Výhody a přednosti vícerozměrné analýzy dat

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice

Souhrn: *Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných,* $y = w_1 x_1 + \ldots + w_m x_m$. *Zdrojová matice dat obsahuje proměnné v m sloupcích a objekty v n řádcích. Data jsou před zpracováním škálována. Cílem je nalézt shluk jako množinu podobných objektů s podobnými proměnnými. Podobnost objektů posuzujeme na základě vzdálenosti (míry) objektů v m-rozměrném prostoru: čím je vzdálenost shluků či objektů větší, tím menší je jejich podobnost. K rychlému posouzení podobnosti slouží grafy exploratorní analýzy vícerozměrných dat: profily, polygony, sluníčka a hvězdičky. Strukturu a vazby mezi proměnnými vystihují metody snížení dimensionality, metoda hlavních komponent (PCA). Důležitou pomůckou je rozptylový diagram, který zobrazuje objekty, rozptýlené v rovině prvních dvou hlavních komponent. Graf komponentních vah porovnává vzdálenosti mezi proměnnými* x_i *a* x_j, *kde krátká vzdálenost značí silnou korelaci. Dvojný graf pak kombinuje oba předchozí grafy. Objekty lze seskupovat do shluků* ***hierarchicky*** *dle předem zvoleného způsobu metriky (průměrově, centroidně, nejbližším sousedem, nejvzdálenějším sousedem, medianově, mezi těžišti a průměrnou vazbou) a* ***nehierarchicky*** *dle uživatelem vybraných objektů-představitelů. Výsledkem je dendrogram. Metoda hlavních komponent a tvorba shluků je demonstrována na dvou vzorových úlohách.*

Vícerozměrná statistická analýza vychází z koncepce latentních proměnných (faktorů, kanonických proměnných) y, které jsou lineární kombinací původních proměnných x s vhodně volenými vazbami. Latentní proměnná y je kombinací m-tice sledovaných (měřených resp. jinak získaných) proměnných $x_1, x_2, \ldots, x_m$ ve tvaru $y = w_1 x_1 + w_2 x_2 + \ldots + w_m x_m$. Jednotlivé vícerozměrné metody využívají různých způsobů stanovení vah $w_1, w_2, \ldots, w_m$.

Zdrojová matice má rozměr $n \times m$. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje

(a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafů,
(b) nalézt *vybočující objekty*, resp. jejich proměnné,
(c) stanovit, zda lze použít předpoklad *lineárních vazeb,*
(d) ověřit *předpoklady o datech* (normalita, nekorelovanost, homogenita).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda se hledají

(a) *struktura a vazby v proměnných* nebo
(b) *struktura a vazby v objektech*:

(1) Hledání struktury *v proměnných* v metrické škále: *faktorová analýza FACT* a *analýza hlavních komponent PCA*.
(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*.
(5) Většina metod vícerozměrné statistické analýzy umožňuje *zpracování lineárních vícerozměrných modelů*, kde závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných resp. vazby mezi proměnnými jsou lineární. V řadě případů se také uvažuje normalita metrických proměnných.

Určením struktury a vzájemných vazeb mezi proměnnými ale i mezi objekty se zabývají techniky redukce proměnných na latentní proměnné, metoda *analýzy hlavních komponent (PCA)* a *metoda faktorové analýzy (FA)*. Důležitou metodou určení vzájemných vazeb mezi proměnnými je i *kanonická korelační analýza CA*, která se používá ke zkoumání závislosti mezi dvěma skupinami proměnných, přičemž jedna ze skupin se považuje za proměnné nezávislé a druhá za skupinu proměnných závislých.

Analýza hlavních komponent (PCA)

Cílem metody je transformace dat z původních proměnných x_j, $j=1, ..., m$, do menšího počtu latentních proměnných y_j. Tyto latentní proměnné mají vhodnější vlastnosti, je jich výrazně méně, vystihují téměř celou *proměnlivost* původních proměnných a jsou vzájemně nekorelované (korelační koeficient mezi latentními proměnnými $y_1, ..., y_m$ je 0). Latentní proměnné jsou u této metody nazvány *hlavními komponentami* a jsou to lineární kombinace původních proměnných: *první hlavní komponenta* y_1 popisuje největší část proměnlivosti čili rozptylu původních dat, *druhá hlavní komponenta* y_2 zase největší část rozptylu neobsaženého v y_1, atd. Matematicky řečeno, *první hlavní komponenta* je takovou lineární kombinací vstupních proměnných, která obsahuje největší proměnlivost mezi všemi lineárními kombinacemi. Má tvar

$$y_1 = \sum_{j=1}^{m} v_{1j}\, x_j = \boldsymbol{v}_1^T\, \boldsymbol{x},$$

kde objekt $\boldsymbol{x}$ obsahuje proměnné $x_1, ..., x_m$. Pro vektor koeficientů $\boldsymbol{v}_1^T = (v_{11}, ..., v_{1m})^T$ platí, že proměnlivost vyjádřená rozptylem $D(y_1) = \boldsymbol{v}_1^T\, \boldsymbol{S}\, \boldsymbol{v}_1$ je maximální $D(y_1), D(y_2), ..., D(y_m)$, přičemž $\boldsymbol{S}$ značí kovarianční matici původních dat. Zcela analogicky jsou konstruovány další hlavní komponenty, jejichž celkový počet je roven menšímu ze dvou čísel, a to n (počet objektů) nebo m (počet proměnných). Protože platí, že součet rozptylů všech hlavních komponent je roven součtu rozptylů vstupujících původních proměnných, můžeme z podílu rozptylů jednotlivých hlavních komponent usuzovat na část proměnlivosti vysvětlenou dotyčnou hlavní komponentou. Jestliže součet prvních (nejvyšších) A podílů proměnlivosti je dostatečně blízký jedné čili 100% (obvykle však stačí 0.9 - 0.95 čili 90% - 95%), postačí brát v úvahu právě těchto prvních A hlavních komponent pro

"dostatečné" vysvětlení variability původních proměnných. Rozdíl mezi souřadnicemi objektů v původních proměnných a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá špatnou těsností proložení modelu PCA nebo také *chybou modelu PCA*. I při velkém počtu původních proměnných (m) může být A velmi malé, často 2 až 5. Volba počtu užitých komponent A představuje vlastní *model hlavních komponent PCA*. Vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních proměnných x_j, $j = 1, ..., m$, k hlavním komponentám y_k, $k = 1, ..., A$, tvoří dominantní součásti zvoleného modelu hlavních komponent PCA.

Vlastní *matematický postup PCA* je následující: maximalizací při zavedení normalizační podmínky $v_1^T v = 1$ vyjde, že

$$(S - \lambda_1 I)\, v_1 = \mathbf{0} ,$$

kde $\mathbf{0}$ označuje nulový vektor, λ_1 je největší *vlastní číslo* a v_1 je odpovídající *vlastní vektor* kovarianční matice S a I je jednotková matice. Po dosazení vyjde

$$D(y_1) = v_1^T S\, v_1\, \lambda_1 .$$

Analogicky lze odvodit, že vektor koeficientů v_2 ve vztahu $y_2 = \sum_{j=1}^{m} v_{2j}\, x_j$, maximalizující $D(y_2)$ za podmínky, že $cov(y_1, y_2) = 0$, odpovídá vlastnímu vektoru, příslušejícímu druhému největšímu vlastnímu číslu λ_2.

Provedeme-li rozklad kovarianční matice S na vlastní čísla $\lambda_1 \geq \lambda_2 ... \geq \lambda_m$, jsou odpovídající vlastní vektory $v_1, v_2, ..., v_m$ přímo koeficienty hlavních komponent $y_1, ..., y_m$. Hlavní komponenty mají řadu zajímavých vlastností. Lze je interpretovat jako hlavní osy m-rozměrného elipsoidu $x^T S^{-1} x$ = konst .

K odstranění závislosti na jednotkách původních proměnných se lépe užívá standardizovaných proměnných x^* s prvky $x_j^* = (x_j - \bar{x}_j)/\sigma_j$. Pro j-tou hlavní komponentu pak platí

$$y_j^* = \sum_{k=1}^{m} v_{jk}^*\, x_k^* ,$$

kde v_j^* je vlastní vektor *korelační* matice R odpovídající j-tému největšímu vlastnímu číslu λ_j^*. Hlavní komponenty y_j^*, určené z korelační matice R, jsou však hůře interpretovatelné. Platí, že $v_j^{*T} v_j^* = \lambda_j$, nikoliv rovno jedné. Pro účely zobrazení vícerozměrných dat různého měřítka jsou však vhodnější standardizované y_j^* než původní y_j.

Graficky lze výsledek analýzy hlavních komponent zobrazit v několika grafech hlavních komponent následujícím způsobem:

(a) **Indexový graf úpatí vlastních čísel** (Scree Plot) je vlastně sloupcový diagram vlastních čísel nebo reziduálního rozptylu proti stoupající hodnotě indexu, pořadového čísla A (obr. 9). Zobrazuje relativní velikost jednotlivých vlastních čísel. Řada autorů ho s oblibou využívá k určení počtu A "užitečných" hlavních komponent. Často je slovo scree vysvětlováno jako zlomové místo mezi kolmou stěnou a vodorovným dnem. Vybrané "užitečné" hlavní komponenty (nebo také faktory) pak tvoří kolmou stěnu a "neužitečné" hlavní komponenty (nebo faktory) představují vodorovné dno. Užitečné komponenty jsou tak odděleny zřetelným zlomovým místem, a x-ová souřadnice tohoto zlomu je hledaná hodnota indexu.

(b) **Graf komponentních vah** (Plot Components Weights) zobrazí komponentní

váhy pro první dvě hlavní komponenty, (obr. 10). V tomto grafu se porovnávají vzdálenosti mezi proměnnými. Krátká vzdálenost mezi dvěma proměnnými znamená silnou korelaci. Lze nalézt i shluk podobných proměnných, jež spolu korelují. Tento graf lze považovat za most mezi původními proměnnými a hlavními komponentami, protože ukazuje, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent. Někdy se podaří hlavní komponenty y_1, y_2, ... pojmenovat, vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam. Pak lze názorně vysvětlit jak jednotlivé původní proměnné x_j, $j = 1,..., m$, přispívají do první hlavní komponenty y_1 nebo do druhé hlavní komponenty y_2. Některé původní proměnné x_j přispívají kladnou vahou, některé zápornou. Bývá zajímavé sledovat kovarianci původních proměnných x_j v prostorovém 3D grafu komponentních vah y_1, y_2 a y_3. Jsou-li původní proměnné x_j, $j = 1,..., m$, blízko sebe v prostorovém shluku, jde o silnou pozitivní kovarianci. Kovariance však nemusí ještě nutně znamenat korelaci. Výklad grafu komponentních vah lze obecně shrnout do následujících bodů:

1. *Důležitost původních proměnných x_j, $j = 1,..., m$,:* proměnné x_j s vysokou mírou proměnlivosti v datech objektů mají vysoké hodnoty komponentní váhy. Ve 2D-diagramu prvních dvou hlavních komponent pak leží hodně daleko od počátku. Proměnné s malou důležitostí leží blízko počátku. Když určíme *důležitost proměnných*, určíme tím také jejich proměnlivost: jestliže, například y_1 objasnuje 70% proměnlivosti a y_2 jenom 5% (přečteno z indexového grafu úpatí vlastních čísel), jsou původní proměnné x_j, $j = 1,..., m$, s vysokou vahou v y_1 tím pádem mnohem důležitější než proměnné x_j s vysokou vahou v y_2.
2. *Korelace a kovariance:* původní proměnné x_j, $j = 1,..., m$, blízko sebe a nebo proměnné x_j s malým úhlem mezi svými průvodiči proměnných a na stejné straně vůči počátku mají vysokou kladnou kovarianci a vysokou kladnou korelaci. Naopak, původní proměnné x_j daleko od sebe anebo s velikým úhlem mezi průvodiči proměnných jsou negativně korelovány.
3. *Spektroskopická data:* ve spektroskopických datech je 1-rozměrný graf komponentních vah často nejvhodnější. I zde platí pravidlo, že vysoké komponentní váhy představují vysokou důležitost proměnných x_j (vlnových délek).

