*Statistické metody II, cvičení č. 9*

# *Shluková analýza*

# *(Cluster analysis)*

Brno, 9.5.2016
Klára Ambrožová

# Použitá data

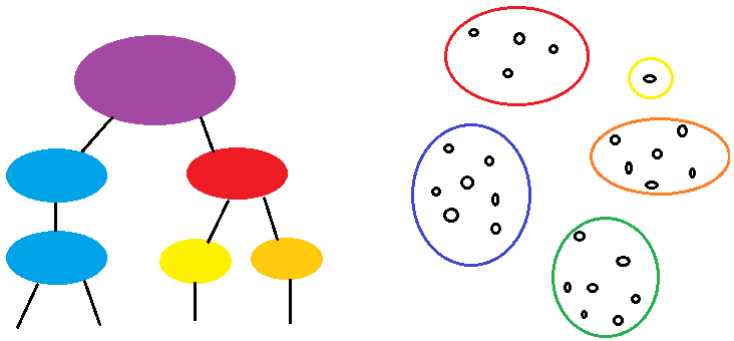
- V této prezentaci budou použita jiná data, než se kterými budete pracovat na cvičení
- Analýza a interpretace dat bude stejná
- Jde o počet obyvatel daného kraje ČR k 31. 12. 2014
- K analýze dat je použit program STATISTICA – karta Statistiky – modul Vicerozměrné/průzkumné techniky – Shluková analýza

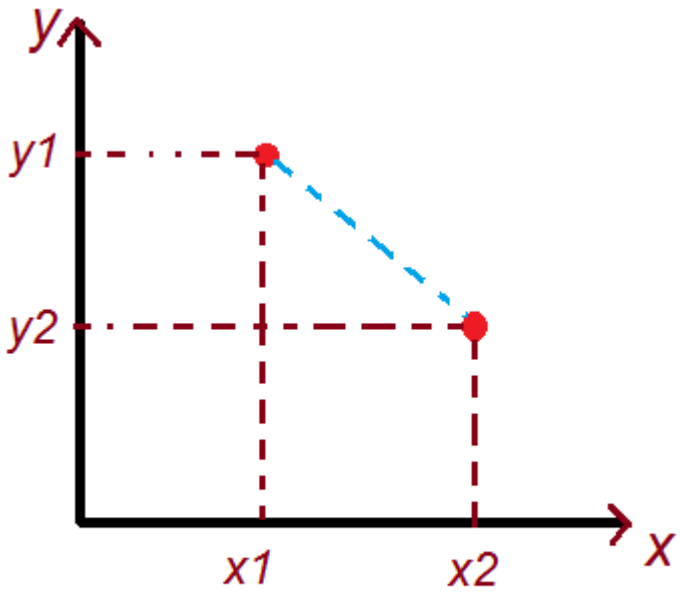
| | C:\Users\Klárka\Desktop\PERSONAL\1jD\Statistika2\cv9\Pocet_o | | | |
|---|---|---|---|---|
| | Počet_obyv | PO_do14let | PO_15-64let | PO_nad65let |
| Hlavní město Praha | 1 | 182 | 846 | 229 |
| Středočeský kraj | 1 | 220 | 874 | 219 |
| Jihočeský kraj | 637 | 96 | 425 | 114 |
| Plzeňský kraj | 575 | 85 | 384 | 105 |
| Karlovarský kraj | 299 | 44 | 202 | 52 |
| Ústecký kraj | 823 | 129 | 553 | 140 |
| Liberecký kraj | 438 | 68 | 292 | 77 |
| Královéhradecký kraj | 551 | 82 | 363 | 104 |
| Pardubický kraj | 516 | 78 | 344 | 93 |
| Kraj Vysočina | 509 | 76 | 340 | 92 |
| Jihomoravský kraj | 1 | 175 | 783 | 213 |
| Olomoucký kraj | 635 | 94 | 424 | 115 |
| Zlínský kraj | 585 | 84 | 393 | 107 |
| Moravskoslezský kraj | 1 | 179 | 824 | 213 |

Trocha teorie:

# Shluková analýza

- Metoda pro zhodnocení vícerozměrných dat
- Snaha o nalezení skupin, kde členové uvnitř skupiny si budou maximálně podobní, zatímco se budou maximálně lišit od členů ostatních skupin
- Existují metody hierarchické (vznikají shluky různých úrovní) a nehierarchické
- Pracujeme s pojmem „vzdálenost“ mezi objekty (často se používá euklidovská vzdálenost)





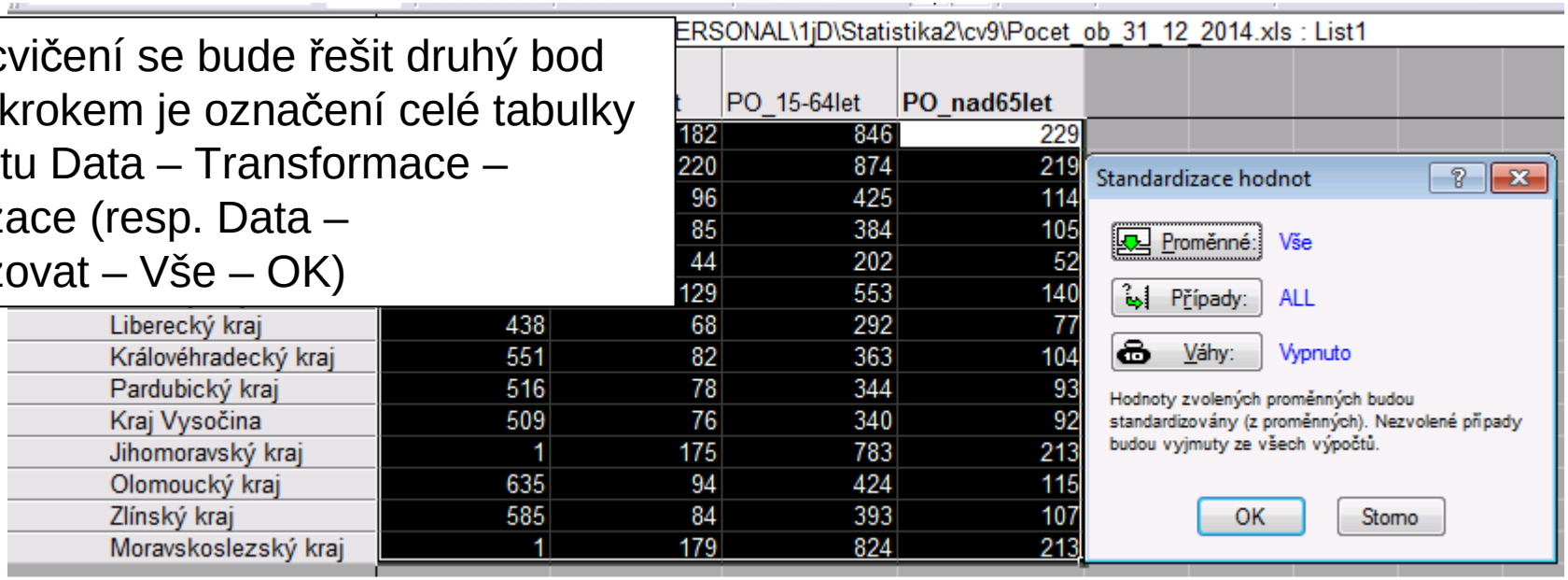

$$d_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2}$$

Trocha teorie:

# Shluková analýza

- Předpoklady shlukové analýzy:
  - Nekorelovanost proměnných (lze zajistit použitím výsledků PCA namísto původních proměnných)
  - Nezávislost na jednotkách (nutná standardizace)
  - Stejný význam proměnných při shlukování (lze vyřešit přidáním váhových koeficientů do výpočtu vzdáleností)

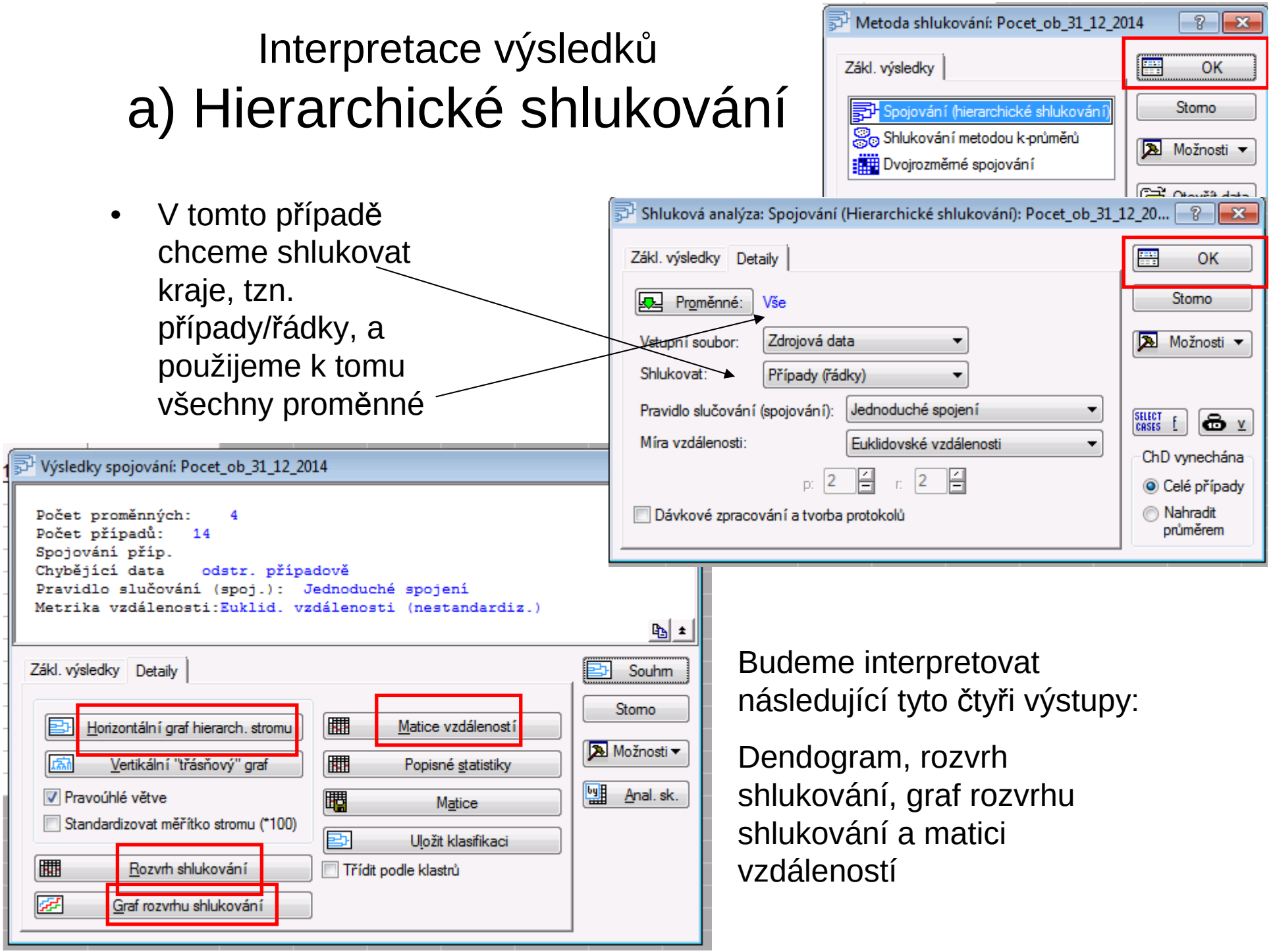
Pozn. Ve cvičení se bude řešit druhý bod → prvním krokem je označení celé tabulky a jít na kartu Data – Transformace – Standardizace (resp. Data – Standardizovat – Vše – OK)



Interpretace výsledků

# a) Hierarchické shlukování

- V tomto případě chceme shlukovat kraje, tzn. případy/řádky, a použijeme k tomu všechny proměnné



