UNIVERZITA PARDUBICE
FAKULTA EKONOMICKO - SPRÁVNÍ

BAKALÁŘSKÁ PRÁCE

2009 JAN ZAJÍC

Univerzita Pardubice
Fakulta ekonomicko - správní

Zpracování podkladů pro praktickou část distanční opory pro předmět KZMSA – část nehierarchické shlukování

Jan Zajíc

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 ZAJÍC**
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 nehierarchické shlukování**

Zásady pro vypracování:

Metody nehierarchického 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**

L.S.

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

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

V Pardubicích dne 6. října 2008

Prohlašuji:

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 1.5.2009 Jan Zajíc

Anotace

Práce popisuje a vysvětluje jednotlivé metody nehierarchického shlukování. Je určena jako podklad pro praktickou část distanční oporu k předmětu KZMSA a cílem práce je poskytnutí přehledu o jednotlivých metodách a popis problematiky analýzy dat od jejich zadání až po závěrečné výstupy. Jednotlivé kroky byly prováděny v různých programech, umožňujících analýzu dat. Nedílnou součástí práce je CD, kde se nachází velké množství grafických výstupů.

Klíčová slova

shluková analýza , nehierarchické shlukovací metody, shlukování dat, analýza shluků, metody Kmeans, analýza dat.

Title

Processing of materials for the practical part of the remote support of the KZMSA subject - part nonhierarchical clustering

Anotation

Particular nonhiearchic clustering methods are described and explained in this thesis. The thesis is destined for remote support of the KZMSA subject. The main purpose of this thesis is to afford a summary of particular methods and description of a question of data analysis from their specification to final output. Particular steps were done in different programs that enable data analysis. The CD, in which we can find a lot of graphic output, is an important part of the thesis.

Keywords

cluster analysis, nonhiearchic clustering methods, clustering, Kmeans methods, data analysis.

1. Úvod

Situací, kdy je potřeba setřídit objekty do skupin na základě jejich podobnosti, se v běžném životě vyskytuje mnoho a není složité je na základě znalostí či intuice řešit bez větších problémů. Pokud je potřeba úlohy automatizovat tak, aby je bylo možné analyzovat pomocí programů, zobrazovat výstupy nebo na jednotlivé postupy sestavovat algoritmy, je to již problém složitější. Danou problematikou se v praxi zabývá shluková analýza. Hlavním cílem shlukové analýzy je shlukování objektů sobě podobných do shluků. Snahou je, aby ve shluku vždy byly objekty sobě podobné, podobnější než objekty v jiných skupinách, shlucích. Pro příklad je uvedena jedna z mnoha definic shlukové analýzy. Jak uvádí R.C.Tryon (1939): „Shluková analýza je obecný logický postup formulovaný jako procedura, pomocí níž seskupujeme objektivně jedince do skupin na základě jejich podobností a rozdílností.“ [16]

V praxi se používá velké množství shlukovacích metod, které se dělí na dvě velké skupiny a to nehierarchické a hierarchické shlukovací metody. Cílem práce je popsat postup při shlukování dat pomocí nehierarchických metod. Objasnění jednotlivých kroků, jež vedou k řešení zadaného problému. Počínaje výběrem dat, jejich přípravou a samotným shlukováním. Porovnání jednotlivých metod a popis získaných výstupů. K analýze dat byl využit program Shluk, který je poskytován školou, a 30denní zkušební verze programu Statistica. Součástí práce je i návrh souhrnného příkladu. Na tomto návrhu je prakticky vysvětlen postup zpracování zadané úlohy. Jsou vysvětleny a prakticky provedeny veškeré nutné úpravy nutné k řešení samotného příkladu. Na navrženém příkladu jsou vysvětleny jednotlivé metody nehierarchického shlukování.

2. Předzpracování dat

2.1. Statistický soubor

Základní soubor obsahuje všechny jednotky, respektive všechna získaná data. V praxi je práce s celým základním souborem velice náročná, a to právě pro jeho obsáhlost. Zpracování velkého objemu dat zvyšuje nároky na software i hardware a snižuje přehlednost dat. V praxi je nejčastěji pracováno s výběrovým souborem, který obsahuje vhodně zvolený vzorek ze základního souboru. Práce s ním je jednodušší a rychlejší. Je velice důležité zvolit vhodný vzorek, náhodný výběr, aby nedošlo ke zkreslení dat. K tomu slouží statistické metody. Základní a výběrový soubor se dohromady nazývá statistickým souborem. Před samotným předzpracováním jsou data tříděna a řazena. Třídění a řazení může být podle zkoumaných dat intervalové (spojitý či diskrétní znak), nebo prosté. Statistické soubory jsou děleny nejen na základní a výběrové, ale dále na soubory jednorozměrné, dvojrozměrné, uspořádané a neuspořádané.

2.1.1. Popis objektů a jejich znaků

Objektem je myšlen útvar, který je popsán konečným počtem charakteristik. Nelze-li objekt takto popsat, je vybrán konečný počet znaků, které nejlépe daný objekt charakterizují. Pří výběru těchto znaků je uvažován cíl projektu a s ohledem na něj jsou znaky vybírány. Musí mít význam a smysl při řešení dané problematiky. Před samotným shlukováním je potřeba objekty změřit, zvážit, určit jejich významné znaky či podobnosti a nesrovnalosti. Objekty které jsou pozorovány jsou předměty nebo jevy. Podle způsobu, jakým jsou data popsána, jsou rozdělena na kvalitativní či kvantitativní. Vstupní matice dat, se kterou je dále pracováno, je tvořena objekty a jejich vlastnostmi. [16]

2.2. Popisná statistika

Popisná statistika zjišťuje informace, které dále zpracovává a analyzuje. Výsledky zobrazuje ve formě grafů a tabulek a vypočítává číselné charakteristiky jednotlivých objektů, např. průměr, rozptyl, rozpětí apod. Popisná statistika umožňuje získat základní přehled o zkoumaných datech ze zvoleného výběru. Vlastnosti statistických jednotek – objektů základního souboru, se nazývají statistické znaky, popřípadě veličiny či proměnné. Tyto veličiny jsou děleny na kvantitativní a kvalitativní. Kvantitativní veličiny jsou popsané číselnou hodnotou (cena, výška, váha) a kvalitativní veličiny vlastnostmi (pohlaví – muž). K nejjednodušším způsobům zobrazení popisné statistiky patří využití nástroje Analýza dat – Popisná statistika, nabízeného v programu MS Excel.

U některých veličin bývá pro lepší vypovídající schopnost místo tabulkového znázornění využíváno grafů. Volba znázornění bývá volena podle vhodnosti aplikace na konkrétní případ a také vzhledem k velikosti základního souboru či zvoleného výběru. Velice přehledným znázorněním dat může být histogram četností. Jedná se o graf, kdy se na vodorovnou osu znázorní třídy a svislou osu četnosti. Jednotlivé charakteristiky popisné statistiky se dělí podle popisované vlastnosti na tři skupiny a to: charakteristika polohy, charakteristiky variability a charakteristiky šikmosti a špičatosti. [11]

Nebude li uvedeno jinak, jsou u vzorců používány tyto zkratky:

n - rozsah základního souboru

N - rozsah výběrového souboru

Xi - pozorovaná hodnota znaku X u i−té statistické jednotky,i = 1,. . ., n.

Wi - váhy, vhodně zvolená reálná čísla, kde alespoň jedno je nenulové a která vyjadřují přesnost či význam hodnot xi. Wi >= 0

s – směrodatná odchylka

s^2 - rozptyl

$\bar{x}$ - aritmetický průměr

2.2.1 Charakteristiky polohy

Charakteristiky polohy jsou hodnoty, které lze považovat za střed, kolem kterého náhodné veličiny kolísají. Nejčastěji používanou charakteristikou polohy je střední hodnota. Často užívanými charakteristikami jsou také medián a modus.

Medián

MS EXCEL = MEDIAN (OBLAST)

Medián je hodnota, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. Platí, že (nejméně) 50% hodnot je menších nebo rovných a (nejméně) 50% hodnot je větších nebo rovných mediánu. Pro nalezení mediánu daného souboru stačí hodnoty seřadit podle velikosti a vzít hodnotu, která se nalézá uprostřed seznamu. Pokud má soubor sudý počet prvků, je za medián označován aritmetický průměr hodnot na místech *n*/2 a *n*/2+1. Za medián se dá označit více čísel. Například ve zmíněném případě sudého počtu prvků neexistuje jedinečná střední hodnota. [7]

Modus

MS EXCEL = MODE (OBLAST)

Modus je hodnota, která se v daném statistickém souboru vyskytuje nejčastěji (hodnota znaku s největší relativní četností). Představuje typickou hodnotu sledovaného souboru a jeho určení předpokládá roztřídění souboru podle obměn znaku. Výhodou modu je, že jej lze využít i v analýze nečíselných dat. [10]

Aritmetický průměr

MS EXCEL = PRŮMĚR (OBLAST)

Aritmetický průměr je statistická veličina, která vyjadřuje typickou hodnotu popisující soubor mnoha hodnot. Aritmetický průměr je zřejmě nejčastěji používaný statistický pojem. S tím ovšem souvisí i fakt, že je velice často využíván chybně. [1]

Nejčastěji vyskytující se chybou je, že je aritmetický průměr používán tam, kde by mnohem přesněji vypovídající hodnotu měla jiná statistická veličina. Například bude-li na vesnici, kde žije 50 obyvatel, bydlet jeden úspěšný podnikatel a zbytek lidi pobírající starobní důchod, bude aritmetický průměr příjmu občana této vesnice vysoké číslo. Ale přitom onoho průměrného platu, dosáhne pouze zmiňovaný podnikatel a dalších 49 lidí ne. Tedy jediná hodnota, která se výrazně odlišuje od ostatních, může velice výrazně zkreslit údaje. Aritmetický průměr je citlivý na extrémní hodnoty. Např. aritmetickým průměrem souboru { 1, 2, 2, 2, 3, 9 } je 3,2, přestože pět ze šesti hodnot tohoto souboru je menších. V obdobných případech je mnohem vhodnější použít pro vyjádření typické hodnoty medián (který je u této množiny roven dvěma, což je mnohem lepší popis typické hodnoty). [1]

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Rovnice 1 – Aritmetický průměr (zdroj: [3])

Useknutý Průměr

MS EXCEL = TRIMMEAN (OBLAST, PROCENTA)

Jak již bylo zmíněno, aritmetický průměr je náchylný na extrémní hodnoty. Zabránit ovlivnění aritmetického průměru těmito hodnotami lze vynecháním např. 10% nejnižších a 10% nejvyšších hodnot. Průměr je poté vypočten ze zbývajících hodnot a je nazýván useknutým aritmetickým průměrem. [3]

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

Rovnice 2 - Useknutý aritmetický průměr (zdroj: [3])

kde m je počet veličin, které zůstali po odebrání dolních x % a horních x %.

Kvantily

Kvantily umožňují posoudit jednak symetrii výběrového rozdělení a jednak procento prvků ve výběru. Kvantil udává, že p% hodnot souboru nabývá hodnoty stejné, nebo menší než je hodnota kvantilu. Kvantil může být použit například u testů, kde je výsledek hodnocen pomocí bodů. Pokud má například kvantil 90 % hodnotu 150 bodů a zkoušený v testu dosáhl 150 bodů, ví, že jeho hodnocení je lepší než 90 % ze všech přítomných (patří tedy mezi 10 % nejlepších). A tedy pokud by do dalšího kola postupovalo 20% zájemců, daný účastník ví, že toto kritérium splňuje).[3]

Kvartily

MS EXCEL = QUARTIL (OBLAST, KVARTIL)

Kvartily oddělují ze statistického souboru čtvrtiny. Rozlišují se dolní a horní kvartil. Udávají, že 25% respektive 75% hodnot souboru, nabývá hodnoty stejné nebo menší než je hodnota kvartilu. Některé kvartily se používají velice často a mají speciální názvy. Kvartily jsou specifickým případem kvantilů. [3]
50%-NÍ KVARTIL = MEDIÁN
75%-NÍ KVARTIL = HORNÍ KVARTIL
25%-NÍ KVARTIL = DOLNÍ KVARTIL

2.2.2. Charakteristiky variability

Charakteristiky variability určují velikost odchylek náhodné veličiny od nějaké charakteristiky polohy. Nejpoužívanějšími charakteristikami variability jsou například rozptyl nebo směrodatná odchylka. Určují také proměnlivost hodnot, některé z nich umožňují srovnání více souborů.

Variační rozpětí

Variační rozpětí je rozdíl mezi nejmenší a největší hodnotou souboru. Jedná se tedy o rozdíl maxima a minima. Stejně jako aritmetický průměr, je i variační rozpětí velice náchylné na extrémní hodnoty. Bude-li tedy základní soubor obsahovat jednu extrémní hodnotu, ať minimum či maximum, může dojít ke zkreslení výsledku a snížení vypovídající hodnoty variačního rozpětí. [3]

Rozptyl

MS EXCEL = VAR(OBLAST) – POPULAČNÍ

MS EXCEL = VAR.VÝBĚR (OBLAST) – VÝBĚROVÝ

Rozptyl je definován jako střední hodnota kvadrátů odchylek od střední hodnoty a je základní mírou rozptýlení znaků. Je běžné tuto míru rozptýlení vyjadřovat ve stejných jednotkách, jako jsou jednotky měřené veličiny. [2] Tedy měří rozptýlenost dat kolem aritmetického průměru – střední hodnoty. Jinak se rozptyl také nazývá průměrná kvadratická odchylka. [6]

Populační rozptyl se vypočte dle vztahu:

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

Rovnice 3 - Populační rozptyl (zdroj: [3])

Nebo lze použít vzorec pro rozptyl výběrový, který vychází pouze z výběru.

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

Rovnice 4 - Výběrový rozptyl (zdroj: [3])

Směrodatná odchylka

MS EXCEL = SMODCH(OBLAST)

Směrodatná odchylka je definována jako odmocnina z rozptylu. Zhruba řečeno vypovídá o tom, jak moc se od sebe navzájem liší typické případy v souboru zkoumaných čísel. Je-li hodnota směrodatné odchylky malá, jsou si prvky souboru většinou navzájem podobné, a naopak velká směrodatná odchylka signalizuje velké vzájemné odlišnosti. [13] Je-li směrodatná odchylka vypočtena z populačního výběru, je nazývána směrodatnou odchylkou populační a je dána vztahem:

$$s = \sqrt{s^2}$$

Rovnice 5 - Směrodatná odchylka (zdroj: [3])

Je-li odchylka počítána z rozptylu výběrového, je nazývána výběrovou směrodatnou odchylkou.

