NOTE SUR L'ANALYSE INTERACTIVE DES DONNÉES STATISTIQUES

EXEMPLES D'UTILISATION DU LANGAGE « A P L »

par **Gérard LACOURLY**(*) et **Ludovic LEBART**

Les méthodes d'analyse de données, ou encore de statistique descriptive multidimensionnelle, qui se sont développées sous l'impulsion du Professeur Benzecri, doivent leur existence à l'apparition du calcul automatique, c'est-à-dire à une innovation technologique. Les applications licites de ces méthodes aux sciences humaines permettent de réaliser d'incontestables progrès dans la connaissance des domaines les plus variés, mais la diffusion rapide de ces techniques n'est pas sans poser de problèmes d'information, de communication. C'est pourquoi toute innovation à caractère technique ayant pour effet de rendre plus accessible l'ordinateur, matériellement et psychologiquement doit vraisemblablement avoir des répercussions sur le mode d'utilisation de ces méthodes.

Le langage de programmation conversationnel « A P L », mis au point au cours des années soixante par le Dr Kenneth Iverson constitue une innovation de ce type. A la fois plus puissant et plus simple que la plupart des langages couramment utilisés actuellement, le langage « A P L », qui a déjà des dizaines de milliers d'utilisateurs au Canada et aux États-Unis, permet un dialogue « statisticien-ordinateur » remarquable de clarté et de concision.

Cette note a pour objet la présentation succincte des diverses caractéristiques de ce langage ainsi que des avantages qui résultent de son utilisation à propos de problèmes d'analyse de données. La présentation du langage s'adresse aux lecteurs non initiés, qui doivent en retirer une impression générale leur permettant de situer les traits les plus remarquables d'« A P L » par rapport aux autres langages. Cette présentation ne peut évidemment se substituer à un cours ou à un stage de formation.

(*) CEGOS. Informatique.

Les remarques et les exemples qui suivent sont également destinés au plus grand nombre de lecteurs statisticiens, mais l'annexe intéressera plus spécialement le lecteur déjà un peu familier du langage « A P L », ou alors le lecteur non initié qui aurait consulté entre temps les ouvrages cités en bibliographie.

Avant la présentation proprement dite, donnons quelques caractéristiques pratiques des modalités d'utilisation des systèmes conversationnels du type de ceux qui nous intéressent ici.

L'utilisateur dispose d'un terminal qui peut être branché sur n'importe quel poste téléphonique. Il entre en contact avec l'ordinateur à la suite d'un simple appel. Il frappe alors sur le clavier du terminal les instructions qu'il désire voir exécuter, et donne la parole à l'ordinateur en frappant, par exemple, « retour chariot ». Celui-ci donne une réponse, signale une erreur, ou indique qu'il attend une autre instruction, selon les cas. La réponse de la machine se distingue, sur le papier enregistreur, des instructions de l'utilisateur, par un décalage, ou, pour certaines installations, par une couleur différente. S'il le désire, et s'il a le téléphone, l'utilisateur peut emporter le terminal chez lui pour terminer le travail en cours.

Tout ceci est fort simple, et dans une certaine mesure, démystificateur. Beaucoup d'intermédiaires sont supprimés.

Ce mode de connection avec la machine sera rentable si le langage laisse certaines libertés à l'utilisateur : Celui-ci doit avoir le droit de se tromper de temps en temps, de ne pas avoir tout prévu à l'avance, d'écrire les choses à peu près comme il les pense.

Les qualités du langage « A P L » vont permettre un tel contact direct avec la machine. Le mode conversationnel, par le dialogue et l'autorégulation qu'il implique, possède des avantages psychologiques indéniables : la sanction immédiate de la machine permet un apprentissage extrêmement rapide, qu'il s'agisse au début d'apprendre le langage lui-même, ou plus tard de faire face à des situations variées (mise au point d'algorithmes, analyse de données statistiques). Ces qualités pédagogiques ne sont pas à négliger puisque notre domaine d'application, les sciences humaines et économiques, suppose des contacts permanents entre personnes de formations très différentes. L'utilisation d'un ordinateur comme prolongement d'un calculateur de bureau, sans la mise à contribution de services et de spécialistes parfois mystérieux, est propre à dissiper bien des malentendus.

En bref, toute innovation susceptible de faire tomber des cloisonnements dans un domaine d'activité interdisciplinaire mérite un accueil, ou au moins un regard favorable.

Examinons maintenant quelles sont les opérations pratiques auxquelles se livre le statisticien désireux d'analyser le contenu de quelques tableaux rectangulaires de données. Pour les tableaux de petites dimensions, il se contente de les lire directement, en calculant éventuellement quelques paramètres sur une machine de bureau (moyennes, totaux marginaux, etc...) ; pour les tableaux de dimensions moyennes, il hésite à mettre en branle le service de calcul, ou à programmer et perforer les instructions nécessaires lui-même. Ceci suppose en effet des démarches et des délais qui peuvent être gênants pour la poursuite de l'étude en cours. Les tableaux importants seront peut-être passés de façon plus systématique à « la moulinette », mais les premiers résultats révèleront que tel codage est défectueux, que deux groupes d'individus gagneraient à faire l'objet d'analyses séparées, qu'une erreur s'était glissée dans le fichier. Ceci suppose des aller et retour, des contacts, en un mot, une mobilisation du chercheur sur un travail de programmation ou de gestion qui peut le

détourner pendant un certain temps des problèmes spécifiques de sa discipline.

A ces va-et-vient parfois accidentels s'ajoutent les tâtonnements inhérents aux méthodes de statistique descriptive.

On désire illustrer une représentation par l'adjonction de variables supplémentaires, effectuer une nouvelle transformation préalable des données, vérifier la validité des résultats par une simulation séquentielle qui permet de faire émerger progressivement les réseaux d'associations significatives. Toutes ces opérations supposent les résultats connus à chaque étape ; elles sont grandement facilitées par une connection directe entre le statisticien et l'ordinateur.

