Stanovení optimálního počtu shluků ve shlukové analýze

Autor: Tomáš Löster
Vysoká škola ekonomická v Praze

Košice, červen 2016

Osnova prezentace

- Shluková analýza a její význam
- Hlavní cíle příspěvku
- Shrnutí a přínosy

Shluková analýza a její význam

- jedna z technik vícerozměrné analýzy dat
- základní cíl analýzy 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**

Hlavní cíle příspěvku

- Stručný popis „běžně“ implementovaných metod a metrik
- Stručný popis „běžně“ implementovaných koeficientů pro hodnocení v SW
- Ilustrace použití koeficientů
- Shrnutí úspěšnosti koeficientů

Vybrané metody shlukování

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

Vybrané míry vzdálenosti

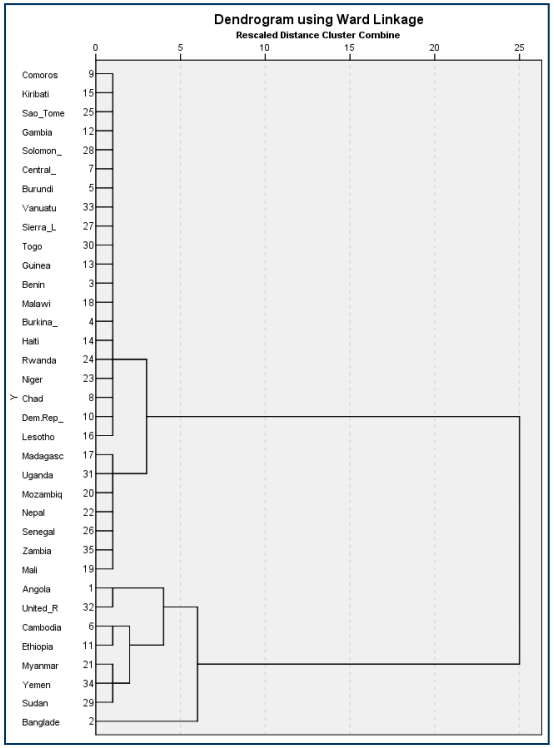
- Euklidovská míra vzdálenosti
- Mahalanobisova míra vzdálenosti

