Bi8600: Vícerozměrné metody – cvičení

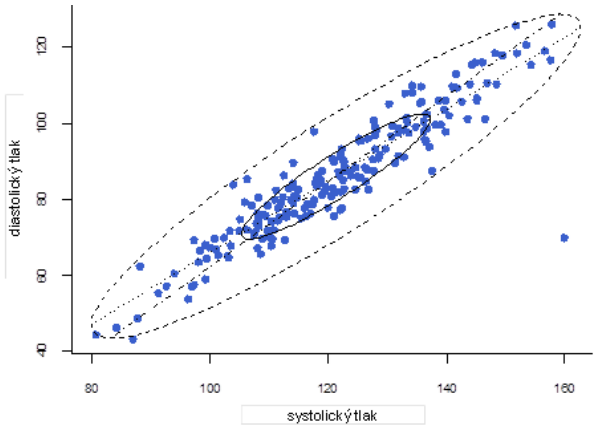
Vícerozměrné rozdělení dat
Koeficienty podobnosti a vzdálenosti
Asociační matice
Shluková analýza

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, S. Littnerová, L. Brožová

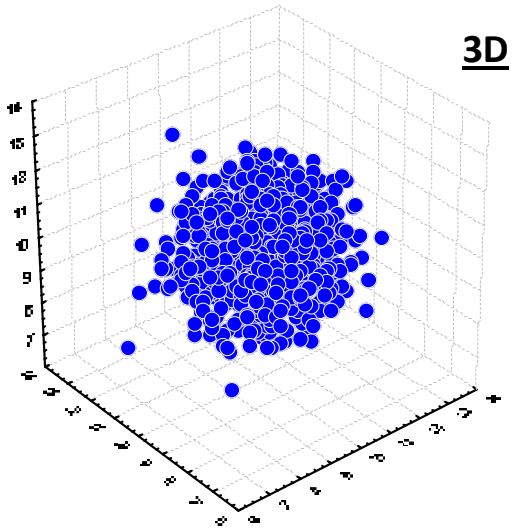
Jak vizualizujeme vícerozměrný prostor?

Jak vizualizujeme vícerozměrný prostor?

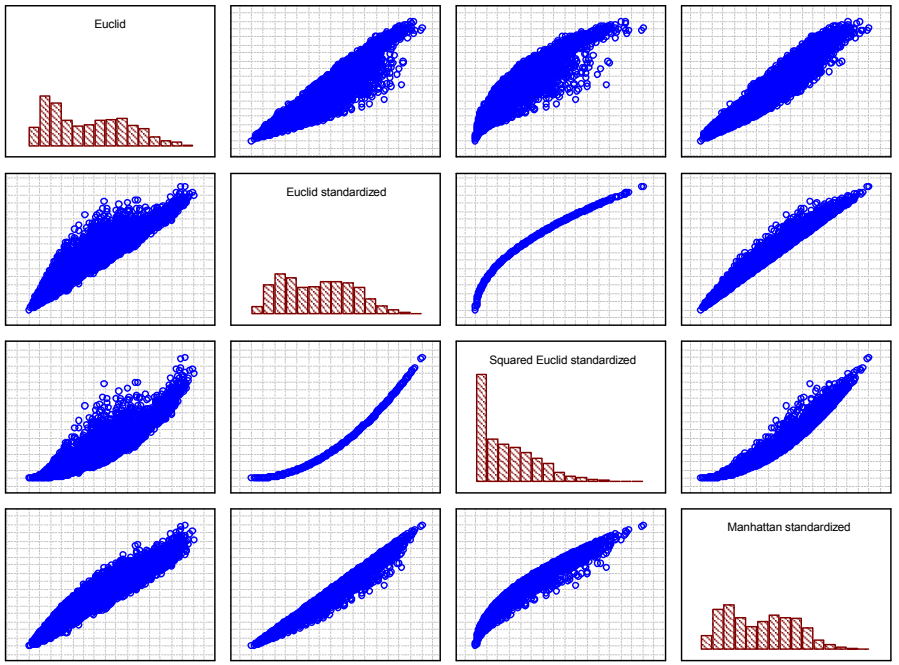
2D



3D



Maticové grafy



Jak popíšeme vícerozměrný prostor?

Popisné statistiky vícerozměrných dat

Charakteristiky polohy středu

- Udávají, kolem jaké hodnoty se data centrují.
- Centroid = vektor průměrných hodnot (mediánů), reprezentuje virtuální střed.
- Medoid = reprezentuje reálný objekt.

Charakteristiky variability

- Zachycují rozptýlení hodnot v souboru (proměnlivost dat).
- Kovarianční matice.

Vícerozměrná normalita dat

- Vícerozměrné normální rozdělení pro proměnné x_1, …, x_p popíšeme vektorem středních hodnot a kovarianční maticí.

Vektor středních hodnot
(odhadem je vektor průměrů)

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Kovarianční matice
(odhadem je výběrová kovarianční matice)

$$C = \begin{pmatrix} D(x_1) & C(x_1,x_2) & \cdots & C(x_1,x_p) \\ C(x_2,x_1) & D(x_2) & \cdots & C(x_2,x_p) \\ \vdots & \vdots & \ddots & \vdots \\ C(x_p,x_1) & C(x_p,x_2) & \cdots & D(x_p) \end{pmatrix}$$

Kovariance páru proměnných

Rozptyl proměnných

- Do popisu dat navíc vstupují charakteristiky vztahu proměnných.
- Kovariance popisuje vztah dvou proměnných; její rozsah závisí na variabilitě dat.

$$C(x_1,x_2) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{n-1}; C \in (-\infty;\infty)$$

Jaký je vztah mezi kovariancí a korelací?

Jaký je vztah mezi kovariancí a korelací?

- **Kovariance** popisuje vztah dvou proměnných; její rozsah závisí na variabilitě dat.

$$C(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2)}{n-1}; C \in (-\infty; \infty)$$

- **Korelace** = kovariance standardizovaná na rozptyl proměnných.

$$r(x_1, x_2) = \frac{C(x_1, x_2)}{\sqrt{D(x_1)}\sqrt{D(x_2)}}; r \in \langle -1; 1 \rangle$$

- Jaké hodnoty se nachází na diagonále korelační matice?
- Má smysl použít metody redukce dimenzionality dat v situaci, kdy jsou hodnoty kovariance/korelace blízké nule?
- Čemu odpovídá kovariance na standardizovaných datech?

Jaký je vztah mezi kovariancí a korelací?

- **Kovariance** popisuje vztah dvou proměnných; její rozsah závisí na variabilitě dat

$$C(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2)}{n-1}; C \in (-\infty; \infty)$$

- **Korelace** = kovariance standardizovaná na rozptyl proměnných.

$$r(x_1, x_2) = \frac{C(x_1, x_2)}{\sqrt{D(x_1)}\sqrt{D(x_2)}}; r \in \langle -1; 1 \rangle$$

- Jaké hodnoty se nachází na diagonále korelační matice?
- Má smysl použít metody redukce dimenzionality dat v situaci, kdy jsou hodnoty kovariance/korelace blízké nule?
- Čemu odpovídá kovariance na standardizovaných datech?

→ Pokud $D(x_1)=D(x_2)=1$ → kovariance = korelace

Jaké jsou dva základní přístupy hodnocení vícer. dat?

Parametry (znaky)

