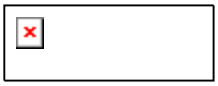
Univerzita Pardubice
Fakulta chemicko-technologická

Licenční studium

Statistické zpracování dat

3.5 Klasifikace analýzou vícerozměrných dat

RNDr. Lada Kovaříková



České technologické centrum pro anorganické pigmenty a.s.

Přerov 20. 4. 2006

Úloha č. 1: Klasifikace diskriminační analýzou

Název úlohy: pigmenty

1.1 Zadání

Data obsahují výsledky analýz tří skupin bílého pigmentu na bázi oxidu titaničitého (chemické a fyzikální vlastnosti pigmentů). Pigmenty jsou komerčně rozděleny do skupin na základě technologie povrchové úpravy. Proveďte **klasifikaci diskriminační analýzou** (DA); ověřte, zda je možné použít dosavadní třídění na základě použitých analýz; proveďte roztřídění (klasifikaci) vzorků, u kterých není znám původ (technologie povrchové úpravy).

1.2 Data

Tabulka č. 1 obsahuje výsledky analýz vzorků, u kterých známe zařazení do tříd (typy RGA, RGX, RGU) – tréninková data. Tabulka č. 2 obsahuje výsledky analýz neznámých vzorků, které potřebujeme zatřídit.

Diskriminátory:

Ti	obsah Ti (%)
TL	obsah těkavých látek (%)
SpO	spotřeba oleje (g/100 g pigmentu)
Barv	barvivost
Podt	podtón
Si	obsah Si (%)
Al	obsah Al (%)

1.3 Program

Úloha byla řešena programem: **R Version 2.1.0**

Tabulka č. 1: data pro úlohu č. 1 – pigmenty – data se známým zatříděním

ID	Ti	TL	SpO	Barv	Podt	Si	Al	klas	ID	Ti	TL	SpO	Barv	Podt	Si	Al	klas	ID	Ti	TL	SpO	Barv	Podt	Si	Al	klas
1	93,78	0,5	24,2	1870	10	2,174	3,511	RGA	51	96,18	0,3	19,1	1860	9	0,749	2,703	RGU	101	97,99	0,2	20	1860	15	0,48	1,115	RGX
2	93,87	0,5	23,5	1870	10	2,082	3,522	RGA	52	96,25	0,3	18,3	1860	9	0,724	2,634	RGU	102	97,99	0,2	19,5	1850	15	0,488	1,114	RGX
3	94,01	0,5	23,4	1860	10	2,008	3,44	RGA	53	96,19	0,3	19	1860	10	0,743	2,67	RGU	103	98,16	0,2	18,1	1850	15	0,405	0,999	RGX
4	93,9	0,5	24,1	1860	10	2,042	3,455	RGA	54	96,16	0,3	19	1860	9	0,735	2,69	RGU	104	98,29	0,2	19,8	1855	15	0,435	0,862	RGX
5	93,71	0,5	23,7	1860	10	2,092	3,577	RGA	55	96,14	0,3	17,9	1870	9	0,745	2,713	RGU	105	97,6	0,2	19,8	1840	13	0,396	1,569	RGX
6	93,81	0,5	24,3	1870	9	2,064	3,543	RGA	56	96,14	0,3	17,9	1870	9	0,745	2,713	RGU	106	97,86	0,2	20,5	1850	13	0,436	1,27	RGX
7	93,85	0,5	24,2	1860	9	2,04	3,519	RGA	57	96,16	0,3	18,6	1870	9	0,745	2,712	RGU	107	98,11	0,2	20	1855	14	0,469	0,983	RGX
8	93,88	0,3	23,1	1850	9	2,004	3,497	RGA	58	96,16	0,4	19,3	1860	9	0,735	2,698	RGU	108	97,95	0,2	20,6	1840	14	0,567	1,081	RGX
9	93,83	0,4	24,2	1860	10	2,052	3,513	RGA	59	96,16	0,4	19,3	1860	9	0,735	2,698	RGU	109	97,72	0,2	20,7	1850	14	0,543	1,317	RGX
10	93,83	0,4	24,2	1860	10	2,052	3,513	RGA	60	96,23	0,2	19,3	1860	10	0,733	2,642	RGU	110	97,76	0,2	20,2	1840	15	0,597	1,251	RGX
11	93,79	0,4	22,9	1860	10	2,107	3,475	RGA	61	96,21	0,3	19,1	1860	10	0,735	2,666	RGU	111	97,88	0,2	21,3	1830	15	0,615	1,103	RGX
12	93,92	0,4	22,9	1860	10	2,041	3,407	RGA	62	96,21	0,3	19,2	1860	10	0,738	2,667	RGU	112	98,17	0,2	19,7	1830	15	0,499	0,941	RGX
13	93,9	0,4	22,7	1860	10	1,999	3,441	RGA	63	96,21	0,3	19,2	1860	10	0,738	2,667	RGU	113	97,36	0,2	20	1840	15	0,389	1,635	RGX
14	93,78	0,4	22,9	1870	10	2,051	3,54	RGA	64	96,19	0,3	19	1860	10	0,74	2,681	RGU	114	98,25	0,2	19,1	1840	15	0,319	0,999	RGX
15	95,13	0,3	21,6	1860	14	1,576	2,724	RGA	65	96,23	0,3	19,1	1870	10	0,747	2,641	RGU	115	98,31	0,2	20,2	1840	15	0,33	0,968	RGX
16	93,56	0,4	24,4	1860	11	2,148	3,714	RGA	66	96,14	0,3	19,5	1860	10	0,748	2,715	RGU	116	98,29	0,1	18,9	1850	15	0,365	0,972	RGX
17	93,68	0,4	23,4	1860	11	2,092	3,656	RGA	67	96,14	0,3	19,5	1860	10	0,748	2,715	RGU	117	98,27	0,2	18,6	1860	15	0,364	0,978	RGX
18	93,67	0,5	24,4	1860	11	2,089	3,667	RGA	68	96,18	0,3	18,3	1860	10	0,748	2,69	RGU	118	98,28	0,2	18,4	1850	15	0,385	0,966	RGX
19	93,62	0,4	23,9	1860	11	2,118	3,696	RGA	69	96,18	0,3	18,3	1860	10	0,748	2,69	RGU	119	98,32	0,2	18,4	1850	15	0,338	0,938	RGX
20	93,86	0,5	24,3	1860	12	2,074	3,483	RGA	70	96,15	0,3	18,2	1860	10	0,757	2,704	RGU	120	98,21	0,2	18,4	1840	15	0,451	0,95	RGX
21	93,79	0,5	24,2	1860	11	2,074	3,575	RGA	71	96,15	0,3	18,2	1860	10	0,757	2,704	RGU	121	98,19	0,2	18,3	1850	15	0,455	0,961	RGX
22	93,97	0,5	24,4	1850	11	1,943	3,515	RGA	72	96,17	0,3	18,8	1860	10	0,746	2,688	RGU	122	98,28	0,2	18,4	1850	14	0,346	0,977	RGX
23	93,92	0,5	24,1	1860	11	2,02	3,483	RGA	73	96,17	0,3	18,8	1860	10	0,746	2,688	RGU	123	98,13	0,2	18,4	1850	14	0,357	1,075	RGX
24	93,79	0,4	23,9	1860	12	2,094	3,501	RGA	74	96,19	0,3	19,3	1860	9	0,747	2,681	RGU	124	98,25	0,2	18,4	1855	14	0,38	0,945	RGX
25	93,69	0,4	23,8	1860	12	2,149	3,546	RGA	75	96,19	0,3	19,3	1860	9	0,747	2,681	RGU	125	98,24	0,2	17,8	1860	14	0,412	0,921	RGX
26	93,75	0,5	24	1860	12	2,138	3,493	RGA	76	96,21	0,3	19	1860	9	0,749	2,657	RGU	126	98,23	0,2	17,8	1860	13	0,444	0,923	RGX
27	96,2	0,3	18,6	1850	8	0,828	2,566	RGU	77	96,17	0,2	20,4	1850	9	0,762	2,687	RGU	127	98,24	0,2	18,3	1880	13	0,402	0,925	RGX
28	96,24	0,3	19,7	1850	8	0,799	2,546	RGU	78	96,18	0,2	20,3	1850	9	0,759	2,683	RGU	128	98,24	0,2	18,3	1880	13	0,402	0,925	RGX
29	96,24	0,3	19,7	1850	8	0,799	2,546	RGU	79	96,14	0,3	19,4	1850	9	0,761	2,691	RGU	129	98,26	0,2	18,7	1880	13	0,412	0,915	RGX
30	96,24	0,3	19,1	1850	8	0,784	2,557	RGU	80	96,15	0,3	19,4	1850	9	0,765	2,693	RGU	130	98,24	0,2	18,7	1890	14	0,417	0,915	RGX
31	96,2	0,3	19,6	1860	8	0,828	2,569	RGU	81	96,15	0,3	19,4	1850	9	0,765	2,693	RGU	131	98,25	0,19	18,3	1880	13	0,429	0,919	RGX
32	96,2	0,3	19,6	1860	8	0,828	2,569	RGU	82	96,16	0,3	19,1	1860	9	0,761	2,681	RGU	132	98,25	0,19	18,3	1885	13	0,429	0,919	RGX
33	96,29	0,2	18,4	1860	8	0,769	2,556	RGU	83	96,19	0,3	19,4	1850	9	0,75	2,676	RGU	133	98,23	0,16	17,9	1870	14	0,411	0,933	RGX
34	96,29	0,3	18,3	1860	8	0,769	2,556	RGU	84	96,18	0,3	19,9	1850	9	0,775	2,658	RGU	134	98,18	0,2	18,2	1860	14	0,423	0,943	RGX
35	96,29	0,3	19,2	1850	9	0,646	2,646	RGU	85	96,2	0,3	19,5	1850	9	0,761	2,654	RGU	135	98,18	0,2	18,3	1860	13	0,456	0,942	RGX
36	96,29	0,3	19,2	1850	9	0,646	2,646	RGU	86	96,18	0,3	19,3	1860	10	0,763	2,667	RGU	136	98,13	0,2	18,4	1860	13	0,476	0,954	RGX
37	96,21	0,3	19,2	1850	9	0,752	2,629	RGU	87	96,14	0,3	19,1	1860	9	0,813	2,65	RGU	137	98,22	0,2	18,4	1860	14	0,399	0,957	RGX
38	96,14	0,3	18,6	1860	10	0,761	2,681	RGU	88	96,14	0,3	19,1	1850	9	0,805	2,641	RGU	138	98,27	0,2	18,4	1860	13	0,333	0,964	RGX
39	96,14	0,3	18,6	1860	10	0,761	2,681	RGU	89	96,23	0,29	19,1	1850	9	0,759	2,633	RGU	139	98,21	0,2	18,2	1860	13	0,41	0,966	RGX
40	96,14	0,3	18,6	1860	10	0,761	2,681	RGU	90	96,17	0,2	19,6	1850	9	0,784	2,643	RGU	140	98,23	0,2	18,3	1850	13	0,377	0,979	RGX
41	96,17	0,3	18,8	1860	10	0,691	2,742	RGU	91	96,17	0,2	19,6	1850	9	0,784	2,643	RGU	141	98,23	0,2	18,3	1860	14	0,377	0,979	RGX
42	96,19	0,3	18,4	1870	10	0,752	2,67	RGU	92	96,17	0,3	19,1	1860	9	0,782	2,672	RGU	142	98,24	0,2	18,8	1860	13	0,352	0,999	RGX
43	96,13	0,3	18,4	1880	10	0,753	2,717	RGU	93	96,17	0,3	19,1	1860	9	0,782	2,672	RGU	143	98,21	0,2	18,4	1840	14	0,371	1,009	RGX
44	96,13	0,3	19,1	1870	10	0,768	2,711	RGU	94	96,19	0,3	18,8	1860	9	0,763	2,649	RGU	144	98,21	0,2	19	1860	13	0,354	1,014	RGX
45	96,15	0,3	18,7	1870	9	0,758	2,708	RGU	95	98,15	0,2	19,2	1860	15	0,333	1,068	RGX									
46	96,15	0,3	18,7	1870	9	0,758	2,708	RGU	96	98,11	0,2	19,5	1860	15	0,357	1,073	RGX									
47	96,15	0,3	19,1	1860	9	0,764	2,69	RGU	97	98,04	0,2	19,5	1860	15	0,383	1,123	RGX									
48	96,18	0,3	19,2	1860	9	0,75	2,68	RGU	98	97,6	0,2	20,2	1855	15	0,504	1,409	RGX									
49	96,21	0,3	19,4	1850	9	0,749	2,659	RGU	99	98	0,2	19,1	1855	15	0,475	1,091	RGX									
50	96,12	0,3	19,5	1860	9	0,756	2,7	RGU	100	98,07	0,2	19,5	1850	15	0,438	1,077	RGX									

