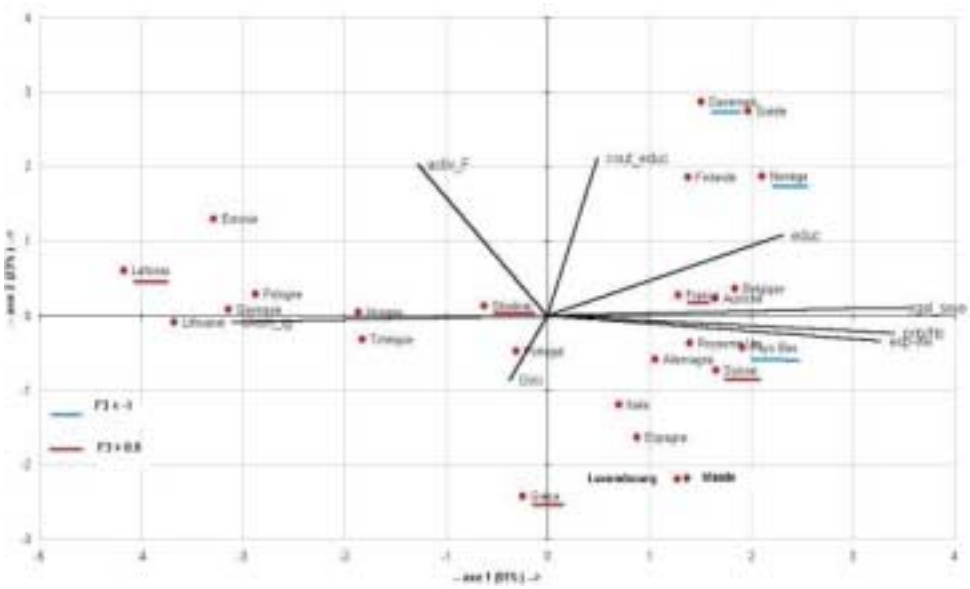
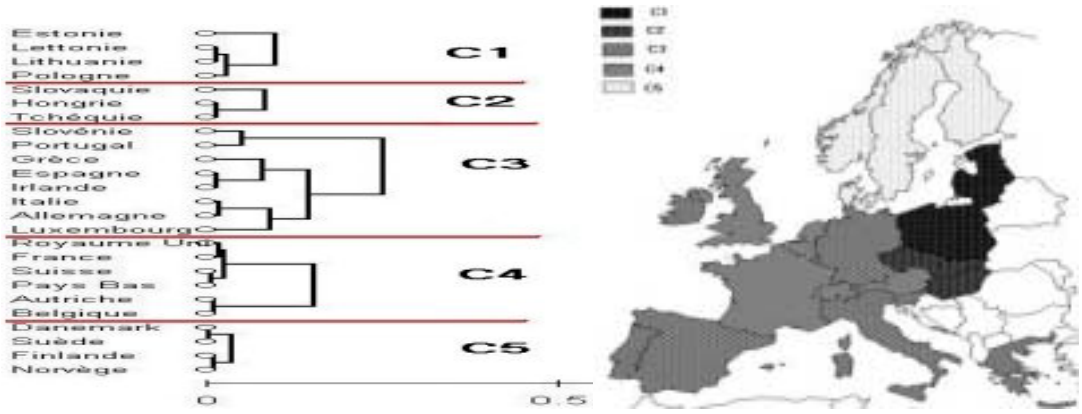
Analyse multivariée



de données



géographiques

Pierre Dumolard

Introduction

Ce manuel a pour objectif de faciliter la compréhension et l'usage des principales méthodes d'analyse statistique multivariée à tous ceux que concerne l'information spatialisée, géographes bien sûr mais, aussi, de plus en plus d'autres scientifiques, de disciplines environnementales aussi bien que sociales.

L'approche spatiale étant, par essence, combinatoire (donc complexe), nécessite des outils dédiés à l'analyse multidimensionnelle et à la représentation synthétique de ses résultats. Que cette approche soit purement exploratoire (comme dans le « data mining » opérant sur de grandes bases de données) ou confirmatoire (d'un modèle sémantique posé a priori pour validation), les méthodes multi-variables ont pour utilité essentielle d'être des « *valoriseurs* » de connaissance disciplinaire et non des ersatz de celle-ci.

Parmi toutes les techniques possibles d'analyse multidimensionnelle, le choix a été fait de ne présenter que :

- des méthodes purement statistiques (alors que d'autres façons de faire se développent, liées à « l'intelligence artificielle » comme les réseaux neuronaux par exemple),

- des méthodes couramment utilisées dont les résultats sont suffisamment stables et bien maîtrisés.

Ce manuel a une optique résolument *appliquée* : plus que d'une formulation mathématique pointue, il part de notions (finalement assez « naturelles ») mises en œuvre via des logiciels courants sur des exemples, complétés par des exercices corrigés. C'est là la structure d'un chapitre type. Bien sûr, la compréhension (à travers exemples et exercices) des notions multivariées implique comme pré-requis une connaissance minimale de la statistique descriptive uni- et bi-variée et, tout autant, une certaine culture disciplinaire.

L' information géographique est nécessairement contextuelle : elle comporte des influences de voisinage et d'interaction à diverses échelles (distances, connexités, concurrences / complémentarités, …). Un certain nombre de méthodes (qu'on peut regrouper sous le terme *d'analyse spatiale des données*) intègrent certaines de ces caractéristiques dans les algorithmes eux mêmes : elles ne sont pas présentées ici vu leur grand nombre et leur caractère assez peu universel (sauf exception). Sont par contre présentées ici des techniques relevant de ce qu'on appelera *analyse des données spatiales* qui ne se préoccupent de contraintes spatiales qu'a posteriori, via l'examen cartographique des résultats par exemple.

On distingue, dans ce manuel, deux grands types d'analyse multi-variables des données:

- des méthodes descriptives (de synthèse numérique)

    - analyses factorielles (chapitres 1, 2, 3, 4)
    - classifications descriptives (chapitre 5)

- des méthodes davantage explicatives.

    - régressions multiples (chapitre 6)
    - classification explicative (chapitre 7)

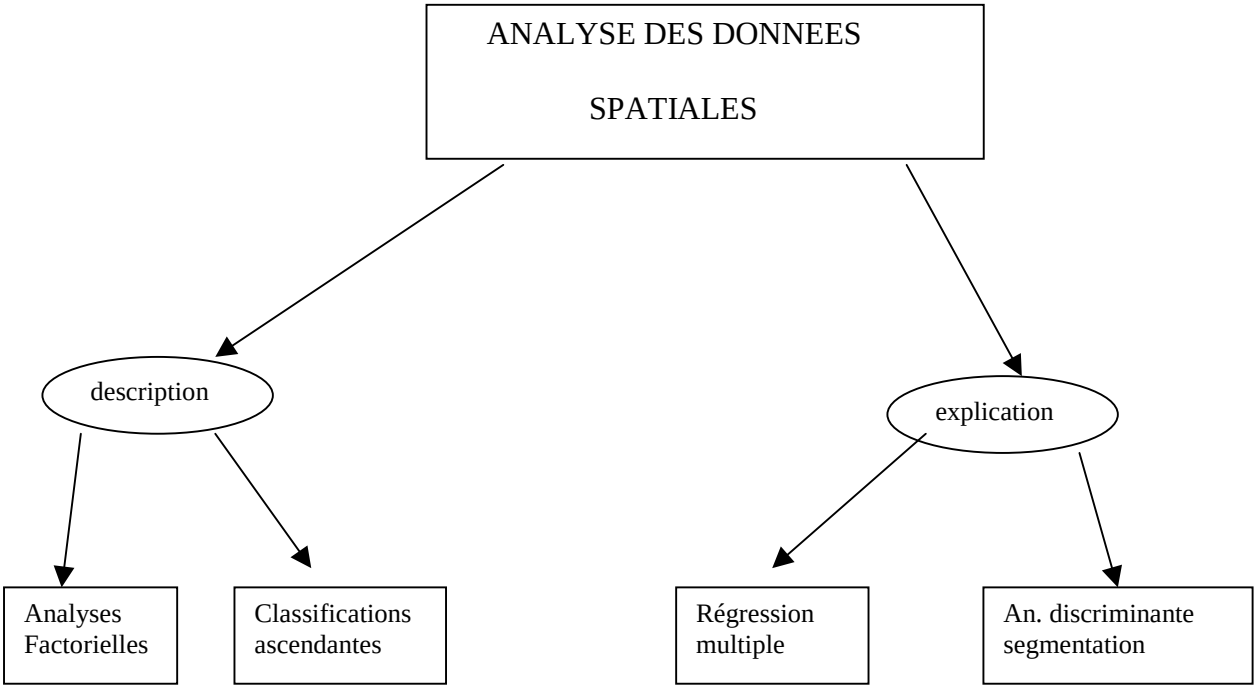


TABLE DES MATIERES

2.3 Aides à l'interprétation d'une AFC

3. **AFC sut tableaux de contingence à plus de 2 caractères**
   3.1 Exemple
   3.2 Interprétation de l'axe 1
   3.3 Interprétation de l'axe 2
   3.4 Plan des axes 1 et 2

B) Exercices corrigés

- **Exercice 1 :** structure d'âge des logements par région française
- **Exercice 2 :** usages de l'eau dans 16 départements du littoral atlantique

▪ Chapitre 4 L'analyse des correspondances multiples (AFCM)

A) Connaissances de base

1. **Généralités**
   1.1 Transformation d'un fichier en tableau de Burt
   1.2 Tableau disjonctif complet
   1.3 Equivalence des AFCM sur ces 2 types de tableau
2. **Résultats sur le tableau binaire 4.3**

C) Exercices corrigés

- **Exercice 1 :** enquête d'opinions aux USA sur les dépenses publiques
- **Exercice 2 :** 5 indicateurs de gestion environnementale de 34 villes françaises

▪ Chapitre 5 Méthodes de classification

A) Connaissances de base

1. **Utilité en géographie**
2. **Méthodes graphiques de classification**
   2.1 Sur graphique cartésien
   2.2 Par arborescence « raisonnée »
   2.3 Sur diagramme triangulaire
   2.4 Par matrice ordonnable de Bertin
3. **Méthodes statistiques de classification**
   3.1 Algorithmes de convergence
   3.2 Classifications arborescentes hiérarchiques (CAH)

B) Exercices corrigés

- **Exercice 1 :** Quelques indicateurs de l'Indice de Développement Humain pour 25 pays européens
- **Exercice 2 :** Recolonisation par le chêne pubescent d'un adret chartrousin

▪ Chapitre 6 Régression multiple

A) Connaissances de base

1. **Le modèle de la régression multiple**
   1.1 Extension du modèle de régression simple à plusieurs variables explicatives
   1.2 Exemple élémentaire
   1.3 En résumé
   1.4 Tests sur données d'échantillon
2. **Corrélations, multiple et partielles**
   2.1 Coefficient R de corrélation multiple
   2.2 Tests de R et R^2
   2.3 Coefficients de corrélation partielle
3. **Régression multiple pas à pas**
4. **Ajout d'une variable catégorielle à une régression multiple**
   4.1 Exemple
   4.2 Conditions de validité

B) Exercices corrigés

- **Exercice 1 :** Explication des températures moyennes de janvier pour un échantillon de villes des U.S.A.
- **Exercice 2 :** Types de contrat de travail de la population active de 20 régions de France métropolitaine

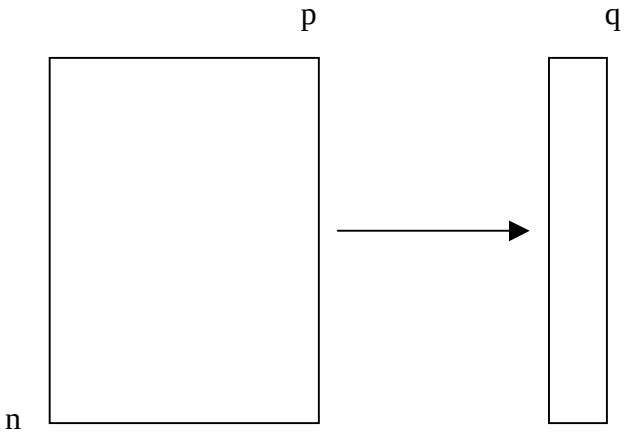
▪ Chapitre 7 Méthodes explicatives : compléments

1. **L'analyse discriminante**
   1.1 Modèle général
   1.2 Deux usages de l'analyse discriminante
   1.3 Exemple : discriminer populations rurales et non rurales en Alaska
2. **La segmentation**
   2.1 L'algorithme
   2.2 Aides à l'interprétation
   2.3 Usages, avantages et limites
   2.4 1^{er} exemple : les femmes suisses prises entre la famille et le travail
   2.5 2^{nd} exemple : la morphologie du terrain sur la planète Mars

Chapitre 1

ANALYSES FACTORIELLES : GENERALITES

Le but des analyses factorielles est de *résumer de grands tableaux numériques* en diminuant leur nombre de colonnes (passant de p colonnes à q « axes factoriels » les résumant).



En géographie, ces tableaux sont fréquemment des tableaux où les lignes repèrent des unités spatiales (par exemple, 96 départements de France métropolitaine) et les colonnes des variables juxtaposées (par exemple, 20 variables socio-économiques). On nomme habituellement « *matrice d'information spatiale* » ce type de présentation de données. Pour en maîtriser l'information, il est impératif de la résumer et il est impossible de le faire sans instrument adapté (dans l'exemple, 96*20 = 1920 nombres !). Faire l'analyse factorielle d'un tel tableau consiste à résumer ses 20 colonnes par 2 ou 3 « facteurs ».

Les expressions « facteurs » et « analyse factorielle » sont d'ailleurs très mal choisies puisqu'on obtient non pas des facteurs explicatifs mais des résumés descriptifs et qu'il ne s'agit pas d'analyse mais de synthèse : c'est l'histoire qui explique ce contresens.

1. Historique des analyses factorielles

Des psychomètres au début du 20ième siècle (Pearson, 1900) ont mis au point les premières analyses factorielles. Ils cherchaient, « cachées » derrière les résultats d'individus à des tests variés, des mesures de capacité intellectuelle (intelligence, mémoire, …) qu'ils ont nommées « facteurs » sous-jacents, explicatifs des résultats fournis par les tests psychologiques.

Avant la 2nde guerre mondiale, des statisticiens (Hotelling, Thurstone, 1934 sqq) ont repris ces travaux dans une perspective descriptive, mettant au point *l'analyse en composantes principales* (A.C.P.), adaptée au résumé, à la synthèse de *variables quantitatives*.

Après la 2nde guerre mondiale, un statisticien français (J.P.Benzecri, 1957 sqq) a adapté, sous le nom d'*analyse factorielle des correspondances* (A.F.C.), cette méthode à la synthèse de tableaux composés de *variables qualitatives*, fréquemment issues d'enquêtes (comme les tableaux de contingence).

Ces deux types d'analyse factorielle ne se sont répandus qu'à partir du moment où l'informatique s'est diffusée car il est à peu près impossible d'en réaliser les calculs à la main.

Bien qu'adaptés à des données de nature différente, ils possèdent de larges traits communs.

2. Traits communs aux analyses factorielles

2.1 Un tableau numérique peut se représenter par un nuage de points

Par exemple, un tableau ayant 96 lignes (départements français métropolitains) et 2 colonnes (par exemple taux de natalité, taux de mortalité) sera représenté graphiquement par un nuage de 96 points-département définis par leurs coordonnées sur deux axes perpendiculaires (l'un représentant le taux de natalité, l'autre le taux de mortalité). Ce graphique est un nuage de 96 points dans un espace géométrique de dimension 2 (un plan).

Si le tableau comporte non plus 2 colonnes mais 4 (en ajoutant, par exemple, taux de fécondité et taux de mortalité infantile), on ne peut plus visualiser directement le nuage des 96 points-département dans l'espace géométrique de dimension 4 mais, s'il n'existe plus graphiquement, cet espace existe *algébriquement*.

Plus généralement, une matrice d'information de *n* lignes et *p* colonnes est un nuage de *n* points-individus dans un espace défini par *p* axes orthogonaux (les p colonnes de la matrice d'information).

2.2 Résumer ce nuage de points : le projeter sur un sous espace

Ce sous espace est de dimension nettement inférieure, idéalement de dimension 2 de façon à pouvoir le représenter graphiquement.

Reprenant l'exemple du tableau à 96 lignes (départements) et 2 colonnes (taux de natalité et de mortalité), résumer ce nuage de dimension 2 consiste à le projeter « le mieux possible » sur une droite (espace de dimension 1). Puisqu'aucune des 2 variables n'est ci à privilégier par rapport à l'autre, la projection des points se fera *perpendiculairement* à cette droite « optimale » qu'on appellera *axe factoriel* et qui représente l'axe de plus grand allongement du nuage de points.

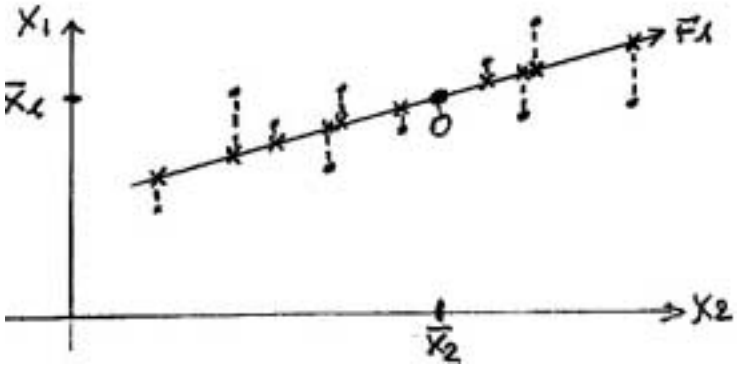


Figure 1.1: Projection des points d'un plan perpendiculairement à un axe factoriel

Chaque département est maintenant représenté par la coordonnée de sa projection sur l'axe factoriel 1, comme le symbolise la figure 1.1. Cet axe passe au plus près de l'ensemble des points, minimisant le carré des écarts entre chaque point et sa projection sur l'axe : c'est une *droite des moindres carrés* comme en régression linéaire, à la différence près que les projections lui sont perpendiculaires et non parallèles à l'un des axes.

L'origine de l'axe F1 (point 0) a pour coordonnées $\overline{xy}$ (moyenne des taux de natalité, moyenne des taux de fécondité). Ce point 0 est le point moyen de l'axe F1 et, par définition du point moyen, la variance des coordonnées des projections <0 est strictement égale à celle des coordonnées des projections >0.

2.3 Axes factoriels

Trouver la droite des moindres carrés la mieux ajustée à l'ensemble des points du nuage consiste à chercher celle qui *minimise* (Figure 1.2) la somme des $(A,A')^2$ ou, ce qui revient au même, celle qui *maximise* la somme des $(0,A')^2$: l'axe factoriel F1 est donc le principal axe d'allongement du nuage de points, celui qui prend en compte le plus possible de sa variance.

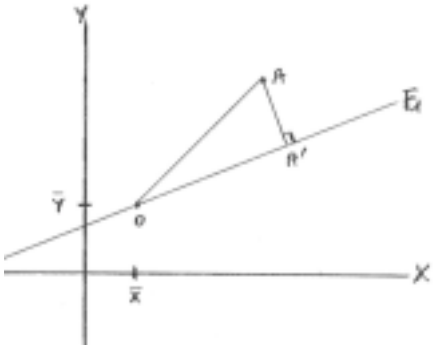


Figure 1.2 : minimiser $(A,A')^2$ = maximiser $(0,A')^2$

Si l'on reprend maintenant l'exemple du tableau à 96 lignes et 4 colonnes, on construira de la même façon un 1er axe factoriel. Considérant les écarts entre les points et cet axe (résidus du 1er axe factoriel), on peut de la même manière extraire F2, un *2ième axe factoriel*, perpendiculaire au 1er, de variance et d'allongement moindres. Le nuage de 96 points dans l'espace de dimension 4 aura ainsi été projeté sur le plan défini par les axes factoriels 1 et 2 et pourra être visualisé graphiquement (comme tout nuage de dimension 2). On pourrait même extraire un 3ième axe factoriel s'il apparaissait qu'il subsiste des résidus importants.

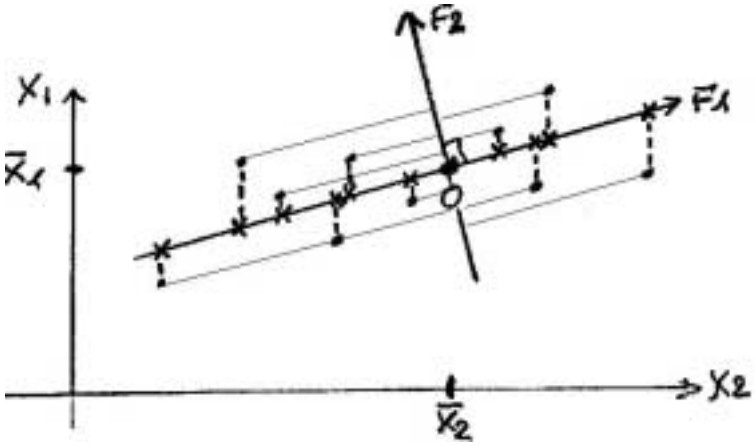


Figure 1.3 : F2 calculé sur les résidus de F1 (écarts des points à F1)

Chaque département est maintenant représenté par 2 coordonnées : celle de sa projection sur l'axe F1 et celle de sa projection sur l'axe F2. L'analyse factorielle a projeté un nuage de points d'un espace de dimension 4 sur un plan (de dimension 2) .

Elle a aussi opéré :
- un changement d'origine (les 2 axes, F1 et F2, sont orthogonaux au point de coordonnées 0,0 (centre du nuage projeté sur F1-F2),
- éventuellement, un changement d'échelle car les unités de mesure des différentes variables sont incompatibles entre elles.

Toute analyse factorielle est donc l'extraction progressive, dans un nuage de points multidimensionnel, de *résumés* unidimensionnels indépendants les uns des autres et d'importance informative (variance) dégressive. Cette technique a clairement une utilité exploratoire, réduisant la complexité, la résumant à ses principales dimensions et les hiérarchisant.

3. Procédure algébrique

Un axe factoriel est un axe d'allongement d'abord du nuage de points puis des résidus par rapport aux axes factoriels successifs. Chacun minimise la somme des carrés des écarts entre points et axe factoriel, ce qui revient au même que maximiser la somme des carrés des écarts entre les coordonnées des points sur l'axe et le point moyen du nuage de points (variance des coordonnées sur chaque axe factoriel). La variance des coordonnées sur l'axe 1 étant supérieure à celle sur l'axe 2, etc…, l'axe 1 est axe d'allongement majeur, l'axe 2 est axe d'allongement secondaire, etc. Cette variance des projections sur un axe est la *quantité d'information* qu'il prend en compte. L'information est la variance, mesurant l'originalité par rapport aux cas moyens.

Les axes factoriels étant orthogonaux donc indépendants les uns des autres, chacun apporte une information complémentaire aux autres. Ils se coupent perpendiculairement au *point moyen* du nuage de points (de coordonnées $\overline{X_1}$, …, $\overline{X_p}$ dans un tableau à p colonnes).

Chaque axe factoriel fait intervenir les p variables du tableau de données mais avec un *poids* différent d'un axe à l'autre (indiqués ci dessous par les coefficients a_j et b_j).

L'équation du 1er facteur peut s'écrire :

$$F_1 = a_1 X_1 + a_2 X_2 + ... + a_p X_p$$

Celle du 2ième facteur perpendiculaire au 1er :

$$F_2 = b_1 X_1 + b_2 X_2 + ... + b_p X_p$$

Où les X_j représentent les variables initiales et les a_j, b_j, … leurs poids.

Ce sont les poids (a_j, b_j, …) des p variables dans la définition des axes factoriels qui permettront de leur donner une signification thématique.

4. Informatiquement

un programme d'analyse factorielle est composé de 3 modules :

- *module 1* : transformation des données et calcul d'une matrice de covariation des variables (module différent selon la nature du tableau de données),

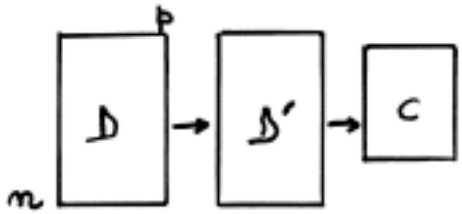


Figure 1.4 : schématisation de l'étape 1 d'une analyse factorielle

Le tableau de données **D** a n lignes et p colonnes. Il est transformé en un tableau **D'** (toujours à n lignes et p colonnes) : le type de transformation opérée sur D dépend de la nature des variables du tableau D. C'est à partir du tableau transformé D' que l'on calcule une matrice **C** (p lignes, p colonnes) de covariation (covariances ou corrélation) entre les p variables prises 2 à 2. La somme des valeurs de la diagonale de la matrice C est la variance totale du nuage de points multidimensionnel.

- *module 2* : extraction des axes factoriels et de la quantité d'information (variance) qu'ils prennent en compte (module commun à toutes les analyses factorielles), La figure 1.5 représente, sous forme de calcul matriciel, l'une des méthodes possibles d'extraction du 1er axe factoriel. Le processus de calcul est itératif pour les axes suivants.

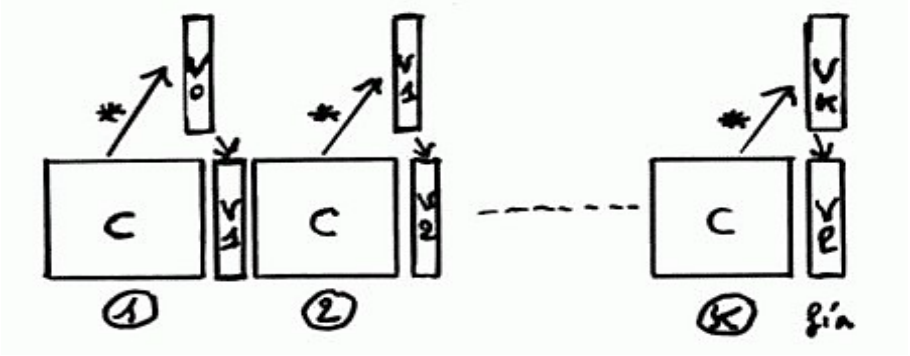


Figure1.5 : Une des méthodes d'extraction des axes factoriels

& *Etape 1* : la matrice C est multipliée par le vecteur unitaire (composé de 1) V_0, le résultat de ce produit matriciel est le vecteur V_1,

& *Etapes suivantes* : C est multiplié (matriciellement) successivement par les vecteurs V_2, …,V_k chacun étant le résultat du produit matriciel **C*V** de l'étape précédente.

& *Etape finale* : les itérations de calcul s'arrêtent quand $\mathbf{C*V_k=\lambda*V_l}$ c'est à dire quand, lors de 2 itérations successives, V_l est égal à V_k à une constante λ près.

Cette valeur λ est appelée *valeur propre* de l'axe factoriel : divisée par la variance totale du nuage de points multidimensionnel, elle indique le % de variance de l'axe (la quantité d'information qu'il prend en compte).

Les p valeurs du vecteur V_l , nommé *Vecteur Propre*, fournissent les coefficients directeurs (les pentes $\mathbf{a_j}$) de l'axe factoriel par rapport aux p axes du nuage de points : c'est sur ce vecteur qu'on projette les points du nuage.

La détermination des axes factoriels suivants procède de la même manière, à ceci près que la matrice C devient la matrice de covariation résiduelle (part de covariation non prise en compte par les axes précédents).

- *module 3* : aides à l'interprétation des axes (partiellement différentes selon la nature du tableau de données et donc le type d'analyse factorielle).

& La première aide à l'interprétation est la liste des *% de variance* pris en compte par chaque axe factoriel.

& La deuxième est constituée par les *coordonnées des projections* des lignes et des colonnes du tableau de données sur chaque axe factoriel.

& La troisième est la *qualité de représentation* (Q.R.) d'une ligne ou d'une colonne du tableau de données par sa projection sur un axe factoriel. Pour faire saisir cette notion, considérons 2 points ayant même coordonnée en projection sur un axe mais distants dans le nuage de dimension p. Sur la figure 1.6, la Q.R. du point P2 par l'axe F est meilleure que celle du point P1, plus éloigné de F. La Q.R. est le cosinus carré des angles P1-0-P' et P2-0-P'. Elle indique la proximité ou l'éloignement du point à l'axe factoriel : la somme des Q.R. d'un point sur les q axes retenus indique s'il est bien ou mal rendu par l'ensemble des q axes.

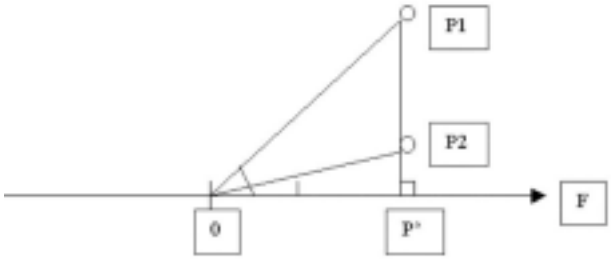


Figure 1.6 : Qualité de représentation des points P1 et P2 par un axe factoriel F

& La *contribution* relative d'une ligne ou d'une colonne de **D'** à la définition d'un axe est le carré de son écart à 0 divisé par la variance de l'axe. Par conséquent, la somme des contributions (lignes ou colonnes) sur chaque axe est égale à 1 ou 100% .

RESUME

En termes géométriques

L'analyse factorielle projette un nuage de n points d'un espace p-dimensionnel sur un sous-espace de dimension inférieure. Les nouvelles données sont les coordonnées des projections de ces n points sur chacun des axes factoriels orthogonaux.

En termes algébriques

Chaque axe factoriel est une combinaison des p variables initiales, chacune dotée d'un poids dans la combinaison. Les axes factoriels sont indépendants les uns des autres.

En termes informatiques

Tout programme d'analyse factorielle est composé de 3 modules :
- le calcul d'une matrice de covariation des p variables 2 à 2,
- le calcul des axes factoriels et de leur % de variance,
- le calcul d'aides à l'interprétation.

En termes d'information

Les axes factoriels sont des résumés des données du tableau initial,
Ils sont hiérarchisés selon la quantité d'information qu'ils prennent en compte,
Chacun porte une information nouvelle.

Ces résumés ne sont que des résumés, *à interpréter géographiquement*.

En ce sens, l'analyse factorielle ne fait que faciliter la description, en la hiérarchisant.

Chapitre 2

L'Analyse en Composantes Principales (ACP)

A) CONNAISSANCES DE BASE

L'ACP est employée pour l'analyse factorielle de *variables quantitatives continues* **(**ou quasi continues, c'est à dire discrètes à valeurs nombreuses). Leur présentation peut donner lieu à divers types de tableaux.

1. Types de tableaux pour l'ACP

La forme prototypique est la *matrice d'information* où :

- les différentes colonnes sont des variables quantitatives juxtaposées,
- les lignes correspondent à des unités statistiques, souvent spatiales en géographie (ponctuelle, linéaire ou surfacique selon l'échelle de leur représentation).

1.1 matrice d'information non spatiale

Il s'agit de mesures effectuées sur des poissons de la même espèce capturés à 5 stations (3, 4, 5, 7, 8) sur une radiale à partir d'un émissaire industriel. Les variables sont le poids du poisson (en kg), les concentrations dans les branchies en calcium, cadmium, cuivre, fer, et zinc et la distance en mètres de l'émissaire.

		Concentrations en					
station_poisson	Poids	Calcium	Cadmium	Cuivre	Fer	Zinc	Distance
3a	0.43	38.01	4.49	27.55	170.55	91.21	1835
3b	0.44	37.89	11.73	74.27	149.09	106.84	1835
3c	0.50	39.67	12.83	32.76	184.63	98.73	1835
3d	0.59	36.58	26.09	42.77	172.19	78.17	1835
3e	0.47	33.46	3.86	45.08	183.11	94.81	1835
4a	0.40	30.68	6.80	40.55	208.76	89.48	1485
4b	0.47	30.58	503.18	25.63	170.27	76.75	1485
4c	0.38	34.56	2.66	68.03	229.48	98.92	1485
4d	0.39	29.54	6.57	42.21	186.76	87.68	1485
4e	0.51	33.08	9.63	11.35	239.21	63.57	1485
5a	0.34	37.53	11.28	162.38	187.04	131.95	1335
5b	0.45	33.24	283.12	34.33	154.02	89.20	1335
5c	0.52	41.92	20.00	5.95	183.01	64.83	1335
5d	0.31	36.38	31.15	48.07	228.00	97.89	1335
5e	0.32	31.37	299.47	50.59	223.25	116.24	1335
7a	0.91	51.23	27.76	14.99	236.84	101.35	610
7b	0.36	66.32	55.54	110.90	332.15	120.87	610
7c	0.48	36.64	55.54	32.23	270.13	81.25	610
7d	0.35	64.61	37.90	16.15	490.95	91.61	610
7e	0.40	54.66	121.64	17.58	655.14	79.28	610
8a	0.40	45.73	391.83	27.56	419.00	105.95	185

Tableau 2.1 : concentrations chez des poissons selon leur distance à un rejet industriel

Les sept variables, issues de ***mesures***, sont toutes quantitatives continues. Juxtaposées en tableau, elles représentent différentes dimensions de la pollution des eaux par rejet d'un émissaire industriel. Les unités statistiques (l'échantillon des poissons capturés) sont des unités statistiques primaires, identifiées par un code. L'hypothèse sous-jacente à la constitution de ce tableau de données est que les différents polluants sont différemment absorbés, hypothèse qu'une analyse confirmerait, infirmerait ou nuancerait.

1.2 matrice d'information spatiale

pays	esp vie F	mort_inf	activF	% chom.	pnb/hb	% education	% santé
Allemagne	74.8	4.4	48.8	8.2	26768	4.3	10.6
Autriche	75.4	4.8	49	4.1	29075	4.9	8
Belgique	75.1	5	42.3	7.3	27952	5.8	8.8
Chypre	75.3	5.6	50.9	3.8	12724	5.8	6
Danemark	74.5	5.3	73.8	4.5	30096	8.1	8.4
Espagne	75.6	3.9	40.3	11.4	22538	5.6	7.7
Estonie	64.9	8.4	52.2	6.8	10201	6.8	5.5
Finlande	74.6	3.8	56.8	9.1	27215	5.6	6.6
France	75.5	4.6	47.8	8.7	27560	5.6	9.4
Grèce	75.4	6.1	37.6	10.3	17670	2.3	9.2
Hongrie	68.4	9.2	45.6	8.4	12733	5.2	5.7
Irlande	73	5.9	47.5	4.4	32549	4.5	7.2
Italie	76.7	4.5	36	9.1	26946	4.6	8
Lettonie	64.5	10.4	49.7	8.5	7809	6.2	4.8
Lituanie	65.9	8.6	54.6	10.9	8359	5.2	5.7
Luxembourg	75.2	5.8	42.5	2.4	50410	4	6.1
Malte	76.2	6	22.9	7.4	9875	5.7	8.9
Norvège	76.2	3.8	69.2	3.9	37070	7.4	8.5
Pays Bas	75.5	5.1	52.9	2.7	29614	5.2	8.2
Pologne	70.3	8.1	49.5	18.1	9852	5.1	4.2
Portugal	72.4	5.5	54.1	5	18500	5.5	8.2
Royaume Uni	75	5.6	53	5.1	26756	4.7	7.3
Slovaquie	69.7	8.6	52.9	17.4	12314	4.3	6.4
Slovénie	72.3	4.9	51.3	11.3	17762	5.2	8.2
Suède	77.5	3.4	76.2	4.9	26849	7.3	7.9
Suisse	77.2	4.9	58.8	3.1	30058	5.1	10.7
Tchéquie	72.2	4.1	51.3	9.8	15011	4.6	7.4

(source : ONU, Trends in Europe & N. America)

Tableau 2.2 : variables démographiques et économiques pour 27 pays européens

On a ici le cas le plus fréquent, celui d'***unités spatiales*** surfaciques (zones de type divers, communes, départements, pays, bassins hydrographiques, …), unités statistiques secondaires (« molécules »), chacune contenant des unités primaires (« atomes ») comme des personnes, des bâtiments, des parcelles, des zones pédologiques,… Dans le tableau 2.2, les différentes variables sont quantitatives, exprimées dans des unités de mesure différentes (taux, fréquences, mesure économique) et les unités statistiques (pays) sont des agrégats.

L'hypothèse qui a présidé à la constitution du tableau 2.2 est celle du lien entre niveaux de vie moyens, % des budgets nationaux consacrés à l'éducation et à la santé et quelques indicateurs démographiques. L'analyse en Composantes Principales permet de quantifier l'importance relative de leurs différents déterminants.

Tous ces tableaux sont ***dissymétriques*** (unités statistiques en lignes, variables en colonnes).

1.3 matrice d'information spatio-chronologique

	Janv.	Févr.	Mars	Avril	Mai	Juin	Juillet	Août	Sept.	Oct.	Nov.	Déc.
Conakry	0.6	1	3.2	20.5	137.1	379.8	1135.1	1109.1	618.3	293.6	69.6	7.9
Kindia	1.2	2	12.2	48	177.7	243.4	371.5	473.5	352.7	224.8	39.2	6.9
Boké	0	0	0	5.1	104.9	255.8	483.5	601	476.9	319	67.5	1.7
Mamou	2	4.6	22.3	73	155.7	207.4	332.7	404.3	337.1	200.1	47.9	4.6
Labé	1.7	3.6	8.7	35.3	140.1	233.1	315.9	339.5	295.8	133.2	133.6	1.8
Koundara	0	0.2	0.3	1.8	49.2	137.2	250.1	305.1	262.7	82.2	9.2	0.1
Kankan	2.1	1	23.8	67.4	134	202.3	261.9	319.4	304.7	130.2	18.1	1.2
Faranah	1.8	2.3	12.8	71.7	137	213.6	261.2	298.9	297.4	175.2	44.1	0.4
Siguiri	0.1	2	4.6	32.4	91.4	170.1	265.1	319.2	242.1	102	8.8	0.7
Kissidougou	6.9	18.4	45.7	129	206.1	257.3	264.8	339.2	328.4	247.1	77.1	15
Macenta	15	56	99.2	204.2	226.9	282	411.9	536.8	462.8	269.9	131.9	39.7
N'Zérékoré	9.6	44.9	112.9	160.6	174	199.2	225.4	298	324.7	188.8	77	23.9

(source : Météo Guinée)

Tableau 2.3 : Précipitations moyennes mensuelles de 12 villes guinéennes

Les tableaux spatio-chronologiques sont une forme particulière de matrice d'information spatiale où les lignes sont des unités spatiales, les colonnes des instants ou des durées : elles ne réfèrent généralement qu'à un seul caractère. Ici ce caractère est le total mensuel moyen de précipitations, les lignes repèrent des stations météorologiques et les colonnes les mois de l'année, le caractère étant les précipitations mesurées.

Ce tableau de valeurs quantitatives (mesures) peut être résumé par ACP, notamment dans le but de régionaliser les régimes de précipitations en Guinée mais, à la différence des tableaux précédents, il n'est pas dissymétrique (unités statistiques en ligne, variables différentes en colonne) : on pourrait fort bien permuter lignes et colonnes. En outre, cela aurait ici un sens d'additionner les précipitations en ligne (total annuel moyen) et en colonnes (divisé par 12 fournit la précipitation moyenne d'un mois en Guinée). Pour cette raison, un autre type d'analyse factorielle est également possible (Analyse Factorielle des Correspondances).

1.4 matrice d'information chronologique multivariée

Date de mesure	DBO5ND (Kg/l)	DCOND (Kg/l)	MES (Kg/l)	NR (Kg/l)
01-01-1999	9	88	4	24
01-02-1999	9	47	5	12
01-03-1999	13	59	3	9
01-04-1999	14	70	6	6
01-05-1999	11	61	12	6
01-06-1999	12	60	6	6
01-07-1999	12	51	5	10
01-08-1999	6	28	3	5
01-09-1999	15	80	16	23
01-10-1999	7	45	5	5
01-11-1999	7	44	8	3
01-12-1999	9	93	6	7

DBO5AD2 : Demande Biologique en Oxygène à 5 jours aprés décantation 2 heures en laboratoire
DBO5ND : Demande Biologique en Oxygène à 5 jours sur effluent non décanté en laboratoire
DCOAD2 : Demande Chimique en Oxygène aprés décantation 2 heures en laboratoire
DCOND : Demande Chimique en Oxygène sur effluent non décanté en laboratoire
MO : Matiéres Oxydables. Ce paramétre est le résultat du calcul : ((2*DBO5AD2) + DCOAD2) / 3
MES : Matiéres en Suspension
MP : Phosphore Organique et Minéral
NR : Azote Organique et Ammoniacal

Tableau 2.4 : Mesures de rejet d'une station d'épuration *(source: Agence de l'eau RMC)*

Le tableau 2.4 est un extrait de tableau chronologique multivarié dont la colonne 1 identifie les jours (unités statistiques), les colonnes 2 à 4 (variables) représentent quelques mesures de qualité de l'eau d'une station d'épuration. Les unités statistiques sont donc chronologiques et les variables (quantitatives) juxtaposées sont de natures différentes.

En résumé, 4 grands types de tableau sont susceptibles d'être résumés par ACP

- une *matrice d'information* (Tab. 2.1) juxtaposant des variables quantitatives pour un ensemble d'unités statistiques primaires quelconques (poissons dans l'exemple),
- une *matrice d'information spatiale* (Tab. 2.2) juxtaposant des variables quantitatives pour des unités spatiales. A des unités spatiales ponctuelles correspondent souvent des mesures, a des agrégats spatiaux surfaciques des variables de comptage,
- une *matrice d'information spatio-chronologique* (Tab. 2.3) où les lignes repèrent des unités spatiales, les colonnes des unités de temps et où, généralement, un seul caractère est présent (précipitations dans l'exemple exposé),
- une *matrice chronologique multivariée* où les lignes sont des unités statistiques chronologiques et les colonnes des variables quantitatives juxtaposées.

Dans tous ces cas, la matrice d'information est considérée de façon *dissymétrique*, unités statistiques en ligne et variables quantitatives (mesures, comptages, taux,…) en colonne.

2. La création d'un tableau de données pour l'ACP

La constitution d'une matrice d'information est, dans tous les cas, le résultat d'un compromis entre une hypothèse sur les dimensions sémantiques d'un phénomène (qu'il s'agit de simplifier et d'ordonner) et les données numériques accessibles pour le caractériser.

L'ACP, comme les autres méthodes d'analyse multivariée, suppose que le tableau à résumer est « sans trou » (sans valeur manquante), ce qui peut être vu comme un inconvénient dans nombre d'applications. Par exemple, il est fréquent en climatologie que les relevés de certaines stations météorologiques connaissent des interruptions et fournissent donc des données chronologiquement incomplètes : il existe, dans ce cas, des moyens pour estimer les valeurs manquantes et produire des tableaux sans absence de valeurs.

Concernant les matrices d'information spatiale, deux grands types peuvent être distingués, selon la nature des unités spatiales (et donc la nature des variables les décrivant):

- si ce sont des *unités statistiques primaires* (relevés en des points précis), les variables quantitatives qui leur correspondent sont issues de *mesures* (surface, poids, volume, précipitation, température, distance, caractéristique physique ou biologique, …) de sorte qu'à chaque lieu échantillon corresponde une grandeur pour chaque variable ; dans ce premier cas (mesures sur unités primaires) se pose souvent pour le géographe le problème de la généralisation (interpolation) spatiale des résultats,

- si ce sont des *unités statistiques secondaires* (agrégats surfaciques ou linéaires contenant des unités primaires), les variables sont généralement quantitatives car elles correspondent à des *comptages*. Par exemple, une variable comme le sexe (qualitative au niveau de personnes observées) devient, au niveau agrégé, deux variables quantitatives de comptage (d'hommes d'une part, de femmes d'autre part) ; ces variables de comptage peuvent être exprimées sous forme d'effectifs, de fréquences (pourcentages d'unités primaires par unité secondaire) ou de taux (rapports d'effectifs ou de fréquences) ; dans ce cas (comptages sur agrégats) existe le

risque « d'erreur écologique », c'est à dire le risque d'affecter à tous les individus d'une zone le comportement moyen observé sur l'agrégat.

3. Les 3 phases d'une ACP sur ces types de tableaux

Les quatre types de matrice d'information distingués au &2.1 peuvent être soumis à une ACP ; comme toute analyse factorielle, celle ci se déroule en trois étapes (&1.2) :

- transformation du tableau de données et calcul des covariances,
- calcul des axes factoriels et de leurs variances,
- calcul d'aides à l'interprétation des résultats.

3.1 Transformation du tableau de données et calcul des covariances

- Les p variables (colonnes) du tableau **D** de données sont **centrées.**

Pour chaque variable j, on remplace chaque valeur par son écart à sa moyenne

$$\mathbf{D'_{ij} = D_{ij} - m_j}$$

Ce centrage des variables correspond, géométriquement, à un changement d'origine du nuage de points qui, une fois projeté sur les axes factoriels, aura pour point moyen le point de coordonnées 0, 0, 0, … Algébriquement, le centrage des variables élimine l'ordonnée à l'origine dans l'équation de chaque axe factoriel.

Si les variables sont exprimées dans la **même unité de mesure** et que l'on désire conserver pour chaque unité statistique les différences de grandeur exprimées par les différentes variables, l'analyse se poursuivra sur le tableau **D'** (variables centrées où l'on a effacé les différences de moyenne mais pas celles de variance). L'exemple suivant illustre ce cas de figure.

distances à	***Bayonne-Biarritz***	***Pamplona***	***Pau***	***San Sebastian***
Anglet	5	106	113	50
Artouste	170	198	58	213
Cambo	21	102	122	46
Gourette	162	86	114	61
Hasparren	29	190	50	205
Hendaye	33	95	106	71
Pierre St Martin	144	78	142	20
Laruns	150	156	75	187
Lourdes	152	141	89	208
Mauléon	93	250	40	194
Navarrenx	89	113	59	135
Oloron Ste Marie	110	136	42	132
Orthez	80	146	35	152
St Jean de Luz	21	87	92	79
St Jean Pied de Port	54	64	74	160
St Palais	61	34	105	95
Salies de Bearn	65	16	128	104
Sauveterre de Béarn	75	20	118	117

(source : ViaMichelin)

Tableau 2.5 : distances aux métropoles régionales de petites villes basco-béarnaises

Faire l'ACP d'un tel tableau conduit à différencier des zones d'influence prioritaires des métropoles régionales en fonction de leurs distances routières aux petites villes mentionnées. On admettra qu'il faille conserver ces distances en km, unité de mesure commune significative.

Le tableau 2.6 est le tableau **D'** (centré) correspondant au Tableau 2.5 : c'est sur celui-ci que l'ACP va se poursuivre. Les différences de distances moyennes aux 4 métropoles ont été effacées (elles étaient de 84.11 km à Bayonne-Biarritz, 112,11 km à Pampelona, 86,77 km à Pau et 123,83 km à San Sebastian). Vous pourrez vérifier par vous même que, aux arrondis de calcul près, la moyenne des 4 variables centrées est égale à 0.

distances centrées à	Bayonne-Biarritz	Pampelune	Pau	San Sebastian
Anglet	-79.11	-6.11	26.23	-73.83
Artouste	85.89	85.89	-28.77	89.17
Cambo	-63.11	-10.11	35.23	-77.83
Gourette	77.89	-26.11	27.23	-62.83
Hasparren	-55.11	77.89	-36.77	81.17
Hendaye	-51.11	-17.11	19.23	-52.83
Pierre St Martin	59.89	-34.11	55.23	-103.83
Laruns	65.89	43.89	-11.77	63.17
Lourdes	67.89	28.89	2.23	84.17
Mauléon	8.89	137.89	-46.77	70.17
Navarrenx	4.89	0.89	-27.77	11.17
Oloron Ste Marie	25.89	23.89	-44.77	8.17
Orthez	-4.11	33.89	-51.77	28.17
St Jean de Luz	-63.11	-25.11	5.23	-44.83
St Jean Pied de Port	-30.11	-48.11	-12.77	36.17
St Palais	-23.11	-78.11	18.23	-28.83
Salies de Bearn	-19.11	-96.11	41.23	-19.83
Sauveterre de Béarn	-9.11	-92.11	31.23	-6.83

Tableau 2.6 : Tableau centré correspondant au tableau 2.5

Il en serait de même si l'on avait construit pour les communes d'un département un tableau de distances aux villes voisines leur fournissant différents niveaux de service (collège, lycée, université, hôpital spécialisé, commerces rares, …).

- La plupart du temps, les variables juxtaposées dans une matrice d'information ont des unités de mesure différentes, qu'il faut donc ramener, pour les combiner, à une unité de mesure commune: c'est accompli en *standardisant* chaque variable (centrage et réduction).

$$D''_{ij} = (D_{ij} - m_j) / \sigma_j$$

Pour chaque variable j (colonne) du tableau, la standardisation exprime l'écart de chaque valeur à sa moyenne en écart type de cette variable. L'écart type devient l'unité de mesure commune à toutes les variables. Ce sont donc les valeurs standardisées des variables (c'est à dire leurs variabilités *relatives*) qui produiront les résultats de l'ACP.

Le tableau 2.7 correspond à la standardisation des variables du tableau 2.2 : vous pourrez vérifier par vous même que (aux arrondis de calcul près) la moyenne des variables standardisées est égale à 0 et leur écart type à 1.

	esp vie F	mort_inf	activF	%chom	pnb/hb	%educ	%santé
Allemagne	0.40	-0.75	-0.17	0.14	0.43	-0.91	1.88
Autriche	0.57	-0.53	-0.15	-0.89	0.65	-0.39	0.28
Belgique	0.48	-0.42	-0.76	-0.09	0.54	0.38	0.77
Chypre	0.54	-0.10	0.02	-0.96	-0.94	0.38	-0.95
Danemark	0.32	-0.26	2.11	-0.79	0.75	2.36	0.53
Espagne	0.62	-1.01	-0.94	0.94	0.02	0.21	0.10
Estonie	-2.27	1.40	0.14	-0.21	-1.18	1.24	-1.25
Finlande	0.35	-1.07	0.56	0.36	0.47	0.21	-0.58
France	0.59	-0.64	-0.26	0.26	0.50	0.21	1.14
Grèce	0.57	0.17	-1.19	0.66	-0.46	-2.63	1.02
Hongrie	-1.33	1.83	-0.46	0.19	-0.94	-0.13	-1.13
Irlande	-0.08	0.06	-0.29	-0.81	0.99	-0.74	-0.21
Italie	0.92	-0.69	-1.33	0.36	0.44	-0.65	0.28
Lettonie	-2.38	2.48	-0.09	0.21	-1.42	0.73	-1.68
Lituanie	-2.00	1.51	0.36	0.81	-1.36	-0.13	-1.13
Luxembourg	0.51	0.01	-0.74	-1.31	2.72	-1.17	-0.88
Malte	0.78	0.11	-2.53	-0.06	-1.22	0.30	0.83
Norvège	0.78	-1.07	1.69	-0.94	1.43	1.76	0.59
Pays Bas	0.59	-0.37	0.21	-1.24	0.70	-0.13	0.40
Pologne	-0.81	1.24	-0.10	2.61	-1.22	-0.22	-2.05
Portugal	-0.25	-0.16	0.31	-0.66	-0.38	0.12	0.40
Royaume Uni	0.46	-0.10	0.21	-0.64	0.43	-0.56	-0.15
Slovaquie	-0.97	1.51	0.21	2.43	-0.98	-0.91	-0.70
Slovénie	-0.27	-0.48	0.06	0.91	-0.45	-0.13	0.40
Suède	1.13	-1.28	2.33	-0.69	0.43	1.67	0.22
Suisse	1.05	-0.48	0.74	-1.14	0.75	-0.22	1.94
Tchéquie	-0.30	-0.91	0.06	0.54	-0.72	-0.65	-0.09

Tableau 2.7 : Tableau standardisé correspondant au tableau 2.2

- Une fois opérée la transformation du tableau initial **D** en tableau **D'** (centré) ou **D''** (standardisé, c'est à dire centré et réduit), on passe au calcul d'une matrice de relations entre les p variables considérées deux à deux.

Cette matrice **C** a donc p lignes et p colonnes. Elle contient dans chaque case ij :

- la *covariance* entre variables i et j si **D** a été centré en **D'**

Le tableau 2.8 fournit la matrice de covariances entre les 4 variables du tableau 2.5

	Bayonne-Biarritz	Pampelune	Pau	San Sebastian
Bayonne-Biarritz	2696.3210	837.8210	-214.3642	1056.9074
Pampelune	837.8210	3665.7654	-1461.4198	2306.5185
Pau	-214.3642	-1461.4198	1069.7284	-1495.2593
San Sebastian	1056.9074	2306.5185	-1495.2593	3618.9167

Tableau 2.8 : Covariances entre variables du tableau 2.5

- la *corrélation* entre variables i et j si **D** a été standardisé en **D''** (la covariance de variables standardisées est leur coefficient de corrélation de Bravais-Pearson).

Le tableau 2.9 fournit, quant à lui, la matrice de corrélations entre les 7 variables du tableau 2.2. Il est symétrique puisque $r_{ij} = r_{ji}$ Seul le triangle supérieur est ici présenté et la diagonale est égale à 1 (corrélations des variables avec elles mêmes).

	esp vie F st	mort_inf st	activF st	%chom st	pnb/hb st	%educ st	%santé st
esp vie F st	1	-0.8621	-0.0035	-0.4116	0.6511	-0.0629	0.7017
mort_inf st		1	-0.1769	0.3783	-0.6167	-0.0862	-0.6746
activF st			1	-0.2426	0.2178	0.6000	-0.0096
%chom st				1	-0.5933	-0.2758	-0.3408
pnb/hb st					1	0.0007	0.4232
%educ st						1	-0.0868
%santé st							1

Tableau 2.9 : Corrélations entre variables du tableau 2.2

NB : Si les variables du tableau initial **D** sont des rangs (variables qualitatives ordinales), la transformation de **D** en **D''** produira des coefficients de corrélation de rangs de Spearman.

En résumé : la 1ère phase d'une ACP consiste à :

- transformer le tableau initial **D** en un tableau **D'** (de variables centrées) ou **D''** (de variables standardisées),
- calculer un tableau **C** de relations entre variables 2 à 2 (covariances sur **D'**, corrélations sur **D''**)

3.2 Calcul des axes factoriels et de leur % de variance

Il est identique dans toutes les analyses factorielles (cf & 1.3, module 2).

- les axes factoriels sont calculés sur le tableau **C** des covariances (ou des corrélations),
- les q (q<p) axes factoriels extraits sont orthogonaux entre eux, définissant ainsi un sous espace euclidien (si q=2, c'est un plan)
- le 1er axe factoriel est le principal axe d'allongement du nuage de points, celui qui prend en compte la plus grande variance des coordonnées de leurs projections sur cet axe,
- Les axes factoriels suivants sont de variance décroissante.

3.3. Aides à l'interprétation des résultats

C'est la phase essentielle pour l'utilisateur. Les différents logiciels offrent, en ce domaine, des indicateurs plus ou moins complets. On n'indique ici que les plus fréquemment rencontrés. Pour être plus compréhensible, on les expose, sur l'exemple du tableau 2.2 (27 pays européens, 7 variables), les résultats fournis par un programme d'ACP

3.3.1 *signification des variables du tableau 2.2*

Il est temps d'être plus précis sur la signification exacte des 7 variables du tableau 2.2 :
esp vie F **:** nombre moyen d'années que vivrait une fille née en 2001 si la mortalité féminine par âge demeurait la même qu'en 2001
Mort_inf **:** nombre d'enfants <1 an morts en 2001 / nombre d'enfants nés vivants en 2001
Activ F **:** nombre de femmes ayant un emploi / nombre de femmes d'âge actif
% chom **:** (nombre de chômeurs / nombre des actifs de plus de 15 ans)*100

Pnb/hb **:** produit national brut annuel par habitant (exprimé en $)
%éducation **:** dépenses d'éducation (de source publique ou privée) en % du Pnb
%santé **:** dépenses de santé (de source publique ou privée) en % du Pnb

3.3.2 *moyennes, écarts type et coefficients de corrélation*

Les variables du tableau 2.2 ont été standardisées (centrées et réduites), ce qui implique que leurs moyennes et variances n'interviendront pas dans l'ACP. Il convient de souligner que ces moyennes et variances sont non pondérées par le poids démographique des différents pays (le Luxembourg pèse autant que l'Allemagne).
Il n'est toutefois pas inutile de les commenter, en calculant notamment (si toutes les valeurs sont supérieures ou égales à 0) leurs *coefficients de variation* (σ_j / m_j). On a ainsi une estimation de la variabilité des valeurs de l'ensemble des pays pour chaque variable (Tab. 2.10).

	moyenne	Ecart type	Coefficient De variation
esp vie F	73.307	3.700	*0.050*
Mort_inf	5.789	1.862	0.322
Activ F	50.65	10.98	0.217
% chom	7.652	4.007	**0.524**
Pnb/hb	22380	10288	**0.460**
%education	5.356	1.163	0.217
%santé	7.541	1.629	0.216

Tableau 2.10 : variabilité relative des variables du tableau 2.2

On remarque immédiatement:
- la très faible différence d'espérance de vie entre pays européens (à l'est comme à l'ouest), tenant à des structures d'âge à peu près uniformément vieillies,
- des différences modérées (variabilité entre 1/5 et 1/3) pour la majorité des variables,
- de fortes différences de revenu moyen (quelle que soit l'imperfection de la mesure) et de chômage (marque de politiques d'emploi différentes).

