UNIVERSITE DE PAU ET DES PAYS DE L'ADOUR

UFR : Droit et Sciences Economiques

Master Economie Appliquée

Parcours : Economie Appliquée

Projet Analyse des données

Analyse multivariée de la base de données du groupe d'étude et de réflexion inter-régional de 1990

Présenté par :
Boubacar KANDE

Année Universitaire : 2022/2023

Table des matières

1 Introduction

Avec le développement de plus en plus de la microéconomie, l'économie géographique et les outils informatiques ; plusieurs chercheurs font recours aux données microéconomiques. Ces outils permettent de mieux caractérisés les régions, les départements, les quartiers, les entreprises etc.

Ces méthodes permettent d'observer la situation socio-économique des zones géographiques. Elles fournissent une source d'information afin de mieux orienter les politiques publiques sectorielles.

Ainsi, les statistiques montrent l'existence d'une hétérogénéité forte dans les régions et les départements Françaises sur le plan économique et sociale. Selon les données de l'Insee 2020, la part de l'industrie dans la valeur ajoutée est beaucoup importante aux régions du nord-Est et au centre nord de la France (18,7% pour Grand Est, 18,3% pour Centre-Val de loire, 17,9% pour Auvergne-Rhones-Alpes et 17,7% Pour la région Bourgogne-Franche-comté). Par contre le tertiaire est plus développé dans les régions du Sud-ouest. Cette hétérogénéité apparait également sur l'indice de vieillissement de la population. Il y'a plus de jeune dans les localités comme Lille, Paris, Lyon et Nante.

Ainsi, nous posons la questions de savoir quelles sont les spécificités de chaque département et région de la France ?

Répondre à cette interrogation permettra aux décideurs de politiques publique de mieux orientés ces politiques afin de réduire certains discriminions ou inégalités.

L'objectif de notre travail est faire une analyse multivariée sur la base de données provenant du Groupe d'Etude et de Réflexion Inter-Regional en 1990 pour caractériser les départements et régions Françaises. Pour atteindre cet objectif, notre travail sera articulé autour de deux grandes sections : (1) données et méthodes, (2) résultats et discussions.

2 Méthodologie

2.1 Données

Notre analyse porte sur des données provenant du groupe d'étude et de réflexion inter-régional (GERI) et décrit quatre grands thèmes à savoir : la démographie, l'emploi, la fiscalité directe locale, la criminalité de chacun des départements français métropolitains et de la corse pendant l'année 1990. Les indicateurs sont calculés relativement à la population totale du département concerné. On a dans la base de données 95 observations représentant les différents départements et régions administratives ; Deux variables qualitatives et 15 variables quantitatives. Nous définissons les variables de la base de données.

- ☞ **depart** : code du département
- ☞ **region** code de la region

- ☞ **TXCR** : taux de croissance de la population sur la période 1882-1990
- ☞ **EXTRA** : part des étrangers dans la population totale
- ☞ **URBR** : indicateur de concentration de la population mesurant le caractère urbain ou rural d'un département
- ☞ **JEUN** : part des 0-19 ans dans la population totale
- ☞ **AGE** : part des plus de 65 ans dans la population totale
- ☞ **CHOM** : taux de chômage
- ☞ Parts de chaque catégorie socioprofessionnelle dans la population active occupée du département :
  - **AGRI** : agriculteurs
  - **ARTI** : artisans
  - **CADR** : cadres supérieurs
  - **EMPL** : employés
  - **OUVR** : ouvriers
  - **PROF** : professions intermédiaires
- ☞ **FISC** : produit, en francs constants 1990 et par habitant des quatre taxes directes locales (professionnelle, habitation, foncier bâti, foncier non bâti).
- ☞ **CRIM** : taux de criminalité (nombre de délits par habitant)
- ☞ **FE90** : taux de fécondité (pour 1000) égal au nombre de naissances rapportés au nombre de femmes âgées de 15 à 49 ans en moyenne triennale

Le tableau 1 présente les résultats de la statistique descriptive obtenus à partir des observations de notre base de données ; Il ressort que le taux de croissance de la population française sur la période 1882-1990 est en moyenne de 3,758%. La part des étrangers dans la population totale française pendant l'année 1990 est en moyenne de 5.1% ; L'indicateur de concentration de la population mesurant le caractère urbain ou rural d'un département est en moyenne de 43.7% en 1990 ; La part des jeune de 0-19 ans dans la population totale est en moyenne de 25.9% en 1990 ; La part des plus de 65 ans dans la population totale française est de 16.3% en moyenne en 1990 ; Le taux de chômage dans la population française est en moyenne de 11.1% en 1990 ; Les agriculteurs représentent en moyenne 7% de la population active française en 1990 ; Les artisans représentent en moyenne 8.6% de la population française en 1990 ; Les cadres supérieurs représentent en moyenne 9.2% de la population française en 1990 ; Les employés représentent en moyenne 25.6% de la

population française en 1990 ; Les ouvriers représentent en moyenne 30.9% de la population française en 1990 ; Les professions intermédiaires représentent en moyenne 18.7% de la population française en 1990 ; Les taxes directes locales (professionnelle, habitation, foncier bâti, foncier non bâti) par habitant en 1990 sont en moyenne de 3,110.259 francs ; Le taux de criminalité en 1990 est en moyenne de 52.057% ; Le taux de fécondité (pour 1000) égal au nombre de naissances rapportés au nombre de femmes âgées de 15 à 49 ans en moyenne triennale s'élève en moyenne de 50.698.

TABLEAU 1 – Statistique descriptives

Statistic	N	Moyenne	Ecart type	Min	Max
txcr	95	3.758	4.910	−5.730	21.870
extra	95	0.051	0.033	0.006	0.189
urbr	95	0.437	0.232	0.000	1.001
jeun	95	0.259	0.027	0.186	0.312
age	95	0.163	0.035	0.088	0.254
chom	95	0.111	0.025	0.063	0.173
agri	95	0.070	0.050	0.000	0.222
arti	95	0.086	0.019	0.051	0.137
cadr	95	0.092	0.040	0.052	0.321
empl	95	0.256	0.024	0.212	0.333
ouvr	95	0.309	0.056	0.134	0.412
prof	95	0.187	0.025	0.144	0.250
fisc	95	3,110.259	535.582	2,216.900	5,029.700
crim	95	52.057	21.095	24.600	139.900
fe90	95	50.698	4.763	39.500	64.400

Le tableau 2 présente l'ensemble des 22 régions que composaient la France lors de l'étude en 1990 ; Ainsi, les régions telles que Rhône-Alpes, Midi-Pyrénées, Ile-de-France sont celles avec le plus grand nombre de département (8 départements) et dont les plus peuplé et les plus développées ; Suivies des régions telles que Provence-Alpes-Côte d'azur, Centre (6 départements chacune) ; Aquitaine, Languedoc-Roussillon, Pays de la Loire (5 départements chacune) ; Ensuite nous avons des régions telles que Pointou-Charentes,

Lorraine, Franche-Comté, Champagne-Ardenne, Bretagne, Bourgogne, Auvergne (4 départements chacune) ; Basse-Normandie, Limousin, Picardie (3 départements chacune) ; Enfin les régions telles que Nord-Pas-de-Calais, Haute Normandie, Alsace (2 départements chacune) et celle de Corse qui est la plus petite des régions avec 1 seul département.

TABLEAU 2 – Nombre de départements par région dans la base de données

Characteristic	N = 95[1]
region	
Alsace	2
Aquitaine	5
Auvergne	4
Basse-Normandie	3
Bourgogne	4
Bretagne	4
Centre	6
Champagne-Ardenne	4
Corse	1
Franche-Comté	4
Haute-Normandie	2
Ile-de-France	8
Languedoc-Roussillon	5
Limousin	3
Lorraine	4
Midi-Pyrénées	8
Nord-Pas-de-Calais	2
Pays de la loire	5
Picardie	3
Pointou-Charentes	4
Provence-Alpes-Côte d'azur	6
Rhône-Alpes	8

2.2 Méthodes

Pour répondre à notre problématique, nous optons les outils d'analyse multivariée. Nous travaillons sur des données provenant du Groupe d'Etude et de Réflexion Inter-Régional en 1990. La base de données est composée de 95 départements de la France métropolitaines soit 22 régions. Ainsi le choix des méthodes d'analyse en économie dépend de la nature des données et les objectifs de l'étude.