Tabulka č. 2: data pro úlohu č. 1 – pigmenty – data s neznámým zatříděním

ID	Ti	TL	SpO	Barv	Podt	Si	Al
1	93,94	0,6	24,6	1850	9	1,548	3,664
2	95,96	0,4	20	1850	9	0,843	2,721
3	93,74	0,4	21,9	1870	12	2,077	3,526
4	98,07	0,2	18,5	1860	14	0,39	1,112
5	95,66	0,4	20,4	1850	10	0,981	2,869
6	95,96	0,4	20	1850	9	0,843	2,721
7	93,2	0,7	24,5	1860	11	2,181	3,904
8	95,96	0,4	20	1850	9	0,843	2,721

1.4 Řešení – EDA, DA

1.4.1 Protokoly

Odhady diskriminačních koeficientů „tréninkových dat“

```
Prior probabilities of groups:
      RGA       RGU       RGX
0.1805556 0.4722222 0.3472222

Group means:
          Ti         TL       SpO     Barv      Podt       Si       Al
RGA 93.85731 0.4423077 23.71923 1860.769 10.615385 2.050885 3.500231
RGU 96.18324 0.2939706 19.05588 1858.529  9.220588 0.756647 2.664824
RGX 98.12160 0.1968000 18.99000 1855.600 14.140000 0.419700 1.036220

Coefficients of linear discriminants:
              LD1          LD2
Ti    4.194004337 -24.18664821
TL   -0.782412815   2.99604972
SpO   0.558599475   0.58557614
Barv -0.002635817   0.01281212
Podt -0.040914643   0.59517322
Si   -8.867375274 -13.35864803
Al    0.446707367 -33.42477854

Proportion of trace:
   LD1    LD2
0.6422 0.3578
```

Tréninková data – zařazení do tříd, klasifikační matice

```
tp    RGA RGU RGX
  RGA  26   0   0
  RGU   0  68   0
  RGX   0   0  50
```

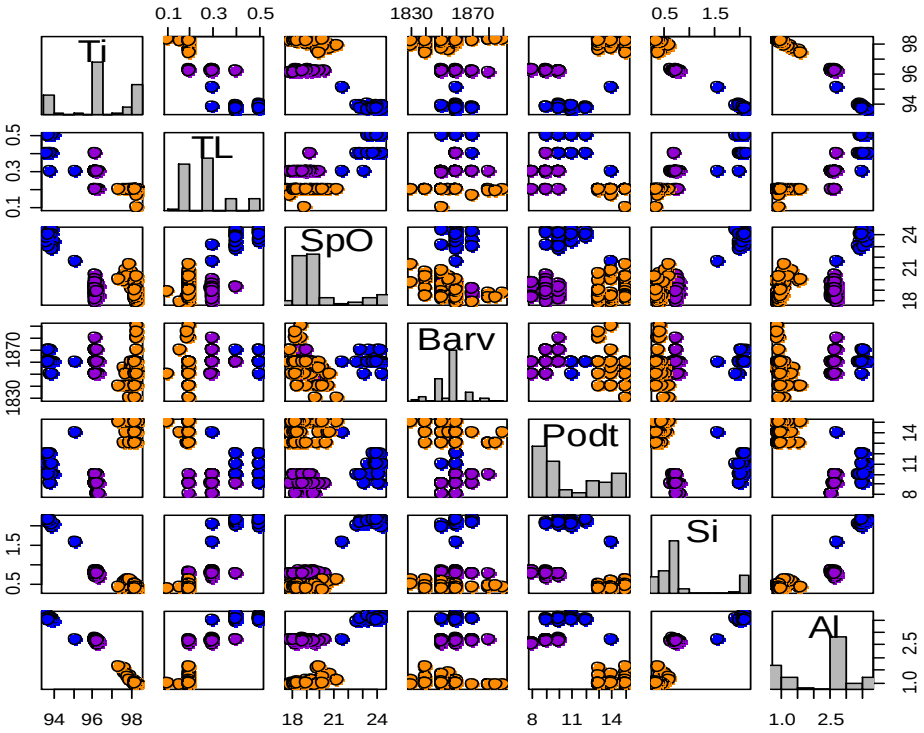
```
 > z<-predict(data2.lda)
> z
$class
  [1] RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA RGA
 [19] RGA RGA RGA RGA RGA RGA RGA RGA RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU
 [37] RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU
 [55] RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU
 [73] RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU RGU
 [91] RGU RGU RGU RGU RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX
[109] RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX
[127] RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX RGX
Levels: RGA RGU RGX
```

Nová data – zařazení do tříd

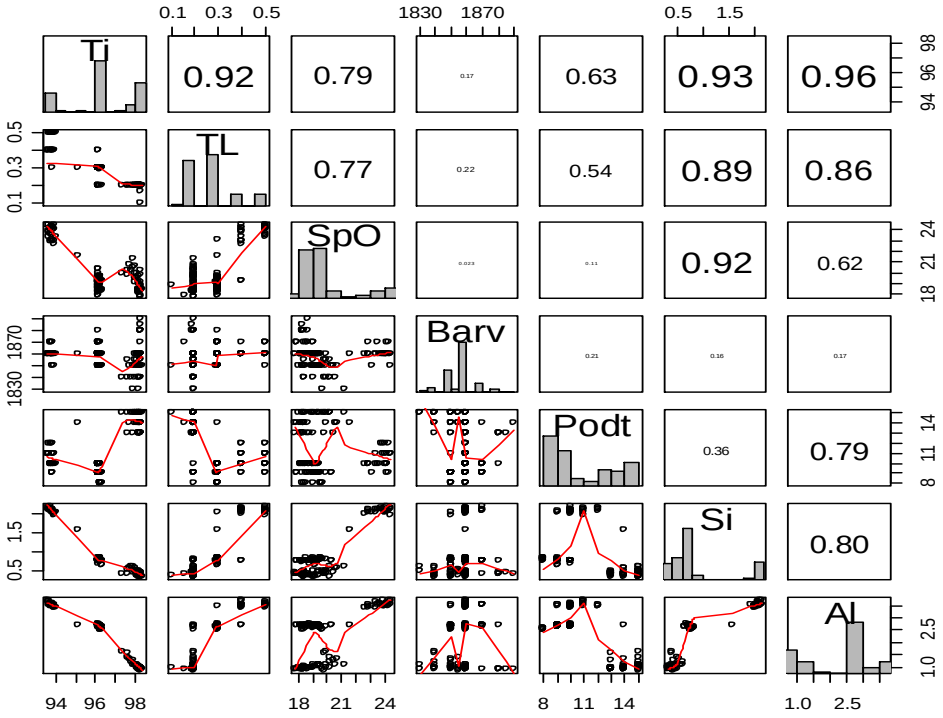
```
upredict
$class
[1] RGA RGU RGA RGX RGU RGU RGA RGU
Levels: RGA RGU RGX

$posterior
             RGA           RGU           RGX
1  1.000000e+00  5.250281e-75 2.909616e-112
2  4.844463e-95  1.000000e+00  2.662608e-56
3  1.000000e+00 1.089714e-146 2.428929e-201
4 2.980455e-177  2.918959e-61  1.000000e+00
5  1.103681e-68  1.000000e+00  5.221942e-58
6  4.844463e-95  1.000000e+00  2.662608e-56
7  1.000000e+00 2.240327e-165 4.557821e-223
8  4.844463e-95  1.000000e+00  2.662608e-56

$x
         LD1       LD2
1 -13.315949  6.413463
2  -1.426880 -4.799013
3 -20.434433  8.658184
4   9.808290  5.625509
5  -3.660142 -3.503976
6  -1.426880 -4.799013
7 -22.167639  9.393127
8  -1.426880 -4.799013
```

1.4.2 Obrázky

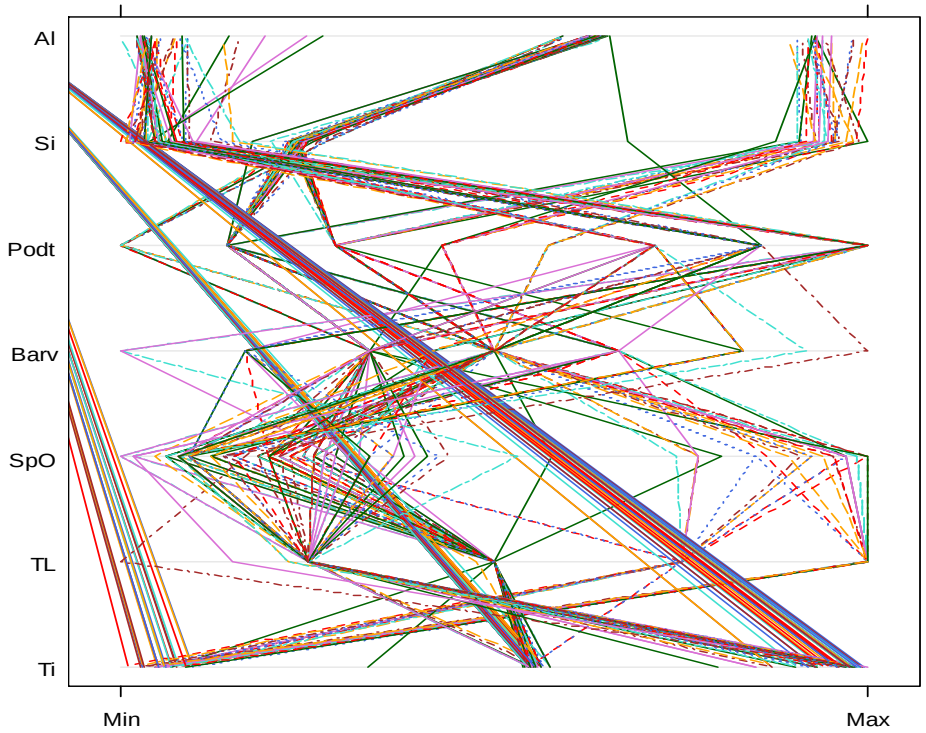
Obrázek č. 1: maticový diagram, histogramy