3.3.3 *tableau de covariances ou de corrélations*

Dans le cas présent (variables standardisées), il s'agit d'un tableau de corrélations (cf Tableau 2.9). Même avec aussi peu de variables, il est peu lisible : il convient donc de le simplifier et, éventuellement, de représenter graphiquement ce tableau simplifié. En décidant (tout à fait empiriquement) de ne retenir que les corrélations supérieures à +0.60 et inférieures à –0.59, on obtient la figure 2.1.

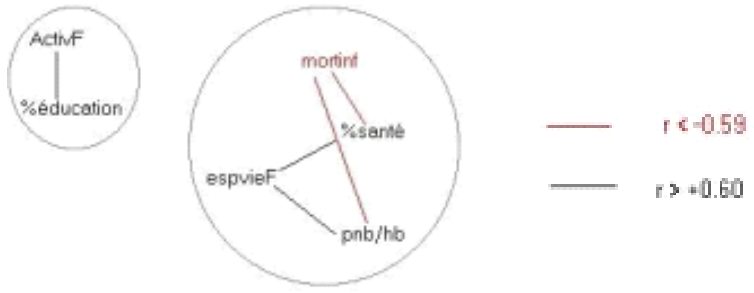


Figure 2.1 : graphe simplifié des corrélations du tableau 2.9

Le principal groupe d'inter-corrélations concerne les variables démographiques, le revenu moyen par habitant et la part du budget dédié aux dépenses de santé. Deux interprétations (éventuellement complémentaires) sont possibles :

& en Europe, l'espérance de vie tend à être proportionnelle au Pnb/hb, aux dépenses de santé et la mortalité infantile à leur être inversement proportionnelle,
& ces variables inter-corrélées seraient en partie liées à la structure d'âge des pays et à leur niveau de développement *social* (et non économique seulement).

Un deuxième groupe de variables corrélées apparaît, connectant taux d'activité féminine et part des budgets allouée à l'éducation (variables *sociales*).

3.3.4 *Résultats concernant l'axe factoriel 1 (47.7% de la variance)*

- *corrélations variables – axe 1*

	Corrélation + axe 1	Qualité de représentation
Esp vie F	***0.86***	**0.73**
Pnb/hb	***0.81***	**0.65**
% santé	***0.78***	**0.60**
Activ F	0.17	0.04
% éducation	0.10	0.01
%chom.	***-0.62***	0.39
mortinf	***-0.90***	**0.81**

Tableau 2.11 : corrélation des variables avec l'axe factoriel 1

L'axe factoriel 1 (représentant presque la moitié de l'information du tableau initial), oppose (pour les 27 pays considérés) l'espérance de vie féminine, le revenu moyen par habitant, le % des budgets consacrés à la santé d'une part aux forts taux de chômage et de mortalité infantile d'autre part. C'est donc un axe de niveau de développement socio-démographique. Les qualités de représentation (comprises entre 0 et 1 puisque ce sont des cosinus carrés) de ces variables par l'axe 1 (cf &1.4) sont correctes, hormis pour la variable « chômage ».

- *coordonnées des pays sur l'axe 1*

	Coordonnées sur l'axe 1	Qualité de représentation
Suisse	**2.51**	**0.773**
Norvège	**2.36**	**0.547**
Suède	**1.96**	0.380
France	**1.62**	**0.614**
Luxembourg	**1.54**	0.184
Allemagne	**1.43**	0.375
Pays Bas	**1.24**	**0.614**
Autriche	**1.21**	**0.693**
Danemark	**1.09**	0.112
Belgique	**1.07**	**0.516**
Italie	**1.03**	0.221
Espagne	0.80	0.136
Finlande	0.76	0.223
Portugal	0.43	0.024
Royaume Uni	0.28	0.078
Irlande	0.15	0.008
Slovénie	-0.03	0.001
Grèce	-0.22	0.005
Malte	-0.23	0.006
Chypre	-0.38	0.047
Tchéquie	-0.45	0.077
Estonie	***-2.53***	**0.631**
Hongrie	***-2.77***	**0.879**
Lituanie	***-2.78***	**0.958**
Slovaquie	***-2.95***	**0.768**
Pologne	***-3.43***	**0.721**
Lettonie	***-3.74***	**0.834**

Tableau 2.12 : Coordonnées des pays sur l'axe 1 et qualité de leur représentation

De façon cohérente avec l'interprétation en termes de variables, s'opposent pays à forts Pnb/hb, espérance de vie féminine et faible mortalité infantile à des pays à faibles Pnb/hb, espérance de vie féminine et mortalité infantile relativement forte. Cette interprétation, s'appuyant sur les coordonnées extrêmes (en positif et négatif), doit tenir compte de la qualité de représentation des pays par leurs coordonnées. Suède, Luxembourg, Allemagne, Danemark, Italie sont assez mal représentés si bien que ce 1er axe n'épuise pas leur originalité.

3.3.5 *Résultats concernant l'axe factoriel 2 (24.5% de la variance)*

- *Corrélations variables – axe 2*

	Corrélations avec l'axe 2	**Qualité de représentation**	**Cumul sur les Axes 1 et 2**
% éducation	**0.83**	**0.6956**	*0.7061*
Activ F	**0.82**	**0.6734**	*0.7102*
Mortinf	0.07	0.0048	*0.8188*
Pnb/hb	0.04	0.0012	*0.6542*
% santé	-0.23	0.0544	*0.6556*
Espvie F	-0.32	0.1050	*0.8387*
% chom	-0.40	0.1605	***0.5498***

Tableau 2.13 : Corrélations variables – axe 2 et qualité de leur représentation

Au vu des corrélations et de leur qualité de représentation, cet axe 2, représentant presque ¼ de l'information initiale, est principalement déterminé par le taux d'activité féminine et la part des budgets nationaux consacrés à l'éducation. Ces deux variables, fortement corrélées entre elles et avec l'axe 2, étaient sur le graphe de la figure 2.1 le second paquet de corrélations. Cet axe met donc en avant les différences inter pays européens d'activité féminine et leurs liens avec les efforts consentis pour l'éducation. Ces différences sont indépendantes du Pnb/hb.

- *Coordonnées sur l'axe 2 et qualité de leur représentation*

	Coordonnées Sur l'axe 2	**Qualité de représentation**	***Cumul Q.R. sur les axes 1 et 2***
Danemark	**2.92**	***0.799***	0.911
Suède	**2.20**	***0.478***	0.858
Norvège	**2.03**	***0.403***	0.950
Portugal	**1.92**	***0.481***	0.505
Estonie	1.63	0.262	0.893
Lettonie	1.25	0.093	0.927
Chypre	0.62	0.127	***0.174***
Lithuanie	0.38	0.018	0.976
Hongrie	0.23	0.006	0.886
Pays Bas	0.23	0.020	0.635
Finlande	0.17	0.011	***0.235***
Royaume Uni	-0.02	0.000	***0.079***
Suisse	-0.12	0.002	0.775
Irlande	-0.17	0.010	***0.017***
Autriche	-0.31	0.046	0.738
Pologne	-0.47	0.014	0.782
Tchéquie	-0.51	0.100	***0.177***
Belgique	-0.51	0.114	0.630
Slovénie	-0.55	0.200	***0.201***
France	-0.73	0.124	0.738
Luxembourg	-0.77	0.047	***0.231***
Slovaquie	-0.84	0.059	0.780
Espagne	-1.15	0.278	***0.413***
Allemagne	-1.24	0.281	0.655
Malte	-1.54	0.300	***0.306***
Italie	***-1.75***	***0.636***	0.856
Grèce	***-2.90***	***0.814***	0.818

Tableau 2.14 : coordonnées des pays et leur qualité de représentation

L'opposition est nette entre pays scandinaves à forts budgets éducatifs, forte participation des femmes au monde du travail et pays méditerranéens (Grèce, Italie et aussi Malte et Espagne, moins bien représentés par l'axe) où ces variables présentent des valeurs plus faibles. Cet axe est, en quelque sorte, celui de l'opposition de deux modèles d'investissement éducatif et de ses prolongements dans la vie active pour les femmes.

3.3.6 *graphiques sur le plan des axes 1 et 2*

C'est cependant sur des combinaisons d'axes factoriels que les résultats de l'ACP sont le mieux interprétables et, notamment par des graphiques du plan des axes 1 et 2 s'ils contiennent une bonne part de l'information du tableau initial (ici presque les ¾).

- *cercle des corrélations sur le plan des axes 1 et 2*

Les corrélations des variables avec les 2 axes sont représentées par un point de coordonnées r_{1j} et r_{2j} et, comme toutes les corrélations sont, par construction, comprises entre –1 et +1, ces points sont tous situés à l'intérieur d'un cercle de rayon 1. Si on les joint à l'origine par un vecteur, on obtient un cercle des corrélations semblable à celui de la figure 2.2

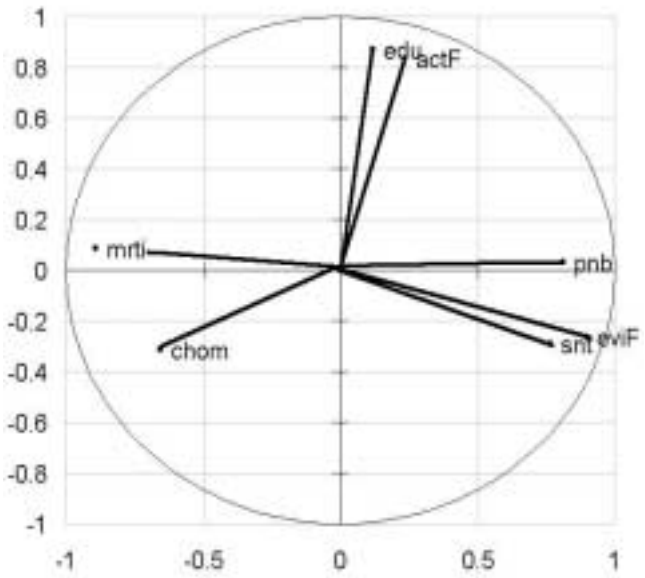


Figure 2.2 : cercle des corrélations variables – axes 1 et 2

Les principes de lecture d'une telle figure sont les suivants :

- un vecteur de faible longueur indique une faible corrélation donc une mauvaise prise en compte de la variable par l'ensemble des deux axes : c'est le cas du taux de chômage,

- un vecteur de grande longueur proche d'un axe indique son poids dans la définition de l'axe : c'est le cas du Pnb/hb pour l'axe 1, des dépenses éducatives pour l'axe 2,

- deux vecteurs longs, proches d'un axe et opposés l'un à l'autre traduisent que l'axe a été construit en opposant ces deux variables : c'est le cas du Pnb/hb et de la mortalité infantile pour ce qui concerne l'axe 1 (qui a « 2 pieds », alors que l'axe 2 est unijambiste !),

- des vecteurs longs voisins les uns des autres indiquent une forte corrélation entre les variables qu'ils représentent : c'est le cas ici de l'espérance vie féminine et des dépenses de santé d'une part, des dépenses d'éducation et du taux d'activité féminine d'autre part.

3.3.7 *graphique des individus et variables sur le plan des axes 1 et 2*

Moyennant une transformation ou une superposition d'échelles pour faire figurer corrélations (ou covariances) et coordonnées sur le même graphique, on peut construire la figure 2.3.

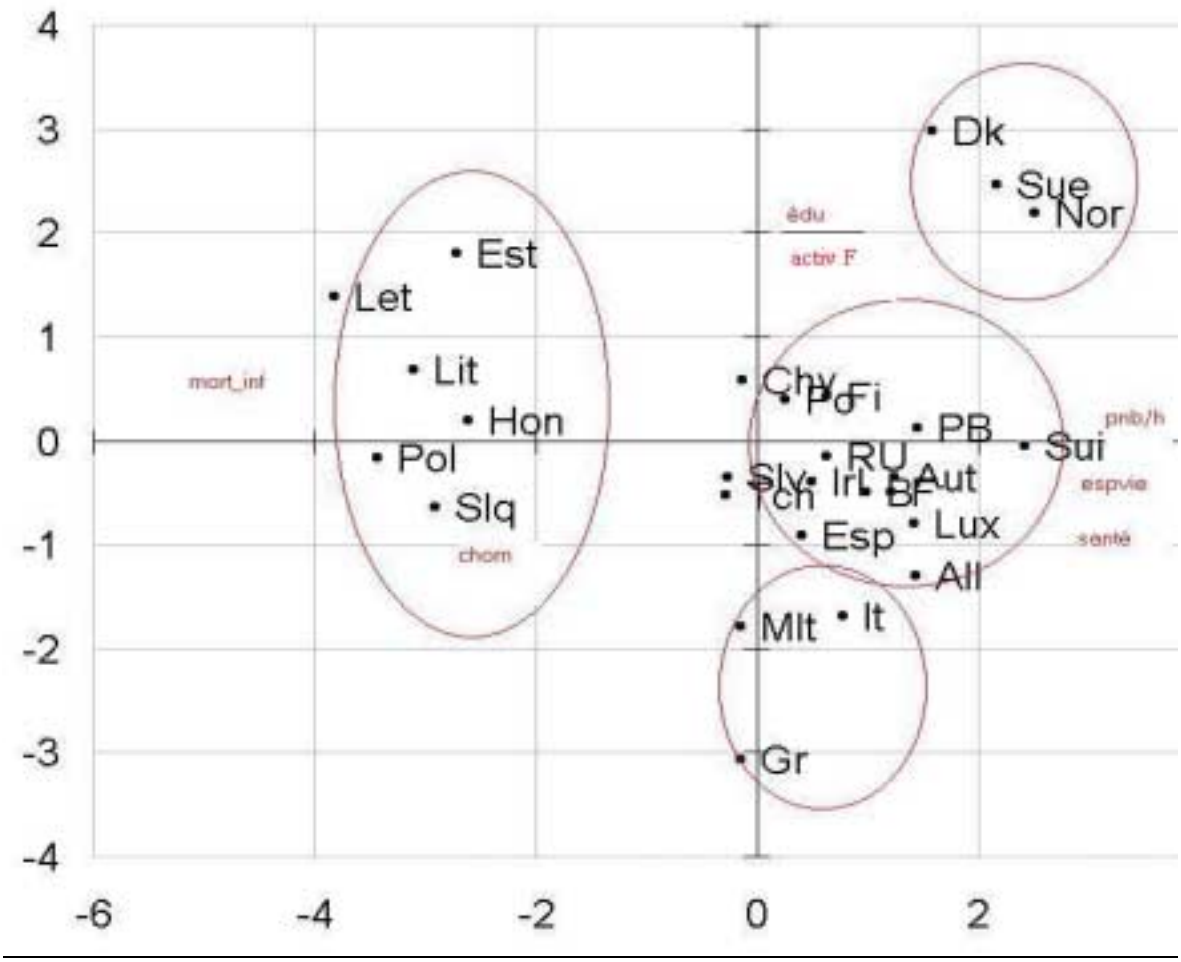


Figure 2.3 : représentation variables-individus sur le plan des axes 1 et 2

Les proximités (et les éloignements) des points pays et des points variables est explicative. Par exemple, le groupe des pays scandinaves est voisin de %éducation, esp_vie F et éloigné de mort_inf, la Suisse est proche du point Pnb/hb, le groupe « méditerranéen (Grèce, Malte, Italie) est, lui, éloigné de tous les points variables.

Les points pays proches de l'origine sont, en général, peu pris en compte par les axes 1 et 2 : ils le seraient par des axes suivants (Chypre, Slovénie, Pologne, Tchéquie, Finlande, Royaume Uni, Irlande, Espagne). Leur exacte qualité de représentation sur le plan F1-F2 est indiquée par la dernière colonne du tableau 2.14. On peut la représenter, en chaque point pays, par un cercle proportionnel au cumul des qualités de représentation sur les axes 1 et 2.

La figure 2.4, lissée et généralisée, les visualise aussi. Elle représente, en quelque sorte, une *carte de fiabilité* de la représentation des pays sur le plan des axes factoriels 1 et 2.

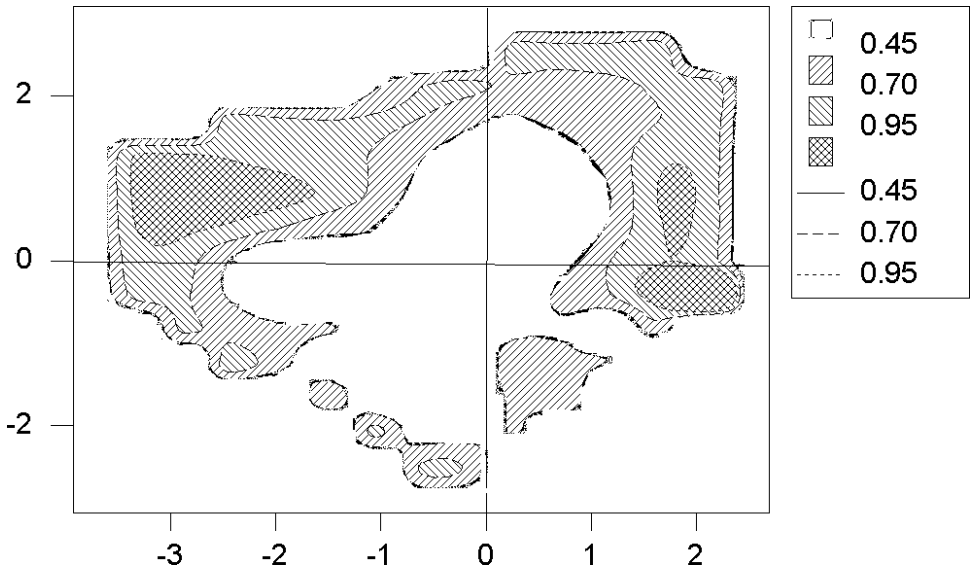


Figure 2.4 : Carte de fiabilité des projections des pays sur le plan F1-F2

Les projections les plus fiables sont, sur la figure 2.4 et en règle générale, périphériques, avec une excellente qualité de représentation aux extrémités de l'axe 1, ce qui est normal puisque les axes sont principalement liés aux coordonnées et corrélations extrêmes.

4. Quelques conseils de bon usage

L'objectif étant de maîtriser la compréhension du problème posé à travers un tableau de données numériques, un certain nombre de façons de procéder sont à éviter ou à recommander : il existe donc quelques principes de bon usage.

- pour un utilisateur, le problème à résoudre est *sémantique* et non technique. Il est abordé via la constitution d'un tableau de données, à transformer en information synthétisée, simplifiée, hiérarchisée mais la plus fidèle possible aux données initiales. En conséquence, l'utilisateur doit avoir une connaissance approfondie de la signification des variables qu'il a mis en tableau ; pas de données sans *métadonnées* (décrivant leur définition, les conditions de relevé, leur champ spatial et temporel, leur source, leur fiabilité, etc). On doit aussi toujours, pour les interprétations, se reporter au tableau des données (tout en étant conscient de leurs limites).

- L'ACP est une méthode de simplification à partir des *relations linéaires* entre variables. Il faut donc absolument éviter la *co-linéarité* entre elles. Par exemple, si pour un ensemble de lieux, certaines variables sont les secteurs d'emploi (primaire, secondaire, tertiaire) en % de l'emploi dans chaque lieu, il conviendrait de n'en retenir que 2 des 3 car le 3^{ième} est toujours complémentaire à 100% (et crée donc une relation artificielle).

- Considérée d'un point de vue géométrique, l'ACP est une méthode de projection d'un nuage de points multidimensionnel « biscornu » sur un sous espace orthonormé de faible dimension : il y a nécessairement déformation et c'est pourquoi il faut prêter grande attention aux qualités de représentation (mesure de distance à un axe).

- Même si la constitution d'un tableau de données reflète des hypothèses sur le phénomène à analyser, l'ACP doit être utilisée comme une méthode ***heuristique***, capable de faire découvrir de l'imprévu par interaction avec les données. Dans cette optique, on peut / doit recommencer l'analyse en supprimant variables et/ou individus ayant trop fortement pesé sur les résultats : le calcul des ***contributions*** à l'axe (variance de la projection / variance de l'axe), fournies par certains logiciels, permet de repérer variables ou unités statistiques trop prégnantes. Recommencer l'analyse sans eux constitue une sorte de ***zoom sémantique et/ou géographique.*** Dans l'exemple développé ci-dessus, on aurait pu retirer Pnb/hb et mort_inf pour mieux différencier au centre du nuage de points. Dans les cas d'études démographiques ou socio-économiques par région, il est ainsi fréquent de retirer (pour des raisons opposées) la Corse et l'Ile de France pour mieux percevoir la position des autres régions.

- On peut retirer des variables ou individus et recommencer l'analyse mais on peut aussi, sans la recommencer, considérer des variables ou individus ***supplémentaires***.

    - Si la variable est quantitative, on calculera sa corrélation (ou covariance) avec chacun des q axes retenus et on en positionnera les vecteurs dans le cercle de corrélations (cf Figure 2.2). On pourrait ainsi dans l'exemple des 27 pays européens représenter le vecteur corrélation avec F1-F2 du taux annuel de croissance démographique.
    - Si la variable supplémentaire est qualitative connue par classes, on fera figurer sur le plan des axes 1 et 2 le point moyen des unités statistiques des classes de la variable qualitative. Toujours dans le même exemple, on pourrait différencier anciens et nouveaux pays de l'Union Européenne et ajouter à la figure 2.3 les points moyens des coordonnées sur F1-F2 des 15 anciens et des 10 nouveaux membres. On pourrait, en outre, tracer autour de ces 2 points moyens un cercle correspondant à un ou deux écarts type.

    - certains logiciels fournissent aussi la possibilité de projeter sur les axes des individus supplémentaires (la Corse et l'Ile de France de l'exemple évoqué).

- Pour faciliter l'interprétation, des graphiques sont toujours préférables à des tableaux de résultats, d'autant plus d'ailleurs que n et/ou p sont grands. Toujours dans le but de faciliter l'interprétation, certains logiciels donnent la possibilité de pratiquer une rotation des axes factoriels (module ***VARIMAX*** par exemple) qui, tout en restant orthogonaux, maximisent les valeurs de quelques éléments et minimisent tous les autres. Ces rotations posant des problèmes théoriques et numériques doivent être considérées avec grande prudence : on peut à la rigueur les pratiquer si les % de variance des quelques premiers axes sont très voisins.

- Enfin, l'ACP (et l'analyse statistique en général) est non un substitut à la culture disciplinaire mais au contraire un ***moyen de la valoriser*** : sans compréhension et connaissance géographique, quelle interprétation géographique ***pertinente*** faire des résultats de l'analyse ? L'ACP (comme toute analyse menée rigoureusement) est un « ***antidéconnomètre*** », évitant le verbiage invalidable d'une géographie » ***gazeuse*** ».

B) EXERCICES CORRIGES

Exercice 1

Le tableau 2.15 fournit, pour 15 pays d'Afrique de l'ouest, 5 variables démographiques (espérance de vie masculine à la naissance, taux de natalité, de mortalité, de mortalité infantile, indice conjoncturel de fécondité), 1 variable sociale (PNB/habitant) et 1 variable sanitaire (VIH% : proportion des 15-54 ans atteint pas le virus du SIDA).

	espvieH	PNB/h	VIH%	Natalité	Mortalité	mort_infant	Fecondité
Bénin	49	920	2.5	45	15	94	6.3
Burkina	47	960	6.4	47	17	105	6.8
Cote d'Ivoire	45	1540	10.8	36	16	112	5.2
Gambie	51	1550	2	43	14	82	5.9
Ghana	56	1850	3.6	32	10	56	4.3
Guinée	43	1870	1.5	41	19	98	5.5
Guinée-Bissau	44	630	2.5	42	20	131	5.8
Liberia	49	1500	2.8	49	17	139	6.6
Mali	45	740	2	50	20	123	7
Mauritanie	49	1550	1.5	43	15	106	6
Niger	41	740	1.4	53	24	123	7.5
Nigeria	52	770	5.1	41	14	75	5.8
Sénégal	51	1400	1.8	41	13	68	5.7
Sierra Leone	42	440	3	47	20	153	6.3
Togo	53	1380	6	40	11	80	5.8

(source : ONU, 2001)

Tableau 2.15 : 7 variables pour 15 pays ouest africains

- *Aides à l'interprétation de l'ACP*

	espvieH	PNB/h	VIH%	Natalité	Mortalité	mort_infant	Fecondité
moyenne	47.80	1189	3.527	43.33	16.333	103.00	6.033
Ecart type	4.39	466	2.578	5.42	3.792	27.73	0.775
Coef de variation	0.092	0.392	0.731	0.125	0.232	0.269	0.128

Tableau 2.16 : paramètres résumant les 7 distributions

r	espvieH	PNB/h	VIH%	Natalité	Mortalité	mort_infant	Fecondité
espvieH	1.00	0.48	0.11	-0.58	***-0.94***	***-0.80***	-0.55
PNB/h		1.00	0.08	-0.59	-0.57	-0.53	-0.62
VIH%			1.00	-0.42	-0.28	-0.08	-0.29
Natalité				1.00	0.73	0.65	**0.98**
Mortalité					1.00	**0.80**	0.69
mort_infant						1.00	0.59
Fécondité							1.00

Tableau 2.17: matrice de corrélations entre variables

- *% de variance*: axe F1=64%, axe F2=16%, axe F3=10%

- *aides à l'interprétation*

r	F1	F2	F3
espvieH	***-0.84***	-0.33	-0.37
PNB/h	***-0.71***	-0.11	**0.50**
VIH%	-0.32	**0.86**	-0.33
Natalité	**0.90**	-0.29	-0.20
Mortalité	**0.93**	0.13	0.27
mort_infant	**0.84**	0.31	0.17
Fecondité	**0.87**	-0.22	-0.32

Tableau 2.18 : corrélations variables-axes factoriels

	F1	QR	CTR	*F2*	QR	CTR	*F3*	QR	CTR
Bénin	0.14	0.02	0.00	-0.67	0.41	0.41	-0.70	0.45	0.67
Burkina	0.90	0.25	0.18	0.69	0.15	0.43	***-1.13***	0.40	1.73
Cote d'Ivoire	-1.41	0.15	0.44	3.22	0.81	9.48	0.40	0.01	0.22
Gambie	***-1.17***	0.52	0.31	***-1.07***	0.44	1.05	0.10	0.00	0.01
Ghana	***-4.61***	0.96	4.75	-0.37	0.01	0.12	0.48	0.01	0.31
Guinée	-0.17	0.01	0.01	-0.16	0.00	0.02	2.24	0.91	6.83
Guinée-Bissau	1.46	0.45	0.48	0.64	0.09	0.37	0.52	0.06	0.36
Liberia	1.09	0.28	0.27	-0.45	0.05	0.18	0.19	0.01	0.05
Mali	**2.51**	0.94	1.41	-0.48	0.03	0.21	-0.37	0.02	0.19
Mauritanie	-0.42	0.11	0.04	-0.82	0.44	0.62	0.60	0.23	0.48
Niger	**3.92**	0.93	3.43	-0.55	0.02	0.27	0.09	0.00	0.01
Nigeria	-1.19	0.35	0.32	0.11	0.00	0.01	***-1.40***	0.49	2.68
Sénégal	***-1.65***	0.65	0.61	***-1.13***	0.30	1.16	-0.07	0.00	0.01
Sierra Leone	**2.77**	0.78	1.71	0.83	0.07	0.63	0.09	0.00	0.01
Togo	***-2.17***	0.78	1.05	0.20	0.01	0.04	-1.03	0.17	1.44

Tableau 2.19 : Coordonnées, Qualité de représentation et Contribution sur F1,F2,F3

Questions

1) **Quelles sont les hypothèses justifiant la construction du tableau 2.16 ?**
2) **Quelles sont, dans les 15 pays ouest africains, les distributions les plus variables ?**
3) **Commentez les corrélations entre les 7 distributions (tab. 2.18)**
4) **Interprétez les 3 premiers axes factoriels**
5) **Construisez, sur le plan des axes 1 et 2, une typologie et cartographiez la**

Ebauches de réponses

Question 1

Le tableau 2.15 a été constitué pour répondre principalement à 3 questions :

- A quel niveau (différent) de transition démographique en sont ces quinze pays en voie de développement ? Y a t'il entre eux un fort différentiel ?
- On présume un fort lien des régimes démographiques avec le revenu moyen dans ces pays : sur quelles variables joue t'il le plus ? le moins ?
- On sait l'Afrique (australe surtout, il est vrai) très touchée par le virus du SIDA et disposant de faibles moyens médicaux pour le combattre : cela se lit il déjà sur les mortalités et espérances de vie ?

Question 2

Pour répondre à cette question, on doit considérer dans le tableau 2.16b non pas les écarts type (car les unités des variables sont fort diverses) mais les coefficients de variation (σ/m) qui traduisent des variabilités rapportées aux moyennes. Il est clair, alors, que la plus grande différence entre les 15 pays est celle des % de personnes atteintes de SIDA (taux très fort en Côte d'Ivoire) et, secondairement, celle du revenu moyen par habitant. Par contre, il y a relative homogénéité démographique (surtout faibles espérances de vie).

Question 3

Le graphe de corrélation (Fig. 2.5) présente les plus fortes corrélations (négatives et positives).

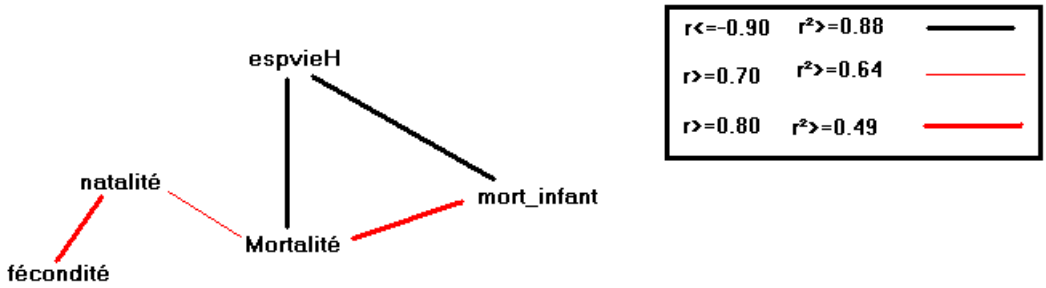


Figure 2.5: graphe des fortes corrélations correspondant au tableau 2.17

Les fortes corrélations concernent les 5 variables démographiques.

- La très forte corrélation entre natalité et fécondité semble aller de soi mais, le taux de natalité étant fonction de la fécondité des femmes *et* de la pyramide des âges de la population, c'est parce que ces 15 pays ont tous une structure d'âge jeune que la corrélation est si forte,
- La très forte corrélation entre taux de mortalités générale et infantile traduit surtout la faiblesse du système sanitaire dans certains des 15 pays,
- La forte corrélation entre taux de natalité et de mortalité (relative proportionnalité entre eux) traduit la plus ou moins grande avancée des 15 pays vers la transition démographique, de pays où elle est à peine entamée (très fortes natalité et mortalité, comme au Niger ou au Mali) à ceux où elle est en cours (Ghana, par exemple où les deux taux sont relativement bas).
- Les très fortes corrélations négatives entre mortalités générale et infantile d'une part, espérance de vie masculine de l'autre est logique.

On est surpris de :

- l'absence de corrélation notable entre proportion de personnes atteintes du SIDA et variables démographiques (essentiellement mortalité et espérance de vie) : il est vrai qu'un seul pays, la Côte d'Ivoire, a un VIH% très important,
- la relative faiblesse des corrélations inverses entre PNB/habitant et variables démographiques (de –0.48 à –0.62 soit de 23% à 38% seulement de variance expliquée) : le PNB/hb n'est pas la mesure idoine des niveaux de vie du plus grand nombre dans les 15 pays considérés et le niveau de vie n'est pas le seul ingrédient des comportements démographiques et des structures d'âge qui en résultent.

Question 4

Interprétation des axes factoriels

Le nombre de pays et de variables étant petit, on peut se passer de construire scalogrammes et cercles de corrélation en consultant, pour chaque axe dans les tableaux 2.18 et 2.19, variables à corrélation notable, pays à coordonnées fortes en valeur absolue avec bonne qualité de représentation et contribution à la variance point trop faible.

- L'axe factoriel 1 représente presque les 2/3 de l'information du tableau 2.15 (64%). Il oppose, en négatif (mais le signe ne signifie rien en analyse factorielle) pays à espérance de vie et PNB/hb relativement forts (Ghana, Togo, Sénégal) et pays à fortes mortalités, natalité et fécondité (Niger, Sierra Leone, Mali, Guinée Bissau). Ce 1er axe est donc clairement un axe opposant relatif développement démographique et économique d'une part, tout début d'évolution vers la transition démographique de l'autre.
- L'axe 2 (seulement 1/6 de l'information du tableau 2.15 et 4 fois moins que l'axe 1) est fondé sur les résidus non pris en compte par l'axe 1 : il met surtout en évidence la variable « virus du SIDA » et des pays, fort atteint comme la Côte d'Ivoire ou faiblement comme Sénégal et Gambie.
- L'axe 3 (10% de l'information) fondé sur des résidus de résidus insiste donc sur un pays mal représenté par les deux premiers axes , la Guinée marquée par un assez fort PNB/hb.

Question 5

1) *Typologie sur le plan des axes F1-F2 et cartographie*

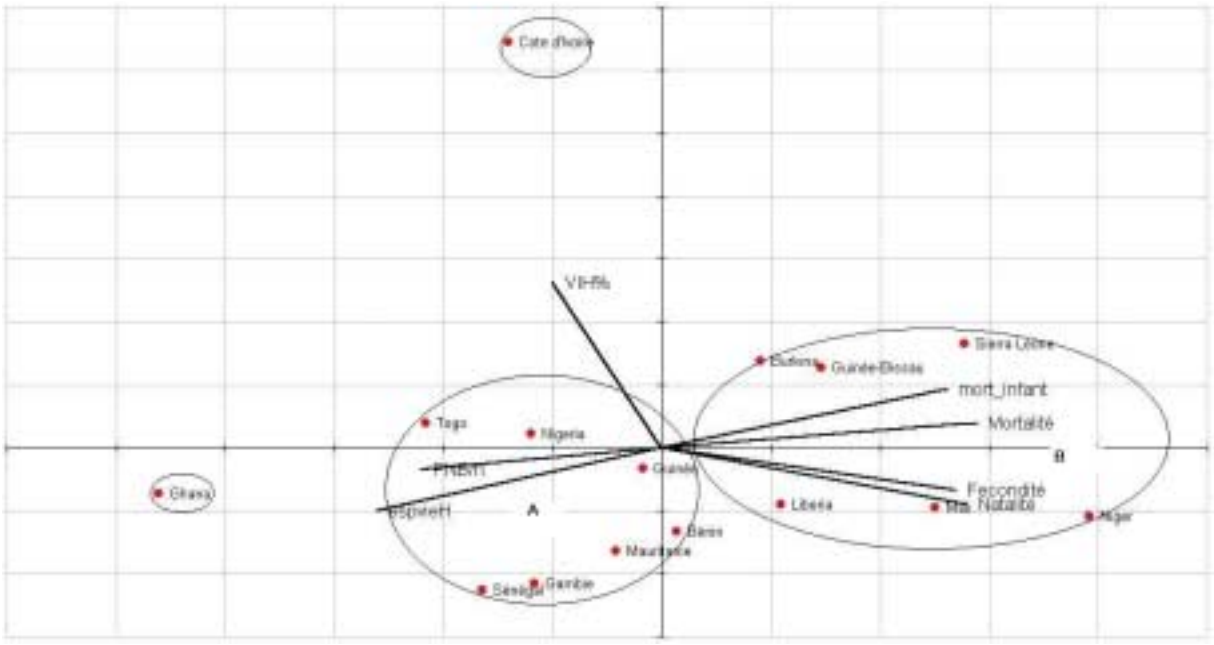


Figure 2.6: projection des pays sur le plan F1-F2 et vecteurs de corrélation des variables

Deux pays apparaissent très particuliers : la Côte d'Ivoire (à cause de sa proportion d'atteints du virus du SIDA) et le Ghana (pour son niveau relativement élevé de développement, démographique et économique).
Deux classes se différencient :

- La *classe A*, globalement caractérisée par des paramètres démographiques plus avancés vers la transition démographique et le développement,

- La *classe B*, globalement caractérisée par une démographie encore largement archaïque et un moindre développement.

La carte correspondant à cette typologie est présentée Figure 2.7.

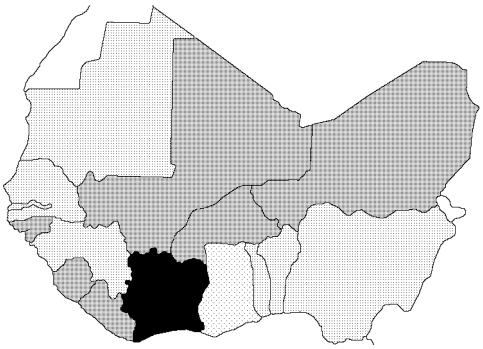


Figure 2.7: cartographie correspondant aux 4 classes de la figure 2.8

La figure 2.8 représente les écarts à la moyenne des 4 classes pour les 7 variables .

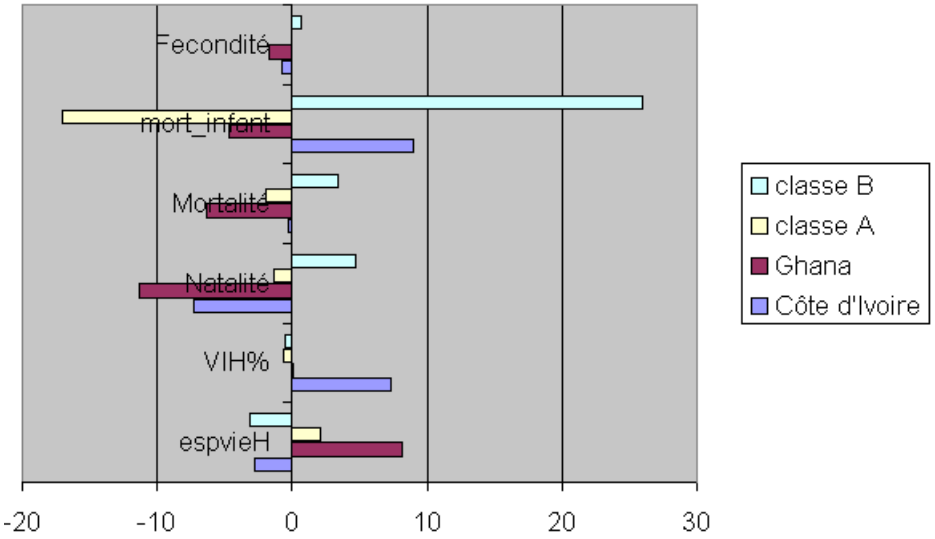


Figure 2.8: écarts aux moyennes des 15 pays pour les 7 variables du tableau 2.15

2) *En guise de conclusion :*

Revenons aux questions initialement posées :

- Oui il existe en Afrique de l'ouest un assez fort différentiel dans l'évolution démographique des pays, de ceux à très fortes fécondité, natalité, mortalité (comme le Niger ou le Mali) à ceux où elles ont déjà notablement évolué (Ghana, Nigeria, Sénégal par exemple).

- Le lien entre PNB/hb et régime démographique n'est pas absent (corrélations de moyenne intensité) mais il n'est pas totalement explicatif (aussi parce que le PNB/hb n'est sans doute pas une bonne mesure du revenu moyen et de sa distribution).

- Un seul pays, la Côte d'Ivoire, affiche un très fort taux d'adultes atteints du virus du SIDA et il est difficile de savoir, à macro-échelle, son rôle dans les mortalités (que seraient elles en Côte d'Ivoire sans cette forte atteinte ?).

Exercice 2

Ce 2nd exercice propose des variables simples de dynamique démographique : taux de croissances naturelle ((naissances - décès)/population) et migratoire ((immigrants - émigrants)/population) pour 6 périodes intercensitaires de 1954 à 1999. Comme ces périodes sont de durées inégales, ces taux sont ramenés à des taux annuels (exprimés en % par an).

Dans le tableau 2.20, les variables, en colonne, sont repérées par deux caractères : le 1er repère le type de croissance (**N** pour naturelle, **M** pour migratoire) et le 2nd la période (**1** : 1954-62, **2** : 1962-68, **3** : 1968-75, **4** : 1975-82, **5** : 1982-90, **6** : 1990-99). Les lignes du tableau sont relatives aux 6 départements de la région Provence-Alpes-Côte d'Azur et aux 8 de la région Rhône-Alpes. Chaque département est identifié par son numéro minéralogique.

	N1	M1	N2	M2	N3	M3	N4	M4	N5	M5	N6	M6
01 : Ain	0.5	0.9	0.4	0.1	0.4	1.0	0.5	1.1	0.4	1.0	0.5	0.6
04 : Alpes de Hte Provence	0.5	1.7	0.2	1.0	0.2	0.8	-0.1	1.0	0.1	1.1	0.0	0.7
05 : Hautes Alpes	0.6	0.2	0.6	0.0	0.4	0.5	0.2	0.9	0.3	0.6	0.2	0.6
06 : Alpes Maritimes	0.0	2.6	0.0	2.4	-0.2	1.9	-0.2	1.3	-0.1	1.3	-0.1	0.5
07 : Ardèche	0.3	0.2	0.2	-0.3	0.2	-0.2	-0.1	0.7	0.0	0.4	0.0	0.3
13 : Bouches du Rhône	0.6	2.1	0.6	1.7	0.5	1.1	0.3	0.5	0.4	-0.1	0.3	0.2
26 : Drôme	0.7	1.3	0.5	0.8	0.5	0.2	0.3	0.7	0.4	0.3	0.4	0.3
38 : Isère	0.8	1.3	0.7	1.2	0.8	0.9	0.6	0.6	0.6	0.4	0.6	0.3
42 : Loire	0.6	0.0	0.5	0.2	0.5	-0.1	0.3	-0.4	0.3	-0.2	0.2	-0.5
69 : Rhône	0.9	1.1	0.6	1.4	0.8	0.3	0.7	-0.5	0.7	-0.2	0.7	-0.2
73 : Savoie	0.7	0.7	0.7	0.1	0.5	0.2	0.4	0.5	0.4	0.5	0.4	0.4
74 : Hte Savoie	0.9	1.4	0.8	1.0	0.8	1.7	0.6	0.8	0.7	1.1	0.7	0.5
83 : Var	0.5	2.4	0.5	1.8	0.3	1.5	0.1	1.7	0.2	1.6	0.1	1.0
84 : Vaucluse	0.7	1.9	0.5	1.2	0.5	0.9	0.3	1.0	0.4	0.7	0.3	0.4

(source : INSEE)

Tableau 2.20 : Croîts naturels et migratoires des départements du Sud Est de 1954 à 1999

- *Moyennes et écarts type*

	N1	N2	N3	N4	N5	N6	M1	M2	M3	M4	M5	M6
Mye	0.59	0.49	0.44	0.28	0.34	0.31	1.27	0.90	0.76	0.71	0.61	0.36
Ect	0.24	0.22	0.27	0.28	0.24	0.26	0.82	0.79	0.65	0.59	0.56	0.37

Tableau 2.21 : moyennes et écarts type des taux de croît naturel et migratoire

- *Aides à l'interprétation*

	N1	M1	N2	M2	N3	M3	N4	M4	N5	M5	N6	M6
N1	1.00	-0.24	**0.89**	-0.16	**0.96**	-0.21	**0.86**	-0.48	**0.94**	-0.41	**0.86**	-0.29
M1		1.00	-0.23	**0.91**	-0.29	**0.80**	-0.23	0.55	-0.14	0.51	-0.12	0.46
N2			1.00	-0.16	**0.90**	-0.08	**0.83**	-0.34	**0.88**	-0.33	**0.81**	-0.20
M2				1.00	-0.16	**0.73**	-0.11	0.26	-0.03	0.28	-0.03	0.16
N3					1.00	-0.21	**0.92**	-0.54	**0.96**	-0.49	**0.91**	-0.40
M3						1.00	-0.08	**0.63**	-0.03	**0.70**	0.02	0.56
N4							1.00	-0.50	**0.97**	-0.41	**0.98**	-0.38
M4								1.00	-0.47	**0.89**	-0.42	**0.93**
N5									1.00	-0.38	**0.98**	-0.32
M5										1.00	-0.33	**0.89**
N6											1.00	-0.29
M6												1.00

Tableau 2.22: Inter-corrélations (en gras si |r| >0.6) des variables du tableau 2.22

F1	F2	F3	F1+F2+F3
55.7%	26.5%	11.5%	93.7%

Tableau 2.23 : % de variance des 3 premiers axes de l'ACP

	N1	M1	N2	M2	N3	M3	N4	M4	N5	M5	N6	M6
axe 1	0.89	-0.50	0.82	-0.33	0.94	-0.44	0.90	-0.75	0.89	-0.7	0.85	-0.62
axe 2	0.31	0.71	0.37	0.63	0.28	0.82	0.35	0.46	0.44	0.52	0.46	0.51
axe 3	-0.11	0.37	-0.19	0.58	-0.06	0.12	-0.03	-0.35	-0.02	-0.34	-0.02	-0.47

Tableau 2.24 : Corrélations des 12 variables avec les 3 premiers axes de l'ACP

	axe 1	axe 2	axe 3	QR1	QR2	QR3	QR1+QR2+QR3
01	-0.13	0.39	***-1.42***	0.00	0.03	*0.46*	***0.50***
04	***-2.95***	-0.38	-0.25	*0.87*	0.01	0.01	0.89
05	0.06	-1.02	***-1.71***	0.00	0.23	*0.65*	0.88
06	***-5.95***	0.42	1.97	*0.88*	0.00	0.10	0.98
07	-1.82	***-3.59***	-0.88	0.19	*0.75*	0.05	0.99
13	0.58	0.64	**1.79**	0.07	0.08	*0.63*	***0.77***
26	0.82	-0.43	0.05	0.42	0.11	0.00	***0.54***
38	**2.46**	1.42	0.14	*0.73*	0.24	0.00	0.98
42	2.14	***-3.33***	1.02	0.26	*0.63*	0.06	0.96
69	**4.23**	0.05	1.75	*0.83*	0.00	0.14	0.97
73	1.47	-0.64	-1.12	0.51	0.09	0.29	0.90
74	2.22	**3.04**	-0.87	0.31	*0.59*	0.05	0.95
83	***-3.09***	2.39	-0.61	*0.58*	0.35	0.02	0.96
84	-0.05	**1.04**	0.14	0.00	*0.71*	0.01	***0.72***

Tableau 2.25 : Coordonnées et qualité de représentation sur les 3 premiers axes de l'ACP

Questions

1) **Quel est le type de ce tableau ? Pourquoi peut on le résumer par une ACP ? Pourquoi n'y avoir pas fait figurer les taux de croissance (naturelle + migratoire) ?**
2) **Commentez tableau de données, moyennes et écarts type des taux de croissance**
3) **Construisez et commentez le graphe des inter-corrélations des variables**
4) **Interprétez les trois premiers axes factoriels**
5) **Faites une typologie des 14 départements et la carte thématique correspondante**

Corrigé suggéré

Question 1

On a affaire ici à 12 variables quantitatives continues : on peut donc les résumer par ACP
Ce sont des taux issus de la division de variables de comptage (naissances, décès, immigrants, émigrants, population). Ces taux sont relatifs à des départements, qui sont des unités statistiques secondaires (agrégats de personnes).
Le tableau lui même est une matrice spatio-chronologique comportant 2 caractères (croissance naturelle, croissance migratoire) connus pour 6 périodes. Ajouter pour chaque période les taux de croissance totale (=croissance naturelle + croissance migratoire) aurait ajouté, à chaque période, un caractère parfaitement co-linéaire avec les deux autres (strictement égal à leur somme), ce qui aurait crée des corrélations artificielles et aurait donc biaisé l'analyse.
L'ACP pratiquée ici est une ACP sur variables standardisées (centrées-réduites) car on s'intéresse aux taux de croissance non pas absolus mais comparativement à ceux des autres

départements à chaque période. Les informations relatives aux différences chronologiques de taux de croissance doivent donc être commentées avant l'ACP, qui effacera leurs effets.

Question 2

Puisque l'ACP centrée-réduite fera disparaître les ordres de grandeur des taux eux mêmes, il est utile de commenter le tableau des données initiales.

Globalement, les deux régions Provence-Alpes-Côte d'Azur et Rhône-Alpes ont été parmi les régions françaises les plus favorisées, en termes de croissance démographique, au cours des 50 dernières années, mais avec de notables différences internes.

Si l'on s'intéresse tout d'abord au tableau 2.21 (moyennes et écarts type des taux départementaux), on observe qu'en 45 ans, de 1954 à 1962 :

- Les taux de croissance naturelle ont baissé d'environ la moitié (de 0.59% à 0.31% par an), ce qui est conforme à ce que l'on sait par ailleurs de l'évolution de la démographie française. Les années 1970 ont été le moment d'une chute relativement brutale.

- La variabilité (écarts type) interdépartementale de ces taux est restée, sur les 45 ans, à peu près constante et relativement faible mais comme les taux eux mêmes ont beaucoup baissé, l'écart *relatif* entre départements s'est accru (sans qu'on puisse calculer de coefficients de variation car certains taux sont négatifs).

- Les taux de croissance migratoire, supérieurs à toute période aux taux de croissance naturelle (mais de moins en moins), ont eux beaucoup baissé. Ils représentent aujourd'hui moins du tiers de ce qu'ils étaient dans les années 1950 et leur diminution brutale des années 1990, est sans doute signe d'une baisse globale d'attractivité.

- La variabilité inter départementale des taux de croissance migratoire a diminué, mais moins que leur moyenne, ce qui implique un accroissement *relatif* des différences.

Commenter moyennes et écarts type permet de caractériser la chronologie mais non la différenciation spatiale ; pour ce faire, on peut constituer, par calcul, les données du tableau 2.26.

Dpts	Croît total en 45ans	Croît naturel	Croît migratoire	Naturel / Migratoire
01	***75.0%***	*22.6%*	*52.4%*	30.1%
04	*48.2%*	*6.9%*	*41.3%*	14.3%
05	*46.7%*	*18.2%*	*28.5%*	39.0%
06	***96.2%***	***-4.4%***	*100.6%*	***-4.6%***
07	*14.3%*	*4.3%*	*10.0%*	30.1%
13	***79.6%***	*22.0%*	*57.6%*	27.6%
26	***60.5%***	*23.2%*	*37.3%*	38.3%
38	***90.0%***	*35.7%*	*54.3%*	39.7%
42	*9.4%*	*19.1%*	*-9.7%*	**203.2%**
69	*57.4%*	*39.1%*	*18.3%*	68.1%
73	*51.3%*	*25.7%*	*25.6%*	50.1%
74	***124.6%***	*39.8%*	*84.8%*	31.9%
83	***135.2%***	*13.0%*	*122.2%*	9.6%
84	***90.4%***	*22.1%*	*68.3%*	24.5%

Tableau 2.26 : Croissances naturelle et migratoire 1954-99

Quelques remarques d'abord sur la constitution de ce tableau :

- Les taux du tableau initial ayant été fortement arrondis (à 1 chiffre après la virgule), nous n'avons là que des valeurs globales approximées,
- Le calcul des taux de croissance sur les 45 ans a été effectué de façon multiplicative (et non additive, comme le font la plupart des étudiants) et en tenant compte de l'inégale durée des périodes intercensitaires (8, 6, 7, 7, 8 et 9 ans). Aux approximations d'arrondi près, les taux résultants sont donc exacts.

Au vu du tableau 2.26, 4 grands groupes peuvent être distingués :

- Un groupe de départements méditerranéens à croissance démographique forte, essentiellement due aux soldes migratoires (Var, Alpes Maritimes, Vaucluse et, dans une moindre mesure, Bouches du Rhône),
- Un groupe de départements situés à l'Est de Lyon, à croissances migratoire et naturelle assez équilibrée (Haute Savoie, Isère, Ain, Drôme) ; le cas du département du Rhône est assez particulier car une bonne partie de la croissance périurbaine de l'agglomération lyonnaise s'est faite sur l'Isère et l'Ain.
- Un groupe de départements « alpin rural sans grande ville », à croissance modérée, comprenant Alpes de Haute Provence, Hautes Alpes (à évolution plutôt « méditerranéenne ») et Savoie (à évolution plutôt de type « Est lyonnais »),
- Les deux départements du rebord du Massif Central (Loire et Ardèche) ont connu une croissance très faible avec des soldes migratoires parfois déficitaires et des soldes naturels très faibles, voire nuls.

Le tableau de données initiales (Tableau 2.20) étant relativement petit, son interprétation directe était possible : sera t'elle confirmée par les aides à l'interprétation de l'ACP ?

Question 3

Le tableau des inter-corrélations entre les 12 variables peut être simplifié et représenté sous forme de graphe (Figure 2.9).

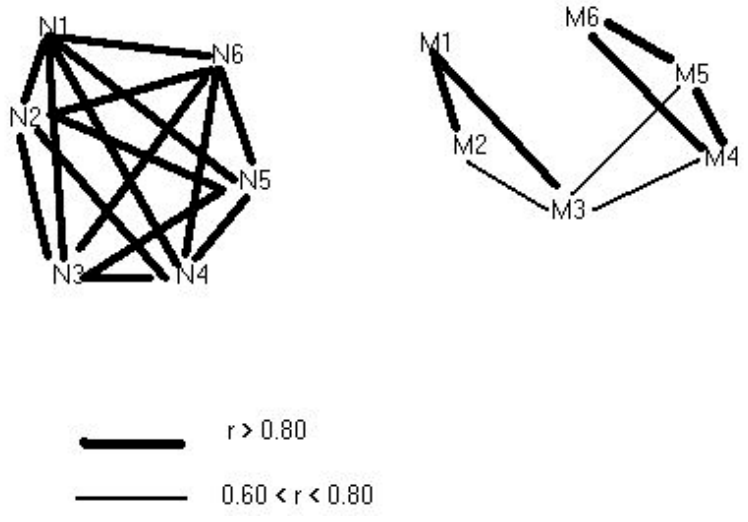


Figure 2.9 : graphe des corrélations principales du tableau 2.22

Le graphe de la figure 2.9 présente clairement deux sous graphes :

- Le 1er, très fortement interconnecté ($r>0.8$ à $r^2>64\%$), concerne tous les taux de croissance naturelle, à toutes les périodes, ce qui signifie que sur 45 ans ce sont pratiquement toujours les mêmes départements qui ont eu les taux relativement les plus forts (et les mêmes qui ont eu les taux les plus faibles). La croissance naturelle a, dans l'espace géographique à l'échelle départementale, une grande *rémanence* car elle dépend en grande partie des structures d'âge, jeunes ou vieillies.

- Le 2nd sous graphe est, lui, bien moins interconnecté et on peut le subdiviser en deux périodes. La 1ère va jusqu'au début des années 1970 (corrélation forte entre 1954-62 et 1962-68 puis plus faible avec 1968-75) : elle correspond à la toute fin de l'ère de l'exode rural. La 2nde période va des années 1970 à la fin du siècle et marque une certaine rupture dans les cartes de croissance migratoire. Les flux ont été partiellement réorientés vers des départements à espaces périurbains (le cas du Rhône est à cet égard significatif) et ruraux sans grande ville (voir l'exemple des hautes Alpes). Les croissances migratoires sont donc chronologiquement plus transitoires et éphémères que les croissances naturelles puisqu'elles dépendent davantage de facteurs liés à la conjoncture économique et à des modifications sociales.

- Normalement, les départements à fort solde migratoire positif en début de période (portant généralement sur les 20-50 ans, sauf migrations de retraite méditerranéenne) auraient dû voir leurs structures d'âge rajeunies et donc avoir des soldes naturels redressés en fin de période. Qu'en est il exactement ? Cela se marque t'il par des corrélations notables entre N1 et N2 avec M5 et M6 ? Cela n'est pas le cas général dans la mesure où ces inter-corrélations sont négatives et faibles. Le raisonnement vaut sans doute pour des espaces plus restreints que des départements, sur des zones très fortement périurbanisées et, peut être, pour des durées plus longues que le demi siècle.

Question 4

L'axe factoriel 1 contient 55.7% de l'information initiale, l'axe 2 26.5% (82.2% d'information sur les 2 premiers axes),l'axe 3 11.5% (93.7% d'information sur les 3 premiers axes).
L'ACP a donc pris en compte presque toute la variance du nuage de points (12-dmensionnel) dans un cube (de dimension 3) : le résumé est donc drastique et ceci est dû aux nombreuses fortes corrélations du tableau 2.23. Il est donc temps d'interpréter chaque axe.

