LLL–Seminari u okviru TEMPUS projekta

Naziv projekta:	**511140 – TEMPUS – JPCR "Master programe in Applied Statistics - MAS"**
Broj projekta:	**511140**
Nosilac projekta:	**Departman za matematiku i informatiku, PMF Novi Sad**
Rukovodilac:	**Prof. dr Andreja Tepavčević**
Vreme trajanja:	**15.10.2010. – 14.10.2013.**
Finansiranje:	**Projekat finansira EU**

SEMINAR 3

METODE KLASTER ANALIZE

U MATEMATIČKOJ TAKSONOMIJI:

klaster metoda prostog povezivanja, klaster metoda kompleksnog povezivanja i metoda proseka

UVOD

Akademski stručnjaci i istraživači tržišta se često sreću sa situacijama koje su najbolje rešene definisanjem grupa homogenih objekata bilo da su firme, proizvodi, individue ili čak njihova ponašanja. Strateška rešenja bazirana na identifikaciji grupa unutar populacije, kao što je segmentacija i ciljni marketing ne bi bila moguća bez objektivne metodologije.

Ova ista potreba se sreće u drugim područjima, od fizičkih do društvenih nauka. U svim slučajevima, istraživači traže **prirodnu strukturu** izmedju observacija zasnovanu na višestrukim kriterijuma. Najčešće korišćena tehnika za ovu namenu je **klaster analiza**. Ona nastoji da **maximizira internu homogenost** i **eksternu heterogenost klastera**.

Važna osobina klaster analize je činjenica da **ona nije metoda strogog statističkog zaključivanja** gde se odabrani uzorak nužno smatra i reprezentativnim za određenu populaciju. **Klaster analiza je metoda kojom se određuju strukturalne karakteristike izmerenih svojstava na strogoj matematičkoj, ali ne i statističkoj utemeljenosti.** Prema tome, da bi rezultati klaster analize bili smisleni potrebno je utvrditi pretpostavke koje se odnose na reprezentativnost uzorka i multikolinearnost varijabli. Pouzdanost rezultata klaster analize zavisi od reprezentativnosti uzorka.

Termin klaster analiza (**prvi put je upotrebio Trion, 1939**) obuhvata niz različitih algoritama i metoda za grupisanje objekata sličnog tipa u odgovarajuće kategorije, a značajnija literatura iz ovog područja razvija se od šezdesetih godina. Brzi razvoj računara i temeljni značaj klasifikacije kao znanstvene procedure doprineli su popularnosti ove metode. Koristi se u različitim oblastima za kategorizaciju, odnosno klasifikaciju pojedinih jedinica analize (objekata ili ispitanika) obzirom na njihovu sličnost, odnosno različitost, prema nekim njihovim mernim obeležjima.

Opšte pitanje sa kojim se suočavaju istraživači u mnogim oblastima ispitivanja jeste kako organizovati posmatrane podatake u smisaone strukture, tj. kako razviti taksonomije. Drugim rečima klaster analiza je istraživačka tehnika za analizu podataka koja ima za cilj da sortira različite objekte u grupe tako da je stepen udruživanja između dva objekta maksimalan ako pripadaju istoj grupi i minimalan ako pripadaju različitoj. **Klaster analiza jednostavno otkriva strukture u podacima ne objašnjavajući zašto one postoje.**

Gotovo svakodnevno se susrećemo sa grupisanjem:

- nekolicina ljudi koja sede za istim stolom u restoranu mogu se svrstati u jednu grupu;
- biolozi moraju da organizuju različite vrste životinja pre nego što se između njih ustanove smislene razlike;
- prilikom segmentacije tržišta formiraju se klasteri potrošača u nekoj zemlji, pa se onda pravi poseban plan poslovnih aktivnosti za svaki klaster pojedinačno;
- u marketingu se klaster analiza koristi prilikom analize karakteristika proizvoda ili usluga, stavova kupaca, demografskih faktora itd.

Klaster analiza se može dobro iskoristiti za redukciju podataka. Ukoliko je, na primer, potrebno izvršiti testiranje novog proizvoda na tržištu po gradovima, naprave se klasteri sličnih gradova pa se iz svakog klastera odabere po jedan grad za testiranje, da se ne bi analizirali svi gradovi.

Pored toga, ako klaster analiza pokaže neko neočekivano grupisanje jedinica posmatranja, onda postoji verovatnoća da su pronađene određene relacije između jedinica posmatranja koje do tada nisu bile poznate i koje treba ispitati.

Vrlo je bitno znati da što je više varijabli uključeno u analizu i što su one više međusobno nezavisne, teže je pronaći odgovarajući model za grupisanje jedinica posmatranja.

1. OSNOVNI POJMOVI ALGEBRE ZA POTREBE KLASTER ANALIZE

1.1. Korespondencije i relacije

Neka su A i B neprazni skupovi. **Korespondencija** R iz skupa A u skup B je podskup proizvoda skupova A i B, tj. $R \subseteq A \times B$. Ako za elemente $a \in A$ i $b \in B$ važi da je $(a,b) \in R$, kaže se da su a i b u korespondenciji i to se zapisuje još i sa aRb. Dakle, korespondencija uspostavlja vezu nekih elemenata skupa A sa nekim elementima iz skupa B i ona se zadaje uređenim parovima elemenata koji su u vezi.

Primer 1. Neka je $A = \{p,q,r\}$ skup tri osobe i neka je $B = \{a,b,c\}$ skup nekih ljudskih osobina (karaktera). Neka je korespondencija R definisana na sledeći način:

Osoba x je u korespondenciji R sa osobinom y, ako x ima osobinu y.

Ako je R kao podskup skupa A×B dat sa:

$$R = \{(p, a), (q, a), (q, b), (q, c), (r, b), (r, c)\},$$

onda se ona obično prikazuje tablično:

R	a	b	c
p	o		
q	o	o	o
r		o	o

Tabela korespondencije R.

Korespondencija iz skupa A u taj isti skup zove se **binarna relacija** (kraće, **relacija**).

Dakle relacija ρ na skupu A je podskup skupa $A \times A$. Kako su i korespondencije i relacije po svojoj prirodi skupovi (u ovom slučaju uredenih parova), za njih važi sve ono što važi i za skupove. Na svakom skupu postoji i prazna relacija i ona ne uspostavlja odnos ni među koja dva elementa skupa.

Primer 2. Navodimo neke poznate relacije u matematici:

- Skup realnih brojeva je R i relacija $\leq$;
- Skup prirodnih brojeva N i na njemu se mogu posmatrati relacije $\leq$, $<$, $\geq$ i $>$, i druge.

1.2. Vektori i matrice

Vektor (realni, *n*-dimenzionalni vektor ili vektor dimenzije *n*) je element iz skupa R^n, gde je R skup realnih brojeva. Vektor je, dakle, uredena *n*-torka realnih brojeva, tj. vektor $a = (a_1,...,a_n)$.

Vektori se **sabiraju** po komponentama:

Ako **su** $a = (a_1,a_2,...,a_n)$ i $b = (b_1,b_2,..,b_n)$ dva vektora, tada je **zbir vektora** a i b takođe vektor koji se definiše sa:

$$a + b = (a_1 + b_1, a_2 + b_2,..., a_n + b_n).$$

Neka je α skalar i $a = (a_1, a_2,...,a_n)$ vektor. Tada se **množenje vektora skalarom** definiše sa:

$$\alpha\cdot(a_1, a_2,...,a_n) = (\alpha\cdot a_1, \alpha\cdot a_2,..., \alpha\cdot a_n).$$

Skalarni proizvod vektora $a = (a_1, a_2,...,a_n)$ i $b = (b_1, b_2,...,b_n)$ je skalar dat sa:

$$a\cdot b = a_1\cdot b_1 + a_2\cdot b_2 + \ldots + a_n\cdot b_n$$

Norma vektora $a = (a_1, a_2,...,a_n)$ je realan broj $||a||$ i

$$||a|| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}\,.$$

Euklidovo rastojanje između dva vektora $a = (a_1, a_2,.....,a_n)$ i $b = (b_1, b_2,...,b_n)$ definiše se sa:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}\,.$$

Matrica nad skupom A je pravougaona šema

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

gde su a_{ij} za $i \in \{1,\ldots,m\}$ i $j \in \{1,\ldots,n\}$ elementi iz skupa A. Elementi a_{ij} se zovu **elementi matrice.**

2. OSNOVNI POJMOVI IZ TAKSONOMIJE

Taksonomija predstavlja hijerarhijsku klasifikaciju pojmova, stvari, objekata, mesta, bića, događaja ili principa koji se klasifikuju.

Reč taksonomija izvodi se iz grčkih reči:

***taxis* = raspored** i ***nomos*= zakon.**

Taksonomija je grana biologije koja se bavi klasifikovanjem organizama, a prema Mayer-u „taksonomija je teorija i praksa klasifikacije organizama".

Pod taksonomijama u najširem smislu se podrazumeva izučavanje opštih principa naučne ili sistematske klasifikacije. Kada se kaže taksonomija ili sistematska klasifikacija, obično se misli na uređenu klasifikaciju biljnog i životinjskog sveta koja je urađena u skladu sa pretpostavljenim prirodnim vezama koje u njemu postoje.

Takson predstavlja taksonomsku grupu bilo kog nivoa koja se dovoljno izdvaja od ostalih tako da bi mogla da se usvoji kao određena kategorija. Da bi se podaci dobijeni registrovanjem i klasifikovanjem mogli koristiti, svakom taksonu daje se određeni naziv. Time se bavi deo sistematike koji se naziva **nomenklatura.**

Identifikacija je proces smeštanja jedinki u jednu od prethodno određenih grupa, tj. određivanje kojoj grupi ta jednika pripada.

U informatici taksonomije su inženjerski proizvod i predstavljaju klasifikaciju informatičkih entiteta u obliku hijerarhije u skladu sa pretpostavljenim vezama koje postoje između objekata u realnom svetu koje ovi informatički objekti predstavljaju.

Konstrukcija taksonomskih struktura se obično sastoji iz četri faze.

U prvoj fazi treba da se odabere objekat koji želimo da klasifikujemo. Pri tome uvek se misli na ranije klasifikacije nastalih tokom istorije, bez obzira da li taksonom ima nameru da ih odbaci i napravi neku svoju klasifikaciju. U **numeričkoj taksonomiji** objekti koji se klasifikuju nazivaju se **taksonomske jedinice**.

Druga faza se sastoji u izboru karaktera na kojima će se bazirati poređenje taksonomskih jedinica koje se posmatraju i odlučuju o važnosti pojedinih karaktera za poređenje. U okviru ove faze taksonom utvrđuje koja su stanja pojedinih karaktera **pleziomorfna (primitivna**), a koja su **apomorfna (izvedena)**, i onda se kod primitivnih karaktera vodi računa o dodeli **što većeg taksonomskog značaja**.

Treća faza se sastoji u utvrdjivanju sličnosti između pojedinih taksonomskih jedinica na osnovu izabranih karaktera.

Poslednja, **četvrta faza** je klasifikacija taksonomskih jedinica na osnovu utvrđene sličnosti.

2.1. Taksonomski karakteri

Važna faza u rešavanju problema klasifikacije neke grupe organizma ili objekata je izbor karaktera na kojima će se bazirati poređenje taksonomskih jedinica.

Karakter (obeležje) je bilo kakva osobina (svojstvo) koja se može razlikovati kod razlièitih taksonomskih jedinica, a **stanja** (realizacije) karaktera su moguće vrednosti koje karakteri uzimaju.

Prilikom izbora karaktera, i prilikom računanja sličnosti između taksonomskih jedinica preko tih karaktera, bitno je da se vodi računa i o taksonomskom značaju pojedninih karaktera.

Poklapanje karaktera sa većim taksonomskim značenjem kod taksonomskih jedinica daje veću informaciju o njihovoj sličnosti, nego ako se poklapaju karakteri sa manjim značajem.

Osnovni tipovi karaktera

Dva osnovna tipa karaktera su **kvalitativni** i **kvantitavni**.

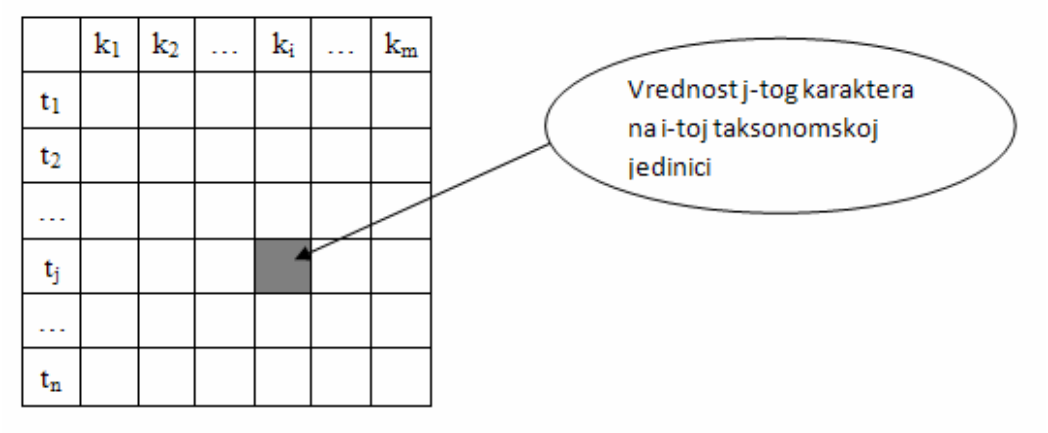
Najjednostavnija vrsta **kvalitativnih** karaktera su binarni karakteri. Karakteri sa samo jednim stanjem u grupi taksonomskih jedinica koja se posmatra očigledno nisu pogodni za posmatranje prilikom klasifikacije tih jedinca. Binarni karakteri imaju samo dva stanja „da“ i „ne“, što znači da binarni karakter predstavlja osobinu koju taksonomska jedinica ili ima, ili nema. Često se stanje prisustva binarnog karaktera obeležava sa 1, a stanja odsustva sa 0.

Kvalitativni karakteri mogu imati više stanja i različite forme predstavljaju stanja karaktera.

Kvantitativni (numerički) karakteri su oni koji uzimaju vrednosti koje se mere. Ako su svi karakteri koji se posmatraju numerički, realizacije karaktera se mogu prikazati u obliku vektora – **vektor realizacije za taksonomsku jedinicu.**

Predstavljanje karaktera u obliku pogodnom za dalju obradu

Najčešće se prikupljeni karakteri predstavljaju tabelarno. U zaglavlju tabele sa leve strane, od gore prema dole, se upisuju taksonomske jedinice koje se posmatraju, a u gornjem delu zaglavlja sa leva na desno se upisuju karakteri. Takva tablica je pogodna i za unošenje podataka u računar.



Tablica za predstavljanje karaktera

Ako podaci o nekom karakteru nisu dostupni za neku taksonomsku jedinicu, unese se posebne oznake, na primer ND (nije dostupno). Ista oznaka se koristi ako se neki par taksonomskih jedinica ne želi upoređivati u odnosu na taj karakter.

Moguće su **dve vrste transformacija prethodne tablice** u oblike pogodne za dalji rad sa njima:

1. Pretvaranje svih tipova karaktera u binarni oblik

2. Standardizacija kvantitativnih mera.

Pretvaranje drugih tipova karaktera u binarni oblik nije komplikovano, a metode računanja sličnosti između taksonomskih jedinica kada su svi karakteri pretvoreni u binarni oblik su jednostavnije od ostalih.

Nedostatak je to da pretvaranje u binarni oblik obično dovodi do gubitka informacije, i time eventualno do lošije procene siičnosti taksonomskih jedinica.

Kvalitativni karakteri sa više stanja i kvantitativni karakteri se mogu kodirati tako da i oni dobiju binarni oblik.

Standardizacija (ujednačavanje) kvantitativnih mera

Jedna od ovakvih transformacija je **rangiranje**. Mera x se pretvara u rangiranu meru x' sledećom formulom:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

gde su $x_{\min}$, odnosno $x_{\max}$ minimalna, odnosno maksimalna vrednost karaktera. Rangirana mera ima uvek vrednost između 0 i 1 (zato što je $x - x_{\min}$ uvek manje od $x_{\max} - x_{\min}$).

Najčešće korišćeni postupak u praksi je postupak **standardizacije.** Ako sa $\overline{x}$ označimo aritmrtičku sredinu, a sa s standardno odstupanje karaktera na grupi taksonomskih jedinica koja se posmatra, tada standardizovani karakter dobijamo sledećom transformacijom:

$$\frac{x_i - \overline{x}}{s},$$

gde je x_i realizacija karaktera x na i-toj taksonomskoj jedinici.

Transformacija tablice u zavisnosti od taksonomskog značaja karaktera.

To se postiže na jednostavan način, tako što se umesto svake kolone koja odgovara karakteru sa taksonomskim značajem k, gde je $k > 1$, tablici doda k istih takvih taksonomskog značaja 1.

2.2. Mere sličnosti i različitosti taksonomskih jedinica

Ovde će pažnja biti ograničena na definisanje sličnosti između dva objekta koristeći neke poznate karakteristike objekata.
Mere sličnosti i različitosti su međusobno povezane, tako što se od svake mere sličnosti na odgovarajući način može dobiti jedna mera različitosti, i obratno.

Postoji više različitih načina za određivanje mere sličnosti. Većina njih uzima vrednost iz intervala [0,1].
Mera različitosti je numerička vrednost koja pokazuje koliko su posmatrani objekti različiti i kod većine mera različitosti minimalna vrednost je 0 i ona pokazuje da ne postoji različitost između posmatranih objekata.

Ne postoji opšte pravilo za odlučivanje koja od metoda za računanje mere sličnosti, odnosno različitosti, je najbolja za primenu u konkretnom slučaju. **Izbor metode je veoma važan, zato što često različite metode daju i različite rezultate.**

Mere sličnosti taksonomskih jedinica kod kojih su svi karakteri predstavljani u binarnom obliku

Najjednostavnije, i možda najviše korišćene, su mere sličnosti za binarne karaktere. One se koriste kada se svi karakteri koji se uzimaju u obzir za određivanje sličnosti između dve taksonomske jedinice ili sami binarni, ili su pretvoreni u binarne karaktere (kodirani sa 0 i 1).

Ako se smatra da su svi karakteri koji se posmatraju ravnopravni za utvrđivanje mere sličnosti između dve taksonomske jedinice t_1 i t_2, onda se odgovarajuća **stanja karaktera za t_1 i t_2 mogu predstaviti sledećom tablicom**:

	0	1
0	m_{00}	m_{01}
1	m_{10}	m_{11}

koja odgovara sledećoj matrici sličnosti 2×2: $\begin{bmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{bmatrix}$.

Ovde je m_{ij} broj karaktera kod kojih taksonomska jedinica t_1 ima vrednost i, a t_2 ima vrednost j. Dakle, m_{00} je broj karaktera koji uzimaju vrednost 0 za obe taksonomske jedinice, m_{01} je broj karaktera koji uzima vrednost 0 za t_1, a 1 za t_2, itd. Prema tome, $m_{00}+m_{11}$ je broj karaktera koji su isti za obe taksonomske jedinice, a $m_{01}+m_{10}$ je broj karaktera koji su različiti za taksonomske jedinice t_2 i t_2. Ukupan broj karaktera koji se posmatra je $M=m_{00}+m_{11}+m_{01}+m_{10}$.

Najpoznatiji su **koeficijent prostog poklapanja** (Sokal & Michener, 1958.) i **Žarakov koeficijent** (Jaccard, 1908.).

