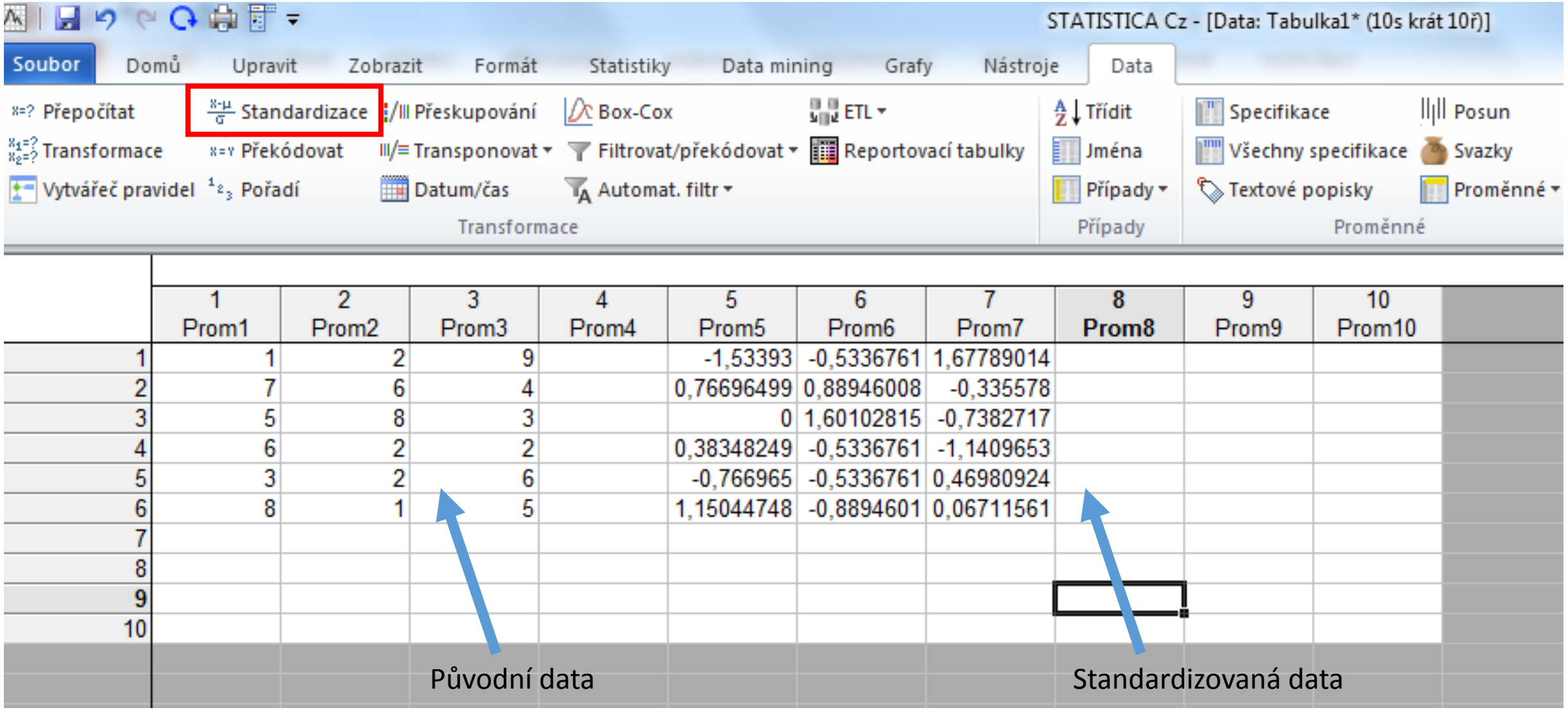
Shluková a PCA analýza

Je dobré (většinou nutné) před vlastní analýzou data standardizovat



	1 Prom1	2 Prom2	3 Prom3	4 Prom4	5 Prom5	6 Prom6	7 Prom7	8 Prom8	9 Prom9	10 Prom10
1	1	2	9		-1,53393	-0,5336761	1,67789014			
2	7	6	4		0,76696499	0,88946008	-0,335578			
3	5	8	3		0	1,60102815	-0,7382717			
4	6	2	2		0,38348249	-0,5336761	-1,1409653			
5	3	2	6		-0,766965	-0,5336761	0,46980924			
6	8	1	5		1,15044748	-0,8894601	0,06711561			
7										
8										
9										
10										

Původní data

Standardizovaná data

Shluková analýza

- Hierarchické shlukování:

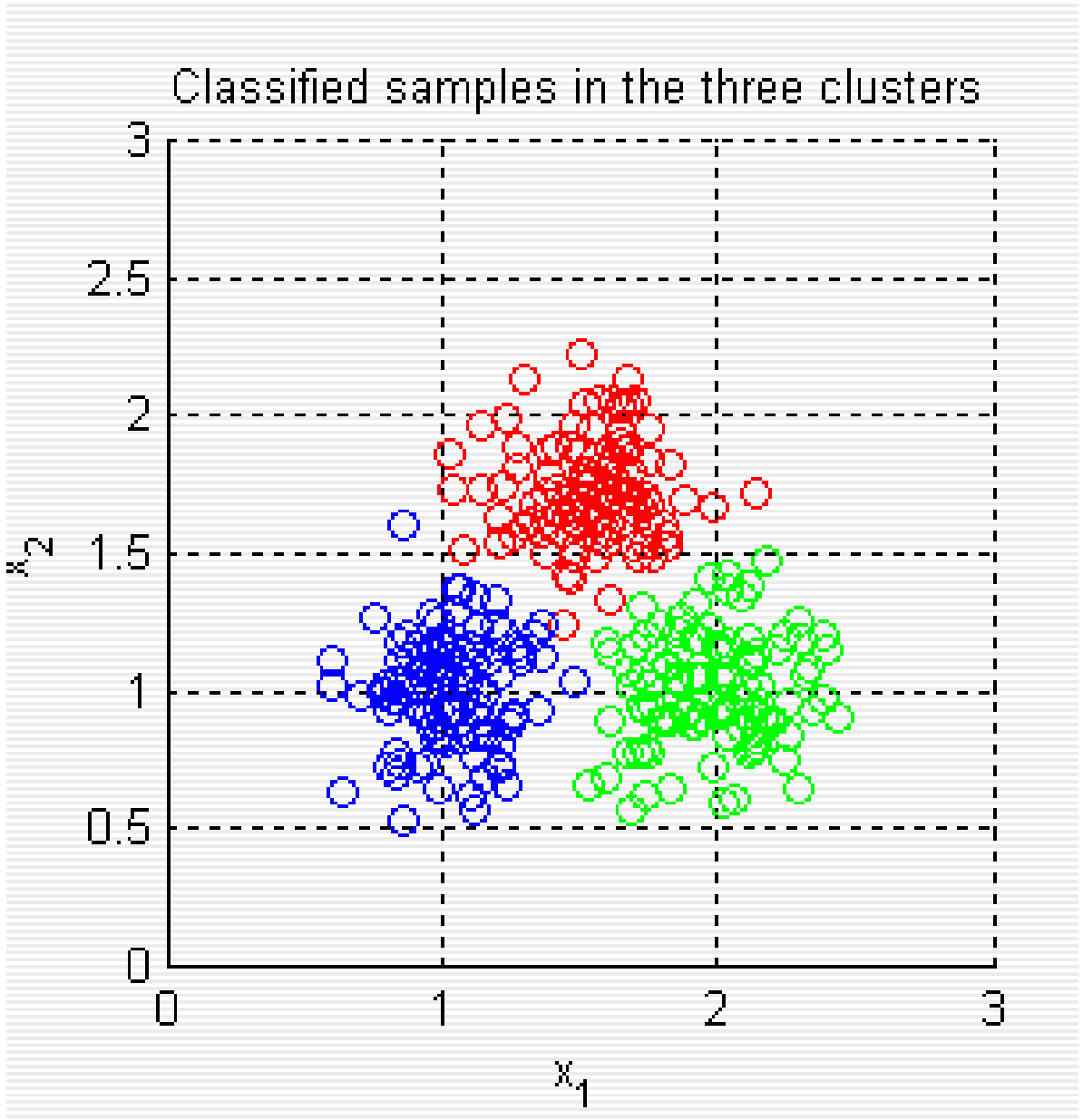
1. Metoda nejbližšího souseda (nebezpečí řetězového efektu)
2. Metoda nejvzdálenějšího souseda (spojení dvou nejbližších shluků měřeno na nejvzdálenějších členech tohoto shluku)
3. Metoda průměrné vzdálenosti
4. Wardova metoda

- Nehierarchické:

Např. metoda zárodečných bodů

- Míru vhodnosti vybrané metody prozradí korelační koeficient CC (čím blíže 1, tím lepší model)

- Otázka optimálního počtu shluků (neexistuje jednoznačný návod, musíme sami určit na základě výsledků)



