

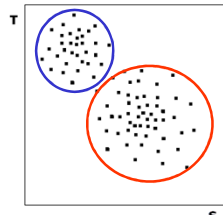

**MASARYKOVA UNIVERZITA**  
**MASARYK JANA BRNO**

Z2069 Statistické metody a zpracování dat II  
**Shluková analýza**

  
 INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

**Ilustrativní případ pro  $m=2$**

Klimatické poměry  $n$  stanic jsou charakterizovány dvěma proměnnými ( $m=2$ ): Průměrnou roční teplotou vzduchu ( $T$ ) a ročním úhrnem srážek ( $S$ )



**INTERPRETACE:** Stanice s vysokými srážkami a nízkými teplotami tvoří shluk stanic vysokohorských, stanice s nízkými úhrny srážek a vysokými teplotami tvoří shluk stanic níže položených. Ve většině případů není vymezení shluků takto triviální.

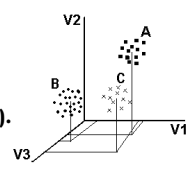
**Shluková analýza (Cluster analysis)**

Je to skupina metod, jejichž cílem je rozdělení souboru jednotek na několik navzájem vylučujících se relativně stejnorodých podmnožin (shluků = clusters).

Rozdělení jednotek je provedeno tak, aby jednotky patřící do téhož shluku si byly co nejvíce „podobné“, zatímco jednotky pocházející z různých shluků by měly být co nejvíce odlišné.

**Charakteristika metody I.**

Shluková analýza je vícerozměrnou metodou. K charakterizování jednotek, kterých je obecně  $n$  využívá většího počtu znaků ( $m \geq 2$ ).



Jednotky představují body v  $m$ -rozměrném prostoru, jehož osy tvoří hodnoty jednotlivých znaků ( $v_1, v_2, v_3$ ).

V takto definovaném prostoru tvoří jednotky s podobnými hodnotami znaků **PŘIROZENÉ** shluky.

Jednotlivé metody shlukové analýzy řeší problém definice a výpočtu „podobnosti“ či „odlišnosti“ jednotek a jejich **PŘÍSLUŠNOST** k určitým shlukům.

**Charakteristika metody II.**

Shluková analýza nesnižuje počet proměnných (jako např. faktorová či komponentní analýza).

Jde jí o shrnutí jednotek do skupin, jejichž počet může nabývat hodnot od 1 do  $n$ , kde  $n$  je počet výchozích jednotek.

Význam má shlukování pro počet shluků výrazně menší než  $n$ . Shlukování může mít vlastnosti hierarchického spojování objektů (skladebnost tříd).

Je-li soubor jednotek postupně pospojován do menšího počtu shluků, jednotky v jednotlivých shlucích jsou si méně „podobné“ a naopak.

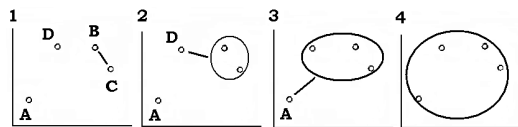
**Charakteristika metody III.**

Analogicky jako v případě faktorové či komponentní analýzy můžeme spojovat (shlukovat) sloupce (proměnné) a nebo řádky (případy) vstupní matice.

Výsledek shlukové analýzy závisí především na těchto parametrech:

- na zvolených znacích a na jejich počtu
- na zvolené míře „podobnosti“ jednotek
- na způsobu shlukování

## Princip shlukování a míry vzdálenosti

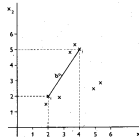


Kritériem víceznakové podobnosti ve shlukové analýze je **VZDÁLENOST**.

Čím blíže se nacházejí body v  $m$ -rozměrném prostoru, tím jsou si podobnější. Nulová vzdálenost znamená identitu – tedy maximální podobnost.

