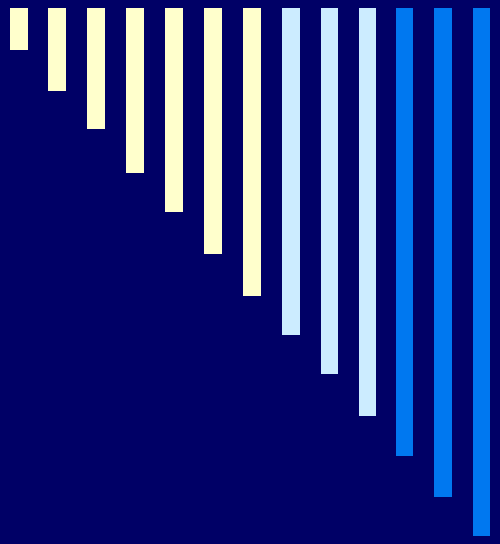


Profilování vzorků heroinu s využitím vícerozměrné statistické analýzy

Autor práce : RNDr. Ivo Beroun,CSc.

Vedoucí práce: prof. RNDr. Milan Meloun, DrSc.

PROFILOVÁNÍ

- Profilování = klasifikace a rozlišování jednotlivých vzorků ilegálně připravených drog na základě zjištění identity a relativního množství různých sloučenin zvolených poté jako markanty (v tomto případě významných doprovodných látek v heroinu – stopových organických nečistot, tzv. by-produktů, pocházejících z opia nebo vzniklých z původních obsahových nečistot v průběhu vlastní výroby heroinu z opia).
- Výskyt a množství doprovodných látek je do značné míry závislé na genetickém založení rostlin opiového máku a na podmínkách prostředí (půdní faktory, klimatické podmínky atd.)

DATA

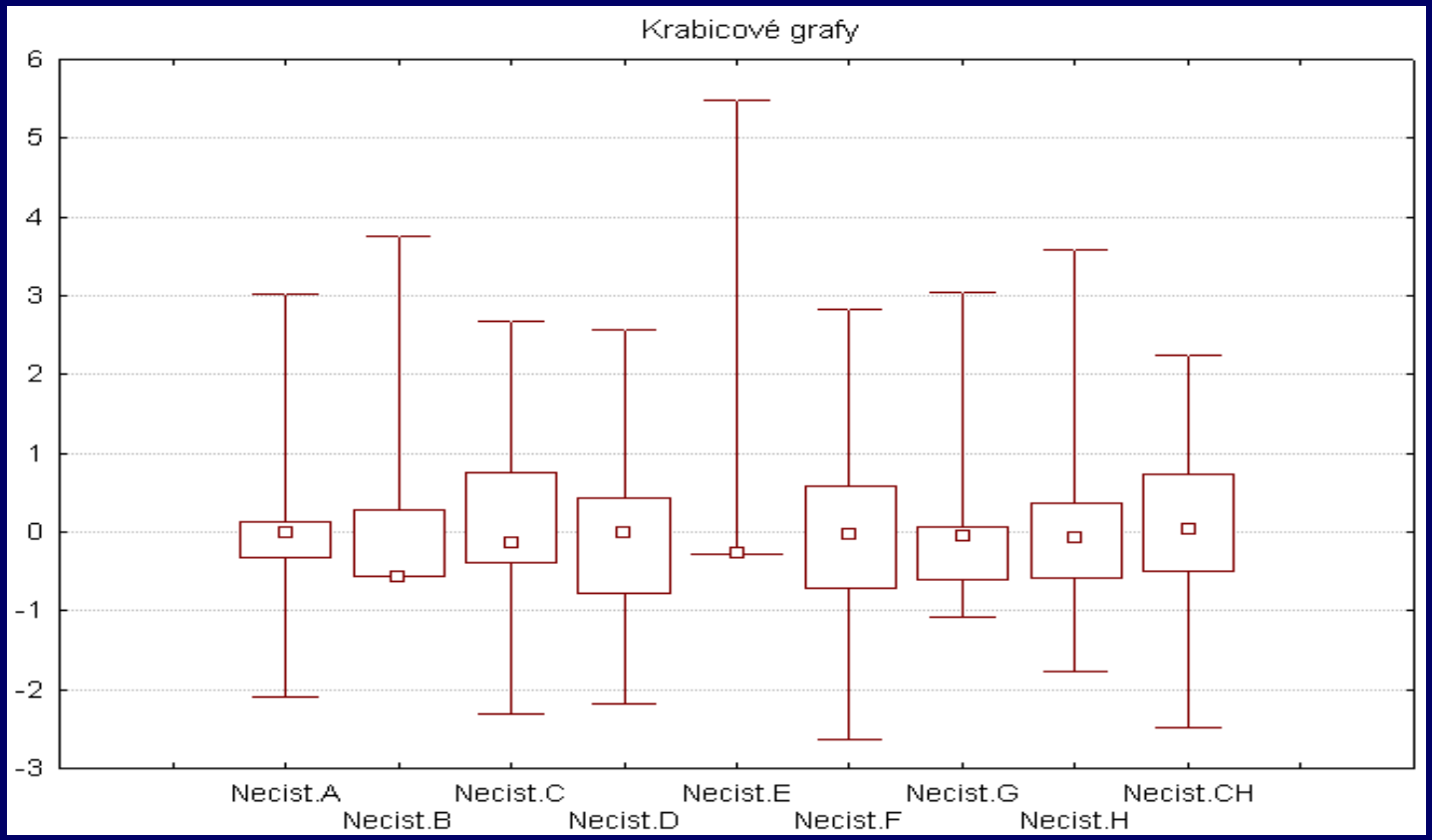
- Vstupní zdrojová matice obsahuje 41 řádků (tj. 41 reálných objektů – vzorků heroinu ze záchytů na území ČR) a 9 sloupců (tj. 9 znaků označených zde písmennými indexy „A“ až „CH“ – zvolených nejčastěji se vyskytujících nečistot, jejichž kvantitativní zastoupení odráží v matici uvedené plochy jejich chromatografických píků získané za shodných podmínek analýzy). Data jsou převzata se svolením autorů z práce realizované na KU Praha v roce 2004 s tím, že původní počet znaků (nečistot) 14 byl zredukován na 9 (byly zvoleny a zahrnuty pouze jednoznačně identifikované nečistoty metodou GC/MSD a nebyly použity ty, které jsou v původní práci zahrnuty s označením „neznámý“). Původní práce nebyla vyhodnocena s využitím vícerozměrné statistické analýzy (použito prvotní rychlé roztřídění dat na principu sortování – „setřásání“ dat, kdy jako kritérium podobnosti byla zvolena hodnota úhlu α mezi vektory - jedna n-tice ploch j-tého vzorku je porovnávána se všemi ostatními a na základě hodnoty α je vybrána nejpodobnější n-tice).

