Fuzzy multivariační analýza chemických dat

Řanda, Bohdan
2002

Vysoká škola chemicko-technologická v Praze

Fakulta chemicko-inženýrská

Ústav analytické chemie

DIPLOMOVÁ PRÁCE

Fuzzy multivariační analýza chemických dat

Vedoucí diplomové práce: **Prof. Ing. Miloslav Suchánek, CSc.**

Konzultant: **Ing. Vojtěch Radil, CSc.**
Ing. Jaromír Kukal Ph.D.

Vypracoval: **Bohdan Řanda**

Praha 2002

Prohlašuji, že jsem předloženou diplomovou práci vypracoval samostatně a použil jen pramenů, které cituji v seznamu literatury.

V Praze 10. května 2002 ……

podpis

Rád bych poděkoval Ing. Vojtěchu Radilovi, CSc a Ing. Jaromíru Kukalovi, Ph.D. za cenné rady a náměty při sepisování této práce. Dále bych rád poděkoval Centru aplikované kybernetiky CAK, při FEL ČVUT v Praze za jejich rady a práci na zdrojovém programu v rámci naší společné spolupráce.

OBSAH:

SEZNAM NEJČASTĚJI POUŽÍVANÝCH ZKRATEK

N	-počet clusterovaných objektů
c	-počet clusterů
p	-dimenze, počet znaků vektorů
u_{ik}	-funkce příslušnosti k-tého objektu vůči i-tému clusteru
a	-základ dekadického logaritmu
e	-Eulerovo číslo
XB	-index validity Xie-Beni
FS	-index validity Fukuyama-Sugeno
Fc	-koeficient rozkladu
Hc	-entropie rozkladu
Hc_norm1	-normalizovaná entropie rozkladu 1
Hc_norm2	-normalizovaná entropie rozkladu 2
Fc_norm	-normalizovaný koeficient rozkladu
Hc_stand	-standardizovaná entropie rozkladu
Fc_stand	-standardizovaný koeficient rozkladu

PŘEDMLUVA

Tématem této diplomové práce je v české literatuře nepříliš popsaná fuzzy shluková analýza. V práci je popsán jeden z nejužívanějších algoritmů fuzzy C-means (FCM).
Této techniky bylo užito na zpracování dat z imisního monitoringu Českého hydrometeorologického ústavu.
Větší část teoretické části je výběrem z mnoha původních článků. U některých tvrzení je uveden jejich důkaz. Pokud je důkaz relativně krátký a srozumitelný, je důležitou součástí výkladu, bez něhož nelze probíranou látku bezezbytku pochopit.
Zdrojový kód je napsán ve výpočetním systému Matlab. Jeho hlavní část vznikla v Centru aplikované kybernetiky při FEL, ČVUT Praha. Jeho upravená verze byla použita při zpracování dat dvou automatických stanic Měděnec a Chomutov v Severních Čechách za rok 1997.
V celém textu je místo shluková, shluk apod. použito výrazu clusterová, cluster apod.
V příloze je z ilustrativních důvodů uveden zdrojový kód programu na úpravu databáze v jazyce C a dále vlastní zdrojový kód FCM.

1. ÚVOD

1.1 FUZZY PŘÍSTUP K PROBLEMATICE

V současné době se v celé řadě lidských činností používá nebo spíše začíná používat teorie fuzzy množin. Začalo to v obecné teorii systémů a v regulační technice a dnes se fuzzy množiny používají v ekonomii, v lékařské diagnostice, k popisu činnosti chemických reaktorů apod. V praktických aplikacích dnes vynikají zejména elektronické konstrukční prvky s fuzzy logikou.
Historie množin obecně začíná ve staré antice, kde skupina řeckých filosofů se zabývala matematickou stránkou výroků z hlediska formální logiky. Od dob Aristotelových se učíme v logice, že tvrzení, může být buď pravdivé nebo nepravdivé, třetí možnost není. První kdo se začal po filozofické stránce zabývat jiným než pravděpodobnostním pohledem na neurčitost, a který zavedl pojem „vagueness“ (vágnost) byl americký filozof Max Black.
V roce 1965 publikoval L. Zadeh[1] svůj článek s názvem „Fuzzy sets“, který dal fuzzy množinám jméno, a který je všeobecně považován za začátek éry fuzzy množin. Celá moderní matematika je založena na teorii množin. Slabým prvkem této teorie je rozhodnutí zda prvek patří či nepatří do dané množiny. Každá množina má svou charakteristickou funkci, která nabývá hodnoty 1 jestliže prvek patří do množiny a 0 jestliže prvek do množiny nepatří. Jenže nalezení hodnot charakteristické funkce u většiny aplikací je problém ležící většinou mimo matematiku a často vůbec těžko rozhodnutelný. Touto oblastí, kdy funkce příslušnosti prvku množiny nabývá hodnot z intervalu od [0,1] a prvek tak do množiny „spíš“ patří – nepatří se zabývá teorie fuzzy množin.
Clusterová analýza je metoda kvantitativního vyjádření podobnosti jevů, objektů a následně jejich zařazení do shluků. Klasifikace je činnost vytvářející rozklad nějaké množiny objektů za účelem vytvoření systému tříd. Chápeme-li vzniklý systém tříd opět jako množinu hodnou klasifikace, jedná se hierarchický přístup shlukování. Naopak nehierarchický přístup hledá rozklad množiny podle vhodně zvoleného kritéria optimality rozkladu nebo využívá pravděpodobnostní přístup.
V klasické clusterové analýze je objekt přiřazen právě jedné třídě dat, právě jednomu clusteru. Rozhodnutím ano/ne u odlehlých a hybridních hodnot velmi ztěžuje zařazení bodu do třídy.
Clusterová analýza s prvky fuzzy dosahuje lepších výsledků ve srovnání s normální clusterovou analýzou díky tomu, že objekt může patřit do vícero shluků zároveň. To jak „hodně“ objekt do clusteru patří, je dáno hodnotou funkce příslušnosti. Při klasifikaci pomocí fuzzy clusterové analýzy lze použít několika algoritmů, někdy se i výrazně lišících. Nejznámější a nejvíce používaný je fuzzy C-means algoritmus FCM.
FCM metoda patří mezi tzv. „unsupervised“ metody bez učitele, které využívají cílové funkce a určení jejího minima. Funkce je vyjádřená v podobě sumy vážených vzdáleností. Váhou je v tomto případě čtverec hodnot funkce příslušnosti. Algoritmus FCM je někdy nazýván ISODATA podle původního algoritmu shlukování bez fuzzy prvků založeném na stejném principu.
Objekt (vektor naměřených proměnných) může patřit do několika tříd (clusterů) zároveň. Příslušnost k danému clusteru je určena hodnotou funkce příslušnosti. Clustery jsou charakterizovány prototypy (centroidy), které určují centry daných clusterů. Tato technika předpokládá apriorní volbu počtu clusterů. Je to iterativní

postup. V každé iteraci je použito předchozích hodnot matice funkcí příslušnosti pro výpočet nové hodnoty. Matice hodnot příslušnosti se použije pro výpočet zlepšené hodnoty centroidů. Jak bylo řečeno, je nutná volba clusterů. Tato skutečnost je slabým místem algoritmu, vhodný počet clusterů je subjektivní prvek. Nejvhodnějším nástrojem pro validitu fuzzy datového souboru jsou indexy validity.

CHEMICKÁ DATA

Analýzou chemických dat se zabývá samostatná disciplína Chemometrie. Chemometrie používá matematické metody s cílem navrhovat optimální experimentální postupy a získávat maximum relevantních informací z pokusných výsledků. Uplatňuje se při získávání ekologických informací, v klinické biochemii a jiných mezioborových praktických problémech. Přehled většiny metod zpracování experimentálních chemických dat podává kniha M. Meloun[2]. Použití „fuzzy-technik" v Chemometrii nalézá široké uplatnění. V současné době je teorie fuzzy množin zavedenou a propracovanou matematickou disciplinou.

PRAKTICKÉ POUŽITÍ FUZZY METOD

Hlavní pole působnosti analýzy dat obecně se dá shrnout do několika oblastí jako je :
Analýza trendu – faktorová analýza, diskriminační analýza, regresní analýza.
Klasifikace – clusterová analýza, neuronové sítě, rozpoznávání obrazu.
Ve všech těchto oblastech nacházejí fuzzy metody velké uplatnění, zejména pak v clusterové analýze.

1.2 FUZZY MNOŽINY V CLUSTEROVÉ ANALÝZE

V roce 1965 vyšel v časopise Information Control článek „Fuzzy sets" od elektroinženýra působícího v Berkeley v Kalifornii L. A. Zadeh[1]. Tento článek odstartoval řadu prací na tomto poli až do dnešních dob. Autor v článku zavádí základní pojmy a vztahy, které jsou přehledně vysvětleny. Fuzzy množina A je množina „x", kde ke každému „x" existuje f(x) z intervalu [0,1], f(x) je funkce příslušnosti. Např. A je množina reálných čísel větších než 1, potom funkce příslušnosti může být: f(0)=0; f(1)=0; f(10)=0,2; f(100)=0,95; f(500)=1.
A je prázdná fuzzy množina, když pro všechny „x" je f(x)=0. Dvě fuzzy množiny A,B se rovnají, je-li pro všechna „x" $f_A(x)=f_B(x)$, zkráceně $f_A=f_B$. Doplněk fuzzy množiny A je fuzzy množina A´ a je definována: $f_{A'} = 1- f_A$.
A je podmnožinou B platí-li:

(1) $$A \subset B \Leftrightarrow f_A \leq f_B$$

Sjednocením fuzzy množin A,B je fuzzy množina C, $C = A \cup B$
jejíž funkce příslušnosti je:

(2) $$f_C(x) = \max\left[f_A(x), f_B(x)\right] \quad zkráceně \quad f_C = f_A \vee f_B$$

Průnikem dvou fuzzy množin A,B je množina A, $C = A \cap B$
pro níž platí:

(3) $$f_C(x) = \min\left[f_A(x), f_B(x)\right] \quad zkráceně \quad f_C = f_A \wedge f_B$$

Pro takto zavedené sjednocení a průnik platí obdobně jako pro „klasické“ množiny Morganova pravidla a distribuční zákon.

Morganova pravidla:

$$(4) \qquad (A \cup B)' = A' \cap B'$$
$$(5) \qquad (A \cap B)' = A' \cup B'$$

Distribuční zákon:

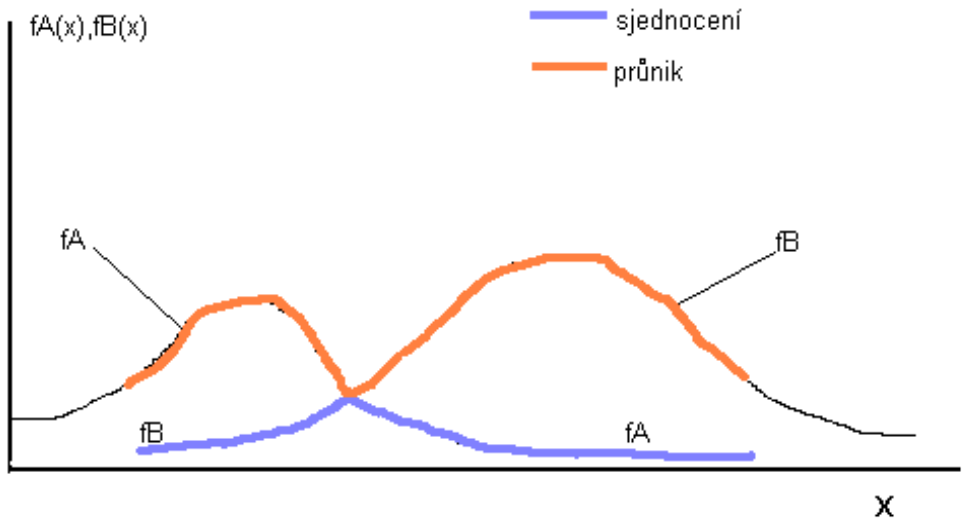
$$(6) \qquad C \cap (A \cup B) = (C \cap A) \cup (C \cap B)$$
$$(7) \qquad C \cup (A \cap B) = (C \cup A) \cap (C \cup B)$$

Pro funkce příslušnosti podle (4) a (7) platí:

$$(8) \qquad 1 - \max[f_A, f_B] = \min[1 - f_A, 1 - f_B]$$
$$(9) \qquad \max[f_C, \min[f_A, f_B]] = \min[\max[f_C, f_A], \max[f_C, f_B]]$$

Na následujícím obrázku obr 1. je ilustrace sjednocení a průniku dvou fuzzy množin.

obr 1.
Sjednocení a průnik fuzzy množin



Konvexní množina
Předpokládejme, že „X“ je n-rozměrný Euklidův prostor E^n. Fuzzy množina A je konvexní množina je-li:

$$(10) \qquad \Gamma_\alpha = \{x\,; \quad f_A(x) \geq \alpha\}$$

také konvexní množina pro všechna alfa na intervalu (0,1). Z formální hlediska je přesnější následující definice konvexnosti.
Množina A je konvexní, platí-li:

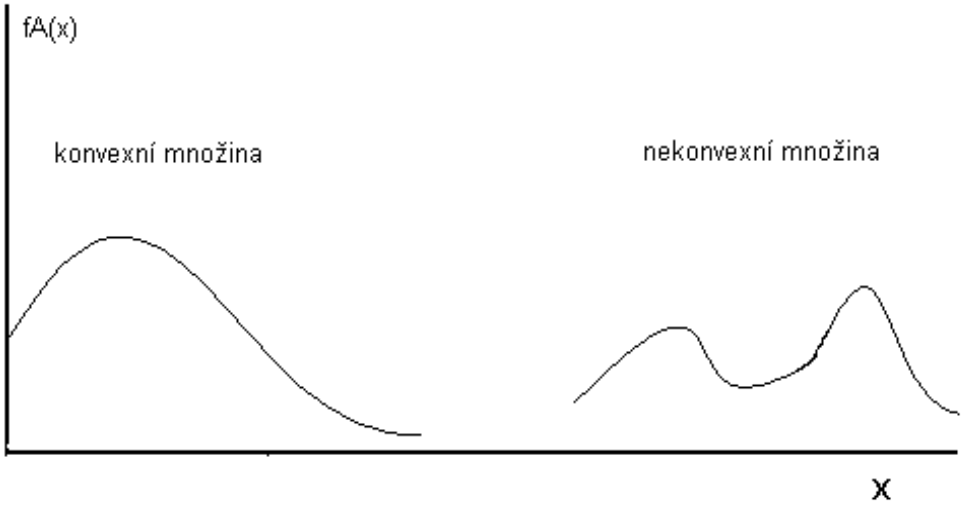
$$(11) \qquad f_A[\lambda x_1 + (1-\lambda)x_2] \geq \min[f_A(x_1), f_A(x_2)]$$

pro všechna α a $\lambda \in [0,1]$.

Jsou-li A,B konvexní, potom také jejich průnik je konvexní množina.

Celou situaci nejlépe vystihuje následující obrázek obr 2.:

obr 2.
Konvexnost fuzzy množin



Po uveřejnění tohoto článku na sebe aplikace ve fuzzy clusterové analýze nenechaly dlouho čekat. V roce 1969 vychází článek od matematika Enrique H. Ruspiny[3,4], který ve své práci zobecnil pojem redukce dat jako zobrazení z množiny dat do množiny reprezentantů. Ukázal jak přístup „fuzzy“ řeší do té doby existující problémy u používaných algoritmů. V zápětí na to vydává roku 1970 práci, kde podrobně rozebírá řešení optimálního rozkladu množiny dat coby hledání lokálních (globálních) extrémů zvolených funkcí (funkcionálů) s využitím pravděpodobnostního přístupu. Místo funkcí příslušnosti fuzzy množin ještě zatím operuje s hustotou pravděpodobnosti. Obě tyto práce obsahují zobecnění principů, ze kterých vychází fuzzy C-means. Bohužel pro kompletní pochopení autorových myšlenek je zapotřebí hlubších matematických znalostí.
Již před publikováním prvních prací z teorie fuzzy množin se jako kritérium kvality rozkladu používala suma čtverců odchylek od centroidu (reprezentanta) shluku tzv. WGSS funkcionál - „withing group of sums squared errors“.

$$J_w(U,\bar{v}) = \sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}(d_{ik})^2$$

kde $\bar{v}$ je matice centroidů (c *x* p), c je počet shluků, p-rozměr shlukovaných objektů, u_{ik} je funkce příslušnosti.

$$d_{ik} = d(x_k, v_i) = \|x_k - v_i\| = \sqrt{\sum_{j=1}^{p} (x_{kj} - v_{ij})^2}$$

$$\bar{v} = (\bar{v}_1 ... \bar{v}_2) \Rightarrow \bar{v}_i \in R^p \quad je\ centroid\ pro\ i-t\acute{y}\ shluk$$

Toto kritérium je výhradně používáno pro „HARD“ rozklady, tj. u_{ik} není funkcí příslušnosti, ale charakteristickou funkcí nabývající hodnoty 1 ; 0.
J_w lze potom přepsat jako:

$$J_w(U, \bar{v}) = \sum_{i=1}^{c} \left(\sum_{x_k \in u_i} \|x_k - v_i\|^2 \right)$$

Takto zavedený J_w hledá optimální rozklad na základě Euklidovy vzdálenosti a jako takový se předem hodí na data sférického (kulového) charakteru. Hledání minima $J_w(U, \bar{v})$ je obecně složitá úloha. Jedním z nejvíce užívaných algoritmů je „iterační optimalizace“ ISODATA. Tento základní algoritmus ISODATA byl zaveden G. H. Ball a D. J. Hall[5] a posléze rozpracován R. Duda a P. Hart[6]. V roce 1974 vydal J. C. Dunn[7] práci, kde uvádí zobecnění ISODATA algoritmu a jeho „verzi fuzzy“ (viz. teoretická část). Spolu s vývojem fuzzy algoritmů shlukové analýzy se zároveň řešila i otázka počáteční volby počtu shluků. Většina algoritmů vyžaduje volený počet shluků. První kdo na to upozornil byl E. Ruspiny, který navrhl zavést veličinu analogickou informační entropii podle C. E. Shannon[8]. Následně autoři A. DeLuca a S. Termini[9] tuto myšlenku rozpracovali a zavedli definici entropie fuzzy množin (viz. teoretická část) na nepravděpodobnostním přístupu s využitím funkcí příslušnosti. Autoři vycházejí ze Shannonovy informační entropie kterou upravily ve smyslu funkce příslušnosti a zavedli pojem normalizované entropie fuzzy množiny jako měřítka její rozmytosti. Rozmytost autoři pokládají za nepřímo úměrnou kvalitě rozkladu, dospívají tak k důležitému závěru, že validní rozklad sebou nese minimalizaci normalizované entropie rozkladu.
Po entropii rozkladu zavedl roku 1974 J. C. Bezdek[10] koeficient rozkladu F_c. Bezdek navazuje na práci J. C. Dunn[7] a místo indexu separace coby obecného vyjádření jakosti rozkladu uvádí koeficient rozkladu, který je relativně dobře počitatelný a jeho maximalizace vede k dobrému rozkladu (viz. teoretická část).
Závažná otázka, která do té doby nebyla zodpovězena, je otázka konvergence funkcionálu fuzzy ISODATA. Hledání minima je iterativní postup, který obecně nemusí dosáhnout svého minima. J. C. Bezdek[11] ukazuje pomocí Zangwillova[12] konvergenčního teorému, že iterace fuzzy ISODATA, tj. obecně Picardova iterace, dosahuje vždy lokálního minima. Tato práce nebyla úplně přesná, a tak v roce 1987 W. T. Tucker[13] dokazuje na příkladech, že Picardova iterace může místo lokálního minima dosáhnout sedlového bodu. V zápětí vycházejí dvě společné práce[14,15], ve kterých je uvedeno za jakých podmínek iterace aproximuje sedlový bod.
V roce 1981 vychází kniha jednoho z nejplodnějších autorů praktických aplikací fuzzy množin J. C. Bezdeka[16], která přehlednou formou probírá problematikou fuzzy kritérií. Autor se v knize věnuje základním technikám clusterové analýzy, které jsou založeny na hledání extrému fuzzy účelové funkce a jejich modifikacím. V prvních částech se věnuje základních pojmům jako fuzzy relace, fuzzy algebra, zobecňuje pojem „HARD“ a „FUZZY“ rozklad množiny dat. Ze základních technik shlukování uvádí Ruspiniho funkcionál hustotu ve srovnání s fuzzy C-means. Rozvádí jejich vlastnosti, problematiku konvergence funkcionálů (viz. teoretická část) apod.

V dalších oddílech se věnuje tolik problematické otázce validity clusterů a s tím související volbě počtu clusterů. Bezdek v knize částečně vychází ze svých dříve vydaných prací a článků.
Ještě předtím, v roce 1980, vyšel článek J. C. Bezdek, R. Ehrlich, W. Full[17] podávající praktickou aplikaci Fuzzy c-means v geostatistické analýze geologických dat. Autoři v práci ukazují zdrojový kód algoritmu v programovacím jazyce fortran.

Algoritmus FCM je běžným standardem ve fuzzy clusterové analýze, avšak není univerzální. Univerzální algoritmus neexistuje. Již z podstaty je FCM vhodný na data sférického charakteru. Na datech se zjevnou tendencí linearity nedosahuje dobrých výsledků.
V roce 1993 vychází od R. Krishnapuram[18,19] práce, v níž autor na základě dříve uvedených prací o teorii „possibility“ (možnosti) zavádí clusterovou analýzu na základě možnosti „possibilistic clustering“ (PCM), jako postup řešící nedostatky FCM. PCM se ukazuje jako zajímavou alternativou FCM, která by se mohla hodit více na data obsahující odlehlé hodnoty.

Fuzzy clusterovou analýzou dat reprezentovaných fyzikálně-chemickými vlastnostmi povrchových vod se zabývají autoři P. Barbieri , G. Adamia, A. Favretto , A. Lutmanc, W. Avoscan , E. Reisenhofer [20]. V severovýchodní části Itálie byly odebrány vzorky ze studní se sladkou vodou pramenící v hloubce, za účelem zjištění znečištěných lokalit. Lokalit bylo 38. Vzorky byly odebírány z hloubek od 20 do 200 m třikrát v průběhu roku. Stanovovaných vlastností bylo 10, byly to: *vodivost, teplota*, *rozpuštěný kyslík, vápník, hořčík, chloridy, dusičnany, sírany, atrazin a desethylatrazin.* Všechny studně byly na území jižních rovin Friuli-Venezia Giulia Region. Odběry probíhaly 1996/1997. Atrazin je v Itálii od roku 1985 zakázán coby herbicid. On a jeho metabolity patří mezi výrazně toxické látky, přesto se v půdě stále objevuje v hojném množství. Vzorky byly odebírány na jaře 1996, na podzim 1996 a na jaře 1997. V každém vzorku bylo stanoveno všech 10 parametrů.
Jako fuzzy algoritmus byl použit algoritmus „FANNY[21]“. Pro určení reprezentativního odběru autoři použili algoritmus „PAM“[21]. Vzorkování bylo provedeno laboratoří A.R.P.A.-FVG, ústav provincie Udine. Na obrázku 3. je geografické rozložení odběrových míst. Zvláštní pozornost byla věnována atrazinu a desethylatrazinu. Dobrá selektivita byla zaručena na plynovém chromatografu s hmotnostní detekcí GC-MS, a extrakcí na pevné fázi SUPELCO, tímto bylo dosaženo detekčního limitu ($0.01\ \mu g.l^{-1}$).

obr 3.
Geografické rozložení odběrových míst

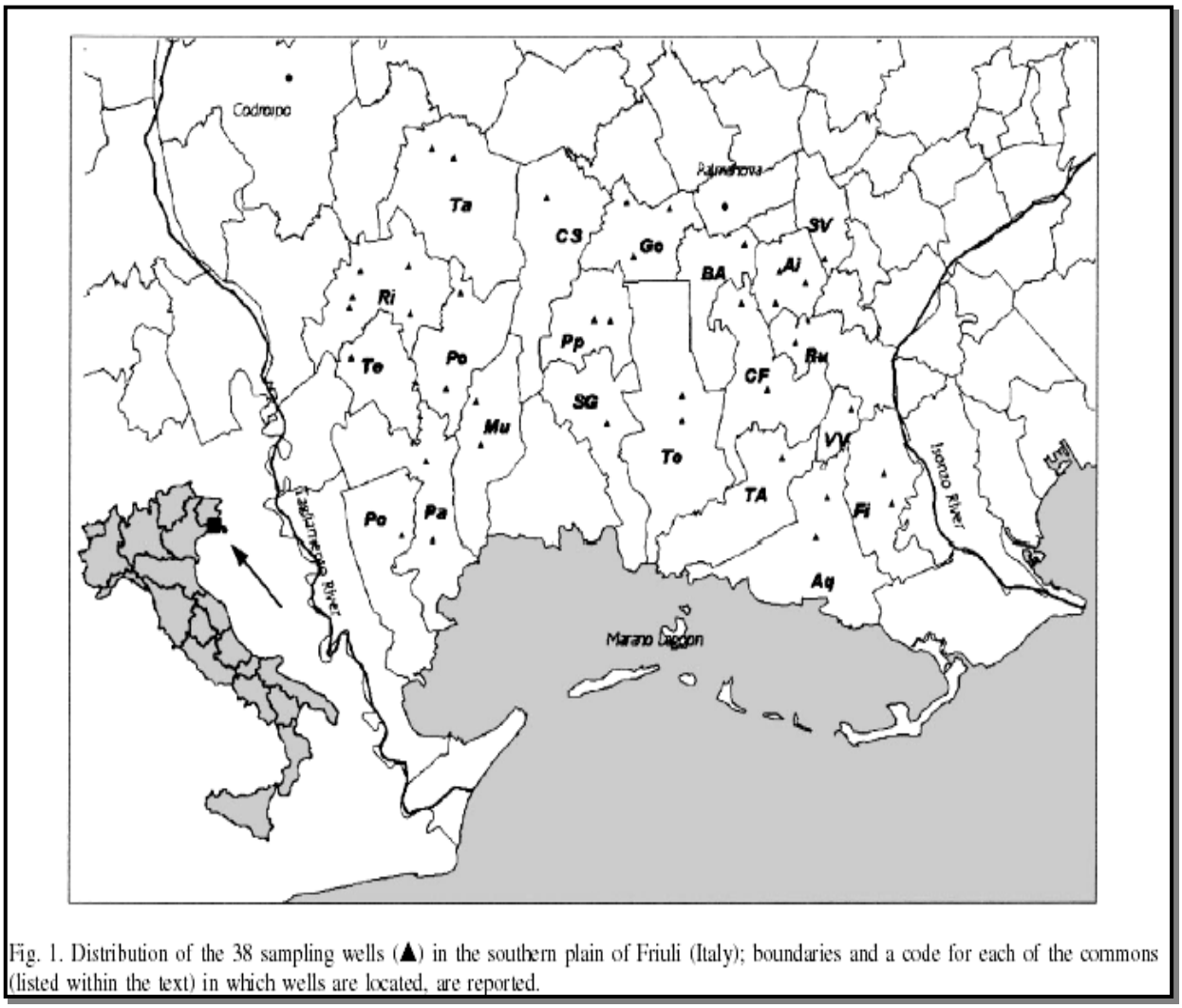


Fig. 1. Distribution of the 38 sampling wells (▲) in the southern plain of Friuli (Italy); boundaries and a code for each of the commons (listed within the text) in which wells are located, are reported.

V obou případech algoritmů PAM i FANNY je nutná apriorní volba počtu clusterů. Pro její správné určení se použila metoda „examination of silhouette indexes“. Pro každý objekt „i“ je počítána hodnota linie (silhouette value) s(i) a graficky vyhodnocena. a(i) značí průměrnou nepodobnost (ve formě Euklidovy vzdálenosti) i-tého objektu vůči ostatním objektům clusteru „a“. d(i,C) je průměrná nepodobnost objektu „i“ vůči ostatním clusterům různým od „a“. Nejmenší z nich je b(i)=d(i,B) a je to pro objekt „i“ druhý nejlepší cluster. Hodnota linie pro objekt „i“ je potom:
a(i)-podobnost
b(i)-nejbližšíi podobný cluster

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Vidíme, že s(i) leží vždy mezi –1 a +1. Hodnotu blízko k –1 indikují špatnou klasifikaci, zatímco hodnotu u +1 značí dobrou. Hodnota kolem 0 vyjadřuje

skutečnost, že objekt nelze jednoznačně klasifikovat do jednoho shluku. Obrazek 4. ukazuje „silhouette“ graf všech objektů pro 6 clusterů a PAM algoritmus.

obr 4.
Silhouette graf objektů

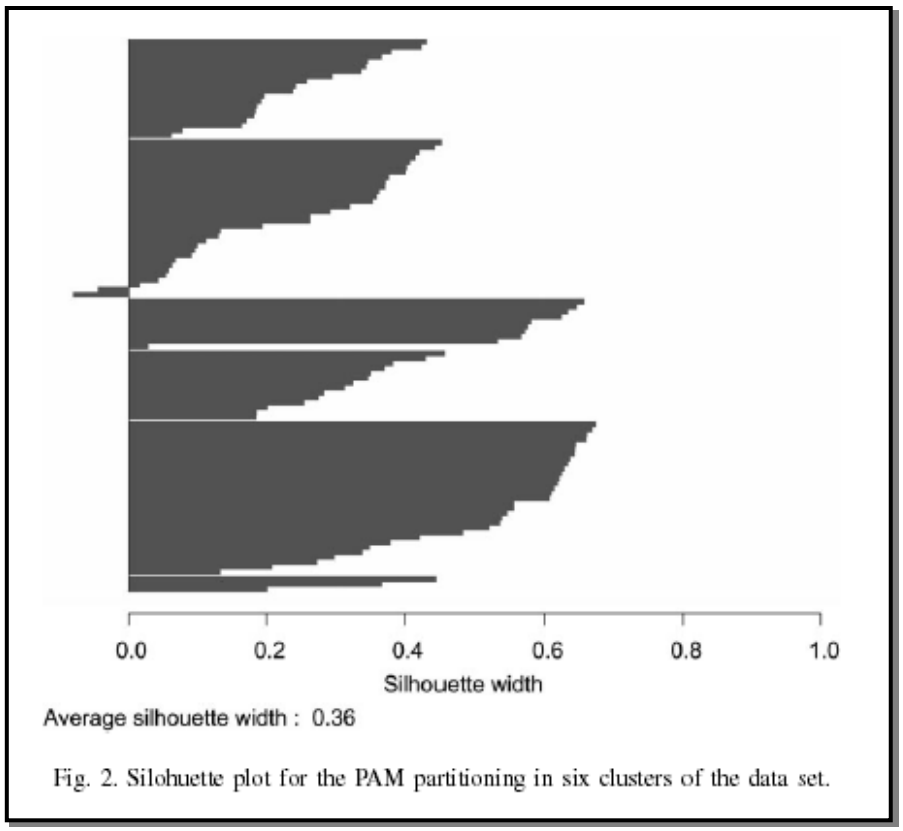


Fig. 2. Silohuette plot for the PAM partitioning in six clusters of the data set.

Následující tabulka obsahuje srovnání algorotmů PAM a FANNY pomocí celkové průměrné „silhouette“ šířky pro počet shluků 2 až 10.

Table 3
Overall average silhouette widths for clustering algorithms PAM and FANNY, number of clusters ranging 2–10

Number of clusters (k)	Overall average silhouette width (PAM)	Overall average silhouette width (FANNY)
2	0.30	0.30
3	0.26	0.11
4	0.31	0.16
5	0.35	0.21
6	0.36	0.21
7	0.27	0.23
8	0.30	0.26
9	0.32	0.26
10	0.31	0.26

Nejlepšího výsledku dosahuje PAM s 6 clustery, průměrná hodnota dosahuje 0,36.

V 89,5% případech (34 z 38) platí, že objekt patří pouze do 1. clusteru. Dá se říci, že vlastnosti vody z hlediska sledovaných parametrů nejsou významněji sezónně závislé. Podle počtu, kolik ze tří vzorků jednoho odběrového místa pro 3 roční období patří do každého clusteru, se dá určit funkce příslušnosti od 0 do 3. Ta je graficky zobrazena na obrázku 5. šesti různými symboly s rozdílnou velikostí podle funkce příslušnosti. Regiony šesti clusterů jsou vyhraničeny obrysovou čarou. Čtyři z objektů-odběrových studní patří více než jednomu shluku, a tak se v regionech překrývají. Jména 5 regionů odpovídají zeměpisné poloze krom šestého, vyznačující ho se výrazným obsahem chloridů, dusičnanů a ATZ, DATZ, tento je nazván „polluted“.

obr 5.
Regiony odběrových míst

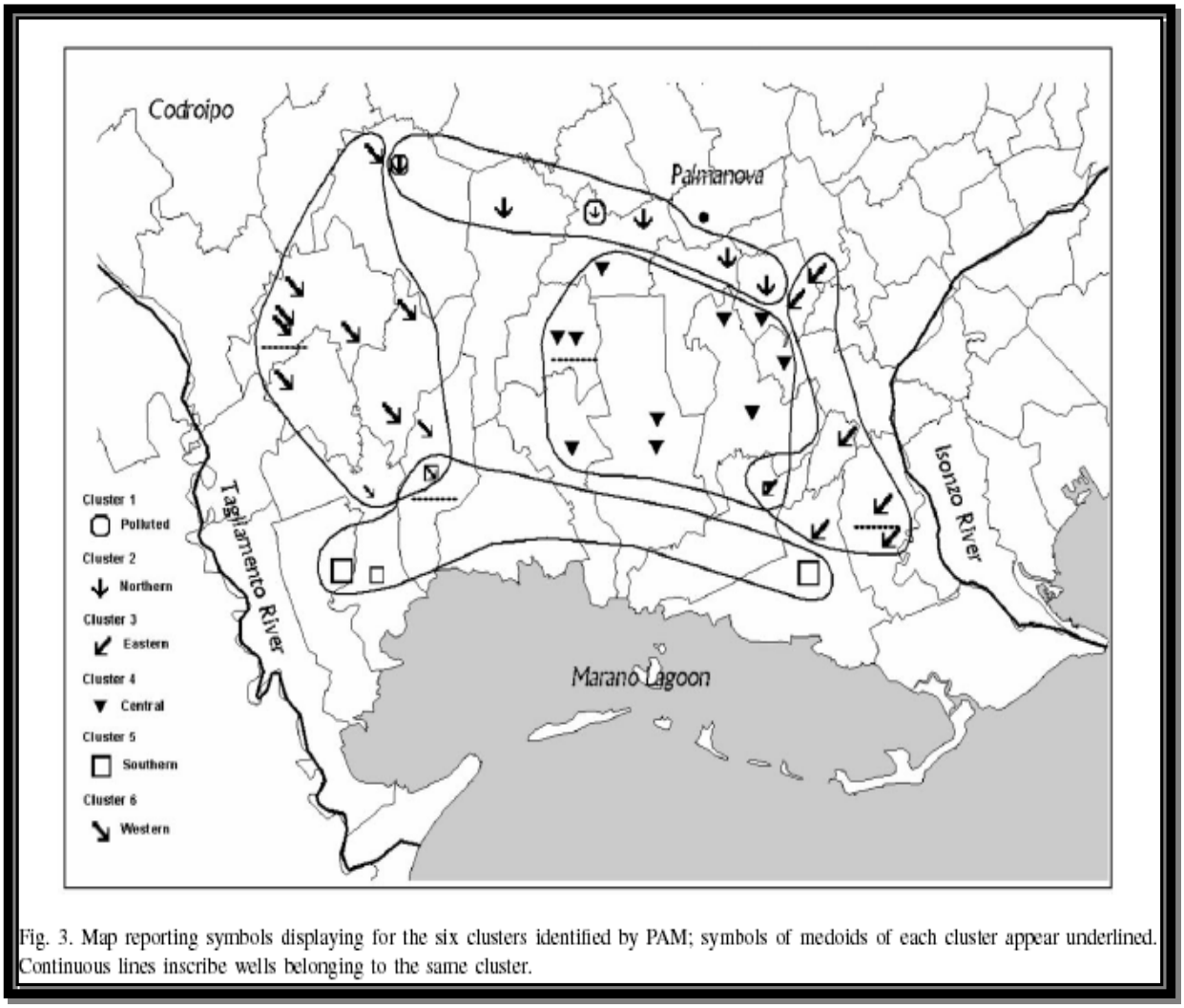


Fig. 3. Map reporting symbols displaying for the six clusters identified by PAM; symbols of medoids of each cluster appear underlined. Continuous lines inscribe wells belonging to the same cluster.

Jižní region je charakterizován velmi nízkým obsahem kyslíku pravděpodobně vlivem redukujícího prostředí rašelinových půd, které v minulosti byly bažinou. Západní region se vyznačuje velkou vodivostí a sírany, což je asi způsobe přítomností sádrovcových skal kudy protéká řeka před zásobováním podzemních pramenů. Dva severní vybočující regiony „northern“ a „polluted“ se odlišují vlivem povrchových vod s malou adsorbcí a výměnou iontů, čili podzemní vody obsahují velké množství iontů.

Další dva odlišující se regiony na jihu jsou opět pod vlivem povrchových vod a to protékajícími řekami Tagliamento a Isonzo River.

Při klasifikaci pomocí fuzzy clusterové analýzy lze použít několika algoritmů někdy se i výrazně lišících. Použitím FCM v řízení procesu se zabývají autoři Pekka Teppola, Satu-Pia Mujunen, Pentti Minkkinen[22]. Použití FCM se stává poněkud obtížnější v případě použití na objekt s mnoha proměnnými, mezi kterými existuje závislost. To je případ čističky odpadních vod a kalů. Autoři se zaměřují na spojení FCM a metody PCA na kontrolu procesu se sezónními trendy. V PCA jsou data projektována na osy hlavních komponent. Každá z nich nese informaci o majoritní části variability. Užitím PCA dojde k dimenzionální redukci a tak k zjednodušení bez ztráty důležitých informací. Ve vlastním algoritmu FCM autoři použili menší obměnu ve smyslu adaptability center shluků (ty se mění podle celkové efektivity algoritmu) a to zavedením parametru α.
Veškeré výpočty byly provedeny ve výpočetním systému matlab 5.1. Následující tabulka ukazuje rozdíly naměřené v létě a v zimě u vybraných veličin. Data byla získána z procesu čištění vod z „Enso Publication Papers Oy paper mill“ in Summa, Finsko. Celkem bylo sledováno 20 proměnných. Data představují naměřené denní hodnoty od 1.ledna 1995 do 28. února 1997, celkem bylo 787 objektů. PCA byla použita na normalizovaná data. Data byla rozdělena do 2 skupin, kalibrační sestávala z 260 objektů a zbývajících 527, měřených od 18. září 1995 do 28. února 1997, sloužilo jako ověřovací část. V PCA bylo použito 6 komponent, které vysvětlovaly postupně 30%, 18%, 12%, 7%, 6%, 5%, celkem 78% variability X. Skóre pro jednotlivé komponenty v kalibrační i v testovací množině jsou zobrazeny na obr 6.

obr 6.
Skóre komponent

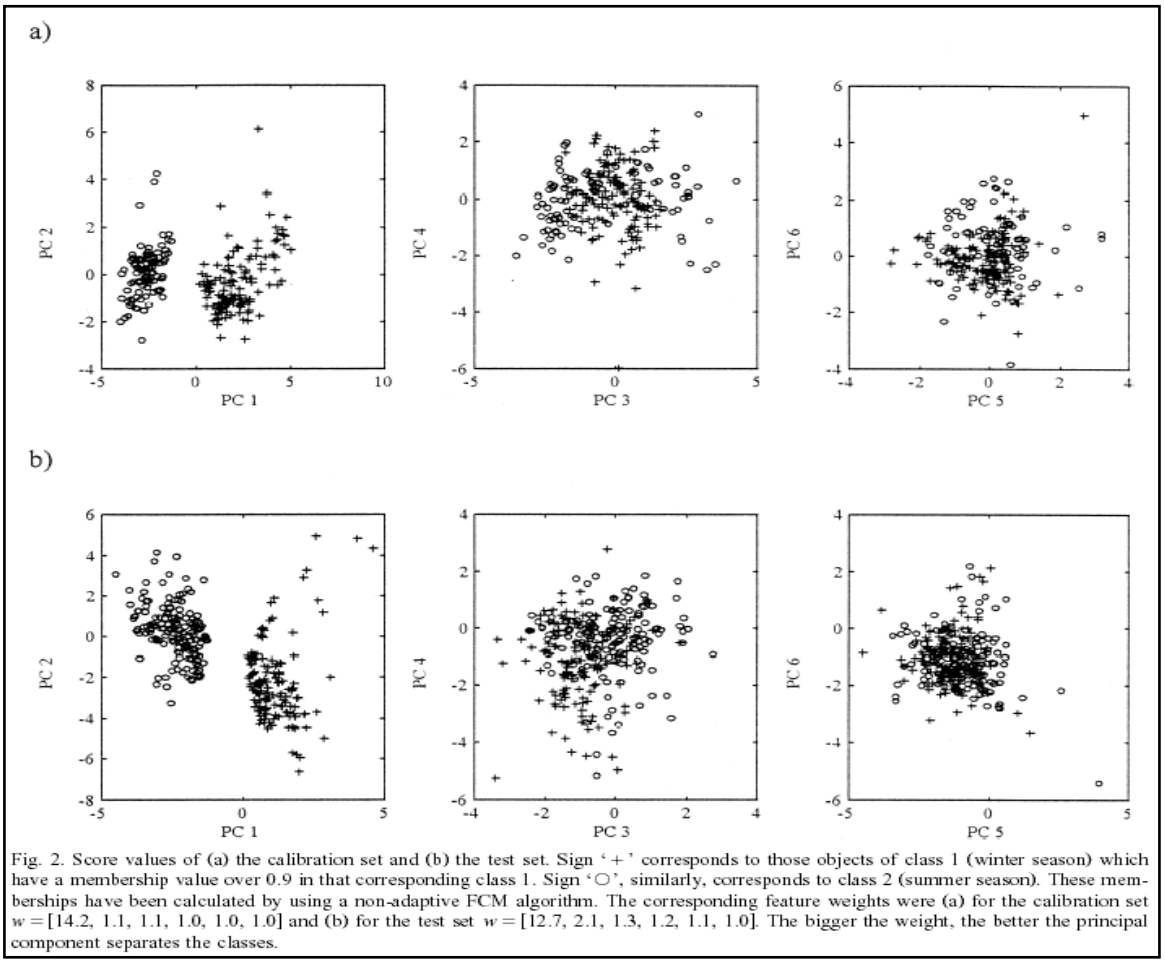


Fig. 2. Score values of (a) the calibration set and (b) the test set. Sign '+' corresponds to those objects of class 1 (winter season) which have a membership value over 0.9 in that corresponding class 1. Sign 'O', similarly, corresponds to class 2 (summer season). These memberships have been calculated by using a non-adaptive FCM algorithm. The corresponding feature weights were (a) for the calibration set $w = [14.2, 1.1, 1.1, 1.0, 1.0, 1.0]$ and (b) for the test set $w = [12.7, 2.1, 1.3, 1.2, 1.1, 1.0]$. The bigger the weight, the better the principal component separates the classes.

Na obrázku 7. autoři ukazují jak se mění poloha centroidu clusteru ve směru PC1, PC2 v závislosti na započítaném vzorku u statického FCM algoritmu s Euklidovou vzdáleností a adaptabilního FCM algoritmu s Mahalanobisovou vzdáleností.

obr 7.
Změna polohy centroidu

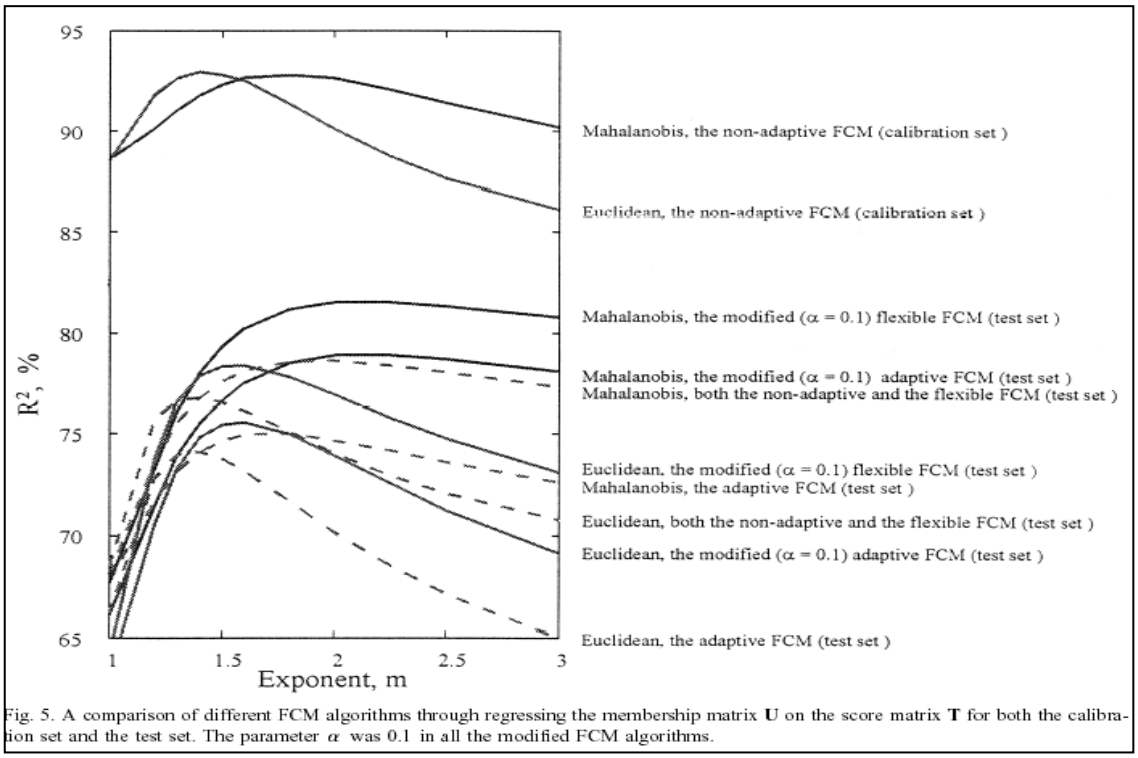


Fig. 5. A comparison of different FCM algorithms through regressing the membership matrix **U** on the score matrix **T** for both the calibration set and the test set. The parameter α was 0.1 in all the modified FCM algorithms.

Analogicky k uvedené práci ti samí autoři provedli fuzzy clusterovou analýzu na skóre z PLS. Obdobným přístupem při řízení čističky odpadních vod a to pomocí FCM a PLS se zabývají Pekka Teppola, Satu-Pia Mujunen, Pentti Minkkinen[23]. Autoři místo PCA jako v předešlém případě použili metodu PLS za účelem extrakce nejdůležitějších informací k sestrojení jedné proměnné nejvíce odpovědné za řízení procesu. Hodnoty skóre byly použity v FCM.

2. TEORETICKÁ ČÁST

2.1 OVZDUŠÍ

2.1.1 ZÁKLADNÍ POJMY – OVZDUŠÍ, ATMOSFÉRA

Ovzduším obecně rozumíme vzdušný obal zeměkoule neboli zemskou atmosféru. Přitom obvykle rozlišujeme tzv. volné nebo venkovní ovzduší, tj. ovzduší mimo uzavřeně nebo jinak obestavěné prostory (budovy, výrobní haly aj.), a vnitřní ovzduší pracovních, obytných a jiných prostorů.
Složení vzduchu, tj. obsah jednotlivých plynných složek v přirozeně atmosféře, není vzhledem k neustálým změnám v ovzduší stálé. Poměrně přesně lze stanovit jen obsahy hlavních složek vzduchu, tj. dusíku, kyslíku a vzácných plynů. Všechny ostatní složky se ve větší, či menší míře zúčastňují různých chemických reakcí a jsou součásti elementárních koloběhů, takže jejich obsah kolísá v závislosti na místu a ročním i denním období. Průměrné chemické složení vzduchu vybraných látek uvádí tab.1.

