Vnitřní vazby a skrytá struktura v hutnických datech vícerozměrnou statistickou analýzou

Milan Meloun[1], Roman Lisztwan[2]
[1]*Katedra analytické chemie, Chemickotechnologická fakulta,*
Univerzita Pardubice, nám. Čs. Legií 565, 532 10 Pardubice,
email: milan.meloun@upce.cz, ICQ: 224-001-003, http: //meloun.upce.cz,

[2]*Třinecké železárny, Průmyslová 1000, 739 70 Třinec, email: roman.lisztwan@trz.cz*

Souhrn: Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných. 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. Výsledkem je dendrogram. Lze formulovat tři hlavní cíle analýzy shluků: popis systematiky, je tradičním využitím shlukové analýzy pro průzkumové cíle a taxonomii, což je empirická klasifikace objektů, zjednodušení dat, kdy analýza shluků poskytuje při hledání taxonomie zjednodušený pohled na objekty, a konečně identifikace vztahu, kdy po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty. Metoda hlavních komponent a tvorba shluků je demonstrována na typické úloze hutní kontrolní laboratoře.

Klíčová slova: metoda hlavních komponent, Shluková analýza, Dendrogram, Analýza ocele, Rozptylový diagram komponentního skóre, Cattelův indexový graf vlastních čísel, Graf komponentních vah, Korelace.

1. Úvod

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 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.

Cíle shlukové analýzy nelze oddělit od hledání a volby vhodných znaků k charakterizování shlukovaných objektů. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky. Volba znaků musí být provedena na základě teoretických, pojmových a praktických hledisek. Vlastní shluková analýza neobsahuje techniku k rozlišení významných a nevýznamných znaků. Provede pouze odlišení shluků. Nesprávné zařazení znaků vede k zahrnutí i odlehlých objektů, které mohou mít rušivý vliv na výsledky analýzy. Měly by být využity pouze takové znaky, které dostatečně rozlišují mezi objekty.

2. Analýza hlavních komponent (PCA)

2.1 Zaměření metody PCA

Metoda hlavních komponent (PCA) je jedna z nejstarších a nejvíce používaných metod vícerozměrné analýzy. Poprvé byla zavedena Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933. Cílem analýzy hlavních komponent je především zjednodušení popisu skupiny vzájemně lineárně závislých čili korelovaných znaků čili rozklad zdrojové matice dat do *matice strukturní* a do *matice šumové.* V analýze hlavních komponent nejsou znaky děleny na závisle a nezávisle proměnné jako v regresi. Techniku lze popsat jako metodu lineární transformace původních znaků na nové, nekorelované proměnné, nazvané *hlavní komponenty*. Každá hlavní komponenta představuje lineární kombinací původních znaků. Základní charakteristikou každé hlavní komponenty je její míra variability čili *rozptyl.* Hlavní komponenty jsou seřazeny dle důležitosti, tj. dle klesajícího rozptylu, od největšího k nejmenšímu. Většina informace o variabilitě původních dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě. Platí pravidlo, že má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.

Standardním využitím PCA je snížení dimenze úlohy čili redukce počtu znaků bez velké ztráty informace, a to užitím pouze prvních několika hlavních komponent. Toto snížení dimenze úlohy se netýká počtu původních znaků. Je výhodné především pro možnost zobrazení vícerozměrných dat. Předpokládá se, že nevyužité hlavní komponenty obsahují malé množství informace, protože jejich rozptyl je příliš malý. Tato metoda je atraktivní především z důvodu, že hlavní komponenty jsou nekorelované. Namísto vyšetřování velkého počtu původních znaků s komplexními vnitřními vazbami analyzuje uživatel pouze malý počet nekorelovaných hlavních komponent.

2.2 Podstata metody PCA

● **Zdrojová matice dat** $\boldsymbol{X}$ ($n \times m$): Zdrojová matice dat $\boldsymbol{X}$ ($n \times m$) obsahuje n objektů a m znaků. *Objekty* jsou pozorování, vzorky, experimenty, měření, pacienti, rostliny, atd., zatímco *znaky* či proměnné jsou druhy signálu měření, měřená veličina, vlastnosti (sladký, kyselý, hořký, slaný, cholerický, atd.), barva, a pod. Důležitá je zde skutečnost, že každý *znak* je znám pro všech n objektů. Správná skladba zdrojové matice $\boldsymbol{X}$ čili volba, které znaky použít a které objekty zařadit je delikátní

úkol silně odvislý od charakteru každé úlohy. Velikou výhodou metody PCA je použití jakéhokoliv počtu proměnných ve zdrojové matici $\boldsymbol{X}$ k vícerozměrné charakterizaci. Cílem každé vícerozměrné analýzy je zpracovat data tak, aby se zřetelně indikoval model a tak odkryl skrytý jev. Myšlenka sledování rozptylu je velice důležitá, protože je vlastně základním předpokladem vícerozměrné analýzy dat, že "nalezené směry maximálního rozptylu" jsou více či méně spjaty s těmito skrytými jevy. Základním cílem PCA je transformace původních znaků x_i, 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 znaků* a jsou vzájemně nekorelované. Latentní proměnné jsou 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 znaků, která pokrývá největší rozptyl mezi všemi ostatními lineárními kombinacemi.

Rozdíl mezi souřadnicemi objektů v původních znacích a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá *mírou těsnosti proložení modelu PCA* nebo také *chybou modelu PCA*. Na obr. 1 je tato situace schematicky znázorněna spolu s použitým označením.

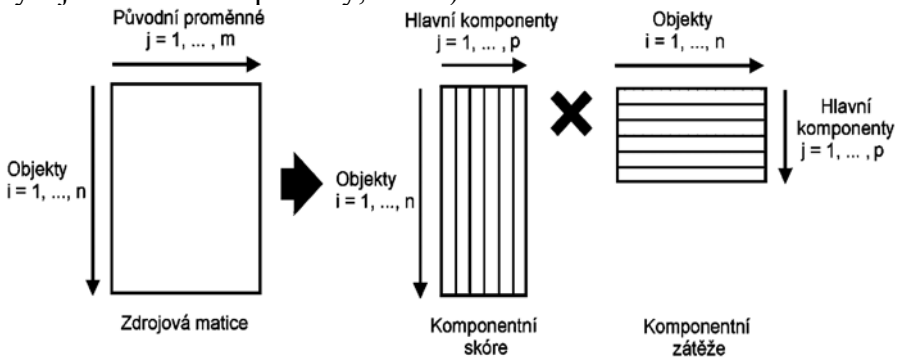
I při velkém počtu původních znaků m může být k velmi malé, běžně 2 až 5. Volba počtu užitých komponent k vede k *modelu hlavních komponent PCA*. Vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních znaků x_j, j = 1, ..., m, k hlavním komponentám y_j, j = 1, ..., k, tvoří dominantní součásti analýzy modelu hlavních komponent PCA.

Z obr. 1 je zřejmé, že zdrojová centrovaná matice $\boldsymbol{X_C}$ se rozkládá na matici komponentních skóre $\boldsymbol{T}$ rozměru $n \times k$ a matici komponentních zátěží $\boldsymbol{P}^{\mathrm{T}}$ rozměru $k \times m$.

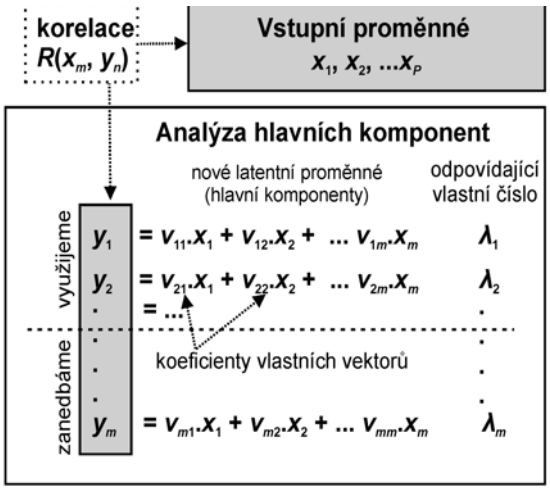
Model PCA odpovídá *aproximaci* zdrojové matice dat $\boldsymbol{X}$, který užijeme místo původní zdrojové matice dat $\boldsymbol{X}$. Aproximace má řadu výhod v interpretaci dat. Nejde zde pouze o změnu systému souřadnic, ale především o nalezení a vypuštění šumu. PCA má proto dvojí cíl: transformace do nového systému os a snížení rozměrnosti úlohy užitím několika prvních hlavních komponent, které vystihují strukturu v datech. Problémem zůstává, kolik hlavních komponent je nutno použít.

● **Maximální počet hlavních komponent:** Existuje horní mez počtu hlavních komponent, které mohou být odvozeny ze zdrojové matice dat $\boldsymbol{X}$. Největší počet hlavních komponent se buď rovná číslu n - 1 nebo m v závislosti na tom, které z těchto dvou čísel je menší. Je-li $\boldsymbol{X}$ složena například n = 40 spekter měřených při m = 2000 vlnových délek, bude maximální počet hlavních komponent 39. Počet efektivních hlavních komponent se rovná *hodnosti zdrojové matice* $\boldsymbol{X}$.