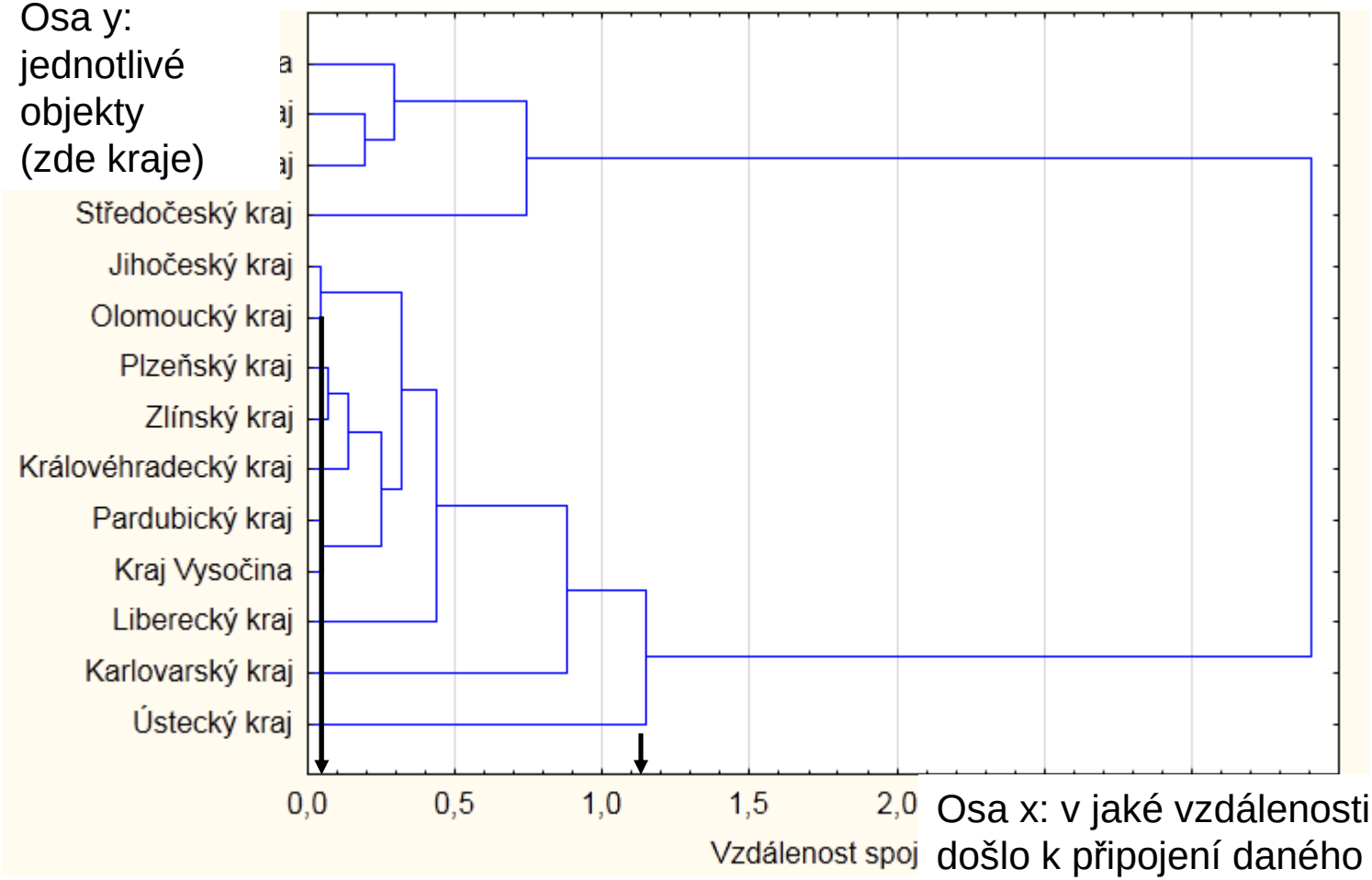

Budeme interpretovat následující tyto čtyři výstupy:

Dendogram, rozvrh shlukování, graf rozvrhu shlukování a matici vzdáleností

Interpretace výsledků

# a) Hierarchické shlukování - dendrogram

Osa y: jednotlivé objekty (zde kraje)



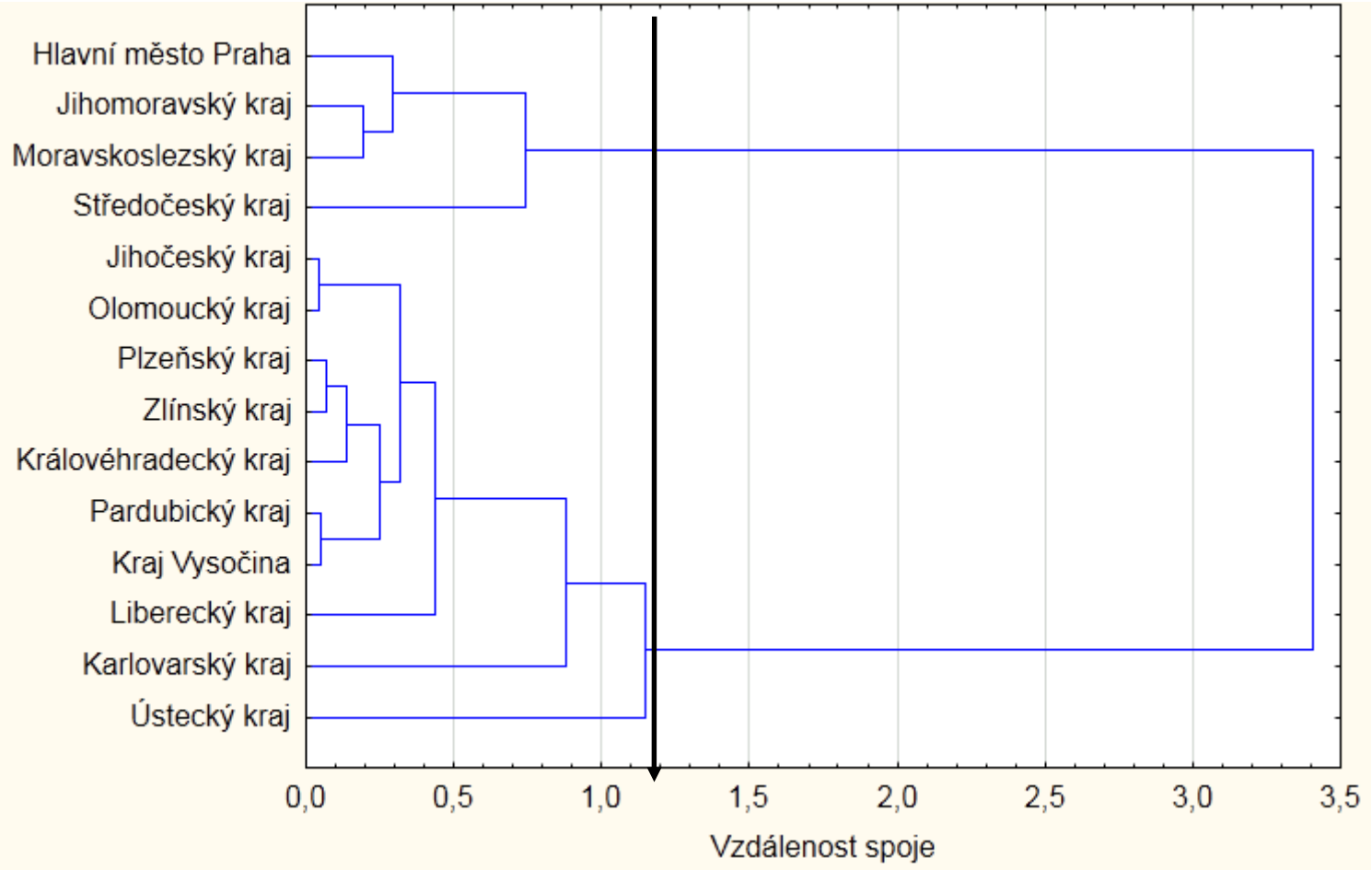

Na grafu vidíme, jak se postupně k sobě připojovaly jednotlivé kraje – čím více vpravo je kraj připojen, tím je méně podobný jádru objektů, které vzniklo jako první

Osa x: v jaké vzdálenosti došlo k připojení daného objektu

Např. první byly spojeny kraje Jihočeský a Olomoucký a téměř ve stejné vzdálenosti i Pardubický a Vysočina. Naopak, jako úplně poslední byl k spodnímu velkému shluku připojen Ústecký kraj, je tedy velmi odlišný.

Interpretace výsledků

# a) Hierarchické shlukování - dendrogram




Fakt, že tyto dva finální shluky se sloučily až ve vzdálenosti cca 3.4 poukazuje na to, že jsou skutečně velmi odlišné.