Zdroj: http://midas.uamt.feec.vutbr.cz/POV/LPOV_Exercise06/content_cz.php

Převod z dat na vzdálenosti

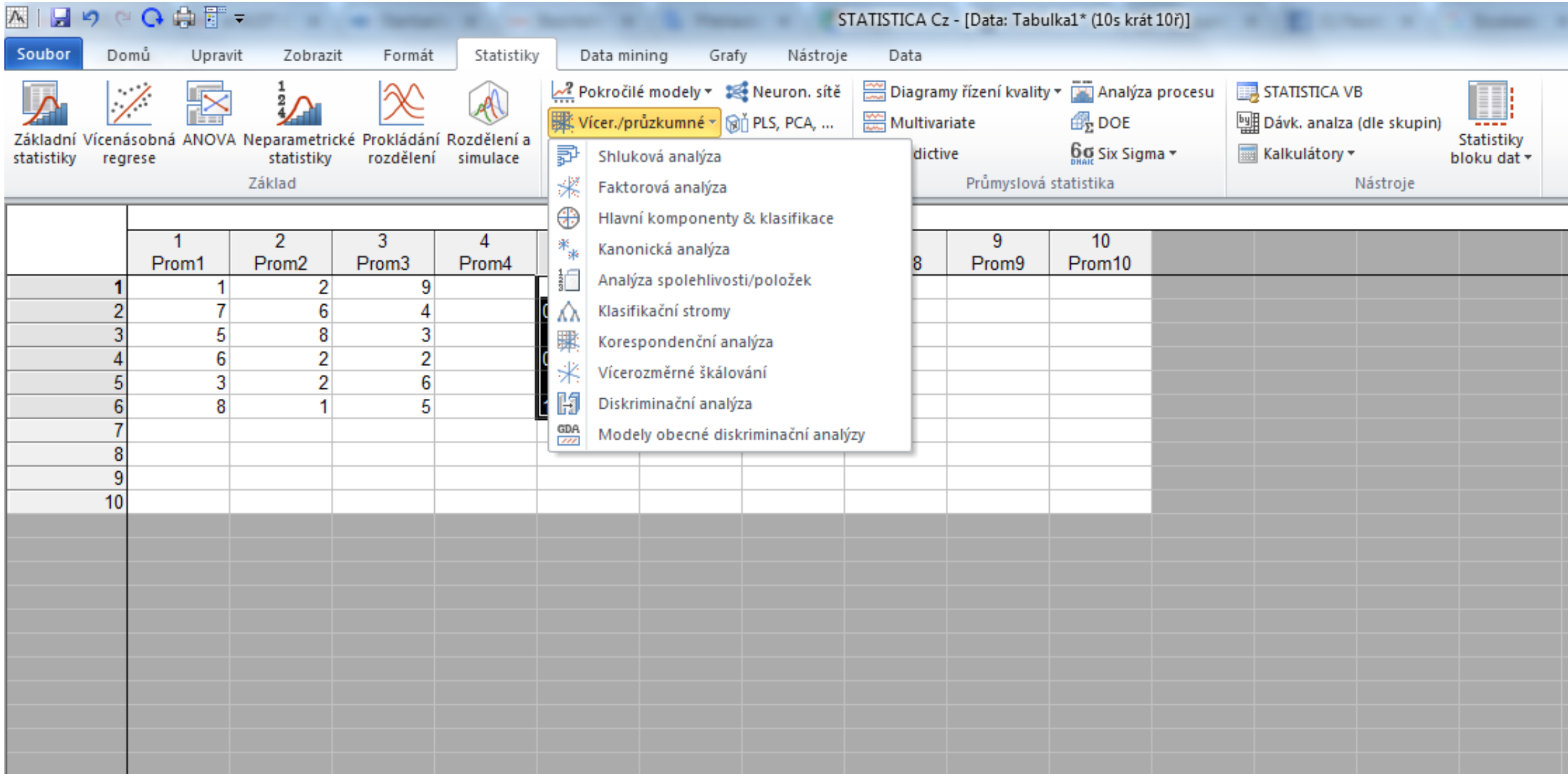
Everitt [88] uvádí data, kdy bylo

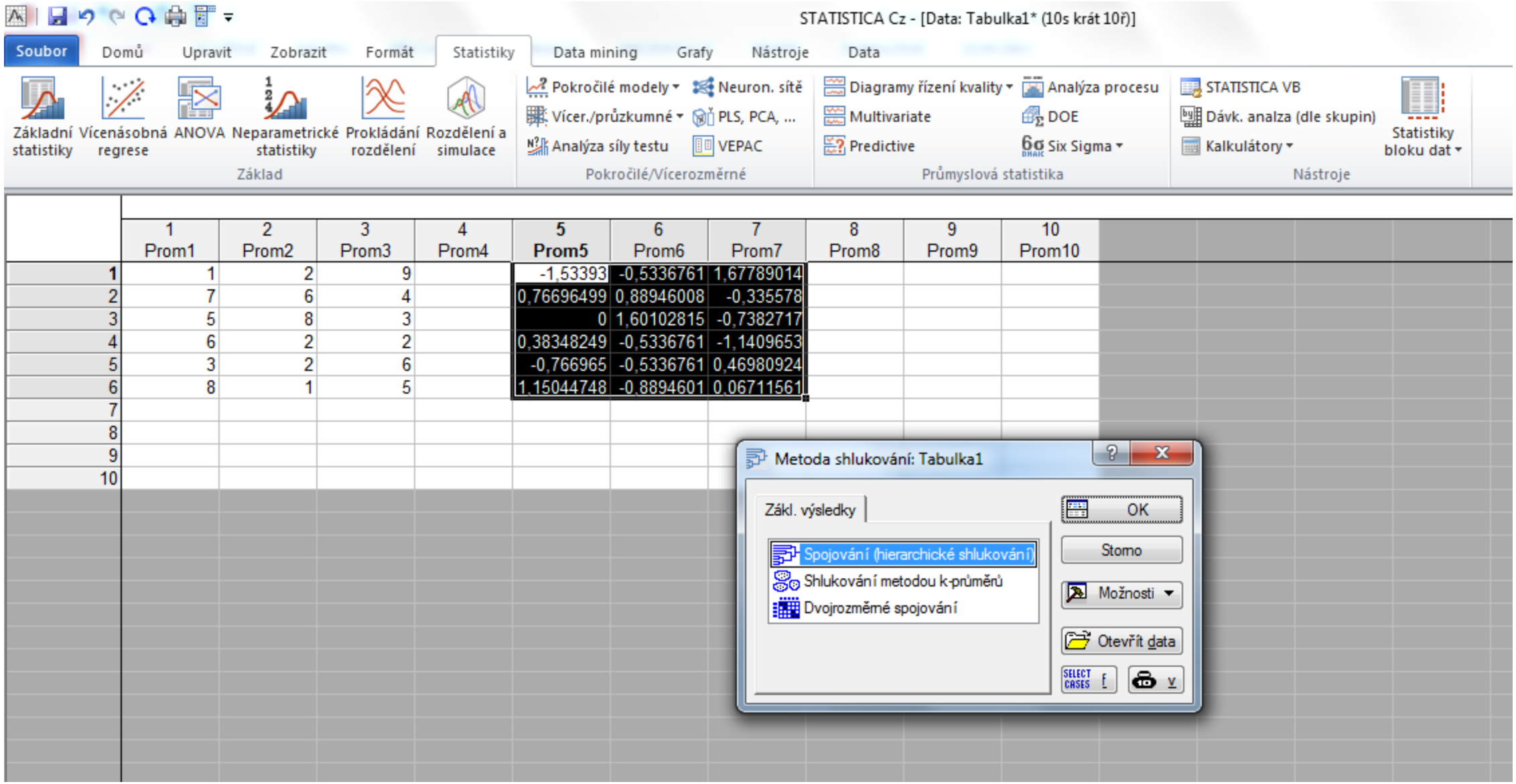
Objekt	První znak	Druhý znak
První (1)	1	1
Druhý (2)	1	2
Třetí (3)	6	3
Čtvrtý (4)	8	2
Pátý (5)	8	0

