

# Odhalení skryté struktury a vnitřních vazeb dat metodami vícerozměrné statistické analýzy

Prof. RNDr. Milan Meloun, DrSc, *Katedra analytické chemie, Univerzita Pardubice,*  
532 10 Pardubice,  
milan.meloun@upce.cz

**Souhrn:** *Diagnostické vlastnosti vybraných metod vícerozměrné statistické analýzy, především metody hlavních komponent, faktorové analýzy, tvorby shluků a diskriminační analýzy jsou ukázány. Lze odhalit vnitřní strukturu v datech, skryté vnitřní vazby a také řadu souvislostí ukrytých mezi znaky a mezi objekty. Vedle PCA, FA a CLU umožňuje také DA hodnocení rozdílů mezi dvěma nebo více skupinami objektů charakterizovaných více znaky-diskriminátory. Dělí se na techniky, které interpretují rozdíly mezi předem stanovenými skupinami objektů, a techniky, kde je cílem klasifikace objektů do skupin. Jsou porovnávány diskriminátory každého objektu (například charakteristiky sloučenin, vlastnosti objektu, vzorku vody, atd.) se znaky ostatních objektů.*

**1. Odhalení struktury ve znacích a objektech:** Zdrojová matice dat  $\mathbf{X}$  má rozměr  $n \times m$ . Před aplikací vícerozměrné statistické analýzy je třeba provést *průzkumovou (exploratorní) analýzu dat*, která umožňuje: a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafu, b) nalézt *vybočující objekty*, resp. jejich znaky, c) ověřit, zda lze použít předpoklad *lineárních vazeb*, d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Techniky vzájemných vazeb se dělí podle toho, zda hledají *strukturu a vazby ve znacích* nebo *strukturu a vazby v objektech*. To znamená a) hledání struktury ve znacích v metrické škále - *faktorová analýza FA, analýza hlavních komponent PCA a shluková analýza*, b) hledání struktury v objektech v metrické škále - *shluková analýza*, c) hledání struktury v objektech v metrické i v nemetrické škále - *vícerozměrné škálování*. d) hledání struktury v objektech v nemetrické škále - *korespondenční analýza*.

**2. Korelační míry:** Mírou podobnosti dvou objektů nebo znaků v kardinální škále je *Pearsonův párový korelační koeficient  $r$* . Objekty jsou si tím podobnější, čím je jejich párový korelační koeficient větší a bližší jedné. U ordinální škály je ale mírou *Spearmanův korelační koeficient*. Vychází se z transponované matice dat  $\mathbf{X}^T$ , kdy sloupce představují objekty a řádky pak znaky. Korelační koeficienty mezi dvěma sloupci matice  $\mathbf{X}^T$  představují korelaci mezi dvojicí objektů. Vysoká korelace prozrazuje vysokou „podobnost“ a nízká korelace pak „nepodobnost“ profilů v profilovém diagramu. Korelační matice se zobrazuje v rozptylovém diagramu dvojic znaků. Vysoké hodnoty korelačního koeficientu vedou ke zřetelné lineární závislosti (přímka), zatímco nízké ukazují, že se nacházejí v chaotickém mraku bodů. Tušíme proto, že by bylo nejlepší proložit objekty novou souřadnou osu tak, která by byla totožná s nejtěsněji prokládanou přímkou. Nová souřadnicová osa, zvaná *PCX*, leží ve směru *největšího rozptylu objektů*. Osa není přitom paralelní s žádnou z původních os znaků  $x_1$ ,  $x_2$  a  $x_3$  a nazveme ji *první hlavní komponentou*. Bude ležet ve směru *maximálního rozptylu zdrojové matice dat  $\mathbf{X}$* . Metoda hlavních komponent PCA nám poskytne nejenom tuto první komponentu, ale také další hlavní komponenty. Obvykle jedna nebo dvě, maximálně však tři, nové souřadnicové

osy  $PC1$ ,  $PC2$  a  $PC3$  vystihnou největší podíl proměnlivosti dat. Existují dvě kritéria konstrukce první hlavní komponenty: směr největšího rozptylu a metoda nejmenších čtverců. Když první hlavní komponenta  $PC1$  nestačí dostatečně popsat rozptýlení objektů, užije se druhá hlavní komponenta  $PC2$ , která je ortogonální vůči první (je na ni kolmá) a je umístěna do směru druhého největšího rozptýlení objektů v rovině. Podobně, vytvoří-li objekty v prostoru přibližný elipsoid, budou do prostorových os elipsoidu umístěny tři osy vystihující největší rozptýlení objektů čili osy tří hlavních komponent  $PC1$ ,  $PC2$  a  $PC3$ . Nové tři osy nekorelují, jsou vzájemně ortogonální, tj. na sebe kolmé.

**3. Metoda hlavních komponent PCA:** Poprvé byla PCA zavedena Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933. Cílem PCA je zjednodušení popisu vzájemně lineárně závislých neboli korelovaných znaků, a tím rozklad zdrojové matice dat do *matice strukturní* a do *matice šumové*. Každá hlavní komponenta představuje lineární kombinaci původních znaků. Základní charakteristikou každé hlavní komponenty je její míra variability čili *rozptyl*. Jde o snížení dimenze úlohy čili redukci počtu znaků bez velké ztráty informace, a to užitím pouze prvních hlavních komponent. Nevyužité hlavní komponenty obsahují malé množství informace, protože jejich míra rozptýlení je malá. Využití první hlavní komponenty jako komplexního ukazatele je běžné v oblasti ekonomie, sociologie a medicíny. První dvě respektive tři hlavní komponenty se využívají jako techniky zobrazení vícerozměrných dat v projekci do roviny nebo do 3D-prostoru. Projekce si zachovává vzdálenosti a úhly mezi jednotlivými objekty.

Zdrojová matice dat  $X$  ( $n \times m$ ) obsahuje  $n$  objektů a  $m$  znaků. *Objekty* jsou pozorování, vzorky, experimenty, měření, pacienti, rostliny atd., zatímco *znaky* či proměnné jsou druhy signálu měření, měřená veličina, vlastnosti (sladký, kyselý, hořký, slaný, cholerický atd.), barva apod. Každý znak je znám pro všech  $n$  objektů. Cílem každé vícerozměrné analýzy je zpracovat data tak, aby se zřetelně detekoval model, a tak se odkryl skrytý jev. Myšlenka sledování rozptyluje je důležitá, protože je základním předpokladem analýzy dat, že „nalezené směry maximálního rozptylu“ jsou spjaty s těmito skrytými jevy.

