Základy zpracování geologických dat

závislosti ve vícerozměrných souborech dat

R. Čopjaková

Matice korelačních koeficientů

- Spočte sílu korelace (Pearsonův korelační koeficient) pro všechny páry
- Využití např. při vyšetřování substitucí v minerálech (ukázka pro plagioklasy)
- V excelu pomocí: Analýza dat-korelace

Matice Pearsonových koeficientů korelace

	Ca^{2+}	K^{+}	Fe^{2+}	Mg^{2+}	Al^{3+}	Sr^{2+}	Na^{+}	Si^{4+}	P^{5+}
Ca^{2+}	1,00								
K^{+}	-0,20	1,00							
Fe^{2+}	0,25	-0,46	1,00						
Mg^{2+}	-0,48	-0,24	0,38	1,00					
Al^{3+}	0,99	-0,15	0,18	-0,55	1,00				
Sr^{2+}	0,45	-0,15	0,12	-0,08	0,42	1,00			
Na^{+}	-0,97	-0,02	-0,14	0,54	-0,98	-0,42	1,00		
Si^{4+}	-0,99	0,18	-0,24	0,49	-0,99	-0,43	0,97	1,00	
P^{5+}	0,56	-0,05	-0,02	-0,31	0,55	0,05	-0,53	-0,58	1,00

Odstranění statisticky nevýznamných koeficientů korelace

	Ca^{2+}	K^{+}	Fe^{2+}	Mg^{2+}	Al^{3+}	Sr^{2+}	Na^{+}	Si^{4+}	P^{5+}
Ca^{2+}									
K^{+}	0,00								
Fe^{2+}	0,00	0,00							
Mg^{2+}	0,00	0,00	0,00						
Al^{3+}	0,99	0,00	0,00	0,00					
Sr^{2+}	0,00	0,00	0,00	0,00	0,00				
Na^{+}	-0,97	0,00	0,00	0,00	-0,98	0,00			
Si^{4+}	-0,99	0,00	0,00	0,00	-0,99	0,00	0,97		
P^{5+}	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	

Al, Ca ⟷ Na, Si

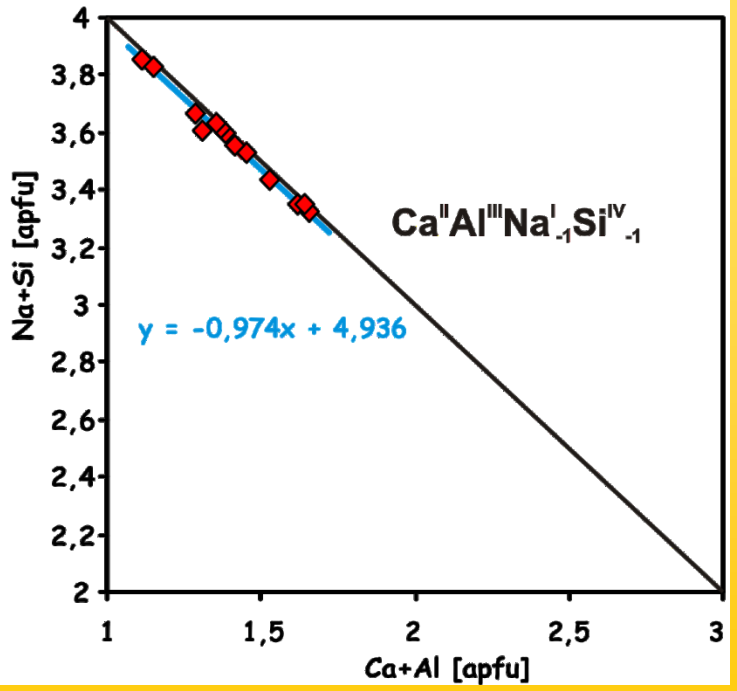
$Ca^{II}Al^{III}_2Si^{IV}_2O^{-II}_8$ anortit

$Na^{I}Al^{III}Si^{IV}_3O^{-II}_8$ albit

Ca-NaAl-Si

$Ca^{II}Al^{III}Na^{I}_{-1}Si^{IV}_{-1}$ substituční vektor

$Ca^{II}Al^{III} \leftrightarrow Na^{I}Si^{IV}$ substituční vektor



- Homovalentní substituce
  - zastupují se ionty o stejné valenci
  - např. $Mg^{2+} \leftrightarrow Fe^{2+}$ - v turmalínech, amfibolech, biotitech, olivínech
- Heterovalentní substituce
  - zastupují se ionty s různou valencí
  - musí být zajištěna elektrická neutralita struktury
  - Např.