Koeficijent prostog poklapanja za taksonomske jedinice t_1 i t_2 je:

$$K_{t_1,t_2} = \frac{m_{00} + m_{11}}{M}.$$

Žakarov koeficijent se definiše sa:

$$J_{t_1,t_2} = \frac{m_{11}}{m_{11} + m_{01} + m_{10}}.$$

Ovi koeficijenti često nisu saglasni i, koeficijent prostog poklapanja je uvek veći ili jednak od Žakarovog koeficijenta.

Mera sličnosti za taksonomske jedinice kod koje su svi karakteri kvalitativni

Ako su podaci za dve taksonomske jedinice predstavljeni smo u obliku kvalitativnih karaktera, najjednostavniji koeficijent sličnosti je onaj koji predstavlja uopštenje koeficijenta prostog poklapanja za binarne karaktere.

On se zove **koeficijent prostog poklapanja za kvalitativne karaktere** i računa se sledećom formulom:

$$K_{t_1,t_2} = \frac{P}{M},$$

gde je P broj karaktera kod kojih se stanje taksonomskih jedinica t_1 i t_2 poklapa, a M ukupan broj posmatranih karaktera.

Mera sličnosti za taksonomske jedinice kod kojih su svi karakteri kvantitativni

I u ovom slučaju se svi karakteri mogu pretvoriti u binarni oblik i zatim računati koeficijent prostog poklapanja ili Žakarov koeficijent.

Jedna od najčešće korišćenih mera sličnosti za ovaj tip karaktera se računa po sledećoj formuli:

$$S(t_1,t_2)=\frac{\sum_{i=1}^{M}\left(1-\frac{|t_1(i)-t_2(i)|}{R_i}\right)}{M},$$

gde je M ukupan broj karaktera, $t_1(i)$ odnosno $t_2(i)$ su vrednosti karaktera i kod taksonomskih jedinica t_1 odnosno t_2, redom, a R_i razlika između najveće i najmanje vrednosti koju uzima karakter i u grupi taksonomskih jedinica koja se posmatra.

Mere sličnosti za mešane tipove karaktera

Moguće je sve vrednosti karaktera prebaciti u binarni oblik pa raditi na već opisani način.

Koeficijent sličnosti može se računati i direktno.

Gouerov (Gower) koeficijent sličnosti (uveden 1971.) je veoma pogodan u slučajevima kada se taksonomske jedinice predstavljaju sa više vrsta karaktera. Ako su t_1 i t_2 taksonomske jedinice, a n ukupan broj karaktera, Gouerov koeficijent sličnosti za njih se obeležava sa $G(t_1, t_2)$ i računa se sledećom formulom:

$$G(t_1,t_2)=\frac{\sum_{i=1}^{n} s_i(t_1,t_2)}{\sum_{i=1}^{n} w_i(t_1,t_2)},$$

gde za svaki karakter i, veličine s_i i w_i zavise od tipa karaktera i, na sledeći način:

– *i* je binarni karakter:

- Vrednost $w_i(t_1,t_2)$ je jednaka nuli u slučajevima kada vrednost za karakter *i* nije dostupna (ND) bar za jednu od taksonomskih jedinica t_1 ili t_2, ili ako karakter *i* uzima vrednost 0 za obe taksonomske jedinice.

- Vrednost $w_i(t_1,t_2)$ je jednaka 1 kada je karakter dostupan za obe taksonomske jedinice, a bar kod jedne od njih uzima vrednost 1.

- Vrednost $s_i(t_1,t_2)$je jednaka 1 samo u slučaju kada karakter *i* ima vrednost 1 za obe taksonomske jedinice, a u svim ostalim slučajevima je $s_i(t_1,t_2)=0$.

– *i* je kvalitativni karakter sa dva ili više stanja:

- $w_i(t_1,t_2)$ je nula u slučaju kada karakter nije dostupan za neku od taksonomskih jedinica ili ako se taksonomske jedinice ne porede u odnosu na taj karakter.
- U svim ostalim slučajevima je $w_i(t_1,t_2)=1$.
- $s_i(t_1,t_2)$ je jedan samo ako taksonomske jedinice imaju istu vrednost karaktera i, a nula u svim ostalim slučajevima.

– *i* je kvantitativni karakter:

- $w_i(t_1,t_2)$ se određuje isto kao u prethodnom slučaju.
- $s_i(t_1,t_2)$ se računa sledećom formulom:

$$s_i(t_1,t_2)=1-\frac{\left|t_1(i)-t_2(i)\right|}{R_i},$$

gde su $t_1(i)$ odnosno $t_2(i)$ vrednosti karaktera *i* kod taksonomskih jedinica t_1 odnosno t_2, redom, a R_i razlika između najveće i najmanje vrednosti koju uzima karakter *i* u grupi taksonomskih jedinica koja se posmatra.

Mere različitosti i rastojanja taksonomskih jedinica

Mere sličnosti i mere različitosti su međusobno povezane, i mogu se izračunati jedne iz drugih. Za svaku meru sličnosti (ako uzima vrednosti između 0 i 1) dobija se mera različitosti kada se od jedan oduzme ta mera sličnosti.

Postoje i mere različitosti koje ne odgovaraju nekoj od već definisanih mera sličnosti. Posebno su značajne one od njih koje zadovoljavaju matematičke uslove koji će se navesti u nastavku. **Takve mere se obično nazivaju mere rastojanja ili metrike.**

Ako se za taksonomske jedinice t_1 i t_2 mera odstojanja obeleži sa d_{t_1,t_2}, tada ona ispunjava sledećih pet uslova:

1. uslov: nenegativnost

$$d_{t_1,t_2} \geq 0,$$

mera različitosti uvek mora biti negativan broj.

2. uslov: simetričnost

$$d_{t_1,t_2} = d_{t_2,t_1},$$

mera rastojanja taksonomske jedinice t_1 od taksonomske jedinice t_2 jednaka je meri rastojanja taksonomske jedinice t_2 od taksonomske jedinice t_1.

3. uslov: ako je $t_1 \neq t_2$ onda je $d_{t_1,t_2} \neq 0$,

ako dve taksonomske jedinice imaju bar jedan karakter u kojem se razlikuju, tada je mera rstojanja sigurno različita od 0.

4. uslov: $d_{t_1,t_1} = 0.$

5. uslov: nejednakost trougla

Ako su t_1, t_2 i t_3 tri taksonomske jedinice, onda zanjihova rastojanja važi sledeća nejednakost:

$$d_{t_1,t_3} \leq d_{t_1,t_2} + d_{t_2,t_3}.$$

U nastavku će se predstaviti nekoliko metoda računanja rastojanja između taksonomskih jedinica.

– **Euklidovo rastojanje** i to sledeći slučajevi:

(I) Euklidovo rastojanje primenjeno direktno na izmerene podatke;

(II) Prosečno (prilagođeno) Euklidovo rastojanje;

(III) Euklidovo rastojanje primenjeno na standardizovane podatke;

(IV) Podela keficijenta Euklidovog rastojanja u dva dela: koeficijent veličine i koeficijent oblika.

– **Apsolutna (blok) metrika (rastojanje).**

– **Rastojanje Minkovskog.**

– **Koeficijenti rastojanja računati preko koeficijenta sličnosti.**

Euklidovo rastojanje

(I) Euklidovo rastojanje primenjeno direktno na izmenjene podatke.

Najpoznatija mera rastojanja je Euklidova mera (ili metrika). Euklidovo rastojanje između dve taksonomske jedinice t_a i t_b, ako se ono računa preko n različitih kvantitativnih karaktera koji su predstavljeni preko vektora na sledeći način: $a=(a_1,a_2,...,a_n)$ i $b=(b_1,b_2,...,b_n)$ je

$$d(t_a,t_b)=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+...+(a_n-b_n)^2}\,.$$

Euklidovo rastojanje se najčešće koristi zato što deluje najprirodnije kada se vektori predstave geometrijski. Međutim, problem kod Euklidovog rastojanja je u tome što se različiti karakteri mere različitim jedinicama, kada se promeni jedinica mere jednog od karaktera, mogu se dobiti sasvim drugačiji rezultati.

(II) Prosečno (prilagođeno) Euklidovo rastojanje

Ako je $d(t_a,t_b)$ Euklidovo rastojanje za taksonomske jedinice t_a i t_b, tada se prosečno (prilagođeno) Euklidovo rastojanje računa sledećom formulom:

$$\overline{d(t_a,t_b)}=\sqrt{\frac{(d(t_a,t_b))^2}{n}},$$

gde je n ukupan broj karaktera koji se posmatra.

(III) Euklidovo rastojanje primenjeno na standardizovane podatke

Kao što je napomenuto u slučaju (I), izbor jedinične mere za merenje pojedinog karaktera ima jak uticaj na Euklidovu meru rastojanja. Da bi se svim karakterima dala ista važnost neophodno je izvršiti standardizaciju svakog karaktera. Neka su poznate realizacije karaktera k na m taksonomskih jedinica. Ako sa $\bar{k}$ označimo aritmetičku sredinu, a sa s standardno odstupanje karaktera k, tada standardizovani karakter dobijamo sledećom transformacijom:

$$\frac{k_i-\bar{k}}{s},$$

gde je k_i realizacija karaktera k na i-toj taksonomskoj jedinici.

(IV) Podela koeficijenata Euklidovog rastojanja u dva dela:
koeficijent veličine i *koeficijent oblika*.

Koeficijent euklidovog rastojanja može se podeliti na dva dela: koeficijent oblika (obeležen sa C_Z^2) i koeficijent veličine (obeležen sa C_Q^2), na sledeći način:

$$(d(t_a,t_b))^2 = (n-1)\cdot C_Z^2 + n\cdot C_Q^2,$$

gde je $d(t_a,t_b)$ Euklidovo rastojanje između taksonomskih jedinica t_a i t_b.

Koeficijent oblika C_Z^2 i veličine C_Q^2 računaju se na sledeći način:

$$C_Z^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(a_i-b_i)^2 - \frac{1}{n(n-1)}(\sum_{i=1}^{n}(a_i-b_i))^2,$$

$$C_Q^2 = \frac{1}{n^2}(\sum_{i=1}^{n}(a_i-b_i))^2.$$

Apsolutna metrika

Apsolutna (blok, Manhattan) metrika (rastojanje) je još jedna mera rastojanja koja se koristi u posebnoj vrsti problema i za taksonomske jedinice t_a i t_b računa se formulom:

$$d(t_a, t_b) = \sum_{i=1}^{n} |a_i - b_i|.$$

Kada se ova mera podeli brojem karaktera, dobija se druga mera koja se zove **srednja razlika karaktera**:

$$\overline{d(t_a, t_b)} = \frac{\sum_{i=1}^{n} |a_i - b_i|}{n}.$$

Rastojanje Minkovskog

Rastojanje Minkovskog između dve taksonomske jedinice t_a i t_b dato je formulom:

$$d(t_a,t_b)=\left(\left|a_1-b_1\right|^r+\left|a_2-b_2\right|^r+\cdots+\left|a_n-b_n\right|^r\right)^{\frac{1}{r}}=\left(\sum_{i=1}^{n}\left|a_i-b_i\right|^r\right)^{\frac{1}{r}},$$

gde je r parameter (obično ceo broj).

Euklidovo rastojanje se dobija ovom formulom za $r=2$, a rastojanje apsolutne metrike za $r=1$, pa su to specijalni slučajevi rastojanja Minkovskog.

Koeficijenti rastojanja računati preko koeficijenata sličnosti

Ako je $C(t_1,t_2)$ bilo koji od koeficijenata sličnosti između taksonomskih jedinica t_1 i t_2 (definisanih u prethodnom delu), tada sledeće formule daju varijante za računanje koeficijenta različitosti:

1. način:

$$R_1(t_1,t_2)=1-C(t_1,t_2).$$

2. način:

$$R_2(t_1,t_2)=\sqrt{1-C(t_1,t_2)}.$$

3. KLASTER ANALIZA

Klaster analiza (razvrstavanje) je postupak podele skupa različitih objekata u grupe (klastere), pri čemu se vodi računa o tome da su objekti u dobijenim grupama međusobno što sličniji, i što više različiti od objekata u ostalim grupama.

Po ovim kriterijumima dobijene grupe se nazivaju **klasteri**. U većini metoda klaster analize krajnji rezultat je jedna **particija (podela) skupa** taksonomskih jedinica.

Termin klaster dolazi od engl. reči ***cluster*** (skupina „istovrsnih stvari“, grozd, skupiti u grupu). Klaster analiza klasifikuje objekte (ispitanike, proizvode ili druge objekte) tako da je svaki objekat veoma sličan drugima u klasteru uz poštovanje nekog unapred određenog kriterijuma selekcije.

Klaster analiza se naziva i

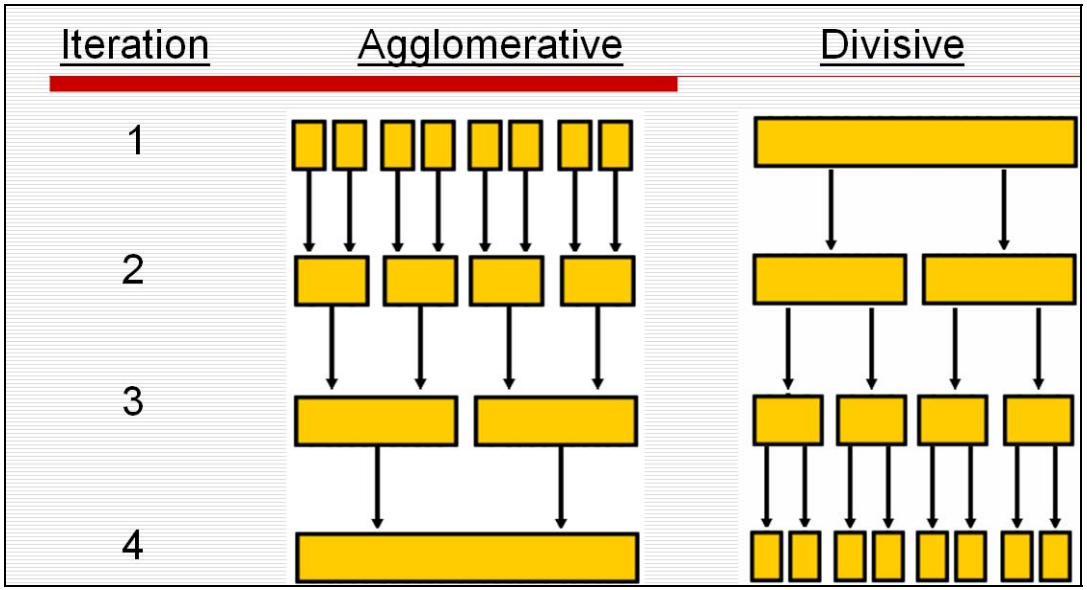
- Q analiza,
- Tipologija gradnje,
- Klasifikacijska analiza i
- Numerička taksonomija.

Ova raznovrsnost u nazivima je zbog korišćenja klaster metode u različitim disciplinama kao što su psihologija, biologija, sociologija, ekonomija. Uprkos različitim nazivima u zavisnosti od discipline, **svi metodi imaju zajedničku dimenziju: klasifikacija u skladu sa prirodnim vezama**.

Klaster analiza je uporediva sa faktor analizom u cilju procenjivanja strukture. Klaster analiza se razlikuje od faktor analize po tome što klaster analiza grupiše objekte, dok je faktor analiza primarno brine o grupisanju varijabli.

Metode klaster analize u matematičkoj taksonomiji mogu se podeliti u dve osnovne grupe:

- **aglomerativne metode** – metode koje grupišu taksonomske jedinice u grupe (klastere) po srodnim osobinama;
- **divizivne metode** – metode koje razbijaju skup taksonomskih jedinica na više grupa (klastera).



Slika 1. Aglomerativne i divizivne metode

U odnosu na karaktere koji se uzimaju u obzir prilikom analize, metode klaster analize se dele na:

- **politetičke** – metode koje uzimaju u obzir sve karaktere;
- **monotetičke** – klasifikacija dobijena ovom metodom se bazira na prisustvu ili odsustvu samo jednog karaktera.

Metode klaster analize se još mogu podeliti i na:

- **hijerarhijske** – metode koje daju niz sukcesivnih particija skupa - podela na klastere, koji se obično prikazuje specijalnim dijagramom - dendrogramom.
- **nehijerarhijske** – metode koje daju samo jednu, optimalnu podelu na klastere, čiji broj može ili ne mora biti unapred zadat.

Klaster analiza daje rezultate raznim algoritmima koji se razlikuju značajno u njihovoj ideji šta predstavlja klaster i kako efikasno da ih pronađemo.

Hijerarhijski metod kao krajnji rezultat ima dendrogram. To je grafički prikaz klastera (grupa) u obliku stabla povezivanja. Prvo se vrše izračunavanja udaljenosti svih jedinica međusobno, a zatim se grupe formiraju putem tehnika spajanja ili razdvajanja.

Nehijerarhijski metod vrši raščlanjivanje tako da jedinice mogu da se kreću iz jedne u drugu grupu u različitim fazama analize. Postoji mnogo varijacija u primeni ove tehnike, ali suština je u tome da se prvo pronađe tačka grupisanja oko koje se nalaze jedinice, na više ili manje proizvoljan način, a zatim se izračunavaju nove tačke grupisanja na osnovu prosečne vrednosti jedinica. Jedinica posmatranja se tada pomera iz jedne u drugu grupu ukoliko je bliža novoizračunatoj tački grupisanja. Proces se odvija iterativno, sve do postizanja stabilnosti.

Nehijerarhijske procedure klasterovanja se često pominju kao ***k*-grupisanje** i oni obično koriste jedan od sledeća dva pristupa dodeljivanja taksonomskih jedinica u jedan od klastera:

Paralelna metoda – ova metoda vrši selekciju u odnosu na nekoliko klaster centara istovremeno i dodeljuje objekte na osnovu praga udaljenosti od najbližeg centra. Kako se proces razvija, prag udaljenosti može biti prilagođen tako da se uključe manje ili više objekata u klastere.

Optimizacija – ovaj metod je sličan prethodnom pristupu stim to što dozvoljava ponavljanje postupka. Ako u tom postupku objekat postaje bliži drugom klasteru kome nije prvobitno dodeljen, tada ih optimizirajući postupak prebacuje sličnijem klasteru.

Nehijerarhijske procedure su dostupne velikom broju računarskih programa, uključujući i sve glavne statističke pakete.

Glavni problem sa kojim su suočeni svi nehijerarhijski postupci klasterovanja je kako odrediti broj klastera. Određivanje početnog klastera može rešiti ovaj problem.

Neke važne napomene o klaster analizi

Prvi oblici klaster analize javljaju se početkom prošlog veka, ali se značajnija literatura iz ovog područja razvija se od šezdesetih godina. Psiholozi je ponekad nazivaju "siromašnom faktorskom analizom".

Važne napomene vezane za korišćenje klaster analize:

1. Većina metoda klaster analize predstavlja relativno jednostavne algoritme, pa nemaju značajniju podršku u standardnom statističkom korišćenju (npr. određivanju značajnosti).
2. Pojedine metode razvijene su i korisne u okviru pojedinih naučnih disciplina , dok u drugima nisu od većeg značaja.
3. Različite metode klasterizacije mogu, a često i dovode, do različitih različitih konačnih rešenja.

Rezultat klaster analize uvek predstavlja klasifikacija objekata u neke grupe, što može dovesti do različitih rešenja. Jedan od važnih kriterijuma može biti i „psihološka“ smislenost dobijene solucije.

