Statistique et analyse des données

Thierry Foucart

Prévision d'une suite de tableaux de probabilités. Ajustement à des marges données

Statistique et analyse des données, tome 4, n° 2 (1979), p. 51-71

<http://www.numdam.org/item?id=SAD_1979__4_2_51_0>

Statistique et Analyse des Données
2 - 1979 pp. 51, 71

PREVISION D'UNE SUITE DE TABLEAUX DE PROBABILITES. AJUSTEMENT A DES MARGES DONNEES.

Thierry FOUCART

Assistant à l'Université de Haute-Bretagne (I.U.T. de Vannes)

Dans un article précédent, nous avons proposé un certain nombre de méthodes permettant la description d'une suite de tableaux de probabilités indexés par le temps.

Nous abordons ici le problème de la prévision d'une telle suite, éventuellement sous contraintes de marges et proposons une procédure conservant certains caractères communs à tous les tableaux.

1 - DESCRIPTION D'UNE SUITE DE TABLEAUX DE PROBABILITES INDEXES PAR LE TEMPS.

Nous avons proposé dans un article précédent [1] plusieurs méthodes permettant de décrire une suite de tableaux de probabilités indexés par le temps. La suite de l'exposé utilisant les objets mathématiques qu'elles introduisent, nous en donnons un aperçu avant de proposer une définition de la structure d'une suite de tableaux de probabilités et une procédure de prévision.

1.1 - Analyse des opérateurs de Burt.

Cette méthode est inspirée de la méthode STATIS [2] ; à chaque tableau de probabilités $P_{IJ}(t)$ défini sur le produit I × J pour les valeurs de t comprises entre 1 et k, on associe le tableau de Burt B(t) :

$$B(t) = \begin{bmatrix} D_{P_I}(t) & P_{IJ}(t) \\ P'_{IJ}(t) & D_{P_J}(t) \end{bmatrix}$$

où : - $P'_{IJ}(t)$ est le tableau transposé de $P_{IJ}(t)$
- $P_I(t)$, $P_J(t)$ sont les marges du tableau $P_{IJ}(t)$
- $D_{P_I}(t)$, $D_{P_J}(t)$ sont les matrices diagonales suivantes :

$$D_{P_I}(t) = \begin{bmatrix} P_{1\bullet}(t) & & \\ & \ddots & \\ & & P_{n\bullet}(t) \end{bmatrix} \qquad D_{P_J}(t) = \begin{bmatrix} P_{\bullet 1}(t) & & \\ & \ddots & \\ & & P_{\bullet m}(t) \end{bmatrix}$$

- n = card I m = card J
- k est le nombre de tableaux de la suite et K = {1,2,...k}

Le tableau de Burt B(t) est la matrice, dans la base canonique de $\mathbb{R}^{n+m}$, d'un opérateur qui est analogue à un opérateur de covariance. L'opérateur "de Burt" ainsi défini, que l'on note également B(t), appartient à l'espace euclidien des opérateurs symétriques de $\mathbb{R}^{n+m}$; le produit scalaire est égal à la trace (tr) du produit de composition [3] :

$$PS(B(t),B(t')) = tr(B(t).B(t'))$$

Pour décrire les angles entre les opérateurs B(t), on effectue l'analyse en composantes principales du tableau X constitué par les opérateurs écrits en lignes et considérés comme s'ils étaient des caractères :

$$X = \begin{bmatrix} B(1) \\ \vdots \\ B(t) \\ \vdots \\ B(k) \end{bmatrix}$$

Procéder ainsi revient à considérer le schéma de dualité [4]

$$\begin{array}{ccc} E = \mathbb{R}^k & \xleftarrow{\quad X \quad} & F^* \\ PS \uparrow \downarrow I & & I \uparrow \downarrow W \\ E^* & \xrightarrow[X']{} & F \subset \mathbb{R}^{(n+m)^2} \end{array}$$

où :

- F est un sous-espace de l'espace des opérateurs symétriques définis sur $\mathbb{R}^{n+m}$
- I désigne la métrique identité
- W = X'X.

Effectuer l'analyse en composantes principales du tableau X consiste à diagonaliser la matrice PS des produits scalaires, qui est analogue à une matrice de variances covariances si les opérateurs ne sont pas normés, ou à une matrice des corrélations si les opérateurs considérés sont unitaires. Les facteurs principaux sont les vecteurs propres associés aux valeurs propres non nulles de la matrice PS et les composantes principales les images des facteurs dans F par l'application X'. Les opérateurs de Burt B(t) sont des combinaisons linéaires des composantes principales normées C(ℓ) et réciproquement. On a donc les relations ci-dessous :

$$B(t) = \sum_{\ell=1}^{p} a(t,\ell)\ C(\ell) \qquad (1)$$

$$C(\ell) = \sum_{t=1}^{k} b(t,\ell)\ B(t) \qquad (2)$$

1.2 - Application de l'analyse factorielle des correspondances.

Les tableaux $P_{IJ}(t)$ définissent des lois de probabilités sur l'ensemble I × J ; Π_{IJ} étant le tableau moyen, on peut calculer les distances du chi-deux de centre Π_{IJ} entre les tableaux ; l'analyse factorielle du tableau de distances obtenu est équivalente à l'analyse factorielle des correspondances du tableau P^K_{IJ} constitué par la juxtaposition des tableaux $P_{IJ}(t)$ vectorialisés :

$$P^K_{IJ} = \begin{bmatrix} P_{1,1}(1) & P_{1,1}(t) & P_{1,1}(k) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ P_{i,j}(1) & P_{i,j}(t) & P_{i,j}(k) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ P_{n,m}(1) & P_{n,m}(t) & P_{n,m}(k) \end{bmatrix}$$

On utilise ici l'isomorphisme classique entre les matrices réelles de taille n × m et l'espace vectoriel $\mathbb{R}^{nm}$.

En effectuant l'analyse factorielle des correspondances de ce tableau, on détermine une base de q facteurs $F_{IJ}(\ell)$ que l'on peut considérer soit comme des vecteurs de $\mathbb{R}^{nm}$, soit comme des tableaux sur I × J, dans laquelle les tableaux $P_{IJ}(t)$ s'écrivent :

$$P_{IJ}(t) = \Pi_{IJ} + \sum_{\ell=1}^{q} u(\ell,t)\ F_{IJ}(\ell) \qquad (3)$$

Sous forme matricielle, on obtient :

$$P_{IJ}^{K} = \Pi_{IJ}^{K} + F_{IJ}^{Q}\ U_{Q}^{K} \qquad (3')$$

avec les notations suivantes :

Π_{IJ}^{K} est le tableau à nm lignes et k colonnes, la t^{e} colonne étant définie par le tableau moyen Π_{IJ} vectorialisé.