Před sloučením do finálního shluku nám zde vznikly dva velké shluky: v prvním je Praha, Jihomoravský kraj, Moravskoslezský kraj a Středočeský kraj, ve druhém shluku ostatní kraje. Tento shluk vzniknul ve vzdálenosti cca 1.2 (jde o bezrozměrné číslo).

# Interpretace výsledků

# a) Hierarchické shlukování – rozvrh shlukování

Zobrazuje téměř totéž jako předchozí graf: na ose y jsou tentokrát přesně číselně vzdálenosti, v nichž došlo ke spojení, na řádcích jsou pak shluky, přičemž poslední objekt na daném řádku byl připojen v tomto kroku.

Např. ve čtvrtém kroku (4. řádek) byl ke shluku Plzeňského a Zlínského kraje připojen Královéhradecký ve vzdálenosti 0, 137.

Rozvrh slučování (Pocet_ob_31_12_2014)
Jednoduché spojení
Euklid. vzdálenosti

| spojení vzdálen. | Obj. č. 1 | Obj. č. 2 | Obj. č. 3 | Obj. č. 4 | Obj. č. 5 | Obj. č. 6 | Obj. č. 7 | Obj. č. 8 | Obj. 9 |
|---|---|---|---|---|---|---|---|---|---|
| **,0418297** | Jihočeský kraj | Olomoucký kraj | | | | | | | |
| ,0509008 | Pardubický kraj | Kraj Vysočina | | | | | | | |
| ,0653331 | Plzeňský kraj | Zlínský kraj | | | | | | | |
| ,1375565 | Plzeňský kraj | Zlínský kraj | Královéhradecký kraj | | | | | | |
| ,1936774 | lihomoravský kraj | Moravskoslezský kraj | | | | | | | |
| ,2494285 | Plzeňský kraj | Zlínský kraj | éhradecký kraj | ardubický kraj | Kraj Vysočina | | | | |
| ,2925919 | avní město Praha | Jihomoravský kraj | Moravskoslezský kraj | | | | | | |
| ,3199249 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | radecký kraj | ardubický kraj | Kraj Vysočina | | |
| ,4390908 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | radecký kraj | ardubický kraj | raj Vysočina | Liberecký kraj | |
| ,7396800 | avní město Praha | Jihomoravský kraj | koslezský kraj | Středočeský kraj | | | | | |
| ,8789036 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | radecký kraj | ardubický kraj | raj Vysočina | Liberecký kraj | Karlova |
| 1,145492 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | radecký kraj | ardubický kraj | raj Vysočina | Liberecký kraj | Karlova |
| 3,408849 | avní město Praha | Jihomoravský kraj | koslezský kraj | edočeský kraj | ihočeský kraj | omoucký kraj | lzeňský kraj | Zlínský kraj | álovéhrade |

# Interpretace výsledků

# a) Hierarchické shlukování – rozvrh shlukování

**Interpretace dendogramu a rozvrhu shlukování:**

Je zjevné, že nejpodobnější jsou si kraje: Jihočeský a Olomoucký; Pardubický a Vysočina; Plzeňský a Zlínský; Jihomoravský a Moravskoslezský (vytvořily jádra shluků). Poslední uvedené jádro k sobě nakonec ještě přidalo Hl. město Prahu a Středočeský kraj, ostatní kraje vytvořily druhý shluk. Finální shluk (14. řádek) vznikl až ve vzdálenosti 3,4, což je rozdíl vůči přechozímu kroku o 3.4-1.14=2.26, tyto dva shluky jsou tedy extrémně rozdílné.

| | Rozvrh slučování (Pocet_ob_31_12_2014)<br>Jednoduché spojení<br>Euklid. vzdálenosti | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| spojení vzdálen. | **Obj. č. 1** | Obj. č. 2 | Obj. č. 3 | Obj. č. 4 | Obj. č. 5 | Obj. č. 6 | Obj. č. 7 | Obj. č. 8 | Obj. 9 |
| **,0418297** | Jihočeský kraj | Olomoucký kraj | | | | | | | |
| ,0509008 | Pardubický kraj | Kraj Vysočina | | | | | | | |
| ,0653331 | Plzeňský kraj | Zlínský kraj | | | | | | | |
| ,1375565 | Plzeňský kraj | Zlínský kraj | Královéhradecký kraj | | | | | | |
| ,1936774 | lihomoravský kraj | Moravskoslezský kraj | | | | | | | |
| ,2494285 | Plzeňský kraj | Zlínský kraj | éhradecký kraj | ardubický kraj | Kraj Vysočina | | | | |
| ,2925919 | avní město Praha | Jihomoravský kraj | Moravskoslezský kraj | | | | | | |
| ,3199249 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | nradecký kraj | ardubický kraj | Kraj Vysočina | | |
| ,4390908 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | nradecký kraj | ardubický kraj | raj Vysočina | Liberecký kraj | |
| ,7396800 | avní město Praha | Jihomoravský kraj | koslezský kraj | Středočeský kraj | | | | | |
| ,8789036 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | nradecký kraj | ardubický kraj | raj Vysočina | Liberecký kraj | Karlova |
| 1,145492 | Jihočeský kraj | Olomoucký kraj | Plzeňský kraj | Zlínský kraj | nradecký kraj | ardubický kraj | raj Vysočina | Liberecký kraj | Karlova |
| 3,408849 | avní město Praha | Jihomoravský kraj | koslezský kraj | edočeský kraj | ihočeský kraj | omoucký kraj | lzeňský kraj | Zlínský kraj | álovéhrade |

