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 4

PREZENTACIJA PRIMENE JOŠ NEKIH
METODA KLASTER ANALIZE
(centroidna metoda, metoda Ward-a, metoda k-unutrašnjih centara i metoda fazi k-unutrašnjih centara)
NA DEFINISANJE GRUPA HOMOGENIH OBJEKATA
(individue, naselja ili proizvodi)

3.1.4. Centroidna klaster metoda

Centroidna klaster metoda se, za razliku od prethodnih, primenjuje direktno na vrednosti karaktera, a ne na već izračunate matrice rastojanja ili sličnosti. Prvo se spajaju u grupu taksonomske jedinice sa najmanjim rastojanjem. U sledećem koraku se računa srednja (prosečna) vrednost karaktera za tako dobijenu grupu, i ta vrednost se uzima kao vrednost karaktera za grupu. Vektor sa tim, prosečnim vrednostima naziva se **centroid.** Dalje se računaju rastojanja izmedu grupe, odnosno centroida, i ostalih taksonomskih jedinica. Ponovo se spajaju grupe, odnosno taksonomske jedinice sa najmanjim rastojanjem, pa se opet računa prosečna vrednost karaktera za tako dobijene grupe, tj. odreduje se novi centroid itd. Ova metoda koristi se za kvanitativne karaktere.

Demostrativni Primer primene centroidne klaster metode. Neka su na pet taksonomskih jedinica posmatrana dva numerička (kvantitativna) karaktera. Za izračunavanje različitosti se koristi Apsolutna (Manhattan) metrika.

	k_1	k_2
t_1	1	1
t_2	3	3
t_3	2	1
t_4	5	5
t_5	4	6

Tablica 38.

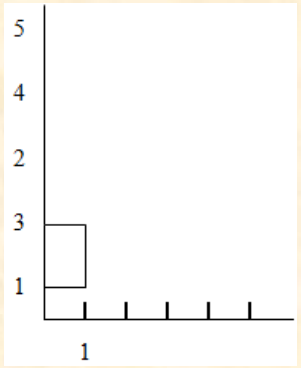
	t_1	t_2	t_3	t_4	t_5
t_1	0	4	1	8	8
t_2	4	0	3	4	*4*
t_3	1	3	0	7	7
t_4	8	4	7	0	2
t_5	8	4	7	2	0

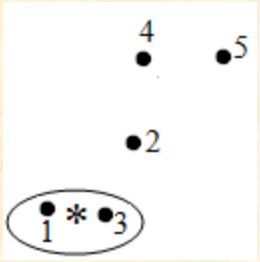
Tablica 39.

Kako je najmanje rastojanje između taksona t_1 i t_3 to se one spajaju u novu grupu. Centroid novodobijene grupe je

$$c_1 = \frac{t_1 + t_3}{2} = \frac{(1,1)+(2,1)}{2} = (1.5,1).$$

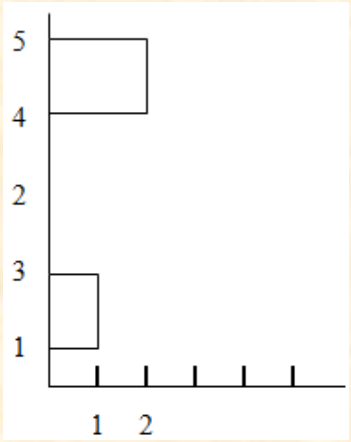
Sada se ponavlja opisani postupak na četiri taksonomske jedinice i jednu grupu koju predstavlja centroid c_1.

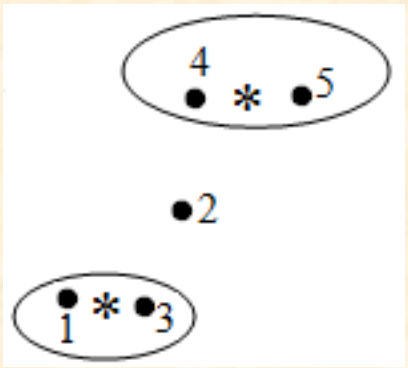




	c_1	t_2	t_4	t_5
c_1	0	3.5	7.5	7.5
t_2	3.5	0	4	4
t_4	**7.5**	4	0	2
t_5	**7.5**	4	2	0

1. iteracija (korak) centrpidnog klaster metoda

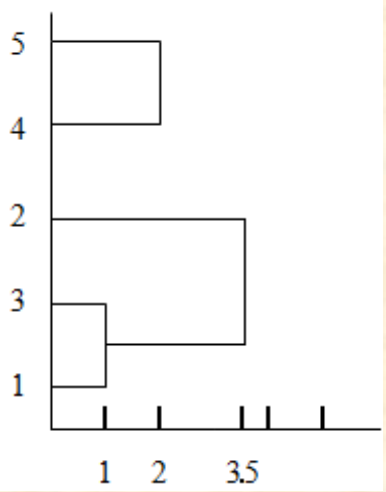


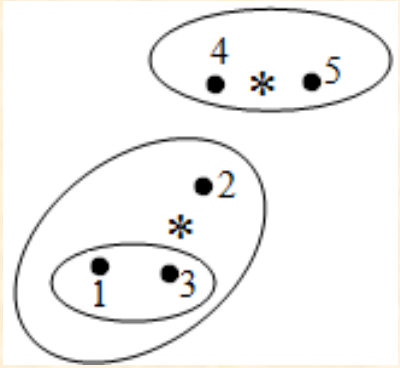


$$c_2 = \frac{t_4 + t_5}{2} = \frac{(5,5)+(4,6)}{2} = (4.5,5.5)$$

	c_1	t_2	c_2
c_1	0	3.5	7.5
t_2	3.5	0	4
c_2	7.5	4	0

2. iteracija centroidnog klaster metoda



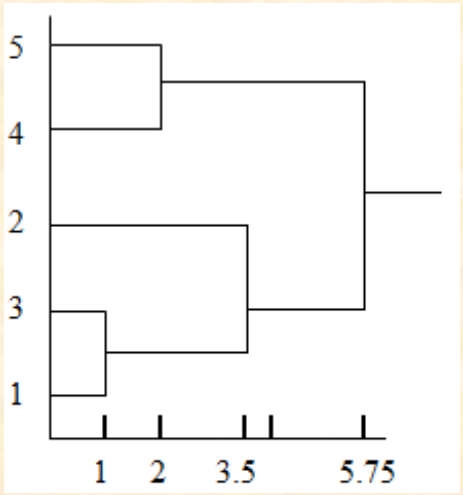


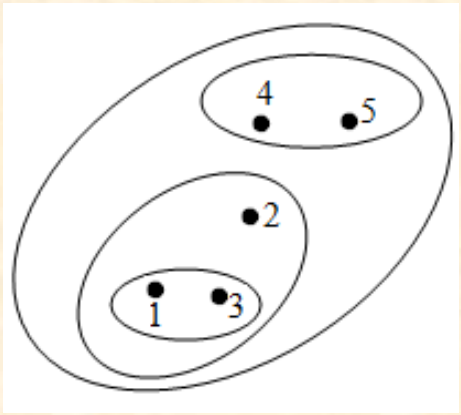
$$c_3 = \frac{t_1 + t_3 + t_2}{3} = (2, 1.6)$$

	c_3	c_2
c_3	0	5.75
c_2	5.75	0

3. iteracija centroidnog klaster metoda

Sledi dendrogram i odgovarajući prikaz skupova nakon primene metode.





Slika 17.

Primer: Primena centroidne klaster metode na izbor virtuelnog tima

Današnja poslovna politika uglavnom je zasnovana na timskom radu. Zato se mora posvetiti naročita pažnja formiranju radnih timova. Karakteristike zaposlenih prikupljaju se kroz njihovo testiranje/ispitivanje. Osobine i sposobnosti članova budućeg virtuelnog tima propratićemo preko promenljivih $k_1, k_2, ..., k_8$; gde je:

k_1 – broj godina zaposlenog i može uzeti vrednosti 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 od 5 do 10, 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, sa 2 zaposleni koji se slabo snalaži 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, 4 oslično znanje engleskog jezike.

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 timski duh zapolenog i može uzeti vrednosti 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 se označava zaposleni bez iskustva, 2 zaposleni sa iskustvom od 1 do 3 godine, sa 3 od 3 do 5 godina i sa 4 za preko 5 godina.

Početna Tabela 40 dobija se posle sprovedenog ispitivanja/testiranja 20 zaposlenih zainteresovanih za rad u timu od 5 članova, **koji ćemo izabrati centroidnom metodom.**

	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	2	1	3	2	2	3	2	4	4	2	2	4	2	2	3	4	3	3	2
k_5	2	4	4	3	4	2	2	2	2	2	2	3	4	3	4	2	4	3	4	3
k_6	2	3	1	3	1	1	2	2	1	2	1	3	2	2	3	2	2	3	2	1
k_7	4	2	4	3	4	3	3	4	2	3	4	3	2	4	3	2	2	4	2	4
k_8	2	2	2	1	4	1	4	2	3	4	1	4	3	2	3	4	2	2	2	3

Tabela 40.

Matrica rastojanja dobijena korišćenjem Apsolutnog (Manhattan) rastojanja data je u Tabeli 41 i može se dobiti na osnovu postupka opisanog u prethodnom delu (demostrativni primer) ili korišćenjem nekog softvera za klaster analizu. U ovom slučaju **koristićemo program Multi-Variate Statistical Package (MVSP) i SYSTAT 13.**

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

Tabela 41.

Slika 18. Dendogram iz Tabele 41.

U tabelama 42 i 43 dat je prikaz rastojanja između zaposlenih na osnovu kojih je i kreiran dendrogram.

Spojeni klasteri		Na rastojanju	Broj članova
KANDIDAT18	KANDIDAT14	2,000	2
KANDIDAT11	KANDIDAT6	3,000	2
KANDIDAT18	KANDIDAT8	3,500	3
KANDIDAT18	KANDIDAT20	3,556	4
KANDIDAT18	KANDIDAT1	3,500	5
KANDIDAT10	KANDIDAT7	4,000	2
KANDIDAT12	KANDIDAT10	4,000	3
KANDIDAT12	KANDIDAT18	3,818	8
KANDIDAT19	KANDIDAT13	4,000	2
KANDIDAT12	KANDIDAT5	4,547	9
KANDIDAT3	KANDIDAT2	5,000	2
KANDIDAT3	KANDIDAT12	3,954	11
KANDIDAT3	KANDIDAT11	5,056	13
KANDIDAT19	KANDIDAT3	5,716	15
KANDIDAT19	KANDIDAT17	5,493	16
KANDIDAT19	KANDIDAT16	5,703	17
KANDIDAT19	KANDIDAT15	6,699	18
KANDIDAT19	KANDIDAT9	7,086	19
KANDIDAT19	KANDIDAT4	7,150	20

Tabela 42.

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

Tabela 43.

Primenom centroidne klaster metode utvrđeno je da će kao tim najbolje funkcionisati kandidati: t_{14}, t_{18}, t_8, t_{20} i t_1.

Zaključak

Nedostatak centroidne klaster metode je u tome što se početna udaljenost dva klastera može smanjiti između dva sukcesivna koraka analize. **Klasteri spojeni u kasnijim fazama su više različiti nego oni spojeni u ranijim koracima.** U centroidnoj metodi udaljenost izmedju dva klastera je udaljenost izmedju njihovih centroida. Centroidni klaster znači srednju vrednost posmatranih varijabli u klaster promenljivama. Po ovoj metodi, svaki put kada su pojedinci grupisani, centroid je preračunat. Postoji promena u klaster centroidu svaki put kada se jedinka ili grupa jedinki doda postojećem klasteru. Ove metode su najpopularnije kod biologa, ali mogu napraviti nered i često daju zbunjujuće rezultate. Konfuzija nastaje zbog inverzije ili obrnutosti koja se javlja kada izmerena udaljenost izmedju jednog para centroida bude manja od nekog ranijeg merenja.

Primer –Taksonomija teritorijalnih lokacija po osnovu klime:

A) Klaster metodom proseka

B) Centroidnom klaster metodom

U ovom primeru posmatrane su neki meteorološki parametri od 14.09.2012.godine za više lokacija/gradova u Srbiji. Podaci su preuzeti sa Automatske MEteorološke Stanice – **AMES**. AMES omogućava da se, bez angažovanja ljudi, neprekidno 24 časa, 365 dana u godini, pet puta u sekundi (tzv. merenje u realnom vremenu) dobiju podaci o meteorloškim parametrima koji se mere. Podaci dobijeni merenjem se mogu automatski slati do udaljenih računara i/ili servera i tako biti dostupni na internetu (Slika 19).