Neke od važnih odluka koje treba doneti pri sprovođenju klaster analiza su:

1) Izbor uzorka kojeg ćemo podvrgnuti klaster analizi;

2) Odrediti skup relevantnih varijabli koje će reprezentovati obeležja objekata (entiteta);

3) Odrediti transformaciju originalnih podataka;

4) Odrediti metodu za određivanje udaljenosti / sličnosti između objekata (entiteta);

5) Odrediti metodu koju ćemo koristiti za povezivanje objekata u klastere;

6) Ocena validnosti dobijenih rezultata.

Uz većinu ovih odluka vezuje se izbor prikladnog statističkog algoritma, odnosno tehnike.

Šta nije klaster analiza?

- ***Klasifikacija pod nadzorom***
  Postoje informacije o oznakama klasa – klasifikacija,
- ***Jednostavna podela*** – Podela studenata po prvom slovu prezimena,
- ***Rezultat ankete –*** Grupisanje je rezultat spoljašnje specifikacije, i drugo.

U marketingu, klaster analiza se koristi za:

- Podelu tržišta na segmente i utvrđivanje ciljnih tržišta,
- Pozicioniranje proizvoda i razvoj novog proizvoda,
- Odabir ispitivanja tržišta.

Ciljevi klaster analize

Pri formiranju homogenih grupa, istraživač može postići bilo koji od sledeća tri cilja:

1. ***Taksonomija opisa***. Najpoznatiji tradicionalini način korišćenja klaster analize je u istraživačke svrhe i za formiranje jednog taksonoma. Ali klaster analiza može takođe generisati hipoteze koje se odnose na strukturu objekata, a može se koristiti i za potvrdu nečega već ustanovljenog.
2. ***Pojednostavljenje podataka.*** U toku izvođenja procesa taksonomije, klaster analiza takođe postiže pojednostavljen način posmatranja. Sa definisanom strukturom, zapažanja mogu biti grupisana u cilju daljih analiza.
3. ***Identifikacija odnosa***. Sa definisanim klasterima i osnovnom strukturom podataka u njima, istraživač objašnjava odnos između posmatranja koje nije bilo moguće sa individualnim posmatranjem. Zato, klaster analiza prikazuje odnos, sličnosti ili razlike koje prehodne analize nisu objavile.

Centralna tema klaster analize

U klaster analizi, *koncept slučajne promenljive* je ponovo centralna tema , ali na potpuno drugačiji način od drugih multivarijacionih tehnika. Fokus klaster analize je na poređenju objekata zasnovanih na slučajnoj promenljivi, a ne na proceni same slučajne promenljive. Ova definicija slučajne varijable od strane istraživača je kritičan korak u klaster analizi.

Ipak, uz pogodnosti klaster analize idu i neke opomene. Klaster analiza može biti okarakterisana kao opisna, ateoretična i noninferentna. **Klaster analiza nema statističku osnovu kod kojih se mogu izvući statistička zaključivanja iz uzorka do populacije i korišćena je prvenstveno kao tehnika istraživanja.** Rešenja nisu jedinstvena i **istraživač mora voditi računa u proceni uticaja svake odluke uključene u izvođenje klaster analize.**

Određivanje zadovoljavajućeg broja klastera

Problem koji zbunjuje istraživače klaster analize je određivanje konačnog broja obrazovanih klastera (poznato kao **stoping pravilo**).

Jedna vrsta stoping pravila je sledeća. Kada usledi jedan jači skok tada istraživači klaster rešenja pribegavaju logici da bi zaključili šta je uzrok znatnog pada u sličnostima. Ova stoping praila pokazuju jednu prilično tačnu odluku u empirijskim studijama.

U praksi se radi tako što se uzme jedan broj klaster rešenja (npr. 2,3,4) i tada donese odluka, sa alternativnim rešenjima, koristeći apriori kriterijume i praktičnu ocenu, zdrav razum ili teorijske ocene.

U društvenim naukama dominiraju dva pristupa određivanju broja klastera: heuristički pristup i formalni testovi. Prvi pristup je najčešći, a odnosi se na subjektivno postavljanje granice na dendrogramu dobijenom hijerahijskom klasterizacijom. Osnovni kriterijum jeste smislenost ili interpretabilnost dobijenog rešenja.

Drugi način, podjednako subjektivan jeste analiza koeficijenata (koeficijenti fuzije) koji pokazuju sličnosti među klasterima pri sukcesivnom spajanju klastera. Naglo opadanje (ili povećanje vrednosti kod mera udaljenosti) ukazuje na manju povezanost među klasterima koji se spajaju. Nagli skok ukazuje na spajanje dva relativno različita klastera.

Izbor uzorka nad kojim se vrši klasterizacija

Izbor uzorka objekata, ispitanika, odnosno jedinica u znatnoj meri determiniše način grupisanja objekata. Istraživač retko ima uvid u populaciju koju koristi u klaster analizi. Obično se dobije uzorak i klasteri su izvedeni u nadi da predstavljaju strukturu populacije. **Istraživač mora biti siguran da je dobijeni uzorak stvarno reprezentativan.**

Izbor varijabli koje će reprezentovati obeležja objekata

Izbor varijabila koji će biti uključeni u klaster mora biti u skadu sa teoretskim i konceptualnim objašnjenjem, kao i sa praktičnim razmatranjem. Klaster analiza mora imati obrazloženje za izabrane varijabile.

Bilo da su razlozi zasnovani na eksplicitnoj teoriji, prethodnim istraživanjima ili pogađanju, istraživač mora shvatiti važnost uključivanja jedino onih promenljivih koje:

1) **karakterišu klasterizovane objekte** i
2) **posebno se odnose na ciljeve klaster analize.**

Tehnika klaster analize ne razlikuje relevantne od nerelavantnih varijabile. Uključivanje jedne nerelevantnie varijabile značajno utiče na rezultate.

Postoje nekoliko mera koje koristimo tokom klasterovanja objekata :

- **mera sličnosti,**
- **mera korelacije,**
- **mera udaljenosti** i
- **mera udruživanja.**

Mere sličnosti

Koncept sličnosti je fundamentalan u klaster analizi. Karakteristike su kombinovane unutar kalkulisanih mera sličnosti za sve parove objekata. Na taj način bilo koji objekat može biti poređen sa drugim kroz mere sličnosti. Procedura klaster analize dalje stavlja grupu sličnih objekata u klaster.

Mere korelacije

Mera povezanosti između objekata koja verovatno prva dolazi u obzir je koeficijent korelacije između objekata zasnovan na paru promenljivih. Visoka korelacija pokazuje sličnost, a slaba korelacija označava nedostatak iste.

Mere udaljenosti

Iako korelacione mere imaju mnoge prednosti i koriste se u drugim multivariacionim tehnikama, nisu najčešće korišćene mere sličnosti u klaster analizi. Razlika između korelacionih i mera udaljenosti jeste ta što mere udaljenosti se zasnivaju na veličini vrednosti i tako povezuju slične slučajeve, mada mogu imati veoma različite dijagrame promenljivih.

Mere udruživanja

Mere udruživanja se koriste za upoređivanje objekata čije su karakteristike merljive jedino u nemetričnim uslovima (nominalana ili ordinalna merenja). Kao na primer, ispitanici mogu odgovoriti sa ***da*** ili ***ne*** na odgovarajući broj pitanja. Mere udruživanja moraju proceniti stepen podudaranja između svakog para ispitanika. Najjednostavniji oblik mere udruživanja može biti procentualno prikazan odgovor sa *da* ili *ne* ispitanika kroz čitav set pitanja. Mere udruživanja imaju ograničenu podršku računarskih programa. **Korelacione i mere udaljenosti zahtevaju metrične podatke dok mere udruživanja su za nemetrične podatke.**

Procena valjanosti klaster rešenja

Procena valjanosti uključuje težnju istraživača da osiguraju da klaster rešenja reprezentuju celokupnu populaciju. Većina direktnih pokušaja odnosi se na to da klaster analiza odvoji uzorke formirajući klastere, kroz procenu sličnosti između podataka. Svaki klaster će se odvojeno analizirati, a zatim će se rezultati uporediti.

Profilisanje klaster rešenja

Analiza profila fokusira se na opis onoga što ne određuje direktno klastere, ali posle identifikacije čini jednu od njegovih karakteristika. Osim toga, naglasak je na onim karakteristikama koje se značajno razlikuju od klastera do klastera i onim koje bitno odlučuju o članstvu u određenom klasteru.

Neke moguće strategije procene valjanosti dobijenih rezultata:

a) Ponavljanje

Verovatno najbolji način provere dobijenog klasterskog rešenja jeste ponavljanje postupka na drugom slučajno odabranom uzorku.

b) Testiranje razlika između klastera na varijablama korišćenim za njihovo formiranje

Ovaj pristup podrazumeva korišćenje multivarijacione analize varijanse, ili više jednostavnih analiza varijanse ili diskriminacione analize zavisno do broja varijabli i klastera. Ozbiljan statistički problem u ovom slučaju predstavlja činjenica da objekti nisu svrstani u klastere po slučaju, već su klasifikovani tako da se maksimizira udaljenost između klastera po korišćenim varijablama. Ovo u statističkom smislu postpak čini neadekvatnim.

c) Testiranje razlika između klastera na nekim relevantnim eksternim varijablama

Postupak je identičan kao i prethodni ali se testiranje razlika među klasterima vrši na nekim relevantnim varijablama koje nisu korišćene u postupku klasterizacije.

d) Monte Carlo metode

Odnose se na složene postupke poređenja dobijenog rešenja sa rešenjem koje predstavlja simulaciju na slučajno odabranim brojevima.

3.1. Aglomerativne metode klaster analize

Metode koje slede spadaju u grupu aglomerativnih i hijerarhijskih metoda. One daju hijerarhijski niz različitih podela taksonomskih jedinica na klastere, polazeći od jednoelementnih klastera koji se dalje spajaju u veće klastere.

Na početku se pretpostavlja da svaka taksonomska jedinica predstavlja klaster za sebe. Prvi korak je grupisanje dve takosonomske jedinice sa najvećim koeficijentom sličnosti (najmanjim koeficijentom rastojanja). Ako ima više parova taksonomskih jedinica sa istim, najvećim koeficijentom sličnosti tada se proizvoljno izabere jedan par taksonomskih jedinica koji se prvo povezuje. Dalje se računa koeficijent sličnosti (rastojanja) između dobijenih grupa taksonomskih jedinica.

Metode se razlikuju po tome na koji način se računa rastojanje (sličnost) između grupa takosnomskih jedinica. Nakon izračunavanja koeficijenta sličnosti (rastojanja) za dobijene grupe taksonomskih jedinica vrši se spajanje grupa sa najvećim koeficijentom sličnosti (najmanjim koeficijentom rastojanja). Taj najveći koeficijent sličnosti (najmanji koeficijent rastojanja) zove se **nivo spajanja**. Zatim se opet računa koeficijent sličnosti (rastojanja) između novodobijenih grupa, itd. Postupak se nastavlja sve dok se sve taksonomske jedinice ne grupišu u jedan klaster.

Većina klaster metoda ovog tipa primenjuju se na već izračunate matrice sličnosti i rastojanja taksonomskih jedinica. Rezultati ovog tipa klaster metoda se obično predstavljaju specijalnim dijagramom koji se naziva **dendogram**. Jedna mogućnost je da se na vertikalnoj osi dendograma prikazuju taksonomske jedinice, a na horizontalnoj koeficijenti sličnosti (rastojanja, nivoi spajanja), a druga mogućnost je da se na horizontalnoj osi prikazuju taksonomske jedinice, a na vertikalnoj nivoi spajanja. Taksonomske jedinice (odnosno grupe taksonomskih jedinica) se povezuju na nivoima spajanja taksonomskih jedinica ili grupa.

Svaki stepen sličnosti odnosno rastojanja (odnosno nivo spajanja) određuje jednu particiju skupa svih taksonomskih jedinica. Svaka takva particija odgovara podeli u grupe taksonomskih jedinica koje se posmatraju.

Dakle, ovaj tip metoda daje više mogućih podela taksonomskih jedinica na klastere. Taksonom može da odluči na osnovu raznih kriterijuma koju od tih podela će da usvoji u svojoj klasifikaciji.

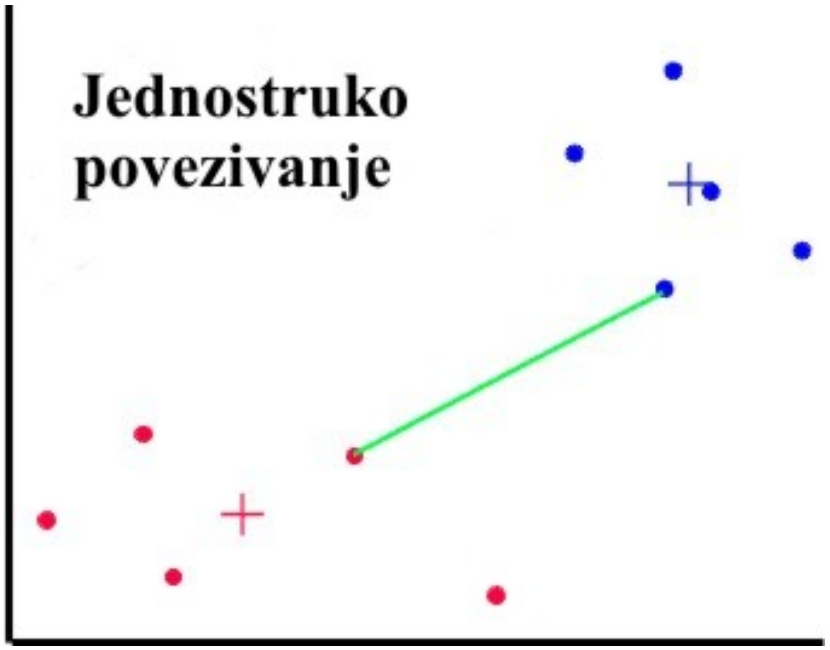
Postoji nekoliko metoda koje spadaju u grupu aglomerativnih, hijerarhijskih metoda. Osnovna razlika između metoda je način određivanja mere sličnosti, odnosno različitosti između grupa taksonomskih jedinica.

Najpoznatije metode ovog tipa su:

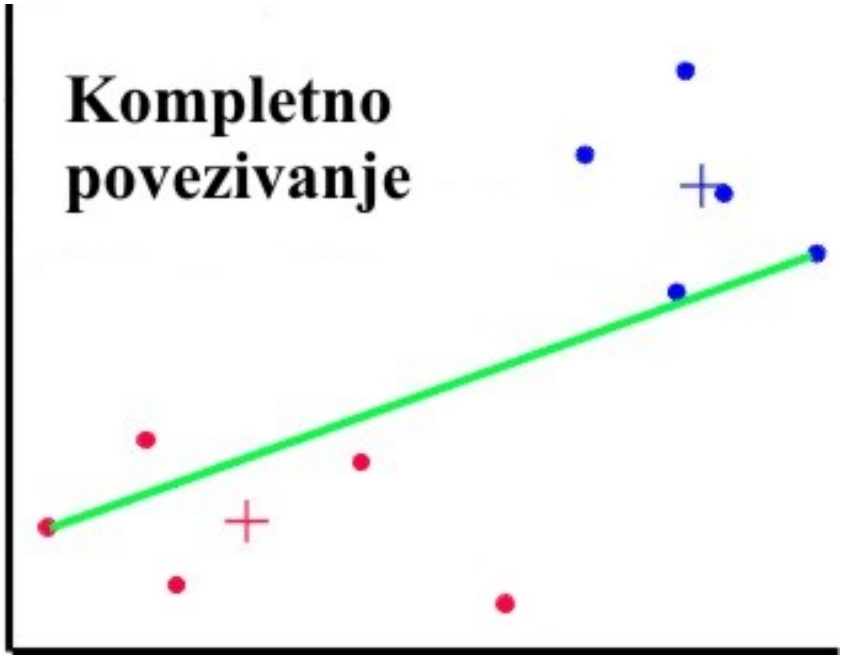
- **klaster metoda prostog povezivanja;**
- **klaster metoda kompleksnog povezivanja;**
- **klaster metoda proseka;**
- **centroidna klaster metoda i**
- **metoda Ward-a.**

One počinju nizom particija podataka: na početku postoji *n* grupa sa po jednim članom, dok na kraju imamo jednu grupu sa *n* članova:

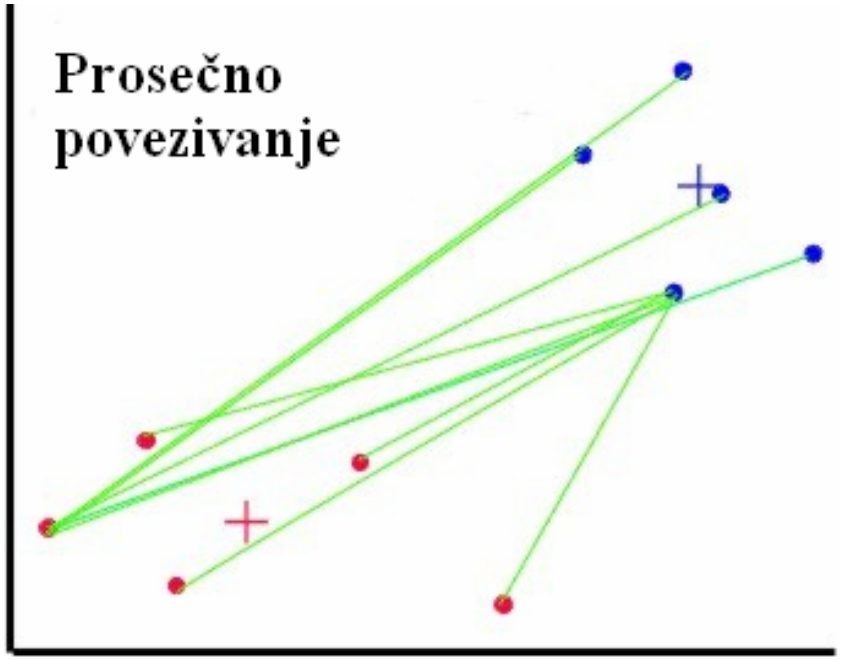
- **metod jednostrukog (prostog) povezivanja** (single linkage) – kao mera rastojanja između dve grupe korisi se najkraće rastojanje između para objakata koji pripadaju ovim grupama;



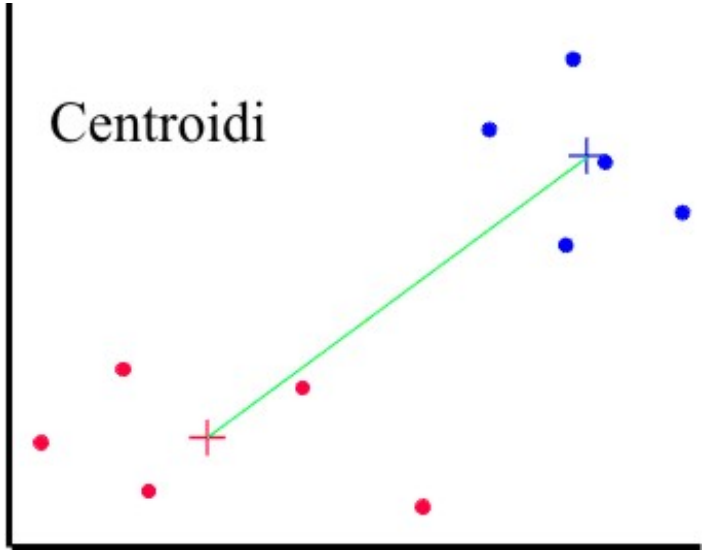
- **metod kompleksnog (potpunog) povezivanja** (complete linkage) – mera rastojanja između dve grupe predstavlja najveće rastojanje između para objekata koji pripadaju tim grupama;



- **metod prosečnog povezivanja (metod proseka)** (average linkage) – rastojanje se određuje prema prosečnom rastojanju svih objekata koji pripadaju dvema grupama tj. saberu se sva rastojanja i podele brojem koliko ih ima.