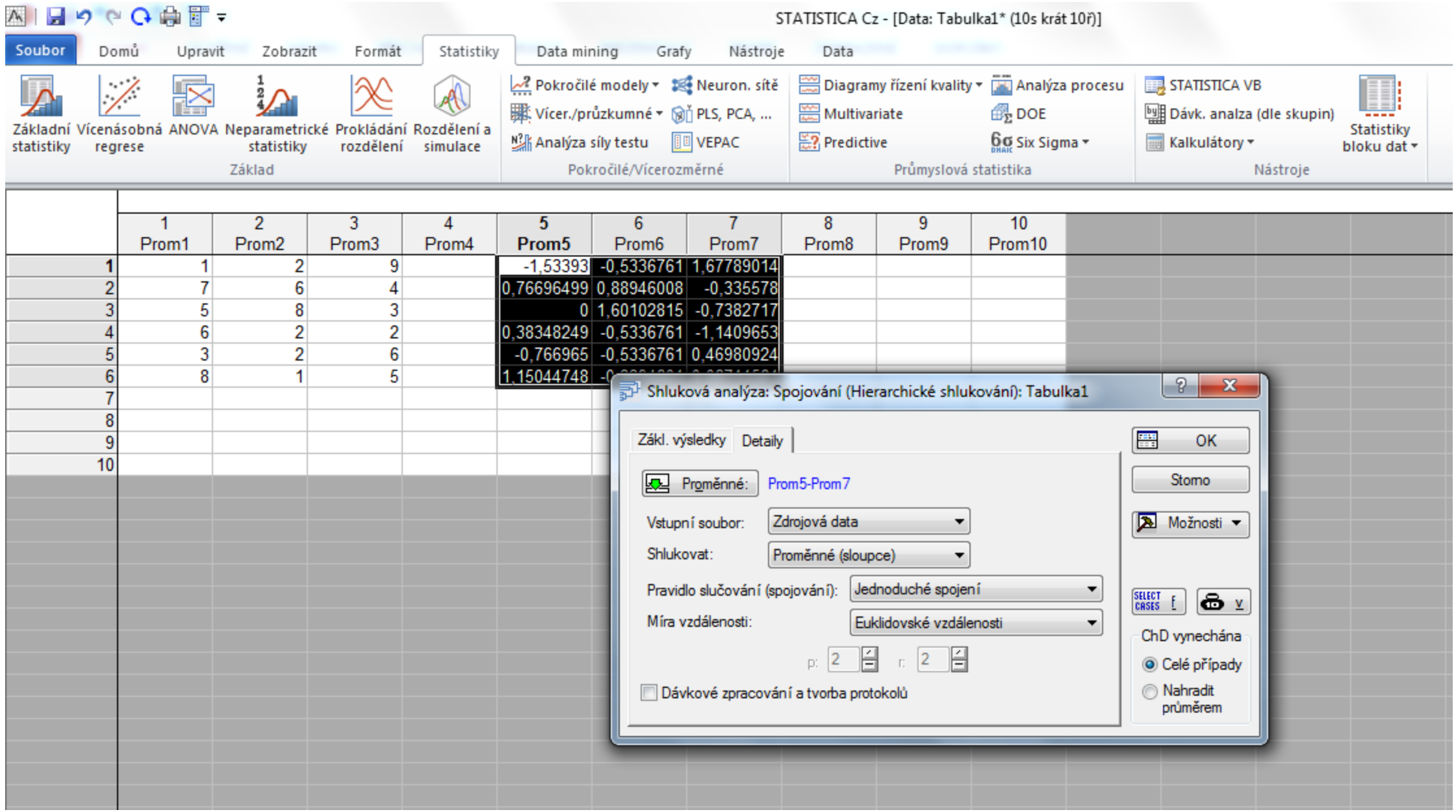
Z tabulky dat se určí matice koeficientů podobnosti $\boldsymbol{E}_1$. Nechť je jako koeficient podobnosti vybrán čtverec eukleidovské vzdálenosti d_E^2. Je zřejmé, že například $d_E^2(1, 3) = (1 - 6)^2 + (1 - 3)^2 = 29$ atd. Matice vzdáleností $\boldsymbol{E}_1$ má pak tvar

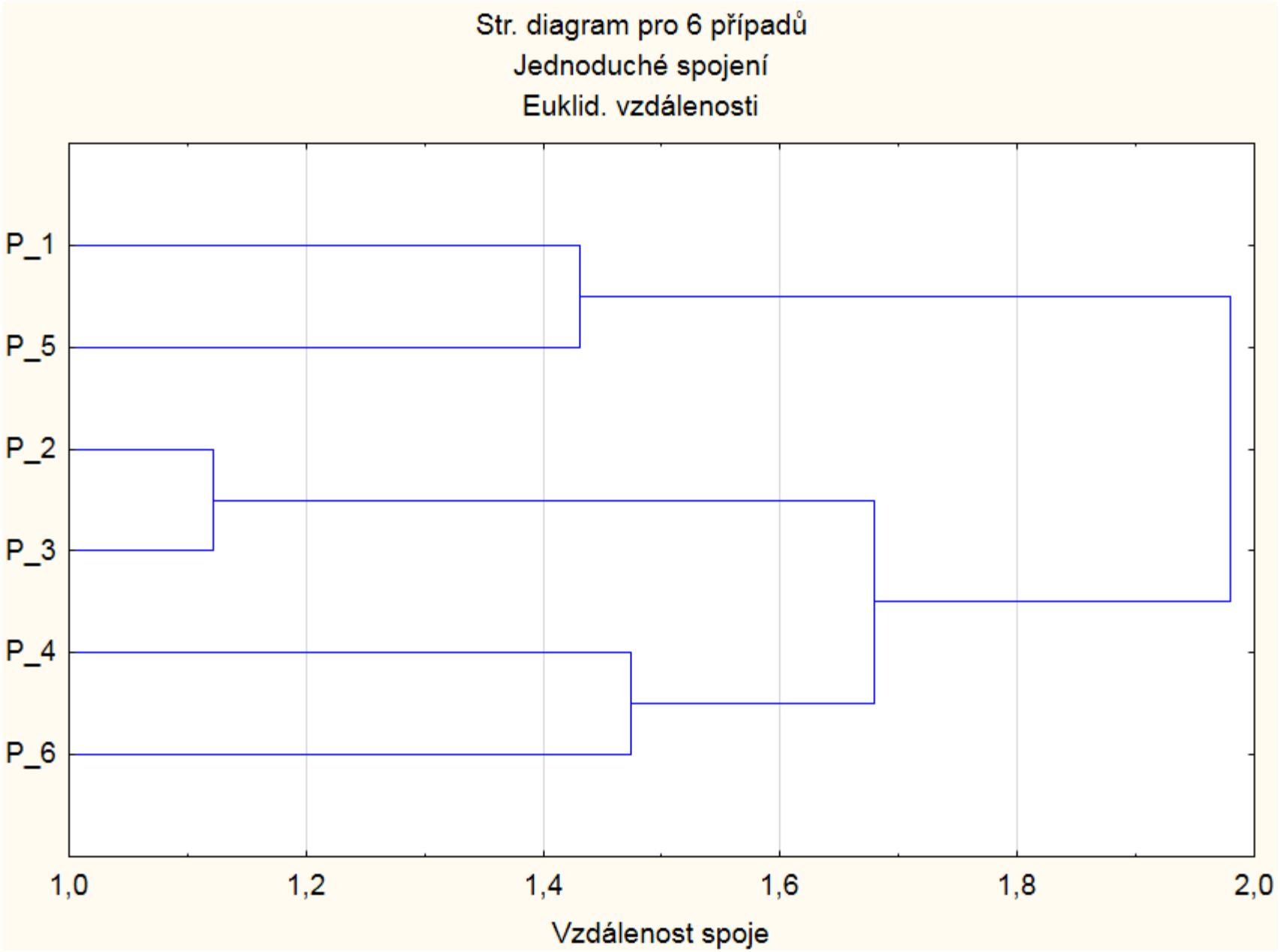
	1	2	3	4	5
1	0	1	29	50	50
2	1	0	26	49	53
3	29	26	0	5	13
4	50	49	5	0	4
5	50	53	13	4	0

