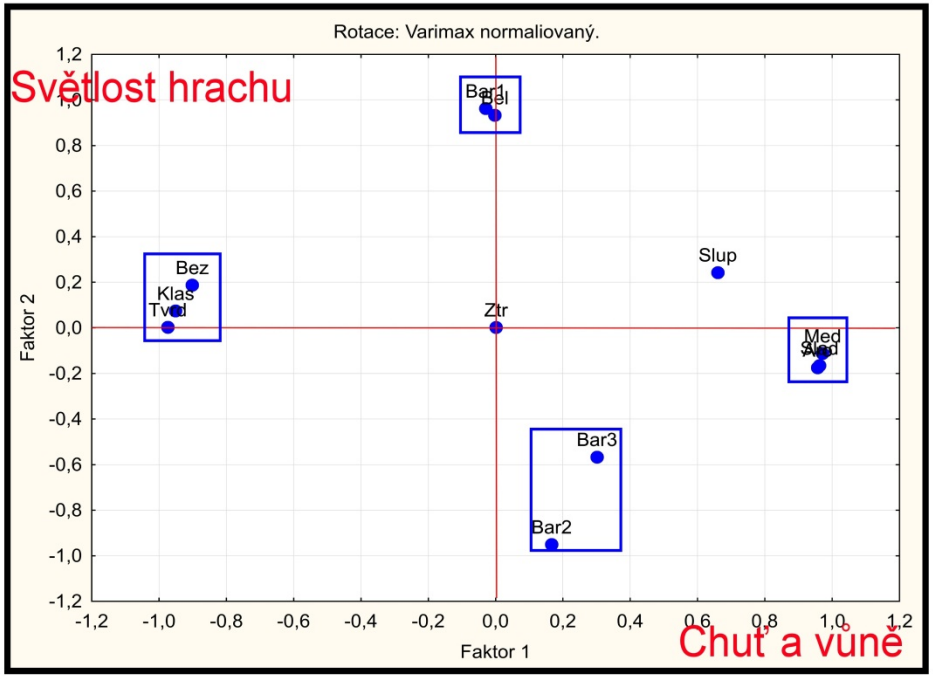
Zdrojová matice dat *Hrách obsahuje znaky smyslového posouzení charakteristik rozličných odrůd hrachu. Objekty zde představují vzorky pěti různých odrůd hrachu A až E, které byly sklízeny v pěti rozličných obdobích 1 až 5. Výsledná zdrojová matice o 12 znacích převážně smyslových charakteristik obsahuje 60 vzorků hrachu. Posouzení každého objektu hrachu bylo provedeno 10 porotci dvojím odhadem tak, že smyslové charakteristiky byly bodovány ve stupnici od 1 (nejhorší) do 9 (nejlepší). Tak bylo získáno 1200 řádků (objektů) postupem: 60 vzorků × 2 hodnocení × 10 porotců. V praxi se data obvykle průměrují, aby se kompenzovaly rozdíly v subjektivní škále přísnosti jednotlivých porotců. Výsledkem je pro každý z šedesáti objektů průměrná hodnota senzorického hodnocení. Cílem úlohy je: 1*. průměrovat data, 2. vynést původní data do grafu a 3. vypočítat popisné jednorozměrné statistiky.

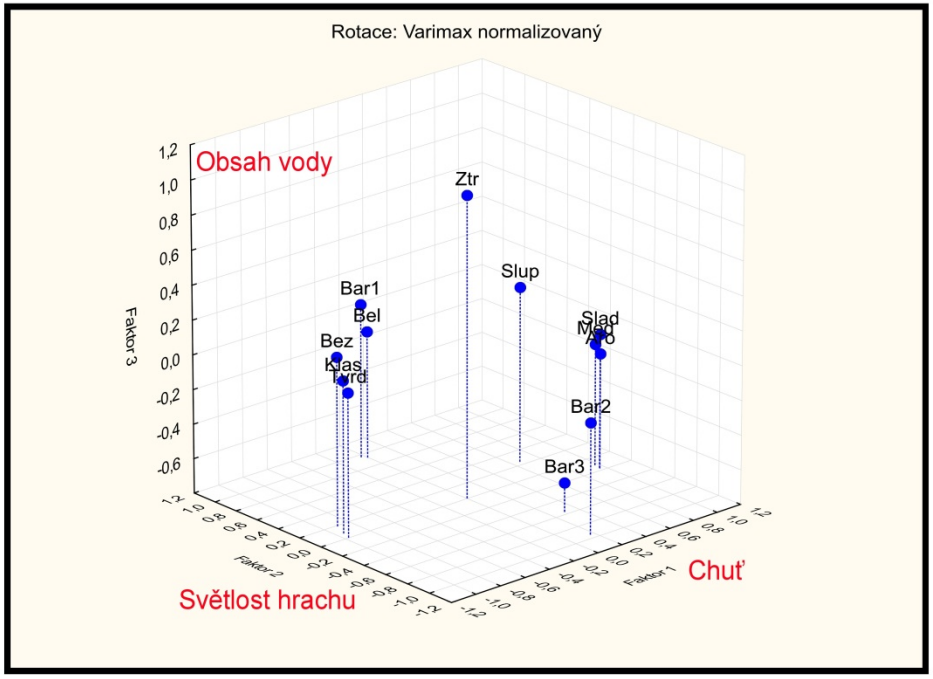
Data: *Zdrojová matice dat n = 1200, m = 12 byla průměrována a výsledkem byla matice 60 × 12. Obsahovala průměrné hodnoty senzorického hodnocení pro znaky ve sloupcích: Aro je aroma, Slad je sladkost, Med je medovost, Bez je bezchuťovost, Klas je klasovost, Tvrd je tvrdost, Bel je bělost, Bar1 je barva1, Bar2 je barva2, Bar3 je barva3, Slup je slupka, Ztr je ztráta.*

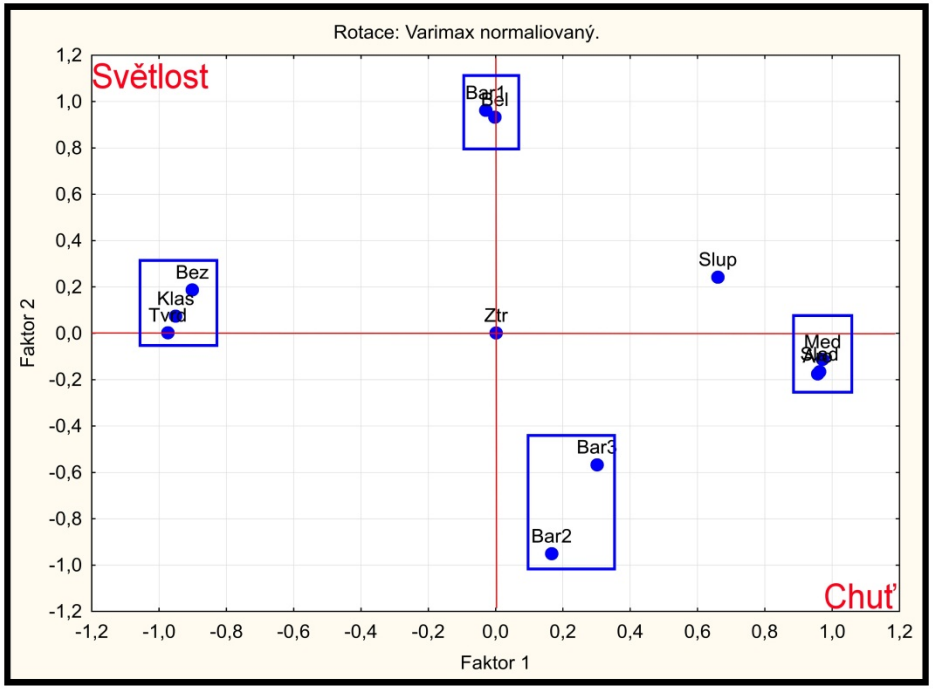
Objekt	*Aro*	*Slad*	*Med*	*Bez*	*Klas*	*Tvrd*	*Bel*	*Bar1*	*Bar2*
	Bar3	*Slup*	*Ztr*						
B5	6.48	6.66	4.56	2.2	2.91	3.47	4.72	5.59	5.73
	5.99	4.26	3.25						
C4	5.75	6.09	3.81	2.32	4.03	3.77	4.17	5.73	5.75
	5.32	3.82	3.38						
B2	3.94	4.12	2.44	3.63	5.77	5.39	4.77	6.67	5.11
	4.60	3.5	3.03						

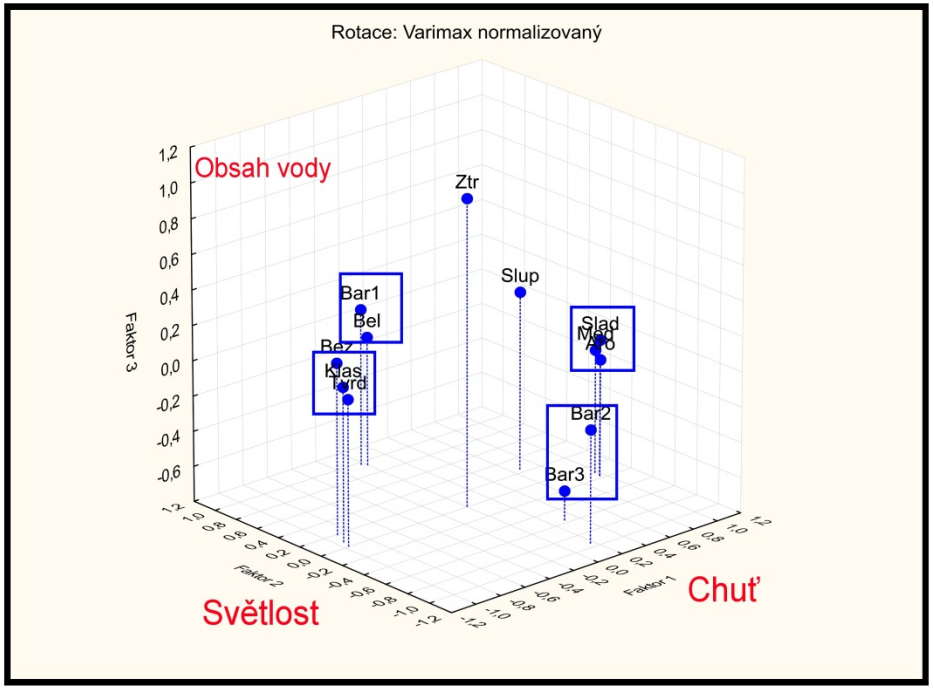

Rotace: Bez rotace
Světlost hrachu
Chuť
Faktor 1
Faktor 2
Bar1
Bel
Slup
Ztr
Med
Slad
Bez
Klas
Tvrd
Bar3
Bar2

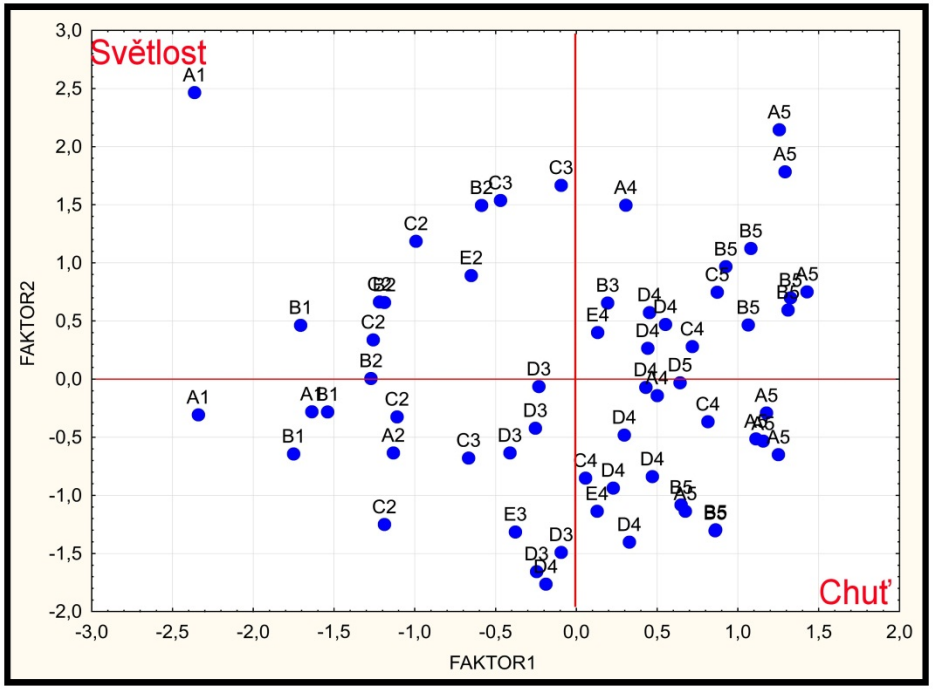

Rotace: Bez rotace
Obsah vody
Světlost hrachu
Chuť
Faktor 1
Faktor 2
Faktor 3
Bel
Bar1
Bar3
Bar2
Ztr

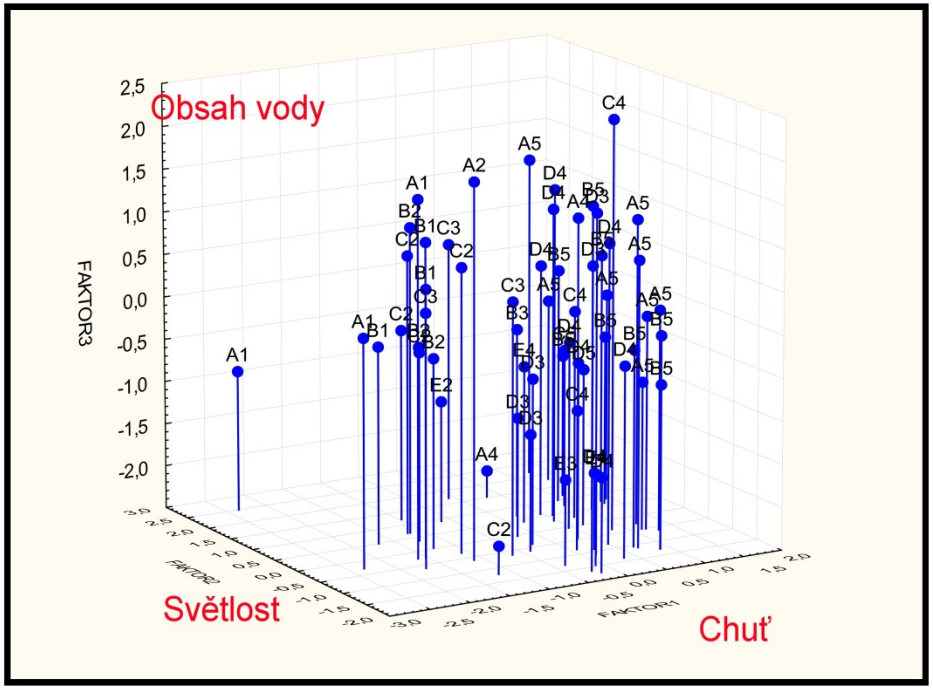

Rotace: Varimax normaliovaný.
Světlost hrachu
Chuť a vůně
Faktor 1
Faktor 2
Bar1
Bel
Bez
Klas
Tvrd
Ztr
Slup
Med
Slad
Bar3
Bar2


Rotace: Varimax normalizovaný
Obsah vody
Světlost hrachu
Chuť
Faktor 1
Faktor 2
Faktor 3
Ztr
Slup
Bar1
Bel
Bez
Klas
Tvrd
Slad
Med
Bar2
Bar3


Rotace: Varimax normaliovaný.
Světlost
Chuť
Faktor 1
Faktor 2
Bar1
Bel
Bez
Klas
Tvrd
Ztr
Slup
Med
Sled
Bar3
Bar2


Rotace: Varimax normalizovaný
Obsah vody
Světlost
Chuť
Faktor 1
Faktor 2
Faktor 3
Ztr
Slup
Bar1
Bel
Bez
Klas
Tvrd
Sled
Med
Bar2
Bar3


Světlost
Chuť
FAKTOR1
FAKTOR2


Obsah vody
Světlost
Chuť
FAKTOR1
FAKTOR2
FAKTOR3

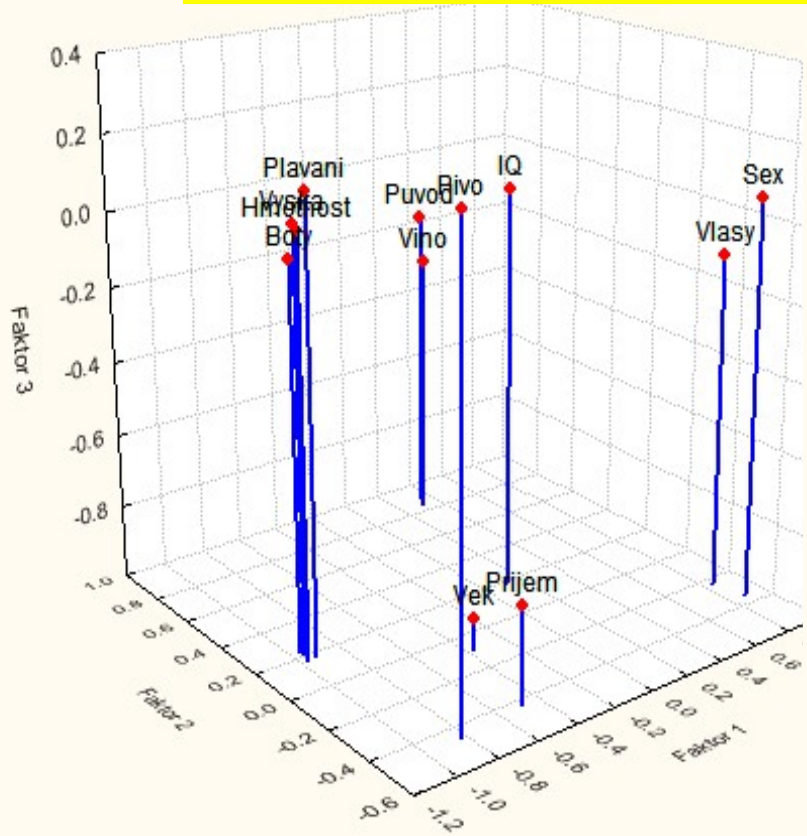
PŘÍKLAD 2.2 ***Průzkumová analýza zdrojové matice dat demografického souboru Lidé***

Vyšetřete grafickými diagnostikami průzkumové analýzy vícerozměrných dat, které ze 12 znaků demografického souboru dat *Lidé jsou nejvýhodnější k charakterizaci osob a které znaky mají největší míru rozptýlení. Matice obsahuje data pro n = 32 osob a m = 12 znaků, kde 16 osob bylo vybráno ze Skandinávie (kód A) a 16 osob ze Středomoří (kód B), 16 osob jsou muži (kód M) a 16 osob jsou ženy (kód F).*

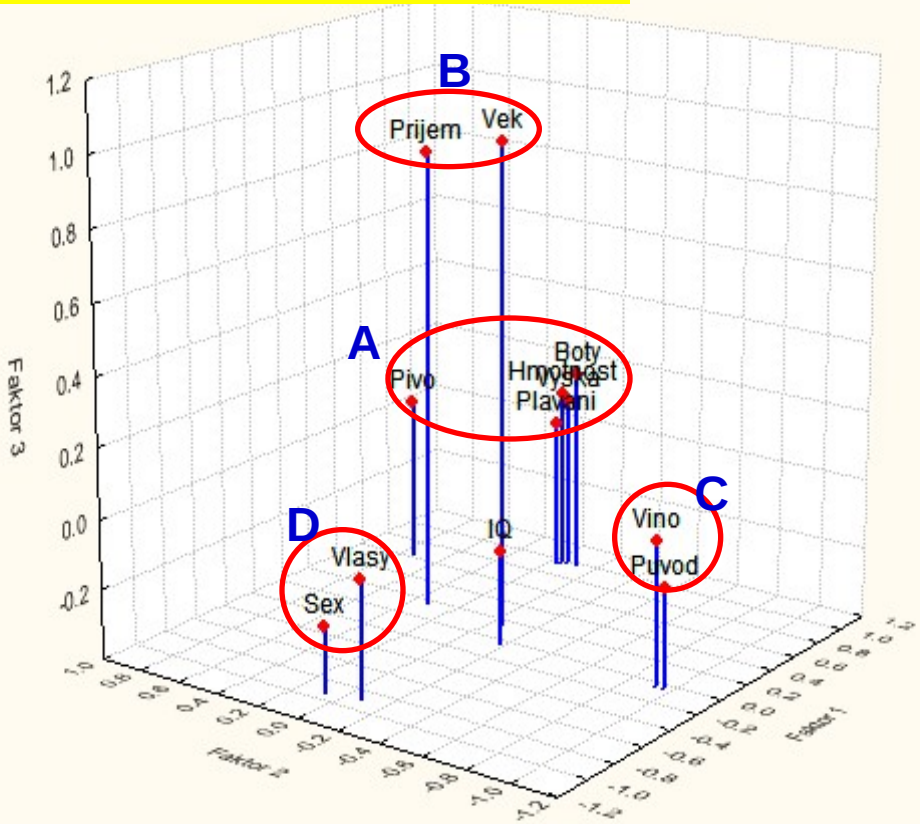
Data: *Znaky obsahují u každé osoby výšku [cm], hmotnost [kg], délku vlasů [krátká: −1, dlouhá: +1], velikost boty [evropský standard], věk [roky], příjem [Euro], spotřeba piva [litry na rok], spotřeba vína [litry na rok], pohlaví [muž: −1, žena: +1], schopnost plavat [naměřený čas na uplavání 500 m], původ [A: −1 Skandinávie, B: +1 Středomoří], inteligenční kvocient IQ [evropský standardizovaný test IQ]. Mezi znaky jsou tři dichotomické, binární proměnné, a to pohlaví, délka vlasů a původ a ostatních 9 znaků nabývá kvantitativních hodnot.*

Osoba	Výška	Hmotnost	Vlasy	Boty	Věk	Příjem	Pivo	Víno	Sex
	Plavání	Původ	IQ						
MA	198	92	−1	48	48	45000	420	115	−1
	98	−1	100						
MA	184	84	−1	44	33	33000	350	102	−1
	92	−1	130						
MA	183	83	−1	44	37	34000	320	98	−1
	91	−1	127						
FA	166	47	−1	36	32	28000	270	78	1
	75	−1	112						