- Sledeća hijerarhijska metoda udruživanja je **metoda centroida**, kod koje se dve grupe udružuju u novu grupu ako su njihovi centroidi najmanje udaljeni međusobno u odnosu na međusobnu udaljenost svih mogućih parova grupa koje postoje na posmatranom nivou udruživanja.



- **Metoda Ward-a** - Ward uvodi tip metoda u kome dolazi do udruživanja grupa ako je njihovim udruživanjem došlo do najmanjeg povećanja sume kvadrata unutar grupa, datog sa:

$$E = \sum_{m=1}^{g} E_m ,$$

gde je:

$$E_m = \sum_{l=1}^{n_m} \sum_{k=1}^{p} \left(x_{m_l,k} - \overline{x}_{m,k} \right)^2 ,$$

pri čemu je $\overline{x}_{m,k} = \frac{1}{n_m} \sum_{l=1}^{n_m} x_{m_l,k}$ (srednja vrednost m-tog klastera za k-tu promenljivu), a $x_{m_l,k}$ je vrednost k-te promenljive $(k = 1,...,p)$ za l-ti objekat $(l = 1,...,n_m)$ u m-tom klasteru $(m = 1,...,g)$.

Primene anglomerativnih metoda klaster analize

3.1.1. Klaster metoda prostog povezivanja

Kod **klaster metode prostog povezivanja** koeficijent sličnosti između dve grupe taksonomskih jedinica jednak je najvećem koeficijentu sličnosti dve taksonomske jedinice od kojih je prva iz prve, a druga iz druge grupe. Ako se računa preko koeficijenta rastojanja, tada je rastojanje dve grupe jednako najmanjem koeficijentu rastojanja, ako se posmatraju svi parovi taksonomskih jedinica od kojih je prva iz prve, a druga iz druge grupe.

Ilustrativni primer. U Tabeli 1. dati su koeficijenti rastojanja na skupu taksonomskih jedinica $T=\{t_1,t_2,t_3,t_4,t_5,t_6\}$.

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	0	0.7	0.6	0.3	0.2	0.1
t_2	0.7	0	0.2	0.4	0.3	0.8
t_3	0.6	0.2	0	0.9	0.8	0.7
t_4	0.3	0.4	0.9	0	0.5	0.2
t_5	0.2	0.3	0.8	0.5	0	0.3
t_6	0.1	0.6	0.7	0.2	0.3	0

Tabela 1. Koeficijenti rastojanja taksonomskih jedinica

Na početku (na stepenu rastojanja 0) sve taksonomske jedinice se posmatraju kao grupa za sebe. Zatim se u prvom koraku spajaju taksonomske jedinice sa najmanjim stepenom rastojanja. To su t_1 i t_6, jer je njihov stepen rastojanja 0.1. Sada se te dve taksonomske jedinice u istoj grupi, a sve ostale se posmatraju kao grupa za sebe. U sledećem koraku se

računa rastojanje nivodobijane grupe $\{t_1,t_6\}$ od ostalih taksonomskih jedinica. Rastojanje te grupe od t_2 jednako je manjem od rastojanje $d(t_1,t_2)$ i $d(t_6,t_2)$ i to je 0.7. Rastojanje između te grupe i t_3 je opet jednako manjem od 0.6 i 0.7, a to je 0.6. Rastojanje te grupe od t_4 je 0.2, kao i rastojanje od t_5. Novoizračunata rastojanja su u Tabeli 2.

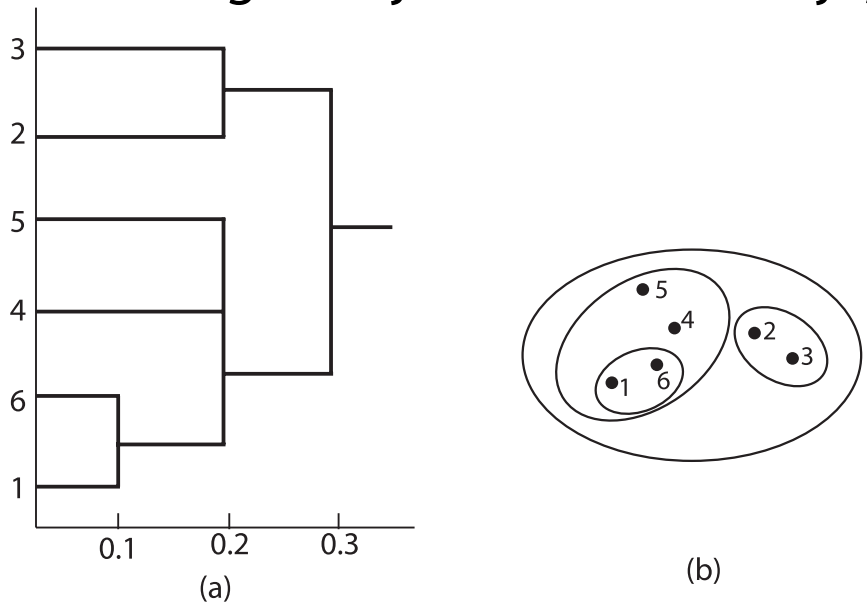
	t_{16}	t_2	t_3	t_4	t_5
t_{16}	0	0.7	0.6	0.2	0.2
t_2	0.7	0	0.2	0.4	0.3
t_3	0.6	0.2	0	0.9	0.8
t_4	0.2	0.4	0.9	0	0.5
t_5	0.2	0.3	0.8	0.5	0

Tabela 2. Novodobijena rastojanja

U sledećem koraku se opet traži najmanji koeficijent rastojanja (različit od nule) u tablici. To je 0.2. Pošto se on nalazi na više mesta u tablici, na tom nivou spajanja spojiće se sve taksonomske jedinice koje se nalaze na

rastojanju 0.2. Dakle, spajaju se t_4 i t_5 sa grupom t_{16}, a takođe i t_2 i t_3 čine grupu za sebe. Sada sve taksonomske jedinice podeljene u dve grupe. Poslednji korak je računanja rastojanja između dve dobijene grupe. To je najmanje rastojanje kada se posmatra svaki element prve sa svakim elementom gruge grupe, pa je poslednji nivo spajanja 0.3 (rastojanje između t_2 i t_5).

Rezultati primene klaster metode prostog povezivanja predstavljeni su u dendrogramu na Slici 2(a) i odgovarajućim Venovim dijagramom na Slici 2(b).



Slika 2(a) Dendrogram i 2(b) Venov dijagram

Primena klaster metode prostog povezivanja na taksonomiju sijalica

Primer A. Na pet različitih proizvođača sijalica (taksonomskih jedinica), ispitano je osvetljenje koje daju sijalice, redom, od 60W, 75W, 100W i 150W. **Na osnovu podataka osvetljena za navedene sijalice** formirajmo hijerarhisku strukturu objekata korišćenjem metode jednostrukog povezivanja, a kao razdaljinu između objekata koristićemo Euklidovo rastojanje.

k_1- sijalica od 60W, k_2- 75W, k_3- 100W, k_4- 150W.

t_1- sijalica proizvođača :G. Electric, t_2- Maxi, t_3- Philips, t_4- JMC, t_5- ΕΛΛΑΣ

	k_1	k_4	k_3	k_2
t_1	145	403	321	321
t_2	119	320	250	189
t_3	20.36	488	274	51
t_4	114	482	252	156
t_5	148	504	266	190

Tabela 3. Matrica podataka.

Ovi karakteri su kvantitativni i za meru različitosti korišćeno je Euklidovo rastojanje (Tabela 4).

	t_1	t_2	t_3	t_4	t_5
t_1	0	112,72	217,87	117,11	101,04
t_2	112,72	0	239,94	165,41	186,95
t_3	217.87	239,94	0	41,67	189,56
t_4	117,11	165,41	41,67	0	54,70
t_5	101.04	186,95	189,56	54,70	0

Tabela 4. Izračunata Euklidova rastojanja

Na početku sve taksonomske jedinice se posmatraju kao grupa za sebe. Zatim se u prvom koraku spajaju taksonomske jedinice sa najmanjim stepenom rastojanja.

To su u ovom slučaju t_3 i t_4, jer je njihov stepen rastojanja 41,67. U sledećem koraku računamo rastojanje novodobijene grupe od ostalih taksonomskih jedinica. Rastojanje te grupe od t_1 jednako je manjem od rastojanja $d(t_3,t_1)$ i $d(t_4,t_1)$, i to je 117,11. Rastojanje te grupe i t_2 je opet jednako manjem od 239,94 i 165,41, i to je 165,41. Rastojanje te grupe od t_5 je 54,70 (Tabela 5).

	t_{34}	t_1	t_2	t_5
t_{34}	0	117,11	165,41	54,70
t_1	117,11	0	112,72	101,04
t_2	165,41	112,72	0	186,95
t_5	54,70	101,04	186,95	0

Tabela 5.

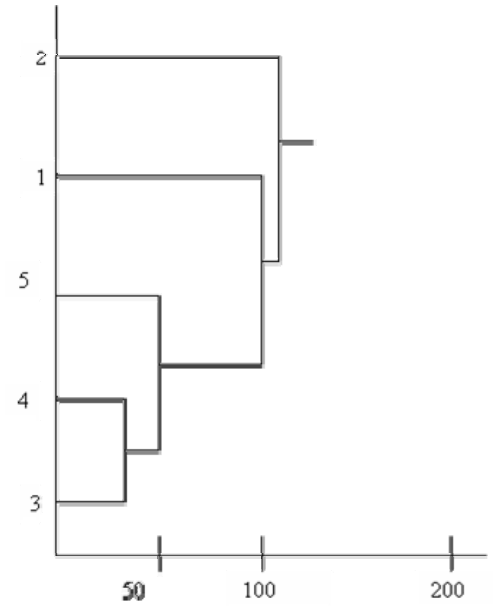
Ponovljenim postupcima dobijaju se Tabele 6 i 7 iz kojih rezultuje dendogram na Slici 3.

	t_{345}	t_1	t_2
t_{345}	0	101,04	165,41
t_1	101,04	0	112,72
t_2	165,41	112,72	0

Tabela 6.

	t_{3451}	t_2
t_{3451}	0	112.72
t_2	112,72	0

Tabela 7.



Slika 3. Dendrogram

Iz ovoga vidimo da su sijalice t_3 i t_4 najsličnije (tj. sijalica proizvođača Philips i sijalica proizvođača JMC imaju imaju najsličnija osvetljenja). Sijalice Philips i JMC imaju najrazličitije osvetljenje sa sijalicom proizvođača Maxi. Dalje sijalice Philips i JMC su najsličije sa sijalicom proizvođača ΕΛΛΑΣ . Sijalica proizvođača ΕΛΛΑΣ je najsličnija sa grupom sijalica Philips i JMC i sa sijalicom G. Electric. Sijalica G. Electric je najsličnija sa sijalicom Maxi.

Realizacija primera hijerarhijske structure na osnovu snage sijalica pomoću SPSS-a

Primer B. Na pet različitih proizvođača sijalica (taksonomskih jedinica), ispitana je snaga koju daju. Sijalice su redom od 60W, 75W, 100W i 150W. **Na osnovu podataka snage za navedene sijalice** formirajmo hijerarhisku strukturu objekata korišćenjem metode jednostrukog povezivanja, a kao razdaljinu između objekata koristićemo Euklidovo rastojanje.

k_1- sijalica od 60W, k_2- 75W, k_3- 100W, k_4- 150W.

t_1- sijalica proizvođača :G. Electric, t_2- Maxi, t_3- Philips, t_4- JMC, t_5- ΕΛΛΑΣ

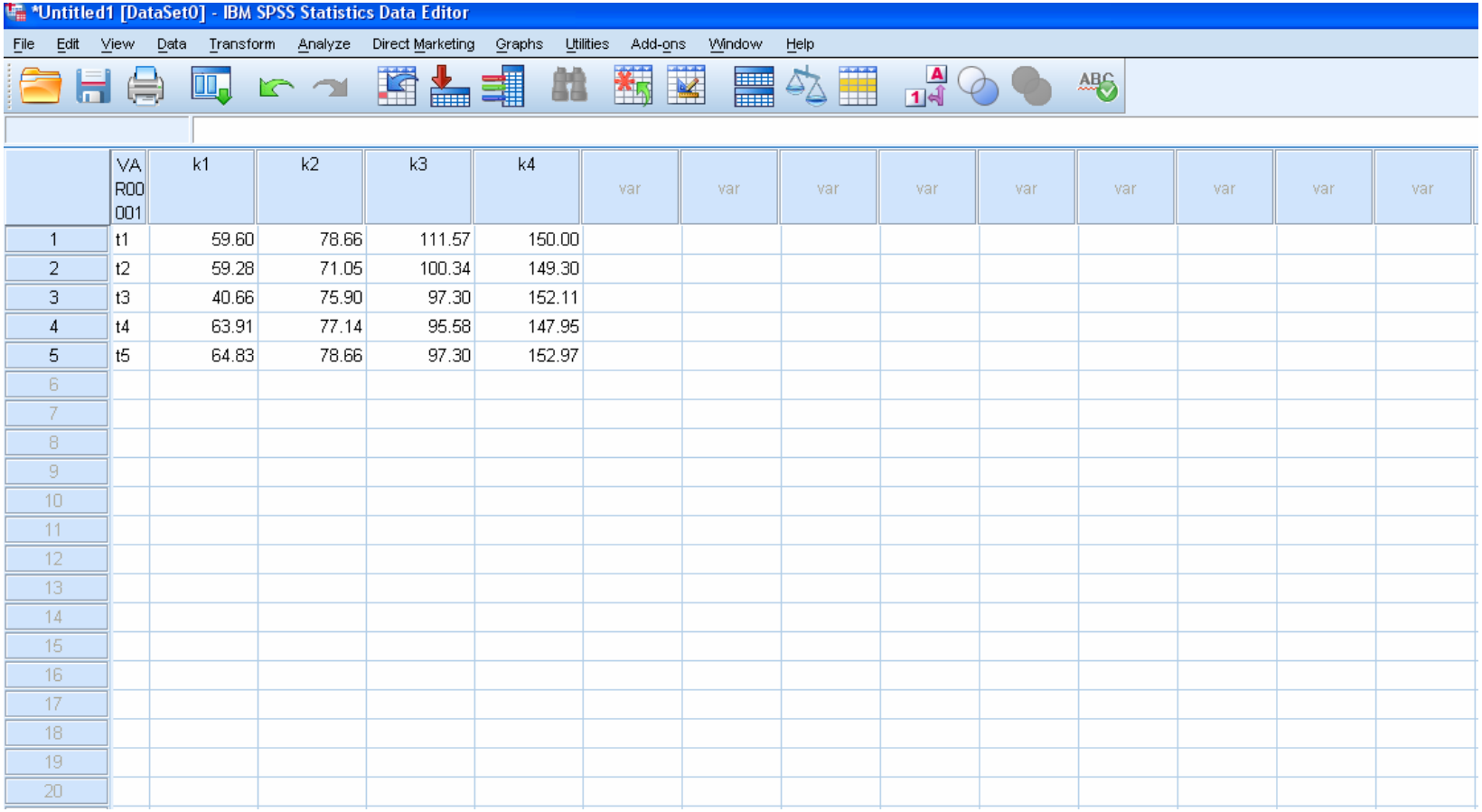
Kod ovog primera se vrši klasifikacija u odnosu na snagu koju daju sijalice, pa se vrednosti u Tabeli 8. razlikuju od Tabele 3, inače proizvođači su isti u oba primera.

Matrica podataka

	k_1	k_4	k_3	k_2
t_1	59,60	150	111,57	78,66
t_2	59,28	149,30	100,34	71,05
t_3	40,66	152,11	97,30	75,90
t_4	63,91	147,95	95,58	77,14
t_5	64,83	152,97	97,30	78,66

Tabela 8.

Nakon unosa podataka u SPSS *Data View* izgleda ovako:



	VAR00001	k1	k2	k3	k4	var	var	var	var	var	var	var	var	var
1	t1	59.60	78.66	111.57	150.00									
2	t2	59.28	71.05	100.34	149.30									
3	t3	40.66	75.90	97.30	152.11									
4	t4	63.91	77.14	95.58	147.95									
5	t5	64.83	78.66	97.30	152.97									
6														
7														
8														
9														
10														
11														
12														
13														
14														
15														
16														
17														
18														
19														
20														

Variable View:

*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

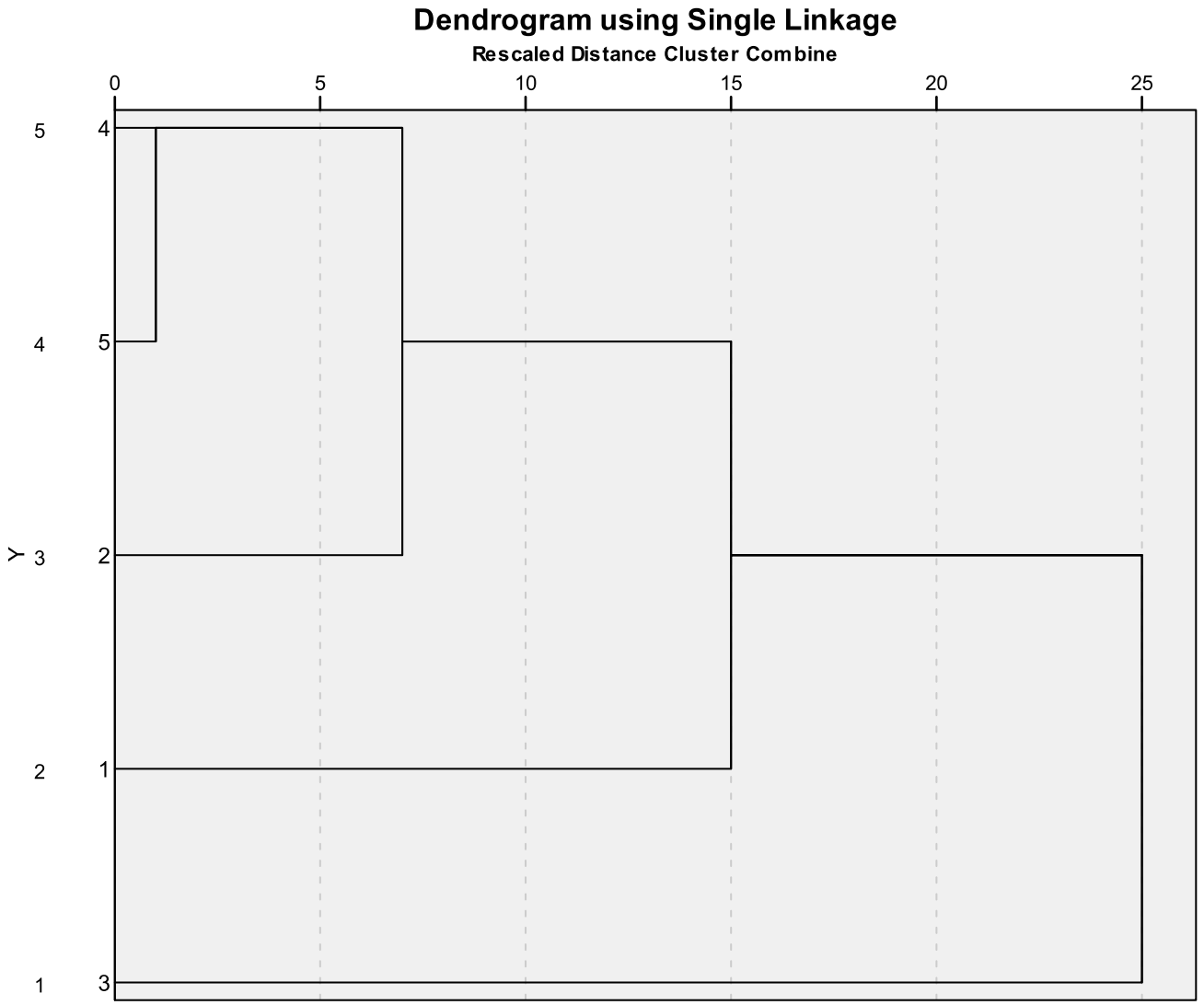
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	VAR00001	String	2	0		None	None	2	Left	Nominal	Input
2	k1	Numeric	8	2		None	None	8	Right	Scale	Input
3	k2	Numeric	8	2		None	None	8	Right	Scale	Input
4	k3	Numeric	8	2		None	None	8	Right	Scale	Input
5	k4	Numeric	8	2		None	None	8	Right	Scale	Input
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											

Rešenje:

Naredbom iz menija: *Analyze→Classify→Hierarhical Cluster…* pokrećemo hijerarhisku klaster analizu. U polje *Variable(s)* ubacujemo promenljive na osnovu kojih se vrši analiza. U polje *Label Cases by* ubacujemo varijablu tipa *String* preko koje identifikujemo objekte (u našem slučaju sijalice). U opcijama *Plots* izaberemo opciju *Dendogram* kako bi na izlazu dobili i dendogram povezivanja objekata. U opcije metoda (*Method*) biramo metod za analizu (u našem slučaju to je jednostruko povezivanje – Nearest neighbor) i kao interval za meru izabiramo Euklidsko odstojanje. Pritiskom na dugme *OK* dobijamo rezultate analize.

