Vícerozměrné metody - cvičení

RNDr. Eva Koriťáková, Ph.D.

Podzim 2017

Shluková analýza – typy metod – opakování

Shluková analýza

Hierarchické
shluky jsou definovány postupným skládáním objektů

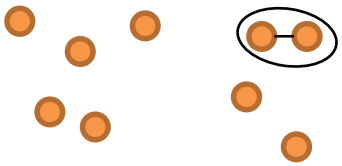
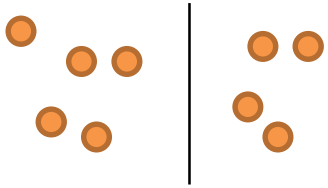
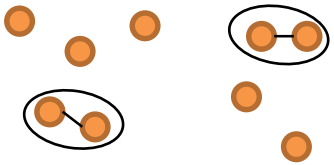
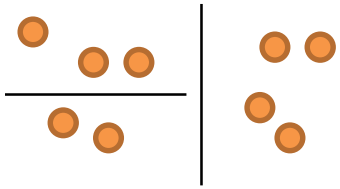
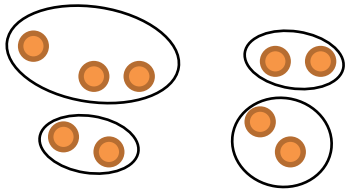
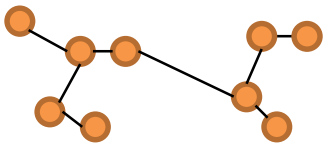
Nehierarchické
shluky jsou definovány v jednom kroku

Aglomerativní
Po spojení první dvojice objektů dochází k postupnému napojování dalších objektů.

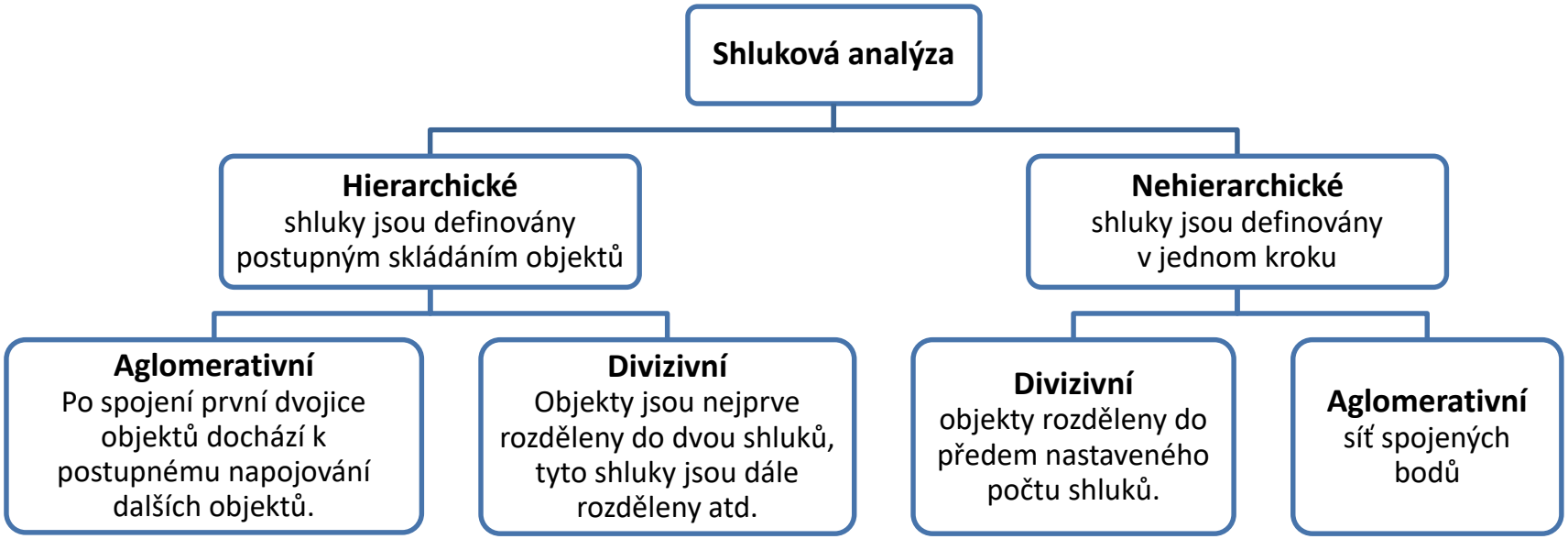
Divizivní
Objekty jsou nejprve rozděleny do dvou shluků, tyto shluky jsou dále rozděleny atd.

Divizivní
objekty rozděleny do předem nastaveného počtu shluků.

Aglomerativní
síť spojených bodů

1. Krok			Kolik shluků chceme definovat? Například 4	Minimum spanning tree, Prime network
2. Krok				
X. Krok	Atd.	Atd.	Výpočet ukončen	Výpočet ukončen

Shluková analýza – typy metod – opakování



Metoda nejbližšího souseda
Metoda nejvzdál. souseda
Metoda průměrné vazby
Centroidová metoda
Wardova metoda

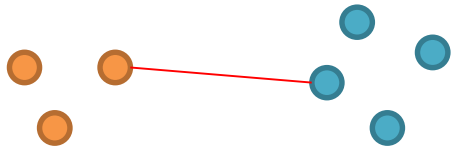
Metody monotetické
Metody polytetické – TWINSPAN

Metoda *k*-průměrů
Metoda *x*-průměrů
Metoda *k*-medoidů

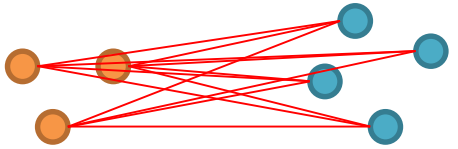
Minimum spanning tree

Shlukovací algoritmy hierarchického aglomerativního shlukování

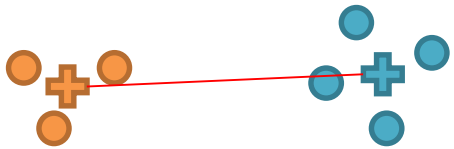
- **Metoda nejbližšího souseda** (jednospojná metoda, metoda jediné vazby, metoda krátké ruky, *nearest neighbour, simple linkage*) – spojení dle nejmenší vzdálenosti mezi objekty shluků



- **Metoda průměrné vazby** (středospojná metoda, *average linkage*) – spojení dle průměrné vzdálenosti mezi objekty shluků
  - Nevážená (*unweighted, UPGMA*) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků
  - Vážená (*weighted, WPGMA*) – odstranění vlivu velikosti shluků, shluky bez ohledu na velikost přispívají k výpočtu spojovací vzdálenosti stejnou vahou



- **Centroidová metoda** (centroidní metoda, metoda středospojné vzdálenosti, Gowerova metoda, *centroid method*) – spojení dle vzdálenosti centroidů shluků
  - Nevážená (*unweighted, UPGMC*) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků
  - Vážená (*weighted, WPGMC*, mediánová metoda, *median method*) – odstranění vlivu velikosti shluků



- **Metoda nejvzdálenějšího souseda** (všespojná metoda, metoda dlouhé ruky, *furthest neigbour, complete linkage*) – spojení dle největší vzdálenosti mezi objekty shluků

Příklad 1

V experimentu byla u 5 buněčných linií zjišťována kvantita membránových markerů popisujících jejich citlivost k chemoterapii. V přiložené tabulce naleznete změřené hodnoty standardizované na referenční buněčnou linii.

Buněčná linie	Marker 1	Marker 2
A	2	4
B	2	8
C	6	10
D	10	14
E	11	13

Vztahy mezi liniemi jsou vyjádřeny následující asociační maticí:

	A	B	C	D	E
A	0.0	4.0	7.2	12.8	12.7
B	4.0	0.0	4.5	10.0	10.3
C	7.2	4.5	0.0	5.7	5.8
D	12.8	10.0	5.7	0.0	1.4
E	12.7	10.3	5.8	1.4	0.0

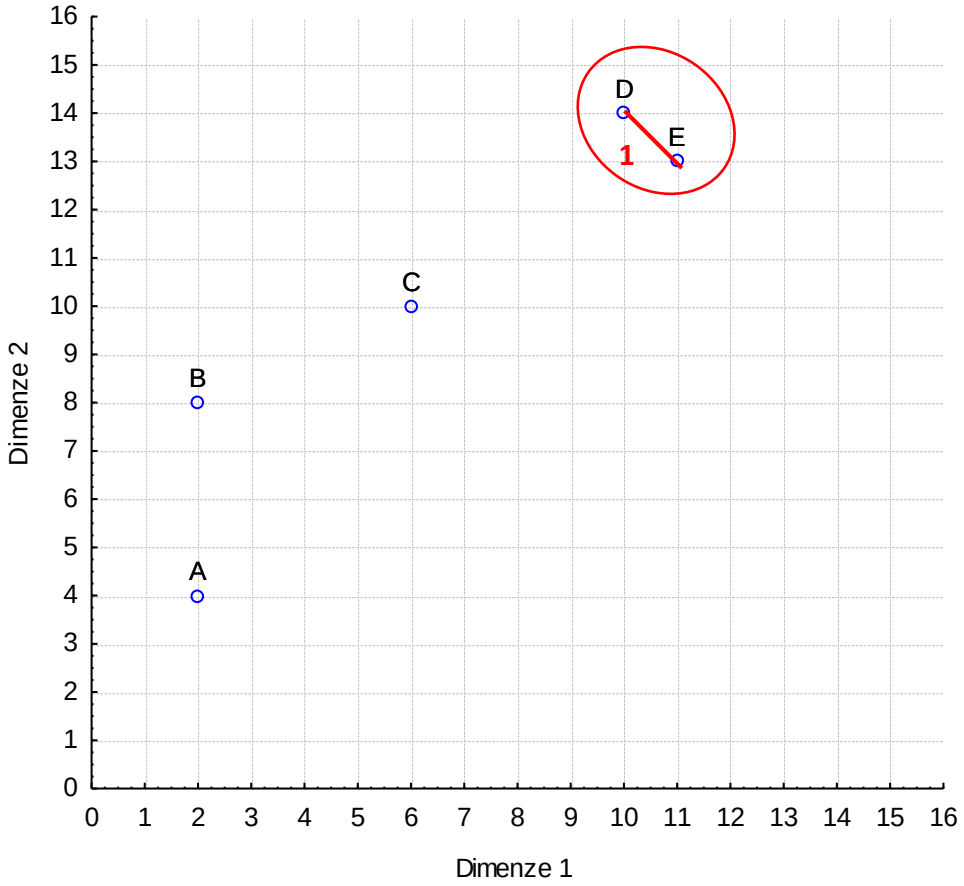
1. Výše uvedená asociační matice vyjadřuje podobnost nebo vzdálenost? A proč?
2. K výpočtu prvků asociační matice byl použit Jaccardův koeficient, Gowerův koeficient, Euklidova metrika nebo Hammingova (manhattanská) metrika?
3. Zdůvodněte vhodnost či nevhodnost použití tohoto koeficientu či metriky v případě těchto dat.
4. Vytvořte dendrogram pomocí algoritmu nejbližšího a nejvzdálenějšího souseda, rozepište jednotlivé kroky výpočtu.

Metoda nejbližšího souseda: 1. krok výpočtu

- Je vypočtena asociační matice

	A	B	C	D	E
A	0.0	4.0	7.2	12.8	12.7
B	4.0	0.0	4.5	10.0	10.3
C	7.2	4.5	0.0	5.7	5.8
D	12.8	10.0	5.7	0.0	1.4
E	12.7	10.3	5.8	1.4	0.0

- Je definován shluk dvou nejbližších objektů
  D-E

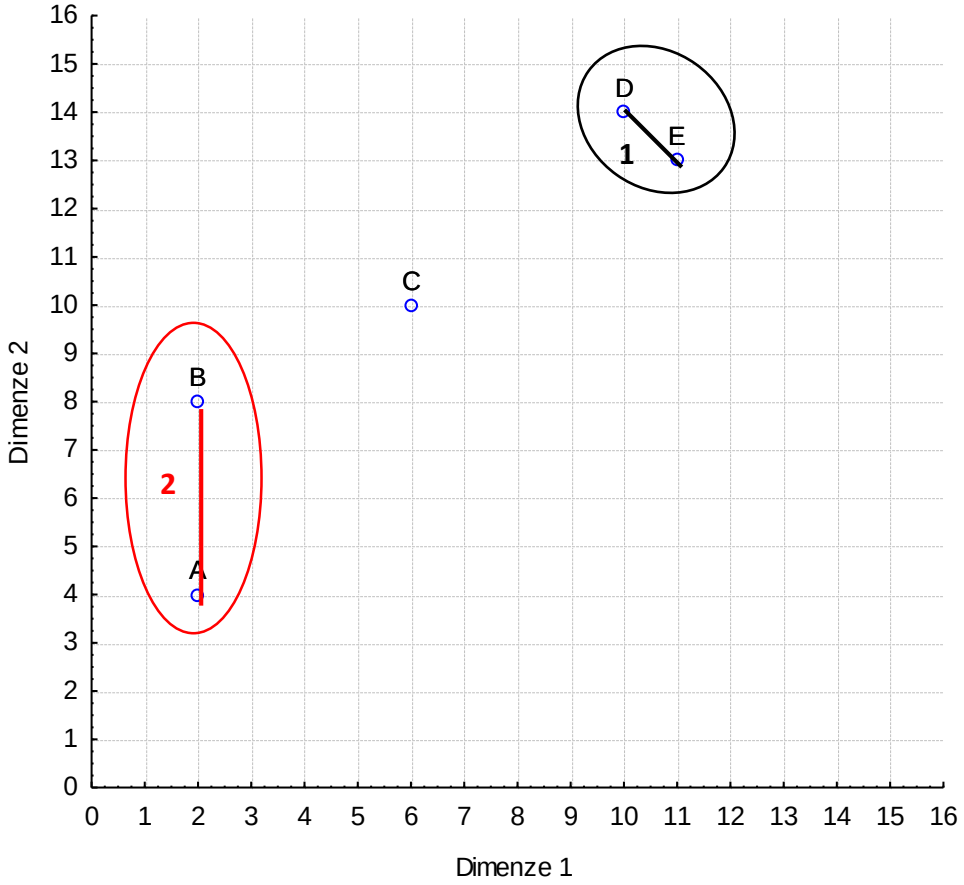


Metoda nejbližšího souseda: 2. krok výpočtu

- Je vypočtena asociační matice, kde objekty D-E již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **nejmenší vzdáleností od jeho členů (D, E)**

