LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI
F. BENZÉCRI

Programmes d'analyse des correspondances et de classification ascendante hiérarchique : notice d'utilisation

Les cahiers de l'analyse des données, tome 14, n° 1 (1989), p. 7-34

<http://www.numdam.org/item?id=CAD_1989__14_1_7_0>

Les Cahiers de l'Analyse des Données
Vol XIV - 1989 - n°1 - pp. 7-34.

PROGRAMMES D'ANALYSE DES CORRESPONDANCES ET DE CLASSIFICATION ASCENDANTE HIÉRARCHIQUE: NOTICE D'UTILISATION

[NOT. CORR. CAH]

J.P. & F. BENZÉCRI

Les programmes considérés ici, exécutables sur l'ordinateur Macintosh+, SE ou Macintosh II, font partie d'une chaîne permettant d'analyser complétement une enquête par questionnaire, ou les résultats d'une expérimentation biologique; et aussi de créer une carte de synthèse d'après l'analyse d'un fichier de données économiques ou épidémiologiques, ventilées par unités territoriales (e.g., dans le cas de la France, par départements). Dans la présente notice, les programmes d'analyse des correspondances et de CAH ne seront pas décrits d'abord avec toutes leurs possibilités; mais seulement présentés sur un exemple simple, afin de permettre à un utilisateur de s'accoutumer à l'enchaînement des étapes du traitement, certains compléments étant donnés en addenda. *Le même exemple sera repris, également sous une forme élémentaire, dans la notice [NOT. PROG. CART.] du programme de cartographie. Le lecteur qui aura assimilé la présente notice et acquis en analyse des données le niveau de base d'un manuel tel que le volume 1 de la série* Pratique de l'Analyse des Données, *pourra, d'après la notice [NOT. CRÉ. TAB.], utiliser nos programmes pour coder des données, découper des variables en classes, créer un tableau de Burt, etc...*

0 Introduction à l'exemple choisi

Nous avons choisi de partir d'un exemple de géographie économique: le tableau de données croisant l'ensemble des départements de la France, avec un ensemble de variables: les consommations mensuelles d'un produit pétrolier, le gazole. En bref, l'analyse de ce tableau montrera d'abord une opposition entre les départements où est très importante la consommation hivernale et très faible la consommation estivale, et les départements où prédomine la consommation estivale, ou tout au moins les départements où la consommation estivale est beaucoup moins réduite qu'il n'est le cas en moyenne. Apparaîtra ensuite le cas des départements betteraviers, à forte consommation automnale.

Les cahiers de l'analyse des données – 0339-3097/89/01 7 28/$ 4.80/

Afin de mettre en évidence avec précision cette stuctrure, (et, ultérieurement, dans le cas de notre exemple, de produire une carte qui, dans son principe, ne surprendra pas les géographes), nous utiliserons l'ensemble des ressources de l'analyse des données:

analyse factorielle, qui assigne, d'après de tableau initial, des coordonnées, ou facteurs, à chacun des départements ainsi qu'à chacun des mois; autrement dit, qui traduit en termes de proximité sur un graphique les ressemblances existant entre individus et entre variables, ainsi que les affinités des uns avec les autres; et ensuite,

classification automatique, qui, au sein de cette représentation géométrique, détermine des classes (dans notre cas, principalement des classes de départements à chacune desquelles sera affectée ultérieurement, dans le tracé de la carte, une trame unique; tandis que les classes de mois, déjà en évidence sur les graphiques, nous sont connues *a priori*).

Dans la présente notice, nous considérerons successivement la création du tableau des données, la méthode d'analyse factorielle des correspondances et la méthode de classification automatique. Ce sera pour nous l'occasion de donner à l'utilisateur quelques notions sur la représentation des données à l'intérieur d'un ordinateur. Des appendices et notes seront consacrés aux fonctions des programmes non exploitées pour traiter l'exemple de base. Ces *addenda* pourront être omis en première lecture.

N.B. Avant d'être rédigé, l'exposé a été fait oralement par un opérateur décrivant les étapes successives du traitement qu'il effectue. Même si le lecteur ne dispose pas d'une disquette ni d'un Macintosh pour parcourir lui-même les étapes du traitement des données, il peut s'imaginer être placé devant un écran, en considérant les graphiques joints au texte. En effet, ces graphiques ne sont autres que des copies d'écran, saisies grâce à une commande dont dispose le Macintosh et imprimées, soit instantanément sur l'imprimante imagewriter II, soit en différé, sur l'imprimante laserwriter d'après un document macpaint (copie de l'écran point par point gardée en mémoire sur disquette).

1 L'entrée du tableau des données

Afin de soutenir l'attention de l'utilisateur, nous avons fixé un exemple concret bien déterminé en nous bornant à signaler les cas de données plus compliquées que l'on peut être amené à utiliser. Notre exemple de la consommation de gazole en France, ventilée par départements et par mois, fait partie de la thèse de M. Moussaoui consacrée plus généralement à l'étude de la variation saisonnière et annuelle de la consommation de tous les produits pétroliers en France sur la période 1972-1981.

Le présent tableau comporte 95 lignes correspondant aux 95 départements français (la Corse étant considérée comme constituant un seul département)

rangés depuis l'Ain jusqu'au Val-d'Oise dans l'ordre de leurs numéros minéralogiques. Le tableau a 12 colonnes qui sont les mois de Janvier à Décembre. A l'intersection, par ex., de la ligne Ardèche et de la colonne Juillet, on lit le nombre 236: ceci signifie que la consommation en gazole dans le département de l'Ardèche au cours du mois de Juillet des années 1972 à 1981 a été au total de 236 centaines de m3.

La première étape d'une analyse de données sera évidemment l'introduction de ce tableau dans l'ordinateur. Cette étape est aujourd'hui grandement facilitée par l'existence de ce qu'on appelle des éditeurs de texte. En bref, en utilisant un éditeur de texte, on tape sur le clavier de l'ordinateur comme on taperait sur le clavier d'une machine à écrire, et l'on voit s'afficher devant soi à l'écran ce que l'on a tapé. Mais, tandis que dans une machine à écrire ordinaire les corrections de symboles erronés, les insertions ou suppressions de lignes ou de mots ne se peuvent faire que par des opérations fastidieuses, sur l'ordinateur muni d'un traitement de texte, ces opérations sont d'une extême facilité. Elles sont particulièrement simples sur l'ordinateur Macintosh où les manœuvres se font en grande partie en manipulant un bouton poussoir sur ce qu'on appelle la souris.

Ce tableau,qui s'affiche à l'écran, est conservé dans la mémoire de l'ordinateur. Plus précisément, dans notre cas, il est conservé sur une disquette où se trouvera l'ensemble du travail de cartographie: tous les calculs intermédiaires et aussi les résultats d'après lesquels sera commandée l'impression de la carte. Disons tout de suite qu'on peut en bref schématiser ainsi le contenu de cette disquette:

il y a des programmes auxquels on fait appel pour traiter les différentes données;

il y a des textes, tableaux et graphiques composés de signes alphabétiques, qui sont accessibles directement à l'utilisateur grâce à l'éditeur de texte qui permet de les rappeler, de les afficher, de les parcourir s'ils comprennent un très grand nombre de lignes ou de pages, et éventuellement de les modifier, qu'ils aient été créés par programme ou au clavier;

et il y a les fichiers numériques internes qui sont utilisés pour faire communiquer les programmes entre eux (c'est à dire communiquer à un programme les résultats de calculs effectués par un autre) et qui éventuellement peuvent être mis sous forme de fichiers de texte que l'éditeur permet d'afficher à l'écran et aussi qu'un programme particulier permet d'imprimer dans un format choisi par l'utilisateur.

Dans notre cas, nous utilisons toujours, avec l'imprimante Imagewriter, plutôt que les commandes d'impression que comporte l'éditeur de texte, un programme nommé 'printX' qui permet d'imprimer, à la suite, un nombre arbitrairement grand de textes, sous la seule restriction que ceux-ci comportent

aucune lettre accentuée, ou autre caractère absent du clavier américain usuel. On évitera donc ces caractères dans le tableau des données et notamment dans les noms abrégés des lignes et colonnes (cf. *infra, addendum* au §1).

Nous pensons utile de présenter dès maintenant ces détails que l'utilisateur retrouvera constamment tout au long de la chaîne de traitement, depuis l'entrée des données dont il s'agit présentement, jusqu'à la création de fichiers qui aboutissent à la réalisation immédiate de la carte, en passant par tous les calculs d'analyse des données.

Affichons donc à l'écran du Macintosh le tableau des données. Apparaît sur cet écran un cadre, et en haut de ce cadre on lit un titre qui est:

```
qp5:gazole
```

Ce titre comporte d'une part le nom de la disquette utilisée (qp5), et d'autre part, après le symbole séparateur (:), le nom (gazole) que nous avons donné au tableau à analyser.

A l'intérieur du cadre se trouve le texte proprement dit que nous avons créé à partir du clavier et dont nous présentons maintenant le contenu.

