

Vícerozměrné statistické metody

Shluková analýza

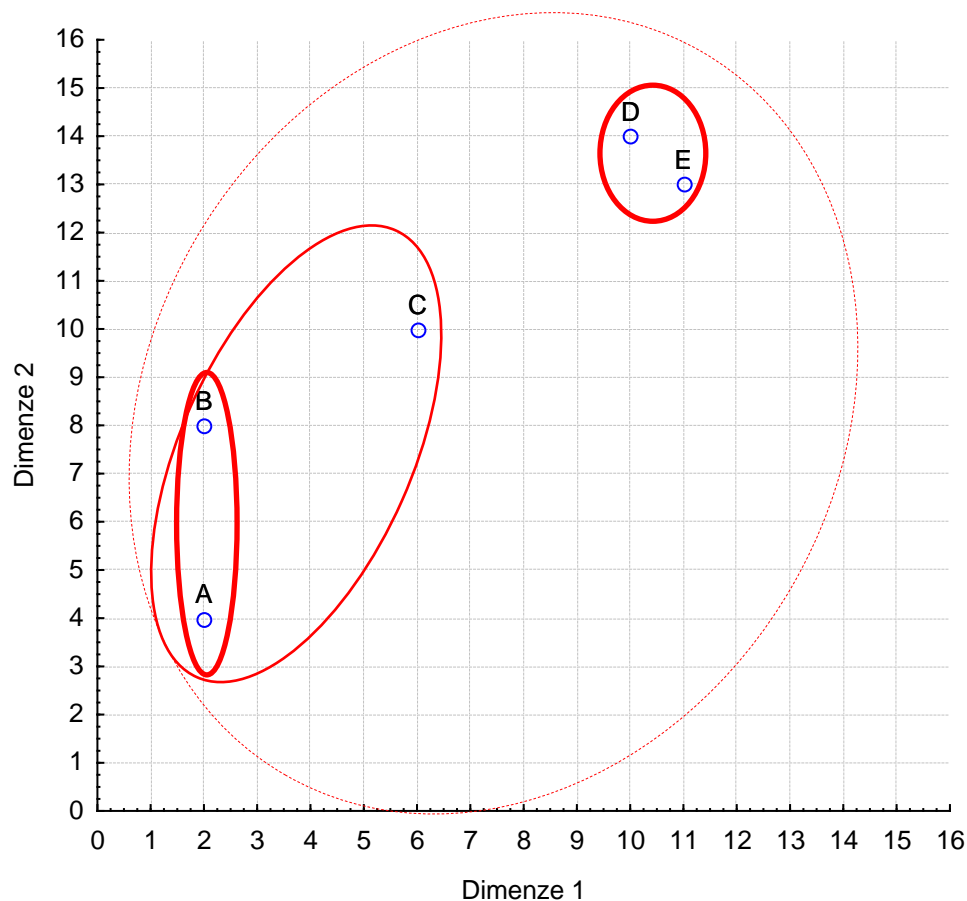
Jiří Jarkovský, Simona Littnerová

Vícerozměrné statistické metody

Typy shlukových analýz

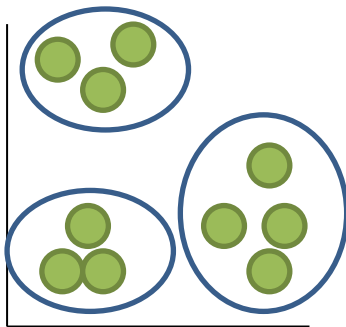
Shluková analýza: cíle a postupy

- Shluková analýza se snaží o identifikaci shluků objektů ve vícerozměrném prostoru a následnou redukce 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

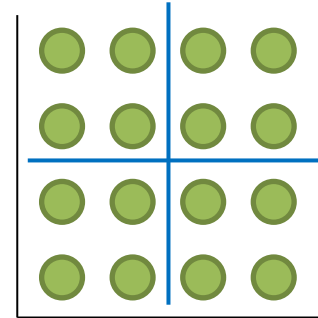


Obecný princip hledání shluků v datech

- Vzájemnou pozici objektů ve vícerozměrném prostoru lze popsat jejich vzdáleností
- Dle vzdálenosti objektů je můžeme slučovat do shluků a přiřazení objektů ke shlukům ve vícerozměrném prostoru následně využít pro zjednodušení jejich x-dimenzionálního popisu
- Smysluplnost výsledků shlukování závisí jednak na objektivní existenci shluků v datech, jednak na arbitrárně nastavených kritériích definice shluků



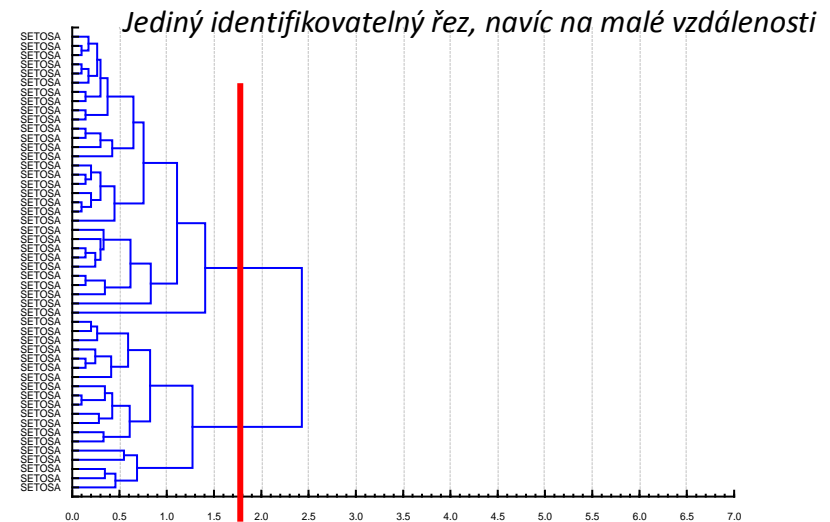
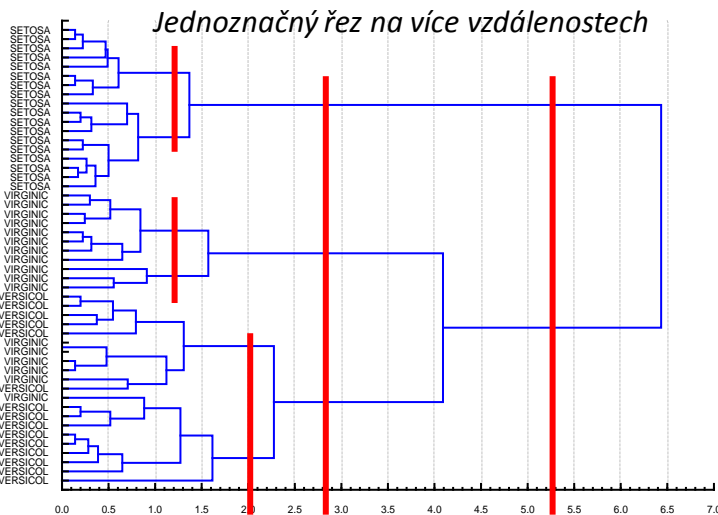
Jednoznačné odlišení existujících shluků v datech (obdoba multimodálního rozložení)



