Anuška Ferligoj

METODE ZA GRAFIČNO PREDSTAVITEV MULTIVARIATNIH PODATKOV

GRAPHICAL METHODS FOR REPRESENTING MULTIVARIATE DATA:
Graphical methods provide a powerful tool for presenting and analysing multivariate data. In the paper several graphical methods for data exploration are presented and compared. Stars, Andrews-plots and Chernoff faces are discussed and used to present European countries according to selected socio-economic indicators.

1. UVOD

Večja zmogljivost računalnikov in dostopnost ustreznih paketov za statistično analizo podatkov uporabnike pogosto zanese, tako da izpustijo duhamorno osnovno obdelavo svojih podatkov in jih takoj "napadejo" z zahtevnejšimi metodami analize podatkov. To je prav gotovo zgrešena pot. Že izbor metode mora biti odvisen od rezultatov osnovnega pregleda podatkov. Tako na primer tujki (outliers), ki jih nismo razkrili s predhodno analizo, lahko popolnoma razmažejo rezultate analize, ki sicer predpostavlja določene porazdelitve spremenljivk in ni dovolj neobčutljiva (robustna) za neizpolnjevanje predpostavk.

Osnovna analiza podatkov, ki jo tvorijo iz pregled frekvenčnih porazdelitev, osnovne statistike posameznih spremenljivk in parov spremenljivk (več o tem npr. Blejec 1973) ter različne grafične predstavitve multivariatnih podatkov, je zelo koristna za postavitev začetnih delovnih domnev in za ustrezen izbor zahtevnejših metod pregledovalne in potrjevalne analize. V nadaljevanju si bomo ogledali nekaj zanimivih in preprostih načinov grafične predstavitve multivariatnih podatkov.

2. GRAFIČNE PREDSTAVITVE PODATKOV

Znana je trditev, da je metoda analize podatkov uspešna, če je mogoče rezultate nazorno grafično predstaviti. Statistična grafika je vizualna predstavitev merjenih količin s kombinacijo točk, koordinatnih sistemov, simbolov, številk, barv itd. S statistično grafiko se z različnih vidikov raziskovalci bavijo že več kot dve stoletji. Računalniška tehnologija pa je odprla nove možnosti grafičnim predstavitvam podatkov. Danes je statistična

grafika pomemben sestavni del analize podatkov. Uporablja se za pregled podatkov, za razkrivanje strukture podatkov, za predstavitev podatkov in njihovo komuniciranje itd. Le redki računalniški programi za statistično analizo podatkov ne vsebujejo tudi preproste grafične ponazoritve ene spremenljivke s histogrami, poligoni, strukturnimi stolpci ali krogi, ter parov spremenljivk s "scattergrami" različnih oblik, barv, z različnimi osenčenji itd. Statistična paketa SAS in STATGRAPHICS sta najpogosteje omenjena kot zelo primerna za grafične ponazoritve podatkov.

Zelo veliko resnega raziskovalnega dela je bilo in se še vedno vlaga v študije, kako kar se da nazorno grafično prikazati rezultate določene analize podatkov. Iščejo se primerni grafični pristopi in ustrezne barvne kombinacije ali osenčenja za predstavitev podatkov in rezultatov statističnih analiz podatkov (npr. Bowman 1968; Ehrenberg 1977; Cleveland in McGill 1984; Becker in sodelavci 1988). Tako na primer dilema, kaj je boljše za predstavitev struktur - stolpci ali krogi, še vedno ni razrešena. V letu 1987 je bil objavljen članek Beckerja in sodelavcev, kjer so z empiričnimi raziskavami potrdili, da ljudje natančnejše razpoznajo strukture, če so predstavljene s stolpci. Po drugi strani pa Spence in Lewandowsky (1987) dokazujeta, da s krogi ljudje pravilneje zaznavajo posredovane strukture.

O multivariatnih podatkih govorimo tedaj, ko je posamezna enota opisana z več spremenljivkami. Znanih je več pristopov, kako grafično predstaviti multivariatne podatke. Cleroux, Lepage in Ranger (1984) so v svojem delu zapisali deset osnovnih načel, ki bi jim naj zadoščale metode grafične predstavitve multivariatnih podatkov. Ti so:

- metoda naj posreduje informacijo preprosto in hitro;
- pomagati mora razumeti informacijo;
- poudariti mora važnejše informacije;
- mnemoničnost metode;
- predstavitev mora biti preprosta, zgoščena in privlačna;
- predstavitev mora biti jasna, točna in nezmaličena;
- predstavitev mora biti lahko in hitro razumljiva;
- temelji naj na običajnih oblikah;
- omogočati mora hkratno predstavitev več razsežnosti;
- omogočati mora primerjave in razvrščanja.

