Bogosavljević Srdjan
Institut za statistiku
Beograd

GRAFICKO PRIKAZIVANJE REZULTATA MULTIVARIJACIONE ANALIZE

GRAPHICAL PRESENTATION OF MULTIVARIATE ANALYSIS RESULTS

Abstarct

Utilization of drawings, diagrams - graphics in statistical literature has become a common practice worldwide, and consumers of statistics are used to perceive various facts through graphical presentations.

The facts comprehended through graphical presentations certainly can not be considered as full scientific proof, but, indisputably, a drawing can point out to the kind of scientific test that ought to be carried out. Conclusions resulting from statistical analysis or the results of statistical data processing can be illustrated very effectively.

This paper deals with application of graphics in presenting the results of multivariate statistical analysis, especially the results of numerical classification, i.e. cluster analysis.

---

1. Uvod

Upotreba crteža, dijagrama - grafike u statističkoj literaturi je odomaćena praksa i svi koji koriste statistiku su navikli da mnoge činjenice sagledaju na osnovu grafičkih prikaza.

Naravno da se činjenice sagledane na grafičkom prikazu ne mogu smatrati dokazanim, ali je nesporno da crtež može da ukaže na vrstu statističkog testa koji treba izvršiti. Veoma se efektno mogu ilustrovati zaključci statističkih analiza ili prikazati rezultati statističke obrade.

U nizu mogućih upotreba grafičkih tehnika u statistici pomenimo dve osnovne:

1. prikaz i opis podataka
2. analiza i interpretacija rezultata

Kad je reč o tzv. jednovarijantnoj (ili jednodimenzionalnoj) statistici odnosno standardnoj statističkoj analizi onda su iskustva sa upotrebom grafike i grafičkog prikazivanja vrlo velika. Korektni prikazi statističkih podataka se mogu naći i u vrlo starim publikacijama.

Danas su nam na raspolaganju ogromne grafičke mogućnosti računara. Na primer samo paket programa STATGRAPHIC - koji predstavlja interaktivni statistički grafički sistem radjen u APL-u, obuhvata oko 200 grafičkih mogućnosti svrstanih u 24 poglavlja koja obuhvataju skoro kompletno područje statističkih metoda od

analize varijansi
klaster analize
ocenjivanja i testiranja
do
uzorka
kontrole kvaliteta i
matematičkog programiranja

Još veće mogućnosti ima statistički analitički sistem SAS koji sadrži kompletan modul SAS-GRAPH posevećen statističkoj grafici i u okviru ostalih modula (BASIC, STATISTICS, ETS, OR ...) sadrži značjne grafičke opcije.

Upotreba grafike u multivarijacionoj statističkoj analizi je novijeg datuma i sve se više upotrebljava. Dva su osnova razloga zašto se statistička grafika u multivarijacionoj analizi šire koristi tek nešto više od 10 godina. Prvi razlog leži u činjenici da je multivarijaciona analiza izašla iz teorijskih okvira u širu praktičnu upotrebu tek masovnom pojavom brzih i moćnih računara početkom šezdesetih godina, i još više, pojavom dobro razradjenih softvera tokom sedamdesetih godina. Drugi razlog se nalazi u problemima vizuelne reprezentacije tačaka iz n-dimenzionalnog prostora. Prevazilaženje ovog problema vodilo je nizu zanimljivih matematičko-statističkih ogleda.

Konačno tek pojava grafičkih paketa omogućava masovnu primenu grafike u statističkim analizama.

U ovom radu će biti više reči o upotrebi grafike u prikazivanju rezultata multivarijacione statističke analize, a posebno rezultata numeričkog klasifikovanja, odnosno klaster analize.