Dendogram je grafički izveštaj rešenja problema. Objekti su poređani po levoj vertikalnoj osi. Horizontalna osa pokazuje razdaljinu između objekata kada se povezuju. Deljenje dendograma kako bi dobili određen broj grupa je subjektivna procena. Najčešće tražimo velike skokove između povezivanja na horizontalnoj osi. Vidimo da je najveći skok kada se povezuje objekat broj 3 (sijalica proizvođača Philips) sa klasterom svih drugih objekata, pa presecanjem dendograma na ovom mestu možemo dobiti dve grupe. Ponavljanjem ovog postupka možemo dobiti više grupa.

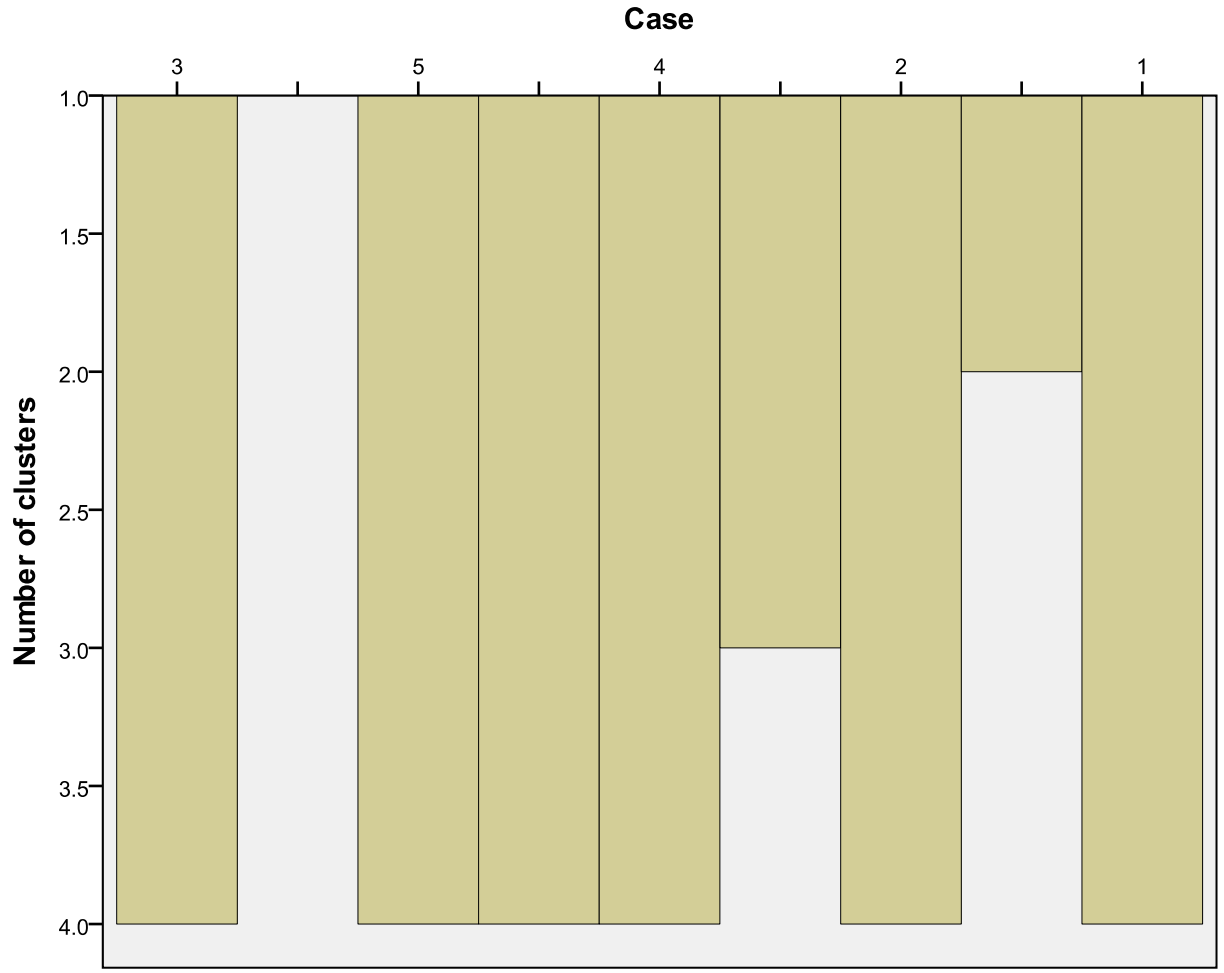

Dendrogram using Single Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25
Y
5
4
3
2
1
4
5
2
1
3

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	5	5.596	0	0	2
2	2	4	9.111	0	1	3
3	1	2	13.587	0	2	4
4	1	3	19.682	3	0	0

U ovoj tabeli vidimo da se u prvoj fazi povezuju objekti 4 i 5 zato što je njihovo rastojanje najmanje. Grupa kreirana njihovim spajanjem se pojavljuje opet u fazi 2 što nam govori zadnja kolona tabele. U fazi 2 klaster se sjedinjuje sa objektom 2. Ako je broj objekata posmatranja veliki lakše je pratiti kolonu koeficijenata i tražiti velike skokove, nego pratiti dendogram.

Na osnovu ovih podataka možemo određivati broj grupa. Za grafčko predstavljanje broja klastera može se koristiti Baner grafik (Banner plot).



Baner grafik hijerarhijske strukture klastera za posmatrane sijalice.

3. 1. 2. Klaster metoda kompleksnog povezivanja

Kod **klaster metode kompleksnog povezivanja** koeficijent sličnosti između dve grupe taksonomskih jedinica je jednak najmanjem koeficijentu sličnosti (najvećem rastojanju) dve taksonomske jedinice od kojih je prva iz prve, a druga iz druge grupe. Dakle, ovo je suprotno u odnosu na klaster metodu prostog povezivanja. Koeficijent rastojanja jedne grupe od sebe same se uzima uvek da je 0 (iako je rastojanje pojedinih elemenata iz te grupe nekad i veće od 0).

Ponovo polazimo od **ilustrativnog primera** i Tabele 1. Prvi korak je da se utvrdi da su najmanje različite taksonomske jedinice t_1 i t_6 sa stepenom rastojanja 0.1, i one se povežu na nivou 0.1. Zatim se računa rastojanje grupe $\{t_1,t_6\}$ od ostalih taksonomskih jedinica, ali ovde je rastojanje jednako najvećem od pojedinačnih rastojanja koja se posmatraju. Tako je rastojanje između grupe t_{16} i t_3 je 0.7, a rastojanje te grupe od t_4 je 0.3, kao i od t_5. Koeficijenti rastojanja su predstavljeni u Tabeli 9.

	t_{16}	t_2	t_3	t_4	t_5
t_{16}	0	0.8	0.7	0.3	0.3
t_2	0.8	0	0.2	0.4	0.3
t_3	0.7	0.2	0	0.9	0.8
t_4	0.3	0.4	0.9	0	0.5
t_5	0.3	0.3	0.8	0.5	0

Tabela 9.

U sledećem koraku se uočava da je najmanje nenula rastojanje u tablici 0.2, i to između taksonomskih jedinica t_2 i t_3 i zato se te taksonomske jedinice spajaju u grupu. U Tabeli 10. je dato rastojanje između grupa dobijenih nakon ovog koraka.

	t_{16}	t_{23}	t_4	t_5
t_{16}	0	0.8	0.3	0.3
t_{23}	0.8	0	0.9	0.8
t_4	0.3	0.9	0	0.5
t_5	0.3	0.8	0.5	0

Tabela 10.

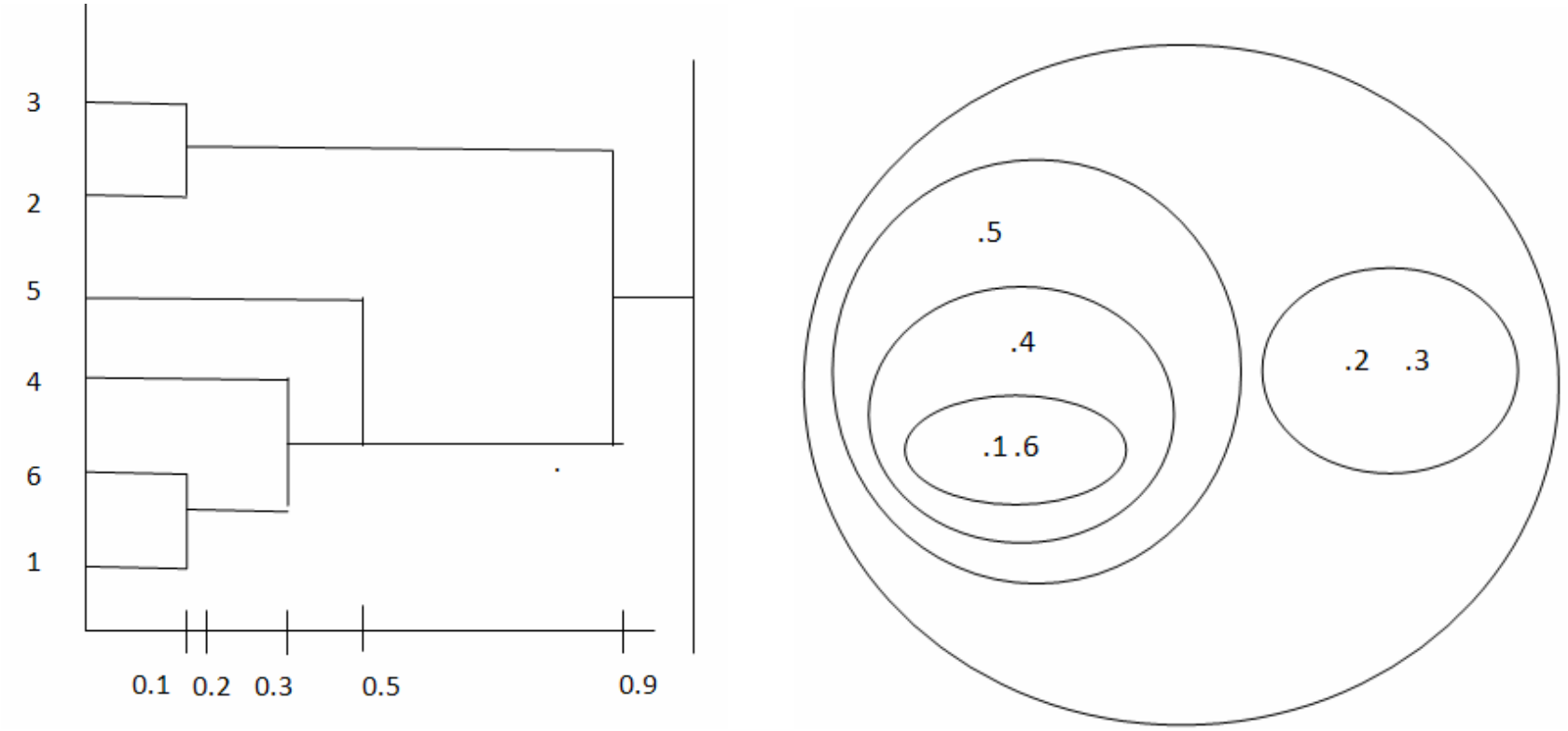
Sledeće najmanje rastojane je 0.3 i to izmeću grupe t_{16} i elemenata t_4 i t_5. Rastojanje između elemenata t_4 i t_5 je 0.5, pa se zato ne spaja grupa sa oba elementa na nivou 0.3, nego se bira jedan od njih (obično se bira element sa manjim indeksom). Na nivou 0,3 dobija se grupa $\{t_1,t_4,t_6\}$. Rastojanja između te grupe i ostalih elemanata prikazana su u Tabeli 11.

	t_{146}	t_{23}	t_5
t_{146}	0	0.9	0.5
t_2	0.9	0	0.8
t_5	0.5	0.8	0

Tabela 11.

U sledećem koraku se t_5 spaja sa grupom t_{146} na nivou 0.5.

Na kraju se računa rastojanje između dve dobijene grupe i ono je jednako najvećem rastojanju između elemenata grupe, a to je 0.9.



Slika 4(a) Dendrogram i 4(b) Venov dijagram

Primena metode kompleksnog povezivanja na razvrstavanje gradova po osnovu njihovih meteoroloških karakteristika

Primer A). U ovom primeru posmatrano je deset virtuelnih gradova. U **dva** fiksirana dana (20.januar i 20.avgust) u toku jedne kalendarske godine izmerene su sledeće karakteristike klime:

- Temperatura - k_1, k_2
- Količina padavina - k_3, k_4
- Brzina vetra - k_5, k_6

Podaci o klimatskim osobinama posmatranih gradova predstavljeni su u sledećoj tabeli:

	k_1	k_2	k_3	k_4	k_5	k_6
t_1	-5	21	0.5	13.6	13	1
t_2	0	32	0.2	0	6	2
t_3	3	19	0.2	10.1	7	1
t_4	11	25	8.2	13.6	8	1
t_5	-1	33	0.2	12.3	12	2
t_6	3	28	5.0	11.5	2	3
t_7	10	40	8.2	48.4	1	1
t_8	-3	26	0.5	10.8	10	1
t_9	7	31	0	13.3	4	2
t_{10}	6	25	8.7	33.8	3	3

Tabela 12.

Na osnovu podataka iz Tabele 12, Euklidovo rastojanje između posmatranih gradova je:

$d(t_1,t_2) = 19.52$, $d(t_1,t_3) = 10.79$, itd. (Tabela 13).

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
t_1	0	19.52	10.79	18.87	13.55	16.21	44.73	10.29	10.06	17.55
t_2	19.52	0	16.8	20.59	13.76	14.05	50.96	13.37	15.20	36.19
t_3	10.79	16.8	0	13.31	15.18	11.63	23.49	9.73	13.43	26.44
t_4	18.87	20.59	13.31	0	17.04	11.30	38.55	16.37	11.67	21.50
t_5	13.55	13.76	15.18	17.04	0	12.87	40.73	8.51	11.54	27.01
t_6	16.21	14.05	11.63	11.30	12.87	0	39.62	11.35	7.63	23.02
t_7	44.73	50.96	23.49	38.55	40.73	39.62	0	43.81	37.41	21.50
t_8	10.29	13.37	9.73	16.37	8.51	11.35	43.81	0	13.76	27.04
t_9	10.06	15.20	13.43	11.67	11.54	7.63	37.41	13.76	0	23.13
t_{10}	17.55	36.19	26.44	21.50	27.01	23.02	21.50	27.04	23.13	0

Tabela 13.

Prvi klaster čine gradovi sa najmanjim stepenom rastojanja, t_6 i t_9 jer je njihov koeficijent rastojanja 7,63. U sledećem koraku se računa rastojanje novonastale grupe gradova t_6 i t_9 od ostalih gradova (Tabela 14).

	t_1	t_2	t_3	t_4	t_5	$t_{6,9}$	t_7	t_8	t_{10}
t_1	0	19.52	10.79	18.87	13.55	16.21	44.73	10.29	17.55
t_2	19.52	0	16.8	20.59	13.76	15.20	50.96	13.37	36.19
t_3	10.79	16.8	0	13.31	15.18	13.43	23.49	9.73	26.44
t_4	18.87	20.59	13.31	0	17.04	11.67	38.55	16.37	21.50
t_5	13.55	13.76	15.18	17.04	0	12.87	40.73	8.51	27.01
$t_{6,9}$	16.21	15.20	13.43	11.67	12.87	0	39.62	13.76	23.13
t_7	44.73	50.96	23.49	38.55	40.73	39.62	0	43.81	21.50
t_8	10.29	13.37	9.73	16.37	8.51	13.76	43.81	0	27.04
t_{10}	17.55	36.19	26.44	21.50	27.01	23.13	21.50	27.04	0

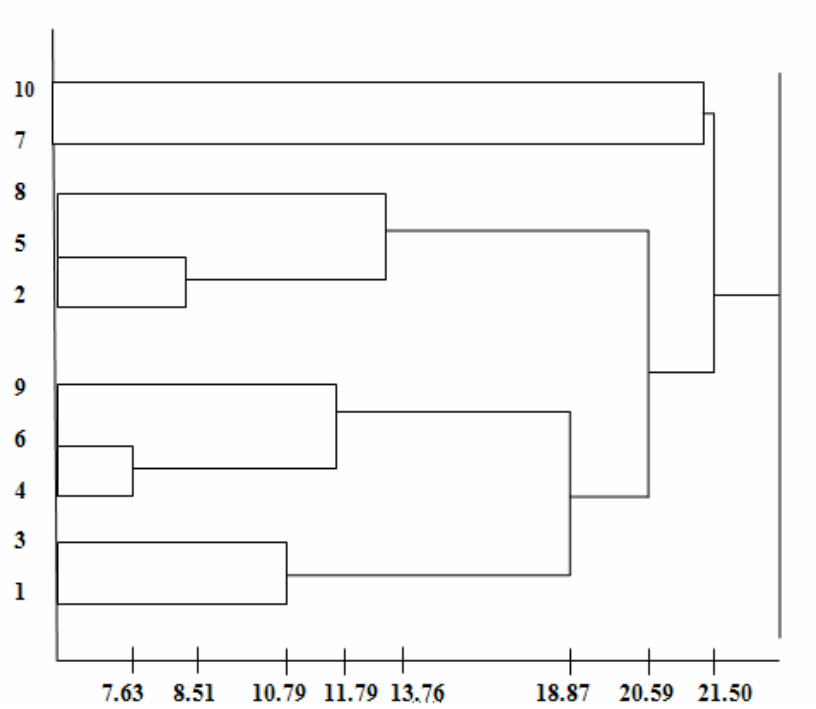
Tabela 14. Klasteri: t_1, t_2, t_3, t_4, t_5, $t_{6,9}$, t_7, t_8, t_{10}

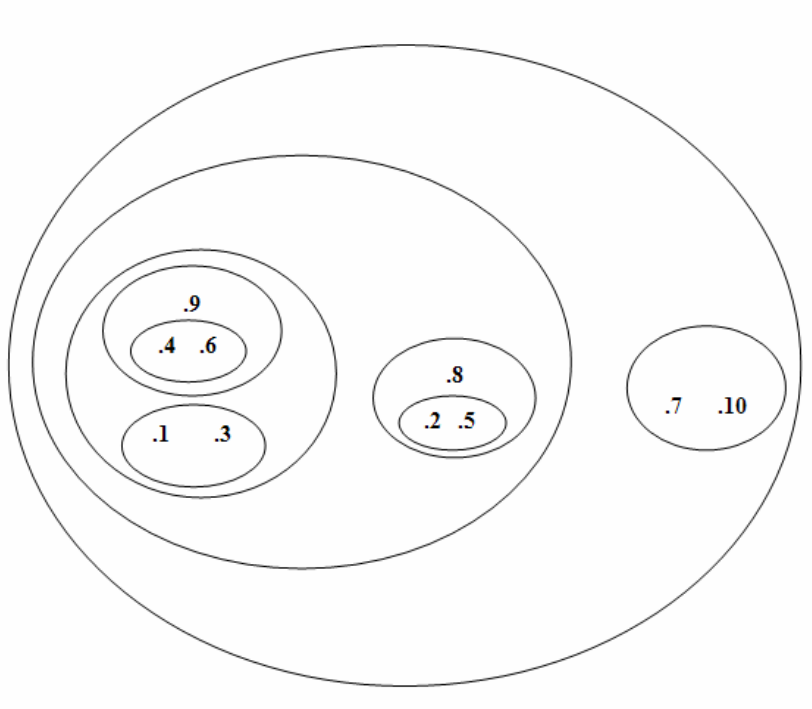
Izrada Tabela 15-21 daje sledeće klasterizacije:

- $t_1, t_2, t_3, t_4, t_{5,8}, t_{6,9}, t_7, t_8, t_{10}$;
- $t_{1,3}, t_2, t_4, t_{5,8}, t_{6,9}, t_7, t_8, t_{10}$;
- $t_{1,3}, t_2, t_{4,6,9}, t_{5,8}, t_7, t_{10}$;
- $t_{1,3}, t_{2,5,8}, t_{4,6,9}, t_7, t_{10}$;
- $t_{1,3,4,6,9}, t_{2,5,8}, t_7, t_{10}$;
- $t_{1,2,3,4,5,6,8,9}, t_7, t_{10}$;
- $t_{1,2,3,4,5,6,8,9}, t_{7,10}$.

