Výhody vícerozměrné statistické analýzy dat

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice,
email: milan.meloun@upce.cz, http://meloun.upce.cz

Souhrn: *Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných,* $y = w_1x_1 + \ldots + w_mx_m$. *Zdrojová matice dat obsahuje proměnné v m sloupcích a objekty v n řádcích. Data jsou před zpracováním škálována. Cílem je nalézt shluk jako množinu podobných objektů s podobnými proměnnými. Podobnost objektů posuzujeme na základě vzdálenosti (míry) objektů v m-rozměrném prostoru: čím je vzdálenost shluků či objektů větší, tím menší je jejich podobnost. K rychlému posouzení podobnosti slouží grafy exploratorní analýzy vícerozměrných dat: profily, polygony, sluníčka a hvězdičky. Strukturu a vazby mezi proměnnými vystihují metody snížení dimensionality, metoda hlavních komponent (PCA). Důležitou pomůckou je rozptylový diagram, který zobrazuje objekty, rozptýlené v rovině prvních dvou hlavních komponent. Graf komponentních vah porovnává vzdálenosti mezi proměnnými* x_i *a* x_j, *kde krátká vzdálenost značí silnou korelaci. Dvojný graf pak kombinuje oba předchozí grafy. Objekty lze seskupovat do shluků hierarchicky dle předem zvoleného způsobu metriky (průměrově, centroidně, nejbližším sousedem, nejvzdálenějším sousedem, medianově, mezi těžišti a průměrnou vazbou) a nehierarchicky dle uživatelem vybraných objektů-představitelů. Výsledkem je dendrogram. Metoda hlavních komponent a tvorba shluků je demonstrována na dvou vzorových úlohách.*

Vícerozměrná statistická analýza vychází z koncepce latentních proměnných (faktorů, kanonických proměnných) y, které jsou lineární kombinací původních proměnných x s vhodně volenými vazbami. Latentní proměnná y je kombinací m-tice sledovaných (měřených resp. jinak získaných) proměnných $x_1, x_2, \ldots, x_m$ ve tvaru $y = w_1x_1 + w_2x_2 + \ldots + w_mx_m$. Jednotlivé vícerozměrné metody využívají různých způsobů stanovení vah $w_1, w_2, \ldots, w_m$. Zdrojová matice má rozměr $n \times m$. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje (a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalézt *vybočující objekty*, resp. jejich proměnné, (c) stanovit, zda lze použít předpoklad *lineárních vazeb,* (d) ověřit *předpoklady o datech* (normalita, nekorelovanost, homogenita).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda se hledají (a) *struktura a vazby v proměnných* nebo (b) *struktura a vazby v objektech*:

(1) Hledání struktury *v proměnných* v metrické škále: *faktorová analýza FACT* a *analýza hlavních komponent PCA*.
(2) Hledání struktury *v objektech* v metrické škále: *shluková analýza*.
(3) Hledání struktury *v objektech* v metrické i v nemetrické škále: *vícerozměrné škálování*.
(4) Hledání struktury *v objektech* v nemetrické škále: *korespondenční analýza*.
(5) Většina metod vícerozměrné statistické analýzy umožňuje *zpracování lineárních vícerozměrných modelů*, kde závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných resp. vazby mezi proměnnými jsou lineární. V řadě případů se také uvažuje normalita metrických proměnných.

Určením struktury a vzájemných vazeb mezi proměnnými ale i mezi objekty se zabývají techniky redukce proměnných na latentní proměnné, metoda *analýzy hlavních komponent (PCA)* a *metoda faktorové analýzy (FA)*. Důležitou metodou určení vzájemných vazeb mezi proměnnými je i *kanonická korelační analýza CA*, která se používá ke zkoumání závislosti mezi dvěma skupinami proměnných, přičemž jedna ze skupin se považuje za proměnné nezávislé a druhá za skupinu proměnných závislých.

