UNIVERZITA PARDUBICE
Fakulta chemicko-technologická
Katedra analytické chemie
Nám. Čs. Legií 565, 532 10 Pardubice

10. licenční studium chemometrie

STATISTICKÉ ZPRACOVÁNÍ DAT

Semestrální práce

KLASIFIKACE ANALÝZOU VÍCEROZMĚRNÝCH DAT

2005/2006

Vedoucí studia a odborný garant:
Prof. RNDr. Milan Meloun, DrSc.

Vyučující:
Prof. RNDr. Milan Meloun, DrSc.

Autor práce:
Ing. Zdeňka Dluhošová

OBSAH

ÚLOHA 2.: KLASIFIKACE ANALÝZOU VÍCEROZMĚRNÝCH DAT – KANONICKÁ KORELAČNÍ ANALÝZA 15

1. Zadání 15
1.1 Data 15
1.2 Užitý program 16

2. Řešení 16
2.1 Exploratorní analýza vícerozměrných dat 16
2.2 Kanonická korelační analýza 20

3. Závěr: 24

ÚLOHA 3.: KLASIFIKACE ANALÝZOU VÍCEROZMĚRNÝCH DAT – KORESPODENČNÍ ANALÝZA 26

1. Zadání 26
1.1 Data 26
1.2 Užitý program 26

2. Řešení 27
2.1 Řádkové profily v procentech 27
2.2 Sloupcové profily v procentech 27
2.3 Hledání počtu projekčních dimenzí 28
2.4 Zobrazení řádkového profilu a příspěvek do inercie 29
2.5 Diagnostické grafy 29

3. Závěr: 30

Úloha 2.: Klasifikace analýzou vícerozměrných dat – kanonická korelační analýza

1. Zadání

V rámci speciálního monitorování ovzduší ostravsko-karvinské oblasti se provádí stanovení základních škodlivin (prašnost PM10, NO_2, NO_x, O3), speciálních anorganických škodlivin (kovy) a speciálních organických škodlivin škodlivin (PAU). Zjistěte, jak souvisí koncentrace základních škodlivin s koncentracemi vybraných speciálních škodlivin.

1.1 Data

Označení vzorku	PM10 [μg.m^{-3}]	NO_2 [μg.m^{-3}]	NO_x [μg.m^{-3}]	O_3 [μg.m^{-3}]	Mn [μg.m^{-3}]	As [μg.m^{-3}]	Pb [μg.m^{-3}]	BaANT [μg.m^{-3}]	BaP [μg.m^{-3}]
MH01	90,4	28,9	42,1	38,9	0,0992	0,0147	0,1728	0,0054	0,0046
MH02	66,6	24,7	31,4	52,3	0,0510	0,0087	0,0515	0,0117	0,0098
MH03	70,7	28,6	35,5	30,9	0,1305	0,0189	0,3630	0,0061	0,0060
MH04	54,4	23,2	31,3	66,8	0,0864	0,0141	0,1950	0,0035	0,0050
MH05	42,9	15,1	20,7	67,6	0,1226	0,0157	0,4101	0,0013	0,0015
MH06	44,1	14,5	20,4	63,2	0,0427	0,0062	0,0575	0,0013	0,0013
MH07	47,5	15,2	21,0	61,1	0,0574	0,0092	0,0789	0,0011	0,0016
MH08	47,6	17,0	23,3	66,1	0,2288	0,0168	0,5313	0,0013	0,0023
MH09	60,8	20,7	30,2	48,9	0,2674	0,0215	0,7979	0,0016	0,0022
MH10	67,3	22,4	35,0	35,9	0,2574	0,0306	0,9075	0,0043	0,0064
MH11	68,1	24,5	38,0	28,0	0,3133	0,0350	0,7527	0,0067	0,0061
MH12	73,0	27,6	44,3	20,6	0,3111	0,0294	0,5574	0,0057	0,0053
P01	54,8	32,3	46,7	33,1	0,0063	0,0062	0,0316	0,0115	0,0095
P02	34,6	27,3	34,9	45,5	0,0133	0,0065	0,0301	0,0091	0,0071
P03	46,9	33,2	42,4	51,9	0,0186	0,0060	0,0352	0,0073	0,0063
P04	44,0	28,7	39,6	61,8	0,0258	0,0062	0,0359	0,0040	0,0047
P05	26,8	19,8	27,1	65,2	0,0153	0,0030	0,0199	0,0018	0,0023
P06	26,5	19,7	27,2	60,8	0,0119	0,0031	0,0183	0,0009	0,0015
P07	29,8	19,6	27,5	58,1	0,0193	0,0047	0,0327	0,0009	0,0020
P08	35,5	24,2	35,0	58,4	0,0255	0,0057	0,0468	0,0016	0,0031
P09	33,1	26,0	40,7	42,6	0,0143	0,0036	0,0257	0,0044	0,0061
P10	46,6	30,2	56,5	29,8	0,0153	0,0045	0,0230	0,0129	0,0128
P11	37,2	29,7	50,0	25,3	0,0138	0,0047	0,0278	0,0135	0,0116
P12	44,5	33,6	62,4	17,4	0,0068	0,0046	0,0231	0,0144	0,0123
B01	80,3	35,3	47,4	35,4	0,0251	0,0101	0,0675	0,0167	0,0046
B02	78,7	33,0	44,3	45,1	0,0554	0,0120	0,1011	0,0227	0,0098
B03	64,7	28,5	36,5	43,9	0,0593	0,0140	0,1559	0,0168	0,0060
B04	55,2	20,5	27,8	64,5	0,0498	0,0127	0,1107	0,0093	0,0050

Označení vzorku	PM10 [μg.m^{-3}]	NO_2 [μg.m^{-3}]	NO_x [μg.m^{-3}]	O_3 [μg.m^{-3}]	Mn [μg.m^{-3}]	As [μg.m^{-3}]	Pb [μg.m^{-3}]	BaANT [μg.m^{-3}]	BaP [μg.m^{-3}]
B05	46,7	17,6	23,0	63,9	0,0670	0,0119	0,1449	0,0040	0,0015
B06	54,9	18,8	25,3	54,2	0,0510	0,0093	0,0810	0,0029	0,0013
B07	48,1	17,4	24,9	55,2	0,0455	0,0079	0,0822	0,0022	0,0016
B08	43,4	16,4	22,8	64,0	0,0719	0,0142	0,1737	0,0032	0,0023
B09	69,4	23,8	35,7	44,1	0,0905	0,0151	0,1681	0,0057	0,0022
B10	69,6	24,8	37,4	31,0	0,2044	0,0223	0,3123	0,0130	0,0064
B11	71,3	26,7	40,7	20,7	0,0840	0,0156	0,1751	0,0124	0,0072
B12	92,2	92,0	48,7	11,7	0,2192	0,0311	0,3078	0,0253	0,0053

Vysvětlivky:
PM10 – frakce prachu: částice prachu o průměru do 10 μm
No_x – suma oxidů dusíku
BaANT – benzo(a)antracen, zástupce polycyklických aromatických uhlovodíků
BaP – benzo(a)pyren, zástupce polycyklických aromatických uhlovodíků

