

REVUE DE STATISTIQUE APPLIQUÉE

HABIB BENALI

BRIGITTE ESCOFIER

Analyse factorielle lissée et analyse factorielle des différences locales

Revue de statistique appliquée, tome 38, n° 2 (1990), p. 55-76

<http://www.numdam.org/item?id=RSA_1990__38_2_55_0>

© Société française de statistique, 1990, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ANALYSE FACTORIELLE LISSÉE ET ANALYSE FACTORIELLE DES DIFFÉRENCES LOCALES

Habib BENALI (1) et Brigitte ESCOPIER (2)

(1) INSERM U287 IGR Rue Camille Desmoulins 94805 Villejuif

(2) IRISA Campus de Beaulieu 35042 Rennes Cedex et IUT Kercado 56008 Vannes

RÉSUMÉ

Nous proposons ici deux méthodes qui permettent de tenir compte, dans une analyse en composantes principales, d'une structure de proximité ou de contiguïté définie sur l'ensemble des lignes d'un tableau de données. La première, l'analyse lissée permet d'analyser les tendances générales des données en éliminant les fluctuations locales. La deuxième, l'analyse des différences locales permet, au contraire, d'analyser ces fluctuations en éliminant les variations générales liées aux positions spatiales. Nous comparons cette dernière à l'analyse locale (cf. Lebart 1984) et à l'analyse en composantes principales pour des unités statistiques corrélées (cf. Caussinus 1980). Des exemples d'applications illustrent ces techniques. Ces méthodes permettent aussi d'étudier les relations entre un ensemble de variables et une structure de contiguïté.

SUMMARY

In this paper, two methods are proposed which allow to take into account a proximity or contiguity structure defined on the set of the data matrix rows in a principal component analysis. With the first method, called smooth factorial analysis, one can study general trends in the data by removing local fluctuations. On the contrary, the second method, called factorial analysis of local differences, is concerned with the analysis of these fluctuations. We compare the latter to local analysis (Lebart 1984) and to Principal Component Analysis for correlated statistical units (Caussinus 1980).

1. Les données

Nous traitons ici le cas de tableaux de données traités habituellement par l'analyse en composantes principales (A.C.P.) c'est-à-dire des tableaux croisant des individus et des variables numériques. Nous exposerons ultérieurement la généralisation des méthodes proposées à d'autres types de tableaux.

On note I l'ensemble des individus et n leur nombre. On note J l'ensemble des variables et p leur nombre. On note X le tableau de données, son terme général x_{ij} représente la valeur de la variable j pour l'individu i et x_j représente la variable j . La plupart du temps, les variables sont centrées et réduites, pour simplifier les notations, X désigne la matrice des données ainsi transformées.

Sur l'ensemble des individus existe, de plus, une structure de contiguïté. Il peut s'agir d'une proximité plane si, par exemple, les individus sont des zones géographiques, ou bien des zones d'une image. Les séries chronologiques multidimensionnelles induisent aussi des structures de proximité : chaque moment peut être relié à son (ou ses) suivant(s). Ce second type de structure se retrouve dès qu'il existe un ordre sur l'ensemble des lignes d'un tableau, ou sur des sous-ensembles de ces lignes, notamment lorsque ces lignes représentent les points de discrétisation d'une courbe. Une contiguïté peut être définie encore sur des régions d'un espace euclidien de dimension quelconque défini, par exemple, par un ensemble de variables.

Nous définissons la contiguïté sur un ensemble I par un graphe dont les sommets sont les éléments de I et les arêtes relient un sommet i à ses voisins. L'ensemble des voisins de i , qui est contenu dans I est noté $V(i)$. On note A la matrice carrée, de dimension égale au cardinal de I , associée au graphe. Son terme général, $a_{ii'}$ (positif ou nul), est le "poids" de l'élément i' dans le voisinage de i . Si i' n'est pas voisin de i , ce poids est nul. Si on veut faire jouer le même rôle à tous les voisins de i , le poids de chacun d'entre eux vaut 1, le graphe est alors un graphe non pondéré. La possibilité de pondérer le graphe assouplit la notion de contiguïté et permet d'introduire des notions de proximité en donnant un poids plus important aux éléments les plus proches. Le graphe n'est pas nécessairement symétrique. Cette définition de la contiguïté par un graphe quelconque est assez générale pour recouvrir la plupart des situations auxquelles on peut être confronté.

On note $n_i = \sum_{i'} a_{ii'}$ la somme des poids des voisins de i , n_i n'est autre que le nombre de ces voisins si le graphe n'est pas pondéré et on note T la matrice diagonale de dimension n dont le coefficient t_{ii} vaut $1/n_i$.

2. Cas d'une partition, analyse inter et analyse intra

Un cas limite, particulièrement simple, nous guidera dans la prise en compte en analyse factorielle de cette notion de contiguïté. C'est le cas d'un graphe symétrique non pondéré induit par une partition de l'ensemble des lignes, où les voisins d'un élément sont les éléments qui appartiennent à la même classe que lui. Nous étudions d'abord ce cas limite et nous le généralisons à un graphe quelconque.

Lorsque la structure de contiguïté est définie par une partition de l'ensemble de I , la matrice du graphe est composée de blocs de 1 (au croisement d'une classe avec elle-même) et de blocs de 0 (un croisement d'une classe avec une autre classe).

Dans ce cas, tenir compte de cette structure de I revient à étudier le comportement des différentes classes pour l'ensemble des variables.

Deux aspects complémentaires de cette étude sont :

- l'analyse du comportement moyen de chaque classe
- l'analyse des différences de comportement à l'intérieur d'une même classe.

La matrice de covariance (de corrélation si les variables sont réduites) se décompose en variance inter-classe (variance des moyennes des classes) et variance intra-classe (variance de chaque classe autour de sa moyenne). Étudier le comportement moyen de classes, c'est analyser la variance inter, tandis que

l'étude du comportement à l'intérieur des classes est l'analyse de la variance intra. Nous appelons respectivement "analyse inter" et "analyse intra" ces deux analyses.

2.1. Analyse inter

2.1.1. Nuage des individus

A une partition de I en K classes notées I_k correspond une décomposition du nuage des individus en K sous-nuages disjoints. Le barycentre d'une classe I_k , dont l'effectif est noté n_k a pour coordonnées :

$$y_{kj} = \sum_{i \in I_k} x_{ij} / n_k$$

Ce qui peut encore s'écrire :

$$y_{kj} = \sum_{i' \in I_k} (a_{ii'} x_{i'j} / n_k) = \sum_{i' \in V(i)} (x_{i'j} / n_k) \text{ pour tout } i \text{ de } I_k$$

Analyser la variance inter revient à placer chaque individu au barycentre de sa classe, donc à faire l'A.C.P. de ces classes en leur affectant le poids n_k ; ou bien encore à faire l'A.C.P. d'un tableau Y de même dimension que X dans lequel x_{ij} est remplacé par la valeur moyenne des voisins de i .

2.1.2. Nuage des variables

Les colonnes du tableau Y définissent de nouvelles variables. Etudions la transformation du nuage des variables lorsque l'on passe du tableau X au tableau Y .