La première ligne de ce texte est le titre:

```
consommation du gazole en France par départements et mois (1972-81)
```

Ce titre peut être composé librement par l'utilisateur; il ne faut pas le confondre avec gazole qui est le nom donné au fichier et qui figure, par ex., dans l'indice des fichiers contenus sur la disquette. Ce titre a toutefois une grande importance pour nous, dans la mesure où la plupart des programmes de traitement affichent à certains moments ce titre *in extenso* afin de rappeler à l'utilisateur dans quelle étude il se trouve. Il peut avoir une longueur quelconque dans la limite d'une ligne; nous conseillons toutefois, du fait des contraintes de l'affichage à l'écran, de ne pas dépasser 70 caractères. L'entrée de cette ligne de titre s'effectue par la frappe de la touche Retour chariot.

Après le titre, viennent les données proprement dites. Nous nous sommes efforcé, dans ce programme, d'imposer à l'utilisateur le moins de contraintes possible dans la présentation de ses données et lui laisser le plus de commodité pour donner à son tableau une forme qui lui permette de le relire et de le corriger facilement. Nous décrirons donc le tableau tel qu'il s'affiche à l'écran, tout en signalant les libertés que l'utilisateur peut prendre et, éventuellement, doit prendre dans la création de son tableau.

Sous la ligne de titre, se trouve une ligne qui commence par le nombre 12 suivi des initiales des mois qui sont présentés comme des mots de quatre caractères majuscules: JANV pour Janvier, FEVR pour Février, etc. jusqu'à DECE pour Décembre. Le nombre 12 est tout simplement le nombre des colonnes du tableau. Ces identificateurs de colonnes, ou sigles des variables, peuvent comporter de 1 à 4 caractères, choisis librement sur le clavier de

l'ordinateur, à ceci près qu'il ne doit pas y avoir d'espace blanc ni, répétons-le, de caractères accentués.

Après la ligne des noms des colonnes, on trouve une suite de lignes, chacune afférente à un département (individu).La première est la ligne Ain qui commence par le nom même de ce département, écrit Ain, suivi de nombres: 681 785 775 etc. jusqu'à 802, afférents aux 12 mois.

Pour les lignes comme pour les colonnes, on doit choisir les sigles avec la même règle: 1 à 4 caractères, sans blanc interposé. Pour les départements, ce choix est relativement facile, même pour ceux dont le nom comporte plus de quatre lettres; par ex., sous Ain, on voit écrit Aisn pour Aisne, puis Alli pour Allier. Le choix des sigles est un art très utile en Analyse des Données et l'on s'habitue progressivement à avoir des sigles aussi évocateurs que possible et ne prêtant à aucune confusion.

L'utilisateur demandera naturellement dans quelles limites sont compris les nombres permis de lignes et de colonnes pour un tableau à analyser. Le programme limite à 300 le nombre des colonnes; le nombre des lignes n'est limité que par la taille de la mémoire. Si le tableau des données est créé comme on l'a expliqué au §1.1, et qu'on utilise un ordinateur Macintosh+ ou Macsintosh SE ayant 1 mégaoctet de mémoire centrale, on peut avoir 150 colonnes et 1000 lignes. Avec deux fois plus de capacité, le nombre des colonnes peut être de 250 avec 1000 lignes; ou de 300 avec 700 lignes.

Quant aux nombres inscrits dans les cases, nous considérons ici le cas de nombres entiers positifs ou nuls, dont le nombre de chiffres est limité à 9. Cette limite n'est aucunement contraignante, car des données supérieures peuvent facilement, par un simple changement d'échelle, être ramenées à des nombres inférieurs à 1 milliard. On peut égalemenent utiliser des nombres réels, ainsi que nous l'expliquerons en *addendum* 2 au §2.1.

L'exemple que nous avons pris pour base de notre exposé est un tableau qui a seulement 12 colonnes. Ce tableau tient dans la largeur de l'écran et, à plus forte raison, dans les limites permises par l'éditeur de texte qui peut faire glisser le texte sur l'écran visible. Mais avec un tableau de 48 colonnes par ex., une telle commodité ne nous est plus offerte. Il est donc indispensable de pouvoir écrire ce qui constitue structurellement une ligne du tableau des données sur plusieurs lignes physiques du tableau que l'on imprime. Le programme d'analyse des correspondances qui consulte le texte créé contenant le tableau des données, est conçu de telle sorte qu'il laisse à l'utilisateur la plus grande liberté dans l'inscription des informations qu'il doit fournir. Nous avons dit qu'il importait de mettre en tête un titre suivi d'un Retour chariot. Cela fait, la liberté est totale quant aux Aller à la ligne. L'utilisateur doit seulement écrire successivement le *nombre de colonnes*, les *sigles des variables* afférentes aux diverses colonnes, puis ce qui constitue les *lignes*, mais disposé, répétons-le, de

façon arbitraire, comme des séquences en tête de chacune desquelles vient le sigle de l'individu, suivi des nombres entiers représentant les valeurs des variables. La seule contrainte à respecter est qu'entre un sigle et un nombre on doit laisser au moins un blanc et de même entre deux nombres successifs. Il va sans dire que cette permission laissée à l'utilisateur de disposer les données dans le plus grand désordre ne doit pas être un encouragement au désordre. Par ex., dans le cas d'un tableau à 48 colonnes donnant successivement les consommations mensuelles de quatre produits, il conviendra d'écrire en tête de la première ligne le nombre 48, puis 12 sigles choisis pour les consommations mensuelles du premier produit, puis sur la ligne suivante, bien à l'alignement, les 12 sigles du produit suivant, etc.; et après ce bloc de quatre lignes d'en-tête donnant l'ensemble des variables, viendront d'autres blocs de quatre lignes avec, sur la première ligne, le nom d'un département suivi des 12 consommations mensuelles du premier produit, puis, sur la ligne suivante, les 12 consommations du second produit, etc.. Mais, encore une fois, ces conseils de bonne disposition ne sont que destinés à permettre à l'utilisateur d'entrer, de relire, éventuellement de corriger son tableau le plus commodément possible. Ils ne correspondent nullement à une contrainte imposée par le programme d'analyse des correspondances. En *addendum* 2 au §2.1 , nous expliquerons comment on peut même introduire des commentaires dans le tableau des données.

Addendum au §1: Impression des fichiers de texte

Si l'on utilise l'imprimante usuelle Imagewriter, (à ruban et aiguille), nous recommandons d'imprimer par le programme 'printX', plutôt que par la commande d'impression d'un éditeur de texte.

Ce programme permet d'imprimer, à la suite, un nombre arbitrairement grand de fichiers de texte, et cela sous un format qui, pour des résultats de calcul, est sans doute le plus commode: le format en lignes de caractères ultracomprimés, ce qui, avec l'imprimante Imagewriter de l'ordinateur Macintosh, correspond à 136 caractères par ligne. La seule condition, essentielle toutefois, pour utiliser 'printX' est que le texte à imprimer ne comporte, répétons le, aucune lettre accentuée, ou autre caractère absent du clavier américain usuel. On évitera donc ces caractères dans le titre du tableau des données ainsi que dans les noms abrégés, (sigles), des lignes et colonnes.

En ouvrant 'printX', on voit s'afficher à l'écran une recommandation très importante :

```
ATTENTION  ce programme fonctionne exclusivement avec
une imprimante Imagewriter; avant de répondre oui(O) à la
question ci-après, vérifier que l'imprimante est allumée
```

puis la question

```
l'imprimante est-elle allumée oui(O) ou non(N)
```

C'est seulement si l'imprimante est bien allumée, et qu'on a répondu "oui", (en entrant, comme ce sera toujours le cas, la lettre O), que l'imprimante est prête à imprimer proprement, en caractères ultracomprimés, les textes dont on lui donne les titres, par groupe de 10 au plus, (en prenant garde de ne point faire d'erreur!, mais avec la faculté de se corriger en utilisant la touche d'effacement).

Si l'on utilise l'imprimante laserwriter, on peut imprimer tout listage grâce à un éditeur de texte tel que 'qued'; nous recommandons de choisir le caractère 'courier', en taille 9, et de réduire s'il y a lieu, ou d'orienter le papier transversalement. Avec des caractères de largeur variable, comme en 'times', les alignements ne seraient pas respectés.

2 L'analyse des correspondances

Nous considérerons successivement dans ce § le programme d'analyse des correspondances proprement dit, puis deux programmes qui permettent de présenter, soit directement à l'écran, soit sous forme de texte imprimé, des graphiques plans où figurent à la fois les sigles des lignes et des colonnes du tableau initial, disposés selon les proximités que l'analyse a fait apparaître entre elles. Nous présenterons dans des *addenda* des fonctions du programme non utilisées pour traiter l'exemple de base.

2.1 Le programme 'qori' d'analyse des correspondances

Proposons d'abord à la curiosité de l'utilisateur une explication sur le nom choisi: 'qori'; explication d'autant plus utile qu'elle apportera déjà quelque lumière sur la structure et les possibilités de ce programme.