Vzorová úloha 1. *Sledování spotřeby proteinů v Evropě*

Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem vyšetření: existuje korelace mezi proměnnými? Budou data vyžadovat standardizaci ? Ukazuje graf komponentních vah na silně korelující proměnné? Jsou některé proměnné redundandní? Lze odhalit v rozptylovém diagramu komponentního skóre odlehlé objekty, výjimečné co do spotřeby proteinů? Které země jsou si podobné ve spotřebě proteinů? Komentujte vzniklé shluky zemí co do spotřeby proteinů.

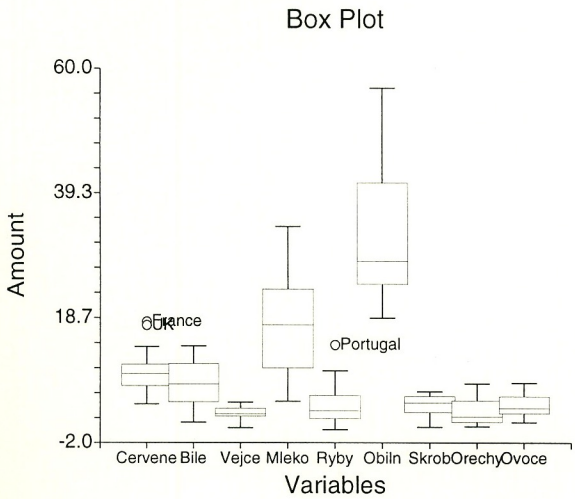
Data: i značí index, ***Cervene*** červené maso, **Bílé** maso, **Vejce, Mléko, Ryby, Obilniny, Škrob, Ořechy, Ovoce** a zelenina,

	Objekty	Proměnné								
i	Stát	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
1	Albania	10.10	1.40	0.50	8.90	0.20	42.30	0.60	5.50	1.70
2	Austria	8.90	14.00	4.30	19.90	2.10	28.00	3.60	1.30	4.30
3	Belgium	13.50	9.30	4.10	17.50	4.50	26.60	5.70	2.10	4.00
4	Bulgaria	7.80	6.00	1.60	8.30	1.20	56.70	1.10	3.70	4.20
5	Czechoslov.	9.70	11.40	2.80	12.50	2.00	34.30	5.00	1.10	4.00
6	Denmark	10.60	10.80	3.70	25.00	9.90	21.90	4.80	0.70	2.40
7	E Germany	8.40	11.60	3.70	11.10	5.40	24.60	6.50	0.80	3.60
8	Finland	9.50	4.90	2.70	33.70	5.80	26.30	5.10	1.00	1.40
9	France	18.00	9.90	3.30	19.50	5.70	28.10	4.80	2.40	6.50
10	Greece	10.20	3.00	2.80	17.60	5.90	41.70	2.20	7.80	6.50
...	.....	...	...	...	...	...	...	...	...	...
...	.....	...	...	...	...	...	...	...	...	...
23	USSR	9.30	4.60	2.10	16.60	3.00	43.60	6.40	3.40	2.90
24	W Germany	11.40	12.50	4.10	18.80	3.40	18.60	5.20	1.50	3.80
25	Yugoslavia	4.40	5.00	1.20	9.50	0.60	55.90	3.00	5.70	3.20

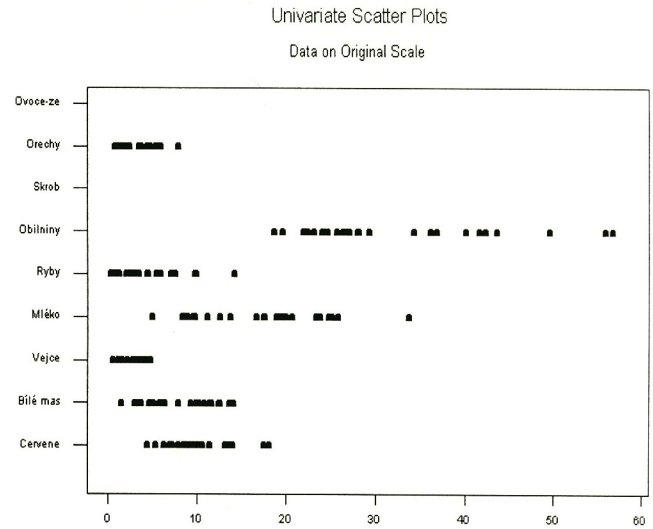
Řešení: k analýze byl použit program NCSS2000.