Nous avons considéré que les variables initiales étaient centrées ; cette transformation conserve le centrage, par contre les nouvelles variables ne sont pas réduites (dans le cas extrême où une variable initiale a la même valeur moyenne pour chaque classe, la nouvelle variable a une variance nulle et doit être supprimée du tableau).

A la décomposition de l'ensemble I en K classes correspond une décomposition de l'espace R^n (muni, au coefficient $1/n$ près, de la métrique usuelle) en K sous-espaces orthogonaux notés R^{nk} :

$$I = U_k(I_k) \quad R^n = \oplus_k (R^{nk})$$

La variable indicatrice de la classe I_k qui vaut 1 sur les individus de cette classe et 0 sur les autres individus appartient à R^{nk} (c'est la "première bissectrice" de R^{nk}). Les nouvelles variables y_j sont les projections des variables x_j , sur le sous-espace E engendré par les variables indicatrices des q classes. En effet, en notant 1_k la variable indicatrice de la classe k , et x_j le vecteur représentant la

variable j , on a :

$$\langle x_j, 1_k \rangle = (1/n) \sum_{i \in I_k} x_{ij} \quad \text{et} \quad \|1_k\|^2 = n_k/n$$

D'où :

$$y_j = \sum_k \langle x_j, 1_k \rangle (1_k / \|1_k\|^2)$$

2.2. Analyse intra

2.2.1. Nuage des individus

Si l'on s'intéresse à l'analyse des différences de comportement entre les individus d'une même classe, il faut supprimer du nuage d'individus associés à X la dispersion inter-classe et ne conserver que la dispersion intra-classe. Pour cela, on translate chaque sous-nuage $N(I_k)$ pour que son centre de gravité g_k coïncide avec l'origine des axes. Un individu i de la classe I_k est représenté par la différence avec la moyenne de sa classe, point dont les coordonnées sont :

$$z_{ij} = x_{ij} - (1/n) \sum_{i' \in I_k} x_{i'j} = x_{ij} - y_{ij}$$

Dans ce nuage deux individus sont proches s'ils s'éloignent de la même façon de la moyenne de leur classe respective.

2.2.2. Nuage des variables

L'ensemble des coordonnées des points de ce nouveau nuage regroupées dans un tableau noté Z définit de nouvelles variables.

Cette transformation consiste tout simplement à centrer les variables sur chaque classe de la partition. Comme pour l'analyse inter, les variables définies dans le tableau Z sont centrées mais non réduites (si une variable initiale est constante sur chaque classe, la nouvelle variable a une variance nulle et doit être supprimée des données).

Puisque $Z = X - Y$ et que la transformation qui permet de passer des variables de X à celles de Y est une projection orthogonale sur E (i.e. le sous-espace engendré par la première bissectrice de chacun des R^{n_k}) la transformation qui permet de passer des variables de X à celles de Z est une projection orthogonale sur l'orthogonal de E . On retrouve ici une généralisation de la propriété classique de l'A.C.P. : le centrage des variables se traduit dans le nuage d'individus par un recentrage du nuage et dans le nuage des variables par une projection sur le sous-espace orthogonal à la première bissectrice.

2.3. Analyse des covariances ou des corrélations

Pour l'analyse inter, comme pour l'analyse intra, une alternative se pose : analyser les variables définies respectivement dans les tableaux Y et Z sans les réduire ou en les réduisant. Les deux solutions sont possibles et se justifient suivant les cas. Les analyses exactes des variances inter et intra correspondent à la première solution. Dans ce cas les résultats des analyses de X, Y, Z sont directement comparables, notamment les valeurs propres, puisque l'inertie des nuages définis par X est la somme des inerties des nuages définis par Y et Z .

Si les variables sont réduites, l'influence de chacune des variables dans l'analyse est équilibré, ce qui est souvent intéressant, mais les analyses de X, Y, Z concernent des objets différents. Par exemple, une variable dont les valeurs sont très proches sur les barycentres des classes (resp. à l'intérieur d'une même classe) aura une importance artificiellement grande dans l'analyse réduite de Y (resp. Z).

3. Généralisation à un graphe quelconque

Dans le cas d'un graphe quelconque définissant une structure de proximité sur les éléments de I , on peut souhaiter éliminer les variations locales pour analyser plus facilement les tendances générales. Dans ce cas, nous proposons l'analyse lissée qui est une généralisation de l'analyse inter. A l'opposé, on peut souhaiter analyser uniquement les variations locales. Nous proposons alors l'analyse des différences locales qui généralise l'analyse intra.

3.1. Analyse factorielle lissée (A.F.L.)

Le terme de lissage est emprunté aux séries chronologiques. Le but principal poursuivi en analyse factorielle lissée est de dégager les grandes tendances, en éliminant l'influence des fluctuations locales. Son principe est très simple, il consiste à remplacer chaque point i représentant une ligne du tableau par le centre de gravité (pondéré par les $a_{ii'}$) de ses voisins. Il est clair que ce lissage diminue l'influence des différences locales, il peut servir à améliorer la fiabilité dans deux circonstances au moins : lorsque les individus ne sont pas définis a priori mais correspondent à un découpage arbitraire (ce qui peut se produire pour des zones géographique par exemple); lorsque les mesures sont entachées d'erreurs et que ces mesures ont des valeurs proches pour des individus voisins au sens du graphe. Cette analyse peut aussi être utilisée avant une étape de classification sur facteurs pour "forcer" la classification et construire des classes plus contigües que celles auxquelles aboutirait une analyse classique.

3.1.1 Nuage des individus

Chaque point i du nuage $N(I)$ défini dans l'A.C.P. de X est remplacé par le barycentre de ses voisins. On étudie alors la dispersion entre ces nouveaux points.

Les coordonnées du point représentant l'individu i s'écrivent :

$$y_{ij} = \sum_{i' \in I} (a_{ii'} x_{i'j} / n_i)$$

Dans le cas d'un graphe non pondéré, cette expression se réduit à :

$$y_{ij} = \sum_{i' \in V(i)} (x_{i'j} / n_i)$$

L'ensemble de ces coordonnées définit un tableau Y qui s'écrit matriciellement :

$$Y = TAX$$

Généralement le nuage d'individus n'est pas centré. En effet, la moyenne d'une colonne j s'écrit :

$$(1/n) \sum_{i \in I} y_{ij} = (1/n) \sum_{i, i'} a_{ii'} (x_{i'j} / n_i) = (1/n) \sum_{i'} \left(\sum_i (a_{ii'} / n_i) \right) x_{i'j}$$

Cette moyenne est nulle (comme celle de la variable x_j) lorsque le coefficient affecté à $x_{i'j}$ est indépendant de i' ce qui se produit notamment dans le cas d'un graphe de partition et dans le cas d'un graphe symétrique ($a_{ii'} = a_{i'i}$) où n_i est constant. Dans les autres cas, la moyenne de y_j peut être différente de 0. Dans beaucoup d'applications pratiques, le nombre et les poids des voisins des éléments varient peu d'un élément à l'autre et cette moyenne est peu différente de 0.

Pour avoir une bonne représentation de ce nuage d'individus, il faut centrer le nuage et donc appliquer une A.C.P. centrée au tableau Y .