Ce nom commence par la syllabe 'qor' parce qu'il s'agit d'un programme d'analyse des correspondances. Mais, curieusement, cette syllabe est écrite avec la lettre q pour rappeler, par la présence des deux lettres 'q' et 'r', que les calculs de diagonalisation de matrices sont effectués par un algorithme, remarquablement performant, qui est l'algorithme 'symqr'. La lettre 'i' qui suit rappelle que, dans son utilisation la plus simple, le programme accepte comme données des entiers longs, c'est-à-dire des nombres allant jusqu'à un milliard et même, en fait, dépassant quelque peu ce nombre. (Pour l'usage de nombre réels quelconques, cf. *addendum* 2 au §3.1).

Ouvrons le programme qori comme il est d'usage de le faire lorsqu'on utilise un ordinateur. Apparaît à l'écran la phrase:

```
le fichier des données est
```

le programme prévoit que l'utilisateur introduise à partir du clavier le nom donné au fichier à l'intérieur de l'ordinateur, c'est-à-dire le nom qp5:gazole que nous avons déjà vu affiché dans le cadre du traitement de texte au §1.1. Nous tapons donc ce nom (en prenant garde de ne point faire d'erreur, mais avec la faculté de nous corriger en utilisant la touche d'effacement):

```
qp5:gazole
```

```
Yvel   947   950  1080  1067  1046  1076   974   706  1030  1142  1049  1116
2Sev   473   465   526   530   543   520   504   403   514   560   493   539
Somm   723   696   761   780   788   787   754   608   743   914   890   921
Tarn   289   306   313   333   354   344   323   262   334   374   336   363
TGar   256   249   285   285   290   292   302   266   296   319   282   300
 Var   629   638   726   741   740   759   841   750   707   725   666   730
Vauc   719   715   794   798   817   810   871   753   842   910   818   873
Vend   668   645   751   740   764   816   833   728   732   771   750   714
Vien   660   638   691   722   801   799   728   581   713   791   725   763
HVie   434   426   479   485   493   482   480   360   480   530   475   496
Vosg   472   471   543   556   575   568   557   425   564   611   547   553
Yonn   819   819   922   906   900   925   935   713   909   990   906   939
Belf   154   152   172   173   175   182   165   129   173   195   170   174
Esso   917   883  1018  1000  1010  1016   959   733   960  1108  1004  1046
HSei   853   837   943   919   881   907   825   603   867   982   903   966
SDen  1281  1230  1360  1324  1300  1321  1221   859  1262  1428  1328  1385
VMrn  1066  1011  1143  1114  1099  1116  1010   786  1073  1216  1107  1191
VOis   720   715   814   796   775   796   739   558   785   882   804   872
nombre de facteurs a garder sur fichier = 20
nombre de facteurs a ecrire sur listage = 20
le nombre des colonnes est  12
y a t il colonne supplementaire(S) ou non(N) N
le nombre des lignes est  95
y a t il ligne supplementaire(S) ou non(N)
```

Fin de l'affichage des données et reprise du dialogue

et ensuite la touche:

```
Retour chariot
```

L'ordinateur répond en affichant la phrase:

```
ce nom est-il confirmé oui (O) ou non (N)
```

phrase, suivie d'une barre verticale qui s'affiche et s'efface alternativement, en signe que l'ordinateur attend une réponse. Nous tapons 'O' majuscule:

```
O
```

puisque le nom du fichier a été entré sans erreur.

Si le nom tapé n'avait pas correspondu à un fichier disponible pour l'ordinateur, le programme aurait affiché un commentaire d'erreur. D'autre part, si nous avions tapé par inadvertance un autre nom que celui du fichier désiré, il nous serait possible de répondre non en tapant: 'N'. Dans l'un et l'autre cas, le dialogue reprendrait au point de départ.

L'ordinateur procède à la lecture du tableau des données, tout en affichant celles-ci. En quelques secondes, toutes les lignes sont passées, d'Ain à Val-d'Oise, et la fin du tableau s'est immobilisée sur l'écran.

L'ordinateur affiche maintenant la phrase:

```
nombre de facteurs à garder sur fichier =
```

suivie de la barre alternative.

N. B. Si nous tapons un nombre trop grand, le programme ramènera ce nombre dans les limites permises (soit 29, dans la présente version). De plus, le

nombre de facteurs gardés ne peut, certes, dépasser le nombre total des facteurs, nombre qui est inférieur ou égal au plus petit des deux nombres de colonnes ou de lignes, moins 1.

Nous demandons 20 facteurs (tout en sachant qu'avec 12 colonnes il ne peut y avoir plus de 11 facteurs):

```
20
```

S'affiche alors la phrase:

```
nombre de facteurs à écrire sur le listage =
```

suivie de la barre alternative.

Si l'on désire consulter les résultats de l'analyse factorielle exclusivement sur le fichier de type texte que le programme 'qori' va créer, c'est-à-dire en passant par l'éditeur de texte, il est préférable de demander le nombre maximum de facteurs prévu par le programme, c'est-à-dire 10. Si l'on s'intéresse à une sortie imprimée des résultats de l'analyse factorielle, il faut savoir que la largeur de l'imprimante Imagewriter avec 136 caractères ne permet pas d'imprimer plus de 8 facteurs. Si toutefois on a dans le fichier de texte des lignes plus longues, le programme printex imprimera le début des lignes où l'on pourra lire les 8 premiers facteurs, en supprimant les fins de lignes contenant les facteurs 9 et 10 qui sortent des limites d'impression possible.(L'imprimante laserwriter permet de sortir 10 facteurs).Nous écrivons donc:

```
10
```

et l'ordinateur affiche immédiatement l'information:

```
le nombre de colonnes est 12
```

et, après un temps d'arrêt consacré à des calculs, il affiche la phrase:

```
y a-t-il colonne supplémentaire (S) ou non (N)
```

avec la barre alternative. Notre réponse sera Non:

```
N
```

nous n'entrerons pas dans le dialogue de création d'éléments supplémentaires (qui fait l'objet de l'*addendum* 1 au §2.1).

Qu'on soit ou non entré dans le dialogue de création de colonnes supplémentaires, on voit s'afficher:

```
y a-t-il ligne supplémentaire (S) ou non (N)
```

Ici encore, notre réponse sera non:

```
N
```

(En répondant 'S', on serait entré dans un dialogue de création de lignes supplémentaires tout analogue à celui brièvement décrit pour les colonnes dans l'*addendum*.)

Le programme affiche alors à nouveau le titre:

Vien	4	0	12	7	24	4	-13	-15	-1	-5
HVie	-13	8	11	-5	0	4	4	-1	-3	0
Vosg	0	-3	26	-2	3	-1	3	1	-3	0
Yonn	-1	11	4	-12	-6	-4	2	1	4	-1
Belf	-15	4	19	7	-6	-2	-9	-4	2	-1
Esso	-25	8	1	0	-1	-4	-3	-4	-4	5
HSei	-52	24	-7	1	-4	-3	-5	3	1	2
SDen	-55	29	-5	-2	-1	-1	0	-7	5	1
VMrn	-41	19	-10	7	-4	2	-4	-2	4	6
VOis	-40	5	-5	-3	-5	0	-6	10	3	3
JANV	-22	20	-12	3	-2	7	2	-10	4	5
FEVR	-17	22	-8	0	-5	1	-1	1	0	-10
MARS	-18	17	0	2	-7	-10	1	6	-1	4
AVRL	-7	9	3	4	2	-2	3	3	-7	2
MAI	1	-2	10	11	15	5	4	-2	-5	-1
JUIN	10	3	10	4	7	-7	-13	-3	4	0
JUIL	45	6	8	-27	2	0	2	-2	-2	0
AOUT	113	-9	-15	11	-6	-1	1	0	0	0
SEPT	3	-5	17	3	-3	7	5	6	9	0
OCTO	-27	-19	8	0	-16	4	-4	-4	-5	0
NOVE	-32	-24	-8	-2	3	-12	7	-4	3	-2
DECE	-26	-12	-19	-7	8	7	-6	7	0	2

Affichage des facteurs: fin des départements, et ensemble des mois

```
consommation du gazole en France par départements et mois (1972-81)
```

Puis l'ordinateur crée des fichiers et calcule quelque temps sans rien afficher.

Nous voyons ensuite s'afficher des lignes de chiffres. D'abord, par lignes de 20, les numéros des départements, de 1 à 95. Puis la ligne de 1 à 10, puis en sens inverse de 11 à 2; et, à partir de ce moment-là, des lignes plus courtes qui débutent successivement par les nombres de 12 à 2 suivi, chacun, d'un nombre variable d'entiers: 1 2 3 4 5; 1 2 3; 1 2; etc.. Cet affichage est destiné seulement au statisticien qui prend ainsi connaissance du nombre d'itérations dont l'algorithme symqr a eu besoin pour calculer chacun des facteurs successifs.