● $\boldsymbol{X}$ **= Struktura + šum:** Všechny hlavní komponenty jsou vzájemně ortogonální a souvisí postupně se snižující hodnotu rozptylu objektů. Poslední hlavní komponenta souvisí s nejmenším rozptylem, tj. *stochastickým rozptylem.* Vyšší hlavní komponenty (obvykle vyšší než 3) se často týkají pouze šumu. Dochází k rozdělení původní zdrojové matice $\boldsymbol{X}$ na *část struktury* (první hlavní komponenty) a *část šumu* (ostatní zbývající hlavní komponenty, obr. 2).



Obr. 1 Schéma maticových výpočtů v PCA.



Obr. 2 Schéma výpočtu hlavních komponent v PCA.

Model hlavních komponent má pak tvar

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} = \textit{struktura dat} + \textit{šum}$$

Zdrojovou matici dat $\boldsymbol{X}$ je třeba nejprve centrovat $x_{ik} = x_{ik} - \bar{x}_k$. V metodě hlavních komponent pracujeme za předpokladu, že zdrojová matice $\boldsymbol{X}$ je rozložena na součin matic $\boldsymbol{TP}^{\mathrm{T}}$ a na matici reziduí $\boldsymbol{E}$, kde matice $\boldsymbol{T}$ je matice komponentního skóre a $\boldsymbol{P}^{\mathrm{T}}$ je transponovaná matice komponentních vah, $\boldsymbol{E}$ je matice reziduí. Cílem metody hlavních komponent je místo $\boldsymbol{X}$ využívat dále jenom součin $\boldsymbol{TP}^{\mathrm{T}}$ a tak oddělit šum od struktury dat $\boldsymbol{TP}^{\mathrm{T}}$. Matici $\boldsymbol{E}$ se nazývá *matice reziduí*, která není objasněna modelem hlavních komponent PCA. Matice $\boldsymbol{E}$ souvisí s "těsností proložení" a ukazuje, jak dobře jsou objekty proloženy modelem hlavních komponent.

● **Komponentní váhy, zátěže - vztah mezi *X* a *PC*:** Hlavní komponenty *PC* jsou vhodně škálované vektory v prostoru znaků. Kterákoliv hlavní komponenta představuje *lineární kombinaci všech m* vektorů v prostoru znaků, tj. jednotkové vektory podél každé osy původního znaku v *m* rozměrném prostoru. Lineární kombinace v každé hlavní komponentě bude obsahovat *m* koeficientů p_{ka}, kde *k* je index *m*-tého znaku a *a* je index směru hlavní komponenty. Například p_{23} znamená koeficient pro druhý znak v lineární kombinaci, která vytvoří *PC*3. Tyto koeficienty se nazývají *komponentní váhy*. Váhy pro všechny hlavní komponenty tvoří matici $\boldsymbol{P}$. Tato matice je vlastně *transformační maticí*, která převádí původní znaky zdrojové matice $\boldsymbol{X}$ do nových latentních proměnných, tj. hlavních komponent. *Vektory vah* čili sloupce v matici $\boldsymbol{P}$ jsou ortogonální.

Váhy informují o vztahu mezi původními *m* znaky a hlavními komponentami. Tvoří tak most mezi prostorem původních znaků a prostorem hlavních komponent. Váhy souvisejí se směrovými kosiny každé hlavní komponenty vzhledem k systému os původních znaků.

● **Komponentní skóre - souřadnice objektů v prostoru hlavních komponent:**
Souřadnice každého objektu na osách hlavních komponent nazýváme *skóre.* Projekce *i*-tého objektu na první hlavní komponentu *PC*1 značí skóre t_{i1}. Projekce téhož objektu na druhou hlavní komponentu *PC*2 značí skóre t_{i2}, ...atd. Každý objekt má svůj soubor komponentních skóre t_{i1}, t_{i2}, ..., $t_{ip\text{-}1}$. Hodnot skóre je stejný počet jako hlavních komponent.

Jedním z nejdůležitějších grafů metody hlavních komponent je *graf komponentního skóre*. Jde o zobrazení dvou skórových vektorů vynesených v systému kartézských os jeden proti druhému. Skórové vektory zde představují znázornění objektů na hlavních komponentách. Vynesení skórových vektorů odpovídá vynesení objektů v prostoru hlavních komponent. Nejužívanějším grafem ve vícerozměrné analýze dat je vektor skóre *PC*1 proti skóre *PC*2. Je snadno k pochopení, protože jde o dva směry, podél kterých shluk objektů vykazuje největší (*PC*1) a druhé největší (*PC*2) rozptýlení.

Graf komponentního skóre t_1 proti t_2 se obvykle vyšetřuje jako první. Do tohoto grafu lze vynášet libovolný pár hlavních komponent. Otázkou však zůstává, které hlavní komponenty jsou ty

nejdůležitější. Je-li počet znaků m menší než počet objektů n, lze vynést $m(m - 1)/2$ možných 2D-grafů komponentního skóre. Počet možných grafů se tak stává nekontrolovatelným a není možné vyšetřovat všechny grafy.

2.3 Grafické diagnostiky metody hlavních komponent

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

(a) **Cattelův indexový graf úpatí vlastních čísel** (Scree Plot) je vlastně sloupcový diagram vlastních čísel nebo reziduálního rozptylu proti stoupající hodnotě indexu, pořadového čísla A. Zobrazuje relativní velikost jednotlivých vlastních čísel. Řada autorů ho s oblibou využívá k určení počtu A "užitečných" hlavních komponent. Cattel vysvětluje scree jako úpatí mořského útesu čili zlomové místo mezi kolmou stěnou a vodorovným dnem. Vybrané "užitečné" hlavní komponenty (nebo také faktory) pak tvoří kolmou stěnu útesu a "neužitečné" hlavní komponenty (nebo faktory) představují vodorovné mořské dno. Užitečné komponenty jsou tak odděleny zřetelným zlomovým místem a souřadnice x tohoto zlomu je hledaná hodnota indexu. Jiným, hrubším kritériem je pravidlo, podle kterého využíváme ty hlavní komponenty, jejichž vlastní číslo je větší než jedna. Pravidlo vychází z myšlenky, že není třeba uvažovat komponenty, jejichž rozptyl je menší než jednotkový rozptyl každého normovaného znaku. Graf úpatí se však jeví objektivnějším a praktičtějším.

(b) **Graf komponentních vah, zátěží** (Plot Components Weights) zobrazí komponentní váhy pro první dvě hlavní komponenty. V tomto grafu se porovnávají vzdálenosti mezi znaky. Krátká vzdálenost mezi dvěma znaky znamená silnou korelaci. Lze nalézt i shluk podobných znaků, jež spolu korelují. Tento graf můžeme považovat za most mezi znaky a hlavními komponentami, protože ukazuje, jakou měrou přispívají jednotlivé znaky 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 znaky x_j, $j = 1,..., m$, přispívají do první hlavní komponenty y_1 nebo do druhé hlavní komponenty y_2. Některé znaky x_j přispívají kladnou vahou, některé zápornou. Bývá zajímavé sledovat kovarianci znaků x_j v prostorovém 3D grafu komponentních vah y_1, y_2 a y_3. Jsou-li znaky 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 znaků x_j, $j = 1,..., m$*: znaky 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. Znaky s malou důležitostí leží blízko počátku. Když určíme *důležitost znaků*, určíme tím také proměnlivost znaků: jestliže například y_1 objasňuje 70 % proměnlivosti a y_2 jenom 5 % (přečteno z indexového grafu úpatí vlastních čísel), jsou znaky x_j, $j= 1,..., m$, s vysokou vahou v y_1 tím pádem mnohem důležitější než znaky x_j s vysokou vahou v y_2. Znaky s úhlem 0° mezi průvodiči jsou zcela pozitivně korelované, znaky s úhlem 90° jsou zcela nekorelované zatímco znaky s úhlem 180° jsou negativně korelované.
2. *Korelace a kovariance:* znaky x_j, $j= 1,..., m$, jsou blízko sebe, anebo znaky x_j s malým úhlem mezi svými průvodiči znaků a na stejné straně vůči počátku mají vysokou kladnou kovarianci a vysokou kladnou korelaci. Naopak, znaky x_j daleko od sebe, anebo s velikým úhlem mezi průvodiči znaků, jsou negativně korelovány.