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

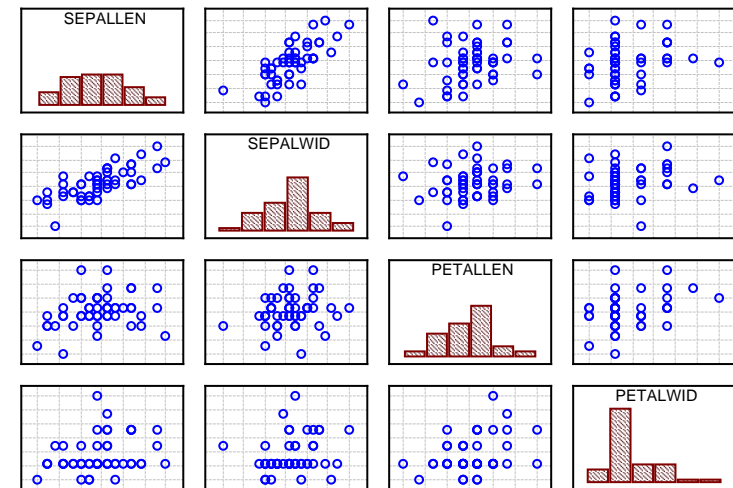
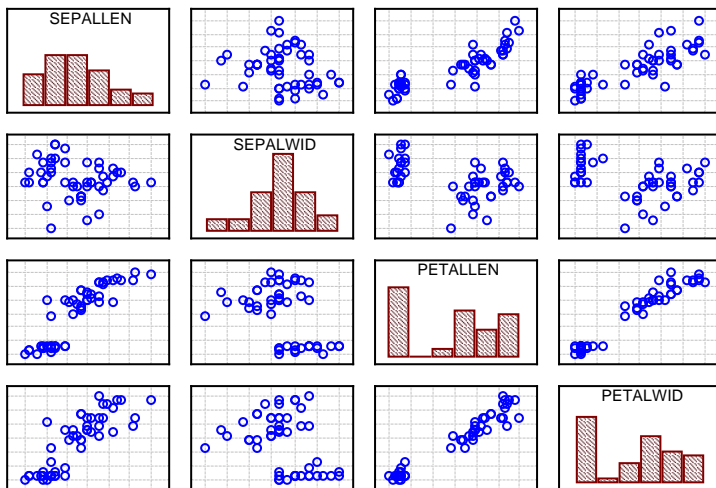
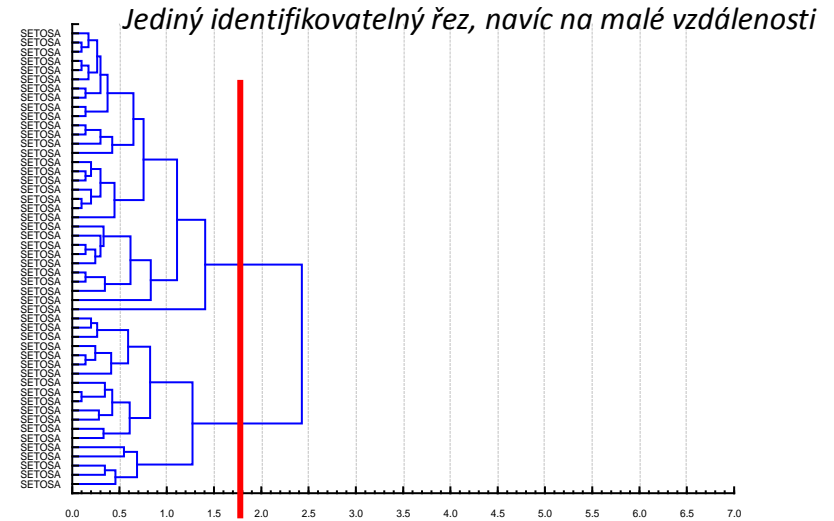
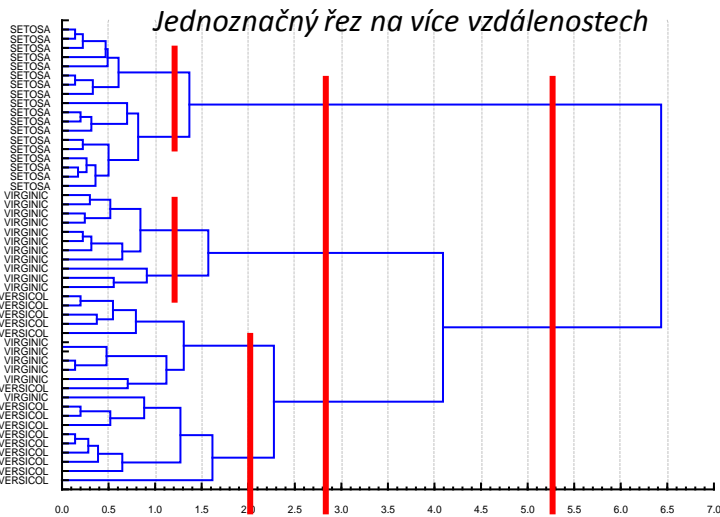
Identifikace optimálního počtu shluků

- 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

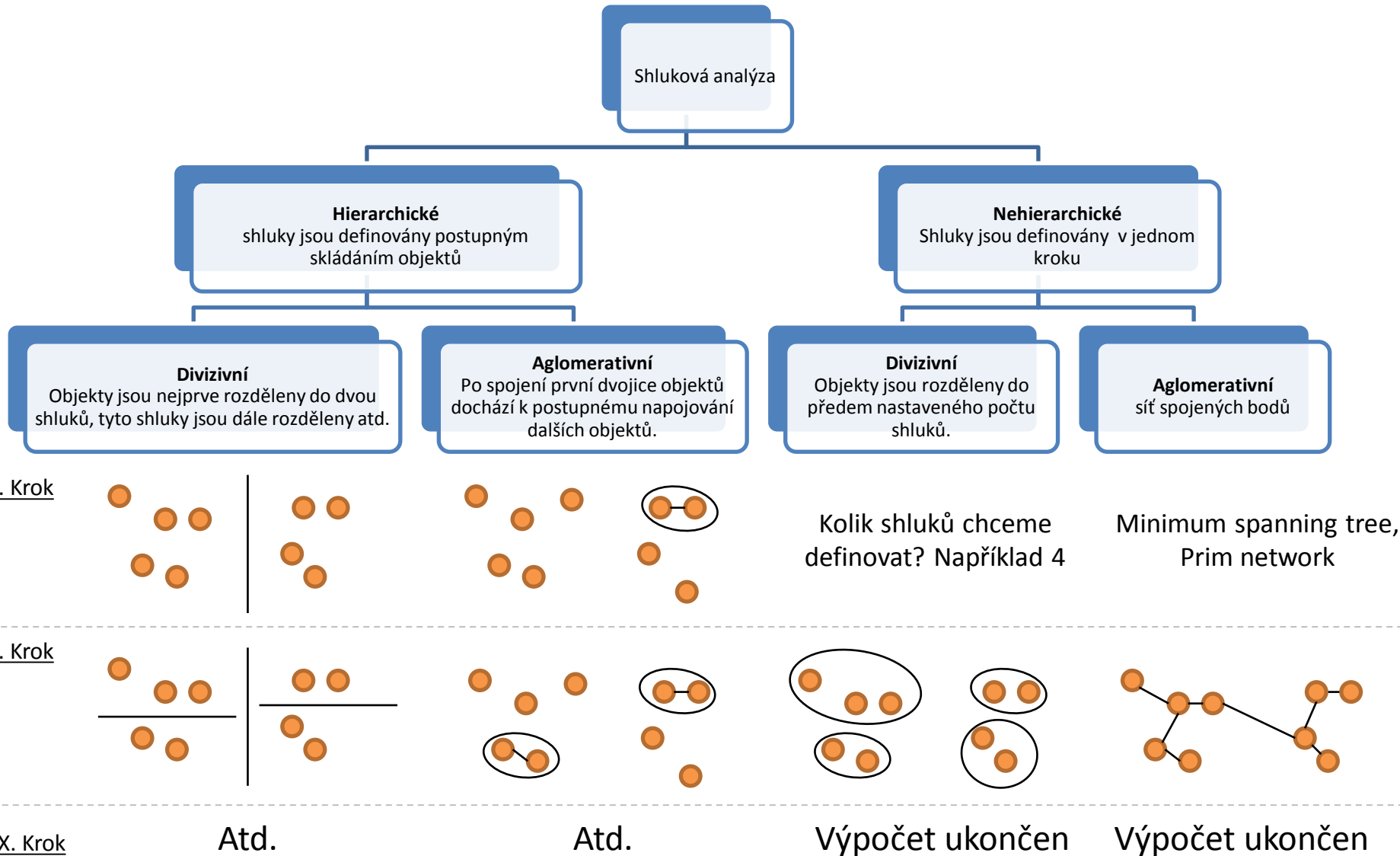


Identifikace optimálního počtu shluků

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



Shluková analýza: typy metod



Vícerozměrné statistické metody

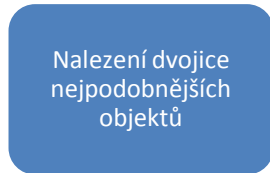
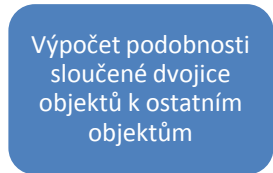
Hierarchické aglomerativní shlukování

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
 - Výběr vhodné metriky vzdálenosti/podobnosti pro výpočet asociační matice (analýza může probíhat na libovolných metrikách vzdálenosti/podobnosti)
 - Výběr shlukovacího algoritmu, který podstatným způsobem ovlivňuje výsledky analýzy a možnosti její interpretace
- Algoritmus výpočtu postupuje v následujícím cyklu
 - Výpočet asociační matice
 - Spojení dvou nejpodobnějších objektů
 - 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)
 - Spojení dvou nejpodobnějších objektů z přepočítané asociační matice
 - Atd. až do spojení všech objektů