3D- graf faktorových vah

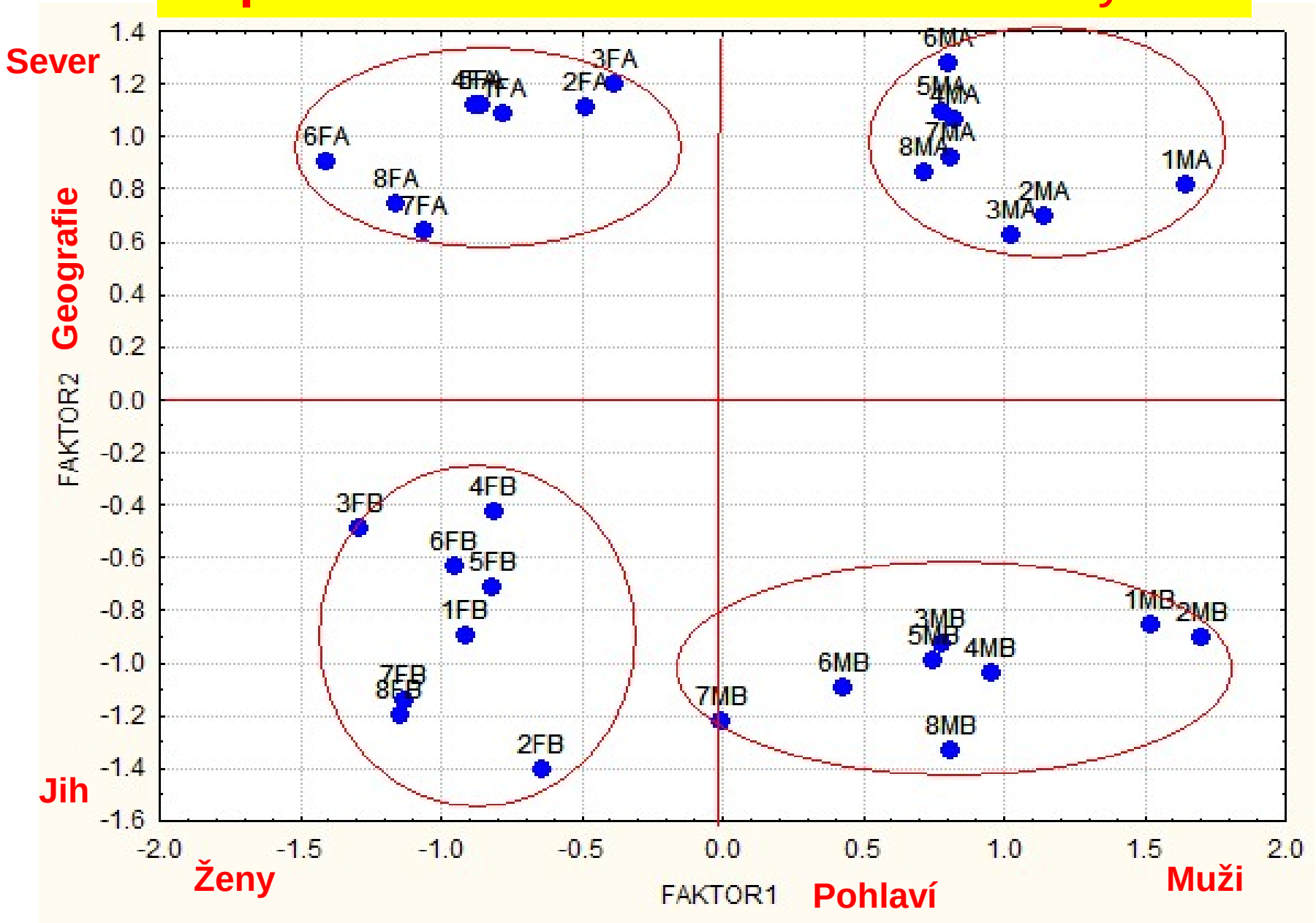


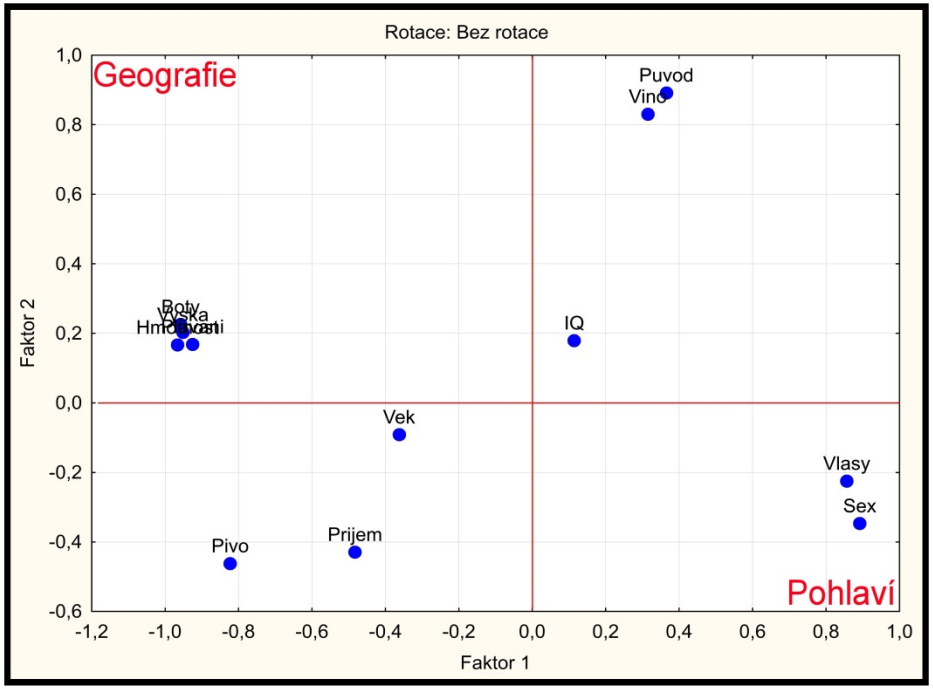
Faktorová analýza:
Graf faktorových vah **bez rotace**

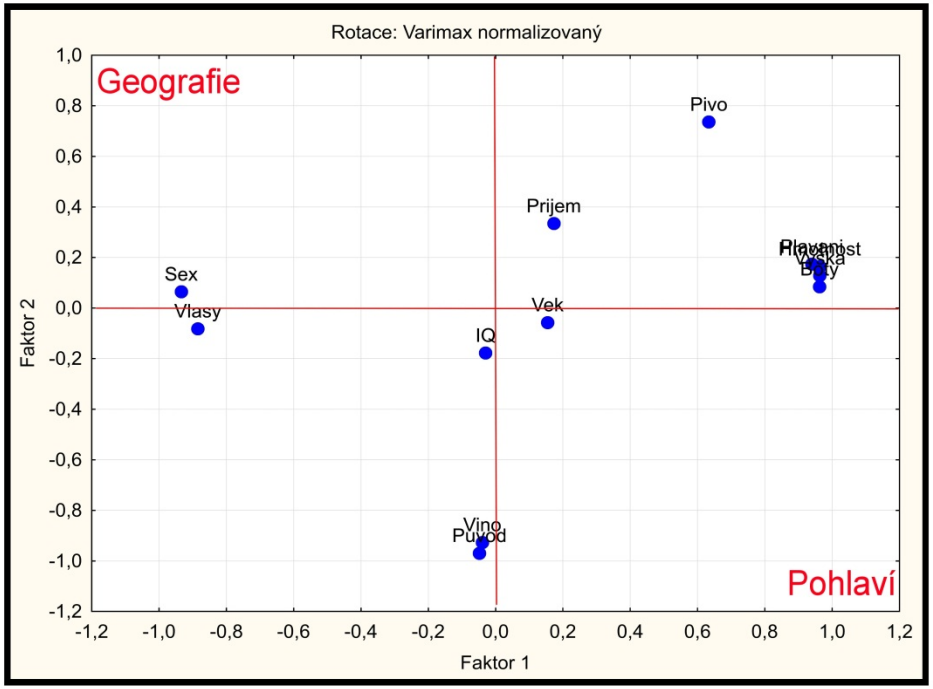


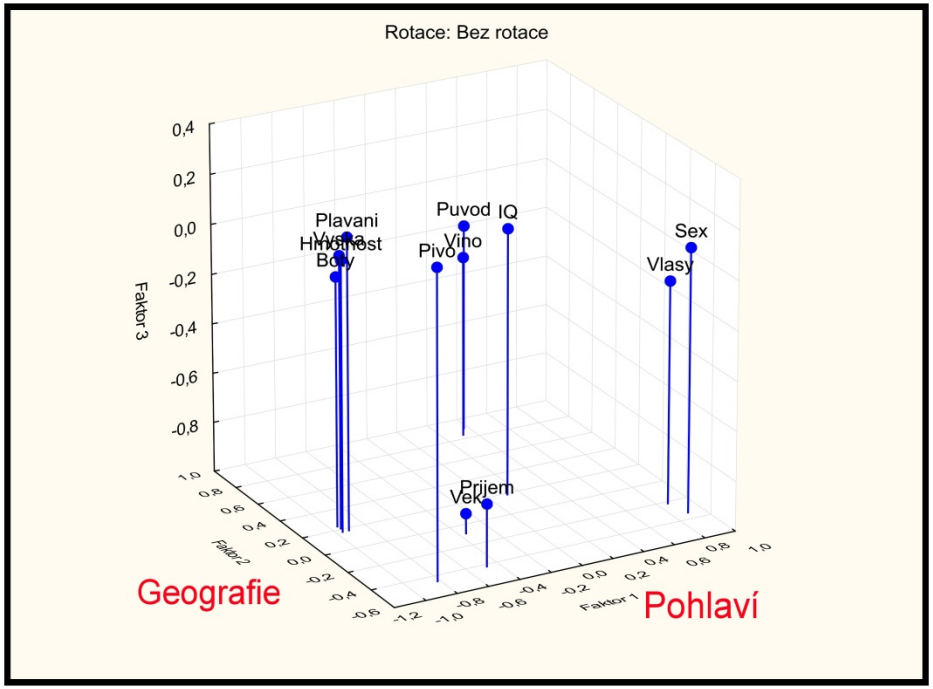
Faktorová analýza:
Graf faktorových vah **po rotaci Varimaxem normalizovaným**

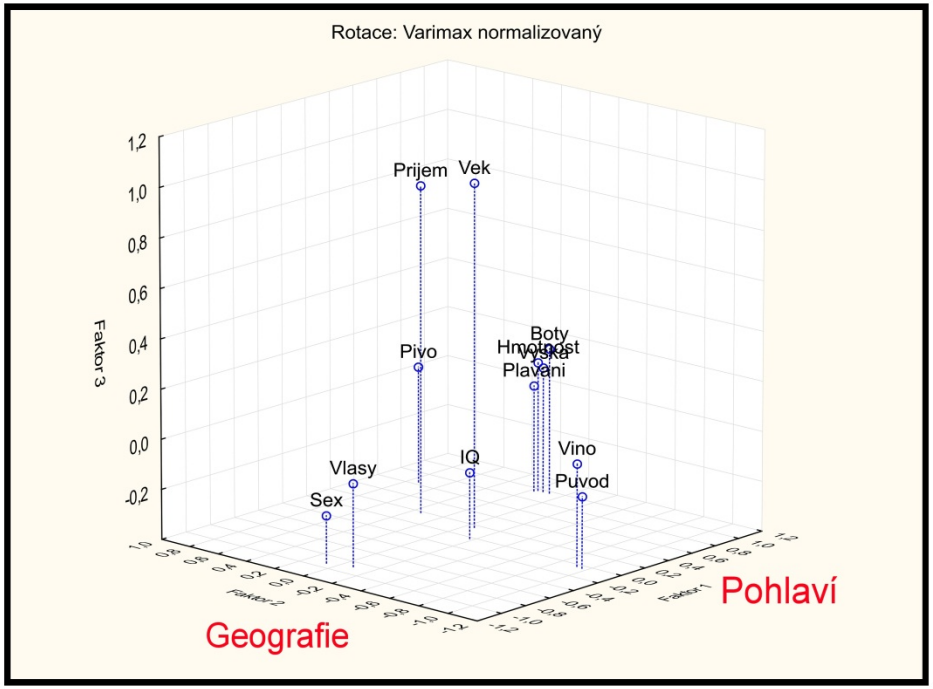
Rozptylový diagram faktorového skóre **po rotaci** Varimaxem normalizovaným

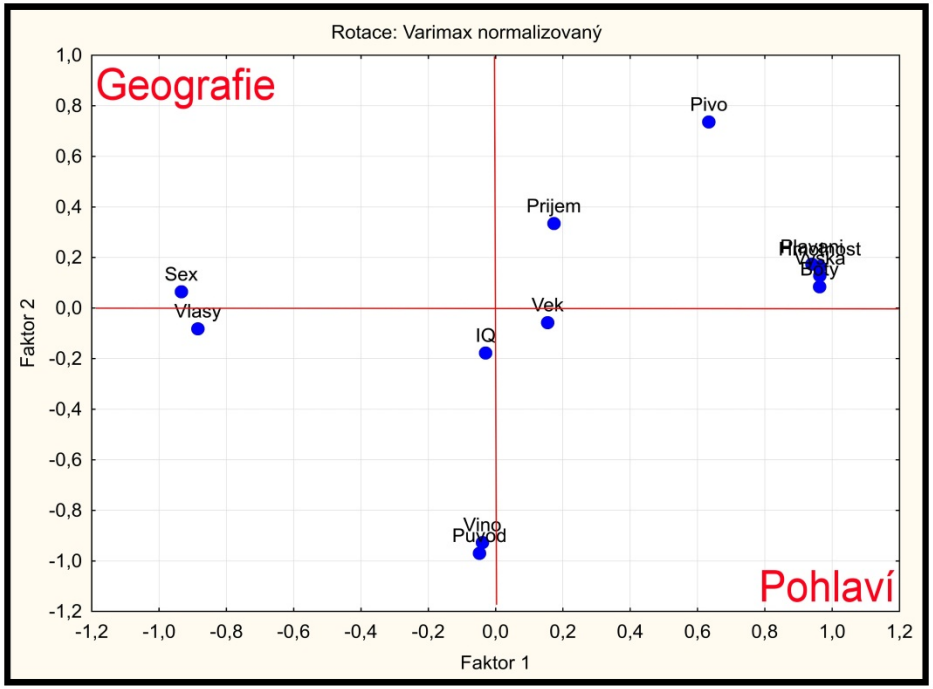


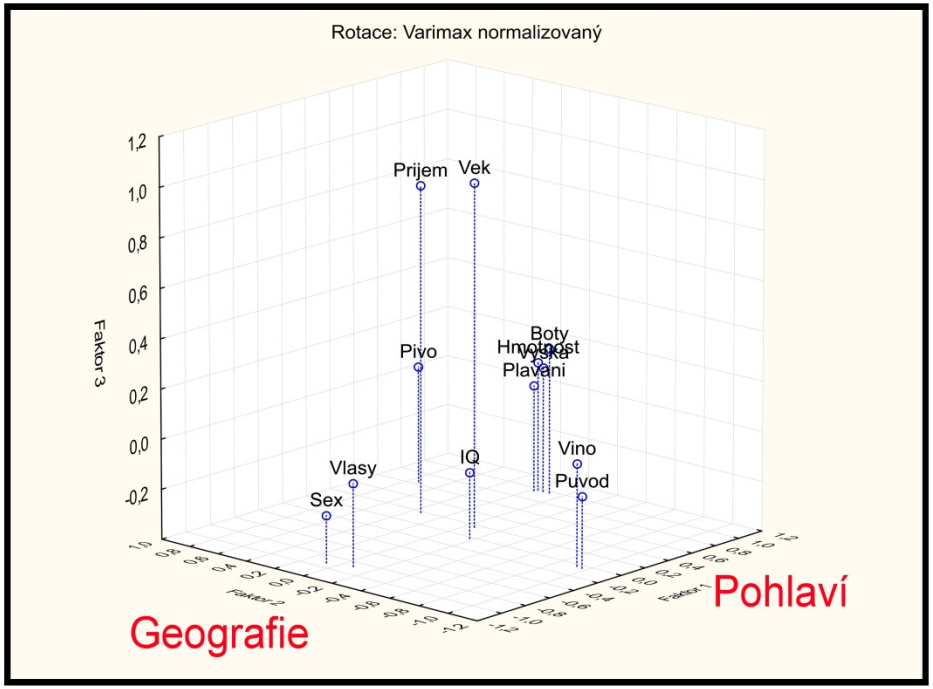

Rotace: Bez rotace
Geografie
Pohlaví
Faktor 1
Faktor 2
Puvod
Vino
IQ
Vek
Vlasy
Sex
Pivo
Prijem
Boty
Vyska
Hmotnost
Plavani

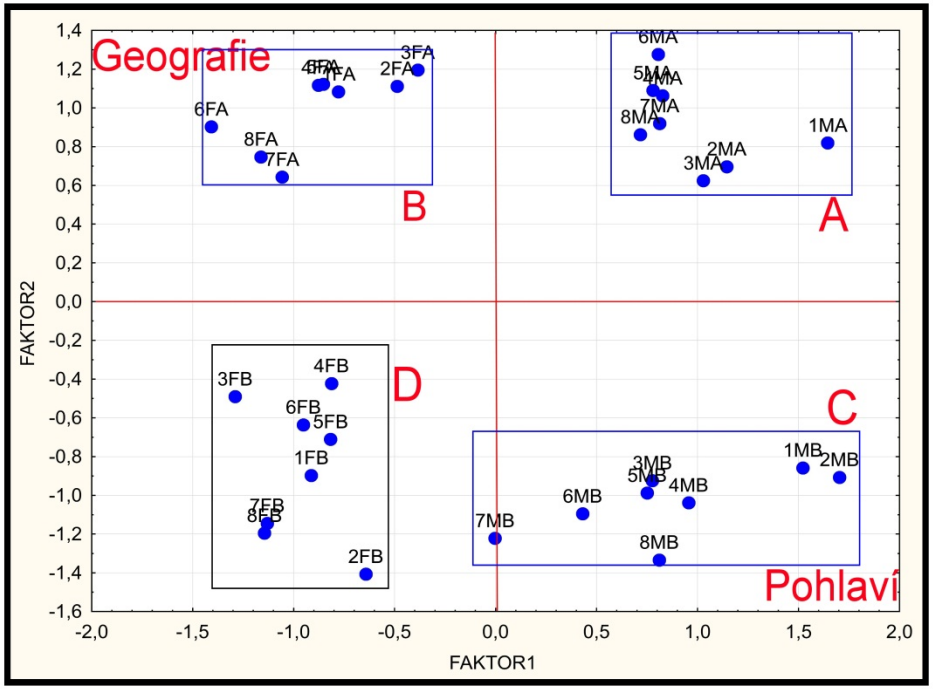

Rotace: Varimax normalizovaný
Geografie
Pohlaví
Faktor 1
Faktor 2
Pivo
Prijem
Sex
Vlasy
Vek
IQ
Vino
Puvod
Plavani
Hmotnost
Vyska
Boty

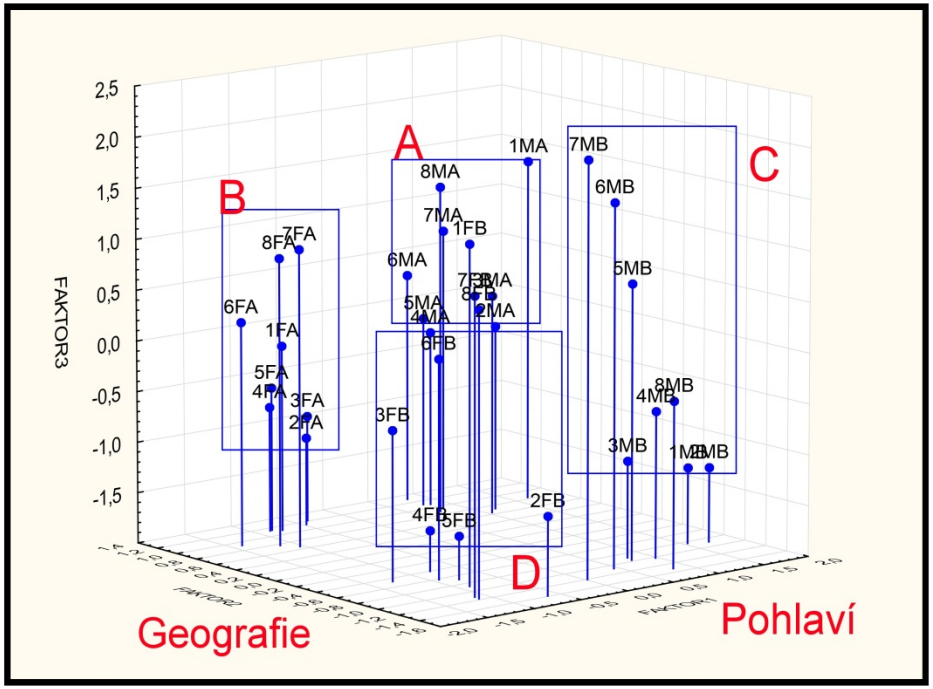

Rotace: Bez rotace
Geografie
Pohlaví
Faktor 1
Faktor 2
Faktor 3
Plavani
Vyska
Hmotnost
Boty
Puvod
IQ
Pivo
Vino
Sex
Vlasy
Prijem
Vek


Rotace: Varimax normalizovaný
Geografie
Pohlaví
Faktor 1
Faktor 2
Faktor 3
Prijem
Vek
Pivo
Boty
Hmotnost
Vyska
Plavani
Vino
Puvod
IQ
Vlasy
Sex


Rotace: Varimax normalizovaný
Geografie
Pohlaví
Pivo
Prijem
Plavani
Hmotnost
Vyska
Boty
Sex
Vlasy
Vek
IQ
Vino
Puvod
Faktor 1
Faktor 2


Rotace: Varimax normalizovaný
Prijem
Vek
Pivo
Boty
Hmotnost
Vyska
Plavani
Vino
Puvod
IQ
Vlasy
Sex
Faktor 3
Pohlaví
Geografie


Geografie
Pohlaví
A
B
C
D
3FA
2FA
6FA
8FA
7FA
6MA
5MA
7MA
8MA
1MA
2MA
3MA
3FB
4FB
6FB
5FB
1FB
7FB
8FB
2FB
1MB
2MB
3MB
5MB
4MB
6MB
7MB
8MB
FAKTOR1
FAKTOR2

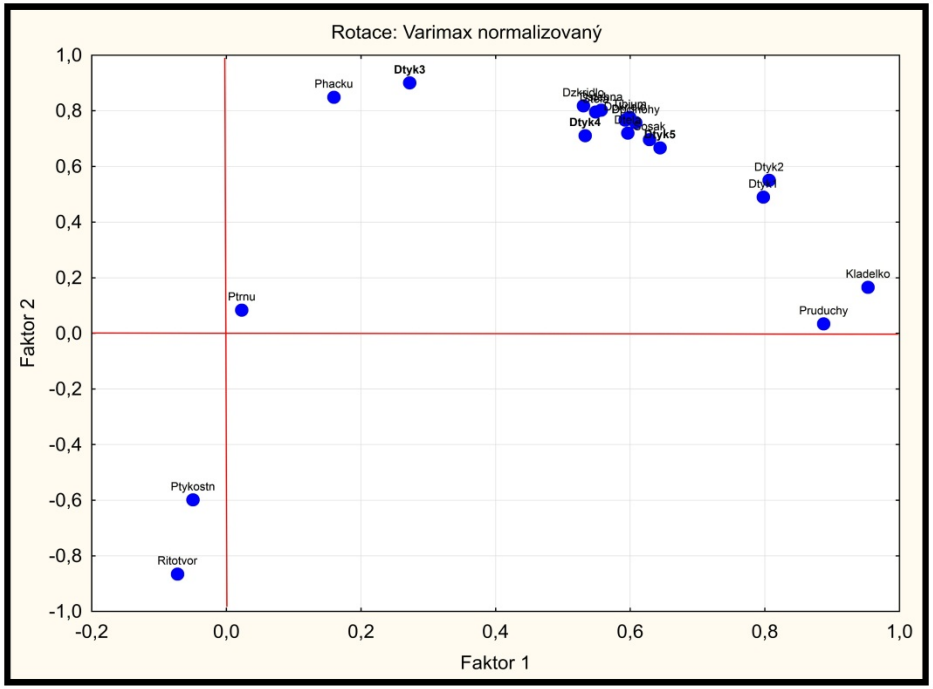

A
B
C
D
1MA
8MA
7MA
1FB
6MA
5MA
4MA
2MA
6FB
7FB
8FB
3FB
2FB
4FB
5FB
7FA
8FA
6FA
1FA
5FA
4FA
3FA
2FA
7MB
6MB
5MB
8MB
4MB
3MB
1MB
2MB
FAKTOR3
Geografie
Pohlaví