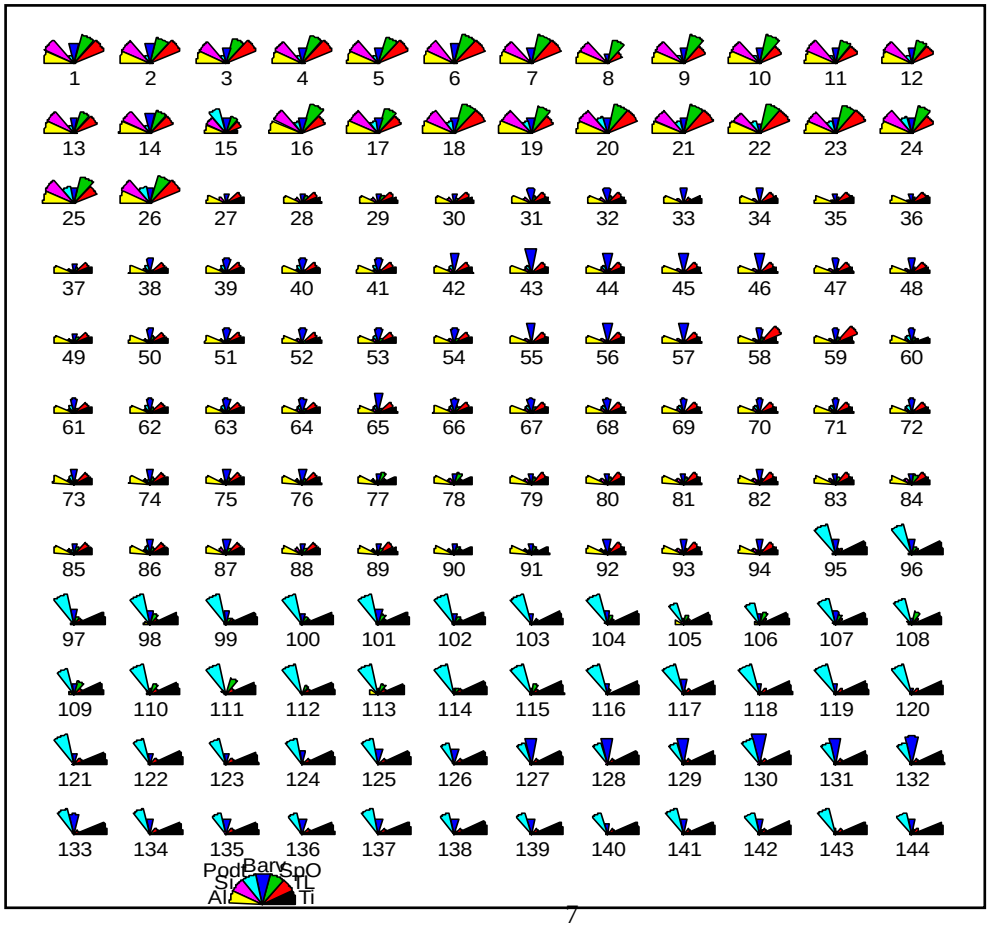
Obrázek č. 2: maticový diagram, korelační koeficienty



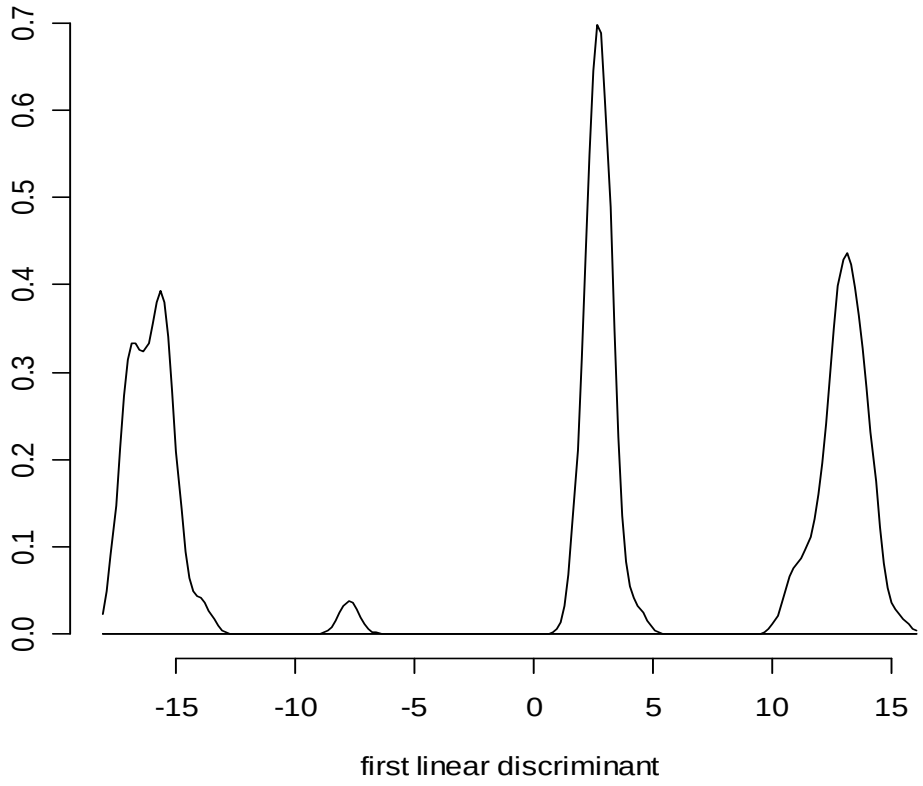
Obrázek č. 3: parallel coordinate graf



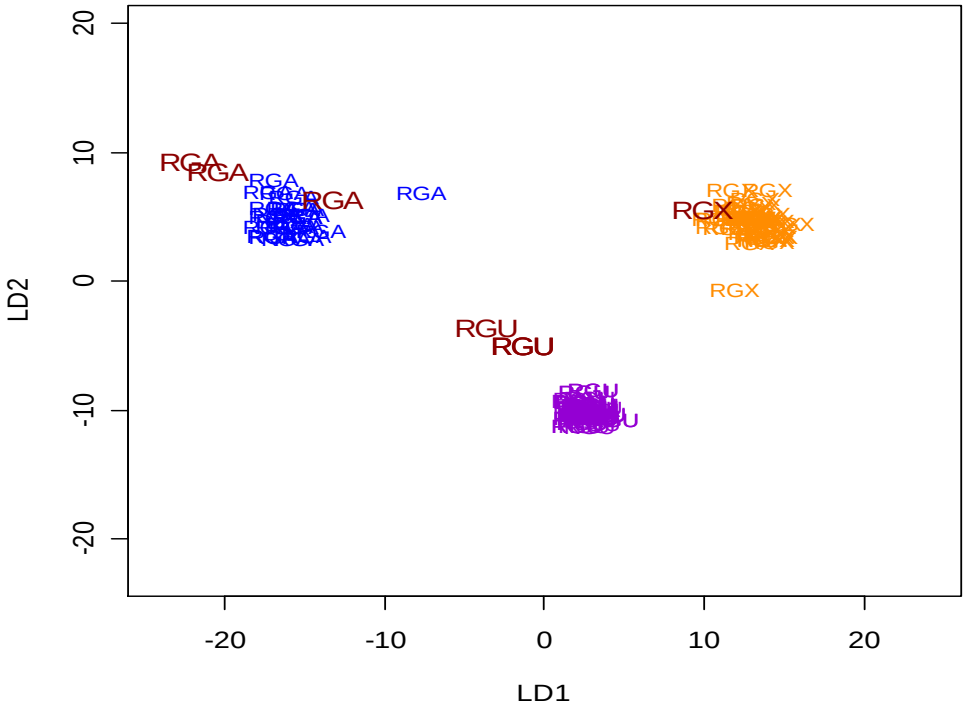
Obrázek č. 4: hvězdicový graf



Obrázek č. 5: Kernelův odhad hustoty



Obrázek č. 6: klasifikační graf



1.5 Závěr

Maticový diagram rozptylových grafů potvrzuje rozdělení pigmentů do tří tříd (tři typy povrchové úpravy). Do diskriminační funkce nejvíce přispívají diskriminátory „obsah Ti“, „obsah Si“ a „obsah Al“. Hvězdicový graf a graf odhadu hustoty rozlišil pigmenty na čtyři třídy (podskupinu ve třídě RGX).

Pro výpočet byla použita lineární diskriminační funkce. Stejného roztřídění bylo dosaženo i s použitím lineární diskriminační funkce s robustním odhadem kovariance a s použitím kvadratické diskriminační funkce. Pigmenty byly roztříděny do tří tříd (RGA, RGU a RGX); všechny objekty byly zařazeny správně.

Na základě vypočtené diskriminační funkce byly neznámé vzorky zařazeny takto:

ID	1	2	3	4	5	6	7	8
třída	RGA	RGU	RGA	RGX	RGU	RGU	RGA	RGU

Úloha č. 2: Logistická regrese

Název úlohy: novorozenci

2.1 Zadání

Cílem úlohy je navrhnout logistický regresní model pro znaky, které ovlivňují nízkou porodní váhu novorozenců a nalézt ty znaky, které jsou v navrženém modelu statisticky významné.

2.2 Data [1]

Data obsahují údaje o 189 novorozencích. V souboru je 8 potenciálních nezávisle proměnných, z toho šest kategorických. Dvě kategorické proměnné (race, ptl) mají více než dvě hladiny. Závisle proměnnou je porodní váha novorozenců. Identifikace proměnných je uvedena v tabulce č. 3. Data **novorozenci** jsou uvedena v tabulce č. 4.

Tabulka č. 3: identifikace proměnných

Popis proměnné	kód / hodnota	název
Porodní váha novorozence	0 = ≥ 2500 1 = <2500	LOW
Věk matky	roky	AGE
Váha matky na počátku těhotenství	libry	LWT
Rasová příslušnost	1 = bílá 2= černá 3 = jiná	RACE
Kouření matky během těhotenství	0 = Ne 1 = Ano	SMOKE
Počet laboratorních vyšetření matky v době těhotenství	0 = žádné 1 = jedno 2 = dvě a více	PTL
Hypertenze	0 = Ne 1 = Ano	HT
Podráždění dělohy	0 = Ne 1 = Ano	UI
Počet lékařských kontrol v 1. trimestru	1 = jedna 2 = dvě a více	FTV

[1] D. W. Hosmer, S. Lemeshow. ***Applied Logistic Regression***. Wiley, New York, 1989.