V nadaljevanju bomo predstavili tri najpogosteje uporabljene grafične predstavitve multivariatnih podatkov: zvezde, Andrewsove krivulje in obraze. Ti grafični postopki zadoščajo večini zgoraj omenjenih zahtev.

2.1 Zvezde

Od omenjenih grafičnih predstavitev multivariatnih podatkov je v standardne statistične pakete (npr. STATGRAPHICS) največkrat vključena metoda zvezd. S tem načinom grafične predstavitve vsaki enoti priredimo zvezdo, ki jo konstruiramo na naslednji način: 360 stopinj razdelimo na toliko enakih delov, kolikor je spremenljivk. Vsak del ločimo z žarkom, na katerega nanašamo ustrezno normalizirano vrednost pripadajoče spremenljivke - vrednosti iste spremenljivke vedno na žarek v isti smeri. Ponavadi spremenljivke normaliziramo tako, da je povprečje spremenljivke na obodu zvezdi pririsanega kroga, konkretne vrednosti, ki jih nanašamo na žarke, pa normaliziramo glede na dolžino od središča zvezde do točke na krožnici. Sosednje točke na žarkih povežemo z ravnimi črtami. Tako dobljene zvezde je mogoče preprosto primerjati: zvezde s podobno obliko pripadajo enotam, ki so si glede na merjene spremenljivke podobne. Na ta način je zelo lahko odkriti tujke. Uporaba zvezd pa je seveda omejena, kajti pri velikem številu spremenljivk (in s tem veliko žarkih) postane zvezda nepregledna. Zvezde tudi niso primerne za predstavitev večjega števila enot.

2.2 Andrewsovi grafi

Andrews (1972) je predlagal za grafično predstavitev enot z m izmerjenimi spremenljivkami, to je vektorjev $X' = (x_1, x_2, ..., x_m)$, krivuljo, določeno s Fourierovo vrsto

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \ldots$$

Ta funkcija je nato narisana za vsako enoto na intervalu $-\pi \leq t \leq \pi$. Podobno kot pri zvezdah imajo glede na merjene spremenljivke podobne enote podobne Andrewsove krivulje. Te namreč zadoščajo nekaj lastnostim, ki so zelo primerne za grafični pregled podatkov. Omenimo nekaj najpomembnejših lastnosti:

- Funkcija $f_x(.)$ ohranja povprečja, kar pomeni naslednje: če je $\overline{X}$ povprečje n enot za spremenljivko X, potem je

$$f_{\overline{x}}(t) = \frac{1}{n} \sum f_{x_i}(t)$$

- $f_x(.)$ ohranja razdalje. Namreč: če je L_2 razdalja oblike

$$f_x(t) - f_y(t) = \int_{-\pi}^{\pi} (f_x(t) - f_y(t))^2 dt$$

je sorazmerna evklidski razdalji med enotama X in Y . Posledica te lastnosti je, da sta krivulji za enoti, ki sta v evklidskem smislu blizu, podobni. Zato so tudi Andrewsovi grafi primerni za določitev skupin podobnih enot in za odkrivanje tujkov.

- Za izbrani t_0 je $f_x(t_0)$ sorazmeren dolžini projekcije vektorja $(x_1, x_2, \ldots, x_m)$ na vektor

$$f_x(t_0) = (\frac{1}{\sqrt{2}}, \sin t_0, \cos t_0, \sin 2t_0, \cos 2t_0, \ldots)$$

Projekcija na ta enorazsežni prostor lahko razkrije podobne skupine enot, tujke itd.

- $f_x(.)$ ohranja variance. Če so komponente enote X nekorelirane s skupno varianco σ^2, potem je $var_x(t)$ enaka $\sigma^2/2$, če je m sodo število, in leži med $\sigma^2(m-1)/2$ in $\sigma^2(m+1)/2$, če je m liho število.