I. PRÉSENTATION SOMMAIRE DU LANGAGE A P L

I. 1. CARACTÉRISTIQUES GÉNÉRALES

Mis au point par K. E. Iverson et A. D. Falkoff, le langage de programmation A P L est principalement destiné à l'exécution de calculs scientifiques en mode conversationnel.

Une des caractéristiques fondamentales de ce langage est l'existence d'un ensemble complet d'opérateurs sur tableau. L'allocation de ces tableaux est dynamique (il n'est pas besoin de spécifier à l'avance les dimensions maxima de ces tableaux). Le foisonnement des opérateurs a conduit à adopter une règle de priorité qui peut paraître surprenante au prime abord : un opérateur s'appliquera à tout ce qui est à sa droite. Ainsi, $1 \times 2 + 3$ signifiera $1 \times (2 + 3)$. En fait, cette règle apparaît bien vite comme étant très féconde, et peu astreignante.

Le plus simple, pour établir un premier contact avec le langage A P L est de s'installer devant le clavier de la machine à écrire du terminal, et de tenter de donner quelques instructions :

Commençons par l'utilisation en calculateur de bureau ; frappons sur le clavier dont l'image figure ci-dessous :

Figure 1. — Clavier d'un terminal A P L


```
      3.1 × 4
12.4
```

l'ordinateur nous a répondu, avec un léger décalage sur la gauche, afin de nous permettre d'identifier les lignes de questions et les lignes de réponses.
Essayons maintenant :

```
      A←3 5 7
```

Pas de réponse. L'ordinateur s'est contenté d'affecter à une zone mémoire qui s'appellera dorénavant A le vecteur à trois composantes 3 5 7. On notera que les « blancs » jouent un rôle important, et qu'il n'est pas nécessaire d'utiliser d'indice pour définir le vecteur A. Si nous voulons vérifier le contenu de la mémoire A, il suffit de frapper la lettre A :

```
      A
3 5 7
```

On remarque qu'il n'a pas été besoin de spécifier le type de A (entier, réel), ni de spécifier un modèle d'impression. Si nous frappons maintenant :

```
      A+2
5 7 9
```

La réponse nous indique que le nombre 2 a été ajouté à chacune des composantes du vecteur A. Ceci est une règle générale en A P L : les opérateurs dyadiques tels que l'addition se généralisent aux tableaux de la façon suivante : si les tableaux ont les mêmes dimensions, l'opérateur agit sur les couples d'éléments terme à terme. Si l'on a un tableau et un scalaire, comme dans notre exemple, l'opérateur agit pour chaque couple formé d'un élément du tableau et du scalaire. Enfin, si l'on a affaire à deux tableaux de dimensions différentes, une erreur est signalée, sans exécution.

Continuons nos essais :

```
      ρA
3
```

Nous avons demandé, à l'aide du symbole ρ, la dimension du tableau A. Cette dimension peut être un vecteur à deux composantes si A est une matrice ou à n composantes si A est un tableau à n dimensions.

```
      +/A
15
```

L'opérateur / appelé « réduction », placé derrière un signe d'opération quelconque correspond à l'insertion de ce signe entre tous les éléments du tableau.

```
      (+/A)÷ρA
5
```

On a donc effectué : $3 + 5 + 7 = 15$. Le signe ÷ étant celui de la division, nous avons divisé la somme des éléments de A par le nombre de ses éléments. Ces instructions simples nous permettent de calculer la moyenne arithmétique des éléments d'un tableau sans même connaître le nombre d'éléments de ce tableau. Construisons une fonction qui calcule cette moyenne.

```
      ∇ MOYENNE A
[1]   (+/A)÷ρA
[2]   ∇
```

Lorsque nous avons frappé le triangle à l'envers, le système a été averti de notre intention de construire une fonction. Il a numéroté lui même les instructions en marge, et s'arrête lorsqu'apparaît un second triangle.

Essayons le programme *MOYENNE* :

```
    MOYENNE 2 3 4 5 6
4
```

Il est certes plus long d'écrire le mot *MOYENNE* que d'écrire les instructions de calcul nécessaires. Indépendamment de l'intérêt pédagogique de cet exemple, l'utilisateur statisticien aura quand même avantage à construire de cette façon un sous-langage adapté à sa sphère d'intérêt. Parmi les opérateurs scalaires dyadiques, il y a, bien sûr, tous ceux existant dans les autres langages (addition, soustraction, division, multiplication, exponentiation, logarithme dans une base donnée, résidu de la division de *B* par *C*, opérations logiques : et, ou, identité, etc...), l'opérateur « plafond » qui donne le plus grand (*A* Γ *B*) de deux éléments *A* et *B*, l'opérateur « plancher », qui donne le plus petit (*A* L *B*) de deux éléments *A* et *B*.

Si *A* et *B* sont des matrices de mêmes dimensions, avec la règle énoncée plus haut, *A* × *B* désigne le produit contracté (*terme à terme*) des matrices *A* et *B*. Le produit matriciel ordinaire existe, lorsque les dimensions des matrices mises en jeu s'y prêtent. On le note : *A* + . × *B*.

Ce produit peut être généralisé, en remplaçant les signes + et × par n'importe quels signes, pourvus que ceux-ci soient relatifs à des opérateurs dyadiques. Ainsi, le produit *A*Γ . × *B* nous donne une matrice dont le terme général est le plus grand élément des produits terme à terme de chaque ligne de *A* par chaque colonne de *B*.

Indépendamment des produits contractés, du produit matriciel généralisé, il existe un produit « extérieur », pour deux matrices de dimensions quelconques, qui effectue une opération dyadique quelconque entre tout élément de *A* et tout élément de *B*. Si l'opération est ×, on obtient des produits tensoriels.

