Univerzita Pardubice

Fakulta chemicko-technologická

Katedra analytické chemie

Semestrální práce

Licenční studium

Statistické zpracování dat při kontrole a řízení jakosti

předmět

3.1. Matematické principy analýzy vícerozměrných dat

leden 2008

Policie ČR, Odbor kriminalistické techniky a expertiz, České Budějovice

Ing. Petr Texl

Obsah

Otázka 1

Najděte vlastní čísla a vlastní vektory, determinant, stopu a odmocninu matice:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 2 \\ 3 & 2 & 10 \end{bmatrix}$$

A. Determinant matice (A)

Determinant je definován pouze pro čtvercové matice.

$$\det(\mathbf{A}) = 1 \det \begin{bmatrix} 8 & 2 \\ 2 & 10 \end{bmatrix} - 2 \det \begin{bmatrix} 2 & 2 \\ 3 & 10 \end{bmatrix} + 3 \det \begin{bmatrix} 2 & 8 \\ 3 & 2 \end{bmatrix}$$

$$\det(\mathbf{A}) = 1 \times 76 \; - \; 2 \times 14 \; + \; 3 \times (-20)$$

det (A) = - 12

B. Stopa matice (A)

Stopa tr (**A**) čtvercové matice **A** je součet jejích diagonálních elementů.

$$\text{tr } (\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

$$\text{tr } (\mathbf{A}) = 1 + 8 + 10$$

tr (A) = 19

C. Vlastní čísla matice (A)

Nechť **A** je čtvercová matice a uvažujeme soustavu rovnic $\mathbf{Av} = \lambda\mathbf{v}$. Hodnota λ, pro kterou má tato soustava řešení ($\mathbf{v} \neq \mathbf{0}$), se nazývá vlastní číslo matice **A**.

Platí: $\mathbf{Av} = \lambda\mathbf{v}$ $\mathbf{Av} - \lambda\mathbf{v} = 0$ $(\mathbf{A} - \lambda\mathbf{E}) = 0$ **E** je jednotková matice

Tato lineární soustava rovnic má řešení pokud je $\det(\mathbf{A} - \lambda\mathbf{E}) = 0$. Pro výpočet vlastních čísel je třeba tuto rovnici řešit.

$$\det(\mathbf{A} - \lambda\mathbf{E}) = \det \left| \begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 2 \\ 3 & 2 & 10 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = \det \begin{bmatrix} 1-\lambda & 2 & 3 \\ 2 & 8-\lambda & 2 \\ 3 & 2 & 10-\lambda \end{bmatrix}$$

$$\det(\mathbf{A} - \lambda\mathbf{E}) = 1 \det \begin{bmatrix} 8-\lambda & 2 \\ 2 & 10-\lambda \end{bmatrix} - 2 \det \begin{bmatrix} 2 & 2 \\ 3 & 10-\lambda \end{bmatrix} + 3 \det \begin{bmatrix} 2 & 8-\lambda \\ 3 & 2 \end{bmatrix}$$

$$\det(\mathbf{A} - \lambda\mathbf{E}) = (8 - \lambda) \times (10 - \lambda) - 4 - 2(20 - 2\lambda - 6) + 3(4 - 24 + 3\lambda)$$

$$\det(\mathbf{A} - \lambda\mathbf{E}) = \lambda^2 - 5\lambda - 12$$

Vlastní čísla potom získáme řešením kořenů této kvadratické rovnice.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$x = \frac{5 \pm \sqrt{-5^2 + 4 \times 12}}{2}$$

$$\mathbf{x_1 = 6{,}772} \qquad \mathbf{x_2 = -\,1{,}772}$$

Hodnoty $\mathbf{x_1}$ a $\mathbf{x_2}$ představují vlastní čísla matice.

Otázka 2

2.1. Roztřídění evropských druhových medů do skupin

Zadání:
Je třeba roztřídit čisté druhové evropské medy dle fyzikálně-chemických parametrů.

