

REVUE DE STATISTIQUE APPLIQUÉE

J.-P. MAILLES

D. MAILLES

Utilisation d'un langage conversationnel en statistique : exemple d'un programme APL d'analyse des données

Revue de statistique appliquée, tome 22, n° 1 (1974), p. 57-65

http://www.numdam.org/item?id=RSA_1974__22_1_57_0

© Société française de statistique, 1974, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

UTILISATION D'UN LANGAGE CONVERSATIONNEL EN STATISTIQUE EXEMPLE D'UN PROGRAMME APL D'ANALYSE DES DONNÉES

J.-P. MAILLES ⁽¹⁾ et D. MAILLES ⁽²⁾

APL (A Programming Language) devant jouer en statistique dans les années qui viennent un rôle de plus en plus important, aussi bien sur le plan de l'enseignement que du traitement des données, nous avons pensé qu'il était bon, dans la Revue de Statistique Appliquée, de faire sentir sur un exemple l'intérêt de ce langage conversationnel.

L'Institut de Statistiques de l'Université de Paris (I.S.U.P.) s'est résolument orienté vers l'utilisation de ce langage. Cette année, des travaux pratiques, tant au CS2 (cycle supérieur 2^{ème} année) qu'au CSA (cycle de statistique appliquée), sont menés avec le soutien d'une console reliée par ligne téléphonique au 360/75 IBM du C.E.A. (SACLAY) ; des séminaires utilisant APL sont organisés par le Bureau Universitaire de Recherche Opérationnelle (B.U.R.O.) dans le cadre de la formation permanente.

De plus, de nombreux utilisateurs venant de différents horizons (Institut de Programmation, CEPREMAP, E.D.F., ...) sont venus sur place s'initier au langage, analyser des données, mettre au point des algorithmes (*).

1 – OBJECTIFS D'UN PROGRAMME EN APL UTILISANT LES RES-SOURCES DU CONVERSATIONNEL

Le programme APATRD est un exemple de programme d'analyse de données en conversationnel.

Il peut être utilisé :

– *pour le traitement de tableaux de dimensions modestes*

Les dimensions sont dictées par la taille de l'espace résiduel disponible pour stocker les données (le programme occupe 45 K-octets). Ainsi, sur le 360/75, le nombre maximum de données à introduire ne doit pas dépasser 700.

(*) Dans la mesure du temps disponible, les consoles de l'I.S.U.P. peuvent être mises à la disposition d'utilisateurs extérieurs.

(1) Université Paris I.

(2) Bureau Universitaire de Recherche opérationnelle (ISUP)

— pour enseigner un chapitre particulier de l'analyse des données

L'analyse en composantes principales

1.1. — Traitement de tableaux de petit gabarit en conversationnel

Le programme APATRD est tellement facile et rapide à mettre en œuvre que le temps gagné sur un système "batch" dans le traitement de petits tableaux justifie déjà son existence ; ce programme a de plus les avantages du conversationnel, ces avantages pouvant être encore largement développés : on retrouve en utilisant APATRD le cheminement séquentiel des statisticiens de l'école classique, tout traitement élémentaire pouvant être repris à chaque instant (on change le pas d'un histogramme, on élimine une observation ou une variable, etc. . .).

L'utilisateur avant de traiter un ensemble vaste de données a intérêt à mettre au point avec APATRD le traitement qu'il préconise sur une partie de ces données ; il peut par exemple, avant le traitement final, ayant regardé l'allure des distributions, effectuer certaines transformations sur les variables considérées.

1.2 — Enseignement

A l'heure actuelle le statisticien, concurrencé sur son terrain par l'informaticien utilisateur des programmes de statistique doit, s'il veut sauver sa profession plus à base de raisonnement que de recettes, connaître toutes les ressources qu'on peut tirer des moyens informatiques.

L'étudiant, mis directement en contact avec ces moyens au cours de ses études, aura toutes les chances d'assimiler sinon d'entrevoir ces ressources.

