K čemu jsou techniky vícerozměrné statistické analýzy?

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

Souhrn: *Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných,* $y = w_1x_1 + \ldots + w_mx_m$. *Zdrojová matice dat obsahuje proměnné v m sloupcích a objekty v n řádcích. Analytická data jsou před zpracováním škálována. Cílem je nalézt shluk podobných objektů s podobnými proměnnými. Podobnost objektů posuzujeme na základě vzdálenosti 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, 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, hlavních komponent) 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 $y = w_1x_1 + w_2x_2 + \ldots + w_mx_m$. 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 FA* 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).*

Úloha 1. *Sledování spotřeby proteinů v Evropě*

Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem vyšetření. Ukazuje graf komponentních vah na silně korelující proměnné? Lze odhalit v rozptylovém diagramu komponentního skóre odlehlé objekty, výjimečné co do spotřeby proteinů? Které země jsou si podobné ve spotřebě proteinů?

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

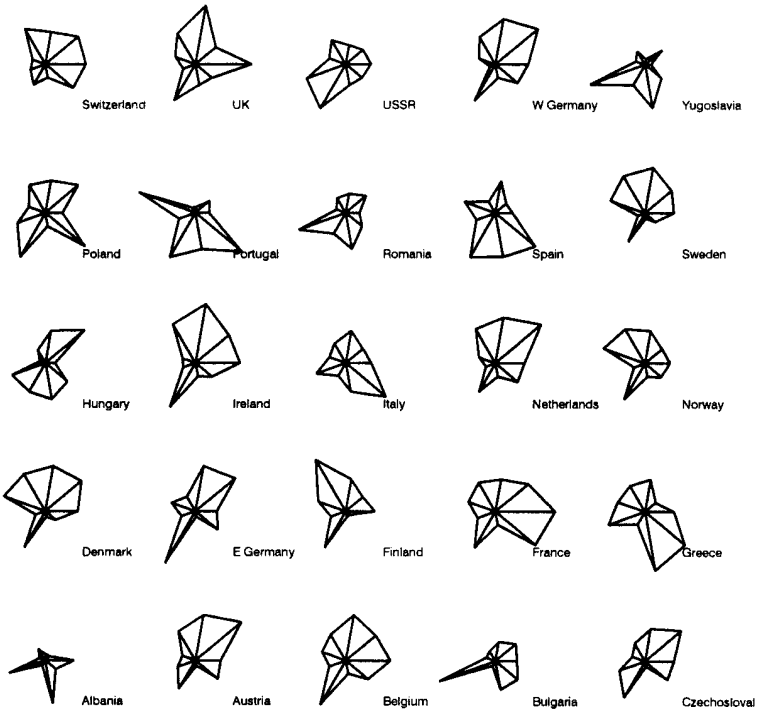
	Objekty	Proměnné								
i	Stát	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
1	Albania	10.10	1.40	0.50	8.90	0.20	42.30	0.60	5.50	1.70
2	Austria	8.90	14.00	4.30	19.90	2.10	28.00	3.60	1.30	4.30
3	Belgium	13.50	9.30	4.10	17.50	4.50	26.60	5.70	2.10	4.00
4	Bulgaria	7.80	6.00	1.60	8.30	1.20	56.70	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:

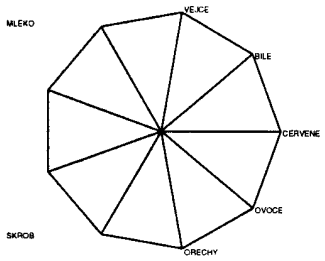
1. Exploratorní analýza: Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především *symbolové grafy.* Jednotlivé proměnné jsou v nich "kódovány" s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů, *symbolů.* Každému objektu x_i odpovídá jistý obrazec zvaný *symbol.* Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi symboly.

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.* Sestávají z paprsků, reprezentujících relativní hodnoty proměnných u jednotlivých objektů, které se pro každý objekt spojují v jednom centrálním bodě.