U_{Q}^{K} est le tableau à q lignes et k colonnes, dont le terme général u(ℓ,t) est la coordonnée du tableau $P_{IJ}(t)$ sur le facteur $F_{IJ}(\ell)$.

Nous savons d'autre part que les facteurs $F_{IJ}(\ell)$ sont des combinaisons linéaires des tableaux $P_{IJ}(t)$:

$$F_{IJ}(\ell) = \sum_{t=1}^{k} v(t,\ell)\ P_{IJ}(t)$$

Sous forme matricielle, on obtient :

$$F_{IJ}^{Q} = P_{IJ}^{K}\ V_{K}^{Q}$$

où V_{K}^{Q} est le tableau à k lignes et ℓ colonnes de terme général v(t,ℓ).

En sommant sur J, on aboutit à la relation (4) suivante :

$$F_{I}^{Q} = P_{I}^{K}\ V_{K}^{Q} \qquad (4)$$

1.3 - Analyse des marges.

A chaque tableau $P_{IJ}(t)$ sont associées deux marges $P_{I}(t)$ et $P_{J}(t)$. Elles définissent deux suites de tableaux de probabilités indexés par le temps, que l'on peut donc décrire par l'une des méthodes précédentes. L'analyse des opérateurs de Burt revenant à définir comme distance entre deux marges la distance euclidienne classique, nous proposons d'utiliser l'analyse factorielle des correspondances.

Soit P_{I}^{K} le tableau défini ci-dessous :

$$P_{I}^{K} = \begin{bmatrix} P_{1}(1) & \cdots & P_{1}(k) \\ \vdots & & \vdots \\ P_{n}(1) & \cdots & P_{n}(k) \end{bmatrix}$$

Effectuer l'analyse factorielle des correspondances de ce tableau revient à déterminer une base de r facteurs $(F_I(h))$ $h = 1,\ldots,r$, dans laquelle les marges $P_I(t)$ s'écrivent :

$$P_I(t) = \Pi_I + \sum_{h=1}^{r} f(h,t)\ F_I(h) \qquad (5)$$

Sous forme matricielle :

$$P_I^K = \Pi_I^K + F_I^R\ F_R^K \qquad (5')$$

où Π_I^K est le tableau à n lignes et k colonnes, chaque colonne étant égale à Π_I.

F_I^R est le tableau à n lignes et r colonnes, la h^e colonne étant définie par le facteur $F_I(h)$.

F_R^K est le tableau à r lignes et k colonnes ; le terme général est la coordonnée de la marge $P_I(t)$ sur le h^e facteur : $f(h,t)$.

En ce qui concerne la suite $(P_J(t))$ $t \in K$, les notations sont les suivantes :

P_J^K est le tableau défini par les marges $P_J(t)$ écrites en colonnes.

Π_J^K est le tableau à m lignes et k colonnes, chaque colonne étant égale à Π_J.

F_J^S est le tableau à m lignes et s colonnes, la h^e colonne étant définie par le facteur $F_J(h)$ et s étant le nombre de facteurs.

σ_S^K est le tableau à s lignes et k colonnes : le terme général est la coordonnée de la marge $P_J(t)$ sur le h^e facteur $\sigma(h,t)$.

Les marges $P_J(t)$ s'écrivent, sous forme matricielle :

$$P_J^K = \Pi_J^K + F_J^R\ \sigma_R^K \qquad (6)$$

2 - STRUCTURE ET PREVISION D'UNE SUITE DE TABLEAUX DE PROBABILITES.

Nous abordons dans ce paragraphe la notion de structure d'une suite de tableaux de probabilités. Introduite de deux façons différentes, cette notion sera utilisée pour effectuer une prévision ou un ajustement.

2.1 - Application de l'analyse des opérateurs de Burt.

Définition.-On appelle structure de la suite $(P_{IJ}(t))$ $t \in K$ la base de l'espace engendré par les opérateurs de Burt associés aux tableaux, définie par les composantes principales obtenues dans l'analyse de ces opérateurs (cf. 1.1).

-Nous disons qu'un tableau de probabilités P_{IJ} défini sur $I \times J$ possède la structure de la suite lorsque l'opérateur de Burt B associé est combinaison linéaire des composantes principales ci-dessus.

Proposition 1. Un tableau de probabilités P_{IJ} possède la structure de la suite si et seulement s'il est combinaison linéaire des tableaux de probabilités $P_{IJ}(t)$.

Soit en effet $\{C(\ell)\ ;\ \ell = 1,\ldots,p\}$ la base des composantes principales obtenue par l'analyse des opérateurs de Burt B(t). Le tableau P_{IJ} possède la structure de la suite si et seulement si l'opérateur de Burt B associé est combinaison linéaire des composantes principales :

$$B = \sum_{\ell=1}^{p} a(\ell)\ C(\ell) \qquad (7)$$

Or, les composantes principales $C(\ell)$ sont elles-mêmes des combinaisons linéaires des opérateurs $B(t)$; l'équation (2) permet de les écrire sous la forme de tableaux $C(\ell)$:

$$C(\ell) = \begin{bmatrix} D_{Q_I(\ell)} & Q_{IJ}(\ell) \\ Q'_{IJ}(\ell) & D_{Q_J(\ell)} \end{bmatrix}$$

où :

$$\begin{aligned} Q_{IJ}(\ell) &= \sum_{t=1}^{k} b(t,\ell)\ P_{IJ}(t) \\ Q_I(\ell) &= \sum_{t=1}^{k} b(t,\ell)\ P_I(t) \\ Q_J(\ell) &= \sum_{t=1}^{k} b(t,\ell)\ P_J(t) \end{aligned} \qquad (8)$$

$D_{Q_I(\ell)}$ et $D_{Q_J(\ell)}$ sont les matrices diagonales associées aux marges $Q_I(\ell)$ et $Q_J(\ell)$ du tableau $Q_{IJ}(\ell)$.

De l'équation (1) on déduit les égalités suivantes :

$$\begin{aligned} P_{IJ} &= \sum_{\ell=1}^{p} a(\ell)\ Q_{IJ}(\ell) \\ P_I &= \sum_{\ell=1}^{p} a(\ell)\ Q_I(\ell) \\ P_J &= \sum_{\ell=1}^{p} a(\ell)\ Q_J(\ell) \end{aligned} \qquad (9)$$

Le rapprochement des équations (8) et (9) montre le tableau P_{IJ} (resp. P_I,P_J) est une combinaison linéaire des tableaux $P_{IJ}(t)$ (resp. $P_I(t)$,$P_J(t)$).