Zdroj: Meloun a kol. Statistická analýza vícerozměrných dat v příkladech


STATISTICA Cz - [Data: Tabulka1* (10s krát 10ř)]
Soubor
Domů
Upravit
Zobrazit
Formát
Statistiky
Data mining
Grafy
Nástroje
Data
Základní statistiky
Vícenásobná regrese
ANOVA
Neparametrické statistiky
Prokládání rozdělení
Rozdělení a simulace
Základ
Pokročilé modely
Neuron. sítě
Vícer./průzkumné
PLS, PCA, ...
Shluková analýza
Faktorová analýza
Hlavní komponenty & klasifikace
Kanonická analýza
Analýza spolehlivosti/položek
Klasifikační stromy
Korespondenční analýza
Vícerozměrné škálování
Diskriminační analýza
Modely obecné diskriminační analýzy
Diagramy řízení kvality
Multivariate
dictive
Analýza procesu
DOE
Six Sigma
Průmyslová statistika
STATISTICA VB
Dávk. analza (dle skupin)
Kalkulátory
Statistiky bloku dat
Nástroje
1 Prom1
2 Prom2
3 Prom3
4 Prom4
9 Prom9
10 Prom10
1 1 2 9
2 7 6 4
3 5 8 3
4 6 2 2
5 3 2 6
6 8 1 5
7
8
9
10