Maintenant, s'affichent à l'écran des nombres qui nous intéressent. Il s'agit de lignes commençant chacune par le nom d'un département suivi de 10 nombres entiers affectés de signes. Ce sont les résultats de l'analyse factorielle pour l'ensemble des lignes. On a à l'écran, par ex., sur la ligne HSei relative aux Hauts-de-Seine, les nombres:

```
-52  24  -7  1  -4  etc..
```

Ces nombres représentent les coordonnées sur les axes factoriels 1, 2, 3, 4, 5, etc. du département des Hauts-de-Seine. Plus précisément, ce sont ses coordonnées imprimées en millièmes. Donc -52 signifie -0,052 etc..

A vrai dire, le défilement des nombres est très rapide et peut seulement rassurer l'utilisateur sur la qualité des résultats obtenus.

En même temps que l'ordinateur effectue cet affichage, il crée un fichier de texte qu'on pourra appeler à l'écran en utilisant l'éditeur de texte et où l'on

pourra lire tout à loisir non seulement les valeurs des facteurs, mais également des informations de corrélations et contributions familières aux praticiens de l'analyse des correspondances et qui permettent de critiquer de façon précise les résultats suivant des principes dont le détail a été donné ailleurs sous une forme simple à l'usage notamment des médecins, des linguistes et des économistes dans la revue C.A.D. et les volumes de la série: Pratique de l'Analyse des Données.

À la suite des lignes relatives aux départements, l'ordinateur a affiché des lignes analogues relatives aux mois: Janvier, Février etc. jusqu'à Décembre. Par ex., sur la ligne Août, on lit d'abord 113 qui est la valeur la plus forte apparue dans la première colonne. Le mois d'Août occupe une position extrême sur l'axe 1 de l'analyse factorielle, cet axe traduisant tout simplement l'opposition entre les autres mois et ce mois de vacance où la consommation de gazole est très forte précisément dans les départements où les activités touristiques sont développées, tandis qu'elle est très faible dans les départements, tels que ceux notamment de la couronne parisienne, désertés de leurs habitants pendant les congés annuels.

Au bas de l'écran s'affiche de nouveau la barre alternative sans question particulière. L'utilisateur doit comprendre que s'il veut quitter le programme 'qori' et l'écran de l'ordinateur, il doit taper un caractère quelconque et l'entrer, ce que nous faisons.

Addendum 1 au §2.1: Dialogue des éléments supplémentaires

Rappelons qu'on appelle élément supplémentaire un élément (ligne ou colonne) qui n'intervient pas dans la détermination des axes factoriels, mais est projeté sur ceux-ci parmi les éléments, dits principaux, par lesquels les axes sont déterminés.

Si à la question:

```
y a-t-il colonne supplémentaire (S) ou non (N)
```

on répond:

```
S
```

l'ordinateur, afin de permettre à l'utilisateur de choisir, affiche les numéros et les sigles des colonnes du tableau; puis est posée la question:

```
les col sup sont-elles désignées une par une (U) ou par blocs (B)
```

Si l'on répond:

```
U
```

apparaît la question:

```
le nombre de col sup sera
```

et, après avoir répondu à cette question par un nombre entier et frappé retour chariot, l'utilisateur se voit demander de désigner successivement par leurs numéros la première col sup, la 2-ème col sup, etc....

Si au contraire on répond:

```
B
```

un dialogue se déroule pour fixer le nombre de blocs de colonnes suppl; puis, pour chacun de ceux-ci, le numéro de sa première et de sa dernière colonne. Cette forme de désignation est très utile car souvent les variables sont rangées en blocs successifs, concernant des thèmes différents.

Il va sans dire que l'utilisateur doit profiter de l'affichage de la liste des colonnes pour arrêter le choix des éléments supplémentaires; mais, d'une part, il peut effacer un nombre erroné non encore entré, grâce à la touche 'recul'; et, d'autre part, si, finalement, les demandes qu'il a faites, (par individus ou par blocs), ne le satisfont pas, la question usuelle

```
ce choix est il confirmé O ou N
```

lui offre l'occasion de revenir sur son choix.

Le choix des lignes supplémentaires font l'objet d'un dialogue tout semblable à celui que l'on vient de décrire pour les colonnes.

Addendum 2 au §2.1: Diverses formes du tableau des données

Au §1, nous avons présenté un tableau de données ne contenant que des données sous forme d'entiers; dans un tel tableau, entre la ligne de titre et le nombre des colonnes (12 dans notre cas) on peut introduire un texte de commentaire quelconque, assujetti à la seule condition de ne comporter aucun des 10 chiffres (de 0 à 9); il est même possible d'introduire un tel commentaire après le sigle d'une ligne, avant les nombres dont se compose cette ligne. Si l'on veut économiser la mémoire, on peut, en tête du tableau, indiquer le nombre des lignes (ou un nombre majorant celui-ci); et, pour cela, à la place du nombre des colonnes, on écrit le nombre:

1000 × nombre des lignes + nombre des colonnes,

soit, dans notre cas, 95012.

Si l'on veut introduire dans le tableau des données réelles quelconques (nombres décimaux), on doit donner au tableau un nom de base suivi du suffixe yy; et c'est le nom de base qui est donné en réponse à la question initiale du programme 'qori'. Les données réelles peuvent être écrites soit avec un point décimal pour séparer la partie entière de la suite; soit avec un exposant suivant le nombre. Voici cinq écritures équivalentes d'un même nombre:

575 ; .575e+3 ; 5.75e+2 ; 5750.00e-1 ; 0.0575e+4.

Si le tableau comporte des données réelles quelconques, on peut y introduire un commentaire après la ligne de titre; mais non après le sigle d'une ligne.

Si l'on utilise un disque de grande capacité, il s'impose de disposer les diverses analyses dans des dossiers séparés; dans un tel cas le nom de base pourra être: D:gaz:gazole, où D est le nom du disque et gaz celui du dossier.

Enfin, nous signalerons que le programme 'zrang' de découpage des variables en modalités crée des tableaux de données qui ne sont pas en format de texte, mais en format numérique; ces tableaux ont des noms terminés par le suffixe 'zz' ou 'ww'; ils sont accessibles au programme 'qori'.

2.2 Le programme 'planF' d'affichage direct des graphiques plans issus de l'analyse factorielle

Outre le fichier de type texte nommé gazolecortx (fichier qui contient sous forme de tableaux des résultats numériques dont une partie s'est affichée tout à l'heure à l'écran, à la fin du déroulement du programme), le programme 'qori' crée des fichiers numériques où sont gardés les mêmes résultats d'analyse des correspondances et qui seront utilisés dans toute la suite des programmes de la chaîne: dans le programme 'planF' qui nous occupe présentement, pour produire les graphiques plans, puis dans la classification automatique, et enfin dans l'exécution de la carte proprement dite.

Nous appelons donc le programme 'planF' qui affiche directement les résultats sur l'écran de l'ordinateur, sans créer aucun fichier. Au prochain §, nous présenterons brièvement un autre programme, 'planX', qui crée un fichier de type texte sur lequel s'inscrivent des informations analogues à celles que nous allons voir maintenant, mais sous une forme telle , qu'on doit y accéder par l'éditeur de texte, avec cet avantage qu'on peut aussi en demander l'impression.

Sur l'écran, s'affiche la phrase:

```
le fichier de base est
```

suivie de la barre alternative. Nous avons déjà vu cette même phrase s'afficher en tête du programme 'qori'. Nous la retrouverons en tête du programme de classification et, sous une forme un peu différente, en tête du programme de cartographie. Dans la mesure du possible, pour la commodité de l'utilisateur, on s'est efforcé de faire en sorte que le seul nom de fichier qu'il ait à fournir soit le nom du fichier de base, c'est-à-dire, dans notre cas, qp5:gazole (qp5 parce que c'est le nom de la disquette que nous utilisons). A partir de ce nom, par simple addition de suffixes, l'ordinateur construit d'une part les noms des fichiers numériques qu'il a à consulter pour produire les résultats que nous lui demandons, et, d'autre part, les noms des fichiers qu'il crée pour y ranger ces résultats,fichiers de type texte, accessibles directement à l'utilisateur et susceptibles d'être imprimés, ou fichiers numériques à l'usage exclusif d'autres programmes qui, eux, produiront des résultats lisibles.

Nous entrons donc:

```
qp5:gazole
```

Comme précédemment, l'ordinateur demande confirmation de ce titre. Nous répondons oui; le titre est confirmé.

Notons ici, comme on l'a signalé au §2.1 à propos du programme 'qori', que le programme 'planF' vérifie l'existence des fichiers numériques nécessaires pour effectuer les opérations qu'il a à effectuer et affiche un commentaire d'erreur si l'un de ces fichiers manque.

