

Graficul matrice – tehnica importanta de elaborare a analizelor statistice multivariate prin imagini

Prof.dr. Vergil VOINEAGU, conf.dr. Emilia TITAN,
lect. Mariana VOINEAGU, ec. Dan TITAN

*În orice domeniu, cunoasterea mai rapida si mai directa a realitatii, surprinderea manifestarii fenomenelor, a succesiunii logice a relatiilor obiective existente între fenomene, explicarea teoriilor si metodologiilor etc. necesita sinteza grafica a informatiilor cantitative si calitative. Aceasta se explica prin faptul ca graficele au calitatea de a traduce prin imagini, într-o forma simpla, sugestiva si atragatoare informatiile referitoare la trasaturile esentiale ale fenomenelor, în conditii determinate de timp si spatiu. În acest scop sinteza grafica, în termenii semiologiei lui Bertin (1967) – acest subiect este pe larg si accesibil prezentat de Bertin în: *Semiologie grafica*, Gauthier – Villars, Paris 1967 – comporta un ansamblu complex de tehnici pentru elucidare, interpretare si analiza a datelor – într-un cuvânt pentru cunoastere – prin in-termediul punctelor, segmentelor de dreapta, ariilor si a figurilor geometrice, a hartilor, simbolurilor si a altor mijloace specifice.*

Cuvinte cheie: sinteza grafica; grafic matrice; bloc matricial; diagrame de împrastiere; diagrame trasate prin linii sau coloane; analiza statistica prin imagini.

Complexitatea unui fenomen poate fi judecata dupa multe criterii: cauzalitate diversificata, efecte multiple, interdependente variate – ca directie si intensitate – cu alte fenomene etc. În conditiile unui pachet specific de programe informatice, sinteza prin imagini a acestei complexitati, sau traducerea vizuala a valentelor diverselor metode utilizate în acelasi scop¹, este posibila prin **graficul matrice**.

Spre exemplificare, consideram distributia bidimensionala $\{x_i, y_j, n_{ij}\}$ (cu $i = \overline{1, m}$; $j = \overline{1, p}$) care are în componenta sa: doua distributii unidimensionale marginale

$\{x_i, n_i\}$ (cu $i = \overline{1, m}$ si $n_i = \sum_{j=1}^p n_{ij}$) si

$\{y_j, n_j\}$ (cu $j = \overline{1, p}$ si $n_j = \sum_{i=1}^m n_{ij}$); p

distributii unidimensionale $\{x_i, n_{ij}\}$ ($i = \overline{1, m}$) conditionate de fiecare y_j ($j = \overline{1, p}$); m distributii unidimensionale $\{y_j, n_{ij}\}$ ($j = \overline{1, p}$) conditionate de fiecare x_i ($i = \overline{1, m}$). Traducerea vizuala simultana a tuturor acestor repartitii cu densitatile, structurile si formele lor specifice nu se poate realiza decât printr-un grafic matrice sau matriceal conceput.

Necesitatea elaborarii acestor grafice matrice este si mai evidenta daca dorim **sinteza vizuala a manifestarii simultane a unor variabile multiple (de natura diferita) în aceeași populație statistică**.

Graficul matrice poate avea un cadru patratic sau dreptunghiular, cu dimensiunea egala cu numarul repartitiilor unidimensionale conditionate sau neconditionate care compun ansamblul, repartitia multi-dimensională.

Elementele matricii grafice sunt **imagini (diagrame de împrastiere, trasate prin linii sau coloane) specifice fiecarui tip de repartitie componenta**.

¹ Exemplificam fara un criteriu anume: Wagner, P.; Stefanescu, D.: *Compararea internationala a PIB*, Ed. Economica, Bucuresti, 1999, p.97-101; Frisch, H.: *Teorii ale inflatiei* Ed. Sedora, Timisoara, 1997, p.97-101; Paun, C.; Paun, L.: *Riscul de tara*, Ed. Economica, Bucuresti, 1999, p.81, 90, 112, 203; Lipsey, R.G., Chrystal, A.C.: *Economia pozitiva*, Ed. Economica, Bucuresti, 1999 etc.