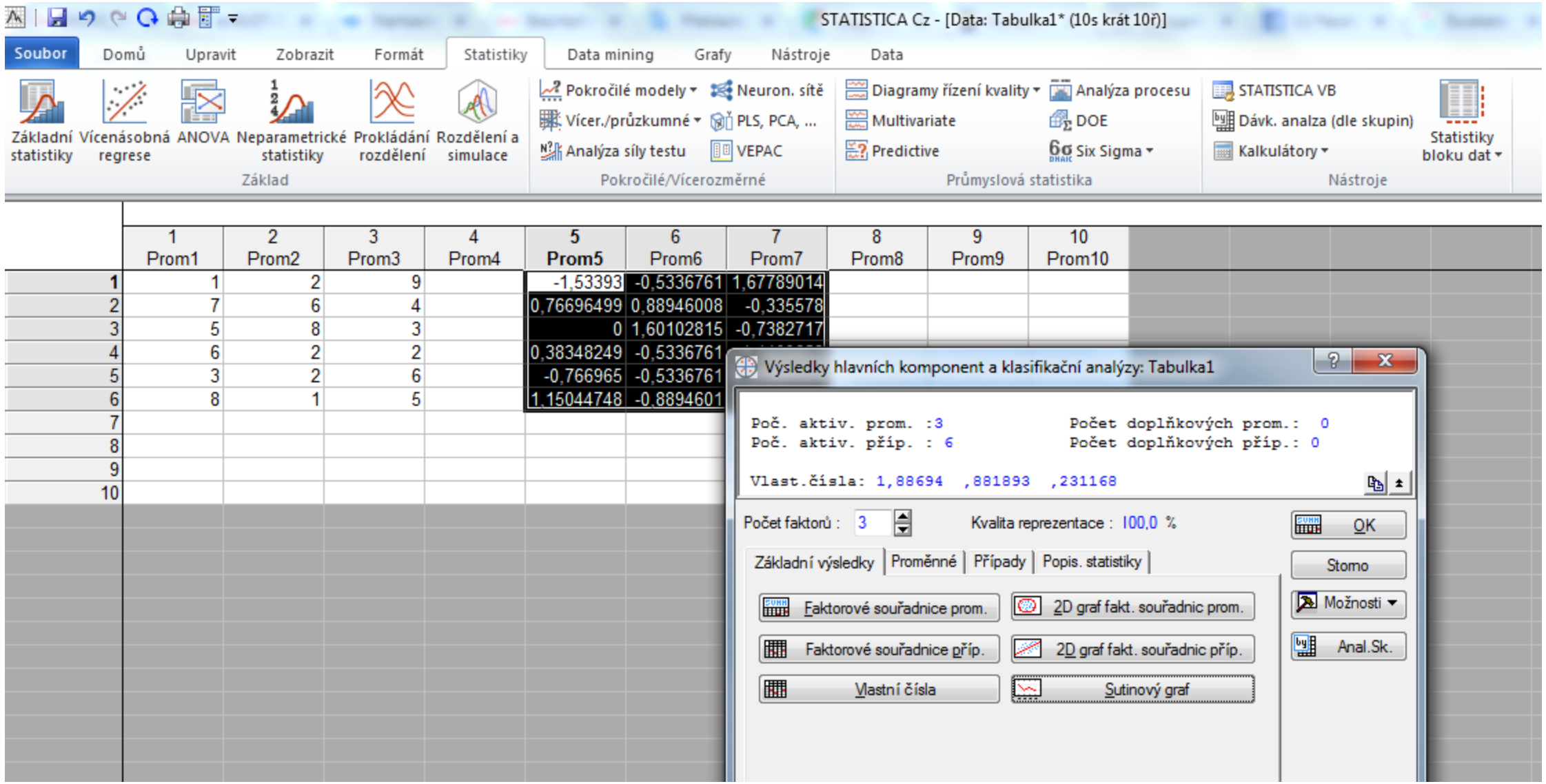

STATISTICA Cz - [Data: Tabulka1* (10s krát 10ř)]
Soubor
Domů
Upravit
Zobrazit
Formát
Statistiky
Data mining
Grafy
Nástroje
Data
Základní statistiky
Vícenásobná regrese
ANOVA
Neparametrické statistiky
Prokládání rozdělení
Rozdělení a simulace
Základ
Pokročilé modely
Neuron. sítě
Vícer./průzkumné
PLS, PCA, ...
Analýza síly testu
VEPAC
Pokročilé/Vícerozměrné
Diagramy řízení kvality
Analýza procesu
Multivariate
DOE
Predictive
Six Sigma
Průmyslová statistika
STATISTICA VB
Dávk. analza (dle skupin)
Kalkulátory
Nástroje
Statistiky bloku dat
1 Prom1
2 Prom2
3 Prom3
4 Prom4
5 Prom5
6 Prom6
7 Prom7
8 Prom8
9 Prom9
10 Prom10
1 1 2 9 -1,53393 -0,5336761 1,67789014
2 7 6 4 0,76696499 0,88946008 -0,335578
3 5 8 3 0 1,60102815 -0,7382717
4 6 2 2 0,38348249 -0,5336761 -1,1409653
5 3 2 6 -0,766965 -0,5336761 0,46980924
6 8 1 5 1,15044748 -0,8894601 0,06711561
7
8
9
10
Metoda shlukování: Tabulka1
Zákl. výsledky
Spojování (hierarchické shlukování)
Shlukování metodou k-průměrů
Dvojrozměrné spojování
OK
Storno
Možnosti
Otevřít data
SELECT CASES
f
v