Podle rostlinného původu se dělí medy na nektarové a medovicové. Nektarové medy vznikají ze sladkých šťáv z nektarií rostlin působením včel. Medovicové medy vznikají také působením včel. Ty ale zpracovávají sladkou šťávu (medovici), kterou vylučují tzv. producenti medovice (mšice, červci a mery) zpravidla na listech a jehličí.

I když je včela medonosná přísně florokonstantní (nalétává pouze na jeden druh rostliny), získat čistý druhový med lze pouze ve velkých lánech a lesních monokulturách. Většinou včelaři získávají medy smíšené s převahou určitého druhu. Čisté druhové medy slouží jako určité standardy s definovanými fyzikálně-chemickými vlastnostmi.

Mezi vlastnosti, které se u medu zjišťují patří: barva (Pfundova stupnice), elektrická vodivost, optická otáčivost, kyselost, aktivita enzymu diastázy, obsah fruktózy, glukózy a vody. Důležité ukazatelé jsou dále: součet obsahu glukózy a fruktózy, poměr obsahu fruktózy a glukózy a poměr glukózy a vody.

Data: Použita data *medy01*

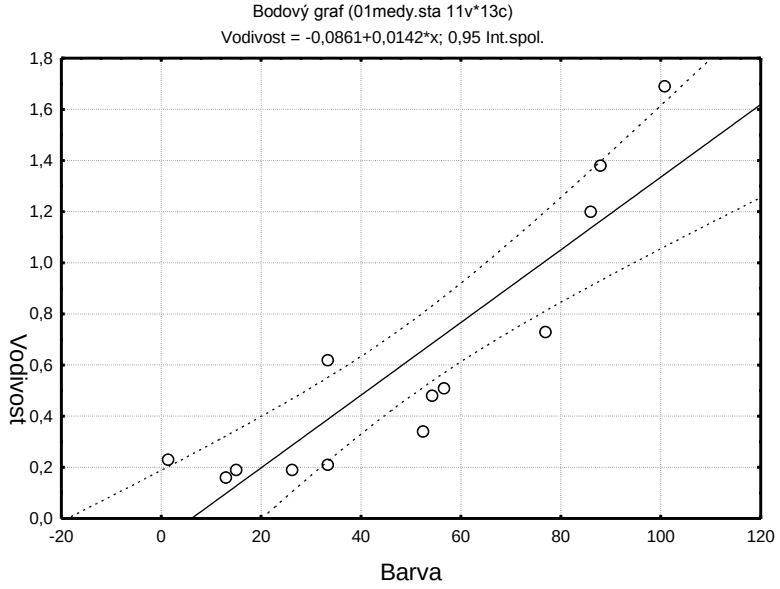
Složení evropských druhových medů (Persano Oddo a Piro, 2004)

Zdroj	Barva (mm Pfund)	Vodivost (mS/cm)	Rotace (stupně)	Kyselost (meq/kg)	Diastáza (DN)	Fruktóza (g/100 g)	Glukóza (g/100 g)	Fru+Gl (g/100 g)	Fru/Gl	Gl/voda
Akát *Robinia*	12,9	0,16	-16,6	11,2	10,5	42,7	26,5	69,2	1,61	1,57
Blahovičník *Eucalyptus*	54,2	0,48	-13,3	19,4	25,5	39,1	33,0	72,1	1,18	2,14
Citrusy *Citrus*	15,0	0,19	-13,4	14,3	9,6	38,7	31,4	70,1	1,23	1,92
Kaštan *Castanea*	87,9	1,38	-16,7	13,0	24,3	40,8	27,9	68,7	1,46	1,62
Levandule *Lavandula*	33,3	0,21	-8,3	17,3	14,1	36,0	30,6	66,6	1,18	1,88
Lípa *Tilia*	33,3	0,62	-12,5	20,8	16,8	37,5	31,9	69,4	1,18	1,93
Medovice *Metcalfa*	100,8	1,69	17,5	37,2	39,3	31,6	23,9	55,5	1,32	1,51
Medovice	86,0	1,20	13,9	26,0	22,6	32,5	26,2	58,7	1,24	1,61
Pampeliška *Taraxacum*	56,6	0,51	-10,0	10,9	11,3	37,4	38,0	75,4	0,98	2,33
Pěnišník *Rhododendron*	1,4	0,23	-5,8	13,3	12,1	39,1	30,4	69,5	1,29	1,79
Řepka *Brassica*	26,2	0,19	-10,3	22,1	26,9	38,3	40,5	78,8	0,95	2,37
Slunečnice *Helianthus*	52,4	0,34	-17,5	23,1	20,8	39,2	37,4	76,6	1,05	2,10
Vřes *Calluna*	76,9	0,73	-32,1	11,1	23,4	40,8	32,5	73,3	1,26	1,76

