*Rivista di Statistica Applicata Vol. 19 n. 2 Giugno 1986*

**STATISTICA COMPUTA-ZIONALE**

*Rubrica a cura di Natale Lauro*

# Il metalinguaggio Genstat e l'analisi multidimensionale dei dati: un programma per l'analisi non simmetrica delle corrispondenze

**N. Lauro, L. D'Ambra**, Dipartimento di Matematica e Statistica Università di Napoli
**A. Calvaruso**, ERA-Elaborazioni, Ricerca e Analisi dei dati-Napoli

## 1. Introduzione

In un recente lavoro presentato alla XXXII Riunione Scientifica della Società Italiana di Statistica, Norbert Victor, interrogandosi sul tema "Computational Statistics-Tool or Science?" concludeva che la Statistica Computazionale non costituisce una disciplina autonoma bensì parte integrante della Statistica (N. Victor 1984).

L'impatto del computer nella Statistica come ben sottolineato da Victor non si è avuto solo nella ricerca applicata ma anche nella ricerca metodologica.

Aggiungiamo da parte nostra che l'Informatica ha profondamente modificato il modo di pensare dello statistico, per cui riteniamo che la Statistica computazionale non possa essere semplicemente intesa come un nuovo capitolo della Statistica ma piuttosto un nuovo modo di fare Statistica nell'era del computer.

Lo sviluppo di metalinguaggi orientati alla Statistica rappresenta concretamente un tentativo di formalizzare il moderno pensiero statistico.

I metalinguaggi si propongono non solo di predisporre strumenti informatici per l'analisi dei dati ma altresì di affrontare nuove forme di comunicazione tra gli statistici e dunque nuovi modi di produrre Statistica.

Obiettivo di questa nota è quello di presentare uno di questi metalinguaggi, il GENSTAT, GENeral STATistical program (Nelder, 1978; Alvey et al. 1982), e di fornire un esempio della sua versatilità per l'analisi multidimensionale dei dati (AMD) benché esso sia di più generale impiego.

Così dopo aver discusso le caratteristiche generali del Genstat ne verranno illustrate le principali funzioni per il calcolo matriciale e le direttive di base per l'AMD con particolare riferimento alla formalizzazione e alla implementazione di un algoritmo per "l'Analisi non simmetrica delle corrispondenze" (Lauro D'Ambra 1984).

## 2. Caratteristiche principali del Genstat

Il Genstat è un programma scritto in Fortran (eccetto alcune subroutine Assembler) articolato in circa 300 sottoprogrammi (*).

Esso è fornito di un linguaggio interpretativo volto a facilitare la gestione, il trattamento e l'elaborazione statistica dei dati.

Il Genstat si colloca in una posizione intermedia tra i packages statistici e i linguaggi simbolici di tipo scientifico. Dai primi si differenzia per la sua struttura flessibile, la possibilità di concatenare analisi e di programmare algoritmi originali. Per contro, rispetto ai linguaggi simbolici, esso è accessibile anche a ricercatori che non hanno elevate esperienze informatiche.

La programmazione Genstat consiste, infatti, nel selezionare quelle parole chiave (generalmente di tipo mnemonico) atte a produrre l'analisi richiesta dal ricercatore.

In particolare la programmazione in Gestat offre la possibilità di concatenare diversi algoritmi, la Fig. 1 evidenzia come l'output di un'analisi può divenire l'input per una analisi successiva.

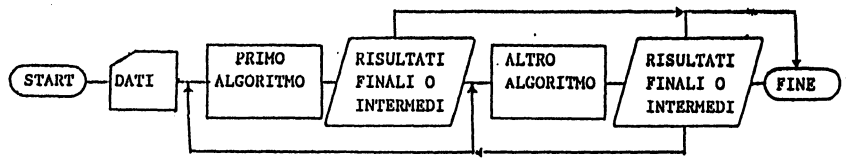



**Fig. 1 - Concatenazione di programmi Genstat**

(*) La versione del Genstat cui facciamo riferimento è la 4.03E implementata su numerosi mainframes e recentemente anche su Pc's IBM e compatibili.

Ciascuna istruzione Genstat agisce come una subroutine Fortran, tuttavia la sequenza di parametri di Input/Output è molto più flessibile. A ciascun parametro di una subroutine Fortran infatti corrisponde in Genstat una lista di parametri i cui identificatori sono, il più delle volte, scelti dall'utente.

Le istruzioni Genstat sono generalmente indicate con il nome di *DIRETTIVE* le quali, a loro volta, si distinguono in DICHIARAZIONI e COMANDI.