În functie de scopul analizei, graficul matrice, dupa cum usor se observa, **nu exclude ci completeaza graficele secventiale (clasice), independente.**

Elaborarea graficului matrice nu poate fi conceputa decât prin mijloace informatice si trebuie sa fie **precedata de parcurgerea mai multor faze (etape)**, din care evidentiem:

- fixarea **obiectului de studiu** – precizia acestuia si calitatea (veridicitatea) ipotezelor formulate sunt decisive pentru urmatoarele etape;
- **selectarea variabilelor** semnificative si elaborarea datelor;
- **înregistrarea** datelor;
- **organizarea si administrarea datelor**;
- **analiza si explorarea datelor.**

Valentele suplimentare ale graficelor matrice sunt în favoarea globalizarii cunoasterii deoarece:

- **armonizeaza** mult mai bine inductia si deductia, raportul particular-general, într-un demers stiintific;
- **compararea secventiala si multipla** are un rol mult mai important în cunoastere din urmatoarele considerente:
 - **interpretarea vizuala a raportului de marime** dintre mai multi indicatori este mult mai obiectiva;
 - se observa si se interpreteaza mai bine **structurile si mutatiile structurale** în dinamica sau în profil teritorial;
 - faciliteaza **interpretarea corelata a densitatilor** diferitelor repartitii de frecvente, comparativ cu cele ale repartitiilor teoretice, cu sau fara restrictii (conditii) suplimentare;
 - faciliteaza interpretarea formelor de realizare a **interdependentelor** dintre doua sau mai multe variabile;
 - faciliteaza **sesizarea similaritatilor si disimilaritatilor** dintre densitatile repartitiilor de frecvente, dintre tendintele de dezvoltare a fenomenelor studiate în dinamica pentru etapa data;

- **completeaza informatiile** obtinute din sinteza numerica (prin calculul unor parametrii);

- permite **identificarea si analiza componentelor principale** si a corespondentelor dintre ele;

- faciliteaza nu numai alegerea si utilizarea metodelor de analiza statistica secventiala dar mai ales a **metodelor de analiza multivariata, factoriala si tipologica**;

- faciliteaza alegerea si utilizarea **metodelor statistice de analiza a variabilelor calitative în interdependenta cu alte variabile de natura diferita**;

- permite **analiza simetriilor** si nu numai fata de diagonala principala a graficului matrice.

Pentru a nu cadea pe panta manipulării sau a erorilor nevoite, sustinem ca, calitatea sintezelor prin graficele matrice, masura în care acestea îndeplinesc functiile amintite si satisfac nevoia cunoasterii realitatii depinde de **gradul respectării unor restrictii constructive, de elementele graficului matrice.** Aceste restrictii sunt subordonate obiectivelor de traducere în imagine a realitatii, iar asupra lor vom face referiri în cele ce urmeaza. Înainte de acestea subliniem, pentru a nu crea confuzii, ca **graficele matrice se deosebesc prin continut de graficele multiple** trasate într-un cadru patratic sau dreptunghic (vezi figurile 1-6, comparativ cu figura 7). În acest sens precizam ca graficul matrice spre deosebire de celalalt, prezinta o **relatie de ordine predefinita între elementele structurale** precum si una de simetrie a elementelor fata de diagonala principala si în consecinta, evita prezentarea unor dependente structurale lipsite de continut, nerealiste.

