LES CAHIERS DE L'ANALYSE DES DONNÉES

L. LEBART
A. MORINEAU

Présentation d'une bibliothèque modulaire de programmes pour l'analyse de grands tableaux

Les cahiers de l'analyse des données, tome 3, n° 3 (1978), p. 269-274

<http://www.numdam.org/item?id=CAD_1978__3_3_269_0>

Les Cahiers de l'Analyse des Données
Vol. III - 1978 - n° 3 - p. 269-274

PRÉSENTATION D'UNE BIBLIOTHÈQUE MODULAIRE DE PROGRAMMES POUR L'ANALYSE DE GRANDS TABLEAUX [BIBL. PROG.]

par L. Lebart (1)

et A. Morineau (2)

I *Introduction*

Un tableau de données n'est grand que mesuré à l'échelle de l'instrument qui doit le traiter. Grand pour l'homme non armé d'outils de calculs, un tableau à 15 lignes et 10 colonnes est petit pour un mini-ordinateur. Pour les calculateurs déjà importants, un tableau ayant quelques milliers de lignes et quelques centaines de colonnes est sans doute un *grand* tableau. D'un autre côté il n'y aura pas de tableaux *trop* grands pour être analysables : ainsi l'étude d'un fichier de plusieurs dizaines de milliers d'observations pourra en général être *approchée* de façon satisfaisante par traitement sur un calculateur de dimension moyenne d'un *échantillon* réduit à quelques milliers d'observations.

Pour le praticien le traitement d'un fichier numérique important pose des problèmes spécifiques, d'une part au niveau des *calculs*, d'autre part au niveau des *éditions* de résultats. En ce qui concerne le traitement numérique, il faut choisir les méthodes et utiliser les algorithmes dont l'exécution s'effectue en une durée raisonnable, tout en n'occupant en mémoire centrale qu'une place restreinte. Le compromis à rechercher entre le temps d'exécution et l'occupation de la mémoire est compliqué enfin par le souci de conserver une bonne précision malgré le volume des calculs à exécuter.

Pour ce qui concerne les *éditions de résultats*, il importe que le statisticien puisse retrouver et identifier aisément malgré leur abondance les individus, les variables ou les divers caractères décrits numériquement dans son fichier initial. Plus le tableau à analyser est important, plus on doit s'ingénier à fournir des illustrations et des aides à l'interprétation, claires et faciles à décrypter. On trouvera intérêt en particulier à élaborer un *dictionnaire* précis des libellés des éléments, même si la gestion de ce lexique entraîne un surcroît de travail au départ. Enfin la lisibilité des illustrations sera obtenue en sélectionnant au moment de la représentation les éléments les plus "significatifs" (par exemple en ne représentant sur un facteur que les éléments ayant les plus fortes contributions), ou en regroupant ces éléments selon leur origine naturelle (par exemple en rangeant par question les éditions des contributions des modalités de réponse dans un questionnaire).

Le plus souvent les grands tableaux que nous avons en vue seront des recueils de *données individuelles* du type des données d'enquête. Ces données étant caractérisées à la fois par une grande variabilité au niveau des observations élémentaires, et par un réseau complexe de dépendance, il est impossible d'extraire "à vue" ou de façon directe les traits essentiels de l'information qu'elles peuvent contenir. Dans ce cas précis, des algorithmes adaptés à la taille des fichiers, les méthodes d'*analyse factorielle* et certaines techniques de *classification* constituent des outils irremplaçables de réduction et de visualisation.

(*) Publiée dans "Technique de la description statistique" (Méthodes et logiciel pour l'analyse de grands tableaux) L. Lebart, A. Morineau, N. Tabard - DUNOD 1977.

(1) Maître de recherches C.N.R.S.

(2) Chargé de recherches CEPREMAP

II Conception d'une bibliothèque modulaire

Les systèmes intégrés de traitements statistiques, où la commande des analyses est gérée par un "méta-langage" propre au système, ne laissent en général pas d'autre initiative à l'utilisateur que le choix parmi les options prévues par le fabriquant. Notre conception d'une programmathèque tend au contraire à donner à l'utilisateur l'autonomie la plus étendue possible. On suppose donc de celui-ci une compétence certaine dans le domaine de la statistique (et peut-être aussi dans le domaine informatique), la programmathèque étant au statisticien ce que la boîte à outils est à l'artisan.

