Pokročilé metody analýzy dat v neurovědách

RNDr. Eva Koriťáková, Ph.D.

doc. RNDr. Ladislav Dušek, Dr.

Blok 4

Shluková analýza

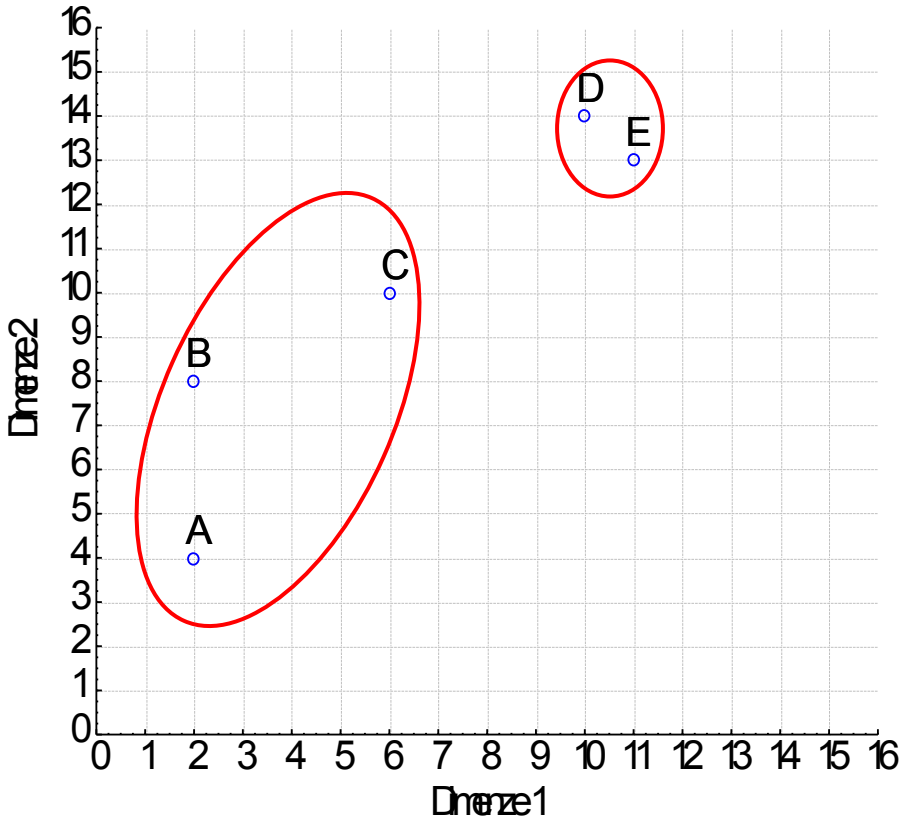
Osnova

1. Podstata a cíle shlukové analýzy dat
2. Shluková analýza hierarchická – hierarchické aglomerativní shlukování
3. Shluková analýza hierarchická – hierarchické divizivní shlukování
4. Shluková analýza nehierarchická
5. Identifikace optimálního počtu shluků

Podstata a cíle shlukové analýzy dat

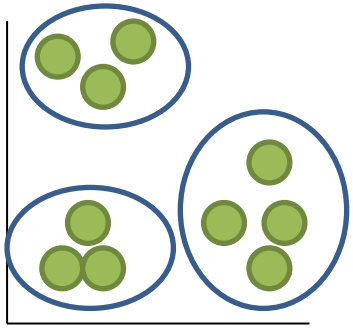
Shluková analýza – cíle a postupy

- Shluková analýza se snaží o identifikaci shluků objektů ve vícerozměrném prostoru a následnou redukci vícedimenzionálního problému kategorizací objektů do zjištěných shluků
- Existuje řada různých metod pro shlukování dat lišících se:
  - Měřením vzdálenosti mezi objekty
  - Algoritmem spojování objektů do shluků
  - Interpretací výstupů
- Každá z metod má své vlastní předpoklady výpočtu a je nasaditelná pro různé typy úloh
- Porušení předpokladů nebo nasazení chybné metody může vést k zavádějícím výsledkům

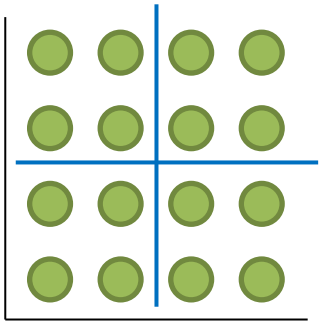


Obecný princip hledání shluků v datech

- Vzájemnou pozici objektů ve vícerozměrném prostoru lze popsat jejich vzdáleností (např. Euklidovou, Čebyševovou apod.)
- Smysluplnost výsledků shlukování závisí jednak na objektivní existenci shluků v datech, jednak na arbitrárně nastavených kritériích definice shluků



Jednoznačné odlišení existujících shluků v datech (obdoba multimodálního rozložení)



Shluková analýza je možná i v tomto případě, nicméně hranice shluků jsou dány pouze naším rozhodnutím.

Shluková analýza – typy metod

Shluková analýza

Hierarchické
shluky jsou definovány postupným skládáním objektů

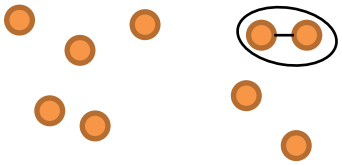
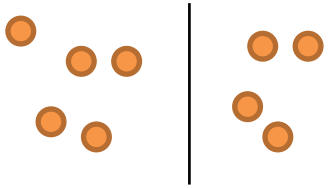
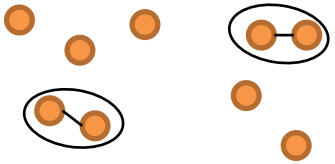
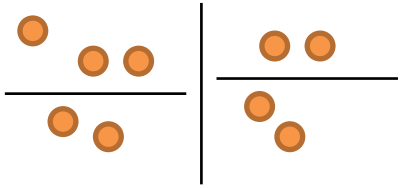
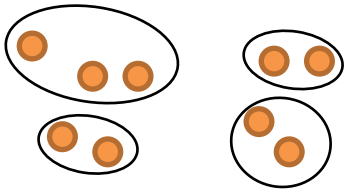
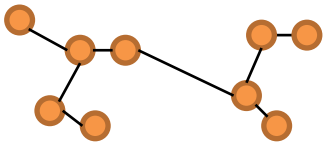
Nehierarchické
shluky jsou definovány v jednom kroku

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

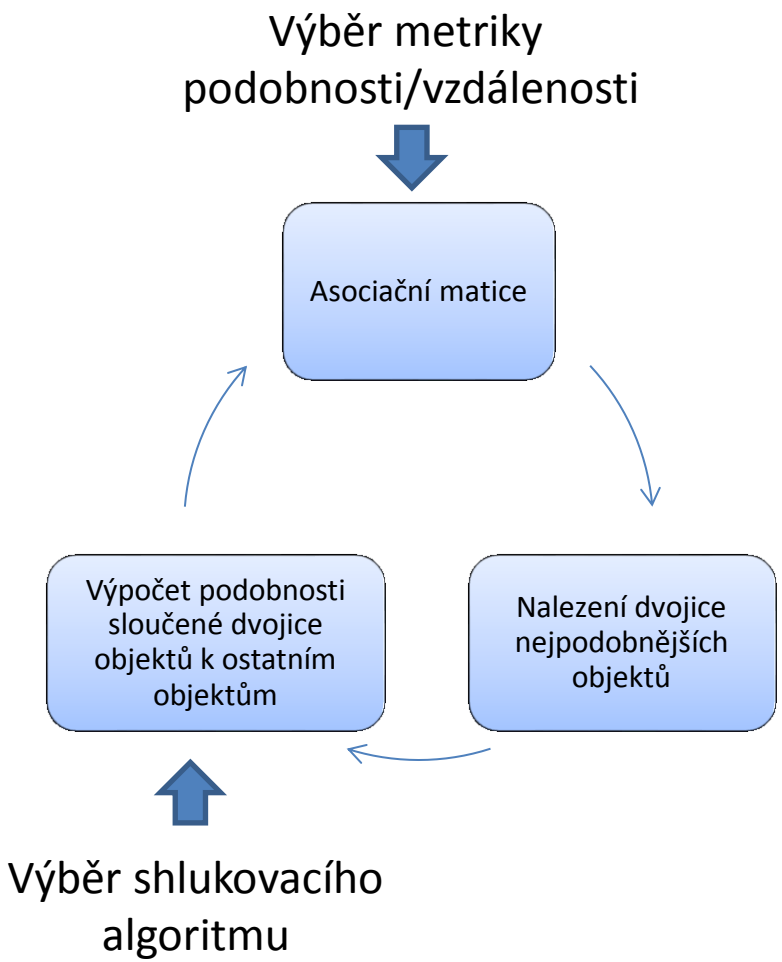
1. Krok			Kolik shluků chceme definovat? Například 4	Minimum spanning tree, Prime network
2. Krok				
X. Krok	Atd.	Atd.	Výpočet ukončen	Výpočet ukončen

Shluková analýza hierarchická – hierarchické aglomerativní shlukování

Hierarchické aglomerativní shlukování

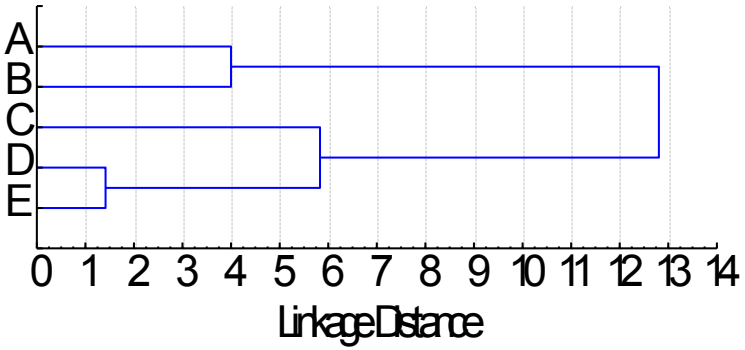
- Při tomto způsobu shlukování jsou postupně shlukovány nejpodobnější objekty až do doby, kdy jsou všechny objekty propojeny do jednoho shluku spojujícího všechny objekty v analyzovaném souboru
- Analýza má dva hlavní kroky:
  1. Výběr vhodné metriky vzdálenosti/podobnosti pro výpočet asociační matice (analýza může probíhat na libovolných metrikách vzdálenosti/podobnosti)
  2. Výběr shlukovacího algoritmu, který podstatným způsobem ovlivňuje výsledky analýzy a možnosti její interpretace
- Algoritmus výpočtu postupuje v následujícím cyklu
  1. Výpočet asociační matice
  2. Spojení dvou nejpodobnějších objektů
  3. Přepočítání asociační matice tak, že spojené objekty již nadále vystupují jako jediný objekt (v tomto kroku se uplatňuje zvolený shlukovací algoritmus, který definuje, jak bude počítána vzdálenost/podobnost spojených objektů vůči ostatním objektům)
  4. Spojení dvou nejpodobnějších objektů z přepočítané asociační matice
  5. Atd. až do spojení všech objektů