#### **Grafické pomůcky PCA:**

##### **1. Cattelův indexový graf úpatí vlastních čísel (Scree Plot)**

Jde o sloupcový diagram vlastních čísel  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$  zdrojové matice dat  $\mathbf{X}$  v závislosti na indexu  $i$ , který zobrazuje relativní velikost jednotlivých vlastních čísel. Úpatí v něm představuje zlomové místo mezi kolmou stěnou a vodorovným dnem. Nevýznamné hlavní komponenty (nebo faktory) představují vodorovné dno. Významné komponenty jsou zde odděleny zřetelným zlomovým místem, úpatím, a hodnota indexu  $i$  tohoto zlomu udává počet významných komponent. Protože jsou hlavní komponenty seřazeny dle klesajícího rozptylu, můžeme vybrat prvních několik hlavních komponent  $k$  jako dominantní představitele  $m$ -tice původních znaků.

##### **2. Graf komponentních vah, zátěží (Plot Components Weights)**

Zobrazuje komponentní váhy, zátěže obvykle pro první dvě hlavní komponenty. Každý bod v grafu odpovídá jednomu znaku a v grafu se především porovnávají vzdálenosti mezi znaky. Krátká vzdálenost mezi dvěma znaky znamená jejich silnou korelaci. Graf je mostem mezi původními znaky a hlavními komponentami, který ukazuje, jakou měrou přispívají jednotlivé znaky do hlavních komponent. Někdy se podaří hlavní komponenty pojmenovat, vysvětlit a

přidělit jim fyzikální, chemický nebo biologický význam. Interpretaci tohoto grafu lze shrnout do bodů:

a) Znaky  $x_j$  s vysokou mírou proměnlivosti v objektech mají vysoké hodnoty komponentní váhy a v 2D-grafu prvních dvou hlavních komponent leží daleko od počátku. Znaky s malou důležitostí leží blízko počátku.

b) Korelace a kovariance. Znaky blízko sebe nebo znaky s malým úhlem mezi svými průvodiči znaků mají vysokou kladnou kovarianci a vysokou kladnou korelaci. Naopak, znaky daleko od sebe nebo s velikým úhlem mezi průvodiči jsou negativně korelovány.

### 3. Rozptylový diagram komponentního skóre (Scatterplot)

Zobrazuje komponentní skóre pro první dvě hlavní komponenty všech objektů. Lze nalézt shluk vzájemně podobných objektů a také objekty odlehlé a silně odlišné od ostatních objektů. Graf se užívá k identifikaci odlehlých objektů, identifikaci trendů, identifikaci tříd, shluků objektů, k objasnění podobnosti objektů atd. Objekty daleko od počátku, respektive centra jsou extrémní. Objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné. Objekty umístěné zřetelně v jednom shluku jsou si podobné a přitom nepodobné objektům v ostatních shlucích. Dobře oddělené shluky prozrazují, že lze nalézt vlastní model pro samotný shluk. Izolované objekty mohou být odlehlými objekty, které jsou silně nepodobné ostatním objektům. V ideálním případě bývají objekty rozptýlené po celé ploše diagramu. Odlehlé objekty jsou totiž schopny zbortit celý diagram, ve srovnání se silně vybočujícím objektem jsou ostatní objekty nakumulovány do jediného úzkého shluku. Výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a mezi pojmenovanými hlavními komponentami. Snadno lze ohraničit shluky podobných objektů nebo nakreslením spojky mezi objekty vystihnout jejich fyzikální či biologický vztah.

### 4. Dvojný graf (Biplot)

Kombinuje předchozí dva grafy. Úhel mezi průvodiči dvou znaků je nepřímě úměrný velikosti korelace mezi těmito dvěma znaky. Čím je menší úhel, tím je větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Délka této souřadnice je úměrná příspěvku původního znaku do hlavní komponenty, čili je úměrná komponentní váze. Kombinace obou grafů do jediného společného přináší cenné srovnání, jeden graf zde působí jako doplněk vůči druhému. Interakce znaků a objektů umožňuje také vysvětlit umístění objektů vpravo od nuly na ose pomocí pozice znaků v tomto grafu, popř. umístění nahoře od nuly.

**4. Analýza shluků (CLU):** Analýza shluků (Cluster analysis, CLU) se zabývá vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd zde nazvaných *shluky*. Hodí se tam, kde objekty projevují přirozenou tendenci se seskupovat. Navzdory starému přísloví, že opaky se přitahují, v přírodě platí spíše pravidlo, že podobné věci se sjednocují. V biologii se proto užívá shluková analýza ke klasifikování živočichů a rostlin v *numerické taxonomii*. V analýze shluků bývá neznámou příslušnost všech objektů do tříd. Dokonce i počet tříd či shluků bývá neznámý. Lze formulovat tři hlavní cíle analýzy shluků: *popis systematiky*, jenž je tradičním využitím shlukové analýzy pro průzkumové cíle a taxonomii čili empirickou klasifikaci objektů, *zjednodušení dat*, kdy analýza shluků poskytuje v taxonomii zjednodušený pohled na objekty,

*identifikaci vztahu*, kdy po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky.

*Podobnost mezi objekty* je kritériem tvorby shluků objektů. Stanovují se zde znaky určující podobnost, které se dále kombinují do podobnostních měr, a tak se porovnává sledovaný objekt s jiným objektem. Analýza shluků vytváří shluky podobných objektů. Mezi-objektová podobnost může být měřena způsoby jako jsou míry korelace, míry vzdálenosti a míry asociace.

*Míry vzdálenosti* představují nejčastěji užívané míry založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Pokud míry splňují požadavky symetrie, jde o tzv. *metriky*. Použití korelace vede na jiné shluky než použití vzdálenostních měr. Nejčastější vzdálenostní mírou je *eukleidovská vzdálenost*, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Vedle této eukleidovské vzdálenosti se užívá také *čtverec eukleidovské vzdálenosti*, který tvoří základ Wardovy metody shlukování. Existuje však ještě několik dalších vzdálenostních měr.

*Míry asociace*, resp. podobnosti (binární) se používají k porovnání objektů, pokud jsou znaky nemetrického charakteru (binární proměnné), kdy respondent odpověděl na řadu otázek odpovědí *ano* nebo *ne*. Míra asociace vyjadřuje stupeň souhlasu každého páru respondentů. Nejjednodušší mírou asociace bude procento souhlasu, kdy oba respondenti na danou otázku odpověděli *ano* nebo *ne*. Rozšíření tohoto jednoduchého „souhlasného koeficientu“ je podstatou míry asociace k vyhodnocování více kategorií nominálních nebo ordinálních znaků.

*Dendrogram shluků* (vývojový strom) se konstruuje pouze v případě, když je k dispozici matice původních znaků. *Dendrogram podobnosti objektů* je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích. *Dendrogram podobnosti znaků* odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) znaků, které jsou si velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku jsou si značně podobné a jsou vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti či znaky není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností. Při sestrojování dendrogramu řadou technik se posuzuje *míra věrohodnosti* těsností proložení. Prvním kritériem při volbě „nejlepšího dendrogramu“ je *kofenetický korelační koeficient CC*, což je Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností v dendrogramu. Čím vyšší hodnota *CC*, tím větší věrohodnost a lepší model shluků. Druhým kritériem těsnosti proložení je *kritérium delta A*, které měří stupeň přetvoření struktury dat spíše než stupeň podobnosti.