	A	B	C	D	E
A	0.0	4.0	7.2	12.8	12.7
B	4.0	0.0	4.5	10.0	10.3
C	7.2	4.5	0.0	5.7	5.8
D	12.8	10.0	5.7	0.0	1.4
E	12.7	10.3	5.8	1.4	0.0

	A	B	C	D+E
A	0.0	4.0	7.2	12.7
B	4.0	0.0	4.5	10.0
C	7.2	4.5	0.0	5.7
D+E	12.7	10.0	5.7	0.0



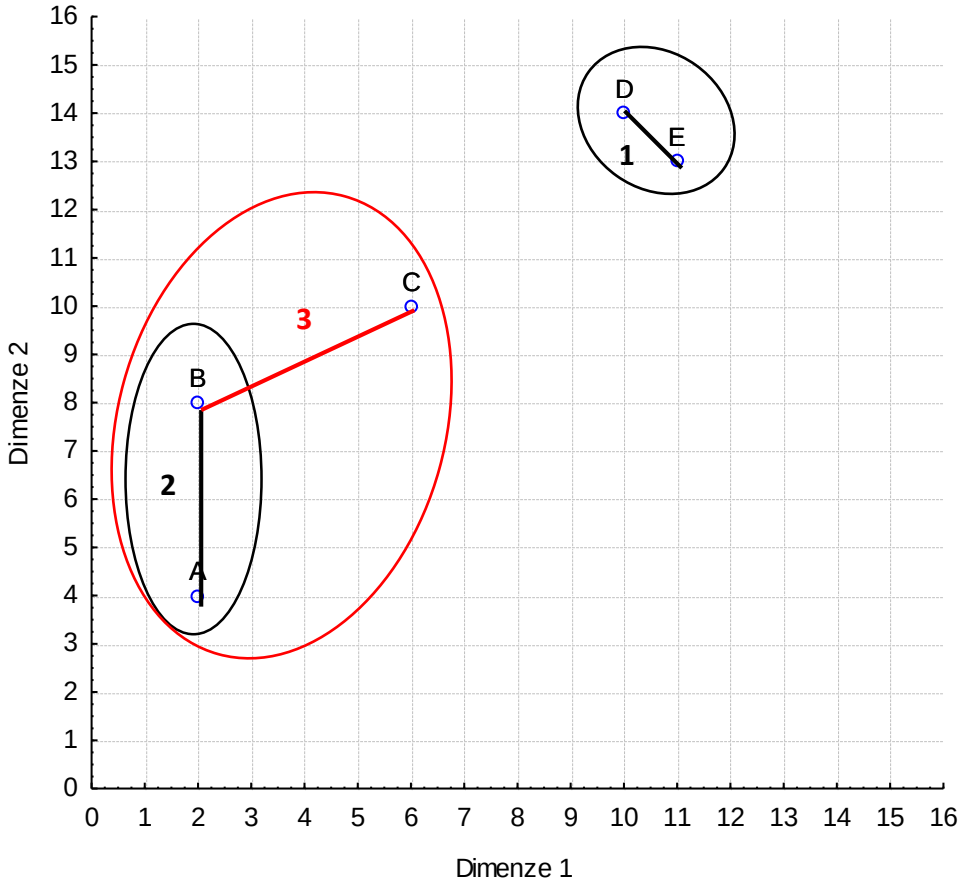
- Je definován shluk dvou nejbližších objektů A-B

Metoda nejbližšího souseda: 3. krok výpočtu

- Je vypočtena asociační matice, kde objekty A-B již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **nejmenší vzdáleností od jeho členů (A, B)**

	A	B	C	D+E
A	0.0	4.0	7.2	12.7
B	4.0	0.0	4.5	10.0
C	7.2	4.5	0.0	5.7
D+E	12.7	10.0	5.7	0.0

	A+B	C	D+E
A+B	0.0	4.5	10.0
C	4.5	0.0	5.7
D+E	10.0	5.7	0.0



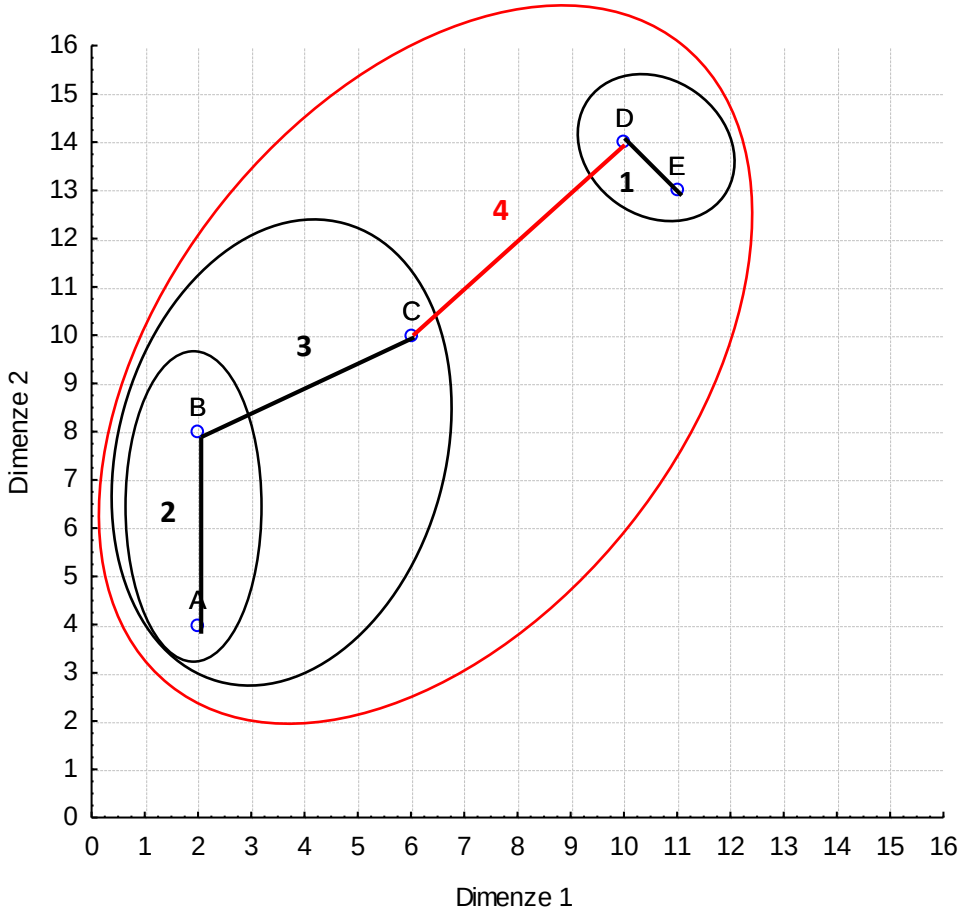
- Je definován shluk dvou nejbližších objektů (A-B)-C

Metoda nejbližšího souseda: 4. krok výpočtu

- Je vypočtena asociační matice, kde objekty (A-B)-C již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **nejmenší vzdáleností od jeho členů (A, B, C)**

	A+B	C	D+E
A+B	0.0	4.5	10.0
C	4.5	0.0	5.7
D+E	10.0	5.7	0.0

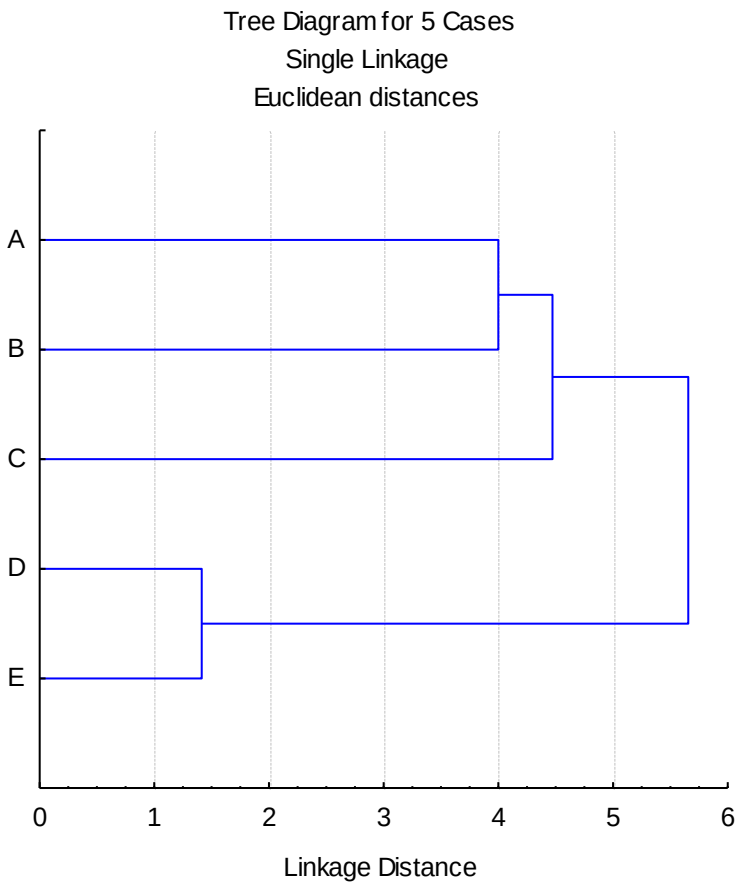
	A+B+C	D+E
A+B+C	0.0	5.7
D+E	5.7	0.0

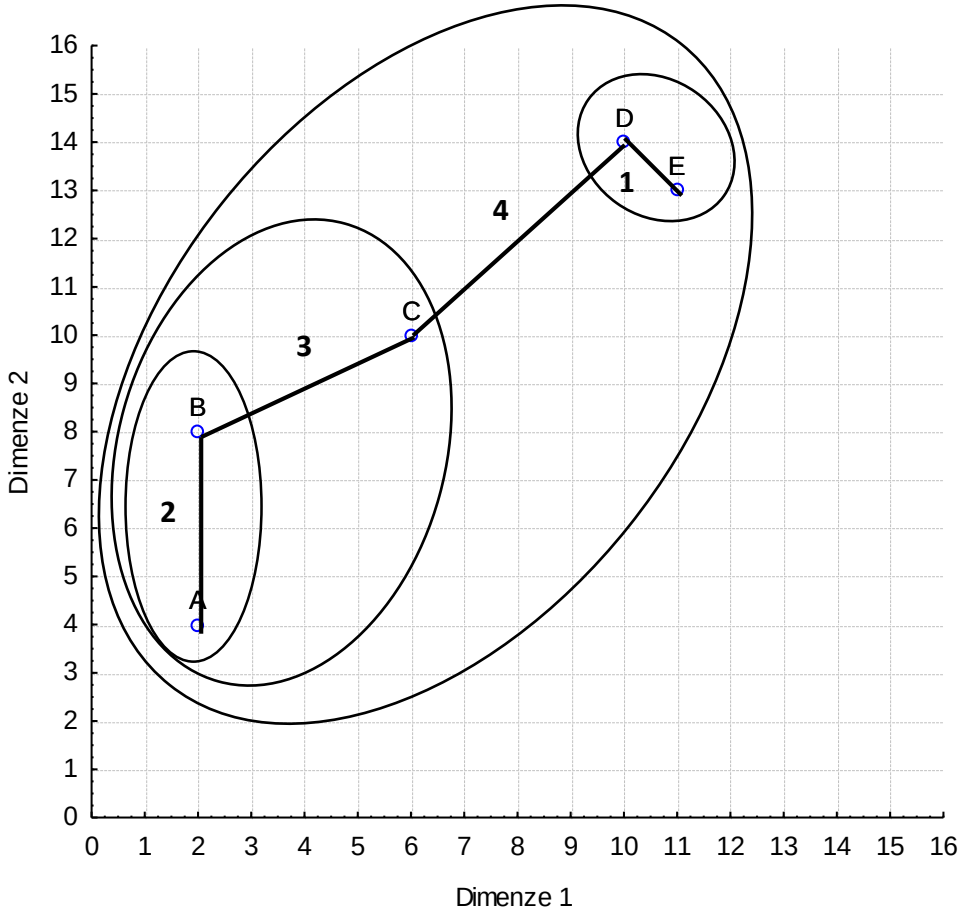


- Je definován shluk dvou nejbližších objektů ((A-B)-C)-(D-E)
- Všechny objekty jsou spojeny, algoritmus je ukončen