ŘEŠENÍ

- Zdrojová data byla statisticky zpracována s využitím softwarových programů „STATISTICA 6“, „NCSS 2002“ a „OPstat“ s následujícím strukturováním postupu:
- 1. GRAFY PŮVODNÍCH DAT
- 2. EXPLORATORNÍ ANALÝZA STRUKTURY OBJEKTŮ
- 3. VYČÍSLENÍ KORELAČNÍ MATICE
- 4. METODA HLAVNÍCH KOMPONENT (PCA)
- 5. FAKTOROVÁ ANALÝZA (FA)
- 6. ANALÝZA SHLUKŮ (CLU)
- 7. VYHODNOCENÍ

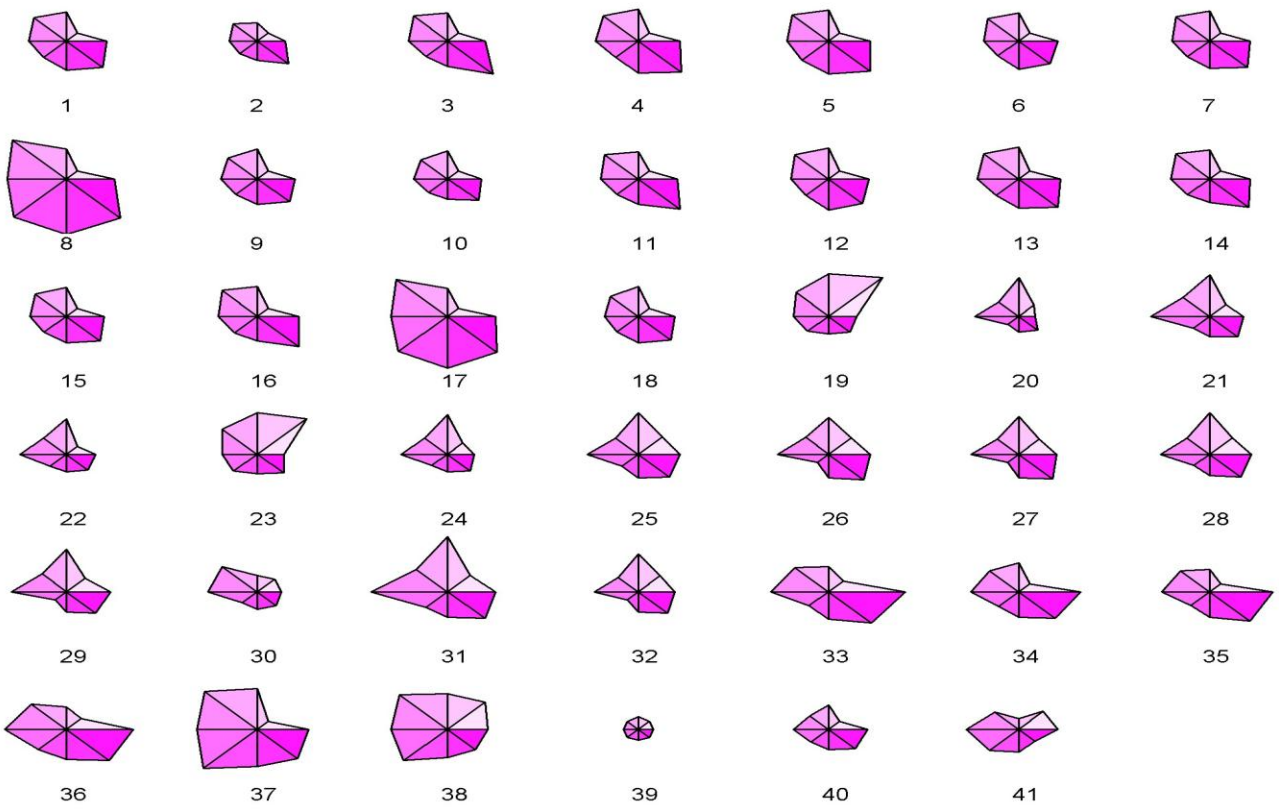
1. GRAFY PŮVODNÍCH DAT

Krabicové grafy proměnlivosti 9 znaků matice zdrojových dat



2. EXPLORATORNÍ ANALÝZA

Symbolový graf (hvězdičky – polygony znaků) objektů



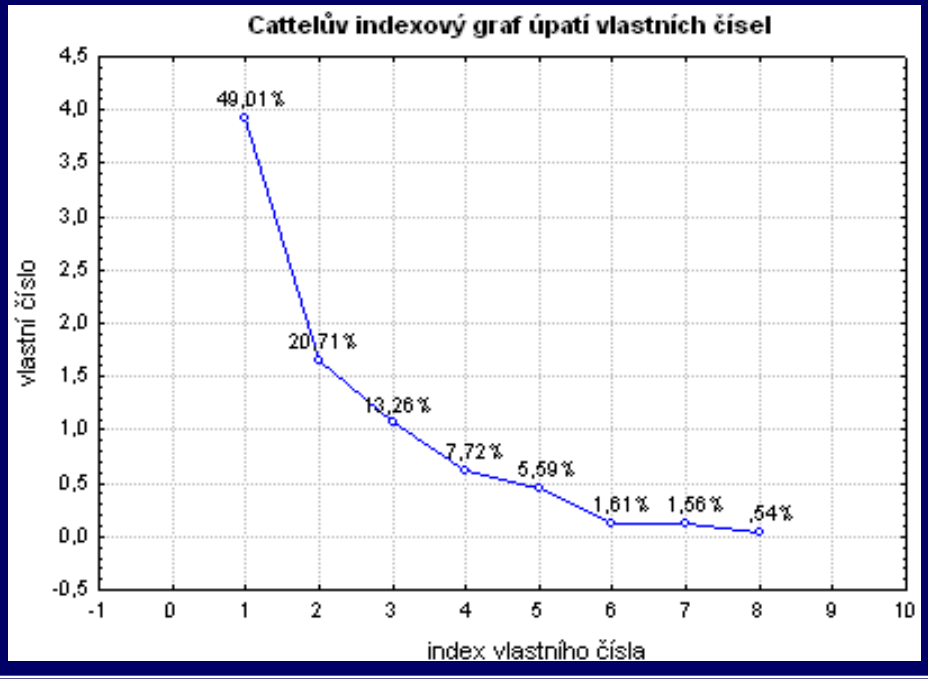
3. KORELAČNÍ MATICE

- Hodnoty párových korelačních koeficientů mezi jednotlivými znaky (nečistotami v heroinu):

znaky	nečist. A	nečist. B	nečist. C	nečist. D	nečist. F	nečist. G	nečist. H	nečist.CH
nečist. A	**1,000000**	-0,190655	-0,037793	0,406090	0,552509	0,340724	0,512593	0,681198
nečist. B	-0,190655	**1,000000**	0,419974	0,003018	0,116207	-0,027592	-0,275146	-0,359852
nečist. C	-0,037793	0,419974	**1,000000**	0,053221	0,415744	-0,046076	-0,005147	0,062775
nečist. D	0,406090	0,003018	0,053221	**1,000000**	0,369257	0,890858	0,816952	0,669025
nečist. F	0,552509	0,116207	0,415744	0,369257	**1,000000**	0,347252	0,381635	0,318156
nečist. G	0,340724	-0,027592	-0,046076	0,890858	0,347252	**1,000000**	0,823067	0,518116
nečist. H	0,512593	-0,275146	-0,005147	0,816952	0,381635	0,823067	**1,000000**	0,801525
nečist.CH	0,681198	-0,359852	0,062775	0,669025	0,318156	0,518116	0,801525	**1,000000**