Un programma Genstat è costituito da una sequenza di direttive relative alle seguenti funzioni:

(a) — DICHIARAZIONE delle strutture su cui si vuole operare
(b) — LETTURA dei valori da inserire nelle strutture dichiarate
(c) — CALCOLI richiesti dall'analisi
(d) — STAMPA dei risultati

(a) — Le strutture riconosciute dal Genstat sono riportate nella figura 2

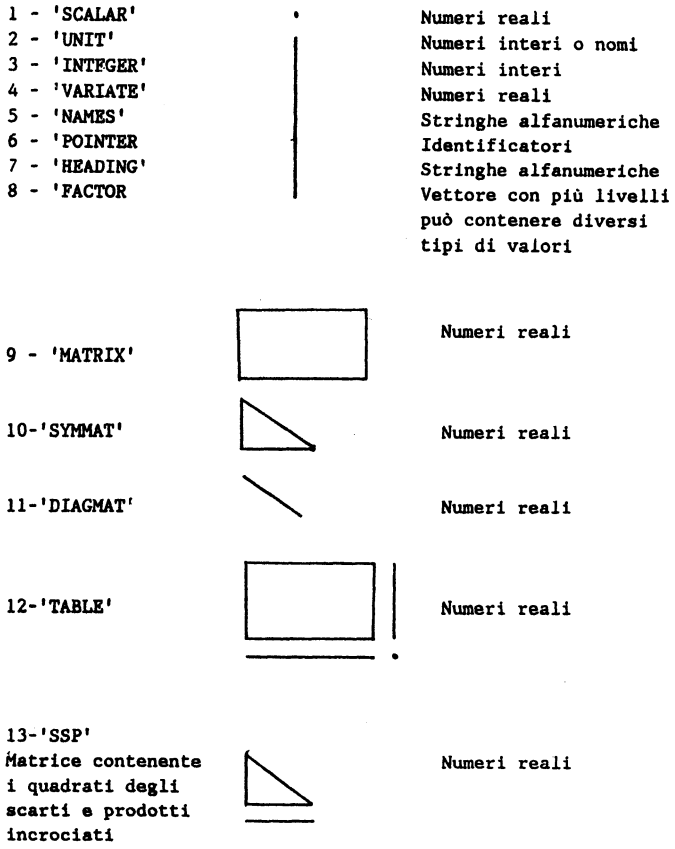



**Fig. 2 - Strutture dati del Genstat**

(b) — Le operazioni di lettura possono esser effettuate o con la direttiva "VALUE" o con la "READ" specificando il nome della struttura in oggetto.

(c) — I calcoli si effettuano utilizzando il comando "CALCULATE" e specificando le liste di strutture su cui deve operare. Oltre a questo comando ne esistono altri in grado di effettuare analisi statistiche quali:

— Analisi di disegni sperimentali
— Modelli lineari generalizzati
— Analisi di regressione
— Analisi fattoriale
— Classificazione automatica
— Analisi delle serie storiche

(d) — La stampa viene effettuata utilizzando il comando "PRINT" per gli output tabulari o "GRAPH" e "HIST" per la stampa di grafici e istogrammi.

Un programma tipo può presentarsi come costituito da diversi job (dove ogni job inizia con la direttiva "REFERENCE" e termina con "CLOSE") i quali, a loro volta, possono essere suddivisi in diversi blocchi (ciascun blocco inizia con "START" e termina con "RUN"). Un blocco può essere definito come un insieme di istruzioni tali da produrre un risultato intermedio dell'analisi, mentre un job è un insieme di blocchi tale da produrre il risultato finale dell'analisi.

In Genstat è possibile assegnare un nome ad uno o più blocchi di istruzioni e farle, quindi, riconoscere dal sistema come un corpo unico. Ciò risulta utile qualora si debbano effettuare delle operazioni di routine all'interno di un programma o si vogliano realizzare delle procedure di calcolo ex-novo.

Il Genstat è dotato di una MACRO-LIBRERIA di sistema che permette di affrontare un elevato numero di analisi statistiche.

Allo sviluppo di tale Macro-libreria, oltre al Dipartimento di Statistica della Rothamsted Experimental Station, partecipano tutti gli utilizzatori del Genstat. In questo senso, il Genstat, rappresenta anche uno strumento di ricerca e di diffusione di idee ed esperienze.

La Gestione dei File

Le strutture dichiarate dall'utente vengono, automaticamente, memorizzate su una particolare area detta MEMORIA CENTRALE DEL SISTEMA. Queste tuttavia possono essere rimosse e trasferite in file sia temporanei che permanenti. È quindi possibile memorizzare dei risultati, delle strutture dati o interi programmi i quali potranno essere richiamati soltanto al momento del loro utilizzo.

Una approfondita trattazione delle capacità informatiche e statistiche del Genstat è stata oggetto di una pubblicazione del Centro di Calcolo Interfacoltà dell'Università di Napoli (A. Calvaruso, 1983).