Vzorové příklady:

### PŘÍKLAD 9.13 *Klasifikace vlastností rozličných druhů kávy*

Byl získán výběr 43 vzorků kávy, pocházejících ze 30 zemí. U každého druhu kávy byly změřeny jeho chemické a fyzikální vlastnosti. Splňují data požadavky na homogenitu a je možné indikovat dvě či více rozličných kategorií? Vytvořte dendrogram klasifikovaných druhů kávy.

○ **Data:** Soubor dat *Kava* obsahuje 2 druhy kávy, Robusta a Arabica ve 43 vzorcích ze 30 zemí a popsanych 13 fyzikálně-chemickými znaky:

*i* značí index kávy,

**Objekt** značí původ kávy,

**Voda** značí obsah vody  $x_1$ ,

**Zrno** značí hmotnost zrn  $x_2$ ,

**Extrakt** značí extrakt  $x_3$ ,

**pH** značí hodnotu pH  $x_4$ ,

**Acidita** značí hodnotu volné acidity  $x_5$ ,

**Mineral** značí obsah minerálů  $x_6$ ,

**Tuky** značí obsah tuků  $x_7$ ,

**Kofein** značí obsah kofeinu  $x_8$ ,

**Trinonelin** značí obsah trinonelinu  $x_9$ ,

**Kchlorogen** značí obsah kyseliny chlorogenikové  $x_{10}$ ,

**Kneochlor** značí obsah kyseliny neochlorogenikové  $x_{11}$ ,

**Kisochlor** značí obsah kyseliny isochlorogenikové  $x_{12}$ ,

**Sumakys** značí sumu kyselin chlorogenikových  $x_{13}$ .

<i>i</i>	Objekt	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$	$x_{11}$	$x_{12}$	$x_{13}$
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1.0	5.4	0.4	0.8	6.6
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
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

○ **Řešení:** Graf komponentních vah znaků odhaluje především korelaci znaků. Je-li úhel mezi průvodiči dvou znaků malý, jsou dva znaky v silné korelaci.

**První shluk** obsahuje znaky *Voda*, *pH*, *Kchlorogen*, *Sumakysel*, *Mineral*, *Kofein*, *Trinonelin*, *Kneochlor*, *Kisochlor*, a *Tuky*.

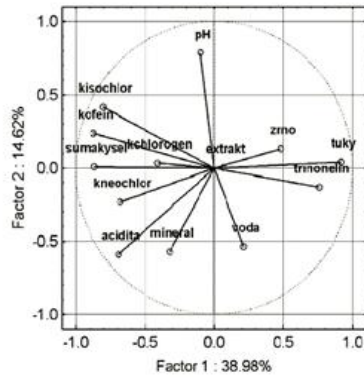
**Druhý shluk** obsahuje dva znaky, *Extrakt* a *Acidita*.

**Vznik shluků druhů kávy** lze sledovat na grafu komponentního skóre objektů a na dendrogramu objektů.

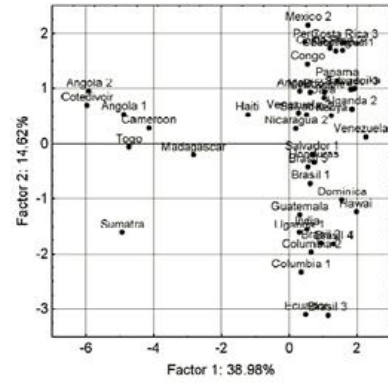
**Graf komponentního skóre** ukazuje, že 43 objektů čili druhů kávy v datovém souboru *Kava* nejsou dostatečně homogenní.

**Objekty** zde lze rozdělit do dvou shluků, v prvním vlevo je 7 objektů a ve druhém svislém shluku vpravo je zbývajících 36 objektů. Klasifikace do těchto shluků je především vlivem znaků *Tuky*, *Kofein*, *Trinonelin* a *Sumakys*.

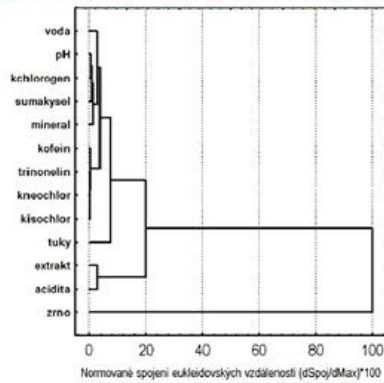
Na **dendrogramu objektů** při postupu zprava doleva je zřejmé, že druhy kávy lze rozdělit do dvou velkých shluků. Větší shluk nazvaný Arabica lze dále rozdělit na dva menší shluky Arabica A a Arabica B a jeden odlehlý objekt. Ve spodní části obrázku zůstává jeden větší shluk 13 druhů kávy, patřících zřejmě do druhu Robusta.



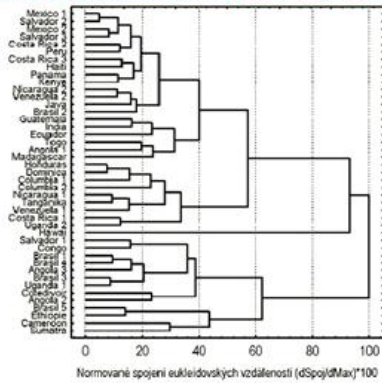
Graf komponentních vah znaku matice dat Káva, (STATISTICA).