Hierarchické aglomerativní shlukování – schéma výpočtu



Ukončení výpočtu po spojení všech objektů

Dendrogram

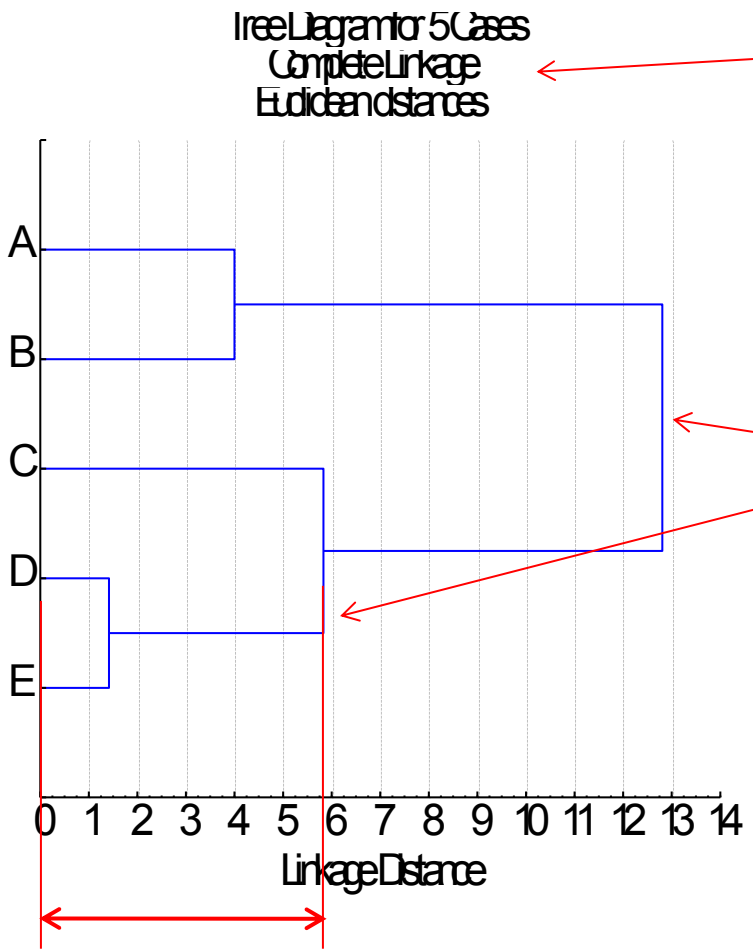


Amalgamation schedule/graph

Amalgamation Schedule (clustering_demo)
Complete Linkage
Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5
1.414214	D	E			
4.000000	A	B			
5.830952	C	D	E		
12.80625	A	B	C	D	E

Popis výstupů – dendrogram



Výstupy shlukové analýzy musí být vždy popsány použitou metrikou vzdáleností a shlukovacím algoritmem

Shlukované objekty - jejich pořadí je dáno přiřazením do shluků, není problém jejich pořadí v grafu měnit (např. v tomto konkrétním grafu prohodit A a B), pouze nesmí dojít ke změně shluků

Propojení shlukovaných objektů

Vzdálenost, na níž došlo ke spojení shluku:

- je v rozměrech použité metriky vzdáleností/podobností a v tomto kontextu ji lze kvantitativně interpretovat
- interpretace vzdálenosti shlukování se liší podle použitého shlukovacího algoritmu
- někdy se uvádí ve škále 0-100%, kde 100% je maximální vzdálenost shlukování

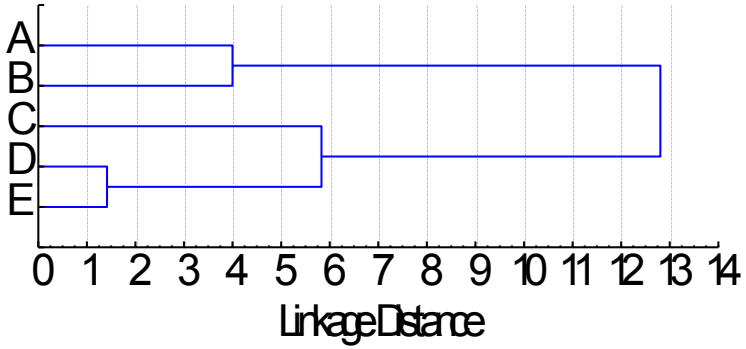
Popis výstupů – Amalgamation schedule/graph

- Popis postupu shlukování; využitelné i pro identifikaci optimálního počtu shluků

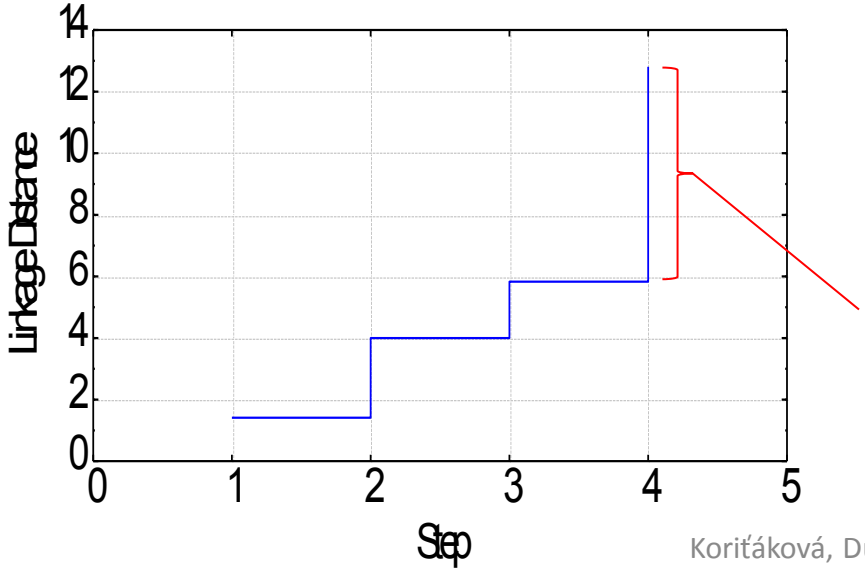
	Amalgamation Schedule (clustering_demo) Complete Linkage Euclidean distances				
linkage distance	**Obj. No. 1**	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5
1.414214	D	E			
4.000000	A	B			
5.830952	C	D	E		
12.80625	A	B	C	D	E

Objekty spojené v jednotlivých krocích shlukování

Souvislost s dendrogramem:



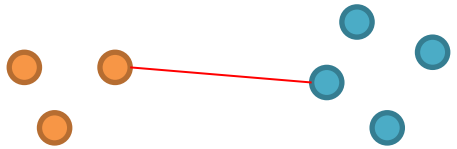
Grafické vyjádření kroků shlukování a vzdálenostech, na nichž došlo k propojení objektů:



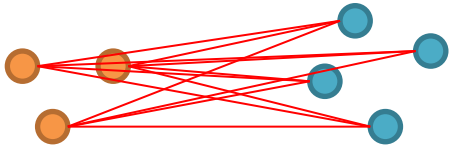
Pokud je v grafu dlouhá vzdálenost bez napojení shluku, jde o možné místo zastavení shlukování a definici finálních shluků

Shlukovací algoritmy hierarchického aglomerativního shlukování

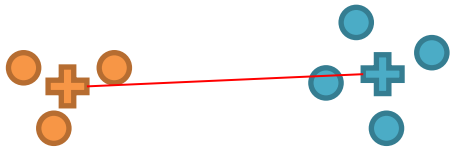
- **Metoda nejbližšího souseda** (jednospojná metoda, metoda jediné vazby, metoda krátké ruky, *nearest neighbour, simple linkage*) – spojení dle nejmenší vzdálenosti mezi objekty shluků



- **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



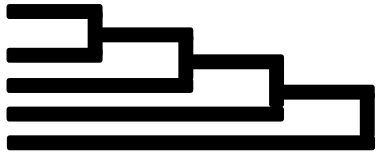
- **Centroidová metoda** (centroidní metoda, metoda středospojné vzdálenosti, Gowerova metoda, *centroid method*) – spojení dle vzdálenosti centroidů shluků
  - Nevážená (*unweighted, UPGMC*) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků
  - Vážená (*weighted, WPGMC*, mediánová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



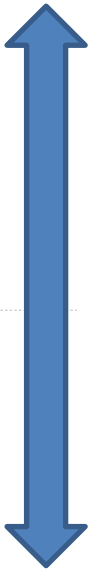
- **Metoda nejvzdálenějšího souseda** (všespojná metoda, metoda dlouhé ruky, *furthest neigbour, complete linkage*) – spojení dle největší vzdálenosti mezi objekty shluků

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

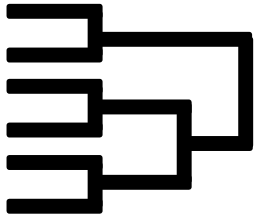


- **Metoda průměrné vazby** (středospojná metoda, *average linkage*) – spojení dle průměrné vzdálenosti mezi objekty shluků
  - Nevážená (*unweighted, UPGMA*) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků
  - Vážená (*weighted, WPGMA*) – odstranění vlivu velikosti shluků, shluky 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



Přechod mezi oběma extrémy (metoda flexible clustering umožňuje dle nastavení zcela plynulý přechod)

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

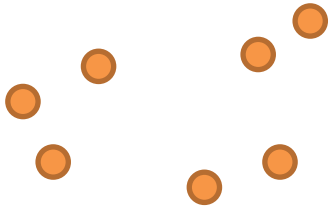


Shlukovací algoritmy hierarchického aglomerativního shlukování – Wardova metoda

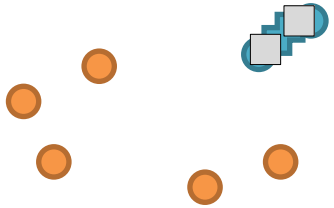
- Principielně podobné ANOVA
- Shluky jsou vytvářeny tak, aby nově vzniklý shluk přispíval co nejméně k sumě čtverců vzdáleností objektů od centroidů jejich shluků
- V počátečním kroku je každý objekt sám sobě shlukem, a tedy vzdálenost od centroidu shluku je rovna 0

- Pro výpočet vzdáleností od centroidu je používána Euklidovská vzdálenost
- Pro popis vzdálenosti shlukování je v dendrogramu možné použít řadu postupů (nezbytné ověřit, jaký přístup je k dispozici v použitém SW):
  - Čtverce vzdáleností
  - Odmocnina čtverce vzdáleností
  - Podíl variability (čtverce vzdáleností) připadající na daný shluk
  - Aj.

