BUREAU DE RECHERCHES GÉOLOGIQUES ET MINIÈRES

MÉTHODES STATISTIQUES ET PROGRAMMES DE TRAITEMENT PAR ORDINATEUR DES DONNÉES NUMÉRIQUES EN GÉOLOGIE

(Texte)

par

J.C. LABROT, Ph. ROLET et P. SOLETY

Département INFORMATIQUE
B.P. 6009 – 45 Orléans (02) – Tél.: (38) 66.06.60

71 INF 001

Juillet 1971

METHODES STATISTIQUES ET PROGRAMMES

DE TRAITEMENT PAR ORDINATEUR

DES DONNEES NUMERIQUES EN GEOLOGIE

TABLE DES MATIERES

RESUME

Le présent rapport comprend trois parties :

1 - Description et classification des principales méthodes de traitement statistique des données numériques en Géologie : méthodes élémentaires d'analyse à une ou deux dimensions ; méthodes multivariables d'analyse de données à plusieurs dimensions (analyses en composantes principales et en correspondances) méthodes multivariables de prévision par modèle linéaires(régression et analyse factorielle en facteurs communs et spécifiques).

2 - Description de l'organisation et du fonctionnement de principe de la chaîne de programmes correspondants,mis au point pour ordinateur I.B.M.1130 et prochainement pour ordinateur I.B.M. 360 - 40.

3 - Illustrations des possibilités offertes aux utilisateurs à l'aide d'une série d'exemples pris dans différentes disciplines de la Géologie (Géochimie, hydrochimie, pétrographie, sédimentologie).

AVANT PROPOS

L'emploi des méthodes statistiques de traitement multivariable tend à se généraliser dans les différents domaines de la Géologie.

De nombreuses publications (ouvrages généraux de statistique appliquée à la Géologie, notes ou communications diverses) sont diffusées.

Elles se caractérisent le plus souvent par leur manque de rigueur dans l'exposé des méthodes utilisées et fréquemment par la confusion qu' elles introduisent entre les fondements des modèles théoriques.

Sans prétendre à l'originalité, ce rapport tente, en première partie, de classer les méthodes d'après leur principe et leur domaine d'emploi.

Dans une deuxième partie,il décrit l'organisation et le fonctionnement schématique de la chaîne de programmes mis au point pour permettre l'emploi et la généralisation des méthodes de traitement multivariable.

Enfin,dans la troisième partie et à titre d'illustration des possibilités offertes aux utilisateurs,il donne un éventail de cas concrets réellement étudiés.

Remerciements :

La méthode d'Analyse Factorielle des correspondances, dont un exposé est donné dans ce rapport, a été mise au point par le LABORATOIRE DE STATISTIQUE DE L'UNIVERSITE DE PARIS dirigé par Monsieur le Professeur BENZECRI.

Nous tenons à remercier vivement Monsieur J.P. BENZECRI et Monsieur P.CAZES, assistant au Laboratoire, d'avoir bien voulu nous guider dans la connaissance et l'utilisation de cette méthode et de l'ensemble des méthodes d'analyse de données.

INTRODUCTION

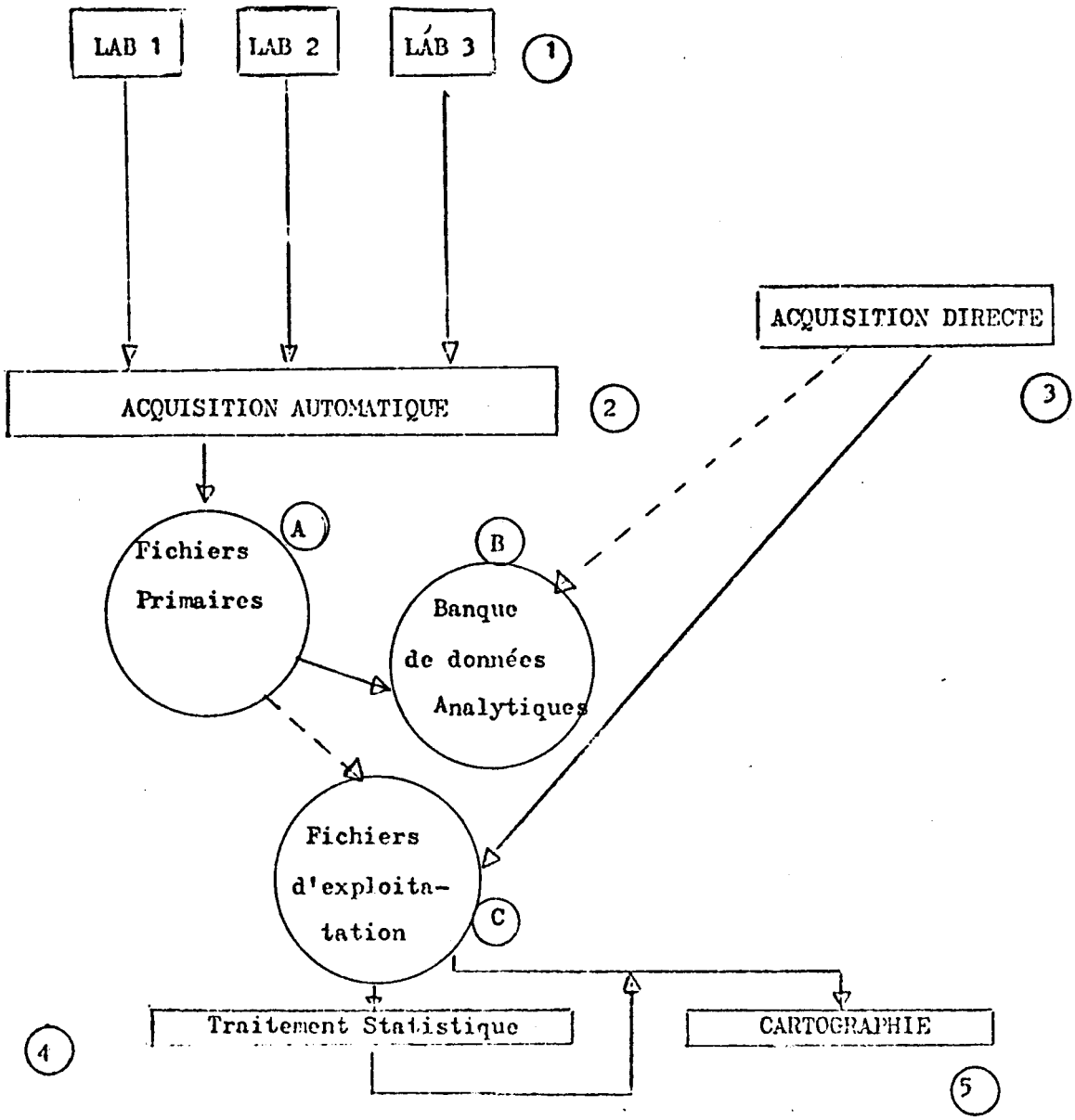
INTERET ET NECESSITE DES METHODES UTILISANT DES PROGRAMMES EN ORDINATEUR

Ce rapport fait le point des études réalisées par le Département Informatique du B.R.G.M. depuis un peu plus d'un an dans le cadre de son opération d'acquisition et traitement automatiques des données de Laboratoire.

Il concerne uniquement le traitement statistique des données ; l'acquisition automatique des résultats de dosage et la représentation cartographique par tracé de courbes isovaleurs en sont exclues.

L'organisation de principe de l'opération peut être représentée par le schéma ci-dessous.

Le rapport traite des fonctions 3 et 4 et des fichiers C



Volume des données :

Le B.R.G.M. réalise annuellement dans ses laboratoires d'ORLEANS plus de 150 000 déterminations analytiques élémentaires sur des échantillons de nature variée (roches, minerais, eaux....).

Le développement de l'acquisition automatique des données en sortie des appareils d'analyse (spectrométrie optique d'absorption, spectrométrie d'émission, spectrométrie par fluorescence X) permet de dépouiller et de conserver simultanément sur support magnétique les résultats de dosage qui sont dès lors disponibles pour traitement en ordinateur.

Les données ne provenant pas des laboratoires automatisés peuvent évidemment être acquises directement sur cartes perforées puis transférées sur support magnétique.

Tout ensemble de résultats analytiques peut se mettre sous la forme d'un tableau de valeurs numériques comprenant autant de lignes(n) que d'échantillons analysés (ou observations) et autant de colonnes (p) que d'éléments dosés (ou variables mesurées) pour n = 500 et p = 20, ce tableau renferme 10 000 nombres.

Les 10 000 nombres représentent le résultat principal d'un travail préalable long et onéreux qui a consisté à localiser sur le terrain, prélever, enregistrer et analyser les 500 échantillons faisant l'objet de l'étude.

L'information qu'ils renferment n'est pas directement accessible.

Il est bien sûr possible d'en extraire des informations partielles en étudiant la distribution des prélèvements en fonction d'un, voire de deux éléments ou en se guidant sur son intuition ou son expérience.

Mais l'intérêt de méthodes d'études globales et indépendantes de la nature physique des résultats à dépouiller est de permettre une approche systématique et synthétique du problème posé.

L'intérêt de telles méthodes devient nécessité lorsque le nombre des données augmente ; leur emploi implique alors un traitement en ordinateur à l'aide de programmes généraux permettant leur utilisation en routine.

Coût du traitement :

Le traitement par ordinateur d'un tableau de 500 échantillons dosés pour 20 élémonts peut être estimé à environ 4 000 F soit 8 F par échantillon.

On peut comparer co chiffre au coût de prélèvement et d'analyse envisagés dans le cadre d'une prospection géochimique : Le coût d'analyse se monte en moyenne à 40 F ;

Le coût de prélèvement peut-être estimé à environ 40 F en EUROPE et 200 F en pays d'outre-mer d'accès difficile. Le coût du traitement par ordinateur s'élève donc respectivement à 10% et 3,3% du coût (prélevement + analyse)

A titre prévisionnel, on peut retenir un chiffre compris entre 5 et 10% du coût total estimé de la campagne d'étude projetée.

Cette dépense permet non seulement d'obtenir les renseignements résultant d'un traitement classique, mais encore de fournir des informations impossibles à dégager sans le recours aux méthodes et programmes de traitement par ordinateur.

---

* Dans les conditions du centre de calcul d'Orléans (I.B.M.1130).

PREMIERE PARTIE

PRESENTATION GENERALE DES METHODES ET PROGRAMMES

1. Représentation des échantillons et des éléments :

Dans le tableau des résultats de mesure, les échantillons (ou observations) sont mis en ligne et caractérisés par l'indice i variant de 1 à n (n = nombre total d'échantillons) ; les éléments dosés (ou variables mesurées) sont mis en colonne et caractérisés par l'indice j variant de 1 à p (p = nombre total de dosages) .

Ce tableau peut être représenté par une matrice X d'élément général x (i,j), ainsi x (i,j) veut dire : teneur de l'échantillon n° i en élément n° j.

Dans la suite, on donnera la même désignation au tableau et à la matrice associée.

	1	2	- - - - -	j	- - - - -	p
1	x(1,1)	x(1,2)		x(1,j)		x(1,p)
2	x(2,1)	x(2,2)		x(2,j)		x(2,p)
"						
i	x(i,1)	x(i,2)		x(i,j)		x(i,p)
"						
n	x(n,1)	x(n,2)		x(n,j)		x(n,p)

fig 1.1.

Tableau des résultats de mesure représenté par une matrice X d'élément général x(i,j).

De la façon la plus générale, tout échantillon i peut être représenté par un vecteur à P composantes de l'espace R^p (espace des éléments) et tout élément j peut être représenté par un vecteur à n composantes de l'espace R^n (espace des échantillons).

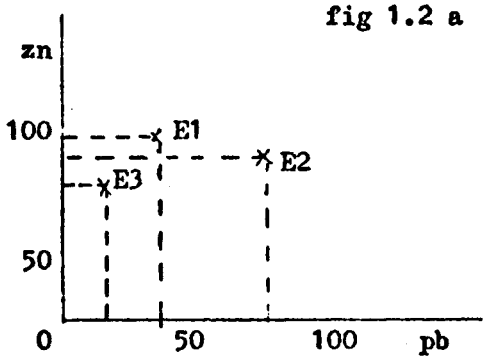
Exemple : Le tableau ci-dessous à 3 échantillons dosés pour 2 éléments :

	pb	zn
E1	50	100
E2	80	90
E3	30	70

a pour matrice associée : $\begin{bmatrix} 50 & 100 \\ 80 & 90 \\ 30 & 70 \end{bmatrix}$

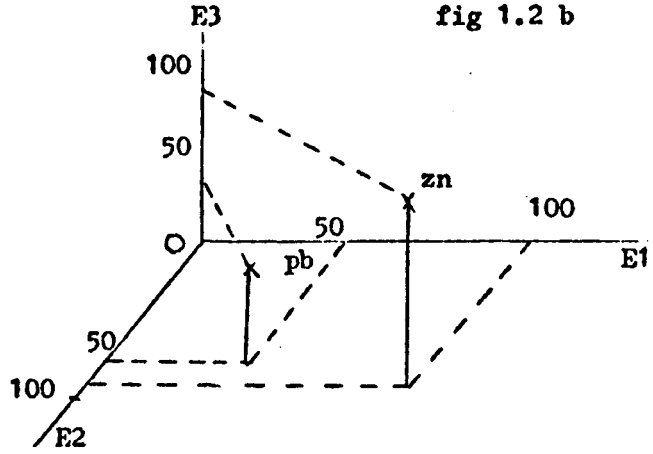
et peut donner lieu aux 2 représentations suivantes :

fig 1.2 a



3 points dans R^2
Représentation des échantillons dans l'espace des éléments

fig 1.2 b



2 points dans R^3
Représentation des éléments dans l'espace des échantillons

Dans la pratique courante, on s'intéresse généralement aux relations (entre éléments ou entre échantillons) existant dans des sous-espaces à 1,2 ou 3 dimensions des espaces de référence R^p ou R^n.

Les méthodes de traitement multivariables permettent d'envisager globalement les relations existant directement dans R^p ou R^n. En langage imagé (donc quelque peu incorrect) , on peut dire que ces méthodes fournissent un moyen de " voir " dans un espace à plusieurs dimensions (20 ou 500 dans l'exemple précité suivant que l'on étudie les vecteurs échantillons dans R^p ou les vecteurs éléments dans R^n).

2. Classification pratique des méthodes et programmes de traitement.

Cette classification est fondée sur leur domaine d'emploi et distingue.

1. Les méthodes descriptives élémentaires.
2. Les méthodes descriptives multivariables.
3. Les méthodes de prévision (recherche de modèle).

2.1. Méthodes descriptives élémentaires.

Elles sont exécutables à la main (avec l'aide souhaitable d'une calculatrice de bureau). Les programmes de traitement permettent seulement de les automatiser et donc de généraliser leur emploi :

Elles comprennent :

- calcul de moyennes et écarts-types par élément ;
- tracé des histogrammes de répartition par élément ;
- calcul des coéfficients de corrélation des éléments deux à deux (matrice de corrélation) ;
- tracé des diagrammes de répartition des éléments deux à deux pour visualiser des relations éventuelles ;
- tracé des diagrammes ternaires (utilisés essentiellement en pétrologie).

2.2. Méthodes descriptives multivariables.

Elles font nécessairement appel à l'ordinateur et permettent de décrire la forme du nuage des points échantillons dans l'espace des éléments (n points dans R^p) et la forme du nuage des points éléments dans l'espace des échantillons (p points dans R^n) ;
elles décrivent parrallèlement les positions respectives des points échantillons entre eux (dans R^p) et les positions respectives des points éléments entre eux (dans R^n).

2.3. Méthodes de prévision.

Elles font également appel à l'ordinateur. Par opposition avec les méthodes descriptives, elles supposent à priori l'existence d'une relation causale de forme linéaire entre tout ou partie des variables de départ et des " facteurs " *, qui sont soit d'autres variables brutes soit de nouvelles variables.

---

* NB. On définira ultérieurement la signification mathématique et géologique du terme facteur (1° partie - §5).

3. Les Méthodes descriptives élémentaires.

Elles permettent d'étudier la distribution d'un élément j pris parmi les p éléments dosés ou les relations entre les distributions d'un couple de deux éléments j et l pris parmi les p éléments.

a - Pour tout élément j, l'ordinateur calcule la moyenne et l'écart-type empiriques de sa distribution et trace l'histogramme de répartition.

b - Pour tout couple d'élément (k,l), l'ordinateur calcule le coéfficient de corrélation linéaire et trace à l'imprimante le diagramme de répartition correspondant (k en fonction de l ou l en fonction de k).

Ce diagramme est important car le coéfficient de corrélation linéaire peut ne résumer qu'imparfaitement la liaison entre deux distributions puisqu'il suppose l'existence d'une relation théorique linéaire et monotone entre les deux variables correspondantes.

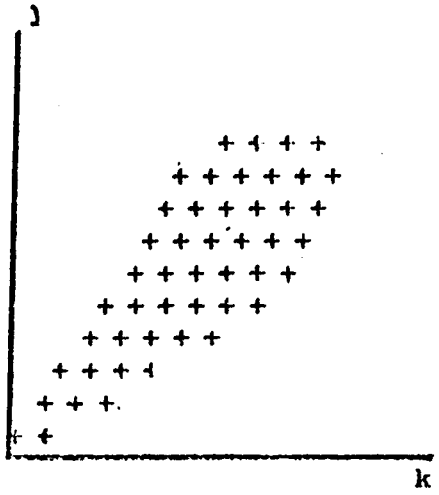


fig 3.1 a

relation linéaire monotone
le coéfficient de corrélation
résume bien la liaison entre
k et l

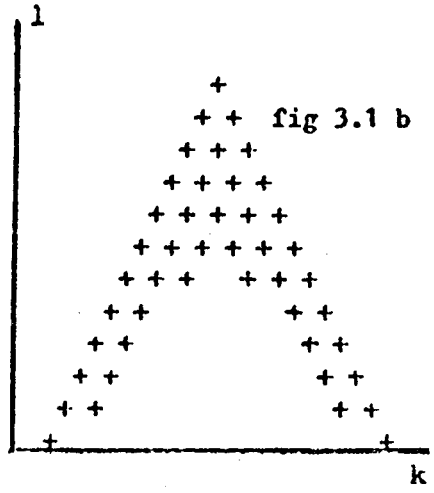


fig 3.1 b

relation non monotone
le coéfficient de corrélation
résume mal la liaison entre
k et l

c - Un programme annexe permet de tracer des diagrammes ternaires, c'est à dire de visualiser, à l'aide d'une figure plane, des échantillons relativement à trois éléments dont la somme est constante.

4. Les méthodes descriptives multivariables.

Ce sont les méthodes d'analyse de données qui ont pour but de fournir des représentations synthétiques et résumées de vastes ensembles de valeurs numériques. Les deux principales méthodes d'étude sont l'analyse en composantes principales et l'analyse des correspondances.

4.1 Cadre général.

Dans la matrice X d'élément général x (i,j), on désigne par : x_{oj} le vecteur colonne des n observations de la variable j ; x_{iv} le vecteur ligne des p variables pour la ième observation.

Avec ces notations, la matrice X s'écrit $X = x_{o1}, x_{o2}, \ldots x_{oj}, \ldots x_{op})$ et sa transposée X' s'écrit $X' = (x_{1v}, x_{2v}, \ldots x_{iv}, \ldots x_{nv})$.

Le tableau des valeurs numériques de X peut donner lieu à deux représentations géométriques :

- Dans R^n, un nuage de p points dont les coordonnées sont les n composantes des vecteurs $x_{o1}, x_{o2}, \ldots x_{oj}, \ldots x_{op}$;
- Dans R^p, un nuage de n points dont les coordonnées sont les p composantes des vecteurs $x_{1v}, x_{2v}, \ldots x_{iv}, \ldots x_{nv}$.

Deux échantillons i et k seront identiques si, dans R^p, leurs vecteurs représentatifs x_{iv} et x_{kv} ont leurs extrémités confondues ; ils seront "proportionnels " si leurs vecteurs représentatifs sont colinéaires.

Deux éléments j et l auront des distributions identiques si, dans R^n leurs vecteurs représentatifs x_{oj} et x_{ol} ont leurs extrémités confondues ; ils seront " proportionnels " si leurs vecteurs représentatifs sont colinéaires.

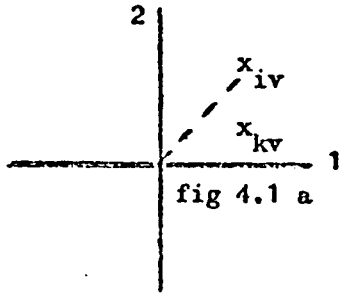


fig 4.1 a

R^p

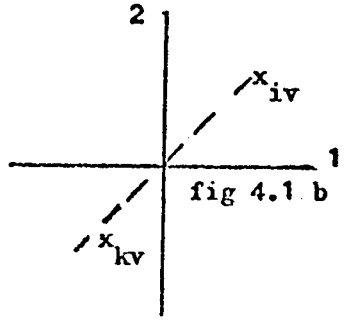


fig 4.1 b

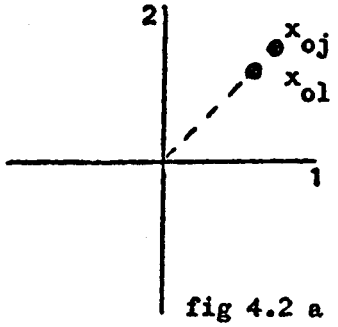
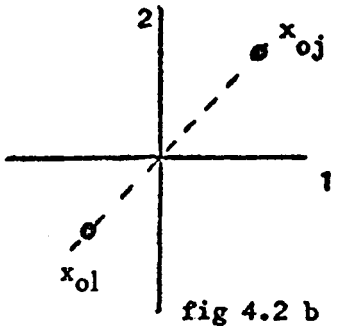


fig 4.2 a



R^n fig 4.2 b

Les proximités entre points représentatifs des échantillons dans R^p indiquent donc la ressemblance de leurs profils de teneurs et les proximités entre points représentatifs des éléments dans R^n indiquent la ressemblance de leurs distributions.

4.1.1.a. Ajustement par un sous-espace dans R^p

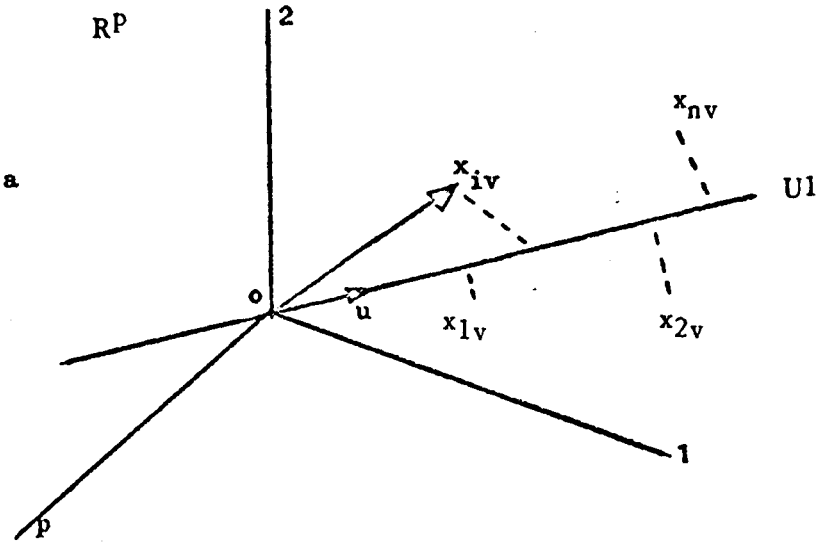
Considérons le nuage des n échantillons dans R^p.

Le problème consiste à décrire la position par rapport à l'origine et la forme du nuage dans un espace de dimension aussi faible que possible.

On cherche d'abord la droite (sous espace à 1 dimension). U1 passant par l'origine qui ajuste au mieux le nuage.

Soit u un vecteur unitaire de cette droite.

fig 4.3.a



La projection du point x_{iv} sur cette droite s'écrit : $Oh_{iv} = u'x_{iv}$ (produit scalaire du vecteur u par le vecteur x_{iv} = produit matriciel du transposé u' de u par x_{iv}).

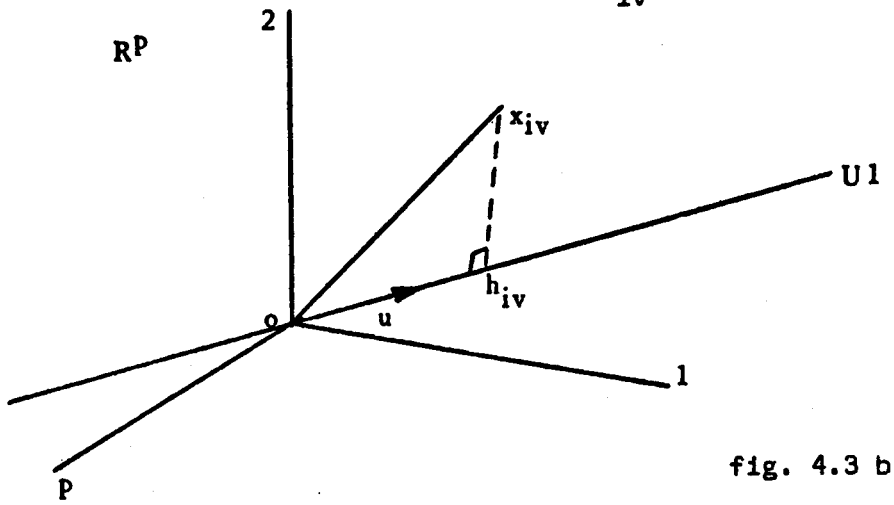


fig. 4.3 b

Le carré de la distance Ox_{iv} du point x_{iv} à l'origine est indépendant de la droite U1 et se décompose en carré de la projection Oh_{iv} sur U1 et carré de la distance $x_{iv}h_{iv}$ à U1.

$$Ox_{iv}^2 = Oh_{iv}^2 + x_{iv}h_{iv}^2$$

La droite U1 qui ajuste au mieux le nuage relativement au point x_{iv} sera celle qui passe le plus près possible de x_{iv} donc celle qui minimise $x_{iv}h_{iv}$. Mais comme la somme

$\overline{Oh_{iv}}^2 + \overline{x_{iv}h_{iv}}^2$ est constante, minimiser $x_{iv}h_{iv}$ revient à maximiser Oh_{iv}.

Il faut donc rechercher le vecteur u_1 définissant la droite U1 tel que, pour l'ensemble des n points du nuage, la quantité :

$$S_1^2 = \sum_{i=1}^{n} \overline{Oh_{iv}}^2 = \sum_{i=1}^{n} \{ u_1'X_{iv} \}^2 \text{ soit maximale.}$$

$$\{ u_1'X_{iv} \}^2 = u_1'X_{iv} \cdot X'_{iv}u_1 = u_1' (X_{iv} X'_{iv}) u_1$$

donc $S_1^2 = u_1' \left(\sum_{i=1}^{n} X_{iv} X'_{iv}\right) u_1$ mais $\sum_{1=1}^{n} X_{iv} X'_{iv}$ est le produit à gauche de la matrice X par sa transposéé X' .

La quantité à maximiser s'écrit donc :

$S_1^2 = u_1' X'X u_1$ avec la contrainte $u_1' . u_1 = 1$ (puisque u_1 est unitaire).

La méthode (méthode du multiplicateur de Lagrange) revient à dériver la quantité matricielle.

$$u_1' X'X u_1 - \lambda (u_1' u_1 - 1)$$

par rapport aux différences composantes u_{1i} du vecteur u_1 puis à annuler chacune des dérivées ;

On obtient alors la relation :

$$X'X - \lambda u_1 = o$$

$$\text{Soit : } X'X u_1 = \lambda u_1$$

qui montre que u_1 est vecteur propre de la matrice X'X

Or
$$S_1^2 = u_1' X'X u_1$$
$$= \lambda u_1' u_1$$
$$= \lambda \quad \text{(puisque } u_1' u_1 = 1\text{)}$$

le maximum cherché est donc une valeur propre de X'X et c'est la plus grande.

Pour rechercher l'espace à deux dimensions qui ajuste le mieux le nuage, il faut trouver une deuxième droite u_2 passant par l'origine, orthogonale à la première, et de vecteur unitaire u_2 tel que la quantité

$$u_2' X'X u_2 \text{ soit maximale}$$

Un raisonnement analogue au précédent montre qu u_2 est le second vecteur propre associé à la seconde valeur propre de X'X.

Il se généralise pour q = 1,2,...m (m étant le rang de la matrice X'X) :

Les q vecteurs propres correspondant aux q plus grandes valeurs propres de la matrice X'X constituent une base (orthonormée) du sous-espace à q dimensions ajustant au mieux au sens des moindres carrés le nuage des n points de R^p.

4.1.2. Ajustement par un sous-espace dans R^n

Considérons le nuage des p éléments dans R^n. On recherche un sous-espace de R^n de dimension aussi faible que possible permettant de décrire au mieux la position par rapport à l'origine et la forme du nuage.

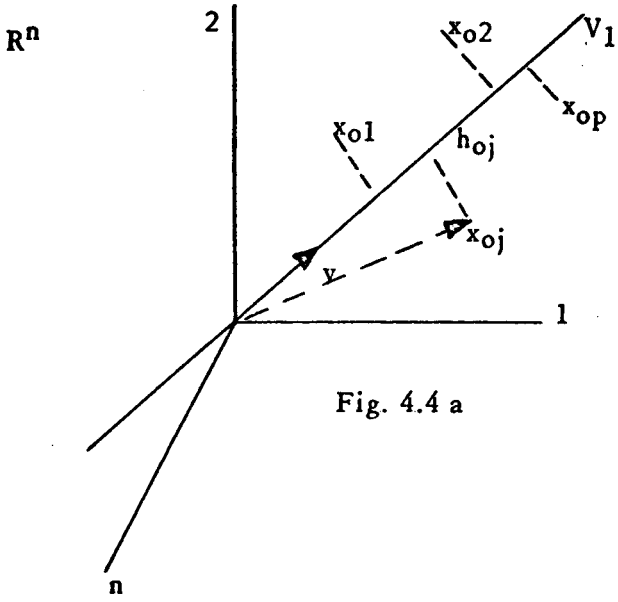


Fig. 4.4 a

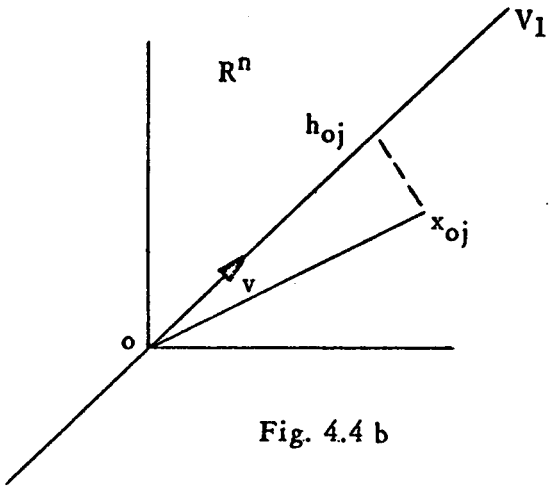


Fig. 4.4 b

On recherche successivement comme précédemment le sous-espace à 1,2,3.., q dimensions qui ajuste au mieux le nuage.

La quantité à maximiser pour trouver la 1° droite V_1, de vecteur unitaire v_1 est :

$$S_1^2 = v_1' XX' v_1, \text{ avec la contrainte } v_1'.v_1 = 1$$

le vecteur v_1 cherché est le vecteur propre de la matrice XX' associé à la plus grande valeur propre de cette matrice.

Le résultat se généralise pour q = 1,2,...m (m étant le rang de la matrice XX' - le même que celui de la matrice X'X) : les q vecteurs propres correspondant aux q plus grandes valeurs propres de la matrice XX' constituent une base (orthonormée) du sous-espace à q dimensions ajustant au mieux au sens des moindres carrés le nuage des p points de R^n.

4.1.3. Relations entre les sous-espaces de R^p et R^n.

Le rang m des matrices X'X et XX' est le même et est égal à la plus petite dimension du tableau X (généralement égale au nombre de dosages).

Les valeurs propres non nulles de XX' sont valeurs propres de X'X (et réciproquement).

A une valeur propre commune λ_q correspondent deux vecteurs propres u_q et v_q liés par les relations :

(1). $$u_q = X'v_q \quad (\text{avec } v'_q v_q = 1)$$

si v_q est unitaire, u_q ne l'est pas.

v_q peut s'exprimer en fonction de u_q par :

(2) $$v_q = \frac{1}{\lambda_q} Xu_q$$

Comme les vecteurs propres d'une matrice ne sont définis qu'à un coefficient près, on peut, à la place de la relation (1), poser :

(3) $$u_q = \frac{1}{\sqrt{\lambda_q}} X'v_q$$

v_q s'exprime alors en fonction de u_q par :

(4) $$v_q = \frac{1}{\sqrt{\lambda_q}} Xu_q$$

Les relations (3) et (4) ont alors une forme symétrique montrant que les analyses dans R^p et dans R^n sont déductibles l'une de l'autre.

4.2. L'analyse en composantes principales.

4.2.1. Analyse sur matrice de covariance (variables non réduites).

Les teneurs des éléments peuvent avoir des ordres de grandeur très distincts, différant souvent par un facteur de 100 ou plus (exemple : teneurs en Pb et en Ag). Les moyennes ont des valeurs très hétérogènes (exemple : 5g/t pour Ag et 500g/t pour Pb).

Le tableau analysé sera donc le tableau Y de terme général :

$$y(i,j) = x(i,j) - m(j)$$

où m (j) est la teneur moyenne empirique de l'élément j.

Cette transformation est dissymétrique par rapport aux lignes et aux colonnes.

la matrice Y'Y dont on recherche les valeurs et vecteurs propres est la matrice de covariance des éléments.

4.2.2. Analyse sur matrice de corrélation (variables reduites).

Les éléments peuvent avoir des teneurs très hétérogènes en moyenne mais aussi en dispersion ; dans le cas où l'on étudie par exemple des majeurs et des traces, les échelles de mesure sont même différentes (% ou g/t) ; la comparaison directe des valeurs brutes n'a alors plus de sens.

Le tableau analysé sera dans ce cas le tableau Y de terme général :

$$y(i,j) = \frac{x(i,j) - m(j)}{s(j)}$$

où m (j) et s (j) sont la teneur moyenne et l'écart type empiriques de l'élément j.

Comme la précédente, cette transformation est dissymétrique par rapport aux lignes et aux colonnes.

La matrice Y'Y dont on recherche les valeurs et vecteurs propres est alors la matrice de corrélation des éléments.

4.3. L'analyse factorielle des correspondances

D'introduction récente (1968), elle étudie à l'origine des tableaux de dépendance formés de lignes correspondant à des catégories d'individus et de colonnes correspondant à des caractères.

A l'intersection de la ligne i et de la colonne j, on porte le nombre de fois ou la catégorie i présente le caractère j (par exemple, le nombre de fois où les roches d'un type donné renferment un minéral donné). Soit k(i,j) ce nombre.

C'est nécessairement un entier positif.

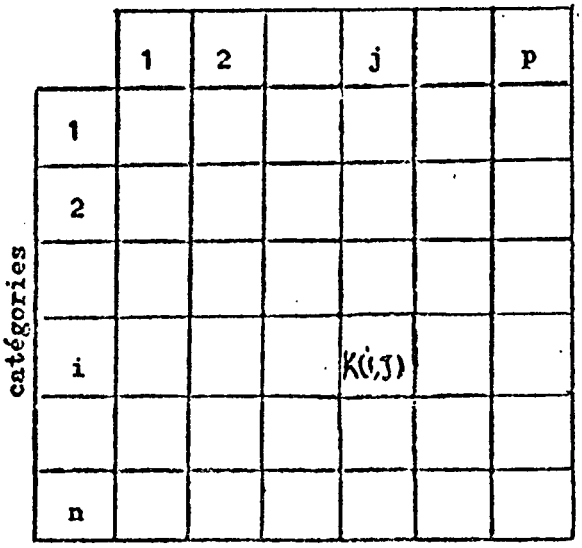


fig 4.5

En désignant par k la somme totale de tous les termes du tableau :

$k = \sum_{i,j} k(i,j)$, on étudie le tableau (à n lignes et p colonnes) d'élément général $p(i,j) = \frac{k(i,j)}{k}$.

p (i,j) est alors une estimation de probabilité puisque :

$\sum_{i,j} p(i,j) = 1$

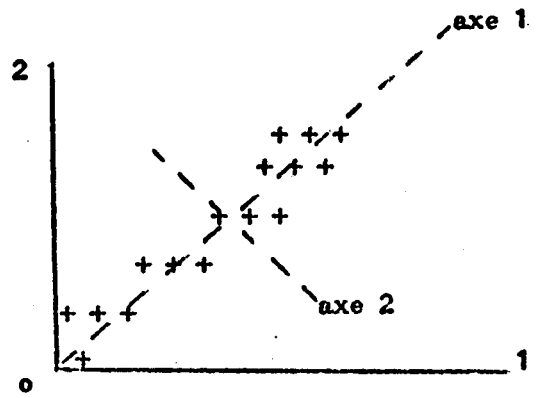


fig 4.7.

4.4.2. Correspondances.

Après transformation des données de base par la formule $z(i,j) = \frac{x(i,j)}{x}$, les coordonnées des échantillons i et i' sont calculées par les formules :

$t(i,j) = \frac{z(i,j)}{z(i)}$ et $t(i',j) = \frac{z(i',j)}{z(i')}$; $t(i,j)$ se déduit donc de $x(i,j)$ par une première division par x, somme totale du tableau de départ, puis par une deuxième division par la somme Z(i) de la ligne i.

Les unités sur les axes de R^p sont donc modifiées par rapport aux axes de départ dans R^p.

Les échantillons i et i' ont alors pour coordonnées dans R^p.

$$x_{iv} \begin{cases} t_{i1} = z(i,1)/z(i) \\ t_{i2} = z(i,2)/z(i) \\ \text{"} \\ t_{ij} = z(i,j)/z(i) \\ \text{"} \\ t_{ip} = z(i,p)/z(i) \end{cases} \qquad \text{et } x_{i'v} \begin{cases} t_{i'1} = z(i',1)/z(i') \\ t_{i'2} = z(i',2)/z(i') \\ \text{"} \\ t_{i'j} = z(i',j)/z(i') \\ \text{"} \\ t_{i'p} = z(i',p)/z(i') \end{cases}$$

Le carré de leur distance dans R^p est :

$$D^2(x_{iv}, x_{i'v}) = (t_{i1} - t_{i'2})^2 + (t_{i2} - t_{i'2})^2 + \ldots + (t_{ij} - t_{i'j})^2 + \ldots +$$

$$(t_{ip} - t_{i'p})^2 = \sum_{j=1}^{p} (t_{ij} - t_{i'j})^2$$

$$= \sum_{j=1}^{p} \left[z(i,j)/z(i) - z(i',j)/z(i') \right]^2$$

Si pour une coordonnée (par exemple la 3ème) , les valeurs absolues des chiffres sont grandes, leur différence peut être grande.

Dans la somme des carrés des différences représentant le carré de la distance D^2, le terme correspondant $(t_{i3} - t_{i'3})^2$ sera prépondérant et " écrasera " les autres termes.

Pour corriger cet inconvénient, chaque différence $(t_{ij} - t_{i'j})^2$ peut être divisée par z(j) (= somme de la colonne correspondant à l'élément j).

La nouvelle distance entre i et i' s'écrit alors *

$$D^2 (x_{iv}, x_{i'v}) = \sum_{j=1}^{p} \frac{1}{z(j)} \left[z(i,j)/z(i) - z(i',j)/z(i') \right]^2$$

Les unités sur les axes sont une nouvelle fois modifiées, mais différemment d'un axe à l'autre. Le nuage est donc déformé par rapport au précédent par un facteur de déformation différent d'un axe à l'autre.

Après avoir défini cette nouvelle distance, on affecte à chaque point échantillon une masse égale à z(i) (somme en lignes); un échantillon sera d'autant plus " pesant " que la somme de ses teneurs sera grande.

L'analyse des correspondances revient à chercher les axes du système des n points de $\mathbb{R}^p$ dans la métrique définie ci-dessus et affectés d'une masse Z(i) = somme des teneurs en ligne

Les axes sont déterminés l'un après l'autre comme indiqué au § IV.1.a et servent de nouveau système de référence. Ils ont la propriété d'expliquer successivement une part décroissante de l'inertie du système des n points.

* <u>NB</u> : $D^2 (x_{iv}, x_{i'v})$ s'écrit aussi :

$$D^2 (x_{iv}, x_{i'v}) = \sum_{j=1}^{p} \left[\frac{z(i,j)}{z(i)\sqrt{z(j)}} - \frac{z(i',j)}{z(i')\sqrt{z(j)}} \right]^2$$

Cette opération est équivalente à celle qui consisterait à diviser $z(i,j)/z(i)$ par $\sqrt{z(j)}$, z(j) étant proportionnel à la moyenne m(j) de

la colonne j:(m (j) = z (j)/n). On s'intéresse donc aux variations relatives de z(i,j)/z(j) par rapport aux racines carrées des teneurs moyennes et non aux variations absolues.

Les propriétés mathématiques de la distance définie en correspondances

$$D^2 (x_{iv}, x_{i'v}) = \sum_{j=1}^{p} \frac{1}{z(j)} \left[z(i,j)/z(i) - z(i'j)/z(i') \right]^2$$

entrainent les deux conséquences fondamentales suivantes :

b1. On compare les profils des deux échantillons x_{iv} et $x_{i'v}$, c'est à dire que les deux échantillons seront confondus si leurs profils de teneurs sont proportionnels (et non nécessairement égaux comme en composantes principales).

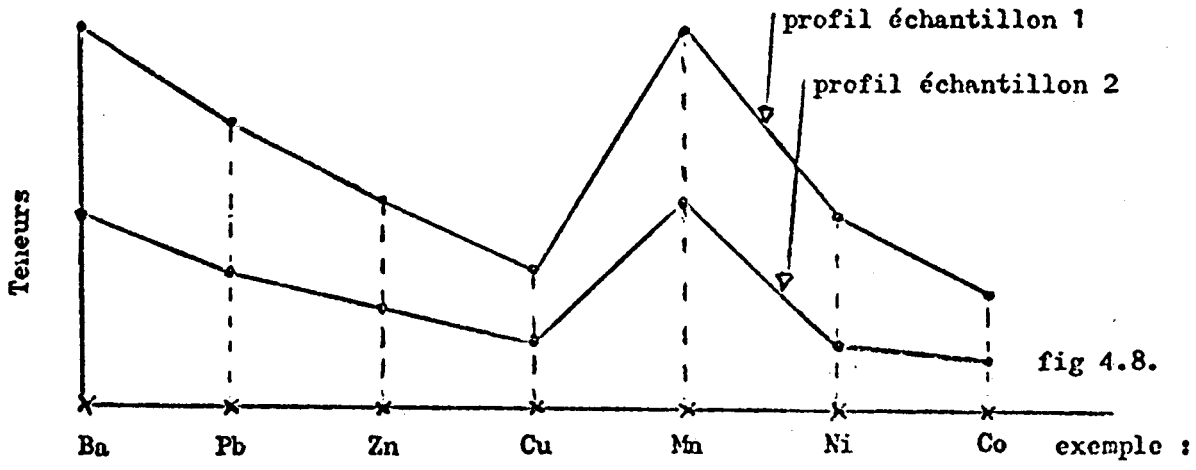


fig 4.8.

b2. On peut remplacer deux échantillons confondus au sens ci-dessus par un troisième échantillon fictif affecté de la somme des masses des deux échantillons (z (i) + z (i')) sans rien changer aux distances entre couples de points. Cette propriété dite " d'équivalence distributionnelle " confère une bonne stabilité aux résultats d'analyse des correspondances

vis à vis des fluctuations de l'échantillonnage (nombre et localisation des échantillons)

Remarque : Ces propriétés relatives aux échantillons, sont évidemment transposables aux éléments.