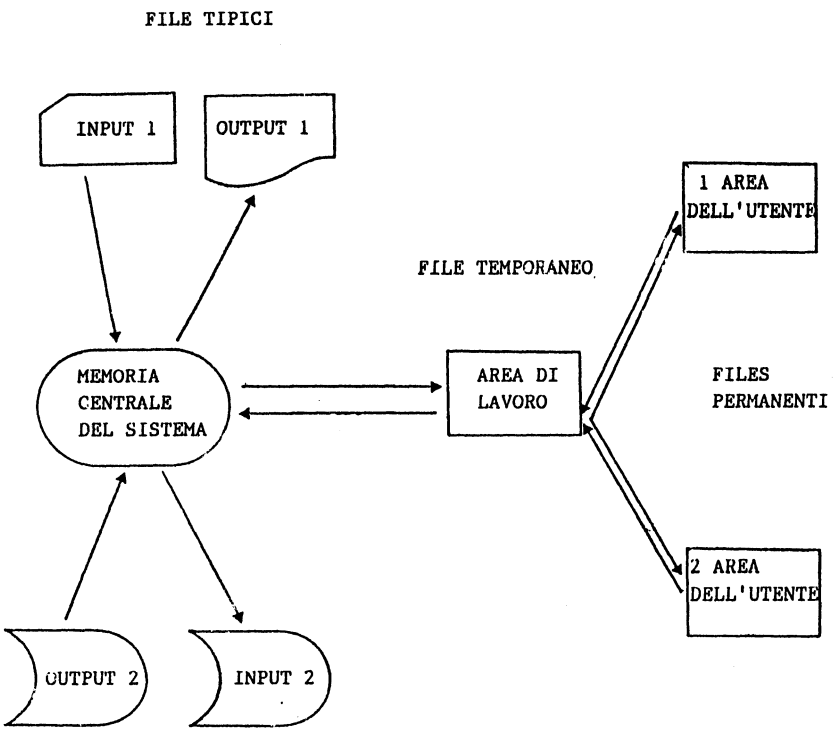



**Fig. 3 - Gestione dei files in Genstat**

## 3. Il Genstat e l'Analisi Multidimensionale dei Dati

Il Genstat permette di affrontare problemi di AMD sia attraverso l'impiego di funzioni di calcolo matriciale che mediante direttive speciali.

Funzioni per il calcolo matriciale

Il comando "CALCULATE" opera indifferentemente sia su strutture scalari che su strutture vettoriali o matriciali. Ciò comporta notevoli facilitazioni specie per la risoluzione di problemi di AMD.

Le principali funzioni matriciali del Genstat definite su matrici di appropriate dimensioni, simbolicamente indicate con X, Y, Z, sono le seguenti:

— TRACE(X) calcola la traccia di una matrice X
— TRANS(X) effettua la trasposizione X′
— INV(X) calcola l'inversa
— DET(X) calcola il determinante
— CORMAT(X) calcola la matrice di correlazione (a partire da una matrice di covarianza)
— CHOL(X) effettua la decomposizione di Choleski X=Z′Z

— PDT(X;Y) effettua il prodotto matriciale XY
— PDTT(X;Y) effettua il prodotto X Y′
— TPDT(X;Y) effettua il prodotto X′ Y
— RSYMRI(X;Y) effettua il prodotto X Y X′

Vi sono inoltre due direttive speciali:

a — "LRV" per estrarre gli autovalori e gli autovettori della matrice X simmetrica e semidefinita positiva;

b — "SVD" per decomporre la matrice X, rettangolare, nel prodotto matriciale *USV*, essendo S la matrice degli autovalori, U la matrice degli autovettori associati alla matrice X X′ e V la matrice degli autovettori associati a X′ X.

Oltre alle funzioni di calcolo matriciale elencate in precedenza il Genstat prevede una serie di direttive che consentono la formalizzazione e l'esecuzione completa delle AMD di base e loro varianti attraverso l'utilizzo di numerose opzioni. Per la analisi fattoriale in particolare si hanno le direttive:

"PCP" per l'analisi in componenti principali
"ROTATE" per l'analisi procusteana
"PCO" per l'analisi delle coordinate principali
"CVA" per l'analisi fattoriale discriminante
"FACROT" per la rotazione dei fattori
"ADPT" per la proiezione di osservazioni supplementari.

Per quanto concerne i metodi di classificazione automatica:
"CLASSIFY" per la formazione di partizioni;
"HIERARCHY" per metodi gerarchici di tipo agglomerativo.

Numerose sono le macro disponibili nella macrolibreria associata al sistema che riguardano l'analisi mutlidimensionale dei dati (Banfield, 1978).

Tra esse citiamo:
"CANCOR" (correlazione canonica); CORRESP (analisi delle corrispondenze); BIPLOTV (metodo del Biplot di Gabriel); INDSCAL (metodo di scaling individuale di Carrol e Chang); ASYMANAL (analisi di matrici non simmetriche di Gower); SVD3 (generalizzazione della decomposizione singolare per tabelle a 3 indici); PCAID (ausilii alla interpretazione di una analisi in componenti principali); DISQUAR (analisi discriminante, con classificazione di soggetti ALLOC e individuazione di soggetti mal classificati MISALLOG MISALLOP); GENPROC (analisi procusteana generalizzata e sua presentazione grafica GPROCPLT e GROCLAB).

Si tratta, a ben vedere, di una lista consistente che comprende gli sviluppi più recenti dell'AMD e che non trova eguale nè nei package general purpose di più larga diffusione nè in biblioteche specializzate (Lauro, Serio 1982).

## 4. Un programma GENSTAT per l'analisi non simmetrica delle corrispondenze (ANSC)

Obiettivo dell'ANSC è lo studio dei legami tra caratteri qualitativi qualora non sia ipotizzabile una relazione simmetrica tra gli stessi. (Lauro, D'Ambra-1984).