Hierarchické aglomerativní shlukování: schéma výpočtu

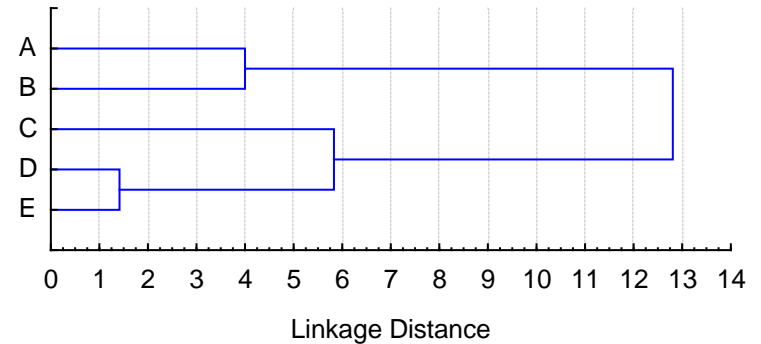
Výběr metriky podobnosti/vzdálenosti



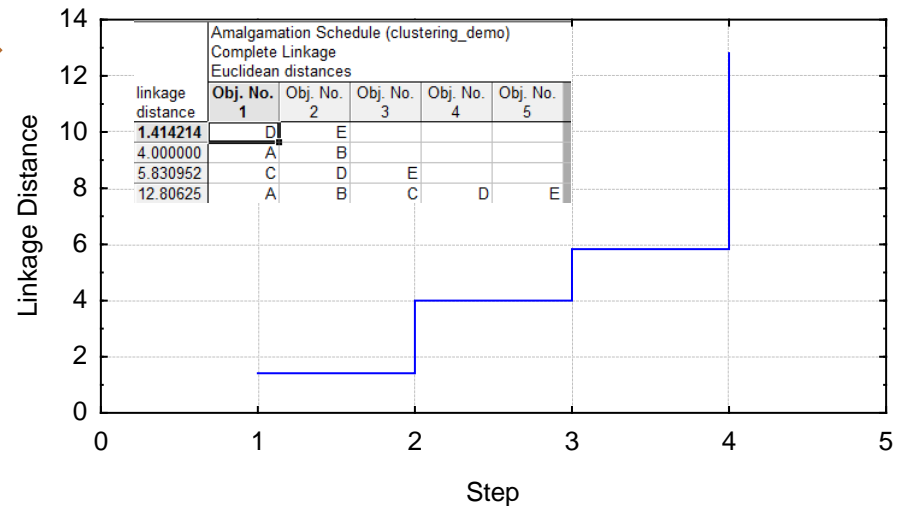
Výběr shlukovacího algoritmu

Ukončení výpočtu po spojení všech objektů

Dendrogram



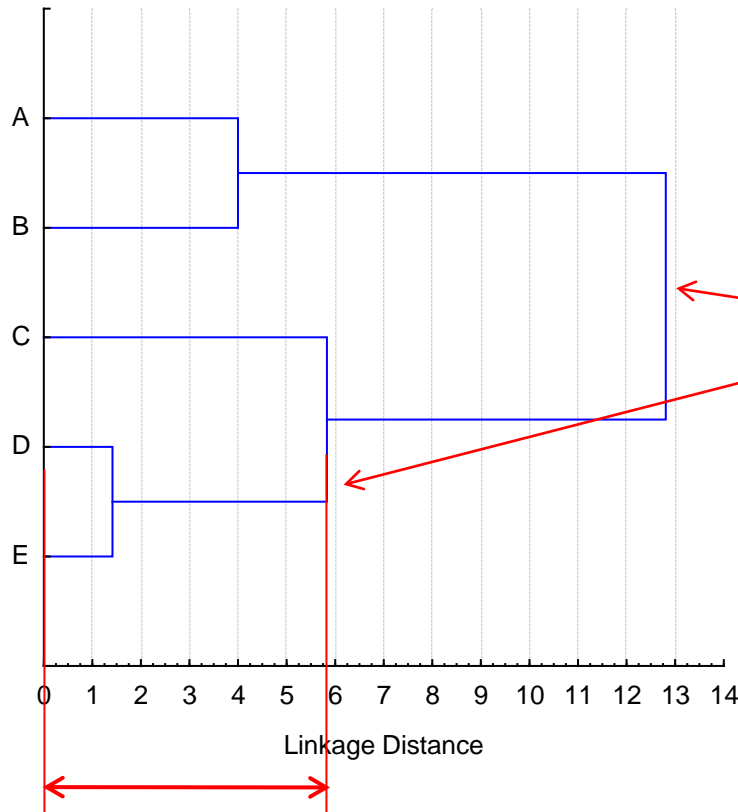
Amalgamation schedule/graph



Popis výstupů: dendrogram

Tree Diagram for 5 Cases
Complete Linkage
Euclidean distances

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



Propojení
shlukovaných
objektů

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

Vzdálenost na níž došlo ke spojení shluku:

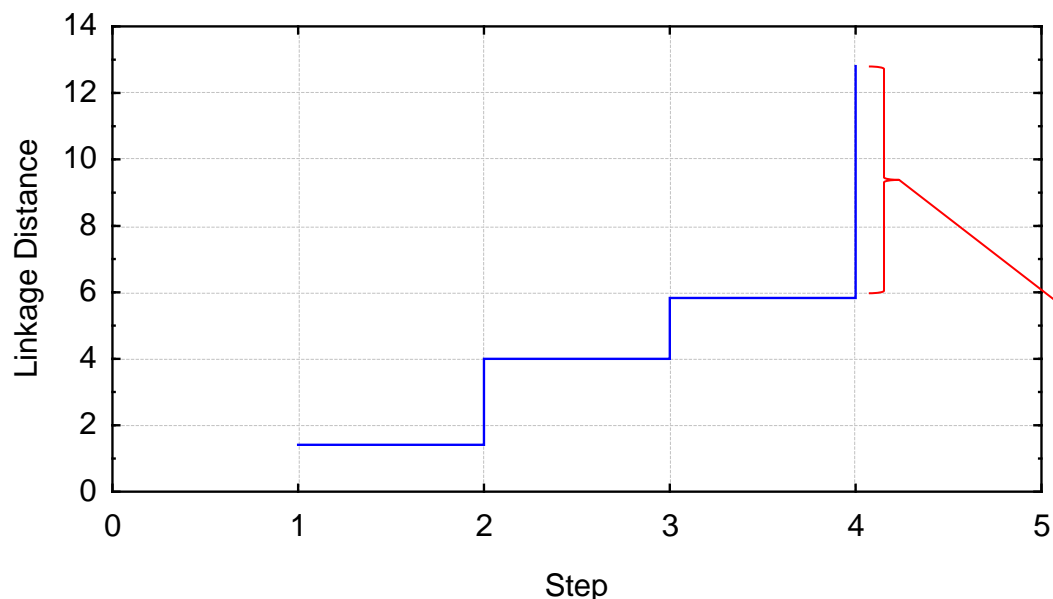
- je v rozměrech použité metriky vzdáleností/podobností a v tomto kontextu ji lze kvantitativně interpretovat
- interpretace vzdálenosti shlukování se liší podle použitého shlukovacího algoritmu
- někdy se uvádí ve škále 0-100%, kde 100% je maximální vzdálenost shlukování

Popis výstupů: Amalgamation schedule/graph

- Popis postupu shlukování
- Využitelné 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



Grafické vyjádření kroků shlukování a vzdálenosti na nichž došlo k propojení objektů

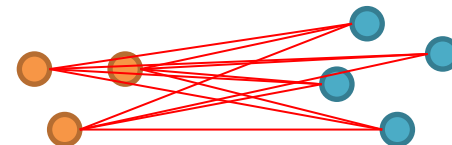
Pokud je v grafu dlouhá vzdálenost bez napojení shluku, jde o možné místo zastavení shlukování a definici finálních shluků

Shlukovací algoritmy hierarchického aglomerativního shlukování I

- **Metoda nejbližšího souseda** (nearest neighbour, simple linkage) – spojení dle nejmenší vzdálenosti mezi objekty shluků



- **Průměrná vzdálenost** (pair group average) – spojení dle průměrné vzdálenosti mezi objekty shluků
 - Vážená (weighted) – odstranění vlivu velikosti shluků, shluky bez ohledu na velikost přispívají k výpočtu spojovací vzdálenosti stejnou vahou
 - Nevážená (unweighted) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků



- **Středospojná vzdálenost** (pair group centroid) – spojení dle vzdálenosti centroidů shluků
 - Vážená (weighted) – odstranění vlivu velikosti shluků, shluky bez ohledu na velikost přispívají k výpočtu spojovací vzdálenosti stejnou vahou
 - Nevážená (unweighted) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků

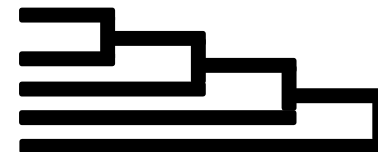


- **Metoda nejvzdálenějšího souseda** (farthest neighbour, complete linkage) – spojení dle největší vzdálenosti mezi objekty shluků



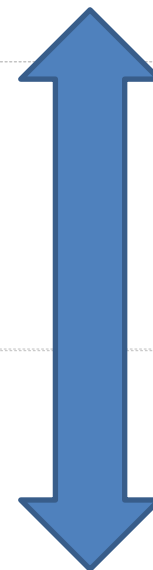
Shlukovací algoritmy hierarchického aglomerativního shlukování II

- **Metoda nejbližšího souseda** (nearest neighbour, simple linkage) – spojení dle nejmenší vzdálenosti mezi objekty shluků – vede na nejvíce zřetěžené dendrogramy



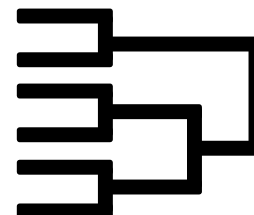
- **Průměrná vzdálenost** (pair group average) – spojení dle průměrné vzdálenosti mezi objekty shluků

- **Středospojná vzdálenost** (pair group centroid) – spojení dle vzdálenosti centroidů shluků



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

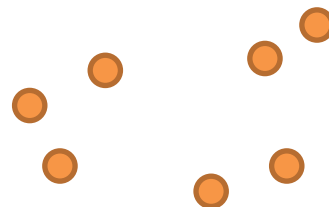
- **Metoda nejvzdálenějšího souseda** (farthest neighbour, complete linkage) – spojení dle největší vzdálenosti mezi objekty shluků – vede na dendrogramy s dobře oddělenými shluky



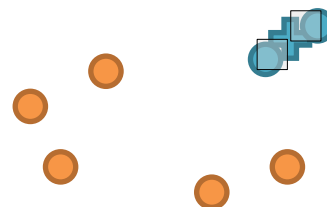
Shlukovací algoritmy hierarchického aglomerativního shlukování III: Wardova metoda

- Principiálně 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 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.