Interprétation de l'axe 1

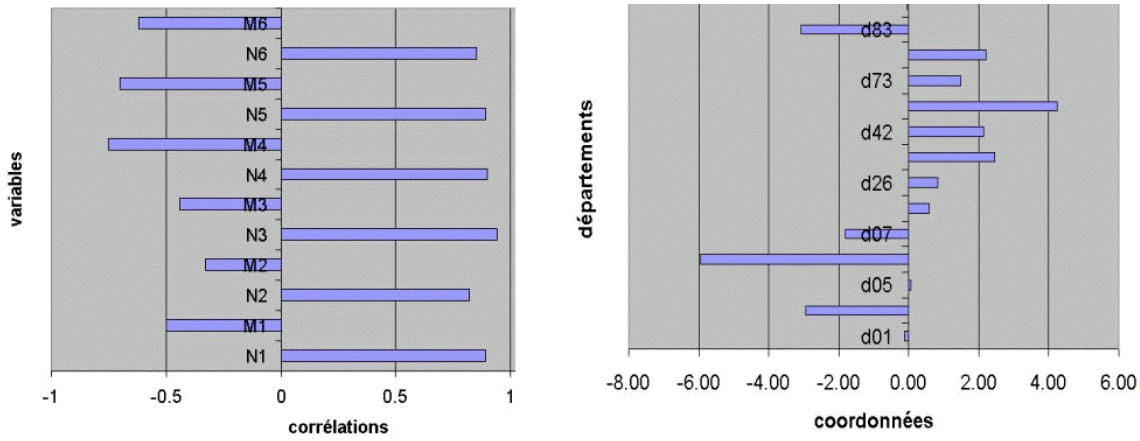


Figure 2.10 : Variables et départements sur l'axe factoriel 1

Si l'on se reporte au tableaux des corrélations des 12 variables avec l'axe 1 et à la figure 2.10, on voit que cet axe oppose tous les taux de croissance naturelle à ceux de croissance migratoire et, principalement à ceux d'après 1975. Cela confirme l'observation faite sur le graphe d'inter-corrélations des variables. Cet axe opposera donc des départements à fort solde migratoire – faible solde naturel à des départements à fort solde naturel – faible solde migratoire : l'opposition principale, si l'on considère aussi les bonnes qualités de représentation, est entre Alpes maritimes, Alpes de Haute Provence, Var d'une part, Rhône et Isère de l'autre.

Interprétation de l'axe 2

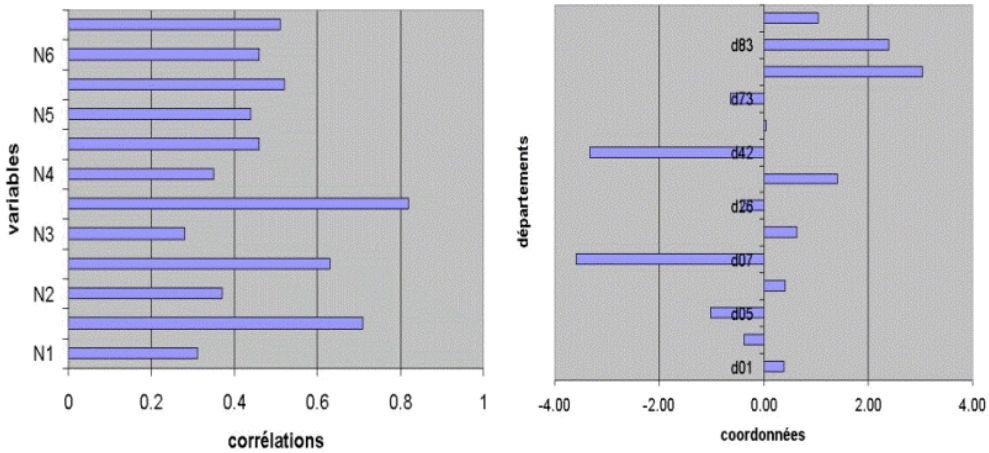


Figure 2.11 : Variables et départements sur l'axe factoriel 2

Toutes les corrélations des variables avec l'axe 2 sont positives (l'axe 2 est « unijambiste ») mais de façon inégale : les plus fortes corrélations concernent les soldes migratoires, principalement ceux d'avant 1975 qui étaient peu apparus sur le 1er axe. Les coordonnées et les qualités de représentation des départements sur cet axe sont moins fortes, ce qui est normal puisqu'on a affaire à des résidus du 1er axe. Néanmoins, cet axe oppose assez clairement des départements à forte croissance migratoire (surtout d'avant 1975) comme la Haute Savoie, le Vaucluse à des départements sans croissance migratoire comme l'Ardèche et la Loire.

Interprétation de l'axe 3

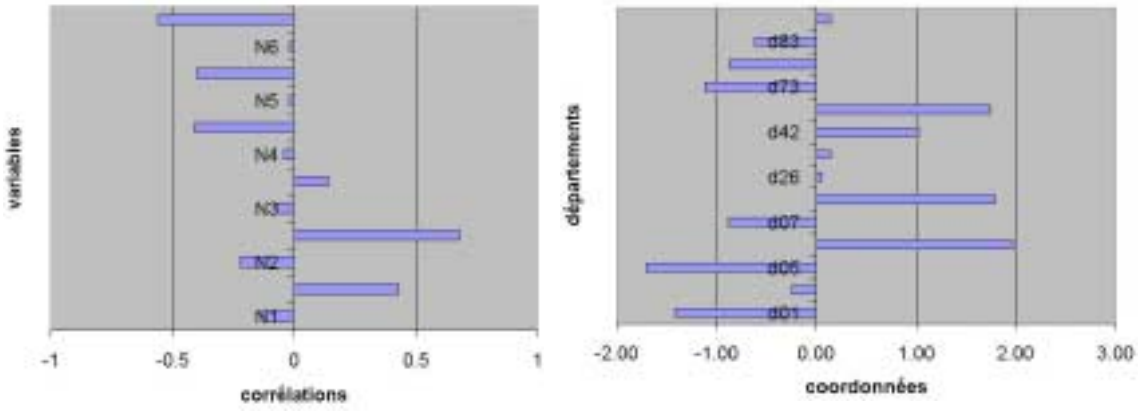


Figure 2.12: Variables et départements sur l'axe factoriel 3

Cet axe oppose les croissances migratoires d'avant 1968 à celles d'après 1982, la période intermédiaire étant de transition entre deux systèmes migratoires. Il oppose donc Bouches du Rhône et Rhône, à croissance migratoire ancienne à Ain et Hautes Alpes, à croissance migratoire récente.

En résumé, l'axe 1 est l'axe des croissances naturelles, l'axe 2 (moitié moins informatif) celui des croissances migratoires, surtout récentes, l'axe 3, quant à lui, oppose croissances migratoires récentes et anciennes.

Représentations synthétiques des 3 axes

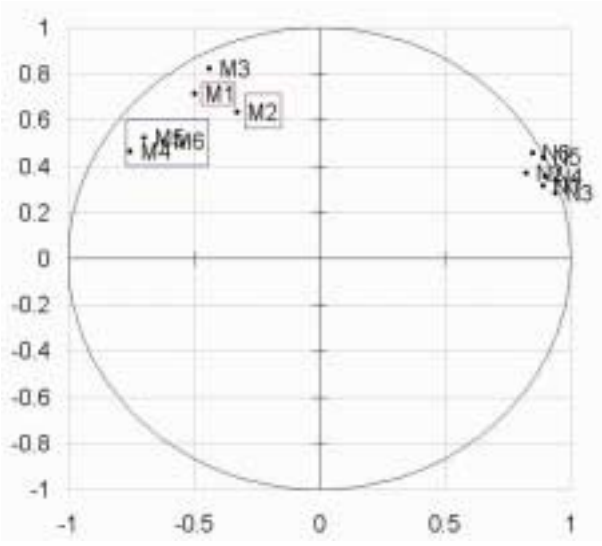


Figure 2.13: Cercle des corrélations sur le plan des axes 1 et 2 (axe 3 : encadrés)

Le cercle des corrélations met bien en évidence l'opposition croissances naturelles et croissances migratoires récentes sur l'axe 1 (abcisses), le rôle des croissances migratoires anciennes sur l'axe 2 (ordonnées), l'opposition des 2 périodes de croissance migratoire sur l'axe 3 (encadré rouge pour les coordonnées les plus positives, bleu pour les plus négatives).

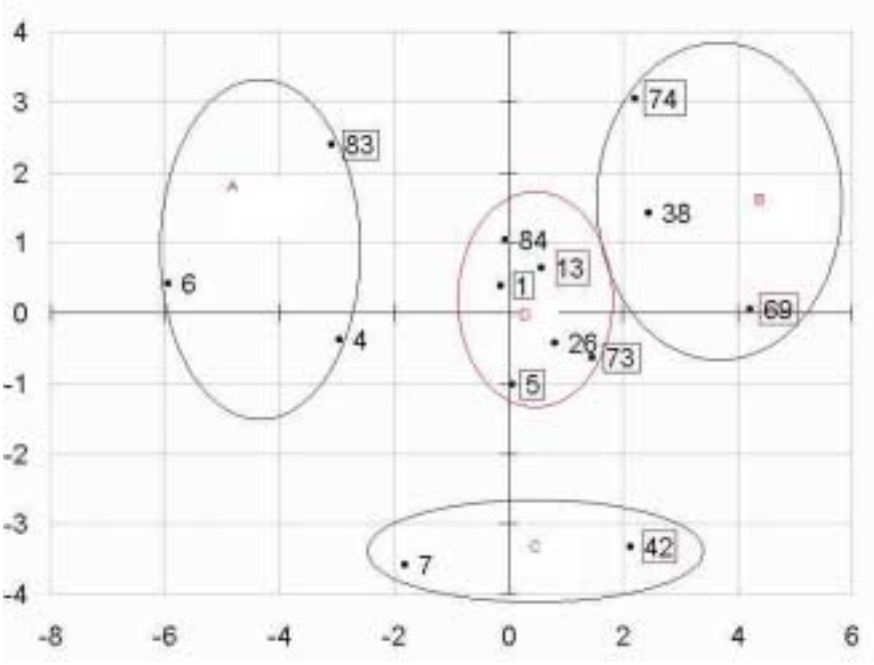


Figure 2.14 : Coordonnées des départements sur le plan des axes 1 et 2

La figure 2.14 permet de faire une typologie empirique des départements en 4 classes :

- *Classe A* : départements à taux de croissances naturelle faible et migratoire forte : Alpes maritimes, Var, Alpes de Haute Provence,
- *Classe B* : départements à fort taux de croissance naturelle : Haute Savoie, Isère, Rhône,
- *Classe C* : départements à taux de croissances naturelle et migratoire faibles : Loire, Ardèche,
- *Classe D* : tous les autres départements, à évolution peu marquée (moyenne) ou mal représentée par l'ACP (cas de la Drôme et de l'Ain).

Cette typologie constitue la légende d'une carte thématique multivariée (Figure 2.15).

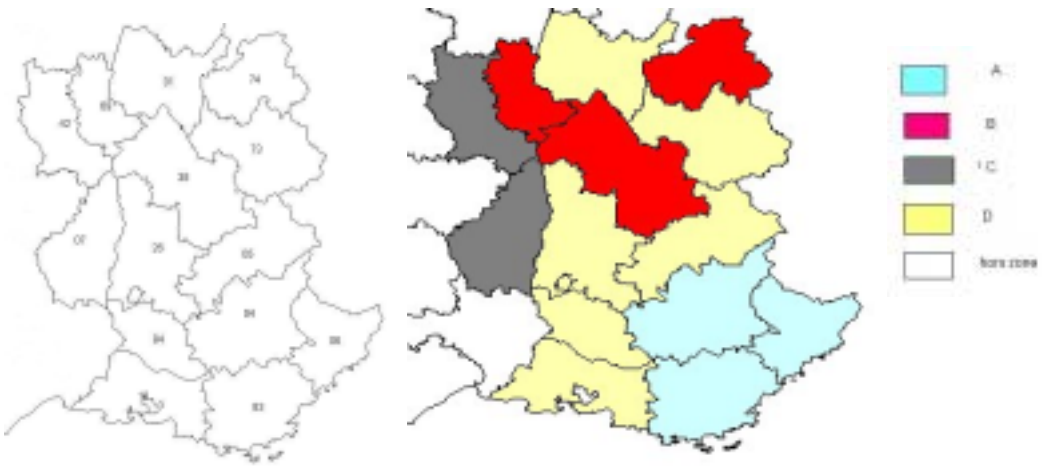


Figure 2.15 : Carte thématique issue des axes 1 et 2 de l'ACP

Chapitre 3

L'analyse factorielle des correspondances (AFC)

A) CONNAISSANCES DE BASE

L'analyse factorielle des correspondances (AFC) est une analyse factorielle adaptée à d'autres formes de tableaux que l'analyse en composantes principales : elle permet, notamment, de résumer des *caractères qualitatifs* (mais pas seulement).

1. Types de tableaux pour l'AFC

La forme prototypale est le tableau de contingence.

1.1 Tableau de contingence

Un tableau de contingence est l'expression d'une *distribution bivariée connue par classes*, croisant les modalités qualitatives ou les classes de valeurs quantitatives de deux caractères, l'un en ligne, l'autre en colonne. Chacune des cases ij du tableau contient *l'effectif* correspondant à la modalité i du caractère en ligne et à la modalité j du caractère en colonne.

- *Exemple*

Une enquête de 1995 auprès d'un échantillon représentatif de 1015 exploitants agricoles français a permis de relever leur revenu mensuel et leur âge, caractères quantitatifs présentés en classes et croisés dans le tableau 3.1.

revenu	< 40 ans	>= 40 ans	tout âge
<8 mFF	143	32	***175***
8-9 mff	237	160	**397**
9-10 m FF	58	165	**223**
10-15 m FF	31	109	**140**
plus	10	70	**80**
tout revenu	***479***	***536***	**1015**

(source : AGREST, 1995)

Tableau 3.1 : Revenu mensuel (en milliers de FF) d'exploitants agricoles selon l'âge

On a bien là un tableau de contingence car :
- il croise 2 caractères connus par classes,
- chaque case contient l'effectif correspondant à une classe de revenus et à une classe d'âges,
- l'addition des effectifs en ligne a un sens : elle fournit la distribution univariée du revenu (quel que soit l'âge),
- l'addition des effectifs en colonne a un sens : elle fournit la distribution univariée par âge (quel que soit le revenu),

- les effectifs concernent ici des personnes (unités statistiques primaires) mais pourraient, dans d'autres cas, concerner des unités statistiques secondaires (comme des agglomérations),
- Le tableau 3.1 est foncièrement *symétrique* : caractères en ligne et en colonne peuvent être permutés et jouer le même rôle.

1.2 Extension de la notion de tableau de contingence

On peut étendre la définition du tableau de contingence à tout tableau :

- Symétrique, croisant 2 caractères jouant un rôle équivalent (on peut donc les permuter)

- contenant des valeurs positives ou nulles

- pour le croisement de caractères où sommes en lignes et en colonnes ont un sens.

C'est le cas par exemple de tableaux de notes pour des élèves ou de tableaux de dépenses de ménages par poste de dépense. Tous ces tableaux relèvent de l'AFC.

1.3 Notations de base

On utilise la notation classique pour les tableaux de contingence (et le test du Khi²). Le tableau de données initial **D** sera donc appelé ici **N**.

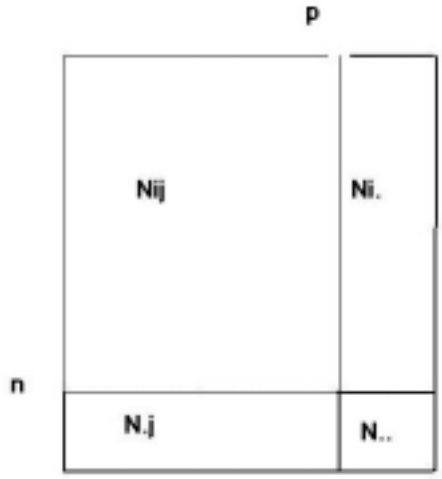


Tableau 3.2 : Notations de base

Le tableau 3.2 contient 4 types d'information : distribution bivariée croisant I et J, distribution univariée de I, distribution univariée de J, effectif total.

- *Valeurs des 4 parties du tableau*

  - $\mathbf{N_{ij}}$ repère la valeur correspondant à la case ligne i - colonne j du tableau (ainsi N_{32}=165 dans le tableau 3.1),
  - $\mathbf{N_{i.}}$ la somme des valeurs d'une ligne, le . signifie « pour toutes ses colonnes » ($N_{5.}$=80 dans cet exemple),
  - $\mathbf{N_{.j}}$ la somme des valeurs d'une colonne, « pour toutes ses lignes » ($N_{.2}$=536),
  - $\mathbf{N_{..}}$ la somme totale des valeurs du tableau, somme des $\mathbf{N_{i.}}$ ou des $\mathbf{N_{.j}}$ ($N_{..}$=1015).

- *Fréquences correspondantes*

  - $\mathbf{f_{ij} = N_{ij} / N_{..}}$ fréquence par rapport au total (f_{32} = 165 / 1015 = 0.163)
  - $\mathbf{f_{i.} = N_{i.} / N..}$ fréquence de la somme des valeurs ligne i ($f_{4.}$ = 140 / 1015 = 0.138)
  - $\mathbf{f_{.j} = N_{.j} / N..}$ fréquence de la somme des valeurs colonne j **(**$f_{.2}$=536/1015=0.528)

- *Fréquences conditionnelles*

  - $\mathbf{f_{ij/i} = N_{ij} / N_{i.}}$ fréquence en proportion du total de ligne i ($f_{32/3}$=165/223=0.740, signifie que par rapport au total de la ligne 3, l case 3-2 représente 74.0%)
  - $\mathbf{f_{ij/j} = N_{ij} / N_{.j}}$ fréquence en proportion du total de colonne j (f_{41}=31/479=0.065, signifie que par rapport au total de la colonne 1, la case 4-1 représente 6.5%).

2. Différences de l'AFC par rapport à l'ACP

L'AFC est, comme l'ACP, une analyse factorielle (cf chapitre 1.1) dont les objectifs généraux sont les mêmes:

- Résumer l'information contenue dans de grands tableaux numériques,
- A partir d'une représentation sous forme de nuage de points multidimensionnel, induire des résumés descriptifs hiérarchisés (axes factoriels),
- Leur donner une signification grâce à des aides à l'interprétation.

Comme l'ACP, on peut en faire un exposé géométrique, algébrique, informatique et/ou informationnel. Mais l'AFC porte sur un autre type de tableaux :

- symétrie et non dissymétrie des lignes et colonnes (ici, on croise 2 caractères connus par classes au lieu de juxtaposer des variables pour des individus),

- possibilité de variables qualitatives et non obligation de variables quantitatives.

En outre, quand il est possible de pratiquer les deux types d'analyse factorielle, les buts ne sont pas les mêmes car, en AFC, on ne corrèle pas des valeurs mais on compare des *profils* en lignes et en colonnes.

Nonobstant ces différences, la démarche générale est la même et, notamment, les trois modules de l'algorithme sont comparables.

2.1 Transformation des données et calcul des covariances

2.1.1 *Transformation du tableau de données*

Puisque les tableaux pour l'AFC sont fondamentalement *symétriques*, lignes et colonnes peuvent être permutées (« transposition » du tableau D de données) avec des résultats identiques. L'analyse procède de même sur les unes ou les autres : ce qui va être détaillé pour une AFC en colonnes peut être transposé pour une AFC en lignes (les logiciels choisissent l'une ou l'autre de façon à minimiser les calculs).

La procédure générale décrite figure 1.4 est la même :

- passage du tableau de données **D** en un tableau transformé **D'**
- puis calcul sur **D'** d'une matrice **C** de covariances

mais la transformation de D en D' est différente.

- *Construction du nuage de points : transformation du tableau **D** de données*

Pour une AFC sur les colonnes de D :

& le *poids* d'une ligne est égal à $\mathbf{f_{i.} = Ni./N..,}$ c'est la fréquence de la somme des valeurs de la ligne i sur l'effectif total. La somme de tous les $\mathbf{f_{i.}} = 1$.

& on transforme les valeurs de chaque ligne en *proportion* de leur total de ligne (fréquences conditionnelles en ligne). $\mathbf{D'_{ij} = f_{ij/i} = D_{ij} / D_{i.}}$ les totaux en ligne deviennent ainsi tous égaux à 1 et chaque ligne devient un profil en fréquence (%)

& La *moyenne* d'une colonne j fait intervenir poids et fréquences conditionnelles : c'est une moyenne pondérée par le poids des différentes lignes (en ACP, pas de pondération) :

$$m_j = \sum_{i=1}^{n} f_{i.} f_{ij/i} \qquad m_j = \sum_{i=1}^{n} f_{i.} f_{ij/i}$$

le centre de gravité **G** (pondéré par les masses des lignes) du nuage des n points-ligne a pour coordonnées $\mathbf{G = m_1, m_2, \ldots, m_p}$ si D a p colonnes et n lignes.

& la *variance* d'une colonne j est également pondérée par le poids des lignes (les $f_{i.}$):

$$\sigma_j^2 = \sum_{i=1}^{n} f_{i.} (f_{ij/i} - G_j)^2 \; o_j^2 = \sum_{i=1}^{n} f_{i.} (f_{ij/i} - G_j)^2$$

la variance totale du nuage de points est la somme de ces variances par colonne.

& Dans le nuage de n points construits pour représenter D', les axes d'allongement sont construits par *moindres carrés* des points aux axes mais en utilisant une autre métrique (mesure des distances). En ACP, on utilisait la métrique euclidienne, en AFC puisqu'on traite de tableaux assimilables aux tableaux de contingence, on utilise la métrique du *Khi²* (celle qu'on utilise pour le test du même nom).

Pour faire intuiter la différence entre métrique du Khi² (utilisée en AFC) et métrique euclidienne (utilisée en ACP), prenons un exemple d'école : considérons la répartition en 3 secteurs d'activité (1re, 2re, 3re) de la population active de 3 communes (A, B, C) fournie par le tableau 3.3.

	1re	**2re**	**3re**	***somme***
A	60	120	220	***400***
B	30	60	110	***200***
C	10	20	40	***70***
Somme	***100***	***200***	***370***	***670***

Tableau 3.3 : Exemple d'école

Calcul des distances euclidiennes AB et AC

Pour reproduire les conditions de l'ACP, nous standardisons les 3 variables (tableau 3.4)

secteur	1re	2re	3re
A	1.30	1.30	1.30
B	-0.16	-0.16	-0.18
C	-1.14	-1.14	-1.12

Tableau 3.4 : Colonnes standardisées du tableau 3.3

La distance euclidienne entre lignes i et k étant $d_{ik} = \sqrt{\sum_{j=1}^{p} (D_{ij} - D_{kj})^2}$, on calcule :

$d_{AB} = \sqrt{(1.30+0.16)^2 + (1.30+0.16)^2 + (1.30+0.18)^2}$ = **2.54**

$d_{AC} = \sqrt{(1.30+1.14)^2 + (1.30+1.14)^2 + (1.30+1.12)^2}$ =**4.21**

- *remarques* :

. les 2 colonnes 1re et 2re, identiques à une multiplication par 2 près, ont les mêmes valeurs standardisées (ce qui est logique),
. on voit de suite que la moyenne de colonnes standardisées =0 (vous pouvez vérifier que leur écart type =1, aux arrondis de calcul près),
. les 2 communes A et B ont même *profil* d'emploi, à une multiplication par 2 près ; leur distance euclidienne est cependant égale à 2.54, ce qui signifie que cette mesure de différence est sensible aux *effets de taille*, même sur variables standardisées,
. En revanche, la commune C a un profil d'emploi légèrement différent et une distance de 4.21 avec A.

La métrique euclidienne de l'ACP, sensible aux effets de taille, n'est donc pas adaptée à la mesure de différences de profils (en ligne ou colonne).

- *Calcul des distances du Khi² d_{AB} et d_{AC}*

Exprimons les effectifs du tableau 3.3 en proportion du total de chaque ligne (tableau 3.5) y compris la ligne « ensemble » (A+B+C).

	1re	**2re**	**3re**	**somme**
A	0.150	0.300	0.550	**1.000**
B	0.150	0.300	0.550	**1.000**
C	0.143	0.286	0.571	**1.000**
A+B+C	**0.149**	**0.299**	**0.552**	**1.000**

Tableau 3.5 : effectifs du tableau 3.3 en proportion du total de ligne

La distance du Khi² entre lignes i et k s'écrit $d_{ik} = \sqrt{\sum_{j=1}^{p} \frac{1}{f.j} (f_{ij/j} - f_{kj/j})^2}$

On peut ainsi calculer :

. $d_{AB} = \sqrt{\frac{1}{0.149}(0.15-0.15)^2 + \frac{1}{0.299}(0.30-0.30)^2 + \frac{1}{0.552}(0.55-0.55)^2}$ = **0**

. $d_{AC} = \sqrt{\frac{1}{0.149}(0.15-0.143)^2 + \frac{1}{0.299}(0.30-0.286)^2 + \frac{1}{0.552}(0.55-0.571)^2}$

$d_{AC} = 6.71 * 0.000049 + 3.34 * 0.000196 + 1.81 * 0.000441 =$ **0.0422**

- *remarques* :

& la différence entre les profils d'emploi des communes A et B est nulle : l'effet de taille des 2 communes a disparu (puisqu'on les a exprimés en fréquences du total de leur ligne). Leur distribution en fréquences de ligne est identique : faire l'AFC du tableau 3.3 en maintenant les lignes A et B distinctes ou en additionnant leurs effectifs en une ligne AB aboutit au même résultat. Cette propriété est appelée « *équivalence distributionnelle* » de l'AFC

& dans le calcul de la différence de profils d'emploi entre communes A et C, la différence de fréquences pour un secteur d'emploi pèse d'autant plus qu'il est plus rare : par exemple ici, le secteur 1re (14.9% de l'emploi des 3 communes) pèse 6.71, le 2re 3.34 (≅4.5 fois moins), le 3re 1.81 (≅8 fois moins). Cette pondération par l'inverse de la fréquence de la somme en colonne en métrique du Khi² « *valorise les différences rares* » car une différence par rapport à une modalité J peu fréquente pèsera beaucoup plus dans la différence globale qu'une modalité J' fréquente.

Toutes les formules données pour les profils en ligne valent pour les profils en colonne à condition de permuter indices de lignes et de colonnes : grâce à la symétrie du rôle des lignes et colonnes, les résultats de l'AFC seront identiques, qu'elle ait été réalisée sur les lignes ou sur les colonnes du tableau N.

2.1.2 *Calcul de la matrice de covariances*

La matrice **C** de covariances est calculée par le produit matriciel (les majuscules désignent des matrices, $\mathbf{F_{ij/i}}$ celle des fréquences en ligne et $\mathbf{F_{ij/j}}$ celle des fréquences en colonne)

$$C = F_{ij/i} * F_{ij/j}$$

2.2 Calcul des Vecteurs Propres et valeurs propres de C

C'est sur cette matrice C que sont calculés Vecteurs Propres et valeurs propres (comme en ACP, cf chapitres 1 et 2), permettant la détermination des aides à l'interprétation.

2.3 Aides à l'interprétation d'une AFC

2.3.1 Généralités

- Comme lignes et colonnes des tableaux de contingence jouent un rôle interchangeable, les aides à leur interprétation sont les *mêmes* (coordonnées des projections, qualités de représentation, contributions relatives) : le graphique des axes et des plans factoriels font cohabiter (sans changement d'échelle) points-ligne et points-colonne.

- En outre, en conséquence de la métrique du Khi², cette cohabitation a un sens : la *proximité* d'un point-ligne i et d'un point-colonne j signifie *sur-représentation* de la modalité j du caractère J dans la modalité i du caractère I et leur *éloignement* signifie *sous-représentation* de l'une par rapport à l'autre.

- L'éloignement d'un point par rapport à l'origine, sur un axe ou un plan factoriel, signifie *écart à l'indépendance*.

On peut d'ailleurs parfaitement interpréter l'AFC comme la décomposition successive (par les vecteurs propres lignes et colonnes) des écarts à l'indépendance d'un calcul de Khi².

2.3.2 Exemple

L'interprétation d'une AFC en est donc facilitée.

Pour le montrer, prenons l'exemple du tableau 3.6.

	50-70m	70-100m	100-150m	150-200m	200-300m	300-500m	500-2M	Paris
Nés Français	714516	571788	727556	626676	1037784	1299808	1822008	3096376
Fr. par acquis.	26872	27288	35152	29272	33888	74064	124756	207072
Italiens	4476	6360	11268	3612	4840	16460	18680	27684
Espagnols	3000	4120	4272	5936	4264	4692	13076	26140
Portugais	8716	5460	8260	3972	12462	9240	17540	100320
Autre CEE	1892	3232	2212	1692	2752	8188	9540	24180
Algériens	5088	8648	9792	6464	8508	18652	39536	83892
Marocains	7140	5092	6628	6136	9364	11260	12536	42884
Tunisiens	1636	1956	1384	792	1764	7504	12484	24736
Turcs	2296	2420	1816	1292	2116	3548	3056	9780
Autres	5372	5840	7060	5516	11968	16680	23556	130784

(source INSEE, RGP 1999)

Tableau 3.6 : Distribution des nationalités par taille de ville (en milliers) en1999

Ce tableau de contingence croise la nationalité des personnes résidant en agglomérations françaises de plus de 50 000 habitants et la taille de celles ci au recensement de 1999.

Le 1er axe factoriel prend en compte 80% de la variance du nuage de points, le 2nd 14% (94% de l'information du tableau 3.6 est donc résumé par 2 axes).

Un rapide examen des premiers résultats nous montre le poids exorbitant de Paris dans l'AFC car, sans doute, la composition ethnique de sa population est fort différente de celle des autres agglomérations de 50 000 habitants ou plus. L'agglomération parisienne pèse à elle seule 1/3 de la population prise en compte, elle contribue (CTR) pour presque les 2/3 à la définition d'un 1er axe représentant 80% de la variance, la coordonnée de sa projection s'y oppose fortement à celle de toutes les autres : ce 1er axe oppose donc nettement la composition ethnique de l'agglomération parisienne à celle de toutes les autres.

Un second essai montre le poids exorbitant de la ligne « Nés Français » : cette origine représente plus de 89% de la population prise en compte : ce poids détermine fortement centre de gravité et variance du nuage de points. Pour y voir plus clair, supprimons la colonne « Paris » et la ligne « Nés Français ». On focalise ainsi l'attention sur la population d'origine étrangère des agglomérations provinciales de plus de 50 000 habitants. Recommençons l'analyse sans cette ligne et cette colonne.

- *% de variance pris en compte*

	Axe 1	Axe 2	Axe 3
% de variance	54%	18%	14%
% cumulé		73%	86%

Tableau 3.7 : % de variance pris en compte par l'AFC du tab. 3.6

Les deux premiers axes résument presque les ¾ et les trois premiers presque les 9/10 de l'information du tableau 3.6 (sans ligne « Paris » et colonne « Nés Français »).

- *Aides à l'interprétation des colonnes*

		Axe 1			*Axe 2*			*Axe 3*			
	poids	**Coord F1**	**CTR F1**	**QR F1**	**Coord F2**	**CTR F2**	**QR F2**	**Coord F3**	**CTR F3**	**QR F3**	**Somme des QR**
50-70m	0.08	**0.27**	0.22	*0.83*	-0.02	0.00	0.01	-0.03	0.01	0.01	**0.85**
70-100m	0.09	0.04	0.00	0.06	-0.04	0.01	0.06	0.08	0.07	0.25	*0.36*
100-150m	0.11	0.08	0.03	0.14	***-0.11***	0.14	0.23	**0.13**	0.29	0.37	0.73
150-200m	0.08	0.10	0.03	0.11	**0.20**	0.36	0.49	**0.15**	0.27	0.29	**0.88**
200-300m	0.11	**0.28**	0.33	*0.78*	0.03	0.01	0.01	***-0.13***	0.28	0.17	**0.96**
300-500m	**0.21**	-0.10	0.08	0.33	***-0.12***	0.32	0.47	-0.02	0.01	0.01	**0.81**
500m-2M	**0.33**	***-0.15***	0.31	*0.78*	0.06	0.16	0.13	-0.03	0.06	0.04	**0.96**

Tableau 3.8 : Aides à l'interprétation des colonnes du tab. 3.6 par l'AFC

- *Aides à l'interprétation des lignes*

		Axe 1			*Axe 2*			*Axe 3*			
	poids	**Coord F1**	**CTR F1**	**QR F1**	**Coord F2**	**CTR F2**	**QR F2**	**Coord F3**	**CTR F3**	**QR F3**	**Somme des QR**
Fr/acquis.	0.43	-0.06	0.06	0.67	0.03	0.03	0.14	0.01	0.00	0.01	**0.81**
Italiens	0.08	-0.04	0.01	0.03	***-0.23***	0.47	0.64	**0.15**	0.27	0.28	**0.94**
Espagnols	0.05	0.09	0.01	0.07	**0.25**	0.35	0.59	**0.18**	0.25	0.31	**0.98**
Portugais	0.08	**0.32**	0.33	0.81	-0.01	0.00	0.00	***-0.11***	0.13	0.08	**0.89**
Autre CEE	0.04	-0.12	0.02	0.25	***-0.12***	0.06	0.26	-0.05	0.01	0.04	*0.55*
Algériens	0.12	***-0.16***	0.12	0.73	0.05	0.03	0.07	-0.01	0.00	0.00	**0.80**
Marocains	0.07	**0.30**	0.24	0.92	-0.01	0.00	0.00	0.02	0.01	0.01	**0.92**
Tunisiens	0.03	***-0.33***	0.14	0.73	-0.05	0.01	0.01	***-0.18***	0.17	0.21	**0.95**
Turcs	0.02	**0.28**	0.06	0.50	***-0.15***	0.05	0.14	0.07	0.02	0.04	*0.68*
Autres	0.09	0.07	0.02	0.17	-0.01	0.00	0.00	***-0.10***	0.15	0.42	*0.59*

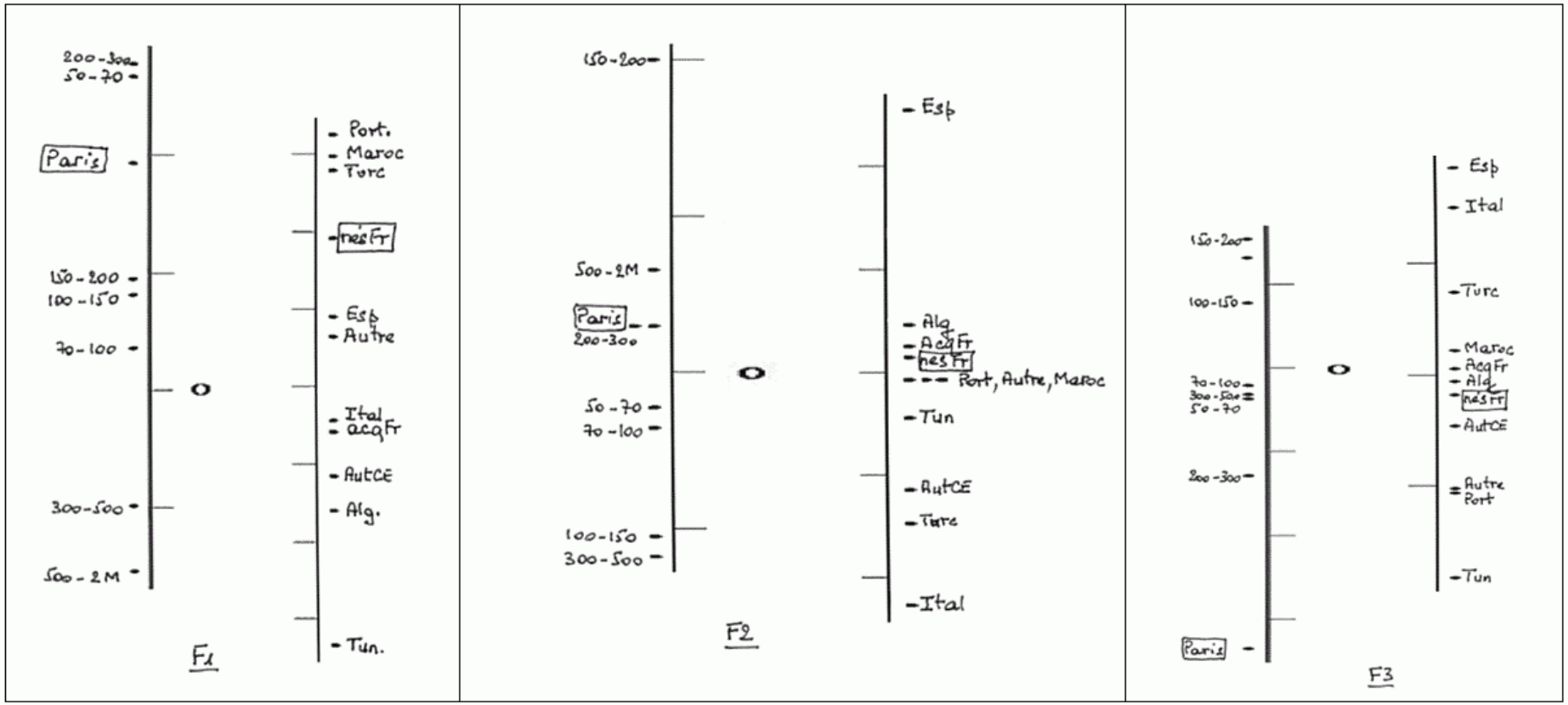
Tableau 3.9 : Aides à l'interprétation des lignes du tab. 3.6 par l'AFC

La Figure 3.1 représente les coordonnées des lignes et des colonnes sur les 3 premiers axes factoriels. La colonne « Paris » et la ligne « Nés Français » y sont présentes : elles ne sont pas intervenues dans les calculs mais ont été projetées a posteriori comme colonne et ligne supplémentaires sur chacun des axes (afin de tester leur comportement par rapport aux autres modalités sans que ne joue leur prépondérance dans l'analyse).

- *Scalogrammes des axes factoriels* (figure 3.1)

Sont graphiqués sur le scalogramme de chaque axe coordonnées des lignes et des colonnes, en vis à vis, puisque la proximité de points ligne et colonne signifie sur-représentation et l'éloignement sous-représentation : c'est là l'information essentielle de l'AFC.

Figure I.3.1 : coordonnées des lignes et colonnes sur les 3 premiers axes factoriels



2.3.3 *Interprétation des axes factoriels*

Elle doit tenir compte des ***scalogrammes*** (coordonnées sur chaque axe des modalités ligne et colonne) mais aussi des ***contributions*** (relatives car elles somment à 1 pour chaque axe) et surtout des ***qualités de représentation*** (distance d'un point à sa projection sur l'axe).

L'information essentielle est fournie par les coordonnées (« coord ») et, principalement par les plus fortes en valeur absolue. Le signe (+ ou -) affecté aléatoirement aux coordonnées ne signifie rien ; ce qui est signifiant, c'est ***l'opposition de modalités*** à fortes coordonnées positives et négatives car ce sont elles qui ont crée la direction de l'axe.

La lecture des contributions (« CTR ») indique si une modalité n'a pas pesé de façon exagérée sur la définition d'un axe, auquel cas il convient de la retirer (ce que nous avons fait avec « Paris » et « nés français ») quitte à la projeter en ligne ou colonne supplémentaire. Elles indiquent, a contrario, les modalités ayant très faiblement participé à la variance de l'axe.

Les qualités de représentation (« QR ») qui, logiquement, diminuent d'axe en axe, sont surtout utiles pour les coordonnées proches de l'origine : si ces modalités sont bien représentées, elles ont réellement une valeur moyenne, sinon un axe ultérieur les représentera mieux. La somme des QR renseigne surtout sur les modalités mal prises en compte par l'ensemble des axes retenus.

& L'axe 1(résumant plus de la moitié de la variance) oppose :

- les plus grandes agglomérations, à sur-représentation des tunisiens et des algériens (ressortissants de la Communauté Européenne mal représentés),
- aux agglomérations moyennes et petites où portugais, marocains et turcs (moins bien pris en compte) sont sur-représentés,
- les français par acquisition (de toute nationalité antérieure) sont bien pris en compte et ont une coordonnée moyenne (plutôt proche de celle des tunisiens et algériens),

La logique de cette opposition tient sans doute à :

- l'***ancienneté*** plus ou moins grande de leurs immigrations (vague migratoire algérienne et tunisienne plus ancienne que celle des portugais et marocains),
- la ***répartition*** des agglomérations dans l'espace français : plus à l'est pour les plus grandes (l'est est plus proche de la Tunisie et de l'Algérie, l'ouest est plus proche du Portugal et du Maroc). Les français par acquisition sont assez bien répartis mais avec une tendance voisine des migrations plus anciennes.

& L'axe 2 (18% de la variance, soit 4/10 environ de la variance résiduelle) oppose :

- les agglomérations de 150 à 200 mille habitants où les espagnols sont plus nombreux qu'ils seraient s'il y avait indépendance entre cette nationalité et cette taille de villes,
- aux autres agglomérations de taille moyenne (300-500 mille habitants) où les italiens sont sur-représentés.

Cet axe met en évidence une différence de distribution entre deux nationalités d'ancienne immigration (du moins la distribution de ceux restés de nationalité étrangère). L'explication tient sans doute à la différence de répartition sur le territoire français des agglomérations (moyennes-petites et moyennes) en fonction de la proximité au pays d'origine (espagnols plus nombreux dans le sud ouest, italiens dans le sud est) : hypothèse à vérifier !

& L'axe 3 (encore 14% de la variance totale) oppose :

- agglomérations de 100 à 200 mille habitants où espagnols et italiens sont sur-représentés,
- agglomérations de 200 à 300 mille habitants où ce sont les nationalités « autres ».

La différenciation, mineure par rapport à la précédente, est néanmoins intéressante : elle porte sur des agglomérations de taille moyenne et oppose des nationalités d'ancienne immigration (espagnols, italiens restés étrangers) diffusées jusque dans les plus petites villes moyennes aux nationalités d'immigration très récente (Afrique, Asie) mieux représentés dans de plus grandes villes (et surtout à Paris, projeté ici en modalité supplémentaire).

& modalités mal représentées

Il s'agit surtout des agglomérations de 70 à 100 mille habitants , des turcs et des autres nationaux de la Communauté Européenne que les latins. Cela signifie probablement qu'ils sont assez également répartis dans les villes de différentes tailles.

2.3.4 *Typologie sur le plan des axes 1 et 2* (Figure 3.2)

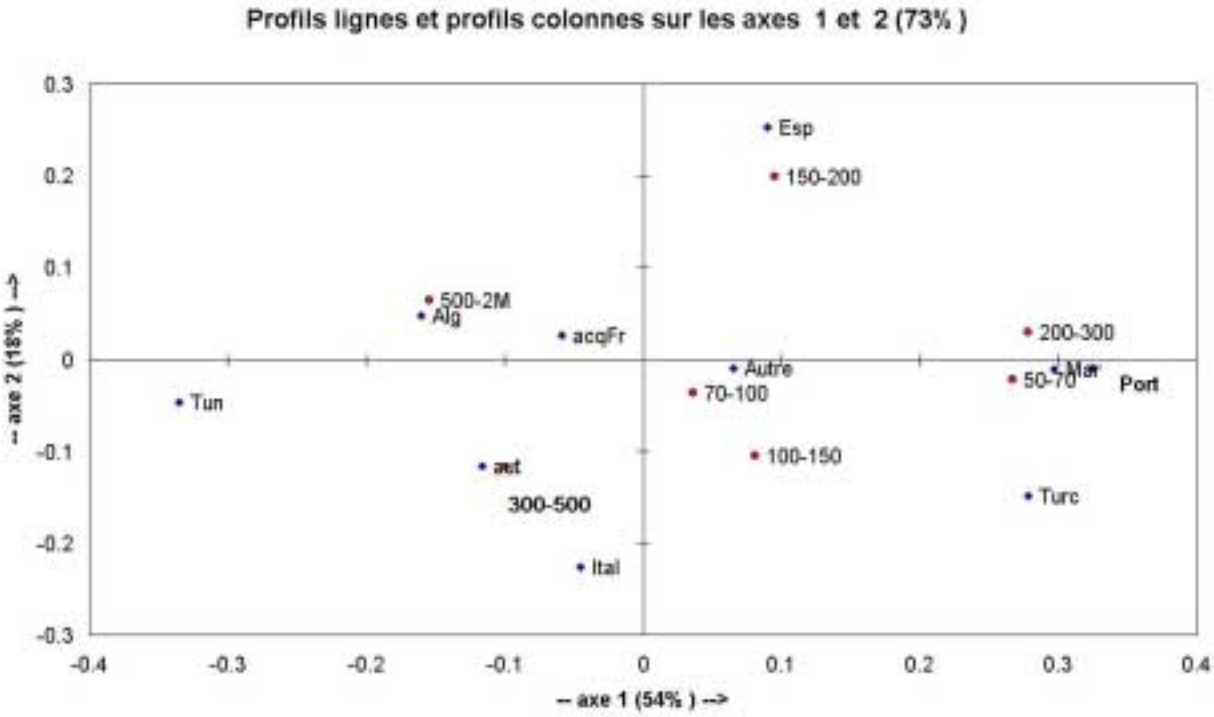


Figure 3.2 : coordonnées sur le plan des axes 1 et 2

- Marocains, Portugais et, dans une moindre mesure, Turcs sont sur-représentés dans les villes de 50 à 70 mille et 200 à 300 mille habitants,
- Les espagnols dans celles de 150 à 200 mille,
- Les nationalités « autres » (essentiellement Afrique noire et Asie) dans celles de 70 à 150 mille habitants,
- Les italiens et les autres nationalités de la C.E. dans celles de 300 à 500 mille,
- Les Algériens et français par acquisition dans les métropoles,
- Les Tunisiens semblent surtout sous-représentés dans les villes petites ou moyennes.

3. AFC sur tableaux de contingence à plus de 2 caractères

Dans la pratique, il existe des tableaux d'effectifs pour le croisement de plus de 2 caractères, leurs lignes et/ou leurs colonnes repérant alors la combinaison de modalités de caractères différents.

- Un 1er exemple est constitué d'un tableau ayant en lignes des unités spatiales (départements, régions, agglomérations, zones d'emploi, …) et en colonnes des classes d'âge par sexe : faire l'AFC d'un tel tableau aboutit à créer une typologie d'unités spatiales en fonction de leurs pyramides d'âge.
- D'autres exemples peuvent être fournis par des tableaux ayant en lignes des unités spatiales et en colonnes les modalités d'un caractère pour des dates ou des périodes différentes.

3.1 Exemple

Le tableau 3.10, fabriqué à partir d'un S.I.G., croise pour le département de la Savoie :
- en ligne, la classe d'altitude et le type d'utilisation du sol,
- en colonne, la classe de pente en degrés,
- chaque case contient un nombre de pixels (carrés de 50*50 mètres au sol).

	code	Pente 0-3°	Pente 3-10°	Pente10-20°	Pente 20-35°	Pente >=35°
<1000-artificiel	**A**	29825	24048	9322	2008	223
<1500-artificiel	**B**	415	1942	2328	452	57
<2000-artficiel	**C**	193	1075	1647	384	43
>=2000-artificiel	**D**	27	261	335	358	229
<1000-agricole	**E**	96869	116123	88654	26308	1628
<1500-agricole	**F**	997	2229017	28486	26347	1493
<2000-agricole	**G**	5994	3588	7390	5604	575
>=2000-agricole	**H**	898	861	1613	2134	592
<1000-nature	**I**	29800	32330	82459	133457	52190
<1500-nature	**J**	3225	18960	71396	180514	72645
<2000-nature	**K**	2387	19544	101355	223841	84178
>=2000-nature	**L**	5940	61243	221364	418133	201882

(sources : IGN, IFEN)

Tableau 3.10 : Tableau croisant pente, altitude, utilisation du sol du département 73

Les classes sont les suivantes :
- *pente* : 0-3°, 3-10°, 10-20°, 20-35°, 35° et plus
- *altitude* : <1000m, 1000-1500m, 1500-2000m, 2000m et plus
- *utilisation du sol* : artificielle (zones urbaines, industrielles ou commerciales, réseaux, chantiers, espaces verts non agricoles), agricole, naturelle (forêts, zones humides, eau).

La Savoie, département le plus élevé de France en moyenne, connaît un fort étagement de ses activités et donc de ses utilisations du sol. De fait, le Khi² calculé est extrêmement fort et confirme une forte relation entre lignes et colonnes du tableau 3.10.

Que nous dit de plus l'AFC du tableau 3.10 ?

& Les 2 premiers axes représentent 100% de l'information (75% pour le 1er, 25% pour le 2nd).

& Les tableaux 3.11 et 3.12 fournissent les autres aides à l'interprétation (coordonnées des projections, Contributions relatives, Qualités de représentation).

- *Aides à l'interprétation des colonnes*

	Poids	Coord F1	CTR F1	QR F1	Coord F2	CTR F2	QR F2
0-3°	0.04	0.34	0.01	0.02	**2.44**	**0.86**	**0.98**
3-10°	**0.53**	***-0.82***	**0.46**	**0.99**	-0.07	0.01	0.01
10-20°	0.13	**0.83**	0.12	**0.88**	0.25	0.03	0.08
20-35°	**0.22**	**1.02**	**0.29**	**0.94**	-0.25	0.05	0.06
>=35°	0.09	**1.09**	0.13	**0.89**	-0.35	0.04	0.09

Tableau 3.11 : Aides à l'interprétation des colonnes du tab. 3.10

- *Aides à l'interprétation des lignes*

	code	Poids	Coord F1	CTR F1	QR F1	Coord F2	CTR F2	QR F2
<1000-artificiel	***A***	0.01	0.01	0.00	0.00	**2.19**	**0.26**	**0.96**
<1500-artificiel	***B***	0.00	0.22	0.00	0.05	0.50	0.00	0.25
<2000-artficiel	***C***	0.00	0.34	0.00	0.09	0.41	0.00	0.14
>=2000-artificiel	***D***	0.00	0.64	0.00	**0.82**	-0.07	0.00	0.01
<1000-agricole	***E***	0.07	0.14	0.00	0.01	**1.45**	**0.57**	**0.99**
<1500-agricole	***F***	**0.48**	***-0.88***	**0.48**	**0.98**	-0.13	0.03	0.02
<2000-agricole	***G***	0.00	0.57	0.00	0.17	**1.24**	0.03	**0.81**
>=2000-agricole	***H***	0.00	**0.70**	0.00	0.58	0.58	0.00	0.40
<1000-nature	***I***	0.07	**0.84**	0.06	**0.93**	0.23	0.01	0.07
<1500-nature	***J***	0.07	**1.00**	0.09	**0.92**	-0.27	0.02	0.07
<2000-nature	***K***	0.09	**1.02**	**0.12**	**0.93**	-0.26	0.02	0.06
>=2000-nature	***L***	**0.19**	**0.97**	**0.23**	**0.94**	-0.24	0.04	0.06

Tableau 3.12 : Aides à l'interprétation des lignes du tab. 3.10

- *Pondérations*

Plus de la moitié de la surface (53%) a des pentes comprises entre 3 et 10° (avant pays, vallées glaciaires) mais 31% a plus de 20°. L'espace agricole compris entre 1000 et 1500m représente presque la moitié (48%) du territoire, l'espace naturel au dessus de 1500m un gros quart (28%).

3.2 Interprétation de l'axe 1(75% de variance)

Il oppose :

- surfaces agricoles situées entre 1000-1500m sur des pentes faibles (3-10°), résidu d'agriculture montagnarde (élevage pour l'essentiel) sur épaulements ou terrasses,
- et espaces naturels à toute altitude mais sur pentes moyennes ou fortes (>10°), espaces restés naturels ou boisés à cause de la pente (et de la déprise agricole),
- on remarque aussi, du côté des pentes fortes ou moyennes de l'axe 1, les surfaces à 2000m ou plus dédiées à l'agriculture (alpage) et, à même altitude, les surfaces artificialisées (stations de sports d'hiver).

3.3 Interprétation de l'axe 2 (25% de variance)

Il met surtout en évidence les espaces artificialisés, inférieurs à 1000m et de pente faible : ce sont les espaces urbanisés des basses vallées (Maurienne, Tarentaise, Sillon Alpin, cluse de Savoie). Apparaissent aussi, sur pentes faibles, les surfaces agricoles situées en dessous de 1000m (toujours dans les basses vallées) et entre 1500 et 2000m (alpages).

3.4 Plan des axes 1 et 2 (100% de variance)

La figure 3.3 révèle quatre grands types, plus ou moins affirmés, de correspondances :

- les espaces agricoles situés entre 1000 et 1500m (F) sur faibles pentes,

- les espaces naturels sur fortes pentes à toute altitude jusqu'à 2000m (I,J,K)
- les espaces urbanisés (A) et les espaces agricoles à basse altitude (E) ou, au contraire, entre 1500 et 2000m mais tous sur faibles pentes,
- les espaces artificialisés au dessus de 1000m (stations de sports d'hiver, B,C,D) ou les alpages d'altitude (H).

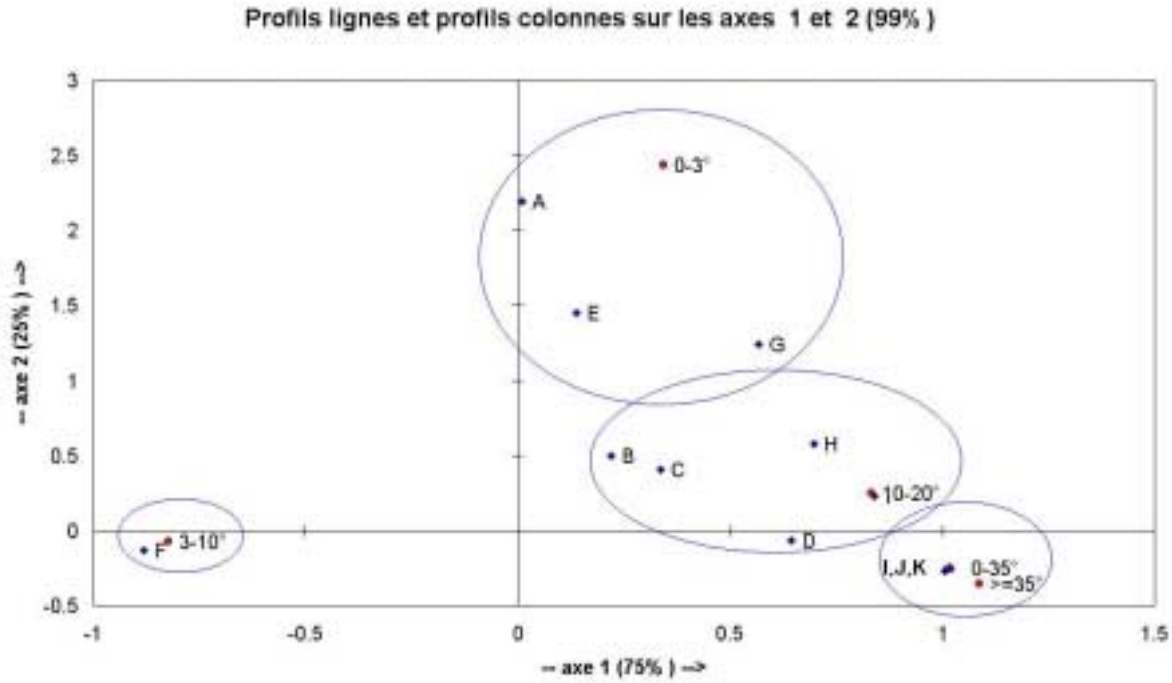


Figure 3.3 : plan des axes 1 et 2 de l'AFC du tableau 3.10

L'Analyse Factorielle des Correspondances :
- est adaptée au résumé de tableaux de contingence (effectifs pour le croisement de caractères *qualitatifs* connus par modalités et/ou *quantitatifs* connus par classes), - fournit les mêmes aides à l'interprétation en lignes et colonnes (mêmes résultats si on transpose le tableau de données), - sur le scalogramme d'un axe ou un plan factoriel, la proximité d'un point-ligne i et d'un point-colonne j signifie sur-représentation de la classe i dans la classe j, leur éloignement signifiant sous-représentation de la classe i dans la classe j.