2. Numeričko klasifikovanje

Procedura numeričkog klasifikovanja obuhvata

- definisanje problema
- izbor jedinica posmatranja (E)
- izbor obeležja (X)
- obezbedjenje matrice podataka (X)
- izbor i formiranje funkcije odstojanja (D)
- izbor strukture klasifikovanja (H(K) ili K)
- izbor i primena klasifikacionog algoritma
- prikaz, interpretacija i evaluacija rezultata

Procedure klasifikovanja mogu biti direktne tipa

$$X \Rightarrow H(K) \Rightarrow K$$

ili indirektne

$$X \Rightarrow D \Rightarrow H(K) \vee K$$

gde je

X	matrica podataka
D	matrica odstojanaj (sličnosti)
H(K)	hijerarhijska klasifikacija
K	nehijerarhijska klasiikacija

Matrica podataka X se dobija merenjem realizacija vektora obeležja

$$X = (X_1\ X_2 \ldots X_p)$$

nad skupom koji se klasifikuje

$$E = \{ e_1, e_2 \ldots e_N \}$$

pa je

$$X = [\ X^{(1)}, X^{(2)}, \ldots X^{(N)}\ ]$$

Jasno je da svako $X^{(1)}$ predstavlja tačku u prostoru E^p

Prikaz tačaka $X^{(1)}$ za p=1 i p=2 pa čak i p=3 je potpuno saglasan sa našim iskustvom, ali zap >3 nije lako ni zamisliti,a kamoli ostvariti taj prikaz.

Postoji niz rešenja kojim se više dimenzionalne tačke predstavljaju u dvodimenzionalnu prostu. Tako se koriste profili (niz paralelnih osa), poligoni (zvezdasti prikazi), face ilidruge figure i td.

3. Grafičko predstavljanje odnosa medju tačkama skupa koji se klasifikuje

Izbor mere odstojanja D predstavlja jedno od najvažnijih odluka u procesu klasifikovanja s obzirom na uticaj koji ima na kompletan izgled klasifikacije. Ne postoje pravila izbora koja bi vodila nekom optimalnom rešenju, ali se mogu sresti mnogi iskustveni kriterijumi koji su povezani sa predmetom istraživanja. Svakako da je najčešće upotrebljavana funkcija odstojanja Euklidsko odstojanje:

$$d_{ij} = \sum w_k (x_{ik} - x_{jk})^2$$

odnosno

$$d_{ij} = \sum (x_{ik} - x_{jk})^2 = (X^{(i)} - X^{(j)})'(X^{(i)} - X^{(j)})$$

Ovim odstojanjem su definisani odnos medju tačkama $X^{(i)}$ i $X^{(j)}$ u nekom p-dimenzionalnom euklidskom prostoru. Kako dovodimenzionalni euklidski prostor omogućava grafički prikaz odnosa potrebno je pronaši ortognalnu transformaciju koja p-dimenzionalne tačke $X^{(i)}$ prevodi u dvodimenzionalne tačke $Y^{(i)}$ uz maksimalno očuvanje odnosa medju tačkama. Problem se tako svodi na klasične probleme teorije skaliranja. Jedno moguće rešenje vodi prostoru definisanom prvim dvema glavnim komponentama.

Ako transformaciju označimo sa L= $L_1 L_2$ onda je

$$Y = L'X$$

uz uslov da je $L_1'L_1 + L_2'L_2 = I$ i $L_1'L_2 = 0$

Minimiziranje razlike izmedju sume svih odstojanja u E^P i E^2 dolazimo do rešenja da je

$$L_1 = P_1 \text{ i } L_2 = P_2 \quad (P_1 = GK_1)$$

pri čemu su P_1 i P_2 prve dve glavne komponente (odnosno karakteristični vektori) korelacione, odnosno kovarijacione matrice, što omogućava grafičko predstavljanje odnosa tačaka kao na slici br.1. Konkretno, radi se o granama delatnosti (skup E) koje karekteriše osam indikatora efikasnosti (vektor X)

O kvalitetu ove reprezentacije se može suditi na osnovu dva kriterijuma:

a) kriterijuma valjanosti glavnih komponenata u smislu očuvanja ukupnog varijabiliteta, datog sa

$$W_1 = (\lambda_1 + \lambda_2)/\Sigma\lambda_i$$

b) krietrijuma prilagodjenosti odnosa tačaka u prostoru niže dimenzije odnosu tačaka u prostoru više dimenzije - primer takvog kriterijuma je Kruskalov koeficijent „stress„:

$$W_2 = \sum_{ij}\{(d_{ij} - d_{ij}^{(2)})/d_{ij}\}^2$$

gde $d^{(2)}$ označava da se radi o meri odstojanja u dvodimenzionalnom prostoru.