Graf komponentního skóre objektů matice dat Káva

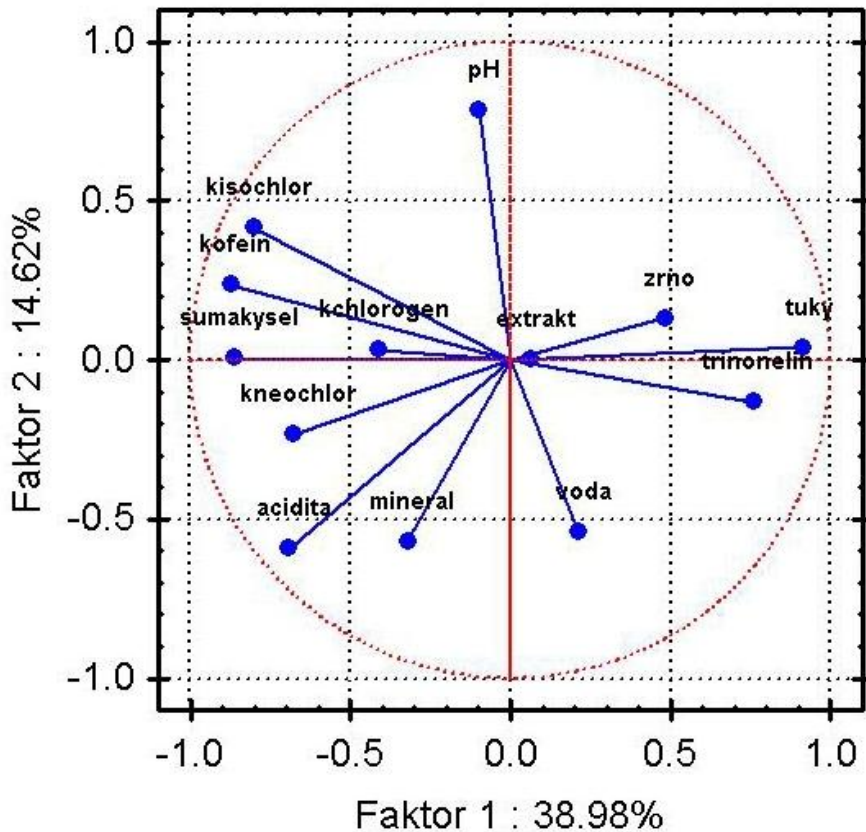


Dendrogram znaku matice dat Káva (STATISTICA).



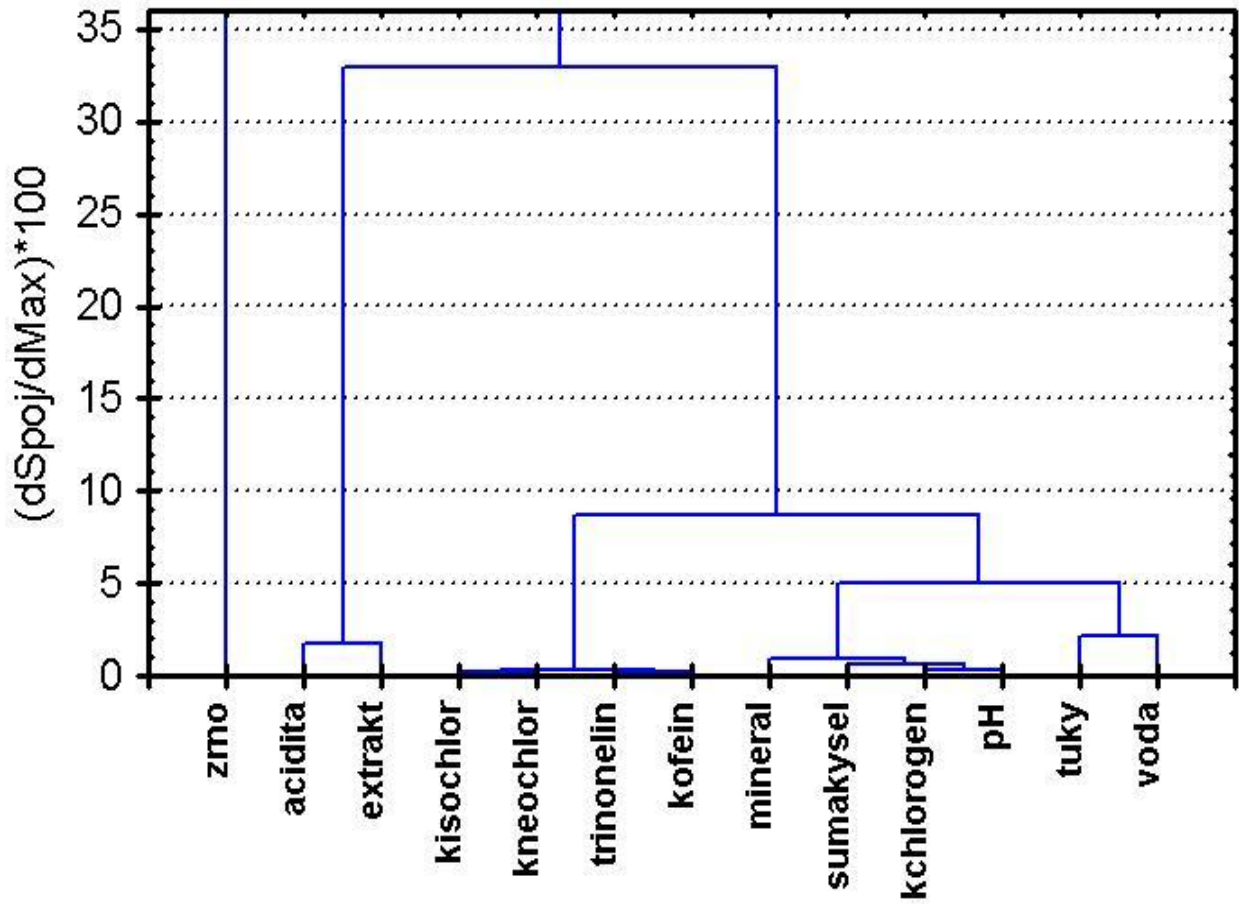
Dendrogram objektů matice dat Káva, (STATISTICA).

