Postup statistické analýzy vícerozměrných dat

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice,
milan.meloun@upce.cz
a
Prof. Ing. Jiří Militký, CSc.,
Katedra textilních materiálů, Technická univerzita Liberec, 461 17 Liberec,
jiri.militky@vslib.cz

Souhrn: *Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných,* $y = w_1x_1 + ... + w_mx_m$. *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 (vzdálenost Euklidovská, Manhattanská, Minkovského a Mahalanobisova), párového korelačního koeficientu a koeficientu asociace (Sokalův-Michelenerův, Russelův-Raoův a Hamanův): čím je vzdálenost shluků či objektů větší, tím menší je jejich podobnost. K rychlému posouzení podobnosti slouží grafy exploratorní analýzy vícerozměrných dat: profily, polygony, tváře, křivky, stromy, sluníčka a hvězdičky. Strukturu a vazby mezi proměnnými vystihují metody snížení dimensionality, metoda hlavních komponent (PCA) a metoda faktorové analýzy. 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.*

V technické praxi se vedle informací, obsažených v náhodném skaláru ξ, vyskytují i vícerozměrné informace, obsažené v náhodném vektoru ξ s m složkami $\xi_1, ..., \xi_m$. Příklady vícerozměrných informací jsou

a) vyjádření vlastností produktů jako jsou potraviny, oleje, slitiny, atd. pomocí řady různých analytických metod,

b) hodnocení spekter pomocí poloh a ploch absorpčních pásů sloužící k charakterizaci a identifikaci chemických sloučenin,

c) sledování složení surovin, produktů, odpadů, v závislosti na čase nebo na místě výskytu,

d) regulace jakosti na základě různých procesních proměnných,

e) stanovení charakteristiky produktu na základě měření souvisejících proměnných, např. spekter (vícerozměrná kalibrace).

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 $x_1, x_2, ..., x_m$ ve tvaru $y = w_1x_1 + w_2x_2 + ... + w_mx_m$. Jednotlivé vícerozměrné metody využívají různých způsobů stanovení vah $w_1, w_2, ..., w_m$.

Zdrojová matice, tj. matice výchozích dat (popisující např. řadu aut) obsahuje **proměnné** v m sloupcích (např. obsah motoru, výkon, spotřeba paliva, hmotnost vozu, zrychlení, výška, šířka, délka, atd.) a **objekty** v n řádcích (např. auta různých výrobců), na nichž jsou tyto

proměnné (vlastnosti) měřeny. Protože měřené proměnné mají různé jednotky, a často se řádově liší, bývá zdrojová matice před zpracováním ještě upravována, *škálována*, a to buď (a) *centrováním*, kdy se od prvků sloupce odečte jejich sloupcový aritmetický průměr, nebo (b) *standardizací* čili *normováním*, kdy se prvky centrovaných sloupců vydělí svou sloupcovou směrodatnou odchylkou. Hodnoty x_{ij} tvoří *náhodný výběr*, který je tvořen n-ticí řádkových vektorů $x_i^T = (x_{i,1}, ..., x_{i,m})$ a lze proto vyjádřit maticí rozměru $(n \times m)$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_i^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,m} \\ \vdots & & \vdots & & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,m} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,m} \end{bmatrix}$$

Řádek zdrojové matice čili i-tý řádkový vektor $x_i^T = (x_{i,1}, ..., x_{i,m})$ nazýváme *objektem* (např. auto určitého typu a modelu) a můžeme ho chápat jako jeden jediný bod v m-rozměrném prostoru. Tento objekt je charakterizován svými *proměnnými*, a to buď *kvantitativními*, metrickými čili číselnými hodnotami nebo proměnnými *kvalitativními* čili nemetrickými. **Metrické proměnné** se vyskytují ve čtyřech škálách:

(a) *Proměnné v absolutní škále* mají na škále přirozený počátek a jediné měřítko, např. obsah uhlíku v %, rychlostní konstanta.

(b) *Proměnné v poměrové škále* mají zachován podíl hodnot charakteristik $c = x_2/x_1$, např. vztah vůči standardní sloučenině, vztah vůči jevu s definovaným nulovým počátkem, parametr σ v Hammettově rovnici.

(c) *Proměnné v intervalové škále* mají zachován podíl rozdílů $c = x_2 - x_1$. Jedná se o poměrovou škálu s přirozeným počátkem pro obě srovnávané hodnoty, např. Poměr absorbancí indikátoru, vztažený na absorbanci nulové linie.

(d) *Proměnné v rozdílové škále* jsou vztahovány k různému počátku, např. hodnoty časových škál, stáří, atd.

Nemetrické proměnné se vyskytují ve dvou škálách:

(a) *Proměnné v ordinální škále* mají svou hodnotu danou pořadím v neklesající posloupnosti proměnných dle nějakého kritéria, např. počet atomů chloru v molekule, žebříček umístění, pořadové číslo.

(b) *Proměnné v nominální škále* jsou nejméně informativní. Obsahují kód, např. barvu kódem 1 až 16, rodinný stav (svobodný 1, ženatý 2, rozvedený 3, vdovec 4).

(c) *Proměnné v alternativní (binární) škále* vyjadřují rovnost či nerovnost vůči nějakému kritériu. Mají binární charakter a relaci můžeme popsat dvojicí 1 (ano), 0 (ne).

Shluk ve vícerozměrné analýze chápeme jako množinu objektů se společnými nebo alespoň blízkými proměnnými, znaky (např. všechna auta typu BMW). Blízkost či podobnost objektů posuzujeme na základě *míry blízkosti* či *vzdálenosti objektů* v m-rozměrném prostoru proměnných. Vyjádření vzdálenosti objektů pro *kvantitativní* proměnné jsou *Euklidova metrika* čili *geometrická vzdálenost* představuje nejjednodušší typ vzdálenosti a definovaný vztahem

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

Hammingova metrika čili *Manhattanská vzdálenost* je definovaná vztahem

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

Zobecněná Minkowskiho metrika je definovaná vztahem

$$d_M(x_k, x_l) = \sqrt[n]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^n},$$

kde pro $n = 1$ jde o Hammingovu metriku a pro $n = 2$ o Euklidovu. Čím je n větší, tím více je zdůrazňována vzdálenost mezi blízkými objekty. Všechny tyto metriky předpokládají nezávislost mezi proměnnými. Zahrneme-li však do vztahu pro vzdálenost i vnitřní vazby mezi proměnnými, vyjádřené kovarianční maticí C dostaneme novou míru, zvanou *Mahalanobisova metrika* nebo *Mahalanobisova vzdálenost*

$$d_{MA}(x_k, x_l) = \sqrt{(x_k - x_l)^T C^{-1} (x_k - x_l)}.$$

Ta se společně s Euklidovou metrikou nejvíce používá v praxi. Ve všech uvedených případech jsou si dva objekty tím bližší, čím je jejich vzdálenost menší.

Mírou podobnosti dvou objektů nebo proměnných x_i a x_j může být *párový korelační koeficient r*. Objekty jsou si tím podobnější, čím je párový korelační koeficient větší a bližší 1. V případě ordinální škály je analogickou mírou podobností *Spearmanův korelační koeficient.* Podobnost binárních nebo nominálních proměnných vyjadřují různé koeficienty asociace. Označíme-li počet případů negativní shody typu 0-0 písmenem a, počet případů s neshodou typu 1-0 písmenem b, počet případů s neshodou typu 0-1 písmenem c a počet případů s pozitivní shodou typu 1-1 písmenem d, dojdeme ke následujícím vzorcům různých koeficientů podobnosti:

(a) *Sokalův-Michelenerův koeficient asociace* je definován vzorcem

$$S_{SM} = \frac{a + d}{a + b + c + d},$$

(b) *Russelův-Raoův koeficient asociace* je definován vzorcem

$$S_{RR} = \frac{d}{a + b + c + d},$$

(c) *Hamannův koeficient asociace* je definován vzorcem

$$S_H = \frac{a + d - b - c}{a + b + c + d},$$

a také lze konstruovat *obdobu korelačního koeficientu* vztahem

$$r_B = \frac{a\,d - b\,c}{\sqrt{(a + b)\,(c + d)\,(a + c)\,(b + d)}}.$$

Míra podobnosti mezi objekty, charakterizovanými různými typy proměnných se vypočte jako vážený průměr jednotlivých měr podobnosti. Na základě měr podobnosti objektů se konstruují míry podobnosti mezi objekty a shluky a míry podobnosti mezi shluky. Jako nejčastější míra podobnosti se používá vzdálenost shluků $d(x_k, x_l)$. Analogicky zde užijeme způsobů vyjádření vzdálenosti objektů, protože objekt můžeme chápat jako shluk o jednom objektu. Čím větší je vzdálenost, tím menší je podobnost:

(a) *Vzdálenost nejbližšího souseda:* nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi dvěma nejbližšími objekty dvou pozorovaných shluků.

(b) *Vzdálenost nejvzdálenějšího souseda:* nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi dvěma nejvzdálenějšími objekty.

(c) *Vzdálenost mezi těžišti tříd:* nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi svými těžišti.

(d) *Vzdálenost průměrné vazby:* nejbližší jsou ty shluky, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jedné a všemi objekty druhého shluku.

1. Postup analýzy vícerozměrných dat

Postup analýzy vícerozměrných dat záleží na typu dat a na druhu požadované informace jež se z dat má získat. Jednotlivé techniky pro *stanovení závislosti* se dále dělí podle počtu závisle proměnných a podle typu či škály měření. Klasická vícerozměrná regrese je případem jedné závisle proměnné v metrické škále.

Schematicky lze vztahy mezi jednotlivými technikami analýzy vícerozměrné závislosti zapsat ve formě těchto přiřazení:

(a) **Kanonická korelace:**

$$\underset{\text{(metrická, nemetrická)}}{y_1 + y_2 + \ldots + y_m} \Leftarrow \underset{\text{(metrická, nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(b) **Vícerozměrná analýza rozptylu:**

$$\underset{\text{(metrická)}}{y_1 + y_2 + \ldots + y_m} \Leftarrow \underset{\text{(nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(c) **Analýza rozptylu:** $$\underset{\text{(metrická)}}{y_1} \Leftarrow \underset{\text{(nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(d) **Diskriminační analýza:**

$$\underset{\text{(nemetrická)}}{y_1} \Leftarrow \underset{\text{(metrická)}}{x_1 + x_2 + \ldots + x_m}$$

(e) **Vícerozměrná regrese a kalibrace:**

$$\underset{\text{(metrická)}}{y_1} \Leftarrow \underset{\text{(metrická, nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(f) **"Conjoint" analýza:**

$$\underset{\text{(metrická, nemetrická)}}{y_1} \Leftarrow \underset{\text{(nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(g) **Strukturní rovnice:**

$$\begin{aligned} y_1 &\Leftarrow x_{11} + x_{12} + \ldots + x_{1m} \\ y_2 &\Leftarrow x_{21} + x_{22} + \ldots + x_{2m} \\ &\ldots\ldots \\ \underset{\text{(metrická)}}{y_n} &\Leftarrow \underset{\text{(metrická, nemetrická)}}{x_{n1} + x_{n2} + \ldots + x_{nm}} \end{aligned}$$

Zdrojová matice má vždy rozměr $n \times m$. Data v nemetrické škále lze kódovat s využitím i umělých (*dummy*) proměnných, nabývajících např. hodnot 1 (přítomnost nominálního znaku) nebo 0 (nepřítomnost nominálního znaku). To umožňuje "rozšíření" faktorové a shlukové analýzy o data v nemetrické škále. Před vlastní vícerozměrnou statistickou analýzou je třeba provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje

(a) posoudit *podobnost objektů* pomocí rozptylových diagramů 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 pro stanovení vzájemných vazeb se dále dělí podle toho, zda se hledají struktury v proměnných nebo v objektech:

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

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

(3) Hledání struktury *v objektech* v obou škálách: *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.