Obr. 1 Hvězdičkový graf

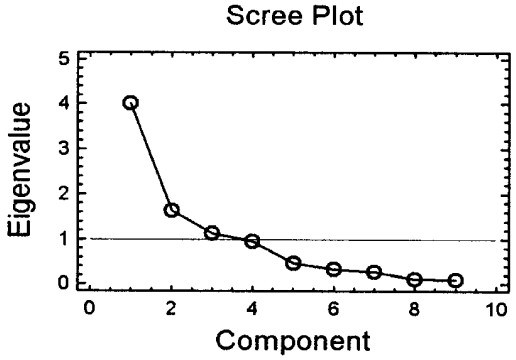


Obr. 2 Klíč

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:

a) Vyšetření indexového grafu úpatí vlastních čísel:

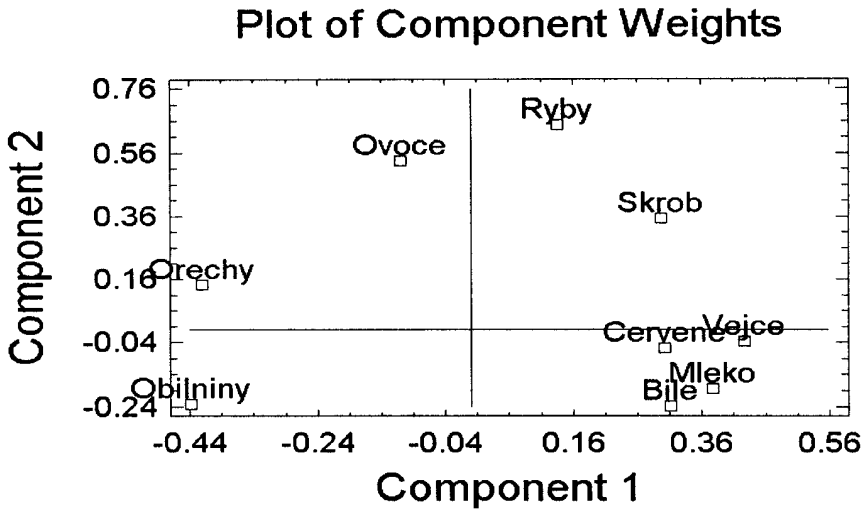


Obr. 3 Indexový graf úpatí vlastních čísel pro 25 objektů a 9 proměnných.

Vlastní čísla slouží k určení počtu "využitelných" hlavních komponent, jež si zvolíme v analýze k dalšímu užívání. Indexový graf úpatí vlastních čísel je vlastně sloupcový diagram velikosti vlastních čísel proti stoupající hodnotě indexu, pořadového čísla. Užitečné

komponenty jsou tak odděleny zřetelným zlomovým místem, a x-ová souřadnice tohoto zlomu je hledaná hodnota indexu.

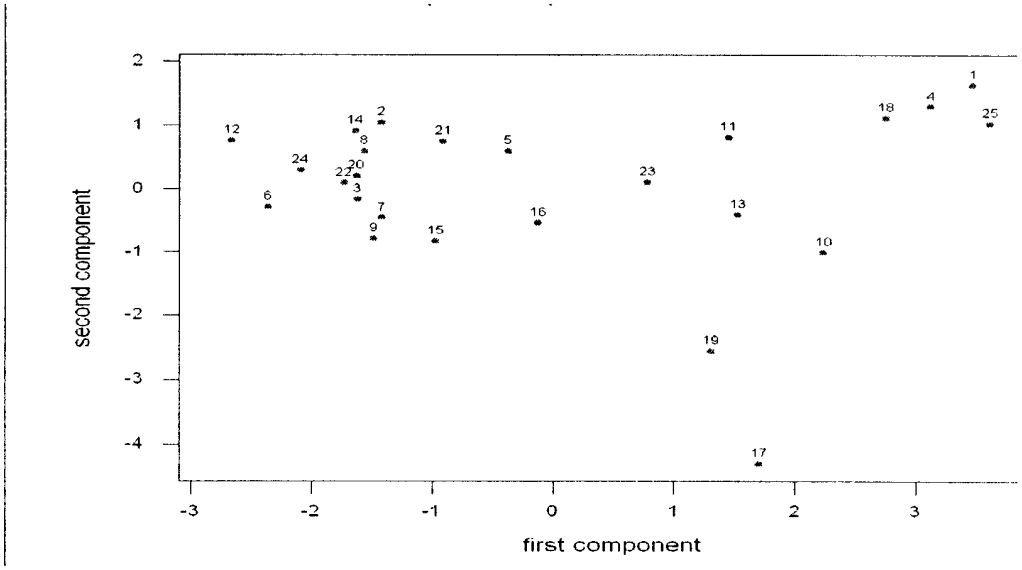
b) Vyšetření grafu komponentních vah: graf ukazuje, že proměnné *Bílé maso-Mléko-Červené maso-Vejce* spolu silně korelují.