Contrariamente all'Analisi delle Corrispondenze, (Benzecri 1972) che fa riferimento alla decomposizione del $\Phi^2$ di Pearson, essa si basa sulla decomposizione di un indice di predicabilità quale il $\tau$ di Goodman-Kruskal (1954) e le sue estensioni al caso di più variabili qualitative dovute a Gray e Williams (1975).

Dal punto di vista algoritmico, l'ANSC si traduce:

(*a*) — In una analisi in componenti principali di una tabella di profili

$$F\ D_j^{-1}$$

nella metrica $I_p$ e secondo il criterio associato alla matrice dei pesi $D_j$.

Essendo $F$ una tabella di contingenza a $p$ righe e $q$ colonne e $D_j$ la matrice delle frequenze marginali di colonna, l'analisi si riconduce alla diagonalizzazione della matrice simmetrica

$$Q_1 = F\ D_j^{-1}\ F' - f_I f_I'$$

essendo $f_I$ il vettore delle frequenze marginali delle $p$ righe.

La traccia di questa matrice, a meno di una costante, pari a $(1\text{-}f_I' f_I)^{-1}$ corrisponde al $\tau$ di Goodman-Kruskal.

(*b*) — In una analisi in componenti principali dell'immagine della matrice $X$ ottenuta per mezzo di un operatore di proiezione ortogonale sul sottospazio di riferimento incentrato sulle colonne di una matrice $Z$, essendo le colonne di $X$ e $Z$ le variabili indicatrici associate alla codifica disgiuntiva completa di due caratteri qualitativi $A_x$ e $A_z$ rispettivamente a $p$ e $q$ modalità (D'Ambra e Lauro 1982).

Quest'ultimo approccio si basa sulla diagonalizzazione della matrice simmetrica:

$$(1/n)\ X'Z\ (Z'Z)^{-1}\ Z'\ X - (1/n)^2\ X'\ u_n u'_n\ X$$

dove $u'_n$ è il vettore riga ad elementi unitari.

La relazione tra i due approcci è evidente ove si tenga conto che:

$$F = X'Z \text{ e } D_j = Z'Z$$

Questo secondo approccio si rivela utile per l'estensione all'analisi di più insiemi di caratteri qualitativi, sia per tener conto di eventuali interazioni tra gli stessi operando su opportune matrici costituite mediante il prodotto cartesiano delle variabili indicatrici associate alle singole modalità.

## 4.1 Le fasi dell'algoritmo dell'ANSC

Elenchiamo le diverse fasi dell'algoritmo identificandole con una numerazione opportuna

(1) Lettura della tabella di contingenza $F$ e costruzione delle matrici dei profili.
(2) Calcolo degli indici di associazione

$$\tau_{xz} \text{ (TAUXZ)}, \ \tau_{zx} \text{ (TAUZX)}, \ \Phi^2 \text{ (PHI2)}$$

(3) Costruzione della matrice da diagonalizzare: $Q1 = F\, D_j^{-1}\, F' \text{-} f_I f'_I$
(4) Calcolo degli autovalori AVA ($\alpha$) e degli autovettori AVE ($\alpha$) e della qualità globale della rappresentazione:

$$\text{TAU}(h) = \Sigma_{\alpha=1}^{h} \text{AVA}(\alpha) \ / \ \Sigma_{\alpha=1}^{p} \text{AVA}(\alpha) \ (h < min \ (\text{p,q}) )$$

(5) Calcolo delle coordinate fattoriali delle modalità delle righe e delle colonne della tabella $F$:

$$\text{COORDX}(\alpha) = \text{AVE}(\alpha) \times \text{AVA}(\alpha)^{1/2}$$

$$\text{COORDZ}(\alpha) = Q_1 \times \text{AVE}(\alpha)$$

(6) Calcolo dei contributi delle modalità delle righe e colonne:

$$\text{CONTRX}(\alpha) = (\text{COORDX}(\alpha))^2 \ / \ \text{AVA}(\alpha)$$

$$\text{CONTRZ}(\alpha) = (D_j^{-1} = DJJ \times (\text{COORDZ}(\alpha)^2) \ / \ \text{AVA}(\alpha)$$

(7) Qualità delle rappresentazioni:

$$\text{COS2X}(\alpha) = (\text{COORDX}(\alpha))^2 \ / \ \Sigma_\alpha(\text{COORDX}(\alpha)^2)$$

$$\text{COS2Z}(\alpha) = (\text{COORDZ}(\alpha))^2 \ / \ \Sigma_\alpha(\text{COORDZ}(\alpha)^2)$$

(8) Rappresentazione sui piani fattoriali $\alpha \neq \alpha'$

$$\text{COORDX}(\alpha), \ \text{COORDZ}(\alpha) \text{ versus } \text{COORDX}(\alpha') \ \text{COORD}(\alpha')$$

Il programma realizzato, abbastanza curato nella stampa, comprende circa 90 istruzioni contro le oltre 1000 del corrispondente programma in FORTRAN. Una versione più essenziale, sottoforma di macro potrebbe essere realizzata con una trentina di istruzioni. A titolo di confronto ricordiamo che la macro e i programmi realizzati per l'analisi delle corrispondenze classica richiedono da un minimo di 42 ad un massimo di 373 istruzioni (Bernard, 1977).