Základní jednotka dat

Pacient	Clovek	aLeu	aTy%	aSe%	aNeu%	aLy%	aTy	aSe	aNeu	aLy	aHtc	aCLsk	aCLNeus	aCLOZ	aCLNeuO
		cell.10^6/	%	%	%	%	cell.10^6/	cell.10^6/	cell.10^6/	cell.10^6/	%	mV.s.10^3	mV.s.10^3	mV.s.10^3	mV.s.10^3
3	1	4									33	72		32	
4	2	7,6	8	58	66	24	0,6	4,4	5,0	1,8	33	95	19	48	10
8	3	4	3	52	55	40	0,1	2,1	2,2	1,6	22	77	35	33	15
11	4	6,1	5	59	64	35	0,3	3,6	3,9	2,1	33	103	26	49	13
12	5	6,9	3	85	88	9	0,2	5,9	6,1	0,6	37	81	13	45	7
14	6	5,9	15	55	70	19	0,9	3,3	4,1	1,1	32	137	33	61	15
16	7	8	18	75	93	7	1,4	6,0	7,4	0,6	34	151	20	59	8
20	8	9,6	3	72	75	23	0,3	6,9	7,2	2,2	40	77	11	38	5
21	9	6	10	67	77	19	0,6	4,0	4,6	1,1	32	120	26	52	11
22	10	3,3	4	55	59	39	0,1	1,8	2,0	1,3	28	81	42	24	12
37	11	3,8	10	60	70	30	0,4	2,3	2,7	1,1	32	111	42	29	11
38	12	6,4	2	76	78	17	0,1	4,9	5,0	1,1	25	366	73	115	23
39	13	6,8	1	57	58	39	0,1	3,9	3,9	2,7	20	234	59	71	18
49	14	8,5	7	67	74	26	0,6	5,7	6,3	2,2	30	156	25	108	17
51	15	9,3	7	57	64	35	0,7	5,3	6,0	3,3	35	129	21	23	4
52	16	2,2	10	56	66	34	0,2	1,2	1,5	0,7	33	46	30	12	8
55	17	9,9	3	78	81	10	0,3	7,7	8,0	0,1	30	189	24	140	18
56	18	5	2	80	82	13	0,1	4,0	4,1	0,7	26	101	25	54	13
6	1	8,8	11	72	83	12	1,0	6,3	7,3	1,1	44	268	36,6	145	19,9
9	2	9,2	2	66	68	28	0,2	6,1	6,3	2,6	42	168	26,9	76	12,2
13	3	10,0	7	83	90	8	0,7	8,3	9,0	0,8	54	181	20,1	81	9
15	4	9,6	1	75	76	23	0,1	7,2	7,3	2,2	45	343	47	124	16,9
17	5	6,0									45	40		21	
19	6	7,2	2	78	80	18	0,1	5,6	5,8	1,3	44	103	17,8	63	10,9
24	7	8,2	1	72	73	25	0,1	5,9	6,0	2,1	41	209	34,9	57	9,6
26	8	10,3	1	85	86	3	0,1	8,8	8,9	0,3	41	364	41,1	112	12,6
29	9	5,0	1	74	75	21	0,1	3,7	3,8	1,1	39	83	22,1	32	8,5
30	10	11,9	1	51	52	47	0,1	6,1	6,2	5,6	33	83	13,4	52	8,4
31	11	7,2	3	53	56	29	0,2	3,8	4,0	2,1	28	109	27,1	63	15,5
32	12	10,8	36	50	76	8	3,9	5,4	9,3	0,9	27	146	15,7	106	11,4
33	13	11,8	22	54	76	16	2,6	6,4	9,0	1,9	45	246	27,4	63	7
34	14	17,0	1	82	83	16	0,2	13,9	14,1	2,7	34	440	31,2	119	8,4
40	15	10,0	8	72	80	4	0,8	7,2	8,0	0,4	37	176	22,0	52	6,5

Co je cílem analýzy?

- **Vstupní matice: řádky = objekty, sloupce = proměnné**

1) **R mode:** hodnotíme závislost …… → ordinační analýzy
2) **Q mode:** hodnotíme vzdálenost/podobnost …… → shluková analýza

Co je cílem analýzy?

- **Vstupní matice: řádky = objekty, sloupce = proměnné**

SHLUKOVÁ ANALÝZA

- **Q mode:** hodnotíme vzdálenost/podobnost **objektů** → shluková analýza
- vytváření shluků objektů na základě jejich podobnosti
- identifikace typů objektů

ORDINAČNÍ METODY

- **R mode:** hodnotíme závislost **proměnných** → ordinační analýzy
- zjednodušení vícerozměrného problému do menšího počtu rozměrů
- principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat

- Základní výběr koeficientu je často spjat s metodou
- Dále je potřeba zohlednit typ vstupních dat: spojitá/kategoriální/mix
- Výběrem metriky ovlivníme výsledky analýz

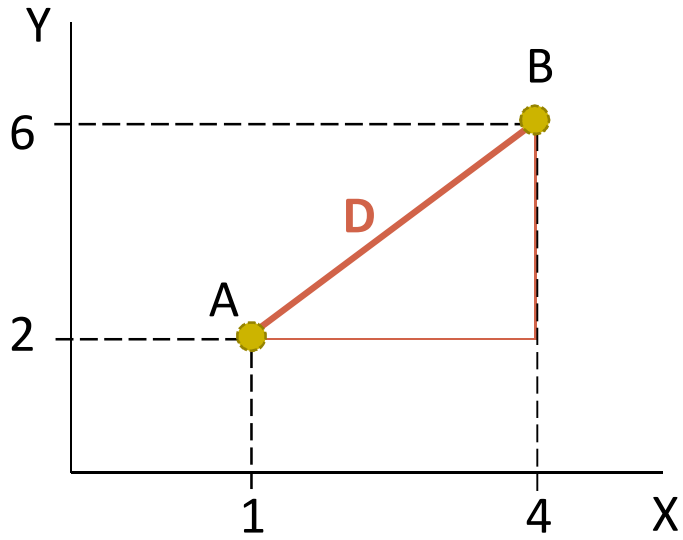
Koeficienty vzdálenosti

Kvantitativní data

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, S. Littnerová, L. Brožová

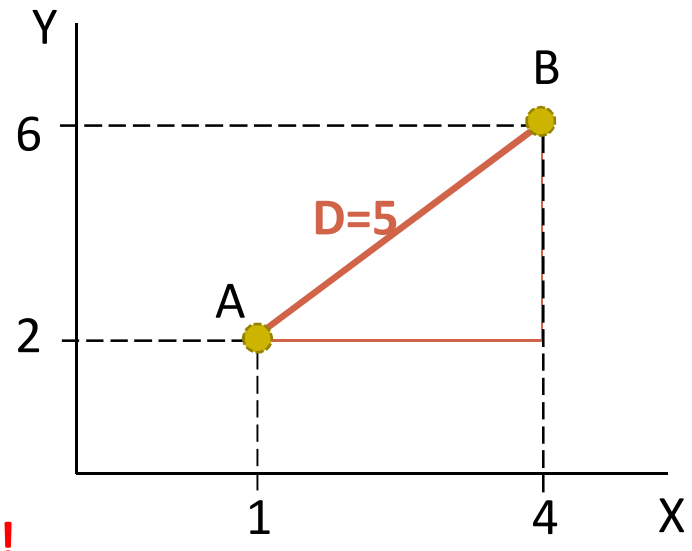
Příklad 1: Euklidova vzdálenost

- Euklidova vzdálenost vychází z Pythagorovy věty
- **Úkol**: spočítejte vzdálenost (D) objektu A[1;2] a B[4;6]



Příklad 1: Euklidova vzdálenost

- Euklidova vzdálenost vychází z Pythagorovy věty
- **Úkol**: spočítejte vzdálenost (D) objektu A[1;2] a B[4;6]



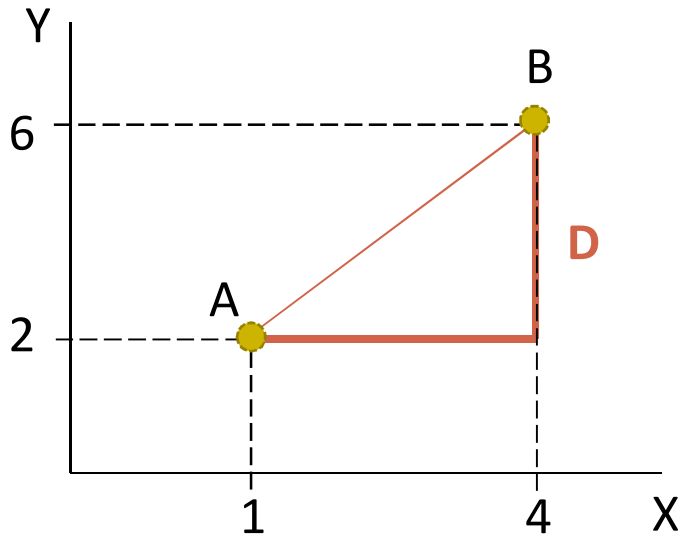
$$D\ (A,B) = \sqrt{(4-1)^2 + (6-2)^2} = 5$$

POZOR!