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

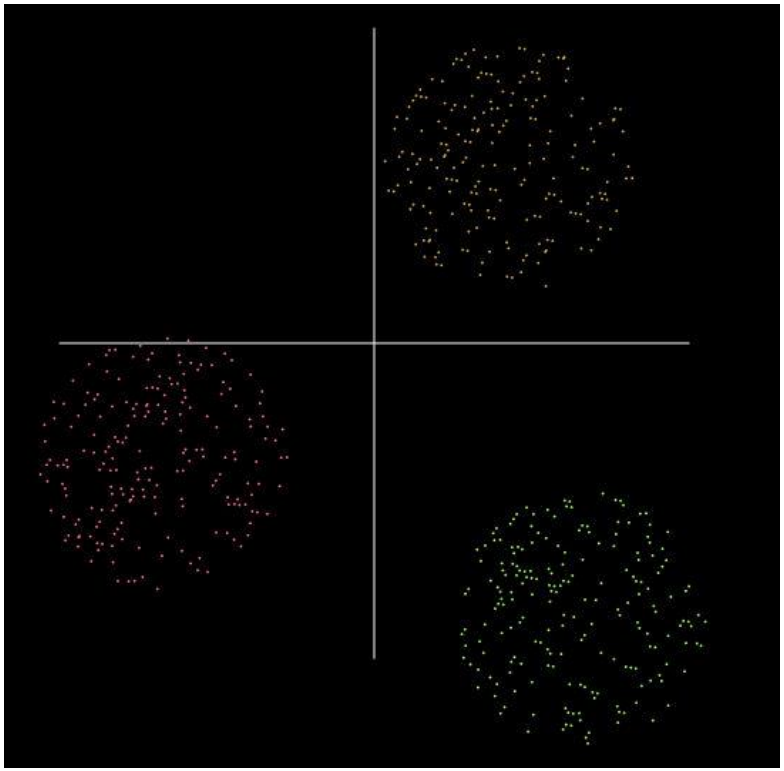
Ilustrace vyhodnocení počtu shluků

- dendrogram

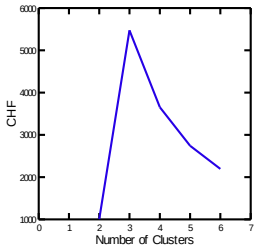


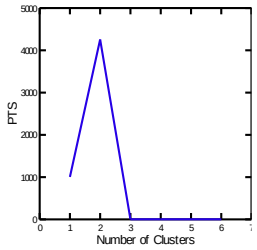
Ilustrace vyhodnocení počtu shluků

- 3 dobře separované shluky