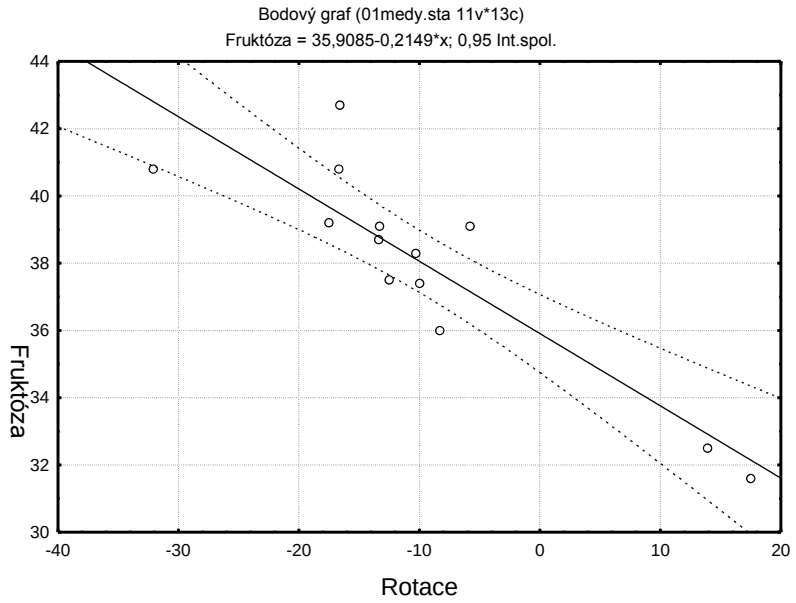
Program:
program STATISTICA

Řešení:

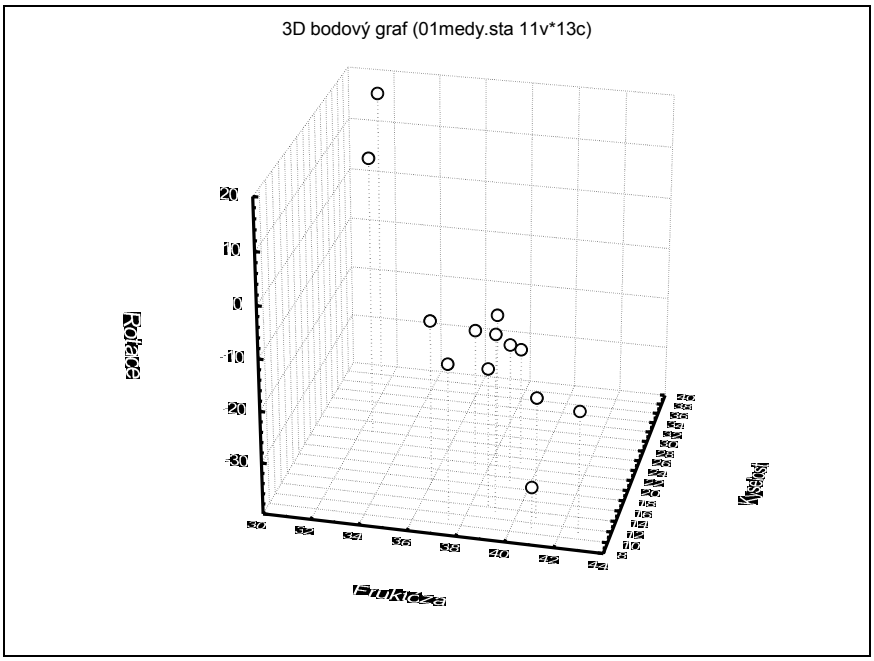
Odhalení struktury ve znacích a objektech