La nature de nos données et nos objectifs nous amènent à choisir trois méthodes d'analyses à savoir l'analyse en composante principale (ACP), l'analyse factorielle discriminante (AFD) et l'analyse typologique autrement appelée méthode de classification (AT).
Ainsi, nous utilisons l'ACP pour caractériser les départements, l'AFD pour créer des groupes de régions homogènes et voir leurs spécificités, et l'AT pour créer des classes de départements et voir les spécificités de chaque classes.

Nous visons plusieurs caractéristiques d'un département notamment : la démographie,

l'emploi, la fiscalité directe locale, la criminalité.

NB : nous avons utiliser le logiciel SAS pour recoder nos données et ensuite R pour effectuer l'analyse (codes voir annexes)

3 Présentation des résultats

3.1 Analyse en composante principale (ACP)

3.1.1 Etude des corrélations entre variables quantitatives

Le tableau 3 est celui des corrélations entre les différentes variables quantitatives de notre base de données lors de l'étude en 1990.

La part des étrangers dans la population totale est fortement corrélé positivement avec l'indicateur de concentration de la population mesurant le caractère urbain ou rural d'un département.

La part des plus de 65 ans dans la population totale est corrélé négativement avec l'indicateur de concentration de la population mesurant le caractère urbain ou rural d'un département.

La part des plus de 65 ans dans la population totale est corrélé négativement avec la part des 0-19 ans dans la population totale.

La part des étrangers dans la population totale est négativement corrélé avec la par agriculteur dans la population active.

L'indicateur de concentration de la population mesurant le caractère urbain ou rural d'un département est négativement corrélé avec la part d'agriculteur dans la population active.

La part des plus de 65 ans dans la population est négativement corrélé avec la part d'agriculteur.

La part artisans est négativement corrélé avec la part des 0-19 ans dans la population totale et positivement corrélé avec la part des plus de 65 ans dans la population totale.

La part des cadres supérieurs dans la population active est positivement corrélé avec la part des étrangers dans la population totale et l'indicateur de concentration de la population mesurant le caractère urbain ou rural d'un département et est négativement corrélé avec la part des agriculteurs.

TABLEAU 3 – Etude des corrélations entre les variables

	txcr	extra	urbr	jeun	age	chom	agri	arti	cadr	empl	ouvr	prof	fisc	crim	fe90
txcr	1	0.31	0.31	0.32	-0.42	-0.07	-0.46	0.04	0.29	0.41	-0.21	0.49	0.41	0.43	0.34
extra	0.31	1	0.71	0.04	-0.46	-0.14	-0.66	-0.27	0.66	0.53	-0.31	0.64	0.5	0.65	0.45
urbr	0.31	0.71	1	0.18	-0.59	0.04	-0.77	-0.41	0.75	0.67	-0.35	0.79	0.6	0.71	0.44
jeun	0.32	0.04	0.18	1	-0.82	-0.17	-0.41	-0.65	-0.07	-0.04	0.52	0.3	-0.09	-0.12	0.75
age	-0.42	-0.46	-0.59	-0.82	1	0.27	0.71	0.75	-0.42	-0.26	-0.18	-0.69	-0.23	-0.24	-0.76
chom	-0.07	-0.14	0.04	-0.17	0.27	1	-0.04	0.36	-0.15	0.22	-0.03	-0.12	0.07	0.34	-0.07
agri	-0.46	-0.66	-0.77	-0.41	0.71	-0.04	1	0.48	-0.57	-0.57	-0.05	-0.78	-0.5	-0.62	-0.58
arti	0.04	-0.27	-0.41	-0.65	0.75	0.36	0.48	1	-0.29	-0.06	-0.36	-0.4	0.05	0.07	-0.55
cadr	0.29	0.66	0.75	-0.07	-0.42	-0.15	-0.57	-0.29	1	0.48	-0.63	0.73	0.51	0.69	0.25
empl	0.41	0.53	0.67	-0.04	-0.26	0.22	-0.57	-0.06	0.48	1	-0.5	0.6	0.48	0.63	0.24
ouvr	-0.21	-0.31	-0.35	0.52	-0.18	-0.03	-0.05	-0.36	-0.63	-0.5	1	-0.38	-0.4	-0.5	0.23
prof	0.49	0.64	0.79	0.3	-0.69	-0.12	-0.78	-0.4	0.73	0.6	-0.38	1	0.58	0.59	0.45
fisc	0.41	0.5	0.6	-0.09	-0.23	0.07	-0.5	0.05	0.51	0.48	-0.4	0.58	1	0.64	0.21
crim	0.43	0.65	0.71	-0.12	-0.24	0.34	-0.62	0.07	0.69	0.63	-0.5	0.59	0.64	1	0.31
fe90	0.34	0.45	0.44	0.75	-0.76	-0.07	-0.58	-0.55	0.25	0.24	0.23	0.45	0.21	0.31	1

3.1.2 Choix du nombre d'axes factoriels optimal

Pour le choix du nombre d'axes à retenir plusieurs méthodes sont à nos dispositions parmi lesquelles nous le critère de Kaiser sur les données normées. on ne retient que les axes dont l'inertie est supérieure à l'inertie moyenne I/p (I : inertie et p variables). Kaiser en ACP normée : I/p= 1 : On ne retiendra que les axes associés à des valeurs propre supérieures à 1. Au regard du tableau 4, nous retenons 3 axes factoriels que nous chercherons à interpréter par la suite. En générale, dans la pratique, on ne retient que les q axes qui fournit près de 80% de l'inertie total (loi de Pareto) ou encore les axes que l'on peut interpréter. Notre cas les trois premiers axes expliquent 75.902% de l'informations.

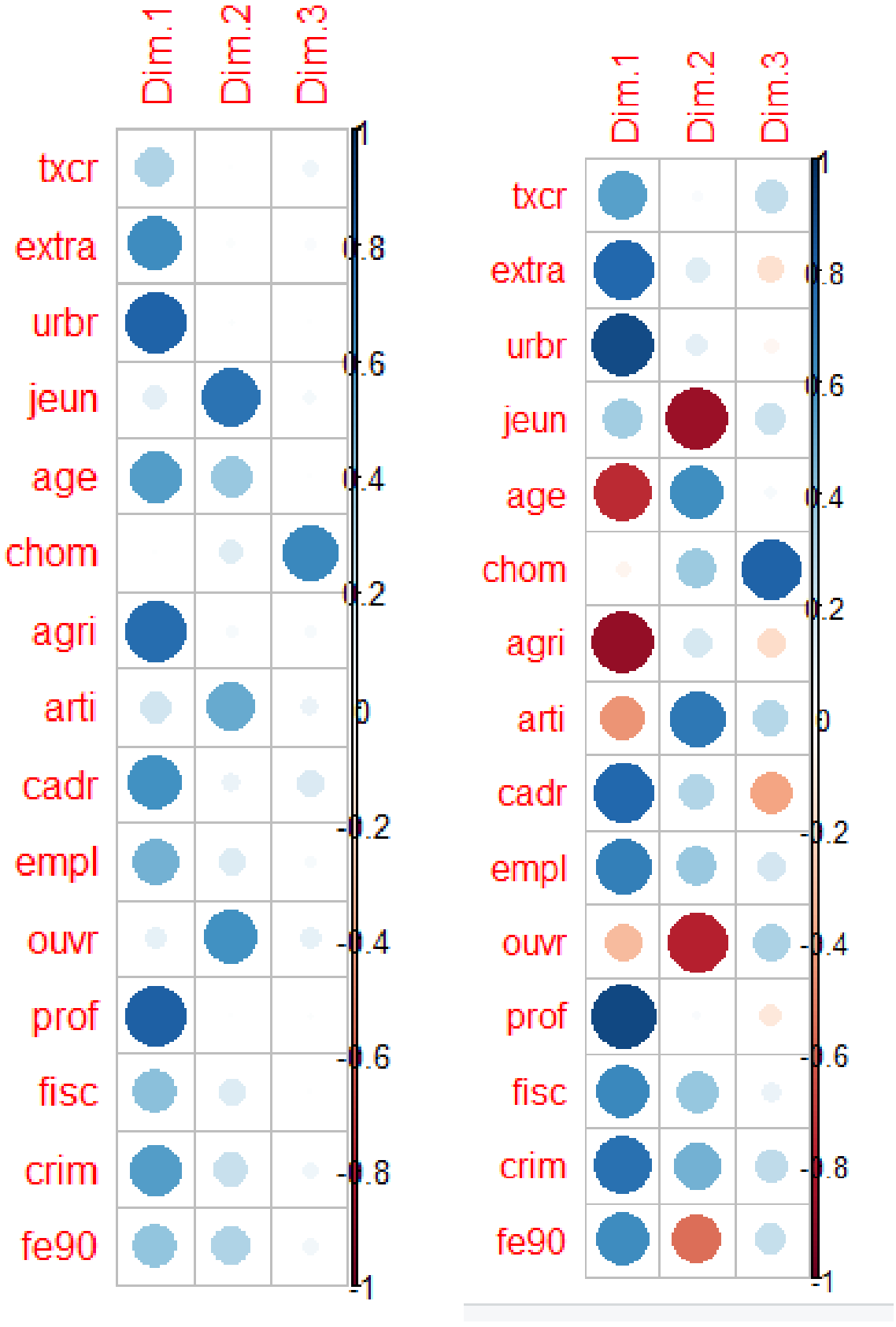
TABLEAU 4 – Valeurs propres de la diagonalisation de la matrice de corrélation