2. Charakteristiky vícerozměrných náhodných veličin

K charakterizaci polohy j-té proměnné ξ_j čili j-tého sloupce matice X se používá **střední hodnota** $E(\xi_j) = \mu_j$ a pro charakterizaci rozptýlení pak **rozptyl** $D(\xi_j) = \sigma_j^2$. Dále je třeba definovat *míru intenzity* vztahu mezi dvěma proměnnými ξ_i a ξ_j, Vhodnou charakteristikou je *druhý smíšený centrální moment*, nazývaný **kovariance** $cov(\xi_i, \xi_j)$, definovaný vztahem

$$cov(\xi_i, \xi_j) = E(\xi_i \xi_j) - E(\xi_i) E(\xi_j)$$

Kovariance má vlastnosti:

a) Její znaménko ukazuje na typ stochastické vazby mezi j-tým a i-tým sloupcem matice.

b) Je v absolutní hodnotě shora ohraničená součinem $\sigma_i \sigma_j$, tj. $|cov(\xi_i, \xi_j)| \leq \sigma_i \sigma_j$.

c) Je symetrickou funkcí svých argumentů.

d) Nemění se posunem počátku, ale změna měřítka se projeví úměrně jeho velikosti. Pro čísla a_1, a_2, b_1, b_2 pak platí, že

$$cov(a_1 \xi_i + b_1, a_2 \xi_j + b_2) = a_1 a_2 cov(\xi_i, \xi_j) .$$

e) Pro nekorelované náhodné veličiny je $cov(\xi_i, \xi_j) = 0$ a mohou nastat dva případy:

1. $E(\xi_i \xi_j) = 0$ a zároveň $E(\xi_i) = E(\xi_j) = 0$, což je případ *centrovaných ortogonálních* náhodných veličin, ne nutně nezávislých.
2. $E(\xi_i \xi_j) = E(\xi_i) E(\xi_j)$, což je případ *nezávislých* náhodných veličin.

f) Je *mírou intenzity lineární závislosti.*

Nevýhodou kovariance je fakt, že její hodnoty závisí na měřítku, ve kterém jsou vyjádřeny proměnné ξ_i a ξ_j. Její velikost lze hodnotit vzhledem k součinu $\sigma_i \sigma_j$. Je proto přirozené provést standardizaci podělením tímto součinem. Vzniklá veličina $\rho_{ij} = \rho(\xi_i, \xi_j)$ se nazývá *párový korelační koeficient*

$$\rho(\xi_i, \xi_j) = \rho_{ij} = \frac{cov(\xi_i, \xi_j)}{\sigma_i \sigma_j}$$

Je zřejmé, že párový korelační koeficient leží v rozmezí $-1 \leq \rho_{ij} \leq 1$. Pokud je $\rho_{ij} > 0$, jde o *pozitivně korelované* náhodné veličiny, a pokud je $\rho_{ij} < 0$, jde o *negativně korelované* náhodné veličiny. Párový korelační koeficient má vlastnosti:

a) Rovnost $|\rho_{ij}| = 1$ ukazuje, že mezi ξ_i a ξ_j existuje přesně lineární vztah.

b) Pokud jsou náhodné veličiny ξ_i a ξ_j vzájemně nekorelované, je $\rho_{ij} = 0$.

c) V případě, že ξ_i a ξ_j pocházejí z vícerozměrného normálního rozdělení a $\rho_{ij} = 0$, znamená to, že jsou *vzájemně nezávislé.*

d) Platí, že i pro nelineárně závislé náhodné veličiny může být $\rho_{ij} = 0$.

e) Korelační koeficient ρ_{ii} náhodné veličiny ξ_i samotné se sebou je roven jedné.

f) Korelační koeficient je invariantní vůči lineární transformaci náhodných proměnných ξ_i, ξ_j. Pro čísla a_1, a_2, b_1, b_2 platí vztah

$$\rho(a_1 \xi_i + b_1, a_2 \xi_j + b_2) = sign(a_1 a_2)\, \rho(\xi_i, \xi_j)$$

kde *sign(x)* je znaménková funkce, pro kterou platí

$$sign(x) = \begin{cases} -1 & \text{pro } x < 0 \\ 0 & \text{pro } x = 0 \\ 1 & \text{pro } x > 0 \end{cases} .$$

Standardizace čili *normování* znamená škálování proměnné, spočívající v jejím převedení na náhodnou veličinu s jednotkovým rozptylem a nulovou střední hodnotou,

$$\xi_1^* = \frac{\xi_1 - E(\xi_1)}{\sqrt{D(\xi_1)}} \quad \text{a} \quad \xi_2^* = \frac{\xi_2 - E(\xi_2)}{\sqrt{D(\xi_2)}} .$$

3. Exploratorní analýza podobnosti objektů (EDA)

Průzkumová analýza vícerozměrných dat je stejně jako u jednorozměrných dat založena na grafických diagnostikách. Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především *symbolové grafy*. Každému objektu x_i (např. autu) 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*.

4. Určení struktury a vazeb v proměnných

Určením struktury a vzájemných vazeb mezi proměnnými se zabývají techniky redukce proměnných na latentní proměnné, metodou *analýzy hlavních komponent (PCA)* a metodou *faktorové analýzy (FA)*.

Metoda analýzy hlavních komponent (PCA): Principem metody je náhrada původních proměnných x_i tzv. *latentními proměnnými* y_i, které mají vhodnější vlastnosti, totiž je jich výrazně menší počet, i když vystihují téměř celou *proměnlivost* původních proměnných a jsou vzájemně nekorelované (korelační koeficient mezi latentními proměnnými $y_1, ..., y_m$ je 0). Latentní proměnné se nazývají hlavní komponenty a jsou lineárními kombinacemi původních proměnných. *První hlavní komponenta*, tj. y_1 popisuje největší část proměnlivosti čili rozptylu původních dat, *druhá hlavní komponenta*, tj. y_2 zase největší část rozptylu neobsaženého v y_1, atd. Zcela analogicky jsou konstruovány další hlavní komponenty, jejichž celkový počet roven menšímu z čísel n (počet objektů) a m (počet proměnných). Protože platí, že součet rozptylů všech hlavních komponent je roven součtu rozptylů vstupujících proměnných, můžeme z podílu rozptylů jednotlivých hlavních komponent usuzovat na část proměnlivosti, vysvětlenou dotyčnou hlavní komponentou. Jestliže součet prvních (nejvyšších) k podílů proměnlivosti je dostatečně blízký jedné (obvykle však stačí 0.9 - 0.95), postačí brát v úvahu právě těchto prvních k hlavních komponent pro "dostatečné" vysvětlení původních proměnných. I při velkém počtu původních proměnných (m) může být počet k velmi malý, často 2 až 5.

(a) **Indexový graf úpatí vlastních čísel** (Scree Plot) je vlastně sloupcový diagram vlastních čísel. Vlastní čísla slouží k určení počtu "využitelných" hlavních komponent, jež si zvolíme v analýze k dalšímu užívání. Procento a kumulativní procento popisuje proměnlivost v

původních proměnných, popsanou dotyčnou hlavní komponentou. Bereme obvykle k dalšímu popisu proměnlivosti tolik hlavních komponent, aby bylo jimi popsáno 90 až 99% celkové proměnlivosti. V tomto případě stačí užít první dvě. Scree Plot zobrazuje relativní velikost jednotlivých vlastních čísel. Vybrané "užitečné" hlavní komponenty (nebo také faktory) pak tvoří kolmou stěnu a "neužitečné" hlavní komponenty (nebo faktory) představují vodorovné dno. Užitečné komponenty jsou tak v tomto grafu odděleny zlomovým místem.

(b) **Graf komponentních vah** (Plot Components Weights) zobrazí komponentní váhy pro první dvě hlavní komponenty. 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.

(c) **Rozptylový diagram** (Scatterplot) zobrazuje *komponentní skóre*, tj. hodnoty dvou hlavních komponent u jednotlivých objektů navzájem. Dokonalé rozptýlení objektů v rovině obvykle 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.

(d) **Dvojný graf** (Biplot) kombinuje předchozí dva grafy. Úhel mezi průvodiči dvou proměnných je nepřímo úměrný velikosti korelace mezi těmito 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.

5. Určení struktury a vzájemných vazeb v objektech

Hledáním struktury a vzájemných vazeb v objektech se zabývají především klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do jedné existující třídy (*diskriminační analýza*) nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých shluků (*analýza shluků*). Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných objektů, např. normální rozdělení náhodného vektoru charakterizujícího objekty, a pak hovoříme o *parametrických klasifikačních metodách.* Není-li klasifikace založena na znalostech rozdělení náhodného vektoru mluvíme o *neparametrických klasifikačních metodách.*

Analýza shluků (Cluster analysis) 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. 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 se vyskytují dva základní problémy: (a) *Způsob měření vzdáleností mezi objekty.* Existuje celá řada měr vzdáleností čili vícerozměrných metrik nejčastěji se však užívá *euklidovská metrika*, která je přirozeným zobecnením běžného pojmu vzdálenosti; (b) *Volba vhodné shlukovací metody* dle zvoleného způsobu metriky, které mohou být:

(1) *Metoda průměrová* (Average): vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy se mezishlukovou vzdáleností objektů 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.

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

(3) *Metoda nejbližšího souseda* (Single): 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. Jsou-li objekty na opačných koncích řetězce zcela nepodobné, řetězování může vést až ke zcela mylným závěrům. Na druhé straně je to jedna z mála metod, která umí roztřídit a rozlišit i neeliptické shluky.

(4) *Metoda nejvzdálenějšího souseda* (Complete): 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 je z jiného shluku. Proto všechny objekty ve shluku jsou klasifikovány na základě maximální vzdálenosti či minimální podobnosti vůči objektům ve druhém shluku.

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

(6) *Wardova metoda* je založena na minimalizaci ztráty informace při spojení dvou shluků. 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 rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů. Místo výchozí matice vzdáleností může být v některých případech ke shlukování použita i *korelační matice.*

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 *shluků.* Analýzou shluků budeme sledovat a vyšetřovat jednak 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 při zadání maticí vzdáleností. Výsledkem je zobrazení hodnot ve dvojrozměrném prostoru, kde osy tvoří zadané proměnné. Objeví se také "obkroužení" objektů v jednotlivých shlucích.

(a) *Dendrogram podobnosti objektů* je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích. Je obvykle cílem naší analýzy.

(b) *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 shluků v dendrogramu: Dendrogram lze sestrojit celou řadou technik. Prvním kritériem věrohodnosti čili těsnosti proložení při volbě "nejlepšího dendrogramu",

jež nejlépe odpovídá struktuře objektů a proměnných mezi objekty, je *kofenetický korelační koeficient CC*. Je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu. Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření, distorze spíše než stupeň podobnosti. Kritérium delta je definováno 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. Jsou žádoucí hodnoty *delta* blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

Postup shlukové analýzy: optimální postup je tvořen následujícími kroky:

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.
7. *Výpočet skutečných a predikovaných vzdáleností v dendrogramu:* jsou porovnány skutečné vzdálenosti mezi objekty a vypočtené vzdálenosti (predikované) v dendrogramu, jejich rozdíl a konečně i procentuální vyjádření tohoto rozdílu.
8. *Hledání nejlepší techniky tvorby dendrogramu:* dle bodu 4. a 5. lze k sestrojení optimálního dendrogramu kombinovat řadu technik. Rozhodčím kritériem věrohodnosti jsou především kofenetický korelační koeficient *CC*, obě míry těsnosti proložení *delta*.
9. *Vysvětlení nejlepšího dendrogramu podobností objektů:* interpretace optimálního dendrogramu podobnosti jednotlivých objektů je prvním a nejdůležitějším cílem shlukové analýzy.
10. *Vysvětlení nejlepšího dendrogramu podobností proměnných:* interpretace optimálního dendrogramu podobnosti jednotlivých proměnných odhalí souvislosti ve struktuře objektů analyzované databáze a je druhým důležitým cílem shlukové analýzy.

Vícerozměrné škálování (MultiDimensional Scaling, MDS) je technika vytvoření diagramu relativního umístění objektů v rovině dvojrozměrného grafu na základě dat vzdáleností mezi objekty, tzv. *matice proximity* (blízkosti). Diagram může obsahovat jeden, dva, tři a zřídka i

více rozměrů, dimenzí. Technika vyčíslí metrické klasické (CMDS) nebo nemetrické (NNMDS) řešení a vychází buď přímo z experimentálních hodnot X, z korelační matice R nebo z matice podobností S či vzdáleností D. Vzdálenost mezi oběma objekty je Eukleidovská, počítaná na základě Pythagorovy věty, $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$, kde m je počet proměnných a x_{ik} jsou data i-tého řádku a k-tého sloupce. I když vynášíme vzdálenosti do dvojrozměrného grafu, může být d_{ij} vyčísleno na základě většího počtu proměnných $m \geq 2$. Matice vzdáleností je potom trojúhelníková a zajímá nás jenom její horní část. S růstem objektů však roste i počet dimenzí, takže pro *tři* objekty je to *dvoj*-rozměrná rovina, pro *čtyři* objekty pak *troj*-rozměrný prostor atd.

Kritérium maximální věrohodnosti MDS: Jak těsně prokládá MDS model vzdáleností daná experimentální data se hodnotí *testem těsnosti proložení* s využitím statistického kritéria *stress*, založeného na rozdílu mezi skutečnou vzdáleností d_{ij} a modelem predikovanou hodnotou $\hat{d}_{ij}$,

$$stress = \sqrt{\frac{\sum_{k=1}^{m} (d_{ij} - \hat{d}_{ij})^2}{\sum_{k=1}^{m} d_{ij}^2}},$$

kde $\hat{d}_{ij}$ je predikovaná vydálenost, založená na MDS modelu. Predikovaná hodnota závisí především na počtu užitých dimenzí a algoritmu, a to metrickém či nemetrickém. Je-li *stress* číslo nízké, blízké nule, jeví se MDS proložení jako nejlepší.

Počet dimenzí v MDS: Důležitým úkolem v MDS je určení počtu dimenzí v MDS modelu. Každá dimenze zde představuje latentní proměnnou. Cílem MDS je udržet počet dimenzí na co možná nejmenší hodnotě. Obvykle volí uživatel dvoj- maximálně trojrozměrný prostor. Vychází-li vyšší počet dimenzí, není MDS technika k analýze dotyčných dat vhodná. Počet dimenzí se volí na základě co nejmenší hodnoty kritéria *stress*. Někteří autoři si pomáhají indexovým grafem relativní velikosti vlastních čísel, která jsou vyčíslována pro rostoucí počet dimenzí, tzv. *grafem úpatí*. Postup a interpretace jsou pak stejné jako u metod PCA nebo FA.

Vstupní data v MDS: Data mohou být trojího typu, mohou obsahovat (1) vzdálenosti mezi objekty D, (2) podobnost mezi objekty S nebo (3) hodnoty proměnných (sloupce) pro jednotlivé objekty (řádky) X. (a) *Vzdálenost (disimilarita)* d_{ij} představující vzdálenost mezi objekty, může být měřena přímo, jako např. vzdálenost dvou měst. MDS užívá vzdálenost v datech přímo a matice vzdáleností D je symetrická. (b) *Podobnost (similarita)* s_{ij} vyjadřuje, jak blízko se nacházejí dva objekty. MDS umožňuje načíst míry podobnosti pro každý pár objektů. Matice podobností S je opět symetrická. Podobnost lze konvertovat do veličiny vzdálenosti vzorcem $d_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$, kde d_{ij} představuje vzdálenost a s_{ij} podobnost. (c) *Hodnoty* x_{ij} proměnných pro jednotlivé objekty představují spíše standardní míry. Z nich se vypočte nejprve korelační matice R a potom matice Eukleidovských či Mahalanobisových vzdáleností D.