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

Rovnice 6 - Výběrová směrodatná odchylka (zdroj: [3])

Variační koeficient

Variační koeficient je podíl směrodatné odchylky (výběrové) a aritmetického průměru. Používá se pro porovnání variability znaků, majících odlišné jednotky nebo lišících se mírou polohy a udává se v procentech. Je-li variační koeficient větší jak 50% jedná se o velice nesourodý soubor. [3]

2.2.3. Charakteristiky rozdělení

Koeficient šikmosti

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 $\hat{x}$. [14]
Vyjde-li šikmost kladná, znamená to, že v souboru převládají nízké hodnoty. Vyjde-li šikmost záporná, převládají naopak hodnoty vysoké. Pro symetrické rozdělení platí, že hodnoty modu, mediánu a průměru jsou si rovné. [3]

Koeficient špičatosti

Charakteristiky špičatosti udávají, jaký průběh má rozdělení hodnot kolem zvoleného středu (rozdělení). Čím je rozdělení špičatější, tím víc jsou hodnoty soustředěny kolem daného středu rozdělení. Na druhé straně, rozdělení s nízkou špičatostí často obsahuje hodnoty velmi vzdálené od středu rozdělení. [14] Koeficient špičatosti udává, jak jsou hodnoty koncentrovány okolo střední hodnoty. Je-li špičatost větší jak nula, jsou hodnoty špičaté a tedy koncentrované okolo středu. V opačném případě jsou hodnoty ploché a rozptýlené okolo středu a může existovat i několik vrcholů. [3]

2.2.4. Vážené charakteristiky

Vážený průměr

Vážený průměr zobecňuje aritmetický průměr a poskytuje charakteristiku statistického souboru v případě, že hodnoty v tomto souboru mají různou důležitost, různou váhu. Používá se především při počítání aritmetického průměru souboru, který je složen z více dílčích celků. Pro výpočet váženého průměru je potřeba jednak hodnot, jejichž průměr bude spočítán, a zároveň jejich váhy. [15]

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Rovnice 7 - Vážený průměr (zdroj: [15])

Vážený rozptyl

$$s^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 * n_i}{\sum_{i=1}^{m} n_i - 1}$$

Rovnice 8 - Vážený rozptyl (zdroj: [3])

2.3. Transformace dat

Transformace dat je často prvním krokem v přípravě dat na samotné shlukování. Hodnoty jednotlivých znaků objektů jsou často v různých jednotkách. To může způsobovat, že se určité znaky jeví jako dominující a jiné znaky jen málo ovlivňují průběh shlukování. Někdy je proto výhodné data upravit tak, aby byly všechny znaky souměřitelné. Způsob, jak docílit této úpravy dat, je jejich standardizace a normalizace. Před vlastní shlukovou analýzou je potřeba řešit otázku, zda je data třeba transformovat. [5]

2.3.1. Standardizace dat

Standardizace se týká jednak znaků a jednak objektů. Standardizace dat znamená odstranění závislosti na jednotkách. Po provedené standardizaci je pomocí vah přiřazena znakům různá důležitost. [8] Pro výpočet standardizovaných dat je třeba nejprve vypočítat střední hodnotu a směrodatnou odchylku dat z původního souboru. Poté je možné hodnoty přepočítat na standardizované. Standardizované hodnoty znaků mají střední hodnotu rovnu 0, rozptyl 1 a jsou bezrozměrné. [5]

2.3.2. Normalizace dat

Objekty pro shlukovou analýzu jsou určeny vektory o p složkách představujících vícerozměrná data. Normy těchto vektorů mohou někdy nežádoucím způsobem ovlivňovat výsledky kvantitativního hodnocení podobnosti objektů. V takových případech je vhodné normalizovat tyto vektory, aby měly stejnou normu (nejlépe jednotkovou). Normalizace vektorů se provádí, aby jednotlivé hodnoty byly souměrné a byla tak odstraněna závislost na velikosti objektů. [5]

3. Podobnost objektů

Podobnost, respektive nepodobnost objektů je ve shlukové analýze základní myšlenka. Ve shlukové analýze jsou vytvářeny shluky sobě podobných objektů. Při tvorbě shluků je podobnost mezi objekty hlavním kritériem jejich tvorby. Nejdříve jsou stanoveny znaky určující podobnost, které se dále kombinují do podobnostních měr. Tímto způsobem se poté porovnávají objekty mezi sebou. Podobnost mezi objekty je měřena rozdílnými způsoby, které se dělí na tři kategorie: míry korelace, míry vzdálenosti a míry asociace. Korelační a vzdálenostní míry jsou míry metrických dat a asociační míry jsou míry určené především pro nemetrická data. [8] Hlavním cílem shlukové analýzy je vytvořit shluky tak, aby objekty tvořící shluk, si byly navzájem co nejvíce podobné a objekty nepatřících do stejného shluku by měly být rozlišné. Při měření se většinou používají míry nepodobnosti, ale není složité převést míru podobnosti na míru nepodobnosti. Při volbě metody také záleží, s jakým typem dat pracujeme. Zdali se jedná o data dichotomická nebo nominální. [10] Podobnost objektů je vyjadřována pomocí koeficientu korelace a asociace. Nepodobnost je určována pomocí metrik.

3.1. Koeficient korelace

Základní mírou vyjádření podobnosti dvou znaků je korelační koeficient. Objekty jsou si podobnější, čím větší jsou jejich párové korelační koeficienty a čím více se blíží jedné. [8] Korelace je ve statistice vzájemný lineární vztah mezi znaky či veličinami. Korelační koeficient může nabývat hodnot od −1 až po +1. Hodnota korelačního koeficientu −1 značí zcela nepřímou závislost, tedy čím více se zvětší hodnoty u jedné zkoumané veličiny, tím více se zmenší hodnoty u druhé (např. vztah mezi uplynulým a zbývajícím časem). Hodnota korelačního koeficientu +1 značí zcela přímou závislost, např. vztah mezi dobou chůze a ujitou vzdáleností za předpokladu konstantní rychlosti. Pokud je korelační koeficient roven 0, pak mezi znaky není žádná statisticky zjistitelná lineární závislost. Je dobré si uvědomit, že i při nulovém korelačním koeficientu na sobě veličiny mohou záviset, pouze tento vztah nelze vyjádřit lineární funkcí. [3] Určit, jaký koeficient korelace lze považovat již za významný, nelze pouze na základě jeho hodnoty, ale musí být posouzen v souvislosti s rozsahem základního souboru. Pro základní soubor, který obsahuje 60 hodnot, bude za velkou závislost považována hodnota koeficientu korelace r=0,3. Je-li koeficient korelace r=0,9, ale u soboru, který obsahuje např. 5 vstupních hodnot, nelze ani takto vyjádřenou závislost považovat za velice silnou.

3.1.1. Korelační matice

Pro n-rozměrnou matici lze vytvořit korelační matici. Korelační matice má na hlavní diagonále samé jedničky a mimo hlavní diagonálu koeficienty korelace. Korelační matice je souměrná. Lze provést jak korelaci objektů tak i vlastností. [3]

3.1.2. Nevýhody koeficientu korelace

Používání koeficientu korelace k hodnocení podobnosti vztahů objektů bývá často kritizováno. Především proto, že střední hodnota znaků téhož objektu, je hodnotou nesprávnou, protože jednotlivé znaky můžou představovat absolutně nesouměřitelné veličiny, vyjádřené ve vzájemně neporovnatelných jednotkách. To samé platí i o směrodatné odchylce. Tento nedostatek lze odstranit standardizací dat, která ale může do jisté míry zkreslit informace obsažené v původních datech. [5]

3.1.3. Výpočet koeficientu korelace

$$\rho(X,Y) = \frac{C(X,Y)}{\sigma(X)\sigma(Y)}$$

Rovnice 9 - Výpočet koeficientu korelace (zdroj: [3])

kde σ (*X*) a σ (*Y*) jsou směrodatné odchylky veličin *X*,*Y*. A C(X,Y) je kovariance. Pro výpočet koeficientu korelace existuje v MS Excel funkce v nástrojích analýzy dat.

3.2. Koeficienty asociace

Tato skupina koeficientů míry podobnosti je určena pro porovnávání objektů, pokud jejich znaky jsou nemetrického charakteru. Tedy jsou-li objekty charakterizovány dichotomickými daty. Tyto koeficienty jsou použity v případě, že dané objekty nelze oklasifikovat přímo číselnými hodnotami. [8]

3.2.1. Asociační tabulka

Asociaci – podobnost, dvojice objektů (Oi a Oj) bývá charakterizována pomocí asociační tabulky, která bude mít dva řádky a dva sloupce, viz obrázek 1.

Asociační tabulka		Oj	
		1	0
Oi	1	a	b
	0	c	d

Obrázek 1 - Asociační tabulka (zdroj: [5])

a – počet případu pozitivní shody, tedy tam, kde znaky objektů Oi a Oj nabývají současně hodnoty 1.

b – případ neshody takové, kde znak objektu Oi nabývá 1 a znak objektu Oj nabývá 0.

c – případ neshody takové, kde znak objektu Oi nabývá hodnoty 0 a znak objektu Oj nabývá hodnoty 1.

d – počet případů negativní shody, tj. počet znaků, v nich oba objekty zároveň nabývají hodnoty 0. [5]

Jako příklad je uvedena situace, kdy si zákazník vybírá z pěti audio sestav a bude chtít vědět, jaké sestavy jsou si podobné. Sestavy nemůžou být ohodnoceny jinak než pomocí dichotomických dat, a to následujícím způsobem. Obsahuje-li sestava daný prvek, například CD přehrávač, je zvolena odpověď ANO a je znázorněna pomocí 1 v asociační tabulce u dané vlastnosti daného objektu. Tímto způsobem je sestavena celá asociační tabulka. Zákazník vybral 5 sestav, které jsou porovnávány a tvoří tabulku, viz obrázek 2, z této tabulky byla pomocí výpočtů sestavena asociační tabulka, viz obrázek 3.

objekt					FM
	MC	CD	DVD	LP	rádio
O1	1	1	0	1	1
O2	1	1	0	0	0
O3	1	0	1	0	0
O4	0	1	1	0	0
O5	1	0	0	1	1
objekt	hodiny	budík	nahrávaní	subwoofer	HDD
O1	0	1	0	0	0
O2	1	1	1	1	0
O3	1	0	1	0	1
O4	0	1	1	0	1
O5	0	1	1	0	0

Obrázek 2 - Tabulka hodnot pro objekty O1-O5

Z dichotomických dat, která jsou zobrazena na obrázku 2, byly vypočteny hodnoty a,b,c,d asociačních tabulek pro každou dvojici objektů. Při výběru objektů O1 a O2, se postupovalo následovně. Hodnoty 1 nabývají oba objekty ve třech případech (MC, CD a budík). Počet případů pozitivní shody je 3 a tedy hodnota a=3. Stejným způsobem se postupuje dále. Počet případů negativní shody je 2. Vlastnosti objektů nabývají 0 ve dvou případech, tedy d=2. Stanoveny jsou i zbylé koeficienty, b=2 a c= 3. Pro větší přehlednost, byly objekty uspořádány do asociační tabulky, zobrazené na obrázku 3.

Asociační tab. objektů	O1		O2		O3		O4		O5	
	5	0	3	2	1	4	2	3	4	1
O1	0	5	2	3	4	1	3	2	1	4
	3	3	6	0	3	3	3	3	3	3
O2	2	2	0	4	2	2	2	2	2	2
	1	4	3	2	5	0	3	2	2	3
O3	4	1	2	3	0	5	2	3	3	2
	2	3	3	2	3	2	5	0	2	3
O4	3	2	3	2	2	3	0	5	3	2
	4	1	3	2	2	3	2	3	5	0
O5	1	4	3	2	3	2	3	2	0	5

Obrázek 3 - Asociační tabulka pro objekty O1-O5

3.2.3. Základní koeficienty asociace

Koeficientů asociace, které vycházejí z hodnot a,b,c,d asociačních tabulek existuje velké množství. V této kapitole jsou uvedeny ty nejvýznamnější a je proveden výpočet jejich hodnot, pro data z příkladu s audio sestavami. [5]

Jacardův koeficient Sj

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

Rovnice 10 - Jacardův koeficient (zdroj: [8])

Koeficient Sj není definován pro dva objekty, jejichž hodnoty jsou nulové, tedy vykazující negativní shodu, u všech vlastností objektu.

	O1	O2	O3	O4	O5
O1	**1.000**	**0.375**	**0.111**	**0.286**	**0.667**
O2	**0.375**	**1.000**	**0.375**	**0.375**	**0.375**
O3	**0.111**	**0.375**	**1.000**	**0.429**	**0.250**
O4	**0.250**	**0.375**	**0.429**	**1.000**	**0.250**
O5	**0.667**	**0.375**	**0.250**	**0.250**	**1.000**

Obrázek 4 - Asociace objektů pro koeficient Sj

Sokalův a Michenerův koeficient Ssm

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

Rovnice 11 - Sokalův a Michnerův koeficient(zdroj: [8])

Ssm vyjadřuje poměr mezi počtem shod a počtem všech případů. [2]

	O1	O2	O3	O4	O5
O1	**1.000**	0.500	0.200	0.500	0.800
O2	0.500	**1.000**	0.500	0.500	0.500
O3	0.200	0.500	**1.000**	0.600	0.400
O4	0.400	0.500	0.600	**1.000**	0.400
O5	0.800	0.500	0.400	0.400	**1.000**

Obrázek 5 - Asociace objektů pro koeficient Ssm

Russellův a Raoův koeficient Srr

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

Rovnice 12 - Russellův a Raoův koeficient (zdroj: [8])

Srr má nevýhodu v tom, že hodnotí závislosti objektů se sebou samým různým způsobem. Tedy nemá, na rozdíl od jiných koeficientů, na hlavní diagonále všechny hodnoty rovné jedné. To nastane pouze v případě, že by se jedné rovnaly všechny hodnoty znaků. Nerovnají-li se, je pak hodnota menší než jedné. [5]

	O1	O2	O3	O4	O5
O1	**0.500**	0.300	0.100	0.200	0.400
O2	0.300	**0.600**	0.300	0.300	0.300
O3	0.100	0.300	**0.500**	0.300	0.200
O4	0.200	0.300	0.300	**0.500**	0.200
O5	0.400	0.300	0.200	0.200	**0.500**

Obrázek 6 - Asociace objektů pro koeficient Srr

Diceův koeficient

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

Rovnice 13 - Diceův koeficient (zdroj: [8])

Sd má stejnou nevýhodu, jako koeficient Sj. Tedy pokud by oba objekty u všech znaků nabývaly hodnoty 0, nastala by negativní shoda u všech znaků, koeficient Sd pro takový případ není definovaný. [5]

	O1	O2	O3	O4	O5
O1	**1.000**	0.545	0.200	0.444	0.800
O2	0.545	**1.000**	0.545	0.545	0.545
O3	0.200	0.545	**1.000**	0.600	0.400
O4	0.400	0.545	0.600	**1.000**	0.400
O5	0.800	0.545	0.400	0.400	**1.000**

Obrázek 7 - Asociace objektů pro koeficient Sd

Nepojmenovaný koeficient 1 Sn_1

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

Rovnice 14 - Nepojmenovaný koeficient 1 (zdroj: [8])

	O1	O2	O3	O4	O5
O1	**1.000**	0.667	0.333	0.667	0.889
O2	0.667	**1.000**	0.667	0.667	0.667
O3	0.333	0.667	**1.000**	0.750	0.571
O4	0.571	0.667	0.750	**1.000**	0.571
O5	0.889	0.667	0.571	0.571	**1.000**

Obrázek 8 - Asociace objektů pro koeficient Sn1

Nepojmenovaný koeficient 2 Sn_2

$$S_{N2} = \frac{a}{a + 2 \cdot (b+c)}$$

Rovnice 15 - Nepojmenovaný koeficient 2 (zdroj: [8])

Sn2 má stejný nedostatek jako koeficienty Sj a Sd. Tedy není definován pro případ, kdy u všech znaků byla zjištěna negativní shoda.

