Cvičení 3: Vícerozměrná jednofaktorová analýza rozptylu (MANOVA)

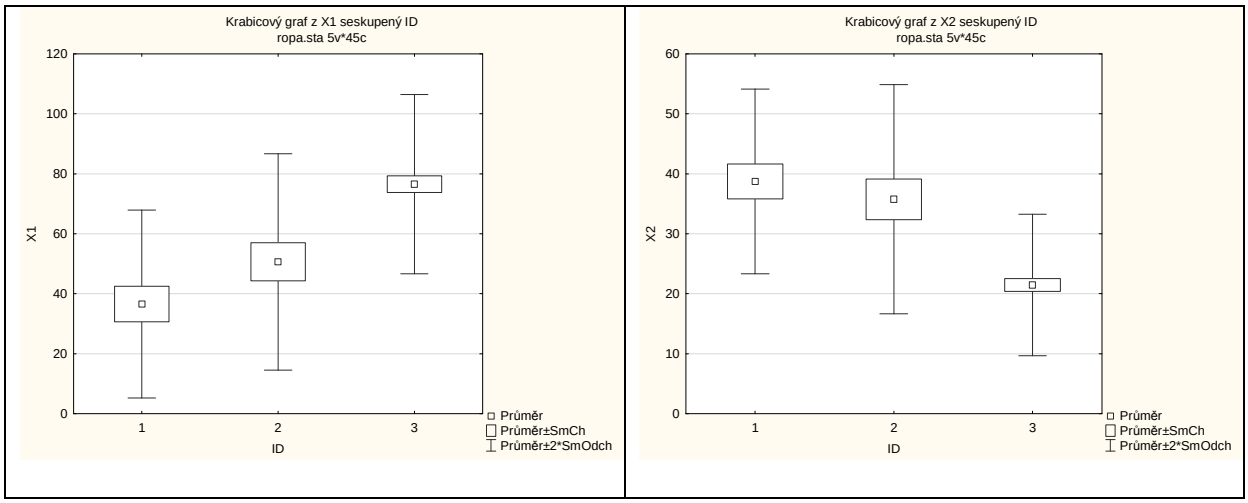
Na 45 vzorcích rudy pocházejících ze tří ložisek byly zjištěny hodnoty těchto čtyř proměnných:
X1 … obsah vanadu v popelu (v promile)
X2 … obsah železa v popelu (v promile)
X3 … obsah nasycených uhlovodíků (v setinách procenta)
X4 … obsah aromatických uhlovodíků (v setinách procenta)
Data jsou uložena v souboru ropa. sta.

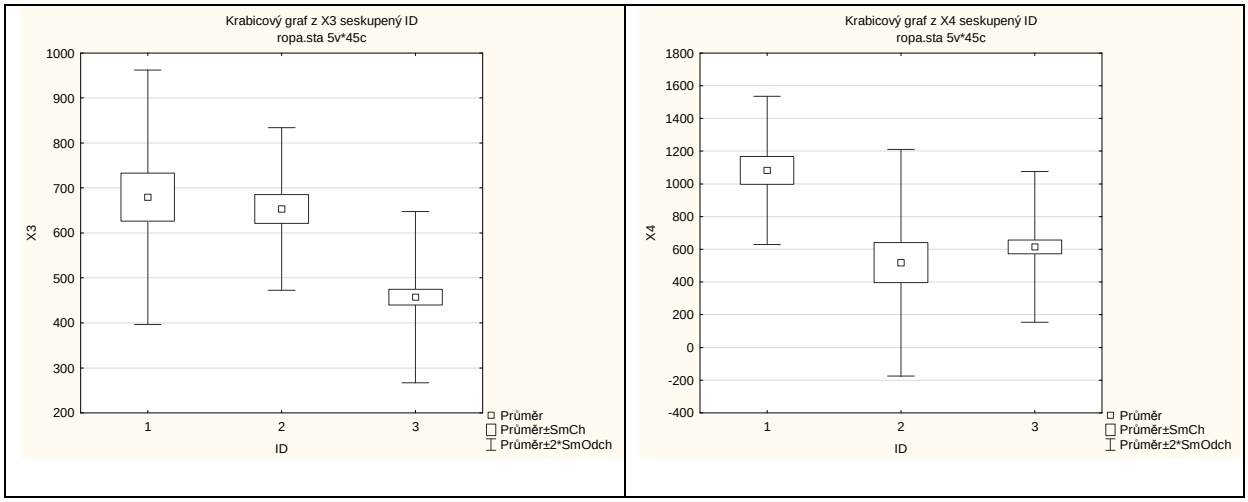
Úkol 1.: Ve všech třech skupinách vypočtěte průměry a směrodatné odchylky proměnných X1, X2, X3, X4. Zjistěte rovněž rozsahy skupin. Vytvořte krabicové grafy proměnné Xi ve všech třech skupinách, i = 1, 2, 3, 4.
Řešení: Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X1, X2, X3, X4 – OK – Anal. skupin – zaškrtneme Zapnuto a Sloučit tabulkové výsledky v jedné tabulce a zrušíme Výsledky za všech. skupiny – zadáme Skupin. proměnná ID – OK – Detailní výsledky – zrušíme Minimum a maximum – Výpočet