B) EXERCICES CORRIGES

Exercice 1

On connaît, grâce au recensement de 1999, la structure d'âges des logements en France (source : INSEE, RGP 1999). Le tableau 3.13 en fournit une représentation simplifiée pour chaque région de France métropolitaine. Les nombres de logements construits à chaque époque sont exprimés en % du total de ligne (la somme pour chaque région est égale à 100).

	av49	49-74	75-89	90-99
alsace	32.70	33.10	21.70	12.50
aquitaine	33.20	28.40	26.40	12.10
auvergne	39.30	28.40	23.60	8.70
b_normandie	35.60	30.60	24.30	9.50
bretagne	25.30	33.10	28.50	13.10
bourgogne	41.40	28.50	22.10	8.10
centre	34.30	30.70	25.40	9.60
champagne	35.80	35.50	21.70	7.00
corse	25.50	31.50	31.80	11.10
f_comté	35.20	32.10	22.70	10.00
h_normandie	32.70	34.50	24.70	8.10
ile-de-France	33.70	37.80	19.40	9.10
languedoc	29.00	27.30	29.40	14.30
limousin	38.30	30.10	23.70	7.90
lorraine	35.20	36.00	21.10	7.70
midi_pyr	31.30	30.00	25.20	13.50
nord	39.90	31.50	21.70	6.80
picardie	40.90	29.20	21.90	8.10
p_loire	29.80	29.70	27.40	13.10
poitou	37.50	25.80	24.90	11.80
provence	26.10	38.10	25.30	10.60
rh-alpes	28.50	34.10	24.90	12.50

Tableau 3.13: structure d'âge des logements par région en 1999 (*source : INSEE)*

Une AFC a été pratiquée pour résumer ce tableau. En voici les aides à l'interprétation :

- *% de variance des axes factoriels*

axe 1 : 65% axe 2 : 27% axes 1+2 : 92%

- *résultats pour les colonnes*

	Poids	F1	ctr	qr	F2	ctr	qr
av49	0.34	***-0.13***	0.48	0.88	-0.05	0.16	0.12
49-74	0.32	0.00	0.00	0.00	**0.10**	0.68	1.00
75-89	0.24	0.10	0.20	0.69	-0.04	0.07	0.10
90-99	0.10	**0.19**	0.32	0.77	***-0.07***	0.09	0.09

Tableau 3.14 : Aides à l'interprétation des colonnes du tab. 3.13

- *résultats pour les lignes (régions)*

	Poids	F1	ctr	qr	F2	ctr	qr
alsace	0.05	0.03	0.00	0.08	0.02	0.00	0.05
aquitaine	0.05	0.06	0.01	0.37	***-0.07***	0.05	0.62
auvergne	0.05	***-0.10***	0.04	*0.69*	***-0.07***	0.04	0.30
b_normandie	0.05	-0.04	0.01	*0.71*	-0.02	0.00	0.22
bretagne	0.05	**0.19**	0.13	*0.97*	0.03	0.01	0.03
bourgogne	0.05	***-0.15***	0.09	*0.84*	***-0.07***	0.04	0.16
centre	0.05	-0.01	0.00	0.09	-0.02	0.00	0.25
champagne	0.05	***-0.11***	0.04	*0.59*	**0.09**	0.07	0.39
corse	0.05	**0.18**	0.12	*0.75*	0.01	0.00	0.00
f_comté	0.05	-0.04	0.01	*0.70*	0.01	0.00	0.03
h_normandie	0.05	-0.02	0.00	0.07	**0.07**	0.04	0.62
ile-de-France	0.05	-0.06	0.02	0.18	**0.13**	0.15	0.69
languedoc	0.05	**0.17**	0.11	*0.76*	***-0.10***	0.08	0.24
limousin	0.05	***-0.10***	0.04	*0.85*	-0.03	0.01	0.06
lorraine	0.05	-0.09	0.03	0.49	**0.10**	0.08	0.51
midi_pyr	0.05	0.09	0.03	*0.65*	-0.04	0.02	0.13
nord	0.05	***-0.16***	0.10	*0.98*	0.00	0.00	0.00
picardie	0.05	***-0.15***	0.08	*0.89*	-0.05	0.02	0.11
p_loire	0.05	**0.12**	0.06	*0.88*	-0.04	0.02	0.11
poitou	0.05	-0.01	0.00	0.01	***-0.13***	0.15	0.95
provence	0.05	0.11	0.04	0.37	**0.14**	0.18	0.63
rh-alpes	0.05	**0.11**	0.04	*0.76*	0.05	0.02	0.16

Tableau 3.15 : Aides à l'interprétation des lignes du tab. 3.13

Questions

1) **Quelle hypothèse a fait choisir un tel tableau ?**
2) **Pourquoi le résumer par une AFC plutôt que par une ACP ?**
3) **Interprétez chacun des 2 axes factoriels**
4) **Construisez une typologie su le plan des axes F1-F2 et cartographiez la**

Question 1

L'hypothèse, fort simple, est la suivante : un fort % de construction de logements à une période donnée date le dynamisme démographique et économique des régions.

Question 2

Une ACP aurait été possible puisqu'on dispose pour 21 agrégats régionaux de 4 variables de comptage (donc quantitatives).

Une AFC a été préférée car :

- le tableau 3.13 est un cas particulier de *tableau de contingence* où les sommes en colonne indiquent la fréquence des différentes dates de construction des logements et où il est licite de calculer les sommes en lignes (égales pour toute région à 100%).
- Ce qu'on désire, c'est comparer les structures d'âge des logements en ayant, pour chaque région, connaissance de la sur- ou sous-représentation de chaque période de construction.

Question 3

- *Interprétation de l'axe 1*

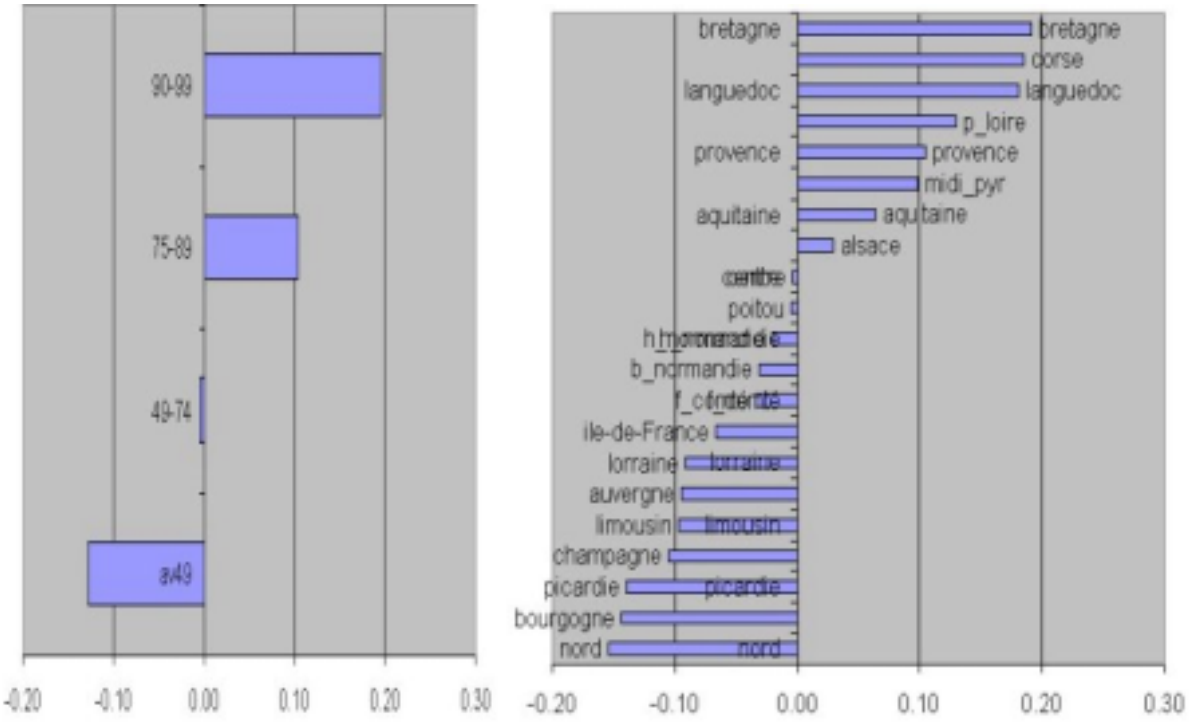


Figure 3.4 : Coordonnées des périodes et des régions sur l'axe F1

Compte tenu des coordonnées et qualités de représentation, l'axe 1 (représentant 65% de l'information du tableau 3.13) oppose :

- des régions (en difficulté industrielle ou peu urbanisées) à surplus de logements d'avant 1949 (Nord, Picardie, Bourgogne, Champagne, Auvergne, Limousin),

- à celles (le plus souvent touristiques et/ou rurbanisées) avec fort surplus de logements récents (Bretagne, Corse, Languedoc, Pays de Loire, Rhône-Alpes).

Ce 1er axe ordonne chronologiquement les périodes de construction immobilière et les régions selon l'ancienneté relative de leurs parcs de logements.

- *Interprétation de l'axe 2*

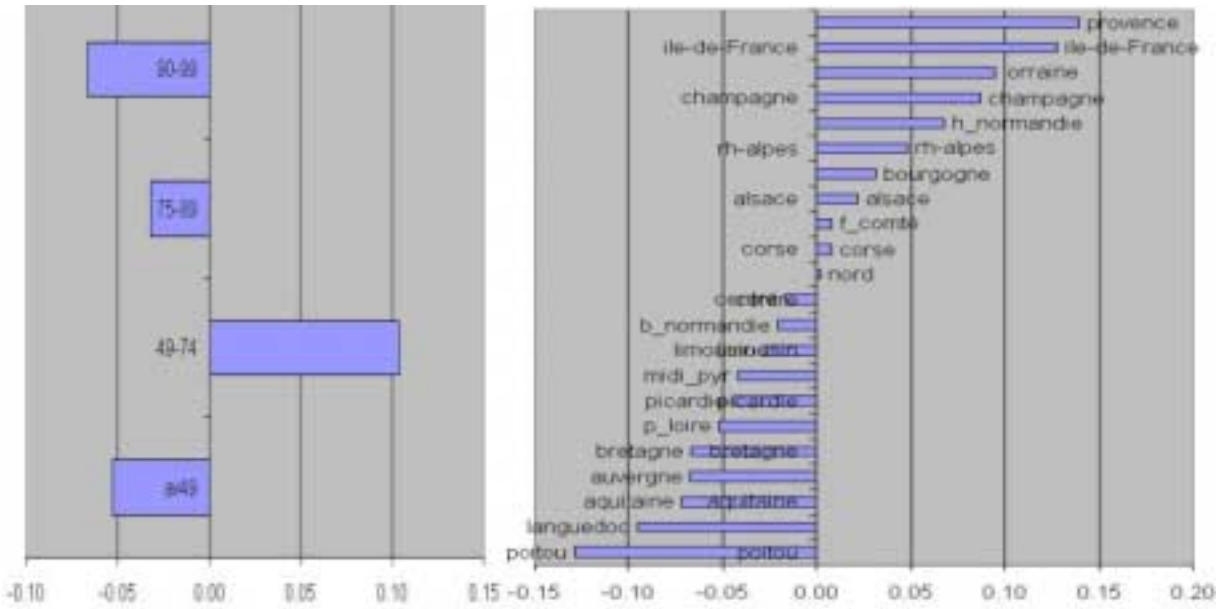


Figure 3.5 : Coordonnées des périodes et des régions sur l'axe F2

L'axe 2 (27% de l'information du tableau 3.13) met en valeur la période 1950-1975 qui était peu représentée par l'axe précédent. Il oppose :

- des régions où cette époque des « 30 glorieuses » est sur-représentée dans le parc actuel de logements (Provence, Ile de France, Lorraine, Champagne, Haute Normandie),

- à des régions où elle est sous-représentée (Poitou, Languedoc, Bourgogne, Aquitaine, Auvergne).

Cet axe met donc en valeur l'opposition entre régions ayant connu un « boom » immobilier lors des 30 glorieuses et régions qui, alors peu dynamiques démographiquement et économiquement, ont alors construit peu de logements neufs.

Deux régions ont été mal prises en compte par les 2 premiers axes de l'AFC : l'Alsace (qualité de représentation sur F1 et F2=0.13) et la région Centre (somme des 2 QR=0.34).

Question 4

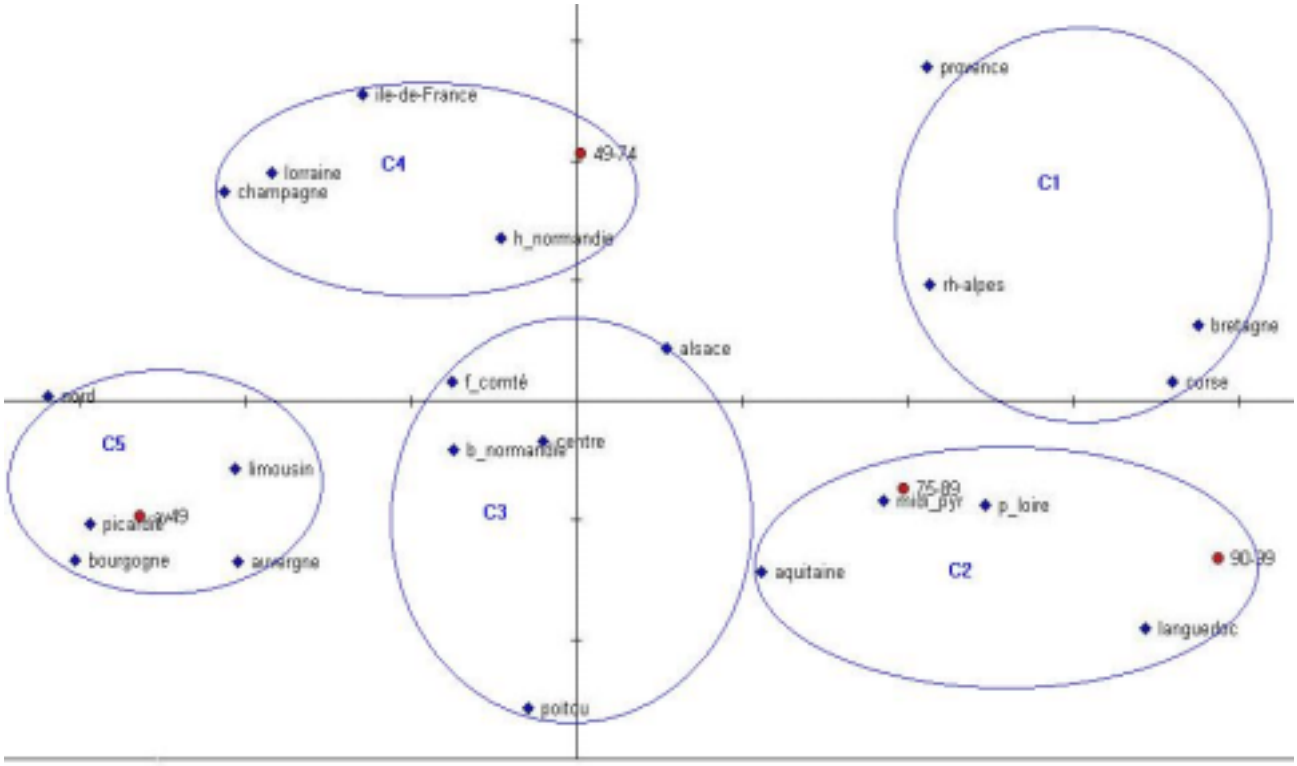


Figure 3.6 : périodes de construction des logements et régions sur le plan des axes F1-F2

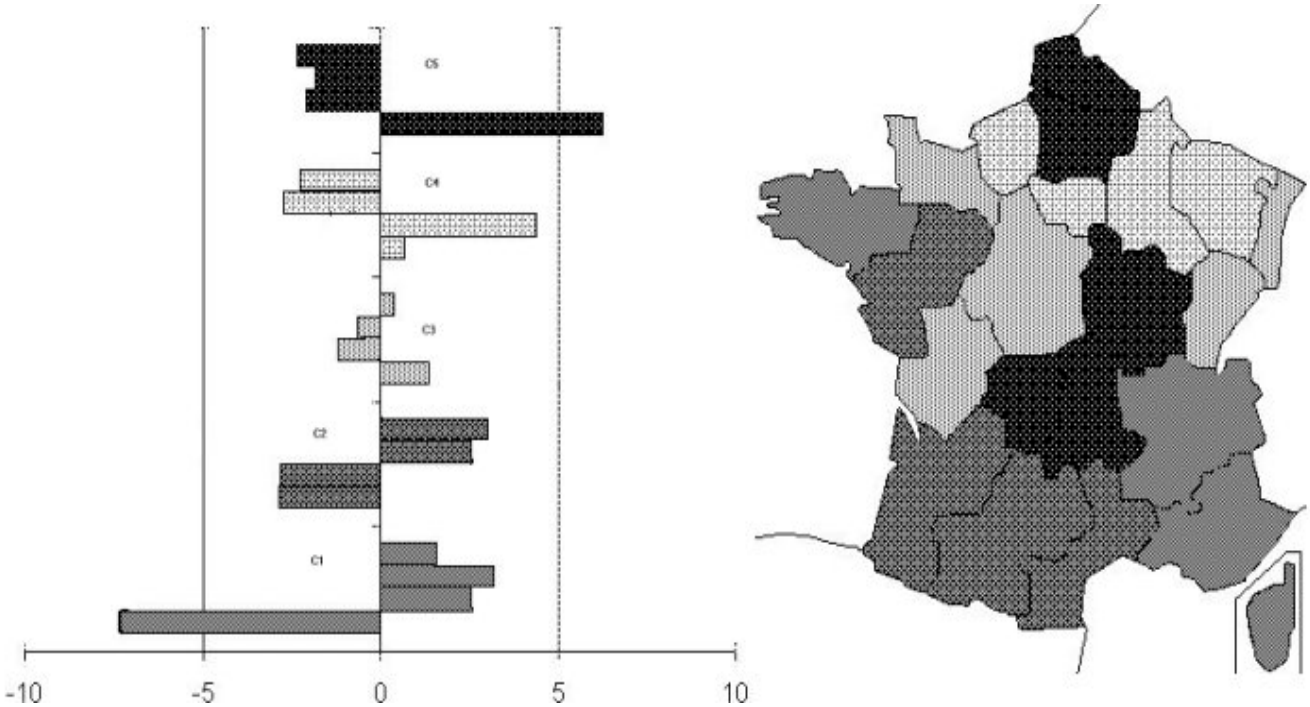


Figure 3.7 : Cartographie des classes de la fig. 3.6 (légende : écarts des 4 périodes à la moyenne nationale)

La figure 3.6 présente, sur le premier plan factoriel, la projection des colonnes (périodes de construction des logements actuels) et des lignes (régions) des tableaux 3.14 et 3.15 : un regroupement en 5 classes, plus ou moins homogènes, y été effectué. La figure 3.7 est la carte thématique correspondante: sa légende (partie gauche) explicite pour chaque période et chacune des 5 classes l'écart à la moyenne des régions françaises métropolitaines.

Ces deux figures autorisent le commentaire suivant :

- *classe C1* (Bretagne, Corse, Rhône-Alpes, Provence) caractérisée par un parc immobilier assez jeune (très faible % de logements d'avant 1949, forts % de plus récents avec un pic datant de la période 1975-89) ; ce profil chronologique s'explique en partie par le caractère touristique de ces régions.
- *Classe C2* (Languedoc, Pays de Loire, Midi Pyrénées, Aquitaine) caractérisée également par un parc immobilier assez récent (faibles % avant 1975, assez forts près) ; par rapport à la classe C1, le pic de construction est plus récent (1990-99).
- *Classe C3* (Poitou, Basse Normandie, Centre, Franche Comté, Alsace) avec un profil moyen très voisin de celui de l'ensemble de la France métropolitaine ; à noter que Alsace et Centre sont mal représentés sur le plan F1-F2 tandis que les 3 autres régions, bien représentées, ont réellement un profil moyen.
- *Classe C4* (Champagne, Lorraine, Ile de France, Haute Normandie) avec une forte sur-représentation de logements datant de la période 1949-74 et un déficit ensuite ; ces régions ont connu leur « âge d'or » immobilier pendant les « 30 glorieuses ».
- *Classe C5* (Nord, Picardie, Bourgogne, Limousin, Auvergne) caractérisée par l'importance des logements les plus anciens (avant 1949) et la sous-représentation des autres périodes ; la raison est double : difficultés économiques ou caractère « rural profond » de ces régions.

Compte tenu de ces profils, les classes 1 et 2 (sur-représentation des logements récents) se différencient assez nettement des classes 4 et 5 (à sous-représentation de logements d'après 1974) tandis que la classe 3 a un profil moyen. Cette chronologie immobilière, avec ses pics et ses creux, donne une des images matérielles possibles des dynamiques territoriales de la France métropolitaine (à granularité spatiale, il est vrai, assez grossière).

Exercice 2

code	Nom de département	pota_sup	pota_sout	indus_sup	indus_sout	irrig_sup	irrig_sout
59	Nord	0	136892	89091	39423	203	1691
62	Pas-de-Calais	16753	135686	76963	31449	853	3648
76	Seine-Maritime	25	105212	199437	54841	116	297
80	Somme	0	47552	13846	34438	260	18327
22	Côtes-d'Armor	42071	10629	1634	1889	147	596
29	Finistère	49122	17289	10490	4855	1552	1498
35	Île-et-Vilaine	37687	16679	2499	2241	1581	1139
44	Loire-Atlantique	66554	22991	18874	1494	11884	3326
56	Morbihan	46933	12474	6136	3282	3978	2433
17	Charente-Maritime	23325	26718	2363	3599	17087	88683
40	Landes	4682	36403	33218	11280	86215	132247
85	Vendée	29525	15884	342	1822	19929	27479
14	Calvados	10801	48054	2205	3444	397	890
33	Gironde	21465	100833	33770	21332	18170	60189
50	Manche	17454	24343	3710	2585	265	2590
64	Pyrénées-Atlantiques	56297	21038	91605	6559	26993	1624

(sources : Agences de l'eau, RNDE, IFEN)

Tableau 3.16 : Principaux usages de l'eau en 2003 dans les départements atlantiques

Le tableau 3.16 récapitule, en milliers de m^3 pour l'année 2003, les principaux usages des eaux (superficielles ou souterraines) pour 16 départements de la façade (Manche et Atlantique) de la France métropolitaine. N'y est pas comptabilisé l'usage des eaux pour la production énergétique (concentrée pour l'essentiel dans d'autres départements).

- *% de variance des deux 1ers axes*

axe F1 : 46% axe F2 : 36% plan F1-F2 : 82%

- *Aides à l'interprétation des colonnes*

	Poids	F1	ctr	qr	F2	ctr	qr
pota_sup	0.17	0.26	0.03	0.04	**1.26**	**0.78**	**0.96**
pota_sout	0.31	-0.34	0.08	0.40	-0.12	0.01	0.05
indus_sup	0.23	***-0.59***	**0.19**	**0.56**	-0.20	0.03	0.07
indus_sout	0.09	-0.51	0.05	0.43	-0.35	0.03	0.20
irrig_sup	0.07	**1.06**	**0.20**	**0.70**	-0.24	0.01	0.04
irrig_sout	0.14	**1.19**	**0.45**	**0.79**	-0.56	0.13	0.18

Tableau 3.17 : Aides à l'interprétation des colonnes de l'AFC du tableau 3.16

- *Aides à l'interprétation des lignes*

	Poids	F1	ctr	qr	F2	ctr	qr
Calvados	0.03	-0.35	0.01	0.13	0.15	0.00	0.02
Charente-Maritime	0.06	**1.11**	0.19	0.78	-0.32	0.02	0.06
Côtes-d'Armor	0.02	0.17	0.00	0.01	**1.53**	0.16	0.96
Finistère	0.03	0.03	0.00	0.00	**1.12**	0.12	0.98
Gironde	0.10	0.19	0.01	0.20	-0.26	0.02	0.36
Île-et-Vilaine	0.02	0.12	0.00	0.01	**1.21**	0.11	0.96
Landes	0.12	**1.07**	0.32	0.72	-0.59	0.13	0.22
Loire-Atlantique	0.05	0.18	0.00	0.03	**0.99**	0.15	0.94
Manche	0.02	-0.12	0.00	0.03	0.54	0.02	0.55
Morbihan	0.03	0.20	0.00	0.03	**1.21**	0.13	0.96
Nord	0.10	***-0.67***	0.11	0.76	-0.32	0.03	0.17
Pas-de-Calais	0.10	***-0.57***	0.08	0.79	-0.16	0.01	0.06
Pyrénées-Atl.	0.08	-0.14	0.00	0.03	0.34	0.03	0.19
Seine-Maritime	0.14	***-0.77***	0.20	0.67	-0.35	0.05	0.14
Somme	0.04	-0.27	0.01	0.08	-0.47	0.03	0.26
Vendée	0.04	0.89	0.07	0.89	0.26	0.01	0.08

Tableau 3.18 : Aides à l'interprétation des lignes de l'AFC du tableau 3.16

Questions

1) **Une ACP était possible pour résumer le tab. 3.16 : quelles auraient été les différences par rapport à l'AFC ici pratiquée ?**
2) **Signification des axes F1 et F2**
3) **Interprétez en géographe le plan F1-F2**

Corrigé

Question 1

Une ACP sur les 6 variables standardisées aurait principalement fait ressortir les départements les plus gros consommateurs d'eau (les plus peuplés, industriels ou irrigateurs) car

standardiser les variables ne fait qu'effacer les différences de moyennes et de variances des descripteurs mais une fois standardisée, une forte valeur reste une forte valeur relative.

Ce qui nous importe ici, c'est de comparer des profils départementaux d'usage des consommations d'eau pour mettre en relief des originalités. L'AFC est donc indiquée.

Question 2

- *Interprétation de l'axe F1*

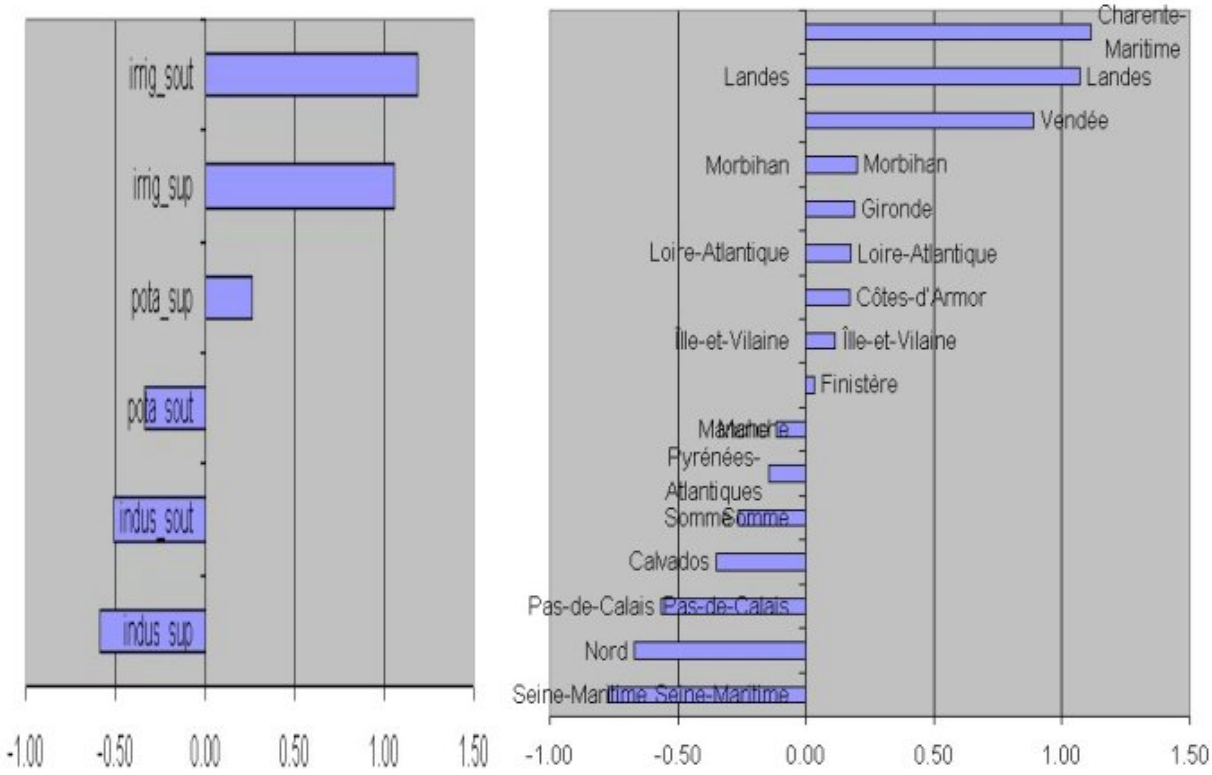


Figure 3.8 : Coordonnées des colonnes et des lignes de l'AFC du tableau 3.16 sur l'axe F1

Compte tenu des coordonnées, contributions et qualités de représentation, l'axe F1 (presque la moitié de l'information du tableau 3.16) oppose nettement :

- Des départements où les eaux à usage d'*irrigation* sont sur-représentées (Charente Maritime, Landes, Vendée mais avec une qualité de représentation médiocre),
- A des départements où ce sont les *usages industriels* des eaux qui prévalent (Seine Maritime, Nord, Pas de Calais).

En somme ce 1er axe met en exergue le contraste entre usages des eaux (par l'industrie et par l'agriculture) dans des départements d'orientations économiques différentes.

- *Interprétation de l'axe F2*

L'axe F2 contient 36% de l'information du tableau 3.16. Rappelons qu'il est construit à partir des résidus non pris en compte par l'axe F1 : il apporte donc une information complémentaire, indépendante de celle de l'axe F1.

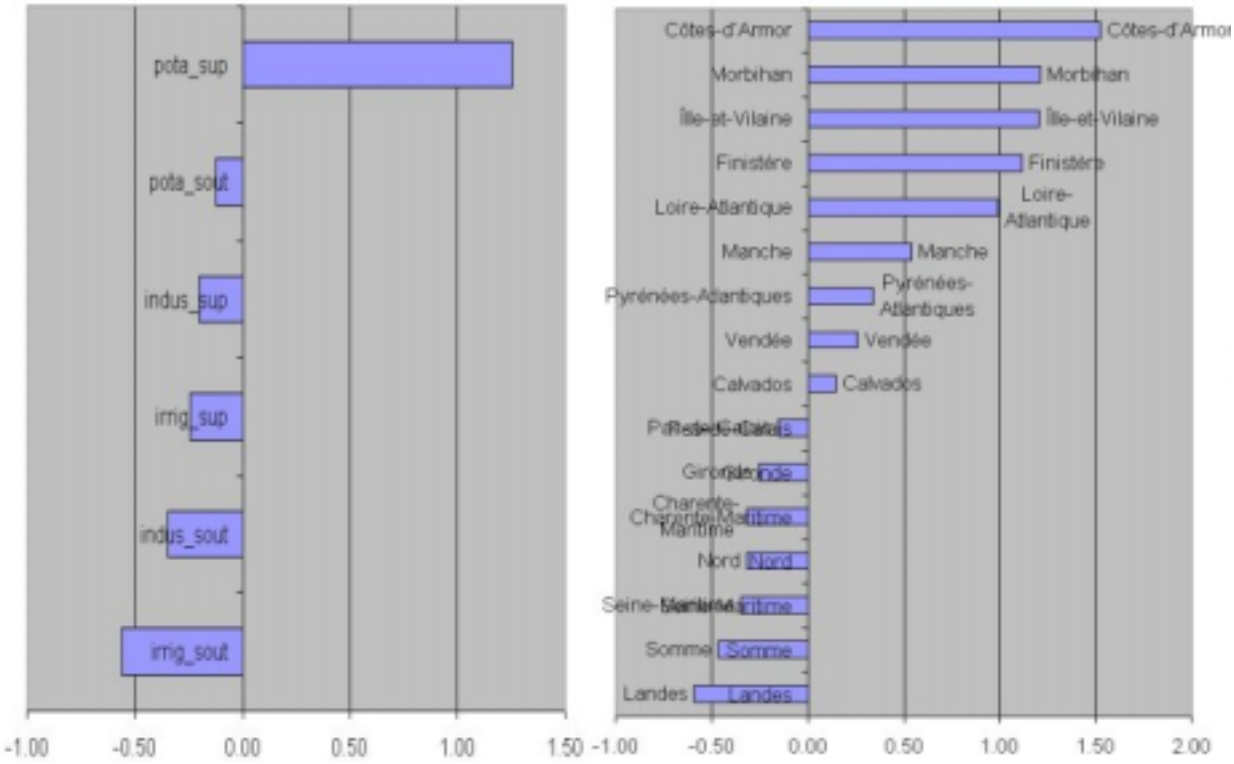


Figure 3.9 : Coordonnées des colonnes et des lignes de l'AFC du tab. 3.16 sur l'axe F2

Les coordonnées significatives sont toutes positives. Tous les départements bretons sont caractérisés, dans leurs consommations, par la sur-représentation, d'eau potable superficielle. En effet, la nature géologique du massif armoricain limite l'importance des nappes d'eau souterraine, la Bretagne est, par ailleurs, faiblement dotée d'industries grosses consommatrices d'eau et l'agriculture intensive bretonne n'est pas fondée sur l'irrigation. On peut donc y parler d'orientation « *démographique* » des usages de l'eau.

Question 3

On s'aidera, pour y répondre, d'une typologie des 16 départements sur le plan F1-F2, de sa cartographie et d'une légende fondée sur le principe suivant : les valeurs moyennes des 6 variables ont été calculées pour chacune des 4 classes, ramenées en % et comparées aux valeurs en % pour l'ensemble des 16 départements, si bien qu'on met ainsi en évidence les écarts (originalités) de chacune des classes par rapport à l'ensemble de la façade atlantique.

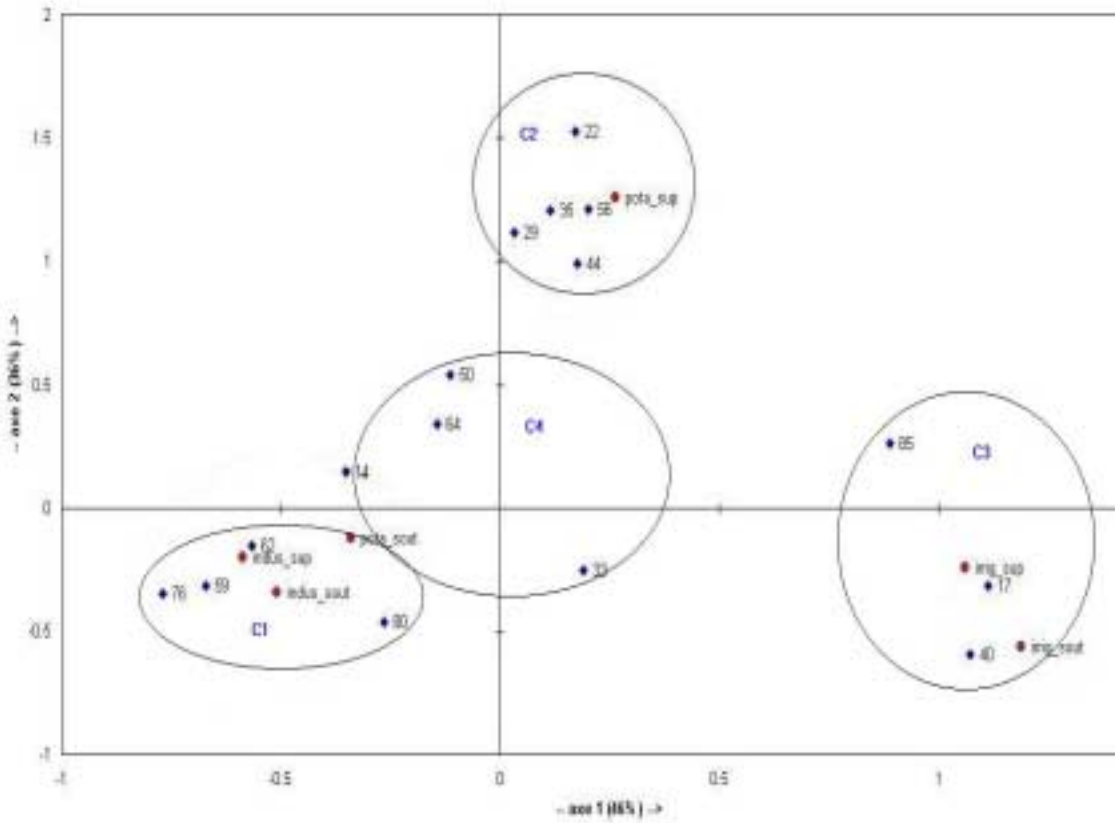


Figure 3.10 : Typologie sur le plan F1-F2 de l'AFC du tableau 3.16

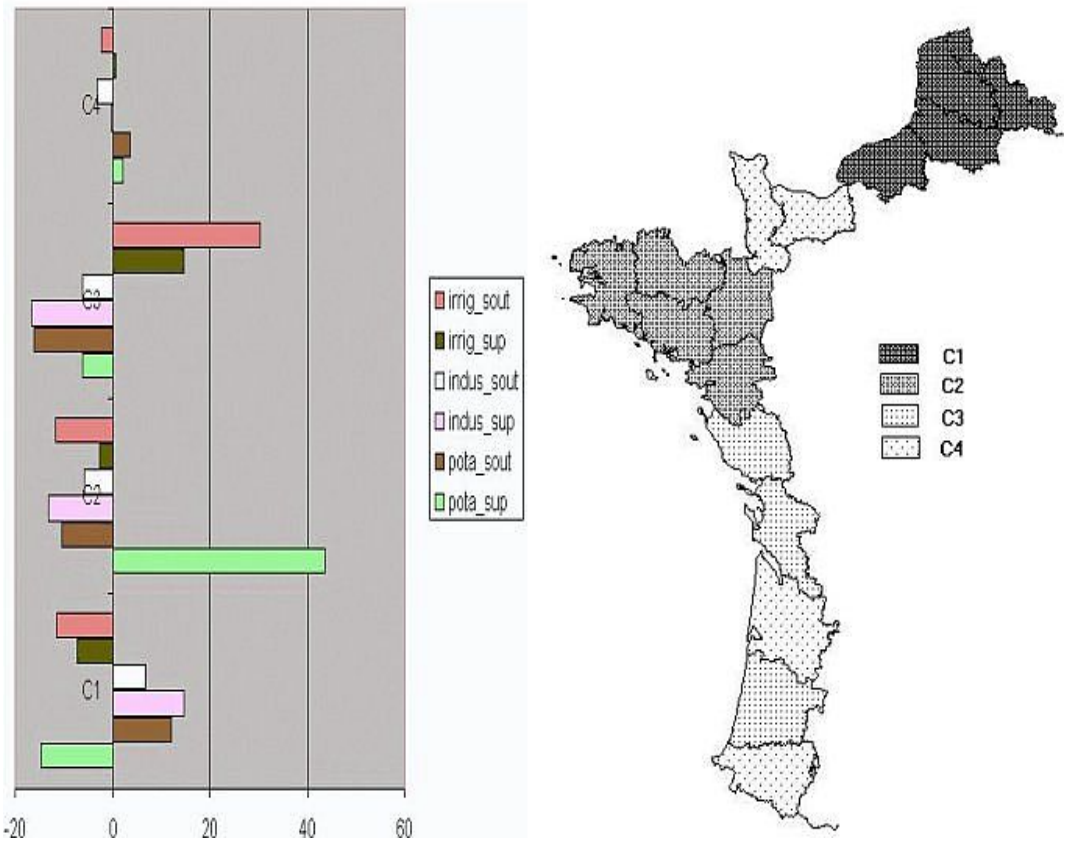


Figure 3.11 : Carte correspondant à la typologie de la fig. 3.10

Les résultats de l'analyse montrent une régionalisation assez nette:

- Nord, Pas de Calais, Seine Maritime où l'usage industriel de l'eau est important,

- Départements bretons où l'usage de l'eau potable est plus fort qu'ailleurs (relative faiblesse des usages industriel et d'irrigation) ; pour des raisons géologiques, la provenance est essentiellement des eaux de surface,

- 3 départements atlantiques (Vendée, Charente Maritime et Landes) font un usage de l'irrigation plus grand qu'ailleurs,

- la classe centrale, regroupant les autres départements, correspondent à deux situations distinctes :
  & 2 départements mal pris en compte par l'analyse (Pyrénées Atlantiques : qr1+qr2=22%, Somme : qr1+qr2=34%),
  & 3 départements (Calvados, Manche, Gironde) de consommations à peu près conformes à la moyenne de la façade atlantique.

Ces résultats, issus d'un tableau suffisamment simple pour donner lieu à un exercice, n'ont rien de révolutionnaires ! Ils ne font qu'ordonner et hiérarchiser des connaissances préalables, ici élémentaires.

Chapitre 4

Analyse des Correspondances Multiples (AFCM)

A) CONNAISSANCES DE BASE

1. Généralités

L'analyse factorielle des correspondances multiples (AFCM) est une extension de l'AFC aux tableaux de contingence multiple, croisant deux à deux m caractères qualitatifs. Ces tableaux sont couramment appelés *tableaux de Burt*.

1.1 Transformation d'un fichier en tableau de Burt

- *Exemple*

	Car1	*Car2*	*Car3*
N°	***opinion***	***sexe***	***csp***
1	**B**	**F**	**Inactif**
11	**M**	**H**	**Indep**
18	**M**	**H**	**Inactif**
25	**B**	**F**	**Inactif**
32	**TM**	**F**	**Inactif**
34	**My**	**F**	**Cadre**
38	**B**	**H**	**Emp-ouvr**
43	**TM**	**F**	**Indep**
64	**My**	**F**	**Indep**
81	**M**	**F**	**Cadre**
91	**B**	**F**	**Inactif**
92	**TM**	**H**	**Inactif**
101	**TB**	**F**	**Cadre**
124	**TB**	**H**	**Inactif**
140	**My**	**F**	**Inactif**
153	**M**	**F**	**Inactif**
158	**My**	**H**	**Inactif**
168	**TB**	**H**	**Indep**
172	**TB**	**H**	**Emp-ouvr**
184	**B**	**H**	**Cadre**
189	**M**	**F**	**Cadre**
211	**My**	**F**	**Indep**
212	**My**	**H**	**Emp-ouvr**
215	**B**	**H**	**Inactif**
231	**B**	**H**	**Inactif**
236	**TM**	**F**	**Cadre**
247	**B**	**H**	**Indep**
269	**B**	**F**	**Cadre**
282	**M**	**F**	**Indep**
283	**B**	**F**	**Inactif**

Tableau 4.1 : Extrait d'une enquête

Le tableau 4.1 est un extrait (échantillon aléatoire au 1/10) d'une enquête d'opinion concernant la réalisation d'un tronçon d'autoroute traversant 7 communes périurbaines.

Seuls 3 caractères qualitatifs, codés alphabétiquement, sont ici conservés :
- son opinion sur le projet (*Car1*): Très Bonne, Bonne, Moyenne, Mauvaise, Très Mauvaise,
- le sexe de l'enquêté (*Car2*): Homme ou Femme,

- sa Catégorie Socio-Professionnelle (*Car3*): Indépendant (exploitant agricole, artisan, commerçant), Cadre (moyen ou supérieur), Employé ou Ouvrier, Inactif (scolaire, retraité, chômeur, femme au foyer),

A chacune de ces questions, chaque enquêté a fourni une réponse et une seule. Le *tableau croisé multiple* (de Burt) correspondant à ce fichier se présente sous la forme du tableau 4.2.

	TB	B	My	M	TM	H	F	Indep	Cadre	Empl	Inactif
TB	***4***	0	0	0	0	*3*	*1*	1	1	1	1
B	0	***10***	0	0	0	*5*	*5*	1	2	1	6
My	0	0	***6***	0	0	*2*	*4*	2	1	1	2
M	0	0	0	***6***	0	*2*	*4*	2	2	0	2
TM	0	0	0	0	***4***	*1*	*3*	1	1	0	2
H	*3*	*5*	*2*	*2*	*1*	***13***	0	*3*	*1*	*3*	*6*
F	*1*	*5*	*4*	*4*	*3*	0	***17***	*4*	*6*	*0*	*7*
Indep	1	1	2	2	1	*3*	*4*	***7***	0	0	0
Cadre	1	2	1	2	1	*1*	*6*	0	***7***	0	0
Emp-ouvr	1	1	1	0	0	*3*	*0*	0	0	***3***	0
Inactif	1	6	2	2	2	*6*	*7*	0	0	0	***13***

Tableau 4.2 : Tableau de Burt correspondant au tableau 4.1

Ce tableau de Burt croise les 3 caractères 2 à 2 (9 sous tableaux de contingence) :

- Les 3 sous tableaux sur la diagonale croisent les 3 caractères avec eux mêmes et ne fournissent donc que les 3 distributions univariées ; seule la diagonale de ces 3 sous tableaux porte un effectif car on ne peut, par exemple, être à la fois homme **et** femme et l'on est soit l'un soit l'autre (les modalités des caractères sont exclusives et exhaustives),
- Les trois sous tableaux de contingence situés au dessus de la diagonale du tableau de Burt croisent opinion et sexe, opinion et csp, sexe et csp,
- Les trois sous tableaux de contingence situés au dessous de la diagonale du tableau de Burt sont identiques aux 3 précédents, sauf que l'on a permuté leurs lignes et colonnes.

L'AFC de ce tableau de Burt donne les résultats résumés suivants :

- *% de variance*

axe 1 : 31%, axe 2 : 19%, axe 3 : 13%, ensemble des 3 axes: 63%

- *l'axe 1*

oppose les opinions extrêmes : très bonne, sur-représentée chez les employés-ouvriers, chez les hommes, versus très mauvaise ou mauvaise chez les cadres et les femmes,

- *l'axe 2*

oppose les très bonnes opinions, fréquentes chez les Indépendants aux opinions bonnes seulement sur-représentées chez les inactifs,

- *l'axe 3*

oppose ceux des cadres qui ont très bonne opinion du projet à tous les indécis.

- *le plan des axes 1 et 2* individualise 4 groupes:

& mauvaise et très mauvaise opinion, sur-représentée chez cadres et les femmes,
& opinion moyenne, sur-représentée chez les indépendants,
& bonne opinion, sur-représentée chez les inactifs,

& très bonne opinion, légèrement sur-représentée chez les employés-ouvriers et les hommes.

Mais, on note immédiatement que l'AFC sur tableau de Burt, partant d'effectifs pour le croisement de modalités identiques en ligne et en colonne, ne nous fournit aucune information sur les individus enquêtés. Or, c'est bien souvent le but d'un dépouillement d'enquête que d'établir une *typologie* d'enquêtés en fonction de leurs profils de réponses aux questions posées. Il faut donc imaginer un codage de l'information du tableau 4.1 qui nous permette de connaître les coordonnées des individus sur les axes factoriels retenus.

1.2 Tableau disjonctif complet

Ce recodage existe sous le nom de *codage disjonctif complet*. Le tableau 4.3 est le tableau disjonctif complet correspondant au tableau 4.1.

N°	TB	B	My	M	TM	H	F	indep	cadre	empl-ouvr	inactif
1	0	1	0	0	0	0	1	0	0	0	1
11	0	0	0	1	0	1	0	1	0	0	0
18	0	0	0	1	0	1	0	0	0	0	1
25	0	1	0	0	0	0	1	0	0	0	1
32	0	0	0	0	1	0	1	0	0	0	1
34	0	0	1	0	0	0	1	0	1	0	0
38	0	1	0	0	0	1	0	0	0	1	0
43	0	0	0	0	1	0	1	1	0	0	0
64	0	0	1	0	0	0	1	0	1	0	0
81	0	0	0	1	0	0	1	0	1	0	0
91	0	1	0	0	0	0	1	0	0	0	1
92	0	0	0	0	1	1	0	0	0	0	1
101	1	0	0	0	0	0	1	0	1	0	0
124	1	0	0	0	0	1	0	0	0	0	1
140	0	0	0	1	0	0	1	0	0	0	1
153	0	0	0	1	0	0	1	0	0	0	1
158	0	0	1	0	0	1	0	0	0	0	1
168	1	0	0	0	0	1	0	1	0	0	0
172	1	0	0	0	0	1	0	0	0	1	0
184	0	1	0	0	0	1	0	0	1	0	0
189	0	0	0	1	0	0	1	0	1	0	0
211	0	0	1	0	0	0	1	1	0	0	0
212	0	0	1	0	0	1	0	0	0	1	0
215	0	1	0	0	0	1	0	0	0	0	1
231	0	1	0	0	0	1	0	0	0	0	1
236	0	0	0	0	1	0	1	0	1	0	0
247	0	1	0	0	0	1	0	1	0	0	0
269	0	1	0	0	0	0	1	0	1	0	0
282	0	0	0	1	0	0	1	0	1	0	0
283	0	1	0	0	0	0	1	0	0	0	1
Σ	**4**	**10**	**5**	**7**	**4**	**13**	**17**	**5**	**9**	**3**	**13**

Tableau 4.3 : Tableau disjonctif complet correspondant au tableau 4.1

Le recodage disjonctif complet de modalités (exclusives et exhaustives) de caractères qualitatifs consiste à créer un tableau où :
- il y a autant de lignes (individus) que dans le tableau initial,
- autant de colonnes qu'il y a de modalités dans l'ensemble des caractères,
- chaque case contient 1 si l'individu présente la modalité et 0 sinon.

Là où il n'y avait que 3 caractères dans le tableau 4.1, il y a 5+2+4 = 11 colonnes dans le tableau 4.3 : on ne peut donc traiter ainsi un très grand nombre de caractères à modalités nombreuses, sauf à créer des tableaux à nombre gigantesque de colonnes.

La somme des cases d'une ligne est, avec ce recodage binaire, toujours égal au nombre de caractères (3 dans l'exemple) et les sommes en colonne indiquent le nombre d'apparitions de chaque modalité (dernière ligne du tableau 4.3). L'effectif total (nombre de 1 dans le tableau) est le nombre d'individus multiplié par le nombre de caractères.

1.3 Equivalence des AFC sur tableau disjonctif complet et tableau de Burt

L'AFCM est une AFC sur tableau binaire (*tableau disjonctif complet*). Par conséquent, elle procède de la même manière que l'AFC:
- calcul des profils-ligne et profils-colonne,
- même ajustement avec pondération par les poids des lignes (constant ici car égal au nombre de caractères) et des colonnes,
- même métrique du Khi² pour déterminer les axes,
- même sur-représentation dans ceux ci des modalités les plus rares.

Les particularités, pour l'interprétation des résultats, tiennent au codage binaire.

& Si tous les caractères présentent seulement 2 modalités, les résultats de l'AFC sur tableaux de contingence, de Burt ou disjonctif complet fournissent exactement les mêmes résultats.

& Si des caractères présentent plus de deux modalités, les résultats de l'AFC sur tableaux de Burt et disjonctif complet sont équivalents *à une dilatation près*. Les facteurs de l'une sont, en effet, co-linéaires de ceux de l'autre. Les % de variance de l'une sont les carrés de ceux de l'autre : par conséquent, l'AFC sur tableaux binaires fournit des % de variance « pessimistes », qu'on ne doit plus strictement interpréter comme *parts d'information* prises en compte par les axes successifs mais comme *hiérarchie de résumés d'information*.

Comme en AFC classique, on peut faire intervenir (a posteriori) des individus supplémentaires, des caractères supplémentaires (ou même seulement telle ou telle de leurs modalités). S'il s'agit d'un caractère quantitatif continu, on le corrélera (comme en ACP) avec les axes factoriels retenus, ce qui fournit sa coordonnée sur chacun d'eux.

2. Résultats de l'AFCM du tableau binaire 4.3

On ne considère ici que la partie des résultats (Tableau 4.4) concernant les modalités des 3 caractères (les individus sont ici anonymes).

- *% de variance*

axe 1 : 21%, axe 2 : 16%, axe 3 :14%, cumul des 3 : 55%, au lieu de 63% pour l'AFCM sur le tableau de Burt (on a donc bien ici une version plus « pessimiste » des quantités d'information prises en compte).

	poids	Coord F1	CTR F1	QR F1	Coord F2	CTR F2	QR F2	Coord F3	CTR F3	QR F3
TB	4.4%	**1.22**	0.23	0.12	**0.86**	0.11	0.08	**1.14**	0.20	0.16
B	11.1%	0.25	0.03	0.01	***-1.03***	0.53	0.28	0.11	0.01	0.00
My	6.7%	0.02	0.00	0.00	**0.76**	0.14	0.09	**-1.23**	0.38	0.28
M	6.7%	***-0.72***	0.13	0.06	0.58	0.08	0.05	0.57	0.08	0.06
TM	4.4%	***-0.79***	0.10	0.05	-0.28	0.01	0.01	-0.43	0.03	0.02
H	14.4%	**0.92**	0.65	0.22	-0.03	0.00	0.00	0.02	0.00	0.00
F	18.9%	**-0.71**	0.65	0.17	0.02	0.00	0.00	-0.01	0.00	0.00
Indep	7.8%	-0.17	0.01	0.00	**1.09**	0.36	0.22	***-0.75***	0.17	0.12
Cadre	7.8%	***-0.81***	0.20	0.09	0.26	0.02	0.01	**1.22**	0.46	0.32
Emp-ouvr	3.3%	**2.13**	0.50	0.27	0.57	0.04	0.03	0.22	0.01	0.01
Inactif	14.4%	0.04	0.00	0.00	***-0.86***	0.56	0.25	-0.31	0.07	0.04

Tableau 4.4 : Résultats de l'AFCM sur les colonnes du tableau 4.3

Les coordonnées des modalités sur les axes factoriels sont les mêmes, *à une homothétie près*, que celles calculées sur le tableau de Burt : par conséquent, l'interprétation des oppositions de modalités sur les différents axes est la même (comme le confirment les valeurs en gras dans le tableau 4.4).

- *Reprenons cette interprétation :*

  - *axe 1*

oppose les opinions extrêmes : très bonne, sur-représentée chez les employés-ouvriers, chez les hommes, très mauvaise ou mauvaise chez les cadres et les femmes,

  - *axe 2*

oppose les très bonnes opinions, fréquentes chez les Indépendants aux opinions bonnes seulement sur-représentées chez les inactifs,

  - *axe 3*

oppose ceux des cadres qui ont très bonne opinion du projet à tous les indécis (opinion moyenne),

- *Représentation graphique sur le plan des axes 1 et 2*

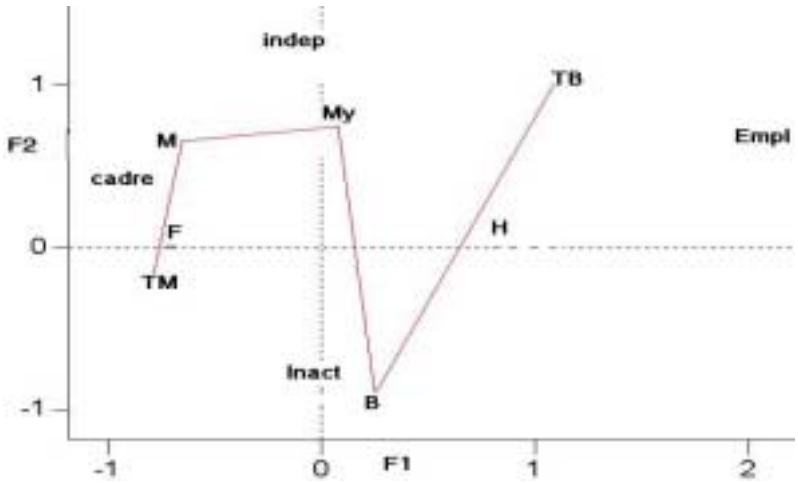


Figure 4.1: modalités du tableau 4.3 sur le plan des axes 1 et 2

L'interprétation est identique à celle faite à partir du tableau de Burt. Rappelons la ici.

Le plan des axes 1 et 2 individualise 4 groupes:
& mauvaise et très mauvaise opinion, sur-représentée chez les cadres et les femmes,
& opinion moyenne, sur-représentée chez les indépendants,
& bonne opinion, sur-représentée chez les inactifs,
& très bonne opinion, légèrement sur-représentée chez les employés-ouvriers et les hommes

D'une façon générale, les représentations et les manipulations graphiques imaginables à partir d'une AFCM sont fort instructives.

- comme en AFC, la proximité entre modalités (de caractères différents) traduit la sur-représentation de l'une dans l'autre (la propension à avoir des profils-ligne comparables) et l'éloignement traduit la sous-représentation.
- La proximité entre individus exprime qu'ils ont des profils-colonne voisins (dans une enquête, patrons comparables de réponses aux différentes questions).
- La proximité de points individus à un point modalité traduit la sur-représentation de celle ci chez ceux là et, de même, l'éloignement traduit leur sous-représentation sur cette modalité.
- Si l'un des caractères est qualitatif *ordinal* (comme l'opinion sur le projet dans notre exemple), il est instructif de relier dans l'ordre les diverses modalités de celui-ci.
- On peut projeter, sur un plan factoriel, une ou plusieurs modalités d'un caractère qualitatif ou une variable quantitative continue et apprécier, comme précédemment décrit, leurs proximité ou éloignement à d'autres points (ligne et/ou colonne). On peut, de même, projeter des individus supplémentaires.
- On peut aussi zoomer sur l'information produite, par exemple en ne représentant que certaines modalités (issues d'un même caractère ou de caractères différents). On peut faire de même pour certains groupes ou types d'individus (par exemple, ne considérer que les femmes).
- Enfin, on doit se souvenir que l'AFCM (comme l'AFC) *valorise les différences rares* (les points modalité sont d'autant plus loin du centre de gravité que leurs fréquences sont faibles).

Au delà de ces avantages, il faut aussi souligner un certain nombre de limites de l'AFCM.

- Le découpage en classes d'une variable quantitative continue (nombre et limites de classes) doit suivre les prescriptions de la statistique univariée,
- Les nombres de modalités des caractères ne doivent pas être trop différents entre eux car cela reviendrait, toutes choses égales par ailleurs, à favoriser implicitement les caractères présentant de nombreuses modalités. Mais, il n'est pas toujours possible d'équilibrer le nombre de modalités des différents caractères: dans notre exemple, le sexe ne peut avoir que 2 modalités tandis qu'il aurait été trop simplificateur de coder les CSP en moins de 4 postes (il y en a 8 dans la plus simple des typologies de l'INSEE). Le codage des opinions en 5 catégories hiérarchisées est classique chez les sociologues (on aurait peut être pu regrouper en 3 modalités : favorable, neutre, défavorable).