	O1	O2	O3	O4	O5
O1	**1.000**	0.231	0.059	0.167	0.500
O2	0.231	**1.000**	0.231	0.231	0.231
O3	0.059	0.231	**1.000**	0.273	0.143
O4	0.143	0.231	0.273	**1.000**	0.143
O5	0.500	0.231	0.143	0.143	**1.000**

Obrázek 9 - Asociace objektů pro koeficient Sn2

Rogersův a Tanimotoův koeficient Srt

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

Rovnice 16 - Rogersův a Tanimotoův koeficient (zdroj: [8])

	O1	O2	O3	O4	O5
O1	**1.000**	0.333	0.111	0.333	0.667
O2	0.333	**1.000**	0.333	0.333	0.333
O3	0.111	0.333	**1.000**	0.429	0.250
O4	0.250	0.333	0.429	**1.000**	0.250
O5	0.667	0.333	0.250	0.250	**1.000**

Obrázek 10 - Asociace objektů pro koeficient Srt

Hamannův koeficient Sh

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

Rovnice 17 - Hamannův koeficient (zdroj: [8])

Koeficient Sh, na rozdíl od zbytku zmíněných koeficientů, má obor hodnot roven intervalu <-1,1>. Hodnoty -1 nabývá u dvojic objektů, které nedosahují shody v žádné ze zkoumaných vlastností. Hodnoty 0, nabývá u dvojice objektů, které mají stejný počet shody a neshody. Hodnoty 1 koeficient nabývá v případě shody u všech znaků. [5]

	O1	O2	O3	O4	O5
O1	**1.000**	0.000	-0.600	0.000	0.600
O2	0.000	**1.000**	0.000	0.000	0.000
O3	-0.600	0.000	**1.000**	0.200	-0.200
O4	-0.200	0.000	0.200	**1.000**	-0.200
O5	0.600	0.000	-0.200	-0.200	**1.000**

Obrázek 11 - Asociace objektů pro koeficient SH

Pro porovnání koeficientu asociace, jsou zobrazeny výsledky jednotlivých koeficientů do jednoho grafu, viz obrázek 14. Obrázek 12 zobrazuje hodnoty všech koeficientů pro každou dvojici objektů.

	Sj	Ssm	Srr	Sd	Sn1	Sn2	Srt	Sh
O11	1.000	1.000	0.500	1.000	1.000	1.000	1.000	1.000
O12	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O13	0.111	0.200	0.100	0.200	0.333	0.059	0.111	-0.600
O14	0.250	0.400	0.200	0.400	0.571	0.143	0.250	-0.200
O15	0.667	0.800	0.400	0.800	0.889	0.500	0.667	0.600
O21	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O22	1.000	1.000	0.600	1.000	1.000	1.000	1.000	1.000
O23	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O24	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O25	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O31	0.111	0.000	0.100	0.200	0.333	0.059	0.111	-0.600
O32	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O33	1.000	1.000	0.500	1.000	1.000	1.000	1.000	1.000
O34	0.420	0.600	0.300	0.600	0.750	0.273	0.429	0.200
O35	0.250	0.400	0.200	0.400	0.571	0.143	0.250	-0.200
O41	0.250	0.400	0.200	0.400	0.571	0.143	0.250	-0.200
O42	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O43	0.420	0.600	0.300	0.600	0.750	0.273	0.429	0.200
O44	1.000	1.000	0.500	1.000	1.000	1.000	1.000	1.000
O45	0.250	0.400	0.200	0.400	0.571	0.143	0.250	-0.200
O51	0.667	0.800	0.400	0.800	0.889	0.500	0.667	0.600
O52	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O53	0.250	0.400	0.200	0.400	0.571	0.143	0.250	-0.200
O54	0.250	0.400	0.200	0.400	0.571	0.143	0.250	-0.200
O55	1.000	1.000	0.500	1.000	1.000	1.000	1.000	1.000

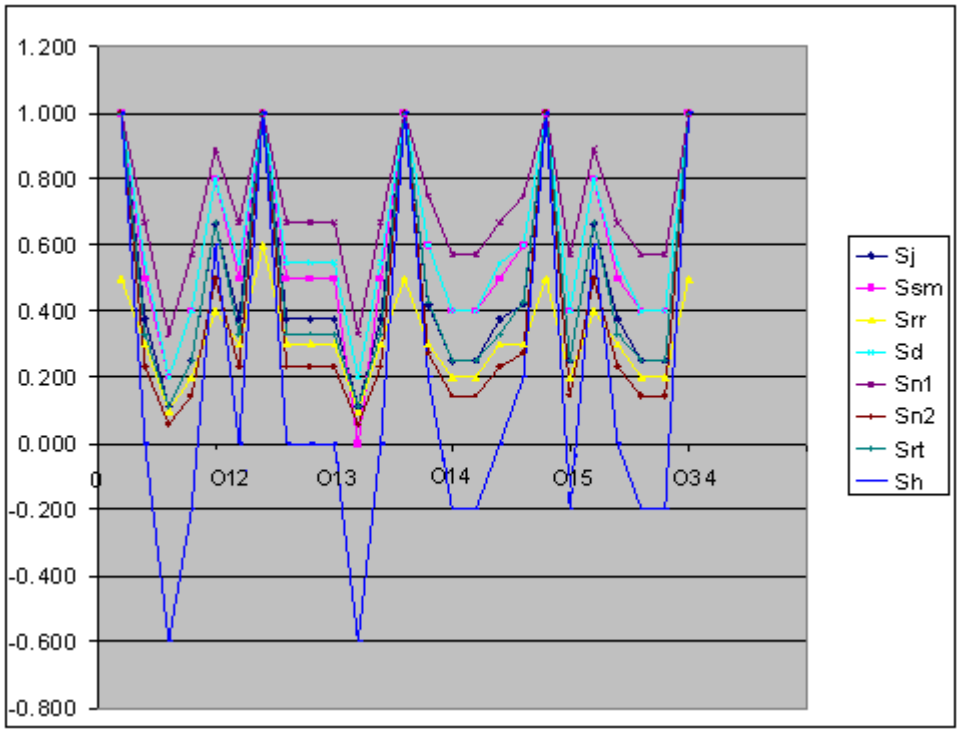
Obrázek 12 - Porovnání koeficientů asociace

Z výsledného grafu, viz obrázek 14, lze vyčíst, že výsledky hodnocení se pro různé koeficienty příliš neliší. Jediná výraznější odchylka byla zaznamenána u koeficientu Srr. Takto zobrazený graf ale není příliš přehledný. Proto z tabulky byly odstraněny všechny duplicity. Například hodnoty objektů O2 a O3 jsou stejné jako hodnoty O3 a O2. Také je vynechána závislost objektů se sebou samým. Upravená data jsou zobrazena na obrázku 13 a graf na obrázku 15 zobrazuje mnohem přehledněji vlastnosti objektů a jejich podobnost. [5]

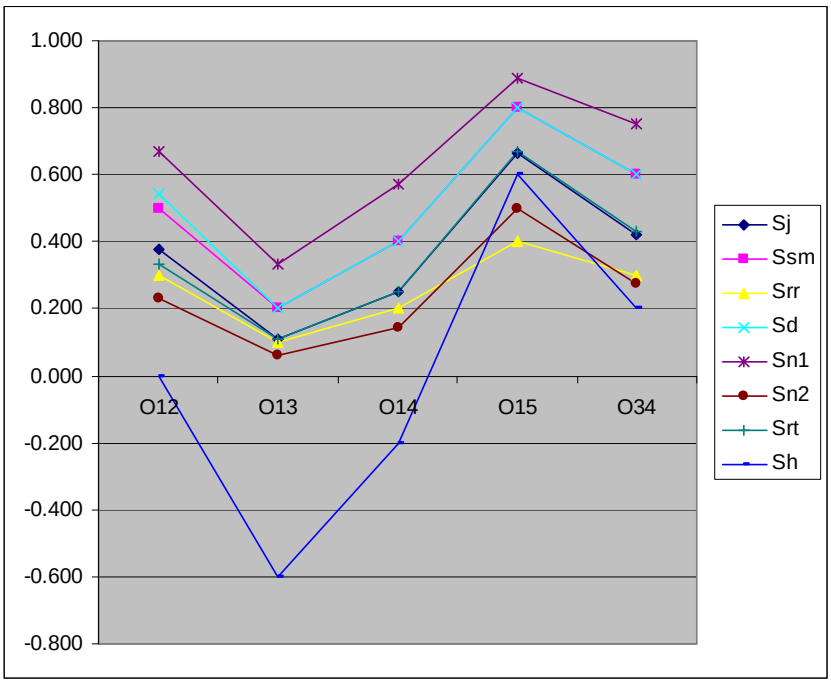
	Sj	Ssm	Srr	Sd	Sn1	Sn2	Srt	Sh
O12	0.375	0.500	0.300	0.545	0.667	0.231	0.333	0.000
O13	0.111	0.200	0.100	0.200	0.333	0.059	0.111	-0.600
O14	0.250	0.400	0.200	0.400	0.571	0.143	0.250	-0.200
O15	0.667	0.800	0.400	0.800	0.889	0.500	0.667	0.600
O34	0.420	0.600	0.300	0.600	0.750	0.273	0.429	0.200

obrázek 13 - upravená tabulka pro koeficienty asociace

Z grafu, viz obrázek 15, je vidět, že maxima dosahují hodnoty pro objekty O1 a O5 a minima pro hodnoty O1 a O3. Tedy objekty O1 a O5 jsou si nejvíce podobné, objekty O1 a O3 jsou si podobné nejméně. [5]



Obrázek 14 - Graf koeficientů asociace pro objekty O1-O5



Obrázek 15 - Graf koeficientů asociace pro objekty O1-O5 bez duplicit

3.3. Metriky

Míry vzdálenosti představují nejčastěji používané míry založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Pokud tyto míry splňují následující požadavky, hovoří se o metrikách. [8]

Podmínky:

Identita ρ (AB)=0 ⇔ A=B

ρ (AB) > 0

Symetrie ρ (AB) = ρ (BA)

Trojúhelníková nerovnost ρ (AC) <= ρ (AB) + ρ (BC)

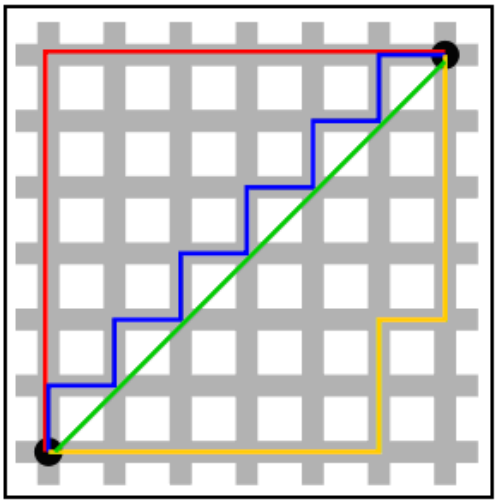
Metriky jsou jeden z nejběžnějších způsobů vyjádření podobnostních vztahů mezi objekty a vycházejí z geometrického modelu dat. [5] Množina reálných čísel spolu s metrikou ρ(x,y) = | x − y |), kde x,y jsou libovolné body množiny reálných čísel , tvoří úplný metrický prostor. U metrik je vyjadřována míra nepodobnosti. Pokud je vzdálenost mezi body (metrika) menší, objekty jsou si více podobné.

3.3.1. Manhattan metrika

Nejjednodušší vícerozměrnou variantou je tzv. součtová či Manhattan metrika, která nese název dle slavné New Yorské čtvrti Manhattan, kde jsou veškeré ulice na sebe kolmé a metrika vyjadřuje vzdálenost, kterou je potřeba ujít mezi dvěma křižovatkami, přičemž se lze pohybovat jen po na sebe kolmých ulicích ve směru obou os. [6] Tato situace je zobrazena na obrázku 16. Manhattan metrika je na množině reálných čísel definována jako:

$$\rho_1(A,B) = \sum_{i=1}^{p} |a_i - b_i|$$

Rovnice 18 - Manhattan metrika (zdroj: [3])



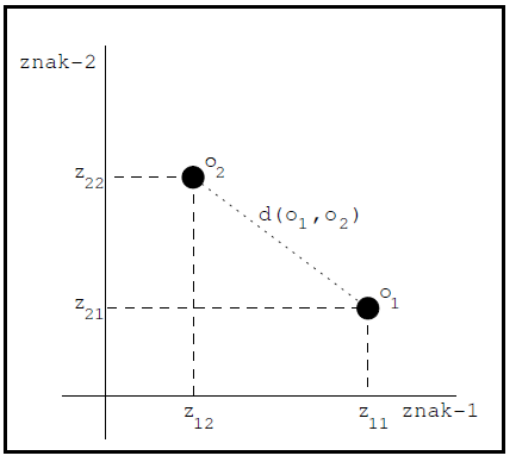
Obrázek 16- Manhattan metrika (zdroj:[6])

3.3.2. Euklidovská metrika

Nejčastější vzdálenostní mírou je Euklidovská vzdálenost, euklidovská metrika. Jinak také zvaná geometrická vzdálenost, která je určena délkou přepony pravoúhlého trojúhelníka. [5] Pro lepší názornost je na obrázku 17 zobrazena Euklidovská metrika graficky. Výpočet euklidovské metriky vychází z Pythagorovy věty a tedy platí:

$$\rho_E(A,B) = \sqrt{\sum_{i=1}^{p} \left(a_i - b_i\right)^2}$$

Rovnice 19 - Euklidovská metrika (zdroj: [3])



Obrázek 17 - Euklidovská metrika (zdroj:[2])

3.3.3. Čtverec euklidovské vzdálenosti

Tato metrika bývá používaná v případě, jsou-li znaky vzájemně nezávislé, tedy dosahují nulové hodnoty koeficientu kovariace. Tato metrika bývá nejčastěji používána u Wardovy shlukovací metody.

$$\rho^2(A,B)=\sum_{i=1}^{p}\left(a_i-b_i\right)^2$$

Rovnice 20 - Čtverec euklidovské vzdálenosti (zdroj: [3])

3.3.4. Sokalova metrika

$$\rho_S(A,B)=\sqrt{\left(\frac{\rho_E^2(A,B)}{p}\right)}$$

Rovnice 21 - Sokalova metrika (zdroj: [3])

3.3.5. Sup metrika

$$\rho_\infty(A,B)=\max_{i=1,..,p}\left(|a_i-b_i|\right)$$

Rovnice 22 - Sup metrika (zdroj: [3])

U všech vzdálenostních měr je potřeba si dát pozor na to, jaká data jsou užívána. Problém nastává v okamžiku, kdy jsou použita data nestandardizovaná, která mohou způsobit rozdíly mezi shluky, a to odlišností jednotek měření. Shluky různých vzdálenostních měr se budou lišit, největší rozptyl mezi shluky bude u čtverce euklidovské vzdálenosti. Standardizace dat by měla být využita vždy, je-li to možné. [5]

4. Analýza problémů pomocí nehierarchického shlukování

4.1. Základní soubor

K vysvětlení postupu při nehierarchickém shlukování bylo využito dat ze serveru www.crohn.cz, který se věnuje zkoumání potravin a hodnot jednotlivých látek, které daná potravina obsahuje. K analyzování bylo vybráno 63 potravin (objektů) a každá byla ohodnocena 10 vlastnostmi. Tyto objekty a jejich vlastnosti tvoří vstupní matici, na které budou prováděny základní analýzy a samotné shlukování pomocí nehierarchických metod. Pro svou obsáhlost je vstupní matice umístěna na CD, které tvoří nedílnou součást práce. Data byla vybrána pro svoji srozumitelnost a názornost. V současné době se zkoumání potravin věnuje velké množství prostoru a základní představu o dané problematice má většina populace. Pro analýzu dat byly vybrány tyto vlastnosti: obsah kalorií, bílkovin, tuků, sacharidů, vápníku, fosforu, železa, nikotinamidu a vitamínu B1 a B2. Jelikož program Statistica nabízí z nehierarchických metod pouze metodu Kmeans, bylo k provedení analýz pomocí jiných metod využito programu Shluk. Ten umožňuje výběr z více metod, ale nedokáže pracovat s natolik rozsáhlou maticí, která je určena ke zpracování. Metody, které budou prováděny v programu Shluk, pracují s výběrem ze základního souboru, viz obrázek 18.