Tabulka č. 4: data **novorozenci**

low	age	lwt	race	smoke	ptl	ht	ui	ftv	low	age	lwt	race	smoke	ptl	ht	ui	ftv	low	age	lwt	race	smoke	ptl	ht	ui	ftv	low	age	lwt	race	smoke	ptl	ht	ui	ftv
0	19	182	2	0	0	0	1	0	0	22	120	1	0	0	1	0	1	0	24	110	1	0	0	0	0	1	1	20	125	3	0	0	0	1	0
0	33	155	3	0	0	0	0	3	0	23	128	3	0	0	0	0	0	0	19	184	1	1	0	1	0	0	1	25	89	3	0	2	0	0	1
0	20	105	1	1	0	0	0	1	0	22	130	1	1	0	0	0	0	0	24	110	3	0	1	0	0	0	1	19	102	1	0	0	0	0	2
0	21	108	1	1	0	0	1	2	0	30	95	1	1	0	0	0	2	0	23	110	1	0	0	0	0	1	1	19	112	1	1	0	0	1	0
0	18	107	1	1	0	0	1	0	0	19	115	3	0	0	0	0	0	0	20	120	3	0	0	0	0	0	1	26	117	1	1	1	0	0	0
0	21	124	3	0	0	0	0	0	0	16	110	3	0	0	0	0	0	0	25	241	2	0	0	1	0	0	1	24	138	1	0	0	0	0	0
0	22	118	1	0	0	0	0	1	0	21	110	3	1	0	0	1	0	0	30	112	1	0	0	0	0	1	1	17	130	3	1	1	0	1	0
0	17	103	3	0	0	0	0	1	0	30	153	3	0	0	0	0	0	0	22	169	1	0	0	0	0	0	1	20	120	2	1	0	0	0	3
0	29	123	1	1	0	0	0	1	0	20	103	3	0	0	0	0	0	0	18	120	1	1	0	0	0	2	1	22	130	1	1	1	0	1	1
0	26	113	1	1	0	0	0	0	0	17	119	3	0	0	0	0	0	0	16	170	2	0	0	0	0	4	1	27	130	2	0	0	0	1	0
0	19	95	3	0	0	0	0	0	0	17	119	3	0	0	0	0	0	0	32	186	1	0	0	0	0	2	1	20	80	3	1	0	0	1	0
0	19	150	3	0	0	0	0	1	0	23	119	3	0	0	0	0	2	0	18	120	3	0	0	0	0	1	1	17	110	1	1	0	0	0	0
0	22	95	3	0	0	1	0	0	0	24	110	3	0	0	0	0	0	0	29	130	1	1	0	0	0	2	1	25	105	3	0	1	0	0	1
0	30	107	3	0	1	0	1	2	0	28	140	1	0	0	0	0	0	0	33	117	1	0	0	0	1	1	1	20	109	3	0	0	0	0	0
0	18	100	1	1	0	0	0	0	0	26	133	3	1	2	0	0	0	0	20	170	1	1	0	0	0	0	1	18	148	3	0	0	0	0	0
0	18	100	1	1	0	0	0	0	0	20	169	3	0	1	0	1	1	0	28	134	3	0	0	0	0	1	1	18	110	2	1	1	0	0	0
0	15	98	2	0	0	0	0	0	0	24	115	3	0	0	0	0	2	0	14	135	1	0	0	0	0	0	1	20	121	1	1	1	0	1	0
0	25	118	1	1	0	0	0	3	0	28	250	3	1	0	0	0	6	0	28	130	3	0	0	0	0	0	1	21	100	3	0	1	0	0	4
0	20	120	3	0	0	0	1	0	0	20	141	1	0	2	0	1	1	0	25	120	1	0	0	0	0	2	1	26	96	3	0	0	0	0	0
0	28	120	1	1	0	0	0	1	0	22	158	2	0	1	0	0	2	0	16	95	3	0	0	0	0	1	1	31	102	1	1	1	0	0	1
0	32	121	3	0	0	0	0	2	0	22	112	1	1	2	0	0	0	0	20	158	1	0	0	0	0	1	1	15	110	1	0	0	0	0	0
0	31	100	1	0	0	0	1	3	0	31	150	3	1	0	0	0	2	0	26	160	3	0	0	0	0	0	1	23	187	2	1	0	0	0	1
0	36	202	1	0	0	0	0	1	0	23	115	3	1	0	0	0	1	0	21	115	1	0	0	0	0	1	1	20	122	2	1	0	0	0	0
0	28	120	3	0	0	0	0	0	0	16	112	2	0	0	0	0	0	0	22	129	1	0	0	0	0	0	1	24	105	2	1	0	0	0	0
0	25	120	3	0	0	0	1	2	0	16	135	1	1	0	0	0	0	0	25	130	1	0	0	0	0	2	1	15	115	3	0	0	0	1	0
0	28	167	1	0	0	0	0	0	0	18	229	2	0	0	0	0	0	0	31	120	1	0	0	0	0	2	1	23	120	3	0	0	0	0	0
0	17	122	1	1	0	0	0	0	0	25	140	1	0	0	0	0	1	0	35	170	1	0	1	0	0	1	1	30	142	1	1	1	0	0	0
0	29	150	1	0	0	0	0	2	0	32	134	1	1	1	0	0	4	0	19	120	1	1	0	0	0	0	1	22	130	1	1	0	0	0	1
0	26	168	2	1	0	0	0	0	0	20	121	2	1	0	0	0	0	0	24	116	1	0	0	0	0	1	1	17	120	1	1	0	0	0	3
0	17	113	2	0	0	0	0	1	0	23	190	1	0	0	0	0	0	0	45	123	1	0	0	0	0	1	1	23	110	1	1	1	0	0	0
0	17	113	2	0	0	0	0	1	0	22	131	1	0	0	0	0	1	1	28	120	3	1	1	0	1	0	1	17	120	2	0	0	0	0	2
0	24	90	1	1	1	0	0	1	0	32	170	1	0	0	0	0	0	1	29	130	1	0	0	0	1	2	1	26	154	3	0	1	1	0	1
0	35	121	2	1	1	0	0	1	0	30	110	3	0	0	0	0	0	1	34	187	2	1	0	1	0	0	1	20	105	3	0	0	0	0	3
0	25	155	1	0	0	0	0	1	0	20	127	3	0	0	0	0	0	1	25	105	3	0	1	1	0	0	1	26	190	1	1	0	0	0	0
0	25	125	2	0	0	0	0	0	0	23	123	3	0	0	0	0	0	1	25	85	3	0	0	0	1	0	1	14	101	3	1	1	0	0	0
0	29	140	1	1	0	0	0	2	0	17	120	3	1	0	0	0	0	1	27	150	3	0	0	0	0	0	1	28	95	1	1	0	0	0	2
0	19	138	1	1	0	0	0	2	0	19	105	3	0	0	0	0	0	1	23	97	3	0	0	0	1	1	1	14	100	3	0	0	0	0	2
0	27	124	1	1	0	0	0	0	0	23	130	1	0	0	0	0	0	1	24	128	2	0	1	0	0	1	1	23	94	3	1	0	0	0	0
0	31	215	1	1	0	0	0	2	0	36	175	1	0	0	0	0	0	1	24	132	3	0	0	1	0	0	1	17	142	2	0	0	1	0	0
0	33	109	1	1	0	0	0	1	0	22	125	1	0	0	0	0	1	1	21	165	1	1	0	1	0	1	1	21	130	1	1	0	1	0	3
0	21	185	2	1	0	0	0	2	0	24	133	1	0	0	0	0	0	1	32	105	1	1	0	0	0	0									
0	19	189	1	0	0	0	0	2	0	21	134	3	0	0	0	0	2	1	19	91	1	1	2	0	1	0									
0	23	130	2	0	0	0	0	1	0	19	235	1	1	0	1	0	0	1	25	115	3	0	0	0	0	0									
0	21	160	1	0	0	0	0	0	0	25	95	1	1	3	0	1	0	1	16	130	3	0	0	0	0	1									
0	18	90	1	1	0	0	1	0	0	16	135	1	1	0	0	0	0	1	25	92	1	1	0	0	0	0									
0	18	90	1	1	0	0	1	0	0	29	135	1	0	0	0	0	1	1	20	150	1	1	0	0	0	2									
0	32	132	1	0	0	0	0	4	0	29	154	1	0	0	0	0	1	1	21	200	2	0	0	0	1	2									
0	19	132	3	0	0	0	0	0	0	19	147	1	1	0	0	0	0	1	24	155	1	1	1	0	0	0									
0	24	115	1	0	0	0	0	2	0	19	147	1	1	0	0	0	0	1	21	103	3	0	0	0	0	0									
0	22	85	3	1	0	0	0	0	0	30	137	1	0	0	0	0	1																		

2.3 Program:

Úloha byla řešena programem: **R Version 2.1.0**

2.4 Řešení

2.4.1 Protokol LR

Call:

bic.glm.formula(f = low ~ age + lwt + race + smoke + ptl + ht + ui + ftv, data = birthwt, glm.family = "binomial")

49 models were selected

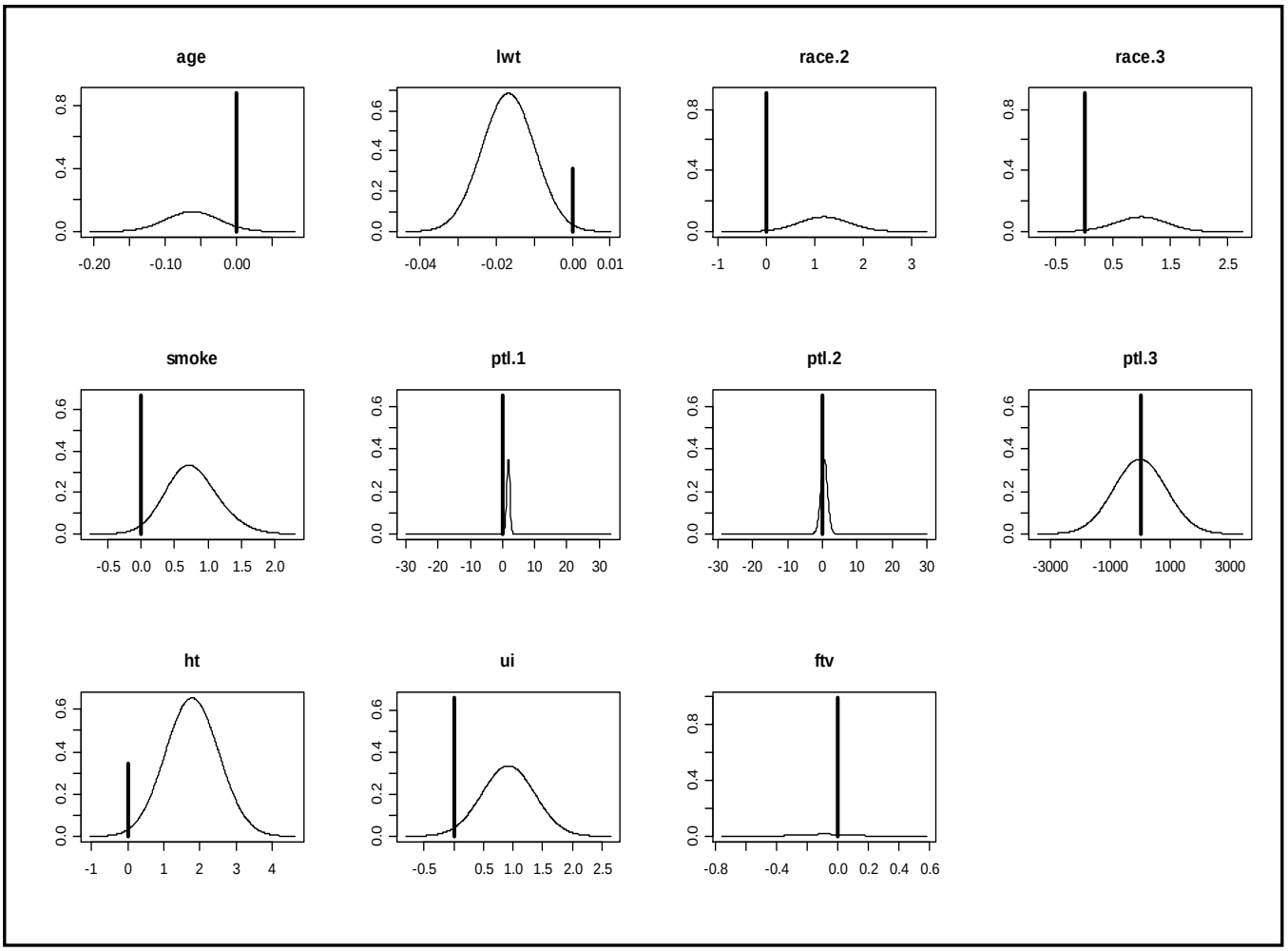
Best 5 models (cumulative posterior probability = 0.37):

	p!=0	EV	SD	cond EV	cond SD	model 1	model 2
Intercept	100	0.4716	1.3e+00	0.472	1.309	1.451	1.068
age	12.6	-0.0078	2.4e-02	-0.062	0.037	.	.
lwt	68.7	-0.0116	9.7e-03	-0.017	0.007	-0.019	-0.017
race	9.6						
.2		0.1153	3.9e-01	1.201	0.547	.	.
.3		0.0927	3.2e-01	0.966	0.461	.	.
smoke	33.2	0.2554	4.3e-01	0.768	0.397	.	.
ptl	35.1						
.1		0.6174	8.9e-01	1.758	8.211	.	.
.2		0.1686	6.1e-01	0.480	7.592	.	.
.3		-4.9110	5.2e+02	-13.988	882.840	.	.
ht	65.4	1.1669	1.0e+00	1.785	0.729	1.856	1.962
ui	33.8	0.3105	5.1e-01	0.918	0.445	.	0.930
ftv	1.5	-0.0013	2.4e-02	-0.087	0.173	.	.
nVar						2	3
BIC						-753.823	-753.110
post prob						0.111	0.078