Le choix d'effectuer une AFCM sur tableau de Burt ou disjonctif complet dépend des données initiales et des objectifs de l'étude :
& Si les lignes du tableau sont des unités spatiales (un géographe voudra souvent en récupérer les coordonnées pour typologie et cartographie), alors il faut partir d'un tableau disjonctif complet (notez que la plupart des logiciels transforment automatiquement les numéros ou intitulés de modalités, comme ceux du tableau .4.1, en codage booléen),

& Si les modalités des caractères sont déjà binaires (présence / absence, dans un ensemble de communes, d'équipements de différents niveaux de rareté spatiale, appartenance ou non à tel ou tel type de zonage territorial), le choix est déjà fait car le tableau des données est déjà disjonctif complet,

& En situation de dépouillement d'enquête à effectif très nombreux, où l'on s'intéresserait surtout à la relation entre modalités de réponses à diverses questions, il serait plus simple de procéder à une AFCM sur tableau de Burt.

B) EXERCICES CORRIGES

Exercice 1

Une enquête sur un échantillon représentatif (1984, 972 personnes) des citoyens de Californie, a permis de recueillir un grand nombre de variables de statut et d'opinion *(source : SDA test data, Survey Research Center, University of Berkeley)*. Du fichier des résultats nous avons extrait 7 variables que nous avons soumises à une AFCM.

Voici les noms de ces 7 questions et des modalités de réponse:

- « *milit* » : opinion sur les dépenses militaires
1 : trop 3 : comme il faut 5 : pas assez
- « *urban* » **:** opinion sur les dépenses liées aux problèmes urbains
1 : trop 3 : comme il faut 5 : pas assez
- « *crime* **» :** opinion sur les dépenses liées à la criminalité
1 : trop 3 : comme il faut 5 : pas assez
- « *welfare* » **:** opinion sur les dépenses de sécurité sociale
1 : trop 3 : comme il faut 5 : pas assez
- « *ideol* **» :** idéologie politique
1 : libérale 3 : conservatrice 5 : modérée 7 : aucune
- « *parti* » **:** identification à un parti
1 : républicain 3 : démocrate 5 : indépendant 6 : pas de préférence
- « *race* » **:** race ou groupe ethnique
1 : blanc 2 : noir 3 : hispanique 4 : autre

Pour des raisons d'encombrement n'est pas fourni ici le très gros tableau de données.. Les résultats concernant les 972 interviewés ne le sont pas davantage (ils ne présentent, en outre, pas d'intérêt individuellement mais seulement comme ensemble représentatif). Les résultats de l'AFCM sur les 24 modalités de réponse aux 7 questions retenues figurent dans le tableau 4.5 (coordonnées sur 4 axes factoriels). L'analyse portera donc sur elles exclusivement.

- *% de variance des 4 axes retenus*

	F1	F2	F3	F4
% de variance	9.5%	7.3%	6.3%	5.8%

- *Coordonnées des modalités sur les 4 axes retenus*

	effectifs	poids	F1	qr1	F2	qr2	F3	qr3	F4	qr4
milit / 3	334.00	4.91	**0.75**	0.29	-0.11	0.01	0.15	0.01	-0.15	0.01
milit / 1	592.00	8.70	***-0.51***	0.41	0.01	0.00	0.05	0.00	0.06	0.01
milit / 5	**46.00**	0.68	1.15	0.07	0.69	0.02	***-1.79***	0.16	0.30	0.00
urban / 5	646.00	9.49	***-0.32***	0.20	-0.08	0.01	***-0.26***	0.13	-0.11	0.03
urban / 3	266.00	3.91	**0.66**	0.16	-0.34	0.04	**0.62**	0.15	0.12	0.01
urban / 1	**60.00**	0.88	0.54	0.02	**2.42**	0.39	0.02	0.00	0.71	0.03
crime / 5	680.00	9.99	0.00	0.00	0.02	0.00	***-0.30***	0.22	***-0.38***	0.34
crime / 1	**33.00**	0.49	-0.68	0.02	**2.42**	0.21	1.12	0.04	**2.15**	0.16
crime / 3	259.00	3.81	0.08	0.00	-0.37	0.05	**0.66**	0.16	**0.73**	0.19
welfare / 1	188.00	2.76	**0.73**	0.13	**0.87**	0.18	-0.48	0.06	-0.35	0.03
welfare / 5	469.00	6.89	***-0.61***	0.34	0.00	0.00	-0.11	0.01	0.14	0.02
welfare / 3	315.00	4.63	0.47	0.10	-0.52	0.13	0.46	0.10	0.00	0.00
ideol / 3	252.00	3.70	**0.89**	0.28	0.43	0.06	***-0.61***	0.13	0.31	0.03
ideol / 1	326.00	4.79	***-0.73***	0.27	-0.11	0.01	-0.14	0.01	**0.44**	0.10
ideol / 5	351.00	5.16	0.09	0.00	-0.44	0.11	0.33	0.06	***-0.54***	0.16
ideol / 7	**43.00**	0.63	-0.41	0.01	1.96	0.18	**1.94**	0.17	-0.75	0.03
parti / 5	263.00	3.87	-0.26	0.03	-0.14	0.01	0.29	0.03	-0.10	0.00
parti / 3	421.00	6.19	***-0.53***	0.22	-0.09	0.01	-0.29	0.06	0.07	0.00
parti / 6	**58.00**	0.85	0.01	0.00	**1.89**	0.23	**1.84**	0.21	-0.94	0.06
parti / 1	230.00	3.38	**1.27**	0.50	-0.15	0.01	-0.27	0.02	0.22	0.02
race / 4	81.00	1.19	-0.07	0.00	0.25	0.01	0.66	0.04	***-1.29***	0.15
race / 3	**52.00**	0.76	-0.59	0.02	0.76	0.03	0.00	0.00	1.06	0.06
race / 2	131.00	1.93	-0.60	0.06	0.52	0.04	-0.58	0.05	***-0.94***	0.14
race / 1	708.00	10.41	0.16	0.07	-0.18	0.09	0.03	0.00	0.24	0.16

Tableau 4.5 : effectifs, coordonnées et qualités de représentation des 24 modalités

Questions

1) **Commentez les effectifs par modalités des 7 questions sélectionnées : comment joueront ils dans l'AFCM ?**
2) **Pourquoi avoir fait une AFCM sur le tableau 4.5 ? Quels en sont les avantages et inconvénients ?**
3) **Interprétez chacun des 4 axes factoriels retenus**
4) **Pour approfondir votre compréhension de cette enquête, construisez quelques tableaux croisant 2 variables (« tris croisés ») et interprétez les.**

Réponses suggérées

Question 1

La colonne 2 du tableau 4.5 fournit les effectifs par modalité de réponse aux 7 questions.

- *dépenses militaires*

la majorité (60.9%) trouve qu'elles sont trop élevées et seulement 4.7% qu'elles ne le sont pas

- *dépenses pour les problèmes urbains*

les 2/3 des interviewés les trouvent insuffisantes, seulement 6.2% insuffisantes

- *dépenses de lute anti-criminalité*

70% les trouvent insuffisantes et seulement 3.4% exagérées

- *dépenses pour la sécurité sociale*

les effectifs sont ici assez bien répartis : une majorité relative pense qu'elles ne sont pas suffisantes et 20% qu'elles sont trop grandes

- *idéologie revendiquée*

idéologies libérale, modérée et conservatrice ont des effectifs notables, moins de 5% des enquêtés n'en affichent aucune

- *parti reconnu*

une majorité relative se reconnaît dans le parti démocrate (au moment de l'enquête), conservateurs (avoués) et indépendants sont de taille comparable tandis que 6% n'affichent pas de préférence

- *répartition ethnique/raciale*

72.8% des enquêtés sont blancs, 13.5% noirs, 5.3% hispanique et 8.3% d'une autre catégorie.

La répartition ethnique est globalement conforme à celle de la population des USA tandis que les répartitions idéologique et politique semblent sur-représenter orientation libérale et parti démocrate, ce qui pourrait expliquer une partie des opinions majoritaires sur les dépenses militaires et pour régler problèmes urbains et sociaux. On doit cependant noter le nombre très grand d'enquêtés pensant qu'il faut dépenser plus pour stopper l'essor de la criminalité.

La métrique du Khi^2 utilisée par AFC et AFCM a pour résultat (à cause de la division par l'effectif en colonnes) d'accorder beaucoup d'importance aux modalités peu fréquentes : opinions de minorités (extrêmistes voulant plus de dépenses militaires et moins de sociales ou gens sans parti ni idéologie ou encore hispaniques et autres ethnies).

Question 2

L'AFCM est la seule analyse factorielle possible sur un tableau de données connues individu par individu, croisant individus en lignes (972 enquêtés ici) et variables qualitatives (ou codées en classes) en colonnes. Dans chaque case d'un tel tableau figure un code, alphabétique ou numérique (repérant alors un numéro de modalité). L'AFCM est opérée sur le tableau de Burt (croisant toutes les variables 2 à 2) ou sur le tableau disjonctif complet (tableau binaire qui, dans cet exemple, a 972 lignes et 24 colonnes).

L'avantage principal de l'AFCM est de créer des résumés hiérarchisés et de permettre une vision beaucoup plus synthétique d'un tableau (comme le 4.5) qu'un grand nombre de tableaux de contingence croisant les variables 2 à 2.

L'inconvénient principal de l'AFCM (à cause du recodage booléen qui est « pauvre » et crée un grand nombre de colonnes) est de générer des axes ne prenant chacun en compte que peu de variance, ce qui offre une vision pessimiste de l'informtion prise en compte. Une seconde propriété, commune à l'AFC et à l'AFCM, est de valoriser les *différences rares* (c'est l'effet de la pondération inverse par la fréquence des modalités). On peut donc s'attendre ici à ce qu'apparaissent beaucoup les opinions de minorités tranchées davantage que celles des majorités « banales ».

Question 3

Pour chaque interprétation d'axe, on ne retiendra que les modalités à coordonnée relativement importante en valeur absolue et dont la qualité de représentation est correcte (les deux étant point trop voisins de 0) : elles sont signalées en gras dans le tableau 4.5.

- *L'axe 1 oppose* :
  & des gens proches de l'aile conservatrice du parti républicain, trouvant correctes les dépenses militaires mais exagérées les dépenses sociales,
  & un segment « libéral » (centriste) de l'opinion publique américaine, appartenant assez souvent au parti démocrate, professant que les dépenses sociales (sécurité sociale, problèmes urbains) sont insuffisantes et exagérées les dépenses militaires;

- *L'axe 2 oppose*:
  & un segment de population sans préférence de parti ou idéologique mais jugeant les dépenses publiques liées à la sécurité sociale trop importantes. On peut parler de minorité exprimant un extrêmisme « droitier » sans conscience politique.
  & le reste de la population

- *L'axe 3 oppose* :
  & encore des sans parti ni idéologie affichés mais plus modérés que les précédents, jugeant correctes les dépenses liées à la criminalité et aux problèmes urbains,
  & un ensemble d'idéologie plutôt conservatrice, jugeant insuffisantes les dépenses militaires, anti-criminalité et dédiées aux problèmes urbains.

- *L'axe 4 oppose* :
  & des interviewés trouvant exagérées les dépenses anti-criminalité (les républicains y sont légèrement sur-représentés)
  & un segment d'idéologie modérée les trouvant insuffisantes : y sont sur-représentés noirs et minorités autres que blancs et hispaniques.

Question 4

Globalement, ces 4 axes factoriels mettent en valeur des opinions et statuts souvent minoritaires, ce qui est fréquemment le propre des AFC. On peut compléter la vision qu'elle fournit par quelques « coups de projecteur » sur des relations 2 à 2 entre des variables, ce qui amène à construire des tableaux de contingence. L'ensemble de ceux ci, avec 7 variables, est de 7*6/2=21 : il ne serait pas raisonnable de tous les considérer et il faut donc choisir. On a ici choisi d'explorer :

- la relation idéologie – parti (y a t'il cohérence de ces déclarations ?),
- la relation race/ethnie – parti (y a t'il vote ethnique ?)
- les relations entre idéologie et opinions sur les dépenses militaires, de sécurité sociale, anti-criminalité (y a t'il des relations quasi mécaniques ou non ?).

Pour chacune de ces relations nous avons calculé les effectifs observés (N_{ij}) et théoriques (N'_{ij}), le Khi² provenant de chaque case ($(N_{ij}/N'_{ij})^2/ N'_{ij}$) et leur somme (Khi² total). Nous avons affecté aux khi² de chaque case le signe – si $N_{ij} < N'_{ij}$ (+ sinon) et représenté ces valeurs (positives ou négatives) sous forme de graphiques en barres (représentant les sur ou sous représentations des modalités de la variable 1 dans celles de la variable 2).

- *relation idéologie – parti*

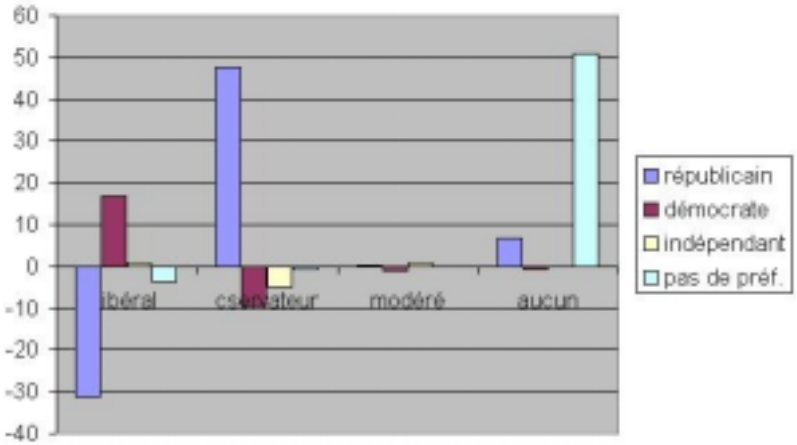


Figure 4.2 : sur/sous représentations des partis selon l'idéologie affichée

Le Khi² calculé valant 175.8 est hautement significatif (avec un risque d'erreur α inférieur à 1 pour mille) : il y a donc forte relation (redondance ?) entre idéologie et parti (les réponses des interviewés sont cohérentes sur ce point). On notera sur la figure 4.2 les forts écarts positifs ou négatifs concernant les républicains (opinions tranchées, militantes) et les écarts plus faibles concernant les démocrates (d'idéologie semble t'il plus mesurée). La forte sur-représentation des sans idéologie chez les sans parti est logique (s'agit il d'un électorat « flottant » ou de non électeurs ?).

- *relation race/ethnie – parti*

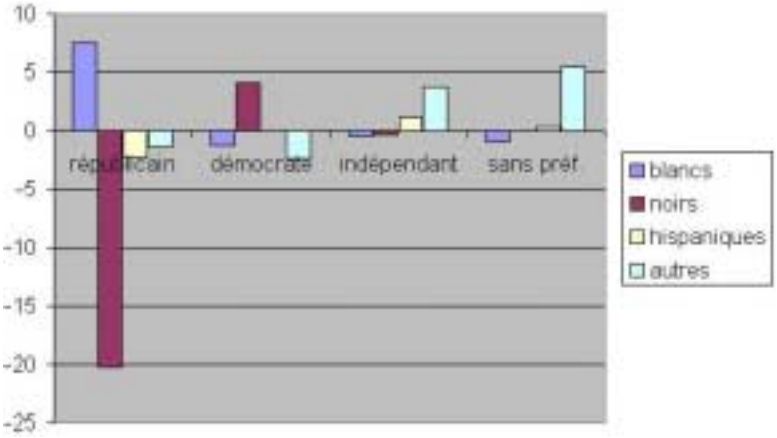


Figure 4.3 : sur/sous représentations des partis selon l'ethnie déclarée

Le Khi² est ici moins fort (61.4) mais toujours significatif d'une relation avec un risque d'erreur α inférieur à 1 pour mille. On notera, parmi les sympathisants républicains, la forte sous-représentation des noirs et la sur-représentation des blancs comme la relative importance des sans parti net chez les hispaniques et autres ethnies (signe qu'ils se sentent moins concernés ?).

- *relation idéologie – dépenses militaires*

Khi² de 73.8, toujours significatif d'une relation, au risque d'erreur α inférieur à 1 pour mille.

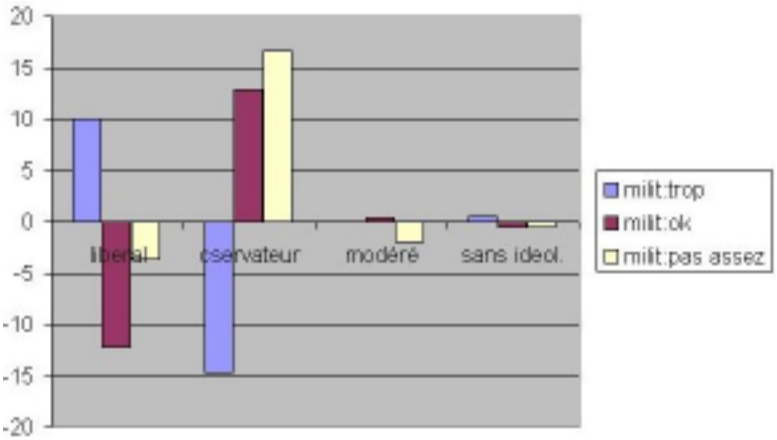


Figure 4.4 : sur/sous représentations des partis pour les dépenses militaires

Les libéraux (au sens américain du terme, quelque chose comme centristes en Europe) trouvent globalement exagérées les dépenses militaires tandis que les conservateurs les trouvent plus souvent convenables ou insuffisantes. Les modérés et les sans idéologie ne manifestent pas d'opinion tranchée sur le sujet.

- *relation idéologie – dépenses de sécurité sociale*

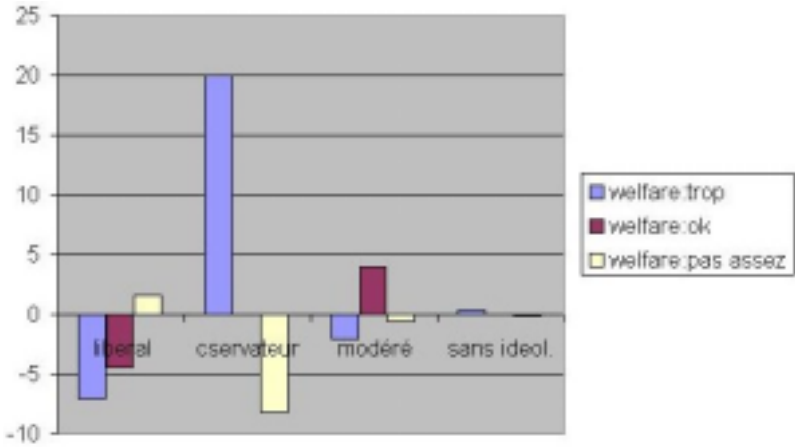


Figure 4.5 : sur/sous représentations des idéologies pour les dépenses sociales

Khi² de 58.4, significatif d'une relation non due au hasard entre les 2 questions (toujours avec un risque d'erreur α inférieur à 1 pour mille). Le principal fait notable ici est la forte sur-représentation des républicains trouvant exagérées les dépenses de sécurité sociale (et le faible soutien des démocrates pour les augmenter !).

- *relation idéologie – dépenses pour diminuer la criminalité*

Khi² de 16.9, significatif au risque d'erreur α égal à 1% mais traduisant une relation bien moins avérée que précédemment. L'opinion sur les dépenses destinées à lutter contre la criminalité ne fait pas apparaître de clivage net entre sympathisants républicains et démocrates (signifiant que le souci sécuritaire est à peu près partagé). Chez les sans idéologie,

la fréquence des opinions « trop d'argent contre la criminalité » concerne trop peu de personnes pour signifier quoi que ce soit de collectif.

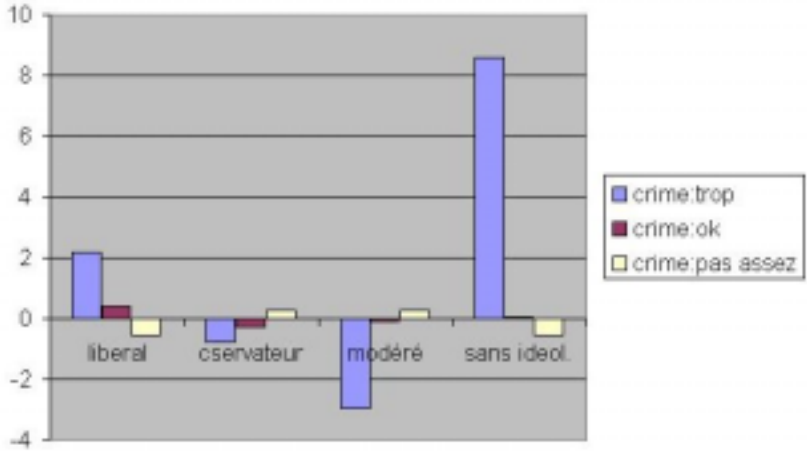


Figure 4.6 : sur/sous représentations des idéologies sur les dépenses anti-criminalité

Exercice 2

Un hebdomadaire a publié en 2003 un classement de 34 villes françaises sur chacun d'une dizaine d'indicateurs de « gestion environnementale ».

Nous en retenons ici 5 :

- *TC* : investissements récents dans les transports en commun,
- *VP* : voies piétonnières,
- *PC* : pistes cyclables,
- *EV* : espaces verts,
- *TS* : tri sélectif.

Pour chacun de ces 5 descripteurs, les données originelles étaient des rangs (de 1 à 34).Pour chacun des 5 descripteurs, nous avons réparti ces rangs en 3 classes d'effectifs voisins:

- *classe 1* : villes de rang 1 à 11 inclus (tiers supérieur),

- *classe 2* : villes de rang 12 à 22 inclus (tiers moyen),

- *classe 3* : villes de rang 23 à 34 inclus (tiers inférieur).

Une AFCM a été pratiquée sur ce tableau de 34 lignes et 5*3 colonnes (modalités). Le tableau 4.6 présente les données qui lui ont été soumises.

	TC	VP	PC	EV	TS
Aix/Prov.	3	1	3	3	2
Amiens	2	2	2	3	3
Angers	3	3	1	1	1
Besançon	2	1	2	1	2
Bordeaux	1	3	1	2	1
Brest	3	3	1	1	2
Caen	2	1	3	1	1
Clermont	2	1	3	2	2
Dijon	3	2	3	1	2
Grenoble	1	2	1	3	1
LeHavre	3	3	2	1	2
LeMans	3	2	2	2	2
Lille	1	2	2	2	3
Limoges	3	1	3	1	1
Lyon	1	2	2	3	3
Marseille	2	3	3	3	2
Metz	3	1	3	1	1
Montpellier	1	1	1	1	1
Mulhouse	2	1	2	3	1
Nancy	1	1	3	2	1
Nantes	2	2	1	2	3
Nice	3	3	3	3	3
Nimes	3	3	3	1	3
Orleans	1	3	2	2	1
Paris	2	3	1	3	2
Perpignan	3	2	3	3	3
Reims	3	3	2	3	3
Rennes	1	2	1	1	2
Rouen	1	2	2	2	3
StEtienne	2	3	2	2	3
Strasbourg	1	1	1	3	1
Toulon	2	3	3	3	3
Toulouse	2	2	1	2	2
Tours	1	1	1	2	3

Tableau 4.6 : répartition en 3 classes de même effectif de 34 villes françaises (5 variables)

De cette AFCM, nous avons retenu 4 axes factoriels. En voici les aides à l'interprétation.

- *% de variance des axes factoriels*

	F1	F2	F3	F4
% de variance	20%	18%	12.5%	11.5%
% cumulé	-	38%	40.5%	52.0%

Tableau 4.7 : % de variance des 4 axes factoriels

- *résultats concernant les modalités*

	F1	*ctr1*	*qr1*	F2	*ctr2*	*qr2*	F3	*ctr3*	*qr3*	F4	*ctr4*	*qr4*
TC / 3	**0.74**	9.57	0.30	***-0.60***	7.17	0.20	0.22	1.43	0.03	***-0.73***	16.47	0.29
TC / 2	-0.16	0.42	0.01	-0.38	2.65	0.07	-0.16	0.71	0.01	**1.23**	42.46	0.72
TC / 1	***-0.65***	6.67	0.20	**1.04**	19.58	0.51	-0.08	0.17	0.00	-0.43	5.18	0.09
VP / 1	**0.68**	7.28	0.22	**0.71**	9.31	0.24	***-0.59***	9.17	0.16	0.53	7.75	0.13
VP / 2	***-0.85***	11.52	0.35	-0.12	0.25	0.01	0.46	5.72	0.10	-0.13	0.45	0.01
VP / 3	0.16	0.44	0.01	***-0.55***	***5.99***	***0.16***	0.11	0.37	0.01	-0.37	4.11	0.07
PC / 3	**0.79**	10.78	0.34	-0.39	3.06	0.08	***-0.55***	8.94	0.17	0.09	0.23	0.00
PC / 2	***-0.76***	9.29	0.28	-0.34	2.09	0.05	-0.06	0.09	0.00	-0.09	0.23	0.00
PC / 1	-0.10	0.15	0.00	**0.77**	10.71	0.28	**0.66**	11.69	0.21	0.00	0.00	0.00
EV / 3	-0.15	0.40	0.01	***-0.53***	5.52	0.15	***-0.69***	13.83	0.26	0.04	0.05	0.00
EV / 1	**1.00**	16.01	0.48	0.15	0.43	0.01	**0.56**	8.31	0.15	-0.30	2.55	0.04
EV / 2	***-0.84***	11.16	0.34	0.42	3.22	0.08	0.19	1.00	0.02	0.26	1.83	0.03
TS / 2	0.36	2.02	0.06	-0.41	3.00	0.08	**0.99**	26.07	0.47	**0.66**	12.14	0.21
TS / 3	***-0.78***	10.47	0.33	***-0.59***	6.93	0.19	-0.46	6.06	0.11	-0.42	5.45	0.10
TS / 1	0.49	3.84	0.12	**1.05**	20.09	0.53	-0.49	6.42	0.12	-0.20	1.10	0.02

Tableau 4.8 : aides à l'interprétation de 4 axes de l'AFCM du tableau 4.6 (variables)

- *résultats concernant les villes*

	F1	*ctr1*	*qr1*	F2	*ctr2*	*qr2*	F3	*ctr3*	*qr3*	F4	*ctr4*	*qr4*
Aix/Prov.	**0.76**	4.14	0.30	-0.41	1.37	0.09	-0.25	0.77	0.03	0.24	0.74	0.03
Amiens	***-0.85***	5.21	0.36	***-0.66***	3.56	0.22	-0.37	1.63	0.07	0.26	0.89	0.03
Angers	**0.72**	3.77	0.26	0.28	0.63	0.04	0.43	2.27	0.09	***-0.67***	5.68	0.22
Besançon	0.35	0.88	0.06	-0.09	0.06	0.00	0.30	1.09	0.04	**0.84**	9.04	0.34
Bordeaux	-0.29	0.62	0.04	**0.91**	6.94	0.41	0.16	0.32	0.01	-0.31	1.22	0.05
Brest	**0.68**	3.34	0.23	-0.21	0.38	0.02	**1.03**	12.93	0.54	-0.31	1.24	0.05
Caen	**0.88**	5.57	0.38	0.38	1.23	0.07	-0.50	3.06	0.12	**0.56**	4.00	0.15
Clermont	0.26	0.48	0.03	-0.01	0.00	0.00	-0.05	0.03	0.00	**1.15**	16.81	0.65
Dijon	**0.64**	2.96	0.21	-0.46	1.73	0.10	**0.68**	5.62	0.23	-0.17	0.38	0.02
Grenoble	-0.39	1.12	0.08	**0.74**	4.56	0.27	-0.05	0.04	0.00	-0.30	1.13	0.04
LeHavre	0.47	1.60	0.11	***-0.58***	2.83	0.17	**0.74**	6.65	0.28	-0.35	1.54	0.06
LeMans	-0.42	1.30	0.09	-0.35	1.01	0.06	**0.74**	6.55	0.26	-0.02	0.00	0.00
Lille	***-1.22***	10.70	0.72	0.14	0.16	0.01	0.03	0.01	0.00	-0.34	1.47	0.06
Limoges	**1.16**	9.75	0.68	0.31	0.80	0.05	-0.35	1.45	0.06	-0.26	0.85	0.03
Lyon	***-1.00***	7.25	0.50	-0.18	0.27	0.02	-0.33	1.34	0.06	-0.43	2.33	0.09
Marseille	0.31	0.70	0.05	***-0.76***	4.75	0.30	-0.12	0.19	0.01	**0.69**	6.03	0.24
Metz	**1.16**	9.75	0.68	0.31	0.80	0.05	-0.35	1.45	0.06	-0.26	0.85	0.03
Montpellier	0.45	1.45	0.10	**1.25**	12.94	0.75	0.03	0.01	0.00	-0.17	0.37	0.01
Mulhouse	0.03	0.01	0.00	0.17	0.25	0.01	***-0.81***	7.89	0.32	**0.63**	5.04	0.19
Nancy	0.15	0.16	0.01	**0.95**	7.48	0.44	***-0.62***	4.59	0.19	0.10	0.13	0.00
Nantes	***-0.85***	5.28	0.36	0.03	0.01	0.00	0.28	0.98	0.04	0.39	1.93	0.07
Nice	0.24	0.41	0.03	***-0.89***	6.60	0.43	***-0.55***	3.72	0.17	***-0.58***	4.29	0.18
Nimes	**0.60**	2.62	0.19	***-0.66***	3.65	0.23	-0.05	0.03	0.00	***-0.72***	6.66	0.28
Orleans	-0.50	1.82	0.12	**0.54**	2.45	0.14	-0.13	0.20	0.01	-0.35	1.52	0.06
Paris	0.03	0.01	0.00	-0.37	1.12	0.07	0.37	1.66	0.07	**0.65**	5.38	0.21
Perpignan	-0.08	0.04	0.00	***-0.75***	4.63	0.30	-0.41	2.06	0.09	-0.48	2.94	0.12
Reims	-0.25	0.44	0.03	***-0.87***	6.34	0.41	-0.35	1.51	0.07	***-0.65***	5.46	0.23
Rennes	-0.07	0.04	0.00	0.48	1.92	0.11	**1.05**	13.44	0.53	-0.08	0.09	0.00
Rouen	***-1.22***	10.70	0.72	0.14	0.16	0.01	0.03	0.01	0.00	-0.34	1.47	0.06
StEtienne	***-0.75***	4.03	0.28	-0.48	1.93	0.12	-0.15	0.27	0.01	0.25	0.81	0.03
Strasbourg	0.09	0.05	0.00	**1.02**	8.64	0.51	-0.48	2.80	0.11	-0.03	0.01	0.00
Toulon	-0.04	0.01	0.00	***-0.82***	5.56	0.36	***-0.71***	6.13	0.27	0.24	0.72	0.03
Toulouse	***-0.50***	1.80	0.12	0.10	0.08	0.00	**0.87**	9.19	0.36	**0.84**	8.97	0.34
Tours	***-0.53***	2.01	0.14	**0.79**	5.15	0.30	-0.11	0.14	0.01	-0.03	0.01	0.00

Tableau 4.9: aides à l'interprétation de 4 axes de l'AFCM du tableau 4.6 (villes)

Questions

1) **Procédez à un examen critique du tableau de données (4.6)**
2) **Quel a été l'effet sur les résultats de l'FCM de répartir les villes en 3 classes d'effectif à peu près égal ?**
3) **Interprétez les 4 axes factoriels retenus**
4) **Construisez sur le plan des axes 1 et 2 une typologie commentée.**

Réponses suggérées

Question 1

Procéder à l'examen critique de données ne signifie pas en faire une critique radicale mais en dégager avantages et limites.

Les limites sautent aux yeux et sont de divers ordres :

& du point de vue géographique, le grand sud ouest est nettement sous représenté (Pau, Biarritz-Bayonne, La Rochelle ne figurent pas dans l'échantllon),
& les descripteurs présents dans le tableau traitent soit de l'état de l'environnement (mais alors manquent, entre autres, des données sur les pollutions et la fréquentation des transports en commun) soit de politiques correctrices,
& ne sont précisées ni les sources de chacun des descripteurs (indice de leur fiabilité) ni les périodes précises d'observation (dates, durées),
& la transformation des valeurs en rangs puis en classes de rangs représentent une perte de précision certaine.

Mais le tableau 4.6 présente aussi un certain nombre de vertus :

& les informations localisées sur l'environnement sont à la fois rares et dispersées dans des publications très disparates, couvrant des zones géographiques rarement complètes (et à des échelles fort différentes),
& or c'est précisément en zone urbaine que l'état de l'environnement pose le plus de problèmes et que les actions correctrices sont le plus nécessaires,
& sur un plan méthodologique, l'exercice (même sur données imparfaites) illustre la possibilité de combiner des typologies spatiales (et éventuellement des zonages).

question 2

On sait (et l'exercice précédent l'a amplement démontré) que les AFC (dont l'AFCM) valorisent les différences rares en « gonflant » l'importance des modalités de faible fréquence. Les 15 modalités du tableau 4.6 étant toutes de fréquences très voisines (11 ou 12 sur 34), cet inconvénient disparaît : l'AFCM du tableau 4.6 comparera sans cet effet là les profils des 34 villes sur les 5 descripteurs retenus. Rappelons aussi qu'en AFCM les sommes en ligne sont toutes égales (au nombre de descripteurs puisqu'une seule modalité est possible pour chacun, (modalité codée 1 dans le tableau disjonctif complet lors que les autres sont codées 0).

question 3

Les interprétations tiendront compte à la fois des coordonnées, des contributions relatives et des qualités de représentation : les plus notables sur chaque axe sont en gras dans les tableaux 4.8 et 4.9. et encadrées dans les figures 4.7 à 4.10. Rappelons que la proximité d'un point modalité et d'un point ville signifie sur représentation de cette modalité dans cette ville, leur éloignement signifiant sous représentation. On tiendra surtout compte des modalités « extrêmes » : classe 1 (11 premiers rangs sur le critère) et classe 3 (12 derniers rangs).

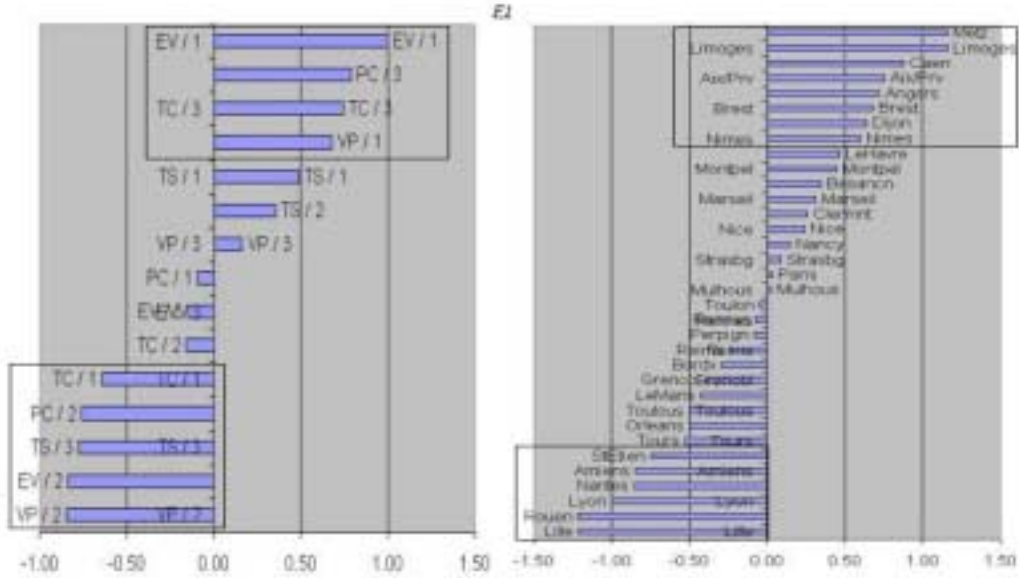


Figure 4.7 : coordonnées des modalités et des villes sur l'axe 1

- L'axe 1 (20% de la variance) oppose :

& des villes (Metz, Limoges, Caen, Aix en Provence, Angers, Brest, Dijon, Nîmes) à faible investissement dans les transports en commun et les pistes cyclables mais bien dotés en voies piétonnières et espaces verts,

& des villes (Lille, Rouen, Lyon, Nantes, Amiens, Saint Etienne) à retard certain en matière de gestion des déchets (tri sélectif) mais effort récent pour les transports en commun.

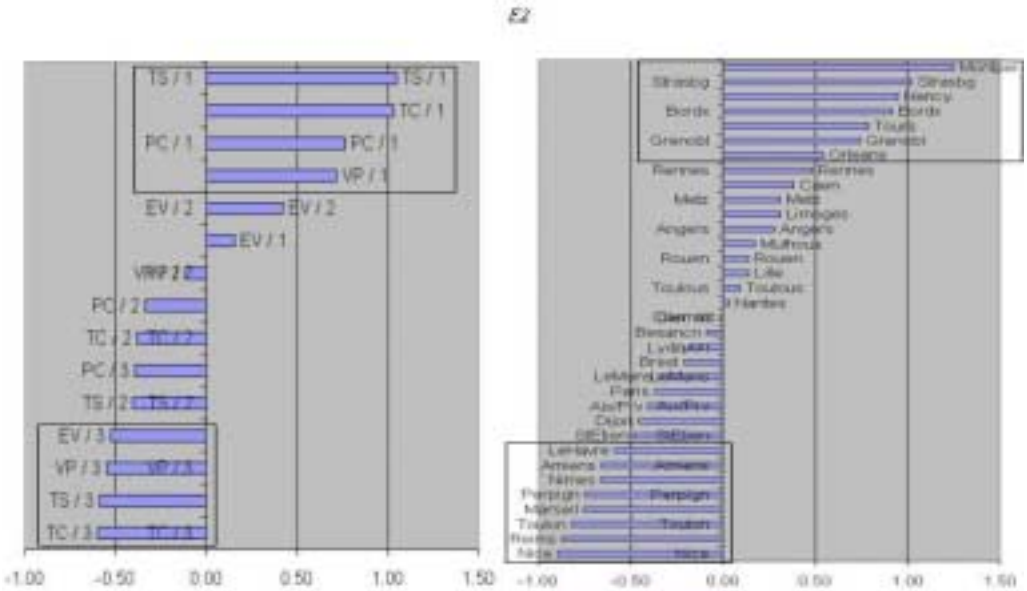


Figure 4.8 : Coordonnées des modalités et les villes sur l'axe 2

- L'axe 2 (18% de la variance) oppose :
& des villes « vertes » à politique environnementale marquée, où tous les indicateurs sont dans le premier 1/3 (tri sélectif, investissement récent dans les transports en commun, pistes cyclables, voies piétonnières) ; il s'agit de Montpellier, Strasbourg, Nancy, Bordeaux, Tours, Grenoble, Orléans,
& des villes en queue de peloton pour les mêmes indicateurs et les espaces verts (Nice, Reims, Toulon, Marseille, Perpignan, Nîmes, Amiens, Le havre) où l'on note la forte présence des villes méditerranéennes (sauf Montpellier).

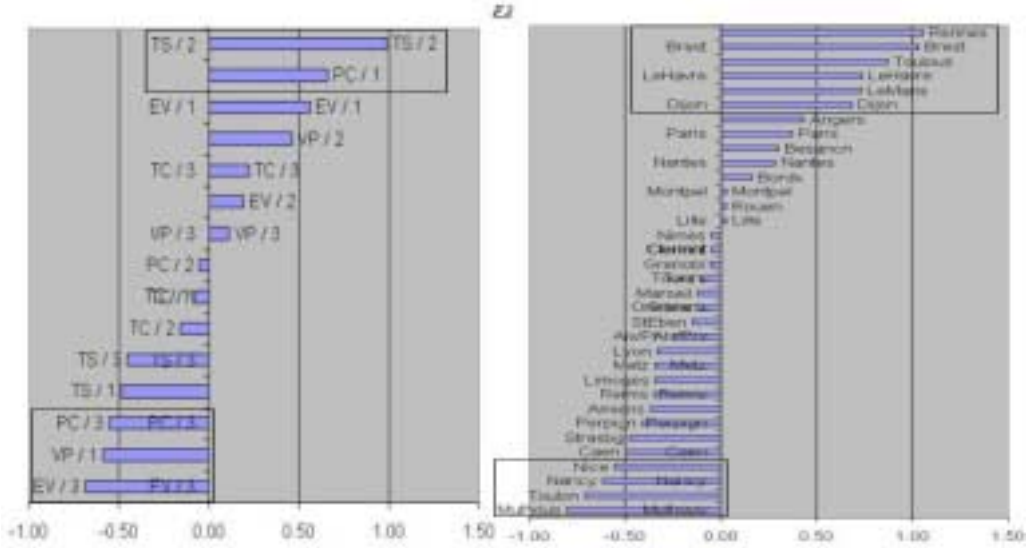


Figure 4.9 : Coordonnées des modalités et les villes sur l'axe 3

- L'axe 3 (12.5% de variance) oppose :
& des villes (Rennes, Brest, Toulouse, Le Havre, Le Mans, Dijon) surtout marquées par l'importance de leur réseau de pistes cyclables,
& à des villes (Mulhouse, Toulon, Nancy, Nice) ayant une forte surface de voies piétonnières mais peu d'espaces verts et de pistes cyclables.

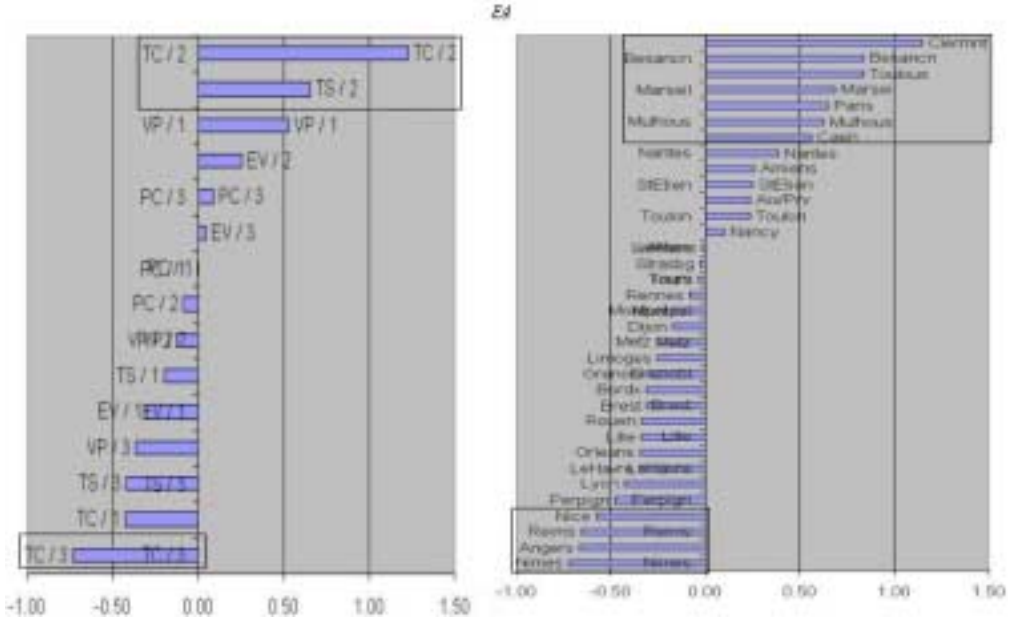


Figure 4.10 : Coordonnées des modalités et les villes sur l'axe 4

- L'axe 4 (11.5% de variance) oppose :

& des villes (Clermont Ferrand, Besançon, Toulouse, Marseille, Paris, Mulhouse, Caen) ayant des scores moyens sur les investissements « transports en commun » et tri sélectif,
& des villes (Nîmes, Angers, Reims, Nice) caractérisées surtout par la faiblesse de leurs investissements récents sur le poste « transports en commun.

Question 4

Les 4 résumés ci-dessus n'offrent qu'une vision analytique : il faudrait pouvoir les combiner pour obtenir une typologie synthétique des villes vis à vis de leurs politiques et état de l'environnement (combinaison que nous apprendrons à maîtriser au chapitre 5). Dans l'état actuel de vos apprentissages, nous pouvons procéder au dessin de « patatoïdes » sur le plan des axes 1 et 2 et commenter (voire cartographier) la typologie obtenue.

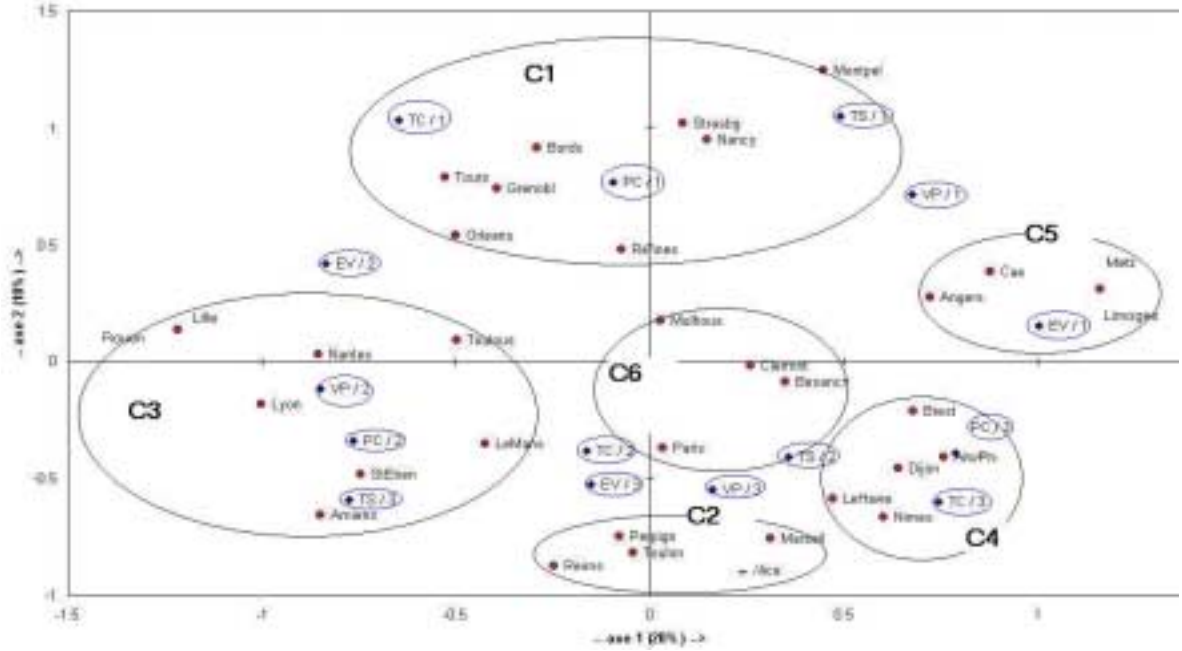


Figure .4.11 : Plan factoriel des axes F1 - F2 et typologie des villes

Six classes, plus ou moins homogènes, ont été dessinées :

- la *classe C1* est surtout celle des villes à bonnes « performance » et action environnementales, notamment pour l'effort récent en faveur de transports moins ou pas polluants (transports en commun, pistes cyclables) et la gestion des déchets,
- la *classe C2* (à majorité de villes méditerranéennes) est caractérisée par la rareté des espaces verts,
- la *classe C3* est surtout marquée par la faiblesse (ou l'absence) de tri sélectif des déchets,
- la *classe C4* a une politique de transports problématique (faible investissement récent pour les transports en commun et les pistes cyclables),
- la *classe 5* est surtout celle des villes à importante surface en espaces verts,
- la *classe 6* est celle de 4 villes mal caractérisées par le 1er plan factoriel (somme des qualités de représentation < 0.07 sur F1 et F2).

On peut tenter une cartographie de cette typologie (la zone de montagne figure sur le fond de carte pour souligner la fragilité des zones de montagne face aux agressions environnementales.

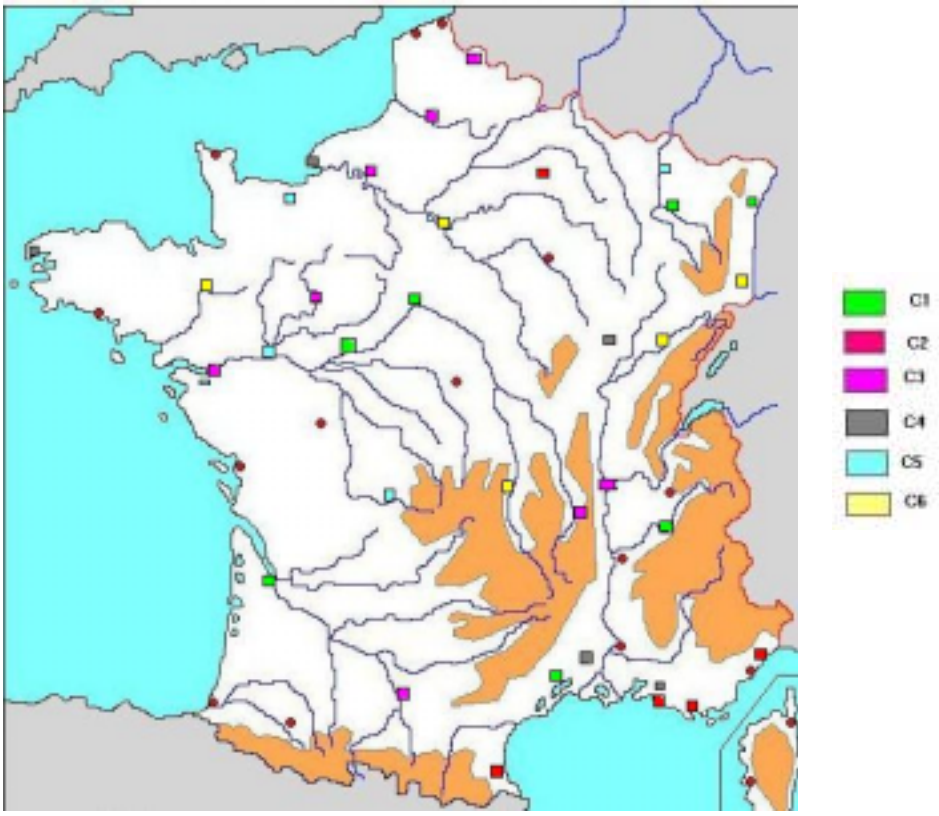


Figure 4.12 : Typologie des villes issue de l'AFCM

- *commentaire*

& Les villes du groupe C3, souvent des villes industrielles et souvent de grandes villes, sont caractérisées par des modes de déplacement non automobile très moyens,

& les villes du groupe C2, presque exclusivement de la façade méditerranéenne, ont de mauvais scores pour pistes cyclables et voies piétonnières (tous aménagements sans doute rendus difficiles par la structure même, très concentrée, de ces villes),

& Les villes de type C4 ont connu de faibles investissements pour les transports en commun,

& Les villes de type C5 ont en commun de grandes surfaces d'espaces verts,

& Les villes de type C1 ont débuté, apparemment, une politique environnementale pour le 21^{ième} siècle : transports en commun, pistes cyclables, tri sélectif,…

Chapitre 5

Méthodes de classification

A) CONNAISSANCES DE BASE

Les analyses factorielles résument, dans les grands tableaux numériques, l'information en colonnes. Les méthodes de classification, elles, ont pour but de résumer celle de leurs lignes. Combiner analyse factorielle et classification permet un résumé complet.

1. Utilité en géographie

Les nombreuses méthodes de classification ont généralement pour but de créer des *typologies*, c'est à dire un ensemble de classes les plus homogènes possible et les plus différentes possible les unes des autres par leurs caractéristiques sur un ensemble de variables. Ce faisant, on s'appuie sur une mesure de *ressemblance* (multivariée) entre unités statistiques. Cette forme de classification a bien sûr tout son intérêt en géographie puisqu'elle constitue les légendes de cartes thématiques combinant des variables. Si leur *autocorrélation* est forte (c'est à dire si les unités spatiales voisines se ressemblent beaucoup), on peut forcer les algorithmes à ne créer que des classes homogènes et spatialement continues, aboutissant ainsi à des *zonations*.

Mais, toutes les zonations géographiques ne s'appuient pas sur un principe de ressemblance. Il est aussi fort utile de créer des régions *fonctionnelles* où les unités spatiales sont interdépendantes. Les mêmes méthodes de classification que dans le cas précédent peuvent être employées, à la condition qu'on puisse mesurer (ou qualifier) différentes formes d'interdépendance fonctionnelle (par exemple par le truchement d'indicateurs économiques, de migrations alternantes, etc.).

Toutes les méthodes de classification ne sont pas statistiques au sens étroit du terme : il existe des méthodes graphiques, des méthodes fondées sur la théorie des graphes ou sur l' »intelligence artificielle » (réseaux neuronaux, par exemple).

2. méthodes graphiques de classification

2.1 graphique cartésien

C'est le graphique d'un nuage de points par rapport à deux axes orthogonaux (Y et X) qui peuvent être les deux premiers axes d'une analyse factorielle (où chaque point est repéré par ses coordonnées en F1 et F2). Si les points du nuage sont peu nombreux, on différencie visuellement des groupes en fonction de leur proximité, créant ainsi des « patatoïdes ». Si les points sont nombreux, on superpose une grille au plan (en 5 ou 9 classes, cf figure 5.1).

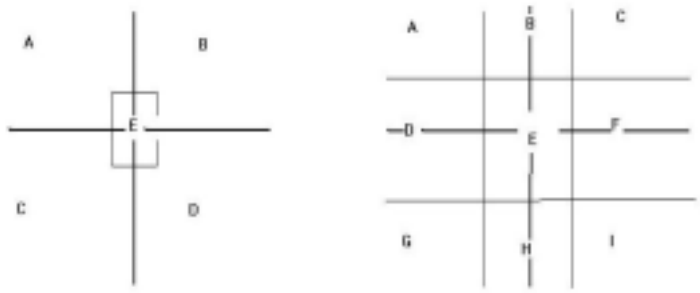


Figure 5.1 : grilles de classification superposées à un plan cartésien

2.2 classification par arborescence « raisonnée »

Cette méthode consiste à créer, par raisonnement, un arbre de tri combinant 2 ou 3 variables ayant chacune peu de modalités. La combinaison peut être graphique et manuelle si le nombre d'unités à trier est petit, informatique sinon (tests emboîtés). Par exemple, un hebdomadaire a publié récemment, pour 34 agglomérations françaises, un classement (cf chapitre 4) pour un certain nombre de critères ayant trait au « développement durable ». Nous en avons extrait 3, relatifs à l'usage des transports en commun, à l'importance relative des pistes cyclables et des voies piétonnières. Pour chacun de ces critères, on a codé 1 les agglomérations de rang inférieur ou égal au rang 17 et 2 les autres. La combinaison de ces 3 critères binaires crée 2^3 classes, identifiées par une lettre de A à H (cf figure 5.2).

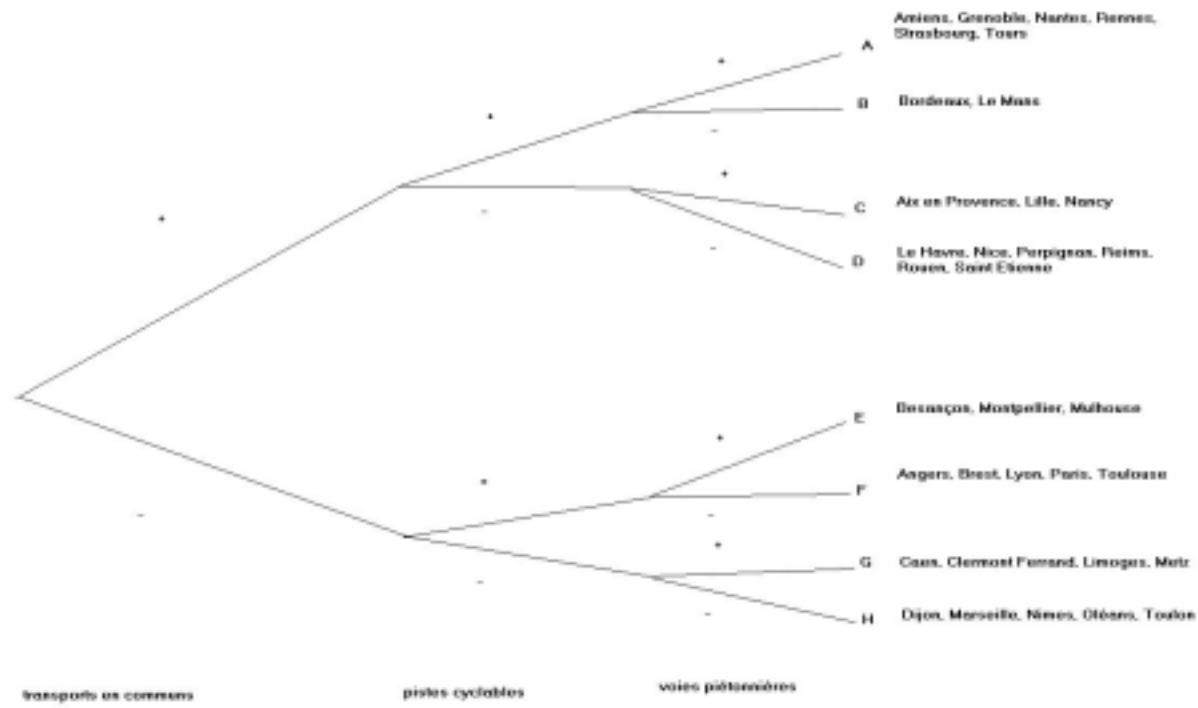


Figure 5.2 : Classification de 34 agglomérations françaises par arbre binaire sur 3 critères

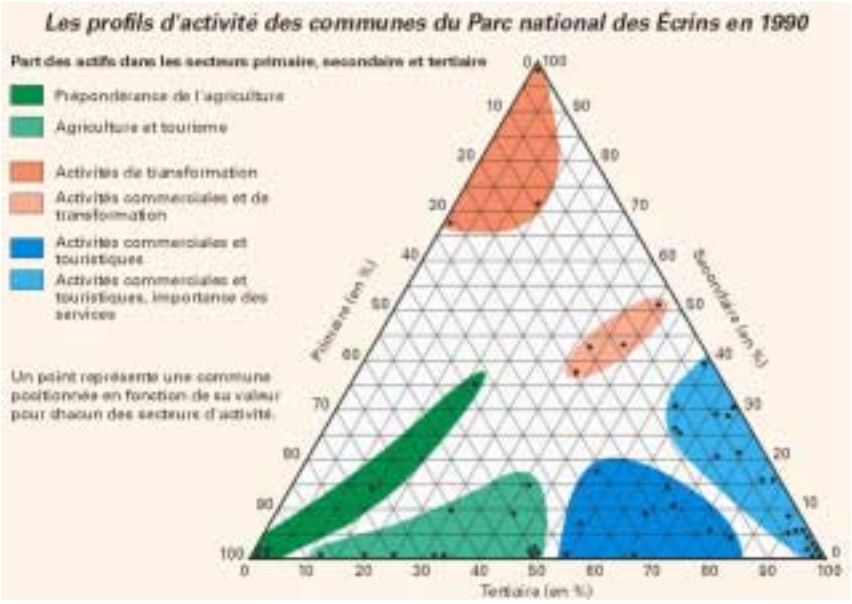
Il est clair qu'avec ce procédé la nature des critères est quelconque mais que le nombre de critères et le nombre de modalités par critère doit être très limité, faute de quoi on aboutit très vite à une « explosion combinatoire » du nombre de classes finales. Par contre, le nombre d'éléments à *trier* est quasiment illimité. Si le manipulateur établit une hiérarchie d'importance entre les critères (ici C1>C2>C3) et entre les modalités de chacun d'eux (ici M1>M2), le résultat final est une *typologie hiérarchique* : dans la figure 5.2, les villes de la

classe A ont un bon équipement général pour tenter de limiter les déplacements en voiture tandis qu'à l'opposé ceux de la classe H en ont un mauvais. S'il n'y a pas de hiérarchie des critères ou des modalités, le procédé crée des classes d'équivalence (différentes seulement).