- Proměnné s číselně většími hodnotami budou mít větší váhu při shlukování!!!
- Např. pokud budeme hodnotit výšku (150–200 cm) a cholesterol (do 5 mmol/l), výška bude mít větší váhu při shlukování – objekty budou rozděleny do shluků podle jejich výšky.
- Data s nesrovnatelnými hodnotami proměnných je potřeba před analýzou **standardizovat**. Jak?
  - standardizace směrodatnou odchylkou nebo rozpětím

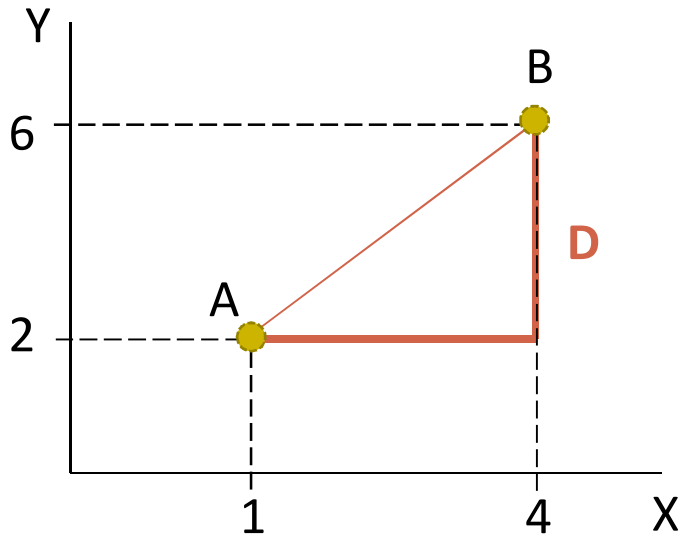
Příklad 2: Manhattanská vzdálenost

- Cesta po Manhattanu
- součet absolutních hodnot rozdílů jednotlivých parametrů popisujících objekty
- **Úkol**: spočítejte vzdálenost (D) objektu A [1;2] a B[4;6]



Příklad 2: Manhattanská vzdálenost

- Cesta po Manhattanu
- součet absolutních hodnot rozdílů jednotlivých parametrů popisujících objekty
- **Úkol**: spočítejte vzdálenost (D) objektu A [1;2] a B[4;6]



$$D\ (A,B) = |4-1| + |6-2| = 7$$

Koeficienty podobnosti

Binární data

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, S. Littnerová, L. Brožová

Koeficienty podobnosti

- Jaké znáte dva hlavní typy koeficientů podobnosti?
- Co znamená double zerow problem?

Koeficienty podobnosti

- Ve vícerozměrné analýze se využívá řada indexů podobnosti založených na přítomnosti/nepřítomnosti kategorií objektů

Lokalita 1 \ Lokalita 2	1	0
1	a	b
0	c	d

a, b, c, d = počet případů, kdy souhlasí binární charakteristika společenstev 1 a 2
a+b+c+d=p

Symetrické binární koeficienty - není rozdíl mezi případem 1-1 a 0-0
Asymetrické binární koeficienty - rozdíl mezi případem 1-1 a 0-0

- Velký počet koeficientů, které dávají různou váhu jednotlivým kombinacím

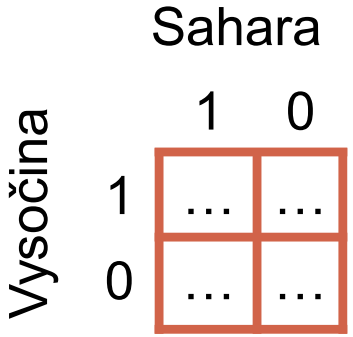
Podrobný přehled koeficientů vzdáleností a podobností najdete v knize **LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elseviere Science BV, Amsterodam.**

Příklad 3: Koeficienty podobnosti

- Tabulka popisuje výskyt (1) nebo nevýskyt (0) živočichů na lokalitách.
- **Úkol**: Pomocí Simple matching, Jaccardova a Sørensenova koeficientu vyhodnoťte, zda si jsou uvedené lokality podobné. Výsledné hodnoty podobností převeďte na vzdálenosti.

	Výskyt/nevýskyt živočicha						
Lokalita	**žralok**	**velryba**	**had**	**ještěrka**	**velbloud**	**varan**	**tučňák**
Vysočina	0	0	1	1	0	0	0
Sahara	0	0	1	1	1	1	0

- Vyplňte počty případů, aby výskyt/nevýskyt živočicha odpovídal vstupní tabulce.



- $S_{simple\ matching}$ = …
- $S_{Jaccard}$ = …
- $S_{Sørensen}$ = …

Příklad 3: Koeficienty podobnosti

- Tabulka popisuje výskyt (1) nebo nevýskyt (0) živočichů na dvou lokalitách.
- **Úkol**: Pomocí Simple matching a Jaccardova koeficientu vyhodnoťte, zda si jsou uvedené lokality podobné. Výsledné hodnoty podobností převeďte na vzdálenosti.

	Výskyt/nevýskyt živočicha						
Lokalita	**žralok**	**velryba**	**had**	**ještěrka**	**velbloud**	**varan**	**tučňák**
Vysočina	0	0	1	1	0	0	0
Sahara	0	0	1	1	1	1	0

- Vyplňte počty případů, aby výskyt/nevýskyt živočicha odpovídal vstupní tabulce.

Vysočina \ Sahara	1	0
1	2	0
0	2	3

- $S_{simple\ matching}$= (2+3)/(2+2+3+0)=0.7 → D=0.3
- $S_{Jaccard}$= 2/(2+2+0)=0.5 → D=0.5
- $S_{Sørensen}$=2*2/(2*2+2+0)=0.7 → D=0.3

Příklad 4: Sørensenův asymetrický koeficient podobnosti pro data abundancí

- Tabulka popisuje abundance živočichů na dvou lokalitách.
- **Úkol**: Pomocí Sørensenova koeficientu vyhodnoťte, zda jsi podobné uvedené lokality.

	Výskyt/nevýskyt živočicha									
Lokalita	**žralok**	**velryba**	**had**	**ještěrka**	**velbloud**	**varan**	**tučňák**	**aN**	**bN**	**jN**
Vysočina	0	0	2	3	0	0	0	5		
Sahara	0	0	4	1	5	6	0		16	
Minimum	0	0	2	1	0	0	0			3

$$C_N = \frac{2\,jN}{(aN + bN)} = \frac{2 \cdot 3}{(5+16)} = 0.3$$

Gowerův obecný koeficient podobnosti

Mix kategoriálních a kvantitativních dat

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, S. Littnerová, L. Brožová

Gowerův obecný koeficient podobnosti

- Kombinuje různé typy deskriptorů.
- Podobnost mezi dvěma objekty je vypočítána jako průměr podobností, vypočítaných pro všechny deskriptory:

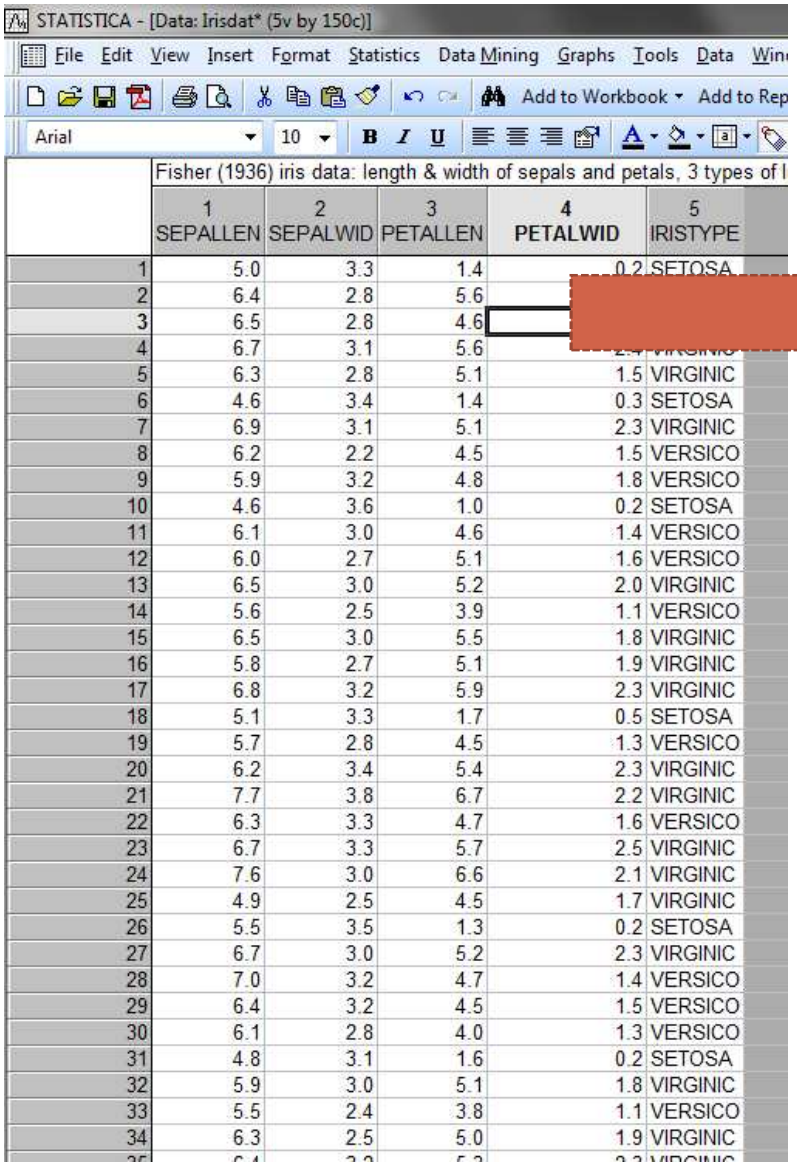
$$S_{15}(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12j}$$

- ✓ Pro kategoriální deskriptory s_j=1 (shoda) nebo 0 (neshoda).
- ✓ Kvantitativní deskriptory (reálná čísla): rozdíl mezi stavy obou objektů je vydělen největším rozdílem (R_j), nalezeným pro daný deskriptor mezi všemi objekty ve studii.