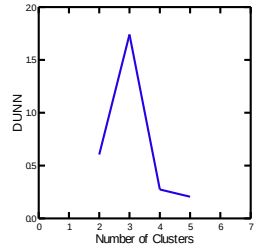






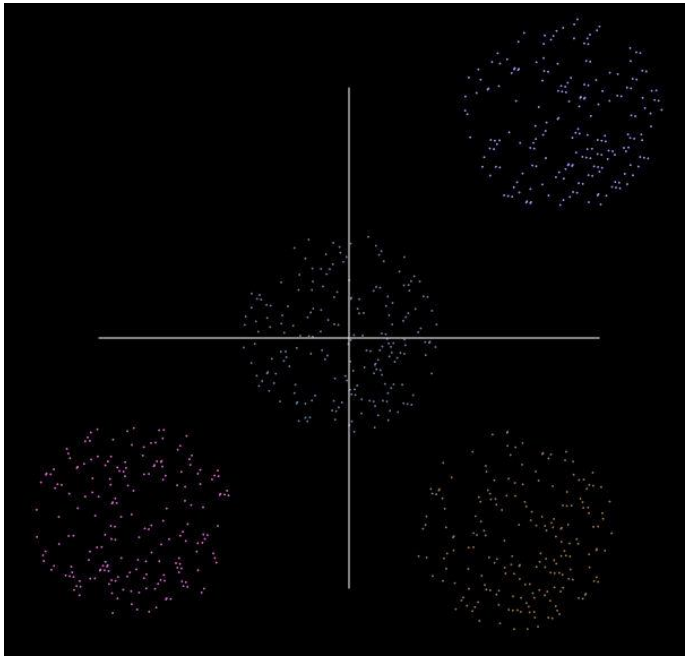




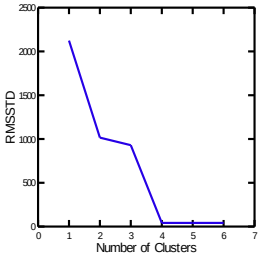


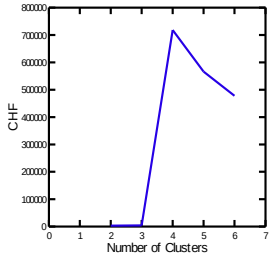
Ilustrace vyhodnocení počtu shluků

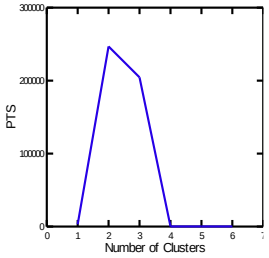
- 4 dobře separované shluky