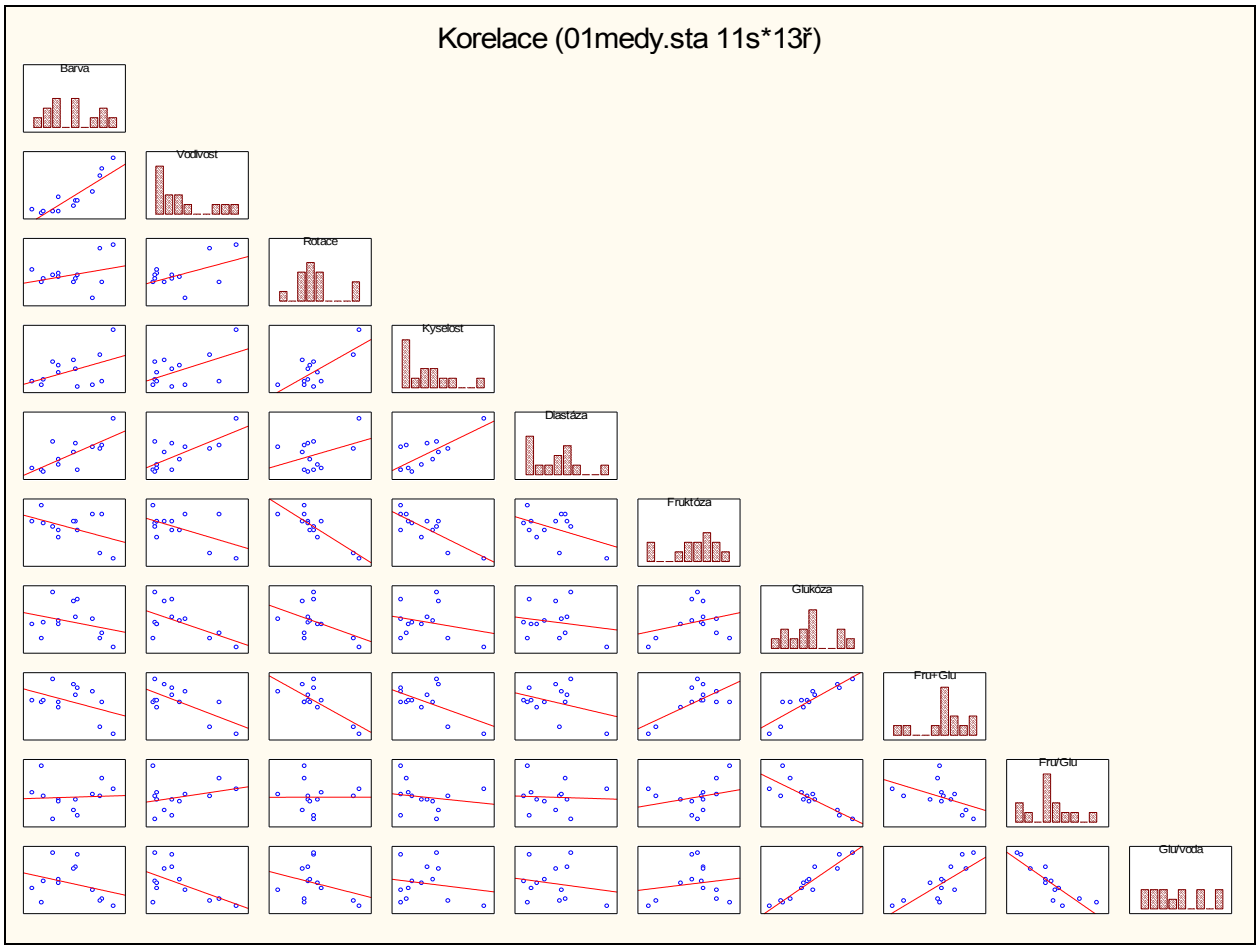
Rozptylový 2D-diagram Vodivost – Barva (STATISTICA)
ukazuje pozitivní korelaci