Dans le cas d'un graphe de partition, on retrouve exactement l'analyse inter. Si le graphe est un graphe complet ($a_{ii'} = 1$ pour tout i et tout i'), tous les points du nuage sont confondus avec le barycentre tandis que dans le cas d'un graphe réduit à l'identité, l'analyse lissée se confond avec l'A.C.P. de X .

3.1.2. Nuage des variables

Les colonnes du tableau Y définissent de nouvelles variables, que nous appellons "variables lissées". Dans le cas de l'analyse inter, à la transformation du nuage d'individus (remplacement de tous les individus d'une classe par le centre de gravité de cette classe) correspond une transformation duale du nuage des variables (projection sur un sous-espace). En dehors de ce cas, les variables lissées ne sont pas des projections orthogonales sur un sous-espace. Cependant, dans le cas où le graphe de contiguïté n'est pas pondéré, les variables lissées se déduisent des variables initiales par une transformation non linéaire appelée "projection sur un graphe". Notons 1_i , la variable indicatrice des voisins de l'élément i , alors :

$$\langle x_j, 1_i \rangle = (1/n) \sum_{i' \in V(i)} (x_{i'j})$$

$$\|1_i\|^2 = n_i/n$$

D'où :

$$y_{ij} = \langle x_j, 1_i \rangle / \|1_i\|^2$$

3.1.3. Projection des individus et des variables du tableau initial

Comme dans toute A.C.P., il est possible d'introduire des individus et des variables en éléments supplémentaires. Il est particulièrement intéressant d'introduire dans l'analyse lissée les éléments du tableau initial X . Ces projections peuvent même être, dans certains cas, les résultats essentiels de l'analyse lissée.

En effet, on souhaite souvent garder exactement les individus initiaux pour lesquels on a la plupart du temps des informations extérieures aux données traitées, tout en les représentant sur des axes de dispersion générale moins dépendants des variations locales que les axes issus de l'A.C.P. de X . La projection en supplémentaires des individus initiaux dans l'analyse lissée répond exactement à ce problème. D'autre part, la comparaison des projections des individus d'origine et des individus lissés permet d'étudier, dans une certaine mesure, l'effet du lissage.

La projection des variables initiales permet de visualiser leur position et de les comparer aux variables d'origine. Ceci permet, dans une certaine mesure, d'étudier les relations entre la structure de proximité et le graphe : une variable très liée au graphe, i.e. dont les valeurs varient assez peu localement est peu modifiée par le lissage ce qui n'est pas le cas pour les variables non liées au graphe. Dans la plupart des applications pratiques les deux projections d'une même variable sont assez proches car, d'une part, les lissages choisis modifient souvent assez peu les variables et, d'autre part, les différences entre variables initiales et variables projetées ne sont pas forcément bien mises en évidence sur les axes d'inertie des variables lissées. Cette situation est particulièrement nette dans le cas d'un graphe de partition (analyse inter) où les projections des variables lissées et des variables initiales se confondent sur tous les axes puisque le nuage des variables lissées est une projection du nuage d'origine.

3.2. Analyse factorielle des différences locales (A.F.D.L.)

Le problème étudié est l'opposé du précédent. On cherche à analyser les différences locales, en se dégageant des tendances générales liées à la contiguïté, (comme par exemple un gradient nord-sud) qui peuvent déterminer les premières composantes principales dans une A.C.P. classique. Le principe de l'analyse des différences locales est exactement le complémentaire de celui de l'analyse lissée. Il consiste à remplacer chaque point représentant une ligne du tableau X par ses différences avec le barycentre de ses voisins. Une analyse des différences locales permet notamment de mettre en évidence des "anomalies" régionales, des accidents dans des séries chronologiques. Elle permet aussi de regrouper des lignes qui diffèrent entre elles, mais dont le point commun est de s'écarter de la même manière de leur voisinage respectif. Une telle analyse peut, comme l'analyse lissée, être utilisée comme première étape d'une classification basée sur les différences locales.

Des solutions différentes sont proposées dans [5] et [2] pour résoudre des problèmes assez proches.

3.2.1. Nuage des individus

Chaque individu i est représenté par un point qui représente sa différence avec le barycentre de ses voisins. Ses coordonnées s'écrivent :

$$z_{ij} = x_{ij} - y_{ij}$$

L'ensemble de ces coordonnées définit un tableau Z qui s'écrit matriciellement

$$Z = X - Y = X - TAX$$

Deux individus sont proches s'ils s'éloignent de la même façon de leurs voisins (dans le nuage d'origine ces mêmes points peuvent être très éloignés). Il résulte de la relation entre les tableaux Y et Z , que ce nuage n'est centré que lorsque le nuage d'individus défini dans l'analyse lissée l'est (graphe de partition ou somme des poids des voisins constants dans un graphe symétrique). Pour avoir une bonne représentation de ce nuage d'individus, il faut le centrer et donc appliquer une A.C.P. centrée au tableau Z .

Dans le cas d'un graphe de partition, on retrouve exactement l'analyse intra. Dans le cas d'un graphe complet, le nuage n'est pas modifié, l'analyse des différences locales se confond avec l'A.C.P. de X tandis que dans le cas d'un graphe réduit à l'identité tous les points du nuage sont confondus.

3.2.2. Nuage des variables

Les colonnes du tableau Z définissent de nouvelles variables que nous appelons différences locales. Notons que ces variables ne se déduisent des variables initiales par une projection sur un sous-espace de R^n que dans le cas d'un graphe de partition. La norme de ces variables est d'autant plus faible que les variables sont plus liées au graphe. Cette analyse mettra donc en évidence les variables peu liées à la structure de contiguïté.

3.2.3. Projection des variables et des individus du tableau initial

Des individus et des variables peuvent être introduits en éléments supplémentaires, mais les projections du tableau initial X ont beaucoup moins d'intérêt que dans l'analyse lissée.

Si dans l'analyse lissée, la projection des individus initiaux est souvent un des résultats essentiels, ceci est rarement le cas en analyse des différences locales. En effet, si "l'individu lissé" a souvent peu de signification concrète et joue plutôt le rôle d'intermédiaire de calcul, "l'individu local" qui traduit l'écart au voisinage garde pleinement son identité propre et s'interprète facilement : la proximité d'individus dans l'analyse locale traduit le même type "d'anomalies" par rapport au voisinage, anomalies que l'on peut relier à des connaissances extérieures.

La position des individus initiaux, qui peuvent être très éloignés des individus locaux n'est guère utile, ni pour l'interprétation d'axes qu'ils ne déterminent pas, ni pour leur étude propre. D'autre part, si la comparaison des positions des individus lissés et des individus initiaux (qui se réfèrent toutes deux à des notions globales) a un sens, celle des positions des individus locaux et des individus initiaux qui se réfèrent à des notions très différentes est beaucoup plus délicate.

La plupart du temps, les projections des variables initiales (très différentes des variables de différences locales) ne présentent guère d'intérêt. Comme pour l'analyse lissée, dans le cas d'un graphe de partition, ces projections sont confondues avec celles des variables des différences locales puisque le nuage analysé est une projection du nuage d'origine sur un sous-espace.