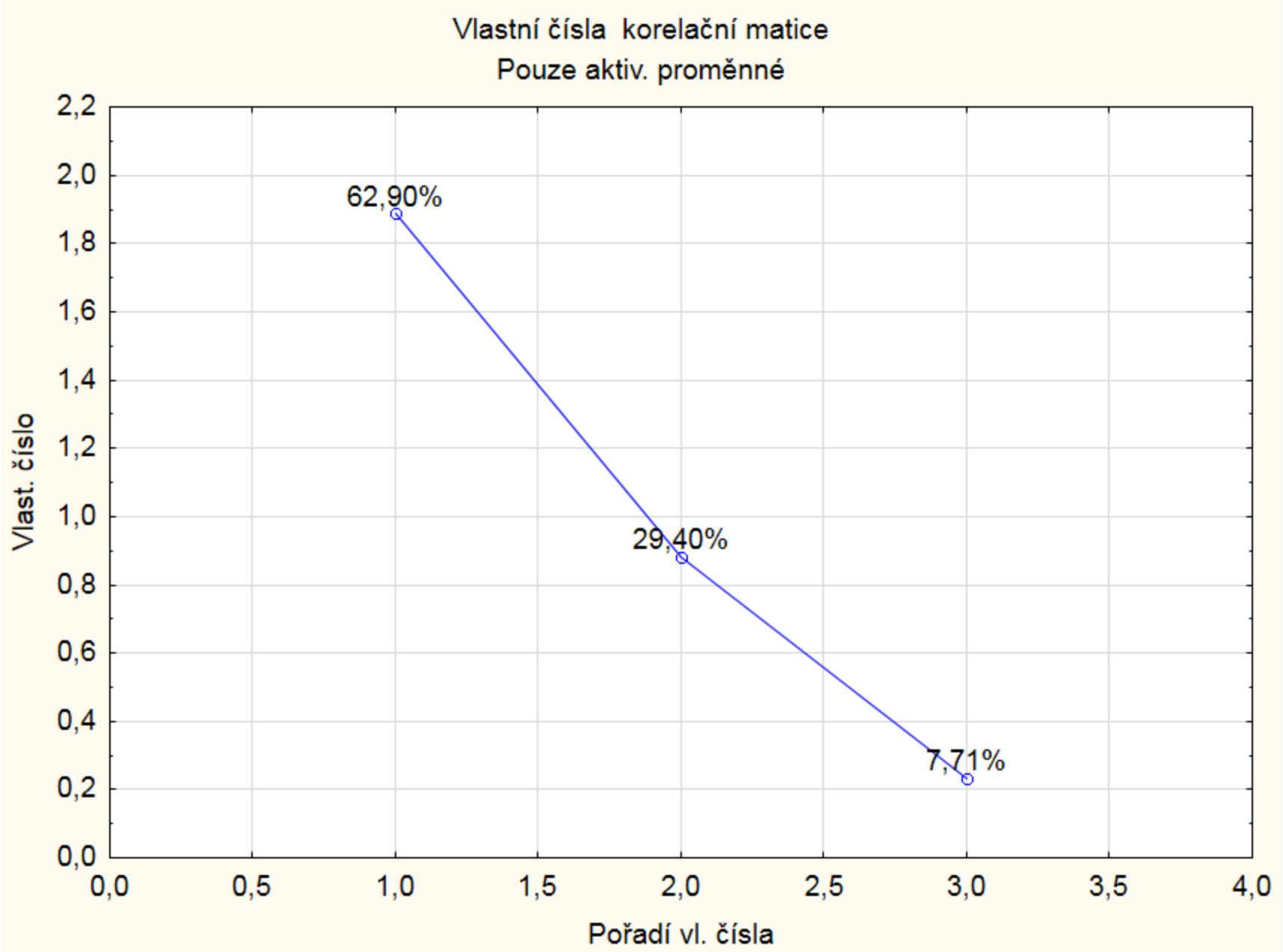

STATISTICA Cz - [Data: Tabulka1* (10s krát 10ř)]
Soubor
Domů
Upravit
Zobrazit
Formát
Statistiky
Data mining
Grafy
Nástroje
Data
Základní statistiky
Vícenásobná regrese
ANOVA
Neparametrické statistiky
Prokládání rozdělení
Rozdělení a simulace
Základ
Pokročilé modely
Neuron. sítě
Vícer./průzkumné
PLS, PCA, ...
Analýza síly testu
VEPAC
Pokročilé/Vícerozměrné
Diagramy řízení kvality
Analýza procesu
Multivariate
DOE
Predictive
Six Sigma
Průmyslová statistika
STATISTICA VB
Dávk. analza (dle skupin)
Kalkulátory
Statistiky bloku dat
Nástroje
1 Prom1
2 Prom2
3 Prom3
4 Prom4
5 Prom5
6 Prom6
7 Prom7
8 Prom8
9 Prom9
10 Prom10
1 1 2 9 -1,53393 -0,5336761 1,67789014
2 7 6 4 0,76696499 0,88946008 -0,335578
3 5 8 3 0 1,60102815 -0,7382717
4 6 2 2 0,38348249 -0,5336761 -1,1409653
5 3 2 6 -0,766965 -0,5336761 0,46980924
6 8 1 5 1,15044748
7
8
9
10
Shluková analýza: Spojování (Hierarchické shlukování): Tabulka1
Zákl. výsledky
Detaily
Proměnné:
Prom5-Prom7
Vstupní soubor:
Zdrojová data
Shlukovat:
Proměnné (sloupce)
Pravidlo slučování (spojování):
Jednoduché spojení
Míra vzdálenosti:
Euklidovské vzdálenosti
p: 2
r: 2
Dávkové zpracování a tvorba protokolů
OK
Storno
Možnosti
SELECT CASES
ChD vynechána
Celé případy
Nahradit průměrem

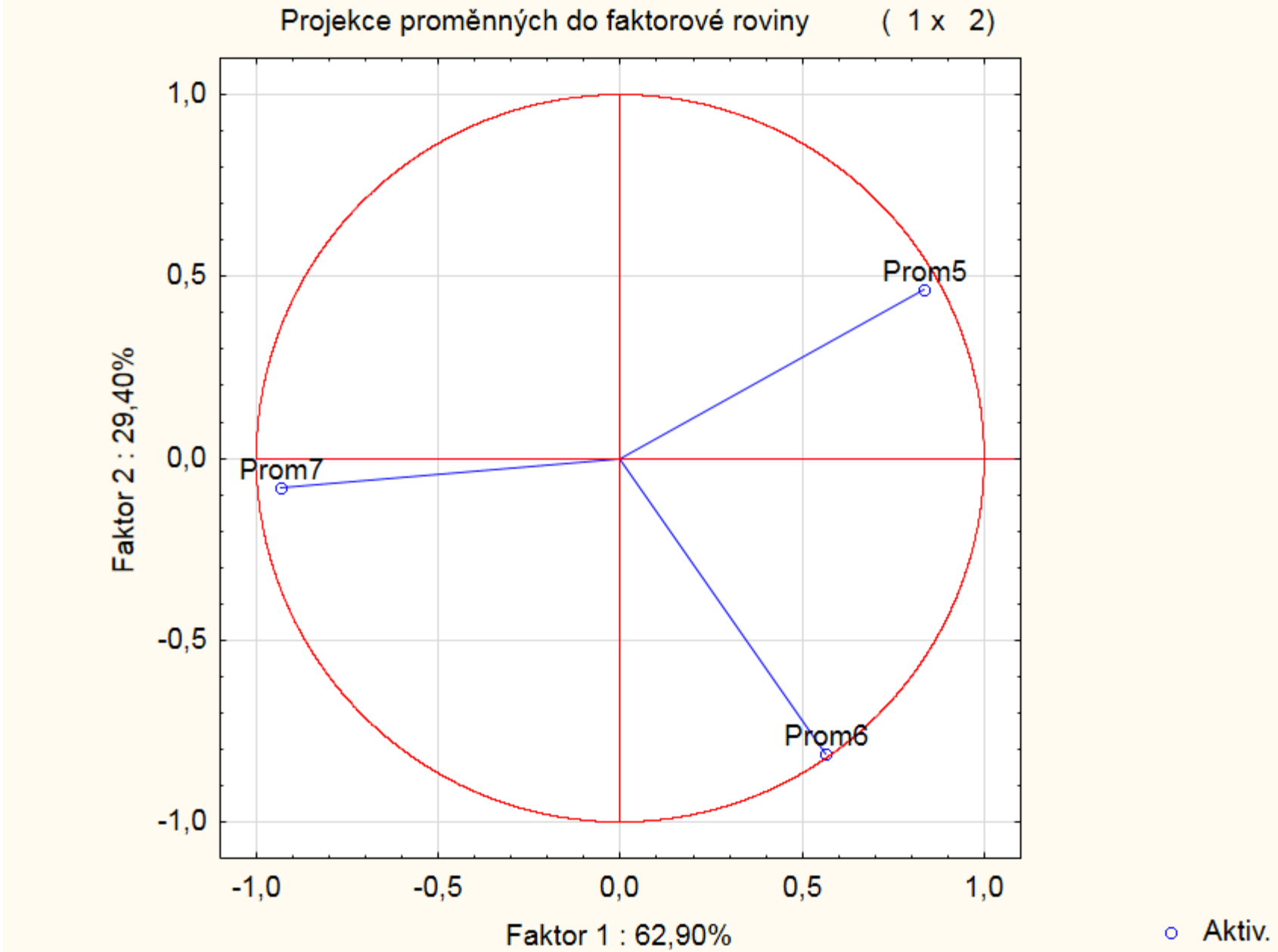

Str. diagram pro 6 případů
Jednoduché spojení
Euklid. vzdálenosti
P_1
P_5
P_2
P_3
P_4
P_6
1,0
1,2
1,4
1,6
1,8
2,0
Vzdálenost spoje