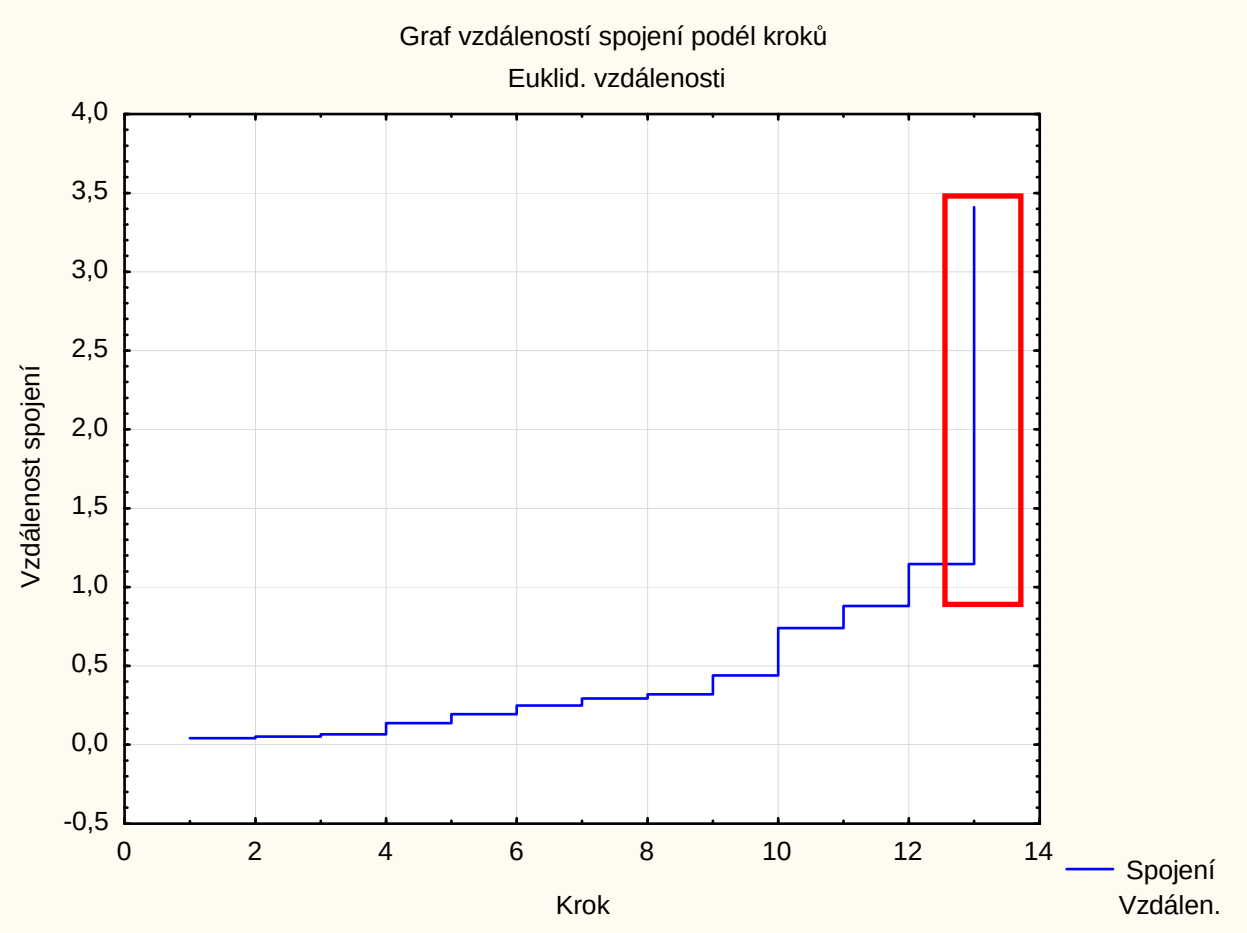
# Interpretace výsledků
# a) Hierarchické shlukování – graf rozvrhu shlukování

Graf rozvrhu shlukování má na ose x krok, v němž došlo k připojení a na ose y vzdálenost (podobně jako byla předtím na ose x u dendogramu).

Hledáme optimální počet shluků: nalezneme největší „schod“ v grafu.

Zde je schod zjevně na 13. kroku, tzn. 14. krok shlukování bychom již neměli provádět. Tzn. ideální jsou v tomto případě **dva shluky**.

Někdy to však z grafu nebude 100 % jasné a bude třeba zjistit, kde došlo k maximální ztrátě informace, z rovzrhu shlukování.



# Interpretace výsledků
# a) Hierarchické shlukování – matice vzdáleností

Matice vzdáleností je čtvercová a určuje vzdálenost mezi objektem na řádku a ve sloupci. Tzn. na diagonále budou nuly (vzdálenost mezi Hlavním městem Prahou a Hlavním městem Prahou je přirozeně 0).

Např. vzdálenost mezi Středočeským krajem a Prahou je 0,739.

| | Pocet_ob_31_12_2014 | | | | | |
|---|---|---|---|---|---|---|
| | **1 Hlavní m** | 2 Středoče | 3 Jihočesk | 4 Plzeňský | 5 Karlovar | 6 Ústeck |
| **Hlavní m** | 0,00000 | 0,73968 | 3,83818 | 3,97614 | 4,95833 | 3,6 |
| Středoče | 0,73968 | 0,00000 | 4,16506 | 4,32545 | 5,33496 | 3,8 |
| Jihočesk | 3,83818 | 4,16506 | 0,00000 | 0,38057 | 2,09765 | 1,1 |
| Plzeňský | 3,97614 | 4,32545 | 0,38057 | 0,00000 | 1,72079 | 1,5 |
| Karlovar | 4,95833 | 5,33496 | 2,09765 | 1,72079 | 0,00000 | 3,2 |
| Ústecký | 3,63727 | 3,87090 | 1,14549 | 1,52598 | 3,23584 | 0,0 |
| Libereck | 4,39730 | 4,75649 | 1,22098 | 0,84734 | 0,87890 | 2,3 |
| Královéh | 4,01630 | 4,37339 | 0,51129 | 0,13756 | 1,59842 | 1,6 |
| Pardubic | 4,13905 | 4,49343 | 0,73813 | 0,36294 | 1,35993 | 1,8 |
| Kraj Vys | 4,16467 | 4,52274 | 0,78601 | 0,40895 | 1,31224 | 1,9 |
| Jihomora | 0,40713 | 0,93480 | 3,52370 | 3,64147 | 4,57323 | 3,4 |
| Olomouck | 3,84344 | 4,17707 | 0,04183 | 0,36299 | 2,08318 | 1,1 |
| Zlínský | 3,96530 | 4,31877 | 0,34223 | 0,06533 | 1,76859 | 1,4 |
| Moravsko | 0,29259 | 0,80253 | 3,63758 | 3,76436 | 4,71406 | 3,4 |
| Průměry | 0,74593 | 0,91156 | -0,04170 | -0,23029 | -1,08678 | 0,5 |
| Sm.Odch. | 1,43456 | 1,55481 | 0,58800 | 0,56876 | 0,49380 | 0,6 |
| Poč.příp | 4,00000 | | | | | |
| Matice | 3,00000 | | | | | |