(c) **Rozptylový diagram komponentního skóre** (Scatterplot) zobrazuje *komponentní skóre* čili hodnoty obyčejně prvních dvou hlavních komponent u všech objektů, (obr. 11). Dokonalé rozptýlení objektů v rovině obou hlavních komponent vede k rozlišení objektů při jejich popisu pomoci y_1 a y_2. Snadno lze v rovině nalézt shluk vzájemně podobných objektů a dále objekty odlehlé a silně odlišné od ostatních. Diagram komponentního skóre však může být i ve 3 či více hlavních komponentách a v rovinném grafu se sleduje pak pouze jeho průmět do roviny. Tento diagram se užívá k identifikaci odlehlých objektů, identifikaci trendů, tříd, shluků objektů, k objasnění podobnosti objektů, atd. Je nemožné analyzovat všechny diagramy, protože jich je velmi mnoho: uvažujme například $m < n$ a pro $m = 10$ proměnných existuje $m(m-1)/2 = 45$ diagramů, pro $m = 11$ pak 55 diagramů, pro $m = 12$ pak 66 diagramů, atd. Obvykle vybíráme diagramy y_1 vs. y_2, y_1 vs. y_3, y_1 vs. y_4, atd. Držíme se první hlavní komponenty y_1, protože v ní bývá

největší míra proměnlivosti v datech. Interpretace rozptylového diagramu komponentního skore lze shrnout do těchto bodů:

1. *Umístění objektů*: objekty daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou typičtější.
2. *Podobnost objektů*: objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné.
3. *Objekty v shluku*: 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. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.
4. *Osamělé objekty:* izolované objekty mohou být odlehlé objekty, které jsou silně nepodobné ostatním objektům.
5. *Odlehlé objekty:* v ideálním případě bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je něco špatného v modelu, obyčejně je přítomen silně odlehlý objekt. Odlehlé objekty jsou totiž schopny zbortit celý diagram, ve srovnání se silně vybočujícím objektem jsou ostatní objekty nakumulovány do jediného úzkého shluku. Po odstranění vybočujícího objektu se ostatní objekty roztřídí po celé ploše diagramu a teprve vypovídají o existujících shlucích.
6. *Pojmenování objektů:* výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a mezi pojmenovanými hlavními komponentami. Snadno obkroužíme shluky podobných objektů nebo nakreslením spojky mezi objekty 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 a pomocí původních proměnných pak i vysvětleno.

(d) **Dvojný graf** (Biplot) kombinuje předchozí dva grafy, (obr. 12). Úhel mezi průvodiči dvou proměnných x_j a x_k je nepřímo úměrný velikosti korelace mezi těmito dvěma proměnnými. Čím menší úhel, tím větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Délka této souřadnice je úměrná příspěvku původní proměnné x_j do hlavní komponety čili je úměrná komponentní váze. Kombinace obou předešlých grafů v jediném přináší cenné srovnání, jeden graf působí zde doplnkově vůči druhému. Když se ve dvojném grafu nachází objekt v blízkosti určité proměnné x_j znamená to, že tento objekt "obsahuje" hodně právě této proměnné a je s ní v interakci. Interakce proměnných a objektů umožňuje také vysvětlit umístění objektů vpravo od nuly na ose y_1 (či vlevo od nuly) pomocí pozice proměnných v tomto grafu, resp. umístění nahoře od nuly (či dole od nuly) na ose y_2.

Diagnostikování častých problémů v PCA: v analýze hlavních komponent PCA se často setkáváme s následujícími problémy:

1. *Data neobsahují předpokládanou informaci:* vysvětlení grafů a diagramů metody PCA nemá smysl, protože bylo provedeno několik závažných chyb.
2. *Užito příliš málo hlavních komponent:* v modelu PCA bylo použito příliš málo latentních proměnných. Nedostatečné vysvětlení dat vede ke ztrátě informace. Problém se může vyřešit opětovným rozborem grafu úpatí vlastních čísel.

3. *Užito příliš mnoho hlavních komponent:* v modelu PCA bylo zahrnuto příliš mnoho latentních proměnných, což může vyvolat vážnou chybu, protože šum je zahrnut do modelu.
4. *Neodstranění odlehlých objektů:* odlehlé objekty mohou být důvodem hrubých chyb v datech. Do modelu jsou vtahovány spíše hrubé chyby než zajímavé proměnlivosti v datech objektů.
5. *Odstraněné odlehlé objekty obsahovaly důležitou informaci:* ztrátou určitých objektů se vytratila důležitá informace z dat a nalezený model je proto zkreslený.
6. *Komponentní skóre je nedostatečně analyzováno:* nedostatečným rozborem důležitého rozptylového diagramu byly zanedbány důležité rysy v datech.
7. *Vysvětlení komponentních vah se špatným počtem hlavních komponent:* může vést k vážnému zkreslení výkladu. Může totiž dojít k vyjmutí důležitých proměnných, protože se zdají býti odlehlými. Tento graf je mostem mezi prostorem původních proměnných a prostorem hlavních komponent PCA. Když zvolíme špatný prostor PCA, tento most už nám mnoho nepomůže.
8. *Přecenění standardních diagnostik v software:* je třeba hodně rozvažovat a přemýšlet o úloze samé a specifickém problému řešeném před pohodlným přebíráním počítačových výsledků.
9. *Užití špatné předzpracování dat:* chybná předúprava dat (ve škálování užité centrování nebo standardizace, transformace logaritmická, mocninná, Box-Coxova, atd.) může vést ke zkresleným závěrům a neporozumění úloze. Způsob předúpravy dat je obecně dán typem úlohy a druhem instrumentálních dat a může vést také ke ztrátě informace.

Postup metody hlavních komponent PCA

Problém musí být správně a přesně definován. Je odpovědností řešitele, aby data obsahovala dostatek relevantní informace k řešení problému. Ani nejlepší počítačová metoda nemůže kompenzovat nedostatek informace v datech. *Maticový graf korelace proměnných* slouží k získání počáteční informace o celém datovém souboru. Odhalí, zda data potřebují škálování. Při prvním seznámení s daty se se v rámci exploratorní analýzy, kam také patří metoda hlavních komponent PCA, aplikuje první výpočet touto metodou. Data je obvykle potřeba škálovat, alespoň centrovat. Lze vyzkoušet i ostatní formy předúpravy dat. V tomto stádiu se vždy vyčíslují všechny hlavní komponenty. První diagramy komponentního skóre slouží k odhalení odlehlých hodnot, tříd, shluků a trendů. Jsou-li objekty roztříděny do dobře oddělených shluků, je třeba určit způsob jak je z dat oddělit a shluky pak analyzovat odděleně. Nikdy se nepokoušíme odhalovat a odstraňovat odlehlé proměnné, mohlo by pak dojít k odstranění cenné informace. Po redukci datového souboru na několik podsouborů, kdy shluky jsou modelovány odděleně se znovu aplikuje metoda hlavních komponent PCA na jednotlivé podsoubory.

1. *Vyšetření indexového grafu úpatí vlastních čísel:* z hrany v tomto diagramu se určí nejlepší počet hlavních komponent.

2. *Výpočet vlastních vektorů pro hlavní komponenty:* vedle číselných hodnot se užívá i názorný čarový diagram hodnot vlastních čísel vektorů, který přehledně informuje o zastoupení původních proměnných x_j, $j = 1, ..., m$, v hlavních komponentách.

3. *Výpočet komponentních vah:* matice párových korelačních koeficientů ukazuje na korelaci původních proměnných s hlavními komponentami. Čarový diagram názorně vysvětluje korelační strukturu mezi oběma druhy proměnných. Uživatel nyní vybere pouze prvních A hlavních komponent a vytvoří tak *PCA model.*

4. *Vyšetření grafu komponentních vah.*

5. *Vyšetření rozptylového diagramu komponentního skóre.*

6. *Vyšetření dvojného grafu.*

7. *Vyšetření grafu úpatí reziduí objektů:* rezidua objektů a rezidua proměnných by měly prokazovat dostatečnou těsnost proložení. Není-li tomu tak, je třeba se navrátit k předúpravě dat a celý výpočet PCA opakovat.

Vzorová úloha 1. *Sledování spotřeby proteinů v Evropě*

Sledováná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem vyšetření: existuje korelace mezi proměnnými? Budou data vyžadovat nějakou úpravu, standardizaci nebo centrování? Ukazuje graf komponentních vah na silně korelující proměnné? Jsou některé proměnné redundandní? Lze odhalit v rozptylovém diagramu komponentního skóre odlehlé objekty, výjimečné co do spotřeby proteinů? Které země jsou si podobné ve spotřebě proteinů? Komentujte vzniklé shluky zemí co do spotřeby proteinů. Dají se ve dvojném grafu odhalit interakce mezi jednotlivými proměnnými-druhy potravin a objekty-zeměmi?