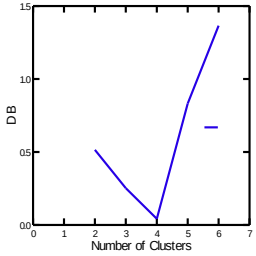


Validity Index Plot





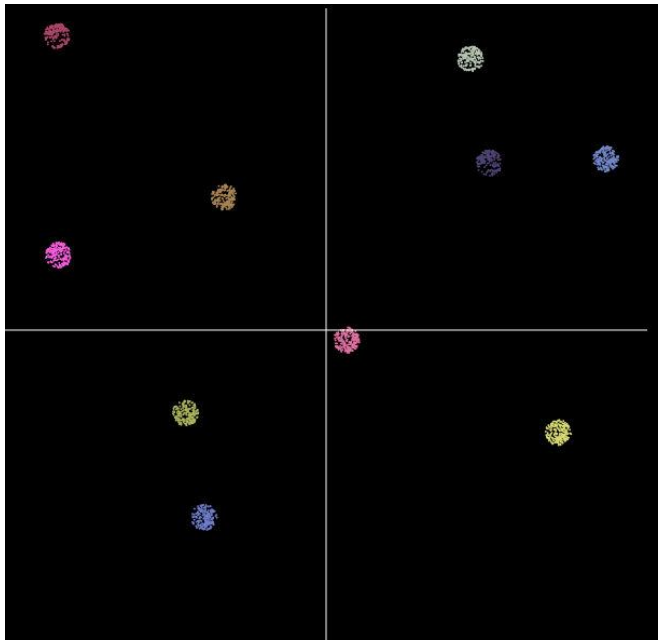


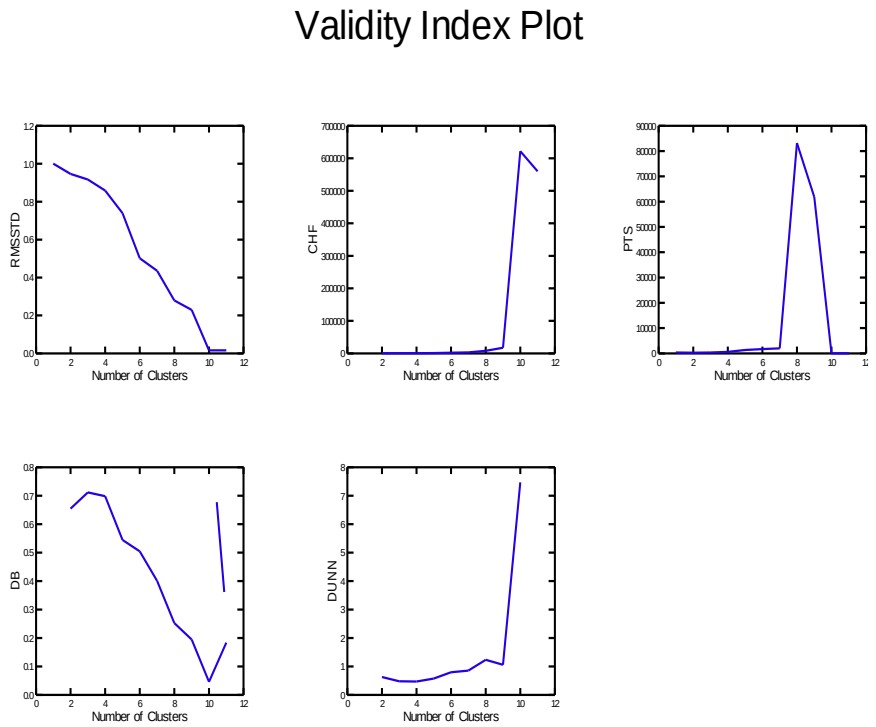




Ilustrace vyhodnocení počtu shluků

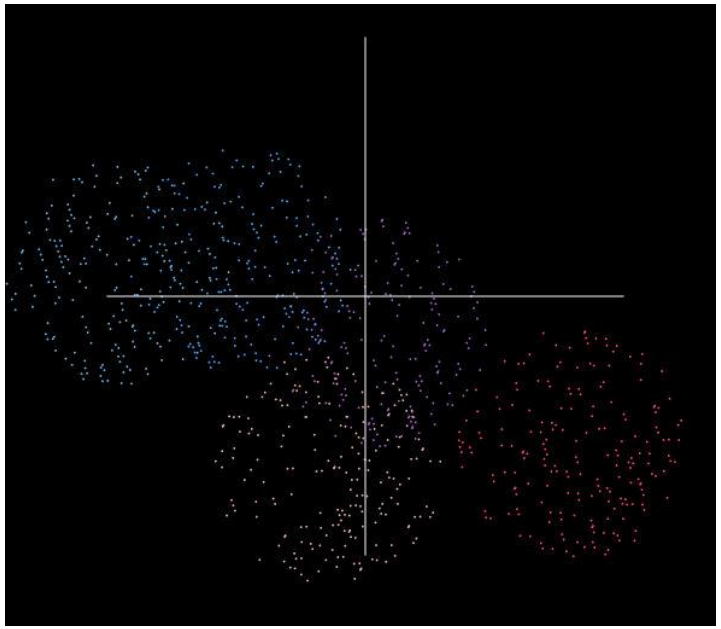
- 10 dobře separovaných shluků

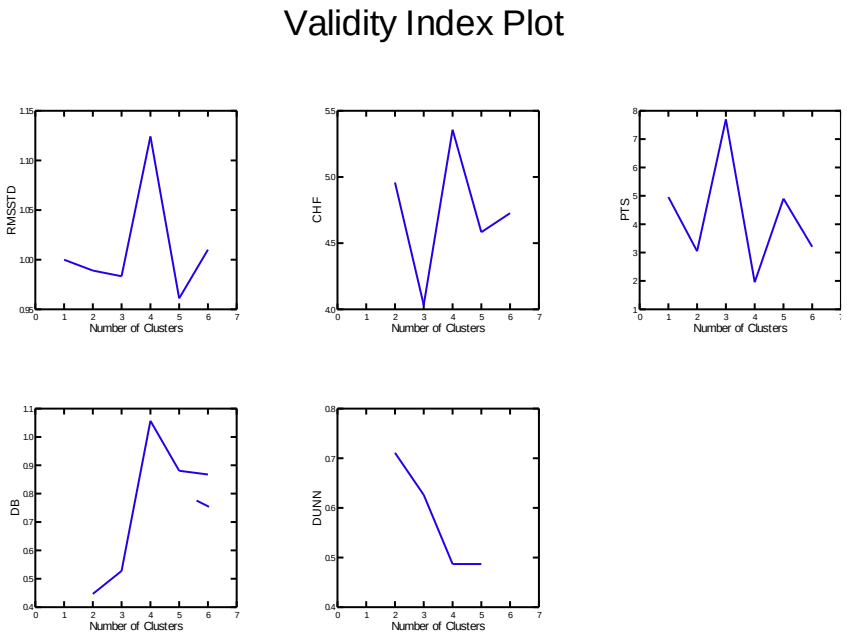




