Shluková analýza dat a stanovení počtu shluků

Autor: Tomáš Löster
Vysoká škola ekonomická v Praze

Ostrava, červen 2017

Osnova prezentace

- Úvod a teorie shlukové analýzy
- Podrobný popis shlukování na příkladu
- Hodnocení shlukovacích postupů
- Závěr

Shluková analýza a její význam

- jedna z technik vícerozměrné analýzy dat
- základní cíl spočívá v rozdělení objektů do co nejhomogennějších shluků
- zastává významnou roli v mnoha odvětvích
- proces shlukování může přinášet různá výsledná rozdělení objektů do shluků v závislosti na použitých metodách a různých specifikacích => **hodnocení výsledků**

Vymezení použitých symbolů

$\mathbf{x}_i$	i-tý objekt (sloupcový vektor)
n	počet objektů
m	počet proměnných, které charakterizují objekty
k	počet shluků
C_h	h -tý shluk
n_h	počet objektů v h-tém shluku
$\mathbf{D}$	matice vzdáleností
D_{ij}	vzdálenost (odlišnost) mezi i-tým a j-tým objektem
$\bar{\mathbf{x}}_h$	centroid h-tého shluku
s_t^2	výběrový rozptyl t-té proměnné
s_{ht}^2	výběrový rozptyl t-té proměnné v h-tém shluku
$\mathbf{S}$	kovarianční matice

Vybrané míry vzdálenosti

- Euklidovská míra vzdálenosti
- Mahalanobisova míra vzdálenosti

Euklidovská vzdálenosti

- = délka přepony pravoúhlého trojúhelníka. Výpočet mezi *i*-tým a *j*-tým objektem je založen na Pythagorově větě:

$$D_{\mathrm{E}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{t=1}^{m} \left(x_{it} - x_{jt}\right)^2}$$

Mahalanobisova vzdálenost

- Odstraňuje problém, který vzniká při použití nestandardizovaných dat, v důsledku odlišností měrných jednotek.

$$D_{\mathrm{Ma}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

Metody shlukování

- Hierarchické metody
  - Aglomerativní (postupné „připojování“ objektů)
  - Divizní (postupné „odpojování“ objektů)
- Nehierarchické metody
  - Např. metody „k-shlukování“

Vybrané metody hierarchického shlukování

- Metoda nejbližšího souseda
- Metoda nejvzdálenějšího souseda
- Centroidní metoda
- Metoda průměrné vazby
- Wardova metoda

Vybrané metody hierarchického shlukování

- **Metoda nejbližšího souseda**

  = postup je založen na minimální vzdálenosti.

  - Nejprve se spojí dva objekty, které mají nejkratší vzdálenost =>shluk.
  - Postupuje se do situace, kdy všechny objekty tvoří 1 shluk.
  - Problém s řetězením: spojují se dva shluky, kde vzdálenost mezi dvěma objekty je nejmenší, ale nemusí se jednat o nejbližší shluky.

Vybrané metody hierarchického shlukování

- **Metoda nejvzdálenějšího souseda**

  = postup je založen na maximální vzdálenosti.

  - Nejdelší vzdálenost mezi objekty ve shluku představuje nejmenší kouli, která obklopuje všechny objekty v obou shlucích.
  - Problém s řetězením: není

Vybrané metody hierarchického shlukování

- **Centroidní metoda**

  = postup je založen na analýze vzdáleností mezi těžišti (centroidy = vektory průměrů) jednotlivých shluků.

  - Výhoda: menší náchylnost k ovlivnění výsledných shluků extrémními či odlehlými objekty

Vybrané metody hierarchického shlukování

- **Metoda průměrné vazby**

  = postup je založen na průměrné vzdálenosti objektů v jednom shluku ke všem objektům ve druhém shluku.

  - Výhoda: kritérium je zaleženo na všech objektech

Vybrané metody hierarchického shlukování

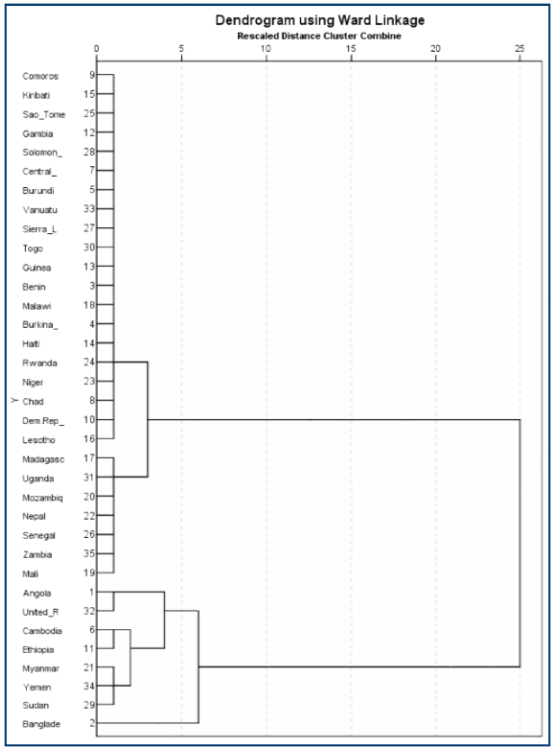
- **Wardova metoda**

  = postup není založen na optimalizaci vzdáleností mezi shluky.

  - Řeší se minimalizace heterogenity shluků podle přírůstku vnitroshlukového součtu čtverců odchylek objektů od těžiště (centroidů) shluků.

Vyhodnocení počtu shluků

- dendrogram



Vybrané koeficienty pro stanovení počtu shluků

- Daviesův-Bouldinův index
- RS (též R-kvadrát, RSQ index)
- RMSSTD (*root-mean-square standard deviation index)*
- CHF (pseudo *F* index)
- PTS (pseudo T-kvadrát index)
- Dunnův index

RS (též R-kvadrát, RSQ index)

- Součty čtverců:

$$SS_{\mathrm{W}} = \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} \sum_{t=1}^{m} (x_{it} - \bar{x}_{ht})^2$$

$$SS_{\mathrm{T}} = \sum_{i=1}^{n} \sum_{t=1}^{m} (x_{it} - \bar{x}_{t})^2$$

$$SS_{\mathrm{B}} = SS_{\mathrm{T}} - SS_{\mathrm{W}}$$

$$I_{\mathrm{RS}} = \frac{SS_{\mathrm{B}}}{SS_{\mathrm{T}}} = \frac{SS_{\mathrm{T}} - SS_{\mathrm{W}}}{SS_{\mathrm{T}}}$$

RMSSTD

- Měří homogenitu nových shluků; je založen pouze na vnitroshlukové variabilitě.

$$I_{\mathrm{RMSSTD}}(k) = \sqrt{\frac{SS_{\mathrm{W}}}{m \cdot (n-k)}}$$

CHF

- analogie *F*-testu (ANOVA):

$$I_{\mathrm{CHF}}(k) = \frac{\frac{SS_{\mathrm{B}}}{k-1}}{\frac{SS_{\mathrm{W}}}{n-k}} = \frac{(n-k)\cdot SS_{\mathrm{B}}}{(k-1)\cdot SS_{\mathrm{W}}}$$

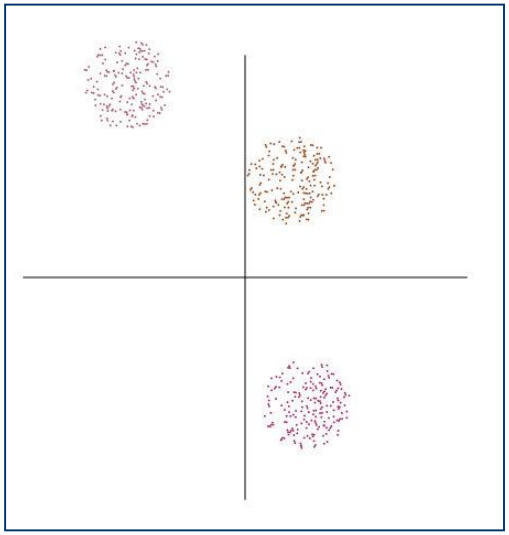
$$I_{\mathrm{CHF}}(k^*) = \max_{2 \le k \le n-1} I_{\mathrm{CHF}}(k)$$