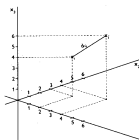
## Eukleidovská vzdálenost

2D prostor



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

3D prostor



$$d_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + (x_{3i} - x_{3j})^2}$$

## Vzdálenost v $m$ - rozměrném prostoru

Pomocí vlastností eukleidovského metrického prostoru lze tento typ vzdálenosti bodů definovat obecně pro  $m$ -rozměrný prostor:

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ki} - x_{kj})^2}$$

Eukleidovská vzdálenost předpokládá ortogonalitu os definovaného prostoru – to znamená vzájemnou nezávislost použitých znaků.

## Předpoklady použití shlukové analýzy

1. Předpoklad vzájemné nezávislosti (nekorelovanosti) proměnných
2. Předpoklad nezávislosti na jednotkách měření
3. Předpoklad stejné významnosti uvažovaných proměnných

ad 1) lze zajistit použitím výsledků faktorové či komponentní analýzy

ad 2) lze zajistit standardizací vstupních dat

ad 3) viz. dále

## Předpoklady použití shlukové analýzy

Předpoklad stejné významnosti uvažovaných proměnných - řeší se přidáním vah do vzorců pro výpočet vzdáleností.

Váhy slouží ke zdůraznění rozdílů mezi znaky důležitými a k potlačení rozdílů mezi znaky málo důležitými.

Vztah pro výpočet vážených vzdáleností:

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ki} - x_{kj})^2 \cdot v_k}$$

kde  $v_k$  je váha příslušející znaku  $k$ .

## Algoritmus výpočtu shlukové analýzy

Pro splnění prvních dvou výše uvedených podmínek vychází vlastní shlukování často ne z původních  $m$  proměnných ale z tzv. komponentních skóre získaných metodou analýzy hlavních komponent.

Algoritmus výpočtu obsahuje dva základní kroky:

1. analýzu vzdáleností
2. vlastní shlukování

## Ilustrativní příklad - vstupní datový soubor

Tzv. komponentní skóre získaná metodou analýzy hlavních komponent.

První čtyři interpretované hlavní komponenty popisující strukturu zaměstnanosti v 10-ti pražských obvodech (viz. Heřmanová, 1991)

	1	2	3	4
	SC I	SC II	SC III	SC IV
P1	7,595	-3,509	0,4	-0,693
P2	3,986	-2,536	-0,717	-0,426
P3	-2,08	-1,203	-2,138	1,894
P4	0,36	2,616	-1,593	-0,878
P5	-5,037	-1,154	2,777	-1,01
P6	10,016	5,088	0,16	0,727
P7	0,386	-1,942	-0,393	0,403
P8	-2,691	1,015	-1,33	0,15
P9	-10,649	0,66	2,703	-0,365
P10	-1,7953	0,964	0,13	0,198

Komponentní skóre poskytují míru vztahu mezi každým pozorováním (případem) a novými komponentami. Jsou důležitá pro interpretaci výsledků v geografii při analýze prostorových struktur. Ukazují do jaké míry je konkrétní pozorování zastoupeno v nových proměnných (komponentách).

## Analýza vzdálenosti

Spočívá ve výpočtu zvoleného typu vzdálenosti mezi všemi jednotkami a v jejich sestavení do symetrické čtvercové matice, která má na diagonále nuly (tj. maximální podobnost).

### Matice vzdáleností

Obvod Prahy	1	2	3	4	5	6	7	8	9	10
1	0	4,01	10,58	9,68	13,06	9,05	7,51	11,39	18,85	10,38
2	4,01	0	6,67	6,31	9,68	9,90	3,66	7,50	15,25	6,69
3	10,58	6,67	0	5,34	6,43	13,88	3,44	3,00	10,27	3,58
4	9,68	6,31	5,34	0	7,90	10,25	4,88	3,60	11,99	3,35
5	13,06	9,68	6,43	7,90	0	16,60	6,47	5,33	5,93	4,88
6	9,05	9,90	13,88	10,25	16,60	0	11,96	13,44	21,32	12,46
7	7,51	3,66	3,44	4,88	6,47	11,96	0	4,36	11,76	3,63
8	11,39	7,50	3,00	3,60	5,33	13,44	4,36	0	8,94	1,75
9	18,85	15,25	10,27	11,99	5,93	21,32	11,76	8,94	0	9,30
10	10,38	6,69	3,58	3,35	4,88	12,46	3,63	1,75	9,30	0
$\Sigma_{i,j}$	94,51	69,68	63,18	63,31	76,29	118,85	57,67	59,32	113,61	56,03