Název	Kal.	Bílk. [g]	Tuky[g]	Sach.[g]	Ca [mg]	P [mg]	Fe [mg]	B1 [mg]	B2 [mg]	PP [mg]
Banány	59	0.8	0.1	15.4	5	19	0.4	0.027	0.034	0.47
Brambory	80	2	0	19	13	72	0.8	0.07	0.04	1.2
Čočka	330	25	1	60	59	423	7.5	0.56	0.24	2.2
Droždí	102	10.6	0.4	13	25	605	4.9	0.45	2.07	28
Fazole	331	21	1.6	62	137	437	6.9	0.67	0.23	3.1
Hrách	332	24	1.4	60	57	388	4.7	0.77	0.28	3.1
Mák	501	19.5	40.8	24.3	1400	610	12	0.45	2.07	28
Pomeranče	32	0.6	0.1	8.1	24	18	0.29	0.058	0.022	0.14
Třešně	54	1	0.4	13.3	16	18	0.36	0.046	0.055	0.36
Vejce 100g	139	11.6	9.8	0	53	196	1.78	0.089	0.267	0.09

Obrázek 18 - Výběrový soubor (výstup: MS Excel)

Nutnou úpravou dat je jejich transformace, tedy standardizace a normalizace. U všech metod je využíváno standardizovaných a normalizovaných dat. Pro příklad bylo v programu Statistica provedeno shlukování pro standardizovaná i normalizovaná data, pouze pro standardizovaná data a pro data bez transformace. Výsledky jsou uvedeny v příloze na CD. Jak se lišily počty objektů ve shlucích je uvedeno na obrázku 19.

n=63	Provedené úpravy			
	STAND+NORM	STAND	BEZ ÚPRAV	
SHLUK1	23	8	4	Počet objektů
SHLUK2	31	47	38	
SHLUK3	9	8	21	

Obrázek 19 - Porovnání dat bez transformace (zdroj: autor)

4.2. Popisná statistika

Dalším krokem předpřípravy dat po jejich transformaci je charakteristika objektů a vlastností pomocí popisné statistiky. Zobrazení popisné statistiky umožňuje program MS Excel pomocí nástroje analýzy dat. Zde je na obrázku 20 zobrazen výstup z programu Statistica. Z obrázku 20 lze vyčíst údaje, které pomohou udělat si o datech základní představu. Tabulka ukazuje, jaké hodnoty jsou průměrné, jak moc se liší průměrná hodnota od mediánu a například jaké jsou maximální hodnoty. Popisná statistika je spíše ukazatelem statistickým, ale o samotných závislostech jednotlivých objektů příliš nenapoví. K určení závislosti objektů a vlastností proto bývá nejčastěji využíváno koeficientu korelace.

	Popisné statistiky (import)												
Proměnná	N platných	Průměr	Medián	Modus	Četnost modu	Minimum	Maximum	Spodní kvartil	Horní kvartil	Rozptyl	Sm. odch.	Šikmost	Špičatost
Kal.	63	149,8571	95,0000	20,00000	3	12,00000	1175,000	36,00000	230,0000	32134,35	179,2606	3,324035	16,46248
Bílk. [g]	63	8,8048	5,2000	,9000000	5	0,20000	29,900	0,90000	16,0000	75,80	8,7066	0,674260	-0,79429
Tuky[g]	62	5,1677	0,6000	,1000000	10	0,00000	40,800	0,20000	7,3000	73,86	8,5940	2,138894	4,64957
Sach.[g]	63	14,0603	5,5000	0,000000	10	0,00000	78,900	1,60000	15,4000	453,72	21,3008	2,004266	2,83129
Ca [mg]	63	88,6984	25,0000	8,000000	5	4,00000	1400,000	13,00000	57,0000	43902,18	209,5285	4,747121	25,97130
P [mg]	63	148,0079	101,0000	18,00000	6	0,50000	610,000	21,00000	200,0000	24423,81	156,2812	1,287106	0,88881
Fe [mg]	63	1,8957	0,6300	,5000000	4	0,00000	17,460	0,40000	2,1000	9,47	3,0771	3,141709	11,52285
B1 [mg]	63	0,1284	0,0510	Vícenás.	3	0,01000	0,770	0,03000	0,1000	0,03	0,1816	2,178361	3,80148
B2 [mg]	63	0,2965	0,0700	,0380000	4	0,01300	3,230	0,03800	0,2670	0,40	0,6293	3,569623	12,63436
PP [mg]	63	2,7094	0,6700	,1000000	4	0,06000	28,000	0,24000	3,1000	29,95	5,4723	3,624622	14,00269

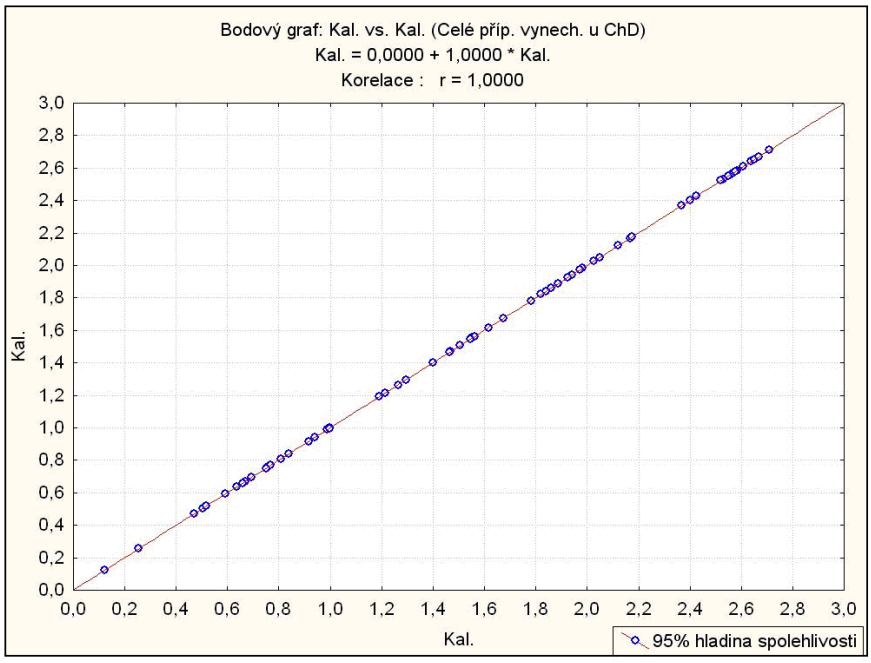
Obrázek 20 - Popisná statistika (výstup: Statistica)

4.3. Koeficient korelace

Proměnná	Korelace (import) Označ. korelace jsou významné na hlad. p < ,05000 N=62 (Celé případy vynechány u ChD) Kal.	Bílk. [g]	Tuky[g]	Sach.[g]	Ca [mg]	P [mg]	Fe [mg]	B1 [mg]	B2 [mg]	PP [mg]
Kal.	1,00	0,51	0,49	0,36	0,47	0,55	0,25	0,32	0,16	0,18
Bílk. [g]	0,51	1,00	0,50	0,11	0,40	0,81	0,48	0,45	0,39	0,34
Tuky[g]	0,49	0,50	1,00	-0,16	0,69	0,51	0,36	0,19	0,24	0,38
Sach.[g]	0,36	0,11	-0,16	1,00	-0,02	0,24	0,14	0,48	-0,04	-0,01
Ca [mg]	0,47	0,40	0,69	-0,02	1,00	0,57	0,30	0,15	0,30	0,41
P [mg]	0,55	0,81	0,51	0,24	0,57	1,00	0,62	0,62	0,59	0,63
Fe [mg]	0,25	0,48	0,36	0,14	0,30	0,62	1,00	0,63	0,74	0,71
B1 [mg]	0,32	0,45	0,19	0,48	0,15	0,62	0,63	1,00	0,38	0,47
B2 [mg]	0,16	0,39	0,24	-0,04	0,30	0,59	0,74	0,38	1,00	0,81
PP [mg]	0,18	0,34	0,38	-0,01	0,41	0,63	0,71	0,47	0,81	1,00

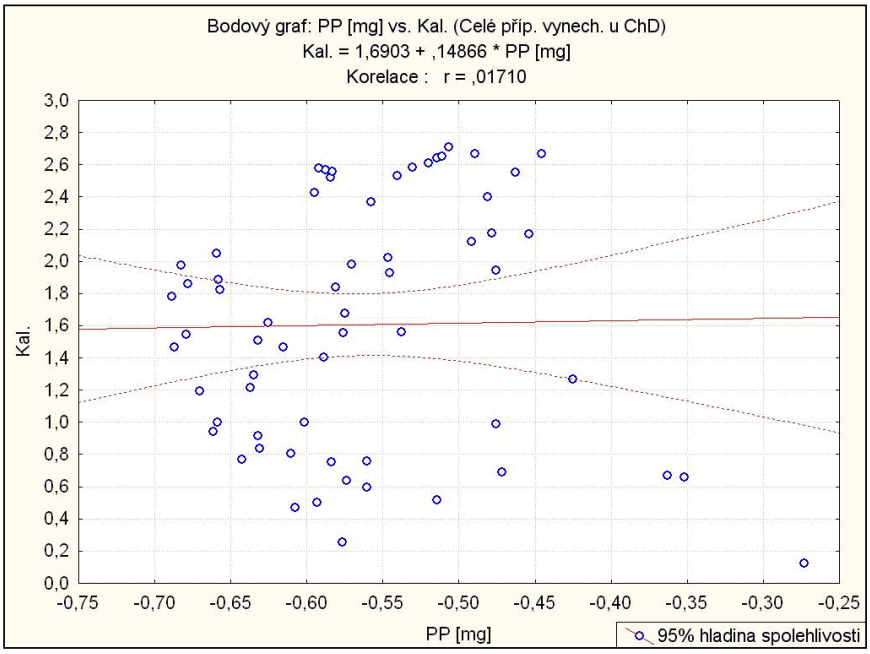
Obrázek 21 - Korelace vlastností (výstup:Statistica)

Na obrázku 21 je zobrazeno, jak jsou jednotlivé vlastnosti na sobě přímo či nepřímo závislé. Na hlavní diagonále jsou hodnoty koeficientu korelace rovny jedné. Jedná se o absolutní závislost vlastnosti se sebou samou. Je logické, že obsah kalorií je absolutně závislý na obsahu kalorií dané potraviny. Tato závislost je znázorněna graficky, viz obrázek 22, kde všechny hodnoty leží na diagonále a jejich rozptyl je nulový. Literatura většinou udává jako silnou závislost, dosahuje-li koeficient korelace hodnoty 0,8. Ovšem jedná se o zavádějící údaj. Hodnotu koeficientu je potřeba analyzovat v souvislosti s rozsahem základního souboru. Pro soubor, kde bylo analyzováno např. 50 vstupních hodnot, může být silná závislost již r=0,3. Naopak, pro soubor s 10 vstupními hodnotami, nemusí ani hodnota koeficientu korelace r=0,8 znamenat vysokou závislost. Jak je vidět na obrázku 21, pro soubor o 62 objektech, program Statistica určil jako významnou závislost již r=0,25.



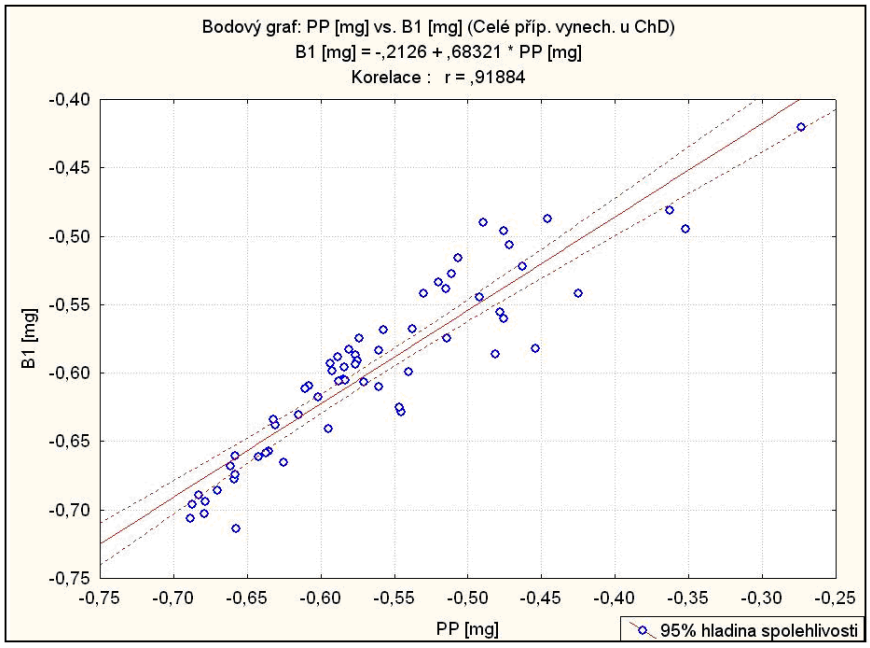
Obrázek 22 - Absolutní korelační závislost (výstup:Statistica)

Dalším případem je velice slabá závislost dvou vlastností. To nastává u kalorií a PP, kde korelační koeficient r=0,02 a závislost mezi těmito vlastnostmi je velice malá. Počet kalorií v dané potravině téměř neovlivňuje hodnotu PP a naopak. Tuto nízkou závislost znázorňuje obrázek 23. Hodnoty jsou velice rozptýleny a oproti předešlému případu jsou velice vzdáleny od osy.



