

# Razvrščanje v skupine

## Multivariatna analiza

# Uvod

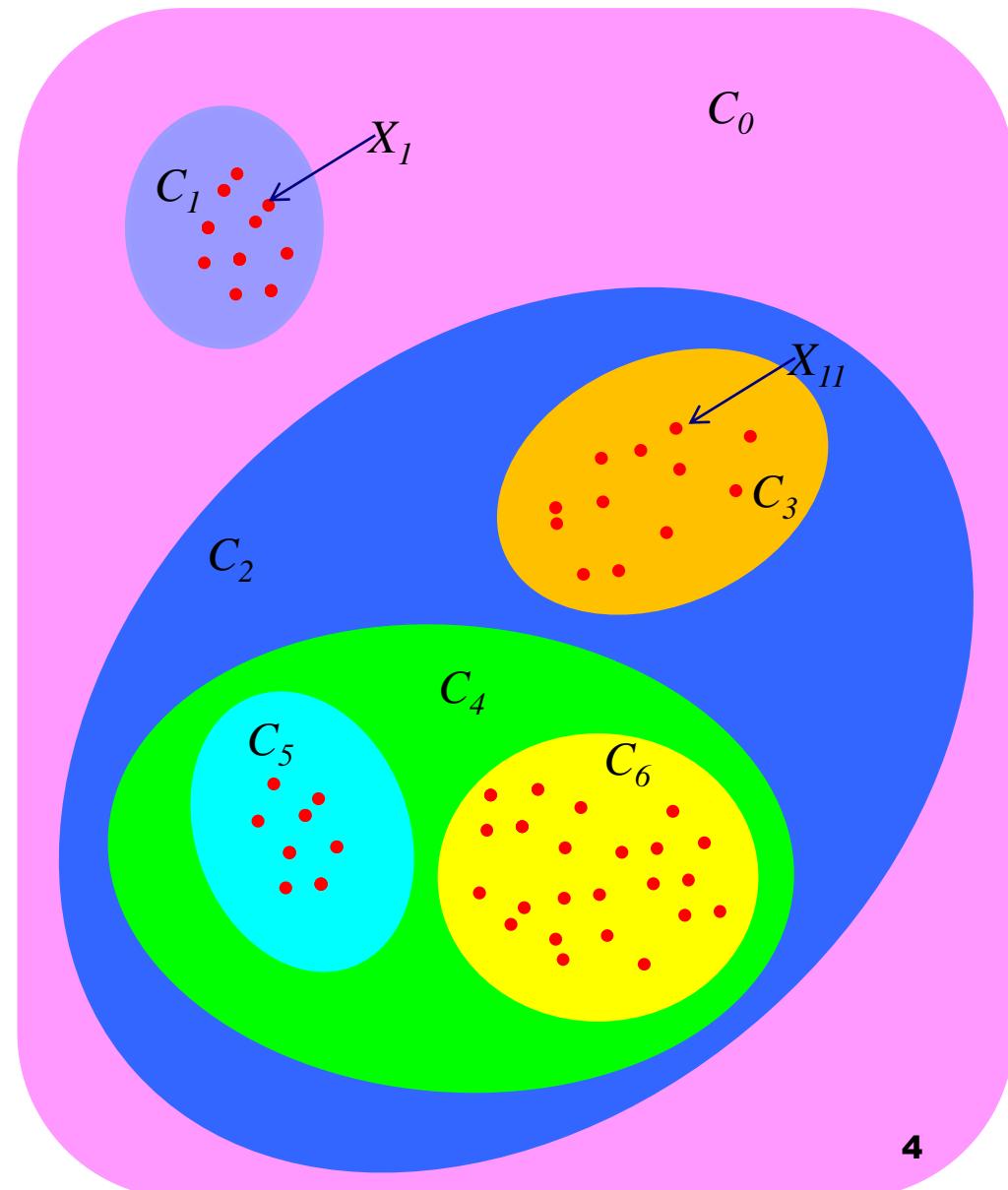
- Cilj razvrščanja v skupine je na podlagi več spremenljivk razvrstiti enot v nekaj skupin tako, da:
  - so si znotraj skupin enote čim bolj podobne
  - so si enote iz različnih skupin čim bolj različne
- Intuitivno preprost problem
- Reševanje še vedno aktualno

# Reševanje

- Formalno je cilj najti tako razvrstitev, ki minimizira vrednost kriterijske funkcije, ki pove, kako dobra je neka rešitev.
- Ker pa je različnih razvrstitev zelo veliko, je problem računsko zelo zahteven.
- Zato večina metod ne zagotavlja, da najdemo globalni optimum.
- Metode poskušajo najti dobro rešitev v sprejemljivem času.

# Osnovni pojmi

- $X$  – enota
- $U$  – končna množica enot
- $C$  – skupina
- $\mathbf{C}$  – razvrstitev,  $\mathbf{C} = \{C_i\}$
- $\Phi$  – množica dopustnih razvrstitev
- $P$  – kriterijska funkcija,  
 $P: \Phi \rightarrow \mathbb{R}_0^+$



# Razvrstitev

Tipi razvrstitev:

- **razbitje**
- **hierarhija**
- piramida
- „fuzzy“ razvrstitev
- razvrstitve s prekrivajočimi skupinami
- ...

# Razvrstitev

- **Razbitje**: Razvrstitev je razbitje, če je vsaka enota v natanko eni skupini.  
 $C = \{C_1, C_2, C_3, \dots, C_k\}$  je razbitje, če  
 $\bigcup_i C_i = U$  in  $i \neq j \Rightarrow C_i \cap C_j = \emptyset$
- **Hierarhija**: Razvrstitev je hierarhija, če za poljubni dve skupini velja, da ali nimata skupnih enot, ali pa je ena vsebovana v drugi.  
 $H = \{C_1, C_2, C_3, \dots, C_k\}$  je hierarhija, če a poljubni par  $C_i$  in  $C_j$  velja:  $C_i \cap C_j = \{C_i, C_j, \emptyset\}$

# Mere podobnosti in različnosti

- Mera *podobnosti*:  $s: (X, Y) \rightarrow R$
- Mera *različnosti*:  $d: (X, Y) \rightarrow R$

Zadoščati mora sledečim pogojem:

- $d(X, Y) \geq 0$                            nenegativnost
- $d(X, X) = 0$
- $d(X, Y) = d(Y, X)$    simetričnosti
- Če veljata tudi spodnja pogoja, je izbrana mera različnosti tudi *razdalja*:
  - $d(X, Y) = 0 \Rightarrow X = Y$
  - $\forall Z: d(X, Y) \leq d(X, Z) + d(Z, Y)$

# Enakovrednost mer različnosti

Meri podobnosti(ali različnosti) sta enakovredni, če je urejenost parov enot, dobljena s prvo mero, enaka urejenosti parov enot z drugo mero.

# Različnosti za številske podatke

$X$  in  $Y$ , merjenimi z  $m$  številskimi spremenljivkami

$$X = (x_1, x_2, x_3, \dots, x_m)$$

$$Y = (y_1, y_2, y_3, \dots, y_m)$$

Najpogosteje uporabljene:

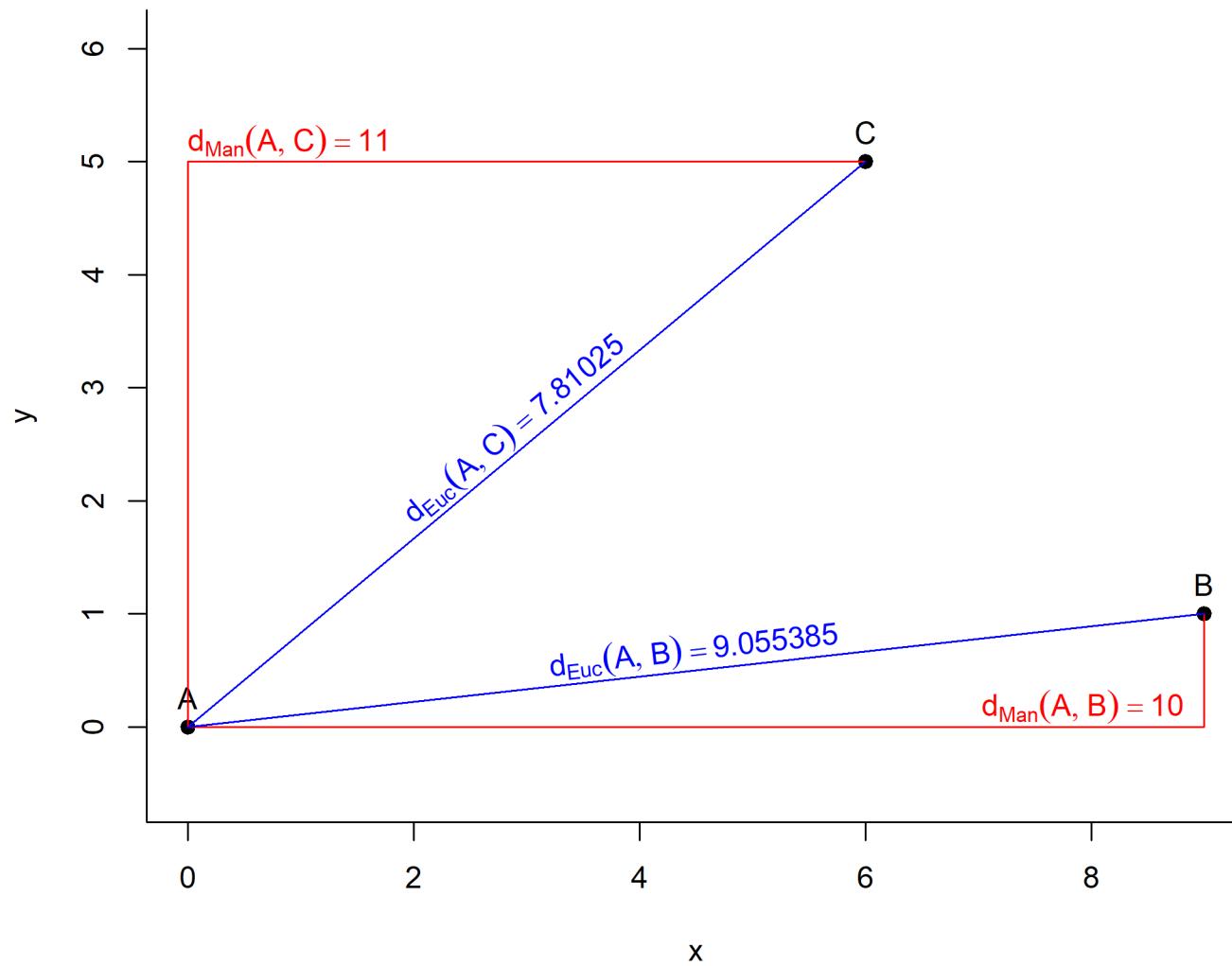
- **Evklidska razdalja:**

$$d_{euc}(X, Y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$$

- **Razdalja Manhattan:**

$$d_{man}(X, Y) = \sum_{i=1}^m |x_i - y_i|$$

# Različnosti za številske podatke



# Različnosti za številske podatke

- **Razdalja Minikowskega:**

$$d(X, Y) = \left( \sum_{i=1}^m |x_i - y_i|^r \right)^{\frac{1}{r}}, r > 0$$

- Če je  $r = 2 \rightarrow$  Evklidska razdalja
- Če je  $r = 1 \rightarrow$  Manhattan razdalja
- Večji kot je  $r$ , večjo težo imajo večje razlike. V limiti, to je pri  $r = \infty$ , dobimo razdaljo Čebiševa oz. trdnjavsko razdaljo (pomembna je le največja razlika):

$$d(X, Y) = \max_i |x_i - y_i|$$

# Različnosti za številske podatke

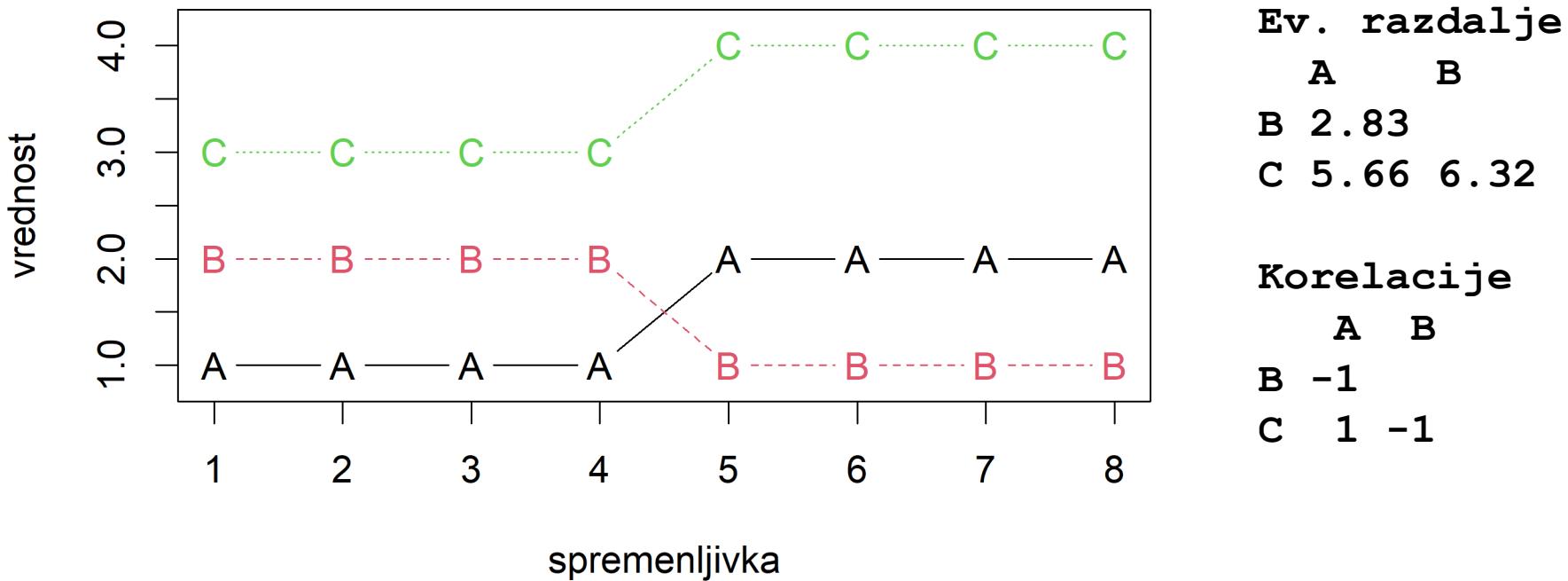
- ***Pearsonov korelacijski koeficient.***

$$r(X, Y) = \frac{\sum_{i=1}^m (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^m (x_i - \mu_x)^2 \sum_{i=1}^m (y_i - \mu_y)^2}}$$

- Z njim ne primerjamo razlike med dejanskimi vrednostnimi spremenljivk pri enotah, ampak med „profili“ enot po spremenljivkah (linearna transformacija vseh vrednosti pri posameznih enotah nima vpliva na rezultat)
- Odločiti se moramo, ali nam je pomembna dejanska ali absolutna vrednost koeficiente

# Različnosti za številske podatke

## ■ Pearsonov korelacijski koeficient



# Standardizacija

- Če so spremenljivke merjene z različnimi merskimi lestvicami, številske spremenljivke pred računanjem različnosti med enotami standardiziramo.
- Najpogostejša standardizacija je:  
$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$
kjer je  $x_{ij}$  vrednost  $j$ -te spremenljivke  $X_j$  za enoto  $i$ ,  $\mu_j$  je aritmetična sredina in  $\sigma_j$  standardni odklon spremenljivke  $X_j$ . Tako standardizirana spremenljivka ima aritmetično sredino 0 in standardni odklon 1.

# Standardizacija

Še nekaj pogostejših standardizacij:

- Deljenje z maksimalno vrednostjo:

$$z_{ij} = \frac{x_{ij}}{\max X_j}$$

- Standardizacija na interval [0,1]:

$$z_{ij} = \frac{x_{ij} - \min X_j}{\max X_j - \min X_j}$$

- Za več glejte Milligan in Cooper (1988)

# Mere podobnosti za binarne podatke

- Mere so določne s frekvencami v kontingenčni tabeli. Kontingenčna tabelo za enoti  $X$  in  $Y$ , kjer so vrednosti vseh  $m$  spremenljivk 1 in 0:
  - $a + b + c + d = m$   
(število vseh spremenljivk)

		Enota $Y$	
		1	0
Enota $X$	1	$a$	$b$
	0	$c$	$d$

# Mere podobnosti in različnosti za binarne podatke

- Sokal-Michenerjeva mera (1958) –  
Matching coefficient:  $\frac{a+d}{a+b+c+d}$
- Russell-Raova mera (1940):  $\frac{a}{a+b+c+d}$
- Jaccardova mera (1908) – ne upošteva spremenljivk, kjer je pri obeh enotah 0:  
$$\frac{a}{a+b+c}$$
- Zgornje mere so definirane na intervalu [0,1]

# Mere podobnosti in različnosti za binarne podatke

- Evklidska razdalja:  $\sqrt{b + c}$
- Kvadrirana evklidska razdalja:  $b + c$
- Za več mer, njihove opise in primerjavo glej Batagelj in Bren (1995): Comparing Resemblance Measures. Journal of Classification 12(1), 73-90.

# Mere podobnosti in različnosti za nominalne spremenljivke

- Najpogosteje se ustvarijo umetne sprem. in nato uporabijo mere za binarne podatke.
- Obstaja tudi več mer, prilagojenih za nominalne sprem., npr. koeficient ujemanja (ali sta dve vrednosti enaki ali ne).
- Bolj kompleksne upoštevajo tudi deleže kategorij (enakost redke kategorije predstavlja večjo podobnost kot enakost pogoste kategorij)

# Mere podobnosti in različnosti za ordinalne spremenljivke

- Običajno se uporabijo ali mere za nominalne ali za intervalne spremenljivke

# Mere podobnosti in različnosti za različne tipe spremenljivk

- V takih primerih se pogosto uporablja uteženo povprečje različnih mer podobnosti (ali različnosti).
- Najbolj zanan je Gowerjev koeficinet:  
Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4), 857–871. doi: [10.2307/2528823](https://doi.org/10.2307/2528823)

# Zveze med merami različnosti in podobnosti

- Praviloma metode razvrščanja v skupine predpostavljajo, mere različnosti.
- Mero podobnosti s je mogoče transformirati v mero različnosti  $d$  in obratno.
- Če je s definirana na  $[0,1]$ :  $d = 1 - s$

# Zveze med merami različnosti in podobnosti

- Če je s definirana na  $[-1, 1]$  (npr. koeficient korelacije), lahko uporabimo:
  - $d = \frac{1-s}{2}$  - primerno, če 1 pomeni največjo podobnost, -1 pa najmanjšo
  - $d = 1 - |s|$  - primerno, če -1 in 1 pomenita največjo podobnost, 0 pa najmanjšo (smer povezanosti ni pomembna)

# Kriterijska funkcija

- Definirana je lahko:
  - Direktno
  - Indirektno (kot funkcija mere različnosti med pari enot)
- Najpogosteje uporabljena pri razbitjih v k skupin je *Wardova kriterijska funkcija*:
$$P(C) = \sum_{C_i \in C} \sum_{X \in C_i} d(X, t_{C_i}),$$
kjer je  $t_{C_i}$  centroid skupine  $C_i$  in  $d$  kvadrat evklidske razdalje.

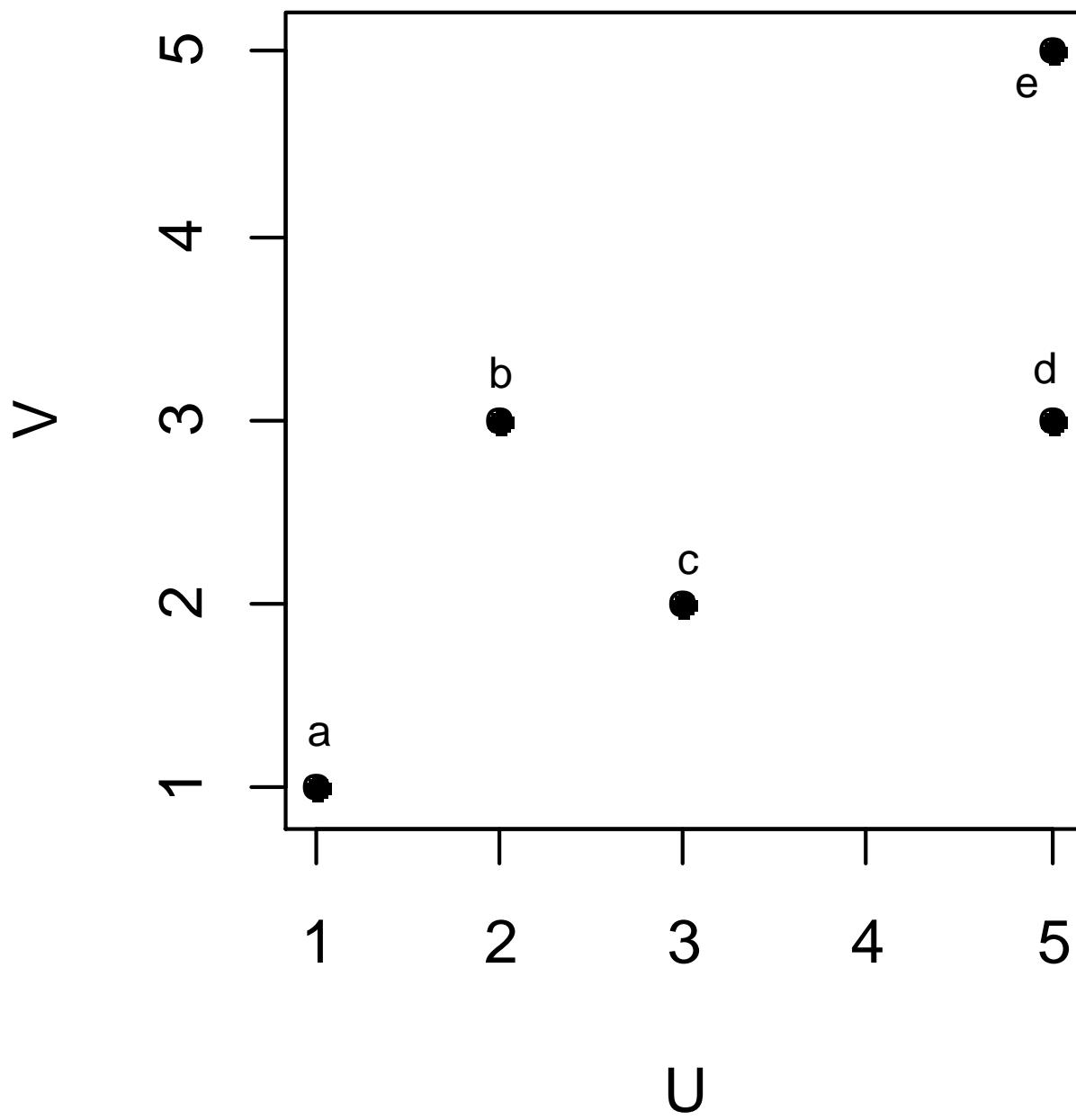
# Problem razvrščanja v skupine

- Problem razvrščanja v skupine lahko definiramo kot iskanje razbitja (razvrstitve), ki minimizira vrednost kriterijske funkcije.
- Teoretično bi lahko optimalno razvrstitev poiskali tako, da bi izračunali vrednosti kriterijske funkcije pri vseh možnih razvrsttvah