Réciproquement, soit P_{IJ} un tableau de probabilités sur $I \times J$ égal à une combinaison linéaire des tableaux $P_{IJ}(t)$:

$$P_{IJ} = \sum_{t=1}^{k} a'(t)\ P_{IJ}(t) \qquad (10)$$

Les opérateurs $B(t)$ sont des combinaisons linéaires des composantes principales $C(\ell)$; l'équation (1) a pour conséquence les équations (11) suivantes :

$$\begin{cases} P_{IJ}(t) = \sum_{\ell=1}^{p} a(t,\ell)\ Q_{IJ}(\ell) \\ P_I(t) = \sum_{\ell=1}^{p} a(t,\ell)\ Q_I(\ell) \\ P_J(t) = \sum_{\ell=1}^{p} a(t,\ell)\ Q_J(\ell) \end{cases} \qquad (11)$$

Des équations (9) et (10), on déduit :

$$P_{IJ} = \sum_{\ell=1}^{p} \sum_{t=1}^{k} a'(t)\, a(t,\ell)\, Q_{IJ}(\ell)$$

Cette relation montre que l'opérateur de Burt B associé au tableau P_{IJ} est combinaison linéaire des composantes principales.

Proposition 2. Soit P_{IJ} un tableau de probabilités sur $I \times J$ possèdant la structure de la suite $(P_{IJ}(t))$ $t \in K$. Alors, la marge P_I (resp. P_J) du tableau P_{IJ} possède la structure de la suite $(P_I(t))$ $t \in K$ (resp. de la suite $(P_J(t))$ $t \in K$).

En effet, l'équation (10) a pour conséquence immédiate :

$$P_I = \sum_{t=1}^{k} a'(t)\, P_I(t) \ .$$

Proposition 3. Soit P_{IJ} un tableau de probabilités sur $I \times J$ possèdant la structure de la suite $(P_{IJ}(t))$ $t \in K$. Ses coordonnées $(a(\ell))$ $\ell = 1,\ldots,p$ vérifient les conditions (12) ci-dessous :

$$\sum_{\ell=1}^{p} a(\ell) \sum_{i,j} Q_{i,j}(\ell) = 1$$
$$\forall (i,j) \in I \times J \qquad \sum_{\ell} a(\ell)\, Q_{i,j}(\ell) \geqslant 0 \qquad (12)$$

Cette proposition est une conséquence immédiate des équations (11).

Les conditions (12) sont caractéristiques d'un simplexe de $\mathbb{R}^p$. Réciproquement, un élément de ce simplexe définit évidemment un tableau de probabilités sur $I \times J$.

2.2 - Application de l'analyse factorielle des correspondances.

La relation (3) nous donne une décomposition des tableaux $P_{IJ}(t)$ en fonction des facteurs $F_{IJ}(\ell)$ et du tableau moyen Π_{IJ}. On peut définir la structure de la suite $(P_{IJ}(t))$ $t \in K$ en utilisant cette décomposition :

Définition. - On appelle structure de la suite $(P_{IJ}(t))$ $t \in K$ l'ensemble formé par le tableau moyen Π_{IJ} et la base des facteurs obtenue par l'analyse factorielle des correspondances du tableau P_{IJ}^{K} (cf. 1.2).

- Nous disons qu'un tableau de probabilités sur $I \times J$ possède la structure de la suite s'il est égal à la somme du tableau moyen et d'une combinaison linéaire des facteurs ci-dessus.

Proposition 4. Les deux notions de structure sont équivalentes.
Soit P_{IJ} un tableau de probabilités sur $I \times J$ tel que :

$$P_{IJ} = \Pi_{IJ} + \sum_{\ell=1}^{q} u(\ell)\, F_{IJ}(\ell)$$

On sait que les facteurs $F_{IJ}(\ell)$ sont des combinaisons linéaires des tableaux de probabilités $P_{IJ}(t)$. Le tableau Π_{IJ} étant le tableau moyen, on en déduit que le tableau P_{IJ} est une combinaison linéaire des tableaux $P_{IJ}(t)$ et donc, en vertu de la proposition 1, qu'il possède la structure de la suite telle que nous l'avons définie au paragraphe 2.1.

Réciproquement, soit un tableau P_{IJ} tel que :

$$P_{IJ} = \sum_{t=1}^{k} \alpha(t)\ P_{IJ}(t)$$

On en déduit, par sommation de chaque membre :

$$\sum_{t=1}^{k} \alpha(t) = 1.$$

En remplaçant $P_{IJ}(t)$ par l'expression donnée en (3), on obtient :

$$P_{IJ} = \sum_{t=1}^{k} \alpha(t)\ (\Pi_{IJ} + \sum_{\ell=1}^{q} u(\ell,t)\ F_{IJ}(\ell))$$

$$= \Pi_{IJ} + \sum_{\ell=1}^{q} \sum_{t=1}^{k} \alpha(t)\ u(\ell,t)\ F_{IJ}(\ell).$$

Le tableau P_{IJ} possède donc la structure de la suite au sens où nous venons de la définir.

2.3 - Prévision d'une suite de tableaux de probabilités.

Deux méthodes, correspondant chacune à une définition de la structure de la suite, peuvent être utilisées pour prévoir le tableau de probabilités. Considérons tout d'abord l'équation (1) :

$$B(t) = \sum_{\ell=1}^{p} a(t,\ell)\ C(\ell).$$

Cette équation permet de reconstruire les opérateurs de Burt B(t) à partir des composantes principales C(ℓ) et des coordonnées a(t,ℓ). Pourquoi ne pas utiliser cette formule pour prévoir l'opérateur à l'horizon h ? Les composantes principales C(ℓ) étant fixées, on propose de régresser les coordonnées a(t,ℓ) par rapport au temps. Si l'on note a(h,ℓ) les prévisions à l'horizon h des coordonnées, la prévision de l'opérateur est donnée par :

$$B(h) = \sum_{\ell=1}^{p} a(h,\ell)\ C(\ell)$$

La prévision du tableau est alors donnée par :

$$P_{IJ}(h) = \sum_{\ell=1}^{p} a(h,\ell)\ Q_{IJ}(\ell).$$

Cette procédure présente deux avantages :

- le nombre de régressions effectuées est égal au nombre de composantes principales retenues, c'est-à-dire en général bien inférieur au nombre de termes des tableaux $P_{IJ}(t)$

- le tableau prévu possède la structure de la suite ; il conserve donc les caractéristiques communes à tous les autres tableaux. Par contre, l'obtention d'un tableau de probabilités n'est pas assurée si les régressions ne sont pas effectuées sous les contraintes (12) ou si certaines composantes principales n'ont pas été retenues.

Dans le cas où la description de la suite est réalisée par l'analyse factorielle des correspondances du tableau P^{K}_{IJ}, on peut utiliser la formule de reconstruction des données (3) pour effectuer la prévision.

Le tableau prévu est alors donné par l'équation ci-dessous :

$$P_{IJ}(h) = \Pi_{IJ} + \sum_{\ell=1}^{q} u(\ell,h)\ F_{IJ}(\ell)$$

où les coefficients $u(\ell,h)$ sont les prévisions à l'horizon h des coefficients $u(\ell,t)$.