1.2 Užitý program

STATISTICA, verze 7

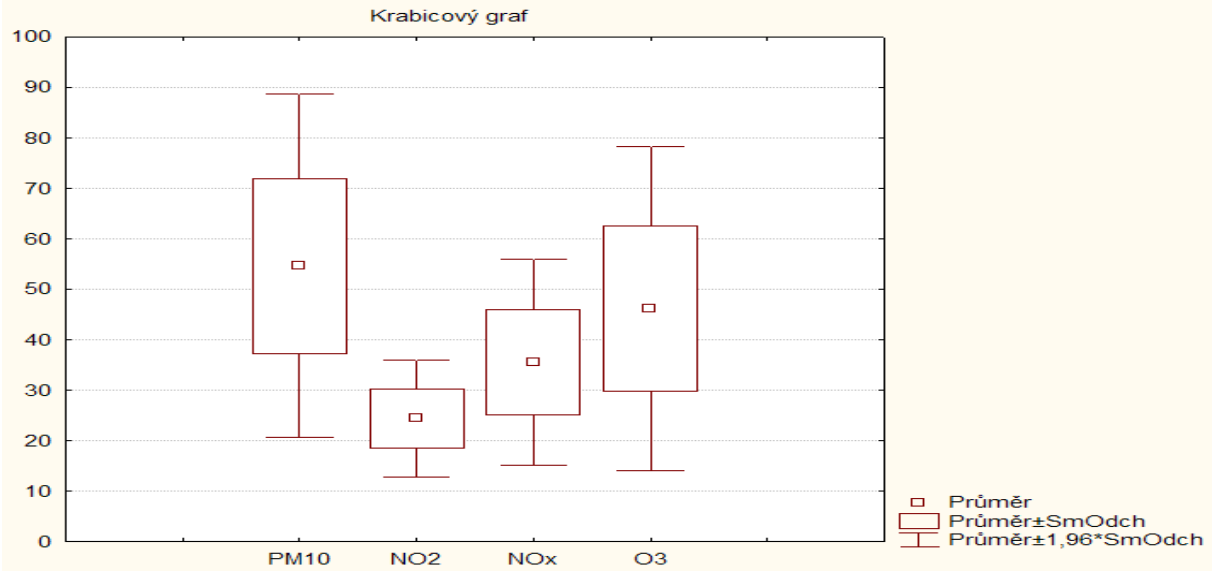
2. Řešení

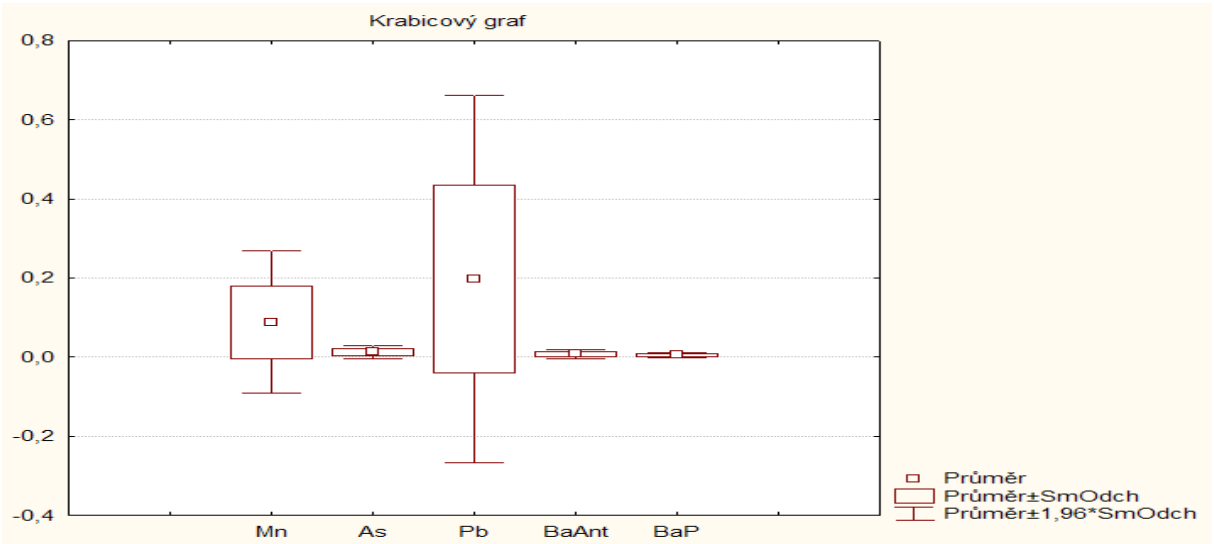
2.1 Exploratorní analýza vícerozměrných dat

a) Parametry polohy a rozptýlení

	Průměr	Směrodatná odchylka
PM10	54,6	17,3
NO_2	24,4	5,8
NO_x	35,4	10,4
O_3	46,2	16,3
Mn	0,0882	0,0916
As	0,0126	0,0085
Pb	0,1973	0,2375
Benzo(a)antracen	0,0074	0,0063
Benzo(a)pyren	0,0051	0,0033

b) Krabicový graf





Pro lepší názornost byly zvoleny 2 krabicového grafy - jeden pro základní škodliviny a jeden pro speciální škodliviny. Největší proměnlivost v datech základních škodlivin vykazují znaky PM10 a O_3, v datech speciálních škodlivin znak Pb.

c) Korelační analýza

Mírou lineární závislosti mezi dvěma náhodnými veličinami je párový korelační koeficient. Korelační koeficient blížící se hodnotě 1.0 značí, že mezi dvěma proměnnými existuje silný pozitivní lineární vztah. Korelační koeficient blížící se hodnotě -1.0 značí, že mezi proměnnými existuje silný negativní lineární vztah. Zda je korelační koeficient statisticky významný, lze usuzovat z hladin významnosti p: je-li $p < 0.05$, je korelační koeficient statisticky významný.

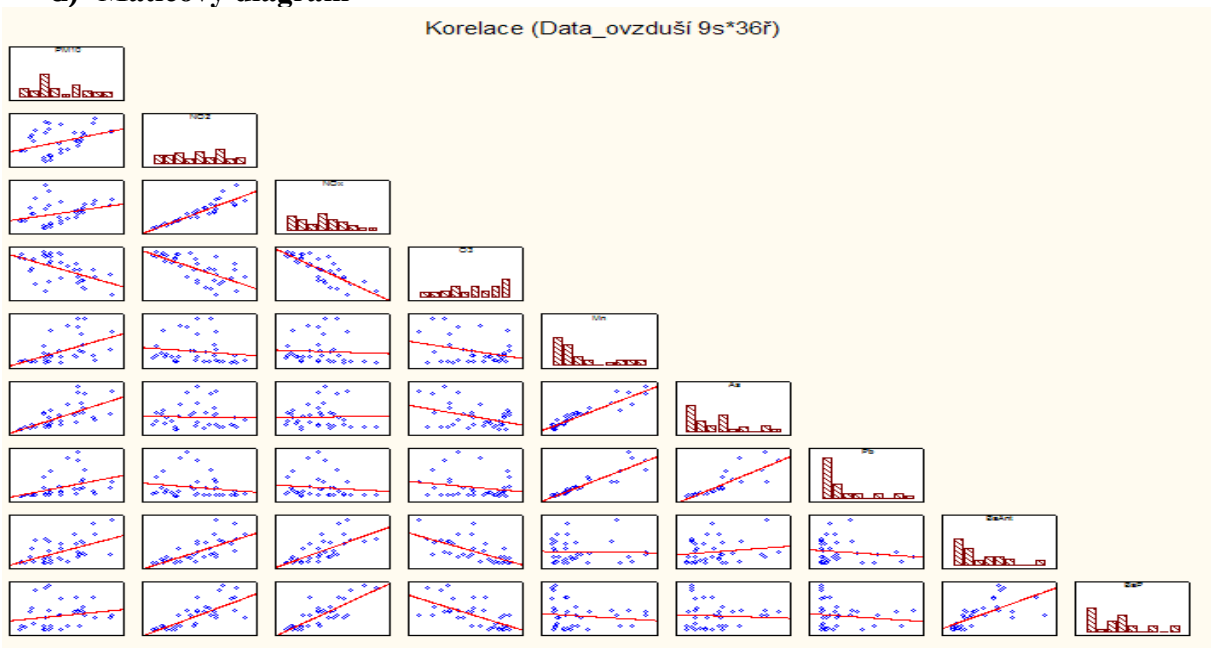
V prvním řádku korelační matice jsou uvedeny korelační koeficienty, v druhém řádku jsou uvedeny hladiny významnosti.