Klasická metrická metoda MDS: Je dána matice vzdáleností D, která vystihuje meziobjektové vzdálenosti objektů X v prostoru spíše nižšího rozměru dle vzorce

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

Jednotlivé kroky klasické MDS jsou následující:

1. Z D se vypočte $A = \{-0.5\ d_{ij}^2\}$.
2. Z A se vypočte $B = \{a_{ij} - a_{i.} - a_{.j} + a_{..}\}$, kde $a_{i.}$ je průměr všech a_{ij} přes j.
3. Nalezne se m největších vlastních čísel $\lambda_1 > \lambda_2 > \ldots > \lambda_m$ matice B a odpovídající vlastní vektory $L = L_{(1)}, L_{(2)}, \ldots, L_{(m)}$, které jsou normovány, takže $L^T_{(i)}\ L_{(i)} = \lambda_i$. Předpokládáme, že m je voleno tak, že vlastní hodnoty jsou relativně velké a kladné.
4. Souřadnicemi objektů jsou řádky matice L.

Klasické řešení je optimalizováno metodou nejmenších čtverců: přímé řešení L minimalizuje sumu čtverců vzdáleností mezi skutečnými prvky matice D, tj. d_{ij} a predikcemi $\hat{d}_{ij}$, založenými na L. Předpokládejme, že experimentální hodnoty vzdálenosti d_{ij} jsou zatíženy náhodnou chybou ε_{ij} dle vzorce $d_{ij} = \delta_{ij} + \varepsilon_{ij}$, kde ε_{ij} představuje kombinaci náhodných chyb z měření, distorze vzdáleností, když MDS model zcela neodpovídá konfiguraci navržených m vzdáleností. Navrhněme model závislosti mezi vzdáleností dvou objektů vztahem $\hat{d}_{ij} = \beta_0 + \beta_1\ \delta_{ij} + \varepsilon_{ij}$ a potom nalezením nejlepších odhadů b_0 pro β_0 a b_1 pro β_1 obdržíme odhad vypočtené vzdálenosti $\hat{d}_{ij} = b_0 + b_1\ \delta_{ij}$. Optimalizační procedura vychází z účelové funkce

$$U = \sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2 \approx \min.$$

Aby byla zajištěna úplná invariantnost vůči transformaci proměnných, užívá se modifikovaná účelová funkce U_{mod} dle vztahu

$$U_{\text{mod}} = \frac{\sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}^{n} d_{ij}^2}$$

a především její druhá odmocnina, zvaná *stress* $= \sqrt{U_{\text{mod}}}$. Je proto výhodné hledat optimální počet dimenzí, která se vezmou k vyčíslení predikce MDS vzdálenosti $\hat{d}_{ij}$ pomocí minimální hodnoty veličiny *stress*. Pro *stress* < 0.05 je těsnost proložení ještě přijatelná a pro *stress* $<$ 0.01 je těsnost proložení výtečná.

Nemetrická MDS: V dosavadnín postupu se předpokládalo, že vzdálenosti jsou vyčísleny metricky. Jsou však situace, kdy jedna hodnota nevystihuje dostatečně skutečnost: např. při porovnávání barev na stupnici může být jedna barva zářivější než druhá, a tento fakt však nikterak neovlivní polohu barvy na stupnici. Predikované vzdálenosti $\hat{d}_{ij}$ jsou vyčíslovány *monotónní regresí:* experimentální vzdálenosti jsou uspořádány vzestupně do řady

$$d_{i1,j1} \le d_{i2,j2} \le \ldots \le d_{iN,jN}, \text{ kde } N = n(n-1)/2$$

a $\hat{d}_{ij}$ jsou odhadovány tak, aby splnily podmínku *slabé monotonicity (WM)*

$$\hat{d}_{i1,j1} \le \hat{d}_{i2,j2} \le \ldots \le \hat{d}_{iN,jN}, \text{ nebo}$$

nebo podmínku *silné monotonicity (SM)*

$$\hat{d}_{i1,j1} < \hat{d}_{i2,j2} < \ldots < \hat{d}_{iN,jN}.$$

Prvním krokem k získání počátečních odhadů predikovaných vzdáleností $\hat{d}_{ij}$ bývá vždy metrické vyčíslení. Pak následuje nemetrický přístup monotónní regrese. Indexový graf úpatí veličiny *stress* je užitečnou pomůckou i u nemetrické metody. Hledá se jednak zlom na tomto

grafu a jednak se vyšetřuje, kdy veličina *stress* nabyde hodnot menších než 0.05, resp. 0.01. Takový index, čili počet dimenzí, se pak jeví jako optimální. Obdobně, jako metrická metoda CMDS, ústí i nemetrická NNMDS ve vícerozměrnou škálovací mapu, na které se sleduje roztřídění vyšetřovaných objektů.

Korespondenční analýza (CA): je grafická metoda k zobrazení vnitřní závislosti, asociace v tabulce četností. Soustředíme se na dvojrozměrnou tabulku četností zvanou *kontigenční tabulka*, která obsahuje *n* řádků a *m* sloupců. Diagram korespondenční analýzy, *subjektivní mapa* obsahuje dva soubory bodů: soubor *n* bodů odpovídajících řádků a soubor *m* bodů, odpovídajících sloupců. Korespondenční analýza je kompozitní technika, protože subjektivní mapa je založena na asociaci mezi souborem objektů a souborem popisných znaků, zadaných člověkem. Polohy bodů pak přímo vyjadřují asociaci. I když je podobná faktorové analýze, zasahuje dále za faktorovou analýzu. Její přímou aplikací je zobrazování korespondence kategorií proměnných, znaků, které jsou měřeny v nominální stupnici. Tato korespondence je základem vytváření subjektivní mapy.

Řádkové body (obr. 26), které jsou těsně u sebe indikují řádky, které mají podobné profily v celém sloupci. *Sloupcové body* (obr. 25), které jsou blízko u sebe indikují sloupce s podobnými profily směrem dolů přes všechny řádky. Konečně řádkové body, které jsou těsně ke sloupcovým bodům představují kombinace, které se objeví častěji než by se očekávalo u nezávislého modelu, ve kterém řádkové kategorie nejsou vztaženy ke sloupcovým.

Běžný výstup z korespondenční analýzy obsahuje "nejlepší" dvojrozměrné zobrazení dat (obr. 27), podél kterého jsou souřadnice zobrazených bodů a dále míru *inertia* vyjadřující množství informace zobrazené v každé dimenzi. Korespondenční analýza přináší řadu výhod:

První výhodou je subjektivní mapa, která zobrazuje tabulku vícerozměrných katego-rických proměnných jako jsou třeba znaky výrobků proti výrobcům. Přístup umožňuje buď analyzovat existující odezvy nebo shromáždit odezvy u nejméně omezeného typu měření nominální nebo kategorické úrovně. Respondent potřebuje například u souboru objektů hodnotit *ano* a *ne* pro celou řadu znaků. Odpovědi jsou shromážděny v tabulce a analyzovány. Ostatní vícerozměrné techniky jako faktorová analýza vyžadují u každého objektu numerickou hodnotu každého znaku.

Druhou výhodou je zobrazení vedle vztahu mezi řádky a sloupci také vztahů mezi kategoriemi buď řádků nebo sloupců. Jsou-li například sloupcemi znaky v těsné blízkosti, budou mít u všech objektů vesměs podobné profily. Tento způsob třídění znaků je velmi podobný faktorové analýze a metodě hlavních komponent.

Třetí výhodou je poskytnutí společného obrazu řádkových a sloupcových kategorií ve stejném počtu dimenzí (obr. 27). Modifikace některých programů dovoluje porovnání vnitřních bodů, ve kterém je relativní blízkost vztažena k vyšší asociaci mezi oddělenými body. U těchto srovnání je umožněno současné vyšetření řádkových a sloupcových kategorií. Analýza umožní identifikovat třídy objektů, charakterizovaných znaků, které jsou v těsné blízkosti.

Vedle výhod přináší korespondenční analýza také nevýhody a omezení. Jde o popisnou techniku, která se nehodí ke statistickému testování hypotéz. Korespondenční analýza se uplatní v rámci exploratorní analýzy dat. Týká se snížení dimensionality, i když nemá proceduru k určení vhodného počtu dimenzí.

Často jsme při analýze vícerozměrných dat postaveni před problémem "kvantifikovat kvalitativní data", nalezená v nominálních proměnných tj. nemetrických datech nebo v kódech. Korespondenční analýza se liší od ostatních vnitřně-závislostních technik ve své schopnosti zpracovávat nemetrická data a i nelineární vztahy. Provádí redukci dimenzí podobnou vícerozměrnému škálování typem subjektivního mapování nebo faktorové analýzy. Ukazuje na míru asociace mezi řádkovými a sloupcovými kategoriemi.

6. Postup analýzy vícerozměrných dat

Při vyšetřování jednotlivých úloh je třeba postupovat dle následujícícho postupu:

1. *Standardizace:* vícerozměrné analýze obvykle předchází standardizace čili škálování proměnných.

2. *Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi proměnnými:* vyčíslení výběrové střední hodnoty každé proměnné, odhad kovarianční matice S a její normované podoby - korelační matice R. Matice R obsahuje Pearsonovy párové korelační koeficienty ρ_{ij}, které se diskutují. Užitečný je především diagram korelační matice.

3. *Exploratorní analýza dat EDA:* (a) posoudí podobnost objektů pomocí vizuálních rozptylových diagramů typu symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy), (b) nalezne vybočující objekty nebo vybočující proměnné, mnohdy k nevhodné analýze, (c) stanoví, zda platí předpoklad lineárních vazeb, (d) testuje všechny předpoklady o datech (normalitu, nekorelovanost, homogenitu).

4. *Určení vhodného počtu latentních proměnných:* matice S nebo R se rozloží na vlastní čísla λ_i a vlastní vektory v_i. Z indexového grafu úpatí vlastních čísel (Scree plot) se určí vhodný počet latentních proměnných (pro zobrazení v rovině se obvykle dává přednost prvním dvěma latentním proměnným), které ještě dostatečně popisují proměnlivost v datech. Když se latentní proměnné podaří pojmenovat a dát jim i fyzikální, biologický či jiný věcný význam, jedná se o faktory. V opačném případě jde o hlavní komponenty.

5. *Určení struktury v proměnných (PCA):* hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu komponentních vah. Hledání struktury v objektech a třídění objektů do shluků se provede v rozptylovém diagramu komponentního skóre. Dvojný graf je přehledným spojením obou předešlých grafů a ukáže interakci objektů a proměnných.

6. *Určení struktury a vzájemných vazeb v objektech:* klasifikační postupy zařadí v diskriminační analýze analyzovaný objekt do jednoho již existujícího a předem zadaného shluku. Neutříděnou skupinu objektů lze uspořádat do shluků a výsledek třídění zobrazit dendrogramem v analýze shluků. V hierarchickém postupu je třeba k vytvoření shluků vybrat vzdálenost mezi objekty (Eukleidovskou, Manhattanovskou, Mahalanobisovu) a jednu z nabídnutých metod: průměrovou, centroidní, nejbližšího souseda, nejvzdálenějšího souseda, mediánovou, Wardovu. Nehierarchické postupy rozdělí objekty do shluků, v nichž jsou předem umístěni typičtí reprezentanti.

7. *Soulad nalezené struktury objektů a vzájemných vazeb v dendrogramu a PCA grafech:* je třeba vyšetřit a komentovat nalezenou strukturu a vazby *jednotlivých proměnných,* nalezenou jednak v PCA a jednak v dendrogramu podobnosti proměnných analýzou vzniklých shluků. Dále je třeba komentovat také strukturu a vazby *klasifikovaných objektů,* nalezenou v PCA a v dendrogramu podobnosti objektů.

Využitím programového systému ADSTAT, resp. programu STATGRAPHICS, SCAN, MINITAB, STATISTICA, S-Plus atd. lze analyzovat dané úlohy.

Úloha 1. *Vícerozměrná analýza dat 38 rozličných vín dle 24 proměnných znaků*
Zdrojová matice dat obsahuje 38 objektů rozličných vín, analyzovaných dle 24 proměnných čili naměřených vlastností a znaků. Je třeba analyzovat vnitřní strukturu vín dle analytické odezvy v proměnných a nalézt především shluky podobných druhů vín. Existují také podobné či korelující proměnné? Které proměnné jsou silně nepodobné s ostatními?

Řešení: Indexový graf úpatí vlastních čísel (obr. 1 a 5) ukazuje, že první zlom na křivce se nachází pro $m = 2$ a pro $m = 7$ je vlastní číslo menší než 1. Byly zvoleny první dvě hlavní komponenty. Složení prvních tří hlavních komponent y_1, y_2 a y_3 z původních 24 proměnných x_1 až x_{24} ukazují diagramy na obr. 2, 3 a 4.