On le note dans ce cas °. ×. On peut construire de la même façon °.+, °.Γ, etc... Par exemple (1 2)°.+ (3 5) donne la matrice :

```
4 6
5 7
```

D'autres opérations sur tableaux sont également cablées :

Ainsi, l'opérateur ⍋ appliqué à un vecteur donne les rangs des éléments de ce vecteur, lorsque ceux-ci sont classés par ordre croissant.

L'inverse de la matrice A se note ⌹ A, le tirage de 100 nombre au hasard entre 1 et 10 000 se note 100 ? 10 000.

Ces quelques informations suffisent à donner une idée, encore vague certes, des possibilités du langage A P L. Ses caractéristiques sont particulièrement favorables aux calculs statistiques usuels, aux calculs portant sur les rangs, à toutes les procédures utilisées en algèbre linéaire. De toute façon, le système de construction des fonctions est tellement souple que le langage s'adapte facilement ; néanmoins, les opérateurs et fonctions « cablés » sont extrêmement performants.

I. 2. MODALITÉS D'IMPLANTATION DU LANGAGE A P L

A P L est réalisé dans un environnement de temps partagé, c'est-à-dire que l'utilisateur est en liaison directe avec l'ordinateur, et bien que n'étant pas seul dans ce cas, a l'impression de disposer de ce dernier continûment et en totalité. A toute question et demande d'exécution de programme, l'ordinateur répond instantanément, sauf évidemment si le calcul demandé mobilise l'unité centrale pendant un temps important.

Les calculs, les définitions de fonctions, les enregistrements de données se font dans une zone appelée « Espace actif de travail » (la taille de cette zone est variable selon les installations). L'utilisateur dispose de plusieurs espaces de travail stockés sur des disques avec lesquels il peut communiquer. Il peut à tout moment copier son espace de travail actuel sur disque, afin de préserver ses résultats, charger un autre espace de travail, copier une fonction ou un fichier situé sur un autre espace de travail, etc...

Bien que les transactions entre espaces de travail soient très simple il n'en existe pas moins une limite en taille. En effet, on ne peut créer ou copier dans l'espace actif de travail un tableau de taille supérieure à celle de cet espace actif. C'est alors qu'il est nécessaire de disposer d'un système de fichiers. Le système de fichier doit être simple et indépendant de la machine et de ses ressources. C'est pourquoi la logique du système doit être aussi simple que son utilisation. Pour l'utilisateur, un fichier est un ensemble de composants, chacun d'eux pouvant être une entité A P L, c'est-à-dire un scalaire, un tableau à *n* dimensions, une chaîne de caractères. Pour avoir accès au composant 3 du fichier appelé par exemple LOISIR, il suffit de frapper : LIRE LOISIR 3.

Pour modifier ou ajouter des éléments il existe des instructions aussi simples. Pour utiliser des fichiers volumineux, il suffit donc d'utiliser des fonctions de gestion de fichier, comme d'autres fonctions A P L.

II. EXEMPLE PRATIQUE D'ANALYSE DE TABLEAU STATISTIQUE

Nous donnons ci-dessous un exemple pratique d'introduction de données, d'appel de sous-programme, enfin d'édition de résultats sur la machine à écrire du terminal. Les programmes utilisés figurent dans leur intégralité en annexe, les données analysées faisant l'objet d'un article beaucoup plus détaillé et documenté dans le même numéro de cette revue, nous nous limiterons à des considérations méthodologiques, sans insister sur l'interprétation économique des représentations obtenues.

Une petite fonction, construite pour l'occasion, nous permet de charger le tableau de données à partir de la machine à écrire, sans grand risque d'erreur (cf. annexe I).

Le tableau, de dimension 15 × 31, nous donne pour les 31 régions d'Europe précédemment étudiées les valeurs de 15 consommations par ménage ouvrier.

Il nous faut introduire ensuite des identificateurs de chacune de ces consommations et de chacune de ces régions : nous chargeons alors un tableau de dimension 5 × 46 caractérisant par 5 lettres les (31 + 15) individus et variables.

Nous choisissons des abréviations pour les régions, faisant précéder les régions allemandes du symbole A* et les régions italiennes du symbole I*.

La liste des variables construites, et des caractères qui les représentent figure ci-dessous :

1) Pain, céréales (*PAIN*).

2) Viandes, Poissons (*VIAND*).

3) Produits laitiers, Corps gras (*LAIT*).

4) Fruits, Légumes, Pommes de terre (*FRUIT*).

5) Produits d'épicerie (*EPICE*).

6) Boissons, Tabac, Repas pris à l'extérieur (*BOISS*).

7) Vêtement, chaussures hommes (*VETEM*).

8) Vêtement, chaussures femmes (*VETEF*).

9) Réparations habillement, soins santé (*SOINS*).

10) Loyers et charges (*LOYER*),

11) Énergie (*ENERG*).

12) Ameublement, Équipement (*MEUBL*).

13) Textiles d'ameublement, articles ménagers, entretien (*MENAG*).

14) Transports (*TRANS*).

15) Éducation, loisirs et divers (*EDUCA*).

Ce tableau étant homogène (l'addition des éléments d'une ligne ou d'une colonne ayant un sens), nous désirons en obtenir une représentation par l'analyse factorielle des correspondances.

Si le tableau des données s'appelle *TAB*, il suffit de frapper (*cf.* annexe III):

2 AFCOR TAB

L'impression des résultats numériques commence aussitôt :

TRACE DE LA MATRICE S 0.04181814

VALEUR PROPRE		*POURCENTAGE DE LA TRACE*
1	0.016	38.8
2	0.011	27.3

VALEURS DES FACTEURS FI

... etc...

Si l'on désire visualiser le plan des deux premiers facteurs, on appelle : (*DES* désignant le tableau des identificateurs, *cf.* annexe VI).

(⍉F. FJ) PLANF1F2 DES

La figure II est une reproduction du graphique qui apparaît alors sur l'imprimante.