5. Les méthodes de prévision :

Contrairement aux méthodes descriptives qui ne posent aucun modèle a priori, elles supposent l'existence de relations, généralement linéaires, entre des variables et d'autres variables ou entre des variables et des facteurs, c'est à dire des variables sous-jacentes aux observations mais non directement observables. Elles permettent ensuite de déterminer les paramètres du modèle choisi.

A cette classe de méthodes appartiennent essentiellement la régression linéaire et l'analyse factorielle en facteurs communs et spécifiques. *

5.1. La régression linéaire.

5.1.1. Modèle général.

Dans le tableau de base X, d'élément x (i,j), on désire étudier la loi conditionnelle d'une variable particulière connaissant les réalisations des autres variables.

Plus précisement, on fait l'hypothèse qu'une variable particulière dépend linéairement des autres variables et on cherche à estimer les coéfficients de cette relation.

Pour clarifier l'exposé, supposons que cette variable particulière soit placée dans la première colonne du tableau de données et désignons la par y.

Les composantes de y sont donc, pour tout i, : y (i) = x (i,1)

Avec cette convention, le tableau de départ s'écrit donc :

---

* NB. L'analyse factorielle en facteurs communs et spécifiques pourrait être rangée parmi les méthodes d'analyse de données (composantes principales et correspondances). Néanmoins elle n'est pas essentiellement descriptive puisq'elle suppose l'existence d'un modèle ; à ce titre nous préférons la considérer comme une méthode de prévision.

		1	2	.....	j	.....	p - 1
1	y(1)	x(1,1)	x(1,2)	.....	x(1,j)	.....	x(1,p-1)
2	y(2)	x(2,1)	x(2,2)	.....	x(2,j)	.....	x(2,p-1)
"	"	"	"		"		"
i	y(i)	x(i,1)	x(i,2)	.....	x(i,j)	.....	x(i,p-1)
"	"	"	"		..."...		"
n	y(n)	x(n,1)	x(n,2)	.....	x(n,j)	.....	x(n,p-1)

fig 5.1.

Dans l'espace R^n des échantillons, les données de ce tableau sont donc représentées par le vecteur y et les (p - 1) vecteurs x_{oj} de composantes :

$$y = \begin{pmatrix} y(1) \\ y(2) \\ " \\ y(i) \\ \\ y(n) \end{pmatrix} \qquad x_{oj} = \begin{pmatrix} x(1,j) \\ x(2,j) \\ " \\ x(i,j) \\ " \\ x(n,j) \end{pmatrix} \quad \text{pour } 1 \leq j \leq p-1$$

Définissons également le vecteur unitaire u de composantes

$$u = \begin{pmatrix} 1 \\ 1 \\ " \\ 1 \\ " \\ 1 \end{pmatrix}$$

Le modèle posé est celui d'une relation théorique de la forme :

1. $y = a_1 x_{o1} + a_2 x_{o2} + \ldots + a_j x_{oj} + \ldots + a_{p-1} x_{op-1} + a_p u + \ell$

Sous forme développée, ce modèle s'écrit :

2. $y(i) = a_1 x(i,1) + a_2 x(i,2) + \ldots + a_j x(i,j) + \ldots + a_{p-1} x(i,p-1) + a_p + \ell i$ pour $1 \leqslant i \leqslant n$.

Les coëfficients du modèle devant être estimés sont les nombres a_j $(1 \leqslant j \leqslant p)$. C'est un vecteur aléatoire résiduel de composantes ℓi telles que tous les ℓi soient indépendants, de moyenne nulle et de même variance.

Supposons le problème résolu, c'est à dire la connaissance de tous les coefficients a_j.

Ces coefficients permettent de définir un vecteur w de R^n, combinaison linéaire des vecteurs x_{oj} et du vecteur u :

$$w = a_1 x_{o1} + a_2 x_{o2} + \ldots + a_j x_{oj} + \ldots + a_{p-1} x_{op-1} + a_p u$$

de composantes :

$$w(i) = a_1 x(i,1) + a_2 x(i,2) + \ldots + a_j x(i,j) + \ldots + a_{p-1} x(i,p-1) + a_p$$

Les vecteurs y et w seront d'autant plus proches que leur distance D sera plus petite.

Le carré de cette distance s'écrit :

$D^2(y,w) = \sum_{i=1}^{n} (y_i - w_i)^2$ soit sous forme développée :

$$D^2(y,w) = \sum_{i=1}^{n} (y_i - a_1 x(i,1) - a_2 x(i,2) \ldots a_j x(i,j) \ldots a_{p-1} x(i,p-1) - a_p)^2$$

que l'on peut écrire : $D^2(y,w) = \sum_{i=1}^{n} r_i^2$; r_i s'appelle le résidu relatif au ième échantillon et $D^2(y,w) = \sum_{i=1}^{n} r_i^2$ la distance résiduelle entre la variable y et son estimation w.

Le problème d'estimation des coéfficients du modèle (1) revient donc à déterminer les coèfficients a_j rendant minimale la distance residuelle $\sum_{i=1}^{n} r_i^2$

A titre d'illustration, reprenons la représentation des éléments Pb et Zn dans l'espace de 3 échantillons (§-1).

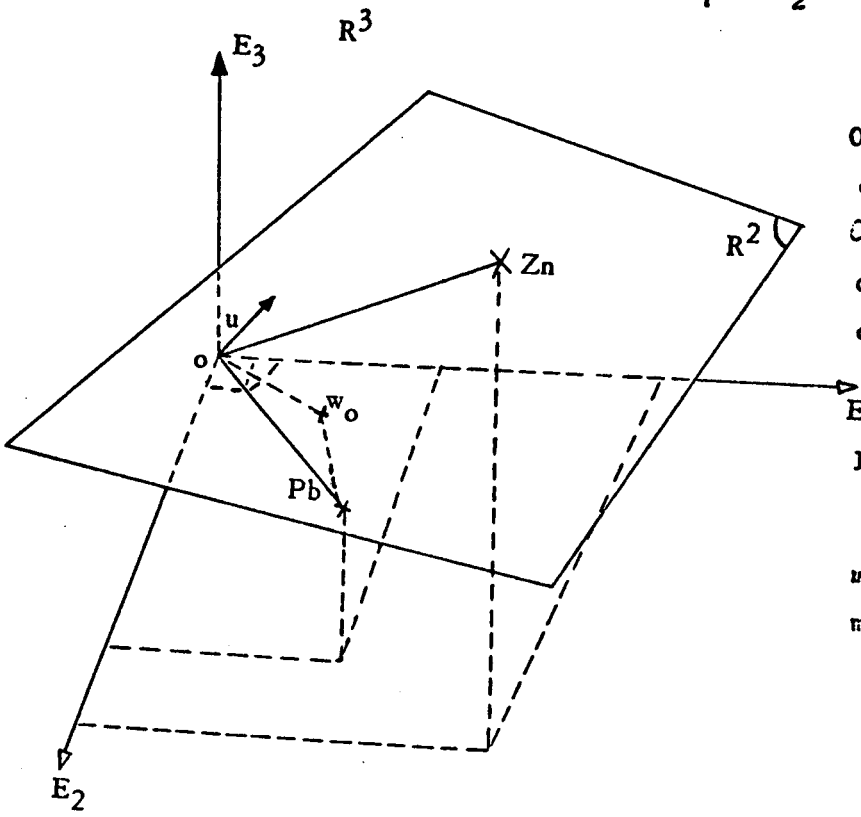
	Pb	Zn
E_1	50	100
E_2	80	90
E_3	30	70

fig 5.2.

Supposons qu'il existe une relation linéaire entre Pb et Zn suivant le modèle :

3. $$Pb = a_1 Zn + a_2 + e$$

dont on désire estimer les coéfficients a_1 et a_2



Les vecteurs Ou de composantes (1,1,1) et OZn de composantes (100,90,70) definissent un plan de R^3 qui contient tous les vecteurs de la forme :

$$Ow = m\,OZn + n\,Ou$$

mais qui ne contient généralement pas les vecteurs OPb.

On recherche un vecteur particulier OWo tel que sa distance à OPb soit minimale. Le vecteur est nécessairement la projection du vecteur OPb sur le plan (OZn,Ou). Les valeurs particulières pour ce vecteur des nombres m et n sont les estimations a_1 et a_2 des coefficients cherchés.

La généralisation à p variables et à n dimensions (échantillons) est la suivante ; Parmi tous les vecteurs de la forme :

$$4. \qquad w = \sum_{j=1}^{p-1} a_j x_{oj} + a_p u$$

qui engendrent un sous-espace à p dimensions de R^n,*

On recherche celui qui est la projection du vecteur y sur ce sous-espace.

Les coefficients de ce vecteur particulier sont les estimations des paramètres du modèle (4).

5.1.2. Régression classique à q variables explicatives.

Lorsque l'on pose le modèle (1), on peut ne pas faire intervenir l'ensemble des p - 1 variables mais seulement q d'entre elles.

On a d'autre part intérêt à centrer les variables par leurs moyennes car le terme constant a_p du modèle (1) est alors nul.

On travaille donc sur les nouvelles variables :

$$y'(i) = y(i) - m_y \text{ et } x'(i,j) = x(i,j) - m_x(j)$$

avec : m_y moyenne de la variable y et $m_x(j)$ moyenne de la jème variable. On recherche alors, parmi tous les vecteurs de la forme

$$5. \qquad w' = \sum_{k=1}^{q} a_k x'_{ok}$$

qui engendrent un sous-espace à q dimensions de R^n **, celui qui est la projection du vecteur y' sur ce sous-espace. Les coefficients de ce vecteur particulier sont les estimateurs de paramètres du modèle (5).

NB. *Sous réserve que les vecteurs $x_{o1},\ldots x_{oj},\ldots x_{op-1}$, u soient linéairement indépendants.

**Sous réserve que les vecteurs $x'_{o1},\ldots x'_{oj},\ldots x'_{oq}$, soient linéairement indépendants.

On montre que dans ce cas, les coéfficients a_k sont les éléments du vecteur colonne a calculé par la relation matricielle :

$$a = V^{-1}_{xx} . V_{xy}$$

dans laquelle

V^{-1}_{xx} est la matrice inverse de la matrice Vxx des covariances expérimentales des q variables explicatives, Vxy est le vecteur colonne des covariances des variables explicatives avec le variable à expliquer.

Remarque.

Cette formule est la généralisation de celle donnant le coéfficient de régression dans le cas d'une seule variable explicative :

$$a_1 = \frac{\text{covariance } (x,y)}{\text{variance } (x)}$$

La qualité de l'ajustement du vecteur y par le vecteur w est mesurée par le rapport de leur longueur (où du carré de leur longueur). Or w est la projection de y sur le sous-espace engendré par les vecteurs observation, on a la relation :

$$\overline{oy}^2 = \overline{ow}^2 + \overline{wy}^2$$

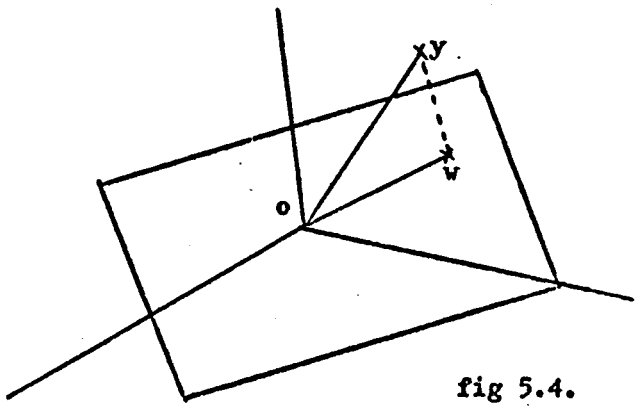


fig 5.4.

La qualité de l'ajustement est donc mesurée par le coéfficient

$$r^2 = \frac{\overline{ow}^2}{oy^2}$$

qui doit être le plus voisin possible de l'unité. r^2 est le carré du coéfficient de corrélation multiple entre la variable expliquée (ou dépendante) y et les q variables explicatives (ou indépendantes) x_{ok}.

5.1.3. Régression étagée ou pas à pas.

Dans le modèle général, on fixe a priori les variables explicatives.

Pour être fondé, ce choix doit donc découler d'hypothèses faites sciemment.

Une variante de la méthode classique permet d' " essayer " successivement pas à pas chacune des p - 1 variables choisies comme variables explicatives.

Cette variante procède par itération :

Désirant à une étape donnée entrer une variable dans le modèle, on détermine si elle améliore la qualité de l'ajustement relativement à un seuil fixé à l'avance. Si non, elle est rejetée. Si oui, on détermine si, une fois cette variable prise en compte, on peut éliminer une des variables entrées précedemment et dont la contribution à la qualité de l'ajustement se trouve désormais inférieure à un seuil donné à l'avance et différent du précédent (plus élevé).

Cette méthode, assez séduisante, doit être employée de préférence à la régression classique quand on ne sait pas quelles variables prendre comme "explicatives".

Il y a néanmoins un choix à faire et un risque à prendre.

Le choix consiste à fixer des valeurs pour les tolérences d'entrée et de sortie des variables ; il ne peut résulter que d'une décision arbitraire.

Le risque est que l'on emploie peut être à tort la régression linéaire : si l'on ne connait pas les variables dites explicatives, connait-on quand même la forme du modèle ?

Celui-ci doit en effet être linéaire et il est indispensable de pouvoir faire cette hyppothèse.

5.2. L'analyse factorielle en facteurs communs et spécifiques.

5.2.1. Modèle général.

Soit le tableau des données de base d'élément général x (i,j), il est constitué de u vecteurs observation à p composantes x_{iv} ou de p vecteurs variable à n composantes x_{oj}

	1	2	.....	j	.....	p
1	x(1,1)	x(1,2)		x(1,j)		x(1,p)
2	x(2,1)	x(2,2)		x(2,j)		x(2,p)
"						
i	x(i,1)	x(i,2)		x(i,j)		x(i,p)
"						
n	x(n,1)	x(n,2)		x(n,j)		x(n,p)

fig 5.5.

Considérons par exemple les p vecteurs variables x_{oj} dans l'espace R^n des échantillons.

On fait l'hypothèse que chaque vecteur variable ne dépend linéairement, à un vecteur spécifique près, que de m facteurs (m n), suivant le modèle

$$x_{o1} = a_{11} f_1 + a_{12} f_2 + \ldots + a_{1k} f_k + \ldots + a_{1m} f_m + e_{o1}$$

"

$$6. \qquad x_{oj} = a_{j1} f_1 + a_{j2} f_2 + \ldots + a_{jk} f_k + \ldots + a_{jm} f_m + e_{oj}$$

"

$$x_{op} = a_{p1} f_1 + a_{p2} f_2 + \ldots + a_{pk} f_k + \ldots + a_{pm} f_m + e_{op}$$

où les vecteurs spécifiques correspondants à chaque vecteur variable sont les vecteurs $e_{o1}, \ldots e_{oj}, \ldots e_{op}$

Exemple : En reprenant l'exemple d'un tableau de 500 échantillons dosés pour 20 éléments et en supposant que chaque élément ne dépende que de 3 facteurs, on aurait le modèle :

$$Cu = a_{11}\, f_1 + a_{12}\, f_2 + a_{13}\, f_3 + e_1$$

$$Pb = a_{21}\, f_1 + a_{22}\, f_2 + a_{23}\, f_3 + e_2$$

$$Zn = a_{31}\, f_1 + a_{32}\, f_2 + a_{33}\, f_3 + e_3$$

$$Ag = a_{41}\, f_1 + a_{42}\, f_2 + a_{43}\, f_3 + e_4$$

Sous forme matricielle, le modèle (6) s'écrit :

7. $$X = A.F + E$$

5.2.2. Représentation dans R^n. Analyse en mode R.

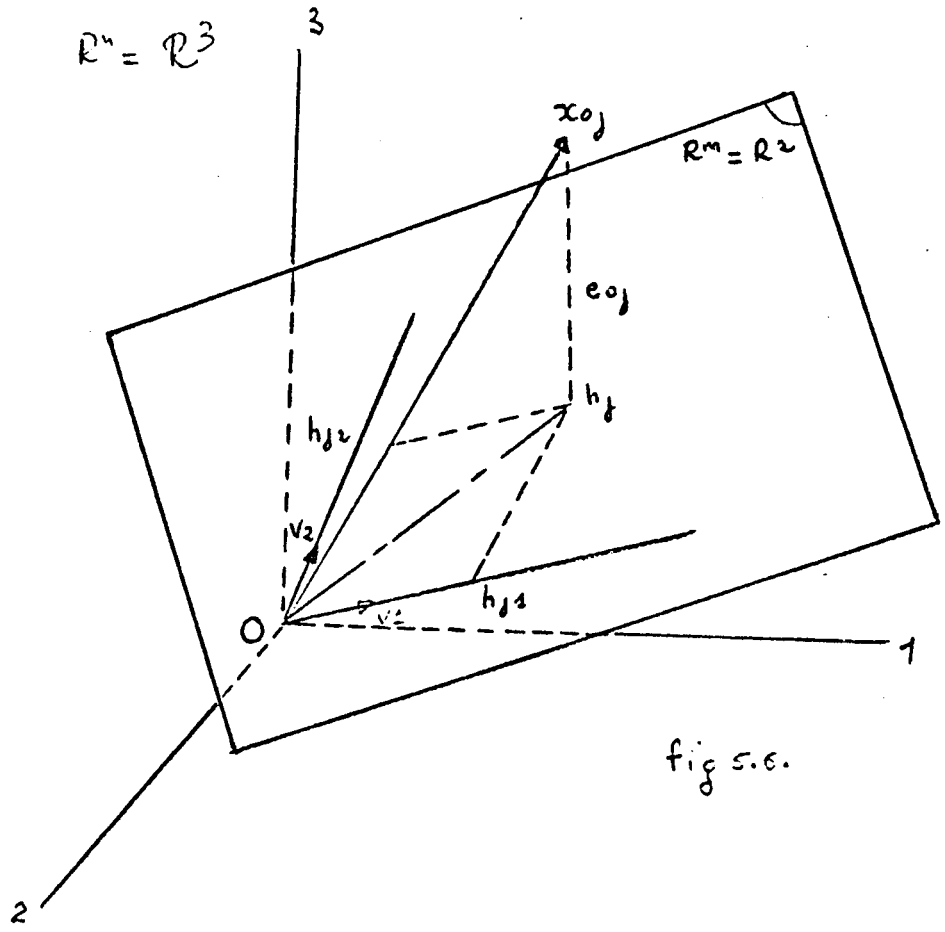


fig 5.6.

L'analyse des relations entre variables est appelée " analyse en mode R ".

Les éléments sont représentés par p vecteurs x_{oj} de R^n.

On désire projeter le nuage qu'ils constituent dans un sous-espace R^m de R^n (de dimension m donnée à l'avance et inférieure à n) contenant l'origine

On définit ce sous-espace à l'aide d'une base orthonormée, c'est à dire d'un système de droites orthogonales de vecteurs unitaires $v_1, v_2, \ldots v_k, \ldots v_m$.

Le point x_{oj} de R^n se projette en $h_{j1}, h_{j2}, \ldots h_{jk}, \ldots h_{jm}$ sur ces droites et en h_j sur le sous-espace qu'elles définissent. Le vecteur x_{oj} se décompose en la somme de sa projection Oh_j sur R^m et du vecteur $(h_j\ x_{oj})$ perpendiculaire à R^m.

$$x_{oj} = Oh_j + (hj\ x_{oj})$$

Oh_j se décompose lui même en la somme de ses projections sur les axes de R^m :

$$Oh_j = Oh_{j1} + Oh_{j2} + \ldots + Oh_{jk} + \ldots + Oh_{jm}$$

donc :

$$x_{oj} = Oh_{j1} + Oh_{j2} + \ldots + Oh_{jk} + \ldots + Oh_{jm} + (h_j\ x_{oj})$$

Le vecteur $(h_j\ x_{oj})$, orthogonal au sous-espace R^m, est appelé le vecteur spécifique ou vecteur résiduel e_{oj}.

Le modèle s'écrit alors, en confondant, pour simplifier l'écriture, les notations Oh_{jk} et h_{jk},

8. $$x_{oj} = h_{j1} + h_{j2} + \ldots + h_{jk} + \ldots + h_{jm} + e_{oj}$$

Le problème revient à rechercher le sous-espace R^m (c'est à dire les vecteurs v_k) qui conduit à des résidus de moyenne nulle (moyenne $(e_{oj}) = o$) et de même variance (variance $(e_{oj}) = s^2$)

Pour le résoudre, on doit supposer que les composantes des vecteurs résiduels sont non corrélées entre elles.

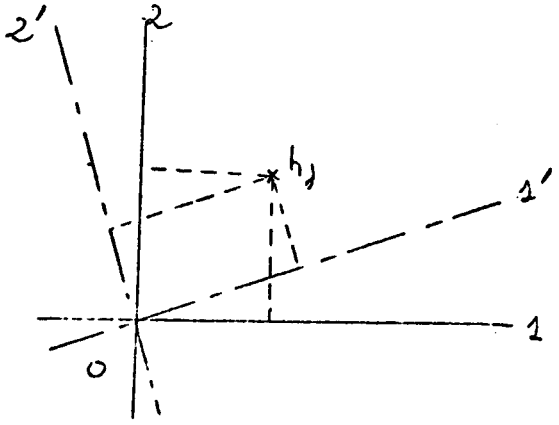
Ayant déterminé le sous-espace R^m correspondant, on assimile, à son résidu près e_{oj}, le vecteur x_{oj} à sa projection h_j sur R^m.

C'est à dire que, au lieu de rapporter le nuage des x_{oj} aux axes initiaux de R^n, on rapporte sa projection sur R^m aux axes de vecteurs unitaires v_k déterminés ci-dessous (base orthonormée de R^m). Les coordonnées de h_j sur ces axes sont les coefficients du modèle (6).

Ces nouveaux axes, communs à toutes les variables (éléments), définissent de nouvelles variables appelées facteurs communs. Le résidu, propre à chaque vecteur variable (élément), est appelé facteur spécifique.

Il représente la " spécificité " du vecteur variable, (c'est à dire la perte d'information entrainée par l'assimilation du vecteur x_{oj} à sa projection h_j.

5.2.3. Rotation des axes factoriels. Le système des m droites définissant R^m n'est pas unique.



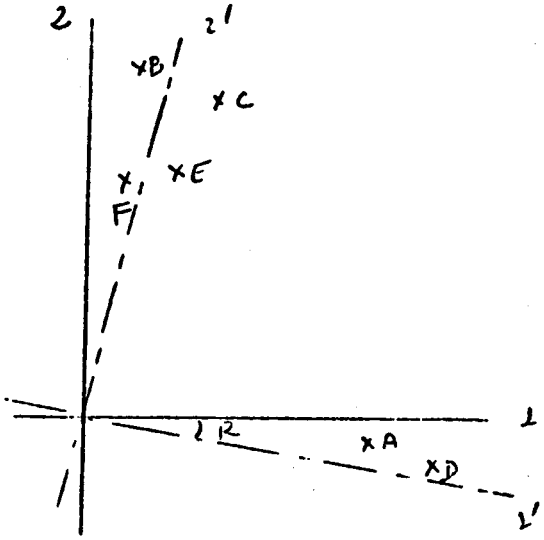
Il n'est déterminé qu'à une rotation près.

Le point h_j peut aussi bien être rapporté aux axes (1,2) ou aux axes (1',2'), déduits de (1,2) par une rotation.

Il n'est d'ailleurs même pas nécessaire que les axes (1',2') demeurent orthogonaux.

Parmi l'infinité de systèmes possibles, on choisira celui qui permet la meilleure " interprétation " des facteurs. Différents méthodes sont proposées pour formaliser ce choix ; le détail n'en sera pas exposé car le terme " meilleure interprétation " ne se traduit pas par un critère mathématique unique.

Le plus couramment utilisé est le critère VARIMAX (rotation orthogonale). Son fonctionnement peut être illustré sur l'exemple à deux dimensions suivant :



Les 6 points A B C D E F sont d'abord rapportés aux axes (1,2).
Les points A et D ont un facteur 2 de l'ordre de 0, tandis que les autres points ont des facteurs 1 et 2 du même ordre de grandeur (au signe près). Les points A et D sont donc bien " expliqués " par le seul facteur 1 alors que les autres points ne le sont pas.
Par rotation d'un angle R, on peut choisir un autre système d'axes (1',2') tel que les 4 points B C E F aient, soit un facteur 1, soit un facteur 2, de l'ordre de 0. On dira que la structure des points A B C D E F par rapport à (1',2') est plus simple que par rapport à (1,2).

D'une façon plus générale dans le cas de m facteurs, on cherchera un système d'axes qui minimise le nombre de facteurs bipolaires (à valeurs positives et négatives élevées) car ceux ci sont généralement plus difficilement interprétables en termes géologiques que des facteurs homopolaires.

5.2.4. Représentation dans R^p. Analyse en mode Q.

L'analyse des relations entre observations est appelée " analyse en mode Q." Les échantillons sont représentés par n vecteurs x_{iv} de R^p. On désire projeter le nuage qu'ils constituentdans un sous-espace R^q (de dimension q donnée à l'avance et inférieure à p), contenant l'origine.

On définit ce sous-espace à l'aide d'une base orthonormée, c'est à dire d'un système de droites orthogonales de vecteurs unitaires $u_1, u_2, \ldots u_\ell, \ldots u_q$.

Par un développement analogue à celui décrit pour les vecteurs variables dans R^n, on confond le vecteur observation x_{iv} de R^p avec sa projection h_i sur le sous-espace R^q au résidu vectoriel aléatoire e_{iv} près, $(e_{iv} = (h_i x_{iv}))$.

Comme précédemment, le système des q droites définissant R^q n'est pas unique et n'est déterminé qu'à une rotation près.

Sous forme algébrique, le modèle analogue au modèle (6) du § 5.2.1. s'écrit :

$$x_{1v} = b_{11} g_1 + b_{12} g_2 + \ldots + b_{1l} g_l + \ldots + b_{1q} g_q + e_{1v}$$

$$\vdots$$

$$x_{iv} = b_{i1} g_1 + b_{i2} g_2 + \ldots + b_{il} g_l + \ldots + b_{iq} g_q + e_{iv} \qquad 9.$$

$$\vdots$$

$$x_{nv} = b_{n1} g_1 + b_{n2} g_2 + \ldots + b_{nl} g_l + \ldots + b_{nq} g_q + e_{nv}$$

où les résidus correspondants à chaque vecteur observation sont les vecteurs spécifiques $e_{1v}, \ldots e_{iv}, \ldots e_{nv}$.

Exemple : Dans le tableau (500 X 20), en supposant que chaque échantillon ne dépende réellement que de 4 facteurs, on aurait le modèle.

$$E_1 = b_{11} g_1 + b_{12} g_2 + b_{13} g_3 + b_{14} g_4 + e_1$$

$$E_2 = b_{21} g_1 + b_{22} g_2 + b_{23} g_3 + b_{24} g_4 + e_2$$

$$E_{500} = b_{500\,1} g_1 + b_{500\,2} g_2 + b_{500\,3} g_3 + b_{500\,4} g_4 + e_{500}$$

5.2.5. Comparaison entre les analyses en mode R et en mode Q.

Elles ne sont pas directement déductibles l'une de l'autre car il n'y a pas de correspondances entre les sous-espaces (R^m et R^q).

Dans l'analyse en mode R, on recherche les valeurs propres de la matrice des coéfficients de corrélation entre variables (éléments).

Dans l'analyse en mode Q, on recherche les valeurs propres de la matrice " coéfficients de corrélation " entre observations (échantillons).

La première est une matrice carrée de rang p, la seconde une matrice carrée de rang n.

Pratiquement, l'analyse en mode Q nécessite une taille mémoire en ordinateur beaucoup plus importante que l'analyse en mode R puisque, le plus souvent en géologie, le nombre n d'échantillons (ex : 500) est très supérieur au nombre p d'échantillons (ex : 20).

Cette limitation pratique en diminue la généralité d'emploi.

DEUXIEME PARTIE

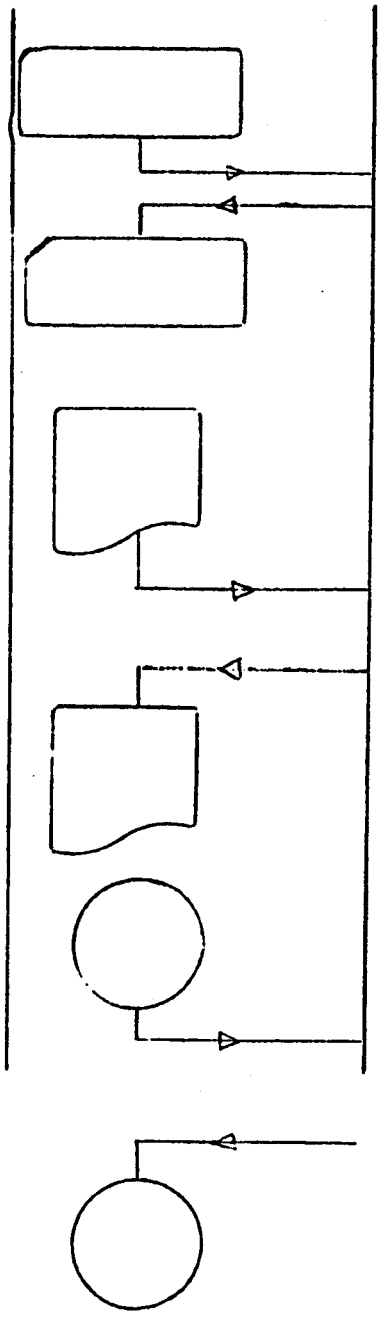
DESCRIPTION DES METHODES

ET PROGRAMMES

ORGANISATION GENERALE ET TERMINOLOGIE.

En guise d'introduction, on va indiquer les symboles utilisés dans les organigramme, les notations rencontrées dans le calcul matriciel et les types de bordereau d'exploitation employés.

Symboles utilisés :



Lecture carte perforée.

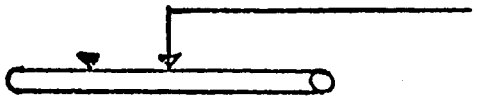
Sortie carte perforée.

Etat de codification en bordereau comportant en général des données à perforer qui seront entrées en machine.

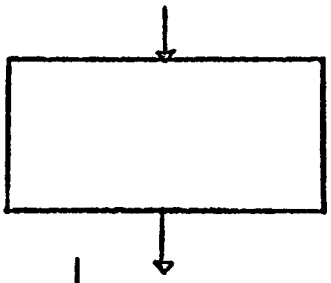
Sortie d'états de résultats à l'imprimante.

Lecture fichier de données sur disque.

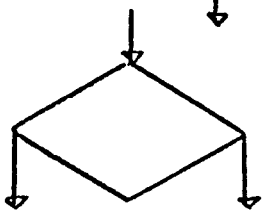
Ecriture fichier de données sur disque.



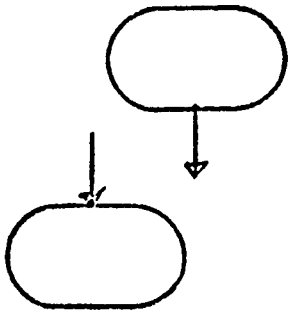
Sortie sur traceur automatique



Module ou ensemble de traitements à exécuter.



Branchement conditionnel.



Début.

Fin.

Notations rencontrées :

$X = (x_{ij}), [i=1,n \ ; \ j=1,p]$ matrice

X' = transposée de $X = (x'_{ij}), [i=1,p \ ; \ j=1,r]$ tel que $\forall\ i,j, x'_{ij} = x_{ji}$

(c'est la matrice de X dans laquelle on a échangé lignes et colonnes)

x_{iv} = matrice ligne extraite de X ou vecteur ligne.

$x_{iv} = (x_{i1}\ x_{i2}\ \ldots x_{ip})$

x_{0j} = matrice colonne extraite de X ou vecteur colonne

$x_{0j} = (x_{1j}\ x_{2j}\ x_{3j}\ \ldots x_{nj})$

X^{-1} = inverse de X. C'est une matrice telle que le produit : $X^{-1}.X = I$ où I est une matrice unitaire (avec des 1 sur la diagonale principale et des 0 ailleurs).

$X.u = \lambda.u$ u est un vecteur propre de X

λ est la valeur propre associée

Types de bordereau :

Bordereau modèle 1 (voir figure tome II)
utilisé pour les programmes statistiques élémentaires.

Bordereau modèle 2 (voir figure tome II)
utilisé pour le programme de création de fichiers.

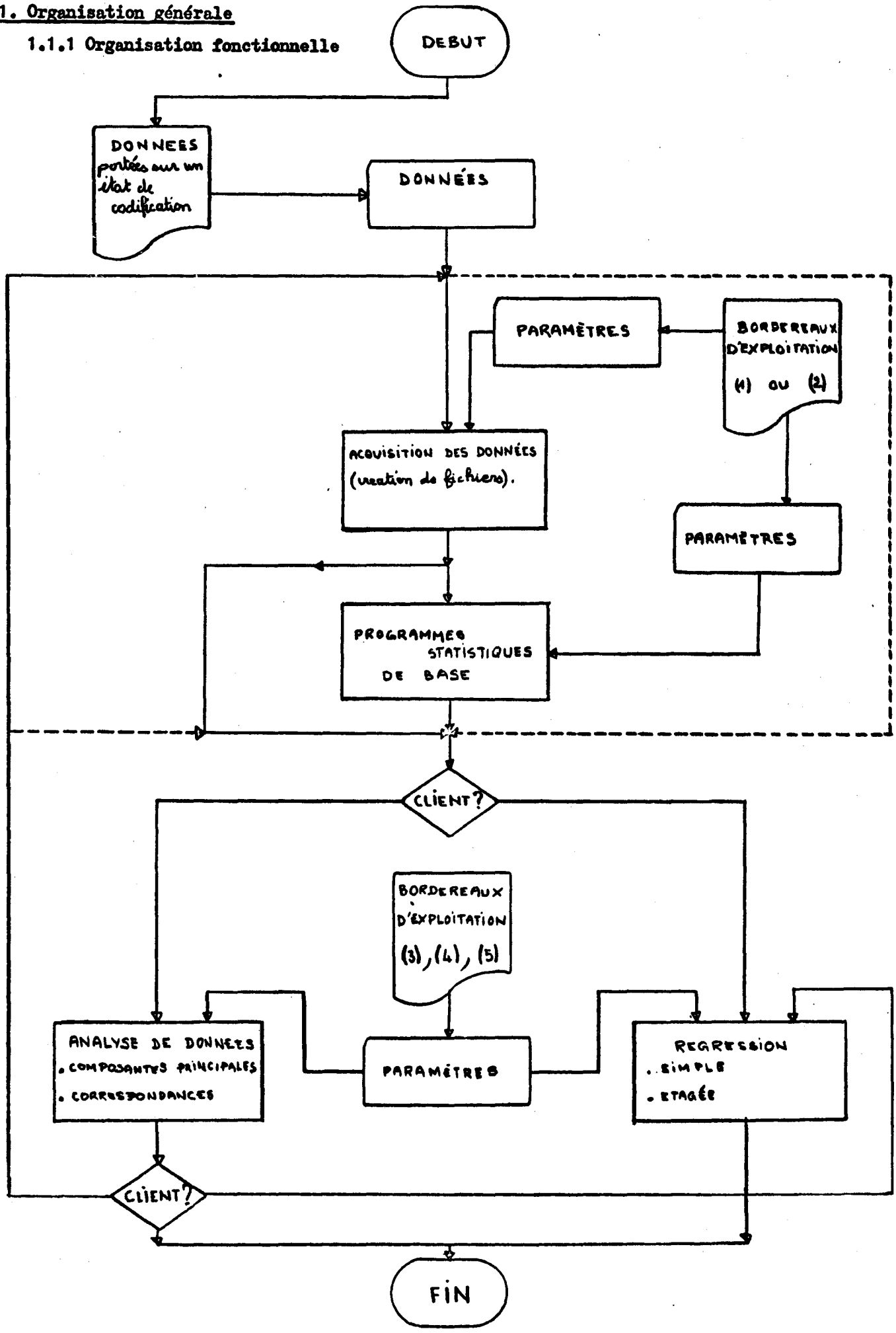
Bordereau modèle 3 (voir figure tome II)
utilisé pour les programmes d'analyse de données et de prévisions (calcul)

Bordereau modèle 4 (voir figure tome II)
utilisé pour les programmes d'analyse de données.

Bordereau modèle 5 (voir figure tome II)
utilisé pour le programme de report de points.

1.1. Organisation générale

1.1.1 Organisation fonctionnelle



Cet organigramme peut s'interpréter de la façon suivante :

. L'utilisateur reporte ses données à traiter sur un état de codification.

. A partir de cet état de codification, il obtient ses données sous forme de carte perforées.

. S'il utilise la procédure la plus courante (fichiers normalisés), il remplit le bordereau d'exploitation modèle (2) (carte paramètre) ; un passage en machine lui permet alors de créer et charger ses fichiers de donnée sur disque.

. A partir de ces fichiers de données normalisés, il peut exécuter en machine des programmes statistiques de base en remplissant le bordereau modèle (1) ou des programmes d'analyse de données, de prévision (régression) en remplissant les bordereaux modèle (3), (4) avec sortie traceur ou non. S'il veut à la suite de ces derniers programmes exécuter un report de points (bordereau modèle (5) à remplir), en utilisant leurs coordonnées, il lui faut créer un nouveau fichier normalisé à partir des facteurs extraits de l'analyse (retour à l'acquisition de données). A partir de ces mêmes facteurs, il peut soit effectuer une nouvelle analyse de données, soit utiliser les programmes de prévision (régression).

. S'il utilise la procédure moins usée (entrée carte), il pourra directement utiliser les programmes d'analyse de données en remplissant les bordereaux correspondants et les programmes de prévision.

1.1.2. Nomenclature des programmes

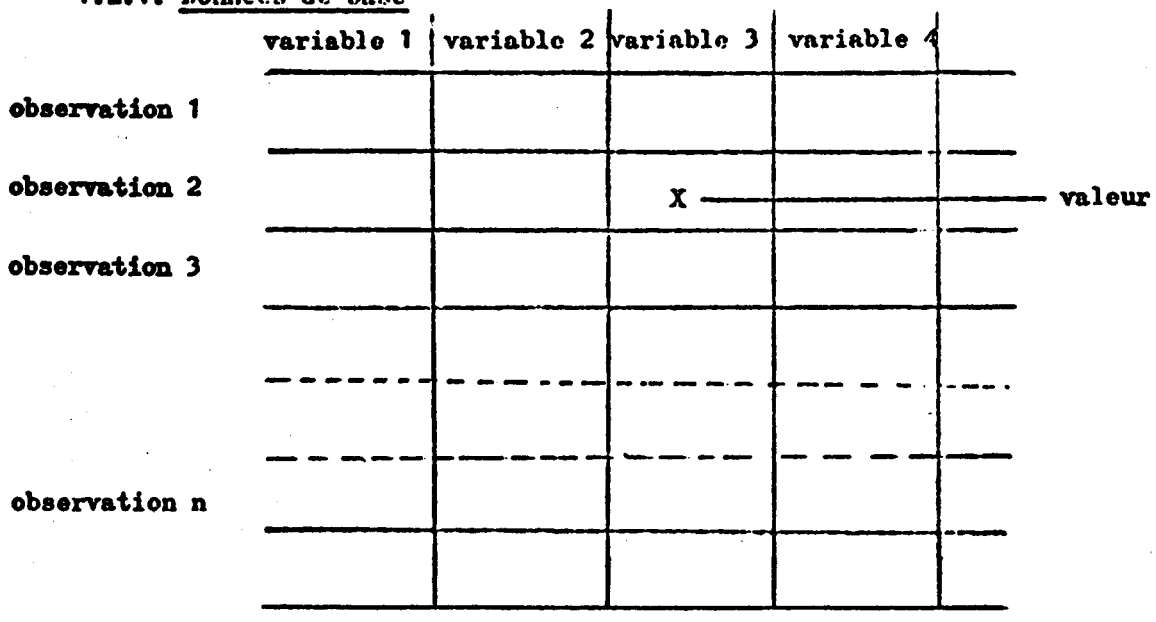
Acquisition des données	GRAD1	: création de fichier
Méthodes descriptives élémentaires	TMET	: teneurs moyennes et écarts-type
	HIST	: histogrammes
	MATCØ	: matrice de corrélation
	GRAD	: graphique de données
	DITRI	: diagrammes triangulaires
	REPP1	: report de points

Méthodes descriptives multivariables d'analyse de données	AFAC1	: correspondances (calcul) (entrée cartes)
	AFAD1	: correspondances (calcul) (entrée disque)
	AFAT1	: correspondances (tracé)
	ACØC1	: composantes principales (calcul) entrée cartes)
	ACØD1	: composantes principales (calcul) entrée disque)
	ACØT1	: composantes principales (tracé)
Méthodes de prévision	REGA1	: régression linéaire
	RTAG1	: régression étagée
	ASIC1	: analyse en facteurs communs et spécifiques (entrée cartes)
	ASID1	: analyse en facteurs communs et spécifiques (entrée disque)
compléménts à l'analyse de données (en annexe)	NUDY1	: nuées dynamiques

1.2. Terminologie

Les données de base se présentent sous forme de tableaux à 2 dimensions. Pour la compréhension des chapitres suivants, on va rappeler les conventions utilisées pour la terminologie.

1.2.1. Données de base



Chaque colonne correspond à une variable, chaque ligne à une observation.
Une valeur x sera la grandeur relative à une observation et à une variable.
On emploiera donc les termes : variables, observations et valeurs.
Dans les problèmes d'analyse de données, on est amené à calculer des facteurs relatifs aux variables et des facteurs relatifs aux observations : les premiers seront appelés facteurs F, les seconds facteurs G .

Nous définirons aussi des facteurs supplémentaires Fs et Gs, relatifs respectivement aux variables et aux observations.