La numerazione adottata per le diverse fasi dell'algoritmo, gli elementi del metalinguaggio richiamati in precedenza e il mantenimento delle stesse

notazioni mnemoniche impiegate per la formalizzazione delle strutture considerate nell'algoritmo, consentono una lettura agevole del programma che non abbisogna di ulteriori commenti. (fig. 4).

Il programma è generalizzato e le uniche linee da sostituire sono quelle relative alle denominazioni delle variabili qualitative diagnosi e malattie date nella istruzione "NAME" e nella "READ" conseguente.

Il programma opera su una tabella di contingenza le cui dimensioni $P$ e $Q$ vanno fornite inizialmente.

L'analisi non simmetrica delle corrispondenze multiple può essere effettuata utilizzando lo stesso programma e fornendo in input una opportuna matrice, giustapposizione degli strati di una tabella multipla. Così ad esempio per una tabella a tre indici di termine generale $f_{ijk}$ la tabella analizzata sarà

$$F = [f_{ij1}\ f_{ij2}\ \ldots\ f_{ijk}]$$

cui vanno associati opportuni pesi derivati dai corrispondenti marginali di colonna. Il passo (2) fornisce in tal caso il TAU multiplo di Gray-Williams.

La centratura della $F$ per mezzo di una opportuna matrice di marginali ($f_{ij\cdot}$) consente di effettuare una analisi nor simmetrica delle corrispondenze parziali.

```
[1] 'REFE' ANSC
    'SCAL' N,P,Q
    'READ' N,P,Q
    'RUN'
    'NAME' MEDICINE = Q
    'NAME' MALATTIE = P
    'READ/S' MEDICINE,MALATTIE
    'RUN'
    'FACT' XTAB = MALATTIE,P = 1...P
    'FACT' ZTAB = MEDICINE,Q = 1...Q
    'TABLE/MARG=N' FREQ = XTAB,ZTAB
    'READ/PRINT=N' FREQ
    'MARG' FREQ
    'PRINT/UNKN=1' FREQ = 8
    'RUN'
[2] 'SCAL' TAUXZ,TAUZX,PHI2,T(1...4)
    'MATR' F = XTAB,ZTAB
    'DIAG' DI = XTAB
    :      DJ = ZTAB
    'EQUA' F,DJ,DI = FREQ = (P,1X)Q,P,1X,((1X)P,1)Q
    'CALC' T(2)=Q-1
    'PRINT' DI,DJ = 8
    'CALC' F,DI,DJ = (F,DI,DJ)/N
    'CALC' DI,DJ = 1/(DI,DJ)
    'CALC' TAUXZ=(TRACE(PDTT(PDT(F;DJ);F))-TRACE((1/DI)**2))/(1-TRACE((1/DI)**2))
    :      TAUZX=(TRACE(PDT(TPDT(F;DI);F))-TRACE((1/DJ)**2))/(1-TRACE((1/DJ)**2))
    :      PHI2 = (TRACE(PDT(PDTT(PDT(F;DJ);F);DI))-1)
    'PAGE'
    'CAPT'  '' *** INDICI DI ASSOCIAZIONE *** ''
    'PRINT' TAUXZ,TAUZX,PHI2 = 10.3
    'PAGE'
    'RUN'

[3] 'SYMM' S = XTAB
    'MATR' Q1 = XTAB,ZTAB
    :      UNOZ = 1,ZTAB = 1(P)
    :      DII = XTAB,1
    'CALC' DI=1/D1
    'EQUA' DII = DI
```