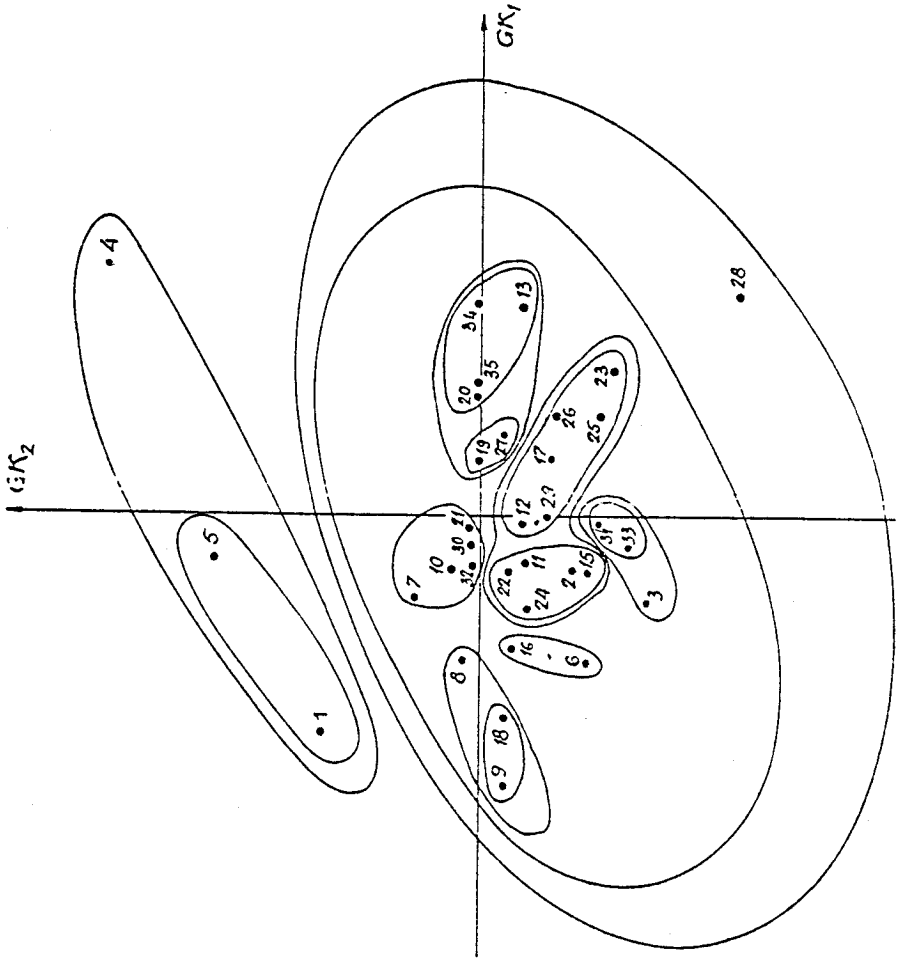
Zahvaljujući rezultatima C.R.Rao o glavnim komponentama neortogonalnih prostora, moguće je pronaći najbolje dvodimenzionalne prostore i za druge vrste odstojanja.

Tehnike multivarijacionog skaliranja nude niz rešenja baziranih na sličnim idejama. Na slikama 2 i 3 su dati primeri prostora prve dve glavne komponenete u kojima se vidi razdvajanje skupa jedinica na nekoliko homogenih grupa i izdvajanje 'outlier'-a.