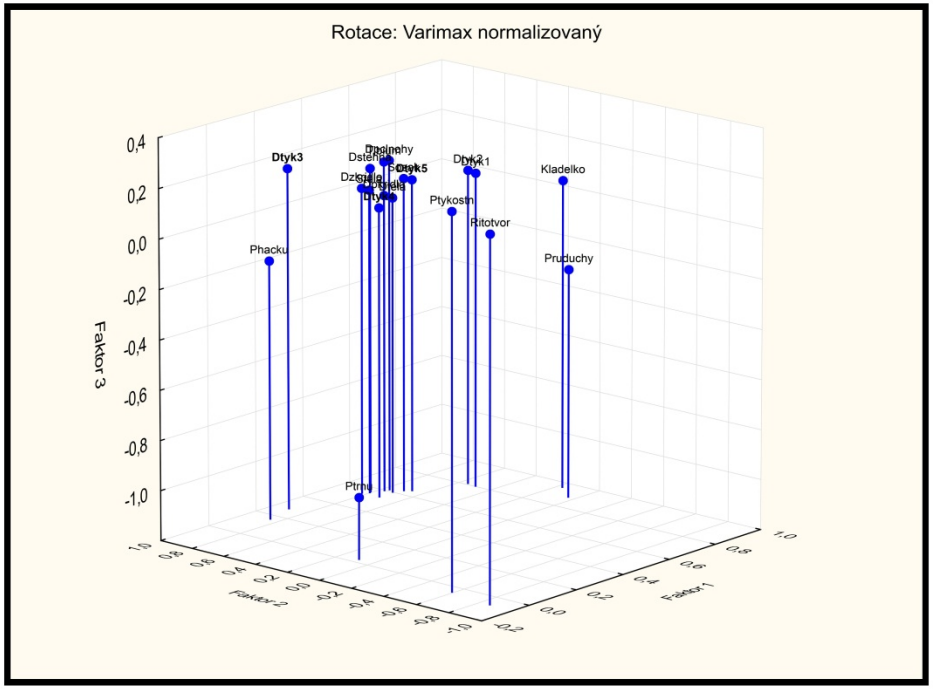
PŘÍKLAD 4.6 *Popis a třídění polétavých mšic*

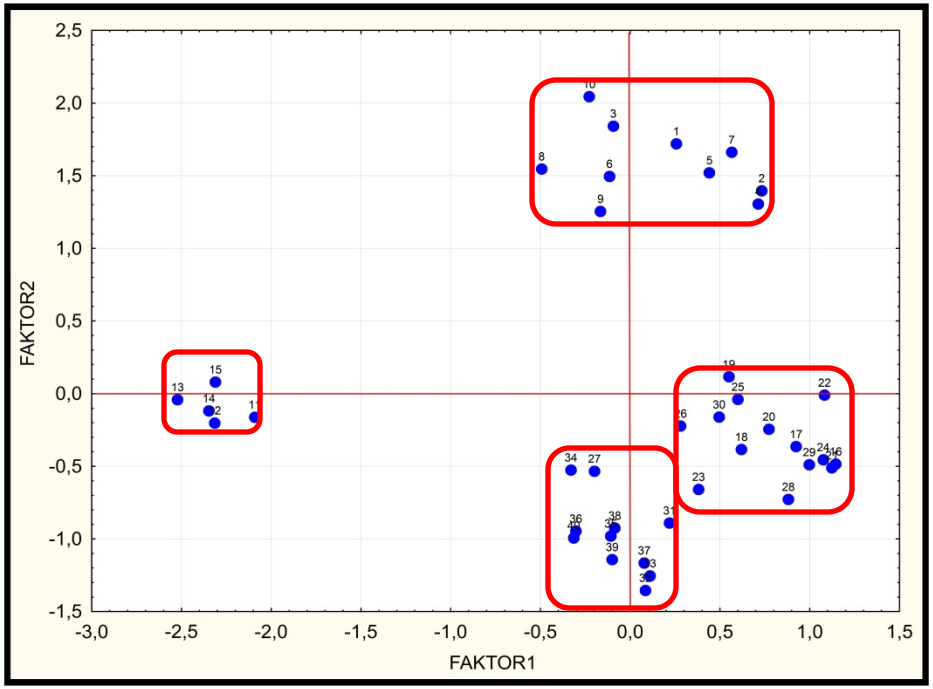
Jeffers (1967) studoval 40 jedinců polétavých mšic *(Alate adelges) za pomoci světelné pasti. U mšic bylo změřeno 19 znaků k rozlišení druhů tohoto hmyzu souboru Mšice: 14 znaků se týká délky nebo šířky, 4 znaky se týkají počtu a 1 znak je binární vyjadřující přítomnost či absenci. Mšice se totiž obtížně rozlišují dle běžných taxonometrických klíčů a obvykle je nutné detailní vyšetření dat, hledání společných rysů a následná klasifikace do větších celků. Užitím metody hlavních komponent PCA je třeba snížit původní počet 19 znaků na menší počet proměnných, který by vystihl co největší množství proměnlivosti mšic. Před užitím PCA je potřeba provést standardizaci dat, protože znaky představují směs délek a počtů.*

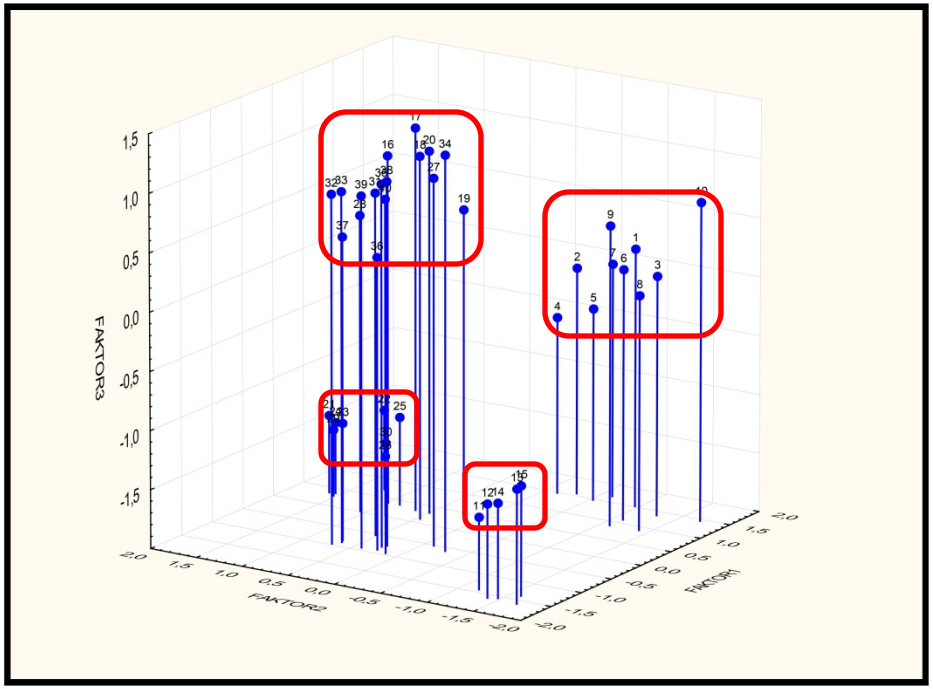
Data: *Charakter znaků − i index jedince, x_1 značí délku těla, x_2 značí šířku těla, x_3 je délka předního křídla, x_4 je délka zadního křídla, x_5 je počet průduchů, x_6 je délka tykadla I, x_7 je délka tykadla II, x_8 je délka tykadla III, x_9 je délka tykadla IV, x_{10} je délka tykadla V, x_{11} je počet tykadlových ostnů, x_{12} je délka posledního článku nohy, x_{13} je délka holeně, tibia, x_{14} je délka stehna, x_{15} je délka sosáku, x_{16} je délka kladélka, x_{17} je počet kladélkových trnů, x_{18} je řitní otvor, x_{19} je počet háčků zadních křídel.*

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
	x_{19}								
1	21.2	11.0	7.5	4.8	5.0	2.0	2.0	2.8	2.8
	3.3	3.0	4.4	4.5	3.6	7.0	4.0	8.0	0.0
	3.0								
2	20.2	10.0	7.5	5.0	5.0	2.3	2.1	3.0	3.0
	3.2	5.0	4.2	4.5	3.5	7.6	4.2	8.0	0.0
	3.0								
3	20.2	10.0	7.0	4.6	5.0	1.9	2.1	3.0	2.5
	3.3	1.0	4.2	4.4	3.3	7.0	4.0	6.0	0.0
	3.0								


Rotace: Varimax normalizovaný
Faktor 2
Faktor 1
Phacku
Dtyk3
Dtyk4
Dtyk5
Dtyk2
Kladelko
Ptrnu
Pruduchy
Ptykostn
Ritotvor


Rotace: Varimax normalizovaný
Faktor 3
Faktor 2
Faktor 1
Dtyk3
Phacku
Kladelko
Pruduchy
Ptykostn
Ritotvor
Ptrnu


FAKTOR2
FAKTOR1


FAKTOR3
FAKTOR2
FAKTOR1

Postup analýzy shluků

Poskytuje empirické a objektivní metody
ke klasifikaci objektů

1. krok: Cíle analýzy shluků

2. krok: Formulace úlohy analýzy shluků

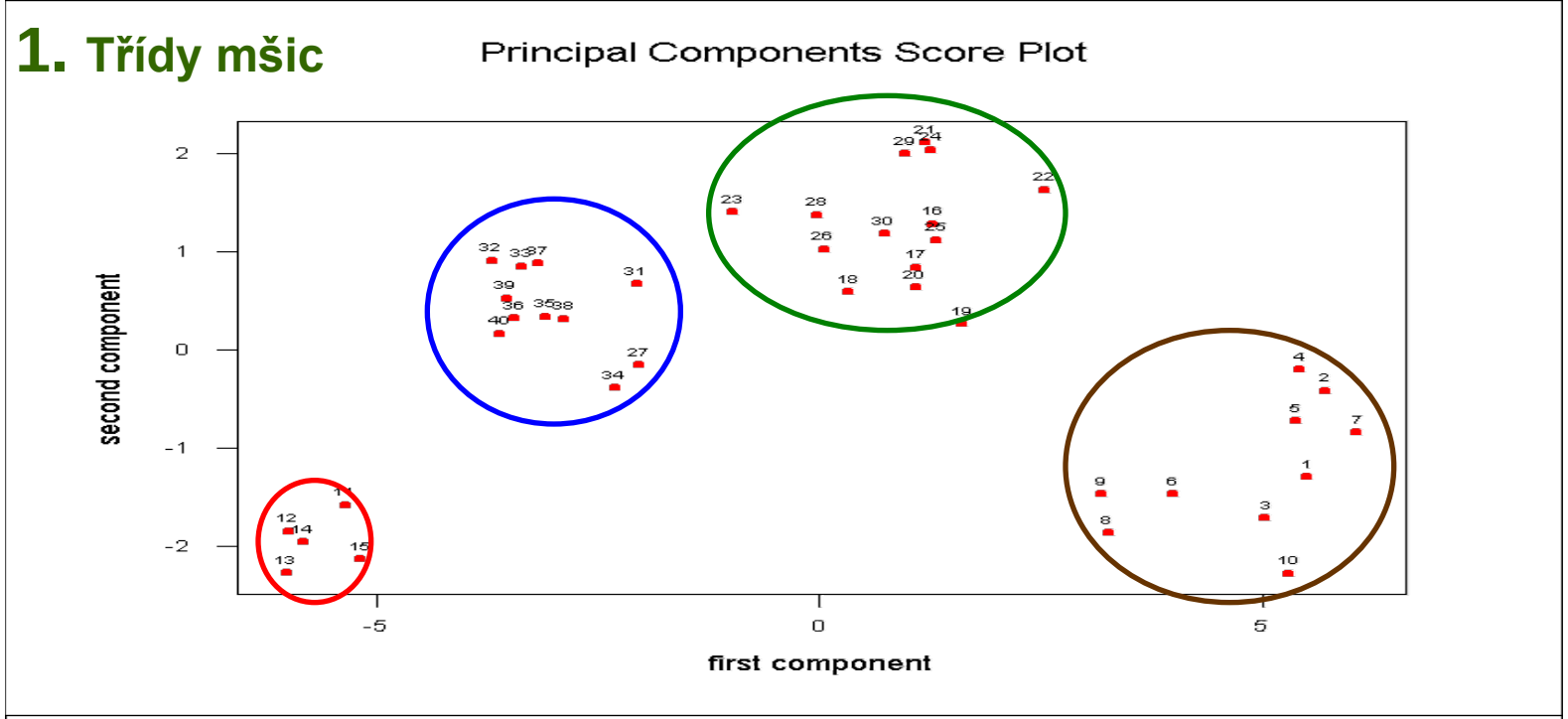
3. krok: Předpoklady úlohy

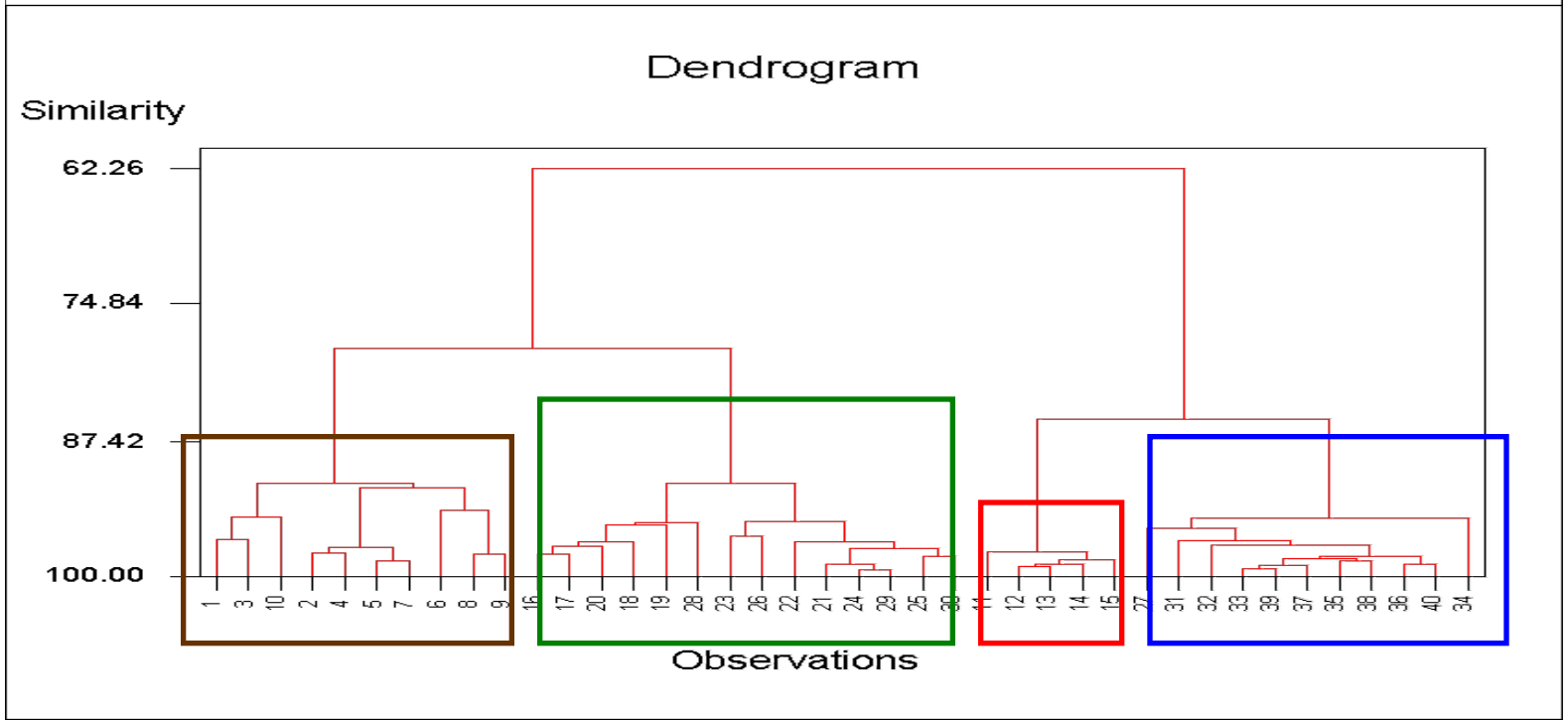
4. krok: Výstavba dendrogramu shluků

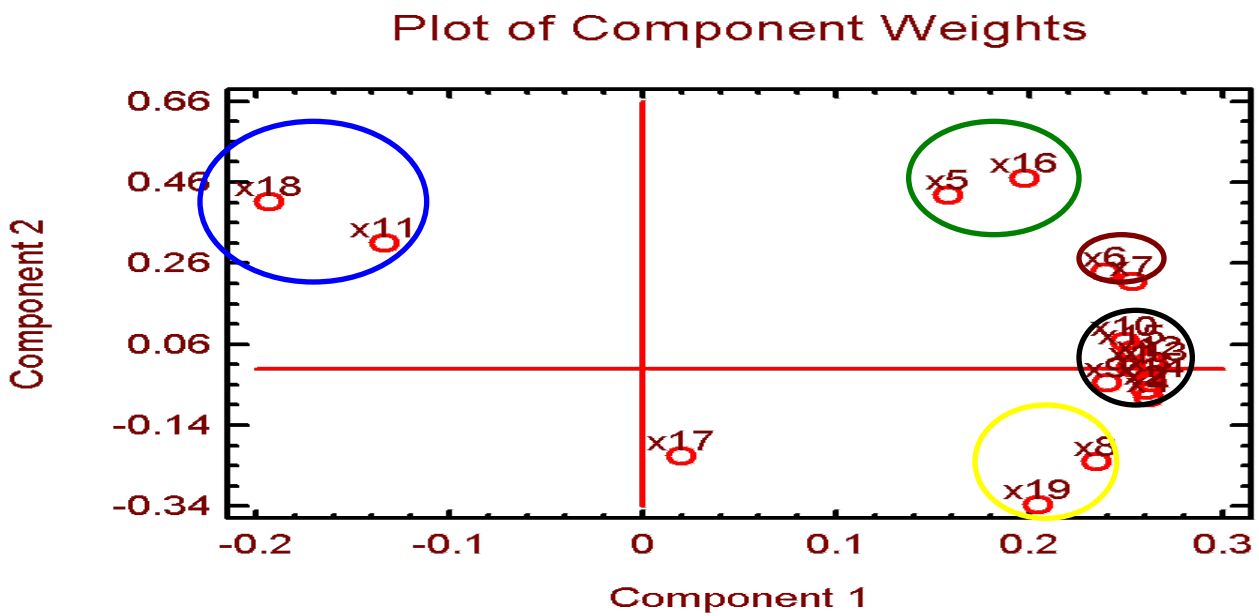
5. krok: Interpretace shluků

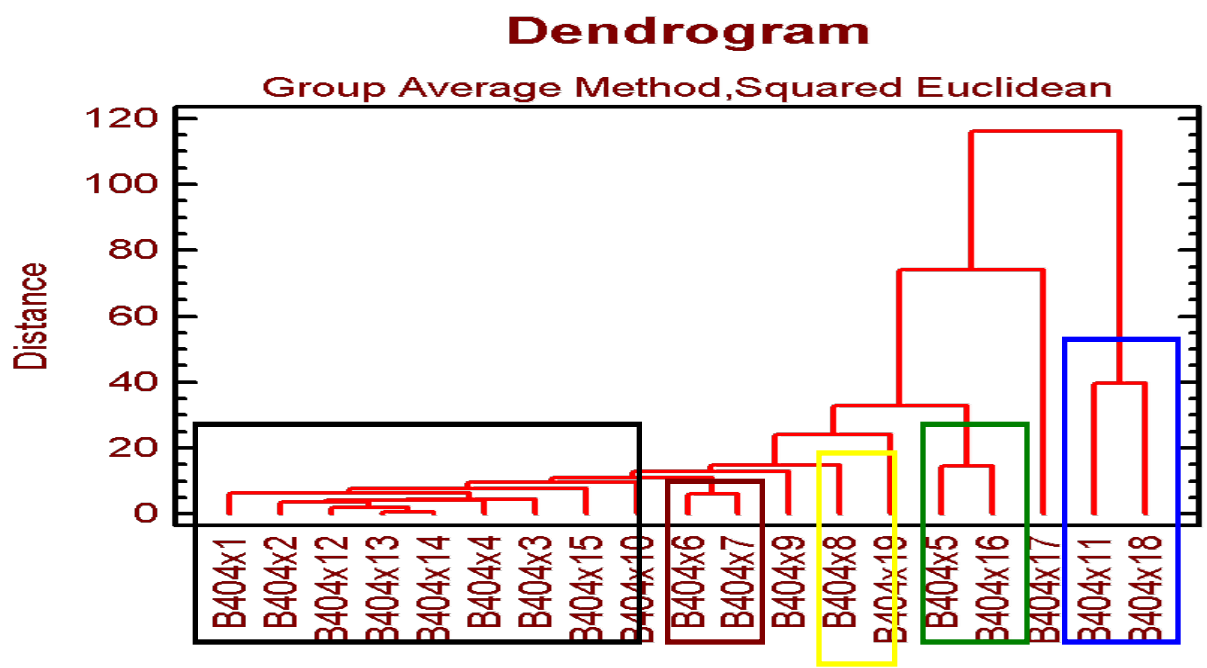
6. krok: Validace a profilování shluků

Úloha 1. Třídy mšic





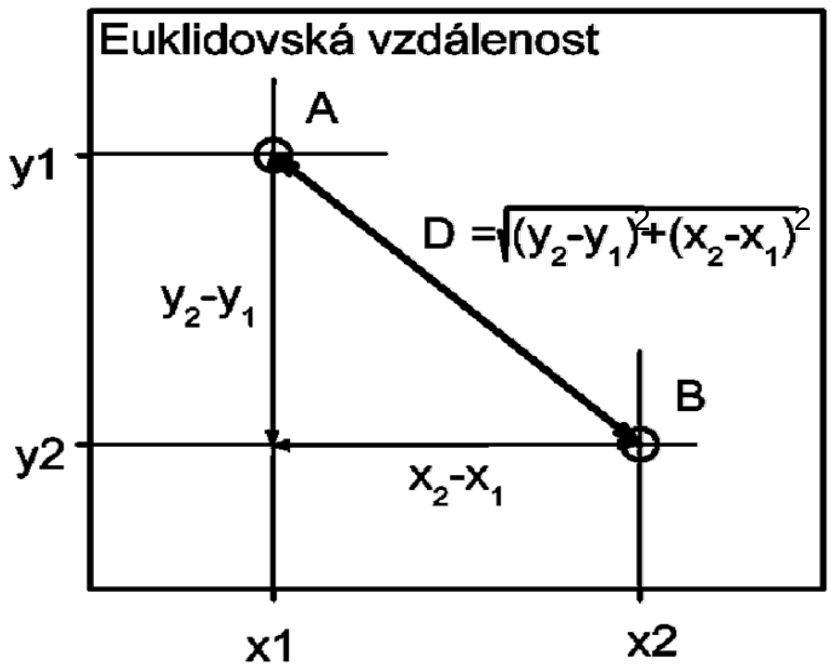

Plot of Component Weights
Component 2
Component 1
0.66
0.46
0.26
0.06
-0.14
-0.34
-0.2
-0.1
0
0.1
0.2
0.3
x18
x11
x5
x16
x6
x7
x17
x8
x19