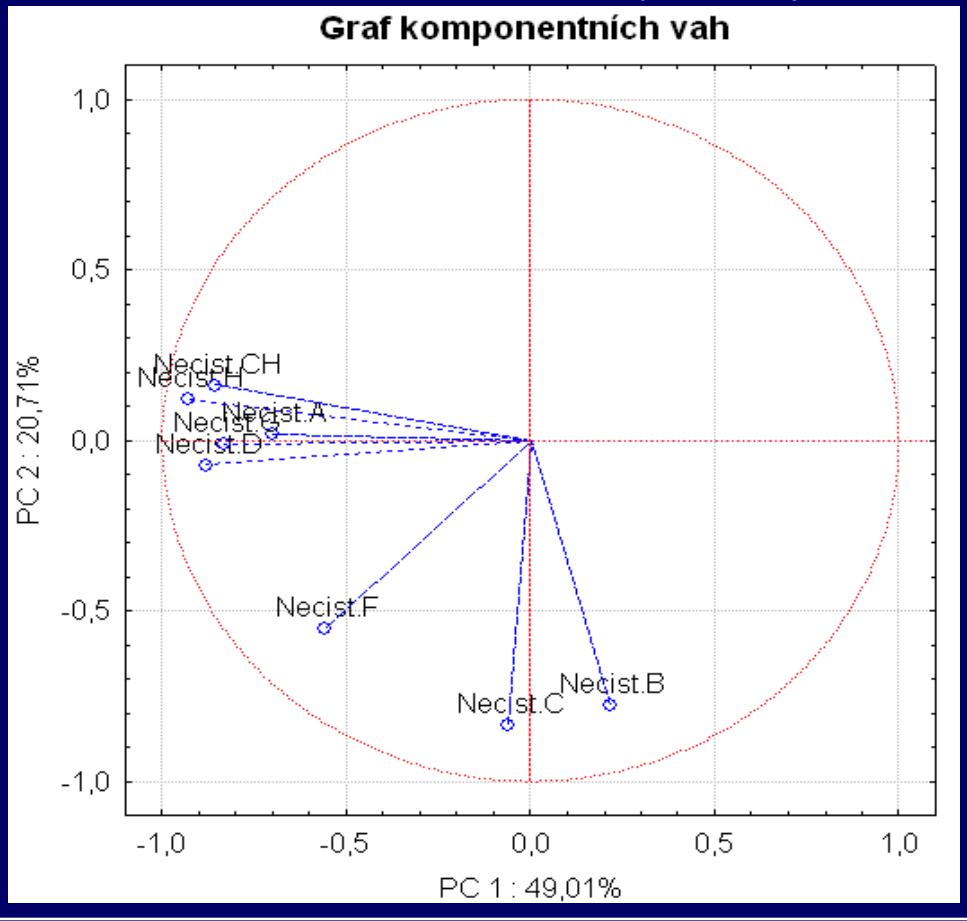
4. METODA HLAVNÍCH KOMPONENT (PCA)

- **Metoda založená na lineární transformaci původních znaků na nové, nekorelované proměnné nazvané hlavní komponenty (základní charakteristikou každé hlavní komponenty je její míra variability). Každá hlavní komponenta představuje lineární kombinaci původních znaků. Výsledek analýzy PCA lze zobrazit v několika typech grafů:**



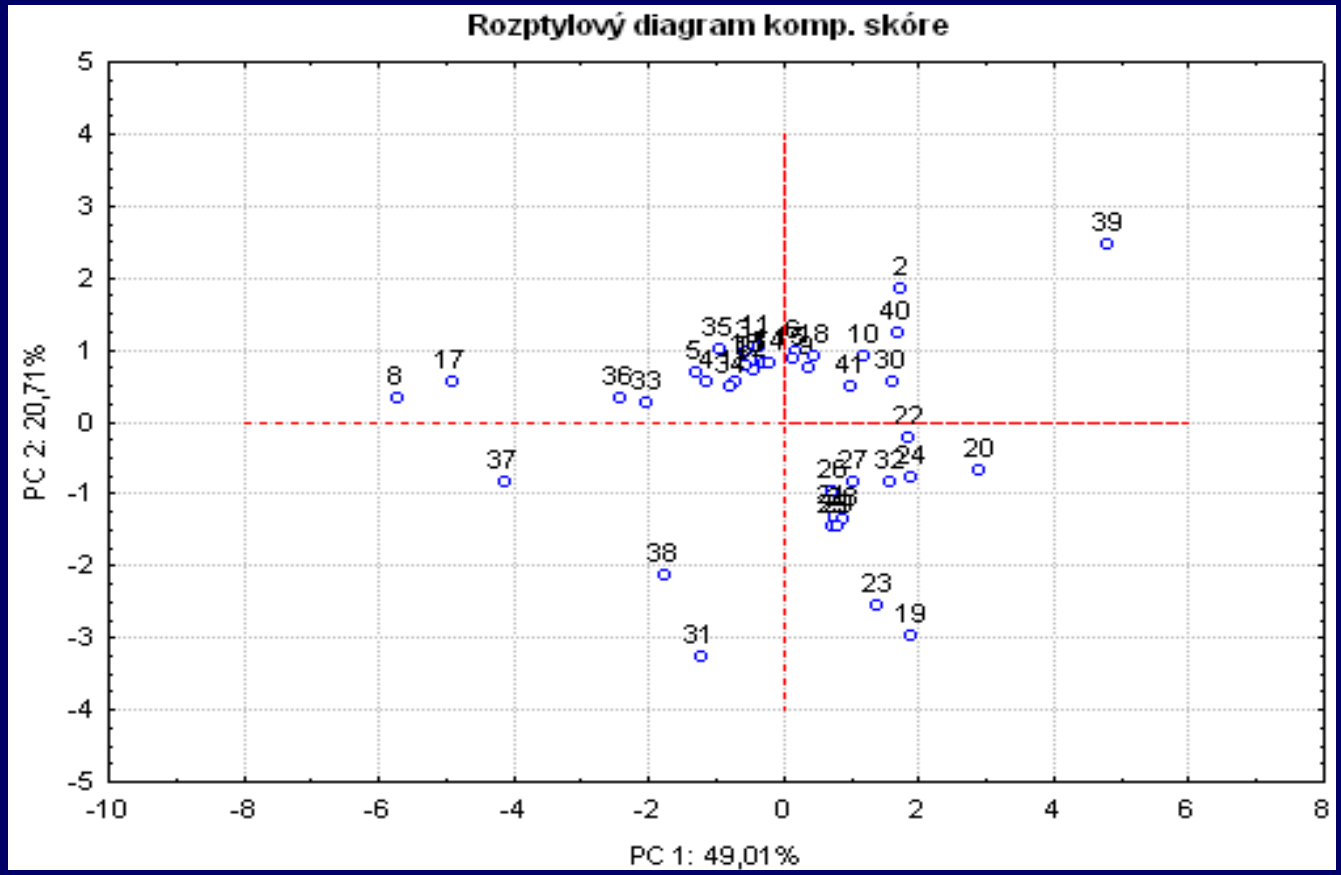
4. METODA HLAVNÍCH KOMPONENT (PCA)