L'ordinateur, en ce qui concerne APATRD, est utilisé non seulement pour permettre à l'étudiant d'appliquer sur un exemple qu'il peut choisir une technique qui lui a été enseignée, mais aussi comme un moyen "expérimental" : ainsi avec APATRD il prendra conscience :

- de l'importance du choix des métriques (il peut facilement en essayer plusieurs sur un même tableau de données)
- de l'arbitraire dans le choix des variables et des unités de mesure
- de l'existence de certains effets (effet taille. . .)
- de ce que donne l'analyse sur des cas particuliers
- de la stabilité de certains résultats, etc. . . .

Mais l'étudiant ne doit pas être uniquement confronté à des produits finis comme APATRD ; au cours de l'année, on lui donnera l'occasion d'écrire lui-même certaines fonctions qui, si elles sont performantes et utiles, seront introduites en bibliothèque.

- - - - -

Nous remercions Melle A. SAUVE (C.I.S.I.) pour l'aide qu'elle a apportée à la réalisation de ce programme.

2 – COMMENT UTILISER APATRD ?

La spécification du programme est stockée dans la fonction APATRDHOW qui est à la disposition de l'utilisateur. Celui-ci tape APATRDHOW ; la console répond en lui donnant le mode d'emploi.

APATRDHOW

ANALYSE EN COMPOSANTES PRINCIPALES

APATRD

ENTRE LE 1/6/73

APATRD EST UN ENSEMBLE DE FONCTIONS POUR LE TRAITEMENT DE DONNEES AVEC ANALYSE EN COMPOSANTES PRINCIPALES.

LES OPERATIONS EFFECTUEES SONT LES SUIVANTES :

- 1) ENTREE DES DONNEES
- 2) IMPRESSION DES DONNEES
- 3) CARACTERISTIQUES DES VARIABLES
- 4) HISTOGRAMMES DES VARIABLES
- 5) CALCUL ET IMPRESSION DES MATRICES DE VARIANCE-COVARIANCE ET DE CORRELATION
- 6) ANALYSE EN COMPOSANTE PRINCIPALE AVEC CHOIX DE LA METRIQUE, REPRESENTATIONS GRAPHIQUES ET QUALITE DE LA REPRESENTATION

POUR UTILISER APATRD, REPONDEZ AUX QUESTIONS PAR OUI OU NON.

IL VOUS EST POSSIBLE D'EFFECTUER PLUSIEURS TRAITEMENTS EN SUPPRIMANT EVENTUELLEMENT DES VARIABLES OU DES INDIVIDUS.

LES DONNEES SONT STOCKEES DANS DON :

DON[I ; J] : MESURE DU CARACTERE J SUR L'INDIVIDU I.

CE TABLEAU EST DETRUIT PAR LE PROGRAMME.

LE TABLEAU DES DONNEES PEUT ETRE SOIT ENTRE PENDANT L'EXECUTION D'APATRD, SOIT PREALABLEMENT AFFECTE A LA VARIABLE DON.

3 – UN EXEMPLE

Les données ont été extraites de l'exemple "Les Poissons d'AMIARD" traité dans le chapitre X de l'ouvrage cité en référence (cf. référence 2).

Le dialogue entre l'utilisateur et l'ordinateur est entièrement reproduit.

APATRD

ANALYSE EN COMPOSANTES PRINCIPALES

=====

ENTREE DES DONNEES

DON EST LE TABLEAU DES DONNEES QUANTITATIVES , EST-IL EN MEMOIRE SOUS CE NOM?
OUI

DON EST UN TABLEAU DE 23 INDIVIDUS DECRITS PAR 6 CARACTERES.

ENTREZ LES IDENTIFICATEURS DES INDIVIDUS

23 IDENTIFICATEURS DE 4 CARACTERES MAXIMUM SEPARES PAR UN BLANC AU MOINS,ENTRE ' '
□:

'I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 I13 I14 I15 I16 I18 I19 I20 I21'
IL RESTE 3 IDENTIFICATEURS A ENTRER

□:

'I22 I23 I24'

ETES-VOUS SATISFAIT ?