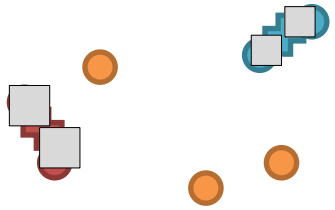
Krok 1: každý objekt je sám sobě centroidem



Krok 2: spojení objektů, které nejméně přispějí k sumě čtverců vzdáleností od centroidu



Krok 3: spojení objektů, které nejméně přispějí k sumě čtverců vzdáleností od centroidu



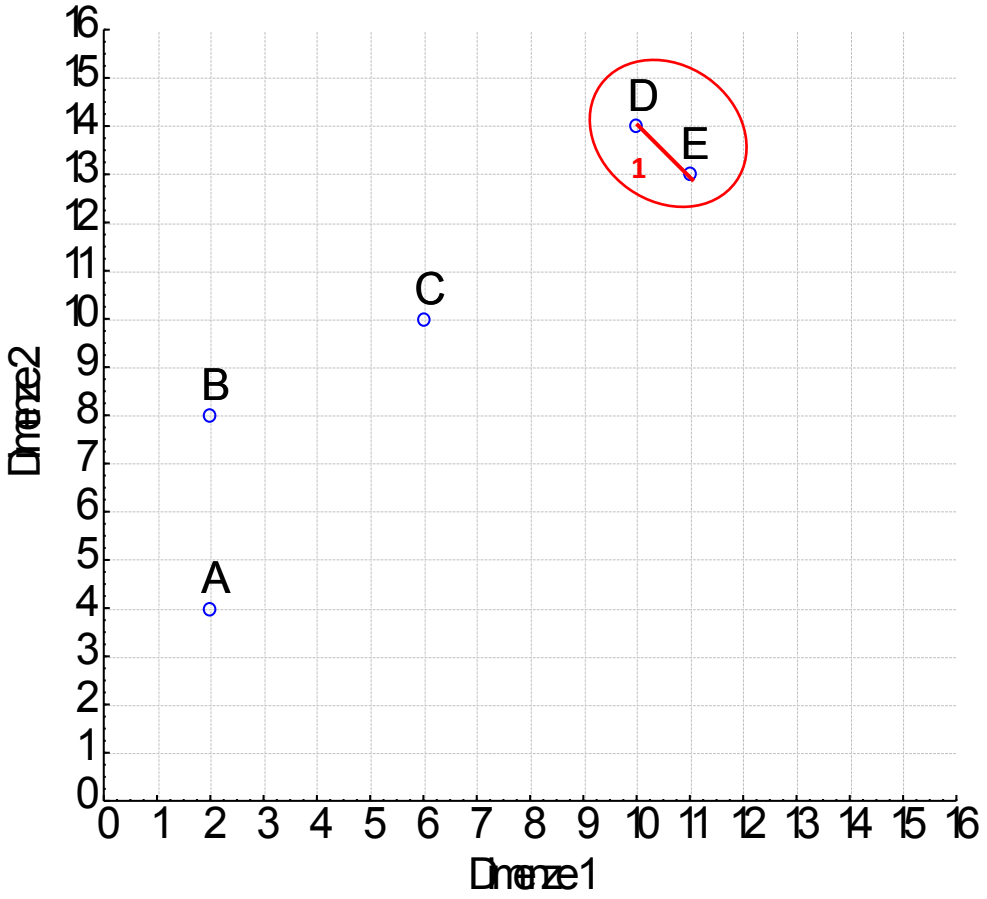
Krok 4: stejný postup až do spojení všech objektů

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

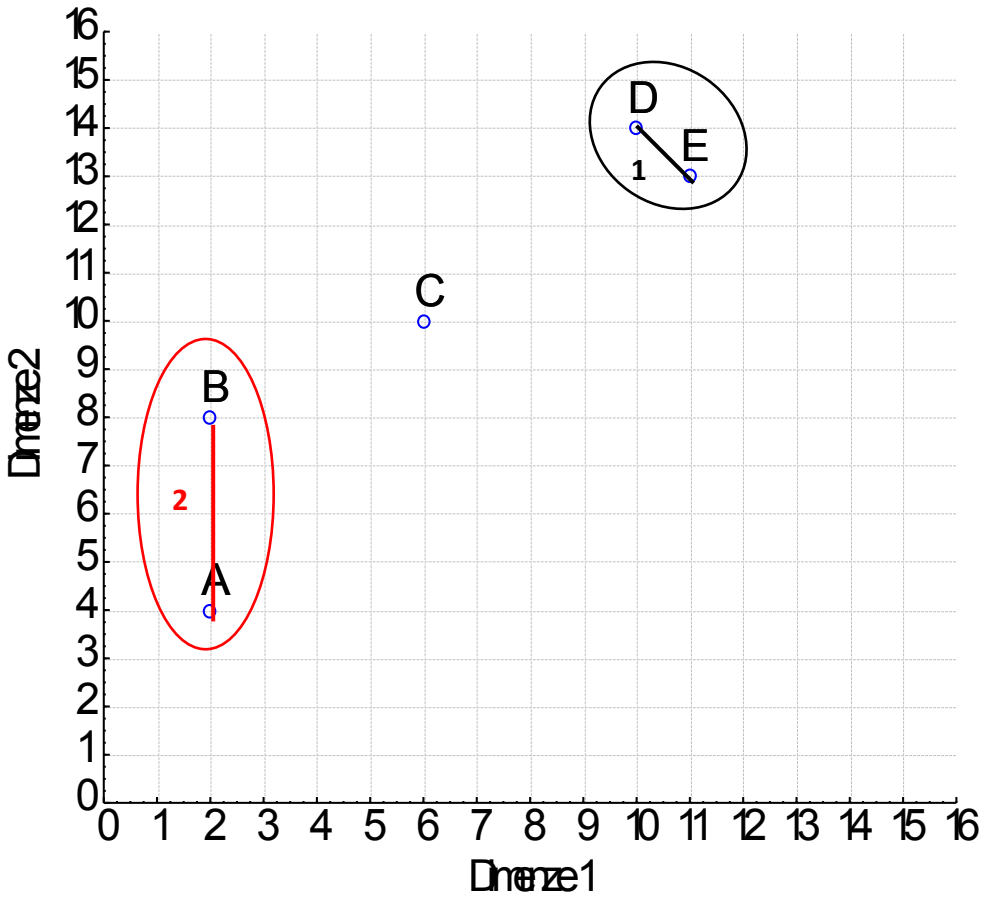


Metoda nejbližšího souseda: 2. krok výpočtu

- 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



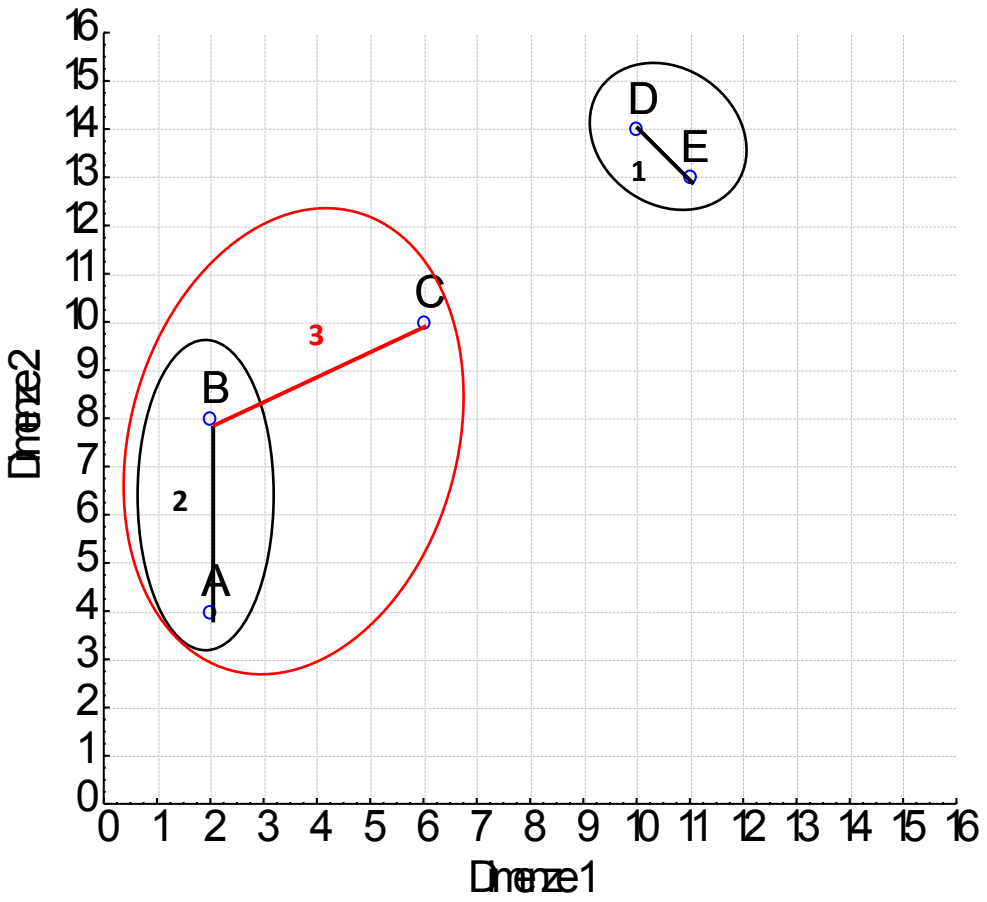
- Je definován shluk dvou nejbližších objektů A-B

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

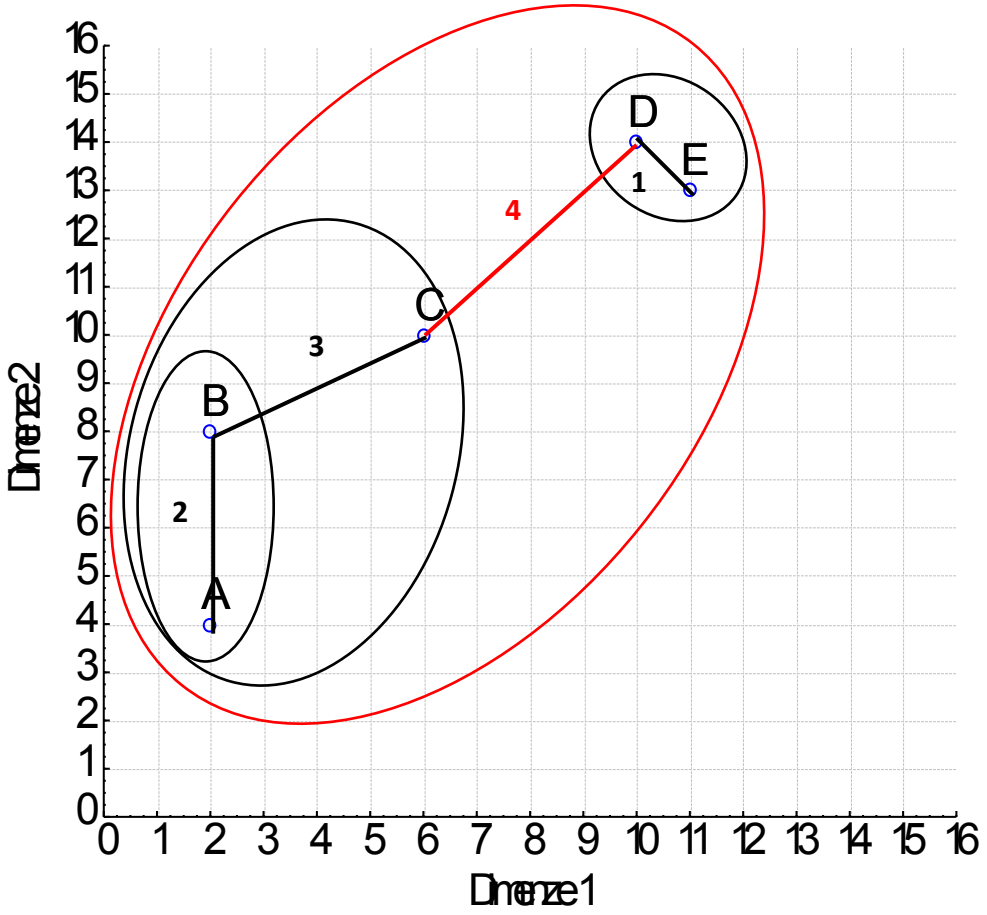
Metoda nejbližšího souseda: 4. krok výpočtu