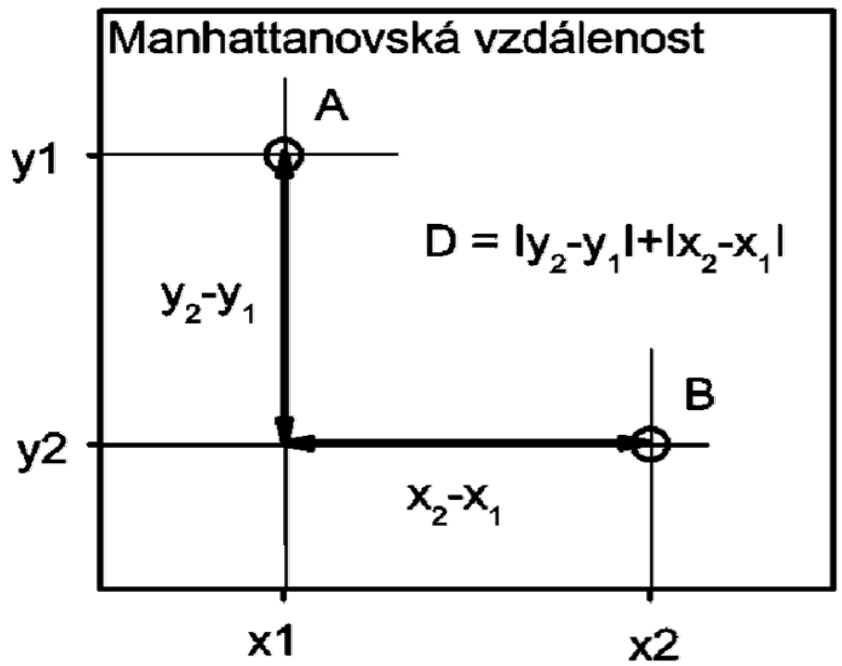

Dendrogram
Group Average Method,Squared Euclidean
Distance
120
100
80
60
40
20
0
B404x1
B404x2
B404x12
B404x13
B404x14
B404x4
B404x3
B404x15
B404x10
B404x6
B404x7
B404x9
B404x8
B404x19
B404x5
B404x16
B404x17
B404x11
B404x18

2.3 Míry podobnosti

Podobnost je měřena rozličnými způsoby

Míry vzdálenosti: nejčastěji užívané míry podobnosti. Vzdálenost je reciproká hodnota podobnosti. Čím větší hodnota vzdálenosti, tím menší podobnost.





Eukleidovská vzdálenost zvaná také *geometrická metrika*

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2}$$

Manhattanská vzdálenost zvaná také *vzdálenost městských bloků* nebo *Hammingova metrika* je definovaná

$$d_H(x_k, x_l) = \sum_{j=1}^{m} |x_{kj} - x_{lj}|$$

Zobecněná Minkovského metrika

$$\sqrt{}$$

kde pro $z = 1$ jde o Hammingovu metriku a pro $z = 2$ o Eukleidovu. Čím je z větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty.

Tětivová vzdálenost (anglicky chord distance) je definovaná

$$\sqrt{\left[\right]}$$

Mahalanobisova metrika pro silně korelované znaky x_k a x_l

$$\sqrt{}$$

vyjadřuje vzdálenost bodů v prostoru, jehož osy nemusí být orthogonální.

Aglomerační způsoby (algoritmy) výstavby dendrogramu shluků:

Metoda nejbližšího souseda: je postavena na minimální vzdálenosti objektů.

Metoda nejvzdálenějšího souseda: je postavena nikoliv na minimální ale na maximální vzdálenosti.

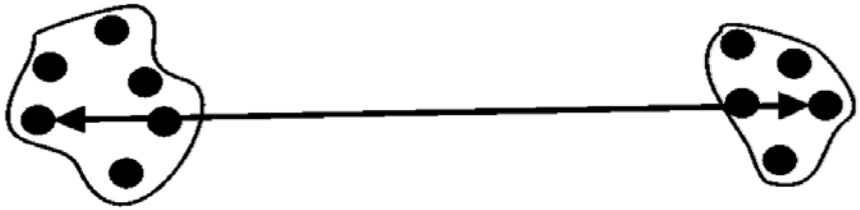
Metoda průměrového linkování: kritériem je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

Wardova metoda: vzdálenost mezi dvěma shluky je tvořena na základě sumy čtverců přes všechny proměnné mezi dvěma shluky.

Metoda těžiště: vzdálenost těžišť shluků spojených Euklidovskou vzdáleností nebo čtvercem Euklidovské vzdálenosti. Těžiště shluku je průměrná hodnota objektů v proměnných, vyjádřená ve shlukových proměnných.

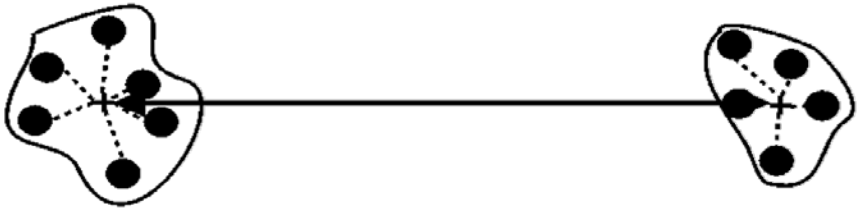
Metoda nejblizšího souseda

Metoda nevzdálenějšího souseda



Metoda průměrné vzdálenosti

Metoda Wardova



Míra věrohodnosti "nejlepšího dendrogramu"

1. kritérium těsnost proložení:

kofenetický korelační koeficient CC

- nejlépe odpovídá struktuře objektů a znaků mezi objekty,
- je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu.

2. kritérium těsnosti proložení:

kritérium delta Δ

- měří stupeň přetvoření struktury dat,
- je žádoucí, aby hodnoty *delta* byly blízké nule,
- je definováno

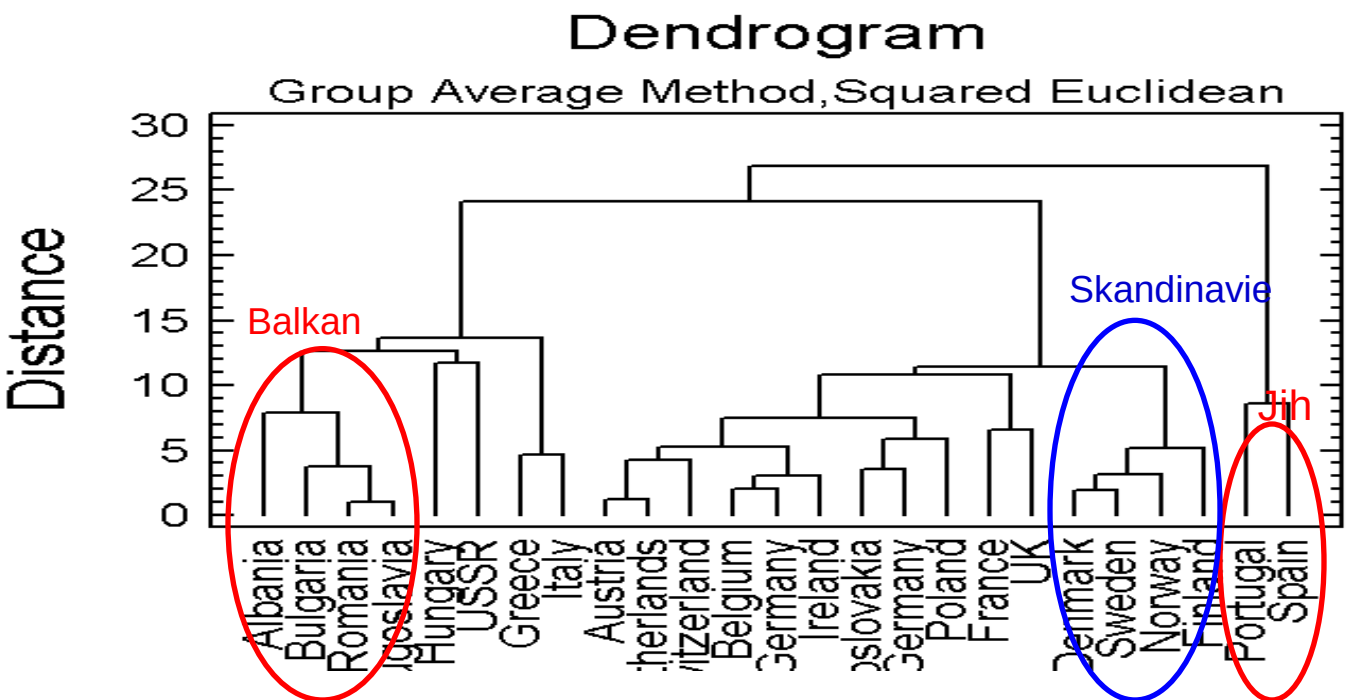
kde A = 0.5 nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^*_{ij} je vzdálenost získaná z dendrogramu.

5. krok: Interpretace shluků

(a) Vyšetření každého shluku v pojmech shlukových proměnných.
(b) Pojmenování shluků nebo jeho označení, které vystihuje jeho podstatu a povahu.

Profilování a interpretace shluků:

(a) Prokazuje popis.
(b) Přidělení korespondence ke shlukům předvídaným z teorie.
(c) V konfirmatorním modu profily přidělují shlukům korespondenci.
(d) Při hledání korespondence nebo praktické významnosti by se měly porovnávat odvozené shluky s předem vytvořenou typologií.



6. krok: Validace a profilování shluků

Existuje subjektivní charakter hledání optimálního shlukového řešení.
Neexistuje jednoduchá metoda, která by zajišťovala validitu a praktický význam.

Validování shluků: znamená, že nalezené shlukové řešení
(a) je reprezentativní,
(b) je zobecnitelné na ostatní objekty v celém původním souboru,
(c) je stabilní i v čase.

Postup: - analyzovat oddělené výběry,
- porovnat nalezená shluková řešení a
- odhadnout shodu výsledků.

Rozdělení výběru dat na dva vzorky: každý vzorek je podroben analýze shluků odděleně a výsledky jsou porovnány:
(1) Modifikovanou formu rozdělení výběru, kdy v prvním vzorku získáme středy shluků a využijeme je k definování shluků ve druhém vzorku objektů a výsledky porovnáme,
(2) Přímá forma vzájemného porovnání (cross-validation).

Způsob vytyčení kritéria: Užijeme takové proměnné, které sice nejsou užity k vytvoření shluků, ale mění se dostatečně od shluku ke shluku.

Úloha 3. Sledování spotřeby proteinů v Evropě (Kompendium B418)

Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem vyšetření.

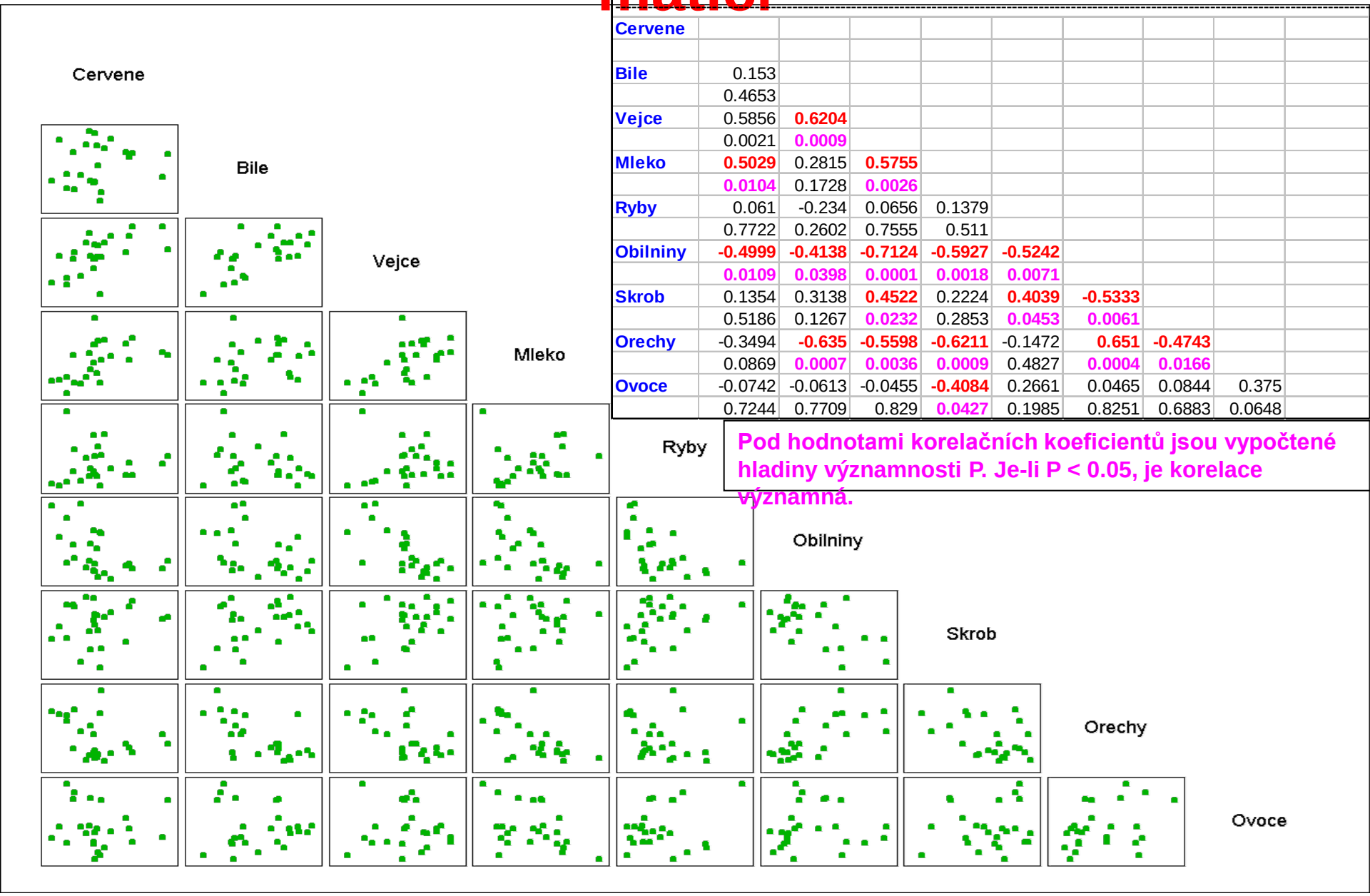
Data: *i* značí index, ***Cervene*** udává červené maso, **Bile** maso, **Vejce, Mleko**, **Ryby**, **Obilniny, Skrob**, **Orechy**, **Ovoce** a zelenina

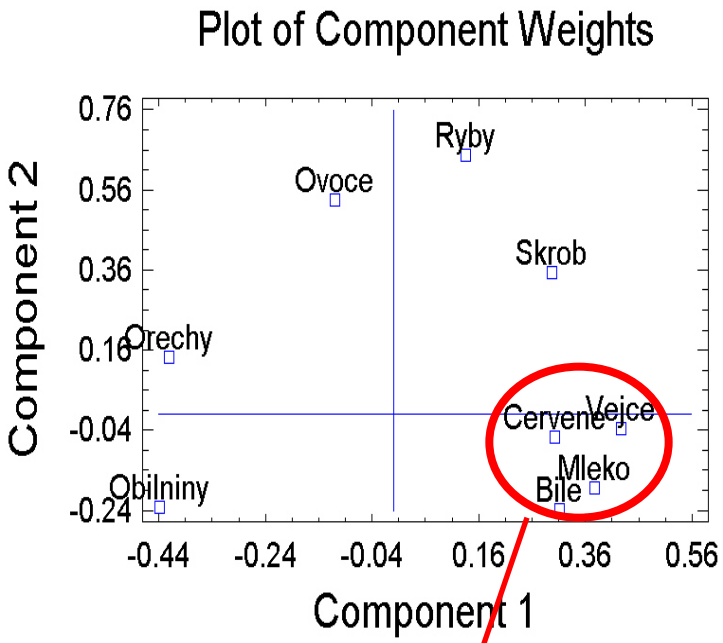
	Objekty	Proměnné								
i	**Stát**	**Cervene**	**Bile**	**Vejce**	**Mleko**	**Ryby**	**Obilniny**	**Skrob**	**Orechy**	**Ovoce**
1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
2	Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
3	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
4	Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
5	Czechoslov.	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
6	Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
7	E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
8	Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
9	France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
10	Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
11	Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
12	Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
13	Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
14	Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
15	Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
16	Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
17	Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
18	Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
19	Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
20	Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
21	Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
22	UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
23	USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
24	W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
25	Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

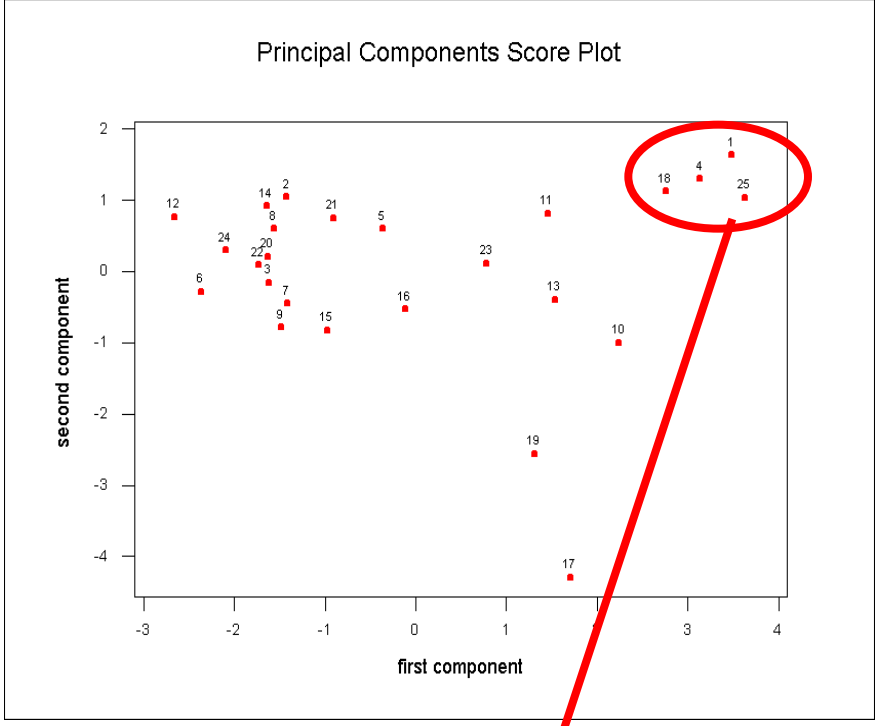
Test významnosti korelace v korelační matici

	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
Cervene									
Bile	0.153								
	0.4653								
Vejce	0.5856	**0.6204**							
	0.0021	**0.0009**							
Mleko	**0.5029**	0.2815	**0.5755**						
	0.0104	0.1728	**0.0026**						
Ryby	0.061	-0.234	0.0656	0.1379					
	0.7722	0.2602	0.7555	0.511					
Obilniny	**-0.4999**	**-0.4138**	**-0.7124**	**-0.5927**	**-0.5242**				
	0.0109	**0.0398**	**0.0001**	**0.0018**	**0.0071**				
Skrob	0.1354	0.3138	**0.4522**	0.2224	**0.4039**	**-0.5333**			
	0.5186	0.1267	**0.0232**	0.2853	**0.0453**	**0.0061**			
Orechy	-0.3494	**-0.635**	**-0.5598**	**-0.6211**	-0.1472	**0.651**	**-0.4743**		
	0.0869	**0.0007**	**0.0036**	**0.0009**	0.4827	**0.0004**	**0.0166**		
Ovoce	-0.0742	-0.0613	-0.0455	**-0.4084**	0.2661	0.0465	0.0844	0.375	
	0.7244	0.7709	0.829	**0.0427**	0.1985	0.8251	0.6883	0.0648	

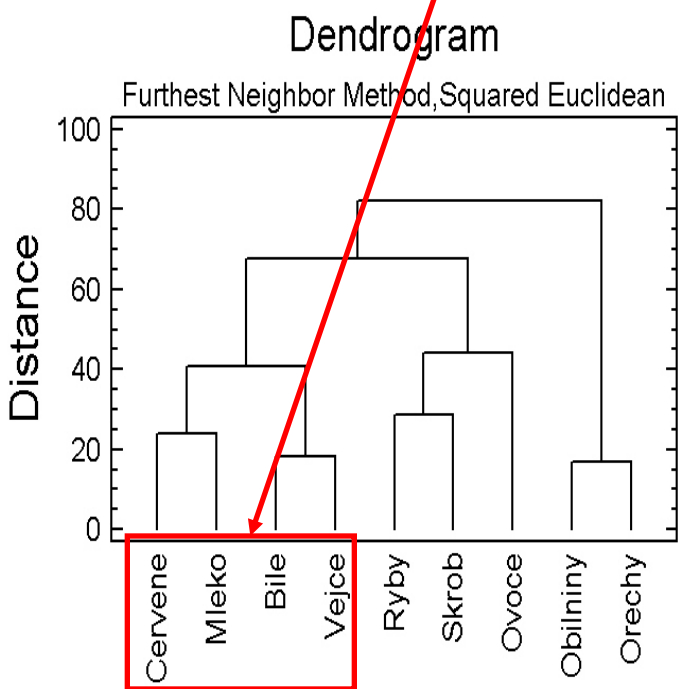
Pod hodnotami korelačních koeficientů jsou vypočtené hladiny významnosti P. Je-li P < 0.05, je korelace významná.