Pour obtenir une analyse de la matrice des corrélations des 15 variables, on appelle maintenant (*cf.* annexe IV) :

2 ACOMP TAB

On obtient le tableau de résultats numériques reproduit sur la figure III. On pourrait, bien sûr, obtenir une représentation graphique analogue à la précédente.

Pour obtenir un classement des variables selon les valeurs croissantes du premier facteur, il suffit d'indexer le tableau d'identificateurs par les indices rangés à l'aide de l'opérateur ⍋ ;

Il en est de même si nous désirons le classement des régions selon le premier axe factoriel (les identificateurs des régions sont décalés de 15 places).

Figure 2. — Exemple de graphique sur clavier de terminal.

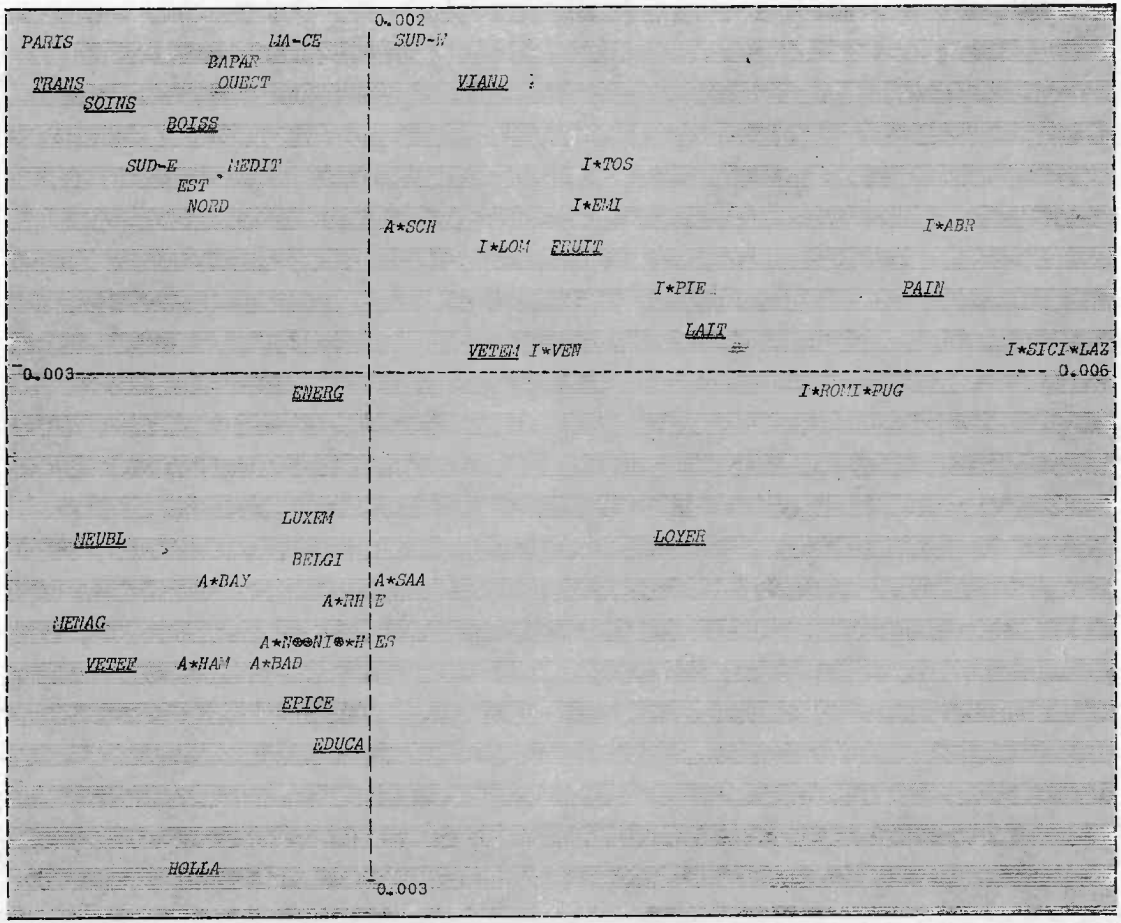

Figure 3. — Exemple d'édition de résultats d'analyse en composantes principales sur clavier de terminal.

2 ACOMP OUV

MATRICE DE CORRELATION

	PAIN	VIAND	LAIT	FRUIT	EPICE	BOISS	VETEM	VETEF	SOINS	LOYER	ENERG	MEUBL	MENAG	TRANS	EDUCA
PAIN	1.000	0.036	0.084	0.094	-0.608	-0.292	-0.143	-0.563	-0.257	0.003	-0.396	-0.571	-0.668	-0.281	-0.452
VIAND	0.036	1.000	0.407	0.678	-0.203	0.780	0.581	0.257	0.784	0.252	0.520	0.335	0.206	0.774	0.096
LAIT	0.084	0.407	1.000	0.427	-0.121	0.149	0.477	0.010	0.193	0.188	0.390	-0.036	-0.231	0.190	-0.178
FRUIT	0.094	0.678	0.427	1.000	0.021	0.576	0.558	0.316	0.664	0.431	0.554	0.290	0.246	0.680	0.083
EPICE	-0.608	-0.203	-0.121	0.021	1.000	0.185	0.188	0.745	0.223	0.116	0.481	0.671	0.776	0.285	0.418
BOISS	-0.292	0.780	0.149	0.576	0.185	1.000	0.430	0.467	0.895	0.087	0.601	0.609	0.580	0.854	0.354
VETEM	-0.143	0.581	0.477	0.558	0.188	0.430	1.000	0.686	0.592	0.406	0.687	0.503	0.397	0.604	0.322
VETEF	-0.563	0.257	0.010	0.316	0.745	0.467	0.686	1.000	0.545	0.360	0.652	0.829	0.858	0.596	0.585
SOINS	-0.257	0.784	0.193	0.664	0.223	0.895	0.592	0.545	1.000	0.106	0.756	0.711	0.629	0.957	0.360
LOYER	0.003	0.252	0.188	0.431	0.116	0.087	0.406	0.360	0.106	1.000	0.286	0.171	0.145	0.194	0.285
ENERG	-0.396	0.520	0.390	0.554	0.481	0.601	0.687	0.652	0.756	0.286	1.000	0.685	0.625	0.710	0.400
MEUBL	-0.571	0.335	-0.036	0.290	0.671	0.609	0.503	0.829	0.711	0.171	0.685	1.000	0.871	0.723	0.578
MENAG	-0.668	0.206	-0.231	0.246	0.776	0.580	0.397	0.858	0.629	0.145	0.625	0.871	1.000	0.639	0.617
TRANS	-0.281	0.774	0.190	0.680	0.285	0.854	0.604	0.596	0.957	0.194	0.710	0.723	0.639	1.000	0.319
EDUCA	-0.452	0.096	-0.178	0.083	0.418	0.354	0.322	0.585	0.360	0.285	0.400	0.578	0.617	0.319	1.000