	PM10	NO_2	NO_x	O_3	Mn	As	Pb	BaANT	BaP
PM10	1,000								
NO_2	**0,402 0,0146**	1,000							
NO_x	**0,331 0,0482**	**0,910 0,0000**	1,000						
O_3	**-0,548 0,0005**	**-0,688 0,0000**	**-0,813 0,0000**	1,000					
Mn	**0,520 0,0012**	-0,138 0,4264	-0,055 0,7458	-0,308 0,0679	1,000				
As	**0,677 0,0000**	-0,022 0,9007	0,026 0,8895	-0,408 0,0137	**0,937 0,0000**	1,000			
Pb	**0,401 0,0155**	-0,178 0,3034	-0,111 0,5185	-0,208 0,2225	**0,933 0,0000**	**0,865 0,0000**	1,000		
BaANT	**0,558 0,0004**	**0,724 0,0000**	**0,700 0,0000**	**-0,663 0,0000**	-0,029 0,8665	0,164 0,3389	-0,142 0,4104	1,000	
BaP	0,192 0,2592	**0,758 0,0000**	**0,820 0,0000**	**-0,649 0,0000**	-0,136 0,4304	-0,062 0,7143	-0,134 0,4341	**0,683 0,0000**	1,000

Lineární vztahy mezi jednotlivými škodlivinami, které vyplývající z korelační matice, jsou znázorněny modře tučně. Obecně lze konstatovat:

- prašnost PM10 pozitivně koreluje s oxidy dusíku, všemi kovy a benzo(a)antracenem, negativně s ozónem
- oxid dusičitý pozitivně koreluje s oxidy dusíku a polyaromatickými uhlovodíky, negativně s ozónem
- ozón negativně koreluje s polyaromatickými uhlovodíky
- kovy pozitivně korelují vzájemně mezi sebou
- benzo(a)antracen pozitivně koreluje s benzo(a)pyrenem

d) Maticový diagram



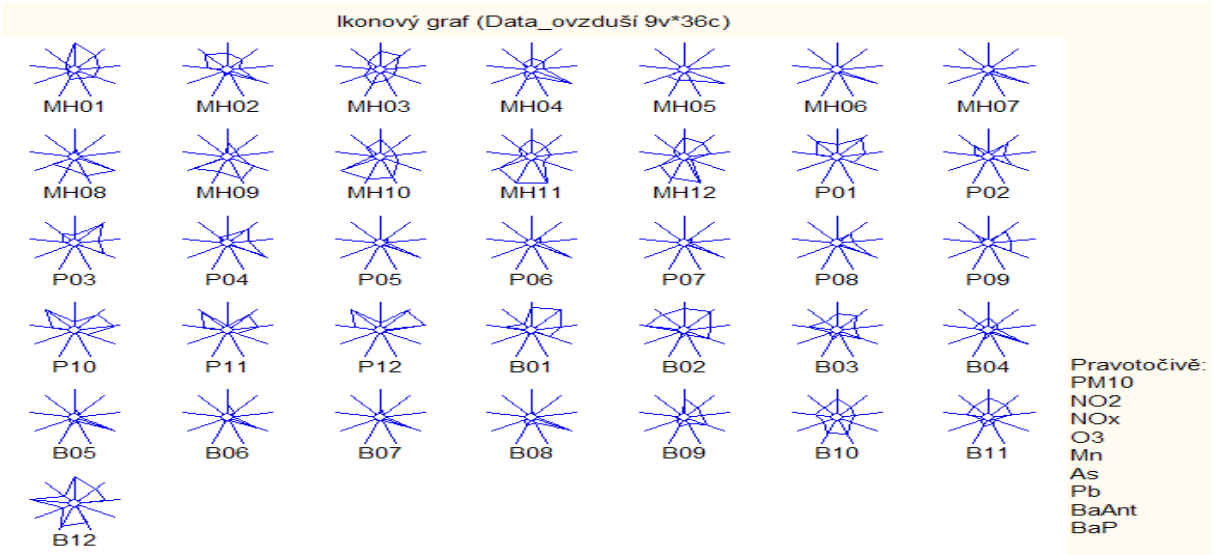
Maticový graf potvrzuje lineární korelace mezi základními a speciálními škodlivinami uvedenými výše.

e) Symbolové grafy

Jednotlivé znaky jsou „kódovány“ s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů či symbolů. Mezi základní typy symbolů patří polygony (sluníčka, hvězdy), tváře, křivky a stromy. Symboly umožňují nalézt podobné objekty.

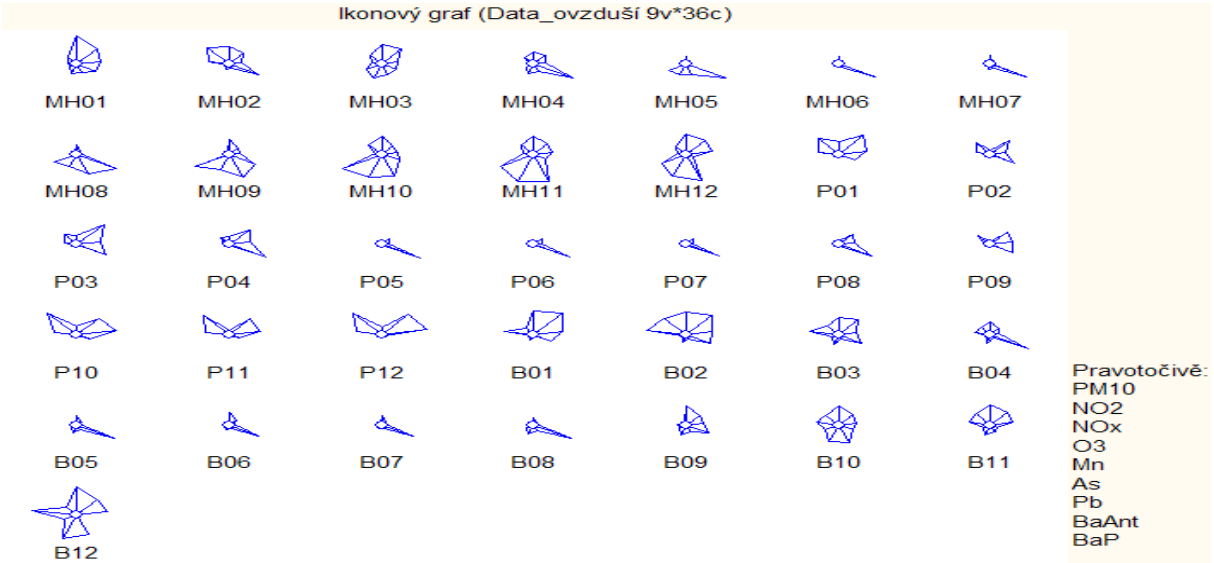
Graf slunečních paprsků

Graf má tvar sluníček, která jsou složena z paprsků spojených do jednoho bodu a úseček spojujících paprsky. Počet paprsků odpovídá počtu proměnných, střed paprsku představuje průměr odpovídající proměnné a délka paprsku 2n násobek směrodatné odchylky.