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

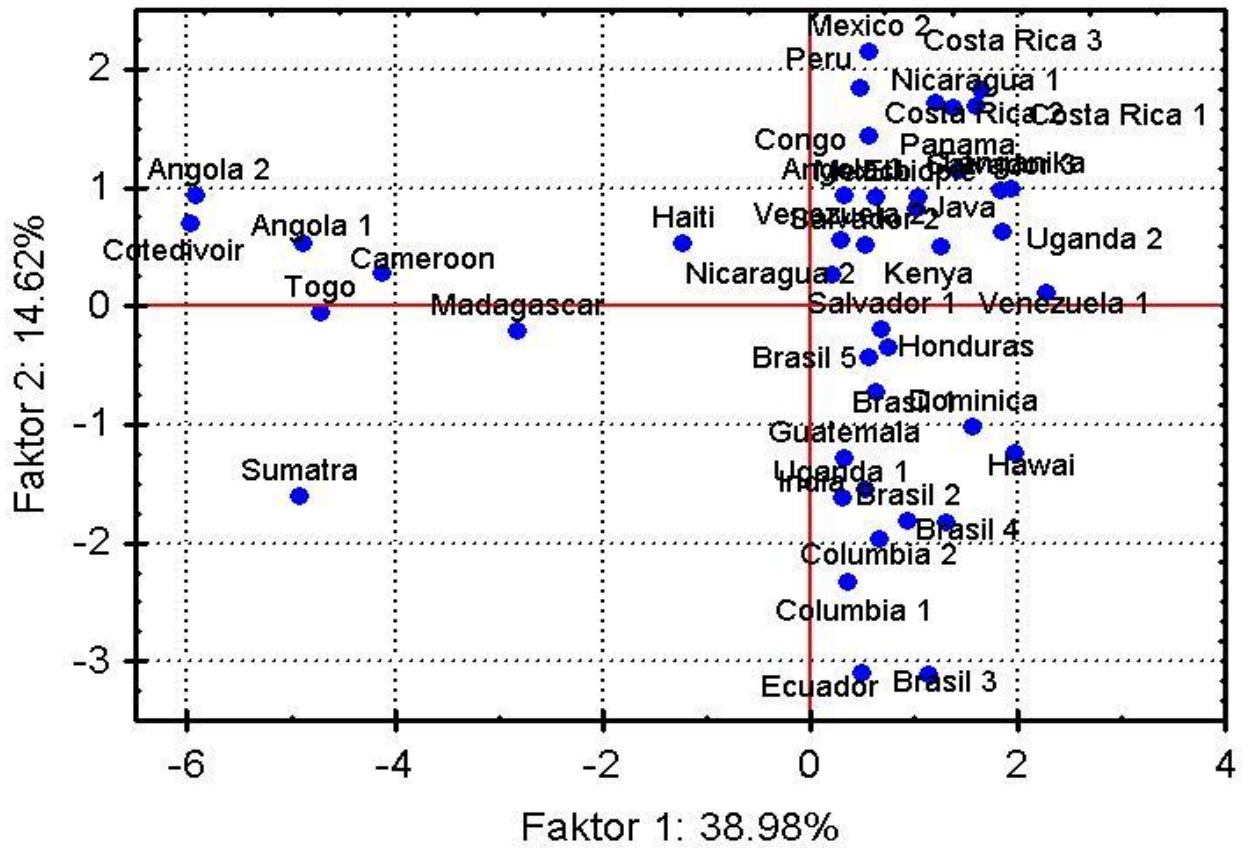


Graf komponentních vah matice dat Káva(Statistica)

### Wardova metoda

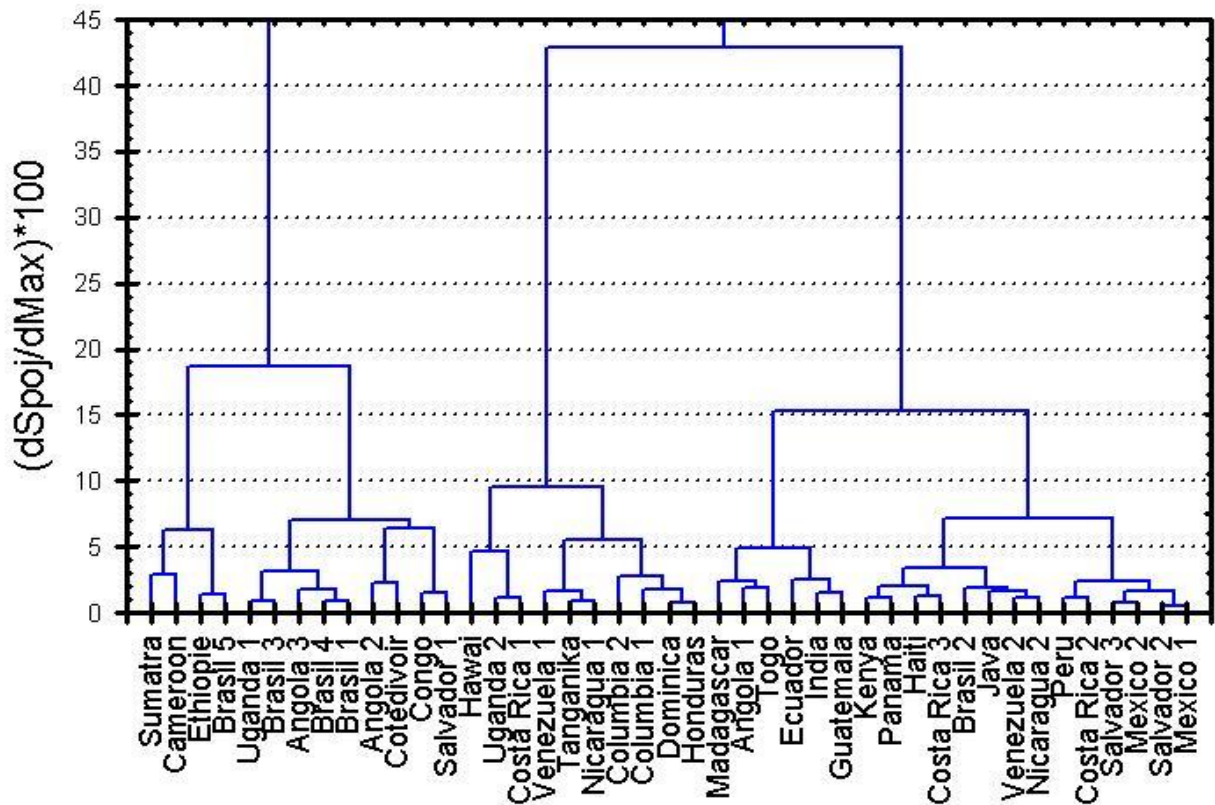


Dendrogram znaků matice dat Káva (Statistica)



Rozptylový diagram komponentního skóre matice dat Káva (Statistica)

### Wardova metoda



Dendrogram objektů matice dat Káva (Statistica)



