Univerzita Pardubice
Fakulta ekonomicko-správní

# Zpracování podkladů pro praktickou část distanční opory pro předmět KZMSA - část hierarchické shlukování

**Jan Míčka**

Bakalářská práce

2009

**Univerzita Pardubice**
**Fakulta ekonomicko-správní**
**Ústav systémového inženýrství a informatiky**
Akademický rok: **2008/2009**

# ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Jan MÍČKA**

Studijní program: **B6209 Systémové inženýrství a informatika**

Studijní obor: **Informatika ve veřejné správě**

Název tématu: **Zpracování podkladů pro praktickou část distanční opory pro předmět KZMSA - část hierarchické shlukování**

Zásady pro vypracování:

Metody hierarchického shlukování.
Předzpracování dat.
Návrh příkladu.
Postup řešení.
Ukázka řešení včetně slovního popisu.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

**KUBANOVÁ J. Statistické metody pro ekonomickou a technickou praxi. Statis Bratislava, 2004.**
**LUKASOVÁ A. - ŠARMANOVÁ J Metody shlukové analýzy. ,Praha, 1985.**
**ŘEZANKOVÁ H., HÚSEK D. Shluková analýza dat. Professional Publishing, Praha, 2007.**
**Zdroje na internetu.**

Vedoucí bakalářské práce: **Ing. Hana Jonášová, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **6. října 2008**

Termín odevzdání bakalářské práce: **1. května 2009**

doc. Ing. Renáta Myšková, Ph.D.
děkanka

L.S.

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 6. října 2008

Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 16.04.2009

Na tomto místě bych rád poděkoval Ing. Haně Jonášové, Ph.D. za cenné rady, poskytnuté informace a připomínky, kterými přispěla k vypracování této bakalářské práce.

**Souhrn**

Bakalářská práce s názvem „ Zpracování podkladů pro praktickou část distanční opory pro předmět KZMSA, část hierarchické shlukování“ se věnuje problematice hierarchických metod shlukové analýzy včetně předzpracování dat.

V první části teoretické práce jsou popsány jednotlivé charakteristiky souboru včetně modelových příkladů. Druhá část se již podrobněji věnuje samotné shlukové analýze, je vysvětleno předzpracování dat, míry podobnosti, resp. nepodobnosti a v poslední části jsou již ukázány samotné metody hierarchického shlukování na navrženém praktickém příkladě.

**Klíčová slova**

Popisná statistika, shluková analýza, standardizace, normalizace, metrika, hierarchické shlukovací metody

**Title**

Data preparation for a practical part of a distance support for subjekt KZMSA, part hierarchical clustering.

**Summary**

Bachelor thesis, titled "Data preparation for a practical part of a distance support for subjekt KZMSA, part hierarchical clustering ” deals with the problem of hierarchical methods of a cluster analysis including the data preparation.

The first part of the theoretical work describes individual charakteristics, including model examples. The second part deals in detail with the cluster analysis itself, explaining data preparation, levels of similarity respectively dissimilarity and in the last part the hierarchical clustering methods themself are already shown on a designed practical example.

**Keywords**

Descriptive statistics, cluster analysis, standardization, normalization, metrics, hierarchical clustering method

**Obsah**

**Úvod** ... 13
1 **Předmět a cíl práce** ... 14
2 **Popisná statistika** ... 14
Základní soubor ... 15
Výběr ... 15
2.1 Charakteristiky polohy ... 15
2.1.1 Aritmetický průměr ... 16
2.1.2 Medián ... 16
2.1.3 Modus ... 17
2.1.4 Useknutý průměr ... 18
2.1.5 Kvantily ... 19
2.1.6 Kvartily ... 19
2.1.7 Minimum ... 20
2.1.8 Maximum ... 20
2.2 Charakteristiky variability ... 20
2.2.1 Variační rozpětí ... 21
2.2.2 Rozptyl základního souboru ... 21
2.2.3 Rozptyl výběru ... 22
2.2.4 Směrodatná odchylka základního souboru ... 23
2.2.5 Směrodatná odchylka výběrová ... 23
2.2.6 Variační koeficient ... 24
2.3 Charakteristiky tvaru rozdělení ... 25
2.3.1 Šikmost ... 25
2.3.2 Špičatost ... 26
3 **Shluková analýza** ... 28
3.1 Úvod ... 28
3.2 Metody shlukové analýzy ... 29
3.3 Klasifikace ... 30
3.4 Standardizace ... 30
3.5 Normalizace ... 32
3.6 Podobnost objektů ... 33
3.6.1 Koeficienty asociace ... 34

3.6.2 Koeficient korelace ..... 38
3.6.3 Metriky ..... 39
4 **Metody hierarchického shlukování** ..... 41
4.1 Metoda nejbližšího souseda ..... 42
4.2 Metoda nejvzdálenějšího souseda ..... 43
4.3 Centroidní metoda ..... 44
4.4 Metoda průměrné vazby ..... 45
4.5 Mediánová metoda ..... 46
4.6 Lanceova a Williamsova „pružná strategie“ ..... 47
4.7 Wardova – Wishartova metoda ..... 47
5 **Příklad** ..... 49
5.1 Výchozí data ..... 49
5.2 Postup řešení ..... 51
5.3 Průběh shlukování ..... 52
5.4 Výsledky shlukování ..... 53
5.5 Dendogram ..... 54
**Závěr** ..... 55
**Použitá literatura** ..... 56

**Seznam obrázků**

Obrázek 1 - rozložení šikmosti....26
Obrázek 2 - rozložení špičatosti....27
Obrázek 3 - rozdělení shlukovacích metod....30
Obrázek 4 - výsledky koeficientů asociace....37
Obrázek 5 - graf míry podobnosti - modelový příklad 4....37
Obrázek 6 - metoda nejbližšího souseda....43
Obrázek 7 - metoda nejvzdálenějšího souseda....44
Obrázek 8 - dendogram použitý z příkladu 3....54

**Seznam tabulek**

Tabulka 1 - modelový příklad 1 - aritmetický průměr....16
Tabulka 2 - modelový příklad 1 - medián....17
Tabulka 3 - modelový příklad 1 - modus....18
Tabulka 4 - modelový příklad 2 - useknutý průměr....18
Tabulka 5 - modelový příklad 2 - kvartily....19
Tabulka 6 - modelový příklad 2 - variační rozpětí....21
Tabulka 7 - modelový příklad 2 - rozptyl základního souboru....22
Tabulka 8 - modelový příklad 2 - výběrový rozptyl....22
Tabulka 9 - modelový příklad 2 - směrodatná odchylka základního souboru....23
Tabulka 10 - modelový příklad 2 - výběrová směrodatná odchylka....24
Tabulka 11 - modelový příklad 2 - variační koeficient....25
Tabulka 12 - modelový příklad 3 - standardizace....31
Tabulka 13 - modelový příklad 3 - vypočtená směrodatná odchylka a střední hodnota....32
Tabulka 14 - modelový příklad 3 - standardizovaná matice....32
Tabulka 15 - modelový příklad 3 - normalizovaná matice....33
Tabulka 16 - asociační tabulka....34
Tabulka 17 - výchozí data pro modelový příklad 4....36
Tabulka 18 - pomocná asociační tabulka....36
Tabulka 19 - modelový příklad 3 - korelace vlastností....38
Tabulka 20 - modelový příklad 3 - korelace objektů....38
Tabulka 21 - základní data pro praktickou část....50
Tabulka 22 - data pro ukázku průběhu shlukování....52
Tabulka 23 - vytvořené shluky metodou nejvzdálenějšího souseda....53

**Seznam použitých parametrů**

| | |
|---|---|
| $\bar{x}$ | Aritmetický průměr |
| $\tilde{x}$ | Medián |
| $\hat{x}$ | Modus |
| $R$ | Variační rozpětí |
| $x_{\max}$ | Maximální hodnota znaku |
| $x_{\min}$ | Minimální hodnota znaku |
| $\sigma^2$ | Rozptyl základního souboru |
| $s^2$ | Rozptyl náhodného výběru |
| $\sigma$ | Směrodatná odchylka základního souboru |
| $s$ | Směrodatná odchylka náhodného výběru |
| $VK$ | Variační koeficient |
| $\alpha$ | Koeficient šikmosti |
| $\beta$ | Koeficient špičatosti |
| $\pi$ | Míra podobnosti |
| $S_j$ | Jaccardův koeficient asociace |
| $S_{SM}$ | Sokalův-Michenerův koeficient asociace |
| $S_{RR}$ | Russellův-Raoův koeficient asociace |
| $S_D$ | Diceův koeficient asociace |
| $S_{RT}$ | Rogersův-Tanimotův koeficient asociace |
| $S_H$ | Hamanův koeficient asociace |
| $S_{N1}$ | N1 koeficient asociace |
| $S_{N2}$ | N2 koeficient asociace |
| $D_{SL}$ | Metoda nejbližšího souseda |
| $D_{CL}$ | Metoda nejvzdálenějšího souseda |
| $D_{WGM}$ | Centroidní metoda |
| $D_{AL}$ | Metoda průměrné vazby |
| $D_{UWGM}$ | Mediánová metoda |

**Seznam použitých vzorců**


(1) Aritmetický průměr

(2) Medián pro sudý počet objektů

(3) Medián pro lichý počet objektů

(4) Kvantity

(5) Variační rozpětí

(6) Rozptyl základního souboru

(7) Rozptyl výběru

(8) Směrodatná odchylka základního souboru

(9) Směrodatná odchylka výběrová

(10) Variační koeficient

(11) Koeficient šikmosti

(12) Koeficient špičatosti

(13) Střední hodnota pro výpočet standardizace

(14) Směrodatná odchylka pro výpočet standardizace

(15) Standardizace

(16) Normalizace

(17) Podmínka míry podobnosti objektů 1

(18) Podmínka míry podobnosti objektů 2

(19) Jaccardův koeficient asociace

(20) Sokalův-Michenerův koeficient asociace

(21) Russellův-Raoův koeficient asociace

(22) Diceův koeficient asociace

(23) Rogersův-Tanimotův koeficient asociace

(24) Hamanův koeficient asociace

(25) Nepojmenovaný koeficient asociace 1

(26) Nepojmenovaný koeficient asociace 2

(27) Podmínka metrik 1

(28) Podmínka metrik 2

(29) Podmínka metrik 3

(30) Podmínka metrik 4

(31) Euklidovská vzdálenost

(32) Čtverec euklidovské vzdálenosti

(33) Hammingova vzdálenost
(34) Sokalova vzdálenost
(35) Čebyšova vzdálenost
(36) Obecné schéma výpočtu nepodobnosti
(37) Metoda nejbližšího souseda
(38) Rekurzivní schéma pro metodu nejbližšího souseda
(39) Metoda nejvzdálenějšího souseda
(40) Rekurzivní schéma pro metodu nejvzdálenějšího souseda
(41) Centroidní metoda
(42) Výpočet první složky centroidů
(43) Výpočet druhé složky centroidů
(44) Rekurzivní schéma pro centroidní metodu
(45) Metoda průměrné vazby
(46) Rekurzivní schéma pro metodu průměrné vazby
(47) Mediánová metoda
(48) Rekurzivní schéma pro mediánovou metodu
(49) Podmínka 1 volby koeficientů pro Lanceovu a Williamsovu „pružnou strategii“
(50) Podmínka 2 volby koeficientů pro Lanceovu a Williamsovu „pružnou strategii“
(51) Podmínka 3 volby koeficientů pro Lanceovu a Williamsovu „pružnou strategii“
(52) Podmínka 4 volby koeficientů pro Lanceovu a Williamsovu „pružnou strategii“
(53) Výsledek plynoucí z podmínek (49), (50), (51) a (52)
(54) Rekurzivní schéma pro Lanceovu a Williamsovu „pružnou strategii“
(55) Přírůstek hodnoty cílové funkce použitý pro Wardovu metodu
(56) Výpočet funkce $E_A$
(57) Výpočet funkce $E_A$
(58) Rekurzivní schéma pro Wardovu – Wishartovu metod

# Úvod

S rozvíjející se vědou a technikou vzrůstá i počet informací, které je třeba nějakým způsobem roztřídit a zpracovat. S příchodem prvních počítačů přišly i první pokusy tento proces hledání podobností mezi objekty a jevy matematicky definovat a zautomatizovat. Bylo potřeba najít techniku, která by tříděná data rozdělila do shluků. Tyto nové metody byly pojmenovány shluková analýza. Pojem shluková analýza se poprvé objevil ve čtyřicátých letech dvacátého století, ale velký rozvoj nastal právě až s rozšířením počítačů. První monografie z oblasti shlukové analýzy byla napsána nikoliv matematikem, ale psychologem kalifornské univerzity R. C. Tryonem roku 1939. Tryon definoval shlukovou analýzu takto: „Shluková analýza je obecný logický postup formulovaný jako procedura, pomocí niž seskupujeme objektivně jedince do skupin na základě jejich podobnosti a rozdílností.“ [8]

Pojem shluková analýza zahrnuje několik různých algoritmů a metod pro seskupování objektů podobného typu do příslušných kategorií. Jinými slovy je shluková analýza datový nástroj, který se zaměřuje na různé objekty a třídí je do skupin (shluků), a to způsobem, že podobnost mezi objekty je maximální, pokud patří do stejné skupiny a minimální, pokud nikoliv. [12]

Dnešní doba, která umožňuje každému vědeckovýzkumnému pracovišti řešit problémy na stále dokonalejších počítačích, dává metodám shlukové analýzy téměř nekonečné možnosti rozvoje. Dnes lze jen těžko najít vědní obor, v němž by tyto metody nenašly své uplatnění. Shluková analýza představuje nezbytnou součást každého automatizovaného systému analýzy vícerozměrných dat. [8]

# 1 Předmět a cíl práce

Předmětem této bakalářské práce je zpracování podkladů pro distanční oporu kombinovaného studia předmětu Základní metody shlukové analýzy dat (dále jen ZMSA). Cílem této práce je objasnění řešení příkladů právě metodami hierarchického shlukování a návrh souhrnného příkladu včetně řešení.