Pro chemika je velmi důležitý *graf prvních dvou komponentních vah* (Plot Components Weights) (obr. 6), který ukazuje proměnné v blízké korelaci a vyznamně ovlivňující proměnlivost v datech. Je-li úhel mezi dvěma průvodiči malý blízký nule, jsou dotyčné původní proměnné v silné korelaci. V korelaci jsou proto proměnné *Region-Body-B*, dále proměnné *P-Cr-Na*, dvojice *K-Sr*, dvojice *Mg-Ba*, silnou korelaci také vykazují proměnné *Mo-Pb-Cu*, dále *Quality-Flavor*. Je-li mezi původními proměnnými velký úhel (např. *Ca-Mn*), je korelace velmi slabá a korelační koeficient se blíží k nule. Je-li dále průvodič původní proměnné x_i dostatečně dlouhý, je tato původní proměnná statisticky významná a vyýznamně také ovlivňuje proměnlivost v datech, např. *Sr, Aroma, Ca, Region, Na,* atd. Je-li naopak průvodič krátký, neovlivňuje významně proměnlivost v datech, např. *Al, Cu.*

Rozptylový diagram komponentního skóre (Scatterplot) (obr. 7) ilustruje klasifikaci vín dle 24 proměnných. Okolo očíslovaných bodů čili typů vín je možné udělat elipsy a vyznačit tak jednotlivé shluky. Některá vína jsou svým charakterem odlehlá od ostatním a vůči nim naprosto nepodobná, např. druhy 14, 12, 24, 20, 1, 9. V diagramu lze označit 3 až 4 shluky nejpodobnějších vín. Interpretace rozptylového diagramu komponentního skore lze shrnout do těchto bodů:

1. Umístění objektů: Objekty daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.
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. Pravidlo platí, pokud se nejedná o zdánlivou nehomogenitu danou sešikmením dat a odstranitelnou transformací proměnných.
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. 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.

Dvojný graf (Biplot) (obr. 8) kombinuje předchozí dva grafy. Úhel mezi průvodiči dvou proměnných x_j a x_k je nepřímo úměrný velikosti korelace mezi těmito dvěma proměnnými. 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 původní proměnné x_j do hlavní komponety, čili je úměrná komponentní váze. Kombinace obou grafů v jediném přináší cenné srovnání, jeden graf působí zde doplňkově 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.

V datech se mohou vyskytovat také odlehlé objekty, které se indikují *Williamsovým grafem* (obr. 9) nebo *grafem Mahalanobisovy vzdálenosti* ve 24 rozměrném prostoru původních proměnných (obr. 10). Obě tyto pomůcky ukazují na odlehlý objekt č. 20. *Diagram proměnlivosti původních proměnných* ukazuje, které proměnné x_i mají velkou míru informace, a tím pádem vysokou proměnlivost.

Obr. 12, 13, 14 a 15 ukazují *dendrogram proměnných.* Interpretace dendrogramu je snadná: objekty blízko sebe či si silně podobné jsou propojeny spojovací úsečkou hodně dole, mají malou vzdálenost, čili značnou podobnost. Objekty propojené hodně vysoko mají malou podobnost a mezi sebou vykazují velkou vzdálenost, např. *Cu* nebo *Al* se velice liší od všech ostatních proměnných. Míra podobnosti nebo naopak míra vzdálenosti dvou objektů se může přečíst přímo na ose. Počet vhodných shluků může být snadno určen zakreslením rovnoběžky s x-ovou osou do diagramu. Vztyčíme-li, například, kolmici v bodě 80 na ose vzdáleností y, dostaneme určitý počet shluků. Nejvhodnější techniku shlukování vybereme na základě dvou rozhodčích kritérií, kofenetické korelace CC a kritéria delta. Na čtyřech dendrogramech jsou uvedeny shluky, vytvořené rozličnými technikami. Protože rozhodčí kritéria nalezla jako nejlepší techniku průměrovou, budeme tento dendrogram považovat za optimální a v naší analýze za výsledný. Uživatel může porovnat jemné rozdíly mezi dendrogramy.

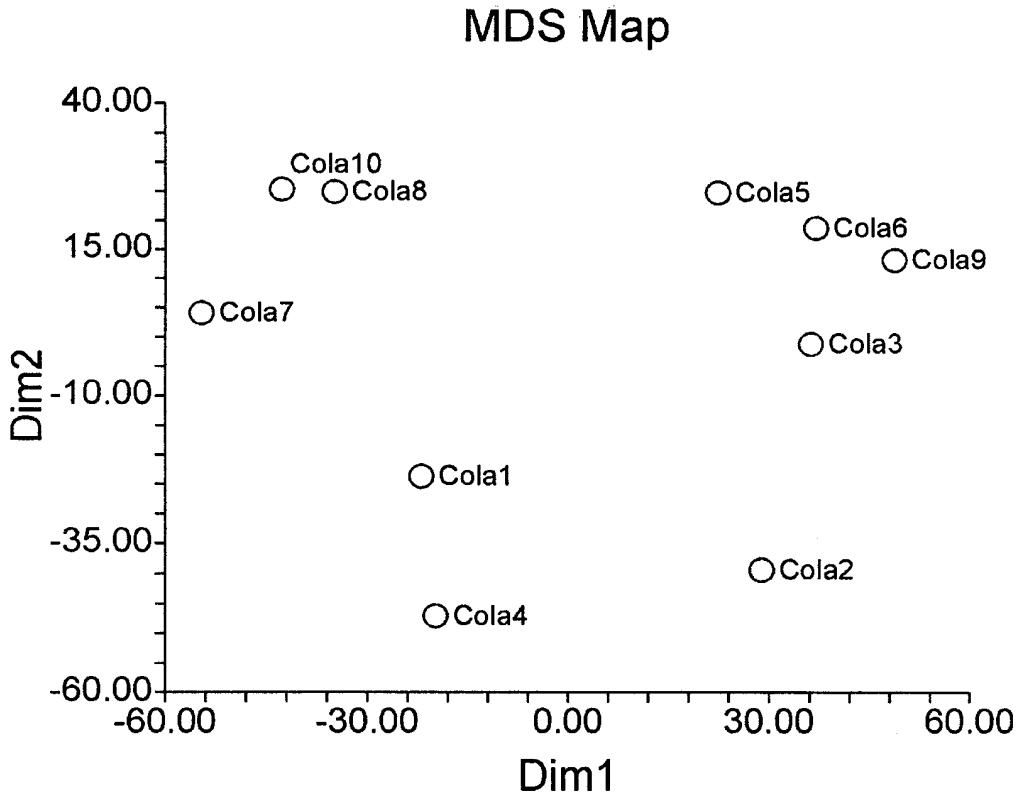
Dendrogram objektů čtyřmi metodami přináší obr. 16, 17, 18 a 19. Dle kofenetického koeficientu a kritéria delta je nejlepší technika tvorby dendrogramu metoda průměrová. Za vybočující je třeba považovat druhy vína 7, 1, 12, 20 a 24. Objekty blízko sebe jsou propojeny spojovací úsečkou nízko, mají značnou podobnost. V dendrogramech lze číst jako v mapě a orientovat se tak, která vína jsou si podobná a která nepodobná. Bývá zvykem, že na určité hladině podobnosti (např. 80%) na y-nové ose se vede rovnoběžka s x osou a pod touto přímkou se identifikují jednotlivé shluky. Na obr. 20, 21, 22 a 23 jsou dendrogramy analyzované jiným software NCSS2000, tyto dendrogramy jsou proti předešlým otočeny o 90°. Jejich výpověď je však analogická.

Dvojrozměrná MDS škálovací mapa (obr. 24) pro 38 druhů vín klasifikovaných dle 24 původních proměnných představuje hlavní výsledek metody vícerozměrného škálování. Mapa umožňuje interpretovat použité proměnné, jejich podobnost, jejich korelaci a nepodobnost. Isolované body proměnných (např. *Ba, Mg, Mn, B, Si*) jsou si nepodobné s ostatními. Naopak blízké body proměnných (např. *Arom, Quality, Flavor*) nebo (*Oak, Mo, Pb, Clar*) vykazují značnou korelaci proměnných. K podobným závěrům dospěla i korespondenční analýza (obr. 25, 26 a 27), která klasifikuje proměnné (obr. 25), objekty - druhy vína (obr. 26) a konečně ve dvojném grafu obojí dohromady (obr. 27).

Úloha 2. *Vícerozměrné škálování u analýzy podobnosti*

Vícerozměrné škálování MDS ukážeme na datech podobnosti 10 výrobků Coly: 50 respondentů hodnotilo a vzájemně porovnalo 10 výrobků Coly (objekty) způsobem "každý s každým" a při dokonalé podobnosti byla přidělena nulová vzdálenost mezi dvěma objekty, zatímco při naprosté nepodobnosti vzdálenost 100. Z hodnot párových vzdáleností od 50 respondentů byla vždy vypočtena střední hodnota a zapsána do buňky vytvořené symetrické čtvercové matice. Z této matice se ve vstupních datech užije pouze trojúhelníková část, tj. prvky nad (nebo pod) diagonálou nul. Je třeba provést *dvojrozměrné škálování* a z výsledného grafu usoudit na podobné a nepodobné výrobky Coly.

Řešení: Veřejnost by v původních datech měla odhadnout míru shodnosti či podobnosti vždy mezi dvěma druhy Coly tak, že odhadne vzdálenost (čili nepodobnost) mezi nimi: naprosto stejné nápoje budou mít mezi sebou vzdálenost nula a naprosto odlišné vzdálenost 100. Těchto 45 hodnot trojúhelníkové matice vede k diagramu na obr. 28. Dvojrozměrný diagram rozložení 10 druhů nápoje Cola představuje hlavní výsledek vícerozměrného škálování. Často je nazýván *vícerozměrnou škálovací mapou* a umožňuje interpretovat matici vzdáleností mezi objekty ve dvojrozměrném diagramu. Neexistuje reálná orientace tohoto diagramu, diagramem je totiž možné libovolně otáčet okolo počátku. Důležité jsou relativní polohy objektů vůči sobě a pak hlavně poloha shluků objektů. Obrázek ukazuje, že jednotlivé druhy Coly jsou zřetelně roztříděny v rovině. Cola 3, 5, 6 a 9 tvoří jeden shluk druhů podobných vlastností, dále Cola 8 a 10 spolu s Colou 7 pak druhý shluk. Odlišné jsou Cola 1, 2, a 4, které se od předešlých dvou velmi odlišují, navíc Cola 1 se značně liší od Coly 2 a Cola 2 se značně liší od Coly 4. První shluk má dominantu Colu 9 a druhý shluk pak Colu 10. Okolo těchto dominant jsou soustředěny ostatní.



Obr. 28 MDS dvojrozměrný diagram rozložení 10 druhů nápoje cola (dvojrozměrná škálovací mapa).

Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.

[2] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods for Data Analysis*. Duxburg Press, Belmont, California 1983.

[3] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 6.

[4] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.

[5] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 12.

[6] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.

[7] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis,* Prentice Hall, 1982

[8] Ajvjazin S., Bežajeva Z., Staroverov O.: *Metody vícerozměrné analýzy,* SNTL Praha 1981

[9] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis,* Ellis Horwood, Chichester 1992.

[10] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,

[11] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,

[12] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994.

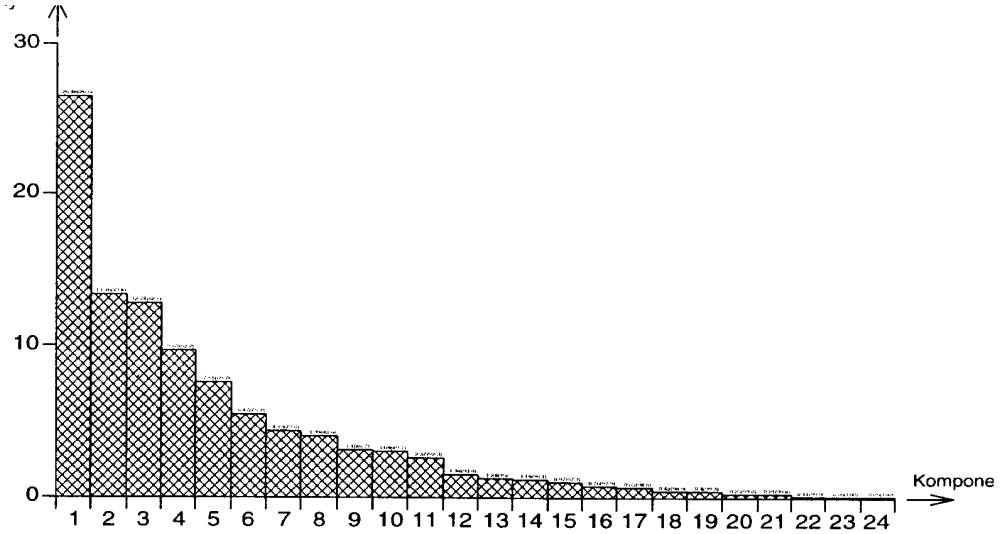
[13] Martens H., Naes T., *Multivariate calibration*, Wiley (1989) Chichester.

[14] Meloun M. , Militký J., *Sbírka úloh - Statistické zpracování experimentálních dat*, Univerzita Pardubice, 1996.

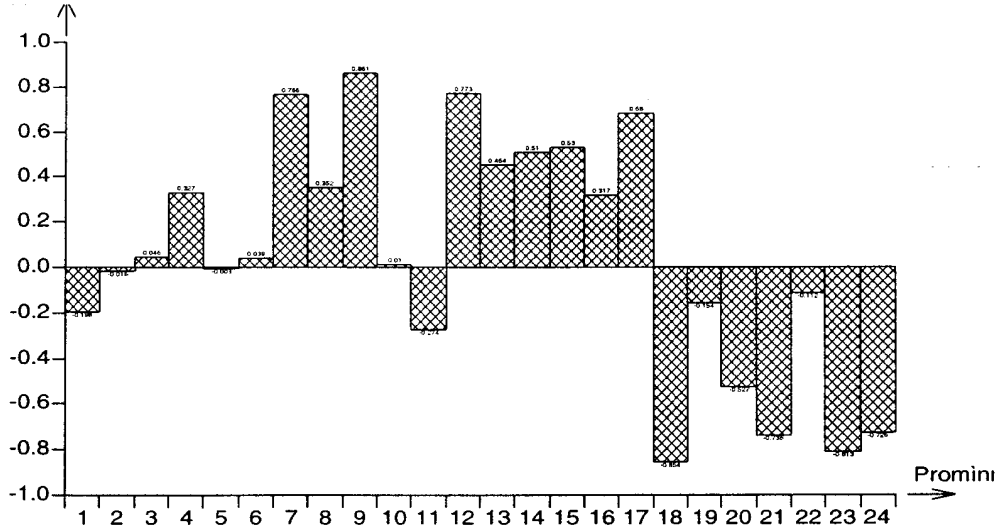
[15] Meloun M. , Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha 2002.