Le logiciel, pour sa partie publiée et actuellement susceptible d'être diffusée, se compose de 68 sous- programmes et 7 exemples de programmes d'appel exécutant des analyses complètes. En volume il représente environ 5000 instructions FORTRAN. Nous allons voir comment les préoccupations évoquées ici se traduisent matériellement dans l'écriture de ces programmes.

Les 7 programmes principaux ne sont que des exemples des assemblages possibles des divers modules. Il a paru opportun d'effectuer du logiciel un découpage en sous-programmes qui donne à chacun de ceux-ci une fonction clairement définie ; d'utiliser du premier au dernier maillon de la chaîne le même système de notations ; et enfin d'accompagner chaque opération de commentaires précis sur les fonctions exécutées ou les étapes d'un algorithme. A l'utilisateur averti, il est possible dans ces conditions de réaliser un nouvel assemblage des modules existant, de remplacer un module ou un algorithme par un autre de son cru, de segmenter où il lui convient une série de calculs, de réécrire certaines procédures dans un autre langage, en bref de travailler de façon autonome et consciente, en adaptant la chaîne des opérations aux situations typiques rencontrées dans son laboratoire.

Pour chacun donc ce logiciel est susceptible de s'enrichir et de se développer dans des directions propres, les 68 sous-programmes initiaux constituant si l'on veut une base de départ. Il peut également être optimisé en fonction du matériel utilisé, étant écrit au départ en un FORTRAN "minimal" qui le rend "transportable".

III Description générale

L'utilisateur doit pouvoir trouver dans sa bibliothèque des programmes, rangés sur divers rayons, un certain nombre d'utilitaires généraux susceptibles de participer à de multiples assemblages : les principales procédures utiles à la gestion simultanée de son fichier de données et de leur dictionnaire ; les programmes essentiels d'éditions graphiques et de tabulation ; les algorithmes fondamentaux de calcul numérique, de statistique classique et de simulation - les algorithmes et leur écriture, conçus pour le traitement des grands tableaux, ne sont pas naturellement définitifs mais doivent pouvoir être échangés contre des programmes plus performants dès l'apparition de ceux-ci. On donnera enfin quelques indications générales sur l'organisation d'une analyse complète, dont le paragraphe suivant fournira sept exemples.

III.1 Les utilitaires généraux

a) Procédures de gestion de fichier : La manipulation d'informations individuelles volumineuses implique des éditions claires et explicites. Lors de l'utilisation de données individuelles, on utilisera un dictionnaire normalisé de libellé. Une variable numérique aura un titre (jusqu'à 40 caractères) ; une variable nominale (dont les valeurs sont les numéros de modalité de réponse) aura un titre analogue, mais chaque modalité de réponse pourra avoir un libellé ayant jusqu'à 20 caractères (procédure LEXIQ). Ceci permettra d'éditer des tabulations aisément lisibles, et d'obtenir des classements des libellés en clair selon les contributions des modalités correspondantes aux axes factoriels (procédure TEXTE).

b) Utilitaires graphiques : L'utilitaire FPLAN effectue la représentation d'objets identifiés par trois caractères dans un plan cartésien. Il contient une procédure EPURE qui ramène sur le cadre les points situés loin du centre évitant ainsi que les points isolés ne déterminent une échelle aberrante pour la représentation. Suivant la demande, le programme peut identifier les points avec un seul caractère ; d'où des graphiques de "densité" : les points sont repérés par le même caractère (graphique de densité simple), ou par des caractères différents identifiant une classe d'appartenance (illustration d'une classification, ou illustration d'un plan factoriel à l'aide d'une variable qualitative, etc...). L'appel des graphiques de densité peut être géré par le programme DENSI, qui lit une carte de commande par graphique demandé.

Dans le cas d'une correspondance multiple l'appel d'un graphique peut être précédé par l'appel de la procédure SELEC qui sélectionne pour la représentation les points ayant les plus fortes contributions sur les premiers axes.