Tab. 1
Průměrné chemické složení vzduchu

Složky vzduchu	**Suchý vzduch**			**Vlhký vzduch**		
	obj (ppm)	mg/m^3	hmotn. (ppm)	obj (ppm)	g/m^3	hmotn. (ppm)
Oxid dusný (N_2O)	0,5	0,90	0,76	0,49	0,87	0,74
Oxid dusičitý (NO_2)	0,001	-	0,003	-	-	-
Oxid siřičitý (SO_2)	0,0002	-	0,0009	-	-	-

Vedle plynných složek obsahuje atmosférický vzduch rovněž kapalné a tuhé složky ve formě aerosolů, tvořených mikroskopickými částicemi hmoty, tuhými i kapalnými, rozptýlenými v plynném prostředí. Kapalné složky se nacházejí v ovzduší ve formě mlhy a mraků a tvoří je drobné kapičky vody, zkondenzované z vodní páry. Součástí ovzduší jsou i další tuhé a kapalné aerosoly, jejichž koncentrace se pohybuje od jednotek nanogramů až po desetiny miligramu. Tyto aerosoly vznikají nad moři i pevninami působením vulkanických erupcí, lesních požárů, působením větru, biologickými aktivitami (pyl, bakterie, spory) i fotochemickými reakcemi z původně plynných složek.

2.1.1.1 ZNEČIŠŤOVÁNÍ OVZDUŠÍ

V užším slova smyslu znečišťováním ovzduší se rozumí vypouštění hmotných látek v tuhém, kapalném nebo plynném skupenství ze zdrojů do ovzduší, které bud' přímo nebo po chemických změnách v atmosféře, nebo ve spolupůsobení s jinou látkou negativně ovlivňují kvalitu a složení venkovního ovzduší.
Vzhledem k tomu, že znečišťující látky jsou z atmosféry postupně odstraňovány, má pro hodnocení účinků znečišťování ovzduší značný význam rovněž doba setrvání jednotlivých znečišťujících látek v atmosféře. V podstatě existují tři základní principy, na jejichž základě jsou tyto látky z atmosféry odstraňovány:

suchá depozice, která představuje záchyt látek při styku se zemským povrchem

mokrá depozice, představující vymývání některých látek deštěm, nebo jejich odstraňování při tvorbě mraků

chemické reakce v troposféře, případně u reaktivnějších látek v nižších vrstvách atmosféry

Průměrné doby setrvání látek v atmosféře se u jednotlivých plynných složek značně odlišují. V tab. 2 jsou uvedeny doby setrvání jednotlivých plynných látek podle MOLDANA[24]. Pro tuhé částice je doba setrvání udávána v rozmezí 1 až 5 dnů pro vrstvu v blízkosti zemského povrchu, 5 až 10 dnů pro dolní část troposféry a přibližně 1 rok pro dolní část stratosféry.

Tab. 2
Doby setrvání jednotlivých látek

Prvek nebo sloučenina		**Průměrná doba setrvaní v atmosféře**
oxid dusný	N_2O	4 roky
oxid dusičitý	NO_2	11 dnů
oxid dusnatý	NO	9 dnů
oxid siřičitý	SO_2	4 dny
jemné tuhé částice v blízkosti zemského povrchu		1-5 dnů

Znečišťování ovzduší označuje určitou činnost či děj, tedy vnášení, či vypouštění *(emisi)* znečišťujících látek do atmosféry.

Znečištění ovzduší označuje naopak určitý stav, který je důsledkem původní činnosti či děje. Rozumí se tím tedy přítomnost neboli obsah *(imisi)* znečišťujících látek v ovzduší v takové koncentraci, při níž dochází k nepříznivému ovlivňování prostředí.

Pojmem emise se tedy rozumí vstup určité látky, příp. skupiny látek do atmosféry. Přítomnost znečišťujících látek v přizemní vrstvě atmosféry je označována jako imise.

Emise = znečišťující látky při vstupu ze zdroje do atmosféry

Imise = znečišťující látky v atmosféře v blízkosti příjemců

Mírou pro znečišťováni ovzduší jsou hmotnostní toky jednotlivých znečišťujících látek na vstupu do atmosféry, vyjádřené bud' v absolutních hodnotách, nebo vztažené na jednotku času, jednotku produkce apod. Mírou znečištění ovzduší jsou pak tzv. imisní koncentrace (vyjádřené obvykle v $\mu g.m^{-3}$ nebo ppb), čímž se rozumí koncentrace znečišťujících látek v ovzduší. Je zřejmé, že emisní koncentrace bývají o několik řádů vyšší než imisní a vyjadřují se zpravidla v $g.m^{-3}$ (nebo $mg.m^{-3}$) nebo v % objemových, případně i v ppm (1 ppm = partes per milion, tj. 1/1 000 000, tedy jedna miliónтina celku; 1 ppb = partes per bilion, tj. jedna miliardtina celku).

2.1.1.2 ČLENĚNÍ ZNEČIŠŤUJÍCÍCH LÁTEK

Při hodnocení znečišťování ovzduší je důležitým kritériem druh znečišťující látky. Znečišťující látky nejčastěji rozlišujeme podle skupenství, chemického složení a podle účinku či míry škodlivosti (nebezpečnosti, rizikovosti) z hlediska příjemců. Podle skupenství se látky znečišťující ovzduší člení na tuhé, kapalné a plynné.

V souvislosti s uvedeným dělením znečišťujících látek podle jejich účinku nutno vymezit i míru škodlivosti či nebezpečnosti jednotlivých látek z hlediska příjemců. Stanovení této míry škodlivosti vychází z hygienického hlediska působení těchto látek na zdraví lidí a vyjadřuje se obvykle hodnotou tzv. nejvyšších přípustných koncentrací (NPK) škodlivin ve venkovním ovzduší. Nejzávažnější škodliviny podle hygienických kritérií jsou uvedeny ve směrnici č. 58 hlavního hygienika ČSR z roku 1981, z níž uvádím v tab. 3 vybrané látky, seřazené podle míry jejich nebezpečnosti. Pořadí těchto látek a součinitel jejich nebezpečnosti byly stanoveny (po zaokrouhlení) bud' pro 30 minutový **(Kmax)** nebo 24 hodinový průměr **(Kd).**

Tab. 3
Nejvyšší přípustné koncentrace

Látky znečišťující ovzduší	**Nejvyšší přípustné koncentrace v $\mu g.m^{-3}$ (0°C; 0,1 Mpa)**	
	K_{max} (30 minut)	**K_d (24 hodin)**
oxid siřičitý .	500	150
nesedimentující (polétavý) prach neobsahující toxické složky biologicky aktivní	500	150
oxidy dusíku (vyjádřeny jako NO_2)	100	100

2.1.2 PRACH

Pojmem prach se označují malé částice tuhých látek, které po rozptýlení v klidném disperzním systému mají pádovou rychlost, která odpovídá zákonům volného pádu.

Dělí se obvykle do tří velikostních skupin: *hrubý prach* (částice větší než 40 μm), *střední prach* (částice velikosti 1 až 40 μm) a *jemný prach* (částice menší než 1μm). Velikost jednotlivých částic se pohybuje ve velmi širokém rozmezí, a to od velikosti řádu 10^{-4} μm až po 10^{3} μm. Tento druh znečištění je vytvářen částicemi tuhých ve značně širokém velikostním spektru od makromolekul až po viditelná zrna. Složeni těchto částic je velmi rozmanité a zahrnuje celou škálu anorganických i organických látek; může přitom jít jak o částice neživé (prach přirozeného původu ze zemského povrchu nebo prach vznikající při mletí surovin a výrobků, mořské soli, částice popela z lesních požárů, vulkanické částice, prachové částice z průmyslových spalovacích procesů, produkty reakcí v ovzduší), tak i o živé částice (pyl rostlin a stromů, viry, bakterie, řasy, prvoky, hmyz, části hmyzích těl apod.).

ÚČINEK PRACHOVÝCH ČÁSTIC NA LIDSKÉ ZDRAVÍ

Vliv prachových a aerosolových částic v ovzduší na lidské zdraví závisí především na velikosti částic, částice větší než 100 μm pro svou značnou hmotnost poměrně rychle sedimentují a mají proto relativně malý přímý zdravotní význam. Daleko závažnější je účinek však jemnějších částic. Tyto mohou být tvořeny nejrůznějšími látkami, a to jak anorganickými prachy (kovovými částicemi, křemičitany, fluoridy, oxidy, dusičnany, chloridy, sírany aj.), tak i prachy organického původu (např. dehty, bakterie, pyly).
Z hlediska ukládání částic v plicích jsou nejnebezpečnější částice střední velikosti, protože jsou až z 90 % zachycovány v plicích. Tato nejnebezpečnější frakce se označuje jako PM_{10}.

2.1.3 PLYNNÉ ZNEČIŠŤUJÍCÍ LÁTKY

Plyny a páry znečišťujících látek se do ovzduší dostávají jednak ze zdrojů přírodních, jednak jako výsledek lidské činnosti, a to především se zplodinami spalování paliv či z průmyslových technologií. Mezi hlavní plynné znečištěniny, s nimiž se ve volném ovzduší setkáváme nejčastěji, patří některé plynné sloučeniny síry a dusíku.

2.1.3.1 SLOUČENINY SÍRY

Oxid siřičitý SO_2 - je vedle aerosolových částic nejrozšířenější látkou znečišťující volné ovzduší. Jediným jeho hlavním přírodním zdrojem je vulkanická činnost. Jeho hlavními zdroji jsou spalné procesy, ve kterých se spaluje palivo s vysokým obsahem síry, tj. především elektrárny, teplárny, domácí topeniště a některé technologické procesy (cca 80 % všech emisí SO_2).
Oxid siřičitý nezůstává v atmosféře beze změn, ale jeho koncentrace rychle klesá, zejména následkem oxidace na oxid sírový, který za přítomnosti vodní páry okamžitě hydratuje za vzniku kyseliny sírové. Zjednodušeně lze tyto reakce popsat následujícími rovnicemi:

$SO_2 + ½\ O_2 + h.v \longrightarrow SO_3$
$SO_3 + H_2O \longrightarrow H_2SO_4$

Přímá oxidace SO_2 na SO_3 probíhá nejčastěji cestou fotooxidace SO_2 v plynné fázi na povrchu tuhých částic. Výsledná rychlost oxidace proto závisí na mnoha aspektech jako povětrnostní podmínky, teplota, sluneční svitu, přítomnost katalyzujících

pevných částic apod. Běžně se oxidací odstraní během 1 hodiny z ovzduší 0,1% až 2% přítomného SO_2. Vzniklý oxid sírový je okamžitě hydratován vzdušnou vlhkostí na aerosol kyseliny sírové, který muže reagovat s prachovými alkalickými částicemi v ovzduši za vzniku síranů. Srážkové vody tak mohou být okyseleny až na pH = 4,0. Kyselé deště uvolňuji z půdy hliníkové a další kovové ionty (Cu, Pb Cd), které dále poškozuji půdní mikroorganizmy, znehodnocují vodu atd.
Oxid siřičitý je vzhledem ke svým redukčním a kyselým vlastnostem dráždivý plyn, který negativně působí na zdraví živočichů, zejména na zdraví lidské a poškozuje především dýchací systém a oční spojivky; akutně vede ke kontrakci hladkých svalů dýchacích cest, zejména u astmatiků, ve vyšších koncentracích působí zánětlivé změny plicního epitelu a poškozuje řasinkové buňky dýchacích cest.
Dlouhodobé působení SO_2 při koncentracích nad 50 $\mu g.m^{-3}$ vede ke zvýšení úmrtnosti na choroby krevního oběhu a chronickou bronchitidu. Chronickými účinky zasahuje intenzívně do metabolizmu a imunitních reakcí organizmů.
Oxid siřičitý už ve velmi malých koncentracích negativně působí rovněž na rostlinstvo. Jedny z nejcitlivějších jsou lišejníky, které rychle hynou. U vyšších rostlin poškozuje jejich fotosyntetický aparát, což vede k odumírání (nejvíce jsou napadeny rostliny s neopadavými listy, tedy jehličnany).

2.1.3.2 SLOUČENINY DUSÍKU

Oxidy dusíku - paleta oxidů dusíku v atmosféře je velmi pestrá; zahrnuje celkem 5 různých oxidů, v nichž dusík vystupuje jako jedno až pětimocný.
(N_7O - oxid dusný, NO **-** oxid dusnatý, N_2O_3 - oxid dusitý, NO_2 - oxid dusičitý a N_2O_5 **-** oxid dusičný).

Nejrozšířenějším oxidem dusíku v atmosféře je bezbarvý oxid dusný N_2O, jehož koncentrace v troposféře dosahuje až 450 $\mu g.m^{-3}$ (zahrnuje 97% hmotnosti všech sloučenin dusíku). Jeho hlavním zdrojem je bakteriální rozklad dusíkatých látek v půdě a v povrchových vrstvách oceánů. Vyznačuje se velmi malou reaktivitou a nemá proto žádný vliv na životní prostředí a není tak považován za škodlivinu. K jeho rozkladu dochází až ve stratosféře, kde se fotochemicky rozkládá na dusík a kyslík.
Z vyšších oxidů dusíku je prakticky bezvýznamný oxid dusitý N_2O_3, který se bezprostředně přeměňuje na ($NO + NO_2$), stejně tak i N_2O_5, jenž vzniká oxidací NO_2 ozonem a rychle reaguje s vodní párou za vzniku kyseliny dusičné.
Z hlediska škodlivého vlivu na životní prostředí je nejvýznamnější výskyt oxidu dusnatého NO a rezavě zbarveného oxidu dusičitého NO_2 v troposféře. Vzhledem k tomu, že většina analytických metod udává sumu těchto oxidů, obvykle tyto dva oxidy shrnujeme pod společný název „suma oxidů dusíku" a označujeme jako NO_x. Při stechiometrických výpočtech uvažujeme NO_x jako NO_2.
Přírodními zdroji NO_x jsou zejména vulkanická činnost, bakterie a elektrické výboje v atmosféře (způsobují přirozené koncentrační pozadí NO: 0 až 7,4 $\mu g.m^{-3}$ a NO_2: 0,4 až 9,4 $\mu g.m^{-3}$). Značná část NO_x pochází ze spalovacích procesů probíhajících při výrobě energie i v dopravě a z chemických výrob (výroba kyseliny dusičné, nitrace organických látek apod). Produkce NO_x je tak soustředěna do průmyslových center a velkých městských aglomerací. Při vysokých teplotách, za nichž probíhají reakce spalování fosilních paliv, vzniká především NO, který je ve směsi NO_x zastoupen z 90 až 95 objemových procent. V ovzduší jsou oxidy dusíku účastny celé řady reakcí. Emitovaný NO je v atmosféře samovolně oxidován na NO_2 s následnou tvorbou kyseliny dusičné podle reakčního schématu:

$2\ NO + O_2 \longrightarrow 2\ NO_2$
$2\ NO_2 + H_2O \longrightarrow HNO_3 + HNO_2$

Kyselina dusičná se po své neutralizaci prachovými alkalickými částicemi, jako jsou CaO, MgO stává ve formě svých solí součástí aerosolových částic, s nimiž je z atmosféry odstraňována prostřednictvím srážek. Množství dusíku, které se tak dostane do půdy a vod mění ekosystém v obou sférách a zejména pak ve vodě dochází k nežádoucímu rozmnoženi některých druhů vodních rostlin, které zvyšují biologickou spotřebu kyslíku a při vyšších koncentracích vedou k úhynu ryb.
Ze zdravotního hlediska působí oxidy dusíku NO_x nepříznivě zejména na dýchací orgány, kde mohou vést k jejich onemocnění, případně i ke vzniku onemocnění nádorových. Při vyšších koncentracích se NO_x váže na hemoglobin a zhoršuje přenos kyslíku z plic do krevního oběhu. Pro eventuální možné akutní poškození se uvádí limitní koncentrace 190 $\mu g.m^{-3}$, která při trvání jedné hodiny může změnit dýchací funkce zdravého člověka. Denní imisní limit pro NO_x je 100 $\mu g.m^{-3}$, krátkodobý třicetiminutový limit je dvojnásobný.

2.1.4 LIMITY ZNEČIŠŤOVÁNÍ OVZDUŠÍ

Přípustnou úroveň znečišťování ovzduší určují dle stávající legislativy (zákon 309/1991 Sb., o ovzduší v platném znění) emisní, imisní a depoziční limity pro jednotlivé znečišťující látky.

Emisní limit je nejvýše přípustné množství znečišťující látky vypouštěné ze zdroje znečišťování do ovzduší, vyjádřené jako (použije se obvykle jedna z uvedených možností):

Tyto emisní limity musí dosahovat hodnoty odpovídající nejlepším dosažitelným prostředkům (technologiím) a jsou pro vybrané znečišťující látky u vybraných zdrojů znečišťování (technologií a zařízení), stejně jako tzv. „obecné emisní limity“ (platí pro zdroje, jež nejsou zahrnuty mezi vybrané zdroje znečišťování), uvedeny ve vyhlášce MŽP ČR. 117/1997 Sb. ve zněni vyhlášky Č. 97/2000 Sb.

Imisní limit je nejvýše přípustná hmotnostní koncentrace znečišťující látky obsažená v ovzduší. Hodnota imisních limitů pro jednotlivé znečišťující látky obvykle odpovídá stanoveným hodnotám nejvyšších přípustných koncentrací (NPK), vyjadřujících škodlivost znečišťujících látek z hygienických hledisek z pohledu lidského zdraví. V legislativě České republiky jsou hodnoty imisních limitů pro vybrané znečišťující látky dány opatřením dřívějšího Federálního výboru pro životní prostředí (Opatření FVZP), pro ostatní škodliviny pak ve Směrnici MZd Č. 58/81/20/. Hodnoty imisních limitů pro vybrané znečišťující látky dle přílohy Opatřeni FVŽP jsou uvedeny v tabulce 4.

Tab. 4
Imisní limity pro znečišťující látky platné pro území České republiky

Polutant	Vyjádření jako	Imisní limity ($\mu g.m^{-3}$)			
		IH_r	IH_d	IH_k	Obecný požadavek
Prašný aerosol	SPM	60	150	500	Koncentrace IH_d a IH_k nesmí být v průběhu roku řekročeny ve více než 5% případů
Oxid siřičitý	SO_2	60	150	500	
Oxidy dusíku	NO_x	80	100	200	

IH_r - průměrná roční koncentrace znečišťující látky. Průměrnou koncentraci se rozumí střední hodnota koncentrace, zjištěná na stanoveném místě v časovém úseku jednoho roku jako aritmetický průměr z průměrných 24-hodinových koncentrací.

IH_d - průměrná denní koncentrace znečišťující látky. Průměrnou denní koncentrací se rozumí střední hodnota koncentrace, zjištěná na stanoveném místě v časovém úseku 24 hodin. Průměrnou denní koncentrací se rozumí též střední hodnota nejméně dvanácti rovnoměrně rozložených měření
průměrných půlhodinových koncentrací v časovém úseku 24 hodin (aritmetický průměr).

IH_k - průměrná půlhodinová koncentrace znečišťující látky. Průměrnou půlhodinovou koncentrací se rozumí střední hodnota koncentrace, zjištěná na stanoveném místě v časovém úseku 30 minut.

V tabulce 5. jsou uvedeny připravované imisní limity podle EU.

Tab. 5
Připravované imisní limity

Znečišťující příměs	Časový interval	Limitní hodnota ($\mu g.m-3$)	Mez tolerance	Max. tolerovaný počet překročení za kalendářní rok
SO_2	kalendářní rok	50	-	0
	24 hod	125	-	3
	1 hod	350	150	24
NO_2	kalendářní rok	40	20	0
	1 hod	200	100	18
PM_{10}	kalendářní rok	40	8	0
	24 hod	50	25	35

Depoziční limit je nejvýše přípustné množství znečišťující látky usazené po dopadu na jednotku plochy zemského povrchu za jednotku času. Depoziční limity nebyly v ČR dosud žádnou vyhláškou ani opatřením stanoveny, v hygienických předpisech je uváděna hodnota depozičního limitu pouze pro spad prachu (tuhých znečišťujících látek). Za hygienicky únosný spad prachu se dle hygienických předpisů považuje hodnota 150 $g.m^{-2}.r^{-1}$.

2.1 CLUSTEROVÁ ANALÝZA

2.2.1 ZÁKLADNÍ POJMY

„Clusterová analýza je obecný logický postup formulovaný jako procedura, pomocí níž seskupujeme objektivně jedince do skupin na základě jejich podobností a rozdílností“

Tryon (1939)

Posuzováním vzájemných podobností věcí a jevů se zabývá oblast aplikované matematiky, clusterová analýza.
Klasifikací je nazývána činnost vytvářející rozklad nějaké množiny objektů, tj. činnost vedoucí k vytvoření systému tříd. Klasifikace se nechápe jako identifikace ve smyslu „rozeznání ke které třídě objekt (vlastnost) patří“, klasifikací jsou třídy vytvářeny. Pokud takto vzniklé třídy dále rozkládáme podle našich požadavků, jedná se o „hierarchickou klasifikaci“, opakem je nehierarchická klasifikace.
Hierarchické metody lze rozdělit na dvě velké skupiny: divizní a aglomerativní metody. Divizní přístup ke shlukování spočívá v postupném rozdělovaní množiny objektů jako celku a naopak aglomerativní přístup seskupuje jednotlivé objekty až ke konečnému stavu, tj. spojení všech objektů do jedné množiny.
Pro nehierarchické metody jsou nejdůležitější optimalizační metody hledající takový rozklad množiny objektů určených ke klasifikaci, který je optimální podle vhodně zvoleného kritéria (nejčastěji funkcionál) optimality rozkladu, u nehierarchických metod napřed zvolíme počet clusterů „c“ a hledáme takový rozklad množiny na clustery, nad kterými předem zvolený funkcionál kvality rozkladu nabývá extrémních hodnot (minim).
Objektem se myslí nějaká množina předmětů, jevů nebo vlastností, které jsou popsány vektorem čísel, tedy p-tící stavů předem určených p-znaků. Např. objektem je zeměpisná poloha, na které se měří 5 chemických veličin v ovzduší. Každé takové místo je popsáno vektorem 5 čísel, kde čísla představují naměřenou hodnotu koncentraci, rozpětí, odchylky apod. Objektem určeným pro shlukovou analýzu je p-tice hodnot vybraných znaků. Důležitým pojmem je matice dat. Matice dat jsou uspořádané výsledky všech měření, je to vstup pro clusterovou analýzu, její rozměr je obvykle (počet objektů *X* počet sledovaných znaků) tedy objekty tvoří řádky a znaky sloupce.

Standardizace
Někdy se určité znaky znaky jeví jako hlavní (dominující) a jiné jako málo důležité. V těchto případech se data upravují tak, aby byly souměřitelné, provede se standardizace dat. Standardizaci se provádí přes k-tý znak, kde k=1…p. Vypočte se střední hodnota a směrodatnou odchylku pro jednotlivé sloupce (znaky) matice dat.

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \qquad s_k = \left[\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2\right]^{\frac{1}{2}}$$

- x_{ik} je členem matice $X=(x_{ik})$ typu (n *X* p)
Standardizovaná hodnota potom je:

$$z_i = \frac{x_i - \bar{x}_k}{s_k}$$

Po standardizaci mají hodnoty znaků střední hodnotu $\bar{x}_k$ rovnu nule a směrodatnou odchylku s_k rovnu jedné.

Normování
Objektem je vektor, je to nejčastěji řádek matice dat. V některých datových souborech je výsledek clusterové analýzy negativně ovlivněn nestejnými normami vektorů. Normováním rozumíme převedení vektorů na vektory jednotkové délky. Norma je nejvhodnější jednotková. Normování se provede vydělením všech složek každého vektoru normou tohoto vektoru.

Normalizování
Normalizování je úprava vektoru, že součet jeho prvků je roven jedné,tj. vydělení prvků vektoru součtem všech jeho prvků.

Vyjádření podobnosti
Mnohé clusterové algoritmy charakterizují podobnost objektů na základě geometrického modelu matice dat, tj. na základě vzdálenosti. Objekt je charakterizován p-znaky které můžeme chápat jako p-rozměrný euklidovský prostor E_p. Euklidovskou metriku definovanou v tomto prostoru využijeme jako míru podobnosti objektů. Objekty jsou tím podobnější, čím je jejich vzdálenost menší.
Euklidovská metrika dvou bodů $Y_a=(a_1 \ldots a_p)$ a $Y_b=(b_1 \ldots b_p)$ je:

$$\|Y_a - Y_b\| = \left(\sum_{i=1}^{p} (a_i - b_i)^2 \right)^{\frac{1}{2}}$$

2.2.2 PROSTOR FUZZY ROZKLADU

2.2.2.1 HARD ROZKLAD

Množina $X=\{x_1 \ldots x_n\}$ je konečná množina. V_{cN} je množina reálných (c *x* N) matic, „c“ je celé číslo, kde $2 \leq c < N$. Prostor hard c-rozkladů množiny „X“ je množina „M_c“, kde:

$$M_c = \left\{ U \in V_{cN} \,\middle|\, u_{ik} \in \{0,1\}\ \forall i,k\ ;\ \sum_{i=1}^{c} u_{ik} = 1\ \forall k\ ;\ 0 < \sum_{k=1}^{N} u_{ik} < N\ \forall i \right\}$$

$u_{ik} = u_i(x_k)$ je jednička nebo nula, podle toho zda „x_k“ je nebo není prvkem i-té podmnožiny „X“.

Výraz $\sum_{i=1}^{c} u_{ik} = 1 \quad \forall k$ říká, že každý „x_k“ patří právě do jednoho clusteru.

Výraz $0 < \sum_{k=1}^{N} u_{ik} < N \ \forall i$ říká, že žádná podmnožina „X“ není prázdná, a že žádná podmnožina není rovna vlastní množině „X“, neboli $2 \leq c < N$.

2.2.2.2 FUZZY ROZKLAD

„X“ je konečná množina. V_{cN} je množina reálných (c *x* N) matic, „c“ je celé číslo, kde $2 \leq c < N$. Prostor fuzzy c-rozkladů množiny „X“ je množina „M_{fc}“, kde:

$$M_{fc} = \left\{ U \in V_{cN} \,\middle|\, u_{ik} \in [0,1] \ \forall i,k \ ; \sum_{i=1}^{c} u_{ik} = 1 \ \forall k \ ; 0 < \sum_{k=1}^{N} u_{ik} < N \ \forall i \right\}$$

Z uvedeného vyplývá, že „M_c“ je podmnožinou „M_{fc}“.

Degenerovaný c-rozklad je zobecnění množin „M_c“ a „M_{fc}“, kdy platí:

$$0 \leq \sum_{k=1}^{N} u_{ik} \leq N \ \forall i$$

Takový rozklad se značí „M_{c0}“ a „M_{fc0}“.

2.2.2.3 VLASTNOSTI PROSTORU FUZZY ROZKLADŮ

Řekli jsme si, že „M_c“ je konečná množina. Dá se ukázat, že jeho velikost (tj. počet všech možných hard-rozkladů) je:

$$|M_c| = \frac{1}{c!}\left[\sum_{j=1}^{c} \binom{c}{j} (-1)^{c-j} \, j^N \right]$$

Např. je-li c = 10 a N = 25, dostáváme 10^{18} hard 10-rozkladů.

Důležitou vlastností prostoru rozkladu, která se často vyskytuje ve vztazích clusterové analýzy je konvexní obal. Hlubší rozbor této problematiky přesahuje rámec této diplomové práce. Konvexní obal nekonvexní množiny „A“ se dá představit jako nejmenší možná konvexní množina „B“, která obsahuje (obaluje) podmnožinou „A“. Graficky si lze množinu „B“ představit tak, že každý její prvek lze spojit přímkou s jakýmkoli dalším a vzniklá úsečka bude celá ležet v množině „B“. Z matematického hlediska množina „S“ uvnitř vektorového prostoru „V“ je konvexní, pokud část přímky spojující dva body z „S“ je také jejím prvkem.

$$S \ je\ konvexní \Leftrightarrow \bar{x}, \bar{y} \in S \Rightarrow \alpha\bar{x} + (1-\alpha)\bar{y} \in S \quad \forall \alpha \in [0,1]$$

Je-li „V“ prostor vektorů a „S“ $\subset$ „V“. (C_i) je množina všech konvexních množin prvkem „V“ takových, že „S“ $\subset$ „C_i“ $\subset$ „V“. Konvexní obal množiny „S“ je potom průnikem „C_i“:

$$conv(S) = \bigcap_i C_i$$

Je-li „S“ konvexní, pak „S“ = conv(S).
Konvexní obal množiny S se zkonstruuje jako:

$$conv(S) = \left\{ \bar{v} \in V \middle| \bar{v} = \sum_{k=1}^{mv} \alpha_k \bar{s}_k ; \sum_{k=1}^{mv} \alpha_k = 1; \alpha_k \geq 0; \bar{s}_k \in S \right\}$$

2.2.3 ZÁKLADNÍ OPERACE V PROSTORU ROZKLADU

R je množina reálných čísel, R^p je lineární prostor p-rozměrných reálných vektorů.
Na tomto prostoru je definováno:

skalární součin vektorů „x“ a „y“

$$(2.2.3.1) \quad \langle x, y \rangle = xy^t = \sum_{i=1}^{P} x_i y_i$$

čtverec délky vektoru „x“, tj. čtverec normy „x“

$$(2.2.3.2) \quad \|x\|^2 = xx^t = \sum_{i=1}^{P} (x_i)^2 = \langle x, x \rangle$$

vzdálenost objektů „x,y“: tj. Euklidova metrika

$$(2.2.3.3) \quad d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

uvedené vztahy platí pro všechny řádkové vektory
$x = (x_1, x_2 \ldots x_p)$, $y = (y_1, y_2 \ldots y_p) \in R^p$.
Předpokládejme množinu $X = \{x_1, x_2 \ldots x_n\} \in R^p$, tedy „X“ obsahuje N-vektorů o dimenzi „p“. Libovolný „hard rozklad“ je „c“ podmnožin (Y_i) množiny „X“, přičemž jednotlivé podmnožiny splňují: $\{Y_i: 1 \leq i \leq c\}$:

$$(2.2.3.4a) \quad X = \bigcup_{i=1}^{c} Y_i$$

$$(2.2.3.4b) \quad 0 = Y_i \cap y_i \qquad \forall i \neq j$$

$$(2.2.3.4c) \quad 0 \neq Y_i \qquad \forall 1 \leq i \leq c$$

Pozn. („0“-značí prázdnou množinu)

se nazývá nedegenerovaný hard c-rozklad množiny „X“ (nedegenerovaný proto, že platí 2.2.3.4c). Označme P_c jako množinu všech rozkladů „c“ podmnožin „X“. Charakteristickou funkcí podmnožiny Y_i je zobrazení w_i: $X \rightarrow \{0,1\}$. Funkce nabývá hodnot 1 pro objekt „x_k“ prvkem podmnožiny Y_i, ($x_k \in Y_i$), jinak 0. Hodnotu charakteristické funkce pro prvek „x_k“ značíme „w_{ik}“. Pro každý prvek $P = \{Y_1 \ldots Y_c\} \in P_c$ existuje množina $Q = \{w_1 \ldots w_c\}$ charakteristických funkcí taková, že:

$$(2.2.3.5a) \qquad 1 = \bigvee_{i=1}^{c} w_i$$

$$(2.2.3.5b) \qquad 0 = w_i \wedge w_j \qquad \forall i \neq j$$

$$(2.2.3.5c) \qquad 0 \neq w_i \qquad \forall 1 \leq i \leq c$$

$w_i \vee w_j$ je charakteristickou funkcí sjednocení $Y_i \bigcup Y_j$ což je maximum z {w_{ik}, w_{jk}} pro každý vektor „x_k“ (neboli sjednocení charakteristických funkcí)

$w_i \wedge w_j$ je charakteristickou funkcí průniku $Y_i \bigcap Y_j$ což je minimum z {w_{ik}, w_{jk}} pro každý vektor „x_k“ (neboli prázdná množina)

Výše uvedené vztahy (2.2.3.4a,b,c) a (2.2.3.5a,b,c) jsou plně srovnatelné (izomorfní), tj. shluk je definován výčtem objektů nebo hodnotou charakteristické (popř. příslušnosti) funkce. Třídu všech c-tic charakteristických funkcí označíme Q_c. Jak P_c tak i Q_c mohou být považovány za prostor nedegenerovaných hard c-rozkladů. Vztahy (2.2.3.5a,b,c) mohou být uvedeny v jednodušší formě:

$$(2.2.3.6a) \qquad \sum_{i=1}^{c} w_{ik} = 1 \qquad \forall 1 \leq k \leq N$$

$$(2.2.3.6b) \qquad 0 \neq w_i \qquad \forall 1 \leq i \leq c$$

Dle Zadeh[1] lze provést fuzzyfikaci třídy Q_c. Označme {u_i: $1 \leq i \leq c$ } fuzzy podmnožiny „X“ („u_i“ dle konvence značí jak fuzzy množinu, v našem případě i-tý shluk, tak zároveň funkci příslušnosti). Analogicky k charakteristické funkci definujeme funkci příslušnosti jako zobrazení u_i: $X \rightarrow [0,1]$. Její hodnoty (u_{ik}) nabývají spojitě hodnot z intervalu [0,1] a jsou nazývány hodnoty (stupně) příslušnosti k i-té fuzzy podmnožiny „X“. Stejně jako w, $u_i(x_k)=u_{ik}$. Rozšířením vztahů (2.2.3.6a,b) získáme „c“ fuzzy podmnožin Q´={u_1,...u_c} jako nedegenerovaný fuzzy c-rozklad splňující:

$$(2.2.3.7a) \qquad \sum_{i=1}^{c} u_{ik} = 1 \qquad \forall 1 \leq k \leq n$$

$$(2.2.3.7b) \qquad 0 \neq u_i \qquad \forall 1 \leq i \leq c$$

2.2.4 FUNKCIONÁL KVALITY

Funkcionál obecně je předpis, který každému prvku z jeho definičního oboru přiřadí číslo. Oproti funkci ovšem jeho definiční obor netvoří čísla z nějaké množiny, ale funkce. Funkcí je „U“, matice příslušnosti, která každému „x_k“ přiřadí u_{ik}. Předpis vychází z již dlouho používaného funkcionálu „WGSS“ - sumy čtverců odchylek od centroidu shluku.

$$(2.2.4.1) \quad J_w(U, \bar{v}) = \sum_{i=1}^{c} \left(\sum_{x_k \in X} \|x_k - v_i\|^2 \right)$$

Hledání minima tohoto funkcionálu je iterativní postup nazývaný podle autorů ISODATA (Iterative Self-Organizing Data Analysis Techniques).

Dunn[7] ve své práci zavedl zobecnění ISODATA a rozšířil pro fuzzy funkce příslušnosti. Jako P(c) označuje množinu všech degenerovaných hard rozkladů a $P_f(c)$ jako množinu všech degenerovaných fuzzy rozkladů. Jako první fuzzy rozšíření uvažuje funkcionál J(U,v):

$$(2.2.4.2) \quad J(U,\bar{v}) = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik} \left\| x_k - v_i \right\|^2$$

Označíme-li „ U_i' “ a „ v_i' “, (v rovnicích toto označení často splývá splívá s „t“ jako transpozice), jako matici příslušnosti a centroid i-tého clusteru takové, že funkcionál „J“ nabývá minima, potom platí:

$$(2.2.4.3a)$$
$$\zeta = \min_{1\le i\le c} \left\| x_k - v_i' \right\|$$
$$I = \left\{ i \middle| \; 1 \le i \le c \middle| \; \left\| x_k - v_i' \right\| = \zeta \right\}$$
$$\widetilde{I} \; = \left\{ i \middle| \; 1 \le i \le c \middle| \; \left\| x_k - v_i' \right\| > \zeta \right\}$$
$$potom$$
$$i \in \widetilde{I} \;\; \Rightarrow u_i'(x_k) = 0 \quad \wedge \quad \sum_{i\in I} u_i'(x_k) = 1$$

Pozn.
ζ nemusí být pouze jedno minimum, existuje-li pouze jedno, pak I sestává z jednoho „i“.

(2.2.4.3b)
Pro všechna $1 \le i \le c$ existuje nějaké $u_i'(x_k) \neq 0$, tj. optimální rozklad dosažený minimalizací „J“ je nedegenerovaný a vždy obsahuje c-neprázdných množin.

(2.2.4.3c)
Centroid se počítá podle:

$$v_i' = \frac{\sum_{k=1}^{N} u_i'(x_k).x_k}{\sum_{k=1}^{N} u_i'(x_k)}$$

Důkaz:

(2.2.4.3a)
Zřejmě platí:

$$(2.2.4.3a1)\quad J_1(U',v') \le J_1(U,v') \Rightarrow \sum_{i=1}^{c} u_i'(x_k)\left\|x_k - v_i'\right\|^2 \le \sum_{i=1}^{c} u_i(x_k)\left\|x_k - v_i'\right\|^2$$

$$(2.2.4.3a2)\quad \sum_{i=1}^{c} u_i'(x_k)\left\|x_k - v_i'\right\|^2 = \min_{w\in U_f} \sum_{i=1}^{c} w_i(x_k)\left\|x_k - v_i'\right\|^2 =$$

$$= \left(\sum_{i\in I} w_i\right)\zeta^2 + \sum_{i\in I^c} w_i(x_k)\left\|x_k - v_i'\right\|^2 \ge \left(\sum_{i=1}^{c} w_i\right)\zeta^2 = \zeta^2$$

Z (2.2.4.3a2) vyplývá že $u_i(x_k)$ na pravé straně (2.2.4.3a1) vyhovuje nerovnosti platí-li:

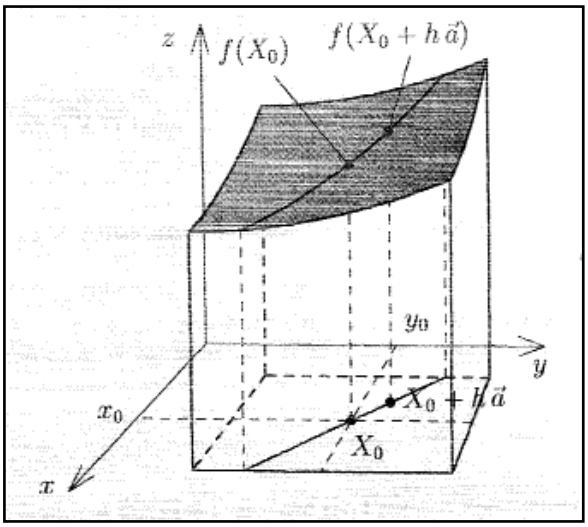
$$u_i'(x_k) = 0 \quad \forall i \in I^c \wedge \quad \sum_{i\in I} u_i'(x_k) = 1$$

(2.2.4.3c)
Důkaz a celé odvození je založeno na derivaci ve směru. O derivaci ve směru se uvažuje v případě, že náhodná veličina je vícerozměrná (více než jeden rozměr). Pro funkci z = f(x,y) udávají její parciální derivace podle „x,y“ rychlost změny funkčních hodnot ve směru osy „x“ a ve směru osy „y“. Tuto úvahu lze zobecnit na jakýkoli směr udaný vektorem ležícím v rovině „xy“ a analogicky ve vícerozměrném prostoru.
Nechť $f(x_1 \ldots x_n)$ je funkcí „n“ proměnných, bod „X“ je z definičního oboru a vektor „a“ (n-rozměrný) má normu jedna: $X \in Df\ ,\ \vec{a} \in V^n \quad \left\|\vec{a}\right\| = 1$, následující limita se potom nazývá derivací funkce „f“ v bodě „X“ a ve směru vektoru „a“ a značí se $Df(X,\vec{a})$:

$$\lim_{h\to 0} \frac{f(X + h\vec{a}) - f(X)}{h}$$

Následující obrázek obr 8. celou situaci dokumentuje:

obr 8.
Derivace ve směru



Nyní zavedeme funkci „g“ pouze jedné proměnné, a to centroidu daného clusteru „v_i“, funkci „g“ ztotožníme s funkcionálem. Funkce příslušnosti pro „x_k“ bereme jako nejlepší možné (a konstantní) z hlediska minimalizace funkcionálu, tedy podle dříve užitého značení máme (matici příslušnosti s čárkou):

$$(2.2.4.3c1)\quad g(v_i) = J_1(U', v_i) = \sum_{k=1}^{N} u_i'(x_k)\left\|x_k - v_i\right\|^2 = \sum_{k=1}^{N} u_i'(x_k)\left\langle x_k - v_i \middle| x_k - v_i \right\rangle$$

Centroid je vektor „p“-čísel ($x_k \in R^p$), je to nezávisle proměnná funkce „g“. Naším cílem je určit takové „v_i“, aby funkcionál nabyl nejmenší hodnoty vůbec, tj. aby derivace funkce „g“ byla rovna nule. Pokud bychom uvažovali derivaci („ g ´ “ = 0) pouze ve směru souřadnicových os (parciální derivace), neměli bychom jistotu, že určená nezávisle proměnná skutečně minimalizuje „g“ (viz. sedlo pro 2D případ). Nejistotu odstraní derivace ve všech možných směrech jednotkového vektoru „w“od nezávisle proměnné „v_i“ (pro 2D případ „w“ leží v jednotkovém kruhu, pro 3D případ „w“ leží v jednotkové sféře…a analogicky pro pD). Vektor „w“ je stejného rozměru jako „v_i“. Derivace ve směru položená rovna nule hledá takové „v_i“, které minimalizuje funkcionál, podle zavedeného značení to je „v_i'“. Dosadíme-li do (2.2.4.3c1) místo „v_i“ nezávisle proměnnou ve směru, dostaneme :

$$(2.2.4.3c2)\quad Dg(v_i', w) = \frac{d}{dh} g(v_i' + hw) = \frac{d}{dh}\left(\sum_{k=1}^{N} u_i'(x_k)\left\langle x_k - v_i' - hw \middle| x_k - v_i' - hw \right\rangle \right)$$

$$(2.2.4.3c3)\quad \frac{d}{dh}\left(\sum_{k=1}^{N} u_i'(x_k)\left\langle x_k - v_i' - hw \middle| x_k - v_i' - hw \right\rangle \right) = \sum_{k=1}^{N} u_i'(x_k)\frac{d}{dh}\left\langle x_k - v_i' - hw \middle| x_k - v_i' - hw \right\rangle$$

výraz na pravé straně (2.2.4.3c2 a 2.2.4.3c3) rozepsán do souřadnic je:

$$\left\langle x_k - v_i' - hw \middle| x_k - v_i' - hw \right\rangle = \left(x_{k1} - v_{i1}' - hw_1\right)^2 + \left(x_{k2} - v_{i2}' - hw_2\right)^2 + ... + \left(x_{kp} - v_{ip}' - hw_p\right)^2$$

derivace potom je:

$$\frac{d}{dh}\left\langle x_k - v_i' - hw \middle| x_k - v_i' - hw \right\rangle = \frac{d}{dh}\left[\left(x_{k1} - v_{i1}' - hw_1\right)^2 + \left(x_{k2} - v_{i2}' - hw_2\right)^2 + ... + \left(x_{kp} - v_{ip}' - hw_p\right)^2\right] =$$
$$= 2\left(x_{k1} - v_{i1}' - hw_1\right)(-w_1) + 2\left(x_{k2} - v_{i2}' - hw_2\right)(-w_2) + ... + 2\left(x_{kp} - v_{ip}' - hw_p\right)(-w_p) =$$
$$= -2\left[\left(x_{k1} - v_{i1}' - hw_1\right)(w_1) + \left(x_{k2} - v_{i2}' - hw_2\right)(w_2) + ... + \left(x_{kp} - v_{ip}' - hw_p\right)(w_p)\right] =$$
$$po\ \ dosazení\ \ h = 0$$
$$= -2\left[\left(x_{k1} - v_{i1}'\right)(w_1) + \left(x_{k2} - v_{i2}'\right)(w_2) + ... + \left(x_{kp} - v_{ip}'\right)(w_p)\right]$$

dosazením do (2.2.4.3c3)

$$(2.2.4.3c4)\quad \sum_{k=1}^{N} u_i'(x_k)\frac{d}{dh}\left\langle x_k - v_i' - hw \middle| x_k - v_i' - hw \right\rangle = -2\sum_{k=1}^{N} u_i'(x_k)\left\langle x_k - v_i' \middle| w \right\rangle$$

Pravou část (2.2.4.3c4) teď položíme rovnou nule a vzniklou rovnici řešíme. Vektor „w“ je jednotkový, a jako takový nemůže být nulový, čili neřeší vzniklou rovnici:

$$-2\sum_{k=1}^{N} u_i'(x_k)\left\langle x_k - v_i' \middle| w \right\rangle = 0 \Leftrightarrow \sum_{k=1}^{N} u_i'(x_k)\left(x_k - v_i'\right) = 0$$

úpravou

$$\left(\sum_{k=1}^{N} u_i'(x_k)\right) v_i' = \sum_{k=1}^{N} u_i'(x_k) x_k$$

a závěrem

$$v_i' = \frac{\sum_{k=1}^{N} u_i'(x_k) x_k}{\sum_{k=1}^{N} u_i'(x_k)}$$

2.2.4.1 ROZŠÍŘENÍ FUNKCIONÁLU

Dunn[7] dále rozšiřuje funkcionál na tvar:

$$(2.2.4.1.1) \quad J(U,\bar{v}) = \sum_{i=1}^{c}\sum_{k=1}^{N} {u_{ik}}^2 \left\|x_k - v_i\right\|^2$$

a dokazuje, že i pro něj platí dříve uvedené vztahy a odvozuje vztah pro výpočet „u_{ik}“ pomocí Lagrangeových multiplikátorů.

V následně vyšlých pracech různých autorů se objevuje výsledný vztah funkcionálu FCM, kde se na rozdíl od (2.2.4.1.1) objevuje m-tá mocnina funkce příslušnosti:

$$(2.2.4.1.2) \quad J_m(U,\bar{v}) = \sum_{k=1}^{N}\sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2 \;\; kde \quad U \in M_{fc} \wedge \bar{v} \in R^{cp} \wedge (d_{ik})^2 = \left\|x_k - v_i\right\|^2$$

$$m \; je \; váhový \; koeficient, m \in [1,\infty]$$

Pro takto zavedený funkcionál zvolme konstantní „m“. Definujme množiny definovaných „i“ ($I_k, \tilde{I}_k$) jako Dunn[7] , pouze místo ζ coby minima uvažujeme přímo nulovou vzdálenost.

$$I_k = \left\{ i \;\middle|\; 1 \le i \le c; d_{ik} = \left\|x_k - v_i\right\| = 0 \right\}$$

$$\tilde{I}_k = \{1,2,....c\} - I_k$$

Potom pro $(U,\bar{v}) \in (M_{fc} \; x \; R^{cp})$ nabývá (2.2.4.1.2) minima pokud platí:

$$(2.2.4.1.3)\quad I_k = 0 \Rightarrow u_{ik} = \frac{1}{\left[\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}\right]}$$

nebo

$$(2.2.4.1.4)\quad I_k \neq 0 \Rightarrow u_{ik} = 0\ \forall i \in \widetilde{I}_k \wedge \sum_{i \in I_k} u_{ik} = 1$$

a potom centroid je:

$$(2.2.4.1.5)\quad v_i = \frac{\sum_{k=1}^{N}(u_{ik})^m x_k}{\sum_{k=1}^{N}(u_{ik})^m}\ \forall i$$

Výše uvedené vztahy tvoří iteraci FCM algoritmu, v té je nejdůležitější vztah pro „u_{ik}“ a vztah pro centroid. Ke vztahu (2.2.4.1.3) dospějeme hledáním vázaného extrému funkcionálu (2.2.4.1.2) metodou LaGrangeovy funkce a multiplikátoru.

ITERACE FCM

Algoritmus Fuzzy C-means, FCM (někdy nazýván fuzzy ISODATA), je obecná Picardova iterace, která iteračně řeší podmínky (2.2.4.1.3),(2.2.4.1.4) a (2.2.4.1.5).