Poděkování:

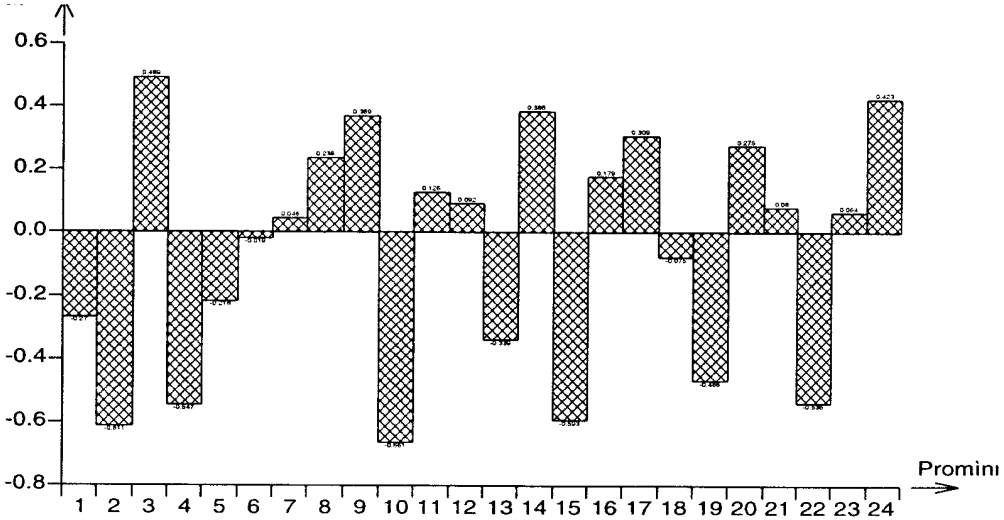
Publikace vznikla za podpory grantu NB/7391-3.



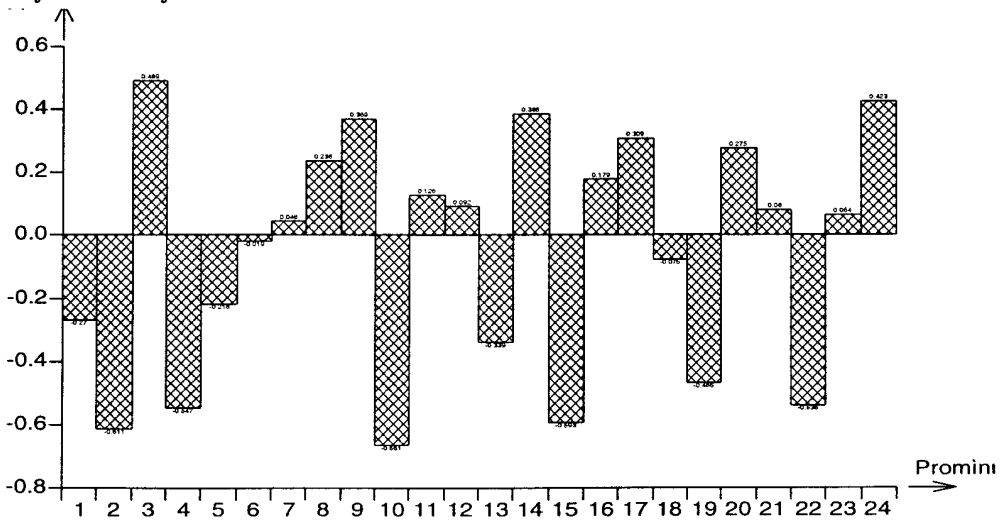
Obr. 1. Sloupcový diagram indexového grafu úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



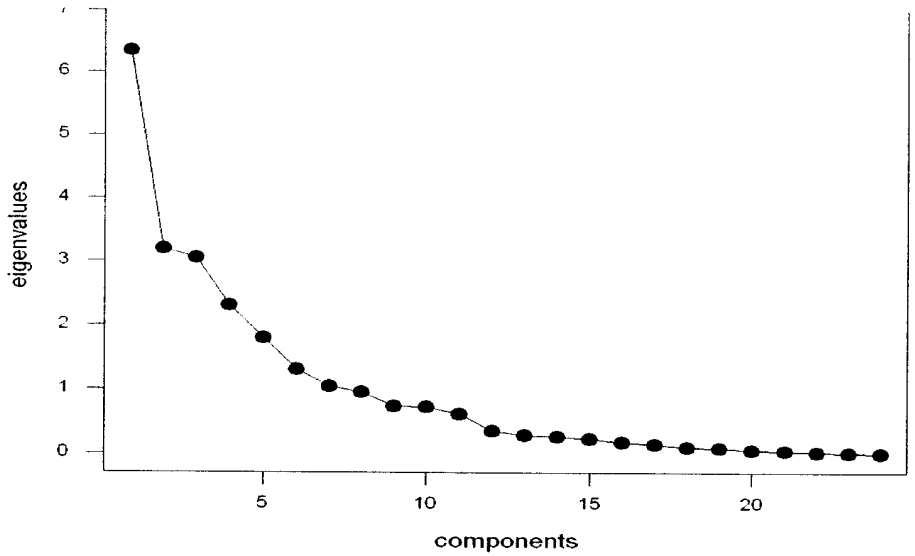
Obr. 2. Složení 1. hlavní komponenty z 24 původních proměnných pro 38 objektů zdrojové matice Wine.



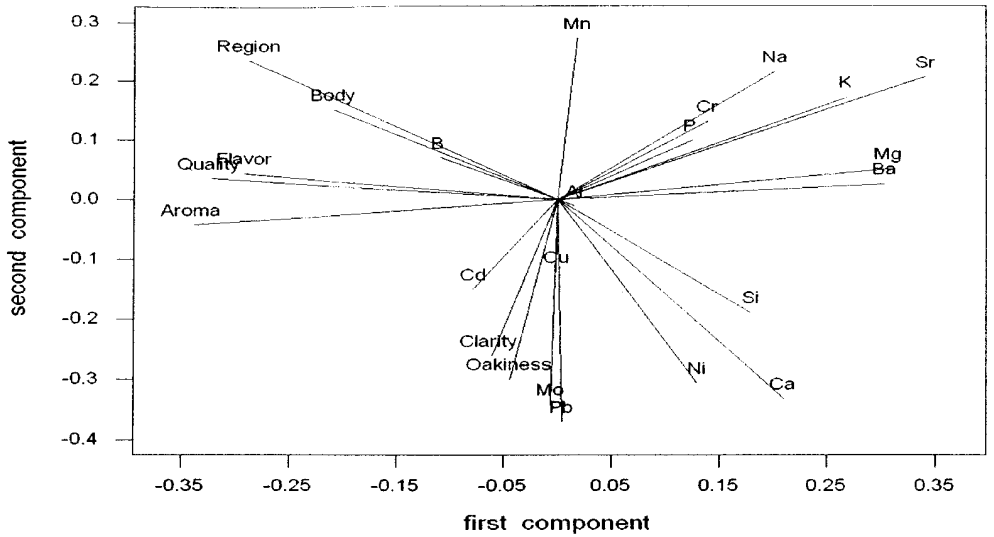
Obr. 3. Složení 2. hlavní komponenty z 24 původních proměnných pro 38 objektů zdrojové matice Wine.



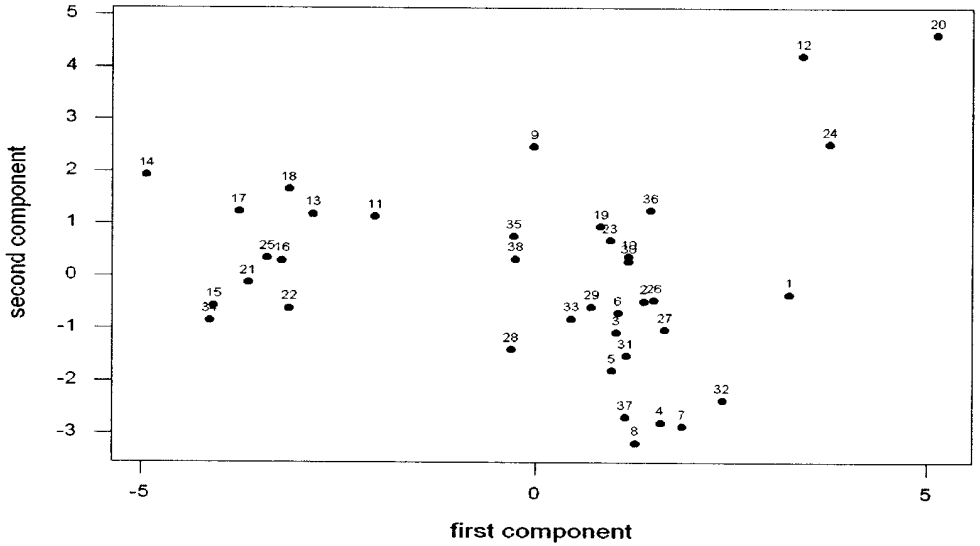
Obr. 4. Složení 3. hlavní komponenty z 24 původních proměnných pro 38 objektů zdrojové matice Wine.



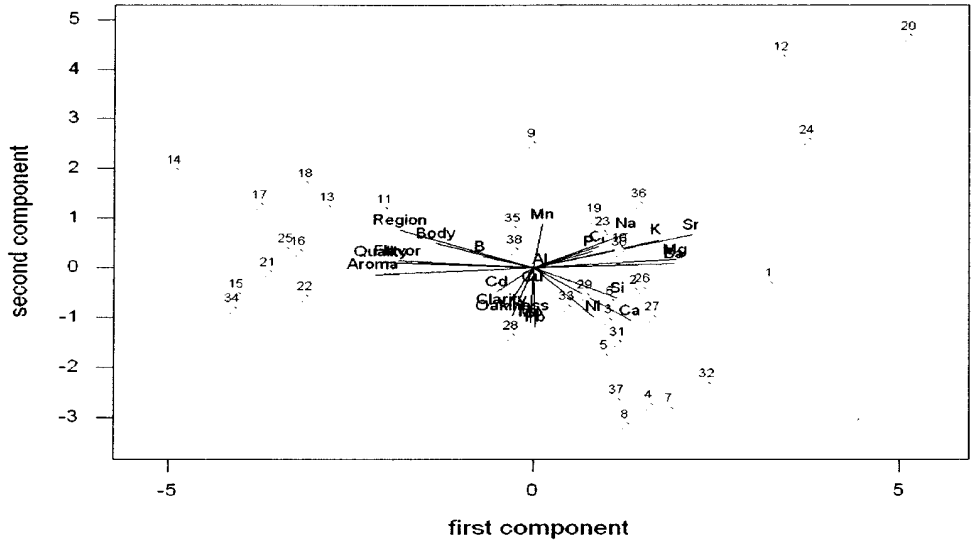
Obr. 5. Indexový graf úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



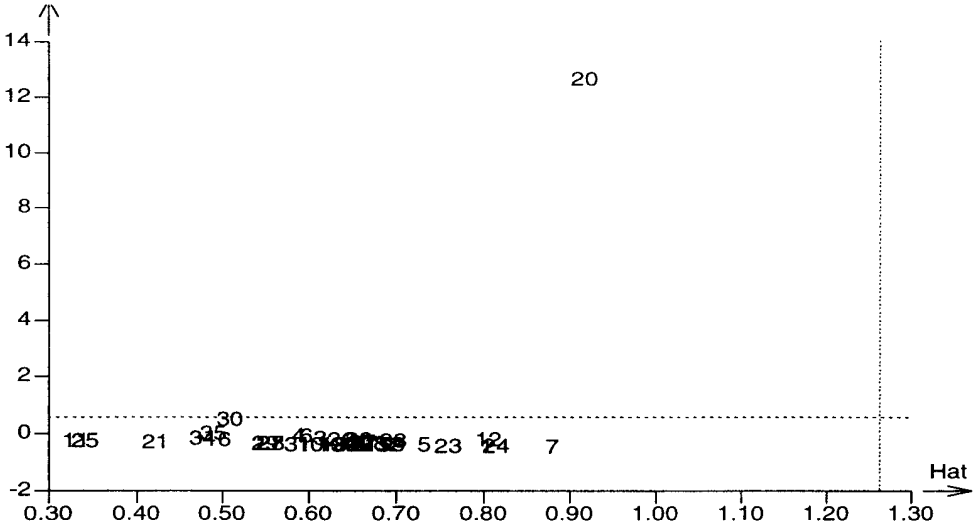
Obr. 6. Graf prvních dvou komponentních vah pro 38 objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.