Pri uporabi Andrewsovih krivulj nastopijo težave, če rišemo krivulje za večje število enot na isti sliki, ker nastane zmešnjava. Druga težava je v tem, da upoštevane spremenljivke v tej predstavitvi nimajo enake teže. Spremenljivke pri členih na začetku Fourierove vrste imajo večjo težo kot pri kasnejših. Zato je priporočljivo, da spremenljivke rangiramo po pomembnosti in v tem vrstnem redu vključimo v vrsto. Problem je tudi v tem, da imajo spremenljivke s povprečno večjimi vrednostmi večjo težo pri obliki krivulje. Zato ponavadi pred določitvijo krivulj spremenljivke standardiziramo.

2.3 Obrazi

Morda najatraktivnejši, a tudi najbolj kritiziran način grafične predstavitve multivariatnih podatkov je predlagal Chernoff (1973). V tem primeru gre za predstavitev vrednosti posameznih spremenljivk na izbranih sestavinah obraza. Tako so na primer velikost oči, nosu, ušes itd., naklon obrvi ali ust sorazmerni vrednostim izbranih spremenljivk. Tudi v tem primeru, kakor pri zvezdah, je mogoče razkriti podobnost enot glede na merjene spremenljivke s podobnostjo pripadajočih obrazov. Ta način grafične predstavitve podatkov je bil zelo kritiziran, češ da subjektivno določimo pomembnost posamezne spremenljivke s tem, da ji priredimo bolj ali manj značilen del obraza. Vendar velja ta kritika tudi za večino drugih metod grafične predstavitve podatkov (npr. Andrewsove krivulje). Po drugi strani pa je bilo več avtorjev navdušenih nad to metodo in nekaj jih je metodo precej izpopolnilo in izdelalo učinkovite računalniške programe, tako da je mogoče z obrazi predstaviti že več kot 20 spremenljivk (Chernoff jih je zmogel predstaviti 18). Flury in Riedwyl (1981) sta z uvedbo nesimetričnih obrazov to število celo podvojila.

Hamner, Turner in Young (1987) so primerjali več znanih metod grafične predstavitve multivariatnih podatkov (med njimi tudi simetrične in nesimetrične obraze, risane s tiskalnikom in risalnikom) s faktorskim načrtom poskusov. Ugotovili so, da je metoda simetričnih Chernoffovih obrazov z zveznimi črtami izrazito najboljša pri razkrivanju skupin podobnih enot. Zato se je tudi Batagelj (1988) odločil, da je v program CLUSE vgradil prav take obraze.

3. PREDSTAVITEV EVROPSKIH DRŽAV

Opisane metode za grafično ponazoritev multivariatnih podatkov, ki so vgrajene v program CLUSE (Batagelj 1988) za računalnik ATARI ST, smo uporabili za predstavitev evropskih držav glede na izbrane kazalce družbeno-ekonomske razvitosti. V tem poglavju so prikazani rezultati za dve skupini evropskih držav:

1. Dansko, Finsko, Island, Norveško in Švedsko
2. Grčijo, Irsko, Porugalsko, Španijo in Jugoslavijo

Države so predstavljene glede na naslednje spremenljivke za leto 1978 (Vir: The Hammond Almanac, 1980):

NARDOH	narodni dohodek na prebivalca
MORDOJ	mortaliteta dojenčkov
NAT	nataliteta
URBAN	odstotek mestnega prebivalstva
AVTO	število osebnih avtomobilov na 100 prebivalcev
TELEFON	število telefonskih priključkov na 100 prebivalcev
ZDRAVNIK	število prebivalcev na zdravnika
VŠOLA	število študentov na visokih šolah na 100 prebivalcev
ČASOPIS	število izvodov dnevnih časopisov na 100 prebivalcev
IND	narodni dohodek industrije glede na celotni narodni dohodek

Na sliki 4 so prikazane zvezde za skandinavsko skupino in južnoevropsko skupino držav tako, da je prva spremenljivka (NARDOH) predstavljena z žarkom, usmerjenim v desno, ostale pa po vrsti v smeri urinega kazalca. Na sliki 5 pa so za ti dve skupini držav predstavljeni Andrewsovi grafi. Na sliki 6 so iste enote prikazane še z obrazi. Posamezne spremenljivke so predstavljene z naslednjimi elementi obraza:

NARDOH	širina obraza
MORDOJ	širina nosu
NAT	dolžina nosu
URBAN	dolžina ust
AVTO	ukrivljenost ust
TELEFON	velikost ušes
ZDRAVNIK	gostota las
VŠOLA	velikost oči
ČASOPISI	naklon obrvi
IND	dolžina las

Spremenljivke so v vseh treh primerih standardizirane. Tudi iz te uporabe grafičnih metod je razvidno, da je ponazoritev multivariatnih podatkov z obrazi najučinkovitejša.

