Klaster analiza

Uvod

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 obično ne zna.

Postoji mnogo razloga za upotrebu klaster analize. Na primer, prilikom segmentacije tržišta kada se formiraju klasteri potrošača u nekoj zemlji, pa se onda pravi poseban plan poslovnih aktivnosti za svaki klaster pojedinačno. Pored toga, 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 obrazac za grupisanje jedinica posmatranja.

Tipovi klaster analize

Mnogi algoritmi su korišćeni za klaster analizu. Ipak, dva pristupa su se izdvojila kao najbolja. Prvi je hijerarhijski metod koji kao krajnji rezultat ima dendogram. 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. Tehnika spajanja (aglomerativni, hijerarhijski metod) polazi od toga da je svaka jedinica sama u grupi od jednog člana. Bliske grupe se postepeno spajaju dok se na kraju ne nađu sve jedinice u jednoj grupi. Kod tehnike razdvajanja ide se obrnutim redosledom, gde se od jedne grupe stvaraju dve, pa od te dve sledeće dve i tako sve dok ne bude svaka jedinica posmatranja posebno. To je takozvani divizioni hijerarhijski metod koji se, ipak, primenuje mnogo ređe nego aglomerativni.

Drugi pristup, nehijerarhijski, je da se 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 poenta je 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 za unapred zadani broj grupa.

Hijerarhijski metod

Aglomerativni, hijerarhijski metod počinje sa matricom udaljenosti između jedinica posmatranja. Sve jedinice su u grupama veličine jedan, a zatim se vrši spajanje u veće grupe koje su blizu jedna druge. Postoji više načina da se definiše šta je to blizu. Najjednostavniji način je preko najbližih suseda. Na primer, u tabeli su date udaljenosti između pet jedinica posmatranja. Grupisanje je zatim prikazano u tabeli.

Tabela: Udaljenosti između jedinica posmatranja

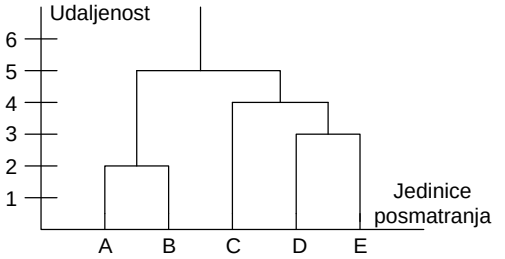
Jedinica posmatranja	Jedinica posmatranja				
	A	B	C	D	E
A	-				
B	2	-			
C	6	5	-		
D	10	9	4	-	
E	9	8	5	3	-

Tabela: Grupisanje na osnovu udaljenosti najbližih suseda

Udaljenost	Grupe
0	A, B, C, D, E
2	(A, B), C, D, E
3	(A, B), C, (D, E)
4	(A, B), (C, D, E)
5	(A, B, C, D, E)

Dve grupe se spajaju u jednu ukoliko je neka jedinica posmatranja iz jedne grupe najbliža nekoj jedinici iz druge grupe. Pri nultoj udaljenosti sve jedinice su posebno, odnosno ima onoliko grupa koliko ima i jedinica posmatranja. Zatim, najmanja udaljenost između dve jedinice je 2 koja postoji

između jedinica posmatranja A i B. Zbog toga na nivou udaljenosti od 2 imamo četiri grupe: (A, B), (C), (D) i (E). Sledeća najmanja udaljenost je 3 i imamo tri grupe: (A, B), (C) i (D, E) itd. Na kraju su sve jedinici u jednoj grupi, na udaljenosti 5. Slika pokazuje dendogram claster analize u kojem se vidi kako je vršeno grupisanje.

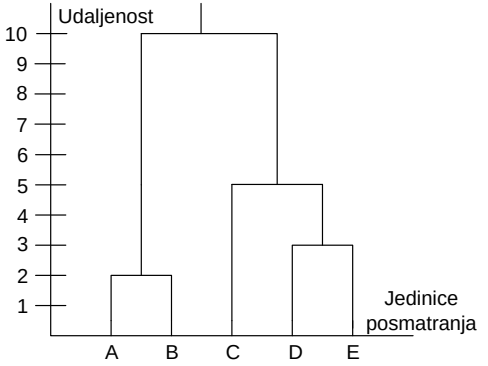


Slika: Grupisanje najbližih suseda

Kod povezivanja najdaljih suseda, dve grupe će se spoijiti ukoliko je udaljenost najdaljih članova iz dve grupe najmanja. Na primer, posle formiranja grupa (A,B), (C) i (D, E), jedinica posmatranja C će biti pridružena grupi (D, E) na nivou udaljenosti od 5 jer je to udaljenost između najudaljenijih jedinica iz dve grupe. Dve poslednje grupe (A, B) i (C, D, E) će biti spojene na nivou udaljenosti 10 jer je to najveća udaljenost između jedinica A i D. Postupak je prikazan u tabeli i na slici.