	Souhrnné výsledky Popisné statistiky (ropa.sta)			
Proměnná	ID	N platných	Průměr	Sm.odch.
X1	1	7	36,571	15,6403
X2	1	7	38,714	7,6966
X3	1	7	679,571	141,4318
X4	1	7	1082,571	226,1260
X1	2	8	50,6250	18,0471
X2	2	8	35,7500	9,5581
X3	2	8	653,2500	90,2754
X4	2	8	518,1250	346,3580
X1	3	30	76,5333	14,9406
X2	3	30	21,4667	5,8882
X3	3	30	457,4667	95,2430
X4	3	30	614,8667	230,5085

Komentář: Počty vzorků z jednotlivých nalezišť se liší. Zatímco z 1. a 2. naleziště bylo odebráno 7 a 8 vzorků, ze třetího pak 30 vzorků.
Obsah vanadu je nejmenší na 1. nalezišti a největší na 3. nalezišti.
U obsahu železa je tomu naopak – nejvíce železa je ve vzorcích z 1. naleziště, naopak nejméně je ho na 3. nalezišti.
Obsah nasycených uhlovodíků se u 1. a 2. naleziště liší jen málo, na 3. nalezišti je nejnižší.
Obsah aromatických uhlovodíků je největší na 1. nalezišti, nejmenší na 2. nalezišti.
Nejvariabilnější obsah vanadu je ve vzorcích z 2. naleziště, naopak nejstabilnější je ve vzorcích z 3. naleziště.
Obsah železa nejvíce kolísá u vzorků 2. naleziště, největší stabilitu obsahu železa vykazují vzorky ze 3. naleziště.
U nasycených uhlovodíků pozorujeme největší variabilitu u vzorků z 1. naleziště, nejmenší u vzorků z 2. naleziště.
Variabilita obsahu aromatických uhlovodíků je u vzorků z 1. a 3. naleziště podobná, největší je u vzorků z 2. naleziště.

Grafy – 2D grafy – Krabicové grafy – Typ grafu: Vícenásobný – Proměnné – Závisle proměnné X1 – Grupovací proměnná ID – Detaily – Střední bod – Průměr – Odlehlé hodnoty – Vypnuto – OK
Tentýž postup zopakujeme pro proměnné X2, X3, X4.





Úkol 2.: Na hladině významnosti 0,05 testujte hypotézu, že proměnné X1, X2, X3, X4 se ve všech třech skupinách řídí normálním rozložením.
Řešení: Statistiky – Základní statistiky a tabulky – Tabulky četností – OK – Proměnné X1, X2, X3, X4 – OK - Anal. skupin – zaškrtneme Zapnuto a Sloučit tabulkové výsledky v jedné tabulce a zrušíme Výsledky za všech. skupiny – zadáme Skupin. proměnná ID – OK – OK – záložka Normalita – zaškrtneme S-W test a zrušíme K-S test – Testy normality