Станица	Температура (°C)	Притисак (hPa)	Правац ветра	Брзина ветра (m/s)	Влажност (%)	Топлотни индекс	Симбол	Опис времена
Палић	19	998.6	N	2	63	21		Претежно облачно
Сомбор	16	1001.0	NE	3	88	19		Облачно
Нови Сад	16	1001.1	S	1	82	19		Сумаглица
Зрењанин	21	1001.2	SW	2	56	23		Претежно облачно
Кикинда	23	1000.6	NW	2	49	25		Претежно облачно
Б. Карловац	20	1000.2	NE	1	60	22		Облачно
Лозница	14	997.3	-	тихо	88	16		Сумаглица
С. Митровица	15	1001.2	SW	1	88	18		Облачно
Ваљево	15	990.6	NW	1	77	17		Облачно
Београд	15	996.0	SW	2	82	17		Облачно
Крагујевац	19	989.7	NE	2	59	21		Претежно облачно
С. Паланка	20	996.8	SW	3	60	22		Претежно облачно
Црни Врх	17	895.3	SE	4	63	18		Облачно
Неготин	26	1004.6	SW	2	47	29		Облачно
Златибор	10	894.0	E	1	100	11		Магла
Сјеница	18	892.7	SE	5	68	20		Облачно
Пожега	15	974.7	SE	1	82	17		Облачно
Краљево	18	985.5	NW	1	68	20		Претежно облачно
Копаоник	12	824.4	SE	5	100	14		Облачно
Куршумлија	23	965.3	NE	2	49	25		Облачно
Крушевац	21	990.0	NE	1	52	23		Претежно облачно

Slika 19.

Početna Tabela 44. je dobijena na osnovu prikazane Slike19. Sa t_n ($n = 1,...,21$) su označeni gradovi, a sa k_i ($i = 1,...,6$) njihove karakteristike na osnovu kojih će se vršiti grupisanje:

k_1 – temperatura [°C];

k_2 – pravac vetra;

k_3 – brzina vetra [m/s];

k_4 –vlažnost [%];

k_5 – toplotni indeks;

k_6 – opis vremena.

	k_1	k_2	k_3	k_4	k_5	k_6
t_1	19	1	2	63	21	1
t_2	16	2	3	88	19	2
t_3	16	4	1	82	19	3
t_4	21	5	2	56	23	1
t_5	23	3	2	49	25	1
t_6	20	2	1	60	22	2
t_7	14	0	0	88	16	3
t_8	15	5	1	88	18	2
t_9	15	3	1	77	17	2
t_{10}	15	5	2	82	17	2
t_{11}	19	2	2	59	21	1
t_{12}	20	5	3	60	22	1
t_{13}	17	6	4	63	18	2
t_{14}	26	5	2	47	29	2
t_{15}	10	7	1	100	11	4
t_{16}	18	6	5	68	20	2
t_{17}	15	6	1	82	17	2
t_{18}	18	3	1	68	20	1
t_{19}	12	6	5	100	14	2
t_{20}	23	2	2	49	25	2
t_{21}	21	2	1	52	23	1

Tabela 44.

U Tabeli 44, na mestima za karakteristike: k_2 – pravac vetra, k_3 – brzine vetra, k_6 – opis vremena, unete su sledeće vrednosti:

k_2 – pravac vetra:

–	N	NE	NW	S	SW	SE	E
0	1	2	3	4	5	6	7

k_3 – brzina vetra:

Tiho	0

k_6 – opis vremena:

Pretežno oblačno	1
Oblačno	2
Sumaglica	3
Magla	4
Slaba kiša	5

Vrednosti iz prethodnih tabela biće korisšćene u obe metode: u klaster metodi proseka i centroidnoj klaster metodi. Tako ćemo moći na kraju da uporedimo rezultate ovih dveju metoda.

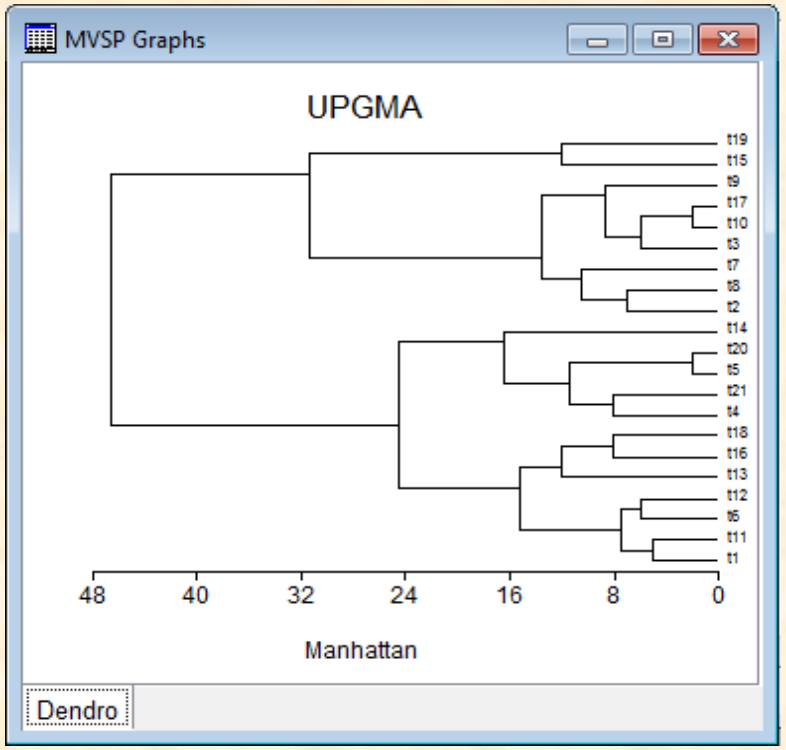
A) Klaster metoda proseka – klimatska taksonomija lokacija/gradova

Primenom klaster metode proseka i Apsolutnog (Manhattan) rastojanja dobijena je Matrica rastojanja prikazana Tabelom 45. Pri tome korišćeni su programi **Multi-Variate Statistical Package (MVSP) i SYSTAT 13.**

	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_{21}
t_1	0,000																				
t_2	33,000	0,000																			
t_3	30,000	11,000	0,000																		
t_4	15,000	46,000	39,000	0,000																	
t_5	24,000	55,000	50,000	13,000	0,000																
t_6	8,000	37,000	32,000	11,000	20,000	0,000															
t_7	40,000	11,000	16,000	55,000	64,000	44,000	0,000														
t_8	38,000	7,000	10,000	45,000	58,000	40,000	10,000	0,000													
t_9	26,000	17,000	10,000	37,000	46,000	28,000	18,000	14,000	0,000												
t_{10}	32,000	13,000	6,000	39,000	52,000	36,000	16,000	8,000	8,000	0,000											
t_{11}	5,000	36,000	33,000	10,000	19,000	5,000	45,000	41,000	29,000	35,000	0,000										
t_{12}	10,000	39,000	34,000	7,000	20,000	6,000	50,000	40,000	32,000	34,000	7,000	0,000									
t_{13}	13,000	32,000	27,000	20,000	33,000	17,000	41,000	31,000	23,000	25,000	16,000	13,000	0,000								
t_{14}	36,000	65,000	58,000	21,000	12,000	30,000	74,000	64,000	56,000	58,000	31,000	28,000	39,000	0,000							
t_{15}	66,000	35,000	36,000	73,000	86,000	68,000	30,000	28,000	40,000	34,000	69,000	68,000	57,000	92,000	0,000						
t_{16}	16,000	29,000	24,000	23,000	36,000	20,000	40,000	30,000	22,000	24,000	19,000	16,000	9,000	42,000	56,000	0,000					
t_{17}	34,000	15,000	6,000	41,000	54,000	36,000	16,000	8,000	8,000	2,000	37,000	36,000	25,000	60,000	32,000	24,000	0,000				
t_{18}	10,000	27,000	20,000	21,000	30,000	14,000	34,000	28,000	16,000	24,000	13,000	16,000	15,000	42,000	56,000	8,000	24,000	0,000			
t_{19}	60,000	27,000	34,000	67,000	80,000	64,000	28,000	24,000	36,000	28,000	63,000	60,000	47,000	86,000	12,000	44,000	28,000	52,000	0,000		
t_{20}	24,000	53,000	50,000	15,000	2,000	18,000	62,000	58,000	46,000	52,000	19,000	22,000	33,000	12,000	86,000	36,000	54,000	32,000	80,000	0,000	
t_{21}	17,000	48,000	43,000	8,000	9,000	11,000	55,000	51,000	39,000	47,000	12,000	15,000	28,000	21,000	79,000	31,000	47,000	23,000	75,000	9,000	0,000

Tabela 45.

Iz Tabele 45 dobijen ja dendogram na Slici 20.



Rastojanje

Slika 20. Dendogram metode proseka za posmatrane lokacije/gradove.

U Tabeli 46 dat je prikaz rastojanja između gradova na osnovu kojih je i kreiran dendrogram.

Spojeni klasteri		Na rastojanju	Broj članova
t_5 – Kikinda	t_{20} – Kuršumlija	2.000	2
t_{10} – Beograd	t_{17} – Požega	2.000	2
t_1 – Palić	t_{11} – Kragujevac	5.000	2
t_3 – Novi Sad	t_{17} – Požega	6.000	3
t_6 – B.Karlovac	t_{12} – S. Palanka	6.000	2
t_2 – Sombor	t_8 – S. Mitrovica	7.000	2
t_1 – Palić	t_{12} – S. Palanka	7.500	4
t_4 – Zrenjanin	t_{21} – Kruševac	8.000	2
t_{16} – Sjenica	t_{18} – Kraljevo	8.000	2
t_3 – Novi Sad	t_9 – Valjevo	8.667	4
t_2 – Sombor	t_7 – Loznica	10.500	3
t_4 – Zrenjanin	t_{20} – Kuršumlija	11.500	4
t_{13} – Crni Vrh	t_{18} – Kraljevo	12.000	3
t_{15} – Zlatibor	t_{19} – Kopaonik	12.000	2
t_2 – Sombor	t_9 – Valjevo	13.500	7
t_1 – Palić	t_{18} – Kraljevo	15.250	7
t_4 – Zrenjanin	t_{14} – Negotin	16.500	5
t_1 – Palić	t_{14} – Negotin	16.500	5
t_2 – Sombor	t_{19} – Kopaonik	31.429	9
t_1 – Palić	t_{19} – Kopaonik	45.593	21

Tabela 46.

Primenom metode proseka utvrđeno je da je najmanja metereološka razlika u klimi u:

1. **Kragujevac, Smederevska Palanka, Banatski Karlovci i Palić.**

Na nešto većem rastojanju, odnosno metereološkoj razlici u klimi, su gradovi:

2. **Valjevo, Požega, Beograd i Novi Sad.**

Sledeća veća grupacija gradova je:

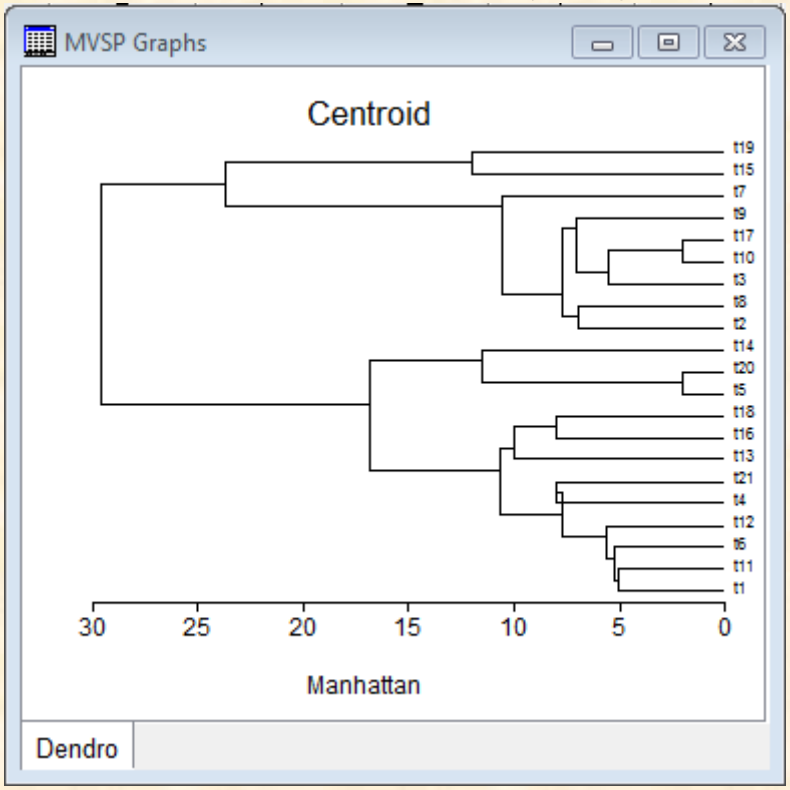
3. **Požega, Valjevo, Beograd, Novi Sad, Loznica, Sremska Mitrovica i Sombor.**

Sa malo većim metereološkim razlikama vremena od prethodne grupacije, u istom klasteru su gradovi:

4. **Kragujevac, Smederevska Palanka, Banatski Karlovac, Palić, Kraljevo, Sjenica i Crni Vrh.**

B) Centroidna klaster metoda –klimatska taksonomija lokacija/gradova

Matrica rastojanja dobijena korišćenjem Apsolutnog (Manhattan) rastojanja data je Tabelom 45 u prethodnom delu. Primenom centroidne metode i programa **MVSP i SYSTAT 13** dobijen je dendogram na Slici 21.