Asociační matice

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, S. Littnerová, L. Brožová

Asociační matice vzdáleností

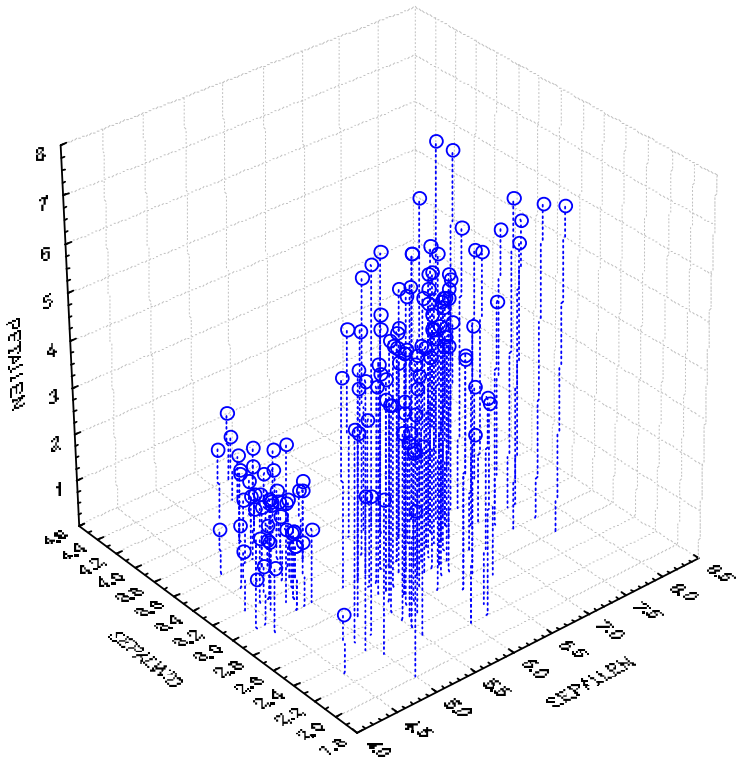


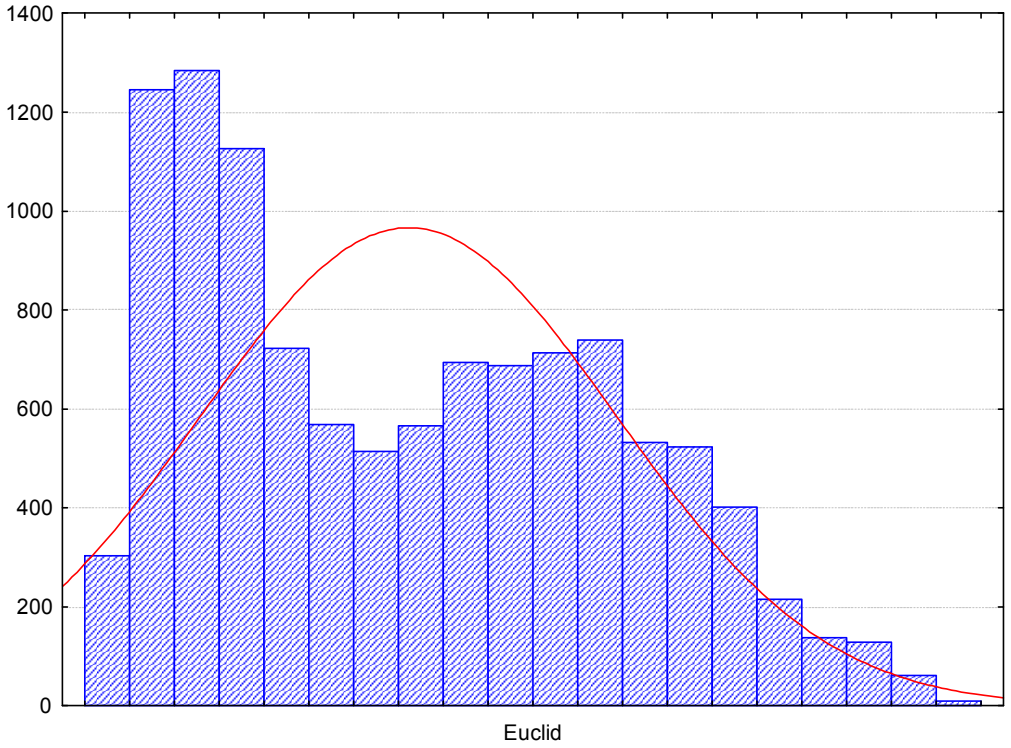
	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
1	5.0	3.3	1.4	0.2	SETOSA
2	6.4	2.8	5.6		
3	6.5	2.8	4.6		
4	6.7	3.1	5.6		
5	6.3	2.8	5.1	1.5	VIRGINIC
6	4.6	3.4	1.4	0.3	SETOSA
7	6.9	3.1	5.1	2.3	VIRGINIC
8	6.2	2.2	4.5	1.5	VERSICO
9	5.9	3.2	4.8	1.8	VERSICO
10	4.6	3.6	1.0	0.2	SETOSA
11	6.1	3.0	4.6	1.4	VERSICO
12	6.0	2.7	5.1	1.6	VERSICO
13	6.5	3.0	5.2	2.0	VIRGINIC
14	5.6	2.5	3.9	1.1	VERSICO
15	6.5	3.0	5.5	1.8	VIRGINIC
16	5.8	2.7	5.1	1.9	VIRGINIC
17	6.8	3.2	5.9	2.3	VIRGINIC
18	5.1	3.3	1.7	0.5	SETOSA
19	5.7	2.8	4.5	1.3	VERSICO
20	6.2	3.4	5.4	2.3	VIRGINIC
21	7.7	3.8	6.7	2.2	VIRGINIC
22	6.3	3.3	4.7	1.6	VERSICO
23	6.7	3.3	5.7	2.5	VIRGINIC
24	7.6	3.0	6.6	2.1	VIRGINIC
25	4.9	2.5	4.5	1.7	VIRGINIC
26	5.5	3.5	1.3	0.2	SETOSA
27	6.7	3.0	5.2	2.3	VIRGINIC
28	7.0	3.2	4.7	1.4	VERSICO
29	6.4	3.2	4.5	1.5	VERSICO
30	6.1	2.8	4.0	1.3	VERSICO
31	4.8	3.1	1.6	0.2	SETOSA
32	5.9	3.0	5.1	1.8	VIRGINIC
33	5.5	2.4	3.8	1.1	VERSICO
34	6.3	2.5	5.0	1.9	VIRGINIC