# Primer

- Kot primer vzemimo 5 enot ( $a, b, c, d, e$ ), ki jih določata dve spremenljivki ( $U$  in  $V$ ).
- Za te enote bomo izračunali vrednosti Wardove kriterijske funkcije za vsa možna razbitja v dve skupini

	$U$	$V$
$a$	1	1
$b$	2	3
$c$	3	2
$d$	5	3
$e$	5	5



<b>C</b>	<b><i>C</i><sub>1</sub></b>	<b><i>C</i><sub>2</sub></b>	<b><i>t</i><sub>1</sub></b>	<b><i>t</i><sub>2</sub></b>	<b>P(C)</b>
<b>1</b>	a	bcde	(1,0; 1,0)	(3,75; 3,25)	11,50
<b>2</b>	b	acde	(2,0; 3,0)	(3,50; 2,75)	19,75
<b>3</b>	c	abde	(3,0; 2,0)	(3,25; 3,00)	20,75
<b>4</b>	d	abce	(5,0; 3,0)	(2,75; 2,75)	17,50
<b>5</b>	e	abcd	(5,0; 5,0)	(2,75; 2,25)	11,50
<b>6</b>	ab	cde	(1,5; 2,0)	(4,33; 3,33)	9,83
<b>7</b>	ac	bde	(2,0; 1,5)	(4,00; 3,67)	11,17
<b>8</b>	ad	bce	(3,0; 2,0)	(3,33; 3,33)	19,33
<b>9</b>	ae	bcd	(3,0; 3,0)	(3,33; 2,67)	21,33
<b>10</b>	bc	ade	(2,5; 2,5)	(3,67; 3,00)	19,67
<b>11</b>	bd	ace	(3,5; 3,0)	(3,00; 2,67)	21,17
<b>12</b>	be	acd	(3,5; 4,0)	(3,00; 2,00)	16,50
<b>13</b>	cd	abe	(4,0; 2,5)	(2,67; 3,00)	19,17
<b>14</b>	ce	abd	(4,0; 3,5)	(2,67; 2,33)	17,83
<b>15</b>	de	abc	(5,0; 4,0)	(2,00; 2,00)	6,00

# Primer

- Kriterijska funkcija ima najmanjšo vrednost pri zadnjem razbitju  $\rightarrow P(C_{15}) = 6,00$
- Pri našem primeru smo dobili 15 različnih razbitji.
- V splošnem lahko  $n$  enot razvrstimo v dve skupini na  $2^{n-1} - 1$  načinov

# Število možnih razvrstitev

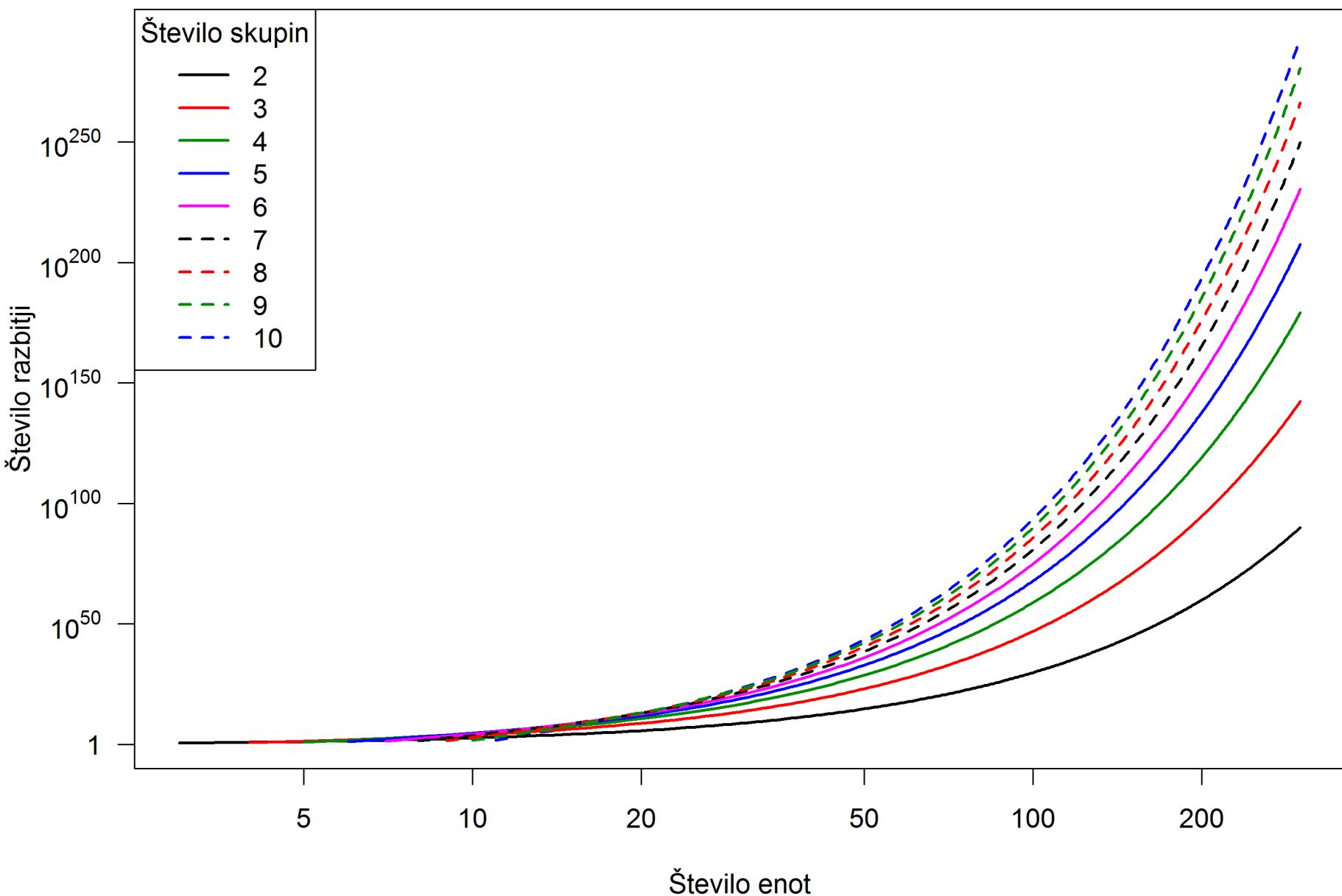
- $n$  enot lahko v  $k$  skupin razvrstimo na:

$$S(n, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^{k-i} \binom{k}{i} i^n$$

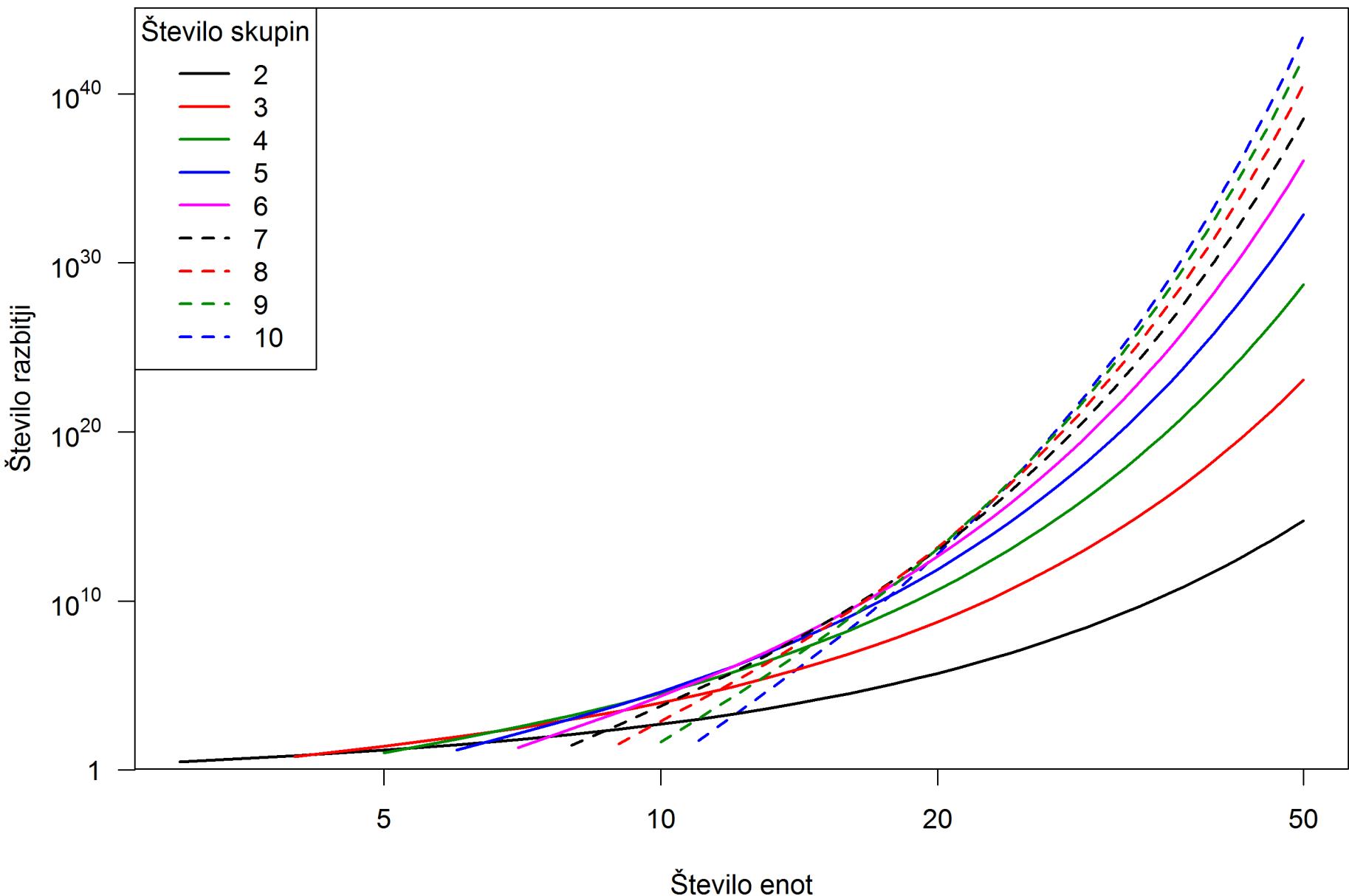
načinov. Številu pravimo Stirlingovo število druge vrste.

- 30 enot lahko v 3 skupine razvrstimo na  $3,4 \cdot 10^{13}$  načinov, 100 enot v 5 skupin pa kar na  $6,6 \cdot 10^{67}$  načinov.

## Število razbitji



## Število razbitji



# Reševanje problema

- Zaradi izjemno velikega števila možnih razbitij problema ne moremo učinkovito izračunati optimalnih rešitev
- Zato obstajajo približne (hevristične) metode, ki v relativno hitrem času dajejo dovolj dobre rešitve.

# Metode

Najpogosteje uporabljane metode so:

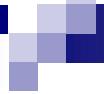
- Hierarhično razvrščanje
- Metoda voditeljev
- Lokalna optimizacija
- Razvrščanje na podlagi modelov

# Hierarhično razvrščanje

- Najpogosteje se enote združujejo
- Ta različica predpostavlja, da so vse pomembne informacije zapisane v matriki različnosti (ki jo je potrebno predhodno izračunati)
- Obstajajo pa tudi metode, ki začnejo z vsemi enotami v eni skupini in nato skupine “delijo/razbijajo” (ne bomo obravnavali)

# Hierarhično razvrščanje - postopek

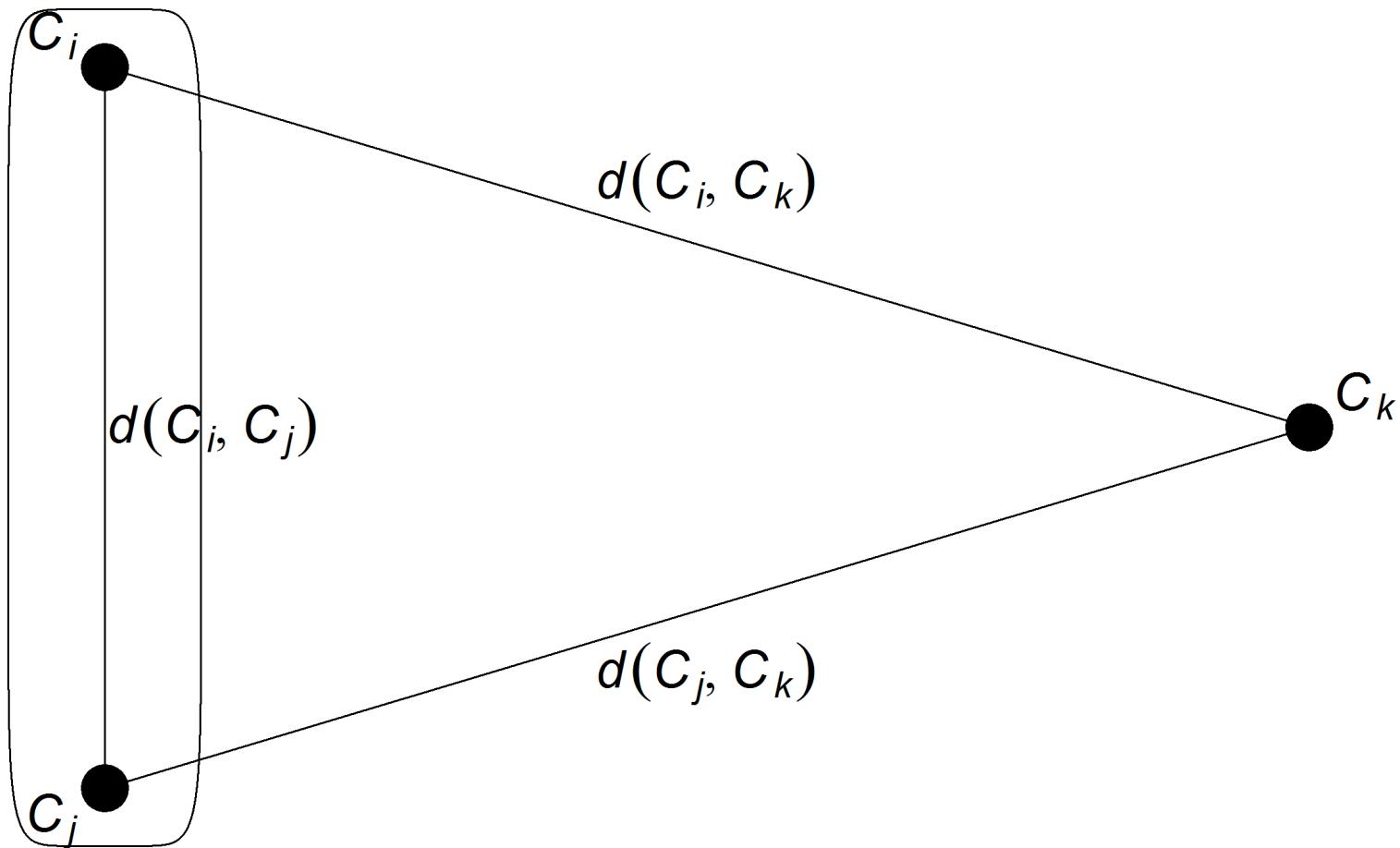
- Začetek: vsaka enota je v svoji skupini (združevalna metoda)
- V vsakem koraku se:
  - združita skupini, ki sta si najbližji
  - izračunajo različnosti te nove (združene) skupine do vseh ostalih skupin
- Postopek se konča, ko so vse enote v eni skupini



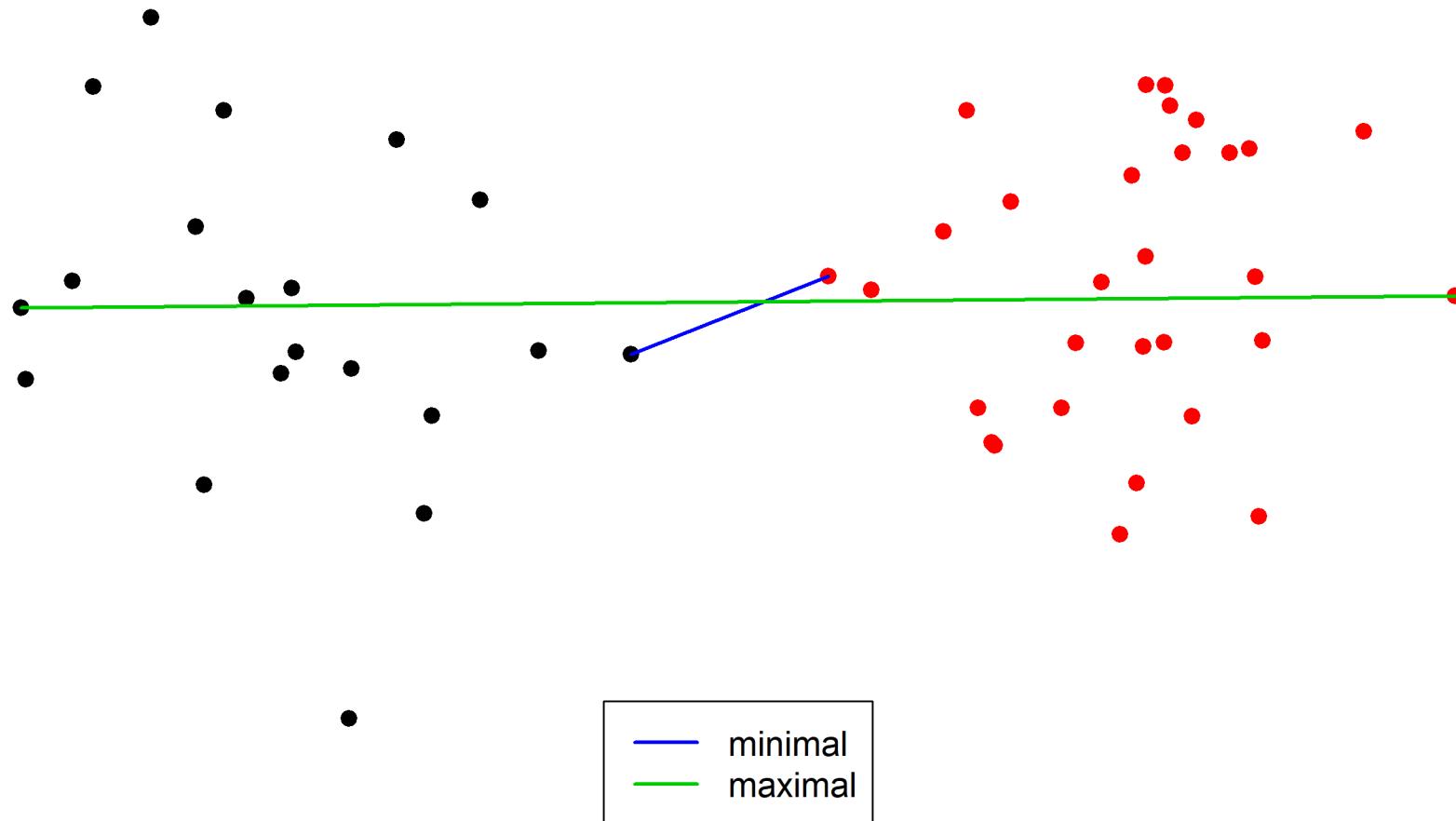
# Hierarhično razvrščanje - postopek

- Imamo matriko različnosti  $D = [d_{ij}]$ , kjer so zapisane mere različnosti med  $n$  enotami iz množice  $\mathbf{U}$ .
- Začetek: Vsaka enota je skupina:  
 $C_i = \{X_i\}, X_i \in \mathbf{U}, i = 1, 2, \dots, n$
- Ponavljam, dokler ne ostane samo ena skupina:
  - določi najbližji si skupini  $C_p$  in  $C_q$ :  
$$d(C_p, C_q) = \min_{u,v} d(C_p, C_q)$$
  - Združi skupini  $C_p$  in  $C_q$  v skupino si skupini  $C_r = C_p \cup C_q$
  - Zamenjaj skupini  $C_p$  in  $C_q$  s skupino  $C_r$
  - Določi mere različnosti  $d$  med novo skupino  $C_r$  in ostalimi skupinami.