Ilustrace vyhodnocení počtu shluků

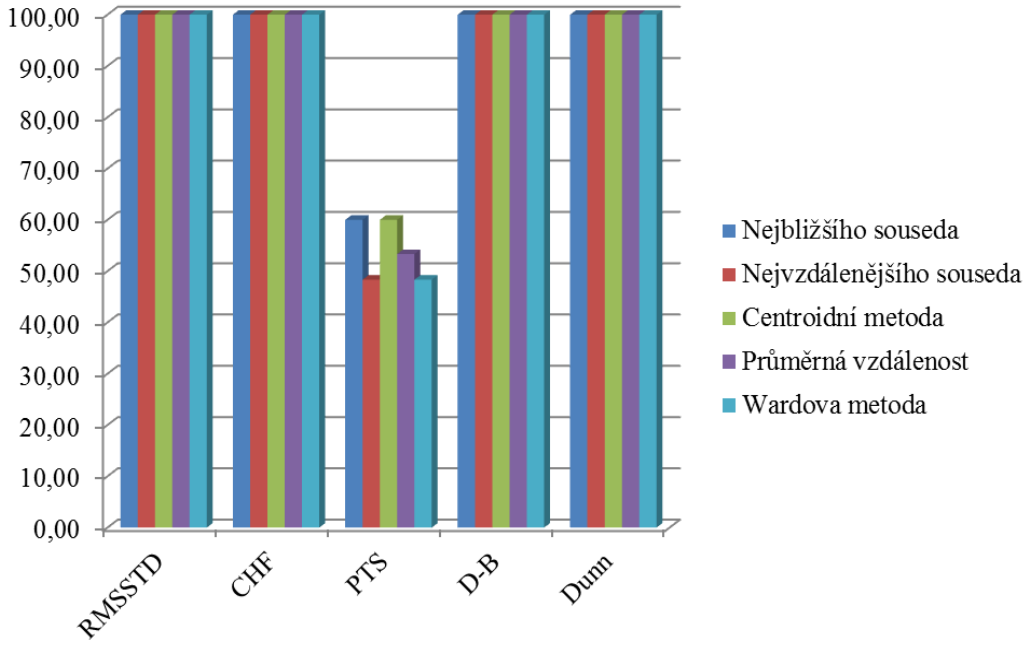
- 5 překrývajících se shluků





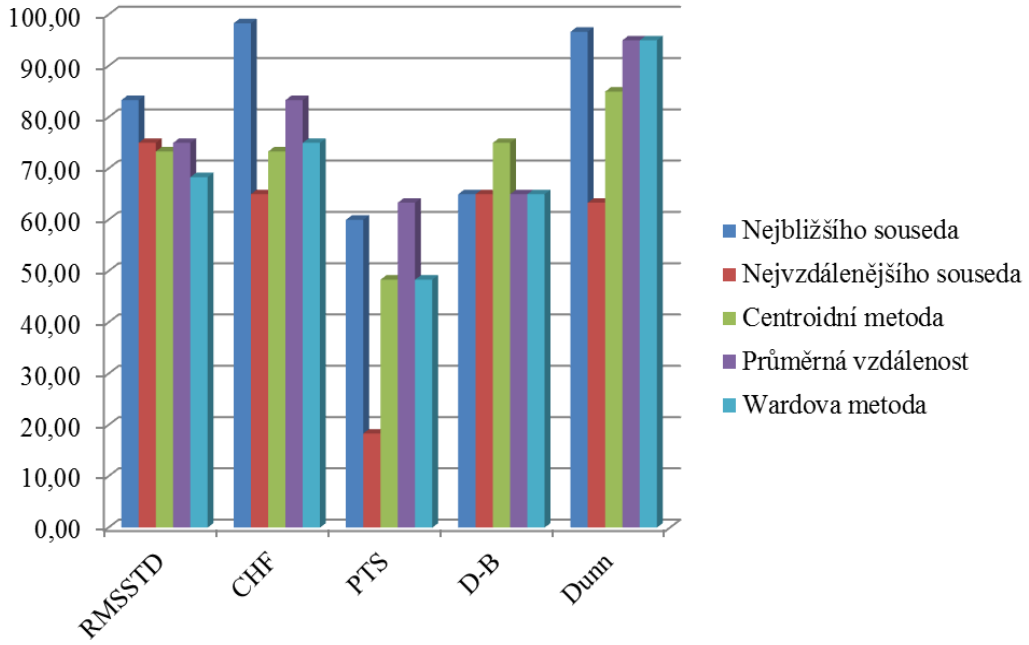
Úspěšnost koeficientů

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



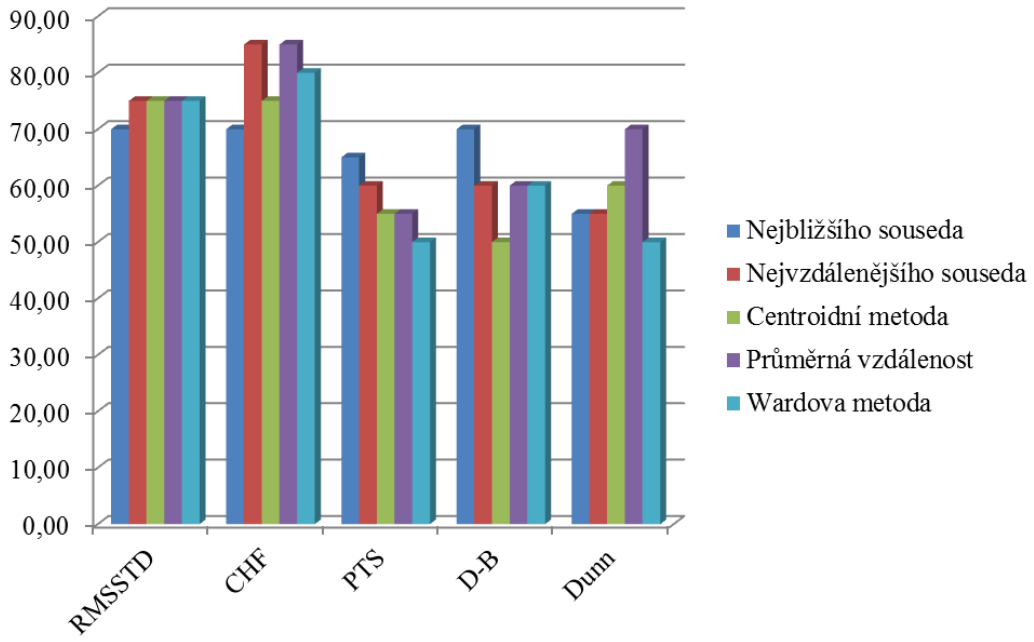
Úspěšnost koeficientů