Obr. 4 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.

c) 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. 5 Rozptylový diagram komponentního skóre (Scatterplot)

3. Klasifikace metodou analýzy 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 *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat.

Hierarchické postupy jsou založeny na postupném spojování objektů a jejich shluků do dalších, větších shluků. Nejprve se vypočte základní matice vzdáleností mezi objekty. 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.

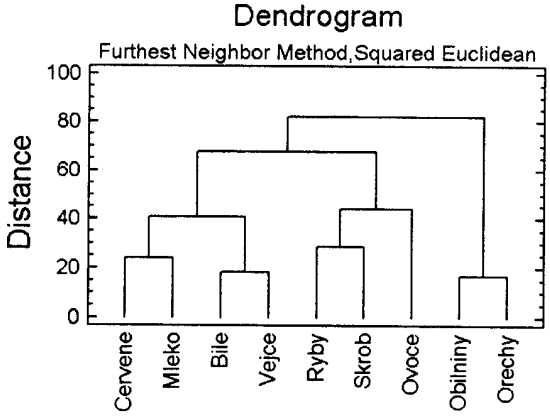
Metody metriky shlukování jsou

Metoda průměrová: 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.

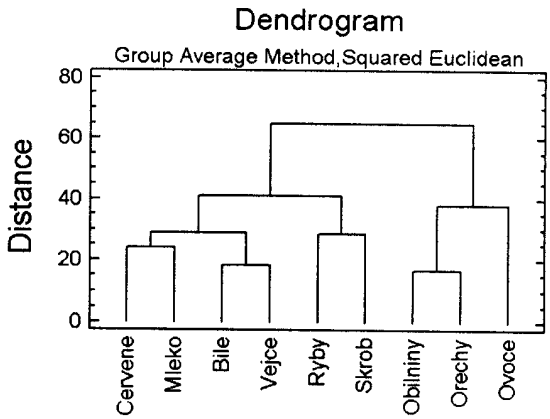
Metoda nejbližšího souseda: kritériem pro vytváření shluků je minimum z možných mezishlukových vzdáleností objektů. Metoda tvoří nový shluk na základě nejkratší vzdálenosti mezi shluky a neumí proto rozlišit špatně separované shluky.

Metoda nejvzdálenějšího souseda: počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů. Proto všechny objekty ve shluku jsou na základě maximální vzdálenosti či minimální podobnosti vůči objektům ve druhém shluku.

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 dosáhla při vzniku shluku svého minima.



Obr. 6 Dendrogram proměnných metodou nejvzdálenějšího souseda



Obr. 7 Dendrogram proměnných metodou průměrovou

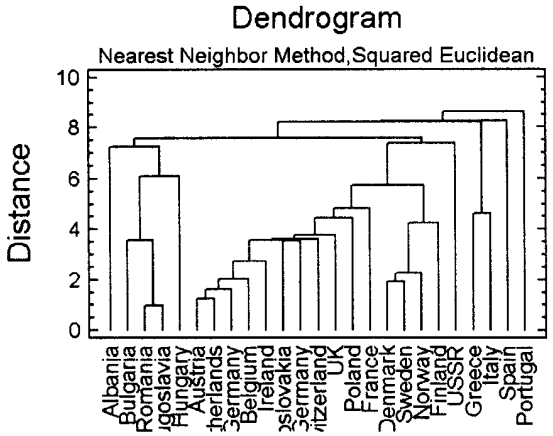
Analýzou shluků budeme sledovat a vyšetřovat podobnost objektů, analyzovanou pomocí *dendrogramu objektů*, a jednak podobnost proměnných analyzovanou pomocí *dendrogramu proměnných.*