Samotná praktická část bude řešena jednak v prostředí Microsoft Excel (dále jen Excel), jednak v prostředí programu Unistat, ale také s pomocí softwaru Statistica 7. Produkt Unistat je začleněn do struktury Microsoft Office, což je jeho nespornou výhodou. To se týká jak komunikace, tak vzájemného přenosu dat, grafů i tabulek do Wordu nebo Excelu. Unistat může být spuštěn samostatně nebo jako Unistat for Excel. V aplikaci Microsoft Excel lze nalézt pod nabídkou Nástroje/Analýza dat řadu statistických procedur, např. popisnou statistiku. Ta je pro shlukování velmi důležitou součástí. [1]

Cílem práce není vysvětlit, jak pracovat v prostředí softwarů Excel, Unistat nebo Statistica, nýbrž pomocí nich zpracovat data a popsat veškeré metodiky práce při shlukování dat hierarchickými metodami.

# 2 Popisná statistika

Ještě než lze přikročit k samotné shlukové analýze, je potřeba se seznámit s charakteristikami zpracovávaného souboru. Základním úkolem popisné statistiky je poskytnout věcně správné informace o průběhu jevů a procesů pomocí číselných charakteristik nebo ukazatelů. [3]

Excel je statistickými funkcemi velmi dobře vybaven a stačí, zvolí-li se v nástrojích analýza dat, Excel pak dle výběru dat automaticky vygeneruje tabulku s nejčastěji používanými statistickými veličinami. [3]

Statistické zpracování dat pomocí tabulek a grafů usnadňuje jejich vizuální analýzu a celkové posouzení datové konfigurace. Pro další zpracování je však potřeba data vhodně upravit. Proto se počítají různé číselné charakteristiky – popisné statistiky, které zachycují různé aspekty dat. [1]

Na úvod před samotnými charakteristikami ještě objasnění základních pojmů. Základní soubor a náhodný výběr.

**Základní soubor**

Základním souborem je určitá množina prvků (osob, automobilů, zvířat, území, podniků, organizací, úřadů, materiálů atd.), které jsou předmětem zkoumání, čili statistického šetření. [3]
Základní soubor je tedy určitá, věcně, prostorově a časově vymezená množina všech zkoumaných prvků, u kterých se zjišťují hodnoty jisté sledované veličiny. Sledovaná veličina se nazývá statistický znak a prvky základního souboru se nazývají statistické jednotky. [3]

**Výběr**

Cílem statistického zkoumání je poznání vlastností základního souboru. Vzhledem k tomu, že základní soubor může mít velmi značný rozsah, bylo by mnohdy zkoumání všech jeho prvků prakticky neuskutečnitelné nebo příliš pracné. Proto se dané zjišťování realizuje jen u vybraných jednotek základního souboru, tj. pouze na jeho vzorku. Tyto vybrané prvky ze základního souboru tvoří výběrový soubor neboli výběr, který by měl být co nejlepším představitelem základního souboru, ze kterého byl utvořen. [3]

Při výběru je potřeba zajistit, aby výsledky získané na základě měření byly platné. Proto je potřeba zaručit, aby byl výběr reprezentativní.
Reprezentativní výběr by měl splňovat následující předpoklady [3]:

- jednotlivé prvky základního souboru jsou vybírány nezávisle na sobě,
- všechny prvky pocházejí ze stejného základního souboru,
- každý prvek musí mít stejnou šanci dostat se do výběru.

## 2.1 Charakteristiky polohy

Charakteristiky polohy se snaží charakterizovat typickou hodnotu dat. Někdy se také nazývají míry střední hodnoty, míry centrální tendence nebo míry polohy, protože určují, kde na číselné ose je vzorek rozložen.

### 2.1.1 Aritmetický průměr

Aritmetický průměr je definován jako součet všech naměřených údajů vydělený jejich počtem. Aritmetický průměr se obvykle značí $\bar{x}$. [1]

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

Aritmetický průměr je pravděpodobně nejpoužívanější a také nejznámější statistickou veličinou. Každý určitě někdy počítal aritmetický průměr. S tímto faktem souvisí skutečnost, že aritmetický průměr je mnohdy požíván zcela nevhodně na místě, kde by bylo třeba využít jiné statistické veličiny [2]. V Excelu se používá funkce PRUMER.

**Př.:** Na modelovém příkladě 1 průměrné mzdy je jasně vidět, jak jedna hodnota ovlivní celý výsledek. Z tabulky plyne, že aritmetický průměr neodpovídá realitě, protože 4 z 5 zaměstnanců mají nižší mzdu, než je aritmetický průměr. Jediný pan Horák má nadprůměrnou mzdu, což ovlivnilo celý výsledek. Pokud se tedy aritmetický průměr používá nesprávně, může značně zkreslit realitu.

**Tabulka 1 – modelový příklad 1- aritmetický průměr, zdroj [autor]**

| Zaměstnanec | Mzda |
|---|---|
| p. Novák | 10500 |
| p. Novotný | 11000 |
| p. Horák | 28000 |
| p. Nový | 11000 |
| p. Sýkora | 13000 |
| **Aritmetický průměr** | **14700** |

### 2.1.2 Medián

Označován *Me* nebo $\tilde{x}$. Medián znamená hodnotu, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. Základní výhodou mediánu jako statistického ukazatele je fakt, že není ovlivněný extrémními hodnotami [1]. Medián je 50% kvantil, dělící soubor na dvě stejně velké poloviny. V Excelu se používá funkce MEDIAN.

Jestliže n je sudé číslo, pak *Me* je jakékoli číslo z intervalu ( $x_{n/2,}x_{n/2+1}$ ).

Jednoznačněji:

$$Me = 0{,}5(x_{n/2} + x_{n/2+1}) \tag{2}$$

Jestliže n je liché číslo, pak:

$$Me = x_{(n+1)/2} \tag{3}$$

**Př.:** Použije-li se medián na hodnoty modelového příkladu 1, je patrné, že odstraňuje extrémní hodnoty a mzda pana Horáka už výsledek neovlivňuje. Pro nalezení mediánu daného souboru stačí hodnoty seřadit podle velikosti a vzít hodnotu, která se nalézá uprostřed seznamu. Takto zjednodušeně lze najít medián.

**Tabulka 2 – modelový příklad 1- medián, zdroj [autor]**

| Zaměstnanec | Mzda |
|---|---|
| p. Novák | 10500 |
| p. Novotný | 11000 |
| p. Horák | 28000 |
| p. Nový | 11000 |
| p. Sýkora | 13000 |
| **Aritmetický průměr** | **14700** |
| **Medián** | **11000** |

### 2.1.3 Modus

Označuje se jako *Mo* nebo $\hat{x}$. Vrátí hodnotu, která se nejčastěji vyskytuje nebo opakuje v poli (matici) nebo oblasti dat (je to hodnota znaku s největší relativní četností). Výhodou modu je, že ho lze snadno použít i pro nečíselná data, kde např. aritmetický průměr použít nelze [2]. V Excelu se používá funkce MODE.

**Př.:** Použije-li se modus na modelovém příkladě 1, je roven mediánu, protože plat 11000 se v poli vyskytuje nejčastěji, přesněji dvakrát.

**Tabulka 3 – modelový příklad 1 – modus, zdroj [autor]**

| Zaměstnanec | Mzda |
|---|---|
| p. Novák | 10500 |
| p. Novotný | 11000 |
| p. Horák | 28000 |
| p. Nový | 11000 |
| p. Sýkora | 13000 |
| **Aritmetický průměr** | **14700** |
| **Medián** | **11000** |
| **Modus** | **11000** |

### 2.1.4 Useknutý průměr

Počítá se z pořádkových statistik (hodnoty seřazené dle velikosti v neklesající řadu). Např. desetiprocentní useknutý průměr znamená, že se vynechá 10% nejnižších výsledků a 10% nejvyšších výsledků a ze zbytku se počítá průměr. Obvykle se volí 5%, 10% nebo 25%. [5]

Ve vzorci se zadává celkový zlomek uříznutých hodnot (nezadává se hodnota v %, ale zlomek; udává se celkový zlomek pro uříznutí nahoře i dole. Např. desetiprocentní uříznutý průměr: zlomek pro 10% je 0,1; počítá se 0,1 dole a 0,1 nahoře, tedy zadaná hodnota je 0,2). [5]

Největší předností useknutého průměru je skutečnost, že zamezuje vlivu extrémních hodnot.

**Př.:** Na modelovém příkladě 2 průměrné spotřeby aut je znázorněn výsledek useknutého desetiprocentního průměru. Jelikož má pole pouze 10 hodnot, znamená to, že 10 % je jedna hodnota. Useknutý průměr tedy ubere nejmenší (A) a největší (J) hodnotu a ze zbytku spočte průměr.

**Tabulka 4 – modelový příklad 2 - useknutý průměr, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

Pro modelový příklad 2 má useknutý průměr hodnotu **5,78**.

### 2.1.5 Kvantily

Kvantil, přesněji řečeno p-procentní kvantil je hodnota, která dělí neklesající řadu pořádkových statistik na dvě části tak, že jedna obsahuje p% hodnot menších než kvantil nebo právě stejných a druhá obsahuje 100-p% (zbytek %) větších nebo právě stejných. Příslušné kvantity nějakého kvantitativního znaku x se značí $\tilde{x}_{0,5}, \tilde{x}_1, \tilde{x}_{10}, \tilde{x}_{20}, \tilde{x}_{25}, \tilde{x}_{50}, \tilde{x}_{95}$ apod. [5]

Kvantil členící soubor na dvě stejné četné poloviny, tedy 50% kvantit, tj. $\tilde{x}_{50}$, se nazývá medián neboli prostřední hodnota a značí se zpravidla jen $\tilde{x}$. Kvantily menší než medián se nazývají dolní kvantily. Naopak kvantily větší než medián se nazývají horní kvantily. [4]

Při výpočtu kvantilů je především nutné najít pořadové číslo jednotky, jejíž hodnota bude hledaný kvantil, popřípadě pořadové číslo jednotek, z jejichž hodnot lze tento kvantil spočítat. Označí-li se toto číslo $Z_p$, pak platí

$$Z_p = \frac{n_p}{100} + 0{,}5 \tag{4}$$

kde n je počet pozorování statistického souboru a p udává relativní četnost nejnižších hodnot, jejichž horní mez je hledaný kvantil. [13]

K nejpoužívanějším kvantilům patří kvartily, decily, percentily a centum. [4]

### 2.1.6 Kvartily

Kvartily jsou tři hodnoty znaku, které rozdělují uspořádanou řadu hodnot na čtyři stejně četné části. První neboli dolní kvartil $\tilde{x}_{25}$ odděluje čtvrtinu prvků s nejnižší hodnotou znaku. Druhým, resp. prostředním kvartilem je medián $\tilde{x}$. Třetí neboli horní kvartil $\tilde{x}_{75}$ odděluje čtvrtinu prvků s nejvyšší hodnotou znaku. [4]

**Př.:** Na modelovém příkladě 2 je znázorněn výpočet kvartilů $\tilde{x}_{25}$ a $\tilde{x}_{75}$.

**Tabulka 5 – modelový příklad 2 – kvartily, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

**Dolním kvartálem** $\tilde{x}_{25}$ z uspořádané řady n hodnot znaku bude člen, jehož pořadové číslo bude $\frac{n}{4}+\frac{1}{2}=\frac{10}{4}+\frac{1}{2}=3$. V tomto případě je tedy dolní kvartil $\tilde{x}_{25}$ = **5,1**.

**Horním kvartálem** $\tilde{x}_{75}$ z uspořádané řady n hodnot znaku bude člen, jehož pořadové číslo bude $\frac{n}{4}*3+\frac{1}{2}=\frac{10}{4}*3+\frac{1}{2}=8$. V tomto případě je tedy horní kvartil $\tilde{x}_{75}$ = **6,3**.

### 2.1.7 Minimum

Minimum vrací nejmenší hodnotu prvku z celého pole. Použití pouze pro číselná data. V Excelu se používá funkce MIN.

### 2.1.8 Maximum

Maximum vrací největší hodnotu prvku z celého pole. Použití pouze pro číselná data. V Excelu se používá funkce MAX.