3.3. Analyse des covariances ou des corrélations

L'analyse lissée et l'analyse des différences locales aboutissent à des A.C.P., l'une du tableau Y et l'autre du tableau $Z = X - Y$. On peut analyser ces deux tableaux, soit en réduisant les variables, soit en ne les réduisant pas.

En dehors du cas de la partition, l'inertie des nuages définis par X n'est pas la somme des inerties des nuages définis par Y et Z , les barycentres des nuages d'individus associés à X et à Y ne sont d'ailleurs même pas toujours confondus. L'analyse lissée et l'analyse des différences locales ne sont donc pas, dans le cas général, les analyses de deux parties exactement complémentaires de l'inertie de X . De ce fait, les résultats de ces différentes analyses, et notamment les valeurs propres, sont à comparer avec précaution, même si les A.C.P. de Y et de Z sont des A.C.P. non normées. Notons que pour mettre en évidence les variables peu liées au graphe, il est préférable de ne pas normer les variables en A.F.D.L.

4. Quelques graphes particuliers

4.1. Cas d'un graphe pondéré

Si le graphe n'est pas pondéré, tous les voisins d'un individu (y compris l'individu lui-même) ont la même influence dans le lissage. Si un individu a beaucoup de voisins, ses caractéristiques peuvent être "couvertes" par celles de ses voisins, ce qui peut être un inconvénient. D'autre part, dans le cas d'une proximité géographique (ou temporelle) par exemple, on peut souhaiter que des zones interviennent dans le lissage avec une importance qui décroît avec la distance. Pour résoudre ces problèmes, un poids différent de 1 peut être attribué à chaque arête (i, i') du graphe. Ce poids est fixé par l'utilisateur en fonction du problème traité.

4.2. Cas d'un graphe non symétrique (orienté)

Un graphe orienté exprime une hypothèse de liaisons non symétrique entre l'ensemble des individus. Donnons deux exemples où il est utile d'introduire un

tel graphe :

– hypothèse Markovienne d'ordre r sur des données évolutives, le graphe associé est de la forme suivante

$$a_{(i,t)(i,t')} = 1 \text{ si } 0 \leq t - t' \leq r \\ = 0 \text{ sinon}$$

– hypothèse sur le sens de la pollution due à un département sur l'écologie des départements voisins, ce graphe est de la forme

$$a_{ii'} = 1 \text{ si } i \text{ pollue } i' \\ = 0 \text{ sinon}$$

4.3. Cas d'un graphe issu de variables instrumentales

La méthodologie présentée précédemment s'applique quand le graphe est construit de façon artificielle. Le graphe peut être déduit de distances ou de similarités sur l'ensemble des individus définis par un groupe de variables "instrumentales". L'analyse factorielle lissée renforce les liaisons qui passent par ces variables, tandis que l'analyse des différences locales neutralise leurs effets. Ces analyses ont à la base les mêmes idées introduites par Rao [7] sur les variables instrumentales et développées depuis dans d'autres contextes. La thèse de R. Sabatier [9] contient une présentation très complète et de nombreuses références à ce sujet. Notons simplement ici que l'analyse sur variables instrumentales (A.C.P.V.I.) classique peut être considérée comme une A.C.P. du tableau déduit de X en projetant les variables sur le sous-espace engendré par les variables instrumentales ou son orthogonal (A.C.P.V.I. orthogonale selon Sabatier). L'introduction de ces variables à travers un graphe est différente, elle est non linéaire et moins stricte : au lieu d'imposer que les composantes principales soient des combinaisons linéaires des variables instrumentales (A.C.P.V.I.), on cherche simplement dans l'analyse lissée des composantes qui tiennent compte de ces variables en neutralisant les dispersions entre éléments proches du point de vue de ces dernières.

5. Comparaison de l'analyse des différences locales avec d'autres techniques

Plusieurs analyses factorielles ont été proposées dans le but d'analyser des liaisons locales entre les variables. La plus connue, dont les programmes ont été implantés dans la bibliothèque Spadn [8], est l'analyse locale introduite par L. Lebart [4] et Aluja [5]. L'analyse en composantes principales des unités statistiques corrélées a été introduite par Aragon et Caussinus [3] selon un point de vue différent. Ces techniques peuvent être regroupées dans un contexte plus général, celui où on introduit des poids éventuellement nuls ou négatifs aux couples d'individus. Nous allons comparer ces techniques à l'analyse des différences locales au niveau de la matrice de covariances diagonalisée, de la représentation des

individus et à celui de la représentation des variables. Nous concluons par la mise en oeuvre et l'interprétation des résultats.

5.1. Les matrices de covariances

L'analyse locale est introduite à partir de la matrice de covariance. Rappelons que le terme général $V_{jj'}$ de cette matrice $V = (1/n)X'X$ s'exprime comme une somme, soit sur l'ensemble des individus (1), soit sur celui des couples d'individus (2) :

$$(1) \quad V_{jj'} = (1/n) \sum_i x_{ij} x_{ij'} \quad \text{car les variables sont centrées}$$

$$(2) \quad V_{jj'} = (1/2n^2) \sum_{i, i'} (x_{ij} - x_{i'j})(x_{ij'} - x_{i'j'})$$

En analyse locale, la contiguïté est définie par un graphe symétrique non pondéré. La variance locale L est définie à partir de la deuxième expression en supprimant de la somme les couples qui ne sont pas reliés par une arête :

$$L_{jj'} = (1/m) \sum_i \sum_{i' \in V(i)} (x_{ij} - x_{i'j})(x_{ij'} - x_{i'j'})$$

$$= (1/m) \sum_i \sum_{i'} a_{ii'} (x_{ij} - x_{i'j})(x_{ij'} - x_{i'j'})$$

où m est le double du nombre d'arêtes.

La matrice L qui est diagonalisée peut s'écrire sous la forme suivante :

$$L = (1/m)X'(T^{-1} - A)X$$

Dans l'analyse des différences locales, la matrice de covariance W est définie par les nouvelles variables. Elle se déduit de la première expression de V en remplaçant les moyennes générales des variables par les moyennes locales y_{ij} :

$$W_{jj'} = (1/n) \sum_i (x_{ij} - y_{ij})(x_{ij'} - y_{ij'}) - y_{jj'}$$

Le dernier terme, le produit des moyennes, découle du centrage.

Dans les deux cas, seules les différences locales (entre les points reliés par le graphe) augmentent les covariances : la covariance entre j et j' est positive si j et j' varient dans le même sens, par rapport à leurs voisins. Mais ces matrices, et donc les analyses, sont différentes.

Cas d'un graphe de partition

Nous avons vu que dans le cas d'un graphe de partition, la variance $W_{jj'}$ se confond avec la variance intra.

Calculons la variance locale L pour une partition :

$$L_{jj'} = (1/m) \sum_k \sum_{i, i' \in Ik} (x_{ij} - x_{i'j})(x_{ij'} - x_{i'j'})$$

qui n'est autre que :

$$L_{jj'} = (2/m) \sum_k \sum_{i \in Ik} n_k (x_{ij} - y_{kj})(x_{ij'} - y_{kj'})$$

La variance locale ne se confond avec la variance intra que dans le cas où toutes les classes ont le même nombre d'éléments puisque chaque individu a un poids proportionnel au nombre d'éléments de sa classe : les individus des classes d'effectif fort interviennent avec une importance plus grande que ceux des classes d'effectif faible. Il en est de même dans un graphe quelconque. Dans la variance locale L , chaque individu intervient avec un poids égal au nombre de ses voisins, les individus ayant un voisinage riche pèseront donc plus que ceux qui ont un voisinage pauvre ; alors que dans W tous les individus gardent le même poids.