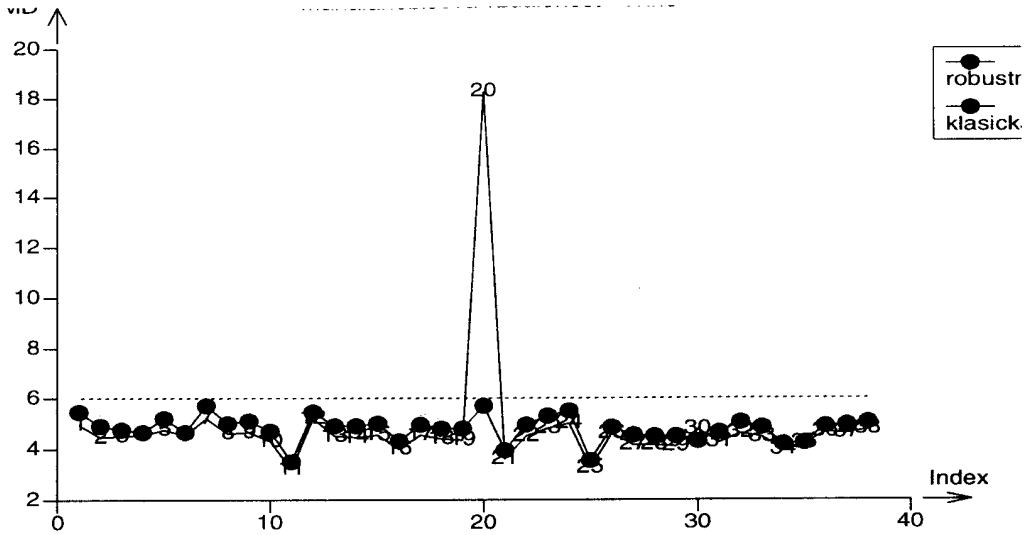
Obr. 7. Rozptylový diagram komponentního skóre pro 38 očíslovaných objektů a 24 původních proměnných zdrojové matice Wine.



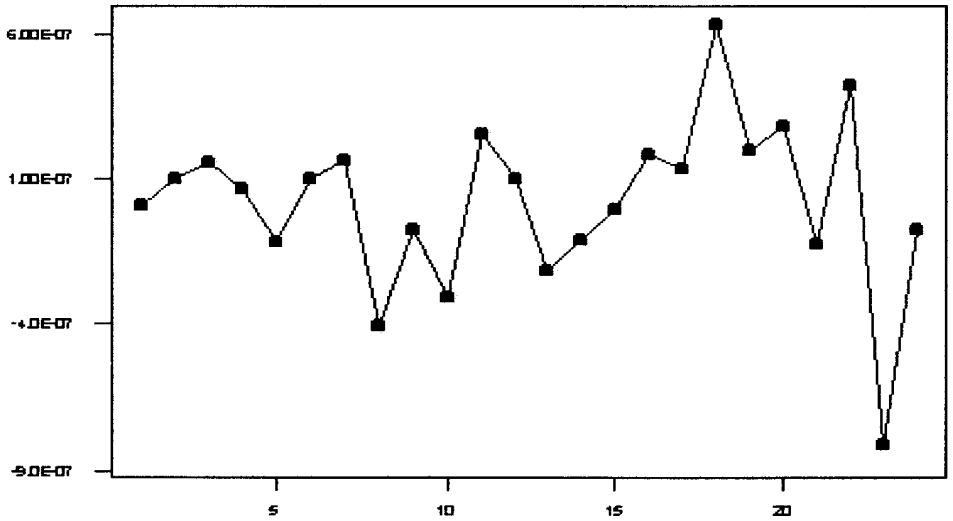
Obr. 8. Dvojný graf pro 38 očíslovaných objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.



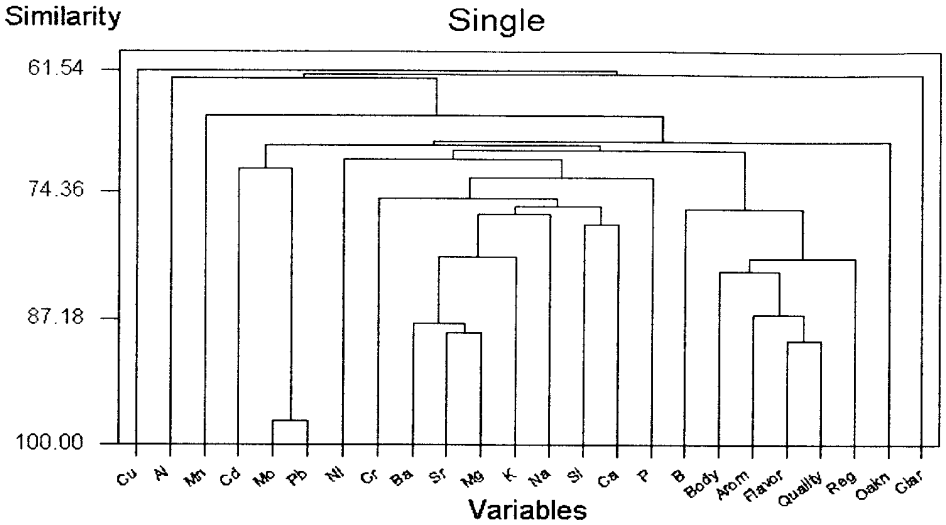
Obr. 9. Určení odlehlých objektů Williamsovým grafem pro 38 očíslovaných objektů a 24 původních proměnných zdrojové matice Wine.



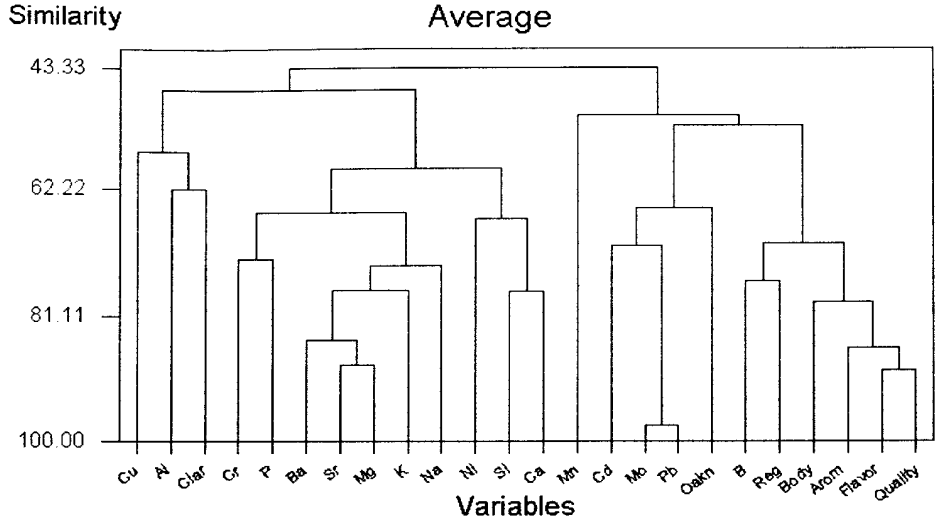
Obr. 10. Graf Mahalanobisovy vzdálenosti pro 38 očíslovaných objektů a 24 původních proměnných zdrojové matice Wine.



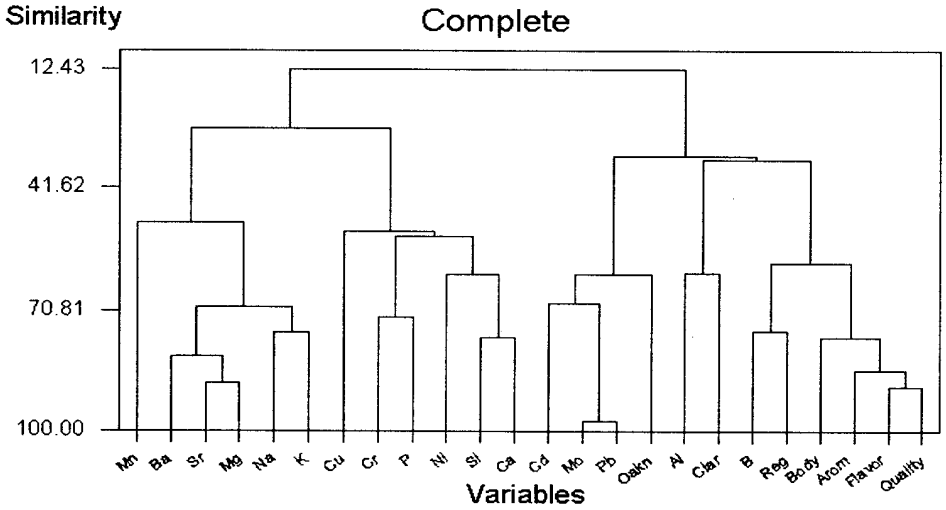
Obr. 11. Diagram proměnlivosti pro 24 původních očíslovaných proměnných zdrojové matice Wine.



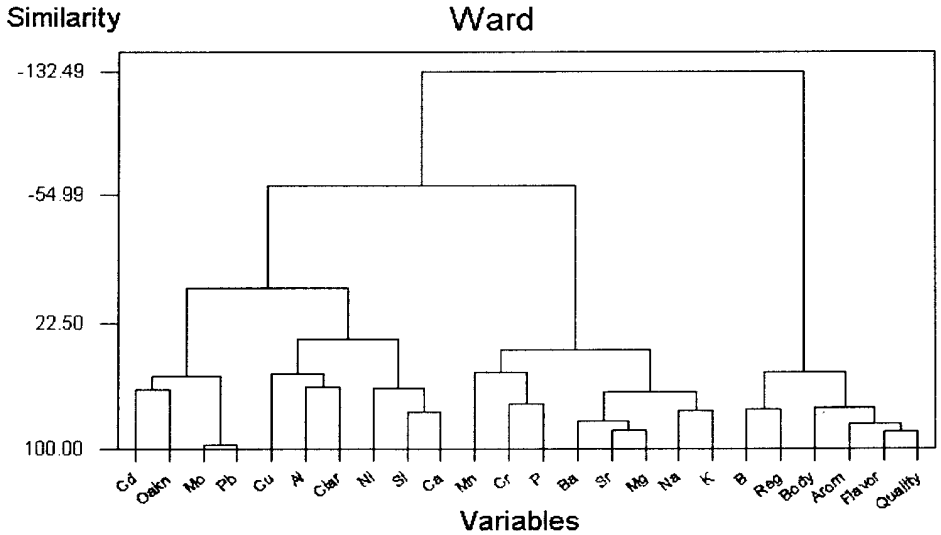
Obr. 12. Dendrogram proměnných metodou nejbližšího souseda (Single) pro 38 objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.



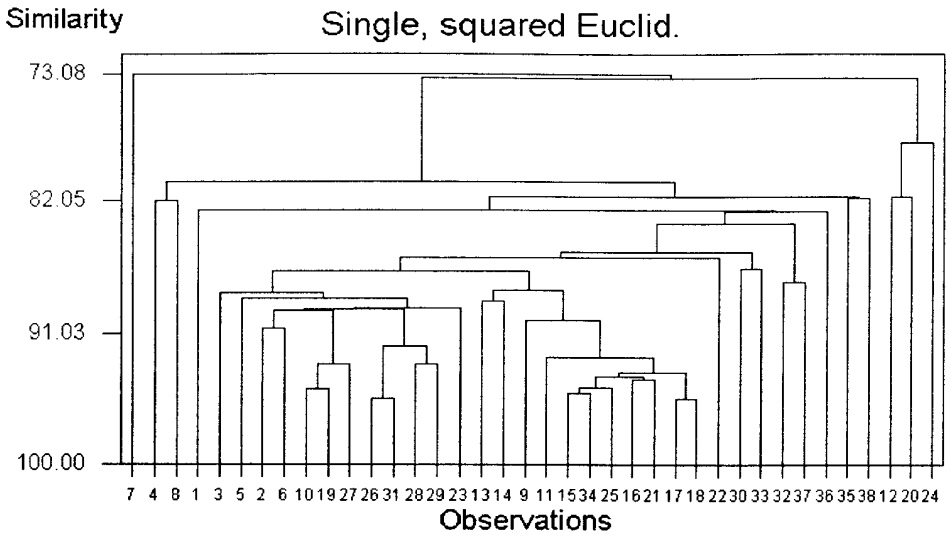
Obr. 13. Dendrogram proměnných metodou průměrovou (Average) pro 38 objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.



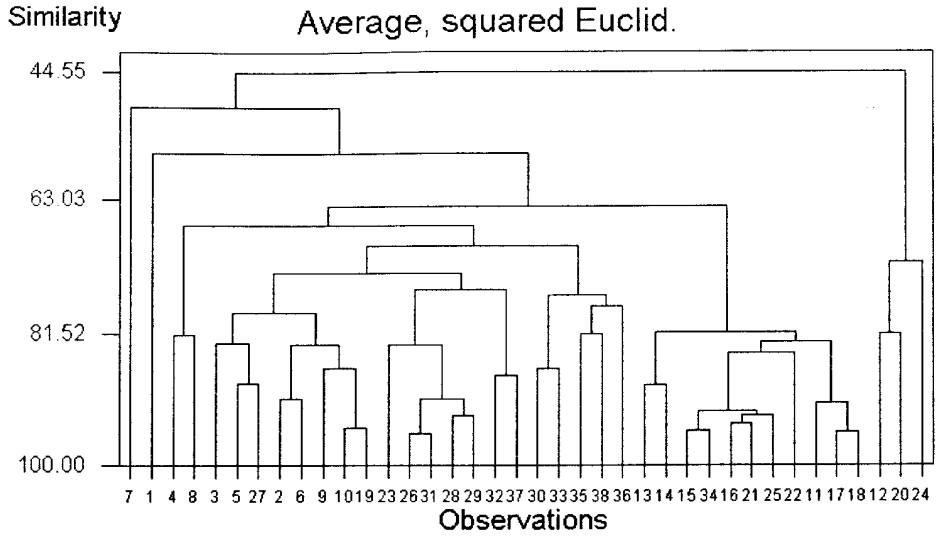
Obr. 14. Dendrogram proměnných metodou nejvzdálenějšího souseda (Complete) pro 38 objektů a 24 pojmenovaných původních proměnných.