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

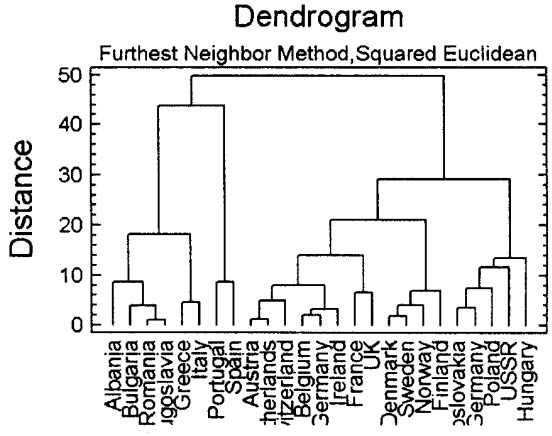
Dendrogram podobnosti proměnných odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) proměnných, které jsou si velmi podobné a silně spolu korelují. Některé vlastnosti (či proměnné) není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

Míra věrohodnosti: kritériem věrohodnosti nalezené struktury objektů a proměnných mezi objekty je *kofenetický korelační koeficient CC*. Je to druh korelačního koeficientu mezi skutečnou a dendrogramem predikovanou vzdáleností. Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření, distorze spíše než stupeň podobnosti. Hodnoty *delta* blízké nule jsou žádoucí. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

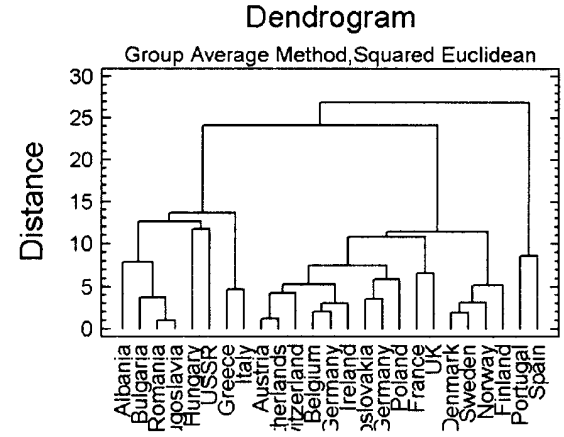
Dendrogram podobnosti proměnných obsahuje dvojice nebo trojice proměnných, které jsou si velmi podobné a silně spolu korelují. Metoda nejvzdálenějšího souseda ukazuje na 4 dvojice: první dvojice *Červené maso-Mléko*, druhá *Bílé maso-Vejce,* třetí *Ryby-Škrob,* čtvrtá *Obilniny-Ořechy.* Nejdůležitějším 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 podobnosti.



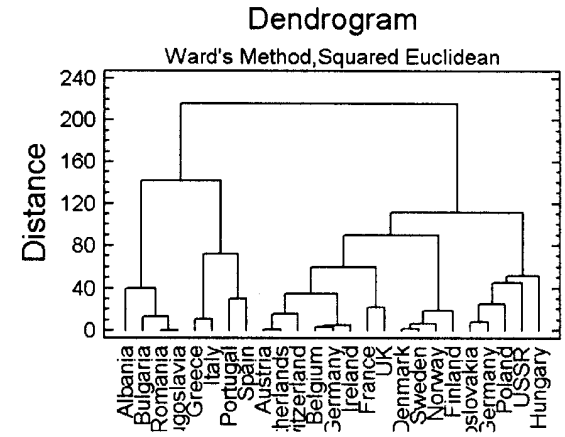
Obr. 8 Dendrogram objektů metodou nejbližšího souseda