- 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

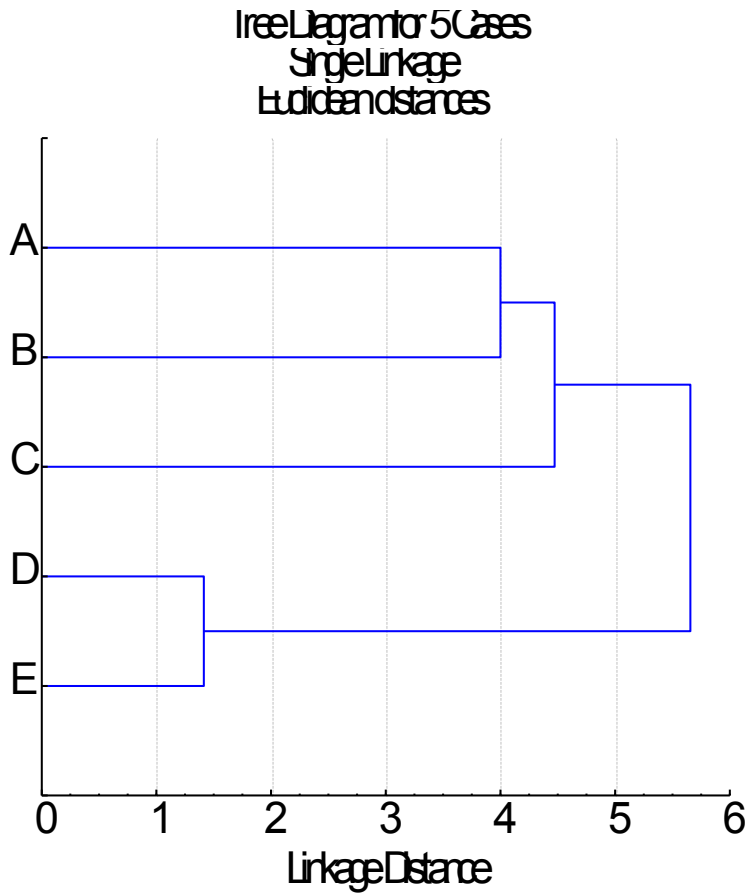
	A+B+C	D+E
A+B+C	0.0	5.7
D+E	5.7	0.0

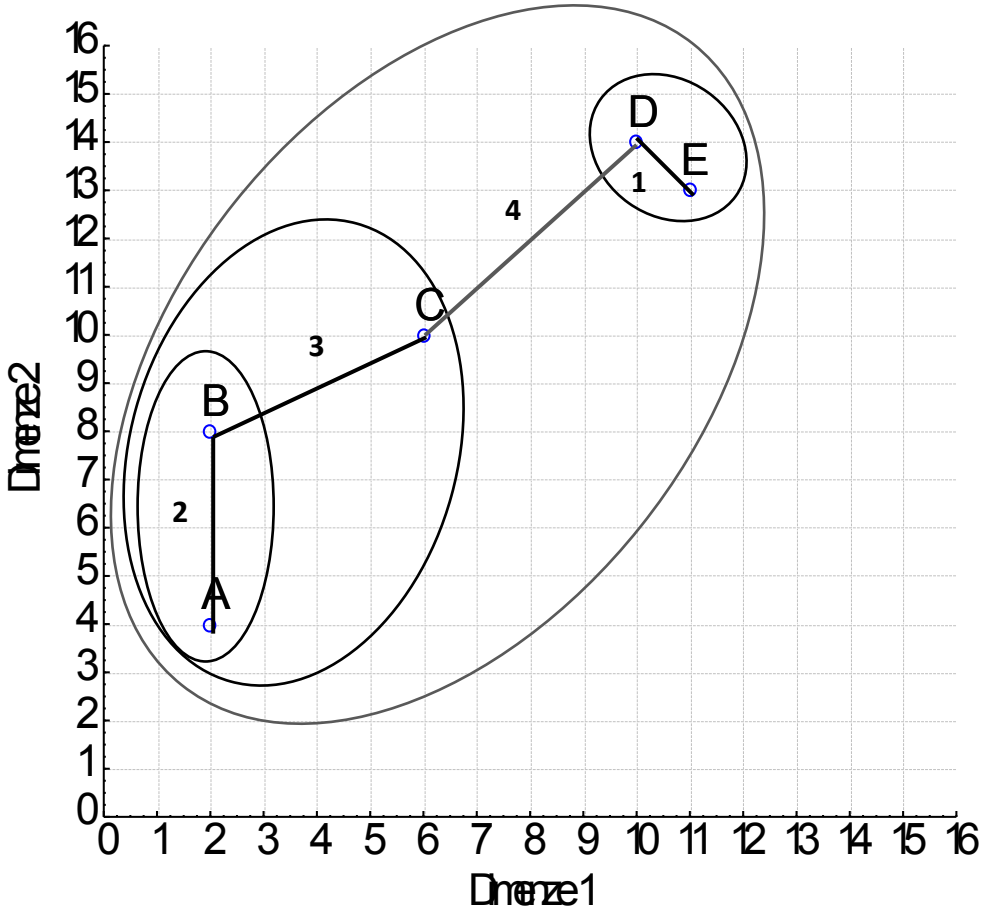
- Je definován shluk dvou nejbližších objektů ((A-B)-C)-(D-E)
- Všechny objekty jsou spojeny, algoritmus je ukončen