## Vlastní postup shlukování I.

1. Zvolení způsobu shlukování. Z řady variant spojování (viz. dále) je zde použita metoda průměrné vzdálenosti.
2. V matici vzdáleností nalezneme minimální prvek  $\min\{d_{ij}\}$  a matici vzdáleností typu  $[n, n]$  redukuje se na typ  $[n-1, n-1]$ .
3. Jednotky, kterým přísluší minimální hodnota vzdálenosti si jsou nejpodobnější.
4. Sloučíme odpovídající si jednotky do prvního shluku.
5. V dalších krocích se mohou slučovat jednotky se shlukem či dva shluky.

## Vlastní postup shlukování II.

6. Vypočtou se nové vzdálenosti mezi vzniklým shlukem a zbylými jednotkami (či shluky). Ty se vypočtou v našem případě jako aritmetický průměr vzdáleností mezi jednotkami, které patří do nově agregovaného shluku a zbylými jednotkami.
7. Vzdálenosti, kterých se slučování netýká jsou pouze přepsány do nové matice vzdáleností.
8. V nové matici se opět hledá prvek  $\min\{d_{ij}\}$
9. Analyzujeme-li  $n$  jednotek, v procesu shlukování je  $n-1$  kroků. V posledním kroku dochází ke sloučení všech jednotek do jednoho shluku.

### 1. krok - nalezení minimálního prvku v původní matici

Obvod Prahy	1	2	3	4	5	6	7	8	9	10
1	0	4,01	10,58	9,68	13,06	9,05	7,51	11,39	18,85	10,38
2	4,01	0	6,67	6,31	9,68	9,90	3,66	7,50	15,25	6,69
3	10,58	6,67	0	5,34	6,43	13,88	3,44	3,00	10,27	3,58
4	9,68	6,31	5,34	0	7,90	10,25	4,88	3,60	11,99	3,35
5	13,06	9,68	6,43	7,90	0	16,60	6,47	5,33	5,93	4,88
6	9,05	9,90	13,88	10,25	16,60	0	11,96	13,44	21,32	12,46
7	7,51	3,66	3,44	4,88	6,47	11,96	0	4,36	11,76	3,63
8	11,39	7,50	3,00	3,60	5,33	13,44	4,36	0	8,94	1,75
9	18,85	15,25	10,27	11,99	5,93	21,32	11,76	8,94	0	9,30
10	10,38	6,69	3,58	3,35	4,88	12,46	3,63	1,75	9,30	0

Ve výchozí matici je minimální vzdálenost mezi prvky 8 a 10:

$$d_{8,10} = 1,75.$$

Tyto dvě jednotky se sloučí.

Vypočítáme vzdálenosti tohoto nového shluku k stávajícím jednotkám (příklad pro jednotku 1):

$$d_{(8+10,1)} = \frac{d_{(8,1)} + d_{(10,1)}}{2} = \frac{11,39 + 10,38}{2} = 10,88$$

Analogicky se vypočtou nové vzdálenosti mezi novým shlukem a zbylými jednotkami, tedy  $d_{(8+10,2)}, d_{(8+10,3)}, \dots, d_{(8+10,9)}$

Výsledkem je nová matice vzdáleností.

## 2. krok - nová matice vzdáleností

8+10	1	2	3	4	5	6	7	9
(1,75)	10,88	7,10	<b>3,29</b>	3,48	5,11	12,95	4,00	9,12
1	0	4,01	10,58	9,68	13,06	9,05	7,51	18,85
2		0	6,67	6,31	9,68	9,90	3,66	15,25
3			0	5,34	6,43	13,88	3,44	10,27
4				0	7,90	10,25	4,88	11,99
5					0	16,60	6,47	5,93
6						0	11,96	21,32
7							0	11,76
9								0

Hodnota v závorce vyjadřuje vzdálenost, při které dochází ke sloučení a využívá se ke konstrukci tzv. dendrogramu (viz. dále).