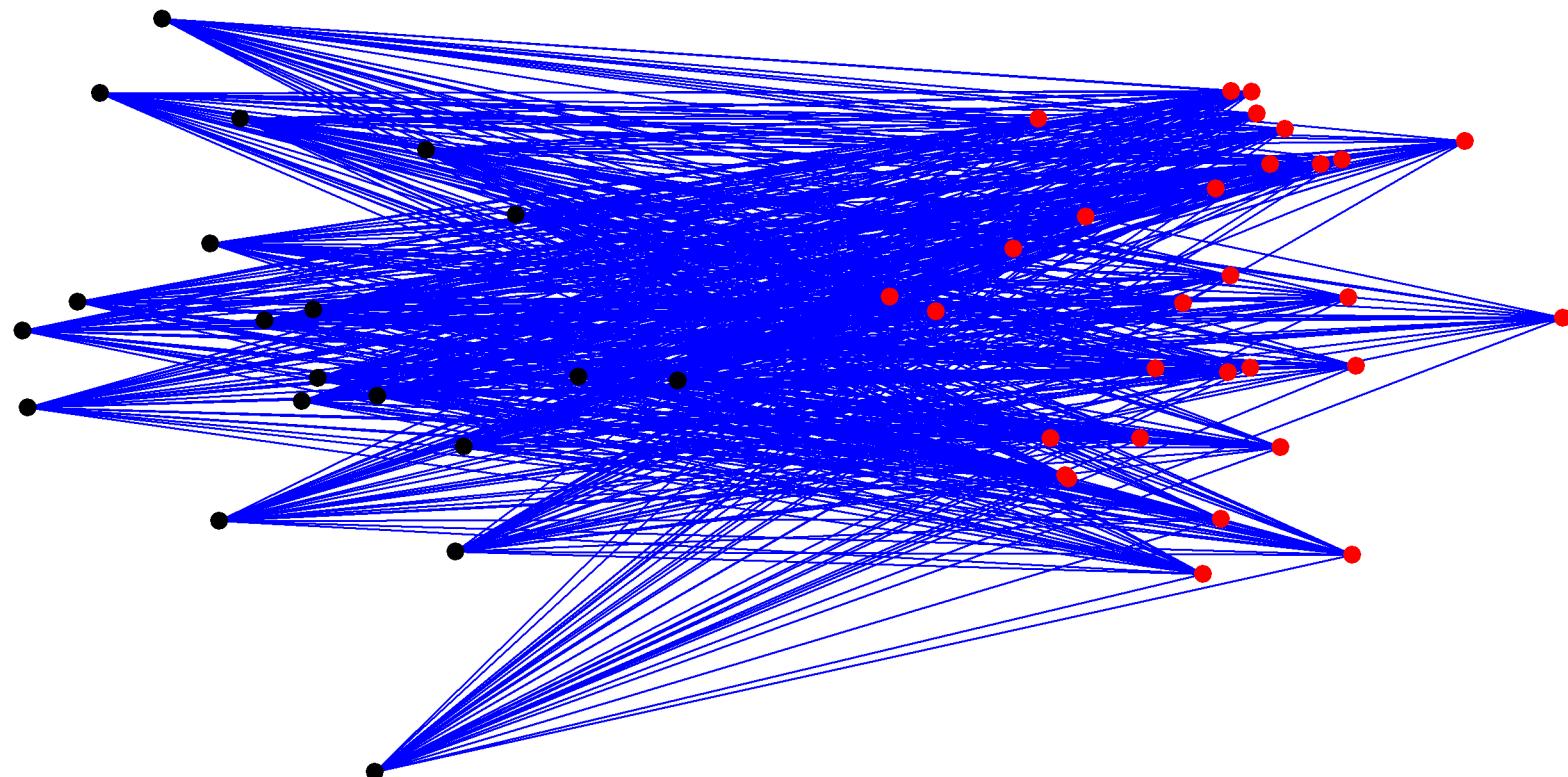
# Združevanje skupin – točke so skupine



# Združevanje skupin – točke so enote



# Združevanje skupin – točke so enote – povprečna metoda



# Metode združevanja skupin

- Minimalna metoda ali enojna povezanost (Florek et al., 1951; Sneath, 1957):

$$d(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

- Maksimalna metoda ali polna povezanost (McQuitty, 1960):

$$d(C_i \cup C_j, C_k) = \max(d(C_i, C_k), d(C_j, C_k))$$

- McQuittijeva metoda (McQuitty, 1966; 1967):

$$d(C_i \cup C_j, C_k) = \frac{d(C_i, C_k), d(C_j, C_k)}{2}$$

- Wardova metoda (Ward, 1963):

$$\begin{aligned} & d(C_i \cup C_j, C_k) \\ &= \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j) = \\ &= \frac{(n_i + n_j)n_k}{(n_i + n_j + n_k)} d(t_{ij}, t_k) \end{aligned}$$

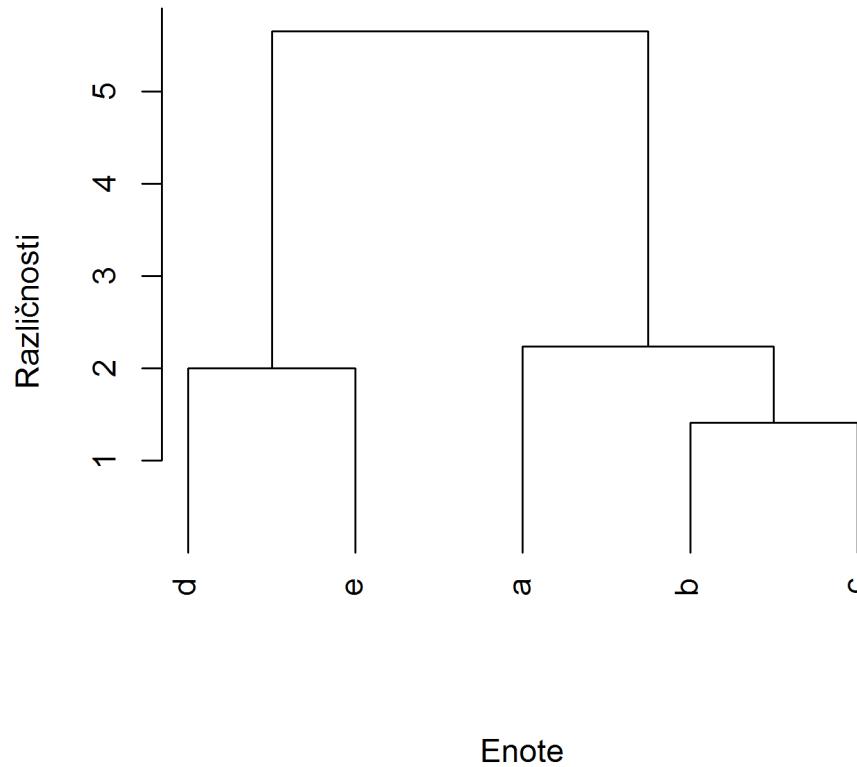
kjer s  $t_{ij}$  označimo težišče (center) združene skupine  $C_i \cup C_j$  in  $t_k$  težišče skupine  $C_k$ .  $n_i$  označuje število enot v skupini  $C_i$ .

# Drevo združevanja

- Z drevesom združevanja ali dendrogramom grafično prestavimo potek združevanja
- Listi so enote, točke združitve pa skupine
- Višina točke, ki jo imenujemo **nivo združevanja**, je sorazmerna meri različnosti med skupinama

# Drevo združevanja

**Primer drevesa**



# Lance in Williamsov obrazec

- Večino metod je mogoče predstaviti s spodnjim obraczem (Lance in Williams, 1967):

$$d(C_i \cup C_j, C_k) = \alpha_1 d(C_i, C_k) + \alpha_2 d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|$$

# Lance in Williamsov obrazec

Metoda	$\alpha_1$	$\alpha_2$	$\beta$	$\gamma$
Minimum	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Maksimum	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Mediana	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	0
Centroidna	$\frac{n_i}{n_k}$	$\frac{n_j}{n_k}$	$-\frac{n_i n_j}{n_k^2}$	0
McQuity	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Povprečna	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Ward	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{n_j + n_k}{n_i + n_j + n_k}$	$\frac{-n_k}{n_i + n_j + n_k}$	0

# Monotonost

- Zaželena lastnost metode/drevesa
- Monotonost velja, če skupini  $C_i$  in  $C_j$  združimo pri različnosti, ki je večja ali enaka različnosti, pri katerih smo združevali elemente posameznih skupin (torej elemente  $C_i$  in elementne  $C_j$ )

# Monotonost

- Definicija:  $X \in \mathbf{U} \Rightarrow h(X) = 0$   
 $C_r = C_i \cup C_j \Rightarrow h(C_r) = d(C_i, C_j)$
- Metoda je monotona, če za vsako skupino  $C_r = C_i \cup C_j$  velja:  $h(C_r) \geq \max(h(C_i), h(C_j))$
- Metoda na podlagi Lance in Williamsovega je monotona, če velja (Batagelj, 1981):

$$\gamma \geq -\min(\alpha_1, \alpha_2)$$

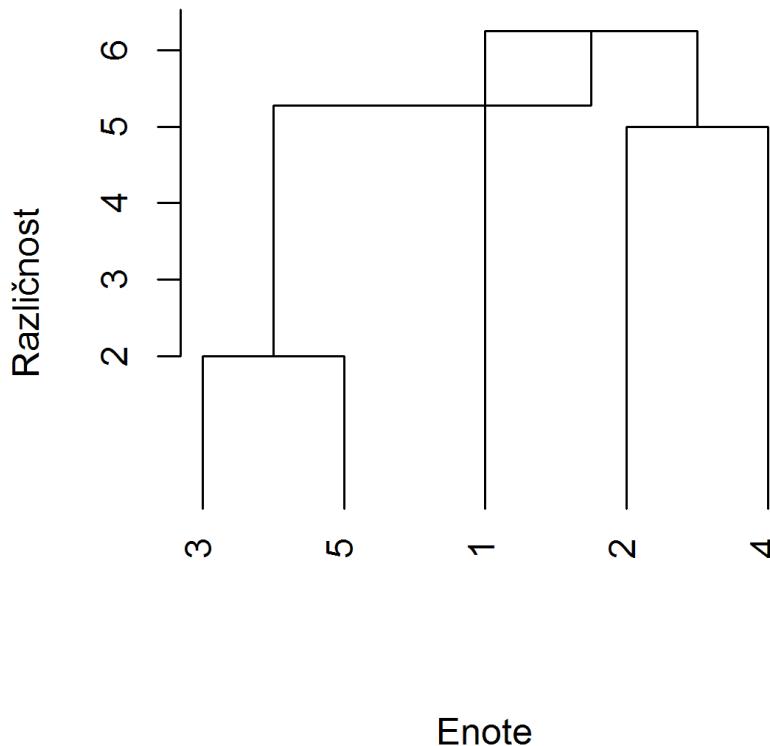
$$\alpha_1 + \alpha_2 \geq 0$$

$$\alpha_1 + \alpha_2 + \beta \geq 1$$

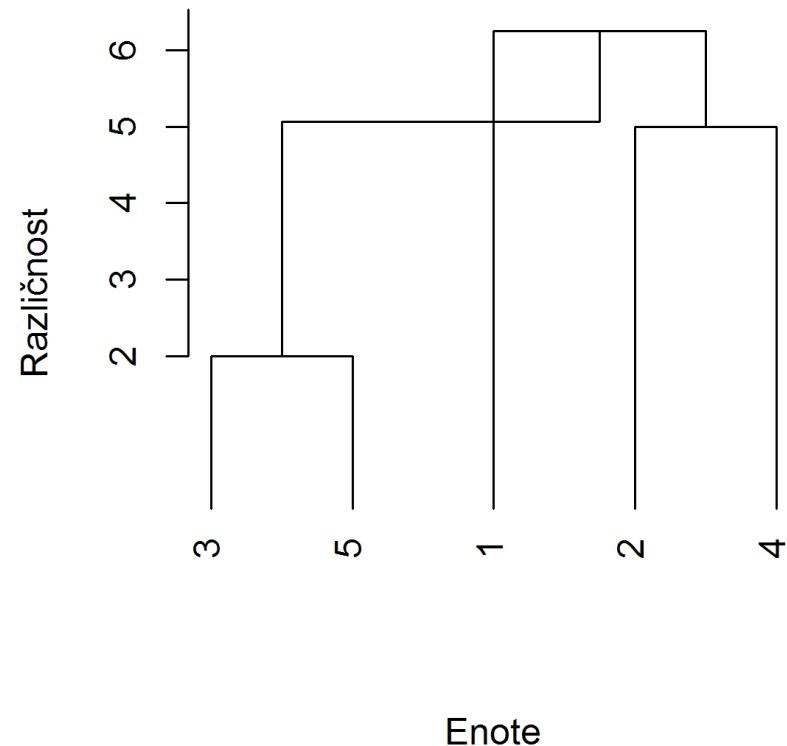
# *Nemonotona drevesa*

Preimer dveh nomonotonih dreves

centroid linkage



median linkage



# Hierarhično razvrščanje

- Te metode so “požrešne” – ko se dve skupini enkrat združita, se jih (oz. enot v njih) ne da več razdružiti
- Zato rešitve načeloma niso optimalne
- Metode/različice se razlikujejo glede na:
  - Izbor mere različnosti
  - Metodo računanja različnosti po združitvi skupin

# Hierarhično razvrščanje

Prednosti:

- Relativno preprost
- Rezultat združevanja je mogoče nazorno prikazati z drevesom združevanja
- Uporabniku ni potrebno v naprej določiti števila skupin (mogoče jih je oceniti tudi iz drevesa združevanja)

# Lastnosti metod hierarhičnega združevanja v skupine

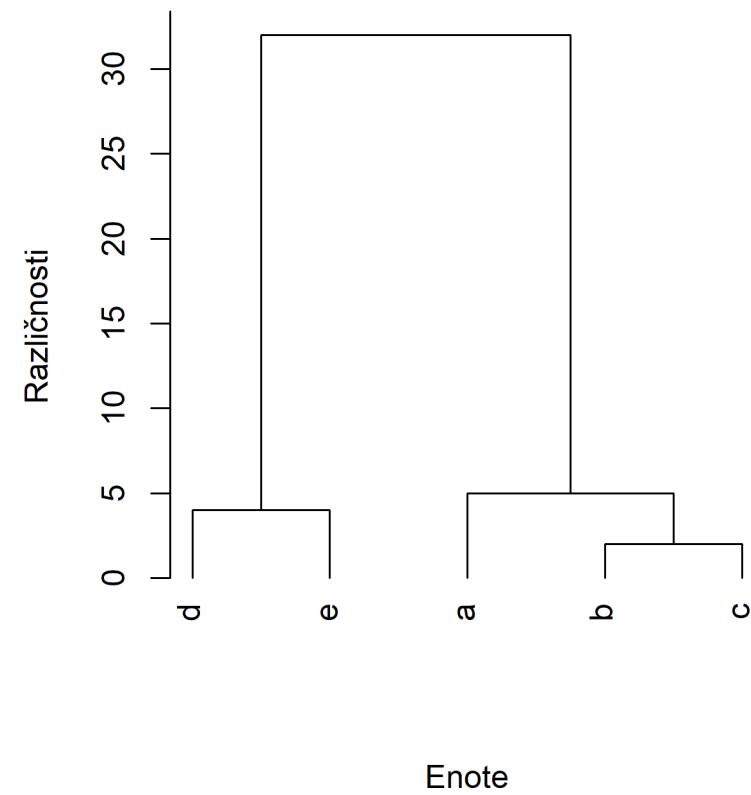
- **Minimalna metoda** je zelo učinkovita pri razkrivanju dolgih „klobasastih“, neeliptičnih skupin, ki so izrazito ločene med seboj. V primeru prekrivajočih se skupin se kaže ‚verižni‘ učinek metode, ko v vsakem koraku združevanja skupini dodaja le posamezno enoto.
- **Maksimalna metoda** dobro razkriva okrogle skupine.
- **Wardova metoda** pa je najprimernejša za eliptično strukturirane podatke

# Primer

## ■ Matrika različnosti:

	a	b	c	d	e
a	0	5	5	20	32
b	5	0	2	9	13
c	5	2	0	5	13
d	20	9	5	0	4
e	32	13	13	4	0

Dendrogram - maksimalna metoda

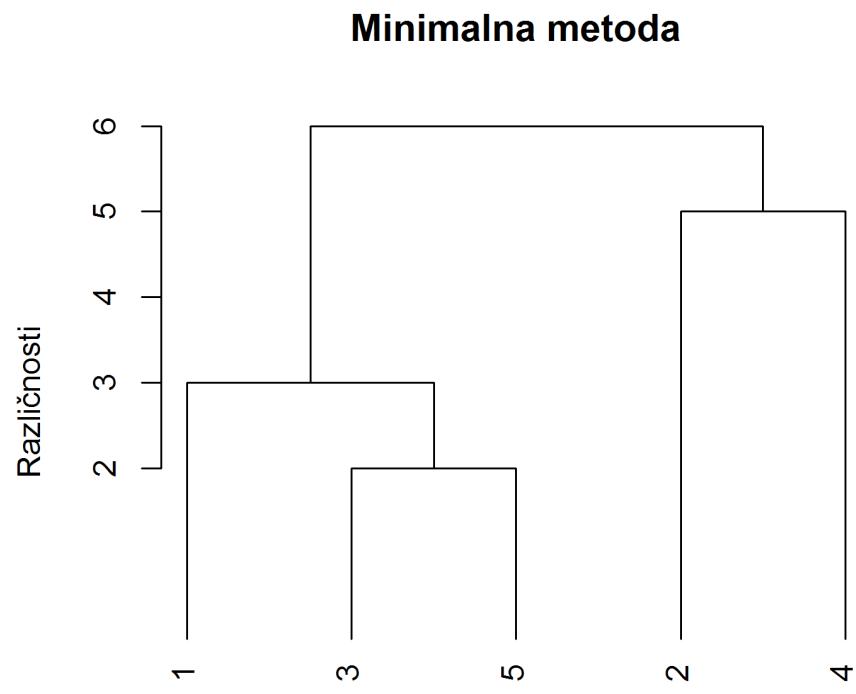


Enote

# Primer 2

## ■ Matrika različnosti:

	1	2	3	4	5
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0



# Metoda voditeljev

- Voditelji so “predstavniki” skupin
- Išče take skupine, da so enote v skupini čim bolj podobne voditelju skupine
- Vsaka enota pripada skupini, katere voditelju je najbolj podobna/blizu

***k-means – posebna različica, kjer:***

- Razdalja je evklidska razdalja
- Voditelji so povprečja skupin

# K-means - postopek

- Spremenljivke morajo biti vsaj intervalne
- Podani so začetni voditelji
- Se ponavlja:
  - vsaka enota se priredi “voditelju” (oz. njegovi skupini), kateremu je najbližja (evklidska razdalja)
  - izračunajo se novi voditelji kot povprečja vseh enot v skupini
  - Oba koraka se ponavljata, dokler novi voditelji niso enaki starim (se voditelji ne ustalijo)

# Metoda voditeljev - značilnosti

- Učinkovita tudi pri velikem številu enot (več 10000, bistveno več kot hierarhične)
- Postopek je **lokalno optimalni postopek**.  
→ Različne začetne množice voditeljev lahko skonvergirajo v lokalno (in ne globalno) optimalne razvrstitve.
- Zato je potrebno postopek **večkrat ponoviti** z različnimi začetnimi voditelji, da bi dobili čim boljšo razvrstitev. (v SPSS-u malce težje - makro)

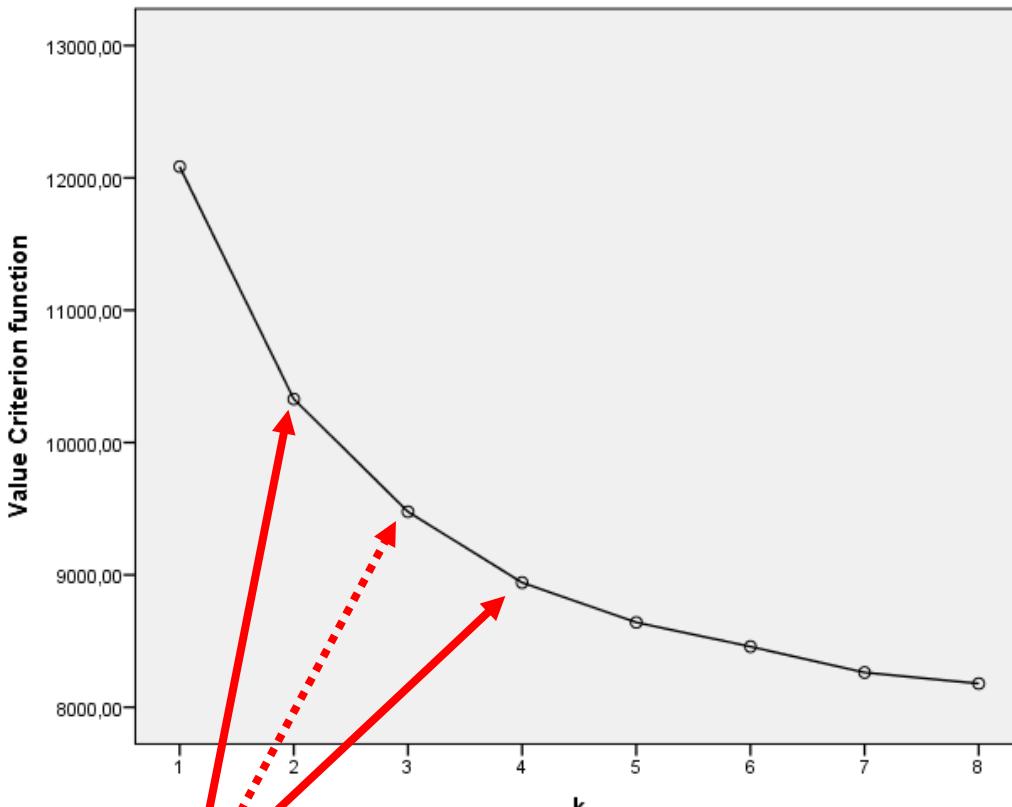
# Določanje števila skupin

Število skupin je potrebno določiti v naprej.

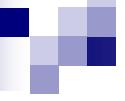
Možni kriteriji:

- Na podlagi dendrograma iz hierarhičnega razvrščanja (še posebej po Wardovi metodi)
- Na podlagi „scree“ diagrama kriterijske funkcije (pri k-means Wardove KF)
- gap statistika
- Silhueta
- drugo (ne bomo obravnavali)

# Scree diagram za število skupin



- Na x os nanašamo število skupin, na y os pa vrednost kriterijske funkcije.
- Število skupin določimo pri kolenu (kjer se krivulja prelomi)
- Pogosto se ne da jasno določiti (kot tu na sliki)

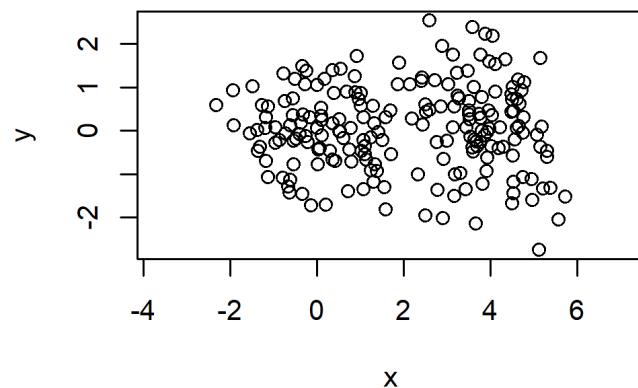


# Gap statistika

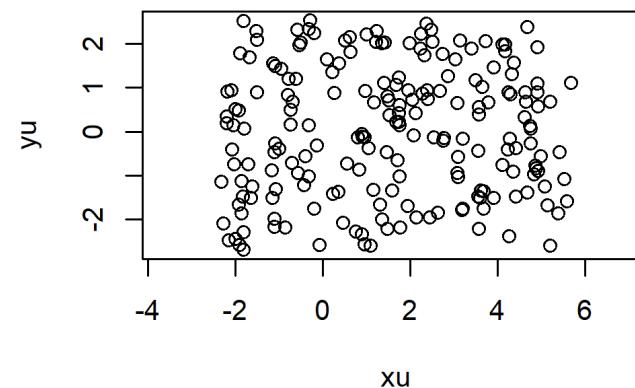
- Iščemo skupine, ki so bolj homogene, kot bi jih našli v podatkih brez skupin.
- Primerja se razdalje znotraj skupin v primerjavi s tistimi, ki bi jih pričakovali glede na porazdelitev iz ničelne hipoteze („referenčni podatki“).
- Ničelna hipoteza predpostavlja, da v podatkih ni skupin. Tako se kot primerna porazdelitev običajno uporabi neko verzijo enakomerne porazdelitve.
- Izberemo tak k (število skupin), kjer je razlika med opaženimi in „referenčnimi“ podatki največja.