Sa presupunem ca într-o populatie statistica P formata din n unitati sunt înregistrate date despre variabilele V_i ($i = 1, p$). Fiecare din aceste variabile este definita printr-un spatiu (O) al observatiilor (valorilor înregistrate), printr-o structura (S) specifica acestui spatiu si printr-o aplicatie de forma:

$V : P \rightarrow O$, înzestrată cu structura S .
 Aceasta înseamnă ca pentru ansamblul celor p variabile $P : \xrightarrow{v_i} O_i$ înzestrată cu structura S_i se poate defini o variabilă vectorială (V) ce aplică P în produsul cartezian $O_1 \times O_2 \times \dots \times O_p$ adică:

$V \equiv (V_1, \dots, V_p) : P \rightarrow O = O_1 \times O_2 \times \dots \times O_p$ **înzestrată cu structura S definită pe baza $S_1, S_2, \dots, S_p$** . Prin urmare, variabila vectorială V asociază fiecărei unități de observare $j \in P$ p -uplul $(V_{1(j)}, V_{2(j)}, \dots, V_{p(j)})$ din O .

În funcție de cardinalul multimii O și de structura S^2 specifică în populația statistică P se disting, după natura lor, următoarele tipuri de variabile:

- **variabile calitative nominale:**
 $V : P \rightarrow O \equiv$ mulțime finită de coduri numerice fără să fie înzestrată cu o anumită structură, dacă exceptăm identitatea și nonidentitatea ($=$ sau $\neq$);

- **variabile calitative ordinale:**
 $V : P \rightarrow O (\equiv N) \equiv$ mulțime finită (sau cel mult numărabilă) înzestrată numai cu o structură de ordine ($\leq$);

- **variabile cantitative ordinale:**
 $V : P \rightarrow O (\equiv R) \equiv$ mulțime finită de numere reale (sau cel mult numărabilă) înzestrată doar cu o structură de ordine ($\leq$) și care nu permite operații algebrice;

- **variabile cantitative (masurabile cardinale):** $V : P \rightarrow O (\equiv R) \equiv$ mulțime continuă înzestrată cu o structură de corp ordonat ($\leq, +, \times$).

² Structura asociată multimii O nu trebuie privită „mecanic” (ca ceva fixat aprioric), deoarece ea este întotdeauna **indusă, în funcție de scopul cercetării, de semantica subiacentă a „parametrului” analizat**. Altfel spus, în funcție de semnificația atribuită unui „parametru”, **acestuia îi pot corespunde variabile cu structuri diferite**. De exemplu, parametrului vechimea în muncă i se poate asocia o variabilă cantitativă (ani de muncă împliniți), o variabilă ordinală (clase de ani de vechime în muncă) dar și o variabilă calitativ nominală (prin regruparea claselor de vechime în muncă, într-o manieră neordinală, se pot forma alte grupe nominale, de exemplu „tineri” și „vârstnici”).

Această formalizare a variabilelor vectoriale cu structurile lor subiacente creează necesitatea și oferă premisele construirii unor **grafice matrice capabile să vizualizeze în același cadru structura S (compusă din combinarea $S_1, S_2, \dots, S_p$) a multimii $O_1 \times O_2 \times \dots \times O_p$** . Mai mult, se evidențiază faptul că **analiza multivariată nu exclude ci, din potrivă, implică și se bazează pe analiza bivariată**. În același mod graficul matrice (conceput ca modalitate de traducere vizuală a relațiilor dintre variabilele V_i ($i = 1, p$) și structurile subiacente acestora) se bazează și are **drept componente graficele utilizate în analiza multivariată**, construite după principiul noncontradicției.

Graficul matrice prezintă într-un **cadru unitar imagini simultane** ale graficelor construite în două dimensiuni (în sistemul de coordonate rectangulare) pentru toate combinațiile variabilelor – luate două câte două – definite în același spațiu sau în două subspații de variabile cu structurile subiacente. Aceasta înseamnă că graficul matrice poate fi construit într-un **cadru patratic sau dreptunghiular**.

Cadrul patratic, de dimensiune $p \times p$, al unui grafic matrice se obține dacă avem la dispoziție variabilele $V_1, V_2, \dots, V_p$ din spațiul observațiilor $O_1 \times O_2 \times \dots \times O_p$ înzestrat cu structura S rezultată din $S_1, S_2, \dots, S_p$. În acest caz, graficul matrice are mai multe componente (blocuri de imagini). Fiecare componentă din matricea de imagini, este un grafic construit adecvat și specific pentru toate perechile p variabile. Pe diagonala principală a graficului matrice sunt dispuse **toate cele p histograme** care vizualizează distribuțiile și densitățile frecvențelor fiecărei variabile V_i ($i = 1, p$).