Obrázek 23 - nízká korelační závislost (výstup:Statistica)

Posledním případem, který může nastat je závislost vysoká, ale ne absolutní. Takový případ nastal u PP a B1. Zde dosahuje korelační koeficient vysokého čísla r=0,92. Závislost mezi obsahem B1 a PP je vysoká a obsah jedné látky přímo ovlivňuje obsah druhé. Tento případ je zobrazen na obrázku 24.



Obrázek 24 - Vysoká korelační závislost (výstup:Statistica)

5. Nehierarchické shlukování

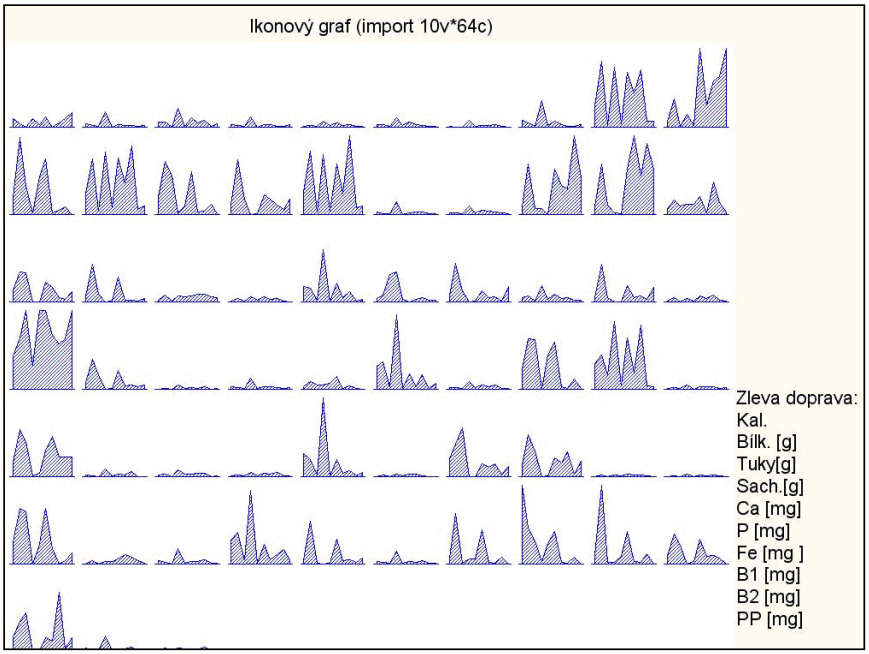
Metody, které lze souhrnně nazvat nehierarchickými shlukovacími metodami, se obecně snaží najít optimální rozklad zadané množiny objektů. Tento rozklad je považován za nejvhodnější klasifikaci zkoumaných objektů. [2] Na rozdíl od hierarchických metod, se tyto metody nezabývají výstavbou dendogramů. Nehierarchické shlukovací metody přidělují objekty do předem známého počtu shluků. Vždy poté co je stanoven střed, neboli těžiště shluku, jsou objekty spadající do předem specifikované vzdálenosti k tomuto shluku přiřazeny. Následuje volba nového těžiště a opětovné přesunutí objektů. Může se tedy stát, že objekt, který již byl v nějakém shluku, bude přesunut znovu do nového shluku, jelikož má k těžišti nového shluku blíže. Nehierarchické shlukové metody se někdy označují jako metody nejbližších těžišť – Kmeans. V nehierarchických shlukových metodách, je počet shluků předem stanoven. Podle toho, zdali se v průběhu výpočtu počet shluků mění či ne, jsou rozlišovány metody s konstantním počtem shluků, kdy se v průběhu výpočtu počet shluků nemění a v opačném případě metody s optimalizovaným počtem shluků, kdy v průběhu výpočtu dochází ke změně počtu shluků. [5]

Největším problémem nehierarchických metod je volba těžišť jednotlivých shluků, ke kterým jsou přiřazovány objekty. Existují tři metody přiřazování objektů do jednotlivých shluků, respektive k jejich typickým bodům.

1) Sekvenční práh – je zvolen jeden zárodek shluku a do něj jsou přiřazeny všechny objekty, uvnitř zvolené vzdálenosti. Poté, co jsou všechny objekty spadající do zvolené vzdálenosti přiřazeny k typickému bodu shluku, je zvolen druhý typický bod se specifikovanou vzdáleností a opět jsou přiřazeny objekty, které do této vzdálenosti spadají. Poté je vybrán třetí typický bod shluku a postup je opakován. Je-li objekt přiřazen k jednomu typickému bodu, v dalších krocích již s ním není počítáno.
2) Paralelní práh – na rozdíl od prvé metody, při této metodě je vybráno na začátku hned několik typických bodů shluku a souběžně jsou zařazovány objekty uvnitř předem specifikované vzdálenosti k nejbližšímu typickému bodu. V průběhu procesu může být vzdálenost změněna tak, aby do ní spadalo více či méně objektů.
3) Optimalizační metoda – tato metoda dovoluje znovuzařazení objektů. Dojde-li v průběhu tvorby shluků k situaci, že některý z objektů se octne blíže jinému shluku, než ve kterém se nachází, dojde k jeho přeřazení do jiného, bližšího shluku. [2]

5.1. Počáteční rozklad

Při zpracování dat nehierarchickými postupy je stanovení optimálního rozkladu na k-shluky velice důležitým krokem. Obzvláště důležitý je tento rozklad u metod s pevným počtem shluků, jelikož v dalším průběhu již nelze tento odhad měnit. U metod s proměnným počtem shluků lze počet „k“ měnit v závislosti na řídících proměnných algoritmů. K stanovení optimálního počtu k-shluků je možné využít znalostí experta dané problematiky, který provede odhad, okolo kterých proměnných se očekává vytváření shluků, a tyto proměnné se označí jako typické body, typické objekty. Častěji je využíváno statistických metod či lze výběr typických bodů provádět náhodně. Náhodný výběr má vyloučit subjektivní rozhodnutí experta o tom, které body či objekty budou vybrány jako typické. Například McQueen navrhuje vybrat za typické body prvních k-bodů z volně uspořádané množiny bodů. McRae navrhuje označit body pořadovými čísly 1,2,….,n a náhodně generovat k pořadových čísel z množiny (1,2,…,n). Dalším způsobem určení počtu shluků je využít některé z metod hierarchického shlukování a podle těchto výsledků volit počet shluků. Ovšem stále je potřeba řešit, jaké body vybrat jako typické. [5] Počáteční rozklad je odvozen z údajů tak, aby hodnota funkcionálu kvality rozkladu byla již v počátečním stádiu co nejlepší. Pro názornost je zobrazen ikonový graf, viz obrázek 25, jenž charakterizuje jednotlivé objekty pomocí tvarů, a umožní tak udělat si základní představu o počtu shluků. Čím si jsou tvary více podobné, tím si jsou podobnější i objekty, které daný tvar reprezentuje. Na obrázku se vyskytují zhruba 3 až 4 tvary objektů, odhadem tedy bude shlukováno do 3 až 4 shluků. Toto je ovšem jen počáteční odhad a nepříliš přesný. Pro přesnější stanovení počtu shluků bylo využito metody hierarchického shlukování v programu Statistica. Optimální je shlukovat objekty do 3 shluků. Tento počet shluků se bude dodržovat u všech metod nehierarchického shlukování. Na serveru www.crohn.cz jsou potraviny rozděleny do pěti skupin. Jsou to skupiny ovoce, zelenina, látky bohaté na sacharidy, látky s extrémními hodnotami a zdroje bílkovin. I přes toto dělení do pěti skupin bude shlukováno pouze do tří shluků. Při bližším pohledu na skupinu látek s extrémními hodnotami lze vypozorovat, že se v této skupině nachází zelenina, ovoce nebo látky, jež by mohly spadat i do jiných skupin. Toto platí i u skupiny látek s vysokým obsahem sacharidů a bílkovin. Dalším důvodem, proč plně nerespektovat při tvorbě shluků počet skupin ze serveru crohn.cz, je skutečnost, že není využito všech objektů a je pracováno pouze s výběrem.



Obrázek 25 - Ikonový graf (výstup:Statistica)

5.2. Funkcionál kvality rozkladu

Hledá-li se optimální rozklad množiny objektů, je potřeba si stanovit, v jakém smyslu má být tento rozklad optimální. Jsou-li požadavky formulovány matematickými prostředky, dospěje se k funkcionálu kvality rozkladu, který je definován na množině všech možných rozkladů množiny objektů. Rozklad je optimální, nabývá-li funkcionál kvality rozkladu extrémních hodnot. Funkcionál kvality rozkladu by měl vyjadřovat některou z následujících vlastností shluků tvořících rozklad:

1) Vzájemnou podobnost objektů uvnitř shluku
2) Míru separace shluků
3) Homogenitu rozložení objektů uvnitř shluků
4) Rovnoměrnost rozložení objektů do různých shluků
5) Kombinace všech (viz body 1-4)

Matematickým prostředkem k vyjádření funkcionálu kvality rozkladu je funkcionál součtu čtverců odchylek, což je odchylka objektů od středu shluku (těžiště). Čím je funkcionál menší, tím jsou si objekty uvnitř shluku podobnější. Funkcionál je vypočten u každého shluku a poté jsou sečteny dohromady. [5]

5.3. Metody s pevným počtem shluků (K-Means)

5.3.1. Forgyova a Janceyova shlukovací metoda

Obě tyto metody jsou založeny na opakovaném provádění dvou kroků. Prvním krokem je výpočet typických bodů z množiny shlukovaných objektů. Krok druhý spočívá v přiřazení každého objektu ke skupině, ke které má daný objekt nejblíže. Tyto dva kroky jsou opakovány tak dlouho, dokud není dosaženo stabilního stavu. To je stavu, kdy následující rozklad je shodný se stavem předešlým. Neboli tak dlouho, až žádný bod během jedné iterace nezmění své členství ve skupině. [2] Forgyova a Janceyova metoda se liší pouze způsobem výpočtu nových typických bodů. U obou metod jsou na začátku zadány nebo odvozeny počáteční typické body, nebo je proveden rozklad na k skupin, kde jsou typické body těchto skupin vypočteny jako těžiště těchto skupin. Poté jsou postupně přiřazeny jednotlivé body k nejbližším typickým bodům. Tím vznikají nové rozklady, ve kterých jsou opětovně vypočteny typické body. Forgyova metoda počítá nový typický bod jako těžiště dané skupiny, kdežto Janceyova metoda umisťuje nový typický bod do bodu souměrně sdruženého s typickým bodem předešlé skupiny. Tento postup je opakován do okamžiku, kdy žádný z bodů během jedné iterace nezmění svojí příslušnost ve skupině. Z praxe vyplývá, že počet iterací je malý, ale záleží také na správném počátečním rozkladu a odhadu shluků. Obě zmíněné metody lokálně minimalizují funkcionář kvality rozkladu. [5] Není-li zmíněno jinak, byly všechny metody provedeny v programu Shluk. Jako vstupní matice byl zvolen výběr ze základního souboru, viz obrázek 18. Prvním krokem, než proběhne samotné shlukování, je provedení transformace dat. Po načtení vstupní matice program Shluk provede standardizaci a poté normalizaci dat. Obě matice jsou programem uloženy pro pozdější použití. Obě matice jsou zobrazeny na obrázku 26.

Jméno matice: STAND
Matice byla vytvořena funkcí: STAN
Postup výpočtu: b00
Řádků: 10 Sloupců: 10

Kal.	Bílk.	Tuky[	Sach.	Ca [m	P [mg	Fe [m	B1 [m	B2 [m	PP [m
-0.84	-1.06	-0.43	-0.51	-0.40	-1.07	-0.90	-1.00	-0.61	-0.55
-0.71	-0.95	-0.44	-0.36	-0.39	-0.85	-0.80	-0.86	-0.60	-0.48
0.82	1.32	-0.36	1.37	-0.28	0.60	0.89	0.83	-0.36	-0.40
-0.58	-0.10	-0.41	-0.61	-0.36	1.35	0.24	0.45	1.88	1.89
0.83	0.93	-0.31	1.45	-0.10	0.65	0.74	1.21	-0.37	-0.32
0.83	1.22	-0.33	1.37	-0.28	0.45	0.19	1.55	-0.31	-0.32
1.87	0.78	2.77	-0.14	2.83	1.37	2.03	0.45	1.88	1.89
-1.01	-1.08	-0.43	-0.82	-0.36	-1.08	-0.93	-0.90	-0.62	-0.58
-0.87	-1.05	-0.41	-0.60	-0.38	-1.08	-0.91	-0.94	-0.58	-0.56
-0.35	-0.00	0.33	-1.16	-0.29	-0.34	-0.55	-0.79	-0.32	-0.58

Jméno matice: NORM
Matice byla vytvořena funkcí: NORM
Postup výpočtu: a00
Řádků: 10 Sloupců: 10

Kal.	Bílk.	Tuky[	Sach.	Ca [m	P [mg	Fe [m	B1 [m	B2 [m	PP [m
-0.34	-0.43	-0.17	-0.21	-0.16	-0.43	-0.36	-0.41	-0.25	-0.22
-0.33	-0.44	-0.20	-0.17	-0.18	-0.40	-0.37	-0.40	-0.28	-0.23
0.32	0.51	-0.14	0.53	-0.11	0.23	0.35	0.32	-0.14	-0.15
-0.18	-0.03	-0.13	-0.19	-0.11	0.42	0.07	0.14	0.59	0.59
0.33	0.36	-0.12	0.57	-0.04	0.26	0.29	0.48	-0.15	-0.12
0.31	0.46	-0.12	0.51	-0.11	0.17	0.07	0.58	-0.12	-0.12
0.33	0.14	0.48	-0.02	0.49	0.24	0.35	0.08	0.33	0.33
-0.39	-0.42	-0.17	-0.32	-0.14	-0.42	-0.36	-0.35	-0.24	-0.22
-0.35	-0.43	-0.17	-0.24	-0.15	-0.44	-0.37	-0.38	-0.24	-0.23
-0.20	-0.00	0.19	-0.65	-0.16	-0.19	-0.31	-0.45	-0.18	-0.33

Obrázek 26 - Standardizovaná a normalizovaná matice (výstup:Shluk)

Data jsou připravena ke shlukování. Následuje výběr typických bodů. Jelikož byl stanoven počet shluků na tři, bude i počet typických bodů roven třem. Typické body jsou zvoleny metodou prvních k bodů, kde k odpovídá počtu shluků. Z vybraných bodů program shluk opět vytvoří samostatnou matici, viz obrázek 27.