Proměnná	Souhrnné výsledky Testy normality (ropa.sta) ID	N	max D	Lilliefors p	W	p
X1: vanad (v promile)	1	7	0,279595	p < ,10	0,837889	0,094950
X2: zelezo (v promile)	1	7	0,256734	p < ,20	0,894783	0,300555
X3: nasyc. uhlovodiky (v des. promile)	1	7	0,164851	p > .20	0,944596	0,680404
X4: arom uhlovodiky (v des. promile)	1	7	0,218850	p > .20	0,886488	0,256840
X1: vanad (v promile)	2	8	0,268226	p < ,10	0,812765	0,039143
X2: zelezo (v promile)	2	8	0,222404	p > .20	0,916727	0,403873
X3: nasyc. uhlovodiky (v des. promile)	2	8	0,222340	p > .20	0,891501	0,241660
X4: arom uhlovodiky (v des. promile)	2	8	0,270404	p < ,10	0,798241	0,027410
X1: vanad (v promile)	3	30	0,114117	p > .20	0,955701	0,239602
X2: zelezo (v promile)	3	30	0,165019	p < ,05	0,939091	0,085977
X3: nasyc. uhlovodiky (v des. promile)	3	30	0,189553	p < ,01	0,884710	0,003623
X4: arom uhlovodiky (v des. promile)	3	30	0,115612	p > .20	0,954858	0,227664

Komentář: Lilieforsův test zamítá na hladině významnosti 0,05 hypotézu o normalitě obsahu železa a obsahu nasycených uhlovodíků u vzorků ze 3. naleziště. S-W test zamítá na hladině významnosti 0,05 hypotézu o normalitě obsahu vanadu a aromatických uhlovodíků u vzorků z 2. naleziště a také obsahu nasycených uhlovodíků u vzorků ze 3. naleziště. Normalita je však porušena jen mírně. Nedopustíme s závažné chyby, budeme-li předpokládat, že datová matice je realizací výběru ze čtyřrozměrného normálního rozložení.

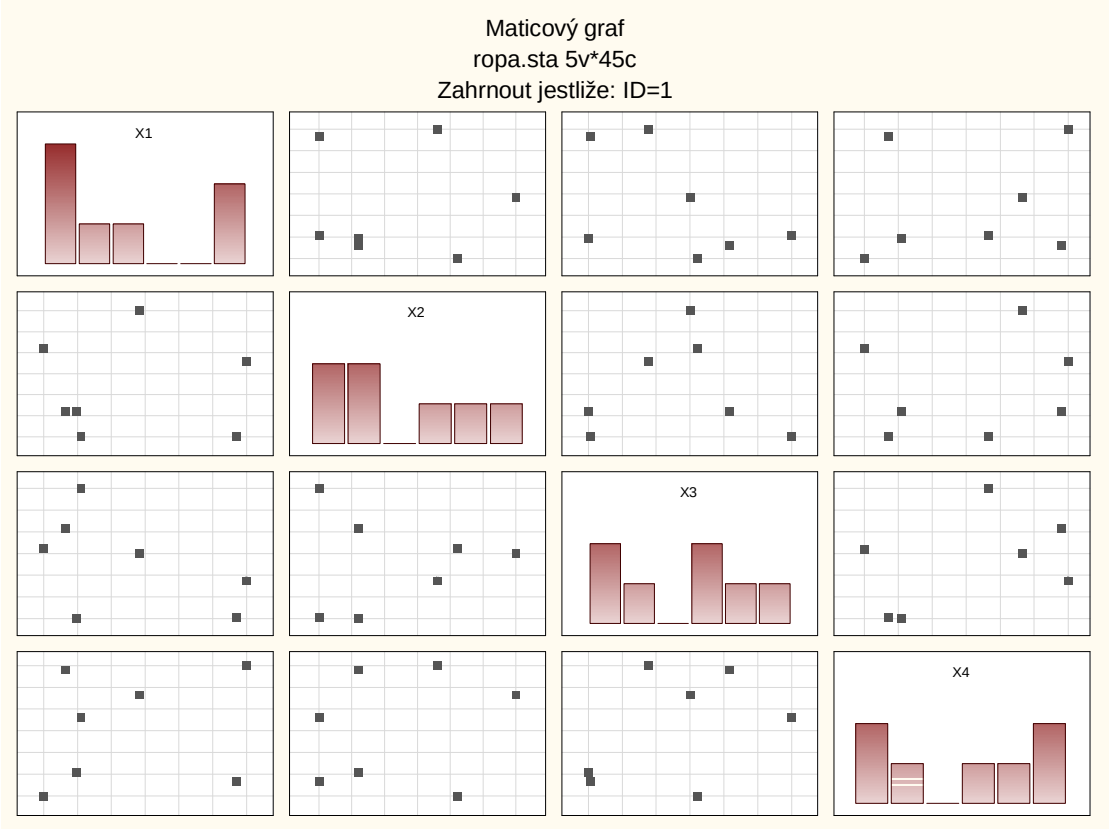
Úkol 3.: Na hladině významnosti 0,05 testujte hypotézu, že varianční matice proměnných X1, X2, X3, X4 jsou ve všech třech skupinách shodné.
Řešení: Statistiky – ANOVA – Jednofaktorová ANOVA – OK – Proměnné – Seznam, závislých proměnných X1, X2, X3, X4 - Kategor. nezávislá proměnná (faktor) ID – OK – OK – Více výsledků – záložka Předpoklady – Boxův M test