La similitude avec la méthode précédente est évidente. On en retrouve les avantages :

- le nombre de régressions effectuées est égal au nombre de facteurs retenus

- le tableau prévu possède la structure de la suite. De plus, la somme du tableau $P_{IJ}(h)$ est égale à 1 puisque les facteurs $F_{IJ}(\ell)$ sont centrés. Les régressions seront donc effectuées sous la seule contrainte de positivité.

3 - AJUSTEMENT D'UN TABLEAU DE PROBABILITES A DES MARGES DONNEES A PRIORI.

Le problème d'ajustement d'un tableau de probabilités à des marges données a priori est très classique. En général, on cherche à déterminer le tableau vérifiant les contraintes de marges le plus proche du tableau donné au sens d'une certaine distance [5]. La procédure consiste alors à résoudre un programme linéaire [6]. Nous proposons ici une méthode différente : il existe une relation linéaire entre un tableau P_{IJ} possèdant une structure donnée et ses marges. Pour déterminer un tableau possèdant la structure et vérifiant les contraintes de marges, il suffit de considérer la relation inverse, dont on discutera de l'existence et des propriétés.

3.1 - Application de l'analyse des opérateurs de Burt à l'ajustement d'un tableau à des marges fixées.

Nous cherchons une relation entre un tableau P_{IJ} possèdant la structure de la suite $(P_{IJ}(t)\ t \in K$ et ses marges.

L'équation (9) s'écrit sous forme matricielle :

$$P_I = Q_I^P A_P$$

où Q_I^P est le tableau à n lignes, chaque colonne étant définie par $Q_I(\ell)$.

A_P est le vecteur colonne des coordonnées de l'opérateur B sur les composantes principales, c'est-à-dire sur la structure.

L'équation (8) s'écrit également sous forme matricielle :

$$Q_I^P = P_I^K B_K^P$$

On en déduit la proposition suivante :

Proposition 5. Les marges P_I, P_J du tableau P_{IJ} et les coordonnées A_P de l'opérateur de Burt B associé au tableau P_{IJ} sur les composantes principales vérifient les relations :

$$P_I = P_I^K B_K^P A_P \qquad P_J = P_J^K B_K^P A_P \qquad (13)$$

Les matrices $P_I^K B_K^P$ et $P_J^K B_K^P$ ne sont guère faciles à manipuler à cause de leur taille ; on peut établir une relation plus simple à l'aide des coordonnées des marges P_I (resp. P_J) sur les r facteurs $F_I(h)$ (resp. les s facteurs $F_J(h)$).

La proposition 2 montre que la marge P_I possède la structure de la suite ; on a donc la relation:

$$P_I = \Pi_I + \sum_{h=1}^{r} f(h)\, F_I(h) \qquad (14)$$

Sous la forme matricielle, cette équation s'écrit :

$$P_I = \Pi_I + F_I^R\, F_R$$

où : F_I^R est le tableau dont les colonnes sont définies par les facteurs $F_I(h)$

F_R est le vecteur colonne des coordonnées de P_I sur les facteurs.

En remplaçant dans l'équation (13) P_I par l'expression (14) et P_I^K par l'expression (5'), on obtient :

$$\Pi_I + F_I^R\, F_R = \Pi_I^K\, B_K^P\, A_P + F_I^R\, F_R^K\, B_K^P\, A_P$$

Le terme $\Pi_I^K\, B_K^P\, A_P$ s'écrit sous la forme suivante :

$$\Pi_I^K\, B_K^P\, A_P = \Pi_I \sum_{\ell=1}^{p} a(\ell) \sum_{t=1}^{k} b(t,\ell)$$

L'équation (8) montre que :

$$\sum_{t=1}^{k} b(t,\ell) = \sum_{i,j} Q_{i,j}(\ell)$$

L'équation (12) montre que :

$$\sum_{\ell=1}^{p} a(\ell) \sum_{t=1}^{k} b(t,\ell) = 1.$$

On en déduit la proposition 6.

Proposition 6. Soit F_R le vecteur des coordonnées de la marge P_I du tableau P_{IJ} dans la base des r facteurs $F_I(h)$; soit G_S le vecteur des coordonnées de la marge P_J du tableau P_{IJ} dans la base des s facteurs $P_J(h)$. Les vecteurs F_R, G_S, A_P vérifient les relations ci-dessous :

$$F_R = F_R^K\, B_K^P\, A_P \qquad G_S = G_S^K\, B_K^P\, A_P \qquad (15)$$

Les matrices $F_R^K\, B_K^P$ et $G_S^K\, B_K^P$ sont respectivement de dimensions (r,p) et (s,p).

3.2. - <u>Application de l'analyse des correspondances à l'ajustement d'un tableau de probabilités à des marges fixées.</u>

Dans le paragraphe précédent, la structure est définie par la base des composantes principales obtenues par l'analyse des opérateurs de Burt. Nous nous plaçons maintenant dans le deuxième cas : la structure de la suite $(P_{IJ}(t))\ t \in K$ est définie par le tableau moyen Π_{IJ} et la base des facteurs $(F_{IJ}(\ell))\ \ell \in Q$ obtenue par l'analyse factorielle des correspondances du tableau P_{IJ}^K. Les relations entre le tableau P_{IJ} et ses marges s'écrivent presque de la même façon que dans le paragraphe précédent.

Soit donc P_{IJ} un tableau de probabilités possèdant la structure de la suite :

$$P_{IJ} = \Pi_{IJ} + \sum_{\ell=1}^{q} u(\ell)\, P_{IJ}(\ell)$$

matriciellement: $$P_{IJ} = \Pi_{IJ} + F_{IJ}^Q\, U_Q$$

Par sommation sur J, on obtient :

$$P_I = \Pi_I + F_I^Q \, U_Q$$

En remplaçant F_I^Q par l'expression donnée en (4), avec les mêmes notations, on aboutit à la proposition 7 :

Proposition 7. Les marges P_I et P_J du tableau P_{IJ} et les coordonnées U_Q du tableau P_{IJ} sur la base des facteurs $(F_{IJ}(\ell))$ $\ell \in Q$ vérifient les relations :

$$P_I = \Pi_I + P_I^K \, V_K^Q \, U_Q \qquad P_J = \Pi_J + P_J^K \, V_K^Q \, U_Q \qquad (16)$$

Proposition 8. Les coordonnées U_Q du tableau P_{IJ} sur la base des facteurs $(F_{IJ}(\ell))$ $\ell \in Q$, les coordonnées F_R de la marge P_I sur la base des facteurs $(F_I(h))$ $h \in R$, et les coordonnées G_S de la marge P_J sur la base des facteurs $(F_J(h))$ $h \in S$ vérifient les relations :

$$F_R = F_R^K \, V_K^Q \, U_Q \qquad G_S = G_S^K \, V_K^Q \, U_Q \qquad (17)$$

La démonstration consiste à remplacer dans les équations (16) les vecteurs P_I et P_I^K par les expressions données en (14) et (5).