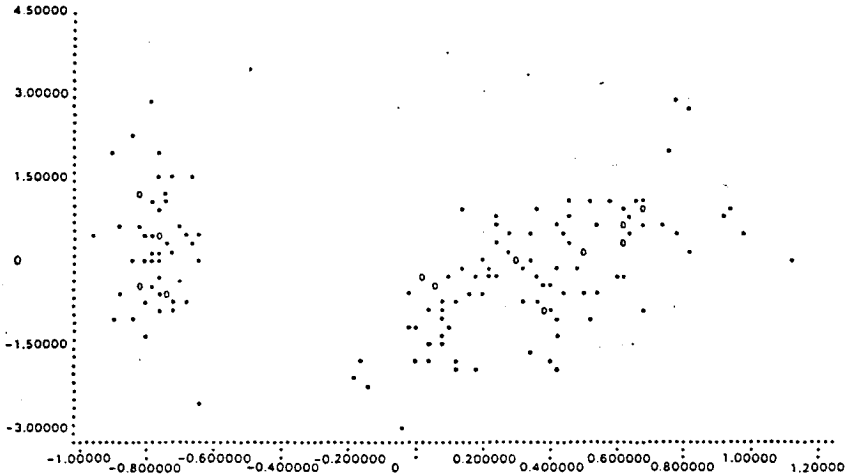
4. Prikazivanje hijerarhijske klasifikacije grafički

Ako imamo na umu da smo tačke iz E^P prostora prikazaliu E^2 prostoru onda je jasna ideja o prikazu hijerarhije u tom istom E^2 prostoru (slika br.1)
Ta ista hijerarhija se prikazuje i na druge načine. Najčešći način prikazivanje hijerarhijske klasifikacije je upotrebom dendrograma (slika br.4)

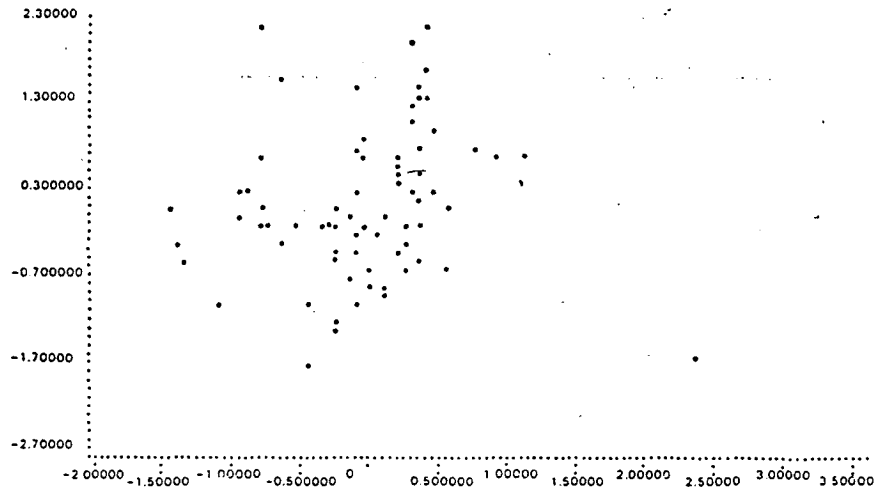
Dendrogram nije slučajno i najpopularniji metod prikaza rezultata numeričkog klasifikovanja. Njime je jasno izražena struktura klasa, a visinom čvora na kom se prvi put dva elementa spajaju (indeksom hijerarhije) je dat i kvalitet pojedinačnih klasa.