Case No.	Euclidean distances (Irisdat)																
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17
C_1	0.00	4.88	3.80	5.04	4.16	0.42	4.66	3.73	3.87	0.64	3.60	4.12	4.47	2.84	4.66	4.19	5.28
C_2	4.88	0.00	1.22	0.47	0.87	4.98	0.77	1.45	1.10	5.39	1.33	0.88	0.50	2.20	0.47	0.84	0.65
C_3	3.80	1.22	0.00	1.39	0.54	3.96	1.07	0.68	0.81	4.35	0.46	0.72	0.81	1.24	0.97	0.95	1.61
C_4	5.04	0.47	1.39	0.00	1.14	5.15	0.55	1.75	1.28	5.54	1.54	1.24	0.61	2.48	0.65	1.21	0.35
C_5	4.16	0.87	0.54	1.14	0.00	4.29	1.04	0.85	0.71	4.69	0.58	0.33	0.58	1.48	0.57	0.65	1.30
	42	4.98	3.96	5.15	4.29	0.00	4.80	3.88	3.94	0.46	3.72	4.22	4.59	2.95	4.78	4.26	5.40
		0.77	1.07	0.55	1.04	4.80	0.00	1.52	1.16	5.17	1.31	1.21	0.52	2.22	0.76	1.24	0.81
	73	1.45	0.68	1.75	0.85	3.88	1.52	0.00	1.13	4.30	0.82	0.81	1.21	0.98	1.35	0.96	1.99
C_9	3.87	1.10	0.81	1.28	0.71	3.94	1.16	1.13	0.00	4.34	0.53	0.62	0.77	1.37	0.94	0.60	1.51
C_10	0.64	5.39	4.35	5.54	4.69	0.46	5.17	4.30	4.34	0.00	4.12	4.64	4.98	3.38	5.17	4.69	5.78
C_11	3.60	1.33	0.46	1.54	0.58	3.72	1.31	0.82	0.53	4.12	0.00	0.62	0.94	1.04	1.06	0.82	1.74
C_12	4.12	0.88	0.72	1.24	0.33	4.22	1.21	0.81	0.62	4.64	0.62	0.00	0.71	1.37	0.73	0.36	1.42
C_13	4.47	0.50	0.81	0.61	0.58	4.59	0.52	1.21	0.77	4.98	0.94	0.71	0.00	1.89	0.36	0.77	0.84
C_14	2.84	2.20	1.24	2.48	1.48	2.95	2.22	0.98	1.37	3.38	1.04	1.37	1.89	0.00	2.03	1.47	2.71
C_15	4.66	0.47	0.97	0.65	0.57	4.78	0.76	1.35	0.94	5.17	1.06	0.73	0.36	2.03	0.00	0.87	0.73
C_16	4.19	0.84	0.95	1.21	0.65	4.26	1.24	0.96	0.60	4.69	0.82	0.36	0.77	1.47	0.87	0.00	1.43
C_17	5.28	0.65	1.61	0.35	1.30	5.40	0.81	1.99	1.51	5.78	1.74	1.42	0.84	2.71	0.73	1.43	0.00
C_18	0.44	4.48	3.41	4.63	3.77	0.62	4.25	3.36	3.46	0.96	3.21	3.73	4.07	2.47	4.26	3.79	4.88
C_19	3.40	1.58	0.83	1.87	0.87	3.49	1.70	0.81	0.73	3.91	0.47	0.74	1.29	0.71	1.39	0.86	2.08
C_20	4.68	0.67	1.32	0.62	1.05	4.75	0.82	1.70	0.86	5.14	1.27	1.05	0.62	2.20	0.71	0.95	0.81

- Typická asociační matice je čtvercová matice symetrická kolem diagonály
- Diagonála obsahuje 0 (v případě vzdáleností) nebo 1 (v případě podobností)

Histogram jako popis asociační matice





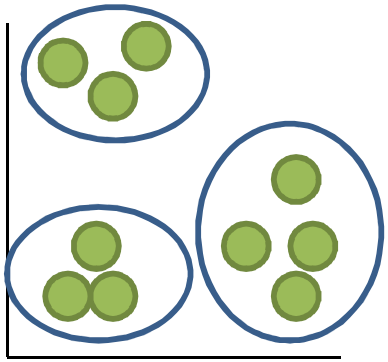
Shluková analýza

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, S. Littnerová, L. Brožová

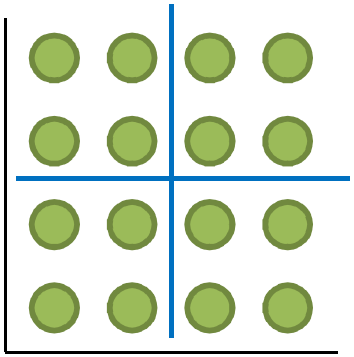
Shluková analýza – jaký je cíl?

Shluková analýza – jaký je cíl?

- Seskupení objektů do shluků podle toho, jak si jsou podobné – chceme co nejpodobnější objekty v rámci shluků a co nejodlišnější mezi shluky.
- Shluková analýza vychází z asociační matice vzdáleností objektů (Q mode) nebo závislosti parametrů (R mode).
- Můžeme provést dvě hlavní chyby: špatný výběr metriky a špatný výběr algoritmu shlukování.
- Smysluplnost výsledků shlukování závisí jednak na objektivní existenci shluků v datech, jednak na arbitrárně nastavených kritériích definice shluků.



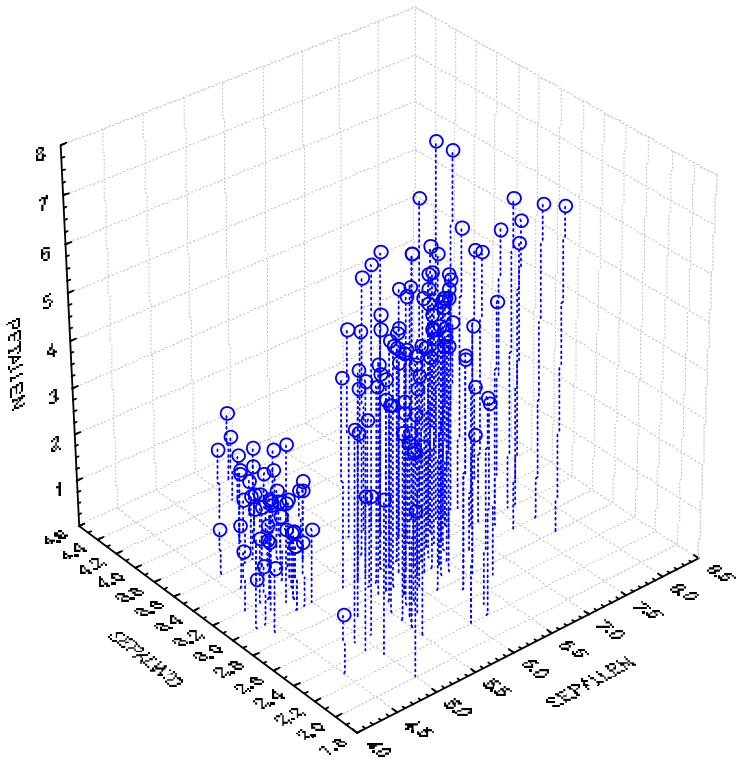
Jednoznačné odlišení existujících shluků v datech