LITERATURA

1. Batagelj V. (1988): Priročnik za program CLUSE. Ljubljana.
2. Andrews D.F. (1972): Plots of high dimensional data. Biometrics, 28, 125-136.
3. Becker R.A., Cleveland W.S., Wilks A.R. (1987): Dynamic graphics for data analysis. Statistical Science, 2, 355-395.
4. Becker R.A., Cleveland W.S., Wilks A.R. (1988): An interacive system for multivariate data display (neobjavljen članek)
5. Blejec M. (1973): Statistične metode za ekonomiste, Ekonomska fakulteta, Ljubljana.
6. Bogosavljevič S. (1986): Grafičko prikazivanje multivarijacionih podataka i rezultata u teoriji klasifikacije. Institut za statistiku SZS, Beograd, 14 str.
7. Bowman W.J. (1968): Graphic Communication. Wiley, New York.
8. Chambers J.M., Kleiner B. (1982): Graphical techniques for multivariate data and for clustering. V: P.R. Krishnaiah in L.N. Kanal (Eds.): Handbook of Statistics, Vol. 2, North-Holland, Amsterdam, 209-244.
9. Chernoff H. (1973): Using faces to represent points in k-dimensional space graphically. JASA, 68, 361-368.
10. Cleroux R., Lepage Y., Ranger N. (1984): La representation graphique de donnees multivariees. Rapports de recherches du departement de mathematiques et de statistique, No 84-2, Montreal.
11. Cleveland W.S., McGill R. (1984): Graphical perception: Theory, experimentation, and application to the development of graphical methods. JASA, 79, 531-554.
12. Dillon W.R., Goldstein M. (1984): Multivariate Analysis: Methods and Applications. Wiley, New York.
13. Ehrenberg A.S.C. (1977): Rudiments of numeracy. J.R. Statist. Soc. A, 140, 277-297.
14. Everitt B.S., Dunn G. (1983): Advanced Methods of Data Exploration and Modelling. Heinemann, London.
15. Flury B., Riedwyl H. (1981): Graphical representation of multivariate data by means of asymmetric faces. JASA, 76, 757-765.

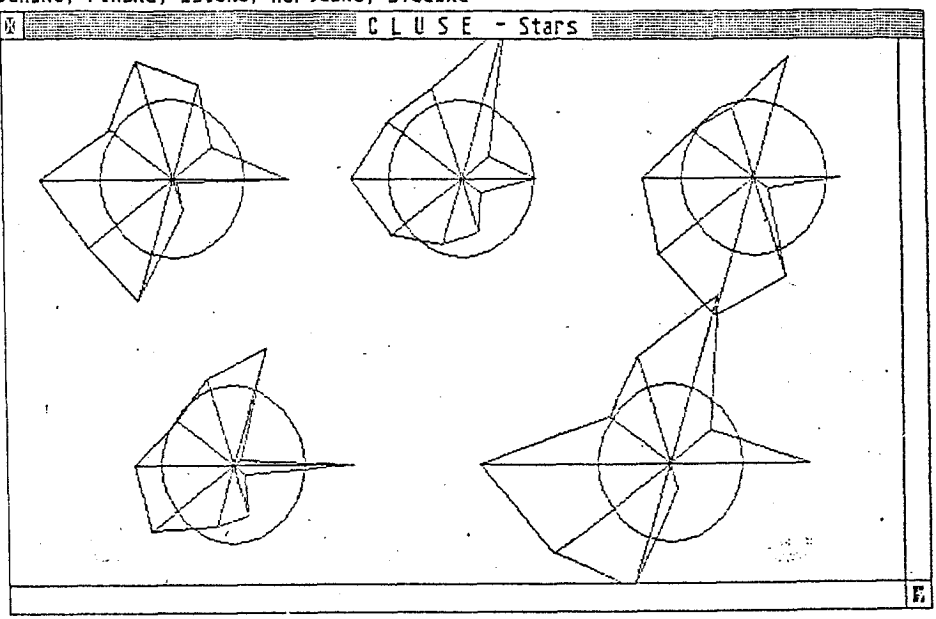
16. Hamner C.G., Turner D.W., Young D.M. (1987): Comparisons of several graphical methods for representing multivariate data. Comput. Math. Applic., 13, 647-655.