Graphe pondéré et non symétrique

Bien que les programmes proposés ne le permettent pas actuellement, l'analyse locale peut être généralisée sans difficulté particulière à un graphe pondéré, comme le propose LeFoll [5] en mettant dans l'expression (2) de la variance, des poids aux arêtes.

Covariance et corrélations

Dans les deux cas, il est possible de considérer une matrice de corrélation au lieu d'une matrice de covariance.

5.2. Représentation des individus

Dans l'analyse locale, les individus du tableau initial X n'apparaissent plus. Mais on voit facilement que l'analyse locale est équivalente à une A.C.P. d'un tableau B dont les lignes sont les couples d'individus reliés par une arête du graphe ; la valeur de la variable j pour le couple (i, i') est la différence $x_{ij} - x_{i'j}$.

Il est tout à fait possible de proposer une représentation des individus dans cette analyse en mettant le tableau X en supplémentaire dans l'A.C.P. de B . Ceci revient à projeter le nuage d'origine sur des axes différents des axes d'inertie (les axes de projection maximisent la somme des carrés des distances entre les couples de points reliés par une arête du graphe). C'est d'ailleurs ainsi que l'A.C.P. sur des

unités statistiques corrélées [6] est présentée : les auteurs proposent d'affecter des poids aux couples de points individus et de chercher des axes qui maximisent la somme ainsi pondérée des carrés des distances des couples de points. Le principe général de cette analyse est donc équivalent à celui de l'analyse locale.

Dans l'analyse des différences locales, les individus "locaux" sont représentés directement et on dispose non seulement de leur qualité de représentation, mais aussi de leur contribution à l'inertie, qui est un indice d'aide à l'interprétation essentiel. La représentation des individus que nous proposons ici s'apparente beaucoup plus à celle obtenue en mettant le tableau X en individus supplémentaires dans l'A.C.P. de Z . Nous avons souligné au paragraphe 3.2.3. l'intérêt limité de ces représentations.

5.3. Représentation des variables

L'analyse locale étant équivalente à une A.C.P. d'un tableau B dont les lignes sont les arêtes du graphe, on peut définir des variables "locales" dont le domaine de définition est l'ensemble des arêtes et non l'ensemble des individus. Ces variables peuvent être représentées dans l'analyse, mais leur domaine de définition n'étant pas l'ensemble classique, l'interprétation est plus délicate que celle des variables de l'analyse des différences locales.

Pour comparer ces variables "locales" aux variables initiales, il faut que les domaines de définition soient identiques. Le tableau B peut être complété par des lignes nulles pour les couples qui ne sont pas reliés par une arête; l'ensemble de définition est alors l'ensemble des couples d'individus. Les variables initiales sont représentées par des variables "générales" définies par le graphe complet. Notons que les variables locales sont des projections orthogonales des variables générales sur le sous-espace engendré par les arêtes du graphe et que dans l'analyse locale les deux types de variables ont leur projections confondues. Pour les comparer les auteurs proposent de projeter en élément supplémentaires les variables locales dans l'analyse des variables générales. Notons que les variables de différences locales (tableau Z) et les variables lissées (tableau Y) peuvent être introduites dans l'A.C.P. du tableau initial.

5.4. A.C.P. des variables avec métrique quelconque

L'analyse locale et l'analyse des différences locales peuvent toutes deux être définies comme une A.C.P. du tableau initial X à condition de considérer encore comme une A.C.P. une analyse du nuage des variables dans laquelle l'espace de représentation est muni d'une métrique (ou d'une semi-métrique) quelconque. Les facteurs sur les variables sont alors vecteurs propres de la matrice $X'QX$ où Q désigne la matrice associée à la métrique.

L'analyse des différences locales étant une A.C.P. classique du tableau $Z = (Id - TA)X$, ses facteurs sont vecteurs propres de la matrice $Z'Z = X'(Id - TA)'(Id - TA)X$ et sont donc facteurs de l'A.C.P. de X avec $Q = (Id - TA)'(Id - TA)$.

Pour l'analyse locale, la métrique Q est définie par $(T^{-1} - A)$ comme on

peut le voir dans l'expression de la matrice des covariances locales donnée dans le paragraphe 5.1.

5.5. Mise en oeuvre et interprétation des résultats

La mise en oeuvre de l'analyse des différences locales consiste à appliquer un programme classique d'A.C.P. à un tableau obtenu par une transformation simple des données tandis que l'analyse locale nécessite un programme spécifique. Dans les deux cas, il est bien entendu nécessaire d'introduire le graphe de contiguïté. L'interprétation des résultats de l'analyse des différences locales est absolument analogue à celle d'une A.C.P. classique et, contrairement à l'analyse locale, peut s'appuyer sur les indices de qualité de représentation et de contribution à l'inertie tant des variables que des individus ainsi que sur la dualité entre les variables et les individus.

6. Exemples d'application

Nous illustrons nos propos sur deux exemples issus des mêmes données. On dispose pour 89 départements français des valeurs des 11 productions principales agricoles mesurées en francs constants, ceci pour chaque année de 1959 à 1985. Ces données nous ont été communiquées par Mr. Fouet de l'INRA de Rennes. Cinq productions sont d'origine végétale : les céréales, les pommes de terre, les légumes, les fruits et le vin. Six productions sont d'origine animale : le porc, le veau, les gros bovins, les ovins-caprins, le lait et les volailles.

6.1. Contiguïté géographique

Dans le premier exemple, nous étudions cette production pour l'année 1985 uniquement. Le graphe considéré est celui des contiguïtés géographiques : deux départements sont contigus s'ils ont une frontière commune. Ce graphe est symétrique et non pondéré.

La méthodologie est la suivante :

a) A.C.P. normée sur le tableau des données initiales (tableau X) en mettant en variables supplémentaires les variables lissées. Analyse lissée (A.C.P. normée du tableau Y) et analyse des différences locales (A.C.P. normée du tableau Z) en mettant dans les deux cas les variables initiales en supplémentaires.

b) Classification ascendante hiérarchique, avec le critère du moment centré d'ordre deux sur les 8 premières composantes de chacune de ces trois analyses.

c) Partitions en six classes issues de chacune de ces hiérarchies et représentation de ces classes sur la carte de France.

6.1.1 Analyse du tableau initial

Les taux d'inertie extraits par les trois premières composantes principales sont respectivement 32.9 pour le premier axe, 18.16 pour le deuxième et 12.45

pour le troisième. Nous commentons le premier plan factoriel représenté dans la figure 1.