În cazul în care din $(V_1, V_2, \dots, V_p)$ se iau în considerare variabilele $(V_1, V_2, \dots, V_r)$ și $(V_{r+1}, V_{r+2}, \dots, V_p)$ (cu $r < p$) cu spațiul $O_1 \times O_2 \times \dots \times O_p$ separat în două subspații $O_1 \times O_2 \times \dots \times O_r$ și $O_{r+1} \times O_{r+2} \times \dots \times O_p$,

cadru construit al matricei grafice va avea forma **dreptunghiulara** de cardinal $[r+1] \times [p-r+1]$. În acest cadru pe axa „absciselor” vor fi reprezentate (individual si succesiv) variabilele $V_{r+1}, V_{r+2}, \dots, V_p$, cu structurile aferente, iar pe axa „ordonatelor” vor fi plasate, în aceeași manieră, variabilele $V_1, V_2, \dots, V_r$ cu structurile aferente. Aceasta înseamnă ca în prima coloana si linie a cadrului dreptunghiular al graficului matrice vor fi plasate imagini ale repartitiilor rezultate pentru fiecare din cele $p-r+(p-r)$ variabile urmarite în P.

Din prezentarea cadrului patrat si dreptunghiular al unui grafic matrice se observa ca dimensiunea acestuia este **functie de numarul variabilelor** V_i

($i = 1, p$) observate, sau de **marimea lui r si p-r**. Aceasta înseamnă ca un numar prea mare de variabile urmarite în P si vizualizate în graficul matrice are drept consecinta **o densitate mare a imaginilor** din matrice, ceea ce creaza dificultati în interpretarea independenta sau în interdependenta a structurilor vizualizate de fiecare **bloc (componenta) matriceal (a)** si implicit afecteaza calitatea informatiilor sintetizate si transmise.

Din acesta cauza sugeram ca elaborarea graficului matrice sa fie precedata de **rationamente, de tehnici statistice adecvate si riguros utilizate pentru selectarea celor p variabile**. De exemplu, în afara criteriilor logice pot fi utilizate teste adecvate de verificare a verosimilitatii diverselor ipoteze statistice (formulate aprioric), referitoare la gradul de semnificatie al variabilelor precizate în programul de observare, pentru un grad de probabilitate ales (acelasi sau diferit). Vor face astfel obiectul graficului matrice numai **variabilele semnificative (cu mare încarcatura informatională)**, numai variabilele pentru care s-a confirmat ipoteza existentei anumitor tipuri de legaturi semnificative.

Titlul general al graficului matrice, **titlurile interioare** acestuia precum si **explicitarea axelor rectangulare** ca elemente ge-

nerale de constructie ale graficelor – sunt alese si precizate în **concordanta cu tipologia** grafica si cu **continutul acesteia** din interiorul matricei.

Scalele utilizate în construirea histogramelor plasate pe prima linie si coloana sau pe diagonala principala pot fi **comune** sau **distincte**. **Scala comuna** stabilita în functie de analiza calitativa prealabila si de amplitudinea variatiei datelor specifice variabilelor luate în considerare faciliteaza comparatii între distributiile vizualizate, între formele si densitatile de repartitie a frecventelor. Precizam, totusi, ca în cazul utilizarii scalelor comune exista riscul ca vizualizarea unor repartitii (formate dupa anumite variabile) sa fie dificila, daca ordinul de marime al valorilor individuale ale unor variabile sau amplitudinea variatiei lor difera foarte mult. Utilizarea **scalelor distincte** permite o mai buna definire a histogramelor individuale cu toate ca îngreuneaza comparabilitatea vizuala a acestora. Indiferent de scalele utilizate, la construirea histogramelor o atentie deosebita trebuie acordata stabilirii numarului si marimii intervalelor de variatie, deoarece acestea trebuie subordonate urmatorului scop: valorile din interiorul intervalelor sa fie relativ omogene si, pe cât posibil, sa urmeze o distributie cât mai apropiata de cea „normala”.