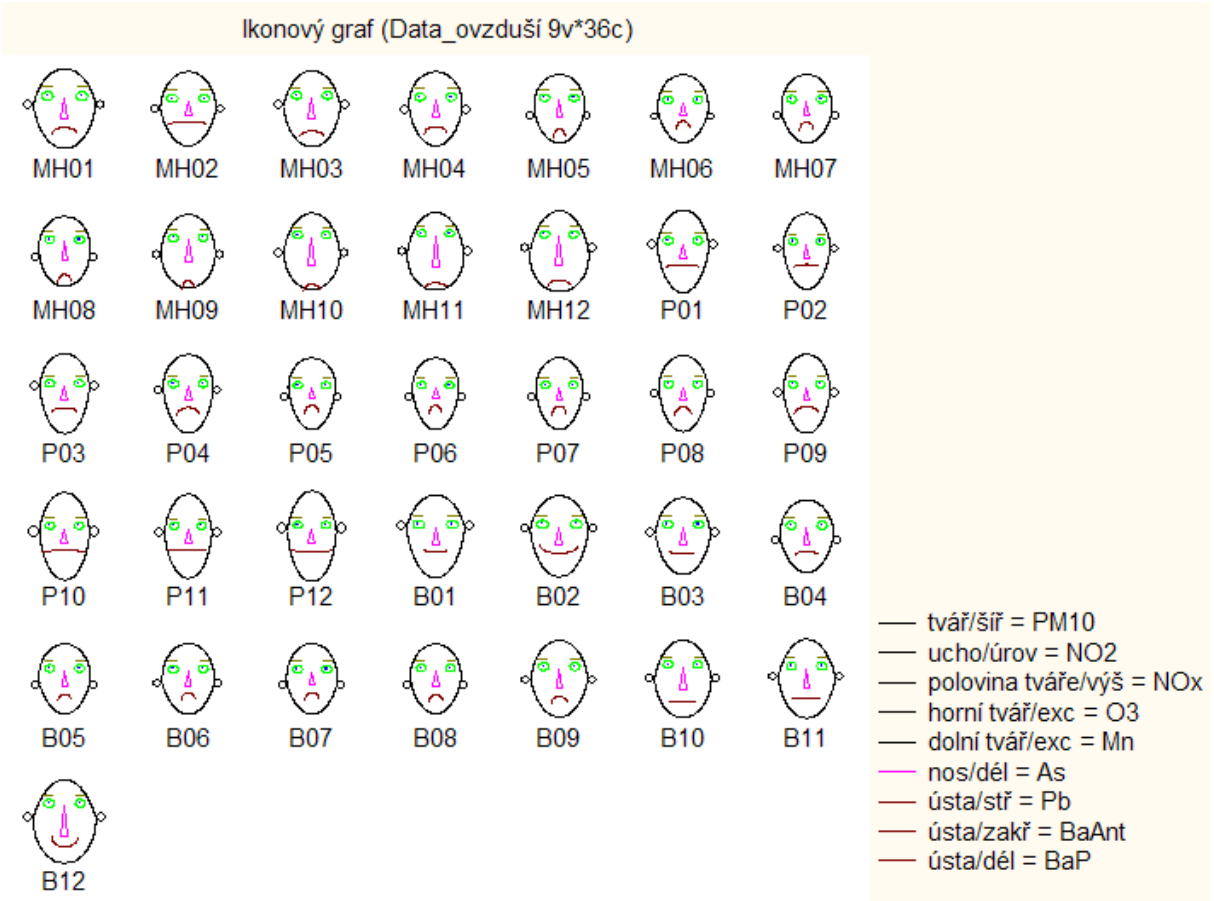


Hvězdicový graf

Podobně jako předchozí graf sestává z paprsků reprezentujících relativní hodnoty proměnných u jednotlivých objektů, které se pro každý objekt spojují v jednom centrálním bodě. Stejně směřující paprsky u různých objektů se liší svoji délkou- nejkratší paprsek indikuje, že u objektu nabývá příslušná proměnná nejmenší hodnoty z celého výběru a naopak



Graf Chernoffových tváří
Graf sestává z obličejů, které charakterizují každou proměnnou objektu nějakým znakem – tvarem tváře, délkou nosu, tvarem úst apod.



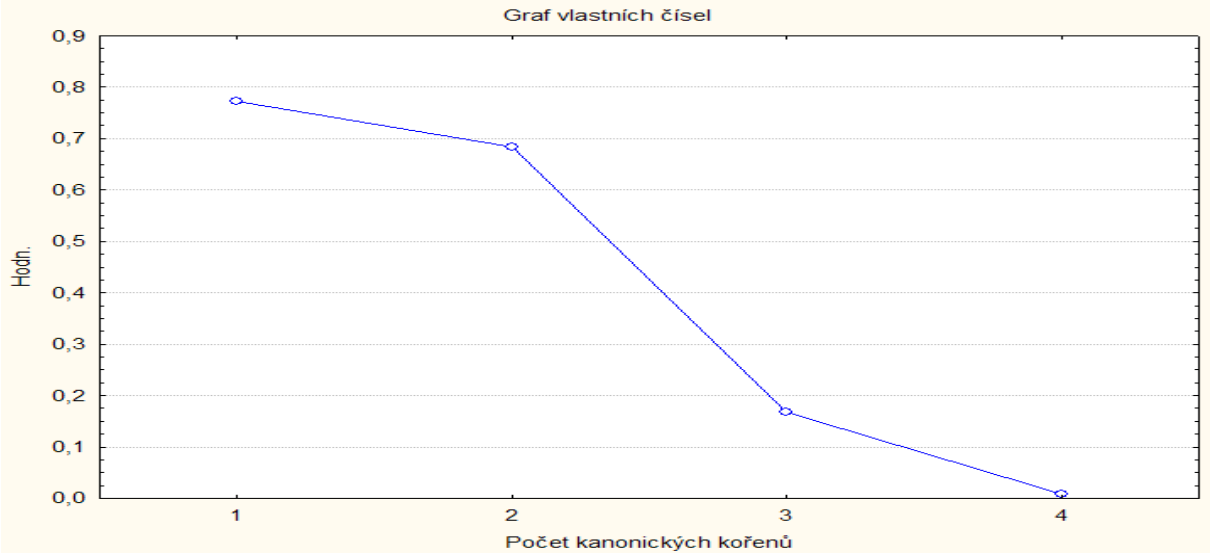
Z vizuálního porovnání symbolových grafů lze usuzovat na nepřítomnost vybočujících objektů.

2.2 Kanonická korelační analýza

a) Zadání kanonických proměnných

Volba znaků a jejich zařazení do obou kanonických proměnných, závisle proměnných U a nezávisle proměnných je v kanonické korelační analýze čistě formální.

Kritériem počtu dvojic kanonických proměnných U a V je Cattelův indexový diagram úpatí vlastních čísel.



Na základě grafu by měly k popisu dat být použity 3 kanonické kořeny.

a) Souhrn kanonické analýzy

Kanonické R: = : 0,8788 p = 0,000

Kanonickou korelací se míní hodnota kanonického korelačního koeficientu R. V našem případě je R statisticky vysoce významné, protože má vypočtenou hladinu významnosti p menší než 0,05.