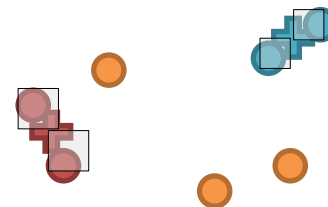
Krok 1: každý objekt je sám sobě centroidem



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 3: stejný postup až do spojení všech objektů

Vícerozměrné statistické metody

Hierarchické aglomerativní shlukování:
Příklad výpočtu metody nejbližšího souseda

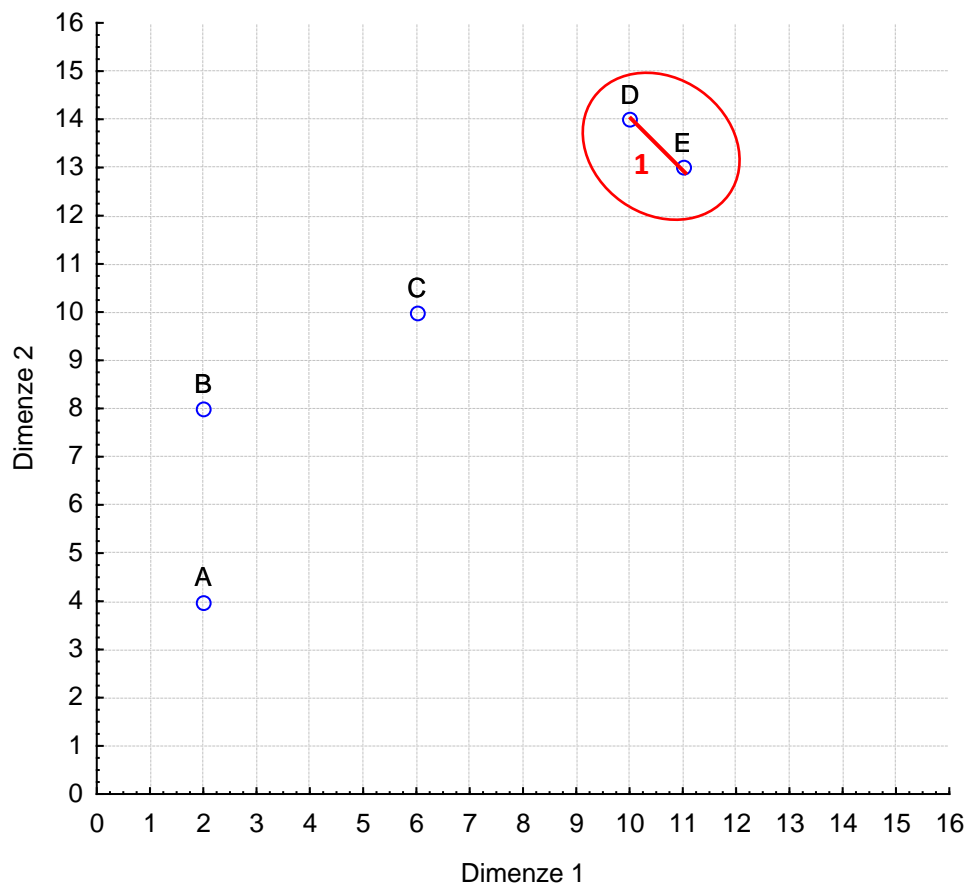
Metoda nejbližšího souseda: 1. krok výpočtu

- 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

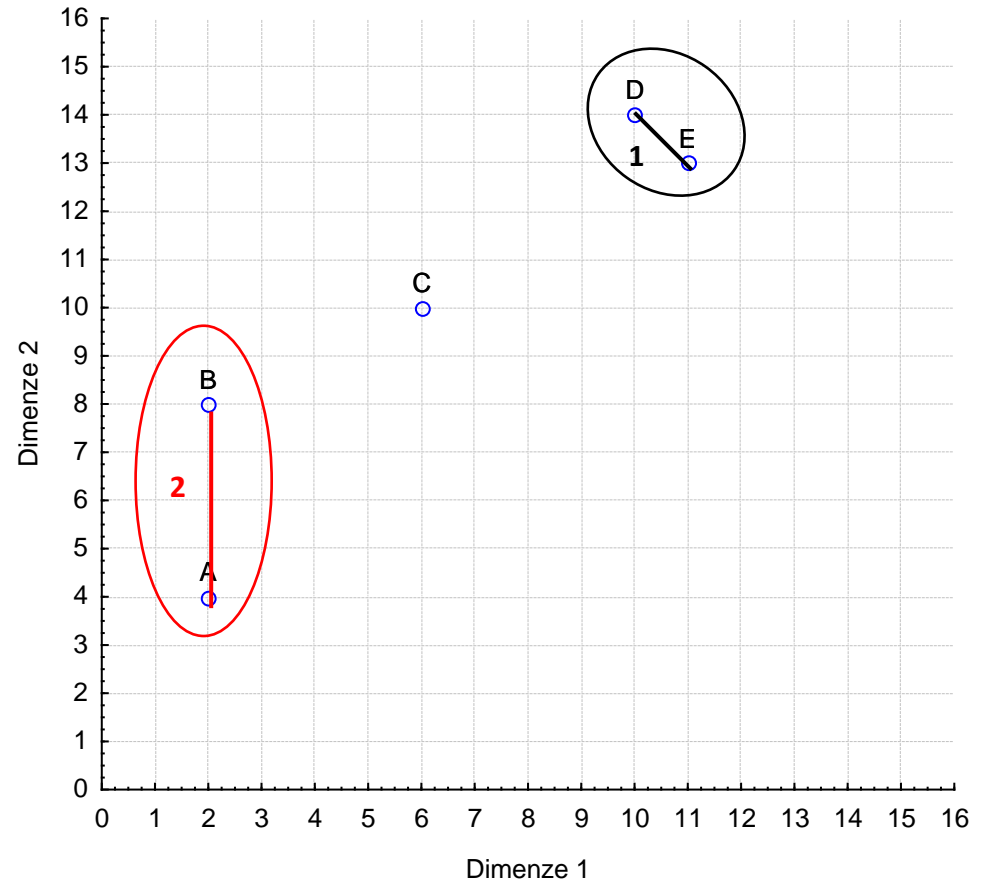


Metoda nejbližšího souseďa: 2. krok výpočtu

- 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	12.7
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

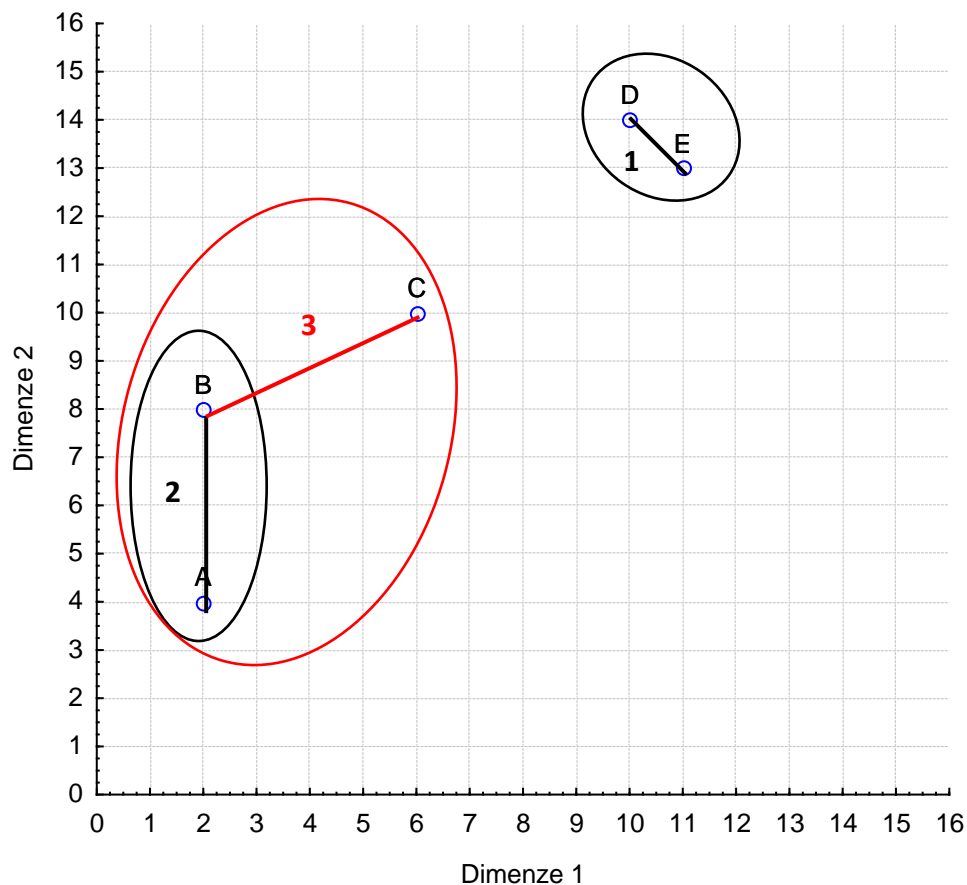


Metoda nejbližšího souseda: 3. krok výpočtu

- 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+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**