Data: i značí index, ***Cervene*** červené maso, **Bílé** maso, **Vejce, Mléko, Ryby, Obilniny, Škrob, Ořechy, Ovoce** a zelenina,

	Objekty	Proměnné								
i	Stát	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
1	**Albania**	10.10	1.40	0.50	8.90	0.20	42.30	0.60	5.50	1.70
2	**Austria**	8.90	14.00	4.30	19.90	2.10	28.00	3.60	1.30	4.30
3	**Belgium**	13.50	9.30	4.10	17.50	4.50	26.60	5.70	2.10	4.00
4	**Bulgaria**	7.80	6.00	1.60	8.30	1.20	56.70	1.10	3.70	4.20
5	**Czechoslov.**	9.70	11.40	2.80	12.50	2.00	34.30	5.00	1.10	4.00
6	**Denmark**	10.60	10.80	3.70	25.00	9.90	21.90	4.80	0.70	2.40
7	**E Germany**	8.40	11.60	3.70	11.10	5.40	24.60	6.50	0.80	3.60
8	**Finland**	9.50	4.90	2.70	33.70	5.80	26.30	5.10	1.00	1.40
9	**France**	18.00	9.90	3.30	19.50	5.70	28.10	4.80	2.40	6.50
10	**Greece**	10.20	3.00	2.80	17.60	5.90	41.70	2.20	7.80	6.50
11	**Hungary**	5.30	12.40	2.90	9.70	0.30	40.10	4.00	5.40	4.20
12	**Ireland**	13.90	10.00	4.70	25.80	2.20	24.00	6.20	1.60	2.90
13	**Italy**	9.00	5.10	2.90	13.70	3.40	36.80	2.10	4.30	6.70
14	**Netherlands**	9.50	13.60	3.60	23.40	2.50	22.40	4.20	1.80	3.70
15	**Norway**	9.40	4.70	2.70	23.30	9.70	23.00	4.60	1.60	2.70
16	**Poland**	6.90	10.20	2.70	19.30	3.00	36.10	5.90	2.00	6.60
17	**Portugal**	6.20	3.70	1.10	4.90	14.20	27.00	5.90	4.70	7.90
18	**Romania**	6.20	6.30	1.50	11.10	1.00	49.60	3.10	5.30	2.80
19	**Spain**	7.10	3.40	3.10	8.60	7.00	29.20	5.70	5.90	7.20
20	**Sweden**	9.90	7.80	3.50	24.70	7.50	19.50	3.70	1.40	2.00
21	**Switzerland**	13.10	10.10	3.10	23.80	2.30	25.60	2.80	2.40	4.90
22	**UK**	17.40	5.70	4.70	20.60	4.30	24.30	4.70	3.40	3.30
23	**USSR**	9.30	4.60	2.10	16.60	3.00	43.60	6.40	3.40	2.90

24	**W Germany**	11.40	12.50	4.10	18.80	3.40	18.60	5.20	1.50	3.80
25	**Yugoslavia**	4.40	5.00	1.20	9.50	0.60	55.90	3.00	5.70	3.20

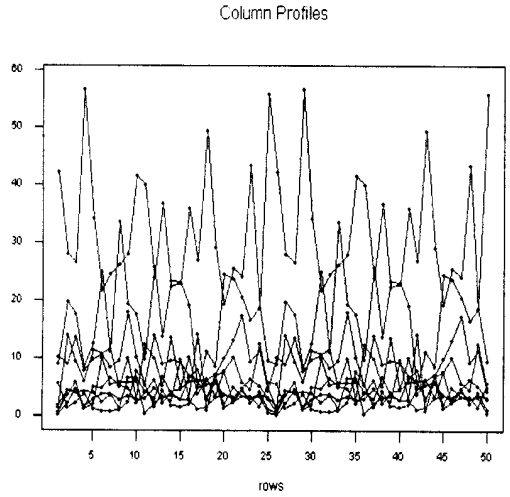
Řešení: k analýze byl použit program NCSS2000.

1. Exploratorní analýza: Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především *symbolové grafy*. Jednotlivé proměnné jsou v nich "kódovány" s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů, *symbolů*. Každému objektu $\boldsymbol{x}_i$ (např. léčivu, autu, sloučenině) tak odpovídá jistý obrazec zvaný *symbol*. Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi symboly. Tím lze v jednom grafu rozlišit *více proměnných* x_j, $j = 1, ..., m$. Prvním krokem před vlastním zobrazením do symbolů je obvykle standardizace. Mezi základní typy zobrazovaných symbolů patří *profily*, *polygony*, *tváře*, *křivky* a *stromy*.

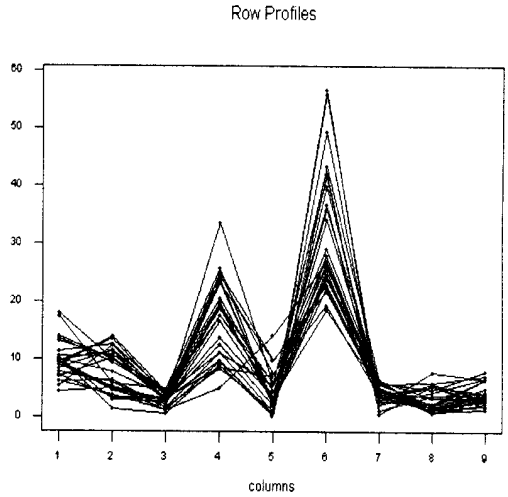
Profily představují dvourozměrné zobrazení m-rozměrných objektů. Každý objekt $\boldsymbol{x}_i$ je charakterizován m proměnnými, zobrazenými zde vertikálními úsečkami. Jejich velikost je úměrná hodnotě odpovídající proměnné x_{ij}, $j = 1, ..., m$. Profil pak vzniká spojením koncových bodů těchto úseček. Je vhodné použít standardizované proměnné dle vzorce

$$x_{ij}^{*} = \frac{x_{ij}}{(\max\limits_{i} |x_{ij}|)} ,$$

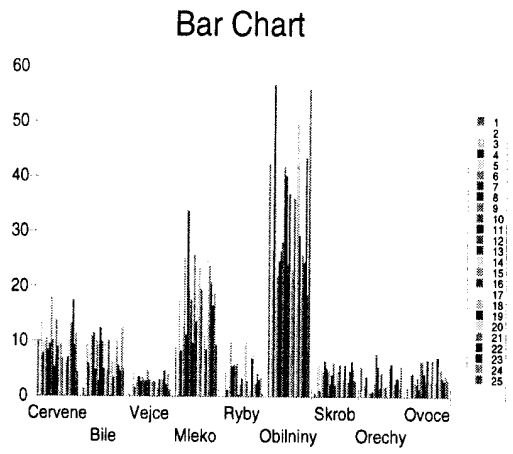
kde max $| x_{ij} |$ je maximální hodnota absolutní velikosti proměnné $\boldsymbol{x}_j$ vektoru $\boldsymbol{x}_i^T$ přes všechny body, $i = 1, ..., n$. Profily jsou jednoduché a umožňují snadné určení rozdílů mezi jednotlivými objekty $\boldsymbol{x}_i$ a $\boldsymbol{x}_k$. Snadno lze takto identifikovat vybočující objekt.



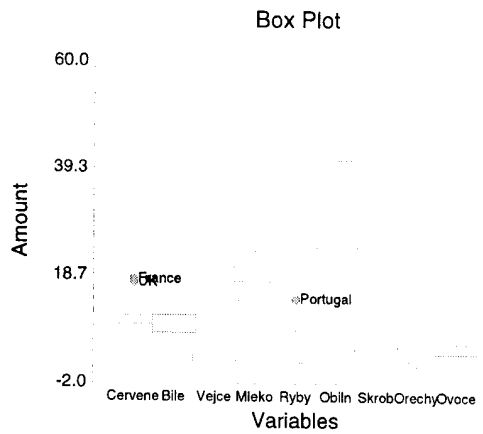
Obr. 1 Sloupcové profily původních dat



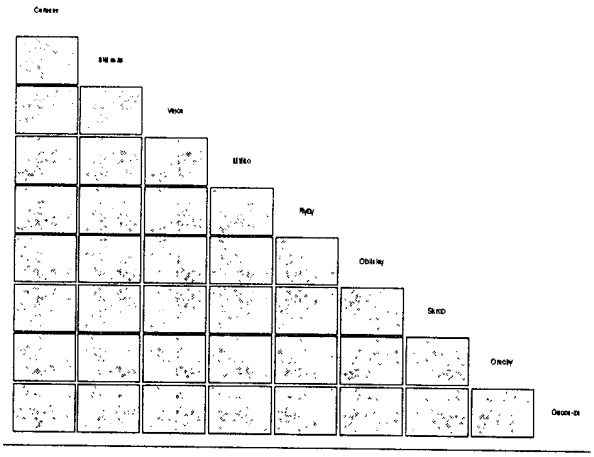
Obr. 2 Řádkové profily původních dat



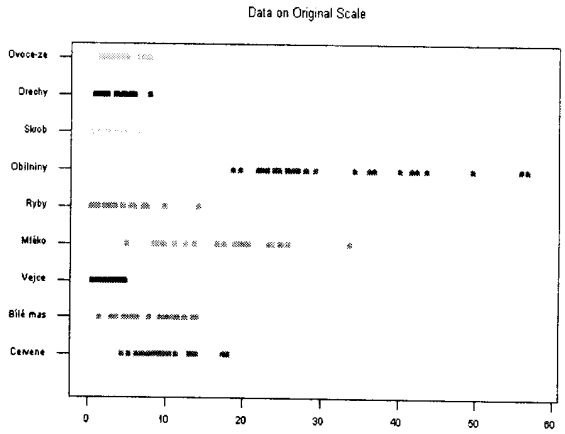
Obr. 3 Sloupcový diagram původních dat



Obr. 4 Krabicové grafy původních dat



Obr. 5 Rozptylový diagram korelační matice



Obr. 6 Rozptylová diagram původních dat

Polygony jsou vlastně profily v polárních souřadnicích, kdy každá proměnná objektu x_i^T, $i = 1, ..., n$, odpovídá délce paprsku vycházejícího ze společného středu. Paprsky dělí kružnici ekvidistantně, proměnné jsou standardizovány do intervalu [0, 1]. Mezi polygony patří *graf slunečních paprsků* a *hvězdicový graf.*

(a) **Graf slunečních paprsků** má tvar "sluníčka", které se skládá z paprsků, začínajících ve společném bodě a spojujících úseček mezi paprsky, které tak tvoří polygon. Zde každá proměnná x_{ij} objektu x_i^T odpovídá délce paprsku vycházejícího ze středu sluníčka. Paprsky jsou rozmístěny ekvidistantně, ve stejných vzdálenostech na kružnici, a proto se provádí lineární transformace do intervalu $[a, 1]$, kde a je zvolená spodní mez, obyčejně $a = 0$. Pro tuto transformaci platí, že

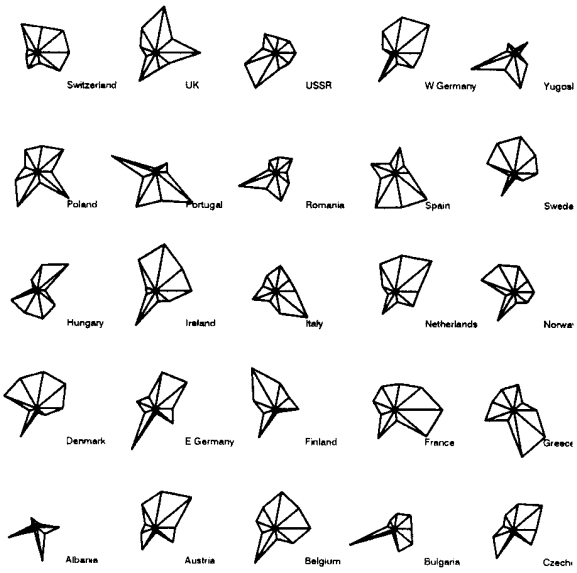
$$x_{ij}^* = \frac{(1 - a)(x_{ij} - \min_i x_{ij})}{\max_i x_{ij} - \min_i x_{ij}} + a$$

kde $\min x_{ij}$ je minimální a $\max x_{ij}$ maximální hodnota j-té proměnné objektu x_i^T přes všechny objekty x_i^T, $i = 1, ..., n$. K určení směrů jednotlivých paprsků se definuje jejich úhel α_j, pro který platí