## 2.2 Charakteristiky variability

V předchozí části byly objasněny charakteristiky polohy. Jejich účelem je umožnit srovnání úrovně hodnot znaku ve dvou či více souborech. Míry polohy však ne vždy stačí. Zcela různé řady hodnot či zcela různá rozdělení četností mohou mít stejné míry polohy. Například řada 4, 4, 4, 5, 5, 5, 5, 6, 6, 6 má aritmetický průměr, medián i modus roven pěti stejně tak jako řada 1, 1, 1, 5, 5, 5, 5, 9, 9, 9. [4]

Z názorného příkladu je zřejmé, že kromě úrovně, na níž se pohybují hodnoty sledovaných znaků, je třeba zkoumat i to, jak se jednotlivé hodnoty znaků vzájemně liší. A právě úroveň odlišnosti hodnot sledovaného znaku v daném souboru se nazývá variabilita. [4]

### 2.2.1 Variační rozpětí

Nejjednodušší a snadno určitelnou charakteristikou variability je variační rozpětí, které je rozdílem nejvyšší a nejnižší hodnoty sledovaného znaku [4]. Značí se písmenem *R*. Přestože se maximální a minimální hodnota uvádějí pravidelně při popisu dat, variační rozpětí *R* se počítá zřídka. Nevýhodou variačního rozpětí je velká citlivost k odlehlým hodnotám [1]. Jelikož závisí na extrémních hodnotách, které mohou být nahodilé, tak neříká nic o skutečné měnlivosti mezi těmito extrémy. Dvě v podstatě stejné řady hodnot znaku, lišící se pouze extrémními hodnotami, mají různé variační rozpětí. Naopak dvě zcela různé řady hodnot znaku, které mají náhodně shodné extrémy, mají variační rozpětí stejné. [4]

$$R = x_{\max} - x_{\min} \tag{5}$$

**Př**.: Na modelovém příkladě 2 je znázorněn výpočet R.

**Tabulka 6 - modelový příklad 2 - variační rozpětí, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

$$R = 8{,}8 - 4{,}1 = 4{,}7$$

Pro modelový příklad 2 má variační rozpětí hodnotu **4,7**.

### 2.2.2 Rozptyl základního souboru

Je to druhá mocnina směrodatné odchylky základního souboru. Ve statistických výpočtech se pracuje s rozptylem často. Stejně jako u směrodatné odchylky se rozeznávají dva rozptyly, rozptyl základního souboru a rozptyl výběrový. Rozptyl základního souboru se obvykle značí symbolem $\sigma^2$ [5]. V Excelu se používá funkce VAR.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \tag{6}$$

**Př.:** Na modelovém příkladě 2 je znázorněn výpočet rozptylu základního souboru.

**Tabulka 7 - modelový příklad 2 – rozptyl základního souboru, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

$$\sigma^2 = \frac{(4{,}1-5{,}91)^2 + (4{,}8-5{,}91)^2 + .... + (8{,}8-5{,}91)^2}{10} = 1{,}465$$

Pro modelový příklad 2 má rozptyl hodnotu **1,465**.

### 2.2.3 Rozptyl výběru

Rozptyl je definován jako průměrná kvadratická odchylka měření od aritmetického průměru, přičemž při průměrování této odchylky se dělí číslem (n-1). Značí se symbolem $s^2$.
Rozptyl se především používá pří výpočtu různých testovacích statistik. Počítá se pomocí čtverců odchylek dat od průměru, proto má jiný rozměr než původní data [1]. V Excelu se používá funkce VAR.VYBER.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{7}$$

**Př.:** Na modelovém příkladě 2 je znázorněn výpočet výběrového rozptylu.

**Tabulka 8- modelový příklad 2 – výběrový rozptyl, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

$$s^2 = \frac{(4{,}1-5{,}91)^2 + (4{,}8-5{,}91)^2 + .... + (8{,}8-5{,}91)^2}{9} = 1{,}628$$

Pro modelový příklad 2 má rozptyl hodnotu **1,628**.

### 2.2.4 Směrodatná odchylka základního souboru

Je spolu s rozptylem nejdůležitější ukazatel variability. Je mírou statistické disperze. Jedná se o kvadratický průměr odchylek hodnot znaku od jejich aritmetického průměru. Zhruba řečeno vypovídá o tom, jak moc se od sebe navzájem liší typické případy v souboru zkoumaných čísel. Je-li malá, jsou si prvky souboru většinou navzájem podobné, a naopak velká směrodatná odchylka signalizuje velké vzájemné odlišnosti [2]. Značí se písmenem s. Směrodatná odchylka je odmocnina z rozptylu a vrací míru rozptýlenosti do měřítka původních dat. Obvykle se značí symbolem $\sigma$ [1]. V Excelu se používá funkce SMODCH.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \tag{8}$$

**Př.:** Na modelovém příkladě 2 je znázorněn výpočet směrodatné odchylky základního souboru.

**Tabulka 9 - modelový příklad 2 - směrodatná odchylka základního souboru, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

$$\sigma = \sqrt{\frac{(4{,}1 - 5{,}91)^2 + (4{,}8 - 5{,}91)^2 + .... + (8{,}8 - 5{,}91)^2}{10}} = 1{,}21$$

Pro modelový příklad 2 má směrodatná odchylka hodnotu **1,21**.

### 2.2.5 Směrodatná odchylka výběrová

Používá se tehdy, počítá-li se z dat pro výběr. Obecně se jako jmenovatel ve vzorci dosazuje tzv. počet stupňů volnosti, což je počet prvků výběru *n* zmenšený o počet charakteristik, které se při výpočtu dle vzorce použijí a které byly již z hodnot naměřených na výběru vypočteny (*k*). Pro počet stupňů volnosti (značí se obvykle písmenem *v* – ný) tedy platí $v = n - k$; zkráceně se často říká jen stupně volnosti. Obvykle se značí symbolem $s$ [5]. V Excelu se používá funkce SMODCH.VYBER.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{9}$$

Odpověď na otázku, kdy dělit počtem *n* a kdy stupni volnosti (*n*-1). Je-li známa střední hodnota základního souboru (nastává jen tehdy, je-li znám přesně celý soubor, např. počet zaměstnanců v podniku), používá se *n*. Je-li znám pouze odhad střední hodnoty, tj. průměr z výběru, což je daleko častěji (např. průměr z 10 měření, která představují výběr z nekonečného počtu možných měření, což je základní soubor), dělí se počtem stupňů volnosti. [5]

**Př.:** Na modelovém příkladě 2 je znázorněn výpočet výběrové směrodatné odchylky.

**Tabulka 10 - modelový příklad 2 - výběrová směrodatná odchylka, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

$$s = \sqrt{\frac{(4{,}1 - 5{,}91)^2 + (4{,}8 - 5{,}91)^2 + .... + (8{,}8 - 5{,}91)^2}{9}} = 1{,}28$$

Pro modelový příklad 2 má výběrová směrodatná odchylka hodnotu **1,28**.

### 2.2.6 Variační koeficient

Jestliže je potřeba posoudit relativní velikost rozptýlenosti dat vzhledem k průměru, použije se koeficient variace neboli variační koeficient *VK*. Počítá se jako podíl směrodatné odchylky k aritmetickému průměru v procentech. [5]

$$VK = \frac{s}{\bar{x}} * 100\% \tag{10}$$

**Př.:** Na modelovém příkladě 2 je znázorněn výpočet variačního koeficientu.

**Tabulka 11 - modelový příklad 2 - variační koeficient, zdroj [autor]**

| Auto | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Spotřeba l/100km | 4,1 | 4,8 | 5,1 | 5,3 | 5,8 | 6,1 | 6,2 | 6,3 | 6,6 | 8,8 |

$$VK = \frac{1{,}21}{5{,}91} * 100 = 20{,}5\%$$

Pro modelový příklad 2 má variační koeficient hodnotu **20,5 %.**

## 2.3 Charakteristiky tvaru rozdělení

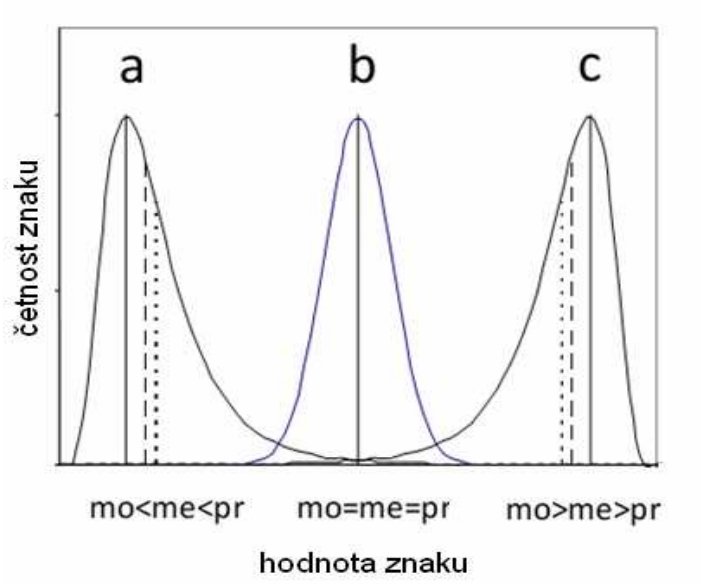
### 2.3.1 Šikmost

Charakteristiky šikmosti udávají, jsou-li hodnoty kolem zvoleného středu rozloženy souměrně nebo je-li rozdělení hodnot zešikmeno na jednu stranu. Všechny charakteristiky šikmosti nějakým způsobem využívají vztahů mezi průměrem $\bar{x}$, mediánem $\tilde{x}$ a modem $\hat{x}$.[6]
Charakteristika šikmosti slouží k jemnějšímu popisu specifických stránek dat [1]. V Excelu se používá funkce SKEW.

Koeficient šikmosti se vypočte podle vzorce [6]:

$$\alpha = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n.s^3} \tag{11}$$

a) $\alpha > 0$ pro kladně sešikmené rozdělení je $\bar{x} > \tilde{x} > \hat{x}$ (převládají nízké hodnoty)
b) $\alpha = 0$ pro symetrické rozdělení je $\bar{x} = \tilde{x} = \hat{x}$ ( hodnoty rovnoměrně rozloženy)
c) $\alpha < 0$ pro záporně sešikmené rozdělení je $\bar{x} < \tilde{x} < \hat{x}$ (převládají vysoké hodnoty)




**Obrázek 1 - rozložení šikmosti; zdroj: podle [5]**

### 2.3.2 Špičatost

Charakteristiky špičatosti udávají, jaký průběh má rozdělení hodnot kolem zvoleného středu (rozdělení). Čím je rozdělení špičatější, tím více jsou hodnoty soustředěny kolem daného středu rozdělení. Na druhé straně, rozdělení s nízkou špičatostí často obsahuje hodnoty velmi vzdálené od středu rozdělení [6]. V Excelu se používá funkce KURT.

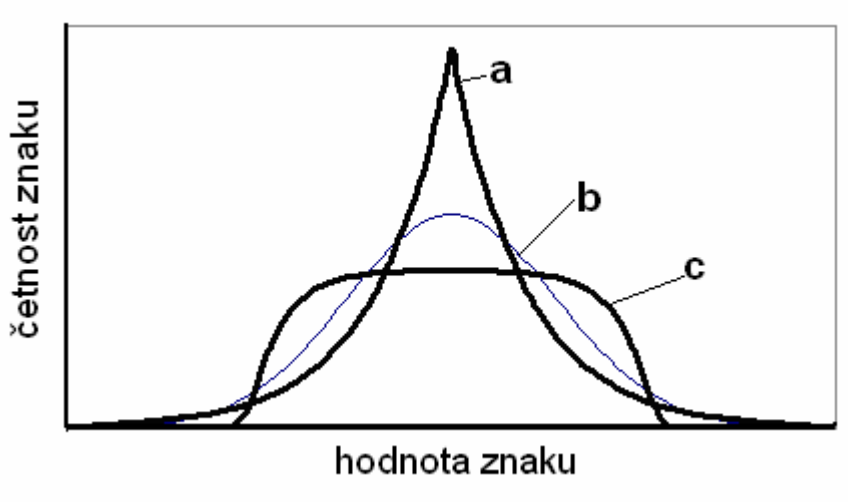
Koeficient špičatosti se vypočte podle vzorce [6]:

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n.s^4} - 3 \qquad (12)$$

a) $\beta$ > 0 špičaté (hodnoty koncentrovány kolem středu)

b) $\beta$ = 0 normální (hodnoty rovnoměrně rozloženy)

c) $\beta$ < 0 ploché (hodnoty nejsou koncentrovány kolem středu)




**Obrázek 2 - rozložení špičatosti; zdroj: podle [5]**

# 3 Shluková analýza

## 3.1 Úvod

Základním cílem shlukové analýzy je zařadit objekty do skupin (shluků). Rozklad množiny dat by měl být proveden takovým způsobem, aby si objekty uvnitř jednotlivých shluků byly co nejvíce podobné. Objekty patřící do různých shluků by si naopak měly být podobné co nejméně. Přitom objekty mohou být různého charakteru. Lze shlukovat živočichy či rostliny, věci, lidi atd. [7]

Vznik shlukové analýzy spadá do čtyřicátých let dvacátého století, ale její bouřlivý rozvoj nastal až s masovým zavedením počítačů. Shluková analýza se stala nedílnou složkou zpracování informací a je obsažena téměř ve všech běžně používaných statistických programech. [3]

Metody shlukové analýzy umožňují rozčlenit zkoumané objekty do vnitřně homogenních skupin, čili shluků. Principem je, že objekty uvnitř shluků jsou si nejvíce podobné a naopak objekty různých shluků jsou navzájem co nejvíce odlišné. Vztah mezi jednotlivými ukazateli je možné definovat pomocí korelační matice. [3]

Právě ona podobnost objektů je hlavním problémem shlukové analýzy. Aby mohla být podobnost měřena, musí být každý objekt charakterizován pomocí svých vlastností [7]. Například auto může být definováno barvou, obsahem motoru, značkou atd.