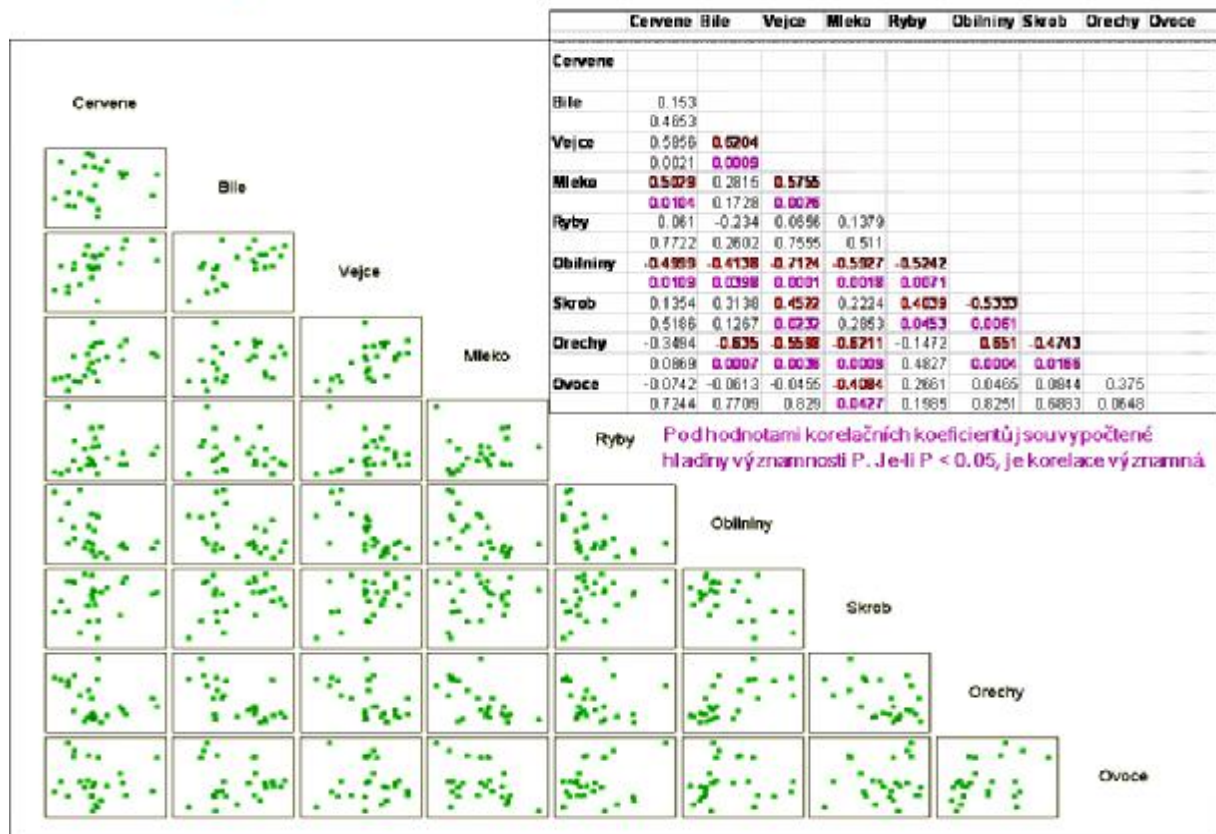
### Úloha 3. Sledování spotřeby proteinů v Evropě (kompendium B418)

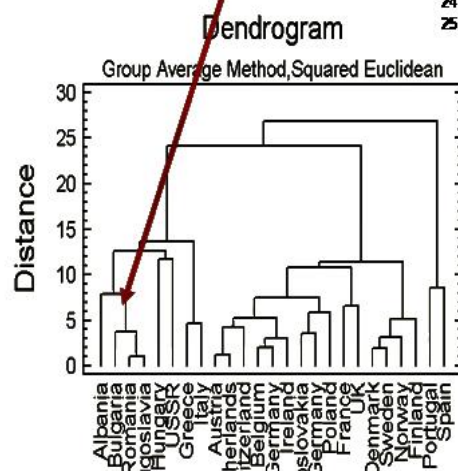
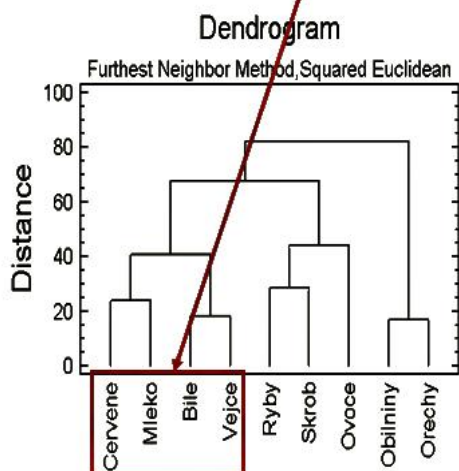
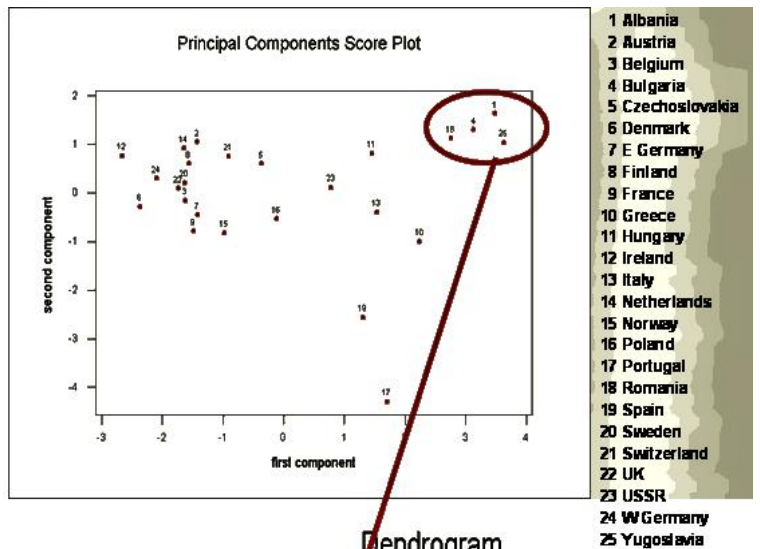
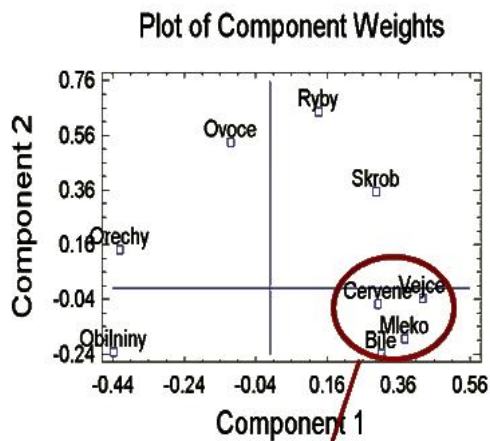
Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 8 druhů potravin je předmětem vyšetření.

**Data:** / značí Index; **C** červená, **B** bílé, **V** vejce, **M** mléko, **R** ryby, **O** obilniny, **S** škrob, **Or** ořechy, **Ov** ovoce a zelenina

i	Objekty Stát	Průměrná Cervena	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Ořechy	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	1.4	4.3	19.9	2.1	28	3.0	1.3	4.3
3	Belgium	13.5	9.3	4.1	17.5	4.5	26.0	5.7	2.1	4
4	Bulgaria	7.8	6	1.0	8.3	1.2	56.7	1.1	3.7	4.2
5	Czechoslov.	9.7	11.1	2.6	12.5	2	34.3	5	1.1	4
6	Denmark	10.8	10.6	3.7	25	8.8	21.9	4.6	0.7	2.4
7	Germany	8.1	11.8	3.7	11.1	5.1	24.6	8.5	0.8	3.8
8	Ireland	9.5	4.9	2.7	33.7	5.8	28.3	5.1	1	1.4
9	France	10	8.9	3.3	19.5	5.7	28.1	4.6	2.4	6.5
10	Greece	10.2	3	2.6	17.6	5.8	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	49.0	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	25.0	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	10.6	3	43.0	6.4	3.4	2.9
24	W Germany	11.1	12.5	4.1	18.8	3.1	18.6	5.2	1.5	3.6
25	Yugoslavia	4.1	5	1.2	9.5	0.8	55.9	3	5.7	3.2