Algoritmus by se dal popsat jako:

I) Volba „c“, volba metriky (většinou Euklidovy), volba váhového exponentu „m“, inicializace „U“ (tj. volba první, náhodné U^0 matice příslušnosti)
II) Výpočet centroidů podle (2.4.1.5)
III) Výpočet nové matice příslušnosti „U“ podle (2.4.1.3)
IV) Porovnání U^{l-1} v předešlé iteraci a U^l v nové, je-li norma jejich rozdílu menší nežli zvolené ε (často 0,01), $\left\|U^l - U^{l-1}\right\| < \varepsilon$, končí výpočetní cyklus

Ve skutečnosti bod IV) neplatí až tak doslova, někdy se bere jako kritérium ukončení rozdíl po sobě jdoucích hodnot funkcionálu. Důležitým parametrem je koeficient „m“. Jeho volba výrazně ovlivňuje výsledek. Čím se více tento koeficient blíží jedničce, tím více tíhne rozklad k „hard“. Naopak s růstem k nekonečnu se jednotlivé u_{ik} blíží 1/c a rozklad je maximálně „fuzzy“.

2.2.4.2 KONVERGENCE FUNKCIONÁLU

Funkcionál FCM obecně konverguje k lokálnímu minimu. Otázka globálního minima nebyla nikdy příliš diskutována, předpokládá se, že dosažený rozklad po mnoha iteracích je ten optimální.

Důkaz konvergence FCM, tj. s každou iterací jsem blíže optimálnímu řešení, je založen za teorému podle Zangwill[12]. V práci J. C. Bezdek[14] autoři ukazují nebezpečí konvergence k sedlovému bodu. Demostrují to na jednorozměrných datech s obráceným FCM, kdy se zvolí jako první centroid a matice příslušnosti se následně dopočítává. Počet clusterů je 4. Vstupní matice dat je řádek: $\{x_1, x_2, x_3, x_4\} = \{-3, -1, 1, 3\}$. Počáteční centroid v^0 je (-1, 0, 1), váhový koeficient m = 2.

Aby autoři ukázali, že výsledkem je sedlový bod, zavedli centroid jako funkci reálné konstanty ε (a-je konstanta z předešlého výpočtu, cca 2,9):

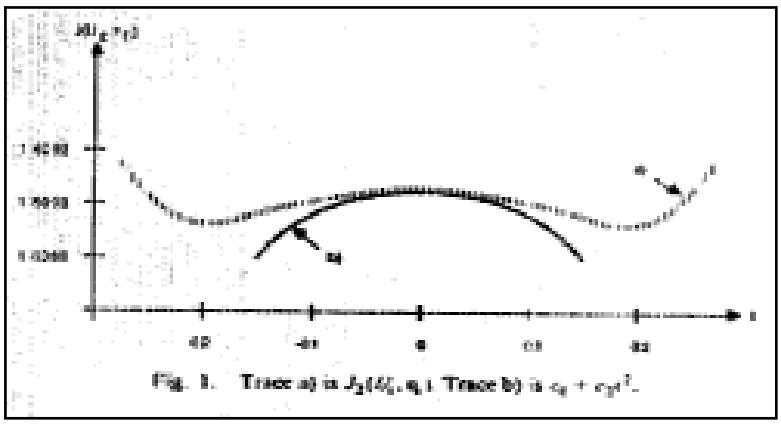
$$v_\varepsilon = (-a, \varepsilon, a)$$

Průběh funkcionálu druhé iterace (první je se zavedeným centroidem) na ε lze ukázat, provede-li se jeho rozvoj podle Taylorova polynomu, dostane se:

$$J_2(U_\varepsilon, v_\varepsilon) = c_0 + c_2\varepsilon^2$$
$$c_0 \cong 1{,}6$$
$$c_2 \cong -0{,}015$$

Z rozvoje jasně plyne, že pro $\varepsilon = 0$ je „J_2“ maximální. Jeho průběh v závislosti na epsilon je na následujícím obrázku 9, obrázek není příliš čitelný, avšak jako demonstrace postačuje. Osa „y“ je funkcionál a „x“ parametr epsilon. Tento příklad jasně dokazuje, že bez ohledu na vstupní matici mohou iterace skončit v sedlovém bodu. Do té doby uvedené práce se zabývaly extrémními případy s neobvyklými vstupními maticemi.

obr 9.
Sedlový bod funkcionálu



Jako druhý příklad sedlového bodu autoři uvádějí případ, kdy po neurčeném počtu iterací je výsledkem matice příslušnosti odpovídající maximální rozmytosti (viz část koeficient rozkladu), tj. :

$$U = \begin{bmatrix} \left(\frac{1}{c}\right) & \cdots & \left(\frac{1}{c}\right) \\ \vdots & & \vdots \\ \left(\frac{1}{c}\right) & \cdots & \left(\frac{1}{c}\right) \end{bmatrix}$$

Potom se dá ukázat, že pro matici D:

$$D = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

existuje epsilon $|\varepsilon| < \frac{1}{c}$ takové, že pro výslednou matici „U_ε“ je „J_m“ minimální:

$$U_\varepsilon = U^* + \varepsilon D$$

$$J_m(U_\varepsilon, v) < J_m(U^*, v^*)$$

$$koeficient m\ musí\ splňovat:$$

$$m < \frac{N}{N-2}$$

Z toho tedy plyne, že J_m nabývá pro (U^*,v^*) hodnoty sedlového bodu. Zkráceně vstupní matice s prvky 1/c vede k iteraci k sedlovému bodu.

Závěrem FCM je tedy jednoznačně konvergentní algoritmus, avšak za určitých podmínek existuje riziko dosáhnutí sedlového bodu.

2.2.5 VALIDITA CLUSTERŮ

Otázka volby optimálního počtu clusterů je jednou z nejdůležitějších, co se týče problematiky clusterové analýzy. Jedna z prvních prací zabývající se objektivními metodami volby je J. C. Bezdek[10] z roku 1974. Posouzení validity rozkladu je založeno na objektivním určení míry překrytí fuzzy množin („rozmytost“), tj. do jaké míry se clustery dělí o prvek „x_k“. Základní úvaha je, čím je rozklad lepší – dosažen optimální rozklad, tím se více blíží k „hard“.

2.2.5.1 KOEFICIENT ROZKLADU

Hodnota funkce příslušnosti prvku „x_k“ vzhledem k množině vzniklé průnikem vícero množin je číslo, které nám řekne s jakou nejmenší možnou příslušností máme chápat prvek „x_k“ jako člena všech sledovaných množin, tj. s jakou největší mírou tíhne prvek „x_k“ ke sledovaným množinám současně. Toto ale nevystihuje rozmytost nejlépe. Účelnější se jeví použití algebraického součinu funkcí příslušnosti.

$Q' = \{u_1 ... u_c\} \in Q_{fc}$, Q_{fc} je třída všech c-tic funkcí příslušnosti.

(2.2.5.1.1) $\left(\frac{1}{N}\right)\sum_{k=1}^{n} u_{ik}u_{jk}$ ... představuje střední vazbu, průměrné propojení mezi množinami u_i, u_j pro $1 \le i, j \le c, kde\ i \ne j$

Říkáme, že množiny u_i, u_j jsou bez vazby (vzájemně se nepronikají) je-li jejich střední vazba nulová, to je zřejmě pouze v případě, že průnik je nulový. Hodnota $u_{ik}u_{jk}$ je přímo úměrná míře, s jakou se „x_k" rozděluje mezi obě množiny.

Označme „V_{cN}" lineární prostor reálných matic rozměru (c *x* N). Na tomto prostoru je definován analogicky ke skalárnímu součinu vektorů skalární součin a metrika matic (A,B) $\in V_{cN}$ jako:

(2.2.5.1.2*a*) $\langle A, B \rangle = tr(AB^t)$

(2.2.5.1.2*b*) $\|A\| = \sqrt{\sum_{ij}(a_{ij})^2} = \sqrt{\langle A, A \rangle}$

(2.2.5.1.2*c*) $d(A, B) = \|A - B\|$

Definujme zobrazení ξ které převádí „Q_{fc}" na „V_{cN}", $\xi : Q_{fc} \rightarrow V_{cN}$. Potom toto zobrazení převede „Q´" na matici hodnot příslušnosti:

(2.2.5.1.3) $\xi(Q') = U = \begin{bmatrix} u_{11} \dots u_{1N} \\ \vdots \quad\quad \vdots \\ u_{c1} \cdots u_{cN} \end{bmatrix}$

(2.2.5.1.4) $S_c(U) = \frac{UU^t}{N}$

Množina matic příslušnosti v prostoru V_{cN} je $M_{fc} = \xi[Q_{fc}]$. Rozmytost lze v maticové podobě zapsat jako:

Pozn. Obvyklý tvar matice příslušnosti je (N *x* c) a pak je v čitateli U^tU.

Matice $S_c(U)$ je rozměru (c *x* c), nazývá se matice podobnosti. Každý prvek této matice odpovídá vztahu (2.2.5.1.1), tedy prvek s_{ij} odpovídá střední vazbě mezi u_i a u_j, a samozřejmě platí $s_{ij} = s_{ji}$, čili matice je čtvercová a symetrická.

Z uvedeného vyplývá:

(2.2.5.1.5) $s_{ij} = 0 \qquad \forall i \ne j \Leftrightarrow U \in m_c\ \ je\ "hard"$

Tedy nediagonální prvek matice je roven nule jedná-li se o „hard" rozklad.

Chování matice S nejlépe odráží její “trace” stopa, tj. součet diagonálních prvků. Je to zobrazení značené jako $F_c : M_{fc} \rightarrow R$.
Platí:

$$(2.2.5.1.6) \quad F_c(U) = tr(S_c(U)) = tr\left(\frac{UU^t}{N}\right) = \frac{\|u\|^2}{N}$$

Stejně jako konvexní funkce i F_c je konvexní a má jedno globální minimum na M_{fc} a lokální maximum na M_c.

Pro $U \in M_{fc}$ platí:

$$(2.2.5.1.7a) \quad \left(\frac{1}{c}\right) \le F_c(U) \le 1 \qquad \forall 2 \le c \le n$$

$$(2.2.5.1.7b) \quad \left(\frac{1}{c}\right) = F_c(U) \Leftrightarrow U = \left[\left(\frac{1}{c}\right)\right]$$

$$(2.2.5.1.7c) \quad 1 = F_c(U) \Leftrightarrow U \in m_c \ \ je\ hard$$

Koeficient $F_c(U)$ se nazývá koeficient rozkladu matice “U”, je ekvivalentní celkové střední vazbě mezi všemi kombinacemi dvojic fuzzy množin. Z toho vyplývá zásadní poznatek, čím je koeficient větší, tím se průniky fuzzy množin zmenšují a rozklad se blíží “hard”. Prvky matice „S_c” poskytují informaci o jednotlivých dvojicích clusterů dat, zatímco koeficient $F_c(U)$ je ukazatelem relativního rozmytí.

Vztah mezi koeficientem rozkladu a validitou clusterů se nejlépe dokumentuje na obecně uznávaném indexu separace. Index separace jasně říká, kdy je provedený rozklad dobrý a objekty jsou dobře separované. Úvahy pro jeho zavedení ihned vyplynou z představ, že objekty jsou separované, jsou-li daleko od sebe v porovnání s centroidy clusterů. Jednoduše jsou-li jakékoli dva objekty z množin “A,B” dál od sebe navzájem než-li průměr (maximální vzdálenost bodů uvnitř množiny) množiny “A” nebo “B”, pak je rozklad dobrý a idnex je větší než jedna. Problém je, že index separace je nepočitatelný pro velká “N”. Index separace pracuje již s rozdělenými objekty, a proto místo funkce příslušnosti pracuje s charakteristickou funkcí.

Dá se ukázat, že když roste index separace a rozklad se tak stává evidentně optimálním, blíží se limitně také koeficient rozkladu k nějaké hodnotě a ta by měla být zřejmě 1. Pro čistý „hard“ rozklad je $F_c = 1$. Hledání maxima F_c s clustery je hledáním optimálního počtu „c“.
Tyto úvahy ovšem stojí na základním předpokladu, a to je existující struktura v datech, tj. že existuje něco, v jehož důsledku jsou data clusterovatelná (blíže viz závěr entropie rozkladu).

Index separace β(c,W) se uvádí jako:

$$\beta(c,W) = \frac{\min\limits_{1\le i\le c}\left\{\min\limits_{\substack{1\le j\le c\\ i\neq j}}\left\{dis(Y_i, conv(Y_j))\right\}\right\}}{\max\limits_{1\le k\le c}\left\{dia(Y_k)\right\}}$$

Index separace je závisle proměnná na "c" a "W", kde W=(Y_1…Y_c) $\in$ M_c, to znamená, že W je libovolný hard c-rozklad.

- dia () je průměr množiny (největší vzdálenost dvou prvků v množině)

- dis () je vzdálenost mezi dvěma množinami (nejmenší vzdálenost mezi libovolným bodem v jedné a libovolným bodem v druhé množině)

- conv() je konvexní obal množiny

Cílem clusterové analýzy je, aby byl index separace po rozkladu co největší. Počítat pro každý člen množiny "W", tj. pro všechny možné hard-rozklady "X", index separace je příliš složité a tak to lze obejít pomocí koeficientu rozkladu.

Důležitý závěr je, že s rostoucím indexem separace se F_c(U) blíží jedničce a v limitním případě "hard" rozkladu, kdy je W = U, je koeficient rozkladu roven jedné. Tedy užití koeficientu rozkladu pro odhad optimálního počtu clusterů je oprávněným předpokladem.

2.2.5.2 ENTROPIE ROZKLADU

N. Wiener a C. E. Shannon, dva američtí vynikající matematici-zakladatelé teoretické kybernetiky, v letech 1948-49 zavedli informační pojem entropie, který vychází ze statistického charakteru sdělení. Došli k zobecnění pojmu entropie chápané jako míra nepořádku vůbec. Entropii chápou jako míra neurčitosti před přijetím zprávy, jež se po příjmu odstraňuje a vyjadřuje tak míru informace. Je to míra neurčitosti, kterou má průměrně jedno písmeno zprávy. Nebo ještě jinak, je to míra informace, kterou v průměru získáme přečtením jednoho znaku.

Podle jejich návrhu je střední hodnota informace připadající na jeden symbol zprávy rovna celkovému množství děleném délkou zprávy. Výsledkem je nejčastěji užívaný vztah pro informační entropii značenou H.

$$(2.2.5.2.1)\quad H = \frac{I}{n} = -K\sum_{i=1}^{s} p_i \ln p_i$$

-p_i je pravděpodobnost výskytu znaku „i" ve zprávě

Z formálního hlediska je entropie diskrétní nebo spojitou funkcí pravděpodobností jednotlivých znaků. Dále uvažujeme pouze jednorozměrnou informační entropii.

$$H(p_1,....p_s) = -K\sum_{i=1}^{s} p_i \ln p_i$$

Informační entropie má tyto vlastnosti:

1) entropie $H(p_1,….p_s)$ je nezáporné číslo, je pozitivní
2) malá změna p_i vyvolá malou změnu entropie, tj. entropie $H(p_1,….p_s)$ je spojitá ve svých proměnných
3) entropie nezávisí na pořadí (je symetrická), v jakém jsou zapsány zdrojové znaky
4) entropie je koherentní vůči zdroji, tj. entropii zdroje (popř. zprávy), který vysílá více než dva znaky je lze určit z entropií menších zdrojů

Pozn.
Logaritmus o základu „r“ se vypočte z přirozeného logaritmu následovně:

$$\log_a x = \frac{\ln x}{\ln a}$$

Vyjádříme-li konstantu K ve vztahu jako: $K = \frac{1}{\ln a}$, dostaneme pro H:

$$(2.2.5.2.2) \quad H(p_1,....p_s) = -\sum_{i=1}^{s} p_i \log_a p_i$$

Velmi důležitá je volba konstanty „r“, která je v podstatě volbou jednotky entropie. Zvolíme-li r = 2, je jednotkou bit, tj. říkáme, že hodnota entropie je tolik a tolik bitů. Právě pro takto zvolenou volbu jednotky entropie si ukážeme její dvě důležité vlastnosti, kdy nabývá maximální a kdy minimální hodnoty.

I) Entropie je rovna nule tehdy, jsou-li všechny pravděpodobnosti znaků kromě jedné z nich rovné nule a jedna pravděpodobnost je rovna jedné.

Důkaz:
Pro všechny p_i platí $0 \le p_i \le 1$, proto jejich logaritmus je záporný, maximálně nulový. Logaritmus se základem větším než 1 je rostoucí funkce, která pro nezávisle proměnné blížící se nule nabývá hodnoty (–∞). Z tohoto důvodu je uzančně definováno 0.log0 = 0. Z výše uvedeného plyne, že suma $-\sum_{i=1}^{s} p_i \log_2 p_i$ nabývá nulové hodnoty tehdy a pouze tehdy, když platí :

$$p_{i\neq k} = 0 \quad \wedge \quad p_k = 1$$

II) Jsou-li pravděpodobnosti výskytu všech hodnot znaku p_i stejné, tj. když

$$p_1 = p_2 = ... = p_s = \frac{1}{s}$$

informační entropie dosahuje maxima. Znaky musí mít samozřejmě rovnoměrné rozdělení.

Pozn.
Obecně pro dvojznakovou zprávu s rovnoměrným rozdělením četnosti, kdy $p(A) = p_1 = p \Rightarrow \quad p(B) = p_2 = 1-p$,

přechází $H(p_1,....p_s) = -\sum_{i=1}^{s} p_i \log_2 p_i$ na tvar:

$$H(p,1-p) = p\log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{(1-p)} = -p\log_2 p - (1-p)\log_2(1-p) \quad bitů$$

Funkce H je definována na intervalu $0 \leq p \leq 1,$, její graf na obrázku 10. ukazuje uvedené vztahy a vlastnosti informační entropie obecně, tj. pozitivitu, spojitost a symetrii.

obr 10.
Graf informační entropie dvouznakové zprávy

Prvním kdo definoval entropii na nestatistickém základu ve vztahu k fuzzy množinám byly autoři A. De Luca a S. Termini[9]. Uvažujme dvě fuzzy množiny s funkcemi příslušnosti f(x) a g(x), obě definované na množině „I“.

Pro jejich sjednocení a průnik platí:
$(f \vee g)(x) = \max(f(x), g(x))$
$(f \wedge g)(x) = \min(f(x), g(x))$

Chceme vyjádřit stupeň rozmytosti množiny tak jako v případě koeficientu rozkladu. Požadujeme, aby veličina která rozmytost měří a je označena d(f) podle fuzzy množiny „f“, splňovala tyto vlastnosti:

a) d(f) musí být rovna nule pouze tehdy, když funkce příslušnosti je 0 nebo1
b) d(f) musí dosáhnout maxima je-li „f“ pro všechna „x“ rovna 1/2
c) d(f) musí být větší nebo rovna $d^*(f)$, kde f^* je zjednodušená funkce příslušnosti, tj. $f^*(x) \geq f(x)$ pokud $f(x) \geq 1/2$ a $f^*(x) \leq f(x)$ pokud $f(x) \leq 0$

Zavedeme funkcionál H(f) s definičním oborem funkcí „f“ definovaných na množině „I“ formálně podobný Shannonově informační entropii, jehož obor hodnot je množina nezáporných reálných čísel.

$$(2.2.5.2.3) \quad H(f) = -K\sum_{k=1}^{N} f(x_k)\ln f(x_k)$$

N…je počet prvků
K…je kladná konstanta

Platí:

H(f) je nezáporná hodnota, pro kterou platí:
$H(f \vee g) + H(f \wedge g) = H(f)+H(g)$

Pokud bychom označili výše zavedený funkcionál jako míru rozmytosti fuzzy množiny, musel by splňovat již uvedené tři vlastnosti (a,b,c), avšak již druhá neplatí:

a) H(f) = 0 pro f(x) = 0 nebo f(x) = 1, toto vyplývá přímo z dosazení do předpisu pro H(f) (samozřejmě i zde předpokládáme 0.ln0=0).

b) maximum H(f) je pro f(x) = 1/e, tj. H(f) = K.N/e což nesplňuje náš požadavek

Z tohoto důvodu zavedeme d(f) jako:

$d(f) = H(f) + H(\bar{f}), kde \quad \bar{f}(x) = 1 - f(x)\ a\ splňuje:$

$i) \quad \bar{\bar{f}} = f$

$ii) \quad \overline{f \vee g} = \bar{f} \wedge \bar{g}$

$iii) \quad \overline{f \wedge g} = \bar{f} \vee \bar{g}$

$zřejmě\ platí:$

$d(f) = d(\bar{f})$

Shannonovu informační entropii pro dva znaky s rovnoměrným rozdělením lze vztáhnout na zavedený funkcionál rozmytosti fuzzy množiny d(f). Tvar Shannonovy entropie pro jednu nezávisle proměnnou je „(H(x) = -x.lnx-(1-x).ln(1-x))“ (to že v předpisu je najednou přirozený logaritmus místo dekadického, není

špatně……pro K = 1/a platí rovnost $\log_a x = \ln x$). Celková entropie fuzzy množiny se odvodí dosazením za d(f) podle definic (2.2.5.2.3):

$$H(f) = -K\sum_{k=1}^{N} f(x_k)\ln f(x_k)$$

$$H(\bar{f}) = -K\sum_{k=1}^{N}(1-f(x_k))\ln(1-f(x_k))$$

$$d(f) = H(f) + H(\bar{f}) = -K\sum_{k=1}^{N} f(x_k)\ln f(x_k) - K\sum_{k=1}^{N}(1-f(x_k))\ln(1-f(x_k)) =$$

$$= K\sum_{k=1}^{N}\left[-f(x_k)\ln f(x_k) - (1-f(x_k))\ln(1-f(x_k))\right] = K\sum_{k=1}^{N} S(f(x_k))$$

Takto zavedené d(f) již krom prvních dvou požadavků a) b) splňuje i třetí c).
Třetí požadavek lze shrnout jako:

1) $0 \le f^*(x) \le f(x) \le 1/2$ pro $0 \le f(x) \le 1/2$
2) $1 \ge f^*(x) \ge f(x) \ge 1/2$ pro $1 \ge f(x) \ge 1/2$

Z průběhu informační entropie dvojznakového zdroje zprávy, tj. funkce je na intervalu (0,1/2) monotónně rostoucí a na (1/2,1) monotónně klesající, lze odvodit že:

$S(f^*(x)) \le S(f(x))$

Z toho následně ihned plyne, že:
$d(f^*) \le d(f)$

Pro d(f) stejně jako pro H(f) platí:
$d(f) + d(g) = d(f \vee g) + d(f \wedge g)$

Pokud je konstanta ve vztahu pro d(f) rovna 1/N, potom d(f) je „normalizovaná entropie“ fuzzy množiny v(f), pro kterou platí:

i) $v(f) = \frac{1}{N}\sum_{k=1}^{N} S(f(x_k))$

ii) $0 \le v(f) \le 1$

iii) $v(f) + v(g) = v(f \vee g) + v(f \wedge g)$

Zavedená normalizovaná entropie fuzzy množiny ještě není entropií rozkladu. Entropie rozkladu je v podstatě entropie více fuzzy množin dohromady jako celku.

Entropie rozkladu je definována jako:

$$(2.2.5.2.4)\quad H(U,c) = -\sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}\log_a(u_{ik})/N \qquad kde\, 1 \le c \le N\,;\, a \in (1,\infty)\,;\, 0.\log_a 0 = 0$$

Je-li c = 2, potom každý objekt „k“ patří pouze do dvou množin s pravděpodobností u_{1k} a u_{2k}, pro které platí:

$$\sum_{i=1}^{c} u_{ik} = 1 \quad \forall k$$

tedy $u_{2k} = 1-u_{1k}$, v takovém případě se vztah entropie rozkladu redukuje na již známý vztah normalizované entropie.

$$c = 2 \Rightarrow v(u) = H(U,2) \Leftrightarrow \frac{1}{N}\sum_{k=1}^{N} S(u(x_k)) = -\sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}\log_a(u_{ik})/N$$

Normalizovaná entropie fuzzy množiny je tak speciálním případem entropie rozkladu.

Pro entropii rozkladu platí:

$i)\quad 0 \le H(U,c) \le \log_a(c)$

$ii)\quad H(U,c) = 0 \Leftrightarrow U \in M_{co},\ rozklad\ je\ "hard"$

$iii)\quad H(U,c) = \log_a(c) \Leftrightarrow U = \frac{1}{c}$

Entropie rozkladu je číslo, které říká, jak je rozklad „dobře udělaný“. Jak „moc“ jsme si jistí, že objekt patří zrovna do tohoto clusteru. Čím je toto číslo větší, tím existuje větší nejistota při rozhodování. Validitou clusterů je obecně myšlena oprávněnost výsledků clusterové analýzy, tedy jistota s jakou tvrdím, že objekt patří do daného clusteru. Chci-li dosáhnout větší jistoty, musím zmenšit entropii rozkladu. Při optimálním počtu clusterů, což má zásadní význam pro výslednou validitu, by měla být entropie rozkladu být minimální, tj. hledám minimum H(U,c) přes všechny „c“, a přes všechny možné matice „U“.

Pozn.
Výše uvedené platí zejména pro „clustrovatelné“ data. Takové datové soubory skutečně obsahují struktury , mezi daty existuje vztah. Minimalizace H(U,c) potom najde optimální řešení, tj. fuzzy rozklad se bude blížit hard rozkladu při $c << N$. Existují ovšem data s tendencí „netvořit clustery“,tj. data prostě v sobě neobsahují informaci o clusterech. V takovém případě by se hledání lokálního minima H(U,c) zvrhlo ve tvorbu příliš mnoha clusterů ze sebeméně „ulétlých“ datových bodů, až by $c \to N$, což jaksi není cílem clusterové analýzy.

Závěrem, u koeficientu rozkladu se snažíme nalézt maximum, kdežto u entropie rozkladu naopak minimum. Mezi oběma indexy platí vztah:

$$(2.2.5.2.5)\quad 0 \le 1 - F(U,c) \le \frac{H(U,c)}{\log_a(e)} \quad kde\, a \in (1,\infty)\,,\ "e"\ je\ Eulerovo\ \ číslo$$

2.2.5.3 NORMALIZACE A STANDARDIZACE $F_C(U)$ A $H_C(U)$

Pozn. Oba koeficienty jsou funkcionály matice příslušnosti, jako jejich značení s důrazem na závislost na počtu clusterů se používá: Fc(U), Hc(U).

Pro „hard“ rozklad je Hc(U) rovna nule. Tedy je-li c =1 nebo c = N, Hc(U) = 0. V reálných obsáhlých datových souborech je nepravděpodobné, aby c $\rightarrow$ N, u takových souborů se projevuje negativní tendence obou koeficientů s rostoucím „c“ k monotonicitě. V takovém případě prakticky není možné určit jakýkoli extrém obou koeficientů a tudíž ani optimální počet clusterů. Pro „N“ řádově stovky jejich hodnota s „c“ poměrně rychle (řekněme už u c = 7 je rozdíl hodnoty od limitní o {5-10}%) stoupá (klesá) a drží se prakticky konstantní, až zase pro c = N nabývá limitních hodnot 0 nebo 1.
Samozřejmě, že tato situace nastane tím spíš, čím data neobsahují clusterovatelné struktury vůbec. Podobné úvahy vedli k zavedení normalizovaných a standardizovaných indexů validity. Standardizované indexy mají střední hodnotu rovnu nule a směrodatnou odchylku jedné. Normalizované indexy nemají tendenci k monotonicitě. Jejich chování je obdobné klasickým indexům, např. optimální "c" je pomocí normalizované entropie určeno opět jako minimum entropie. Odvození vztahů takto pozměněných koeficientů není až tak jednoduché, pochopení vyžaduje rozsáhlejší matematický aparát.

NORMALIZOVANÉ KOEFICIENTY

Normalizované koeficienty $\widetilde{Hc}(U)$, $\widetilde{Fc}(U)$ entropie a rozkladu, nabývají hodnot z intervalu [0,1] a jsou definovány pro $2 \le c < N$.

$$(2.2.5.3.1a) \quad \widetilde{Hc}(U) = \frac{Hc(U)}{\log_a c} \quad \forall\, 1 < a < \infty$$

$$(2.2.5.3.1.2) \quad \widetilde{Fc}(U) = \frac{c}{c-1}\left(1 - Fc(U)\right)$$

Pozn. Někdy se jako normalizovaná entropie uvádí jiný vztah, jehož chování je podobné již uvedenému.

$$(2.2.5.3.1b) \quad \widetilde{H}^{*}c(U) = \frac{N.Hc(U)}{N-c}$$

Střední hodnota μ a směrodatná odchylka σ^2 takto zavedených koeficientů potom je:

$$\mu\left[\widetilde{H}c(U)\right] = \frac{\left(\sum_{i=2}^{c}\frac{1}{i}\right)}{\log_a c}$$

$$\sigma^2\left[\widetilde{H}c(U)\right] = \frac{\left[\sum_{i=2}^{c}\frac{1}{i^2} - \left(\frac{c-1}{c+1}\right)\left(\frac{\pi^2}{6}-1\right)\right]}{\left(\log_a c\right)^2}$$

$$\mu\left[\widetilde{H}^* c(U)\right] = \left(\frac{N}{N-c}\right)\left(\sum_{i=2}^{c}\frac{1}{i}\right)$$

$$\sigma^2\left[\widetilde{H}^* c(U)\right] = \left(\frac{N}{N-c}\right)^2\left[\sum_{i=2}^{c}\frac{1}{i^2} - \left(\frac{c-1}{c+1}\right)\left(\frac{\pi^2}{6}-1\right)\right]$$

$$\mu\left[\widetilde{F}c(U)\right] = \frac{c}{c+1}$$

$$\sigma^2\left[\widetilde{F}c(U)\right] = \frac{4c^2}{N(c-1)(c+2)(c+3)(c+1)^2}$$

STANDARDIZOVANÉ KOEFICIENTY

Standardizované koeficienty $\widehat{H}c(U)$ a $\widehat{F}c(U)$ jsou definovány pro $2 \le c < N$. Takto upravené koeficienty mají pro velké datové soubory N="stovky" charakter náhodné veličiny s normálním rozdělením:

$\widehat{H}c(U) \approx N(0,1)$

$\widehat{F}c(U) \approx N(0,1)$

Standardizovaná entropie často vychází z normalizované entropie $\widetilde{H}^* c(U)$.

$$(2.2.5.3.3)\quad \widehat{H}c(U) = \frac{Hc(U) - \sum_{i=2}^{c}\frac{1}{i}}{\sqrt{\left(\sum_{i=2}^{c}\frac{1}{Ni^2} - \left(\frac{c-1}{c+1}\right)\left(\frac{\pi^2-6}{6N}\right)\right)}}$$

$$(2.2.5.3.4)\quad \widehat{F}c(U) = \sqrt{\left(\frac{N(c+2)(c+3)}{c-1}\right)}\left(\frac{(c+1)Fc(U)}{2}-1\right)$$

2.2.5.4 INDEXY VALIDITY XIE-BENI A FUKUYAMA-SUGENO

V posledních letech se objevují návrhy jak vytvořit indexy, které lépe odrážejí kvalitu rozkladu, indexy, které nejeví rysy monotonicity se vzrůstajícím počtem clusterů. Oba zmíněné koeficienty, rozkladu a entropie, počítají kvalitu rozkladu pouze na základě funkce příslušnosti. Takový přístup se ukazuje jako nevhodný pro data s relativně

velkým rozptylem, což většinou chemická data jsou. Zkrátka, čím je v datech obsažena méně výrazná clusterovatelná struktura, tím méně se hodí použití Fc(U) a Hc(U), další nevýhodou obou koeficientů je skutečnost, že když „zkolabuje" jeden, potom s největší pravděpodobností zkolabuje i druhý. Aby se toto odstranilo, měl by být index validity krom funkce příslušnosti závislý i na hodnotě samotného funkcionálu, na matici centroidů, na váhovém koeficientu apod.
Toto splňují nové indexy Xie-Beni a Fukuyama-Sugeno. V práci Xuanli Lisa Xie a Gerardo Beni[25] autoři bez uvedení úvah, které je k tomu vedli, zavádějí nový funkcionál validity „S".

$$(2.2.5.4.1) \quad S(U,\bar{v},x) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u^{2}{}_{ij}\left\|v_i - x_k\right\|^2}{N \min_{i,j}\left\|v_i - v_j\right\|^2}$$

Funkce příslušnosti v čitateli „S" nezávisí na algoritmu jakým byla vypočítána. Volíme-li váhový koeficient u FCM m = 2, dostává „S" tvar:

$$S(U,\bar{v},x) = \frac{J_2}{N \min_{i,j}\left\|v_i - v_j\right\|^2}$$

Minimalizací takto zavedeného funkcionálu minimalizujeme čitatele J_2 a zároveň maximalizujeme jmenovatele tj. minimální vzdálenost centroidů. Intutitivně chápeme, že minimalizací „S" dostáváme optimální rozklad.
V části o koeficientu rozkladu byl ukázán index separace podle Dunn $\beta(c,W)$, který je zjevně pro optimální rozklad >1. Dá se ukázat, že nový funkcionál splňuje:

$$S(U,\bar{v},x) \le \frac{1}{\left(\beta(c,W)\right)^2}$$

Nikhil R. Pal a J. C. Bezdek[26] ve své práci porovnávají 4 indexy validity, koeficient rozkladu+entropie rozkladu+Xie-Beni+Fukuyama-Sugeno, na stejném datovém souboru. Ukazují, jak se indexy chovají s proměnným váhovým koeficientem „m". Podle této práce je index Fukuyama-Sugeno následující:

$$(2.2.5.4.2) \quad FS(U,\bar{v},x) = \sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik})^m\left(\left\|x_k - v_i\right\|^2 - \left\|v_i - \bar{x}\right\|^2\right) =$$

$$= J_m(U,\bar{v}) - \sum_{i=1}^{c}\left\{\left(\sum_{k=1}^{N}(u_{ik})^m\right)\left\|v_i - \bar{x}\right\|^2\right\}$$

kde $\bar{x}$ celkový průměr přes všechna „k".

V současné literatuře se indexy validity Xie-Beni a Fukuyama-Sugeno označují:

$v_{XB,m}(U,V;X)$

$v_{FS,m}(U,V;X)$

2.2.6 MODIFIKACE FCM, GUSTAFSSON-KESSELOVA METODA

Obecně funkcionál kvality FCM se dá zapsat jako:

$$J_m(U,\bar{v},A)=\sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik})^m\left\|x_k-v_i\right\|^2{}_A=\left\langle x_k-v_i\middle|x_k-v_i\right\rangle_A=(x_k-v_i)^T A(x_k-v_i)$$

„A“ je matice rozměru (p *x* p). Zůstává-li „A“ stejné pro každou iteraci a jedná-li se o pozitivně definitní matici, potom hovoříme o klasickém algoritmu FCM. Clustery jsou potom ovlivňovány rozptylem jednotlivých znaků v původních datech. Nejčastěji se „A“ volí jako matice identity, tj. jedničky na diagonále, a potom nejsou vzdálenosti ničím váženy. Volím-li „A“ jako symetrickou pozitivně definitní (výsledný skalární součin musí být kladný) matici s převrácenými rozptyly znaků na diagonále (mimodiagonální prvky jsou nulové), potom vzdálenost vážená touto maticí je ovlivňována rozptyly znaků. Částečně tak odstraňuji velkou variability v datech. Při třetí variantě se jako „A“ používá kovarianční matice původních dat, tato volba potlačuje rozptyl znaků a zároveň bere v úvahu vzájemné vztahy mezi znaky. Jestliže místo kovarianční matice A se použije fuzzy kovarianční matice, dostaneme Gustafsson-Kesselovu metodu. U této metody je však nutno počítat fuzzy kovarianční matice pro jednotlivé shluky a iterace. Výsledkem nejsou kulové clustery nýbrž elipsoidní, které lépe vystihují lineárnost v datech (např. 2D data mají tendenci tvořit přímku, křivku, elipsu nebo jiné lineární tvary).
U lineárních dat je tedy vhodnější metoda Gustafsson-Kesselova, která lépe vystihuje přirozenou povahu dat. Její hlavní nevýhodou je, že na datových souborem s extrémními hodnotami dochází k výpočetním komplikacím, protože je problém zvolit optimální tvar elipsoidu. Jeho výpočet je oproti FCM složitější a jeho teoretické zdůvodnění je poměrně komplikované. Tohoto způsobu bylo použito např. v práci Rousseeuw[27], kde této metody bylo použito k optimální přípravě jaderného paliva na bázi obohaceného uranu.

3. PRAKTICKÁ ČÁST

3.1 ČESKÝ HYDROMETEOROLOGICKÝ ÚSTAV

Historie ochrany ovzduší v HMÚ v Praze začala v r. 1967, kdy byla zřízena speciální složka čistoty ovzduší jako odborná základna objektivního sledování a hodnocení vývoje znečištění ovzduší pro potřeby ministerstev. Již od počátku činnosti oboru ochrana čistoty ovzduší v HMÚ byla hlavní pozornost zaměřena na monitorování kvality ovzduší. Již na počátku sedmdesátých let pracovalo v sítích HMÚ více než sto stanic pravidelně měřících znečištění ovzduší převážně diskontinuálními manuálními metodami. Některé stanice byly později vybavovány automatickými analyzátory. Soustavné sledování kvality ovzduší v sítích stanic představuje obrovské množství údajů. Kvalitativní změnou ve zpracování výsledků měření byla realizace Interního informačního systému. Tento systém zahájil činnost v r. 1971 a jeho výsledkem bylo vytvoření banky dat na nových počítačích a vydávání tištěných ročenek čistoty ovzduší. Vytvoření úseku ochrany čistoty ovzduší v ČHMÚ počátkem devadesátých let posílilo postavení ústavu v této oblasti tak, že v současnosti je ČHMÚ vedoucí institucí na úseku sledování a vyhodnocování kvality ovzduší v ČR. Mezi stěžejní činnosti úseku ochrany prostředí patří Automatizovaný imisní monitoring (AIM) , který se stal páteřním systémem imisního monitorovacího systému ČR. AIM tvoří 95 stanic na celém území ČR vybavených analyzátory SO_2, NO, NO_2, NO_x,polétavého prachu. Na 33 stanicích jsou analyzátory O_3, CO, na vybraných stanicích meteorologická čidla. Dále existuje manuální imisní síť, která slouží pro doplnění a zahuštění měřící sítě na území celé republiky a pro rozšíření spektra měřených komponent i na složky, které nejsou měřitelné automatickými analyzátory v on-line režimu.
Od roku 1992 je Imisní informační systém vedle ostatních informačních agend kvality ovzduší, integrální součástí Informačního systému kvality ovzduší (ISKO). Každoročně jsou do této imisní databáze ukládána kromě údajů ze síti ČHMÚ a hygienické služby i data ze stanic sítí Výzkumného ústavu lesního hospodářství a myslivosti (VÚLHM), Organizace pro racionalizaci energetických závodů (ORGREZ) a řady institucí a ústavů resortu zemědělství, především z Výzkumného ústavu rostlinné výroby a ze sítí společnosti Ekotoxa.
Jak bylo uvedeno v předmluvě, naměřená data pocházejí ze stanic v Severních Čechách. V následující tabulce je uveden přehled počtu měřících míst registrovaných v IIS-ISKO v Severočeském kraji podle vlastníka v roce 1997.

Rok 1997

V uvedeném čísle je zahrnuto několik typů stanic, ať už podle měřené sloučeniny nebo podle druhu provozu stanice. V Severočeském kraji jsou zahrnuty tyto okresy: Chomutov, Most, Teplice, Ústí nad Labem.

Tab. 6
Přehled měřících míst registrovaných v IIS-ISKO v Severočeském kraji

	1997				
	Označení vlastníka				
region	ČHMÚ	HS	VÚRV	ORG	(celkem)
Severočeský	25	17	5	25	72

Vlastníci:

CHMÚ	Český hydrometeorologický ústav
HS	Hygienická služba
VURV	Výzkumný ústav rostlinné výroby
ORG	Organizace pro racionalizaci energetických závodů

Nejpodstatnější částí imisního monitoringu jsou automatické stanice měřící zároveň SO_2, NO_x, a PM_{10}. Z celkového počtu 72 stanic ve vybraných okresech to činí 13 stanic. V následujících tabulkách jsou rozepsány tyto AIM-stanice, jejich umístění, vlastník a identifikační číslo, vše stejně pro roky 1997.

Tab. 7
AIM stanice vybraných polutantů v Severočeském kraji

Okres Most		
ID	Umístění	Vlastník
1004	**Fláje**	**ČHMÚ**
1005	**Most**	**ČHMÚ**
1317	**Rudolice v Horách**	**ČHMÚ**

Okres Chomutov		
ID	Umístění	Vlastník
1001	**Chomutov**	**ČHMÚ**
1000	**Měděnec**	**ČHMÚ**
1002	**Tušimice**	**ČHMÚ**

Okres Teplice		
ID	Umístění	Vlastník
1007	**Krupka**	**ČHMÚ**
1008	**Teplice**	**ČHMÚ**
1009	**Všechlapy**	**ČHMÚ**
1226	**Bílina**	**HS**

Okres Ústí nad Labem		
ID	Umístění	Vlastník
1010	**Chabařovice**	**ČHMÚ**
1011	**Ústí n. Labem-Kočkov**	**ČHMÚ**
1012	**Ústí n. Labem-město**	**ČHMÚ**

3.2 METODY MĚŘENÍ NA AIM-STANICÍCH

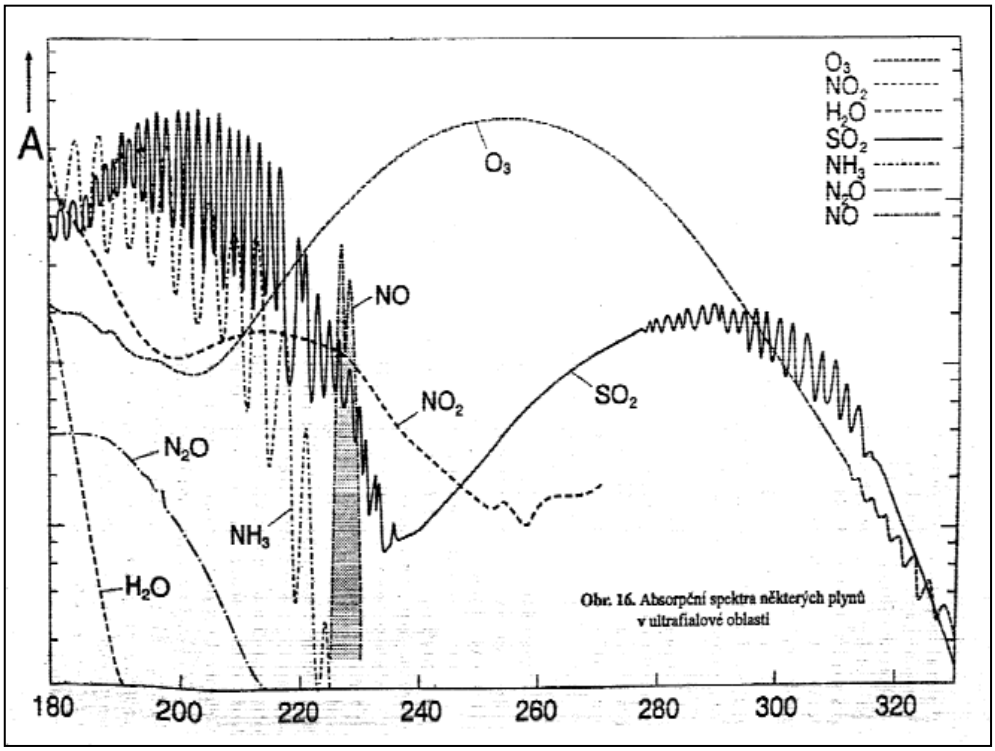
Všechny tyto stanice pracují v kontinuálním režimu s automatizovaným sběrem a vyhodnocením po 30 minutách. Automatizované monitorovací stanice na měření

SO_2, NOx a PM_{10} pracují na všeobecně známých a přijatých principech. V průběhu let dochází k jejich částečné obměně a tím ke změně výrobce a jejich parametrů. V letech 1997 to byly ve vybraných okresech pro PM_{10} stanice firmy VEREWA využívající radiometrické stanovení prachu, pro SO_2 fluorescenční analyzátor Thermo Environmental Instruments a pro NO_x chemiluminiscenční analýzatory firmy Thermo Environmental Instruments.