Opět se najde minimální hodnota a celý výpočet se opakuje tak, jak je naznačeno v dále uvedených maticích vzdáleností ...

## 3. krok

8+10+3	1	2	4	5	6	7	9
(2,78)	10,78	6,96	4,10	5,55	13,26	3,81	9,50
1	0	4,01	9,68	13,06	9,05	7,51	18,85
2		0	6,31	9,68	9,90	<b>3,66</b>	15,25
4			0	7,90	10,25	4,88	11,99
5				0	16,60	6,47	5,93
6					0	11,96	21,32
7						0	11,76
9							0

## 4. krok

8+10+3	2+7	1	4	5	6	9
(2,78)	5,38	10,78	<b>4,10</b>	5,55	13,26	9,50
2+7	(3,66)	5,76	5,60	8,08	10,93	13,50
1		0	9,68	13,06	9,05	18,85
4			0	7,90	10,25	11,99
5				0	16,60	5,93
6					0	21,32
9						0

## 5. krok

8+10+3+4	2+7	1	5	6	9
(3,44)	<b>5,44</b>	10,51	6,14	12,51	10,12
2+7	(3,66)	5,76	8,08	10,93	13,50
1		0	13,06	9,05	18,85
5			0	16,60	5,93
6				0	21,32
9					0

## 6. krok

8+10+3+4+2+7	1	5	6	9
(4,52)	8,92	6,78	11,98	11,25
1	0	13,06	9,05	18,85
5		0	16,60	<b>5,93</b>
6			0	21,32
9				0

## 7. krok

8+10+3+4+2+7	5+9	1	6
(4,52)	9,02	<b>8,92</b>	11,98
5+9	(5,93)	15,96	18,96
1		0	9,05
6			0

## 8. krok

8+10+3+4+2+7+1	5+9	6
(5,78)	<b>10,01</b>	11,56
5+9	(5,93)	18,96
6		0

## 9. krok

8+10+3+4+2+7+1+5+9	6
(7,43)	13,20
6	0

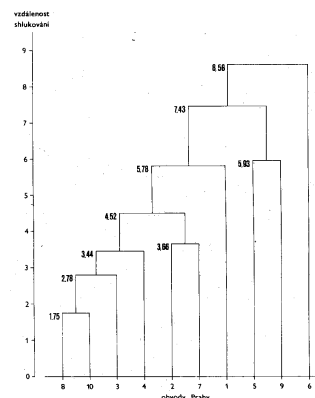
## Ukončení shlukování a prezentace jeho průběhu

V posledním kroku dochází ke sloučení všech jednotek do jednoho shluku na vzdálenosti 8,58.

To je průměrná vzdálenost mezi dvěma jednotkami.

Průběh shlukování se obvykle zaznamenává do tzv. dendrogramů - hierarchicky uspořádaných „stromů“).

## Dendrogram



## Charakteristiky procesu shlukování

- Koefficient ztráty informace
- Graf ztráty informace

Slouží ke stanovení optimálního rozdělení souboru na skupiny.

Při každém kroku shlukování dochází ke ztrátě informace o výchozím souboru tím, že používáme jistá generalizující vyjádření shlukovaných jednotek (např. průměrná vzdálenost v uvedeném příkladě).

Tedy při rozdělení na  $n$  jednotek (na počátku shlukování) je ztráta informace nulová, při spojení do jedné jednotky (na konci) je ztráta maximální (100%).