Shlukování založené na měření podobnosti se nazývá konvenční. Označí-li se dva objekty jako *A* a *B*, pak lze symbolicky zapsat, že

$$\text{Podobnost}(A,B) = f(\text{vlastnost}(A), \text{vlastnost}(B)),$$

tedy podobnost dvou objektů je funkcí jejich vlastností. Toto je však značně zjednodušené, ve skutečnosti může být interpretace skupin velmi obtížná. [7]

Kromě konvenčního shlukování se používá též shlukování konceptuální. Vytvářené shluky jsou zde založeny na konceptuální soudržnosti, která je funkcí jednak vlastností objektů, jednak popisného jazyka *L* a okolí *E*. Popisný jazyk je způsob, jakým jsou popsány třídy (skupiny) objektů, a okolí je množina sousedících vzorů. Symbolicky lze napsat, že

$$\text{konceptuální soudržnost}(A,B) = f(\text{vlastnosti}(A), \text{vlastnosti}(B), L, E).$$ [7]

Definic shlukové analýzy existuje celá řada, proto nelze v práci uvést všechny. Byly vybrány dvě, pravděpodobně nejsrozumitelnější definice. Tryonova definice již byla zmíněna v úvodu.

Tryon, 1939.
„Shluková analýza je obecný logický postup formulovaný jako procedura, pomocí niž seskupujeme objektivně jedince do skupin na základě jejich podobnosti a rozdílností.“ [8]

Bonner, 1964.
„Je dána množina objektů, z nichž každý je definován pomocí množiny znaků s ním souvisejících. Tato množina znaků je pro každý objekt stejná. Máme nalézt shluky objektů ( podmnožiny původní množiny objektů) tak, aby si členové shluku byli vzájemně podobní, ale nebyli si příliš podobní s objekty mimo tento shluk.“ [8]

## 3.2 Metody shlukové analýzy

Je známa řada metod shlukové analýzy a je obtížné je rozumným způsobem utřídit. Nejčastější způsob utřídění metod je nikoliv podle použitých matematických metod, ale podle systému použité klasifikace. Dle tohoto kritéria se dělí metody shlukové analýzy na dvě základní skupiny [3]:

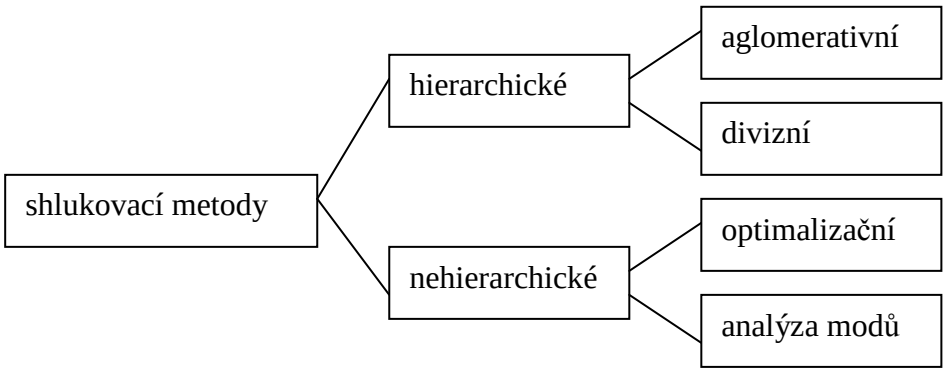
- hierarchické metody,
- nehierarchické metody.

Hierarchické metody lze charakterizovat tak, že každý shluk je současně podmnožinou jiného shluku s výjimkou samotné množiny objektů, která je považována za maximálně možný shluk. Tyto metody se dělí na dvě základní skupiny, lišící se způsobem shlukování. Jedná se o aglomerativní a divizní přístup. [3]

- Aglomerativní přístup je charakteristický tím, že vychází od jednotlivých objektů a jejich postupným seskupováním buduje hierarchický systém podmnožin až dospěje ke konečnému spojení všech objektů do množiny objektů O. [3]
- Divizní přístup shlukování je založen na tom, že vychází z množiny objektů určených ke klasifikaci jako celku a jejich postupným rozdělováním získává hierarchický systém podmnožin. [3]

Nehierarchické metody se dělí také na dvě základní skupiny [8]:

- optimalizační metody hledají takový rozklad množiny objektů určených ke klasifikaci, který je optimální podle vhodně zvoleného kritéria optimality rozkladu,
- metody analýzy modů používají pravděpodobnostní přístup.




**Obrázek 3 - rozdělení shlukovacích metod; zdroj: podle [8]**

Před samotným shlukováním je potřeba si předpřipravit data tak, aby se dala lépe zpracovávat. Je možné data klasifikovat, standardizovat a normalizovat.

## 3.3 Klasifikace

Klasifikací se nazývá činnost vytvářející rozklad nějaké množiny objektů, činnost vedoucí k vytvoření systému tříd. Tento systém tříd však rovněž bývá nazýván klasifikací. Jde-li tedy o výsledek činností, je vhodnější nazývat ho klasifikační systém. [8]

Klasifikací nebo klasifikačním systémem se tedy rozumí, jak je uvedeno výše, rozklad množiny objektů na třídy. Přitom každou z těchto tříd lze považovat za výchozí množinu objektů a v ní rovněž provádět klasifikaci. Takto vytvořený systém rozkladů se nazývá hierarchickým klasifikačním systémem nebo hierarchickou klasifikací. [8]

## 3.4 Standardizace

Velmi často jsou získané hodnoty jednotlivých znaků jako výsledky měření v různých jednotkách. V takových případech se může stát, že objekt určený ke shlukování je např. popsán údaji v gramech, stupních, milimetrech apod. Metody shlukové analýzy, které vycházejí z podobnostních vztahů mezi objekty, by neměly pracovat s daty závislými na jednotkách měření. Proto je vhodné před shlukováním provést standardizaci dat. Znamená to, že všechny znaky budou souměřitelné. [8]

Po standardizaci jsou data bezrozměrná, jejich střední hodnota se rovná 0 a směrodatná odchylka se rovná 1.

Je dána základní matice $Z = (z_{ij})$ dat typu $n$ x $p$, kde $n$ je počet objektů a $p$ je počet vlastností. Pro výpočet standardizace je potřeba spočítat střední hodnotu $\bar{z}_j$ j-tého znaku $z_j$ a směrodatnou odchylku $s_j$ pro $j = 1, 2, \ldots, p$ podle vzorců [8]:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_{ij} \tag{13}$$

$$s_j = \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\left(z_{ij} - \bar{z}_j\right)^2\right]} \tag{14}$$

Původně naměřené hodnoty $z_{ij}$ j-tého znaku i-tého objektu se přepočtou na tzv. standardizované hodnoty [8]:

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j} \tag{15}$$

**Př**. Modelový příklad 3. V tabulce jsou uvedena jména pěti hokejistů HC Pardubice a k nim přiřazených 6 vlastností. Toto je tedy výchozí matice dat.

- Z – počet odehraných zápasů
- G – počet vstřelených branek
- A – počet gólových přihrávek
- B – počet nasbíraných bodů celkem
- +/- – hodnocení hráče ( + za účast na ledě při vstřelené brance, - při obdržené)
- Tmin – počet trestných minut, které hráč obdržel

**Tabulka 12 - modelový příklad 3 – základní data, zdroj [autor]**

| | Z | G | A | B | +/− | Tmin |
|---|---|---|---|---|---|---|
| Kolář Jan | 37 | 11 | 22 | 33 | 6 | 22 |
| Pivko Libor | 37 | 14 | 17 | 31 | 10 | 46 |
| Divišek Tomáš | 36 | 12 | 18 | 30 | 15 | 40 |
| Starý Jan | 34 | 19 | 9 | 28 | 6 | 16 |
| Koukal Petr | 38 | 11 | 16 | 27 | 7 | 28 |

Jak je uvedeno v charakteristice standardizace, tak k výpočtu standardizované matice je zapotřebí směrodatná odchylka a střední hodnota.

**Tabulka 13 - modelový příklad 3 – vypočtená směrodatná odchylka a střední hodnota, zdroj [autor]**

| | | | | | | |
|---|---|---|---|---|---|---|
| Směrodatná odchylka | 1,36 | 3,01 | 4,22 | 2,14 | 3,43 | 11,13 |
| Střední hodnota | 36,4 | 13,4 | 16,4 | 29,8 | 8,8 | 30,4 |

Podle dosazení do vzorce (14) byla vypočítána standardizovaná data.

**Tabulka 14 – modelový příklad 3 - standardizovaná matice, zdroj [autor]**

| | Z | G | A | B | +/− | Tmin |
|---|---|---|---|---|---|---|
| Kolář Jan | 0,44 | -0,8 | 1,33 | 1,5 | -0,8 | -0,75 |
| Pivko Libor | 0,44 | 0,2 | 0,14 | 0,56 | 0,35 | 1,402 |
| Divišek Tomáš | -0,3 | -0,5 | 0,38 | 0,09 | 1,81 | 0,863 |
| Starý Jan | -1,8 | 1,86 | -1,8 | -0,8 | -0,8 | -1,29 |
| Koukal Petr | 1,18 | -0,8 | -0,1 | -1,3 | -0,5 | -0,22 |

## 3.5 Normalizace

Objekty pro shlukovou analýzu jsou určeny vektory o *p* složkách představujících hodnoty vybraných *p* znaků. Normy vektorů mohou někdy nežádoucím způsobem ovlivňovat výsledky kvantitativního hodnocení podobnostních vztahů mezi objekty. V takových případech je vhodné normalizovat vektory tak, aby měly stejnou normu (nejlépe jednotkovou). [8]

Tato procedura normalizuje všechny vektory (řádky) matice dat na jednotkovou normu tak, že všechny složky každého z vektorů dělí normou tohoto vektoru. Normalizace se provádí ze standardizovaných dat. [8]

$$a_i = \sqrt{\left[\sum_{j=1}^{p} \left(z_{ij}\right)^2\right]} \tag{16}$$

**Př.** Pro ukázku bude použit stejný příklad jako u standardizace, tedy modelový příklad 3. Normalizace se provádí ze standardizovaných dat. Pro výpočet ze standardizovaných je použit vektor normy. Výsledná znormalizovaná matice vypadá následovně.

**Tabulka 15 - modelový příklad 3 – normalizovaná matice, zdroj [autor]**

| | Z | G | A | B | +/− | Tmin |
|---|---|---|---|---|---|---|
| Kolář Jan | 15 | 4,46 | 8,93 | 13,4 | 2,43 | 8,928 |
| Pivko Libor | 22,7 | 8,59 | 10,4 | 19 | 6,13 | 28,21 |
| Divišek Tomáš | 17 | 5,68 | 8,51 | 14,2 | 7,1 | 18,92 |
| Starý Jan | 9,53 | 5,33 | 2,52 | 7,85 | 1,68 | 4,486 |
| Koukal Petr | 18,8 | 5,45 | 7,92 | 13,4 | 3,47 | 13,86 |

| Vektor normy |
|---|
| 2,464 |
| 1,631 |
| 2,114 |
| 3,567 |
| 2,020 |

## 3.6 Podobnost objektů

V souvislosti s metodami shlukové analýzy je důležitý pojem podobnosti, resp. nepodobnosti mezi jednotlivými objekty či znaky. V jednotlivých krocích algoritmů se posuzuje podobnost, resp. vzdálenost dvou objektů, objektu a shluku nebo dvou shluků. V některých případech je způsob hodnocení dán přímo shlukovací metodou, často jsou ale tyto kroky nezávislé a je potřeba vybrat nejvhodnější míru podobnosti, a to jak z hlediska shlukovaných objektů, tak i z hlediska použité metody shlukování. [3]

Ve všech případech je potřeba najít vhodný předpis $\pi$, který dvojici objektů *Oi*, *Oj* přiřadí číslo $\pi(O_i, O_j)$, které vyjadřuje míru podobnosti objektů. Tento předpis by navíc měl splňovat alespoň dvě podmínky [9]:

$$\pi(O_i, O_j) \geq 0 \tag{17}$$

$$\pi(O_i, O_j) = \pi(O_j, O_i) \tag{18}$$

Pokud je uvažováno $\pi$ jako míra podobnosti, pak by kromě podmínek (17) a (18) mělo platit, že hodnota $\pi(O_i, O_i)$, tedy podobnost objektu se sebou samým, je maximální možnou. Jednoduše řečeno, čím více jsou si objekty podobné, tím více hodnota $\pi$ narůstá. [9]

Pro účely shlukování se však jeví mnohem lepší použití míry k $\pi$ duální – míry nepodobnosti $\nu$. Podmínky (17) a (18) musí být splněny i v tomto případě, ale pro podobnost totožných objektů platí $\nu(O_i, O_i) = 0$. [9]

Přestože jsou pravidla, která musí míra podobnosti či nepodobnosti splňovat dostatečně přesně definována, neexistuje žádná univerzální míra, kterou by bylo možné použít pro všechny typy úloh a všechny typy dat. [9]

Míry podobnosti :

- Koeficienty asociace
- Koeficient korelace

Míry nepodobnosti:

- Metriky

### 3.6.1 Koeficienty asociace

K vyjádření vztahu mezi kvalitativními znaky byly ve statistice zavedeny koeficienty asociace neboli kontingenční koeficienty. Ve shlukové analýze se míra podobnosti nazývá koeficientem jen tehdy, kdy jsou objekty charakterizovány dichotomickými znaky. Při určování prvků v matici vzdáleností resp. podobností objektů $O_i$ a $O_j$ bude pozorována shoda či neshoda výsledků u p proměnných. Asociace dvou objektů je pak vyjádřena asociační čtyřpolní tabulkou se dvěma možnými výsledky, 1 a 0. [3]

**Tabulka 16 - asociační tabulka, zdroj [autor]**

| | | Oj | |
|---|---|---|---|
| | | 1 | 0 |
| Oi | 1 | a | b |
| | 0 | c | d |

**a – pozitivní shoda** – značí počet znaků, kde mají objekty $O_i$ a $O_j$ hodnotu 1
**b – negativní shoda** – značí počet znaků, kde má objekt $O_i$ hodnotu 1 a $O_j$ hodnotu 0
**c – negativní shoda** – značí počet znaků, kde má objekt $O_i$ hodnotu 0 a $O_j$ hodnotu 1
**d – neshoda** – značí počet znaků, kde mají objekty $O_i$ a $O_j$ hodnotu 0 [10]

Nyní budou uvedeny ty koeficienty, které jsou považovány za nejdůležitější, a současně budou zmíněny jejich vlastnosti a nedostatky. [8]

**Jaccardův koeficient asociace**

$$S_j = \frac{a}{a+b+c} \tag{19}$$

Koeficient $S_j$ není definován pro dvojice objektů vykazující negativní shodu ve všech znacích.