Pentru a sintetiza vizual informatiile furnizate de diferitele tipuri de repartitii multivariate, graficele în matrice pot avea urmatoarea tipologie: a) **diagrame de împrastiere**; b) **diagrame trasate prin linii**; c) **diagrame prin coloane**.

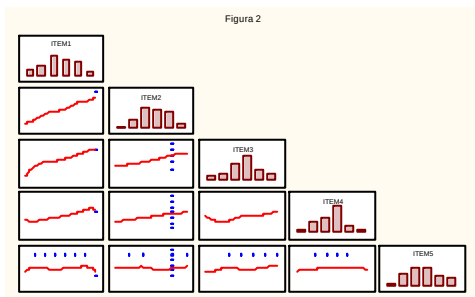
a) **Diagramele de împrastiere** sunt utilizate pentru descrierea legaturilor dintre variabilele $V_1, V_2, \dots, V_p$ sau între ($V_1, V_2, \dots, V_r$) si ($V_{r+1}, V_{r+2}, \dots, V_p$) înzestrate cu structurile corespunzatoare. În aceasta situatie, matricea diagramelor de împrastiere poate fi considerata corespondenta în imagini a matricii coeficientilor de corelatie.

În **matricea patratica a diagramelor de împrastiere** fiecare relatie (cuplu) dintre

(de) variabile (luate doua câte doua) din spatiile respective, înzestrate cu structuri corespunzatoare $[(V_i, V_j): P \rightarrow O_i \times O_j]$ înzestrat cu S_i si S_j ($i, j = \overline{1, p}$)] este dublu si simetric reprezentata fata de diagonala principala. În consecinta, graficelor din interiorul matricei le este atasata **o ordine** si **o simetrie fata de prima diagonala** numita si diagonala de histograme (fig. 1).

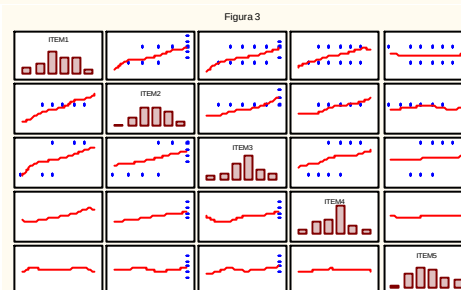
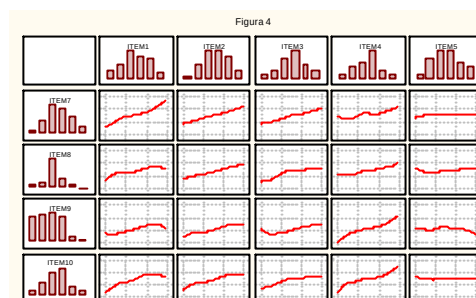


Din aceasta cauza pentru analiza legaturilor dintre variabile (multivariata) se poate construi si utiliza doar **matricea triunghiulara de sub diagonala principala** (fig.2).



În cazul în care, potrivit obiectivelor cercetării, se stabilesc anumite ipoteze referitoare la dependenta manifestării unei variabile de cele ale celorlalte variabile interesează, în consecinta, vizualizarea dependentei formale dintre variabile. În acest caz, utilizându-se o functie de ajustare într-un grafic matrice cu cadru patratic, **funcțiile folosite pentru graficele simetrice sunt diferite** în functie de relatia de cauzalitate, în functie de variabila dispusa pe axa absciselor si de cea plasata pe axa ordonatelor (fig.3). Prin urmare, într-o

matrice de diagrame de împrastiere se poate identifica vizual, cu usurinta, variabila care nu este corelata cu niciuna din cele $(p-1)$ variabile sau una din cele $V_1, V_2, \dots, V_r$ variabile necorelata cu niciuna din cele $(p-r)$ variabile si de asemenea se pot identifica vizual acele variabile (histograme) ale caror repartitii nu manifesta tendinta de normalitate (fig.4).