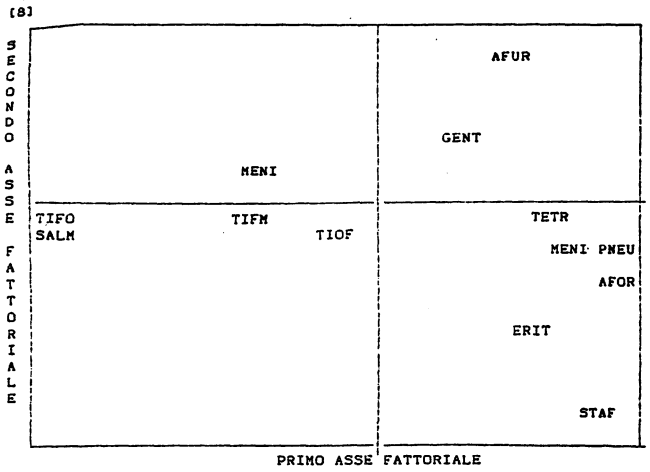
**Fig. 4 - PROGRAMMA ANSC** **(continua a pag. 134)**

```
      'PRINT' DII $ 8.2
      'CALC' Q1 = PDT(F;DJ)-PDT(DII;UNOZ)
      :    DJ = 1/DJ
      'VARI' DJJ $ ZTAB
      'CALC' S = PDTT(PDT(Q1;DJ);Q1)
      :    DI = 1/DI
      'EQUA' DJJ = DJ
      'PAGE'

      'CAPT'  "          *** MATRICE DA DIAGONALIZZARE *** "
      'PRINT/LABC=1' S $ 10.3
      'PAGE'
      'RUN'


[4]   'DIAG' AVAC $ Q
      'MATR' AVE   $ Q,Q
      'LRV' S; AVE,AVAC,T(1)
      'CALC' AVAC = ABS(AVAC)
      'SCAL' AVAB(1...P)
      'EQUA' AVAB(1...Q)=AVAC
      'VARI' AVA(1...3) $ Q
      'EQUA' AVA(1) = AVAB(1...Q)
      'CALC' T(3) = SUM(AVA(1))
      :    AVA(2) = (AVA(1)/T(3))*100
      :    AVA(3) = CUM(AVA(2))
      'PAGE'
      'CAPT' '' ***  AUTOVALORI   PERCENTUALE   PERC.CUM *** ''
      'PRINT/P' AVA(1...3) $ 10.3,2(10)
      'PAGE'
      'RUN'

[5]   'CALC' AVAC = SQRT(AVAC)
      'MATR' XCOOR $ XTAB,Q
      'CALC' XCOOR = PDT(AVE;AVAC)
      'MATR' ZCOOR $ ZTAB,Q
      'CALC' AVAC = (1/(AVAC*1000000))*1000000
      'CALC' ZCOOR = TPDT(Q1;AVE)
      'VARI' X2,COS2X(1...Q),CONTRX(1...Q),COORDX(1...Q) $ XTAB
      :    Z2,COS2Z(1...Q),CONTRZ(1...Q),COORDZ(1...Q) $ ZTAB
      'EQUA' COORDX(1...P) = XCOOR $ (1,(1X)T(2))Q,1X
      :    COORDZ(1...P) = ZCOOR $ (1,(1X)T(2))P,1X
      'PAGE'
      'CAPT'" *** COORDINATE FATTORIALI DELLE MODALITA' RIGA   ***"
      'PRINT'AVE $ 10.3
      'PRINT'XCOOR $ 10.3
      'PRINT/P' MEDICINE,COORDX(1...Q) $ 10.3
      'PAGE'
      'CAPT' " *** COORDINATE FATTORIALI DELLE MODALITA' COLONNA  ***"
      'PRINT'ZCOOR $ 10.3
      'PRINT/P' MALATTIE,COORDZ(1...Q) $ 10.3
      'PAGE'

[6]   'CALC' CONTRX(1...Q) =((COORDX(1...Q)**2)/SUM(COORDX(1...Q)**2)*1000)
      'PAGE'
      'CAPT' " *** CONTRIBUTI DELLE MODALITA' RIGA ***"
      'PRINT/P' MEDICINE,CONTRX(1...Q) $ 10.3
      'PAGE'
      'CALC' COS2X(1...Q) =(COORDX(1...Q)**2)
      :    COS2Z(1...Q) =(COORDZ(1...Q)**2)
      CALC' CONTRZ(1...Q)= (COS2Z(1...Q)*DJJ)/AVAB(1...Q)
      'CAPT' " *** CONTRIBUTI DELLE MODALITA'COLONNA ***"
      'PRINT/P' CONTRZ(1...Q) $ 10.3
      'CALC' X2 = VSUM(COS2X(1...Q))
      :    Z2 = VSUM(COS2Z(1...Q))
      :    COS2X(1...Q) =(COS2X(1...Q)/X2*1000)
      :    COS2Z(1...Q) =(COS2Z(1...Q)/Z2*1000)

[7]   'CAPT' " *** QUALITA' DELLA RAPPRESENTAZIONE DELLE MODALITA' RIGA ***"
      'PRIN/P' MEDICINE,COS2X(1...Q) $ 10.3
      'PAGE'
      'CAPT' " *** QUALITA' DELLA RAPPRESENTAZIONE DELLE MODALITA' COLONNA ***"
      'PRIN/P ' MALATTIE,COS2Z(1...Q) $ 10.3
      'PAGE'
      'RUN'

[8]   'HEAD' AX = " PRIMO ASSE FATTORIALE "
      :    AY = " SECONDO ASSE FATTORIALE "
      'GRAP/ATY=AY,ATX,EQXY=NCF=120,NRF=54' COORDX(2),COORDZ(2);
      COORDX(1),COORDZ(1) $ ; XTAB,ZTAB
      'RUN'
```

```
[1]        TIFO SALM AFOR PNEU MENI AFUR STAF

      PENI  0    0    8    7    2    4    3      24
      TIFM  4    2    0    0    2    0    0       8
      TETR  0    0    5    5    0    2    1      13
      ERIT  0    0    3    2    0    0    3       8
      TIOF  2    1    0    0    0    0    0       3
      GENT  0    0    3    3    1    6    0      13

            6    3   19   17    5   12    7      69

[2]   INDICI DI ASSOCIAZIONE

      TAUXZ  0.180

      TAUZX  0.154

      PHI2   1.160

[3]   MATRICE DA DIAGONALIZZARE

      PENI  0.019
      TIFM -0.029  0.056
      TETR  0.011 -0.022  0.012
      ERIT  0.009 -0.013  0.004  0.015
      TIOF -0.015  0.024 -0.008 -0.005  0.013
      GENT  0.005 -0.016  0.003 -0.010 -0.008  0.003

[4]   AUTOVALORI   PERCENTUALE    PERC.CUM

      0.099           70             70
      0.031           23             93
      0.006            4             97
      0.004            3            100

[5]   COORDINATE FATTORIALI DELLE MODALITA' RIGA

      PENI   0.128  -0.022   0.033  -0.036   0.001
      TIFM  -0.235   0.004   0.019  -0.021  -0.001
      TETR   0.090  -0.013  -0.056  -0.020  -0.001
      ERIT   0.054  -0.104   0.021   0.037  -0.001
      TIOF  -0.106  -0.006  -0.030   0.020   0.002
      GENT   0.070   0.141   0.013   0.021   0.000

[5]   COORDINATE FATTORIALI DELLE MODALITA' COLONNA

      TIFO  -0.767  -0.023  -0.054   0.029   0.000
      SALM  -0.767  -0.023  -0.054   0.029   0.000
      AFOR   0.153  -0.064  -0.028  -0.024   0.004
      PNEU   0.155  -0.027  -0.061  -0.045  -0.003
      MENI  -0.248   0.092   0.214  -0.137   0.000
      AFUR   0.138   0.319   0.021   0.073   0.000
      STAF   0.133  -0.341   0.106   0.110   0.001

[6]   CONTRIBUTI DELLE MODALITA' RIGA

      PENI   165.28  15.50  178.11  295.28  179.18
      TIFM   560.04   0.39   59.36  105.39  108.14
      TETR    82.13   5.55  510.20   95.10  140.35
      ERIT    29.48 342.55   70.67  314.53   76.09
      TIOF   114.09   1.05  151.90   88.36  477.94
      GENT    48.98 634.95   29.76  101.34   18.30
```