Graf komponentních vah (zátěží) 1 a 2



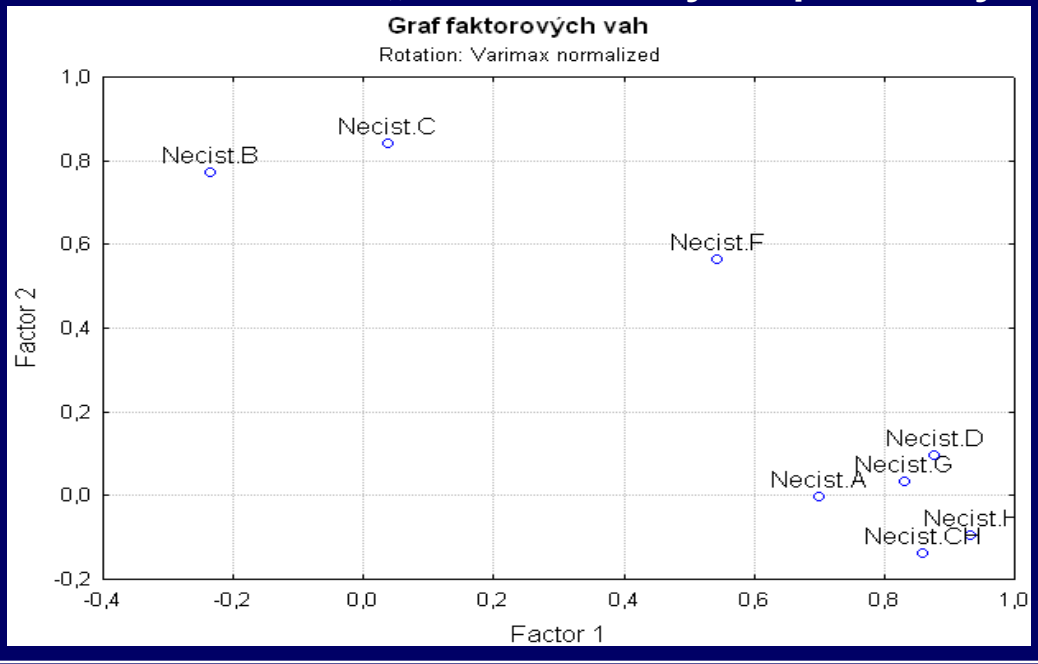
4. METODA HLAVNÍCH KOMPONENT (PCA)