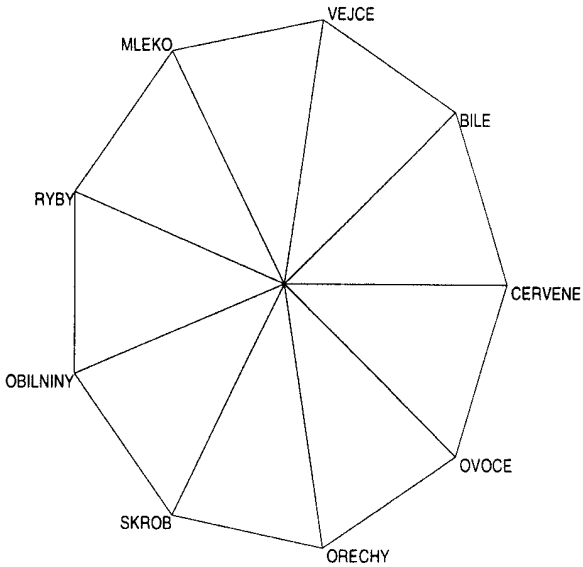
$$\alpha_j = \frac{2\pi(j - 1)}{m}, \qquad j = 1, ..., m$$

Za společný střed paprsků se obyčejně volí počátek souřadnic. Pokud má být maximální délka paprsků rovna R, je polygon pro objekt $\boldsymbol{x}_i^{\mathbf{T}}$ spojnicí m bodů p_{ij} o souřadnicích $p_{ij} = (x_{ij}^* R \cos\alpha_j, x_{ij}^* R \sin\alpha_j)$. Aby vznikl uzavřený obrazec, spojují se ještě první a poslední bod p_{i1} a p_{im}. Vzájemné porovnání polygonů slouží k vizuálnímu posouzení podobnosti objektů. V případě velkého počtu proměnných, např. $m > 6$, bývá však výsledný obrázek polygonů nepřehledný.

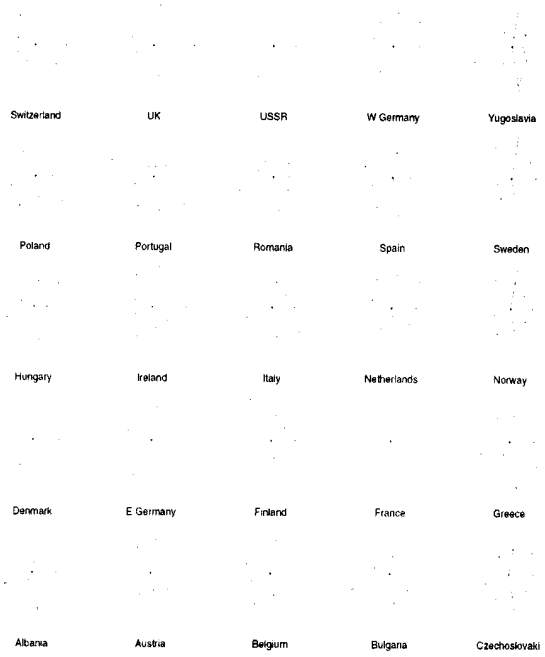
(b) **Hvězdicový graf** vypadá na první pohled jako předchozí graf sluníčka. Sestává z paprsků, reprezentujících relativní hodnoty proměnných u jednotlivých objektů, které se pro každý objekt spojují v jednom centrálním bodě.



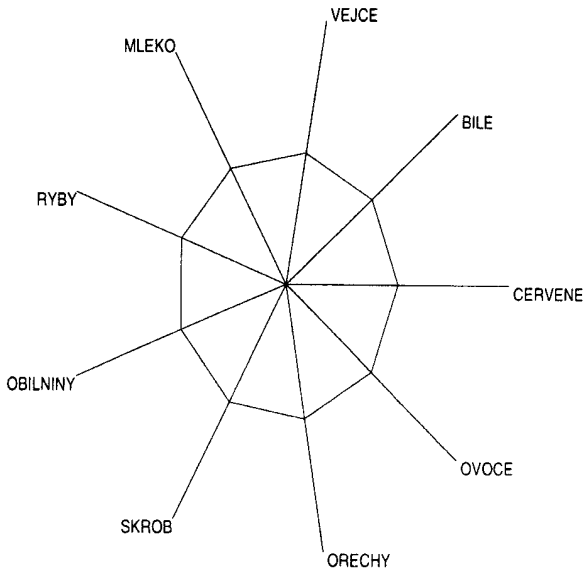
Obr. 7a Hvězdičkový graf



Obr. 7b Klíč k hvězdičkovému grafu



Obr. 8a Sluníčkový graf

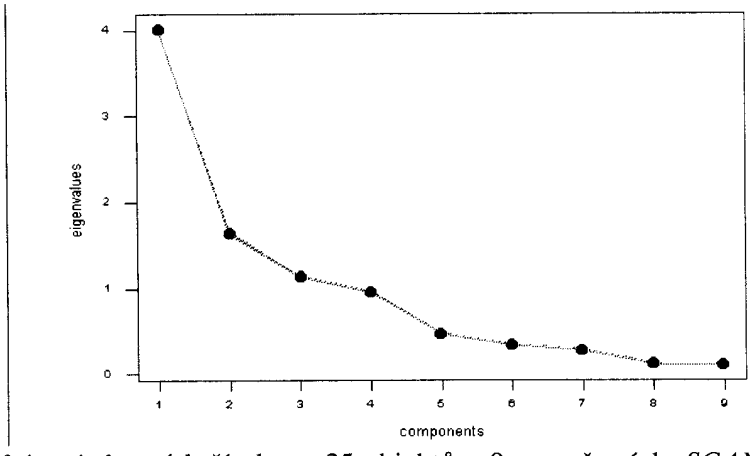


Obr. 8b Klíč ke sluníčkovému grafu

Stejně směřující paprsky u různých objektů se liší svojí délkou. *Nejkratší paprsek* indikuje, že u objektu nabývá příslušná proměnná nejmenší hodnoty z celého výběru. Podobně *nejdelší paprsek* informuje o nejvyšší hodnotě příslušné proměnné. Délky ostatních paprsků se pohybují podle relativní velikosti hodnot proměnné u příslušného objektu mezi těmito dvěma krajními mezemi.

2. Metoda hlavních komponent:

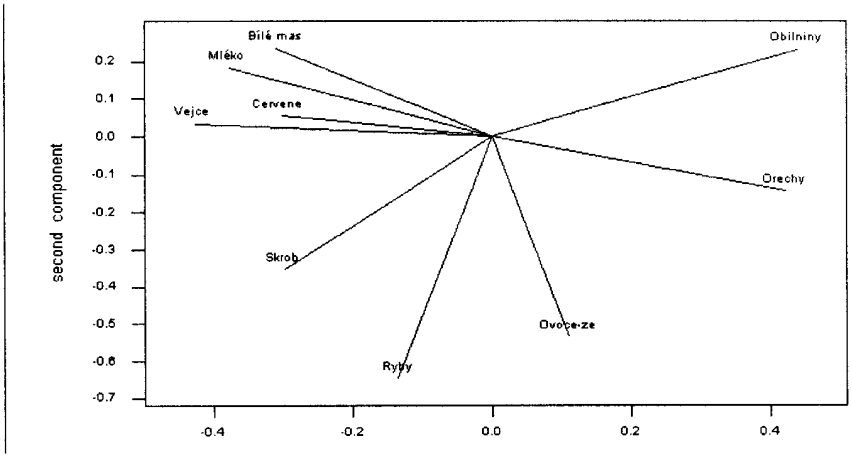
1. Vyšetření indexového grafu úpatí vlastních čísel:



Obr. 9 Indexový graf úpatí vlastních čísel pro 25 objektů a 9 proměnných, *SCAN*

Vlastní čísla slouží k určení počtu A "využitelných" hlavních komponent, jež si zvolíme v analýze k dalšímu užívání. Procento a kumulativní procento popisuje proměnlivost v původních proměnných, popsanou dotyčnou hlavní komponentou. K dalšímu popisu proměnlivosti bereme obvykle tolik hlavních komponent, aby bylo jimi popsáno 90 až 99% celkové proměnlivosti. V tomto případě stačí užít první dvě. Indexový graf úpatí vlastních čísel je vlastně sloupcový diagram velikosti vlastních čísel proti stoupající hodnotě indexu, pořadového čísla. Zobrazuje relativní velikost jednotlivých vlastních čísel. Užitečné komponenty jsou tak odděleny zřetelným zlomovým místem, a x-ová souřadnice tohoto zlomu je hledaná hodnota indexu.

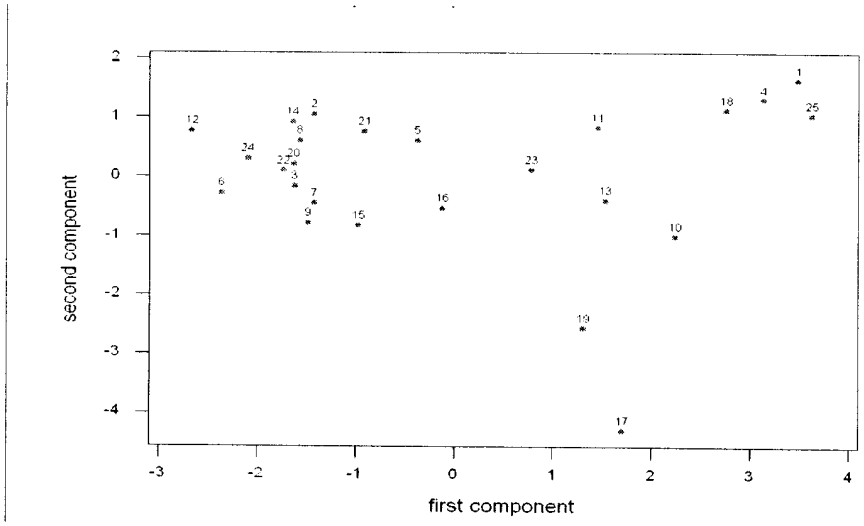
2. Vyšetření grafu komponentních vah: tento graf ukazuje, že proměnné *Bílé maso-Mléko-Červené maso-Vejce* spolu silně korelují.