1.2.2. exemple de tableau

soit en général le tableau de données (x_{ij}) i=1,7 ; j=1,4

	variable 1	variable 2	variable 3	variable 4	Somme en ligne
observation 1	X11	X12	X13	X14	T1
observation 2	X21	X22	X23	X24	T2
observation 3	X31	X32	X33	X34	T3
observation 4	X41	X42	X43	X44	T4
observation 5	X51	X52	X53	X54	T5
observation 6	X61	X62	X63	X64	T6
observation 7	X71	X72	X73	X74	T7
Somme en colonne	S1	S2	S3	S4	AKT

ses caractéristiques sont :

7 observations (i=1,7)

4 variables (j=1,4)

On définit les quantités suivantes :

somme en colonnes (ou poids des colonnes) : Sj, j=1,4

somme en lignes (ou poids des lignes) : Ti, i=1,7

somme en lignes et colonnes (ou poids total) : AKT

1.2.3. Structure d'une observation

Exemple : Une observation, constituée par un identificateur et 8 valeurs correspondant à 8 variables, est disposée sur cartes perforées suivant le schéma ci-dessous.

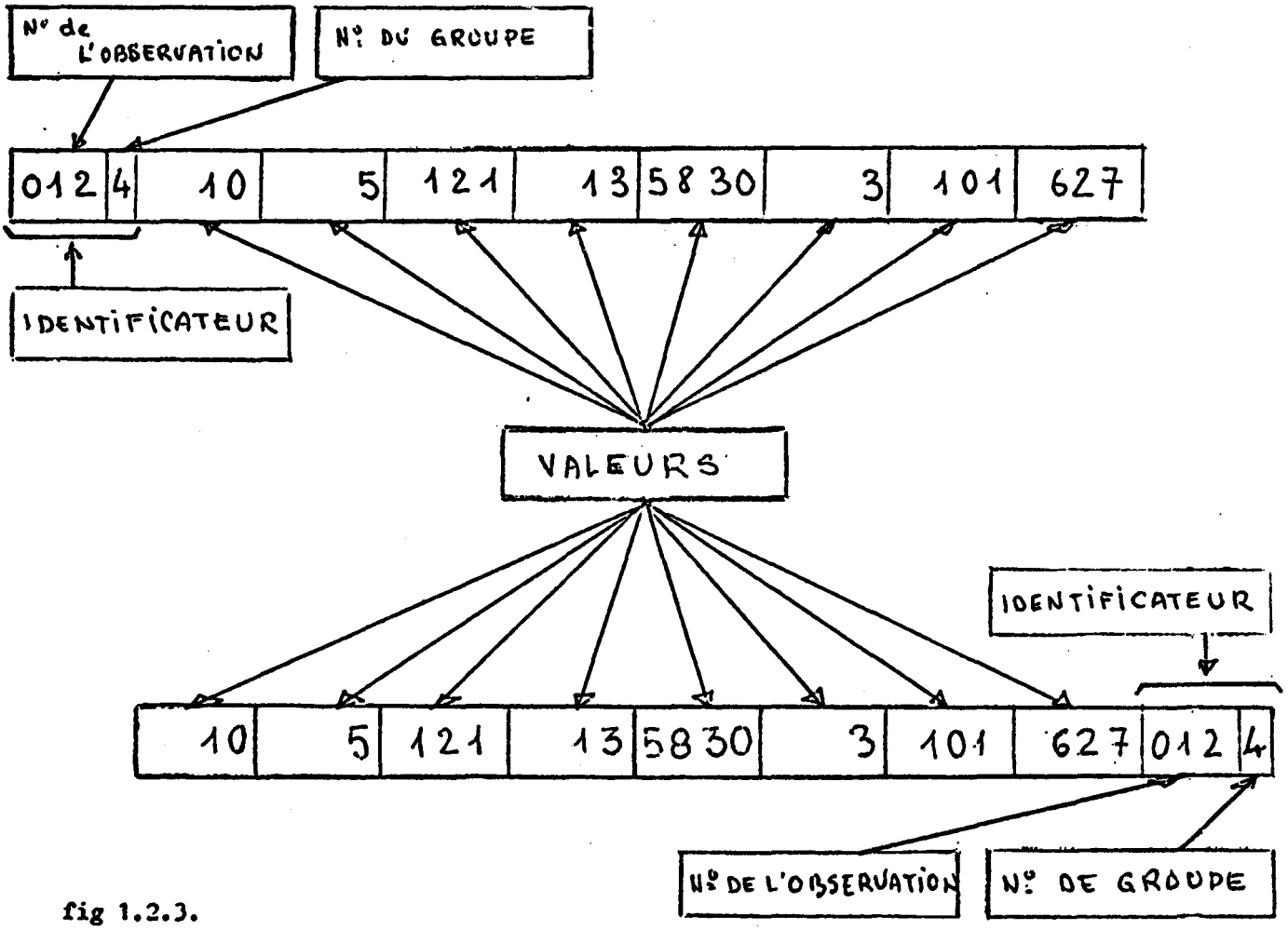


fig 1.2.3.

croquis du haut : l'identificateur précède les valeurs
(il est en début de carte)

croquis du bas : l'identificateur suit les valeurs
(il est en fin de carte)

l'identificateur est l'association d'un numéro de référence (ex:012) et d'un caractère facultatif de groupage (ex:4) qui peuvent être considérés globalement ou séparément.

1.2.4. Facteurs multiplicatifs

Ils permettent d'enregistrer les données sur disque magnétique en format entier et donc d'économiser une place importante. Ce sont les puissances de 10 par l'opposé desquelles il faut multiplier les valeurs enregistrées sur disque pour obtenir les vraies valeurs (voir fig. 1.2.4)

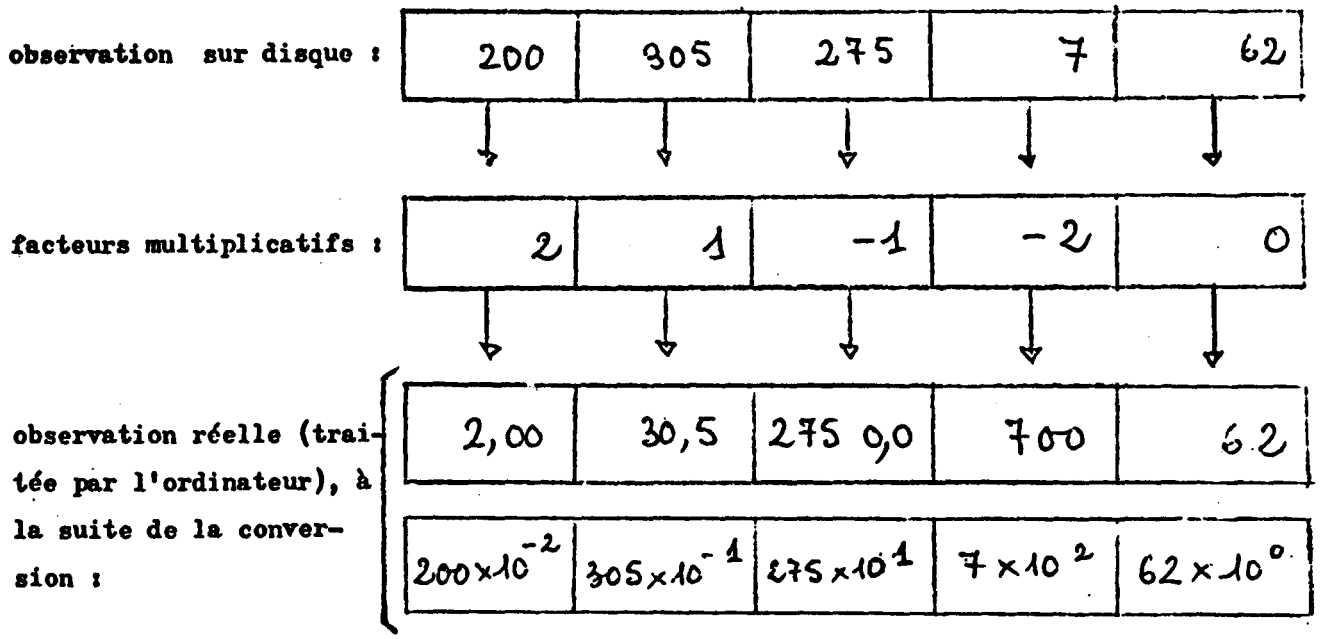
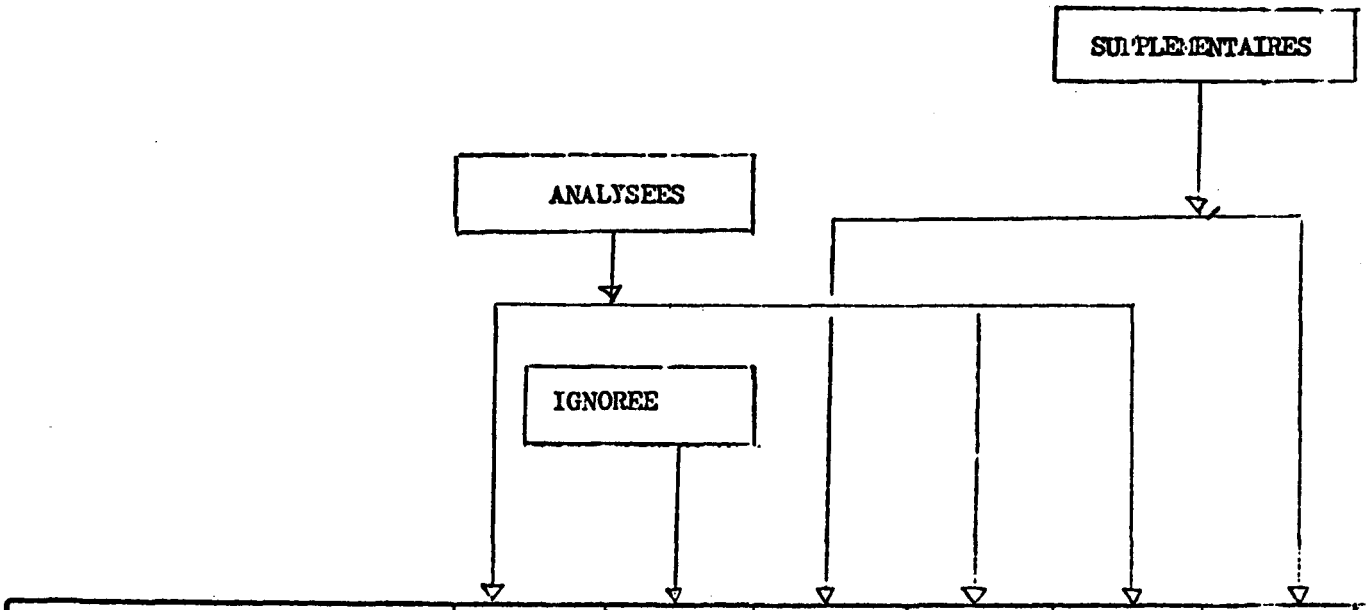


fig 1.2.4.

1.2.5. Variables supplémentaires



	variable1	variable2	variable3	variable4	variable5	variable6
observation 1						
observation 2						
observation 3						
observation 4						
observation 5						

fig 1.2.5.

Parmi l'ensemble des colonnes "variables", certaines sont prises en compte et participent à l'analyse (variables analysées) ; d'autre sont totalement ignorées ; d'autres ne participent pas à l'analyse mais sont néanmoins représentables dans le sous espace déterminé par l'analyse (variables supplémentaires).

1.2.6. Observations supplémentaires

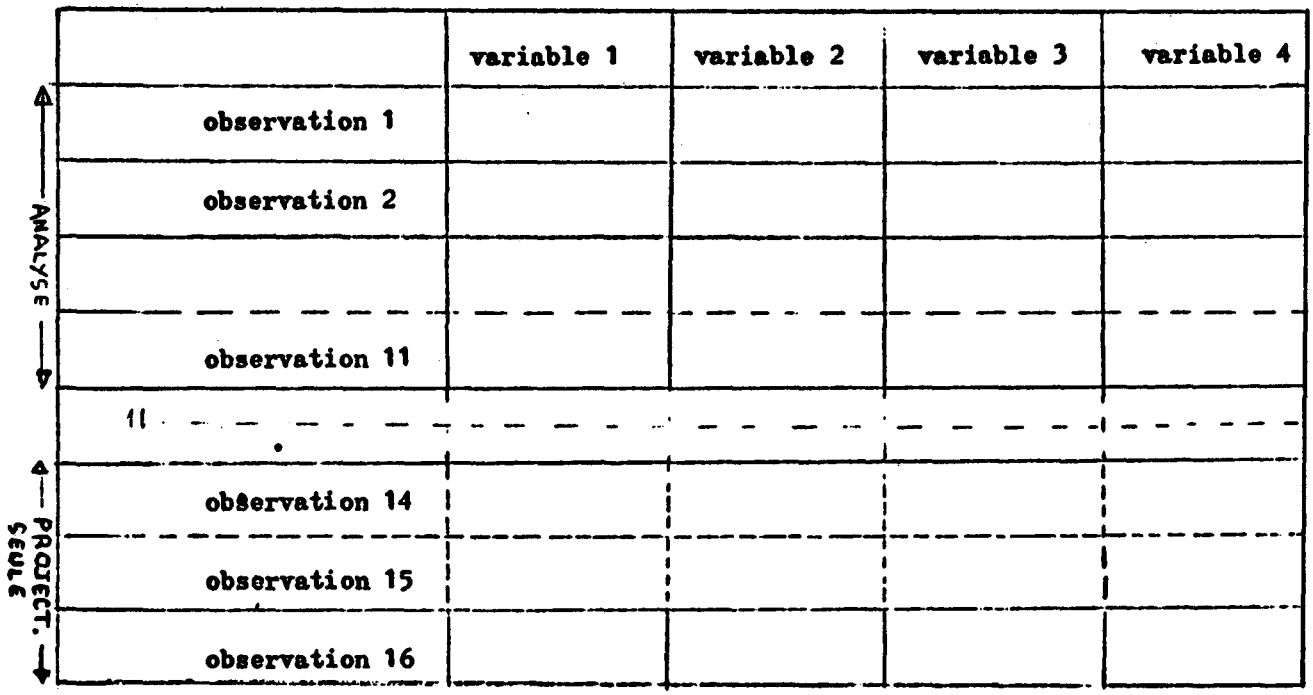


fig 1.2.6.

Les observations 14-15-16 seront dites supplémentaires.

Parmi l'ensemble des lignes "observations", certaines sont prises en compte et participent à l'analyse (variables analysées) ; d'autres sont totalement ignorées ; d'autres ne participent pas à l'analyse mais sont néanmoins représentables par leur projection dans le sous espace déterminé par l'analyse (observations supplémentaires) .

1.2.7. Contraintes

Tous les programmes passent actuellement sur un ordinateur IBM-1130 de capacité mémoire 8 K mots et muni d'un seul disque. Ils seront prochainement adaptés à un ordinateur IBM de la série 360 modèle 40 de 128 K octets de mémoire centrale et équipé de bandes et disques.

Il est clair que les contraintes différeront suivant la version utilisée

contraintes 1130		contraintes 360	
Nombre de variables	$\leq$ 24	Nombre de variables	$\leq$ 50
Nombre d'observations	$\leq$ 2000	Nombre d'observations	illimité
Nombre de facteurs	$\leq$ 5	Nombre de facteurs	$\leq$ 5

2. Collecte des données et organisation des fichiers

2.1. Acquisition directe des données.

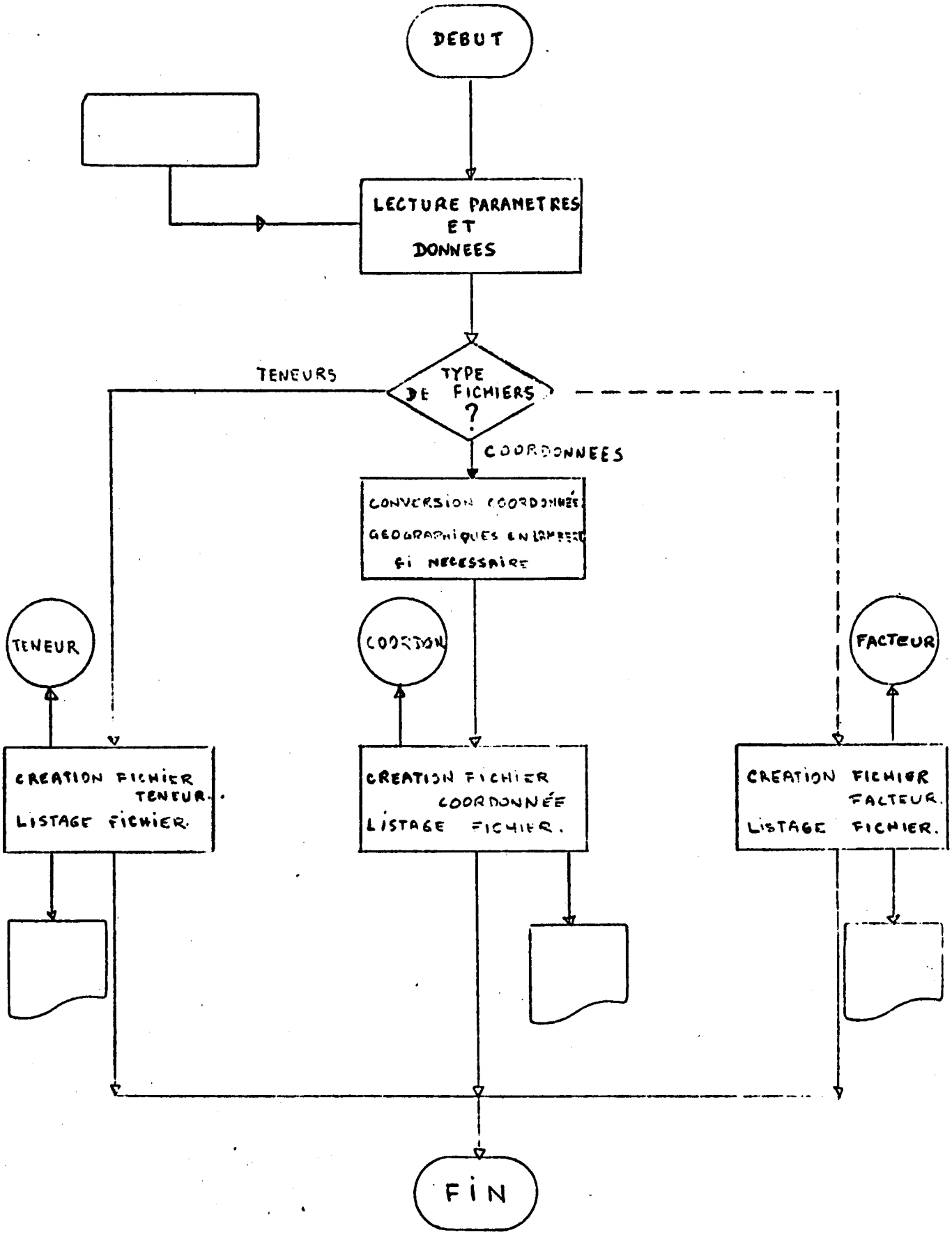
Les données de mesure peuvent se présenter à l'origine sous plusieurs formes :

- bordereau normalisé ou quelconque, à partir duquel des cartes sont perforées ;
- cartes directement perforées par un programme indépendant.
- fichier général préalablement constitué sur disque magnétique.

Quelques soient leur support ou leur format, ces données sont d'abord réorganisées dans un fichier disque normalisé, accessible par la totalité des programmes.

Remarque Le terme "données" est pris dans son sens le plus général, c'est à dire : "résultats de mesures effectuées sur des échantillons, ou valeurs d'autres variables telles que coordonnées, ou paramètres calculés à un stade intermédiaire".

2.2. Organigramme de principe.



Cet organigramme de principe peut s'interpréter de la façon suivante :

- Lecture données sous forme de cartes perforées et lecture des paramètres obtenus à partir du bordereau modèle (2).

- Au premier passage, on peut créer soit le fichier teneur seul, soit les fichiers coordonnées et teneurs dans un même passage. Pour le fichier coordonnées, une conversion est prévue dans le cas où celles-ci sont géographiques ; elle permet de les transformer en coordonnées Lambert. Au fur et à mesure de sa création, le fichier est listé à l'imprimante.

- Après une analyse de données, on peut créer un fichier à partir des facteurs extraits de cette analyse. Le fichier facteur est listé de la même façon au fur et à mesure de sa création.

2.3. Présentation du bordereau de description de données : modèle 2

On rencontre dans le bordereau cinq types de paramètres: description physique des données, description logique des données, transformation des données, caractéristiques du traitement et nature du (ou des) fichier (1).

- Description physique des données.

Il convient de préciser :

- les positions respectives de l'identificateur, éventuellement des coordonnées et (ou) des teneurs.

- les noms des variables

- les facteurs multiplicatifs (puissances de 10 par l'opposé desquelles il faut multiplier les valeurs sur carte pour restituer les vraies valeurs)

- Description logique des données

Il s'agit de définir :

- la largeur de l'identificateur ;

Pour des raisons de programmation, la longueur à indiquer sur la carte paramètre n'est que la moitié de la longueur réelle (voir bordereau) ou nombre de caractères

- le nombre de colonnes ou variables.

- le nombre de coordonnées et leur nature (géographiques ou Lorbert)

- Indirectement le nombre d'observations, en plaçant une carte ne contenant que des 9 à la fin du paquet des données.

- Transformation des données

Les coordonnées figurant dans le fichier doivent être toujours rectangulaires.

Le programme admet des coordonnées géographiques qui sont alors transformées en coordonnées rectangulaires Lambert avant d'être écrites dans le fichier.

- Caractéristiques du traitement.

Suivant la valeur du paramètre "nature du fichier" (voir fig.2.5.2.) on peut créer 1 ou plusieurs fichiers :

- Nature du fichier
  - teneurs
  - coordonnées
  - facteurs

Les fichiers teneurs et coordonnées peuvent être créés simultanément.

Le fichier coordonnées peut être créé seul à condition que le fichier facteur l'ait été au préalable.

Remarque : Il est possible de porter des corrections sur un fichier grâce à un programme de mise à jour.

2.4. Description des fichiers .

2.4.1. Fichier "Teneurs.

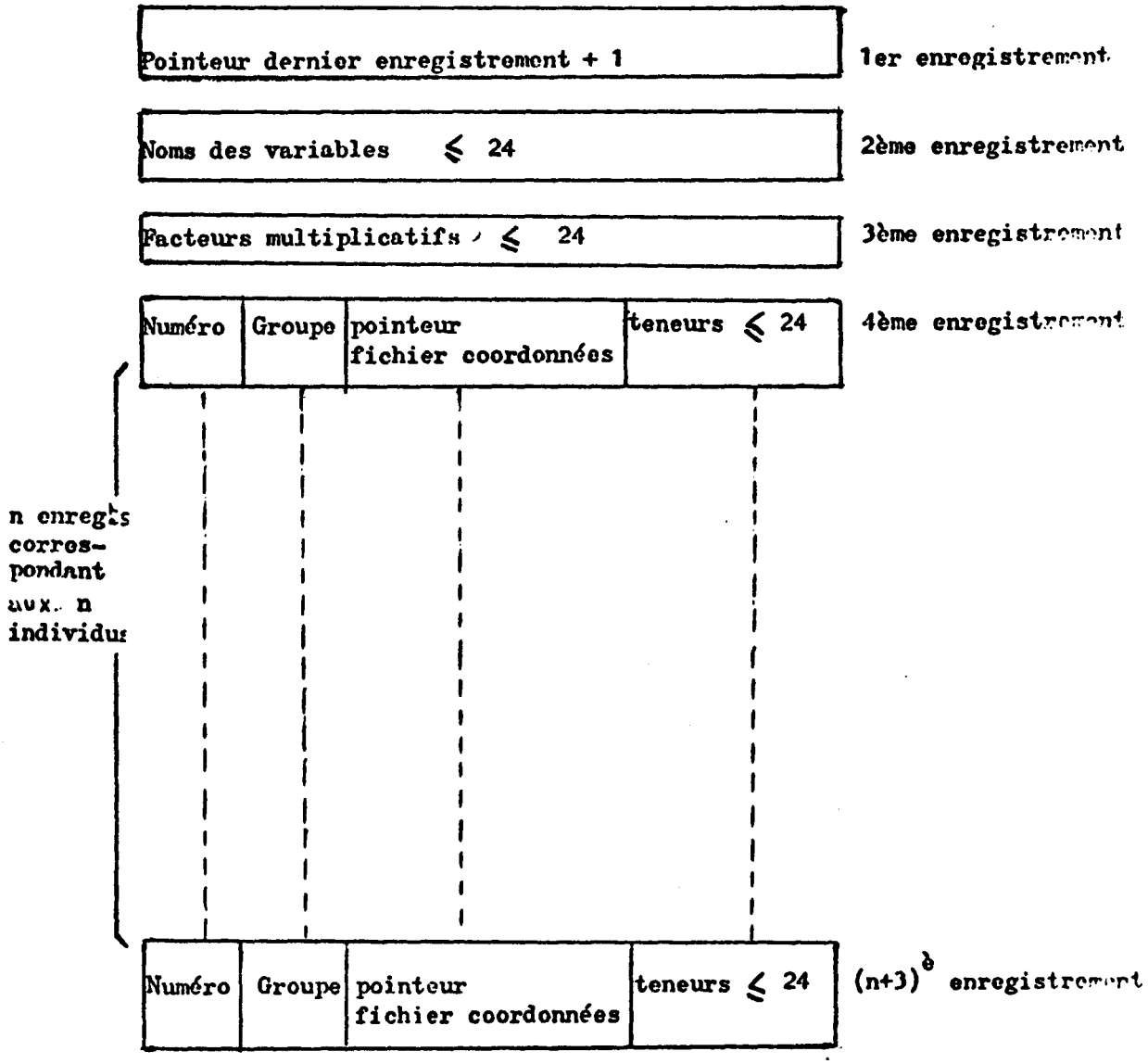


fig. 2.4.1.

2.4.2. Fichier "coordonnées".

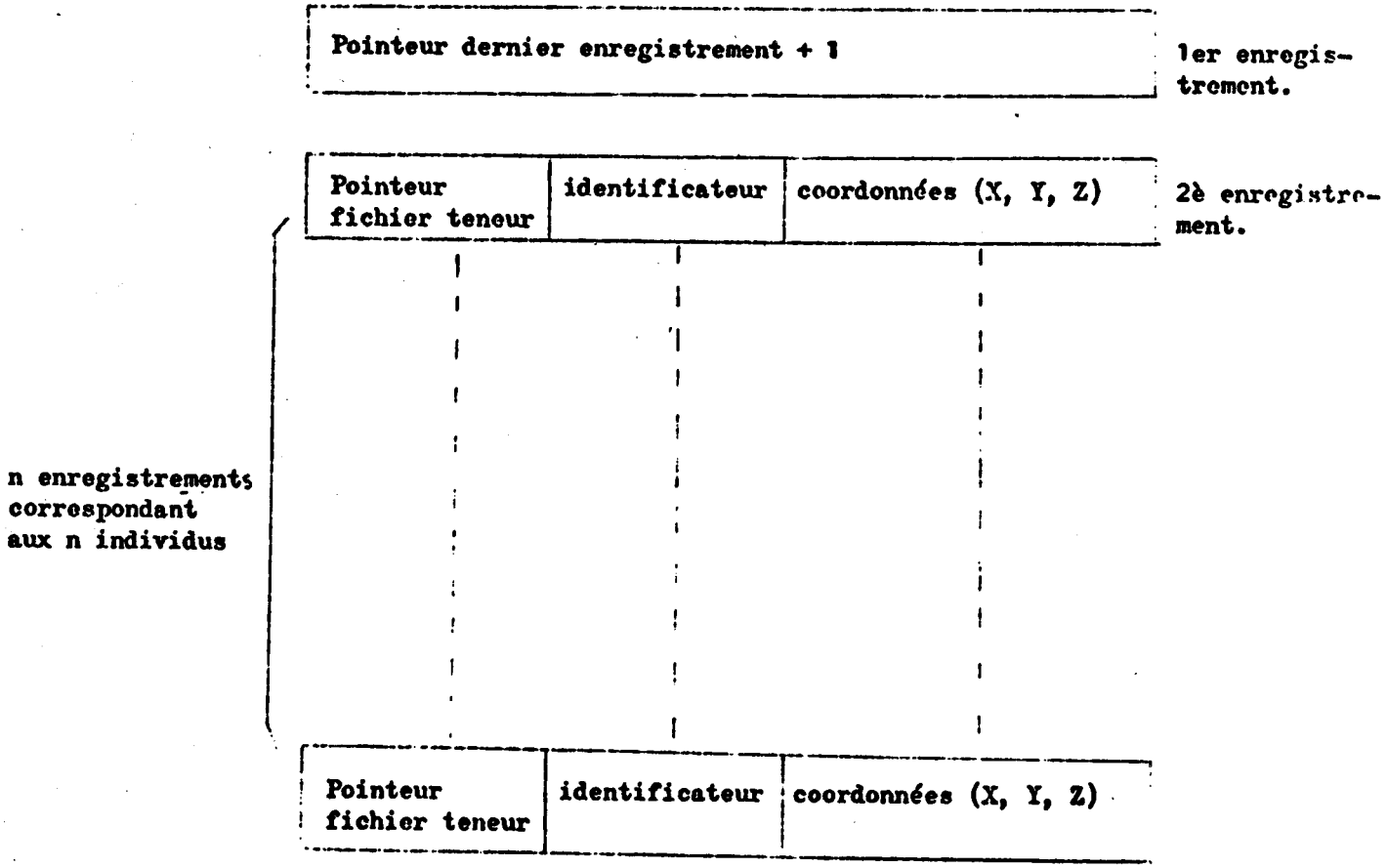


fig. 2.4.2.

2.4.3. Fichier "facteurs".

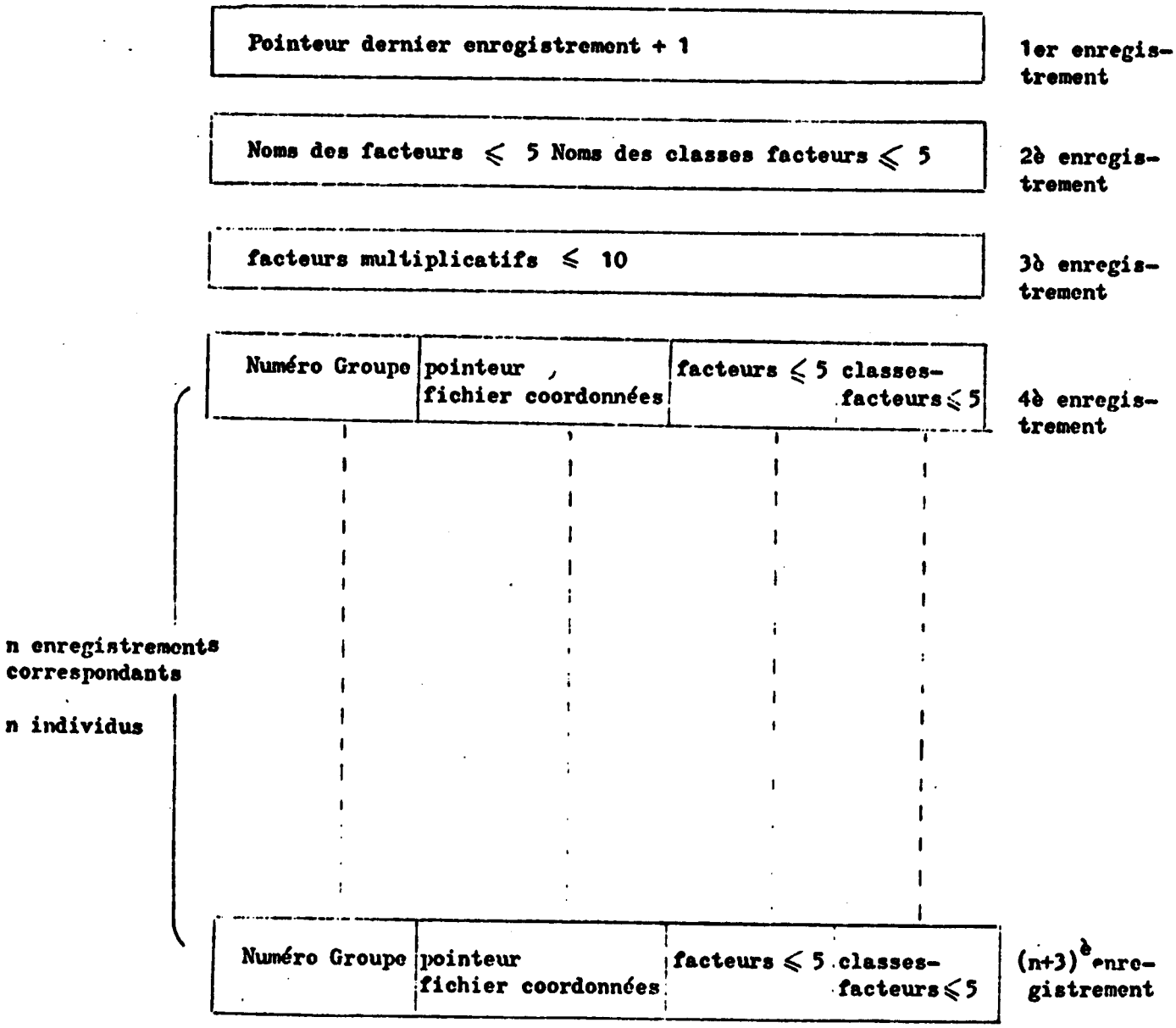


fig. 2.4.3.

2.5. Programmes utilisant ces fichiers.

2.5.1. Fichier "teneurs".

Programme statistiques élémontaires.

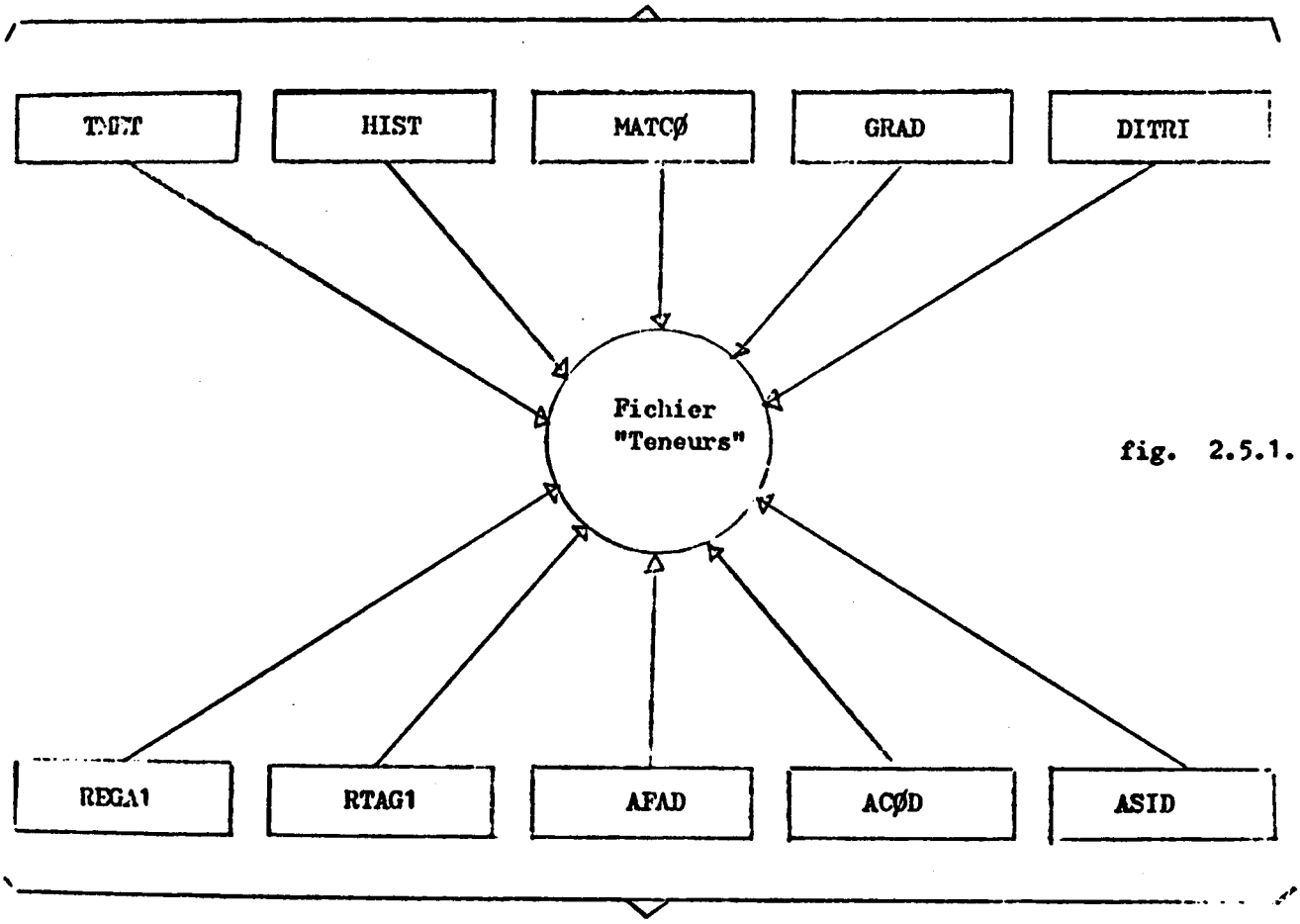


fig. 2.5.1.

Programmes statistiques multivariables. fig 2.5.1.

Le fichier normalisé "Teneurs" est à la base de l'organisation des traitements. Tous les programmes lui ont accès et l'utilisent en entrée.

2.5.2. Fichiers "coordonnées" , "teneurs" et "facteurs".

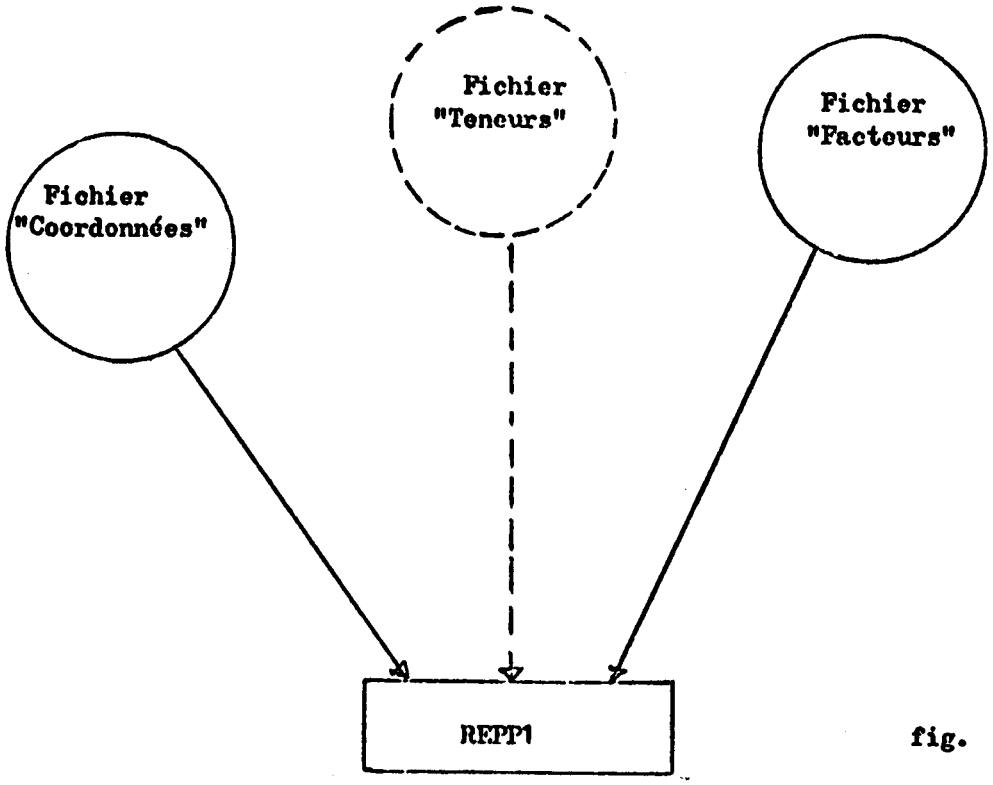
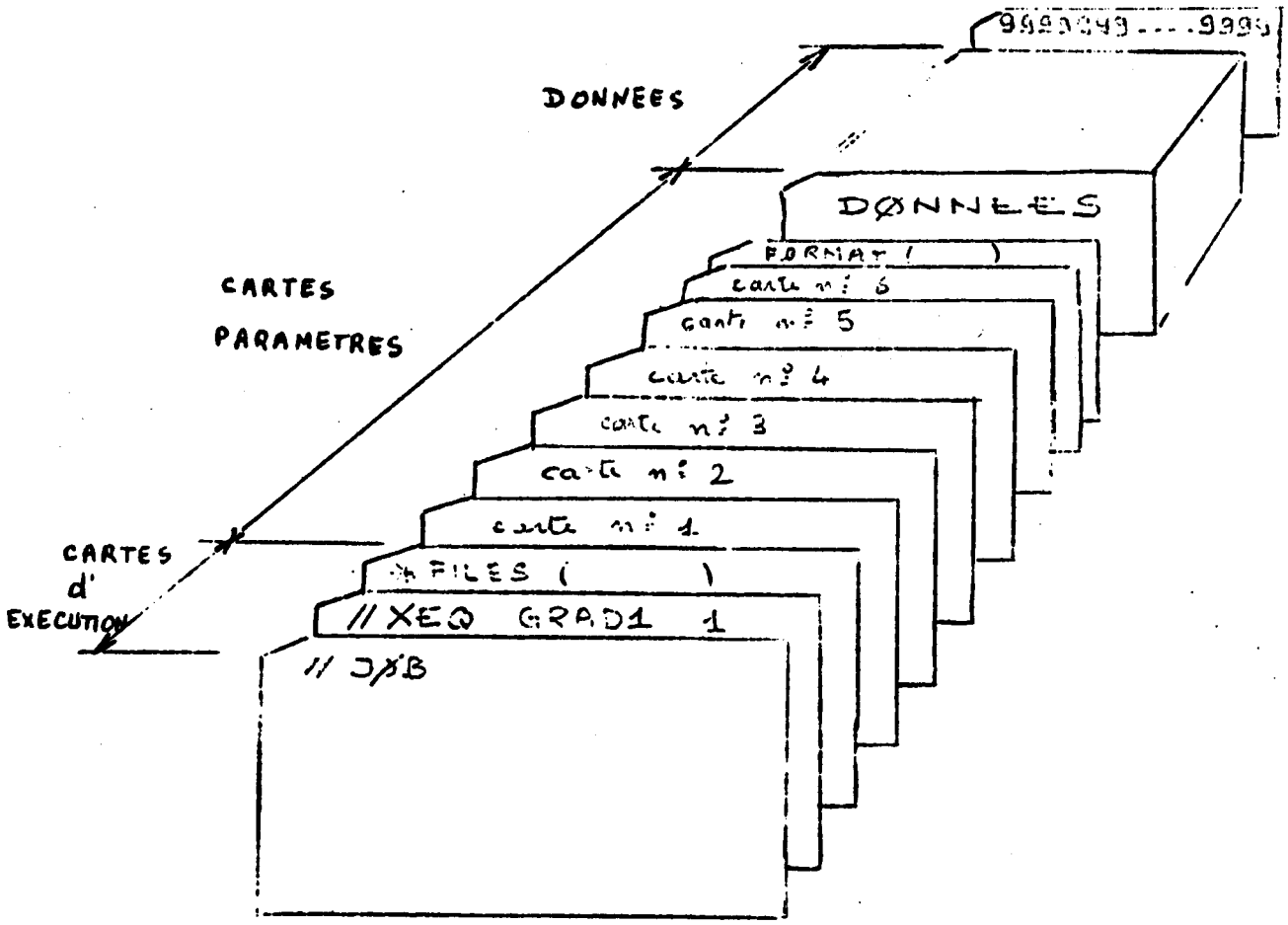


fig. 2.5.2.