S'affiche alors à l'écran la phrase:

```
le nombre (de 1 à 20) des ensembles considérés est:
```

Jusqu'à présent, il n'existe que deux ensembles: l' ensemble des lignes, qui est désigné par la lettre i, et l'ensemble des colonnes, qui est désigné par la lettre j. L'ensemble i est, dans notre cas, l'ensemble des 95 départements français, et l'ensemble j, celui des 12 mois. Mais il pourrait y avoir des ensembles, is et js, de lignes et colonnes supplémentaires; et, si l'on appelle le programme planF après avoir fait tourner le programme de classification automatique, on peut aussi demander l'affichage des ensembles iq, ou ensemble des classes de départements qui auront été construites, et jq, ou ensemble des classes de mois; sans parler d'autres tableaux de données que l'on peut adjoindre en supplémentaires à l'analyse du tableau de base. Voilà pourquoi on prévoit ici un nombre d'ensembles allant de 1 à 20. Notre réponse sera:

```
2
```

L'ordinateur affiche alors la question:

```
le sigle de l'ensemble n°1 est
```

Gardons-nous de répondre i, car l'affichage des deux ensembles se fera en deux étapes, dans l'ordre des numéros qu'il s'agit ici de leur attribuer. Si nous demandions i sous le n°1, l'ensemble des 95 départements se projetant d'abord sur l'écran créerait un certain encombrement, une accumulation fâcheuse. Nous demanderons plutôt que s'affiche d'abord l'ensemble des colonnes, c'est-à-dire l'ensemble des 12 mois, et par là-dessus s'afficheront tant bien que mal, éventuellement en surcharge, les sigles des 95 départements.

C'est ici qu'apparaît au mieux la nécessité de la classification automatique. Quand nous aurons effectué une partition de l'ensemble des départements en 8 ou 9 classes, il sera facile de voir clairement sur l'écran à la fois les sigles de ces classes et les sigles des 12 mois.

Cependant, sur un graphique plan couvrant une surface plus grande que celle de l'écran, il est possible, même avec 95 départements, de présenter l'ensemble des individus, presque sans perte de points. C'est ce que permet le programme 'planX', objet du §2.3 suivant.

Nous demandons donc, comme ensemble n°1:

```
j
```

et à la question suivante;

```
le sigle de l'ensemble n°2 est
```

nous répondons:

```
i
```

Suit la ligne invariable de confirmation:

```
ce choix est-il confirmé oui (O)ou non (N)
```

Nous répondons oui:

```
O
```

L'ordinateur affiche alors la phrase:

```
afin de préparer le tracé, on calcule les bornes des facteurs
```

Ce calcul des bornes est toutefois très rapide et l'utilisateur n'a pas à s'impatienter: il voit un nouveau dialogue à l'écran. Sous le titre:

```
choix des ensembles à représenter parmi ceux considérés
```

apparaît la phrase:

```
faut-il représenter l'ensemble j oui (O) ou non (N)
```

Nous répondons: oui. Suit la phrase:

```
faut-il représenter l'ensemble i oui (O) ou non (N)
```

à laquelle nous répondons également oui.

L'intérêt de ce dialogue est qu'il permet à l'utilisateur de n'afficher qu'un seul des ensembles qu'il a prévu de considérer au départ, ou, dans le cas où 4 ensembles sont considérés, de n'en considérer que 2 et deux différents selon les essais que le programme 'planF' permet de multiplier.

Maintenant apparaît la phrase:

```
NB afin que s'affiche chaque ensemble, il faut taper Retour chariot;
de même,on met fin à l'affichage en tapant Retour chariot
```

puis sous le titre:

```
choix des axes du plan
```

la question:

```
le n° de l'axe horizontal est ah =
```

Nous tapons:

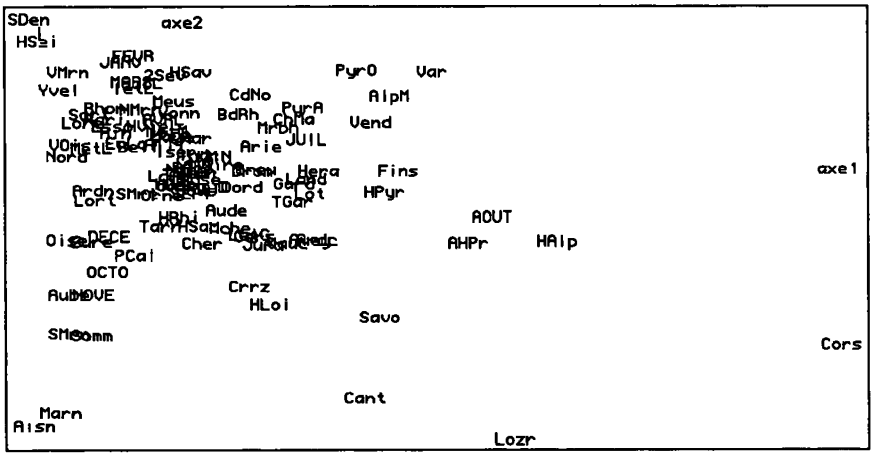
```
1
```

Suit la question:

```
le n° de l'axe vertical est av =
```

Nous tapons:

```
2
```



Affichage du plan (1,2) par le programme 'planF'

Cela signifie que les sigles des mois et des départements s'afficheront en ayant pour abscisse le premier facteur calculé par le programme 'qori', et pour ordonnée, le second facteur. Evidemment, d'autres choix seraient possibles.

Nous voyons maintenant à l'écran une page blanche avec seulement une croix qui représente l'origine, et aux extrémités de l'horizontale et de la verticale passant par cette croix, les mentions: axe 1 (sur l'horizontale) et axe 2 (sur la verticale). Pour l'instant, aucun sigle ne figure sur le plan.

Frappons la touche Retour chariot: s'affichent sur l'écran les noms des mois avec, comme nous l'avons déjà annoncé, dans la partie droite de l'écran, le plus à droite des mois: Août, qui n'est toutefois pas à la limite de l'écran. Août est suivi de Juillet et de Juin; et sur une ligne, de bas en haut, dans la partie gauche de l'écran, les mois de Novembre, Octobre, Décembre,Septembre, Mai, Avril, Mars, Janvier et Février. Ces noms de mois se lisent nettement, bien que Février soit en surimpression par rapport à Janvier. On voit clairement confirmée l'interprétation donnée plus haut de l'axe 1: le facteur 1 positif s'identifie à la période de vacances. Sur l'axe 2, en revanche, on a une opposition entre l'automne et l'hiver. Cette opposition sera plus clairement expliquée lorsqu'on pourra considérer le plan des axes 2, 3 et qu'on avancera dans l'interprétation des résultats.

Il est temps maintenant d'appuyer une nouvelle fois sur la touche Retour chariot.

Les noms des départements s'affichent. Au centre, le nuage est particulièrement dense et peu lisible, mais à la périphérie, en revanche, une

quarantaine de noms de départements se lisent clairement. A l'extrême droite, associé au mois d'Août, on lit: Cors. Le département de la Corse (que nous n'avons pas divisé dans cette étude en nord et sud) est un département où l'activité de vacances a une importance toute particulière. Suivent les départements des Hautes-Alpes et des Alpes-de-Haute-Provence et, en bas de l'écran, mais également vers la droite, les départements de la Lozère, de la Savoie et du Cantal. Sur le bord gauche de l'écran, on trouve en haut, associés à l'hiver, les départements de la couronne: SDen (Seine-Saint-Denis), HSei (Hauts-de-Seine), VMrn (Val-de-Marne), Yvel (Yvelines) etc.. Au contraire, dans le coin inférieur gauche, on lit Aisne, Marne, Somme, Seine-et-Marne. Ces départements, qui ont une consommation en gazole particulièrement forte en Octobre, sont, en fait, les départements où la récolte des betteraves sucrières et leur élaboration occupent plusieurs semaines en automne..

Nous quittons ce plan en tapant Retour chariot. L'ordinateur demande:

```
faut-il afficher un autre plan
```

Nous répondons oui en demandant le plan (2,3) où l'on peut voir s'afficher le cycle des mois, à peu près dans son ordre naturel, avec le mois d'Août dans une position non significative, puisque ce mois est déjà sorti très à l'écart sur le premier axe.

Nous répondons maintenant non à toutes les demandes que nous fait l'ordinateur pour poursuivre éventuellement l'utilisation du programme 'planF', et quittons ce programme pour indiquer rapidement l'usage du programme 'planX'.

2.3 Le programme 'planX' de création de graphiques plans sous forme de fichiers de textes

Inutile de dire qu'apparaît à l'écran la phrase:

```
le fichier de base est
```

à laquelle nous répondons:

```
qp5:gazole
```

Nous entrons et confirmons cette réponse et, comme avec le programme planx, nous précisons que le nombre des ensembles considérés est 2 et que ceux-ci sont désignés par les sigles: j, ensemble des colonnes (les mois) et i, ensemble des lignes (les départements).

Comme dans le programme précédent, on voit s'afficher la phrase:

```
afin de préparer le tracé, on calcule les bornes des facteurs
```

Ce calcul est rapide, et s'affiche maintenant la phrase:

```
choix des ensembles à représenter parmi les ensembles considérés
```

Nous répondons oui à la représentation de j et à celle de i.