Rozptylový diagram komponentního skóre pro PC 1 a 2



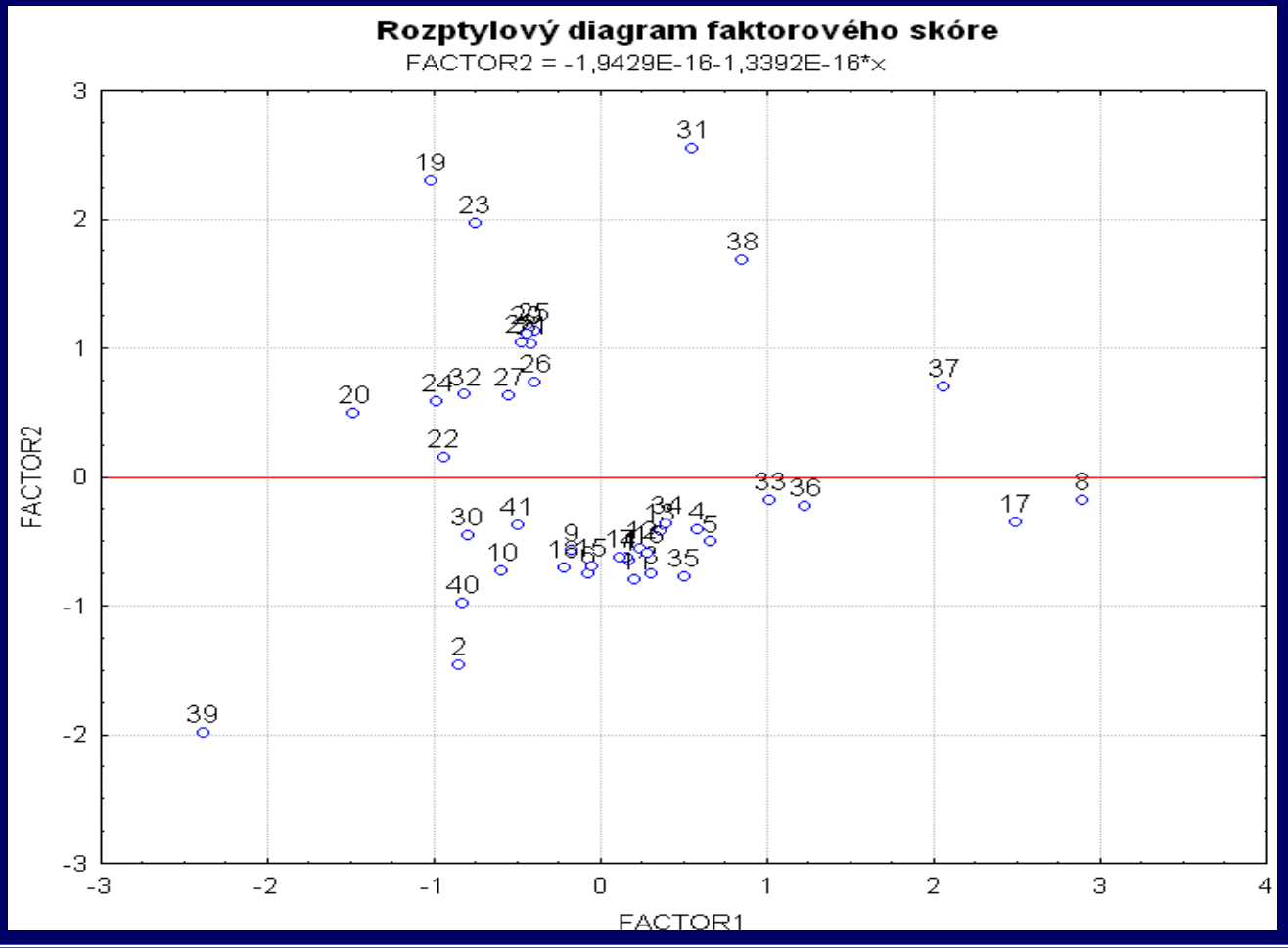
5. FAKTOROVÁ ANALÝZA (FA)

- **Podobně jako metoda PCA náleží faktorová analýza mezi metody snížení dimenze, tedy redukce počtu původních znaků. Tato metoda předpokládá, že každou vstupující proměnnou lze vyjádřit jako lineární funkci nevelkého počtu latentních proměnných – faktorů a jediného specifického faktoru. Významnou nabídkou FA je rotace faktorů (maximalizace zisku „faktorově čistých“ proměnných).**