Rastojanje

Slika 21. Dendogram centroidne metode za taksonomiju gradova

U Tabeli 47 dat je prikaz rastojanja izmedju gradova na osnovu kojih je i kreiran dendrogram (Slika 21).

Spojeni klasteri		Na rastojanju	Broj članova
t_5 – Kikinda	t_{20} – Kuršumlija	2.000	2
t_{10} – Beograd	t_{17} – Požega	2.000	2
t_1 – Palić	t_{11} – Kragujevac	5.000	2
t_1 – Palić	t_6 – B. Karlovac	5.250	3
t_3 – Novi Sad	t_{17} – Požega	5.500	3
t_1 – Palić	t_{12} – S.Palanka	5.667	4
t_2 – Sombor	t_8 – S. Mitrovica	7.000	2
t_3 – Novi Sad	t_9 – Valjevo	7.111	4
t_2 – Sombor	t_9 – Valjevo	7.750	6
t_4 – Zrenjanin	t_{21} – Kruševac	8.000	2
t_1 – Palić	t_{21} – Kruševac	7.688	6
t_{16} – Sjenica	t_{18} – Kraljevo	8.000	2
t_{13} – Crni Vrh	t_{18} – Kraljevo	10.000	2
t_2 – Sombor	t_7 – Loznica	10.528	7
t_1 – Palić	t_{18} – Kraljevo	10.639	9
t_5 – Kikinda	t_{14} – Negotin	11.500	3
t_{15} – Zlatibor	t_{19} – Kopaonik	12.000	2
t_1 – Palić	t_{14} – Negotin	16.840	2
t_2 – Sombor	t_{19} – Kopaonik	23.735	9
t_1 – Palić	t_{19} – Kopaonik	29.590	21

Tabela 47.

Primenom centroidne metode utvrđeno je da je najmanja metereološka razlika u vremenu u:

1. **Kragujevcu, Smederevskoj Palanci, Banatskim Karlovcima i Paliću** (isto kao kod metode proseka).

Sa nešto većim metereološkim razlikama vremena, u istom klasteru su:

2. **Požega, Valjevo, Beograd i Novi Sad** (isto kao kod metode proseka).

Sledeća, šira grupacija gradova po osnovu klime, je:

3. **Požega, Valjevo, Beograd, Novi Sad, Sremska Mitrovica i Sombor** (isto kao kod metode proseka, samo bez Loznice).

Sa nešto većim razlikama u klimi od klastera pod 3. je sledeći:

4. **Kragujevac, Smederevska Palanka, Banatski Karlovci, Palić, Zrenjanin i Kruševac** (prva 4 grada su u ovom klasteru i kod metode proseka).

Zaključak

Dobijeni rezultati pokazuju da metode klaster analize daju jednu empirijsku i objektivnu klasifikaciju. Ali ova tehnika zahteva oprez i odgovornost istraživača pri njenom korišćenju.

3.1.5. Metoda Ward-a

Vardova metoda, poznata i kao inkrementalna suma kvadrata, upotrebljava kvadrirane udaljenosti unutar klastera i kvadrirane udaljenosti između klastera. Za svaki klaster izračunaju se aritmetičke sredine za svaku varijablu. Zatim se za svaki objekt računa kvadratna Euklidska udaljenost do aritmetičke sredine klastera. Sumiraju se ove udaljenosti za sve članove klastera. Spajaju se oni klasteri za koje je ukupna (zajednička) suma ovih odstupanja najmanja. U ovoj metodi razdaljina izmedju dva klastera je ustvari suma kvadrata rastojanja između svih promenljivih u ta dva klastera.

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:

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

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

Ovde ćemo koristiti prosečno Euklidovo rastojanje između taksonomskih jedinica:

$$\overline{d(t_a, t_b)} = \sqrt{\frac{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}{n}} .$$

Praktično polazimo od formirane tablice sa rastojanjima, kao i dve taksonomske jedinice između kojih je rastojanje najmanje a koje se grupišu formirajući jedan klaster. Stanja novog klastera se dobijaju kao aritmetička sredina stanja klastera od kojih je novi klaster nastao. Nakon prvog grupisanja, vrši se ponovno grupisanje što sličnijih taksonomskih jedinica i tako se grupisanje može ponavljati $n-1$ puta, gde je n broj taksonomskih jedinica. Na kraju su sve jedinice grupisane u jedan klaster.

Primer – primena metode Ward-a na izbor virtuelnog tima

Za obradu sledećeg primera metodom Ward-a koristiće se i programski paketi MYSTAT 12 i STATISTICA 8.

Primer. Od 10 visokoobrazovanih kandidata treba formirati tim od 6 članova. Kriterijumi su:

- prosek ocena u toku studija;
- poznavanje engleskog jezika;
- poznavanje rada na računaru;
- radno iskustvo i
- komunikativnost.

Uzećemo da je:

k_1 – predstavlja prosek ocena u toku studija;

k_2 – predstavalja poznavanje engleskog jezika, a stanja ovog karaktera su ocene postignute na ispitu iz engleskog jezika (ako je bilo više ispita, računa se prosečna ocena);

k_3 – poznavanje rada na računaru, i sa 1 označimo stanje ne pozanavanja rada na računaru, sa 2 srednji nivo znanja, a sa 3 stanje odličnog poznavanja rada na računaru.

k_4 – radno iskustvo, pa su moguća stanja godine iskustva;

k_5 – komunikativnost, stanje „nije komunikativan“ označimo sa 1, sa dobrim komunikativnim sposobnostima označimo sa 2, a stanje „veoma komunikativan“ označimo sa 3 (Tabela 48).

	k_1	k_2	k_3	k_4	k_5
t_1	8,65	8	2	2	3
t_2	7,96	7	3	4	2
t_3	9,02	8	2	1	2
t_4	8,67	8	2	2	3
t_5	7,35	7	3	0	1
t_6	8,74	9	3	3	3
t_7	7,07	8	1	5	3
t_8	9,22	9	2	1	3
t_9	8,21	8	3	3	2
t_{10}	7,67	8	3	2	3

Tabela 48. Podaci o kandidatima

Sada ćemo kreirati matricu rastojanja, na osnovu koje ćemo vršiti grupisanje jedinica u klastere (Tabela 49). Za određivanje rastojanja koristimo prosečno Euklidovo rastojanje:

$$\overline{d(t_a,t_b)}=\sqrt{\frac{(a_1-b_1)^2+\cdots+(a_n-b_n)^2}{n}},$$

pa dobijamo:

$$\overline{d(t_1,t_2)}=\sqrt{\frac{(8.65-7.96)^2+(8-7)^2+(2-3)^2+(2-4)^2+(3-2)^2}{5}}=1.046,$$

$$\overline{d(t_2,t_3)}=1.557,$$

itd. (Tabela49).

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
t_1	0									
t_2	1.046	0								
t_3	0.654	1.557	0							
t_4	0.009	1.225	0.651	0						
t_5	1.529	1.864	1.165	1.532	0					
t_6	0.776	1.150	1.190	0.775	1.946	0				
t_7	1.581	1.248	2.088	1.585	2.611	1.535	0			
t_8	0.682	1.821	0.639	0.678	1.643	1.023	2.127	0		
t_9	0.799	0.642	1.063	0.801	1.532	0.675	1.435	1.266	0	
t_{10}	0.626	1.103	0.982	0.632	1.349	0.793	1.635	1.039	0.677	0

Tabela 49. Matrica rastojanja

Vidimo da je koeficijent rastojanja taksonomskih jedinica t_1 i t_4 najmanji i iznosi 0.009. Dakle, t_1 i t_4 se grupišu u jedan klaster. Sledeći najmanji koeficijent rastojanja je između t_3 i t_8 i iznosi 0.639, pa se i one spajaju u jedan klaster. Sledeći najmanji koeficijent rastojanja je između t_2 i t_9 i iznosi 0.642. Posle njega najmanji je koeficijent između t_6 i t_{10} i iznosi 0.793. Kako bismo kreirali novu tabelu stanja grupisanih klastera, računamo aritmetičku sredinu stanja od kojih je nastao novi klaster (Tabela 50).

	k_1	k_2	k_3	k_4	k_5
$t_{1,4}$	8,66	8	2	2	3
$t_{2,9}$	8.085	7.5	3	3.5	2
$t_{3,8}$	9.12	8.5	2	1	2.5
t_5	7.35	7	3	0	1
$t_{6,10}$	8.205	8.5	3	2.5	3
t_7	7.07	8	1	5	3

Tabela 50. Stanja grupisanih taksonomskih jedinica

Kreiramo novu matricu (Tabela 51) rastojanja po uzoru na prethodnu.

	$t_{1,4}$	$t_{2,9}$	$t_{3,8}$	t_5	$t_{6,10}$	t_7
$t_{1,4}$	0					
$t_{2,9}$	0.983	0				
$t_{3,8}$	0.585	1.383	0			
t_5	1.531	1.676	1.388	0		
$t_{6,10}$	0.584	0.776	0.931	1.627	0	
t_7	1.583	1.306	2.083	2.610	1.535	0

Tabela 51. Matrica rastojanja u drugom koraku

Vidimo da je koeficijent rastojanja najmanji za klastere $t_{1,4}$ i $t_{6,10}$ i iznosi 0.584. Dakle, grupišemo klastere $t_{1,4}$ i $t_{6,10}$ u jedan klaster $t_{1,4,6,10}$. Pošto je sledeće manje rastojanje 0.585, a njime ne možemo da kreiramo klaster sa četiri elementa, prelazimo na sledeći korak.

	k_1	k_2	k_3	k_4	k_5
$t_{1,4,6,10}$	8.4325	8.25	2.5	2.25	3
$t_{2,9}$	8.085	7.5	3	3.5	2
$t_{3,8}$	9.12	8.5	2	1	2.5
t_5	7.35	7	3	0	1
t_7	7.07	8	1	5	3

Tabela 52. Stanja grupisanih klastera

Ponovo kreiramo matricu rastojanja (Tabela 53).

	$t_{1,4,6,10}$	$t_{2,9}$	$t_{3,8}$	t_5	t_7
$t_{1,4,6,10}$	0				
$t_{2,9}$	0.836	0			
$t_{3,8}$	0.731	1.383	0		
t_5	1.552	1.676	1.388	0	
t_7	1.532	1.306	2.083	2.610	0

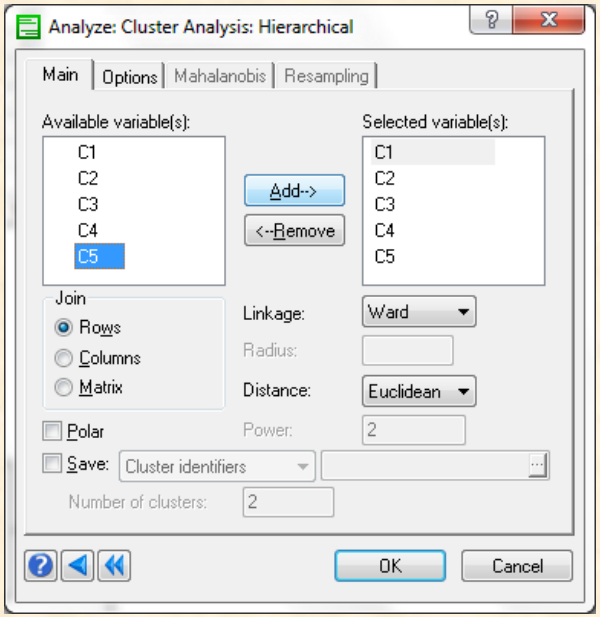
Tabela 53. Matrica rastojanja u trećem koraku

Vidimo da je najmanji koeficijent rastojanja 0.731, i odnosi se na klastere $t_{1,4,6,10}$ i $t_{3,8}$. To znači da grupišemo klestere $t_{1,4,6,10}$ i $t_{3,8}$ u jedan klaster. Konačno smo dobili grupu od šest najsličnijih kandidata, a to su: t_1, t_3, t_4, t_6, t_8 i t_{10}.

Rešenje prethodnog problema u paketu MYSTAT 12

Isti primer obradićemo i u programskom paketu MYSTAT 12.

Nakon unosa podataka u program, iz menija *Analyze* biramo *Cluster Analyzis/Hierarchical*. Zatim se otvara prozor u kome biramo odgovarajuću metodu i vrstu rastojanja, kao i karaktere prema kojim se vrši grupisanje (Slika 22).