Rozptylový 2D-diagram Fruktóza – Rotace (STATISTICA)
naznačuje negativní korelaci



Rozptylový 3D-diagram znaků Rotace – Fruktóza – Kyselost (STATISTICA)
Graf naznačuje silnou korelaci a je zde patrná nová souřadnicová osa – hlavní komponenta (viz. metoda PCA)



Graf korelační matice (STATISTICA)

2.2. Analýza hlavních komponent (PCA)

Cílem metody je transformace dat z původních proměnných do menšího počtu latentních proměnných. Tyto proměnné mají vhodnější vlastnosti, je jich výrazně méně, vystihují téměř celou proměnlivost původních proměnných a jsou vzájemně nekorelované. Latentní proměnné jsou označovány jako hlavní komponenty

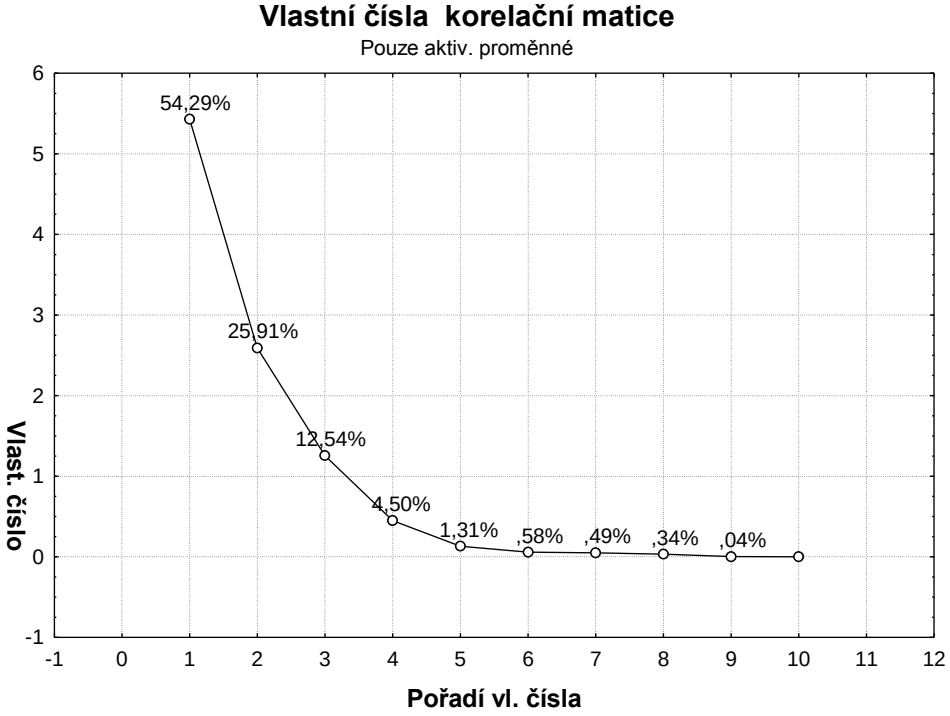
	Korelace (01medy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=13 (Celé případy vynechány u ChD)									
Proměnná	Barva	Vodivost	Rotace	Kyselost	Diastáza	Fruktóza	Glukóza	Fru+Glu	Fru/Glu	Glu/voda
Barva	1,00	0,89	0,32	0,46	0,73	-0,46	-0,31	-0,46	0,06	-0,32
Vodivost	0,89	1,00	0,52	0,54	0,72	-0,54	-0,58	-0,69	0,28	-0,56
Rotace	0,32	0,52	1,00	0,76	0,40	-0,89	-0,47	-0,78	0,00	-0,31
Kyselost	0,46	0,54	0,76	1,00	0,76	-0,80	-0,25	-0,57	-0,17	-0,17
Diastáza	0,73	0,72	0,40	0,76	1,00	-0,47	-0,19	-0,36	-0,05	-0,18
Fruktóza	-0,46	-0,54	-0,89	-0,80	-0,47	1,00	0,30	0,70	0,27	0,16
Glukóza	-0,31	-0,58	-0,47	-0,25	-0,19	0,30	1,00	0,89	-0,83	0,95
Fru+Glu	-0,46	-0,69	-0,78	-0,57	-0,36	0,70	0,89	1,00	-0,49	0,78
Fru/Glu	0,06	0,28	0,00	-0,17	-0,05	0,27	-0,83	-0,49	1,00	-0,86
Glu/voda	-0,32	-0,56	-0,31	-0,17	-0,18	0,16	0,95	0,78	-0,86	1,00