TRACE DE LA MATRICE S 15.00000000

	VALEUR PROPRE	POURCENTAGE DE LA TRACE
1	7.390	49.3
2	2.963	19.8

VALEURS DES FACTEURS-VARIABLES

FACTEUR NO 1

0.4931 -0.6357 -0.2037 -0.6142 -0.5358 -0.8023 -0.7244 -0.8378 -0.8868 -0.3328 -0.8544 -0.8656 -0.8236 -0.8936 -0.5602

FACTEUR NO 2

0.6007 0.6644 0.6077 0.5865 -0.6781 0.2431 0.2936 -0.3628 0.2696 0.1620 0.0862 -0.3342 -0.5025 0.2488 -0.4224

VALEURS DES FACTEURS-OBSERVATIONS

FACTEUR NO 1

0.3431 -0.2861 0.1794 -0.1565 0.1136 0.1146 0.0789 0.1623 -0.0891 -0.3476 -1.6078 -0.5792 -0.8172 -0.8506 -0.0235
-0.1588 -0.0885 -0.7823 -0.5976 -0.7524 0.3604 0.0632 0.4852 0.6509 0.5868 1.3409 1.3536 1.1104 1.2455
0.1242 -1.1757

FACTEUR NO 2

-0.3284 -0.7282 -0.6400 -0.5706 -0.5543 -0.5386 -0.6877 -0.7295 -0.2758 -0.8141 0.5586 0.3935 0.2339 0.3066 0.0995
0.3248 0.5685 0.1936 0.3015 -0.1435 0.3729 0.3624 0.0681 0.1514 0.2594 0.2899 0.1863 0.2429 0.3270
0.7440 0.0197

Les instructions et les résultats obtenus figurent ci-dessous.

On remarque que de telles éditions complémentaires facilitent grandement l'interprétation des directions principales des nuages analysés.

On peut, tout aussi facilement, obtenir des histogrammes recommencer l'analyse en supprimant un pays ou un groupe de produits :

```
      DES[ ⍋FI[1 ;] ;]
TRANS
SOINS
MEUBL
ENERG
VETEF
MENAG
BOISS
VETEM
VIAND
FRUIT
EDUCA
EPICE
LOYER
LAIT
PAIN
      DES [15 + ⍋FJ [1 ;] ;]
PARIS
LUXEM
EST
NORD
SUD-E
BELGI
MEDIT
BAPAR
HOLLA
A* HAM
MA-CE
A* NOE
A* SAA
SUD-W
OUEST
I* LOM
A* BAD
A* HES
A* RHE
I* ROM
A* BAY
A* NIE
A* SCH
I* PIE
```

** VEN*
** TOS*
** EMI*
I PUG*
I SIC*
I LAZ*
I ABR*

Remarques

On peut faire plusieurs remarques à la suite de l'exemple ci-dessus.

1) Le terminal pouvait sembler être un goulot d'étranglement de l'information. En fait, les volumineux listages de résultats que l'on obtient lors des exploitations traditionnelles n'étaient justifiés que parce qu'il fallait tout prévoir à l'avance. En mode conversationnel, on édite seulement ce dont on a besoin, au fur et à mesure de ces besoins.

2) La contrainte qui sera le plus rapidement ressentie par le statisticien sera vraisemblablement la taille des espaces de travail. Dans les installations actuelles, on ne dispose souvent que de 32 K octets. Ceci permet de faire des calculs très importants, car l'occupation de cet espace est gérée « en temps réel » et il est toujours possible de translater ce qui n'est pas immédiatement utile sur d'autres espaces de travail. Cependant, certaines opérations relativement élémentaires, telles que les diagonalisations de matrice, sont grandement facilitées par la mobilisation d'un gros volume de mémoire rapide. Il faut donc souhaiter que les installateurs mettent à la disposition des utilisateurs des zones-mémoires plus importantes.

3) Pour les dépouillements d'enquêtes, il sera exclu d'introduire les données à partir du terminal, dans l'état actuel de la diffusion de ces terminaux. Les bandes magnétiques peuvent être domiciliées directement à l'ordinateur, puis copiées sur disques sous forme de fichiers directement interprétables et analysables à partir des terminaux.

4) Le contrôle continu du processus de travail donne la possibilité de comparer immédiatement deux méthodes ou techniques, de les combiner éventuellement, enfin de les critiquer. De nombreuses conjectures seront suscitées par ces manipulations.