5. FAKTOROVÁ ANALÝZA (FA)

Rozptylový diagram faktorového skóre pro faktory 1 a 2 po rotaci



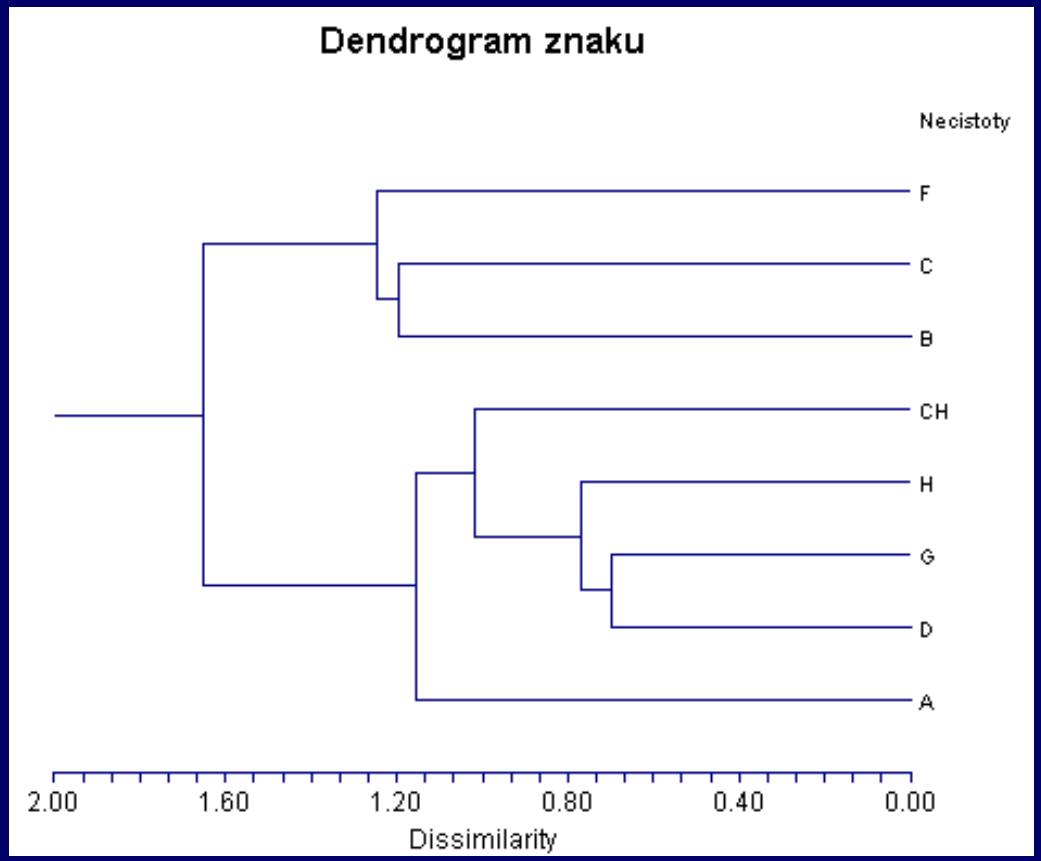
6. ANALÝZA SHLUKŮ (CLU)

- **Analýza shluků je metodou zabývající se vyšetřováním podobnosti vícerozměrných objektů a jejich klasifikací do tříd (shluků). Diagram shluků se nazývá dendrogram (vývojový strom). Při konstrukci dendrogramů je potřebné zvolit algoritmus a typ shlukování a dále vyšetřit, jakou nejvhodnější metriku shlukování použít (posouzení míry věrohodnosti na základě kofenetického korelačního koeficientu CC a kritéria Δ):**