Metoda nejbližšího souseda: výsledek analýzy

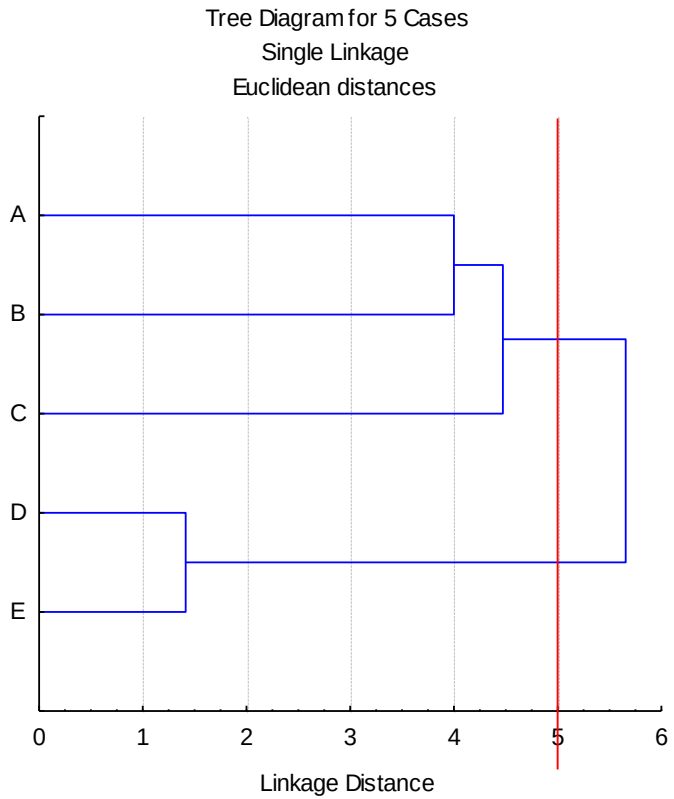
- Výsledek analýzy je vizualizován ve formě dendrogramu

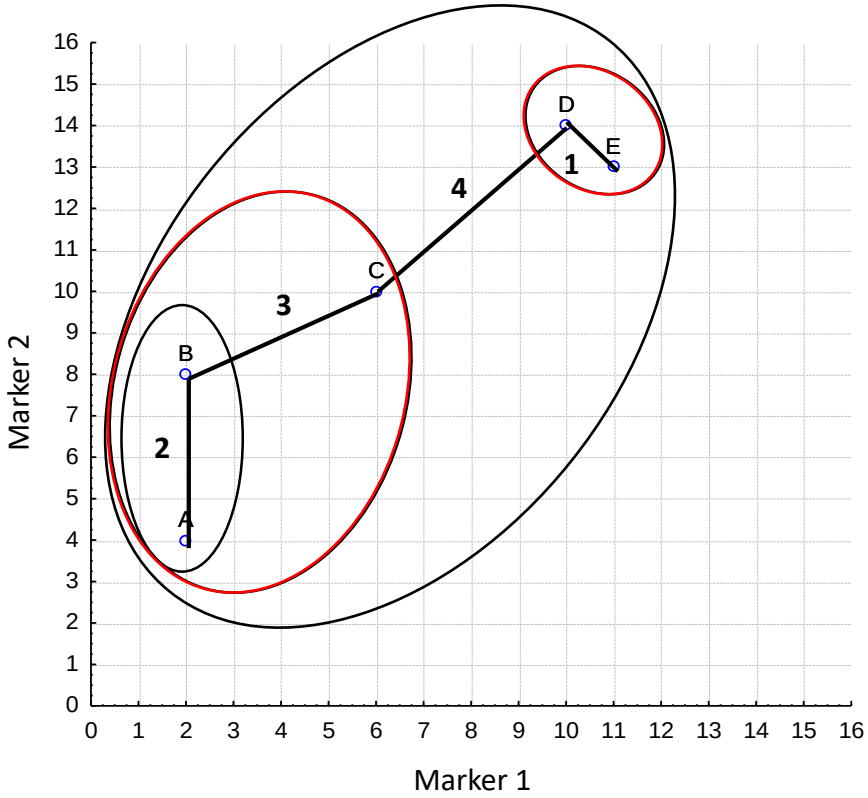




Metoda nejbližšího souseda: výsledek analýzy

Pokud bychom v dendrogramu provedli řez na podobnosti/vzdálenosti 5, kolik dostaneme shluků? Které buněčné linie budou v jednotlivých shlucích? Výsledek interpretujte.





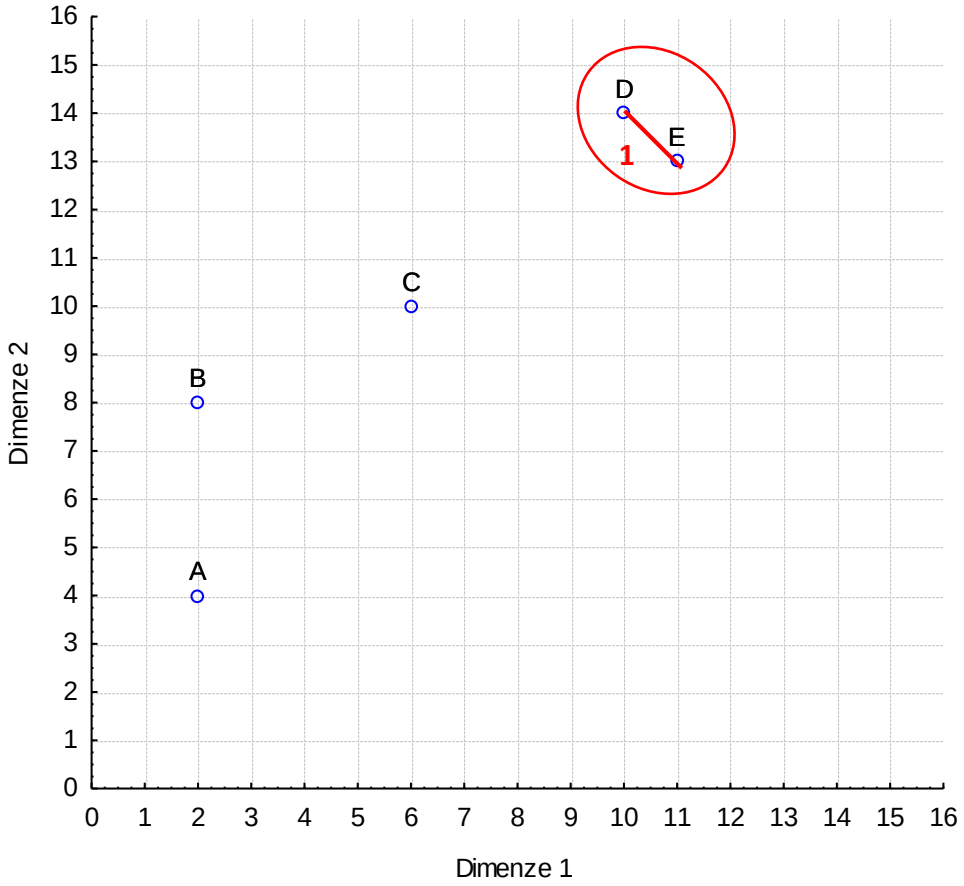
→ dostaneme 2 shluky: (A+B+C) a (D+E); přičemž linie D a E mají mnohem vyšší hodnoty obou markerů než linie A, B a C

Metoda nejvzdálenějšího souseda: 1. krok výpočtu

- Je vypočtena asociační matice

	A	B	C	D	E
A	0.0	4.0	7.2	12.8	12.7
B	4.0	0.0	4.5	10.0	10.3
C	7.2	4.5	0.0	5.7	5.8
D	12.8	10.0	5.7	0.0	1.4
E	12.7	10.3	5.8	1.4	0.0

- Je definován shluk dvou nejbližších objektů
  D-E

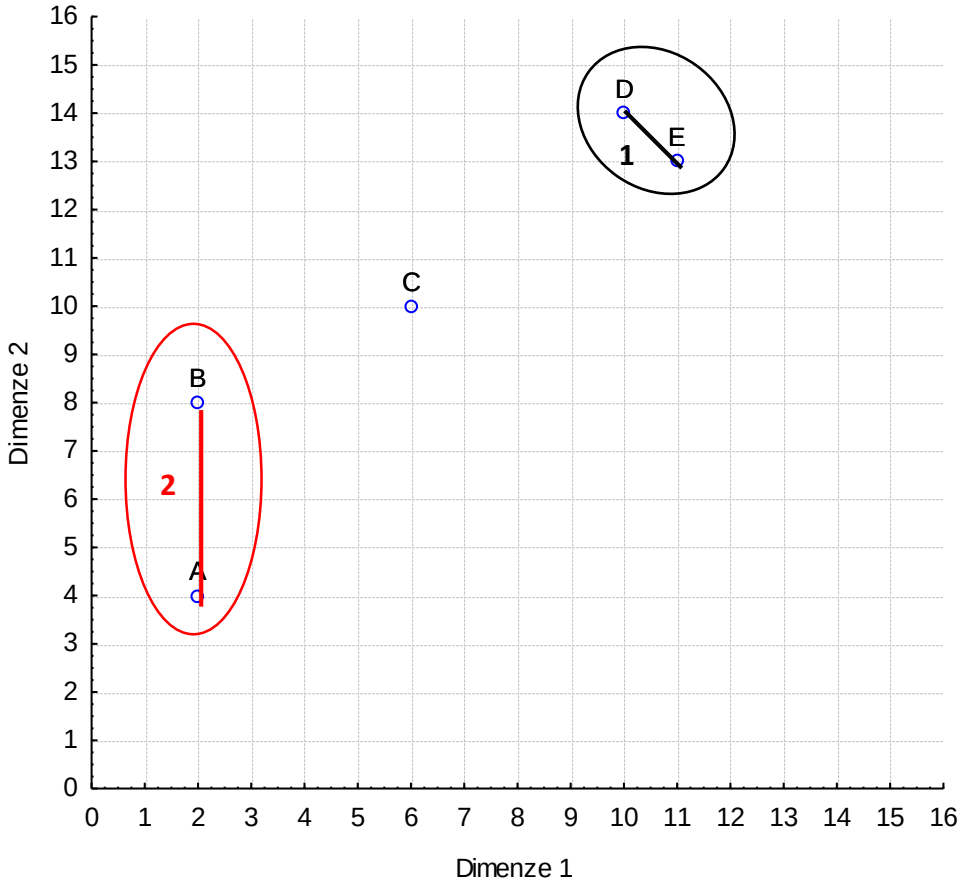


Metoda nejvzdálenějšího souseda: 2. krok výpočtu

- Je vypočtena asociační matice, kde objekty D-E již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **největší vzdáleností od jeho členů (D, E)**

	A	B	C	D	E
A	0.0	4.0	7.2	12.8	12.7
B	4.0	0.0	4.5	10.0	10.3
C	7.2	4.5	0.0	5.7	5.8
D	12.8	10.0	5.7	0.0	1.4
E	12.7	10.3	5.8	1.4	0.0

	A	B	C	D+E
A	0.0	4.0	7.2	12.8
B	4.0	0.0	4.5	10.3
C	7.2	4.5	0.0	5.8
D+E	12.8	10.3	5.8	0.0



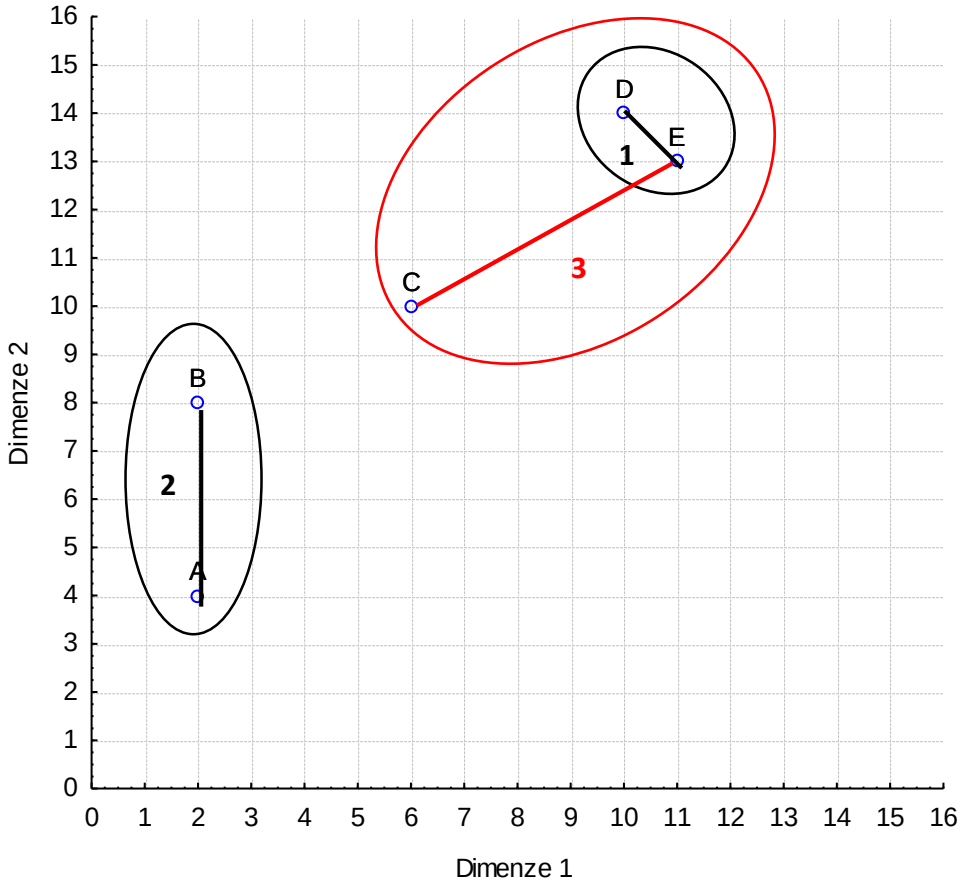
- Je definován shluk dvou nejbližších objektů
  A-B