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Valeurs propres	6.724	3.333	1.329	0.985	0.599	0.494	0.461
Variance en %	44.824	22.220	8.857	6.568	3.994	3.294	3.070
Variance Cumulé en %	44.824	67.045	75.902	82.470	86.465	89.758	92.828

3.1.3 Définitions des axes factoriels

Pour définir les axes nous observons la figure 1a et pour savoir leurs positions 1b.

FIGURE 1 – Représentation des cos^2 et cor entre les variables et les axes choisis



(a) Cosinus carré variables/axes

(b) Corrélation variables/axes

Axe1 est caractérisé par les variables : extra, urbr, age, agri, cadr, empl, prof et crim
Axe2 est caractérisé par les variables : Jeun, arti et ouvr
Axe3 est caractérisé par la varible : chom.

L'axe 1 oppose les variables age, agri et les variables extra, urbr, cadr, empl, prof, crim.

L'axe 2 oppose les variables jeun, ouvr et la variable arti.

L'axe 3 est défini par une variable qui est chom.

TABLEAU 5 – Caractéristiques des axes

Négatifs	**Positifs**
Axe 1	
- age - agri	- extra - urbr - cadr - empl - prof - crim
Axe 2	
- jeun - ouvr	- arti
Axe 3	
	- chom

3.1.4 Interprétations des résulats de l'ACP

Les 2 premiers axes de l' analyse expriment **67.04%** de l'inertie totale du jeu de données ; cela signifie que 67.04% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage assez important.

☞ **Plan 1 , 2**

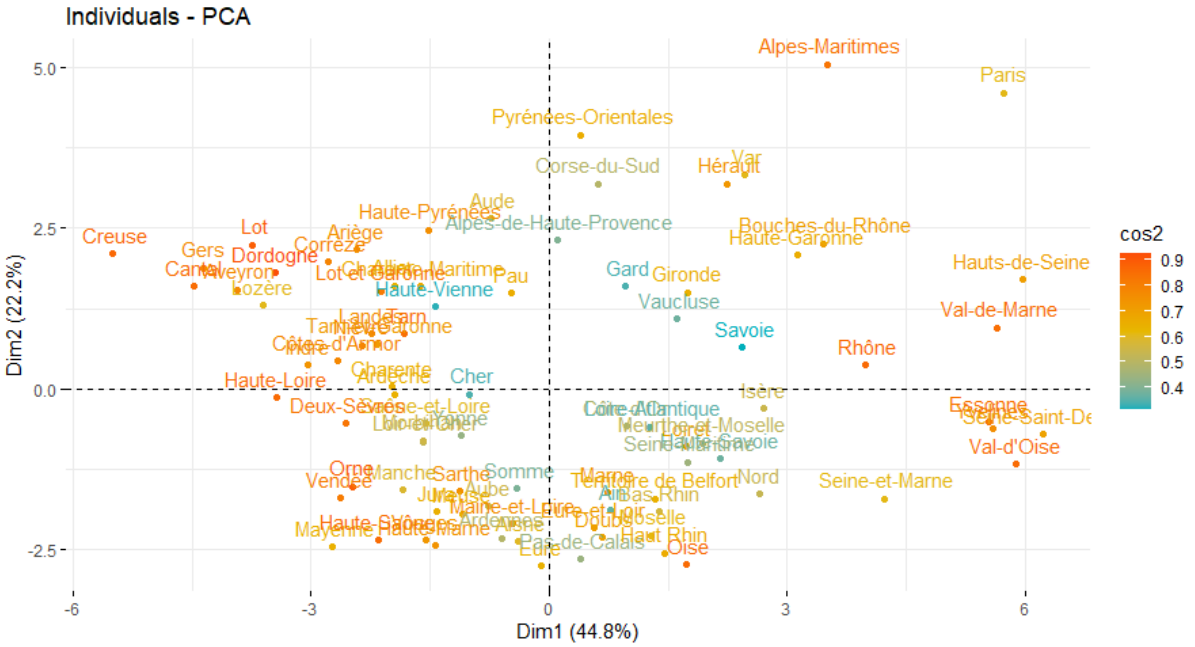
Les départements comme Hauts-de-Seine, Val-de-Marne, Seine-Saint-Denis, val-d'Oise et Paris se sont des départements une part importante d'étrangers dans la population totale, plus de criminalité, une forte indice de concentration de la population et plus d'emplois qualifiés (cadre, profession et employés).

Les départements comme Creuse, Gers, Lot, Cantal et Dordogne se caractérisent par une vieillissement de la population et plus d'agriculteurs.

Les départements comme Alpes-Maritimes et Pyrénées-Orientales se caractérisent par le développement de l'artisanat.

Les départements comme Oise, Haute-Saone et Eure se caractérisent par une population jeune et la part très important d'ouvriers dans la population active.

FIGURE 2 – Représentation des individus sur le plan (1,2)