Tyto řádky nebudou v následujícím kroku potřeba a je třeba je **odstranit!**

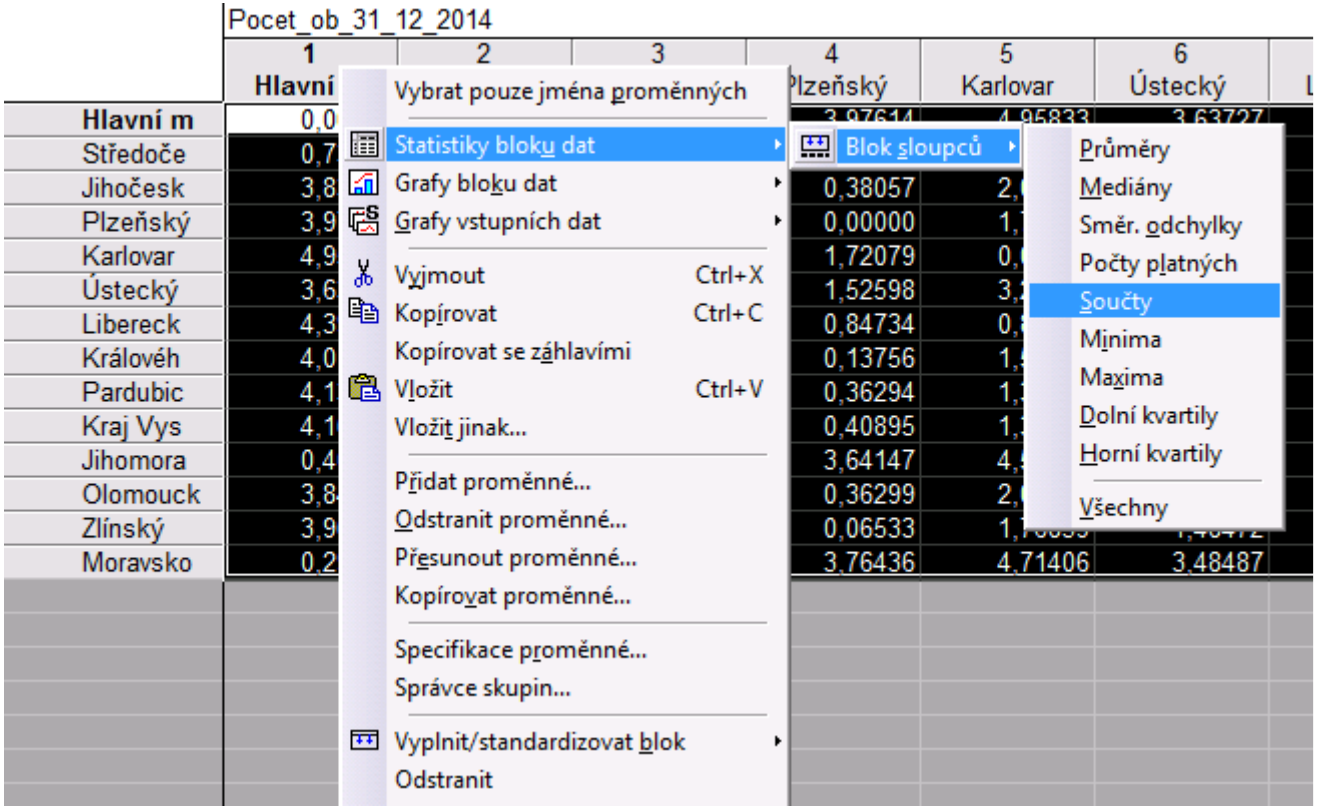
# Interpretace výsledků
# a) Hierarchické shlukování – matice vzdáleností

Chceme odpověď na otázku: Který kraj je „typický“ (nejbližší všem ostatním)? A které jsou naopak zcela atypické?

1) Odstranit poslední čtyři případy

2) Označit všechny sloupce v matici vzdáleností, kliknout na hlavičku a dát „Statistiky bloku dat – Blok sloupců – Součty“