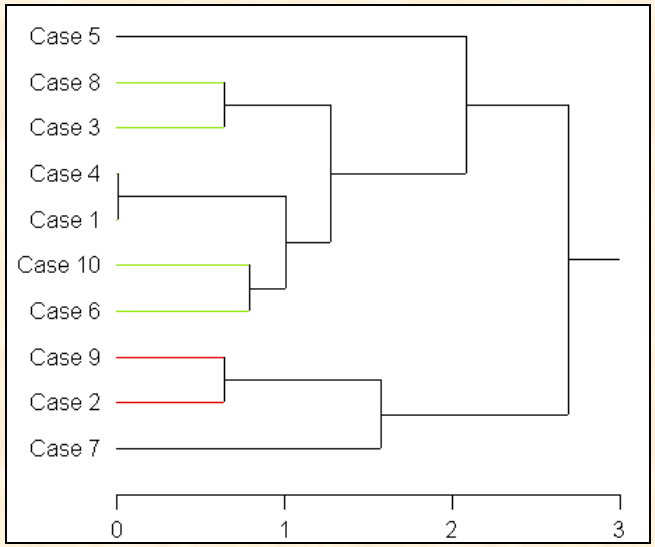


Slika 22. Izbor metode

Potvrdimo sa *OK* i dobijamo dendogram (Slika 24) koji prikazuje ceo postupak grupisanja, kao i tabelu sa rastojanjima između klastera koji se grupišu.

Clusters Joining		at Distance	No. of Members
Case 4	Case 1	0,009	2
Case 8	Case 3	0,639	2
Case 9	Case 2	0,642	2
Case 10	Case 6	0,793	2
Case 10	Case 4	1,004	4
Case 10	Case 8	1,272	6
Case 9	Case 7	1,575	3
Case 5	Case 10	2,088	7
Case 5	Case 9	2,691	10

Slika 23. Tabela grupisanja sa rastojanjima

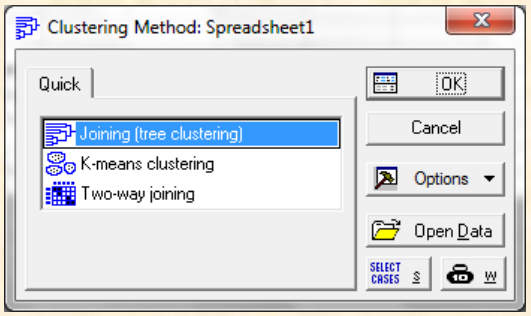


Slika 24. Dendogram

Sa dendograma se vidi da su najsličniji klasteri 1, 4, 3, 8, 6, 10, što smo dobili i u prethodnom delu.

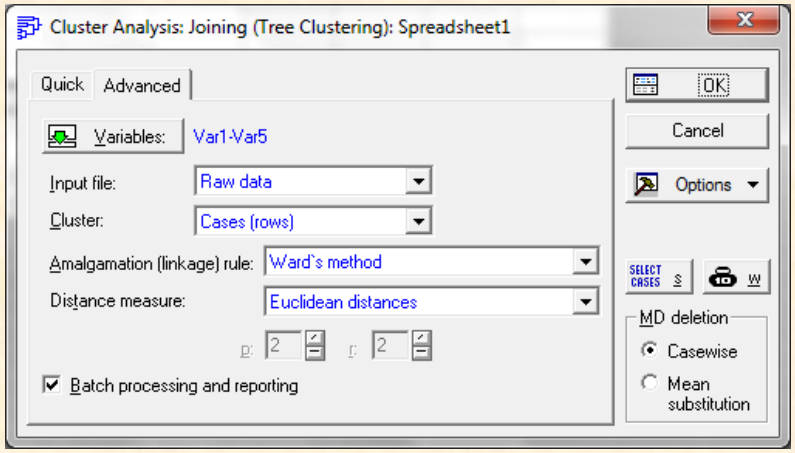
Rešenje polaznog problema i u paketu Statistica 8

Nakon unošenja podataka u program, preko menija *Statistics* biramo opciju *Multivariate Exploratory Techniques/Cluster Analysis*. Otvara se prozor (Slika 25) u kome biramo *Joining (tree clustering)*.



Slika 25. Clustrering method

Nakon toga, otvara se novi prozor za izbor klaster metode i rastojanja, kao i za odabir karaktera po kojim će se vršiti grupisanje (Slika 26).



Slika 26. Izbor metode

Kao rezultat dobijamo dendogram, matricu rastojanja i tabelu grupisanja klastera po koracima.

	Euclidean distances (Spreadsheet1)									
Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10
C_1	0.00	2.73	1.46	0.02	3.42	1.73	3.54	1.52	1.79	1.40
C_2	2.73	0.00	3.48	2.74	4.17	2.57	2.79	4.07	1.44	2.47
C_3	1.46	3.48	0.00	1.46	2.61	2.66	4.67	1.43	2.38	2.20
C_4	0.02	2.74	1.46	0.00	3.43	1.73	3.54	1.52	1.79	1.41
C_5	3.42	4.17	2.61	3.43	0.00	4.35	5.84	3.67	3.43	3.02
C_6	1.73	2.57	2.66	1.73	4.35	0.00	3.43	2.29	1.51	1.77
C_7	3.54	2.79	4.67	3.54	5.84	3.43	0.00	4.76	3.21	3.66
C_8	1.52	4.07	1.43	1.52	3.67	2.29	4.76	0.00	2.83	2.32
C_9	1.79	1.44	2.38	1.79	3.43	1.51	3.21	2.83	0.00	1.51
C_10	1.40	2.47	2.20	1.41	3.02	1.77	3.66	2.32	1.51	0.00

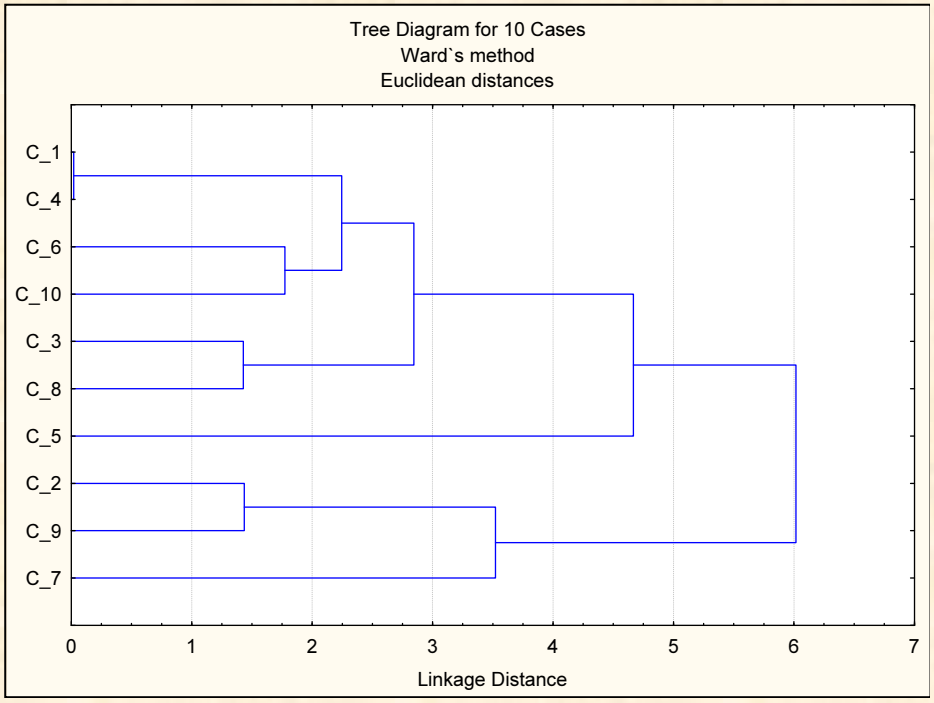
Slika 27. Matrica rastojanja

linkage distance	Amalgamation Schedule (Spreadsheet1) Ward`s method Euclidean distances Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	**Obj. No. 10**
,0200000	C_1	C_4								
1,428286	C_3	C_8								
1,436141	C_2	C_9								
1,773387	C_6	C_10								
2,244411	C_1	C_4	C_6	C_10						
2,844789	C_1	C_4	C_6	C_10	C_3	C_8				
3,521775	C_2	C_9	C_7							
4,667952	C_1	C_4	C_6	C_10	C_3	C_8	C_5			
6,016669	C_1	C_4	C_6	C_10	C_3	C_8	C_5	C_2	C_9	C_7

Slika 28. Tabela grupisanja

Napomena: Rastojanja između klastera se razlikuju od rastojanja koja smo dobili proračunima, zbog toga što smo mi koristili prosečno Euklidovo rastojanje, dok program koristi Euklidovo rastojanje primenjeno direktno na izmerene podatke.

Na dendogramu vidimo da su šest međusobno najsličnijih kandidata, kandidati klastera sa elementima: 1, 4, 6, 10, 3 i 8, što smo dobili i u prethodna dva slučaja (Slika 29).



Slika 29. Dendogram

3.2. DIVIZIVNE METODE

Pored standardnih metoda klaster analize koje se bave tačnom podelom skupa taksonomskih jedinica na klastere, postoje i drugačije metode koje se razvijaju poslednjih godina.

Kod ovog tipa metoda kreće se od pretpostavki da su **sve taksonomske jedinice u jednom klasteru**, pa se zatim taj klaster deli na više manjih klastera. Ponovo se svaki od dobijenih klastera može podeliti na više klastera itd. Od mnogobrojnih metoda ovog tipa navodima samo neke:

- Monotetičke metode
- **Metoda „k-unutrašnjih centara“**
- Metoda particije oko medoida
- **Metoda fazi „k-unutrašnjih centara“**

3.2.1. Metoda „*k* – unutrašnjih centara“

U ovoj metodi **se unapred određuje broj klastera (*k*).** Često taksonom ima intuitivnu predstavu na koliko klastera se skup taksonomskih jedinica, na prirodan način, može podeliti. Za proveru te hipoteze koristi se ova metoda. Kako ne postoji početna vrednost za *k*, metoda se može ponoviti za razne prirodne brojeve *k*, i usvojiti podela na tačno *k* klastera koji su međusobno najviše razdvojeni.

Kod ove metode polazi se od *k* proizvoljnih centara klastera, kojima odgovara *k* klastera. Svaki centar se predstavlja vektorom sa *n* kordinata, gde je *n* broj karaktera koji se posmatra. Sa c_i, $i = 1,\ldots,k$ označeni su centri, a sa $K(c_i)$ označeni su klasteri određeni cetrom c_i. Redom se posmatraju taksonomske jedinice i utvrđuje rastojanje svake taksonomske jedinice od svakog od *k* centara.

Ovo rastojanje se računa preko neke mere sličnosti ili rastojanja. Utvrđuje se kom centru je taksonomska jedinica najbliža tj. računaju se koeficijenti sličnosti (rastojanja) za svaki od centara i taksonomsku jedinicu koja se posmatra, pa se izdvoji centar za koji je koeficijent sličnosti najveći, odnosno koeficijent rastojanja najmanji.

Zatim se taksonomska jedinica pridružuje klasteru koji odgovara tom centru. Ako je taksonomska jedinica podjednako udaljena od dva ili više centara, tada se ona pridruži jednom od njih. Ovo se postiže tako što se klasteri i centri numerišu (tj. pridruže im se brojevi od 1 do k), pa se taksonomska jednica uvek dodeljuje klasteru kome je pridružen najmanji broj.

Na kraju prvog koraka su sve taksonomske jedinice podeljene u *k* klastera. Sada se računaju nove koordinate centra za svaki klaster. Nove koordinate centra će biti jednake prosečnim koordinatama svih taksonomskih jedinica koje su pridružene tom klasteru.

Dalje postupak se ponavlja sa novim koordinatama centra. Ponovo se posmatraju sve taksonomske jedinice i pridružuju onom klasteru od čijeg centra su najmanje udaljene. Može da se dogodi da su klasteri isti kao i u prethodnom koraku algoritma, i ako je to slučaj, postupak se prekida i usvaja se dobijena podela po klasterima. Ako se dobiju drugačiji klasteri od onih koji su bili posle prvog koraka, ponovo se računaju centri novo-dobijenih klastera i raspoređuju se taksonomske jedinice po klasterima.

Dalji postupak se nastavlja sve dok se u jednom koraku raspoređivanja ne dobiju isti klasteri kao i u prethodnom, ili dok se ne dostigne unapred određeni broj iteracija. Kad se to desi, usvaja se dobijena podela.

Dakle, kod ove metode se kreće od *k* proizvoljnih klastera i zatim se premeštaju taksonomske jedinice iz klastera u klaster sa ciljem da je varijabilnost karaktera unutar svakog klastera što manja, a varijabilnost karaktera između svaka dva klastera što veća.

Različiti rezultati se i ovde mogu dobiti u zavisnosti od toga koja mera sličnosti, odnosno rastojanja se uzima prilikom računanja odstojanja taksonomskih jedinica od centra klastera.

Izbor početnih centara je proizvoljan. Često se početni centri biraju na slučajan način. **Međutim njihov izbor je od velikog uticaja na konvergenciju** opisanog postupka i valjanosti dobijenih klastera. Iz tih razloga se pri primeni ove metode **preporučuje da se postupak ponovi više puta** sa različitim početnim centrima. **Druga mogučnost je da se izdvoji uzorak** iz taksonomskih jedinica na kojima će se primeniti neka od hijerarhijskih metoda pa se dobijeni rezultati primene prilikom određivanja početnih centara.