Obr. 10 Graf komponentních vah (Components Weights Plot)

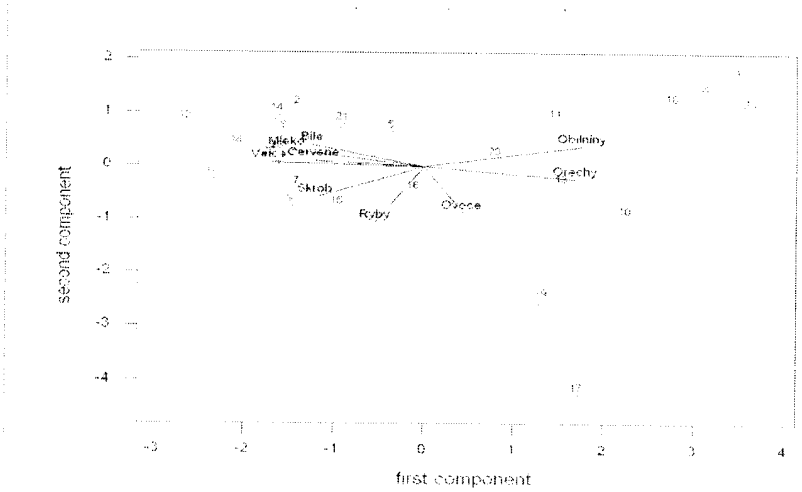
Ukazuje se, že by bylo vhodné jejich počet zredukovat na dvě proměnné, např. *Bílé maso-Vejce.* Mezi průvodiči ostatních proměnných je dostatečně velký úhel, a tím malá korelace. Proměnné, které spolu nekorelují mohou být ponechány ve vstupních datech.

3. Vyšetření rozptylového diagramu komponentního skóre: nejdůležitější diagram metody hlavních komponent ukazuje celou vyšetřovanou strukturu objektů, tzn. shluky objektů, izolované objekty, odlehlé objekty, anomálie, atd. Objekty mohou být označeny textovým popisem nebo číselně indexem. V pravém horním rohu se dobře oddělil shluk objektů: *1(Albanie), 4 (Bulharsko), 18 (Rumunsko), 25(Jugoslávie)*, který pokrývá země Balkánu. Vyjímečné postavení mají *17(Portugalsko), 19(Španělsko)*. Ostatní státy jsou kromě *11, 13, 23* v jednom společném shluku.



Obr. 11 Rozptylový diagram komponentního skóre (Scatterplot)

4. Vyšetření dvojného grafu: je důležité sledovat interakci objektů a proměnných. Je-li některý objekt umístěn ve dvojném grafu na stejném místě nebo alespoň poblíž místa proměnné, jsou spolu v interakci. Interakce poslouží interpretaci objektů.



Obr. 12 Dvojný graf (Biplot)

Klasifikace objektů analýzou shluků

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do jedné existující třídy (*diskriminační analýza DA*), nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd či shluků (*analýza shluků CLU*).

Analýza shluků patří mezi metody, které se zabývají vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Podle způsobu shlukování se postupy dělí na *hierarchické* a *nehierarchické*. Hierarchické se dělí dále na *aglomerativní* a *divizní.*

Hierarchické postupy jsou založeny na postupném spojování objektů a jejich shluků do dalších, větších shluků. Nejprve se vypočte základní matice vzdáleností mezi objekty. U *aglomerativního shlukování* se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako celek. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. *Divizní postup* je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů. Výhodou hierarchických metod je nepotřebnost informace o optimálním počtu shluků v procesu shlukování, tento počet se určuje až dodatečně. Při shlukování vznikají dva základní problémy:

(a) *způsob měření vzdáleností mezi objekty:* i když existuje celá řada měr vzdáleností (vícerozměrných metrik), nejčastěji se užívá *euklidovská metrika*, která je přirozeným zobecnením běžného pojmu vzdálenosti;

(b) *volba vhodné shlukovací procedury* dle zvoleného způsobu metriky. Metody metriky shlukování jsou

Metoda průměrová (Average): vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy mezishlukovou vzdáleností objektů se rozumí vzdálenost dvou objektů, z nichž každý patří do jiného shluku. Nejbližší jsou shluky, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jednoho a všemi objekty druhého shluku. Dendrogramy mají strukturu podobnou dendrogramům metody nejvzdálenějšího souseda, pouze spojení je provedeno při obvykle vyšších vzdálenostech.

Metoda centroidní (Centroid): vzdálenost shluků se počítá jako euklidovská vzdálenost jejich těžišť. Nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi těžišti.

Metoda nejbližšího souseda (Single, Nearest): kritériem pro vytváření shluků je minimum z možných mezishlukových vzdáleností objektů. Metoda tvoří nový shluk na základě nejkratší vzdálenosti mezi shluky (či objekty) a neumí proto rozlišit špatně separované shluky. Je zde silná tendence ke tvorbě řetězců. Řetězování může vésti až ke zcela mylným závěrům, jsou-li objekty na opačních koncích řetězce zcela nepodobné. Na druhé straně je to jedna z mála metod, která umí roztřídit, rozlišit i neeliptické shluky.

Metoda nejvzdálenějšího souseda (Complete, Furthest): počítá vzdálenost

dvou shluků jako maximum z možných mezishlukových vzdáleností objektů. Probíhá podobně jako metoda Single s jednou důležitou výjimkou: vzdálenost (či nepodobnost) mezi shluky je určována vzdáleností (či nepodobností) mezi dvěma nejvzdálenějšími objekty, každý přitom z jiného shluku. Proto všechny objekty ve shluku jsou na základě maximální vzdálenosti či minimální podobnosti vůči objektům ve druhém shluku.

Metoda mediánová (Median): jde o jisté vylepšení centroidní metody, neboť se snaží odstranit rozdílné "váhy", které centroidní metoda dává různě velkým shlukům.

Wardova metoda je založena na minimalizaci ztráty informace při spojení dvou tříd. V každém kroku je uvažován takový možný pár objektů (či shluků), aby suma čtverců odchylek od střední hodnoty ESS = $\sum_{i=1}^{n} (x_i - \bar{x})^2$ dosáhla při vzniku shluku svého minima.

Nehierarchické shlukovací metody: u *metody typických bodů* (Seeded) uživatel na základě svých věcných znalostí určí, které objekty mají být "typickými" představiteli nově vytvořených shluků a systém pak rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů. V nehierarchických shlukovacích metodách je počet shluků obvykle předem dán, i když se v průběhu výpočtu může změnit. Zůstává-li počet shluků zachován, hovoříme o nehierarchických metodách s *konstantním počtem shluků*, v opačném případě o nehierarchických metodách s *optimalizovaným počtem shluků*. Nehierarchické metody zahrnují dvě základní varianty - optimalizační metody a analýzu módů, medoidů. *Optimalizační nehierarchické metody* hledají optimální rozklad přeřazováním objektů ze shluku do shluku s cílem minimalizovat nebo maximalizovat nějakou charakteristiku rozkladu. Metody, označované jako *analýza módů, medoidů* představují hledání rozkladu do shluků, kde shluky jsou chápány jako místa se zvýšenou koncentrací objektů v m-rozměrném prostoru proměnných.

Místo výchozí matice vzdáleností může být v některých případech ke shlukování použita i *korelační matice.*

(a) Hierarchické shlukování

Analýza shluků patří mezi metody, zabývající se vyšetřováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich roztříděním do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Analýzou shluků budeme sledovat a vyšetřovat podobnost objektů, analyzovanou pomocí *dendrogramu objektů*, a jednak podobnost proměnných analyzovanou pomocí *dendrogramu proměnných.*

Dendrogram, diagram shluků nebo *vývojový strom* se objeví pouze v případě zadání hodnot původních proměnných a nikoli maticí vzdáleností. Výsledkem je zobrazení hodnot ve dvojrozměrném prostoru, kde osy tvoří zadané proměnné. Objeví se také "obkroužení" objektů v jednotlivých shlucích.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Dendrogram podobnosti proměnných odhaluje nejčastěji dvojice či trojice (obecně m-tice) proměnných, které jsou si velmi podobné a silně spolu korelují.

Odhaluje proměnné, které jsou ve společném shluku, které jsou si tím pádem značně podobné a které jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti (či proměnné) není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

Míra věrohodnosti: dendrogram lze sestrojit celou řadu technik. Prvním kritériem jeho věrohodnosti čili těsnosti proložení, jež nejlépe odpovídá struktuře objektů a proměnných mezi objekty je *kofenetický korelační koeficient CC*. Je to druh korelačního koeficientu mezi skutečnou a dendrogramem predikovanou vzdáleností. Je-li tato hodnota větší než 0.75, je obvykle nulová hypotéza o dané struktuře zamítnuta. Hodnota 0.9 svědčí, že dendrogram vůbec neodpovídá skutečné struktuře dat.

Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření, distorze spíše než stupeň podobnosti. Kritérium delta je definované vztahem

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d_{jk}^*|^{1/A}}{\sum_{j<k}^{N} (d_{jk}^*)^{1/A}} \right]^A$$

kde $A = 0.5$ nebo 1 a d^*_{ij} je vzdálenost, získaná z dendrogramu. Hodnoty *delta* blízké nule jsou žádoucí. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

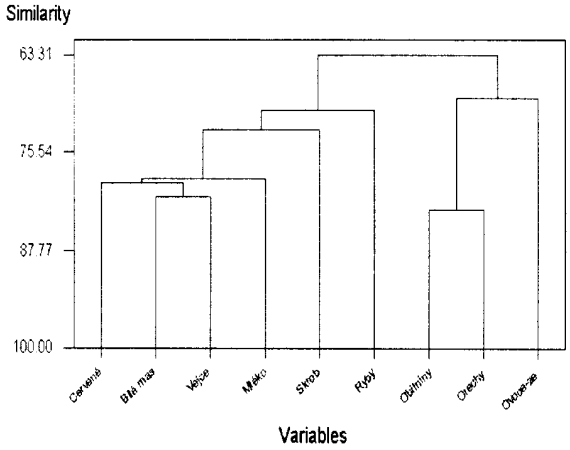
Postup shlukové analýzy

1. *Volba vstupní databáze*: zadává se typ dat (a) proměnných (sloupců) analyzovaných objektů (řádků), (b) sloupců matice vzdáleností, (c) sloupců korelační matice.
2. *Volba druhu veličin:* zadává se typ užitých veličin v datech, která mohou být (a) intervalová, (b) ordinální, (c) nominální, (d) symetrická binární, (e) asymetrická binární, (f) poměrová.
3. *Název objektů:* zadání pojmenování či jmen jednotlivých objektů, umístěných v řádcích, které se mohou objevit v dendrogramu místo indexů (pořadových čísel) objektů.
4. *Typ shlukovací techniky:* volba metody z možností: jednoduchá průměrová (Average), skupinového průměru, centroidní (Centroid), nejbližšího souseda (Single, Nearest), nejvzdálenějšího souseda (Complete, Furthest), mediánová (Median), Wardova, a flexibilní.
5. *Volí se druh užité vzdálenosti:* vzdálenosti mohou být Eukleidova metrika čili geometrická vzdálenost, Hammingova metrika čili Manhattanská vzdálenost, zobecněná Minkowskiho metrika a Mahalanobisova metrika.
6. *Postup linkování a zařazení do shluků:* tabelární výpočet vzdáleností (nebo podobností) mezi objekty a shluky a postupné vytváření dendrogramu. Postupy jsou (1) metodou hierarchického shlukování, (2) shlukování metodou nejbližších středů, (3) shlukování metodou středů-medoidů, a (4) metodou fuzzy shlukování.