Metoda nejbližšího souseda: výsledek analýzy

- Výsledek analýzy je vizualizován ve formě dendrogramu



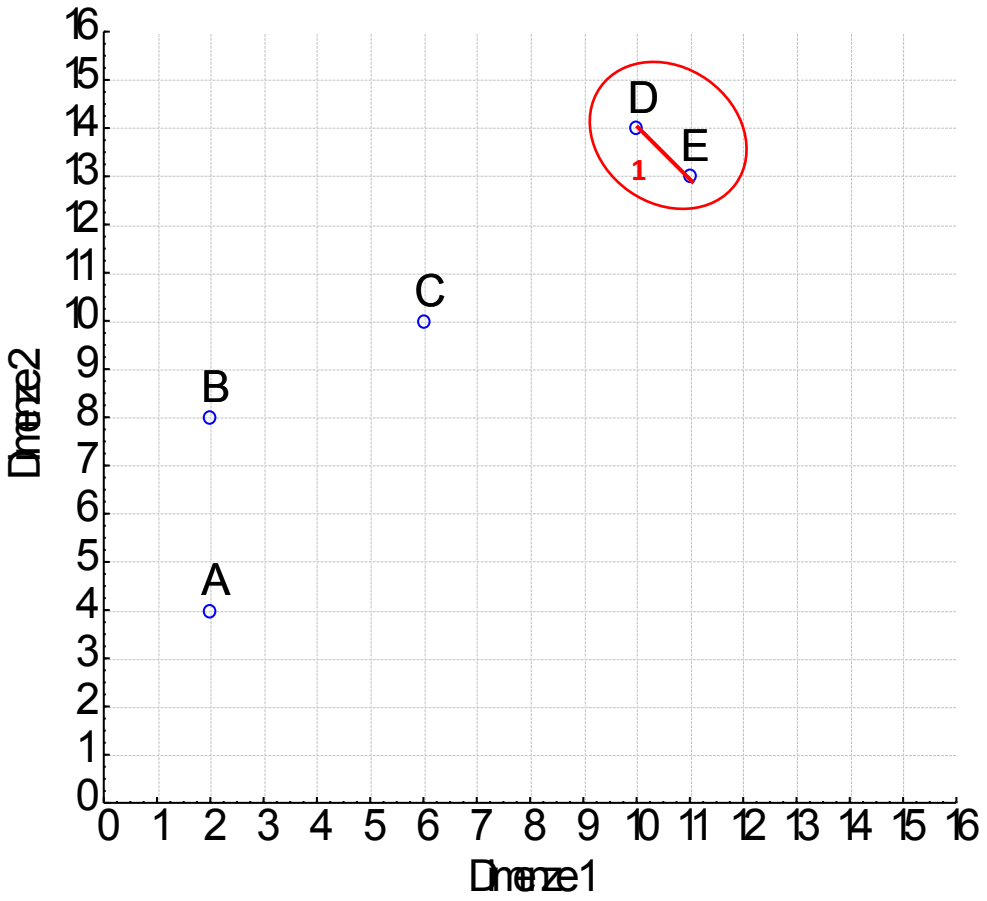


Metoda nejvzdálenějšího souseda: 1. krok výpočtu

- 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

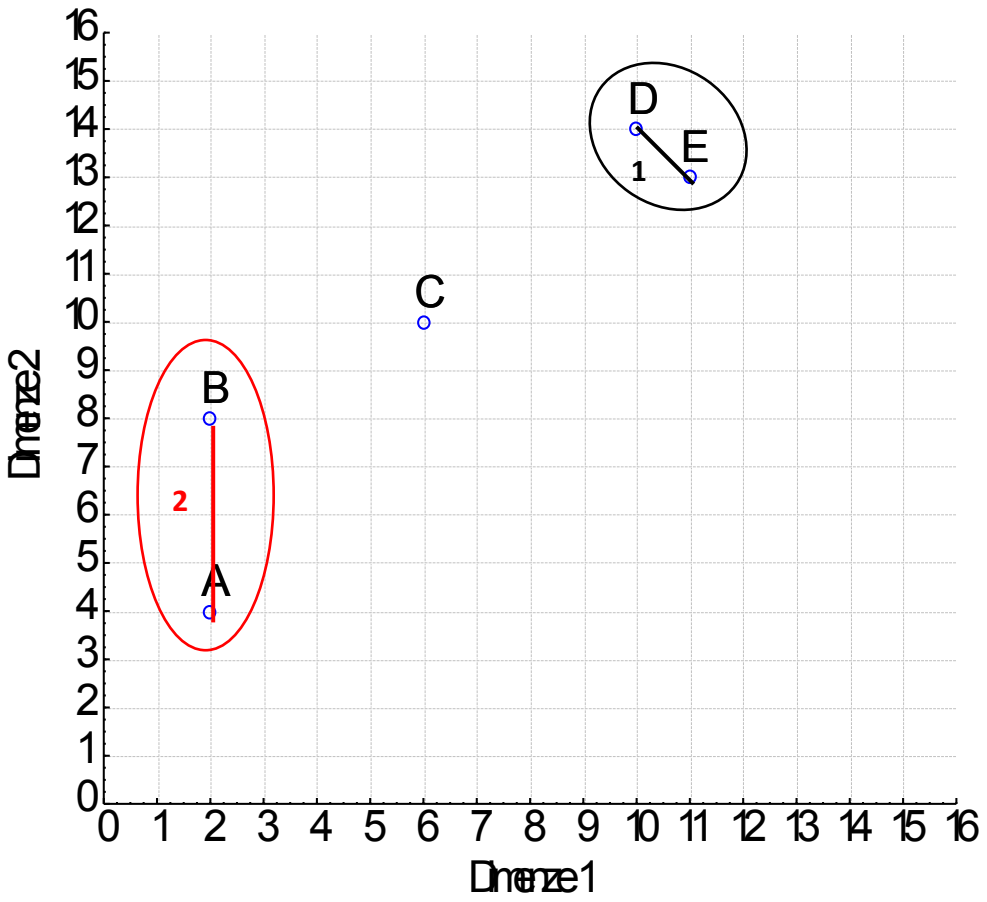


Metoda nejvzdálenějšího souseda: 2. krok výpočtu

- 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



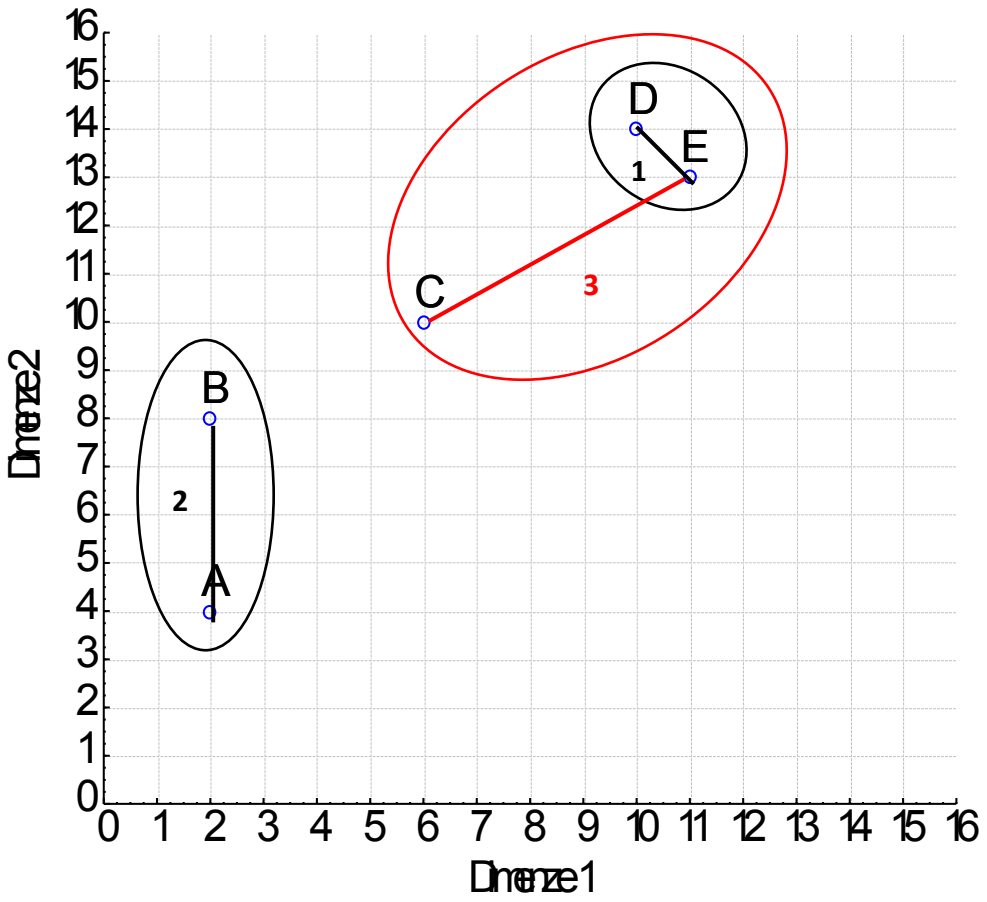
- Je definován shluk dvou nejbližších objektů A-B

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

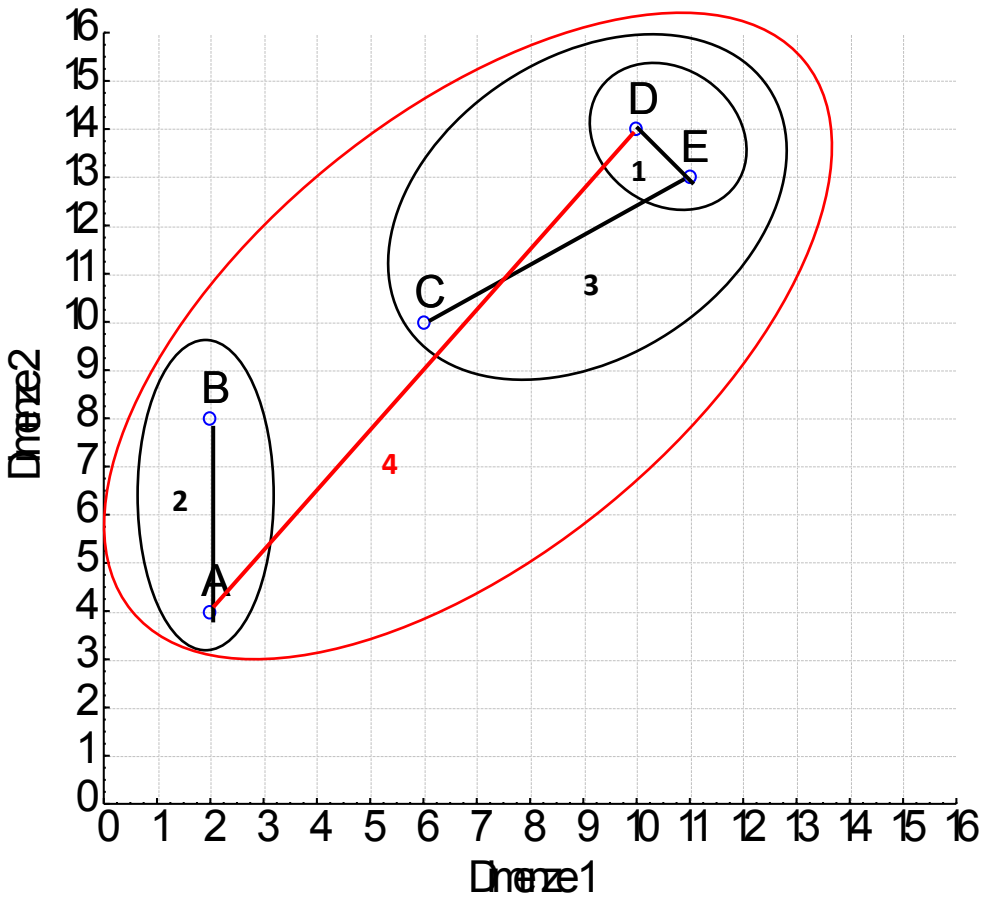
Metoda nejvzdálenějšího souseda: 4. krok výpočtu

- 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

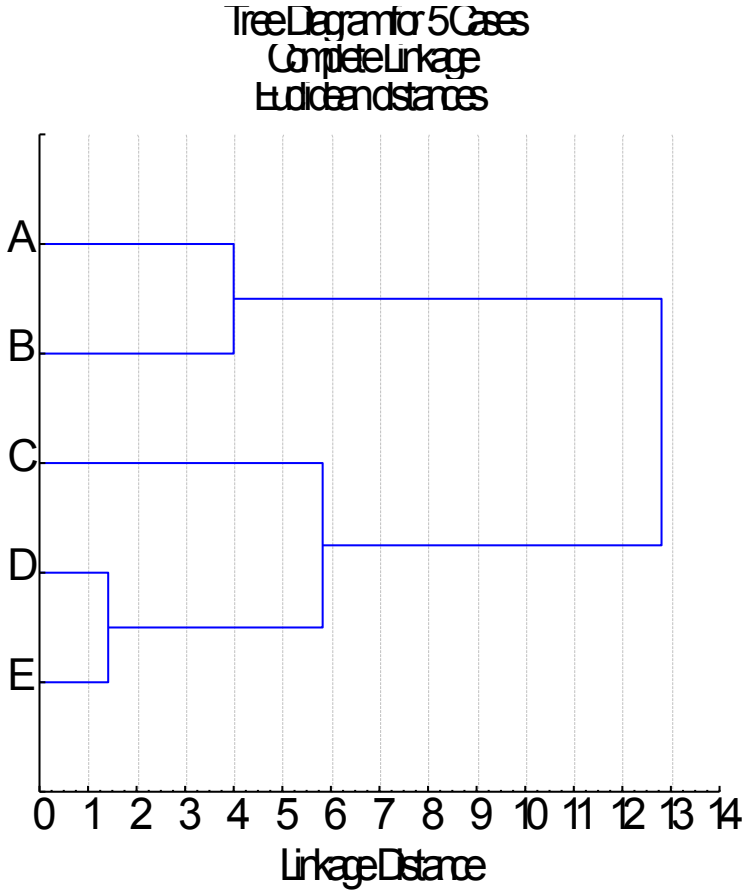
	A+B	D+E+C
A+B	0.0	12.8
D+E+C	12.8	0.0

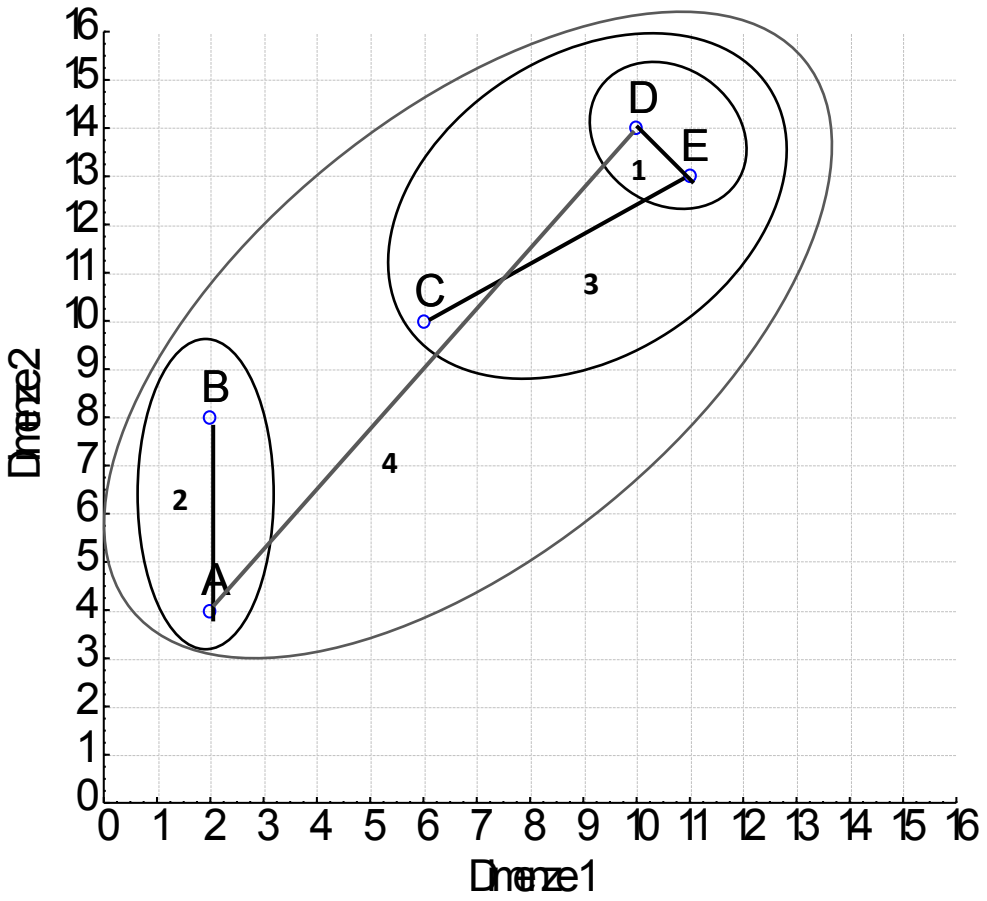
- Je definován shluk dvou nejbližších objektů ((D-E)-C)-(A-B)
- Všechny objekty jsou spojeny, algoritmus je ukončen