Na kraju svi elementi (gradovi) su u jednom klasteru.

Na osnovu prethodnih tabela i izračunatih rastojanja podatke smo predstavili dendogramom i Venovim dijagramom gde su pregledno predstavljeni rezultati primenjene metode (povezani su slični gradovi u klimatskom smislu).





Slika 5(a) Dendrogram i 5(b) Venov dijagram posmatranih gradova

Primer B). KK Partizan Beograd bira idealni tim u prethodnoj deceniji. Najuži krug je činilo 11 igrača koji su obeležili prethodno navedeni period. U obzir je uziman broj:

- k_1 Poena
- k_2 Skokova
- k_3 Asistencija
- k_4 Ukradenih lopti
- k_5 Blokada

Uzeti su pravi podaci sa sajta www.euroleague.net i dati su u Tabeli 22:

t_1 - Nikola Peković
t_2 - Novica Veličković
t_3 - Jan Veseli
t_4 - Uroš Tripković
t_5 - Miloš Vujanić
t_6 - Nenad Krstić
t_7 - Lorens Roberts
t_8 - Dušan Kecman
t_9 - Vlado Šćepanović
t_{10} - Vule Avdalović
t_{11}- Petar Božić

	k_1	k_2	k_3	k_4	k_5
t_1	16.4	6.8	0.6	0.4	0.4
t_2	14.6	7.9	1.6	0.9	0.6
t_3	10	3.6	1	1.2	0.8
t_4	7.1	1.8	0.9	0.2	0
t_5	25.8	3	3.2	1	0
t_6	15.4	6.5	0.9	1.2	1.3
t_7	7	5.3	1.1	0.4	0.2
t_8	13.4	4.9	2	1.6	0.5
t_9	19	1.5	1.3	0.5	0
t_{10}	13.1	2.5	2	1	0
t_{11}	3.8	0.9	1.2	0.5	0

Tabela 22.

Euklidovo rastojanje primenjeno na prethodnu tabelu daje: $d(t_1,t_2) = 2.39$, $d(t_1,t_3) = 7.22$, itd. (Tabela 23).

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}
t_1	0	2.39	7.22	10.57	10.49	1.62	9.53	4	5.95	5.64	13.93
t_2	2.39	0	6.33	9.73	12.34	1.91	8	3.33	7.8	5.65	12.8
t_3	7.22	6.33	0	3.64	15.98	6.15	3.59	3.8	9.3	3.53	6.84
t_4	10.57	9.73	3.64	0	18.89	9.67	3.51	7.26	11.91	6.19	3.41
t_5	10.49	12.34	15.98	18.89	0	11.28	19.06	12.62	7.23	12.76	22.19
t_6	1.62	1.91	6.15	9.67	11.28	0	8.59	2.92	6.34	4.92	12.96
t_7	9.53	8	3.59	3.51	19.06	8.59	0	6.59	12.58	6.8	5.44
t_8	4	3.33	3.8	7.26	12.62	2.92	6.59	0	6.69	2.54	10.5
t_9	5.95	7.8	9.3	11.91	7.23	6.34	12.58	6.69	0	6.04	15.21
t_{10}	5.64	5.65	3.53	6.19	12.76	4.92	6.8	2.54	6.04	0	9.48
t_{11}	13.93	12.8	6.84	3.41	22.19	12.96	5.44	10.5	15.21	9.48	0

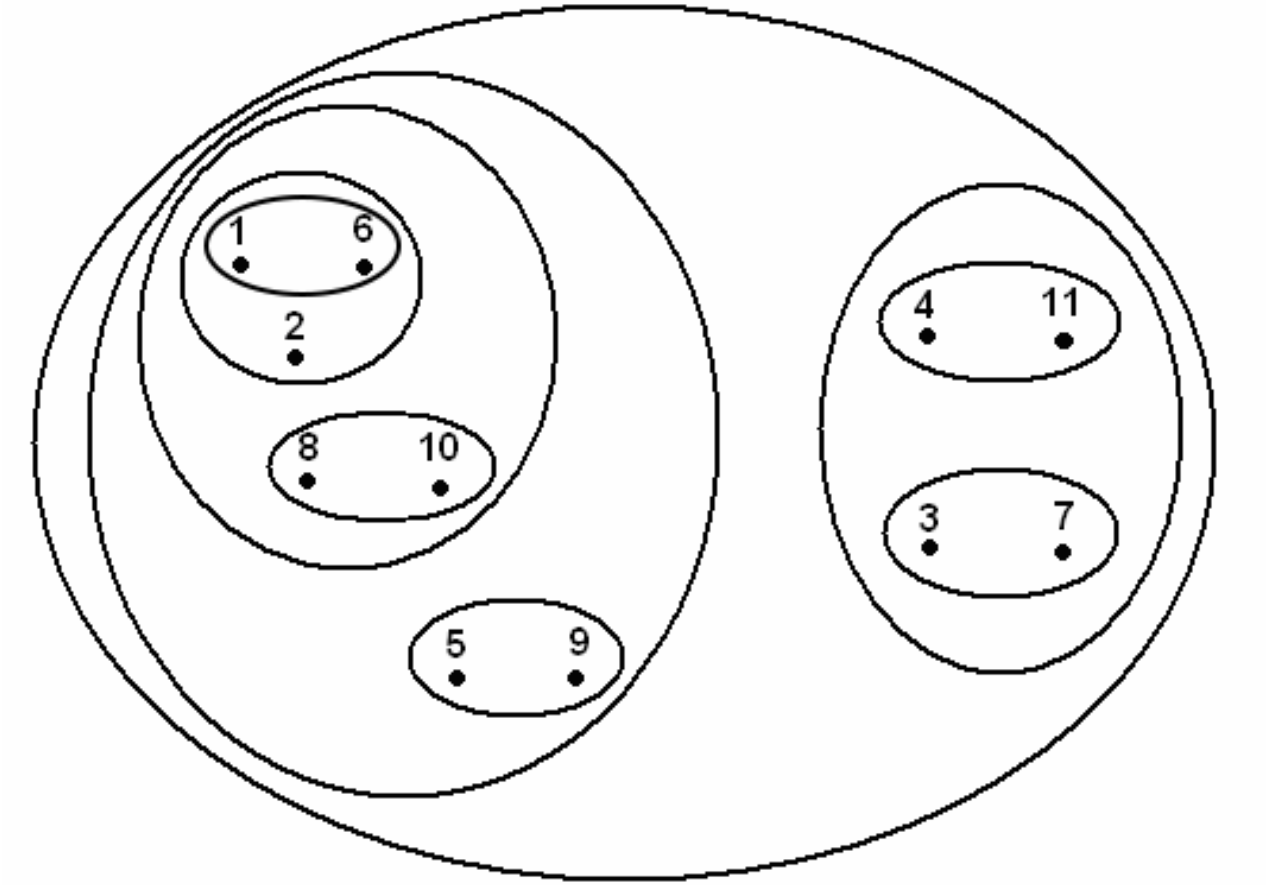
Tabela 23.

Najmanji stepen rastojanja su igrači t_1 i t_6, njihov koeficijent rastojanja je 1.62. U sledećem koraku se računa rastojanje novonastale grupe $\{t_1,t_6\}$od ostalih igrača, pri čemu se bira veće od rastojanja ovih dveju taksonomskih jedinica prema nekoj drugoj, posmatranoj, taksonomskoj jedinici (Tabela 24).

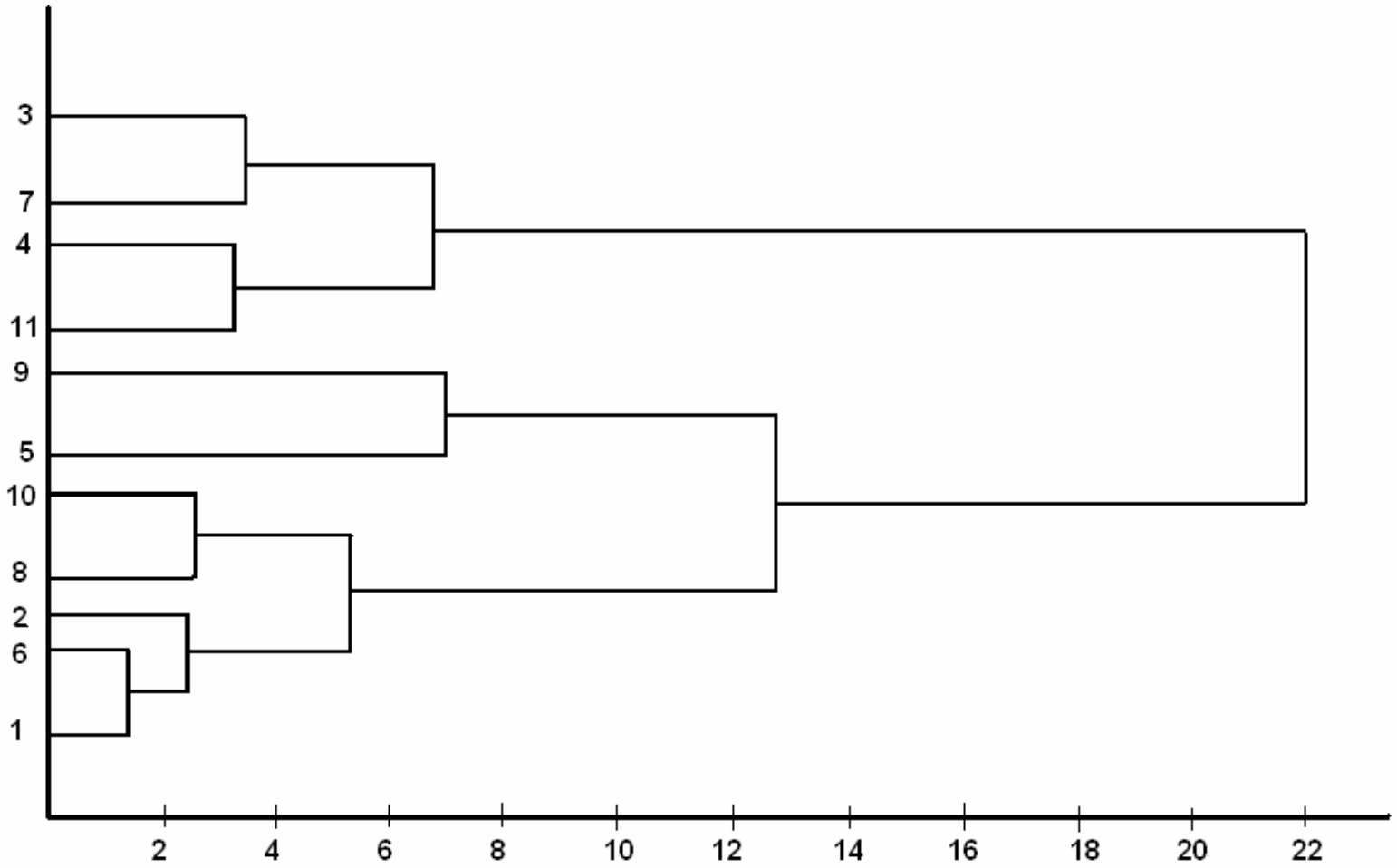
	$t_{1,}\, t_6$	t_2	t_3	t_4	t_5	t_7	t_8	t_9	t_{10}	t_{11}
$t_{1,6}$	0	2.39	7.22	10.57	11.28	9.53	4	6.34	5.64	13.93
t_2	2.39	0	6.33	9.73	12.34	8	3.33	7.8	5.65	12.8
t_3	7.22	6.33	0	3.64	15.98	3.59	3.8	9.3	3.53	6.84
t_4	10.57	9.73	3.64	0	18.89	3.51	7.26	11.91	6.19	3.41
t_5	11.28	12.34	15.98	18.89	0	19.06	12.62	7.23	12.76	22.19
t_7	9.53	8	3.59	3.51	19.06	0	6.59	12.58	6.8	5.44
t_8	4	3.33	3.8	7.26	12.62	6.59	0	6.69	2.54	10.5
t_9	6.34	7.8	9.3	11.91	7.23	12.58	6.69	0	6.04	15.21
t_{10}	5.64	5.65	3.53	6.19	12.76	6.8	2.54	6.04	0	9.48
t_{11}	13.93	12.8	6.84	3.41	22.19	5.44	10.5	15.21	9.48	0

Tabela 24. Klasteri: $t_{1,6}$, t_2, t_3, t_4, t_5, t_7, t_8, t_9, t_{10}, t_{11}

Izrada Tabela 25-32 određuje klaserizaciju na odgovarajućim nivoima rastojanja. Na osnovu tih rezultata napravljen je Venov dijagram (Slika 6.1) i dendogram idealnog košarkaškog tima (Slika 6.2).



Slika 6.1. Venov dijagram



Slika 6.2 Dendrogram idealnog košarkaškog tima

Rezultat primenjene metode daje idealni tim KK Partizan za proteklu deceniju i to je:

1. Vule Avdalović (t_{10});
2. Nikola Peković (t_1);
3. Dušan Kecman (t_8)**;**
4. Novica Veličković (t_2)**;**
5. Nenad Krstić (t_6)**.**

Specijalizovani navijački sajt www.partizan.net izabrao je sličan tim kao i metoda kompleksnog povezivanja koja je ovde primenjena (Link: http://partizan.net/srpski/?p=5957):

1. Miloš Vujanić (on je uvršćen u tim kao najbolji igrač);
2. Miroslav Berić (nije bio posmatrani kandidat pošto nije igrao u navedenom periodu);
3. Dušan Kecman (t_8)**;**
4. Novica Veličković (t_2)**;**
5. Nenad Krstić (t_6)**.**

Obrada Primera B) pomoću MVSP (*Multi Variate Statistical Package*) programa

Postoji mnogo brži i jednostavniji način od ručnog računanja a to je pomoću MVSP programa. MVSP je program koji obavlja veliki broj numeričkih analiza koje se mogu koristiti u mnogim naučnim oblastima, od ekologije, geologije, sociologije do istraživanja tržišta. Koristi se na preko stotinu sajtova u preko 50 zemalja. Rezultati analiza urađenih u MVSP-u objavljivani su u mnogim poznatim novinama.

Kada su podaci uneti, biramo koordinatnu osu i očitavamo podatke sa dijagrama (svi dijagrami imaju i opciju zumiranja radi preglednosti). Rezultati klaster analize u vidu denrograma se pojavljuju automatski, i grafik se dalje može štampati na odgovarajućem uređaju ili sačuvati za dalju upotrebu.
Broj promenjljivih koje se mogu analizirati je ograničen samo brojem slobodne memorije na sistemu (ram memorije i memorije na hard disku) i dostiže broj od 2 milijarde promenljivih. Podaci se mogu ubacivati preko različitih formata uključujući *Excel*, *Quattro*, *xBase*, *Paradox*… Poseduje i prilagodljiv tulbar, editor za izmenu podataka, opciju vraćanja unazad ukoliko napravimo grešku.

Da bismo ubacili podatke u program možemo kopirati našu tabelu u Excel odakle je možemo prebaciti u program, da ne bi kucali svaku vrednost posebno (Slika 7).