(c) **Rozptylový diagram komponentního skóre** (Scatterplot) zobrazuje *komponentní skóre* čili hodnoty obyčejně prvních dvou hlavních komponent u všech objektů. Lze snadno nalézt shluk vzájemně podobných objektů a dále objekty odlehlé a silně odlišné od ostatních. Diagram komponentního skóre však může být prostorový ve třech hlavních komponentách a v rovinném grafu se pak sleduje pouze jeho průmět. Tento diagram se užívá k identifikaci odlehlých objektů, identifikaci trendů, tříd, shluků objektů, k objasnění podobnosti objektů atd. Je často nemožné analyzovat všechny diagramy, protože jich je velmi mnoho: pro $m = 10$ znaků 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 objasňuje největší míru 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 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. To platí jen v případech, kdy se nejedná o zdánlivou nehomogenitu danou sešikmením dat a odstranitelnou transformací znaků.
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 jejich fyzikální či biologickou podobnost.
7. *Vysvětlení místa objektu.* Umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami znaků ve dvojném grafu a pomocí znaků pak i vysvětleno.

2.4 Diagnostika metody hlavních komponent

Maticový graf rozptylových diagramů znaků slouží k získání počáteční informace o datech. Odhalí, zda data potřebují škálování. Při prvním seznámení s daty se v rámci exploratorní analýzy použije standardní metoda hlavních komponent PCA. Data je obvykle potřeba škálovat nebo alespoň centrovat. V tomto stadiu 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. Po redukci dat na několik podvýběrů, kdy jsou shluky modelovány odděleně, se znovu aplikuje metoda hlavních komponent PCA na jednotlivé dílčí výběry, kdy postupně provádíme:

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

2. *Výpočet vlastních vektorů* – vedle číselných hodnot se užívá i názorný čárový diagram hodnot vlastních vektorů, který přehledně informuje o relativním zastoupení původních znaků x_j , $j = 1, ..., m$, v hlavních komponentách.

3. *Výpočet komponentních vah* – matice párových korelačních koeficientů obsahující korelace původních znaků s hlavními komponentami. Čárový diagram názorně vysvětluje korelační strukturu mezi oběma druhy znaků. Uživatel nyní vybere pouze prvních *k* hlavních komponent a vytvoří tak model PCA.

4. *Vyšetření grafu komponentních vah.*

5. *Vyšetření rozptylového diagramu komponentního skóre.*

6. *Vyšetření dvojného grafu.*

7. *Vyšetření reziduí* – rezidua objektů a rezidua proměnných by měla prokazovat dostatečnou těsnost proložení. Není-li tomu tak, je třeba se vrátit k předúpravě dat a celý výpočet PCA opakovat.

8. *Určení významných znaků* – v některých případech je výhodné vyhledávat také významné znaky, protože klasická metoda PCA umožňuje sice redukci počtu hlavních komponent, ale každá komponenta zůstává stále kombinací všech původních znaků. Nalezení podmnožiny znaků, které

obsahují téměř všechny informace jako původní znaky, je poměrně zajímavé v řadě praktických aplikací.

3 Vlastnosti metody shlukování

3.1 Míry podobnosti

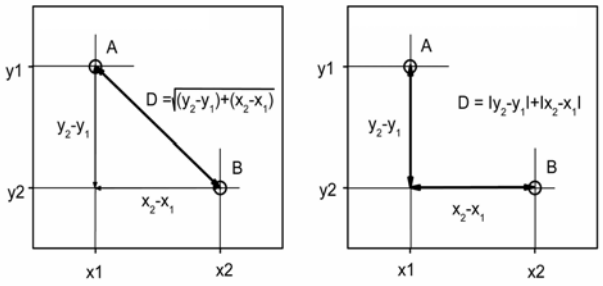
Podobnost mezi objekty je užita jako kritérium tvorby shluků objektů. Nejdříve se stanovují znaky, určující podobnost, které se dále kombinují do podobnostních měr. Tímto způsobem pak může být objekt porovnán s jiným objektem. Analýza shluků vytváří shluky podobných objektů. Podobnost může být měřena rozličnými způsoby, které se dají obyčejně zařadit do jedné ze tří základních skupin:

(1) *Korelační míry:* Základní mírou podobnosti dvou objektů či znaků x_i a x_j, vyjádřených v kardinální škále může být *Pearsonův párový korelační koeficient r*. Objekty jsou si tím podobnější, čím je jejich párový korelační koeficient větší a bližší jedničce. V případě ordinální škály (pořadová čísla) je analogickou mírou podobností *Spearmanův korelační koeficient.* Vysoká korelace prozrazuje vysokou "podobnost" a nízká korelace pak "nepodobnost" profilů.

(2) *Míry vzdálenosti:* Představují nejčastěji užívané míry, založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Nejčastější vzdálenostní mírou je *Eukleidovská vzdálenost* zvaná také *geometrická metrika*, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Platí, že vzdálenost

$$d_E(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2} \qquad (1)$$

představuje standardní typ vzdálenosti.



Obr. 3 Míry vzdálenosti: Euklidovská (vlevo), manhattanská (vpravo).

Vedle Eukleidovské vzdálenosti se užívá také *čtverec Eukleidovské vzdálenosti,* který tvoří základ Wardovy metody shlukování. Často je užívaná *Manhattanská vzdálenost* zvaná také *vzdálenost městských bloků* nebo *Hammingova metrika*, definovaná vztahem

$$d_H(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sum_{j=1}^{m} |x_{kj} - x_{lj}| \qquad (2)$$

Před užitím této vzdálenosti se musíme ujistit, že znaky spolu nekorelují. Když tato podmínka není splněna, shluky jsou nesprávné. Další mírou je zobecněná *Minkovského metrika,* pro kterou platí

$$d_M(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt[z]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^z} \qquad (3)$$

kde pro $z = 1$ jde o Hammingovu metriku a pro $z = 2$ o Eukleidovu. Čím je z větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty. V některých případech se používá také *tětivová vzdálenost* (anglicky chord distance), definovaná vztahem

$$d_{CH}(x_k, x_l) = \sqrt{2\left[1 - \frac{\sum_{j=1}^{m} x_{kj}\, x_{lj}}{\sum_{j=1}^{m} x^2_{kj} \sum_{j=1}^{m} x^2_{lj}}\right]} \qquad (4)$$

V případě třech znaků je tětivová vzdálenost přímou vzdáleností dvou bodů na povrchu koule s jednotkovým poloměrem a počátkem v těžišti.

Problém všech vzdálenostních měr vzniká při použití nestandardizovaných dat, které mohou způsobit rozdíly mezi shluky, díky často veliké odlišnosti jednotek měření. Shluky různých vzdálenostních měr se budou lišit, největší rozptýlení mezi shluky bude u čtverce Eukleidovské vzdálenosti. Pořadí podobností se významně změní se změnou měřítka nebo změnou jednotek jednoho ze znaků.

Všechny dosud uvedené metriky neuvažují závislost mezi znaky. Zahrneme-li do vztahu pro vzdálenost také vazby mezi znaky, vyjádřené kovarianční maticí $\boldsymbol{C}$, dostaneme novou statistickou míru, zvanou *Mahalanobisova metrika*

$$d_{Ma}(x_k, x_l) = \sqrt{(x_k - x_l)^{\mathrm{T}}\, C^{-1}\, (x_k - x_l)} \qquad (5)$$

Jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být orthogonální. Vysoce korelovaný výběr znaků může skrytě převážit celý soubor znaků shlukování.

(3) *Míry asociace:* Míry asociace podobnosti se používají k porovnání objektů, pokud jsou jejich znaky nemetrického charakteru (například binární proměnné). Uveďme příklad, kdy respondent odpověděl na řadu otázek odpovědí *ano* nebo *ne*.

3.2 Standardizace dat

Před vlastní shlukovou analýzou je třeba řešit otázku, zda je třeba data standardizovat. Musí se respektovat fakt, že většina měr vzdálenosti je velmi citlivá na měřítka (stupnice), vedoucí k různé numerické velikosti znaků. Obecně platí pravidlo, že znaky s větší mírou proměnlivosti čili větší směrodatnou odchylkou mají větší vliv na míru podobnosti.

(1) *Standardizování znaků*: Nejužívanější formou standardizace je *normalizace každého znaku* do svého *Z-skóre*, tj. odečtením průměru a dělením směrodatnou odchylkou. Tato standardizace je známa pod názvem *normovací Z-funkce*.

(2) *Standardizace objektů:* Když chceme identifikovat shluky dle vzdálenosti pak standardizace není vhodná. Standardizace objektů nebo-li řádková standardizace může být však efektní ve speciálních případech.