Jedná se o poměr mezi počtem pozitivních shod a počtem případů, kde alespoň jeden z objektů danou vlastnost má. [9]

**Sokalův-Michenerův koeficient asociace**

$$S_{SM} = \frac{a+d}{a+b+c+d} \tag{20}$$

Koeficient $S_{SM}$ vyjadřuje poměr mezi počtem shod a počtem všech případů. [9]

**Russellův-Raoův koeficient asociace**

$$S_{RR} = \frac{a}{a+b+c+d} \tag{21}$$

Koeficient $S_{RR}$ má tu nevýhodu, že různě hodnotí podobnost objektu se sebou samým a na diagonále nemusí vždy vyjít 1. [9]

**Diceův koeficient asociace**

$$S_D = \frac{2a}{2a+b+c} \tag{22}$$

Koeficient $S_D$ má stejná omezení jako Jaccardův koeficient, tedy že nepočítá s negativní shodou všech znaků. [9]

**Rogersův-Tanimotův koeficient asociace**

$$S_{RT} = \frac{a+d}{a+d+2(b+c)} \tag{23}$$

**Hamanův koeficient asociace**

$$S_H = \frac{(a+d)-(b+c)}{a+b+c+d} \tag{24}$$

Koeficient $S_H$ má jako obor hodnot interval <-1,1> . Hodnota -1 nastane v případě, kdy nedojde k žádné shodě, hodnota 1 v případě shody ve všech znacích a hodnota 0 v případě stejného počtu shod i neshod. [9]

**Nepojmenovaný koeficient asociace 1**

$$S_{N1} = \frac{2(a+d)}{2a+b+c+2d} \tag{25}$$

**Nepojmenovaný koeficient asociace 2**

$$S_{N2} = \frac{a}{a+2b+2c} \tag{26}$$

**Př.:** Výchozí data pro modelový příklad 4 na téma koeficienty asociace obsahují 3 objekty s 5 vlastnostmi. Jako objekty jsou zde 3 různé druhy automobilů označené jako O1, O2 a O3. Mezi vlastnosti jsem vybral to, zda má auto ve své výbavě klimatizaci, palubní počítač, automatickou převodovku, xenony a metalízu.

**Tabulka 17 - výchozí data pro modelový příklad 4, zdroj [autor]**

| objekt | klimatizace | palubní počítač | automatická převodovka | xenony | metalíza |
|---|---|---|---|---|---|
| O1 | 1 | 1 | 1 | 0 | 0 |
| O2 | 1 | 0 | 0 | 1 | 1 |
| O3 | 0 | 1 | 1 | 0 | 1 |

Podle tabulky 16 je sestrojena pomocná asociační tabulka, která znázorňuje počty shod a neshod mezi jednotlivými objekty.

**Tabulka 18 - pomocná asociační tabulka, zdroj [autor]**

| | | O1 | | O2 | | O3 | |
|---|---|---|---|---|---|---|---|
| | | 1 | 0 | 1 | 0 | 1 | 0 |
| O1 | 1 | 3 | 0 | 1 | 2 | 2 | 1 |
| | 0 | 0 | 2 | 2 | 0 | 1 | 1 |
| O2 | 1 | 1 | 2 | 3 | 0 | 1 | 2 |
| | 0 | 2 | 0 | 0 | 2 | 2 | 0 |
| O3 | 1 | 2 | 1 | 1 | 2 | 3 | 0 |
| | 0 | 1 | 1 | 2 | 0 | 0 | 2 |

| | |
|---|---|
| a | b |
| c | d |

Po sestrojení pomocné asociační tabulky již následuje samotný výpočet jednotlivých koeficientů. Výsledné koeficienty asociace vypočítané podle tabulky 17.

**Jaccardův koeficient**

| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 1,000 | 0,200 | 0,500 |
| O2 | 0,200 | 1,000 | 0,200 |
| O3 | 0,500 | 0,200 | 1,000 |

**Nepojmenovaný 1**

| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 1,000 | 0,333 | 0,750 |
| O2 | 0,333 | 1,000 | 0,333 |
| O3 | 0,750 | 0,333 | 1,000 |

**Sokalův a Michenerův koeficient**

| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 1,000 | 0,200 | 0,400 |
| O2 | 0,200 | 1,000 | 0,200 |
| O3 | 0,400 | 0,200 | 1,000 |

**Nepojmenovaný 2**

| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 1,000 | 0,111 | 0,333 |
| O2 | 0,111 | 1,000 | 0,111 |
| O3 | 0,333 | 0,111 | 1,000 |

**Russellův a Raoův koeficient**

| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 0,600 | 0,200 | 0,400 |
| O2 | 0,200 | 0,600 | 0,200 |
| O3 | 0,400 | 0,200 | 0,600 |

**Rogersův a Tanimotoův koeficient**

| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 1,000 | 0,111 | 0,429 |
| O2 | 0,111 | 1,000 | 0,111 |
| O3 | 0,429 | 0,111 | 1,000 |

**Diceův koeficient**

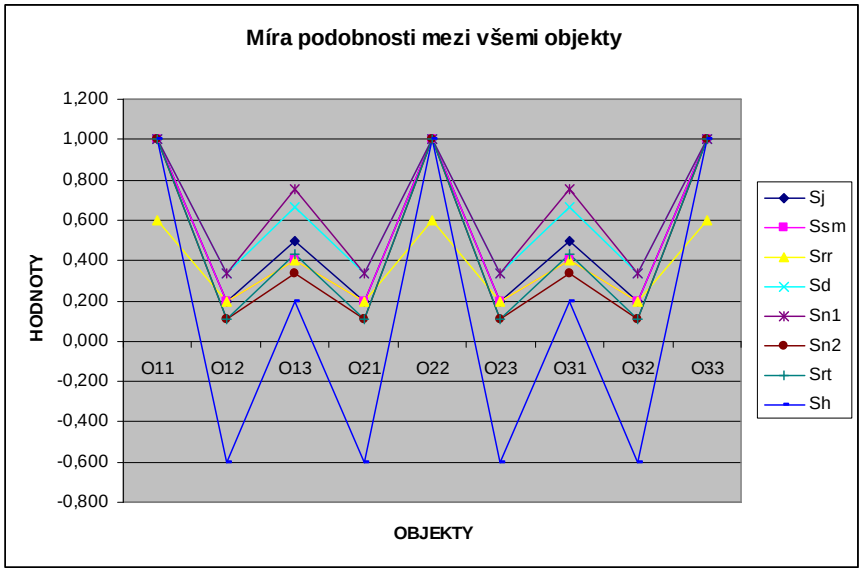
| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 1,000 | 0,333 | 0,667 |
| O2 | 0,333 | 1,000 | 0,333 |
| O3 | 0,667 | 0,333 | 1,000 |

**Hamannův koeficient**

| | O1 | O2 | O3 |
|---|---|---|---|
| O1 | 1,000 | -0,600 | 0,200 |
| O2 | -0,600 | 1,000 | -0,600 |
| O3 | 0,200 | -0,600 | 1,000 |

**Obrázek 4 - výsledky koeficientů asociace, zdroj [autor]**

Ze získaných výsledků z jednotlivých koeficientů asociace byl sestrojen pro lepší přehled graf podobnosti mezi jednotlivými objekty. Z grafu je evidentní, že nejpodobnější jsou si objekty O1 a O3, pomine-li se podobnost objektu se sebou samým, protože tato podobnost by měla být maximální možnou.




**Obrázek 5 - graf míry podobnosti - modelový příklad 4, zdroj [autor]**

### 3.6.2 Koeficient korelace

Korelace je ve statistice vzájemný vztah mezi znaky či veličinami. Korelační koeficient může nabývat hodnot od −1 až po +1. [2]

Hodnota korelačního koeficientu −1 značí nepřímou závislost, tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků, např. vztah mezi uplynulým a zbývajícím časem. Hodnota korelačního koeficientu +1 značí přímou závislost, např. vztah mezi rychlostí bicyklu a frekvencí otáček kola bicyklu. [2]

**Př.:** Na modelovém příkladě 3 je znázorněna korelační matice vztahů mezi vlastnostmi hokejistů. V tabulce je barvami znázorněna nejmenší a největší hodnota pole. Nejmenší hodnotu, v tabulce znázorněnou žlutou barvou a tedy i nepřímou závislost, mají vlastnosti počet vstřelených gólů a počet asistencí. Je jasné, že čím dá hráč více gólů, tím menšího může dosáhnout počtu asistencí a naopak. Největší hodnotu, v tabulce znázorněnou zelenou barvou a tedy přímou závislost, mají vlastnosti počet zápasů a celkový počet bodů. Hráč hrající méně zápasů, získá celkově méně bodů, naopak hráč hrající zápasů více, má na svém kontě více bodů. To samozřejmě platí pro tento příklad, ve skutečnosti to tak být nemusí. Co se týče korelace objektů, tak nejvíce jsou si podobní Tomáš Divíšek s Liborem Pivkem, naopak nejmenší podobnost je mezi Tomášem Divíškem a Janem Starým.

**Tabulka 19 – modelový příklad 3 – korelace vlastností, zdroj [autor]**

| | Z | G | A | B | +/? | Tmin |
|---|---|---|---|---|---|---|
| Z | 1 | | | | | |
| G | 0,71 | 1 | | | | |
| A | 0,89 | **0,435** | 1 | | | |
| B | **0,96** | 0,727 | 0,93 | 1 | | |
| +/? | 0,7 | 0,6 | 0,65 | 0,7 | 1 | |
| Tmin | 0,91 | 0,863 | 0,77 | 0,9 | 0,87 | 1 |

**Tabulka 20 - modelový příklad 3 - korelace objektů, zdroj [autor]**

| | Kolář | Pivko | Divišek | Starý | Koukal |
|---|---|---|---|---|---|
| Kolář Jan | 1 | | | | |
| Pivko Libor | 0,706 | 1 | | | |
| Divišek Tomáš | 0,741 | **0,981** | 1 | | |
| Starý Jan | 0,819 | 0,578 | **0,576** | 1 | |
| Koukal Petr | 0,924 | 0,883 | 0,906 | 0,817 | 1 |

### 3.6.3 Metriky

Nejobvyklejší způsob vyjádření míry podobnosti mezi objekty vychází z jejich geometrické reprezentace v prostoru. Je-li libovolné množství navzájem různých objektů, které jsou popsány *p* různými reálně-hodnotovými atributy, lze tyto objekty zobrazit v *p*-rozměrném prostoru pomocí diskrétních bodů. Situaci v jedno-, dvou a tří-rozměrném prostoru lze ještě posoudit okem, avšak těchto případů je málo, jelikož jednotlivé objekty bývají popsány větším množstvím měření. [9]

Ve vícerozměrném prostoru se lze těžko obejít bez vhodného nástroje, který by objektivně posoudil vzájemný vztah dvou bodů tohoto prostoru. Pravidla, která musí tento nástroj splňovat lze vyjádřit pojmem metrika. [9]

Metrikou *d* je funkce definovaná na kartézském součinu p-rozměrného prostoru $E_p$ x $E_p$. Metrika $_p$ přiřazuje každé dvojici bodů A,B, tohoto prostoru, reálné číslo p (A,B), které splňuje čtyři podmínky $(\forall A, B, C \in E_p)$ [9]:

$$d(A,B) = 0 \Leftrightarrow A = B \tag{27}$$

$$d(A,B) \geq 0 \tag{28}$$

$$d(A,B) = d(B,A) \tag{29}$$

$$d(A,C) \leq d(A,B) + d(B,C) \tag{30}$$

Dále bude uvedeno několik nejznámějších a nejpoužívanějších metrik.

**Euklidovská vzdálenost** je nejčastější mírou, která je někdy nazývána jako geometrická metrika a představuje délku přepony pravoúhlého trojúhelníka. Její výpočet je založen na Pythagorově větě. [10]

$$d(A,B) = \sqrt{\left[\sum_{i=1}^{p}(a_i - b_i)^2\right]} \tag{31}$$

**Čtverec euklidovské vzdálenosti** je používán méně, tvoří základ Wardovy metody shlukování. [10]

$$d^2(A,B)=\left[\sum_{i=1}^{p}(a_i-b_i)^2\right] \tag{32}$$

**Hammingova vzdálenost**, zvaná také Manhattanská vzdálenost nebo-li vzdálenost městských bloků. [10]

$$d(A,B)=\sum_{i=1}^{p}|a_i-b_i| \tag{33}$$

**Sokalova vzdálenost**

$$d(A,B)=\sqrt{\left[\frac{\sum_{i=1}^{p}(a_i-b_i)}{p}\right]} \tag{34}$$

**Čebyšova vzdálenost**

$$d(A,B)=\max|a_i-b_i| \tag{35}$$

Hammingova, Eukleidovská a Čebyšova vzdálenost mají podobné vlastnosti. Jejich společnou nevýhodou je to, že jsou závislé na jednotkách, ve kterých byly měřeny jednotlivé znaky. Pokud mají tyto znaky i různorodý charakter z hlediska jejich významnosti, není to žádným způsobem zohledněno při výpočtu vzdáleností. [3] Ukázkový příklad na metriky přiložen na CD.