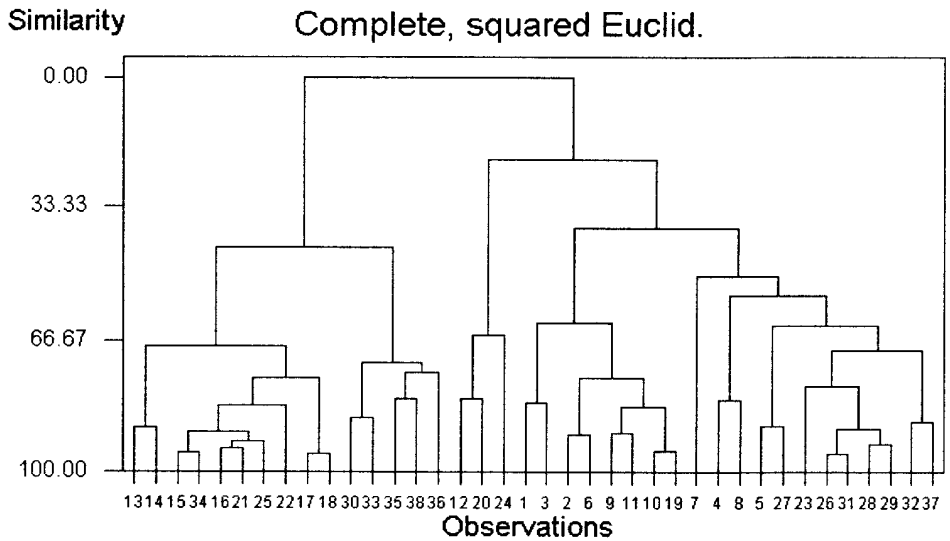
Obr. 15. Dendrogram proměnných metodou Wardovou (Ward) pro 38 objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.



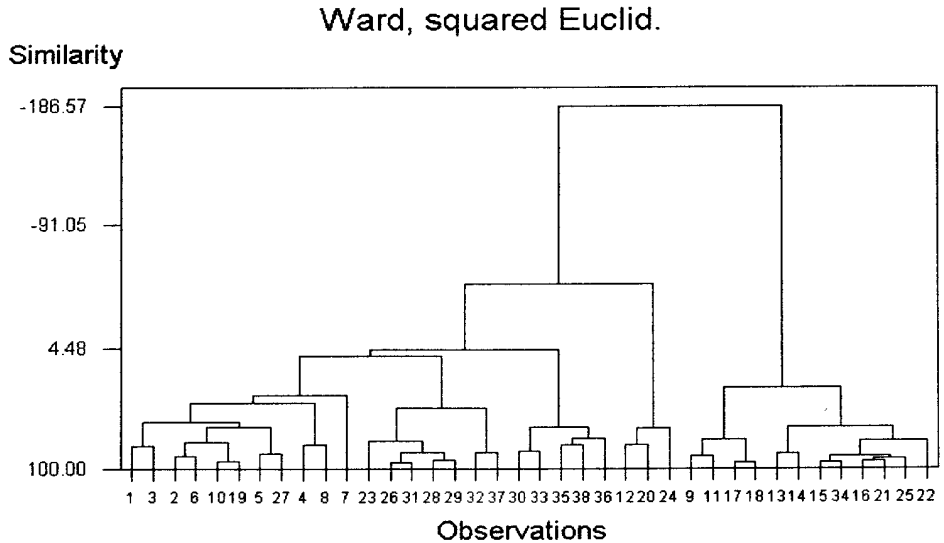
Obr. 16. Dendrogram objektů metodou nejbližšího souseda (Single) pro 38 očíslovaných objektů a 24 původních proměnných zdrojové matice Wine.



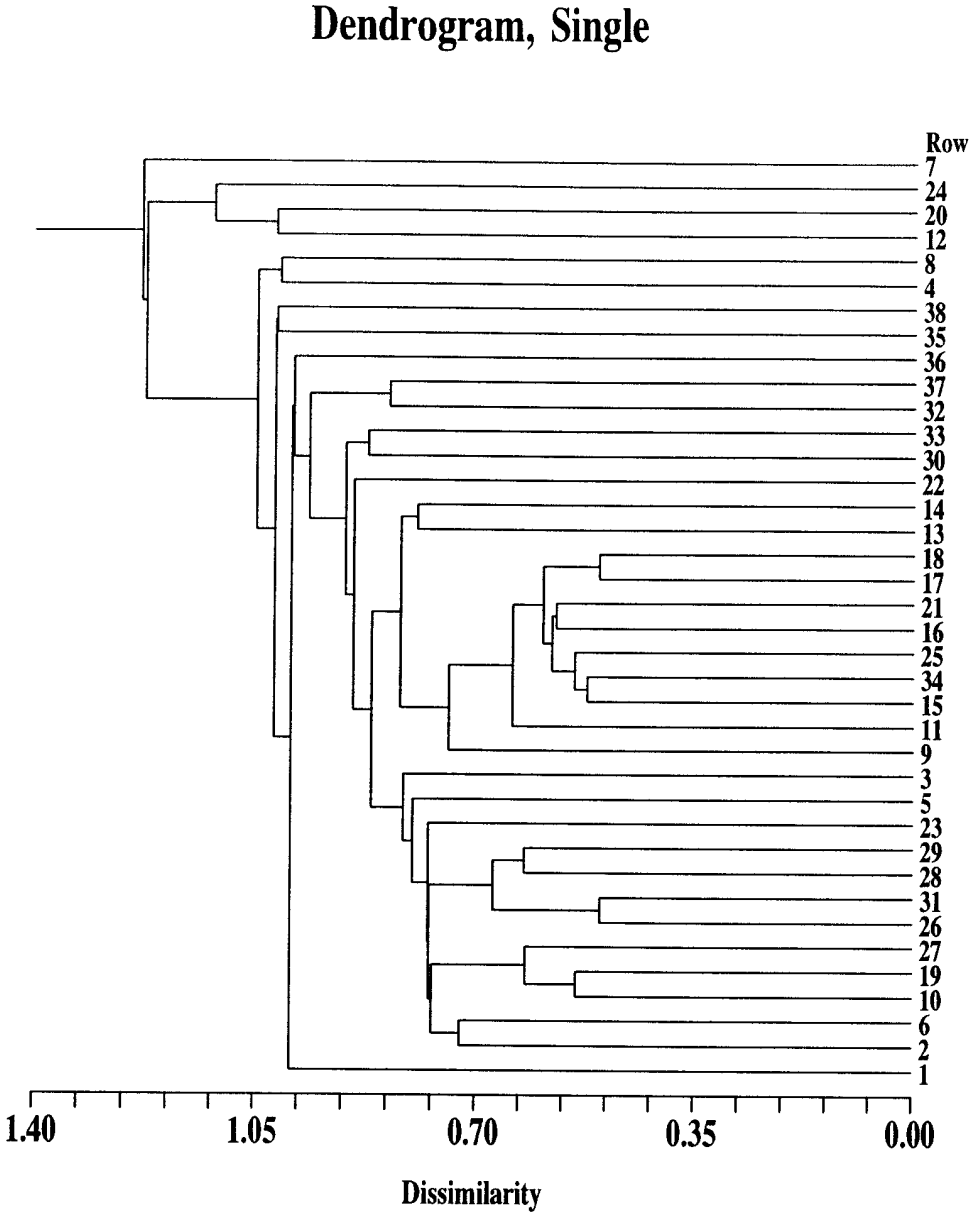
Obr. 17. Dendrogram objektů metodou průměrovou (Average) pro 38 očíslovaných objektů a 24 původních proměnných zdrojové matice Wine.



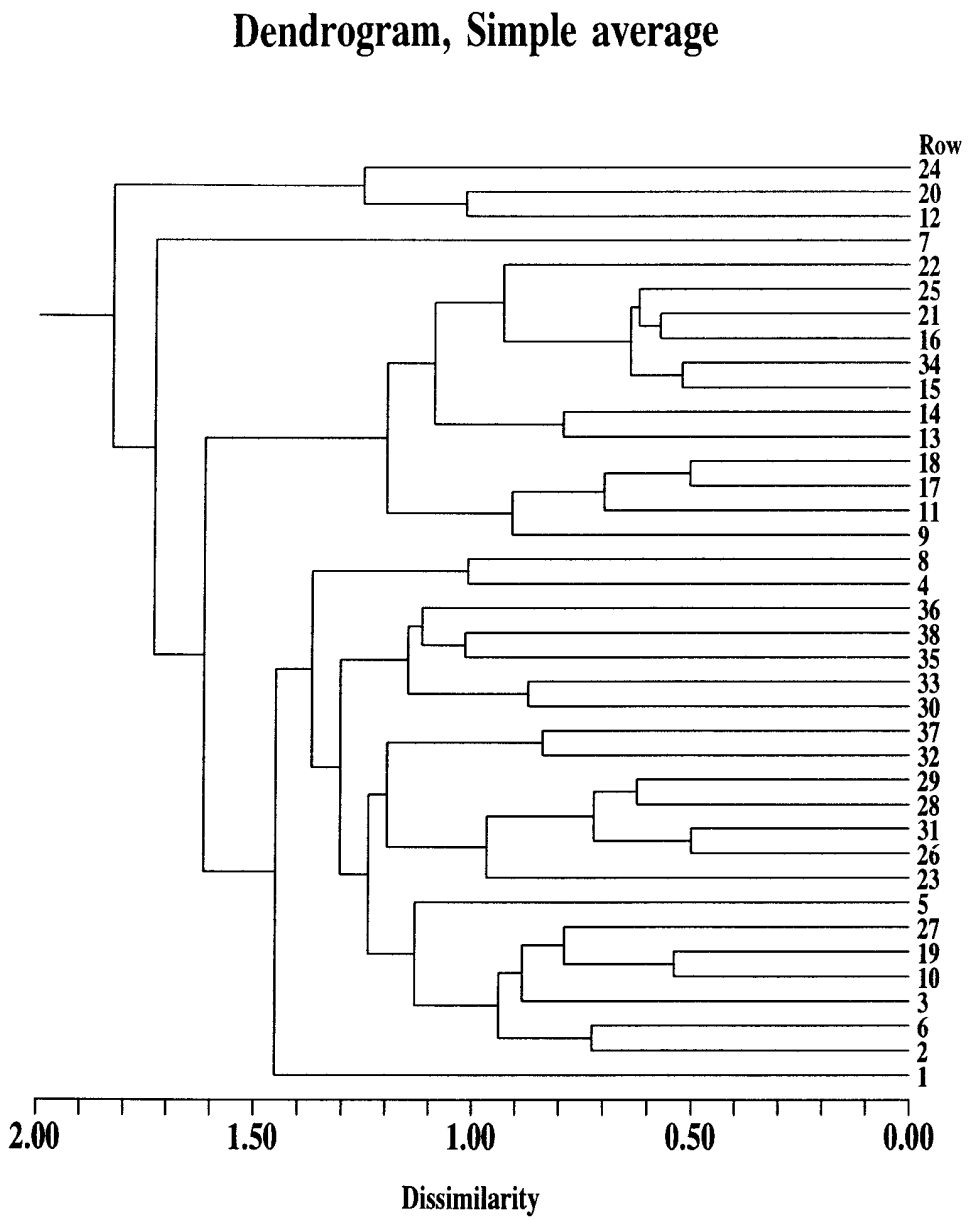
Obr. 18. Dendrogram objektů metodou nejvzdálenějšího souseda (Complete) pro 38 očíslovaných objektů a 24 proměnných zdrojové matice Wine.



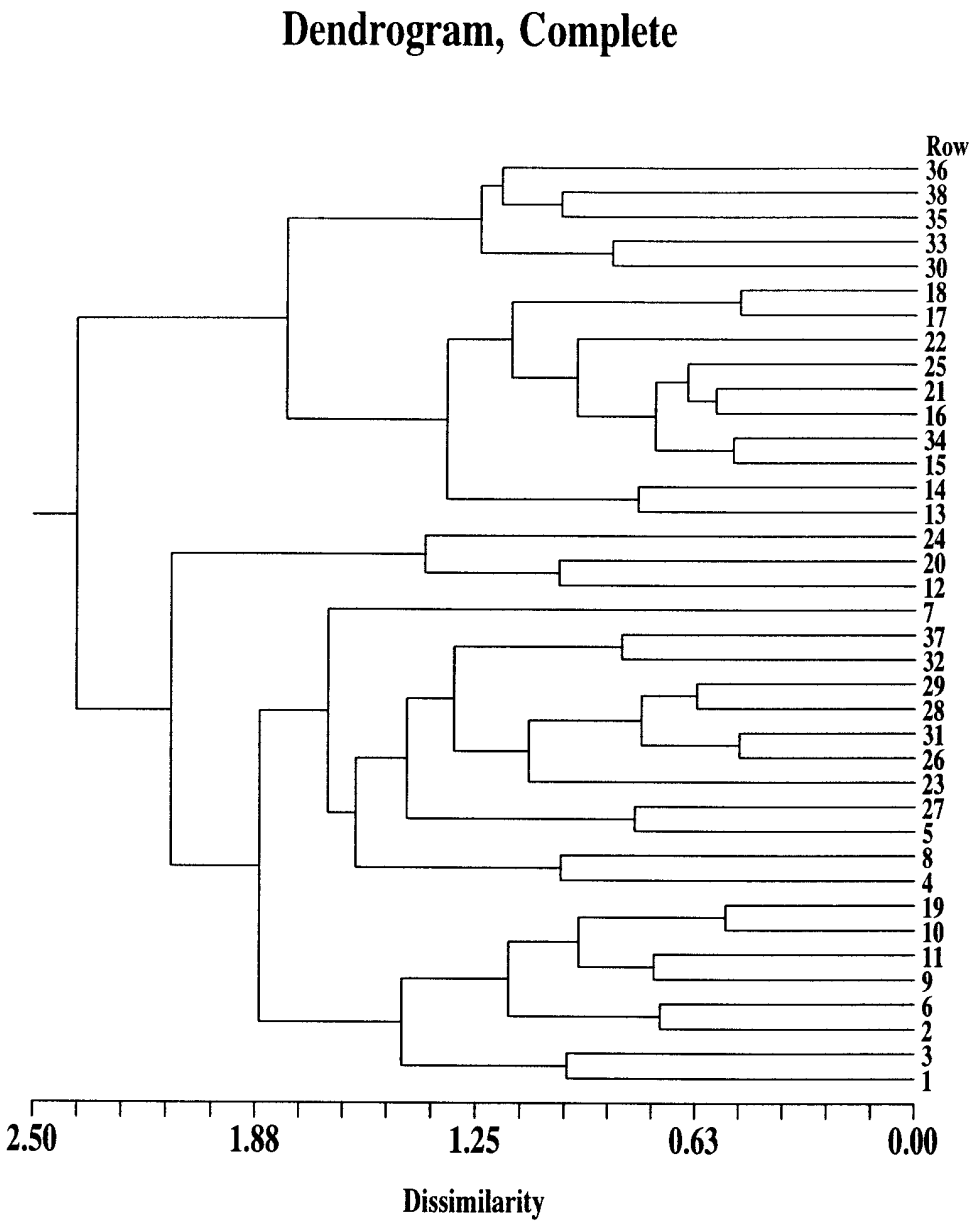
Obr. 19. Dendrogram objektů metodou Wardovou (Ward) pro 38 očíslovaných objektů a 24 původních proměnných zdrojové matice Wine.