Jméno matice: TYPB
Matice byla vytvořena funkcí: TYP1
Postup výpočtu: ap0
Řádků: 3 Sloupců: 10

-0.34	-0.43	-0.17	-0.21	-0.16	-0.43	-0.36	-0.41	-0.25	-0.22
-0.33	-0.44	-0.20	-0.17	-0.18	-0.40	-0.37	-0.40	-0.28	-0.23
0.32	0.51	-0.14	0.53	-0.11	0.23	0.35	0.32	-0.14	-0.15

Obrázek 27 - Matice typických bodů (výstup: Shluk)

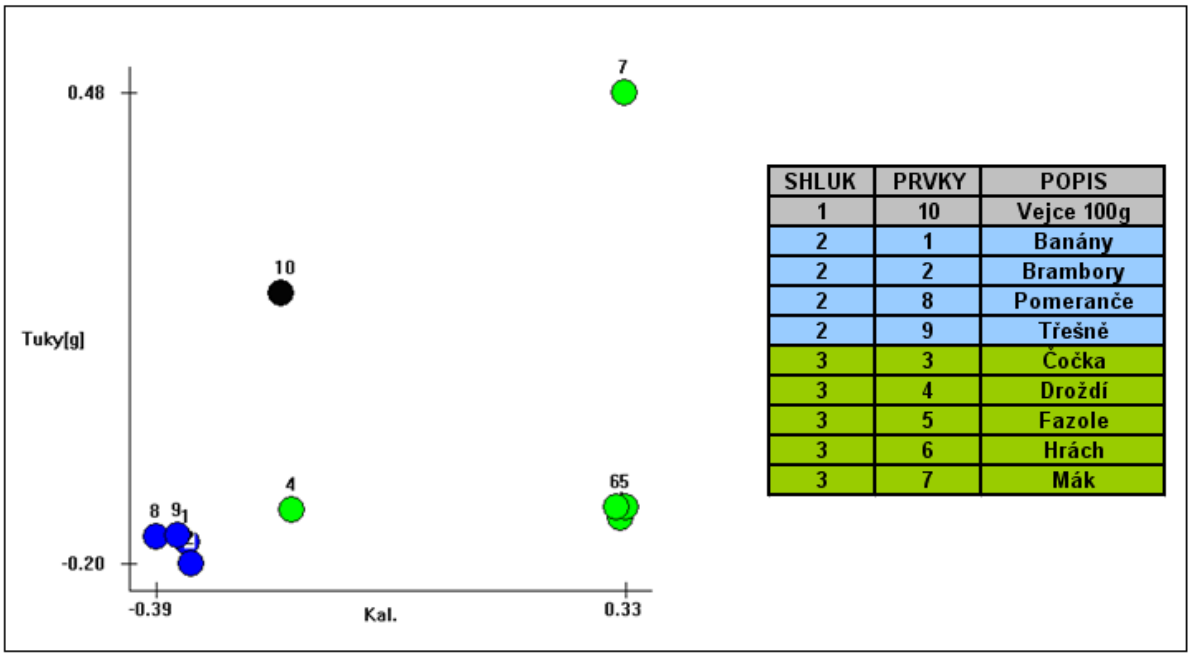
V programu Shluk jsou vybrána data, která budou shlukována, poté je načtena matice typických bodů a je shlukováno do tří shluků. Výsledek shlukování ukazuje obrázek 28.

```
Jméno matice: FORG

Počet iterací =4

Shluk: 1
Teziste: -0.20 -0.00 0.19 -0.65 -0.16 -0.19 -0.31 -0.45 -0.18 -0.33
Prvky:
10;
Součet čtverců odchylek: 0.00

Shluk: 2
Teziste: -0.35 -0.43 -0.18 -0.23 -0.16 -0.42 -0.37 -0.38 -0.25 -0.22
Prvky:
1; 2; 8; 9;
Součet čtverců odchylek: 0.02

Shluk: 3
Teziste: 0.22 0.29 -0.01 0.28 0.03 0.26 0.23 0.32 0.10 0.10
Prvky:
3; 4; 5; 6; 7;
Součet čtverců odchylek: 2.72


Celkový součet čtverců odchylek =2.74

Vzájemná vzálenost těžišt:
0.00
0.77 0.00
1.60 1.65 0.00

Průměrná vzdálenost těžišť = 1.34
```

Obrázek 28 - Sumarizace Forgyovy shlukovací metody (výstup:Shluk)

Na obrázku 28 je zobrazen souhrn celého shlukování. Proběhly 4 iterace a objekty byly rozděleny do shluků následovně. V prvním shluku se ocitl pouze objekt č. 10, tedy vejce 100g. Shluk, který je tvořen pouze jediným objektem, je neuspokojivým. Příčin vzniku takového shluku může být více. Důvodem by mohl být malý počet objektů a nevhodný výběr ze základního souboru. Při takto malém počtu objektů je pravděpodobné, že jeden z objektů se bude od ostatních lišit natolik, že jeho vzdálenost od těžiště ostatních shluků bude natolik velká, že do žádného z nich nebude spadat. Takovýto objekt je označen jako odlehlý. Možným řešením by byla změna počtu shluků na dva, nebo rozšíření základní matice o další objekty. Ve druhém shluku se vyskytují objekty 4. Jsou to banány, brambory, třešně a pomeranče. Tento shluk má velice nízký součet čtverců odchylek, a tedy tyto objekty jsou si velice podobné, co se týče složení jednotlivých látek. Je poměrně překvapující, že do stejného shluku byly zařazeny například brambory a pomeranč. Tyto potraviny na první pohled spolu moc společného nemají, ale jak ukázalo shlukování, co se týče hodnot jednotlivých zkoumaných látek, jsou si velice podobné.

Lze očekávat, že objekty druhého shluku budou velice blízko u sebe, což potvrzuje i graf na obrázku 29. Třetí shluk obsahuje nejvíce prvků. Jsou to čočka, droždí, fazole, hrách a mák. Součet čtverců odchylek je oproti shluku č. 2 mnohem větší, a dá se tedy očekávat, že objekty budou od sebe více vzdáleny. V tomto shluku se nachází potraviny, které jsou si velice podobné. Jedná se o luštěniny. Mák a droždí se nejvíce odchýlily od těžiště shluku.



SHLUK	PRVKY	POPIS
1	10	Vejce 100g
2	1	Banány
2	2	Brambory
2	8	Pomeranče
2	9	Třešně
3	3	Čočka
3	4	Droždí
3	5	Fazole
3	6	Hrách
3	7	Mák

Obrázek 29 - Graf závislosti mezi tuky a kaloriemi (výstup: Shluk)

Obrázek 29, zobrazující graf závislosti tuků a kalorií, potvrzuje, že členy druhého shluku jsou si velice blízké a jsou soustředěny těsně okolo těžiště shluku. O objektech ve shluku č. 2 lze tvrdit, že nízký obsah tuků zároveň znamená i nízký obsah kalorií. U shluku 2 a objektů do něj patřících je sledována vysoká závislost mezi obsahem tuků a kalorií. Prvky hrách, fazole a čočka, patřící do shluku 3, jsou si velice blízké. Jelikož se jedná o potraviny velice podobné a spadající do luštěnin, není tento výsledek příliš překvapující. Další objekty tohoto shluku jsou již rozptýleny a například droždí (4) má mnohem blíže ke shluku číslo dvě. Prvky 3, 5 a 6, tedy čočka, fazole a hrách, mají vysoký obsah kalorií, ale při srovnatelném obsahu tuků jako členy shluku 2. Tedy jsou co se týče energetické hodnoty vysoce výživné, přitom ale hodnota tuků je na velice nízké úrovni. Objekt č.10, 100g vajec tvoří samostatný shluk s vyšším obsahem tuků a celkem nízkým obsahem kalorií.

Vytváření shluků o jednom prvku nemá moc velký význam a proto některé metody dovolují stanovit minimální počet objektů ve shluku. V našem případě by bylo řešením zahrnout do analýzy více objektů, nebo změnit typické body a sledovat, zdali došlo k nějaké výrazné změně. Ideálním shlukem je shluk č.2, kde objekty do něj patřící jsou velice blízko sebe a soustředěny okolo těžiště. U shluku č.3 je pravděpodobné, že prvky mák a droždí by v případě většího počtu objektů, příslušely k jinému shluku, jelikož jejich vzdálenost od těžiště je poměrně značná. Stejný postup byl zvolen u Janceovy metody. Zde se neprojevil v příslušnosti objektů do shluků žádný rozdíl, pouze bylo shlukování ukončeno již po třech iteracích.

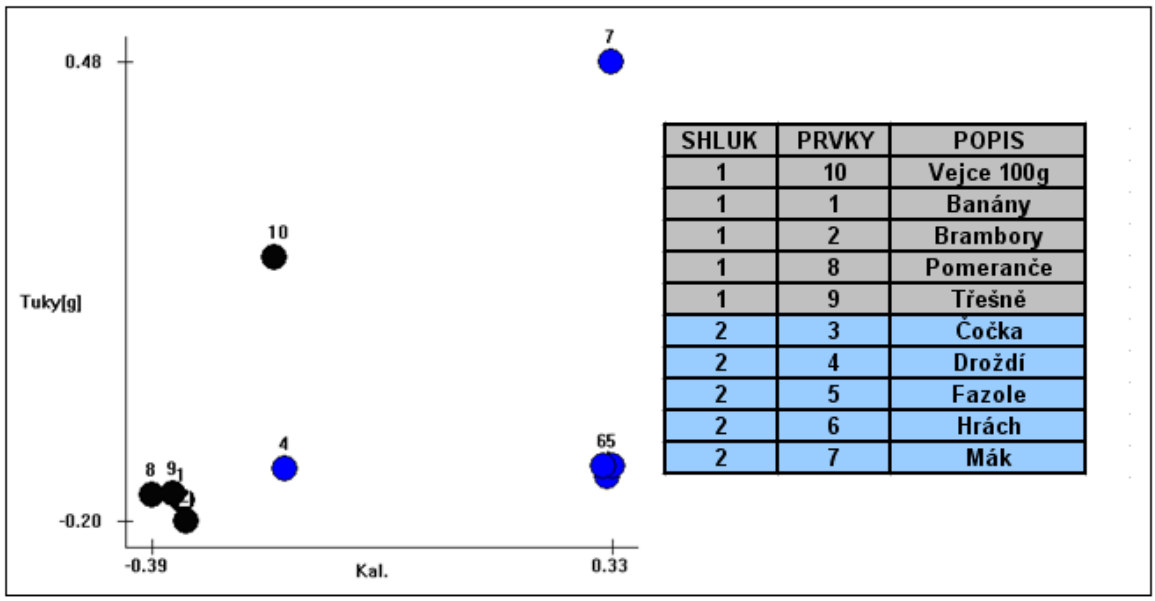
5.3.2. MacQueenova a Wishartova shlukovací metoda

Jak Wishartova, tak i MacQueenova metoda se od předešlých metod liší v tom, že po každém přemístění bodů přepočítávají typický bod skupiny. Tento postup má však za následek závislost výsledku shlukování na uspořádání objektů původní množiny. Nejdříve je popsána MacQueenova metoda. Nejprve je určeno k počátečních bodů. Jak bylo zmíněno v kapitole věnované počátečnímu rozkladu, MacQueen doporučuje vybrat prvních k bodů z libovolně uspořádané množiny objektů. Objekty jsou postupně po jednom přiřazeny k nejbližšímu typickému bodu. Po každém přiřazení nového bodu je přepočítán typický bod skupiny, která se zvětšila o právě přiřazený objekt. Po projití všech objektů a jejich přiřazení k jednotlivým skupinám jsou považována těžiště skupin za typické body. Poté jsou znovu dané objekty přiřazovány k jednotlivým typický bodům, které jsou po každém přiřazení přepočítany. [2]
Wishartova metoda začíná stejným postupem, a tedy zadáním k počátečních typických bodů. Poté jsou objekty přemisťovány vždy k nejbližšímu typickému bodu, přičemž dojde-li k přemístění objektů do jiných skupin, jsou vypočteny nové typické body a to jak u nové zvětšené skupiny, tak i u skupiny zmenšené. Wishart používá konvergentní variantu MacQueenovy metody. Dochází zde k urychlení a zlepšení konvergence, jelikož typické body nejsou přepočítávány vždy, ale pouze dojde-li k přesunu objektů. Obě metody mají stejnou nevýhodu, a to, že výsledek shlukování je ovlivněn počátečním uspořádáním objektů. Rozdíl mezi metodami spočívá v tom, že u MacQueenovy metody je ukončeno shlukování v druhé iteraci, kdežto u Wishartovy metody v okamžiku dosažení stabilního stavu. [5] Vstupní matice je již připravena z předešlých metod. U všech metod je pracováno se standardizovanými i normalizovanými daty. Mezi Wishartovou a MacQuenovou metodou, co se týče výsledku shlukování, nejsou žádné rozdíly. Výsledky jsou přehledně shrnuty na obrázku 30.

```
Jméno matice: WISH
Počet iterací=3

Shluk: 1
Těžiště
-0.32 -0.34 -0.10 -0.32 -0.16 -0.38 -0.36 -0.40 -0.24 -0.25
Prvky:
1; 2; 8; 9; 10;
Součet čtverců odchylek: 16.20

Shluk: 2
Těžiště
0.22 0.29 -0.01 0.28 0.03 0.26 0.23 0.32 0.10 0.10
Prvky:
3; 4; 5; 6; 7;
Součet čtverců odchylek: 16.20
```

Obrázek 30 - Sumarizace Wishardovy metody (výstup:Shluk)

Rozdílem proti předešlým metodám je rozdělení objektů do shluků. Místo tří shluků, které byly pozorovány u Forgyovy a Janceovy metody, zde vznikly pouze dva shluky. Došlo k přesunu objektu č. 10 (100g vejce) do shluku č. 1. Tím došlo k odstranění shluku, jenž byl tvořen pouze jediným objektem, oproti tomu se však zvýšila hodnota součtu čtverců odchylek. Je to logickým následkem přiřazení prvku, jehož vzdálenost od těžiště byla značná, a jiné metody ho proto umisťovaly do samostatného shluku. U MacQuenovy metody je dosaženo totožných výsledků, odlišnost je pouze v počtu iterací, které u MacQuenovy metody byly pouze dvě. Na závěr je pro porovnání opět uveden graf znázorňující závislosti tuků a kalorií, viz obrázek 31.



SHLUK	PRVKY	POPIS
1	10	Vejce 100g
1	1	Banány
1	2	Brambory
1	8	Pomeranče
1	9	Třešně
2	3	Čočka
2	4	Droždí
2	5	Fazole
2	6	Hrách
2	7	Mák

Obrázek 31 - Graf závislosti tuků a kalorií pro proceduru Wish (výstup: Shluk)

Rozložení prvků je stále stejné, pouze prvek 10 – 100g vejce, patří nyní ke shluku č.1. Jelikož jeho vzdálenost od těžiště je poměrně velká, zvyšuje se hodnota součtu čtverců odchylek. Je velice pravděpodobné, že pokud by vstupní matice obsahovala více dat – objektů, tak by prvky 4,10 a 7 spadaly do jiných shluků, než je tomu v dané chvíli.