Après le choix des axes, viennent des choix particuliers par lesquels le programme 'planX' diffère du programme 'planF'. Il s'agit du

```
choix des dimensions du graphique
```

A la question :

```
largeur du graphique en caractères =
```

nous répondons en entrant le nombre:

```
150
```

L'ordinateur répond en affichant à l'écran:

```
largeur = 136
```

136 est le nombre maximum permis par l'imprimante Imagewriter, que nous utilisons, avec le programme d'impression 'printX'. (Avec l'imprimante laserwriter, nous avons dit qu'on peut imprimer en passant par un éditeur de texte, utiliser la police courier 9 qui est la plus étroite; et réduire, ou, afin d'avoir des lignes plus longues, prendre les pages dans le sens transversal).

Suit la phrase:

```
afin d'avoir même échelle sur les deux axes,
sur l'écran,  demander une hauteur de 19 lignes,
sur le listage, demander une hauteur de 14 lignes
```

puis la question:

```
hauteur du graphique en lignes =
```

La question qui nous est ici posée est beaucoup plus embarrassante que la première. En effet, il apparaît que ces deux nombres, 14 ou 19, sont très faibles et que si nous demandons une hauteur de graphique qui permette que l'échelle sur les axes 1 et 2 soit la même, soit à l'imprimante, soit à l'écran, nous aurons un graphique extrêmement étroit sur lequel il sera impossible que s'inscrivent tous les points. Nous devons donc faire ce qu'a fait, sans nous en demander la permission, le programme 'planF': dilater largement l'échelle de l'axe 2. Ceci est dû à une particularité des données que nous utilisons: l'axe 1 a, ici, une importance beaucoup plus grande que tous les autres axes, l'opposition entre mois de vacances et mois d'activité économique étant prédominante dans cette analyse, tandis que l'opposition entre automne et hiver, liée surtout à l'activité des départements betteraviers, ne se traduit que par des distances assez faibles qui s'inscrivent sur l'axe 2.

Nous répondons donc 50 lignes, escomptant que cette hauteur de graphique permettra d'inscrire sans superposition la plupart des points.

L'ordinateur crée alors un fichier de texte qui contient précisément le plan que nous avons demandé. Quelques instants suffisent et apparaît à l'écran la question:

```
faut-il tracer un autre plan oui (O) ou non (N)
```

Nous pourrions répondre oui et et demander par ex. le plan (1, 3) qui se mettrait sur le même fichier de texte. Nous sortirons plus rapidement du programme 'planX' en répondant non, et nous bornerons à appeler le plan qui a été créé, sur l'éditeur de texte (qui permet de l'afficher à l'écran) .

L'ordinateur a placé ce fichier de texte sur la disquette qp5 que nous utilisons et lui a donné pour titre 'gazoleplantx'; le suffixe 'plantx' a été ajouté au nom du tableau des données.

L'éditeur de texte affiche d'abord à l'écran ce titre:

```
qp5:gazoleplantx
```

puis, après une double ligne de séparation, les indications:

```
ensembles représentés: j, i
axe horizontal: 1;min=-5,47e-2 ; max=2,41e-1;lam=1,36e-3;tau=6,89e-1
```

Ce sera l'occasion de commenter brièvement quelques résultats d'analyse factorielle. Le min et le max sont simplement la valeur minima et la valeur maxima du premier facteur pour les ensembles des départements et des mois réunis.Le lam désigne la valeur propre, (habituellement désignée par la lettre grecque λ), qui est ici très faible: 1,36 millièmes seulement, parce que les contrastes entre départements sont peu prononcés, même s'ils sont très signicatifs. Le tau désigne le taux: rapport de la première valeur propre à la somme de toutes les valeurs propres relatives aux différents axes factoriels de 1 à 11. (Rappelons que le nombre total de facteurs est égal au nombre d'éléments du plus petit des deux ensembles en correspondance, moins 1; ici: 12-1= 11.) Ce taux est de 6,89 dixièmes, c'est à dire 0,70 ou encore 69%. Il se confirme que l'essentiel de l'information contenue dans le tableau analysé est représenté sur l'axe 1.

Suivent des informations semblables relatives à l'axe vertical. Puis les mentions:

```
éléments j non représentés: 0
éléments i non représentés: Avey Calv Doub Drom HteG,…
```

en tout, 20 départements n'ont pas été placés sur le graphique, parce qu'ils se seraient superposés à d'autres points. L'utilisateur peut toutefois savoir la place de ces départements; d'une part en consultant le fichier de texte 'gazolecortx' qui en donne les coordonnées numériques, d'autre part en consultant ultérieurement les résultats de la classification automatique, qui mettront ces départements dans la même classe que d'autres effectivement représentés, dont ils sont très proches et auxquels précisément ils se seraient superposés si on avait tenté de les imprimer.

Somme toute, cette présentation graphique montre combien nous avons eu raison de demander 50 lignes et suggère peut-être que nous aurions eu avantage à en demander 70 pour avoir encore plus d'espace pour disposer les sigles des 95 départements avec le minimum d'omissions.

Quoi qu'il en soit de ces indications préliminaires, on trouve ensuite, dans un cadre où figurent deux axes à l'extrémité desquels sont les noms: axe 1 et axe 2, les sigles des mois et des départements sans superposition ni empiètement, parce que la création du fichier de texte a été faite de telle sorte que rien de tel ne se produise, avec cet inconvénient, bien sûr, que certains départements manquent. Le programme 'planF' d'affichage direct à l'écran opérait tout autrement. Il mettait sans vérification tous les sigles, quitte à créer, au centre du graphique, après l'affichage des départements, un fouillis inextricable.

3 Le programme 'CAH2' de classification ascendante hiérarchique

Les initiales CAH indiquent qu'il s'agit d'un programme de Classification Ascendante Hiérarchique. Mais, dans sa présente version, le programme a de multiples fonctions; notamment la création de tableaux d'aides à l'interprétation d'une grande utilité pour le praticien de la classification automatique.

Bien que toutes les fonctions soient intégrées dans un programme unique, nous distinguerons trois § consacrés respectivement à la CAH proprement dite, au tracé de l'arbre et aux aides à l'interprétation.

3.1 La Classification Ascendante Hiérarchique

Nous supposons dans ce § que le programme de classification n'a pas encore tourné sur les données qp5:gazole. Pourtant, on peut être amené à l'appeler une seconde fois, après un premier passage, notamment pour changer la partition choisie. Le dialogue à l'écran lors de ce second passage est un peu différent de celui que nous allons présenter ici, et fera l'objet d'une note, en fin de §, où nous indiquerons rapidement les changements.

En ouvrant le programme 'CAH2', on voit encore une fois la question:

```
le fichier de base est
```

Nous répondons qp5:gazole et l'ordinateur affiche:

```
NB il n'y a pas de classification déjà faite pour les i
NB il n'y a pas de classification déjà faite pour les j
```

Suit la demande:

```
le nom du fichier de base est-il confirmé oui (O) ou non (N)
```

Nous répondons oui.

A la question suivante:

```
faut-il classer l'ensemble des i oui (O) ou non (N)
```

nous répondons oui, et s'affichent alors trois lignes:

```
cardinal = 95
```

(cette indication rappelle le nombre d'éléments à classer)

```
le nombre des facteurs disponibles est 11
nombre de facteurs à utiliser =
```

suivi de la barre alternative invitant à répondre. Nous tapons:

```
11
```

Le programme 'CAH2' consulte alors le fichier numérique des facteurs qui a été créé par le programme 'qori' d'analyse des correspondances sans lequel le programme de classification automatique ne pourrait s'exécuter; il lit ces facteurs et, par des calculs de distances, il range les uns avec les autres les éléments qui sont le plus proches.

Sans entrer dans le détail du programme, décrivons ce qui s'affiche à l'écran.

Nous voyons défiler des lignes successives afférentes chacune à un nœud, du 1-er au 94-ème qui est le sommet de l'arbre. L'algorithme de la classification ascendante hiérarchique met d'abord deux individus ensemble pour constituer le 1-er nœud, puis deux autres individus pour constituer un 2-ème nœud, à moins que celui-ci ne soit constitué en agrégeant au 1-er nœud un 3-ème individu; et ainsi de suite. Ainsi, par 94 (nombre des départements -1) agrégations successives, se constitue la hiérarchie complète au sommet de laquelle le 94-ème nœud contient l'ensemble des départements. Le défilement des lignes que nous voyons à l'écran donne des indications sur le déroulement de cet algorithme. C'est en nous plaçant à un certain niveau dans cet arbre que nous pourrons choisir une partition de l'ensemble des départements en un ensemble de 8 ou 9 classes, à chacune desquelles sera affectée une trame dans la cartographie.