4 Způsoby shlukování

Shluk (cluster) je skupina objektů, jejichž vzdálenost je menší než vzdálenost s objekty do shluku nepatřícími. Podle způsobu shlukování se postupy dělí na *hierarchické* a *nehierarchické*. Hierarchické se dělí dále na *aglomerační* a *divizní.*

4.1 Hierarchické shlukovací postupy

Postupy jsou založeny na hierarchickém uspořádání objektů a jejich shluků. Graficky se hierarchicky uspořádané shluky zobrazují formou *vývojového stromu* nebo *dendrogramu*. U *aglomeračního 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 objekt. 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,

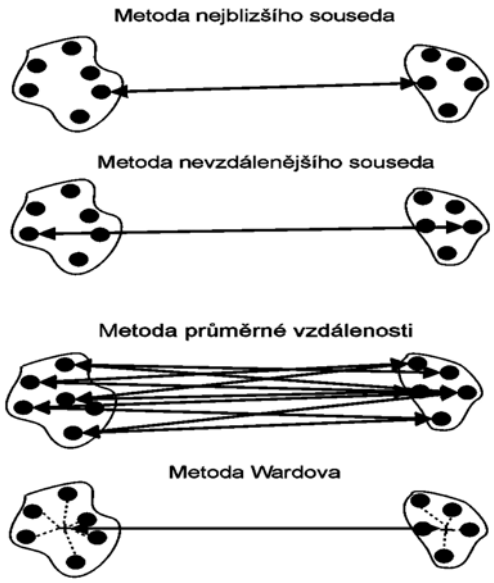
až skončíme ve stadiu jednotlivých objektů.
Mezi metody metriky shlukování patří:

(1) *Metoda nejbližšího souseda:* Postup je postaven na minimální vzdálenosti. Naleznou se dva objekty, oddělené nejkratší vzdáleností a umístí se do shluku. Další shluk je vytvořen přidáním třetího nejbližšího objektu. Proces se opakuje až jsou všechny objekty v jednom společném shluku. Vzdálenost mezi dvěma shluky je definována jako nejkratší vzdálenost libovolného bodu v prvním shluku vůči libovolnému bodu v druhém. Dva shluky jsou propojeny v libovolném stádiu nejkratší spojkou.

(2) *Metoda nejvzdálenějšího souseda:* Kritérium je postaveno nikoliv na minimální ale na maximální vzdálenosti. Nejdelší vzdálenost mezi objekty v každém shluku představuje nejmenší kouli, která obklopuje všechny objekty v obou shlucích. Metoda se také nazývá *metodou úplného propojení,* protože všechny objekty ve shluku jsou propojeny každý s každým při maximální vzdálenosti čili minimální podobnosti.

(3) *Metoda průměrné vzdálenosti:* Kritériem vzniku shluků je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

(4) *Wardova metoda:* Principem není optimalizace vzdáleností mezi shluky ale minimalizace heterogenity shluků podle kritéria minima přírustku vnitroskupinového součtu čtverců odchylek objektů od těžiště shluků. V každém kroku se pro všechny dvojice odchylek spočítá přírůstek součtu čtverců odchylek, vzniklý jejich sloučením a pak se spojí ty shluky, kterým odpovídá minimální hodnota tohoto přírůstku.



Obr. 4 Nejčastěji užívané metriky shlukování.

(5) *Metoda těžiště:* Jde o vzdálenost dvou těžišť shluků, vyjádřených Eukleidovskou vzdáleností nebo čtvercem Eukleidovské vzdálenosti. Těžiště shluku má souřadnice odpovídající průměrným hodnotám objektů pro jednotlivé znaky. Po každém kroku shlukování se počítá nové těžiště. Poloha těžiště shluku poněkud migruje tak jak se připojují nové objekty a vznikají větší shluky.

(6) *Metoda mediánová*: Jde o jisté vylepšení metody těžiště, neboť se snaží odstranit rozdílné významnosti, které metoda těžiště dává různě velkým shlukům.

4.2 Nehierarchické shlukovací postupy

U *metody zárodečných bodů* (Seeded) uživatel na základě svých věcných znalostí určí, které objekty mají tvořit zárodky nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich Eukleidovské vzdálenosti od těchto typických objektů. Existuje několik postupů zadávání zárodků shluku a zařazování objektů do shluku. Těmto metodám se říká *K-means shlukování.*

4.3 Dendrogramy hierarchického shlukování

Analýzou shluků je možné hodnotit jednak podobnost objektů, analyzovanou pomocí *dendrogramu*

objektů, a jednak podobnost znaků analyzovanou pomocí *dendrogramu znaků.*

Dendrogram shluků (vývojový strom) se konstruuje pouze v případě, kdy je k dispozici matice původních znaků. Dendrogram podobnosti znaků ukazuje rozlišení znaků ve shlucích. Jeho interpretace je snadná: znaky blízko sebe jsou propojeny spojovací úsečkou hodně nízko, mají malou vzdálenost čili značnou vzájemnou podobnost. Znaky propojené hodně vysoko mají malou podobnost a mezi sebou vykazují velkou vzdálenost.

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

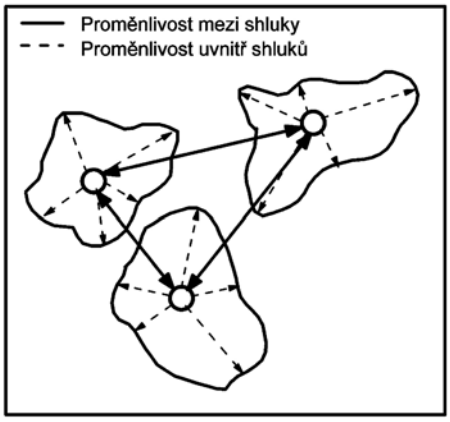
Dendrogram podobnosti znaků odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) znaků, které jsou si velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku si jsou značně podobné a jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti či znaky není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

Míra věrohodnosti: Dendrogram lze sestrojit celou řadou technik. Prvním kritériem těsnosti proložení při volbě "nejlepšího dendrogramu", jež nejlépe odpovídá struktuře objektů a znaků 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 struktury dat spíše než stupeň podobnosti. Kritérium delta je definováno vztahem

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d_{jk}^*|^{1/A}}{\sum_{j<k}^{N} (d_{jk}^*)^{1/A}} \right]^A \qquad (6)$$

kde $A = 0.5$ nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^*_{ij} je vzdálenost získaná z dendrogramu. Je žádoucí, aby hodnoty *delta* byly blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.



Obr. 5 Porovnání vzdáleností mezi shluky a uvnitř shluku.

5 Postup obecné analýzy shluků

Analýza shluků je vždy silnou analytickou pomůckou k účelům zjednodušení, průzkumu a potvrzení struktury.

1. krok: Cíle analýzy shluků: Primárním cílem je rozdělení souboru objektů do dvou nebo více skupin, tříd či shluků. Sledujeme tři cíle:

(a) *Popis systematiky:* Tradičním využitím jsou průzkumové cíle a popis systematiky - *taxonomie*, tj. empirická klasifikace objektů. Analýzou shluků se dospěje k určitým shlukům objektů, které jsou pak porovnány s jejich teoreticky odvozenou typologií.

(b) *Zjednodušení dat:* Při hledání taxonomie poskytuje analýza shluků zjednodušený pohled na objekty. Zatímco faktorová analýza se snaží nalézt strukturu znaků, analýza shluků činí totéž ale pro objekty. Na objekty se pak už nehledí jako na jeden společný soubor ale na oddělené shluky objektů, rozčleněné dle jejich vlastností.

(c) *Identifikace vztahu:* Po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty, což by bylo mezi samotnými objekty daleko obtížnější. Shluky mohou být předmětem dalšího kvalitativního uvažování.

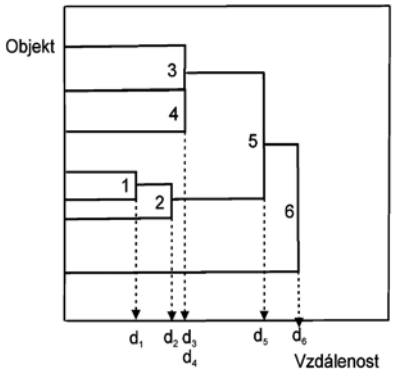
2. krok: Formulace úlohy analýzy shluků: S vybranými znaky budeme shlukovat vyšetřované objekty. Nejprve však je třeba odpovědět na tři otázky: (1) Mohou být v datech nějaké odlehlé objekty, které mohou být posléze odstraněny? (2) Jak vyjádříme podobnost objektů? (3) Měla by být data před analýzou shluků standardizována?

3. krok: Předpoklady analýzy shluků: Analýza shluků objektů není charakteru statistického testování. Požadavky normality, linearity, homoskedasticity, které jsou tolik důležité v ostatních vícerozměrných technikách nemají zde význam. Přesto existují dva kritické předpoklady: reprezentativnost a vliv multikolinearity.

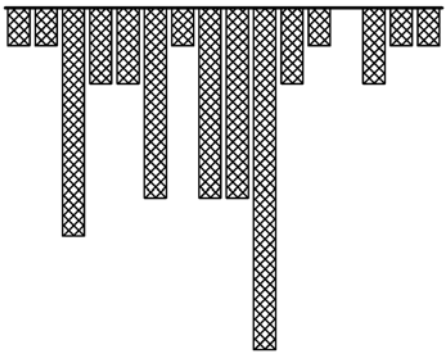
4. krok: Výstavba shluků a celková těsnost proložení. Za vhodné rozlišovací kritérium je možné použít maximalizaci rozdílů mezi shluky, a toto porovnávat vůči proměnlivosti uvnitř shluků. Testování poměru roztylu mezi jednotlivými shluky vůči průměru rozptylu uvnitř shluku lze porovnat s *F* kritériem v analýze rozptylu.