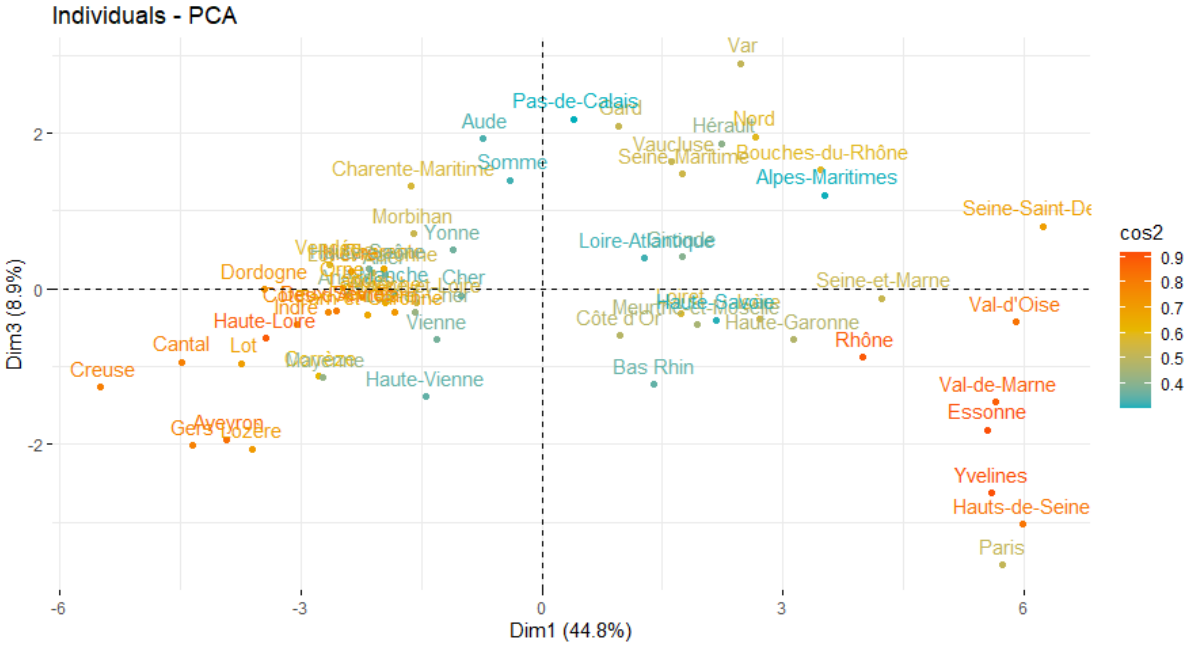
☞ **Plan 1 , 3**

Nous avions vu que l'axe 3 est défini par une seule variable qui est le taux de chômage.

Les départements comme Var, Nord, Charente Maritime et Gard se caractérisent un taux de chômage important.

Les départements comme Yvelines, Hauts-de-seine, Paris, Aveyron, Gers et Lozere se caractérisent par un faible taux de chômage.

FIGURE 3 – Représentation des individus sur le plan (1,3)

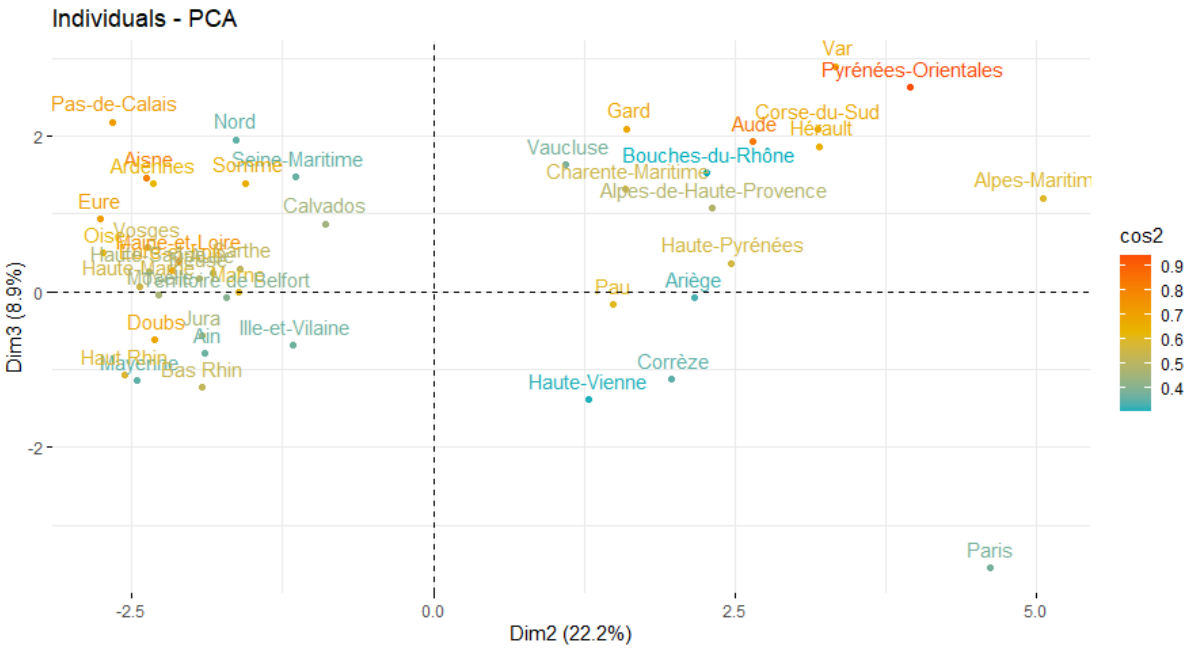


☞ **Plan 2 , 3**

Les départements comme Var, Pyrénées-orientales, corse du sud, Aude, Gard et Hérault se caractérisent par une part important d'artisan et un taux de chômage élevé.

Les départements comme Pas de calais, Aisne, Somme, Ardennes et Eure se caractérisent par une forte part d'ouvriers dans la population active, une population jeune et un taux de chômage très élevé.

FIGURE 4 – Représentation des individus sur le plan (2,3)



3.2 Analyse typologique (AT)

Dans le but de construire des classes pour les départements, nous adoptons la méthode de l'analyse typologique. Cette méthode est scinder deux sous méthodes telle que la méthode k-means et celle par Classification Ascendante Hiérarchique (CAH). Dans le cadre de notre travail, nous utilisons la méthode de classification ascendante hiérarchique car nus avons que 95 observations.

Avant tout, nous commençons par sélectionner le nombre de classe. Pour ce faire nous observons le graphique de l'inertie. Au regard du graphique 5, nous sélectionnons quatre classes.