D'autres exemples d'arbres de tri peuvent être trouvés dans des *nomenclatures* où chaque niveau de l'arborescence représente une subdivision plus détaillée du niveau précédent. Ainsi la nomenclature européenne de l'utilisation du sol (Corine Land Cover) comprend 3 niveaux. Au 1er niveau, on a 5 types d'utilisation du sol (territoires artificialisés, agricoles, forêts et milieux ½ naturels, zones humides, surfaces en eau), le 2ième niveau du type 1 étant subdivisé en 4 sous types (zones urbanisées, zones industrielles ou commerciales, mines, décharges et chantiers, espaces verts non agricoles) et le 3ième niveau du sous type 1.1 en tissu urbain continu ou discontinu. Au total, Corine Land Cover comprend 44 classes d'utilisation du sol.

2.3 Diagramme triangulaire

Pour que la construction d'un diagramme triangulaire soit possible, il faut qu'existent, pour un ensemble d'unités statistiques, un caractère dont les 3 modalités somment à 1 (ou 100%) : il en est ainsi, par exemple, pour la classification de sols en fonction de leur granulométrie (% d'éléments grossiers, moyens, fins) ou pour une typologie d'unités spatiales en fonction de la répartition de leur population active en 3 secteurs (cf Figure 5.3).

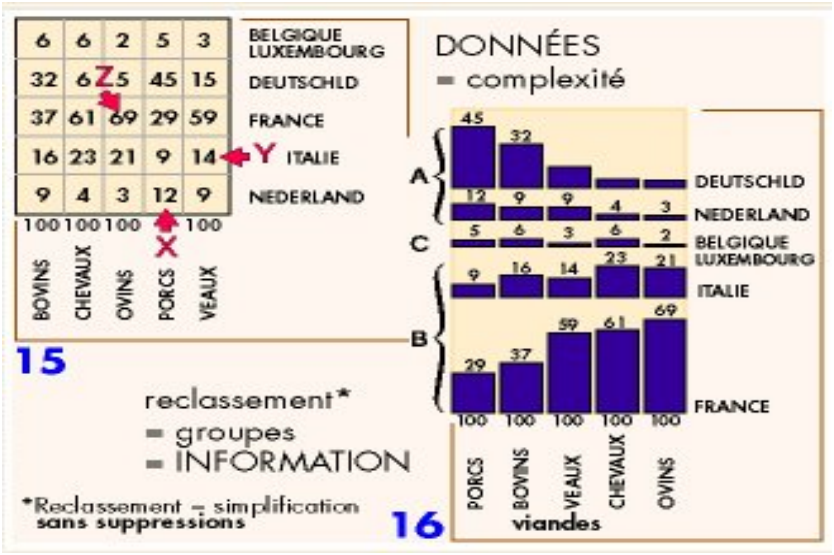


(source : Parc National des Ecrins et INSEE)

Figure 5.3 : Typologie des communes du P.N. des Ecrins en selon les 3 secteurs d'emploi

2.4 Matrice ordonnable de J.Bertin

Une matrice d'information de n unités statistiques en ligne et p variables en colonne peut être représentée visuellement sous forme de matrice de Bertin où chaque case ij porte un niveau de gris ou un rectangle proportionnel à la valeur de la variable j pour l'unité i. La permutation (manuelle ou informatique) des lignes et/ou des colonnes de la matrice permet de créer des groupes visuellement homogènes. La figure 5.4 représente un exemple particulièrement simple où l'on classifie quelques pays en fonction de la composition de leur cheptel.



(source : http://www.sciences-po.fr/)

Figure 5.4 : exemple de matrice ordonnable de Bertin

Le grand avantage des méthodes graphiques de classification est leur côté artisanal : le manipulateur apprend / comprend pendant la réalisation du graphique, surtout si celle ci est opérée manuellement. En somme, les méthodes graphiques présentent un intérêt *didactique* mais aussi nombre de limitations : faiblesse opératoire en termes de nombre d'unités statistiques et/ou de variables, absence (parfois) de critère explicite de rangement, absence (souvent) d'aides à l'interprétation, résultats variables selon les utilisateurs, …

3. Méthodes statistiques de classification

Comme les analyses factorielles, les méthodes classificatoires de la statistique descriptive sous entendent l'existence d'un nuage de n points par rapport à p variables et de métriques (adaptées à la nature des variables) pour mesurer la distance multivariée entre points du nuage. Il existe, en statistique descriptive, un grand nombre de ces méthodes, dont on n'évoque ici que les plus éprouvées.

3.1 Algorithmes de convergence

L'idée générale de tous ces algorithmes est d'itérer un procédé de classification jusqu'à stabilité des résultats lors de deux itérations successives (on parle alors de convergence).

3.1.1 *Agrégation autour de centres mobiles*

& Phase d'initialisation

Le nombre k de classes est fixé a priori et une unité statistique représentative de chacune est fournie par l'utilisateur ou tirée au hasard. Le choix de ces représentants n'est qu'une initialisation de l'algorithme car ils seront remis en cause ultérieurement, par contre le choix du nombre de classes est définitif (si le résultat est à cartographier, ce nombre ne doit pas être trop grand). Le calcul de la distance (euclidienne ou du Khi² selon la nature des variables) permet d'affecter chacune des n-k unités statistiques aux k représentants de groupes.

& Phases ultérieures

Le centre de gravité de chacun des k groupes est recalculé. On détermine alors la distance des n éléments aux k centres de gravité, ce qui permet d'affecter chacun de ces n éléments au groupe dont il est le plus proche (avec la métrique choisie).

Ces calculs sont recommencés jusqu'à ce que deux itérations successives produisent la même partition en classes : le résultat final est alors obtenu.

& <u>Aides à l'interprétation</u>
Tous les logiciels fournissent l'appartenance des n unités statistiques aux k classes (qu'on peut stocker comme une variable qualitative nominale codée de 1 à k) et le rapport variance intra-classes / variance totale, qui permet de juger de la qualité globale de la classification. Beaucoup fournissent la valeur sur les p variables du centre de gravité de chaque groupe, ce qui permet de différencier thématiquement les k classes. Certains logiciels fournissent aussi l'écart type sur les p variables de chaque groupe et, parfois, la somme des p écarts types pour chaque classe, ce qui indique leur homogénéité ou hétérogénéité (à la condition que les p variables soient non corrélées, ce qui est le cas des axes d'une analyse factorielle).

Ces diverses informations permettent non seulement de cartographier la typologie multivariée obtenue mais aussi de superposer à cette carte une représentation de la plus ou moins grande fiabilité (homogénéité) de chaque type.

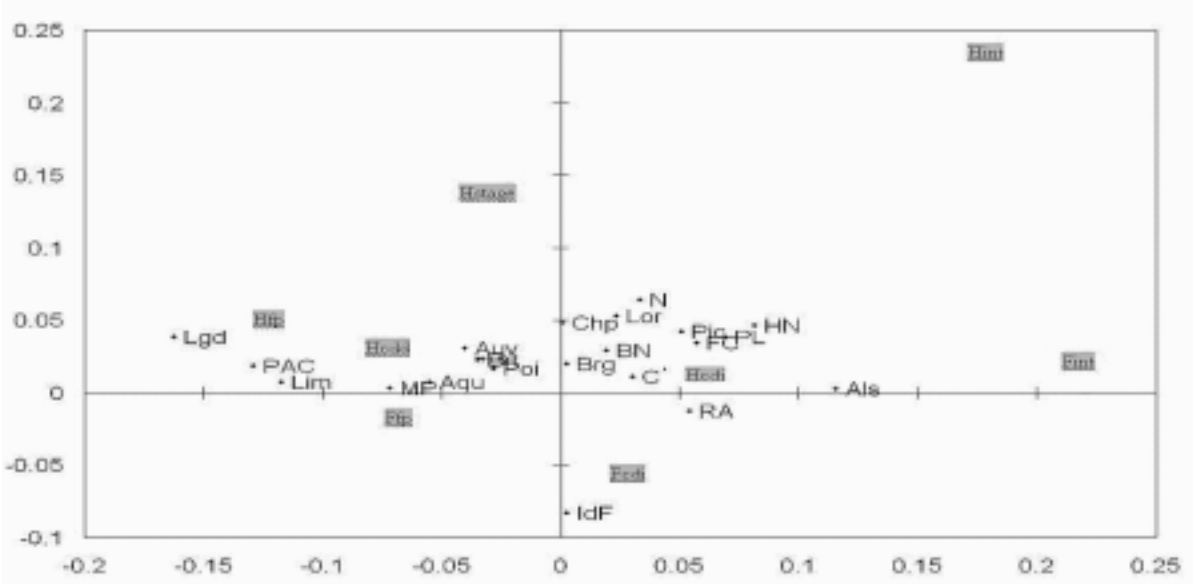
- *<u>Exemple</u>*

De source INSEE, on dispose de données sur le type de contrat de travail de la population active masculine et féminine des régions françaises (hormis la Corse). Les variables retenues sont les suivantes, pour les hommes comme pour les femmes :

H : Hommes **F** : Femmes
cdi : Contrat à Durée Indéterminée **cdd** : Contrat à Durée Déterminée
Int : Intérim **Stage** : stage **f p** : fonction publique

NB : On pourrait déduire de ces indicateurs les % d'emplois précaire et stable (CDI + f.p.) par région : ce n'est pas ici le but. Nous illustrerons les différents procédés de classification par convergence à partir de ce même tableau.

Nous avons effectué sur ce tableau d'effectifs croisant 21 régions et 10 variables une AFC : nous en avons retenu les deux premiers axes (F1 : 52% de variance, F2 : 30%). Le plan des axes 1 et 2 est représenté figure 5.5.



<u>Figure 5.5 *:*</u> 1er plan factoriel de l'AFC sur l'exemple

La figure 5.5 met en évidence :

- la sur-représentation des emplois de la fonction publique et des CDD Hommes en Languedoc, Provence, Limousin, Midi Pyrénées,
- l'opposition avec des régions industrielles où les CDI et l'interim sont relativement abondants comme la Haute Normandie, les Pays de Loire, la Franche Comté, la Picardie, Rhône-Alpes,
- l'Ile de France est, elle, originale par la place que tiennent les CDI féminins.

On procède d'abord à une classification par *agrégation autour de centres mobiles*.
Les variables sont les coordonnées des régions sur les axes 1 et 2 de l'AFC. Le nombre de classes choisi est de 4 et 4 représentants sont tirés au hasard pour initialiser l'algorithme. Le rapport variance inter-classes / variance totale est de 79.6%, ce qui indique une assez bonne séparation en classes, globalement relativement homogènes (variance intra-classes = 20.4%).

Les classes obtenues sont les suivantes :

- *classe 1* : Ile de France et Rhône-Alpes (classe globalement homogène, marquée par l'importance relative des emplois en CDI),
- *classe 2* : Auvergne, Basse Normandie, Bourgogne, Bretagne, Centre, Champagne, Lorraine, Poitou (classe nombreuse, hétérogène et assez moyenne car située au centre du nuage de points),
- *classe 3* : Alsace, Franche Comté, Haute Normandie, Nord, Pays de Loire, Picardie (classe de régions industrielles où sont relativement abondants l'intérim et les CDI masculins),
- *classe 4* : Aquitaine, Languedoc, Limousin, Midi Pyrénées, Provence (classe où sont surreprésentés l'emploi dans la fonction publique et les CDD masculins).

Cette typologie n'est pas optimale (aucune d'ailleurs ne peut l'être !), notamment à cause de l'hétérogénéité de la classe 2. Il faut donc envisager des méthodes mieux discriminantes.

Il existe des variantes de ces algorithmes d'agrégation autour de centres mobiles : par exemple, celle dénommée « *k means* » diffère légèrement de celle décrite ci-dessus en ce qu'elle recalcule le centre de gravité d'un groupe chaque fois qu'un nouvel élément lui est affecté au lieu d'attendre la fin des affectations pour tous les recalculer. Cet ajustement permanent des centres de gravité des classes peut mener à des groupes plus homogènes.

Les méthodes d'agrégation autour de centres mobiles présentent le grand avantage de l'efficacité : on peut, en effet, les employer sur de très grands tableaux avec des temps de calcul raisonnables. Par contre, la qualité de leur résultat final dépend des choix initiaux (nombres de classes, représentant de chacune). D'où l'idée de combiner plusieurs partitions initialisées par des choix différents afin d'obtenir des classes plus homogènes.

3.1.2 *Algorithme des nuées dynamiques*

L'idée de base est d'effectuer plusieurs partitions par agrégation autour de k centres mobiles (représentants choisis ou tirés au hasard) et de les « *intersecter* », ce qui affecte aux classes finales les éléments ayant toujours figuré dans la même classe au cours des diverses partitions effectuées. Naturellement, les classes finales doivent comporter plus d'1 élément (sinon, ce ne

sont pas des classes mais des éléments isolés). L'objectif des nuées dynamiques est donc de rechercher des groupements stables et plus fiables (mais à nombre de classes plus grand ou à éléments isolés).

Pratiquons, sur les mêmes 2 premiers axes factoriels de l'ACP, une seconde classification par agrégation autour de 4 centres mobiles (autres représentants tirés au hasard) ; les classes obtenues (avec une variance expliquée, 80.8%, légèrement meilleur que le précédent) sont :

- *classe 1* : Ile de France seule
- *classe 2* : même composition que la classe 4 précédente,
- *classe 3* : même composition que la classe 3 précédente, moins le Nord et plus Centre et Rhône-Alpes,
- *classe 4* : même composition que la classe 2 précédente, plus le Nord et moins le Centre.

Intersectant les 2 classifications, on obtient donc 4 éléments isolés, d'appartenance instable (Ile de France, Rhône-Alpes, Centre et Nord), et 3 classes stables :
- *groupe A* composé de Alsace, Franche Comté, Haute Normandie, Pays de Loire, Picardie (sur-représentation de l'intérim et des CDI masculins),
- *groupe B* composé de Aquitaine, Languedoc, Limousin, Midi Pyrénées, Provence (sur-représentation de l'emploi dans la fonction publique),
- *groupe C* composé de Auvergne, Basse Normandie, Bourgogne, Bretagne, Champagne, Lorraine, Poitou (groupe plus hétérogène, de caractéristiques moyennes).

L'algorithme des nuées dynamiques présente l'avantage d'aboutir à des classes fiables mais il a aussi, surtout s'il est appliqué à de très grands tableaux, des inconvénients : celui de créer un grand nombre de classes (à effectif inégal) et/ou de nombreux éléments isolés (qu'on peut tenter de rattacher au groupe le plus proche si l'on désire une partition ou laisser comme éléments hors classe).

3.2 Classification Arborescente Hiérarchique (CAH)

Le principe général de l'algorithme est d'agréger progressivement tous les éléments deux à deux en fonction de leur ressemblance multivariée. Au début du processus, les éléments sont des unités statistiques individuelles puis ce sont ensuite des groupes d'éléments. Si le tableau initial comporte p variables et n individus statistiques, après n-1 étapes tous les éléments ne forment plus qu'un seul groupe. L'utilisateur dispose alors d'une *arborescence* des regroupements successifs : c'est sur cette arborescence qu'il choisira le nombre de classes et, donc, leur composition.

A partir de cette architecture générale existent des variantes car l'utilisateur doit choisir :
- une distance multivariée entre individus statistiques,
- un critère d'agrégation des groupes d'individus.

3.2.1 *choix d'une distance entre individus statistiques*

Ce choix est fonction de la nature des variables et des intentions de l'utilisateur :

- *distance du khi²* sur des tableaux de contingence (comme en AFC),

- *distance euclidienne* multivariée pour des variables quantitatives (comme en ACP) ou distance euclidienne au carré (pour « éloigner » les individus à valeurs très différentes),
- *distance rectilinéaire* (« de Manhattan ») pour, au contraire, les « rapprocher »,
- *autre combinaison* de différences partielles.

Les logiciels offrent en général le choix entre distances euclidienne ou du Khi². Ces distances multivariées étant des *additions* de différences sur chacune des p variables, il faut que ces p variables soient *indépendantes* (non corrélées), faute de quoi l'addition n'a guère de sens. C'est pourquoi la CAH a souvent pour variables d'entrée les axes factoriels d'une ACP ou d'une AFC et, alors, la distance entre individus est la distance euclidienne.

3.2.2 *choix d'un critère d'agrégation*

Là encore l'utilisateur devra choisir une des variantes suivantes :

- *plus proche voisin* : la distance (multivariée) entre classes Ci et Cj est la plus petite distance séparant un individu de Ci d'un individu de Cj (les groupes crées avec cette stratégie sont de longues chaînes, peu homogènes),
- *diamètre maximum* : la distance (multivariée) entre Ci et Cj est la plus grande distance séparant un individu de Ci d'un individu de Cj (stratégie fonctionnant bien si le nuage de points est constitué de groupes bien distincts et fonctionnant mal si ces groupes sont « allongés »),
- *distance moyenne* entre tous les individus de Ci et de Cj (méthode efficace si les groupes sont bien distincts dans le nuage de points) ; cette distance moyenne entre groupes peut être pondérée (par leur taille) ou non,
- *distance au centre de gravité* des groupes (pondérée ou non par leur taille), méthode bien adaptée aux cas de groupes de taille très différente,
- *méthode de Ward* : elle cherche à minimiser la variance intra-classes (donc à maximiser la variance inter-classes) pour créer des groupes homogènes les plus différents possibles les uns des autres.

3.2.3 *Exemple*

Considérons le tableau 5.1 qui nous fournit plusieurs informations :

- le régime des précipitations mensuelles moyennes de 12 villes de Guinée (en % de leur total annuel moyen de précipitations),
- le total annuel moyen de précipitations (en millimètres d'eau),
- le rang des 12 villes selon leur total annuel moyen de précipitations,
- la classe à laquelle appartiennent ces 12 villes en fonction de leur régime mensuel (en %) de précipitations.

La méthode de classification utilisée est la CAH avec distance euclidienne et critère d'agrégation de Ward (minimisation de la variance intra-classes).

	janvier	février	mars	avril	mai	juin	juillet	août	sept.	oct.	nov.	dec.	total P.	rang	cls
Mamou	0.1	0.3	1.2	4.1	8.7	11.6	18.6	22.6	18.8	11.2	2.7	0.3	**1791.7**	**7**	**A**
Kissidougou	0.4	1.0	2.4	6.7	10.7	13.3	13.7	17.5	17.0	12.8	4.0	0.8	**1935.0**	**5**	**A**
Macenta	0.5	2.0	3.6	7.5	8.3	10.3	15.1	19.6	16.9	9.9	4.8	1.5	**2736.3**	**2**	**A**
Kankan	0.1	0.1	1.6	4.6	9.1	13.8	17.9	21.8	20.8	8.9	1.2	0.1	**1466.1**	**10**	**B**
Faranah	0.1	0.2	0.8	4.7	9.0	14.1	17.2	19.7	19.6	11.6	2.9	0.0	**1516.4**	**9**	**B**
Labé	0.1	0.2	0.5	2.1	8.5	14.2	19.2	20.7	18.0	8.1	8.1	0.1	**1642.3**	**8**	**B**
N'Zérékoré	0.5	2.4	6.1	8.7	9.5	10.8	12.3	16.2	17.7	10.3	4.2	1.3	**1839.0**	**6**	**B**
Kindia	0.1	0.1	0.6	2.5	9.1	12.5	19.0	24.2	18.1	11.5	2.0	0.4	**1953.1**	**4**	**B**
Koundara	0.0	0.0	0.0	0.2	4.5	12.5	22.8	27.8	23.9	7.5	0.8	0.0	**1098.1**	**12**	**C**
Siguiri	0.0	0.2	0.4	2.6	7.4	13.7	21.4	25.8	19.5	8.2	0.7	0.1	**1238.5**	**11**	**C**
Boké	0.0	0.0	0.0	0.2	4.5	11.0	20.9	26.0	20.6	13.8	2.9	0.1	**2315.4**	**3**	**C**
Conakry	0.0	0.0	0.1	0.5	3.6	10.1	30.1	29.4	16.4	7.8	1.8	0.2	**3775.8**	**1**	**D**

(source : Météo-Guinée)

Tableau 5.1 : Régime moyen de précipitations pour 12 villes de Guinée

La figure 5.6 fournit l'arbre complet des regroupements successifs (jusqu'à une seule classe finale). La longueur d'un lien entre deux nœuds de l'arbre indique l'augmentation de la variance intra-classes lors d'un regroupement : pour obtenir des classes assez homogènes, il convient donc de couper les branches provoquant une forte augmentation de variance. On obtient ainsi 4 classes (A, B, C et D).

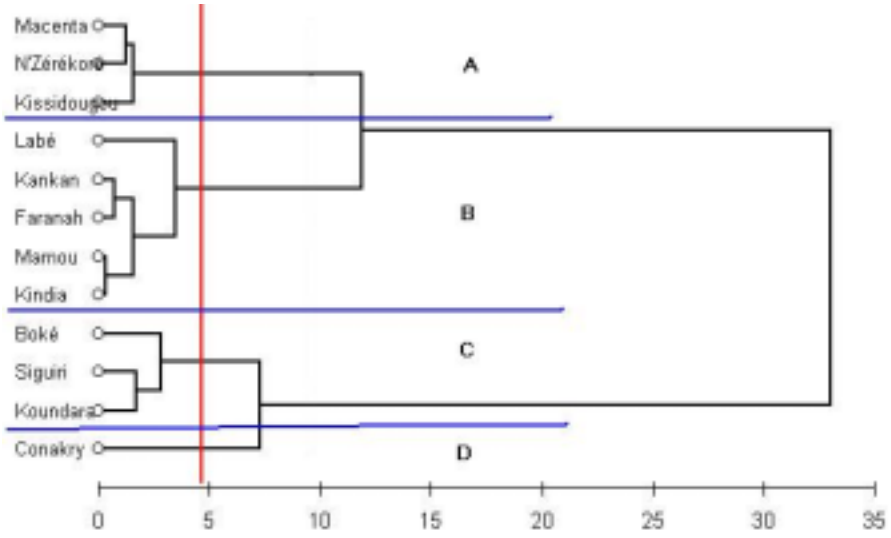


Figure 5.6 : arbre des regroupements successifs (critère d'agrégation de Ward)

Figure 5.7 : carte issue de la CAH (et carte physique explicative)

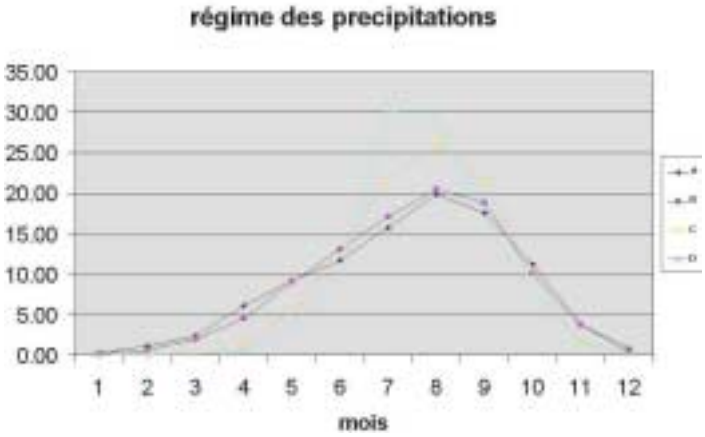


Figure 5.8 : régimes mensuels de précipitations des 4 classes de la CAH

Des figures 5.7 et 5.8, on peut tirer les observations suivantes :

- la classe A, qui correspond pour l'essentiel à la dorsale guinéenne, a des moyennes annuelles de précipitations de l'ordre de 2 m et ses maxima de saison humide (été) sont moins accusés qu'ailleurs,
- la classe B, correspondant aux plaines et massif (Fouta Djalon) intérieurs, a des précipitations annuelles de l'ordre de 1,7 m avec des maxima d'été moins accusés qu'ailleurs (comme la classe A),
- la classe C, correspondant au Nord du pays, plus sec (précipitations de l'ordre de 1,5m) a un régime plus contrasté : saison sèche plus longue et maxima relatifs d'été plus nets,
- la classe D, représentée ici par Conakry seulement, correspond à la zone côtière, fortement arrosée (3,8 m) et aux saisons sèche et humide nettement marquées.

B) EXERCICES CORRIGES

Exercice 1

Le service statistique des Nations Unies calcule chaque année un ensemble d'indicateurs qui, combinés, définissent pour tous les pays du monde un « *Indice de Développement Humain* » plus parlant, en termes de bien être social, que le simple Produit National Brut par habitant. Nous avons extrait du rapport 2004 du Développement Humain de l'ONU 8 indicateurs :

- *esp_vie*, espérance de vie à la naissance exprimée entre 0 et 1 (maximum possible),
- *educ*, combinant taux d'alphabétisation des adultes et % des jeunes scolarisés (exprimée aussi entre 0 et 1),
- *cout_educ*, % du PNB national dédié à l'éducation,
- *egal_sexe*, indice (0-1) mesurant le niveau d'égalité hommes – femmes,
- *activ_F*, % de femmes adultes ayant une activité économique,
- *Pnb/hb*, indice (0-1) de Produit National Brut par habitant,
- *Gini*, indice mesurant l'inégalité de distribution des revenus dans la population de chaque pays (0 : égalité absolue, 100 :inégalité absolue),
- *Chom_lg*, % de la population adulte en chômage de longue durée.

Le tableau 5.2 fournit les valeurs de ces indicateurs pour 25 pays européens : Union Européenne moins Chypre et Malte (données manquantes) plus Suisse et Norvège.

	esp-vie	educ	cout_educ	egal_sexe	activ_F	pnb/hb	Gini	chom_lg
Allemagne	0.88	0.97	4.6	0.916	69	0.91	30.0	4.1
Autriche	0.88	0.96	5.9	0.915	65	0.92	23.1	0.8
Belgique	0.89	0.99	5.8	0.928	65	0.92	25.0	3.4
Danemark	0.85	0.98	8.3	0.92	84	0.93	24.7	0.8
Espagne	0.89	0.97	4.4	0.901	55	0.87	32.5	4.6
Estonie	0.76	0.94	7.4	0.800	82	0.74	37.6	7.9
Finlande	0.87	0.99	6.3	0.923	86	0.91	25.6	2.2
France	0.89	0.97	5.7	0.922	76	0.91	32.7	3.0
Grèce	0.89	0.92	3.8	0.874	57	0.84	32.7	5.0
Hongrie	0.77	0.93	5.1	0.826	72	0.79	24.4	2.6
Irlande	0.86	0.96	4.3	0.908	51	0.93	35.9	1.2
Italie	0.89	0.94	5.0	0.903	58	0.9	27.3	5.3
Lettonie	0.75	0.93	5.9	0.789	81	0.69	32.4	9.8
Lithuanie	0.78	0.93	5.0	0.801	79	0.70	32.4	9.5
Luxembourg	0.87	0.9	4.1	0.907	57	1.00	26.9	0.7
Norvège	0.89	0.98	6.8	0.937	84	0.94	25.8	0.2
Pays Bas	0.88	0.99	5.0	0.926	66	0.92	32.6	0.8
Pologne	0.80	0.94	5.4	0.826	80	0.74	31.6	9.6
Portugal	0.84	0.93	5.8	0.870	70	0.85	35.6	1.8
Royaume Uni	0.87	0.99	4.6	0.920	74	0.90	36.1	1.2
Slovaquie	0.80	0.91	4.1	0.829	84	0.78	19.5	11.1
Slovénie	0.84	0.94	5.0	0.871	80	0.85	28.4	4.1
Suède	0.91	0.99	7.6	0.931	89	0.90	25.0	1.1
Suisse	0.90	0.94	5.6	0.918	66	0.94	33.1	0.6
Tchéquie	0.83	0.89	4.4	0.842	84	0.81	25.4	3.7

(source : ONU, 1999-2002)

Tableau 5.2 : Quelques indicateurs pour le calcul de l'indice de développement humain

Les 8 variables sont quantitatives, exprimées dans des unités de mesure différentes : on les a donc résumées par une ACP centrée-réduite. L'examen de la matrice de corrélation entre variables fait apparaître les principales liaisons suivantes (tableau 5.3).

Fortes corrélations positives	**Fortes corrélations négatives**
Esp_vie – egal_sexe	Esp_vie – chom_lg
Esp_vie – pnb/hb	Egal_sexe – chom_lg
Educ – egal_sexe	Pnb/hb – chom_lg
Pnb/hb – egal_sexe	

Tableau 5.3 : Principales corrélations des 8 variables du tableau 5.2

On a conservé les 3 premiers axes de l'ACP dont les % de variance sont :
- axe 1 : 51%
- axe 2 : 23%
- axe 3 : 13%

Au total, ils résument donc 87% de l'information du tableau 5.2.

Sur les coordonnées des 25 pays sur ces trois premiers axes, nous avons pratiqué une Classification Arborescente Hiérarchique (CAH) selon la méthode de Ward (maximisation de la variance inter-classes).

On fournit sous forme graphique (figures 5.9 et 5.10) les éléments permettant l'interprétation de l'ACP (coordonnées et vecteurs de corrélation sur le plan F1-F2) et de la CAH (dendrogramme des regroupements avant forte augmentation de variance intra – classes).

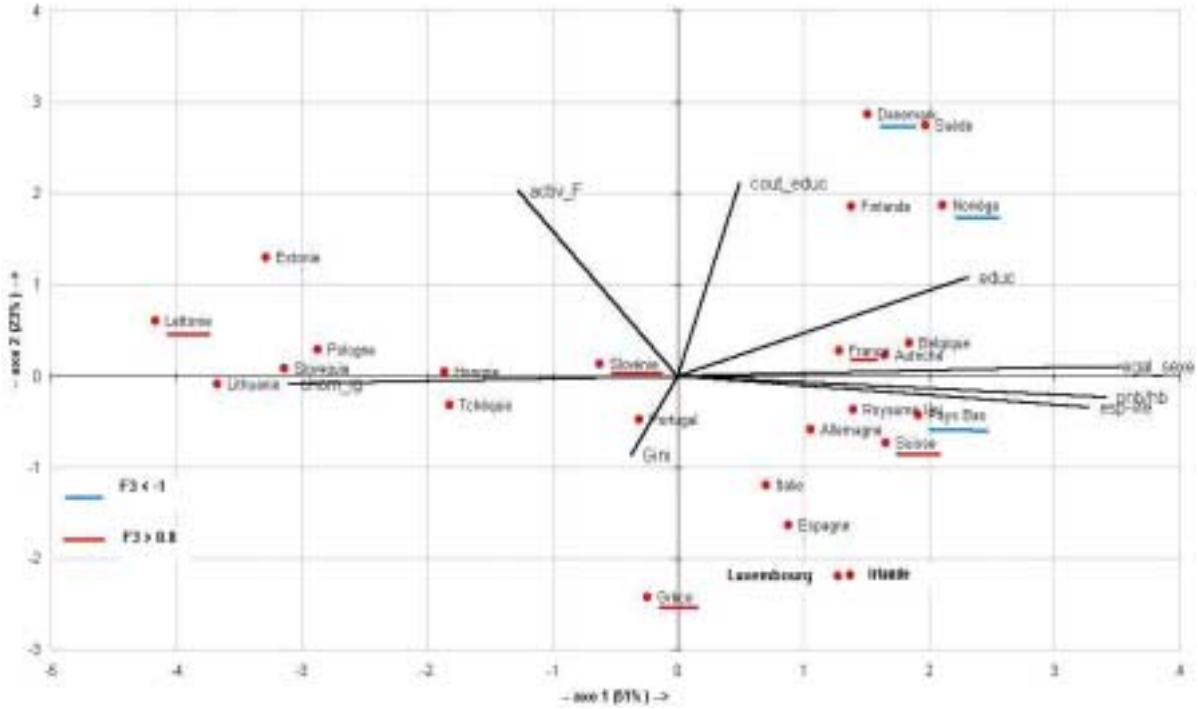


Figure 5.9: Coordonnées et vecteurs de corrélation sur le plan des axes F1 et F2

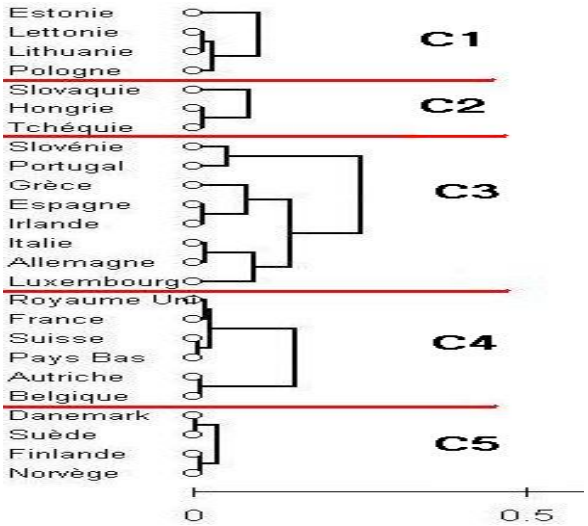


Figure 5.10: Dendrogramme de la CAH opérée sur les 3 axes F1,F2,F3

Questions

1) **Quelles sont les hypothèses sous jacentes à la sélection des variables ?**
2) **Pourquoi avoir fait une CAH sur les axes de l'ACP (et non directement sur les 8 variables) ? Pourquoi une ACP et non une AFC ? Pourquoi sur variables standardisées ?**
3) **Interprétez les axes factoriels à l'aide de la figure 5.9**
4) **Un algorithme de convergence autour de 5 centres mobiles a aussi été utilisé avec les coordonnées des 25 pays sur F1, F2 et F3. En combinant les 2 classifications:**
   Portugal et Allemagne sont passés de la classe 3 à la classe 4
   La Slovénie est passée de la classe 3 à la classe 2
   qu'en induire sur les résultats de la CAH ?
5) **Interprétez les 5 classes de la CAH (crées sur les coordonnées de F1, F2, F3).**

Réponses suggérées

Question 1

Les 8 variables retenues ressortent de 5 thèmes :
& durée moyenne de vie, paramètre démographique essentiel du développement social d'un pays,
& effort pour l'éducation, appréhendé par le % du revenu national qui lui est consacré et ses effets (alphabétisation, % de la population jeune scolarisée) : cet effort est clairement un gage de développement culturel,
& égalité entre les sexes devant le travail et les fruits du travail,
& revenu moyen par habitant, donnant plus ou moins de possibilités d'atteindre des formes de bien être social,
& indices d'inégalité et d'inefficacité économiques vus à travers l'indice de Gini de concentration des revenus et l'exclusion du marché du travail.

Les statisticiens de l'ONU ont combiné ces indices (et d'autres) pour créer par addition une variable numérique, l'indice de développement humain (à valeurs comprises entre 0 et 1), suggérant une hiérarchie unique entre les pays du monde. Notre hypothèse est que, sans nier cette hiérarchie, on peut affirmer aussi qu'il en existe des formes *différentes* et qu'il n'y a pas nécessairement sens à pratiquer des additions entre indicateurs : par exemple, une grande inégalité hommes – femmes est elle compensée par un haut revenu moyen et aboutit on ainsi à un bon niveau de développement humain ?

Question 2

Pratiquer une CAH sur un tableau de données implique que ses variables sont indépendantes (leurs inter - corrélations montrent qu'elles ne le sont pas). En effet, toute CAH commence par le calcul, entre individus statistiques, de distances multivariées qui sont des additions de différences sur chaque variable. Si les variables sont redondantes, le résultat de la typologie est difficilement interprétable. Faire une CAH sur des axes factoriels garantit l'indépendance : additionner des différences partielles a donc un sens.

Ici, l'analyse factorielle pratiquée est une ACP car les variables sont quantitatives ; exprimées dans des unités de mesure différentes, elles ont été centrées – réduites (=standardisées).

Question 3

Aucune variable ni aucun pays n'étant mal représentés par l'ACP, seules les corrélations variables – axes et les coordonnées des pays ont été fournies (figure 5.9).

& *L'axe 1* (la moitié de l'information du tableau 5.2) oppose des pays d'Europe du Nord et alpine (Norvège, Suède, Pays Bas, Belgique, Suisse, Autriche, Danemark) à bonne égalité hommes – femmes, fort Pnb/hb et espérance de vie longue à des pays d'Europe centrale (Pays Baltes, Slovaquie, Pologne, Hongrie, Tchéquie) ayant un fort % de chômeurs de longue durée et de plus faibles valeurs sur espérance de vie, Pnb/hb et égalité des sexes. Cet axe 1 pourrait s'intituler « du modèle scandinave aux effets du protectorat soviétique ».
& *L'axe 2* (environ ¼ de l'information) met en valeur le taux d'activité féminine et l'éducation opposant les pays nordiques (Danemark, Suède, Norvège, Finlande) à des pays

(Grèce, Irlande, Luxembourg, Espagne) où taux d'activité féminine et effort pour l'éducation sont relativement plus faibles.
& *L'axe 3* (13% de l'information) met surtout en évidence l'indice de Gini qui, ici, mesure l'inégalité de distribution des revenus : Lettonie, Slovénie, Grèce, Suisse, France présentent une distribution assez inégalitaire alors que celles du Danemark, de Norvège, des Pays Bas sont bien plus égalitaires.

Ces 3 résumés nous mènent loin d'un indicateur hiérarchique unique : l'axe 1 résume un état de développement en termes de revenu moyen, d'espérance de vie, d'égalité des sexes, de faible taux de chômage longue durée (peu d'exclusion sociale), l'axe 2 insiste plus sur les ferments de développement futur (effort pour l'éducation, taux d'activité féminine) tandis que l'axe 3 met en valeur l'inégalité plus ou moins forte de distribution des revenus (complément indispensable aux Pnb/habitant qui sont des revenus moyens).

Question 4

Globalement, la typologie issue de la CAH est fiable puisque 22 pays sur 25 se retrouvent dans la même classe avec l'algorithme de nuées dynamiques. L'analyse de variance sur les résultats de celui-ci donne 87% de variance inter-classes (13% de variance intra-classes) : les 5 classes de la CAH sont donc globalement homogènes et bien distinctes les unes des autres. Seule la classe 3 est moins homogène puisque Slovénie, Portugal et Allemagne s'en sont séparés. La relative hétérogénéité de cette classe se lit bien, d'ailleurs, sur la figure 5.9. Les 4 autres classes sont, quant à elles, restées de même composition.

Question 5

Pour interpréter la CAH, il faut exprimer l'originalité de chaque classe et ses principales différences aux autres. On a calculé pour chaque classe sa moyenne sur les 8 variables du tableau 5.2, moyennes que l'on a standardisées pour pouvoir les comparer (figure 5.11).

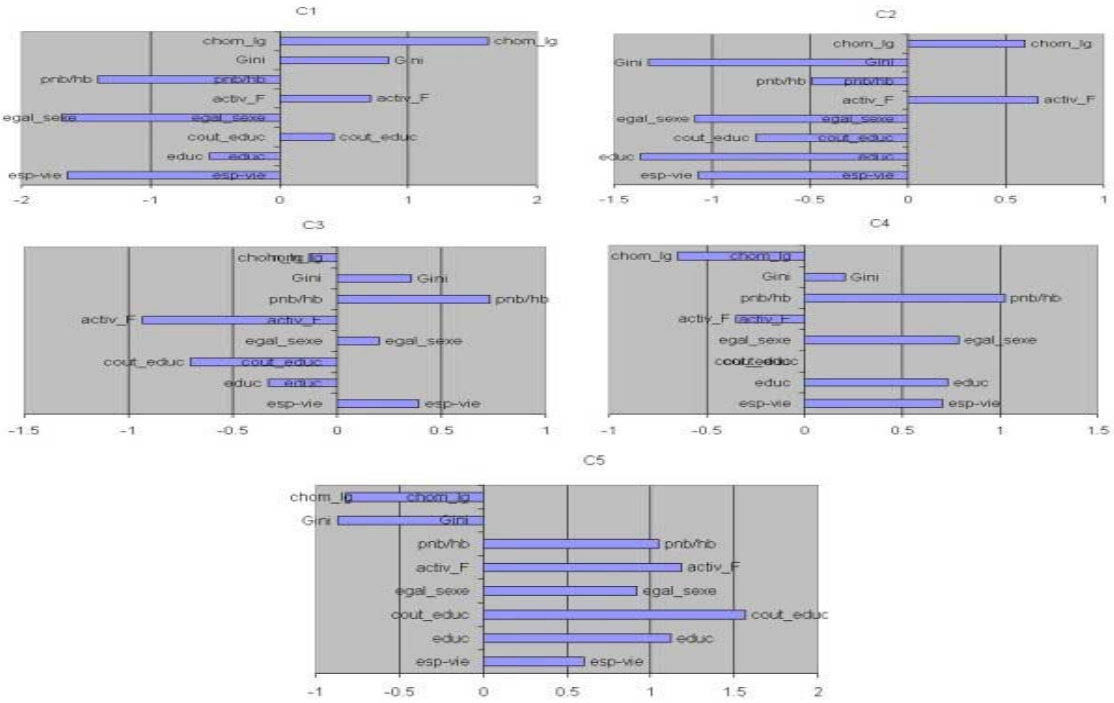


Figure 5.11: Ecarts aux moyennes des 8 variables pour les 5 classes de la CAH

& Classes 1 et 2 sont assez semblables du point de vue de leurs principales caractéristiques :
- faible espérance vie,
- faible égalité des sexes,
- assez grande importance du chômage de longue durée,

mais elles diffèrent par :
- leur Pnb/hb, inférieur en classe 1,
- l'effort pour éducation, scolarisation et alphabétisme, inférieurs en classe 2,
- le taux d'activité féminine, inférieur en classe 2.

& Classes 3 et 4 présentent, de même, un certain nombre de similitudes :
- forts Pnb/hb et espérance vie,
- assez bonne égalité de la distribution des revenus,
- assez faible taux d'activité féminine,

mais diffèrent par :
- taux d'activité féminine plus faible en classe 3,
- effort pour éducation, scolarisation et alphabétisme inférieurs en classe 3.

& La classe 5 a tous les voyants du développement au vert et, principalement, tout ce qui touche à l'éducation, au taux d'activité féminine et au revenu moyen par habitant.

Une carte de cette typologie (figure 5.12) montre que ces classes sont assez bien « zonées ».

Figure 5.12 : Cartographie de la typologie obtenue par CAH

La carte montre bien l'existence de 3 zones principales de développement humain en Europe:
- les pays nordiques,
- l'Europe centrale des nouveaux membres,
- l'Europe des quinze (à laquelle s'agrège la Slovénie).

Exercice 2

Des étudiants ont effectué des relevés de terrain sur le Mont Rachais (1046 m) qui surplombe directement la ville de Grenoble. Le Rachais, retombée méridionale du massif préalpin calcaire de la Chartreuse, était autrefois cultivé (vigne, notamment). Aujourd'hui, des quartiers résidentiels ont envahi le bas des pentes et une forêt de recolonisation leur succède en altitude. L'espèce arborée dominante est le chêne pubescent. Le tableau 5.4 présente 2 facteurs qui lui sont favorables (altitude, ancienne utilisation du sol) pour un échantillon représentatif de 30 placettes en position d'adret entre 215 m et 700 m(source : G.Rovera & C.Corona). La Figure 5.13 localise ces 30 placettes par rapport aux principales courbes de niveau et parcelles cadastrales.

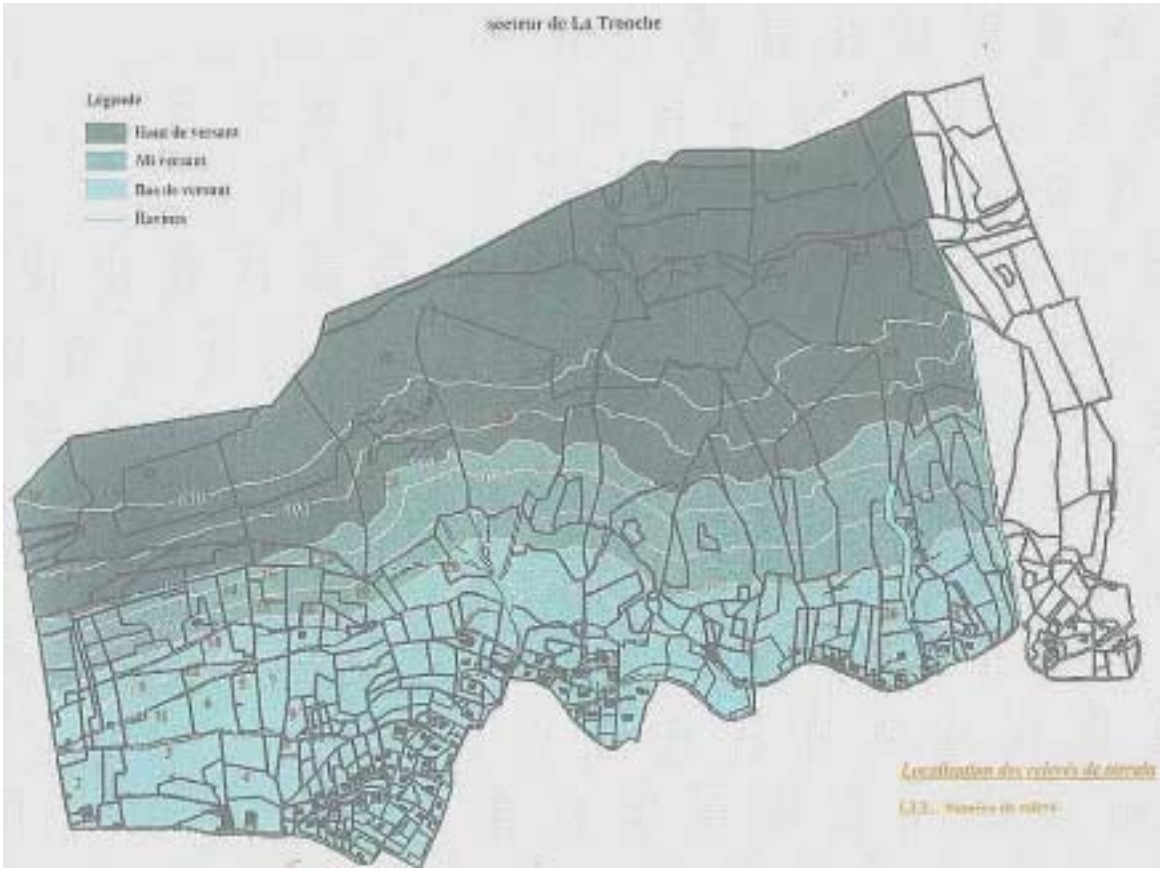


Figure 5.13 : Localisation des 30 parcelles de relevé *(source : G.Rovera & C.Corona)*

Les caractères du tableau 5.4 ont été codés numériquement en 3 modalités :

- *altitude* : **1** : 215-400 m, **2** : 401-550 m **3** : 551-700 m

- *utilisation du sol en 1850* : **1** : vigne-verger **2** : forêt **3** : cultures

- *couverture par chênes pubescents* : **1** : 1-25% **2** : 26-50% **3** : >50%

Les 9 modalités du tableau 5.4 ont été résumés par les 3 premiers facteurs d'une AFCM qui prennent respectivement en compte 41%, 28% et 18% de la variance (donc 87% au total).

N° site	altitude	Sol 1850	chênes
1	3	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	2	3	1
14	2	2	2
15	2	3	1
16	2	3	3
17	3	2	2
18	2	3	3
19	2	3	3
20	2	3	2
21	3	2	3
22	3	1	3
23	3	2	3
24	3	2	2
25	1	3	1
26	1	3	1
27	3	2	3
28	3	2	3
29	3	1	1
30	3	3	3

(source : G.Rovera & C.Corona)

Tableau 5.4 : N° de classes des parcelles enquêtées

Le tableau 5.5 fournit les aides à l'interprétation des 9 modalités et leur effectif. Les valeurs jugées représentatives (coordonnées, contributions à la variance de l'axe, qualités de représentation) sont en caractère gras.

modalité	n	f1	ctr1	qr1	f2	ctr2	qr2	f3	ctr3	qr3
alt. / 3	10.00	-0.83	9.29	0.34	***-0.92***	16.98	0.42	-0.37	4.18	0.07
alt. / 1	13.00	**1.02**	18.35	0.80	-0.10	0.27	0.01	0.11	0.52	0.01
alt. / 2	7.00	-0.71	4.82	0.16	**1.50**	31.75	0.68	0.31	2.13	0.03
1850 / 1	14.00	**0.89**	14.94	0.69	-0.37	3.85	0.12	-0.03	0.04	0.00
1850 / 3	9.00	-0.38	1.72	0.06	**1.35**	32.92	0.78	-0.22	1.38	0.02
1850 / 2	7.00	***-1.30***	15.85	0.51	***-0.99***	13.93	0.30	0.35	2.59	0.04
chên / 1	17.00	**0.81**	15.06	0.86	0.04	0.05	0.00	0.07	0.25	0.01
chên / 2	4.00	-1.25	8.42	0.24	-0.18	0.26	0.00	**2.14**	56.48	0.71
chên / 3	9.00	***-0.98***	11.56	0.41	0.01	0.00	0.00	***-1.08***	32.44	0.50

Tableau 5.5 : Résultats de l'AFCM du tableau 5.4 (modalités)

Le tableau 5.6 fournit les mêmes informations pour les 30 placettes, plus :

- un numéro de classe obtenu par algorithme de convergence,
- un numéro de classe obtenu par classification arborescente hiérarchique sur les 3 premiers axes de l'AFCM.

Site N°	f1	ctr1	qr1	f2	ctr2	qr2	f3	ctr3	qr3	conv.	cah
1	0.32	0.41	0.08	-0.56	1.90	0.24	-0.18	0.31	0.03	1	4
2	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
3	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
4	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
5	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
6	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
7	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
8	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
9	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
10	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
11	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
12	**1.00**	4.05	0.93	-0.19	0.23	0.04	0.09	0.07	0.01	1	4
13	-0.10	0.04	0.00	**1.30**	10.17	0.79	0.09	0.07	0.00	3	3
14	***-1.20***	5.80	0.33	0.15	0.13	0.00	**1.55**	22.33	0.55	4	2
15	-0.10	0.04	0.00	**1.30**	10.17	0.79	0.09	0.07	0.00	3	3
16	-0.76	2.33	0.22	**1.28**	9.95	0.62	-0.55	2.79	0.11	3	3
17	***-1.24***	6.22	0.39	***-0.94***	5.34	0.22	**1.18**	12.78	0.35	4	2
18	-0.76	2.33	0.22	**1.28**	9.95	0.62	-0.55	2.79	0.11	3	3
19	-0.76	2.33	0.22	**1.28**	9.95	0.62	-0.55	2.79	0.11	3	3
20	-0.86	2.99	0.18	**1.20**	8.69	0.35	**1.24**	14.17	0.38	3	2
21	***-1.14***	5.25	0.51	***-0.86***	4.43	0.29	-0.61	3.47	0.15	2	1
22	-0.34	0.46	0.06	-0.57	2.00	0.18	***-0.82***	6.23	0.37	2	1
23	***-1.14***	5.25	0.51	***-0.86***	4.43	0.29	-0.61	3.47	0.15	2	1
24	***-1.24***	6.22	0.39	-0.94	5.34	0.22	**1.18**	12.78	0.35	4	2
25	0.54	1.16	0.20	0.58	2.01	0.23	-0.02	0.00	0.00	1	3
26	0.54	1.16	0.20	0.58	2.01	0.23	-0.02	0.00	0.00	1	3
27	***-1.14***	5.25	0.51	***-0.86***	4.43	0.29	-0.61	3.47	0.15	2	1
28	***-1.14***	5.25	0.51	***-0.86***	4.43	0.29	-0.61	3.47	0.15	2	1
29	0.32	0.41	0.08	-0.56	1.90	0.24	-0.18	0.31	0.03	1	4
30	-0.80	2.60	0.29	0.20	0.23	0.02	***-0.93***	7.97	0.39	2	1

Tableau 5.6 : Résultats de l'AFCM du tableau 5.4 (placettes)

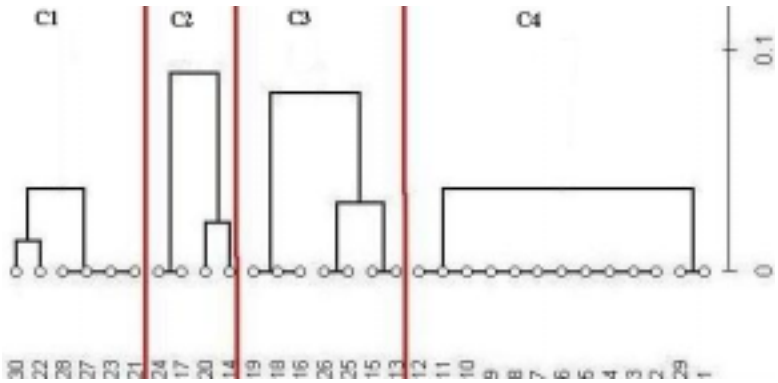


Figure 5.14 : Dendrogramme de la CAH sur les 3 axes factoriels du tableau 5.6

Questions

1) **Pourquoi avoir procédé à une AFCM avant de classifier les 30 sites ?**
2) **Interprétez les 3 axes de cette AFCM**
3) **Interprétez les 4 classes issues de la CAH**
4) **Comparez les résultats de la CAH avec ceux (en 4 classes aussi) issus de l'algorithme de convergence : qu'en déduire ?**
5) **Cartographiez l'interpolation spatiale issue de cette comparaison.**

Réponses suggérées

Question 1

Une analyse factorielle a été pratiquée avant la CAH pour rendre les super-variables (que sont les axes factoriels) indépendantes et pouvoir donc calculer des différences globales entre les 30 placettes échantillon en additionnant des différences partielles.

L'analyse factorielle adaptée est ici l'AFCM car les 3 caractères (tranche d'altitude, utilisation du sol en 1850, taux de couverture par chênes pubescents) sont des variables *catégorielles* c.a.d. des variables à peu de modalités connues individu par individu. Une ACP, adaptée à des variables quantitatives continues, était exclue de même qu'une AFC simple, adaptée à une distribution bivariée (croisement de seulement 2 caractères connus par modalités).

Question 2

L'interprétation des coordonnées des placettes sur les axes factoriels n'est pas ici fondamentale dans la mesure où elles ne sont que représentatives de leur voisinage (et non des entités significatives en elles mêmes). On peut donc se contenter de n'interpréter ces axes qu'en termes de modalités (tableau 5.5).

& *L'axe 1* (41% de la variance) oppose les placettes jadis en vigne-verger, situées à basse altitude (N° 2 à 12), faiblement colonisées par le chêne pubescent à celles fortement couvertes de chênes pubescents et situées au dessus de 550 mètres d'altitude.

& *L'axe 2* (28% de la variance) oppose des parcelles d'altitude moyenne (400-550 mètres) cultivées en 1850 à des parcelles d'altitude plus élevée déjà en forêt en 1850.