	Boxův M test (ropa.sta) Efekt: ID (Vypočteno pro všechny proměnné) Boxovo M	Chí-kv.	SV	p
Boxovo M	35,34766	27,23627	20	0,128747

Komentář: Test shody tří variančních matic poskytl p-hodnotu 0,1287, což je větší než 0,05, tedy dále budeme varianční matice považovat za shodné.

Lze konstatovat, že důležité předpoklady vícerozměrné analýzy rozptylu jsou splněny.

Úkol 4.: Pomocí maticových grafů prověřte, že vztahy mezi proměnnými X1, X2, X3, X4 jsou ve všech třech skupinách přibližně lineární.
Řešení: Grafy – Maticové grafy - Proměnné X1, X2, X3, X4 – OK – Filtr případů – Zapnout filtr ID=1 – OK – OK
(Analogicky pro 2. a 3. naleziště, zadáme ID=2 resp. ID=3)



Vidíme, že pro vzorky ropy z 1. naleziště je v některých případech linearita porušená. Podobně to dopadne i pro data z 2. a 3. naleziště. Musíme si být vědomi toho, že výskyt nelinearit snižuje sílu testů v MANOVĚ.

Úkol 5.: Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty proměnných X1, X2, X3, X4 jsou ve všech třech skupinách shodné. Použijte Wilksův, Pillaiův, Hotellingův – Lawleův a Royův test.
Řešení: Návrat do ANOVA – záložka Detaily – zaškrtneme vš. Vícerozměrné testy – Test všech efektů

Efekt	Vícerozměrné testy významnosti. (ropa.sta) Sigma-omezená parametrizace Dekompozice efektivní hypotézy Test	Hodnota	F	Efekt SV	Chyba SV	p
Abs. člen	Wilksův	0,01616	593,4657	4	39	0,000000
	Pillaiův	0,98384	593,4657	4	39	0,000000
	Hotellng	60,86828	593,4657	4	39	0,000000
	Royův	60,86828	593,4657	4	39	0,000000
ID	Wilksův	0,17959	13,2570	8	78	0,000000
	Pillaiův	1,08176	11,7808	8	80	0,000000
	Hotellng	3,11290	14,7863	8	76	0,000000
	Royův	2,53997	25,3997	4	40	0,000000

Komentář: Všechny čtyři testy zamítají na hladině významnosti 0,05 hypotézu, že střední hodnoty proměnných X1, X2, X3, X4 jsou ve všech třech skupinách shodné. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že aspoň mezi dvěma nalezišti existuje rozdíl z hlediska obsahu sledovaných látek.

Úkol 6.: Pomocí simultánního testu založeného na Wilksově statistice testujte na hladině významnosti 0,05 hypotézu, že proměnné X1, X2, X3, X4 nezpůsobují rozdíly mezi skupinami.
Řešení: Simultánní testy STATISTICA neposkytuje. Můžeme však s její pomocí vypočítat matici **E** reziduální variability a matici **T** celkové variability. Z těchto matic použijeme diagonální prvky pro výpočet všech čtyř testových statistik

$K_j = -\left(n - \frac{p+r}{2} - 1\right)\ln\frac{e_{jj}}{t_{jj}}$, j = 1, 2, 3, 4. Platí-li nulová hypotéza, K_j se asymptoticky řídí rozložením $\chi^2(p(r-1))$. Nulovou hypotézu o proměnné Xj tedy zamítneme na asymptotické hladině významnosti α, když $K_j \in \langle \chi^2_{1-\alpha}(p(r-1)), \infty)$. V našem případě n = 45, p = 4, r = 3.