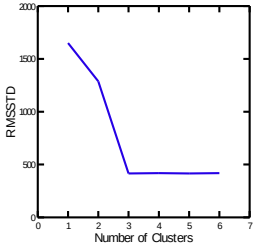
Implementované koeficienty v SW

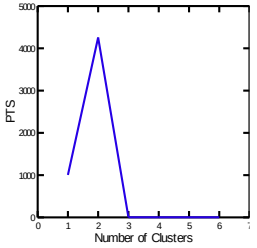
Koeficient	Hledaný extrém	Software
CHF index (pseudo F)	maximum	SAS system LE, SYSTAT
PTS index (T-kvadrát)	minimum	SAS system LE, SYSTAT
RS (R-kvadrát, RSQ)	maximum	SAS system LE
SPRS (SPRSQ)	minimum	SAS system LE
BIC, AIC	minimum	SPSS
RMSSTD	minimum	SYSTAT
Daviesův-Bouldinův (DB)	minimum	SYSTAT
Dunnův separační index	maximum	SYSTAT

Ilustrace vyhodnocení počtu shluků

- 3 dobře separované shluky

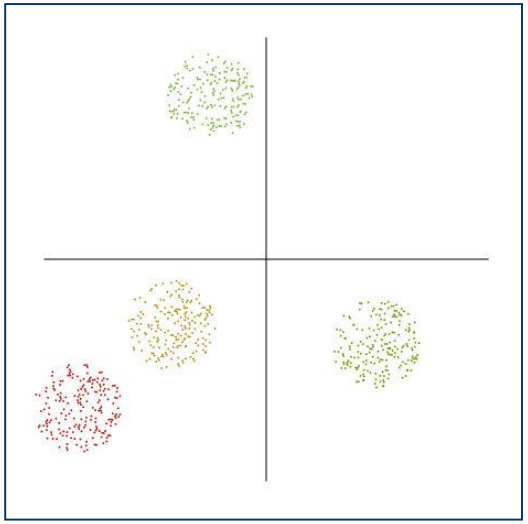




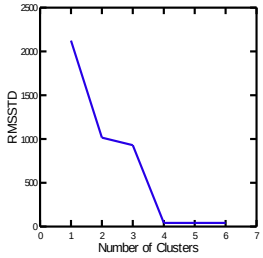


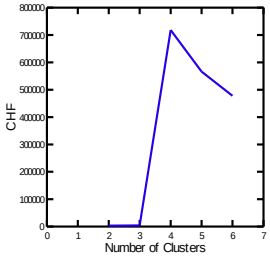
Ilustrace vyhodnocení počtu shluků

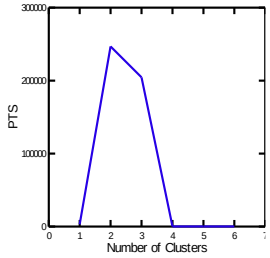
- 4 dobře separované shluky

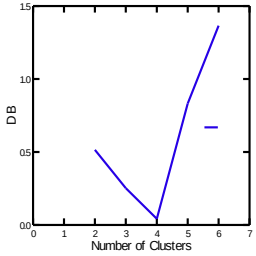


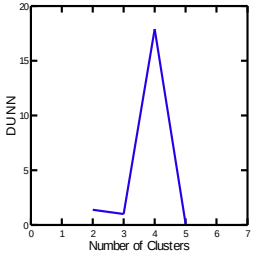
Validity Index Plot





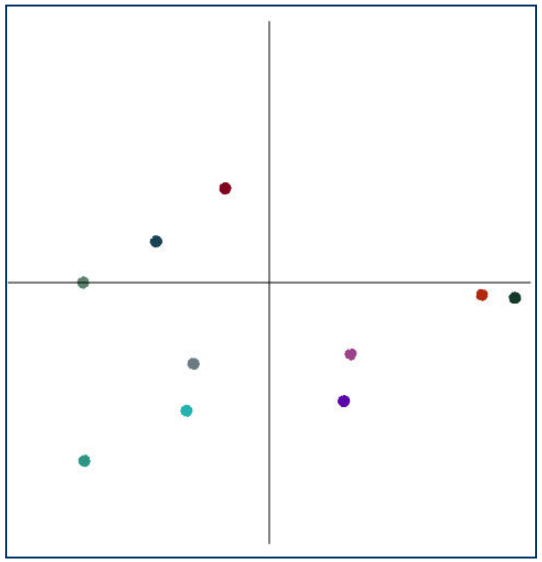




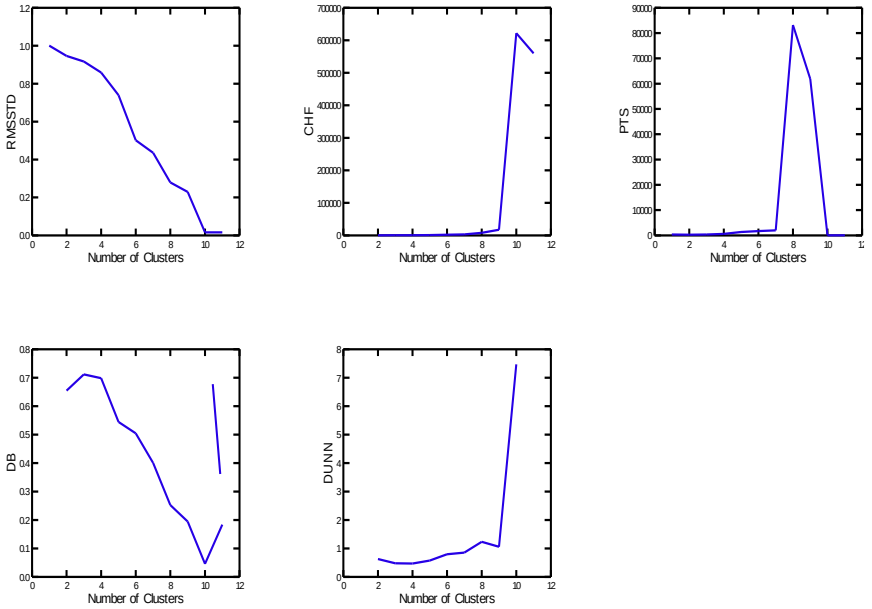


Ilustrace vyhodnocení počtu shluků

- 10 dobře separovaných shluků

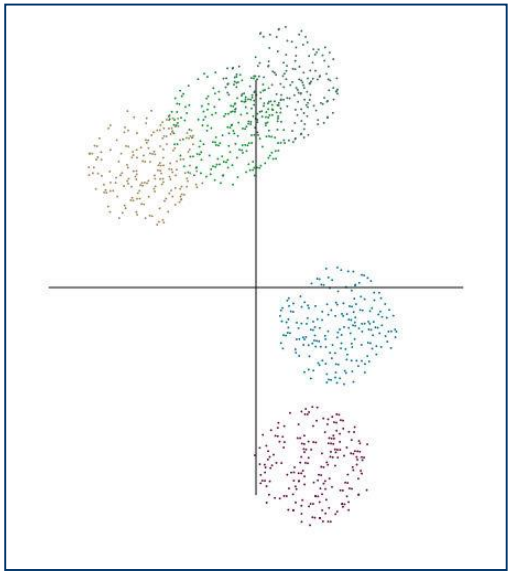


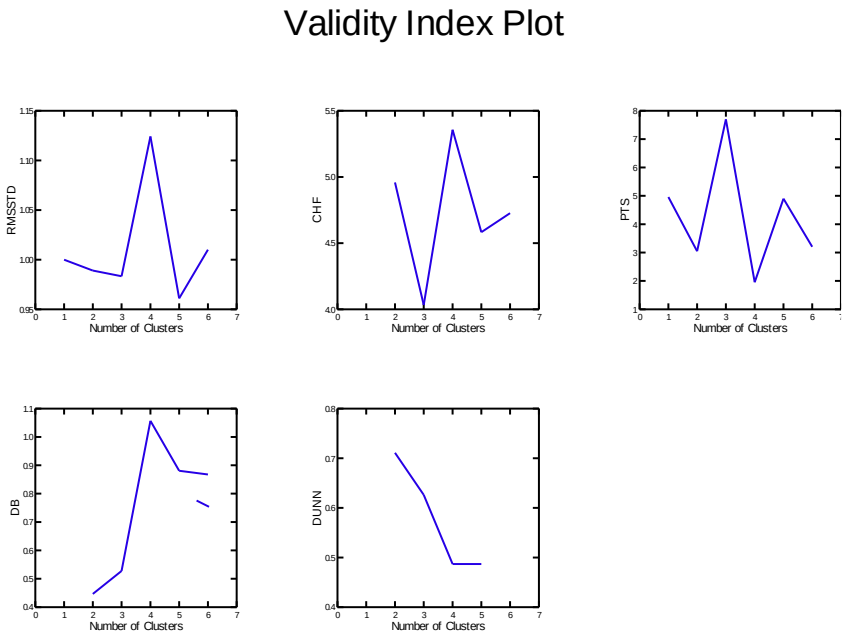
Validity Index Plot



Ilustrace vyhodnocení počtu shluků

- 5 překrývajících se shluků





Proces shlukování – Soubor Wine

- Pochází z roku 1991; 178 vzorků vín; 13 kvant. proměnných; 3 různé odrůdy

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Alkohol	178	11,03	14,83	13,0006	,81183
Kyselina_jablecna	178	,74	5,80	2,3363	1,11715
Popel	178	1,36	3,23	2,3665	,27434
Stupen_zasaditosti	178	10,60	30,00	19,4949	3,33956
Magnesium	178	70,00	162,00	99,7416	14,28248
Fenoly_celkem	178	,98	3,88	2,2951	,62585
Flavanoidy	178	,34	5,08	2,0293	,99886
Fenoly_2	178	,13	,66	,3619	,12445
Proanthokyanidiny	178	,41	3,58	1,5909	,57236
Intenzita_barvy	178	1,28	13,00	5,0581	2,31829
Odstín	178	,48	1,71	,9574	,22857
Variable	178	1,27	4,00	2,6117	,70999
Aminokyseliny	178	278,00	1680,00	746,8933	314,90747
Valid N (listwise)	178				