1. Exploratorní analýza:

Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především *symbolové grafy.*

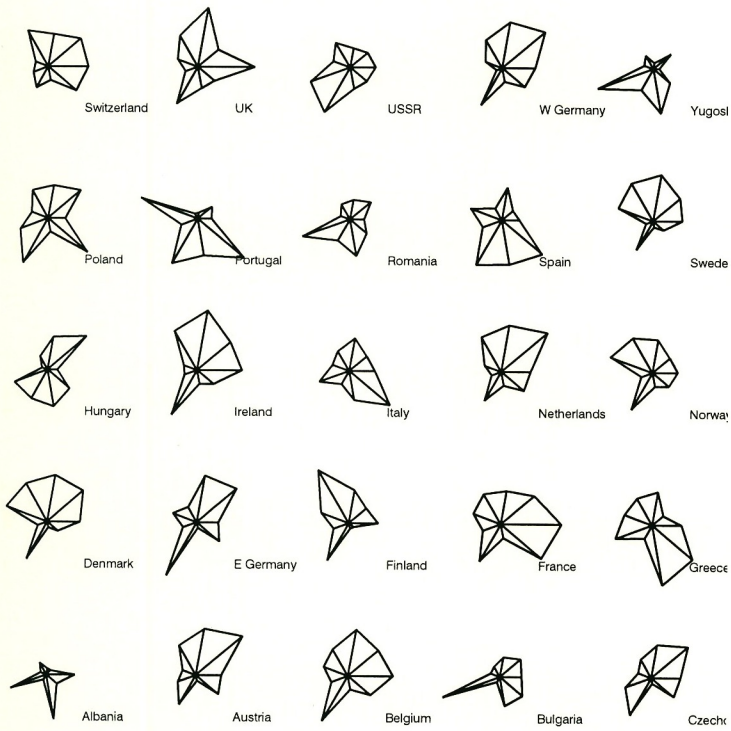


Krabicové grafy původních dat

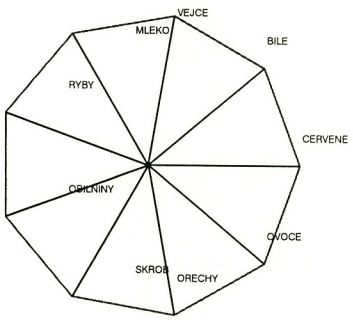


Rozptylová diagram původních dat

Polygony jsou vlastně profily v polárních souřadnicích, kdy každá proměnná objektu x_i^T, $i = 1, ..., n$, odpovídá délce paprsku vycházejícího ze společného středu.