Metoda nejvzdálenějšího souseda: výsledek analýzy

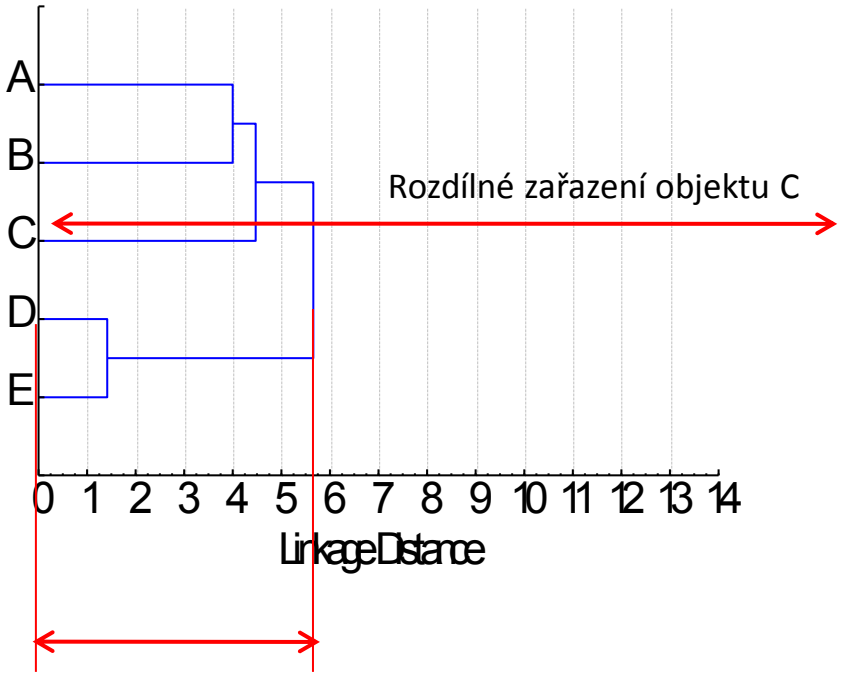
- Výsledek analýzy je vizualizován ve formě dendrogramu





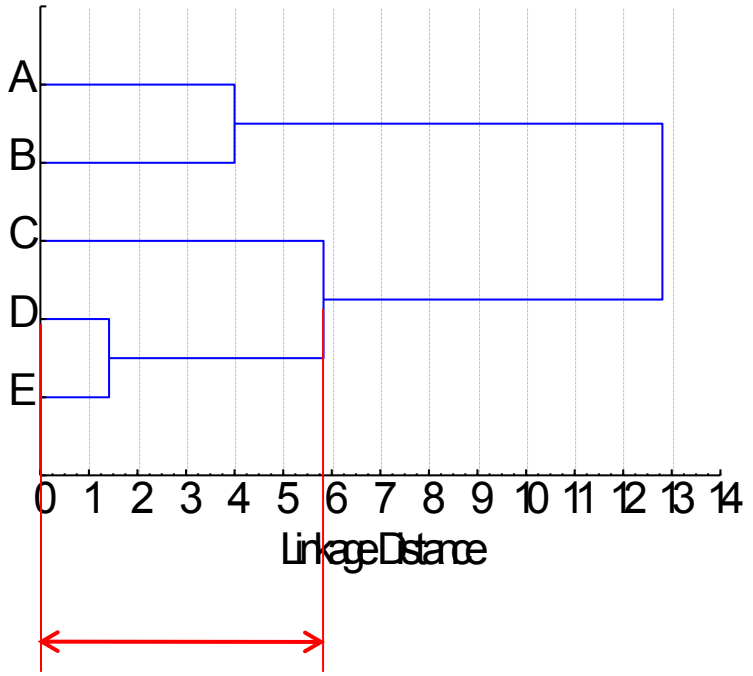
Metoda nejbližšího a nejvzdálenějšího souseda – interpretace výsledků

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

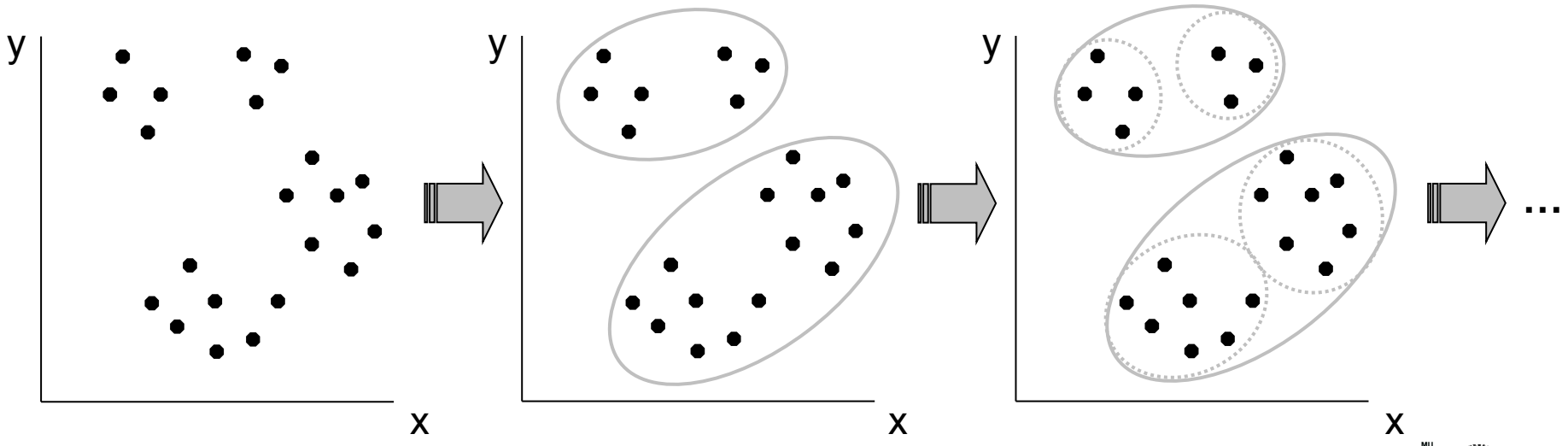


Vzdálenost, na níž došlo ke spojení shluku:
• u metody nejvzdálenějšího souseda znamená největší vzdálenost objektů shluku, tedy objekty ve shluku už mohou být k sobě pouze blíže nebo stejně vzdálené jako je tato vzdálenost

Shluková analýza hierarchická – hierarchické divizivní shlukování

Hierarchické divizivní shlukování – postup

- divizivní metody pracují ze začátku se všemi objekty jako s jednou skupinou
- nejdříve je tato skupina rozdělena do dvou menších skupin
- dělení podskupin pokračuje dále, dokud není splněno alespoň jedno z kritérií, které ukončí analýzu:
  - předem definovaný počet kroků
  - rozklad na samostatné objekty
  - dosažení kritéria minimálního rozdílu mezi shluky



Hierarchické divizivní shlukování – poznámky

- výhoda oproti hierarch. aglomerativnímu shlukování: vhodné pro objemné datové soubory
- výhodou rovněž, že ke každému dělení je připojeno kritérium, podle kterého dělení proběhlo

- typy divizivních metod:
  - monotetické (dělení souboru podle jediné proměnné)
  - polytetické (dělení souboru podle komplexní charakteristiky získané na základě všech proměnných) – např. metoda TWINSPAN

Shluková analýza nehierarchická

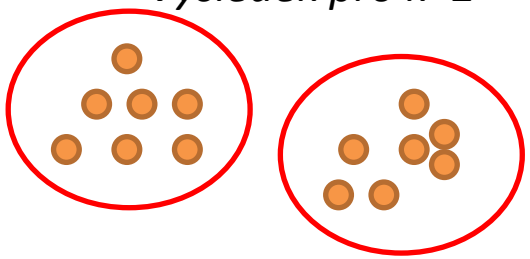
Nehierarchické shlukování

- pokud data nevykazují hierarchickou strukturu, je často vhodnější používat nehierarchické shlukování namísto hierarchického shlukování
- výstupem vytvoření skupin stejného řádu
- skupiny uvnitř co nejvíce homogenní a mezi sebou co nejvíce odlišné
- nehierarchické metody vhodné pro velmi objemná data
- metody nehierarchického divizivního shlukování:
  - metoda *k*-průměrů (*k-means clustering*)
  - metoda *x*-průměrů
  - metoda *k*-medoidů
- metody nehierarchického aglomerativního shlukování:
  - Minimum spanning tree

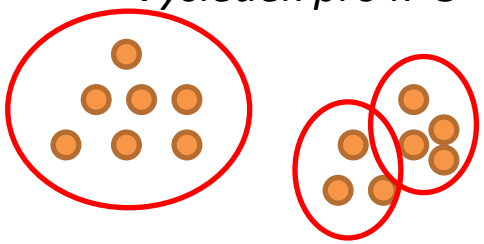
Nehierarchické divizivní shlukování – metoda k-průměrů

- Metoda zařazuje objekty do shluků na principu ANOVA, analogií je Wardova metoda shlukování v hierarchickém aglomerativním shlukování
- Počet shluků je předem definován, výběr nejvhodnějšího počtu shluků je prováděn buď expertně, nebo pomocí matematických metod výběru optimálního počtu shluků (analýza vnitro a mezishlukových vzdáleností)
- Postup:
  1. V prvním kroku je určeno *k* objektů jako počáteční středy shluků (výběr může být náhodný, daný uživatelem nebo maximalizující počáteční vzdálenosti *k* objektů)
  2. Následně jsou objekty zařazeny do *k* shluků tak, aby byla minimalizována suma čtverců vzdáleností objektů k centroidům jejich shluků