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



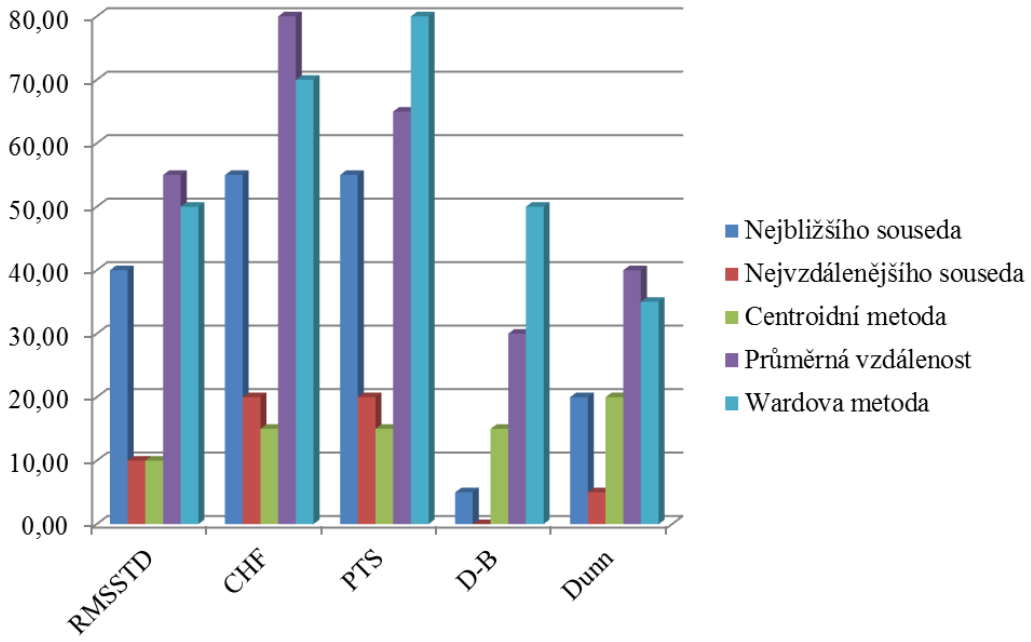
Úspěšnost koeficientů

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



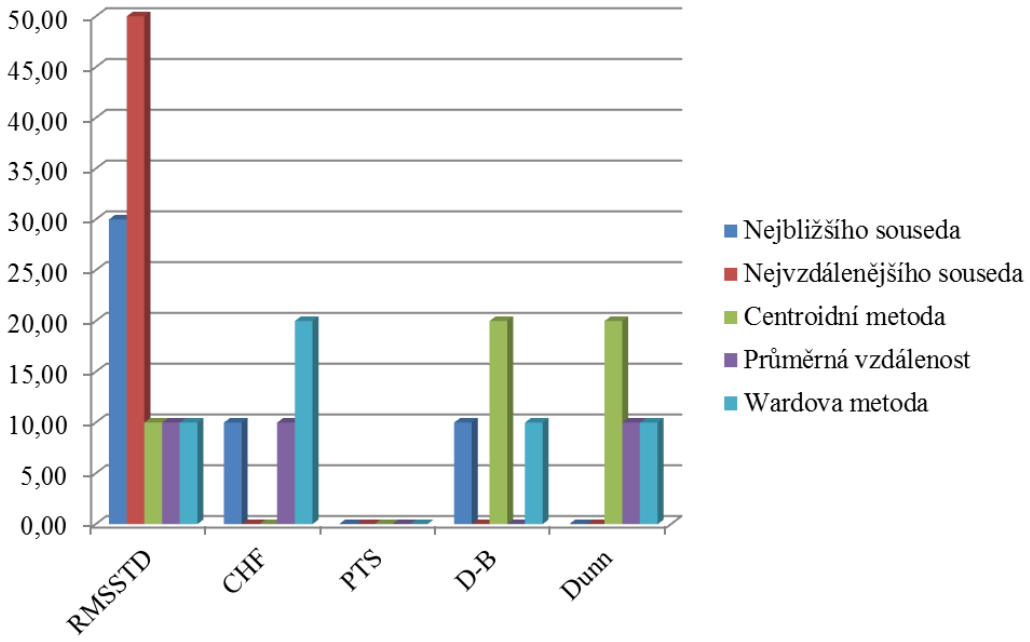
Úspěšnost koeficientů

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