FIGURE 5 – Perte d'inertie en fonction du nombre de classes

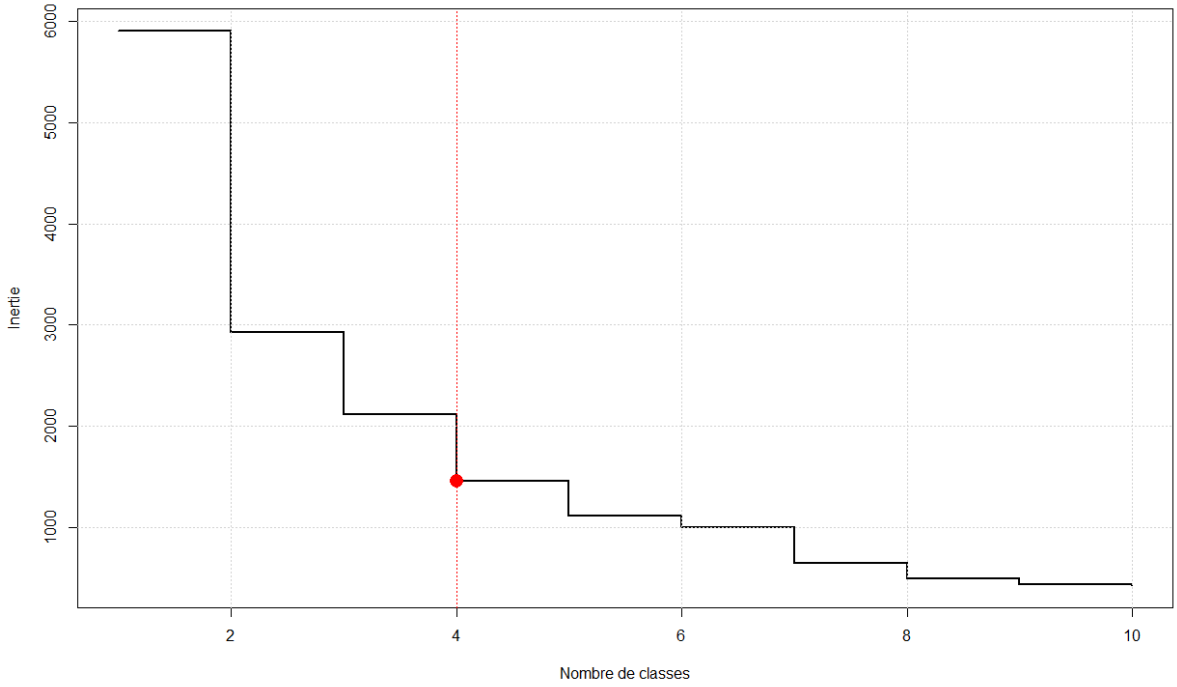
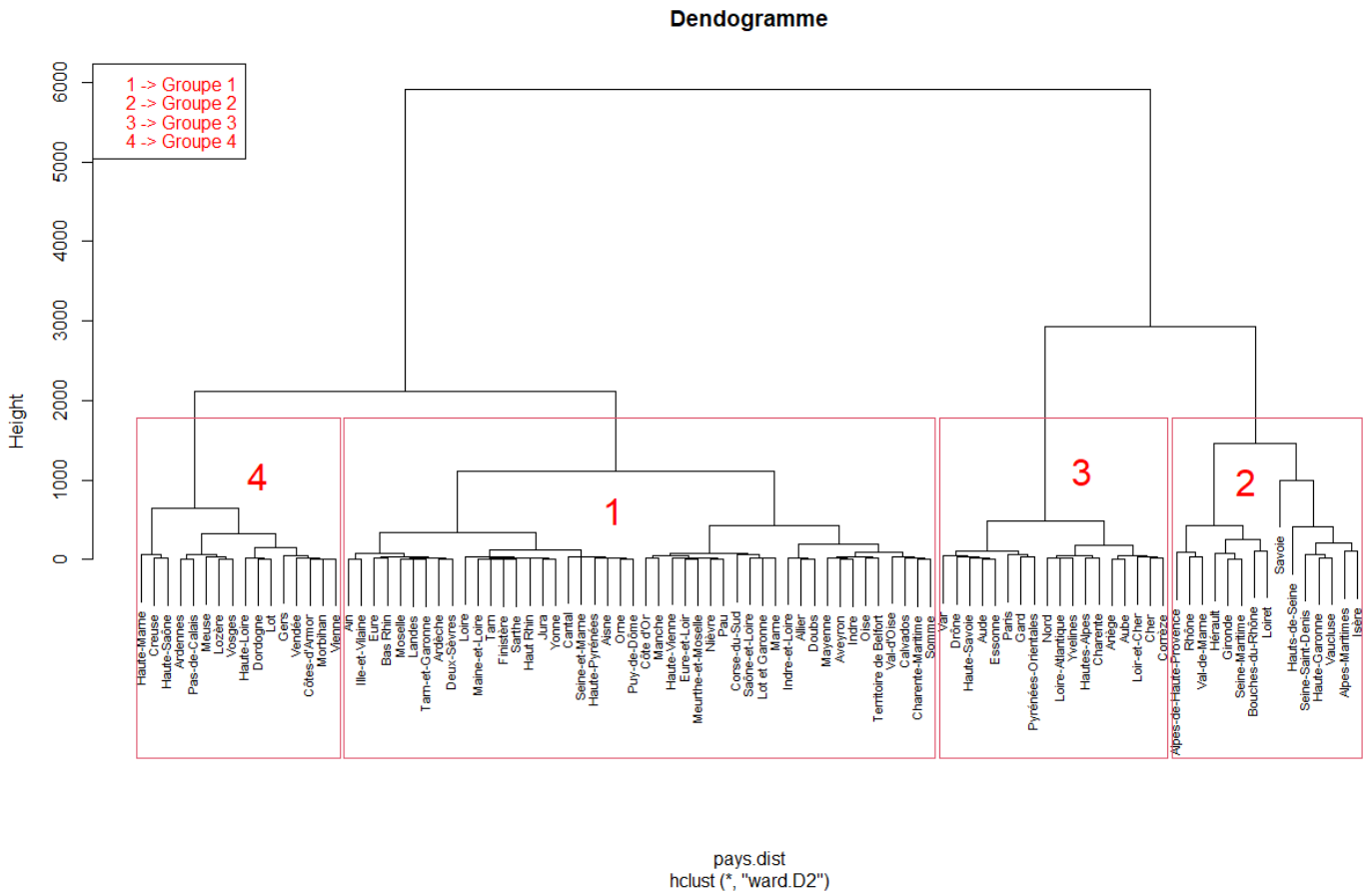


FIGURE 6 – Dendogramme des individus



Le graphique ci-dessus (graphe 6) nous montre les résultats de la classifications. Ensuite pour caractériser ces classes, nous observons le tableau 6. Ce tableau présente les moyennes de chaque de chaque variable de la base selon les groupes. Pour déterminer les spécificités de chaque classe, il suffit de faire une comparaison de leurs moyennes. Le classement des individus en groupes de cinq permet de mieux concevoir la ressemblance et la dissemblance entre les classes.

Le groupe 1 caractérise les départements qui ont une forte valeurs pour les variables comme : la part des jeunes dans la population totale, part des ouvriers dans la population active. Dans ces départements, il y a peu d'individus dans l'artisanat.

Le groupe 2 regroupe les départements qui se caractérisent par une forte démographie (taux de croissance de la population, part des étrangers, indice de concentration de la population, taux de fécondité), une forte part des emplois qualifiés (cadre,employés,professions), la fiscalité élevée et taux de criminalité important. Par contre il se caractérise par une faible part des personnes âgés, d'agriculteurs et d'ouvriers.

Le groupe 3 : se caractérise par un taux de chômage élevé, et plus d'artisans et moins de jeune.

Le groupe 4 : se caractérise par plus de personnes âgés, plus d'agriculteurs, un taux important d'artisans.

TABLEAU 6 – Caractéristique des groupes

Groupes	1	2	3	4
txcr	3.354	6.851	5.392	0.179
extra	0.045	0.082	0.062	0.028
urbr	0.398	0.694	0.502	0.233
jeun	0.263	0.255	0.252	0.257
age	0.160	0.147	0.166	0.181
chom	0.106	0.116	0.117	0.112
agri	0.074	0.027	0.053	0.116
arti	0.083	0.086	0.090	0.090
cadr	0.082	0.128	0.110	0.067
empl	0.252	0.277	0.262	0.239
ouvr	0.326	0.268	0.290	0.320
prof	0.182	0.214	0.194	0.168
fisc	2909.843	4091.887	3368.594	2475.550
crim	44.980	75.680	63.950	36.875
fe90	50.254	52.993	50.917	49.575

3.3 Analyse factorielle discriminant (AFD)

Cette méthode nous permettra de caractériser les régions selon leurs positions géographiques. Notre variable de regroupement est celle des quatre points cardinaux. Nous constituons ainsi quatre groupes régions : Nord, Sud, Est et Ouest .

Avant tout nous observons le coefficient de corrélation canonique au carré qui mesure la qualité de la discrimination. Dans notre cas il est de **92,88%** (tableau 7). Celui ci s'interprète comme le coefficient de détermination dans l'économétrie classique.

TABLEAU 7 – Coefficient de corrélation canonique

```
dis1$eig
] 0.9288171 0.6940942 0.2384922
```

Le choix du nombres d'axes optimaux se fait comme dans la méthode précédente (ACP). Nous observons les valeurs propres et nous choisissons les axes discriminants dont la valeur propre est supérieure à 1. Nous rappelons que nous travaillons avec les données normées. Le tableau 8 nous présente les valeurs propres des trois premiers axes discriminants. Au regard de ces résultats, nous choix les deux premiers axes (1 et 2) qui ont des valeurs propres supérieure à 1. Par la suite nous interprétons ces deux axes.

TABLEAU 8 – Valeurs propres

Axe 1	**Axe 2**	**Axe 3**
13.0483134	2.2689800	0.3131843

Nous définissons les deux axes discriminants sélectionnés en observant les résultats du tableau 9. Premier axe discriminant est défini par les variables : jeun, age, arti, ouvr et fe90. Deuxième axe est défini par une seule variable qui extra.

TABLEAU 9 – Définition des axes discriminants choisis

	CS1	**CS2**
txcr	0.39581505	-0.082204158
extra	0.05937577	**0.762352728**
urbr	-0.08386769	0.128988398
jeun	**-0.90827629**	-0.258952408
age	**0.80525052**	0.031233867
chom	0.25354612	-0.191380489
agri	0.41012109	-0.305907236
arti	**0.88253923**	-0.079726579
cadr	0.25894789	0.144219403
empl	0.41007787	0.071538057
ouvr	**-0.90262287**	0.126860461
prof	-0.08752931	0.233180499
fisc	0.41630159	-0.005010105
crim	0.34328920	0.069629436
fe90	**-0.77138020**	-0.179440316

☞ **Interprétations des résultats de l'AFD**

Pour interpréter ce figure ci-dessous, nous nous référons du tableau 9.