3) Ve vzniklé tabulce hledáme minima a maxima



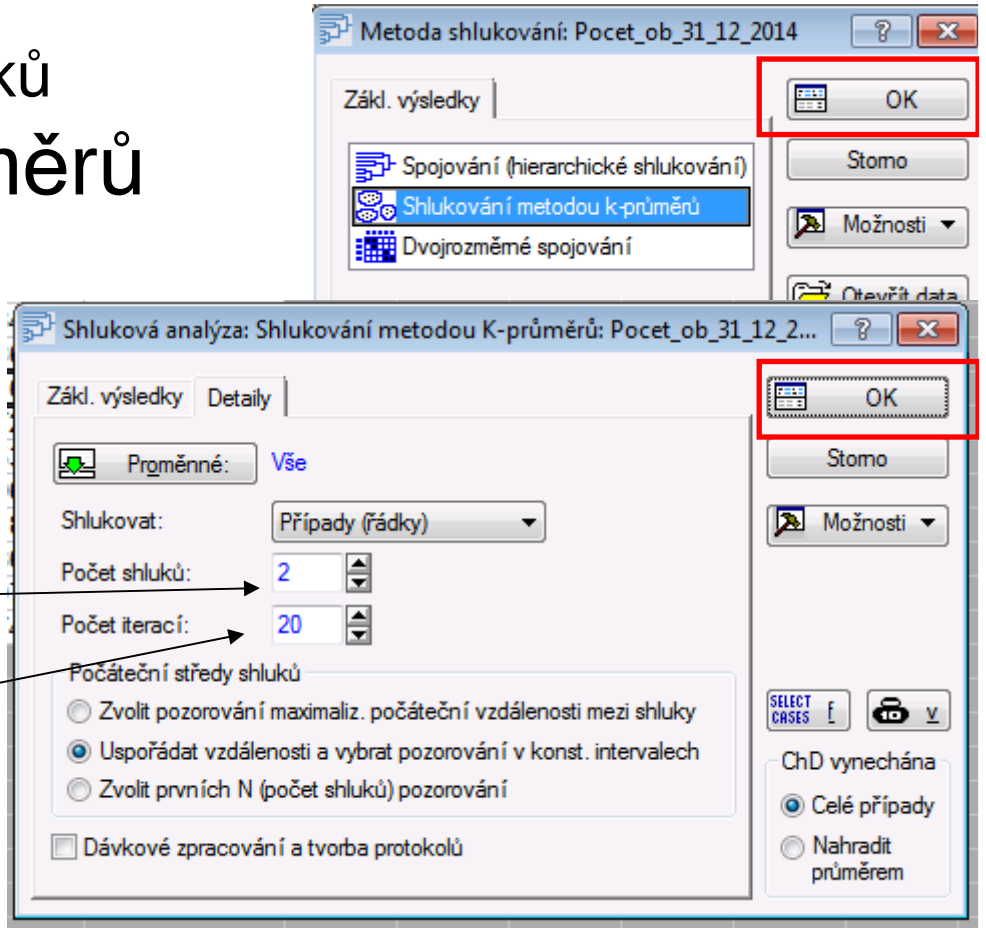

**Nejtypičtější je zde Zlínský kraj = je to takový „nejprůměrnější kraj“ v ČR z hlediska obyvatelstva v r. 2014**

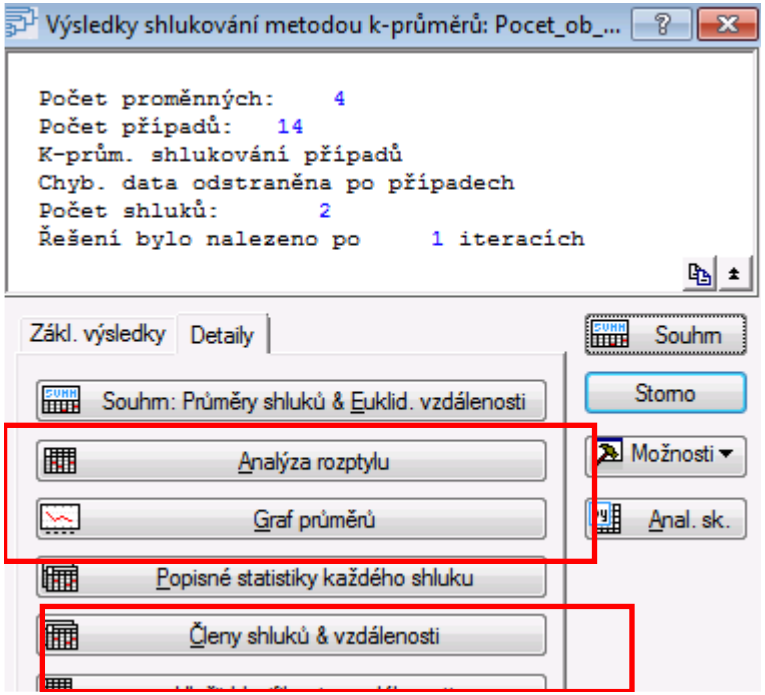
Pocet_ob_31_12_2014

| | 1 Hlavní m | 2 Středoče | 3 Jihočesk | 4 Plzeňský | 5 Karlovar | 6 Ústecký | 7 Libereck | 8 Královéh | 9 Pardubic | 10 Kraj Vys | 11 Jihomora | 12 Olomouck | 13 Zlínský | 14 Moravsko |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SOUČET případy 1-14 | 42,3753909 | 46,8152429 | 22,4286976 | 21,5198785 | 35,636125 | 30,7820631 | 26,4089376 | 21,72157 | 22,6015561 | 22,8891578 | 39,1483657 | 22,363226 | 21,618127 | 40,1113631 |

# Interpretace výsledků
# b) Metoda k-průměrů

- V tomto případě chceme shlukovat kraje, tzn. případy/řádky, a použijeme k tomu všechny proměnné
- Jelikož mi v předchozím případě vyšel ideální počet shluků jako 2, nastavím, že chci 2 shluky
- Minimální počet iterací je 20 (je možno nastavit i víc)



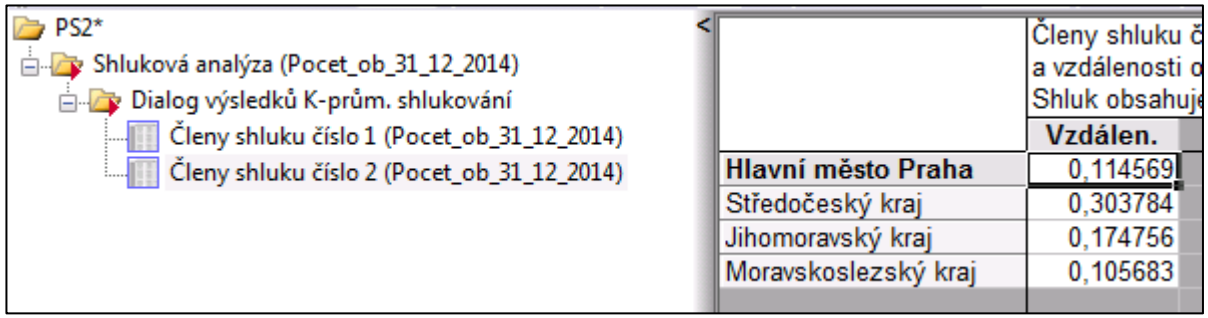




Budeme interpretovat následující tyto čtyři výstupy:

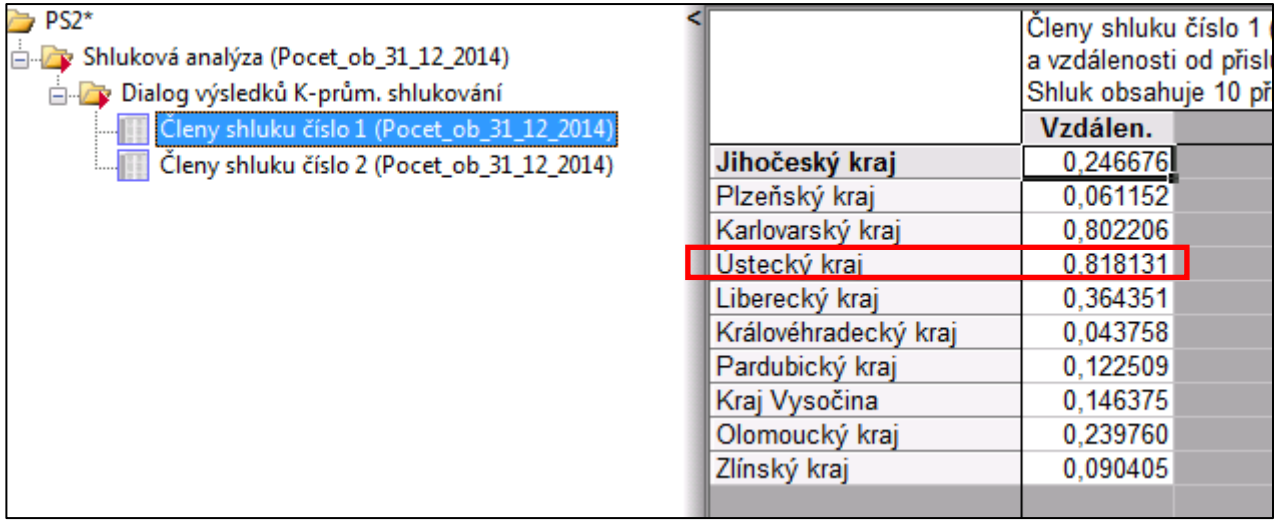
Členy shluků a vzdáleností,
graf průměrů, analýzu rozptylu

# Interpretace výsledků
# b) Metoda k-průměrů – členy shluků a vzdálenosti

- Vzniknou (v mém případě) dvě tabulky, z jedné na druhou lze překliknout v nabídce vlevo




| | Členy shluku č a vzdálenosti o Shluk obsahuje |
|---|---|
| | Vzdálen. |
| **Hlavní město Praha** | 0,114569 |
| Středočeský kraj | 0,303784 |
| Jihomoravský kraj | 0,174756 |
| Moravskoslezský kraj | 0,105683 |




| | Členy shluku číslo 1 ( a vzdálenosti od přísl Shluk obsahuje 10 př |
|---|---|
| | Vzdálen. |
| **Jihočeský kraj** | 0,246676 |
| Plzeňský kraj | 0,061152 |
| Karlovarský kraj | 0,802206 |
| Ústecký kraj | 0,818131 |
| Liberecký kraj | 0,364351 |
| Královéhradecký kraj | 0,043758 |
| Pardubický kraj | 0,122509 |
| Kraj Vysočina | 0,146375 |
| Olomoucký kraj | 0,239760 |
| Zlínský kraj | 0,090405 |

Je zjevné, že členy shluků jsou v tomto případě úplně stejné jako v předchozím.

Vzdálenost ve sloupci je vzdálenost od středu daného shluku – tzn. čím menší, tím „typičtější“ je daný kraj pro tento shluk.

Např. největší vzdálenost má Ústecký kraj – potvrzuje se, že je velmi odlišný.