Tabela: Grupisanje na osnovu udaljenosti najbližih suseda

Udaljenost	Grupe
0	A, B, C, D, E
2	(A, B), C, D, E
3	(A, B), C, (D, E)
5	(A, B), (C, D, E)
10	(A, B, C, D, E)

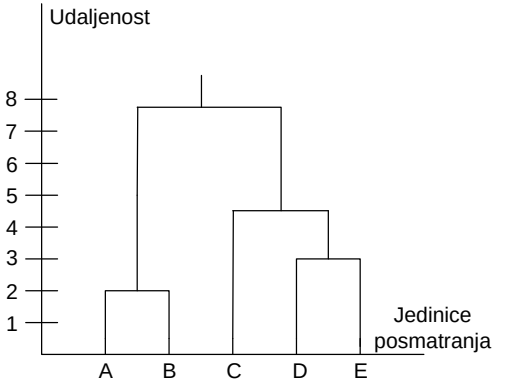


Slika: Grupisanje najdaljih suseda

Treća tehnika grupisanja je na osnovu prosečnih udaljenosti, odnosno kada je prosečna udaljenost između dve grupe najmanja u odnosu na sve ostale prosečne udaljenosti između grupa. Ovo grupisanje je predstavljeno u tabeli i na slici. Na primer, grupe (A, B) i (C, D, E) su spoljene na nivou udaljenosti od 7,8 jer je to prosečna udaljenost koja je izračunata na sledeći način: Udaljenosti između jedinice A iz prve grupe i jedinica C, D i E iz druge grupe su 6, 10 i 9 respektivno, dok su udaljenosti između jedinice B iz prve grupe i jedinica C, D i E iz druge grupe 5, 9 i 8 respektivno. Prosečna vrednost je aritmetička sredina: (6+10+9+5+9+8):6=7,8.

Tabela: Grupisanje na osnovu prosečne udaljenosti grupa

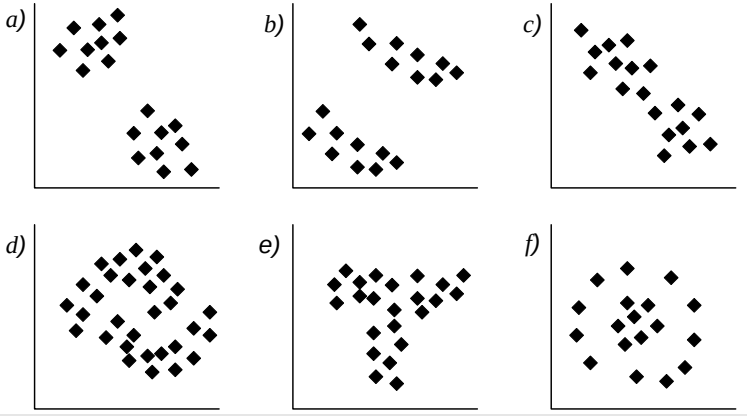
Udaljenost	Grupe
0	A, B, C, D, E
2	(A, B), C, D, E
3	(A, B), C, (D, E)
4,5	(A, B), (C, D, E)
7,8	(A, B, C, D, E)



Slika: Grupisanje na osnovu prosečne udaljenosti grupa

Razlike između tehnika grupisanja

Nijedna od postojećih tehnika grupisanja podataka u klastere se nije izdvojila kao najbolja i najupotrebljavanija. Problem je u tome što različite tehnike mogu da daju sasvim različite rezultate. Veliki uticaj ima priroda originalnih podataka. Ukoliko postoji određeni stepen preklapanja između podataka, vrlo je moguće da će različite tehnike grupisanja dati i različite rezultate.



Slika: Mogući rasporedi originalnih podataka kada postoje dva klastera (grupe)

Na slici su prikazani neki od mogućih rasporeda originalnih podataka. U slučaju *a*) i *b*) sve tehnike će verovatno dati iste klastere. U slučaju *c*) neke tehnike možda neće uspeti da izdvoje dva klastera. U slučajevima *d*), *e*) i *f*) će većina tehika imati teškoća u definisanju dva klastera.

Veliku ulogu u grupisanju ima subjektivni faktor, odnosno sam pristup u određivanju varijabli na osnovu kojih će se vršiti analiza. Izabrane varijable moraju biti relevantne u odnosu na klasifikaciju

koja se traži. U slučaju klaster analize proizvođača mobilnih telefona sa aspekta potrošača bilo bi najverovatnije nepotrebno kao jednu od varijabli navesti broj radnika muškog i ženskog pola.

Merenje udaljenosti

Originalni podaci na osnovu kojih se vrši klaster analiza obično predstavljaju raspored *n* jedinca posmatranja u odnosu na *p* varijabli. Na primer, napravljena je lista kupaca automobila i evidentiran je njihov pol, stručna sprema, boja vozila koje su kupili, kubikaža, plaćena cena, itd.

Za izračunavanje udaljenosti jedinica posmatranja obično se koristi Euklidova funkcija:

$$d_{ij} = \left\{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right\}^{\frac{1}{2}}$$

gde je x_{ik} vrednost jedinice posmatranja x_i za varijablu X_k, a x_{jk} vrednost jedinice posmatranja x_j za varijablu X_k..