    $Na^{+}\ Si^{4+} \leftrightarrow Ca^{2+}\ Al^{3+}$ v plagioklasech

    $Al^{3+}\ F^{-} \leftrightarrow Ti^{+4}\ O^{2-}$ v titanitech

    $Ca^{2+}\ Th^{4+} \leftrightarrow 2REE^{3+}$ v monazitech

Koeficienty podobnosti v geologii

- slouží pro stanovení míry podobnosti objektů (znaků).
- Koeficienty podobnosti lze dělit do tří skupin.

  - vlastní korelační koeficienty
  - asociační koeficienty
  - míry vzdálenosti

Vlastní korelační koeficienty

- **Pearsonův korelační koeficient**
- **Spearmanův koeficient pořadové korelace**
- **Binární korelační koeficient** r_D pro metalogenetické studie je počítán z kvalitativních geologických údajů. Spočtení síly závislosti mezi výskytem charakteristiky A a B.

kde a (b) je počet ložisek obsahujících charakteristiku A (B),
n je počet případů v souboru,
e_{ab} je počet případů obsahujících zároveň charakteristiky A a B.

Všechny korelační koeficienty nabývají hodnot $\langle -1;1 \rangle$

Asociační koeficienty

- Asociační koeficienty jsou založeny na počtu shod a rozdílů mezi znaky srovnávaných objektů. Užití pro binární data.
- Znaky považujeme za shodné, jsou-li jejich stavy u srovnávaného páru sobě rovny. Rozlišujeme shodu pozitivní (oba znaky mají hodnotu jedna) a negativní (oba znaky nulové).
- Asociačních koeficientů byla navržena celá řada, běžně se jich užívá v ložiskové geologii,
- Asociační koeficienty nabývají hodnot od nuly do jedné <0;1>.
- V geologických vědách patří mezi nejčastěji užívané koeficient Jaccardův a Sokalův-Michenerův. Volba mezi Jaccardovým a Sokal-Michenerovým koeficientem závisí na tom, jestli pro dané znaky má nebo nemá smysl negativní shoda

Kde n_{JK} - počet pozitivních shod mezi znaky jedinců J a K

n_{jk} - počet negativních shod mezi znaky jedinců J a K

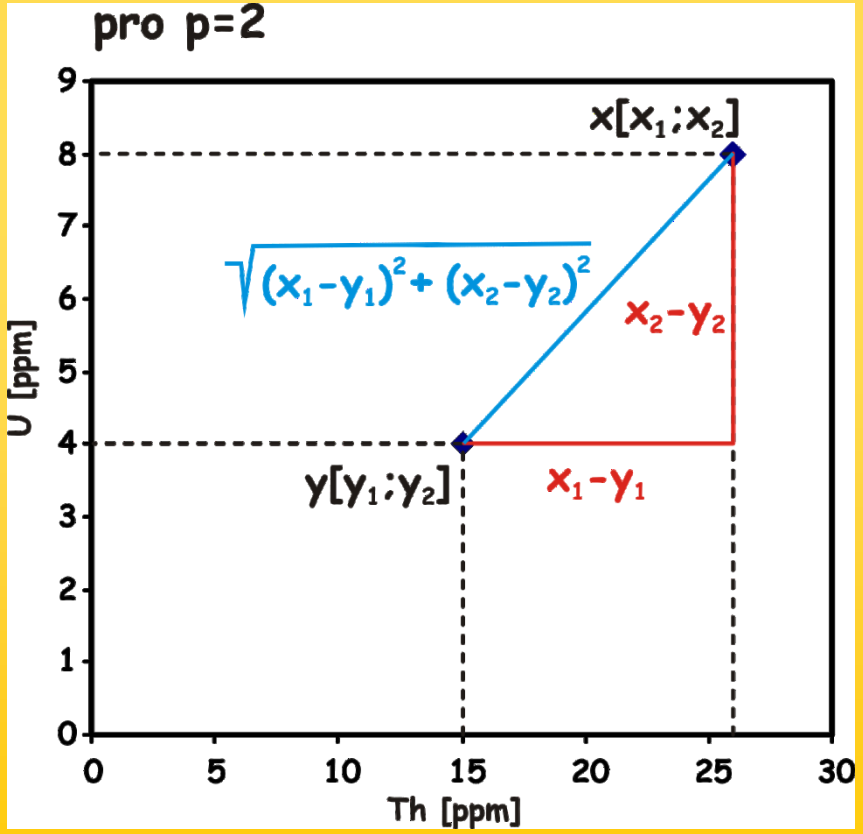
m - počet všech shodných párů znaků ($m = n_{JK} + n_{jk}$)