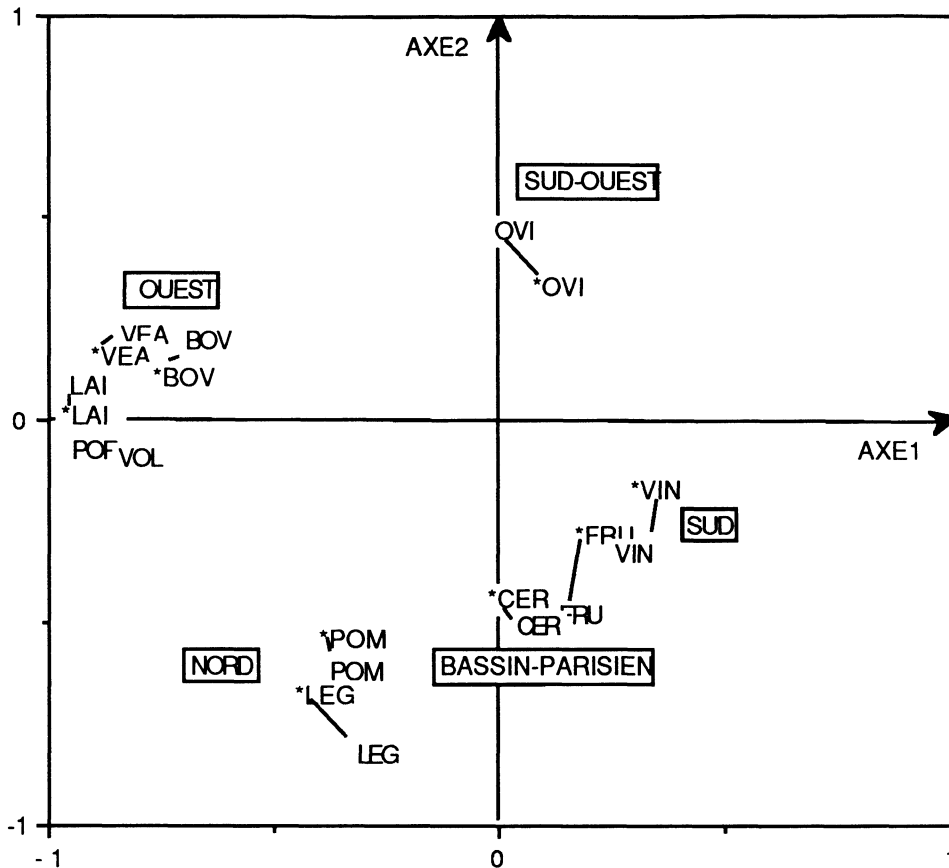


FIGURE 1

Plan 1-2 de l'A.C.P. * = variable lissée

Les variables (les productions)

Il est intéressant de noter les positions relatives des différentes variables sur ce plan et plus particulièrement de comparer les variables d'origine et les variables lissées. Une production ayant une faible variation d'un département à ses voisins, sera représentée par deux points proches tandis que la variable lissée d'une production ayant de fortes variations locales s'éloignera de la variable d'origine. On remarque ici la forte corrélation entre les variables initiales et les variables lissées. Certaines, comme le porc et les volailles sont même confondues sur le plan. Notons aussi que ces dernières qui n'ont pas déterminé les axes sont moins bien représentées.

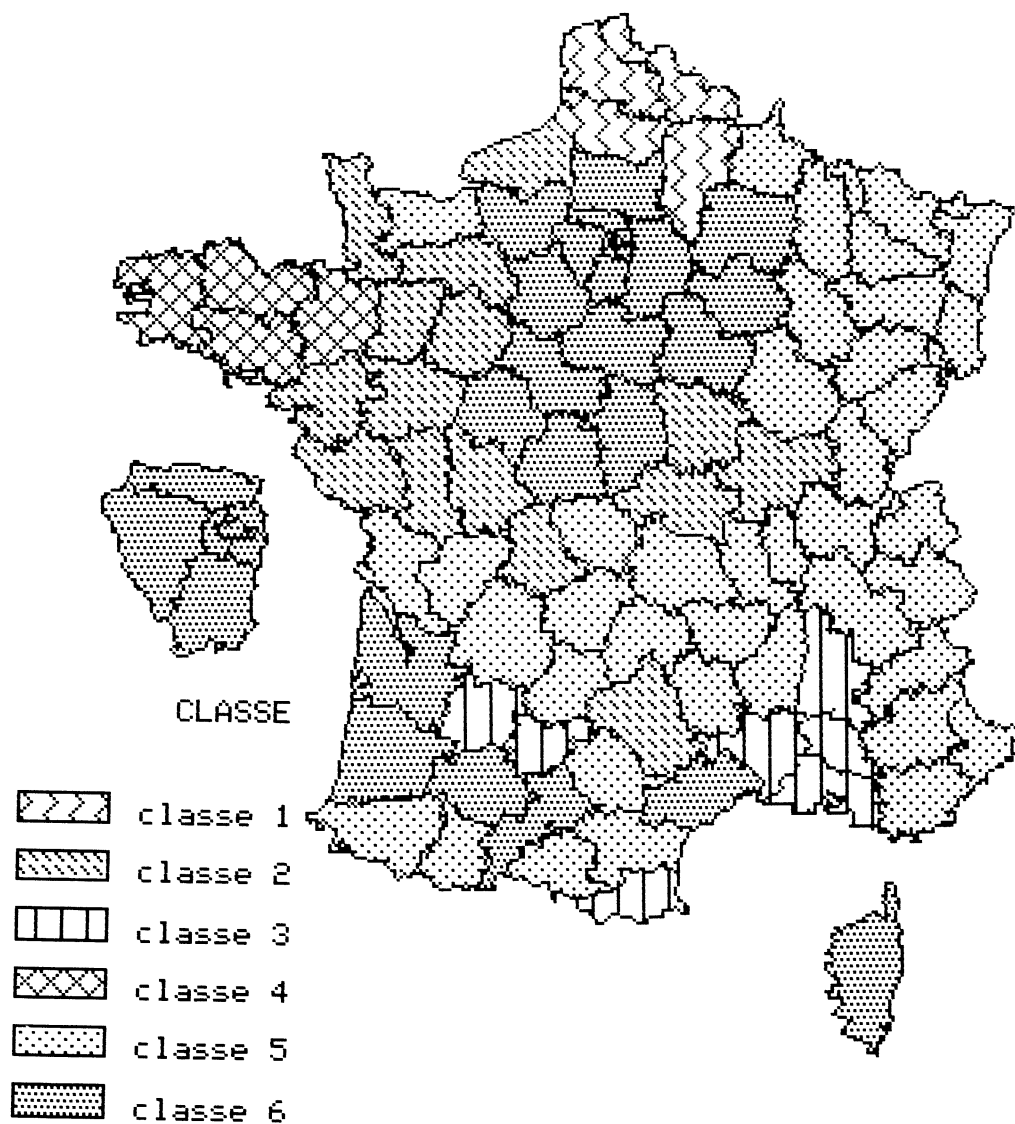


FIGURE 2
Partition issue de l'A.C.P.