Obr. 9 Dendrogram objektů metodou nejvzdálenějšího souseda



Obr. 10 Dendrogram objektů metodou průměrovou



Obr. 11 Dendrogram objektů metodou Wardovou

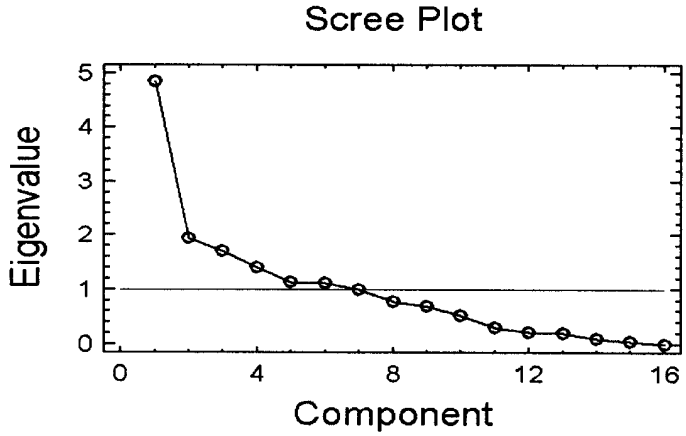
Úloha 2. *Klasifikace zdrojů pitné vody*

Na 62 vzorcích zdrojů pitné vody bylo stanoveno 16 proměnných kvality. Je třeba nalézt *vybočující objekty*, resp. jejich proměnné, zda ukazuje graf komponentních vah na korelující proměnné, zda lze odhalit v rozptylovém diagramu komponentního skóre odlehlé objekty, zda lze posoudit *podobnost objektů* shlukovou analýzou klasifikaci zdrojů.

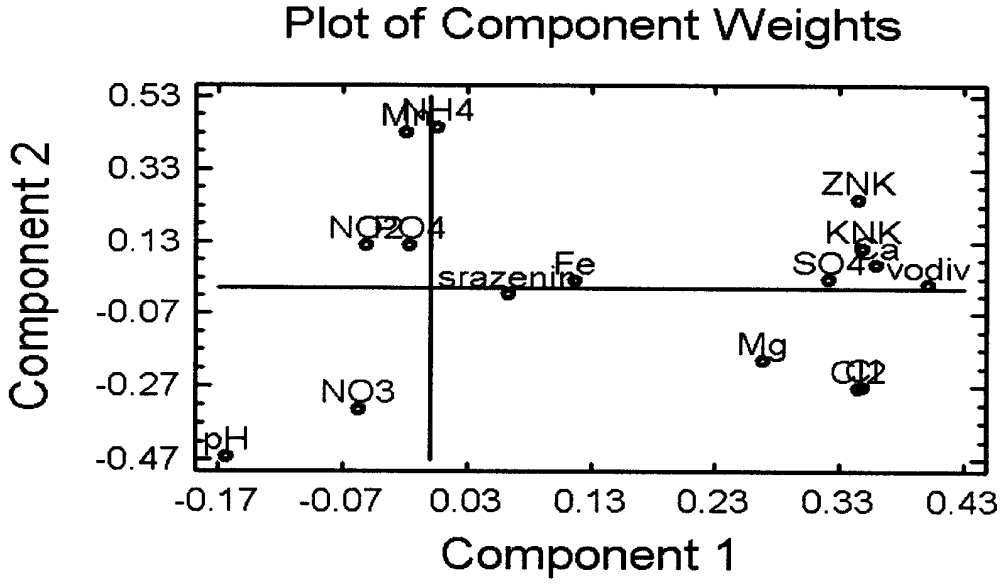
Data: i index vzorku, x_1 obsah dusičnanů [mg/l], x_2 obsah dusitanů [mg/l], x_3 obsah chloridů [mg/l], x_4 obsah celkového chloru [mg/l], x_5 obsah síranů [mg/l], x_6 obsah fosforečnanů [mg/l], x_7 obsah amonných solí [mg/l], x_8 obsah vápníku [mg/l], x_9 obsah hořčíku [mg/l], x_{10} obsah železa (celkového) [mg/l], x_{11} obsah manganu [mg/l], x_{12} pH, x_{13} KNK, x_{14} ZNK, x_{15} vodivost, x_{16} 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	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	0.1	0.02	102	12	0.016	0.05	7.69	2.6	0.65	436	0.05