# Gap statistika

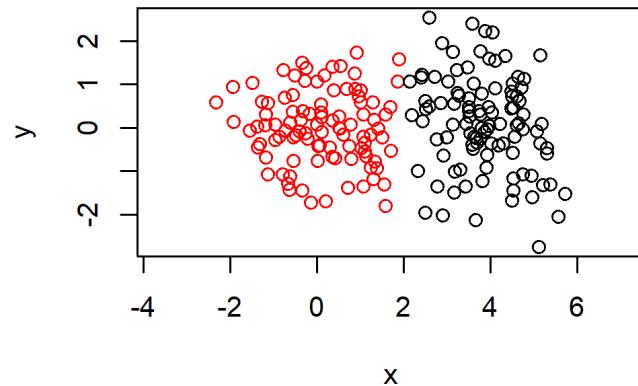
Podatki



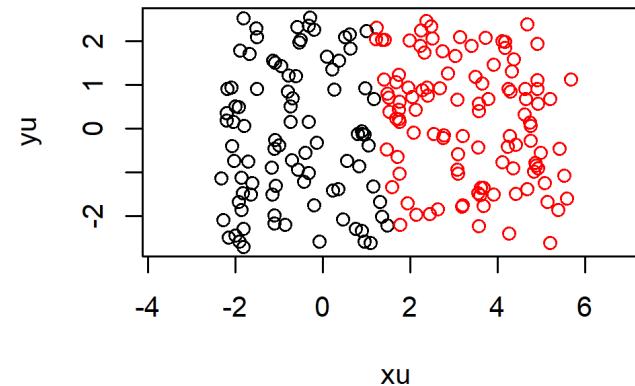
Referenčni podatki



Podatki + skupine



Referenčni podatki + skupine



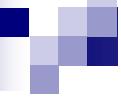
# Gap statistika

- Izračuna se kot:

$$GAP_n(k) = E_n^*\{\log(W_k)\} - \log(W_k) \quad (1)$$

Kjer je:

- $n$  število enot v vzorcu
- $k$  število skupin
- $W_k = \sum_{r=1}^k \frac{D_r}{2n_r}$  kjer je  $D_r = \sum_{i,i' \in C_r} d_{i,i'}$   
( $d_{i,i'}$  je lahko npr. kvadrirana Evklidska razdalja – potem je  $W_k$  vsota kvadratov znotraj skupine)
- $E_n^*$  je pričakovana vrednost  $\log(W_k)$  izračunana na referenčnem podatkovju velikosti  $n$



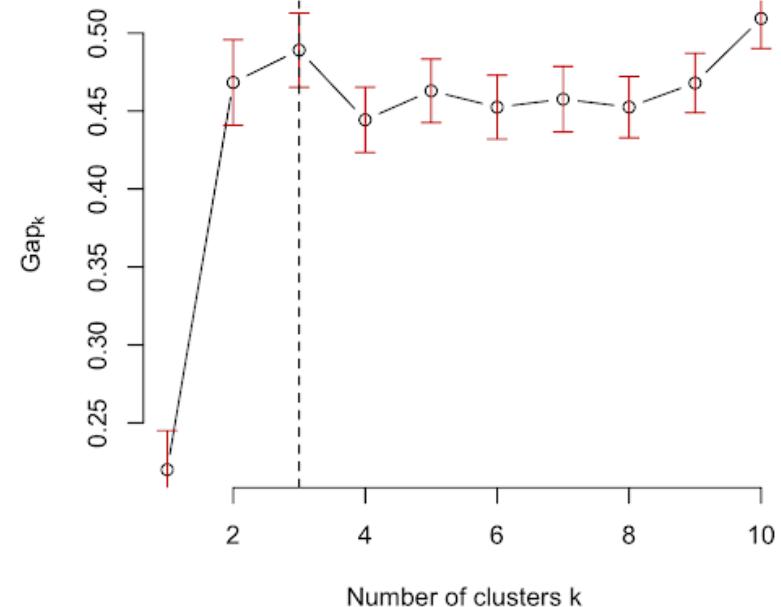
# Gap statistika

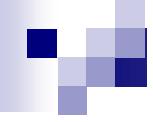
POSTOPEK:

1. Razvrsti enote v  $1, \dots, K$  skupin in izračunaj  $W_k$ .
2. Generiraj  $B$  referečnih podatkovij in na vsakem naredi enako (točko 1).
3. Izračunaj pričakovano GAP statistiko po formuli (1).
4. Izračunaj  $SE(Gap(k))$ , nato izberi najmanjše število skupin, tako da velja:

$$\begin{aligned}\hat{k} &= \text{najmanjši } k \text{ tako da } Gap(k) \\ &\geq [Gap(k+1) - SE(Gap(k+1))]\end{aligned}$$

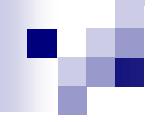
Včasih SE tudi ignoriramo in iščemo največji  $Gap(k)$ .





# Silhueta (Silhouette)

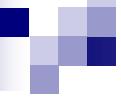
- Silhuite naj bi ocenjevale, kako enote „pašejo“ v skupino, v katero so razvrščene v primerjavi z najbližjo skupino.
- Za vsako enoto  $i$  izračunamo število  $s(i)$ , ki meri, koliko bližje je enota drugim enotam iz svoje skupine kot enotam iz druge najbližje skupine.
- Za izračun potrebujemo razbitje in vse parne razdalje med enotami.
- Izberemo tisto število skupin, kjer je povprečna  $s(i)$  največje.



# Silhueta (Silhouette) - Izračun

- Označimo z  $i$  neko enoto in z  $A$  skupino, v kateri je  $i$ .
- $a(i)$  = povprečna razdalja od  $i$  do vseh ostalih enot iz  $A$ .
- $d(i, C)$  = povprečna razdalja od  $i$  do vseh enot iz  $C$ .
- $b(i) = \min_{C \neq A} d(i, C)$
- $s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$ , če je  $i$  edina enota v  $A$ , je  $s(i) = 0$

Povprečna vrednost  $\bar{s} = \frac{1}{n} \sum_{i=1}^n s(i)$  se uporablja kot mera primernosti razbitja in tudi za oceno števila skupin (kjer je največja).

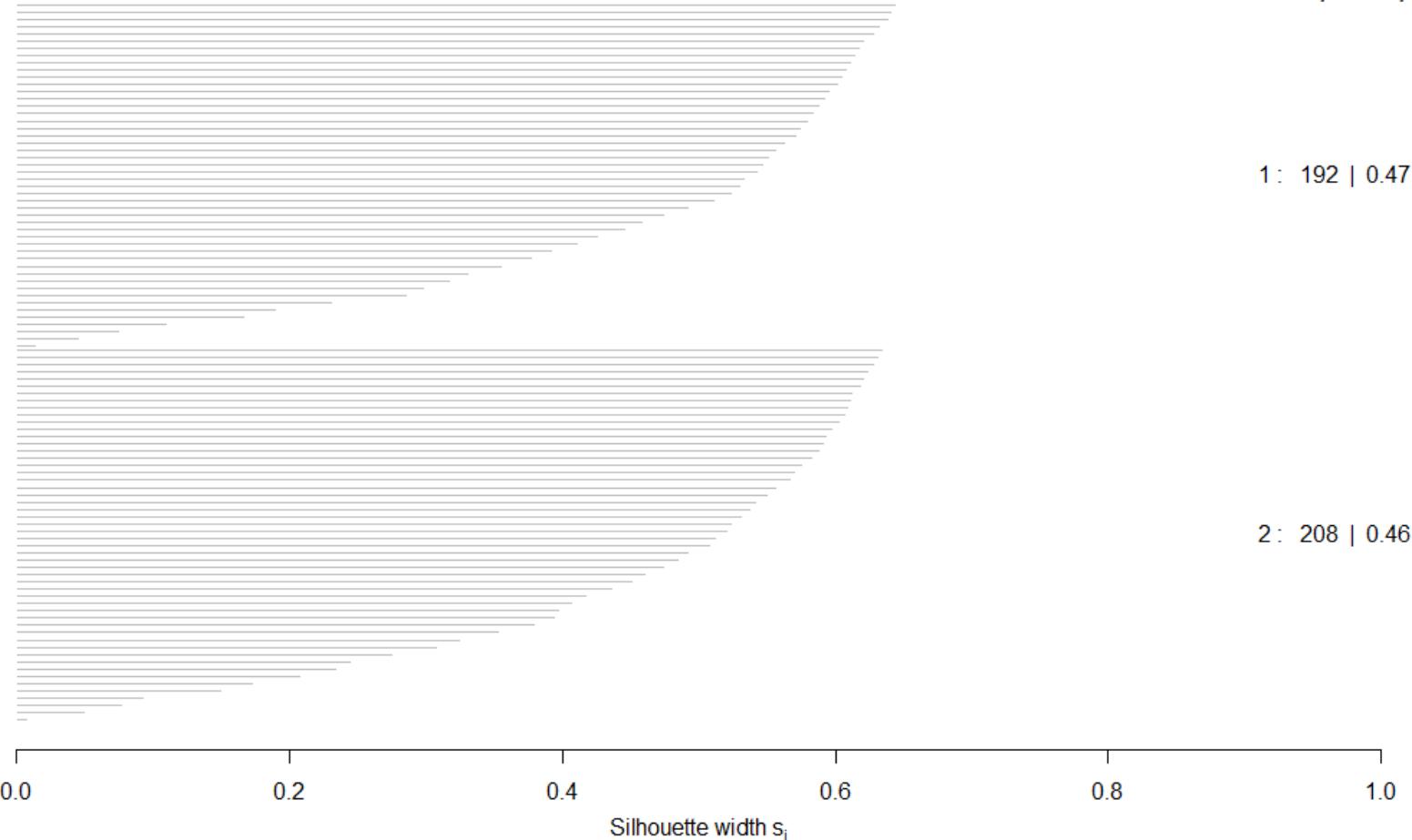


# Silhueta (Silhouette)

Silhouette plot of (x = tclu, dist = d)