Výpočet matice **E** reziduální variability:
Návrat do ANOVA – záložka Matice – v části ozn. Meziskupinové efekty vybereme SČ chyb.

	Matice SSCP (Z' Z) reziduí (ropa.sta) Sigma-omezená parametrizace Dekompozice efektivní hypotézy				
Efekt	proměnné	X1	X2	X3	X4
Chyba	X1	10221,1	-1826,1	-16205,0	47988
	X2	-1826,1	2000,4	9266,1	-15263
	X3	-16205,0	9266,1	440130,7	403609
	X4	47988,2	-15262,7	403609,3	2687436

Výpočet matice **T** celkové variability (je to matice v pravém dolním rohu):
Návrat do ANOVA – záložka Matice – v části ozn. Meziskupinové schéma vybereme Z'Z odchylek.

	Matice SSCP (Z' Z) odchylek (ropa.sta) Matice SSCP (Z' Z) odchylek vektorů matice v matici schématu X									
Efekt	Úroveň	Sloupec	Efekt (P/N)	Sloup.1 Abs.člen	Sloup.2 ID	Sloup.3 ID	Sloup.4 X1	Sloup.5 X2	Sloup.6 X3	Sloup.7 X4
Abs. člen		1	Pevný							
ID	1	2	Pevný		25,244	18,756	-528,6	240,84	3149,9	4552
ID	2	3	Pevný		18,756	27,244	-445,4	229,16	3092,1	448
X1		4			-528,644	-445,356	21499,2	-7068,04	-85138,3	-35738
X2		5			240,844	229,156	-7068,0	4487,64	42154,5	17095
X3		6			3149,911	3092,089	-85138,3	42154,51	875634,6	805853
X4		7			4551,711	448,289	-35737,5	17094,91	805853,4	4154653

K dalším výpočtům použijeme STATISTIKU jako inteligentní kalkulačku. Otevřeme nový datový soubor o jednom případu a s pěti proměnnými K1, K2, K3, K4 a kvantil.
Do Dlouhého jména proměnné K1 napíšeme: =-40,5*log(10221,1/21499,2)
Do Dlouhého jména proměnné K2 napíšeme: =-40,5*log(2000,4/4487,64)
Do Dlouhého jména proměnné K3 napíšeme: =-40,5*log(440130,7/875634,6)
Do Dlouhého jména proměnné K4 napíšeme: =-40,5*log(2687436/4154653)
Proměnná kvantil obsahuje kvantil $\chi^2_{0,95}(8)$, tedy do jejího Dlouhého jména napíšeme: =VChi2(0,95;8)

Vypočtené testové statistiky pro simultánní testy a kritická hodnota:

	1 K1	2 K2	3 K3	4 K4	5 kvantil
1	30,114241	32,723182	27,859025	17,643474	15,507313

Komentář: Vidíme, že všechny čtyři statistiky se realizují v kritickém oboru $W = \langle 15{,}5073, \infty)$. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že všechna tři naleziště se liší v obsahu všech čtyř zkoumaných látek.

Úkol 7.: Na hladině významnosti 0,05 proveďte vícerozměrnou obdobu mnohonásobného porovnávání, tj. zjistěte, které dvojice skupin se liší.
Řešení: Vícenásobnou obdobu mnohonásobného porovnávání STATISTICA neposkytuje. Problém vyřešíme tak, že provedeme všechna tři porovnání (1-2, 1-3, 2-3) pomocí vícerozměrného dvouvýběrového t-testu založeného na Hotellingově statistice T^2 a získané p-hodnoty porovnáme s hladinou významnosti korigovanou podle Bonferroniho, tj. s číslem $\frac{\alpha}{\binom{3}{2}} = \frac{0{,}05}{3} = 0{,}01\overline{6}$.
Statistiky – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK – Proměnné – Závisle proměnné X1, X2, X3, X4 – Grupovací proměnná ID – OK – Kód pro skup. 1: 1, Kód pro skup. 2: 2 – na záložce Možnosti zaškrtneme Vícerozměrný test (Hotellingovo T^2) - Výpočet