# 4 Metody hierarchického shlukování

Pro hierarchické metody je společné to, že shlukovací proces má charakter posloupnosti rozkladů množiny objektů. Dle postupu algoritmu lze dělit tento typ metod na aglomerativní a divizní. [11]

V případě aglomerativních algoritmů je každý shluk tvořen právě jedním objektem. Počáteční stav, resp. rozklad množiny objektů na *n* jednoprvkových shluků, se nazývá nultý rozklad $\Xi_0$. V následujícím kroku aglomerativního algoritmu jsou vybrány dva jednoprvkové shluky, které jsou si z jistého pohledu nejpodobnější. Tyto objekty jsou sloučeny do nového shluku. Takto vzniklý shluk, spolu s ostatními. Jednoprvkovými shluky, tvoří rozklad $\Xi_1$. Uvedený postup se opakuje. V pořadí s-tý rozklad je tedy tvořen sloučením dvou nejpodobnějších si shluků předcházejícího s-1 rozkladu a ostatních nezměněných shluků s-1 rozkladu. Postup je ukončen při dosažení n-1 rozkladu, kdy shlukované objekty tvoří jediný shluk. [11]

Hierarchických shlukovacích algoritmů existuje celá řada. Aglomerativní algoritmy se mezi sebou liší především stanovením vzdálenosti mezi jednotlivými shluky v daném p-rozměrném prostoru. S ohledem na různě definovanou vzdálenost mezi shluky mohou různé algoritmy vést k naprosto rozdílným výsledkům i v případě, že jsou aplikovány na stejnou množinu dat. [11]

Pánové Lance a Williams poprvé uvedli obecnější přístup ke skupině hierarchických aglomerativních shlukovacích metod (strategií). Zavedli pojem koeficient nepodobnosti shluků a zabývali se způsobem výpočtu nepodobnosti nového shluku vzniklého sloučením dvou shluků s nejmenší nepodobností s ostatními shluky rozkladu. Dospěli k závěru, že pro některé strategie je možno tyto hodnoty vypočítat z hodnot už vypočtených v rámci předcházejícího rozkladu bez použití původní matice dat. Tyto strategie nazvali kombinatorickými na rozdíl od ostatních nekombinatorických strategií, které toto postupné přepočítávání vzájemných nepodobností shluků z předcházejícího kroku neumožňují a vyžadují pro výpočet nepodobnosti shluků původní data. [8]

Autoři Lance a Williams rovněž uvedli obecné schéma, podle něhož mohou kombinatorické metody vypočíst nepodobnost nově vzniklých shluků s ostatními shluky na základě uchovávaných hodnot koeficientů nepodobnosti už existujících shluků. Pro výpočet nepodobnosti *D(U,R)* shluku *U* se shlukem $R = P \cup Q$ odvodili obecné schéma [8]:

$$D(U,R) = \alpha_i D(U,P) + \alpha_j D(U,Q) + \beta D(P,Q) + \gamma D(U,P) - D(U,Q) \qquad (36)$$

kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. Koeficienty $\alpha_i, \alpha_j, \beta$ a $\gamma$ se mění v závislosti na metodě (strategii), jakou je definována nepodobnost shluků. Lance a Williams odvodili tyto koeficienty pro metodu nejbližšího souseda, nejvzdálenějšího souseda, centroidní, mediánovou, průměrné nepodobnosti a navíc navrhli tzv. pružnou strategii, v níž jsou koeficienty voleny libovolně, ale musí splňovat určité podmínky. Wishart rozšířil skupinu kombinatorických metod o metodu Wardovu. Všechny zde uvedené metody přiblížím v následujících odstavcích. [8]

V souvislosti s hierarchickým shlukováním je ještě potřeba objasnit pojem dendogram. Dendogram představuje grafické znázornění výsledků hierarchické shlukové analýzy. Představuje vlastně jakýsi strom, ve kterém dochází ke sloučení dvou shluků ve výsledný jeden, tedy spojení dvou větví stromu v jednu. Ve výsledku by měla zůstat právě jediná větev. Dendogram může být vertikální nebo horizontální. V praktické části je použit dendogram vertikální, kde na ose x jsou jednotlivé shlukovací sekvence a na ose y je nanesena vzdálenost mezi shluky. [12]

## 4.1 Metoda nejbližšího souseda

Metodu nejbližšího souseda poprvé uvedl Sneath pod názvem „simple linkage“. [8]

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 stadium nejkratší spojkou. [10]

Metodu nejbližšího souseda lze definovat tímto způsobem [11]:

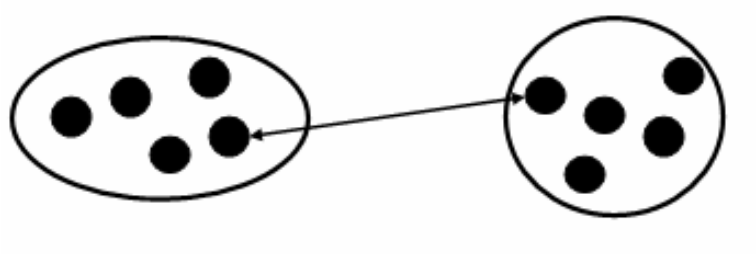
Jestliže D je libovolný koeficient nepodobnosti, symboly A, B jsou dva různé shluky, které náleží do rozkladu $\Xi_k$, objekt $x_i$ patří do shluku A a objekt $x_j$ do shluku B, pak je

$$D_{SL}(A, B) = \min\left\{D(x_i, x_j)\right\} \tag{37}$$

předpisem určujícím vzdálenost shluků A a B v metodě nejbližšího souseda. [11]

Rekurzivní schéma, kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. [8]:

$$D(U,R) = 0{,}5D(U,P) + 0{,}5D(U,Q) - 0{,}5\left|D(U,P) - D(U,Q)\right| \tag{38}$$



**Obrázek 6 - metoda nejbližšího souseda; zdroj: podle [10]**

## 4.2 Metoda nejvzdálenějšího souseda

Tato metoda, jejímž autorem je Sorrensen, je známa též pod názvem „complete linkage“. [8]

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 taky někdy nazývá metodou úplného propojení, protože všechny objekty ve shluku jsou propojeny každý s každým při maximální vzdálenosti čili minimální podobnosti. [10]

Metoda nejvzdálenějšího souseda je tedy ve své podstatě „opakem“ metody nejbližšího souseda. Lze ji definovat tímto způsobem [11]:

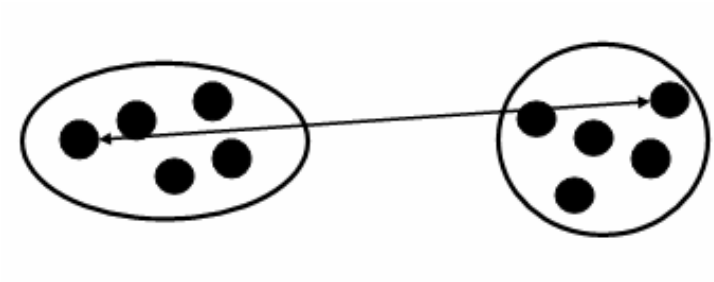
Jestliže D je libovolný koeficient nepodobnosti, symboly A, B jsou dva různé shluky, které náleží do rozkladu $\Xi_k$, objekt $x_i$ patří do shluku A a objekt $x_j$ do shluku B, pak je

$$D_{CL}(A,B) = \max\left\{D(x_i, x_j)\ \right\} \tag{39}$$

předpisem určujícím vzdálenost shluků pro metodu nejvzdálenějšího souseda. [11]

Rekurzivní schéma, kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. [8]:

$$D(U,R) = 0{,}5D(U,P) + 0{,}5D(U,Q) + 0{,}5\left|D(U,P) - D(U,Q)\right| \tag{40}$$



**Obrázek 7 - metoda nejvzdálenějšího souseda; zdroj: podle [10]**

## 4.3 Centroidní metoda

Této metody poprvé použili Sokal a Michener pod názvem „weighted group method“. Autoři vyšli z geometrického modelu v prostoru $\Xi_k$ a vyjádřili nepodobnost dvou shluků euklidovskou vzdáleností jejich těžišť. [8]

Jde o vzdálenost dvou těžišť shluků, vyjádřených euklidovskou vzdáleností nebo čtvercem euklidovské 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 [10]. Předpis určující vzdálenost dvou navzájem různých shluků lze definovat následujícím způsobem [11]:

Nechť symbol $q_E^2$ představuje čtverec euklidovské metriky a vektory $\bar{x}^{(A)}$ a $\bar{x}^{(B)}$ představují centroidy shluků A resp. B. Pak

$$D_{WGM}(A,B) = q_E^2(\bar{x}^{(A)}; \bar{x}^{(B)}) \tag{41}$$

je předpisem určujícím vzdálenost shluků A a B pro centroidní metodu. Jednotlivé složky centroidů

$\bar{x}^{(A)}$ a $\bar{x}^{(B)}$ ve vzorci se určí jako:

$$\bar{x}_j^{(A)} = \frac{1}{n_A}\sum_{i=1}^{n_A} x_{ij}^{(A)} \quad \text{pro j = 1,2,…,p} \tag{42}$$

a

$$\bar{x}_j^{(B)} = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{ij}^{(B)} \quad \text{pro j = 1,2,…,p.} \tag{43}$$

Symboly $n_a$ a $n_b$ označují počty jednotlivých objektů v daných shlucích. Symbol $x_{ij}^{(A)}$ představuje hodnotu j-tého znaku i-tého objektu, který patří do shluku A. Analogicky pak pro $x_{ij}^{(B)}$. [11]

Rekurzivní schéma, kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. [8]:

$$D(U,R) = \frac{|P|}{|R|}D(U,P) + \frac{|Q|}{|R|}D(U,Q) - \frac{|P|*|Q|}{|R|^2}D(P,Q) \tag{44}$$

## 4.4 Metoda průměrné vazby

V anglické literatuře nazývána jako „average linkage method“. [11]

Kritérium vzniku shluků je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku [10]. Vzdálenost mezi dvěma shluky lze definovat následujícím způsobem [11]:

Jestliže D je libovolný koeficient nepodobnosti, $n_a$ a $n_b$ jsou počty objektů ve shlucích A a B, objekt $x_i$ patří do shluku A a objekt $x_j$ do shluku B, pak

$$D_{AL}(A,B) = \frac{1}{n_A n_B}\sum_{x_i \in A}^{n_A}\sum_{x_{ji} \in B}^{n_B} D(x_i; x_j) \tag{45}$$

Je předpisem určujícím vzdálenost shluků A a B pro metodu průměrné vazby. Jinými slovy, vzdálenost mezi jednotlivými dvěma shluky je definována jako průměrná vzdálenost mezi všemi dvojicemi objektů, které náleží do dvou různých shluků. [11]

Rekurzivní schéma, kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. [8]:

$$D(U,R) = \frac{|P|}{|R|} D(U,P) + \frac{|Q|}{|R|} D(U,Q) \tag{46}$$

## 4.5 Mediánová metoda

Tuto metodu uvedl poprvé Gower pod názvem „unweighted group method“. Důvodem pro zavedení metody byla snaha odstranit jistý nedostatek centroidní metody. Gower totiž konstatoval, že vlivem rozdílných počtů objektů ve shlucích vede k potlačení vlastností malých shluků. [8] Vzdálenost mezi dvěma shluky lze definovat následovně [11]:

Nechť D představuje libovolný koeficient nepodobnosti a vektory $\bar{x}_{50}^{(A)}$ a $\bar{x}_{50}^{(B)}$ představují mediány shluků A resp. B. Pak

$$D_{UWGM}(A,B) = D(\bar{x}_{50}^{(A)}; \bar{x}_{50}^{(B)}) \tag{47}$$

je předpisem určujícím vzdálenost shluků A a B pro mediánovou metodu. U mediánové metody je tedy za míru vzdálenosti dvou shluků považována vzdálenost jejich mediánů. [11]

Rekurzivní schéma, kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. [8]:

$$D(U,R) = 0{,}5D(U,P) + 0{,}5D(U,Q) - 0{,}25D(P,Q) \tag{48}$$

## 4.6 Lanceova a Williamsova „pružná strategie“

Lance a Williams provedli řadu pokusů s různými hodnotami koeficientů $\alpha_i, \alpha_j, \beta$ a $\gamma$ rekurzivního schématu. Dospěli k závěru, že nejlepší výsledky dává metoda, v níž je možno nepodobnost nově vzniklého shluku s ostatními shluky definovanou tímto schématem měnit vhodnými volbami koeficientů $\alpha_i, \alpha_j, \beta$ a $\gamma$, přičemž je třeba dodržet tyto podmínky [8]:

$$\alpha_i = \alpha_j \tag{49}$$

$$\alpha_i + \alpha_j + \beta = 1 \tag{50}$$

$$\beta < 1 \tag{51}$$

$$\gamma = 0 \tag{52}$$

Z uvedených podmínek plyne, že

$$\alpha_i = \alpha_j = \frac{1-\beta}{2} \tag{53}$$

Rekurzivní schéma, kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. [8]:

$$D(U,R) = \frac{1-\beta}{2} D(U,P) + \frac{1-\beta}{2} D(U,Q) + \beta D(P,Q) \tag{54}$$