Proces shlukování – Soubor Wine

- Korelační matice

Correlations

		Alkohol	Kyselina_jablecna	Popel	Stupen_zasaditosti	Magnesium	Fenoly_celkem	Flavanoidy	Fenoly_2	Proanthokyanidiny	Intenzita_barvy	Odstín	VAR00013	Aminokyseliny
Alkohol	Pearson Correlation	1	,094	,212**	-,310**	,271**	,289**	,237**	-,156*	,137	,546**	-,072	,072	,644**
	Sig. (2-tailed)		,210	,005	,000	,000	,000	,001	,038	,069	,000	,341	,337	,000
Kyselina_jablecna	Pearson Correlation	,094	1	,164*	,289**	-,055	-,335**	-,411**	,293**	-,221**	,249**	-,561**	-,369**	-,192*
	Sig. (2-tailed)	,210		,029	,000	,469	,000	,000	,000	,003	,001	,000	,000	,010
Popel	Pearson Correlation	,212**	,164*	1	,443**	,287**	,129	,115	,186*	,010	,259**	-,075	,004	,224**
	Sig. (2-tailed)	,005	,029		,000	,000	,086	,126	,013	,898	,000	,322	,959	,003
Stupen_zasaditosti	Pearson Correlation	-,310**	,289**	,443**	1	-,083	-,321**	-,351**	,362**	-,197**	,019	-,274**	-,277**	-,441**
	Sig. (2-tailed)	,000	,000	,000		,269	,000	,000	,000	,008	,804	,000	,000	,000
Magnesium	Pearson Correlation	,271**	-,055	,287**	-,083	1	,214**	,196**	-,256**	,236**	,200**	,055	,066	,393**
	Sig. (2-tailed)	,000	,469	,000	,269		,004	,009	,001	,001	,007	,463	,381	,000
Fenoly_celkem	Pearson Correlation	,289**	-,335**	,129	-,321**	,214**	1	,865**	-,450**	,612**	-,055	,434**	,700**	,498**
	Sig. (2-tailed)	,000	,000	,086	,000	,004		,000	,000	,000	,465	,000	,000	,000
Flavanoidy	Pearson Correlation	,237**	-,411**	,115	-,351**	,196**	,865**	1	-,538**	,653**	-,172*	,543**	,787**	,494**
	Sig. (2-tailed)	,001	,000	,126	,000	,009	,000		,000	,000	,021	,000	,000	,000
Fenoly_2	Pearson Correlation	-,156*	,293**	,186*	,362**	-,256**	-,450**	-,538**	1	-,366**	,139	-,263**	-,503**	-,311**
	Sig. (2-tailed)	,038	,000	,013	,000	,001	,000	,000		,000	,064	,000	,000	,000
Proanthokyanidiny	Pearson Correlation	,137	-,221**	,010	-,197**	,236**	,612**	,653**	-,366**	1	-,025	,296**	,519**	,330**
	Sig. (2-tailed)	,069	,003	,898	,008	,001	,000	,000	,000		,738	,000	,000	,000
Intenzita_barvy	Pearson Correlation	,546**	,249**	,259**	,019	,200**	-,055	-,172*	,139	-,025	1	-,522**	-,429**	,316**
	Sig. (2-tailed)	,000	,001	,000	,804	,007	,465	,021	,064	,738		,000	,000	,000
Odstín	Pearson Correlation	-,072	-,561**	-,075	-,274**	,055	,434**	,543**	-,263**	,296**	-,522**	1	,565**	,236**
	Sig. (2-tailed)	,341	,000	,322	,000	,463	,000	,000	,000	,000	,000		,000	,002
VAR00013	Pearson Correlation	,072	-,369**	,004	-,277**	,066	,700**	,787**	-,503**	,519**	-,429**	,565**	1	,313**
	Sig. (2-tailed)	,337	,000	,959	,000	,381	,000	,000	,000	,000	,000	,000		,000
Aminokyseliny	Pearson Correlation	,644**	-,192*	,224**	-,441**	,393**	,498**	,494**	-,311**	,330**	,316**	,236**	,313**	1
	Sig. (2-tailed)	0	0	0	0	0	0	0	0	0	0	0	0	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

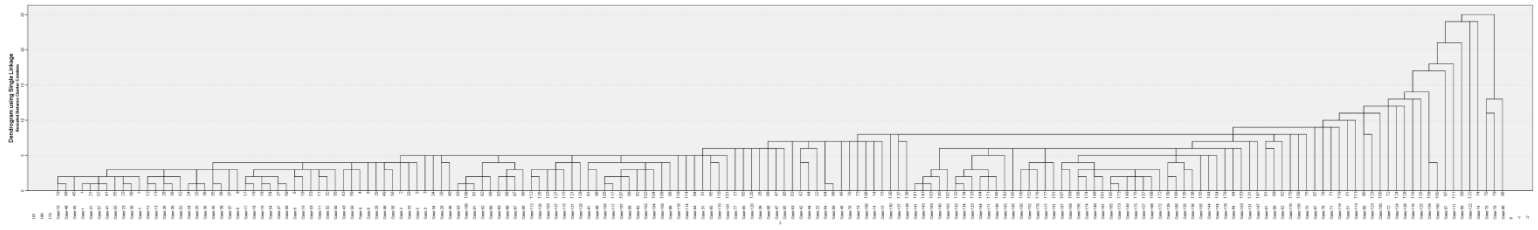
Proces shlukování – Soubor Wine

- **EUKLIDOVSKÁ VZDÁLENOST**

Proces shlukování – Soubor Wine

- Metoda nejbližšího souseda – s transformací

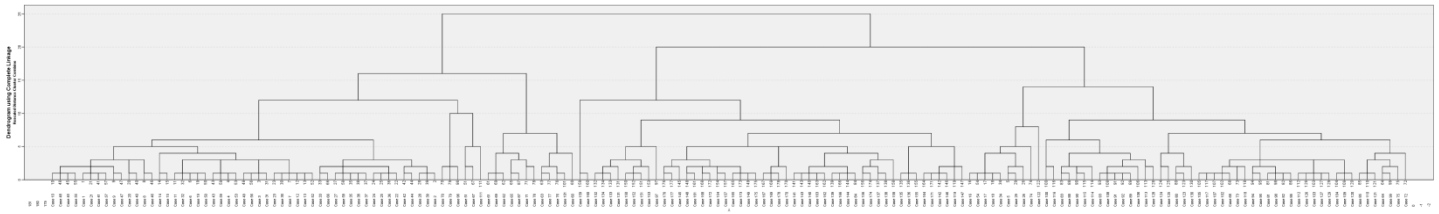
Odrůda/Shluk	1	2	3	Součet
1	0	59	0	59
2	3	67	1	71
3	0	48	0	48
Součet	3	174	1	178



Proces shlukování – Soubor Wine

- Metoda nejvzdálenějšího souseda – s transformací

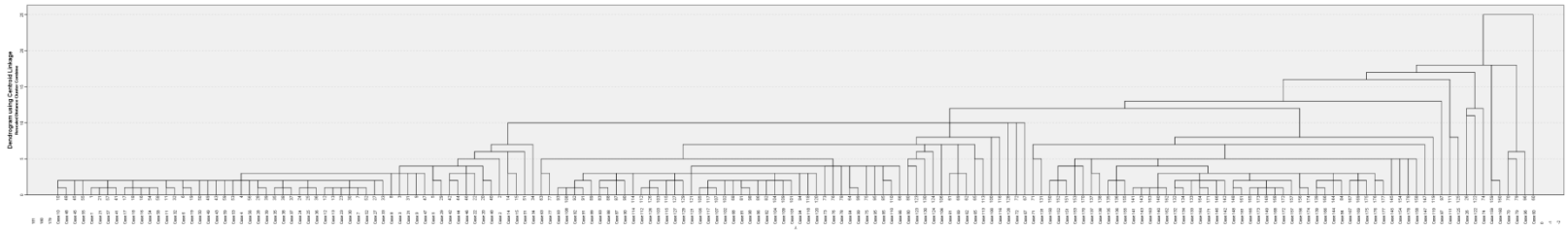
Odrůda/Shluk	1	2	3	Součet
1	51	8	0	59
2	18	50	3	71
3	0	0	48	48
Součet	69	58	51	178