Les régions de l'Est se caractérisent par une forte part des jeunes dans la population totale, une forte taux de fécondité, une forte part d'ouvriers dans la population active et un taux important d'étrangers dans la population totale.

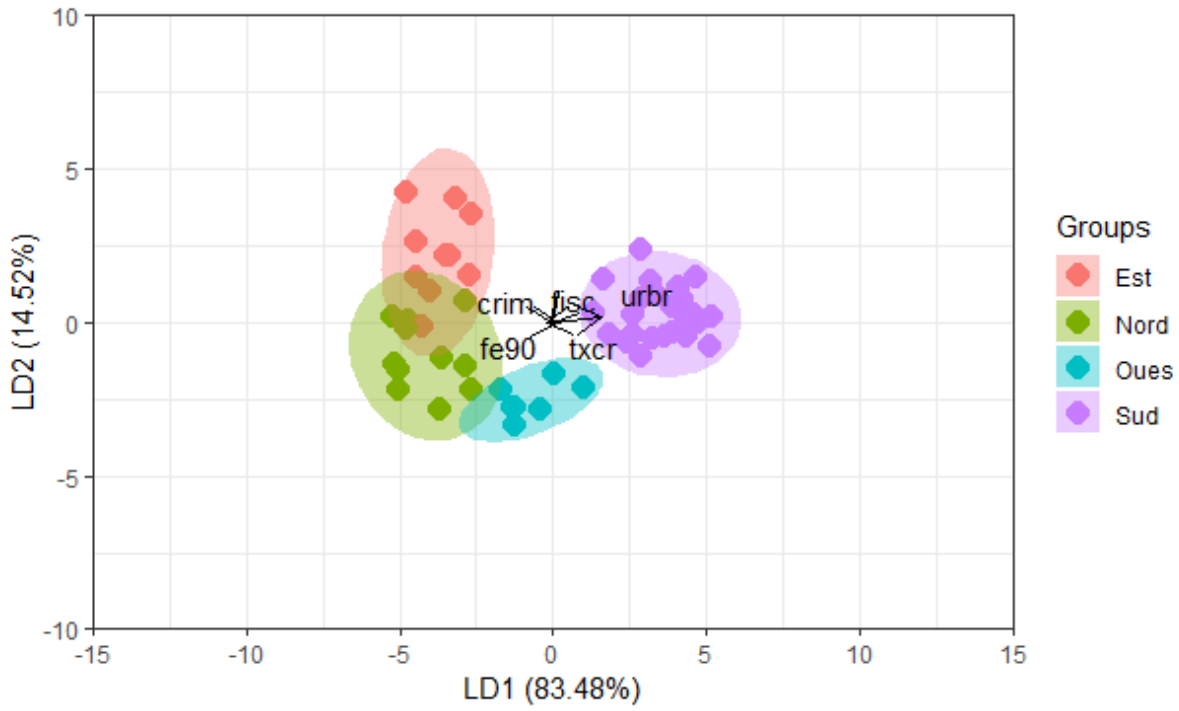
Les régions du Nord se caractérisent par un taux de fécondité élevé, part des ouvriers également importante et une population qui est jeune.

Les régions de L'Ouest se caractérisent par une faible part d'étrangers dans la population totale autrement dit il y a moins d'étrangers à l'Ouest.

Les régions du Sud se caractérisent une vieillissement de la population et une part importante d'individus qui s'activent dans les métiers d'artisans.

Ce modèle de discrimination prévoit exactement le regroupement que nous avons fait à priori (voir Annexe 10)

FIGURE 7 – Graphique de la discrimination des groupes



4 Conclusion

Notre objectif principal était de caractériser les départements et régions de la France métropolitain. Pour se faire, nous avons utiliser l'analyse multivariée à savoir l'analyse en composante principale, l'analyse factorielle discriminante et la méthode de classification ascendante hiérarchique.

Les résultats de l'analyse à composante principale et la classification montrent l'exis-

tence d'une forte hétérogénéité des départements. Mais néanmoins, il y a des groupes homogènes. Ainsi, la méthode CAH nous a permis de construire 4 classes de département. La classe 1 se caractérise par une population jeune mais peu d'emplois qualifiés ; la classe 2 par une forte démographie, d'emplois qualifiés et de l'insécurité ; la classe 3 un fort taux de chômage, emploi non qualifié et une part important de jeune ; et enfin la classe 4 vieillissement de la population et emplois non qualifiés.
L'analyse factorielle discriminant nous montre que la position géographique de la région est une bonne variable discriminante. Il y a une hétérogénéité selon que la région se trouve au l'Est, l'Ouest, Sud ou Nord.

Ces méthodes ont sans doute des limites car elles ne permettent pas de voir les interactions entre variables. Dans le futur, nous souhaiterons d'effectuer une étude économétrique afin de voir les relations entre variables.

5 Annexes

5.1 ACP

5.2 AFD

Tableau 10 – Prédiction de l'AFD

	Est	Nord	Oues	Sud	class	group	depart1	region region
1	0.99968	0.00032	0.00000	0.00000	Est	Est	Bas Rhin	Alsace
2	1.00000	0.00000	0.00000	0.00000	Est	Est	Haut Rhin	Alsace
3	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Dordogne	Aquitaine
4	0.00000	0.00000	0.00045	0.99955	Sud	Sud	Gironde	Aquitaine
5	0.00000	0.00000	0.00076	0.99924	Sud	Sud	Landes	Aquitaine
6	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Lot et Garonne	Aquitaine
7	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Pau	Aquitaine
8	0.00000	0.00105	0.99895	0.00000	Oues	Oues	Calvados	Basse-Normandie
9	0.00000	0.00043	0.99957	0.00000	Oues	Oues	Manche	Basse-Normandie
10	0.00002	0.04162	0.95837	0.00000	Oues	Oues	Orne	Basse-Normandie
11	0.00000	0.00342	0.99658	0.00000	Oues	Oues	Ille-et-Vilaine	Bretagne
12	0.00000	0.00000	0.98176	0.01824	Oues	Oues	Côtes-d'Armor	Bretagne
13	0.00000	0.00003	0.99928	0.00069	Oues	Oues	Finistère	Bretagne
14	0.00000	0.00015	0.99983	0.00001	Oues	Oues	Morbihan	Bretagne
15	0.02535	0.97465	0.00000	0.00000	Nord	Nord	Ardennes	Champagne-Ardenne
16	0.36651	0.63329	0.00020	0.00000	Nord	Nord	Aube	Champagne-Ardenne
17	0.21750	0.78250	0.00000	0.00000	Nord	Nord	Haute-Marne	Champagne-Ardenne
18	0.00163	0.95041	0.04797	0.00000	Nord	Nord	Marne	Champagne-Ardenne
19	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Corse-du-Sud	Corse
20	0.99629	0.00371	0.00000	0.00000	Est	Est	Doubs	Franche-Comté
21	0.84186	0.15814	0.00001	0.00000	Est	Est	Haute-Saône	Franche-Comté

TABLEAU 11 – Prédiction de l'AFD (suite)