Závislost (Mn, As, Pb, BaANT, BaP) = f (PM10, NO_2, NO_x, O_3)

		Pravá sada nezávislých kanonických proměnných V	Levá sada závislých kanonických proměnných U
Počet proměnných		4	5
Získaný rozptyl		100,0%	91,2%
Celková redundance		68,4%	57,2%
Proměnné	1	PM10	Mn
	2	NO_2	As
	3	NO_x	Pb
	4	O_3	BaANT
	5		BaP

Získaný rozptyl značí průměrné množství proměnlivosti vybrané ze znaků v obou souborech kanonickými proměnnými. 4 kanonické proměnné určují 100 % rozptylu na pravé straně souboru, to je ze čtyř znaků PM10, NO_2, NO_x a O_3 a 91,2 % rozptylu na levé straně rovnice. Celková redundance ukazuje na velikost celkové korelace mezi znaky na pravé straně rovnice (68,4 %) a na levé straně rovnice (57,2 %). Výsledky potvrzují silný celkový vztah mezi položkami obou souborů.

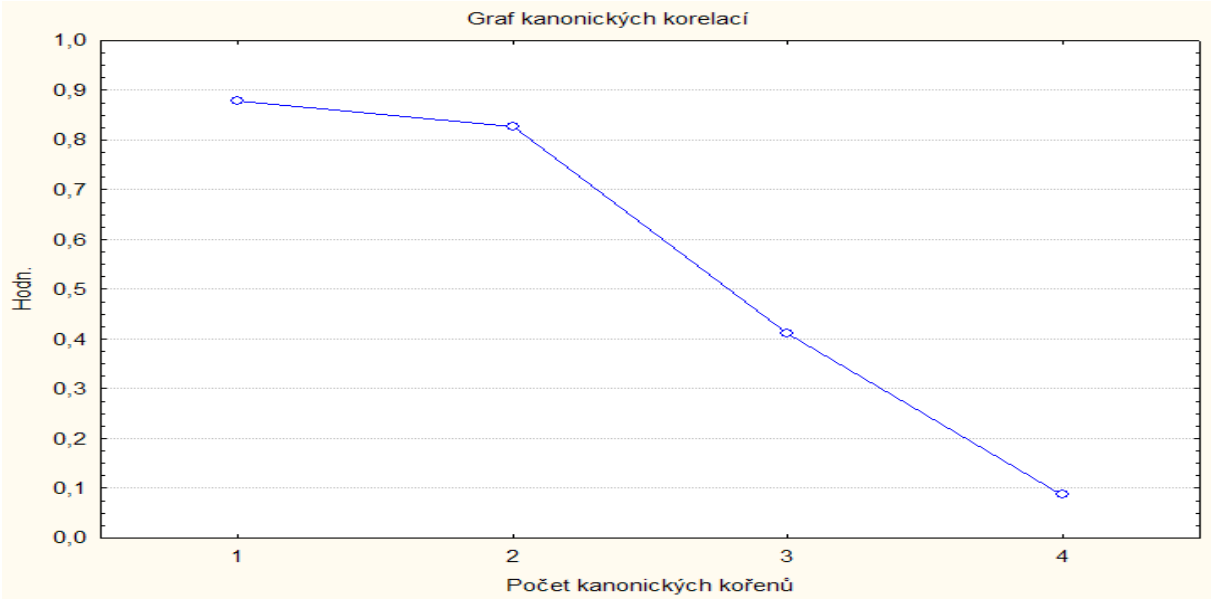
b) Test významnosti kanonických kořenů

Tento test vyšetřuje , zda všechny čtyři kanonické kořeny jsou statisticky významné. Maximální počet kanonických kořenů, který může být z dat vybrán, je roven nejmenšímu počtu znaků užitých v jednotlivých souborech na levé a pravé straně rovnice. Protože v pravém souboru jsou 4 znaky a v levém 5 znaků, budou v kanonické analýze užity 4 kanonické kořeny.

Vypuštěný kořen	Kanonické R	Kanonické D	Chí-kvadrát	df	P	Lambda
žádný	0,878879	0,772428	84,70635	20	0,000000	0,059395
první U_1 a V_1	0,826862	0,683701	40,29772	12	0,000065	0,260994
první U_1 a V_1 a druhý U_2 a V_2	0,410650	0,168634	5,76566	6	0,449965	0,825151
první U_1 a V_1 a druhý U_2 a V_2 a třetí U_3 a V_3	0,086463	0,007476	0,22512	2	0,893544	0,992524

První řádek uvádí situaci, když nebyl odstraněn žádný kanonický kořen - testy jsou vysoce statisticky významné. V druhém řádku je ukázána situace, když byl odstraněn první a nejvýznamnější kořen – testy jsou vysoce statisticky významné. V třetím řádku je ukázána situace, když byl odstraněn první a druhý kořen – testy už nejsou statisticky významné. Můžeme tedy testování ukončit a uzavřít, že statisticky významné jsou první dva kanonické kořeny U_1 a V_1, U_2 a V_2.

Pro hledání počtu potřebných kanonických kořenů slouží i graf kanonických korelací. Výrazný zlom nastává u druhého kanonického kořene.



c) Struktura kanonických faktorů a redundance

Faktorová struktura v pravém souboru

Proměnná	Kořen 1	Kořen 2	Kořen 3	Kořen 4
PM10	0,611957	0,583427	0,44027	-0,97827
NO_2	-0,083574	-0,464411	1,69743	2,10896
NO_x	0,131247	-0,778836	-1,09851	-3,25018
O_3	-0,485682	-0,318143	0,82719	-1,94368

První kanonický kořen se vyznačuje vysokou vahou u znaku PM10, rovněž druhý kanonický kořen se vyznačuje vysokou vahou u znaku PM10.

Tabulka extrahovaného rozptylu na pravé straně

	Získaný rozptyl	Redundance
Kořen 1	0,587770	0,453171
Kořen 2	0,319127	0,217740
Kořen 3	0,072756	0,008995
Kořen 4	0,020346	0,000265

První kanonický kořen vyčísluje 59 % rozptylu znaků PM10, NO_2, NO_x a O_3, 45 % rozptylu znaků Mn, As, Pb, BaANT a BaP. Druhý kořen vyčísluje 32 % rozptylu znaků PM10, NO_2, NO_x a O_3 a 22 % rozptylu znaků Mn, As, Pb, BaANT a BaP Celkem vystihují první dva kanonické kořeny 94 % rozptylu znaků PM10, NO_2, NO_x a O_3.

Faktorová struktura v levém souboru

Proměnná	Kořen 1	Kořen 2	Kořen 3	Kořen 4
Mn	0,198816	-0,248974	2,22019	-3,36551
As	0,789136	0,899465	-0,62533	2,95578
Pb	-0,381288	-0,083784	-1,11609	0,66755
BaANT	0,411912	-0,044727	-1,05679	-1,24068
BaP	0,300314	-0,763190	0,95347	0,82181

První kanonický kořen se vyznačuje vysokou vahou u znaku As, rovněž druhý kanonický kořen se vyznačuje vysokou vahou u znaku As.

Tabulka extrahovaného rozptylu na levé straně

	Získaný rozptyl	Redundance
Kořen 1	0,370973	0,286550
Kořen 2	0,383910	0,262479
Kořen 3	0,137337	0,023160
Kořen 4	0,020404	0,000153

První kanonický kořen vyčísluje 37 % rozptylu znaků Mn, As, Pb, BaANT a BaP a 29 % rozptylu znaků PM10, NO_2, NO_x a O_3. Druhý kořen vyčísluje 38 % rozptylu znaků Mn, As, Pb, BaANT a BaP a 26 % rozptylu znaků PM10, NO_2, NO_x a O_3. Celkem vystihují první dva kanonické kořeny 75 % rozptylu znaků Mn, As, Pb, BaANT a BaP.