Metoda nejvzdálenějšího souseda: 3. krok výpočtu

- Je vypočtena asociační matice, kde objekty A-B již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **největší vzdáleností od jeho členů (A, B)**

	A	B	C	D+E
A	0.0	4.0	7.2	12.8
B	4.0	0.0	4.5	10.3
C	7.2	4.5	0.0	5.8
D+E	12.8	10.3	5.8	0.0

	A+B	C	D+E
A+B	0.0	7.2	12.8
C	7.2	0.0	5.8
D+E	12.8	5.8	0.0



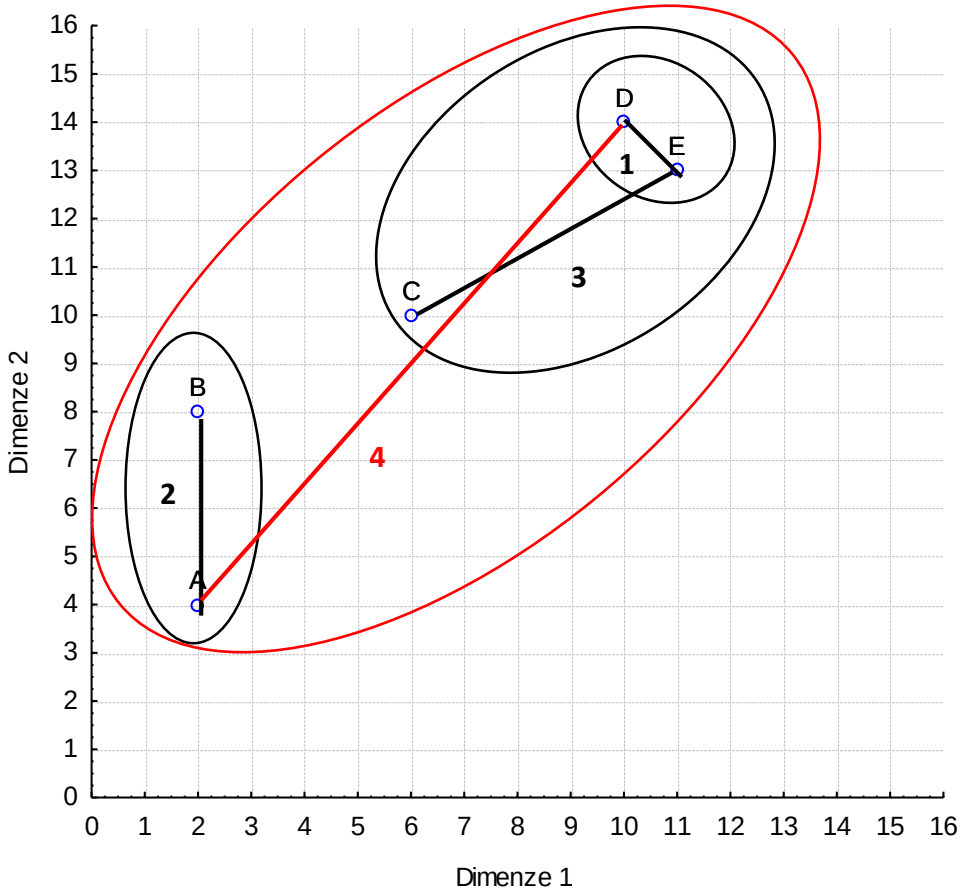
- Je definován shluk dvou nejbližších objektů (D-E)-C

Metoda nejvzdálenějšího souseda: 4. krok výpočtu

- Je vypočtena asociační matice, kde objekty (D-E)-C již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **největší vzdáleností od jeho členů (D, E, C)**

	A+B	C	D+E
A+B	0.0	7.2	12.8
C	7.2	0.0	5.8
D+E	12.8	5.8	0.0

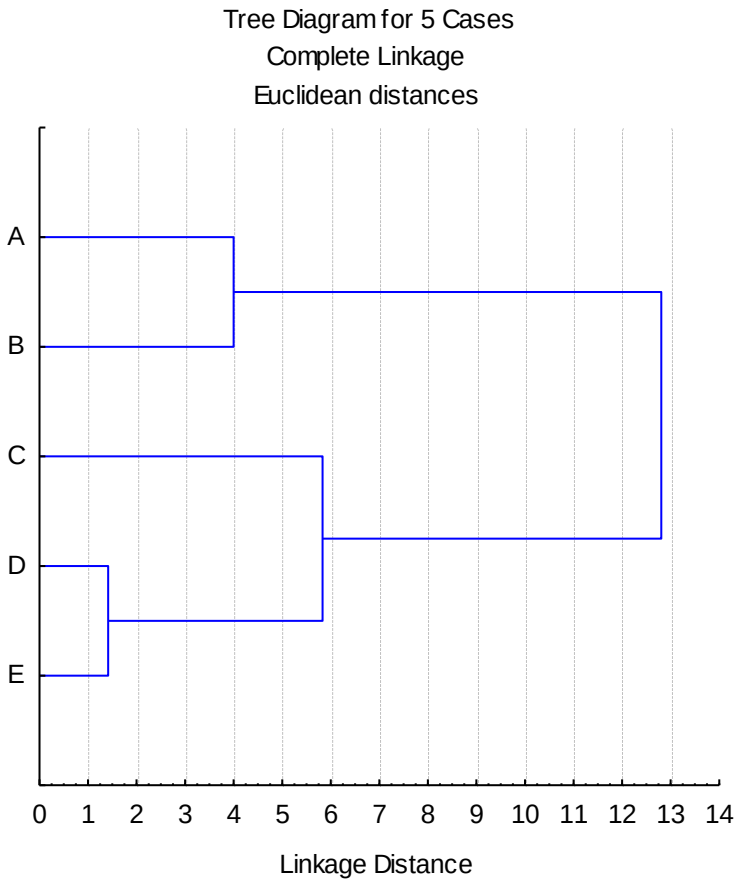
	A+B	D+E+C
A+B	0.0	12.8
D+E+C	12.8	0.0

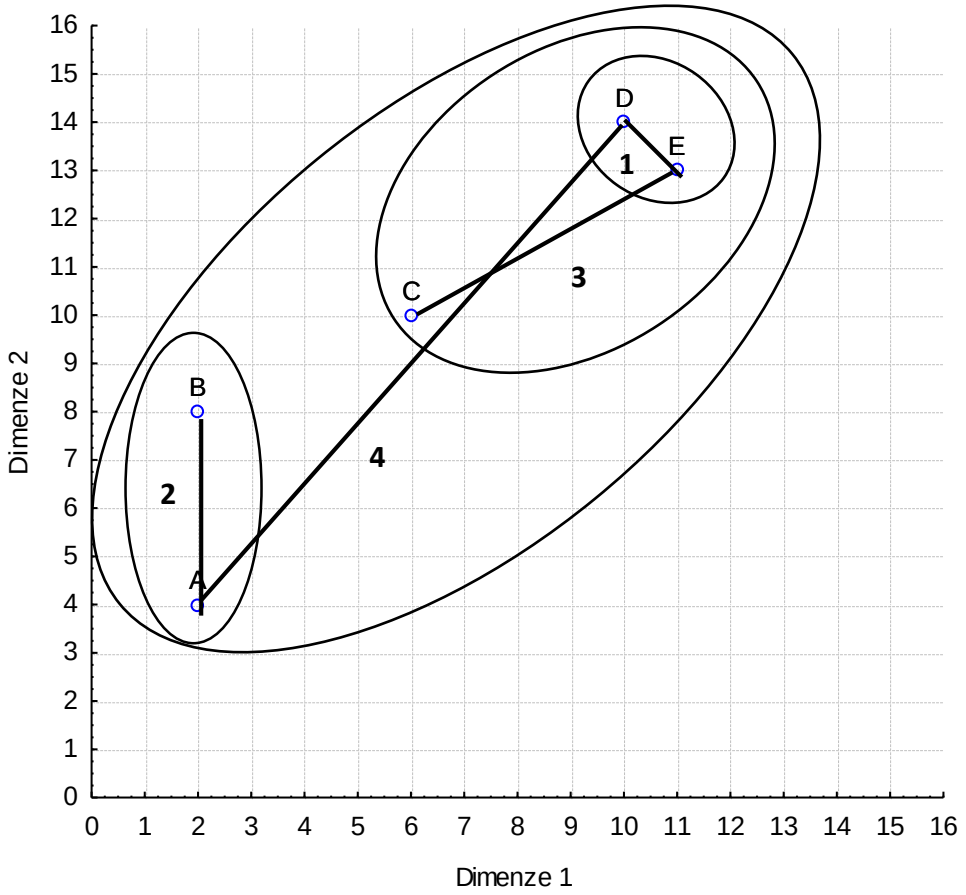


- Je definován shluk dvou nejbližších objektů ((D-E)-C)-(A-B)
- Všechny objekty jsou spojeny, algoritmus je ukončen

Metoda nejvzdálenějšího souseda: výsledek analýzy

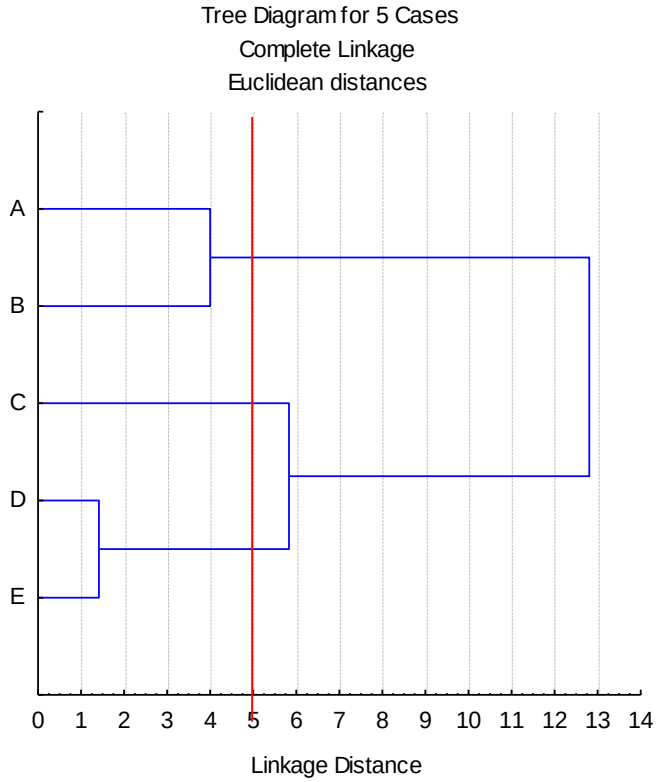
- Výsledek analýzy je vizualizován ve formě dendrogramu

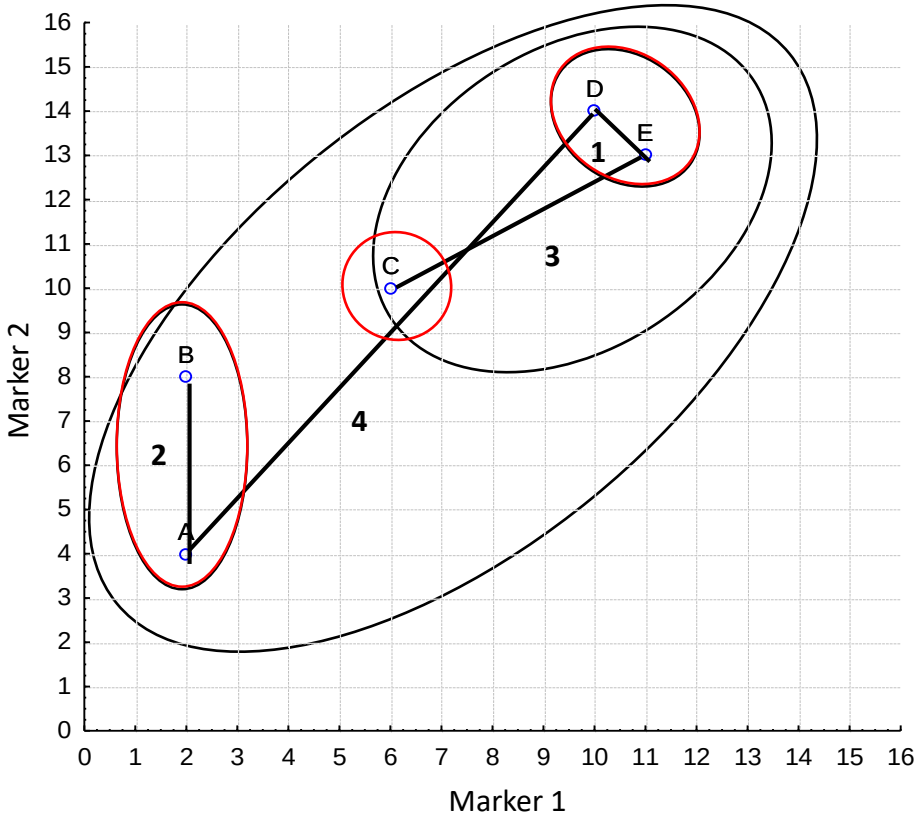




Metoda nejvzdálenějšího souseda: výsledek analýzy

Pokud bychom v dendrogramu provedli řez na podobnosti/vzdálenosti 5, kolik dostaneme shluků? Které buněčné linie budou v jednotlivých shlucích? Výsledek interpretujte.