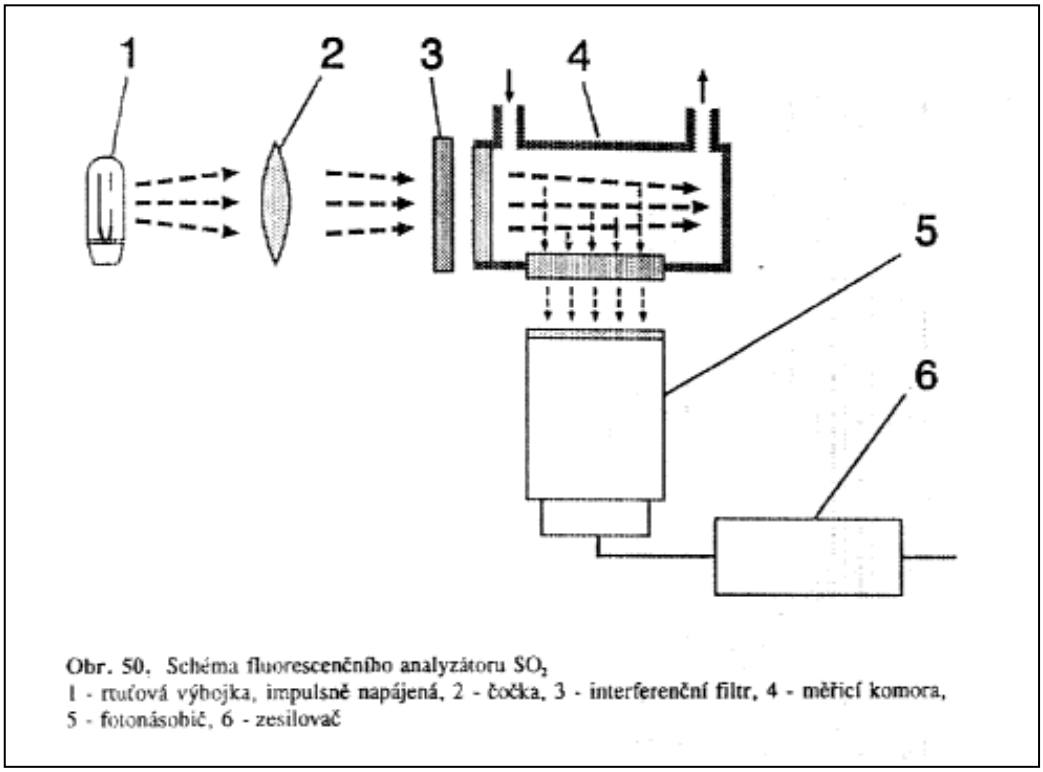
3.2.1 MĚŘENÍ SO_2

Princip měření všech automatických stanic TEI-43, Thermo Environmental Instruments SO_2, by se dal shrnout jako měření fluorescenčního záření. Vzorek (plynná směs) je ozařován UV lampou, přitom dochází k energetické excitaci molekul SO_2. Při zpětném přechodu molekuly do základního stavu dochází k uvolnění energie ve formě fluorescenčního záření. Intenzita fluorescence, která se detekuje fotonásobičem je pak přímo úměrná koncentraci oxidu v měřící komůrce. Ve fluorescenční komůrce se molekuly oxidu obvykle budí zářením o vlnové délce 215 nm. Fluorescenční záření se snímá kolmo na budící, protože intenzita budícího záření je několikanásobně vyšší a docházelo by ke zkreslení. Následující obrázek 11. ukazuje absorbanci běžných plynů v závislosti na vlnové délce, vše v ultrafialové oblasti. Nevýhodou této detekce je snížení emise v důsledku zhášení fluorescence některými sloučeninami.

obr 11.
Absorbance běžných plynů



obr 12.
Obecné schéma s popisem fluorescenčního analyzátoru SO_2

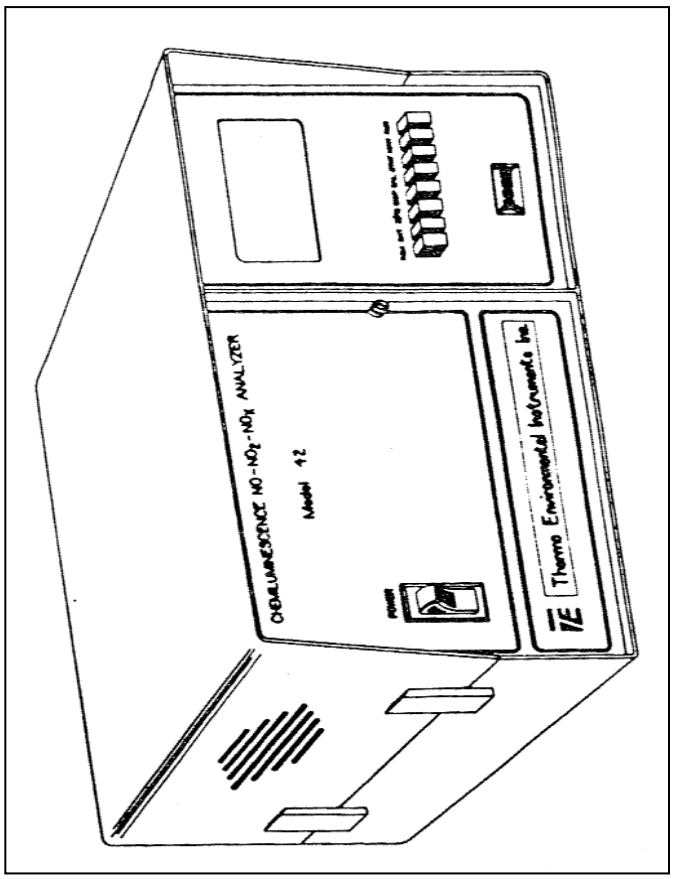


Obr. 50. Schéma fluorescenčního analyzátoru SO_2
1 - rtuťová výbojka, impulsně napájená, 2 - čočka, 3 - interferenční filtr, 4 - měřicí komora, 5 - fotonásobič, 6 - zesilovač

3.2.2 MĚŘENÍ NO_x

Měření koncentrace NO_x se provádí chemiluminiscenčním analyzátorem pro měření nižších koncentraci NO, NO_2 a NO_x. Princip metody je založen na excitaci molekul dusíku ozónem. Základem je chemická reakce oxidace oxidu dusnatého ozónem doprovázená emisí světelného záření. Jedná se o model 42 firmy Thermo Environmental Instruments (viz. obrázek 12.), je to přístroj druhé generace schopný stanovit oxidy dusíku v ovzduší o koncentraci od jednotek ppb až po 20 ppm. Model používá fotonásobič o malém průměru a jednu reakční komoru, které jsou časově sdíleny pro měření NO a NO_x . Rozdíl obou měření tak umožňuje generaci tří kontinuálních signálů pro NO, NO_2 tj. (NO_x-NO) a NO_x tj. (NO+NO_2).

obr 12.
Analyzátor NO_x model 42



Zmiňovaná reakce s ozónem se dá zapsat jako:

$NO + O_3 \rightarrow NO_2^* + O_2$

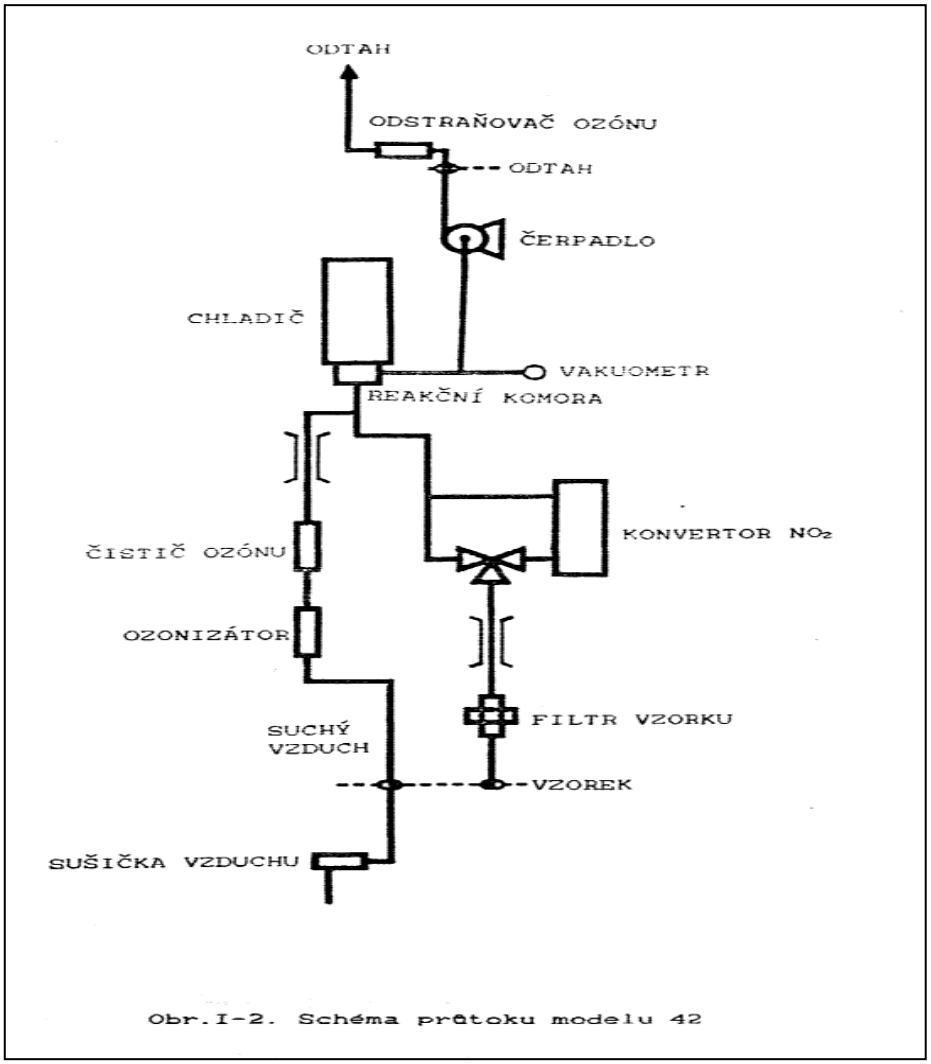
$NO_2^* \rightarrow NO_2 + hv$

Aby analyzátor mohl měřit koncentraci NO_x, musí umět měřit i oxid dusičitý, který se ovšem reakce s ozonem neúčastní, protože je jejím produktem. Aby se mohl NO_2 měřit, musí se převést před vstupem do reakční komory na NO. U modelu 42 se převod uskutečňuje v redukčním konvertoru na bázi Mo. V konvertoru vyhřívaném na 325^oC probíhá redukce podle schématu:

$3NO_2 + Mo \rightarrow 3NO + MoO_3$

Vzorek okolního vzduchu vstupuje do modelu 42 kontrolní kapilárou průtoku a je veden do solenoidového ventilu. Solenoidový ventil směruje vzorek buď do konvertoru NO_2 (tj. NOx režim) nebo mimo konvertor NO_2 (tj. NO režim). Při průchodu vzorku konvertorem představuje chemiluminiscence změřená v reakční komoře koncentraci NO_x. Při obejití konvertoru je možné měřit pouze koncentraci NO. Následující obrázek 13. celou situaci dokumentuje schématem.

obr 13.
Schéma průtoku plynu analyzátorem



Signály generované v těchto dvou režimech provozu jsou ukládány a uchovávány v paměti mikropočítače modelu 42. Jejich rozdíl je využíván k tvorbě signálu NO_2. Číslicově analogový převodník pak převádí tyto tři uložené hodnoty na analogové signály, a ty jsou vedeny na výstupy na zadním panelu přístroje.

Energie vyzařovaného fotonu odpovídá blízké infračervené oblasti mezi 600 až 2500 nm. Aby bylo dosaženo citlivosti v infračervené oblasti je použit multialkalický typ fotonásobiče. Aby byl snížen proud za tmy, je fotonásobič ochlazován přibližně na -3^{o}C. Vlastní reakční komora je zevnitř pozlacený termostatovaný kvádr. Změna teploty je hlavním rušivým elementem stanovení, a tak měření je prováděno při 50^{o}C. Ozonizátor dodává přibližně konstantní množství ozonu. Je napájen pulsy vysokého napětí, které způsobí tichý výboj ve vzduchu, jenž se tak obohacuje ozonem na 0,5% obj. při průtoku 120 cm^{3}/min. Důležitý je přebytek ozonu vůči stechiometrii, jeho mírný nadbytek nevadí.

V následující tabulce 8. jsou uvedeny technické parametry analyzátoru:

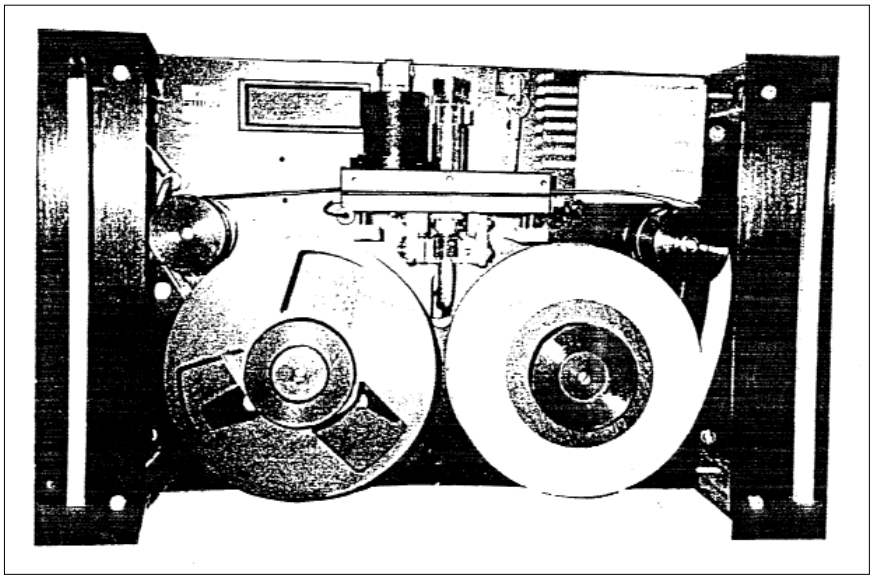
Tab. 8
Technické parametry analyzátoru

Mez stanovitelnosti	0,50 ppb
Přesnost	± 0,5 ppb
Linearita	± 1% z rozsahu
Průtok vzorku	0,65 l/min
Průtok ozonizovaného vzduchu	0,12 l/min
Provozní teplota	(5 – 40)^{o}C
Hmotnost	45 kg

3.2.3 MĚŘENÍ PRAŠNÉ FRAKCE PM_{10}

Měření koncentrace PM_{10} (fakce prašného aerosolu do 10 μm) se provádí radiometrickou metodou, která je založena na absorpci beta záření ve vzorku zachyceném na filtračním materiálu. *Z* rozdílu absorpce beta záření mezi exponovaným a neexponovaným filtračním materiálem, který je úměrný hmotnosti zachyceného prašného aerosolu je odvozen údaj o jeho koncentraci. Analyzátor je vyráběn firmou VEREWA Mess-und Regeltechnik GmbH, Mulheim/Ruhr, typové označení je „beta prachoměr“ F 703 V/R II (viz. obrázek 14.).

obr 14.
Betaprachoměr



Beta prachoměr VEREWA F 703 V/R II měří koncentraci prachu v jednotkách mg/m^3 vlhkého plynu. Pro stanovení koncentrace prachu je měřen objemový průtok plynu a je stanoveno množství prachu v něm obsažené. Množství prachu usazené na filtru ze skelné tkaniny, se stanovuje na základě zeslabení záření uhlíku ^{14}C, měřeného Geiger-Müllerovým čítačem. Radiometrická metoda měření je všeobecně použitelná, protože stanoví množství prachu v širokých mezních hodnotách bez ohledu na chemické a fyzikální vlastnosti prachu a nosného plynu. Jestliže usazený prach o množství M je homogenně rozložen na ploše filtru „a“, platí následující vztah přibližně až do *5* mg/cm^2:

$$\ln\left(\frac{n_8}{n}\right) = \left(\frac{\mu}{\rho}\right).d$$

μ – je lineární koeficient zeslabení použitého beta záření v cm^{-1}

ρ – je hustota absorpčního materiálu v g/cm^3

μ/ρ - je koeficient zeslabení záření hmotou, který je prakticky nezávislý na chemickém složení v cm^2/g

d = M/a - je hustota prachu na jednotku plochy filtru v mg/cm^2 (mg/cm^2) pro usazený prach M(mg) na konstantní ploše usazování „a“ (cm^2)

n_8 , n - jsou částice beta zaznamenané čítačem jako napěťové pulsy bez a nebo s hustotou „d“ za minutu. Hustota impulzů je mírou intenzity záření. Jednotka plochy filtru záleží na měnitelné sondě, běžně to je 2,54 cm^2.

Koeficient zeslabení hmotou je úměrný poměru:
(Atomové číslo Z) / (nukleonové číslo A)
Tento poměr je pro většinu prachů konstantní okolo 0,5, čili zeslabení beta záření nezávisí na chemickém složení. Pouze vodík, pro který je poměr roven jedné, výrazně ovlivňuje měření avšak nutno dodat, že vodík není běžnou součástí prachů. Tedy nemění-li se plocha filtru, můžeme množství prachu vypočíst z uvedené rovnice.

Filtr je vyroben ze skelných vláken o velikosti zrna 0,3 μm. V analyzátoru je ve formě kotouče o délce 45 m a šíři 45 mm, který se postupně odvíjí podle časového cyklu. Běžně se používá měření jednou za 30 minut, tj. filtr vystačí na 260 dní.

Jak jsem uvedl, plocha filtru závisí na použité sondě dosedající na filtr, přes kterou čerpadlo nasává plyn. Běžně se požívá jednotka 7HE s plochou 2,54 cm^2, která je schopna zadržet od 1,5 mg do max. 10 mg absolutně.

V následující tabulce 9. jsou uvedeny vybrané parametry analyzátoru.

Tab. 9
Vybrané parametry analyzátoru

Měřící rozsah [µg/m^3]	nastavitelný – interně jednorozsahový omezení množstvím navzorkovaného aerosolu v závislosti na době vzorkování a koncentraci min. navážka 1,5 mg; max. navážka 10 mg
Používané měř. rozsahy	500, 1000 µg/m^3
Mez detekce	30 µg tj. při 3 hod a 3 m^3/hod: 3,33 µg/m^3
Stabilita nuly	2 % měřícího rozsahu
Přesnost	2 % měřícího rozsahu

3.3 VÝPOČETNÍ PROSTŘEDÍ MATLAB

MATLAB je interaktivní systém pro vědecké a technické výpočty založený na maticovém počtu. Umožňuje řešit velkou oblast numerických problémů, aniž bychom museli programovat vlastní program. Název matlab vznikl zkrácením **MATrix LABoratory**.
Výpočetní systém MATLAB se během uplynulých let stal celosvětovým standardem v oblasti technických výpočtů a simulací nejen ve sféře vědy, výzkumu a průmyslu, ale i v oblasti vzdělávání. Nejnovější verzí je verze MATLAB 6. Výrobcem je americká firma **the MathWorks, Inc**. Tento "chytrý maticový software" pracuje v podstatě pouze s jedním typem dat a tím je obdélníková matice s reálnými nebo komplexními prvky. Skaláry potom vyjadřujeme jako matice rozměru (1 x 1) a vektory jako matice s jedním řádkem (1 x n) nebo jedním sloupcem (n x 1).

Vlastností, která patrně nejvíce přispěla k rozšíření MATLABu, je jeho otevřená architektura. Uživatel se může vytvořit funkce podle sebe, tyto funkce jsou normálně zařazeny mezi knihovní funkce. Otevřená architektura MATLABu vedla ke vzniku knihoven funkcí, nazývaných ***toolboxy***, které rozšiřují použití programu v příslušných vědních a technických oborech. Toolbox je soubor nadefinovaných funkcí, které jsou určeny pro práci a výpočty v určité vědní disciplíně. Např. toolbox řekněme pro „neuronové sítě“ obsahuje funkce, které vykonají se vstupními parametry složité operace s jednoduchým pro uživatele důležitým výsledkem. Pokud bychom tyto operace chtěli naprogramovat v jazyce nižší úrovně, museli bychom vynaložit mnohem více úsilí a každou, i základní operaci, algoritmizovat. Tyto knihovny, navržené a v jazyce MATLABu napsané nejvýznačnějšími světovými odborníky, nabízejí předzpracované specializované funkce tzv. „m-files s příponou m“, které je možno rozšiřovat vlastními funkcemi a předefinované funkce modifikovat.

3.4 VLASTNÍ ALGORITMUS FCM

Jak bylo uvedeno, úloha minimalizace funkcionálu je iterativní postup. Napsat celý algoritmus tak, aby po jednotlivých iteracích dával výstup není v klasických programovacích jazycích až tak jednoduché. V každé iteraci se mění matice funkcí příslušnosti a zároveň matice centroidů. Nejlépe je ukázat celý algoritmus rozepsaný postupně do dvou iterací. Předpokládejme, že máme veškerá data převedená do vstupní datové matice, tj. matice rozměru (n *x* c) = (počet shlukovaných bodů *x* počet clusterů) popř. transponovaná (c *x* n). Řekněme, že se jedná o matici dat z odběrového místa „Měděnec" v okrese Chomutov za měsíc únor. Sloupce matice jsou měřené veličiny tedy SO_2, NO_x a PM_{10}. Data byla předem upravena aby byla databáze kompaktní, některé dny nefungovala stanice na měření SO_2 a jiné zase na PM_{10}, proto výsledný počet objektů u kterých znám koncentrace všech tří polutantů nemusí být roven počtu dnů v měsíci. Dále předpokládejme (z analýzy validity clustrů), že jako nejvhodnější počet clusterů je 3. Taková matice má rozměr 24 *x* 3.

PRVNÍ ITERACE

Prvním krokem je matice náhodných funkcí příslušnosti „C" rozměru datové matice. V matlabu je jednoduchý příkaz „rand", který generuje čísla v intervalu (0,1) s rovnoměrným rozdělením. Dále se „C" transponuje, aby se s ní lépe pracovalo tedy C (3 *x* 24).

```
C=rand(24,3)
C=C´
```

Podle teorie musí být suma funkcí příslušnosti objektu přes clustry rovna jedné, proto musíme „C" upravit. Sečteme funkce příslušnosti objektu přes všechny clustery, to celé pro všechny objekty. Dostaneme vektor s(24) o 24 prvcích. Příkaz sum(C(:,i)) je suma přes první rozměr „:" matice C, tj. všechny řádky, pro i-tý sloupec. Chceme-li násobit prvek matice A s prvkem B, to celé pro obě matice stejného rozměru zároveň, použije se příkaz „A .* B". Úpravu matice C si lze představit jako:

$$\frac{X}{X+Y+Z}+\frac{Y}{X+Y+Z}+\frac{Z}{X+Y+Z}=\frac{X+Y+Z}{X+Y+Z}=1$$

Výsledkem je upravená matice „U".

```
for i=1:24
s(i)=sum(C(:,i))
end;
s1=s.^-1
a=ones(3,1)*s1
U=C'
U=U.*a'
```

Dále se postupuje podle vzorce pro výpočet centroidu pro i-tý cluster. V mém případě je takový centroid vektorem 3 souřadnic, každá pro SO_2, NO_x a PM_{10}. Cílem je dostat matici centroidů pro všechny clustery. Volitelný parametr „m" se často volí 2 (viz. teoretická část).

$$\bar{v}_i = \frac{\sum_{k=1}^{N} (u_{ik})^m \bar{x}_k}{\sum_{k=1}^{N} (u_{ik})^m}$$

V proměnné medenecunor je vstupní matice dat, e1, e2, e3 jsou pomocné proměnné rozměru (24 *x* 3…..3 za proměnné) postupně pro 1., 2. a 3. cluster.

```
e1=medenecunor.*(ones(3,1)*U(:,1).^2')'
e2=medenecunor.*(ones(3,1)*U(:,2).^2')'
e3=medenecunor.*(ones(3,1)*U(:,3).^2')'
```

Vektor v11 je rozměru (1 *x* 3). Sum(U(:,1).^2) ve jmenovateli je podle vzorce $\sum_{k=1}^{N} (u_{ik})^m$, sum(e1(:, **j**)) je podle $\sum_{k=1}^{N} (u_{ik})^m x_{kj}$ pro **j** rovno postupně 1,2 a 3 podle SO_2, NO_x a PM_{10}. Výsledkem je matice „v“ mající v prvním řádku vektor v11, ve druhém v21 a analogicky v31, tedy např. v21 je centroid druhého clusteru a v první iteraci.

```
v11=[ sum(e1(:,1)) sum(e1(:,2)) sum(e1(:,3)) ]/sum(U(:,1).^2)
v21=[ sum(e2(:,1)) sum(e2(:,2)) sum(e2(:,3)) ]/sum(U(:,2).^2)
v31=[ sum(e3(:,1)) sum(e3(:,2)) sum(e3(:,3)) ]/sum(U(:,3).^2)
v=[v11; v21; v31]
```

Nyní vypočítáme matici nových funkcí příslušností podle vzorce:

$$u_{ik} = \left(\sum_{j=1}^{C} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}} \right)^{-1}$$ i ≠ j….jsou clustery, k…objekt

d1,d2,d3 jsou matice (24 x 3….SO_2,NO_x,PM_{10}), kde d1 představuje 24 diferencí mezi koncentracemi 3 polutantů daného objektu a prvního clustru postupně, analogicky platí i d2,d3.

```
d1=medenecunor-ones(24,1)*v11
d2=medenecunor-ones(24,1)*v21
d3=medenecunor-ones(24,1)*v31
```

Příkaz diag(*matice*) vrací sloupcový vektor diagonálních prvků. Diagonální prvky matice ($d1*d1^t$) tvoří vektor 24 prvků (za každý objekt) sum čtverců **odchylek** koncentrací polutantů od **prvního** (d**1**) clusteru, tj.

24 *x* **(delta**^{2}c(SO_2)+**delta**^{2}c(NO_X)+**delta**^{2}c(PM_{10}) **)**

```
k11=diag(d1*d1')'
k21=diag(d2*d2')'
k31=diag(d3*d3')'
```

U1 je matice funkcí příslušnosti po první iteraci. C**1**, C**2**, C**3** jsou sloupcové vektory

funkcí příslušnosti **1. 2.** a **3.** clusteru.

(k11.*e) v předpise pro C1,2,3 je vlastně $\sum_{j=1}^{C}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}$

e(24) je řádkový vektor 24 prvků, kde každý prvek je převrácená suma čtverců vzdáleností daného objektu od centrů jednotlivých clusterů. Vzdálenost k-tého bodu od prvního clusteru je v podstatě:

$$d_{1k} = \sqrt{\mathrm{delta}^2\mathrm{c}(\mathrm{SO}_2) + \mathrm{delta}^2\mathrm{c}(\mathrm{NO}_\mathrm{X}) + \mathrm{delta}^2\mathrm{c}(\mathrm{PM}_{10})}$$

čtverec vzdálenosti potom je:

$$(d_{1k})^2 = \mathrm{delta}^2\mathrm{c}(\mathrm{SO}_2) + \mathrm{delta}^2\mathrm{c}(\mathrm{NO}_\mathrm{X}) + \mathrm{delta}^2\mathrm{c}(\mathrm{PM}_{10})$$

Uvážíme-li, že parametr „m“ jsme zvolili 2, potom výraz **1/k11(i)+1/k21(i)+1/k31(i)** (**i**-je objekt) představuje:

$$\sum_{j=1}^{C}\left(\frac{1}{d_{ji}}\right)^{\frac{2}{m-1}}$$

Součin **(1/k11(i)+1/k21(i)+1/k31(i))*k21(i)** pro nějaké dané „i“, třeba i=5, je:

$$\sum_{j=1}^{C}\left(\frac{d_{25}}{d_{j5}}\right)^{\frac{2}{m-1}}$$

Pokud budou výsledky součinů **(1/k11(i)+1/k21(i)+1/k31(i))*k21(i)** pro zvyšující se „i“ od 1 do 24 zapsány pod sebe, dostane se sloupcový vektor, který je **druhým** sloupcem matice U1 (24 *x* 3).

```
for i=1:24
e(i)=1/k11(i)+1/k21(i)+1/k31(i)
end
c1=((k11.*e).^(-1))'
c2=((k21.*e).^(-1))'
c3=((k31.*e).^(-1))'
U1=[c1 c2 c3]
```

Následně mohu podle teoretické části vypočíst koeficient rozkladu v první iteraci Fc1. Příkaz „trace(matice)“ počítá stopu (matice).

```
SU1=(U1'*U1)/24
Fc1=trace(SU1)
```

Entropie rozkladu se počítá podle:

$$H(U,c) = -\sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}\log_a(u_{ik})/N$$

člen a1 odpovídá sumě přes objekty, pro první cluster. V matlabu je log přirozeným

logaritmem.

$$\sum_{k=1}^{N} u_{1k} \ln(u_{1k})$$

H1 je potom sumou přes všechny clustery.

```
a1=sum(U1(:,1).*log(U1(:,1)))
a2=sum(U1(:,2).*log(U1(:,2)))
a3=sum(U1(:,3).*log(U1(:,3)))
H1=-(a1+a2+a3)/24
```

V teoretické části bylo ukázáno, jak krom indexů validity clusterů je možné za měřítko jakosti rozkladu brát i vlastní hodnotu funkcionálu rozkladu.
A1 je suma přes objekty. Výsledkem je hodnota funkcionálu J**1** po **první** iteraci.

$$\sum_{k=1}^{N} (u_{1k})^m \left\| x_k - v_i \right\|^2$$

```
A1=sum(k11'.*U1(:,1).^2)
A2=sum(k21'.*U1(:,2).^2)
A3=sum(k31'.*U1(:,3).^2)
J1=A1+A2+A3
```

Takto se dá zapsat první iterace algoritmu FCM. Následně v druhé iteraci je použito stejného značení až na některé proměnné, které se v každé iteraci přepisují.

DRUHÁ ITERACE

```
e1=medenecunor.*(ones(3,1)*U1(:,1).^2')'
e2=medenecunor.*(ones(3,1)*U1(:,2).^2')'
e3=medenecunor.*(ones(3,1)*U1(:,3).^2')'
v12=[ sum(e1(:,1)) sum(e1(:,2)) sum(e1(:,3)) ]/sum(U(:,1).^2)
v22=[ sum(e2(:,1)) sum(e2(:,2)) sum(e2(:,3)) ]/sum(U(:,2).^2)
v32=[ sum(e3(:,1)) sum(e3(:,2)) sum(e3(:,3)) ]/sum(U(:,3).^2)
v=[v12; v22; v33]
d1=medenecunor-ones(24,1)*v12
d2=medenecunor-ones(24,1)*v22
d3=medenecunor-ones(24,1)*v32
k12=diag(d1*d1')'
k22=diag(d2*d2')'
k32=diag(d3*d3')'
for i=1:24
e(i)=1/k12(i)+1/k22(i)+1/k32(i)
end
c1=((k12.*e).^(-1))'
c2=((k22.*e).^(-1))'
c3=((k32.*e).^(-1))'
U2=[c1 c2 c3]
SU2=(U2'*U2)/24
Fc2=trace(SU2)
a1=sum(U2(:,1).*log(U2(:,1)))
a2=sum(U2(:,2).*log(U2(:,2)))
a3=sum(U2(:,3).*log(U2(:,3)))
H2=-(a1+a2+a3)/24
```

Pro druhou iteraci lze ukázat výpočet indexu validity Xie-Beni. Hodnota funkcionálu

J2 pro výpočet indexu Xie-Beni je již vypočítaná.

```
A1=sum(k11'.*U1(:,1).^2)
A2=sum(k21'.*U1(:,2).^2)
A3=sum(k31'.*U1(:,3).^2)
J2=A1+A2+A3
```

Podle vzorce

$$v_{XB}(U,V:X)=\frac{J_m(U,V:X)}{n\left(\min\limits_{i\neq j}\left\{\left\|v_i-v_j\right\|^2\right\}\right)}$$

se musí určit minimum ze čtverce vzdálenosti, v našem případě ze tří čísel.

```
V12=(norm(v12-v22))^2
V22=(norm(v12-v32))^2
V32=(norm(v22-v32))^2
```

Určí se jako nejmenší např. V32, tedy index Xie-Beni potom je:

```
vxb=J1/(24*V32)
```

Výše uvedené výpočty byly uvedeny jako ukázka algoritmu FCM po iteracích. Samozřejmě, že se to jako postup výpočtu FCM nedá použít, psát (programovat) každou iteraci je neefektivní. V centru aplikované kybernetiky při FEL, ČVUT Praha vznikl program, který řeší FCM nástroji systému matlab a umožňuje tak získat výsledky mnohem efektivněji a snáze. Součástí programu jsou i grafické funkce prezentující výsledky.

3.4.1 POPIS ZDROJOVÉHO PROGRAMU FCM, MATLAB VER. 6

Slovní popis řádků zdrojového textu velice usnadňuje orientaci a celkové pochopení programu. Z tohoto důvodu jsou dále popsány nejdůležitější příkazy matlabu, které byly použity v programu FCM. Program FCM sestává z několika funkcí, z nichž nejdůležitější jsou tyto: *cluster_graf.m, cluster_color3.m, barvicky.m* a *fuzzyCMA1.m.*

Vstupní datová matice má jeden z rozměrů roven počtu znaků, tyto se někdy nazývají dimenze. Častokrát je počet dimenzí větší než tři a v takovém případě již nelze jednoduše graficky prezentovat výsledky jako zobrazení ukazující seskupení objektů matice na základě clusterové analýzy. V případě imisního monitoringu byly sledovány pouze tři proměnné, tři dimenze SO_2, NO_x a PM_{10}. Součástí zdrojového programu jsou i funkce pro grafickou prezentaci výsledků v 3D prostoru. Funkce *cluster_graf.m, cluster_color3.m* umožňují zobrazit jak 2D tak i 3D grafy, v grafu jsou body patřící k příslušným clusterům odlišeny barevně. Funkce *barvicky.m* slouží k identifikaci clusterů a současně tak i k identifikaci jednotlivých bodů v clusteru.

Nejdůležitější funkcí provádějící vlastní FCM je *fuzzyCMA1.m*.

fuzzyCMA1.m – funkce fuzzy C-means

Funkce obecně má vstup a výstup. Proměnné definované ve funkci jsou lokální, po opuštění funkce jsou zapomenuty. V běžných programovacích jazycích je problém, aby funkce vracela více hodnot. V matlabu to sice jde, ale kvůli přehlednosti je lepší používat strukturované proměnné. Strukturovaná proměnná je v matlabu označena jejím obecným názvem následovaným tečkou, která specifikuje jednotlivé proměnné dále. Funkce *fuzzyCMA1.m* má jako vstup i výstup strukturovanou proměnnou. Její plný název s parametry je:

function [output]=fuzzyCMA1(input)

Input je strukturovaná proměnná. Má několik hodnot, nejdůležitější jsou:

input.data	vstupní matice dat v transponované poloze, tj. rozměru (dimenze *x* objekty
input.n_clust	apriorní volba clusterů
input.expon	váhový koeficient
input.steps	počet iterací, z formálního hlediska je lepší mít za „stoping“ kritérium objektivní hodnotu buďto funkcionálu nebo normu rozdílu matic příslušnosti ve dvou po sobě jdoucích iteracích. Při výpočtech bylo jako stopovací kritérium vzata norma matic příslušností. Jako hodnotu hraniční hodnota byla zvolena hodnotu 0,01. Jakmile hodnota normy klesne pod 0,01, výpočet se zastaví.

Output je strukturovaná proměnná výstupu funkce. Opět má několik hodnot, nejdůležitější jsou:

output.Fc;
Vektor hodnot koeficientu rozkladu po iteracích.

output.Hc;
Vektor entropie rozkladu po iteracích.

output.U;
Pole buněk matic funkcí příslušnosti po iteracích. Pole buněk je zde vlastně vektor objektů, kde každý objekt je matice.

output.C;
Pole buněk matic centroidů po iteracích. Pole buněk je zde vlastně vektor objektů, kde každý objekt je matice.

output.Obj;
Vektor hodnot funkcionálu po iteracích.

output.XB;
Vektor hodnot indexu validity Xie-Beni po iteracích.

output.FS;
Vektor hodnot indexu validity Fukuyama-Sugeno po iteracích.

output.stoping
Vektor hodnot „stoping“ kriteria po iteracích.

output.Hc_norm1
Vektor hodnot normalizované entropie rozkladu po iteracích.

output.Hc_norm2
Vektor hodnot normalizované entropie (její druhý tvar-viz teoretická část) po iteracích

output.Hc_stand
Vektor hodnot standardizované entropie rozkladu po iteracích.

output.Fc_norm
Vektor hodnot normalizovaného koeficientu rozkladu po iteracích.

output.Fc_stand
Vektor hodnot standardizovaného koeficientu rozkladu po iteracích.

Ihned po definici vnější funkce fuzzyCMA1 následuje předání dat ze vstupu.

```
data     =  input.data;
n_clust  =  input.n_clust;
```

Následuje volání funkce *fcma,* která je vnořenou funkcí funkce *fuzzyCMA1* a provádí iterativní výpočet. Funkce má pochopitelně stejný vstup jako *fuzzyCMA1* a výstup také. Za voláním funkce s parametry následuje vytvoření strukturované proměnné výstupu.

```
[U,C,Obj,mess,Fc,Hc,XB,FS,stoping,Hc_norm1,Hc_norm2,Fc_norm,Hc_stand,Fc_
stand]=fcma(data, n_clust, expon, steps, stop, Anorm, c_in, zerod)
```

```
output.Fc=Fc;
output.Hc=Hc;
output.XB=XB;
output.FS=FS;
output.U=U;
output.C=C;
output.Obj=Obj;
output.norm=Anorm;
output.stoping=stoping;
output.Hc_norm1=Hc_norm1;
output.Hc_norm2=Hc_norm2;
output.Hc_stand=Hc_stand;
output.Fc_norm=Fc_norm;
output.Fc_stand=Fc_stand;
```

Pak následuje definice funkce fcma se zdrojovým kódem funkce.

```
Function     [U,C,Obj,mess,Fc,Hc,XB,FS,stoping,Hc_norm1,Hc_norm2,Fc_norm,
Hc_stand, Fc_stand]=fcma(data, n_clust, expon, steps, stop, A, c_in, zerod)
```

Někdy vhodné sledovat jak se mění matice „U“ a „C“ po iteracích, proto jsou definovány jako „cell array“ o předepsaném počtu buněk. Při zpracování dat výpočet vždy skončil do 100 iterací, tak stačí nadefinovat počet buněk 200, což je implicitně daná proměnná „steps“.

```
C=cell(1,steps);
U=cell(1,steps);
```

Ze vstupní matice se zjistí její rozměr.

```
PocetDat=size(data,2);
Dimenze=size(data,1);
```

Dále se zvolí náhodná matice funkcí příslušnosti tak, že součet pro všechny clustery je roven jedné. Velmi důležitým příkazem, který zjednodušuje celý výpočet FCM vůbec je „repmat“ který opakuje matici podle zadaných rozměrů, blíže Help Matlab. Dist je jednotková matice zadaného rozměru, která představuje matici vzdálenosti.

```
Umat=rand(PocetDat,n_clust);
Umat=Umat./repmat(sum(Umat,2),1,n_clust);
Dist=ones(PocetDat,n_clust);
```

Následuje „for“ cyklus po iteracích do maximálního počtu 200 iterací. V něm se provádí výpočet centroidů a matic příslušnosti.

```
for iterace=1:steps

Cmat=data*(Umat.^expon)./repmat(sum(Umat.^expon,1),Dimenze,1);
C{1,iterace}=Cmat;
```

„C“ je pole buněk „cell array“ matic centroidů po iteracích.

```
U{1,iterace}=Umat;

for cluster=1:n_clust

Dist(:,cluster)=sum( (data-repmat(Cmat(:,cluster),1,PocetDat))' *A .* (data-
repmat(Cmat(:,cluster),1,PocetDat))' ,2);

end
```

Výsledkem je matice čtverců vzdáleností rozměru (PocetDat *x* n_clust), „A“ je matice identity s jedničkami na diagonále. Volba matice „A“ umožňuje počítat FCM variantu Gustafsson-Kessel pro eliptické clustery.

```
Dist=Dist.^0.5;
```

Matici čtverců vzdáleností je nutno pro další výpočty odmocnit.

```
[idato_nula,jshluk_nula]=find(Dist<=zerod);
```

Najde souřadnice nulových prvků v matici „Dist“.

```
idato_nula_os=unique(idato_nula);
```

Odstraní duplicitní prvky.

```
idato_spocti=setdiff([1:PocetDat]',idato_nula_os);
```

Určí objekty matice dat, které nejsou současně centroidem.

```
nulove_prvky=find(Dist<=zerod);
```

Najde prvky matice „Dist“, které jsou nulové.

```
Umat(idato_spocti,:)=1./( Dist(idato_spocti,:).^(2/(expon-1)) .* repmat(
sum( 1./(Dist(idato_spocti,:) .^ (2/(expon-1)) ),2) ,1,n_clust) );
```

Výsledkem je matice „Umat“ funkcí příslušnosti pro danou iteraci rozměru (idato_spocti *x* n_clust), kde „idato_spocti“ označuje nenulové řádky, tj. objekty, které nejsou zároveň centroidem. Je-li objekt centroidem, jemu přiřazena hodnota příslušnosti rovna jedné a

```
Umat(idato_nula_os,:) = 0;
Umat(nulove_prvky) = 1;
Umat(idato_nula_os,:) = Umat(idato_nula_os,:) ./
repmat(sum(Umat(idato_nula_os,:),2),1,n_clust);
```

Dopočítá funkce příslušnosti u „problémových objektů“, které byly sami centroidy.

```
Obj(iterace)= sum(sum((Umat.^expon) .* Dist.^expon ));
```

Výpočet hodnoty funkcionálu (objective-function).

```
for f=1:(n_clust-1)
for sloupec=(f+1):n_clust
pomoc(sloupec-f)=(norm(Cmat(:,f)-Cmat(:,sloupec)))^2;
end
norma(f)=min(pomoc);
end
```

Pomocná proměnná pro určení Xie-Beni.

```
ahoj=sum(Umat.^expon);
for g=1:n_clust
soucet(g)=ahoj(g)*((norm(Cmat(:,g)-mean(data')'))^2);
end
celkovysoucet=sum(soucet);
```

Pomocná proměnná pro určení Fukuyama-Sugeno.

```
tebuh=0;
nazdar=0;
for g=2:n_clust
tebuh=tebuh+1/g^2;
nazdar=nazdar+1/g;
end
```

Pomocné proměnné pro určení standardizovaných koeficientů rozkladu a entropie.

```
XB(iterace)=Obj(iterace)/(PocetDat*nejmensi);
Fc(iterace)=sum(sum(Umat.^2))/size(data,2);
Hc(iterace)=-sum(sum(Umat.*log(Umat)))/size(data,2);
```

```
FS(iterace)=Obj(iterace)-celkovysoucet;
Hc_norm1(iterace)=Hc(iterace)/log(n_clust);
Hc_norm2(iterace)=PocetDat*Hc(iterace)/(PocetDat-n_clust);
Fc_norm(iterace)=(n_clust/(n_clust-1))*(1-Fc(iterace));
Hc_stand(iterace)=(Hc(iterace)-nazdar)/((((1/PocetDat)*tebuh-(n_clust-
1)/(n_clust+1))*(pi^2-6)/(6*PocetDat))^0.5);
Fc_stand(iterace)=((PocetDat*(n_clust+2)*(n_clust+3)/(n_clust-
1))^0.5)*(((n_clust+1)*Fc(iterace)/2)-1);

if iterace > 1
stoping(iterace)=norm(matice{1,iterace}-matice{1,iterace-1});
if (stoping(iterace) < 0.001)
iterace
break;
end
end
```

Výpočet indexů validity a vyhodnocení „stoping“ kriteria.

```
end
```

Konec celého „for“ cyklu po iteraci.

cluster_graf.m-funkce grafického výstupu

Definice funkce grafického výstupu. Jako vstupní parametry má „in, out, IDdata´ “, což je vstup, výstup fuzzyCMA1 a IDdata je stejná matice jako vstupní datová matice, navíc má ovšem 2 sloupce „měsíc a den“ kvůli identifikaci objektů po clusterování (viz dále).

```
function h=cluster_graf(in,out,IDdata)
```

Definice globálních proměnných na identifikaci clusterovaných objektů.

```
global pp;
global id;
```

Stanovení dimenzí, (sloupců vstupní matice), které chci vykreslit.

```
dimenze=input('Zadej dimenze v kterych chces vykreslit vysledek (napr. [2 3
5])   ');
```

Pro můj případ 3D dat, funkce volá další funkci *cluster_color3* pro práci ve 3D prostoru.

```
if size(dimenze,2)==3
h=cluster_color3(out.U,out.C(dimenze,:),in.data(dimenze,:),IDdata);
end
```

cluster_color3-funkce pro práci ve 3D prostoru

Jako vstupní parametry má již známé parametry, záleží na jejím volání z funkce

Cluster_graf.m.

function h=cluster_color3(U,C,data,IDdata);

Určení clusterů, ke kterým má objekt největší funkci příslušnosti, „I“ je vektor čísel představující clustery ve kterých má příslušný objekt největší příslušnost.

[X,I]=max(U');

PP{ii} a id{ii} jsou pole buněk obsahující množiny objeků patřících do “ii-tého” clusteru, id{ii} navíc od pp{ii} obsahuje identifikační sloupce.

```
for ii=1:size(C,2)
pp{ii}=data(:,find(I==ii));
id{ii}=IDdata(:,find(I==ii));
end;
```

Funkce končí Matlabovskými funkcemi, které podle parametrů vykreslí požadovaný graf.

```
str='plot3(';
strC='plot3(';
for ii=1:size(C,2)
str=sprintf('%spp{%d}(1,:),pp{%d}(2,:),pp{%d}(3,:),''.'',',str,ii,ii,ii);
strC=sprintf('%sC(1,%d),C(2,%d),C(3,%d),''*'',',strC,ii,ii,ii);
end;
str =sprintf('%s,''MarkerSize'',5);',str(1:end-1));
strC=sprintf('%s,''MarkerSize'',20);',strC(1:end-1));

h=figure;
eval(str);
hold on;
eval(strC);
grid on;
```

3.5 IMISNÍ DATABÁZE

Český hydrometeorologický ústav zajišťuje sběr imisních dat z území celé republiky. V průběhu roku 1995, v říjnu, došlo ke změně měření prašných imisí. Místo SPM coby prachu bez rozlišení frakcí se měří pro člověka nejnebezpečnější frakce PM_{10}. V práci jsou zpracována data z imisního monitoringu 1997. Všechny měřicí stanice jsou plně automatické AIM-stanice. Stanice sama měří a předává naměřená data. Občas se vyskytne porucha na stanici. Příkladem je neměnnost naměřených dat. Do souhrnné databáze je stanice zařazena, pokud vypočtená imisní charakteristika splňuje následující kritéria:

Pro 30 minutové data (tj. každých 30 minut jedno měření) a počítaný průměr za 24 hod., musí být doba nejdelšího souvislého výpadku dat 8 půlhodin a minimální počet naměřených dat 24 půlhodin.

Naskýtá se jeden problém a to opakování dat popř. nesouvislý výpadek. Jestliže je několik hodnot za sebou stejných, je velká pravděpodobnost, že stanice nefunguje dobře. To samé platí, jedná-li se o nesouvislý výpadek nepřekračující hranici 8 měření. Problém lze částečně odstranit úpravou databáze (viz. příloha).

MĚDĚNEC A CHOMUTOV

V oddíle 3.1 je uveden přehled stanic imisního monitoringu v Severočeském kraji. Ke každé stanici byla zpracovávána její databáze za rok 1997. V průběhu zpracování se ukázalo, že zbrát v úvahu všechny tyto stanice je zbytečné a zmatečné. Z těchto důvodů byly vybrány dvě stanice ve stejném okrese Chomutov a to stanice s názvem Měděnec a Chomutov. Stanice Měděnec leží v nadmořské výšce 827 m. Stanice chomutov leží v nadmořské výšce 344 m. Bližší specifikace obou míst je snadno dostupná v informačních stanicích ČHMÚ.

Z databáze tabelárního přehledu ČHMÚ byla vybrána databáze již spočítaných průměrů. Dále byly dotazem vybrány ty datumy, kdy odběrové stanice poskytovaly data o všech třech sledovaných polutantech SO_2, NO_x, PM_{10}. Vzniklá databáze měla několik sloupců z nichž nejdůležitější jsou koncentrace polutantů v $\mu g.m^{-3}$ a $\mu g.cm^{-2}$, den, měsíc a jméno stanice.

Tab. 10
Vzor zpracované databáze

SO2	NOX	PM10	DEN	MESIC	ROK	STA_NAZ
218.743	60.922	91.765	1	1	1997	Měděnec
207.595	80.047	131.080	2	1	1997	Měděnec
96.980	46.003	22.158	3	1	1997	Měděnec
107.145	50.583	54.375	4	1	1997	Měděnec
49.465	31.615	79.333	5	1	1997	Měděnec
246.253	79.375	35.721	6	1	1997	Měděnec

Prvek, který měl vliv na clusterování je čas. Clusterová analýza hledá časovou podobnost v datech. U takto rozsáhlých databází je obtížné rozhodnout, co se má clusterovat. Je zřejmé, že jedním z největších vlivů na koncentraci je roční období.

Clusterovat data po měsíci nemá příliš cenu. Soubor je malý a rozptyl dat přebije časové trendy, které nemohou v měsíci vyniknout. Po několika zkouškách byl vybrán soubor po čtvrtletí začínající od ledna. Data byla upravena jako soubory 1 až 4 pro stanici, kde 1 značí data za leden, únor a březen. Výsledkem byly soubory (v Matlabu proměnné):

medenec1_97
medenec2_97
medenec3_97
medenec4_97
chomutov1_97
chomutov2_97
chomutov3_97
chomutov4_97

Toto jsou názvy proměnných obsahující data, s kterými se prováděla analýza. Z tabulky Ms-Excel se data převedly do matlabu pomocí příkazu

```
kanal=ddeinit('excel','nazev souboru.xls');
medenec2_97=ddereq(kanal,'r92c1:r177c3');
```

Po provedení clusterové analýzy je potřeba identifikovat objekty, které k sobě patří. Proto byly ještě vytvořeny proměnné např. **IDmedenec2_97**. Tyto proměnné obsahují ty samé 3 sloupce polutantů jako **medenec2_97** navíc s dvěma sloupci za den a měsíc. Při převodu z "Excelu" stačilo načíst o dva sloupce dat více.

```
medenec2_97=ddereq(kanal,'r92c1:r177c5');
```

Pro každou proměnnou byla sledována závislost 10 charakteristik s cílem určit optimální počet clusterů. Při všech výpočtech byl použit váhový koeficient m = 2. Počet clusterů obecně není libovolné číslo. Běžně se uvádí jako číslo v rozmezí 2 až $\sqrt{N}$. Tři měsíce představují cca 90 dní, (v průměru méně vzhledem k homogennitě databáze), takže jako maximální počet clusterů bohatě stačí 8. Jako kritérium ukončení vpočtu byla zvolena norma rozdílu matic příslušnosti v po sobě jdoucích iteracích. Hodnota kritéria byla zvolena 0,01.

Sledovaných 10 charakteristik bylo:

XB	koeficient validity Xie-Beni
Fc	koeficient rozkladu
Hc	entropie rozkladu
FS	koeficient validity Fukuyama-Sugeno
Hc_norm1	normalizovaná entropie podle (2.2.5.3.1a)
Hc_norm2	normalizovaná entropie podle (2.2.5.3.1b)
Fc_norm	normalizovaný koeficient rozkladu
Hc_stand	standardizovaná entropie rozkladu
Fc_stand	standardizovaný koeficient rozkladu
Obj	funkcionál kvality

TABULKY CHARAKTERISTIK

V následující tabulkách je uveden přehled charakteristik pro jednotlivé proměnné. Každá tabulka je označena názvem podle proměnné. Všechny mají maximální počet clusterů 8 a ukazují ty samé charakteristiky.