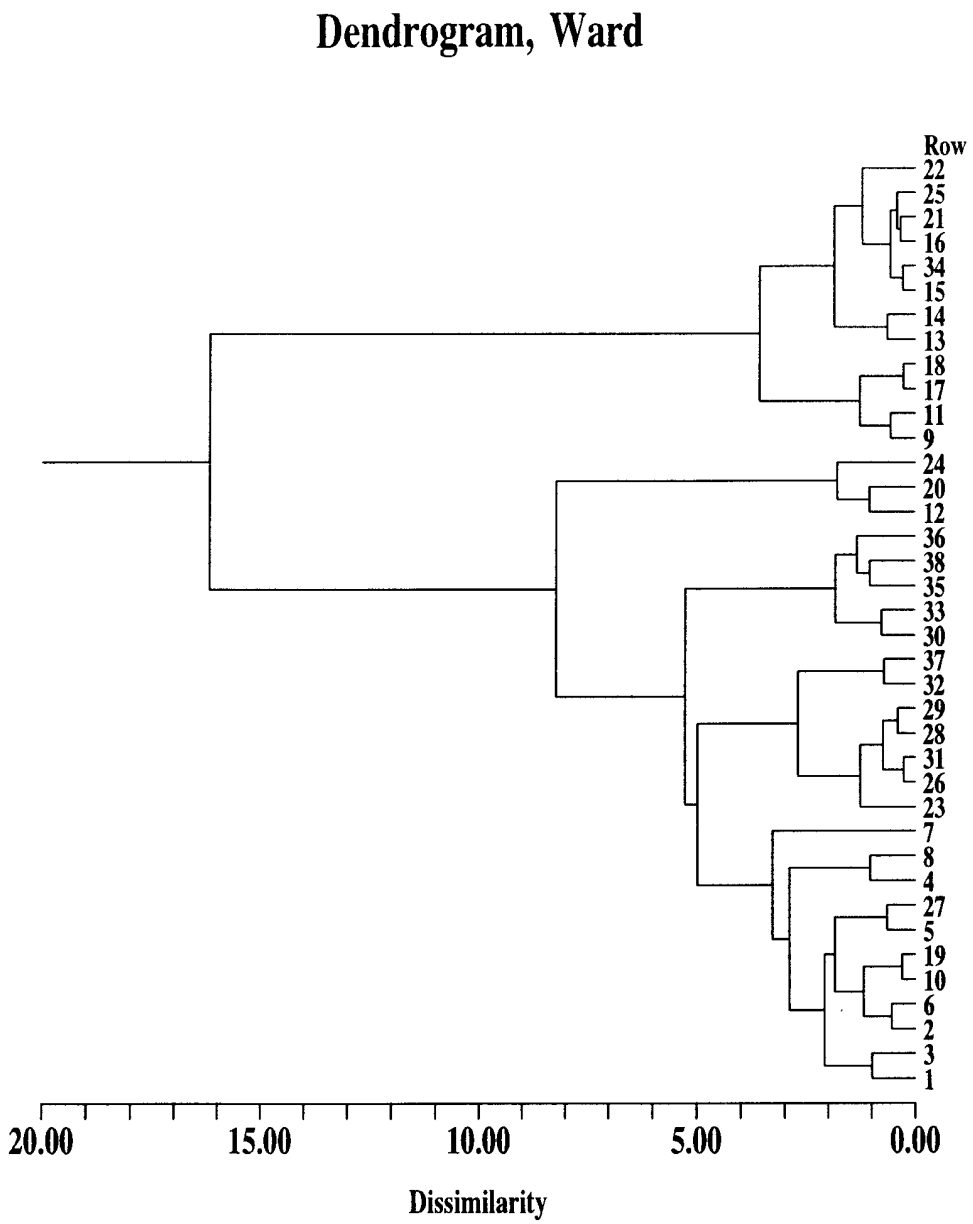
Obr. 20. Diagram objektů metodou nejbližšího souseda (Single) pro 38 očíslovaných objektů a 24 původních proměnných Wine (NCSS2000).



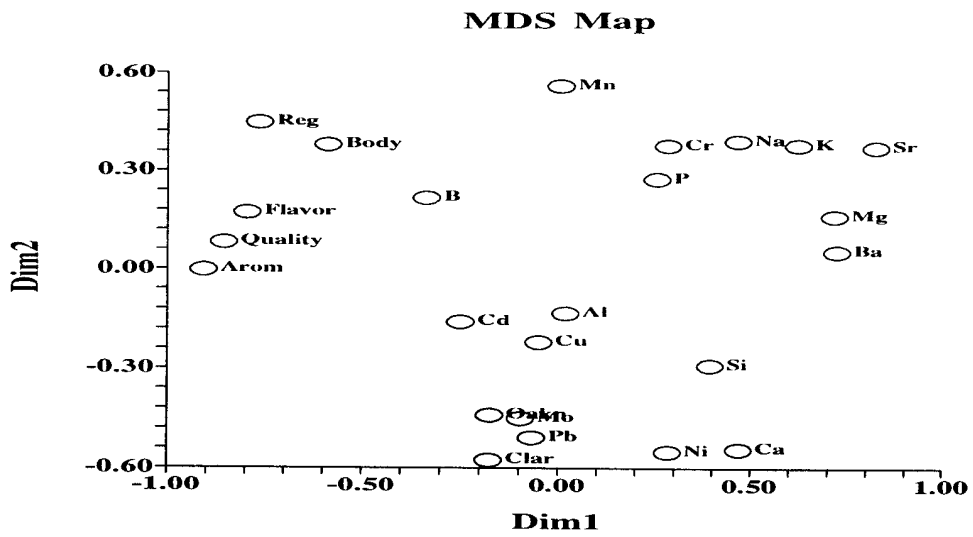
Obr. 21. Dendrogram objektů metodou průměrovou (Average) pro 38 očíslovaných objektů a 24 původních proměnných Wine (NCSS2000).



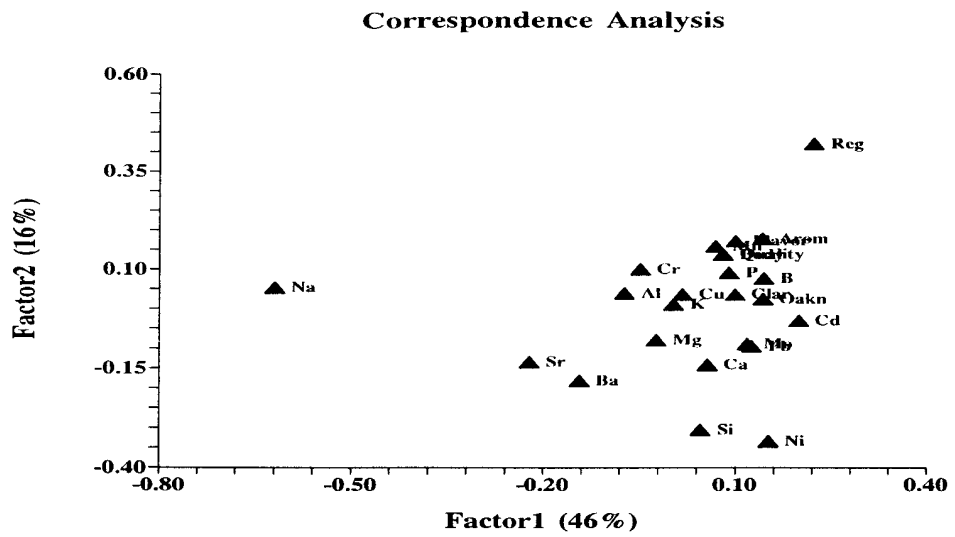
Obr. 22. Dendrogram objektů metodou nejvzdálenějšího souseda (Complete) pro 38 očíslovaných objektů a 24 původních proměnných Wine (NCSS2000).



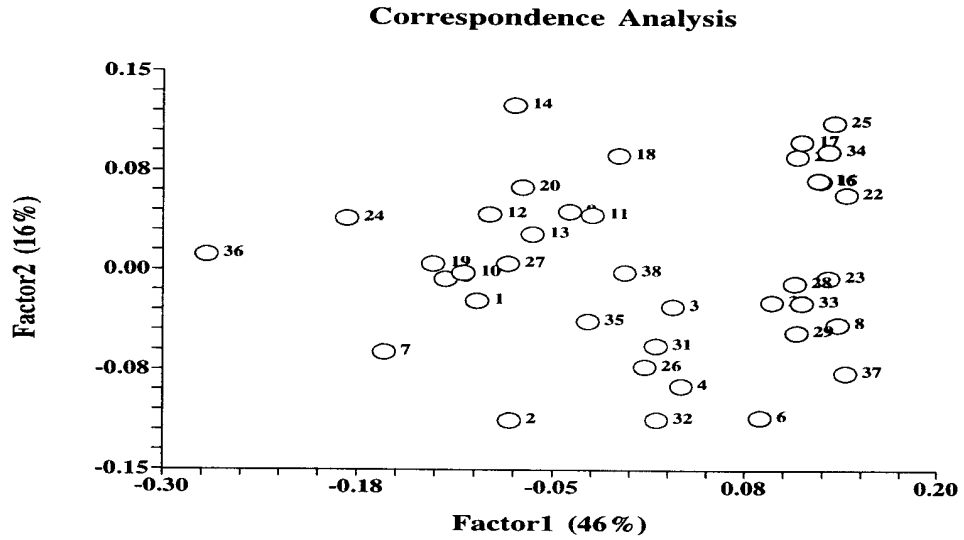
Obr. 23. Dendrogram objektů metodou Wardovou (Ward) pro 38 očíslovaných objektů a 24 původních proměnných Wine (NCSS2000).



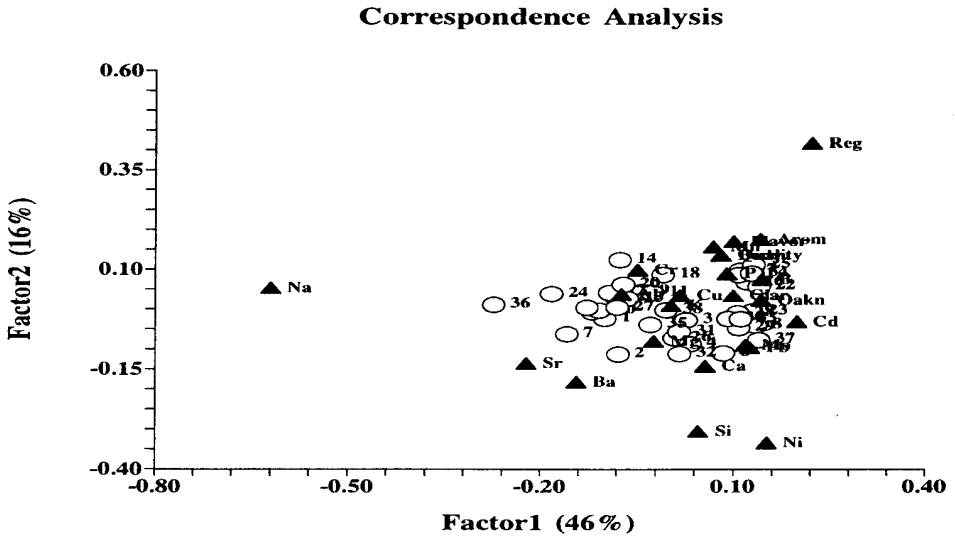
Obr. 24. Dvojrozměrná MDS škálovací mapa pro 38 objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.



Obr. 25. Graf sloupcových profilů korespondenční analýzy pro 38 objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.



Obr. 26. Graf řádkových profilů korespondenční analýzy pro 38 očíslovaných objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 27. Dvojný graf sloupcových i řádkových profilů pro 38 očíslovaných objektů a 24 pojmenovaných původních proměnných zdrojové matice Wine.