Proces shlukování – Soubor Wine

- Centroidní metoda – s transformací

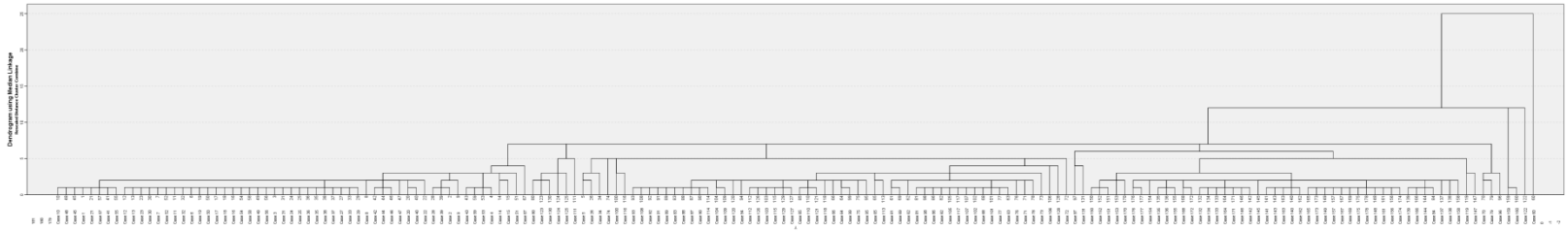
Odrůda/Shluk	1	2	3	Součet
1	0	59	0	59
2	1	67	3	71
3	0	48	0	48
Součet	1	174	3	178



Proces shlukování – Soubor Wine

- Mediánová metoda – s transformací

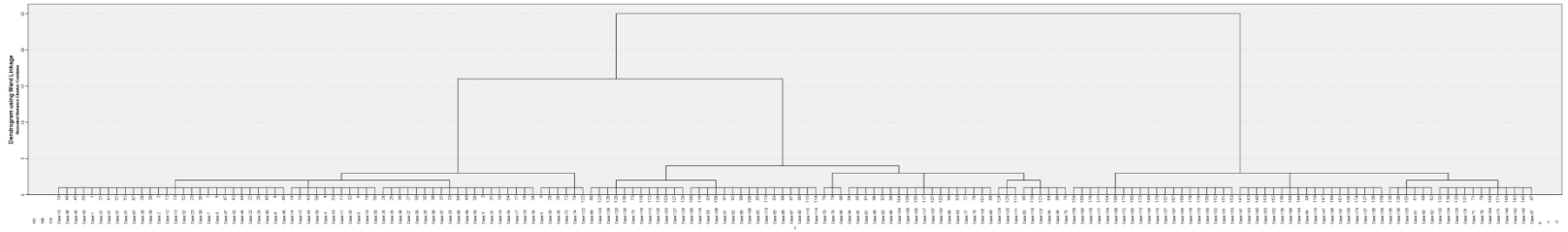
Odrůda/Shluk	1	2	3	Součet
1	0	59	0	59
2	1	69	1	71
3	0	48	0	48
Součet	1	176	1	178



Proces shlukování – Soubor Wine

- Wardova metoda – s transformací

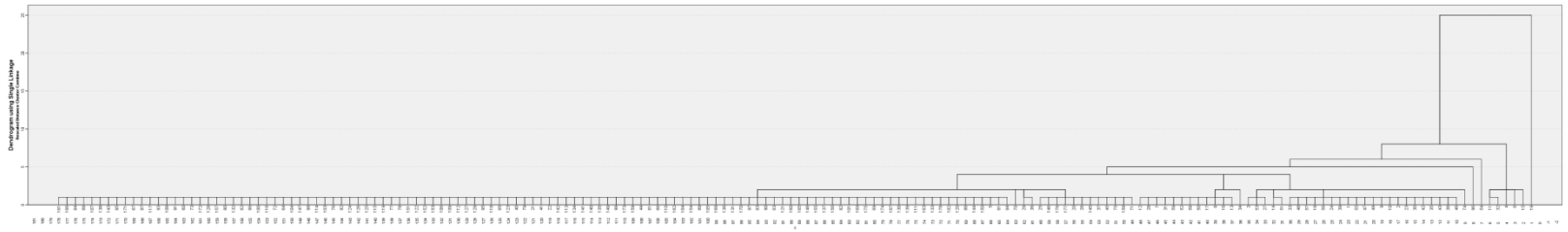
Odrůda/Shluk	1	2	3	Součet
1	59	0	0	59
2	5	58	8	71
3	0	0	48	48
Součet	64	58	56	178



Proces shlukování – Soubor Wine

- Metoda nejbližšího souseda – bez transformace

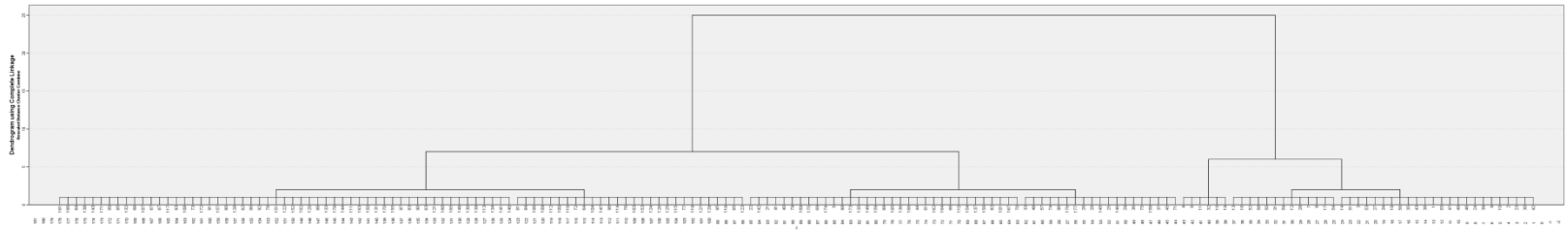
Odrůda/Shluk	1	2	3	Součet
1	5	53	1	59
2	0	71	0	71
3	0	48	0	48
Součet	5	172	1	178



Proces shlukování – Soubor Wine

- Metoda nejvzdálenějšího souseda – bez transformace

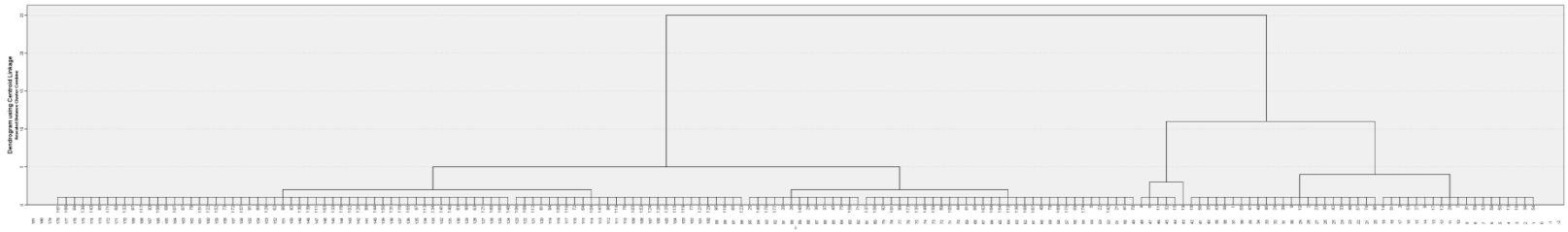
Odrůda/Shluk	1	2	3	Součet
1	43	0	16	59
2	0	56	15	71
3	0	27	21	48
Součet	43	83	52	178