OUI

ENTREZ LES IDENTIFICATEURS DES VARIABLES

6 IDENTIFICATEURS DE 4 CARACTERES MAXIMUM SEPARES PAR UN BLANC AU MOINS,ENTRE ' '

□:

'V1 V2 V3 V4 V5 V6'

ETES-VOUS SATISFAIT ?

NON

6 IDENTIFICATEURS DE 4 CARACTERES MAXIMUM SEPARES PAR UN BLANC AU MOINS,ENTRE ' '

□:

'V5 V6 V9 V10 V11 V12'

ETES-VOUS SATISFAIT ?

OUI

TABLEAU DES DONNEES.

DESIREZ VOUS L'IMPRESSION DU TABLEAU DES DONNEES ?

OUI

	V5	V6	V9	V10	V11	V12
I1	7.000E00	7.600E01	1.000E00	1.320E02	2.140E02	1.970E02
I2	2.900E01	1.130E02	2.000E00	1.220E02	2.200E02	1.980E02
I3	1.100E01	1.920E02	2.000E00	1.290E02	2.200E02	1.980E02
I4	8.000E00	3.100E02	2.000E00	1.330E02	2.250E02	1.990E02
I5	1.400E01	2.890E02	1.000E00	5.700E01	1.680E02	1.490E02
I6	1.700E01	1.150E02	1.000E00	5.900E01	1.780E02	1.600E02
I7	1.600E01	1.000E02	1.000E00	5.900E01	1.760E02	1.560E02
I8	2.000E01	1.060E02	4.000E00	4.700E01	1.760E02	1.650E02
I9	4.200E01	1.920E02	3.000E00	7.200E01	1.820E02	1.640E02
I10	4.900E01	2.290E02	5.000E00	7.900E01	2.000E02	1.790E02
I11	2.200E01	5.900E02	2.000E00	8.000E01	1.850E02	1.630E02
I12	2.100E01	3.090E02	5.000E00	7.200E01	1.750E02	1.580E02
I13	3.400E01	5.230E02	1.100E01	7.500E01	1.890E02	1.690E02
I14	4.900E01	4.590E02	2.000E00	5.200E01	1.640E02	1.470E02
I15	2.000E01	3.180E02	4.000E00	8.600E01	1.950E02	1.750E02
I16	3.100E01	1.150E02	6.000E00	8.700E01	2.100E02	1.700E02
I18	3.600E01	1.070E02	3.000E00	7.200E01	1.810E02	1.640E02
I19	2.500E01	8.840E02	2.000E00	6.300E01	1.750E02	1.600E02
I20	7.300E01	1.090E02	1.100E01	4.900E01	1.700E02	1.540E02
I21	2.300E01	9.900E01	2.000E00	1.070E02	2.040E02	1.850E02
I22	1.200E01	1.020E02	3.000E00	8.300E01	1.900E02	1.760E02
I23	5.100E01	1.031E03	2.000E00	8.200E01	1.940E02	1.680E02
I24	1.600E01	1.090E02	1.000E00	9.100E01	1.900E02	1.720E02

Y-A-T-IL DES ERREURS DANS VOTRE TABLEAU?

NON

CONTINUE?

OUI

POIDS DES INDIVIDUS.

LES INDIVIDUS ONT-ILS TOUS LE MEME POIDS ?

SI OUI TAPEZ 0 , SINON TAPEZ LE VECTEUR DES POIDS (23 NOMBRES).

□:

0

CARACTERISTIQUES DES VARIABLES.

CARACTERISTIQUES DES VARIABLES?

OUI

NOM	MOY	VAR	E-T	MAX	MIN
V5	2.722E01	2.594E02	1.611E01	7.300E01	7.000E00
V6	2.816E02	6.421E04	2.534E02	1.031E03	7.600E01
V9	3.304E00	7.516E00	2.742E00	1.100E01	1.000E00
V10	8.209E01	6.643E02	2.577E01	1.330E02	4.700E01
V11	1.905E02	3.065E02	1.751E01	2.250E02	1.640E02
V12	1.707E02	2.354E02	1.534E01	1.990E02	1.470E02

HISTOGRAMMES DES VARIABLES.

DESIREZ-VOUS LES HISTOGRAMMES DES VARIABLES?