ANNEXE

LISTAGES DE PROGRAMMES D'ANALYSE DE DONNÉES USUELS

Nous donnons ci-dessous des exemples de listages de programmes d'un emploi courant en analyse des données. On pourra noter que l'on obtient une édition relativement soignée, avec des instructions réduites et simples. Par suite de la règle de priorité des opérateurs, et de la possibilité de faire plusieurs affectations (symbole ←) dans une même ligne de calcul, on aura intérêt à lire les lignes de la droite vers la gauche. Par exemple, le symbole ⍉ désignant la transposition de matrice, la ligne suivante :

$$M \leftarrow S + .xT \leftarrow ⍉S$$

Peut se lire : Je mets la transposée de la matrice *S* dans *T* que je prémultiplie matriciellement par S, le produit obtenu étant mis dans *M* (ou étant baptisé *M*).

Il existe un nombre extrêmement grand de versions A P L d'un même algorithme mathématique. Le caractère synthétique du langage et la possibilité d'affectations répétées dans une même instruction permet d'écrire des versions très condensées de certains programmes, qui sont surtout des exercices de style. Nous préférerons une écriture progressive, plus facile à lire.

De toute façon, il s'agit d'exemples qui ne prétendent pas à l'optimalité.

I. EXEMPLE DE FONCTION DESTINÉE A FACILITER L'INTRODUCTION DE DONNÉES

```
     ∇ ENTREE
[1]    →(15 1 =⍴I←,⍞)/ 4 0
[2]    'VOUS VOUS ETES TROMPE'!'
[3]    →1
[4]    TAB←TAB,[1] I
[5]    →1
     ∇
```

Commentaires : Cette fonction est destinée à permettre un chargement pratique, par le clavier du terminal, du tableau de dimensions 15 × 31 de l'exemple traité plus haut.

Les données sont introduites par séries de 15 (les 15 consommations relatives à une région donnée). L'instruction 1 est un branchement multiple : la paranthèse sera un vecteur à deux composantes prenant les valeurs (1 0) si 15 nombres ont été frappés, (0 1) si un seul nombre a été frappé, (0 0) si le nombre de données frappées est différent de 15 ou de 1. En effet, le vecteur des données frappées, ⍞ est mis dans *I* dont on prend la dimension ⍴*I* que l'on compare logiquement au vecteur (15 1). La flèche → signifie « aller à » et le symbole / utilisé maintenant pour désigner un opérateur dyadique a pour fonction de contracter le vecteur situé à sa droite, selon le vecteur logique de même dimension situé à sa gauche.

On ira donc en 4 si on a bien frappé 15 données, en 0, c'est-à-dire en fin d'exécution, si on a frappé une seule donnée, et en séquence si on s'est

trompé, ce que notre fonction nous signale en clair. En 4, on construit le tableau *TAB* de proche en proche, en le complétant à chaque fois du vecteur qui vient d'être lu, et dont la longueur vient d'être contrôlée.

Le volume du commentaire peut paraître important par rapport au volume et à l'intérêt de la fonction : nous serons plus elliptiques par la suite.

II. EXEMPLE DE PROGRAMME D'EXTRACTION DES PLUS GRANDES VALEURS PROPRES D'UNE MATRICE SYMÉTRIQUE

On trouvera les algorithmes du programme *DIAG* ci-dessous, et de son auxiliaire *HIL* dans (Benzecri, ref. 1).

DIAG extrait les plus grandes valeurs propres d'une matrice symétrique en itérant la transformation linéaire associée, et en orthonormant les vecteurs trouvés à chaque pas, à l'aide de *HIL*.

```
    ∇ Z←CV DIAG S;LV;VI
[1]   VI←S[⍳CV;]
[2]   IT←1
[3]  L1:LV←+/IX×VI←(IX←HIL VI)+.×S
[4]   →L1×⍳(MAXIT≥IT←IT+1)∧ANG<|(H-LV*2)÷H←+/VI*2
[5]   Z←LV,HIL VI
    ∇


    ∇ Z←HIL VI;V;VX
[1]   V←1
[2]   →L3,⍴Z←VI
[3]  L1:VX←Z[⍳V-1;]+.×Z[V;]
[4]   Z[V;]←Z[V;]-VX+.×Z[⍳V-1;]
[5]  L3:Z[V;]←Z[V;]÷(+/Z[V;]*2)*0.5
[6]   →L1×(⍴VI)[1]≥V←1+V
    ∇
```

Commentaires : On désire les *CV* premiers vecteurs propres de la matrice *S*. *MAXIT* désigne le nombre maximum d'itérations désirées, et *ANG* le carré du sinus du plus grand angle de rotation des vecteurs propres au cours de la dernière itération. On peut raisonnablement fixer pour les applications usuelles *MAXIT* à 50 et *ANG* à 10^{-6}.

On notera que les variables qui ne figurent pas sur la ligne du titre de la fonction ne sont pas locales (comme elles le seraient en *FORTRAN* par exemple). En tête de ligne, « *L1* » : ou « *L3* » : permettent d'identifier la ligne, quelle que soit la place de cette ligne dans la fonction, ce qui donne plus de souplesse à l'écriture et au test des programmes.

L'opérateur ⍳ (Générateur d'indice), appliqué à un nombre entier *K*, génère un vecteur dont les *K* composantes sont 1, 2, 3, *K*-1, *K*.

Ainsi, dans *DIAG*, la matrice *VI*, qui sert au « démarrage » de l'itération, est formée des *CV* premières lignes de *S*.