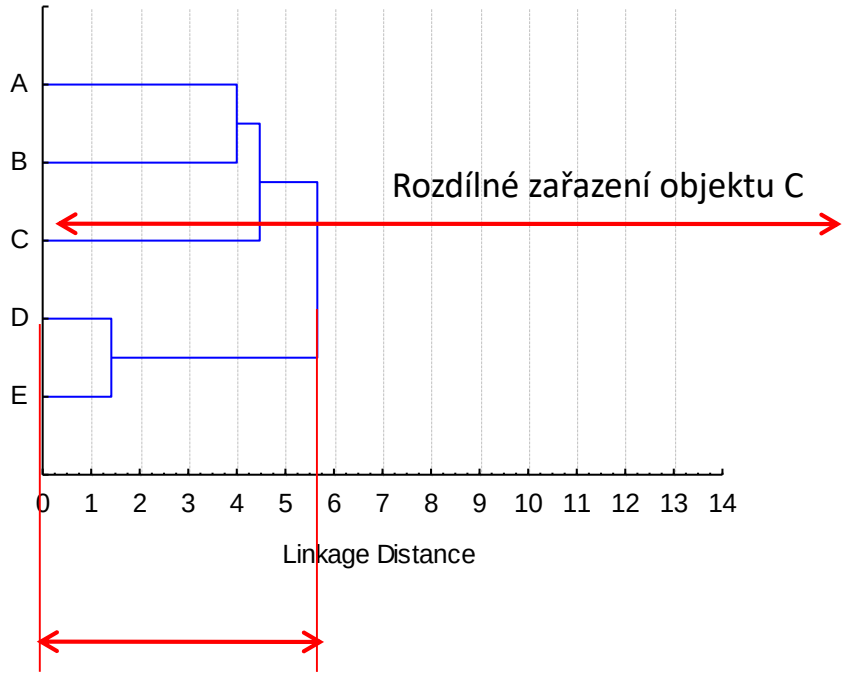




→ dostaneme 3 shluky: (A+B), (C) a (D+E); přičemž linie D a E mají vysoké hodnoty obou markerů, A a B mají nízké hodnoty obou markerů a linie C má střední hodnoty markerů

Metoda nejbližšího a nejvzdálenějšího souseda – interpretace výsledků

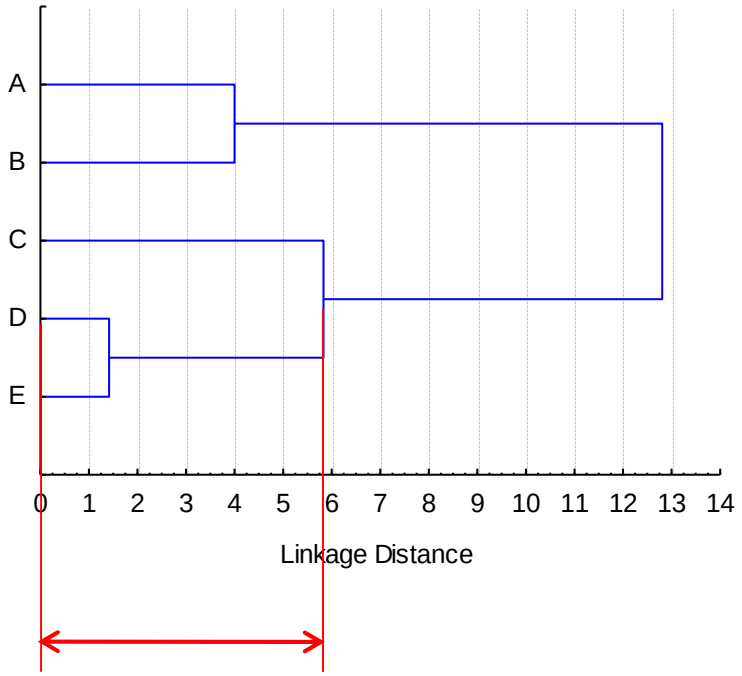
Metoda nejbližšího souseda



Vzdálenost, na níž došlo ke spojení shluku:

• u metody nejbližšího souseda znamená nejmenší vzdálenost objektů shluku, tedy ve shluku mohou existovat objekty s větší vzdáleností

Metoda nejvzdálenějšího souseda



Vzdálenost, na níž došlo ke spojení shluku:

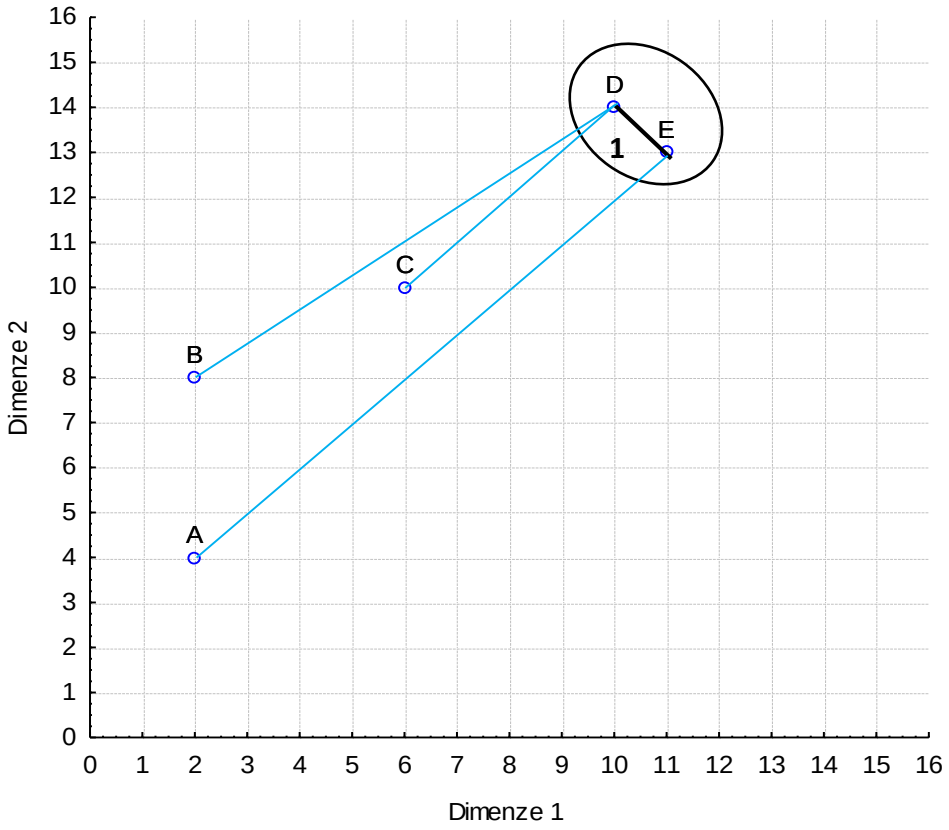
• u metody nejvzdálenějšího souseda znamená největší vzdálenost objektů shluku, tedy objekty ve shluku už mohou být k sobě pouze blíže nebo stejně vzdálené jako je tato vzdálenost

Metoda nejbližšího souseda – doplnění

- Je vypočtena asociační matice, kde objekty D-E již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **nejmenší vzdáleností od jeho členů (D, E)**

	A	B	C	D	E
A	0.0	4.0	7.2	12.8	12.7
B	4.0	0.0	4.5	10.0	10.3
C	7.2	4.5	0.0	5.7	5.8
D	12.8	10.0	5.7	0.0	1.4
E	12.7	10.3	5.8	1.4	0.0

	A	B	C	D+E
A	0.0	4.0	7.2	12.7
B	4.0	0.0	4.5	10.0
C	7.2	4.5	0.0	5.7
D+E	12.7	10.0	5.7	0.0



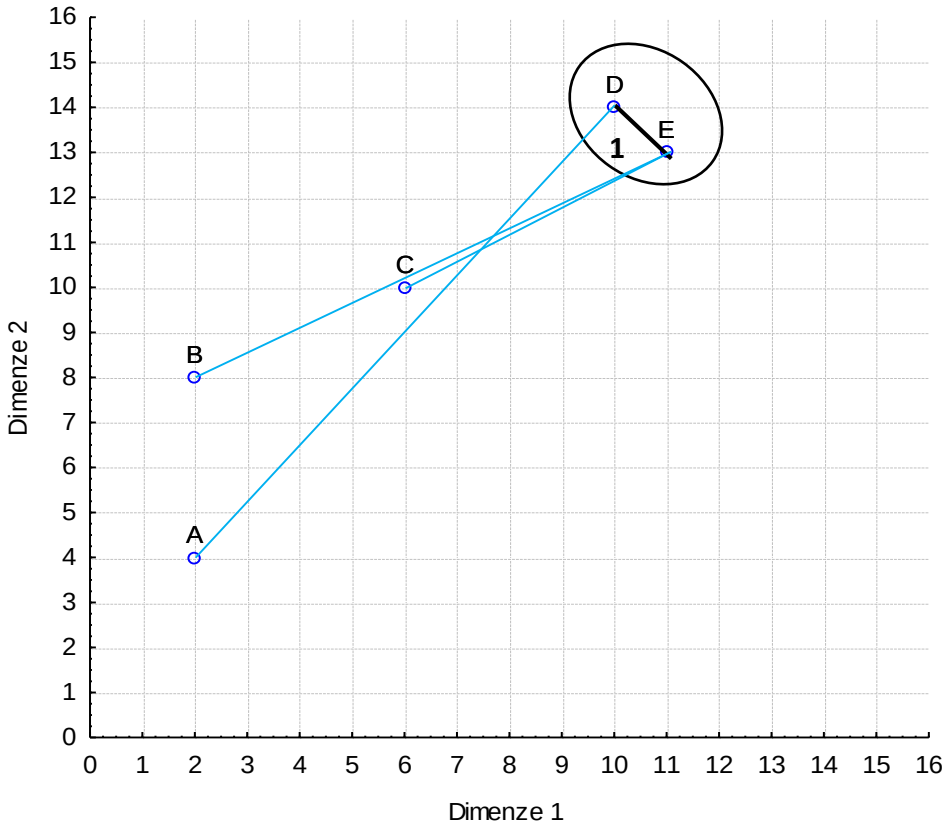
→ vzdálenost A od shluku D+E je dána vzdáleností A od E, protože je menší než vzdálenost A od D
→ vzdálenost B od shluku D+E je dána vzdáleností B od D, protože je menší než vzdálenost B od E
→ vzdálenost C od shluku D+E je dána vzdáleností C od D, protože je menší než vzdálenost C od E

Metoda nejvzdálenějšího souseda – doplnění

- Je vypočtena asociační matice, kde objekty D-E již vystupují jako jeden objekt, jehož vzdálenost od ostatních objektů je dána **největší vzdáleností od jeho členů (D, E)**

	A	B	C	D	E
A	0.0	4.0	7.2	12.8	12.7
B	4.0	0.0	4.5	10.0	10.3
C	7.2	4.5	0.0	5.7	5.8
D	12.8	10.0	5.7	0.0	1.4
E	12.7	10.3	5.8	1.4	0.0

	A	B	C	D+E
A	0.0	4.0	7.2	12.8
B	4.0	0.0	4.5	10.3
C	7.2	4.5	0.0	5.8
D+E	12.8	10.3	5.8	0.0



→ vzdálenost A od shluku D+E je dána vzdáleností A od D, protože je větší než vzdálenost A od E
→ vzdálenost B od shluku D+E je dána vzdáleností B od E, protože je větší než vzdálenost B od D
→ vzdálenost C od shluku D+E je dána vzdáleností C od E, protože je větší než vzdálenost C od D

Příklad 2

Bylo provedeno měření objemu hipokampu a mozkových komor (v cm^3) u 5 pacientů se schizofrenií. Naměřené hodnoty objemu hipokampu a mozkových komor byly zaznamenány do matice $\mathbf{X}_D$:

$$\mathbf{X}_D = \begin{bmatrix} 4{,}6 & 3{,}4 \\ 6{,}1 & 3{,}0 \\ 6{,}7 & 3{,}1 \\ 6{,}2 & 2{,}3 \\ 6{,}9 & 3{,}1 \end{bmatrix}.$$

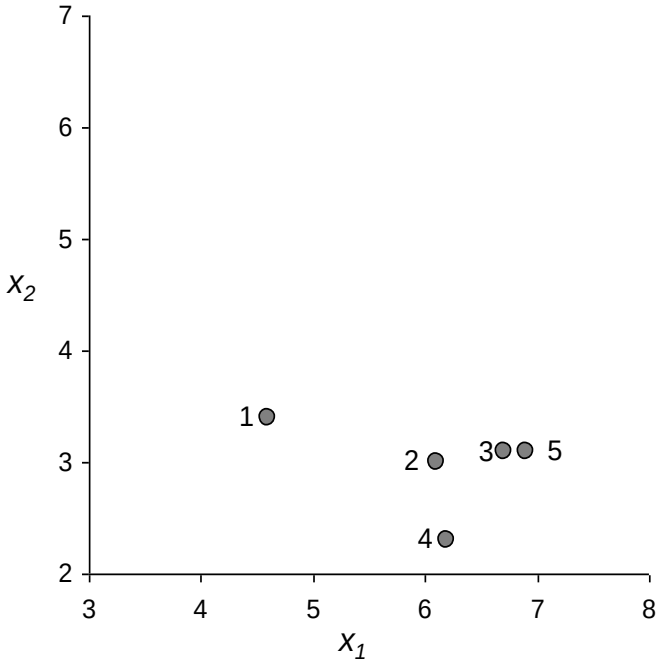
Určete podobnost pěti pacientů na základě naměřených charakteristik pomocí hierarchické shlukové analýzy, použijte metodu nejbližšího a nejvzdálenějšího souseda.