În diagramele de împrastiere dintr-o matrice pot fi reprezentate diferite functii de ajustare: liniare, logaritmice, exponentiale, polinomiale etc.

Vizualizarea în graficul matrice a relațiilor de dependenta formală dintre variabile (a raportului cauzal dintre fenomene) este deseori afectata de **prezența unor date atipice sau a unor valori individuale extreme**. Mai mult, exista cazuri când unele date din cei doi vectori $(V_1, V_2, \dots, V_r)$ si $(V_{r+1}, V_{r+2}, \dots, V_p)$ sunt **absente** – adica, pot exista unitati statistice din P la care nu s-au înregistrat date la cel puțin una din variabilele luate în considerare. În consecinta, în reprezentarea grafica a diagramele de împrastiere în matrice trebuie **sa se tina seama de datele lipsa**, iar acest lucru se poate face în doua variante:

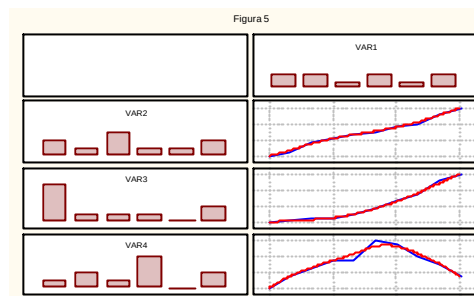
– fie se elimina, din toate graficele, unitatea statistica pentru care exista cel puțin o valoare lipsa la una dintre variabile;

– fie se elimina unitatea statistica doar în diagrama de împrastiere pentru care nu avem valide ambele coordonate.

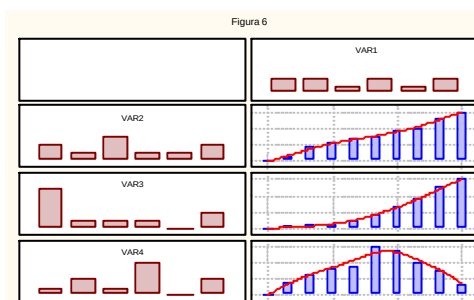
În primul caz, analiza prin imagini va face obiectul doar acelor unitati de observare care au un set complet de date pentru toate variabilele considerate. În acest caz exista riscul ca dependentele dintre variabile sa apara mult mai intense decât sunt în realitate, sau exista riscul sa se ajunga astfel la situatia în care nu exista **date suficiente, valide**, pentru $(V_1, V_2, \dots, V_p)$.

În al doilea caz, exista riscul ca intensitatea corelatiei sa fie influentata de prezenta unor valori atipice sau a unor valori extreme la unitatile la care nu s-au înregistrat date pentru toate variabilele. Si într-un caz si în celalalt **verificarea prealabila a calitatii datelor este decisiva** pentru minimizarea riscului deformarii realitatii pe cale vizuala, prin graficul matrice.

b) O alta structura atasata graficului matrice poate fi formata din **diagramele trasate prin linii**. Aceasta structura poate fi atasata pentru a sintetiza vizual atât informatiile privind **relatiile de cauzalitate (de dependenta)** dintre variabile cât si informatiile privind **evolutia unor fenomene sau procese** urmarite în populatia statistica P. În acest caz variabilele $(V_1, V_2, \dots, V_p)$ sunt plasate distinctiv pe axele ordonatelor diagramelor din matrice, iar pe axa absciselor fiecarei diagrame se reprezinta variabila timp. Prin liniile care unesc punctele de coordonate XY se vizualizeaza în fiecare diagrama evolutia în timp a fiecareia din cele p variabile (fig.5).