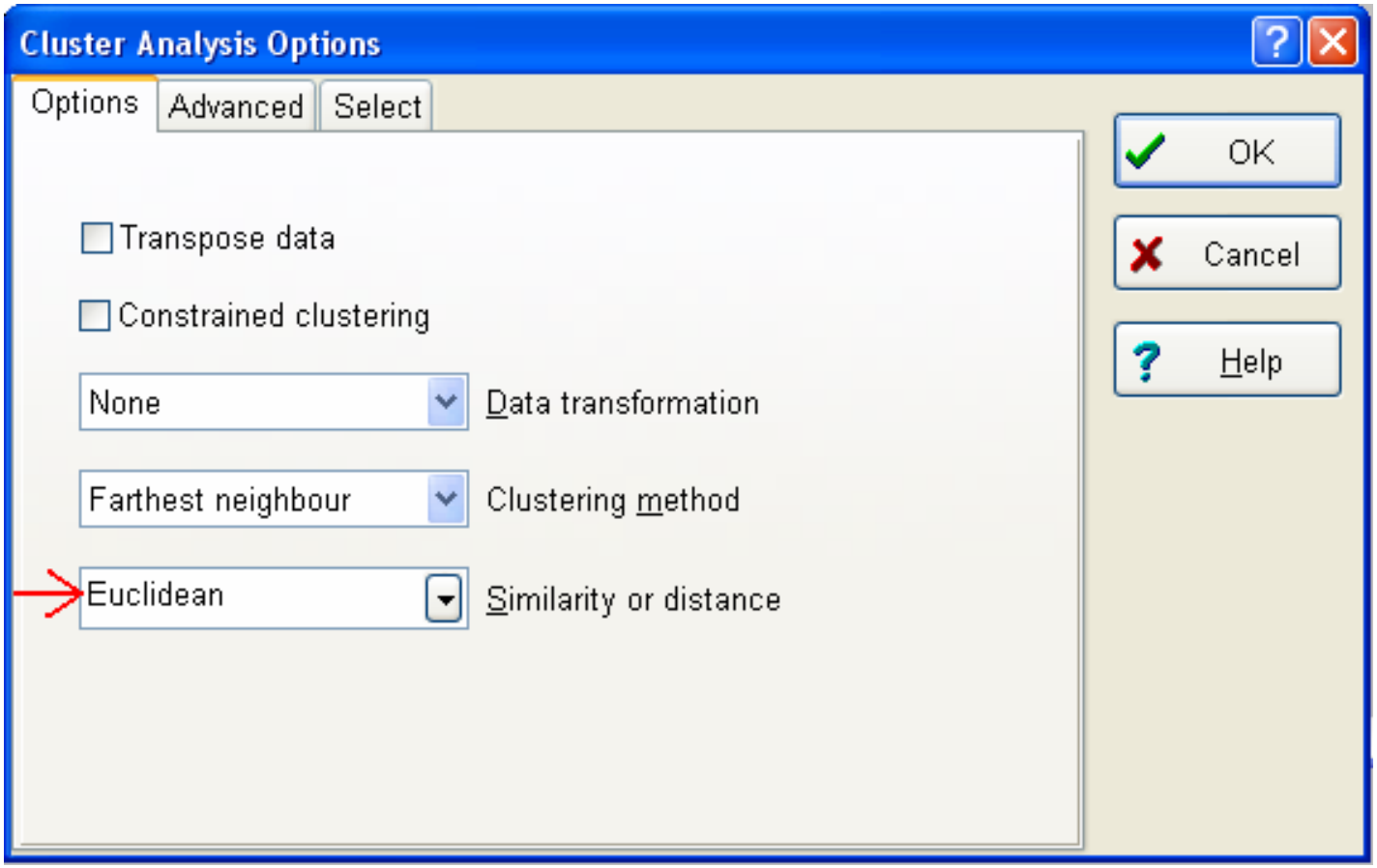
MVSP 3.21 - [Data Editor]

File Edit Data Analyses Graphs Options Window Help

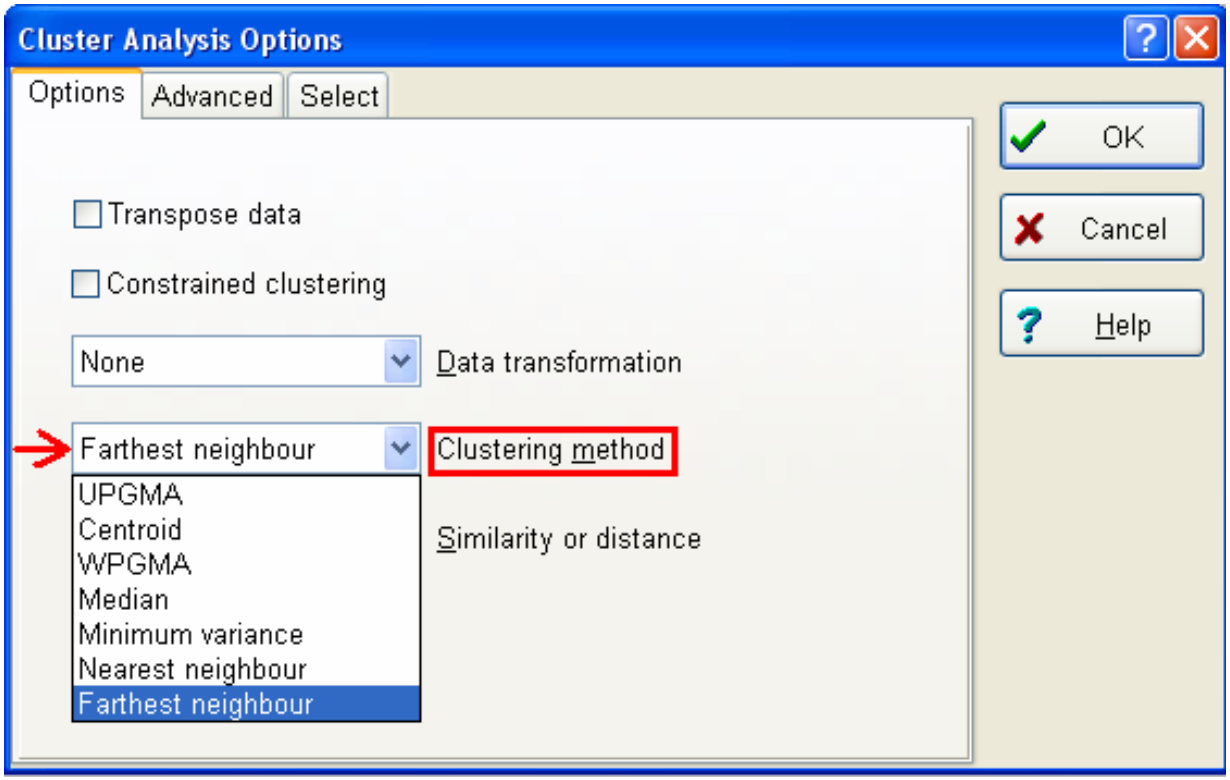
	A	B	C	D	E	F
1		k1	k2	k3	k4	k5
2	t1	16,400	6,800	0,600	0,400	0,400
3	t2	14,600	7,900	1,600	0,900	0,600
4	t3	10,000	3,600	1,000	1,200	0,800
5	t4	7,100	1,800	0,900	0,200	0,000
6	t5	25,800	3,000	3,200	1,000	0,000
7	t6	15,400	6,500	0,900	1,200	1,300
8	t7	7,000	5,300	1,100	0,400	0,200
9	t8	13,400	4,900	2,000	1,600	0,500
10	t9	19,000	1,500	1,300	0,500	0,000
11	t10	13,100	2,500	2,000	1,000	0,000
12	t11	3,800	0,900	1,200	0,500	0,000

Slika 7. Podaci u MVSP-u

Sledeći korak je Analyses → Cluster Analyses. Za rastojanje taksonomskih jedinica smo koristili Euklidovo rastojanje, a za klaster analizu, klaster metodu kompleksnog povezivanja (Slika 8. i 9.).

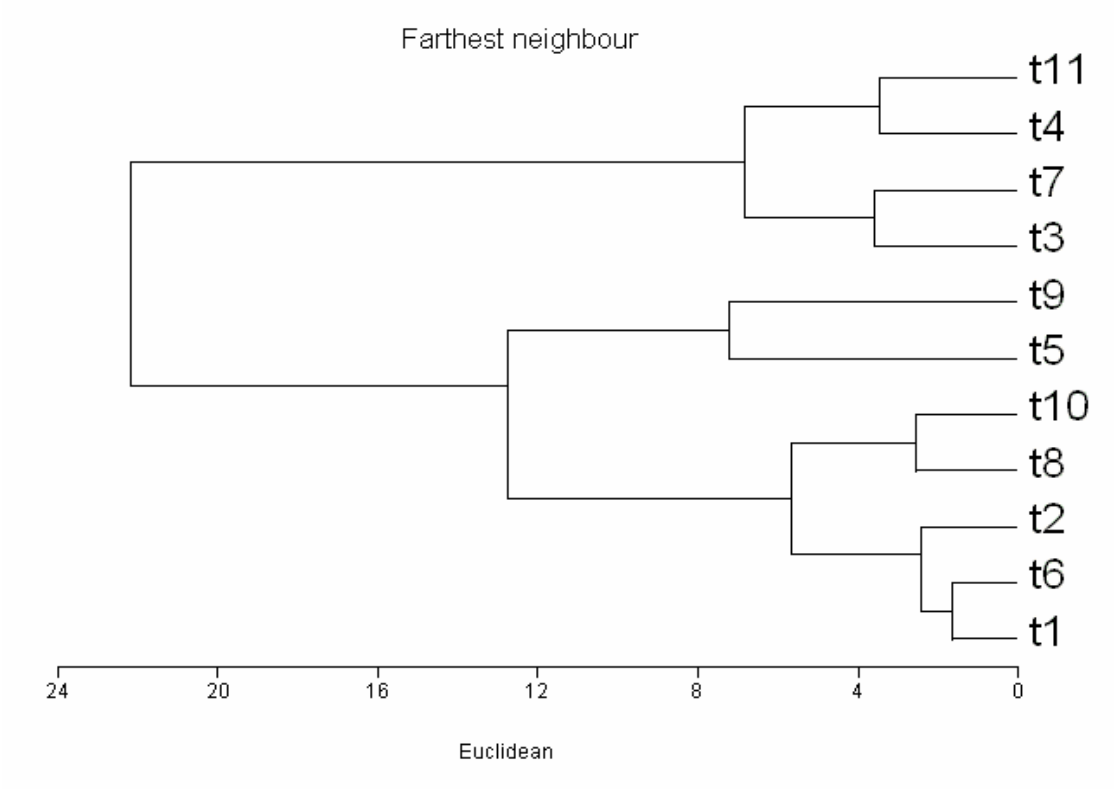


Slika 8. Biranje Euklidovog rastojanja



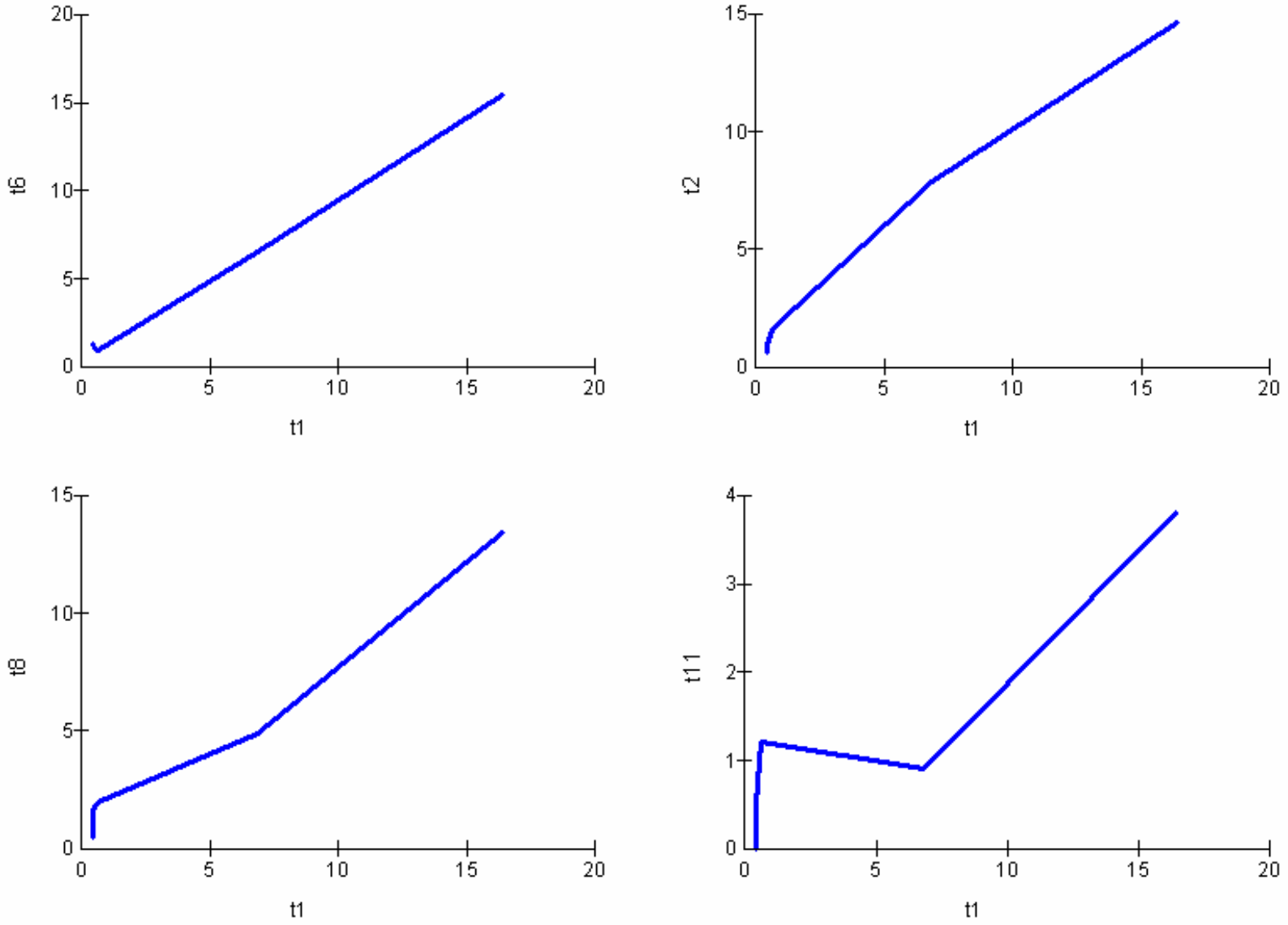
Slika 9. Klaster metoda kompleksnog povezivanja

Na osnovu unetih podataka dobijamo dendrogram (Slika 10) identičan prvom, na Slici 6.2.

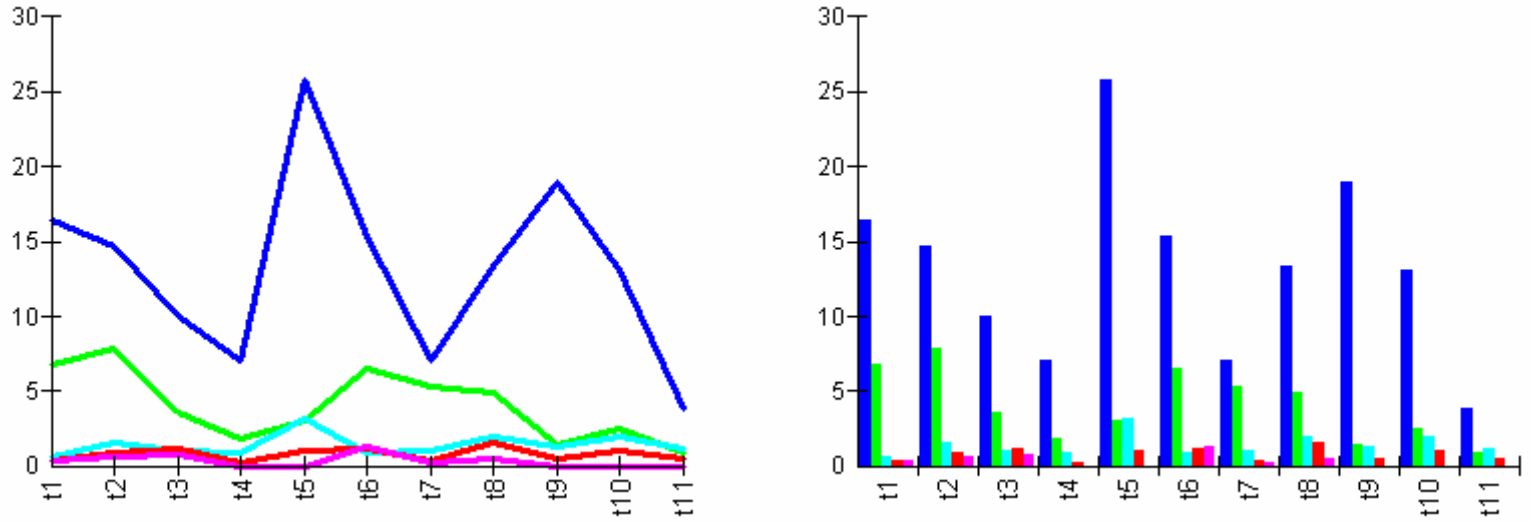


Slika 10. Dendrogram košarkaškog tima dobijen primenom MVSP programa.

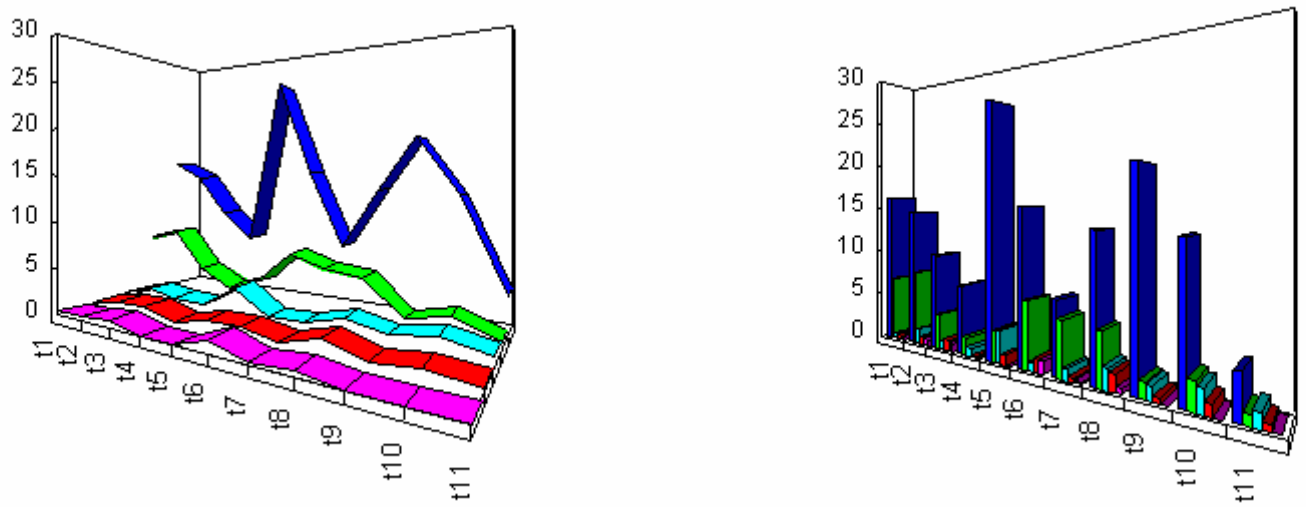
Možemo vršiti pojedinačno (svaka sa svakom) upoređivanje taksonomskih jedinica (Slika 11) ili grupno (Slike 12 i 13) u 2D ili 3D obliku. Iz samih dijagrama se tačno može videti koje taksonomske jedinice (igrači) imaju slične karakteristike.



Slika 11. Grafičko upoređivnje pojedinačnih taksonomskih jedinica



Slika 12. Grafičko upoređivnje svih taksonomskih jedinica- 2D



Slika 13. Grafičko upoređivnje svih taksonomskih jedinica- 3D

3.1.3. Klaster metoda proseka

Kod klaster metode proseka koeficijent sličnosti, odnosno rastojanja između dve grupe taksonomskih jedinica se računa kao prosek svih koeficienata, odnosno rastojanja, pri čemu se posmatraju odgovarajući koeficijenti između svake dve taksonomske jedinice od kojih je prva iz prve, a druga iz druge grupe.

Sledi primena klaster metode proseka na **ilustrativni primer** iz Tabele1.

Prvi korak je isti kao i kod prethodnih metoda - spajanje elemenata sa najmanjim rastojanjem – t_1 i t_6. Rastojanje te grupe od svake druge taksonomske jedinice se računa kao prosek rastojanja t_1 od te jedinice i rastojanja t_6 od te jedinice. Na primer, rastojanje te grupe od t_2 je jednako proseku 0.7 i 0.8, što je 0.75. Rastojanje te grupe od t_3 je 0.65, od t_4 i od t_5 je 0.25 (Tabela 33).

	t_{16}	t_2	t_3	t_4	t_5
t_{16}	0	0.75	0.65	0.25	0.25
t_2	0.75	0	0.2	0.4	0.3
t_3	0.65	0.2	0	0.9	08
t_4	0.25	0.4	0.9	0	0.5
t_5	0.25	0.3	0.8	0.5	0

Tabela 33.