4 - APPLICATION PRATIQUE DES METHODES PROPOSEES.

Nous avons appliqué les méthodes proposées dans les paragraphes précédents de la façon suivante : la suite des tableaux de probabilités est décrite par l'analyse des opérateurs de Burt (cf. 1.1) ; retenant alors les deux premières composantes principales (cf. 2.1), nous avons effectué une prévision pour janvier 1979 (cf. 2.3) ; les marges du tableau prévu ont été projetées en éléments supplémentaires [7] sur les plans principaux obtenus par l'analyse factorielle des correspondances des suites des marges (cf. 1.3) ; enfin, nous avons reconstruit ce tableau à partir de ses marges (cf. 3.2).

Les tableaux analysés donnent la répartition des chercheurs en Sciences Humaines au C.N.R.S. suivant le grade et la section ; ces tableaux, échelonnés de façon irrégulière entre 1968 et 1978 sont au nombre de 9. L'annexe 1 donne la répartition des chercheurs en janvier 1978, avec la nomenclature des grades et des sections, et précise les dates des différents tableaux.

4.1 - Description de la suite. Choix de la structure.

Nous avons effectué l'analyse de la suite considérée par la méthode décrite au paragraphe 1.1. La matrice des produits scalaires, les facteurs et les valeurs propres sont donnés en annexe 2. Les produits scalaires étant positifs, les coordonnées du premier facteur sont positives [1]. La figure 1 donne la représentation des opérateurs dans le plan des deux premières composantes

[1] Théorème de Frobenius [8] : les coordonnées du vecteur propre associé à la plus grande valeur propre d'une matrice semi définie positive à termes positifs sont toutes de même signe .

principales normées : l'évolution apparaît régulière, sauf en 1975 et 1977. Les points représentatifs des opérateurs sont alignés le long d'une droite, qui correspond à l'hyperplan défini au paragraphe 2.1 (le coefficient de corrélation linéaire entre les coordonnées sur chaque composante principale est égal à -0.97).
Le point 79 correspond à la prévision que nous avons effectuée.

Le choix de la structure est facile : on ne retient que les deux premières composantes principales, qui représentent 99,9 % de l'inertie totale.

4.2 - Prévision de la répartition des chercheurs en janvier 1979.

La procédure utilisée est décrite au paragraphe 1.3. Etant donné le petit nombre des tableaux, des projections empiriques des coordonnées des opérateurs ont été jugées suffisantes. On comparera cette prévision avec la répartition effectuée en janvier 1979 donnée ci-après.

Prévision des coordonnées :

- 1er axe : 0.7
- 2ème axe : 0.125

effectif total (fixé a priori) : 1 350

	Directeur de Recherche	Maître de Recherche	Chargé de Recherche	Attaché de Recherche	marge (sections)
Anthropologie (an)	9,72	45,55	109,39	44,17	208
Sociologie (so)	11,65	24,92	99,28	40,45	176
Géographie (gé)	2,51	8,77	32,62	35,57	81
Sciences Economiques (sé)	4,15	26,62	40,86	49,70	122
Sciences juridiques (sj)	8,36	15,41	37,92	39,59	101
Linguistique générale (lg)	8,96	10,67	33,29	45,09	98
Linguistique France (lf)	3,21	8,96	38,85	17,14	68
Civilisation classique (cc)	2,13	8,68	23,96	23,72	59
Civilisation orientale (co)	12,06	19,40	78,49	54,02	163
Histoire médiévale (mé)	5,76	12,43	42,59	32,94	93
Histoire moderne (mo)	1,95	9,96	19,41	47,38	78
Philosophie (ph)	5,84	11,47	41,82	43,58	103
Marge (grades)	77	202	597	474	1 350