## Koefficient ztráty informace

krok shlukování	shlukované jednotky	$\sum$ vzdáleností	$\sum$ kumulované vzdáleností	% ztráty informace
1	8,10	1,75	1,75	0,452
2	8,10+3	6,58	8,33	2,156
3	8,10,3+4	12,30	20,63	5,340
4	2,7	3,66	24,29	6,288
5	8,10,3,4+2,7	43,48	67,77	17,547
6	8,10,3,4,2,7+1	53,55	121,32	31,413
7	5,9	5,93	127,25	32,949
8	8,10,3,4,2,7,1+5,9	140,12	267,37	69,228
9	8,10,3,4,2,7,1,5,9+6	118,85	386,22	100,000

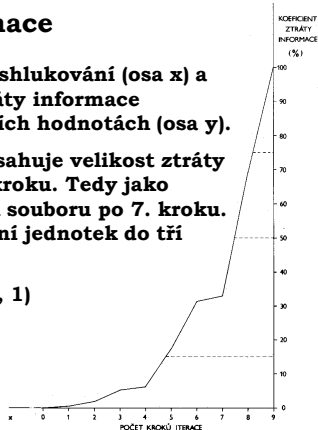
Koefficient ztráty informace se počítá po každém kroku jako součet vzdáleností právě agregovaných jednotek v původní matici  $[n,n]$ . Potom stav shlukování před krokem, kterému přísluší maximální hodnota koeficientu představuje optimální dělení souboru.

## Graf ztráty informace

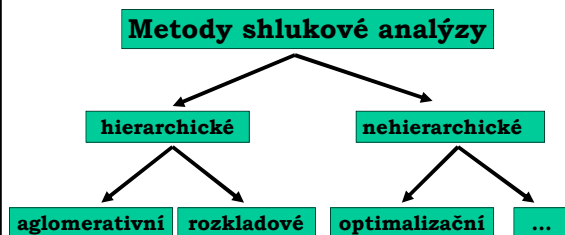
Znázorňuje počet kroků shlukování (osa x) a hodnotu koeficientu ztráty informace v absolutních či relativních hodnotách (osa y).

V uvedeném příkladě dosahuje velikost ztráty informace maxima v 8. kroku. Tedy jako optimální se jeví členění souboru po 7. kroku. Výsledkem je tedy členění jednotek do tří shluků:

- 1: (8, 10, 3, 4, 2, 7, 1)
- 2: (5, 9)
- 3: (6)



## Rozdělení metod shlukové analýzy



## Rozdělení metod shlukové analýzy

Výše uvedený text prezentuje postupy **aglomerativní hierarchické** (jednotky se postupně spojují do shluků)

Hierarchické metody rozkladové naopak postupně dělí vstupní soubor do 2, 3, 4, ... skupin.

**Nehierarchické (optimalizační) metody** hledají takový rozklad množiny objektů, který je optimální podle vhodně zvoleného kritéria.

Mohou být založeny na předem daném (přibližném) počtu shluků a jejich postupném „zlepšováním“ převodem vybraných jednotek mezi shluky, na eventuelním spojování či rozdělování shluků.

Výpočty využívají iteračního počtu.

**Řešené příklady:** Statistica – Elektronický manuál

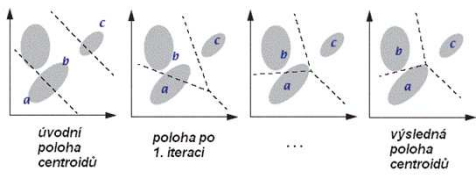
## Shluková analýza metodou k- průměrů (k-means) jako příklad nehierarchického třídění

Algoritmus předpokládá, že dopředu známe počet shluků, do kterého si přejeme rozdělit vstupní soubor.

Výpočet začne s  $k$  náhodnými shluky. Jednotky se poté postupně přesouvají mezi jednotlivými shluky a to tak, aby:

- 1) minimalizovaly variabilitu mezi jednotkami uvnitř jednoho shluku
- 2) Maximalizovaly variabilitu mezi jednotlivými shluky

## Postup shlukování metodou k- průměrů



1. Definování požadovaného počtu výsledných shluků.
2. Určení počáteční polohy středu shluku (centroidu) pro každý shluk (a, b, c).
3. Postupně přiřazení všech jednotek ke shluku, k němuž mají v analyzovaném prostoru nejbliže.
4. Výpočet nové polohy centroidu pro každý shluk na základě přiřazených jednotek.
5. Opakování kroku 3 a 4 do té doby, dokud se poloha shluku či počet jednotek zařazených do shluku výrazně nemění (tzv. stabilní shluky).