Les individus (les départements)

La plupart des départements se regroupent par grandes régions dans lesquelles les productions sont analogues. Ce regroupement géographique explique la forte corrélation entre les variables initiales et lissées. Seules les grandes régions regroupant des départements contigus ont été indiquées sur la figure 1. Une nette opposition entre les départements de l'ouest caractérisés par une forte production animale (porc, volailles-oeuf, lait, gros bovins et veau) et une faible production de fruits, légumes et vin. Ces dernières productions caractérisent plutôt les départements du sud. Les départements du sud-ouest se distinguent par la production d'ovins-caprins, et ceux du nord par une production de légumes et pommes de terre.

Classification ascendante hiérarchique et partition

Nous avons effectué une C.A.H. à partir des huit coordonnées factorielles totalisant une inertie de 95.8 %. Une forte variation de l'indice de la hiérarchie conduit à une partition en six classes. La figure 2 qui représente ces six classes sur la carte montre des regroupements régionaux importants. Cependant, sur les 6 classes, seules 2 petites classes de 4 éléments sont connexes. Les autres classes sont composées de 3 ou 4 composantes connexes. Nous illustrons ces classes par les productions les plus caractéristiques :

classe 1 (4 éléments) Pommes de terre, légumes, céréales

classe 2 (15 éléments) Bovins, Ovins, Lait

classe 3 (7 éléments) Fruits, Légumes

classe 4 (4 éléments) Porc, Volaille, Veaux

classe 5 (41 éléments) Très faible production agricole

classe 6 (18 éléments) Céréales

6.1.2 Analyse lissée

Les taux d'inertie des trois premiers facteurs de l'analyse lissée valent respectivement 40.5, 19.6 et 15.0. Ils sont supérieurs à ceux de l'analyse des variables initiales : le lissage en supprimant les irrégularités renforce l'importance relative des tendances générales. Sur le plan 1-2 de la figure 3 on retrouve l'association entre la plupart des régions et leur production prépondérante. Cependant le plan est assez différent du précédent. Ceci s'explique car le lissage non pondéré effectué est assez brutal puisqu'il ne donne pas plus d'importance à un département qu'à chacun de ses voisins. Le poids peut bien entendu être modulé. Dans le choix de la pondération se pose aussi le problème des départements très entourés et des départements extrêmes comme le Finistère qui n'ont que 2 voisins et qui ont été moins modifiés que les premiers dans le lissage effectué.

Classification et partition

Pour comparer les résultats obtenus à partir de l'analyse lissée à ceux de l'analyse du tableau initial, nous avons effectué une classification sur les 8 pre-

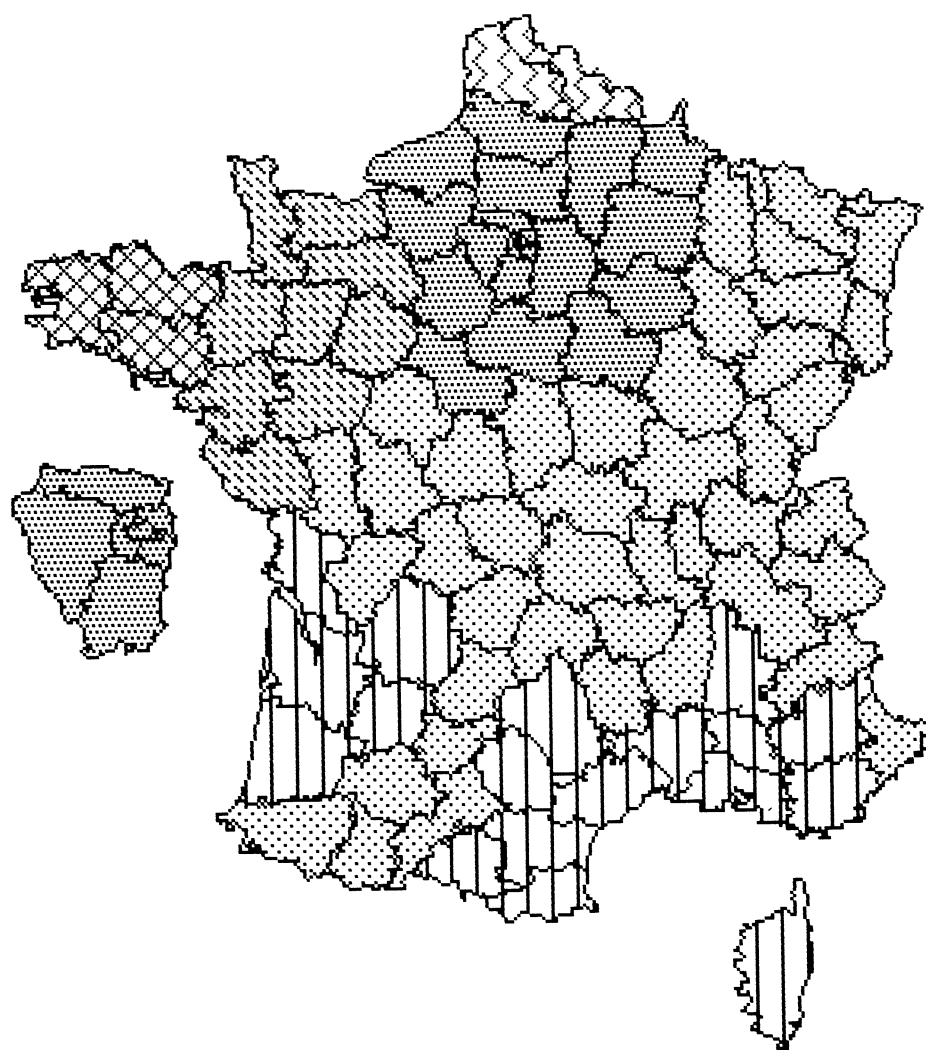


FIGURE 4
Partition issue de l'analyse lissée

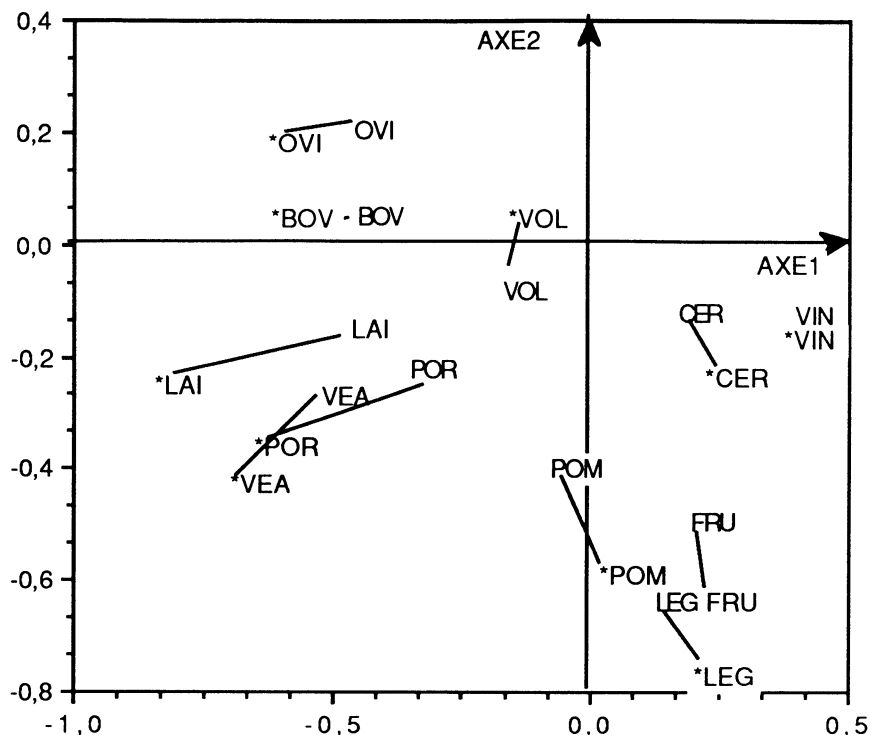


FIGURE 5
Plan 1-2 de l'A.F.D.L. * = variable résidu

des départements ayant une faible production par rapport à leurs voisins. Parmi les départements de l'est on trouve le Jura, la Savoie et les Hautes-Alpes.