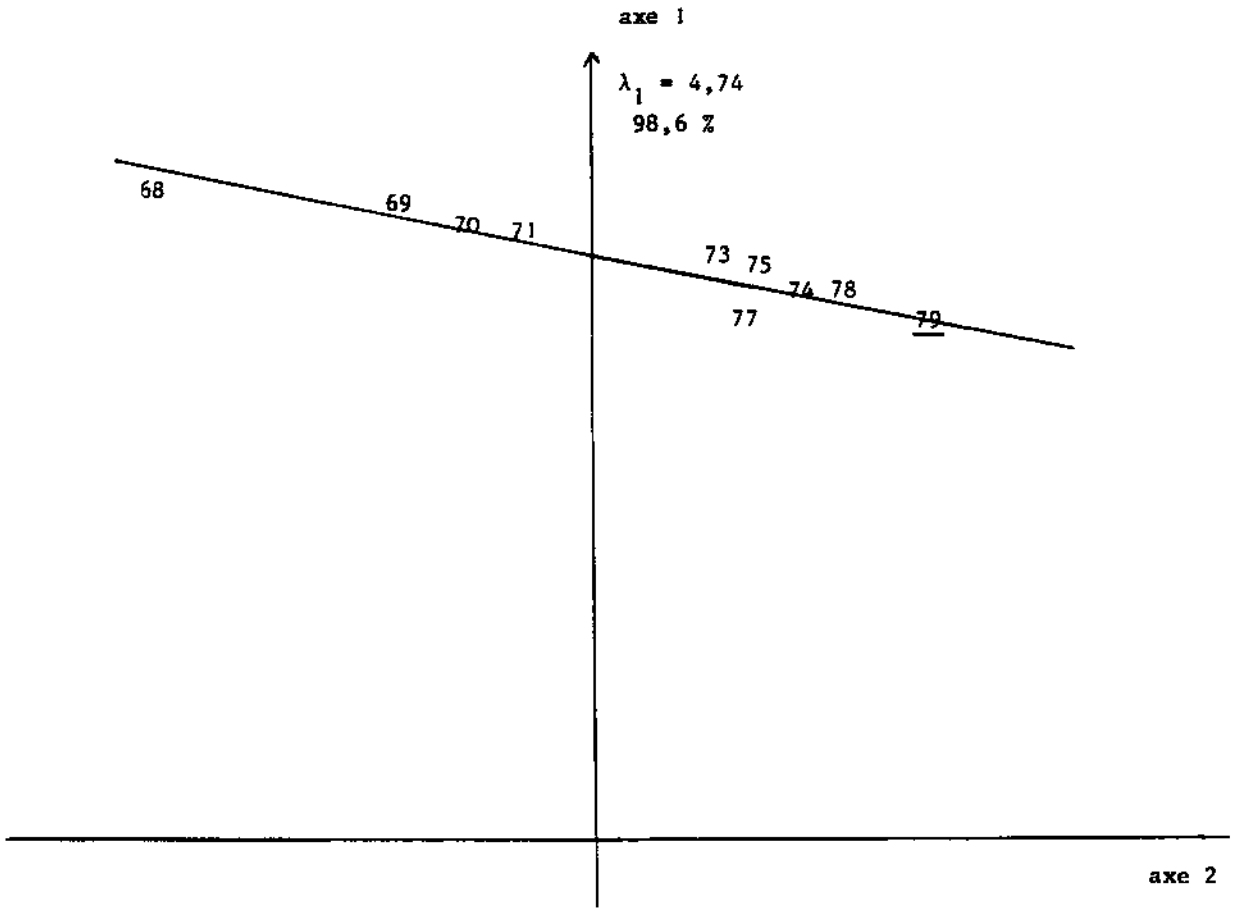

Figure 1 : Analyse des opérateurs de Burt

Effectifs réels des chercheurs du C.N.R.S. en Sciences Humaines au 22 janvier 1979 :

	Directeur de Recherche	Maître de Recherche	Chargé de Recherche	Attaché de Recherche	Marge (sections)
Anthropologie	9	44	99	55	207
Sociologie	11	32	113	63	219
Géographie	4	6	31	33	74
Sciences économiques	4	25	48	46	123
Sciences juridiques	6	13	36	39	94
Linguistique générale	6	10	38	40	94
Linguistique France	3	10	36	16	65
Civilisation classique	2	7	27	22	58
Civilisation orientale	7	20	77	50	154
Histoire médiévale	4	13	38	32	87
Histoire moderne	1	7	21	44	73
Philosophie	6	11	47	34	98
Marge (grades)	63	198	611	474	1 346

4.3 - Etude des marges.

Nous avons effectué l'analyse factorielle des correspondances de chacun des tableaux constitués par les suites des marges (cf. 1.3). Les marges du tableau prévu pour janvier 1979 ont été projetées en éléments supplémentaires et sont représentées par les points 79.

En ce qui concerne les marges correspondant aux grades (figure 2) : leur évolution dans le temps apparaît assez régulière, sauf en 1975 et 1977 ; elles s'ordonnent de façon presque chronologique le long de l'axe 1, qui peut être considéré comme axe des temps. Cette tendance s'explique par une opposition entre le grade attaché de recherche (AR) et les trois autres grades chargé de recherche (CR), maître de recherche (MR), directeur de recherche (DR) : la proportion d'attachés de recherche a diminué au cours de la période étudiée au profit des trois autres grades.

En ce qui concerne les marges correspondant aux sections (figure 3) : l'année 1974 est la seule à présenter une particularité dans l'évolution des sections. A droite de l'axe 1 se trouvent les sections dont la proportion de chercheurs a diminué : c'est le cas des sections histoire moderne (mo) et anthropologie (an). Inversement, la proportion a augmenté dans les sections placées à gauche de l'axe, par exemple en géographie (gé) et sciences économiques (sé).

Les marges du tableau prévu se projettent dans les deux cas de façon homogène avec celles des tableaux analysés. La prévision que nous avons réalisée au paragraphe 4.2 apparaît très compatible avec l'évolution des marges.

4.4 - Relations entre les opérateurs et les marges des tableaux associés.

Au paragraphe 4.1 nous avons défini la structure par les deux premières composantes principales ; en ne conservant que les deux premiers facteurs dans l'analyse des marges, et en appliquant la