Proces shlukování – Soubor Wine

- Centroidní metoda – bez transformace

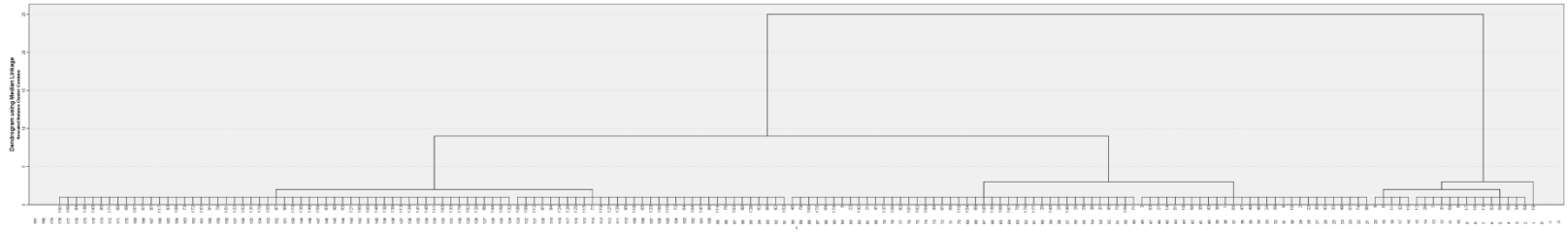
Odrůda/Shluk	1	2	3	Součet
1	40	13	6	59
2	2	69	0	71
3	0	48	0	48
Součet	42	130	6	178



Proces shlukování – Soubor Wine

- Mediánová metoda – bez transformace

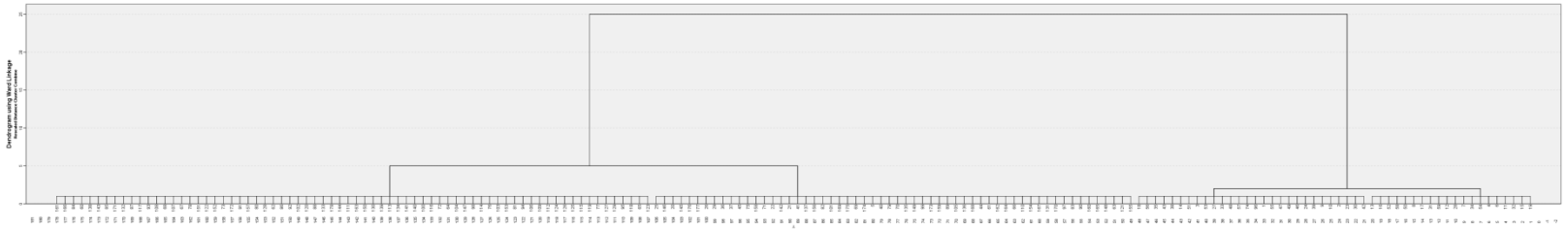
Odrůda/Shluk	1	2	3	Součet
1	39	0	20	59
2	14	57	0	71
3	17	31	0	48
Součet	70	88	20	178



Proces shlukování – Soubor Wine

- Wardova metoda – bez transformací

Odrůda/Shluk	1	2	3	Součet
1	46	0	13	59
2	2	51	18	71
3	0	21	27	48
Součet	48	72	58	178



Proces shlukování – Soubor Wine

- Vyhodnocení úspěšnosti klasifikace objektů

Metoda/způsob	S transformací	Bez transformace	Rozdíl
Nejbližšího souseda	37,64%	42,70%	**5,06%**
Nejvzdálenějšího souseda	83,71%	67,42%	**16,29%**
Centroidní metoda	37,64%	61,24%	**23,60%**
Mediánová metoda	38,76%	53,93%	**15,17%**
Wardova metoda	92,70%	69,66%	**23,03%**

Proces shlukování – Soubor Wine

- Počty shluků: Euklidovská vzdáleností

Metoda/koeficient	RMSSTD	CHF	PTS	D-B	Dunn	Správný počet shluků
Nejbližšího souseda	6	2	6	2	2	3
Nejvzdálenějšího souseda	4	**3**	2	2	4	3
Centroidní metoda	4	4	4	**3**	**3**	3
Mediánová metoda	5	5	2	**3**	**3**	3
Wardova metoda	4	6	2	2	2	3

Proces shlukování – Soubor Wine

- **MAHALANOBISOVA VZDÁLENOST**

Proces shlukování – Soubor Wine

- Metoda nejbližšího souseda – bez transformace

Odrůda/Shluk	1	2	3	Součet
1		59		59
2	1	69	1	71
3		48		48
Součet	176	1	1	178

Proces shlukování – Soubor Wine

- Metoda nejvzdálenějšího souseda – bez transformace

Odrůda/Shluk	1	2	3	Součet
1		59		59
2	6	64	1	71
3	2	46		48
Součet	169	8	1	178

Proces shlukování – Soubor Wine

- Centroidní metoda – bez transformace

Odrůda/Shluk	1	2	3	Součet
1		59		59
2	1	69	1	71
3		48		48
Součet	176	1	1	178

Proces shlukování – Soubor Wine

- Mediánová metoda – bez transformace

Odrůda/Shluk	1	2	3	Součet
1		59		59
2	1	69	1	71
3		48		48
Součet	176	1	1	178

Proces shlukování – Soubor Wine

- Wardova metoda – bez transformací

Odrůda/Shluk	1	2	3	Součet
1	27	32		59
2		71		71
3		25	23	48
Součet	27	128	23	178

Proces shlukování – Soubor Wine

- Vyhodnocení úspěšnosti klasifikace objektů

Metoda/způsob	Bez transformace
Nejbližšího souseda	38,76%
Nejvzdálenějšího souseda	35,96%
Centroidní metoda	38,76%
Mediánová metoda	38,76%
Wardova metoda	70,79%

Proces shlukování – Soubor Wine

- Charakteristiky jednotlivých odrůd

Proměnná/Odrůda	Odrůda 1		Odrůda 2		Odrůda 3	
	Průměr	Směr. Odchylka	Průměr	Směr. Odchylka	Průměr	Směr. Odchylka
Alkohol	13,7447	,46213	12,2787	,53796	13,1538	,53024
Kyselina_jablecna	2,0107	,68855	1,9327	1,01557	3,3338	1,08791
Stupen_zasaditosti	17,0373	2,54632	20,2380	3,34977	21,4167	2,25816
Magnesium	106,3390	10,49895	94,5493	16,75350	99,3125	10,89047
Fenoly_celkem	2,8402	,33896	2,2589	,54536	1,6788	,35697
Flavanoidy	2,9824	,39749	2,0808	,70570	,7815	,29350
Proanthokyanidiny	1,8993	,41211	1,6303	,60207	1,1535	,40884
Intenzita_barvy	5,5283	1,23857	3,0866	,92493	7,3962	2,31094
Odstín	1,0620	,11648	1,0563	,20294	,6827	,11444
Aminokyseliny	1115,7119	221,52077	519,5070	157,21122	629,8958	115,09704

odrůda/zařazení	1	2	3
1	0,66613	0,04134	0,29253
2	0,02399	0,69697	0,27904
3	0,01907	0,432910	0,54802

Hodnocení úspěšnosti

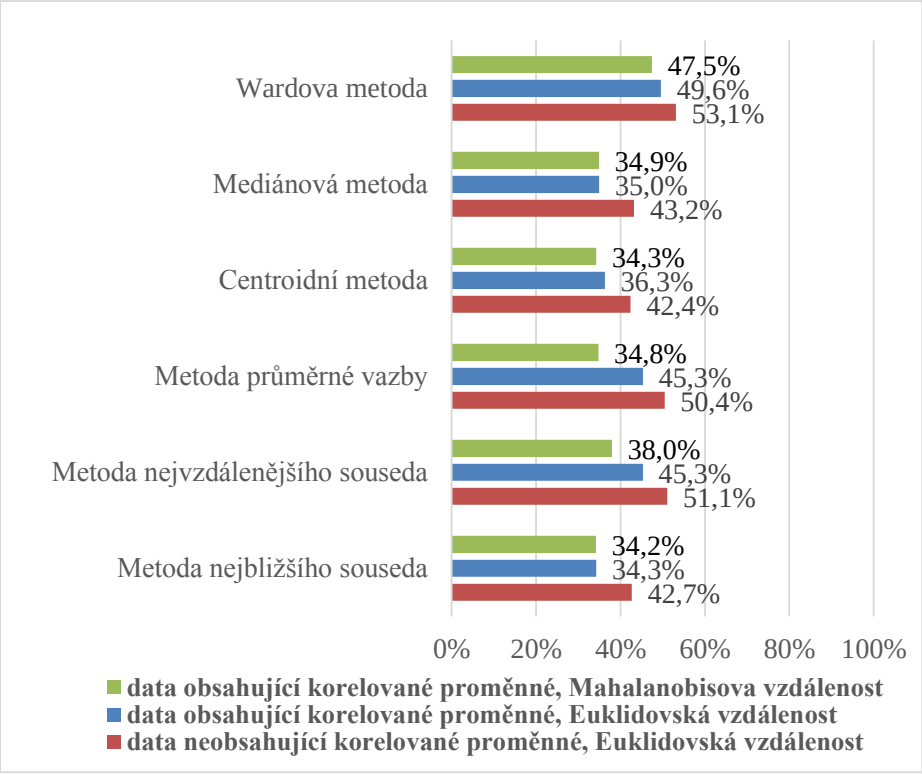
- Úspěšnost zařazení objektů do shluků
- Úspěšnost stanovení počtu shluků