u - počet všech nesouhlasných párů znaků

n - celkový počet porovnávaných znaků mezi jedinci J a K

Míry vzdálenosti

- Míry vzdálenosti lze považovat spíše za míry nepodobnosti (dissimilarity).
- Nejčastěji se užívá **euklidovská vzdálenost**, kterou můžeme chápat jako geometrickou vzdálenost dvou bodů (jedinců *x*, *y* charakterizovaných pomocí *p* znaků) v *p*-dimenzionálním prostoru.



- **Vzdálenost Manhattan**
  snižuje vliv odlehlých hodnot

- **Čebyševova vzdálenost**

Multivariační statistické metody

- Multivariační statistické metody umožňují analýzu mnohorozměrných souborů dat.
- Příkladem takového vícerozměrného souboru dat může být např. chemická analýza hornin, kde každý objekt (horninový vzorek) je popsán pomocí *p* charakteristik (procentuálních obsahů hlavních oxidů).
- Pro zpracování vícerozměrných dat lze použít „tradiční" metody, např. chemické přepočty (CIPW klasifikace), znázornění pomocí trojúhelníkových diagramů, které však redukují počet charakteristik původních dat.
- Multivariační statistické metody pracují se všemi zadanými údaji a jejich matematicko-statistický aparát umožňuje nalézt a vyjádřit charakter variability objektů (proměnných), hlavní faktory, které tuto variabilitu způsobují, i kvantifikovat podíl jednotlivých faktorů na variabilitě.
- Obvykle předtím, než přistoupíme k multivariačním statistickým metodám, vyhodnotíme jednotlivé proměnné jednoduchými statistickými metodami (výpočet průměru, rozptylu, utvoření histogramu, testování typu rozdělení apod.).

Multivariační statistické metody

V rámci multivariačních statistických metod můžeme rozlišit:

- metody, které umožňují třídění objektů nebo znaků (zejména **shlukové analýzy**)
- metody provádějící zatřídění neznámých objektů do předem definovaných skupin (různé druhy **diskriminační analýzy**)
- metody snižující dimenzi prostoru a tvořící faktory ovlivňující variabilitu souboru (**faktorové analýzy**)

- Při vícerozměrných analýzách pracujeme se souborem dat obsahujícím *n* objektů (pozorování), kde každý objekt je charakterizován *p* znaky (proměnnými). Jednotlivé objekty si lze představit jako body v *p*-dimenzionálním prostoru, kde pozice bodu v prostoru je určena pomocí *p* úseků na jednotlivých souřadných osách.
- Můžeme rozlišit Q způsob analýzy, při níž sledujeme vztahy mezi *n* objekty popsanými prostřednictvím *p* znaků, a R způsob analýzy, kterým zkoumáme vztah mezi *p* proměnnými určenými pomocí *n* objektů.

Transformace dat

V případě geologických dat vybrané znaky či objekty nebývají vždy vzájemně souměřitelné, což může vést ke špatným výsledkům analýz. Některé možné příčiny:

- znaky mohou být měřeny v různých jednotkách
- značné rozdíly absolutních hodnot znaků (i řádové rozdíly)
- ovlivnění výsledků analýzy znakem, jenž má výrazně větší variabilitu (směrodatnou odchylku)

Tyto nedostatky lze odstranit transformací výchozích dat tak, aby veškeré znaky (objekty) měly stejnou váhu a variabilitu. Hovoříme o tzv. standardizaci (normování) dat. Existují různé způsoby transformace vstupních dat. Jedním z často užívaných způsobů je tato úprava:

kde x_{ij} je hodnota i-tého objektu a j-tého znaku, kterou chceme standardizovat, x_j je střední hodnota j-tého znaku, S_j je směrodatná odchylka j-tého znaku. Pro takto standardizovaná data je směrodatná odchylka rovna 1 a střední hodnota 0.