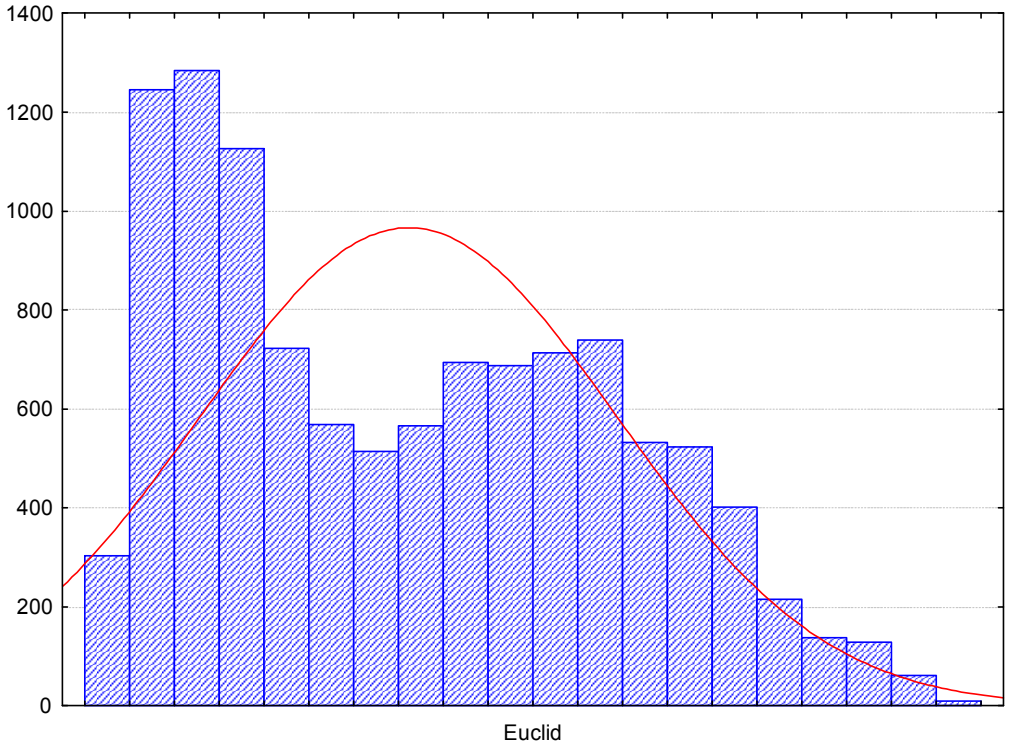


Shlukovou analýzu lze provést i na datech bez objektivní existence shluků

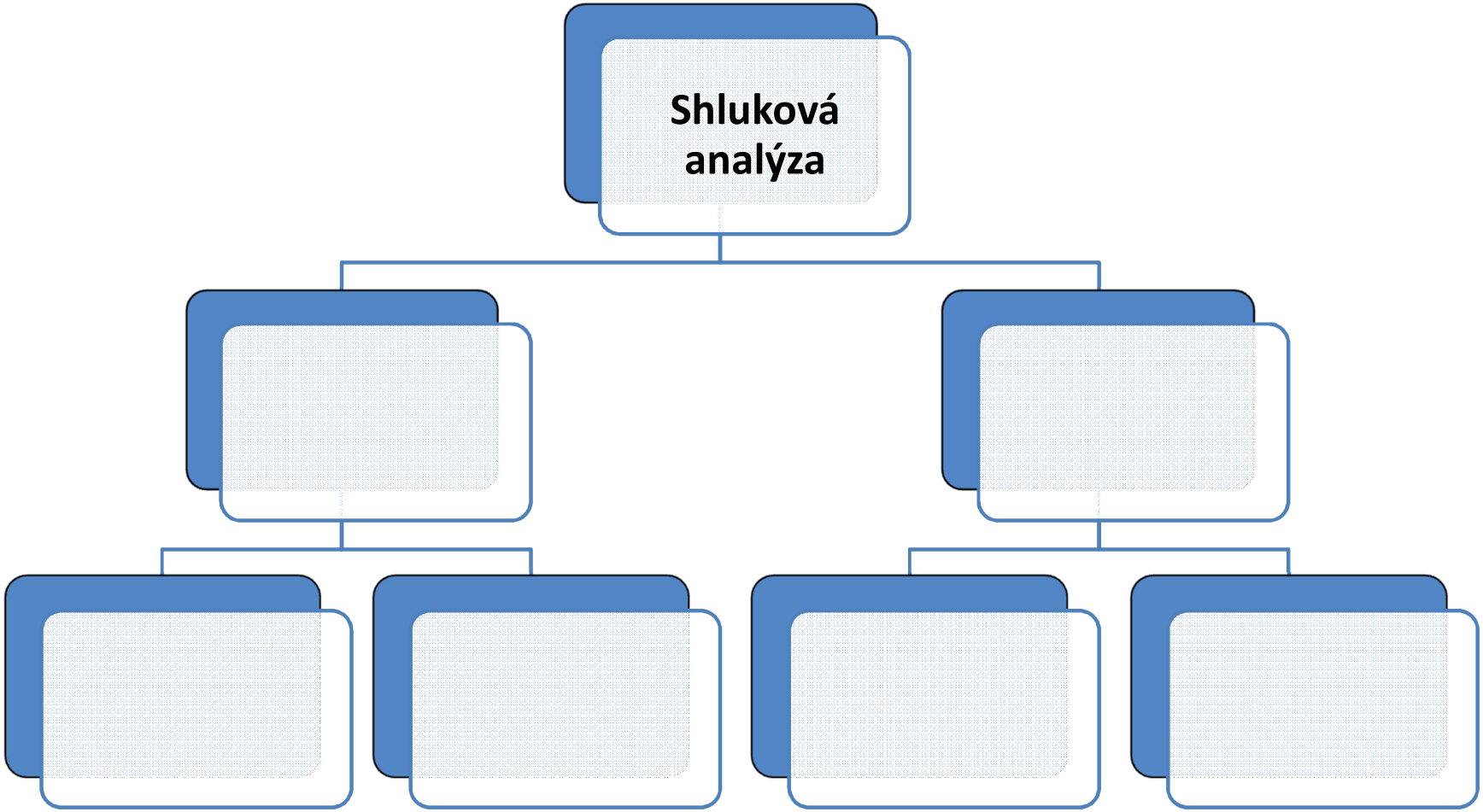
Jak ověříme, že se v datech vyskytují shluky?

Jak ověříme, že se v datech vyskytují shluky?

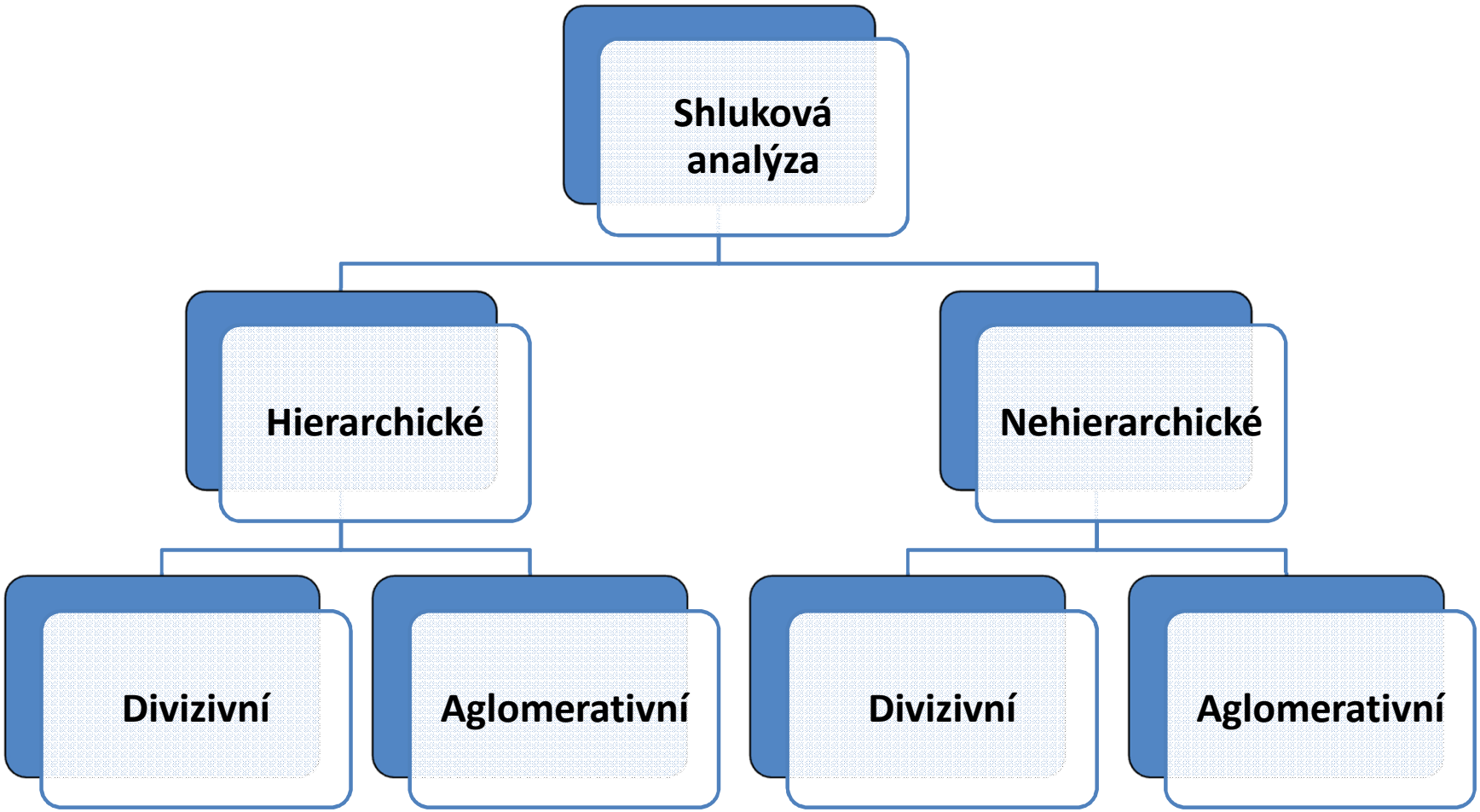




Shluková analýza: typy metod



Shluková analýza: typy metod



Shluková analýza: typy metod

Hierarchické
shluky jsou definovány postupným skládáním objektů

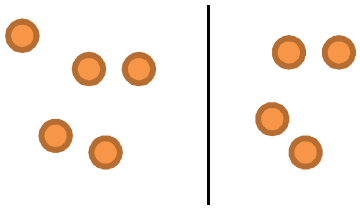
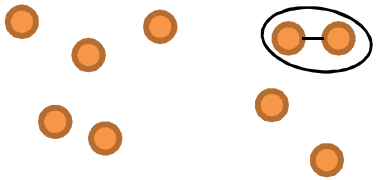
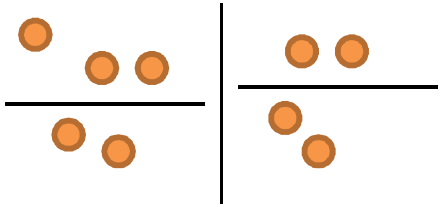
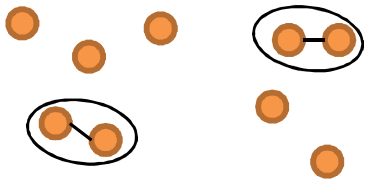
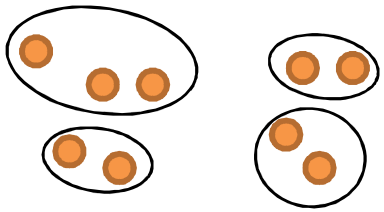
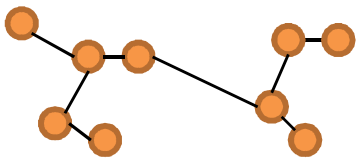
Divizivní
Objekty jsou nejprve rozděleny do dvou shluků, tyto shluky jsou dále rozděleny atd.