& *L'axe 3* (18% de la variance) oppose des parcelles à densité moyenne (25-50%) de chênes pubescents (N° 14, 17, 20, 24) à d'autres (N° 22, 30) où cette densité est supérieure à 50%.

Question 3

Pour interpréter les 4 classes issues de la CAH, il est bon de construire un tableau représentant, pour chacune, la fréquence des 9 modalités (tableau 5.7) et, éventuellement, une représentation graphique (figure 5.15) qui en facilite la lecture.

	alt-1	alt-2	alt-3	1850-1	1850-2	1850-3	chen-1	chen-2	chen-3
c1	0	0	100	16.5	67	16.5	0	0	100
c2	0	50	50	0	75	25	0	100	0
c3	71.5	28.5	0	0	0	100	57	43	0
c4	84.5	15.5	0	100	0	0	100	0	0

Tableau 5.7 : Fréquence des 9 modalités dans chacune des 4 classes de la CAH

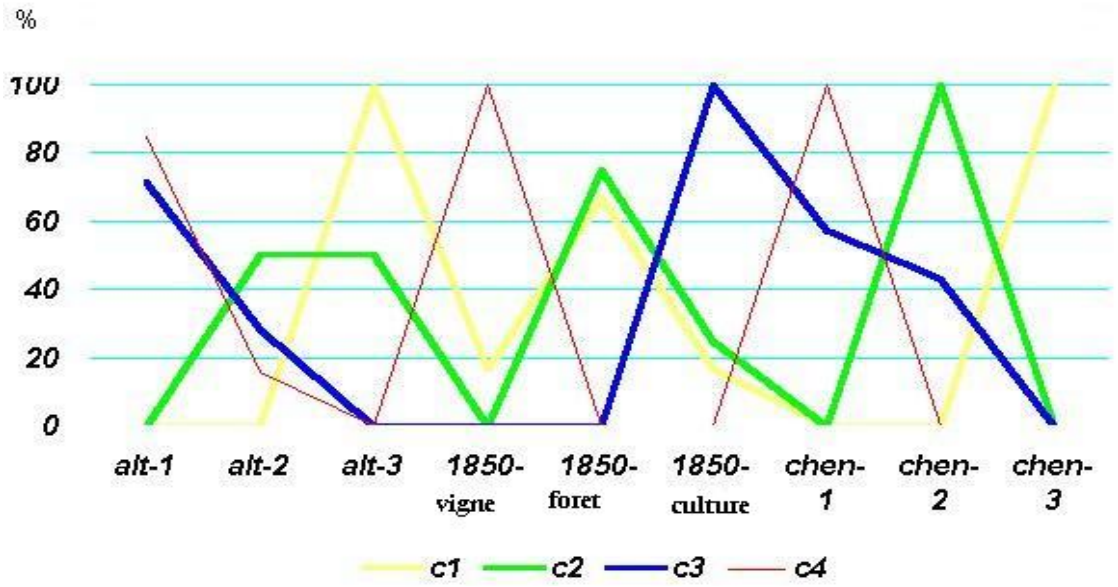


Figure 5.15 : Représentation graphique de ces fréquences

& *La classe C1* est composée de 6 parcelles toutes à plus de 550 mètres et forte densité de chênes pubescents dont l'utilisation du sol en 1850 était au 2/3 de la forêt.

& *La classe C2*, comprenant 4 parcelles, a pour point commun entre elles un taux moyen de recouvrement en chênes pubescents : 3 d'entre elles étaient en forêt en 1850. Cette classe est hétérogène du point de vue des altitudes.

& *La classe C3* est composée de 7 parcelles, la plupart à basse altitude et toutes en cultures en 1850 : elles sont aujourd'hui en voie de colonisation par le chêne pubescent.

& *La classe C4* est composée de 13 placettes (43% de l'effectif total). Elle regroupe des parcelles jadis en vigne-verger, aujourd'hui à faible proportion de chênes pubescents et très majoritairement de faible altitude (donc proches des habitats).

Si l'on focalise l'attention sur la densité de chênes pubescents :

& il est assez faiblement représenté dans les placettes de type C2 ou C4 (d'altitude moyenne ou haute),
& dans celles de type C3 (jadis totalement cultivées), la recolonisation est moyenne,
& dans celles de type C1, recolonisation par le chêne pubescent dense de toutes les parcelles qui n'étaient pas déjà en forêt il y a 150 ans.

Question 4

Comparer la classification issue de la CAH avec celle issue de l'algorithme de convergence permet de voir quelles placettes figurent ensemble dans la même classe avec les deux méthodes : ces classes « croisées » sont fiables par leur composition tandis que celles contenant des parcelles qui ont changé de classe d'une méthode à l'autre le sont beaucoup moins.

Pour croiser les 2 typologies en 4 classes, on construit le tableau 5.8 où les cases contiennent les numéros de parcelles de la typologie croisée.

	CAH-C1	CAH-C2	CAH-C3	CAH-C4	effectifs
Conv-C1			25,26	**1➔12, 29**	15
Conv-C2	**21,22,23,27,28,30**				6
Conv-C3		20	**13,15,16,18,19**		6
Conv-C4		**14,17,24**			3
effectifs	9	4	7	13	30

Tableau 5.8 : croisement des 2 typologies

Le croisement des 2 typologies est globalement révélatrice d'une bonne homogénéité des classes dans la mesure où 10% seulement (3/30) des placettes ont changé d'affectation d'une typologie à l'autre : les placettes numéro 20, 25 et 26 (ces 2 dernières étant isolées dans l'angle sud est). La question qui se pose maintenant est de savoir si les placettes classées dans le même type sont géographiquement proches les unes des autres (formant des zones homogènes) ou dispersées sans ordre spatial visible.

Question 5

Il existe des méthodes d'interpolation spatiale mais elles sont relativement complexes : on procédera donc à une interpolation graphique « à main levée » sur la figure 5.13.

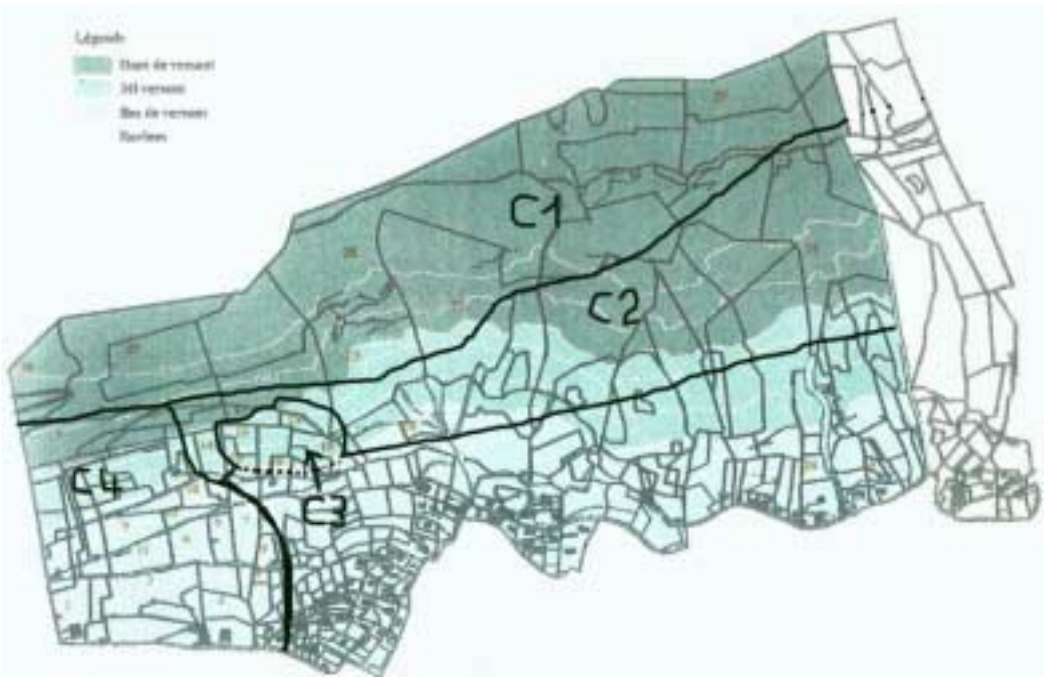


Figure 5.16 : Projection de la typologie par CAH sur le croquis de relevés de terrain

Globalement, il y a une forte autocorrélation spatiale sur la figure 5.16 : les placettes appartenant à la même classe sont géographiquement voisines, ce qui autorise un essai de hiérarchie et de combinaison des facteurs expliquant le taux de recouvrement en chênes pubescents.

Le recouvrement maximum est celui de la zone C1 correspondant à la plus haute tranche d'altitude, la plus éloignée de l'habitat urbain et à fort taux de boisement ancien, le recouvrement minimum est celui de la zone C4, à la fois d'altitude assez basse, la plus proche de l'habitat urbain dense (situé au sud est) et jadis entièrement couverte de vignes et vergers. La zone C2, d'altitude intermédiaire, a un taux de recouvrement relativement important sur d'anciennes parcelles en forêt ou en culture. La zone C3, d'altitude comparable à C4, est plus

éloignée qu'elle du bâti très dense et a de plus forts taux de recouvrement par le chêne pubescent.

A la consultation de la figure 5.16, les facteurs majeurs de recolonisation d'un adret calcaire jadis boisé en haut et cultivé en bas sont l'altitude (présente dans les données) et la proximité du bâti urbain dense (non présent dans les données). L'ancienne utilisation du sol explique aussi, surtout en altitude, la progression du chêne pubescent, espèce de type subméditerranéen (présente ici en position d'adret très protégé sur des sols minces issus du sous sol calcaire).

Référence citée dans ce chapitre

CORONA Christophe, 1998.- L'évolution paysagère post-viticole des bas-versants adrets de la Chartreuse méridionale depuis le XIXe siècle : le cas des communes de La Tronche et de Crolles. (Mémoire de maîtrise sous la direction de G.Rovera, IGA-UJF, 180 p).

Chapitre 6

Régression multiple

A) CONNAISSANCES DE BASE

La géographie offre très peu d'exemples d'un phénomène en « expliquant » directement un autre (cas qui correspond à une régression simple). Science thématiquement combinatoire sur une base spatiale, elle offre beaucoup plus d'exemples où une répartition géographique est « explicable » par la conjonction de plusieurs facteurs : il faut, par conséquent, passer d'un modèle de régression simple à un modèle de régression multiple, où plusieurs variables « explicatives » notées $X_1,\ldots,X_p$ rendent compte de la variabilité de Y, variable « à expliquer » (Y et les X_j étant des variables *quantitatives continues connues par individu*).

1. Le modèle de la régression multiple

C'est une extension du modèle de régression simple, à une différence près : alors que la régression simple est symétrique (on peut permuter les rôles de Y et X, tour à tour variables à expliquer et explicative), la régression multiple est, elle, dissymétrique : c'est bien la distribution de Y qu'il s'agit d'expliquer par celles des X_j.

1.1 Extension du modèle de régression simple à plusieurs variables explicatives

La régression simple, rendant compte de la variabilité de Y en fonction de celle de X, procède à l'ajustement dans un nuage de points bi-dimensionnel d'une droite d'équation **Y'=aX+b** où **Y'** est la valeur ajustée de Y, **a** le coefficient directeur (pente) de la droite et **b** son ordonnée à l'origine (valeur de Y quand X=0).

La régression multiple consiste à projeter les points d'un nuage multidimensionnel sur un *hyperplan* (généralisation d'un plan à plus de 2 dimensions). Comme en régression simple, l'ajustement des projections est réalisé par *moindres carrés*, tels que soit minimale la somme des carrés des projections des Y_i sur l'hyperplan (parallèlement à l'axe des Y):

$$\sum(Y-Y')^2=\min$$

1.1.1 *Coefficients de l'équation de régression*

L'équation de régression multiple (avec p variables explicatives) est :

$$Y'=a_1X_1+\ldots+a_pX_p+b$$

où les a_j sont les coefficients de régression et **b** est l'ordonnée à l'origine (valeur de Y quand tous les $X_j=0$).

1.1.2 *Coefficients de régression standardisés*

b disparaît si Y et les X_j sont *standardisés*, puisque la standardisation (centrage et réduction des variables) opère un changement d'origine (origine nouvelle=0,…,0) et d'échelle (nouvelle unité=1,…,1).
L'équation de régression devient alors : $\mathbf{Y'=\alpha_1 Z_1+\ldots+\alpha_p Z_p}$
où Y et les Z_j sont les variables standardisées (centrées – réduites) et les α_j sont les coefficients de régression standardisés, comparables entre eux car débarrassés des effets de différences de moyenne, d'écart type et d'unité de mesure.

On peut obtenir les coefficients de régression α_j sans passer par la standardisation des variables grâce à la relation : $\mathbf{\alpha_j = a_j (\sigma_{Xj} / \sigma_Y)}$

1.1.3 *Indépendance des variables explicatives*

Pour qu'on puisse additionner les effets des variables explicatives et, donc, connaître la part d'explication de Y par chacune des variables explicatives X_j, il faut qu'elles soient *indépendantes* les unes des autres ;

Ce qui est souhaitable, c'est donc que :

- les variables explicatives X_j soient *très peu corrélées* entre elles,
- les variables explicatives X_j soient *bien corrélées* avec Y, la variable à expliquer.

Ce sont des conditions à vérifier avant de poursuivre.

& *Si (et seulement si) l'indépendance des X_j est vérifiée, alors les coefficients* :

- $\mathbf{a_j}$ s'interprètent comme en régression simple (quand X_j augmente de 1, Y augmente de a_j)
- $\mathbf{\alpha_j=a_j*(\sigma_{Xj}/\sigma_Y)}$ indiquent la part de variance de Y due à chacun des X_j.

& *Si l'indépendance des X_j n'est pas vérifiée* , il faudra se débarrasser de l'effet de leurs redondances.

1.1.4 *Résidus de la régression*

Les *résidus* de la régression ($\mathbf{E_i=Y_i-Y'_i}$) doivent être considérés comme en régression simple et, comme en régression simple, il y a intérêt à étudier leur distribution (histogramme de Y-Y') et à les cartographier, par exemple avec une légende en 3 classes :
- (Y_i-Y_i') très inférieur à 0, le modèle sous estime la valeur Y_i observée,
- (Y_i-Y_i') voisin de 0, le modèle estime bien la valeur Y_i observée,
- (Y_i-Y_i') très supérieur à 0, le modèle sur estime la valeur Y_i observée.

Les résidus, s'ils sont assez importants, peuvent traduire :
- la nécessité d'ajouter une variable explicative oubliée,
- l'existence d'individus « hors norme », situés loin de l'hyperplan,
- des particularités locales,
- l'effet d'une erreur aléatoire (d'échantillonnage ou sur les mesures).

1.1.5 *Conditions de validité d'une régression multiple*

- la relation entre chaque variable explicative X_j et la variable à expliquer Y doit être *linéaire* ; si ce n'est pas le cas, il faut pratiquer une transformation des variables en relation non linéaire avec Y (carrés, log, …) ou utiliser d'autres techniques (réseaux « neuronaux », par exemple),
- il ne doit pas y avoir de variables *colinéaires*, c'est à dire de variables dont la somme des valeurs est égale à une constante ; par exemple, dans une régression entre revenu moyen par habitant en Y et pourcentages d'emploi dans les 3 secteurs primaire, secondaire et tertiaire, l'une de ces 3 variables explicatives doit être enlevée (car son % se déduit de 100% moins la somme des 2 autres) et les résultats n'en seront pas changés,
- les variables explicatives doivent être *indépendantes* (avoir de très faibles corrélations entre elles) ; dans le cas contraire, il peut aussi être fait appel aux réseaux neuronaux
- il est par contre souhaitable que chacune ait une bonne corrélation avec Y

En cas d'*erreur aléatoire*, d'échantillonnage ou de mesure, sur Y (mais pas sur les X_j, considérés comme dénués d'erreur aléatoire), on pourra procéder à des tests supposant, comme en régression simple,
- la *normalité* des résidus Y-Y'
- leur *homoscédasticité* (variance à peu près égale quelque soit l'intervalle de valeurs de Y').

1.2 Exemple élémentaire

Il est fourni par le tableau 6.1, « expliquant » Y, la température moyenne annuelle de 6 villes du fossé rhénan et de ses abords lorrains par leurs latitude et longitude (X_1 et X_2). L'exemple n'a d'autre utilité que calculatoire.

	Lat.	Long.	Temp	Temp'	résidu
strasbourg	48.55	7.6	9.6	10.200	-0.600
bale	47.60	7.5	10.6	10.574	0.026
fribourg	48.00	7.8	11.3	10.576	0.724
nancy	48.70	6.2	9.5	9.269	0.231
belfort	47.63	6.8	9.5	10.129	-0.629
luxeuil	47.78	6.3	10.0	9.752	0.248

Tableau 6.1 : quelques variables explicatives des températures moyennes annuelles

Le tableau 6.2 fournit les coefficients de détermination entre les variables.

	Temp	Long
Lat	0.139	**0.016**
Long	0.389	

Tableau 6.2 : r^2 entre variables du tableau 6.1

Les deux variables explicatives, latitude et longitude des 6 stations, sont indépendantes (coefficient de détermination r^2 très voisin de 0) ; en revanche, elles sont assez faiblement corrélées avec les températures moyennes annuelles observées, variable Y à expliquer.

L'équation de régression est : **Temp'=-0.4588Lat+0.6160Long+27.79**

Quand la latitude augmente de 1 (vers le nord), la température *baisse* en moyenne de 0.46°
Quand la longitude augmente de 1(vers l'est), la température *augmente* en moyenne de 0.62° au dessus d'une valeur de 27.79°.

Les coefficients de régression standardisés sont :
- $\alpha_1=a_1(\sigma_{X1}/\sigma_Y)$ = -0.4588 (0.47466/0.73052) = **-0.298**
- $\alpha_2=a_2(\sigma_{X2}/\sigma_Y)$ = 0.616 (0.69474/0.73052) = **0.586**

Un coefficient de régression standardisé exprime l'augmentation moyenne de Y quand une variable explicative augmente de un écart type et que les autres variables explicatives sont *maintenues constantes*. Ici, les coefficients de régression standardisés indiquent, pour les 6 villes considérées, l'influence sur leurs températures moyennes annuelles :
- de la latitude à longitude constante,
- de la longitude à latitude constante.

Les deux dernières colonnes du tableau 6.1 indiquent les températures prédites par le modèle de régression linéaire multiple (Temp') et les résidus de la régression (différence entre températures réelles et prédites par l'équation de régression).

Par exemple pour Strasbourg, Temp'=(-0.4588*48.55)+(0.616*7.6)+27.79=10.197≈10.2
Et le résidu est de 9.6-10.2=-0.6° : le modèle surestime donc la température de Strasbourg.

Il est clair que l'on a ici un exercice d'école et que l'étude thermique du fossé rhénan et de ses abords nécessiterait bien d'autres stations et variables (altitude, par exemple). Le but, ici, n'est que d'illustrer les principales aides à l'explicitation des résultats.

1.3 En résumé

1.3.1 *corrélations simples entre variables*

Elles doivent être minimales entre les X_j, variables explicatives (indépendance) et bonnes entre variables explicatives X_j et variable à expliquer Y.

La contrainte d'indépendance entre variables explicatives (latitude et longitude des 6 villes) est ici respectée puisque leur coefficient de détermination r^2 (variance commune) est de **0.02**.

Le coefficient de détermination r^2 (cf tableau 6.2) entre :
- Températures et latitude est de **0.14** (r=-0.37) : les températures moyennes tendent légèrement à être + chaudes au Nord, où les villes sont d'altitude plus basse)
- Températures et longitude est de **0.39** (r=0.62) : les températures moyennes tendent à être plus chaudes à l'Est, situé dans le fossé rhénan).

1.3.2 *équation de régression*

L'ajustement par moindres carrés nous fournit une équation de régression multiple. L'interprétation des coefficients est la même qu'en régression simple ; si les variables explicatives sont indépendantes (très peu corrélées), les coefficients de régression standardisés indiquent l'importance de chacune dans « l'explication » de Y.

1.3.3 *résidus de la régression*

On vérifie sur le tableau 6.1 que la moyenne des résidus est nulle (aux arrondis de calcul près). L'importance des écarts Y-Y' est un premier indicateur de la qualité de l'ajustement. Il faut donc regarder de près, cartographier et interpréter les résidus les plus forts (<0 et >0).

Les résidus du tableau 6.1 (dernière colonne) semblent forts, notamment 3 d'entre eux :

- La température moyenne annuelle est nettement surestimée par l'équation de régression à Belfort (altitude plus élevée : 422 mètres) et à Strasbourg,
- Elle est nettement sous estimée à Fribourg.

Reste à trouver l'explication à ces forts résidus (altitude, topographie locale, échantillon de villes non significatif,…).

1.4 Tests sur données d'échantillon

Si (et seulement si) les données proviennent d'un échantillon *représentatif* dont on veut généraliser les résultats à toute la population mère (toute la zone dans l'exemple), on procédera à des tests de significativité des résultats de la régression (dont les éléments sont fournis par la plupart des logiciels statistiques).

1.4.1 *Résidus comme erreur aléatoire du modèle de régression ?*

Comme en régression simple, la distribution des résidus ($E_i=Y_i-Y_i'$), exprimés dans l'unité de mesure de Y (en degrés Celsius dans l'exemple), doit alors donner lieu à examen :

- la distribution des E_i doit être normale (gaussienne),
- le nuage de points E (en ordonnées) – Y' (en abcisses) ne doit pas montrer de nettes croissance ou décroissance des valeurs de E en fonction de celles de Y',

Si la distribution de l'erreur aléatoire est gaussienne, on peut donc utiliser la distribution de probabilités de la loi de Gauss pour extrapoler les résultats.

1.4.2 *Significativité de l'ensemble des variables explicatives*

On effectue une analyse de variance et un test F de Fisher Snedecor car la Somme des Carrés des Ecarts (SCE) totale de Y = SCE expliquée par la régression + SCE des résidus Y-Y'.

$$SCE_Y = \sum_{i=1}^{n}(Y_i - \bar{y})^2 \quad , \quad SCE_{residu} = \sum_{i=1}^{n}(Y_i - Y_i')^2 \quad \text{et} \quad SCE_{regr} = SCE_Y - SCE_{residu}$$

Carré Moyen dû à la régression : $CM_{regr}=SCE_{regr}/p$ p variables explicatives
Carré Moyen des résidus : $CM_{residu}=SCE_{residu}/(n-p-1)$ n individus

F calculé = $CM_{regr}/CM_{résidu}$

On lit dans la table du F de Fisher Snedecor (pour un risque d'erreur choisi) la valeur de F correspondant à p et n-p-1 degrés de liberté. Si F calculé > F lu, on accepte (au risque d'erreur choisi) l'hypothèse que la régression est généralisable à la population mère (toute la zone rhénane – vosgienne ici). Le tableau 6.3 fournit les valeurs pour cette analyse de variance.

SCE due à	SCE	Degrés de Liberté	Carré Moyen
régression	1.2721	P = 2	0.6360
Résidus	1.3962	n-p-1 = 3	0.4654
totale	2.6683	n-1 = 5	

Tableau 6.3 : analyse de variance relative à la régression du tableau 6.1

F calculé vaut ici 0.6360/0.4654=**1.37**

Au risque d'erreur de 5%, F lu dans la table pour 2 et 3 degrés de liberté vaut **19.16**
F calculé < F lu : *on ne peut généraliser la régression à toute la zone*.

On vérifie par ailleurs que, sur l'échantillon des 6 villes, l'intensité de la relation est faible:
$I=SCE_{regr}/SCE_{tot}$ = 1.2721/2.6683 =**0.477**
Latitude et longitude n'expliquent, dans l'échantillon, que 47.7% des variations inter-cités de températures annuelles moyennes.

1.4.3 *Significativité de chacune des variables explicatives*

Même si l'ensemble des variables explicatives est significatif, il se peut que certaines d'entre elles ne le soient pas, d'où le test de chaque coefficient $\mathbf{a_j}$ de régression à l'aide d'un t de Student qui vaut **t calculé = a_j/erreur type de a_j** (erreur type d'échantillon de a_j).

Les logiciels fournissent généralement les éléments de ces tests. Le tableau 6.4 les récapitule.

Var. explicative	Coefficient a de régression	Erreur type de a	T calculé	Risque d'erreur
Latitude	-0.4588	0.6481	-0.71	0.530
longitude	0.6160	0.4428	1.39	0.258

Tableau 6.4 : test de Student des coefficients de régression

T calculé pour la latitude = -0.4588/0.6481= **-0.71**
T calculé pour la longitude=0.6160/0.4428= **1.39**

On pourrait comparer avec des valeurs de t lues dans la table de Student mais nous sont ici donnés les risques d'erreur des tests. Ainsi, au vu de ceux ci (53.0% et 25.8%), ni a_1 ni a_2 ne sont significativement différents de 0 dans la population mère : ni la latitude ni la longitude ne jouent un rôle généralisable à toute la zone dans l'explication de ses températures annuelles moyennes (en outre, très petit échantillon, probablement non représentatif).

1.4.4 *Eléments « hors norme » (« outliers »)*

Au risque d'erreur de 5%, ce sont les éléments dont le résidu Y_i-Y_i' standardisé (divisé par l'écart type de Y-Y') est inférieur à –1.96 ou supérieur à +1.96 (inférieur à –2.57 ou supérieur à +2.57 au risque d'erreur de 1%).

Il est évidemment essentiel d'examiner de près d'éventuels outliers, voire de les supprimer.

Dans l'exemple, en standardisant les résidus (les divisant par leur écart type=0.5284), on observe qu'aucun n'est inférieur à –1.96 ou supérieur à 1.96.

2. Corrélations multiple et partielles

2.1 Coefficient de corrélation multiple

C'est le coefficient de corrélation de Bravais Pearson entre Y et Y', c'est à dire entre valeurs observées et prédites par le modèle de régression : il est noté $\mathbf{R_{Y,Xj}}$.

Comme en régression simple, c'est le carré du coefficient de corrélation (**R^2** : coefficient de détermination) qui exprime le *pourcentage de variance pris en compte par le modèle* et qui mesure donc la qualité de l'ajustement linéaire. Si les variables explicatives X_j sont parfaitement indépendantes les unes des autres (aucune redondance entre elles), R^2 multiple est la somme des r^2 simples entre chaque X_j et Y.

Dans l'exemple du tableau 6.1, le coefficient de corrélation multiple (corrélation simple entre les variables Temp et Temp') est de **0.691** et le coefficient de détermination de **0.477.** R^2 mesure la variance expliquée par la régression

$$R^2= I=SCE_{regr}/SCE_{tot} = 1.2721/2.6683 =\mathbf{0.477}$$

Comme l'analyse de variance l'avait déjà révélé, l'équation de régression multiple n'explique que 48% des différences de température moyenne annuelle entre les 6 villes de l'échantillon tandis que 52% de celle ci est inexpliquée (et due à d'autres facteurs). Analyse de variance et R^2 fournissent donc la même information (variation de Y explicable par l'ensemble des X_j).

2.2 Tests sur R et R^2

Si (et seulement si) les données de Y proviennent d'un échantillon, on procédera à des tests de généralisation à la population mère.

2.2.1 *test de significativité de R, exactement comme en régression simple :*

- hypothèse H_0 d'indépendance entre Y, la variable à expliquer, et l'ensemble des variables explicatives X_1, …, X_p
- lecture dans la table du r de Bravais Pearson d'une *valeur plafond* pour un risque d'erreur α choisi et un nombre de degrés de liberté ν**=n-p-1** (p variables explicatives),
- si R calculé <= R lu dans la table, on accepte l'hypothèse d'indépendance,
- si R calculé > R lu dans la table, on rejette l'hypothèse d'indépendance (en pratique, on conclut à une relation significative, au risque d'erreur choisi).

Dans l'exemple du tableau 6.1 :
- R calculé = 0.69,
- R lu dans la table (pour α=0.05 et ν=6-2-1=3) = 0.88
- R calculé est inférieur à R lu.

On ne peut pas rejeter l'hypothèse d'indépendance entre températures moyennes annuelles des 6 villes, leurs latitude et longitude : on conclut en pratique que leur relation n'est pas significative. Il est clair, en outre, que l'échantillon des 6 villes a peu de chances d'être représentatif de la zone rhénane (l'exercice n'est qu'un exercice d'école !).

2.2.2 *test de R^2 par analyse de variance*

Déjà effectué pour tester la significativité de l'ensemble des variables explicatives, l'analyse de variance du &1.4.2 est le test de R^2 (et a donc déjà fourni l'information).

2.3 Coefficients de corrélation partielle

Quand les variables explicatives ne sont *pas indépendantes*, l'explication de la variance de Y par chacun des X_j est partiellement redondante. On procède alors au calcul de coefficients de *corrélation et de détermination partielles* entre chaque X_j et Y.

Quand il n'y a que deux variables explicatives X_1 et X_2, la formule est, pour X_1 par exemple :

$$r_{YX_1/X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{(1-r^2_{YX_2})(1-r^2_{X_1X_2})}$$ où les r et les r^2 sont les coefficients simples.

Le coefficient de détermination partielle est le carré du coefficient de corrélation partielle. Il représente la part de la variance de Y qu'explique une variable explicative X_j quand les autres variables explicatives sont présentes mais *maintenues constantes* par rapport à Y et X_j. En d'autres termes, le coefficient de détermination partielle explique la part de variance de Y non déjà prise en compte par les autres variables explicatives : c'est l'apport spécifique de X_j à l'explication de Y, complémentaire à celui des autres variables explicatives.
Par exemple, si l'on a mesuré, pour n parcelles témoin sur un mois donné, la production végétale Y, la pluviométrie X_1 et le nombre d'heures d'ensoleillement X_2, le coefficient de détermination partielle entre production végétale et ensoleillement indique quelle est la part de variance de Y qu'explique l'ensoleillement *comme si la pluviométrie avait été partout la même* (« à pluviométrie constante » donc).

3. Régression multiple pas à pas

Avec l'accessibilité croissante de grandes bases de données numériques, il devient fréquent d'adopter une démarche exploratoire recherchant, parmi les variables stockées, la « meilleure » combinaison des X_j disponibles pour expliquer Y : cela participe de ce que l'on appelle *fouille de données* (« data mining » en anglais). Dans ce contexte, des stratégies d'ajout progressif au modèle de variables explicatives afin de maximiser R^2 sont devenues d'usage courant sous le nom de *régression pas à pas* (« stepwise regression » en anglais).

La procédure, itérative, est la suivante :
- régression simple entre Y et celle (X_k par exemple) ayant avec Y le plus fort R^2,
- ajout, parmi toutes les combinaisons de variables explicatives, de celle (X_l) conduisant au plus fort R^2 avec Y (équation : $Y' = a_1X_k + a_2X_l + b$),
- itération du procédé jusqu'à une condition d'arrêt des ajouts.

Naturellement, les précautions habituelles pour une régression multiple sont à respecter (notamment l'indépendance des X_j progressivement ajoutées) mais des questions supplémentaires se posent aussi:
- on ne peut itérer le procédé jusqu'à ce que **p**, le nombre de variables explicatives, soit disproportionné par rapport au nombre d'individus **n** (on considère, en général, que n doit être au moins égal à 10 fois p).
- Même en ajoutant des variables « bidon » (composées de nombres tirés au hasard), on a toute chance d'augmenter la valeur de R^2 puisqu'il est rarissime qu'une corrélation soit exactement égale à 0. Il faut donc choisir où arrêter l'ajout de variables.
- Le gain de variance expliquée entre deux itérations successives doit être conséquent. Dans ces cas là, la variable ajoutée doit amener un surplus significatif de R^2 (on considère empiriquement que la variable ajoutée doit être significative avec un risque d'erreur inférieur à 15%).

La procédure ci dessus décrite est *ascendante* (ajout progressif de variables explicatives). Il existe aussi des procédures *descendantes* (retranchant progressivement les variables les moins significatives).

4. Ajout d'une variable catégorielle à une régression multiple

Pour en faire saisir l'intérêt (et les limites), prenons un exemple.

4.1 Exemple

Considérons, pour un ensemble de pays européens en 2003, la régression simple entre Y : espérance de vie à la naissance des hommes (en années) et X: Produit National Brut par habitant (en milliers de dollars). La figure 6.1 fournit le graphique du nuage de points correspondant.

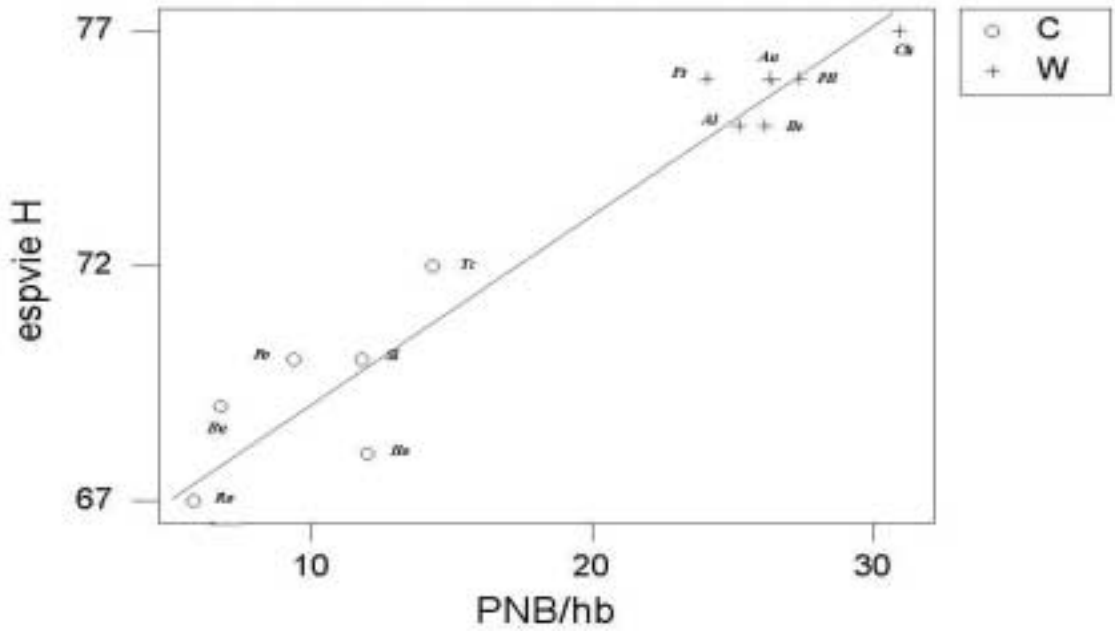


Figure 6.1 : Nuage de points croisant espérance de vie masculine et PNB/habitant

Deux remarques peuvent y être faites :
- La relation est apparemment linéaire (r^2=0.93),
- Cette linéarité tient essentiellement à l'appartenance des pays à l'Europe occidentale (W) ou centrale (C) qui apparaît comme une variable explicative pertinente.

Or, cette variable (Z) est qualitative et ne peut être traitée par régression classique : elle peut l'être par *ANCOVA* (Analyse de covariance) dont l'idée directrice est la suivante :
- calculer la régression entre variables quantitatives continues (ici régression simple entre Y : espérance de vie et X : PNB/hb),
- faire une analyse de variance (ANOVA) entre les résidus Y-Y' de la régression (variable à expliquer) et la variable qualitative Z à k modalités (2 seulement ici).
- Le commentaire des résultats est donc le même qu'en analyse de variance.

Détaillons quelque peu à l'aide des tableaux 6.5 et 6.6.

		Y : espvie H	X : PNB/hb	Z :region
Al	Allemagne	75	25.24	W
Au	Autriche	76	26.38	W
Be	Belgique	75	26.15	W
Fr	France	76	24.08	W
PB	Pays Bas	76	27.39	W
Ch	Suisse	77	30.97	W
Bu	Bulgarie	69	6.74	C
Ho	Hongrie	68	11.99	C
Po	Pologne	70	9.37	C
Ro	Roumanie	67	5.78	C
Sl	Slovaquie	70	11.78	C
Tc	Tchéquie	72	14.32	C

Source : INED 2003

Tableau 6.5 : tableau pour l'ANCOVA

& *Résultats de la régression*

L'équation de régression est : **Esp vie H=0.383 PNB/hb + 65.6**

Pour l'ensemble des 12 pays considérés en 2003, l'espérance de vie masculine augmente, au dessus de la valeur 65.6 ans, d'1 an pour chaque augmentation du PNB/hb de 383 dollars.

Seule la Hongrie est un outlier: son résidu standardisé vaut –2.32 (son espérance de vie est nettement inférieure à ce que laisserait attendre son PNB/habitant).

Coefficients r de corrélation linéaire : 0.97 et r^2 de détermination : 0.93.

& *Résultats de l'analyse de variance résidus – région d'Europe*

	Degrés de liberté	Somme des carrés	Carrés moyens
Modèle	2	135.46	67.73
Résidus	9	9.45	1.05
Total	11	144.92	

Tableau 6.6 : ANOVA entre résidus de la régression et la variable « région »

La variance expliquée par le modèle est de 135.46 / 144.92 = **93.5%**
Il n'y a pas lieu ici de procéder à un test probabiliste puisque Y, issue d'une comptabilité exhaustive et non d'un échantillon, est sans erreur aléatoire. Dans le cas contraire (si Y provenait d'un échantillon représentatif ou était affecté d'erreurs de mesure aléatoires), on procéderait à un test F de Fisher Snedecor.

4.2 Conditions de validité de l'ANCOVA

& *Ce sont celles de la régression et de l'analyse de variance.*

- les variables explicatives X_j (quantitative(s) et qualitative) doivent être sans erreur aléatoire et seule la variable à expliquer Y peut l'être éventuellement (Y est sans erreur aléatoire dans l'exemple),
- les variables explicatives quantitatives doivent être linéairement liées à Y (ce qui est le cas ici, voir Figure 6.1),

- il ne doit pas y avoir de nette différence de pente ou d'ordonnée à l'origine des régressions dans chacun des groupes définis par les modalités de Z (condition respectée dans l'exemple ci-dessus). Si ce n'est pas le cas, il faut faire une régression par modalité de Z.
- Il ne doit pas y avoir dans la régression d'outlier (condition assez mal respectée dans notre exemple puisque la Hongrie a un résidu réduit de –2.32),
- S'il y a plusieurs variables explicatives quantitatives,elles ne doivent pas être co-linéaires (sommant à une constante).

& *S'il y a erreur aléatoire sur Y, pour que le test soit valide :*
- la distribution des résidus doit être (à peu près) gaussienne,
- leur variance doit être (à peu près) constante quelque soit la valeur de Y,
- dans chaque groupe défini par les modalités de Z, la distribution des valeurs de Y-Y' doit être (à peu près) gaussienne et de même variance (condition qui serait mal respectée ici puisque les résidus sont plus importants dans le groupe C que dans le groupe W).

B) EXERCICES CORRIGES

Exercice 1

Le tableau 6.7 présente, pour 33 villes des USA :
- Y: température de janvier, en degrés Celsius (moyenne sur 31 ans, de 1960 à 1990),
- 3 variables explicatives : X_1 latitude, X_2 longitude, X_3 altitude,
- Températures T'jv prédites par la régression multiple, résidus (T-T') et résidus réduits.

	T° jv	Lat.	Long.	Alt.	T'jv	T-T' en °C	T-T' reduits
Montgomery,AL	3.3	32.9	86.8	67.4	4.69	-1.39	-0.45
Phoenix,AZ	1.7	33.6	112.5	363.3	6.01	-4.31	-1.39
LittleRock,AR	-0.6	35.4	92.8	78.6	2.67	-3.27	-1.06
LosAngeles,CA	8.3	34.3	118.7	38.4	7.92	0.38	0.12
SanFrancisco,CA	5.6	38.4	123.0	3.4	4.09	1.51	0.49
Denver,CO	-9.5	40.7	105.3	1793.1	-10.97	1.47	0.47
Washington,DC	-1.1	39.7	77.5	4.9	-4.03	2.93	0.95
KeyWest,FL	18.3	25.0	82.0	1.5	13.31	4.99	1.61
Miami,FL	14.5	26.3	80.7	3.4	11.63	2.87	0.93
Atlanta,GA	2.8	33.9	85.0	312.7	1.94	0.86	0.28
Boise,ID	-5.6	43.7	117.1	871.1	-7.56	1.96	0.63
Chicago,IL	-7.2	42.3	88.0	188.7	-6.46	-0.74	-0.24
Indianapolis,IN	-6.1	39.8	86.9	242.9	-4.09	-2.01	-0.65
DesMoines,IA	-11.7	41.8	93.6	291.7	-5.65	-6.05	-1.96
NewOrleans,LA	7.2	30.8	90.2	1.2	7.93	-0.73	-0.24
Boston,MA	-5.0	42.7	71.4	6.1	-8.33	3.33	1.08
Detroit,MI	-6.1	43.1	83.9	194.8	-8.00	1.90	0.61
Helena,MT	-13.3	47.1	112.4	1180.5	-13.81	0.51	0.16
Concord,NH	-11.7	43.5	71.9	342	-11.02	-0.68	-0.22
Albany,NY	-10.0	42.6	73.7	86.9	-8.33	-1.67	-0.54
NewYork,NY	-2.8	40.8	74.6	4.0	-5.70	2.90	0.94
Raleigh,NC	-0.6	36.4	78.9	133.2	-0.80	0.20	0.06
Bismarck,ND	-17.8	47.1	101.0	511.1	-11.78	-6.02	-1.94
OklahomaCity,OK	-2.2	35.9	97.5	394.7	1.04	-3.24	-1.05
Harrisburg,PA	-4.4	40.9	77.8	94.5	-5.84	1.44	0.47
Philadelphia,PA	-4.4	40.9	75.5	6.4	-5.69	1.29	0.42
Charleston,SC	3.3	33.3	80.8	7.3	3.69	-0.39	-0.13
Nashville,TN	-0.6	36.7	87.6	177.4	-0.11	-0.49	-0.16
Houston,TX	6.7	30.1	95.9	29.9	9.40	-2.70	-0.87
SaltLakeCity,UT	-7.8	41.1	112.3	1288.4	-7.62	-0.18	-0.06
Seattle,WA	0.6	48.1	122.5	130.8	-7.68	8.28	***2.67***
Madison,WI	-12.8	43.4	90.2	262.7	-7.80	-5.00	-1.62
Cheyenne,WY	-10.0	41.2	104.9	1876.0	-12.05	2.05	0.66

(source : http://name.math.univ-rennes1.fr/bernard.delyon/tp/temp.dat)

Tableau 6.7 : Températures moyennes de janvier de 33 villes des USA

Figure 6.2 : localisation des villes de l'échantillon

L'équation de régression est : $Y' = -1.134\ X_1 + 0.146\ X_2 - 0.006\ X_3 + 29.681$

Le tableau 6.8 vous fournit les coefficients de détermination entre les 4 variables.

	T°jv	**Lat**	**Long**
Lat	0.752		
Long	0.000	0.032	
Alt	0.242	0.124	0.211

Tableau 6.8 : r² entre variables

On considère, dans cet exercice, que Y est entaché d'une erreur aléatoire (d'échantillon). Par conséquent, les tableaux 6.9 et 6.10 fournissent l'analyse de variance de la régression et le test t de Student sur chacun de ses coefficients avec leurs risques d'erreur.

source	Degrés de liberté	Somme des carrés	Carrés moyens	F de Fisher	Risque d'erreur
Régression	3	1797.96	599.32	56.68	0.0001
Résidus	29	306.66	10.57		
Total	32	2104.62			

Tableau 6.9 : analyse de variance de la régression multiple

	Valeur	erreur-type	t de Student	Risque d'erreur
Constante b	29.681	5.396	5.501	0.000
Lat	-1.134	0.108	-10.527	0.000
Long	0.146	0.041	3.529	0.001
Alt	-0.006	0.001	-4.062	0.000

Tableau 6.10 : significativité des coefficients de la régression

La figure 6.3 présente l'histogramme des résidus et leur distribution par rapport à Y'.

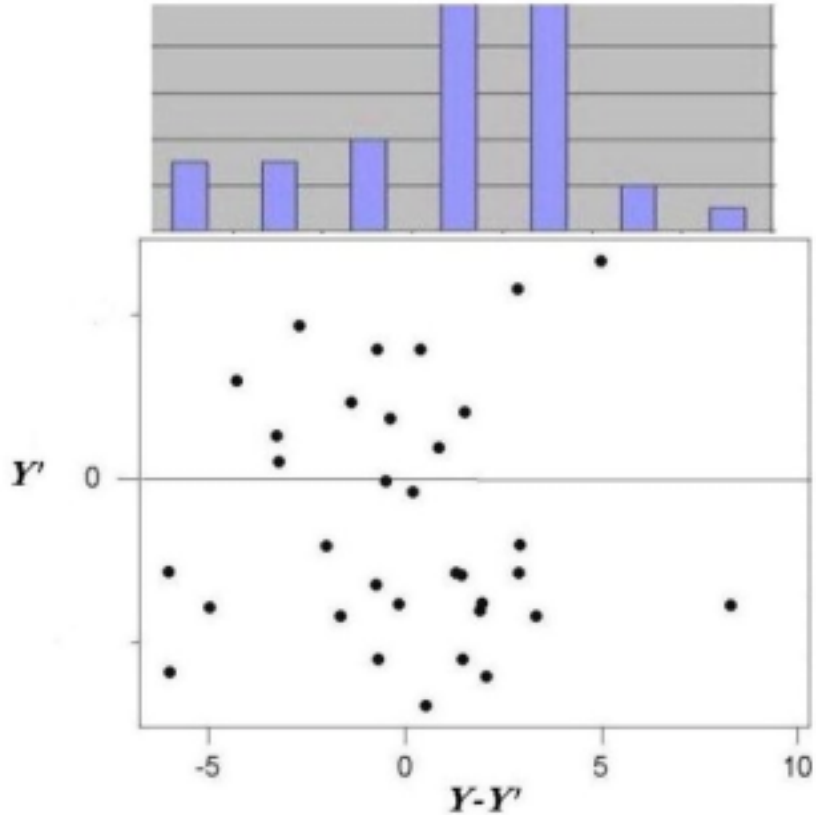


Figure 6.3 : distributions des résidus de la régression multiple

Questions

Les principales variables explicatives des températures à l'échelle continentale, notamment en hiver, sont bien l'altitude (diminution moyenne à cette échelle de 0.6° tous les 100 mètres), la latitude (plus ou moins proche du pôle ou de l'équateur) et la longitude (continentalité, opposition des façades maritimes est et ouest). La régression devrait donc logiquement conduire à un fort R^2 et permettre de juger l'effet de ces 3 facteurs, à condition que les principales contraintes d'une régression multiples soient respectées.

1) **Ces principales contraintes sont elles ici respectées ?**
Pour ce faire, examinez les coefficients de détermination entre les variables (Tableau 6.8)

2) **Quelle part de variance de Y expliquent les 3 variables explicatives ?**

3) **Commentez leur importance relative dans la différenciation des températures moyennes de janvier.**

4) **L'échantillon des 33 villes vous semble t'il spatialement représentatif ? de quoi ?**

5) **Faisons comme s'il l'était. Commentez les résultats des tests (R^2, R, coefficients de l'équation de régression) et le graphique des résidus : peut on extrapoler à tout le territoire des USA ?**

Réponses suggérées

Question 1

La première condition de validité à examiner porte sur l'*indépendance* des 3 variables explicatives, car l'équation de régression multiple additionne leurs valeurs (pondérées par les a_j) pour expliquer celles de Y. On peut vérifier l'indépendance des X_j en observant les coefficients de détermination du tableau 6.8. Il y a réellement indépendance ($r^2 \cong 0$) entre latitude et longitude des 33 villes (traduisant la bonne répartition planimétrique de l'échantillon) et r^2 faible entre latitude et altitude. Par contre, il y a quelque dépendance ($r^2=0.211$ soit 21% de redondance) entre altitude et longitude, ce qui traduit une petite concentration des villes retenues à certaines altitudes et latitudes (les principales zones montagneuses, les Rocheuses, sont à l'ouest des USA). Cela ne nous empêchera pas de garder l'altitude comme variable explicative mais rendra prudent dans l'interprétation.

La seconde condition fondamentale à vérifier est la *linéarité* des régressions simples entre les X_j et Y : on peut, indirectement, apprécier si les relations sont bien ajustées par des droites (ou pas) en examinant les coefficients de détermination r^2 du tableau 6.8 (colonne T°jv). Si le r^2 est assez voisin de 1, il y a bonne relation linéaire (c'est le cas entre T°jv et latitude). S'il est voisin de 0, ou bien il n'y a pas de corrélation linéaire ou bien pas de corrélation du tout (rappelons qu'il est souhaitable que les variables explicatives X_j soient bien corrélées avec la variable Y à expliquer). Dans l'échantillon des 33 villes (figure 6.2), la température moyenne de janvier est :

- indépendante de la longitude ($r^2=0.00$), ce qui traduit sans doute la sur représentation dans l'échantillon des villes de 2 façades maritimes (est et ouest), très différentes du point de vue thermique et la sous représentation des villes du « middle west »,
- faiblement dépendante de l'altitude ($r^2=0.24$), ce qui s'explique par la forte proportion dans l'échantillon de villes côtières, de très faible altitude (où le facteur thermique majeur n'est pas l'altitude).

Pour vérifier la linéarité des relations entre Y et les X_j , nous avons graphiqué les 3 nuages de points (figure 6.4).

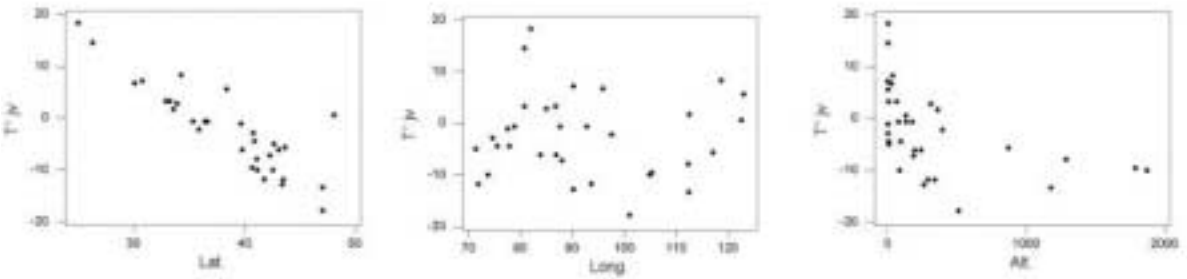


Figure 6.4 : nuages de points température de janvier – variables explicatives

Il en ressort que :

- mis à part un net « outlier », la relation entre T°jv et latitude est raisonnablement linéaire,
- il y a réellement indépendance (dans l'échantillon) entre T°jv et longitude,
- il y a tendance à relation (mais de type exponentielle négative et non linéaire) entre T°jv et altitude (grand nombre de villes de faible altitude).

Si la condition d'indépendance entre elles des X_j est « raisonnablement » respectée, on ne peut en dire autant de celle de linéarité entre les X_j et Y : on doit s'attendre à ce que le facteur majeur de variation des températures hivernales soit la latitude (fréquentes coulées d'air polaire dans le centre et l'est des USA, en outre pas ou peu montagneux).

Question 2

La part de variance de Y qu'explique l'addition des 3 variables explicatives peut être tirée du tableau 6.9 puisque :

- R^2 = somme des carrés due à la régression / somme des carrés totale
- Soit, ici, R^2 = 1797.96 / 2104.62 = 0.854 = **85.4%**

La conjonction des 3 variables latitude, longitude, altitude explique 85% des différences de température moyenne des 33 villes de l'échantillon, mais fort probablement l'influence de chacune est fort diverse : l'examen de l'équation de régression permettra de le constater.

Question 3

L'équation de régression est :

$$Y' = -1.134\ lat + 0.146\ long - 0.006\ alt + 29.681$$

Si les variables explicatives étaient parfaitement indépendantes, la température de janvier :

- diminuerait de 1.134° celsius pour tout déplacement en latitude vers le nord de 1°,
- augmenterait de 0.146° celsius pour tout déplacement en longitude vers l'ouest de 1°,
- diminuerait de 0.6° celsius pour toute augmentation d'altitude de 100 mètres,

par rapport à une température constante de 29.681° celsius. La latitude a clairement une importance majeure, au moins pour les villes de l'échantillon.

Si les X_j étaient parfaitement indépendantes (r^2=0 exactement pour tous les couples de variables explicatives), la somme des coefficients de régression standardisés serait de 1 et chacun fournirait (en % de variance de Y) l'importance relative de chaque X_j sur les variations de Y. Ce n'est pas tout à fait le cas ici.

Question 4

La consultation de la carte de répartition des 33 villes sur le territoire des USA (figure 6.2) et les remarques faites (Question 1) nous indiquent assez clairement que l'échantillon n'est pas tout à fait spatialement représentatif (cela pourrait d'ailleurs être testé). En effet, les villes retenues sont soit des capitales d'état soit de grandes villes : leur répartition reflète surtout l'histoire du peuplement des USA (états plus petits et nombreux à l'est) et la densité de population plus qu'une répartition satisfaisante du point de vue climatologique (comme aurait pu l'être une grille régulière de points de mesure).

La figure 6.2 fait clairement ressortir :

- la sur représentation de la partie est du territoire, la plus anciennement colonisée et la plus basse, et donc la sous représentation de la partie ouest, la plus montagneuse,

- la très forte sous représentation du « middle west » (S. Dakota, Nebraska, Kansas).

Nous allons néanmoins traiter cet échantillon comme représentatif, pour exemplifier le traitement de données comportant une *erreur aléatoire* sur Y (erreur provenant de l'échantillonnage et/ou d'erreurs non systématiques de mesure de Y). Considérons ici que l'erreur aléatoire provient de l'imprécision des mesures de température.

Question 5

Les tests doivent porter, dans ce cas, sur :

- la significativité de la régression,
- la significativité de chacune de ses variables explicatives,
- la conformité des résidus.

- *significativité de la régression*

L'analyse de variance (tableau 6.9) qui a permis de calculer R^2, permet aussi de juger de la validité globale de la régression. Le F calculé de Fisher Snedecor est égal au carré moyen dû à la régression divisé par celui dû aux résidus (un carré moyen est la somme des carrés divisée par le nombre de degrés de liberté correspondant). Ici F = 599.32 / 10.57 = **56.7** (la variance moyenne des résidus est 56 fois plus petite que celle prise en compte par la régression). On peut tester ce F calculé à l'aide d'une table mais nous est directement fourni le risque d'erreur de 0.0001 (0.01% soit 1 pour 10 000) : il y a une chance sur 10 000 que la corrélation entre température moyenne de janvier et latitude, longitude, altitude ne soit pas avérée. En d'autres termes, les erreurs de mesure des températures sont insignifiantes et pas de nature à contester la validité de l'explication, au moins pour l'ensemble des 33 villes retenues. Mais, est ce à dire que toutes les variables explicatives sont significatives ?

- *significativité de chacune des variables explicatives*

Le tableau 6.10 permet d'en juger à travers les calculs du t de Student (=valeur/erreur type). Tous les paramètres de la régression sont significatifs (avec un risque d'erreur maximum de 1 pour mille) : toutes les variables explicatives jouent un rôle quasi certain dans la détermination des températures moyennes de janvier des 33 villes.

- *conformité des résidus*

Encore faut il, pour cela, que les résidus Y-Y' aient une distribution conforme aux conditions
1) de normalité (l'histogramme des Y-Y' doit suivre à peu près une loi de Gauss),
2) d'homoscédasticité (la variance des résidus doit être à peu près égale pour tout intervalle de valeurs de Y'),
3) d'« outliers » peu nombreux (un outlier a un résidu réduit (standardisé) important).

La figure 6.3 permet d'en juger. Le diagramme en bâtons (du haut de la figure) n'est pas trop dissymétrique et peut être approximé par une gaussienne (condition 1). Le nuage de points Y' par rapport aux résidus Y-Y' a une variance à peu près constante pour tout intervalle de valeurs de Y', sauf pour les plus fortes (condition 2). La dernière colonne du tableau 6.7 permet de repérer un seul outlier, Seattle, avec une valeur supérieure à 1.96 (condition 3). La température de janvier à Seattle est « anormalement » élevée (courant chaud des façades ouest de continent aux latitudes tempérées ?). Les 3 conditions sont à peu près remplies : si l'échantillon avait été spatialement représentatif, on aurait pu interpoler ses résultats à tout le

territoire continental des USA. Cette représentativité n'étant pas assurée, la régression ne vaut que pour l'échantillon des 33 villes.

- *compléments*

Si l'on effectue une régression multiple pas à pas ascendante, les variables explicatives progressivement intégrées sont la latitude, l'altitude et, en dernier lieu, la longitude qui améliore peu la valeur de R^2 et qui, en outre, est corrélée à l'altitude et très peu corrélée avec la température de janvier : on peut donc l'ôter de la régression.

Comme latitude et altitude sont partiellement corrélés, il est bon de calculer les coefficients de corrélation partielle. Celui entre température et latitude *à altitude constante* vaut **-0.71**, celui entre température et altitude *à latitude constante* vaut **–0.17**, ce qui fixe l'ordre d'importance des facteurs de température hivernale, pour les 33 villes considérées.