Užití chci-li dát stejnou váhu všem znakům (např. hlavním i stopovým prvkům)

Transformace dat

Některé další možné způsoby transformace dat (normování) jsou:

normování průměrem

normování dat s lognormálním rozdělením pravděpodobností

Časté jsou transformace dat v paleontologii a biologii, z důvodu vyhnutí se velikostním rozdílům způsobeným věkem jedinců - hojně užívaná standardizace dat užitím vzájemných poměrů měřených biometrických parametrů.

Shluková analýza

- shluková (sdružovací, klastrová) analýza se užívá při řešení klasifikace objektů (Q způsob) nebo jejich znaků (R způsob), vyjádřených kvantitativně nebo i kvalitativně

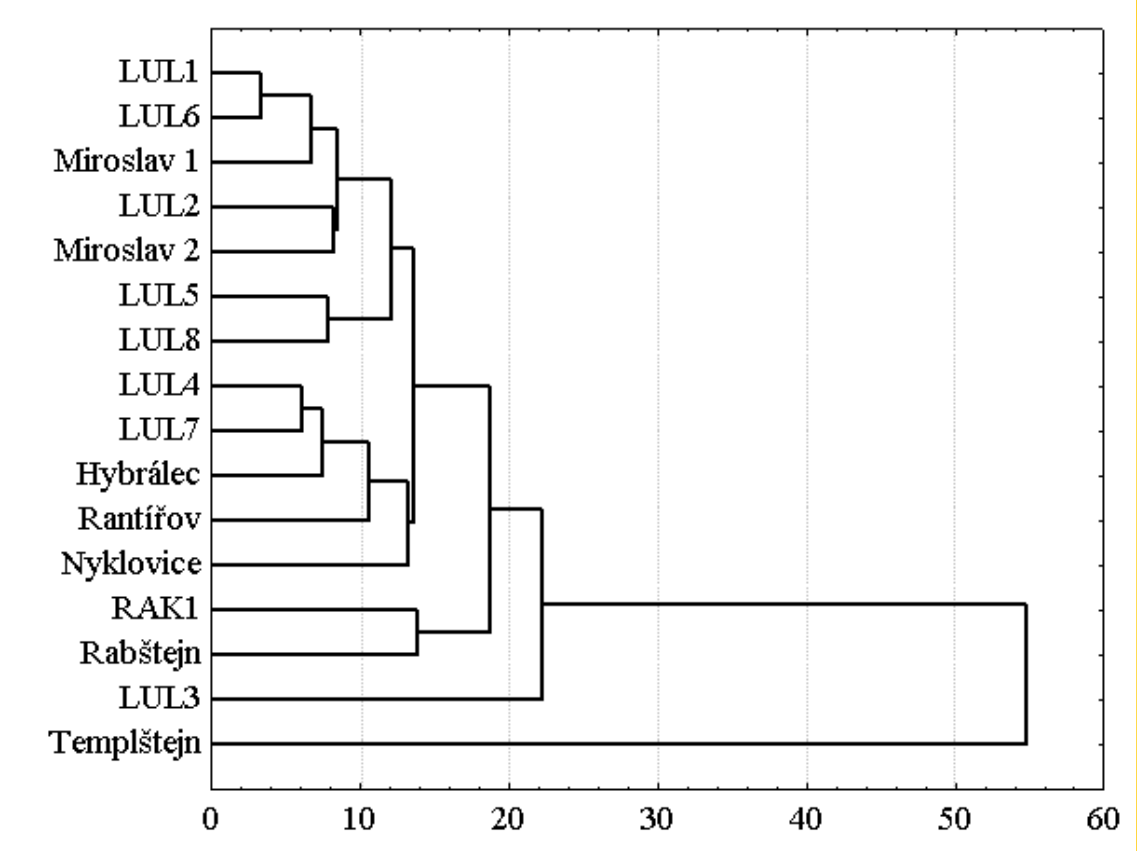
- V závislosti na výchozích datech se volí postup, jak sdružit jedince (znaky) do navzájem oddělených skupin pomocí některého z koeficientů podobnosti.