typ metriky shlukování	kofenetická korelace CC	kritérium $\Delta_{0.5}$	kritérium Δ_1
metoda nejbližšího souseda	0,868	0,548	0,629
metoda nejvzdálenějšího souseda	0,708	0,436	0,517
metoda párového průměru	0,848	0,181	0,241
metoda skupinového průměru	**0,897**	**0,152**	**0,199**
metoda mediánu	0,661	0,506	0,614
metoda těžiště	0,815	0,471	0,628
Wardova metoda	0,622	0,818	0,828

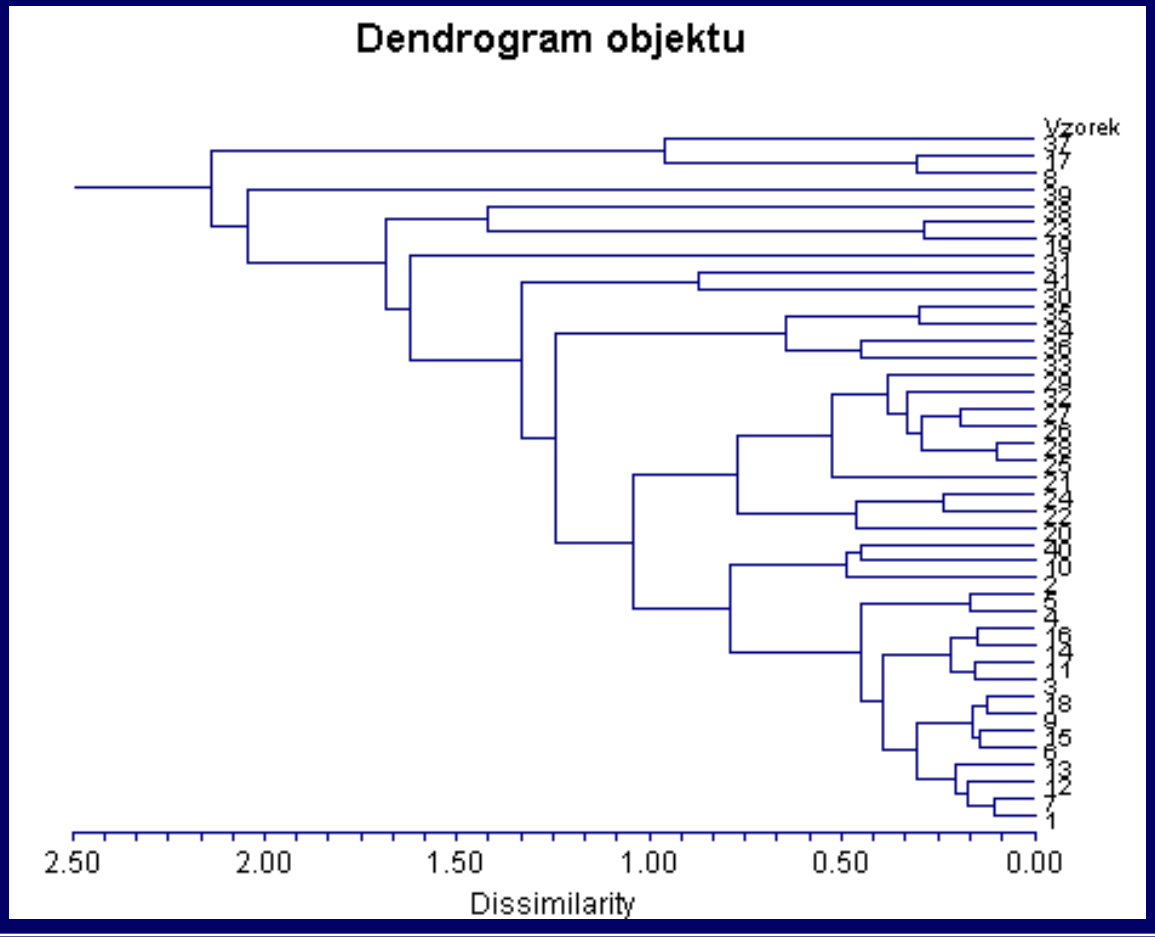
6. ANALÝZA SHLUKŮ (CLU)