MĚDĚNEC1_97

Počet clusterů	2	3	4	5	6	7	8
XB	0,0442	0,1683	0,1587	0,1082	0,343	0,4042	0,4916
Fc	0,9076	0,8134	0,7691	0,7401	0,6813	0,6689	0,6321
Hc	0,1637	0,3414	0,4629	0,5336	0,6451	0,6848	0,7685
FS	-2,44E+05	-3,36E+05	-3,49E+05	-3,44E+05	-3,47E+05	-3,43E+05	-3,45E+05
Hc_norm1	0,2362	0,2362	0,3108	0,3339	0,3316	0,36	0,3519
Hc_norm2	0,1674	0,1674	0,3532	0,4844	0,565	0,6911	0,7425
Fc_norm	0,1849	0,1849	0,2799	0,3079	0,3249	0,3824	0,3863
Hc_stand	-17,0476	-17,0476	-23,7385	-30,7458	-38,769	-43,5671	-51,3962
Fc_stand	15,3304	15,3304	23,03	32,7533	43,3179	49,8477	61,5589
Obj	1,75E+05	9,28E+04	6,59E+04	4,82E+04	3,73E+04	3,20E+04	3,20E+04

MĚDĚNEC2_97

Počet clusterů	2	3	4	5	6	7	8
XB	0,0912	0,1505	0,2586	0,2119	0,2325	0,5503	0,5404
Fc	0,8608	0,7583	0,7152	0,6868	0,6353	0,5714	0,5684
Hc	0,2394	0,4386	0,5506	0,6312	0,7514	0,888	0,9163
FS	-2,11E+04	-4,75E+04	-5,49E+04	-5,46E+04	-6,16E+04	-6,12E+04	-6,69E+04
Hc_norm1	0,3454	0,3454	0,3989	0,3971	0,3922	0,4194	0,4563
Hc_norm2	0,2451	0,2451	0,4541	0,5774	0,6702	0,8078	0,9667
Fc_norm	0,2783	0,2783	0,3626	0,3878	0,3915	0,4377	0,5
Hc_stand	-12,913	-12,913	-18,6381	-25,8076	-32,9633	-36,9617	-38,9963
Fc_stand	12,0797	12,0797	18,5512	27,3425	36,7927	43,0527	46,178
Obj	3,71E+04	2,03E+04	1,34E+04	9,95E+03	8,04E+03	6,62E+03	5,29E+03

MĚDĚNEC3_97

Počet clusterů	2	3	4	5	6	7	8
XB	0,1166	0,1449	0,3128	0,2673	0,2196	0,1924	0,4056
Fc	0,68231	0,7546	0,6643	0,6289	0,6261	0,6185	0,9199
Hc	0,2948	0,4412	0,6336	0,7521	0,784	0,8206	0,865
FS	-1,34E+04	-3,99E+04	-4,40E+04	-4,28E+04	-4,87E+04	-5,83E+04	-4,92E+04
Hc_norm1	0,4253	0,4253	0,4016	0,457	0,4673	0,4375	0,4217
Hc_norm2	0,3014	0,3014	0,4561	0,6624	0,7953	0,8386	0,8882
Fc_norm	0,3537	0,3537	0,3681	0,4476	0,4639	0,4486	0,4451
Hc_stand	-10,515	-10,515	-19,132	-22,56323	-27,7784	-36,4482	-44,19
Fc_stand	10,0684	10,0684	18,9139	23,7157	31,8238	43,368	54,75
Obj	3,83E+04	2,07E+04	1,42E+04	1,12E+04	8,39E+03	6,49E+03	5,80E+03

MĚDĚNEC4_97

Počet clusterů	2	3	4	5	6	7	8
XB	0,0825	0,1637	0,1676	0,2657	0,2771	0,3674	0,3025
Fc	0,8514	0,7507	0,711	0,64	0,593	0,6037	0,6018
Hc	0,2499	0,4516	0,5535	0,7051	0,8271	0,8387	0,8768
FS	-3,71E+04	-6,60E+04	-8,82E+04	-9,22E+04	-9,51E+04	-1,06E+05	-1,09E+05
Hc_norm1	0,3606	0,3606	0,4111	0,3993	0,4381	0,4616	0,431
Hc_norm2	0,2557	0,2557	0,4674	0,5796	0,747	0,8869	0,9102
Fc_norm	0,2972	0,2972	0,3739	0,3853	0,45	0,4884	0,4623
Hc_stand	-12,6054	-12,6054	-18,3174	-26,1072	-29,7374	-33,5244	-42,448
Fc_stand	11,6897	11,6897	18,3228	27,4469	32,4751	38,5054	51,6955
Obj	5,75E+04	3,23E+04	2,15E+04	1,60E+04	1,28E+04	9,77E+03	8,03E+03

CHOMUTOV1_97

Počet clusterů	2	3	4	5	6	7	8
XB	0,0378	0,1699	0,2964	0,2673	0,2503	0,3012	0,6485
Fc	0,9138	0,773	0,6934	0,665	0,6199	0,5916	0,5437
Hc	0,1599	0,3963	0,5678	0,6469	0,7681	0,8456	0,967
FS	-3,01E+05	-5,40E+05	-5,65E+05	-5,79E+05	-5,53E+05	-5,44E+05	-5,43E+05
Hc_norm1	0,2307	0,2307	0,3607	0,4096	0,4019	0,4287	0,4396
Hc_norm2	0,1636	0,1636	0,4101	0,5945	0,6854	0,8236	0,9178
Fc_norm	0,1724	0,1724	0,3404	0,4088	0,4168	0,4562	0,4764
Hc_stand	-17,1455	-17,1455	-20,9744	-25,4037	-32,7298	-36,7042	-42,0571
Fc_stand	15,641	15,641	19,9544	25,8905	35,2839	41,8697	49,931
Obj	3,02E+05	1,45E+05	9,78E+04	7,03E+04	5,72E+04	4,67E+04	4,14E+04

CHOMUTOV2_97

Počet clusterů	2	3	4	5	6	7	8
XB	0,1907	0,1914	0,2385	0,507	0,4356	0,4496	0,4148
Fc	0,7698	0,6863	0,5987	0,5213	0,5225	0,4974	0,4975
Hc	0,3686	0,5734	0,7672	0,9434	0,9787	1,0761	1,1082
FS	2,26E+03	-7,95E+03	-1,10E+04	-1,14E+04	-1,85E+04	-1,88E+04	-2,19E+00
Hc_norm1	0,5318	0,5318	0,5219	0,5534	0,5861	0,5462	0,553
Hc_norm2	0,3774	0,3774	0,5941	0,8046	1,0016	1,0521	1,1715
Fc_norm	0,4604	0,4604	0,4706	0,535	0,5984	0,573	0,5864
Hc_stand	-6,5091	-6,5091	-12,2645	-15,3154	-17,1853	-24,9361	-28,5895
Fc_stand	6,4152	6,4152	13,3812	17,2319	19,5631	29,1677	35,5436
Obj	2,00E+04	1,21E+04	8,62E+03	6,68E+03	5,00E+03	4,15E+03	3,38E+03

CHOMUTOV3_97

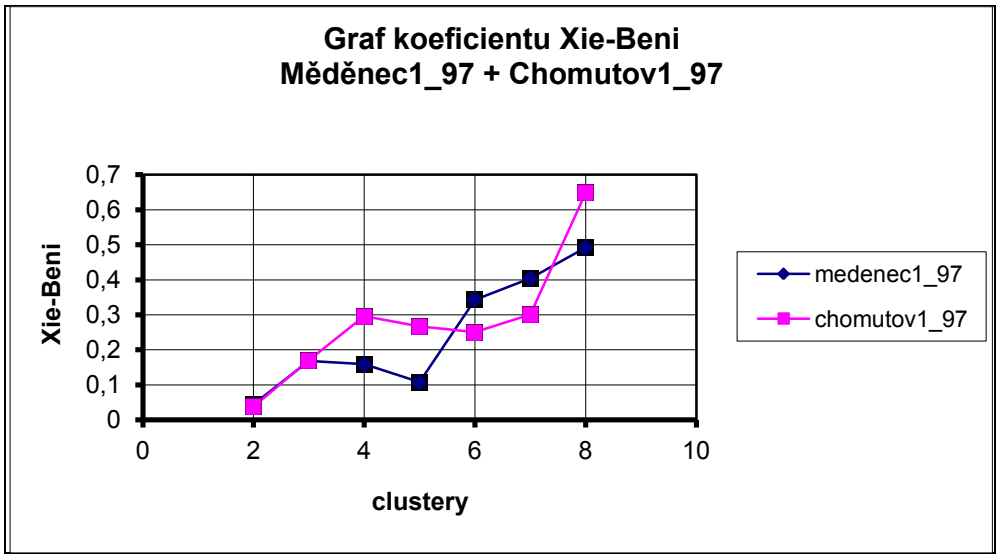
Počet clusterů	2	3	4	5	6	7	8
XB	0,1673	0,3205	0,2818	0,2858	0,3446	0,3452	0,37
Fc	0,7678	0,6321	0,5554	0,4995	0,4636	0,4599	0,4349
Hc	0,3731	0,6447	0,8413	0,9943	1,113	1,1607	1,1264
FS	-2,48E+03	-9,02E+03	-1,13E+04	-1,20E+04	-1,24E+04	-1,63E+04	-1,33E+04
Hc_norm1	0,5382	5,38E-01	5,87E-01	6,07E-01	6,18E-01	6,21E-01	5,97E-01
Hc_norm2	0,3819	0,3819	0,668	0,8823	1,0056	1,1964	1,2636
Fc_norm	0,4644	0,4644	0,5518	0,5928	0,6256	0,6437	0,6301
Hc_stand	-6,2902	-6,2902	-8,8985	-4,7248	-14,6127	-17,8317	-23,9067
Fc_stand	6,289	6,289	9,4924	13,4809	17,3006	21,91	30,1545
Obj	1,74E+04	1,06E+04	7,70E+03	5,96E+03	4,82E+03	3,96E+03	3,48E+03

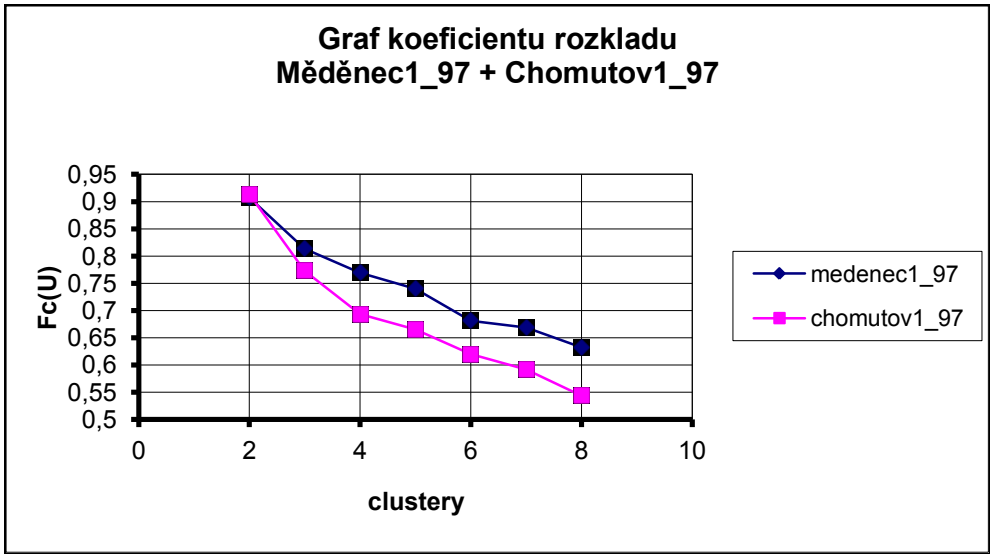
CHOMUTOV4_97

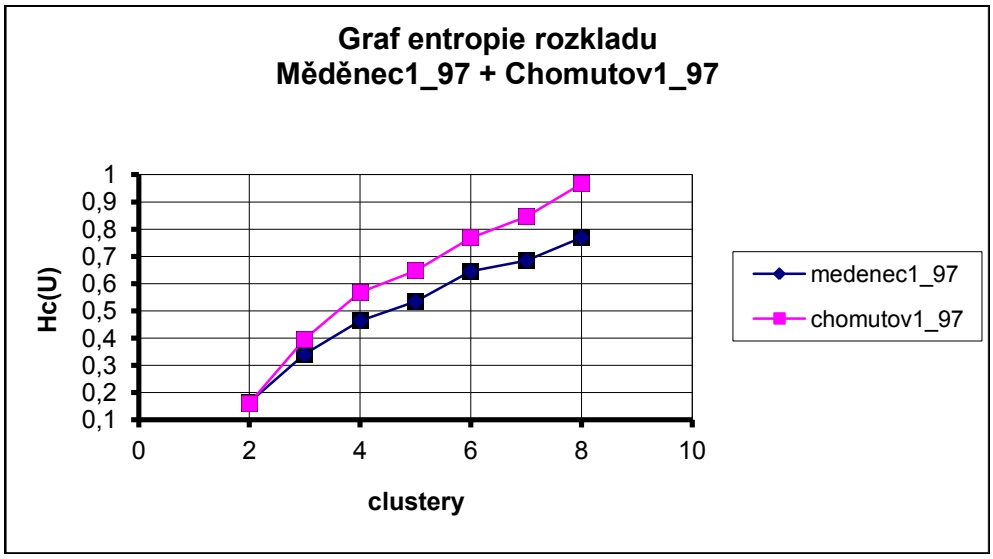
Počet clusterů	2	3	4	5	6	7	8
XB	0,1865	0,2532	0,2224	0,2627	0,3	0,2303	0,2039
Fc	0,7576	0,632	0,5679	0,5385	0,5079	0,4936	0,4881
Hc	3,87E-01	6,48E-01	8,19E-01	9,19E-01	1,03E+00	1,11E+00	1,15E+00
FS	2,63E+03	-1,97E+04	-3,30E+04	-4,87E+04	-4,34E+04	-4,67E+04	-5,92E+04
Hc_norm1	0,5579	0,5579	0,5898	0,5909	0,5707	0,5746	0,5684
Hc_norm2	0,3954	0,3954	0,6701	0,8568	0,972	1,1021	1,198
Fc_norm	0,4849	0,4849	0,552	0,5762	0,5769	0,5905	0,5908
Hc_stand	-5,7739	-5,7739	-8,9931	-13,1638	-18,9672	-22,8878	-27,7003
Fc_stand	5,8168	5,8168	9,7536	14,98	21,9648	28,1479	36,0014
Obj	6,15E+04	3,80E+04	2,69E+04	2,03E+04	1,65E+04	1,38E+04	1,10E+04

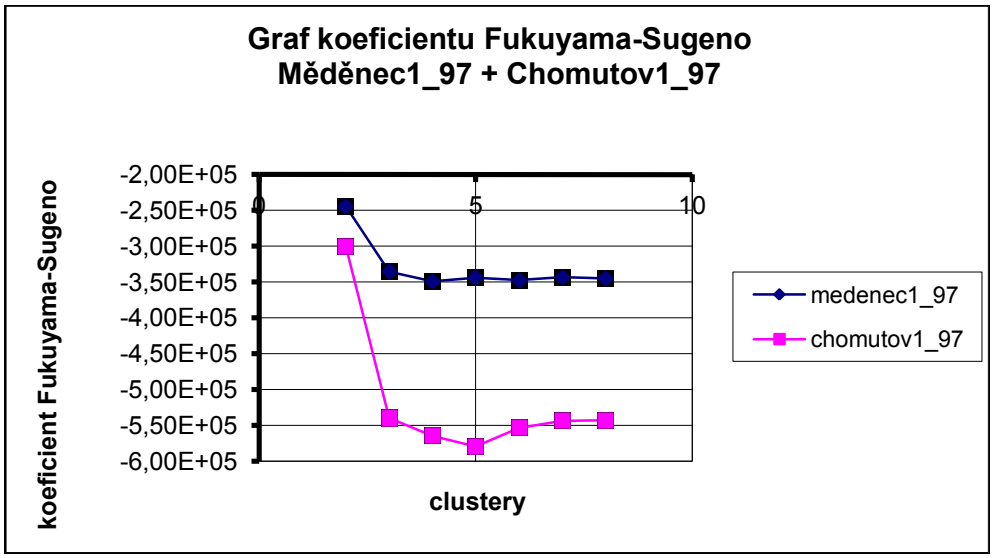
ZÁVISLOSTI CHARAKTERISTIK PRO OBĚ STANICE

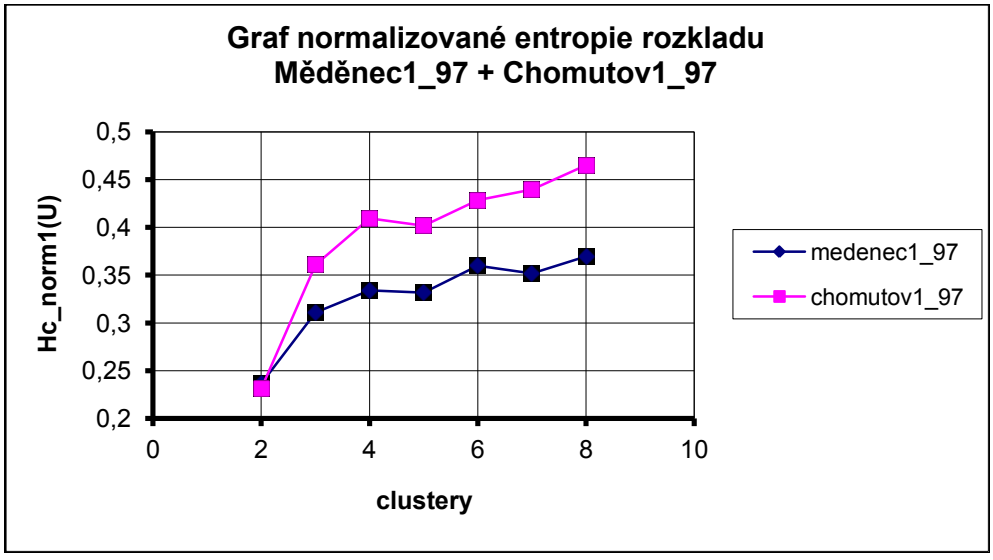
V následujících grafech jsou uvedeny závislosti sledovaných charakteristik pro obě stanice. Každý graf porovnává průběh dané charakteristiky pro obě stanice. Osa „x“ představuje počet clusterů. Osa „y“ je hodnota charakteristiky.

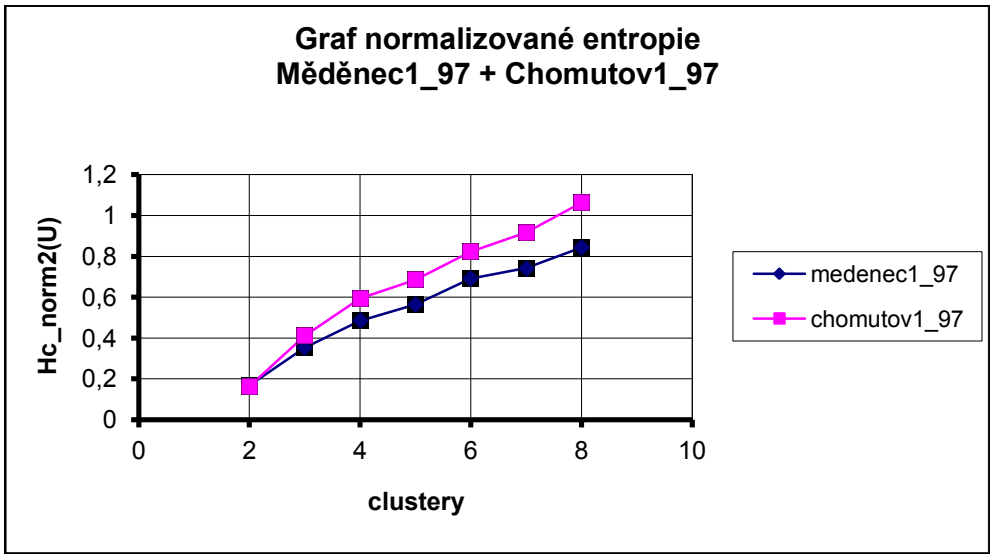


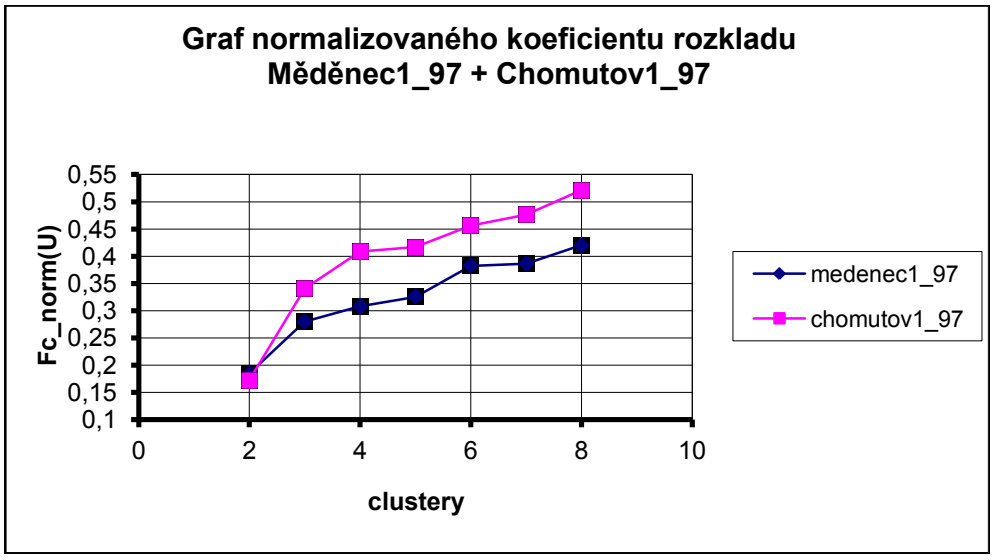


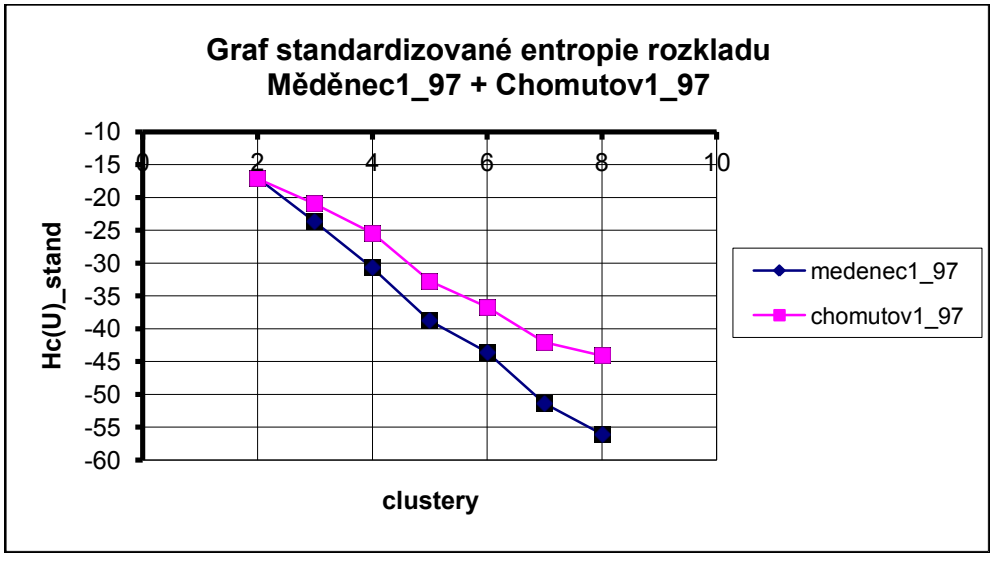

Graf entropie rozkladu
Měděnec1_97 + Chomutov1_97
Hc(U)
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
2
4
6
8
10
clustery
medenec1_97
chomutov1_97

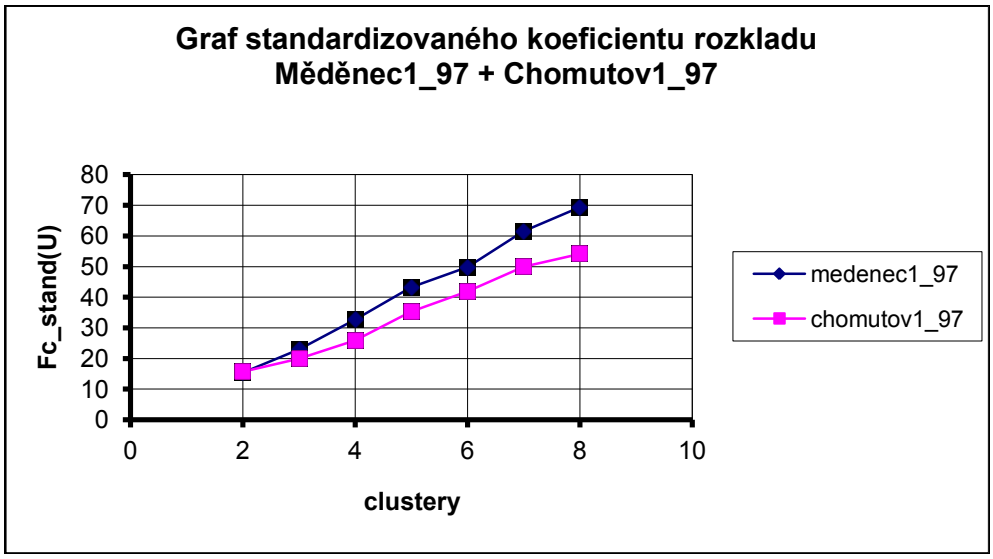

Graf koeficientu Fukuyama-Sugeno
Měděnec1_97 + Chomutov1_97
koeficient Fukuyama-Sugeno
-2,00E+05
-2,50E+05
-3,00E+05
-3,50E+05
-4,00E+05
-4,50E+05
-5,00E+05
-5,50E+05
-6,00E+05
0
5
10
clustery
medenec1_97
chomutov1_97

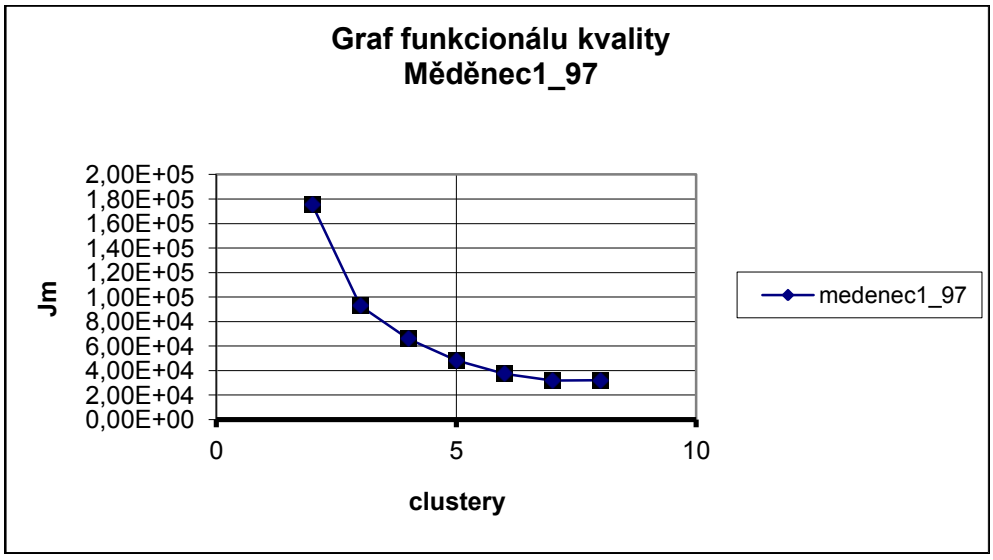

Graf normalizované entropie rozkladu
Měděnec1_97 + Chomutov1_97
Hc_norm1(U)
0,5
0,45
0,4
0,35
0,3
0,25
0,2
0
2
4
6
8
10
clustery
medenec1_97
chomutov1_97

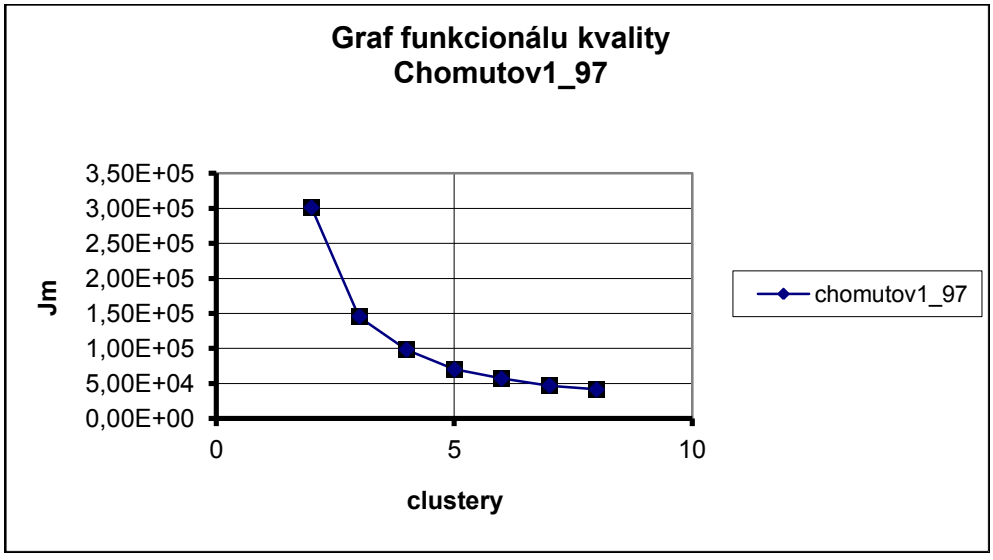

Graf normalizované entropie
Měděnec1_97 + Chomutov1_97
Hc_norm2(U)
1,2
1
0,8
0,6
0,4
0,2
0
0
2
4
6
8
10
clustery
medenec1_97
chomutov1_97

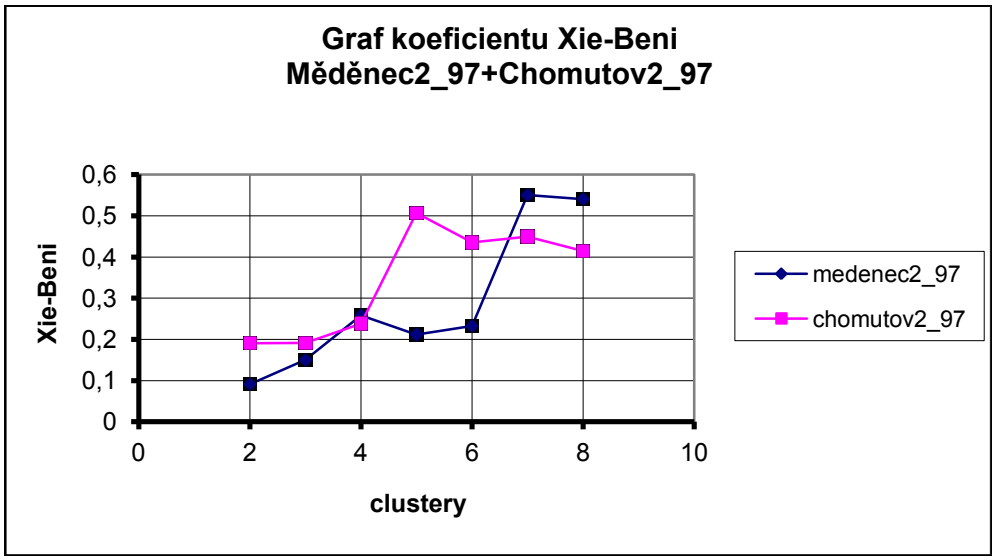

Graf normalizovaného koeficientu rozkladu
Měděnec1_97 + Chomutov1_97
Fc_norm(U)
0,55
0,5
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0
2
4
6
8
10
clustery
medenec1_97
chomutov1_97

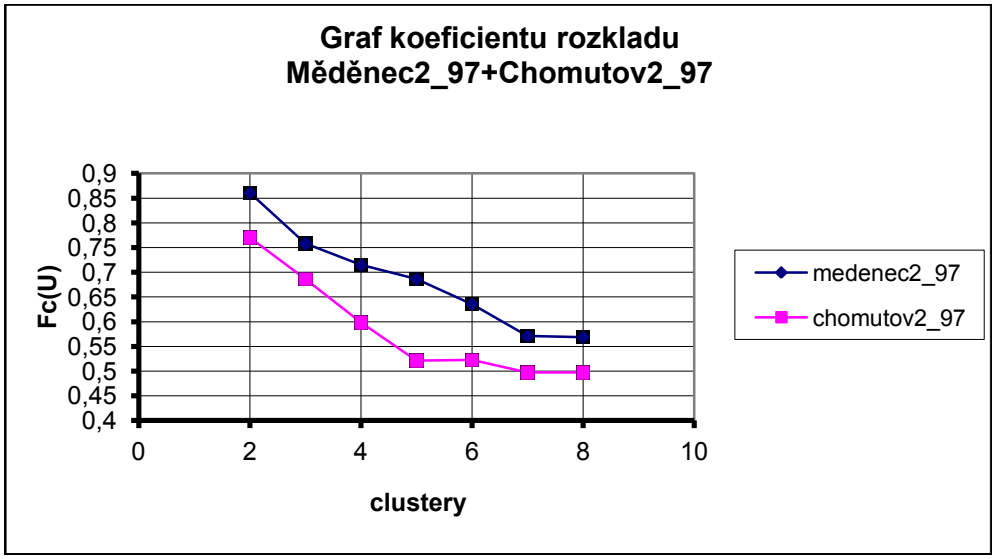

Graf standardizované entropie rozkladu
Měděnec1_97 + Chomutov1_97
Hc(U)_stand
-10
-15
-20
-25
-30
-35
-40
-45
-50
-55
-60
0
2
4
6
8
10
clustery
medenec1_97
chomutov1_97

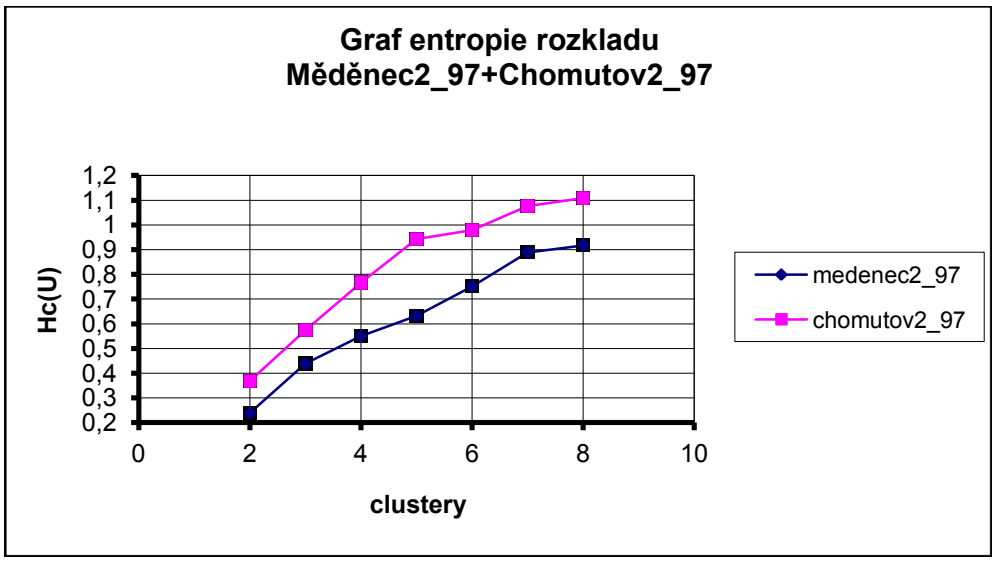

Graf standardizovaného koeficientu rozkladu
Měděnec1_97 + Chomutov1_97
Fc_stand(U)
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
clustery
medenec1_97
chomutov1_97

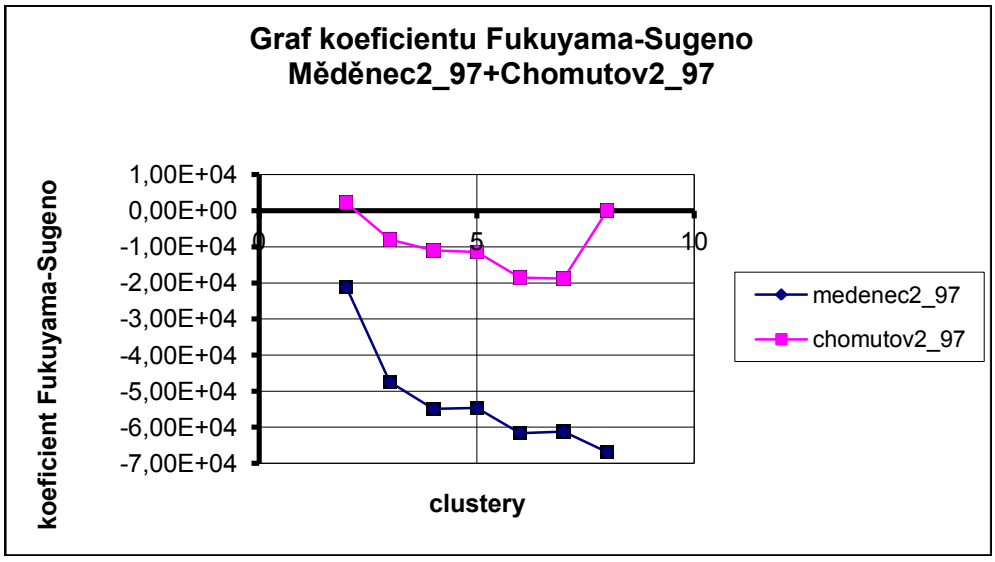

Graf funkcionálu kvality
Měděnec1_97
Jm
2,00E+05
1,80E+05
1,60E+05
1,40E+05
1,20E+05
1,00E+05
8,00E+04
6,00E+04
4,00E+04
2,00E+04
0,00E+00
0
5
10
clustery
medenec1_97

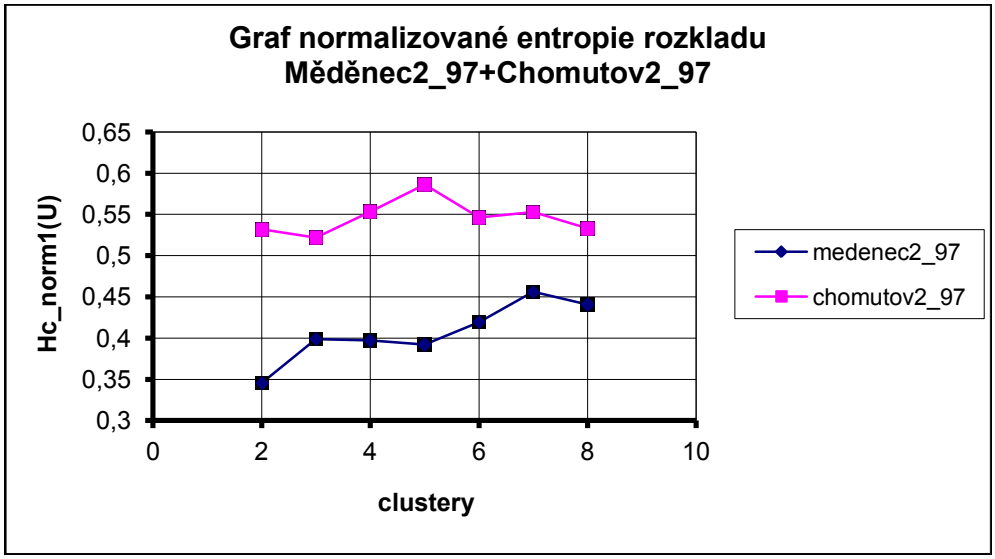

Graf funkcionálu kvality
Chomutov1_97
Jm
3,50E+05
3,00E+05
2,50E+05
2,00E+05
1,50E+05
1,00E+05
5,00E+04
0,00E+00
0
5
10
clustery
chomutov1_97

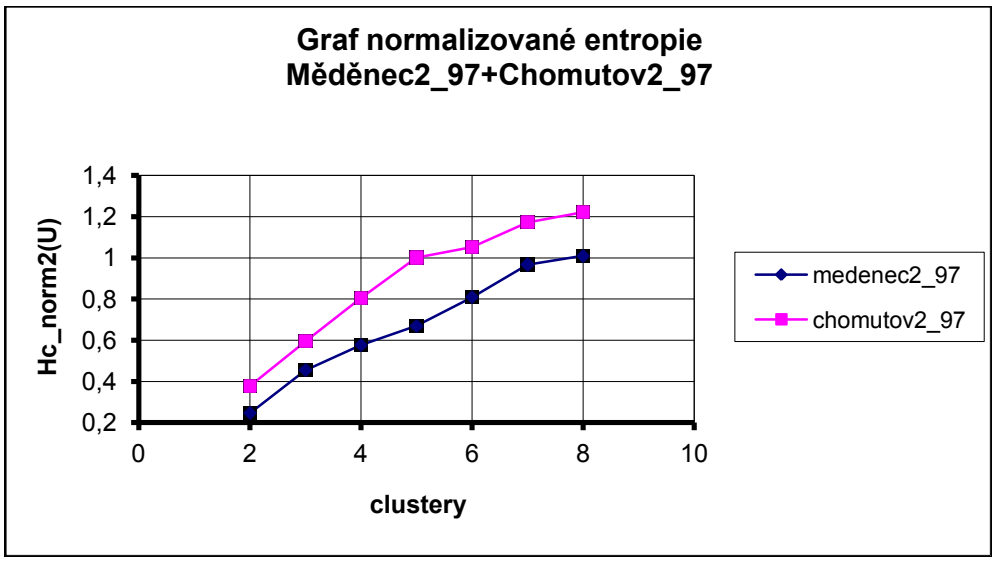

Graf koeficientu Xie-Beni
Měděnec2_97+Chomutov2_97
Xie-Beni
0,6
0,5
0,4
0,3
0,2
0,1
0
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

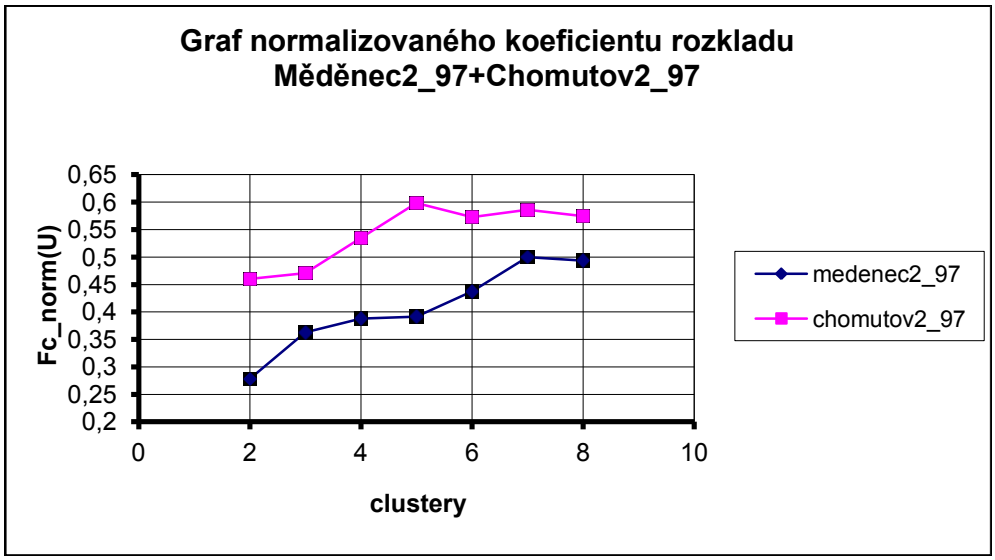

Graf koeficientu rozkladu
Měděnec2_97+Chomutov2_97
Fc(U)
0,9
0,85
0,8
0,75
0,7
0,65
0,6
0,55
0,5
0,45
0,4
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

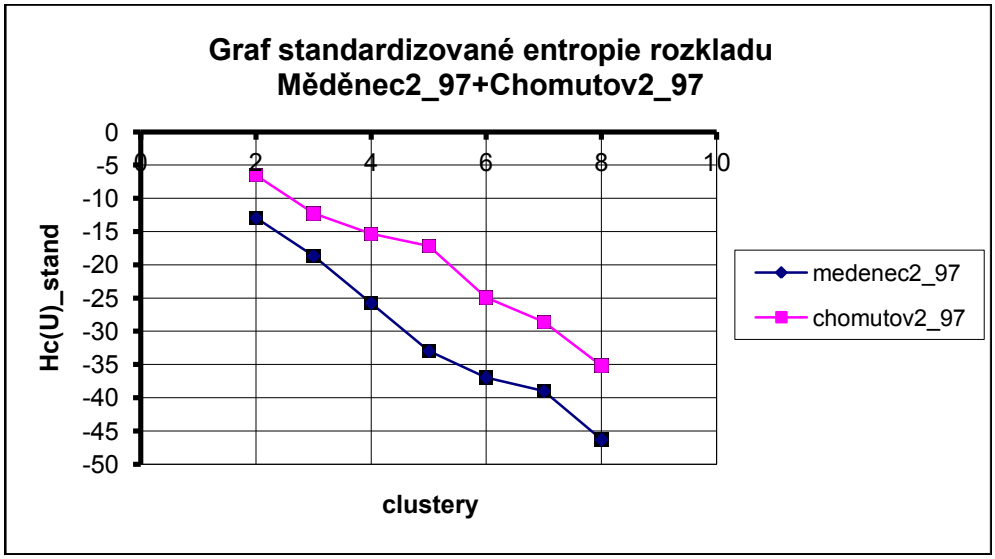

Graf entropie rozkladu
Měděnec2_97+Chomutov2_97
Hc(U)
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

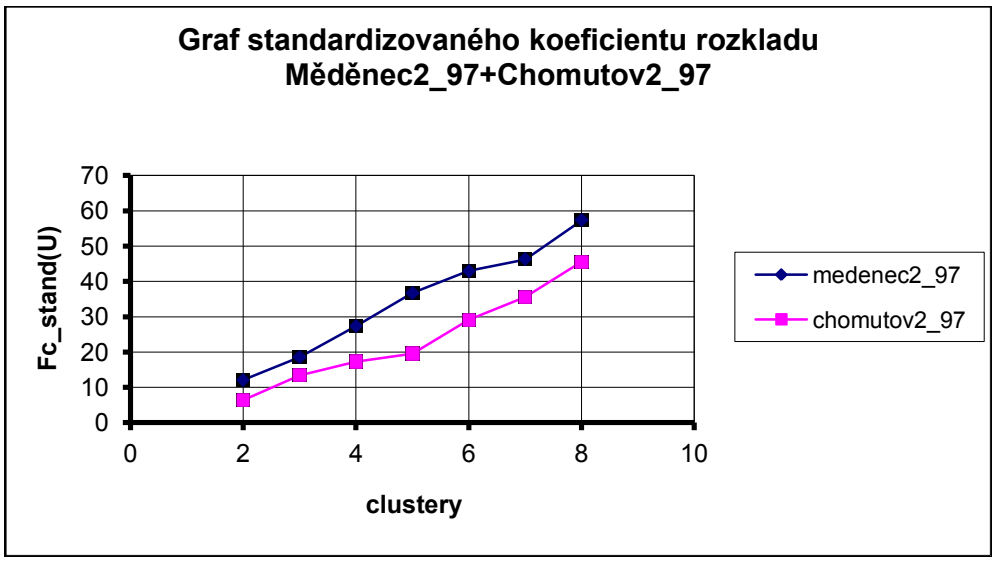

Graf koeficientu Fukuyama-Sugeno
Měděnec2_97+Chomutov2_97
koeficient Fukuyama-Sugeno
1,00E+04
0,00E+00
-1,00E+04
-2,00E+04
-3,00E+04
-4,00E+04
-5,00E+04
-6,00E+04
-7,00E+04
0
5
10
clustery
medenec2_97
chomutov2_97