Výsledek pro 1. a 2. skupinu:

Proměnná	t-testy; grupováno: ID: naleziste (ropa.sta) Skup. 1: 1; Skup. 2: 2 Hotellingovo 45,6734 F(4,10)=8,7833 p<,00261 Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr Rozptyly	p Rozptyly
X1	36,571	50,6250	-1,59930	13	0,133764	7	8	15,6403	18,0471	1,331443	0,743087
X2	38,714	35,7500	0,65470	13	0,524074	7	8	7,6966	9,5581	1,542203	0,613888
X3	679,571	653,2500	0,43578	13	0,670148	7	8	141,4318	90,2754	2,454458	0,265396
X4	1082,571	518,1250	3,67238	13	0,002814	7	8	226,1260	346,3580	2,346116	0,318519

Vypočtenou p-hodnotu (tj. 0,00261) porovnáme s $0{,}01\overline{6}$. Vidíme, že 1. a 2. skupina se liší.

Výsledek pro 1. a 3. skupinu

Proměnná	t-testy; grupováno: ID: naleziste (ropa.sta) Skup. 1: 1; Skup. 2: 3 Hotellingovo 125,397 F(4,32)=28,662 p<,00000 Průměr 1	Průměr 3	t	sv	p	Poč.plat 1	Poč.plat. 3	Sm.odch. 1	Sm.odch. 3	F-poměr Rozptyly	p Rozptyly
X1	36,571	76,5333	-6,32043	35	0,000000	7	30	15,6403	14,9406	1,095851	0,776819
X2	38,714	21,4667	6,58961	35	0,000000	7	30	7,6966	5,8882	1,708565	0,309137
X3	679,571	457,4667	5,05771	35	0,000013	7	30	141,4318	95,2430	2,205100	0,142430
X4	1082,571	614,8667	4,84954	35	0,000025	7	30	226,1260	230,5085	1,039138	1,000000

I v tomto případě nulovou hypotézu zamítáme na hladině významnosti 0,05.

Výsledek pro 2. a 3. skupinu:

Proměnná	t-testy; grupováno: ID: naleziste (ropa.sta) Skup. 1: 2; Skup. 2: 3 Hotellingovo 44,5444 F(4,33)=10,208 p<,00002 Průměr 2	Průměr 3	t	sv	p	Poč.plat 2	Poč.plat. 3	Sm.odch. 2	Sm.odch. 3	F-poměr Rozptyly	p Rozptyly
X1	50,6250	76,5333	-4,17559	36	0,000180	8	30	18,0471	14,9406	1,459063	0,441637
X2	35,7500	21,4667	5,31026	36	0,000006	8	30	9,5581	5,8882	2,634953	0,061803
X3	653,2500	457,4667	5,21782	36	0,000008	8	30	90,2754	95,2430	1,113082	0,958255
X4	518,1250	614,8667	-0,94544	36	0,350739	8	30	346,3580	230,5085	2,257752	0,116036

Vidíme, že i 2. a 3. skupina se liší na hladině významnosti 0,05.

Úkol 8.: Na hladině významnosti 0,05 zjistěte, které proměnné způsobují rozdíly mezi jednotlivými dvojicemi skupin. (Těchto testů je nutno provést $\frac{pr(r-1)}{2}$, v našem případě tedy $\frac{4 \cdot 3(3-1)}{2} = 12$.)

Řešení: Posouzení rozdílů mezi jednotlivými proměnnými v rámci skupin STATISTICA neposkytuje. Pro každou proměnnou tedy provedeme dvouvýběrový t-test, abychom ji porovnali ve dvojicích skupin 1-2, 1-3, 2-3 a zjistíme, zda vypočtené p-hodnoty jsou menší nebo rovny korigované hladině významnosti $\frac{0,05}{12} = 0,0042$.

Podíváme-li se na tabulky v úkolu 7, můžeme konstatovat, že:

- naleziště 1 a 2 se liší pouze v obsahu aromatických uhlovodíků
- naleziště 1 a 3 se liší v obsahu všech čtyř látek
- naleziště 2 a 3 se neliší pouze v obsahu aromatických uhlovodíků.