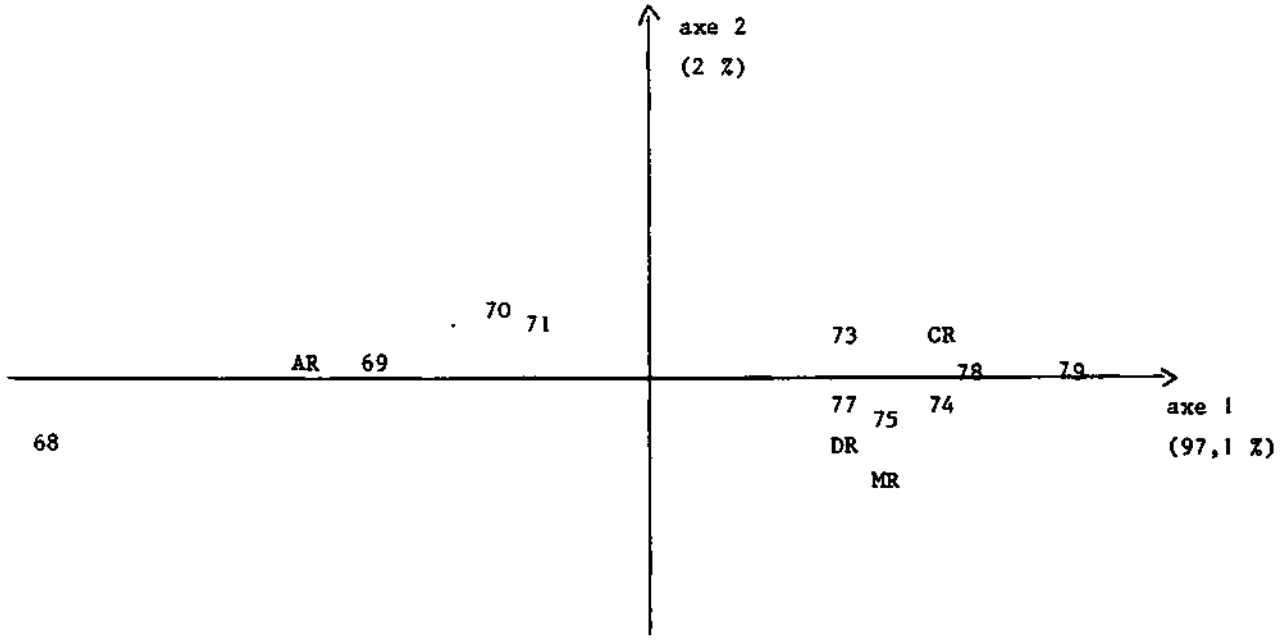

Figure 2 : Evolution des grades

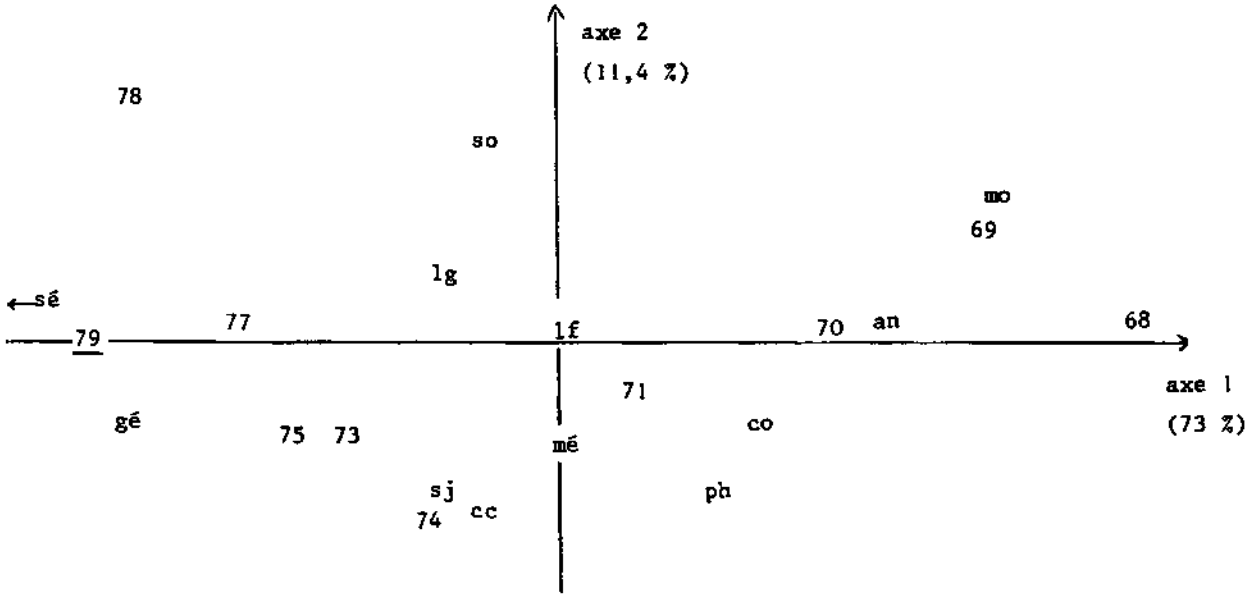

Figure 3 : Evolution des sections

procédure décrite au paragraphe 2, nous obtenons les matrices suivantes :

$$M = F_R^K B_K^P = \begin{bmatrix} -3.12 & 10^{-3} & 1.395 & \\ 1.13 & 10^{-4} & 2.07 & 10^{-3} \end{bmatrix} \qquad N = G_S^K = B_K^P = \begin{bmatrix} 1.36 & 10^{-3} & -0.591 \\ 6.59 & 10^{-5} & -0.022 \end{bmatrix}$$

La matrice M (resp. N) donne les deux premières coordonnées de la marge correspondant aux grades (resp. sections) en fonction des deux premières coordonnées de l'opérateur. Nous donnons en annexe 3 les coordonnées réelles des marges sur les deux premiers axes et les coordonnées calculées à l'aide des matrices ci-dessus. On constate que les coordonnées sur le premier axe sont en général bien reconstruites, contrairement aux coordonnées sur le second axe. Dans le cas des marges correspondant aux sections, la qualité de la reconstruction est moins bonne : cela peut s'expliquer par la différence entre les contributions à l'inertie des deux premières valeurs propres de chaque analyse (99 % dans le premier cas, 84 % dans le deuxième).

On peut inverser ces matrices pour trouver un tableau possèdant des marges données. Les résultats obtenus ne sont guère acceptables, à cause sans doute de la relation liant les coordonnées des opérateurs (cf. 2.1 relation 12).

4.5 - Reconstruction du tableau prévu à partir de ses marges.

Nous nous proposons dans ce paragraphe d'étudier la reconstruction d'un tableau à partir de marges données ; pour cela, nous avons calculé les matrices considérées au paragraphe 3.2, permettant de déduire des coordonnées d'un tableau sur un système d'axes factoriels les coordonnées de ses marges sur les systèmes d'axes correspondant :

$$M_1 = F_R^K V_K^Q = \begin{bmatrix} -0.70814 & 0.27704 \\ 0.01637 & 0.27837 \end{bmatrix} \qquad N_1 = G_S^K V_K^Q = \begin{bmatrix} 0.38059 & -0.19545 \\ 0.02179 & 0.38019 \end{bmatrix}$$

L'annexe 4 donne la reconstruction des marges de la suite par ces matrices, à l'aide donc des deux premiers facteurs.

Les résultats apparaissent satisfaisants.

En inversant la matrice M_1, on peut reconstruire un tableau dont la marge associée aux grades coïncide sur les deux premiers axes factoriels du référentiel correspondant avec une marge fixée à l'avance ; on pourrait procéder de la même façon à propos de la répartition suivant les sections.

Nous avons effectué le calcul en utilisant les marges du tableau prévu ; quelle que soit la marge considérée, ce tableau est reconstruit de façon presque parfaite.

Il est remarquable qu'une seule marge suffise à déterminer le tableau de probabilités ; cette propriété indique la "rigidité" de la structure.

CONCLUSION.

Les méthodes de prévision de tableaux et d'ajustement à des marges données que nous venons de développer présentent l'avantage sur les méthodes classiques de conserver ce que l'on a appelé la structure de la suite et de nécessiter un nombre faible de régressions. Il semble toutefois indispensable de préciser quelles sont les propriétés des tableaux contenues dans cette structure, et de choisir avec précaution les facteurs ou les composantes principales à retenir.

D'autre part, les tableaux reconstruits ne vérifient pas nécessairement les contraintes de départ ; il peut être nécessaire de faire appel aux méthodes classiques telles que, par exemple, l'algorithme RAS.

Enfin, l'efficacité des méthodes proposées ne peut être jugée qu'après de nombreuses applications ; c'est ce que nous tentons de faire actuellement avec nos modestes moyens (informatiques).

BIBLIOGRAPHIE.

[1] T. FOUCART
"Sur les suites de tableaux de contingence indexés par le temps".
Statistique et Analyse des données n° 2, 1978.

[2] H. L'HERMIER des PLANTES
"Structuration des Tableaux à Trois Indices de la Statistique : théorie et application d'une méthode d'analyse conjointe".
Thèse de 3ème cycle présentée à l'Université des Sciences et Techniques du Languedoc Montpellier ; 1976.

[3] Y. ESCOUFIER
"Opérateur associé à un tableau de données".
Annales de l'INSEE n°22-23 ; 1976.

[4] F. CAILLIEZ, J.P. PAGES
"Introduction à l'analyse des données".
SMASH, 1976.

[5] P. THIONET
"Construction et reconstruction de tableaux statistiques".
Annales de l'INSEE n°22-23 ; 1976.

[6] E. STEMMELEN
"Tableaux d'échanges : description et prévision".
Cahiers de BURO, série Recherche, n° 28 ; 1977.

[7] J.P. BENZECRI
"L'Analyse des données", tome 2 : "L'Analyse des correspondances".
Dunod, 1973.

[8] F.R. GANTMACHER
"Théorie des matrices" tome 2.
Dunod, 1966.