6.2. Contiguïté temporelle

Dans cet exemple dont les résultats ne sont pas détaillés ici on étudie l'évolution de la production agricole entre 1959 et 1985. On dispose de données cubiques X_{IJT} où I représente l'ensemble des départements, J l'ensemble des productions et T les différentes années. Dans le tableau analysé les lignes représentent les départements par année. L'objectif est de détecter des périodes "types" de production, de voir les départements dont les productions sont restées stables dans le temps, ceux qui ont changé ou diversifié leur production etc. Le graphe que nous considérons dans cette étude est un graphe temporel qui relie les lignes représentant le même département aux temps $t-1$, t et $t+1$. Ce graphe permet de lisser les trajectoires des départements et l'analyse lissée a permis de réduire les "biais" dus à des accidents climatiques ou autres.

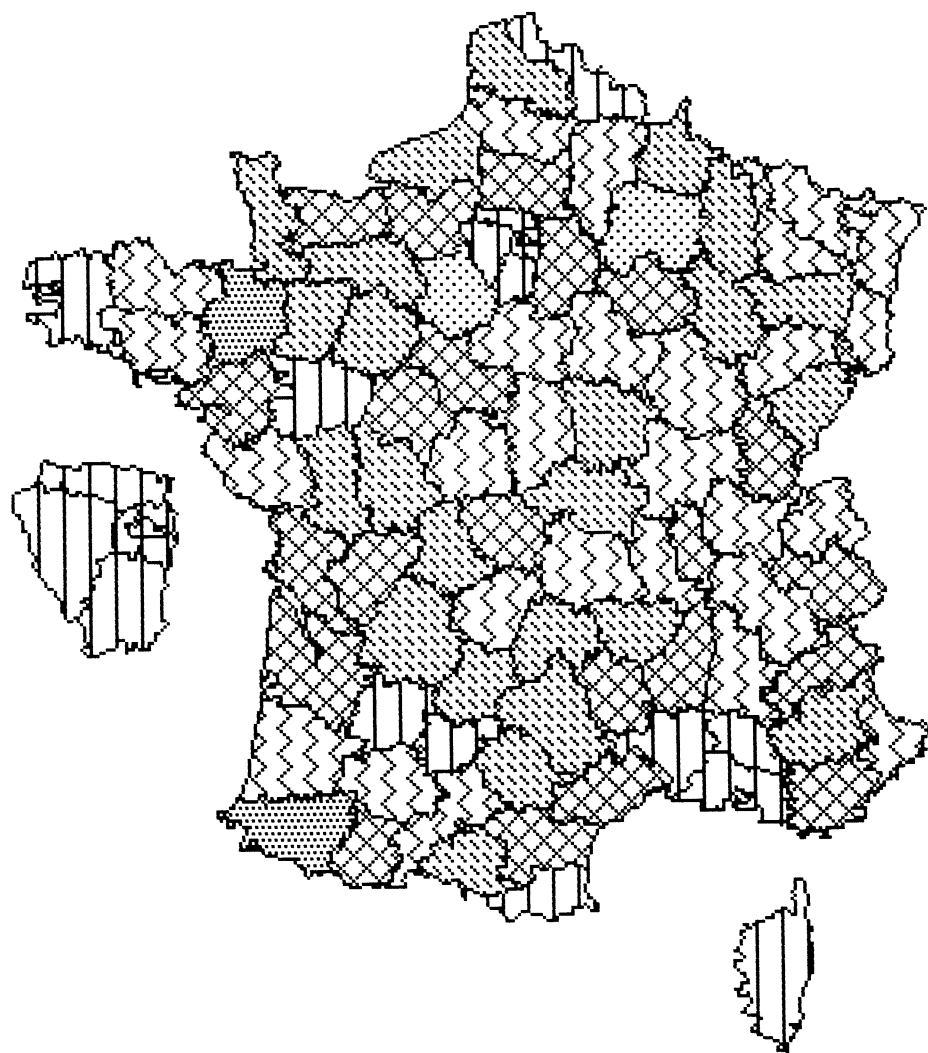


FIGURE 6
Partition issue de l'analyse des différences locales

7. Conclusion

L'introduction de la notion de proximité par un graphe pondéré, éventuellement non symétrique, est très souple et permet de traduire la plupart des situations concrètes. Les deux méthodes proposées, l'analyse lissée et l'analyse des différences locales sont faciles à appliquer car elles ne nécessitent qu'une transformation simple du tableau des données avant une A.C.P. classique. L'interprétation des résultats est absolument analogue à celle d'une A.C.P. classique. La projection en éléments supplémentaires des individus et des variables du tableau initial enrichit l'analyse et permet de mesurer l'influence des transformations du tableau de données. Dans le cas particulier d'une proximité induite par une partition de l'ensemble des individus en classes, l'analyse lissée se confond avec l'analyse interclasses tandis que l'analyse des différences locales se confond avec l'analyse intraclasse.

Bibliographie

- [1] BENALI H. et ESCOFIER B. 1988 Smooth factorial analysis and factorial analysis of local differences. Analysis of Multiway Data Matrices Edit. R.Coppi and S.Bolasco Amsterdam
- [2] CARLIER A. 1985 Applications de l'analyse factorielle de l'évolution et de l'analyse intrapériode. Statistique et Analyse des données Vol.10
- [3] CAUSSINUS H. et ARAGON Y. 1980 Une analyse en composantes principales pour des unités statistiques corrélées. Data analysis and Informatics North Holland
- [4] LEBART L. 1984 Correspondance Analysis of graph structure. Bulletin technique du centre de statistique et informatique appliquées. Vol. 2 p.5-19
- [5] ALUJA T. et LEBART L. 1985 Factorial analysis upon graph Bulletin technique du centre de statistique et informatique appliquées. Vol. 3 p.4-34
- [6] FOLL Y. 1972 L'analyse des évolutions. Actes du petit séminaire sur l'analyse Factorielle Paris
- [7] RAO C.R. 1964 The use and interprétation of Principal Component Analysis in applied research. Sankhya A 26,4 p.329-358
- [8] SPADN 1989 Manuel de référence CISIA
- [9] SABATIER R. 1987 Méthodes factorielles en analyse des données. Approximation et prise en compte de variables concomitantes. Thèse. Université de Montpellier
- [10] ESCOFIER B. et PAGES J. Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation. Dunod 1988