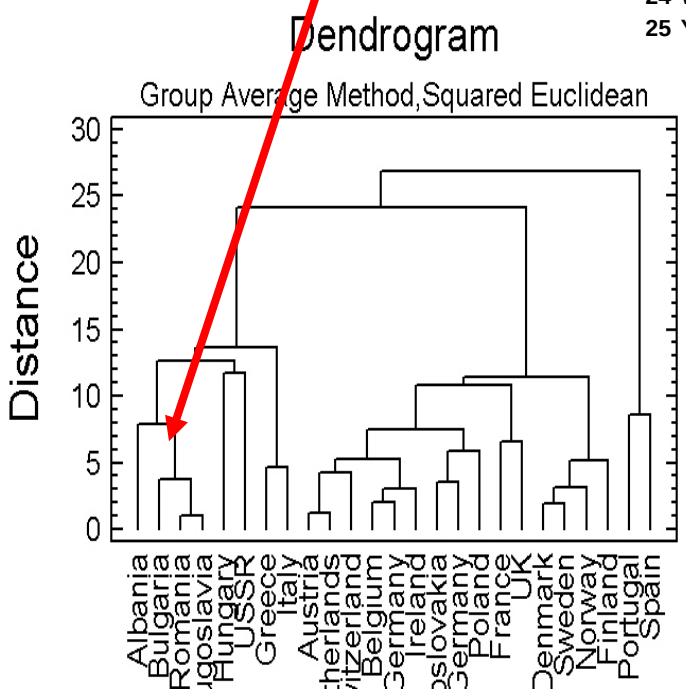






1 Albania
2 Austria
3 Belgium
4 Bulgaria
5 Czechoslovakia
6 Denmark
7 E Germany
8 Finland
9 France
10 Greece
11 Hungary
12 Ireland
13 Italy
14 Netherlands
15 Norway
16 Poland
17 Portugal
18 Romania
19 Spain
20 Sweden
21 Switzerland
22 UK
23 USSR
24 W Germany
25 Yugoslavia



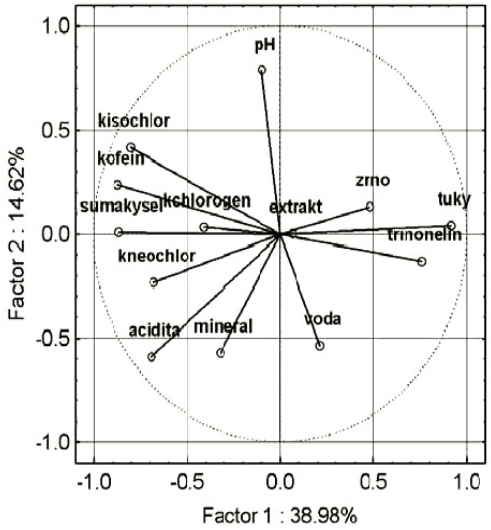


Úloha 6. Klasifikace vlastností rozličných druhů kávy (Kompendium E406)

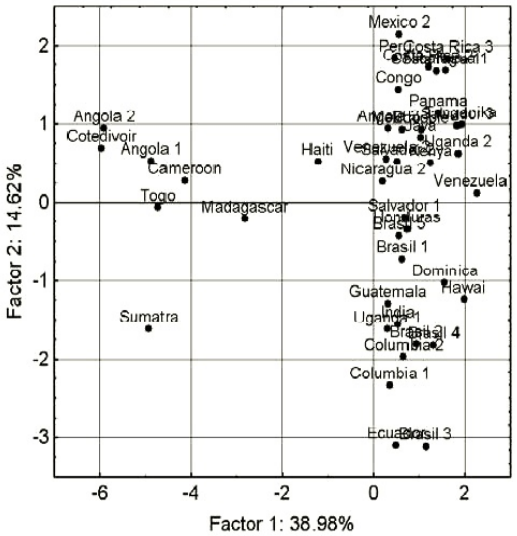
U 43 vzorků kávy ze 30 zemí byly změřeny chemické a fyzikální vlastnosti. Nalezněte shluky podobných vlastností a shluky podobných prvků.

Data: 13 proměnných (sloupce): **i** index kávy, **j** je původ kávy, **x1** obsah vody, **x2** hmotnost zrn, **x3** extrakt, **x4** pH, **x5** volná acidita, **x6** obsah minerálů, **x7** tuky, **x8** kofein, **x9** trinonelin, **x10** kyselina chlorogeniková, **x11** kyselina neochlorogeniková, **x12** kyseliny isochlorogeniková, **x13** suma kyselin chlorogenikových.

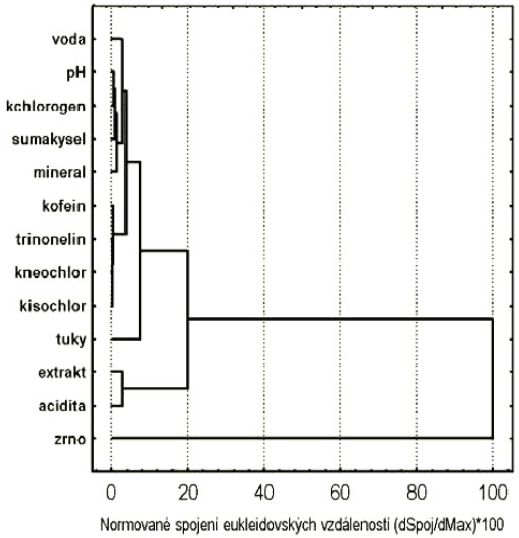
i	ii	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1	5.4	0.4	0.8	6.6
2	Mexico 2	7.4	157.3	32.1	5.8	30.8	3.7	15	1.3	1	5.1	0.3	1	6.4
3	Guatemala	9.7	152.9	33.1	5.3	36.7	4.2	16.1	1.2	1	5.9	0.2	0.8	6.9
4	Honduras	10.4	174	31.5	5.6	34.2	3.9	15.8	1.1	0.9	5.9	0.4	0.6	6.8
5	Salvador 1	10.5	145.1	35.2	5.8	31.8	4.1	15.2	1.1	1	5.1	0.5	0.7	6.3
6	Salvador 2	10	156.4	34.5	5.8	32.6	3.9	15.4	1.2	0.8	5.3	0.4	0.7	6.4
7	Salvador 3	8.2	155.2	32.4	5.6	29.7	3.8	15.6	1.3	1.2	4.8	0.3	0.7	5.9
8	Nicaragua 1	9.2	167.8	30.6	5.9	28.9	3.8	15.1	1.3	1	5	0.3	0.7	5.9
9	Nicaragua 2	9.3	165.4	35.3	5.8	32.6	4.2	14.3	1.2	1	5.5	0.4	0.8	6.7
10	Costa Rica 1	7.1	180.3	33	5.8	29.3	4	15.1	1.3	1	5.1	0.3	0.7	6.1
11	Costa Rica 2	7.6	153.2	36	5.9	30.5	3.9	16.8	1.4	1.1	5.3	0.3	0.7	6.3
12	Costa Rica 3	7.3	159.6	35	5.8	29.9	3.7	16.5	1.2	1.2	5.5	0.3	0.7	6.5
13	Panama	9.3	161.8	32.4	5.8	31	3.7	15.5	1.3	1.2	5.6	0.3	0.6	6.6
14	Haiti	8.3	160.8	35.7	5.9	30	4.4	13	1.3	1	6.1	0.6	0.8	7.5
15	Dominica	11.6	174.8	32.5	5.4	35.2	3.7	14.5	1	1	5.7	0.3	0.5	6.5
16	Venezuela 1	9.7	169.1	34	5.8	31.6	4	15.7	1.3	1.3	5.1	0.3	0.3	6.2
17	Venezuela 2	10.6	163.7	35	5.8	35	3.8	15.8	1.2	1.1	6.1	0.3	0.9	7.3
18	Columbia 1	12	178.8	32.9	5.3	36.2	4.4	15.6	1.3	1	5.6	0.4	0.7	6.7
19	Columbia 2	10.6	169.1	33	5.3	37.5	4.4	15.1	1.2	1	6.1	0.1	0.6	6.9
20	Ecuador	11.6	148.5	34.6	5.3	39.4	4.2	14.6	1	1.1	5.7	0.5	0.4	6.6
21	Peru	10.1	153.7	34.5	6	28.4	3.7	15.9	1.3	1.1	6.1	0.4	0.8	7.3
22	Brasil 1	10.7	134.5	29.8	5.4	34.1	3.7	15.8	1.2	0.9	5.4	0.4	0.6	6.4
23	Brasil 2	9.7	160.7	33.8	5.3	37.2	4.2	15.2	1.1	0.9	5.4	0.3	0.5	6.2
24	Brasil 3	10.8	133.2	35	5.2	34.7	4.5	15.1	1.2	1.4	5	0.5	0.5	6
25	Brasil 4	11.1	131.7	29.8	5.4	33	4.1	15.8	1.1	1.2	5.1	0.5	0.5	6
26	Brasil 5	10.1	121.6	33.6	5.4	34.7	3.5	15.4	1.1	0.9	5.5	0.4	0.6	6.5
27	Cotedivoir	8	141.8	33.7	5.8	41.9	4.2	11	2	0.5	6.4	0.6	1.5	8.5
28	Togo	9	144.6	29.9	5.6	38	3.9	7.5	1.9	0.3	5.4	0.8	0.9	7.1
29	Cameroon	10.3	119.2	35.5	6.1	41.7	4.1	9.8	1.8	0.8	6	0.5	1.1	7.6
30	Congo	10	143.2	31.7	6.1	29.3	4.1	17	1.2	0.6	5.4	0.3	0.7	6.4
31	Angola 1	9.2	150.4	31.5	5.7	36.4	4.2	8.5	1.9	0.6	5.9	0.6	1.4	7.9
32	Angola 2	9.6	136.6	33.9	5.6	38.2	4	7.2	2.2	0.5	6.2	0.4	1.6	8.3
33	Angola 3	9.5	136.5	32	5.8	31.2	3.8	14.6	1.3	1	5.2	0.4	0.8	6.4
34	Ethiopie	9.3	124.2	35.6	5.8	31.8	3.8	15.7	0.9	0.9	5.5	0.2	0.8	6.5
35	Uganda 1	10.5	132.9	36.2	5.4	36.7	4	15.6	1	1	5.9	0.4	0.6	6.9
36	Uganda 2	10.7	181.2	33.1	5.8	30.7	3.9	15.8	1.3	1.1	5.3	0.3	0.6	6.2
37	Kenya	10.5	159.1	30.3	5.6	31.5	3.7	15.2	1.3	0.9	5.1	0.3	0.7	6
38	Tanganika	9.9	169.4	29	5.6	30.2	3.7	16.5	1.3	0.9	5	0.2	0.7	5.9
39	Madagascar	5	152	30.6	5.3	40.5	3.9	9.6	1.6	0.7	5.3	0.6	0.8	6.7
40	India	11.5	156.8	30.8	5.5	37.5	3.9	14.3	1.2	1	5.8	0.4	0.4	6.6
41	Sumatra	8.4	110.8	31.6	5.7	43.4	4.5	10.1	1.7	0.8	6.3	0.7	0.9	7.9
42	Java	5.6	163.1	34.5	5.5	33.3	4	16	1.2	1.1	5.1	0.3	0.8	6.3
43	Hawai	9.7	191.2	35.1	5.6	34.6	4.2	14.2	1.1	0.9	0.7	0.5	0.3	6.5



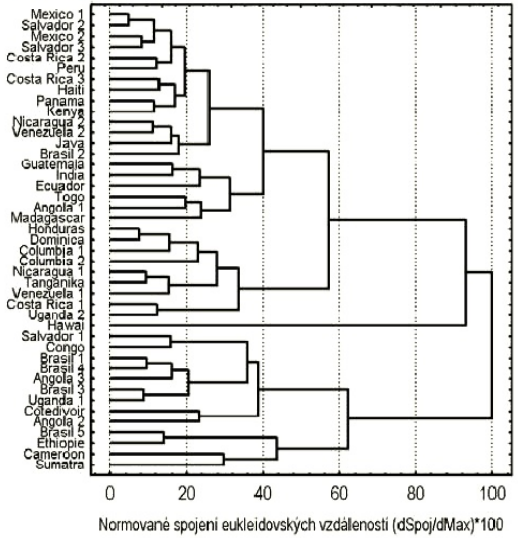
Graf komponentních vah znaků matice dat *Kava*, (STATISTICA).



Graf komponentního skóre objektů matice dat *Kava*



Dendrogram znaků matice dat *Kava* (STATISTICA).



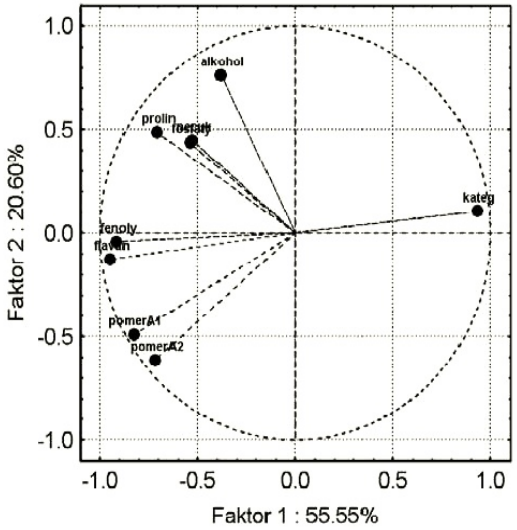
Dendrogram objektů matice dat *Kava*, (STATISTICA).

○ ***Závěr:*** Dendrogram znaků ukazuje shluky podobných vlastností kávy, zatímco dendrogram objektů klasifikuje podobné druhy kávy do shluků.

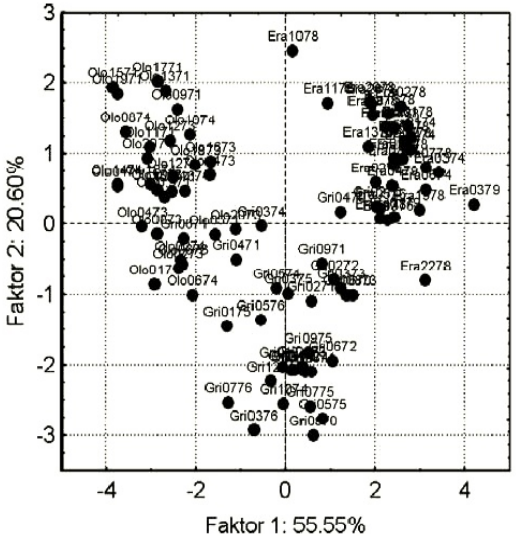
Úloha 4. Faktorová analýza při klasifikaci vzorků vín (Kompendium E408)

Pro 38 vzorků vín bylo nalezeno 24 analytických obsahů stopových prvků a charakteristických fyzikálně-chemických vlastností. Utvořte shluky podobných vlastností a dále shluky podobných vín.

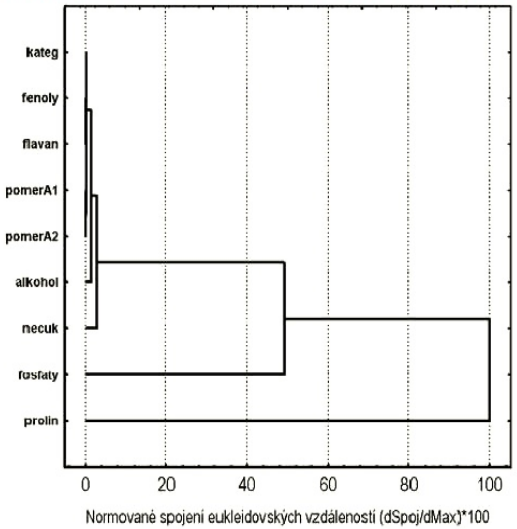
Index	Cd	Mo	Mn	Ni	Cu	Al	Ba	Cr	Sr	Pb	B	Mg	Si	Na	Ca	P	K	Arom	Clar	Body	Flavor	Oakn	Quality	Reg
1	0.005	0.044	1.51	0.122	0.83	0.982	0.387	0.029	1.23	0.561	2.63	128	17.3	66.8	80.5	150	1130	3.3	1	2.8	3.1	4.1	9.8	1
2	0.055	0.16	1.16	0.149	0.066	1.02	0.312	0.038	0.975	0.697	6.21	193	19.7	53.3	75	118	1010	4.4	1	4.9	3.5	3.9	12.6	1
3	0.056	0.146	1.1	0.088	0.643	1.29	0.308	0.035	1.14	0.73	3.05	127	15.8	35.4	91	161	1160	3.9	1	5.3	4.8	4.7	11.9	1
4	0.063	0.191	0.959	0.38	0.133	1.05	0.165	0.036	0.927	0.796	2.57	112	13.4	27.5	93.6	120	924	3.9	1	2.6	3.1	3.6	11.1	1
5	0.011	0.363	1.38	0.16	0.051	1.32	0.38	0.059	1.13	1.73	3.07	138	16.7	76.6	84.6	164	1090	5.6	1	5.1	5.5	5.1	13.3	1
6	0.05	0.106	1.25	0.114	0.055	1.27	0.275	0.019	1.05	0.491	6.56	172	18.7	15.7	112	137	1290	4.6	1	4.7	5	4.1	12.8	1
7	0.025	0.479	1.07	0.168	0.753	0.715	0.164	0.062	0.823	2.06	4.57	179	17.8	98.5	122	184	1170	4.8	1	4.8	4.8	3.3	12.8	1
8	0.024	0.234	0.906	0.466	0.102	0.811	0.271	0.044	0.963	1.09	3.18	145	14.3	10.5	91.9	187	1020	5.3	1	4.5	4.3	5.2	12	1
9	0.009	0.058	1.84	0.042	0.17	1.8	0.225	0.022	1.13	0.048	6.13	113	13	54.4	70.2	158	1240	4.3	1	4.3	3.9	2.9	13.6	3
10	0.033	0.074	1.28	0.098	0.053	1.35	0.329	0.03	1.07	0.552	3.3	140	16.3	70.5	74.7	159	1100	4.3	1	3.9	4.7	3.9	13.9	1
11	0.039	0.071	1.19	0.043	0.163	0.971	0.105	0.028	0.491	0.31	6.56	103	9.5	45.3	67.9	133	1090	5.1	1	4.3	4.5	3.6	14.4	3
12	0.045	0.147	2.76	0.071	0.074	0.483	0.301	0.087	2.14	0.546	3.5	199	9.2	80.4	66.3	212	1470	3.3	0.5	5.4	4.3	3.6	12.3	2
13	0.06	0.116	1.15	0.055	0.18	0.912	0.166	0.041	0.578	0.518	6.43	111	11.1	59.7	83.8	139	1120	5.9	0.8	5.7	7	4.1	16.1	3
14	0.067	0.166	1.53	0.041	0.043	0.512	0.132	0.026	0.229	0.699	7.27	107	6	55.2	44.9	148	854	7.7	0.7	6.6	6.7	3.7	16.1	3
15	0.077	0.261	1.65	0.073	0.285	0.596	0.078	0.063	0.156	1.02	5.04	94.6	6.3	10.4	54.9	132	899	7.1	1	4.4	5.8	4.1	15.5	3
16	0.064	0.191	1.78	0.067	0.552	0.633	0.085	0.063	0.192	0.777	5.56	110	7	13.6	64.1	167	976	5.5	0.9	5.6	5.6	4.4	15.5	3
17	0.025	0.009	1.57	0.041	0.081	0.655	0.072	0.021	0.172	0.232	3.79	75.9	6.4	11.6	48.1	132	995	6.3	1	5.4	4.8	4.6	13.8	3
18	0.02	0.027	1.74	0.046	0.153	1.15	0.094	0.021	0.358	0.025	4.24	80.9	7.9	38.9	57.6	136	876	5	1	5.5	5.5	4.1	13.8	3
19	0.034	0.05	1.15	0.058	0.058	1.35	0.294	0.006	1.12	0.206	2.71	120	14.7	68.1	64.8	133	1050	4.6	1	4.1	4.3	3.1	11.3	1
20	0.013	0.03	2.82	0.058	0.05	0.623	0.349	0.082	2.91	0.171	3.54	208	9.3	79.2	66.4	266	1430	3.4	0.9	5	3.4	3.4	7.9	2
21	0.043	0.268	2.32	0.066	0.314	0.627	0.099	0.045	0.36	1.28	5.68	98.4	9.1	19.5	64.3	176	945	6.4	0.9	5.4	6.6	4.8	15.1	3
22	0.061	0.245	1.61	0.07	0.172	2.07	0.071	0.053	0.186	1.19	4.42	87.6	7.6	11.6	70.6	156	820	5.5	1	5.3	5.3	3.8	13.5	3
23	0.047	0.161	1.47	0.154	0.082	0.546	0.181	0.06	0.898	0.747	8.11	160	19.3	12.5	82.1	218	1220	4.7	0.7	4.1	5	3.7	10.8	2
24	0.048	0.146	1.85	0.092	0.09	0.889	0.328	0.1	1.32	0.604	6.42	134	19.3	125	83.2	173	1810	4.1	0.7	4	4.1	4	9.5	2
25	0.049	0.155	1.73	0.051	0.158	0.653	0.081	0.037	0.164	0.767	4.91	86.5	6.5	11.5	53.9	172	1020	6	1	5.4	5.7	4.7	12.7	3
26	0.042	0.126	1.7	0.112	0.21	0.508	0.299	0.054	0.995	0.686	6.94	129	43.6	45	85.9	165	1330	4.3	1	4.6	4.7	4.9	11.6	2
27	0.058	0.184	1.28	0.095	0.058	1.3	0.346	0.037	1.17	1.28	3.29	145	16.7	65.8	72.8	175	1140	3.9	1	4	5.1	5.1	11.7	1
28	0.065	0.211	1.65	0.102	0.055	0.308	0.206	0.028	0.72	1.02	6.12	99.3	27.1	20.5	95.2	194	1260	5.1	1	4.9	5	5.1	11.9	2
29	0.065	0.129	1.56	0.166	0.151	0.373	0.281	0.034	0.889	0.638	7.28	139	22.2	13.3	84.2	164	1200	3.9	1	4.4	5	4.4	10.8	2
30	0.068	0.166	3.14	0.104	0.053	0.368	0.292	0.039	1.11	0.831	4.71	125	17.6	13.9	59.5	141	1030	4.5	1	3.7	2.9	3.9	8.5	2
31	0.067	0.199	1.65	0.119	0.163	0.447	0.292	0.058	0.927	1.02	6.97	131	38.3	42.9	85.9	164	1390	5.2	1	4.3	5	6	10.7	2
32	0.084	0.266	1.28	0.087	0.071	1.14	0.158	0.049	0.794	1.3	3.77	143	19.7	39.1	128	146	1230	4.2	0.8	3.8	3	4.7	9.1	1
33	0.069	0.183	1.94	0.07	0.095	0.465	0.225	0.037	1.19	0.915	2	123	4.6	7.5	69.4	123	943	3.3	1	3.5	4.3	4.5	12.1	1
34	0.087	0.208	1.76	0.061	0.099	0.683	0.087	0.042	0.168	1.33	5.04	92.9	7	12	56.3	157	949	6.8	1	5	6	5.2	14.9	3
35	0.074	0.142	2.44	0.051	0.052	0.737	0.408	0.022	1.16	0.745	3.94	143	6.8	36.8	67.6	82	1170	5	0.8	5.7	5.5	4.8	13.5	1
36	0.084	0.171	1.85	0.088	0.038	1.21	0.263	0.072	1.35	0.899	2.38	130	6.2	101	64.4	99	1070	3.5	0.8	4.7	4.2	3.3	12.2	1
37	0.106	0.307	1.15	0.063	0.051	0.643	0.29	0.031	0.885	1.61	4.4	151	17.4	7.3	103	177	1100	4.3	0.8	5.5	3.5	5.8	10.3	1
38	0.102	0.342	4.08	0.065	0.077	0.752	0.366	0.048	1.08	1.77	3.37	145	5.3	33.1	58.3	117	1010	5.2	0.8	4.8	5.7	3.5	13.2	1



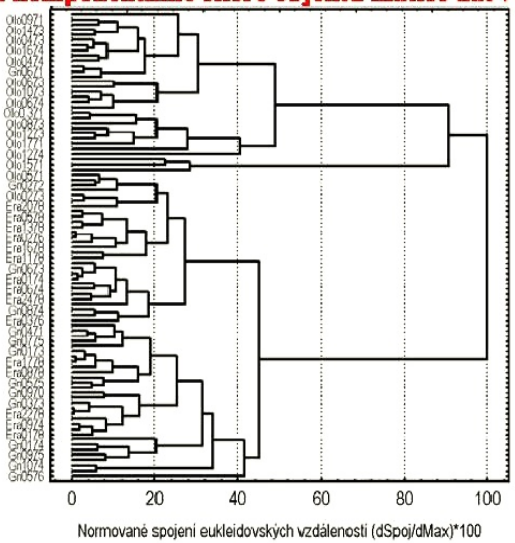
Graf komponentních vah znaků matice dat *Vina*, (STATISTICA).