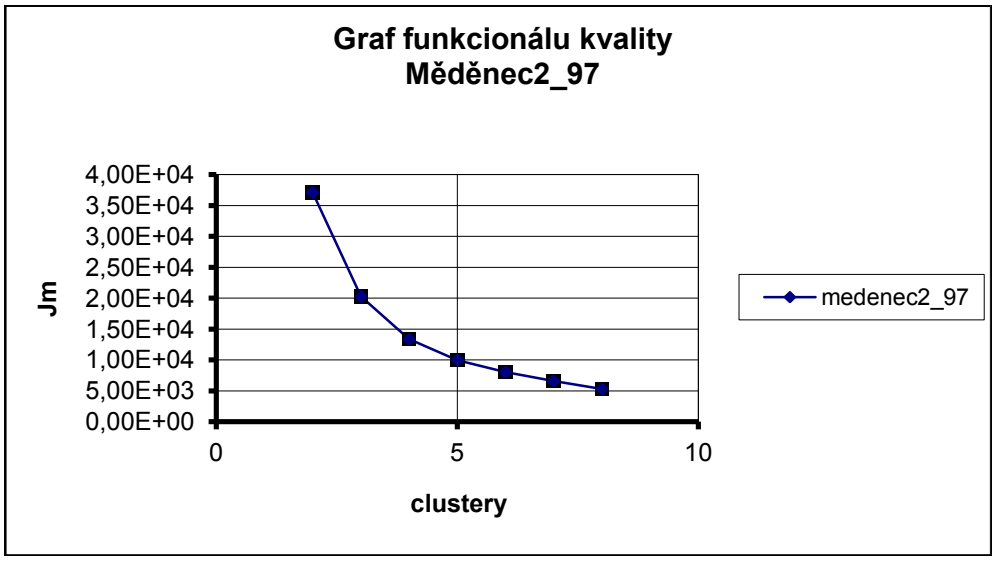

Graf normalizované entropie rozkladu
Měděnec2_97+Chomutov2_97
Hc_norm1(U)
0,65
0,6
0,55
0,5
0,45
0,4
0,35
0,3
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

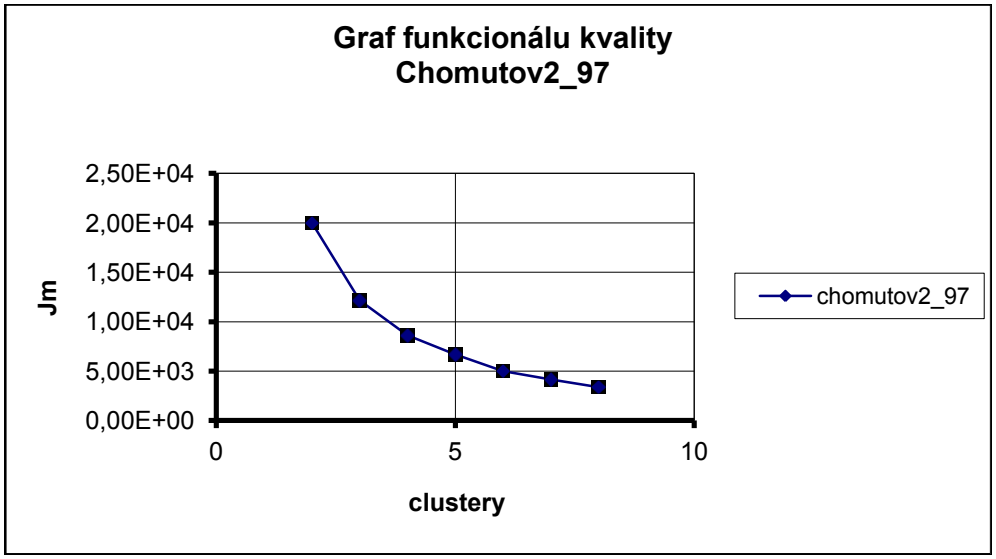

Graf normalizované entropie
Měděnec2_97+Chomutov2_97
Hc_norm2(U)
1,4
1,2
1
0,8
0,6
0,4
0,2
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

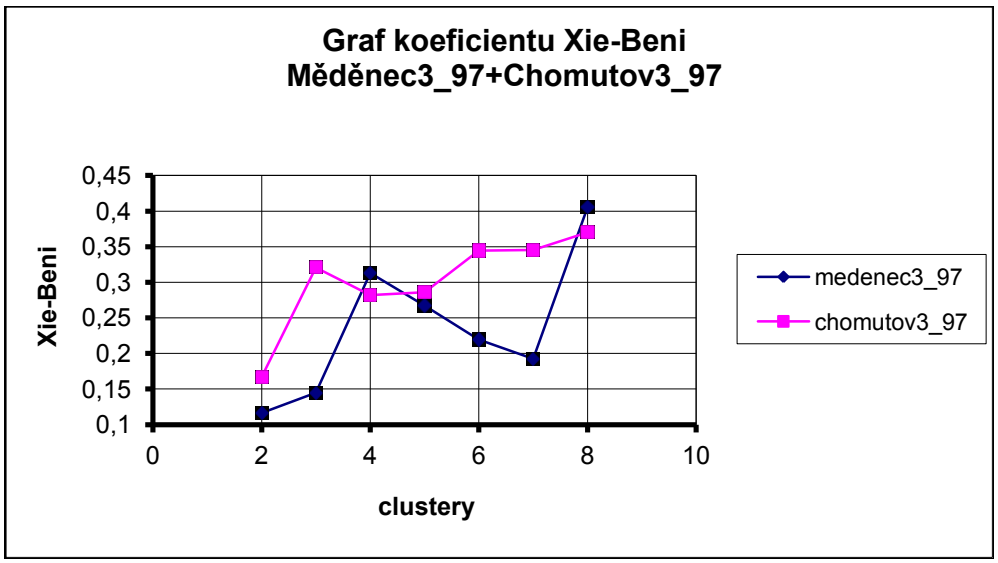

Graf normalizovaného koeficientu rozkladu
Měděnec2_97+Chomutov2_97
Fc_norm(U)
0,65
0,6
0,55
0,5
0,45
0,4
0,35
0,3
0,25
0,2
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

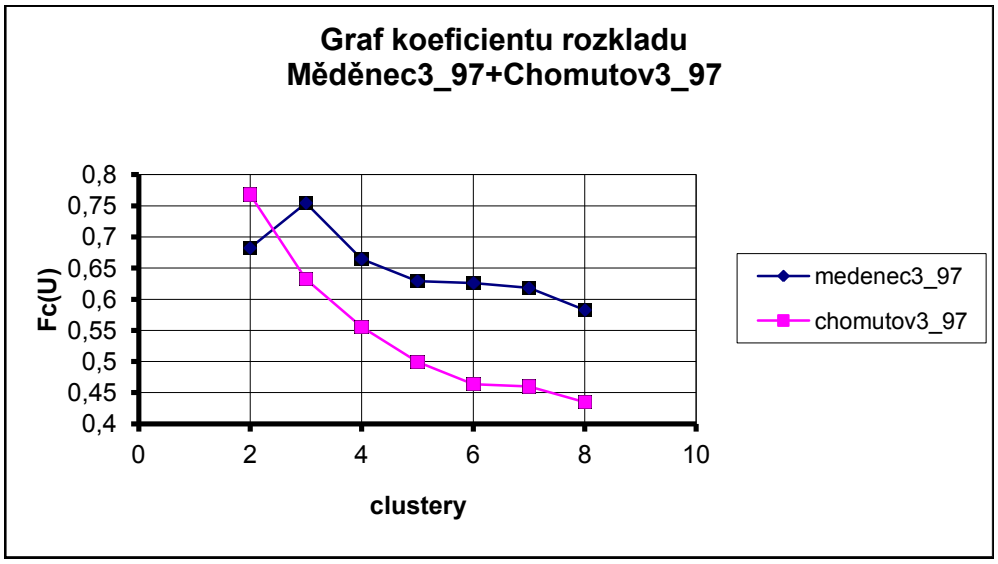

Graf standardizované entropie rozkladu
Měděnec2_97+Chomutov2_97
Hc(U)_stand
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

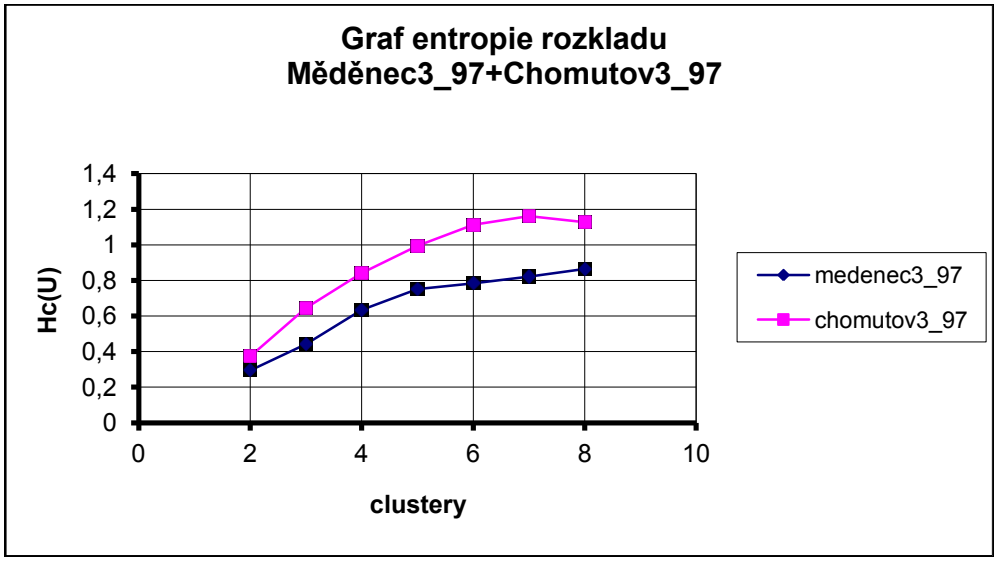

Graf standardizovaného koeficientu rozkladu
Měděnec2_97+Chomutov2_97
Fc_stand(U)
70
60
50
40
30
20
10
0
0
2
4
6
8
10
clustery
medenec2_97
chomutov2_97

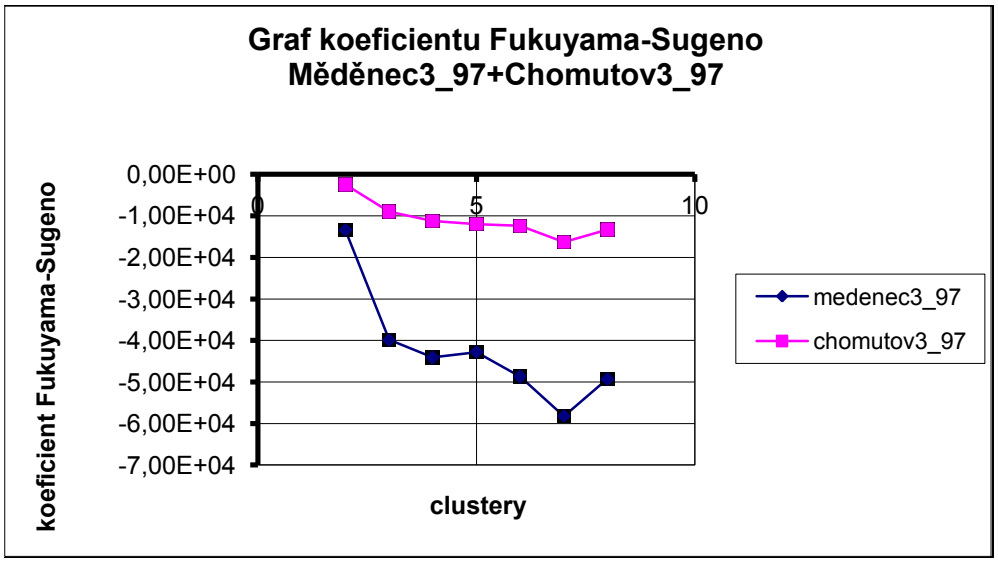

Graf funkcionálu kvality
Měděnec2_97
4,00E+04
3,50E+04
3,00E+04
2,50E+04
2,00E+04
1,50E+04
1,00E+04
5,00E+03
0,00E+00
Jm
0
5
10
clustery
medenec2_97

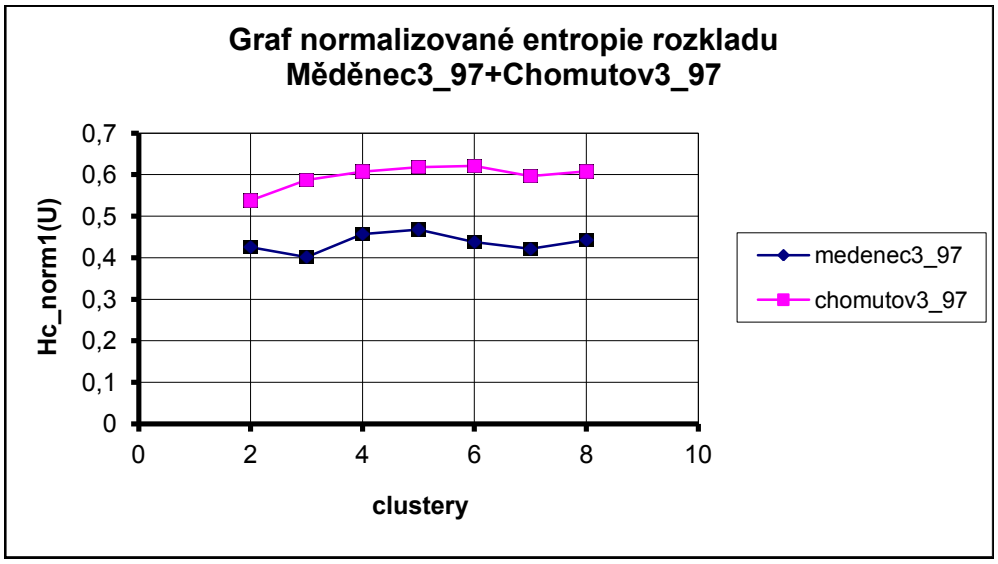

Graf funkcionálu kvality
Chomutov2_97
2,50E+04
2,00E+04
1,50E+04
1,00E+04
5,00E+03
0,00E+00
Jm
0
5
10
clustery
chomutov2_97

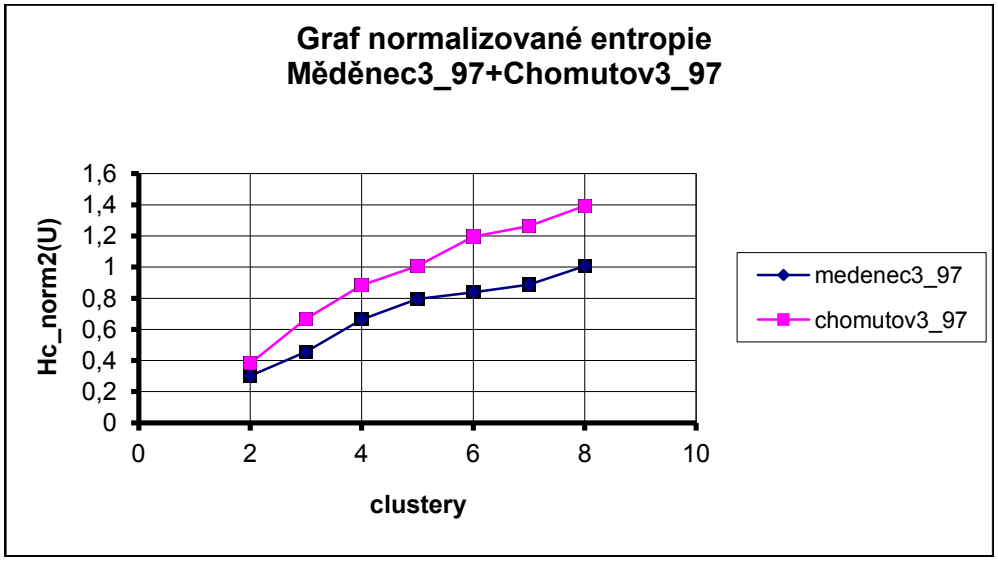

Graf koeficientu Xie-Beni
Měděnec3_97+Chomutov3_97
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
Xie-Beni
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

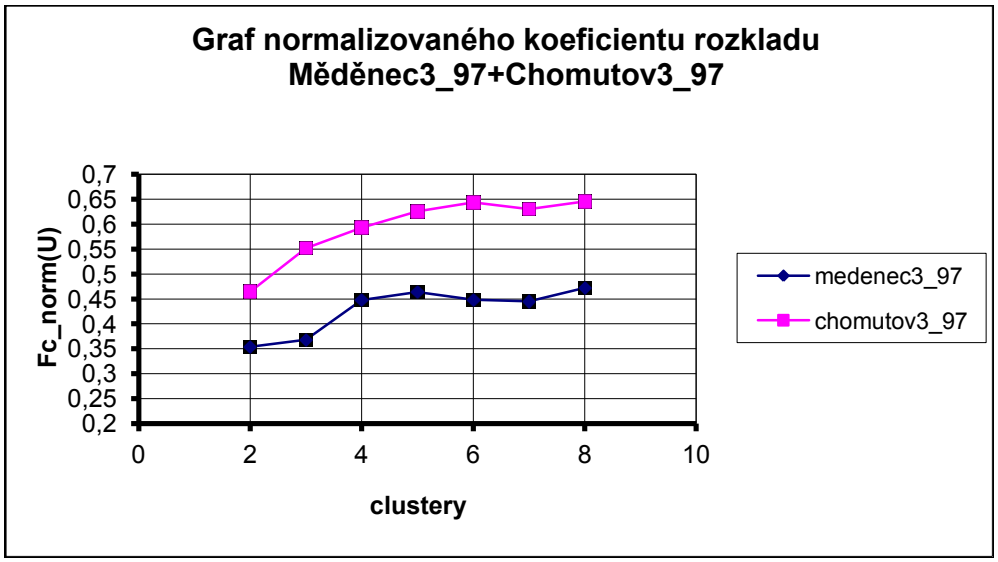

Graf koeficientu rozkladu
Měděnec3_97+Chomutov3_97
Fc(U)
0,8
0,75
0,7
0,65
0,6
0,55
0,5
0,45
0,4
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

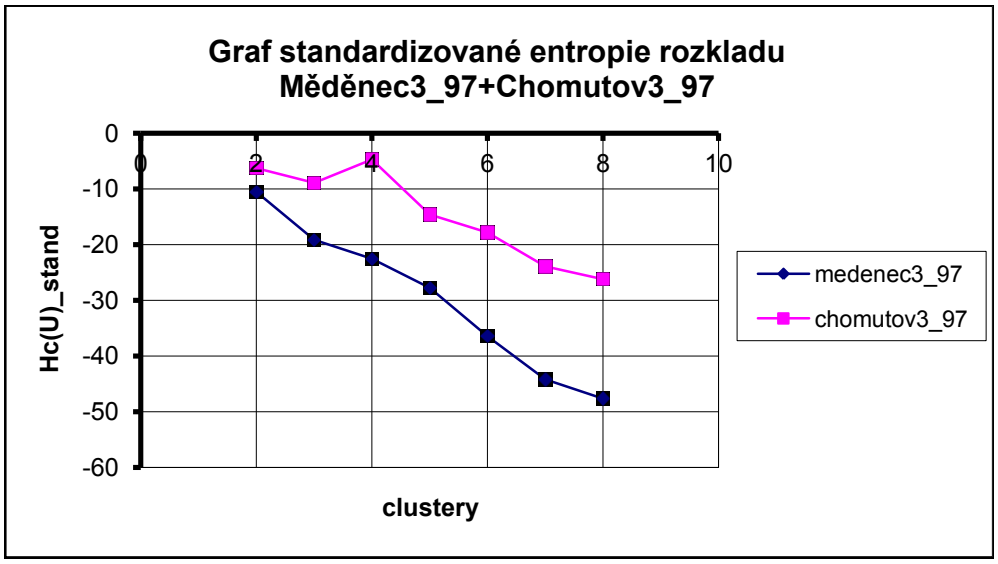

Graf entropie rozkladu
Měděnec3_97+Chomutov3_97
Hc(U)
1,4
1,2
1
0,8
0,6
0,4
0,2
0
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

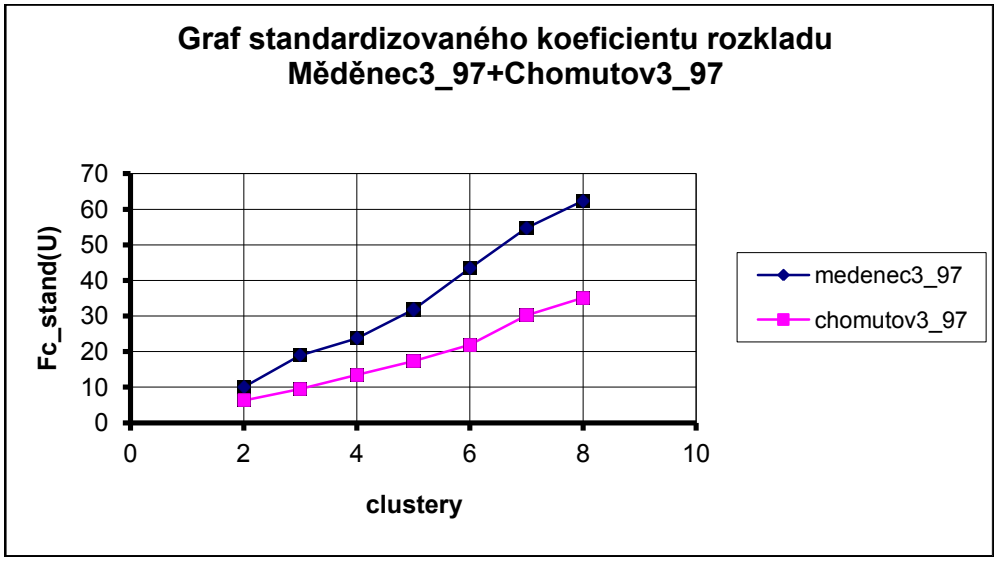

Graf koeficientu Fukuyama-Sugeno
Měděnec3_97+Chomutov3_97
koeficient Fukuyama-Sugeno
0,00E+00
-1,00E+04
-2,00E+04
-3,00E+04
-4,00E+04
-5,00E+04
-6,00E+04
-7,00E+04
0
5
10
clustery
medenec3_97
chomutov3_97

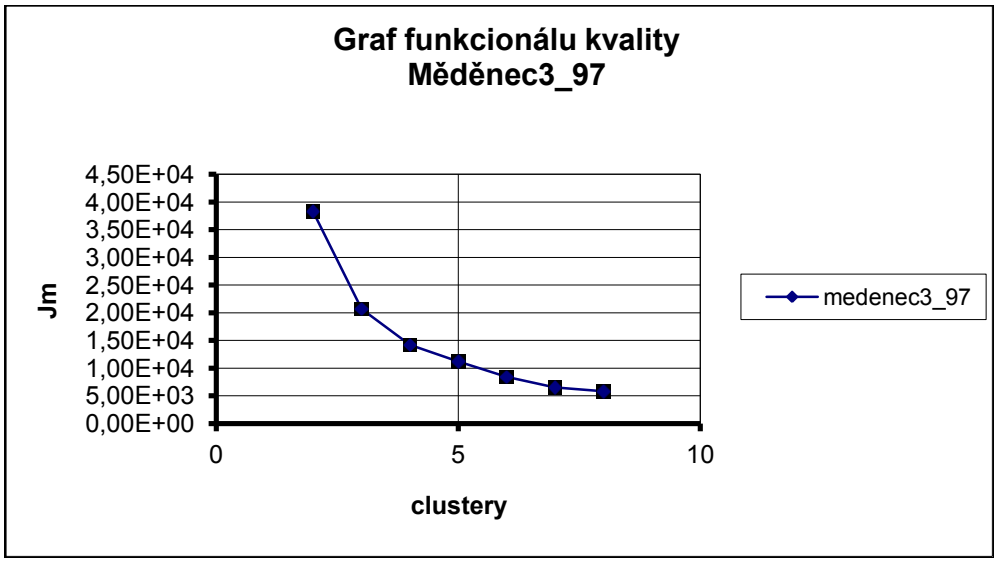

Graf normalizované entropie rozkladu
Měděnec3_97+Chomutov3_97
Hc_norm1(U)
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

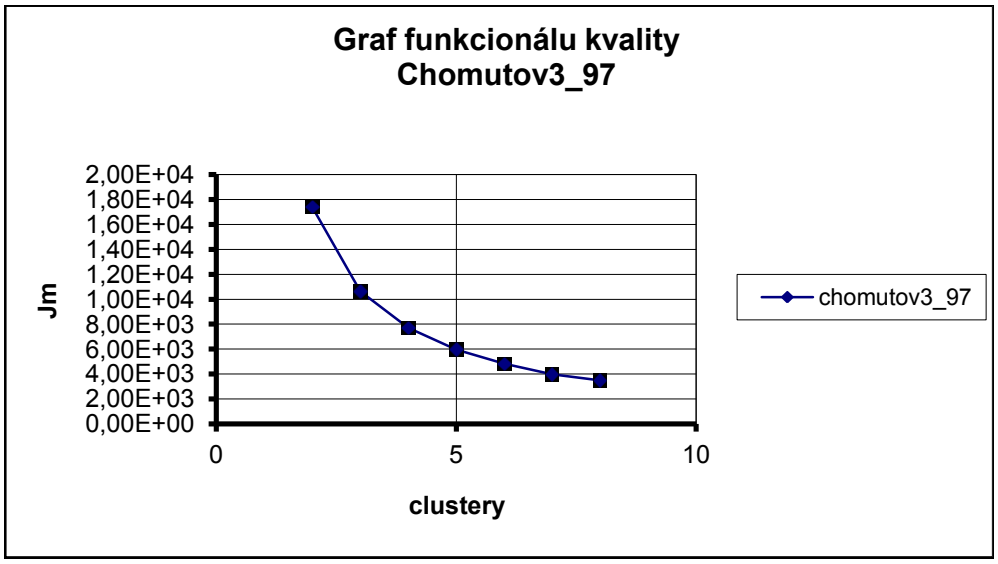

Graf normalizované entropie
Měděnec3_97+Chomutov3_97
Hc_norm2(U)
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

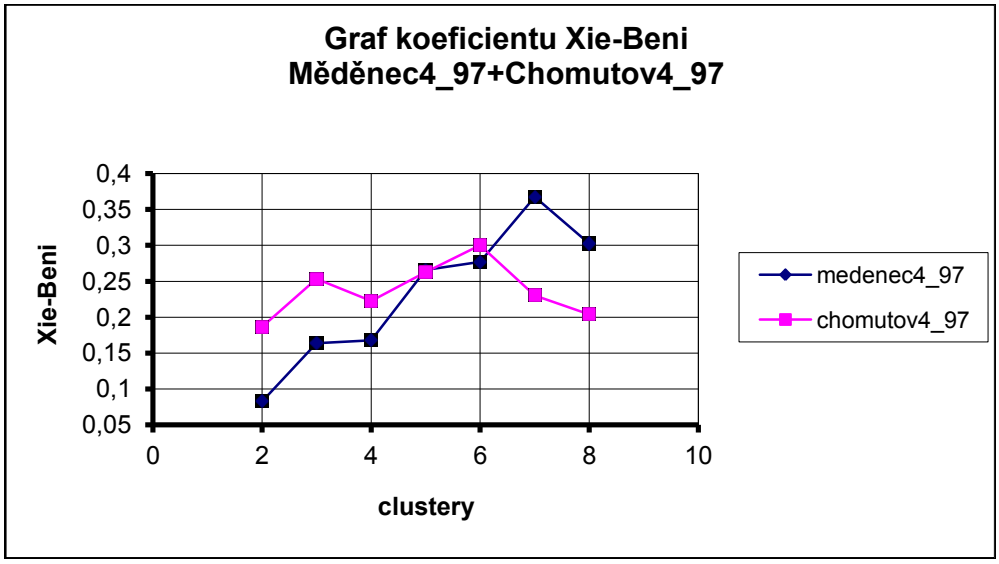

Graf normalizovaného koeficientu rozkladu
Měděnec3_97+Chomutov3_97
Fc_norm(U)
0,7
0,65
0,6
0,55
0,5
0,45
0,4
0,35
0,3
0,25
0,2
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

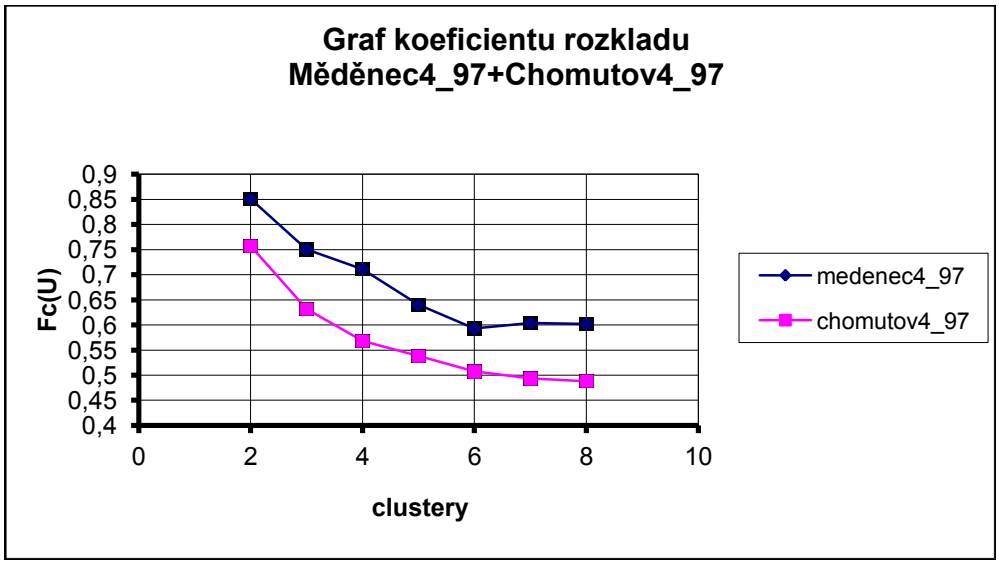

Graf standardizované entropie rozkladu
Měděnec3_97+Chomutov3_97
Hc(U)_stand
0
-10
-20
-30
-40
-50
-60
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

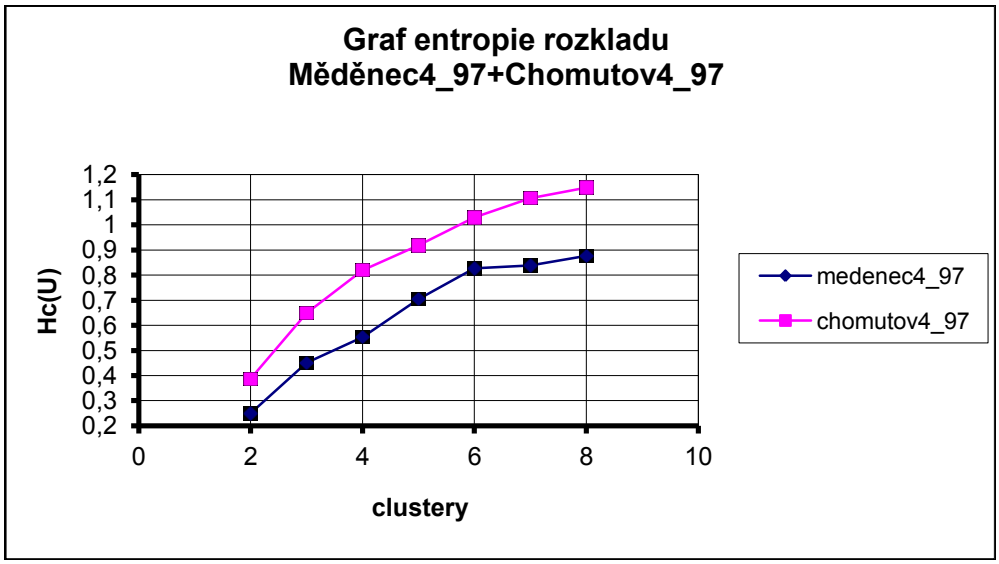

Graf standardizovaného koeficientu rozkladu
Měděnec3_97+Chomutov3_97
Fc_stand(U)
70
60
50
40
30
20
10
0
0
2
4
6
8
10
clustery
medenec3_97
chomutov3_97

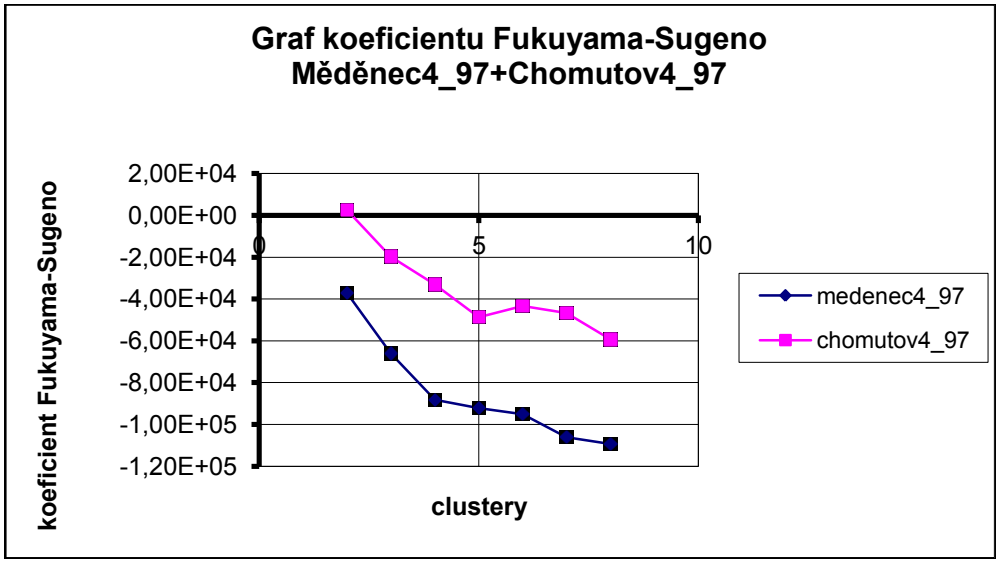

Graf funkcionálu kvality
Měděnec3_97
Jm
4,50E+04
4,00E+04
3,50E+04
3,00E+04
2,50E+04
2,00E+04
1,50E+04
1,00E+04
5,00E+03
0,00E+00
0
5
10
clustery
medenec3_97

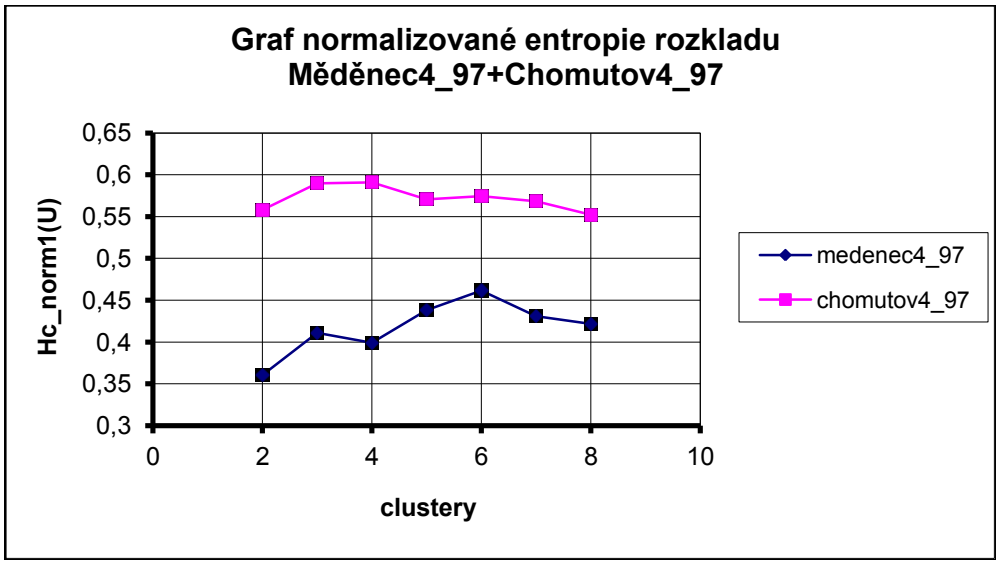

Graf funkcionálu kvality
Chomutov3_97
Jm
2,00E+04
1,80E+04
1,60E+04
1,40E+04
1,20E+04
1,00E+04
8,00E+03
6,00E+03
4,00E+03
2,00E+03
0,00E+00
0
5
10
clustery
chomutov3_97

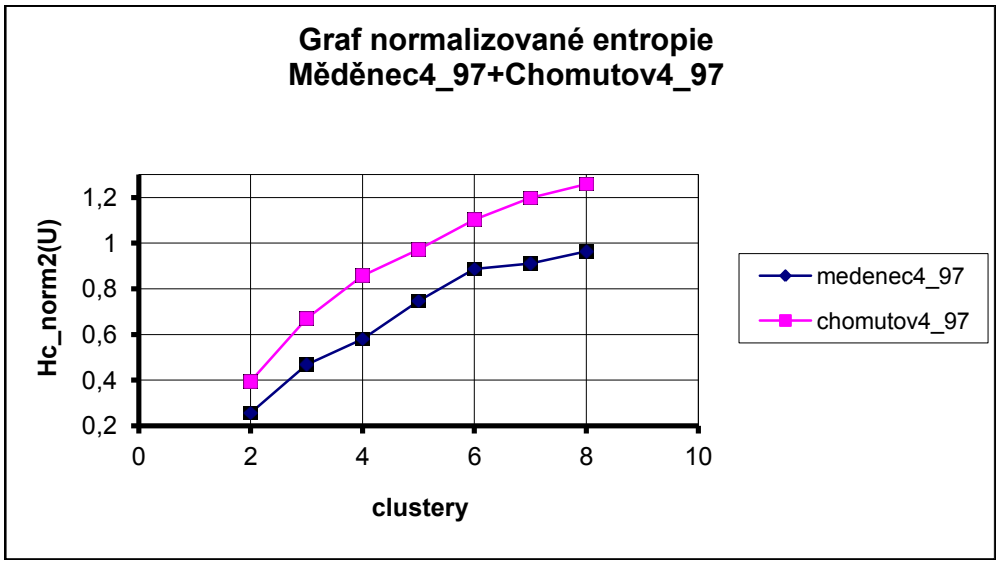

Graf koeficientu Xie-Beni
Měděnec4_97+Chomutov4_97
Xie-Beni
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97

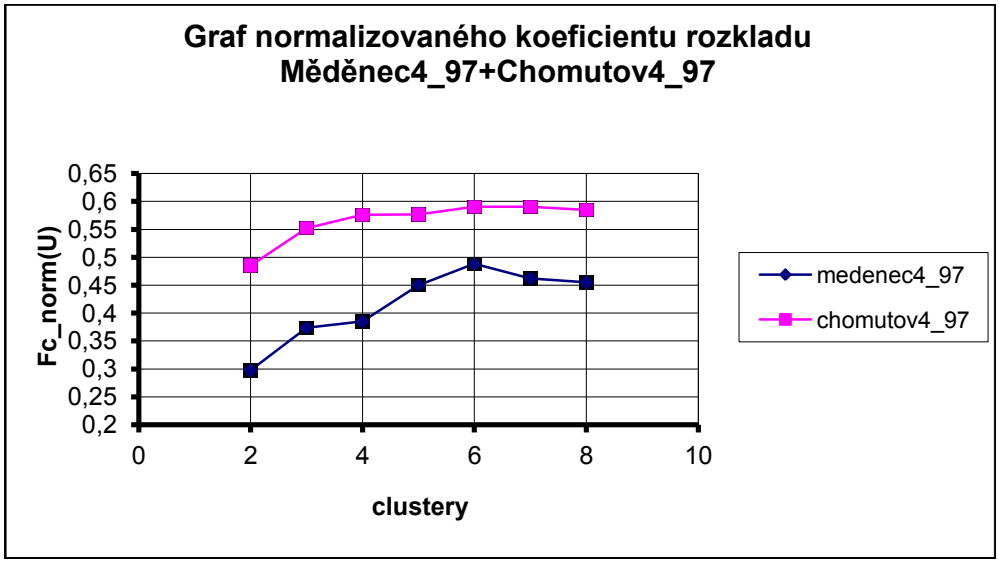

Graf koeficientu rozkladu
Měděnec4_97+Chomutov4_97
Fc(U)
0,9
0,85
0,8
0,75
0,7
0,65
0,6
0,55
0,5
0,45
0,4
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97

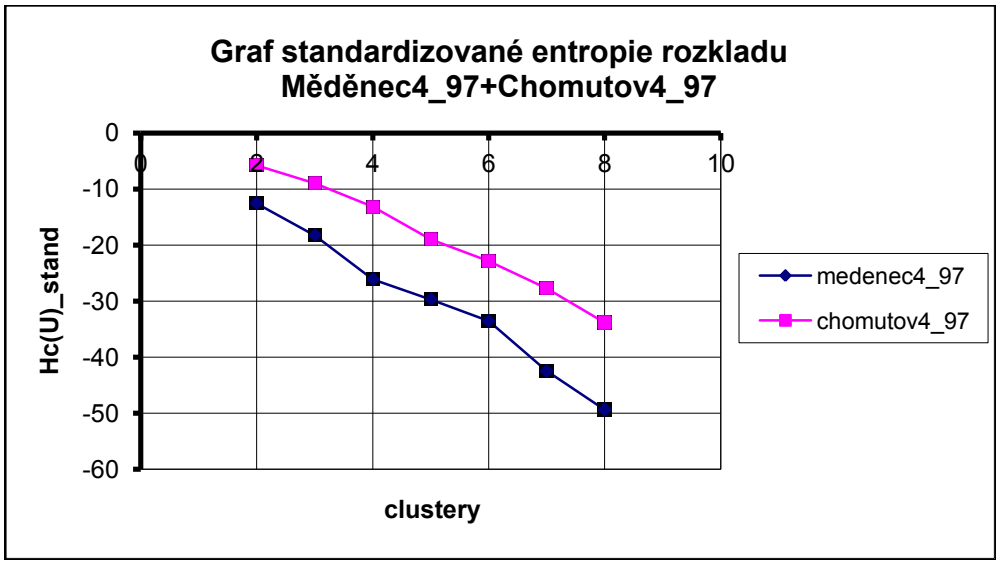

Graf entropie rozkladu
Měděnec4_97+Chomutov4_97
Hc(U)
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97

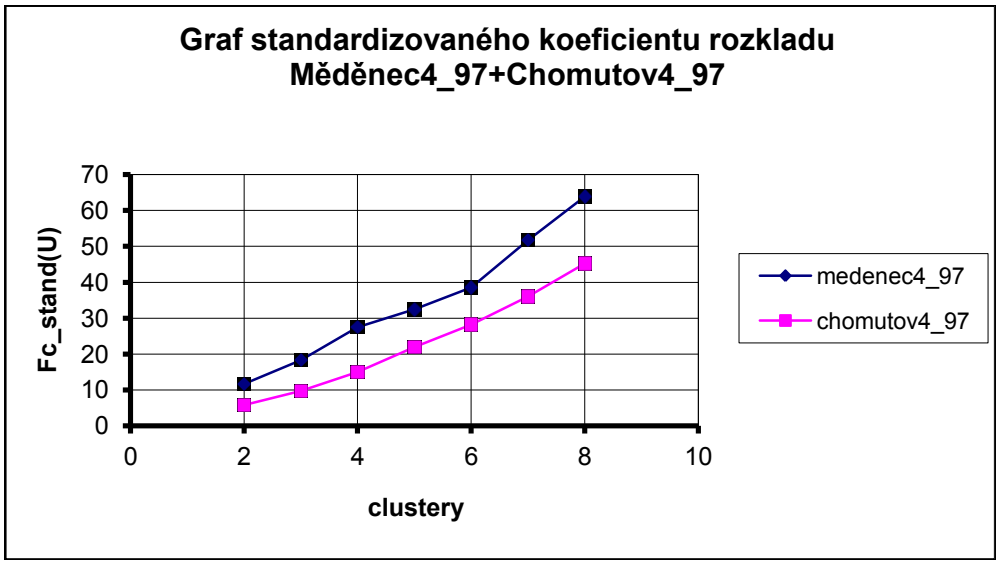

Graf koeficientu Fukuyama-Sugeno
Měděnec4_97+Chomutov4_97
koeficient Fukuyama-Sugeno
2,00E+04
0,00E+00
-2,00E+04
-4,00E+04
-6,00E+04
-8,00E+04
-1,00E+05
-1,20E+05
0
5
10
clustery
medenec4_97
chomutov4_97

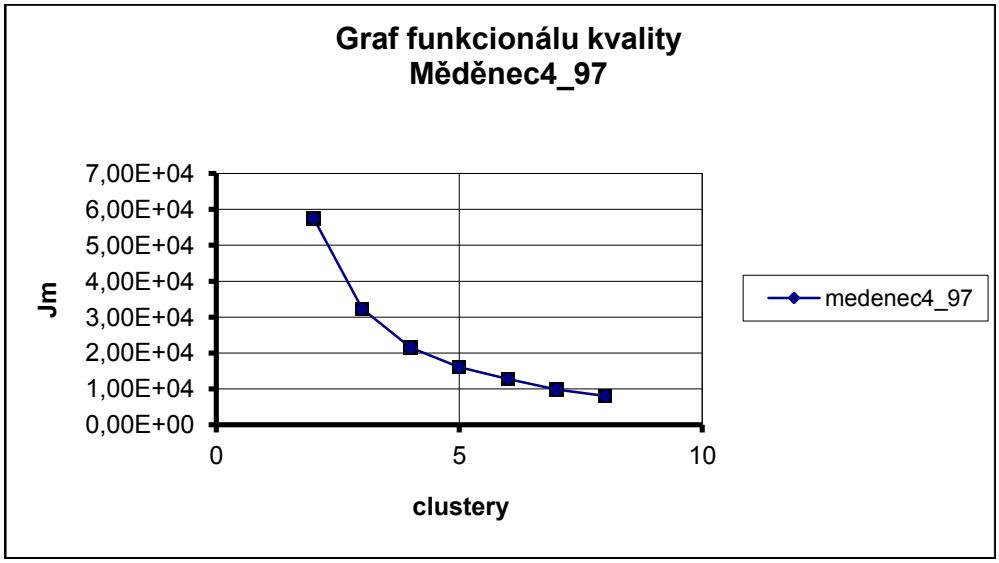

Graf normalizované entropie rozkladu
Měděnec4_97+Chomutov4_97
Hc_norm1(U)
0,65
0,6
0,55
0,5
0,45
0,4
0,35
0,3
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97

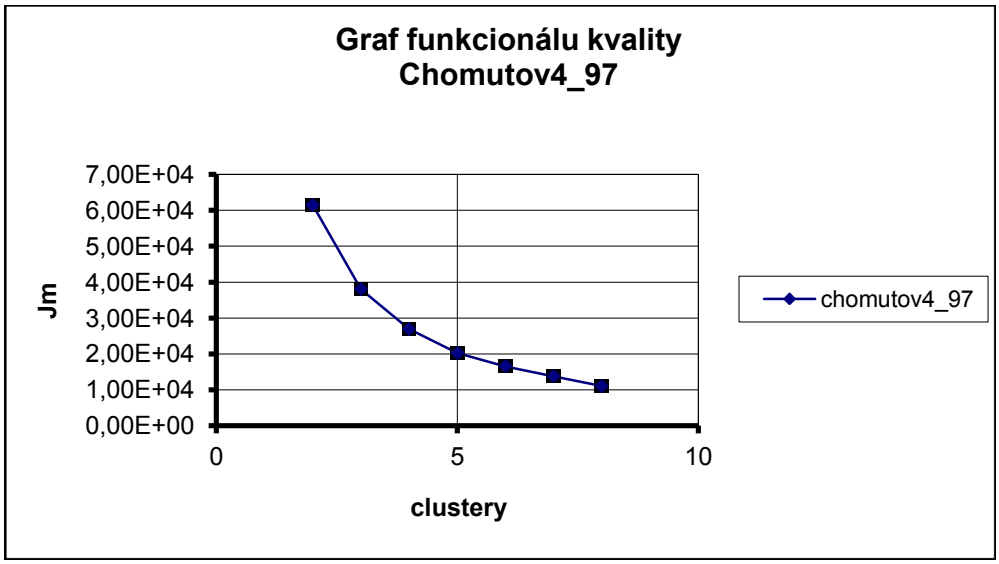

Graf normalizované entropie
Měděnec4_97+Chomutov4_97
Hc_norm2(U)
1,2
1
0,8
0,6
0,4
0,2
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97


Graf normalizovaného koeficientu rozkladu
Měděnec4_97+Chomutov4_97
Fc_norm(U)
0,65
0,6
0,55
0,5
0,45
0,4
0,35
0,3
0,25
0,2
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97


Graf standardizované entropie rozkladu
Měděnec4_97+Chomutov4_97
Hc(U)_stand
0
-10
-20
-30
-40
-50
-60
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97


Graf standardizovaného koeficientu rozkladu
Měděnec4_97+Chomutov4_97
Fc_stand(U)
70
60
50
40
30
20
10
0
0
2
4
6
8
10
clustery
medenec4_97
chomutov4_97


Graf funkcionálu kvality
Měděnec4_97
Jm
7,00E+04
6,00E+04
5,00E+04
4,00E+04
3,00E+04
2,00E+04
1,00E+04
0,00E+00
0
5
10
clustery
medenec4_97


Graf funkcionálu kvality
Chomutov4_97
Jm
7,00E+04
6,00E+04
5,00E+04
4,00E+04
3,00E+04
2,00E+04
1,00E+04
0,00E+00
0
5
10
clustery
chomutov4_97

VOLBA POČTU CLUSTERŮ

Podle uvedených závislostí indexů validity přichází na řadu nejdůležitější krok, a to volba optimálního počtu clusterů. Ze všech uvedených grafů vyplývá několik věcí. Předně standardizované koeficienty rozkladu a entropie nejsou zdaleka závislé na počtu clusterů tak, aby bylo možné jednoznačně určit optimální počet clusterů.
Otimální rozklad je v minimu klasické entropie rozkladu a v maximu koeficientu rozkladu. Oba datové soubory medenec97 a chomutov97 neobsahují zřejmě dostatečně clusterovatelné struktury. Projevuje se u nich tendence monotonicity pro oba koeficienty. V takovém případě se optimální rozklad určuje jako průnik dvou přímek, dvou tečen na začátku průběhu pro 2 clustery a na konci pro 8 clusterů. Pro naše účely stačí graficky pravítkem vyznačit obě tečny v již vytisknutých grafech. Koeficienty Fc a Hc se hodí na soubory s velmi dobře clusterovatelnými daty. Pokud se v průběhu obou koeficientů objeví jednoznačný extrém, potom je podle toho určené „c“ možno brát jako nejlepší. Koeficient Xie-Beni je nejvíce reagující charakteristika na změnu počtu clusterů, avšak nutno dodat že občas se nechová nejlépe. Jeho minimum by mělo určit optimum, ale občas koeficient také tíhne k monotonicitě a mívá extrémně odlišné hodnoty. Koeficient Fukuyama-Sugeno se chová nejlépe, jeho tendence se příliš nemění. Z počátku ostře klesá a posléze nabývá tendenci konstantní funkce. Jako nejlepší počet clusterů se bere proměnná, od které koeficient začíná jevit rysy konstantnosti. Koeficient normalizované entropie má význam brát v úvahu pouze Hc_norm1(U). Koeficienty normalizované entropie a rozkladu nemají obecně příliš velkou vypovídací hodnotu. Dle mého názoru by se měli případně brát v úvahu jejich lokální extrémy. Ohledně hodnoty funkcionálu je zjevné, že se jedná jednoznačně o monotónně klesající funkci. Teoreticky je možné posuzovat dvě hodnoty funkcionálu lišící se o jeden cluster po sobě a je-li rozdíl menší než zvolená hodnota, našli jsme optimální rozklad. Takový postup ovšem není příliš věrohodný ve srovnání s koeficienty validity vzhledem k evidentní monotonicitě funkcionálu. Jako lokální extrém popsaných diskrétních charakteristik je bod, jehož nejbližší sousedi mají hodnotu charakteristiky menší(větší).
Jako lokální extrém popsaných diskrétních charakteristik je bod, jehož nejbližší sousedi mají hodnotu charakteristiky menší(větší).