Lorsque les coordonnées des échantillons sont connues, on peut faire de la cartographie par report de points, soit à partir des teneurs brutes, soit à partir des facteurs extraits par l'analyse et voir ainsi la répartition géographique des variables considérées.

NB Le programme REPP1 ne fait pas le tracé des courbes iso-valeurs.

2.6. Dessin du jaquet de cartes. (IBM 1130)



Les cartes paramètres correspondent au bordereau.

3. Méthodes descriptives élémentaires

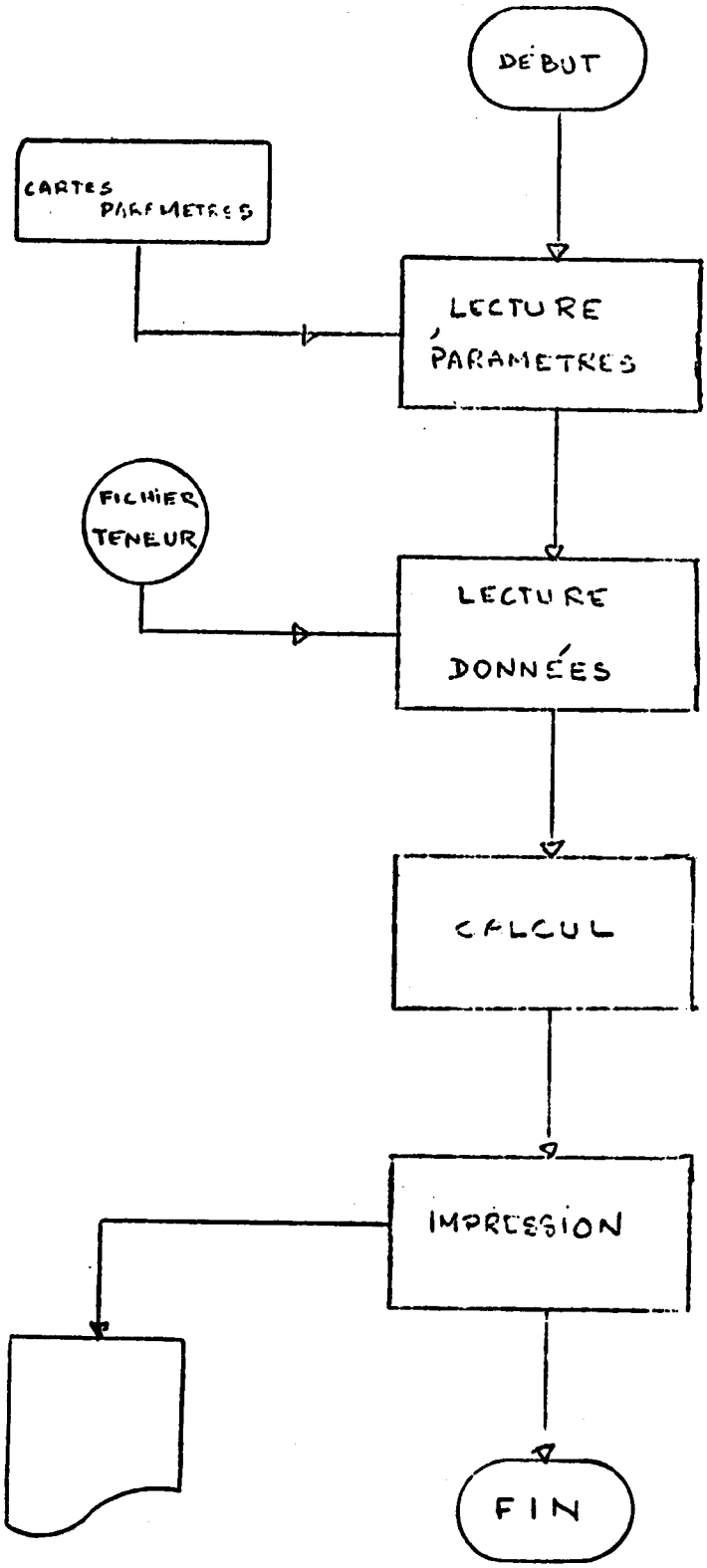
3.1. Introduction

Le bordereau modèle 1 est un bordereau d'exploitation des programmes statistiques élémentaires. Il est rempli par l'utilisateur et fournit à l'opérateur tous les renseignements nécessaires à l'exploitation des programmes suivants:

- Teneurs moyennes et écarts-types : TMET
- Histogrammes : HIST
- Matrice de corrélation : MATCØ
- Graphiques de données : GRAD
- Diagrammes triangulaires : DITRI
- Report de points : REPP1

3.2. Programme de calcul des moyennes et écarts-types (TMET)

3.2.1. Organigramme de Principe



3.2.2. Méthode utilisée dans le programme

Lorsqu'une série d'observations (caractère quantitatif) comporte N valeurs Q(1), Q(2), ..., Q(n) , la valeur arithmétique moyenne empirique s'écrit :

$$mQ = \frac{Q(1) + Q(2) + \ldots + Q(n)}{}$$

L'écart type (déviation standard) empirique est égale à la racine carrée de la moyenne quadratique des écarts à la moyenne ; c'est un paramètre de dispersion qui tient compte des écarts de toutes les valeurs observées :

$$\text{Ecart-type} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{n} \left(Q(i) - mQ\right)^2}$$

3.2.3. Dessin des cartes

Carte n° 1

Colonnes	1 2	5 8	9 12	17 18	
variables	Nbre de teneurs	Rang du premier échantillon à considérer	n° du dernier échantillon à considérer	Nbre de groupes retenus	

Carte n° 2 KTSUP = borne supérieure au delà de laquelle les valeurs ne sont pas prises en compte.

Colonnes	1	80
variables	KTSUP(I), I=1,20	

(Si plus de 20 variables carte n° 2 bis)

Colonnes	1	24
variables	KTSUP(I), I=21,24	

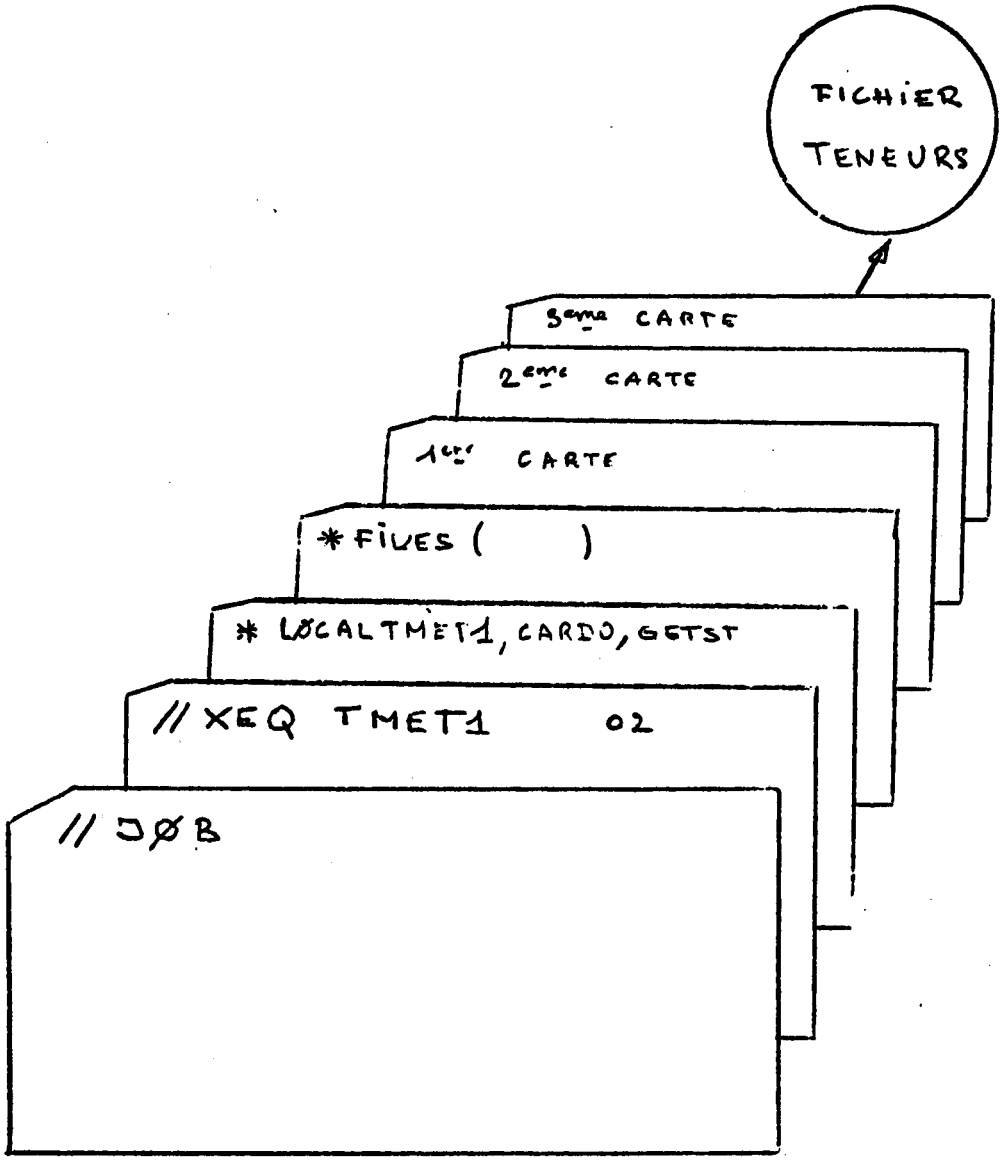
Carte n° 3 KTINF borne inférieure au dessous de laquelle les teneurs ne sont pas prises en compte

Colonnes	1	80
variable	KTINF(I), I=1,20	

(Si plus de 20 variables carte n° 3 bis)

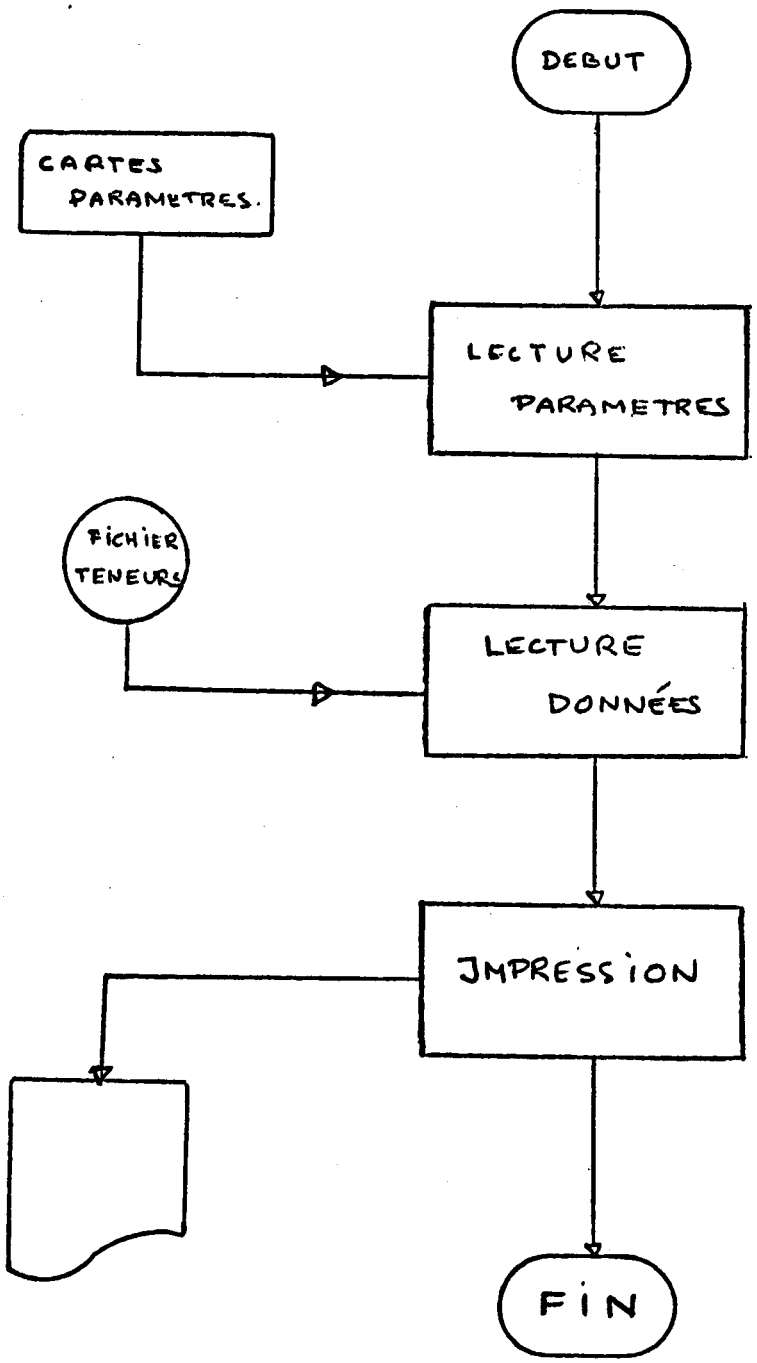
Colonnes	1	24
variables	KTINF(I), I=21,24	

3.2.4. Dessin du paquet de cartes (IBM 1130)



3.3.Programme de tracé d'histogramme (HIST)

3.3.1. Organigramme de Principe



3.3.2. Méthode utilisée dans le programme

On définit un intervalle constant comme largeur de classe ; en prenant chaque largeur de classe comme base, on construit un rectangle dont l'aire est proportionnelle à la fréquence de cette classe.

La réunion de plusieurs rectangles constitue l'histogramme des fréquences.

Chaque rectangle peut être matérialisé de trois façons différentes.

- par les identificateurs des individus appartenant à chacune des classes .
- par les numéros de groupe des individus appartenant à chacune des classes.
- par des astérisques.

3.3.3. Dessin des cartes .

Carte n° 1

Colonnes	1 2	3 4	5 8	9 12	13 16	80
variables	Nbre de teneurs	Nbre de classes	Rang du 1 échantillon à considérer	Numéro du dernier échantillon à considérer	paramètre permettant de matérialiser l'histogramme	option

option = 0 Sélection des variables sur le clavier
option = 1 Sélection des variables par carte paramètre
Dans les colonnes 13 à 16 0001 = Numéro 1000 = ***

Carte n° 2 KTSUP : borne au dela de laquelle les teneurs ne sont pas prises en compte.

Colonnes	1	80
variables	KTSUP(I), I=1,20	

(Si plus de 20 variables carte 2 bis)

Colonnes	1	24
variables	KTSUP(I), I=21,24	

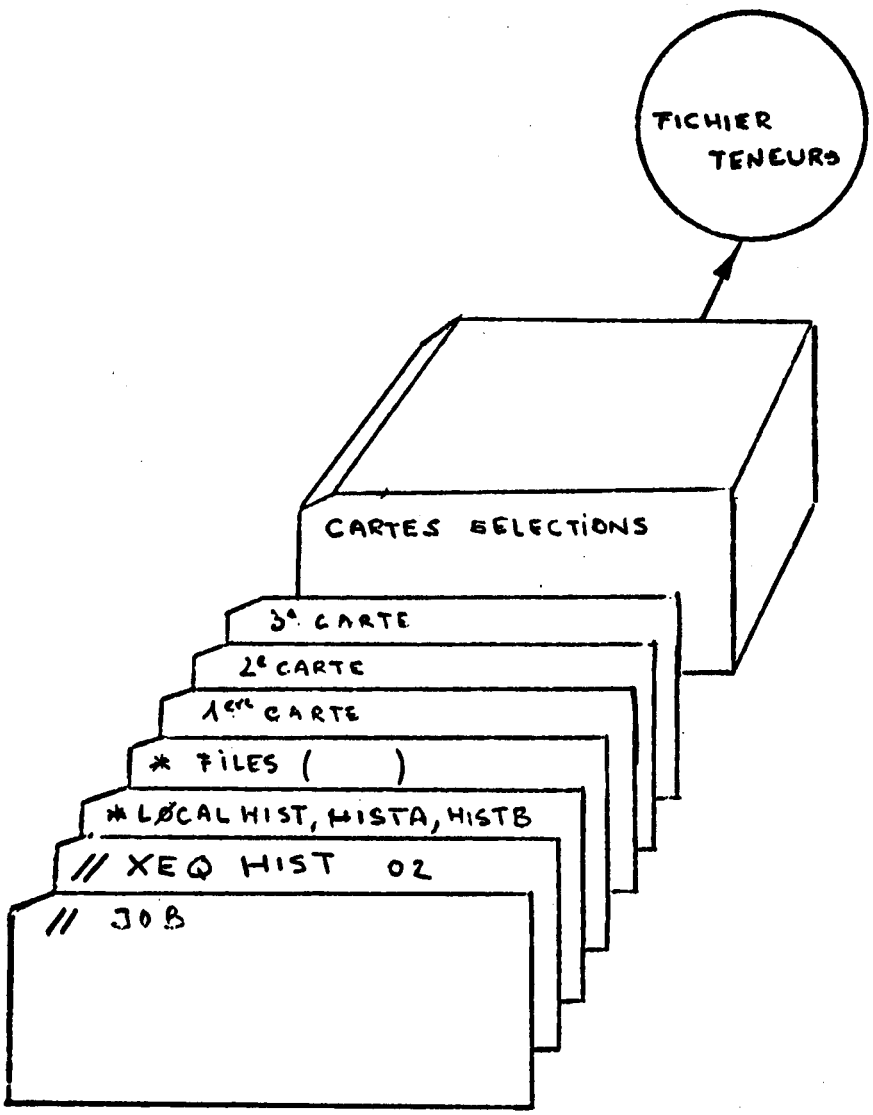
Carte n° 3 KTINF : borne inférieure au dessous de laquelle les teneurs ne sont pas prises en compte.

Colonnes	1	80
variables	KTINF(I), I=1,20	

(si plus de 20 variables carte 3 bis)

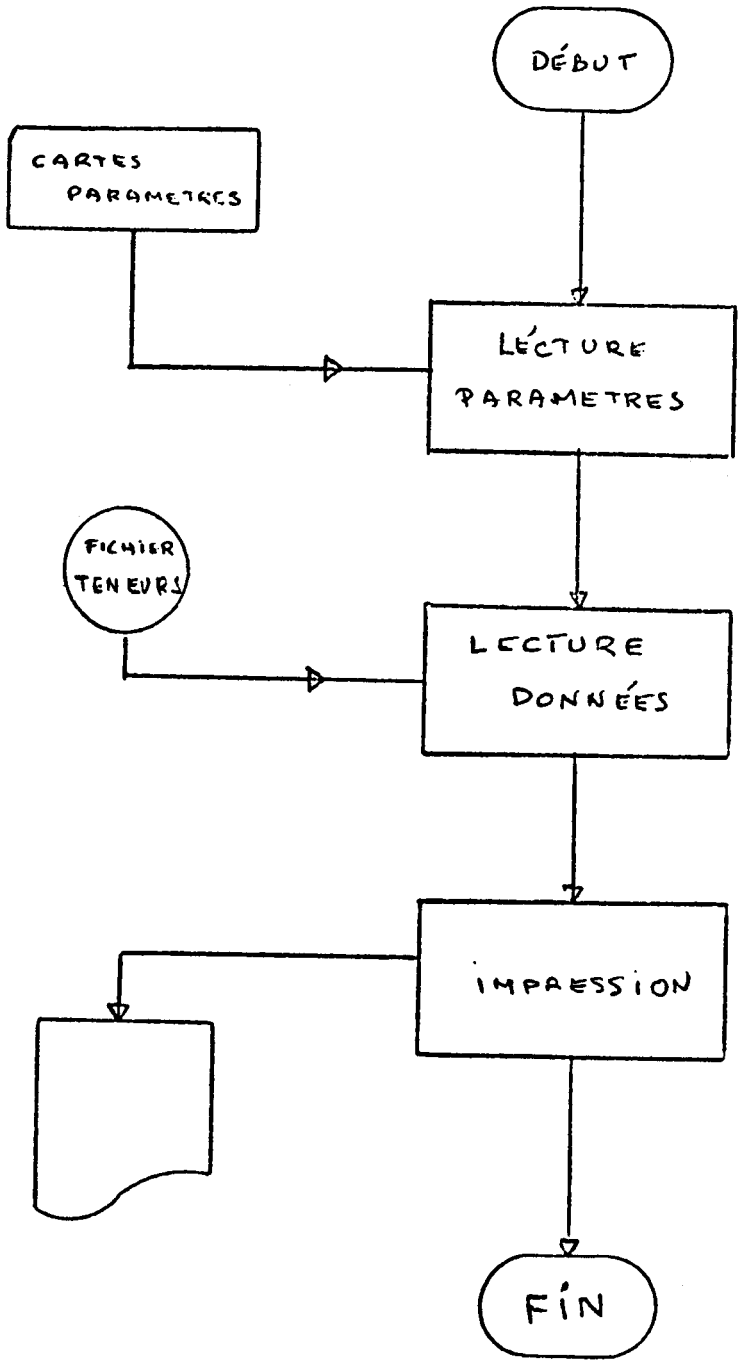
Colonnes	1	24
variables	KTINF(I), I=21,24	

3.3.4. Dessin du paquet de cartes (IBM 1130)



3.4. Programme de tracé de graphiques de données (GRAD)

3.4.1. Organigramme de principe.



3.4.2. Méthode utilisée dans le programme.

Etant donnée une série d'observations à plusieurs variables, ce programme permet de voir s'il y a une relation entre ces variables prises deux à deux.

La représentation par un nuage de points done une idée de l'existence possible de ces relations.

Il existe trois versions différentes.

- Graphique donnant les numéros de groupe des points observation: en fonction des valeurs des deux variables prises en compte.
- Graphique donnant la position des points observation en fonction des valeurs des deux variables prises en compte.
- Graphique donnant les identificateurs des points observation en fonction des valeurs des deux variables prises en compte.

3.4.3. Dessin des cartes

Carte n° 1

Colonnes	1 2		5 8	9 11		79 80
variables	Nbre de teneurs		Rang du 1 échantillon à considérer	n° du dernier échantillon à considérer		option indic

Option = 0 Sélection des variables sur le clavier.
Option = 1 Sélection des variables par carte paramètre.
Indic = 0 échelle calculée.
Indic = 1 échelle paramétrée.

Carte n° 2 KTSUP borne supérieure au dela de laquelle les teneurs ne sont pas prises en compte.

Colonnes	1	80
variables	KTSUP(I), I=1,20	

(si plus de 20 variables carte n° 2 bis)

Colonnes	1	24
variables	KTSUP(I), I=21,24	

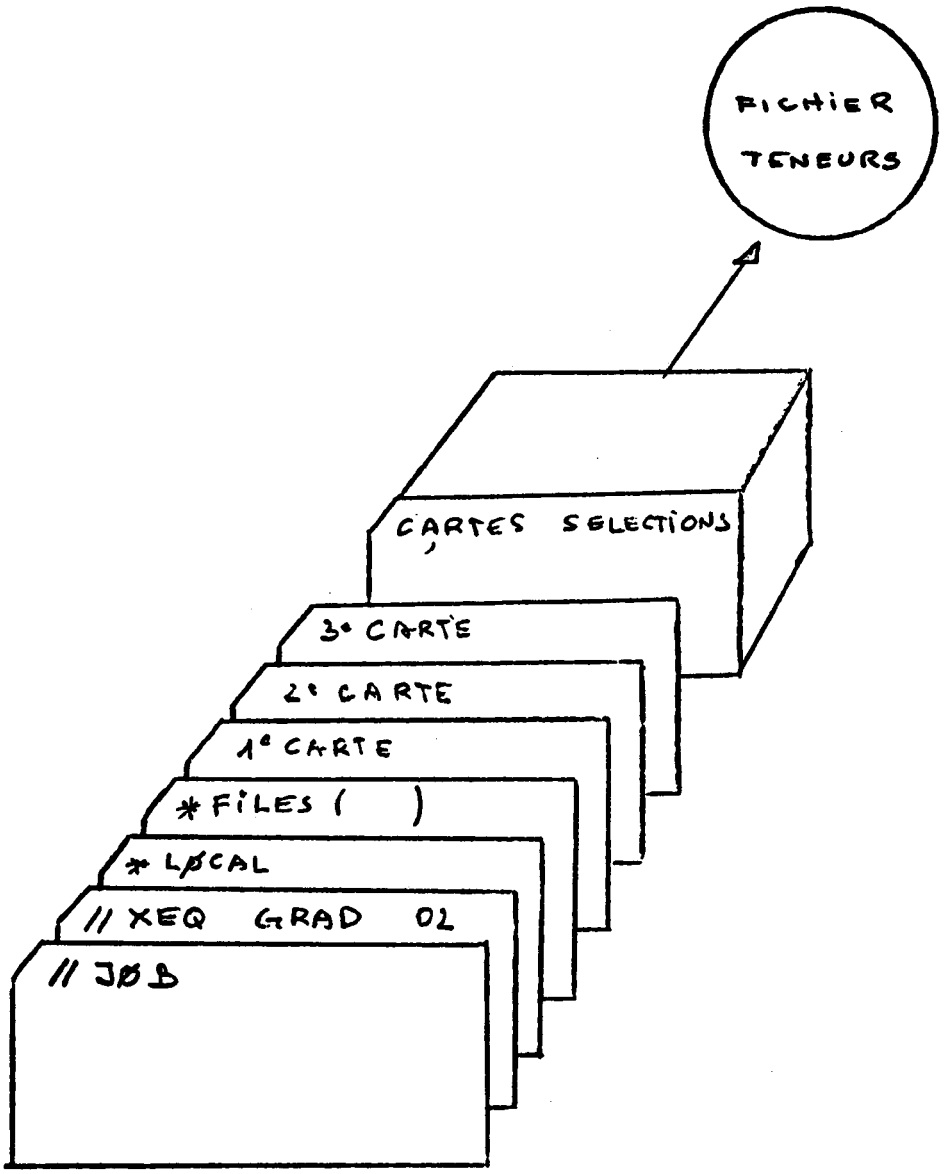
Carte n° 3 KTINF borne inférieure au dessous de laquelle les teneurs ne sont pas prises en compte.

Colonnes	1	80
variables	KTINF(I), I=1,20	

(Si plus de 20 variables carte n° 3 bis)

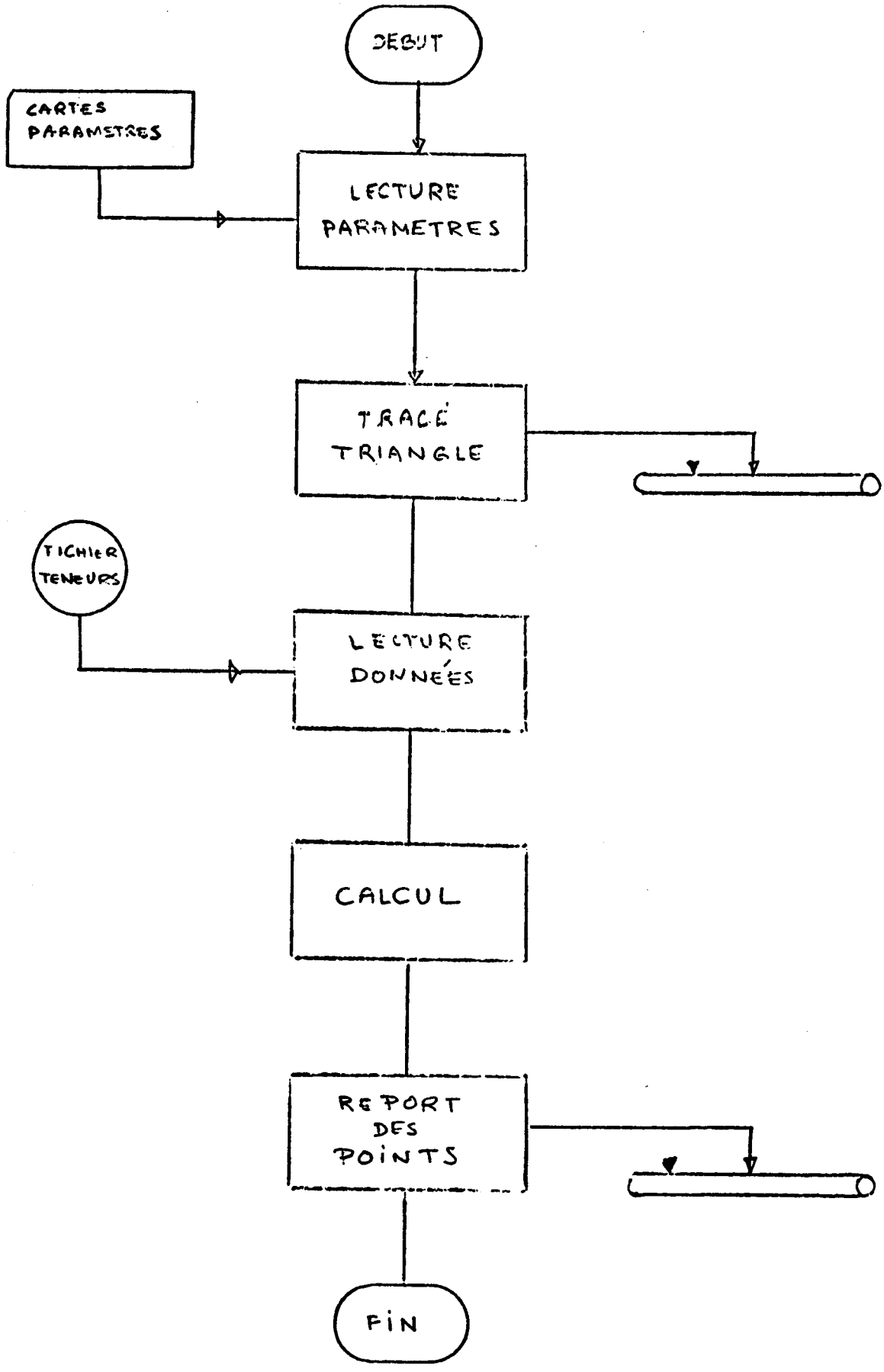
Colonnes	1	24
variables	KTINF(I), I=1,24	

3.4.4. Dessin du paquet de cartes (IBM 1130)



3.5. Programme de tracé de diagrammes ternaires (DITRI)

3.5.1. Organigramme de principe.



3.5.2. Méthode utilisée dans le programme

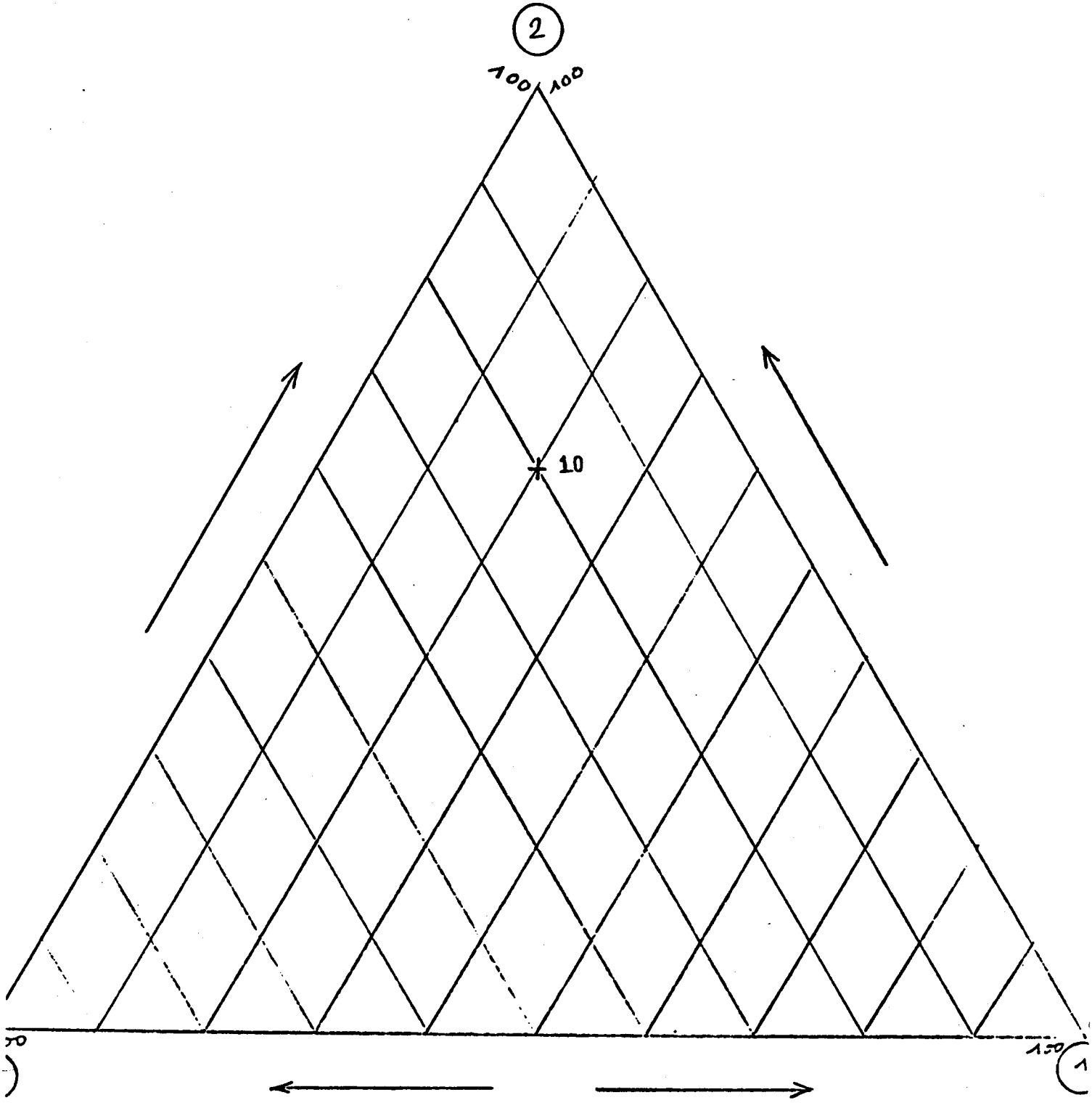
Etant donnée une série d'observations à plusieurs variables, ce programme permet de représenter des observations par rapport à trois variables dont la somme est constante. (voir fig 3.5.3.)

La représentation par un nuage de points montre, les groupements, de points-observations et donne une idée de l'existence possible de relations entre les variables sélectionnées.

Trois options sont possibles :

- Chaque point est représenté par son identificateur.
- Chaque point est représenté par son numéro de groupe.
- Chaque point est représenté par une croix.

3.5.3. Diagramme ternaire (exemple)



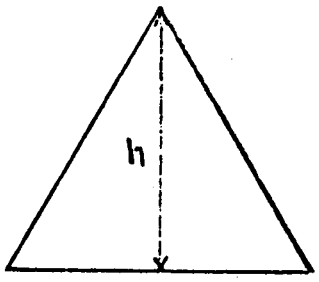
Ex : l'échantillon n°10 possède pôle 2 : 60 %
1 : 20 %
3 : 20 %

3.5.4. Dessin des cartes.

carte n° 1

1 4	5 8	9 10	11 12	13 36
I4	I4	I2	I2	24I1
hauteur du triangle en mm	Nombre d'observations	Nombre de variables	Forme d'édition des points	variables choisies comme pôle (ou sommets du triangle).

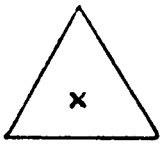
hauteur du triangle équilatéral :

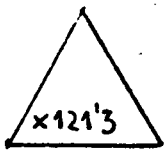


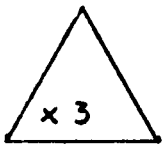
Nombre de variables : il s'agit de toutes les variables correspondant à une observation.

Forme d'édition des points :

valeur du paramètre :

1 ⟶  croix seule

0 ⟶  croix + identificateur complet

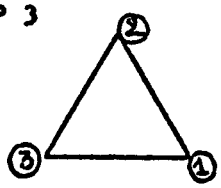
-1 ⟶  croix + n° de groupe seul.

- Variables choisies comme pôle

Cette sélection est opérée par un masque pouvant prendre quatre valeurs :

0	La variable ne sera pas retenue
1	La variable sera choisie comme pôle n° 1
2	La variable sera choisie comme pôle n° 2
3	La variable sera choisie comme pôle n° 3

La position des pôles (numérotation) est immuable :



Si on rencontre dans le masque, plusieurs fois la valeur 1 (ou 2, ou 3) cela signifie que la variable correspondant au pôle 1 (ou 2, ou 3) sera la somme de toutes les variables affectées de la valeur 1 (ou 2, ou 3).

Exemple :

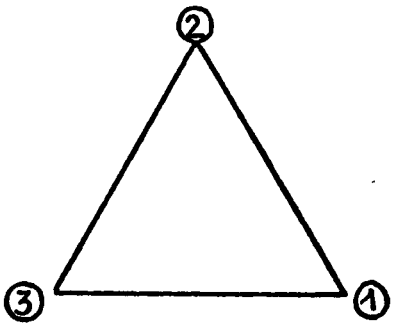
soit le masque

1	1	2	2	0	3	1	0	0	0	0	0

correspondant aux variables

CA	MG	NA	K	CL	Si	AL	MN	W	Ti	BA	GA

les pôles auront la structure :



(1) = CA + MG + AL
(2) = NA + K
(3) = SI

carte n° 2

1 2	3 12
I2	10A1
Nombre de diagrammes à tracer	type de chaque diagramme à tracer

- Nombre de triangles

On a vu que l'identificateur d'une observation comporte un numéro de groupe qui peut être égal à (0, 1, 2, 3, ..., 9).

On peut représenter toutes les observations relatives à un même groupe, dans un diagramme ternaire et faire donc autant ou moins de diagrammes que de groupes.

Le groupe 0 permet de traiter toutes les observations, quelque soit leur numéro de groupe dans un même diagramme.

On indiquera ici le nombre de diagrammes (chacun correspondant à un groupe) à tracer.

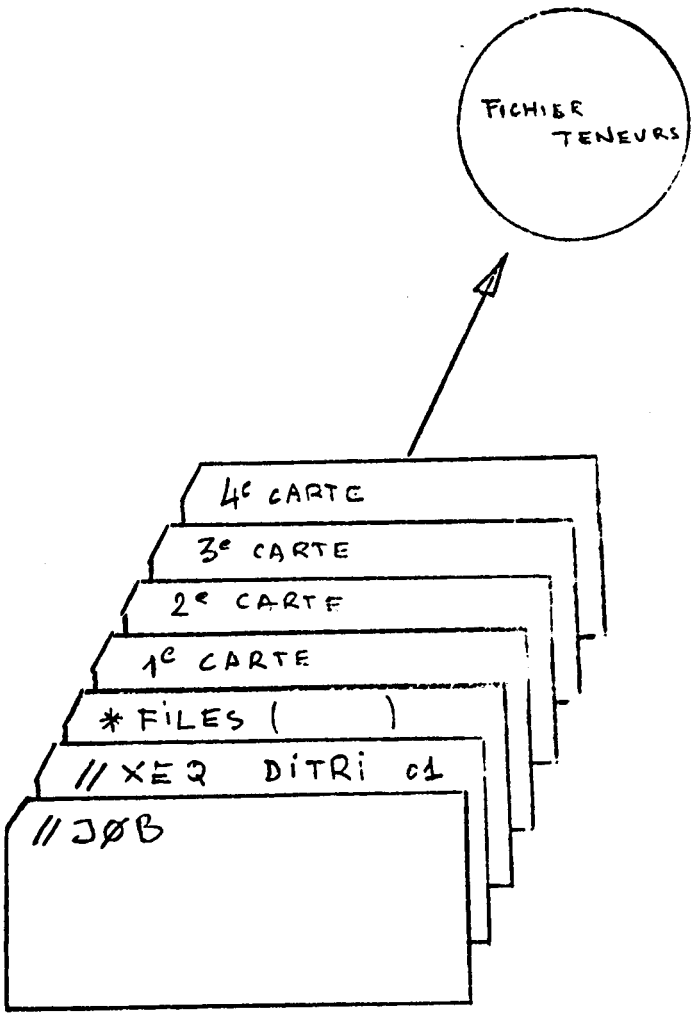
- Type de chaque diagramme.

On indiquera ici les n° de groupe successifs correspondants à chaque diagramme que l'on veut tracer.

carte n° 3 - carte n° 4 (du même type)

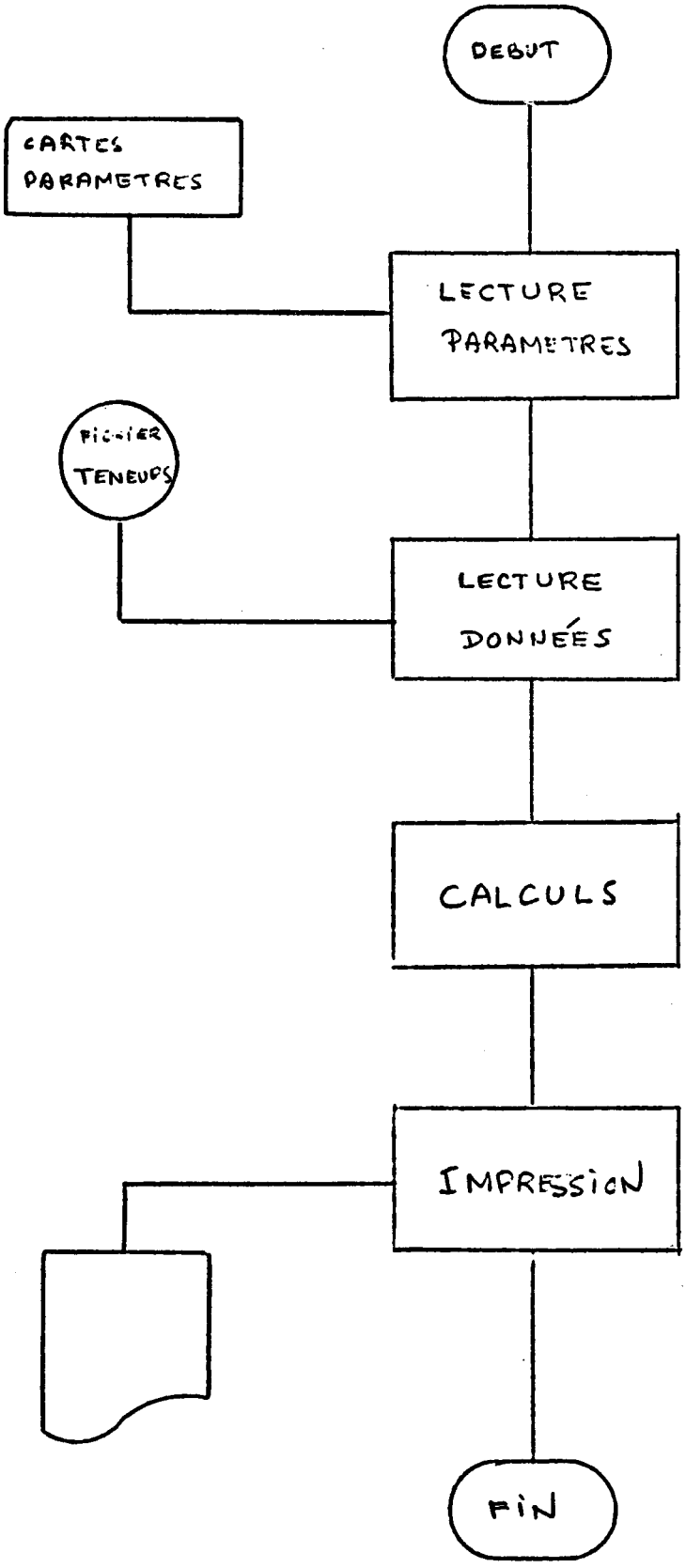
1 60	
30 A2	
Titre et nom du diagramme.	