## 4.7 Wardova – Wishartova metoda

V anglické literatuře také „Ward’s error of squares method“. [11]

Ward navrhl měření nepodobnosti shluků přírůstkem $I_{pq}$ hodnoty „cílové funkce E“, který provází sjednocení shluků *P* a *Q*. Označí-li se $P \cup Q = R$, je přírůstek cílové funkce

$$I_{pq} = E_r - E_p - E_q \tag{55}$$

kde $E_p$, $E_q$, $E_r$ jsou hodnoty cílové funkce shluků *P, Q, R*. Funkce $E_A$ shluku $A = \{A_1, A_2, ..., A_i\}$ tvořeného t objekty $A_i = (a_{i1}, a_{i2}, ..., a_{ip}), i = 1,2,...,t$, je přitom definována takto [8]:

$$E_A = \sum_{i=1}^{t} \sum_{j=1}^{p} \left(a_{ij} - \overline{a_j}\right)^2 \tag{56}$$

$$\overline{a_j} = \frac{1}{t} \sum_{i=1}^{t} a_{ij} \tag{57}$$

Rekurzivní schéma, kde *D* je nepodobnost, *U* původní shluk, *R* je nový shluk sloučený ze shluků *P* a *Q*. [8]:

$$D(U,R) = \frac{1}{|R| + |U|}\left[(|U| + |P|)D(U,P) + (|U| + |Q|)D(U,Q) - |U|D(P,Q)\right] \tag{58}$$

# 5 Příklad

## 5.1 Výchozí data

Jak je uvedeno v kapitole Předmět a cíl práce, tak samotný příklad bude řešen za pomocí softwaru MS Excel, Unistat a Statistica 7 CZ. Pro modelový příklad byla vybrána data, která se týkají obsahu jednotlivých látek v konkrétních potravinách. Samotná data byla vybrána z webových stránek týkajících se Crohnovy nemoci www.crohn.cz. Data byla vybrána z toho důvodu, že je z nich patrné rozčlenění do různých kategorií a proto bude zajímavé srovnání s nově vzniklými shluky. Dalším důvodem bylo to, že jsou tato data srozumitelná pro všechny studenty.

Základní data obsahují 40 objektů s 11 vlastnostmi. Všechny vlastnosti udávají obsah na 100 gramů potraviny nebo v případě nápoje na 100 ml. Vlastnosti jsou tyto:

- Kalorie – uvádí počet kalorií
- Bílkoviny – uvádí obsah bílkovin v g
- Tuky – uvádí obsah tuků v g
- Sacharidy – uvádí obsah sacharidů v g
- Ca – uvádí obsah vápníku v mg
- P – uvádí obsah fosforu v mg
- Fe – uvádí obsah železa v mg
- B1 – uvádí obsah vitamínu B1 v mg
- B2 – uvádí obsah vitamínu B2 v mg
- PP – uvádí obsah vitamínu PP v mg
- C – uvádí obsah vitamínu C v mg

Objekty jsou rozděleny do 4 kategorií. Jsou to kategorie:

- Zdroje sacharidů
- Zdroje živočišných bílkovin
- Maso
- Ovoce a zelenina

**Tabulka 21 - základní data pro praktickou část**

| Č. | Potravina | Kalorie | Bílkoviny g | Tuky g | Sacharidy g | Ca mg | P mg | Fe mg | B1 mg | B2 mg | PP mg | C mg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Brambory | 80 | 2 | 0 | 19 | 13 | 72 | 0,8 | 0,07 | 0,04 | 1,2 | 10 | Zdroje sacharidů |
| 2 | Čočka | 330 | 25 | 1 | 60 | 59 | 423 | 7,5 | 0,56 | 0,24 | 2,2 | 0 | |
| 3 | Fazole | 331 | 21 | 1,6 | 62 | 137 | 437 | 6,9 | 0,67 | 0,23 | 3,1 | 0 | |
| 4 | Jáhly | 356 | 10,6 | 2,9 | 71,4 | 33 | 269 | 5,9 | 0,26 | 0,1 | 3,6 | 0 | |
| 5 | Mouka hlad. | 354 | 10,4 | 1,3 | 74,3 | 25 | 121 | 0,6 | 0,15 | 0,03 | 2 | 0 | |
| 6 | Oves.vločky | 386 | 13 | 7,5 | 67,8 | 56 | 397 | 3,8 | 0,63 | 0,14 | 0,9 | 0 | |
| 7 | Rýže | 354 | 6,7 | 0,7 | 78,9 | 24 | 135 | 0,8 | 0,07 | 0,03 | 1,6 | 0 | |
| 8 | Kmín.chléb | 237 | 5,2 | 0,8 | 51,8 | 20 | 143 | 0,9 | 0,11 | 0,04 | 1 | 0 | |
| 9 | Těstoviny | 366 | 12 | 2,2 | 74,1 | 25 | 153 | 1 | 0,1 | 0,6 | 1,5 | 0 | |
| 10 | 0,5 l Piva 10 | 140 | 1,5 | 0 | 8,5 | 45 | 75 | 0 | 0,05 | 0,35 | 5 | 0 | |
| 11 | Tvaroh tučný | 1175 | 13,7 | 12 | 2,8 | 366 | 253 | 0,3 | 0,02 | 0,28 | 0,1 | 0 | Zdroje živočišných bílkovin |
| 12 | Eidam 30% | 259 | 29,2 | 14,6 | 1,8 | 669 | 427 | 0,29 | 0,048 | 0,332 | 0,1 | 0 | |
| 13 | Hermelín | 270 | 20,2 | 20,2 | 1,6 | 157 | 330 | 0,5 | 0,04 | 0,42 | 0,3 | 0 | |
| 14 | Jogurt | 101 | 5,7 | 4,5 | 9,7 | 180 | 135 | 0,4 | 0,32 | 0,48 | 0,8 | 0 | |
| 15 | Mléko 2% | 48 | 3,2 | 2 | 4,4 | 112 | 101 | 0,1 | 0,04 | 0,06 | 0,1 | 0 | |
| 16 | Vejce 1 kus | 79 | 6,5 | 5,5 | 0 | 30 | 111 | 1,01 | 0,05 | 0,151 | 0,05 | 0 | |
| 17 | Hovězí libové | 159 | 20,8 | 7,8 | 0 | 8 | 152 | 3,3 | 0,1 | 0,22 | 5,5 | 0 | Maso |
| 18 | Kuře (s kostí) | 78 | 14,2 | 2 | 0 | 8 | 126 | 0,94 | 0,063 | 0,101 | 5,1 | 0 | |
| 19 | Kachna-kost | 183 | 11,4 | 15 | 0 | 8 | 150 | 3,25 | 0,052 | 0,124 | 3,7 | 0 | |
| 20 | Králík | 113 | 14,8 | 5,5 | 0 | 12 | 88 | 0,91 | 0,056 | 0,042 | 5,32 | 0 | |
| 21 | Makrela | 114 | 11,4 | 7,3 | 0 | 23 | 146 | 0,73 | 0,049 | 0,128 | 1,65 | 0 | |
| 22 | Sardinky | 335 | 21,1 | 27 | 0 | 354 | 434 | 3,5 | 0,01 | 0,14 | 3,9 | 0 | |
| 23 | Celer | 26 | 0,9 | 0,2 | 5,5 | 31 | 31 | 0,31 | 0,031 | 0,025 | 0,25 | 2,7 | Ovoce a zelenina |
| 24 | Cibule | 19 | 0,4 | 0,1 | 4,4 | 33 | 10 | 0,37 | 0,025 | 0,016 | 0,2 | 7,4 | |
| 25 | Česnek | 110 | 1,7 | 0,1 | 26 | 28 | 46 | 0,46 | 0,019 | 0,028 | 0,74 | 16,7 | |
| 26 | Kedlubny | 25 | 1,6 | 0,2 | 4,6 | 34 | 38 | 0,45 | 0,045 | 0,038 | 0,15 | 30 | |
| 27 | Kopr | 66 | 2,6 | 14,1 | 30 | 50 | 0,5 | 0,5 | 0,05 | 0,08 | 0,5 | 100 | |
| 28 | Papriky zel. | 20 | 0,9 | 0,2 | 4 | 4 | 19 | 0,6 | 0,03 | 0,038 | 0,68 | 90 | |
| 29 | Špenát | 17 | 1,5 | 0,2 | 2,7 | 57 | 38 | 2,1 | 0,077 | 0,14 | 0,42 | 31,5 | |
| 30 | Zelí | 20 | 1,4 | 0,3 | 3,4 | 45 | 18 | 0,4 | 0,056 | 0,04 | 0,24 | 17,7 | |
| 31 | Banány | 59 | 0,8 | 0,1 | 15,4 | 5 | 19 | 0,4 | 0,027 | 0,034 | 0,47 | 6,7 | |
| 32 | Broskve | 41 | 0,7 | 0,2 | 10,4 | 7 | 18 | 0,53 | 0,018 | 0,044 | 0,79 | 7 | |
| 33 | Citróny | 23 | 0,2 | 0 | 6,3 | 21 | 9 | 0,3 | 0,036 | 0 | 0,06 | 24 | |
| 34 | Jablka | 50 | 0,3 | 0,4 | 12,9 | 6 | 10 | 0,44 | 0,035 | 0,026 | 0,18 | 6,2 | |
| 35 | Jahody | 36 | 0,8 | 0,5 | 8 | 27 | 29 | 0,77 | 0,029 | 0,067 | 0,29 | 57,6 | |
| 36 | Melouny | 16 | 0,5 | 0,1 | 4 | 13 | 11 | 0,26 | 0,033 | 0,02 | 0,39 | 5,3 | |
| 37 | Pomeranče | 32 | 0,6 | 0,1 | 8,1 | 24 | 18 | 0,29 | 0,058 | 0,022 | 0,14 | 37,4 | |
| 38 | Třešně | 54 | 1 | 0,4 | 13,3 | 16 | 18 | 0,36 | 0,046 | 0,055 | 0,36 | 7,3 | |
| 39 | Víno hrozny | 61 | 0,7 | 0,4 | 15,5 | 19 | 18 | 0,55 | 0,055 | 0,037 | 0,18 | 3,7 | |
| 40 | Švestky | 59 | 0,7 | 0,2 | 15,4 | 16 | 21 | 0,56 | 0,056 | 0,038 | 0,47 | 3,8 | |

## 5.2 Postup řešení

Data z tabulky 20 byla zpracována čtyřmi různými metodami hierarchického shlukování:

- metodou nejbližšího souseda,
- metodou nejvzdálenějšího souseda,
- mediánovou metodou,
- centroidní metodou.

Pro všechny metody byla použita k výpočtům Euklidovská vzdálenost, v případě metody nejbližšího souseda pro srovnání i čtverec euklidovské vzdálenosti. Všechny výpočty byly prováděny z normalizovaných dat, pouze v případě metody nejbližšího souseda byly výpočty pro porovnání provedeny zároveň i bez normalizace, tedy ze standardizovaných dat.
Z toho plyne, že celkem bylo provedeno 6 shlukovacích procesů (příkladů) se 6 výsledky.

**Příklad 1:** Metoda nejbližšího souseda – Euklidovská vzdálenost (normalizovaná matice)
**Příklad 2:** Metoda nejbližšího souseda – čtverec euklidovské vzdálenosti (normalizovaná matice)
**Příklad 3:** Metoda nejvzdálenějšího souseda - Euklidovská vzdálenost (normalizovaná matice)
**Příklad 4:** Metoda centroidní - Euklidovská vzdálenost (normalizovaná matice)
**Příklad 5:** Metoda mediánová - Euklidovská vzdálenost (normalizovaná matice)
**Příklad 6:** Metoda nejbližšího souseda – Euklidovská vzdálenost (jako výchozí data byla použita standardizovaná matice)

Jednotlivé příklady jsou uloženy v samostatných sešitech MS Excel. Každý sešit obsahuje 9 listů.:

- **Základní data** – obsahuje výchozí data (viz. tabulka 20)
- **Popisná statistika** – ze základních dat byla provedena jejich analýza (konkrétně popisná statistika)
- **Korelace** – korelace vlastností a korelace objektů
- **Standardizace** – ze základních dat vypočtená standardizovaná matice
- **Normalizace** – ze standardizovaných dat vypočtená normalizovaná matice
- **Vzdálenost** – použitá vzdálenost pro shlukování
- **Metoda** – tento list obsahuje průběh a rozvrh shlukování
- **Rozdělení do shluků** – v tomto listě jsou uvedeny výsledné shluky
- **Dendogram** – grafické znázornění shlukování

## 5.3 Průběh shlukování

Ukázka průběhu shlukování bude vysvětlena na příkladu 2. Bude vysvětlen postup 10 krocích shlukování.