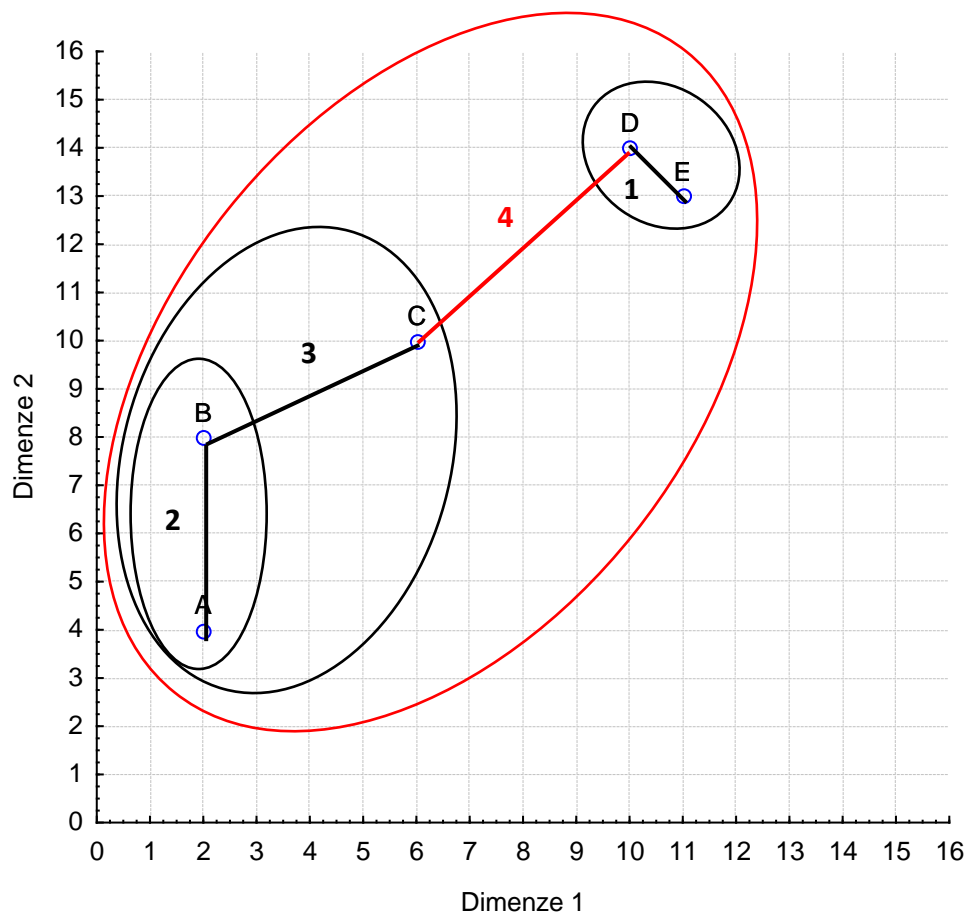


Metoda nejbližšího souseda: 4. krok výpočtu

- 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)**

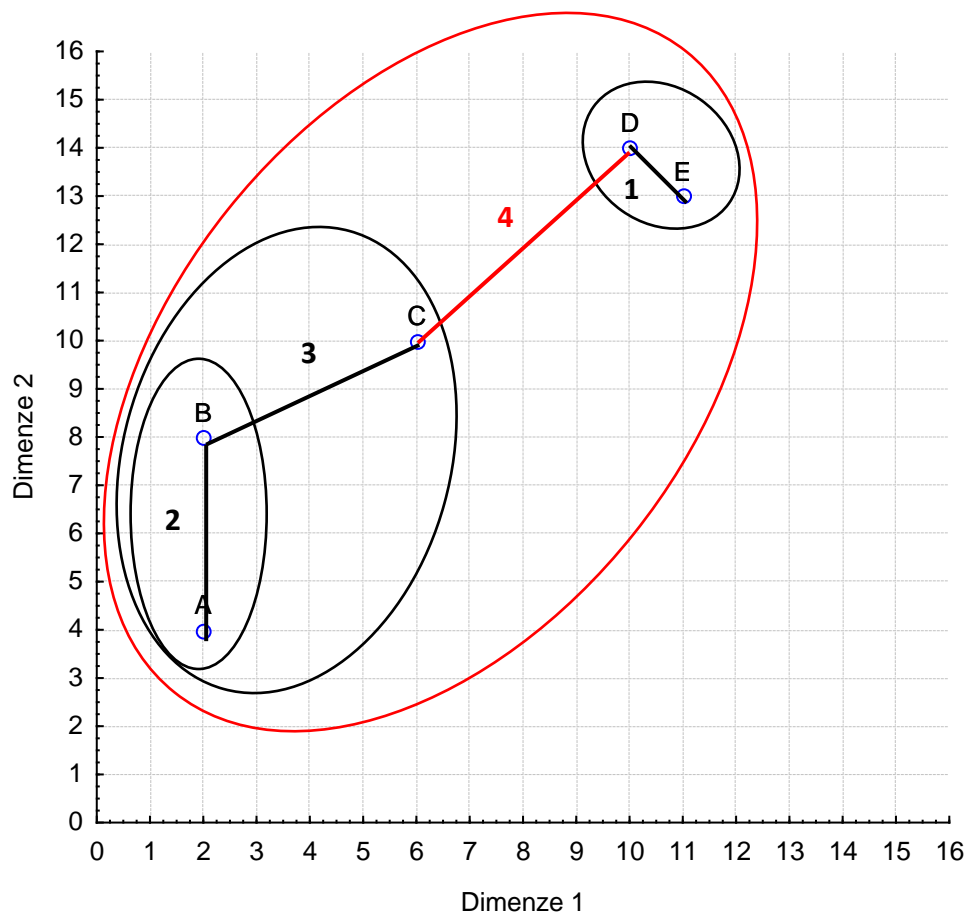
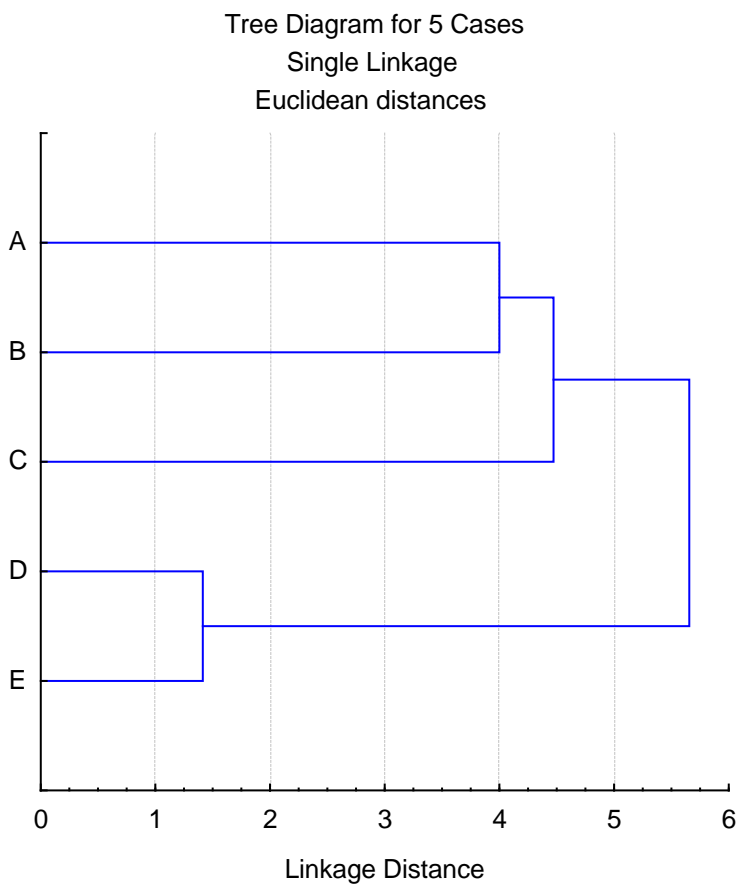
	A+B+C	D+E
A+B+C	0.0	5.7
D+E	5.7	0.0