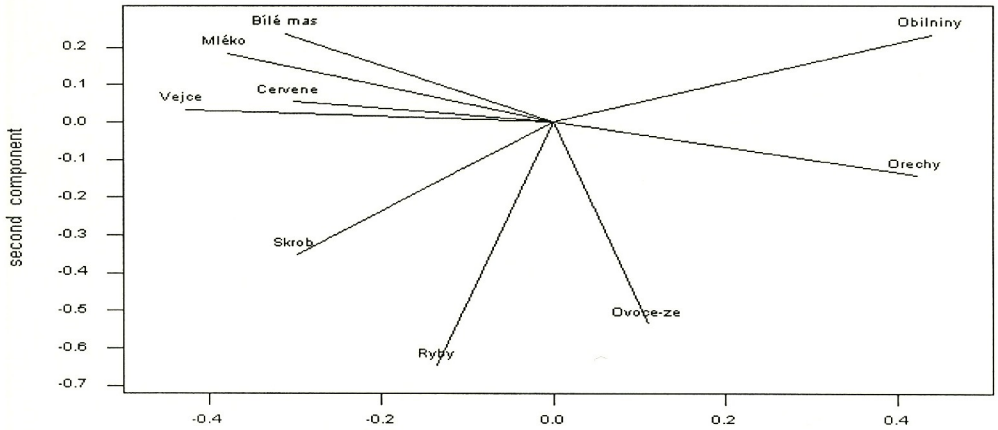
Hvězdičkový graf



Klíč k hvězdičkovému grafu

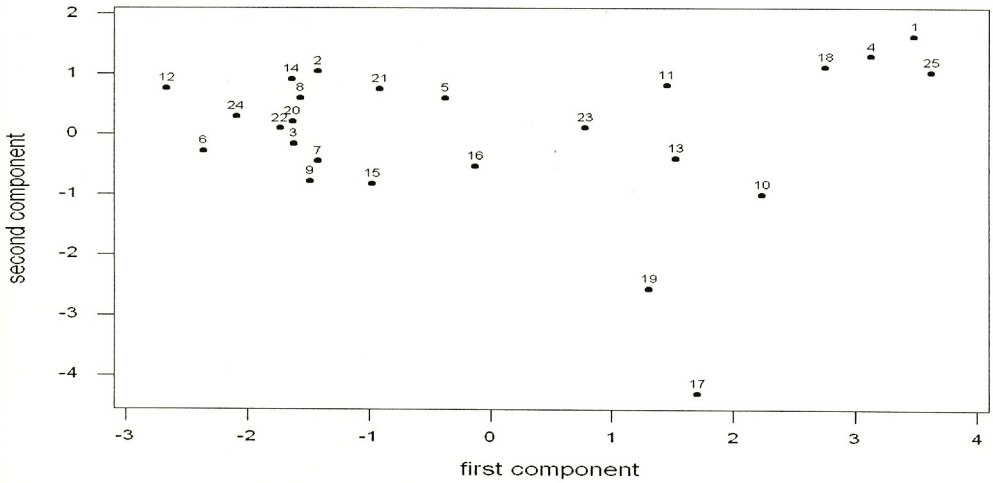
2. Metoda hlavních komponent:

1. **Vyšetření indexového grafu úpatí vlastních čísel:** Vlastní čísla slouží k určení počtu A "využitelných" hlavních komponent, jež si zvolíme v analýze k dalšímu užívání.
2. **Vyšetření grafu komponentních vah:** tento graf ukazuje, že proměnné *Bílé maso-Mléko-Červené maso-Vejce* spolu silně korelují. Ukazuje se, že by bylo vhodné jejich počet zredukovat na dvě proměnné, např. *Bílé maso-Vejce.* Mezi průvodiči ostatních proměnných je dostatečně velký úhel, a tím malá korelace. Proměnné, které spolu nekorelují mohou být ponechány ve vstupních datech.



Graf komponentních vah (Components Weights Plot)

3. **Vyšetření rozptylového diagramu komponentního skóre:** nejdůležitější diagram metody hlavních komponent ukazuje celou vyšetřovanou strukturu objektů, tzn. shluky objektů, izolované objekty, odlehlé objekty, anomálie, atd. Objekty mohou být označeny textovým popisem nebo číselně indexem. V pravém horním rohu se dobře oddělil shluk objektů: *1(Albanie)*, *4 (Bulharsko), 18 (Rumunsko), 25(Jugoslávie)*, který pokrývá země Balkánu. Vyjímečné postavení mají *17(Portugalsko), 19(Španělsko)*. Ostatní státy jsou kromě *11, 13, 23* v jednom společném shluku.



Rozptylový diagram komponentního skóre (Scatterplot)

4. **Vyšetření dvojného grafu:** je důležité sledovat interakci objektů a proměnných. Je-li některý objekt umístěn ve dvojném grafu na stejném místě nebo alespoň poblíž místa proměnné, jsou spolu v interakci. Interakce poslouží interpretaci objektů.