	Est	Nord	Oues	Sud	class	group	depart1	region region
22	0.99696	0.00304	0.00000	0.00000	Est	Est	Jura	Franche-Comté
23	0.99999	0.00001	0.00000	0.00000	Est	Est	Territoire de Belfort	Franche-Comté
24	0.00006	0.89111	0.10883	0.00000	Nord	Nord	Seine-Maritime	Haute-Normandie
25	0.01068	0.97169	0.01763	0.00000	Nord	Nord	Eure	Haute-Normandie
26	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Aude	Languedoc-Roussillon
27	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Gard	Languedoc-Roussillon
28	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Hérault	Languedoc-Roussillon
29	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Pyrénées-Orientales	Languedoc-Roussillon
30	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Lozère	Languedoc-Roussillon
31	1.00000	0.00000	0.00000	0.00000	Est	Est	Moselle	Lorraine
32	0.98836	0.01159	0.00005	0.00000	Est	Est	Meurthe-et-Moselle	Lorraine
33	0.22554	0.77412	0.00034	0.00000	Nord	Est	Meuse	Lorraine
34	0.90550	0.09450	0.00000	0.00000	Est	Est	Vosges	Lorraine
35	0.00000	0.00000	0.00001	0.99999	Sud	Sud	Aveyron	Midi-Pyrénées
36	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Haute-Pyrénées	Midi-Pyrénées
37	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Lot	Midi-Pyrénées
38	0.00000	0.00000	0.00008	0.99992	Sud	Sud	Tarn	Midi-Pyrénées
39	0.00000	0.00000	0.00002	0.99998	Sud	Sud	Tarn-et-Garonne	Midi-Pyrénées
40	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Ariège	Midi-Pyrénées
41	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Gers	Midi-Pyrénées
42	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Haute-Garonne	Midi-Pyrénées

TABLEAU 12 – Prédiction de l'AFD (suite)

	Est	Nord	Oues	Sud	class	group	depart1	region region
43	0.00002	0.99995	0.00003	0.00000	Nord	Nord	Pas-de-Calais	Nord-Pas-de-Calais
44	0.00006	0.99994	0.00000	0.00000	Nord	Nord	Nord	Nord-Pas-de-Calais
45	0.00020	0.99980	0.00000	0.00000	Nord	Nord	Aisne	Picardie
46	0.02546	0.97454	0.00000	0.00000	Nord	Nord	Oise	Picardie
47	0.00000	0.99777	0.00223	0.00000	Nord	Nord	Somme	Picardie
48	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Alpes-Maritimes	Provence-Alpes-Côte d'azu
49	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Alpes-de-Haute-Provence	Provence-Alpes-Côte d'azu
50	0.00000	0.00000	0.00245	0.99755	Sud	Sud	Bouches-du-Rhône	Provence-Alpes-Côte d'azu
51	0.00000	0.00000	0.00357	0.99643	Sud	Sud	Hautes-Alpes	Provence-Alpes-Côte d'azu
52	0.00000	0.00000	0.00000	1.00000	Sud	Sud	Var	Provence-Alpes-Côte d'azu
53	0.00000	0.00000	0.00030	0.99970	Sud	Sud	Vaucluse	Provence-Alpes-Côte d'azu

5.3 AT

TABLEAU 13 – Les individus de la classe 1

```
> dep1$depart1[groups.3 == 1]
 [1] " Ain"                 " Aisne"               "Ardennes"
 [4] "Aube"                 "Calvados"             "Côte d'Or"
 [7] " Doubs"               "Eure"                 "Eure-et-Loir"
[10] " Ille-et-Vilaine"     "Indre-et-Loire"       "Jura"
[13] " Loir-et-Cher"        " Loire"               " Loire-Atlantique"
[16] "Loiret"               "Maine-et-Loire"       "Manche"
[19] "Marne"                "Haute-Marne"          " Mayenne"
[22] "Meurthe-et-Moselle"   "Meuse"                " Moselle"
[25] "Nord"                 " Oise"                "Orne"
[28] " Pas-de-Calais"       "Bas Rhin"             "Haut Rhin"
[31] "Haute-Saône"          " Sarthe"              "Haute-Savoie"
[34] " Seine-Maritime"      " Somme"               "Vendée"
[37] "Vosges"               "Yonne"                "Territoire de Belfort"
>
```

TABLEAU 14 – Les individus de la classe 2

```
> dep1$depart1[groups.3 == 2]
 [1] "Allier"              "Ardèche"             "Ariège"
 [4] " Aveyron"            "Cantal"              "Charente"
 [7] " Charente-Maritime" " Cher"               " Corrèze"
[10] "Côtes-d'Armor"       " Creuse"             "Dordogne"
[13] "Finistère"           "Gers"                "Indre"
[16] "Landes"              "Haute-Loire"         " Lot"
[19] "Lot et Garonne"      "Lozère"              "Morbihan"
[22] "Nièvre"              "Puy-de-Dôme"         "Pau"
[25] " Haute-Pyrénées"     "Saône-et-Loire"      " Deux-Sèvres"
[28] " Tarn"               " Tarn-et-Garonne"    "Vienne"
[31] "Haute-Vienne"
>
```

TABLEAU 15 – Les individus de la classe 3

```
> dep1$depart1[groups.3 == 3]
 [1] " Alpes-de-Haute-Provence" " Hautes-Alpes"
 [3] " Alpes-Maritimes"         " Aude"
 [5] " Bouches-du-Rhône"        "Corse-du-Sud"
 [7] " Drôme"                   " Gard"
 [9] "Haute-Garonne"            "Gironde"
[11] " Hérault"                 "Isère"
[13] " Pyrénées-Orientales"     " Savoie"
[15] " Var"                     " Vaucluse"
>
```

TABLEAU 16 – Les individus de la classe 4

```
> dep1$depart1[groups.3 == 4]
[1] " Rhône"           "Paris"            " Seine-et-Marne"   "Yvelines"
[5] "Essonne"          "Hauts-de-Seine"   "Seine-Saint-Denis" " Val-de-Marne"
[9] "Val-d'Oise"
```