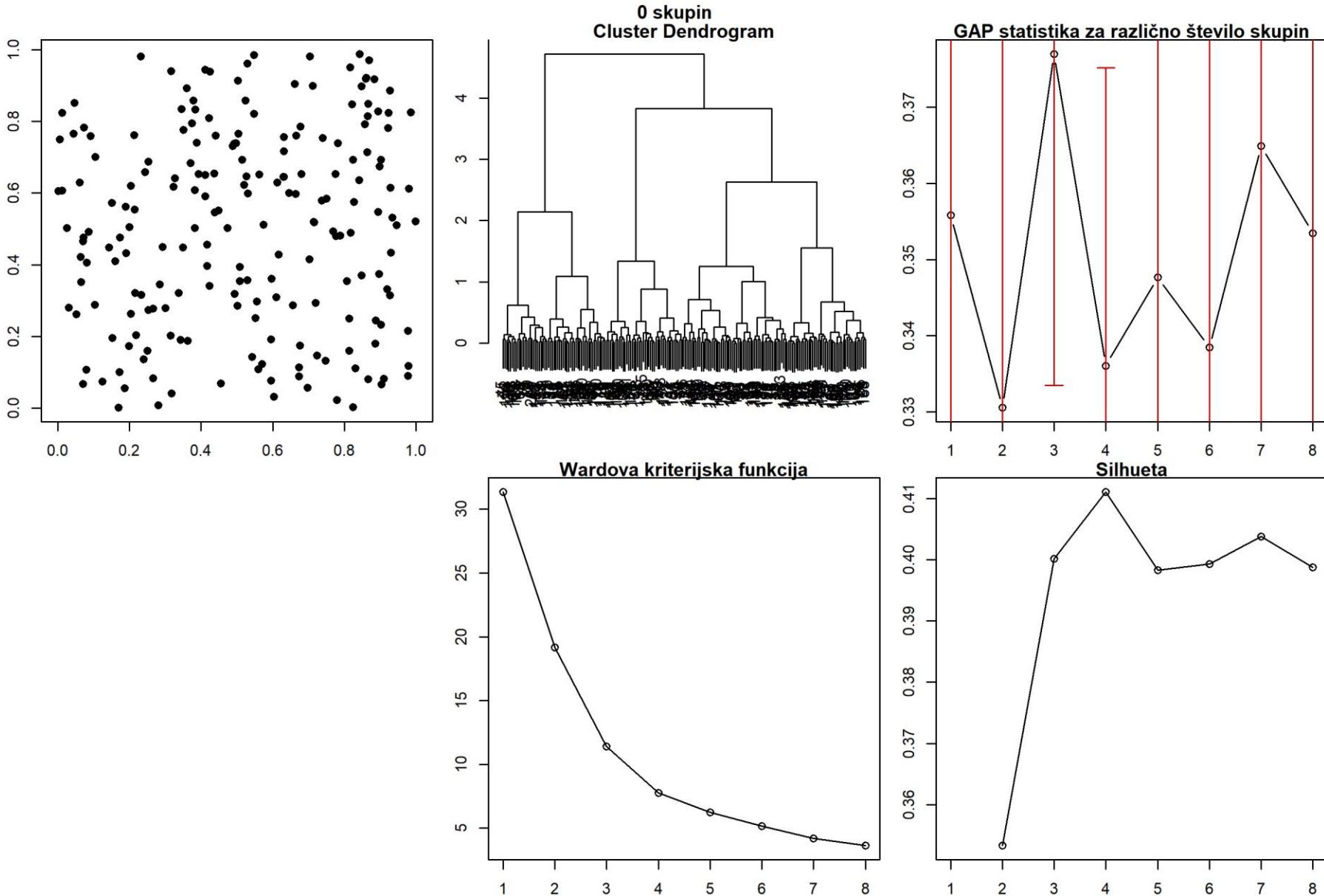
n = 400

2 clusters  $C_j$   
 $j : n_j | \text{ave}_{i \in C_j} s_i$

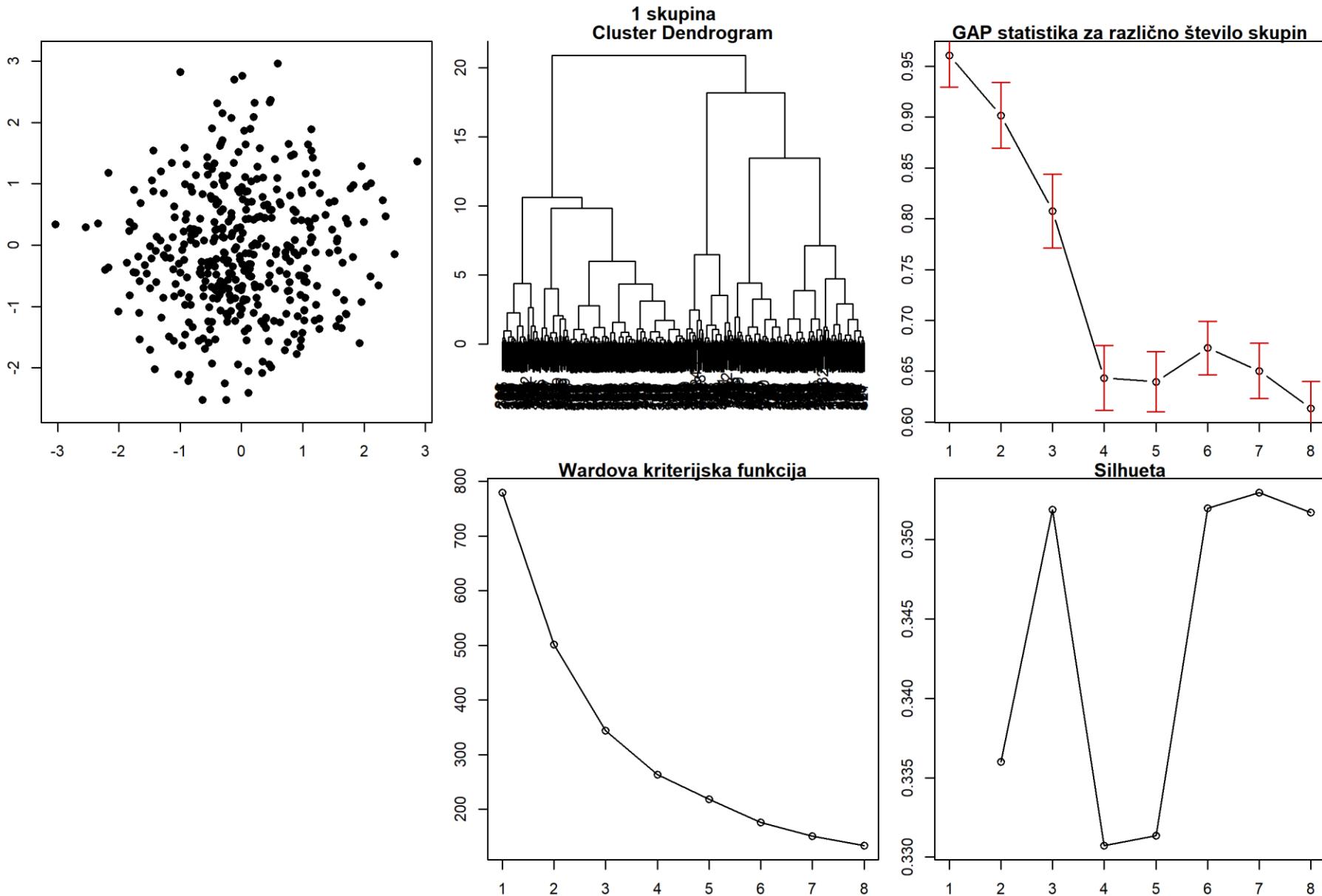


Average silhouette width : 0.46

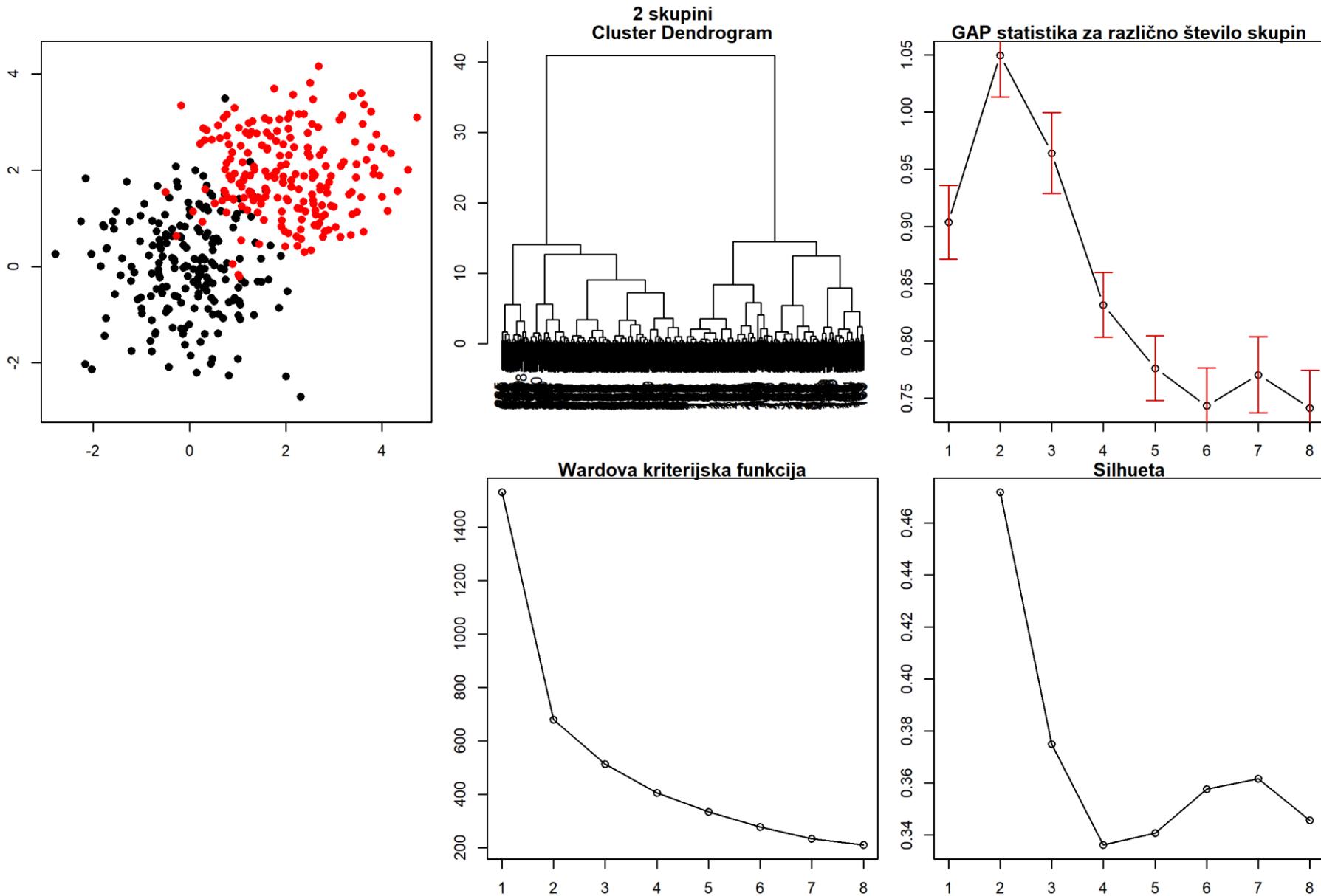
# Določanje števila skupin



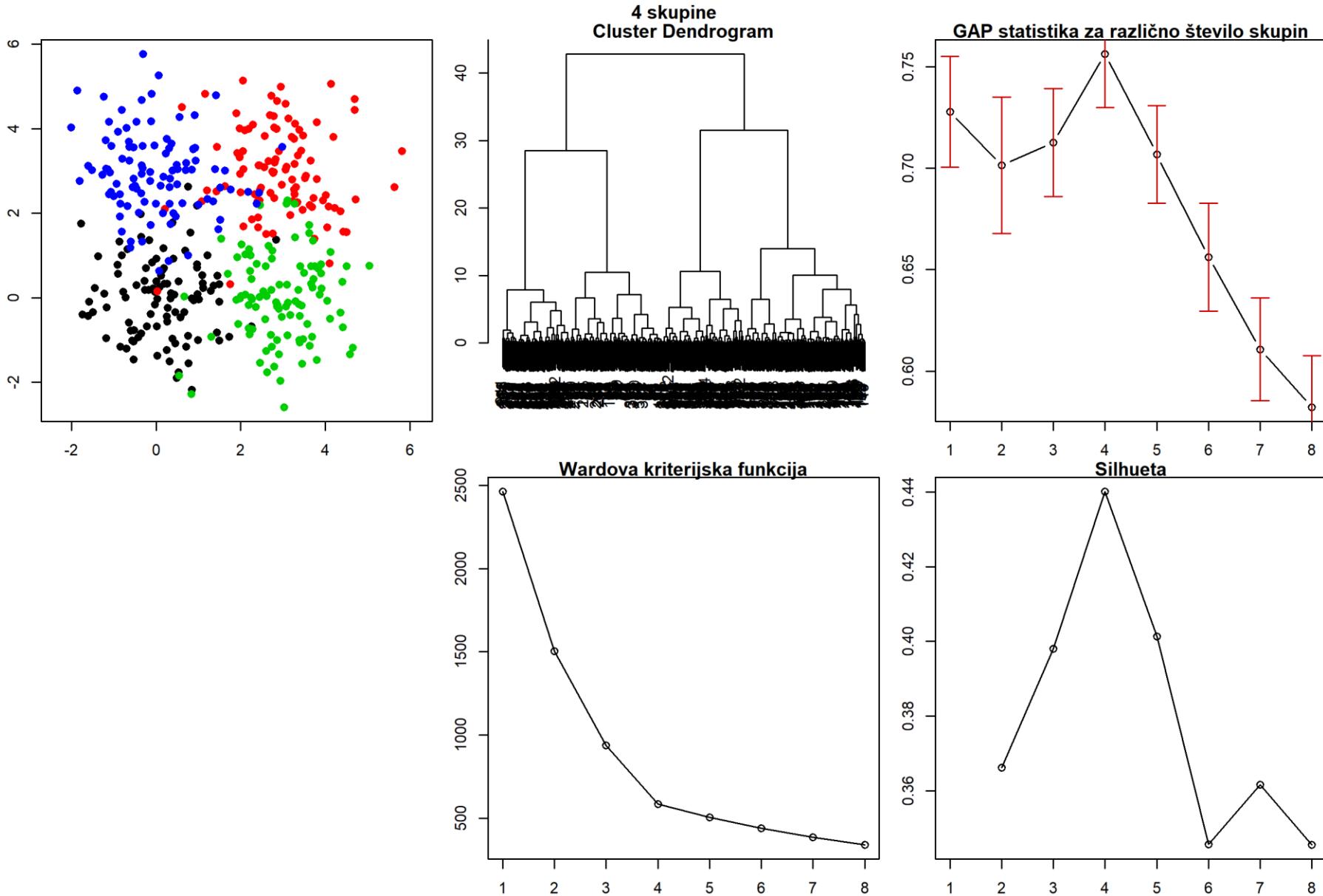
# Določanje števila skupin



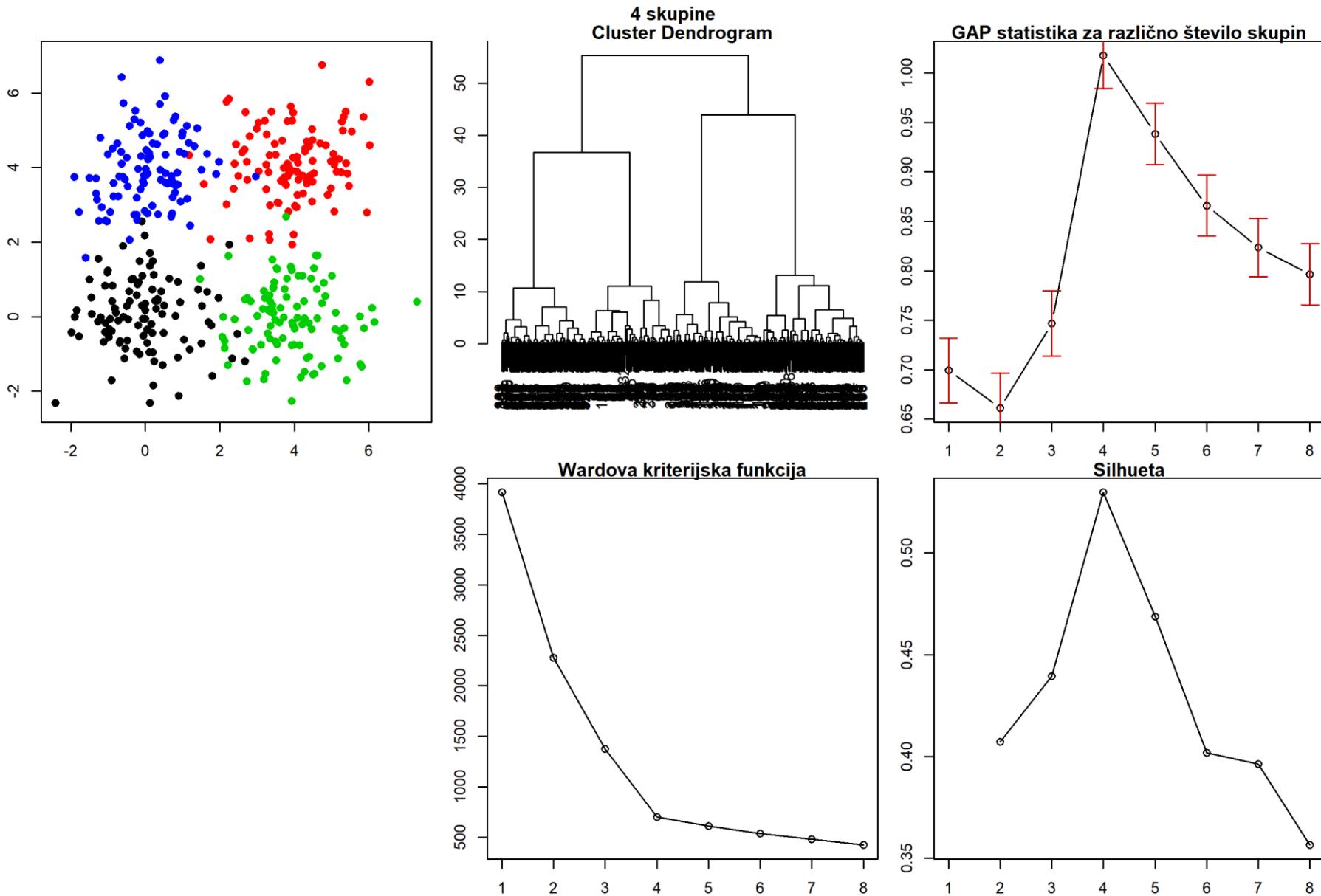
# Določanje števila skupin



# Določanje števila skupin



# Določanje števila skupin



# K-means „na roke“

Sprogramirajmo k-means algoritem v R-ju

# Lokalna optimizacija

## ■ Postopek:

1. Začnemo z nekim začetim razbitjem (skupinami)
2. Izračunamo vrednost kriterijske funkcije ( $KF$ ) za to razbije
3. Pregledamo razbitja v sosedstvu in za njih izračunamo vrednost  $KF$
4. Če najdemo razbitje z manjšo vrednostjo  $KF$ , se "premaknemo" vanj
5. Ponavljamo 3. in 4. korak, dokler še najdemo razbitje z manjšo vrednostjo  $KF$

# Razvrščanje na podlagi modelov (Fraley in Raftery, 2002)

- Metoda temelji na modelu, na podlagi katerega predvidevamo, da so generirani podatki
- To je vedno neka “mešanica” (ang. mixture) porazdelitev (npr. multivariatnih normalnih) z različnimi parametri (imenujemo jih komponente)
- Z metodo ocenimo:
  - Parametre “mešanice” (število skupin, parametre za vsako skupino)
  - Kateri skupini pripada posamezna enota oz. iz katere porazdelitve je bila generirana

# Razvrščanje na podlagi modelov

- V praksi ponavadi predpostavljamo mešanico **multivariatnih normalnih porazdelitev**
- Multivariatna normalna porazdelitev ima dva parametra:
  - Vektor povprečji  $\mu_k$
  - Kovariančno matriko:  $\Sigma_k$
- Glede na omejitve na kovariančni matriki lahko imamo različno kompleksne modele

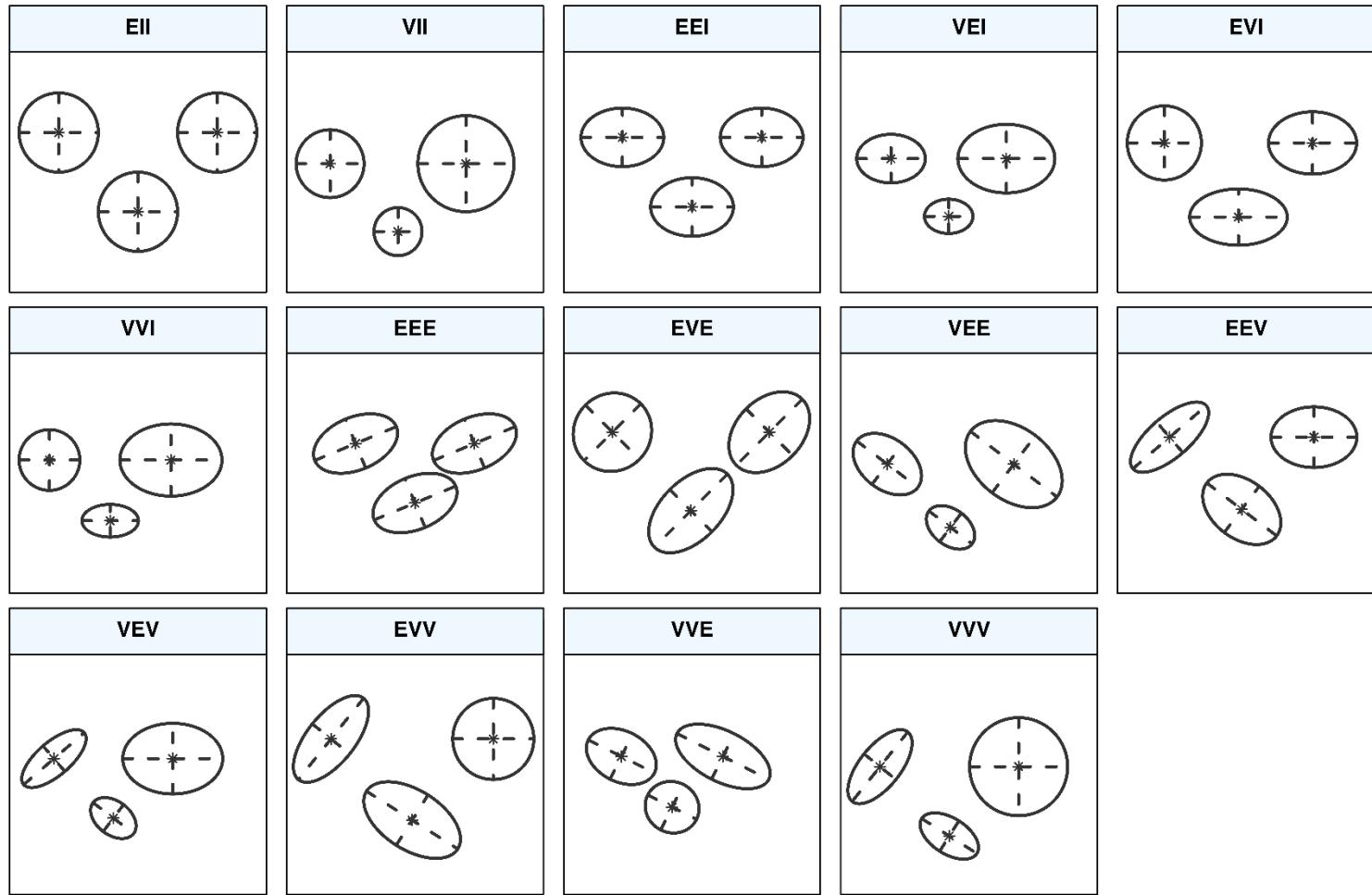
# Razvrščanje na podlagi modelov

- Kovariančno matriko razbijemo:

$$\Sigma_k = \lambda_k D_k A_k D_k^T$$

- $\lambda_k$  - “volumen“ komponente
  - $A_k$  - oblika komponente (v smislu okrogle oz. (kako eliptična)
  - $D_k$  - „smer“ oz. rotacija komponente
- Z omejevanjem teh parametrov na določeno vrednost ali na enakost med komponentami (skupinami) dobimo različno kompleksne modele

# Razvrščanje na podlagi modelov



**Figure 2:** Ellipses of isodensity for each of the 14 Gaussian models obtained by eigen-decomposition in case of three groups in two dimensions.

# Razvrščanje na podlagi modelov

Model	$\Sigma_k$	Distribution	Volume	Shape	Orientation
EII	$\lambda I$	Spherical	Equal	Equal	—
VII	$\lambda_k I$	Spherical	Variable	Equal	—
EEI	$\lambda A$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k A$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda A_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda DAD^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda DA_k D^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k DAD^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k DA_k D^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda D_k AD_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k D_k AD_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda D_k A_k D_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k D_k A_k D_k^\top$	Ellipsoidal	Variable	Variable	Variable

# Razvrščanje na podlagi modelov

- Parametre porazdelitev ocenimo z metodo največjega verjetja, natančneje z EM algoritmom (na podlagi predhodnih ocen)
- Na podlagi BIC (Bayesian information criterion) lahko tudi izberemo najprimernejši model (z največjo vrednostjo BIC), tako v smislu kompleksnosti modela kot tudi števila skupin. Uporabljajo formulo:

$$\text{BIC}_{\mathcal{M}, \mathcal{G}} = 2\ell_{\mathcal{M}, \mathcal{G}}(x|\widehat{\Psi}) - \nu \log(n)$$

- BIC temelji na logaritmu verjetja, ki ga “popravimo” za kompleksnost modela („število parametrov“\* $\log(n)$ )

# Razvrščanje na podlagi modelov

Rezultat modela je:

- Ocene parametrov porazdelitev za vsako komponento (povprečja, kovariančne matrike)
- Ocene deležev komponent v vzorcu
- Za vsako enoto verjetnost, da izhaja iz (oz. pripada) posamezni komponenti/skupini

# Razvrščanje na podlagi modelov

- Teoretično najprimernejša metoda, če so predpostavke (multivariatna normalna porazdelitev) izpolnjene
- Uporaba na diskretnih spremenljivkah še ni dovolj preizkušena.

# Primerjava razbitij

- S pomočjo kontingenčne tabele in ustreznih indeksov
- Najpogosteje uporabljena: Randov indeks (Rand, 1971) in Popravljen Randov indeks (popravljen za slučajnost) (Hubert in Arabie, 1985)
- Uporabimo lahko tudi mere povezanosti za nominalne spremenljivke (manj priporočljivo), npr. Cramerjev V (ali  $\alpha$ )

# Primerjava raz. – Randov indeks

- Randov indeka (Rand, 1971): Vrednost predstavlja delež parov enot, ki sta v obeh razbitjih „usklajena“ (ali v obeh v isti skupini, ali v obhe v različnih skupinah).
- Definiran je na razponu med 0 in 1
- Njegova težava je, da že pri dveh slučajnih razbitjih lahko dosega precej velike vrednosti.

# Popravljen Randov indeks

- Popravljen Randov indeks (Hubert in Arabie, 1985) je Randov indeks, popravljen za slučajnost.
- Izračunamo ga kot:

$$ARI = \frac{RI - E(RI)}{RI_{max} - E(RI)}$$

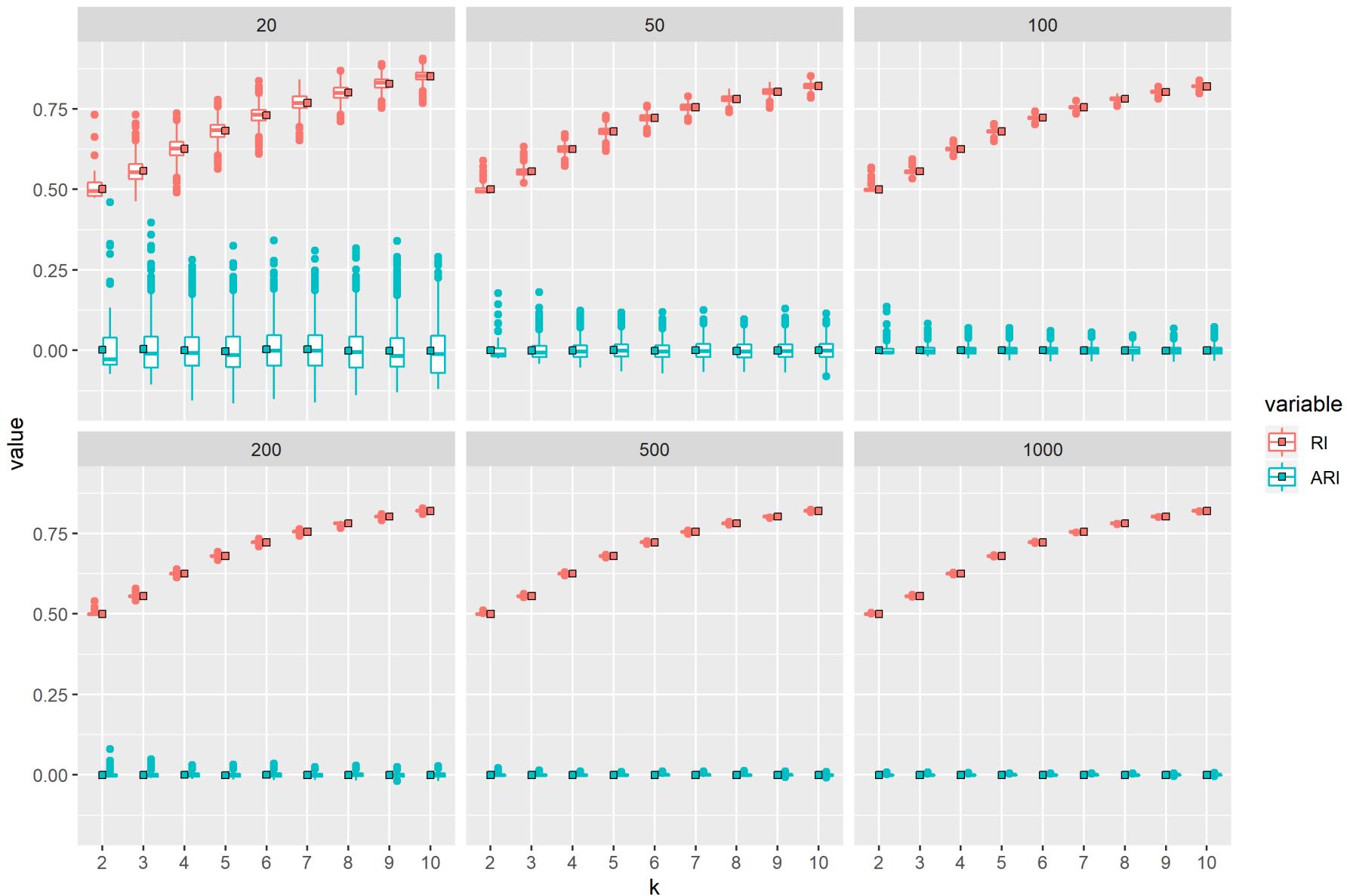
- 1 pomeni identični razbitji, 0 pa, da sta razbitji tako podobni, kot bi pričakovali po slučaju.
- Navzdol ni omejen, a vrednosti pod 0 so redke.

# Popravljen Randov indeks

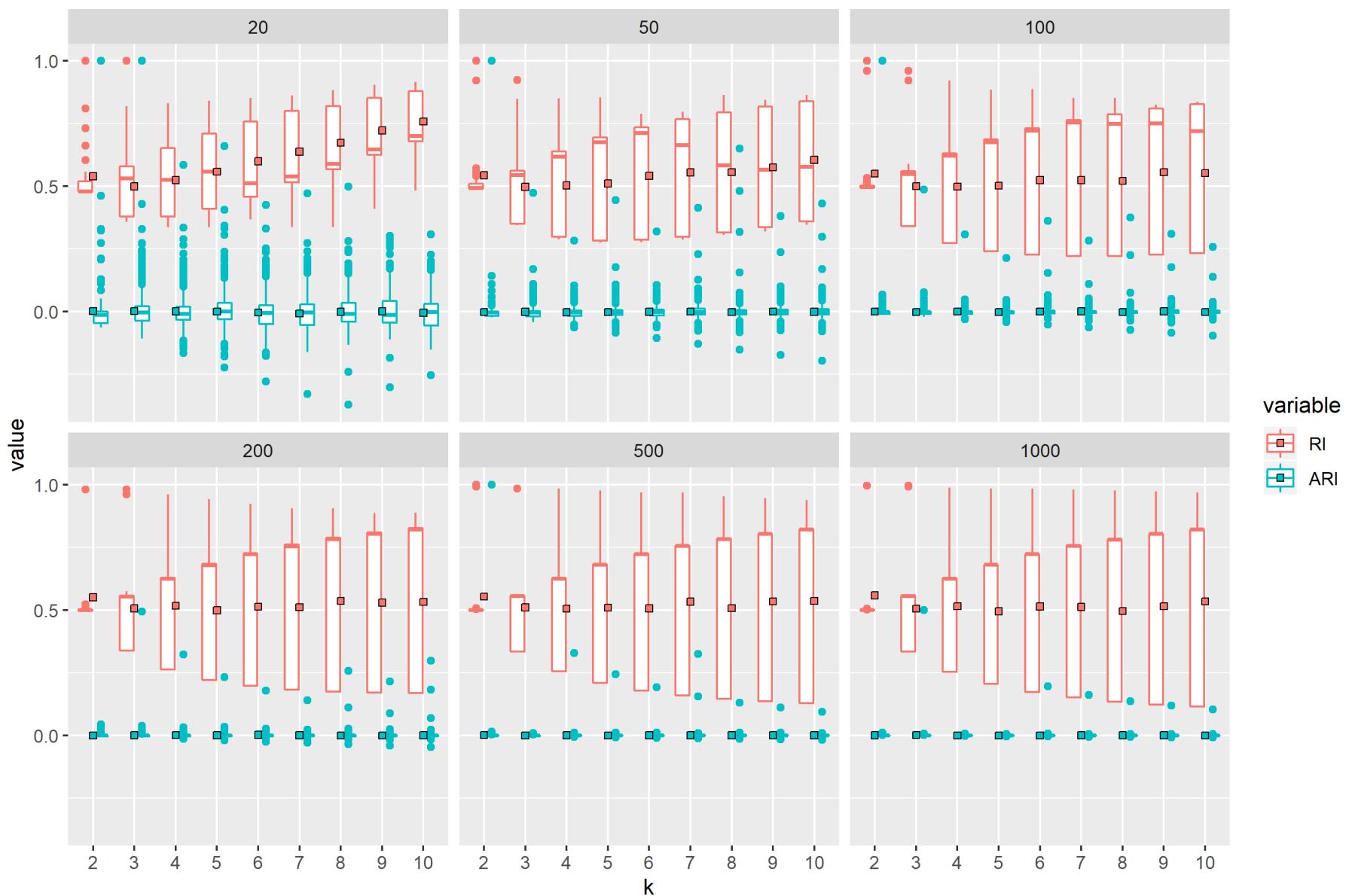
## Smernice za interpretacijo

$ARI = 0$	Razbitji si podobni le toliko, kot bi lahko pričakovali po slučaju
$0 < ARI \leq 0.2$	Razbitji sta zanemarljivo podobni
$0.2 < ARI \leq 0.4$	Razbitji sta malo podobni
$0.4 < ARI \leq 0.6$	Razbitji sta srednje podobni
$0.6 < ARI \leq 0.8$	Razbitji sta precej podobni
$0.8 < ARI < 1$	Razbitji sta zelo podobni
$ARI = 1$	Razbitji sta identični

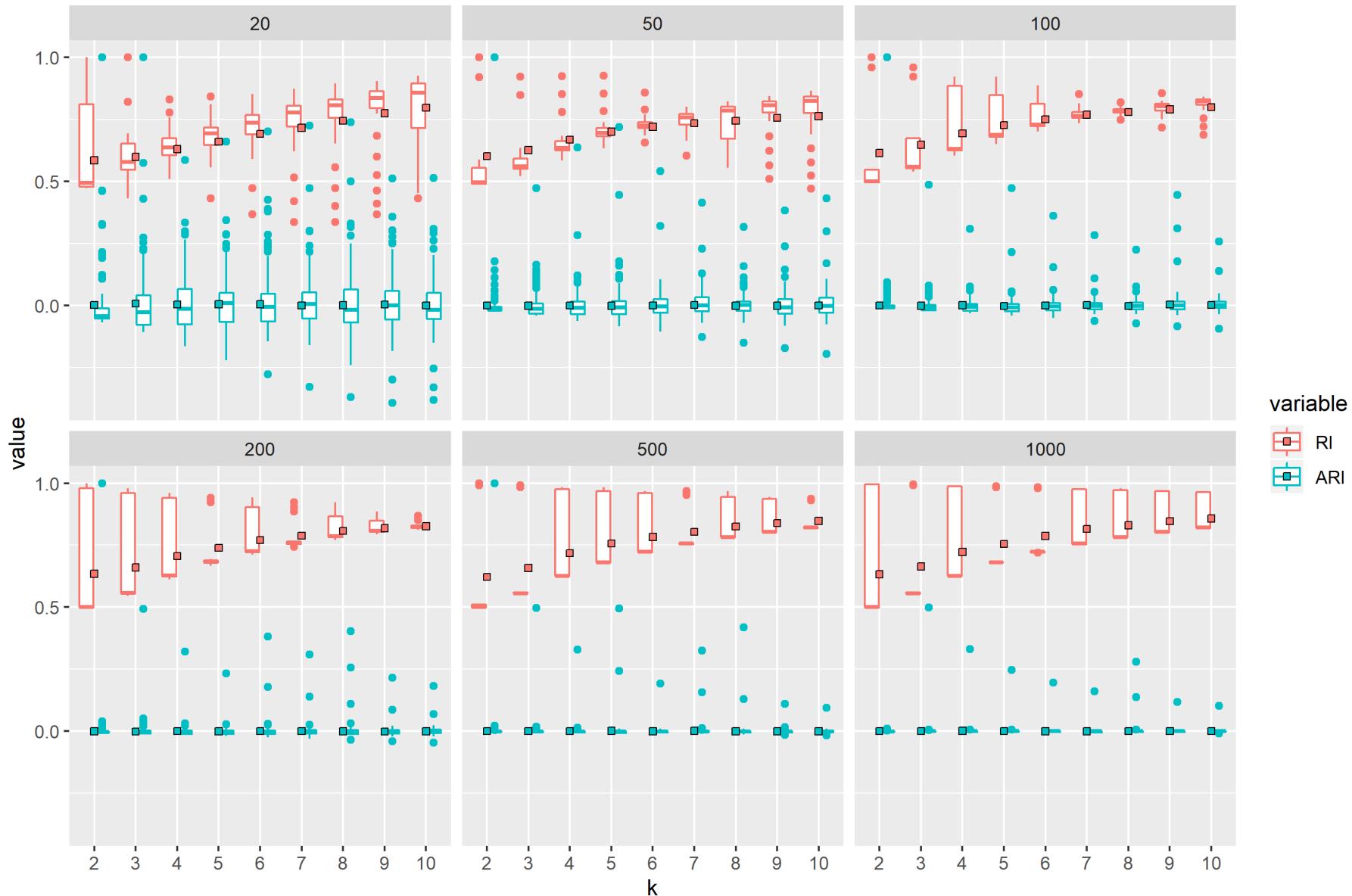
# (A) RI– slučajna razbitja



# (A) RI – slučajna razbitja „Pajek“



# (A) RI – slučajni razbitji „Pajek“ – obe na enak sistem



# Koraki reševanja (prirejeno po Hansen, Jaumard, Sanlaville 1993)

1. Izberi enote in ustrezne spremenljivke.
2. Izberi problemu in tipu spremenljivk ustrezeno definicijo/mero različnosti.
3. Izberi ustrezeni tip razvrstitev (npr. razbitje, hierarhija).
4. Izberi ustrezeno kriterijsko funkcijo (e.g., Wardova kriterijska funkcija) ali kriterij za razvrščanje. → Izberi postopek za razvrščanje v skupine glede na postavljeni problem.
5. Določi razvrstitev (razvrstitev) z izbranim postopkom.
6. Z metodami opisne statistike poišči lastnosti dobljenih skupin in z ustreznimi postopki razišči, če je dobljena razvrstitev razkrila naravno strukturo podatkov.
7. Opiši skupine glede na ostale spremenljivke.