Exercice 2

Le tableau 6.11 présente, pour les actifs ayant un emploi en France métropolitaine au recensement de population de 1999, leur répartition régionale par type de contrat professionnel et le PIB/habitant de chaque région. Rappelons que le Produit Intérieur Brut est l'ensemble des richesses produites une année donnée (en 2003 ici) par l'ensemble des agents économiques présents dans la région (originaires ou non de celle ci) : divisé par le nombre d'habitants, il fournit un indicateur de richesse moyenne produit dans chaque région.

L'Ile de France et la Corse ayant des répartitions tout à fait particulières ont été supprimées du tableau si bien que l'exercice ne porte que sur 20 régions de « province ».

	CDI	CDD	Intérim	Emploi aidé	stage	Fonction publique	PIB/hb
Alsace	276965	22779	12671	3660	10023	51569	24.804
Aquitaine	325976	37591	12540	10106	13394	94930	22.475
Auvergne	154119	15293	5871	5158	6933	41442	21.011
B.Normandie	170365	18281	7901	5400	8043	39725	20.599
Bourgogne	199534	19129	8570	4850	8524	49584	22.511
Bretagne	335711	40080	13868	7689	14052	95480	21.402
Centre	318317	30240	15117	6151	13020	75847	22.192
Champagne	170825	17979	8063	4885	6778	43046	22.926
F.Comté	154957	13318	7059	3079	6151	33719	21.897
H.Normandie	240019	22760	12742	6430	10533	45338	23.013
Languedoc	207669	30979	6022	10349	11132	77163	19.416
Limousin	76230	7358	2617	2769	3200	25468	20.592
Lorraine	315154	28443	15084	7239	11320	78272	20.297
Midi Pyrénées	282218	32731	9444	8820	12054	84443	22.025
Nord	487541	49067	23790	22398	16640	110868	19.835
Pays de Loire	419802	45992	22254	8702	21859	82104	22.300
Picardie	250523	24527	11723	8930	9088	54358	19.932
Poitou	184493	20427	7147	6390	9425	48438	20.325
Provence	457278	65501	13433	13262	20768	155312	22.901
Rhône Alpes	771394	84230	29968	13732	25111	152309	25.153

(*Source : INSEE, RGP 1999)*

Tableau 6.11 : population employée par type de contrat et PIB/habitant

On fait l'hypothèse qu'il existe une relation entre PIB/habitant et fréquence des types de contrat de travail mais il n'est pas envisageable de faire directement une régression multiple expliquant le PIB/hb par les types de contrat. Ceux ci constituant un tableau de contingence, on l'a résumé par AFC (2 axes retenus, prenant en compte 70% et 16% de la variance soit près des 9/10^{ième} à eux deux). La figure 6.5 en présente le nuage de points (région et statut).

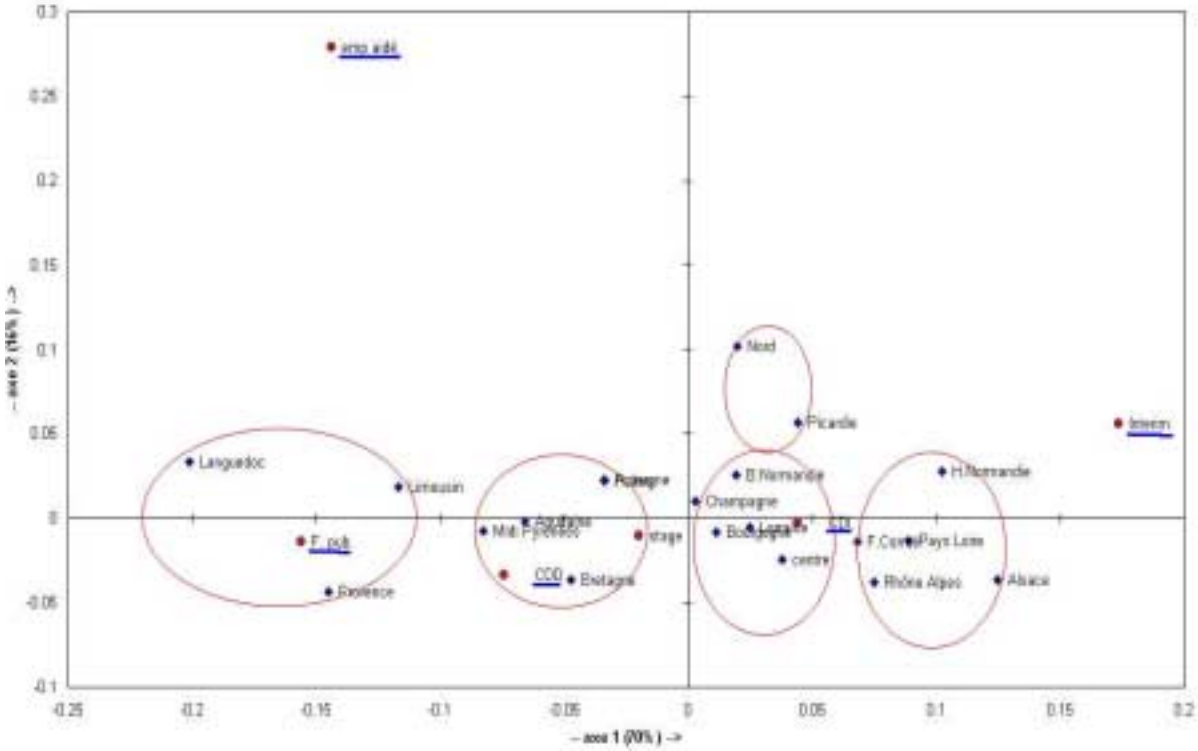


Figure 6.5 : Plan des axes 1 et 2 de l'AFC portant sur les types de contrats de travail

On peut maintenant procéder à une régression expliquant le PIB/hb (exprimé en milliers de francs 1999) par les axes F1 et F2. Le tableau 6.12 fournit les r² simples entre les variables.

	PIB/hb	F1
F1	0.198	
F2	0.471	0.005

Tableau 6.12 : r² simples entre les 3 variables

Les nuages de points des régressions simples PIB/hb – F1 et PIB/hb – F2 sont représentés par la figure 6.6.

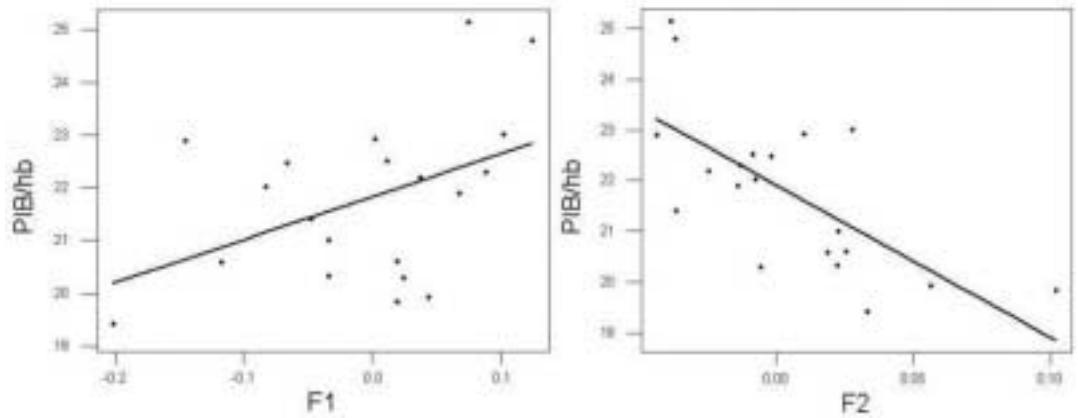


Figure 6.6 : régressions simples entre PIB/hb et F1 puis F2

L'équation de régression est : **PIB/HB'=7.3F1-28.9F2+21.9**

Elle conduit à un R^2 de 0.63, résultat assez conforme à l'hypothèse émise. Une régression pas à pas indique que F2 est une meilleure variable explicative du PNB/hb régional que F1.

Pour aller plus loin, on tente de valider l'hypothèse que la relation entre PIB/habitant et facteurs résumant les contrats de travail est sensible au fait qu'une région est située ou non au nord ou à l'est de la fameuse ligne « Marseille – Le Havre » sensée partager l'hexagone entre une France industrielle et une France qui l'est peu. Une analyse de covariance intégrant cette variable qualitative donne un R^2 de 0.67, à peu de chose près identique à celui de la régression multiple sans cette variable qualitative.

Séparant les 20 régions en 2 paquets de 10 régions selon leur situation géographique (Nord et Est ou Sud et Ouest de la ligne Le Havre – Marseille), on a effectué 2 régressions multiples dont les résultats sont les suivants :

- zone N et E : R^2=0.63, équation : **PIB/hb'=7.7F1-31.2F2+22.2**
- zone S et O : R^2=0.66, équation : **PIB/hb'=3.9F1-26.4F2+21.5**

Dans les 2 paquets la variable F2 explique mieux le PIB/habitant que la variable F1.

Questions

1) **Interprétez le résumé effectués par les 2 premiers axes de l'AFC (figure 6.5)**

2) **A votre avis, les conditions de validité sont elles (à peu près) réunies pour expliquer le PIB/habitant par les 2 facteurs de l'AFC ?**

3) **Si oui, interprétez l'équation de régression**

4) **Que penser de la ligne Le Havre – Marseille (au vu des données de cet exercice et à l'échelle régionale) ?**

Réponses suggérées

Question 1

L'interprétation des 2 axes de l'AFC (F1 :70%, F2 :16% de l'information initiale) se fera ici à partir de le figure 6.5.

- Le 1er axe oppose des régions où l'emploi dans la fonction publique est sur représenté (Languedoc, Provence, Limousin) à des régions où est sur représenté l'emploi en CDI et en intérim (régions industrielles : Alsace, Haute Normandie, Pays de Loire, Franche Comté, Rhône Alpes).
- Le 2nd axe met en relief deux régions où l'emploi précaire (emploi aidé, intérim) est sur représenté (Nord, Picardie).

Une typologie effectuée sur le plan des axes 1 et 2 aboutit à 5 classes :

- forte importance relative de la fonction publique (Languedoc, Provence, Limousin), faute d'emploi industriel suffisant,
- importance relative de l'emploi à durée déterminée (Midi Pyrénées, Aquitaine, Bretagne, Auvergne, Poitou),
- importance relative des CDI en Basse Normandie, Champagne, Lorraine, Bourgogne, Centre,
- relative importance des emplois aidés et de l'intérim dans le Nord et en Picardie,
- relative importance des CDI et de l'intérim dans les régions les plus industrielles (Rhône Alpes, Franche Comté, Pays de Loire, Haute Normandie et Alsace).

Trois facteurs majeurs de différenciation : l'emploi industriel, l'emploi dans la fonction publique, des formes d'emploi précaire.

Question 2

La 1ère condition à vérifier porte sur l'*indépendance* des variables explicatives : par construction, les axes factoriels sont orthogonaux donc indépendants (le r^2 de 0.005 est dû aux arrondis de calcul). Il est également souhaitable que les variables explicatives soient bien corrélées avec la variable à expliquer, ce qui est surtout le cas de F2 (cf tableau 6.12).

La 2nde condition à vérifier impérativement porte sur la *linéarité* des relations entre variable à expliquer et variable explicative : la figure 6.6 montre que c'est à peu près le cas (bien que d'autres ajustements que linéaires auraient mieux ajusté les 2 nuages de points).

On peut donc considérer qu'il est envisageable d'ajuster un modèle de régression multiple sans dénaturer le problème posé.

Question 3

L'équation de régression est : **PIB/HB'=7.3F1-28.9F2+21.9**

Rappelons que F1 oppose régions à surplus relatif de fonction publique et régions à surplus d'emploi en CDI et intérim. F2 met principalement en lumière les emplois précaires aidés (par les collectivités publiques).

En général, pour les 20 régions (moins l'Ile de France et la Corse), la richesse produite par habitant augmente, au dessus de 21 900 francs 1999, de 7 300 Francs quand F1 (CDI et intérim, à forte orientation industrielle) augmente de 1 unité et diminue de 28 900 quand F2 (emploi précaire, généralement peu rétribué) augmente de 1 unité. F1 et F2 étant indépendants, leurs effets s'additionnent directement, expliquant tous deux ensemble 63% des différences régionales de PIB/hb. Plus du 1/3 de celles ci sont à rechercher dans d'autres facteurs.

Question 4

En effectuant 2 régressions multiples, une pour les 10 régions situées au nord et à l'est de la ligne Le Havre – Marseille, l'autre sur les 10 régions situées au sud et à l'ouest, on obtient 2 équations de régression aux coefficients a_j différents (donc empêchant de faire une même analyse de covariance sur les 20 régions ensemble).

$Y'_{NE} = 7.7\ F1 - 31.2\ F2 + 22.2$ et **$Y'_{SO} = 3.9\ F1 - 26.4\ F2 + 21.5$**

Globalement, une augmentation unitaire de l'emploi en CDI ou interim fait augmenter bien plus le PIB/hb et l'emploi précaire le fait moins diminuer à l'est et nord qu'au sud et ouest. Mais la variance expliquée n'a guère augmenté (63% et 66% au lieu de 63%).

L'explication du PIB/habitant bute sur ce plafond des 2/3 et la raison en est simple :

- si l'on considère les 5 régions à plus faible PIB/hb (<20 500 Francs 1999), 3 sont situés d'un coté de la ligne Le Havre – Marseille (Nord, Picardie, Lorraine) et 2 de l'autre (Languedoc, Poitou) et n'appartiennent pas aux mêmes classes de profil d'emploi,
- si l'on considère les 5 régions à plus fort PIB/hb (>22 500 Francs 1999), elles sont toutes à l'est de Marseille – Le Havre mais 3 (Haute Normandie, Alsace, Rhône Alpes) figuraient dans la classe 5 (prédominance CDI et intérim, grande industrie) et 2 (Bourgogne, Champagne) dans d'autres classes.

Par conséquent, la ligne Marseille – Le Havre, qui avait une validité descriptive certaine en termes d'intensité industrielle, n'a aujourd'hui qu'une validité restreinte en termes de richesse produite par habitant dans la mesure où la productivité industrielle (grande créatrice de PIB) s'est fortement différenciée, par branche et, donc, par région.

Chapitre 7

Méthodes explicatives : compléments

Sous ce titre, on présente ici deux méthodes (parmi beaucoup d'autres possibles) :

- analyse discriminante,
- segmentation (ou « arbre de décision »).

Elles ont en commun de viser le même but général « explicatif » (au sens statistique du terme) que la régression multiple mais ne portant pas sur des variables exclusivement quantitatives. On les présente ici dans un objectif de compréhension générale, à travers des exemples, plutôt que d'application à l'aide d'exercices corrigés. On entre donc très peu, ici, dans les détails procéduraux.

1. L'analyse discriminante

1.1 modèle général

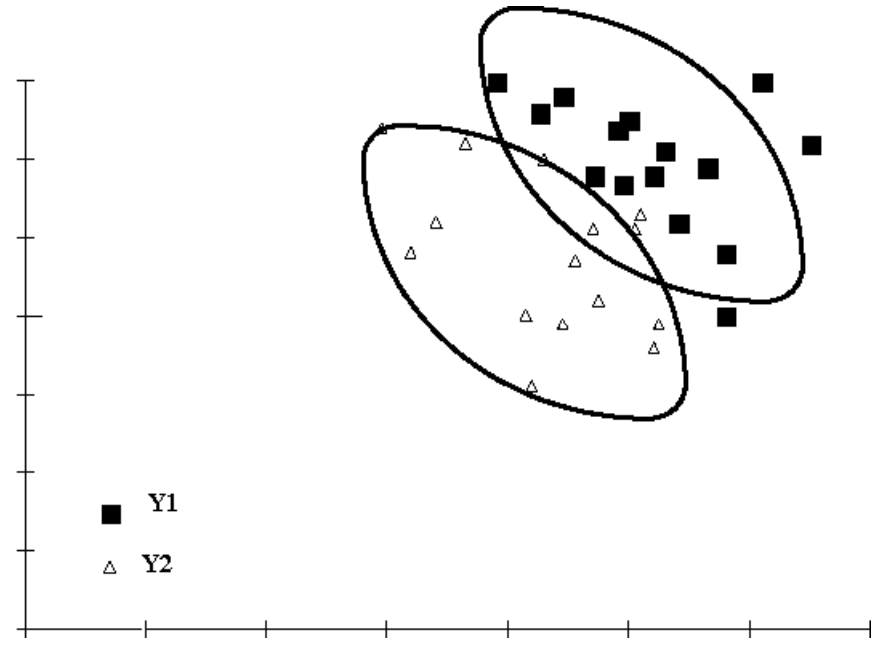
L'idée de base est d' »expliquer» Y, variable *catégorielle* (qualitative ou quantitative à classes peu nombreuses) par un ensemble $X_1,\ldots,X_p$ de variables *quantitatives*. Le but de l'analyse discriminante est de remplacer les $X_1,\ldots,X_p$ par un nombre réduit de *fonctions discriminantes* Z qui différencient au mieux les classes de Y par leurs valeurs sur $X_1,\ldots, X_p$. Comme en ACP, ces fonctions, orthogonales les unes aux autres et de variance décroissante, sont obtenues par combinaison linéaire des variables $X_1,\ldots,X_p$.

L'équation de ces fonctions linéaires discriminantes, sur variables centrées, sont de la forme:

$$Z = a_1 X_1 + \ldots + a_p X_p$$

Les coordonnées des points sur une droite discriminante sont les projections orthogonales sur celle-ci, de façon à minimiser la superposition entre groupes (maximiser la variance expliquée par le modèle).

Considérons un exemple élémentaire (figure 7.1): soit une variable catégorielle Y à 2 modalités (individus représentés par un rectangle noir ou un triangle blanc) et 2 variables quantitatives X_1 (en ordonnées) et X_2 (en abcisses). Dans une application géographique, la variable Y pourrait être l'appartenance des pays européens au bloc de l'ouest ou de l'est et les 2 variables quantitatives leurs taux de natalité et de mortalité.

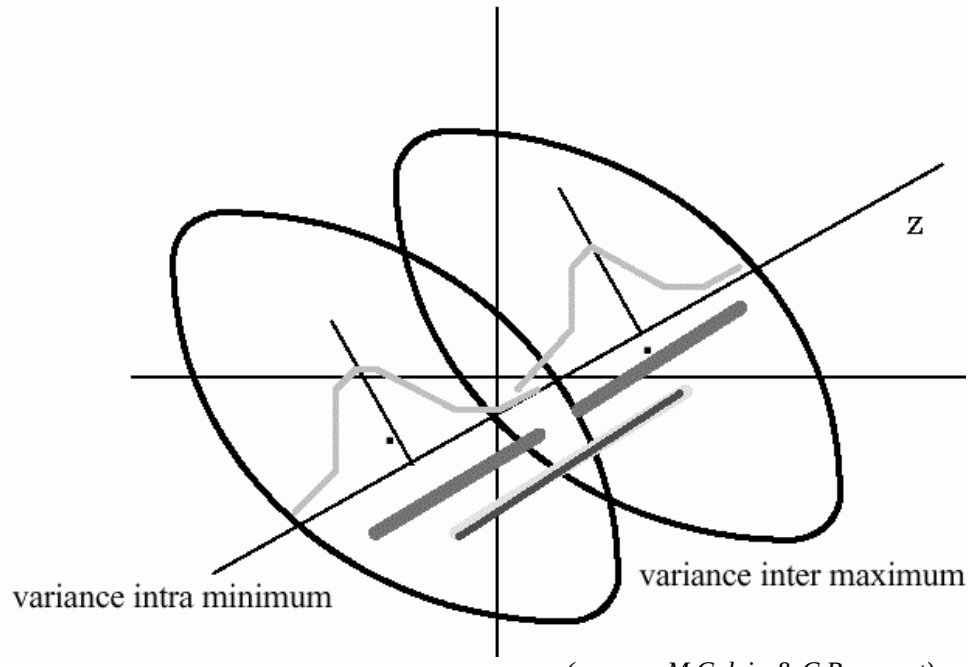


(source : M.Calciu & C.Benavent)

Figure 7.1 : nuage de points sur X1-X2 avec différenciation des 2 classes de Y

La figure 7.1 présente le nuage de points X1-X2 où l'on a différencié les individus appartenant aux classes Y1 et Y2 de Y. La plupart d'entre eux sont à l'intérieur d'ellipses permettant de vérifier si les distributions de X1 et X2 sont à peu près gaussiennes dans chaque classe de Y. Ces ellipses se chevauchent partiellement, définissant une zone d'incertitude.

La recherche de fonctions discriminantes (une seule dans cet exemple) a pour but de minimiser cette zone de chevauchement en maximisant la variance expliquée (variance inter-classes / variance totale sur X1-X2) comme le montre la figure 7.2.



(source : M.Calciu & C.Benavent)

Figure 7.2 : fonction discriminante Z maximisant la variance expliquée

1.2 Deux usages de l'analyse discriminante

Selon que l'accent est mis sur ses résultats en termes de variables ou d'individus, on insiste

- sur l'analyse sémantique des fonctions discriminantes et de leurs paramètres,
- sur la construction d'une classification explicative / prédictive,

les deux usages étant complémentaires pour le géographe (classer / expliquer ou vice versa).

1.2.1 *analyse des fonctions discriminantes*

La question prioritaire est alors : « *qu'est ce qui rend les classes de Y différentes les unes des autres* » ? Dans l'exemple élémentaire évoqué ci-dessus, on peut déterminer le profil de chacun des deux groupes de pays quant à leurs natalité et mortalité.

Les coefficients $a_1,\ldots,a_p$ de la fonction discriminante sont calculés de façon à maximiser le rapport *variance inter-groupes / variance totale* c'est à dire à séparer au mieux des classes les plus homogènes possibles. Ils indiquent le poids des variables dans la différenciation des groupes de Y mais coefficients *bruts* (exprimés dans l'unité de mesure des X_j), ils n'indiquent pas l'importance *relative* de chaque X_j dans la construction de la fonction discriminante. Pour éviter l'effet d'unités de mesure différentes des variables explicatives, on calcule des coefficients $\alpha_1,\ldots, \alpha_p$ standardisés (a_j / σ_{Xj}) pour connaître la contribution de chacune.

L'intensité de la relation entre Y et les X_j (équivalent du R^2 de la régression multiple) se détermine en rapportant la variance inter-groupes à la variance totale.

Si le jeu de données comporte une erreur aléatoire, on procède à un test F de Fisher-Snedecor pour savoir si la fonction discriminante est probablement significative (c'est à dire si les groupes de Y sont probablement distincts ou non quant aux variables explicatives retenues).

1.2.2 *Construction d'une classification explicative / prédictive*

Les fonctions discriminantes constituent des frontières entre classes de Y sur la base de combinaisons linéaires des variables quantitatives continues $X_1,\ldots,X_p$. Ce sont des « classifieurs » géométriques : un individu est affecté à la classe dont le centre de gravité lui est le plus proche (au sens de la distance de Mahalanobis).

L'analyse discriminante peut donc aussi être vue comme une méthode de *classification explicative*. Le profil de chaque classe est connu par les valeurs de la projection de son point moyen sur les fonctions discriminantes, son homogénéité par sa variance intra-classe et la différence entre classes par la variance inter-classes.

L'analyse discriminante peut surtout être utilisée comme une méthode de *classification prédictive*. De nouveaux individus peuvent être introduits : ils seront affectés à la classe dont le centre de gravité leur est le plus proche. On peut alors observer l'évolution des variances intra et inter-classes pour voir si la classification est stable ou non et a encore une qualité acceptable. Dans une perspective de validation d'une typologie, le géographe peut partager sa

population ou son échantillon en deux parties d'effectifs égaux, faire l'analyse sur un des deux groupes puis, indépendamment ou en individus supplémentaires, sur l'autre et comparer.

En général, les groupes s'intersectent partiellement : un individu appartient dans ce cas à plusieurs groupes avec des probabilités plus ou moins grandes selon ses distances à leurs centres de gravité. Les groupes sont alors des sous ensembles « flous » : on peut connaître les probabilités d'appartenance d'un élément à tel ou tel d'entre eux. Des probabilités voisines indiquent une zone de chevauchement et on peut alors :

- Ou bien laisser ces individus hors classification (pour n'avoir que des classes homogènes),
- Ou bien faire des zones de chevauchement des classes mixtes, considérées à part.

1.2.3 *Conditions de validité du modèle*

L'analyse discriminante suppose :
- l'égalité des variances et covariances dans les groupes,
- une distribution normale des variables explicatives X_j dans chaque classe : le modèle est assez robuste quand on s'en éloigne mais il y a dégradation de sa qualité.

Evidemment, les variables explicatives X_j ne doivent pas être co-linéaires. Si elles sont fortement corrélées, une ACP préalable est recommandée : l'analyse discriminante s'effectue alors sur les 1^{ers} axes factoriels, parfaitement indépendants.

1.3 Exemple d'analyse discriminante

Cet exemple est tiré d'un rapport de 2003 (dû à Robert J.Wolfe et Victor Fisher) au « US Fish & Wildlife Service » d'Alaska. L'objectif de l'étude est d'y discriminer populations rurales (bénéficiant d'un soutien financier) et non rurales.

Pour ce faire, l'analyse discriminante a été mise en œuvre sur une sélection de 195 populations (ayant plus de 49 personnes et une production vivrière annuelle inférieure à 450 kilos par tête). Les populations de moins de 50 personnes ou dont la production vivrière annuelle par tête est supérieure à 450 kilos ont, en effet, été directement classées comme rurales. La variable (qualitative) à expliquer a donc deux modalités (rural / non rural) et deux variables (quantitatives) explicatives ont été retenues : la production vivrière annuelle par tête et la densité de population.

La figure 7.3 présente le nuage de points de la relation entre production vivrière (Y sur la figure) et la densité (X sur la figure) ainsi qu'une droite de séparation entre rural (faible densité, production vivrière relativement forte) et non rural (densité relativement forte, faible production vivrière). Cette droite de séparation est encadrée par un intervalle situé à 1 écart type de chacun des points moyens. Cette première analyse, exploratoire, atteste le pouvoir discriminant des 2 variables explicatives retenues.

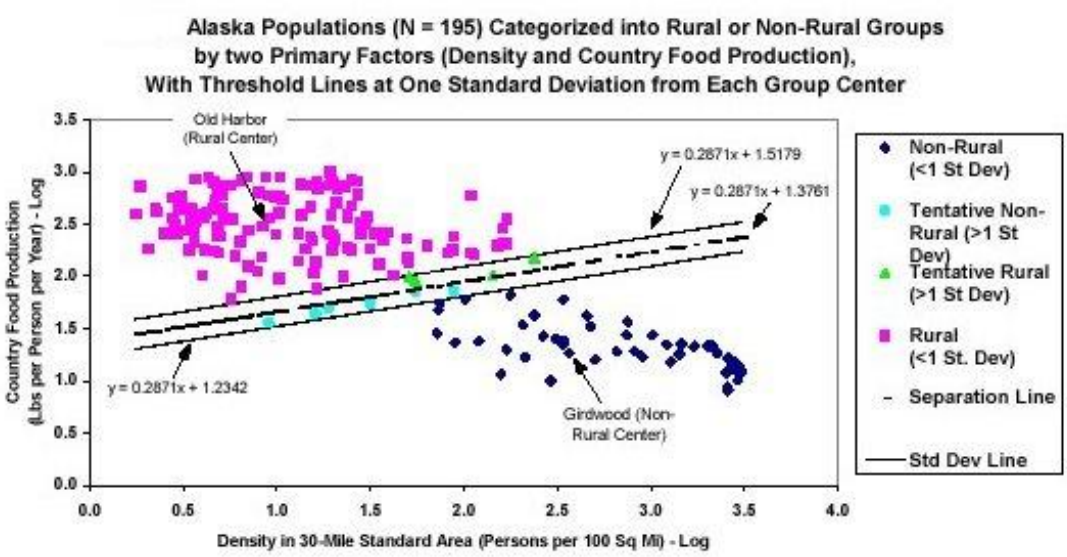


Figure 7.3 : analyse du pouvoir discriminant des 2 variables explicatives retenues *(source : R.J.Wolfe & V.Fisher)*

Une analyse discriminante a donc été menée avec ces deux variables explicatives pour valider cette première typologie. L'équation de la fonction discriminante calculée est :

Y'=2.828 *food* –0.812 *density* –4.882.

La variance expliquée par ce modèle est de **0.832** (sa racine, appelée ici *corrélation canonique* est de **0.907**), ce qui valide encore le choix des 2 variables explicatives.

L'équation de la fonction discriminante permet de calculer le score Y'_i de chacune des 195 populations et de la qualifier de rurale ou non rurale.

De la comparaison avec la première analyse, il ressort que :

- 95% des populations (185 sur 195) ont été classées de la même façon par les 2 méthodes (134 comme rurales, 51 comme non rurales),

- pour les 10 autres populations, d'appartenance douteuse, l'adjonction d'autres variables discriminantes (utilisations extensives du sol, productions spécialisées) a permis d'en affecter 6 au rural et 4 au non rural.

2. La segmentation (ou arbre de classification)

Son idée directrice est voisine de celle de la régression multiple ou de l'analyse discriminante. Comme ces dernières, elle cherche à expliquer une variable par une combinaison de variables explicatives. Comme l'analyse discriminante, elle permet aussi de répartir en classes les éléments d'une population ou d'un échantillon. A la différence des deux méthodes précédentes, elle porte sur des *variables catégorielles* (qualitatives ou quantitatives à peu de classes). Comme la CAH (cf chapitre 5), c'est une méthode construisant une *arborescence* mais qui divise successivement le tout en parties (méthode « descendante ») au lieu d'agréger progressivement les éléments en classes (méthode « ascendante »).

2.1 L'algorithme

La segmentation (aussi appelée, selon les usages qu'on en fait, « arbre de classification » ou « arbre de décision »), produit un arbre binaire cherchant à expliquer les modalités d'une variable ***catégorielle*** Y par un ensemble de variables $X_1,\ldots,X_p$ de nature quelconque mais que l'algorithme transforme, à chaque étape, en variables ***binaires***. Si X_j est quantitative, il recherche le seuil qui en partage au mieux les valeurs en 2 classes, si elle est catégorielle à plus de 2 modalités, il en recherche la meilleure combinaison en 2 classes.

On appelle ***segment*** d'abord l'ensemble de la population ou de l'échantillon puis ses sous ensembles successifs. Le segment initial est appelé « ***racine*** », les segments intermédiaires « ***nœuds*** » et les segments terminaux « ***feuilles*** » de l'arbre.

L'algorithme est itératif : à chacune de ses itérations, il partage les éléments d'un segment en 2 classes en fonction du plus fort gain d'information ou de la meilleure liaison entre variable à expliquer et variables explicatives .

Pour chaque segment (nœud de l'arbre binaire en construction), l'algorithme construit p tableaux de contingence (s'il y a p variables explicatives) : chacun croise la variable à expliquer avec une variable explicative. Le tableau manifestant entre Y et X_j le plus fort gain d'information ou la plus forte liaison permet de retenir, pour ce segment, l'une des variables explicatives dont les modalités ou les valeurs sont partitionnées en 2 parties pour définir 2 sous segments (selon la valeur ou la modalité prise par chaque individu du segment).

Pour que la méthode soit opérationnelle, il faut définir :

- un ***critère*** mesurant le gain d'information ou la relation entre variables à expliquer et explicatives,
- une condition d'***arrêt*** de la construction progressive de l'arbre.

2.1.1 *critère de choix d'une variable explicative*

Plusieurs critères sont possibles. Les plus fréquemment utilisés sont :

- des mesures dérivées du Khi² (V de Cramer, T de Tschuprow, Phi², …),
- des indicateurs fondés sur les distributions de fréquences conditionnelles,
- et, le plus souvent, des mesures dérivées de la théorie de l'information (entropie conditionnelle, redondance, perte d'entropie) bien adaptées aux variables catégorielles.

Le processus de partitions successives garantit que les segments sont de plus en plus homogènes et de mieux en mieux explicatifs de la variable Y. On dit que les segments sont, au fur et à mesure, de plus en plus « ***purs*** » : un segment est pur si tous ses éléments appartiennent à la même classe de Y, impur si leur distribution par classe de Y est uniforme.

2.1.2 Conditions d'arrêt de construction de l'arbre

Deux façons de faire existent (selon le logiciel utilisé) :

- *ou bien, vérifier avant chaque partition d'un segment que* :

- on ne dépasse pas le nombre de niveaux initialement fixé à l'arbre,
- chaque sous segment à créer comprend un nombre minimum d'éléments,
- le gain de pureté est suffisant pour qu'on coupe le segment en 2 sous segments.

- *ou bien créer l'arborescence entière puis élaguer les branches*

  - celles qui ne correspondraient pas à l'une des conditions ci-dessus.

2.2 Aides à l'interprétation

La première aide à l'interprétation est le graphique de l'arborescence.

Pour en faire saisir l'intérêt, considérons l'exemple des motivations de migration résidentielle vers le périurbain à partir d'une enquête menée en 1980 auprès de 1513 chefs de ménage du triangle Lyon – Grenoble – Valence. Pour expliquer l'intensité de leur mobilité (Y, variable à expliquer réduite à 2 modalités : moins de 3 logements occupés, 3 ou plus), on a fait appel à 10 variables explicatives (tableau 7.1), de natures diverses (nominale, ordinale, quantitative) mais ramenées, pour chaque nœud, à 2 catégories par recherche de seuil ou association de modalités.

N°	*Variables*	*Modalités*
1	Age du Chef de Ménage (CM)	9 tranches d'âge décennales (< 20, 20-29, 30-39…)
2	Type de commune	Urbaine, péri-urbaine, bourg, village
3	Type de logement actuel	Ferme, maison ou villa, collectif, autre
4	Type localisation dans commune	Centrale, périphérique
5	Niveau enseignement du CM	Zéro, CEP, BEPC, BAC, plus
6	Niveau enseignement professionnel du CM	Zéro, CAP., Brevets, BAC ou DUT, autre
7	Niveau formation continue	Idem
8	Nombre d'emplois du CM	De zéro à neuf
9	CS initiale du CM	Agriculteur, artisan-commerçant, cadre, ouvrier-employé, non actif
10	CS actuelle du CM	Idem

Tableau 7.1 : variables explicatives de l'arbre de segmentation des 1513 chefs de ménage *(source : P.Dumolard)*

De ces 10 variables explicatives, 4 seulement ont joué un rôle majeur avant que les conditions d'arrêt n'interrompent les partitions. Il s'agit de :

- la CSP (en 1980) des chefs de ménage (sans profession et agriculteurs sont peu mobiles)
- leur âge (carrières résidentielles à peu près achevées à partir de la cinquantaine)
- le statut du logement (individuel, il est souvent en accession à la propriété dont la dette freine la mobilité ultérieure)
- la mobilité professionnelle.

Les feuilles de l'arbre (identifiées sur la figure 7.4 par une lettre de A à F et dont l'effectif est indiqué) sont plus homogènes que les nœuds intermédiaires : de A à F, l'intensité migratoire

des chefs de ménage est croissante : % minimal de « mobiles » pour les sans profession et agriculteurs ayant occupé moins de 3 emplois, maximal pour les professions non agricoles ayant plus de 50 ans et ayant occupé au moins 5 emplois successifs.

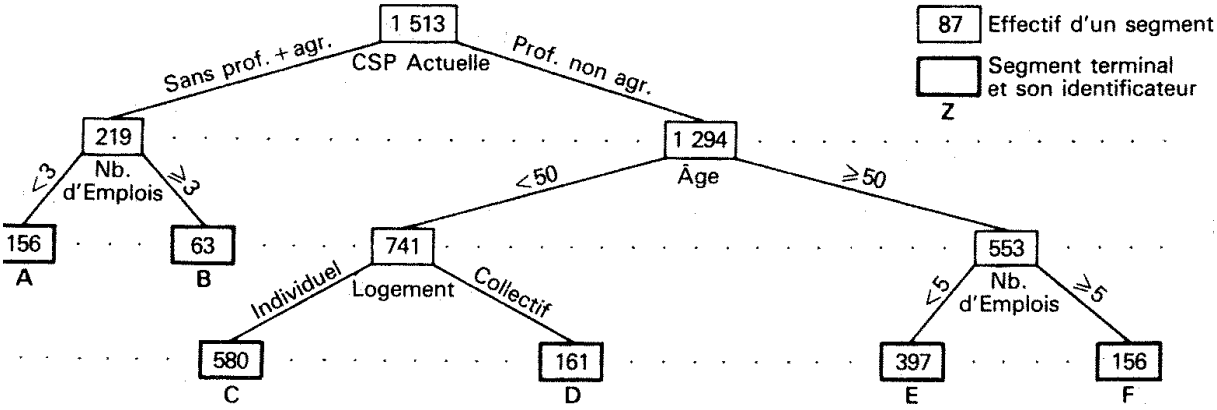


Figure 7.4 : arbre de segmentation des 1500 chefs de ménage *(source : P. Dumolard)*

Le graphique de l'arbre fournit donc aussi un profil explicatif d'une mobilité résidentielle faible ou plus forte, profil explicatif *propre à chaque segment terminal*.

D'autres aides à l'interprétation peuvent aussi être fournies :

- pour chaque segment, les proportions rencontrées de chaque modalité de la variable à expliquer (indiquant le gain d'homogénéité des feuilles par rapport à la racine),
- un indicateur global de qualité de la partition.

2.3 Usages, avantages et limites de la segmentation

2.3.1 *Deux types d'usage*

La segmentation peut être utilisée de deux façons :

- ou bien comme méthode de classification explicative (cf l'exemple ci dessus) sur l'ensemble des éléments,
- ou bien comme méthode de prédiction d'appartenance à une classe. Un arbre est construit sur un échantillon représentatif puis des éléments supplémentaires sont introduits : on observe alors les modifications de la pureté de ses feuilles.

2.3.2 Principales utilisations actuelles

Les principales applications concernent :

- le marketing et la gestion financière,
- le diagnostic médical,
- l'interprétation d'images satellite,
- le traitement d'enquêtes,
- des études environnementales variées (pédologie, géomorphologie, climatologie,…).

2.3.3 Avantages et limites de la méthode

Les principaux avantages sont les suivants :

- en tant que méthode *non paramétrique*, la segmentation n'implique aucune hypothèse sur la distribution (gaussienne) des variables,
- les variables explicatives (binarisées par l'algorithme) peuvent y être de toute nature,
- pour chaque feuille de l'arbre, on dispose de la hiérarchie de leur importance,
- la segmentation crée des règles de partition compréhensibles qui peuvent, par la suite, servir à interroger des bases de données (via un langage de requêtes dénommé SQL),
- l'algorithme, économe en temps calcul, permet de traiter un grand nombre d'éléments.

Mais il y a aussi un certain nombre de limites :

- pour chaque segment à chaque étape de partition, on ne tient compte que d'une seule des variables explicatives et non d'une combinaison de celles ci,
- pour les variables explicatives quantitatives, la perte d'information peut être grande à les réduire à 2 catégories seulement,
- la méthode n'est pas fiable si le nombre d'éléments N est petit et le nombre p de variables explicatives est grand,
- géométriquement, la segmentation crée des *boîtes rectangulaires* de partition de nuages de points, ce qui peut en « caricaturer » certains.

2.4 1er Exemple : les femmes suisses entre la famille et le travail

Cet exemple provient d'une étude due à F.B.Losa et P.Origoni sur les compromis entre carrières familiales et professionnelles des femmes de 20 à 61 ans nées et présentes en Suisse au recensement de 2000. Ces compromis sont dictés à la fois par l'offre du marché de travail en Suisse et la demande de travail des femmes, tenant compte de leurs impératifs familiaux.

Une des hypothèses de base de l'étude est qu'il existe, entre les 3 régions de Suisse (italienne, romande, alémanique) des différences culturelles qui concourent à expliquer la faible participation des femmes au marché du travail en Suisse italienne et la prédilection de celles de Suisse alémanique pour les emplois à temps partiel court.

Pour mettre en évidence ce facteur culturel, 9 arbres de segmentation (dans chaque région distinction entre femmes sans enfant / mariées ou veuves / célibataires ou divorcées) ont été crées et comparés (sur les données du recensement de 2000) à partir des mêmes critères :

- variable catégorielle à expliquer : 4 modalités (pas d'emploi, emploi à temps partiel court, à temps partiel long, à temps complet),

- variables explicatives (binarisées par l'algorithme):
  - fait d'être mère ou pas,
  - existence ou non d'un partenaire actif dans le ménage,
  - niveau de formation de la femme,
  - âge de la femme,
  - profession de la femme,
  - âge du dernier enfant,
  - nombre d'enfants.

A titre d'illustration, nous reproduisons (figure 7.5) l'arbre binaire correspondant aux femmes célibataires ou divorcées de Suisse romande.

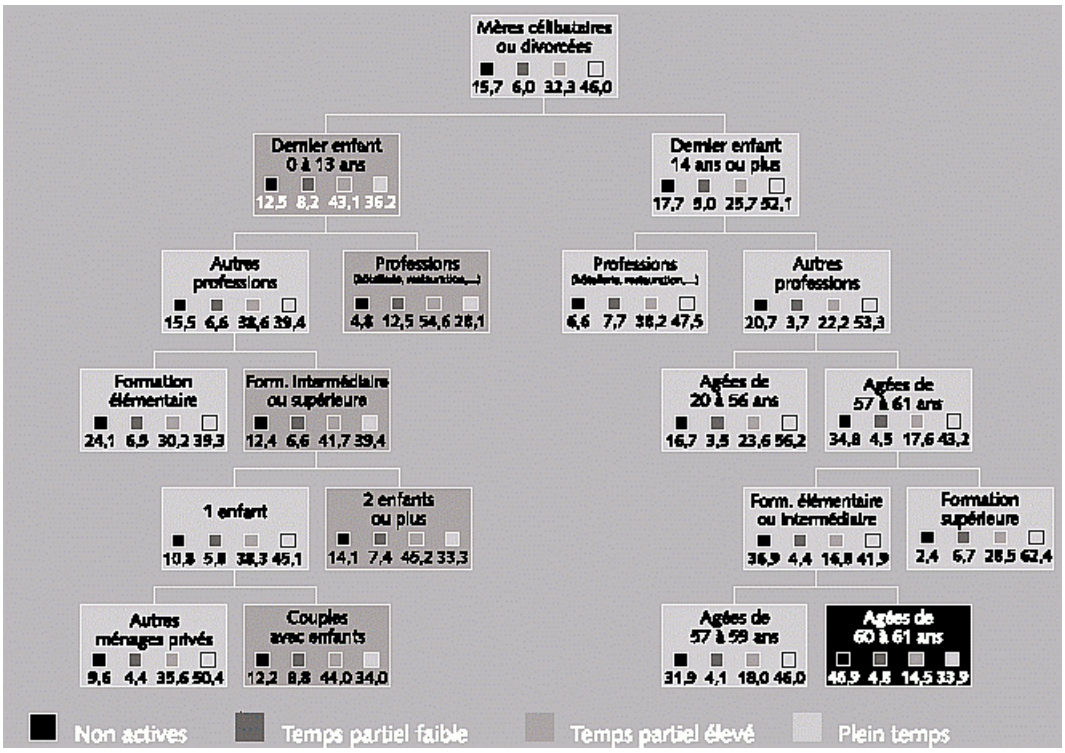


Figure 7.5 : taux d'occupation en 2000 des femmes célibataires ou divorcées de Suisse romande *(source : F.B.Losa & P.Origoni)*

Sur la figure 7.5, chaque segment de l'arbre binaire est accompagné des fréquences correspondant aux 4 modalités de la variable à expliquer, ce qui permet de juger leurs gains de « pureté » au fur et à mesure de la descente de la racine aux feuilles. Ces gains de pureté peuvent être résumés par des calculs d'entropie de la distribution en fréquence des modalités de la variable à expliquer à la racine et pour chaque feuille.

Rappelons ici que l'entropie est, en statistique, une mesure de dispersion de la distribution d'une variable catégorielle. Elle est maximale quand toutes les modalités ont même fréquence (égale à 1/k s'il y a k classes) et minimale si une seule classe a une fréquence de 1.

L'entropie d'une distribution de fréquences vaut :

$$H = -\sum_{j=1}^{k} f_j * \log_e(f_j)$$

L'entropie maximale possible vaut :

$$H_{\max} = -\sum_{j=1}^{k} \frac{1}{k} * \log_e(\frac{1}{k})$$

Et donc, aux fins de comparaison, l'entropie relative vaut :

$$H_r = H / H_{\max}$$

L'entropie relative à la racine est ici de **0.848**, celle pour le segment « dernier enfant >13 ans / autre profession / plus de 56 ans / formation supérieure » de **0.665**. La perte d'entropie mesure donc le gain d'homogénéité de cette feuille quant aux modalités de la variable à expliquer.

Pour une feuille (segment terminal) donné, le chemin depuis la racine indique la hiérarchie de ses variables explicatives et les seuils permettant de les binariser. Par exemple, le % de femmes célibataires ou divorcées employées à plein temps est de 46.0% en Suisse romande, la 1re variable explicative est l'âge du dernier enfant (% montant à 52.1% s'il a plus de 13 ans). Parmi celles ci, il monte à 53.3% si la profession est autre que hotellerie – restauration et, pour ces dernières, à 62.4% pour les plus de 56 ans de formation supérieure : on a là un *profil* particulier de romandes, dont le taux d'occupation à plein temps est supérieur de plus d'1/3 à la moyenne.

Pour en venir à l'hypothèse d'une différence culturelle en termes de taux d'occupation des femmes dans les 3 Suisses, elle semble avérée. Les femmes de Suisse « latine » et particulièrement italienne semblent plus attachées à un modèle traditionnel de famille (impliquant de moindres taux d'occupation) que leurs consoeurs de Suisse alémanique où les stratégies de compromis entre travail et famille semblent traduire un modèle plus « moderne » avec fréquent recours au temps partiel. Cette différence culturelle semble, bien sûr, à combiner avec d'éventuels effets de structure (différences des 3 marchés du travail, compositions par âge des 3 populations expliquant une évolution inter-générationnelle plus ou moins prononcée des mentalités).

2.5 2nd Exemple : la morphologie du terrain sur la planète Mars

L'étude (due à R.Vilalta et T.F.Stepinski) ici résumée expérimente la fabrication d'une carte géomorphologique de la vallée Tisia sur la planète Mars pour en inférer les mécanismes ayant crée les formes.

Les auteurs disposent d'un Modèle Numérique de Terrain (MNT) de la zone crée par le dispositif MOLA (« Mars Orbiter Laser Altimeter »), représenté sur la figure 7.6A : on y remarque visuellement l'importance surfacique de cratères, de pédiments inter-cratères et l'existence de chenaux. Ce MNT (figure 7.6A) est une grille d'altitude de 161 626 cases (on nomme ces cases « *pixels* » en traitement d'images).

A partir de ce MNT, 6 variables ont été calculées pour chaque pixel :

- son altitude moyenne,
- la différence entre son altitude et son altitude comme si toute la zone était drainée,
- sa pente principale,
- sa pente principale comme si toute la zone était drainée,
- le nombre de pixels amont de son « bassin versant »,
- ce nombre comme si toute la zone était drainée.

La masse d'information représentant presque un million de valeurs (161 626 * 6), les auteurs ont constitué, pour les 6 variables, un échantillon spatialement représentatif de 40 000 pixels sur 161 626 (≅1 pixel sur 4). Un algorithme de classification ascendante sur cet échantillon a permis de créer 12 classes de formes (tableau 7.2) elles mêmes regroupées en 4 grandes catégories :

- plaines inter-cratères (classes 1, 5, 8, 11),
- intérieur des cratères (classes 3, 4, 7, 10),
- crêtes (classes 2 et 6),
- chenaux (classe 9).

classe	signification
1	Plaines intra-cratère à altitude moyenne
5	Plaines intra-cratère à altitude élevée
8	Plaines intra-cratère à altitude basse
11	Plaines intra-cratère à altitude moyenne-basse
3	Terrain intra-cratère assez profond sous la crête
4	Terrain intra-cratère très profond sous la crête
7	Terrain intra-cratère juste sous la crête
10	Terrain intra-cratère intermédiaire entre classes 3 et 4
12	Terrain intra-cratère sur un haut fond partiellement enseveli
2	Terrain dominant une vallée
6	Terrain situé juste sous le bord extérieur du cratère
9	chenal

Tableau 7.2 : principales formes de la vallée Tisia sur Mars *(source : R.Vilalta & T.F.Stepinski)*

Cette classification ascendante sur l'échantillon représentatif a donc joué le rôle de méthode d'*apprentissage* des formes de terrain en jouant à la fois sur des variables propres à chaque pixel (variables 1,3, 5) et sur des variables simulées, géographiquement contextuelles (variables 2, 4, 6). Un arbre de décision a ensuite permis, sur la base de leurs 6 valeurs, de répartir au mieux les 121 626 autres pixels aux 12 classes précédemment crées.

La figure 7.6B représente la carte géomorphologique issue de ce traitement de l'information. La comparaison des figures 7.6A et B montre une bonne discrimination des formes, ce qui permettra de remonter aux processus géomorphologiques producteurs.

A B

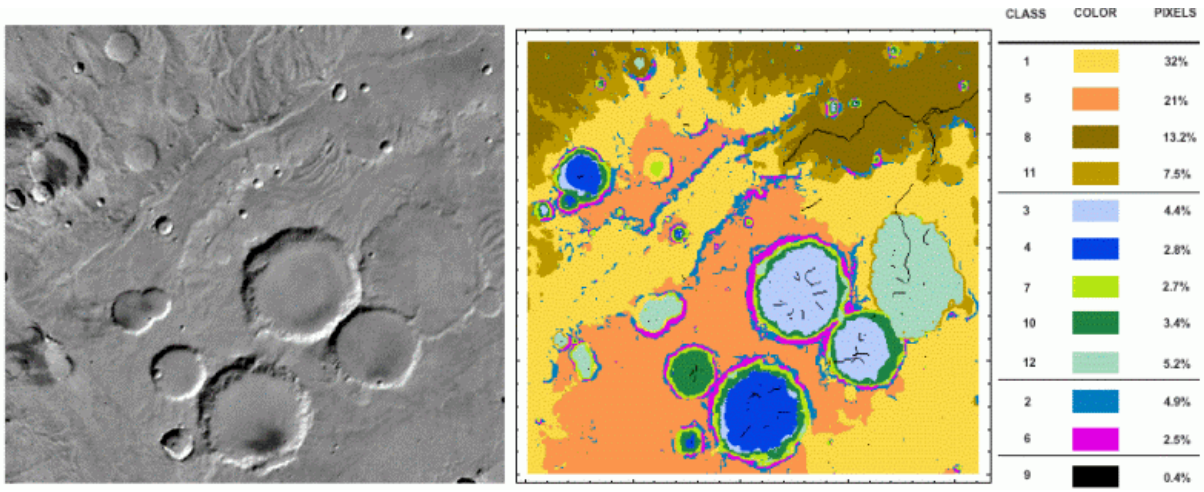


Figure 7.6 : carte géomorphologique de la vallée Tisia sur Mars *(source : R.Vilalta & T.F.Stepinski)*

L'étude, outre son intérêt thématique d'actualité, présente nombre d'intérêts méthodologiques (traitement d'une grande masse de données, utilisation d'un échantillon d'apprentissage, combinaison de classifications ascendante puis descendante, création d'une carte support d'un raisonnement inductif en géomorphologie, …).

Références citées dans ce chapitre

- M.Calciu, C.Bénavent : http://www.christophe.benavent.free.fr/cours/stat/AFD.PDF
- R.J.Wolfe, V.Fisher : « *Discriminating rural populations in Alaska* » (US Fish & Wildlife Service, 2003, rapport)
- P.Dumolard : « *Migrations et mobilité en bas Dauphiné* » (thèse, Université Lyon 2, 1983, 575 pages)
- R.Vilalta, T.F.Stepinski : http://www.geocities.com/js_source/tab04.html

A noter également

- M.Phipps : « *Espace, entropie, environnement* » (5ième rencontres Théo Quant, Besançon, 2001)
- L'existence d'un logiciel libre de segmentation (CTREE, add in Excel) téléchargeable à http://www.geocities.com/adotsaha/Ctree/

Bibliographie sommaire

- ***Prérequis de 1er cycle***

- Dumolard P., Dubus N., Charleux L. : *« Les statistiques en géographie »* (Belin, coll. Atouts, 2003)
- Minvielle E., Souiah S.A. : *« L'analyse statistique et spatiale »* (Ed. du Temps, 2003)
- Groupe Chadule : *« Initiation aux pratiques statistiques en géographie »* (A.Colin, coll. U, 1997, 4ième ed.)

- ***En lien avec la cartographie thématique***

- Béguin M., Pumain D. : *« La représentation des données géographiques »* (A.Colin, 2000)
- Godard V. : http://margaux.ipt.paris8.fr/vgodard/
- Grasland C. : http://www.grasland.cicrp.jussieu.fr/grasland/agreg/module4/

- ***Statistique multivariée géographique***

- Sanders L. : *« L'analyse statistique des données en géographie »* (Reclus modes d'emploi, 1990)
- Charre J. : *« Statistique et territoire »* (Reclus modes d'emploi, 1995)
- Cybergeo : http://www.cybergeo.presse.fr/

- ***Statistique multivariée***

- Lebart L., Morineau A., Piron M. : « *Statistique exploratoire multidimensionnelle »* (Dunod, 2002, 3ième ed.)
- Escofier B., Pages J. : *« Analyses factorielles simples et multiples »* (Dunod, 1998, 3ième ed.)
- Statsoft : http://www.statsoft.com/textbook/stathome.html/

LOGICIELS UTILISES

L'usage de la statistique multivariée ne s'est répandu dans toutes les disciplines qu'à partir de la large diffusion de l'informatique (et, spécialement, de la micro informatique). C'est assez dire que sa mise en œuvre implique l'utilisation d'un ou plusieurs logiciels. On a choisi, dans ce manuel, de n'utiliser que des logiciels commerciaux de coût abordable et de manipulation aisée (afin de ne pas ajouter de difficultés de manipulation informatique à celles des méthodes statistiques multivariées).

On a donc utilisé :

- ***EXCEL***

Pour sa gestion simple de données numériques tabulées, ses quelques possibilités graphiques et statistiques et, surtout, le fait qu'il est présent sur à peu près tous les micro ordinateurs. En outre, pratiquement tous les logiciels statistiques importent et exportent des données au format Excel (ce qui en fait un choix « d'entrée de gamme).

- ***XLSTAT***

Un des grands intérêts de ce logiciel est que c'est un « add in » d'Excel (une fois installé, il lui est intégré), qu'il est d'usage simple et qu'il contient la très grande majorité des méthodes exposées dans ce manuel : analyses factorielles, classifications (convergence, CAH), régression multiple, analyse discriminante, quelques possibilités graphiques, …

La version utilisée (4.4) est un peu ancienne mais les nouvelles versions (7.5) coûtent beaucoup plus cher et n'ajoutent pas de notables méthodes multivariées (hormis dans le domaine des « chroniques »).

- ***MINITAB***

Autre logiciel de manipulation aisée qui, outre les méthodes présentes dans XLSTAT, comporte aussi des facilités de gestion des données, de calculs simples, une rubrique « chroniques » et une analyse de variance multiple.

On a, exceptionnellement, utilisé d'autres logiciels entre autres :

- *Idrisi kilimandjaro* (S.I.G. raster comportant notamment des fonctions de statistique spatiale et d'analyse d'images)

- *Statistica* (logiciel assez complexe mais produisant des graphiques de grande qualité)

On terminera en signalant un site répertoriant un certain nombre de logiciels statistiques téléchargeables (d'usage gratuit pour une période limitée ou « ad aeternam »:

http://freestatistics.altervista.org/stat.php/

Exemples et exercices traités

Chapitre 2 : Analyse en Composantes Principales

- 7 variables démographiques et sociales pour 27 pays européens
- 5 variables démographiques pour 15 pays ouest africains
- Croîts naturel et migratoire (1954-99) de 14 départements du S.E. français

Chapitre 3 : Analyse des Correspondances

- Nationalités par taille de villes françaises de plus de 50 000 habitants en 1999
- Pentes, altitudes et utilisations du sol du département de la Savoie
- Structure d'âge des logements des régions françaises en 1999
- Principaux usages de l'eau dans les départements français atlantiques en 2003

Chapitre 4 : Analyse des Correspondances Multiples

- Extrait d'une enquête préalable à la construction d'un tronçon autoroutier
- Enquête d'opinion sur les dépenses publiques aux U.S.A.
- 5 indicateurs de politique environnementale de 34 villes françaises

Chapitre 5 : Classifications

- Types de contrats de travail de la population active par région française
- Précipitations moyennes mensuelles de 12 villes guinéennes
- 8 indicateurs de développement humain de 25 pays européens
- recolonisation par le chêne pubescent d'un adret calcaire de Chartreuse

Chapitre 6 : Régression multiple

- Températures moyennes annuelles de 6 villes du fossé rhénan et abord vosgien
- Relation espérance de vie – PNB/hb des pays d'Europe de l'ouest ou du centre
- Explication des températures moyennes de janvier d'un échantillon de 33 villes US
- Relation PIB/hb – types de contrat de travail (régions françaises, 1999)

Chapitre 7 : **Analyse discriminante et segmentation**

- Discrimination des populations rurales et urbaines en Alaska
- Segmentation d'une population par rapport aux facteurs de sa mobilité résidentielle
- La population féminine suisse : les compromis de carrières familiale et professionnelle
- Etablissement d'une carte géomorphologique de la vallée Tisia (planète Mars)