On rangera également sur ce rayon l'utilitaire HISTZ qui engendre, en une seule lecture du fichier, tous les histogrammes des variables continues, et fournit une description élémentaire classique pour chacune d'elles.

c) Calcul numérique et combinatoire : Diagonalisation d'une matrice symétrique (S/P VPROP). On utilise une des méthodes les plus performantes actuellement disponibles pour diagonaliser une matrice symétrique (stockée en mémoire centrale) : tridiagonalisation selon la méthode de HOUSEHOLDER, puis calcul des éléments propres de cette matrice par la méthode "QL implicite". Tel qu'il est écrit, le programme est accepté sans modification par la plupart des compilateurs.

Orthonormalisation d'un tableau (S/P GSMOD). On détermine une base orthonormale du sous-espace engendré par les colonnes d'un tableau en utilisant la méthode de GRAM-SCHMIDT "modifiée", numériquement plus stable que la méthode usuelle.

Rangement des éléments d'un vecteur (S/P SHELK). Ici comme pour la plupart des algorithmes, il n'y a pas de méthode universellement optimale. L'algorithme de SHELL a l'avantage d'être simple, compact et rapide.

Permutation "in situ" d'un tableau (S/P PERMX). L'algorithme permet de permuter les lignes et les colonnes d'un tableau en minimisant les opérations d'affectation, et sans utiliser de tableaux auxiliaires. On l'appelle par exemple pour ranger une matrice de corrélations dans l'ordre des variables projetées sur un premier axe factoriel, ou un tableau de correspondances en réordonnant les lignes et les colonnes suivant le premier axe, (ce qui constitue souvent une précieuse aide à l'interprétation).

d) Utilitaires de "probabilités". On range sur ce rayon divers utilitaires du domaine de la statistique ou des probabilités. Suivant la commande, la *fonction* RNKSF retourne la probabilité d'être inférieur à une valeur donnée pour une variable suivant une loi normale, ou une loi du χ^2, de FISHER, ou de STUDENT. Le *programme* PICON calcule la statistique du χ^2 sur un tableau de contingence, la probabilité de dépasser cette valeur, et le nombre de cases où l'effectif théorique est inférieur à un seuil déterminé. La *fonction* PINLG retourne la "probabilité inverse" d'une loi normale ; on l'utilise par exemple pour effectuer des analyses en "composantes robustes", en remplaçant chaque observation brute par l'espérance de la variable normale ayant même rang dans l'échantillon. A noter encore diverses *fonctions de tirages pseudo-aléatoires* : loi uniforme, loi normale, simulation d'un tableau de correspondances. Ces fonctions ne sont pas optimisées mais fournissent les mêmes séries quelle que soit la machine ; de plus les séries obtenues jouissent de bonnes qualités.

III.2 Organisation d'une analyse

Le programme principal : Il effectue la réservation globale de mémoire centrale utile, en fonction des dimensions du tableau à analyser. Un seul vecteur est dimensionné, à l'intérieur duquel on localise les divers tableaux nécessaires à l'aide d'un calcul simple d'indices. Le programme appelle ensuite la lecture des données, leur recodage éventuel et leur écriture par ligne sur une mémoire auxiliaire (S/P DONN x). Le fichier ainsi créé est ensuite traité par un "exécutant" (S/P DCAL x).

Préparation du fichier normalisé : Le fichier est "normalisé" lorsque le dictionnaire standard des variables est créé (S/P LEXIQ), et après épuration éventuelle des données (par la procédure SAVON, qui élimine les modalités à faibles effectifs et remet à jour le dictionnaire en conséquence).

Appel de l'exécutant : C'est le véritable programme principal de l'analyse ; on y trouve, en dimensions paramétrées, la liste de tous les tableaux utiles en mémoire centrale pour l'exécution. Il comprend deux phases plus ou moins séparables : la phase *calcul* (diagonalisation et calcul des coordonnées sur les axes pour une analyse factorielle ; construction des classes pour une partition) ; et la phase des *éditions* illustrant les résultats (contributions, graphiques, tabulations, etc...).