Algoritam *k*-unutrašnjih centara

Najpopularniji algoritam za grupisanje rasturenih (razbacanih) podataka je algoritam „*k* – unutrašnjih centara“. Ovu metodu grupisanja kreirao je **Joseph C. Dunn 1973.** godine, a dopunio je **James C. Bezdek 1981.** Svaki objekat može biti svrstan u više klasa s varirajućim stepenom pripadnosti toj klasi. Sličnost objekata se određuje određenim merama koje pokazuju relevantnost, odnosno daju određene garancije da taj dokument pripada toj klasi.

Alogirtam se odvija u nekoliko koraka:

Zadaju se **početne vrednosti**: broj klastera ***k*** sa centrima: $c_1,\ldots,c_k$

1.korak: **korak dodeljivanja**

Svaka taksonomska jedinica pridružuje se klasteru od čijeg centra je najmanje udaljena. Specijalno, ako centar nije jednoznačno određen taksonomska jedinica se pridružuje klasteru sa manjim indeksom.

***2. korak:* korak ažuriranja**

Određivanje novog centra klastera $K_i = \{t_{p_1}, t_{p_2}, \ldots, t_{p_i}\}, \quad i = 1, \ldots, k$.

Novi centar klastera jednak je aritmetičkoj sredini vektora realizacija taksonomskih jedinica koje se nalaze u tom kalsteru. Specijalno, ako nekom centru nije pridružena nijedna taksonomska jedinica, tada centar ostaje nepromenjen.

I na kraju imamo **uslov zaustavljanja:** koraci pod 1. i 2. se ponavljaju dok se u dve uzastopne iteracije ne dobiju isti klasteri ili dok se ne dostigne unapred zadati broj iteracija.

Primer- primene metode „k-unutrašnjih centara“ na izbor radnog tima

Kriterijumi za izbor radnog tima u ovom slučaju **preuzeti su iz javnog konkursa** koji je početkom juna 2012. godine objavila **kompanija „Zlatiborac“** za radno mesto: „Menadžer za odonose s javnošću“ sa mestom rada u Beogradu. Kompanija je potraživala kandidate koji bi radili u timu.

Uslovi konkursa su sledeći:

- univerzitetsko obrazovanje (prosečna ocena),
- poželjno radno iskustvo na istim ili sličnim poslovima (u godinama),
- izražene organizacione sposobnosti kao i sposobnost planiranja (ocena na skali od 1 do 10),
- izražene komunikacione veštine (ocena na skali od 1 do 10)
- odlično znanje engleskog jezika (ocena na skali od 6 do 10)
- poznavanje rada na računaru (ocena na skali od 6 do 10)

Neka se na konkurs javilo 10 kandidata čije karakteristike su prikazane u Tabeli 54.

	k_1	k_2	k_3	k_4	k_5	k_6
t_1	8,1	3	6	6	8	10
t_2	6,55	2	4	5	7	8
t_3	9,50	6	9	10	10	10
t_4	8,69	1	7	7	8	9
t_5	7,39	4	5	4	6	6
t_6	9,75	5	10	9	10	10
t_7	7,99	2	8	8	9	9
t_8	7,61	1	3	6	6	8
t_9	6,87	1	2	4	7	7
t_{10}	9,06	5	10	9	9	10

Tabela 54. Podaci prijavljenih kandidata na konkurs

Kvantitativni karakteri k_1 do k_6 označavaju, istim redom, navedene kriterijume konkursa.

Podatke iz prethodne tabele obradićemo primenom programskog paketa STATISTICA i metode „k-unutrašnjih centara (eng. K-means clustering)".

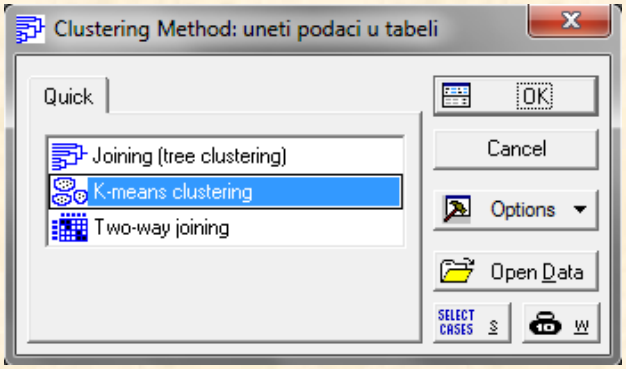
Na sledećoj Slici 30 prikazani su podaci uneti u program Statistika 8.

	1 k1	2 k2	3 k3	4 k4	5 k5	6 k6
t1	8,1	3	6	6	8	10
t2	6,55	2	4	5	7	8
t3	9,5	6	9	10	10	10
t4	8,69	1	7	7	8	9
t5	7,39	4	5	4	6	6
t6	9,75	5	10	9	10	10
t7	7,99	2	8	8	9	9
t8	7,61	1	3	6	6	8
t9	6,87	1	2	4	7	7
t10	9,06	5	10	9	9	10

Slika 30. Prikaz unetih podataka u programu Statistica

Postupak za dobijanje rezultata metodom „k-unutrašnjih centara“ je sledeći.

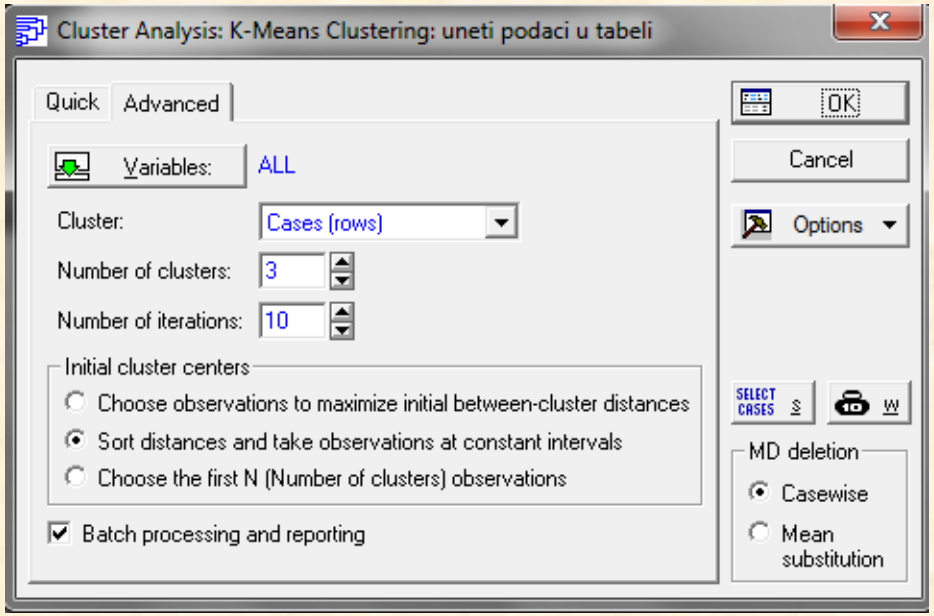
Na glavnom meniju izaberemo padajući meni ***Statistics***, zatim idemo na opciju ***Multivariate Exploratory Techniques***, i zatim odaberemo ***Cluster Analysis***. Sledi prozor prikazan na Slici 31.



Slika 31. Prozor za izbor metode k-unutrašnjih centara

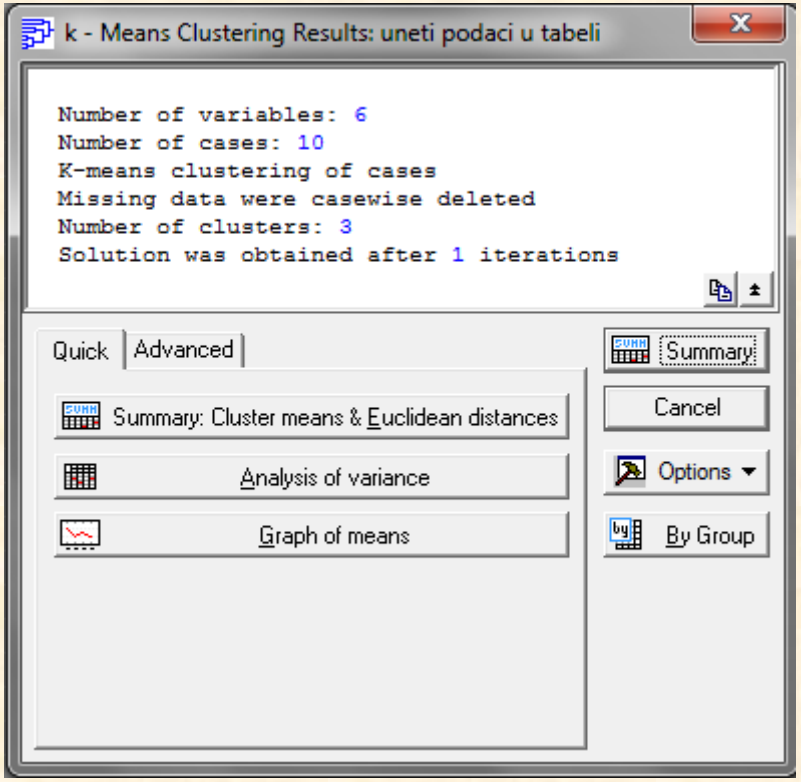
U programu zatim izvršimo selekciju svih karatera k_1 - k_6 i upišemo broj željenih klastera (grupa) u koje želimo da razvrstamo kandidate. Unesemo i broj iteracija do koje bi se išlo u slučaju da se ne nađe rešenje koje bi se poklopilo sa prethodnom iteracijom, što je i uslov zaustavljanja kako bi dobili rešenje problema.

Ovaj postupak prikazan je Slikom 32.



Slika 32. Poslednji korak za dobijanje rešenja metodom k-untrašnjih centara

Sa prethodne slike se vidi da su odabrana 3 klastera i 10 iteracija. Potrvdom na *OK* dobijamo prozor dobijamo prozor sa Slike 33 koji nam nudi moguće opcije.



Slika 33. Izgled prozora sa opcijama za dalje korake

Potvrdom na opciju *Summary* na Slici 33 dobijamo rezultate koji su prikazani na sledećim slikama.

Variable	Analysis of Variance (uneti podaci prijavljenih kandidata)					
	Between SS	df	Within SS	df	F	signif. p
k1	9,37092	2	1,226967	7	26,73114	0,000528
k2	23,33333	2	8,666666	7	9,42308	0,010339
k3	66,73334	2	7,666663	7	30,46523	0,000351
k4	36,18333	2	5,416666	7	23,38000	0,000797
k5	17,66667	2	2,333334	7	26,49999	0,000542
k6	14,68333	2	3,416667	7	15,04146	0,002922

Sika 34. Izgled dobijenih rezultata analize varijanse

Analiza varijanse (Slika 34) nam daje prikaz najzastupljenijih kriterijuma prilikom izbora kandidata, a to su karakteristike koje u koloni *F* imaju najveću vrednost, tj. k_1, k_3, k_5.

Na sledećoj, Slici 35, vidimo dobijena rastojanja između klastera.

Cluster Number	Euclidean Distances between Clusters Distances below diagonal Squared distances above diagonal		
	No. 1	No. 2	No. 3
No. 1	0,000000	4,391319	4,54557
No. 2	2,095547	0,000000	15,52880
No. 3	2,132035	3,940660	0,00000

Slika 35. Rastojanja između klastera

Ovi rezultati su dobijeni primenom Euklidovog rastojanja i najudaljeniji su klasteri 2 i 3.

Konačnu tabelu možemo dobiti tako što iz padajućeg menija izaberemo opciju:

Data mining → Generalized EM & k-Means Cluster Analysis, izvršimo izbor kriterijuma (varijable), odredimo broj klastera i iteracija i potvrdimo na ***OK*** (Slika 36).