3.5.5. Dessin de paquet de cartes (IBM 1130)



3.6. Programme de calcul des coefficients de corrélations. (MATCØ)

3.6.1. Organigramme de principe.



3.6.2. Méthode utilisée dans le programme

Le coefficient de corrélation est un indice numérique qui mesure l'intensité de la liaison linéaire entre deux grandeurs. Supposons que ces grandeurs correspondent aux colonnes k et l du tableau de données x(i,j). Soient mx(k) et mx(l) les moyennes des grandeurs k et l.

Le coefficient de corrélation entre les grandeurs xk et xl est défini par :

$$\frac{\text{covariance (xk, xl)}}{\sqrt{\text{variance (xk) * variance (xl)}}}$$

Cette relation s'écrit sous forme développée :

$$\frac{\sum_{i=1}^{n} (x(i,k) - mx(k)).(x(i,l) - mx(l))}{\sqrt{\sum_{i=1}^{n} \left[x(i,k) - mx(k)\right]^2 \sum_{i=1}^{n} \left[x(i,l) - mx(l)\right]^2}}$$

3.6.3. Dessin des cartes

carte n° 1 (à remplir par l'opérateur)

1 3	4 6
Numéro du fichier	Nouvelle longueur

carte n° 2

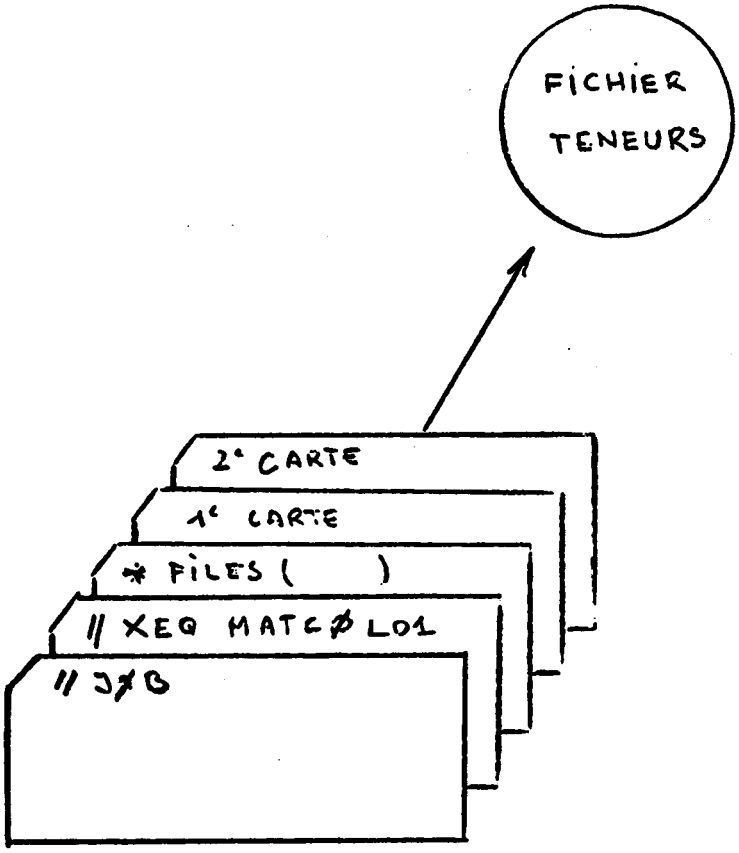
1 3	5 8	9 10	11	12 21
Nombre d'observations	Nombre de variables	Nbre de sélection	ØPTION	Numéro des groupes sélectionnés

Le nombre de sélections correspond au nombre de groupes pour lesquelles nous voulons une matrice de corrélation.

OPTIØN = 0 teneurs naturelles

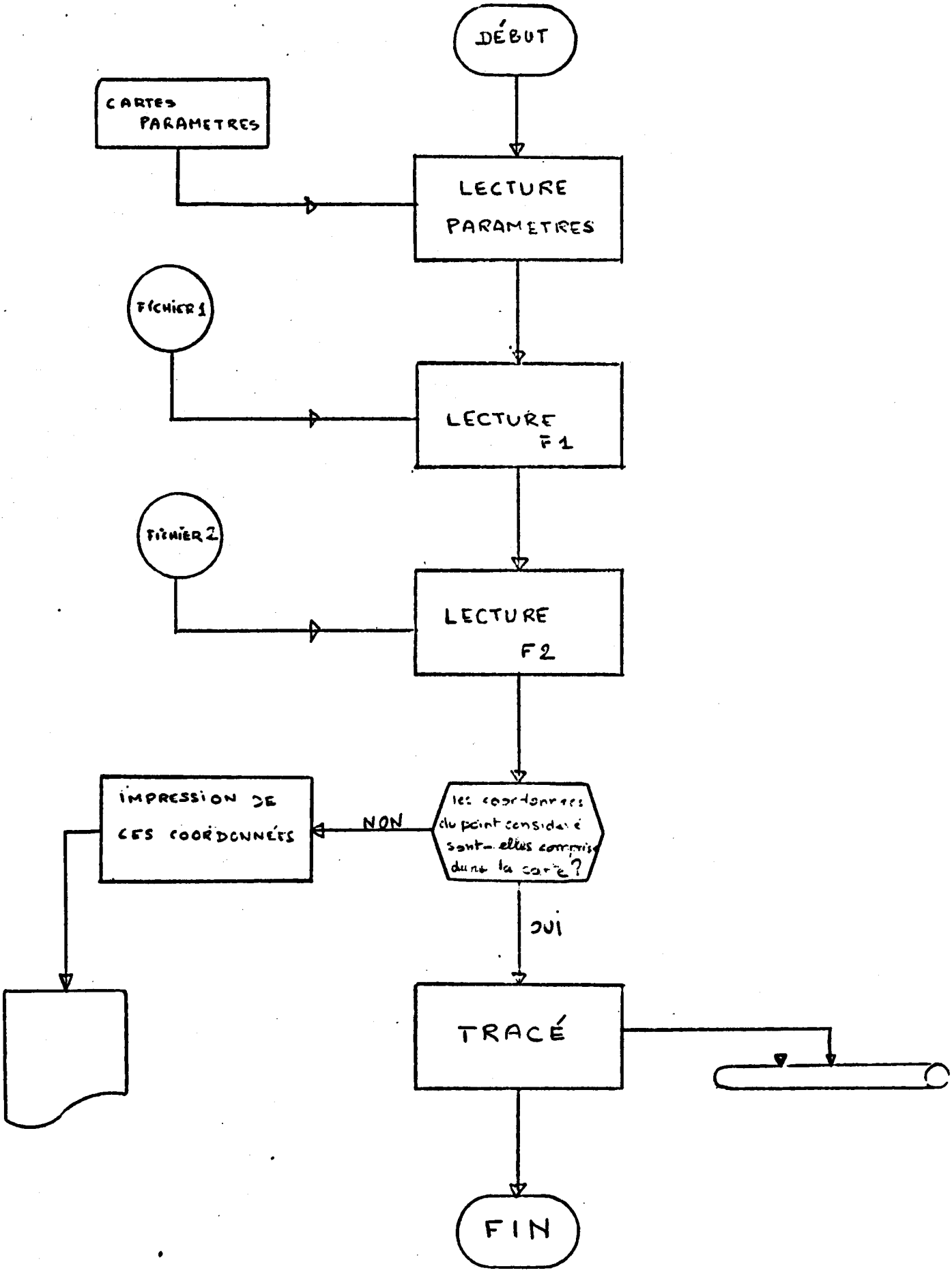
OPTIØN = 1 logarithmes des teneurs.

3.6.4. Dessin du paquet de cartes (IBM 1130)



3.7. Report de points (REP1)

3.7.1. Organigramme de principe.



3.7.2. Description du bordereau Modèle 5

On rencontre dans le bordereau trois types de paramètres : mise en page, description logique des données, mode de représentation des paramètres.

- Mise en page

On définira tout d'abord

- Les coordonnées de l'origine
- Les coordonnées maximales
- La distance entre axes à reporter.
- L'échelle en X
- L'échelle en Y
- Distance d'écriture des libellés de coordonnées par rapport au cadre.
- Taille en cm des caractères.
- Angle d'écriture des caractères par rapport à l'axe horizontal.

- Description logique des données.

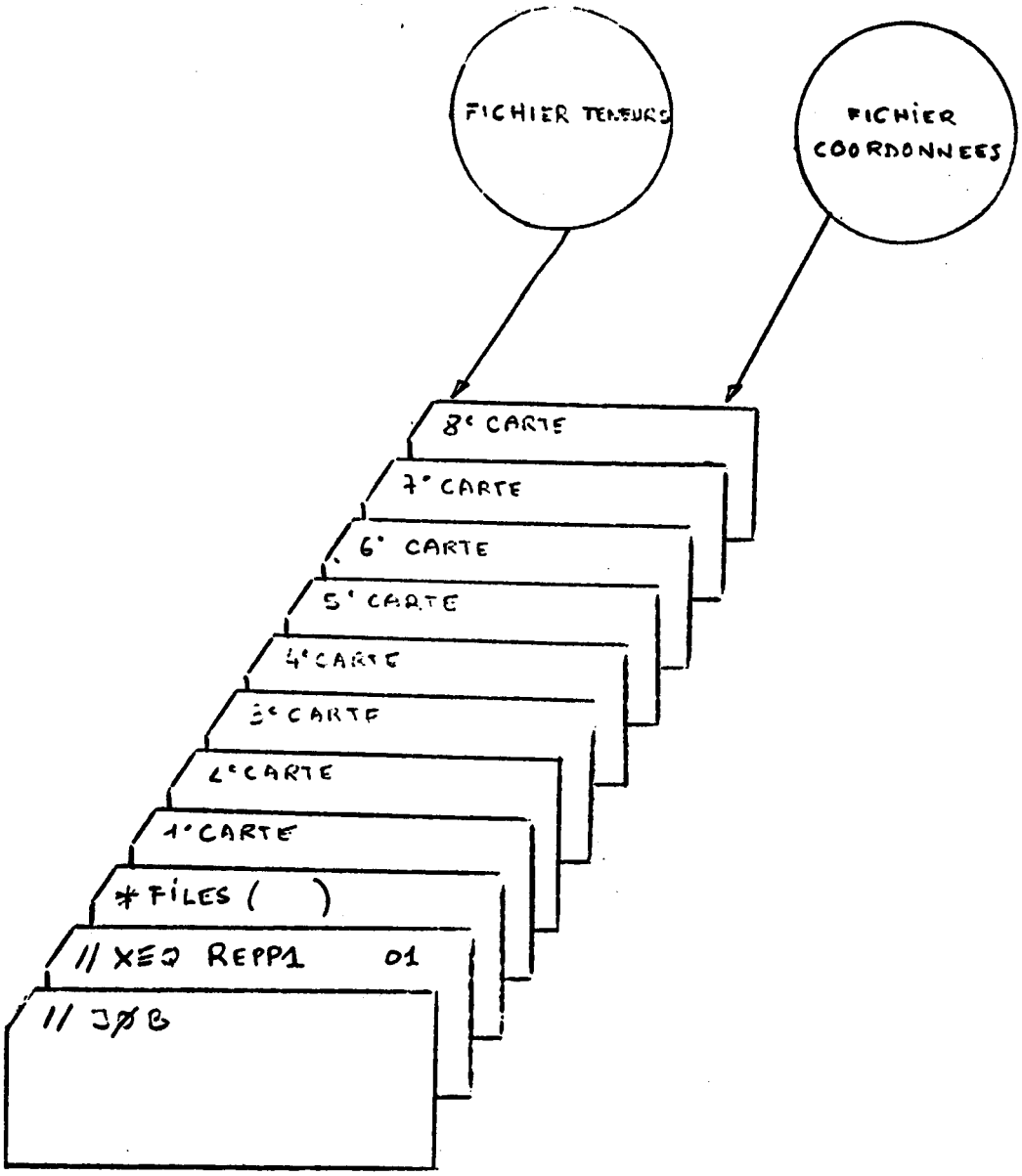
- Numéro du 1er enregistrement à prendre en compte.
- Numéro du dernier enregistrement à prendre en compte.
- La longueur de l'identificateur du fichier coordonnées.
- Le nombre de variables.
- Le numéro du groupe sélectionné.
- Un masque permet de reporter l'identificateur en totalité ou seulement en partie.

- Mode de représentation des paramètres.

Trois possibilités de report.

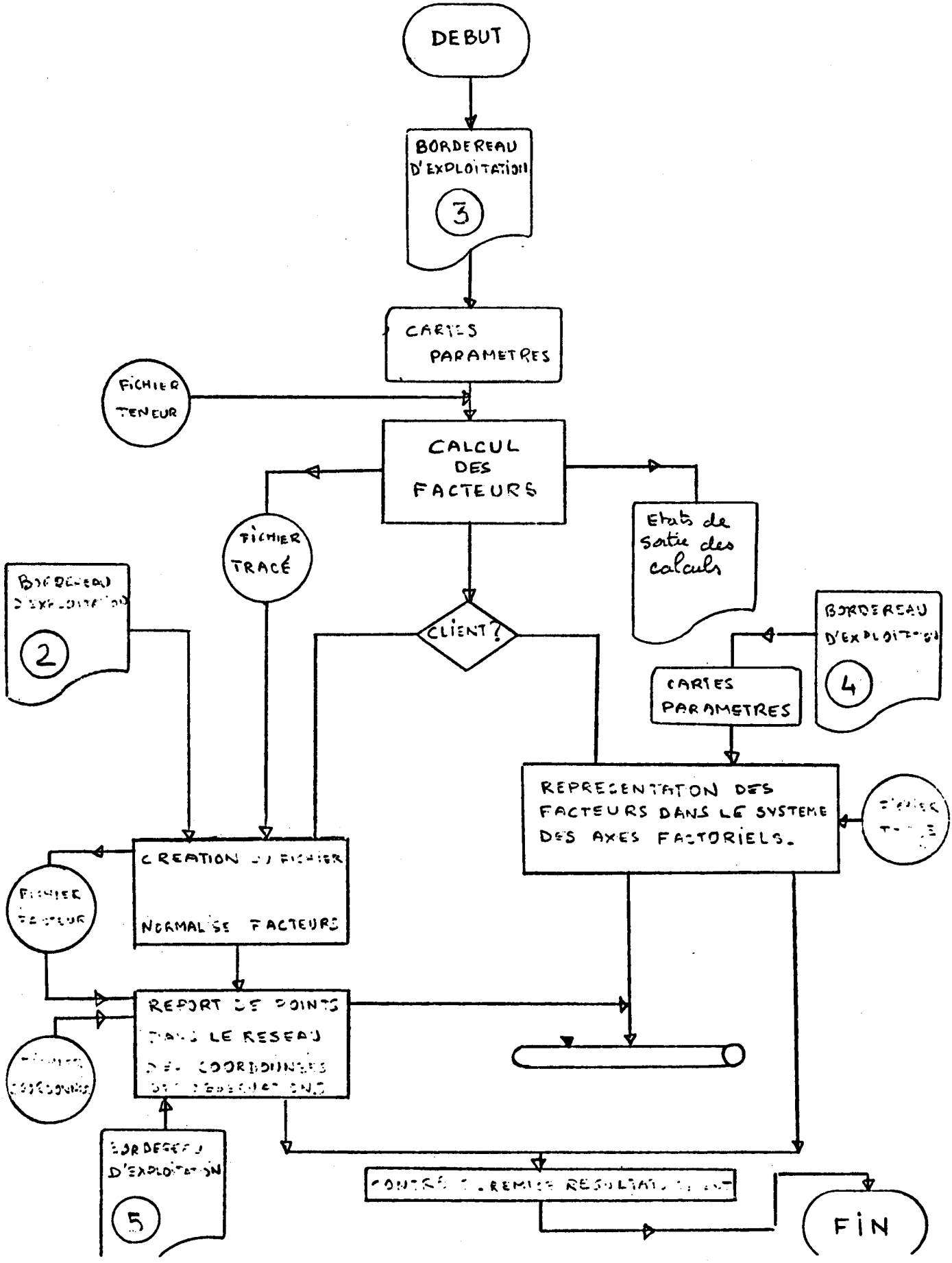
- report numéro seul.
- report numéro et paramètre.
- report paramètres seuls .
- pour 1 point donné on pourra reporter jusqu'à cinq paramètres.

3.7.3. Dessin du paquet de cartes (IBM 1130)



4.1. Analyse de données

4.1.1. Organigramme fonctionnel



Cet organigramme peut s'interpréter de la façon suivante :

- L'utilisateur remet à l'exploitation le bordereau modèle 3 qui fournit les cartes paramètres. En entrée, il dispose soit de son fichier teneur, crée au préalable, soit de son fichier de données sur carte.

- Un premier passage en machine lui fournit axes factoriels, facteurs F et G (quelque soit le mode d'analyse utilisé).

- Il récupère en sortie un état de résultats à l'imprimante et tous les renseignements pour la représentation graphique ultérieure, chargés sur un fichier tracé.

- Selon ce qu'il souhaite pour interpréter ces résultats, il utilise les bordereaux modèle 4 ou 5 : représentation des facteurs dans le système des axes factoriels ou report de points dans le réseau des coordonnées des observations (avec au préalable, création d'un fichier normalisé "facteurs" à partir du fichier tracé.

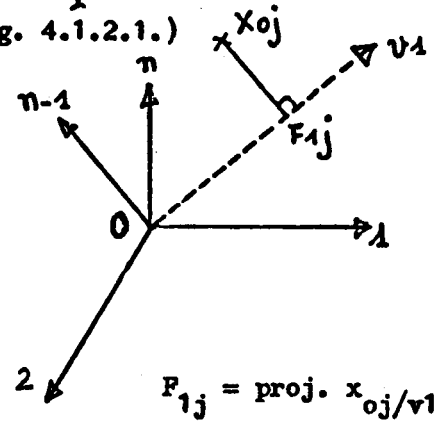
4.1.2. Facteurs, pourcentage d'explication, qualité de la représentation

On a vu, dans la première partie, que l'on pouvait reconstituer les positions des points (x_{ij}) d'un nuage de R^p ou de R^n , à partir d'un nombre restreint de facteurs. Ces facteurs sont les coordonnées des points initiaux sur la base d'un sous-espace qui ajuste au mieux ce nuage.

4.1.2.1. Définition des facteurs

- Facteurs F

Ce sont les composantes des P points variables de R^n dans le sous-espace R^l qui ajuste au mieux le nuage (R^l a pour base les vecteurs unitaires $v_1, v_2, \ldots, v_l$). L est le nombre d'axes extraits et conservés par l'utilisateur (voir fig. 4.1.2.1.)



On a vu que le vecteur v_ℓ est colinéaire à un vecteur propre de la matrice xx' (matrice de covariance).
Si x_{oj} est le vecteur variable associé à la variable j, dont on cherche la projection sur v_1, on a :

$$F_{1j} = \text{proj. } x_{oj/v1} = v'_1 \cdot x_{oj} = x'_{oj} \cdot v1$$

fig 4.1.2.1.

C'est encore la composante du vecteur $X' . V_1$ (sous forme matricielle, avec x' = transposée de X) .

Mais on sait que $U_1 = \frac{1}{\sqrt{\lambda_1}} \cdot x' V_1$ (voir première partie § 4.1.3.)

Par suite $F(1,j) = U(1,j) \sqrt{\lambda_1}$

Soit en généralisant : $\boxed{F(k,j) = U(k,j) \sqrt{\lambda k}}$ $k=1,L$, $j=1,P$

- Facteurs G

Ce sont les composantes des n points observations de R^p dans le sous-espace R^m qui ajuste au mieux le nuage (R^m a pour base les vecteurs U1, U2, ..., U_m). m est le nombre d'axes extraits et conservés par l'utilisateur (voir fig. 4.1.2.2.)

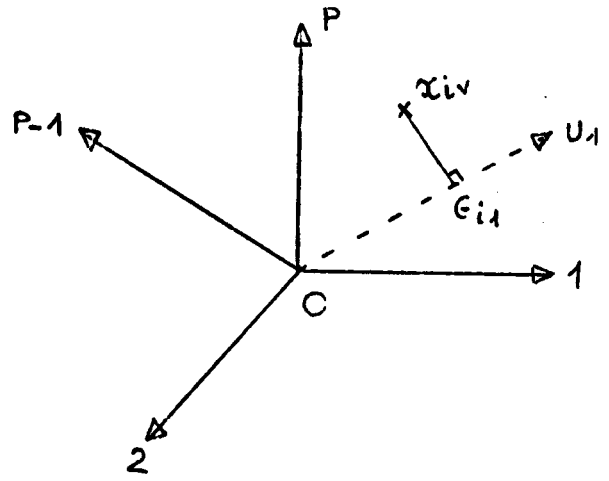


fig 4.1.2.2.

On a vu que le vecteur U_k est colinéaire à un vecteur propre de la matrice X'.X.

Si x_{iv} est le vecteur observation associé à l'observation i, dont on cherche la projection sur U_1 par exemple, on a :

$$G(i,1) = \text{proj. } x_{iv/U1} = U'_1 \cdot x_{iv} = x'_{iv} \cdot U_1$$

en effectuant le produit indiqué, on obtient :

$$G(i,1) = \sum_{j=1}^{p} x(i,j) \cdot U(1,j)$$

soit en généralisant :

$$\boxed{G(i,k) = \sum_{j=1}^{p} x(i,j) \cdot U(k,j)} \qquad k=1,m \quad i=1,n$$

4.1.2.2. Facteurs supplémentaires.

Dans certains cas, il peut être intéressant de projeter des points - variables ou des points - observations dans les plans factoriels de R^L ou R^m définis plus haut sans pour autant que ces points aient contribué à déterminer les axes factoriels. On dit qu'ils n'ont pas participé à l'analyse mais on souhaite quand même les représenter par leur projection dans R^1 ou R^m

Il peut s'agir de points différents des autres par nature ou correspondant à des anomalies qui déformeraient l'analyse.

- Facteurs F supplémentaires

Ce sont les composantes des points - variables supplémentaires de R^n dans le sous espace vectoriel R^1

Si x_{ojs} est le vecteur variable associé à la variable supplémentaire js, dont on cherche la projection sur v1 par exemple, on obtient :

$$F(_{1,js}) = \text{proj. } x_{ojs/v1} = v'_1 \cdot x_{ojs} \cdot v_1$$

or $$v1 = \frac{1}{\sqrt{\lambda_1}} \; x.u_1$$ (voir § 4.1.3. 1ère partie)

et sachant que $G(i,1) = x'_{iv} \cdot v_1$

Il vient : $F(1,js) = \sum_{i=1}^{i=p} x(i,js) \cdot G(i,1) / \sqrt{\lambda_1}$

en généralisant : $$\boxed{F(k,js) = \sum_{i=p}^{i=p} x(i,js) \cdot G(i,k) / \sqrt{\lambda_k}} \qquad k=1,m$$

- Facteurs G supplémentaires

Ce sont les composantes de points-observations supplémentaires de R^p dans le sous-espace vectoriel R^m

Si $x_{is,v}$ est le vecteur observation supplémentaire associé à l'observation is, dont on cherche la projection sur U_1 par exemple, on a :

$$G(is,1) = \text{proj. } x_{isv/U1} = U'_1 \cdot x_{is,v} = x'_{is,v} \cdot U_1$$

en effectuant le produit indiqué, on obtient :

$$G(is,1) = \sum_{j=1}^{p} x(is,j) \cdot U(1,j)$$

en généralisant :

$$\boxed{G(is,k) = \sum_{j=1}^{p} x(is,j) \cdot U(k,j)} \qquad k=1,m$$

4.1.2.3. Pourcentage d'explication

Considérons par exemple l'espace R^p des points observation : l'axe U_1 peut permettre de localiser les points observations de façon satisfaisante si la première valeur propre λ_1 est très grande par rapport aux autres. En effet λ_1 mesure en projection la somme des carrés des distances à l'origine. Ainsi un repère constitué par les k premiers axes factoriels permettra de reconstituer correctement les points initiaux si $\lambda_1 + \lambda_2 + \ldots + \lambda_k$ représente une proportion importante de la trace (=variance totale).

Cette somme partielle ramenée à 100 indiquera donc le pourcentage d'explication pour les axes considérés.

4.1.2.4. Reconstitution du tableau initial

Pour reconstituer le tableau de départ, on utilise les coordonnées des points sur les axes factoriels et les cosinus directeurs des axes factoriels :

On a vu que : $X \cdot u_q = \sqrt{\lambda_q} \cdot v_q$

multiplions à droite par u'_q : $X \cdot u_q \cdot u'_q = \sqrt{\lambda_q} \cdot V_q \cdot U'_q$

sommons par rapport à P : $X\left\{\sum_1^p u_q \cdot u'_q\right\} = \sum_1^p \sqrt{\lambda_q} \cdot v_q \cdot u'_q$

Les u_q étant orthogonaux : $\sum_1^p u_q \cdot u'_q = 1$

Par suite $X = \sum_1^p \sqrt{\lambda_q} \cdot v_q \cdot u'_q$

Si on se limite aux k premiers facteurs ; et si

$$\lambda_{s+1}, \ldots, \lambda_q < \varepsilon$$

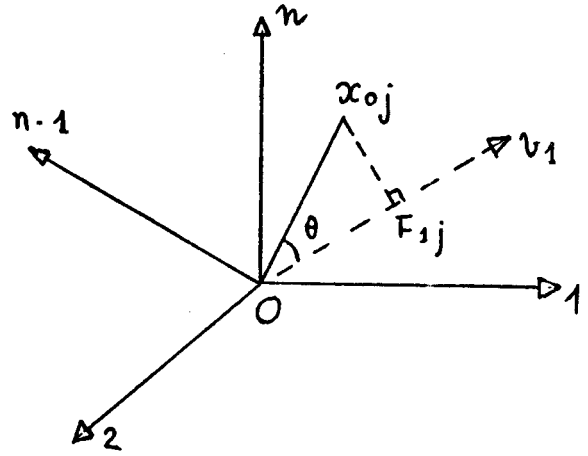
$$\boxed{X \# \sum_{q=1}^{k} \sqrt{\lambda_q} \cdot v_q \cdot u'_q}$$

4.1.2.5. Qualité de la représentation

Pour voir si un point variable ou point observation est bien représenté par sa projection (le facteur correspondant) sur les axes factoriels, il suffit de déterminer l'angle du vecteur joignant l'origine au point avec chacun des axes factoriels.

Angle des points variables

Plaçons nous encore dans R^n ; et cherchons l'angle du point variable j avec l'axe factoriel v_1 (voir fig.)



Il est clair que

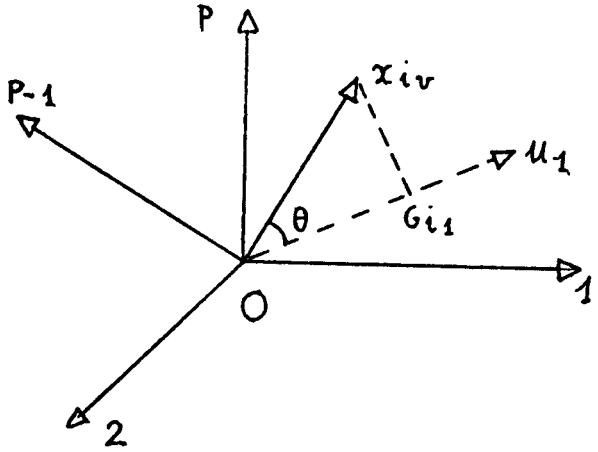
$$\cos^2\theta\ (j,1) = \frac{F^2(1,j)}{\|x(i,j)\|_2}$$

C'est à dire :

$$\cos^2\theta\ (j,i) = \frac{\lambda_1 u^2(1,j)}{\sum_{i=1}^{i=n} x^2(j,j)}$$

Angle des points observation

Plaçons nous dans R^p et cherchons l'angle du point observation i avec l'axe factoriel U_1 (voir fig.)



Il est clair que

$$\cos^2\theta\ (i,1) = \frac{G^2\ (1,i)}{\|x(i,j)\|_2}$$

C'est à dire :

$$\cos^2\theta\ (i,1) = \frac{G^2(1,i)}{\sum_{j=1}^{j=p} x^2\ (i,j)}$$

4.1.3. Représentation des échantillons et des éléments

Il y a deux représentations possibles : la représentation des facteurs F dans un sous espace vectoriel R^k (des axes factoriels) de R^n et celle des facteurs G dans un sous-espace vectoriel R^l de R^p.

Dans certaines analyses que nous verrons plus loin, on peut parler de représentation simultanée parce que les deux sous espaces R^k et R^l vectoriels commutent et que les facteurs F et G sont liés dans l'espace commun par des relations simples.

Nous nous bornerons ici à évoquer la représentation des facteurs F et G dans les deux espaces R^k et R^l distincts.

4.1.3.1. Représentation des facteurs F dans $R^k \oplus R^n$

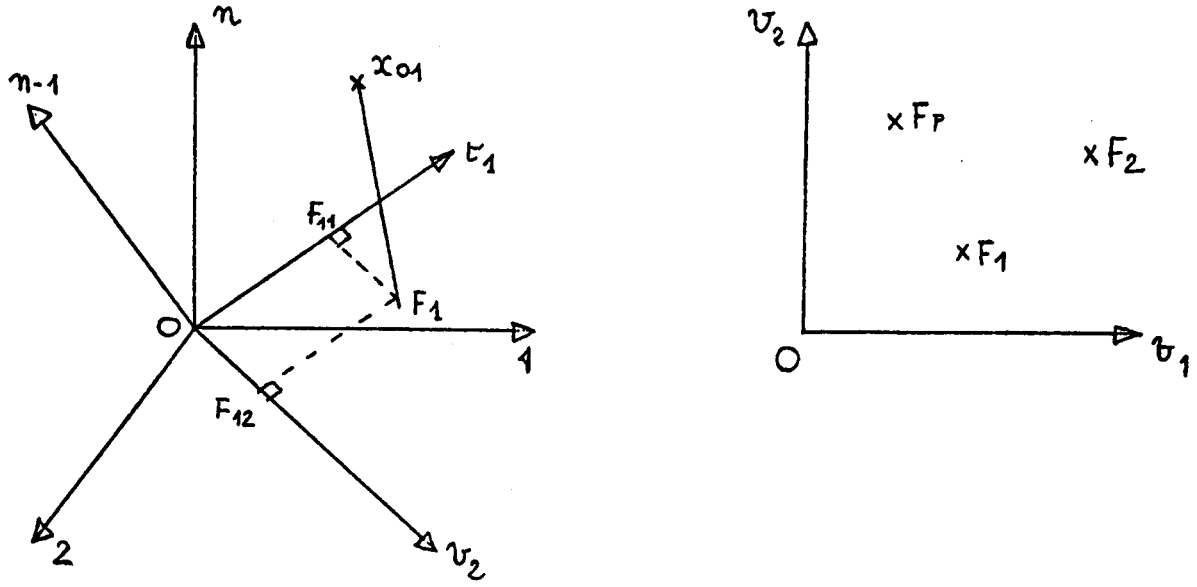


fig 4.1.3.1.

Les points variables se projettent en (F1, ... , Fp) sur les plans factoriels (V1, V2) (V1, V3) ... (V2, V3) (voir fig. 4.1.3.1.)

En étudiant ces graphes simultanément, on peut aisément reconstituer la position des points variables dans R^n et mettre en évidence le rôle des axes factoriels (V1,...,Vk). La qualité de la représentation fournit de précieux renseignements pour l'interprétation.

4.1.3.2. Représentation des facteurs G dans $R^l \subset R^p$

Les points observations se projettent en (G1, G2...Gn) sur les plans factoriels (u1, u2) (u1, u3) ... (u2, u3) (voir fig. 4.1.3.2.) En étudiant les groupements entre les points observations, on peut mettre en évidence le rôle des axes factoriels.

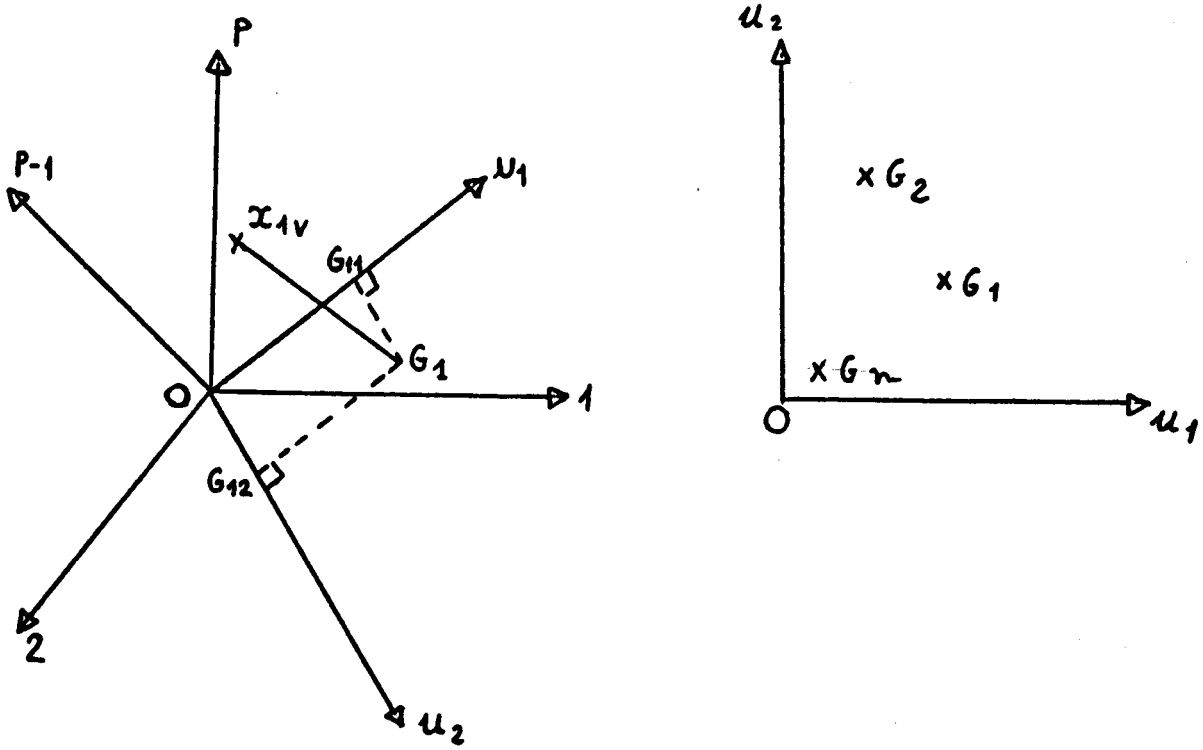
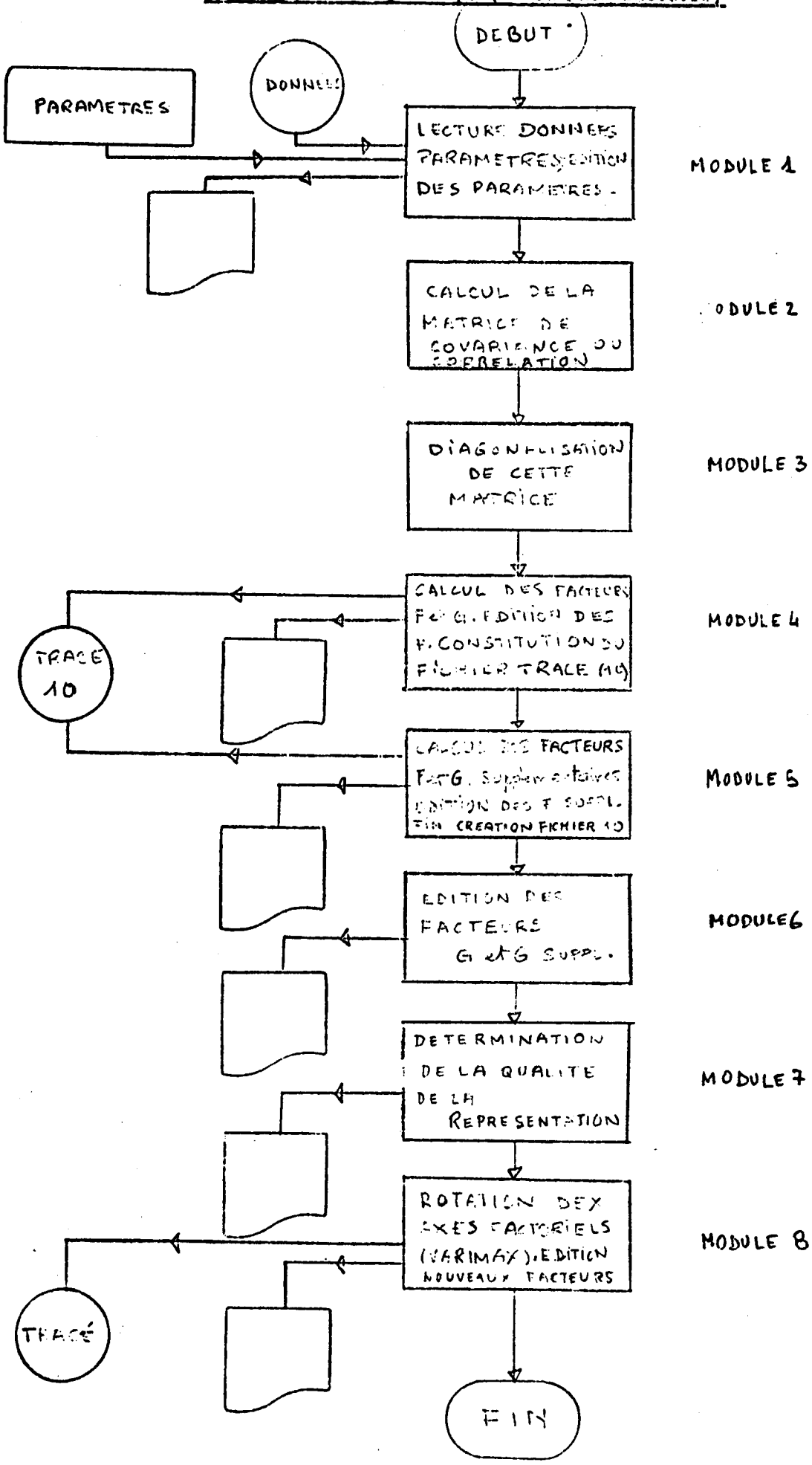


fig 4.1.3.2.

4.2. Analyse factorielle en composantes principales.

4.2.1. Organigramme de principe (calcul des facteurs)



4.2.2. Méthode utilisée dans le programme

Considérons toujours le tableau de données x(i,j) avec i=1,n et j=1,p.

L'analyse en composantes principales peut être effectuée sous trois formes selon le type de changement de variable adopté.

- variables centrées : on pose : $y(i,j) = x(i,j) - mx(j)$
- variables centrées et réduites par la moyenne : on pose : $y(i,j) = (x(i,j) - mx(j))/mx(j)$
- variables centrées et réduites par l'écart-type : on pose : $y(i,j) = (x(i,j) - mx(j))/s(j)$

$mx(j)$ est la moyenne de la variable x(j)

$s(j)$ est son écart-type.

Dans les trois cas, les facteurs sont obtenus à partir des valeurs propres de la matrice de covariance associé à x(i,j).

Toutes les définitions et les résultats du chapitre "analyse de données sont directement applicables pour l'analyse en composantes principales, pourvu que l'on exécute l'un des trois changements de variables ci-dessus et que l'on diagonalise la matrice de covariance associée.

4.2.3. Description du bordereau (modèle 3)

On rencontre dans le bordereau quatre types de paramètres : description physique des données, description logique des données, transformation des données, caractéristiques du traitement.

- Description physique des données

Lorsque les données sont lues sur carte, il convient de préciser :

- La position de l'identificateur par rapport aux valeurs (avant ou après). (voir § et fig 1.2.3.)
- Les noms des variables
- Les facteurs multiplicatifs (puissances de 10 par l'opposé desquelles il faut multiplier les valeurs sur carte pour restituer les vraies valeurs). (voir § et fig.1.2.4.)

Lorsque les données sont lues sur disque, noms des variables et facteurs multiplicatifs sont écrits sur le fichier.

- Description logique des données

Il s'agit de définir les dimensions du tableau de données :

- Le nombre de colonnes ou variables.
- Le nombre de lignes ou observations : pour les données sur carte, on indiquera le nombre exact d'observations à analyser. Pour les données sur disque, on indiquera le nombre d'observations à analyser augmenté de 3. Ceci s'explique par la structure du fichier (voir § et fig 2.4.1.). Ce nombre correspond en fait au pointeur de la dernière observation sur laquelle on effectue l'analyse.
- Le nombre de lignes ou observations supplémentaires : on indique alors les rangs de la première et de la dernière s'il s'agit de cartes ; les valeurs des pointeurs début et fin (rang augmenté 3) s'il s'agit de données sur disque. (voir § et fig 1.2.6.)

- Transformation des données

- Changement de variable : il est possible d'exécuter le traitement soit sur les valeurs centrées, soit sur les valeurs centrées et réduites par l'écart-type, soit sur les valeurs centrées et réduites par la moyenne. (voir fig 1 Tome II)
- Coefficients de pondération : chaque colonne peut être multipliée par un nombre compris entre 0,01 et 99.
- Logarithmes : on peut transformer toutes les données en logarithmes
- On a vu que le numéro d'identification contenait un numéro de groupe. Il est possible de n'exécuter l'analyse que sur un groupe bien déterminé.