- Aglomerativní metody, pracují na principu postupného sjednocování shluků (častěji používané). V rámci aglomerativních metod se rozlišují tzv. *hierarchické metody* (optimalizují postup, kterým získáváme skupiny) a *nehierarchické metody* (optimalizují vnitřní homogenitu skupin).

Hierarchické shlukovací metody

- Na počátku sdružování jsou všechny objekty (znaky) považovány za jednoprvkové shluky. Na základě podobnosti (vzdálenosti) se slučují dva nejpodobnější (nejbližší) shluky, čímž se vytvoří shluk nový a k tomuto novému shluku je nutno přepočítat podobnosti ostatních nezměněných shluků. Proces slučování se opakuje tak dlouho, až jsou všechny objekty (znaky) sloučeny do jediného shluku.
- Několik různých strategií tvorby shluků.
- Použitý koeficient podobnosti může být různý - závisí na povaze dat. Mezi osvědčené koeficienty podobnosti v geologii patří euklidovská vzdálenost, Sokalův-Michenerův koeficient, či Manhattan vzdálenost.
- Výsledky se znázorňují graficky pomocí tzv. dendrogramů či dendrografů
- Dendrogram vytváří hierarchickou strukturu, kde taxony můžeme srovnávat na určité zvolené hladině podobnosti. Skupiny vzniklé na určité hladině podobnosti se snažíme geologicky interpretovat.
- Dendrograf je tvarem shodný s dendrogramem, avšak sledujeme jím nejen vztahy mezi skupinami, nýbrž i uvnitř skupin mezi jedinci. Ve směru jedné osy vytváří hierarchickou strukturu jako dendrogram, ve směru druhé osy ukazuje vzdálenost mezi jedinci uvnitř skupin.
- Hierarchická shlukovací analýza se často užívá jako výchozí metoda usnadňující orientaci ve vícerozměrných souborech dat. Z jejích výsledků mohou vycházet další metody, jejichž aplikace vyžaduje předběžnou znalost struktury souboru (např. nehierarchické sdružovací metody).

Hierarchické shlukovací metody



Dendrogram pro chemického složení granátu z valounů granulitů z lulečských slepenců a granulitů Českého masivu (použity hlavní i stopové prvky – jejich normalizované obsahy). Použita vzdálenost Manhattan a strategie nevážená pár-skupinová.

Nehierarchické shlukovací metody

- Vycházejí ze stanovení počátečního rozkladu souboru subjektivně, uživatelem. Na počátku se definuje *k* skupin