17. Lewandowsky S., Spence I. (1987): Discriminating strata in scatterplots. Paper presented at the European Meeting of the Psychometric Society, Enschede.

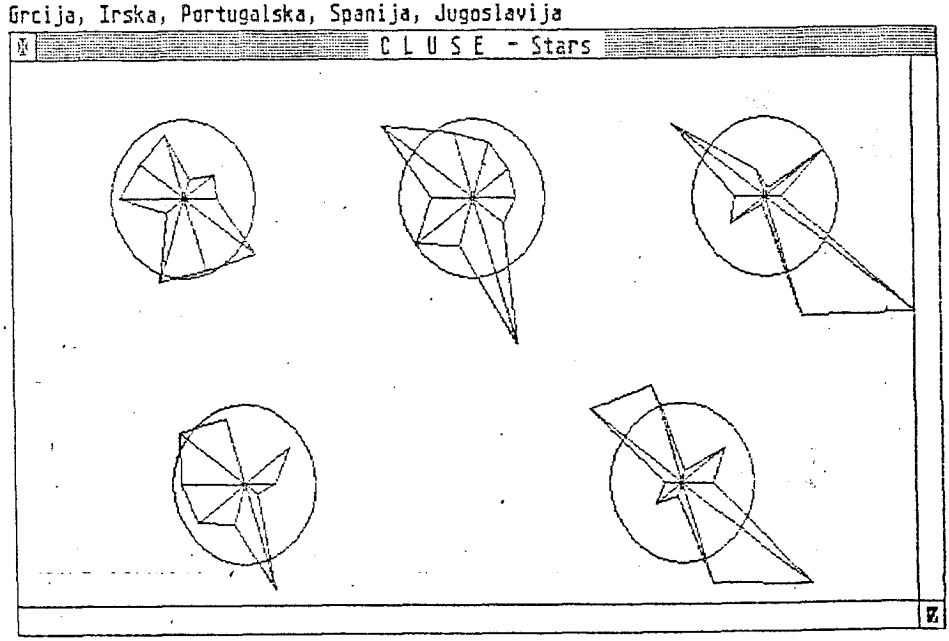
18. Nagel M., Dobberkau H.J. (1988): Graphical methods of exploratory data analysis: An overview. V: H.H. Bock (Ed.): Classification and Related Methods of Data Analysis. North-Holland, Amsterdam, 633-640.

19. Spence I., Lewandowsky S. (1987): Displaying proportions and percentages. Paper presented at the European Meeting of the Psychometric Society, Enschede.

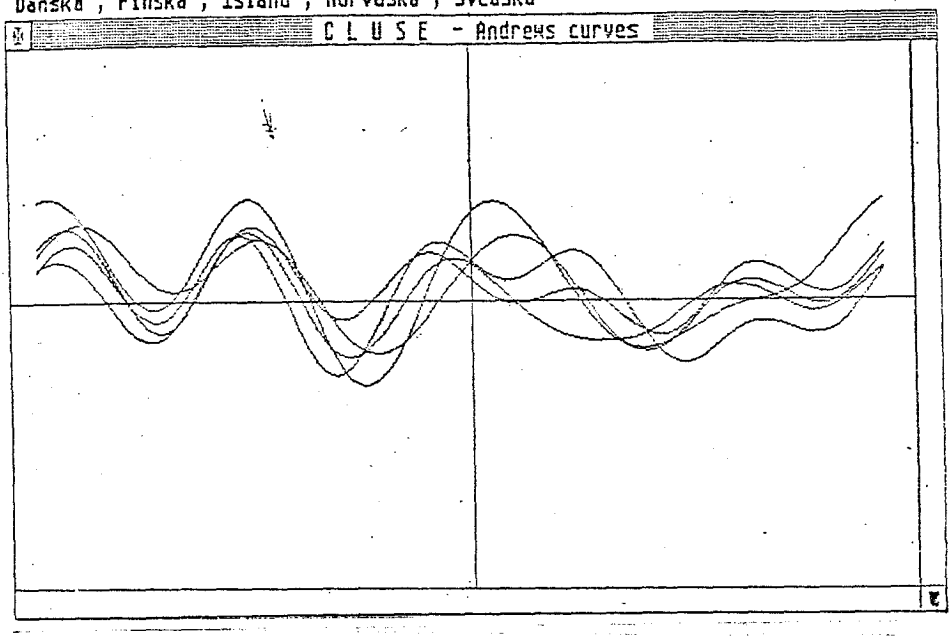
Danska, Finska, Island, Norveska, Svedska