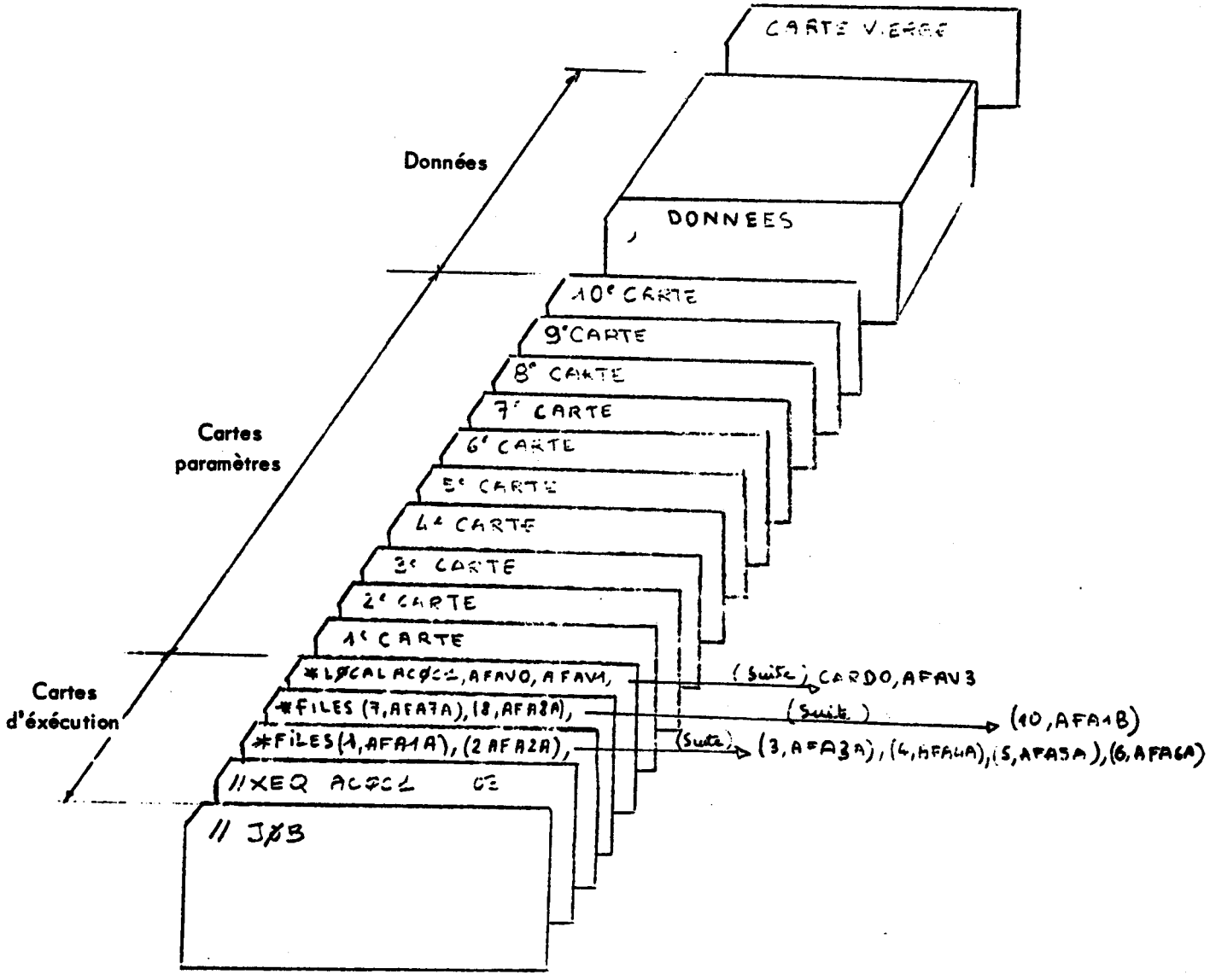
- Caractéristiques du traitement

- Le traitement peut être interrompu ou poursuivi après le calcul de la matrice de covariance ou corrélation.
- L'impression de cette même matrice peut être réalisée ou non sur option.
- Dans certains cas, on est amené à ignorer ou à représenter seulement certaines variables dans l'espace des axes factoriels déterminés par l'analyse.

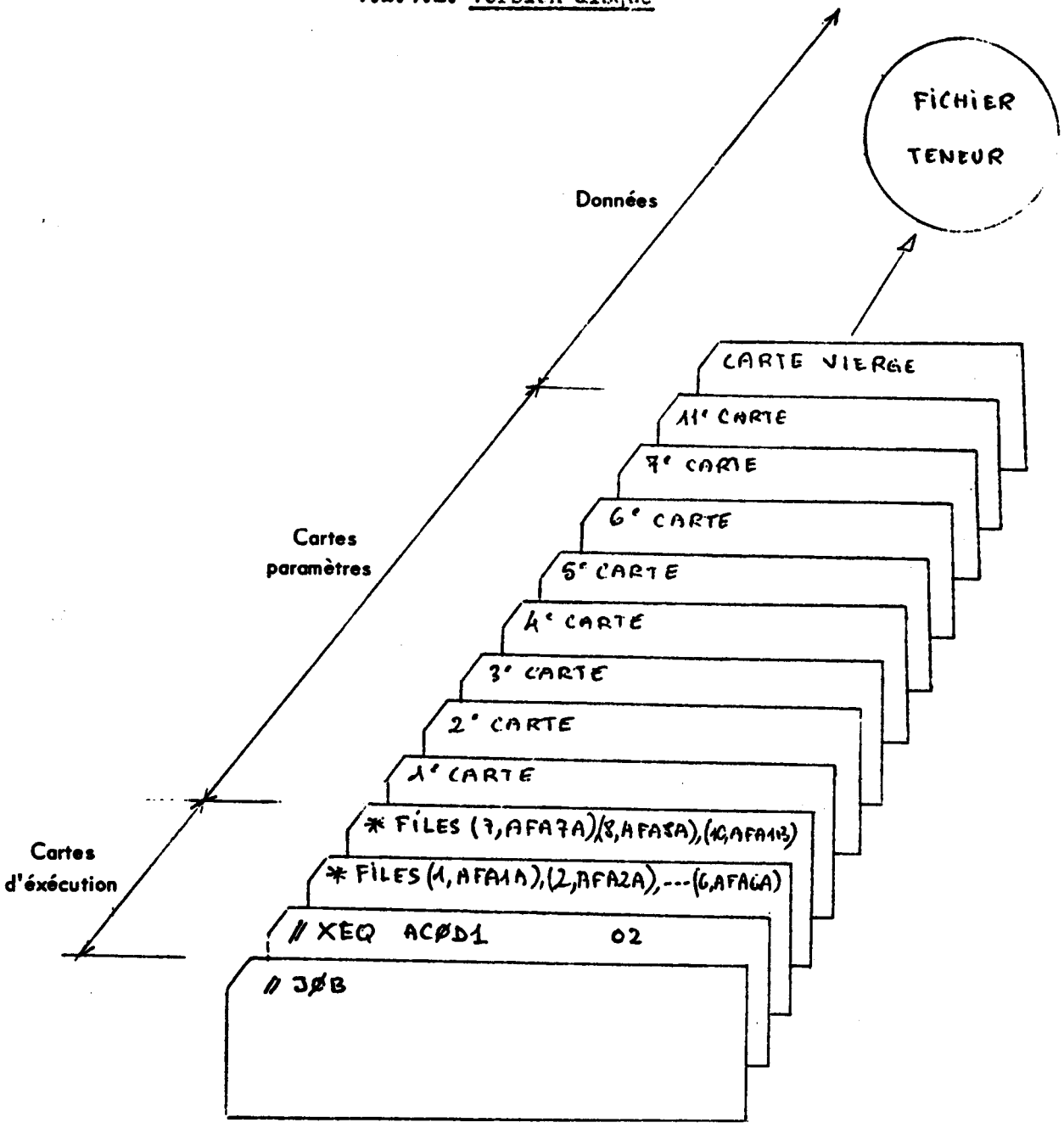
- Il faut indiquer le nombre de facteurs que l'on souhaite extraire. Plus ce nombre est élevé, plus le temps de calcul est important.
- Si l'on veut estimer la qualité de la représentation, on peut demander le calcul des angles des facteurs avec les axes factoriels.
- On peut faire tourner les axes factoriels, à la demande, en utilisant la méthode VARIMAX.

4.2.4. Dessin du paquet de cartes (IBM 1130)

4.2.4.1. Version carte

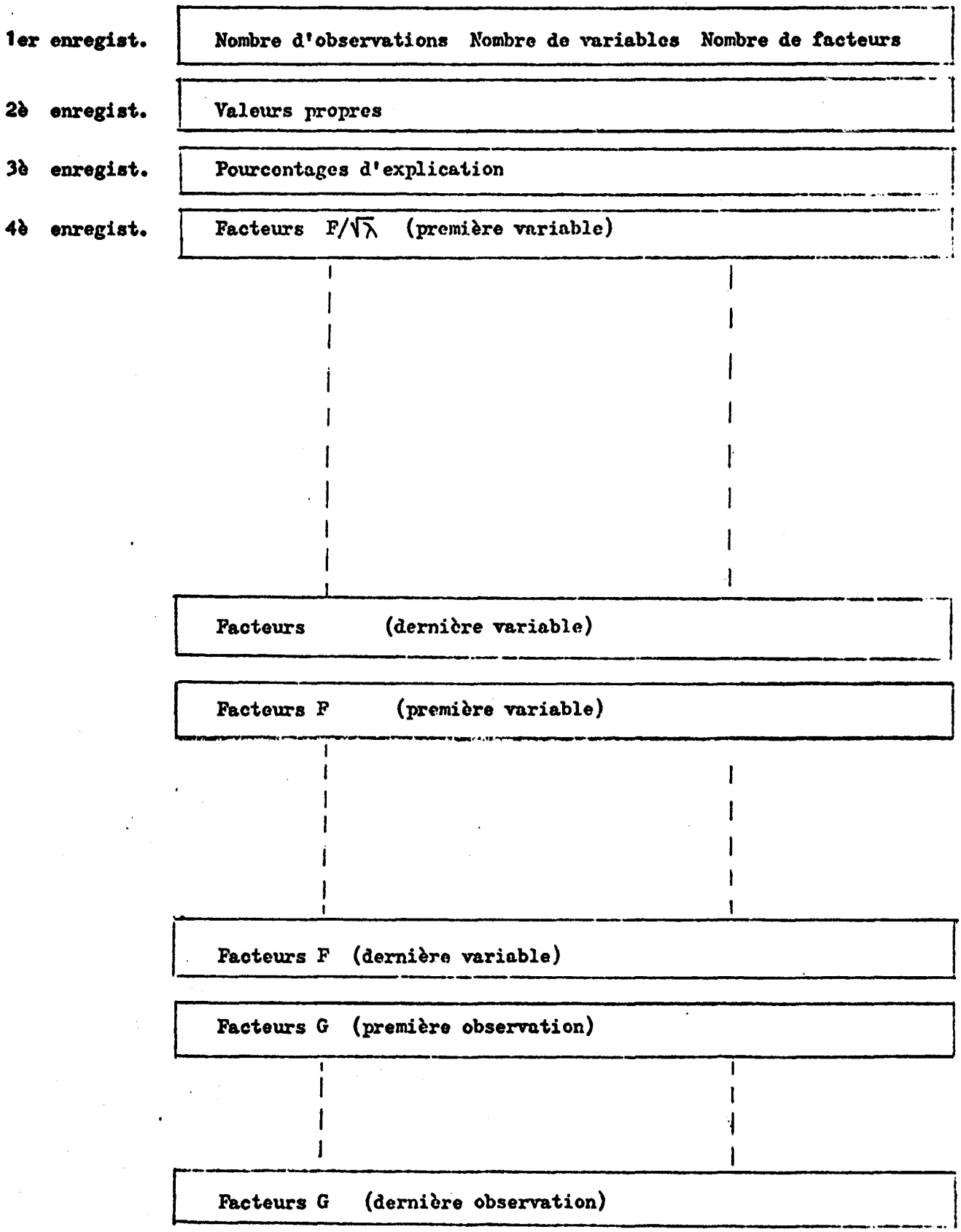


4.2.4.2. Version disque

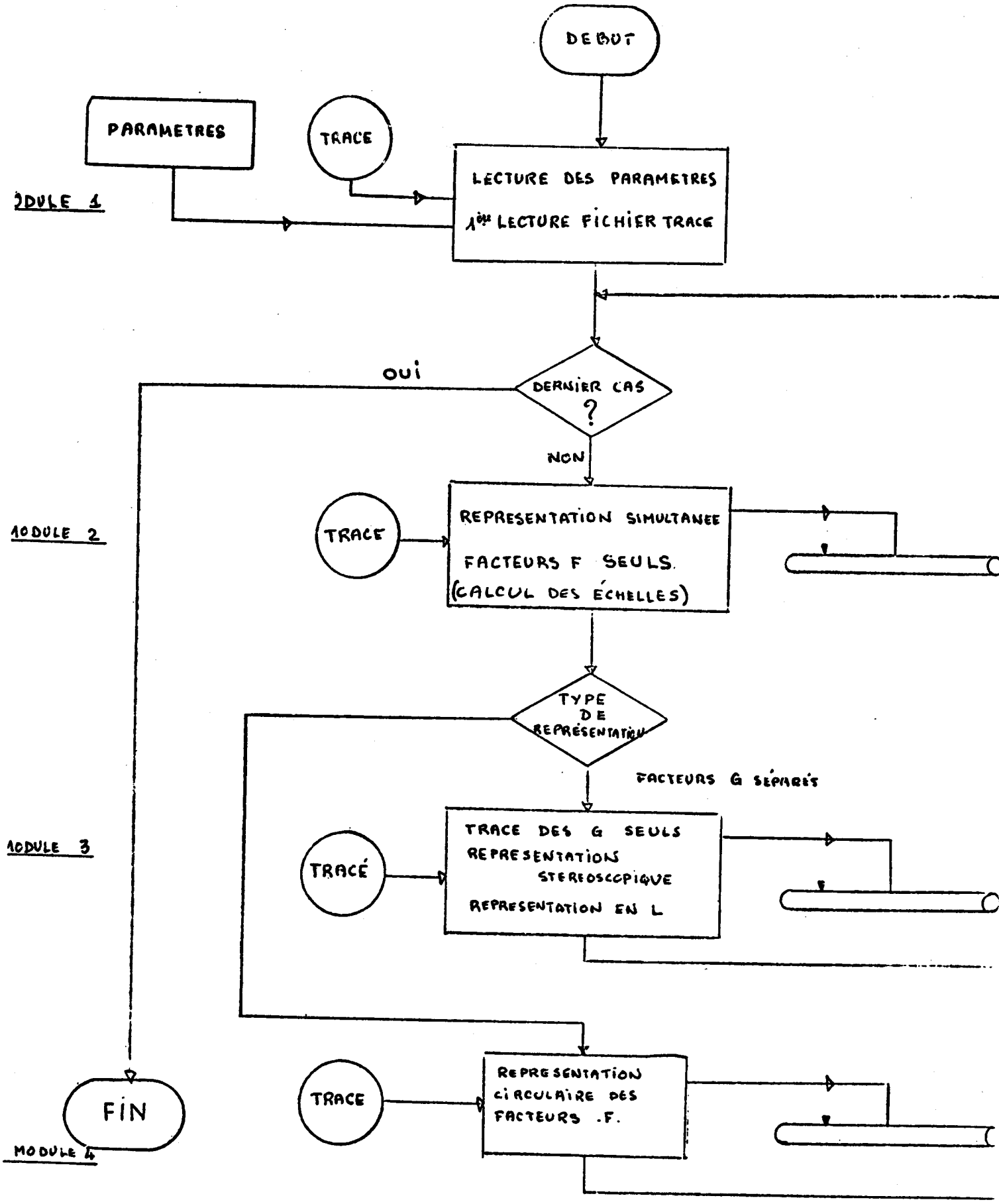


4.2.5. Tracé graphique

4.2.5.1. Dessin du Fichier Tracé



4.2.5.2. Organigramme de principe (Programme de Tracé)



4.2.5.3. Représentation

Si on porte sur doux axes les P vecteurs ayant pour coordonnées (U_{1j}, U_{2j}) , jième composantes des premiers et seconds axes factoriels, on obtiendra dans le plan P vecteurs variables. On porte ces P vecteurs-variables sur le même graphe que les n points-observations ; au vu des directions des vecteurs-variables, on a aussitôt une idée de la signification des facteurs, et l'on sait quelles variables sont responsable de la proximité entre telle ou telle observation.

On a ainsi effectué une représentation dite-simultanée des variables et des observations (voir fig 4.2.5.3.)

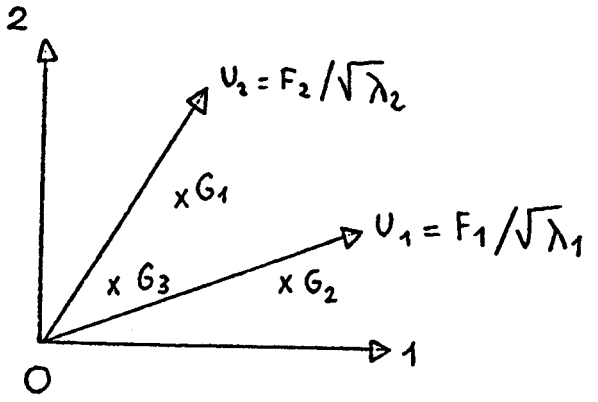


fig 4.2.5.3.

Remarque

Dans la pratique les extrémités des vecteurs (U1, U2) sont très voisines de l'origine : on multiplie leurs coordonnées par une même constante pour obtenir un graphe plus lisible.

Dans le cas ou l'analyse est effectuée sur des variables centrées et réduites par l'écart-type, le coefficient de corrélation entre une variable j et un facteur k est égal à Fkj.

Si on porte par rapport à deux axes (k=1, k'=2) les points Aj de composantes F_{kj} et $F_{k'j}$ ils seront disposés à l'intérieur d'un cercle de rayon unité l'élément j sera d'autant mieux représenté qu'il sera plus proche du bord du cercle. Le cosinus de l'angle de OAj avec l'axe 1 sera le coefficient de corrélation entre la variable j et le premier facteur. Le cosinus de l'angle de OAj avec OAj' est le coefficient de corrélation entre les deux variables j et j'.

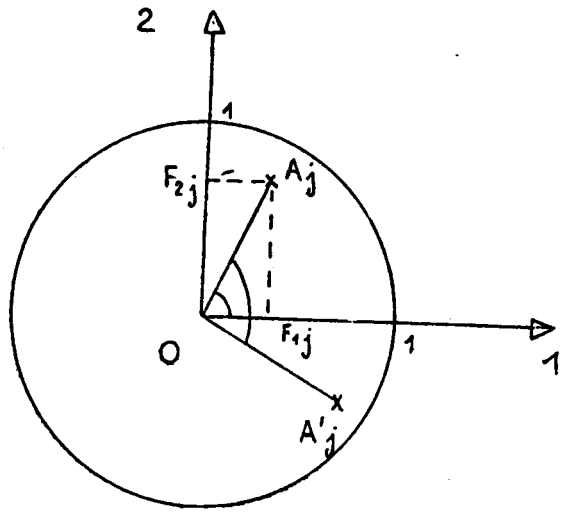


fig. 4.2.5.4.

Remarque

Si on n'a pas effectué l'analyse sur des variables centrées et réduites par l'écart-type, cette représentation circulaire est abandonnée ; Dans le cas le coefficient de corrélation entre une variable k et une variable j n'est plus égal à Fkj.

4.2.6. Description bordereau (modèle 4)

On rencontre dans le bordereau, trois types de paramètres : mise en page, type des cas traités, mode de représentation des facteurs.

- Mise en page

  - Chaque graphe se présente comme un semi de points, ayant pour coordonnées les facteurs, dans un système d'axes orthogonaux à deux dimensions. On indique la longueur en cm de chaque demi axe.

- Les échelles adoptées pour chaque axe sont définies de la façon suivante :

  Echelle en module : on détermine le facteur dont le module est le plus grand affecté d'un coefficient multiplicatif tel que son module devienne égal à 25 cm. Ce coefficient est l'échelle commune aux deux axes.

  Echelle classique : le facteur qui possède la plus grande composante se projette à l'extrémité du demi-axe correspondant, ce qui fournit une échelle pour chaque axe. Si on veut une échelle commune, on prend la plus petite des deux. Un paramètre permet d'opérer ce choix.

- Chacun des deux axes choisis comme repère correspond à une direction propre de la matrice de corrélation (ou covariance) (axe factoriel). Il faudra indiquer les 2 directions propres axe des abcisses et axe des ordonnées.

- <u>types de ces traités</u>

  - On définira, tout d'abord, le nombre de cas à traiter.
  - Il faudra ensuite indiquer le type des cas que l'on désire traiter, ceci pour chaque couple de directions propres :

    Facteurs F/$\sqrt{\lambda}$ seuls

    Facteurs F/$\sqrt{\lambda}$ et G sur un même graphe (pseudo-représentation simultanée).

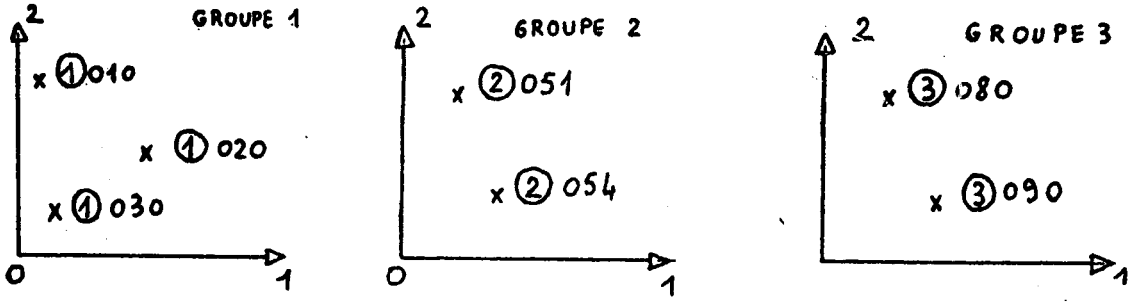
    Facteurs F/$\sqrt{\lambda}$ et G sur deux graphes différents.

    Facteurs G seuls

    Facteurs F situés à l'intérieur d'un cercle.

  - La représentation stéréoscopique engendrera un dédoublement des graphes G seuls. La direction adoptée comme axe de visée et celle adoptée comme axe de rotation du plan défini par les axes des abcisses et des ordonnées devront être indiquées.
  - Si on a effectué l'analyse sur des observations appartenant à plusieurs groupes (traitement tout groupes), on pourra faire autant ou moins de graphes que de groupes, dans le même cas (Facteurs C), en indiquant la valeur du dernier groupe que l'on souhaite représenter.

Si on a indiqué comme valeur du dernier groupe à représenter 3, on a donc 3 graphes



- Mode de représentation des facteurs

La position des facteurs G sera définie soit par une croix (x), soit par un L .

- On pourra inscrire la x seulement, la x et le numéro d'observation, la x et le numéro de groupe, la x et l'identificateur complet.
- Le L suivi de l'identificateur permettra de représenter sur le même graphe, deux directions propres supplémentaires, les branches du L étant proportionnelles aux facteurs sur les deux directions et indiquant leur signe par leur orientation (L, J, Γ, ⅂)
  On indiquera la direction adoptée comme base du L et celle comme hauteur du L .
  Prenons quelques exemples :

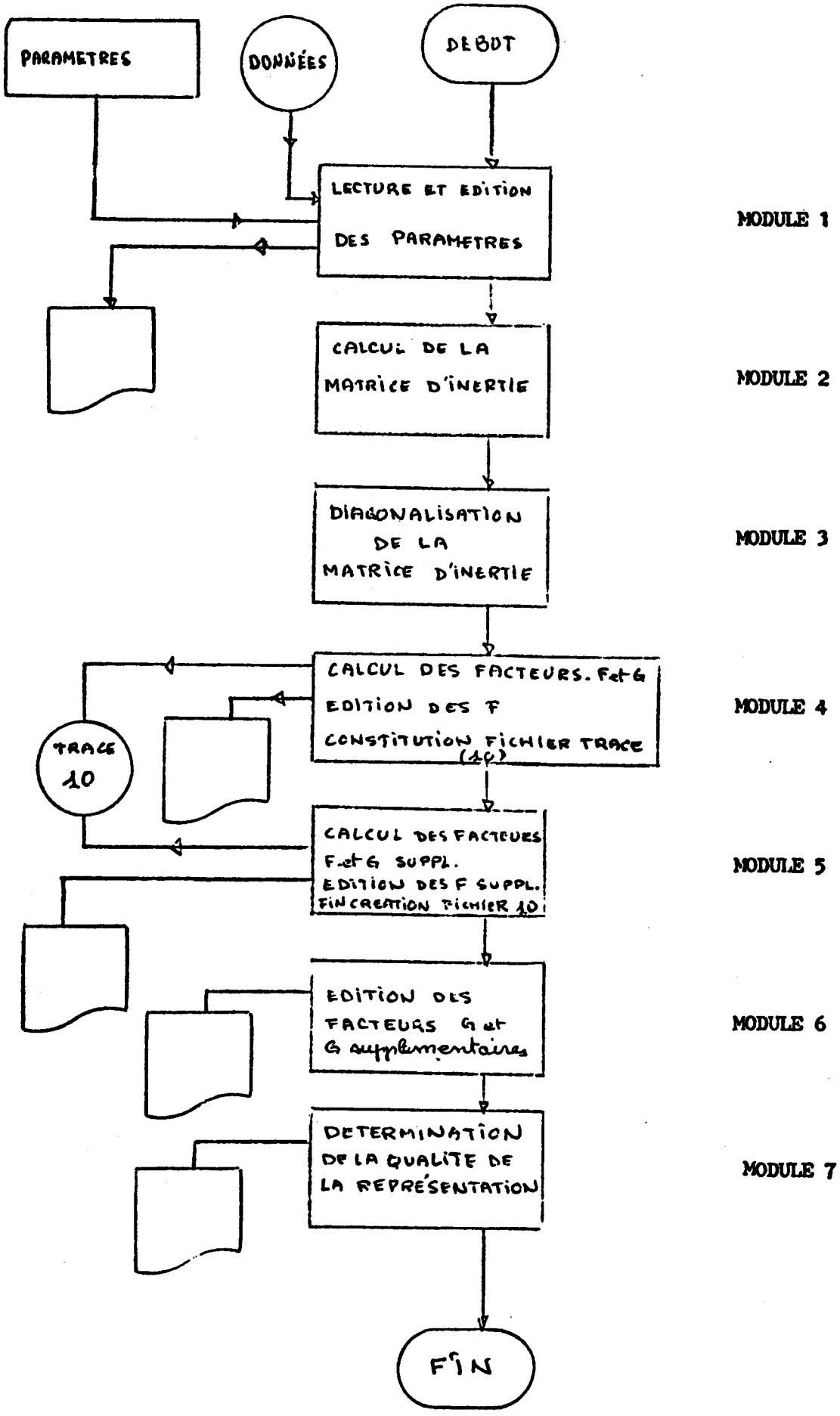
Symbole seul	Numéro d'observation	Groupe seul	Groupe + n° d'observation
X	X 012 L 012	X 4 L 4	X 4012 L 4012

fig. 4.2.6.2.

012 = n° d'observation
4 = numéro de groupe
} ⟶ 012 4 identificateur complet

4.3. Analyse factorielle des correspondances.

4.3.1. Organigramme de Principe (calcul des facteurs)



4.3.2. Méthode utilisée dans le programme

Considérons toujours le tableau de données : $x(i,j)$, $i=1,n$, $j=1,P$

Exécutons les transformations suivantes :

$$Z(i,j) = \frac{x(i,j)}{X} \qquad X = \sum_{i,j} x(i,j)$$

$$Z(i) = \frac{1}{X} \cdot \sum_{j} x(i,j) = \sum_{j} Z(i,j)$$

$$Z(j) = \frac{1}{X} \cdot \sum_{i} x(i,j) = \sum_{i} Z(i,j)$$

On peut se ramener à une analyse simple, en faisant le changement de variables $Y(i,j) = \dfrac{Z(i,j)}{Z(i) \cdot Z(j)}$ et en cherchant les valeurs et vecteurs propres de la matrice $Y' \cdot Y$

Toutes les définitions établies dans l'analyse simple restent valables et les relations deviennent :

Facteurs F : $F(k,j) = u(k,j) \cdot \sqrt{\lambda_k} / Z(j)$

Facteurs G : $G(i,k) = \sum_{j=1}^{p} u(k,j) \cdot Z(i,j) / (Z(i) \cdot \sqrt{Z(j)})$

Facteurs F supplémentaires : $F(k,js) = \sum_{i=1}^{p} Z(i,js) \cdot G(i,k) / (Z(j) \cdot \sqrt{\lambda_k})$

Facteurs G supplémentaires : $G(is,k) = \sum_{j=1}^{p} Z(is,j) \cdot F(k,j) / (Z(i) \cdot \sqrt{\lambda_k})$

Angles des points variables : $\cos^2 \theta\ (j,k) = F^2(k,j) / \sum_{i=1}^{p} \left[\left(\frac{Z^2(i,j)}{Z^2(j)} \cdot \frac{1}{Z(i)} \right) - 1 \right]$

Angles des points observations :

$$\cos^2 \theta\ (i,k) = G^2(i,k) / \sum_{j=1}^{p} \left[\left(\frac{Z^2(i,j)}{Z^2(i)} \cdot \frac{1}{Z(j)} \right) - 1 \right]$$

4.3.3. Description du bordereau (modèle 3)

On rencontre dans le bordereau, quatre types de paramètres : description physique des données, description logique des données, transformation des données, caractéristiques du traitement.

- Description physique des données

Lorsque les données sont lues sur carte, il convient de préciser :

- La position de l'identificateur par rapport aux valeurs (avant ou après) (voir § et fig 1.2.3.)
- Les noms des variables
- Les facteurs multiplicatifs (puissances de 10 par l'opposé desquelles il faut multiplier les valeurs sur carte pour restituer les vraies valeurs). (voir § et fig 1.2.4.)

Lorsque les données sont lues sur disque, noms des variables et facteurs multiplicatifs sont écrits sur le fichier.

- Description logique des données

Il s'agit de définir les dimensions du tableau de données :

- Le nombre de colonnes ou variables.
- Le nombre de lignes ou observations :

  pour les données sur carte, on indiquera le nombre exact d'observations à analyser.

  pour les données sur disque, on indiquera le nombre d'observations à analyser augmenté de 3. Ceci s'explique par la structure du fichier (voir § et fig 2.4.1.) Ce nombre correspond en fait au pointeur de la dernière observation sur laquelle on effectue l'analyse.
- Le nombre de lignes ou d'observations supplémentaires :on indique alors les rangs de la première et de la dernière s'il s'agit de cartes, les valeurs des pointeurs début et fin (rang augmenté de 3) s'il s'agit de données sur disque. (voir § et fig 1.2. 6.)

- Transformation des données

- Changement de variable : il est possible d'exécuter le traitement soit sur les valeurs initiales, soit sur les valeurs divisées par la somme des colonnes, (voir fig 2 tome II), dans le cas où les données sont très hétérogènes.
- Coefficients de pondération : chaque colonne peut être multipliée par un nombre compris entre 0,01 et 99.

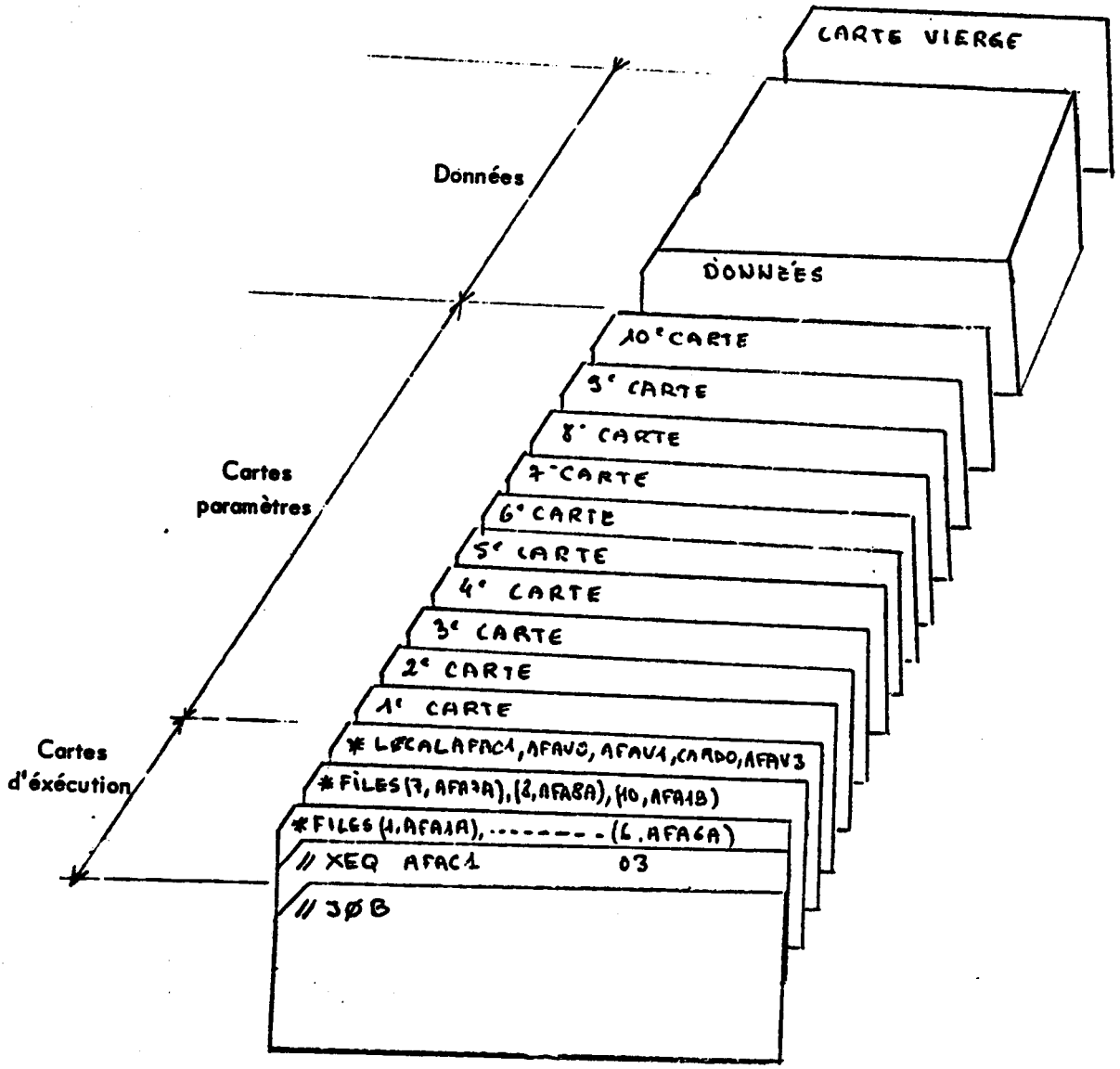
- Logarithmes : on peut transformer toutes les données en logarithmes.
- On a vu que le numéro d'identification contenait un numéro de groupe. Il est possible de n'exécuter l'analyse que sur un groupe bien déterminé.

- Caractéristiques du traitement

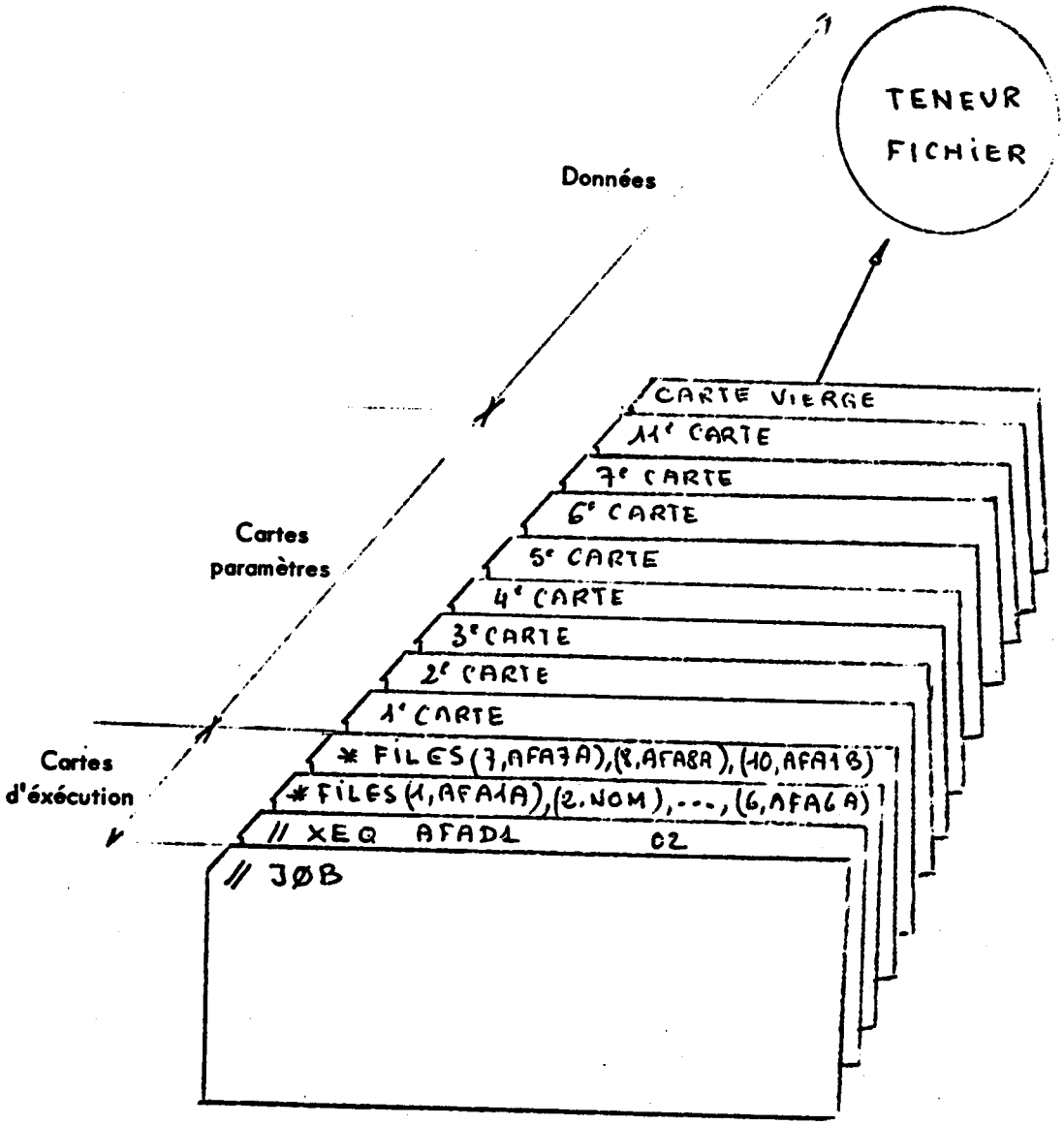
- Le traitemont peut être poursuivi ou interrompu après le calcul de la matrice d'inertie.
- L'impression de cette même matrice peut être réalisée ou non sur option.
- Dans certains cas, on est amené à ignorer ou à représenter seulement certaines variables dans l'espace des axes factoriels déterminés par l'analyse.
- Il faut indiquer le nombre de facteurs que l'on souhaite extraire. Plus ce nombre est élevé, plus le temps de calcul est important.
- Si on veut estimer la qualité de la représentation, on peut demander le calcul des angles des facteurs avec les axes factoriels.

4.3.4. Dessin du paquet de cartes

4.3.4.1. Version carte



4.3.4.2. Version disque

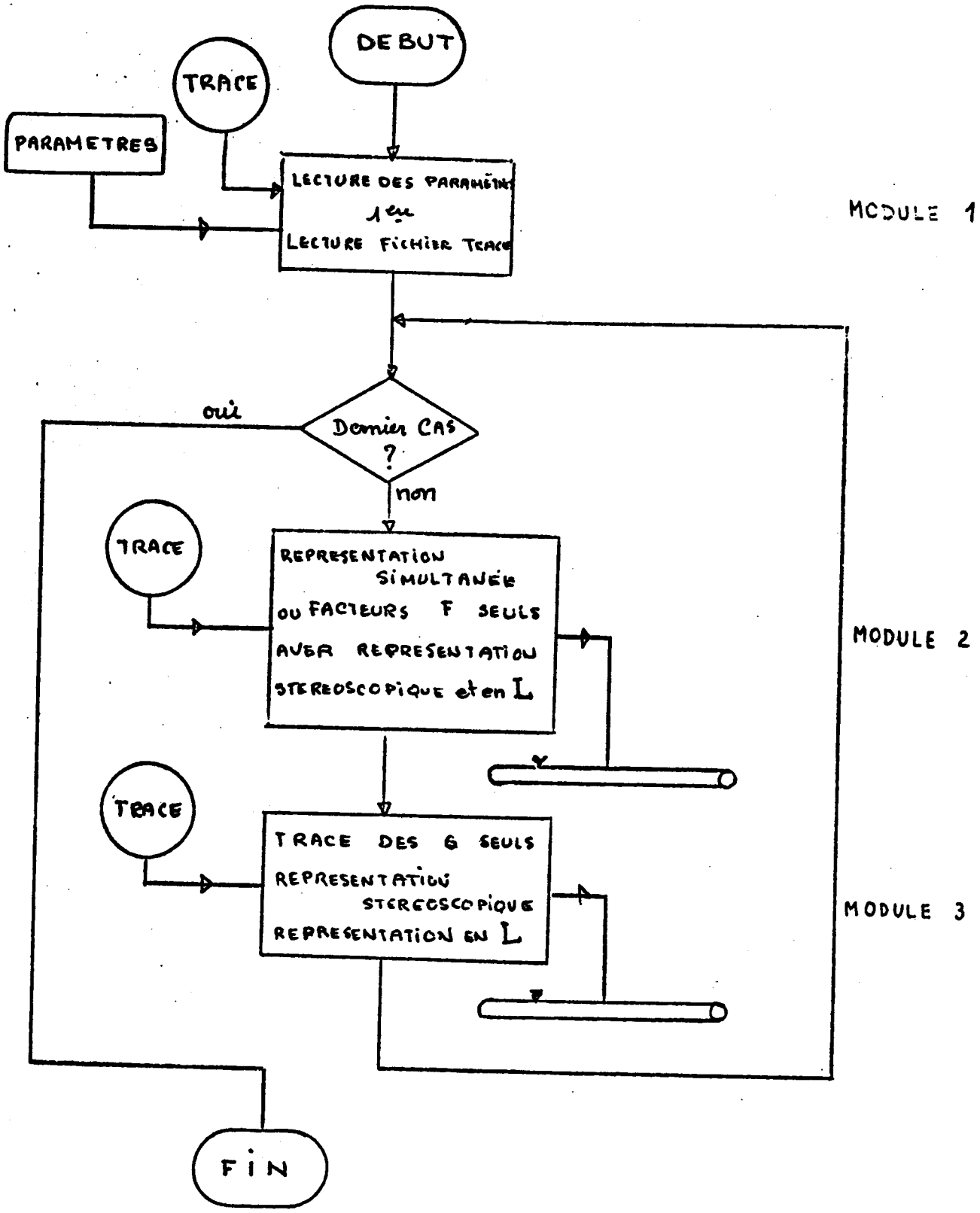


4.3.5. Tracé graphiques

4.3.5.1. Dessin du fichier tracé

1er enregist.	Nombre d'observations Nombre de variables Nombre de facteurs
2è enregistr.	Valeurs propres
3è enregist.	Pourcentages d'explication
4è enregist.	Valeurs maximales des facteurs F
5è enregist.	Echelles des facteurs F
6è enregist.	Facteurs F (première variable)
	Facteurs F (dernière variable)
	Facteurs G (première observation)
	Facteurs G (dernière observation)

4.3.5.2. Organigramme de principe (programme de Tracé)



4.3.5.3. Représentation

On peut représenter les proximités entre les individus dans le plan de deux axes factoriels (par exemple le premier et le second). Les coordonnées des points Ai sont alors les nombres G(i,1) et G(i,2).