Varijable su obično standardizovane pre izračunavanja udaljenosti da bi svih p varijabli bilo u jednakom položaju. To znači da će aritmetička sredina za svaku varijablu biti jednaka nuli, a standardna devijacija jedinici. Na žalost, standardizacija ima i jedan negativan efekat, a to je što se na taj načini minimiziraju razlike između klastera.

Neke klaster analize započinju sa izračunavanjem glavnih komponenti da bi se smanjio broj originalnih varijabli. Na ovaj način se smanjuje računski deo posla u klaster analizi ali se na taj način dobijaju i drugačiji rezultati. Danas se ipak analiza glavnih komponenti uglavnom izbegava zbog toga.

Primer: Klaster analiza evropskih zemalja – hijerarhijski metod

Za klaster analizu će poslužiti podaci o zaposlenosti u evropskim zemljama (tabela).

Tabela: Procenat radne snage zaposlen u devet grana industrije u 30 zemalja Evrope

Country	Group	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
Belgium	EU	2.6	0.2	20.8	0.8	6.3	16.9	8.7	36.9	6.8
Denmark	EU	5.6	0.1	20.4	0.7	6.4	14.5	9.1	36.3	7
France	EU	5.1	0.3	20.2	0.9	7.1	16.7	10.2	33.1	6.4
Germany	EU	3.2	0.7	24.8	1	9.4	17.2	9.6	28.4	5.6
Ireland	EU	22.2	0.5	19.2	1	6.8	18.2	5.3	19.8	6.9
Greece	EU	13.8	0.6	19.8	1.2	7.1	17.8	8.4	25.5	5.8
Italy	EU	8.4	1.1	21.93	0	9.1	21.6	4.6	28	5.3
Luxembourg	EU	3.3	0.1	19.6	0.7	9.9	21.2	8.7	29.6	6.8
Netherlands	EU	4.2	0.1	19.2	0.7	0.6	18.5	11.5	38.3	6.8
Portugal	EU	11.5	0.5	23.6	0.7	8.2	19.8	6.3	24.6	4.8
Spain	EU	9.9	0.5	21.1	0.6	9.5	20.1	5.9	26.7	5.8

U.K.	EU	2.2	0.7	21.3	1.2	7	20.2	12.4	28.4	6.5
Austria	EFTA	7.4	0.3	26.9	1.2	8.5	19.1	6.7	23.3	6.4
Finland	EFTA	8.5	0.2	19.3	1.2	6.8	14.6	8.6	33.2	7.5
Iceland	EFTA	10.5	0	18.7	0.9	10	14.5	8	30.7	6.7
Norway	EFTA	5.8	1.1	14.6	1.1	6.5	17.6	7.6	37.5	8.1
Sweden	EFTA	3.2	0.3	19	0.8	6.4	14.2	9.4	39.5	7.2
Switzerland	EFTA	5.6	0	24.7	0	9.2	20.5	10.7	23.1	6.2
Albania	Eastern	55.5	19.4	0	0	3.4	3.3	15.3	0	3
Bulgaria	Eastern	19	0	35	0	6.7	9.4	1.5	20.9	7.5
Czech/Slovak Rep.	Eastern	12.8	37.3	0	0	8.4	10.2	1.6	22.9	6.9
Hungary	Eastern	15.3	28.9	0	0	6.4	13.3	0	27.3	8.8
Poland	Eastern	23.6	3.9	24.1	0.9	6.3	10.3	1.3	24.5	5.2
Romania	Eastern	22	2.6	37.9	2	5.8	6.9	0.6	15.3	6.8
USSR (form.)	Eastern	18.5	0	28.8	0	10.2	7.9	0.6	25.6	8.4
Yugoslavia (form.)	Eastern	5	2.2	38.7	2.2	8.1	13.8	3.1	19.1	7.8
Cyprus	Other	13.5	0.3	19	0.5	9.1	23.7	6.7	21.2	6
Gibraltar	Other	0	0	6.8	2	16.9	24.5	10.8	34	5
Malta	Other	2.6	0.6	27.9	1.5	4.6	10.2	3.9	41.6	7.2
Turkey	Other	44.8	0.9	15.3	0.2	5.2	12.4	2.4	14.5	4.4

Klaster analiza treba da pokaže koje zemlje imaju sličnu situaciju u strukturi zaposlenosti. Određena podela već postoji na zemlje evropske unije, zemlje EFTA, zemlje istočne evrope i grupu koju čine Kipar, Gibraltar, Malta i Turska. Interesantno će biti da se utvrdi da li će se rezultati klaster analize poklopiti sa ovom podelom.

Prvi korak u analizi je standardizacija originalnih podataka, odnosno transformacija podataka u seriju koja ima aritmetičku sredinu jednaku nuli i standardnu devijaciju jednaku jedinici. Na primer, za Belgiju vrednost obeležja varijable AGR iznosi 2,6. Aritmetička sredina varijable AGR je 12,19 a standardna devijacija 12,31. Nova standardizovana vrednost se dobija na sledeći način:

$$u = \frac{2{,}6 - 12{,}19}{12{,}31} = -0{,}78$$

itd.