Hierarchické shlukování: Tyto algoritmy se týkají konstrukce stromovité struktury shluků, dendrogramu. Existují v zásadě dva typy hierarchického shlukování, *aglomerační* a *divizní*. Grafickým zobrazením růstového stromu je diagram, také zvaný *dendrogram*. Jinou grafickou metodou je *vertikální rampouchovitý diagram*.

Když shlukovací proces probíhá v opačném směru než aglomerační, označuje se jako *metoda divizní.* Postup začíná z jednoho velkého shluku, ve kterém jsou všechny objekty. V následujících krocích jsou nepodobné objekty odloučeny ze společného shluku a vzniká tím menší shluk. Proces probíhá tak dlouho, až je ve shluku jediný objekt. Pět nejpoužívanějších aglomeračních algoritmů výstavby shluků jsou metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, metoda průměrová, Wardova metoda, a metoda těžiště.



Obr. 6 Postupná výstavba dendrogramu.



Obr. 7 Rampouchovity diagram shluků.

Nehierarchické shlukování: Na rozdíl od předešlých hierarchických metod se tyto metody netýkají výstavby stromu, dendrogramu. Místo toho se přidělují objekty do předem známého počtu shluků. Prvním krokem je zadání *zárodečného shluku* jako počátečního středu shluku a všechny objekty nacházející se uvnitř předspecifikované vzdálenosti pak budou do výsledného shluku zařazeny. Pak je zvolen *zárodek* jiného shluku a zařazování do shluku pokračuje až jsou všechny objekty zařazeny.

Přednost hierachickým nebo nehierarchickým metodám? Výzkumný problém musí jednoznačně vést na jednu nebo na druhou metodu. Metodu zvolíme dle obsahu úlohy.

(1) *Důvody pro a proti hierarchickým metodám:* V minulosti byly hierarchické shlukové

techniky populárnější. Wardova a průměrová metoda byly považovány za nejlepší techniky. Hierarchické metody mají výhodu, že jsou rychlé a spotřebují méně strojového času.

(2) *Užívání nehierarchických metod:* Nehierachické metody se v poslední doběvyužívají stále více. Jejich kvalita závisí na schopnosti uživatele, jeho praktických zkušenostech a objektivní teorii jak si vybrat zárodkové body.

(3) *Kombinace obou metod:* Jiným přístupem je užití *obou* metod, hierarchické a nehierarchické, abychom využili výhod obou. Nejprve hierarchickou metodou určíme počet shluků, profily shlukovaných center a identifikují se zřetelné odlehlé body.

Počet vytvářených shluků: Snad nejvíce matoucí otázkou v analýze shluků je dosažení konečného počtu shluků, známého také pod názvem *terminační kritérium*. Neexistuje žádný objektivní způsob určení tohoto kritéria. Jedno z terminačních kritérií se týká relativně jednoduchého vyšetření měr podobnosti mezi shluky v každém kroku, když totiž míra podobnosti překročí předdefinovanou velikost nebo když následné hodnoty se skokově změní. Je vhodné postupovat tak, že se určí rozličný počet shluků např. 2, 3 a 4 a na základě úvah o alternativním řešení, praktickém úsudku a teoretických základech úlohy samé se rozhodne. Když se objeví jednoobjektový shluk nebo shluk o poměrně malé velikosti, uživatel musí rozhodnout, zda tento představuje strukturního člena vzorku nebo zda ho lze označit jako nedostatečně reprezentativní pro soubor dat.

5. krok: Interpretace shluků: Interpretace shluků se týká vyšetření každého shluku v pojmech shlukových znaků a především pojmenování shluků, které vystihuje podstatu a povahu shluků. Při zahájení interpretace si uvědomíme, že ve shlukové analýze je často užívanou mírou těžiště shluků. Uživatel se *musí* vrátit k původním datům v původních znacích a vyčíslit průměrové profily pro původní data. Vyšetříme proto profily průměrových skóre a označíme popisným nadpisem každý shluk.

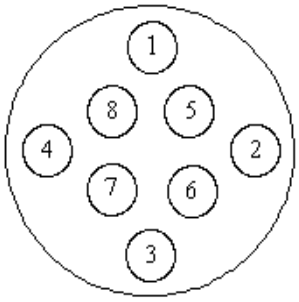
6. krok: Validace a profilování shluků: Existuje poněkud subjektivní charakter analýzy shluků stran hledání optimálního shlukového řešení.

Validace shluků: Validace čili potvrzení shluků zahrnuje pokusy zajistit, že shlukové řešení je reprezentativní v celé obecné populaci, a je proto zobecnitelné i na ostatní objekty a dále je i stabilní v čase.

Profilování shlukového řešení: Profilování se týká popisu vlastností každého shluku, aby se objasnilo jak se vlastnosti liší ve významných dimenzích. Profilová analýza se zaměřuje na popis ne toho, co konkrétně odhalují shluky ale na vlastnosti shluků, na jejímž základě byly identifikovány. Kromě toho je kladen důraz na vlastnosti, ve kterých se shluky vzájemně odlišují a na vlastnosti, které mohou predikovat účast v dotyčném shluku.

Vzorová úloha 1. *Posouzení chemické homogenity v kruhové tyči CrNi oceli* (H404)

Kruhová tyč chromniklové oceli o rozměrech 50 × 1000 mm, vyrobená v Poldi Kladno, byla rozřezána na 32 zkušebních vzorků o rozměrech 50 × 30 mm. Každý vzorek byl označen pořadovým číslem za současného zachování původní orientace v tyči. K testování homogenity bylo náhodně vybráno 16 vzorků. Metodou optické emisní spektroskopie s jiskrovým buzením byla na každém vzorku v předem určených místech provedena analýza na 8 expozicích. Cílem bylo vybrat k vyhodnocení homogenity materiálu omezený počet znaků a odhalit trendy v chemickém složení. Prvním číslem v matici dat je pořadí vzorku a číslicí po mezeře je pak umístění expozice. Je třeba vyšetřit, kolik latentních proměnných popisuje alespoň 66%ní proměnlivost v datech? Z grafu komponentních vah pak vyšetřit korelaci znaků, důležité znaky ale také redundandní znaky. Kolik odlehlých objektů se nachází v rozptylovém diagramu komponentního skóre a kolik je jich podobných ve shlucích. Komentujte počet shluků v rozptylovém diagramu komponentního skóre a pokuste se nalézt interakci objektů a znaků ve dvojném grafu. Kolik shluků lze nalézt v dendrogramu podobnosti objektů nejlepší shlukovací procedurou? Komentujte vzniklé shluky dle jejich vlastností. Graf ukazuje volbu expozic na vzorku:



○ Data: Výběr *TYC* se týká 16 x 8 vzorků v řádcích. Obsahuje znaky *i* pořadové číslo oceli a C je obsah uhlíku, MN značí obsah manganu, SI značí obsah křemíku, P značí obsah fosforu, S značí obsah síry, CU značí obsah mědi, CR značí obsah chromu, NI značí obsah niklu, AL značí obsah hliníku, MO značí obsah molybdenu, TI značí obsah titanu, B značí obsah boru.

I	C	MN	SI	P	S	CU	CR	NI	AL	MO	TI	B
31	0,0510	0,1531	1,4636	27,1810	28,5060	0,0980	1,7085	0,7280	1,9752	1,8461	2,5478	0,0620
32	0,0510	0,1537	1,4523	27,1220	29,2860	0,0990	1,7089	0,7240	1,9524	1,8466	2,5678	0,0620
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
288	0,0500	0,1531	1,4548	27,7460	29,4650	0,0980	1,7127	0,7283	1,9348	1,8348	2,4939	0

Řešení: 1) EDA -průzkumová analýza dat:

K popisu vícerozměrného náhodného výběru se využívá řada grafických technik mezi které patří především: 1. Rozptylové grafy, které zobrazují všechny dvojice proměnných daného výběru umožňují hledání odlehlých bodů, shluků a míry párové závislosti mezi dvojicemi složek, 2. Symbolové grafy usnadňují detekovat rozdíly mezi obrazci nebo symboly jednoduchým dvourozměrným zobrazením vícerozměrných dat. Z rozptylových grafů nejsou patrné výrazné odlehlé body (snad do určité míry body 51, 52, 55, 191, 192, 193 a 194. Dále jsou z grafů postřehnutelné významné korelace mezi proměnnými Mn-P, Mn-S, Mn-Cu, Mn-Cr, Mn-Mo, Mn-B, P-S, P-Cu, P-Cr, P-Al, P-Mo, P-B, S-Cu, S-B, Cu-Mo, Cu-B, Cr-Si, Cr-Cu, Cr-Mo, Mo-Al, Mo-B, Mo-Ti, Al-Ti a B-Ti. Zcela bez korelace se jeví proměnné C, Ni a Si (pouze z Cr). Na základě vizuálního posouzení symbolových grafů (sluníčka, hvězdičky) lze s velkou mírou pravděpodobnosti konstatovat:

Expozice v místech 1, 2, 3 a 4 vykazují na první pohled vyšší hodnoty obsahů proměnných než v místech 5, 6, 7 a 8 a to především u C, Si, Ni, Mo, Ti a Al u ostatních to není tak znatelné

1) Je pozorovatelná tendence zvyšování obsahů z místa č. 1 do místa č. 4 hlavně u Mn, Si, P, S, Cu, Cr a B a naopak opačný trend u Ni, Mo, Al a Ti

2) Od místa č. 4 do č. 5 dochází k prudkému poklesu v obsazích C, Mn, Si, P, S, Cu, Cr a B a zároveň k růstu u Mo, Ni, Al a Ti

3) Od místa č. 5 se opět obsahy C, Mn, P, S, Cu, Cr a B trvale zvyšují až do místa č. 8 a u Mo, Al a Ti se obsahy posunují k nižším hodnotám

4) Největší rozdíly v obsazích jsou patrné mezi místy č. 4 a č.5 u prvků C, Mn, Si, P, S, Cu a B a dále mezi místy č. 1 a č. 8 u Si, Cr, Mo, Al a Ti

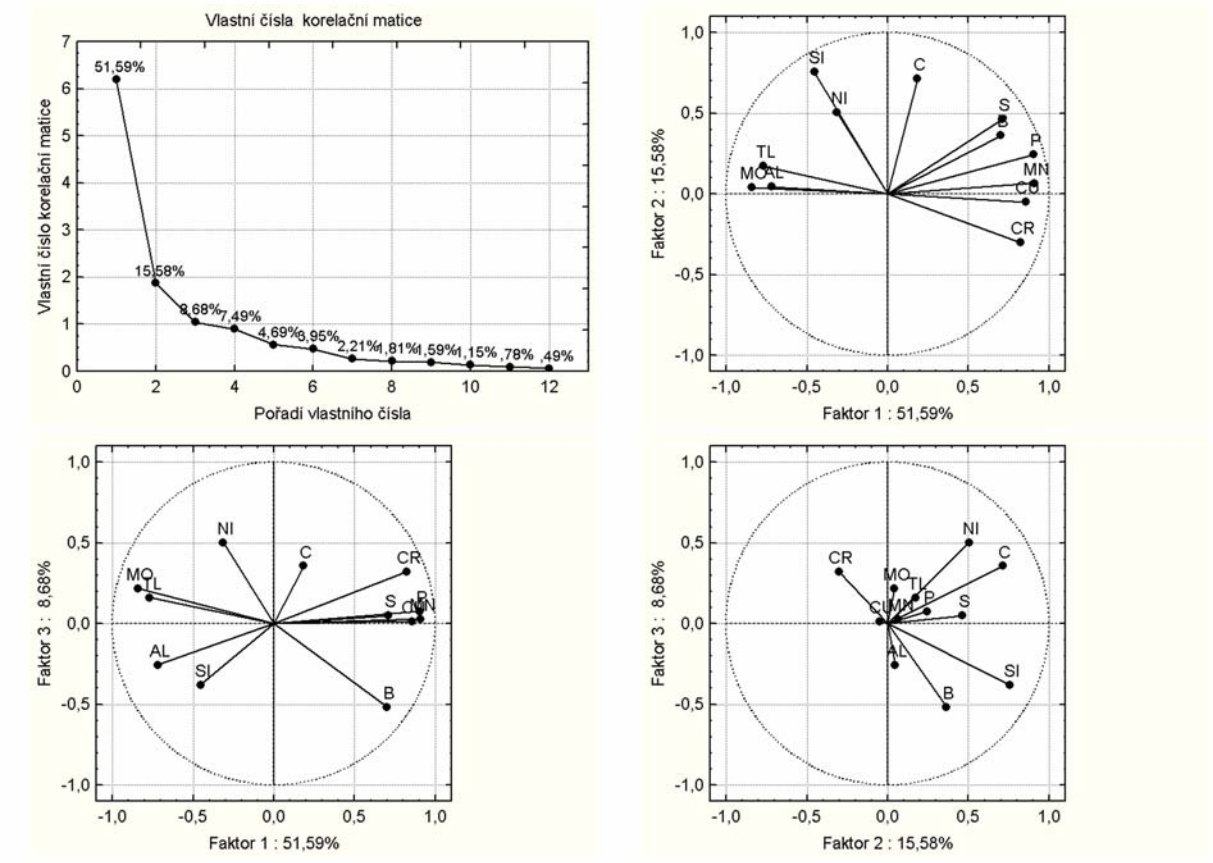
5) Mezi jednotlivými místy není pozorovatelný žádný trend v rozptylu hodnot

6) U vzorku č. 22 jsou symboly odpovídající jednotlivým místům expozic nejpodobnější, což vypovídá o nízkém rozptylu proměnných C, Mn, Si, P, S, Cu, Cr, Ni, Al a Ti, podobný charakter mají vzorky č. 15, 16 a 28

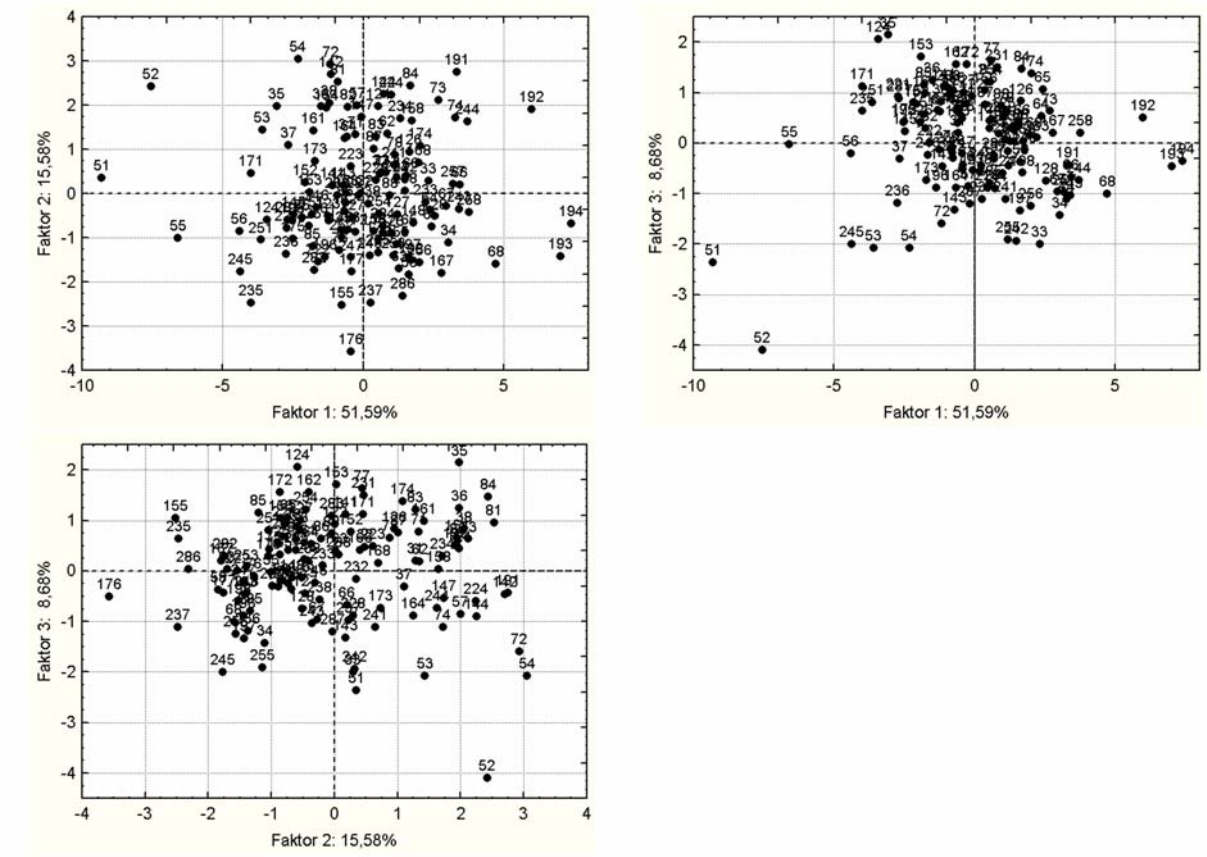
7) Mezi nejhorší vzorek lze zařadit vzorek č. 5 (C, Mn, Cu, Cr, Ni, Mo, Al, B a Ti) a č. 19 (Mn, Si, P, S, Cu, Cr, Ni a B)

Závěr EDA: Na základě symbolových grafů lze jednotlivé proměnné předběžně zařadit do 3 podobnostních skupin: 1. skupina Ni, Al, Mo a Ti, 2. skupina C a Si tvoří určitý přechod mezi skupinou 1 a 3, 3. skupina Mn, P, S, Cu, Cr a B. Po vyloučení sporných bodů na základě rozptylových grafů dospějeme ke stejným závěrům.