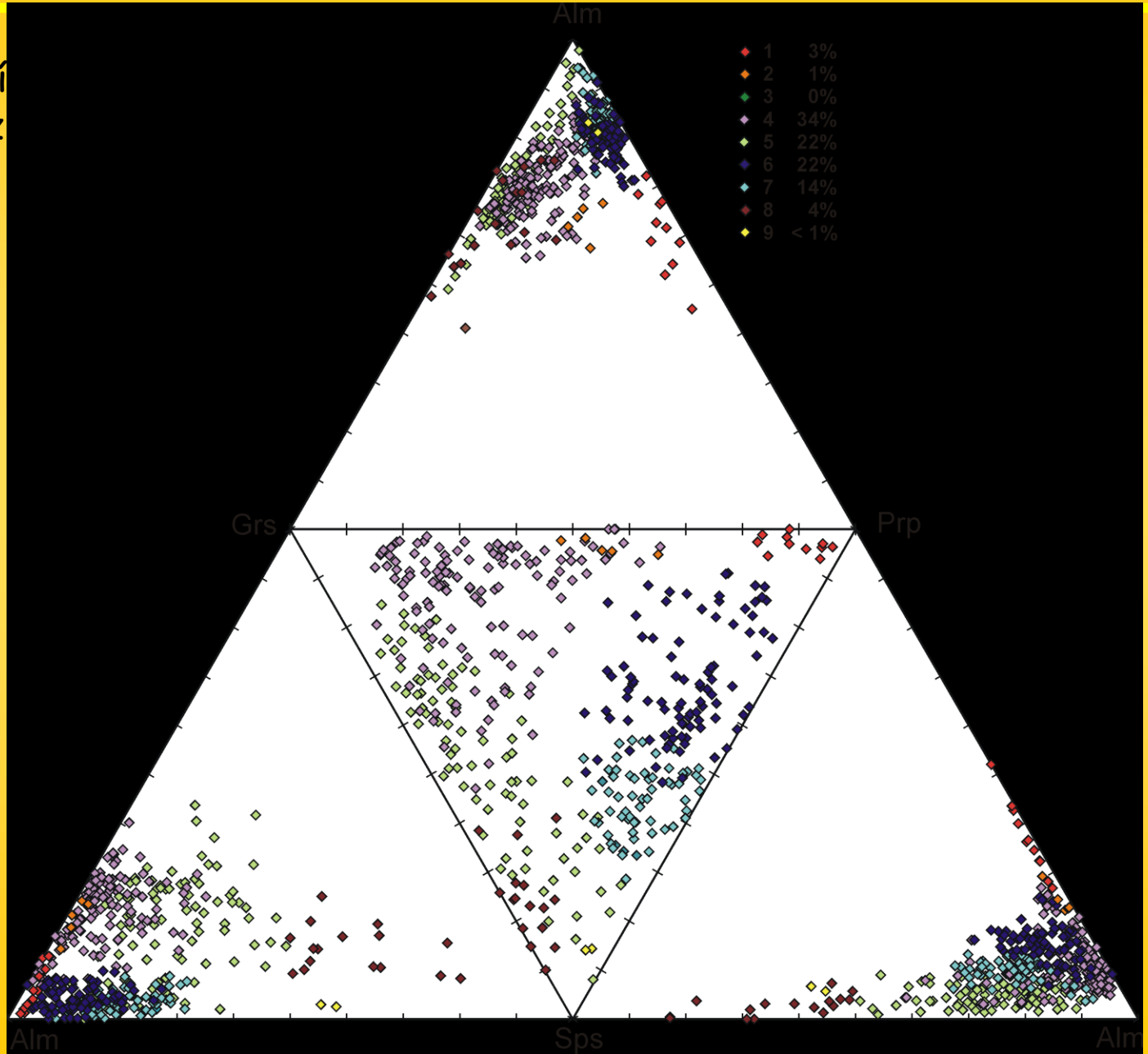
- Užívají se v případech, kdy počet shluků můžeme předběžně určit na základě dřívějších geologických zkušeností. V případě práce s neznámým souborem dat je nutno navázat na výsledky jiné metody vícerozměrné analýzy – např. dendrogramu.

Příklad výsledků nehierarchického sdružování. Zařazení vzorků asociací těžkých minerálů do tří klastrů s uvedením vzdálenosti každého vzorku od typického bodu, kde počet klastrů byl stanovený na základě studia tvaru dendrogramu.

vzorek č.	klastr	vzdálenost
2609	1	3.02
2427	3	1.75
2424	2	5.69
2377	2	2.96
2356	2	3.27
2171	3	1.78
2147	1	2.15
2119	3	0.96
1819	3	0.70
1726	2	3.55
1725	2	2.81
1605	1	1.42
1584	1	2.79
1583	1	2.97
1530	1	2.18
1529	1	1.60
1528	1	2.75
1527	1	2.92
1522	3	1.33
1521	3	1.27
1500	3	1.39
1499	3	0.64
1498	3	1.45
1497	3	1.43

Nehierarchické shlukovací metody

Vyčlenění granátů z souvrství



Příklady použití sdružovacích analýz v geologii

- klasifikace belemnitů rodu Duvalia na základě biometrických měření roster belemnitů
- rozlišení pectinoidních fosilních mlžů podle tvarové variability schránek
  klasifikace důlních otřesů na dole ČSA pro stanovení porušování horninového masívu (tzv. ploch porušení) s postupem důlních prací
- rozdělení pyroxenů a amfibolů hornin těšínitové asociace na základě chemického složení
- klasifikace šedých a pestrých vrstev OKR na základě chemického složení
- třídění fluviálních jeskynních štěrků na základě valounové analýzy
- vytvoření etalonů typů důlních vod pro jednotlivé oblasti OKR podle základních hydrogeochemických charakteristik
- klasifikace metamorfovaných hornin domažlické oblasti vycházející z analýzy geochemických dat