- 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



Vícerozměrné statistické metody

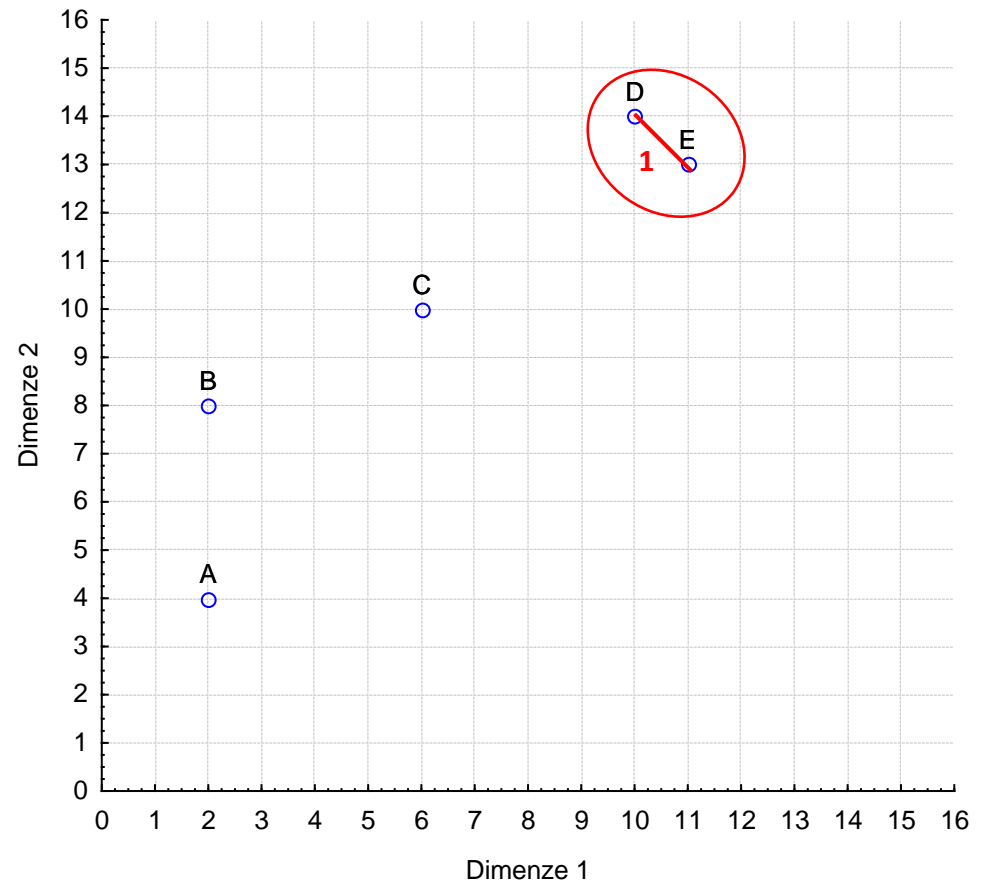
Hierarchické aglomerativní shlukování:
Příklad výpočtu metody nejvzdálenějšího souseda

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

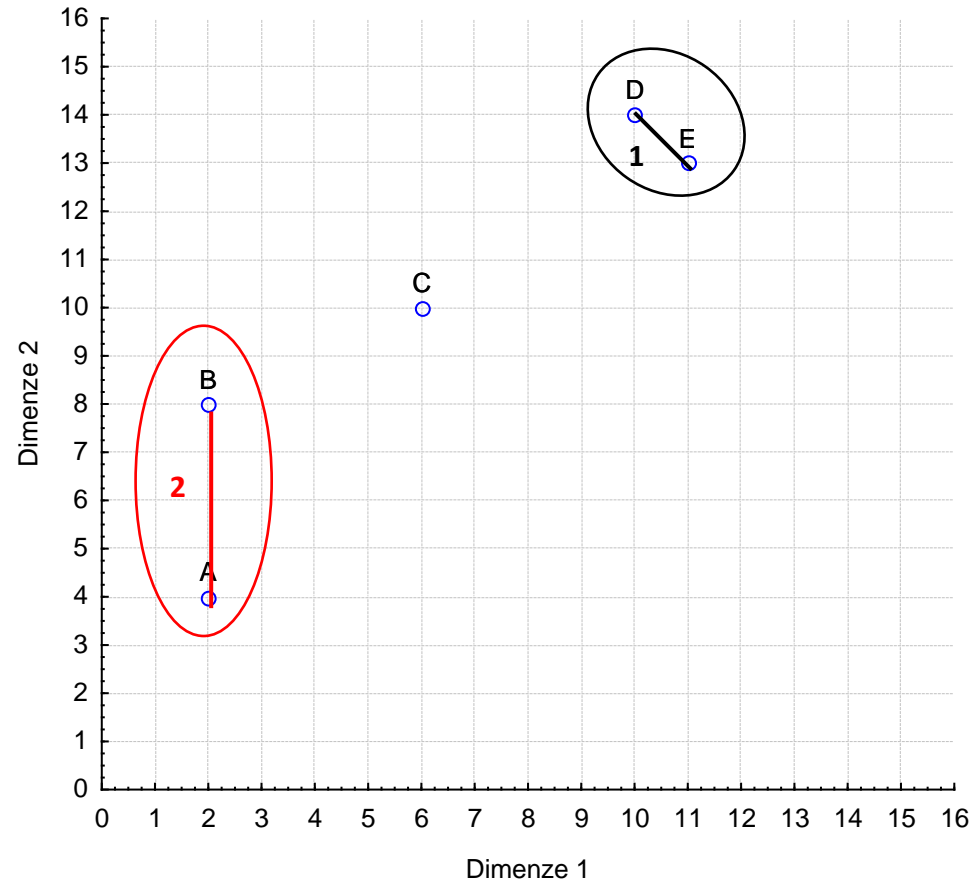


Metoda nejvzdálenějšího souseda: 2. krok výpočtu

- 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
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

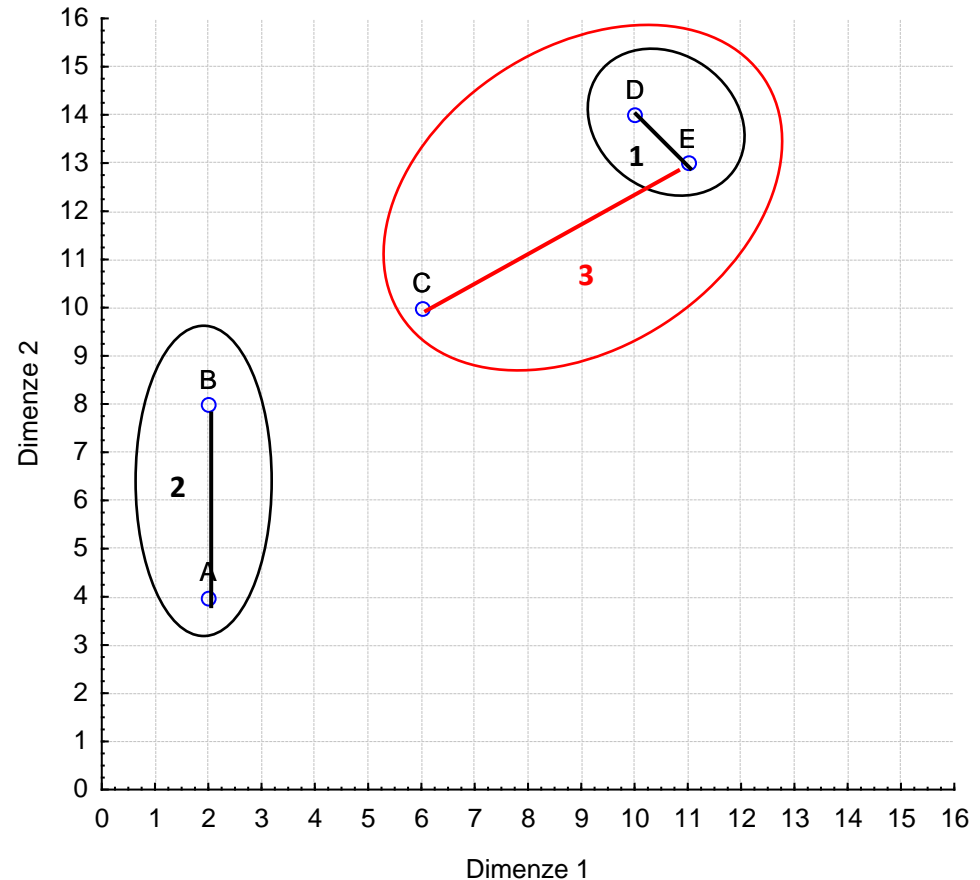


Metoda nejvzdálenějšího souseda: 3. krok výpočtu

- 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+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**