2) PCA - metoda hlavních komponent - hledání struktury v proměnných: Metoda PCA je založena na hledání takových lineárních kombinací původních proměnných, které extrahují a popisují co nejvíce jejich celkové variability, a to rozkladem původní-zdrojové matice na součin dvou matic - matice latentních proměnných, hlavních komponent a matice zátěží. Většinou stačí k popisu variability původní matice první dvě komponenty, které vysvětlují 85-95 % variability. Graf komponentních vah zobrazuje komponentní váhy pro první dvě hlavní komponenty u jednotlivých proměnných. Jednotlivé body představující proměnné jsou tím podobnější, čím jsou v prostoru k sobě blíže. Proměnné ve skupinách (Al, Mo, Ti) a (Mn, P, S, Cu, B) je možno nahradit jedinou proměnnou ze středu skupin, která bude reprezentovat chování celé skupiny. Největší přínos pro danou komponentu mají proměnné, které se v grafu nacházejí u souřadnice dané komponenty a co nejdále od nuly. Tak například největší přínos k 1. hlavní komponentě mají Mn, Cu, P a Cr a nejmenší mají Mo, Ti a Al. K 2. hlavní komponentě mají největší přínos Si, Ni, C a nejmenší Cu, Cr a Al. V bodovém diagramu komponentního skóre jsou proti sobě vyneseny první dvě hlavní komponenty, což díky jejich ortogonalitě usnadňuje geometrickou interpretaci jednotlivých objektů po stránce jejich strukturních vazeb. Objekty 51, 52, 55, 191, 192, 193 a 194 jsou strukturně vzdáleny od ostatních objektů i od sebe samých. Jelikož jednotlivé body jsou zhruba rozmístěny v kruhu je možno rozdělení dat považovat za normální. Výpočtem vyšlo, že první dvě hlavní komponenty popisují 68,5 % variability dat, první tři 77,5 % a čtyři 84,9 %. Po zredukování počtu proměnných vyjde procento vysvětlené variability pro první dvě hlavní komponenty 80,5 % a první tři 93 %. Odstraníme Mn – vykazuje podobné vlastnosti jako Cr. K popisu variability zdrojové matice tedy postačují tři hlavní komponenty.



Obr. 8 Cattelův indexový graf vlastních čísel a tři grafy komponentních vah.



Obr. 9 Rozptylové diagramy komponentního skóre v PCA.

Tabulka 1. Vysvětlené variability pro všechny proměnné a pro redukovaný počet

Číslo komponenty	Variabilita %	Kumulativní variabilita %		Variabilita % red. prom.	Kumul. variab. %	Variabilita % red. prom.	Kumul. variab. %
1	52,57	52,57		55,60	55,60	52,97	52,97
2	15,90	68,48		24,89	80,49	27,42	80,39
3	9,00	77,48		12,56	93,05	15,27	95,66
4	7,37	84,86		4,87	97,92	4,34	100
5	4,48	89,33		2,08	100,00		
6	3,66	93,00					
7	2,00	95,00					
8	1,54	96,55					
9	1,37	97,92					
10	1,01	98,93					
11	0,65	99,58					
12	0,42	100,00					

Tabulka 1 ukazuje jak se zmenšuje celková vysvětlená variabilita (i když nepatrně) a zároveň grafy naznačují na nekorelovanost zbývajících proměnných. Příspěvky jednotlivých proměnných do hlavní komponenty:

$$y_1 = -0{,}510*Mo - 0{,}578*Si - 0{,}960*C + 0{,}630*Cr$$

$$y_2 = 0{,}242*Mo - 0{,}272*Si - 0{,}911*C - 0{,}192*Cr$$

Metoda PCA zredukovala počet proměnných na 4 základní, postačující k testování homogenity materiálu. Vzhledem k vlastnostem a použití tohoto materiálu (odolnost proti korozi, žáruvzdornost) byly vybrány prvky:

1) C - nebezpečí tvorby karbidů s Cr, Mn, Si a Mo, což vede k místnímu poklesu obsahu uvedených prvků, zvláště chromu a vzniku mezikrystalové koroze. Čím nižší obsah C tím vyšší kvalita materiálu zvláště pro antikorozní účely.

2) Si - feritotvorný prvek, zvyšuje stabilitu feritu, zvyšuje odolnost proti oxidaci ve vysoce oxidačních prostředích vznikem vrstvy SiO_2 na povrchu materiálu.

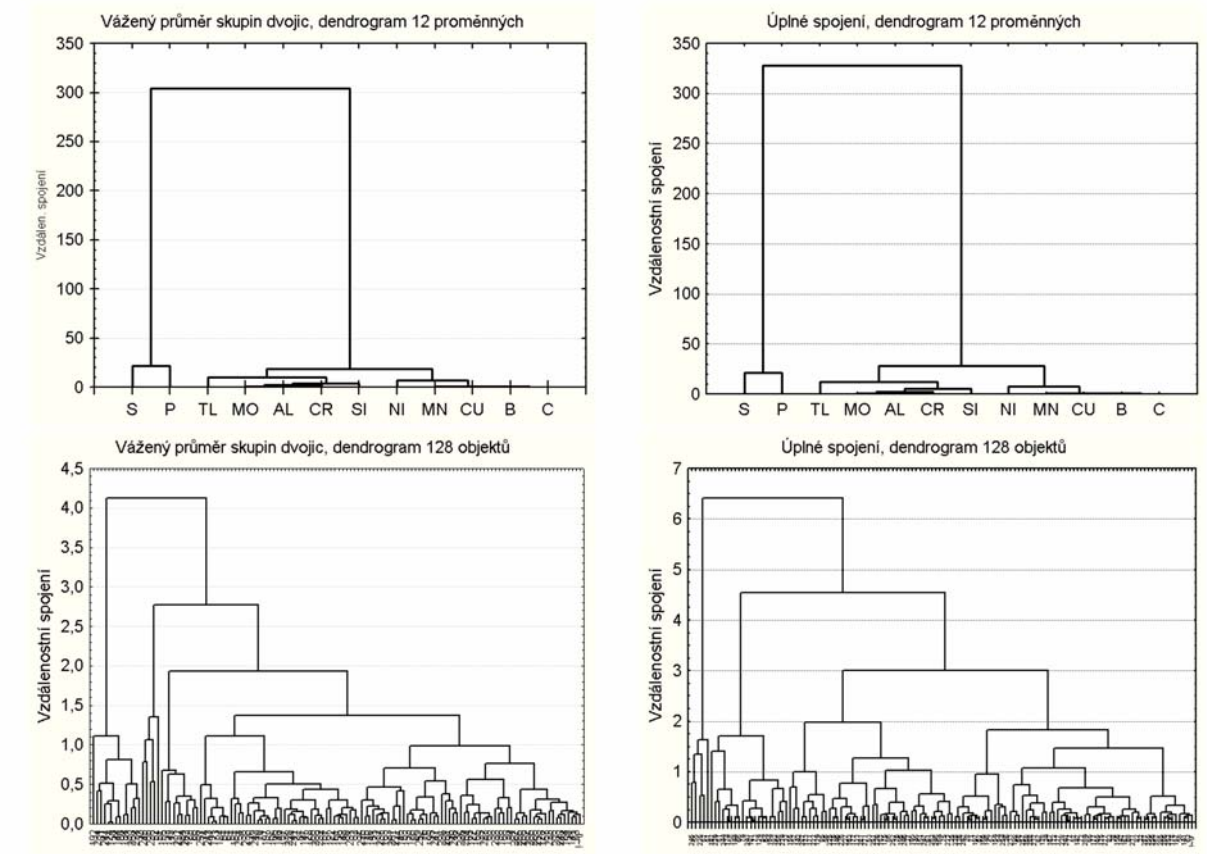
3) Cr - feritotvorný prvek, hlavní účel-odolnost proti korozi, oxidaci a opotřebení.

4) Mo - feritotvorný prvek, odolnost proti korozi, oxidaci a zvýšení žáropevnosti.

Tabulka 2 poskytuje informaci o prvních odhadech komunalit, charakteristických číslech a vysvětlené variabilitě. Komunalita představuje váhu s jakou jednotlivé manifestní proměnné přispívají ke konstrukci příslušných latentních proměnných (nejvíc přispívají Cr a Si a nejméně C a Mo). K vysvětlení více jak 85 % variability dat je zapotřebí min. 3 faktorů. Ty společně vysvětlují 93 % variability dat. Tabulka 3 udává matici faktorových zátěží, které představují kovariance resp. korelační koeficienty mezi jednotlivými proměnnými a společnými faktory.

1. faktor je silně korelován především Cr, Si a Mo,
2. faktor proměnnou C,
3. faktor hlavně proměnnou Mo.