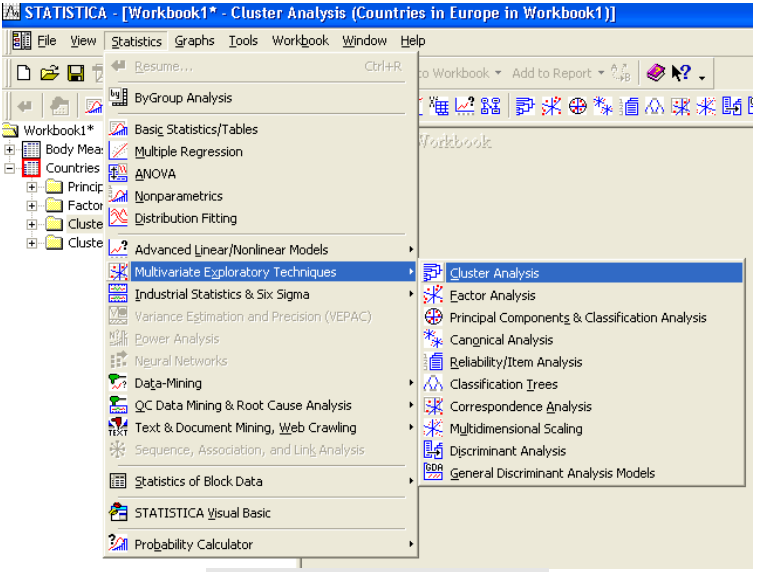
Sledeći korak je izračunavanje Euklidovih rastojanja između jedinica za svaki par standardizovanih vrednosti. Nakon toga se pristupa izradi dendograma koji predstavlja grafički prikaz rezultata klaster analize.

Koraci za izvođenje analize u programu su sledeći:

Pokretanje analize:

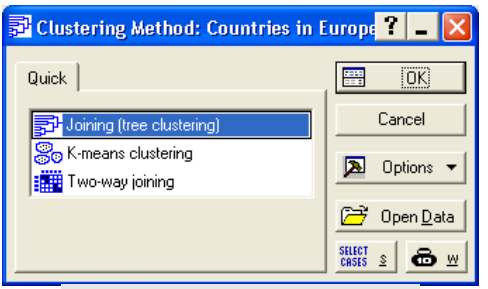
Statistics ▸ Multivariate Exploratory Technique ▸ Cluster Analysis

Dobija se početni meni za analizu.



Slika: Početni meni za analizu

Pokretanjem klaster analize dobija se meni iz kojeg se bira tip klaster analize (slika).



Slika: Izbor tipa klaster analize

Pritiskom na „OK“ dobija se glavni meni za klaster analizu.

Slika: Kratki meni za klaster analizu

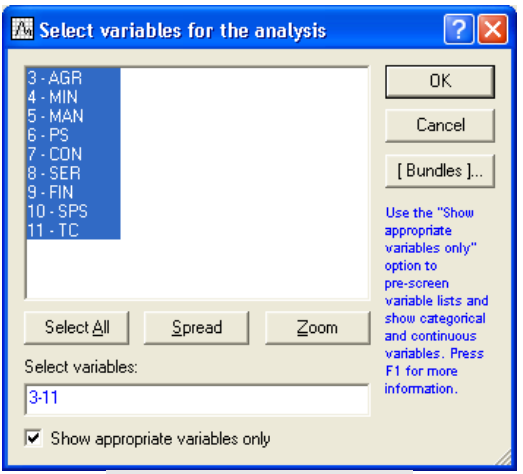
U glavnom meniju za klaster analizu potrebno je definisati opcije onako kako je prikazano na slici broj (Row data; Cases (rows); Single Linkage; Euclidean distances).

Definisanje varijabli:

Quick (ili Advanced) ▸ Variables

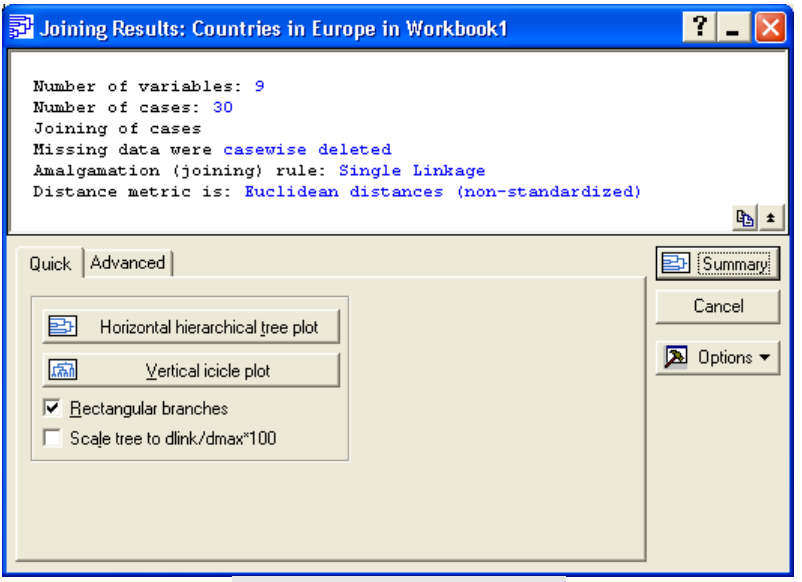
Otvara se prozor sa spiskom varijabli od kojih treba odabrati one koje će biti uvrštene u analizu.