**Fig. 5 - OUTPUT DELL'ANSC** **(continua a pag. 136)**

```
[6]  CONTRIBUTI DELLE MODALITA' COLONNA

     TIFO  518.00    2.00   42.00   18.00    0.00
     SALM  259.00    1.00   21.00    9.00    0.00
     AFOR   65.00   36.00   35.00   38.00  550.00
     PNEU   60.00    6.00  153.00  118.00  416.00
     MENI   45.00   19.00  547.00  314.00    2.00
     AFUR   34.00  562.00   13.00  218.00    0.00
     STAF   18.00  374.00  188.00  286.00   32.00

[7]  QUALITA' DELLA RAPPRESENTAZIONE DELLE MODALITA' RIGA

     PENI  851.72   25.43   56.34   66.45   0.06
     TIFM  985.26    0.22    6.41    8.10   0.01
     TETR  687.98   14.82  262.33   34.79   0.08
     ERIT  188.19  696.40   27.69   87.69   0.03
     TIOF  893.86    2.61   73.05   30.23   0.24
     GENT  190.27  785.44    7.09   17.19   0.00

[7]  QUALITA' DELLA RAPPRESENTAZIONE DELLE MODALITA' COLONNA

     TIFO   992.63    0.92   4.98   1.46   0.00
     SALM   992.63    0.92   4.98   1.46   0.00
     AFOR   808.94  143.34  26.88  20.40   0.45
     PNEU   785.48   24.13 122.77  67.26   0.36
     MENI   457.76   62.55 340.58 139.12   0.00
     AFUR   151.33  802.53   3.50  42.65   0.00
     STAF   112.90  738.38  71.38  77.33   0.01
```




## 4.2 L'Output del programma di ANSC

L'esempio di output riguarda la rielaborazione nell'ottica dell'ANSC di uno studio già utilizzato da F. Benzecri (1980) per una presentazione dell'analisi classica delle corrispondenze. I dati (output[1] ) riguardano lo studio della associazione tra 7 malattie (Tifo (TIFO), Salmonellosi (SALM), Affezioni orali (AFOR), Pneumopatia (PNEU), Meningite (MENI), Affezioni

vie urinarie (AFUR), Stafilococco (STAF) ) e 6 farmaci (Penicillina (PENI), Tifomicina (TIFM), Tetraciclina (TETR), Eritromicina (ERIT), Tiofemicina (TIOF), Gentalina (GENT) ).

La scelta dell'analisi non simmetrica è qui giustificata non solo dalla natura dei dati considerati dove una variabile qualitativa (la diagnosi) è antecedente logicamente rispetto all'altra (la scelta del farmaco), ma anche dal calcolo degli indici di associazione

$$\text{TAUXZ} = 0.180 \quad \text{TAUZX} = 0.154 \quad \text{(output [2] )}$$

La stampa della matrice da diagonalizzare (output [3] ) i cui elementi rappresentano varianze e covarianze spiegate permette di verificare la precisione dei calcoli, tra cui una inversione di matrice, poiché la somma per riga (colonna) è zero.

Dalla stampa degli autovalori (output [4] ) si evince che una rappresentazione sul primo piano fattoriale rappresenta ben il 93% dell'informazione dei dati originari. Seguono le coordinate fattoriali (output [5] ), ed altri ausilii all'interpretazione dei risultati come il calcolo dei contributi (output [6] ) e della qualità delle rappresentazioni delle diverse modalità, (output [7] ).