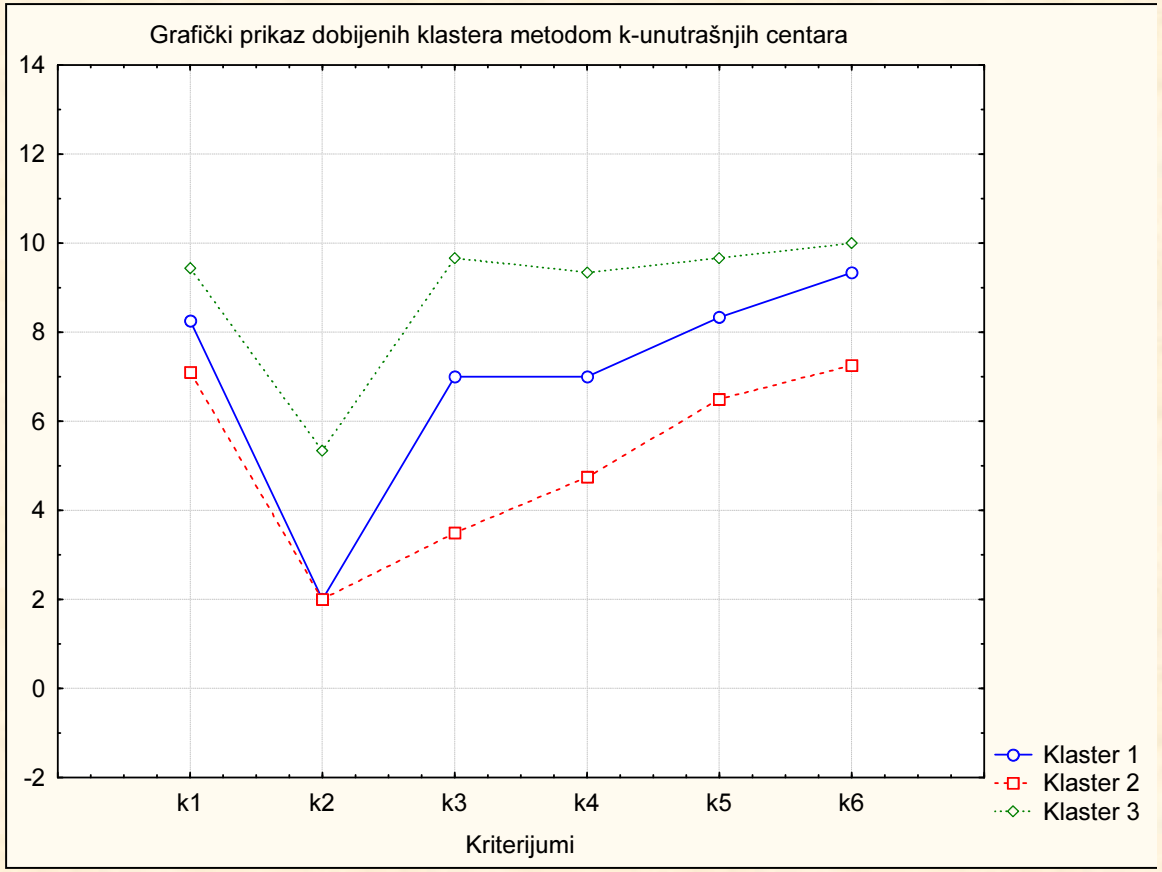
Cluster members (uneti podaci prijavljenih kandidata)
Number of clusters: 3
Total number of training cases: 10

kandidati	**Konačna klasifikacija**	k1	k2	k3	k4	k5	k6	rastojanje od centroida
t1	1	8,100000	3,000000	6,00000	6,00000	8,00000	10,00000	0,347311
t4	1	8,690000	1,000000	7,00000	7,00000	8,00000	9,00000	0,268227
t7	1	7,990000	2,000000	8,00000	8,00000	9,00000	9,00000	0,291966
t2	2	6,550000	2,000000	4,00000	5,00000	7,00000	8,00000	0,294116
t5	2	7,390000	4,000000	5,00000	4,00000	6,00000	6,00000	0,576190
t8	2	7,610000	1,000000	3,00000	6,00000	6,00000	8,00000	0,403726
t9	2	6,870000	1,000000	2,00000	4,00000	7,00000	7,00000	0,340155
t3	3	**9,500000**	**6,000000**	**9,00000**	**10,00000**	**10,00000**	**10,00000**	**0,210723**
t6	3	**9,750000**	**5,000000**	**10,00000**	**9,00000**	**10,00000**	**10,00000**	**0,160621**
t10	3	**9,060000**	**5,000000**	**10,00000**	**9,00000**	**9,00000**	**10,00000**	**0,225610**

Slika 36. Izgled rezultata dobijenih klastera

Sa Slike 36 vidimo da smo dobili konačnu klasifikaciju od tri klastera i dobili smo rastojanje od centra svakog klastera (“rastojanje od centroida”). Na osnovu ove slike može se zaključiti koji je klaster najbolji tj. u kom klasteru su kandidati sa najboljim karakteristikama.

Grafički prikaz klastera, dobijen klaster metodom „k-unutrašnjih centara”, dat je na Slici 37 i on nam daje uvid u to koja grupa kandidata će najbolje odgovoriti na potrebe kompanije “Zlatiborac“.



Slika 37. Grafički prikaz klastera

U donjem desnom uglu vidimo koji klaster je kako označen, i na osnovu toga sa slike (dijagrama) čitamo srednje vrednosti određenog klastera sa njegovim karaketristikama.

Očigledno je da je klaster 3 najbolji. Njega čine kandidati označeni sa t_3, t_6 i t_{10}, i ovi kandidati imaju najbolje karakteristike potrebne za traženi posao. Nešto lošiju grupu čine kandidati koji se nalaze u klasteru 1 tu su kandidati t_1, t_4 i t_7, a sa najlošijim karakeristikama su kandidati t_2, t_5, t_8 i t_9 iz klastera 2.

Zaključak

Metoda „k-unutrašnjih centara“ je jedna od najboljih metoda iz grupe divizivnih metoda zbog njene jednostavnosti i njena primena je veoma široka.

Međutim njen nedostatak je problem oko odabira centara klastera, jer je njihov izbor proizvoljan a bitno utiče na valjanost dobijenih klastera. Takođe, broj klastera kod ove metode se unapred određuje. Ova metoda pokazuje slabije rezultate kada su klasteri različite veličine, gustine i oblika (ne može se pronaći klaster čiji oblik nije konveksan). Ovom metodom nije poželjno klasifikovati podatke koji imaju veliko odstupanje od drugih taksonomskih jedinica. Takve taksonomske jedinice nazivaju se – **samci**.

3.2.2. Metoda fazi *k*- unutrašnjih centara

Ova metoda je bazirana na teoriji rasplinutih (fazi) skupova. Ona razvrstava taksonomske jedinice u klastere sa nekim stepenom pripadanja, tako da taksonomske jedinice mogu pripadati i različitim klasterima u nekoj meri.

Fazi logika

Svedoci smo naglog porasta upotrebe fazi logike u veoma raznovrsnim komercijalnim aplikacijama i industrijskim sistemima. Neki primeri kao što su veš mašine, klima uređaji, usisivači, navigacioni uređaji, kao i mnogi drugi, dovoljan su dokaz velike rasprostranjenosti i primenjivosti ove tehnike. Fazi tehnologija je našla i primenu u informacionim tehnologijama i ekspertskim sistemima, gde se koristi kao podrška pri odlučivanju.

Fuzzy skupove definisao je **1965-e godine Lotfi Zadeh** kao ***matematički formalizovan način predstave i modeliranja neodređenosti u lingvistici***. U teoriji klasičnih, jasnih skupova, neki određeni element ili pripada ili ne pripada nekom definisanom skupu. Fazi skup je, u tom smislu, generalizacija klasičnog skupa, budući da se pripadnost (tj. stepen pripadnosti) elementa fazi skupu može okarakterisati brojem iz intervala [0,1].

Drugim rečima, ***funkcija pripadnosti*** (membership function) fazi skupa preslikava svaki element univerzalnog skupa u pomenuti interval realnih brojeva. Jedna od najvećih razlika između klasičnih i fazi skupova jeste u tome što klasični skupovi uvek imaju jedinstvenu funkciju pripadnosti, dok se za fazi skup može definisati beskonačno mnogo različitih funkcija pripadnosti kojima se on može opisati (sam izraz ***fazi (fuzzy)*** predstavlja nešto nejasno, zamagljeno, lepršavo).

Osnovna razlika između fazi logike i teorije verovatnoće sastoji se u tome da fazi logika operiše sa determinističkim nedorečenostima i neodređenostima, dok se verovatnoća bavi verodostojnošću stohastičkih događaja i iza nje suštinski stoji eksperiment.

Fazi logika ima za cilj prevazilaženje problema u komunikaciji vezanih za razlike između pravila koja nameću formalne teorije i načina razmišljanja koji opisuju ponašanje ljudskog uma, **dok se verovatnoća** bavi fenomenom ponavljanja koji se simbolizuje slučajnošću (slučajnim promenljivama i slučajnim procesima). Drugim rečima, fazi i slučajni su dva atributa koji se razlikuju u svojoj prirodi, odnosno, oni opisuju drugačiji aspekt neodređenosti.

Dakle, fazi logika pokriva subjektivnost ljudskog mišljenja, osećanja, jezika, dok verovatnoća pokriva objektivnu statistiku u prirodnim naukama.

Pojam fazi skupova

Kod klasičnih skupova svaki element ili pripada ili ne pripada skupu. U praksi se javljaju i drugačije situacije: želimo, naprimer, da izdvojimo podskup nekog skupa koji sadrži one elemente toga skupa koji imaju neku osobinu. Problem je u tome što je ta osobina neprecizno definisana.

Ovakvi i slični problemi se matematički prevazilaze uvođenjem nove vrste skupova, tzv. **rasplinutih** ili **fazi skupova**. To su skupovi čije elemente karakteriše **stepen pripadanja** elementa skupu. Stepen pripadanja je neki realni broj između 0 i 1.

Uvodimo formalnu **definiciju pojma rasplinutog fazi skupa (podskupa).**

Skup X od kojeg se polazi naziva se **univerzum**. **Rasplinuti podskup** A skupa X je određen funkcijom μ_A: $X \rightarrow [0,1]$, gde je $[0,1]$ interval realnih brojeva. Ova funkcija se zove **funkcija pripadanja.** Za svako $x \in X$, vrednost $\mu_A(x)$ naziva se stepen pripadanja elementa X skupu A. Dakle, elementi pripadaju rasplinutom skupu sa većim ili manjim stepenom pripadanja.

Rasplinuta (fazi) particija

Particija (podela) nekog skupa *A* je kolekcija nepraznih podskupova tog skupa, takva da je unija te kolekcije skup *A* i presek svaka dva skupa iz te kolekcije prazan skup. Dakle, u pitanju je tačna podela tog skupa na podskupove takva da svaki element skupa *A* pripada tačno jednom podskupu particije.

Za razliku od obične particije, u rasplinutoj fazi particiji skupa A jedan element može pripadati većem broju podskupova, s različitim stepenima pripadanja.

Dakle, rasplinuta (fazi) particija skupa A je familija njegovih rasplinutih podskupova $\{A_1, A_2, \ldots, A_k\}$, takvih da je za svaki element $x \in A$, zbir stepena pripadanja tim rasplinutim podskupovima jednak 1, odnosno

$$A_1(x) + A_2(x) + \ldots + A_k(x) = 1.$$

Ovde se, zbog jednostavnosti, funkcija pripadanja nekom rasplinutom skupu obeležava isto kao i rasplinuti skup.

Algoritam metode fazi *k*-unutrašnjih centara

Ova metoda je poznata pod nazivom FCM metoda - **Fuzzy C means (metoda fazi k sredina)** i zasnovana je na **Bazdekovom algoritmu**. Razlika između ove i svih drugih pomenutih metoda klaster analize je u tome što se ovde vrši razvrstavanje taksonomskih jedinica u klastere tako da jedna taksonomska jedinica može pripadati više nego jednom klasteru (sa određenim stepenom pripadanja). Precizno, kod svih drugih metoda, ono što se dobija je particija (podela) taksonomskih jedinica u klastere, a kod ove metode dobija se fazi particija. Definisane su funkcije koje svakoj taksonomskoj jedinici dodeljuju stepen pripadanja svakom od klastera. Za svaku taksonomsku jedinicu zbir svih stepena pripadanja klasterima je 1.

Slično kao i kod metode k-unutrašnjih centara, polazi se od skupa taksonomskih jedinica $T=\{t_1,t_2,\ldots,t_n\}$ koje želimo da razvrstamo po osnovu *N* karaktera. Vektor koji predstavlja realizaciju karaktera za neku taksonomsku jedinicu i samu tu jedinicu obeležavamo istim slovom, tj. sa $\{t_1,t_2,\ldots,t_n\}$. Unapred se zadaje broj klastera *k*. Kreće se od proizvoljne rasplinute particije $\{A_1,A_2,\ldots,A_k\}$ (familije od *k* rasplinutih skupova na skupu *T*). Svakom od tih

rasplinutih skupova odgovara funkcija pripadnosti koja se obeležava istim simbolom i definiše stepene pripadanja svake od taksonomskih jedinica iz T tim rasplinutim skupovima. Zatim se računaju centri datih klastera. Centri klastera $c_1,\ldots,c_k$ su takođe vektori sa istim brojem koordinata kao i vektori realizacija $t_1,t_2,\ldots,t_n$. Centre klastera možemo da izračunamo preko sledeće formule:

$$c_j = \frac{\sum_{i=1}^{N} U_{ij}^m x_i}{\sum_{i=1}^{N} U_{ij}^m}, \quad (j = 1,2,\ldots,k),$$

gde je U_{ij} stepen pripadanja karaktera x_i klasteru j, a m parameter tako da je $m > 1$.

U sledećem koraku se definiše nova rasplinuta particija formulom koja zavisi od rastojanja pojedinih taksonomskih jedinica od centara klastera. Postupak se nastavlja sve dok se ta particija ne poklopi sa prethodno izračunatom particijom (bar približno). Ova metoda se zasniva na minimizaciji funkcije J_m:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{k} U_{ij}^{\ m} \left\| x_i - c_j \right\|^2,$$

gde važi $1 < m < \infty$.