Graf komponentního skóre objektů matice dat *Vina*.



Dendrogram znaků matice dat *Vina* (STATISTICA).



Dendrogram objektů matice dat *Vina*, (STATISTICA).

○ ***Závěr:*** Graf komponentního skóre objektů a dendrogram objektů shodně vykazují tři větší shluky vín ve shodě s jejich kategoriemi ***Olo, Era*** a ***Gri***.

Diskriminační analýza DA

hodnotí rozdíly mezi dvěma nebo více skupinami objektů charakterizovaných pomocí několika znaků.

Dělí se na

1) Techniky, které *vysvětlují rozdíly* mezi předem danými shluky objektů.
2) Techniky, u kterých je cílem *klasifikace objektů do skupin*. Jsou porovnávány znaky jednoho objektu (např. charakteristiky sloučenin, vlastnosti pacientů, ...) se znaky ostatních.

Příklad: Analyzovaný výběr obsahuje *třídu přežívajících* pacientů a *třídu nepřežívajících*. Dle klinických vyšetření chceme rozhodnout, kam nového pacienta z klasifikovaného výběru zařadit. Potřebujeme znaky k rozlišení dosud nezařazených pacientů do tříd.

Klasická DA (Ronald Fisher, 1936): zkoumání vztahu mezi skupinou p nezávislých znaků (*diskriminátory*) a jednou kvalitativní závisle proměnnou y, nabývající hodnotu 0 (objekt je v 1. třídě), nebo hodnotu 1 (objekt je ve 2. třídě).

Třídy jsou zřetelně *odlišené*. Každý objekt patří do jedné ze dvou tříd.

Cílem je nalézt znaky, které úspěšně přispívají do procesu *třídění*.

DISKRIMINAČNÍ ANALÝZA

TRÉNOVÁNÍ S UČITELEM (SUPERVISED LEARNING)

$$\boxed{y} \quad \boxed{x_1 \; x_2 \; x_3 \; x_4 \; \cdots\cdots \; x_m} \longrightarrow \text{MODEL (FISHEROVA FUNKCE)}$$

1. TRÉNOVACÍ DATA
(ANALYZOVANÝ SOUBOR)

$$\boxed{} \quad \boxed{x_1 \; x_2 \; x_3 \; x_4 \; \cdots\cdots \; x_m} \longleftarrow \text{VYUŽITÍ MODELU}$$

2. KLASIFIKACE OBJEKTŮ
(KLASIFIKOVANÝ SOUBOR)

V datech **analyzovaného výběru** jsou pomocí znaků objekty umístěny do tříd.
V **klasifikovaném výběru** je třeba nalézt *predikční model,* který umístí **nové objekty** do stávajících tříd.

Řešení: v analyzovaném výběru

π_1 značí apriorní pravděpodobnost příslušnosti objektu do první třídy,
$\pi_2 = 1 - \pi_1$ značí pravděpodobnost příslušnosti objektu do druhé třídy.

Pomocí Bayesovy věty pak určíme aposteriorní pravděpodobnost příslušnosti k j-té skupině ($j = 1, 2$)

$$P(G = j \mid \boldsymbol{x}) = \frac{f_j(\boldsymbol{x})\ \pi_j}{\sum_{i=1}^{2} f_i(\boldsymbol{x})\ \pi_i}, \quad j = 1,2$$

Zápis se dá snadno rozšířit i pro více kategorií.

Třídění:

Objekt zařadíme do první skupiny, pokud $\pi_1 f_1(x) > \pi_2 f_2(x)$.
Pro zařazení objektu do první skupiny musí platit, že

$$f_1(x)/f_2(x) > \pi_2/\pi_1$$

Pro vícerozměrné Gaussovo rozdělení v i-té třídě má odpovídající hustota pravděpodobnosti tvar

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}\,\det(\boldsymbol{C}_i)^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \mu_i)^T \boldsymbol{C}_i^{-1} (\boldsymbol{x} - \mu_i)\right]$$

kde $m = 2$ je počet znaků.

Lineární diskriminační funkce LDA

Pro dvě třídy 1 a 2 vychází LDA z předpokladu, že $\boldsymbol{C}_1 = \boldsymbol{C}_2 = \boldsymbol{C}$.

Logaritmus poměru aposteriorních pravděpodobností je

$$\ln\frac{P(G = 1 \mid x)}{P(G = 2 \mid x)} = \ln\left(\frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x})} + \ln\frac{\pi_1}{\pi_2}\right) =$$

$$= \ln\frac{\pi_1}{\pi_2} - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{C}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \boldsymbol{x}^T \boldsymbol{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Vzhledem k $\boldsymbol{x} = (x_1, x_2)^T$ jde o lineární funkci a dělicí funkcí bude přímka.

Dělicí funkce je množinou bodů $\boldsymbol{x}$, pro které platí

$$P(G = 1/\boldsymbol{x}) = P(G = 0/\boldsymbol{x}).$$

Pro více znaků jsou dělicí čáry lineárními úseky procházejícími průsečíky jednotlivých elips konstantní hustoty pro jednotlivé třídy.

Zařazovací pravidla DA

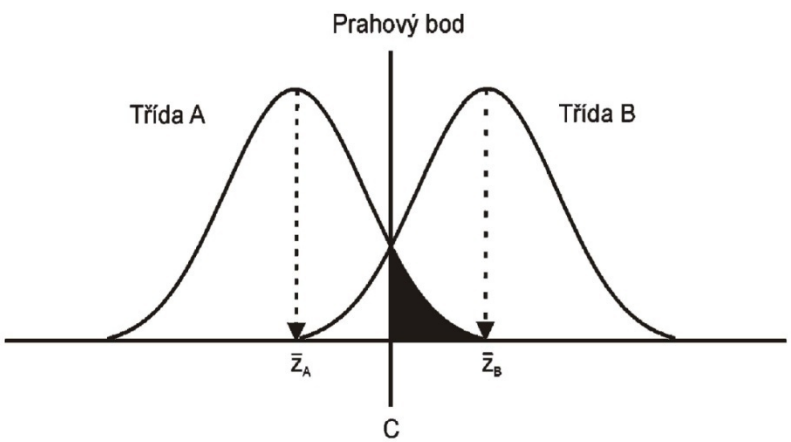
účelem je klasifikace do jedné ze dvou tříd, A nebo B

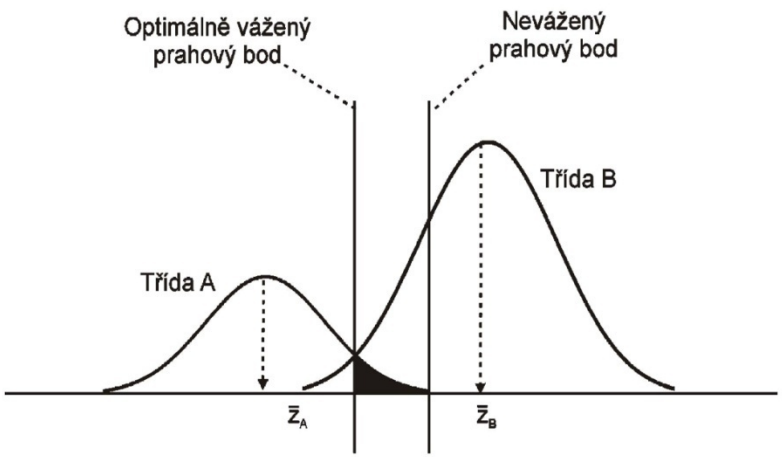
a) Klasifikace na základě jediného znaku *x* (s normálním rozdělením):

1) ve třídě A jde o rozdělení $N(\mu_A, \sigma_A^2)$,
2) ve třídě B jde o rozdělení $N(\mu_B, \sigma_B^2)$,
3) nový objekt má hodnotu jediného znaku x,
4) vybrat třídu, pro kterou je x blíže ke střední hodnotě dané třídy μ_A či μ_B.

b) Určíme *prahový bod* $C = (\mu_A + \mu_B)/2$:

- je-li $x < C$, pak se objekt zařadí do první třídy, A,
- je-li $x \geq C$ pak se objekt zařadí do druhé třídy, B.





1) $\sigma_A^2 = \sigma_B^2$ pravděpodobnost nesprávné klasifikace je šrafovaně. Tato souvisí se vzdáleností obou rozdělení a s jejich rozptyly.
2) $\sigma_A^2 < \sigma_B^2$ je pravděpodobnost nesprávné klasifikace pro třídu A větší než pro třídu B. Bude proto třeba penalizovat C s ohledem na nestejný rozptyl.

Příklad 4.22 ***Třídění lebek Tibeťanů lineární diskriminační funkcí***

Databáze lebek na pohřebištích v Tibetu svědčí o dvou skupinách lidí:

prvních 13 bylo nalezeno v hrobech v Sikkimu a okolí,

druhých 15 lebek na bojištích okolo Lhasy.

Předpokládejme, že máme data o 2 třídách tibetských lebek.

Data: *i* index lebky,

x1 největší délka lebky [mm],

x2 největší horizontální šířka lebky [mm],

x3 výška lebky [mm],

x4 výška horní části obličeje [mm],

x5 šířka obličeje mezi body lícních kostí [mm].

x1	***x2***	***x3***	***x4***	***x5***
190.5	152.5	145	73.5	136.5
...	...	...	...	...
...	...	...	...	...
172.5	132	125.5	63	121
167	130	125.5	69.5	119.5
1825	131	135	68.5	136

Řešení:

1) **Analyzovaný výběr**: na trénovacích datech vytvoříme třídy

První třída (13 lebek z hrobů v Sikkimu) vede ke středním hodnotám

[174.82, 139.35, 132.00, 69.82, 130.35]

a kovarianční matici $S_1 = \begin{vmatrix} 45.53 & & & & \\ 25.22 & 57.81 & & & \\ 12.39 & 11.88 & 36.09 & & \\ 22.15 & 7.52 & -0.31 & 20.94 & \\ 27.97 & 48.06 & 1.41 & 16.77 & 66.21 \end{vmatrix}$.

Druhá třída (15 lebek z bojišť v Lhase) vede ke středním hodnotám

[185.73, 138.73, 134.77, 76.47, 137.50]

a kovarianční matici $S_2 = \begin{vmatrix} 74.42 & & & & \\ -9.52 & 37.35 & & & \\ 22.74 & -11.26 & 36.32 & & \\ 17.79 & 0.70 & -10.72 & 15.30 & \\ 11.13 & 9.46 & 7.20 & 8.66 & 17.96 \end{vmatrix}$.

Koeficienty diskriminační funkce $a_1, \ldots, a_5$, jsou vyčísleny podle vztahu

$$\boldsymbol{a} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = [-0.09, 0.16, 0.01, -0.18, -0.18]$$

a vedou k *průměrům obou tříd*: $\bar{Z}_1 = -28.71$ a $\bar{Z}_2 = -32.21$.

Optimální prahový bod C,

dle kterého se budou nezařazené objekty třídit do první nebo druhé třídy, se vyčíslí jako polosuma obou průměrů dle vztahu

$$C = (\bar{Z}_1 + \bar{Z}_2)/2 = (-28.71 + (-32.21))/2 = -30.46.$$

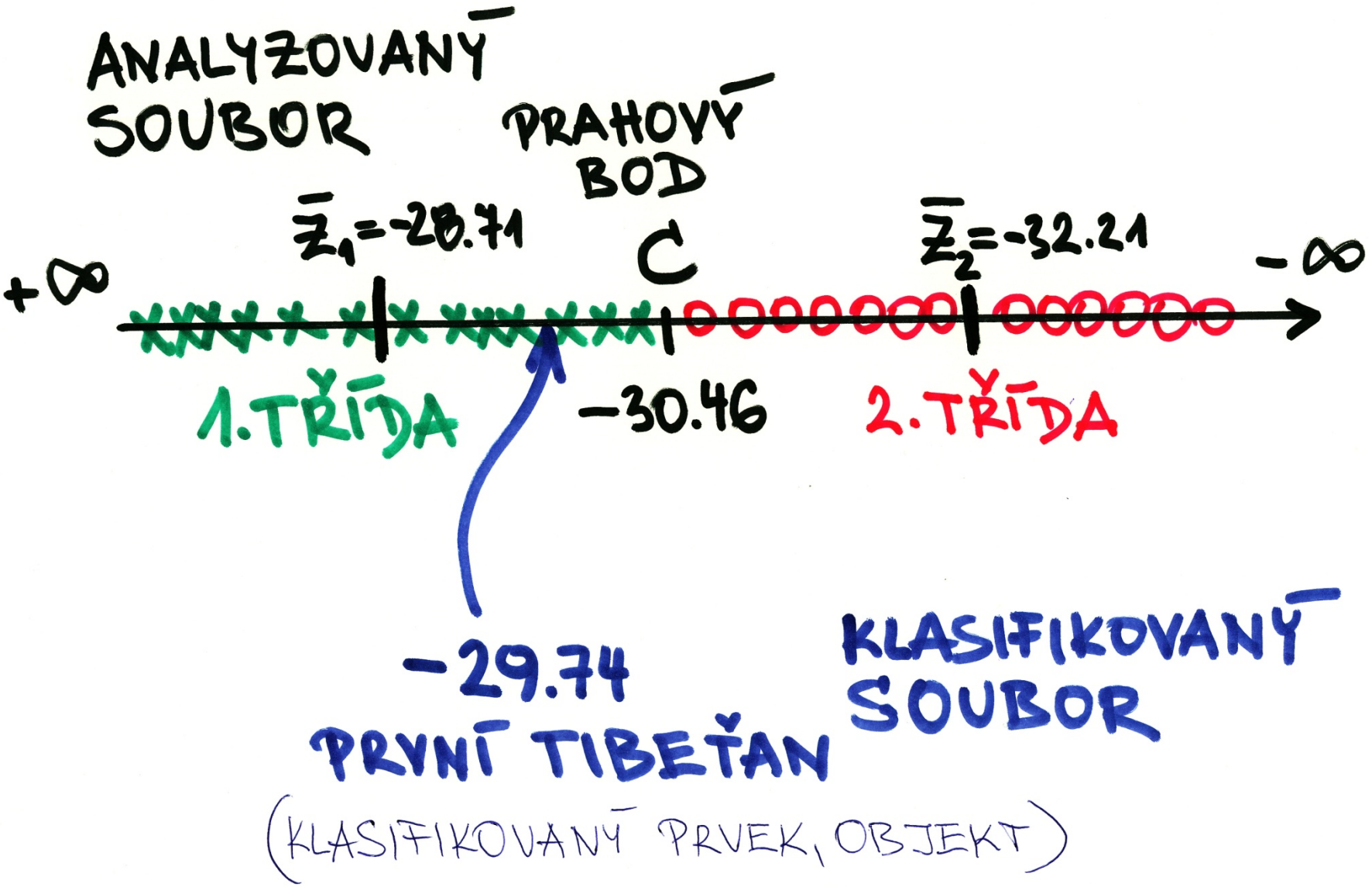
Diskriminace:

2) **Klasifikovaný výběr:** na nových datech provedeme zařazení dosud nezařazených objektů

Vezmeme *data pro lebku prvního Tibeťana* a pokusíme se ji zařadit do 1. nebo do 2. třídy. Vyčísleme proto pro ni hodnotu *lineární diskriminační funkce*

$$Z_1 = -0.09 \times 190.5 + 0.16 \times 152.5 + 0.01 \times 145.0 - 0.18 \times 73.5 - 0.18 \times 136.5 = -29.74,$$

Závěr: Protože lineární diskriminační funkce $Z_1 = -29.74$ je menší než optimální prahový bod $C = -30.46$, *patří lebka prvního Tibeťana do první třídy*.


ANALYZOVANÝ SOUBOR
PRAHOVÝ BOD
$\bar{z}_1 = -28.71$
C
$\bar{z}_2 = -32.21$
$+\infty$
$-\infty$
1. TŘÍDA
-30.46
2. TŘÍDA
-29.74
PRVNÍ TIBEŤAN
(KLASIFIKOVANÝ PRVEK, OBJEKT)
KLASIFIKOVANÝ SOUBOR

Kritéria k vybírání znaků

Wilkovo kritérium λ: když znak poskytuje v diskriminační funkci nejmenší hodnotu Wilkova kritéria λ, je zahrnut do modelu:
- k zavedení nebo odstranění znaku je dovolen jeden krok.
- maximální počet kroků je roven dvojnásobku počtu znaků.

- ***tolerance*** je mírou stupně lineární asociace mezi znaky. Pro *i*-tý znak platí vztah $1 - R_i^2$, kde R_i^2 je čtverec vícenásobného korelačního koeficientu, když je uvažován *i*-tý znak za závisle proměnnou a když je uvažována regresní rovnice mezi tímto *i*-tým znakem a ostatními znaky.

- ***využití tolerance***: malé hodnoty tolerance indikují, že *i*-tý znak je lineární kombinací ostatních znaků. Znaky s tolerancí menší než 0.001 nejsou do modelu zařazeny.

Testování: významnost změny Wilkova kritéria λ po zavedení znaku do modelu nebo odstranění z modelu je založena na testačním kritériu *F*.

Na začátku procesu vybírání znaků: tolerance a minimum tolerance jsou položeny rovny 1, protože v modelu zatím nejsou znaky.

F-test významnosti každého znaku: hodnota F pro změnu Wilkova kritéria λ při přidání znaku do modelu se vyčíslí dle

$$F_{\text{změny}} = \frac{n - g - p}{g - 1} \left(\frac{\dfrac{1 - \lambda_{p+1}}{\lambda_p}}{\dfrac{\lambda_{p+1}}{\lambda_p}} \right)$$

kde n je celkový počet objektů,
g udává počet tříd,
p je počet znaků,
λ_p značí Wilkovo lambda před přidáním a
λ_{p+1} je Wilkovo lambda po přidání znaku do modelu.

Testování: Do modelu je zařazen ten znak, který způsobuje nejmenší hodnotu Wilkova kritéria λ

Raovo *V* kritérium, známé jako ***Lawley-Hotellingova stopa***, je

$$V = (n - g) \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \sum_{k=1}^{g} (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j)$$

kde p udává počet znaků v modelu,

g značí počet tříd,

n_k je velikost k-té třídy,

$\bar{x}_{ik}$ je střední hodnota i-tého znaku v k-té třídě,

$\bar{x}$ je průměr i-tého znaku pro všechny třídy kombinované dohromady a

w^*_{ij} představuje prvek inverzní kovarianční matice mezi třídami $\boldsymbol{B}^{-1}$.

Testování: čím větší jsou rozdíly mezi středními hodnotami (průměry) tříd, tím větší je hodnota Raova V.

Může se však stát, že znak po zařazení do modelu sníží hodnotu Raova V.

Mahalanobisova vzdálenost $D_{1,2}^2$ je zobecněná míra vzdálenosti mezi dvěma třídami 1 a 2 definovaná vztahem

$$D_{1,2}^2 = (n - g) \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} (\bar{x}_{i1} - \bar{x}_{i2})(\bar{x}_{j1} - \bar{x}_{j2})$$

kde m udává počet znaků v modelu,

$\bar{x}_{i1}$ je průměr i-tého znaku ve třídě 1,

w_{ij} je prvek inverzní kovarianční matice $\boldsymbol{B}^{-1}$.

Testování: kritérium všech párů tříd vyčísleno je jako první. Znak, který měl pro dvě od začátku nejtěsnější třídy, nejmenší hodnotu $D_{1,2}^2$, je zařazen do modelu.

***F*-testační kritérium mezi třídami** testuje nulovou hypotézu H_0: *dva vektory středních hodnot tříd objektů jsou stejné.*

Kritérium je postaveno na Mahalanobisově vzdálenosti $D_{1,2}^2$.

Testační kritérium *F* je definováno vztahem

$$F = \frac{(n - 1 - p)\, n_1\, n_2}{p(n - 2)\,(n_1 + n_2)}\, D_{1,2}^2$$

a může být použito k výběru znaků.

Testování: V každém kroku je zařazen do modelu ten znak, který vykazuje největší hodnotu kritéria F.

Po zavedení dalších znaků do modelu jsou sledovány změny hodnoty testačního kritéria $F_{\text{změny}}$.

Proces přidávání znaků do modelu buď pokračuje, nebo je zastaven terminačním kritériem.