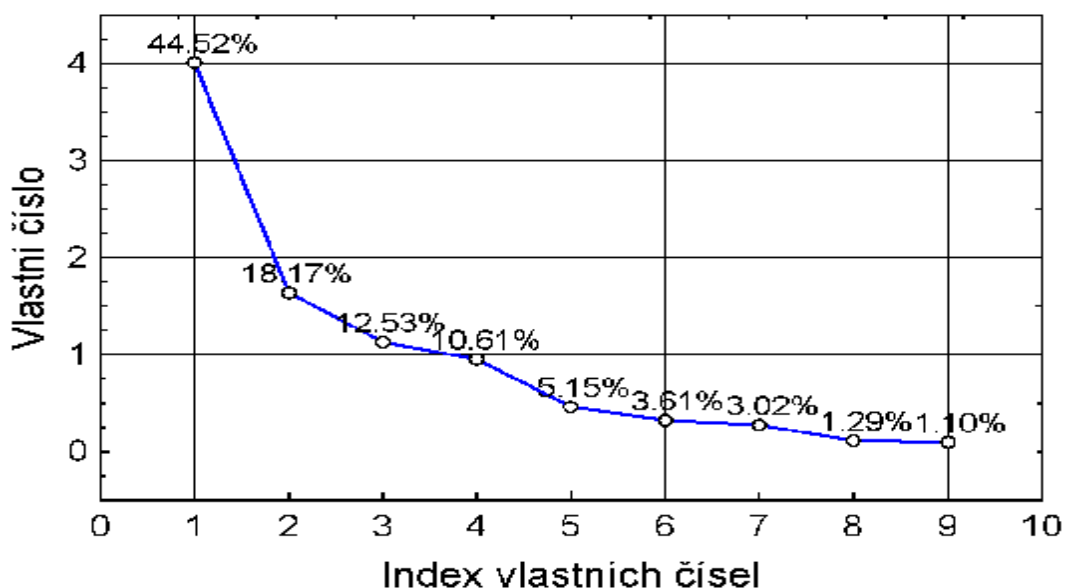
# Test významnosti korelace v korelační matici



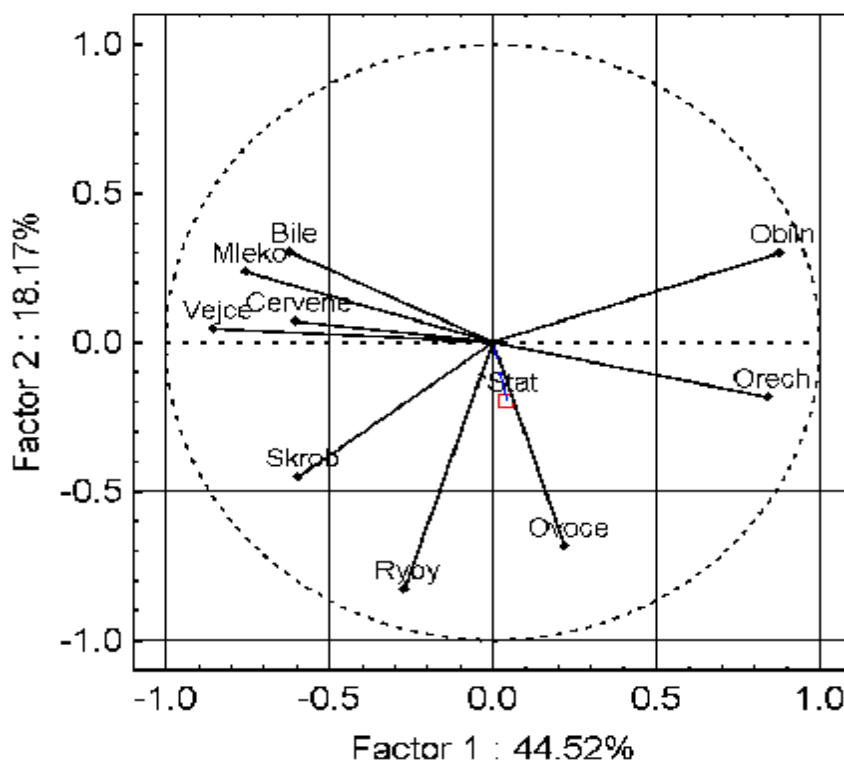


Z **Cattelova grafu** vyplynulo, že první hlavní komponenta (44.52 % celkové proměnlivosti) a druhá hlavní komponenta (18.17 % celkové proměnlivosti) dohromady dostatečně popíší proměnlivost v datech. V krajním případě by bylo možné dle Kaiserova kritéria uvažovat ještě i třetí hlavní komponentu, která zahrnuje 12.53 % celkové proměnlivosti.

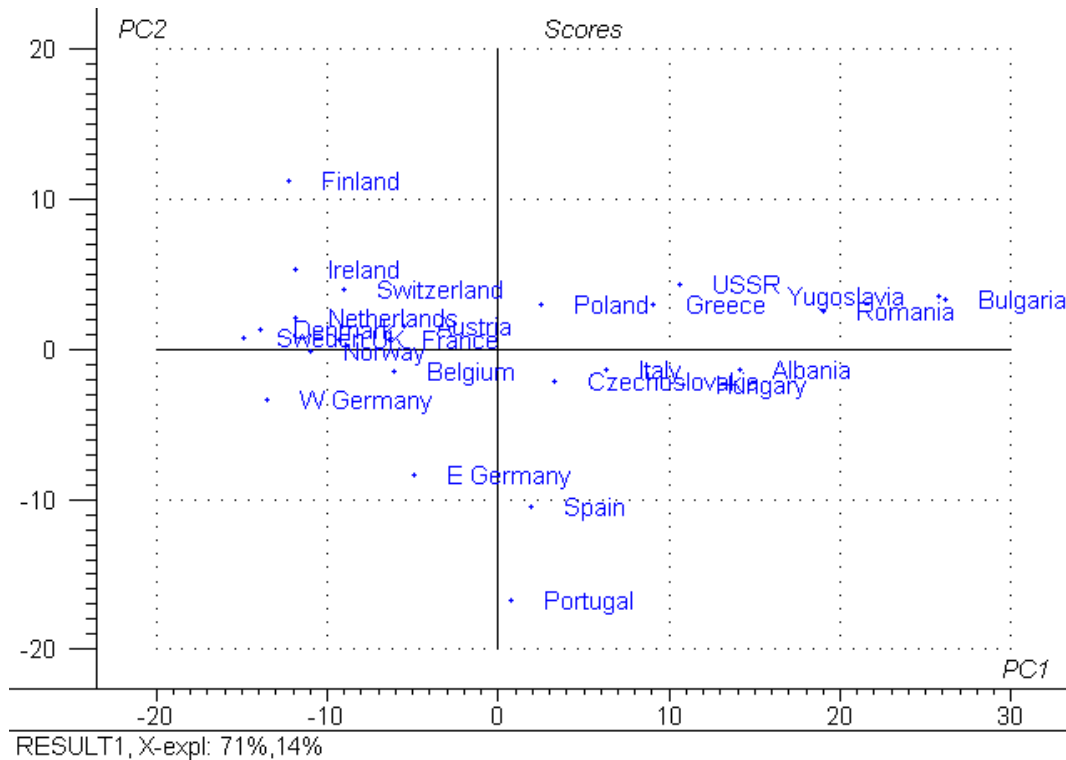
Cattelův indexový graf úpatí vlastních čísel