5.4. Metody s proměnným počtem shluků

Jak již bylo popsáno v předešlé kapitole, počáteční odhad nemusí vždy být ideální odhad počtu shluků. Metody s proměnným počtem shluků se snaží tento problém odstranit, a tedy počet shluků na konci nemusí být roven počátečnímu odhadu počtu shluků. Tedy tyto metody umožňují během shlukování slučování a rozdělování skupin objektů. U metod s proměnným počtem shluků nestačí pouze zadání počátečního počtu k skupin, ale musí být určeny další řídící parametry. Na základě těchto parametrů je rozhodováno zdali dojde k sloučení či rozdělení skupiny objektů. Tyto parametry je možné zadat dvojím způsobem. Buď za pomoci analytika, který využije svých znalostí s daným typem úloh, nebo provede několik testovacích pokusů. Tyto parametry se během shlukování nemění. Dalším způsobem je přímý výpočet parametrů z dat, která jsou analyzována. Těchto parametrů využívá například metoda CLASS. [5] Program Shluk nabízí z metod s proměnným počtem shluků pouze metodu CLASS. Existují další metody, mezi které patří metoda RELOC, ISODATA či McQuenova metoda se dvěma parametry. Tyto metody jsou popsány na přiloženém CD.

5.4.1. Metoda CLASS

Metoda CLASS je modifikací metody ISODATA a od této i jiných metod se liší hlavně tím, že většina parametrů není zadávána analytikem, nýbrž jsou počítány automaticky ze zadaných dat. Jediné tři parametry zadává analytik. Jsou to parametry GAMA, který určuje maximální počet iterací, THETAN definující minimální počet objektů ve skupině a parametr Sn udávající počáteční rozdělovací práh. Parametr Sn tedy řídí rozklad shluků a při každém opakování cyklu se zvýší a tedy kritérium rozkladu se zpřísňuje. Metoda CLASS končí při dosažení stabilního rozkladu, nebo v případě, že je dosaženo GAMA iterací. [4] Před samotnou iterační částí je potřeba vygenerovat 2s+1 typických bodů. Tyto body generuje metoda CLASS sama. Samotný postup je poté následující. Prvním krokem je vyloučení malých shluků, to je skupin, kde je počet objektů menší, nežli parametr Thetan a zároveň u nich během posledních dvou iterací nedošlo ke změně. Tyto skupiny jsou zrušeny. [5] Následuje druhý krok, a to rozdělení skupin. V i-té iteraci je vypočten rozdělovací práh.

$$S_n = S_{n-1} + \frac{1 - S_0}{GAMA}$$

Rovnice 23 - Výpočet rozdělovacího prahu (zdroj: [5])

Pro každý shluk jsou vypočteny nové souřadnice objektů, a to jako odchylky od souřadnic těžišť, tedy typických bodů. Nyní je vypočtena průměrná odchylka, j-té souřadnice pro objekt ležící vpravo od těžiště (Dj1) a vlevo od těžiště (Dj2). [5]

$$\mathrm{D}_{j1} = \frac{1}{k_1}\sum_{i=1}^{k_1} x_{ij} \quad -\mathrm{D}_{j2} = \frac{1}{k_2}\sum_{i=1}^{k_2} x_{ij}$$

Rovnice 24 - Výpočet průměrné odchylky (zdroj: [5])

Kde k1 a k2 udávají počet prvků vlevo, respektive vpravo od objektu. Dále jsou vypočteny parametry a1 a a2, pro body, které se nacházejí vpravo či vlevo od těžiště.

$$a_1 = \max_j \frac{\mathrm{D}_{j1}}{\max x_{ij}} \; a_2 = \max_j \frac{\mathrm{D}_{j2}}{\max x_{ij}}$$

Rovnice 25 - Výpočet parametrů A1, A2 (zdroj: [5])

Parametr j nabývá hodnot 1,2,….,p,j a prochází všechny indexy bodů dané skupiny. Je-li v n-té iteraci počet shluků menší, než 2K a Sm < a1 nebo Sm < a2 a zároveň počet objektů větší jak 2*(THETAN +1), rozdělí se shluk na část vlevo a vpravo od j-té souřadnice těžiště, kdy a1 nebo a2 nabylo svého maxima. Třetím krokem je zrušení skupin. Zde je nejprve potřeba spočítat průměrnou minimální vzdálenost skupin, označovanou TAU. [4]

$$TAU = \frac{1}{h}\sum_{i=1}^{h} \mathrm{D}_i$$

Rovnice 26 - Výpočet parametru TAU (zdroj: [5])

h – současný počet skupin (shluků)
Di – minimální vzdálenost i-tého shluku (i-té skupiny) od ostatních h-1 shluků, která vyjadřuje vzdálenosti jednotlivých těžišť.

Pokud je minimální vzdálenost těžišť menší jak parametr TAU a zároveň počet shluků větší jak K/2, i-tý shluk je zrušen. Objekty, které do daného shluku patřily, jsou přiřazeny do shluku s nejbližším těžištěm. Stejně jako metoda ISODATA, i metoda CLASS prošla vývojem a i zde se jednotlivé uváděné postupy mohou trochu lišit. [5]

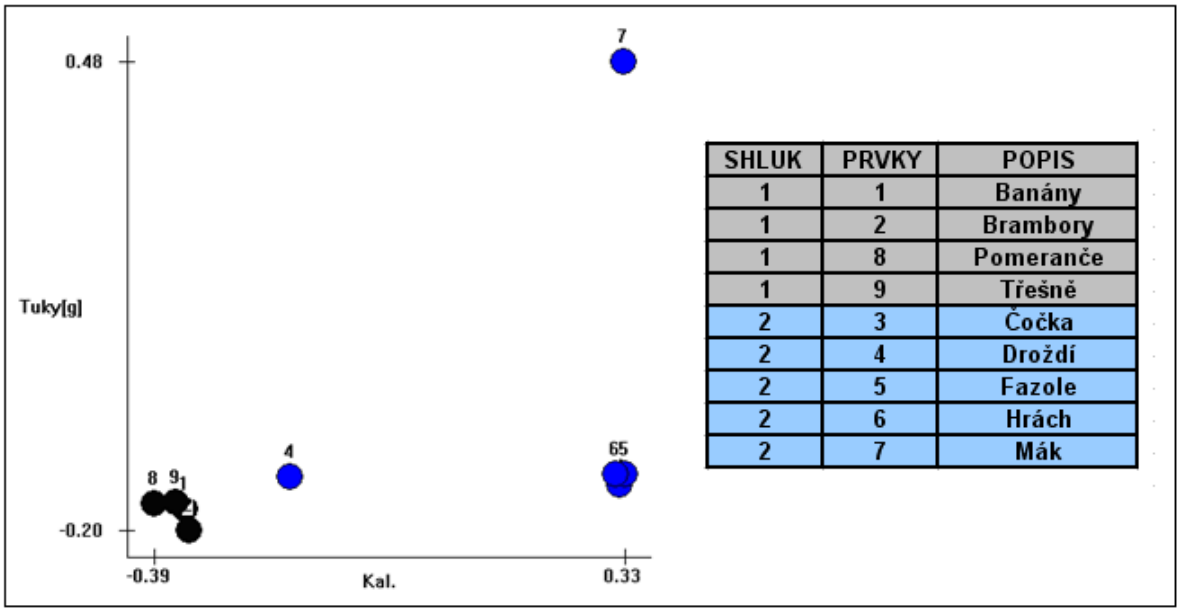
Literatura [5] udává, že na rozdíl od algoritmu, který byl popsán výše a generuje 2S+1 typických bodů, může být zadána libovolná skupina typických bodů. K jejímu určení může být použit libovolný způsob. Viz kapitola věnující se počátečnímu rozkladu.

Prvním krokem této metody je volba minimálního počtu iterací. Podle zkušeností z předešlých metod jsou nastaveny minimálně tři iterace a ve stejném kroku je zvolen minimální počet objektů ve shluku na dva. Tím je vyloučena možnost, vzniku shluku o jednom objektu. Poté je provedeno samotné shlukování a jeho výsledky zobrazeny, viz obrázek 32.

```
Jméno matice: CLASS

Počet iterací: 4

Shluk: 1
Těžziště:
 A 1: -0.35  A 2: -0.43  A 3: -0.18  A 4: -0.23  A 5: -0.16  A 6: -0.42  A 7: -0.37  A 8: -0.38  A 9: -0.25  A 10: -0.22
Prvky
1; 2; 8; 9;
Počet prvků ve shluku: 4
Součet čtverců odchylek: 0.02
Průměrná vzdálenost od těžiště: 0.07

Shluk: 2
Těžziště:
 A 1:  0.22  A 2:  0.29  A 3: -0.01  A 4:  0.28  A 5:  0.03  A 6:  0.26  A 7:  0.23  A 8:  0.32  A 9:  0.10  A 10:  0.10
Prvky
3; 4; 5; 6; 7;
Počet prvků ve shluku: 5
Součet čtverců odchylek: 2.72
Průměrná vzdálenost od těžiště: 0.74


Celkový součet čtverců odchylek = 2.74
Vzájemné vzdálenosti těžišť
0.00
1.65  0.00
```

Obrázek 32 - Sumarizace metody Class (Výstup: Shluk)

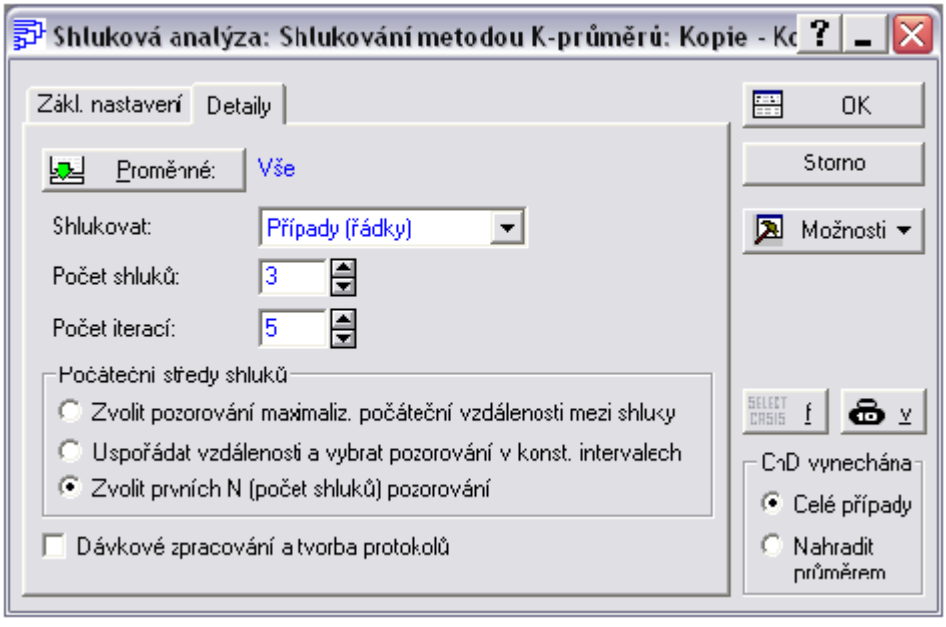
Shlukování bylo ukončeno po čtvrté iteraci a objekty byly rozděleny do dvou shluků. Objekt 100g vejce, nebyl přiřazen ani do jedno ze shluků. To je rozdíl oproti Forgyově a Janceově metodě, kde tento objekt tvořil samostatný shluk. Jinak se rozdělení objektů do shluků v ničem neliší. Tedy i graf závislosti tuků na kaloriích bude totožný, pouze s tím rozdílem, že se v něm nebude vyskytovat prvek 10 – 100g vejce, tato skutečnost je zobrazena na obrázku 33.



SHLUK	PRVKY	POPIS
1	1	Banány
1	2	Brambory
1	8	Pomeranče
1	9	Třešně
2	3	Čočka
2	4	Droždí
2	5	Fazole
2	6	Hrách
2	7	Mák

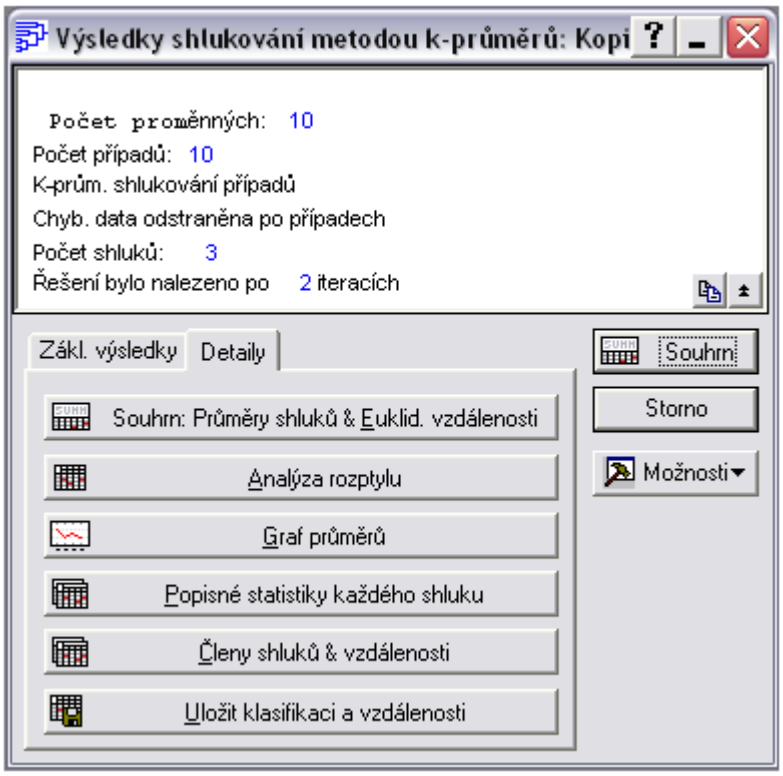
Obrázek 33 - Graf závislosti tuků a kalorií pro metodu Class (Výstup: Shluk)

Výhodou této metody je možnost nastavení parametru, minimálního počtu objektů ve shluku. Tím se eliminuje možnost vzniku shluků s jediným objektem. V porovnání s ostatními metodami se výsledky shlukování neliší, je dosahováno stejných hodnot, ať už u vzdálenosti od těžišť či součtu čtverců odchylek. Pro názornost bylo provedeno shlukování programem Statistica a výsledky porovnány s výstupy z programu Shluk. Shlukována byla data ze základního souboru. Vstupní matice je načtena z programu MS Excel. Jedná se o standardizovaná a normalizovaná data.



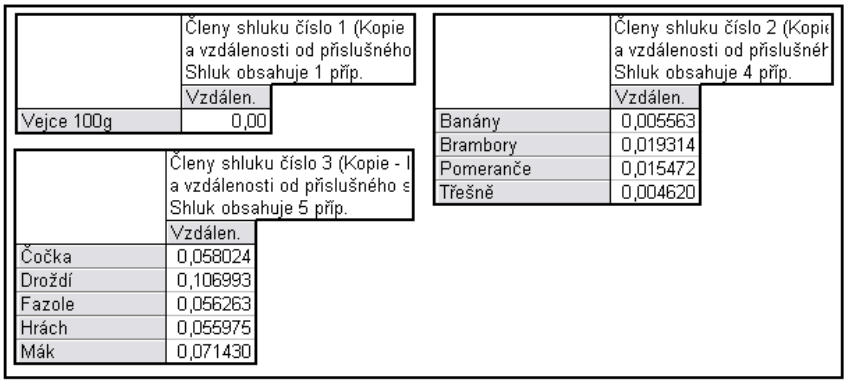
Obrázek 34 - Zadávání vstupních parametrů (Výstup: Statistica)

Obrázek 34 ukazuje počáteční nastavení před samotným zahájením shlukování. Je zvoleno shlukování do 3 shluků a maximální počet iterací bude pět. Z předešlých metod víme, že takovýto počet iterací je dostačující. Výběr typických bodů bude proveden metodou prvních n bodů.