On peut représenter sur le même graphique les proximités entre les variables (les coordonnées des points Bj étant les nombres F(1,j) et F(2,j). (voir fig 4.3.5.3.)

Cette représentation simultanée est justifiée par le fait que les Ai apparaissent comme les barycentres des Bj, chaque F(k,j) étant affecté du poids $\frac{Z(i,j)}{Z(i)}$ et réciproquement .

$$G(i,k) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{p} \frac{Z(i,j)}{Z(i)} \cdot F(k,j)$$

$$F(k,j) = \frac{1}{\sqrt{\lambda_k}} \sum_{l=1}^{n} \frac{Z(i,j)}{Z(j)} \cdot G(i,k)$$

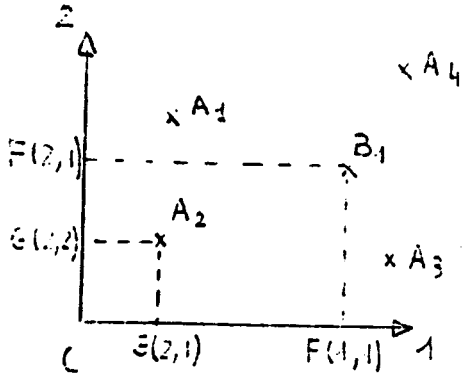


fig. 4.3.5.3.

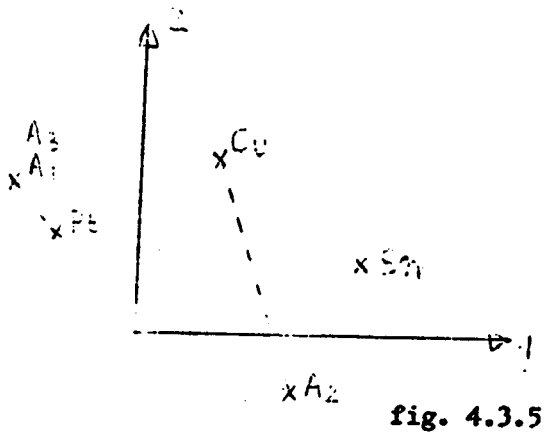


fig. 4.3.5.4.

Prenons un exemple pour montrer l'importance de ce résultat : on dose un certain nombre d'échantillons renfermant CU, Pb, Sn . Sur le graphe la variable Pb est d'autant plus proche d'un échantillon A_1 que celui-ci renferme plus de plomb ; la variable CU est d'autant plus lointaine d'un échantillon A_2 que celui-ci renferme moins de cuivre.

Si les deux échantillons A_1 et A_3 ont des points confondus sur le graphe, cela voudra dire que leurs teneurs en CU, Pb, Sn sont proportionnelles (voir fig 4.3.5.4.)

4.3.6. Description du bordereau (modèle 4)

On rencontre dans le bordereau, trois types de paramètres : mise en page, type des cas traités, mode de représentation des facteurs.

- Mise en page

  - Chaque graphe se présente comme un semi de points, ayant pour coordonnées les facteurs, dans un système d'axes orthogonaux à deux dimensions. On indique la longueur en cm de chaque demi-axe.
  - Les échelles adoptées pour chaque axe sont définies de la façon suivante : le facteur qui possède la plus grande composante se projette à l'extrémité du demi-axe correspondant, ce qui fournit une échelle pour chaque axe. Si on veut une échelle commune on prend la plus petite des deux. Un paramètre permet d'opérer ce choix.
  - Chacun des deux axes choisis comme repère correspond à une direction propre de la matrice d'inertie (axe factoriel). Il faudra indiquer la direction propre axe des abcisses et celle axe des ordonnées.

- Types des cas traités

  - Tout d'abord, on définit le nombre de cas à traiter.
  - ensuite on indique le type des cas que l'on désire traiter, ceci pour chaque couple de directions propres :

    Facteurs F seuls

    Facteurs F et G sur un même graphe (représentation simultanée)

    Facteurs F et G sur deux graphes différents.

    Facteurs G seuls.
  - on observe fréquemment, dans la représentation simultanée, que les facteurs G sont très concentrés autour du centre ; on pourra les représenter à plus grande échelle sur un graphe distinct, dans le même cas (option agrandissement).
  - la représentation stéréoscopique engendrera un dédoublement des graphes F seuls, G seuls. La direction adoptée comme axe de visée, celle adoptée comme axe de rotation du plan défini par les axes des abcisses et des ordonnées, devront être indiquées.

- Si on a effectué l'analyse sur des observations appartenant à plusieurs groupes (traitement tout groupes), on pourra faire autant ou moins de graphes que de groupes, dans le même cas (Facteurs G), en indiquant la valeur du dernier groupe que l'on souhaite représenter. (voir fig 4.2.6.1.)

- Mode de représentation des facteurs.

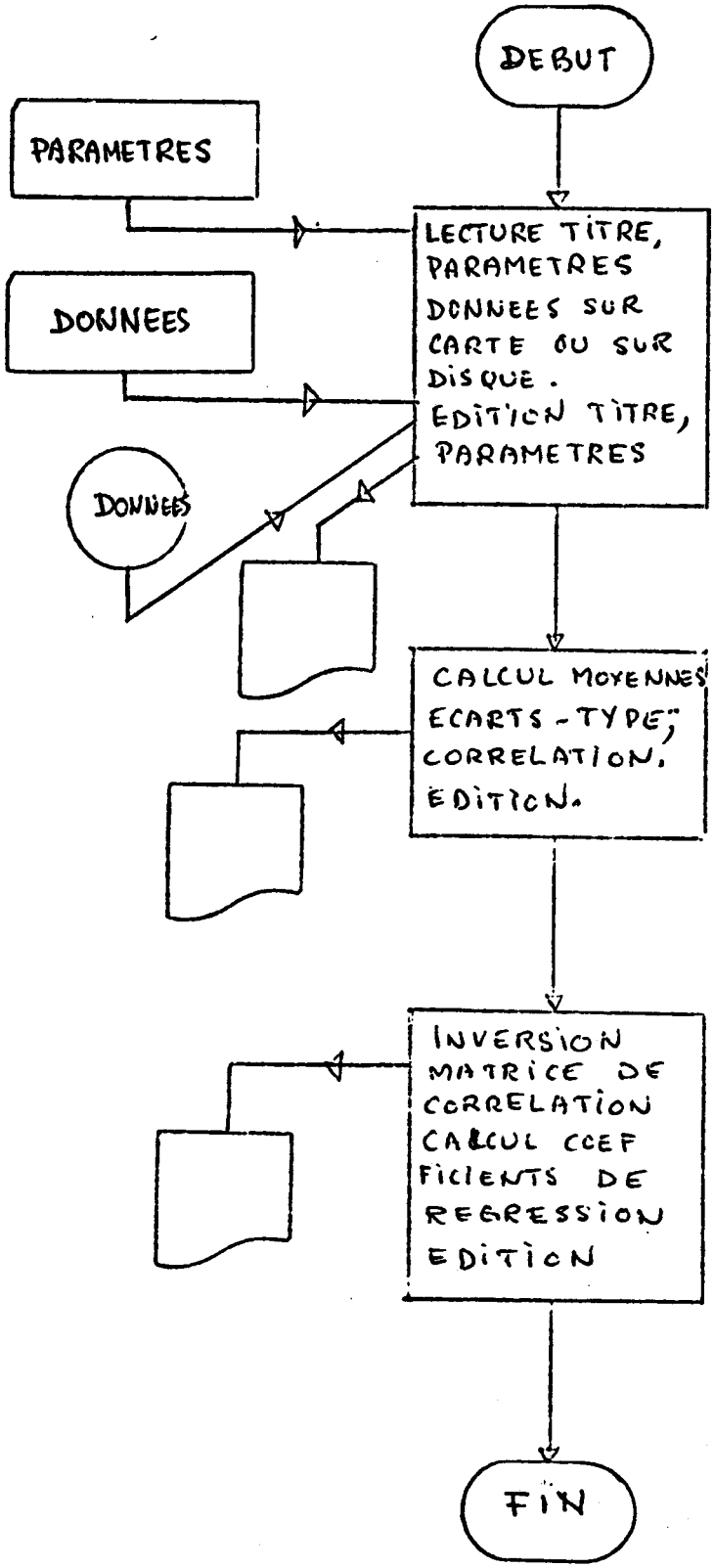
La position des facteurs sera définie soit par une x, soit par un L .

- On pourra inscrire la x seulement, la x et le n° d'observation, la x et le n° de groupe, la x et l'identificateur complet (pour les facteurs G)
- Le L suivi de l'identificateur complet permettra de représenter sur le même graphe, deux directions propres supplémentaires, les branches du L étant proportionnelles aux facteurs sur les deux directions et indiquant leur signe par leur orientation (L, ⌋, ⌜, ⌝). On indiquera la direction adoptée comme base du L, et celle comme hauteur du L. (voir fig 4.2.6.2.)

5. Méthodes de prévision.

5.1. Régression linéaire .

5.1.1. Organigramme de principe.



5.1.2. Coefficients de régression, de corrélation multiple, test de snedecor.

5.1.2.1. Coefficient de régression

Considérons le tableau de n observations à P variables : x(i,j)

Etant donné le modèle linéaire vu au § 5.1.1. de la première partie (notation matricielle) :

$Y = X.A + E$ (Y, A, E sont des estimations des valeurs théoriques)

$$Y = \begin{bmatrix} y(1) \\ \vdots \\ y(n) \end{bmatrix} \quad X = \begin{bmatrix} x(1,1) & \ldots. & x(1,P-1) & U \\ \vdots & & \vdots & \vdots \\ x(n,1) & \ldots. & x(n,P-1) & U \end{bmatrix} \quad A = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{P-1} \\ \alpha_P \end{bmatrix} \quad E = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

Il s'agit de calculer les termes de la matrice A qui sont les coefficients de régression.

La méthode des moindres carrés revient à minimiser la quantité :

$E'.E = (Y-XA)' \cdot (Y-XA)$

ou $E'.E = Y'Y - A'X'Y - Y'XA + AX'XA$

Dérivons cette expression matricielle par rapport à A :

$$\frac{D\,(E'.E)}{DA} = -2X'Y + 2X'XA$$

Annulons cette expression pour trouver la condition qui minimise E'.E :

$-2X'Y + 2X'XA = 0$

$X'XA = X'Y$

multiplions les deux membres de cette égalité, à gauche par $(X'X)^{-1}$, il vient : $A = (X'X)^{-1} X'Y$

Si on partitionne la matrice des covariances V des variables, on obtient :

$$V = \begin{bmatrix} V_{yy} & V'_{xy} \\ V_{xy} & V_{xx} \end{bmatrix}$$

V_{yy} = variance de la variable dépendante

V_{xx} = matrice des covariances des variables indépendantes

V_{xy} = matrice colonne des covariances de la variable dépendante avec chacune des variables indépendantes.

La relation ci-dessus s'écrit encore :

$A = Vxx^{-1} \cdot Vxy$

Le terme constant s'obtient aisément :

$Qp = my - a_1 \cdot mx_1 - a_2 \cdot mx_2 - \ldots\ldots - a_{p-1} \cdot mx_{p-1}$

my, mx_i représentent les moyennes respectives de la variable dépendante et des variables indépendantes.

Complétons cette étude en explicitant un certain nombre de coefficients et de tests.

5.1.2.2. Matrice des covariances du vecteur a

La matrice des covariances théoriques du vecteur à P-1 composantes des coefficients de régression est proportionnelle à l'inverse de la matrice des covariances expérimentales des variables explicatives.

$$V(a) = \frac{S^2}{n} \quad Vxx^{-1}$$

S^2 est appelée la variance résiduelle théorique que l'on peut estimer par la relation :

$$S^2 = \frac{1}{n-P-1} \sum_{i=1}^{n} \gamma_i^2$$

si γ_i est le résidu attaché à l'observation i .

5.1.2.3. Coefficient de corrélation multiple.

Par définition, le carré du coefficient de corrélation multiple est le rapport de la variance expliquée par la régression à la variance totale de la variable expliquée.

$$R^2 = \frac{\text{variance expliquée}}{\text{variance totale}} = \frac{\text{variance totale - variance résiduelle}}{\text{variance totale}}$$

En explicitant numérateur et dénominateur, on obtient :

$$R^2 = \frac{V'xy \cdot a}{Vyy}$$

5.1.2.4. Test de Fischer - Snedecor

La quantité $\frac{1}{S^2} \cdot R^2$ est un χ^2 à P degrés de liberté

La quantité $\frac{1}{S^2}(1-R^2)$ est un χ^2 à n-1-P degrés de liberté ; Ces deux χ^2 sont indépendants.

Le quotient $\frac{n-P-1}{P} \cdot \frac{R^2}{1-R^2}$ que nous désignerons par F est donc une variable aléatoire de Fischer-Snedocor à P et n-P-1 degrés de liberté.

Ce test sur R^2 permet d'apprécier la validité de l'ensemble de la régression.

Si la valeur F dépasse la valeur trouvée dans une table de Snedecor, on estime que F n'est sans doute pas une variable aléatoire de Fischer-Snedecor : Dans ce cas il existe au moins un coefficient de régression significatif.

5.1.3. Description du bordereau (modèle 3)

On rencontre dans le bordereau, quatre types de paramètres : description physique des données, description logique des données, transformation des données, caractéristiques du traitement.

- Description physique des données

Lorsque les données sont sur carte, il convient de préciser :

- la position de l'identificateur par rapport aux valeurs (avant ou après) (voir § et fig 1.2.3.)
- les noms des variables.
- les facteurs multiplicatifs (puissances de 10 par l'opposé desquelles il faut multiplier les valeurs sur carte pour restituer les vraies valeurs) (voir § et fig 1.2.4.)

Lorsque les données sont lues sur disque, noms des variables et facteurs multiplicatifs sont écrits sur le fichier.

- Description logique des données

Il s'agit de définir les dimensions du tableau de données :

- le nombre de colonnes ou variables.
- le nombre de lignes ou observations:

  pour les données sur carte, on indiquera indirectement le nombre d'observations, en plaçant une carte ne contenant que des 9 à la fin du paquet de données.

  pour les données sur disque, le pointeur - prochain enregistrement à écrire - indiquant (à une constante près) le nombre d'observations, figure sur le premier enregistrement du fichier. voir annexe (§ et fig 2.4.1.)

- Transformation des données

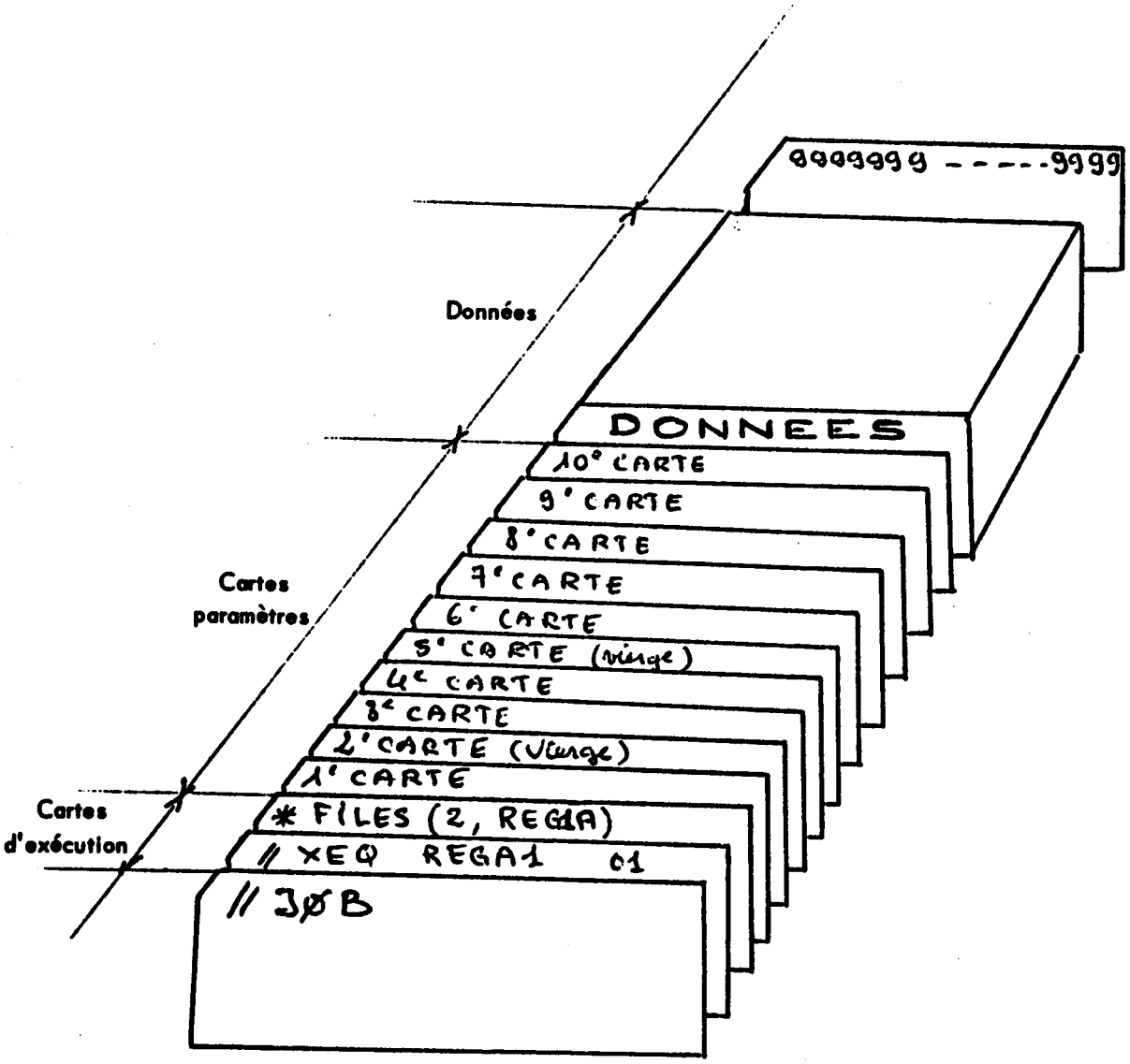
  - Logarithmes : on peut transformer les variables que l'on veut en logarithmes.
  - On a vu que le numéro d'identification contenait un numéro de groupe. Il est possible de faire le traitement sur chaque groupe successivement, dans le même JOB.

- Caractéristiques du traitement

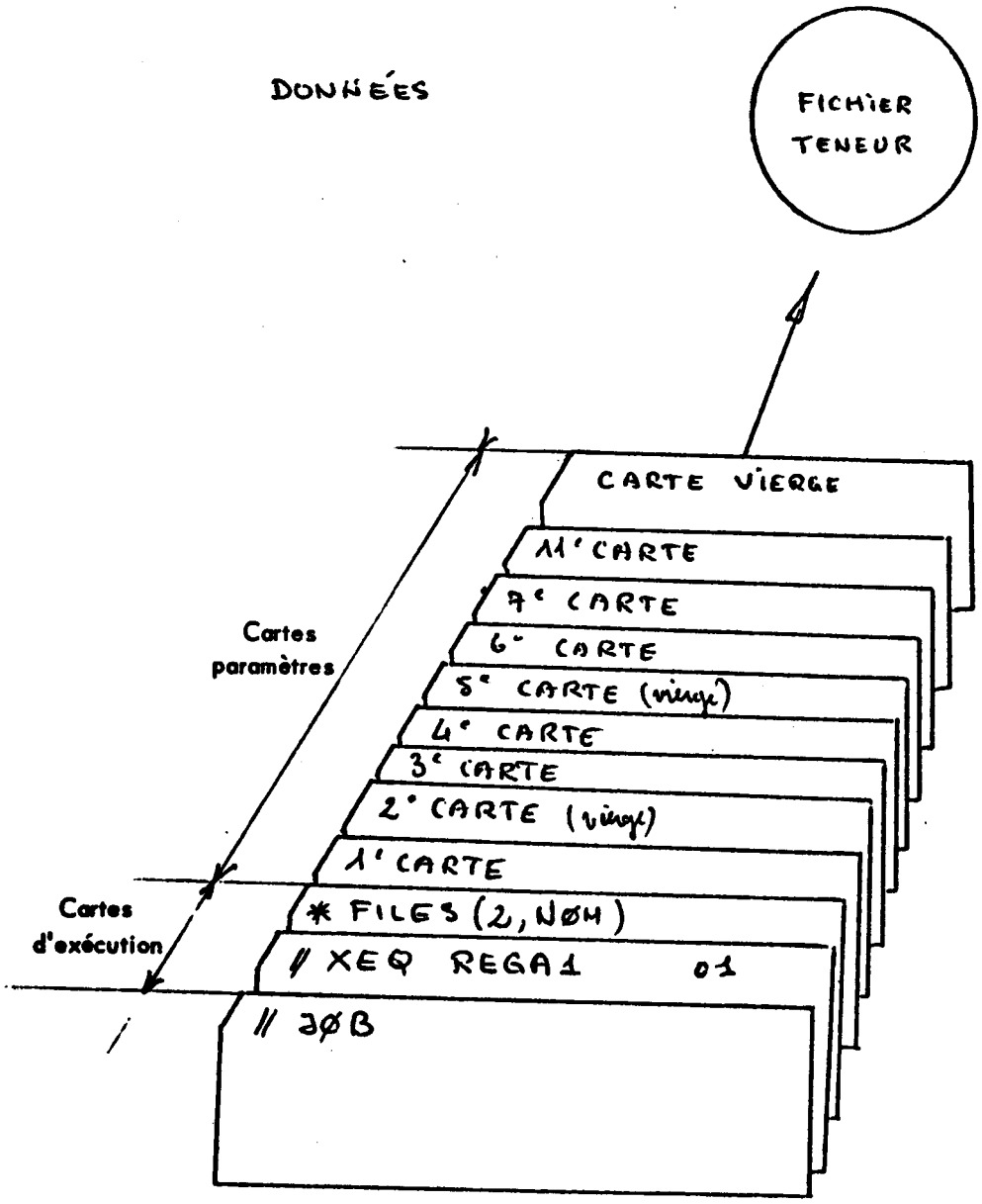
  - Le traitement peut être poursuivi ou interrompu après le calcul de la matrice de corrélation.
  - L'impression de cette même matrice peut être réalisée ou non sur option.
  - Il faut indiquer quelles sont les variables choisies comme variables indépendantes, la variable dépendante et les variables dont on ne veut pas tenir compte dans la régression.

5.1.4. Dessin du paquet de cartes (IBM 1130)

5.1.4.1. Version carte

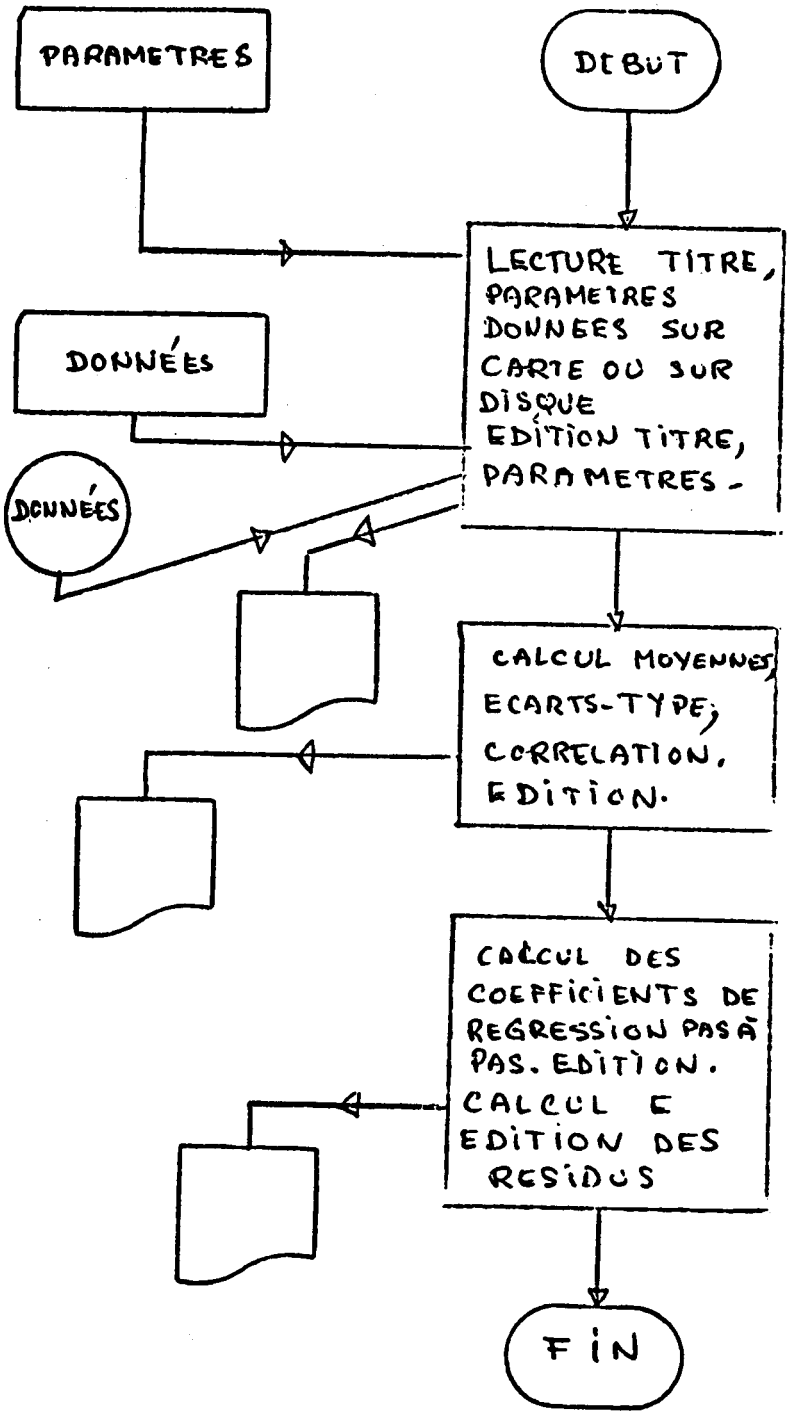


5.1.4.2. Version disque



5.2. Régression étagée

5.2.1. Organigramme de principe



5.2.2. Méthode utilisée

La régression étagée consiste à entrer pas à pas des variables, en commançant par une seule variable ..., suivant un critère que l'on exposera plus bas et à déterminer à chaque étape les coefficients de régression, le coefficient de corrélation multiple, etc. ; Ce qui a été dit pour la régression multiple est valable ici, à chaque étape. Le traitement peut être décomposé en quatre phases :

5.2.2.1. On sélectionne la variable indépendante dont le coefficient de corrélation avec la variable dépendante (y) est le plus fort. On appelera cette variable (x1). On détermine si le coefficient de régression totale b1 est significatif, en le comparant à un seuil (ε1) fixé. S'il ne l'est pas, on dit qu'il n'y a pas de régression significative pour l'ensemble des variables étudiées. S'il l'est, on passe au § 5.2.2.2.

5.2.2.2. On introduit successivement chacune des autres variables indépendantes en plus de (x1).

On conserve parmi ces variables, celle dont le coefficient de corrélation multiple est le plus grand. Il revient au même de dire que la variable conservée donne une somme des carrés résiduelle minimum.

On appelle (x2) cette variable.

On détermine si le coefficient de régression partielle dans la régression de y sur x1 et x2 est significatif, en le comparant à (ε1).

S'il ne l'est pas, on ne conservera que la variable (x1).

S'il l'est, on passe au § 5.2.2.3.

5.2.2.3. On teste la signification de x1, après x2 en utilisant le seuil (ε2) ($\varepsilon 2 \geqslant \varepsilon 1$)

- dans le cas où (x1) n'est pas significatif après x2, on garde la variable (x2) seule et on recommence le traitement à partir du § 5.2.2.2 .

- dans le cas où (x1) est significatif après x2, on garde les variables (x1) et (x2) et on passe au §5.2.2.4 .

5.2.2.4 On introduit successivement, chacune des autres variables indépendantes en plus de (x1) et (x2) et on retient la variable qui fournit la plus forte corrélation multiple. On appelle (x3) cette variable. Si (x3) n'est pas significative après (x1) et (x2) en utilisant le seuil (ε1), on garde la formule avec (x1) et (x2). Si (x3) est significative, on teste celle des variables (x1) et (x2) qui est la moins significative, en utilisant le seuil (ε2).

- Si les deux variables (x1) et (x2) ont des influences significatives, on poursuit le traitement en essayant d'introduire une quatrième variable.

- Si l'une des deux par exemple (x2) ne l'est pas, on la supprime et on recommence le traitement à partir du § 5.2.2.4.

5.2.2.5. On arrête le traitement quand il ne reste plus de variable ni à entrer, ni à supprimer. On édite en même temps les résidus.

Remarques :

- Les seuils $\mathcal{C}_1$, $\mathcal{C}_2$ sont évidemment des nombres qui déterminent les variables à entrer et à supprimer. Leurs valeurs sont données par une table de Snedecor et sont fonctions des degrés de liberté. L'utilisateur indique leurs valeurs par carte paramètre (voir bordereau).

- Le calcul des coefficients de régression et des autres grandeurs associées est plus simple (du point de vue programmation) que dans le cas de la régression multiple. Il n'y a pas d'inversion globale de matrice ; celle-ci s'opère progressivement d'abord quand on entre la 1ère variable, puis la seconde ... et ainsi de suite.

5.2.3. Description du bordereau (modèle 3)

On rencontre dans le bordereau, quatre types de paramètres : description physique des données, description logique des données, transformation des données, caractéristiques du traitement.

- Description physique des données

Lorsque les données sont sur carte, il convient de préciser :

- La position de l'identificateur par rapport aux valeurs (avant ou après). (voir § et fig 1.2.3.)
- Les noms des variables.
- Les facteurs multiplicatifs (puissances de 10 par l'opposé desquelles il faut multiplier les valeurs sur carte pour restituer les vraies valeurs). (voir § et fig 1.2.4.)

Lorsque les données sont lues sur disque, noms des variables et facteurs multiplicatifs sont écrits sur le fichier.

- Description logique des données

Il s'agit de définir les dimensions du tableau des données :

- Le nombre de colonnes ou variables.
- Le nombre de lignes ou observations.

pour des données sur carte, on indiquera indirectement le nombre d'observations, en plaçant une carte ne contenant que des 9 à la fin du paquet de données.

pour des données sur disque, le pointeur - prochain enregistrement a écrire - indiquant (à une constante près) le nombre d'observations, figure sur le premier enregistrement du fichier (voir § et fig 2.4.1.)

- Transformation des données

- Logarithmes : on peut transformer les variables que l'on veut en logarithmes.
- On a vu que le numéro d'identification contenait un numéro de groupe. Il est possible de faire le traitement sur chaque groupe successivement, dans le même JOB.

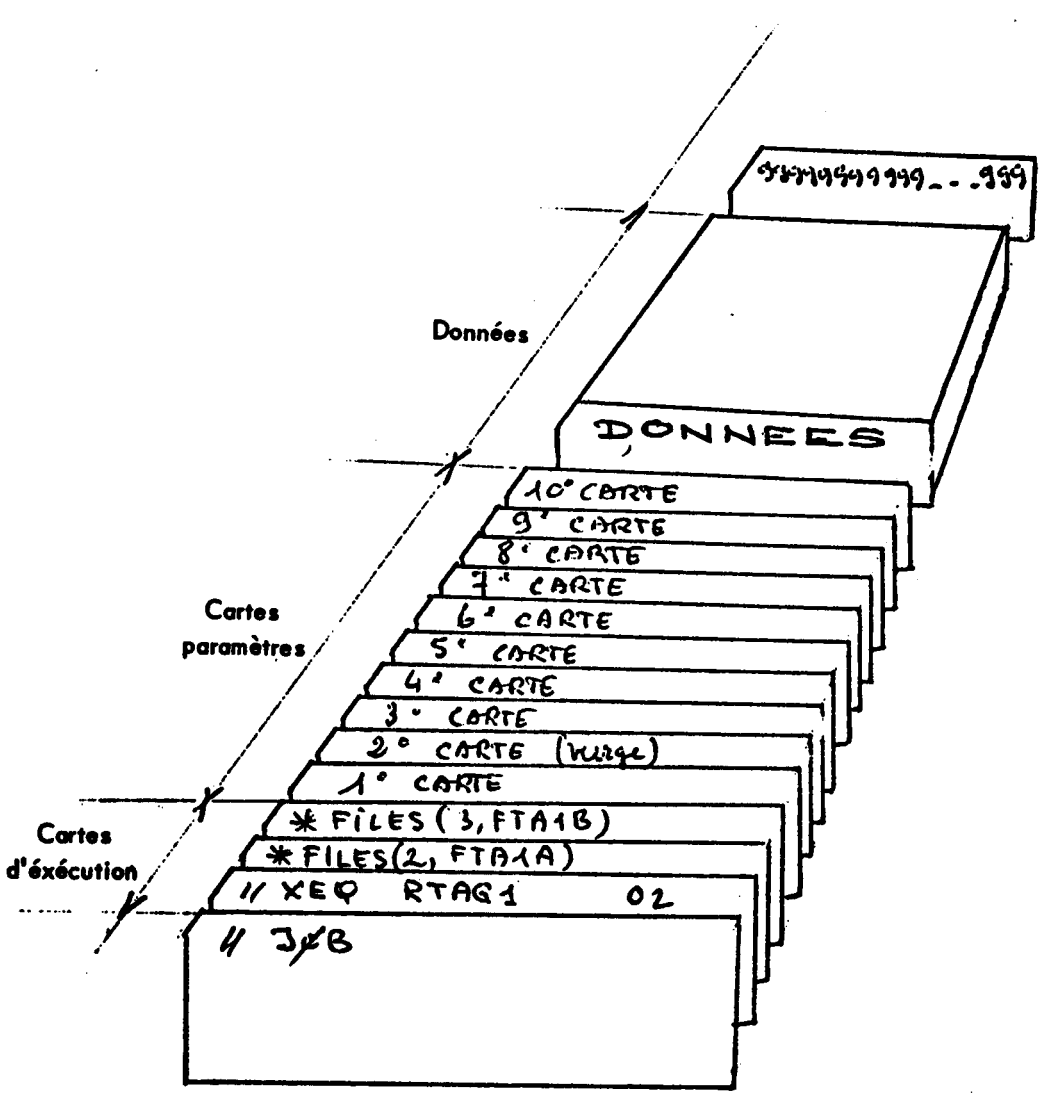
- Caractéristiques du traitement

- Le traitement peut être poursuivi ou interrompu après le calcul de la matrice de corrélation.
- L'impression de cette même matrice peut être réalisée ou non sur option.
- Il faut indiquer quelles sont les variables choisies comme variables indépendantes, la variable dépendante et les variables dont on ne veut pas tenir compte dans la régression.
- Il y aura lieu de préciser les seuils au delà et en deçà desquels on entre et on sort les variables. Ces seuils sont donnés dans une table de Snedecor.

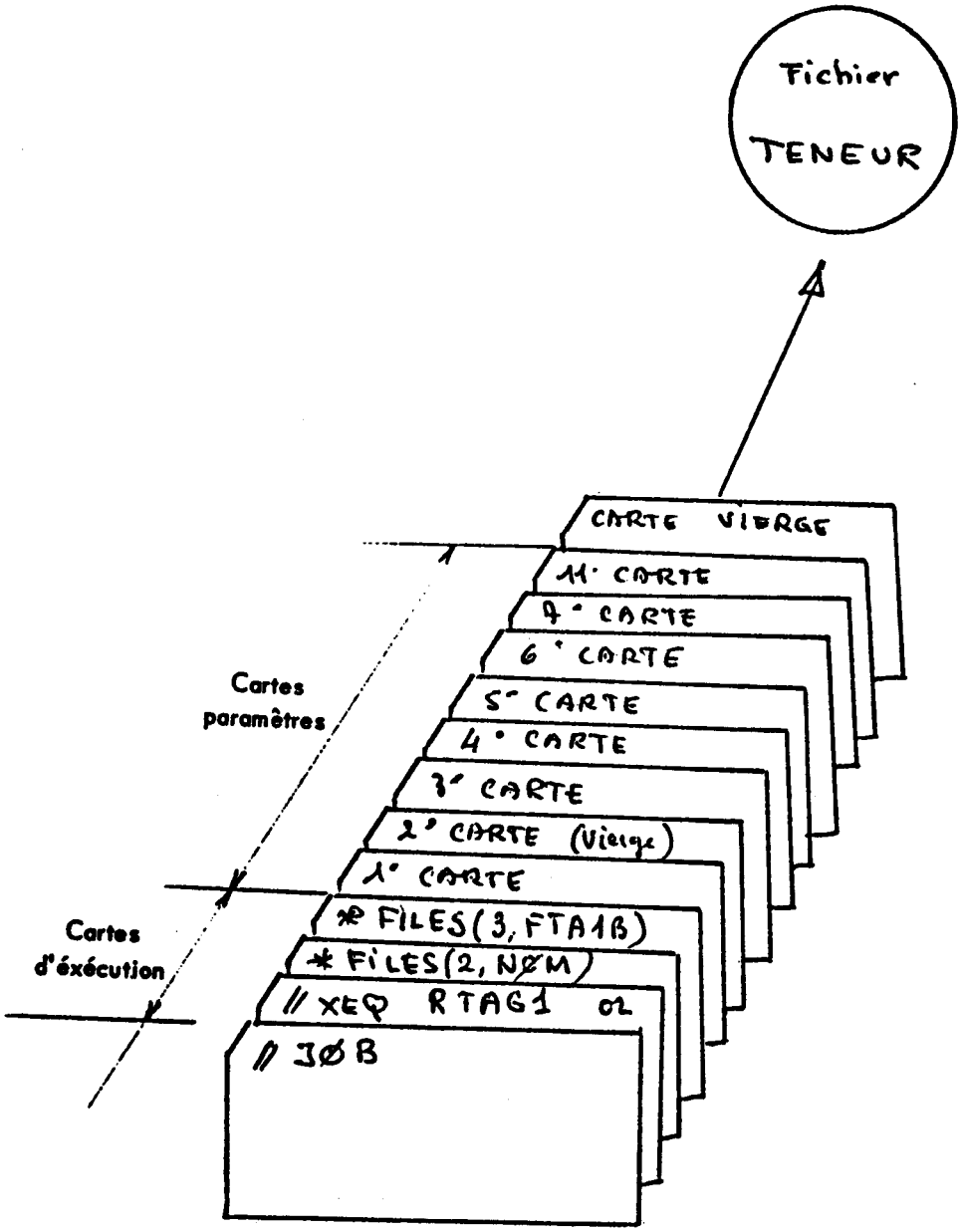
A titre indicatif, on peut prendre 0,05 et 0,025 comme seuils d'entrée et de sortie, valeurs qui correspondent à des degrés de liberté infinis (il est impératif que le seuil d'entrée soit supérieur au seuil de sortie).

5.2.4. Dessin du paquet de cartes (IBM 1130)

5.2.4.1. Version carte

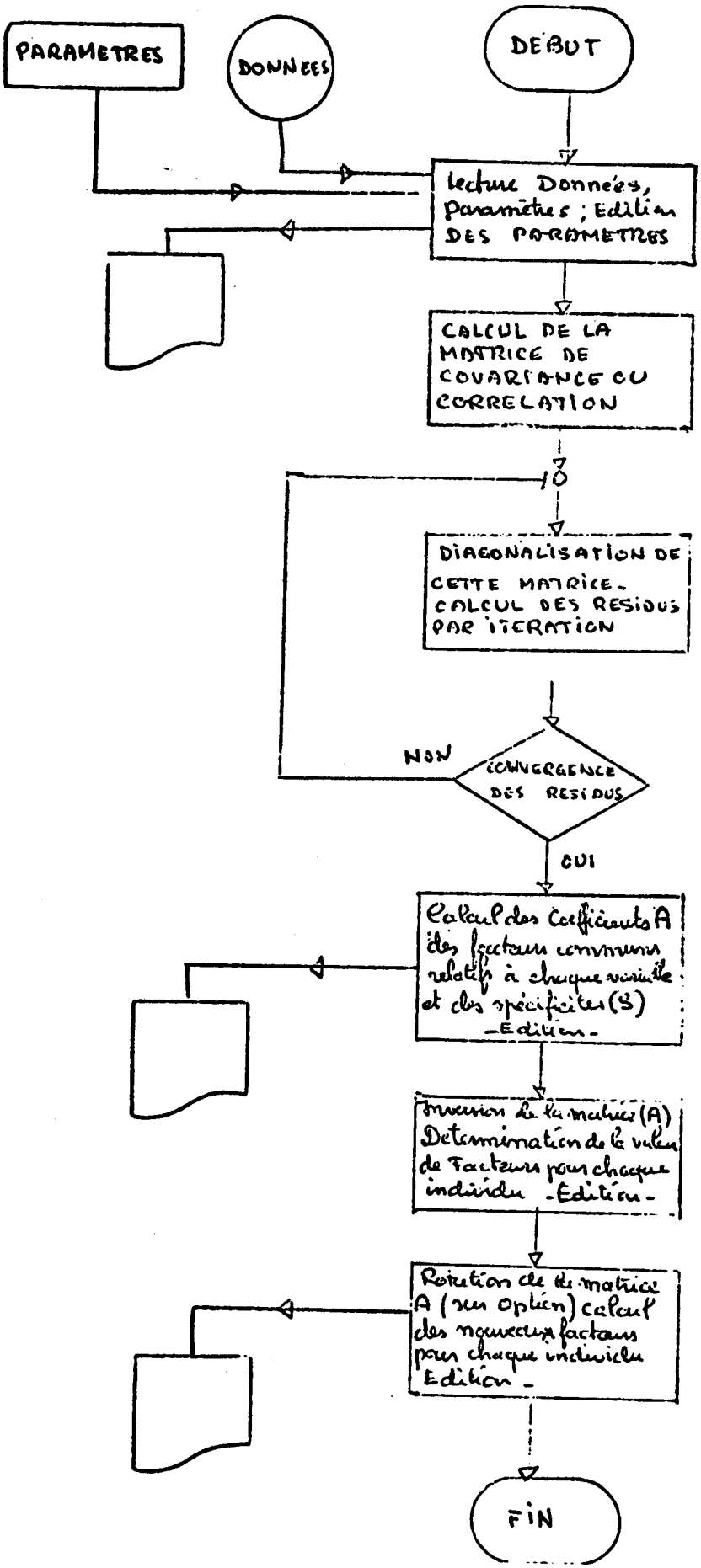


5.2.4.2. Version disque



5.3. Analyse factorielle en facteurs communs et spécifiques.

5.3.1. Organigramme de principe.



5.3.2. Méthode utilisée

On a vu dans les généralités qu'il convenait de poser un modèle à priori (analyse en mode R) qui s'écrit sous forme matricielle :

$$(1) \qquad \underset{(P,1)}{X} = \underset{(P,q)}{\Lambda} \cdot \underset{(q,1)}{F} + \underset{(P,1)}{A}$$

X = matrice de données

Λ = coefficients constants

F = facteurs

A = matrice des résidus

Dans ce modèle, seule X est connue. Pour le résoudre, il convient de poser des conditions supplémentaires :

- Les résidus ont une moyenne nulle et une variance constante
- Les facteurs (F) ne sont pas corrélés et sont de variances égales à 1

Si on introduit la matrice des covariances de (x), on se ramène au modèle équivalent :

$$(2) \qquad V = \Lambda\Lambda' + \Delta$$

avec $\Delta = \frac{1}{n} A'A$ matrice diagonale dont les éléments diagonaux sont les variances des résidus.