Výsledek pro k=2



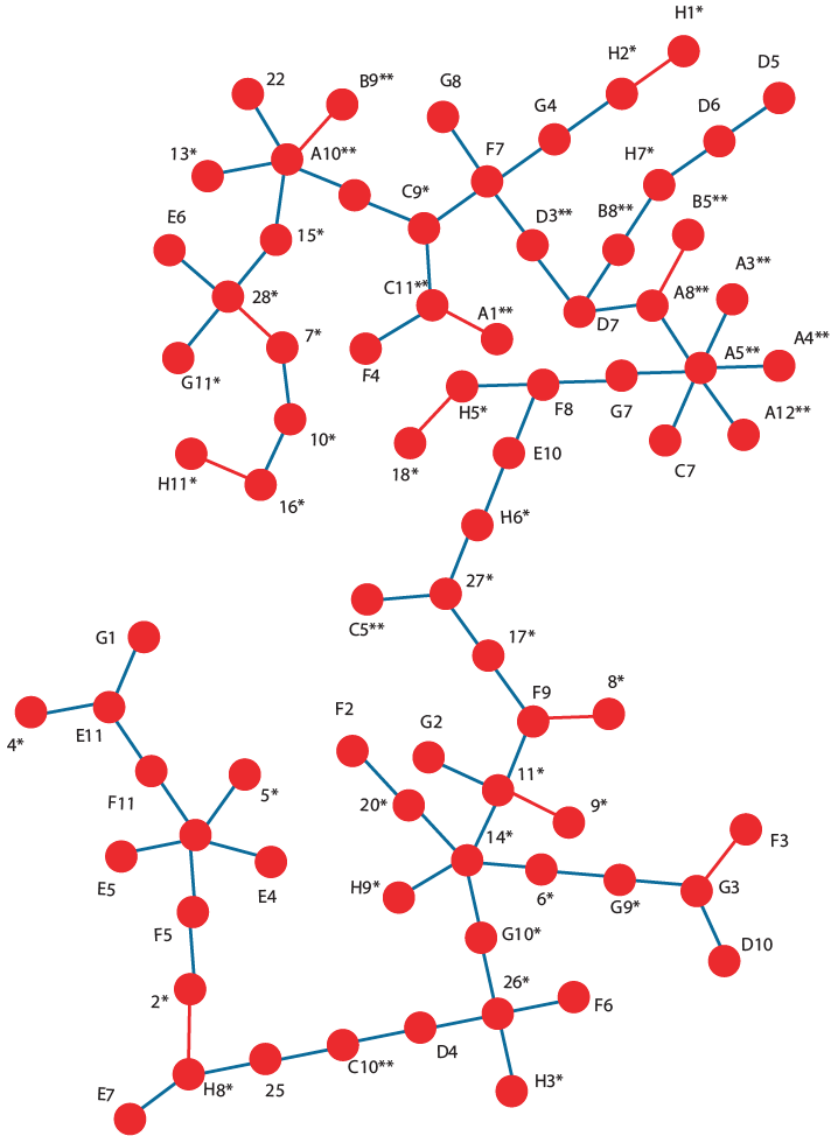
Výsledek pro k=3



- Upozornění: Analýza vždy nalezne zadaný počet shluků, i když výsledek nemusí být vždy prakticky smysluplný!

Nehierarchické aglomerativní shlukování - postup

- Do této skupiny lze zařadit metody hledající nejkratší spojnici mezi objekty ve vícerozměrném prostoru (i když lze vznést námitky proti nazývání těchto metod nehierarchickými)
- Metody hledají v asociační matici (prvním krokem je tak vždy výběr vhodné metriky vzdáleností/ podobností) propojení všech objektů s nejmenší sumou vzdáleností mezi propojenými objekty
- Na rozdíl od klasického hierarchického aglomerativního shlukování může být na jeden objekt napojeno několik dalších objektů
- Minimum spanning tree

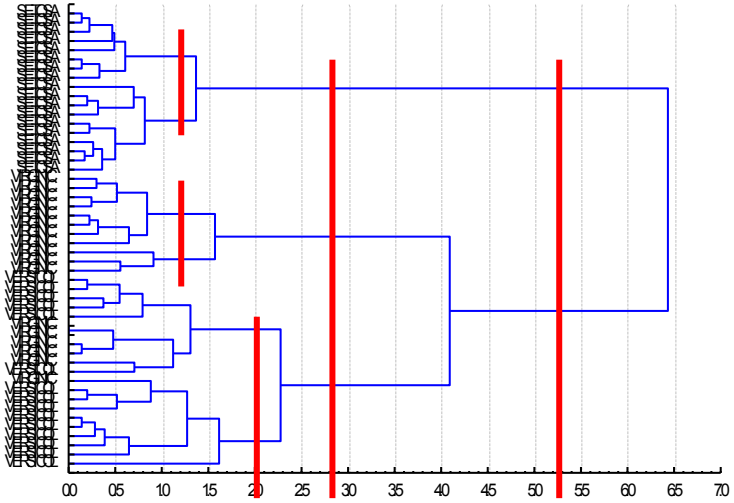


Identifikace optimálního počtu shluků

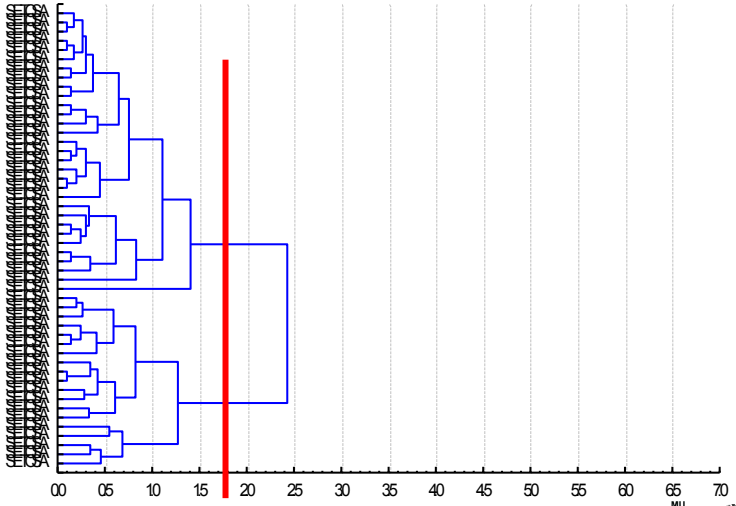
Identifikace optimálního počtu shluků

- Cílem analýzy může být jednak zjistit vazby mezi objekty (dostatečným výstupem je dendrogram) nebo identifikovat v datech shluky, které budou využity v další analýze jako zjednodušení vícedimenzionálního problému
- Identifikace shluků ve výsledcích shlukové analýzy:
  - Expertní/intuitivní – hranice oddělení shluků je určena podle zkušeností analytika a praktického významu výstupu
  - Matematické metody (analýza mezishlukových/vnitroshlukových vzdáleností; silhouette metoda aj.) fungují dobře v případě existence přirozených shluků
  - V některých případech (při neexistenci přirozených shluků) je rozdělení souboru pouze arbitrární

Jednoznačný řez na více vzdálenostech



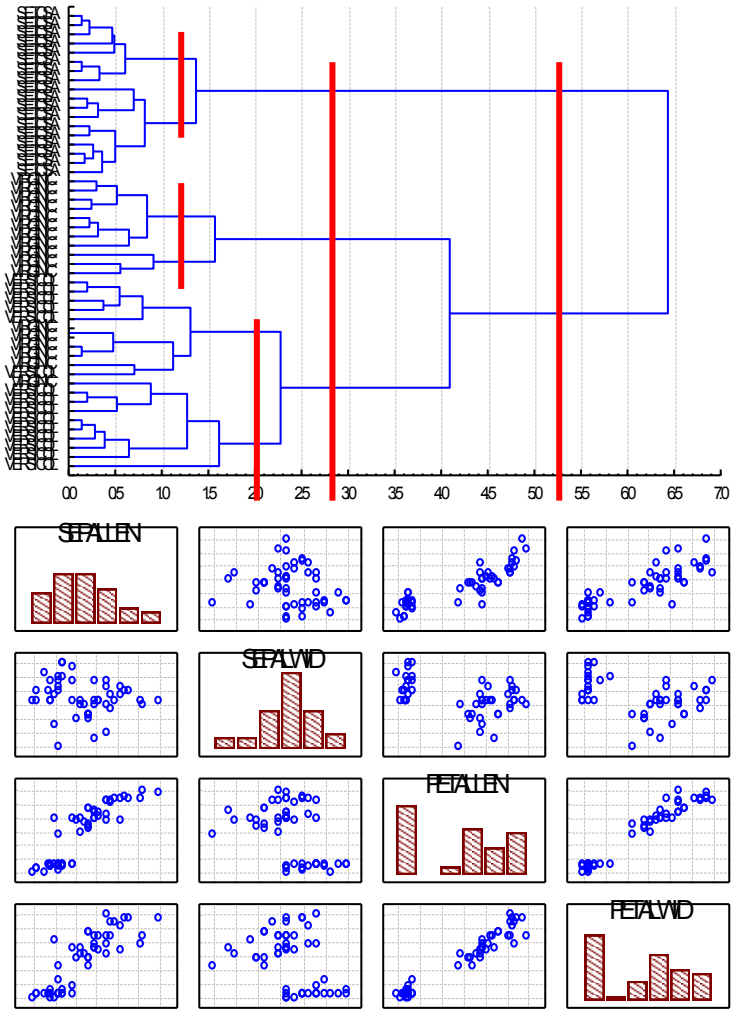
Jediný identifikovatelný řez, navíc na malé vzdálenosti



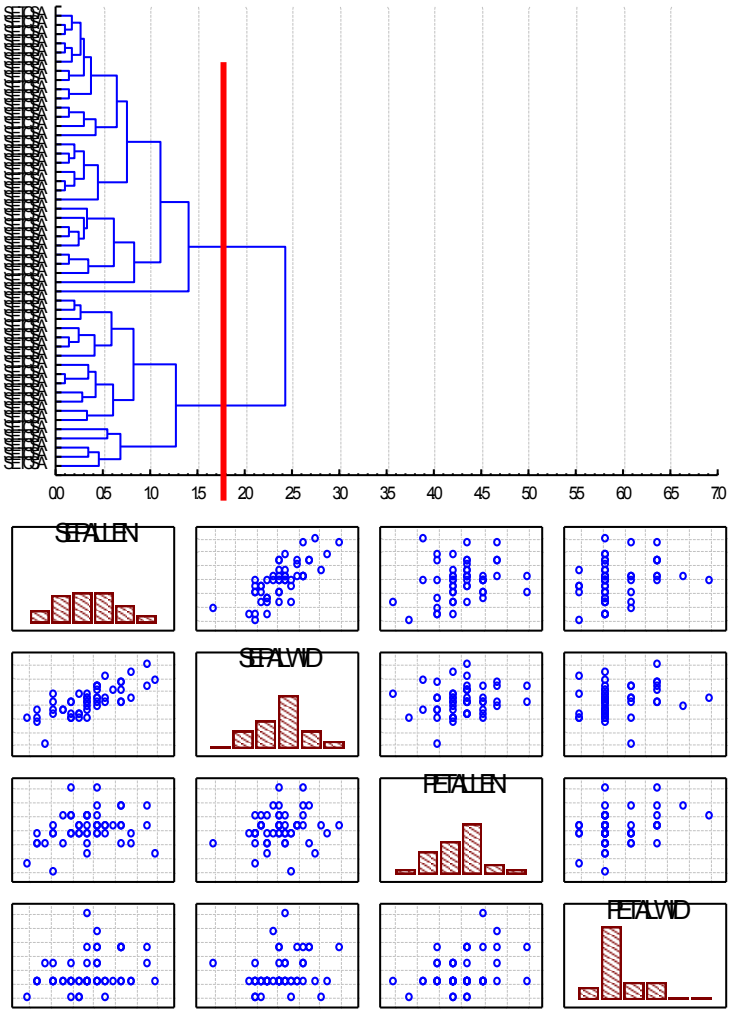
Identifikace optimálního počtu shluků

- Mezi shlukovou analýzou a pozicí objektů ve vícerozměrném prostoru existuje vztah