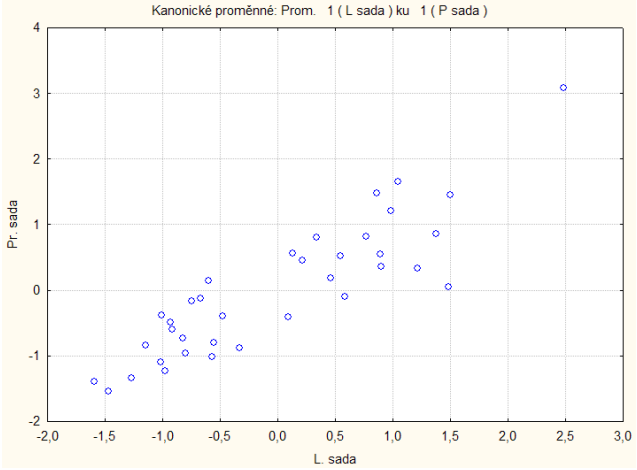
d) Kanonické skóre

Proměnná	Kořen 1	Kořen 2	Kořen 3	Kořen 4
Mn	0,324004	-0,432188	2,97135	-2,41188
As	0,711982	1,002765	-1,02136	1,95088
Pb	-0,289930	-0,232336	-0,17335	-1,54186
BaANT	-0,162863	0,255584	-1,42953	2,45810
BaP	0,390687	0,003998	-1,06761	-0,79626

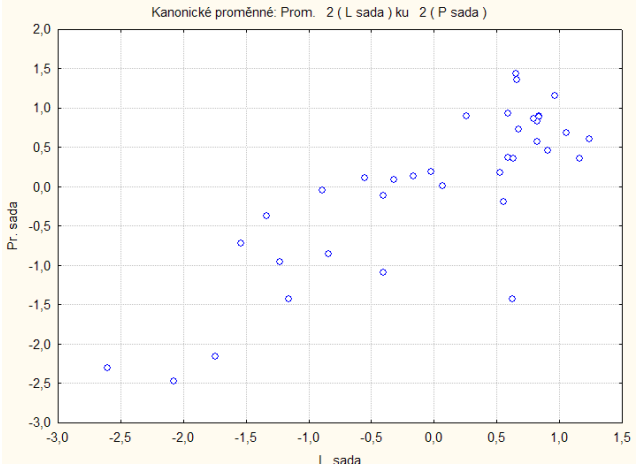
Proměnná	Kořen 1	Kořen 2	Kořen 3	Kořen 4
PM10	0,585698	-0,150261	-0,254826	1,110100
NO_2	0,554020	-0,296481	1,073709	-0,566164
NO_x	-0,493738	1,204014	0,539173	1,189875
O_3	-0,350612	0,039448	1,249104	1,523186

e) Grafy kanonických skóre

Rozptylový diagram kanonických korelací 1 a 1



Rozptylový diagram kanonických korelací 2 a 2



V grafu nenacházíme žádné vybočující body, objekty leží na přímce, nikoliv na křivce tvaru S nebo U. Rovněž nejsou indikovány žádné shluky.

3. *Závěr:*

Pro statistické zpracování souboru koncentrací základních škodlivin (prašnost PM10, NO_2, NO_x, O3), speciálních anorganických škodlivin (kovy) a speciálních organických škodlivin škodlivin (PAU) v ovzduší byla použita metoda exploratorní analýza dat a metoda kanonické korelační analýzy.

Exploratorní analýza dat prokázala existenci lineární vztahů:

- prašnost PM10 pozitivně koreluje s oxidy dusíku, všemi kovy a benzo(a)antracenem, negativně s ozónem
- oxid dusičitý pozitivně koreluje s oxidy dusíku a polyaromatickými uhlovodíky, negativně s ozónem
- ozón negativně koreluje s polyaromatickými uhlovodíky
- kovy pozitivně korelují vzájemně mezi sebou
- benzo(a)antracen pozitivně koreluje s benzo(a)pyrenem

Největší proměnlivost v datech základních škodlivin vykazují znaky PM10 a O_3, v datech speciálních škodlivin znak Pb. Z vizuálního porovnání symbolových grafů lze usuzovat na nepřítomnost vybočujících objektů.

Metoda kanonické korelační analýzy slouží pro vyhledání lineárních kombinací znaků dvou skupin, tj. hypotetických kanonických proměnných, které vedou k maximálním vzájemným korelacím. V našem případě byly jako statisticky významné určeny první dva kanonické kořeny U_1 a V_1, U_2 a V_2..
První kanonický kořen vyčísluje 59 % rozptylu znaků PM10, NO_2, NO_x a O_3, 45 % rozptylu znaků Mn, As, Pb, BaANT a BaP. Druhý kořen vyčísluje 32 % rozptylu znaků PM10, NO_2, NO_x a O_3 a 22 % rozptylu znaků Mn, As, Pb, BaANT a BaP Celkem vystihují první dva kanonické kořeny 94 % rozptylu znaků PM10, NO_2, NO_x a O_3.
První kanonický kořen U_1 vyčísluje 37 % rozptylu znaků Mn, As, Pb, BaANT a BaP a 29 % rozptylu znaků PM10, NO_2, NO_x a O_3. Druhý kořen vyčísluje 38 % rozptylu znaků Mn, As, Pb, BaANT a BaP a 26 % rozptylu znaků PM10, NO_2, NO_x a O_3. Celkem vystihují první dva kanonické kořeny 75 % rozptylu znaků Mn, As, Pb, BaANT a BaP.
Hodnota kanonického korelačního koeficientu je R=0,879, nazývá se první kanonickou korelací a představuje nejsilnější korelaci mezi lineárními kombinacemi znaků dvou skupin. První kanonický kořen se vyznačuje vysokou vahou u znaku PM10 a As, rovněž druhý kanonický kořen se vyznačuje vysokou vahou u znaku PM10 a As.

Úloha 3.: Klasifikace analýzou vícerozměrných dat – korespondenční analýza

1. Zadání

Ve statistické ročence byly vydány výsledky sčítání lidu. Jedna z oblastí se týkala vzdělání podle deklarované národnosti obyvatelstva. Tabulka obsahuje počet obyvatel ze skupiny 1000 obyvatel s danou národností a nejvyšším ukončeným vzděláním. Cílem úlohy je klasifikovat obyvatele z těchto dvou hledisek.

1.1 Data

Národnost	Nejvyšší ukončené vzdělání				
	základní	střední bez maturity	úplné střední s maturitou	vyšší odborné	vysokoškolské
česká	227	385	253	35	88
moravská	216	398	254	35	92
slezská	194	373	267	40	118
slovenská	359	314	181	24	94
polská	290	347	226	31	93
německá	376	382	140	27	57
romská	654	173	43	7	16
maďarská	418	313	141	21	79
ukrajinská	200	272	301	54	124
ruská	130	88	253	84	407
vietnamská	315	217	289	33	65

1.2 Užitý program

STATISTICA, verze 7

2. Řešení

2.1 Řádkové profily v procentech

Národnost	Nejvyšší ukončené vzdělání					Součet
	základní	střední bez maturity	úplné střední s maturitou	vyšší odborné	vysokoškolské	
česká	22,98	38,97	25,61	3,54	8,91	100,00
moravská	21,71	40,00	25,53	3,52	9,25	100,00
slezská	19,56	37,60	26,92	4,03	11,90	100,00
slovenská	36,93	32,30	18,62	2,47	9,67	100,00
polská	29,38	35,16	22,90	3,14	9,42	100,00
německá	38,29	38,90	14,26	2,75	5,80	100,00
romská	73,24	19,37	4,82	0,78	1,79	100,00
maďarská	43,00	32,20	14,51	2,16	8,13	100,00
ukrajinská	21,03	28,60	31,65	5,68	13,04	100,00
ruská	13,51	9,15	26,30	8,73	42,31	100,00
vietnamská	34,28	23,61	31,45	3,59	7,07	100,00
Průměr	31,84	30,74	22,12	3,68	11,62	100,00

Tabulka přináší řádkové profily a celkový řádkový profil (v tabulce označený heslem Průměr) vyjádřené v procentech.