Matice korelačních koeficientů znaků zdrojové matice dat *medy01* (STATISTICA)

(červeně jsou vyznačeny statisticky významné korelace, kde hodnota p < 0,05.

	Vlastní čísla korelační matice a související statistiky (01medy.sta) Pouze aktiv. proměnné			
Pořadí vl.č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	5,429321	54,29321	5,42932	54,2932
2	2,590613	25,90613	8,01993	80,1993
3	1,254380	12,54380	9,27431	92,7431
4	0,449557	4,49557	9,72387	97,2387
5	0,131389	1,31389	9,85526	98,5526
6	0,057912	0,57912	9,91317	99,1317
7	0,049152	0,49152	9,96232	99,6232
8	0,034074	0,34074	9,99640	99,9640
9	0,003602	0,03602	10,00000	100,0000

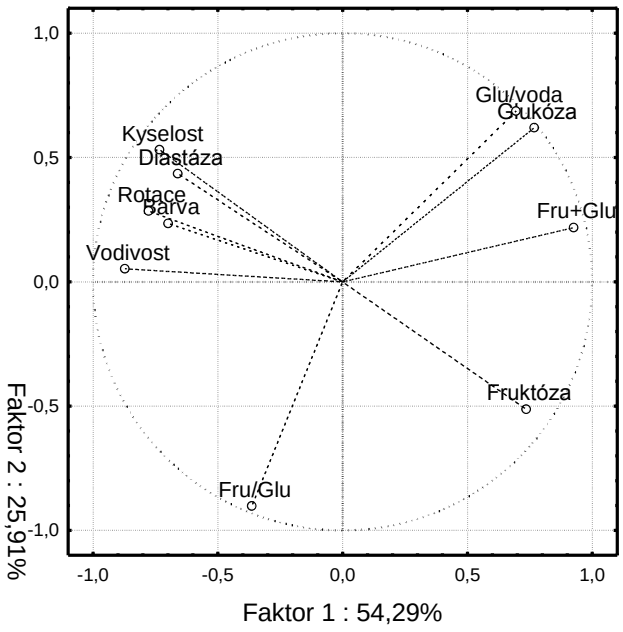
Tabelární podoba Cattelova indexového diagramu vlastních čísel (STATISTICA)
První dvě vlastní čísla pokrývají téměř 80,2 % proměnlivosti v datech.



Cattelův indexový graf úpatí vlastních čísel Scree Plot zdrojové matice dat medy01 (STATISTICA)

Podle Kaiserova kriteria 1.0 první tři hlavní komponenty dosahují vlastních čísel větších než 1.0 - tzn. že zdrojová matice je rovna 3. (Největší vlastní číslo pokrývá 54,29 % proměnlivosti v datech. Druhé největší vlastní číslo pokrývá 25,91 % proměnlivosti, což je dohromady 80,20 %.)

Protože graf nevykazuje znatelný zlom (koleno) mělo by se zahrnout i třetí největší vlastní číslo s pokrytím 12,54 % proměnlivosti. V tomto případě by bylo celkové pokrytí proměnlivosti v datech 92,74 %.