▸ OK

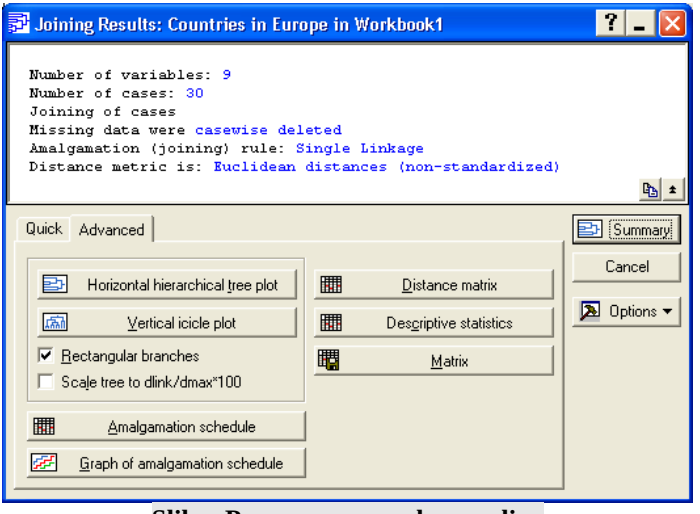


Slika: Definisanje varijabli

Brza analiza može da se obavi preko kratkom menija („Quick“) ili preko naprednog menija („Advanced“).



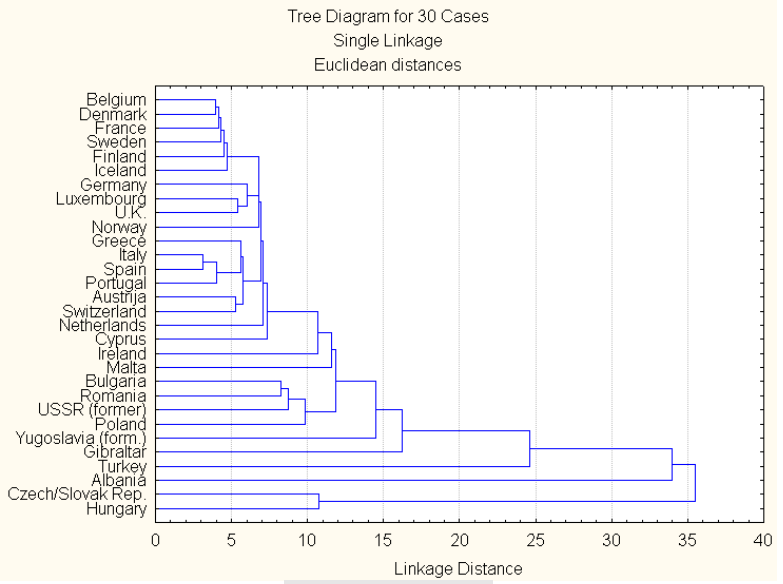
Slika: Prozor za brzu analizu



Slika: Prozor za naprednu analizu

Izrada dendograma:

Advanced (ili Quick) ▸ Horizontal hierarchical tree plot (ili Vertical icicle plot)



Slika: Dendogram

Uočava se da su dve najbliže i najsličnije zemlje Italija i Španija, a zatim slede Belgija i Danska, kojima se kasnije pridružuje Francuska pa Švedska itd.

Izračunavanje Euklidovih rastojanja:

Advanced ▸ Distance matrix

Case No.	Euclidean distances (Countries in Europe in Workbook1)								
	Belgium	Denmark	France	Germany	Ireland	Greece	Italy	Luxembourg	Netherlands
Belgium	0.0	3.9	4.9	10.0	26.3	16.1	12.8	9.3	7.1
Denmark	3.9	0.0	4.2	10.3	24.0	14.1	12.7	10.4	7.9
France	4.9	4.2	0.0	7.3	22.3	11.8	10.1	6.8	8.7
Germany	10.0	10.3	7.3	0.0	22.2	12.4	9.0	6.9	14.7
Ireland	26.3	24.0	22.3	22.2	0.0	10.7	16.9	22.0	27.3
Greece	16.1	14.1	11.8	12.4	10.7	0.0	8.6	12.2	17.6
Italy	12.8	12.7	10.1	9.0	16.9	8.6	0.0	7.4	16.3
Luxembourg	9.3	10.4	6.8	6.9	22.0	12.2	7.4	0.0	13.4
Netherlands	7.1	7.9	8.7	14.7	27.3	17.6	16.3	13.4	0.0
Portugal	16.1	15.0	12.4	10.2	12.9	5.6	5.7	11.1	18.7
Spain	13.7	12.8	10.1	9.2	14.7	6.0	3.1	8.0	16.9
U.K.	9.9	10.9	7.0	6.1	23.1	12.9	10.5	5.4	12.3
Austrija	16.1	15.7	12.9	7.9	17.2	10.2	7.9	11.0	19.5
Finland	7.5	4.5	4.6	10.0	19.8	10.1	10.6	9.7	10.4
Iceland	11.2	8.5	7.4	10.4	17.0	7.7	9.4	10.0	14.6
Norway	7.3	7.0	7.9	14.6	24.7	15.5	13.8	11.1	8.8
Sweden	4.3	4.3	7.4	13.3	28.0	18.0	16.1	12.7	7.7
Switzerland	15.5	15.5	11.9	6.9	18.9	10.8	8.9	8.9	18.5
Albania	72.2	69.2	68.1	69.1	50.8	58.4	65.6	69.1	71.5
Bulgaria	28.9	26.7	26.3	23.4	18.8	20.1	22.4	27.6	31.5
Czech/Slovak Rep.	47.0	45.9	45.4	46.8	43.6	43.1	44.5	45.6	47.9
Hungary	40.0	38.8	38.6	41.1	36.6	36.1	37.6	39.0	40.5
Poland	26.8	23.9	23.8	23.8	11.9	15.3	20.1	25.7	28.5
Romania	36.2	34.0	33.4	30.0	22.9	26.2	29.2	34.4	38.5
USSR (former)	24.7	22.0	22.2	20.9	16.8	16.7	19.3	24.0	28.0
Yugoslavia (form.)	26.3	26.0	24.6	18.6	26.6	23.0	21.2	24.0	30.1

Slika: Euklidova rastojanja između jedinica posmatranja

Tabelarni prikaz toga grupisanja:

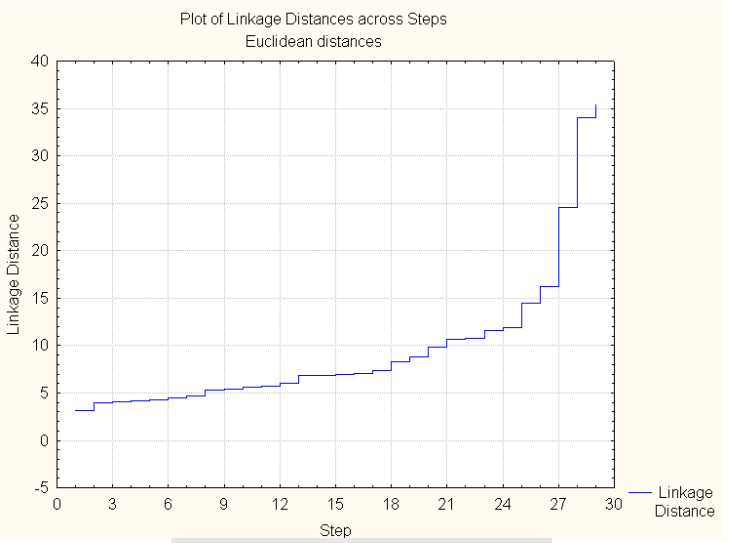
Advanced ▸ Amalgamation schedule

	Euclidean distances								
linkage distance	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
3.114306	Italy	Spain							
3.938274	Belgium	Denmark							
4.021194	Italy	Spain	Portugal						
4.184495	Belgium	Denmark	France						
4.269660	Belgium	Denmark	France	Sweden					
4.489989	Belgium	Denmark	France	Sweden	Finland				
4.689350	Belgium	Denmark	France	Sweden	Finland	Iceland			
5.304715	Austrija	Switzerland							
5.416641	Luxembourg	U.K.							
5.605355	Greece	Italy	Spain	Portugal					
5.742822	Greece	Italy	Spain	Portugal	Austrija	Switzerland			
6.058053	Germany	Luxembourg	U.K.						
6.813956	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
6.816891	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
6.917369	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
7.070361	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
7.376991	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
8.239539	Bulgaria	Romania							
8.769264	Bulgaria	Romania	USSR (former)						
9.850381	Bulgaria	Romania	USSR (former)	Poland					
10.70187	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
10.76801	Czech/Slovak Rep.	Hungary							
11.57800	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
11.88192	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
14.50586	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
16.22745	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.
24.61138	Belgium	Denmark	France	Sweden	Finland	Iceland	Germany	Luxembourg	U.K.

Slika: Tabelarni prikaz toka grupisanja jedinica posmatranja

Za grafički prikaz grupisanja:

Advanced ▸ Graph of amalgamation schedule



Slika: Grafički prikaz grupisanja

Nehijerarhijski metod ili metod raščlanjivanja (Partition Clustering)

Hijerarhijski metod gradi klastere korak po korak, sve dok se sve jedinice posmatranja ne nađu na dendogramu. Tek nakon toga se pristupa određivanju broja klastera koji imaju značaja za istraživača. Nehijerarhijski metod ili metod raščlanjivanja polazi od unapred određenog broja klastera koji istraživač sam definiše na osnovu iskustva, ranijih analiza ili preporuke statističkog softvera. Nakon toga se pristupa razvrstavanju jedinica posmatranja.