V tabulce 3 jsou uvedeny také odhady komunalit odpovídající faktorové matici. Je zřejmé, že došlo k celkovému zvýšení míry a srovnání podílů společné vysvětlené variability jednotlivých proměnných odpovídající 3 faktorům (bližší hodnoty ukazují na podobné vlastnosti). Na celkové společné vysvětlené variabilitě se nejvíce podílí C dále Mo.



Obr. 10 Dendrogramy proměnných (horní řada) a dendrogramy objektů (dolní řada) průměrovou metodou (vlevo) a metodou nejvzdálenějšího souseda (vpravo).

Tabulka č. 2

Proměnná	Komunalita	Faktor	Char. hodnota	Variabilita	Kumul. var.
C	0,225	1	2,119	53,0	53,0
Si	0,617	2	1,097	27,4	80,4
Cr	0,696	3	0,611	15,3	95,7
Mo	0,373	4	0,173	4,3	100,0

Tabulka č. 3. Vypočtené faktory

Proměnná/Faktor	1	2	3	komunality
C	0,139	-0,954	0,242	0,989
Si	0,841	-0,285	-0,390	0,940
Cr	-0,916	-0,201	0,173	0,910
Mo	0,743	0,254	0,609	0,987

Závěr PCA a FA: Faktorově nejčistší vychází proměnná C (druhý faktor, po rotaci třetí), což znamená, že je nejméně závislá a podobná ostatním proměnným. Stejně vychází Cr (1. faktor). Si a Mo vycházejí po rotaci faktorově čisté (1. faktor a 2. faktor). Tím, že proměnná C jeví zcela odlišné vlastnosti od ostatních proměnných je zapotřebí dalšího samostatného faktoru k vysvětlení jejího chování. K vysvětlení chování proměnných bez uhlíku by stačily pouze dva faktory. Čím odlišnější vlastnosti mezi jednotlivými proměnnými tím více latentních proměnných je zapotřebí k popisu chování proměnných.

3) CLU - Analýza shluků - hledání struktury v objektech: Shluková analýza umožňuje na základě podobnosti vícerozměrných objektů, jejich třídění do skupin (shluků). Hierarchickým postupem dochází k postupnému spojování objektů do shluků a dále do větších shluků. Z matice vzdálenosti se určí nejmenší vzdálenost dvou objektů, jež se spojí ve shluk, který oba objekty nahradí novým. Následně se spočte nová matice vzdáleností a postup se opakuje až vznikne jeden velký shluk. Vzdálenost mezi objekty se měří tvz. euklidovskou metrikou a vzdálenost mezi shluky se měří několika metodami (průměru, centroidu, nejbližšího a nejvzdálenějšího souseda). U metody průměru se vypočte vzdálenost mezi 2 shluky jako průměr ze všech mezishlukových vzdáleností objektů patřících každý do jiného shluku. Horní dva dendrogramy na obr. 10 znázorňují strukturu tvorby výsledného shluku proměnných, když jsou zahrnuty všechny proměnné, a to dvěma rozličnými shlukovacími metodami. Dolní dva dendrogramy na obr. 10 pak ukazují tvorbu dendrogramu objektů, kdy v prvních 19 krocích jsou spojeny dvojice objektů do 19 shluků. Pak následuje střídavé spojování shluků do větších shluků, shluků a objektů a objektů do shluku až do kroku 127, kde proces končí vytvořením 1 shluku. Obrázek představuje schéma vzniku finálního shluku. Číslo pod označením objektů uvádí pořadí spojení dvou objektů nebo shluku a objektu. Při zhruba 90 %ní podobnosti existují dva velké shluky c112 (podobnost objektů 91 %) a c 115 (podobnost objektů 89 %), které obsahují 43 a 47 objektů. Tabulka 4 udává četnosti jednotlivých umístění expozic v daných shlucích. Z tabulky je patrné, že v obou shlucích dominují určité expozice a naopak. Oba shluky jsou chudé na objekty s umístěním expozic na 4. místě (pouze 19 %). Poměr vnějších expozic k vnitřním v obou shlucích je 41:59, což znamená, že podobnější jsou vnitřní expozice. Shluky c117, c122 a c125 se zapojují do spojování až od 81 %ní hladiny.do konečného shluku 127. Je patrné, že shluk 124 obsahuje nejvíc objektů s umístěním expozic na 4. místě (shluky obsahující tyto objekty se zapojují do tvorby konečného shluku až v posledních fázích.

Tabulka 4. Četnosti jednotlivých umístění v daných shlucích

Expozice	1	2	3	4	5	6	7	8
c112	4	8	2	2	3	3	10	11
c115	7	4	9	1	8	10	5	3
%								
c112	25	50	12,5	12,5	18,8	18,8	62,5	68,8
c115	44	25	56	6,25	50	62,5	31,25	18,8

Závěr CLU: Z korelační analýzy vyplynulo, že existuje určitá rozdílnost v chemickém složení mezi jednotlivými místy. Strukturně odlehlé body se jeví 51, 52, 55 a 192, což ukázaly i symbolové a PCA grafy. Do výpočtu byly zahrnuty všechny proměnné. Determinant korelační matice je velice blízký nule (-0,0046), proto je možno předpokládat multikolinearitu (závislost mezi proměnnými). Párové korelační koeficienty a příslušné kovariance naznačují významnou míru vztahu mezi některými proměnnými. Korelační analýza potvrdila významné korelace mezi proměnnými z rozptylových grafů. Prvky ve skupinách (Al, Mo a Ti) a (Mn, P, S, Cu, Cr a B vykazují v rámci skupiny pozitivní korelace a naopak meziskupinově negativní korelace. C, Ni a Si (pouze Cr) se jeví jako nezávislé.

Závěr úlohy 1: Pomocí vizuálního posouzení vícerozměrných dat je možné předběžně činit závěry o chování jednotlivých proměnných. Společně s metodou PCA byl počet proměnných snížen na základních 4 (C, Si, Cr a Mo), které reprezentují chování všech proměnných. Metodou FA vyšel C jako faktorově nejčistší tzn., že je nejméně závislý na ostatních proměnných, i když s určitou vazbou na Si, který tvoří jistý předěl mezi C a Mo. Chrom a molybden vykazují navzájem opačné chování. Shlukovou analýzou byly identifikovány strukturně odlehlé body 51, 52, 55 a 192. Zároveň byla indikována určitá nehomogenita mezi analyzovanými místy ve vzorcích. Mezi vzorky samotnými nebyl nalezen žádný koncentrační trend. Celá analýza byla provedena ještě jednou s totožnými závěry na datech získaných stejným postupem měření 16 vzorků. Před detailnějším statistickým zpracování je třeba brát v úvahu tyto závěry.

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

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, Academia Praha 2004.
[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.
[31] Meloun M. , Militký J., *Kompendium statistického zpracování dat*, Academia Praha 2002, Academia Praha 2006.

Internal bounds and hidden structure of the metallurgic data with the use of Multivariate Data Analysis MDA

Milan Meloun[1], Roman Lisztwan[2]

[1]*Katedra analytické chemie, Chemickotechnologická fakulta,*
Univerzita Pardubice, nám. Čs. Legií 565, 532 10 Pardubice,
email: milan.meloun@upce.cz, ICQ: 224-001-003, http: //meloun.upce.cz,
[2]*Třinecké železárny, Průmyslová 1000, 739 70 Třinec, email: roman.lisztwan@trz.cz*

Summary: *Multivariate data analysis **MDA*** deals with ***Objects*** (samples, individuals, molecules, ...) described by ***Variables*** (quantities, parameters, biological activities, etc.). It searches for relationship among objects, among variables, and among objects and variables. MDA is based on the latent variables being formed as the linear transformation of the original variables. The source data matrix contains variables in *m* columns and *n* rows. Before a data treatment the data are scaled. The ***principal components analysis*** reduces dimensionality and presents objects in two or three dimensions. The *plot of components weight* shows hidden structure among variables while the *scatterplot of component score* shows the hidden structure of objects. Clustering means searching for groups of similar objects or similar variables. When objects are known to exist in several categories, clustering techniques can be used as classification techniques. A fundamental concept of clustering techniques is ***similarity.*** Similarity of objects and variables is considered on base on *Mahalonobis distance* or *Euclidean distance* in the *m*-dimensional space. The ***cluster analysis*** leads to clusters which may be plotted in ***dendrogram***. There are two dendrograms available, the dendrogram of variables and the dendrogram of objects. The *agglomerative hierarchical methods* start with many clusters as objects. Clusters are progressively linked to form bigger clusters, until a single big cluster, of all the objects, is obtained. *Divisive methods* start instead with a single big cluster, and this is progressively subdivided into smaller cluster, until as many clusters as objects have been obtained. Both statistical techniques are demonstrated on the analysis and classification of the homogeneity of steel material.

Key Words: PCA, Principal Components Analysis, Cluster Analysis, Dendrogram, Scatterplot, Scree Plot, Components Weight Plot, Steel analysis, Correlation matrix.