ANNEXE 2 : ANALYSE DES OPERATEURS DE BURT (cf. paragraphe 4.1)

MATRICE DES PRODUITS SCALAIRES

0.59414	0.56917	0.56125	0.55479	0.52963	0.52131	0.52334	0.52516	0.51855
0.56917	0.55527	0.55097	0.54559	0.52807	0.52147	0.52189	0.52363	0.51985
0.56125	0.55097	0.54855	0.54375	0.52897	0.52277	0.52266	0.52431	0.52161
0.55479	0.54559	0.54375	0.53974	0.52596	0.51988	0.51978	0.52159	0.51894
0.52963	0.52807	0.52897	0.52596	0.51988	0.51551	0.51447	0.51578	0.51550
0.52131	0.52147	0.52277	0.51988	0.51551	0.51195	0.51047	0.51168	0.51177
0.52334	0.52189	0.52266	0.51978	0.51447	0.51057	0.50993	0.51099	0.51069
0.52516	0.52363	0.52431	0.52159	0.51578	0.51168	0.51099	0.51293	0.51237
0.51855	0.51985	0.52161	0.51894	0.51550	0.51177	0.51069	0.51237	0.51357

ANALYSE DES OPERATEURS PAR STATIS

VECTEURS PROPRES (en colonnes)

0.3447	-0.6991	0.5312
0.3403	-0.3130	-0.1958
0.3395	-0.1676	-0.5041
0.3370	-0.1089	-0.4474
0.3302	0.2219	-0.1417
0.3268	0.2916	-0.0039
0.3267	0.2457	0.2881
0.3277	0.2399	0.2860
0.3265	0.3481	0.2010

VALEURS PROPRES	POURCENTAGE D'INERTIE
4.7390	98.6 %
0.0621	1.3 %
0.0022	0.0 %

ANNEXE 3 : RECONSTRUCTION DES MARGES (paragraphe 4.4)

La première ligne associée à chaque tableau correspond aux coordonnées réelles sur les axes factoriels, la deuxième aux coordonnées reconstruites à l'aide des matrices M et N du paragraphe 4.4.

	Grades		Sections	
tableau	1er axe	2ème axe	1er axe	2ème axe
69	- 0.2464	- 0.0256	0.0886	4.57 E-04
	- 0.2454	- 2.76 E-04	0.1040	3.96 E-03
70	- 0.1048	2.97 E-03	0.0646	0.0200
	- 0.1111	- 7.81 E-05	0.0471	1.80 E-03
71	- 0.0610	0.0289	0.0416	2.33 E-03
	- 0.0606	- 3.21 E-06	0.0257	9.86 E-04
73	- 0.0460	0.0258	0.0104	- 0.0102
	- 0.0401	2.65 E-05	0.0170	6.57 E-04
74	0.0729	0.0136	- 0.0313	- 0.0172
	0.0749	1.96 E-04	- 0.0317	- 1.19 E-03
75	0.1057	- 8.07 E-03	- 0,0195	- 0.0270
	0.0992	2.31 E-04	- 0.0420	- 1.58 E-03
76	0.0851	- 0.0219	- 0.0400	- 0.0183
	0.0832	2.07 E-04	- 0.0352	- 1.32 E-03
77	0.0764	- 0.0126	- 0.0493	8.40 E-03
	0.0812	2.04 E-04	- 0.0343	- 1.29 E-03
78	0.1183	- 3.07 E-03	- 0.0652	0.0415
	0.1188	2.60 E-04	- 0.0503	- 1.90 E-03

ANNEXE 4 : RECONSTRUCTION DES MARGES (paragraphe 4.5)

La première ligne associée à chaque tableau correspond aux coordonnées réelles sur les axes factoriels, la deuxième aux coordonnées reconstruites à l'aide des matrices M_1 et N_1 du paragraphe 4.5.

	Grades		Sections	
tableau	1er axe	2ème axe	1er axe	2ème axe
68	- 0.2464	- 0.0256	0.0886	4.57 E-04
	- 0.2410	- 0.0288	0.0936	4.39 E-03
69	- 0.1048	2.97 E-03	0.0646	0.0200
	- 0.1190	8.74 E-03	0.0547	1.38 E-03
70	- 0.0610	0.0289	0.0416	2.33 E-03
	- 0.0672	0.0273	0.0392	8.44 E-06
71	- 0.0460	0.0258	0.0104	- 0.0102
	- 0.0247	0.0273	0.0207	- 5.89 E-04
73	0.0729	0.0136	- 0.0313	- 0.0172
	0.0684	3.81 E-03	- 0.0282	- 1.10 E-03
74	0.1057	- 8.07 E-03	- 0.0195	- 0.0270
	0.0845	- 5.05 E-03	- 0.0384	- 1.02 E-03
75	0.0851	- 0.0219	- 0.0400	- 0.0183
	0.0862	- 0.0146	- 0.0427	- 7.16 E-04
77	0.0764	- 0.0126	- 0.0493	8.40 E-03
	0.0932	- 0.0118	- 0.0447	- 9.10 E-04
78	0.1183	- 3.08 E-03	- 0.0652	- 0.0415
	0.1196	- 6.88 E-03	- 0.0543	- 1.45 E-03

ANNEXE 1 : DESCRIPTION DES DONNEES.

Les 9 tableaux de répartition des chercheurs en Sciences Humaines au C.N.R.S. sont datés successivement du 15 juillet 1968, 15 juin 1969, 15 juin 1970, 30 juin 1971, 15 décembre 1973, 18 décembre 1974, 24 décembre 1975, 14 janvier 1977, 13 janvier 1978.

Nous donnons ci-dessous le tableau de 1978, avec les abréviations des sections et des grades :

	Directeurs de Recherche (DR)	Maîtres de Recherche (MR)	Chargés de Recherche (CR)	Attachés de Recherche (AR)	Marge
Anthropologie (an)	9	43	94	52	198 (15,7 %)
Sociologie (so)	11	25	90	56	182 (14,4 %)
Géographie (gé)	3	7	26	34	70 (5,5 %)
Sciences Economiques (sé)	4	23	40	45	112 (8,9 %)
Sciences Juridiques (sj)	7	12	33	38	90 (7,1 %)
Linguistique générale (lg)	7	11	31	44	93 (7,3 %)
Linguistique France (lf)	3	9	33	20	65 (5,1 %)
Civilisation classique (cc)	2	7	24	20	53 (4,2 %)
Civilisation orientale (co)	10	18	70	51	149 (11,8 %)
Histoire médiévale (mé)	4	13	34	32	83 (6,6 %)
Histoire moderne (mo)	2	7	20	48	77 (6,1 %)
Philosophie (ph)	6	10	39	38	93 (7,3 %)
Marge	68(5,4%)	185(14,6%)	534(42,2%)	478(37,8%)	1 265