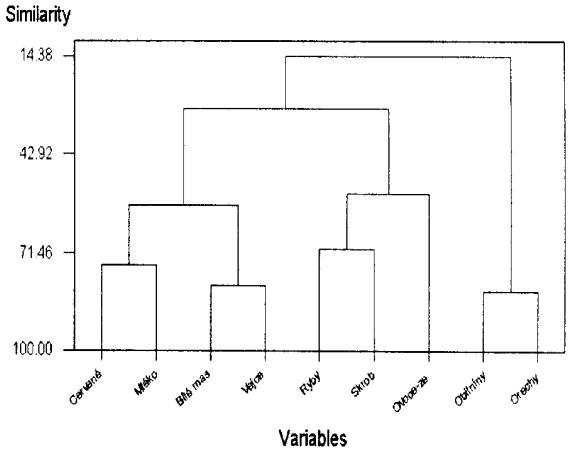
Vzorová úloha 2. *Klasifikace spotřeby proteinů v Evropě*

Sledováná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem klasifikace. Které země jsou si podobné ve spotřebě proteinů? Komentujte vzniklé shluky zemí co do spotřeby proteinů.

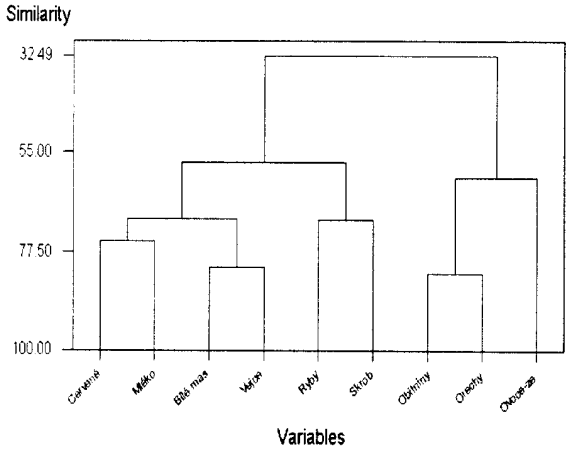
Řešení: Dendrogram podobnosti proměnných obsahuje dvojice nebo trojice proměnných, které jsou si velmi podobné a silně spolu korelují. Metodou nejbližšího souseda jsou to trojice *Červené maso-Bílé maso-Vejce*, k tomu se přidružuje také proměnná *Mléko.* Další dvojice je *Obilniny-Ořechy* (obr. 13a). Metoda nejvzdálenějšího souseda ukazuje na 4 dvojice: první dvojice *Červené maso-Mléko*, druhá *Bílé maso-Vejce,* třetí *Ryby-Škrob,* čtvrtá *Obilniny-Ořechy* (obr. 13b). Matoda Wardova poskytla stejné shluky jako metoda nejvzdálenějšího souseda.



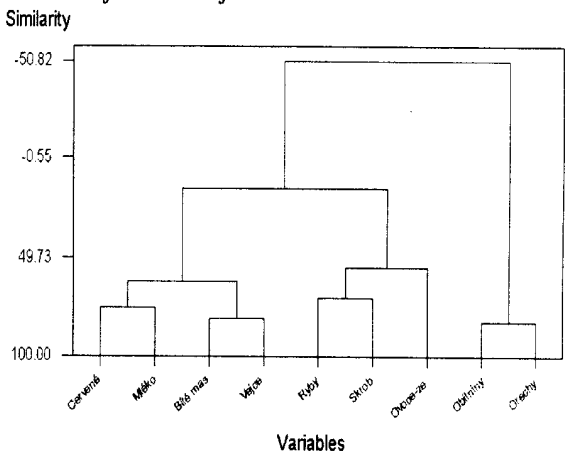
Obr. 13a Dendrogram proměnných metodou nejbližšího souseda



Obr. 13b Dendrogram proměnných metodou nejvzdálenějšího souseda

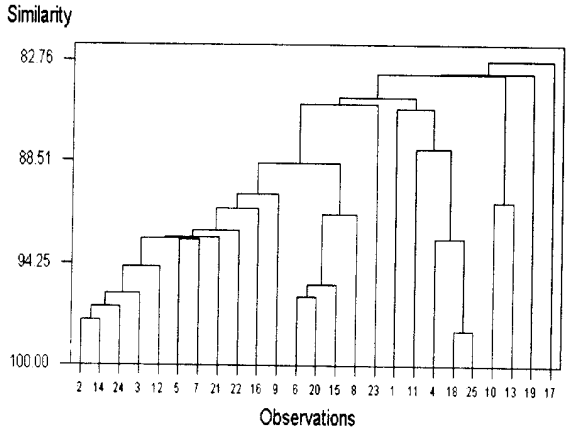


Obr. 13c Dendrogram proměnných metodou průměrovou

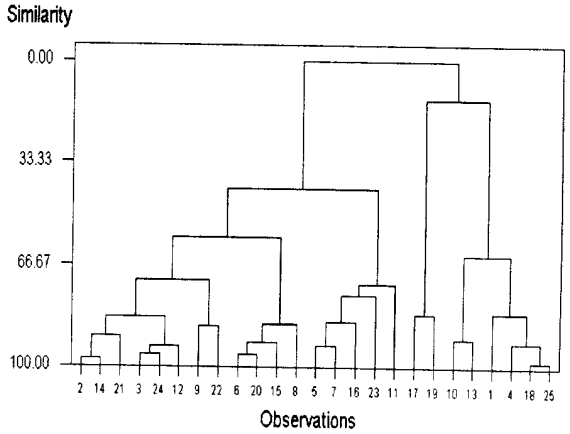


Obr. 13d Dendrogram proměnných metodou Wardovou

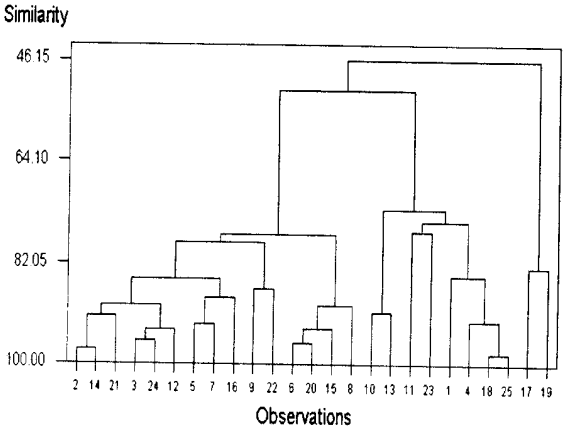
Nejdůležitějším dendrogramem je dendrogram podobnosti objektů, ze kterého je patrná struktura objektů ve shlucích, roztřídění států Evropy dle spotřeby proteinů na základě 9 kritérií a vzájemné podbnosti.



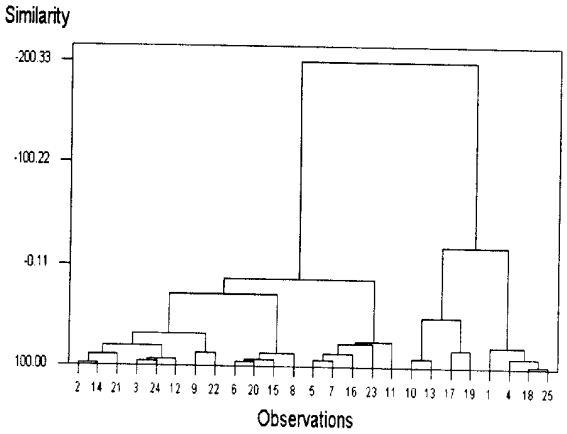
Obr. 14a Dendrogram objektů metodou nejbližšího souseda, Minitab



Obr. 14b Dendrogram objektů metodou nejvzdálenějšího souseda, Minitab



Obr. 14c Dendrogram objektů metodou průměrovou, Minitab



Obr. 14d Dendrogram objektů metodou Wardovou, Minitab

Vzorová úloha 3. *Klasifikace zdrojů pitné vody*

Na 62 vzorcích zdrojů pitné vody bylo stanoveno 16 proměnných kvality. Je třeba vyšetřit, zda krabicový graf ukazuje nutnost data standardizovat, zda lze nalézt *vybočující objekty*, resp. jejich proměnné, zda existuje korelace mezi proměnnými, zda ukazuje graf komponentních vah na korelující proměnné, zda jsou některé proměnné redundandní, zda lze odhalit v rozptylovém diagramu komponentního skóre odlehlé objekty, zda lze posoudit *podobnost objektů* shlukovou analýzou klasifikaci zdrojů.

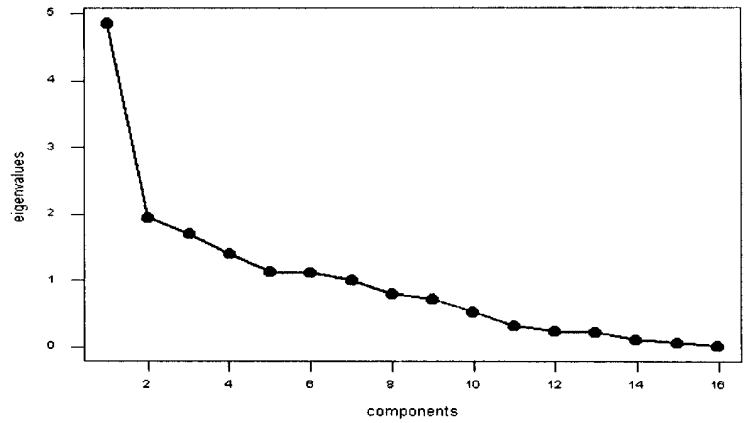
Data: *E404i* index vzorku, *E404x1* obsah dusičnanů [mg/l], *E404x2* obsah dusitanů [mg/l], *E404x3* obsah chloridů [mg/l], *E404x4* obsah celkového chloru [mg/l], *E404x5* obsah síranů [mg/l], *E404x6* obsah fosforečnanů [mg/l], *E404x7* obsah amonných solí [mg/l], *E404x8* obsah vápníku [mg/l], *E404x9* obsah hořčíku [mg/l], *E404x10* obsah železa (celkového) [mg/l], *E404x11* obsah manganu [mg/l], *E404x12* pH, *E404x13* KNK, *E404x14* ZNK, *E404x15* vodivost, *E404x16* nerozpuštěné látky [mg/l].