IV Description de sept pré-assemblages des modules

IV.1 Analyse en composantes principales (ACOMP)

Les données sont lues avec les libellés nécessaires aux éditions. On transforme éventuellement les observations en rangs (S/P SHELK), et les rangs sont "normalisés" (fonction PINLG) si on désire effectuer une analyse en "composantes robustes". Les sorties de ACOMP peuvent comprendre simultanément les éditions suivantes :

- dictionnaire des libellés des variables nominales et continues (LEXIQ)
- moyennes, écart-types, *extrema* des variables continues
- matrice des corrélations des variables actives
- valeurs propres, histogramme, pourcentages
- matrice des corrélations des variables actives rangées suivant le premier facteur (PERMX)
- coordonnées des variables actives et supplémentaires, et (éventuellement) des individus actifs et supplémentaires
- classement des libellés en clair (variables et individus), suivant chacun des six premiers facteurs
- graphiques des plans factoriels (FPLAN)
- partition (agrégation autour de centres variables, avec groupements stables) des individus d'après leurs coordonnées sur les premiers facteurs, et graphique de densité représentant le nuage des individus dans le (ou les) premiers plans factoriels, identifiés par un caractère représentant une classe de la partition (PARTS)
- graphiques de densité complémentaires, identifiant les individus par les modalités d'une des variables nominales (DENSI).

IV.2 Analyse des correspondances binaires (sur tableau de contingence) (ACOBI)

Après l'édition des *taux d'inertie* associés à chaque axe factoriel, on calcule les *contributions absolues* et *relatives* pour chaque élément. Sur *chaque facteur séparément* on effectue la projection simultanée des points des deux nuages, chaque élément étant repéré par un libellé complet. On réalise ensuite les graphiques-plans des projections des deux nuages sur deux facteurs consécutifs, où les points lignes et colonnes sont repérés par un nom à trois caractères (programme FPLAN).

Si le tableau des effectifs est de dimension acceptable pour entrer en mémoire centrale de l'ordinateur, on peut enrichir l'interprétation des résultats en effectuant le rangement de ses lignes et de ses colonnes selon l'ordre des points sur le premier facteur (programme PERMX). L'édition des profils mettra en évidence des blocs qui éclairent la nature de la *correspondance* entre les deux ensembles. Eventuellement, on peut obtenir une *classification* des points-colonnes par la méthode des "nuées dynamiques" appliquée aux coordonnées sur les facteurs (programme PARTS). Ces classifications sont illustrées par des graphiques de densité en repérant chaque point par le numéro de sa classe.

IV.3 Analyse des correspondances multiples (MULTM, MULTS, MULTK, KMULT)

L'analyse des tableaux mis sous forme disjonctive complète peut être réalisée par le programme décrit en IV.2. Au niveau des calculs et des éditions, il est cependant beaucoup plus intéressant d'utiliser des procédures spécifiques. Les calculs s'effectuent directement sur le codage réduit (numéros de modalité) et non sur le codage binaire obtenu après éclatement des variables. Le fichier réduit réside sur un support périphérique (disque, bande, cassette).

MULTM réalise une analyse complète, avec réduction de la matrice à diagonaliser par utilisation du codage particulier des modalités. MULTK réalise la même analyse avec des éditions abrégées, puis fait une partition sur les coordonnées factorielles avec calcul de groupements stables (S/P PARTS), et croise cette partition avec toutes les variables actives et illustratives du fichier. MULTS réalise la même analyse que MULTM, mais la diagonalisation s'effectue par *approximation stochastique* sans utilisation importante de mémoire centrale. Ainsi, pour un fichier donnant les 600 modalités de réponses à 100 questions de 2000 individus, l'encombrement mémoire est de 18.600, contre plus de 400.000 pour un programme conventionnel. Les temps de calcul ne sont guère faciles à comparer, car ils dépendent linéairement du nombre de questions, et du nombre d'individus (et non du carré du nombre de modalités, comme les programmes usuels).

Pour le fichier cité ci-dessus, le temps de calcul est sensiblement inférieur à celui de la procédure MULTM, elle-même plus rapide qu'un programme classique. Enfin KMULT effectue une partition *préalable* des individus, et ensuite analyse le tableau agrégé. (Les deux derniers programmes, travaillant en lecture directe ligne-à-ligne sur le fichier des données, permettent de traiter des données très vastes sur des ordinateurs de dimension moyenne).