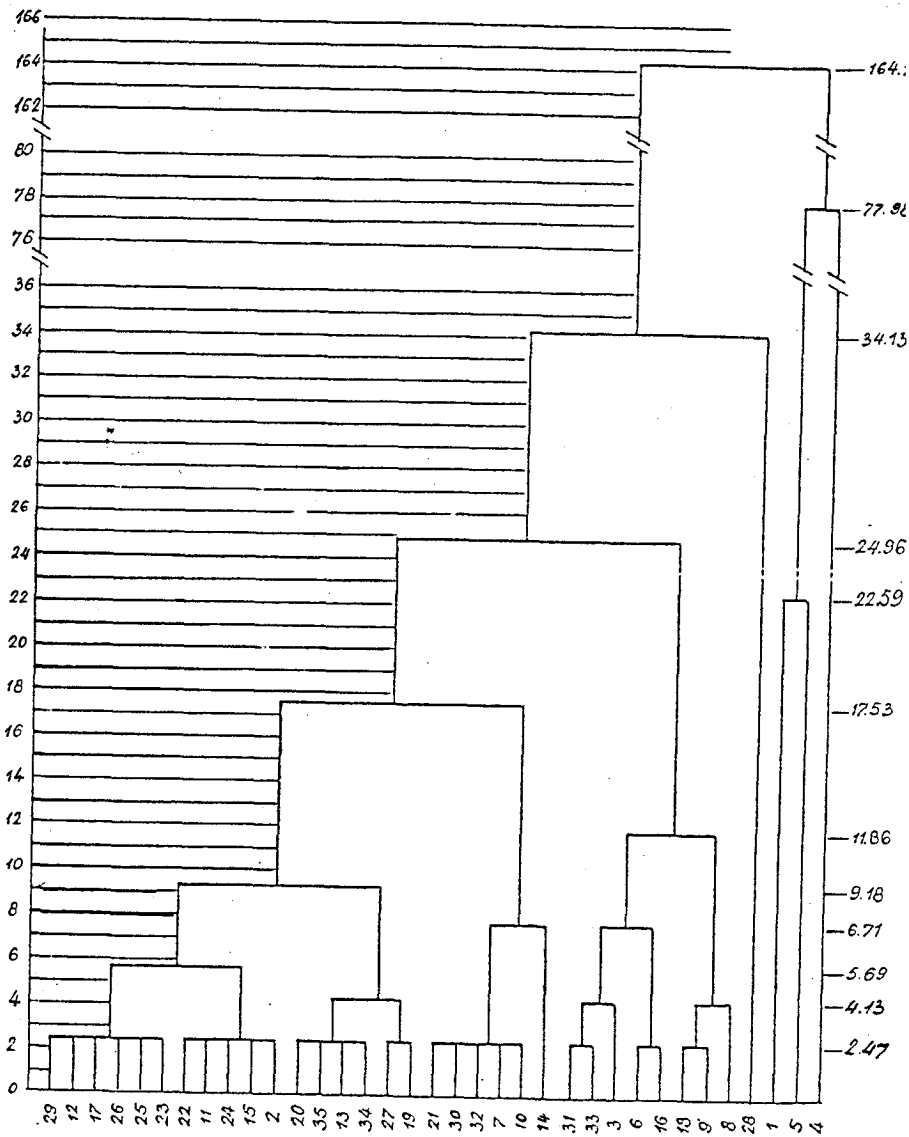

GK1
GK2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

Slika br.1


4.50000
3.00000
1.50000
0
-1.50000
-3.00000
-1.00000
-0.800000
-0.600000
-0.400000
-0.200000
0
0.200000
0.400000
0.600000
0.800000
1.00000
1.20000

Slika br. 2


2.30000
1.30000
0.300000
-0.700000
-1.70000
-2.70000
-2.00000
-1.50000
-1.00000
-0.500000
0
0.500000
1.00000
1.50000
2.00000
2.50000
3.00000
3.50000


166
164
162
80
78
76
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
164.
77.9
34.13
24.96
22.59
17.53
11.86
9.18
6.71
5.69
4.13
2.47
29 12 17 26 25 23 22 11 24 15 2 20 35 13 34 27 19 21 30 32 7 10 14 31 33 3 6 16 18 9 8 28 1 5 4

Slika br.4

Kod ove vrste (slično većini ostalih) grafičkog prikaza hijerarhije uočavamo jedan nedostatak. On se odnosi jednoznačnost prezentacije, koja može da se postigne tzv.uredjenjem klasifikacije.