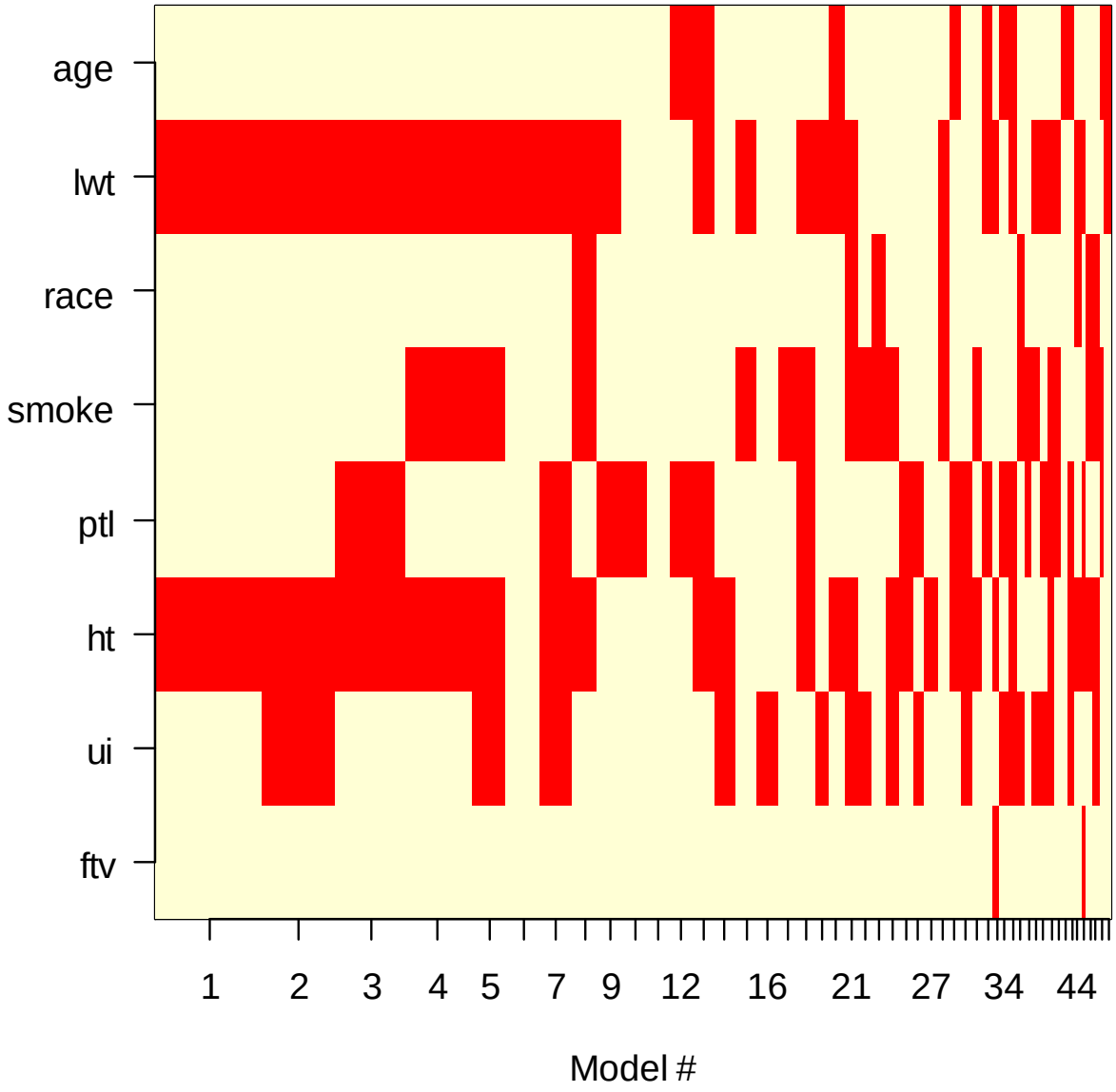
	model 3	model 4	model 5
Intercept	1.207	1.084	0.722
age	.	.	.
lwt	-0.019	-0.018	-0.016
race			
.2	.	.	.
.3	.	.	.
smoke	.	0.684	0.653
ptl			
.1	1.743	.	.
.2	0.501	.	.
.3	-13.986	.	.
ht	1.924	1.822	1.922
ui	.	.	0.896
ftv	.	.	.
nVar	3	3	4
BIC	-752.998	-752.865	-751.656
post prob	0.073	0.069	0.037

2.4.2 Obrázky

Obrázek č. 7: aposteriorní distribuce parametrů logistické regrese



Obrázek č. 8: grafická sumarizace navržených modelů (statisticky významný parametr je v daném modelu vyznačen červeně)



2.5 Závěr

Bylo nalezeno 49 logistických regresních modelů. Na základě hodnot posteriorní pravděpodobnosti a hodnot Bayesova informačního kritéria (BIS) bylo nalezeno 5 nejlepších modelů, ve kterých jsou uvedeny pouze statisticky významné parametry.

model 1	***LOW = 1,451 – 0,019 LWT + 1,856 HT***
model 2	***LOW = 1,068 – 0,017 LWT + 1,962 HT + 0,93 UI***
model 3	***LOW = 1,207 – 0,019 LWT + PTL.1+ 0,501 PTL.2 - 13,986 PTL.3+1,924 HT***
model 4	***LOW = 1,084 – 0,018 LWT + 0,684 SMOKE + 1,822 HT***
model 5	***LOW = 0,722 – 0,016 LWT + 0,653 SMOKE + 1,922 HT + 0,896 UI***

Ve shodě s grafem aposteriorní distribuce parametrů logistické regrese (obr. č. 1) jsou statisticky nejvýznamnější parametry LWT (váha matky na počátku těhotenství), SMOKE (kouření), a HT (hypertenze); tyto parametry jsou statisticky nejvýznamnější pro odhad, zda hrozí nebezpečí nízké porodní váhy novorozence.

Úloha č. 3 A: Vícerozměrné škálování

Název úlohy:emise

3.A.1. Zadání

Vícerozměrným škálováním a analýzou klastrů posuďte podobnost evropských zemí z hlediska množství emisí produkovaných při výrobě energie.

3.A.2. Data

Data **emise** jsou v tabulce č. 5. Jsou převzata z databáze Evropské komise [2]. Obsahují emise v jednotlivých evropských zemích za rok 2003; v jednotkách 1000 tun. Emise jsou vyjádřeny v těchto kategoriích:

sox	oxidy síry
nox	oxidy dusíku
pm10	pevné částice pod 10 μm
co2	oxid uhličitý
ap	kyselé emise
top	emise prekurzorů troposférického ozónu

Pro názvy zemí jsou použity zkratky:

be	Belgie	cz	Česká republika	dk	Dánsko	de	Německo
ee	Estonsko	gr	Řecko	es	Španělsko	fr	Francie
ie	Irsko	it	Itálie	cy	Kypr	lv	Litva
hu	Maďarsko	nl	Holandsko	at	Rakousko	pl	Polsko
pt	Portugalsko	sl	Slovinsko	sk	Slovensko	fi	Finsko
se	Švédsko	uk	Velká Británie	hr	Chorvatsko	no	Norsko

3.A.3. Program:

Úloha byla řešena programem: **R Version 2.1.0**

[2] http://epp.eurostat.cec.eu.int

Tabulka č. 5: data **emise**

země/emise	sox	nox	pm10	co2	ap	top
be	56562,1	41641	4064,3	29141192	2674	54945,5
cz	144396,4	104724,4	4999,3	58924308	6790,8	136188,9
dk	17461	64507,7	1141,1	31401903	1948	84571,7
de	339921,8	267707	21471	362581556	16608,8	348633,8
ee	86290	15680	11433	15854745	3038	23259,3
gr	383330	89340	15828	56100103	13921	120134,1
es	985198,1	331984	26312	105332266	38010,3	417908,6
fr	205136,7	133192	13127	63802190	9306	170889,2
ie	45040	34713	1332,1	15480300	2162,2	43050,1
it	373763	139310	16560	160882830	1471,2	181393,4
cy	28920	5620	440	3214423	1025,9	9395,3
lv	3827,4	7611	1114,1	2416412,7	285,1	11724,8
hu	227080	33080	4605	20501452	7821,9	42844,7
nl	37580,7	60570,5	505,9	67347391	2491,3	79651,9
at	8430,2	16958,3	1541,4	16030352	650	21923,5
pl	823500	264528,6	301693,7	183069328	31543,8	340803,9
pt	186663	68096,9	3515,6	20009037	7313,6	84818,7
sl	51100	16520	866,9	6159862,3	1956	21927,5
sk	60121,7	22692,2	5206,4	13373548	2377,1	29947,1
fi	54034	60782	2704	36047271	3010,7	76603,3
se	15684,7	16854,3	5074,7	12768834	891,8	26575
uk	745447,9	468988,7	12786	212728525	33492,1	591358
hr	23289,8	13192,7	970	7873853,1	1015,3	16501,6
no	1387,5	45799,5	711,3	12713884	1039	59310,2

3.A.4 Řešení – MDS, CLU

3.A.4.1 Protokoly

Pro řešení bylo použité klasické metrické vícerozměrné škálování (CMDS), Sammonovo nelineární mapování (SMDS) a nemetrické vícerozměrné škálování (NNMDS). Všechny techniky vycházely z matice vzdáleností. Protokoly uvádějí tabelární podobu 2D mapy objektů a hodnoty koeficientu stress pro jednotlivé techniky.K posouzení úspěšnosti metody byl použit koeficient těsnosti proložení stress.