Cependant que nous exposons ces généralités, s'achève l'affichage à l'écran de la création des nœuds. Sur chaque ligne est écrit, après le n° du nœud, le nombre de maillons utilisés dans la chaîne de création (carm= ; où car est mis pour cardinal et m pour maillon), et le niveau de création du nœud (niv=). Ces indications s'adressent au statisticien, mais il importe de signaler ici que le programme de classification ascendante hiérarchique que nous utilisons procède par recherche en chaîne, pour chaque individu ou chaque nœud, de ses plus proches voisins, et que le niveau que nous voyons s'écrire indique, en bref, le degré de proximité des individus que nous avons pu agréger: plus le niveau d'agrégation est faible et plus sont proches les individus agrégés.

Voici maintenant que l'écran est occupé par un tableau constitué de blocs de 4 lignes marquées c, T, A, B:

c désigne tout simplement le n° de la classe;

T le taux d'inertie afférent au nœud auquel est constituée la classe;

A et B désignent les numéros des deux descendants de ce nœud.

```
n= 92 carm=  2  niv= 1.51e-4
n= 93 carm=  2  niv= 3.53e-4
n= 94 carm=  1  niv= 6.90e-4
c     189   188   187   186   185   184   183   182   181   180   179   178   177   176   175
T    3493  1789   762   600   350   243   227   222   172   153   112    92    87    82    75
A     186   187   182   183    20   175   165   171   177   169    29   173   167    50   163
B     188   185   184   174   180   178   172   181   176   179   161   168   170   158   166
c     174   173   172   171   170   169   168   167   166   165   164   163   162   161   160
T      72    68    66    63    63    59    55    55    49    48    43    42    41    38    33
A     164   155   152    62   151    73    56   143    11   149    80   138   126    85   139
B     136   160   145    76   154   147   159   157   146   162   156   124   153   148   144
c     159   158   157   156   155   154   153   152   151   150   149   148   147   146   145
T      30    30    30    27    26    26    26    25    24    22    20    18    18    17    17
A     130   150    86   127    15   123   142   128   137   133   132   131     4    22    94
B     134    57    35   121   140   141   111    59   120    67   101    83   125    13   114
c     144   143   142   141   140   139   138   137   136   135   134   133   132   131   130
T      16    15    15    14    13    13    13    13    12    12    11    11    11    10    10
A     109    81    69   115   110    39    31    79    51     7    65    54   119    66    40
B     129    44   116   117   135   122    14    74     2    12    34    55    97     6   106
c     129   128   127   126   125   124   123   122   121   120   119   118   117   116   115
T      10    10    10    10     9     9     9     8     8     8     8     7     7     7     7
A      47    95    27   112    48    24    61    30    10   108     1    32    68    49    63
B     118    78    60    75     5    33   102    26    77   113   107   105    18    42   103
```

Affichage du tableau des résultats de la CAH des départements

Sans reprendre la théorie de la classification automatique, lisons les premières lignes de ce tableau.

c = 189 est le n° de la classe constituée au sommet de l'arbre, c'est-à-dire de la classe comprenant l'ensemble des 95 départements. Le sommet, en tant que nœud, est le 94-ème nœud, mais si l'on range les classes constituées aux différents nœuds à la suite des 95 classes comprenant un seul département, la classe constituée comme 1-er nœud porte le n° 96 et ainsi de suite; la classe constituée comme 94-ème nœud, ou sommet, porte le n° 95+94= 189. Puis viennent les classes n° 188, 187, 186 etc. afférentes aux nœuds plus bas que le sommet.

L'indication T est très importante. Nous voyons apparaître, sous 189, le nombre 3493; sous 188, 1789; et ensuite, des nombres tous inférieurs à 1000. En fait, ce T représente un taux de représentation de l'information, en dix millièmes. La mention T= 3315 signifie donc, en bref, que 1/3 de l'inertie, ou de l'information, contenue dans le tableau des données peut être schématisé par la partition de l'ensemble des départements en deux classes. Ces deux classes ont d'ailleurs leurs numéros dans les lignes suivantes A et B: Ce sont les classes 186 et 188 dont nous pourrons à loisir, ultérieurement, considérer le contenu. Il va sans dire que cette opposition est une opposition entre mois où la consommation de gazole liée aux activités de vacances est forte: Août (ainsi que Juillet et Juin) et mois où cette consommation est faible, à l'inverse de celle liée aux activités économiques, importante en automne et en hiver.

Il est intéressant de comparer ce taux d'environ 1/3 au taux de 70% trouvé pour le 1-er axe en analyse factorielle. Le 1-er axe représente une opposition graduée sur l'ensemble des départements. Il donne donc plus d'information que ne peut en donner une simple coupure en deux classes: départements du côté du mois d'Août et autres départements.

Dans la suite du programme, l'utilisateur aura à choisir le nombre de classes de la partition désirée. La décroissance des taux peut guider, dans une certaine mesure, son choix.

Sous le tableau, on voit la barre alternative qui demande à l'utilisateur d'entrer un caractère quelconque quand il aura fini de consulter le tableau, ce que nous faisons.

Addendum au §3.1: Second passage dans 'CAH2'

Signalons ici comment se présente le dialogue au début du programme 'CAH2', dans un second passage, après qu'une classification a déjà été effectuée pour l'ensemble i.

```
le fichier de base est
```

Réponse:

```
qp5:gazole
le nom du fichier de base est-il confirmé oui (O) ou non (N)
```

Réponse:

```
O
faut-il utiliser une classification déjà faite
pour l'ensemble i oui (O) ou non (N)
```

Si l'on répond non, le programme enchaîne comme au premier passage:

```
faut-il classer l'ensemble i oui (O) ou non (N)
cardinal= 95
```

etc.

Si l'on répond oui (il faut utiliser une classification déjà faite), le programme reprend à la question:

```
faut-il faire arbre et partition oui (O) ou non (N)
```

par laquelle débute le §3.2 qui suit.

Cette possibilité d'utiliser une classification déjà faite est intéressante dans le cas de grands ensembles de données (une CAH pour 1000 individus demande 3 heures, tandis que les calculs de partition et d'aide à l'interprétation, qui font l'objet des §§1.3.2 et 1.3.3, ne demandent que 15 minutes pour ces mêmes données. Dans le cas présent, avec 95 individus, la CAH demande 5 minutes et il importe peu qu'on la refasse ou non.

3.2 Arbre et partition

Vient alors la question:

```
faut-il faire arbre et partition des i oui (O) ou non (N)
```

Il va sans dire que nous répondons oui, car, jusqu'à présent, les calculs effectués par le programme, tout en représentant l'essentiel de la classification, n'a pourtant créé que des fichiers numériques inaccessibles à l'utilisateur.

Vient alors la question:

```
la partition à écrire est-elle définie par
les noeuds les plus hauts (H) ou par des noeuds spécifiés (S)
```

En bref, il s'agit de retenir de la classification ascendante hiérarchique, qui comporte au dessus des 95 éléments (départements), 94 nœuds, une sous-hiérarchie à la base de laquelle se trouvera une partition, par ex. en 9 classes, avec, au dessus de ces 9 classes, des réunions successives en classes supérieures, jusqu'à parvenir au sommet. Pour effectuer un tel choix de façon tout à fait conséquente, il faut avoir une connaissance approfondie des résultats de la classification, et c'est pourquoi l'on prévoit de rentrer, après une première consultation , dans le programme 'CAH2 'pour demander à bon escient une sous-hiérarchie. Au point où nous en sommes, nous répondons que la partition à écrire est définie par les nœuds les plus hauts, c'est-à-dire par ceux auxquels correspondent les pourcentages d'explication les plus forts. Nous répondons:

```
H
```

Vient alors la question:

```
le nombre de classes (de 1 à 100) de la partition à écrire est
```

Nous pourrions, dans une certaine mesure, déterminer ce nombre d'après les valeurs des taux. Mais notre choix est ici guidé, en fait, par une connaissance préalable des données et aussi par la commodité cartographique. Nous demandons 9 classes:

```
9
```

Le programme va créer un fichier de type texte donnant l'arbre de cette sous-hiérarchie retenue. C'est pourquoi viennent les questions:

```
on donnera ci-après les dimensions  de l'arbre de la partition
la largeur (en caractères) choisie pour tracer l'arbre est
```

à quoi nous répondons:

```
80
```

(Si la réponse sort des limites permises, le programme la corrige automatiquement)

```
faut-il passer une ligne entre deux classes oui (O) ou non (N)
```

à quoi nous répondons non. (L'utilisateur peut répondre oui s'il a l'intention d'ajouter lui-même des commentaires sur le dessin de l'arbre).