Řešení: **Metoda hlavních komponent - Vyšetření indexového grafu úpatí vlastních čísel:**



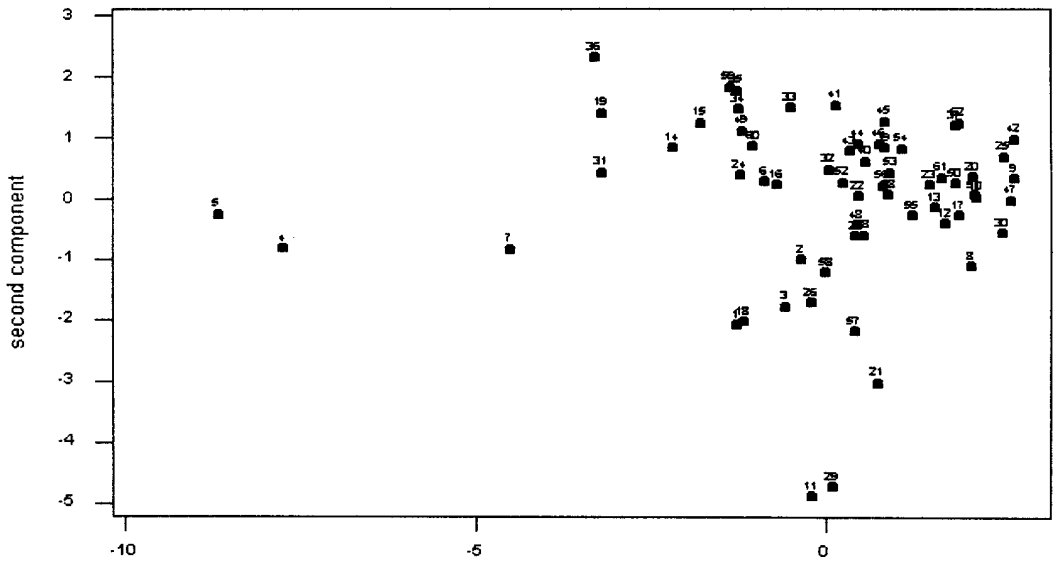
Obr. 12 Indexový graf úpatí vlastních čísel pro 62 objektů a 16 proměnných
Užitečné komponenty jsou odděleny zřetelným zlomovým místem a hledaný počet je 2.



Obr. 13 Graf komponentních vah (Components Weights Plot)

Vyšetření grafu komponentních vah: porovnáním vzdáleností mezi proměnnými a se dospěje k závěru, že krátká vzdálenost mezi dvěma proměnnými znamená silnou korelaci. Graf ukazuje, jakou měrou přispívají jednotlivé původní proměnné x_j, j= 1,..., 16, 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 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.

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. 14 Rozptylový diagram komponentního skóre (Scatterplot)

Objekty mohou být označeny textovým popisem nebo jako na obr. 14 číselně indexem:

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.

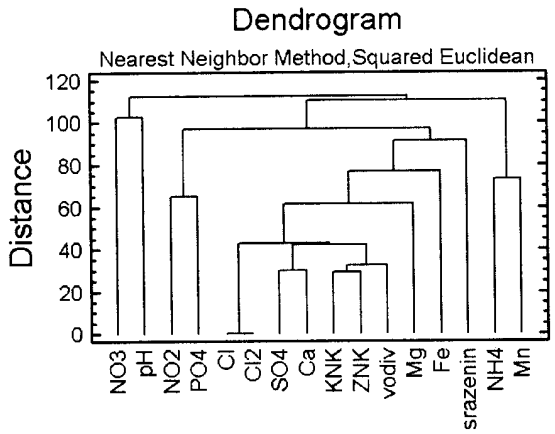
2. *Podobnost objektů*: objekty blízko sebe (např. 53 a 54 a 44 a 45) si jsou podobné, objekty daleko od sebe (např. 5 a 45, 4 a 30, atd) jsou si nepodobné.