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}
1	2.2	0.00	6.	6.	103.5	0.03	0.02	181	17	0.016	0.05	7.08	8.1	3.40	855	0.09
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
62	32.8	0.01	25.	25.	115.5	0.05	0.02	102	12	0.016	0.05	7.69	2.6	0.65	436	0.05

Řešení:

A. Metoda hlavních komponent

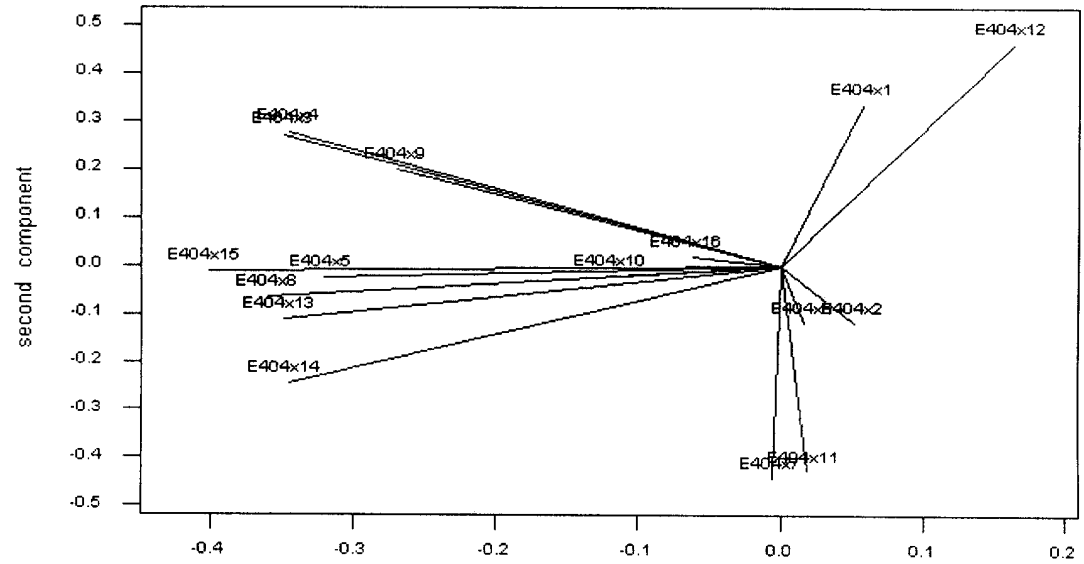
1. Vyšetření indexového grafu úpatí vlastních čísel:



Obr. 15 Indexový graf úpatí vlastních čísel pro 62 objektů a 16 proměnných, *SCAN*

Užitečné komponenty jsou na obr. 15 odděleny zřetelným zlomovým místem, a x-ová souřadnice tohoto zlomu je hledaná hodnota indexu 2.

2. Vyšetření grafu komponentních vah: provede se porovnáním vzdáleností mezi proměnnými a dospěje se k závěru, že krátká vzdálenost mezi dvěma proměnnými znamená silnou korelaci.

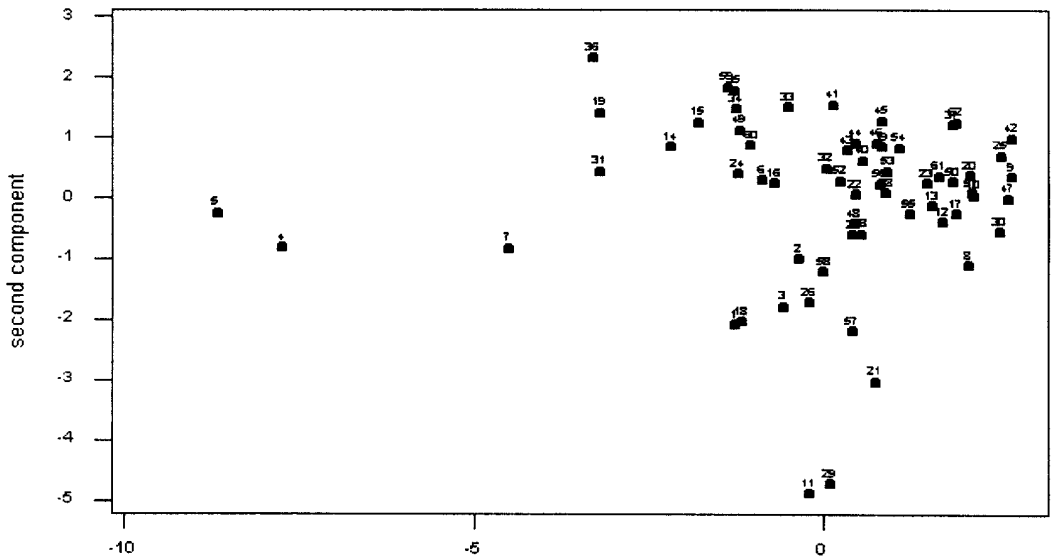


Obr. 16 Graf komponentních vah (Components Weights Plot)

Graf ukazuje, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent. Lze názorně vysvětlit, jak jednotlivé původní proměnné x_j, $j = 1,..., 16$, přispívají do první hlavní komponenty y_1 nebo do druhé hlavní komponenty y_2. Některé původní proměnné x_j přispívají kladnou vahou, některé zápornou. Původní proměnné x_j, $j = 1,..., 16$, blízko sebe a nebo proměnné x_j s malým úhlem mezi svými průvodiči proměnných a na stejné straně vůči počátku mají vysokou kladnou kovarianci a vysokou kladnou korelaci. Naopak, původní proměnné x_j daleko od sebe anebo s velikým úhlem mezi průvodiči proměnných jsou negativně

korelovány. Ukazuje se užitečné, aby původní proměnné, které spolu silně korelují byly z dat odstraněny.

3. Vyšetření rozptylového diagramu komponentního skóre: nejdůležitější diagram metody hlavních komponent ukazuje celou vyšetřovanou strukturu objektů, tzn. shluky objektů, izolované objekty, odlehlé objekty, anomálie, atd.



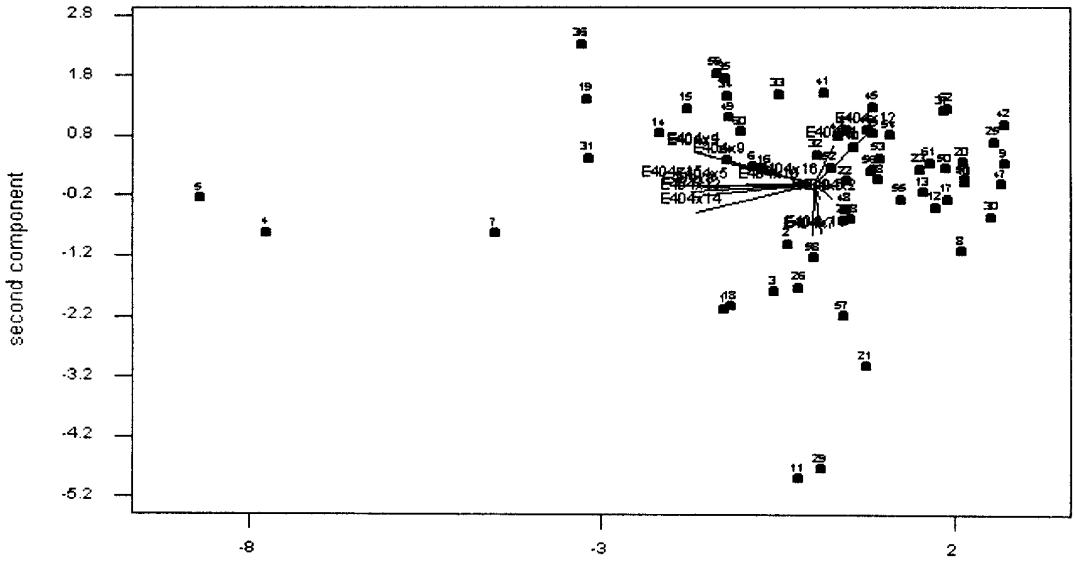
Obr. 17 Rozptylový diagram komponentního skóre (Scatterplot)

Objekty mohou být označeny textovým popisem nebo jako na obr. 17 číselně indexem. Lze dospět k těmto závěrům:

1. *Umístění objektů*: objekty daleko od počátku (4, 5, 7, 11, 29, atd.) jsou extrémy. Objekty nejblíže počátku (22, 52, 53, 54, 45, atd.) jsou typičtější.
2. *Podobnost objektů*: objekty blízko sebe (např. 53 a 54 a 44 a 45) si jsou podobné, objekty daleko od sebe (např. 5 a 45, 4 a 30, atd) jsou si nepodobné.
3. *Objekty v shluku*: objekty umístěné zřetelně v jednom shluku (např. 12, 13, 17, 55, 23, 61, 50, 20, atd) jsou si podobné a přitom nepodobné objektům v ostatních shlucích (např. 14, 15, 24, 49, atd.). Dobře oddělené shluky prozrazují, že lze nalézt vlastní model pro samotný shluk. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.
4. *Osamělé objekty:* izolované objekty (4, 5, 7) mohou být odlehlé objekty, které jsou silně nepodobné ostatním objektům.
5. *Odlehlé objekty:* v ideálním případě bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je něco špatného v modelu, obyčejně je přítomen silně odlehlý objekt (4, 5, 11, 29). Odlehlé objekty jsou totiž schopny zbortit celý diagram, ve srovnání se silně vybočujícím objektem jsou ostatní objekty nakumulovány do jediného úzkého shluku. Po odstranění vybočujícího objektu se ostatní objekty roztřídí po celé ploše diagramu a teprve vypovídají o existujících shlucích.

6. Vyšetření dvojného grafu: je důležité sledovat interakci objektů a proměnných na obr. 18. Je-li některý objekt umístěn ve dvojném grafu na stejném místě nebo

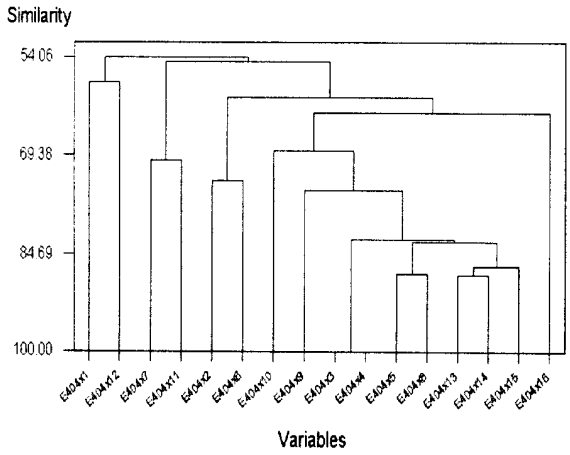
alespoň poblíž místa proměnné, jsou spolu v interakci. Interakce poslouží interpretaci objektů. Dokonalé rozptýlení objektů v rovině obou hlavních komponent vede k rozlišení objektů při jejich popisu pomoci y_1 a y_2.