Hodnocení úspěšnosti

- V rámci výzkumu byly uvažovány potenciální vlivy:
  - překrytí shluků
  - počtu shluků
  - korelace mezi proměnnými
  - vliv použité metody shlukování
  - vliv použité míry vzdálenosti
  - nestejné měrné jednotky (=> transformace)

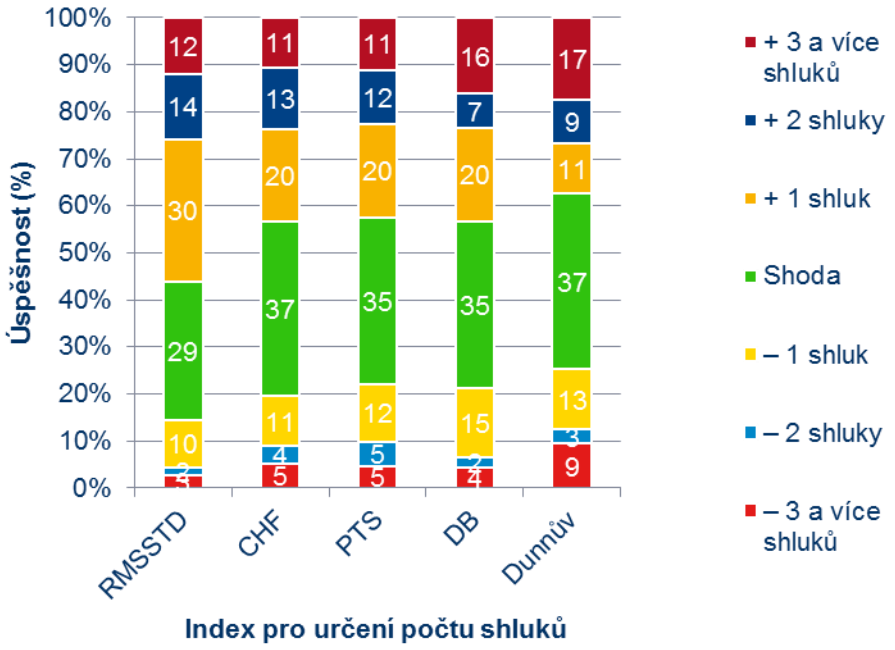
Úspěšnost metod shlukování

- Analyzováno celkem 23 různých souborů z UCI (Maršálková)



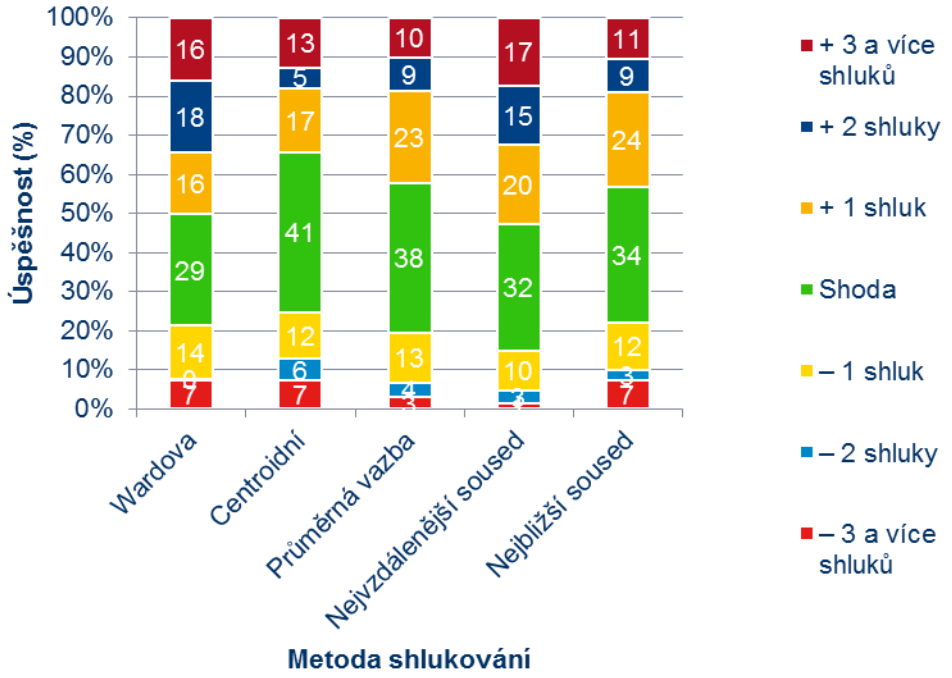
Úspěšnost koeficientů

- Analyzováno celkem 18 různých souborů z UCI (Hammerbauer) – celková úspěšnost bez ohledu na metody



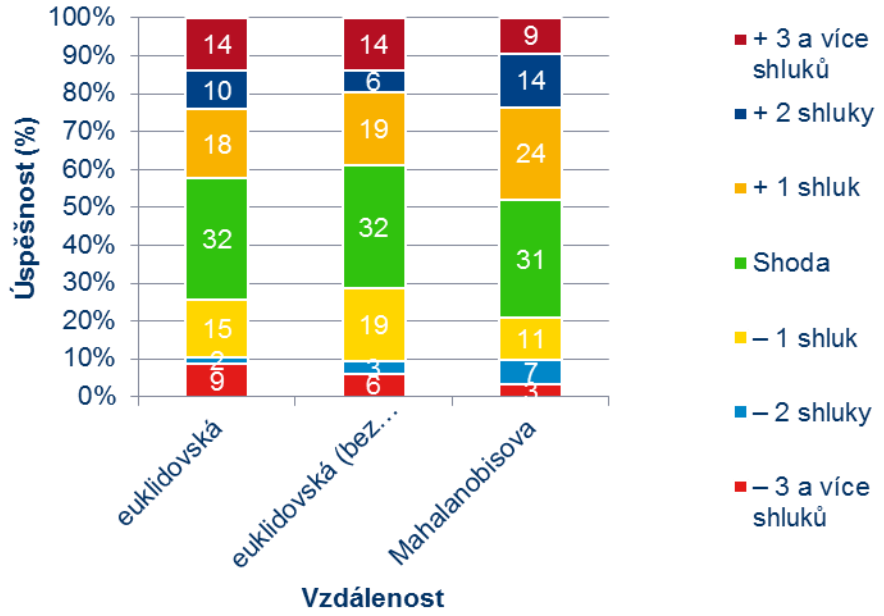
Úspěšnost koeficientů

- Analyzováno celkem 18 různých souborů z UCI (Hammerbauer) – úspěšnost koeficientů podle metody



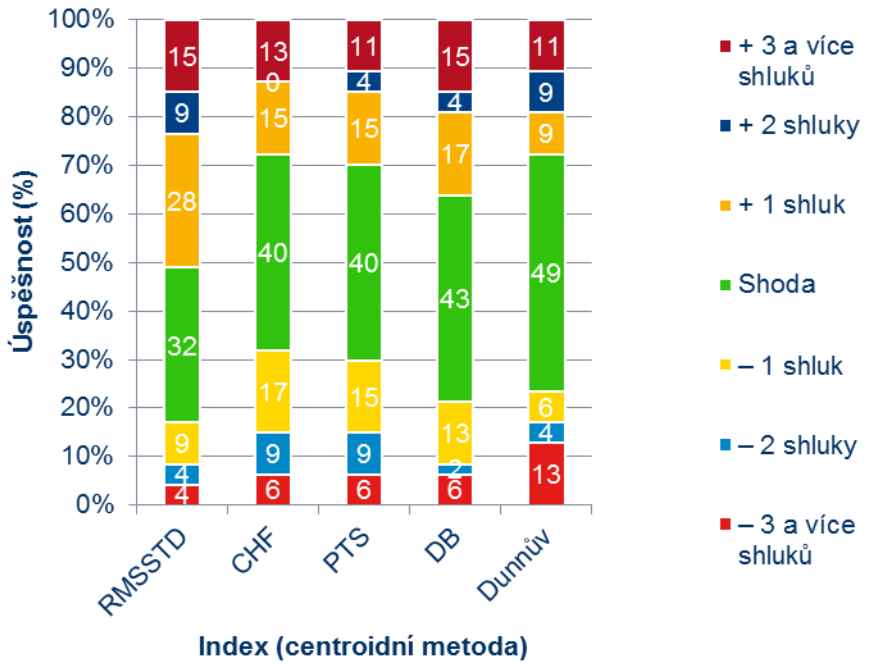
Úspěšnost koeficientů

- Analyzováno celkem 18 různých souborů z UCI (Hammerbauer) – úspěšnost koeficientů podle korelací