L'ordinateur crée alors un fichier de texte et, cependant qu'il effectue cette opération, nous voyons s'afficher à l'écran:

```
identificateurs des départements
```

```
Char ChMa Cher Crrz Cors CdOr CdNo Creu Dord Doub Drom Eure EuLo Fins Gard
HGar Gers Giro Hera letU Indr letL Iser Jura Land LoCh Lore HLoi LoAt Lort
Lot LetG Lozr MetL Mche Marn HMrn Maye MetM Meus Mrbh Mose Niev Nord Oise
Orne PCal PuyD PyrA HPyr PyrO BRhi HRhi Rhon HSao SaoL Sart Savo HSav Pari
SMrt SMrn Yvel 2Sev Somm Tarn TGar Var Vauc Vend Vien HVie Vosg Yonn Belf
Esso HSei SDen VMrn VOis
faut il tracer l arbre de la CAH generale oui(O) ou non(N) O
on donnera ci apres les dimensions de l arbre de la CAH
la largeur (en caracteres) choisie pour tracer l arbre est  159
faut il passer une ligne entre 2 individus oui(O) ou non(N) N
*
faut il faire Facor pour i oui(O) ou non(N) O
le listage Facor est cree
le fichier x:gaz:gazoleiqFacww est cree
le fichier x:gaz:gazoleinFacww est cree
faut il faire Vacor pour l oui(O) ou non(N) O
NB Vacor ne sera fait que si les donnees le permettent
fin de lecture de x:gaz:gazoleicqki
fin de lecture de x:gaz:gazolekij
fin d'ecriture de x:gaz:gazolekjiq
fin de lecture de x:gaz:gazoleiFacww
fin d'ecriture de x:gaz.gazoleiVacorj
*
faut il classer l ensemble des j oui(O) ou non(N)
```

Dialogue: fin de la création de l'arbre et aides à l'interprétation

et suivent les sigles de tous les départements, ce qui est une vérification utile: le programme lit correctement le fichier contenant ces sigles.

Puis vient la question:

```
faut-il tracer l'arbre de la CAH générale oui (O) ou non (N)
```

Nous répondons oui, car la consultation de cet arbre, bien qu'il soit assez long et occupe 95 lignes de texte, est très instructive.

A la demande de l'ordinateur:

```
on donnera ci-après les dimensions de l'arbre de la CAH
la largeur (en caractères) choisie pour tracer l'arbre est
```

nous répondons en demandant la largeur maxima permise par l'imprimante Imagewriter que nous utilisons, soit 136 caractères. (En fait, si l'utilisateur demande un nombre plus élevé, l'ordinateur ramène automatiquement ce nombre à la valeur maxima permise).

Quelques secondes se passent; l'arbre est maintenant créé sur le fichier de texte (cet arbre pourra être affiché à l'écran ou imprimé dès que sera achevé le déroulement du programme 'CAH2').

3.3 Aides à l'interprétation de la Classification Ascendante Hiérarchique

Voici que s'affiche à l'écran la question:

```
faut-il faire facor pour i oui (O) ou non (N)
```

A cette question, l'utilisateur répondra oui (O); car, d'une part, les listages d'aide à l'interprétation sont d'une véritable utilité pour connaître de façon fine les caractéristiques des classes d'unités territoriales (ici, de départements) que l'on a créées; et, d'autre part, c'est en calculant ces aides à l'interprétation que le programme crée le fichier numérique des facteurs pour l'ensemble iq des centres de gravité des classes; fichier indispensable pour représenter cet ensemble dans les graphiques produits par les programmes 'planF' et 'planX', (cf. §§2.2 et 2.3), ainsi que par le programe 'carthage', (cf. [NOT. PROG. CART.], §2, présentation des résultats de l'analyse factorielle).

Nous répondons donc oui:

```
O
```

Bientôt, nous sommes avisés de la création du listage 'Facor', puis du fichier des facteurs pour les classes (iq) et les nœuds (in). Et c'est la question:

```
faut-il faire Vacor pour i oui (O) ou non (N)
```

nous répondons oui, et s'affiche la note:

```
N.B. Vacor ne sera fait que si les données le permettent
```

Expliquons les conditions auxquelles fait allusion cette note: il s'agit des dimensions du tableau de données: nombre de colonnes et nombre de lignes. Dans une classification sur l'ensemble des i, les j (dans notre cas, les 12 mois) jouent le rôle de variables explicatives; et un nombre tel que 12 n'est pas excessif.

Mais dans la classification des j, qui sera faite ensuite, les variables explicatives sont les i (95, dans notre cas): et quiconque a consulté un listage d'aide à l'interprétation sait que tant de variables ne peuvent servir ensemble à l'interprétation. C'est pourquoi on proposera, dans la suite, de faire pour j un Vacor sur iq... Et c'est précisément pourquoi il peut être bon de demander Vacor, même pour j: car, à défaut du listage Vacor, le programme créera des fichiers qui serviront à faire Vacor pour j sur iq. Pour i, Vacor est effectivement créé, avec un autre fichier, comme nous en avertissent des messages sur l'écran.

Le traitement de l'ensemble des i (ensemble des départements) est maintenant achevé: la présence de la barre alternative l'indique.

Nous tapons une étoile et entrons ce caractère en tapant Retour chariot.

Apparaît à l'écran la question:

```
faut-il classer l'ensemble j oui (O) ou non (N)
```

Sans reprendre le détail des réponses et des traitements qui sont identiques à ceux effectués pour l'ensemble des i, (à ceci près que nous devons noter qu'il n'est pas strictement indispensable à la cartographie que ces traitements soient effectués sur l'ensemble des variables, parce qu'il suffit d'avoir une classification de l'ensemble que l'on veut représenter sur la carte, c'est-à-dire

sur l'ensemble des départements), nous passons rapidement, jusqu'à voir apparaître à nouveau à l'écran des questions qui concernent l'aide à l'interprétation.

S'affiche à l'écran la phrase:

```
faut-il faire Vacor sur jq pour i oui (O) ou non (N)
```

Le programme propose ici un Vacor expliquant la hiérarchie des départements i, non par les variables j elles-mêmes (les mois), mais par les classes jq de la partition de l'ensemble des variables qu'il vient d'obtenir par classification des j (mois). Dans le cas présent, avec un ensemble de 12 variables seulement, les 12 mois de l'année, l'interprétation de la classification, et donc celle de la cartographie, directement en termes de ces variables est possible, et même facile. Mais prenons dans le même domaine un tableau à 48 colonnes: le tableau donnant les consommations mensuelles de quatre produits pétroliers. L'interprétation directe en termes de ces 48 variables devient difficile, sinon impossible. Il faut donc procéder à l'agrégation de ces variables, et, donc, considérer en quelque sorte un tableau intermédiaire donnant, par ex., au lieu des consommations en gazole, fuel lourd, fuel domestique, carburant auto mois par mois, les consommations de chacun de ces produits suivant des périodes plus longues choisies selon le produit: saison entière, moitié de l'année, même dans certains cas trois quarts de l'année opposés à la période d'été etc.; en sorte que, finalement, placé en présence d'un tableau d'une dizaine de colonnes, et de pourcentages calculés sur ce tableau par le programme d'aide à l'interprétation, ainsi que de graphiques plans où s'affichent les sigles de ces agrégats de variables, l'utilisateur comprenne sans peine l'interprétation des classes de départements créées.

Cette expression "faire Vacor sur jq pour i" signifie donc, tout simplement, en bref, représenter, interpréter le tableau dont les lignes sont les individus, les départements i, et dont les colonnes sont, non les 12 mois, mais, dans le cas présent, selon les demandes qui ont été faites, 6 classes de mois reconnues pertinentes par classification ascendante hiérarchique.

Nous répondons non:

```
N
```

De la même manière vient maintenant la question:

```
faut-il faire Vacor sur iq pour j oui (O) ou non (N)
```

Cette question est évidemment ici tout-à-fait secondaire, la représentation de l'ensemble j étant claire. Toutefois, si l'on voulait interpréter les classes de l'ensemble j, il est certain que (comme on l'a dit plus haut) cette interprétation ne pourrait se faire directement en termes de départements, mais en termes de groupes de départements, ce qui justifie alors la considération d'un tableau dont les colonnes sont les j eux-mêmes (les colonnes primitives du tableau de

données) mais dont les lignes sont réalisées par cumul de départements suivant les classes créées par la CAH.

3.4 Conclusion de la notice du programme 'CAH2'

Outre les fichiers numériques dont certains sont indispensables au déroulement du programme 'carthage', le programme 'CAH2' a créé des fichiers de type texte contenant d'une part l'arbre de la classification ascendante hiérarchique, et, d'autre part, les listages Vacor et Facor. Nous ne donnerons pas dans cette notice de commentaires sur ces listages, renvoyant comme précédemment l'utilisateur à ce qui est publié, par ex. dans les volumes de la série: *Pratique de l'Analyse des Données*, ou dans des articles équivalents de la revue *Cahiers de l'Analyse des Données*.

N.B. Les programmes décrits dans la présente notice font partie du Logiciel **Mac-SAIF**, ('Système d'Analyse des InFormations'); les lecteurs désireux d'acquérir ce Logiciel s'adresseront à la

Société **STATMATIC**: 4, rue de Fécamp 75012 Paris;
Téléphone: (16.1) 43.42.48.19 / (16.1) 47.98.77.39