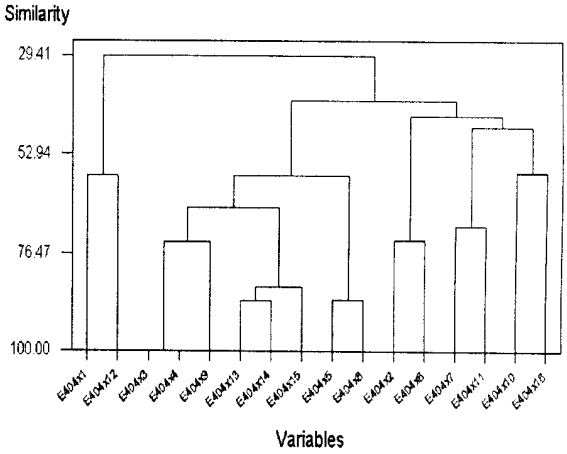
Obr. 18 Dvojný graf (Biplot)

B. Klasifikační metoda shluků

V prvém stádiu se vyšetřuje podobnost proměnných, a tím se také odhaluje silná korelace proměnných. Odhalí se, která původní proměnná je nadbytečná a kterou lze vynechat a nahradit jinou. Čím nižší je spojka dvou objektů, tím jsou si objekty podobnější.

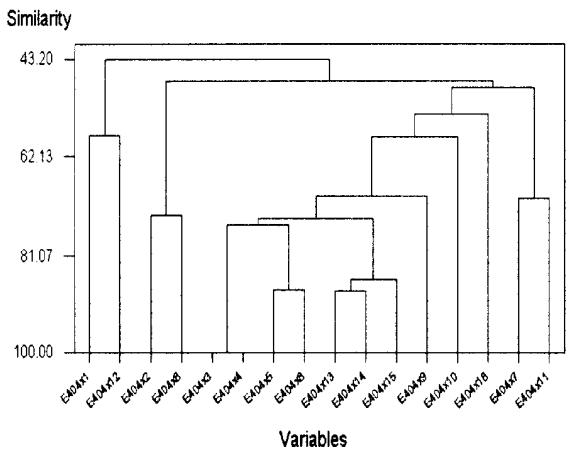


Obr. 19a Dendrogram proměnných metodou nejbližšího souseda

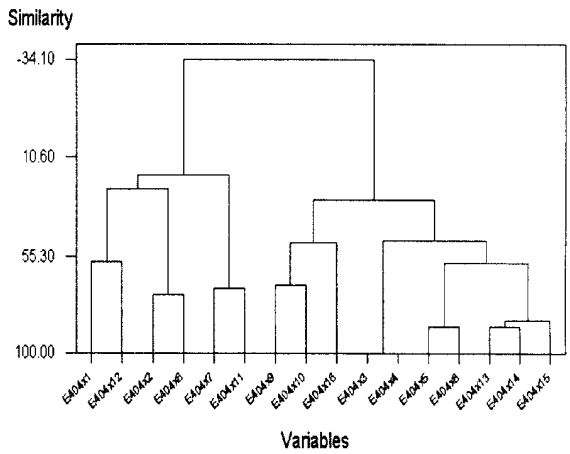


Obr. 19b Dendrogram proměnných metodou nejvzdálenějšího souseda

Metodou nejbližšího souseda lze nalézt šest dvojic velice si podobných proměnných. Nejvíce jsou si podobné proměnné ve dvojicích: *x13(KNK)-x14(ZNK)* a *x5(SO_4^{2-})-x8(Ca)*. Poněkud ménějsou si podobné proměnné ve dvojici *x2(NO_2^-)-x6(PO_4^{3-})* a také ve dvojici *x7(NH_4^+)-x11(Mn)*. Nejméně jsou si podobné proměnné ve dvojici *x1(NO_3^-)-x12(pH)*. Výjimečné postavení má proměnná *x16(#sraženina)*, která si není podobná s žádnou jinou proměnnou.



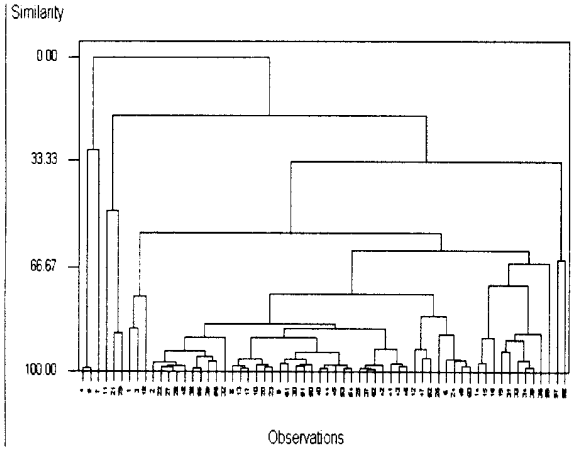
Obr. 19c Dendrogram proměnných metodou průměrovou



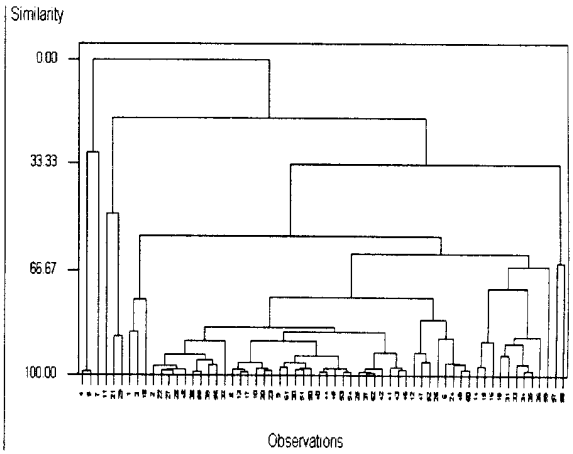
Obr. 19d Dendrogram proměnných metodou Wardovou

Wardova metoda je jednou z nejnáročnějších ve výpočtu a také nejpřísnější na tvorbu shluků. Odhalila šest shluků, šest dvojich podobných proměnných. Dospěla k podobným závěrům jako metoda nejbližšího souseda.

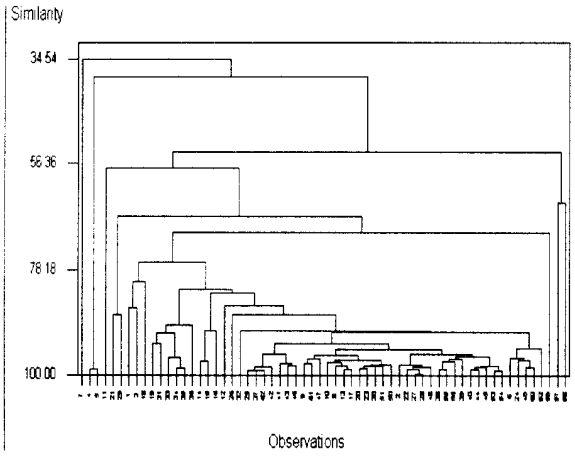
V druhém stádiu klasifikace tvorbou shluků se vyšetřuje podobnost objektů-zdrojů pitné vody, jako nejdůležitější část klasifikační analýzy. Jde o odhalení vybočujících zdrojů, které jsou silně nepodobné ostatním, které mají anomální hodnoty proměnných.



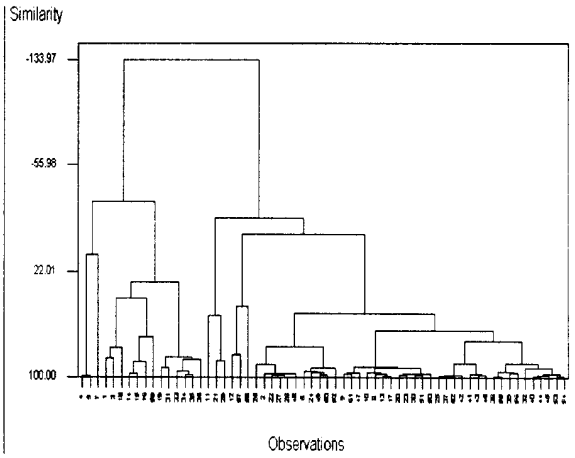
Obr. 20a Dendrogram objektů metodou nejbližšího souseda, Minitab



Obr. 20b Dendrogram objektů metodou nejvzdálenějšího souseda, Minitab



Obr. 20c Dendrogram objektů metodou průměrovou, Minitab



Obr. 20d Dendrogram objektů metodou Wardovou, Minitab

Poděkování:
Autoři vyslovují svůj dík za finační podporu vědeckého záměru č. MSM253100002.

Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.
[2] Kendall M. G., Stuart A.: *The Advanced Theory of Statistics*, Vol. III. New York 1966.
[3] James W., Stein C.: *Estimation with Quadratic Loss*, Proceed. 4th Berkeley Symp. on Math. Statist., p. 361, 1961.
[4] Guanadeskian R., Kettenring J. R.: *Biometrics* **28**, 80 (1972).
[5] Campbell N. A.: *Appl. Statist.*, 29, 231 (1980).
[6] Hu J., Skrabal P., Zollinger H.: *Dyes and Pigments*, **8**, 189 (1987).
[7] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods for Data Analysis*. Duxburg Press, Belmont, California 1983.
[8] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 6.
[9] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.
[10] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 12.
[11] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.
[12] Andrews D. F.: *Biometrics*, **28**, 125 (1972).
[13] Kulkarni S. R., Paranjape S. R.: *Commun. Statist.*, **13**, 2511 (1984).
[14] Guanadeskian R.: *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York 1977.
[15] Kleiner B., Hartigan J. A., *J. Amer. Statist. Assoc.*, **76**, 260 (1981).
[16] Kres H.: *Statistical Tables for Multivariate Analysis*. Springer, New York 1983.
[17] Seber G. A. F.: *Multivariate Observations*. Wiley, New York 1984.
[18] Stryjewska E., Rubel S., Henrion A., Henrion G.: *Z. Anal. Chem.*, **327**, 679 (1987).
[19] Mudholkar G. S., Trivedi M. S., Lin T. C.: *Technometrics*, **24**, 139 (1982).
[20] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis*, Prentice Hall, 1982

[21] Ajvjazin S., Bežajeva Z., Staroverov O.: *Metody vícerozměrné analýzy,* SNTL Praha 1981
[22] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis,* Ellis Horwood, Chichester 1992.
[23] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
[24] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,
[25] Jeffers J. N. R., *Applied Statistician*, **16**, 225 (1967).
[26] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994.
[27] Martens H., Naes T., *Multivariate calibration*, Wiley (1989) Chichester.
[28] Thomas E. V., *Anal. Chem.*, **66** (1994) 795A-804A.
[29] Malinowski F., Howery D., *Factor Analysis in Chemistry*, Wiley (1980) New York.
[30] Meloun M. , Militký J., *Sbírka úloh - Statistické zpracování experimentálních dat*, Univerzita Pardubice, 1996.