Klasifikace objektů analýzou shluků

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do jedné existující třídy (*diskriminační analýza DA*), nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd či shluků (*analýza shluků CLU*). Analýza shluků patří mezi metody, které se zabývají vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili *shluků*. Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Podle způsobu shlukování se postupy dělí na *hierarchické* a *nehierarchické*. Hierarchické se dělí dále na *aglomerativní* a *divizní*.

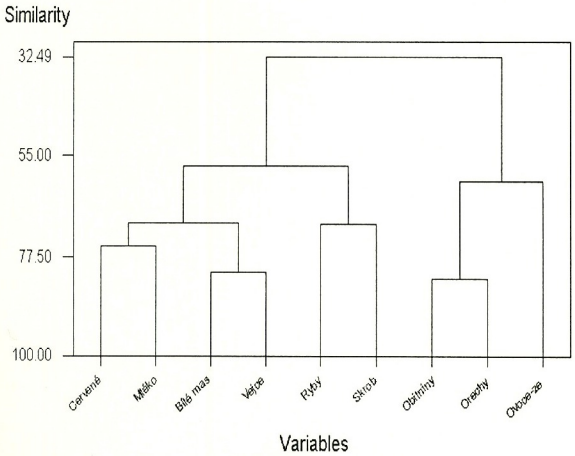
Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Dendrogram podobnosti proměnných odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) proměnných, které jsou si velmi podobné a silně spolu korelují.

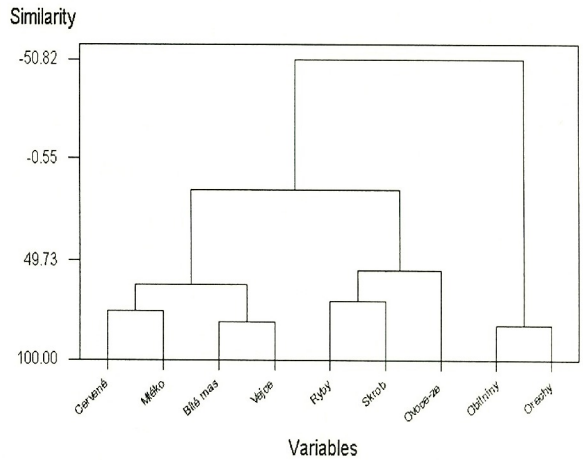
Vzorová úloha 2. *Klasifikace spotřeby proteinů v Evropě*

Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem klasifikace. Které země jsou si podobné ve spotřebě proteinů? Komentujte vzniklé shluky zemí co do spotřeby proteinů.

Řešení: Dendrogram podobnosti proměnných obsahuje dvojice nebo trojice proměnných, které jsou si velmi podobné a silně spolu korelují. Metodou nejbližšího souseda jsou to trojice *Červené maso-Bílé maso-Vejce*, k tomu se přidružuje také proměnná *Mléko*. Další dvojice je *Obilniny-Ořechy* (obr. 13a). Metoda nejvzdálenějšího souseda ukazuje na 4 dvojice: první dvojice *Červené maso-Mléko*, druhá *Bílé maso-Vejce*, třetí *Ryby-Škrob*, čtvrtá *Obilniny-Ořechy* (obr. 13b). Matoda Wardova poskytla stejné shluky jako metoda nejvzdálenějšího souseda.