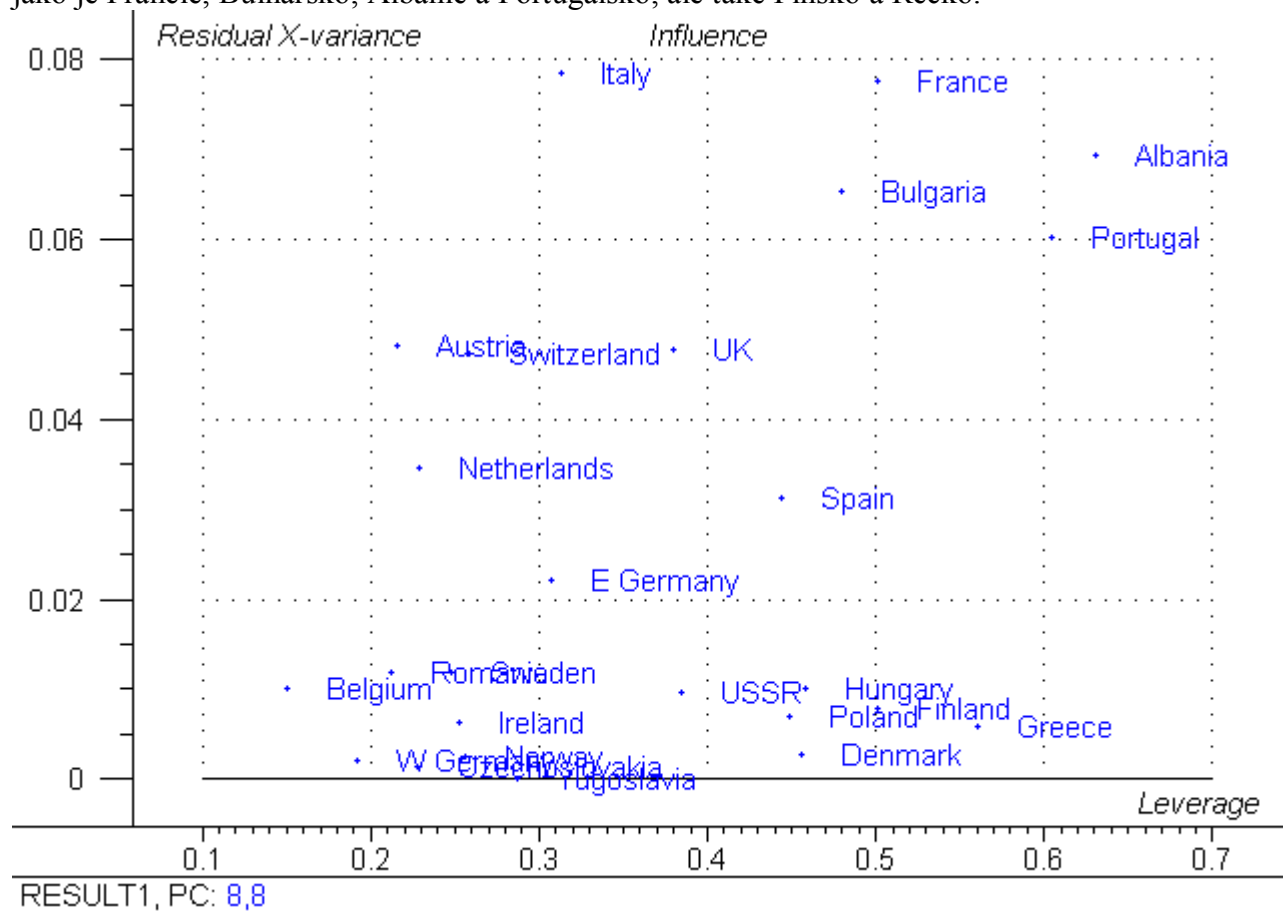
Z grafu komponentních vah PC1 a PC2 plyne, že například *Mléko* a *Obilniny* spolu vzhledem obsahu proteinů nekorelují. Výjimečně si stojí i znak *Ryby*. Okolo počátku je shluk znaků, které jsou spolu v silné korelaci, jsou to *Červené maso*, *Bílé maso*, *Vejce*, *Škrob*, *Ořechy* a *Ovoce* a zelenina.



**Rozptylový diagram komponentního skóre** roztřídil státy dle spotřeby proteinů na základě podobnosti konzumované potravy do několika shluků. Vyčlenil se shluk balkánských zemí (Bulharsko, Rumunsko, Albánie, Jugoslávie). Další shluk s pozitivní korelací tvoří země Polsko, Řecko, SSSR, Československo, Itálie a Maďarsko. Španělsko koreluje s Portugalskem a NDR. Velký shluk obsahuje státy západní Evropy, ze kterých však vybočuje Finsko a částečně i NSR.



**Graf vlivných bodů** ukazuje postavení vybočujících či odlehlých objektů země umístěné v horní části grafu (Itálie, Francie, Bulharsko, Albánie a Portugalsko). Za extrémě ale ne odlehlé lze z regresního hlediska v tomto grafu označit země při pravém okraji grafu, jako je Francie, Bulharsko, Albánie a Portugalsko, ale také Finsko a Řecko.



**Poděkování:** Autoři vyslovují svůj dík za finanční podporu vědeckého záměru č. MSM0021627502.

### **Doporučená literatura**

- [1] ČHMÚ – databáze jakosti vody (<http://hydro.chmi.cz/ojv2/>)
- [2] STATISTICA 7.0 (<http://www.statsoft.cz>)
- [3] Statistická ročenka životního prostředí České republiky 2006. MŽP ČR, Praha 2006.
- [4] NCSS 2000 (<http://www.ncss.com/>)
- [5] Meloun M., Militký J., Hill M.: Počítačová analýza vícerozměrných dat v příkladech. Academia, Praha 2005.
- [6] Pytela O.: Chemometrie pro organické chemiky. Univerzita Pardubice, skripta, Pardubice 2003.
- [7] Meloun M., Militký J.: Kompendium statistického zpracování experimentálních dat. Academia, Praha 2002 (1. vydání), 2006 (2. rozšířené vydání).
- [8] Meloun M., Militký J.: Statistická analýza experimentálních dat. Academia, Praha 2004.