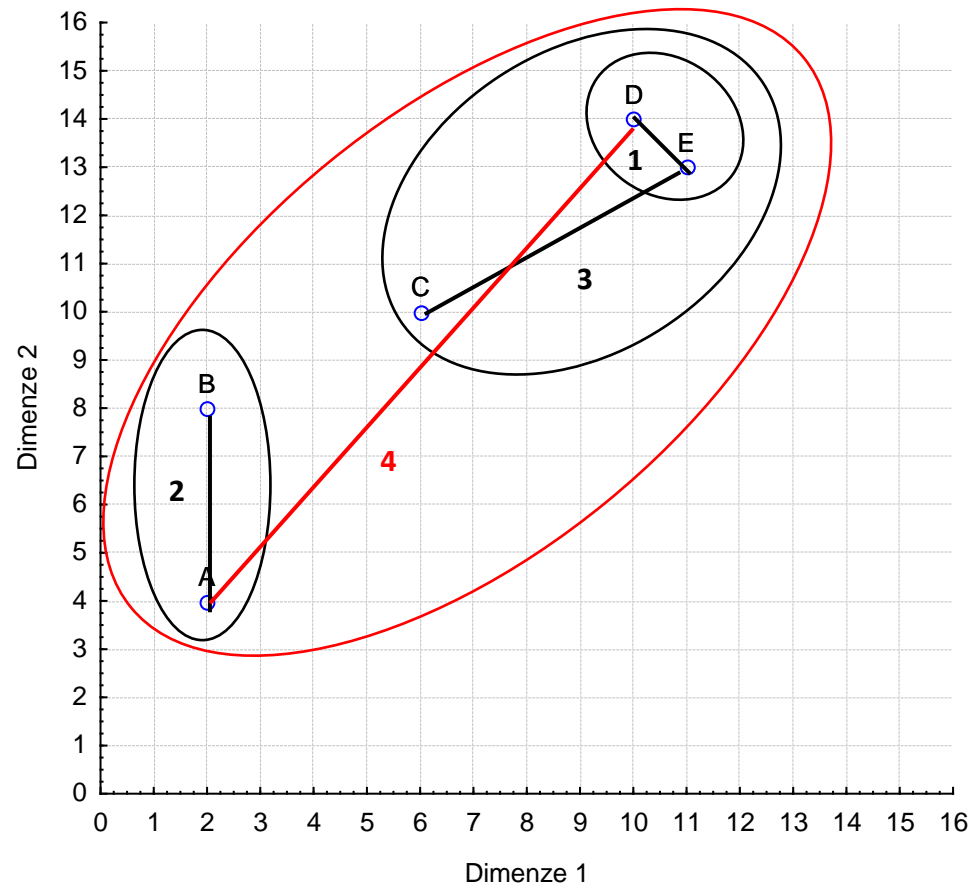


Metoda nejvzdálenějšího souseda: 4. krok výpočtu

- 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)**

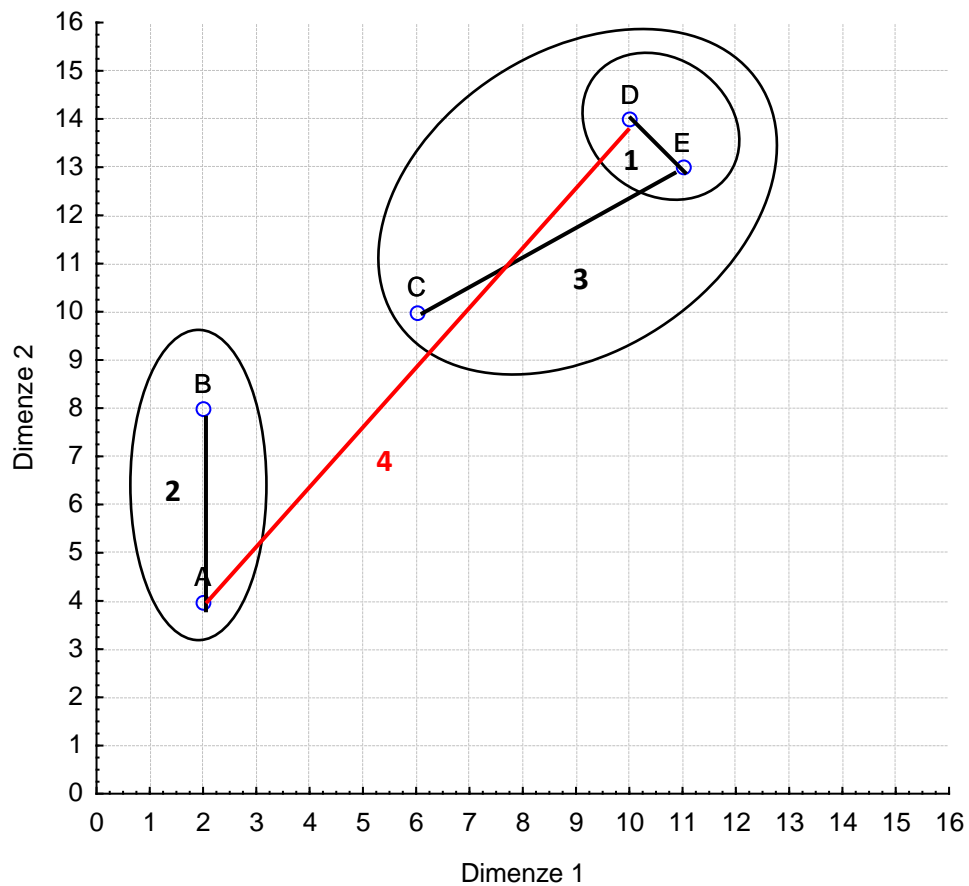
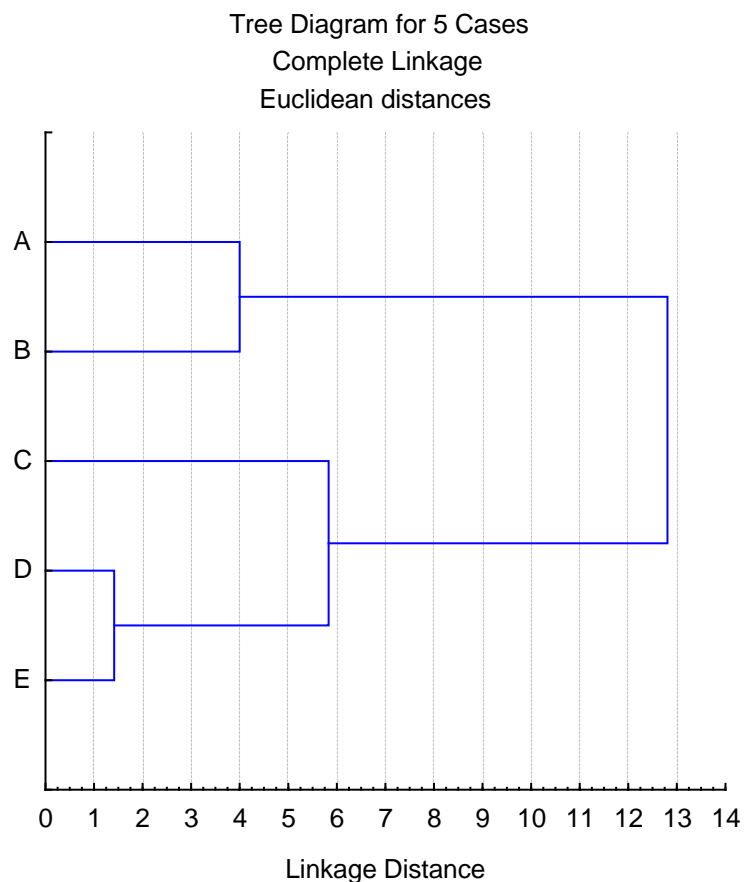
	A+B	D+E+C
A+B	0.0	12.8
D+E+C	12.8	0.0

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



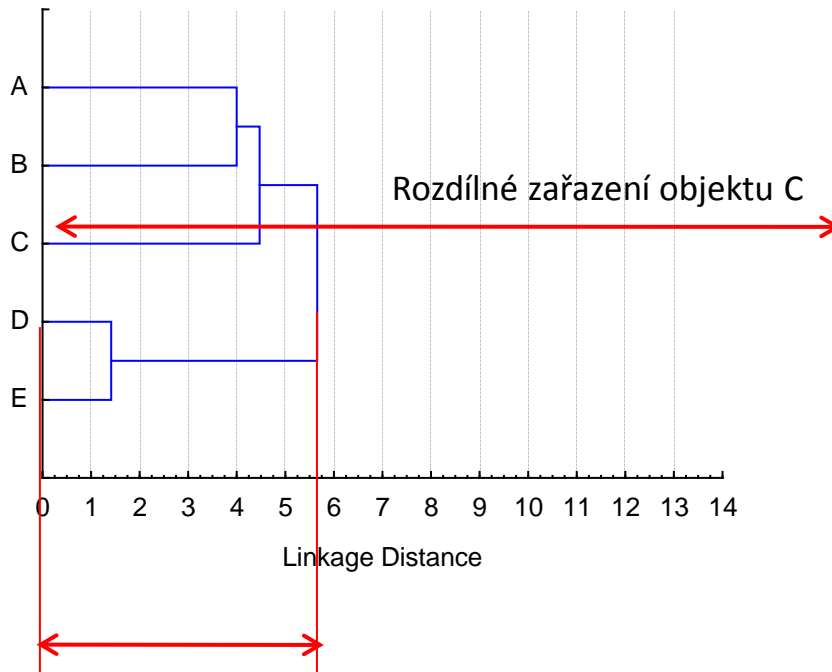
Metoda nejvzdálenějšího souseda: výsledek analýzy