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.

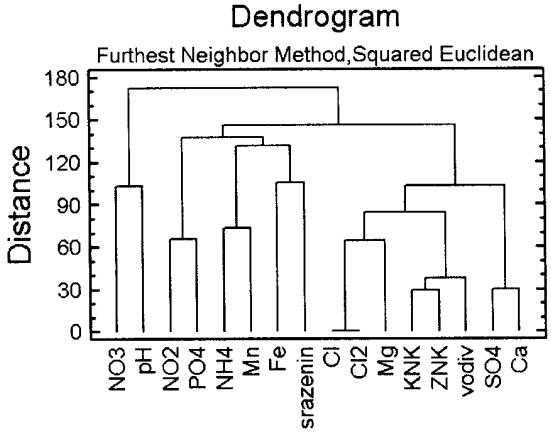
4. *Osamělé objekty:* izolované objekty (4, 5, 7) mohou být odlehlé objekty, které jsou silně nepodobné ostatním objektům.

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.

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.

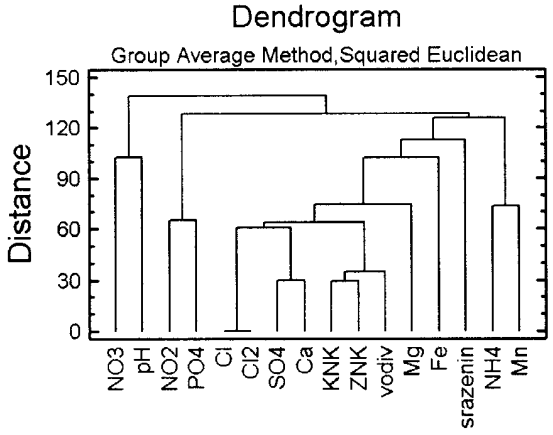


Obr. 15 Dendrogram proměnných metodou nejbližšího souseda

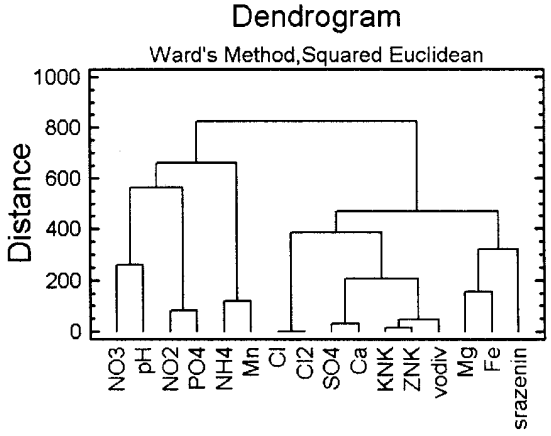


Obr. 16 Dendrogram proměnných metodou nejvzdálenějšího souseda

Metodou nejbližšího souseda lze nalézt šest dvojic velice si podobných proměnných. Nejvíce jsou si podobné proměnné ve dvojicích: $x_{13}(KNK)$-$x_{14}(ZNK)$ a $x_5(SO_4^{2-})$-$x_8(Ca)$. Poněkud ménějsou si podobné proměnné ve dvojici $x_2(NO_2^-)$-$x_6(PO_4^{3-})$ a také ve dvojici $x_7(NH_4^+)$-$x_{11}(Mn)$. Nejméně jsou si podobné proměnné ve dvojici $x_1(NO_3^-)$-$x_{12}(pH)$. Výjimečné postavení má proměnná x_{16}*(#sraženina)*, která si není podobná s žádnou jinou proměnnou.



Obr. 17 Dendrogram proměnných metodou průměrovou



Obr. 18 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.