A titre d'exemple, donnons quelques indications complémentaires concernant les programmes MULTM et KMULT.

Programme MULTM : La lecture se fait dans un sous-programme isolé (DONNB) qui peut aisément être modifié pour permettre éventuellement des lectures simultanées de plusieurs supports ainsi que des recodages ; le dictionnaire est lu, puis remis à jour quand le fichier est modifié (suppression des modalités à faibles effectifs, qui réapparaîtront en variables supplémentaires : procédures LEXIQ, SAVON).

Désignons par Q le nombre de questions actives et J le nombre cumulé de modalités de réponses correspondant. Le tableau de Burt, noté B est calculé à partir du codage réduit (temps d'exécution proportionnel à Q^2, et non à J^2), et éventuellement édité. La matrice à diagonaliser (B prémultipliée par $(\text{diag } B)^{-1}$, où diag B s'obtient en annulant les éléments non-diagonaux de B) mérite ici d'être éditée, car elle est formée des Q(Q-1)/2 tableaux des profils-lignes, et des Q(Q-1)/2 tableaux de profils-colonnes correspondant aux divers tableaux de contingence qui composent B (tableaux habituellement consultés lors d'une analyse manuelle). La matrice à diagonaliser d'ordre (J,J) est ramenée par projection à une matrice d'ordre (J-Q , J-Q).

Les coordonnées et les contributions sont regroupées par questions (décrites par leur intitulé en clair) pour lesquelles sont calculées des

contributions absolues cumulées. SELEC sélectionne les points les mieux représentés. On calcule les coordonnées des individus supplémentaires et celles des variables supplémentaires. Enfin TEXTE édite les classements des libellés complets des réponses sur les premiers facteurs, qui permettent en particulier de guider la lecture des graphiques-plans réalisé par FPLAN. Ainsi sont évités, dans le cas de plusieurs centaines de modalités, les graphiques surchargés et illisibles.

Programme KMULT : On veut classer un ensemble d'individus caractérisés par des variables qualitatives. Le programme KMULT effectue la partition avec recherche des groupements stables, en travaillant en lecture directe ligne-à-ligne du fichier. Les groupements stables (ou formes fortes selon la terminologie de E. Diday) sont les groupes d'individus toujours classés ensemble lors d'une série de partitionnement obtenus à partir de germes aléatoires différents. (L'algorithme est une adaptation de PARTS au codage disjonctif et à la distance du χ^2).

Le tableau obtenu en agrégeant les lignes du fichier initial suivant cette partition est soumis à l'analyse des correspondances (programme allégé CPLUM). On obtient une estimation des axes factoriels, sur lesquels on projette les individus ce qui donne lieu à un graphique de densité.

Ce programme KMULT tient donc lieu de programme de partition, l'analyse des correspondances du tableau agrégé permettant une représentation simultanée très parlante des classes et des variables de base. Cependant la partition pourra, dans certains cas, n'être qu'un intermédiaire de calcul, l'objectif final étant une analyse approchée d'un tableau trop grand pour être traité directement.

IV.4 Description élémentaire d'un fichier de données (DSCRI)

Le programme DSCRI permet d'obtenir une description élémentaire des données (calculs de moyennes, corrélations, édition d'histogrammes), des tabulations relativement élaborées (utilisant jusqu'à 7 modalités non consécutives de deux variables *filtres*, soit en inclusion, soit en exclusion), et des graphiques de densité. L'ensemble de ces opérations s'applique au fichier standard dont le dictionnaire est lu par LEXIQ.

On utilise DSCRI en amont d'une analyse (aide à la préparation et au codage des données) ou en aval (vérifications d'associations par tabulation, graphiques de densité sur les plans factoriels, etc...).

N.B. : Le programme KMULT réalise ce que d'autres auteurs appellent une Analyse en Données Groupées. Sur ce point, cf l'article de J.P. Briane, J.J. Lazare et R. Salanon dans les Cahiers, n° 2, 1978, pp 167 - 173.

Adapté au codage disjonctif, KMULT utilise pour l'agrégation autour de centres mobiles les numéros de modalités et non le codage binaire éclaté : le temps de calcul ne dépend que du nombre Q de questions, et non du nombre J de modalités de réponse.