Příklad 2 – asociační matice

Nejprve vypočteme matici vzdáleností mezi objekty založenou na Euklidovské vzdálenosti:

	1	**2**	**3**	**4**	**5**
1	0,0	1,6	2,1	1,9	2,3
2	1,6	0,0	0,6	0,7	0,8
3	2,1	0,6	0,0	0,9	0,2
4	1,9	0,7	0,9	0,0	1,1
5	2,3	0,8	0,2	1,1	0,0

Pro snadnější představu postupu výpočtu si jednotlivé objekty vykreslíme do jednoduchého xy grafu.

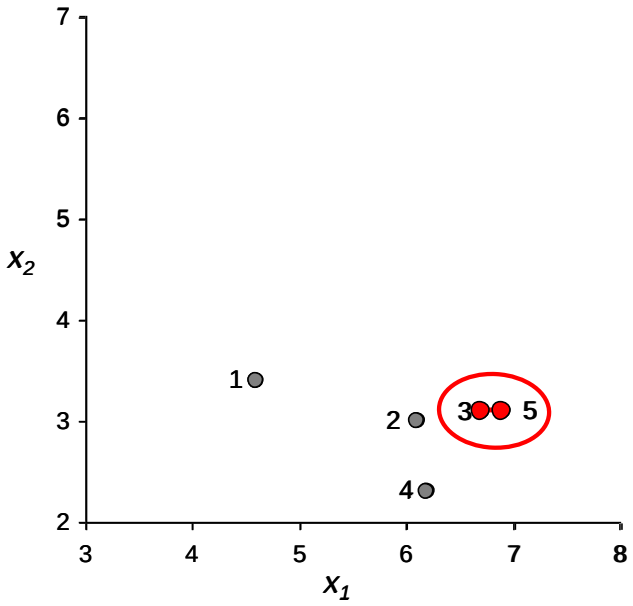


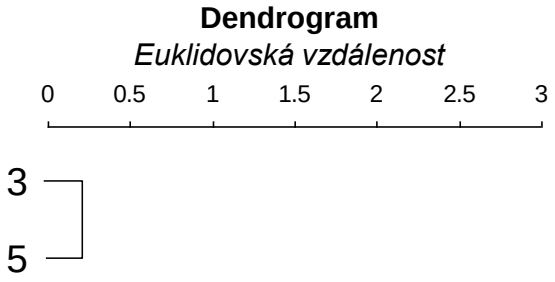
Metoda nejbližšího souseda – krok 1

	1	2	3	4	5
1	0,0	1,6	2,1	1,9	2,3
2	1,6	0,0	0,6	0,7	0,8
3	2,1	0,6	0,0	0,9	0,2
4	1,9	0,7	0,9	0,0	1,1
5	2,3	0,8	0,2	1,1	0,0

→

	1	2	3+5	4
1	0,0	1,6	2,1	1,9
2	1,6	0,0	0,6	0,7
3+5	2,1	0,6	0,0	0,9
4	1,9	0,7	0,9	0,0



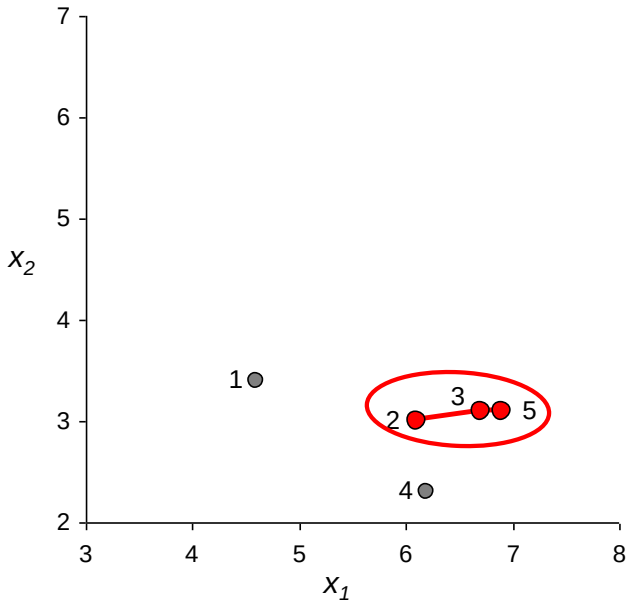


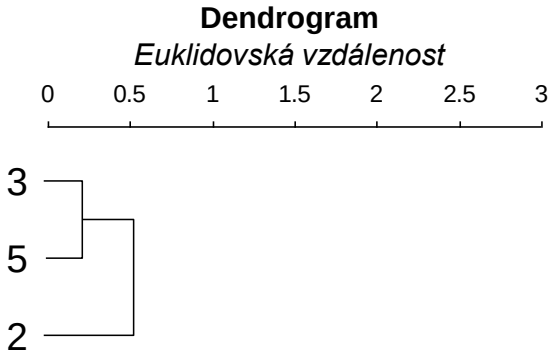
Metoda nejbližšího souseda – krok 2

	1	2	3+5	4
1	0,0	1,6	2,1	1,9
2	1,6	0,0	0,6	0,7
3+5	2,1	0,6	0,0	0,9
4	1,9	0,7	0,9	0,0

→

	1	2+3+5	4
1	0,0	1,6	1,9
2+3+5	1,6	0,0	0,7
4	1,9	0,7	0,0



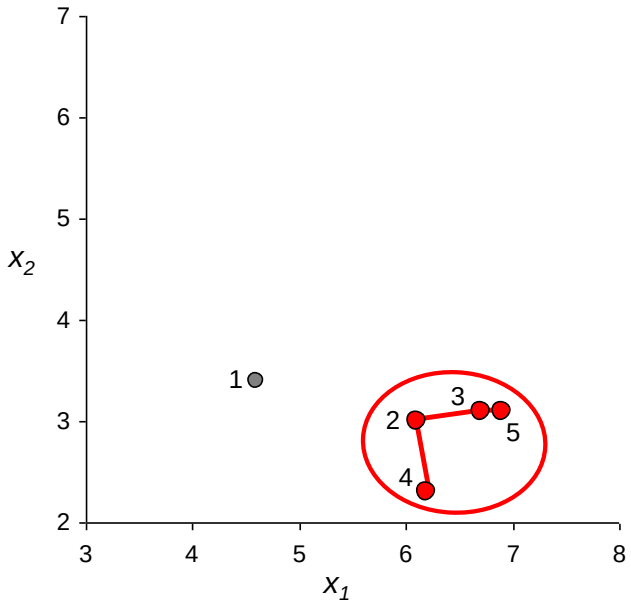


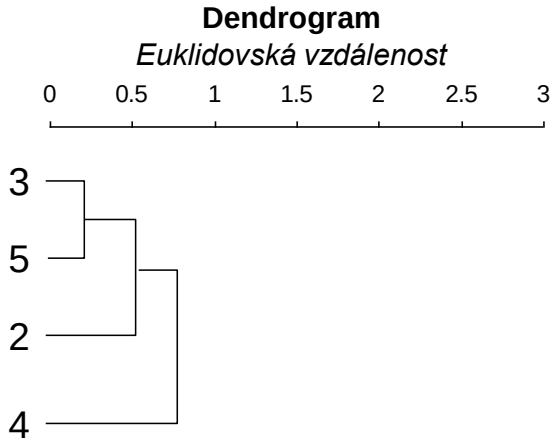
Metoda nejbližšího souseda – krok 3

	1	2+3+5	4
1	0,0	1,6	1,9
2+3+5	1,6	0,0	0,7
4	1,9	0,7	0,0

→

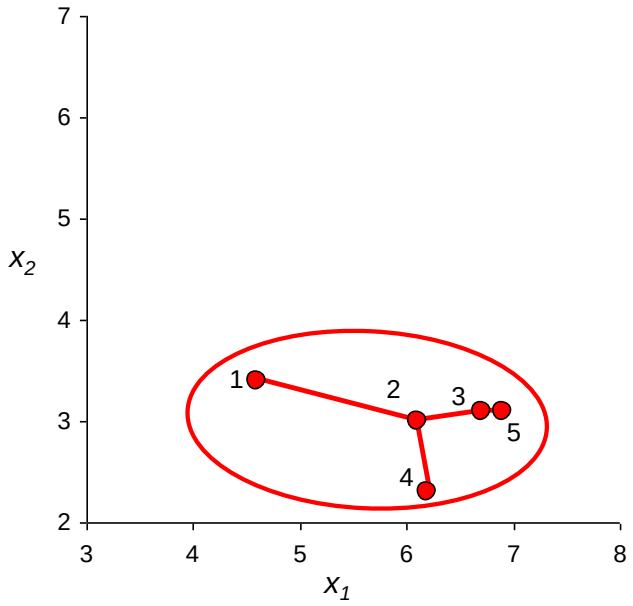
	1	4+2+3+5
1	0,0	1,6
4+2+3+5	1,6	0,0

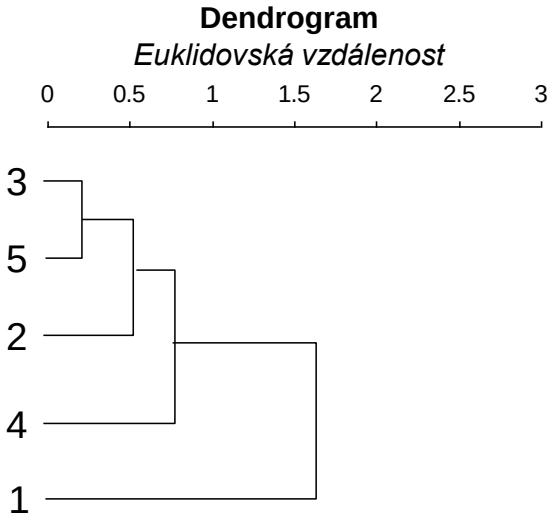




Metoda nejbližšího souseda – krok 4

	1	4+2+3+5
1	0,0	1,6
4+2+3+5	1,6	0,0



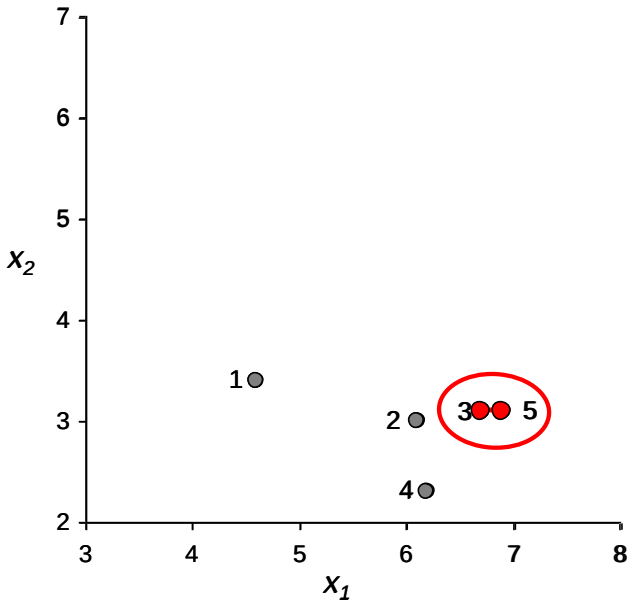


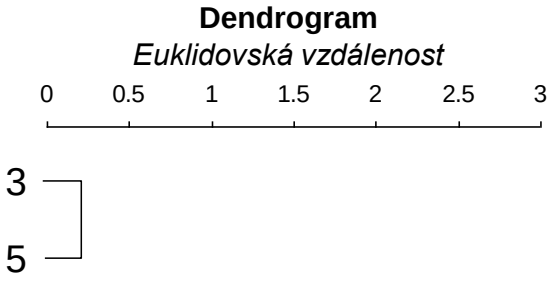
Metoda nejvzdálenějšího souseda – krok 1

	1	2	3	4	5
1	0,0	1,6	2,1	1,9	2,3
2	1,6	0,0	0,6	0,7	0,8
3	2,1	0,6	0,0	0,9	0,2
4	1,9	0,7	0,9	0,0	1,1
5	2,3	0,8	0,2	1,1	0,0

→

	1	2	3+5	4
1	0,0	1,6	2,3	1,9
2	1,6	0,0	0,8	0,7
3+5	2,3	0,8	0,0	1,1
4	1,9	0,7	1,1	0,0