OUI

HISTOGRAMMES DE TOUTES LES VARIABLES ?

NON

TAPEZ LES NUMEROS DES VARIABLES DONT VOUS DESIREZ LES HISTOGRAMMES.

□:

2 1
6

VARIABLE V6

ETENDUE 955.000 NB DE CLASSES PROPOSE 4

DESIREZ-VOUS CHANGER LE NB DE CLASSES?

SI NON TAPEZ 0,SI OUI TAPEZ LE NB DE CLASSES.

□:

5

MAX 1031.00

MIN 76.00

PAS 191.00

14		I24				
13		I22				
12		I21				
11		I20				
10		I18				
9		I16				
8		I10				
7		I9				
6		I8				
5		I7				
4		I6	I15			
3		I3	I12	I14		
2		I2	I5	I13	I23	
1		I1	I4	I11	I19	

CLASSE		1	2	3	4	5
FREQ.		14	4	3	0	2
CUMUL		14	18	21	21	23

DESIREZ-VOUS UN AUTRE HISTOGRAMME DE CETTE VARIABLE?

NON

VARIABLE V12

ETENDUE 52.000 NB DE CLASSES PROPOSE 4

DESIREZ-VOUS CHANGER LE NB DE CLASSES?

SI NON TAPEZ 0,SI OUI TAPEZ LE NB DE CLASSES.

□:

6

MAX 199.00

MIN 147.00

PAS 8.67

7			I19			
6			I18			
5			I12	I24		
4			I11	I23		I4
3		I20	I9	I16	I22	I3
2		I14	I7	I13	I15	I2
1		I5	I6	I8	I10	I21
						I1

CLASSE		1	2	3	4	5	6
FREQ.		3	7	5	3	1	4
CUMUL		3	10	15	18	19	23

NON

SI NON	TAPEZ 0
NORMALITE	TAPEZ 1
LOG-NORMALITE	TAPEZ 2

0

QUI

CHOIX DE LA METRIQUE.

□:

2

NON

OUT

MATRICE DE CORRELATION.

1.00	.24	-.57	-.44	-.33	-.40
.24	1.00	-.01	-.16	-.18	-.24
-.57	-.01	1.00	-.28	-.14	-.18
-.44	-.16	-.28	1.00	.94	.94
-.33	-.18	-.14	.94	1.00	.95
-.40	-.24	-.18	.94	.95	1.00

ANALYSE EN COMPOSANTES PRINCIPALES.

□:

4

ETUDE DE LA MATRICE DE CORRELATION

TRACE DE LA MATRICE DE CORRELATION: 6.0000E00

	VALEUR PROPRE	POURCENTAGE D'INERTIE	CUMUL
1	0.9768	97.68%	97.68%
2	0.0232	2.32%	100%

1	3.288	.548	.548
2	1.277	.213	.761
3	.977	.163	.924
4	.364	.061	.984

DESIREZ-VOUS TOUJOURS CALCULER 4 COMPOSANTES PRINCIPALES ?
SI OUI TAPEZ 0 SINON TAPEZ LE NOUVEAU NOMBRE DE COMPOSANTES PRINCIPALES
DESIREES.

3

NON

DESIREZ-VOUS L'IMPRESSION DES CORRELATIONS ENTRE LES VARIABLES INITIALES ET LES COMPOSANTES PRINCIPALES ?

QUI

CORRELATIONS ENTRE LES VARIABLES INITIALES ET LES COMPOSANTES PRINCIPALES.

V5	.614	-.637	.138
V6	.300	-.012	-.942
V9	.397	-.817	.182
V10	-.956	-.169	.147
V11	-.923	-.325	.114
V12	-.947	-.265	.052

DESIREZ-VOUS L'IMPRESSION DES CORRELATIONS MULTIPLES ?
OUI

CORRELATIONS MULTIPLES ENTRE LES VARIABLES INITIALES ET LES COMPOSANTES PRINCIPALES.