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

Nehierarchické
Shluky jsou definovány v jednom kroku

Divizivní
Objekty jsou rozděleny do předem nastaveného počtu shluků.

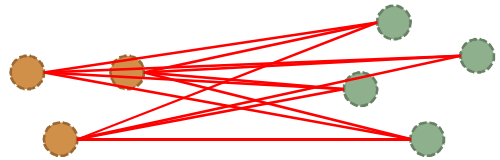
Aglomerativní
síť spojených bodů

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

Shlukovací algoritmy hierarchického aglomerativního shlukování I

1)

2)



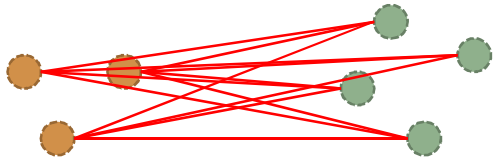
3)

4)

Shlukovací algoritmy hierarchického aglomerativního shlukování I

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

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



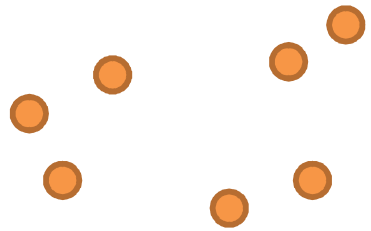
3) **Středospojná vzdálenost** (pair group centroid) – spojení dle vzdálenosti centroidů shluků

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

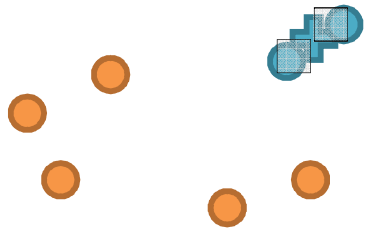
Shlukovací algoritmy hierarchického aglomerativního shlukování II: Wardova metoda

- 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 0
- Pro výpočet vzdáleností od centroidu je používán čtverec Euklidovské vzdálenosti
- Nedoporučuje se používat při hodnocení binárních dat – pracuje se vzdálenostmi v Euklidovském prostoru

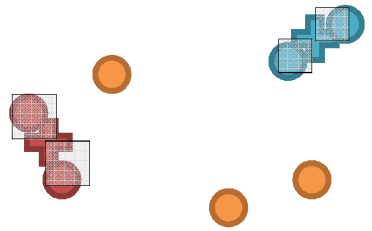
Krok 1: každý objekt je sám sobě centroidem



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

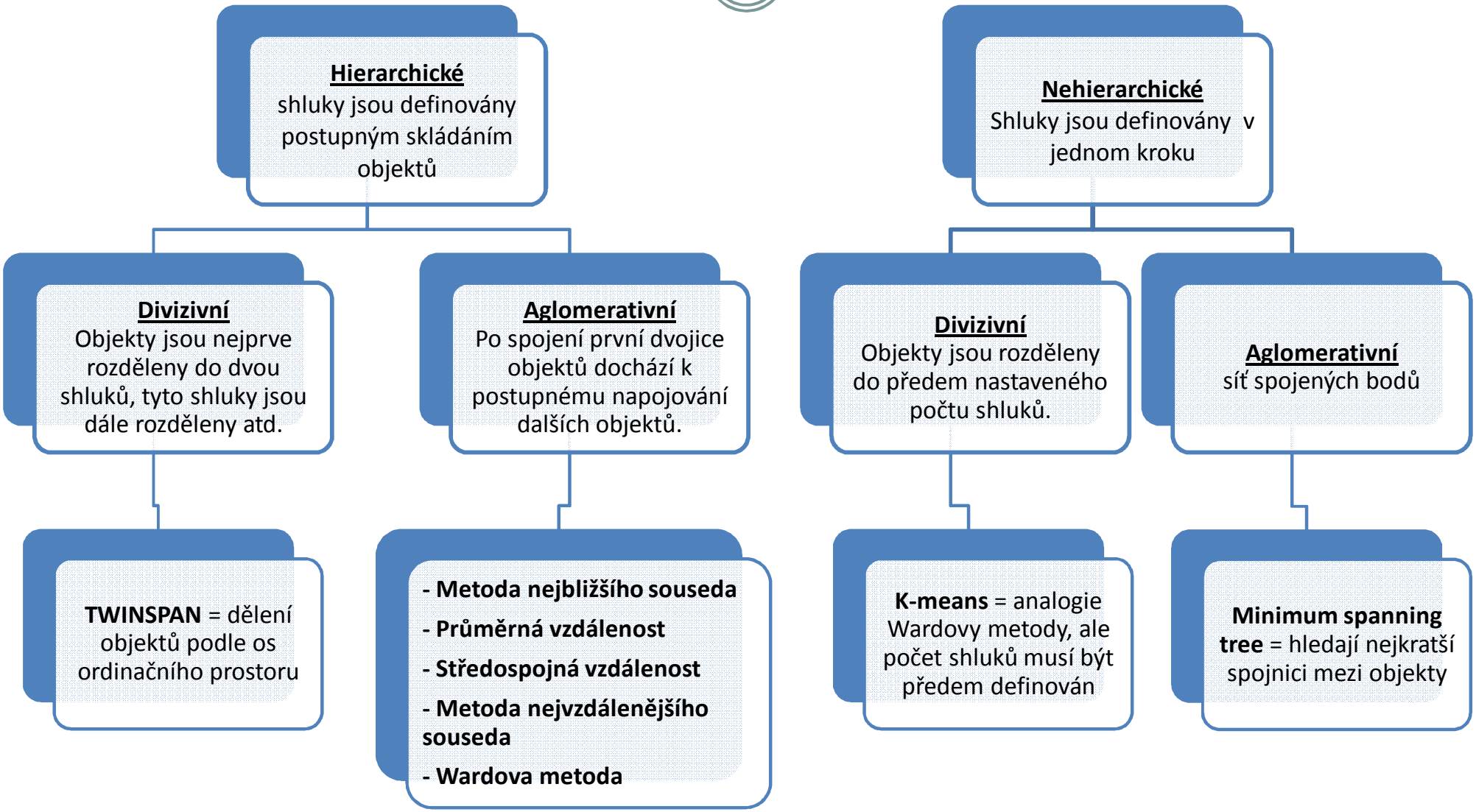


Krok 3: spojení objektů, které nejméně přispějí k sumě čtverců vzdáleností od centroidu