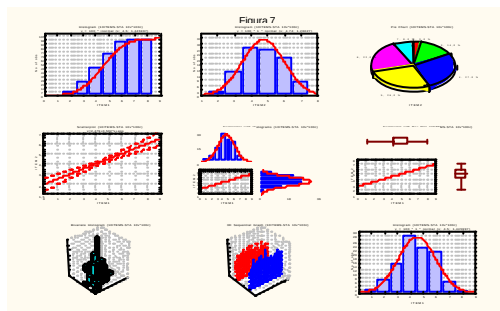


c) **Diagramele prin coloane** formeaza structura unui alt grafic matrice. Ceea ce este specific acestui tip de grafic matrice este faptul ca fiecare punct plasat într-o diagrama poate fi reprezentat, în doua dimensiuni, printr-o coloana. Punctele trasate pe axa absciselor corespund unitatilor de observare (momentelor sau intervalelor de timp). Din aceste puncte se ridica coloane până în dreptul valorii individuale a fiecărei variabile înregistrate. În felul acesta fiecare diagrama vizualizeaza evolutia fiecărei variabile $(V_1, V_2, \dots, V_p)$, în aceeași perioada de timp. Un astfel de grafic matrice sintetizeaza vizual evolutia ansamblului de variabile, sugereaza direct **comparatii ale amplitudinilor de variatie, ale formelor si directiilor de evolutie, tipul functiilor matematice care pot fi asociate evolutiilor respective** (fig.6).



Graficele matrice (indiferent de tipologie) construite după reguli exacte (care au fost anterior precizate) prezintă - în scopul cunoașterii prin sinteze vizuale - avantaje multiple și complementare față de alte tipuri de reprezentări grafice utilizate frecvent în practică. Dar ceea ce este mai important și cu mai directă și sugestivă aplicabilitate în cunoașterea și funda-

mentarea deciziilor este faptul ca **graficele matrice ofera premisele analizei statistice multivariate prin imagini**. Aceasta nu poate fi concretizata decât prin utilizarea unui **pachet adecvat de programe informatice**, chiar cu posibilitati de abordari interactive. Atributul de „**adecvat**” are în vedere nu numai **calitatea tehnicilor** de elaborare a programelor cât si **unele aspecte – care nu sunt deloc secundare** – cum ar fi: sa tina seama de **natura datelor**; sa pastreze **proportiile reale** dintre date; sa evidentieze **relatiile obiective** de cauzalitate; sa faciliteze **comparatiile vizuale** cu repartitiile teoretice; sa **minimizeze optim densitatea** imaginilor în matrice; sa evite aparitia în matricea grafica a **elementelor „nule”**; sa permita **localizarea** valorilor sintetice calculate pentru anumiti parametri etc.



Aceasta concluzie nu trebuie privita „**me-canic**”, deoarece graficele matrice si analiza statistica multivariata prin imagini **nu exclud ci presupun** efectuarea în etape premergatoare a reprezentarilor grafice si analizelor statistice unidimensionale. **Mai mult, calitatea informatiilor transmise de graficele matrice si de analizele statistice multivariate prin imagini este decisiv conditionata de calitatea datelor înregistrate, de calitatea si gradul de adecvare ale operatiilor precedente de prelucrare a datelor.**

Bibliografie:

1. Crauser, J.P.- Ghide pratique d'analyse des donnés, Ed. D'organisation, Paris, 1989.
2. Dreesbeke, J.-J.- Elements de statistique, Ed. a II-a, Dunod, Paris, 1988.
3. Hasiganu, D.; Marinescu, I.- Grafice si elemente de calcul grafic, Ed. Stiintifica, Bucuresti, 1968.
4. Isaac-Maniu, Al.; Mitrut, C.; Voineagu, V. - Statistica pentru managementul afacerilor, Ed. a II-a, Ed. Economica, Bucuresti, 1999.
5. Jambu, M.- Exploration informatique et statistique des données, Dunod, Paris, 1970.
6. Morisson, D.F.- Multivariate Statistical Methods, London, 1967.
7. xxx - Statistica for windows (Vol.II) Graphics (2nd Edition), 1998 by StatSoft, Inc.