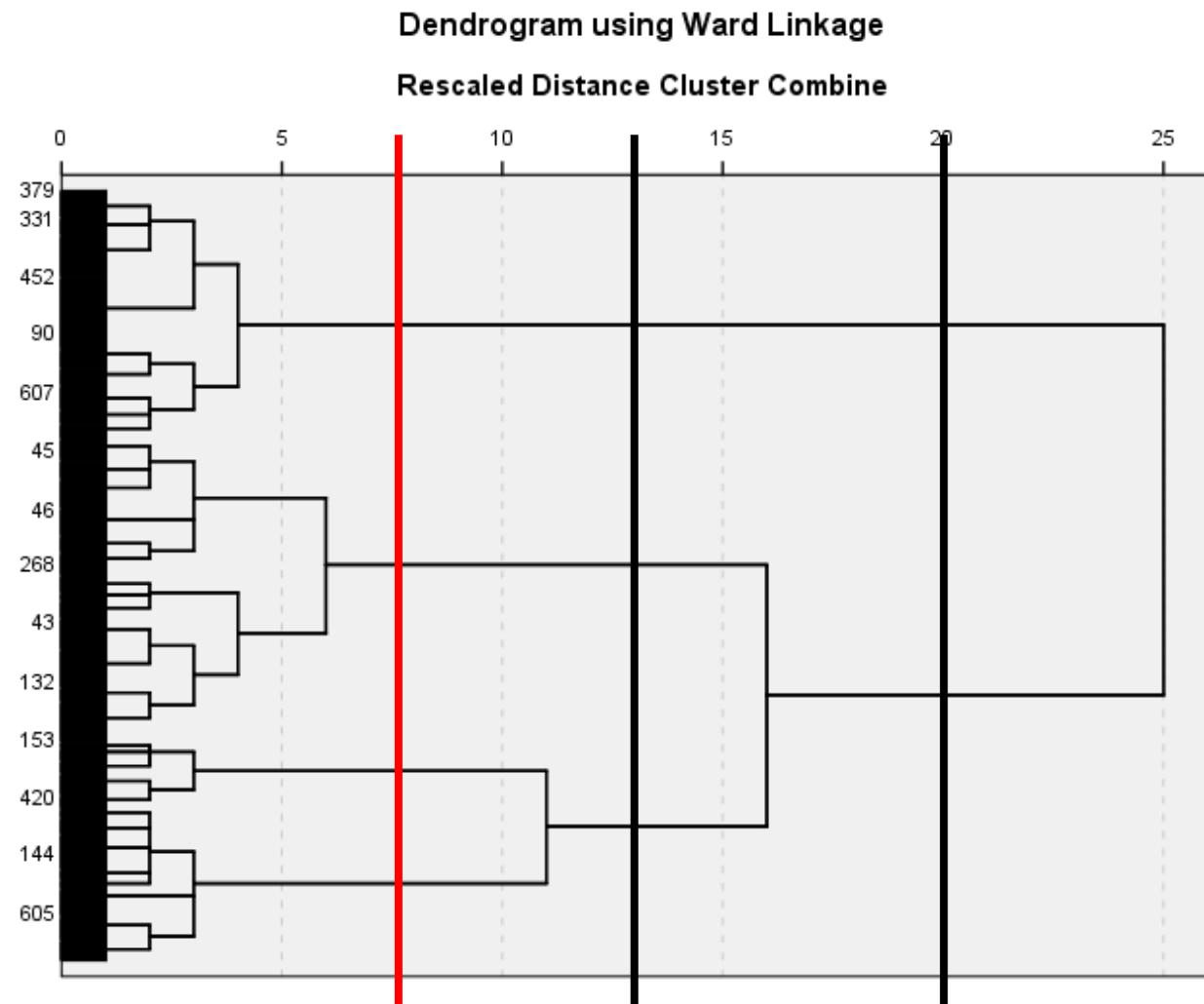
# Primer

- Podatki so bili zbrani v okviru raziskave *Kakovost merjenja egocentričnih socialnih omrežij* (Ferligoj in drugi, 2000) leta 2000. Vzorec vsebuje 1033 prebivalcev Ljubljane. Analiza je narejena le na 631 prebivalcih, ki so bili osebno anketirani.
- Za razvrščanje smo izbrali spremenljivke, ki merijo Ekstravertiranost in Emocionalno stabilnost.

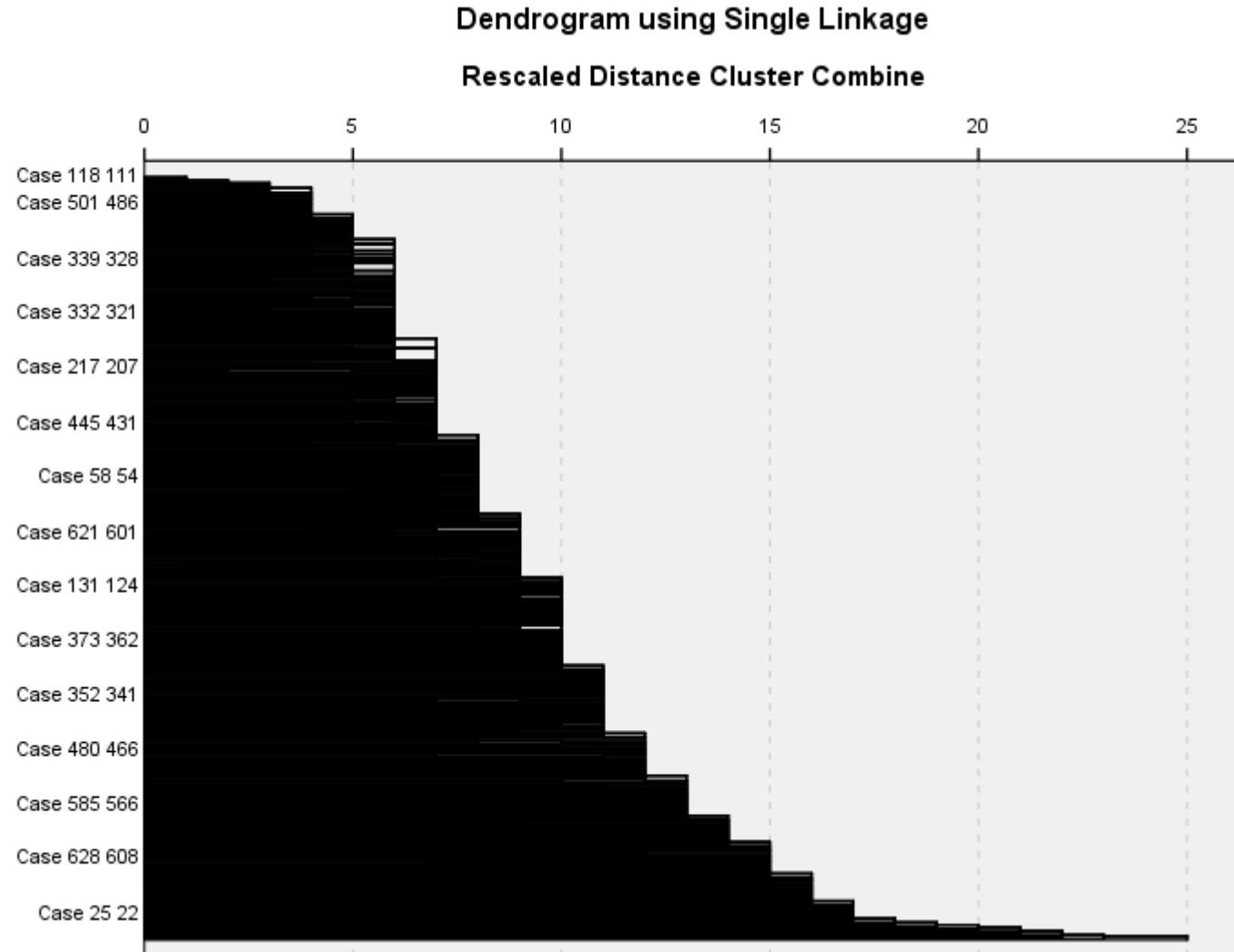
# Primer: Hierarhično razvrščanje

- Standardizacija spremenljivk
- Kvadrirana evklidska razdalja
- Wardova metoda (za primerjavo tudi minimalna in maksimalna povezanosti)

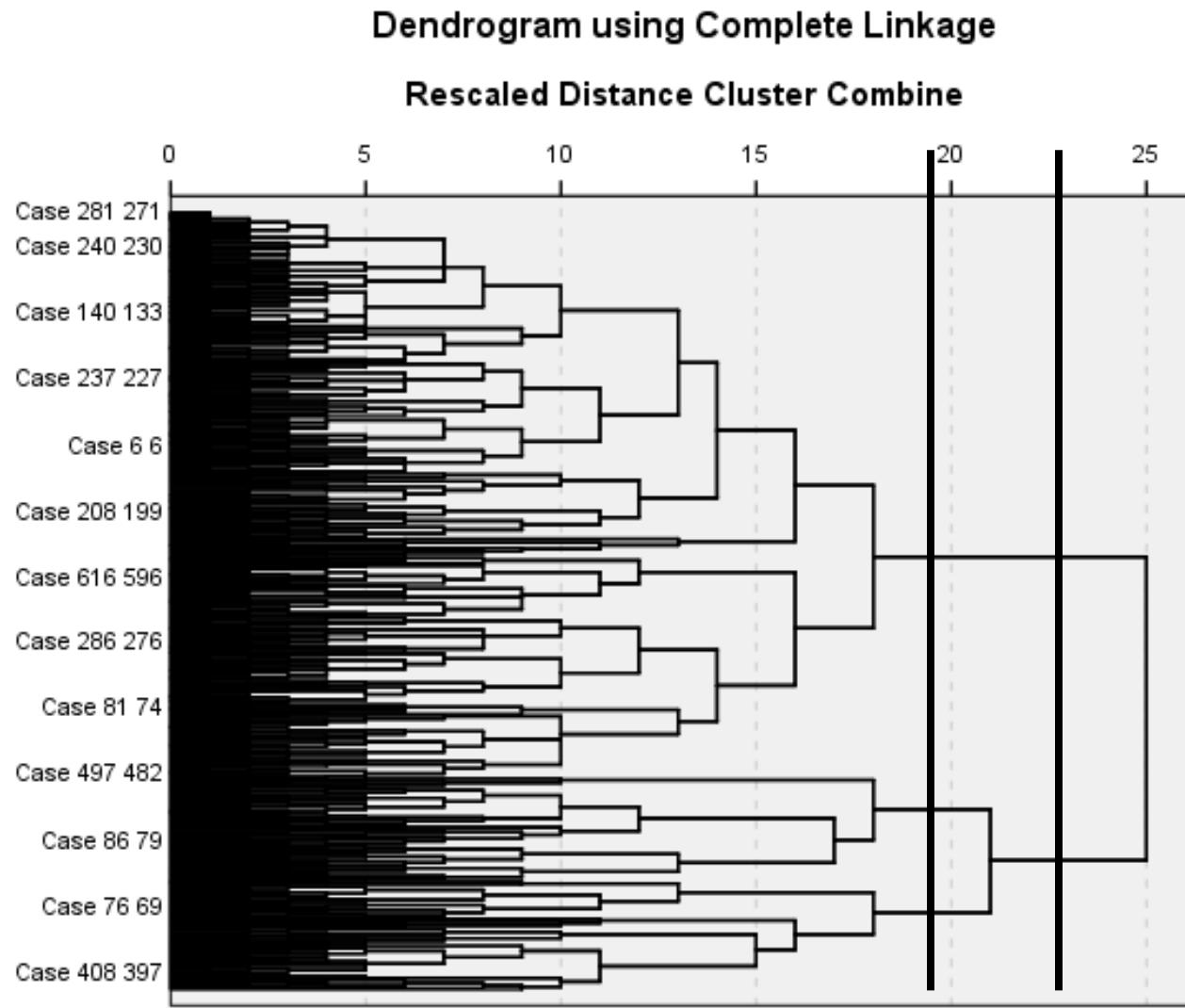
# Primer: dendrogram



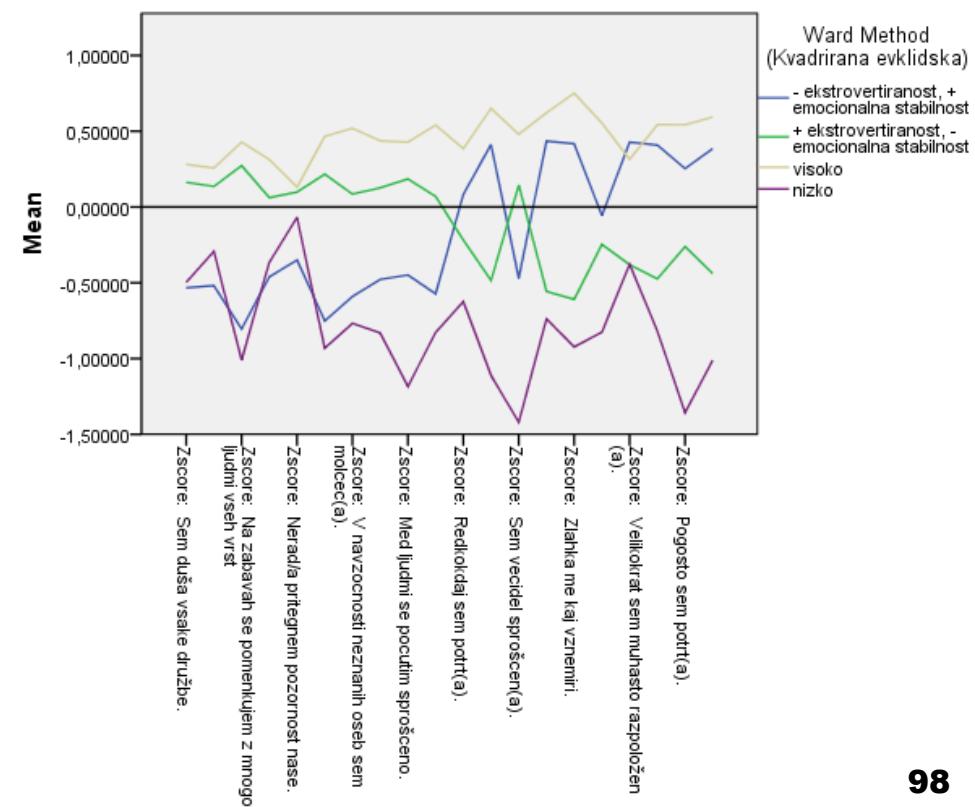
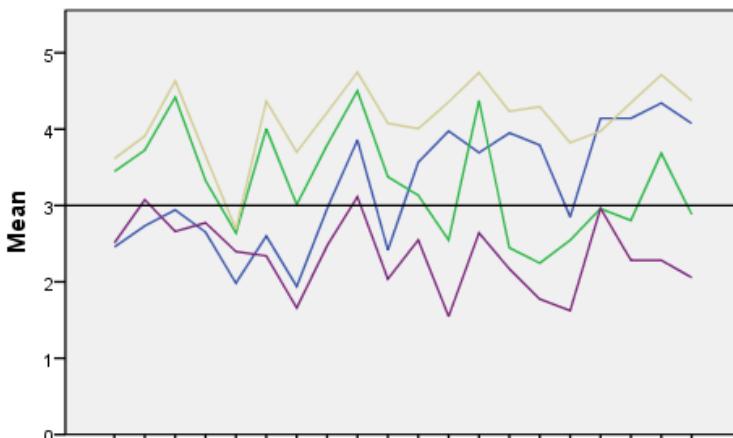
# Primer: dendrogram



# Primer: dendrogram



# Primer: opis skupin (kvadratna evklidska, Ward)



# Primer: opis skupin (kvadratna evklidska, Ward)

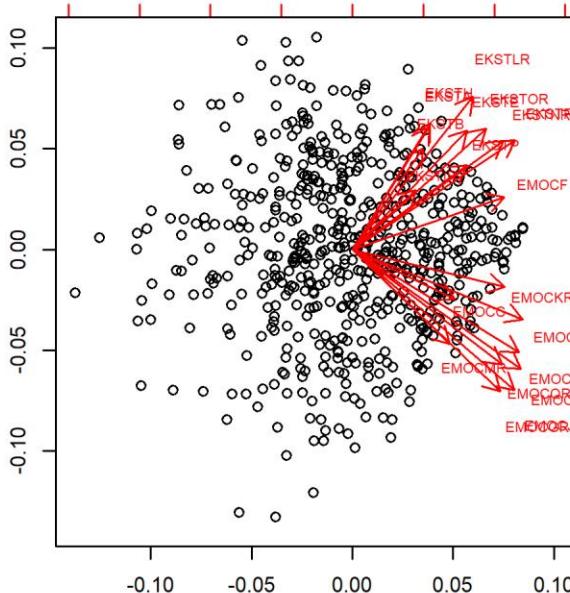
	CLU4_1 Ward Method			
	1	2	3	4
Frekvenca	120	240	197	53
EKSTA Sem duša vsake družbe.	2,46	3,45	3,61	2,51
EKSTB Ne moti me, ce sem v središcu pozornosti.	2,73	3,73	3,91	3,08
EKSTE Na zabavah se pomenkujem z mnogo ljudmi vseh vrst	2,94	4,42	4,63	2,66
EKSTH Pogovore nacenjam jaz.	2,65	3,33	3,65	2,77
EKSTIR Nerad/a pritegnem pozornost nase.	1,98	2,64	2,69	2,40
EKSTLR Sem redkobeseden(a).	2,60	4,00	4,37	2,34
EKSTNR V navzočnosti neznanih oseb sem molčeč(a).	1,94	3,01	3,70	1,66
EKSTOR Imam malo povedati.	2,96	3,79	4,22	2,47
EKSTP Med ljudmi se pocutim sprošceno.	3,86	4,50	4,75	3,11
EKSTRR Zadržujem se v ozadju.	2,42	3,38	4,08	2,04
EMOCC Redkokdaj sem potrt(a).	3,57	3,13	4,01	2,55
EMOCDR Zlahka me kaj vrže iz tira.	3,98	2,55	4,36	1,55
EMOCF Sem vecidel sprošcen(a).	3,69	4,37	4,74	2,64
EMOCGR Zlahka me kaj razdraži.	3,95	2,45	4,23	2,17
EMOCJR Zlahka me kaj vznemiri.	3,79	2,25	4,29	1,77
EMOCKR Sem zaskrbljene narave.	2,85	2,55	3,82	1,62
EMOCMR Velikokrat sem muhasto razpoložen(a).	4,14	2,96	3,97	2,96
EMOCQR Moje razpoloženje se pogosto menja.	4,14	2,80	4,35	2,28
EMOCSR Pogosto sem potrt(a).	4,34	3,68	4,71	2,28
EMOCTR Zlahka se me poloti napetost.	4,08	2,88	4,38	2,06

# Interpretacija skupin

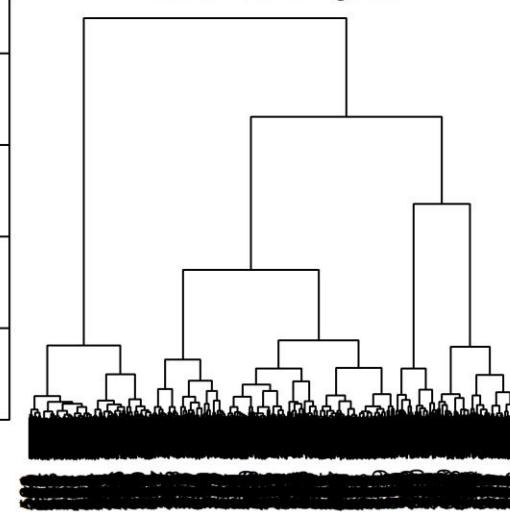
Po skupinah:

- - ekstrovertiranost, + emocionalna stabilnost (120 enot). V povprečju so ljudje iz te skupine emocionalno stabilni in niso ekstrovertirani (niso pa popolnoma introvertirani oz. ne-ekstrovertirani) oz. se rahlo ne strinjajo z izjavami, ki merijo ekstrovertiranost. So nadpovprečno ekstrovertirani in podpovprečno emocionalno stabilni.
- + ekstrovertiranost, - emocionalna stabilnost (240 enot). Gre za največjo skupino, ki je na nek način ravno obratna prejšnji. Osebe iz te skupine so precej ekstrovertirane, emocionalno stabilnost pa ocenjujejo okoli srednje vrednosti.
- Visoko (197 enot). Precej velika skupina, kjer so povprečne vrednosti vseh spremenljivk precej visoke (skoraj povsod najvišje med skupinami) in seveda precej nadpovprečne.
- Nizko (53 enot). Najmanjša skupina, kjer so povprečne vrednosti pod ali se kvečjemu približajo srednji vrednosti. Nizke so predvsem pri nekaterih spremenljivkah emocionalne stabilnosti.