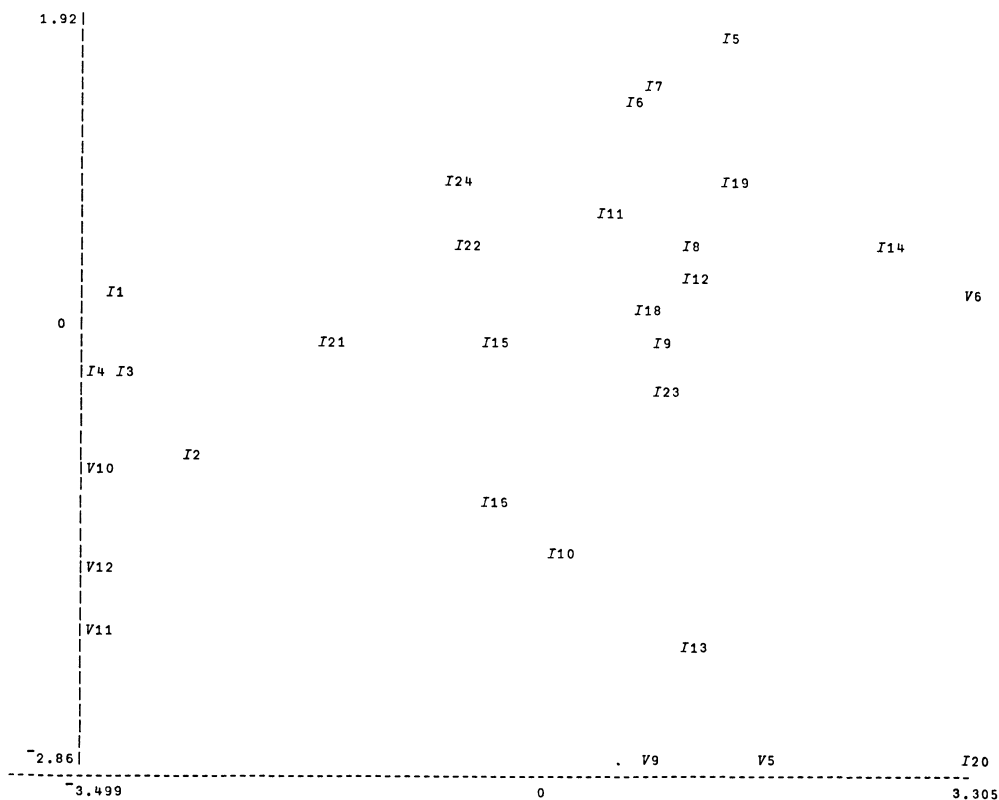
V5	.614	.885	.895
V6	.300	.300	.989
V9	.397	.908	.926
V10	.956	.971	.982
V11	.923	.978	.985
V12	.947	.983	.985

DESIREZ-VOUS L'IMPRESSION DES COSINUS DES ANGLES DES AXES INITIAUX ?
NON

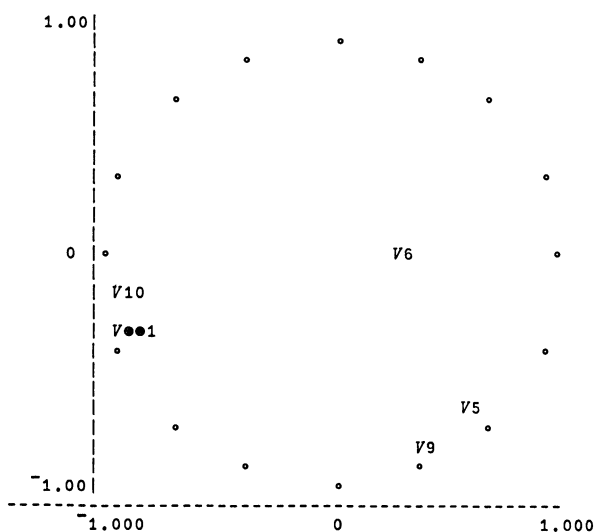
CONTINUEZ ?
OUI

DESIREZ-VOUS LES REPRESENTATIONS GRAPHIQUES ?
OUI

PLAN PRINCIPAL 1-2.



CERCLE DES CORRELATIONS.