# Interpretace výsledků
# b) Metoda k-průměrů – analýza rozptylu

- Analýzou rozptylu zde testujeme,
  které proměnné nejvíce ovlivnily rozřazení do shluků

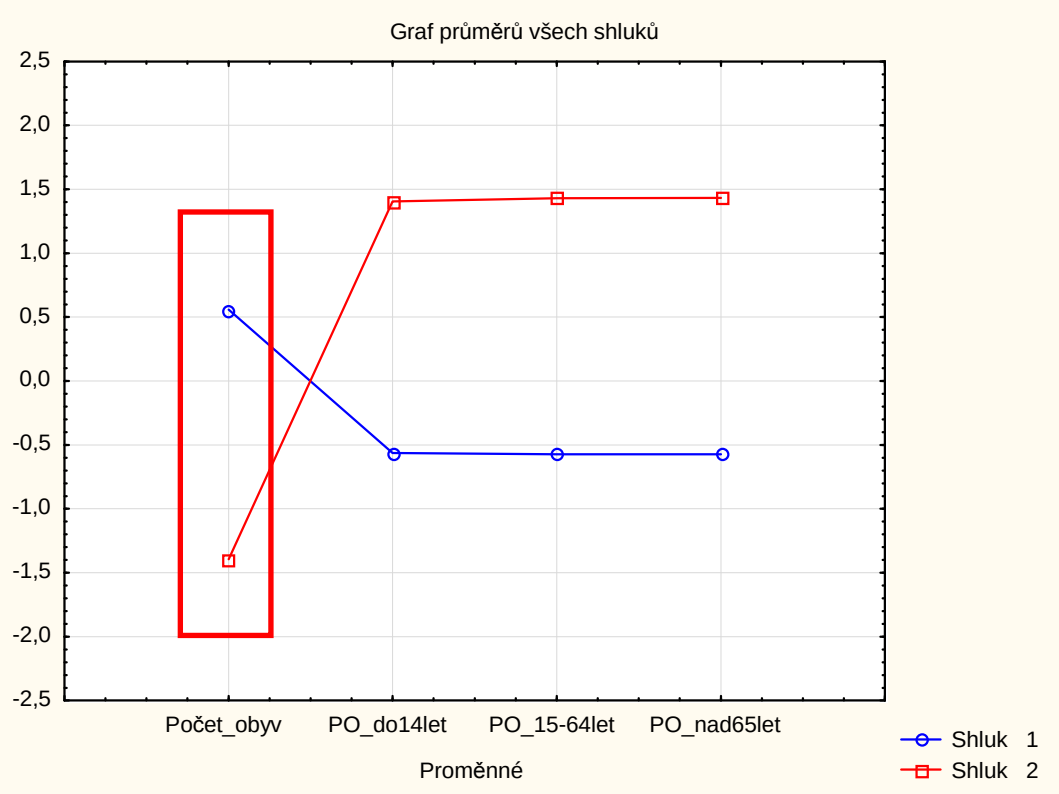
| Proměnná | Analýza rozptylu (Pocet_ob_31_12_2014) | | | | | |
|---|---|---|---|---|---|---|
| | **Mezisk. SČ** | sv | Vnitřní SČ | sv | F | význam. p |
| **Počet_obyv** | 10,90985 | 1 | 2,090148 | 12 | 62,63586 | 0,000004 |
| PO_do14let | 11,06982 | 1 | 1,930181 | 12 | 68,82146 | 0,000003 |
| PO_15-64let | 11,47141 | 1 | 1,528587 | 12 | 90,05502 | 0,000001 |
| PO_nad65let | 11,50356 | 1 | 1,496436 | 12 | 92,24768 | 0,000001 |

Hodnotu testové statistiky F zde můžeme interpretovat tak, že čím vyšší, tím větší byl vliv.

Tzn. proměnné počet obyvatel v kategorii 15–64 let a nad 65 let ovlivnily rozdělení víc, než zbylé dvě proměnné.

# Interpretace výsledků
# b) Metoda k-průměrů – graf průměrů

- Graf průměrů ukazuje průměrnou hodnotu (osa y, bezrozměrné číslo) dané proměnné (osa x) pro konkrétní shluk (rozlišeny barevně).




Zde vidíme, že pro shluk číslo 1 je typický vysoký počet obyvatel, do druhého státy s nízkým počtem obyvatel.

Teoreticky by se také dalo říct, že druhý shluk má obecně více obyvatel ve všech třech kategoriích (0-14, 15-64 a 65+ let), což je samozřejmě nesmysl.

Jde o důsledek toho, že na začátku nebyla použita průzkumová analýza dat, která by prokázala korelaci mezi proměnnými → správně bylo třeba provést nejdřív PCA!

# Interpretace výsledků

- Pomineme-li špatné použití proměnných ke shlukování z důvodu chybějící průzkumové analýzy dat:
  - Pomocí obou  metod vznikly dva shluky:
    - Praha, Jihomoravský kraj, Moravskoslezský kraj a Středočeský kraj
    - Jihočeský, Olomoucký, Plzeňský, Zlínský, Královéhradecký, Pardubice, Vysočina, Liberecký, Karlovarský, Ústecký
  - V prvním shluku jsou kraje s velmi mladou věkovou strukturou (nadprůměrný počet jedinců v kategorii do 14 let – lze zjistit po prohlédnutí tabulky standardizovaných dat)
  - Nejtypičtější v ČR je Zlínský kraj
  - Naopak nejméně typické byly Středočeský a Hlavní město Praha (z dat vyplynulo, že jde o důsledek enormně vysokého počtu obyvatel vůči ostatním krajům – zejména krajům 2. shluku)

# Poznámka ke cvičení

- Cvičení bude obsahovat:
  - Pro hierarchické shlukování:
    - Dendogram,
    - rozvrh shlukování,
    - graf rozvrhu shlukování
    - matici vzdáleností
  - Pro nehierarchické shlukování:
    - Členy shluků a vzdáleností,
    - graf průměrů,
    - analýzu rozptylu
  - Výstupy budou ve vhodné podobě (tzn. čitelné a popsané, abych nemusela hádat, co vlastně je na obrázku)

- Pozor!
  - Máte jiná data, může vám jako optimální vyjít jiný počet shluků než 2
  - Pro napsání závěru si přečíst požadavky profesora Dobrovolného v pdf zadání (body 6–10)
  - Není potřeba popisovat teorii nebo „na ose x obr. 1 je tohle a tohle…“ – v tomhle cvičení je tolik výsledků, které se dají popsat a interpretovat, že by to klidně vyšlo na stránku

# Poznámka ke cvičení

- Data ve cvičení: data o evropských zemích z r. 1979
- Zastoupení činného obyvatelstva v kategoriích:
  - Zemědělství
  - Těžba
  - Průmyslová výroba
  - Energetika
  - Stavebnictví
  - Místní hospodářství
  - Finance
  - Služby
  - Doprava a komunikace

Zdroje:


- BUDÍKOVÁ, Marie. Shluková analýza (přednáška). Brno: Masarykova univerzita, 9.5. 2016.

- DOBROVOLNÝ, Petr.  Z2069 Statistické metody a zpracování dat II: Shluková analýza (přednáška) Brno: Masarykova univerzita,9.5. 2016.