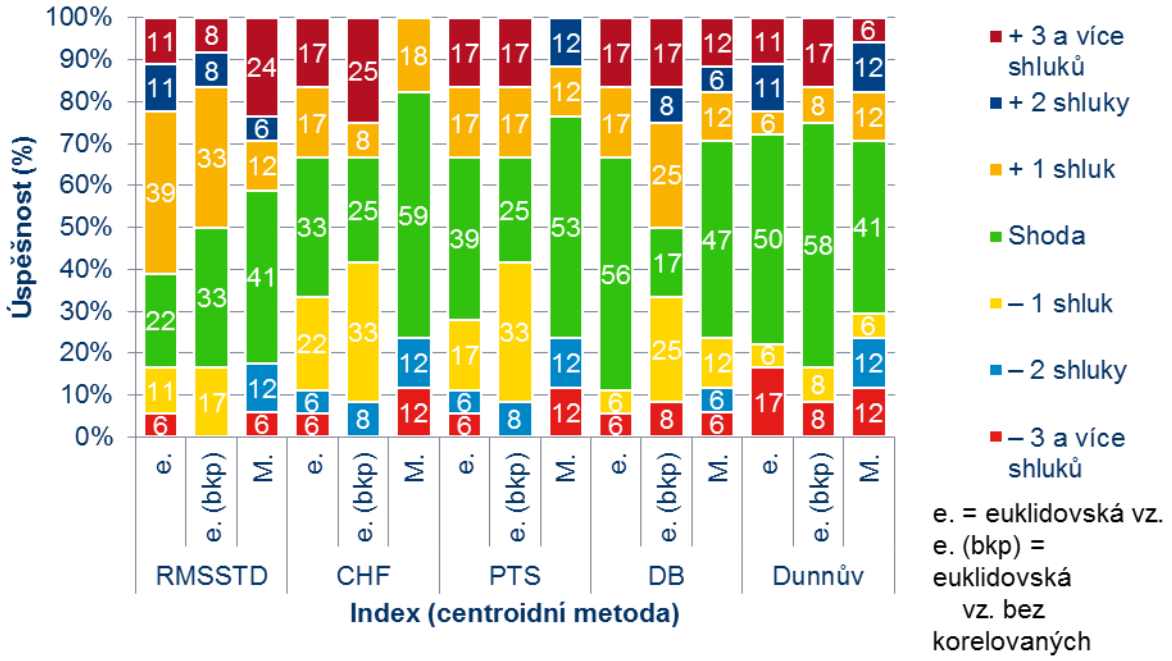
Úspěšnost koeficientů

- Analyzováno celkem 18 různých souborů z UCI (Hammerbauer) – úspěšnost koeficientů podle metody



Úspěšnost koeficientů

- Analyzováno celkem 18 různých souborů z UCI (Hammerbauer) – úspěšnost koeficientů podle metody vše

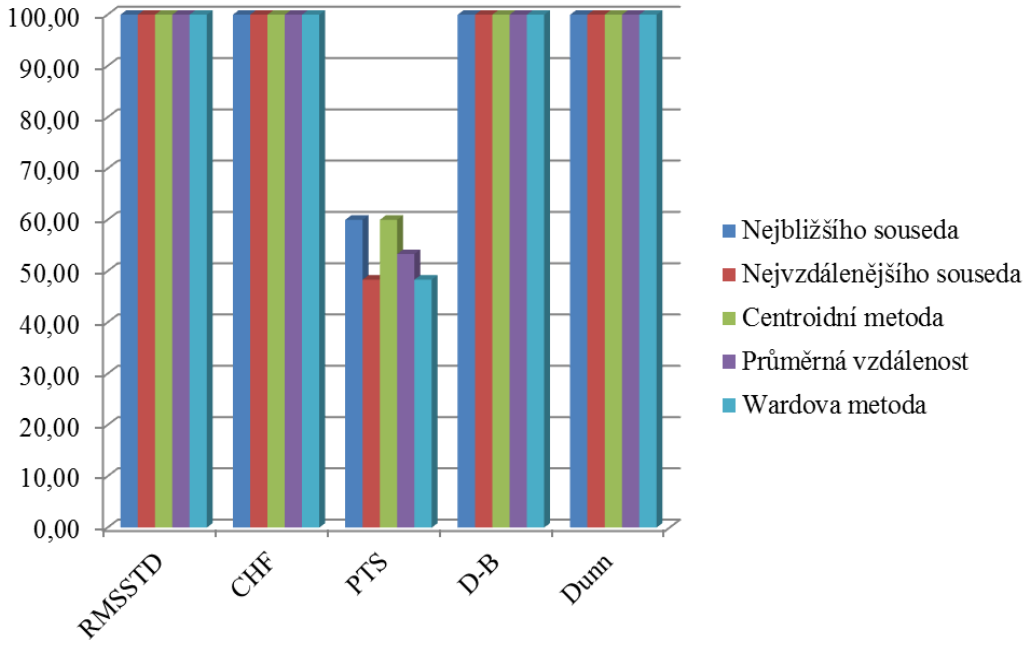


Úspěšnost koeficientů

- Generátor shluků
- Umělé soubory
- Stejné podmínky pro jednotlivé skupiny

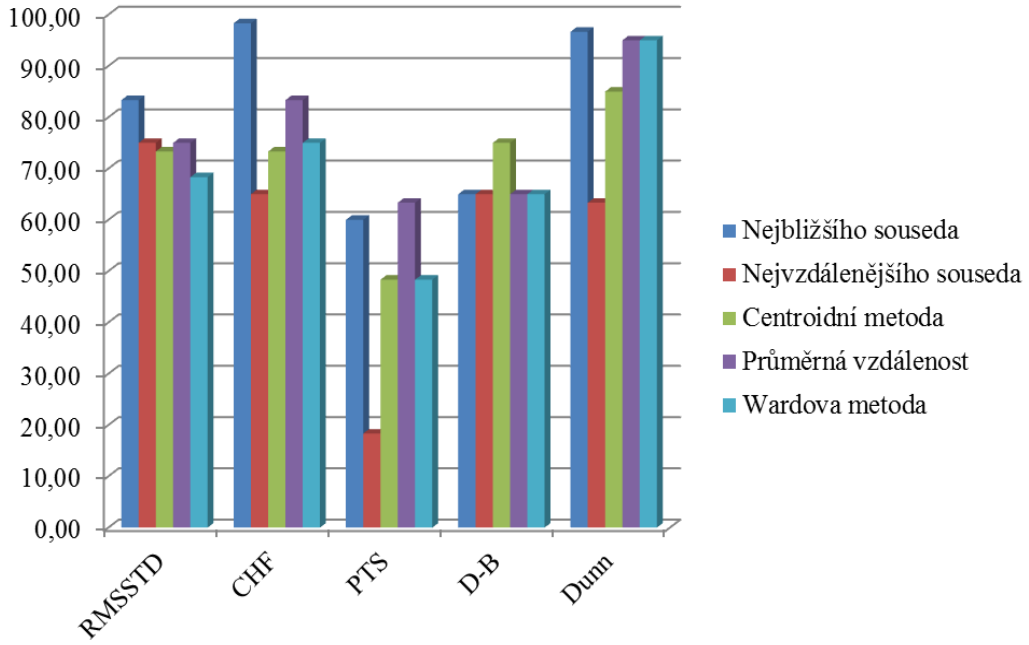
Úspěšnost koeficientů

- Shluky jsou dobře separované; Euklid. vzdálenost
- Analyzováno celkem 60 různých souborů