Měděnec1_97

Pro medenec1_97 bylo vybráno 5 clusterů. Xie-Beni (dále XB) má jednoznačné minimum, Fukuyama-Sugeno (FS) volí již 4, ale 5 vyhovuje taktéž. Průnik tečen Fc a Hc je mezi 4 a 5. Normalizované Hc_norm1(U) a Fc_norm(U) nabývají oba lokálních minim pro počet clusterů c = 5.

Měděnec2_97

Pro medenec2_97 bylo vybráno c = 5. XB má pro c =5 lokální minimum, FS není příliš zřejmý stejně jako Fc a Hc, Hc_norm1(U) a Fc_norm(U) mají pro c = 5 lokální extrém.

Měděnec3_97

Pro medenec3_97 bylo vybráno podle Fc c = 3. Všechny charakteristiky krom XB a Fc jsou nejednoznačné. Fc má maximum v c = 3.

Měděnec4_97

Pro medenec4_97 bylo vybráno c = 4. XB má téměř lokální minimum. XB v tomto případě není nejlepší, protože nemá výrazné extrémy. Tečny Fc a Hc také volí c = 4.

Hc_norm1(U) a Fc_norm(U) mají pro c = 4 lokální extrémy.

Chumutov1_97
Pro chomutov1_97 bylo vybráno c = 6. XB má jednoznačné minimum, FS je zřejmě konstantní, Hc_norm1(U) a Fc_norm(U) mají lokální extrémy.

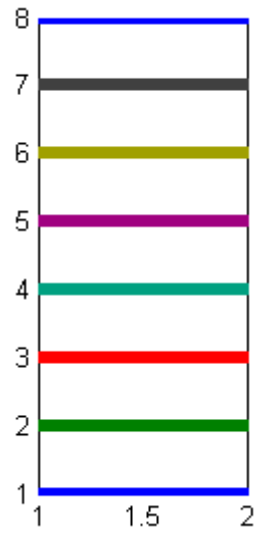
Chumutov2_97
Pro chomutov2_97 bylo vybráno c = 6. U tohoto datového souboru je rozhodnutí velmi nejednoznačné. Podle XB může být c = 6 nebo 8. Fc, Hc a FS jsou nejednoznačné. Podle Hc_norm1(U) a Fc_norm(U) je to c = 5 nebo 7.

Chumutov3_97
Pro chomutov3_97 bylo vybráno c = 4. Nejvěrohodnější je XB a metoda tečen pro Fc a Hc.

Chumutov4_97
Pro chomutov4_97 bylo vybráno c = 4. Nejvěrohodnější je opět XB a metoda tečen. Také Hc_norm1(U) ukazuje na c = 4.

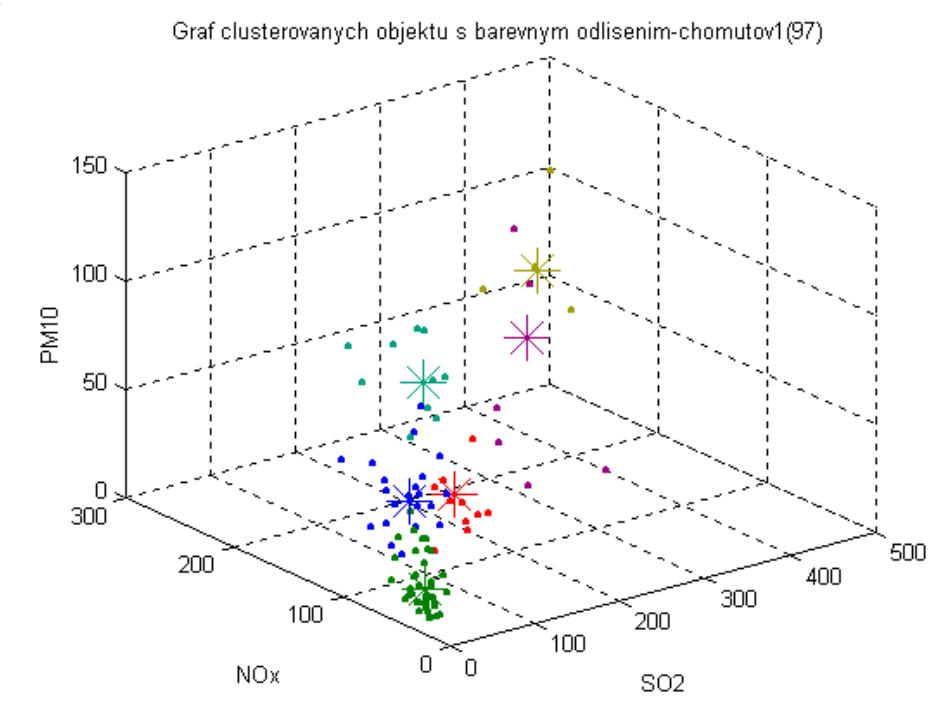
BAREVNÉ OZNAČENÍ CLUSTERŮ

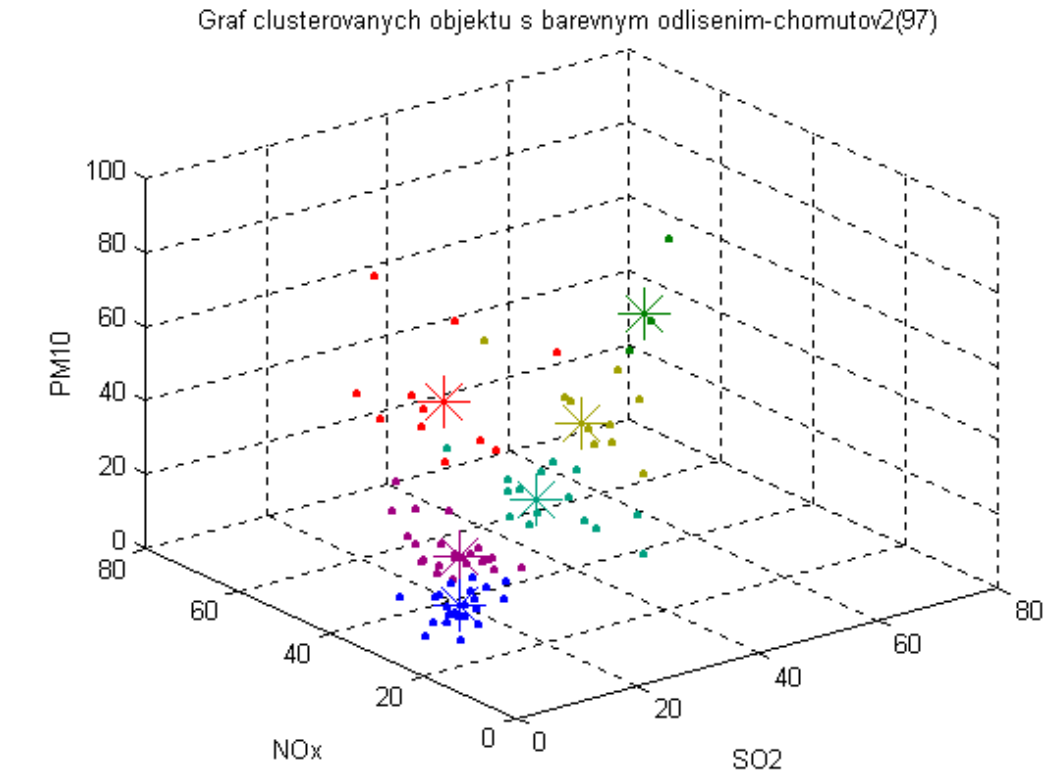
Po volbě počtu clusterů a clusterové analýze vůbec, je důležitým výstupem matice funkcí příslušnosti. Maximum funkce pro daný objekt přes všechny clustery rozhoduje kam objekt zařadit. Tímto způsobem jsou objektům přiřazena čísla 1 až 8, kde číslo značí pořadové číslo clusteru. Jako výsledek clusterové analýzy je uveden 3D graf clusterovaných objektů v barvě v prostředí Matlabu. Co barva to cluster s pořadovým číslem. Abychom věděli které barvě vždy přísluší jisté pořadí, existuje pomůcka ve formě obrázku. Toto číslo označuje pořadí clusteru a vždy mu přísluší určitá barva.

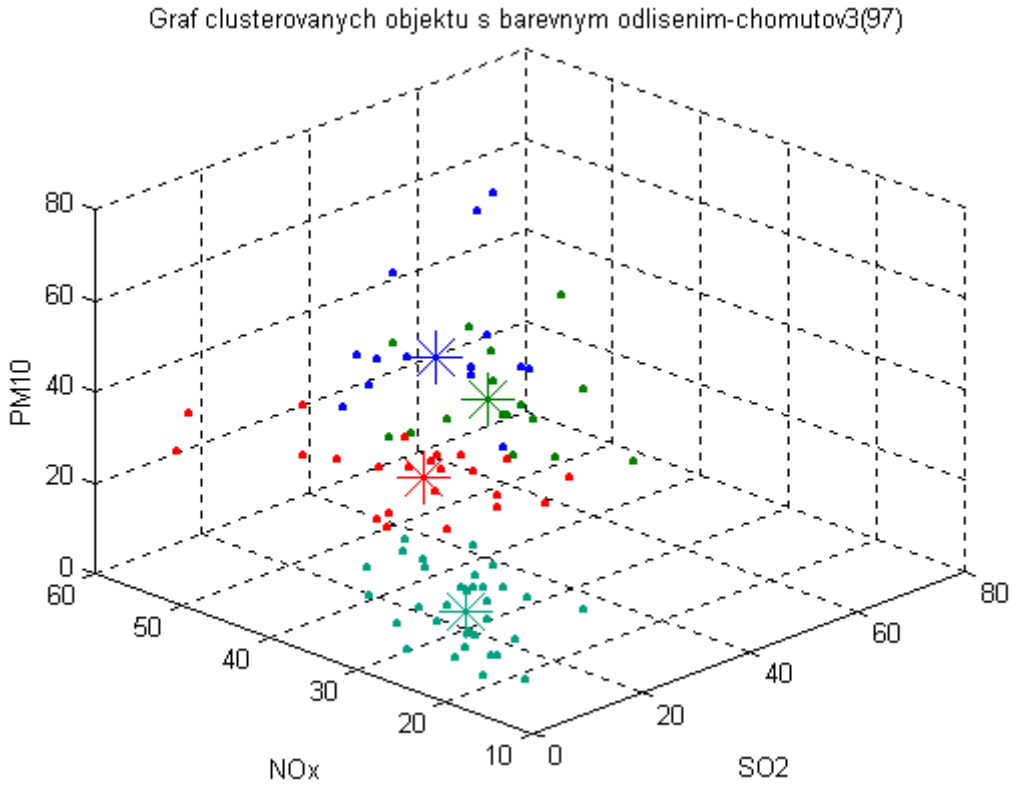


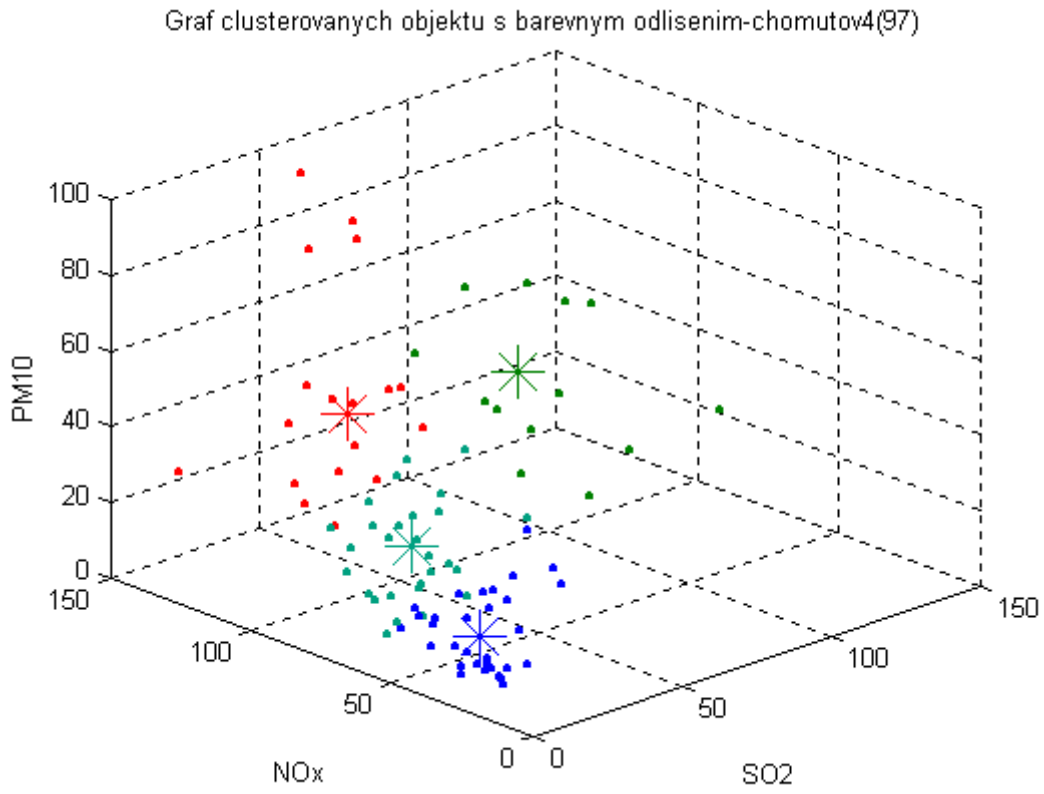
3D GRAFY CLUSTEROVANÝCH OBJEKTŮ

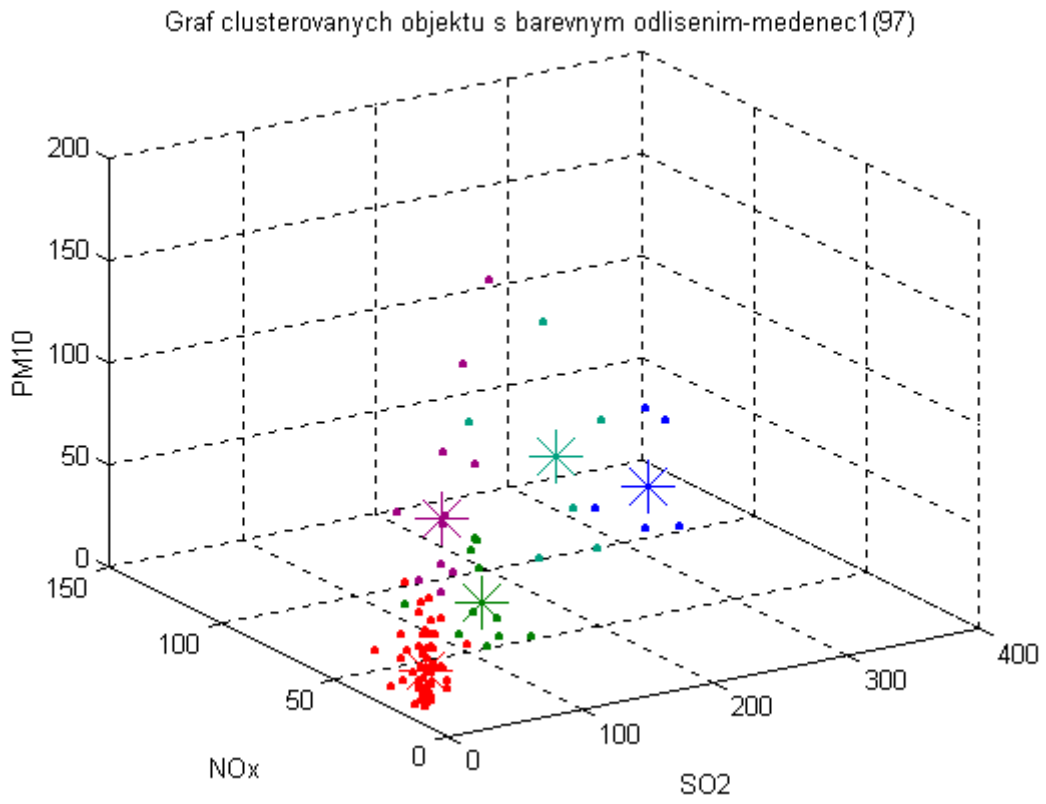
Prostředí Matlabu umožňuje zobrazení 3-dimenzionálních grafů s barevným odlišením jednotlivých clusterů. 3D grafy pomáhají při grafické interpretaci výsledků. Osy grafů představují koncentrace polutantů. Clusterovaný objekt je barevně zvýrazněný bod v grafu. Každá barva označuje jeden cluster.

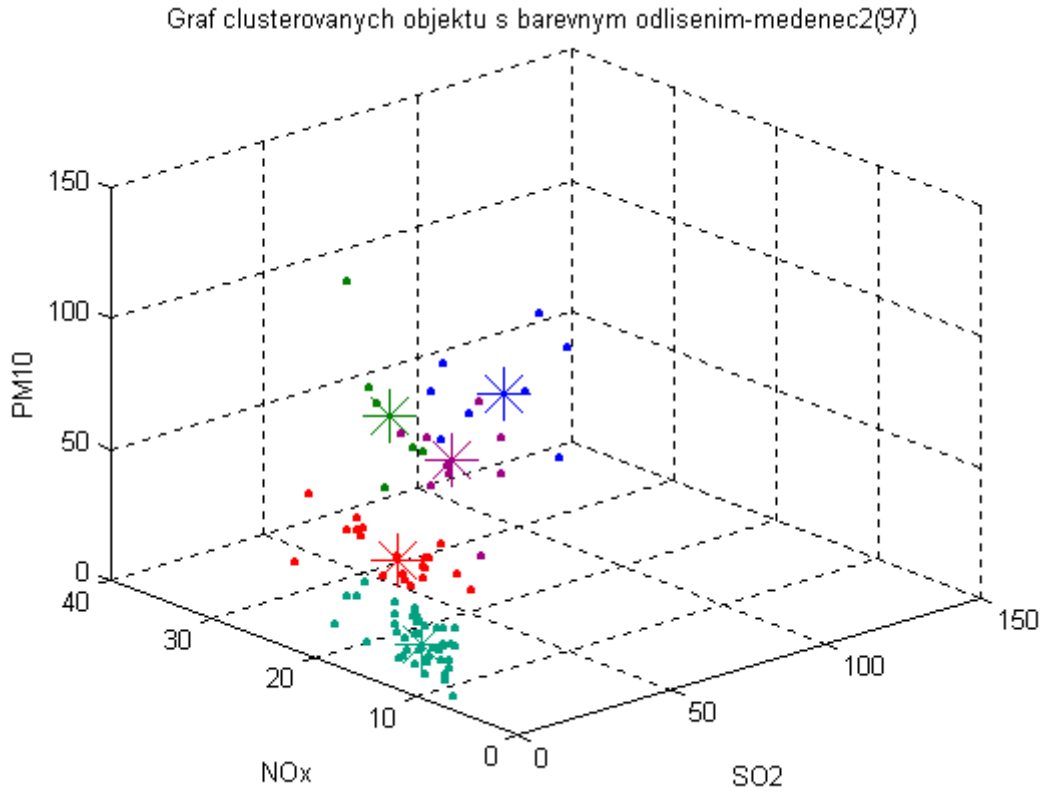


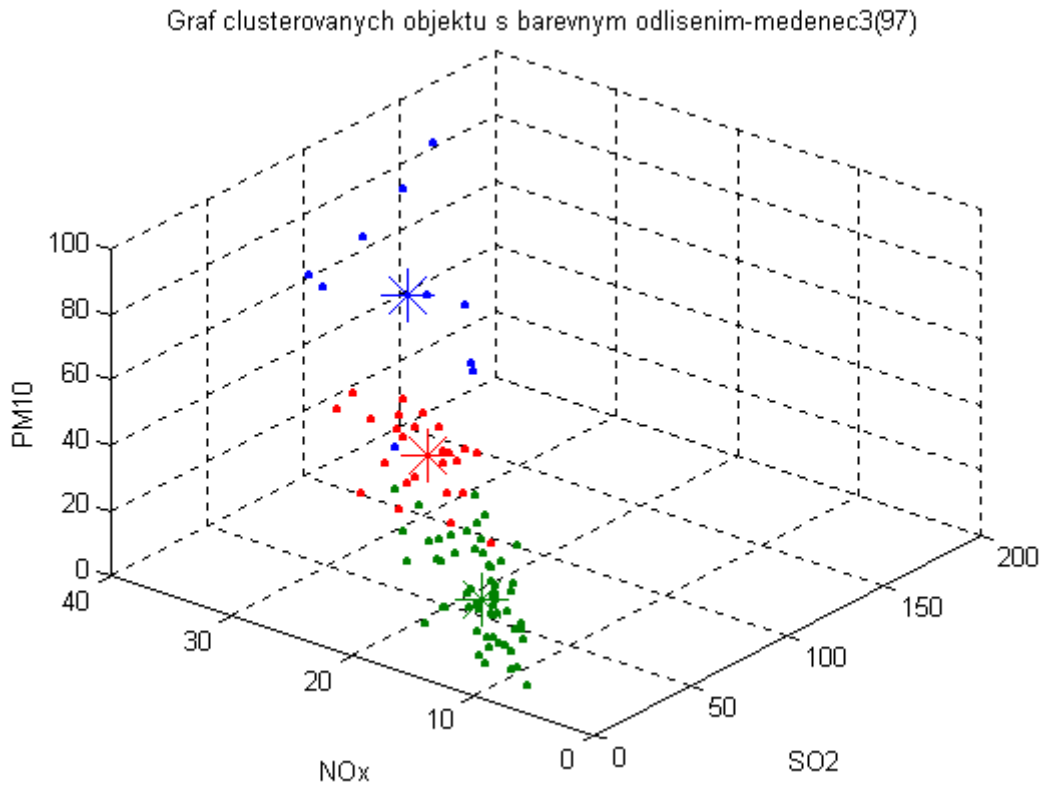


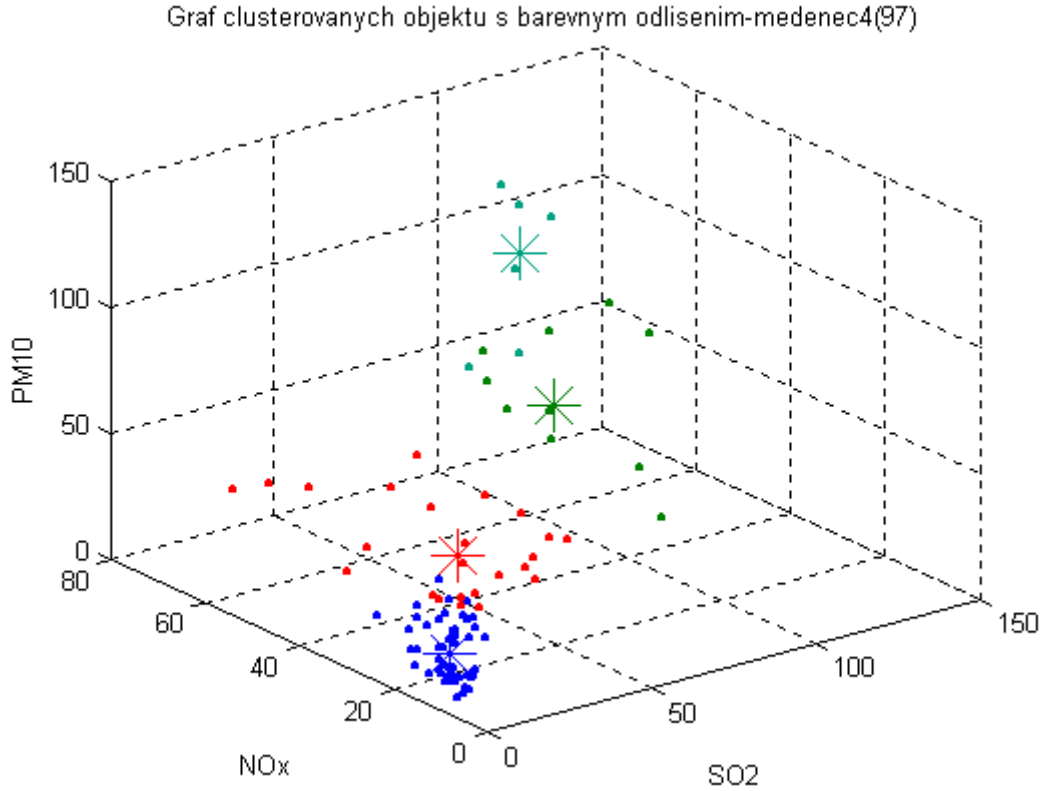

Graf clusterovanych objektu s barevnym odlisenim-chomutov3(97)
PM10
0
20
40
60
80
NOx
60
50
40
30
20
10
SO2
0
20
40
60
80


Graf clusterovanych objektu s barevnym odlisenim-chomutov4(97)
PM10
0
20
40
60
80
100
NOx
150
100
50
0
SO2
0
50
100
150


Graf clusterovanych objektu s barevnym odlisenim-medenec1(97)
PM10
200
150
100
50
0
150
100
50
0
NOx
0
100
200
300
400
SO2


Graf clusterovanych objektu s barevnym odlisenim-medenec2(97)
PM10
150
100
50
0
40
30
20
10
0
NOx
0
50
100
150
SO2


Graf clusterovanych objektu s barevnym odlisenim-medenec3(97)
PM10
100
80
60
40
20
0
40
30
20
10
0
NOx
0
50
100
150
200
SO2


Graf clusterovanych objektu s barevnym odlisenim-medenec4(97)
PM10
150
100
50
0
80
60
40
20
0
NOx
0
50
100
150
SO2

IDENTIFIKACE OBJEKTŮ

Nezbytnou součástí clusterové analýzy je identifikace jednotlivých bodů tak, jak byly zařazeny do clusterů. V následujících tabulkách jsou uvedeny všechny clusterované objekty s identifikací a příslušným clusterem, ke kterému patří dle maximální funkce příslušnosti. Ne všechny tabulky mají stejný počet řádků (počet objektů v clusterech se samozřejmě liší), a proto nejsou stejného rozměru. První sloupec každé tabulky "cl." udává číslo clusteru podle barevné identifikace 3D grafů. Každá tabulka je označena názvem datového souboru.

Měděnec1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	246,25	79,38	35,72	6	1
1	278,57	61,34	32,24	7	1
1	275,28	74,62	24,11	8	1
1	285,44	71,94	77,67	11	1
1	374,91	133,00	37,99	19	1

Měděnec1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
2	218,74	60,92	91,77	1	1
2	207,60	80,05	131,08	2	1
2	211,05	68,59	45,64	10	1
2	209,57	113,68	63,05	17	1
2	182,84	66,78	26,35	24	1
2	166,18	31,88	51,98	2	2

Měděnec1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
3	96,98	46,00	22,16	3	1
3	107,15	50,58	54,38	4	1
3	81,85	30,65	15,64	9	1
3	110,22	27,87	18,29	14	1
3	67,72	35,03	21,80	18	1
3	67,84	59,36	22,53	21	1
3	93,73	32,04	18,36	22	1
3	130,97	63,90	42,83	25	1
3	78,54	35,56	60,94	26	1
3	82,85	26,70	31,53	24	3
3	82,79	34,85	51,55	25	3

Měděnec1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
4	49,47	31,62	79,33	5	1
4	53,86	29,14	56,58	27	1
4	28,50	29,45	56,31	3	2
4	42,86	28,65	48,94	7	3
4	36,02	22,43	89,94	8	3
4	73,34	30,89	105,97	9	3
4	34,20	22,70	121,63	10	3
4	111,02	47,34	182,45	11	3
4	81,67	41,53	147,99	12	3
4	2,24	23,47	96,34	13	3
4	13,95	11,60	75,09	31	3

Měděnec1_97						Měděnec1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
5	4,71	10,12	24,18	22	2	5	31,61	25,37	16,41	12	1
5	17,61	17,53	37,95	23	2	5	7,12	17,33	16,87	13	1
5	20,30	23,61	49,34	24	2	5	23,58	17,30	21,63	15	1
5	2,42	13,09	9,93	25	2	5	16,90	17,32	22,03	16	1
5	0,48	9,65	11,14	26	2	5	20,07	44,01	15,07	20	1
5	5,10	11,42	14,03	27	2	5	17,54	19,92	24,50	23	1
5	4,35	11,84	16,95	28	2	5	8,43	25,40	23,12	28	1
5	8,51	16,42	30,67	1	3	5	6,87	16,67	14,82	29	1
5	8,85	11,88	34,95	2	3	5	2,59	9,37	11,84	30	1
5	2,30	10,05	14,74	3	3	5	10,05	13,01	17,83	31	1
5	14,64	14,05	24,07	4	3	5	45,99	18,75	28,33	1	2
5	17,09	21,35	35,54	5	3	5	8,40	23,04	27,55	4	2
5	7,65	24,86	35,07	6	3	5	4,91	14,10	11,89	5	2
5	1,64	20,02	64,07	14	3	5	7,89	12,69	22,95	6	2
5	2,05	21,21	15,25	15	3	5	16,66	21,23	25,43	7	2
5	7,00	13,87	8,93	16	3	5	4,50	13,52	20,06	8	2
5	6,54	15,39	24,92	17	3	5	5,36	10,47	37,32	9	2
5	21,69	18,65	35,95	18	3	5	14,87	21,62	46,33	10	2
5	9,72	14,70	33,30	19	3	5	6,30	11,66	9,94	11	2
5	15,92	12,03	48,70	20	3	5	4,54	13,04	6,47	12	2
5	13,64	18,09	39,49	21	3	5	3,59	16,86	5,55	13	2
5	13,30	10,34	25,66	22	3	5	4,95	10,65	12,39	14	2
5	16,81	10,52	15,35	23	3	5	6,32	11,30	18,27	15	2
5	23,89	22,39	51,96	26	3	5	20,33	12,19	17,68	16	2
5	3,88	14,67	35,15	27	3	5	34,37	28,04	36,32	17	2
5	3,48	26,83	8,51	28	3	5	13,91	23,75	15,75	18	2
5	8,95	13,52	16,08	29	3	5	8,85	17,84	12,72	19	2
5	7,12	8,09	27,32	30	3	5	1,65	10,43	13,73	20	2
						5	1,24	13,28	15,46	21	2

Měděnec2_97						Měděnec2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	98,22	25,59	33,68	8	4	2	24,50	16,63	75,07	1	4
1	121,56	31,80	58,10	9	4	2	25,53	24,44	128,40	2	4
1	81,99	32,27	36,47	21	4	2	15,06	17,56	62,99	25	4
1	76,95	31,75	57,00	13	5	2	9,36	17,33	103,84	3	5
1	125,12	35,52	64,22	10	6	2	24,87	21,20	86,75	15	5
1	84,21	32,76	63,81	11	6	2	29,18	19,03	71,45	16	5
1	93,85	33,09	40,74	27	6						
1	89,49	26,17	61,22	28	6						

Měděnec2_97						Měděnec2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
3	4,42	17,14	56,29	3	4	3	22,63	14,38	44,05	18	5
3	38,05	16,06	18,62	7	4	3	24,97	16,56	34,06	26	5
3	7,06	22,60	56,37	10	4	3	23,19	16,32	27,85	1	6
3	3,59	22,94	31,25	11	4	3	25,95	13,78	32,21	2	6
3	17,51	22,07	39,47	17	4	3	31,29	18,02	29,97	8	6
3	13,59	19,36	42,50	22	4	3	14,31	17,43	29,98	12	6
3	14,40	15,57	33,28	23	4	3	10,76	15,15	42,39	16	6
3	8,25	18,18	49,00	24	4	3	13,43	14,46	31,01	17	6
3	5,53	10,82	45,77	1	5	3	11,63	14,57	33,80	18	6
3	17,64	14,62	36,65	4	5	3	21,85	21,78	39,05	25	6

Měděnec2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
4	1,23	11,55	12,78	4	4
4	2,59	15,50	11,47	5	4
4	3,95	8,76	14,14	6	4
4	3,00	8,55	18,89	12	4
4	3,56	7,13	22,57	13	4
4	1,78	18,39	14,56	14	4
4	6,36	13,85	19,53	15	4
4	3,89	10,88	15,12	16	4
4	3,74	17,89	25,28	18	4
4	6,55	9,26	18,88	19	4
4	11,10	11,23	20,27	20	4
4	17,18	20,98	16,01	26	4
4	8,56	13,63	14,04	27	4
4	4,32	13,09	18,57	28	4
4	0,32	11,77	11,79	29	4
4	11,15	15,43	19,10	30	4
4	1,07	10,23	27,57	2	5
4	14,95	14,49	21,96	5	5
4	1,35	9,53	8,78	6	5
4	15,13	13,62	16,22	7	5
4	6,74	8,72	13,04	19	5

Měděnec2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
4	2,30	7,80	11,02	20	5
4	2,65	7,12	3,38	21	5
4	1,28	10,37	11,09	22	5
4	6,19	10,18	15,84	23	5
4	7,66	8,45	25,77	24	5
4	2,54	9,84	12,87	27	5
4	8,51	9,93	16,59	28	5
4	9,07	10,03	22,78	29	5
4	4,00	8,31	6,92	30	5
4	3,16	7,41	22,77	31	5
4	4,47	11,22	27,03	3	6
4	2,84	10,84	21,29	9	6
4	8,36	14,46	26,61	13	6
4	3,12	11,70	14,17	14	6
4	1,82	10,78	28,92	15	6
4	2,45	10,19	17,02	20	6
4	5,21	10,38	10,94	21	6
4	7,71	9,03	9,71	22	6
4	5,06	11,07	15,20	23	6
4	4,88	11,21	20,66	24	6
4	29,06	23,69	13,22	26	6
4	4,39	9,94	13,78	30	6

Měděnec2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
5	69,94	24,83	65,83	14	5
5	41,92	23,98	65,32	17	5
5	51,34	18,99	22,64	25	5
5	56,56	23,69	44,87	4	6
5	54,59	25,41	56,67	5	6
5	44,90	22,03	47,08	6	6
5	66,30	21,68	44,78	7	6
5	75,93	24,52	50,42	19	6
5	46,72	21,02	55,51	29	6

Měděnec3_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	107,42	27,64	64,97	17	7
1	104,81	26,56	47,19	12	8
1	79,52	35,03	70,20	18	8
1	80,93	29,32	27,67	19	8
1	66,91	34,24	78,94	20	8
1	54,29	22,46	90,66	25	8
1	70,66	30,32	93,70	26	8
1	101,27	31,83	97,64	27	8
1	80,45	22,97	61,47	1	9
1	152,91	37,66	88,54	3	9

Měděnec3_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
2	3,29	10,07	10,70	1	7
2	13,70	12,69	23,64	2	7
2	11,12	9,87	27,70	3	7
2	3,64	9,17	14,20	4	7
2	2,63	8,73	28,01	5	7
2	4,64	7,27	21,93	6	7
2	8,83	9,48	26,89	7	7
2	4,92	9,62	16,65	8	7
2	6,19	9,42	33,83	9	7
2	7,45	7,84	34,32	12	7
2	5,94	8,09	26,27	13	7
2	18,65	11,68	32,03	14	7
2	3,96	8,56	25,45	15	7
2	9,53	12,94	37,10	16	7
2	5,91	9,13	32,58	18	7
2	4,48	6,48	20,12	19	7
2	4,77	6,22	6,16	20	7
2	7,03	9,47	22,60	21	7
2	13,89	10,65	33,76	22	7
2	20,83	10,89	33,47	23	7
2	5,29	10,15	45,66	24	7
2	7,11	9,46	28,85	25	7
2	3,25	6,44	23,64	26	7
2	4,72	6,79	24,28	27	7
2	23,29	9,99	38,99	28	7
2	26,42	13,28	31,03	30	7
2	7,82	10,87	24,57	31	7
2	2,07	7,64	17,48	1	8

Měděnec3_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
2	11,95	8,66	29,76	2	8
2	8,22	10,97	48,81	13	8
2	5,17	9,75	53,97	14	8
2	8,54	11,15	57,00	15	8
2	6,63	11,47	46,14	23	8
2	4,94	10,66	42,10	24	8
2	1,26	8,02	17,98	29	8
2	1,32	7,09	11,02	30	8
2	12,44	10,46	20,42	31	8
2	9,70	15,97	47,58	5	9
2	1,88	12,11	45,95	6	9
2	1,81	8,65	18,91	7	9
2	1,84	9,77	25,76	8	9
2	1,63	9,83	18,85	9	9
2	2,54	9,37	9,83	10	9
2	22,98	16,46	32,89	11	9
2	8,44	13,94	33,62	12	9
2	1,36	6,60	12,01	13	9
2	2,47	7,17	16,03	14	9
2	7,19	11,38	22,72	15	9
2	14,32	17,66	27,10	17	9
2	5,29	14,52	39,99	18	9
2	3,37	14,40	15,64	19	9
2	27,90	14,64	18,06	20	9
2	8,67	13,70	19,53	23	9
2	5,64	13,28	20,89	24	9
2	21,54	16,40	27,31	26	9
2	9,97	18,07	50,09	29	9
2	6,65	16,86	39,97	30	9

Měděnec3_97						Měděnec3_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
3	50,25	18,74	35,70	10	7	3	29,34	17,50	63,65	16	8
3	37,58	18,50	49,21	11	7	3	62,27	29,79	49,26	17	8
3	40,02	18,38	51,50	29	7	3	37,93	22,07	59,44	21	8
3	43,36	16,54	52,81	3	8	3	28,61	18,63	67,35	22	8
3	32,76	16,95	34,38	4	8	3	31,27	22,24	46,66	28	8
3	27,50	12,91	34,64	5	8	3	51,39	23,92	57,80	2	9
3	43,71	19,12	37,01	6	8	3	22,31	19,25	60,95	4	9
3	47,96	20,13	47,69	7	8	3	27,16	20,36	35,90	16	9
3	42,39	25,11	52,90	8	8	3	66,66	31,82	40,51	21	9
3	41,04	17,08	54,27	9	8	3	62,06	24,62	29,94	22	9
3	37,97	22,33	54,71	10	8	3	51,28	27,46	24,66	25	9
3	44,78	21,87	54,09	11	8	3	17,93	18,21	49,27	27	9
3	29,34	17,50	63,65	16	8	3	28,18	15,71	56,10	28	9

Měděnec4_97						Měděnec4_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	4,92	13,48	20,42	1	10	1	5,72	11,82	24,65	2	11
1	2,83	9,83	12,32	2	10	1	24,55	20,20	19,56	3	11
1	2,31	6,45	11,32	3	10	1	16,43	18,15	17,50	9	11
1	4,97	8,44	37,58	4	10	1	10,88	22,21	27,68	13	11
1	10,50	11,69	37,90	5	10	1	11,19	16,01	35,10	14	11
1	16,44	15,12	18,62	8	10	1	1,67	16,28	11,75	15	11
1	2,24	8,31	27,62	9	10	1	7,60	14,12	32,13	16	11
1	2,35	9,89	12,70	10	10	1	22,72	18,31	18,33	17	11
1	2,59	7,06	8,27	11	10	1	4,04	19,23	22,75	21	11
1	4,10	6,67	9,52	12	10	1	6,02	14,36	14,22	1	12
1	2,02	7,78	6,84	13	10	1	15,16	20,86	21,27	2	12
1	4,70	11,39	14,02	14	10	1	9,61	29,83	17,55	3	12
1	6,51	12,34	12,31	15	10	1	6,80	19,72	13,35	4	12
1	12,88	13,17	29,09	16	10	1	4,58	10,56	15,19	5	12
1	5,67	16,66	26,32	20	10	1	3,18	12,41	13,18	6	12
1	13,44	19,70	39,31	22	10	1	5,53	12,95	20,36	7	12
1	3,57	11,41	15,23	23	10	1	16,80	26,35	17,34	9	12
1	2,59	9,13	13,86	24	10	1	6,89	20,85	12,29	10	12
1	3,80	9,68	10,74	25	10	1	4,33	9,94	10,10	11	12
1	5,24	6,54	15,52	26	10	1	6,36	10,54	10,29	12	12
1	9,42	13,47	25,66	27	10	1	9,31	8,89	13,92	13	12
						1	11,18	8,02	27,00	14	12
						1	4,35	11,89	8,60	24	12
						1	5,49	16,25	7,40	25	12
						1	7,06	14,77	8,13	26	12
						1	3,67	6,12	15,24	27	12
						1	5,88	7,13	16,35	28	12
						1	7,31	9,09	12,07	29	12

Měděnec4_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
2	58,74	40,63	83,99	7	10
2	125,16	60,92	74,22	30	10
2	89,51	48,86	85,35	31	10
2	65,89	41,33	69,88	4	11
2	75,54	39,37	66,79	20	11
2	96,65	67,98	59,19	22	11
2	115,35	47,87	23,72	24	11
2	126,16	53,42	68,49	16	12
2	82,44	43,89	49,56	19	12
2	90,73	26,14	30,92	31	12

Měděnec4_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
4	47,68	36,85	96,73	17	10
4	66,22	40,12	125,92	18	10
4	84,74	45,39	135,65	19	10
4	53,35	30,13	105,55	1	11
4	83,02	50,92	135,92	5	11
4	89,21	59,20	134,57	6	11

Měděnec4_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
3	30,52	21,66	64,55	6	10
3	46,90	22,43	25,19	21	10
3	35,51	27,05	19,63	28	10
3	34,32	28,89	29,87	7	11
3	45,68	29,33	20,99	8	11
3	25,96	23,66	21,13	10	11
3	41,82	21,17	32,63	18	11
3	55,86	31,78	39,63	19	11
3	43,03	44,66	56,32	23	11
3	30,14	32,38	15,83	25	11
3	50,59	47,22	30,63	26	11
3	49,29	54,74	33,25	27	11
3	18,79	66,78	32,35	28	11
3	19,53	43,34	19,63	29	11
3	31,75	27,67	18,92	30	11
3	25,22	27,79	20,40	8	12
3	46,69	19,15	44,84	15	12
3	33,49	27,83	39,44	17	12
3	66,45	29,32	27,86	18	12
3	62,56	33,53	18,60	20	12
3	31,61	59,70	34,55	21	12
3	40,87	74,45	20,76	22	12
3	22,67	41,26	30,35	23	12
3	27,22	20,57	22,24	30	12

Chomutov1_97						Chomutov1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	36,54	45,58	50,18	5	1	1	38,02	87,84	47,33	23	2
1	43,94	44,03	41,09	22	1	1	33,90	98,23	58,53	24	2
1	53,75	70,41	48,31	23	1	1	37,24	74,17	21,42	28	2
1	44,94	66,77	53,14	4	2	1	22,55	72,04	27,27	1	3
1	21,18	117,52	57,11	7	2	1	24,74	80,17	56,01	5	3
1	12,54	83,50	34,43	8	2	1	44,25	65,15	44,87	7	3
1	14,75	63,13	49,85	9	2	1	39,96	56,91	60,94	9	3
1	36,50	60,00	37,95	10	2	1	45,19	68,70	78,59	10	3
1	32,79	64,19	50,99	17	2	1	40,73	34,46	58,08	24	3
1	17,81	72,94	37,85	18	2	1	56,48	70,70	88,86	25	3
						1	53,85	50,77	70,10	26	3