III. EXEMPLE DE PROGRAMME D'ANALYSE DES CORRESPONDANCES

Comme le programme d'analyse en composantes principales que l'on trouvera ci-dessous, *AFCOR* utilise, outre *DIAG*, deux petits programmes auxiliaires d'édition : *EDIFAC* et *EDIPROPRE*, dont la liste est donnée plus loin.

```
    ∇ Z←NF AFCOR P;S;K;PI;PJ
[1]   PI←+/P←P÷K←+/+/P
[2]   PJ←+⌿P
[3]   S←(PI∘.×PJ)*0.5
[4]   P←(P-S)÷S
[5]   Z←NF DIAG S←S+.×⍉S
[6]   (+/ 1 1 ⍉S) EDIPROPRE Z[;1]
[7]   'VALEURS DES FACTEURS FI'
[8]   '──────────────────────'
[9]   EDIFAC FI←(0 1↓Z)×(Z[;1]∘.÷PI)*0.5
[10]  'VALEURS DES FACTEURS FJ'
[11]  '──────────────────────'
[12]  EDIFAC FJ←(FI+.×P)÷(Z[;1]*0.5)∘.×PJ
    ∇
```

Commentaires

L'analyse du tableau de nombres positifs *P* donne lieu à l'extraction de *NF* facteurs. On diagonalise la Matrice *S* « symétrisée », calculée à partir des inerties par rapport au centre de gravité, et non par rapport à l'origine, afin d'éliminer le facteur trivial.

IV. EXEMPLE DE PROGRAMME D'ANALYSE EN COMPOSANTES PRINCIPALES

```
    ∇ Z←NF ACOMP P
[1]   Q←⍴P[;1]
[2]   M←(+/P)÷N←⍴P[1;]
[3]   S←(((+/P*2)÷N)-M*2)*0.5
[4]   P←(P-M∘.×(N⍴1))
[5]   P←P÷S∘.×N⍴1
[6]   Z←NF DIAG COR←(P+.×⍉P)÷N
[7]   'MATRICE DE CORRELATION'
[8]   '──────────────────────'
[9]   ((9⍴' '),¯2↓,M),[1]' ',[1](M←1⍕' ',IDENT,' '), 7 3 DFT COR
[10]  (+/ 1 1 ⍉COR) EDIPROPRE Z[;1]
[11]  'VALEURS DES FACTEURS-VARIABLES'
[12]  '──────────────────────────────'
[13]  EDIFAC FI←(0 1↓Z)×(Z[;1]∘.×Q⍴1)*0.5
[14]  'VALEURS DES FACTEURS-OBSERVATIONS'
[15]  '─────────────────────────────────'
[16]  EDIFAC FJ←(FI+.×P)÷(Z[;1]*0.5)∘.×N⍴Q*0.5
    ∇
```

Commentaires

Le tableau initial *P* contient les variables ou attributs en ligne et les observations ou individus en colonnes. La première instruction désigne par *Q* le nombre de variables (dimension d'un vecteur colonne). La seconde désigne par *N* le nombre d'observations, et calcule les moyennes relatives

à chacune des variables. Le vecteur des écarts-types est calculé en 3, le tableau *P* est centré en 4, réduit en 5 ; la matrice des corrélations *COR*, calculée et diagonalisée en 6, est imprimée en 9. La fonction *DFT*, disponible sur la plupart des installations A P L, permet de fixer le nombre de décimales à l'impression. Pour plus de détail concernant le cadrage variable-observation, on pourra se reporter à réf. 5.

V. EXEMPLES DE PROGRAMMES D'ÉDITION DE RÉSULTATS EN ANALYSE DES DONNÉES

```
     ∇ I EDIPROPRE J;K
[1]    2ρRC;'TRACE DE LA MATRICE S' ; 20 8 DFT I
[2]    K←1
[3]    '        VALEUR PROPRE       POURCENTAGE DE LA TRACE'
[4]   L1:K; 15 3 20 1 DFT J[K],100×J[K]÷I
[5]    →L1×(ρJ)≥K←K+1
     ∇
```

```
     ∇ EDIFAC I;K
[1]    K←1
[2]   L1:'FACTEUR NO ';K
[3]    8 4 DFT I[K;]
[4]    →L1×(ρI)[1]≥K←K+1
     ∇
```

Commentaires : Ces deux petites fonctions, appelées par *AFCOR* et par *ACOMP*, utilisent également la fonction d'édition *DFT*. Pour la présentation des résultats, on pourra se reporter à l'exemple traité plus haut.

VI. EXEMPLES DE PROGRAMMES DE REPRÉSENTATION GRAPHIQUE SUR MACHINE A ÉCRIRE DU TERMINAL

La fonction *PLANF1F2*, qui appelle la fonction *FUSION*, a servi à tracer le graphique de l'exemple ci-dessus. Elle utilise une autre fonction auxiliaire, *D4 I*, qui n'est autre que 6 3 *DFT I*, c'est-à-dire une écriture du tableau *I* avec des nombres de 6 caractères, dont 3 après la virgule.

```
     ∇ TT←XY PLANF1F2 DES;IO;IA;SO;SA;J;T;TT;K
[1]    IO←⌊/XY[;2]
[2]    SO←⌈/XY[;2]
[3]    IA←⌊/XY[;1]
[4]    SA←⌈/XY[;1]
[5]    J←⌈1+(XY[;2]-IO)×(L←1)÷SO-IO
[6]    T←(-⌈1+(XY[;1]-IA)×(C←1)÷SA-IA)⌽DES,((1↑ρDES),C)ρ' '
[7]    TT←(L,(ρT)[2])ρ' '
[8]    I←L
[9]   L1:→(1 0 =ρK←(J=I)/⍳ρJ)/L2,L4
[10]   TT[I;]←FUSION T[K;]
[11]   →L4
[12]  L2:TT[I;]←T[K;]
[13]  L4:→(1≤I←I-1)/L1
[14]   T←⍳0
[15]   TT←⊖TT
[16]   TT←(-J)⌽[1]((D4 IA),(((ρTT)[2]-12)ρ'-'),D4 SA),[1](J←⌈L×SO÷SO-IO)⌽[1] TT
[17]   J←⌈C×-IA÷SA-IA
[18]   TT←(-J)⌽(((ρTT)[2]+1)↑D4 SO),[1]('|',J⌽TT),[1]((ρTT)[2]+1)↑D4 IO
[19]   TT←'_',[1]('|',TT,'|'),[1] '_'
     ∇
```

```
     ∇ Z←FUSION M;I;J;D
[1]    Z←(D←⍴M)[2]⍴' '
[2]    J←D[2]|I←(M≠' ')/⍳⍴M←,M
[3]    Z[J+D[2]×0=J]←M[I]
[4]    Z[J[(1≠+/J∘.=J)/⍳⍴J]]←'⊗'
     ∇
```