N=143	Výsledky diskriminační funkční analýzy (B23) Počet prom. v modelu: 5; grupovací: B23x6-GROUP (3 skup) Wilk. lambda: ,10863 přibliž F (10,272)=55,328 p<0,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,136)	p-hodn.	Toler.	1-toler. R^2
B23x1-RELWT	0,117426	0,925075	5,50754	0,005012	0,745200	0,254800
B23x2-GLUFAST	0,149708	0,725594	25,71637	0,000000	0,149227	0,850773
B23x3-GLUTEST	0,199867	0,543500	57,11504	0,000000	0,143100	0,856900
B23x4-INSTEST	0,122473	0,886949	8,66731	0,000286	0,813954	0,186046
B23x5-SSPG	0,112112	0,968922	2,18109	0,116853	0,562221	0,437779

Test statistické významnosti dotyčného diskriminátoru

N=150	Výsledky diskriminační funkční analýzy (73Iris) Počet prom. v modelu: 4; grupovací: Trida (3 skup) Wilk. lambda: ,02344 přibliž F (8,288)=199,15 p<0,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,144)	p-hodn.	Toler.	1-toler. R^2
SEPALLENGTH	0,024976	0,938464	4,72115	0,010329	0,347993	0,652007
SEPALWIDTH	0,030580	0,766480	21,93593	0,000000	0,608859	0,391141
PETALLENGTH	0,035025	0,669206	35,59018	0,000000	0,365126	0,634874
PETALWIDTH	0,031546	0,743001	24,90433	0,000000	0,649314	0,350686

Příklad 4.23 *Logistická diskriminace k určení rakoviny prostaty*

Rakovina lymfatických uzlin. Rozhodující metodou vyšetření je laparotomie, vyjádřená proměnnou *B412x6,* kde 0 značí nepřítomnost rakoviny a 1 přítomnost rakoviny. Brownovo vyšetření pěti diskriminantů u 53 pacientů mělo nahradit toto obtížnější vyšetření:

Data:

B412x1 značí věk pacienta,
B412x2 je hladina sérové kyselé fosfatázy v Kingových-Armstrongových jedn.,
B412x3 rentgenového vyšetření: 0 značí nepřítomnost a 1 přítomnost tumoru,
B412x4 velikost tumoru rektálně: 0 = malý, 1 = velký,
B412x5 závěr patologického bodování z biopsie: 0 méně vážný, 1 velmi vážný.

B412x1	***B412x2***	***B412x3***	***B412x4***	***B412x5***	***B412x6***
66	0.48	0	0	0	0
..	..	..	..	..	..
..	..	..	..	..	..
64	0.89	1	1	0	1
68	1.26	1	1	1	1

Řešení:

1) **Analyzovaný výběr**: odhady parametrů (směrodat. odchylky v závorce) jsou b_0 1.52 (3.56), b_1 0.10 (0.06), b_2 2.64 (1.33), b_3 1.68 (0.80), b_4 2.04 (0.83), b_5 0.35 (0.80).

Dvě třídy: 1. třída (rakovinu má), 2. třída (rakovinu nemá).

Pacient rakovinu lymfatických uzlin má (je zařazen do 1. třídy), je-li splněna nerovnost

$$1.52 - 0.10\, x_1 + 2.64\, x_2 + 1.68\, x_3 + 2.04\, x_4 + 0.35\, x_5 > 0.$$

vykazuje-li pacient **kladnou hodnotu** v této nerovnosti, má rakovinu uzlin a je zařazen do 1. třídy, zatímco vykazuje-li **zápornou hodnotu** v této nerovnosti, nemá rakovinu a je zařazen do 2. třídy.

2) **Klasifikovaný výběr:** dosadíme hodnoty prvního pacienta do této nerovnosti a vyčíslíme

$$1.52 - 0.10 \times 66 + 2.64 \times 0.48 + 1.68 \times 0 + 2.04 \times 0 + 0.35 \times 0 = \mathbf{-3.81.}$$

Závěr:

Protože **výsledek v nerovnosti -3.81 je záporný**, je pacient zařazen do 2. třídy a **nemá rakovinu** lymfatických uzlin.

Příklad 4.24 *Botanické třídění kosatců diskriminační analýzou*

Ve Fisherově úloze o rozměrech okvětních lístků u 150 kosatců se analyzují květy tří základních tříd: (1) *Iris setosa,* (2) *Iris versicolor,* (3) *Iris virginica.* Květy kosatců jsou popsány čtyřmi znaky: délkou kališních lístků v mm anglicky *lsepal* a jejich šířkou *wsepal,* dále délkou korunních plátků v mm *lpetal* a jejich šířkou *wpetal*.

Každý objekt je popsán $p = 4$ znaky, a to *SepalLength, SepalWidth, PetalLength, PetalWidth*.

Data: Data jsou v pořadí proměnných *E418a, E418b, E418c, E418d:*

5.1	3.5	1.4	0.2	5.4	3.9	1.3	0.4	5.2	4.1	1.5	0.1	5.3	3.7	1.5	0.2
4.9	3.0	1.4	0.2	5.1	3.5	1.4	0.3	5.5	4.2	1.4	0.2	5.0	3.3	1.4	0.2
4.7	3.2	1.3	0.2	5.7	3.8	1.7	0.3	4.9	3.1	1.5	0.2	7.0	3.2	4.7	1.4
4.6	3.1	1.5	0.2	5.1	3.8	1.5	0.3	5.0	3.2	1.2	0.2	6.4	3.2	4.5	1.5
5.0	3.6	1.4	0.2	5.4	3.4	1.7	0.2	5.5	3.5	1.3	0.2	6.9	3.1	4.9	1.5
5.4	3.9	1.7	0.4	5.1	3.7	1.5	0.4	4.9	3.6	1.4	0.1	5.5	2.3	4.0	1.3
4.6	3.4	1.4	0.3	4.6	3.6	1.0	0.2	4.4	3.0	1.3	0.2	6.5	2.8	4.6	1.5
5.0	3.4	1.5	0.2	5.1	3.3	1.7	0.5	5.1	3.4	1.5	0.2	5.7	2.8	4.5	1.3
4.4	2.9	1.4	0.2	4.8	3.4	1.9	0.2	5.0	3.5	1.3	0.3	6.3	3.3	4.7	1.6
4.9	3.1	1.5	0.1	5.0	3.0	1.6	0.2	4.5	2.3	1.3	0.3	4.9	2.4	3.3	1.0
5.4	3.7	1.5	0.2	5.0	3.4	1.6	0.4	4.4	3.2	1.3	0.2	6.6	2.9	4.6	1.3
4.8	3.4	1.6	0.2	5.2	3.5	1.5	0.2	5.0	3.5	1.6	0.6	5.2	2.7	3.9	1.4
4.8	3.0	1.4	0.1	5.2	3.4	1.4	0.2	5.1	3.8	1.9	0.4	5.0	2.0	3.5	1.0
4.3	3.0	1.1	0.1	4.7	3.2	1.6	0.2	4.8	3.0	1.4	0.3	5.9	3.0	4.2	1.5
5.8	4.0	1.2	0.2	4.8	3.1	1.6	0.2	5.1	3.8	1.6	0.2	6.0	2.2	4.0	1.0
5.7	4.4	1.5	0.4	5.4	3.4	1.5	0.4	4.6	3.2	1.4	0.2	6.1	2.9	4.7	1.4

Kosatce (Iris)

Iris Setosa

Wild Iris - *Iris setosa*

Iris Versicolor

Iris Virginica

Kosatce (Iris)

Iris Setosa

Iris Versicolor

Iris Versicolor

Iris Virginica

Iris Virginica

Iris Virginica

Řešení: zdrojová matice 150 řádků a 4 sloupce neobsahuje žádné chybějící prvky.

1. Znaky

1.1 Aritmetický průměr [mm] objektů ve třídě Setosa G_1, Versicolor G_2, Virginica G_3.

Znak	**Setosa G_1**	**Versicolor G_2**	**Virginica G_3**	**Celkově $G_1+G_2+G_3$**
SepalLength	50.06	59.36	65.88	58.43333
SepalWidth	34.28	27.7	29.74	30.57333
PetalLength	14.62	42.6	55.52	37.58
PetalWidth	2.46	13.26	20.26	11.99333
Počet	50	50	50	150

Tabulka obsahuje průměry každého znaku pro každou třídu kosatců G_1, G_2 a G_3.

1.2 Směrodatné odchylky [mm] objektů ve třídě Setosa G_1, Versicolor G_2, Virginica G_3.

Znak	Setosa G_1	Versicolor G_2	Virginica G_3	Celkově $G_1+G_2+G_3$
SepalLength	3.524897	5.161712	6.358796	8.280662
SepalWidth	3.790644	3.137983	3.224966	4.358663
PetalLength	1.73664	4.69911	5.518947	17.65298
PetalWidth	1.053856	1.977527	2.7465	7.622377
Počet	50	50	50	150

Lineární diskriminační analýza předpokládá, že kovarianční matice jsou stejné pro každou třídu. Tabulka posoudí předpoklad, zda jsou směrodatné odchylky v třídách zhruba stejné.

1.5 Vnitrotřídní korelace a kovariance znaků.

Znak	SepalLength	SepalWidth	PetalLengthPetalWidth
SepalLength	26.50082	9.272109	16.751433.840136
SepalWidth	0.530236	11.53878	5.5243543.27102
PetalLength	0.756164	0.377916	18.518784.266531
PetalWidth	0.364506	0.470535	0.4844594.188163

V datech byly třídní průměry objektů odečteny. Korelace jsou v dolní levé části, vnitrotřídní kovariance jsou na diagonále a v pravé horní části matice.

1.6 Vyšetření vlivu jednotlivých znaků.

	Při odstranění dotyčného znaku		Pro dotyčný samotný znak		R^2		
Diskrimin.	Wilk. λ	*F*-test	Spočt. α	Wilk. λ	*F*-test	Spočt. α	Ostatní *X*
SepalLength	0.938463	4.720	0.010329	0.381294	119.26	0.000000	0.858612
SepalWidth	0.766480	21.94	0.000000	0.599217	49.16	0.000000	0.524007
PetalLength	0.669206	35.59	0.000000	0.058628	1180.2	0.000000	0.968012
PetalWidth	0.743001	24.90	0.000000	0.071117	960.01	0.000000	0.937850

Vliv jednotlivých znaků na diskriminační analýzu: *spočtená hladina významnosti při odstranění dotyčného znaku* je vypočtená u *F*-testu při odstranění dotyčného znaku. Znak je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$. Zde jsou všechny čtyři znaky statisticky významné. *Wilk. λ pro dotyčný samotný znak* značí hodnotu za použití jediného znaku, který je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$.

1.7 Automatická volba účinných znaků

Úkolem v diskriminační analýze je výběr znaků. Z velké skupiny možných znaků je třeba vybrat menší výběr těch nejlepších, maximálně 8 proměnných.

Krok	Činnost v kroku	Diskrimin.	% změny lambda	v *F*-test	Spočtená hladina α	Wilkovo lambda
0	Žádné					1.000000
1	Zavedeno	PetalLength	94.14	1180.16	0.000000	0.058628
2	Zavedeno	SepalWidth	37.09	43.04	0.000000	0.036884
3	Zavedeno	PetalWidth	32.29	34.57	0.000000	0.024976
4	Zavedeno	SepalLength	6.15	4.72	0.010329	0.023439
.. .		.....			...	.....

Detail ve 4. kroku automatického volby znaku:

Status	Diskrimin.	% změny lambda	v *F*-test	Spočtená hladina α	R^2 na ostatních *X*
Uvnitř	SepalLength	6.15	4.72	0.010329	0.858612
Uvnitř	SepalWidth	23.35	21.94	0.000000	0.524007
Uvnitř	PetalLength	33.08	35.59	0.000000	0.968012
Uvnitř	PetalWidth	25.70	24.90	0.000000	0.937850

Celkové Wilkovo $\lambda = 0.023439$

Automatický výběr znaků se plní krokově - nejprve se nalezne nejlepší znak a potom druhý nejlepší. Postupný krokový proces přidávání nejlepšího zbývajícího znaku s ověřením, zda by jeden aktivní znak mohl být odebrán, pokračuje, dokud je ještě nějaký znak k dispozici. *% změny v lambda* značí procento snížení v hodnotě lambda, které je výsledkem tohoto kroku. Wilkovo kritérium λ je analogické (1 - R^2) ve vícenásobné regresi. Vhodnější model sníží Wilkovo kritérium λ .

2. Diskriminační funkce.

2.1 Odhady diskriminačních koeficientů $a_0, a_1, \ldots, a_p$ diskriminační funkce Z_k pro třídy Setosa G_1, Versicolor G_2, Virginica G_3.

Znak	Setosa G_1	Versicolor G_2	Virginica G_3
Absolutní člen	-85.20985	-71.754	-103.2697
SepalLength	2.354417	1.569821	1.244585
SepalWidth	2.358787	0.707251	0.3685279
PetalLength	-1.643064	0.5211451	1.276654
PetalWidth	-1.739841	0.6434229	2.107911

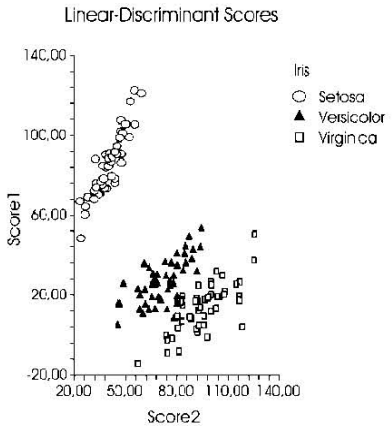
Tabulka obsahuje celkem tři diskriminační funkce, jednu pro každou třídu ve sloupci. Tato diskriminační funkce je určena k zařazování nových, dosud nezařazených objektů do tříd čili ke klasifikaci, a proto se v této formě nazývá *klasifikační funkce.*

Klasifikační funkce *pro první třídu* G_1: -85.20985+2.35442*SepalLength +2.35879*SepalWidth-1.64306*PetalLength-1.73984*PetalWidth.

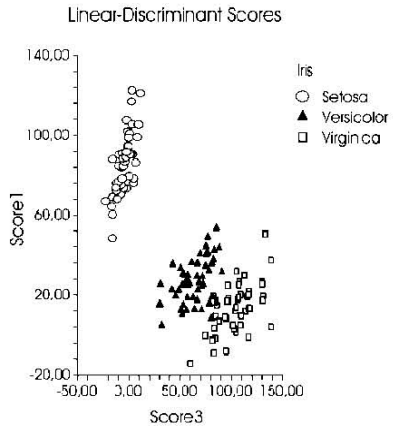
Tyto tři funkce mohou být použity k zařazení nových kosatců do navržených tříd.

6. Klasifikační grafy.

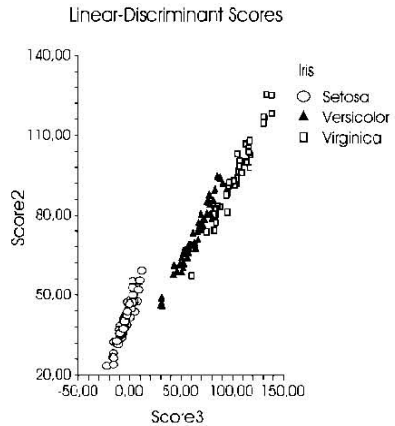
Užijeme grafy: (a) lineárních diskriminačních skóre, (b) regresních skóre nebo (c) kanonických skóre.



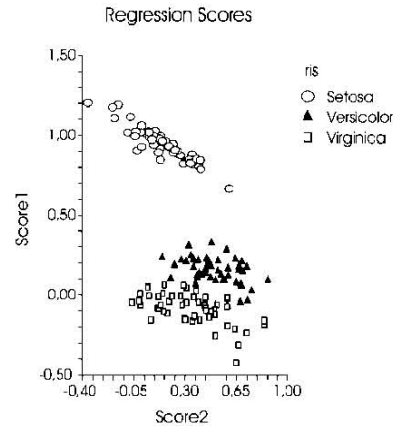
Obr. 7.9a Graf lineárního diskriminačního skóre 1 a 2 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



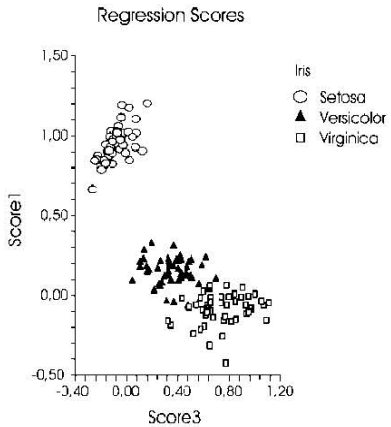
Obr. 7.9b Graf lineárního diskriminačního skóre 1 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



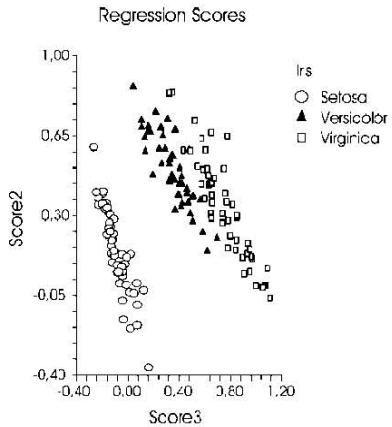
Obr. 7.9c Graf lineárního diskriminačního skóre 2 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



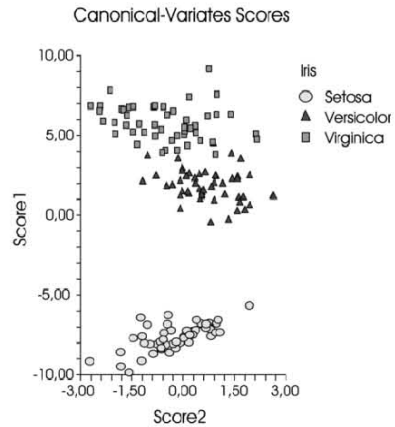
Obr. 7.9d Graf regresního skóre 1 a 2 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



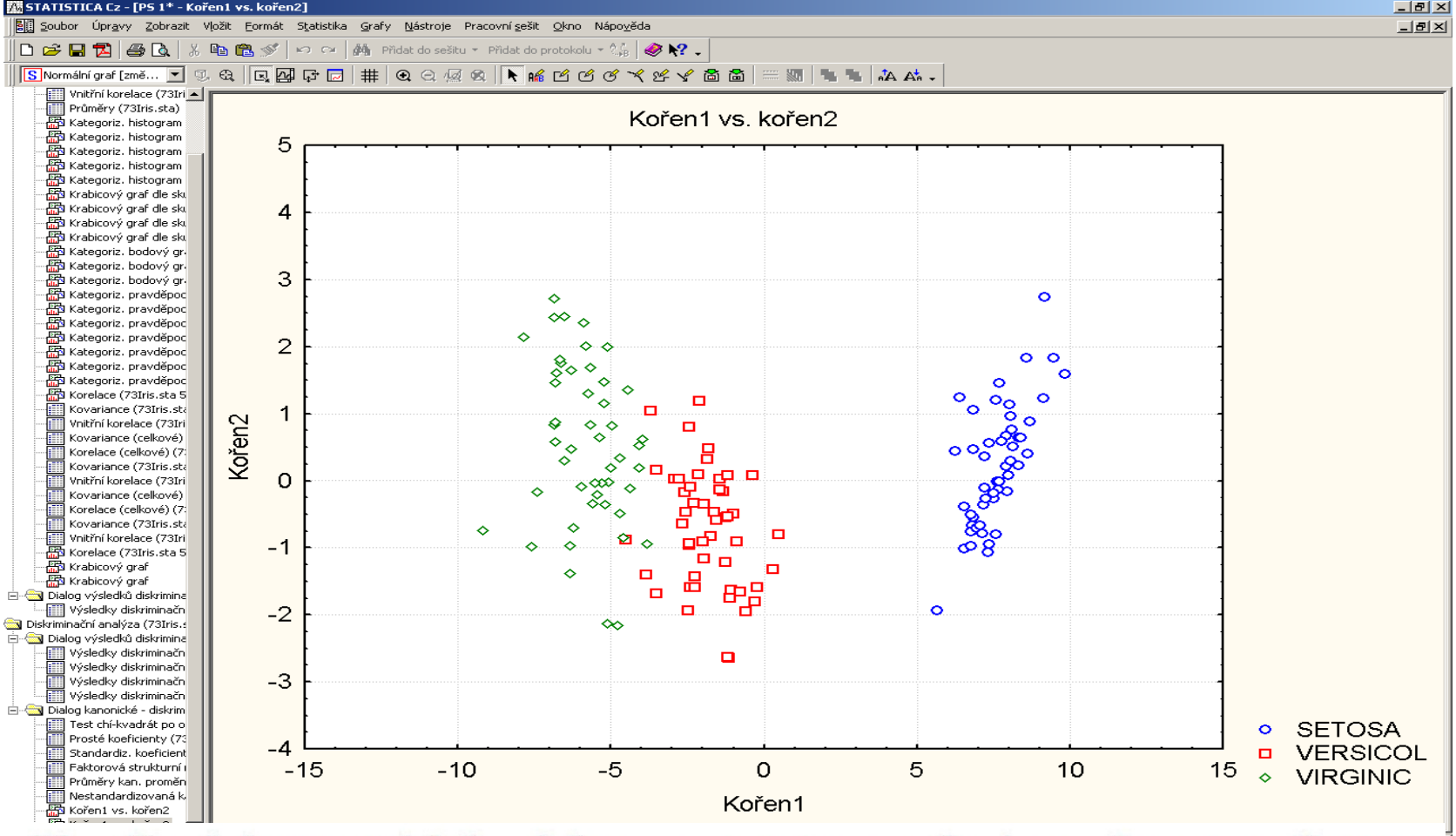
Obr. 7.9e Graf regresního skóre 1 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



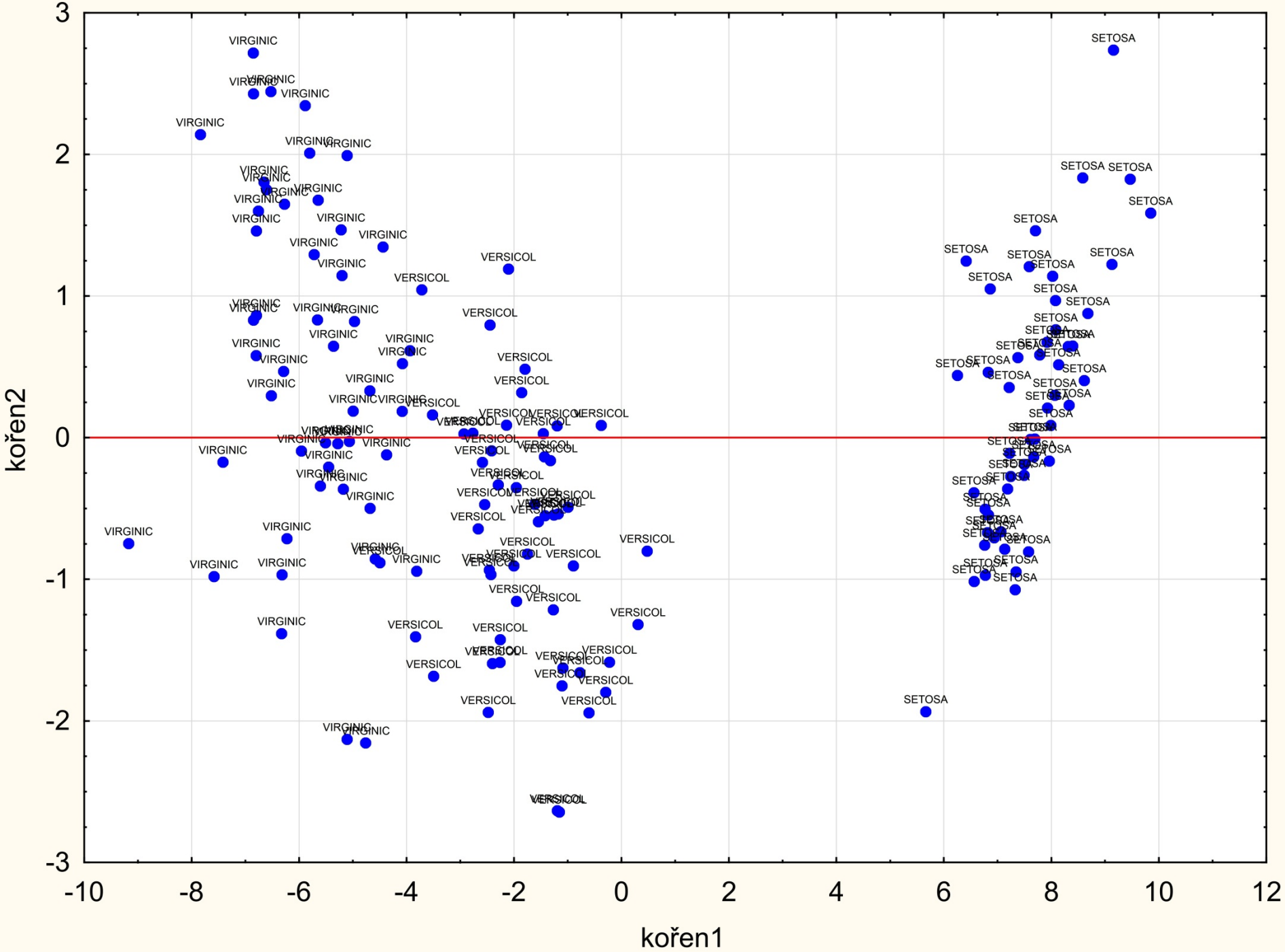
Obr. 7.9f Graf regresního skóre 2 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