Postupak za nalaženje stepena pripadnosti taksonomskih jedinica nekom klasteru je sledeći:

1) Izračunamo udaljenosti pojedinih taksonomskih jedinica od centara klastera preko Euklidskog rastojanja:

$$d_j = \sqrt{\sum_{i=1}^{N} \left| x_i - y_i \right|^2}, \quad (j = 1,2,\ldots,k).$$

2) Izračunamo vrednost p:

$$p = \frac{2}{m-1}.$$

3) Zatim računamo sumu nađenih rastojanja:

$$sumD = \sum_{j=1}^{k} d_j^p.$$

4) Izračunamo vrednosti V_j:

$$V_j = \frac{1}{d_j^p \cdot sumD},$$

5) Suma tih vrednosti V je:

$$sumV = \sum_{j=1}^{k} V_j.$$

6) Sada je stepen pripadnosti pojedinim klasterima dat sa:

$$U_j = \frac{V_j}{sumV}.$$

Primer A) - primene metode FCM

Preko metode fazi k-unutrašnjih centara izvršimo grupisanje u dva klastera, tj. $k = 2$, 6 različitih modela automobila koji su međusobno sličnih performansi. Njihova svojstva pratićemo preko: jačine motora i prosečne potrošnje goriva. Podaci o automobilima su dati u Tabeli 55.

REDNI BROJ	MODEL AUTOMOBILA	JAČINA MOTORA	PROSEČNA POTROSNJA (l)
1	Opel Corsa	1.3	3.6
2	Chevrolet Spark	1.0	5.1
3	Citroen C3	1.4	6.1
4	Toyota Yaris	1.0	7.3
5	Fiat 500 L	1.4	3.2
6	Volkswagen Lupo	1.4	5.5

Tabela 55.

Ovde je skup taksonomskih jedinica $T=\{1,2,3,4,5,6\}$, gde su 1,2,3,4,5,6 njihovi redni brojevi u Tabeli 55. Iterativni postupak ćemo započeti sa proizvoljnom rasplinutom particijom $\{A_1, A_2\}$ (familija od dva rasplinuta skupa na skupu T). Svakom od tih rasplinutih skupova odgovara funkcija pripadnosti koju označavamo istim simbolom. Uzećemo da je:

$$A_1 : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.3 & 0.7 & 0.7 & 0.7 & 0.3 & 0.7 \end{pmatrix}$$

i

$$A_2 : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.7 & 0.3 & 0.3 & 0.3 & 0.7 & 0.3 \end{pmatrix}.$$

Sada računamo početne centre c_1 i c_2 prvog i drugog klastera.

Za vrednost parametara m uzećemo $m = 2$. Iz prethodno navedenih formula, sledi

$$c_1 = \left(\frac{0.3^2 \cdot 1.3 + 0.7^2 \cdot 1 + 0.7^2 \cdot 1.4 + 0.7^2 \cdot 1 + 0.3^2 \cdot 1.4 + 0.7^2 \cdot 1.4}{0.3^2 \cdot 2 + 0.7^2 \cdot 4}, \right.$$

$$\left. \frac{0.3^2 \cdot 3.6 + 0.7^2 \cdot 5.1 + 0.7^2 \cdot 6.1 + 0.7^2 \cdot 7.3 + 0.3^2 \cdot 3.2 + 0.7^2 \cdot 0.5}{0.3^2 \cdot 2 + 0.7^2 \cdot 4} \right),$$

$$c_1 = (1.21,\ 4.64),$$

$$c_2 = \left(\frac{0.7^2 \cdot 1.3 + 0.3^2 \cdot 1 + 0.3^2 \cdot 1.4 + 0.3^2 \cdot 1 + 0.7^2 \cdot 1.4 + 0.3^2 \cdot 1.4}{0.7^2 \cdot 2 + 0.3^2 \cdot 4}, \right.$$

$$\left. \frac{0.7^2 \cdot 3.6 + 0.3^2 \cdot 5.1 + 0.3^2 \cdot 6.1 + 0.3^2 \cdot 7.3 + 0.7^2 \cdot 3.2 + 0.3^2 \cdot 0.5}{0.7^2 \cdot 2 + 0.3^2 \cdot 4} \right),$$

$$c_2 = (1.31,\ 3.76).$$

Sada definišimo novu rasplinutu particiju koja zavisi od rastojanja pojedinih taksonomskih jedinica od centara klastera, tj. od c_1 i c_2.

Za postizanje toga cilja, primenićemo prethodno opisani postupak za određivanje stepena pripadnosti taksonomskih jedinica iz skupa $T=\{1,2,3,4,5,6\}$, a samim tim i svakog njihovog karaktera, prvom i drugom klasteru.

- Za taksonomsku jedinicu pod rednim brojem 1 je:

  1) Udaljenost od centara c_1 i c_2 data sa:

$$d_1=\sqrt{(1.3-1.21)^2+(3.6-4.64)^2}=1.04$$

$$d_2=\sqrt{(1.3-1.31)^2+(3.6-3.76)^2}=0.16.$$

  2) Izračunajmo parametar *p*:

$$p=\frac{2}{m-1}=2.$$

  3) Suma novoizračunatih rastojanja je:

$$sumD=d_1^p+d_2^p=1.04^2+0.16^2=1.11$$

4) Izračunajmo vrednosti V_j ($j = 1,2$):

$$V_1 = \frac{1}{d_1^p \cdot sumD} = 0.83 \quad \text{i} \quad V_2 = \frac{1}{d_2^p \cdot sumD} = 35.71.$$

5) Suma V vrednosti je:

$$sumV = V_1 + V_2 = 36.54.$$

6) Stepen pripadanja taksonomske jedinice pod rednim brojem 1, a samim tim i svakog njenog karaktera, prvom i drugom klasteru nakon prve iteracije je:

$$U_{11} = \frac{V_1}{sumV} = 0.0227 \text{ i } U_{12} = \frac{V_2}{sumV} = 0.9773.$$

- Za taksonomsku jedinicu 2, nakon prve iteracije dobijamo:

$U_{21} = 0.8798$ i $U_{22} = 0.1202$, itd. (Tabela 56.)

REDNI BROJ	KLASTER 1	KLASTER 2
1	0.0227	0.9773
2	0.8798	0.1202
3	0.7172	0.2828
4	0.6384	0.3616
5	0.1339	0.8661
6	0.4885	0.5115

Tabela 56.

Možemo primetiti da je za svaku taksonomsku jedinicu, zbir stepena pripadnosti prvom i drugom klasteru jednak 1.

Ponavljanjem postupka, uz novodobijene rasplinute particije skupa *T*, date funkcijama pripadnosti u Tabeli 56, dolazimo do novih centara klastera:

$$c_1 = (1.16,\ 5.00) \text{ i } c_2 = (1.33,\ 3.42).$$

REDNI BROJ	KLASTER 1	KLASTER 2
1	0	1
2	0.73	0.27
3	1	0
4	0.91	0.09
5	0.01	0.99
6	0.92	0.08

Tabela 57.

Iz Tabele 57 očigledno je da jedinice pod rednim brojem 2,3,4,6 pripadaju prvom klasteru, a jedinice pod rednim brojem 1 i 5 pripadaju drugom klasteru. Krajnji klaster centri koje dobijamo nakon 6. iteracije su C_1(1.23, 6.09) i C_2 (1.34, 3.49).

To znači da prvoj grupi (klasteru) prupadaju automobili Chevrolet Spark, Citroen C3, Toyota Yaris i Volkswagen Lupo, a drugom klasteru automobili Opel Corsa i Fiat 500L.

Zadatak možemo rešiti i u programskom paketu Matlab kompanije Mathworks. Najpre ćemo da se pozovemo na našu metodu (Fuzzy C means) tako što ćemo otkucati kod sa Slike 38 u komandnom prozoru (prethodno smo naše podatke uneli i kreirali fajl podaci.txt):

Command Window

```
>> data = load('podaci.txt');  % ucitavanje podataka
n_clusters = 2;                % broj klastera
[center,U,obj_fcn] = fcm(data, n_clusters);
```

Slika 38. Zadavanje komandi za Fuzzy C Means Clustering u Matlab-u

Zatim pritiskom na taster Enter, dobijamo prikaz funckije FCM i njene minimizacije, odnosno kako se smanjuje iz iteracije u iteraciju, dok ne ostane nepromenjena u bar dve iteracije. Takođe, u gornjem delu prozora, možemo videti podatke o centrima klastera, stepenu pripadanja klasterima.

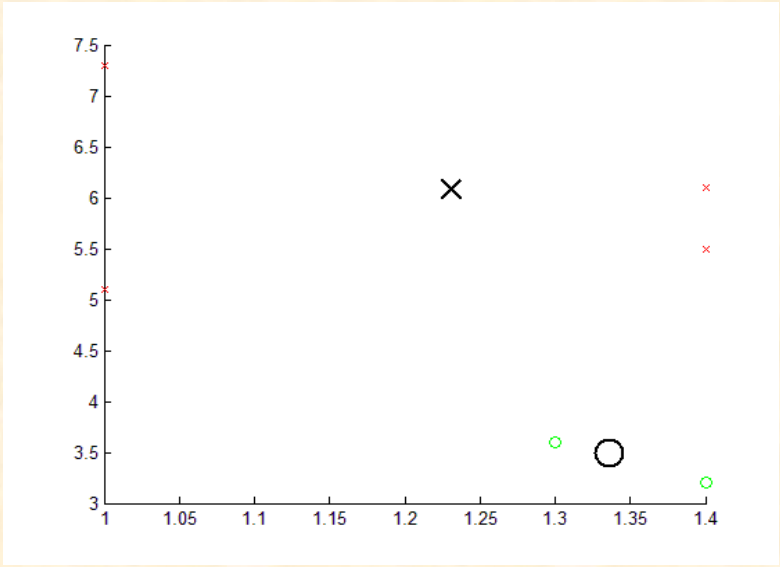
```
Iteration count = 1, obj. fcn = 6.520414
Iteration count = 2, obj. fcn = 6.026234
Iteration count = 3, obj. fcn = 6.015211
Iteration count = 4, obj. fcn = 5.929401
Iteration count = 5, obj. fcn = 5.353621
Iteration count = 6, obj. fcn = 3.670934
Iteration count = 7, obj. fcn = 2.656330
Iteration count = 8, obj. fcn = 2.597265
Iteration count = 9, obj. fcn = 2.596348
Iteration count = 10, obj. fcn = 2.596323
Iteration count = 11, obj. fcn = 2.596320
```

Slika 39. Prikaz minimizacije funkcije

Zatim želimo da nam se centri klastera grafički prikažu. Zadaćemo da nam centar jednog klastera bude označen sa *X* a drugog sa *O*, a jedinice (automobili) koji pripadaju klasteru čiji je centar *X* da budu označeni sa malim crvenim iks-evima, a oni koji pripadaju klasteru čiji je centar označen sa *O* da budu označeni sa malim zelenim kružićima, kao što možemo videti na Slici 40.

Command Window

```
>> load podaci.txt
plot(podaci(:,1),podaci(:,2),'o')
[center,U,objFcn] = fcm(podaci,2);
maxU = max(U);
index1 = find(U(1, :) == maxU);
index2 = find(U(2, :) == maxU);
figure
line(podaci(index1, 1), podaci(index1, 2), 'linestyle',...
'none','marker', 'o','color','g');
line(podaci(index2,1),podaci(index2,2),'linestyle',...
'none','marker', 'x','color','r');
hold on
plot(center(1,1),center(1,2),'ko','markersize',15,'LineWidth',2)
plot(center(2,1),center(2,2),'kx','markersize',15,'LineWidth',2)
```



Slika 40. Zadavanje komandi i grafički prikaz centara i pripadnosti klasterima

Na Slici 41 možemo videti prikaz raspodele jedinica po klasterima, krajnje klaster centre, kao i stepen pripadnosti klasterima.