Uobičajno se hijerarhijska klasifikacija predstavlja kao uredjeni niz klasifikacija tj. kao

$$H(K) = < K_1, K_2, K_3 \ldots K_N >$$

pri čemu je

$$K_1 = \{E_1^1, E_2^1, E_3^1 \ \ldots \ E_1^1\}$$

odnosno, pri čcemu je K_1 skup klasa. Medjutim, K_1 može biti i uredjeni niz klasa, odnosno

$$K_1 = <E_1^1, E_2^1, \ldots E_1^1>$$

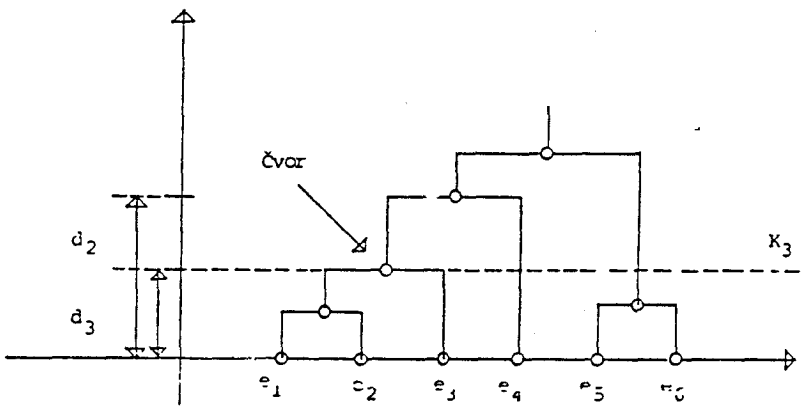
Zbog čega je potrebno i eksplicitno definisati relaci j u poretka ®

$$E_1^s \circledR E_1^m$$

kako bi bilo moguće konstruisati i alogoritam uredjenja hijerarhije koji bi se upotrebio prilikom grafičke prezentacije hijerarhije dendrogramom.

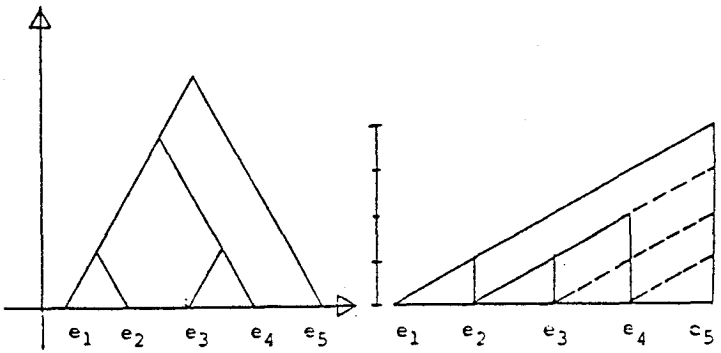
Sam oblik relacije uredjenja je obično diktiran prirodom problema, mada može biti baziran i na nekim opštim zahtevima.

Sa dedrograma (slika 5) se lako čita indeks hijerahije,koji obično predstavlja homogenost klase, odnosno klasifikacije, i koji je odredjen visinom (d_2, d_3) na kojoj se klasa prpoznaje u dendrogramu. Sve jedinice skupa E koje se sučeljavaju u jednom čvoru čine klasu sa indeksom hijerarhije jednakom visini tog čvora.

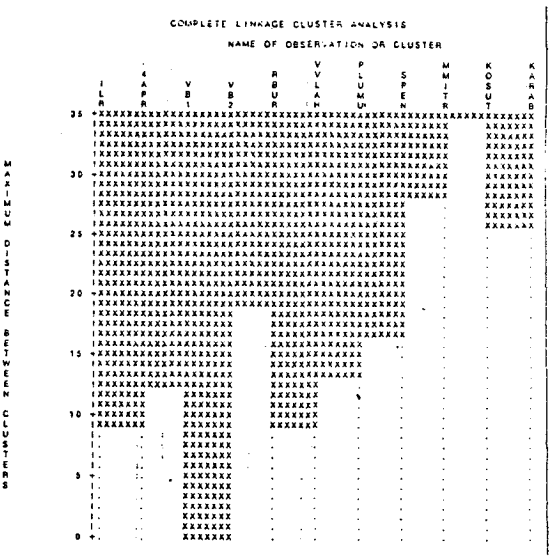


Slika br.5

Osim dendrograma na slikama 4 i 5 za grafičko prikazivanje hijerarhijskih klasifikacija se koristi jos niz drvolikih prikaza. Na slici 6 je dat oblik koji se koristi za prikaz hijerarhije u okviru programskog paketa BMDP, a na slici 7 ledenični dijagram (icily plot) koji se koristi u proceduri TREE programskog sistema SAS.