**Tabulka 22 - data pro ukázku průběhu shlukování, zdroj [autor]**

| Krok | Spojení1 | Spojení2 | Vzdálenost |
|---|---|---|---|
| 1 | Víno hrozny | Švestky | 0,04399 |
| 2 | Broskve | Víno hrozny | 0,08235 |
| 3 | Papriky zelené | Banány | 0,22276 |
| 4 | Kedlubny | Broskve | 0,24217 |
| 5 | Cibule | Jablka | 0,27560 |
| 6 | Česnek | Jahody | 0,28427 |
| 7 | Špenát | Zelí | 0,30141 |
| 8 | Králík | Makrela | 0,46608 |
| 9 | Celer | Pomeranče | 0,49445 |
| 10 | Celer | Kedlubny | 0,58817 |

**Krok 1** – spojením objektů víno hrozny a švestky vznikne nový shluk víno hrozny (1)

**Krok 2** – spojením objektu broskve s nově vytvořeným shlukem víno hrozny (1) vznikne nový shluk broskve (1) obsahující objekty broskve, víno hrozny, švestky

**Krok 3** – spojením objektu kedlubny a nově vytvořeného shluku broskve (1) vznikne nový shluk kedlubny (1) obsahující objekty kedlubny, broskve, víno hrozny, švestky

**Krok 4** – spojením objektů papriky zelené a banány vznikne nový shluk papriky zelené (1)

**Krok 5** - spojením objektů cibule a jablka vznikne nový shluk cibule (1)

**Krok 6** - spojením objektů česnek a jahody vznikne nový shluk česnek (1)

**Krok 7** – spojením objektů špenát a zelí vznikne nový shluk špenát (1)

**Krok 8** – spojením nového shluku kedlubny (1) a objektu pomeranče vznikne nový shluk kedlubny (2) obsahující objekty kedlubny, broskve, víno hrozny, švestky, pomeranče

**Krok 9** – spojením nového shluku špenát (1) a objektu citróny vznikne nový shluk špenát (2) obsahující objekty špenát, zelí, citróny

**Krok 10** - spojením nového shluku cibule (1) a shluku kedlubny (2) vznikne nový shluk cibule (2) obsahující objekty cibule, jablka, kedlubny, broskve, víno hrozny, švestky, pomeranče

## 5.4 Výsledky shlukování

Vzhledem k rozdělení základních dat do 4 kategorií bylo uvažováno během shlukování u všech metod rozdělení do 4 shluků. Dalo by se říct, že samotné výsledky shlukování jsou velmi překvapivé, protože původní rozložení dat na kategorie se v žádném případě nepotvrdilo. Kromě metody nejvzdálenějšího souseda, kdy byla data rozdělena do 4 poměrně velikostně souměrných shluků, u všech ostatních metod vznikly jako výsledek velmi nepravděpodobné a nesouměrně veliké shluky, kdy například jeden shluk obsahuje 37 objektů a zbylé tři shluky mají shodně po jednom objektu.

Co se týká srovnání shlukování z normalizovaných nebo standardizovaných dat, tak v tomto případě nebyl rozdíl zcela jednoznačný. Výsledné rozdělení do shluků bylo pro oboje data velmi podobné, rozdíl byl pouze v několika objektech. Nicméně vždy je lepší provádět normalizaci dat.

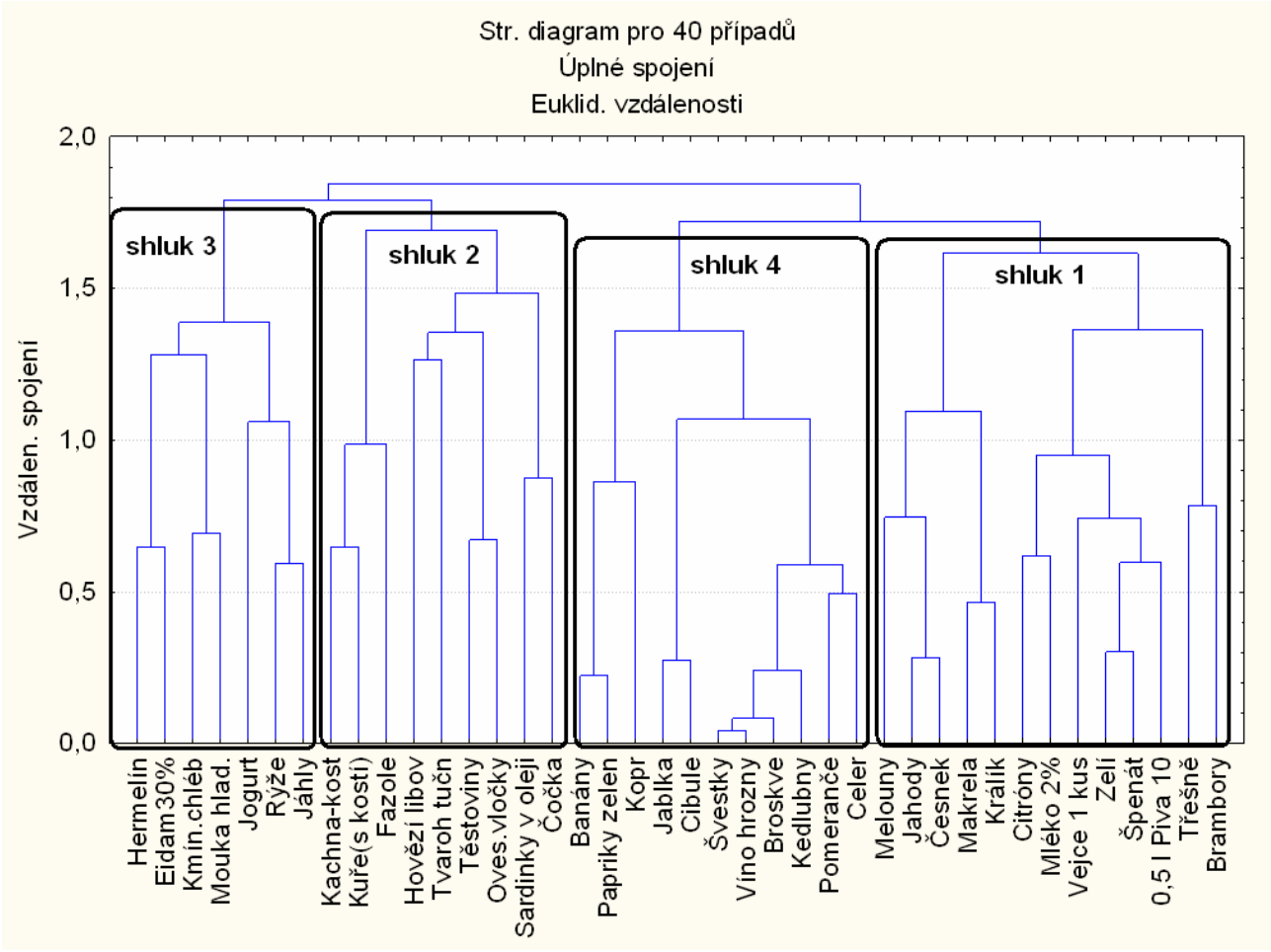
Ze všech shlukovacích procedur vyšla pro normalizovaná data nejlépe metoda nejvzdálenějšího souseda, kdy rozdělení do shluků bylo celkem podobné původnímu předpokladu. Proto je pro ukázku vložena tabulka výsledných shluků provedena metodou nejvzdálenějšího souseda. Všechny ostatní výsledky včetně postupů jsou uvedeny v příloze na CD.

**Tabulka 23 - vytvořené shluky metodou nejvzdálenějšího souseda, zdroj [autor]**

| shluk 1 | shluk 2 | shluk 3 | shluk 4 |
|---|---|---|---|
| Brambory | Čočka | Jáhly | Celer |
| 0,5 l Piva 10 | Fazole | Mouka hlad. | Cibule |
| Mléko 2% | Oves.vločky | Rýže | Kedlubny |
| Vejce 1 kus | Těstoviny | Kmín.chléb | Kopr |
| Králík | Tvaroh tučný | Eidam30% | Papriky zelené |
| Makrela | Hovězí libové | Hermelín | Banány |
| Česnek | Kuře(s kostí) | Jogurt | Broskve |
| Špenát | Kachna-kost | | Jablka |
| Zelí | Sardinky v oleji | | Pomeranče |
| Citróny | | | Víno hrozny |
| Jahody | | | Švestky |
| Melouny | | | |
| Třešně | | | |

## 5.5 Dendogram

Jak bylo uvedeno v teoretické části, dendogram je grafické znázornění postupu shlukování. Na obrázku 8 je graficky pomocí dendogramu znázorněn výsledek shlukování v příkladě 3, tedy metody nejvzdálenějšího souseda ve spojení s Euklidovskou vzdáleností. Pro lepší orientaci jsou do dendogramu zakresleny jednotlivé shluky.




**Obrázek 8 - dendogram použitý z příkladu 3, zdroj [autor]**

# Závěr

Cílem této práce bylo vytvořit pomocné materiály ke studiu předmětu KZMSA, který by se dal zařadit vzhledem ke své obtížnosti do kategorie náročnějších. Tím spíše je složitější pro studenty, kteří studují formou dálkového studia, tudíž nemají možnost danou látku dostatečně procvičit na seminářích. Snaha při zpracovávání této bakalářské práce byla taková, aby text byl co nejvíce srozumitelný a modelové příklady co nejnázornější.

V praktické části bylo provedeno 6 shlukovacích procesů, z nichž vyšly poměrně překvapivé výsledky. Kromě metody nejvzdálenějšího souseda, kde rozložení výsledných shluků bylo pravidelné a pravděpodobné, u ostatních metod převažoval jeden shluk svou velikostí nad ostatními. Proto je dobré pro řešení příkladu zkusit více metod a vybrat tu nejvhodnější. V případě používání různých metrik bylo zjištěno, že na výsledných shlucích se nic nemění, to znamená, že volba metriky neovlivňuje správnost shlukování. Pro některé metody je však konkrétní metrika doporučena (viz. kapitola 4). Pro ukázku bylo provedeno shlukování metodou nejbližšího souseda s Euklidovskou vzdáleností a s čtvercem euklidovské vzdálenosti. Změní se pouze vzdálenosti mezi jednotlivými objekty, jinak je výsledek shlukování totožný.

Pro samotné shlukování byly použity dva druhy softwarů, Unistat a Statistica 7. Práce v prostředí Unistatu byla jednodušší pro seznámení s prostředím, protože Unistat je vlastně nadstavbou Excelu, jehož základní znalost je dnes považována za samozřejmost. Je zde dobře zpracováno exportování dat, na druhou stranu výstupy z Unistatu nejsou mnohdy úplně ideální, protože nejsou dostatečně podrobné. Software Statistica 7 je o hodně náročnější na samotnou práci a orientaci v jeho prostředí, nicméně nabízí mnohem více funkcí a výstupy z něj jsou precizně zpracované. Nabízejí se i možnosti volby nejrůznějších rozvrhů shlukování, je k dispozici většina známých metrik a mnoho jiných užitečných funkcí, které byly využity k řešení této bakalářské práce. Při zpracovávání praktické části práce byly použity výstupy z obou výše uvedených softwarů doplněné analýzou dat vytvořenou pomocí Excelu.

Pro větší část široké veřejnosti je shluková analýza zcela neznámým pojmem, přestože je jen těžko možné v dnešní době najít vědní obor, v němž by metody shlukové analýzy nenašly své opodstatněné uplatnění.

# Použitá literatura


[1] HENDL, Jan. *Přehled statistických metod zpracování dat : Analýza a metaanalýza dat*. Praha : Portál s.r.o., 2004. 583 s

[2] *Wikipedie* [online]. 2001 [cit. 2008-11-25]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Hlavn%C3%AD_strana>.

[3] KUBANOVÁ, Jana. *Statistické metody pro ekonomickou a technickou praxi*. Bratislava : Statis, 2004. 249 s.

[4] CYHELSKÝ, Lubomír. *Úvod do teorie statistiky*. Praha : SNTL, 1981. 347 s.

[5] SYNEK. *Statistika* [online]. c2004 [cit. 2009-03-30]. Dostupný z WWW: <fzp.ujep.cz/~synek/statistika/skripta/DiscStat2.doc>.

[6] ŠŤASTNÝ, František. *Zpracování experimentálních dat. Skripta* [online]. 1997 [cit. 2008-11-30]. Dostupný z WWW: <http://amper.ped.muni.cz/jenik/nejistoty/frst_zed.pdf>.

[7] ŘEZÁNKOVÁ, Hana, HÚSEK, Dušan, SNÁŠEL, Václav. *Shluková analýza dat*. Příbram : Professional Publishing, 2007. 196 s.

[8] LUKASOVÁ, Alena, ŠARMANOVÁ, Jana. *Metody shlukové analýzy*. [s.l.] : [s.n.], 1985. 145 s.

[9] HYNAR, Martin. *Metody shlukování* [online]. 2003 [cit. 2009-03-31]. Dostupný z WWW: <http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0304/Shlukovani1-text.pdf>.

[10] MELOUN, Milan, MILITKÝ, Jiří. *Přednosti analýzy shluků ve vícerozměrné statistické analýze* [online]. 2005 [cit. 2009-03-30]. Dostupný z WWW: <http://meloun.upce.cz/docs/publication/152.pdf>.

[11] Metody shlukové analýzy. *Katedra aplikované matematiky a informatiky*. [Online] 2005-2006. [cit. 2009-02-21]. Dostupný z WWW: <http://www2.zf.jcu.cz/public/departments/kmi/MSMT_05/metody%20shlukove%20analyzy.pdf >

[12] Cluster Analysis. *Elektronická učebnice StatSoft*. [Online] StatSoft, 2008. [cit. 2009-02-14]. Dostupný z WWW: <http://www.statsoft.com/textbook/stcluan.html >

[13] SEGER, Jan. *Statistické metody pro ekonomy průmyslu*. Praha : SNTL, 1988. 548 s.