Postoje dva načina za razvrstavanje jedinica posmatranja. Prvi je da se privremeno, na slučajan načini, odrede jedinice koje predstavljaju tačke grupisanja, pa se na osnovu udaljenosti od tih jedinica sve ostale jedinice smeštaju u odgovarajući klaster. Određuje se onoliko tački grupisanja koliko je unapred definisano klastera. Nakon toga računarski program premešta jedinice iz jednog u drugi klaster da bi bili što homogeniji, Taj postupak se ponavlja nekoliko puta.

Drugi način raščlanjivanja je da se razvrstavanje odvija na osnovu nekog a priori zadatog kriterijuma.

Tipični algoritam za metod raščlanjivanja podrazumeva sledeće korake:

1. Proizvoljno određivanje privremenih tačaka grupisanja.
2. Program pronalazi tačku u prostoru unutar svakog klastera tako da su udaljenosti jedinica svedene na minimum. Ova tačka se naziva klasterov centroid. Centroidi se uglavnom nalaze tamo gde je najveća gustina jedinica.
3. Centroidi se koriste kao nove tačke grupisanja za nove klastere jer su mnogo relevantniji za formiranje klastera nego inicijalne, proizvoljne tačke.

4. Izračunavanje udaljenosti svih jedinica posmatranja u odnosu na centroide radi započinjanja nove iteracije.
5. Određivanje novih centroida unutar klastera.
6. Nastavljanje iteracija sve do situacije kada preseljavanje jedinica posmatranja iz jednog u drugi klaster više ne doprinosi poboljšanju homogenosti unutar klastera.

Različitost klastera

Da bi se utvrdilo da li se dobijeni klasteri zaista međusobno razlikuju koriste se tri tehnike:

- Vizuelno posmatranje grafičkih prikaza (dendograma, različitih linijskih dijagrama itd.).
- Testiranje statističke značajnosti razlika između klastera preko, na primer, analize varijanse.
- Upoređivanje dobijenih rezultata sa rezultatima diskriminacione analize nad istim podacima.
- Posmatranje klastera u kontekstu karakteristika originalnih varijabli na osnovu kojih je izvršena analiza.

Određivanje relevantnog broja klastera

Nakon klaster analize postavlja se pitanje koji broj klastera je od najvećeg značaja. Istraživač sam treba da prosudi, u kontekstu svog istraživanja, koji broj klastera i sa kakvim karakteristikama mu je potreban. Na donošenje odluke uticaj mogu da imaju sledeći faktori:

- statistička značajnost razlike između klastera,
- veličina klastera,
- veličina uzorka,
- dekompozicija klastera na nove klastere,
- osobine klastera koje imaju smisla u kontekstu karakteristika originalnih varijabli i samog istraživanja,
- karakteristike klastera koje su od interesa za istraživača (menadžment preduzeća),
- karakteristike klastera u kontekstu varijabli koje nisu uvrštene u analizu (demografske, političke karakteristike i slično).

Primer: Klaster analiza evropskih zemalja – nehijerarhijski metod

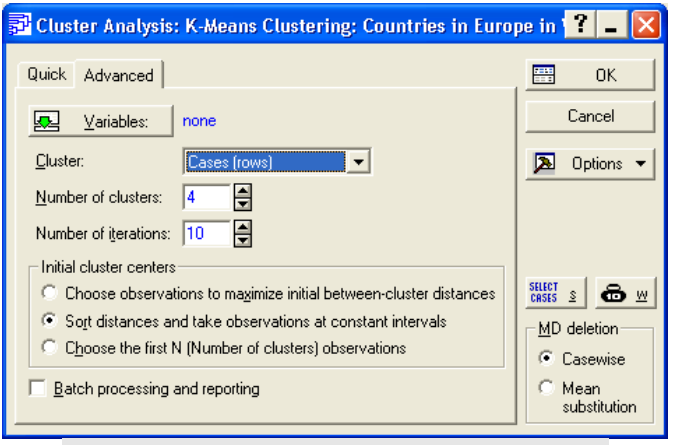
Na istim podacima biće prikazan i nehijerarhijski metod k-sredina (k-means).