Obr. 7.9g Graf kanonických proměnných 1 a 2 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



58. *Graf kanonických skóre:* v tomto grafu jsou kosatce Setosa zobrazeny modře daleko vpravo od ostatních. Kořen 1 (čili první diskriminační funkce) silně diskriminuje Setosu od ostatních kosatců. Kořen 2 (druhá diskriminační funkce) poskytuje diskriminaci mezi Versicolor se zápornými hodnotami Kořene 2 a ostatními kosatci, převážně kladného Kořene 2. Diskriminace však není tak zřetelná jako je tomu u prvního Kořene 1.


kořen2
kořen1
3
2
1
0
-1
-2
-3
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
VIRGINIC
VERSICOL
SETOSA

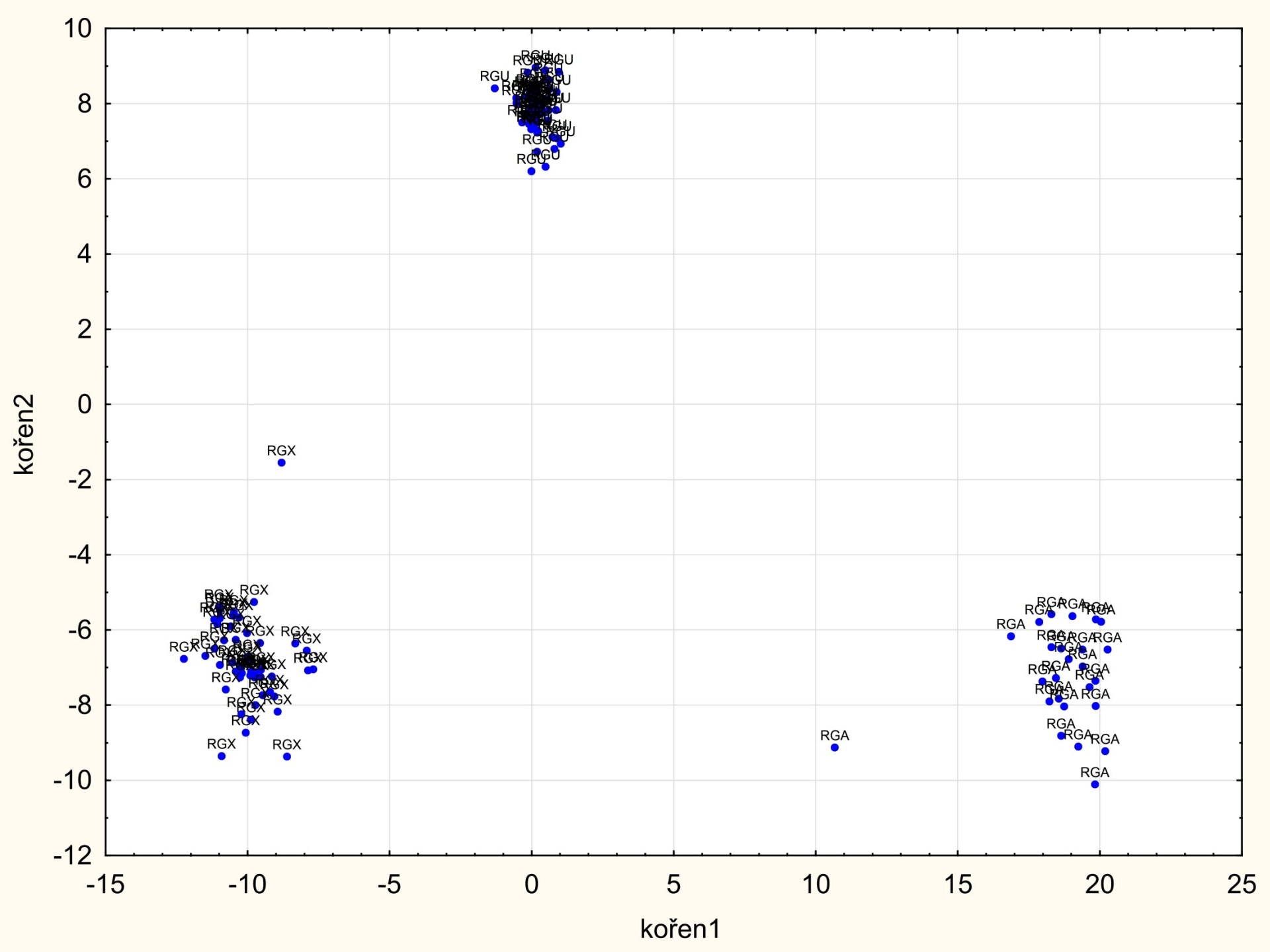
Úloha C32 ***Klasifikace a struktura znaků oxidu titaničitého dle chemických a fyzikálních vlastností*** **(DA)**

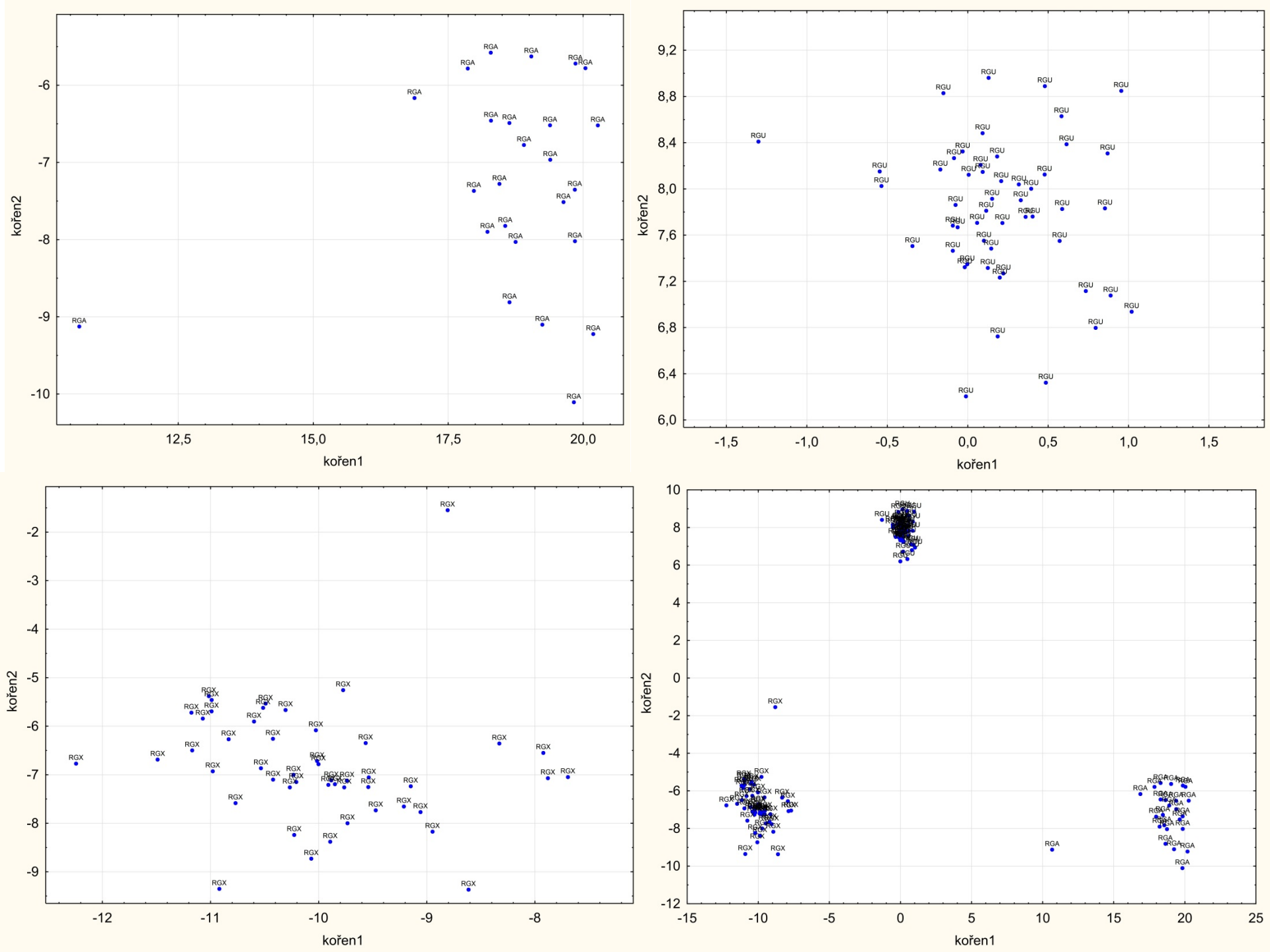
Výsledky analýz tří skupin bílého pigmentu oxidu titaničitého. Proveďte klasifikaci diskriminační analýzou (DA).

◯ ***Data:*** Popis znaků ve sloupcích: Data **ID**, Zařazení do tříd **Klas**, tj. typy RGA, RG*X*, RGU tvoří tréninková data. Diskriminátory:

Ti obsah Ti (%),
TL obsah těkavých látek (%),
SpO je spotřeba oleje (g/100 g pigmentu),
Barv značí barvivost,
Podt značí podtón,
Si značí obsah Si (%),
Al značí obsah Al (%).

ID	Ti	TL	SpO	Barv	Podt	Si	Al	Klas
1	93.78	0.5	24.2	1870	10	2.174	3.511	RGA
.....	.....	.....	.....	.....	.....	.....	.....	.....
144	98.21	0.2	19	1860	13	0.354	1.014	RGX


kořen2
kořen1
10
8
6
4
2
0
-2
-4
-6
-8
-10
-12
-15
-10
-5
0
5
10
15
20
25
RGU
RGX
RGA


kořen2
kořen1
RGA
RGU
RGX

N=144	Výsledky diskriminační funkční analýzy (C32) Počet prom. v modelu: 7; grupovací: klas (3 skup) Wilk. lambda: ,00018 přibliž F (14,270)=1436,8 p<0,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,135)	p-hodn.	Toler.	1-toler. R^2
Ti	*0,000268*	*0,654340*	*35,65741*	*0,000000*	*0,019606*	*0,980394*
TL	0,000177	0,988528	0,78338	0,458925	0,892277	0,107723
SpO	*0,000210*	*0,836343*	*13,20851*	*0,000006*	*0,538857*	*0,461143*
Barv	0,000177	0,989290	0,73079	0,483425	0,735078	0,264922
Podt	*0,000223*	*0,785127*	*18,47337*	*0,000000*	*0,903931*	*0,096069*
Si	*0,000205*	*0,857679*	*11,20079*	*0,000032*	*0,134136*	*0,865864*
Al	*0,000350*	*0,501522*	*67,09041*	*0,000000*	*0,029380*	*0,970620*

Kořeny odstraněny	Test chí-kvadrát po odstranění post. kořenů (C32)					
	Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	99,74799	0,995025	0,000175	1193,457	14	0,00
1	55,57775	0,991123	0,017675	556,915	6	0,00

Případ	Klasifikace případů (C32) Nesprávná klasifikace je označena *			
	Pozorova Klasif.	1 p=,18056	2 p=,47222	3 p=,34722
2	RGA	RGA	RGU	RGX
3	RGA	RGA	RGU	RGX
4	RGA	RGA	RGU	RGX
5	RGA	RGA	RGU	RGX
6	RGA	RGA	RGU	RGX
7	RGA	RGA	RGU	RGX
8	RGA	RGA	RGU	RGX
9	RGA	RGA	RGU	RGX
10	RGA	RGA	RGU	RGX
11	RGA	RGA	RGU	RGX
12	RGA	RGA	RGU	RGX
13	RGA	RGA	RGU	RGX
14	RGA	RGA	RGU	RGX
15	RGA	RGA	RGU	RGX
16	RGA	RGA	RGU	RGX
17	RGA	RGA	RGU	RGX
18	RGA	RGA	RGU	RGX
19	RGA	RGA	RGU	RGX
20	RGA	RGA	RGU	RGX
21	RGA	RGA	RGU	RGX
22	RGA	RGA	RGU	RGX
23	RGA	RGA	RGU	RGX
24	RGA	RGA	RGU	RGX
25	RGA	RGA	RGU	RGX
26	RGA	RGA	RGU	RGX
27	RGA	RGA	RGU	RGX
28	RGU	RGU	RGA	RGX
29	RGU	RGU	RGA	RGX
30	RGU	RGU	RGA	RGX
31	RGU	RGU	RGA	RGX
32	RGU	RGU	RGA	RGX
33	RGU	RGU	RGA	RGX
34	RGU	RGU	RGA	RGX
35	RGU	RGU	RGA	RGX
36	RGU	RGU	RGA	RGX
37	RGU	RGU	RGA	RGX
38	RGU	RGU	RGA	RGX
39	RGU	RGU	RGA	RGX
40	RGU	RGU	RGA	RGX
41	RGU	RGU	RGA	RGX
42	RGU	RGU	RGA	RGX
43	RGU	RGU	RGA	RGX
44	RGU	RGU	RGA	RGX
45	RGU	RGU	RGA	RGX
46	RGU	RGU	RGA	RGX
47	RGU	RGU	RGA	RGX
48	RGU	RGU	RGA	RGX
49	RGU	RGU	RGA	RGX
50	RGU	RGU	RGA	RGX
51	RGU	RGU	RGA	RGX

Případ	Mahalanobisovy vzdálenosti (C32) Nesprávná klasifikace je označena *			
	Pozorova Klasif.	RGA p=,18056	RGU p=,47222	RGX p=,34722
2	RGA	10,1351	623,1396	900,3110
3	RGA	10,0932	541,3123	852,3933
4	RGA	8,0935	500,6362	783,0421
5	RGA	5,0655	574,8432	802,9057
6	RGA	4,8171	614,9479	883,0971
7	RGA	8,5472	550,8557	828,4168
8	RGA	7,8208	536,6371	808,1469
9	RGA	23,7401	525,8623	823,0444
10	RGA	3,3003	562,5682	813,3089
11	RGA	3,3003	562,5682	813,3089
12	RGA	8,3897	641,5818	899,2757
13	RGA	6,4779	599,4259	834,2990
14	RGA	10,8236	590,3975	827,4419
15	RGA	5,9304	590,4635	869,6200
16	RGA	79,4508	407,4983	444,9338
17	RGA	8,4340	612,0465	922,2669
18	RGA	8,1832	572,3218	897,4604
19	RGA	4,8689	575,7915	867,9421
20	RGA	9,1453	581,9042	909,0862
21	RGA	7,2044	619,5528	829,3455
22	RGA	4,3358	565,1535	839,9256
23	RGA	11,4189	479,1127	730,1782
24	RGA	4,1241	548,6656	787,1074
25	RGA	7,0487	650,5714	864,3236
26	RGA	10,4597	692,3977	921,9067
27	RGA	13,0826	708,4438	904,7653
28	RGU	521,2330	9,3222	318,7957
29	RGU	536,7763	8,9696	275,9360
30	RGU	536,7763	8,9696	275,9360
31	RGU	543,4659	7,2903	293,4542
32	RGU	519,7687	6,6857	287,3059
33	RGU	519,7687	6,6857	287,3059
34	RGU	593,7277	10,4871	342,0955
35	RGU	575,4103	4,3821	331,6899
36	RGU	648,0285	6,1246	311,1339
37	RGU	648,0285	6,1246	311,1339
38	RGU	563,7344	1,9133	301,4111
39	RGU	530,8040	2,9600	311,8223
40	RGU	530,8040	2,9600	311,8223
41	RGU	530,8040	2,9600	311,8223
42	RGU	615,5059	3,3827	344,1681
43	RGU	558,3248	3,6562	330,2850
44	RGU	553,2376	8,2193	340,4370
45	RGU	550,5502	3,6501	321,8599
46	RGU	581,4843	2,4549	352,4435
47	RGU	581,4843	2,4549	352,4435
48	RGU	567,2060	0,1514	327,9353
49	RGU	582,9557	0,3958	326,2716
50	RGU	594,5043	1,5918	322,6992

Případ	Aposteriorní pravděpodobnosti (C32) Nesprávná klasifikace je označena *			
	Pozorova Klasif.	RGA p=,18056	RGU p=,47222	RGX p=,34722
2	RGA	1,000000	0,000000	0,000000
3	RGA	1,000000	0,000000	0,000000
4	RGA	1,000000	0,000000	0,000000
5	RGA	1,000000	0,000000	0,000000
6	RGA	1,000000	0,000000	0,000000
7	RGA	1,000000	0,000000	0,000000
8	RGA	1,000000	0,000000	0,000000
9	RGA	1,000000	0,000000	0,000000
10	RGA	1,000000	0,000000	0,000000
11	RGA	1,000000	0,000000	0,000000
12	RGA	1,000000	0,000000	0,000000
13	RGA	1,000000	0,000000	0,000000
14	RGA	1,000000	0,000000	0,000000
15	RGA	1,000000	0,000000	0,000000
16	RGA	1,000000	0,000000	0,000000
17	RGA	1,000000	0,000000	0,000000
18	RGA	1,000000	0,000000	0,000000
19	RGA	1,000000	0,000000	0,000000
20	RGA	1,000000	0,000000	0,000000
21	RGA	1,000000	0,000000	0,000000
22	RGA	1,000000	0,000000	0,000000
23	RGA	1,000000	0,000000	0,000000
24	RGA	1,000000	0,000000	0,000000
25	RGA	1,000000	0,000000	0,000000
26	RGA	1,000000	0,000000	0,000000
27	RGA	1,000000	0,000000	0,000000
28	RGU	0,000000	1,000000	0,000000
29	RGU	0,000000	1,000000	0,000000
30	RGU	0,000000	1,000000	0,000000
31	RGU	0,000000	1,000000	0,000000
32	RGU	0,000000	1,000000	0,000000
33	RGU	0,000000	1,000000	0,000000
34	RGU	0,000000	1,000000	0,000000
35	RGU	0,000000	1,000000	0,000000
36	RGU	0,000000	1,000000	0,000000
37	RGU	0,000000	1,000000	0,000000
38	RGU	0,000000	1,000000	0,000000
39	RGU	0,000000	1,000000	0,000000
40	RGU	0,000000	1,000000	0,000000
41	RGU	0,000000	1,000000	0,000000
42	RGU	0,000000	1,000000	0,000000
43	RGU	0,000000	1,000000	0,000000
44	RGU	0,000000	1,000000	0,000000
45	RGU	0,000000	1,000000	0,000000
46	RGU	0,000000	1,000000	0,000000

Postup analýzy vícerozměrných dat

1. Standardizace: analýze vždy předchází standardizace čili škálování proměnných.

2. Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi proměnnými:

Vyčíslení výběrové střední hodnoty každé proměnné.

Odhad kovarianční matice ***S*** a její normované podoby - korelační matice ***R*.**

Odhadu vícerozměrné šikmosti a vícerozměrné špičatosti.

Matice ***R*** obsahuje Pearsonovy párové korelační koeficienty, které se diskutují.

3. Exploratorní analýza dat EDA:

(a) Hledání podobnosti objektů vizuálními rozptylovými diagramy typu casement plot, draftsman plot, dále symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy),

(b) Nalezení vybočujících objektů nebo vybočujících proměnných, mnohdy nevhodných k analýze,

(c) Testy předpokladů lineárních vazeb,

(d) Testy předpokladů o datech (normalitu, nekorelovanost, homogenitu).
Ověřování normality založené na vícerozměrné šikmosti a vícerozměrné špičatosti.

4. Určení vhodného počtu latentních proměnných:

a) Matice ***S*** nebo ***R*** se rozloží na vlastní čísla a vlastní vektory.

b) Indexový graf úpatí vlastních čísel (Scree plot): určí vhodný počet latentních proměnných, které ještě dostatečně popisují proměnlivost v datech.

c) Když se latentní proměnné podaří pojmenovat a dát jim i fyzikální, biologický či jiný věcný význam, jedná se o faktory. Jinak jde o hlavní komponenty.

5. Určení struktury v proměnných (PCA a FA):

a) Graf komponentních vah (Plot of components weights, loadings): hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu

b) Rozptylový diagram komponentního skóre (Scatterplot): hledání struktury v objektech a třídění objektů do shluků.

c) Dvojný graf (Biplot) je přehledným spojením obou předešlých grafů a ukáže interakci objektů a proměnných.

6. Určení struktury a vzájemných vazeb v objektech:

a) Klasifikační postupy zařadí analyzovaný objekt do jednoho již existujícího a předem zadaného shluku.

b) Neutříděnou skupinu objektů lze uspořádat do shluků a výsledek třídění zobrazit dendrogramem v analýze shluků. V hierarchickém postupu je třeba k vytvoření shluků vybrat vzdálenost mezi objekty (Eukleidovskou, Manhattanovskou, Mahalanobisovu) a jednu z nabídnutých metod: průměrovou, centroidní, nejbližšího souseda, nejvzdálenějšího souseda, mediánovou, Wardovu.

c) Nehierarchické postupy rozdělí objekty do shluků, v nichž jsou předem umístěni typičtí reprezentanti.

7. Vysvětlení souladu nalezené struktury objektů a vzájemných vazeb v dendrogramu a PCA (či FA) grafech:

a) Vyšetřit a vysvětlit nalezenou strukturu a vazby jednotlivých proměnných nalezenou jednak v PCA (či FA) a jednak v dendrogramu podobnosti proměnných analýzou vzniklých shluků.

b) Vysvětlit strukturu a vazby klasifikovaných objektů nalezenou v PCA a v dendrogramu podobnosti objektů.

Děkuji za pozornost!

http://meloun.upce.cz