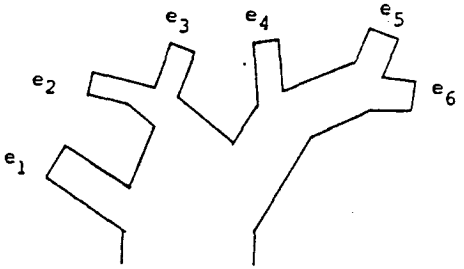


Slika br.6

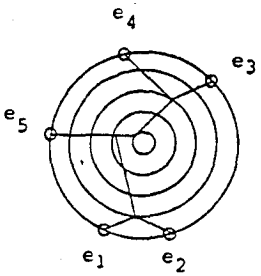


Slika br.7

Razvijen je još niz prikaza drvolikog tipa (sklika 8 i 9) ali kako nisu inkorporirani u neki od popularnih paketa programa redje se susreću u literaturi.



Slika br.8



Slika br.9

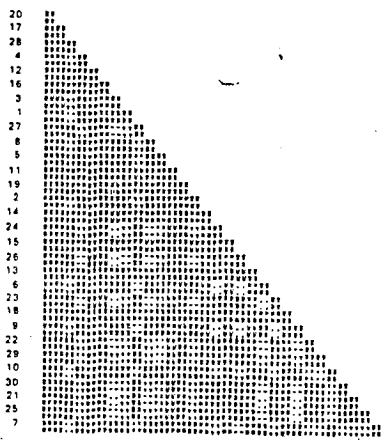
Upotreba ovakvih dijagrama (drvolikih) veoma odgovara prirodi hijerarhijskog klasifikovanja, a vidljiv je uticaj bioloških nauka odakle je i započeo razvoj klaster anlize.

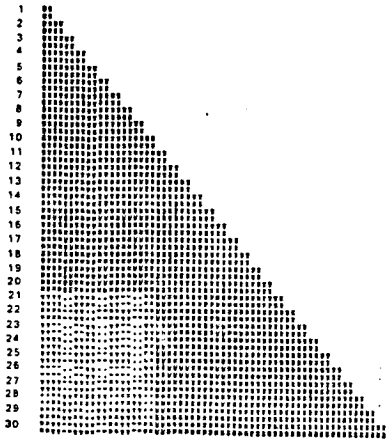
5.Grafički metodi u korelacionoj i regresionoj analizi

Polazeci od skupa jedinic E na kojima se realizuje vektor slučajnih obelezja X medjusobne odnose obeležja prikazuje korelaciona (ili kovariaciona matrica). Sto je dimenzija vektora X viša to je teže snalaženje u korelacionoj matrici. Programski paket BMDP nudi tehniku SHADE za grafički prikaz korelacione matrice (ideja potiče od Linga).

Ideja je sasvim jednostavna - da se nivo koreliranosti iskazuje stepenom zatamnjenja.Ista ideja se sasvim jednostavno prenosi i na bilo koju simetričnu matricu npr. matricu distanci.

Na slikama 10 i 11 je data ista matrica generisanih podataka jedanput u slučajnom rasporedu, a jedanput sortiranu tako da se prepoznaju homogene grupacije. Ako bi se radilo o korelaciji 30 obelezja onda bi ovakav grafikon ukazivao na postojanje vrlo malog broja faktora (2 ili 3).