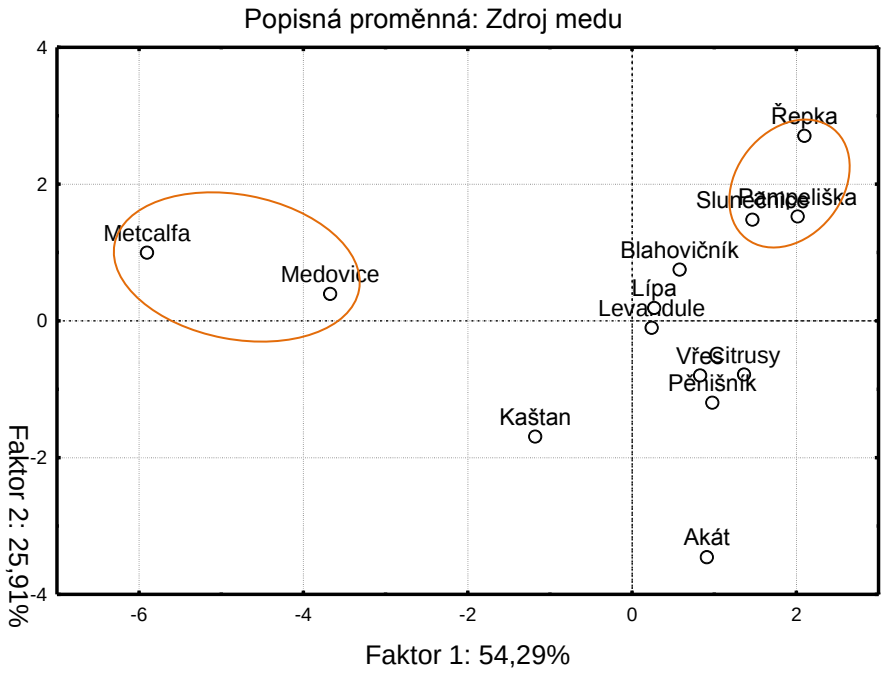


Graf komponentních vah znaků 1. a 2. hlavní komponenty (STATISTICA)

2.3. Shrnutí (popis grafu komponentních vah znaků 1. a 2. hlavní komponenty

Souřadnice obou hlavních komponent jednotlivých znaků představují vlastně korelační koeficient mezi znakem a hlavní komponentou. Záporné znaménko je u negativně korelovaných znaků (čím více tím méně) s 1. hlavní komponentou a to jsou znaky Vodivost, Barva, Rotace, Diastáza a Kyselost. Kladné je u pozitivně korelovaných znaků (čím více tím více) s 1. hlavní komponentou a to jsou znaky Fruktóza, Glukóza, Glu/voda, Fru+Glu.

2. hlavní komponenta negativně koreluje se znaky Fruktóza a Fru/Glu. Znaky blízko sebe značí, že znaky spolu korelují (Rotace-Barva, Kyselost-Diastáza, Glu/voda-Glukóza). Znaky vzdálené nekorelují a nejsou si podobné. Čím více se blíží poloha znaku k jednotkové kružnici, tím lepší je jeho zobrazení v tomto systému hlavních komponent. Komponentní průvodiče 3 znaků se dotýkají jednotkové korelační kružnice, 7 znaků dosahuje téměř hodnoty korelace 0,8 až 0,9 .



Rozptylový graf komponentního skóre komponenty 1 a 2 (STATISTICA)

Graf ukazuje na komponentní skóre všech zdrojů medu. Jsou patrné shluky druhových medů podle zdroje (rostliny). Medovicové medy (Medovice a Metcalfa) jsou negativně korelovány s 1. hlavní komponentou. Podobné nektarové medy (Řepka, Slunečnice, Pampeliška) jsou pozitivně korelovány s 1. hlavní komponentou. Zajímavá je skupina květových medů (Pampeliška, Řepka, Slunečnice), ve které medy velice rychle krystalizují oproti akátu, který nekrystalizuje vůbec.

Med z akátu je zřetelně oddělen od ostatních medů (liší se největším obsahem Fruktózy, nejnižším obsahem Glukózy a nejnižším poměrem Glukóza/voda).