Dendrogram znaků (nečistot heroinu) matice dat



6. ANALÝZA SHLUKŮ (CLU)

Dendrogram objektů (vzorků heroinu) matice dat



7. VYHODNOCENÍ

- vhodnost použití uvedeného postupu je zde dokumentována zjištěným výskytem vzorků heroinu prakticky shodného původu (složení) v téže kmenové skupině související s jednou vyšetřovanou trestní věcí (= 1 dealer) a na druhé straně i výskytem jednotlivých vzorků nebo jejich skupinek vyčleněných z rámce kmenové skupiny, které se vymykají většinovému složení;
- z výsledků provedené analýzy vícerozměrných dat vyplývá, že je ji možno aplikovat na jednotlivé soubory vzorků heroinu (zajištěné k různým trestním kauzám) s uspokojivými výsledky a že je tedy možné metodiku profilování zařadit do rutinní práce se vzorky heroinu;
- důležitým výstupem se jeví i možné propojení mezi trestními kauzami původně řešenými jako naprosto nezávislé a to na základě zjištěných vzorků podobných nebo shodných vlastností (možná shoda v původu).

Jaké jsou výhody a nevýhody při použití standardizovaných dat u PCA

- **Předúprava dat standardizací může být důležitým krokem v řadě technik vícerozměrné analýzy dat PCA nevyjímaje. Standardizací dat rozumíme lineární transformaci dat odstraňující jejich závislost na měřících jednotkách a na parametru polohy, popř. i rozptýlení (obecný termín „škálování“ označuje operaci týkající se jak jednotek, tak i počátku stupnice). Existuje celá řada forem standardizace dat.**
- **Výhody: právě odstranění závislosti na jednotkách a tím je daná možnost zahrnutí znaků s nejrůznějším fyzikálním, resp. i jiným „obsahem“ (znaky se stávají souměřitelné) a dále následné přiřazení vhodné předem dané důležitosti všem znakům.**
- **Nevýhoda: neuniversálnost (nepoužívat standardizaci plošně, ale s ohledem na charakter analyzovaných dat, aby nedošlo ke zkreslení závěrů – např. je nevhodná pro znaky blízké nebo na úrovni experimentálního šumu, tj. s nízkým podílem signál/šum, kdy dochází k nežádoucímu zvýraznění významnosti těchto znaků).**

Jak je definována Mahalanobisova vzdálenost objektů od průměrného objektu charakterizovaného střední hodnotou **m** a kovarianční maticí **D**

- Mahalanobisova vzdálenost pro prvek x_i náhodného výběru je definována vztahem:
- $MD(x_i) = [(x_i - m)^T \cdot (D)^{-1} \cdot (x_i - m)]^{1/2}$
- kde **m** je odhad vektoru středních hodnot (výběrových aritmetických průměrů) a **D** je odhad kovarianční matice (výběrová kovarianční matice), prvky inverze kovarianční matice se (na rozdíl od kovarianční matice) mění přidáním dalších znaků a její prvky jsou tak funkcí počtu znaků.
- Mahalanobisova („zobecněná“) vzdálenost se používá hlavně při indikaci vlivných bodů a při porovnání jednotlivých objektů se skupinami objektů (rozpoznávání objektů). Existuje i v robustní variantě (**m** nahrazuje verze robustních výběrových průměrů a **D** nahrazuje robustní verze kovarianční matice)