CMDS:

```
> euro.clas
$points
             [,1]          [,2]
 [1,] -1.0711080 -0.013694674
 [2,] -0.1563282  0.259364662
 [3,] -0.9904244  0.147024089
 [4,]  3.1768471  1.416530203
 [5,] -1.2272454 -0.275909897
 [6,]  0.4336987 -0.089254750
 [7,]  4.4003272  0.512726871
 [8,]  0.3064869  0.229750258
 [9,] -1.2422486 -0.037668945
[10,]  0.7499024  0.458152802
[11,] -1.5795853 -0.159355459
[12,] -1.6347744 -0.151024417
[13,] -0.7004640 -0.170816165
[14,] -0.8133822  0.232278914
[15,] -1.4865747 -0.087422850
[16,]  4.9391595 -3.526220036
[17,] -0.5347042 -0.001708189
[18,] -1.4146354 -0.128369995
[19,] -1.2863992 -0.152764593
[20,] -0.9012032  0.098123405
[21,] -1.4508277 -0.145916374
[22,]  5.2806672  1.678206497
[23,] -1.5167458 -0.125424735
[24,] -1.2804382  0.033393375

$eig
[1] 107.11108  18.18389

classical.stress
0.07891955
```

SMDS:

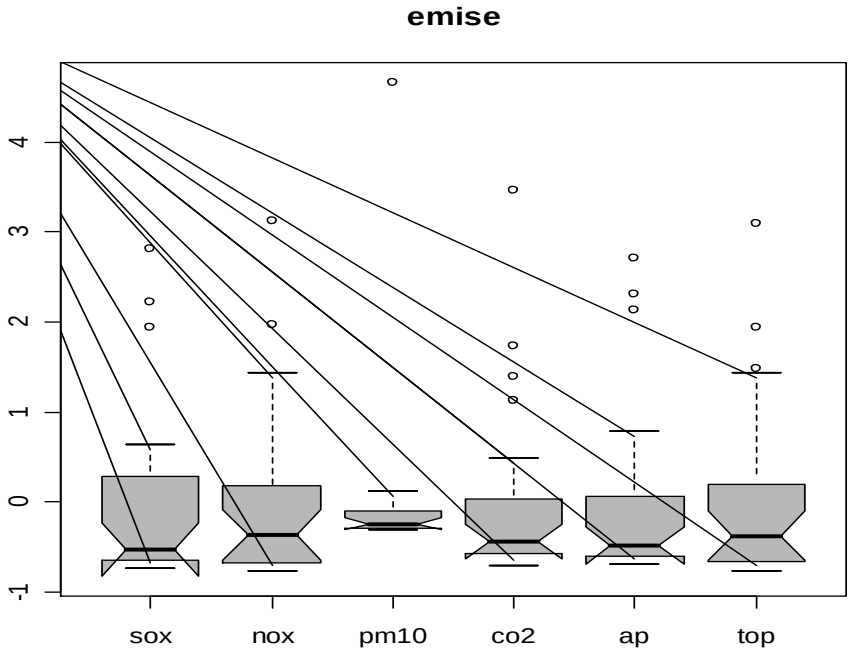
```
> euro.sam
$points
             [,1]         [,2]
 [1,] -1.0844644  0.03643953
 [2,] -0.1650030  0.27583446
 [3,] -1.0428791  0.32924940
 [4,]  3.0307644  2.86680997
 [5,] -1.2289408 -0.30956399
 [6,]  0.5036039 -0.55791622
 [7,]  4.7191528 -0.51912262
 [8,]  0.3065339  0.21114133
 [9,] -1.2573271 -0.01566610
[10,]  0.7838790  1.28974043
[11,] -1.5844805 -0.21582690
[12,] -1.6560315 -0.12873596
[13,] -0.6467138 -0.59789403
[14,] -0.8385710  0.51756263
[15,] -1.5258637 -0.01115898
[16,]  4.6199104 -3.99606326
[17,] -0.4840235 -0.25237159
[18,] -1.4170506 -0.20443053
[19,] -1.2960040 -0.15961267
[20,] -0.9237540  0.19053572
[21,] -1.4746340 -0.05133824
[22,]  5.5236879  1.22652294
[23,] -1.5314091 -0.13648202
[24,] -1.3303823  0.21234671

$stress
[1] 0.002364262
```

NNMDS:

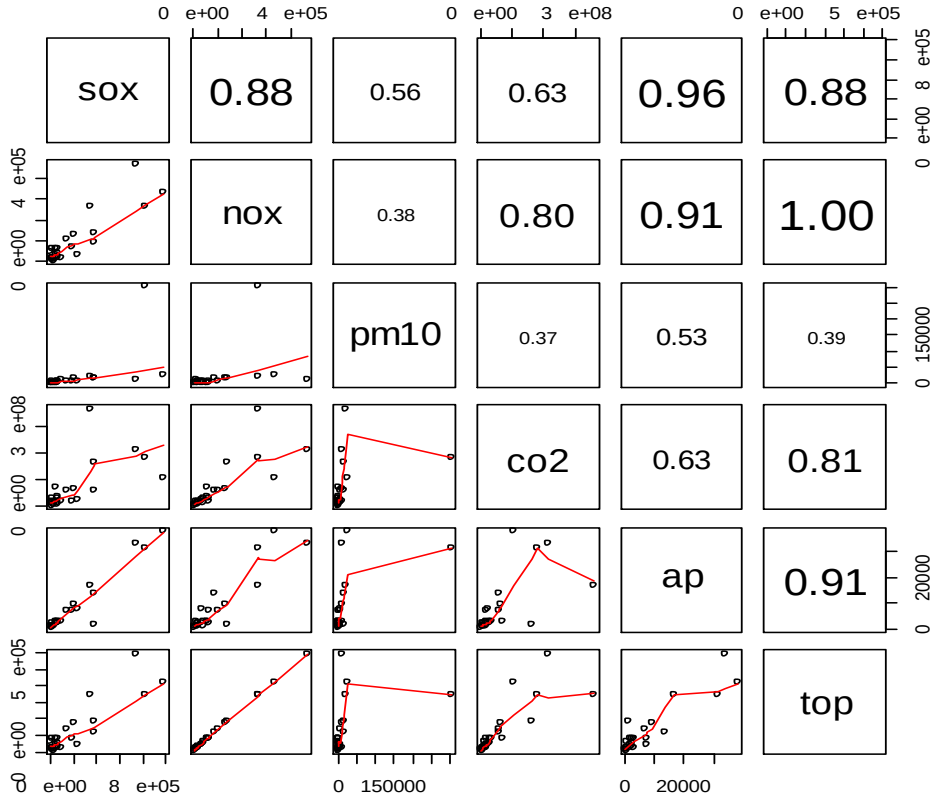
```
> euro.iso
$points
             [,1]          [,2]
 [1,] -1.0667998 -0.003795473
 [2,] -0.1818565  0.264760733
 [3,] -1.0400766  0.247935984
 [4,]  2.9473780  2.369738078
 [5,] -1.2753000 -0.313544726
 [6,]  0.5652853 -0.441207861
 [7,]  4.5327116 -0.236660841
 [8,]  0.3175326  0.168128186
 [9,] -1.2708984 -0.057240115
[10,]  0.8181163  1.136518242
[11,] -1.5433669 -0.231716424
[12,] -1.6227513 -0.204074543
[13,] -0.6326118 -0.553965271
[14,] -0.8688780  0.460605472
[15,] -1.4978403 -0.086011132
[16,]  4.9525808 -3.537915471
[17,] -0.4684727 -0.197483045
[18,] -1.4197390 -0.217370680
[19,] -1.3166115 -0.195196100
[20,] -0.9624780  0.157677370
[21,] -1.4621673 -0.109957287
[22,]  5.3177340  1.606510255
[23,] -1.4988360 -0.158877184
[24,] -1.3226544  0.133141830

$stress
[1] 2.772679
```