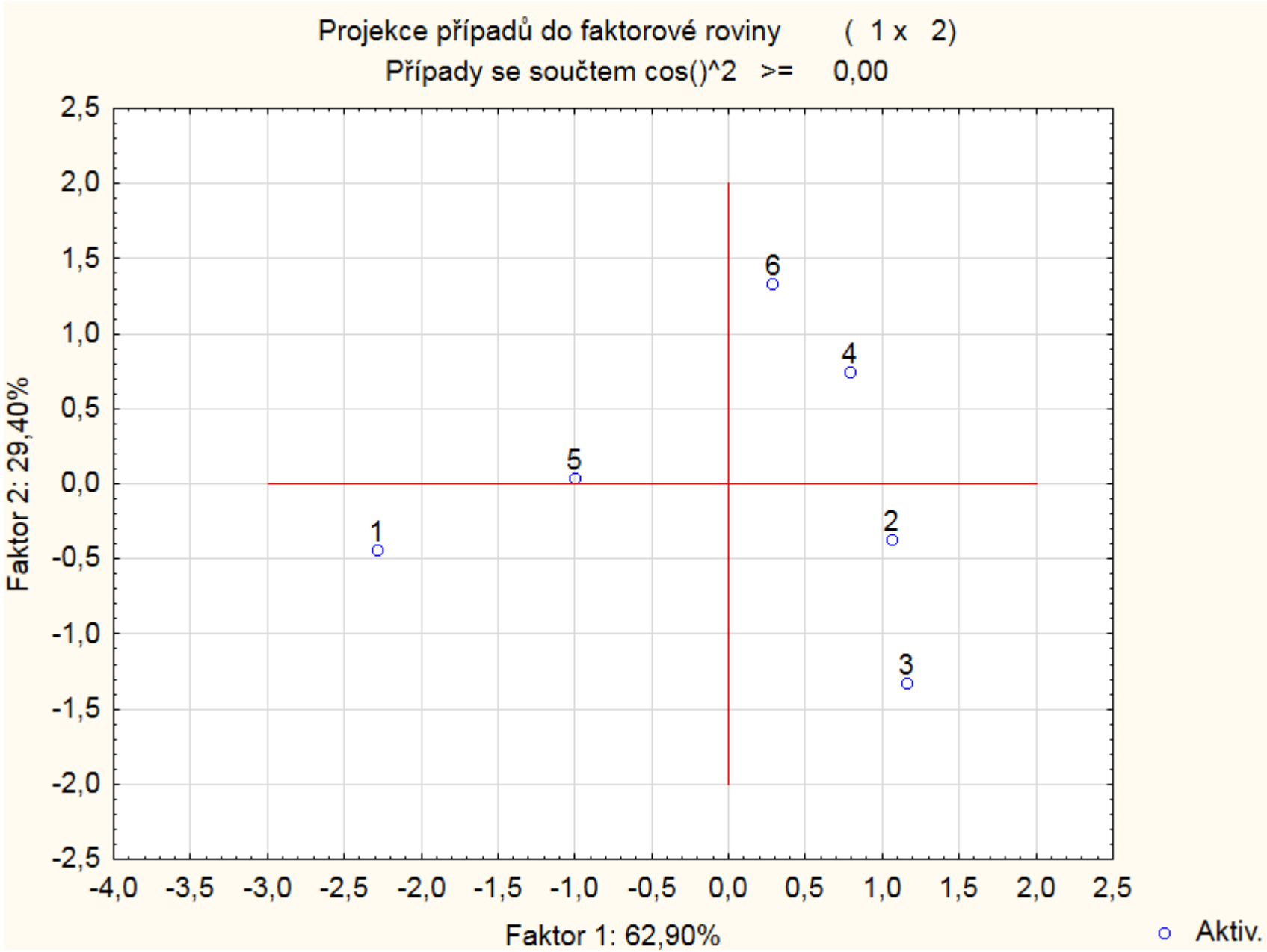
PCA
Analýza

Cílem je nahradit velké množství původních znaků několika hlavními komponentami, které vystihují většinu rozptylu.


STATISTICA Cz - [Data: Tabulka1* (10s krát 10ř)]
Soubor
Domů
Upravit
Zobrazit
Formát
Statistiky
Data mining
Grafy
Nástroje
Data
Základní statistiky
Vícenásobná regrese
ANOVA
Neparametrické statistiky
Prokládání rozdělení
Rozdělení a simulace
Základ
Pokročilé modely
Neuron. sítě
Vícer./průzkumné
PLS, PCA, ...
Analýza síly testu
VEPAC
Pokročilé/Vícerozměrné
Diagramy řízení kvality
Analýza procesu
Multivariate
DOE
Predictive
Six Sigma
Průmyslová statistika
STATISTICA VB
Dávk. analza (dle skupin)
Kalkulátory
Nástroje
Statistiky bloku dat
1 Prom1
2 Prom2
3 Prom3
4 Prom4
5 Prom5
6 Prom6
7 Prom7
8 Prom8
9 Prom9
10 Prom10
1 1 2 9 -1,53393 -0,5336761 1,67789014
2 7 6 4 0,76696499 0,88946008 -0,335578
3 5 8 3 0 1,60102815 -0,7382717
4 6 2 2 0,38348249 -0,5336761
5 3 2 6 -0,766965 -0,5336761
6 8 1 5 1,15044748 -0,8894601
7
8
9
10
Výsledky hlavních komponent a klasifikační analýzy: Tabulka1
Poč. aktiv. prom. :3
Počet doplňkových prom.: 0
Poč. aktiv. příp. : 6
Počet doplňkových příp.: 0
Vlast.čísla: 1,88694 ,881893 ,231168
Počet faktorů : 3
Kvalita reprezentace : 100,0 %
OK
Storno
Možnosti
Anal.Sk.
Základní výsledky
Proměnné
Případy
Popis. statistiky
Faktorové souřadnice prom.
2D graf fakt. souřadnic prom.
Faktorové souřadnice příp.
2D graf fakt. souřadnic příp.
Vlastní čísla
Sutinový graf


Vlastní čísla korelační matice
Pouze aktiv. proměnné
Vlast. číslo
Pořadí vl. čísla
62,90%
29,40%
7,71%
2,2
2,0
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,5
1,0
1,5
2,0
2,5
3,0
3,5
4,0


Projekce proměnných do faktorové roviny (1 x 2)
1,0
0,5
0,0
-0,5
-1,0
Faktor 2 : 29,40%
Prom5
Prom7
Prom6
-1,0
-0,5
0,0
0,5
1,0
Faktor 1 : 62,90%
Aktiv.


Projekce případů do faktorové roviny (1 x 2)
Případy se součtem cos()^2 >= 0,00
Faktor 2: 29,40%
Faktor 1: 62,90%
2,5
2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-4,0 -3,5 -3,0 -2,5 -2,0 -1,5 -1,0 -0,5 0,0 0,5 1,0 1,5 2,0 2,5
1
2
3
4
5
6
Aktiv.