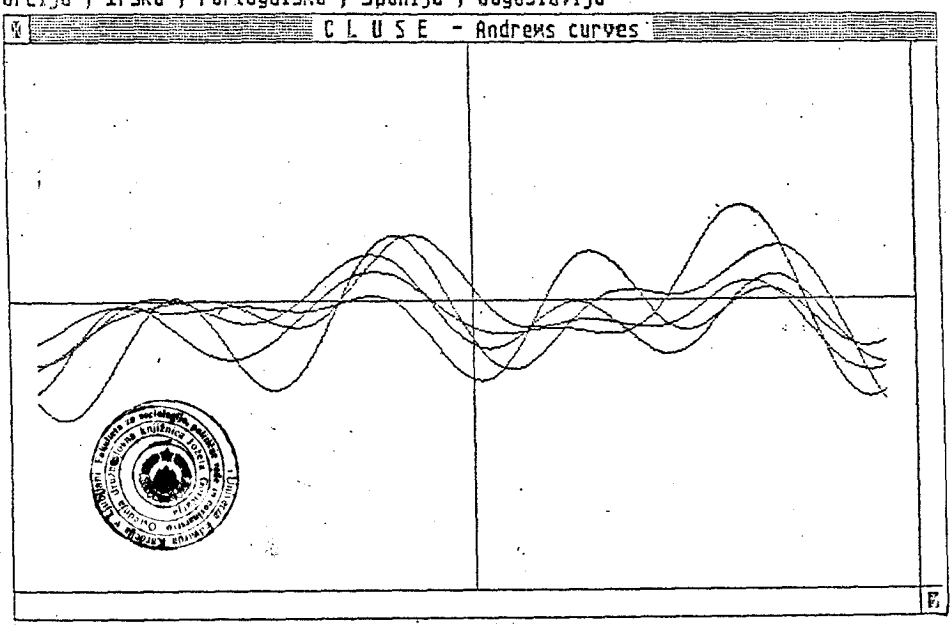
Grcija, Irska, Portugalska, Spanija, Jugoslavija



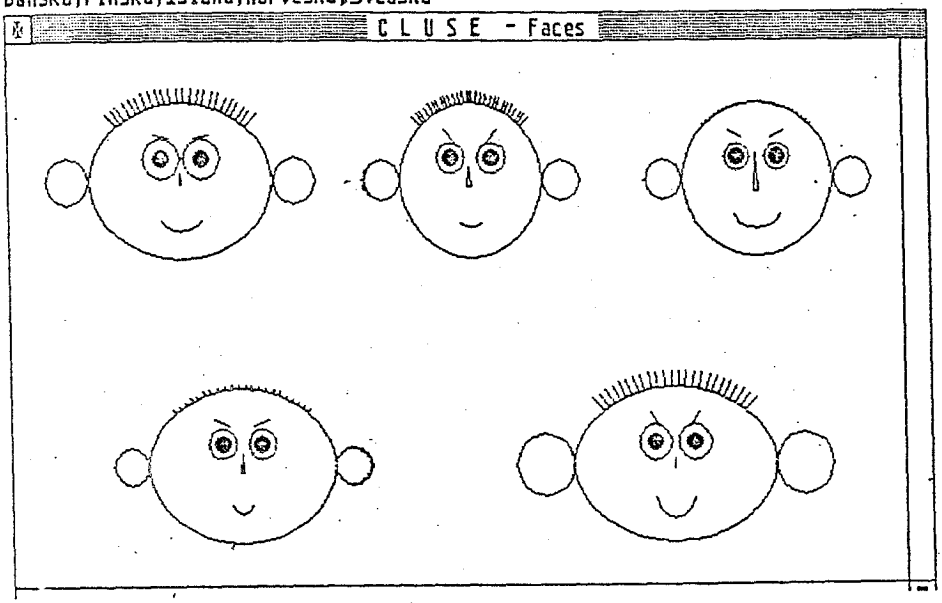
Danska , Finska , Island , Horvaska , Svedska



Grcija , Irska , Portugalska , Spanija , Jugoslavija



Danska,Finska,Island,Horveska,Svedska



Grcija, Irska, Portugalska, Spanija, Jugoslavija