Sledeće najmanje rastojanje je 0.2 između t_2 i t_3, pa se sada na nivou 0.2 spajaju u grupu t_2 i t_3. Rastojanje između grupa dobijenih u ovom koraku računa se kao prosek rastojanja svake taksonomske jedinice iz prve grupe sa svakom taksonomskom jedinicom iz druge grupe (Tabela 1):

$$d(t_{16}, t_{23}) = \frac{0.7 + 0.6 + 0.8 + 0.7}{4} = 0.7.$$

Ostala rastojanja data su u Tabeli 34.

	t_{16}	t_{23}	t_4	t_5
t_{16}	0	0.7	0.25	0,25
t_{23}	0.7	0	0.65	0.55
t_4	0.25	0.65	0	0.5
t_5	0.25	0.55	0.5	0

Tabela 34.

Najmanje rastojanje je sada 0.25, i to između t_4 i grupe t_{16}, kao i između t_5 i grupe t_{16}. Kako je rastojanje t_4 i t_5 jednako 0.5, to se obično se bira element sa manjim indeksom, tako da se na nivou 0.25 spajaju t_4 i t_{16}. Rastojanja te grupe od ostalih data su u Tabeli 35.

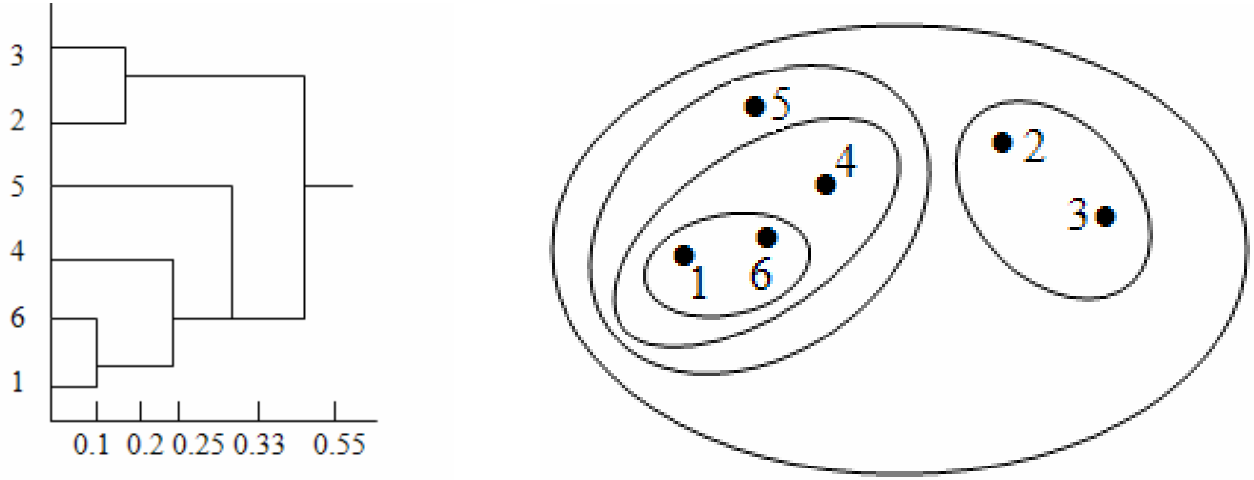
	t_{146}	t_{23}	t_5
t_{146}	0	0.683	0.333
t_{23}	0.683	0	0.55
t_5	0.333	0.55	0

Tabela 35.

U sledećem koraku se spaja t_5 sa grupom t_{146} na stepenu rastojanja 0.333. Rastojanje grupe $\{t_1, t_4, t_5, t_6\}$ od preostale grupe se računa kao prosek svih međusobnih rastojanja:

$$d\left(t_{1456}, t_{23}\right) = \frac{0.7+0.6+0.4+0.9+0.3+0.8+0.8+0.7}{8} = 0.65.$$

Rezultati primene klaster metode proseka predstavljeni su dendrogramom na Slici 14.



(a) (b)

Slika 14.

Primena klaster metoda proseka na izbor virtuelnog radnog tima

U poslovnoj praksi često je prisutan timski rad. Radna organizacija formira tim iz raspoloživih ili novoprimljenih kadrova. Bitno je da novoformirani tim ima što sličnije karakteristike kako bi optimalno funkcionisao. Karakteristike koje su utvrđenje prilikom testiranja/ispitivanja mogu biti brojne i one predstavljaju varijable. U ovom slučaju to su: k_1, k_2,..., k_8. Na osnovu ovih karakteristika, primenom klaster metode proseka, može se doći do kompaktnog tima.

Uzećemo da je:

***k*₁ – broj godina zaposlenog** i staviti:

1, ako zaposleni ima do 20 godina;
2, ako ima 21 do 30;
3, ako ima od 31 do 40;
4, ako ima od 41 do 50;
5, ako ima od 51 do 60.

***k*₂ – obrazovanje zaposlenog** čije vrednosti mogu biti:

1, ako zaposleni ima završenu osnovnu školu,
2, ako ima završenu srednju školi,
3, višu školu,
4, fakultet i
5, ako zaposleni ima doktorat.

***k*₃ – radno iskustvo** gde:

1, predstavlja zaposlenog sa iskustvom od 1 do 2 godine,
2, zaposlenog od 2 do 5 godina,
3, zaposlenog od 5 do 10 godina,
4, od 10 do 20 i
5, preko 20 godina radnog iskustva.

***k*₄ –znanje rada na računaru** gde se sa:

1, označava zaposleni bez znanja rada na računaru,
2, zaposleni koji se slabo snalazi u radu na računaru,
3, zaposleni koji se dobro snalazi i
4, zaposleni poseduje odlično znanje rada na računaru.

***k_5* –znanje engleskog jezika** pri čemu se sa:

1, obeležava zaposleni koji ne govori engleski jezik,
2, slabo znanje,
3, dobro i
4, odlično znanje engleskog jezika.

***k_6* – komunikativnost zaposlenog** gde:

1 predstavlja zaposlenog sa slabo izraženim komunikativnim sposobnostima,
2, sa dobrim i
3, sa odličnim komunikativnim sposobnostima.

***k_7* – karakteriše sklonost zaposlenog ka timskom radu** i uzećemo:

1, ako zaposleni nema želju za radom u timu,
2, ako zaposleni može da radi u timu i
3, ako zaposleni odlično funkcioniše u timu.

***k_8* – iskustvo na sličnim poslovima** i sa:

1, označava zaposlenog bez iskustva,
2, zaposleni sa iskustvom od 1 do 3 godine,
3, zaposleni sa iskustvom od 3 do 5 godina i
4, zaposleni sa iskustvom preko 5 godina.

Neka je broj zainteresovanih kandidata za predstojeći projekat 20, a potrebno je izabrati 5 koji će najbolje funkcionisati kao tim.

Početna Tabela 36 dobijena je nakon sprovedenog ispitivanja/testiranja zainteresovanih i u nju se smeštaju karakteristike koje su utvrđene u tom procesu. Sa t_n su označeni kandidati ($n = 1,...,20$), a sa k_i ($i = 1,...,8$) njihove karakteristike na osnovu kojih će se vršiti grupisanje.

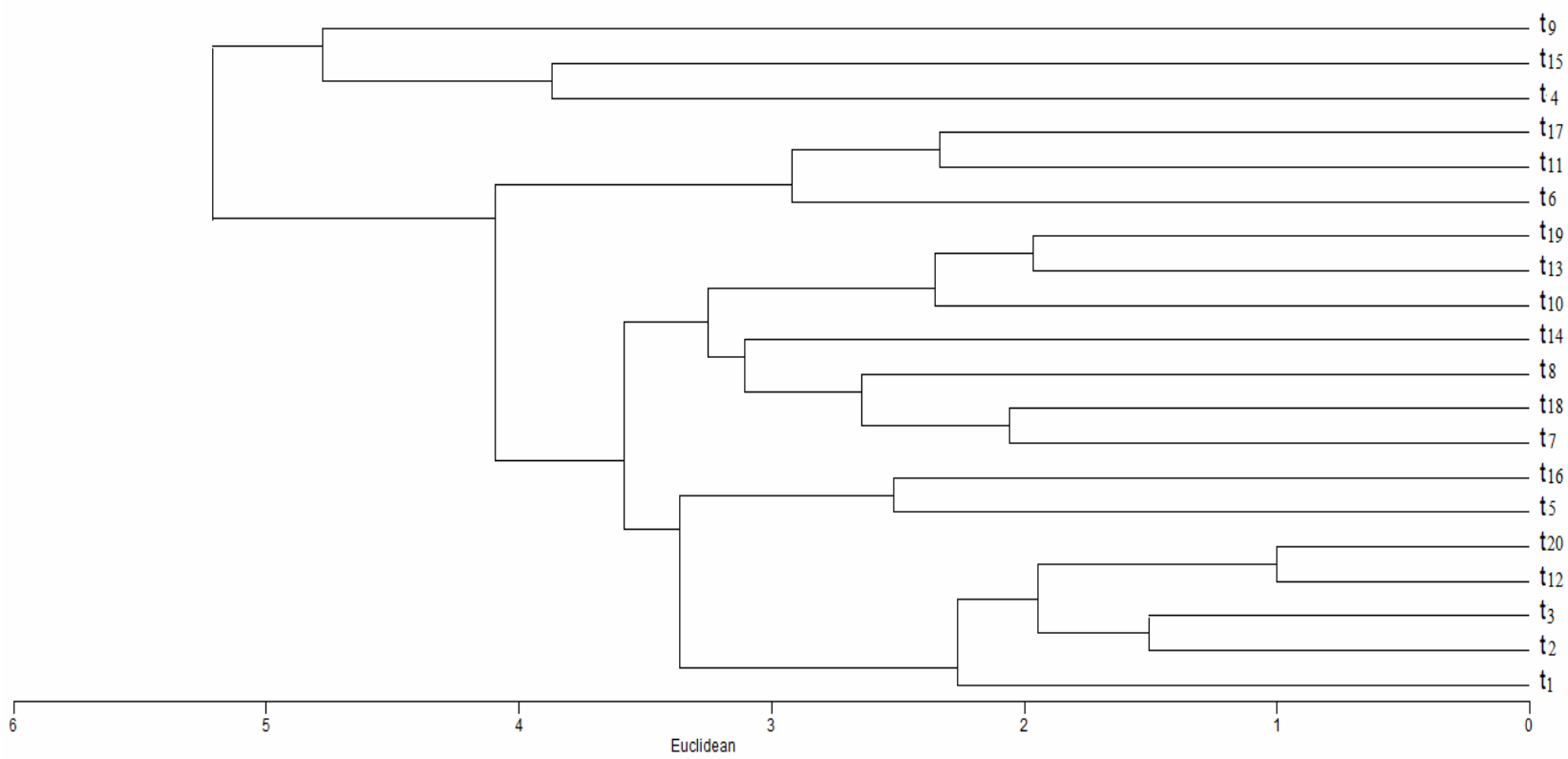
	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}
k_1	2	2	2	5	3	1	2	3	5	4	1	2	4	2	5	4	1	2	3	2
k_2	1	2	2	4	3	1	3	4	1	3	2	3	2	4	5	2	1	4	2	3
k_3	3	2	2	5	2	1	2	4	4	3	2	3	5	3	3	1	2	3	4	3
k_4	2	4	2	3	3	3	1	1	3	1	3	3	2	3	2	3	2	2	2	3
k_5	3	2	3	4	2	4	2	2	2	4	3	2	3	4	2	3	2	3	4	3
k_6	3	2	2	2	1	2	1	3	1	2	1	2	2	2	1	2	1	1	1	2
k_7	2	2	2	3	4	1	1	2	2	3	4	2	3	3	4	3	3	1	3	1
k_8	3	4	4	4	4	1	1	2	4	3	1	3	2	2	1	3	2	2	2	3

Tabela 36.

Matrica Euklidovog rastojanja data je u Tabeli 37 i može se dobiti na osnovu postupka opisanog u prethodnom delu (klaster metoda proseka) ili korišćenjem nekog softvera za klaster analizu. Upotreba softvera za obradu podataka je bolje rešenje jer se izbegavaju greške u računu i znatno se skraćuje vreme analize. U ovom slučaju korišćen je program Multi-Variate Statistical Package (MVSP).

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}
t_1	0																			
t_2	2,3	0																		
t_3	1,9	1,5	0																	
t_4	5,3	5,3	5	0																
t_5	3,9	2,8	2,6	4,3	0															
t_6	3,3	3,6	3,6	7,3	5,1	0														
t_7	3,3	3,7	3,2	6,1	4,2	3,6	0													
t_8	3,7	4,3	3,8	4,4	4,3	5,5	2,9	0												
t_9	3,8	4	4,1	4,4	4,2	6,2	5	4,5	0											
t_{10}	3,4	4,1	3,2	3,4	3,4	5,1	3,5	2,9	3,9	0										
t_{11}	3,7	4	3,5	6,1	3,8	3	3,5	4,8	6,1	4,2	0									
t_{12}	2,4	1,6	1,9	4,6	2,9	4,1	3	3	3,9	3,5	4	0								
t_{13}	3,4	4,4	4,1	3,5	4,2	5,6	4,1	2,9	2,9	2,7	4,7	3,7	0							
t_{14}	3,6	3,5	3,3	4,3	3,4	4,1	3,4	3,1	5,5	3,2	3	2,8	3,9	0						
t_{15}	5,8	5,5	5,2	3,9	3,9	6,7	4,5	3,5	5,2	3,2	5,3	4,7	4	3,9	0					
t_{16}	3,5	2,8	2,5	4,7	2,5	4,2	4,1	4,8	3,9	3	4,1	3,4	4,4	3,9	4,5	0				
t_{17}	2,7	3,1	2,7	6,5	3,7	2,9	2,9	4,7	5	4,3	2,3	3,3	4,5	4	5,9	3,9	0			
t_{18}	3,5	3,6	3,1	4,9	4	4,5	2,1	2,4	4,9	3	4,2	2,3	3,9	2,8	4,3	4,3	3,8	0		
t_{19}	2,6	3,7	3	3,6	3,5	4,4	3,6	3,3	3,7	2	3,4	3,1	2	2,9	4,2	3,5	3,4	3,2	0	
t_{20}	2,4	2	2,3	4,7	3,5	3,6	2,9	3,1	4,2	3,5	3,9	1	3,7	2,5	4,9	3,5	3,5	2,1	3	0

Tabela 37.



Slika 15. Dendrogram kandidata dobijen iz Tabele 37.

Na Slici 16 dat je prikaz rastojanja između kandidata na osnovu kojih je i kreiran dendrogram.

Node	Group 1	Group 2	Dissimil.	Objects in group
1	t12	t20	1,0	2
2	t2	t3	1,5	2
3	Node 2	Node 1	1,9	4
4	t13	t19	2,0	2
5	t7	t18	2,1	2
6	t1	Node 3	2,3	5
7	t11	t17	2,3	2
8	t10	Node 4	2,4	3
9	t5	t16	2,5	2
10	Node 5	t8	2,6	3
11	t6	Node 7	2,9	3
12	Node 10	t14	3,1	4
13	Node 12	Node 8	3,3	7
14	Node 6	Node 9	3,4	7
15	Node 14	Node 13	3,6	14
16	t4	t15	3,9	2
17	Node 15	Node 11	4,1	17
18	Node 16	t9	4,8	3
19	Node 17	Node 18	5,2	20

Slika 16.

Primenom klaster metode proseka utvrđeno je da će kao tim najbolje funkcionisati kandidati t_1, t_2, t_3, t_{12}, t_{20}.

Karakteristike metoda klaster analize i njihov izbor

Osobine pojedinih metoda

Različite klaster metode primenjene na iste podatke često daju različite rezultate, tj. različite klasifikacije. Ni jedna metoda nije najbolja u svim primenama, i svaka od njih ima svoje prednosti i nedostatke. U nastavku su navedene neke osobine pojedinih metoda, kao i njihove prednosti i nedostaci.

1. osobina:

Klaster metoda prostog povezivanja i klaster metoda kompleksnog povezivanja daju iste rezultate u odnosu na monotone transformacije matrica sličnosti i rastojanja. To znači da je hijerarhijska podela na klastere uvek ista (iako se taksonomske jedinice mogu grupisati u isti klaster na različtim nivoima sličnosti). Ostale pomenute metode nemaju ovu osobinu.

2. osobina:

Nedostatak klaster metode prostog povezivanja je osobina tzv. lančanog povezivanja, pa dve relativno udaljene grupe imaju između sebe lanac bliskih taksonomskih jedinica. Tada će ova metoda povezati te dve grupe na malom stepenu rastojanja, iako su one kao grupe možda udaljenije nego svaka od njih od nekih drugih grupa. Ova osobina može ponekad da dovede do pogrešne klasifikacije taksonomskih jedinica.

U vezi sa ovim je jedan nedostatak svih hijerarhiskih metoda. Može da se dogodi da se zbog nekog povezivanja na nižem nivou, na višem nivou zajedno se nađu taksonomske jedinice koje ne bi trebalo da su zajedno u istoj grupi. Zato je ponekad potrebno izvršiti premeštanje nekih taksonomskih jedinica radi pravilnije klasifikacije.

3. osobina:

Klaster metode prostog povezivanja, kompleksnog povezivanja i proseka imaju osobinu da su **nivoi spajanja grupa u prethodnom koraku uvek manji od onih na sledećem**. Ovo nije uvek slučaj kod centroidne klaster metode, i to može biti problem kod pravljenja dendrograrna i utvrđivanja klastera.

Procena rezultata dobijenih kaster analizom

Postoji više razloga zbog kojih taksonom, prilikom primene metoda klaster analize, treba da bude veoma pažljiv.

Prvo, u zavisnosti od izbora metode klaster analize i izbora mere sličnosti ili rastojanja koje se primenjuju u računu, **često se dobijaju različite klasifikacije.** Potrebno je poznavati suštinu svake metode, da bi dobijeni rezultati mogli da se protumače i da se usvoji rezultat dobijen metodom koja najviše odgovara prirodi samog problema. Često su najbolji oni rezultati koji se poklapaju sa intuicijom taksonoma. Tada matematička metoda služi samo za potvrdu hipoteze koja je unapred postavljena. Neki od zaključaka su da metod prostog povezivanja često daje loše rezultate, a metoda proseka, kao i metod Warda dobre.

Druga činjenica je da čak i **ako ne postoji nikakva prirodna podela grupe** taksonomskih jedinica na manje grupe, svaka od ovih metoda će te grupe ipak podeliti. Tako da taksonom može da dođe do pogrešnih zaključaka da postoji podela taksonomskih jedinica na manje grupe, a u prirodi takva podela, a ni bilo kakva druga, ne postoji.

Postoji nekoliko metoda uz pomoć kojih se proverava da li je podela koja je dobijena nekom od metoda klaster analize veštačka, ili prihvatljiva.

Pre svega, postoji više statističkih metoda za proveru hipoteze da se taksonomske jedinice na prirodan način dele na klastere. Te metode kao podatke koriste dobijene matrice sličnosti ili rastojanja, ili same vrednosti karaktera. Ako se nakon provere te hipoteze dobije rezultat da je velika verovatnoća da ne postoji podela taksonomskih jedinica, tada ni nema smisla koristiti metode klaster analize za pravljenje klastera.

Taksonom na subjektivan način odlučuje o tome koju će particiju iz dendrograma odabrati za klasifikaciju. Najjednostavnije je uoči skok na dendrogramu na kome postoji velika razlika izmedu susednih nivoa spajanja i izabere se odgovarajuća podela.