En analyse en composantes principales ,

Si V est la matrice de covariance extraite des données,

B est la matrice des vecteurs propres (orthogonaux deux à deux),

D la matrice diagonale des valeurs propres correspondantes,

On partait de la forme quadratique :

$$(3) \qquad V = B\,D\,B^{-1} \quad \text{ou } V = B.D\,B'$$

car $B' = B^{-1}$ (orthogonalité des vecteurs propres)

La relation (3) peut encore d'écrire

$$(4) \quad V = (B.D^{\frac{1}{2}}) \cdot (D^{\frac{1}{2}}.B') = \Gamma.\Gamma'$$

Les valeurs propres de V étant toutes positives

On revient maintenant à l'analyse en facteurs communs et spécifiques ; rapprochant les deux relations (2) et (4) :

En retranchant une matrice diagonale à éléments positifs (Δ) de la matrice des covariances, on obtiendra une décomposition analogue à celle donnée par la relation (4)

$$\underset{(q,q)}{V - \Delta} = \underset{(q,p)}{\Lambda} \cdot \underset{(p,q)}{\Lambda'}$$

Pour déterminer la matrice $V - \Delta$, on procède par itérations :

- on pose au départ $\Delta = 0$
- on calcule les vecteurs propres de V, rangés en colonne dans la matrice B : $V_0 = BDB' = \Gamma\Gamma'$
- on estime Δ par les éléments diagonaux de $V_0 - \Gamma\Gamma'$; on cherche les vecteurs propres de $V_1 = V_0 - \Delta^*$ et on estime de nouveau
- Si le processus converge raisonnablement, on a bien obtenu une décomposition du type $V = \Lambda\Lambda' + \Delta$

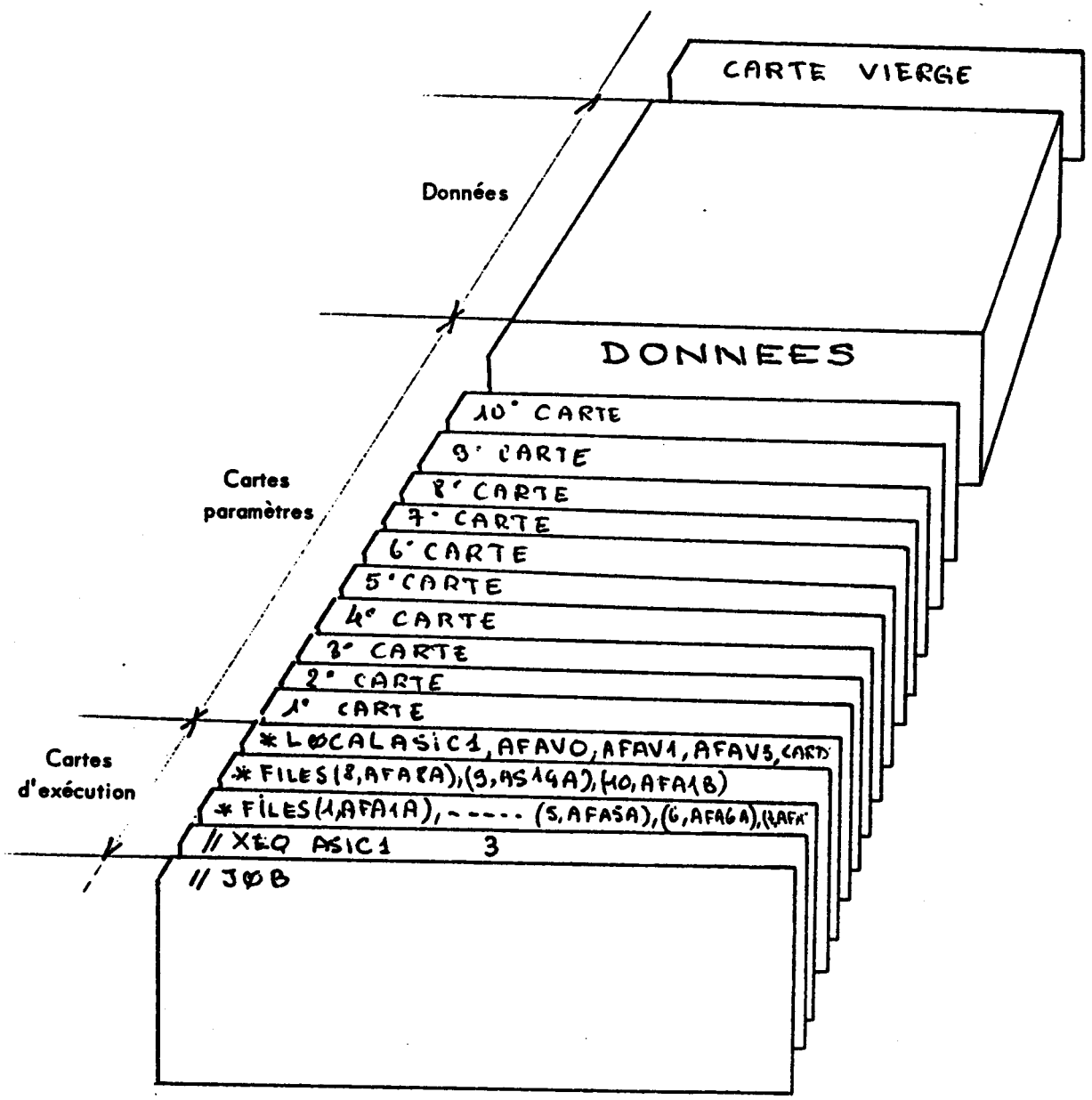
A partir de ce résultat, on poursuit comme pour une analyse en composantes principales.

5.3.3. Description du bordereau (modèle 3)

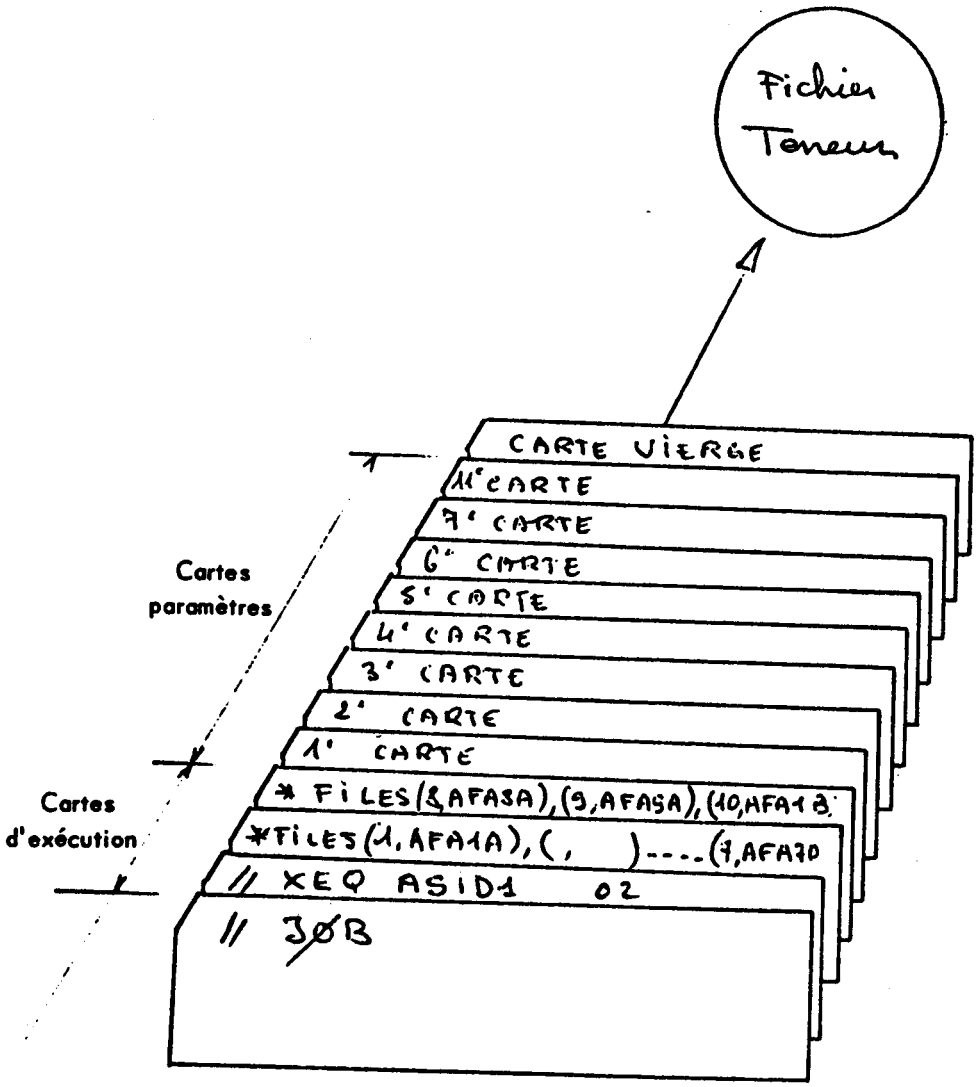
Le bordereau déja utilisé pour les programmes de régression et d'analyse de données peut être utilisé à nouveau pour l'analyse en facteurs communs et spécifiques. Nous n'en ferons donc pas une nouvelle présentation. En effet il se remplit rigoureusement de la même façon que pour le programme d'analyse en composantes principales.

5.3.4. Dessin du paquet de cartes (IBM 1130)

5.3.4.1. Version carte.



5.3.5.1. Version disque.



TROISIEME PARTIE

EXEMPLES D'APPLICATION

1. Choix des exemples.

Les quelques exemples faisant l'objet de cette partie sont donnés à titre d'illustration et portent essentiellement sur les tableaux et graphiques fournis par l'ordinateur.

Les conclusions d'ordre géologique sont volontairement réduites et pourront être obtenues de façon détaillée auprès des Départements responsables des études correspondantes :

- Calcaire d'AVETA (Minéralurgie).
- Dolérites de l'ANTARCTIQUE (Géologie).
- Nappes aquifères du NORD de la FRANCE (Hydrogéologie).
- Prospection géochimique de BELLE-ISLE-EN-TERRE (Géophysique / Géochimie).
- Sédiments marins de la BAIE DE LA VILAINE (Laboratoires).
- Comparaison de courbes granulométriques (Géologie).

Un rapport ultérieur traitera d'un exemple complet permettant de comparer sur les mêmes données les différentes méthodes de traitement. Cet exemple portera vraisemblablement sur l'étude de la distribution des éléments dans des échantillons en roche provenant du granite de LA MARCHE.

2. Commentaires.

Préliminaire : transformation des données.

Les données analysées ne sont généralement pas les mesures de base, mais les mesures transformées par centrage ou par centrage et réduction.

Les figures 1 et 2 illustrent à l'aide d'un exemple numérique les transformations qui peuvent être exécutées avant exécution de l'analyse.

Dans la partie gauche des figures, les tableaux reprennent les données initiales (dosages bruts) ; dans la partie droite, ils représentent les données transformées et réellement analysées.

Figure 1

Analyse en composantes principales.

Trois transformations possibles :

- variables centrées : $yij = xij - mj$;
- variables centrées et réduites par la moyenne : $yij = \frac{xij - mj}{mj}$
- variables centrées et réduites par l'écart-type : $yij = \frac{xij - mj}{sj}$

Figure 2

Analyse des correspondances.

Deux transformations possibles :

- variables non transformées : $yij = xij$
- variables réduites par la somme des colonnes : $yij = \frac{xij}{Nx\ mj}$

2.1. Calcaire d'AVETA

2.1.1. Nature des échantillons.

Ce sont des carrottes de sondages effectués dans un gisement subhorizontal de calcaire pour cimenterie au TOGO. Les dosages suivants ont été effectués : CaO, pertes au feu, P205, Si02, Al203, Fe203, Na20, K20, Cl2.

Le problème posé était d'étudier la distribution verticale et horizontale des différents échantillons dans le gisement.

2.1.2. Exploitations effectuées et sorties graphiques.

Figure 3

Plan factoriel déterminé par les axes factoriels 1 et 2.

Représentation des échantillons seuls. Visualisation sur un même graphique des directions factorielles 3 et 4 à l'aide du symbole L (2è partie §)

Deux groupements se manifestent :

- celui de droite rassemble les échantillons riches en Si02, Al203 et Fe203.

- celui de gauche correspond aux échantillons riches en CaO et pertes au feu. Ce dernier se subdivise lui-même en 3 sous-groupements moins nettement séparés.

La représentation en L montre par exemple que les échantillons 2302 et 2402 ont un facteur 3 de même sens mais un facteur 4 de sens opposé.

Figure 4

Plan factoriel (1-2). Représentation circulaire et détermination des corrélations entre éléments. La proximité du cercle et des points représentatifs des éléments Si, Al, Fe montre que ceux ci sont bien représentés dans le plan (1-2). Il en est de même des points Ca et PF.

La proximité des points Si, Fe, Al montre que les éléments correspondants sont corrélés positivement ; ils sont par contre corrélés négativement à Ca et PF.

Analyse des correspondances

Figure 5

Plan factoriel déterminé par les axes factoriels 1 et 2 . Représentation des échantillons seuls. Visualisation sur un même graphique des directions factorielles 3 et 4 à l'aide du Symbole L . La distinction entre les différents groupements apparait plus nettement qu'en composantes principales.

Figure 6

Plan factoriel (1-2) . Représentation simultanée des échantillons et des éléments.

Elle permet de caractériser chacun des groupements par les associations d'éléments qui se projettent à lour proximité.

Le groupement de droite est caractérisé par Si02, Al203 et Fe203 ; celui de gauche est surtout caractérisé par CaO et perte au feu. P205 introduit une discrimination suivant l'axe 2.

Analyse en facteurs communs et spécifiques.

Figure 7

On rappelle le modèle linéaire adopté par cette analyse pour mieux comprendre les résultats obtenus.

$$Z_{ji} = Q_{j1}F_{1i} + Q_{j2}F_{2i} + Q_{j3}F_{3i} + Q_{j4}F_{4i} + Q_{j5}F_{5i} + Q_{j}U_{ji}$$

Le premier tableau indique les coefficients des facteurs (Loading-factors en termes anglo-saxon) pour chacune des variables (j) analysées (Q_{j1}, Q_{j2}, Q_{j3}, Q_{j4}, Q_{j5}), ainsi que les spécificités Q_j.

Le second tableau indique la valeur de chacun des mêmes coefficients après exécution d'une rotation orthogonale (Q'_{j1}.... Q'_{j5}), (méthode varimax).

Le dernier tableau indique la valeur de chacun des facteurs pour chacune des observations i (factor-scores, en terme anglo-saxon) (F_{1i}, F_{2i}, F_{3i}, F_{4i}, F_{5i}).

Régression étagée

Figure 8

La moitié supérieure de la figure est relative à la première étape de la régression étagée (première variable entrée)

Tout d'abord une série de coefficients indique les limites de la validité de la régression avec notamment le coefficient de corrélation multiple. On trouve ensuite sur une même ligne le coefficient de régression l'écart-type de

la régression, le coefficient bêta et le coefficient T de Student.

Au dessous le terme constant de la régression et enfin un tableau d'analyse de variance.

La moitié inférieure de la figure est relative à la deuxième étape de la régression (2ème variable entrée). On retrouve tous les coefficients énumérés plus haut.

Figure 9

Analyse des correspondances.

Etat de sortie donnant les coordonnées des éléments et des échantillons sur les axes factoriels.

Figure 10

Analyses des correspondances.

Etat de sortie permmettant de voir la qualité de la représentation des échantillons par leur projection sur chaque axe ou plan factoriel.

Figure 11

Analyse des correspondances. Plan factoriel (1-2). Représentation stéréoscopique à l'aide des coordonnées sur le 3è axe factoriel prises comme cotes. L'examen des deux graphiques à l'aide d'un stéréoscope permet de visualiser les points dans l'espace à 3 dimensions des 3 premiers axes factoriels.

2.2 Dolérite de l'ANTARCTIQUE

2.2.1. Nature des échantillons

Ce sont des échantillons dont les dosages proviennent de la bibliographie (majeurs et traces).

2.2.1. Exploitations effectuées et sorties graphiques.

Analyse en composantes principales à l'aide des éléments moyens seuls.

Figure 12

Plan factoriel (1-2). Représentation des échantillons seuls. Le graphique donne le classement des échantillons en fonction des 10 majeurs pris en compte (SiO_2, Al_2O_3, Fe_2O_3, CaO, MgO, Na_2O, K_2O, TiO_2, MnO, H_2O^+, H_2O^-).

Analyse en composantes principales à l'aide des éléments en trace seuls.

Figure 13

Plan factoriel (1-2). Représentation des échantillons seuls. Le graphique donne le classement en fonction des traces suivantes : Cr, Ni, Cu, Zn, Sr, Ba, Zr, Rb.

Ce classement est analogue au précédent, assurant ainsi que la distribution des traces est fortement liée à celle des majeurs.

Diagramme triangulaire.

Figure 14

Il montre la répartition des roches entre les trois pôles Fe, Mg, Na^+, K et met en évidence une différenciation avec enrichissement en fer (type tholéïtique très net).

2.3. Nappes aquifères du NORD de la FRANCE.

2.3.1. Nature des échantillons

Ce sont des prélèvements d'eau provenant de deux nappes superposées (région de NORD-MELANTOIS) et analysés pour les ions Ca^{++}, Mg^{++}, Na^+, K^+, Cl^-, $CO3^{--}$, $SO4^{--}$, $NO3^-$.

l'étude avait pour buts :

- de déterminer si les deux nappes avaient un chimisme différent.
- de permettre d'affecter des prélèvements d'origine mal connue à l'une ou l'autre des nappes.

2.3.2. Exploitations effectuées et sorties graphiques.

Analyse des correspondances.

Figure 15

Plan factoriel (1-2). Représentation des échantillons seuls.

Le graphique montre que :

- les prélèvements des deux nappes forment deux groupements séparés correspondant à deux chimismes différents ;
- les prélèvements de la nappe 1 sont très groupés alors que ceux de la nappe 2 sont fortement dispersés.

- la nappe 1 est caractérisée par un faciès sulfate (+ carbone) calcique tandis que la nappe 2 évolue d'un pôle carbonaté magnésien vers un pôle carbonaté sodique ; Ces deux pôles correspondent à des profondeurs différentes de la nappe.

2.4. Prospection géochimique de BELLE-ISLE-EN TERRE.

2.4.1. Nature des échantillons

Ce sont 500 prélèvements en sol sur berges provenant d'une prospection stratégique de la région de POULLAOUEN - BELLE ISLE EN TERRE (BRETAGNE). Les dosages suivants ont été effectués : Ba, Sr, Ga, Pb, Zn, Cu, Ag, Ni, Co, Cr, V, Sc, Mn, Y, Yb, B, Be, Sn, Bi .

2.4.2. Exploitations effectuées et sorties graphiques

Histogramme

Figure 16 Histogramme de répartition du baryum. l'amplitude de variation retenue est de 0 à 720 g/t. (intervalle de classe : 36g/t). Cette répartition, donnée à titre d'illustration, correspond à une distribution sensiblement normale ; seuls 5 échantillons ont une teneur anormale.

Figure 17

Graphique représentant les teneurs en vanadium en fonction des teneurs en scandium. ("+" signale un seul point ; "*" signale deux ou plusieurs points confondus). Le coefficient de corrélation entre les deux éléments, calculé par ailleurs, est de 0.81.

Analyse des correspondances

Figure 18

Plan factoriel (1-2). Représentation des éléments seuls. L'axe 1 est caractérisé par le manganèse qui s'oppose principalement à Sr, Ba, Ga, Sn, Be, B. L'axe 2 est déterminé par Pb.

Figure 19

Plan factoriel (1-3). Représentation des éléments seuls. L'axe 3 est caractérisé par Cr, Sc, V, Cu.

L'examen des deux graphiques permet de supposer des relations de corrélation entre les associations d'éléments suivantes : V - Cr - Sc

Ni - Co

Sn - Be

Zn - Y - Yb - Bi

2.5. Sédiments marins de la BAIE DE LA VILAINE

2.5.1. Nature des échantillons.

Ce sont des sédiments marins de faible profondeur provenant de la BAIE DE LA VILAINE (120 échantillons). Ils ont été classés à priori en 4 groupes en fonction de leurs caractéristiques granulométriques :

groupe 1 : graviers

" 2 : sables

" 3 : vases sableuses

" 4 : vases

Le problème était de déterminer si ces quatre groupes se caractérisaient par des distributions géochimiques différentes.

2.5.2. Exploitations effectuées et sorties graphiques.

Analyse des correspondances.

Figure 20

Plan factoriel (1-2). Représentation simultanée des échantillons et des éléments.

Les échantillons forment trois groupements :

- à gauche : quelques sables, très riches en Sr.
- au centre : la majorité des graviers et des sables
- à droite : les vases et les vases sableuses.

La représentation simultanée montre que :

- le groupement des sables riches en Sr est essentiellement caractérisé par des éléments Sr et Mn.
- le groupement des graviers et sables est essentiellement caractérisé par les éléments Mn, Ba, Pb.
- le groupement des vases et vases sableuses est essentiellement caractérisé par les associations Cu, Co, Ni et V, Cr.

On note de plus l'existence de correlations possibles entre Pb - Ba,

V - Cr, Cu - Ni .

Figure 21

Plan factoriel (1-3). Représentation simultanée des échantillons et des éléments.

Deux groupements seulement apparaissent car les projections sur ce plan des points sables à Sr et des points graviers et sables sont partiellement confondues.

Les proximités des points représentatifs des éléments confirment les corrélations ci-dessus.

Figure 22

On a reporté une carte, les échantillons (à droite su symbole +) à l'endroit du prélèvement, ainsi que leurs teneurs en chrome et en vanadium (à gauche du symbole +).

On observe ainsi que les variables chrome et vanadium sont fortement corrélées : si un échantillon renferme beaucoup de chrome, il renferme beaucoup de vanadium.

2.6. Comparaison de courbes granulométriques.

2.6.1. Nature des échantillons.

Ce sont des sédiments marins de profondeur comprise entre 50 et 250 m provenant des plateau NORD ESPAGNOL (DE SANTANDER au CAB PENAS).

Les données analysées correspondent aux 8 fréquences principales des courbes granulométriques intéressant la fraction Arénites et correspondant aux dimensions suivantes:

2. - 1.25 mm ; 1.25 - 0.800 mm ; 0.800 - 0.500 mm ; 0.500 - 0.315 mm ; 0.315 - 0.200 mm ; 0.200 - 0.125 mm ; 0.125 - 0.080 mm ; - 0.080 mm .

2.6.2. Exploitations effectuées et sorties graphiques.

Figure 23

Plan factoriel (1-3). Représentation des échantillons seuls. L'ensemble des 223 échantillons se répartit en plusieurs classes correspondant à des histogrammes de forme différente. Les histogrammes types de chaque classe ont été figurés manuellement et représentés sur le graphique sorti par l'ordinateur.

La classification obtenue a été confirmée par la méthode de classification des "nuées dynamiques" qui est décrite en annexe.

3. Remarques générales

Comme le montrent les exemples précédents, la base de toute étude de données numériques en Géologie est constituée par les méthodes d'analyse de données. (composantes principales et correspondances).

Pour préparer leur interprétation et contrôler les résultats obtenus, il est indispensable de passer par l'étape des méthodes statistiques élémentaires fournissant moyennes, variances, histogrammes de distribution et coefficients de corrélation.

Lorsque les résultats d'analyse de données ont été dépouillés, en particulier lorsque l'on a mis en évidence des associations d'éléments caractéristiques et interprétables, en termes géologiques, il peut être intéressant d'utiliser les méthodes de prévision.

La régression permettra d'exprimer un ou plusieurs éléments en fonction d'autres considérés comme indépendants. L'analyse factorielle en facteurs communs et spécifiques, en se fixant le nombre de facteurs recherchés égal au nombre des associations d'éléments trouvées à l'étape précédente, permettra de préciser l'interprétation géologique des facteurs extraits.

Remarque

Il faut bien remarquer que le nombre de facteurs déterminés par l'analyse en facteurs communs et spécifique doit être fixé a priori, puisqu'il représente la dimension du sous espace dans lequel on désire se placer. La meilleur façon de choisir ce nombre est probablement de déterminer le nombre d'associations interprétables fournies par une méthode purement descriptive d'analyse de données.

CONCLUSION

Les méthodes de traitement multivariables doivent nécessairement se développer car l'ordinateur permet d'affranchir l'utilisateur des contraintes de calcul liées au nombre des échantillons analysés et des dosages effectués.

Pourquoi en effet vouloir par exemple classer des individus humains uniquement par leur taille et leur poids en négligeant entre autres renseignements la couleur de leurs yeux ou de leurs cheveux, la forme de leur visage ou leur tour de taille ?

Un échantillon géologique n'est vraiment bien analysé que lorsque tous les éléments du tableau de MENDELEIEFF ont été déterminés. Le fait que des considérations d'ordre économique interdisent une détermination aussi poussée ne doit pas être assimilé à une limitation du nombre des dosages pour une cause d'impossibilité pratique à les dépouiller.

N'est-il pas plus rentable, à coût global comparable, de diminuer le nombre des échantillons prélevés et d'augmenter le nombre des déterminations effectuées ?

Comme le montrent les exemples décrits précédemment, les Départements installés à ORLEANS utilisent de plus en plus les méthodes et programmes de traitement multivariables. En moyenne, l'ordinateur IBM 1130 du Centre de Calcul est employé à raison de 15-20 heures par mois d'unité centrale, essentiellement pour des applications en Géochimie et en Hydrogéologie.

Les utilisateurs affectés à des services décentralisés ou basés à l'étranger disposent dans chacun des départements d'ORLEANS d'un ou plusieurs correspondants susceptibles de faciliter la prise en charge, le traitement et la résolution des problèmes qu'ils souhaiteraient éventuellement étudier.

A titre méthodologique, une étude détaillée du comportement des programmes en face d'un cas concret est actuellement en cours au Département Informatique. Elle a pour but de définir à l'aide d'un exemple une procédure d'approche générale aidant à la résolution des problèmes de traitement statistique des données numériques en Géologie. Ses conclusions pratiques, complétées par un mode d'emploi des programmes, feront l'objet d'un prochain rapport.

ANNEXE

CLASSIFICATION NON HIERARCHIQUE

PAR LA METHODE DES " NUEES DYNAMIQUES "

Présentation sommaire

Cette nouvelle méthode de classification non hiérarchique est due à Monsieur E. DIDAY, Ingénieur à l'IRIA. Elle est extrèmement ingénieuse et peut rendre de grands services dans de nombreuses applications géologiques. Nous remercions vivement Monsieur DIDAY d'avoir bien voulu nous communiquer cette méthode et nous permettre de la programmer pour ordinateur IBM 1130.

Algorithme de principe.

La méthode permet de générer une partition de l'ensemble des n échantillons du tableau x(i,j) en fonction des teneurs des p éléments dosés et en un nombre k fixé de classes.

Son algorithme schématique est le suivant :

1 - On se fixe le nombre k de classes désirées.

2 - On choisit pour chaque classe l (ou en l'absence de critère de choix on tire au hasard) un nombre nl d'individus qui sont considérés comme étalons pour la lème classe.

3 - On calcule les distances de tous les échantillons aux centres de granite des étalons de classes.

4 - En fonction de ces distances, on affecte chaque échantillon i à la classe (définie par ses étalons) dont il est le plus proche.

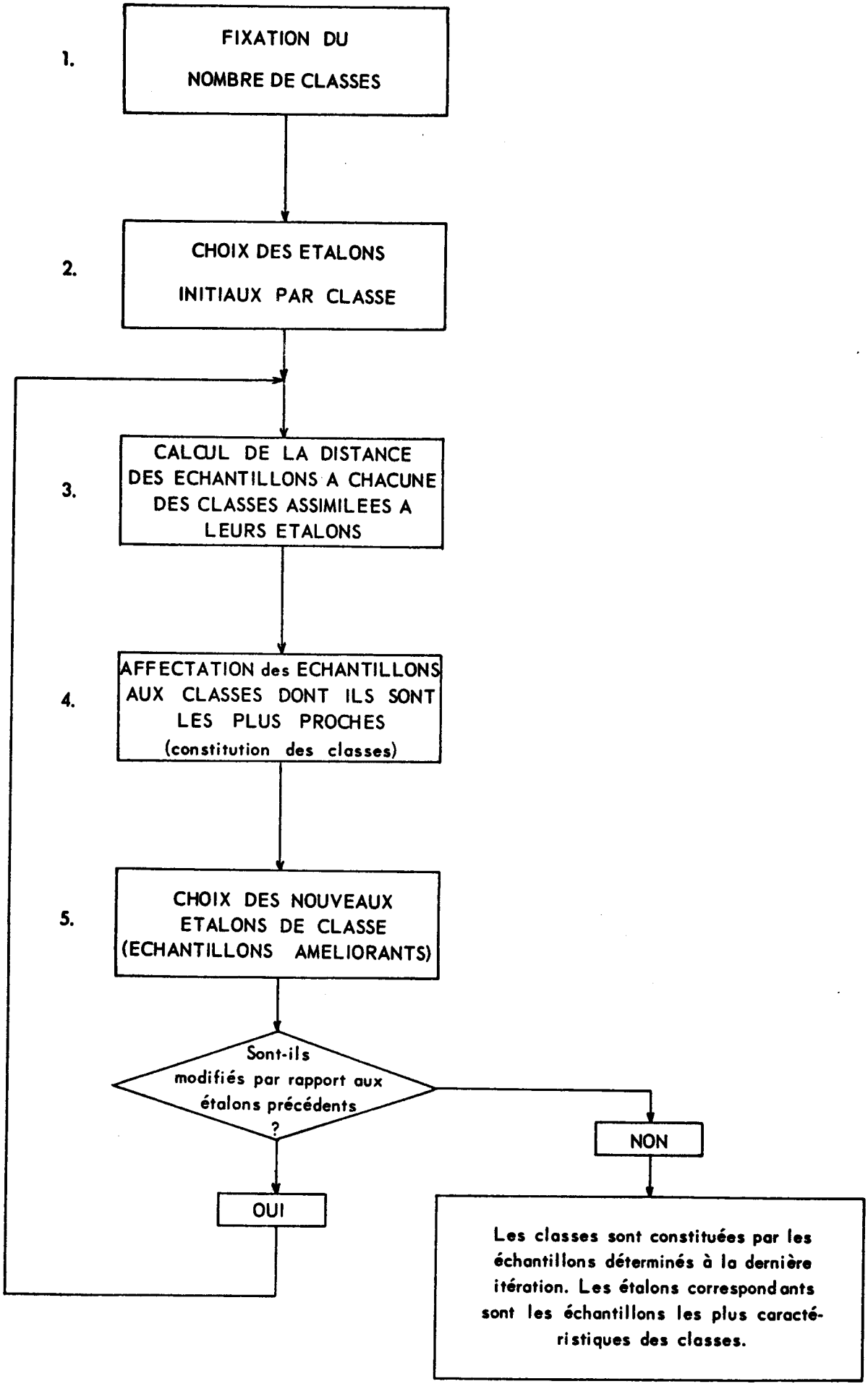
5 - On prend comme nouveaux groupes étalons les échantillons les plus proches des centres de gravité des classes ; les nl nouveaux étalons de la classe l sont les nl échantillons dont les distances à la classe l sont les plus faibles.

6 - On recommence à partir du point 3 jusqu'à ce que les étalons ne varient plus entre deux itérations successives.

7 - A ce niveau, les derniers étalons trouvés sont les étalons terminaux et les n échantillons sont affectés aux classes suivant la procédure du point 4.

La partition recherchée est ainsi déterminée.

L'ensemble de la procédure est résumée par le schéma suivant :


1.
FIXATION DU
NOMBRE DE CLASSES
2.
CHOIX DES ETALONS
INITIAUX PAR CLASSE
3.
CALCUL DE LA DISTANCE
DES ECHANTILLONS A CHACUNE
DES CLASSES ASSIMILEES A
LEURS ETALONS
4.
AFFECTATION des ECHANTILLONS
AUX CLASSES DONT ILS SONT
LES PLUS PROCHES
(constitution des classes)
5.
CHOIX DES NOUVEAUX
ETALONS DE CLASSE
(ECHANTILLONS AMELIORANTS)
Sont-ils
modifiés par rapport aux
étalons précédents
?
OUI
NON
Les classes sont constituées par les échantillons déterminés à la dernière itération. Les étalons correspondants sont les échantillons les plus caractéristiques des classes.

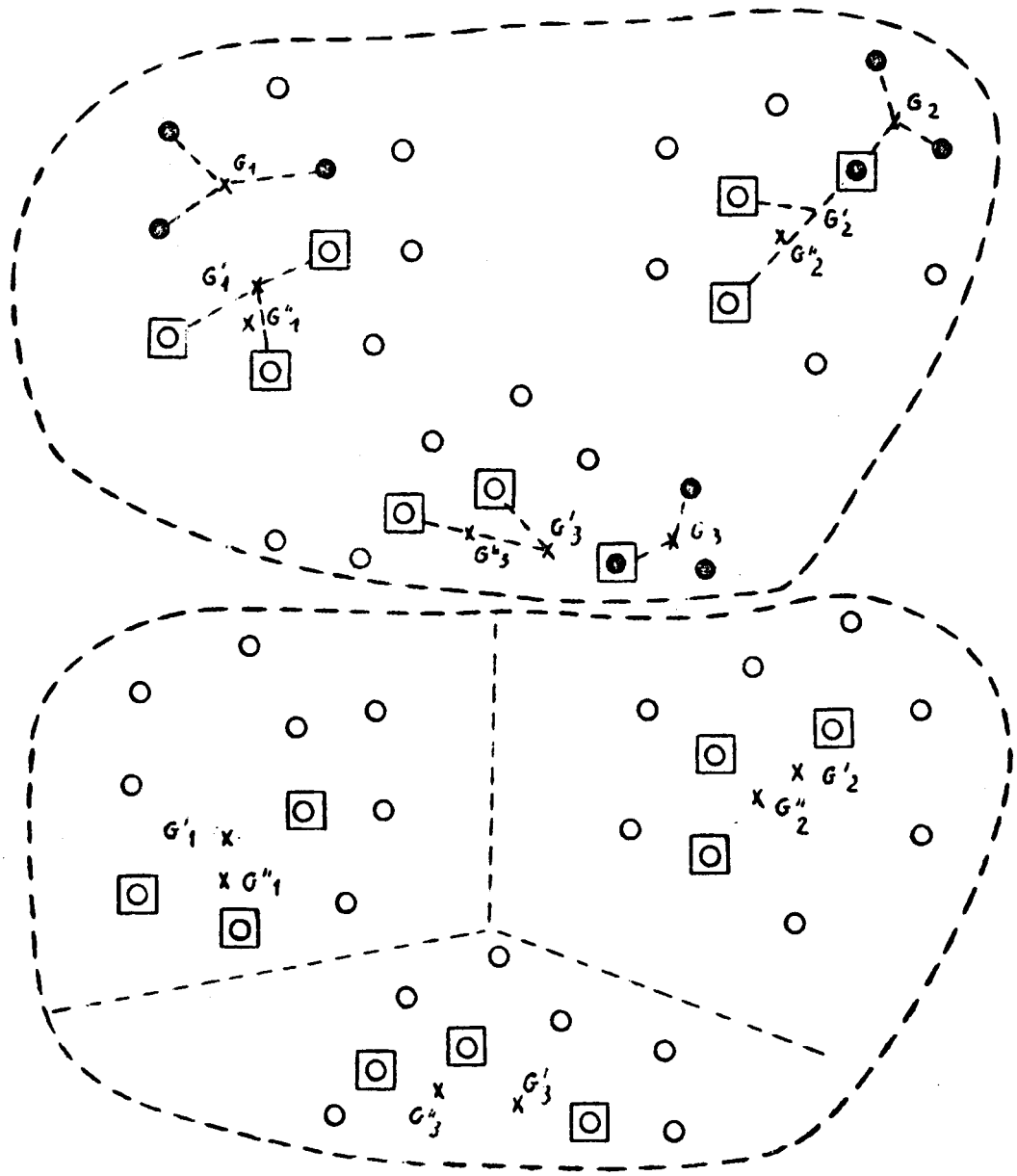
Exemple élémentaire (figure 24)

On dispose de points dans un plan que l'on désire répartir en trois classes. On suppose que l'on connaisse les étalons de départ (ronds noirs). On affecte chaque point à l'une des trois classes en fonction de sa distance aux centres de gravité G1, G2, G3 des trois groupes d'étalons.

On prend comme nouveaux étalons (ronds inscrits dans un carré) les échantillons les plus proches des centres de gravité G'1, G'2, G'3 des classes qui viennent d'être constituées. On recommence la procédure en répartissant à nouveau les points en fonction de leur distance aux contres de gravité G"1, G"2, G"3 des trois nouveaux groupes d'étalons.

Lorsque les étalons ne varient plus, la dernière partition obtenue correspond à la classification recherchée.

Cet exemple élémentaire montre la procédure de recherche des formes qui ne fait appel à aucune hypothèse a priori mais relève d'une technique purement exploratoire.



Généralisation

Dans l'exemple ci-dessus, la distance entre deux individus (les points) est la distance géométrique classique à 2 dimensions.

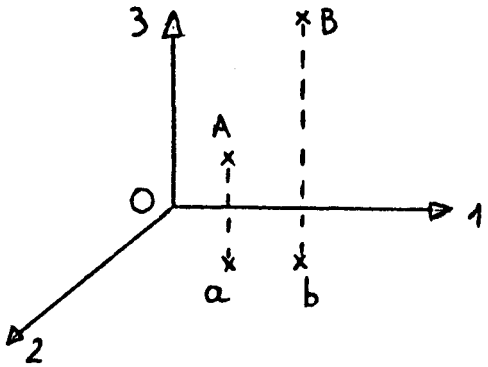
Cette distance se généralise à p dimensions et permet de partitionner suivant une procédure analogue à un ensemble de n observations à p variables (1è partie § 4.4.1.). Au lieu d'utiliser une distance enclidienne, on peut utiliser une distance du CHI^2 (1è partie § 4.4.2.) ou même d'autres distances dotées de propriétés particulières.

Place de la méthode des "nuées dynamiques" dans l'analyse de données.

Elle est complémentaire des méthodes d'analyse factorielle en composantes principales (distance euclidienne) et des correspondances (distance du CHI^2).

Lorsque le nombre n des échantillons analysés est grand, les sorties graphiques sont fréquemment assez difficiles à interpréter par suite de la superposition des projections des points représentatifs sur les plans factoriels. Pour déterminer les groupements éventuels de points dans l'espace des axes factoriels, il faudrait pouvoir examiner simultanément tous les plans factoriels.

Deux points peuvent en effet avoir des projections proches sur le plan (1,2) et avoir une cote très différente sur l'axe 3.



La méthode permet de pallier à cet inconvénient et de fournir les groupements existants indépendemment d'une sélection visuelle.

Une procédure d'utilisation constituée des méthodes de nuées dynamiques et d'analyse de données peut être la suivante :

1° Analyse de données sur l'ensemble des échantillons conduisant à la détermination de k groupements approximatifs.

2° Emploi des nuées dynamiques en fixant k classes et sélection des échantillons caractéristiques des k classes.

3° Analyse de données sur les seuls échantillons caractéristiques conduisant à l'obtention de graphiques clairs facilitant l'interprétation.

4° Extension de l'interprétation à la population globale

Formes fortes et formes faibles.

Lorsque l'on choisit au hasard les étalons initiaux des classes, on obtient une certaine partition (P1) de l'ensemble des n échantillons. Si l'on choisit d'autres étalons par un autre tirage au hasard, on obtient une partition (P2) et ainsi de suite.

Si après plusieurs tirages distincts, on obtient sensiblement les mêmes classes finales, on peut affirmer que les échantillons sont classifiables.

Une classe de la partition sera appelée "forme forte" si elle apparait distinctement quelque soit le tirage de départ ; elle sera appelée "forme faible" si elle évolue d'un tirage à l'autre. Les étalons des formes fortes représentent donc les "pôles" de la classification entre lesquels se répartit l'ensemble des échantillons que l'on peut alors comparer aux types fondamentaux.

Etat actuel des travaux

Le programme est disponible dès à présent sur IBM 1130. Il demande cependant un temps d'exécution assez long du fait de la faible capacité mémoire de cet ordinateur ; il sera prochainement adapté pour ordinateur IBM 360/40.

Les essais réalisés à ce jour (classification de courbes teneur/profondeur dans les sondages verticaux ; classification de courbes granulométriques ; recherche d'associations géochimiques) fournissent des résultats très encourageants. Ils feront l'objet d'une prochaine publication.

BIBLIOGRAPHIE SOMMAIRE

1° Ouvrages généraux

* STATISTIQUE ET INFORMATIQUE APPLIQUEES
  L. LEBART et J.P. FENELON DUNOD - 1971
* MODERN FACTOR ANALYSIS
  H.H. HARMAN THE UNIVERSITY OF CHICAGO PRESS. 1960
* MULTIVARIATE STATISTICAL METHODS
  D.F. MORRISSON Mc GRAW HILL - 1967
* MULTIVARIATE PROCEDURES IN THE BEHAVIORAL SCIENCES
  W.W. COOLEY et P.R. LOHNES JOHN WILEY - 1962
* AN INTRODUCTION TO REGRESSION AND CORRELATION
  K.W. SMILIE ACADEMIC PRESS - 1966

2° Applications géologiques

* METHODES MODERNES DE TRAITEMENT DE L'INFORMATION GEOLOGIQUE SUR ORDINATEUR
  TECHNIP - 1969
* JOURNAL OF THE INTERNATIONAL ASSOCIATION FOR MATHEMATICAL GEOLOGY
  - CRITICAL REVIEW OF SOME MULTIVARIATE PROCEDURES IN THE ANALYSIS OF GEOCHIMICAL DATA - Vol 1 n°2 - 1969
  - ORDINATION OF SEDIMENTS FROM THE CAPE HATTERAS CONTINENTAL MARGIN - Vol 2 n°2 - 1970
  - PRINCIPAL COMPONENTS ANALYSIS IN THE GEOCHEMISTRY AND MINERALOGY OF THE PORTA SKAIG TILLITE AND KILTY FANNEI SCHIST (Co DONEGAL-EIRE) - Vol 2 n°3 - 1970
* JOURNAL OF SEDIMENTARY PETROLOGY
  - THE USE OF FACTOR ANALYSIS IN DETERMINING DEPOSITIONAL ENVIRONMENTS FROM GRAIN - SIZE DISTRIBUTIONS - Vol 36 n°1 - 1966