Jednoznačný řez na více vzdálenostech



Jediný identifikovatelný řez, navíc na malé vzdálenosti



Identifikace optimálního počtu shluků - metody

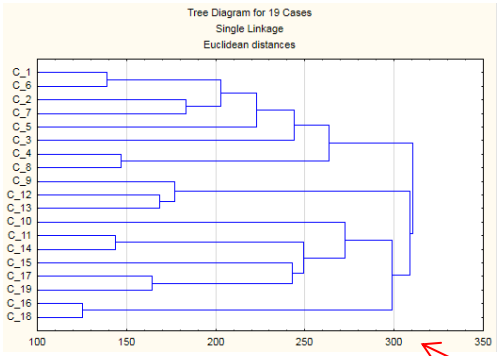
- Dunnův validační index
- Daviesův-Bouldinův validační index
- Metoda siluety
- Izolační index
- C-index
- Goodmanův-Kruskalův index

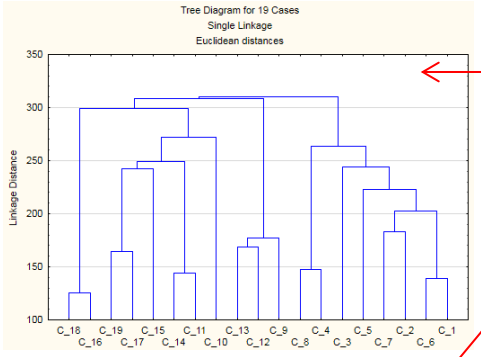
Výpočet shlukové analýzy v softwarech

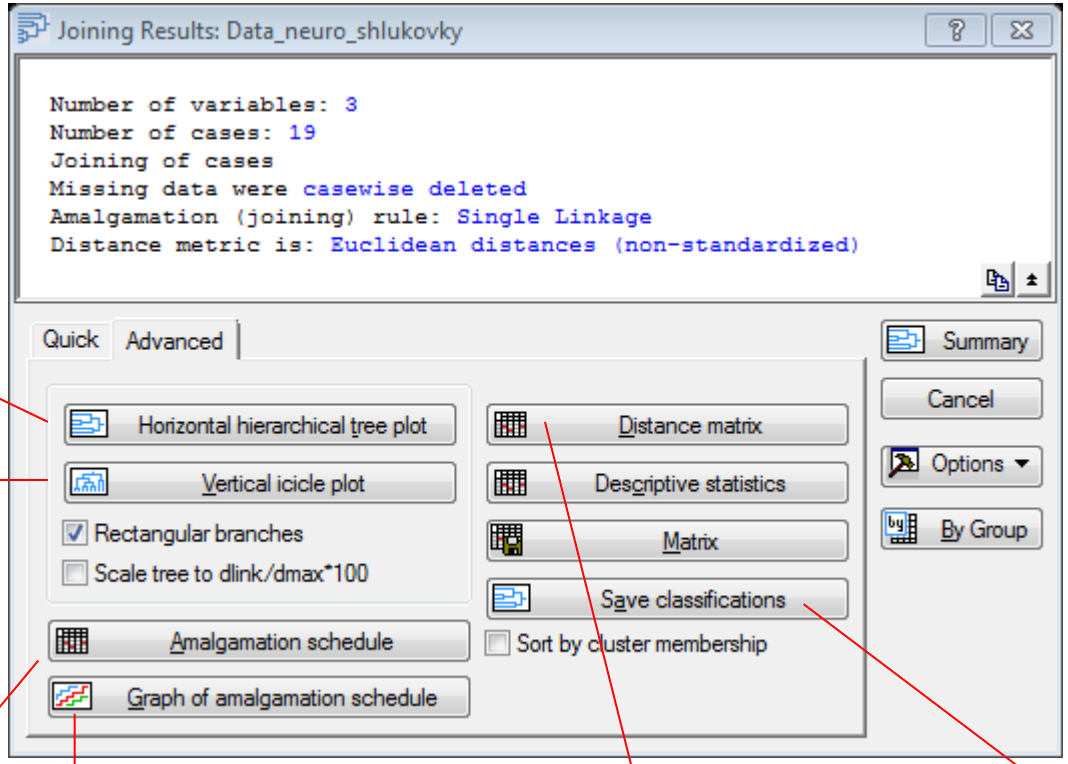
STATISTICA – hierarchické aglomerativní shlukování

- Statistics – Multivariate Exploratory Techniques – Cluster Analysis – Joining (tree 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))
- 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

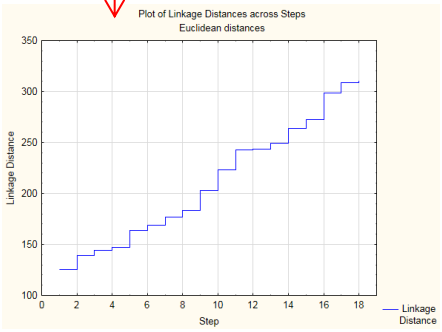






Amalgamation Schedule (Data_neuro_shlukovky)
Single Linkage
Euclidean distances

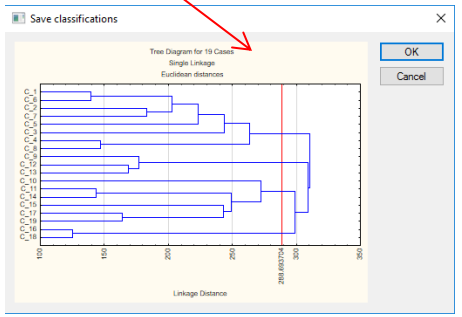
linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	O
125.1972	C_16	C_18				
139.2640	C_1	C_6				
143.9270	C_11	C_14				
147.0873	C_4	C_8				
164.1363	C_17	C_19				
168.6528	C_12	C_13				
176.9954	C_9	C_12	C_13			
183.2707	C_2	C_7				
202.7584	C_1	C_6	C_2	C_7		
223.0460	C_1	C_6	C_2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



asociační matice
Euklidových vzdáleností

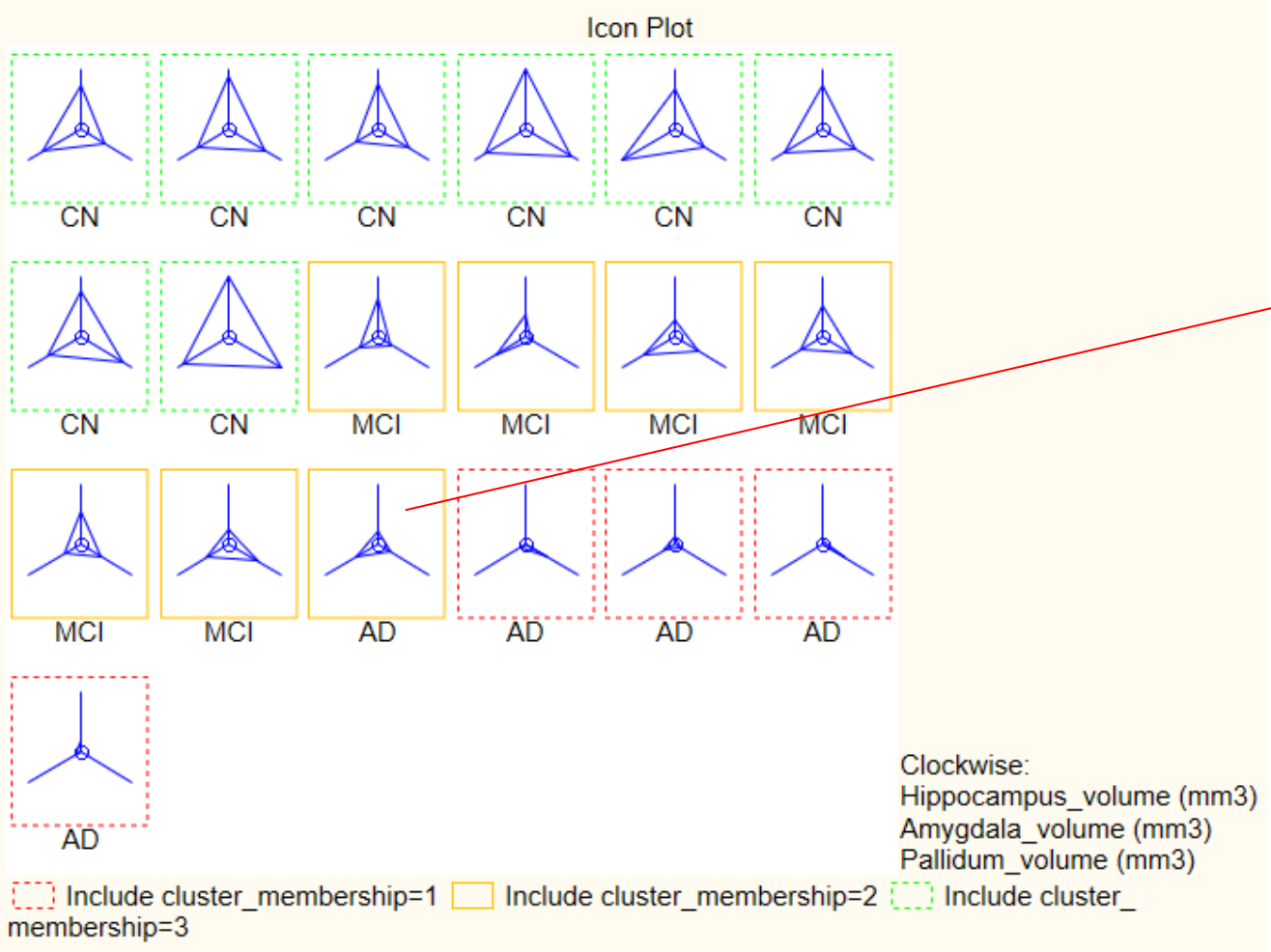
Euclidean distances (Data_neuro_shlukovky)

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



STATISTICA – hierarch. aglom. shluk. – pokračování

Příslušnosti do skupin lze vykreslit pomocí ikonového grafu



tento subjekt klasifikován špatně

Patrné, že:

- CN největší hodnoty objemu všech 3 struktur
- AD nejmenší hodnoty objemu všech 3 struktur

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
```

Poděkování

Příprava výukových materiálů předmětu
„DSAN02 Pokročilé metody analýzy dat v neurovědách“
byla finančně podporována prostředky projektu FRMU
č. MUNI/FR/0260/2014 „Pokročilé metody analýzy dat
v neurovědách jako nový předmět na LF MU“