3.A.4.2 Obrázky

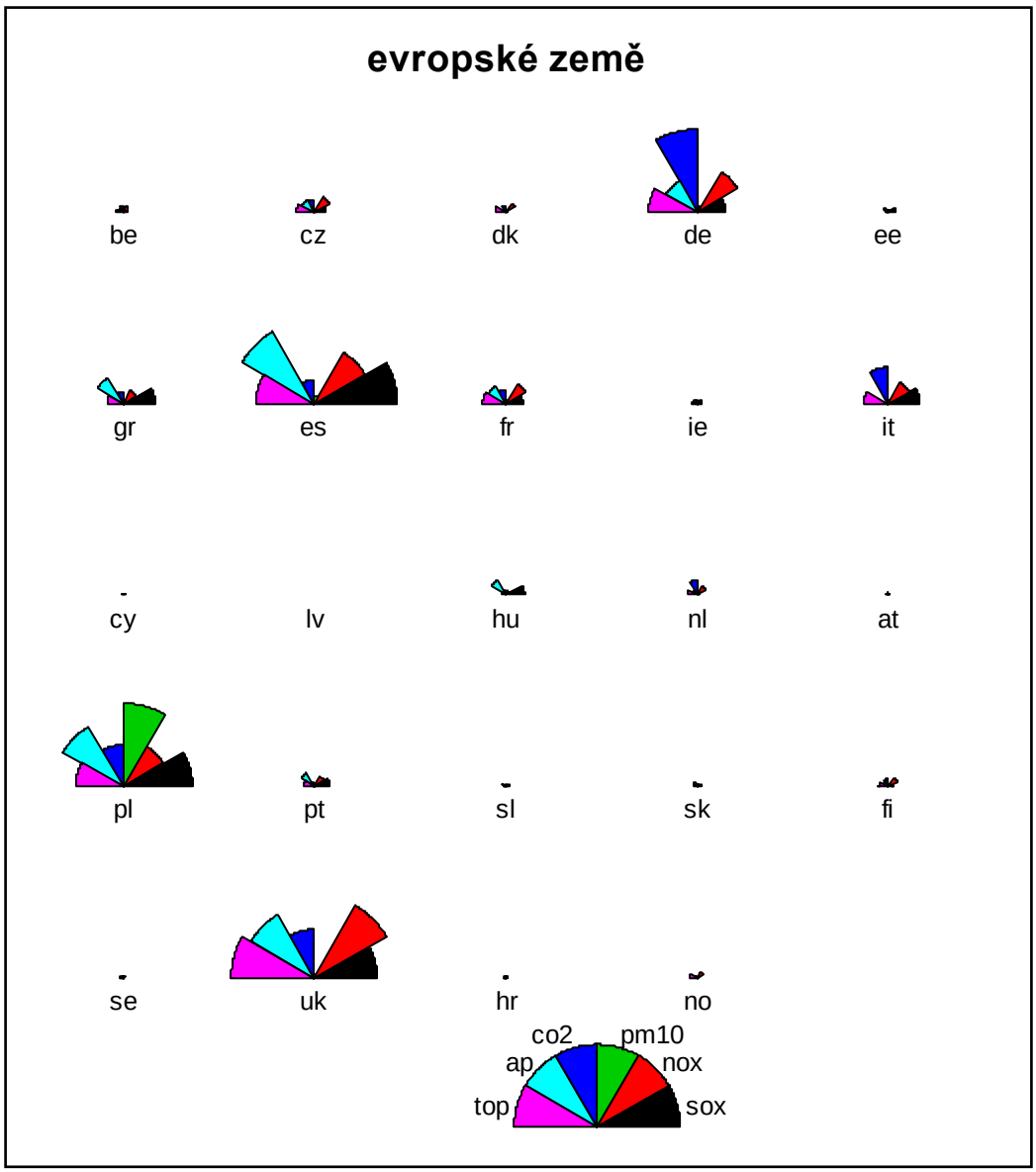
Obrázek č. 9: krabicové grafy znaků



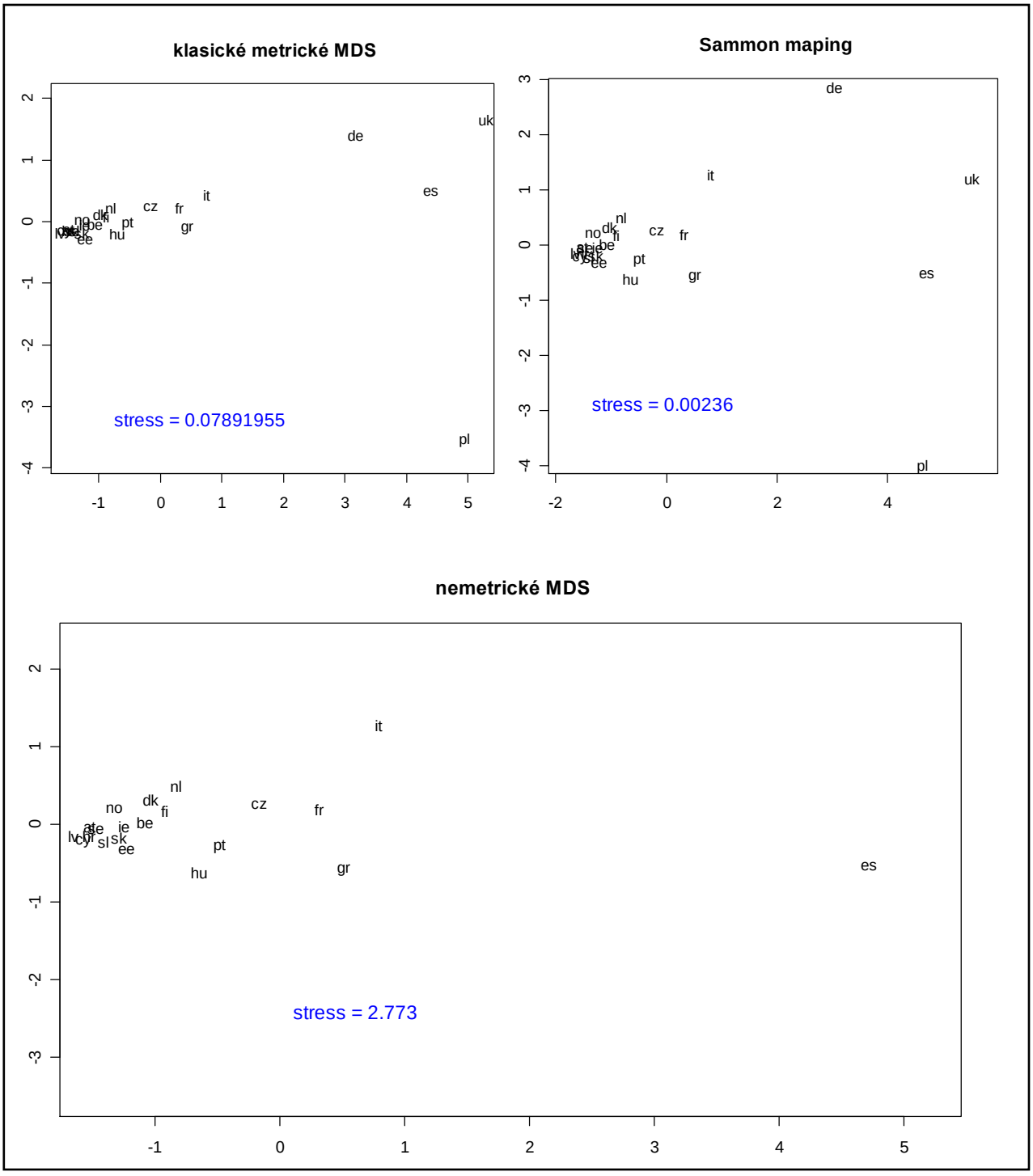
Obrázek č. 10: Maticový diagram, korelační matice



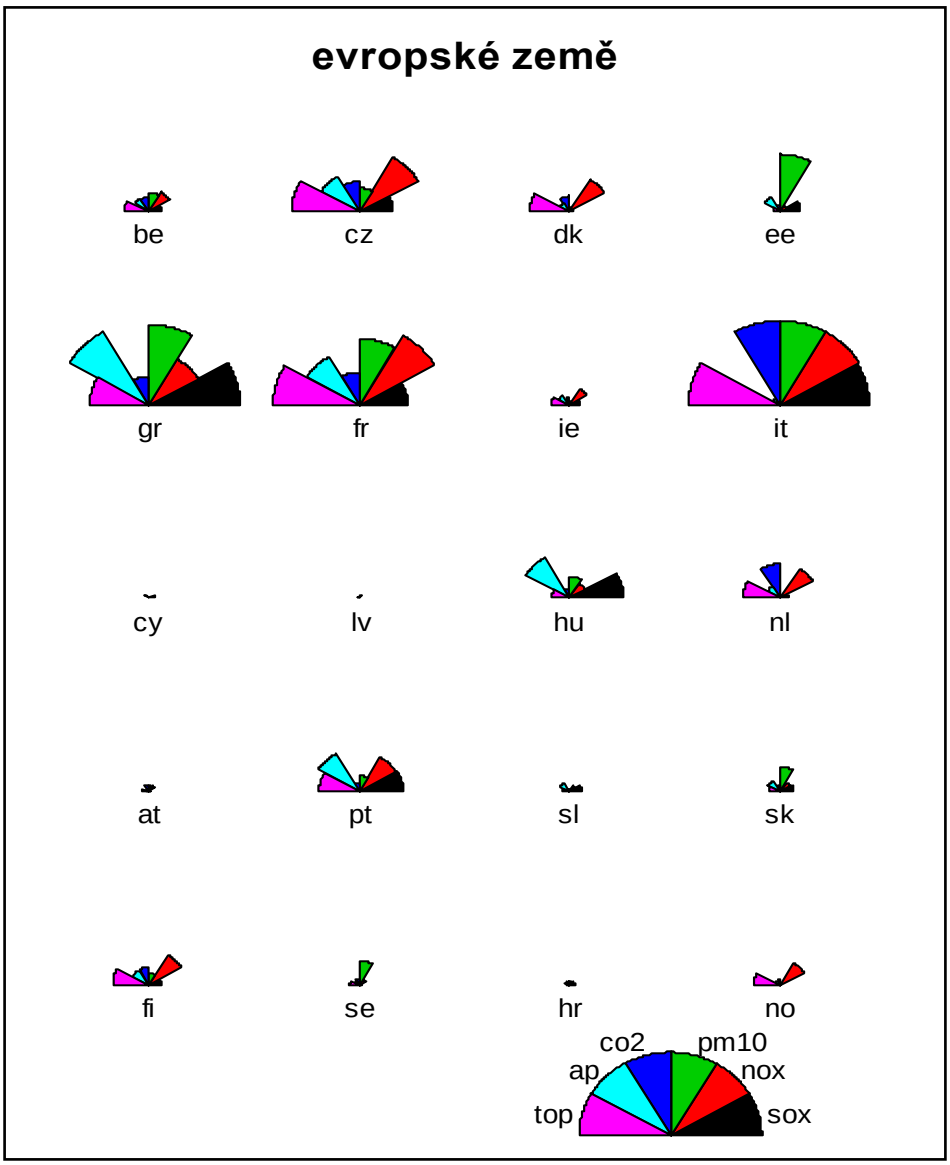
Obrázek č. 11: hvězdicový graf



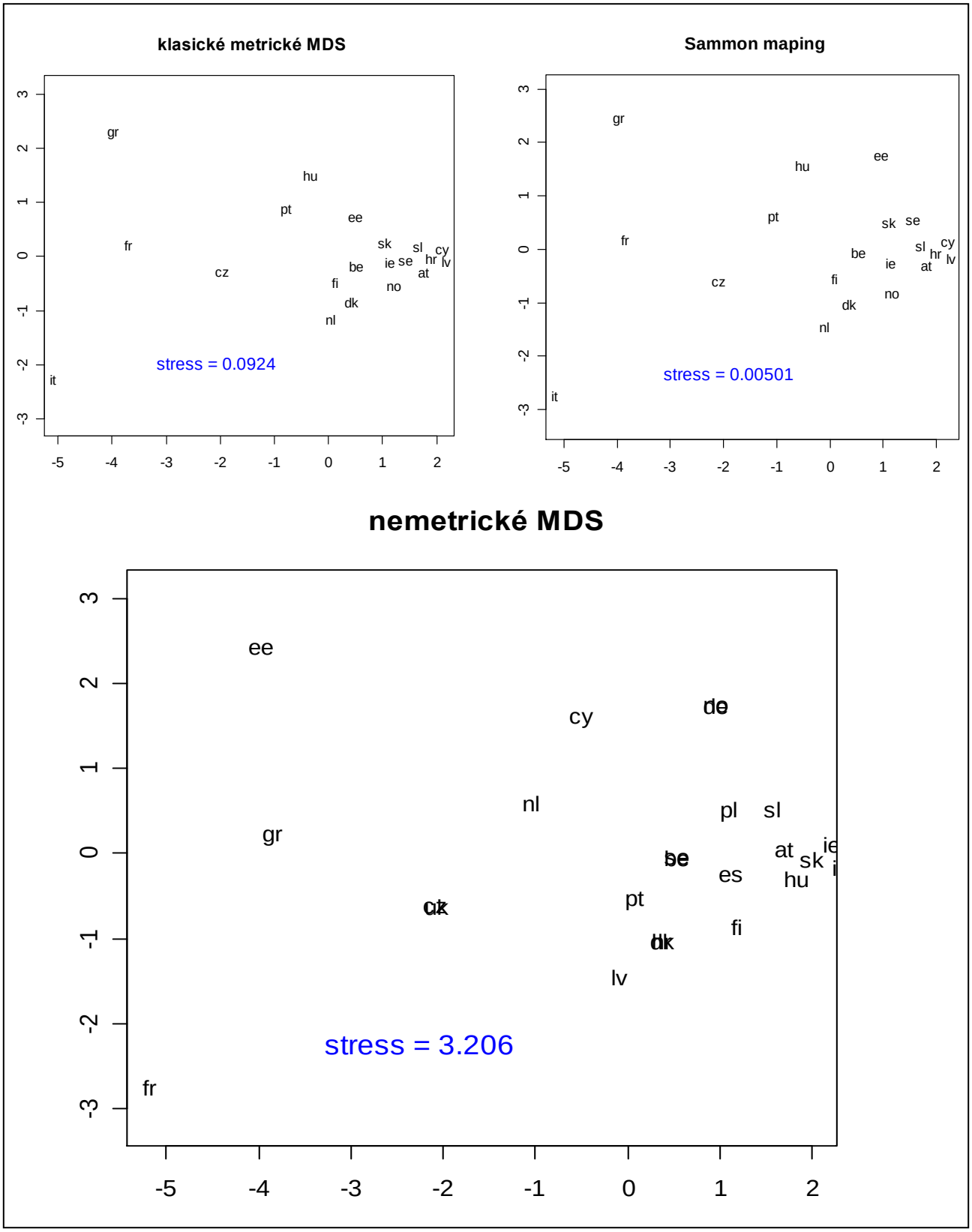
Obrázek č. 12: Mapy objektů pro původní data