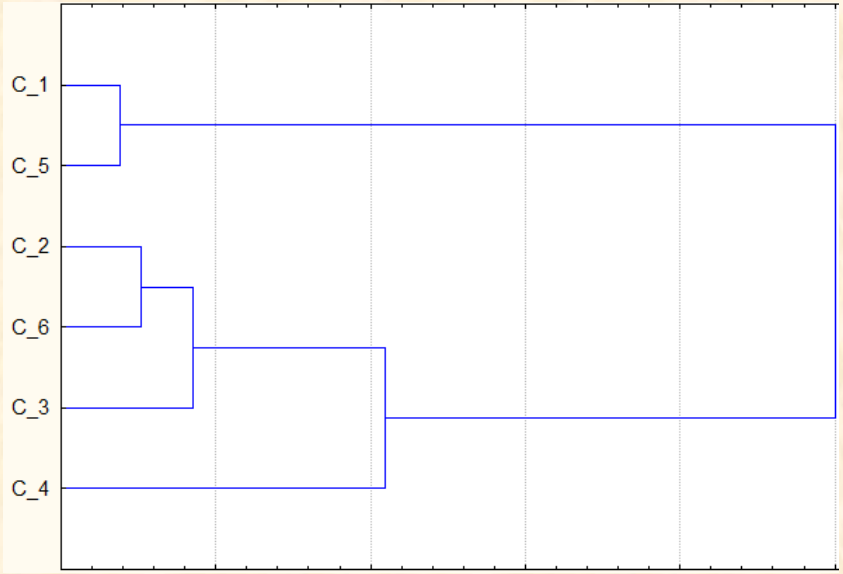
Name	Value	Min	Max
U	<2x6 double>	0.0022	0.9978
center	[1.2305,6.0855;1.3355,3.4875]	1.2305	6.0855
index1	[2,3,4,6]	2	6
index2	[1,5]	1	5
maxU	[0.9978,0.7259,0.9958,0.9055,0.9897,0.9161]	0.7259	0.9978
objFcn	<11x1 double>	2.5963	6.5204
podaci	<6x2 double>	1	7.3000

U <2x6 double>

	1	2	3	4	5	6
1	0.0022	0.7259	0.9958	0.9055	0.0103	0.9161
2	0.9978	0.2741	0.0042	0.0945	0.9897	0.0839

Slika 41. Razvrstavanje podataka po klasterima i stepen pripadanja prvom i drugom klasteru

U programskom paketu Statistica grafički prikaz raspodela ovih 6 modela automobila po klasterima dat je dendrogramom na Slici 42.



Slika 42. Dendrogram grupisanja podataka u klastere

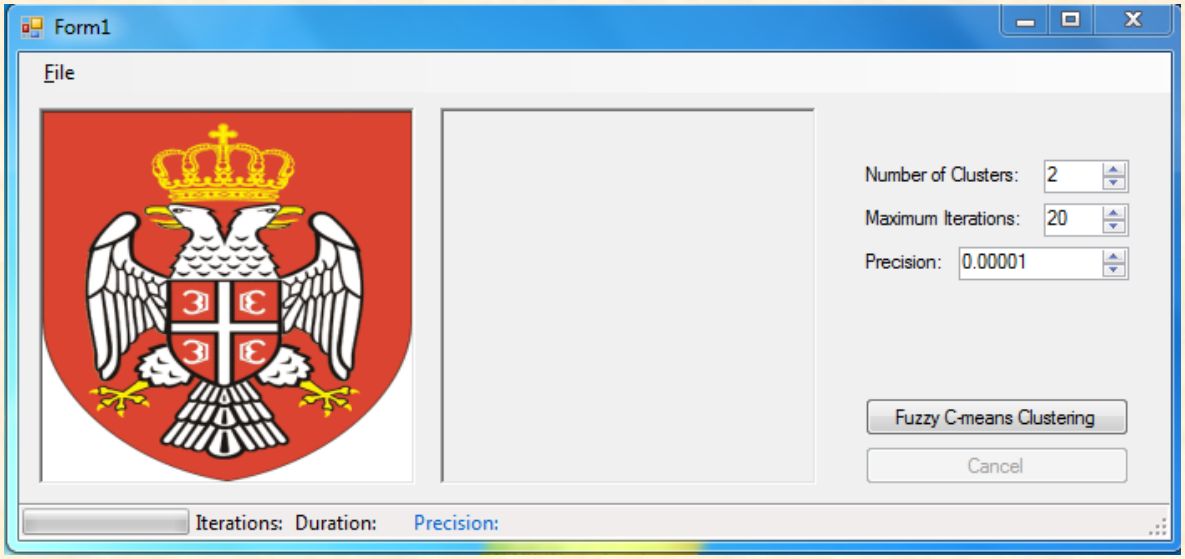
I sa Slike 42 možemo još jednom videti da su jedinice pod rednim brojem 1 i 5 svrstane u jedan klaster (Opel Corsa i Fiat 500L), a jedinice pod rednim brojem 2,3,4 i 6 u drugi klaster (Chevrolet Spark, Citroen C3, Toyota Yaris i Volkswagen Lupo).

Primer B) Segmentacija slike preko FCM klasteringa

Segmentacija slike, podela slike u homogene (slične) regione na osnovu skupa karakteristika, ključni je element u analizi slike i računarske vizije. Grupisanje (klastering) je jedan od metoda koji se koristi za ovu svrhu. Klastering je proces koji se može koristiti za klasifikaciju piksela na osnovu sličnosti po boji piksela ili po nivou intenziteta sive boje.

K-means algoritam (metod k-sredina) je korišćen za brze i oštre segmentacije. Teorija fazi skupova poboljšava ovaj proces tako što postoji koncept delimičnog članstva u kome piksel može pripadati više nego jednom klasteru. Ovo **„meko“ grupisanje** omogućava preciznije računanje klaster članstva i uspešno se koristi za grupisanje slika i za segmentacije medicinskih, geoloških i satelitskih snimaka.

Na jednom primeru pokazaćemo kako se vrši segmentovanje slike i grupisanje u određeni broj klastera. Koristimo aplikaciju koja je kreirana u programskom jeziku C#. Otvorićemo aplikaciju, učitati jednu sliku po izboru, kao što možemo videti sa Slike 43.



Slika 43. Učitavanje slike u aplikaciji za FCM klastering

Grafički korisnički interfejs je prilično jednostavan, ali proračuni mogu biti veoma intenzivni. Opcije koje možemo da menjamo su broj klastera (mi smo podesili 2 klastera), maksimalni broj iteracija i preciznost. Algoritam će prestati pre maksimalnog broja iteracija, kada je postignuta zadata preciznost. Klikom na „Fuzzy C-means Clustering“ počinje računanje. Program počinje kreiranjem klastera početnih tački za svaki piksel na slici (Slika 44).

```
List<ClusterPoint> points = new List<ClusterPoint>();
for (int row = 0; row < originalImage.Width; ++row)
{
    for (int col = 0; col < originalImage.Height; ++col)
    {
        Color c2 = originalImage.GetPixel(row, col);
        points.Add(new ClusterPoint(row, col, c2));
    }
}
```

Slika 44. Kreiranje klaster tački za svaki piksel

Zatim se kreira zahtevani broj centara klastera (Slika 45).

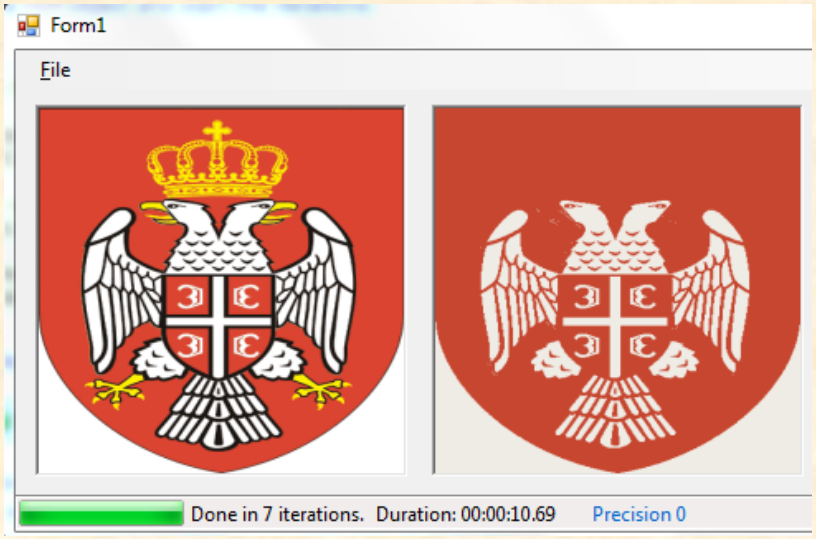
```
List<ClusterCentroid> centroids = new List<ClusterCentroid>();

//Create random points to use a the cluster centroids
Random random = new Random();
for (int i = 0; i < numClusters; i++)
{
    int randomNumber1 = random.Next(sourceImage.Width);
    int randomNumber2 = random.Next(sourceImage.Height);
    centroids.Add(new ClusterCentroid(randomNumber1, randomNumber2,
                  filteredImage.GetPixel(randomNumber1, randomNumber2)));
}
```

Slika 45. Kreiranje zahtevanog broja centroida (centara) klastera

Klaster centri su nasumice izabrani za prvu iteraciju, a u nastavku su prilagođeni prema algoritmu.

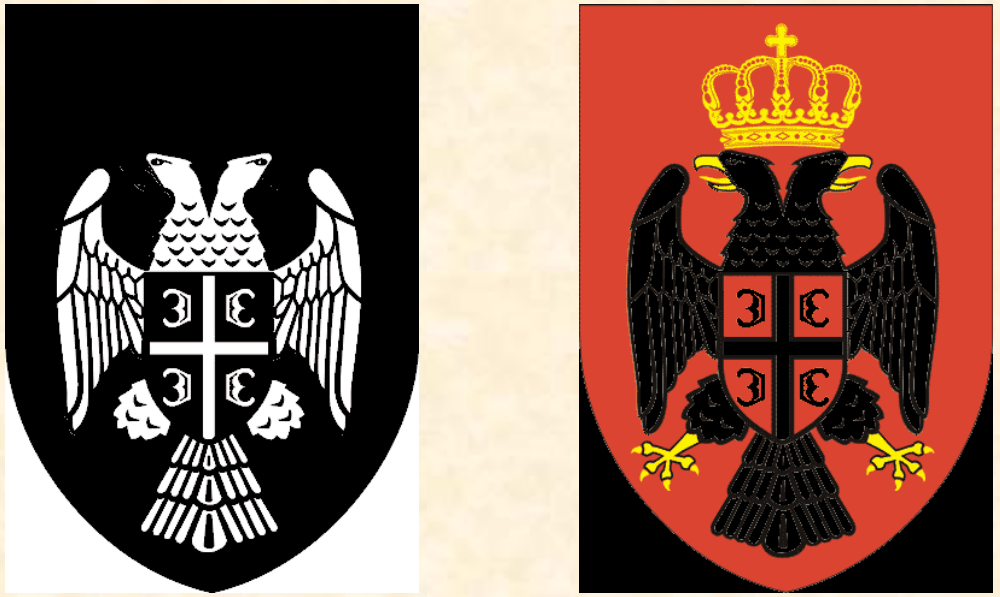
Konačno, kreira se FCM objekat i počinje se sa iteracijama. Napredak se prikazuje na statusnoj traci (Slika 46).

```
FCM alg = new FCM(points, centroids, 2, filteredImage, (int)numericUpDown2.Value);
k++;
alg.J = alg.CalculateObjectiveFunction();
alg.CalculateClusterCentroids();
alg.Step();
double Jnew = alg.CalculateObjectiveFunction();
Console.WriteLine("Run method i={0} accuracy = {1} delta={2}",
                  k, alg.J, Math.Abs(alg.J - Jnew));
backgroundWorker.ReportProgress((100 * k) / maxIterations, "Iteration " + k);
if (Math.Abs(alg.J - Jnew) < accuracy) break;
```



Slika 46. Kreiranje FCM objekta

Kada se iteracija izvrši, algoritam će izvršiti proces defazifikacije (izvođenje tačnog rezultata iz fazi skupa) i dodeliti piksel klasteru kome on pripada u najvećem stepenu. Za svaki klaster, program će zatim sačuvati segmentiranu sliku koja sadrži piksele iz originalne slike.
Kao primer, prikazaćemo dve segmentirane slike, od kojih jedna pripada prvom klasteru (slika levo), a druga drugom klasteru (slika desno) što možemo videti na Slici 47.



Slika 47. Prikaz segmentiranih slika koje pripadaju različitim klasterima

Zaključak

Klaster analiza vrši grupisanje jedinica posmatranja u grupe ili klase tako da se slične jedinice nađu u istoj klasi (klasteru). Grupisanje se vrši na osnovu rezultata (skora) koji se izračunava na osnovu vrednosti obeležja po svim varijablama, za svaku jedinicu posmatranja posebno. Metod koji se koristi za klasifikaciju mora biti potpuno numerički, a broj klasa se unapred ne zna kod hijerarhijskih metoda, dok se kod nehijerarhijskih metoda određuje unapred.

Postoji mnogo razloga za upotrebu klaster analize. Na primer, 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. Klaster analiza ima veliku primenu i u biologiji, medicini kada se vrše neka odvajanja u grupe. Na primer, u medicini, kada je potrebno snimke razdvojiti na celine radi lakšeg posmatranja i postavljanja dijagnoze. Vrlo je bitno znati da što je više varijabli uključeno u analizu i što su one više međusobno nezavisne, teže je pronaći odgovarajući obrazac za grupisanje jedinica posmatranja.

Nema jednostavnog pravila za donošenje odluke o tome koju metodu klaster analize je najbolje koristiti. Nekad je bolje koristiti hijerarhijske metode gde se ne zna unapred broj klastera, a nekad je bolje koristiti nehijerarhijske metode gde zadamo broj klastera unapred. Stiče se utisak da je pravi izbor kombinovanje više metoda i logično donošenje relevantnog zaključka za posmatranu analizu.