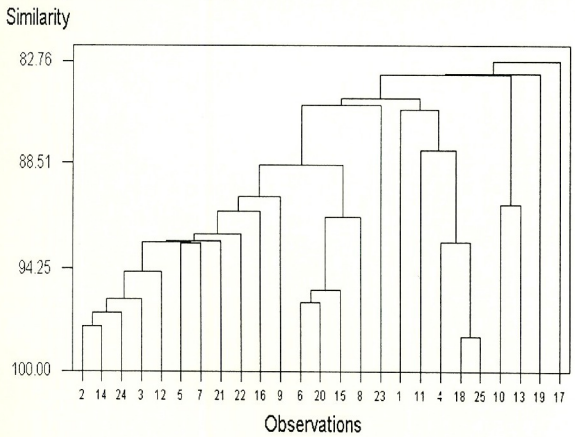


Dendrogram proměnných metodou průměrovou

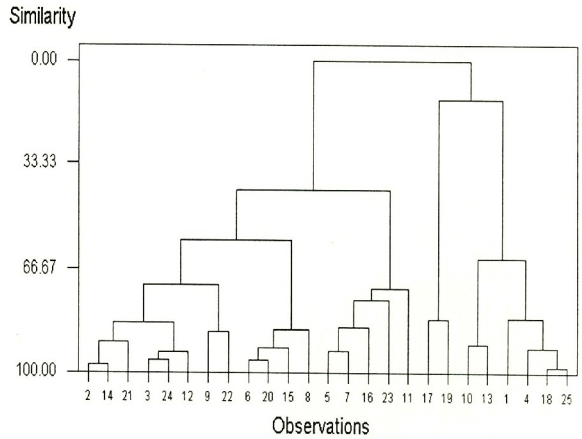


Dendrogram proměnných metodou Wardovou

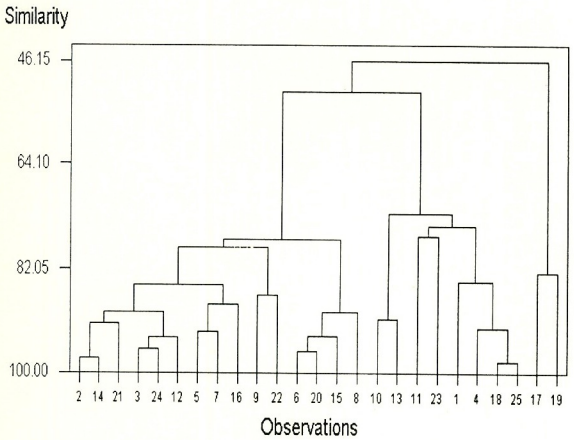
Nejdůležitějším dendrogramem je dendrogram podobnosti objektů, ze kterého je patrná struktura objektů ve shlucích, roztřídění států Evropy dle spotřeby proteinů na základě 9 kritérií a vzájemné podbnosti.



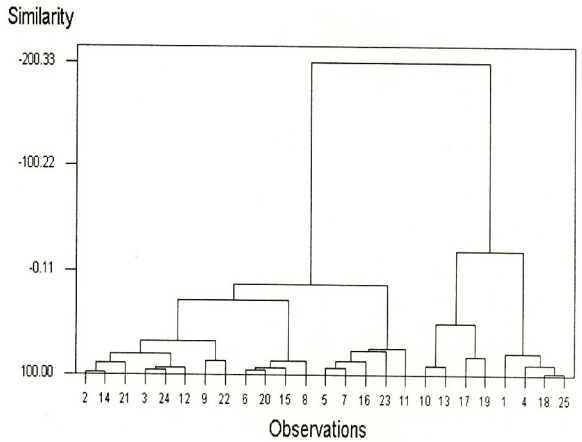
Dendrogram objektů metodou nejbližšího souseda, Minitab



Dendrogram objektů metodou nejvzdálenějšího souseda, Minitab



Dendrogram objektů metodou průměrovou, Minitab



Dendrogram objektů metodou Wardovou, Minitab

Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.

[2] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis,* Ellis Horwood, Chichester 1992.

[3] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,

[4] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,

[5] Jeffers J. N. R., *Applied Statistician*, **16**, 225 (1967).

[6] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994.

[7] Martens H., Naes T., *Multivariate calibration*, Wiley (1989) Chichester.

[8] Malinowski F., Howery D., *Factor Analysis in Chemistry*, Wiley (1980) New York.

[9] Meloun M. , Militký J., *Sbírka úloh - Statistické zpracování experimentálních dat*, Univerzita Pardubice, 1996.