DESIREZ-VOUS L'IMPRESSION DES NOUVELLES COORDONNEES DES OBSERVATIONS ?
OUI

COORDONNEES DES OBSERVATIONS DANS LE NOUVEAU SYSTEME D'AXES.

POUR TOUTES LES OBSERVATIONS ?
SI OUI TAPEZ 0, SINON TAPEZ LES NUMEROS DES OBSERVATIONS.

□:

1 10 20

I1	-3.344E00	2.293E-01	-2.600E-01
I10	6.276E-02	-1.477E00	-5.175E-02
I21	-1.702E00	-1.012E-01	-3.539E-01

DESIREZ-VOUS L'IMPRESSION DES COSINUS DES ANGLES ENTRE LES VECTEURS OBSERVATIONS
ET LES SOUS-ESPACES PRINCIPAUX?
OUI

CARRES DES COSINUS DES ANGLES FORMES PAR LES VECTEURS OBSERVATIONS ET LES
DIFFERENTS SOUS-ESPACES PRINCIPAUX.

POUR TOUTES LES OBSERVATIONS ?
SI OUI TAPEZ 0, SINON TAPEZ LES NUMEROS DES OBSERVATIONS.

□:

3 4 18 12

I3	.989	.996	.998
I4	.952	.958	.990
I19	.272	.396	.960
I12	.571	.606	.640

DESIREZ-VOUS REFAIRE UNE ANALYSE?
NON

4 – ANNEXE

Dans le cadre de l'I.S.U.P., le Groupe de Travail sur la programmation APL, animé par M. Valette, centralise les travaux en ce qui concerne les fonctions statistiques APL.

Le listing du programme APATRD peut être obtenu auprès de ce Groupe de Travail.

On donne ici, à titre d'exemple, le listing de l'une des fonctions d'APATRD.

```

      ∇ METRIQUE [M] ∇
      ∇ METRIQUE ; VAR ; I ; M
[1] L2 : LINES 2
[2] ' : CHOIX DE LA METRIQUE :
[3] '
[4] ' '
[5] ' : METRIQUE I                      TAPEZ 1'
[6] ' : METRIQUE D(1/SIGMA 2)          TAPEZ 2'
[7] ' : METRIQUE QUE VOUS FOURNISSEZ  TAPEZ 3'
[8] ICOR ← 0
[9] → (ICOR ≠ 3)/L7
[10] L6 : '
      ENTREZ VOTRE METRIQUE LIGNE PAR LIGNE.'
[11] METR ← (NV, NV) ρ 0
[12] I ← 1
[13] L10 : M ← 0
[14] → ((ρ M) = NV)/L11
[15] ' ERREUR DE DIMENSION RETAPEZ LA LIGNE'
[16] → L10
[17] L11 : METR [I:] ← M
[18] → (NV ≥ I ← I + 1)/L10
[19] L7 : ' '
[20] VAR → P VARPON2 DON
[21] → ('N' = (0, 0 ← ' DESIREZ-VOUS L'IMPRESSION DE LA MATRICE DE COVARIANCE ?') [1])/L30
[22] '
      MATRICE DE VARIANCE-COVARIANCE.
      '
[23] 12 4 EFT VAR
[24] L30 : ' '
[25] → ('N' = (0, 0 ← ' DESIREZ-VOUS L'IMPRESSION DE LA MATRICE DE CORRELATION ?') [1])/L31
[26] '
      MATRICE DE CORRELATION.
      '
[27] 7 2 DFT VARCOR VAR
[28] ' '
[29] L31 : LINES 2
      ∇

```

BIBLIOGRAPHIE

- [1] P.S. ABRAMS, G. LACOURLY – Informatique par téléphone, Hermann.
- [2] F. CAILLIEZ, J.P. MAILLES, J.P. NAKACHE, J.P. PAGES – Analyse des données multidimensionnelles, C 3 E
- [3] Colloque APL, Paris 9-10 septembre 1971, I.R.I.A.
- [4] K.E. IVERSON – A programming Language, John Wiley and Sons, New-York (1962).
- [5] K.W. SMILLIE – An APL Statistical Package, Publication No. 9, University of Alberta, Edmonton, Alberta.