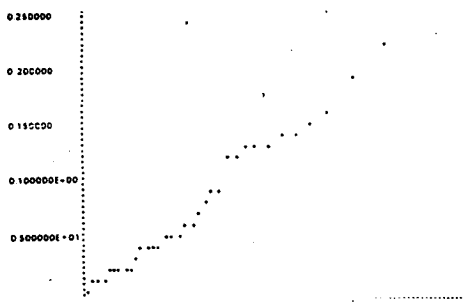


Slika br.11

Konačno za upotrebu regresionih tehnika je poznato da treba poznavati raspodelu podataka na kojima se radi. Ako se realizuje slučajana velična po normalnoj raspodeli onda je dozvoljiv rad sa modelima koji polaze od te predpostavke. Radi provere predpostavke o raspodeli populaciji na jednostavan i brz način razvijen je metod ctreža verovatnoca (probability plots). I ovde se radi o jednostavnoj ideji da statistike poretka (ili kvantili) iz uzorka mora da odgovaraju teorijskim vrednostima ukoliko je ispravna predpostavka o raspodeli. Ucrtavanjem parova uzoračkih i teorijskih vrednosti hipotezu potvrdjuje prava linija koja prolazi kroz koordinatni početak. Veće odstupanje od prave linije znači i veće odstupanje od predpostavljene raspodele.

Najčešće su provere predpostavke o normalnosti populacije mada ni druge raspodele nisu retke (slika 12) .

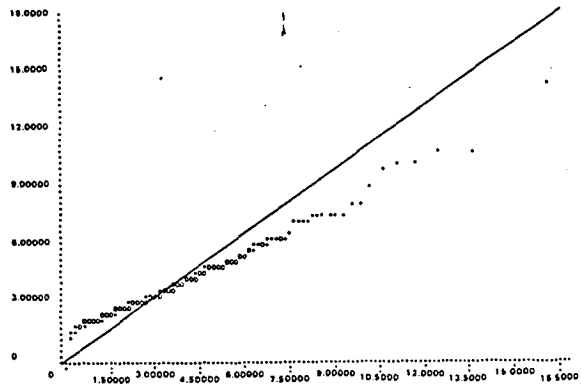
Mada su radovi o robusnosti donekle smanjili intetres za ovu tehniku ipak postoji veliki broj paketa programa koji daju mogućnost i izračunavanja i crtanja statistika poretka.



Kada je reč o multivarijacionoj analizi još je izrazeniji zahtev za normalnoscu osnovnoh skupa da bi se primenjena tehnika opravdala. Kako generalizovano odstojanje

$$d_1 = (X_1^{(1)} - \bar{X})' S^{-1} (X^{(1)} - \bar{X})$$

pod predpostavkom da su podaci multivarijaciono normalni ima hi-kvadrat raspodelu sa p stepeni slobode moguća je primena ideja o 'probability' grafikonima. Na slici 13 je prikazano da se može ostati pri predpostavci o multivarijacionoj normalnosti vektora od četiri varijabli koji se realizuje na 180 psihijatrijskih pacijenata.



Slika br.13

Napomena 1: U radu je napravljen pokušaj sistematizovanja obimne gradje koja, očigledno, zahteva seriozniji pristup.

Napomena 2:Jedan broj slika-grafikona je preuzet iz Everitt (4)

L i t e r a t u r a

1. *Andrews, D.F (1972) - PLOTS OF HIGH DIMENSIONAL DATA - Biometrics 28*
2. *Bogosavljevic,S. (1984) - ASPRIORNE METODE KLASIFIKACIJE EKONOMSKIH POJAVA - nepublikovana doktorska disertacija*
3. *Chernoff, H (1973) - USING FACES TO REPRESENT A POINTS IN K-DIMENSIONAL SPACE GRAPHICALLY - JASA 68*
4. *Everitt, B.S (1977) - GRAPHICAL TEHNILUES FOR MULTIVARIATE DATA - Heinemann Educational Books*
5. *Ivanovic, B. (1977) - TEORIJA KLASIFIKACIJE - Institut za ekonomiku industrije*
6. *Inselberg, A (1981) - N-DIMENSIONAL GRAPHICES - IBM Los Angeles Scienetific Center*
7. *Marinkovic D, Salcberger, B. (1982) - MOGUCNOSTI I KORISCENJE PROGRAMA "LICA" - IRD VII/19 - SZS*
8. *C.R. Rao (1964) - THE USE AND INTERPRETATION OF PRINCIPAL COMPONENTS - Sankhya 26*
9. *Wainer H, (1984) - HOW TO DISPLAY DATA BADLY - The American Statistician 38*