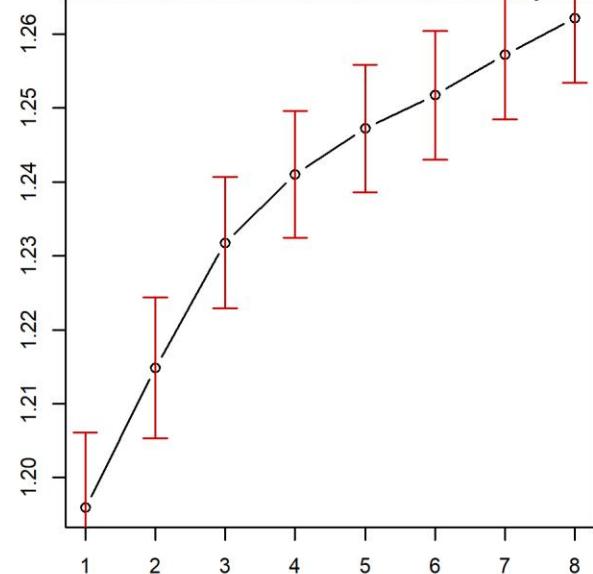
# Primer: Število skupin?



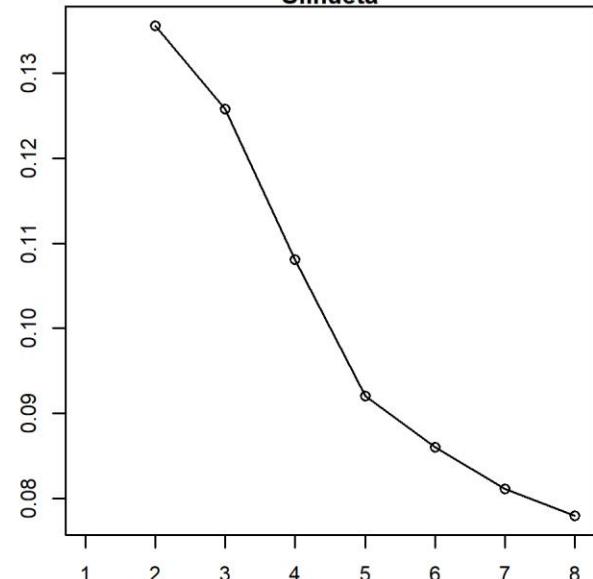
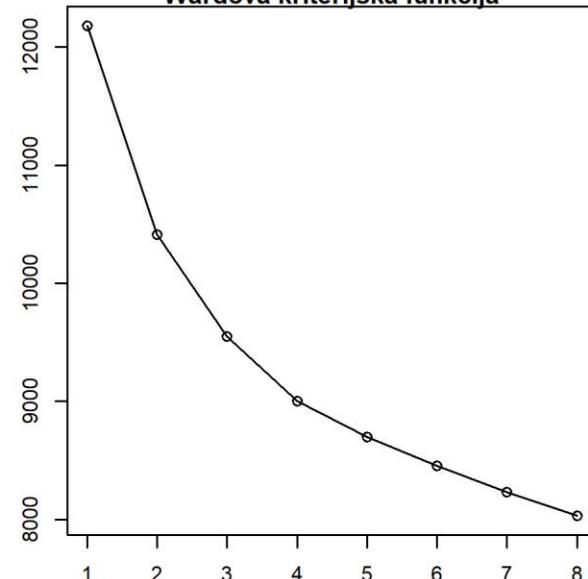
Cluster Dendrogram



GAP statistika za različno število skupin



Wardova kriterijska funkcija



Ni mogoče  
jasno razbrati  
pravega  
števila.

# Primer: Metoda voditeljev

- Standardizacija spremenljivk
- 4 skupine (na podlagi dendrograma)
- Voditelji:
  - SPSS-ovi: SPSS izbere čim bolj oddaljene enote
  - Vsiljeni (na podlagi centrov iz hierarhičnega razvrščanja)
  - Veliko (npr. 1000) slučajno izbranih (makro)

# Primer: Vrednost kriterijske funkcije

**Descriptive Statistics**

	N	Sum
d2ward	610	8935,91
d2spss	610	8933,13
Valid N (listwise)	610	

Hierarhično razvrščanje: 9354,03

# Primer: Slučajno izbrani voditelji

- Uporabil sem makro, s katerim sem kmeans v SPSS-u pognal 1000-krat, vsakič z slučajno izbranimi začetnimi voditelji
- Nato sem poiskal razvrstitev, ki ima najmanjšo vrednost kriterijske funkcije.

# Primer: Slučajno izbrani voditelji

- Najboljša je 527-ta ponovitev
- Vrednost KF je **8928,56**, kar je znatno manj kot z voditelji, ki jih je izbral SPSS
- Pogledali bomo rešitev za to ponovitev

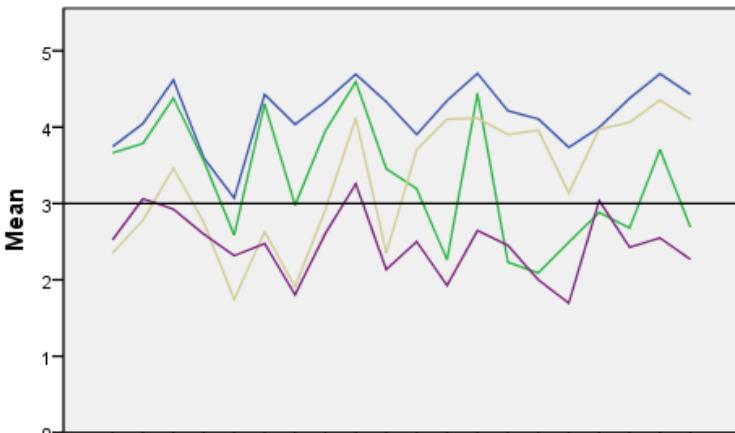
Descriptive Statistics		
	N	Sum
KMDISTSQ527	610	8928.56
KMDISTSQ416	610	8928.59
KMDISTSQ871	610	8928.59
KMDISTSQ699	610	8928.61
KMDISTSQ129	610	8928.71
KMDISTSQ404	610	8928.71
KMDISTSQ347	610	8928.77
KMDISTSQ128	610	8928.77
KMDISTSQ256	610	8928.81
KMDISTSQ731	610	8928.99

Rows 1 through 10 of 1001

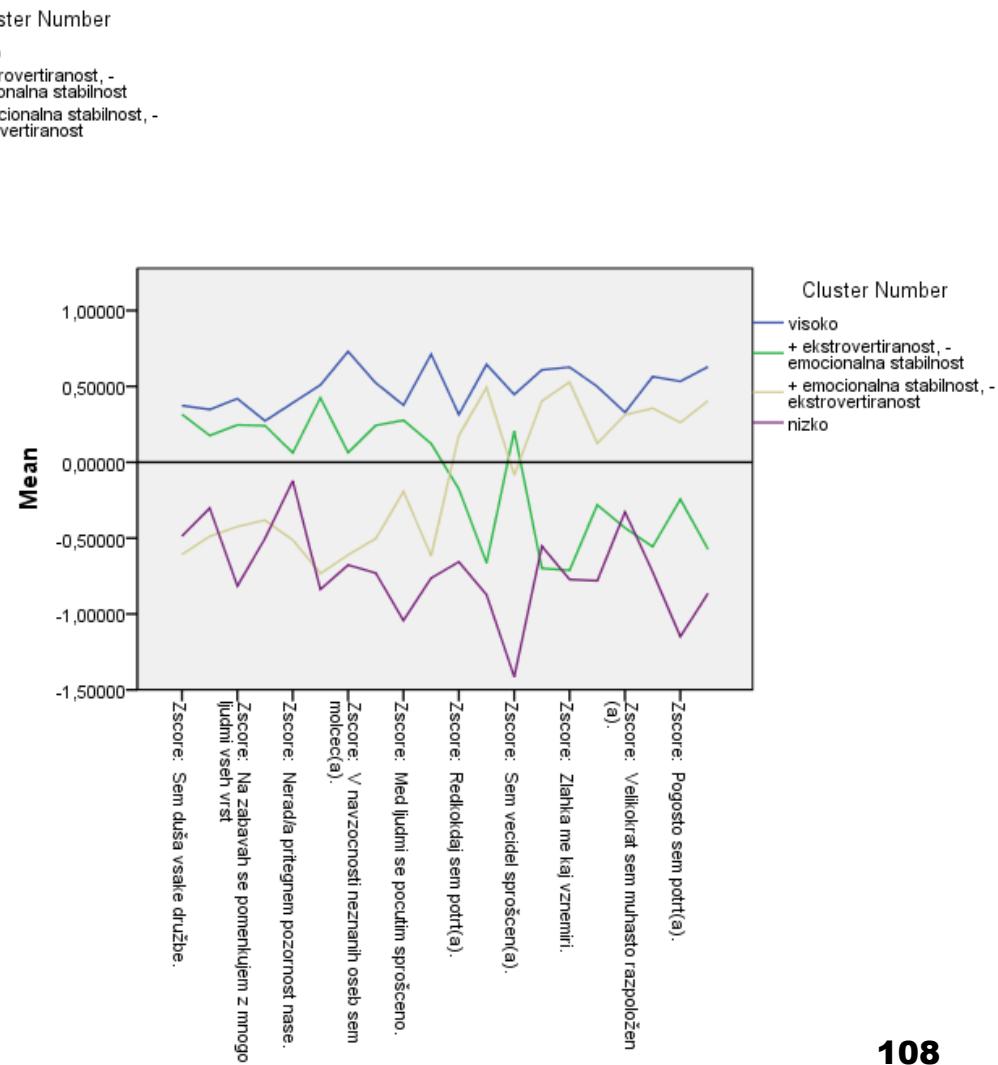
↓ ↓

# Primer: opis skupin

## (metoda voditeljev – slučajni)



- Zlahka se me poloti napetost.
- Pogosto sem potrta(a).
- Male razpoloženje se pogosto menja.
- Včiljkrat sem muhasto razpoložen(a).
- Sem zasitiljene narave.
- Zlahka me kaj vzremi.
- Zlahka me kaj razdraži.
- Sem večidel sproščen(a).
- Zlahka me kaj vrže iz tira.
- Redkolokaj sem potrta(a).
- Zadružujem se v ozadju.
- Imed ljudmi se pocutim sproščeno.
- Imam malo povediti.
- V navzročnosti neznanih oseb sem molce(a).
- Ssem redkolobeseden(a).
- Nerada pri temem pozornost nase.
- Pogovore načenjam laž.
- Na zabavah se pomenujkujem z mnogo ljudmi vsem vrst.
- Ne moti me, ce sem v srediscu pozornosti.
- Sem duša vsake družbe.



# Primer: opis skupin (metoda voditeljev)

	KMCLU527 Cluster Number			
	1	2	3	4
Frekvenca	196	187	145	82
EKSTA Sem duša vsake družbe.	3,74	3,66	2,35	2,52
EKSTB Ne moti me, ce sem v središcu pozornosti.	4,05	3,79	2,78	3,06
EKSTE Na zabavah se pomenujem z mnogo ljudmi vseh vrst	4,62	4,38	3,46	2,93
EKSTH Pogovore nacenjam jaz.	3,60	3,56	2,75	2,60
EKSTIR Nerad/a pritegnem pozornost nase.	3,07	2,59	1,74	2,32
EKSTLR Sem redkobeseden(a).	4,43	4,30	2,63	2,48
EKSTNR V navzocnosti neznanih oseb sem molcec(a).	4,04	2,98	1,91	1,80
EKSTOR Imam malo povedati.	4,34	3,95	2,92	2,61
EKSTP Med ljudmi se pocutim sprošceno.	4,69	4,59	4,12	3,26
EKSTRR Zadržujem se v ozadju.	4,33	3,45	2,35	2,13
EMOCC Redkokdaj sem potrt(a).	3,90	3,19	3,70	2,50
EMOCDR Zlahka me kaj vrže iz tira.	4,35	2,26	4,10	1,93
EMOCF Sem vecidel sprošcen(a).	4,70	4,44	4,12	2,65
EMOCGR Zlahka me kaj razdraži.	4,21	2,23	3,90	2,45
EMOCJR Zlahka me kaj vznemiri.	4,11	2,09	3,96	2,00
EMOCKR Sem zaskrbljene narave.	3,73	2,49	3,14	1,70
EMOCMR Velikokrat sem muhasto razpoložen(a).	3,99	2,88	3,97	3,04
EMOCQR Moje razpoloženje se pogosto menja.	4,38	2,68	4,06	2,43
EMOCSR Pogosto sem potrt(a).	4,70	3,71	4,35	2,55
EMOCTR Zlahka se me poloti napetost.	4,43	2,69	4,10	2,27

# Primer: opis skupin (metoda voditeljev)

	KMCLU527 Cluster Number			
	1	2	3	4
Frekvenca	196	187	145	82
ZEKSTA Zscore: Sem duša vsake družbe.	0,37	0,32	-0,61	-0,49
ZEKSTB Zscore: Ne moti me, ce sem v središcu pozornosti.	0,35	0,18	-0,49	-0,30
ZEKSTE Zscore: Na zabavah se pomenujem z mnogo ljudmi vseh vrst	0,42	0,25	-0,42	-0,82
ZEKSTH Zscore: Pogovore nacenjam jaz.	0,27	0,24	-0,38	-0,50
ZEKSTIR Zscore: Nerad/a pritegnem pozornost nase.	0,39	0,06	-0,51	-0,12
ZEKSTLR Zscore: Sem redkobeseden(a).	0,51	0,42	-0,73	-0,84
ZEKSTNR Zscore: V navzocnosti neznanih oseb sem molcec(a).	0,73	0,06	-0,61	-0,68
ZEKSTOR Zscore: Imam malo povedati.	0,52	0,24	-0,50	-0,73
ZEKSTP Zscore: Med ljudmi se pocutim sprošceno.	0,38	0,28	-0,19	-1,04
ZEKSTRR Zscore: Zadržujem se v ozadju.	0,71	0,12	-0,62	-0,76
ZEMOCC Zscore: Redkokdaj sem potrt(a).	0,31	-0,18	0,18	-0,66
ZEMOCDR Zscore: Zlahka me kaj vrže iz tira.	0,65	-0,66	0,49	-0,87
ZEMOCF Zscore: Sem vecidel sprošcen(a).	0,45	0,21	-0,08	-1,42
ZEMOCGR Zscore: Zlahka me kaj razdraži.	0,61	-0,70	0,40	-0,55
ZEMOCJR Zscore: Zlahka me kaj vznemiri.	0,63	-0,71	0,53	-0,77
ZEMOCKR Zscore: Sem zaskrbljene narave.	0,50	-0,28	0,12	-0,78
ZEMOCMR Zscore: Velikokrat sem muhasto razpoložen(a).	0,33	-0,43	0,31	-0,33
ZEMOCQR Zscore: Moje razpoloženje se pogosto menja.	0,56	-0,56	0,36	-0,72
ZEMOCSR Zscore: Pogosto sem potrt(a).	0,53	-0,24	0,26	-1,15
ZEMOCTR Zscore: Zlahka se me poloti napetost.	0,63	-0,57	0,40	-0,86

# Interpretacija skupin

Po skupinah:

- Visoko (196 enot). Največja skupina, kjer so povprečne vrednosti vseh spremenljivk precej visoke (povsod najvišje med skupinami) in seveda precej nadpovprečne.
- + ekstrovertiranost, - emocionalna stabilnost (187enot). Precej velika skupina. Osebe iz te skupine so v povprečju precej ekstrovertirane, emocionalno stabilnost pa ocenjujejo okoli srednje vrednosti. Pri tej skupini spremenljivka „Večidel sem sproščen“ še posebej izstopa. So nadpovprečno ekstrovertirane in podpovprečno emocionalno stabilne.
- - ekstrovertiranost, + emocionalna stabilnost (145 enot). V povprečju so ljudje iz te skupine emocionalno stabilni in niso ekstrovertirani (niso pa popolnoma introvertirani oz. ne-ekstrovertirani) oz. se rahlo ne strinjajo z izjavami, ki merijo ekstrovertiranost. So nadpovprečno ekstrovertirani in podpovprečno emocionalno stabilni.
- Nizko (82 enot). Najmanjša skupina, kjer so povprečne vrednosti pod ali se kvečjemu približajo srednji vrednosti.

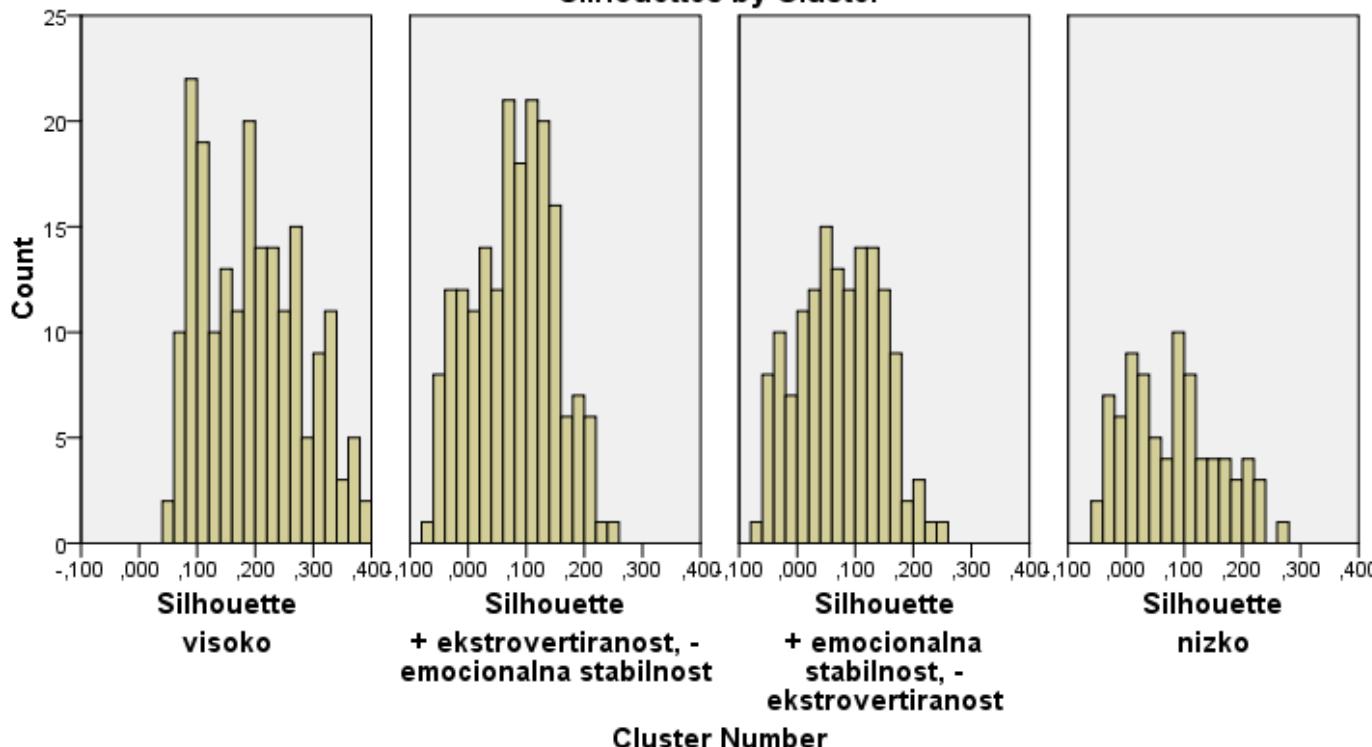
# Silhueta - KM slučajni

Statistics

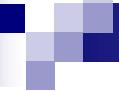
Cluster	Case Count	Mean	Minimum	Maximum
1	196,000	,195	,045	,381
2	187,000	,078	-,074	,240
3	145,000	,072	-,070	,244
4	82,000	,078	-,044	,271
Total	610,000	,114	-,074	,381

Dissimilarity measure = Euclid

Silhouettes by Cluster



Dissimilarity measure = Euclid



# Primer: primerjava KM SPSS – slučajni

**CLUkmeansSPSSk4 Cluster Number (SPSS) \* KMCLU527 Cluster Number Crosstabulation**

Count

		KMCLU527 Cluster Number				Total
		1 visoko	2 + ekstrovertiranost, - emocionalna stabilnost	3 + emocionalna stabilnost, - ekstrovertiranost	4 nizko	
CLUkmeansSPSSk4	1 nizko	0	0	2	81	83
Cluster Number (SPSS)	2 visoko	186	2	0	0	188
	3 + ekstrovertiranost, - emocionalna stabilnost	0	185	6	1	192
	4 + emocionalna stabilnost, - ekstrovertiranost	10	0	137	0	147
Total		196	187	145	82	610

**Directional Measures**

		Value	Asymp. Std. Error <sup>a</sup>	Approx. T <sup>b</sup>	Approx. Sig.
Nominal	Uncertainty Coefficient	.889	,019	43,497	,000 <sup>c</sup>
by		.888	,019	43,497	,000 <sup>c</sup>
Nominal					
	Symmetric				
	CLUkmeansSPSSk4 Cluster				
	Number (SPSS)				
	Dependent				
	KMCLU527 Cluster Number				
	Dependent				

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Likelihood ratio chi-square probability.

**Symmetric Measures**

	Value	Approx. Sig.
Nominal by Nominal	Phi	,1,656 ,000
	Cramer's V	,956 ,000
N of Valid Cases		610

Randov index: 0.964

Popravljen Randov indeks: 0.910

# Interpretacija podobnosti razbitij

Razvrstitvi sta si zelo podobni. To se vidi iz:

- V kontingenčni tabeli je v vsaki vrstici in v vsakem stolpcu le ena večja vrednost, vse ostale pa so zelo majhne. To pomeni, da lahko za vsak razred iz ene porazdelitve jasno najdemo ujemajoči razred iz druge porazdelitve, kjer je večino enot iz tega razreda.
- Vsi indeksi ujemanja razbitij so zelo visoki. Popravljen Randov indeks 0.91, koeficient negotovosti 0.89, Cramerjev  $V/\alpha$  0.96.

# Primer: primerjava Ward – KM slučajni

CLUwardE2k4 Ward Method (Kvadrirana evklidska) \* KMCLU527 Cluster Number Crosstabulation

Count

		KMCLU527 Cluster Number				Total
		1 visoko	2 + ekstrovertiranost, - emocionalna stabilnost	3 + emocionalna stabilnost, - ekstrovertiranost	4 nizko	
CLUwardE2k4 Ward Method (Kvadrirana evklidska)	1 - ekstrovertiranost, + emocionalna stabilnost	6		6	96	12
	2 + ekstrovertiranost, - emocionalna stabilnost	30		171	19	20
	3 visoko	160		10	27	0
	4 nizko	0		0	3	50
Total		196		187	145	82
						610

Directional Measures

			Value	Asymp. Std. Error <sup>a</sup>	Appro x. T <sup>b</sup>	Appro x. Sig.
Nominal by Nominal	Uncertainty Coefficient	Symmetric CLUwardE2k4 Ward Method (Kvadrirana evklidska)	,485 ,499	,027 ,027	17,393 17,393	,000 <sup>c</sup> ,000 <sup>c</sup>
		Dependent KMCLU527 Cluster Number				
		Dependent	,472	,027	17,393	,000 <sup>c</sup>

Symmetric Measures

	Value	Approx. Sig.
Nominal by Nominal	Phi	,1,231
	Cramer's V	,711
	N of Valid Cases	,610

Randov indeks: 0.795

Popravljen Randov indeks: 0.500

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Likelihood ratio chi-square probability.