- Výsledek analýzy je vizualizován ve formě dendrogramu



Metoda nejbližšího a nejvzdálenějšího souseda: interpretace výsledků

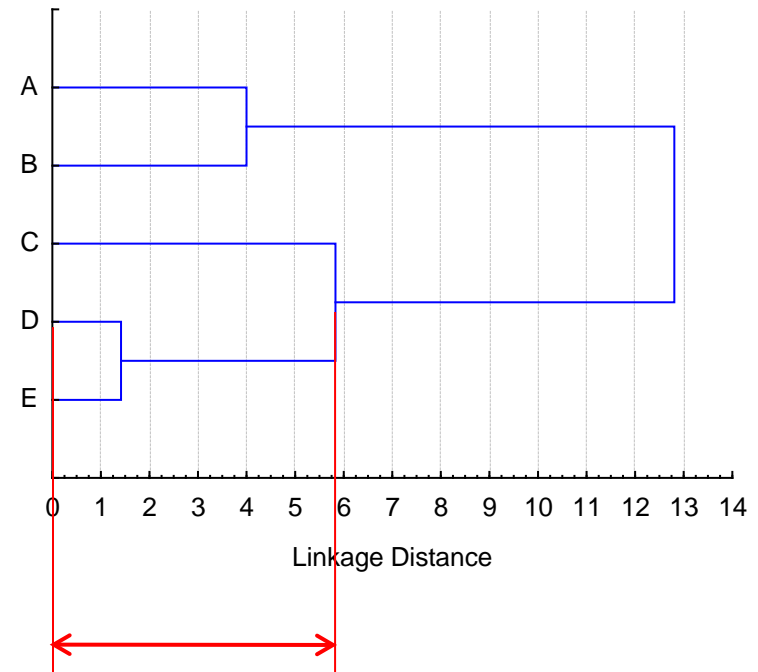
Metoda nejbližšího souseda



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 než je tato vzdálenost

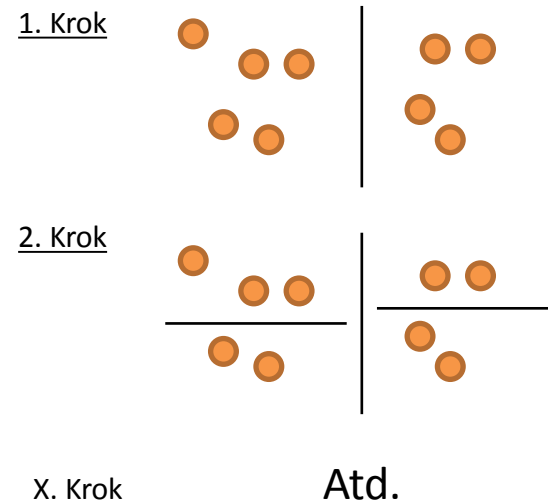
Vícerozměrné statistické metody

Hierarchické divizivní shlukování

Hierarchické divizivní shlukování: postup

- Hierarchická divizivní shlukování fungují na principu výpočtu ordinační analýzy a dělení objektů podle os ordinačního prostoru, tedy dle směrů největší variability v datech
- Shlukování může být zastaveno po rozdělení všech objektů do shluků, po předem daném počtu kroků nebo po dosažení kritéria minimálního rozdílu mezi shluky
- Typickým příkladem je metoda TWINSpan používaná v analýzách biologických společenstev

Obecný postup hierarchického divizivního shlukování

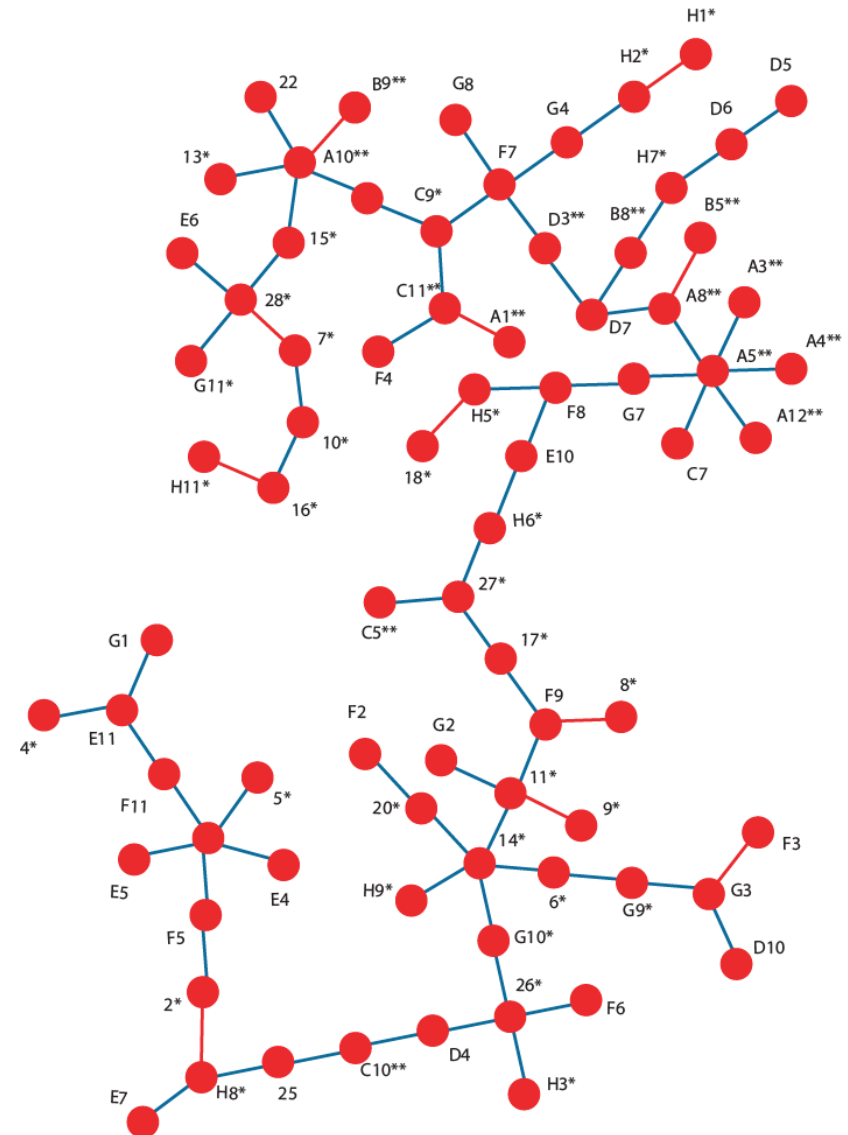


Vícerozměrné statistické metody

Nehierarchické aglomerativní shlukování

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 (Prim network)



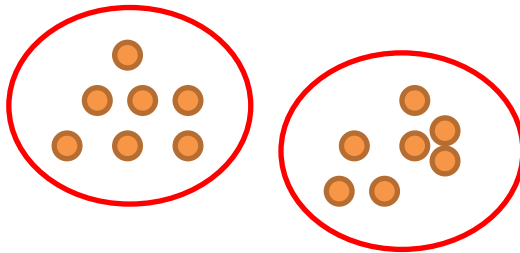
Vícerozměrné statistické metody

Nehierarchické divizivní shlukování

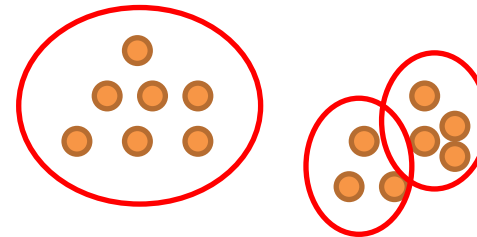
Nehierarchické divizivní shlukování: postup

- Nejběžnější metodu je tzv. k-means clustering
- 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)
- V prvním kroku je určeno k objektů jako počáteční středy shluků (výběr může být náhodný, daný uživatelem nebo maximalizující počáteční vzdálenosti k objektů)
- Následně jsou objekty zařazeny do k shluků tak, aby byla minimalizována suma čtverců vzdáleností objektů k centroidům jejich shluků
- Výpočet vzdáleností probíhá na bázi Euklidovské vzdálenosti, pro k-means clustering na jiné metrice vzdálenosti/podobnosti je nezbytná kombinace s jinými metodami

K-means k=2



K-means k=3



Analýza vždy nalezne zadaný počet shluků, i když výsledek nemusí být vždy prakticky smysluplný