2.2 Sloupcové profily v procentech

Národnost	Nejvyšší ukončené vzdělání					Průměr
	základní	střední bez maturity	úplné střední s maturitou	vyšší odborné	vysokoškolské	
česká	6,72	11,80	10,78	8,95	7,14	9,31
moravská	6,39	12,20	10,82	8,95	7,46	9,38
slezská	5,74	11,43	11,37	10,23	9,57	9,35
slovenská	10,62	9,63	7,71	6,14	7,62	9,16
polská	8,58	10,64	9,63	7,93	7,54	9,30
německá	11,13	11,71	5,96	6,91	4,62	9,25
romská	19,35	5,30	1,83	1,79	1,30	8,41
maďarská	12,37	9,60	6,01	5,37	6,41	9,16
ukrajinská	5,92	8,34	12,82	13,81	10,06	8,96
ruská	3,85	2,70	10,78	21,48	33,01	9,06
vietnamská	9,32	6,65	12,31	8,44	5,27	8,66
Součet	100,00	100,00	100,00	100,00	100,00	100,00

Tabulka přináší sloupcové profily a celkový sloupcový průměr (v tabulce označený heslem Průměr) vyjádřené v procentech.

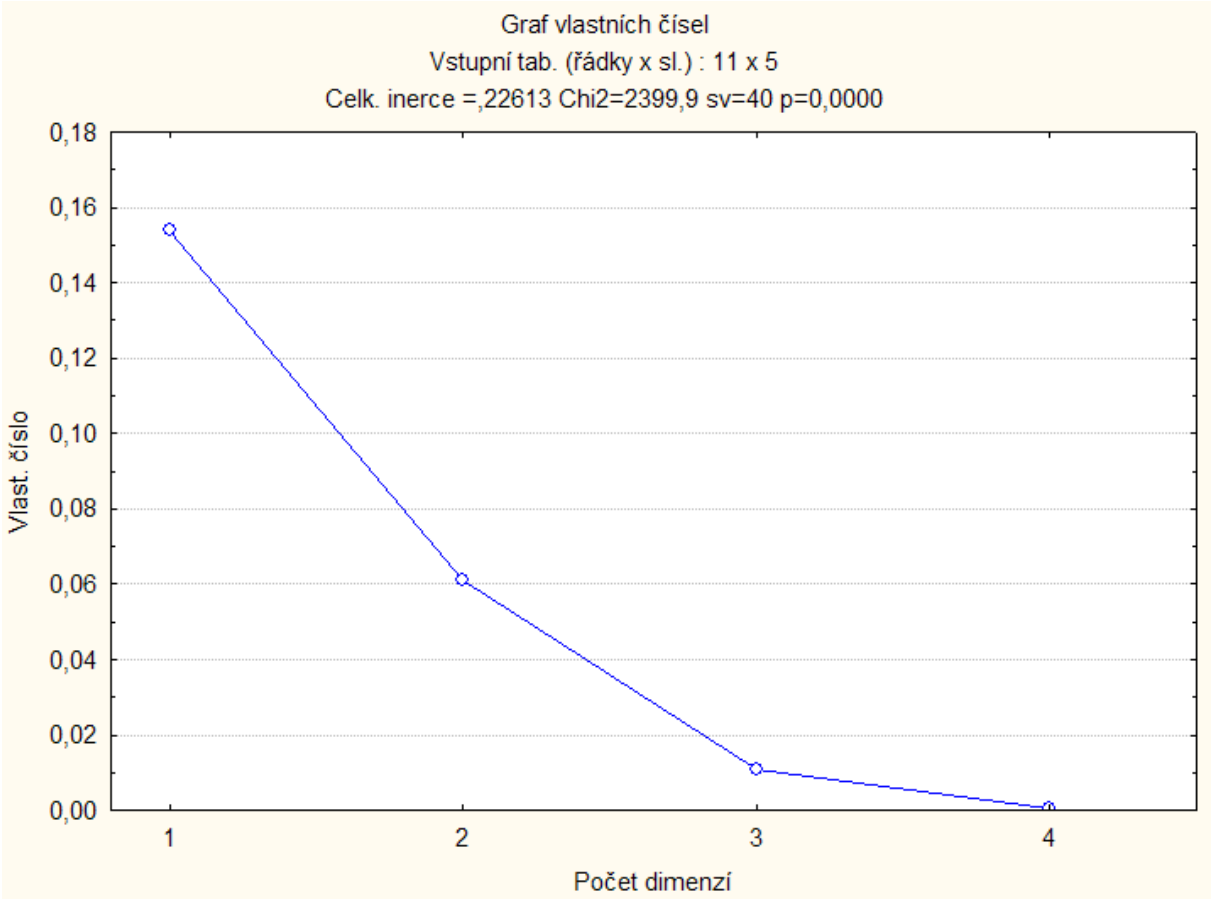
2.3 Hledání počtu projekčních dimenzí

Protože korespondenční analýza promítá řádkové a sloupcové profily do sníženého počtu dimenzí, obvykle do dvojrozměrné roviny, je třeba určit, jak dobře se podařilo snížení dimenze provést.

Počet dimenzí	Singul. číslo	Vlastní číslo	Procent inerce	Kumulované procento	Chí kvadrát
1	**0,392147**	**0,153780**	**68,00540**	**68,0054**	**1632,063**
2	**0,246791**	**0,060906**	**26,93423**	**94,9396**	**646,395**
3	0,104993	0,011024	4,87494	99,8146	116,994
4	0,020477	0,000419	0,18543	100,0000	4,450

Kumulativní % představuje objem celkové informace, zobrazený daným počtem dimenzí. Platí pravidlo, že první dvě dimenze by měly pokrývat alespoň 90 % celkové variability v datech. V našem případě první dvě dimenze pokrývají 95 % celkové informace, což znamená, že redukce dimenzí na dvě způsobí pouze 5 % ztracené informace.

Počet dimenzí je možné určit pomocí Cattelova indexového grafu úpatí vlastních čísel.



V našem případě je charakteristický zlom u počtu dimenzí rovnajícímu se 2.

2.4 Zobrazení řádkového profilu a příspěvek do inercie

Kategorie	Souřad. Dim. 1	Souřad Dim. 2	Masa	Kvalita	Relat. inerce	Inerce Dim. 1	Cos^2 Dim. 1	Inerce Dim. 2	Cos^2 Dim. 2
základní vzdělání	0,416	-0,240	0,318	0,997	0,326	0,358	0,747	0,302	0,250
střední bez maturity	0,106	0,259	0,307	0,868	0,122	0,022	0,125	0,338	0,743
úplné střední s maturitou	-0,247	0,183	0,221	0,784	0,118	0,088	0,507	0,121	0,277
vyšší odborné	-0,516	-0,068	0,037	0,957	0,046	0,064	0,940	0,003	0,016
vysokoškolské	-0,786	-0,352	0,116	0,985	0,387	0,467	0,821	0,236	0,164

Masa : váha, která představuje procentuální podíl z celé tabulky v dané kategorii, tj. u řádkové váhy představovaný tímto řádkem. Tato váha vyjadřuje poměr řádkové četnosti a celkové četnosti tabulky.