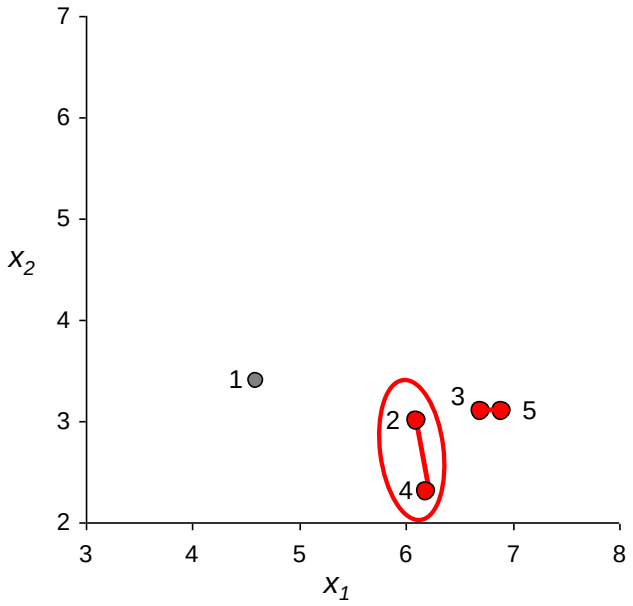


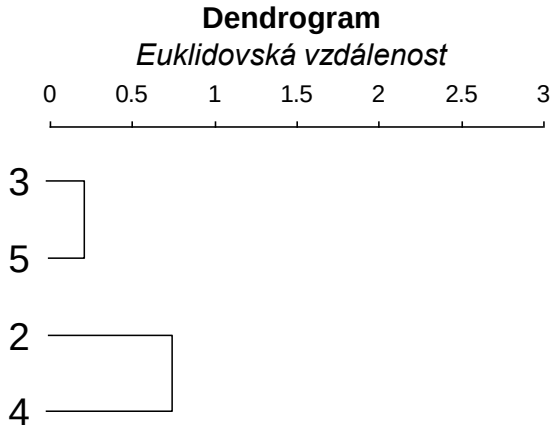
Metoda nejvzdálenějšího souseda – krok 2

	1	2	3+5	4
1	0,0	1,6	2,3	1,9
2	1,6	0,0	0,8	0,7
3+5	2,3	0,8	0,0	1,1
4	1,9	0,7	1,1	0,0

→

	1	2+4	3+5
1	0,0	1,9	2,3
2+4	1,9	0,0	1,1
3+5	2,3	1,1	0,0



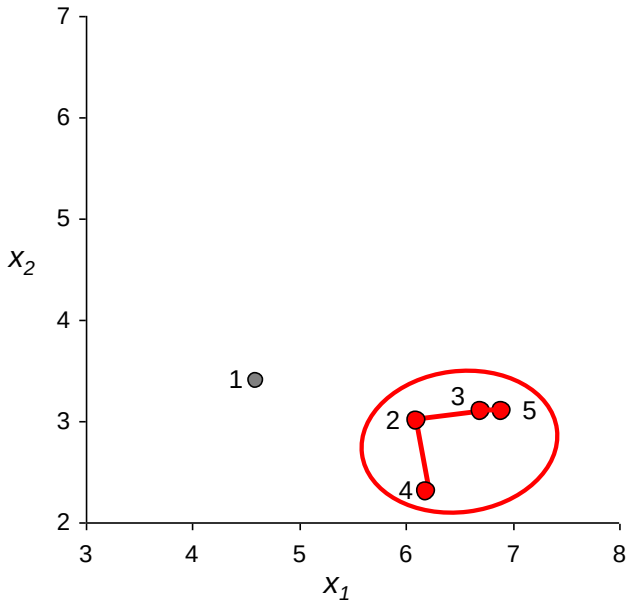


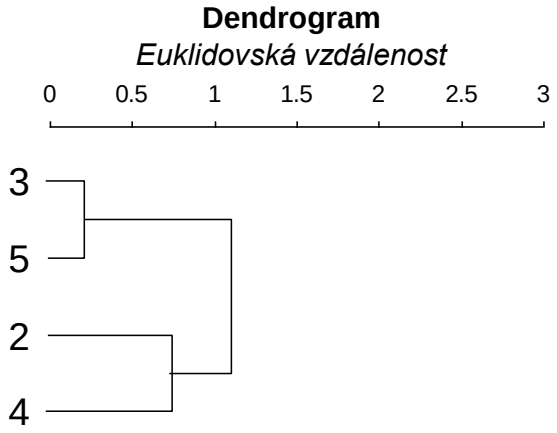
Metoda nejvzdálenějšího souseda – krok 3

	1	2+4	3+5
1	0,0	1,9	2,3
2+4	1,9	0,0	1,1
3+5	2,3	1,1	0,0

→

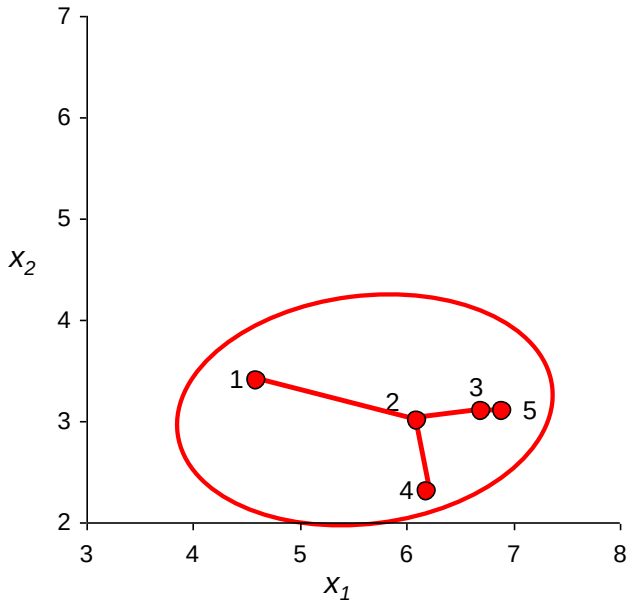
	1	4+2+3+5
1	0,0	2,3
4+2+3+5	2,3	0,0

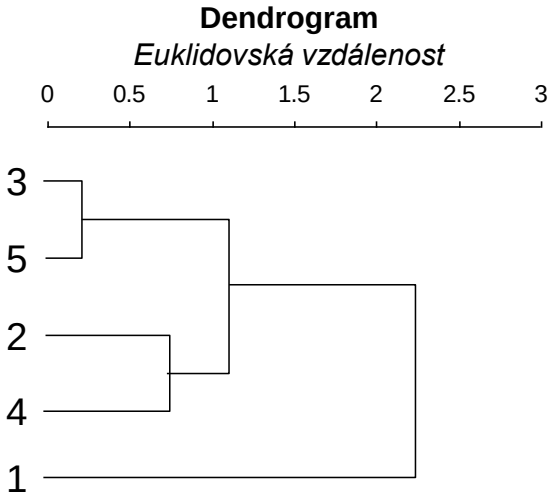




Metoda nejvzdálenějšího souseda – krok 4

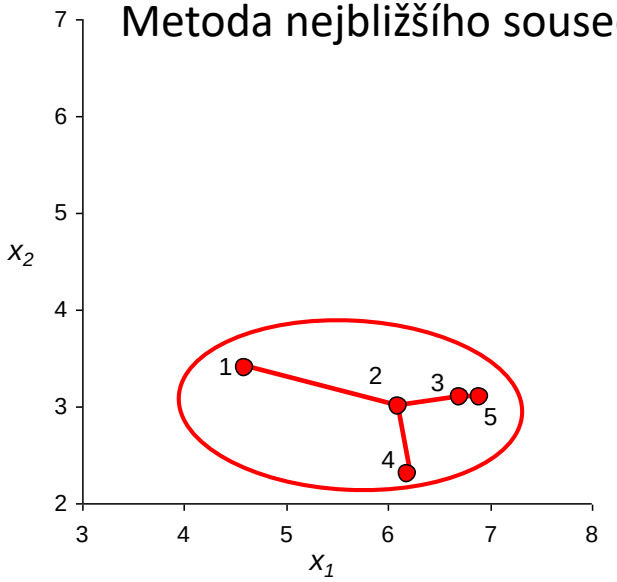
	1	4+2+3+5
1	0,0	2,3
4+2+3+5	2,3	0,0

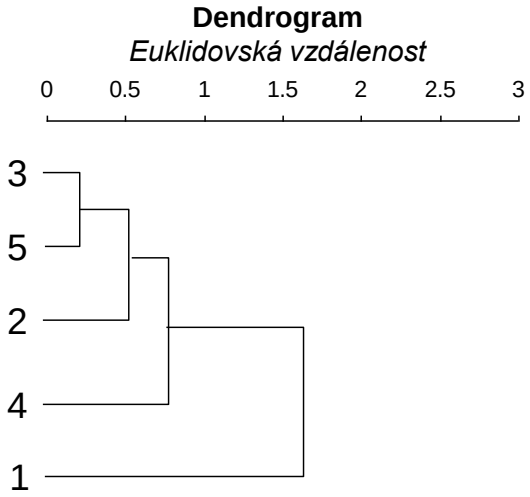




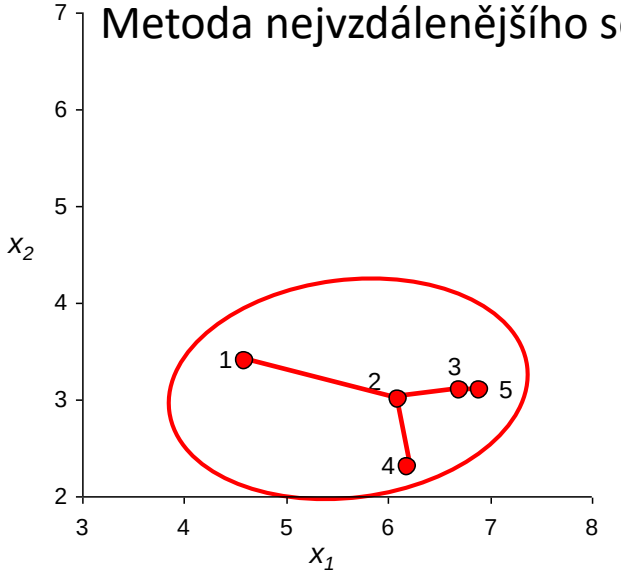
Srovnání metody nejbližšího a nejvzdálenějšího souseda

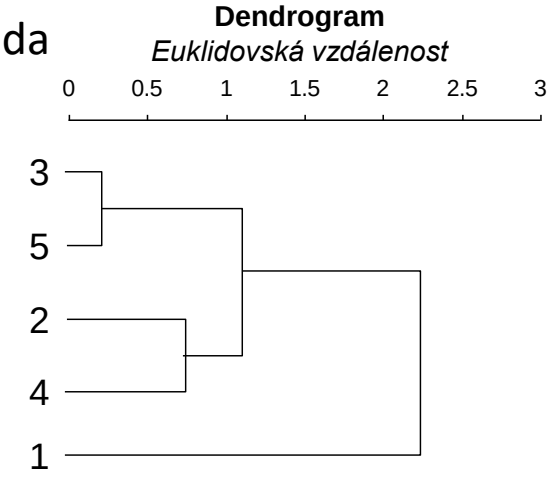
Metoda nejbližšího souseda





Metoda nejvzdálenějšího souseda





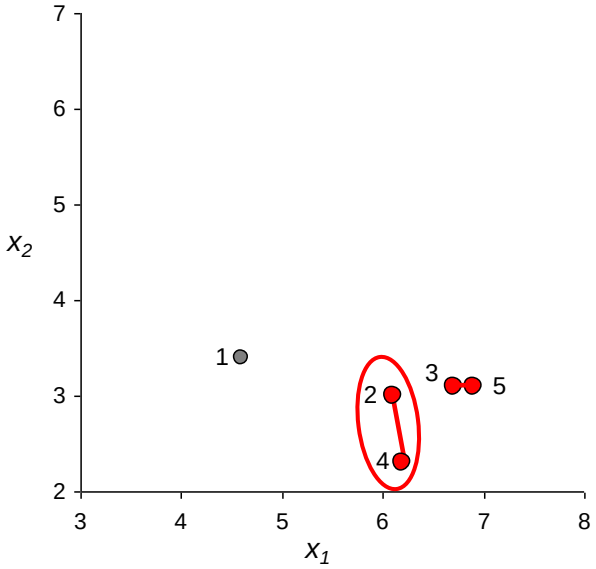
→ metoda nejbližšího souseda má tendenci vytvářet protáhlé shluky

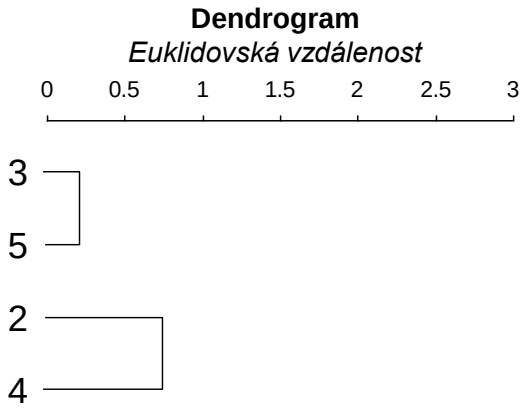
Metoda nejvzdálenějšího souseda – doplnění

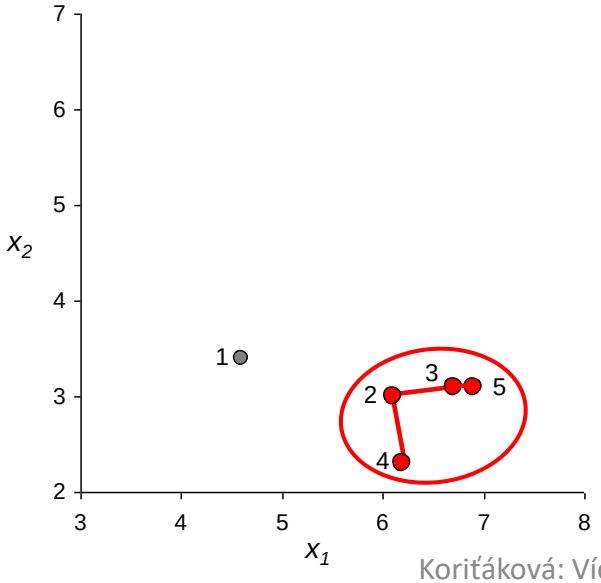
	1	2+4	3+5
1	0,0	1,9	2,3
2+4	1,9	0,0	1,1
3+5	2,3	1,1	0,0