Commentaires

Le tableau *DES* contient les identificateurs des points (par exemple, cinq lettres par point, comme dans notre exemple). S'il y a *N* points, *DES* a alors pour dimension (*N*, 5). Les deux colonnes du tableau *XY* sont les valeurs numériques des abscisses et des ordonnées des points du nuage à représenter. Les paramètres *L* et *C*, qui ne sont pas des variables locales, sont les nombres de lignes et de colonnes du graphique désiré. La fonction *FUSION* sert à marquer d'un symbole particulier les points multiples, repérés au cours de l'instruction 9.

Références bibliographiques

[1] **Algorithmes de géométrie euclidienne en analyse des données,** 1970. J.-P. BENZECRI.

[2] **Leçons sur l'analyse statistique des données multidimensionnelles,** 1970. J. P. BENZECRI, L.S.M. (I.S.U.P.), 9 quai Saint-Bernard, Paris (5e).

[3] **A Programming language,** K. E. IVERSON. John Wiley and Son, New York, 1962.

[4] **Une introduction au langage A P L,** P. S. ABRAMS et G. LACOURLY, Hermann, à paraître.

[5] **Statistique et Informatique appliquées,** L. LEBART et J. P. FENELON, Dunod, 1971.

Le directeur de la publication : P. BORDAS.

Dépôt légal : 1er trimestre 1972. Numéro 7405. Imprimé en France.

Imprimerie Nouvelle, Orléans. — N° 6535.

CONSOMMATION (ANNALES DU C.R.E.D.O.C.)

1967

N° 1. — Une étude économétrique de la demande de viande. — La consommation des Français en 1965. — Intégration des méthodes d'approche psycho-sociologiques à l'étude de l'épargne.

N° 2. — Un indicateur de la morbidité appliqué aux données d'une enquête sur la consommation médicale. — La diffusion des services collectifs : phénomène économique ou social ? — Les travaux de préparation du V^e Plan et l'élaboration d'un modèle national de fonctionnement du marché du logement. — Les conditions de vie des familles.

N° 3. — L'épargne des exploitants agricoles. — Structure et équilibre du marché du textile. — Les dépenses touristiques.

N° 4. — L'appareil commercial et les circuits de distribution. — Le développement de la radiologie.

1968

N° 1. — Étude critique de méthodes d'enquête. — Étude sur l'offre et la demande de créance.

N° 2. — Théorie et politique de l'épargne. — Un modèle prévisionnel de la demande de logements. — L'évolution de la consommation de viande.

N° 3. — La consommation et la demande de monnaie. — Valeur prédictive des intentions d'achats au niveau du ménage pris individuellement.

N° 4. — Quelques éléments sur le comportement des propriétaires vis-à-vis du prix du logement acheté et de la mise de fonds versée. — Facteurs « irrationnels » de l'offre d'épargne (recherches allemandes).

1969

N° 1. — L'offre de monnaie par les banques commerciales. — L'économie des services de soins médicaux en France. — L'évolution de la consommation de produits laitiers de 1950 à 1966.

N° 2. — L'économie des services de soins médicaux en France. — La formation de l'épargne liquide (l'exemple du Crédit Mutuel). — Consommation individuelle et consommation collective. — Étude sur la demande en logement des ménages.

N° 3. — Les prix de détail en France par rapport aux autres pays de la Communauté. — La consommation des ménages en France et en Hongrie. — Introduction à l'analyse des données.

N° 4. — Durée d'observation et précision dans les enquêtes de consommation. — Un essai de classification de titres boursiers fondée sur l'analyse factorielle. — Introduction à l'analyse des données.

1970

N° 1. — La fréquentation des équipements collectifs. — La supériorité de la gestion collective de l'épargne mobilière : analyse méthodologique et application aux SICAV. — Le comportement des exploitants agricoles en Eure-et-Loir et en Ille-et-Vilaine.

N° 2-3. — L'Évolution de la consommation des ménages de 1959 à 1968.

N° 4. — Les services médicaux en Suède et en France. — Proposition pour une méthodologie de l'étude de la redistribution. — La consommation des boissons dans quelques pays d'Europe.

1971

N° 1. — Les familles devant l'éducation des enfants. — Nouvelle évaluation de la fortune des ménages (1959-1967). — Budget-temps et choix d'activité.

N° 2. — Enquête sur les loisirs et mode de vie du personnel de la Régie Nationale des Usines Renault. — Étude des effets différentiels des impôts sur la consommation. — La morphologie sociale des communes urbaines.

N° 3. — La consommation élargie. — Étude économique de l'activité des médecins. — Possibilités et difficultés de la régulation des problèmes d'environnement et de nuisance par entente spontanée entre les intéressés.

N° 4. — Nature et prix des soins médicaux en ville. — Quelques résultats de l'étude des bilans de petites et moyennes entreprises.

SOMMAIRE DES PROCHAINS NUMÉROS

Qualité de la vie et choix collectifs. Consommation et statut social. Tests d'hypothèses linéaires sur un modèle de régression. Les sciences humaines devant la ville et le logement. Modèles de projection de consommation médicale. Le système fiscal aux U.S.A. Les indicateurs sociaux. Les dépenses de santé au Canada. Consommation des ménages, séries chronologiques.

sommaire

ÉTUDES

MÉTHODOLOGIE

CENTRE DE RECHERCHES ET DE DOCUMENTATION SUR LA CONSOMMATION
45, boulevard de la Gare, PARIS-13e
Tél. POR. 97-59

1972 n° 1
Janvier Mars