# Interpretacija podobnosti razbitij

Razvrstitvi sta si precej podobni. To se vidi iz:

- V kontingenčni tabeli je v vsaki vrstici in v vsakem stolpcu le ena večja vrednost, a ponekod so tudi druge vrednosti nezanemarljive. Še vedno lahko za vsak razred iz ene porazdelitve jasno najdemo ujemajoči razred iz druge porazdelitve, kjer je večino enot iz tega razreda.
- Vsi indeksi ujemanja razbitij so vsaj srednje visoki. Popravljen Randov indeks 0.50, koeficient negotovosti 0.49, Cramerjev  $V/\alpha$  0.71.

# Povezanost z drugimi sprem.

E\_SPOL spol ega \* KMCLU527 Cluster Number Crosstabulation

		KMCLU527 Cluster Number				Total		
		1 visoko	2 + ekstrovertiranost, - emocionalna stabilnost	3 + emocionalna stabilnost, - ekstrovertiranost	4 nizko			
E_SPOL spol ega	1 moški	Count	88	75	71	27	261	
		% within KMCLU527 Cluster Number	44,9%	40,1%	49,0%	32,9%	42,8%	
	2 ženski	Count	108	112	74	55	349	
		% within KMCLU527 Cluster Number	55,1%	59,9%	51,0%	67,1%	57,2%	
Total		Count	196	187	145	82	610	
		% within KMCLU527 Cluster Number	100,0%	100,0%	100,0%	100,0%	100,0%	

## Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,423 <sup>a</sup>	3	,093
N of Valid Cases	610		

## Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,103	,093
	Cramer's V	,103	,093
	Contingency Coefficient	,102	,093
	N of Valid Cases	610	

a. 0 cells (.0%) have expected count less than 5.  
The minimum expected count is 35,09.

# Povezanost z drugimi sprem.

## Descriptives

VEL\_OM

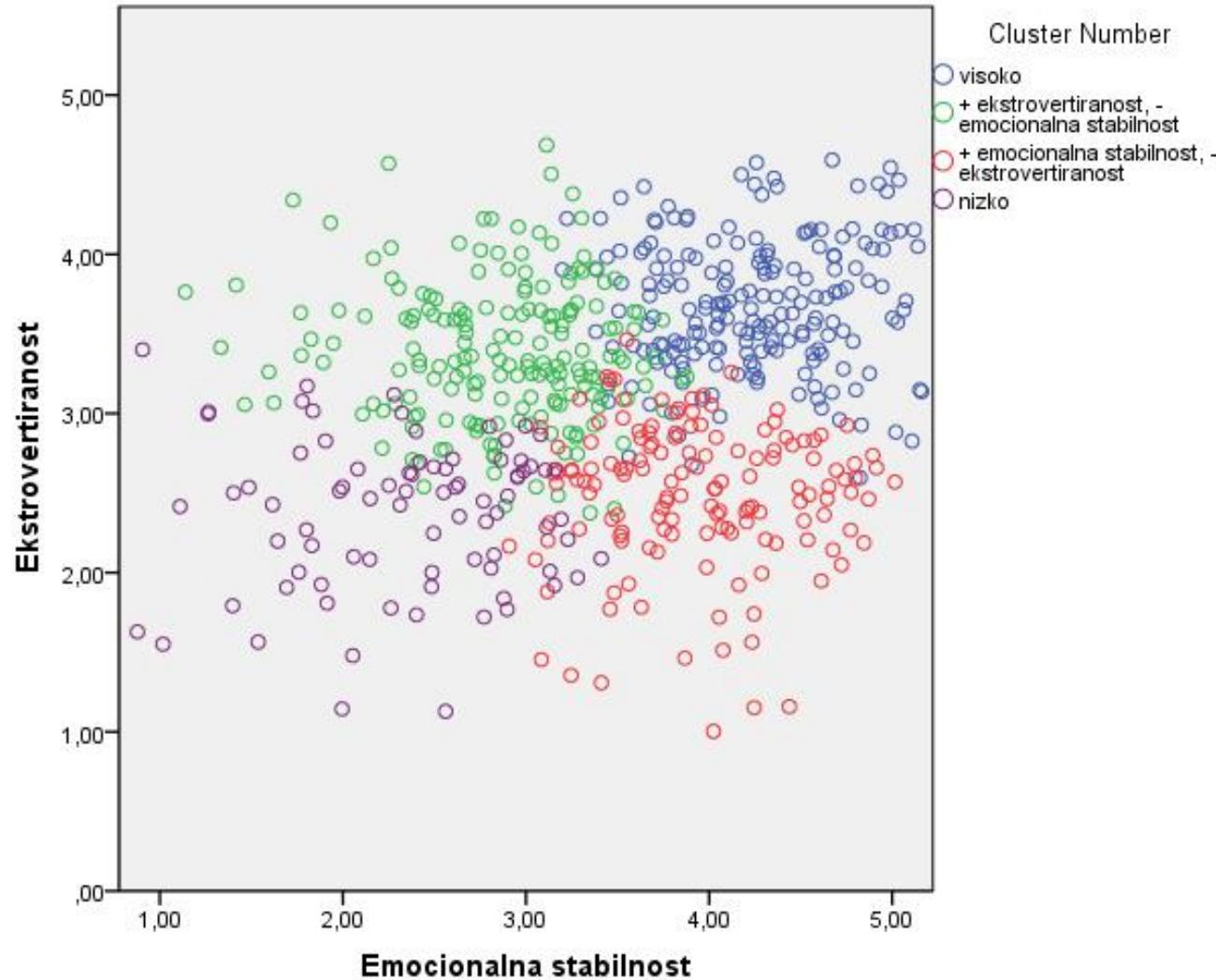
	N	Mean	Std. Deviation	Std. Error
1 visoko	196	7,06	2,914	,208
2 + ekstrovertiranost, - emocionalna stabilnost	187	7,16	3,401	,249
3 + emocionalna stabilnost, - ekstrovertiranost	145	6,51	2,754	,229
4 nizko	82	6,49	2,768	,306
Total	610	6,88	3,025	,122

## ANOVA

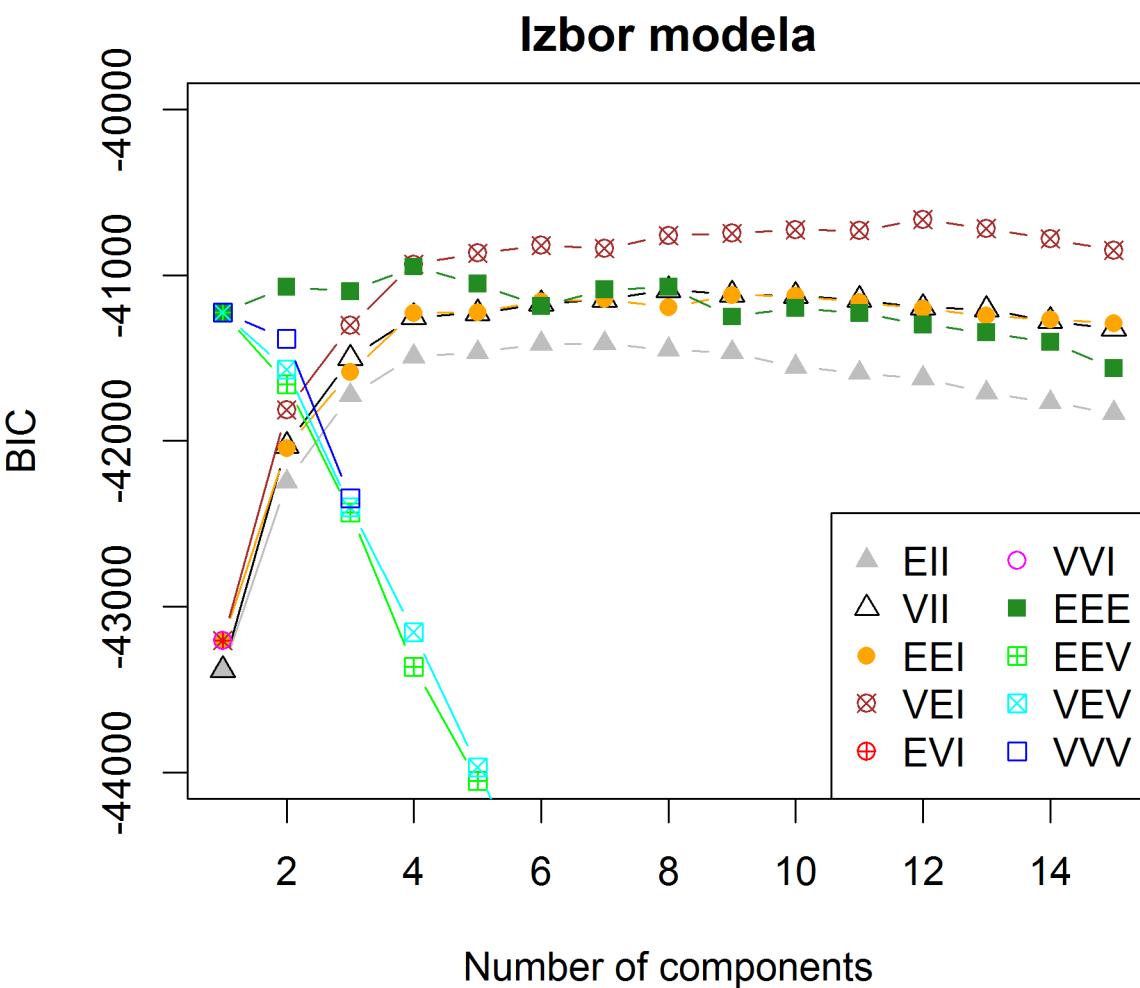
VEL\_OM

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	53,210	3	17,737	1,947	,121
Within Groups	5520,292	606	9,109		
Total	5573,502	609			

# Skupine (slučajni voditelji) v prostoru likertovih lestvic.



# Primer: Razvrščanje na podlagi modelov



- Najbolj se prilega model VEI (različno „razpršene“ okrogle skupine) z 12 komponentami
- Naraščanje BIC se ustali pri 4 komp.
- Od drugih modelov se najbolj prilega model EEE (Enake kov. matrike) s 4 komponentami

# Primer: Razvrščanje na podlagi modelov

---

Gaussian finite mixture model fitted by EM algorithm

---

Mclust VFI (diagonal, equal shape) model with 4 components:

log.likelihood	n	df	BIC	ICL
-20126.77	610	106	-40933.37	-41043.98

Clustering table:

1	2	3	4
74	140	194	202

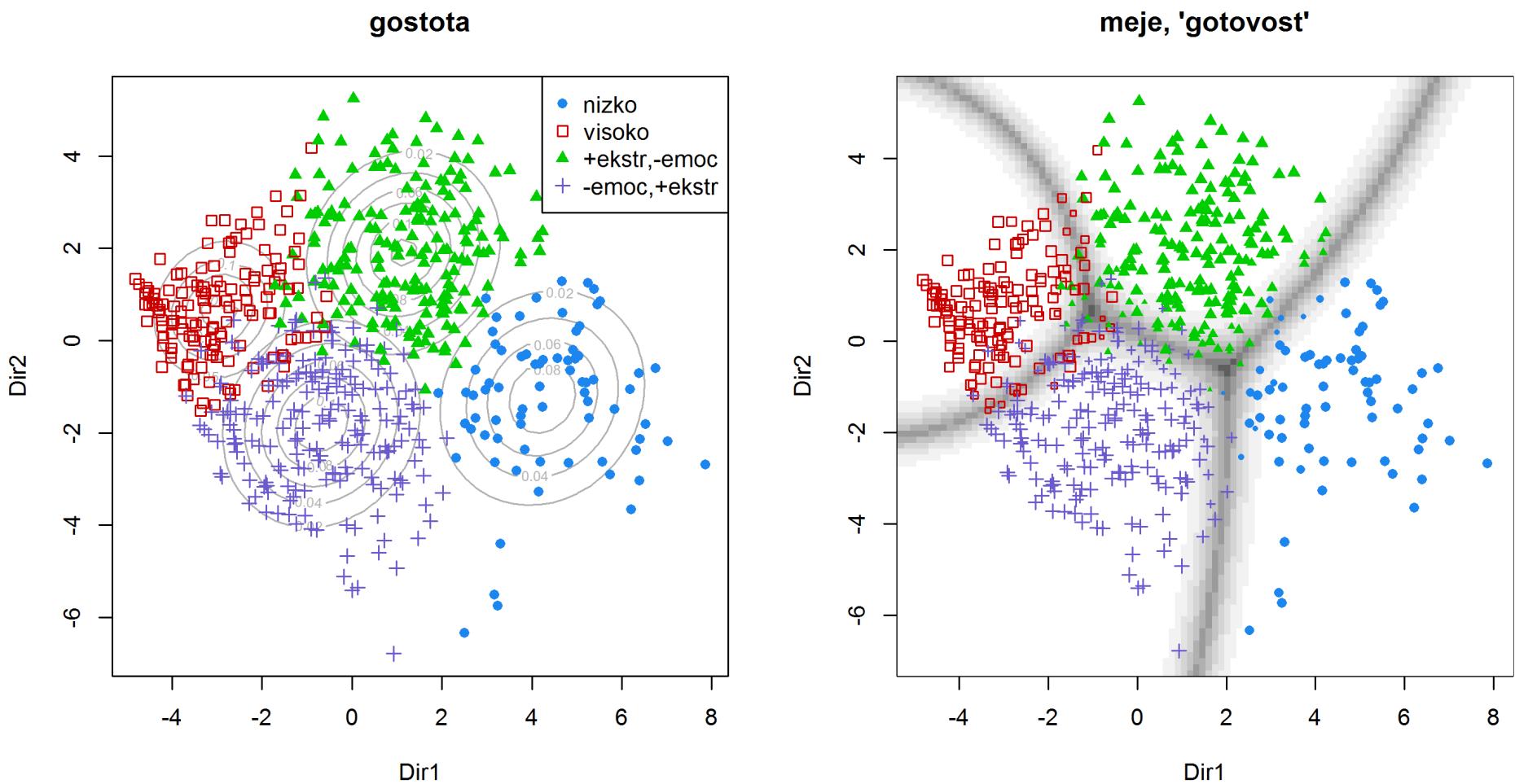
# Primer: Razvrščanje na podlagi modelov

Skupina	$\mu$ , originalni				$\mu$ , standardizirani			
	1	2	3	4	1	2	3	4
Deleži	0,13	0,23	0,32	0,33	0,12	0,23	0,32	0,33
EKSTA	2,48	3,8	3,63	2,73	-0,53	0,40	0,29	-0,35
EKSTB	3,05	4,19	3,78	3,03	-0,32	0,44	0,16	-0,33
EKSTE	2,91	4,69	4,4	3,69	-0,82	0,47	0,26	-0,26
EKSTH	2,57	3,7	3,59	2,87	-0,53	0,35	0,27	-0,30
EKSTIR	2,22	3,01	2,52	2,26	-0,20	0,35	0,01	-0,17
EKSTLR	2,45	4,64	4,3	2,96	-0,87	0,65	0,42	-0,51
EKSTNR	1,71	4,2	3,12	2,27	-0,75	0,81	0,14	-0,40
EKSTOR	2,49	4,56	4	3,14	-0,85	0,66	0,26	-0,37
EKSTP	3,31	4,74	4,56	4,22	-1,03	0,41	0,24	-0,11
EKSTRR	1,97	4,52	3,54	2,74	-0,89	0,82	0,16	-0,38
EMOCC	2,56	3,98	3,28	3,58	-0,62	0,37	-0,12	0,09
EMOCDR	1,77	4,47	2,16	4,24	-0,98	0,72	-0,73	0,57
EMOCF	2,66	4,78	4,45	4,17	-1,41	0,52	0,22	-0,03
EMOCGR	2,43	4,26	2,18	4,03	-0,57	0,64	-0,74	0,48
EMOCJR	1,95	4,25	2,02	3,99	-0,81	0,72	-0,77	0,55
EMOCKR	1,61	3,97	2,56	3,09	-0,83	0,65	-0,24	0,10
EMOCMR	3,05	4,07	3,07	3,75	-0,32	0,37	-0,31	0,16
EMOCQR	2,38	4,43	2,86	3,95	-0,75	0,60	-0,43	0,29
EMOCSR	2,53	4,73	3,76	4,36	-1,17	0,56	-0,20	0,26
EMOCTR	2,23	4,57	2,81	3,99	-0,90	0,73	-0,50	0,32

# Primer: Razvrščanje na podlagi modelov

Skupina	povprečja po org. sprem.				povprečja po stand. sprem.			
	1	2	3	4	1	2	3	4
Deleži	0,13	0,23	0,32	0,33	0,12	0,23	0,32	0,33
EKSTA	2,45	3,82	3,61	2,73	-0,55	0,42	0,27	-0,35
EKSTB	2,97	4,17	3,76	3,07	-0,37	0,42	0,15	-0,3
EKSTE	2,85	4,72	4,39	3,68	-0,87	0,5	0,25	-0,27
EKSTH	2,57	3,71	3,60	2,84	-0,53	0,36	0,27	-0,32
EKSTIR	2,27	3,01	2,49	2,26	-0,16	0,35	-0,01	-0,17
EKSTLR	2,38	4,64	4,34	2,93	-0,92	0,65	0,44	-0,54
EKSTNR	1,69	4,22	3,11	2,24	-0,77	0,83	0,13	-0,42
EKSTOR	2,49	4,57	3,98	3,12	-0,85	0,67	0,24	-0,38
EKSTP	3,24	4,74	4,58	4,21	-1,11	0,42	0,25	-0,13
EKSTRR	2,01	4,51	3,49	2,73	-0,86	0,82	0,13	-0,38
EMOCC	2,54	3,98	3,27	3,59	-0,63	0,37	-0,13	0,1
EMOCDR	1,72	4,47	2,11	4,28	-1,01	0,72	-0,76	0,6
EMOCF	2,58	4,79	4,44	4,17	-1,47	0,53	0,21	-0,03
EMOCGR	2,46	4,24	2,17	4,03	-0,55	0,62	-0,74	0,48
EMOCJR	1,89	4,21	2,02	4,02	-0,85	0,7	-0,77	0,57
EMOCKR	1,61	4,00	2,53	3,08	-0,83	0,67	-0,26	0,09
EMOCMR	3,15	4,04	3,06	3,74	-0,25	0,36	-0,32	0,15
EMOCQR	2,42	4,44	2,84	3,94	-0,73	0,61	-0,45	0,28
EMOCSR	2,49	4,73	3,76	4,35	-1,21	0,55	-0,2	0,25
EMOCTR	2,19	4,58	2,78	4,01	-0,93	0,73	-0,52	0,33

# Primer: Razvrščanje na podlagi modelov



# Primer iz publikacije (Ziherl et. al, 2006)

- Ziherl, P., Iglič, H., & Ferligoj, A. (2006). Research Groups' Social Capital: A Clustering Approach. *Metodološki Zvezki*, 3(2), 217–237.
- Cilj je poiskati skupine raziskovalnih skupin glede na značilnosti omrežij in jih primerjati z rezultati mladih raziskovalcev v njih.

# Primer iz publikacije (Ziherl et. al, 2006)

Spremenljivke:

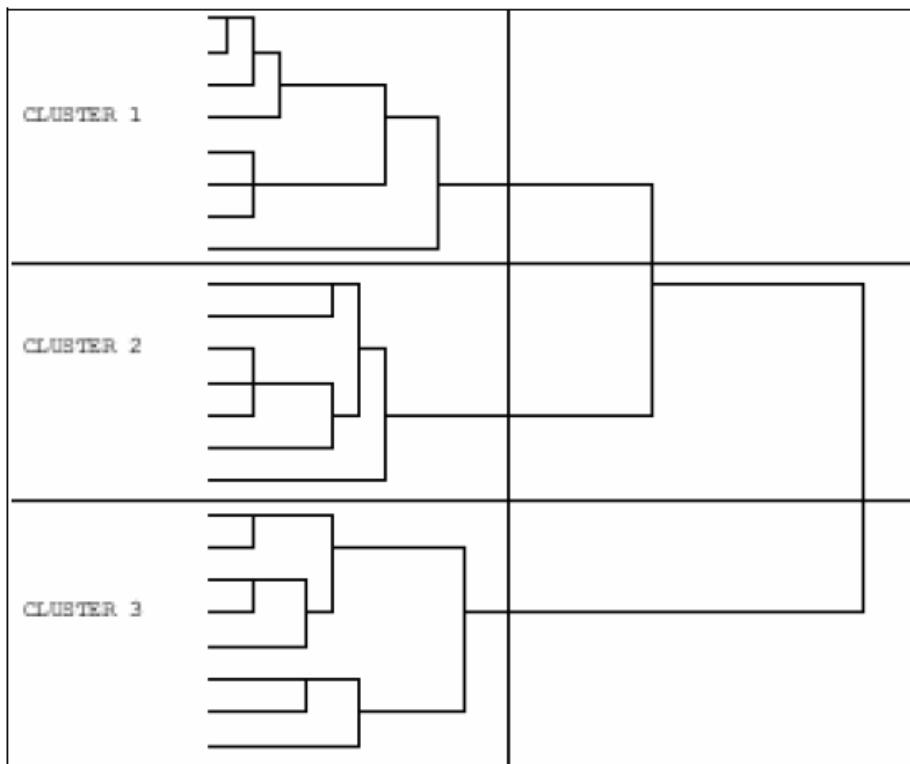
- Tie strength – povprečna moč povezave med MR-jem in člani skupine
- Cohesion - povprečna moč povezave med vsemi članci skupine
- Size – velikost skupine
- Others – Število ljudi, s katerimi MR sodeluje izven skupine
- Institutions – Število različnih institucij, s katerimi sodeluje
- Constraint – Burtova mera omejenosti

# Primer iz publikacije (Ziherl et. al, 2006)

Metoda:

- Vse spremenljivke so standardizirane (z-score)
- Evklidska razdalja (kvadrirana?)
- Wardova metoda hierarhičnega razvrščanja

# Primer iz publikacije (Ziherl et. al, 2006)



**Figure 1:** Dendrogram.

# Primer iz publikacije (Ziherl et. al, 2006)

**Table 2:** The characteristics of clusters.

Clusters		Tie strength	Cohesion	Size	Others	Institutions	Constraint
1 - Weak social capital cluster	Mean	1,67	2,36	5,5	0,5	1,88	0,73
	Std.dev.	1,08	1,05	1,6	0,75	0,64	0,13
2 - Bonding social capital cluster	Mean	3,66	4,47	4 0,6	2,57	1,71	0,84
	Std.dev.	1,97	0,86	9	1,51	0,75	0,12
3 - Bridging social capital cluster	Mean	2,69	3,07	5 1,9	4,13	3,5	0,55
	Std.dev.	0,83	1,02	1	2,99	0,75	0,09
Total	Mean	2,64	3,25	4 2,2	2,39	2,39	0,71
	Std.dev.	1,52	1,29	6	2,46	1,07	0,16

# Primer iz publikacije (Ziherl et. al, 2006)

**Table 3:** Average index of PhD students' performance in three clusters.

Clusters		Index of performance
1 – Weak social capital cluster	Mean	<b>6,63</b>
	Std.dev.	5,65
2 – Bonding social capital cluster	Mean	<b>9,29</b>
	Std.dev.	4,07
3 – Bridging social capital cluster	Mean	<b>22,13</b>
	Std.dev.	8,87
Total	Mean	<b>12,83</b>
	Std.dev.	9,44