Pokretanje analize:

Statistics ▸ Multivariate Exploratory Technique ▸ Cluster Analysis ▸ K-means clustering

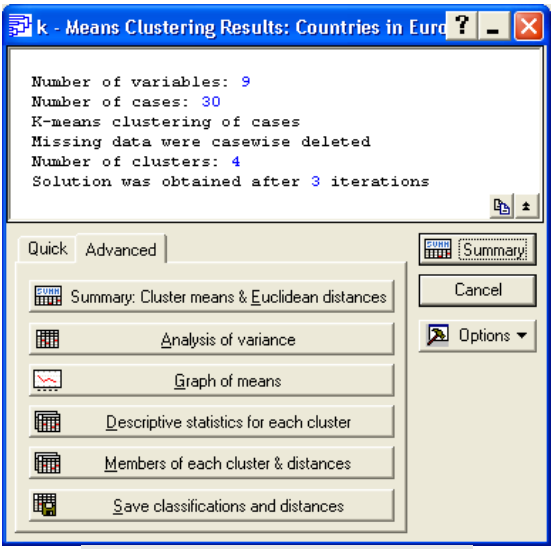
Slika: Izbor tipa klaster analize

U prozoru „Advanced“ definisati opcije onako kako je prikazano na slici. Odlučeno je da se formira četiri klastera, što svakako zavisi od želje istraživača.



Slika: Definisanje opcija u prozoru „Advanced“

Pritiskom na „OK“ otvara se prozor za naprednu analizu („Advanced“).



Slika: Prozor za naprednu analizu

Analiza varijanse nam pokazuje da li postoje statistički značajne razlike između formiranih klastera (grupa) s obzirom na svaku varijablu posebno. Rezultati analize varijanse se dobijaju preko sledećih opcija:

Advanced ▸ Analysis of variance

Variable	Analysis of Variance (Countries in Europe in Workbook1)					
	Between SS	df	Within SS	df	F	signif. p
AGR	3802.267	3	590.0678	26	55.84609	0.000000
MIN	2057.951	3	221.4837	26	80.52769	0.000000
MAN	1783.481	3	810.1115	26	19.07988	0.000001
PS	2.686	3	8.4938	26	2.74083	0.063587
CON	24.373	3	192.2503	26	1.09872	0.367387
SER	404.229	3	367.9603	26	9.52092	0.000204
FIN	264.719	3	196.1963	26	11.69353	0.000049
SPS	1332.251	3	878.9678	26	13.13606	0.000021
TC	21.640	3	22.4745	26	8.34493	0.000473

Slika: Rezultati analize varijanse

Za pregled svih klastera i međusobnih udaljenosti jedinica posmatranja:

Advanced ▸ Members of each cluster&distances

	Members of Cluster Number 1 (Countries in Europe in Workbook1) and Distances from Respective Cluster Center Cluster contains 2 cases
	Distance
Czech/Slovak Rep.	1.794668
Hungary	1.794668

Slika: Članovi prvog klastera

	Members of Cluster Number 2 (Countries in Europe in Workbook1) and Distances from Respective Cluster Center Cluster contains 20 cases
	Distance
Belgium	2.439646
Denmark	2.248458
France	1.163981
Germany	2.123542
Greece	3.119150
Italy	2.243078
Luxembourg	1.681020
Netherlands	3.702787
Portugal	3.107061
Spain	2.218758
U.K.	2.257343
Austrija	3.426621
Finland	1.710178
Iceland	2.072898
Norway	3.048906
Sweden	3.402699
Switzerland	3.232288
Cyprus	4.527058
Gibraltar	6.378330
Malta	5.551452

Slika: Članovi drugog klastera

	Members of Cluster Number 3 (Countries in Europe in Workbook1) and Distances from Respective Cluster Center Cluster contains 2 cases
	Distance
Albania	5.665539
Turkey	5.665539

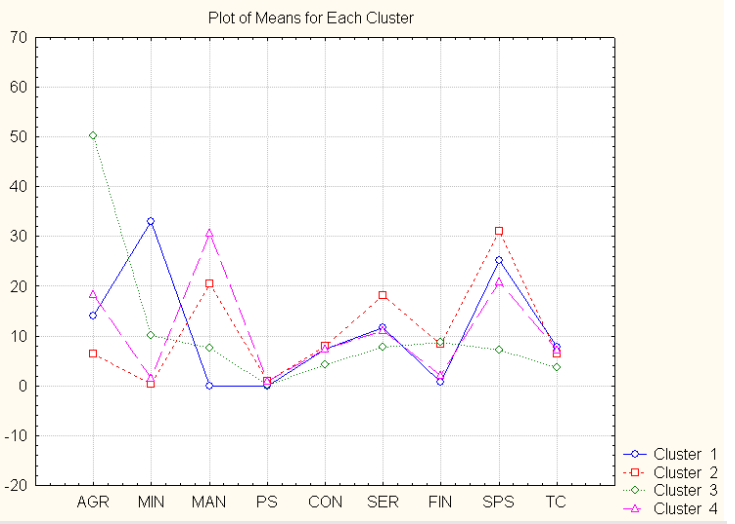
Slika: Članovi trećeg klastera

	Members of Cluster Number 4 (Countries in Europe in Workbook1) and Distances from Respective Cluster Center Cluster contains 6 cases
	Distance
Ireland	4.813423
Bulgaria	1.721586
Poland	3.237650
Romania	3.670625
USSR (former)	2.389593
Yugoslavia (form.)	5.364196

Slika: Članovi četvrtog klastera

Za grafičko upoređivanje klastera:

Advanced ▸ Graph of means



Slika: Dijagram srednjih vrednosti svih klastera po varijablama