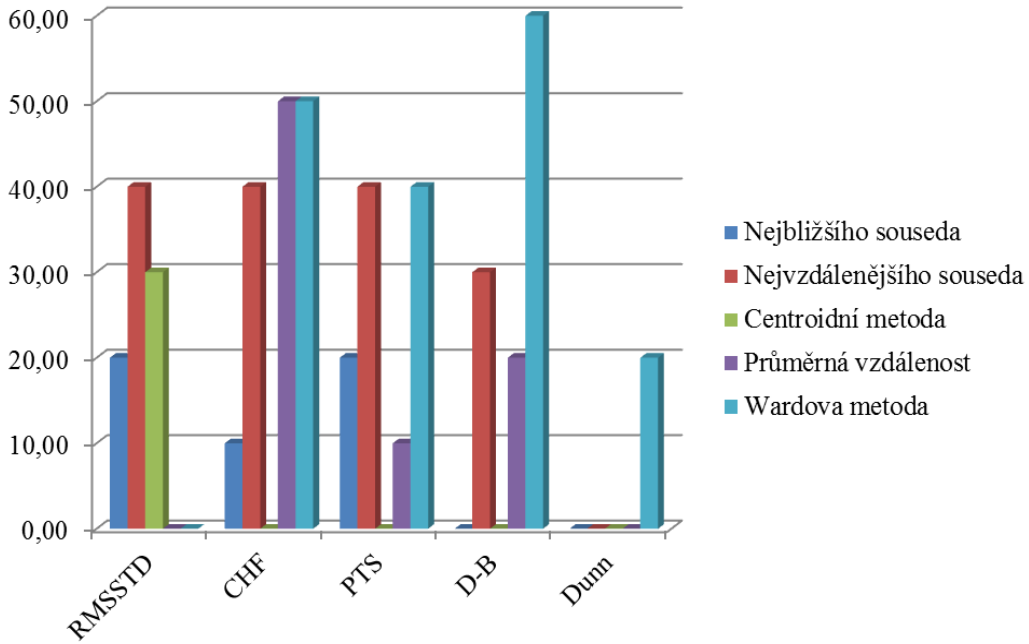
Úspěšnost koeficientů

- 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.
- Počet shluků nejen na základě dendrogramu.
- Na základě vlastních zkušeností je „nejúspěšnější“ CHF koeficient.
- 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 ☺