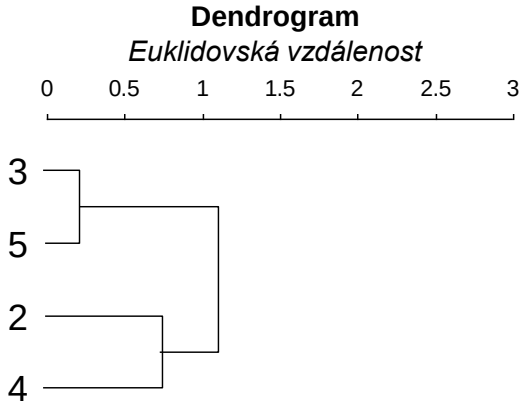
	1	4+2+3+5
1	0,0	2,3
4+2+3+5	2,3	0,0

→ došlo ke spojení shluku 2+4 a 3+5 na vzdálenosti 1,1, což je vzdálenost subjektu 4 a subjektu 5, protože tato vzdálenost je ze všech vzdáleností 4→5, 2→5, 4→3, 2→3 největší







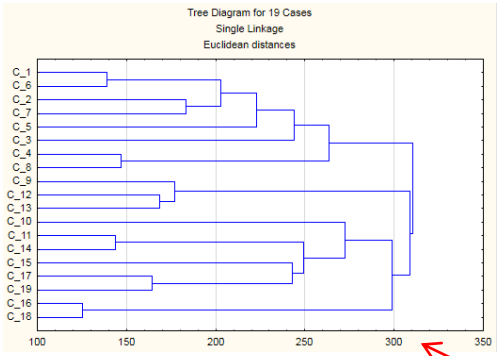


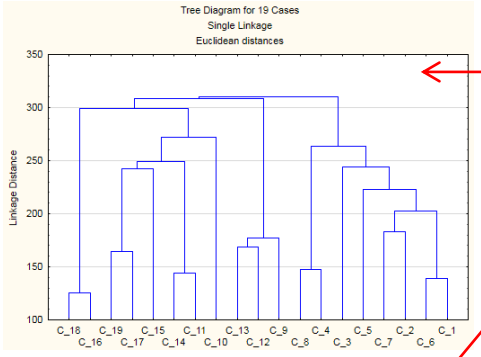
Výpočet shlukové analýzy v softwarech

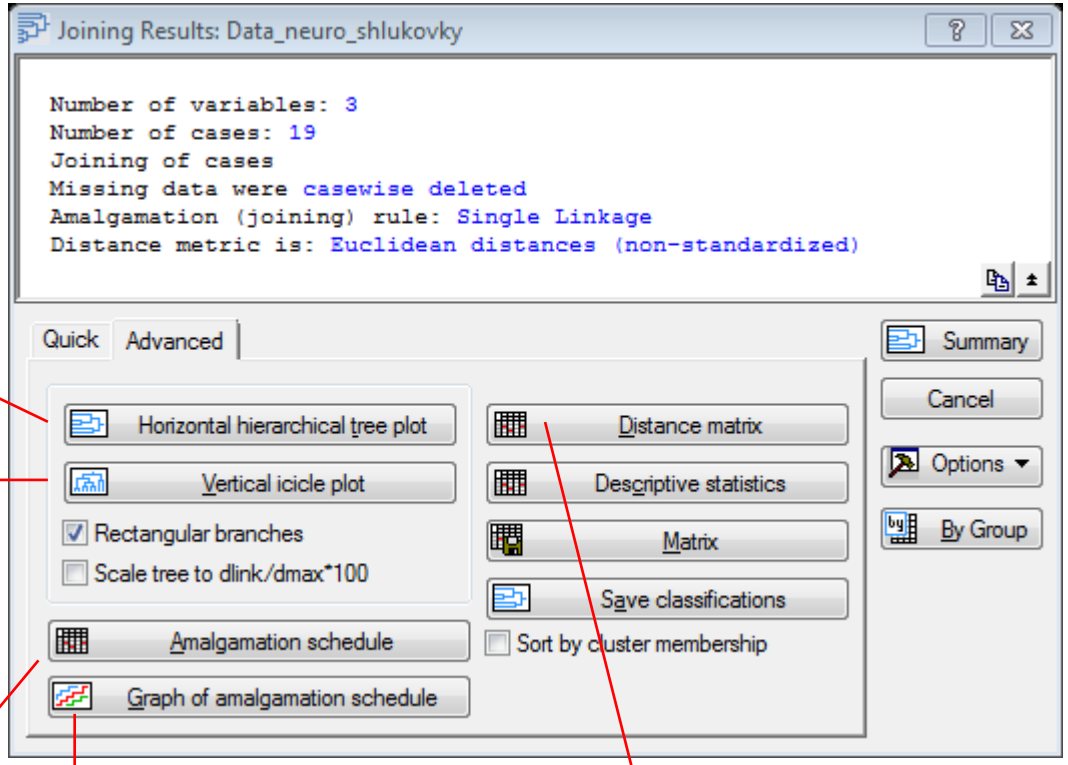
STATISTICA – hierarchické aglomerativní shlukování

- Statistics – Multivariate Exploratory Techniques – Cluster Analysis – Joining (tree clustering) – OK
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables (columns)) či subjekty (Cases (rows))
- Amalgamation (linkage) rule = volba shlukovacího algoritmu:
  - Single Linkage – metoda nejbližšího souseda
  - Complete Linkage – metoda nejvzdálenějšího souseda
  - Unweighted pair-group average – metoda průměrné vazby (nevážená)
  - Weighted pair-group average – metoda průměrné vazby (vážená)
  - Unweighted pair-group centroid – centroidová metoda (nevážená)
  - Weighted pair-group centroid (median) – centroidová metoda (vážená) = mediánová metoda
  - Ward's method – Wardova metoda
- Distance measure = volba metrik vzdáleností objektů (subjektů):
  - Squared Euclidean distances – čtverec Euklidovy vzdálenosti
  - Euclidean distances – Euklidova metrika
  - City-block (Manhattan) distances – Hammingova (manhattanská) metrika
  - Chebychev distance metric – Čebyševova metrika
  - Power: SUM(ABS(x-y)**p)**1/r – pokud r=p, jde o Minkovského metriku
  - Percent disagreement
  - 1-Pearson r – jedna mínus Pearsonův korelační koeficient

STATISTICA – hierarch. aglom. shluk. – pokračování

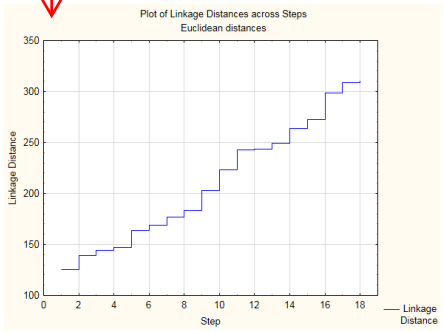






Amalgamation Schedule (Data_neuro_shlukovky)
Single Linkage
Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	O
125.1972	C_16	C_18				
139.2640	C_1	C_6				
143.9270	C_11	C_14				
147.0873	C_4	C_8				
164.1363	C_17	C_19				
168.6528	C_12	C_13				
176.9954	C_9	C_12	C_13			
183.2707	C_2	C_7				
202.7584	C_1	C_6	C_2	C_7		
223.0460	C_1	C_6	C_2	C_7	C_5	



asociační matice Euklidových vzdáleností

Euclidean distances (Data_neuro_shlukovky)

Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C
C_1	0	291	299	490	271	139	
C_2	291	0	244	264	454	251	
C_3	299	244	0	500	527	311	
C_4	490	264	500	0	535	410	
C_5	271	454	527	535	0	223	
C_6	139	251	311	410	223	0	
C_7	307	183	262	328	399	203	

STATISTICA – nehierarchické shlukování

- Statistics – Multivariate Exploratory Techniques – Cluster Analysis – K-means clustering – OK – přepnout se na záložku Advanced
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables (columns)) či subjekty (Cases (rows))
- Number of clusters: zvolit počet shluků (např. 3)
- Number of iterations: volba počtu iterací (metoda *k*-průměrů je iterativní metoda)
- Initial cluster centers: volba počátečních středů shluků

- příslušnost jednotlivých subjektů do shluků nalezneme na záložce Advanced v „Members of each cluster & distances“

SPSS – hierarchické aglomerativní shlukování

- Analyze – Classify – Hierarchical Cluster...
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables) či subjekty (Cases)
- Statistics...: zatrhnout Proximity matrix (= asociační matice vzdáleností či podobností)
- Plots...: zatrhnout Dendrogram (možnost volby Vertical či Horizontal)
- Method...:
  - Cluster Method = volba shlukovacího algoritmu:
    - Between-groups linkage – metoda průměrné vazby mezi skupinami
    - Within-groups linkage – metoda průměrné vazby uvnitř skupin
    - Nearest neighbor – metoda nejbližšího souseda
    - Furthest neighbor – metoda nejvzdálenějšího souseda
    - Centroid clustering – centroidová metoda (nevážená)
    - Median clustering – centroidová metoda (vážená) = mediánová metoda
    - Ward's method – Wardova metoda
  - Distance measure: volba metrik vzdáleností objektů (subjektů):
    - Euclidean distance – Euklidova metrika
    - Squared Euclidean distance – čtverec Euklidovy vzdálenosti
    - Cosine – kosinová metrika
    - Pearson correlation – Pearsonův korelační koeficient
    - Chebychev – Čebyševova metrika
    - Block – Hammingova (manhattanská) metrika
    - Minkowski – Minkovského metrika
    - Customized – výpočet pomocí SUM(ABS(x-y)**p)**1/r
  - Transform Values, Transform Measure – je možno transformovat původní data nebo vypočtené vzdálenosti

SPSS – nehierarchické shlukování

- Analyze – Classify – K-Means Cluster...
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Number of clusters: zvolit počet shluků (např. 3)
- Method: přepnout na „Classify only“ v případě, že známe středy shluků, které můžeme načíst pomocí „Read initial“
- Iterate... – Maximum Iterations (volba počtu iterací – metoda *k*-průměrů je iterativní metoda)
- Options... – zatrhnout „Cluster information for each case“, abychom získali tabulku, do kterého shluku patří který subjekt

Software R – hierarchické aglomerativní shlukování

- funkce *dist* na výpočet vzdáleností objektů (či subjektů) :
  - „euclidean“ – Euklidovska metrika
  - „maximum“ – Čebyševova metrika
  - „manhattan“ – Hammingova (manhattanská) metrika
  - „canberra“ – Canberrská metrika
  - „minkowski“ – Minkovského metrika
- funkce *hclust* na výpočet shlukové analýzy:
  - „ward.D“ a „ward.D2“ – dva algoritmy pro Wardovu metodu
  - „single“ – metoda nejbližšího souseda (single linkage)
  - „complete“ – metoda nejvzdálenějšího souseda (complete linkage)
  - „average“ – metoda průměrné vazby (nevážená) (average linkage)
  - „mcquitty“ – metoda průměrné vazby (vážená)
  - „median“ – centroidová metoda (vážená) = mediánová metoda
  - „centroid“ – centroidová metoda (nevážená)
- podrobná ukázka v souboru Shlukovky_skript.R

Software R – nehierarchické shlukování

- funkce *kmeans*
- ukázka:

```
cl <- kmeans(data.vyber, 3)    # provedeni shlukove analyzy
table(cl$cluster,groupCodes) # zjisteni, kolik subjektu bylo spatne zarazenych
```

Matlab – hierarchické aglomerativní shlukování

- funkce *linkage*, která umožňuje volbu shlukovacího algoritmu i volbu metriky vzdálenosti mezi objekty (subjekty)
- volba shlukovacího algoritmu:
  - „average“ – metoda průměrné vazby (nevážená) (average linkage)
  - „centroid“ – centroidová metoda (nevážená)
  - „complete“ – metoda nejvzdálenějšího souseda (complete linkage)
  - „median“ – centroidová metoda (vážená) = mediánová metoda
  - „single“ – metoda nejbližšího souseda (single linkage)
  - „ward“ – Wardova metoda
  - „weighted“ – metoda průměrné vazby (vážená)
- volba metriky vzdáleností – stejná nabídka jako u funkce *pdist*
- ukázka:

```
[num, txt] = xlsread('Data_neuro_shlukovky.xlsx',1);
data=num(:,[23,24,26]);

Z=linkage(data,'complete','euclidean');    % provedeni shlukove analyzy
dendrogram(Z)                               % vykresleni dendrogramu

c=cluster(Z,'maxclust',3);  % vytvoreni definovaneho poctu shluku
crosstab(c,num(:,3))        % zjisteni, kolik subjektu bylo spatne zarazenych
```

Matlab – nehierarchické shlukování

- funkce *kmeans*
- ukázka:

```
[idx,C]=kmeans(data,3);   % provedeni shlukove analyzy (matice C – centroidy skupin)
crosstab(idx,num(:,3))  % zjisteni, kolik subjektu bylo spatne zarazenych
```

- funkce *kmedoids*
- bohužel není ve starých verzích Matlabu
- ukázka:

```
[idx,C]=kmedoids(data,3);   % provedeni shlukove analyzy (matice C – medoidy skupin)
crosstab(idx,num(:,3))  % zjisteni, kolik subjektu bylo spatne zarazenych
```