Il calcolo dei contributi (output [6] ) consente di valutare l'importanza giocata da ciascun elemento nell'orientamento degli assi non essendo sufficienti nel caso di un sistema di punti pesanti, come quello in studio, l'analisi delle coordinate e dunque delle rappresentazioni fattoriali.

Il primo asse risulta spiegato dai forti contributi del Tifo (560/1000) e dal farmaco corrispondente Tifomicina (518/1000). Il secondo asse dalle Affezioni urinarie (562/1000) e dalla Gentalina sul fronte dei farmaci (634/1000).

La qualità delle rappresentazioni (output [7] ) valutate come rapporto tra la rappresentazione dei profili riga (colonna) nello spazio originario e la immagine corrispondente sugli assi fattoriali consente di stabilire l'ottima rappresentazione per alcuni punti già sul primo asse fattoriale (Tifomicina, Penicillina, Tifo, Salmonellosi), e che le rappresentazioni sono generalmente più che adeguate nei primi due assi fattoriali eccetto che per Meningite e Tetraciclina la cui posizione riceve ulteriore spiegazione dal terzo asse fattoriale.

Si ricorda che le prossimità tra punti sui piani fattoriali possono valutarsi solo per quei punti che vi sono ben rappresentati.

La lettura delle prossimità sulle rappresentazioni dell'ANSC (output [8] ) va fatta, contrariamente a quanto avviene nella analisi delle corrispondenze, tenendo conto della direzione dei legami espressi ovvero della capacità predittiva dei punti diagnosi e non viceversa.

I risultati complessivamente mostrano una buona coerenza nelle rappresentazioni fattoriali dei punti malattie e dei farmaci associati.

Si noti ad esempio la forte associazione indotta lungo il primo asse tra Tifo e Salmonellosi da una parte e l'uso di Tifomicina e Tiofemicina. Tifo e

Salmonellosi risultano coincidenti in quanto rappresentazioni di profili uguali come del resto avviene nell'analisi delle corrispondenze.

Si osservi ancora la posizione baricentrica della Eritrocina rispetto alle malattie per le quali è impiegata: Pneumopatie (2), Affezioni orali (3), Staffilococco (2).

In altra sede (Lauro D'Ambra 1984) si è mostrato che l'analisi della corrispondenza classica non è adeguata per studiare la struttura delle associazioni della tabella qui analizzata a causa della simmetria con cui intervengono righe e colonne che spesso porta a ravvicinare le proiezioni di alcuni punti giocando a sfavore della loro interpretabilità.

Si tratta ovviamente di un confronto puramente indicativo, poiché evidentemente i metodi poggiano su presupposti teorici diversi, ma tale confronto è nondimeno utile per mettere in guardia anche sul piano pratico coloro che giustificano la scelta di un metodo solo perché il programma è disponibile.

Ricerca svolta con i contributi
MPI 40% Valutazione e documentazione software (coord. N. Lauro)
MPI 60% Analisi di dati qualitativi (coord. L. D'Ambra).

**Riferimenti bibliografici**

Alvey N. et al., 1982, *An introduction to Genstat*. Academic Press.
Benzecri, J.P. et al., *L'analyse des données: l'analyse des correspondance*. (Dunod, 1973).
Benzecri F., Introduction à l'analyse des correspondances d'apres un example de donnèes mèdicales. *Les cahiers de l'analyse des donnèes*, vol. 15, n. 3, 1980.
Bernard G., Ecriture d'un algorithme en language Genstat: example de l'analyse des correspondances. in First International Symposium on data analysis and Informatics, INRIA, 1977.
Banfield C.F., Gower J.C., Macros in Genstat with special reference to multivariate analysis. in COMPSTAT 78 Phisica Verlag 1978.
Calvaruso A., Introduzione al linguaggio Genstat e alle sue applicazioni. Centro di Calcolo Elettronico Interfacoltà della Università di Napoli, 1983.
D'Ambra L., Lauro N., Analisi in componenti principali in rapporto ad un sottospazio di riferimento. Rivista di Statistica Applicata, vol. 15, n. 1, 1982.
Goodman L.A., Kruskal W.K., Measure of association for cross-classification. J. Amer Statist. Assoc, V 49, 1954.
Gray L.N., Williams J.S. Godman and Kruskal's taub: multiple and partial analogs. Proc. Soc. Statist. Section of the Amer. Statist. Assoc. 1975.
Lauro N. D'Ambra L., L'analyse non symétrique des correspondance, Third international Symposium on Data Analysis and Informatics, North Holland 1984.
Lauro N. Serio G., Criteria for evaluating and comparing statistical software: A multidimensional data analysis approach. Statistical Software Newsletter, n. 3, 1982.
Victor N., Computational Statistics: strumento o scienza? in "Atti della XXXII Riunione Scientifica della Società Italiana di Statistica", vol II, Liguori, 1984.

**Summary**

The Authors discuss in this paper the role of metalanguages like Genstat as a vehicle of the modern statistical thinking as well as a significant tool of the methodological progress.

Main statistical and computing features of Genstat are presented with special reference to their performances in multidimensional data analysis.

In this context an algorithm for non symmetrical correspondence analysis is formalized to give and idea of Genstat capability, conciseness, flexibility and last but not least its easy handling for the statistical user not very experienced in informatics.