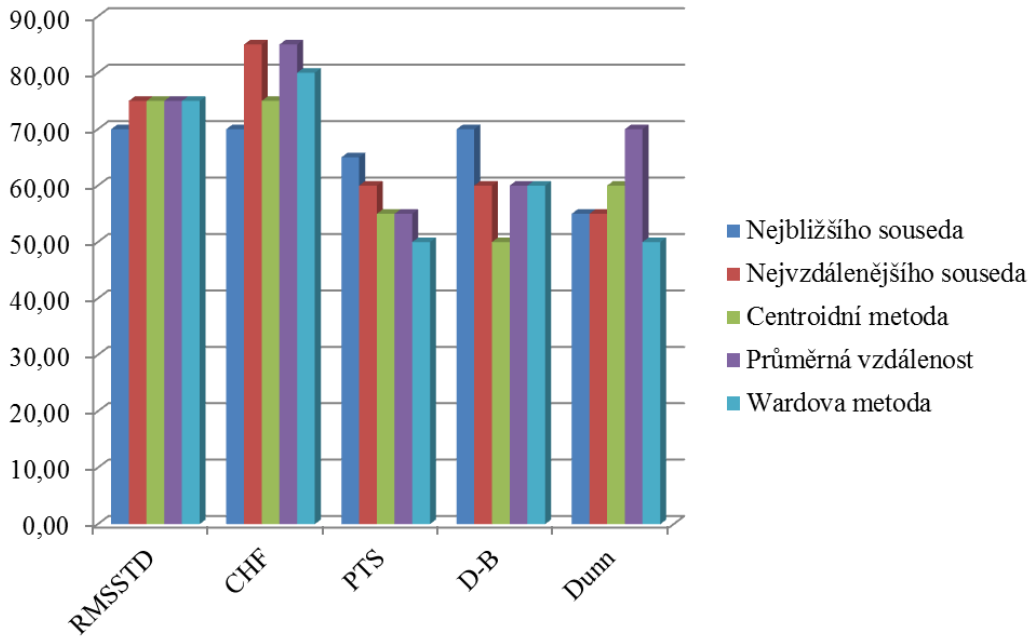
Úspěšnost koeficientů

- Shluky jsou dobře separované; Mahal. vzdálenost
- Analyzováno celkem 60 různých souborů



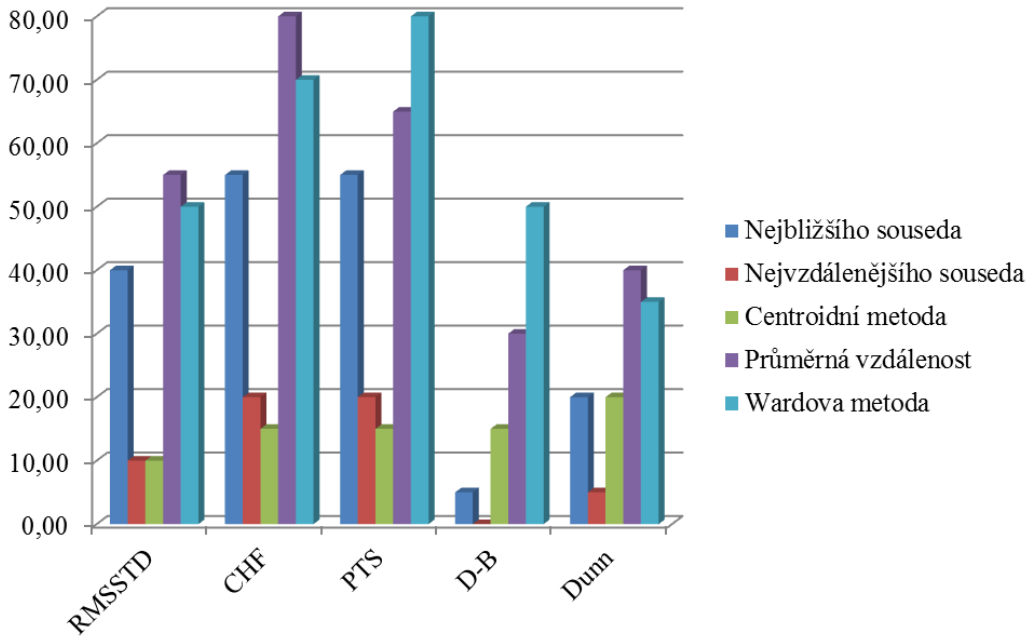
Úspěšnost koeficientů

- Shluky se částečně překrývají; Euklid. vzdálenost
- Analyzováno celkem 20 různých souborů



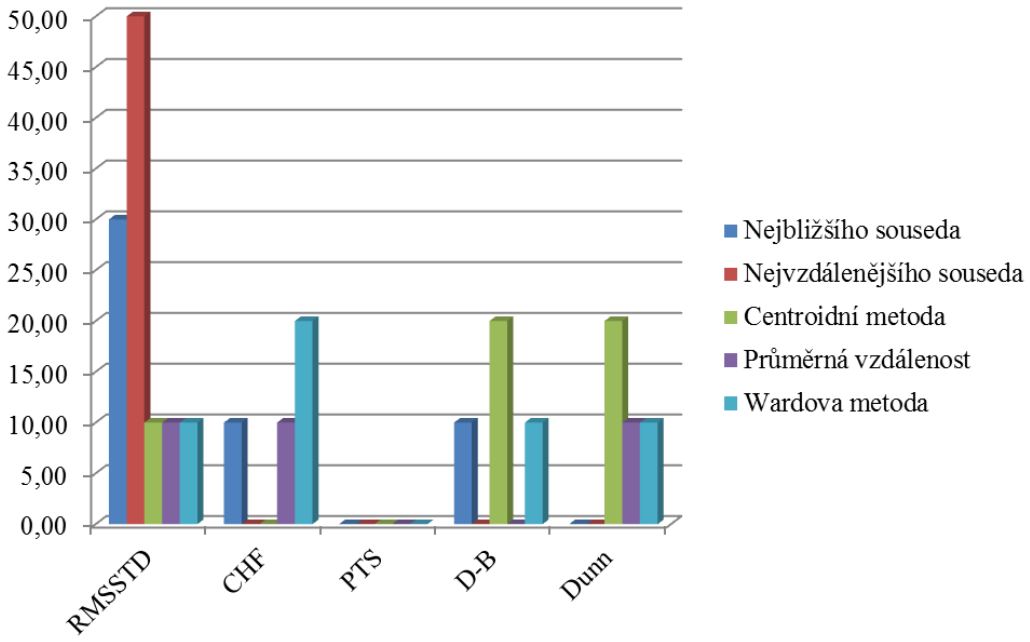
Úspěšnost koeficientů

- Shluky se částečně překrývají; Mahal. vzdálenost
- Analyzováno celkem 20 různých souborů



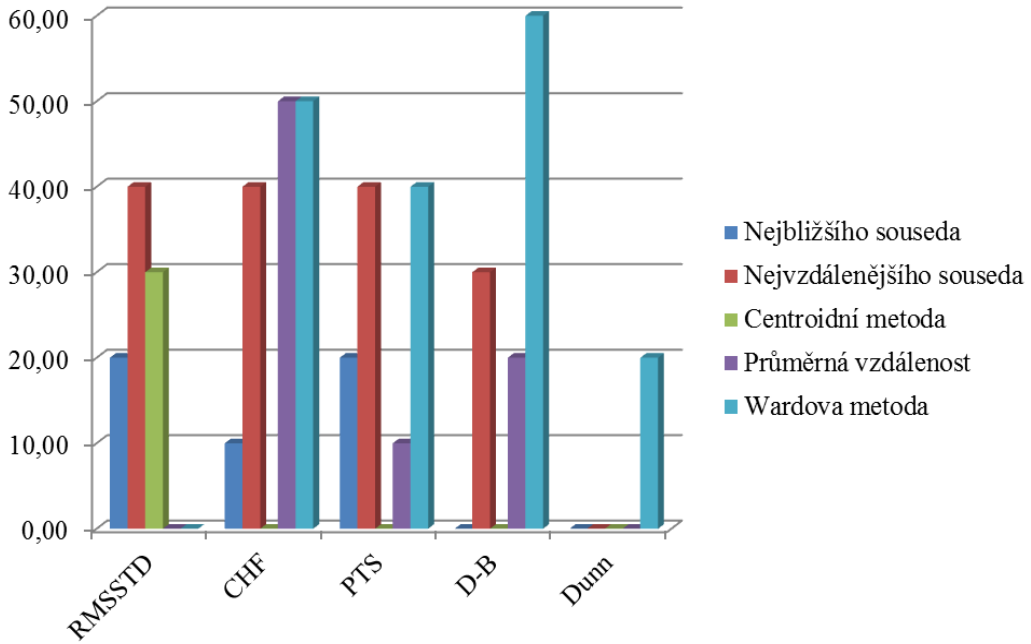
Úspěšnost koeficientů

- Shluky se překrývají; Euklid. vzdálenost
- Analyzováno celkem 10 různých souborů



Úspěšnost koeficientů

- Shluky se překrývají; Mahal. vzdálenost
- Analyzováno celkem 10 různých souborů



Shrnutí

- Různé metody přináší různé výsledky, je třeba hodnotit.
- Stanovení počtu shluků a výsledná klasifikace jsou samostatné úlohy.
- Počet shluků nejen na základě dendrogramu.
- Na základě vlastních zkušeností je „nejúspěšnější“ CHF koeficient.
- Kromě situace značně překrytých shluků je vhodnější Euklidovská vzdálenost.
- Probíhající výzkum: modifikace CHF koeficientu

Kontakt

Tomáš Löster, Ing., Ph. D.
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
tomas.loster@vse.cz

Děkuji za pozornost ☺