Chomutov1_97						Chomutov1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
2	18,22	46,94	19,38	28	1	2	21,31	37,89	34,13	2	3
2	12,34	31,82	12,72	29	1	2	3,73	40,26	13,87	3	3
2	6,36	24,98	9,69	30	1	2	9,63	55,03	36,47	6	3
2	7,48	27,78	10,44	31	1	2	4,23	40,31	52,82	13	3
2	24,44	36,52	16,64	1	2	2	6,41	38,56	43,74	14	3
2	9,35	42,46	13,24	5	2	2	3,24	24,67	13,20	15	3
2	13,49	62,04	25,39	6	2	2	8,03	28,05	15,93	16	3
2	11,25	36,64	10,80	11	2	2	5,46	32,26	20,27	17	3
2	20,88	53,45	12,43	12	2	2	19,00	31,98	34,51	18	3
2	10,42	27,73	6,07	13	2	2	16,03	44,94	21,04	19	3
2	5,10	44,82	11,29	14	2	2	14,82	34,93	29,09	20	3
2	14,83	26,99	15,10	15	2	2	22,86	40,55	38,29	21	3
2	6,17	19,62	15,18	16	2	2	14,51	28,13	21,56	22	3
2	7,94	27,86	11,11	19	2	2	17,78	20,39	26,55	23	3
2	7,53	36,44	11,38	20	2	2	11,24	40,95	33,12	27	3
2	3,23	56,26	16,94	22	2	2	3,64	15,07	10,30	28	3
2	19,08	29,43	9,70	25	2	2	6,87	14,08	10,74	29	3
2	4,45	22,11	7,82	26	2	2	16,63	16,44	19,27	30	3
2	5,77	32,87	7,80	27	2	2	7,17	31,13	41,80	31	3
2	21,31	37,89	34,13	2	3						

Chomutov1_97						Chomutov1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
3	91,52	45,92	40,89	3	1	4	110,70	109,56	109,07	10	1
3	127,38	78,07	64,45	4	1	4	115,99	106,50	85,85	11	1
3	100,33	64,62	31,94	8	1	4	119,09	105,29	68,38	19	1
3	68,54	53,00	47,47	9	1	4	80,85	156,69	94,13	20	1
3	129,08	66,00	33,07	12	1	4	75,30	95,49	66,56	21	1
3	80,88	77,01	47,14	25	1	4	69,86	135,47	83,66	24	1
3	66,21	35,99	38,50	26	1	4	88,13	89,21	80,52	27	1
3	101,79	68,21	39,59	2	2	4	84,33	96,38	115,82	3	2
3	66,32	66,46	22,14	4	3	4	40,96	83,97	115,78	11	3
3	70,25	61,05	55,54	8	3	4	91,81	76,06	97,45	12	3

Chomutov1_97						Chomutov1_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
5	230,51	106,79	118,65	1	1	6	345,56	239,85	73,07	15	1
5	212,80	107,79	145,01	2	1	6	463,27	270,33	108,52	16	1
5	195,91	81,81	34,71	6	1	6	414,95	213,48	62,84	17	1
5	179,56	95,59	53,39	7	1	6	341,53	188,42	96,07	18	1
5	319,84	107,14	23,48	13	1						
5	269,73	168,27	42,55	14	1						

Chomutov2_97						Chomutov2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	2,51	22,66	7,89	4	4	1	6,10	16,91	19,51	27	5
1	8,57	24,16	10,57	5	4	1	2,34	20,80	12,98	28	5
1	3,21	15,15	17,06	6	4	1	8,86	17,94	21,03	29	5
1	7,61	18,42	15,65	12	4	1	2,42	14,94	11,40	30	5
1	4,14	20,31	11,98	14	4	1	4,15	13,24	15,68	31	5
1	4,04	16,23	19,34	15	4	1	3,07	20,49	20,23	13	6
1	4,06	17,77	15,39	16	4	1	5,54	21,90	14,73	14	6
1	8,78	20,62	22,37	30	4	1	2,17	16,70	17,13	15	6
1	11,78	17,70	16,59	6	5	1	5,66	24,74	16,10	20	6
1	1,44	26,78	16,69	22	5	1	11,30	26,99	7,58	22	6
1	6,97	23,14	19,69	23	5	1	10,73	23,85	15,82	23	6
1	11,45	17,12	21,73	24	5	1	4,51	20,31	13,98	24	6

Chomutov2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
2	61,53	55,53	40,30	8	4

2	73,89	63,04	60,74	9	4
2	64,93	55,31	47,17	21	4

Chomutov2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
3	10,30	44,00	89,75	2	4
3	11,28	37,05	61,13	3	4
3	8,17	30,97	57,03	10	4
3	15,60	54,66	49,36	24	4
3	9,72	28,02	48,88	25	4
3	16,48	29,23	50,74	26	4
3	17,11	51,39	43,73	27	4
3	8,39	30,89	61,96	3	5
3	14,91	32,76	81,76	16	5
3	25,16	24,01	73,80	17	5
3	15,11	23,98	51,72	29	6

Chomutov2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
4	21,99	29,96	27,59	17	4
4	19,20	39,70	41,64	22	4
4	32,22	27,15	23,39	7	5
4	36,51	21,38	26,58	8	5
4	22,30	21,16	46,97	18	5
4	31,11	28,89	29,33	26	5
4	20,15	27,87	36,27	6	6
4	24,38	33,57	34,48	12	6
4	22,62	26,62	26,68	16	6
4	23,15	25,66	30,60	17	6
4	33,60	26,39	21,03	18	6
4	33,74	30,75	34,33	19	6
4	37,68	21,64	15,17	21	6
4	26,06	32,90	31,42	25	6
4	34,90	39,84	28,00	27	6

Chomutov2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
5	18,51	32,37	19,04	7	4
5	6,81	20,26	28,33	11	4
5	12,24	20,63	22,83	13	4
5	13,91	28,76	18,76	20	4
5	11,92	41,40	34,66	23	4
5	11,47	36,67	20,65	28	4
5	13,00	30,33	14,24	29	4
5	4,92	29,79	29,88	1	5
5	5,91	34,52	33,45	2	5
5	11,38	36,33	30,53	4	5
5	6,62	28,74	23,04	5	5
5	19,01	23,61	18,75	9	5
5	14,82	29,08	21,18	10	5
5	12,47	30,75	32,48	11	5
5	7,67	23,22	27,34	12	5
5	7,34	29,86	21,65	19	5
5	16,06	26,16	21,10	20	5
5	8,74	28,52	18,49	21	5
5	14,93	26,59	20,47	25	5
5	7,48	25,80	28,98	3	6
5	7,30	22,70	27,10	8	6
5	4,72	22,88	25,66	9	6
5	16,65	27,40	19,50	26	6

Chomutov2_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc
6	32,91	49,63	59,12	1	4
6	48,96	37,13	43,42	13	5
6	43,19	28,62	30,62	1	6
6	42,34	34,46	36,11	2	6
6	37,95	38,74	47,81	4	6
6	36,04	24,90	63,97	5	6
6	37,60	28,39	46,16	7	6
6	37,12	32,53	42,92	10	6
6	35,50	34,32	50,09	11	6
6	37,97	32,36	38,35	28	6

Chomutov3_97						Chomutov3_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	41,77	35,94	43,02	17	8	2	8,64	31,65	46,01	1	7
1	66,46	55,21	57,28	18	8	2	35,36	28,41	67,74	2	7
1	28,83	46,41	38,28	19	8	2	18,65	24,69	51,33	3	7
1	33,89	45,16	42,83	20	8	2	29,95	31,18	41,56	17	7
1	41,34	36,54	43,33	22	8	2	42,12	24,34	30,90	30	7
1	58,44	52,19	59,14	27	8	2	31,43	26,75	34,93	6	8
1	44,71	42,65	44,83	3	9	2	35,71	31,78	37,82	7	8
1	39,20	37,37	26,34	12	9	2	29,13	32,31	48,74	8	8
1	45,18	57,76	29,58	17	9	2	38,66	27,95	45,98	9	8
1	34,94	52,94	26,10	22	9	2	25,27	41,36	52,55	10	8
1	55,54	51,01	25,02	25	9	2	29,74	35,38	58,22	11	8
1	46,77	45,76	34,43	26	9	2	22,94	33,80	41,93	12	8
1	37,47	50,67	37,00	28	9	2	34,93	36,21	50,22	13	8
1	47,56	55,12	48,10	29	9	2	3,37	25,87	53,43	14	8
						2	14,19	20,76	47,23	16	8
						2	22,31	25,05	51,38	28	8

Chomutov3_97						Chomutov3_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
3	23,06	28,21	26,64	10	7	4	5,59	21,62	21,37	4	7
3	30,80	24,68	32,02	11	7	4	4,22	14,42	15,35	5	7
3	12,52	34,02	25,94	23	7	4	15,41	13,66	17,54	6	7
3	17,48	31,77	30,20	24	7	4	7,50	15,10	22,85	7	7
3	27,15	25,13	27,84	29	7	4	11,05	20,14	20,13	8	7
3	30,05	31,35	31,57	3	8	4	10,40	20,76	24,79	9	7
3	24,14	42,19	24,98	4	8	4	10,62	23,19	27,22	12	7
3	12,25	27,11	27,11	5	8	4	11,19	22,44	18,05	13	7
3	9,16	19,57	41,62	15	8	4	4,78	31,86	19,10	15	7
3	17,02	34,95	39,84	21	8	4	9,45	30,35	23,93	16	7
3	20,55	33,48	35,35	23	8	4	6,44	19,04	19,49	18	7
3	14,29	36,53	21,92	31	8	4	2,84	12,46	8,65	19	7
3	30,91	37,14	28,31	1	9	4	5,17	18,86	4,25	20	7
3	15,94	26,57	38,70	2	9	4	5,92	23,21	16,16	21	7
3	21,70	49,49	34,67	4	9	4	7,29	29,14	23,10	22	7
3	19,67	36,09	31,19	5	9	4	10,85	23,09	20,72	25	7
3	31,12	40,56	24,62	11	9	4	2,42	16,07	11,69	26	7
3	30,96	51,22	17,53	15	9	4	8,69	20,46	13,80	27	7
3	9,82	56,62	24,93	16	9	4	5,82	26,50	13,50	28	7
3	15,05	58,48	29,48	18	9	4	2,73	19,32	10,83	1	8
3	33,65	41,05	21,27	21	9	4	6,26	20,48	21,88	2	8
3	18,12	47,30	27,02	23	9	4	4,60	23,72	12,83	29	8
3	8,43	31,75	26,07	27	9	4	5,92	20,08	11,70	30	8
						4	3,06	24,00	25,00	6	9
						4	2,74	19,01	14,09	7	9
						4	2,73	18,82	14,88	8	9
						4	2,50	15,44	12,02	9	9
						4	1,68	25,37	6,84	10	9
						4	3,24	20,67	7,57	13	9
						4	3,55	27,63	9,86	14	9
						4	3,11	30,55	14,43	19	9
						4	10,89	24,00	16,31	20	9
						4	16,27	32,48	15,09	24	9

Chomutov4_97						Chomutov4_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
1	2,91	13,86	9,07	2	10	1	27,16	38,07	18,26	17	11
1	7,43	24,66	8,54	3	10	1	35,72	27,67	23,01	18	11
1	11,42	29,37	27,00	4	10	1	19,91	34,89	23,55	19	11
1	14,37	30,50	21,15	9	10	1	6,26	48,45	18,39	1	12
1	9,65	26,49	9,15	10	10	1	9,62	37,85	10,77	3	12
1	2,58	17,76	11,89	11	10	1	4,63	28,36	13,10	4	12
1	5,17	17,37	9,52	12	10	1	6,33	41,41	17,74	5	12
1	3,11	28,69	7,37	13	10	1	9,42	45,35	14,19	12	12
1	7,45	33,02	7,35	14	10	1	16,30	18,87	9,49	13	12
1	5,49	52,57	12,43	15	10	1	16,64	22,39	17,35	14	12
1	4,87	21,15	13,00	24	10	1	28,73	23,19	30,25	17	12
1	4,43	15,57	9,63	25	10	1	26,20	29,68	39,37	18	12
1	5,27	25,21	10,58	26	10	1	7,01	24,63	8,29	25	12
1	15,55	39,83	15,59	3	11	1	11,65	21,32	9,01	26	12
1	10,07	36,83	24,34	13	11	1	4,66	40,97	10,98	27	12
1	23,30	31,80	27,24	16	11	1	6,73	47,11	16,65	29	12

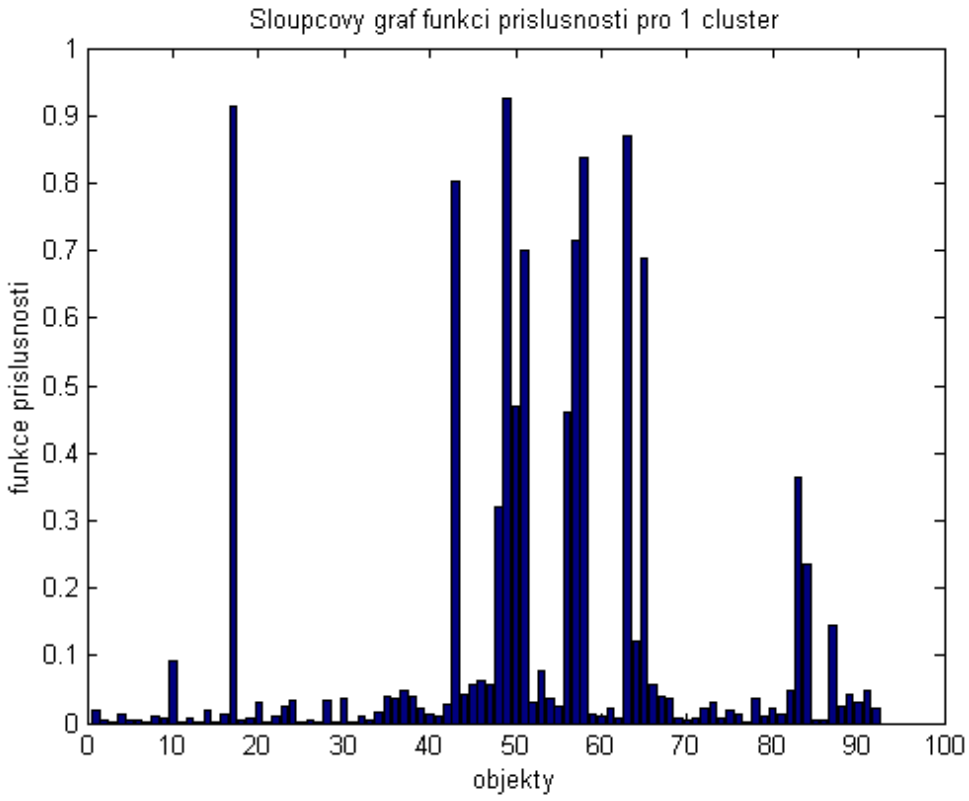
Chomutov4_97						Chomutov4_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
2	102,84	88,00	62,73	30	10	3	33,84	86,78	58,50	7	10
2	48,52	51,46	53,94	31	10	3	17,99	105,37	48,40	17	10
2	34,92	78,57	70,17	1	11	3	29,49	77,58	62,84	18	10
2	38,85	65,28	89,99	5	11	3	57,69	123,10	81,90	19	10
2	67,35	73,04	81,40	22	11	3	14,27	86,00	61,08	20	10
2	75,01	67,22	76,19	23	11	3	29,97	86,98	35,69	22	10
2	75,67	70,52	50,89	24	11	3	37,06	107,82	29,98	29	10
2	57,46	72,96	50,59	26	11	3	56,43	141,73	94,69	6	11
2	54,14	61,40	38,04	27	11	3	39,49	104,33	36,93	7	11
2	61,73	30,87	50,84	15	12	3	38,78	121,24	48,80	8	11
2	103,68	42,89	46,61	16	12	3	11,26	137,10	28,78	10	11
2	38,16	57,44	62,24	19	12	3	42,03	82,98	47,04	11	11
2	64,27	47,83	33,28	20	12	3	28,32	93,63	54,14	20	11
						3	36,78	102,36	97,94	21	11
						3	22,01	93,08	23,80	9	12
						3	24,24	110,19	29,23	22	12
						3	12,11	93,74	31,84	30	12
						3	36,81	117,84	86,28	31	12

Chomutov4_97						Chomutov4_97					
cl.	SO_2	NO_x	PM_{10}	den	měsíc	cl.	SO_2	NO_x	PM_{10}	den	měsíc
4	9,88	46,94	32,02	5	10	4	37,78	41,75	36,13	25	11
4	19,98	69,11	44,44	6	10	4	24,57	70,50	47,08	28	11
4	14,81	54,31	12,54	8	10	4	18,21	77,23	35,63	29	11
4	7,74	73,85	20,60	16	10	4	11,08	62,28	16,49	30	11
4	27,45	85,75	24,45	21	10	4	18,89	46,84	25,80	2	12
4	9,63	49,64	23,69	23	10	4	15,82	52,79	24,36	6	12
4	3,22	55,09	10,67	27	10	4	7,32	65,61	17,04	7	12
4	29,45	54,53	13,80	28	10	4	13,45	78,64	23,89	8	12
4	29,89	55,79	52,09	2	11	4	22,92	53,85	24,17	10	12
4	26,22	60,37	39,97	4	11	4	14,36	70,77	12,24	11	12
4	27,59	70,11	25,10	9	11	4	19,03	67,35	31,90	21	12
4	29,81	64,68	33,40	12	11	4	17,46	90,10	25,43	23	12
4	16,91	68,48	28,68	14	11	4	15,63	56,97	19,02	24	12
4	23,03	66,53	33,64	15	11	4	5,19	53,46	13,46	28	12

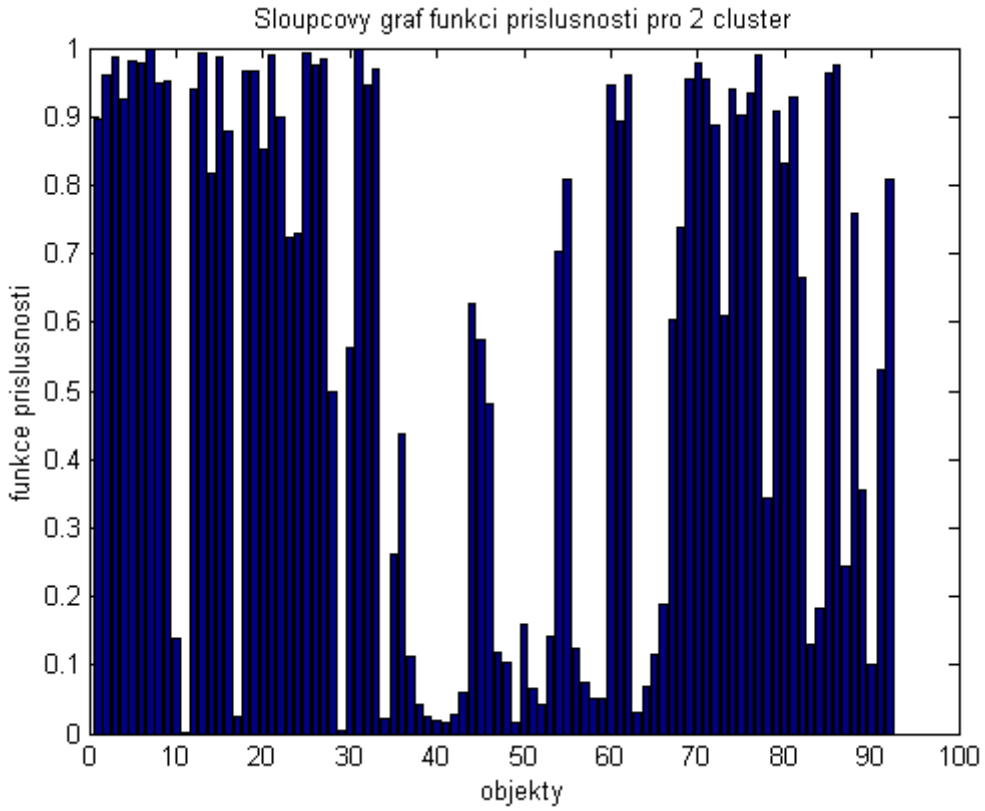
SLOUPCOVÝ GRAF-MĚDĚNEC3_97

Rychlou představu o tom, které objekty kam patří, lze získat ze sloupcového grafu. Z ilustrativních důvodů je uveden sloupcový graf pro medenec3_97, který obsahuje 3 clustery. Na ose „x“ jsou pořadí objektů od července do září 1997 postupně. Tento soubor jeví výrazné rysy linearity. Na jeho clusterovou analýzu by se spíš hodil algoritmus FCM-Gustafsson-Kessel. Je zřetelně vidět, jak 1. cluster má ze všech nejblíž k „hard“ rozkladu a tudíž je nejvěrohodnější.

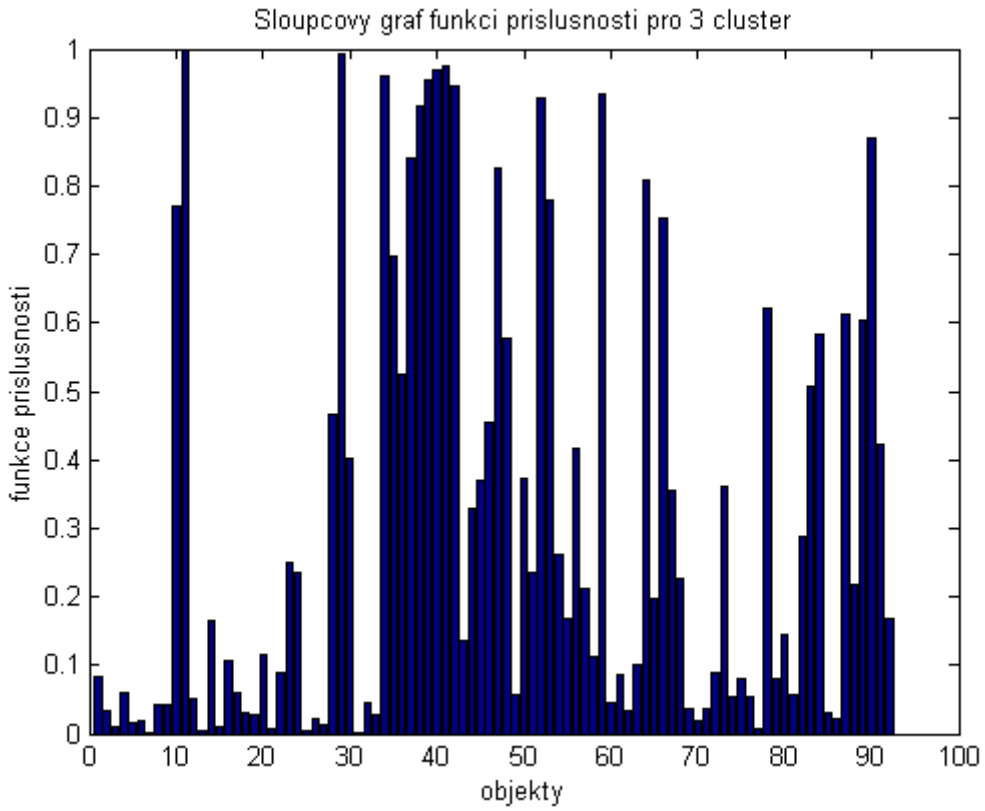
Měděnec3_97



Měděnec3_97



Měděnec3_97



3.6 HODNOCENÍ VÝSLEDKŮ

Z výsledků se určí objekty, které k sobě patří. Podle jejich identifikace, tj. datumu měření, je potřeba určit prvek, který měl vliv na jejich podobnost. Cílem je určit proč koncentrace polutantů v daných dnech jsou podobné. K tomu je potřeba předně znát lokální znečišťovatele, jejich emise v závislosti na čase, dále povětrnostní podmínky a chování ovzduší v okolí měřicí stanice. Tyto doprovodné údaje nebyly k dispozici a tak hodnocení výsledků se omezuje na výčet zřejmých faktů plynoucích z dat.

Měděnec a Chomutov se ve smyslu stejného počtu clusterů nechovají nejlépe. Z tohoto hlediska si sobě odpovídají pouze soubory Měděnec4_97 a Chomutov4_97. Krom indexu validity XB jsou průběhy ostatních indexů sobě velmi podobné. Stejně tak si odpovídají procentuální zastoupení říjnů, listopadů a prosinců ve všech clusterech. Z toho se dá vyvodit závěr, že povětrnostní podmínky a emise znečišťovatelů v oněch třech měsících byly stejné pro obě lokality.

Velice špatně porovnatelné jsou soubory Měděnec2_97 a Chomutov2_97. Počtem clusterů jsou rozdílné a navíc ani neexistují clustery, které by svým složením ve smyslu datumu odpovídaly. Procentuální zastoupení měsíců duben, květen a červen se v clusterech liší pro obě lokality. V měsících duben, květen a červen nelze určit jakékoli závislosti.

Drtivá většina února a března v souboru Měděnec1_97 leží v jednom, 5. clusteru. U souboru Chomutov1_97 jsou únory a březny výhradně ve dvou clusterech, 1. a 2. Navíc rozepsáním dnů únoru Chomutova1_97 je zřetelně vidět trend střídání cca 3 denní periody. Únor je mezi 1. a 2. cluster Chomutova1_97 rozložen v průměru po 3 dnech, tj. tři dny po sobě jsou v 1. clusteru a další 3 po sobě ve 2. clusteru atd.

Březen je naopak ve většině v 2. clusteru. V 1. clusteru je 8 březnů s výrazně vyššími koncentracemi všech polutantů a to na začátku měsíce a na konci. Větší část března, a to dni ze střední části měsíce, leží ve 2. clusteru. Zprvu to ukazuje, že u stanice Chomutov došlo pravděpodobně v únoru k periodicitám v povětrnostních podmínkám, kdy se střídaly vyšší koncentrace s nižšími po cca 3 dnech. Aritmetické průměrůy ukazují, že střídání vyšší a nižší koncentrace bylo pro všechny polutanty.

Aritmetické průměry únorů a březnů 1. clusteru Chomutov1_97 jsou:

	SO_2	NO_x	PM_{10}
únor	28,9662	78,8284	44,8592
březen	40,96713	62,362	60,59025

Aritmetické průměry únorů a březnů 2. clusteru Chomutov1_97 jsou:

	SO_2	NO_x	PM_{10}
únor	10,928	37,014	12,727
březen	11,189	32,401	27,194

U souboru Měděnec3_97 leží celý červenec ve 2. clusteru. U Chomutova3_97 tato tendence již není tak zřetelná, přesto větší část července leží ve 4.clusteru. Aritmetické průměry července napovídají, že chování povětrnostních podmínek a ostatních činitelů bylo stejné z hlediska SO_2.

Aritmetické průměry července stanice Měděnec a Chomutov

	SO_2	NO_x	PM_{10}
červenec Měděněc3_97	8,862	9,453	26,830
červenec Chomutov3_97	7,665	21,178	17,984

4. ZÁVĚR

Práce se zabývá použitím metody clusterové analýzy fuzzy C-means při interpretaci údajů monitorovací sítě ČHMU. Práce svou teoretickou částí dostatečně vysvětluje algoritmus a problematiku validity clusterů. Vychází z původních článků a postihuje nejdůležitější indexy validity. Během vypracování :

- byla rozebrána problematika validity clusterů
- byly popsány nejdůležitější indexy validity
- byl detailně rozpracován zdrojový kód FCM v prostředí Matlab
- byla provedena clusterová analýza na datech dvou vybraných stanic monitorovací sítě ČHMU
- byly částečně interpretovány výsledky

5. LITERATURA

[1] Zadeh L. : A. Fuzzy Sets, Inform. Control., Vol. 8 (1965) 338-353

[2] Meloun M., Militký J. : *Chemometrie-zpracování experimentálních dat na IBM-PC*, SNTL, Praha 1990

[3] Ruspini E. : A New Approach to Clustering, Inf. Control., Vol. 15 (1969) 22-32

[4] Ruspini E. : Numerical Methods for Fuzzy Clustering, Inf. Sci., Vol. 2 (1970) 319-350

[5] Ball G. H., Hall D. J. : A Clustering Technique for Summarizing Multivariate Data, Behav. Sci., Vol. 12 (1967) 153-155

[6] Duda R., Hart P. : *Pattern Classification and Scene Analysis*, Wiley New York 1973

[7] Dunn J. C. : A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well Separated Clusters, J. Cyber., Vol. 3 (1974) 32-57

[8] Shannon C. E. : A Mathematical Theory of Communication, Bell Syst. Tech. J., Vol. XXVII-3 (1948) 379-423

[9] DeLuca A., Termini S. : A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory, Inf. Control., Vol. 20 (1972) 301-312

[10] Bezdek J. C. : Cluster Validity with Fuzzy Sets, J. of Cyber., Vol. 3 (1974) 58-73

[11] Bezdek J. C. : A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 2 (1980) 1-8

[12] Zangwill W. : *Nonlinear Programming: A Unified Approach*, Englewood Cliffs, NJ:Prentice –Hall 1969 ch 4

[13] Tucker W. T. : Counterexamples to the Convergence Theorem for Fuzzy ISODATA Clustering Algorithms, in The Analysis of Fuzzy Information. Bezdek J. C., Ed. Boca Raton, FL: CRC Press Vol. 3 (1987) ch 7

[14] Bezdek J. C., Hathaway R. H., Sabin M. J., Tucker W. T. : Convergence Theory for Fuzzy C-Means: Counterexamples and Repairs, IEEE Trans. Systems, Man, and Cybernetics, Vol. 17 (1987) 873-877

[15] Hathaway R., Bezdek J. C., Tucker W. : An Improved Convergence Theorem for the Fuzzy C-Means Clustering Algorithms, in The Analysis of Fuzzy Information. Bezdek J. C., Ed. Boca Raton, FL: CRC Press Vol. 3 (1987) ch 8

[16] Bezdek J. C. : *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum 1981

[17] Bezdek J. C., Trivedi M., Ehrlich R., Full W. : Fuzzy Clustering: A New Approach for Geostatical Analysis, J. Math. Geo. (1980)

[18] Krishnapuram R., Keller J. :A Possibilistic Approach to Clustering, IEEE Trans. Fuzzy Systems, Vol. 1 (1993) 98-110

[19] Krishnapuram R., Keller J., The Possibilistic C-Means Algorithm: Insights and Recommendations, in A Possibilistic Approach to Clustering, IEEE Trans. Fuzzy Systems, Vol. 4 (1996) 358-393

[20] P. Barbieri,G. Adamia,A. Favretto,A. Lutmanc,W. Avoscan, E. Reisenhofer: Robust cluster analysis for detecting physico-chemical typologies of freshwater from wells of the plain of Friuli northeastern Italy, Analytica Chimica Acta 440 (2001) 161–170

[21] Rousseeuw P. J., Kaufman L., *Finding Groups in Data*, J. Wiley and Sons 1990

[22] Pekka Teppola, Satu-Pia Mujunen, Pentti Minkkinen: Adaptive Fuzzy C-Means clustering in process monitoring,Chemometrics and Intelligent Laboratory Systems 45 1999 23–38

[23] Pekka Teppola,Satu-Pia Mujunen,Pentti Minkkinen:A combined approach of partial least squares and fuzzy c-means clustering for the monitoring of an activated-sludge waste-water treatment plant, Chemometrics and Intelligent Laboratory systems 41 1998 95–103

[24] Moldan B. : *Geochemie atmosféry*, Academia, Praha 1977

[25] Xie X. L., Beni G. : A Validity Measure for Fuzzy Clustering, IEEE Trans. Pattern Anal. Machine Intell., Vol. PAMI-13,no. 8 (1991) 841-847

[26] Bezdek J. C., Pal N. R. : On cluster validity for the fuzzy C-means model, IEEE Trans. Fuzzy Systems, Vol. 3, no. 3 (1995) 375-379

[27] Rousseeuw P. J., Kaufman L., Trauwaert F. : Fuzzy Clustering Using Scatter Matrices, Comp. Statistics and Data Anal., Vol. 23 (1996) 135-151

[28] Windham M. P. : Cluster Validity for the Fuzzy c-Means Clustering Algorithm, IEEE Pattern Analysis and Machine Intelligence, Vol. PAMI-4 (1982), 357-363.

[29] Dunn J. C., Well Separated Clusters and Optimal Fuzzy Partitions, J. Cybern, Vol. 4-1 (1974) 95-104

[30] Gustafson E. E., Kessel W. C. : Fuzzy Clustering with a Fuzzy Covariance Matrix, IEEE CDC, San Diego, California (1979) 761-766

[31] Lukasová A., Šarmanová J. : *Metody shlukové analýzy*, SNTL, Praha 1985

[32] Adámek J. : *Kódování a teorie informace*, Ediční středisko ČVUT, Praha 1991

[33] Vysoký P. : *Fuzzy řízení*, Ediční středisko ČVUT, Praha 1996

[34] Obroučka K. : *Látky znečišťující ovzduší*, VŠB Ostrava, Ostrava 2001

[35] Bartovský T. : *Analyzátory emisí*, VUSTE SERVIS s.p., Praha 1994

[36] Zimmermann H.J : *Fuzzy set theory and its applications*, Kluwer, Boston 1994

[37] Materiály ČHMÚ

PŘÍLOHY

ÚPRAVA DATABÁZE

Veškeré výsledky prezentované v praktické části nebyly získány z upravené databáze jak je to popsáno dále. Tato úpravu může sloužit jako návod pro další práci.

Data, která se zjevně opakují, jsou problém. Došlo-li v průběhu dne k nesouvislému výpadku třeba 20 hodnot, což je poměrně mnoho vzhledem k 48 měřením, je lepší měření vynechat. Způsob jakým jsou data databáze ISKO spravována je poměrně složitý, chyby jsou odděleny od normálních dat. Není problém získat databázi v řádkovém tvaru po půlhodinách. Řádek (záznam) je den a sloupec (pole) je naměřená koncentrace. Následující část tabulky vše dokumentuje.

ECS	STA_NAZ	SPZ_OKR	VEL_NAZ	ROK	P_MESIC	P_DEN	H_00_00	H_00_30	H_01_00	H_01_30
192	Chomutov	CV	SO2	1997	1	1	70,000	28,000	43,000	73,000
192	Chomutov	CV	SO2	1997	1	8	2,000	3,000	2,000	3,000
192	Chomutov	CV	SO2	1997	1	9	72,000	65,000	42,000	70,000
192	Chomutov	CV	SO2	1997	1	10	77,000	91,000	87,000	96,000
192	Chomutov	CV	SO2	1997	1	11	137,000	141,000	136,000	138,000
192	Chomutov	CV	SO2	1997	1	12	111,000	113,000	114,000	121,000
192	Chomutov	CV	SO2	1997	1	13	101,000	121,000	110,000	99,000
192	Chomutov	CV	SO2	1997	1	14	308,000	284,000	312,000	328,000
192	Chomutov	CV	SO2	1997	1	15	242,000	233,000	237,000	231,000
192	Chomutov	CV	SO2	1997	1	16	229,000	237,000	223,000	203,000
192	Chomutov	CV	SO2	1997	1	17	-1	-1	-1	386,000
192	Chomutov	CV	SO2	1997	1	18	337,000	308,000	338,000	489,000
192	Chomutov	CV	SO2	1997	1	19	285,000	255,000	202,000	187,000
192	Chomutov	CV	SO2	1997	1	20	116,000	112,000	104,000	105,000
192	Chomutov	CV	SO2	1997	1	21	88,000	87,000	108,000	98,000
192	Chomutov	CV	SO2	1997	1	22	59,000	53,000	52,000	52,000
192	Chomutov	CV	SO2	1997	1	23	6,000	6,000	5,000	5,000
192	Chomutov	CV	SO2	1997	1	24	8,000	7,000	7,000	7,000

V poli VEL_NAZ je pouze SO_2, ale celá tabulka je velmi rozsáhlá, protože daná stanice měří všechny tři polutanty SO_2, NO_x a PM_{10}. P_MESIC je pro každý polutant od 1 do 12, to celé pro danou stanici STA_NAZ v daném okrese SPZ_OKR. Posledních 48 sloupců jsou půlhodinová měření. „-1“ značí nefunkčnost (nedošlá data). Např. pro 17. ledna je to v prvních čtyřech půhodinách 3-krát. Co když pro daný řádek se takto vyskytuje nesouvisle 20 mínus jedniček. Pak by bylo zřejmě lépe celý řádek vynechat.

Dále je uveden způsob jak tento problém vyřešit. Tento způsob byl vyzkoušen, avšak na vybraných místech Měděnci1_97 a Chomutovu1_97 neměl příliš význam, protože „-1“ se nevyskytovaly tolikrát. Pokud by byl zvolen počet mínus jedniček příliš malý na pro odstranění řádku, došlo by ke kontrakci dat a ztrátě informace v datech.

Nejprve je nutné převést databázi na sloupcový tvar. To není úplně jednoduché. Databáze se např. z Excelu vyexportuje do textového souboru „txt“ s pevnými oddělovači jako je mezera. Z tohoto souboru je možné potom číst programem v jazyce C data po řádku a převádět na sloupce. Na výslednou databázi už lze položit dotaz, který nechtěné řádky odstraní.

Zdrojový kód programu je:
(v programu předpokládám, že databáze nebude mít více než 55 sloupců, 48 za odběry a 7 na další sloupce)

```
#include <stdio.h>
#include <stdlib.h>

main()
{
	FILE *fr,*fw,*ff;
	int i=0;
	int k=0;
	int j=0;
	int m=0;
	char *identifikace[55];
	char radka[1000];
	int mezera=' ';
	int enter='\n';

	for (m=0;m<55;m++)
		identifikace[m]=(char*)malloc(11);
	if ((ff=fopen("test00.txt","wt+"))==NULL) {
		printf("soubor test00.txt se nepodarilo otevrit \n");
		return;
	}
	if ((fr=fopen("test0.txt","rt"))==NULL) {
		printf("soubor test0.txt se nepodarilo otevrit \n");
		return;
	}
	if ((fw=fopen("test1.txt","wt"))==NULL) {
		printf("soubor test1.txt se nepodarilo otevrit \n");
		return;
	}
	while(fgets(radka,1000,fr)!=NULL) {
		fputs(radka,ff);
		fseek(ff,0,SEEK_SET);
		while (j<55) {
			fscanf(ff,"%s",identifikace[j]);
			j++;
		}
		j=0;
		fseek(ff,0,SEEK_SET);
		for (k=0;k<48;k++) {		//49-ty je enter '\n'

			for (i=0;i<7;i++) {
				fputs(identifikace[i],fw);
				putc(mezera,fw);
			}
			fputs(identifikace[i+k],fw);
			putc(enter,fw);
		}
	}
	for (m=0;m<55;m++)
		free((void*) identifikace[m]);
	fclose(fr);
```

```
        fclose(ff);
        fclose(fw);
}
```

Obsah test00.txt např. může být:

```
192    Chomutov CV SO2 1997   1  1  70,000  28,000  43,000  73,000  39,000  64,000
```

Obsah test1.txt potom je:

```
192 Chomutov CV SO2 1997 1 1 70,000
192 Chomutov CV SO2 1997 1 1 28,000
192 Chomutov CV SO2 1997 1 1 43,000
192 Chomutov CV SO2 1997 1 1 73,000
192 Chomutov CV SO2 1997 1 1 39,000
192 Chomutov CV SO2 1997 1 1 64,000
```

Řekněme, že pole (sloupce) takto vzniklé databáze „TEST1“ pojmenujeme (viz. test1.txt):
ID, STA_NAZEV, SPZ_OKRES, VELICINA, ROK, MESIC, DEN, KONCENTRACE

Na takovou databázi se již může položit položit dotaz, který odstraní řádky s více „-1“, něž je požadováno. Pro ilustraci je požadováno odstranění řádků obsahujících „-1“ více než 5-krát.

```
select ID, STA_NAZEV, SPZ_OKRES, VELICINA, ROK, MESIC, DEN, avg(KONCENTRACE)
from TEST1
where KONCENTRACE <>-1
group by DEN, MESIC, STA_NAZEV, SPZ_OKRES, VELICINA
having count(VELICINA)>=(48-5)
ordered by VELICINA, SPZ_OKRES, STA_NAZEV, MESIC, DEN
```

ZDROJOVÝ KOD FCM

```
function [output]=fuzzyCMA(input)

global Fc
global Hc
global XB
global FS
global Obj
global Hc_norm2
global Hc_stand
global Fc_norm
global Fc_stand

clear XB Fc Hc FS Hc_norm1 Hc_norm2 Fc_norm Hc_stand Fc_stand Obj;

data      =  input.data;
n_clust   =  input.n_clust;

try, expon  =  input.expon;     catch,  expon  =  2;            end;
try, c_in   =  input.c_in;      catch,  c_in   =  [];           end;
try, Anorm  =  input.norm;      catch,  Anorm  =  eye(size(data,1)); end;
try, steps  =  input.steps;     catch,  steps  =  100;          end;
try, stop   =  input.stop;      catch,  stop   =  0;            end;
try, zerod  =  input.zerod;     catch,  zerod  =  0;            end;

[U,C,Obj,mess,Fc,Hc,XB,FS,stoping,Hc_norm1,Hc_norm2,Fc_norm,Hc_stand,Fc_
stand]=fcma(data, n_clust, expon, steps, stop, Anorm, c_in, zerod);

output.Fc=Fc;
output.Hc=Hc;
output.XB=XB;
output.FS=FS;
output.U=U;
output.C=C;
output.Obj=Obj;
output.norm=Anorm;
output.stoping=stoping;
output.Hc_norm1=Hc_norm1;
output.Hc_norm2=Hc_norm2;
output.Hc_stand=Hc_stand;
output.Fc_norm=Fc_norm;
output.Fc_stand=Fc_stand;

function
[U,C,Obj,mess,Fc,Hc,XB,FS,stoping,Hc_norm1,Hc_norm2,Fc_norm,Hc_stand,Fc_
stand]=fcma(data, n_clust, expon, steps, stop, A, c_in, zerod)

clear XB Fc Hc FS Hc_norm1 Hc_norm2 Fc_norm Hc_stand Fc_stand Obj;

C=cell(1,steps);
U=cell(1,steps);

PocetDat=size(data,2);
Dimenze=size(data,1);

Umat=rand(PocetDat,n_clust);
```

```
Umat=Umat./repmat(sum(Umat,2),1,n_clust);
Dist=ones(PocetDat,n_clust);

for iterace=1:steps

U{1,iterace}=Umat;
if ~isempty(c_in)
Cmat=c_in;
c_in=[];
else
Cmat=data*(Umat.^expon)./repmat(sum(Umat.^expon,1),Dimenze,1);
C{1,iterace}=Cmat; %Cmat je matice (Dimenze x pocet shluku)
end;

for cluster=1:n_clust

Dist(:,cluster)=sum( (data-repmat(Cmat(:,cluster),1,PocetDat))' *A .* (data-
repmat(Cmat(:,cluster),1,PocetDat))' ,2); %matice vzdalenosti

end

Dist=Dist.^0.5;

[idato_nula,jshluk_nula]=find(Dist<=zerod);
idato_nula_os=unique(idato_nula);
idato_spocti=setdiff([1:PocetDat]',idato_nula_os);
nulove_prvky=find(Dist<=zerod);
Umat(idato_spocti,:)=1./( Dist(idato_spocti,:).^(2/(expon-1)) .* repmat(
sum( 1./(Dist(idato_spocti,:) .^ (2/(expon-1)) ),2)  ,1, n_clust) );

Umat(idato_nula_os,:) = 0;
Umat(nulove_prvky) = 1;
Umat(idato_nula_os,:) = Umat(idato_nula_os,:) ./
repmat(sum(Umat(idato_nula_os,:),2),1,n_clust);

Obj(iterace)= sum(sum((Umat.^expon) .* Dist.^expon ));
for f=1:(n_clust-1)
for sloupec=(f+1):n_clust
pomoc(sloupec-f)=(norm(Cmat(:,f)-Cmat(:,sloupec)))^2;
end
norma(f)=min(pomoc);
end

nejmensi=min(norma);
ahoj=sum(Umat.^expon);
for g=1:n_clust
soucet(g)=ahoj(g)*((norm(Cmat(:,g)-mean(data')'))^2);
end
celkovysoucet=sum(soucet);

tebuh=0;
nazdar=0;
for g=2:n_clust
tebuh=tebuh+1/g^2;
nazdar=nazdar+1/g;
end
```

```
XB(iterace)=Obj(iterace)/(PocetDat*nejmensi);
Fc(iterace)=sum(sum(Umat.^2))/size(data,2);
Hc(iterace)=-sum(sum(Umat.*log(Umat)))/size(data,2);
FS(iterace)=Obj(iterace)-celkovysoucet;
Hc_norm1(iterace)=Hc(iterace)/log(n_clust);
Hc_norm2(iterace)=PocetDat*Hc(iterace)/(PocetDat-n_clust);
Fc_norm(iterace)=(n_clust/(n_clust-1))*(1-Fc(iterace));
Hc_stand(iterace)=(Hc(iterace)-nazdar)/(((1/PocetDat)*tebuh-(n_clust-
1)/(n_clust+1)*(pi^2-6)/(6*PocetDat))^0.5);
Fc_stand(iterace)=((PocetDat*(n_clust+2)*(n_clust+3)/(n_clust-
1))^0.5)*(((n_clust+1)*Fc(iterace)/2)-1);

if iterace > 1
stoping(iterace)=norm(U{1,iterace}-U{1,iterace-1});
if (stoping(iterace) < 0.01)
iterace
break;
end
end


end
```

ZDROJOVÝ KOD cluster_graf.m

```
function h=cluster_graf(in,out,IDdata)
global pp;
global id;
dimenze=input('Zadej dimenze v kterych chces vykreslit vysledek (napr. [2 3
5])    ');
if size(dimenze,2)==2
volba=input('zobrazit shluky (kovariancni matice vysledku) elipsoidem?  0=ne
1=ano ');
if volba
for cluster=1:in.n_clust
F{cluster}= ( repmat(out.U(:,cluster)'.^2,2,1) .* (in.data(dimenze,:)-
repmat(out.C(dimenze,cluster),1,size(in.data,2))) )...
* (in.data(dimenze,:)-repmat(out.C(dimenze,cluster),1,size(in.data,2)))' /
sum(out.U(:,cluster).^2);
end
h=cluster_color2F(out.U,out.C(dimenze,:),F,in.data(dimenze,:));
else
h=cluster_color2(out.U,out.C(dimenze,:),in.data(dimenze,:));
end

end

if size(dimenze,2)==3
h=cluster_color3(out.U,out.C(dimenze,:),in.data(dimenze,:),IDdata);
end
```

ZDROJOVÝ KOD cluster_color3

```
function h=cluster_color3(U,C,data,IDdata);
global pp;
global id;
[X,I]=max(U');

for ii=1:size(C,2)
   pp{ii}=data(:,find(I==ii));
   id{ii}=IDdata(:,find(I==ii));
end;

str='plot3(';
strC='plot3(';
for ii=1:size(C,2)
   str=sprintf('%spp{%d}(1,:),pp{%d}(2,:),pp{%d}(3,:),''.'',',str,ii,ii,ii);
   strC=sprintf('%sC(1,%d),C(2,%d),C(3,%d),''*'',',strC,ii,ii,ii);
end;
str =sprintf('%s,''MarkerSize'',5);',str(1:end-1));
strC=sprintf('%s,''MarkerSize'',20);',strC(1:end-1));
h=figure;
eval(str);
hold on;
eval(strC);
grid on;

title('Graf clusterovanych objektu s barevnym odlisenim')
xlabel('SO2')
ylabel('NOx')
zlabel('PM10')
```

ERRATA

str. 2 chemometrie

str. 6 Fuzzy clusterovou analýzou dat, reprezentovaných fyzikálně-chemickými vlastnostmi povrchových vod, se zabývají autoři P. Barbieri, G. Adamia, A. Favretto, A. Lutmanc, W. Avoscan, E. Reisenhofer[20].

Odběry probíhaly v roce 1996/1997.

str. 7 ….., který může reagovat s prachovými alkalickými částicemi v ovzduší za vzniku síranů.

str. 18 ***Emisní limit*** je nejvýše přípustné množství znečišťující látky vypouštěné ze zdroje znečišťování do ovzduší.

str. 20 Standardizace se provádí přes k-tý znak, kde k = 1…p. Vypočte se směrodatná odchylka pro jednotlivé sloupce (znaky) matice dat.

str. 64 Pro 30 minutová data….

Naskýtá se jeden problém, a to opakování dat popř. nesouvislý výpadek.

str. 84 Dle mého názoru by se měly případně brát v úvahu jejich lokální extrémy.

Vynechat souvětí:
Jako lokální extrém popsaných diskrétních charakteristik je bod, jehož nejbližší sousedi mají hodnotu charakteristiky menší (větší).