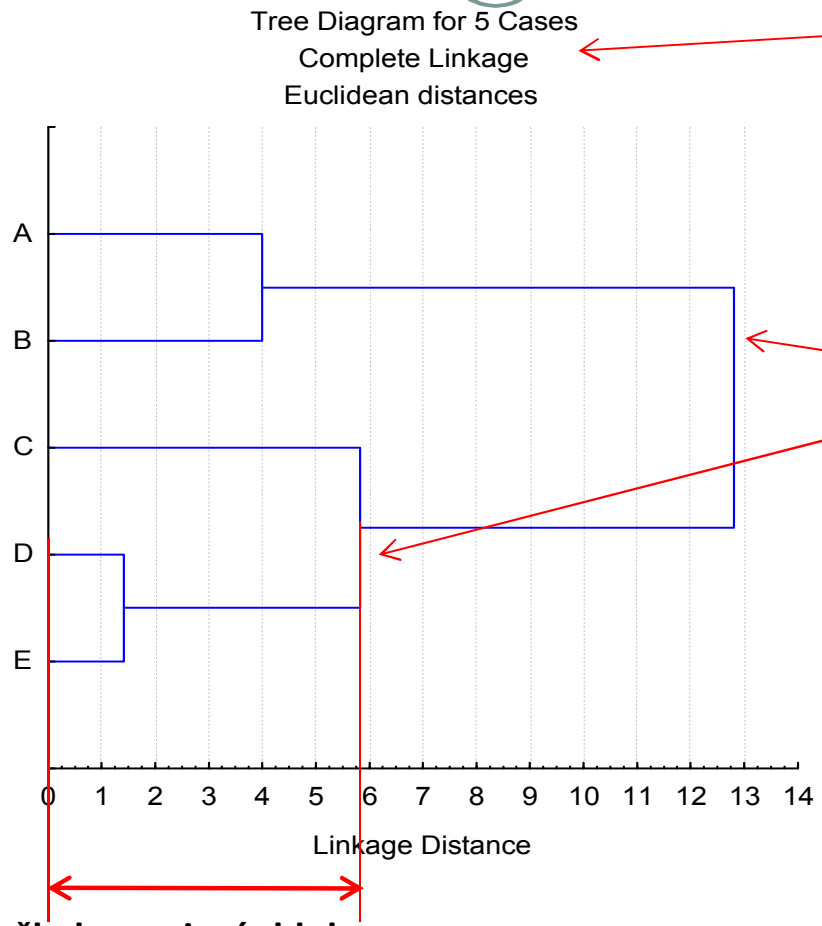


Krok 4: stejný postup až do spojení všech objektů

Shluková analýza: přehled metod



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

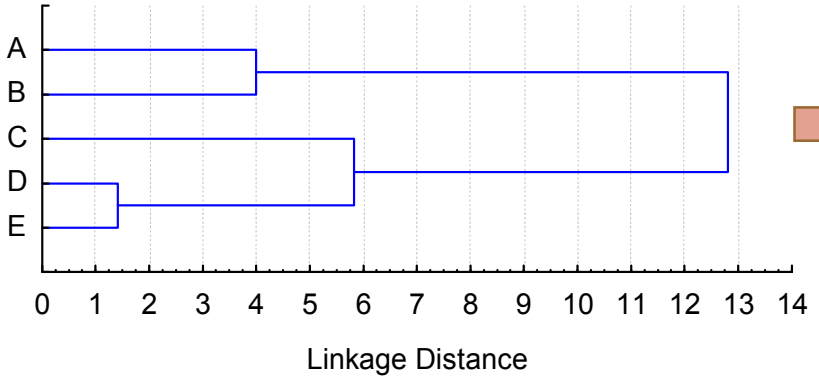
Vzdálenost, na které došlo ke shlukování

Vzdálenost na níž došlo ke spojení shluku:

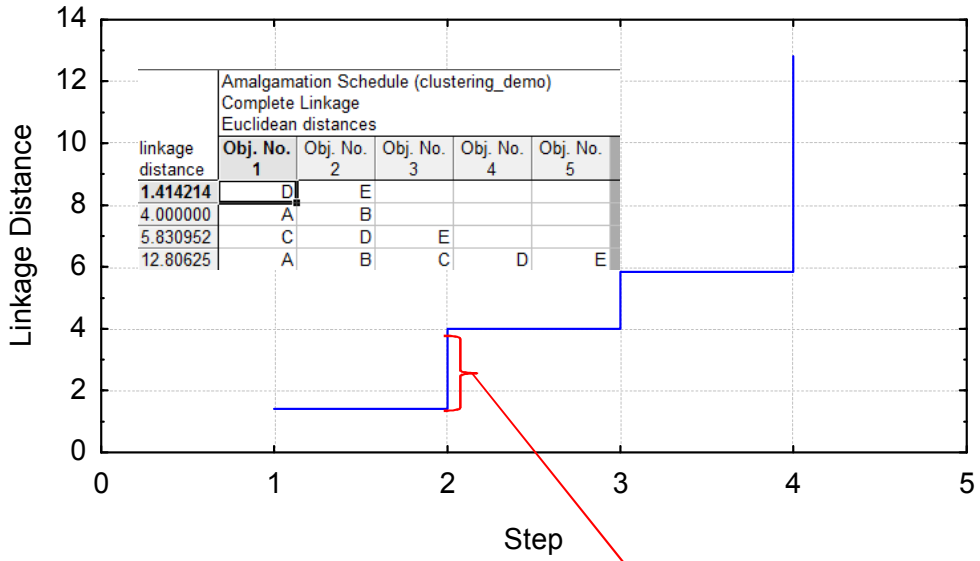
- je v rozměrech použité metriky vzdáleností/podobností a v tomto kontextu ji lze kvantitativně interpretovat
- interpretace vzdálenosti shlukování se liší podle použitého shlukovacího algoritmu
- někdy se uvádí ve škále 0-100%, kde 100% je maximální vzdálenost shlukování

Amalgamation schedule/graph

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

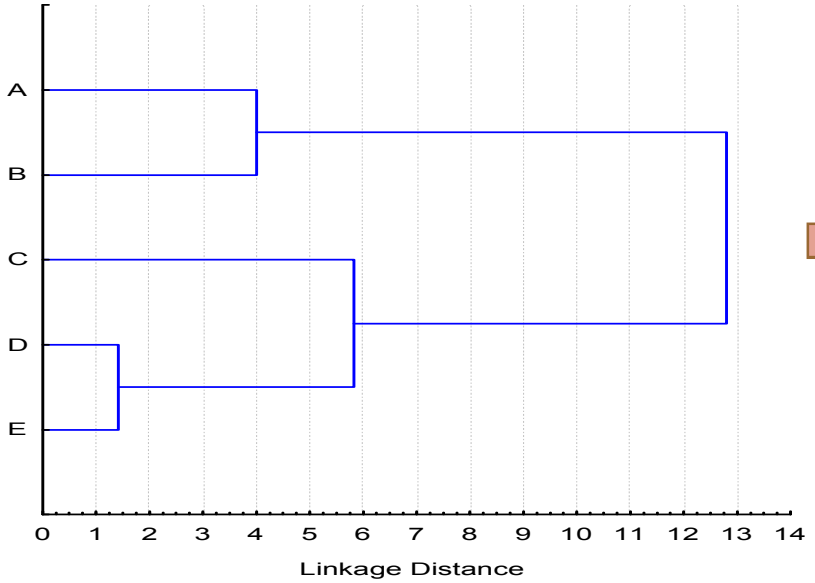
Čím delší bude svislá čára – tím více si jsou shluky spojené v tomto kroku odlišné

Výběr vhodného algoritmu

Kofenetická matice

- Matice dimenze n x n (n = počet objektů), popisující vzdálenost, kdy byl objekt poprvé zařazen do shluku.
- Hodnoty kofenetické matice závisí na typu algoritmu shlukování.

Dendrogram



Kofenetická matice

	A	B	C	D	E
A	0	4.0	12.7	12.7	12.7
B		0	12.7	12.7	12.7
C			0	5.7	5.7
D				0	1.4
E					0

Matice je symetrická podél diagonály

Vzdálenost, kdy došlo k prvnímu spojení D+C

Kofenetický index

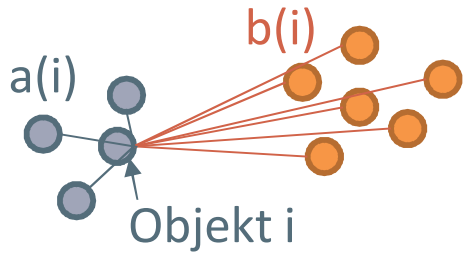
- Korelace kofenetické matice s původní maticí vzdáleností.
- Čím vyšší korelace, tím lepší algoritmus (-> více odpovídá realitě).

Určení optimálního počtu shluků I

- **Subjektivní** rozhodování podle:
  1) počtu objektů ve shluku,
  2) vzdálenosti shluků,
  3) na základě charakteru dat.

- Objektivní např. pomocí **Silhouette indexu**, kde *a(i)* je průměrná vzdálenost objektu ke všem ostatním objektům v daném shluku a *b(i)* je nejmenší průměrná vzdálenost objektu *i* k objektům ostatních shluků (odkazuje tedy na vzdálenost k sousednímu shluku).

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$



- Platí: -1 ≤ s(i) ≤ 1.
- s(i) blízké -1 značí špatné zařazení do shluku, blízké 1 správné zařazení do shluku, hodnoty blízké 0 značí, že objekt leží na hranici dvou shluků.
- Počítá se průměr *s(i)* v rámci shluků a do grafu vykreslujeme průměr *s(i)* pro všechny shluky. Počet shluků s nejvyšší hodnotou celkového *s(i)* odkazuje na nejlepší dělení souboru.

Nakonec ale stejně může vyhrát naše subjektivní rozhodnutí ☺

Určení optimálního počtu shluků II

- Objektivní pomocí **Mantelova testu.**

- Hodnotíme korelaci původní asociační matice vzdáleností a asociační matice (vypočítanou pomocí Gowerova indexu), která obsahuje 1, pokud jsou spolu objekty ve shluku a 0 pokud nejsou. R si matici určující současný výskyt ve shluku převede na vzdálenosti – tedy 0 pokud jsou spolu objekty ve shluku a 1 pokud nejsou.
- Kladná korelace (nízká vzdálenost → objekty jsou spolu ve shluku) nám říká, že objekty sobě podobné leží spolu ve shluku.