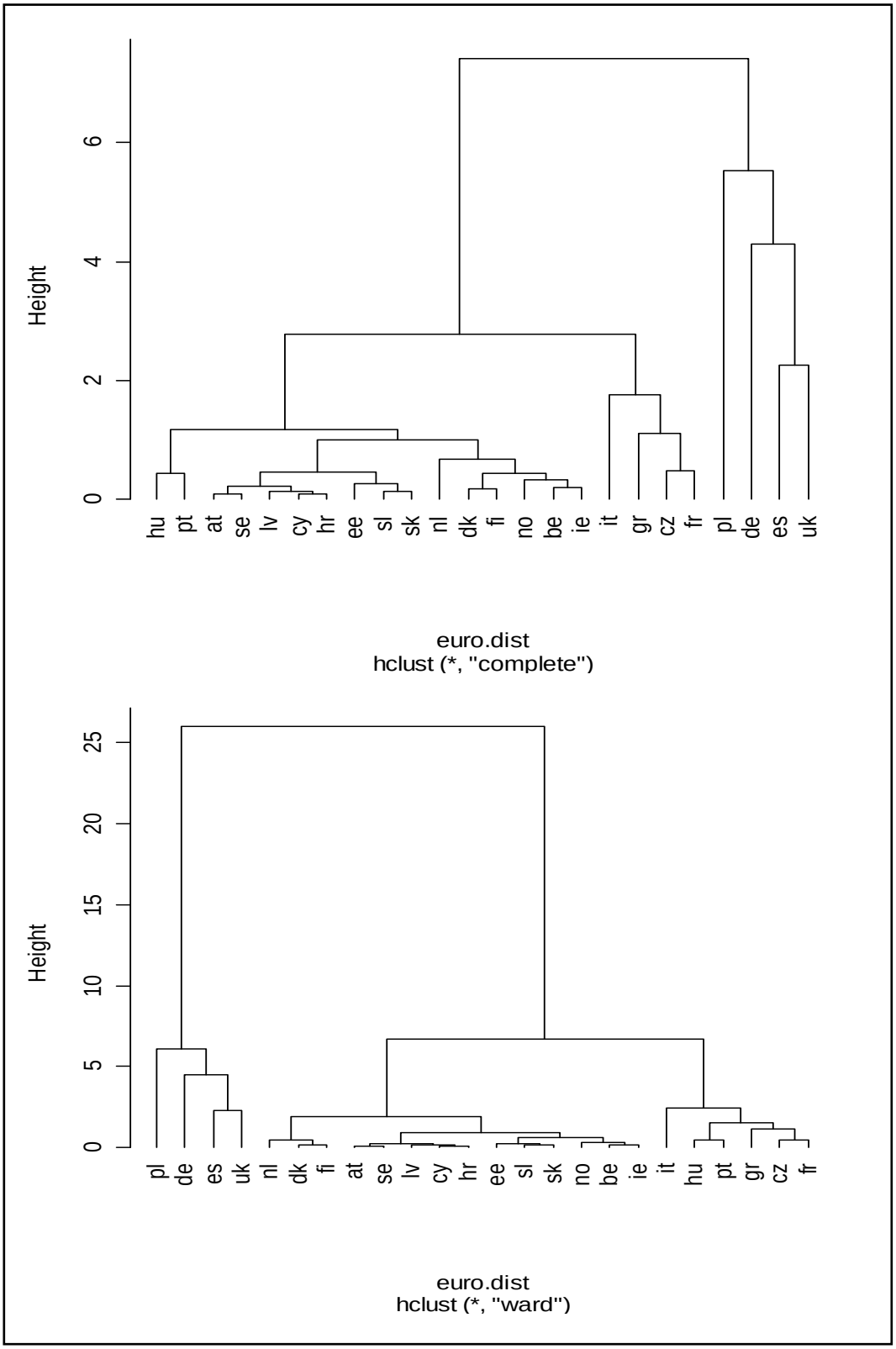
Obrázek č. 13: Hvězdicový graf po odstranění objektů de, es, pl, uk



Obrázek č. 14: Mapy objektů po odstranění objektů de, es, pl, uk



Obrázek č. 15: Dendrogramy



3.A.5 Závěr

Metodou MDS i CLU byly evropské země rozděleny do tří skupin:

1. Dánsko, Španělsko, Polsko, Velká Británie
2. Itálie, Maďarsko, Řecko, Česká republika, Francie
3. ostatní země

Nejlepší výsledek (nejnižší hodnotu koeficientu stress poskytla metoda SMDS.

Je ale nutné konstatovat, že podobnost evropských zemí na základě sledovaných parametrů (emise způsobené výrobou energie) je nízká a rozlišení do skupin je diskutabilní (v první skupině jde spíše o odlehlé objekty). Souřadnice map objektů se nepodařilo identifikovat. Zajímavé výsledky by mohlo přinést srovnání s analýzou zdrojů energie, používaných v jednotlivých zemích.

Úloha č. 3 B: Korespondenční analýza kategorických dat

Název úlohy:slova

3.B.1. Zadání

Ověřte apriorní rozdělení uživatelů anglického jazyka (anglický jazyk pro ně není jazyk mateřský) do dvou skupin na základě frekvence nejčastěji používaných anglických slov. Pět skupin uživatelů angličtiny jako cizího jazyka je na základě geografické polohy rozděleno do dvou oblastí takto:

1. Arabský poloostrov - ***RAK*** (emirát Ras al Khaimah)
2. Sino-pacifická skupina – ***THAI*** (Thajsko)
   TAIWAN
   KOREA (Severní Korea)
   JAPAN

3.B.2. Data [3]

Data **slova** jsou uvedena v tabulce č. 6. Tabulka uvádí četnost používání 19-ti nejfrekventovanějších anglických slov v jednotlivých studovaných skupinách uživatelů anglického jazyka. Četnost používaných slov byla zjišťována v textu o celkové průměrné délce 10 324 slov.

[3] P. J. Hassall, S. Ganesh: *Correspondence analysis of international relative deviance* http:// fccl.ksu.ru/winter.99/lang_typ/hassal/reldev.pdf

Tabulka č. 6: data **slova**

Word	*THAI*	*TAIWAN*	*RAK*	*KOREA*	*JAPAN*
IN	279	468	373	569	392
A	279	389	237	559	493
THE	228	528	359	461	260
CITY	184	340	206	573	433
AND	150	254	273	451	309
TO	182	305	159	336	293
LARGE	1	259	163	449	342
OF	183	182	121	389	275
CAN	94	210	181	346	233
MANY	111	140	135	408	194
IS	101	187	105	310	260
ARE	175	120	119	282	230
I	62	110	42	276	377
WE	7	141	34	375	223
THERE	117	90	103	198	245
YOU	124	216	213	66	65
PEOPLE	90	118	98	209	127
IT	42	127	84	147	118
LIVING	71	140	69	151	75

3.B.3. Program:

Úloha byla řešena programem: **R Version 2.1.0**

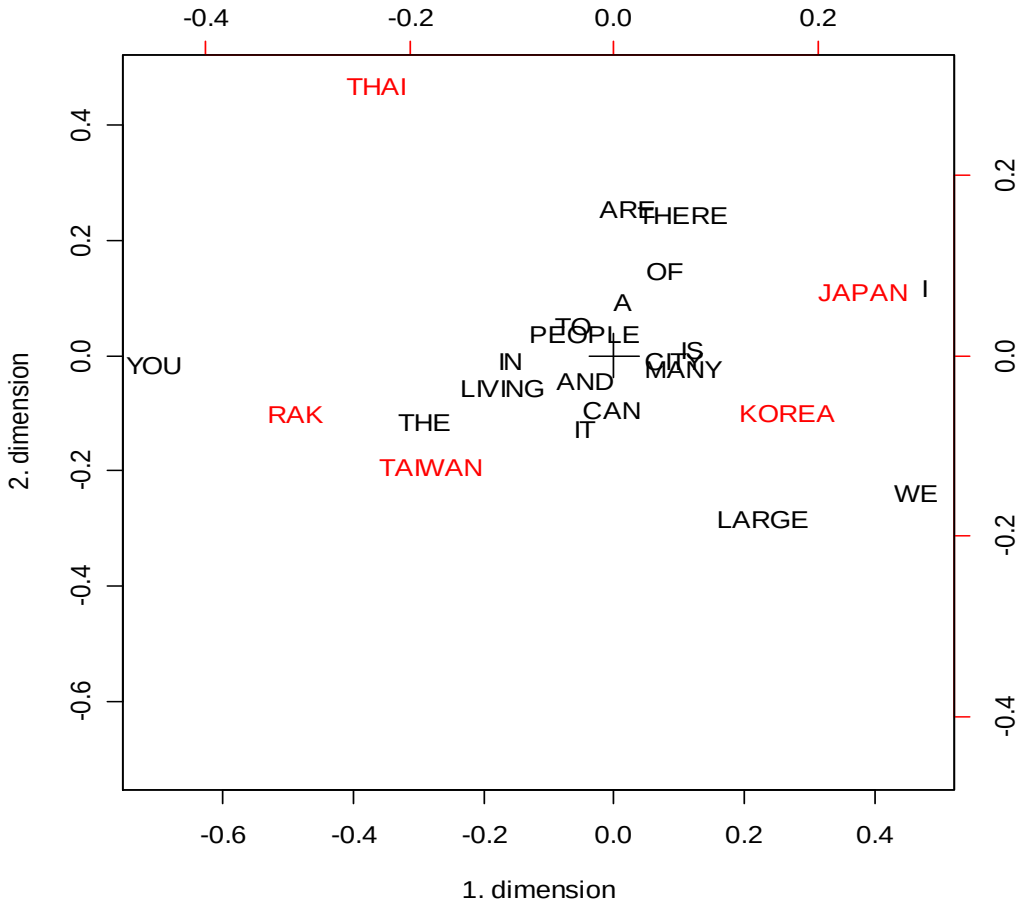
3.B.4. Řešení

3.B.4.1 Protokol CA

```
Importance of dimensions:
                          Dimen.1    Dimen.2      Dimen.3
Standard deviation     217.8279232 102.9599019 54.65635579
Proportion of Variance   0.7450746   0.1664596  0.04690876
Cumulative Proportion    0.7450746   0.9115342  0.95844297

First canonical correlation(s): 0.22209658 0.12828752 0.08492008

 Row scores:
              [,1]        [,2]        [,3]
IN      -0.71770981 -0.03719102  0.10028950
A        0.06089127  0.76594450 -0.29462794
THE     -1.30938508 -0.86865503 -0.08355481
CITY     0.39985109 -0.04734538  0.07194198
AND     -0.20034584 -0.29587700  0.31914457
TO      -0.29469713  0.43939842 -0.57054702
LARGE    1.03022344 -2.16880482 -0.98207132
OF       0.34779167  1.18194480  1.13828447
CAN     -0.01519128 -0.69211000  0.17301311
MANY     0.47499211 -0.13250751  2.48071769
IS       0.54569240  0.10103855 -0.33589603
ARE      0.09333473  2.01820580  0.93733545
I        2.13321049  0.95322051 -2.59415168
WE       2.08553043 -1.82239542  1.07128554
THERE    0.47367150  1.93493888 -1.09686233
YOU     -3.17585230 -0.10716017 -1.54233423
PEOPLE  -0.20347747  0.33121970  1.11858522
IT      -0.20267349 -0.96006973 -0.94807859
LIVING  -0.77623788 -0.41251027  0.86224873

 Column scores:
             [,1]       [,2]       [,3]
THAI    -1.0492966  2.3559075  0.7702019
TAIWAN  -0.8067538 -0.9490640 -0.6210297
RAK     -1.4067583 -0.4813943 -0.2129575
KOREA    0.7618618 -0.4660735  1.2085319
JAPAN    1.0964867  0.5655380 -1.3131196
```

3.B.4.2 Graf

Obrázek č. 16: graf řádkových a sloupcových profilů



3.B.5 Závěr

Z protokolu CA plyne, že první dvě komponenty popisují 91 % variability datech; dvojrozměrný graf je tedy postačující. Korespondenční analýza na základě frekvence nejpoužívanějších slov rozdělila skupiny uživatelů angličtiny takto (v závorce jsou nejfrekventovanější slova příslušející skupině:

1. JAPAN + KOREA
(I, WE, IS, CITY, MANY)

2. RAK + TAIWAN, THAI
(YOU, THE, IN, LIVING, IT, AND)

Nejzajímavější je naprostá rozdílnost v používání osobních zájmen I, WE ⇔ YOU. Rozborem analyzovaných textů bylo zjištěno, že osobní zájména jsou často používána nesprávně; tato skutečnost může být zdrojem nedorozumění při komunikaci v cizím (anglickém) jazyce.