Obrázek 35 - Výsledný přehled shlukování (Výstup: Statistica)

Obrázek 35 ukazuje, že shlukování bylo ukončeno po dvou iteracích a objekty byly rozděleny na základě vstupního požadavku do tří shluků. Nyní se nabízí možnosti zobrazit si více detailů ze samotného shlukování. V tuto chvíli je dostačující zobrazení členů shluků, viz obrázek 36 a grafické znázornění závislosti tuků a kalorií, která je zobrazena na obrázku 37.



Členy shluku číslo 1 (Kopie a vzdálenosti od příslušného Shluk obsahuje 1 příp.

	Vzdálen.
Vejce 100g	0,00

Členy shluku číslo 2 (Kopie a vzdálenosti od příslušnéh Shluk obsahuje 4 příp.

	Vzdálen.
Banány	0,005563
Brambory	0,019314
Pomeranče	0,015472
Třešně	0,004620

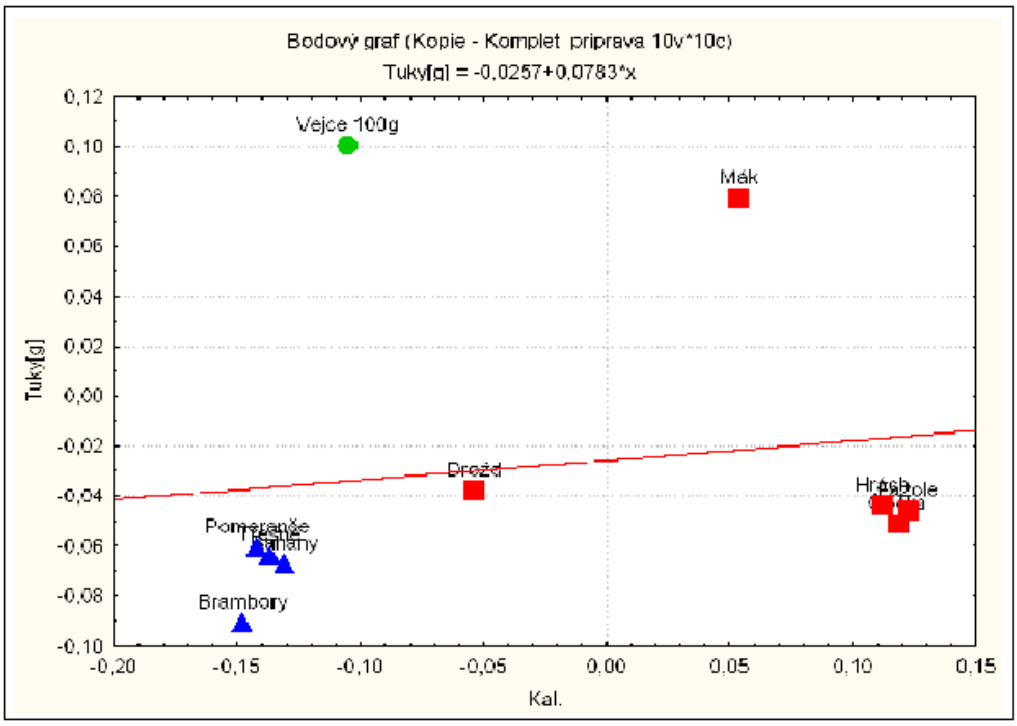
Členy shluku číslo 3 (Kopie - I a vzdálenosti od příslušného s Shluk obsahuje 5 příp.

	Vzdálen.
Čočka	0,058024
Droždí	0,106993
Fazole	0,056263
Hrách	0,055975
Mák	0,071430

Obrázek 36 - Členy shluků (Výstup:Statistica)

Je dosaženo totožných výsledků, jako u předešlých metod, kde bylo shlukováno do třech shluků. Výhodou programu Statistica oproti programu Shluk je možnost pracovat s rozsáhlejšími vstupními maticemi. Nevýhodou je omezený počet metod ve zkušební verzi tohoto programu. Co se týče grafického výstupu, umožňuje program Statistica přímo popisovat jednotlivé objekty, ale v případě, že jsou tyto objekty blízko sebe, nebo je-li jejich počet vysoký, jsou popisky nepřehledné.

Na obrázku 37, je zobrazeno rozložení objektů do shluků. Opět není pozorována výraznější změna, oproti výstupu z programu Shluk.



Obrázek 37 - Závislost tuků a kalorií (Výstup:Statistica)

Při větším počtu objektů a jejich vzájemné blízkosti dochází ke ztrátě přehlednosti. Jednotlivé objekty se navzájem překrývají a jejich identifikace není možná ani za pomoci popisků, které danou situaci tvoří ještě více nepřehlednou. Proto je důležité vnímat výstupy spíše z hlediska celkového rozložení shluků. Při velkém počtu objektů je potřeba daný výstup prezentovat na dostatečně velké ploše. Proto jsou výstupy v práci uloženy v elektronické podobě na CD, aby jejich zkoumání bylo možné na monitoru či projektoru. V tištěné verzi mají spíše informativní charakter.

6. Metodika zpracování komplexního příkladu

Pro svoji obsáhlost je příklad přiložen na CD. Prvním krokem zpracování zadané úlohy je výběr dat, která budou shlukována. Data tvoří vstupní matici, na které jsou prováděny veškeré úpravy. Tato data byla stažena ze serveru www.crohn.cz a zabývají se energetickými hodnotami jednotlivých potravin. Veškeré údaje jsou uloženy v podobě tabulky v MS Excel s názvem Vstupní data.xls na CD ve složce Vstupní data. Listy nejsou uzamčeny a je tedy možné na nich provádět veškeré úpravy. U dat byla provedena jejich transformace, standardizace a normalizace. Tím byla odstraněna závislost na jednotkách a rozdílech hodnot u jednotlivých vlastností zkoumaných objektů. Transformace byla provedena v programu MS Excel a transformovaná data byla využita jako vstup do programu Statistica. Tato data jsou umístěna ve složce Předzpracování dat v souboru Transformace dat.xls. Ve stejném souboru jsou uložena i netransformovaná data, je tedy možné provádět analýzy i na nich a sledovat rozdíly při tvorbě shluků. Ve složce Předzpracování dat jsou dále umístěny výstupy z programů Statistika a Shluk. Jedná se obrázky a grafy, které jsou vždy pojmenovány podle zkoumané problematiky. Například soubor Korelace tuky kalorie.jpg zobrazuje závislost korelací pomocí grafického výstupu z programu Statistica. Poté co byla data vybrána a transformována následovalo jejich shlukování. Metoda Kmeans byla prováděna v programu Statistica. Ostatní metody byly prováděny v programu Shluk. Jako vstupní data byla využita Vstupní matice, která je umístěna v souboru Transformace dat (soubor Předzpracování dat), na listu Vstupní matice. Jednotlivé kroky pro práci s programem jsou vždy reprezentovány pomocí obrázků a popisu. Tyto výstupy jsou uloženy ve složce Výstupy z programů. Tato složka je dále rozdělena na další soubory, podle toho, z jakého programu daný výstup pochází a jakou metodou byla data zpracována. Například grafické výstupy Forgyovy metody prováděné v programu Shluk jsou umístěny ve složce Výstupy z programů, Shluk a Forgyova metoda. Ve složce Metriky se nachází soubor Metriky.xls, kde jsou vypočteny ze základních dat nejčastěji používané metriky. Zde je možné měnit jednotlivé hodnoty a sledovat změny u jednotlivých metrik. V souboru Komplexní příklad je umístěn dokument Analýza příkladu.doc, který vysvětluje celkový postup zpracování dat od jejich zisku až po výsledky shlukování a slouží jako návod pro zpracování dat, která jsou na CD umístěna. Ve stejné složce je umístěn i soubor Komplexní příklad.st, který je určen k otevření v programu Statistica. Složka Ostatní metody obsahuje popis dalších nehierarchických metod, které nejsou v práci uvedeny. Pomocí těchto metod neumožňoval shlukovat žádny z testovaných programů.

7. Závěr

Cílem práce bylo seznámení s postupy nehierarchických metod shlukové analýzy a navrhnout praktickou ukázku řešení příkladu pomocí těchto metod. Metody byly vysvětleny na příkladu zabývajícím se shlukováním potravin dle hodnot obsahu jednotlivých látek. Vybraná data byla použita pro svoji jednoduchost a srozumitelnost. Původním záměrem práce bylo provést shlukování za pomoci jediného programu a v daném programu porovnat jednotlivé shlukovací metody. Prvním programem, který byl vyzkoušen, byl program Unistat. Zkušební verze tohoto programu však neumožňuje práci s vlastními daty. Analýzy tedy lze provádět pouze na datech dodaných Unistatem, jež jsou ovšem bez jakéhokoliv vysvětlení. Další nevýhodou tohoto programu je nemožnost přímého popisu výstupů. Druhým testovaným programem byl program Statistica. Tento program nabízí velké množství nástrojů, a to již ve zkušební verzi. K programu lze později dokupovat velké množství komponent. Bohužel ve zkušební verzi nabízí pouze jedinou nehierarchickou metodu. Ohledně práce s výstupy nabízí program Statistica mnoho možností. Pro účely práce byly důležité především grafy závislostí. Oproti Unistatu, program Statistica umožňuje jednotlivé objekty popisovat, volit barvu i tvar značek zobrazující objekty. Bohužel při velkém množství dat se grafy stávají nepřehlednými a jednotlivé objekty se překrývají. Zároveň, při zapnutých popiscích dat, nelze při velkém množství prvků tyto popisky přečíst, jelikož jsou velice blízko u sebe. Grafy jsou nečitelné jak v tištěné podobě, tak i na monitoru. Řešením je možnost výřezu shuků, ale tím se ztrácí celkový přehled o shlucích. Druhou možností je prezentace výsledků na větších zobrazovacích plochách, například promítání na velká plátna. Z důvodu nedostatku metod v programu Statistica bylo pro vyzkoušení více nehierarchických metod využito programu Shluk. Tento program sice nabízí více metod, ale neumožňuje pracovat s rozsáhlými vstupními maticemi. Program Shluk je schopen pracovat maximálně s maticí 10x10. Ohledně grafických výstupů dopadl program Shluk velice dobře. Jednotlivé shluky automaticky barevně odlišuje, u ostatních programů toto muselo být prováděno ručně. Jednotlivé prvky popisuje pomocí čísel, která vždy přísluší danému objektu. Pokud má být přímo vidět o jaký objekt se jedná, je potřeba vedle grafu umístit legendu. Posledním programem, jenž byl na analyzování využíván, byl MS Excel. Ten byl využíván především na základní úpravy, transformace dat nebo výpočet metrik. Z programů, které byly vyzkoušeny, nabízí nejvíce možností program Statistica. Je potřebné dokoupit komponenty, aby tento program disponoval více metodami. Program Shluk nabízí velké množství metod, ale limituje ho omezení pro vstupní data. Při samotném shlukování bylo vyzkoušeno více metod, aby mohly být sledovány rozdíly mezi jednotlivými metodami. Byly provedeny veškeré kroky od výběru dat, jejich přípravy až po samotné shlukování. Jednotlivé úpravy byly popisovány a veškeré důležité grafické výstupy umístěny na CD. Během práce nenastaly větší problémy. Více času zabralo seznámení se s programem Statistica a vysledování možností jednotlivých programů.

Seznam vzorců:

Seznam obrázků:

Použitá literatura:

[1] – *Aritmetický průměr* [online]. Wikipedie, 25.2.2009 [cit. 2009-04-19]. Český. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Aritmetický_průměr>.

[2] – HYNAR, Martin. Metody shlukování. Ostrava, 2003. 23 s. Bakalářská práce.

[3] – JONÁŠOVÁ, Hana. *Zpracování dat metodami shlukové analýzy*. Učební tet k předmětu Zpracování dat metodami shlukové analýzy na FES-UPCE

[4] –KUČERA, Jiří. *Metody katagorizace dat*. Brno, 2008. 36 s. Vedoucí bakalářské práce Matěj Štefánik. DostupnýWWW: <http://is.muni.cz/th/172767/fi_b/Metody_kategorizace_dat.txt>.

[5] – LUKASOVÁ, Alena, ŠARMANOVÁ, Jana*. Metody shlukové analýzy*. Praha : SNTL, 1985. 208 s.

[6] – *Manhattan-Metrik* [online]. Wikipedie, 17.7.2008 [cit. 2009-04-19]. Německý. Dostupný z WWW: <http://de.wikipedia.org/wiki/Manhattan-Metrik>.

[7] – *Medián* [online]. Wikipedie, 23.2.2009 [cit. 2009-04-19]. Český. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Medián>.

[8] – MELOUN, Milan, MILITKÝ, Jiří, HILL, Martin. *Počítačová analýza vícerozměrných dat v příkladech*. 1. vyd. Praha : Academia, 2005. 449 s. ISBN 80-200-1335-0.

[9] – MELOUN, Milan, MILITKÝ, Jiří. *Kompedium statistického zpracování dat : Metody a řešené úlohy.* 2. přepracované a rozšířené vyd. Praha : Academia, 2006. 984 s., CD. ISBN 80-200-1396-2.

[10] – *Modus* [online]. Wikipedie, 14.2.2009 [cit. 2009-04-19]. Český. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Modus>.

[11] – *Popisná statistika* [online]. [cit. 2009-04-19]. Český. Dostupný z WWW: <www.kvetakov.net/down.php?file=./FEIKRT/1/zs/INTPE/u11.pdf>.

[12] – ŘEZANKOVÁ, Hana, HÚSEK, Dušan, SNÁŠEL, Václav. *Shluková analýza dat*. 1. vyd. Praha : Professional Publishing, 2007. 196 s. ISBN 978-80-86946-9

[13] – *Směrodatná odchylka* [online]. Wikipedie, 9.4.2009 [cit. 2009-04-19]. Český. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Směrodatná_odchylka>.

[14] – ŠŤASTNÝ, František. *Popisné statistiky* [online]. 1997 [cit. 2009-04-19]. Český. Dostupný z WWW: <http://amper.ped.muni.cz/jenik/nejistoty/html_tree/node13.html>.

[15] – *Vážený průměr* [online]. Wikipedie, 15.2.2009 [cit. 2009-04-19]. Český. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Vážený_průměr>.

[16] – ŽÁK, Libor. Shluková analýza I.. *Automatizace*. 2004, roč. 47, č. 3, s. 180-182.