Kvalita: díl variability v dotyčném profilu, který je reprodukován oběma osami. Ukazuje, jak dobře je profil zobrazen v prostoru definovaných dvou os.

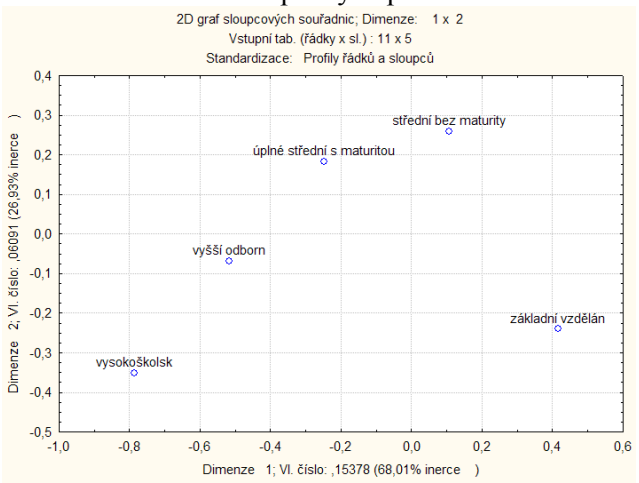
Relat. inerce: podíl celkové inercie na profilu.

Inercie: Pearsonovo χ^2 dělené n (sumou všech četností prvků tabulky). Je to vážený průměr χ^2 vzdáleností mezi řádkovými profily a jejich průměrným profilem.

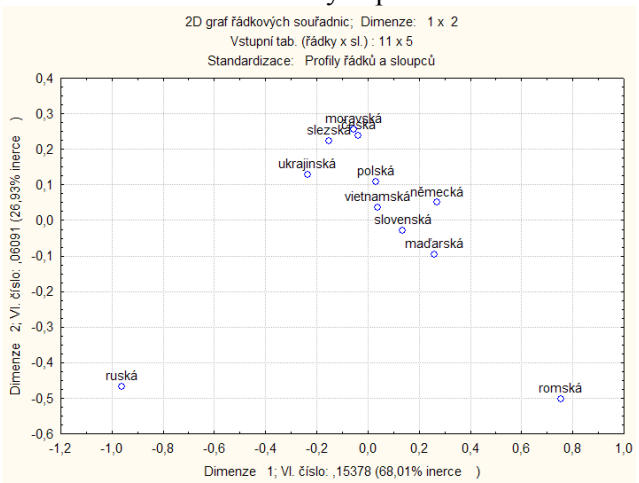
2.5 Diagnostické grafy

Grafy korespondenční analýzy zobrazují a diagnostikují sloupcové profily a řádkové profily. Každý bod v těchto grafech diagnostikuje celý profil promítnutý do roviny vybraných dvou os.

Graf sloupcových profilů



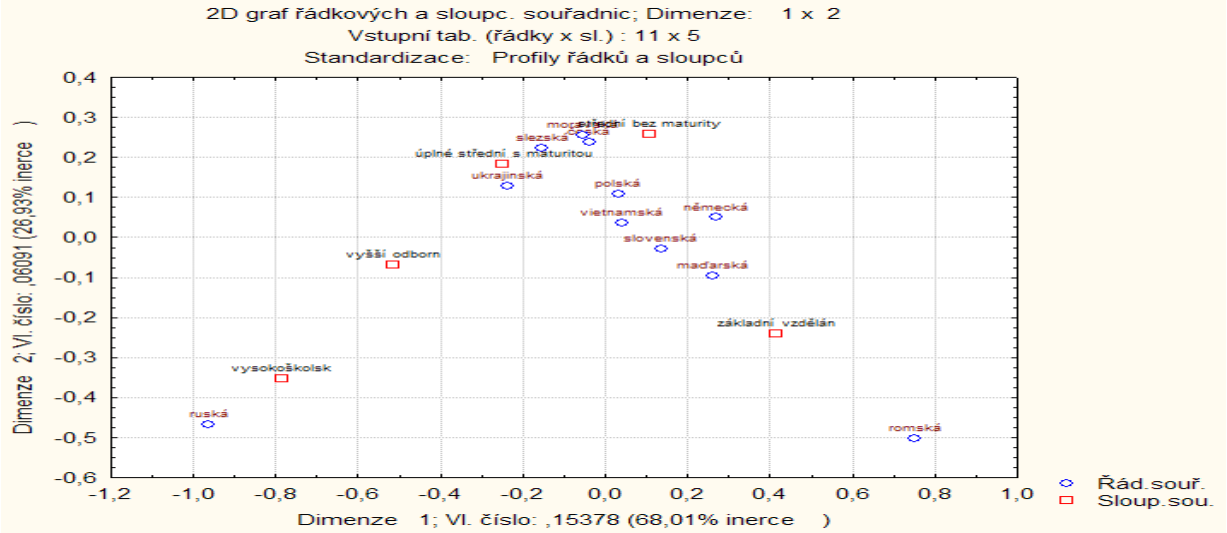
Graf řádkových profilů



Graf sloupcových profilů zobrazuje první hlavní komponentu, které odděluje obyvatele s vysokoškolským vzděláním od obyvatel se základním vzděláním. Druhá hlavní

komponenta odděluje obyvatele s vysokoškolským vzděláním od obyvatel se středním vzděláním.
Graf řádkových profilů zobrazuje první hlavní komponentu, které odděluje obyvatele s ruskou národností od romské národnosti, druhá hlavní komponenta odděluje obyvatele s ruskou národností od romské národnosti a skupiny obyvatel se zbývajícím národnostmi.
Ze vzájemného posouzení grafů sloupcových a řádkových profilů lze vyčíst nižší vzdělání u obyvatel romské národnosti, naopak vysokoškolské vzdělání u obyvatel s ruskou národností.

Graf sloupcových a řádkových profilů



V grafu sloupcových a řádkových profilů nemůžeme vysvětlovat vzdálenost mezi řádkovými a sloupcovými body (kategoriemi). Můžeme si všimnou jenom určité analogie z předešlých grafů. Lze posoudit jenom body (kategorie), které se nacházejí blízko sebe – kategorie obyvatelé se středním vzdělání bez maturity je podobná svou polohou kategorii obyvatelé s českou a moravskou národností, kategorie obyvatelé s neúplným středním vzděláním je podobná svou polohou kategorii obyvatelé s ukrajinskou národností.

3. Závěr:

Korespondenční analýza promítá řádkové a sloupcové profily do sníženého počtu dimenzí, obvykle do dvojrozměrné roviny. Jak dobře se podařilo snížení dimenze provést, ukazuje např. kumulativní % inerce . Platí pravidlo, že první dvě dimenze by měly pokrývat alespoň 90 % celkové variability v datech. V našem případě první dvě dimenze pokrývají 95 % celkové informace, což znamená, že redukce dimenzí na dvě způsobí pouze 5 % ztracené informace. Počet dimenzí – 2 – bylo určeno i pomocí Cattelova indexového grafu úpatí vlastních čísel.
Cílem korespondenční analýzy je sestrojení grafů ve dvojrozměrném prostoru a grafická prezentace řádků a sloupců . V našem případě grafy CA názorně klasifikují obyvatele podle vzdělání, hlavně rozlišuje obyvatele s vysokoškolským vzděláním a základním vzděláním. Dále rozlišuje obyvatele s národností ruskou od národnosti romské a ostatních národností. Převládající základní vzdělání bylo potvrzeno u obyvatel s romskou národností etnika, naopak převládající vysokoškolské vzdělání bylo potvrzeno u obyvatel s ruskou národností. U ostatních národností převládá střední vzdělání.