5.4 Les programmes : SAS et R

Code 1: Recodage des données avec SAS

```
1
2   libname bk "/home/u43748914/UPPA";
3   FILENAME REFFILE '/home/u43748914/UPPA/departement.xlsx';
4
5   PROC IMPORT DATAFILE=REFFILE
6   DBMS=XLSX
7   OUT=WORK.IMPORT;
8   GETNAMES=YES;
9   RUN;
10  data bk.dep1;
11  set import;
12  run;
13
14  proc sort data=bk.dep1;
15  by  region;
16  run;
17
18  data BK.DEP1;
19  length depart1 $ 60;
20  set BK.DEP1;
21
22  select (depart);
23  when ('Rb') depart1='Bas Rhin';
24  when ('Rh') depart1='Haut Rhin';
25  when ('Dd') depart1='Dordogne';
26  when ('Gi') depart1='Gironde';
27  when ('La') depart1='Landes';
28  when ('Lg') depart1='Lot et Garonne';
29  when ('Al') depart1='Allier';
30  when ('Cl') depart1='Cantal';
31  when ('Lh') depart1='Haute-Loire';
32  when ('Pd') depart1='Puy-de-D me';
33  when ('Mc') depart1='Manche';
34  when ('Or') depart1='Orne';
35
36  when ('Ni') depart1='Ni vre';
37  when ('Sl') depart1='Sa ne-et-Loire ';
38  when ('Yo') depart1='Yonne ';
39
40  when ('Fi') depart1='Finist re';
41  when ('Iv') depart1=' Ille-et-Vilaine';
42
43  when ('Ce') depart1=' Cher';
44  when ('El') depart1='Eure-et-Loir ';
45  when ('In') depart1='Indre ';
46  when ('Il') depart1='Indre-et-Loire ';
47  when ('Lc') depart1=' Loir-et-Cher';
48  when ('Lr') depart1='Loiret ';
49  when ('Ab') depart1='Aube ';
50  when ('Ad') depart1='Ardennes ';
51  when ('Ma') depart1='Marne ';
52
53  when ('Cs') depart1='Corse-du-Sud ';
54  when ('Db') depart1=' Doubs';
55  when ('Ju') depart1='Jura ';
56
57  when ('TB') depart1='Territoire de Belfort ';
```

```
  when ('Eu') depart1='Eure ';
  when ('Sm') depart1=' Seine-Maritime';
  when ('Es') depart1='Essonne ';
  when ('HS') depart1='Hauts-de-Seine ';
  *when ('Pa') depart1='Paris ';
  when ('SM') depart1=' Seine-et-Marne';
  when ('SS') depart1='Seine-Saint-Denis ';
  when ('VM') depart1=' Val-de-Marne';
  when ('VO') depart1="Val-d'Oise";
  when ('Yv') depart1='Yvelines ';
  when ('AD') depart1=' Aude';
  when ('Ga') depart1=' Gard';
  when ('He') depart1=' H rault';
  when ('Lz') depart1='Loz re ';
  when ('Po') depart1=' Pyr n es-Orientales';

  when ('Cr') depart1=' Creuse';
  when ('VH') depart1='Haute-Vienne ';
  when ('Mm') depart1='Meurthe-et-Moselle ';
  when ('Mo') depart1=' Moselle';
  when ('Mu') depart1='Meuse ';
  when ('Vo') depart1='Vosges ';
  when ('AG') depart1='Ari ge ';
  when ('AV') depart1=' Aveyron';
  when ('Ge') depart1='Gers ';
  when ('Hg') depart1='Haute-Garonne ';
  when ('Lt') depart1=' Lot';
  when ('Ph') depart1=' Haute-Pyr n es';
  when ('TG') depart1=' Tarn-et-Garonne';
  when ('Ta') depart1=' Tarn';

  when ('No') depart1='Nord ';
  when ('Pc') depart1=' Pas-de-Calais';
  when ('Lm') depart1=' Loire-Atlantique';
  when ('Ml') depart1='Maine-et-Loire ';
  when ('My') depart1=' Mayenne';
  when ('Sa') depart1=' Sarthe';
  when ('Ve') depart1='Vend e ';
  when ('As') depart1=' Aisne';
  when ('Oi') depart1=' Oise';
  when ('So') depart1=' Somme';
  when ('2S') depart1=' Deux-S vres';
  when ('Ch') depart1='Charente ';
  when ('Cm') depart1=' Charente-Maritime';
  when ('Vi') depart1='Vienne ';
  when ('AH') depart1=' Hautes-Alpes';
  when ('Am') depart1=' Alpes-Maritimes';
  when ('Ap') depart1=' Alpes-de-Haute-Provence';
  when ('Br') depart1=' Bouches-du-Rh ne';
  when ('Va') depart1=' Var';
  when ('Vc') depart1=' Vaucluse';
  when ('Ai') depart1=' Ain';
  when ('Ar') depart1='Ard che ';
  when ('Dr') depart1=' Dr ne';
  when ('Is') depart1='Is re ';
  when ('Lo') depart1=' Loire';
  when ('Ro') depart1=' Rh ne';
  *when ('Sh') depart1=' Haute-Savoie';
```

```
116   when ('Sv') depart1=' Savoie';
117
118   otherwise depart1=depart;
119   end;
120   if region="Basse-Normandie" and depart="Ca" then depart1='Calvados';
121   if region="Bretagne" and depart="Ca" then depart1="C tes-d'Armor ";
122   if region="Limousin" and depart="Co" then depart1=" Corr ze";
123   if region="Bourgogne" and depart="Co" then depart1="C te d'Or ";
124   if region="Champagne-Ardenne" and depart="Mh" then depart1="Haute-Marne
          ";
125   if region="Bretagne" and depart="Mh" then depart1="Morbihan ";
126
127   if region="Aquitaine" and depart="Pa" then depart1="Pau";
128   if region="Ile-de-France" and depart="Pa" then depart1="Paris ";
129   if region="Franche-Comt " and depart="Sh" then depart1="Haute-Sa ne ";
130   if region="Rh ne-Alpes" and depart="Sh" then depart1="Haute-Savoie ";
131
132
133   run;
134   data bk.dep1;
135   set bk.dep1;
136   drop OBS depart;
137   run;
138   proc sort data=bk.dep1;
139   by depart1;
140   run;
141   title "Base de donn es des 95 d partements fran ais recoder sous SAS";
142   footnote "@Boubacar KANDE";
143   proc print data=bk.dep1;
144
145   run;
146   proc sort data=bk.dep1;
147   by region;
148   run;
149   /* regrouper les r gions selon leurs positions g ographique*/
150   data bk.dep3;
151
152   set bk.dep1;
153
154   select;
155
156   when(region in ("Nord-Pas-de-Calais","Picardie","Haute-Normandie","
          Champagne-Ardenne"))
157   group="Nord";
158   when(region in ("Alsace","Lorraine","Franche-Comt ")) group="Est";
159   when(region in ("Basse-Normandie","Bretagne","  Pays de la loire")) group
          ="Ouest";
160   /*when(region in("Centre","Bourgogne")) group="CN";
161   when(region in ("Pointou-Charentes","Limousin")) group="CO";
162   when(region in ("Auvergne","Rh ne-Alpes")) group="CE";*/
163   when(region in ("Aquitaine","Midi-Pyr n es","Languedoc-Roussillon","
          Provence-Alpes-C te d'azur", "Corse")) group="Sud";
164   otherwise delete;
165   end;
166   run;
167   proc print data=bk.dep3; run;
```

Code 2: Analyse de données avec R

```r
library("FactoMineR")
library("factoextra")
library("corrplot")
library("dplyr")
library(tidyverse)
library(gtsummary)
library("gplots")
library("MASS")
library(haven)
library("ade4")
library("ggord")
library(readxl)
dep <- read_excel("//profils.uppa.univ-pau.fr/folderredir/bkande/Desktop/departement.xlsx")
View(dep)
attach(dep)
library(readxl)
depar<- read_excel("//profils.uppa.univ-pau.fr/folderredir/bkande/Downloads/departement(1).xlsx")
View(depar)
depar <- data.frame(depar)
library(stargazer)
d=summary(depar[,4:18])
d=data.frame(d)

stargazer(depar[,4:18],out="sum.tex")

table(depar$region)

depar%>%select(region)%>%tbl_summary(statistic=list(all_categorical()~" {n}"))



# ACP



cor <- cor(dep[,4:18])
View(cor)
cor <- round(cor,3)

# pour exporter la matrice de cor
setwd("//profils.uppa.univ-pau.fr/folderredir/bkande/Desktop")
write.csv(cor, "cor.csv")







dep <- dep[,-c(1,3)]
res$cor
acp <- PCA(dep,quali.sup =1:2, scale.unit=T,ncp = 15)
summary(acp,nbelements = Inf)
dimdesc(acp,axes = 1:3,prob=0.05)
fviz_eig(acp, addlabels = T,ylim=c(0,50))
res <- get_pca_var(acp)
```

```r
print(res)

corrplot(res$cor)
corrplot(res$cor, order = "AOE",method = "color")
corrplot(res$cos2)
# dim 1 et dim 2

fviz_pca_var (acp, col.var = "black", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), select.var =  list(name = NULL, cos2 = 0.6, contrib = NULL))
fviz_pca_ind (acp, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), select.ind = list(name = NULL, cos2 = 0.6, contrib = NULL))

# dim 1 et dim 3

fviz_pca_var (acp,axes = c(1,3), col.var = "black", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), select.var =  list(name = NULL, cos2 = NULL, contrib = 10))
fviz_pca_ind (acp, axes=c(1,3),col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), select.ind = list(name = NULL, cos2 = NULL, contrib = 50))

# dim 2 et dim 3
fviz_pca_var (acp,axes=c(2,3), col.var = "black", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), select.var =  list(name = NULL, cos2 = NULL, contrib = 10))
fviz_pca_ind (acp,axes = c(2,3), col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), select.ind = list(name = NULL, cos2 = NULL, contrib = 50))


# AFD

library(haven)
dep <- read_sas("//profils.uppa.univ-pau.fr/folderredir/bkande/Desktop/dep.sas7bdat",
NULL)
View(dep)

dis1 <-discrimin(dudi.pca(dep[,4:18], scan = FALSE), factor(dep$groupe), scannf = FALSE,nf=4)
#R2
dis1$eig

#valeurs propres
dis1$eig/(1-dis1$eig)
#cosines between the variables and the canonical scores (correlation)
dis1$va
#Coefficients of linear discriminants
lda<-lda(dep[,4:18],dep$groupe)
lda$scaling
#Graghiques
ggord(lda, factor(dep$groupe), ylim = c(-10, 10),xlim=c(-15,15),c("1", "2"),repel=TRUE)

ggord(lda, factor(dep$groupe), ylim = c(-10, 10),xlim=c(-15,15),c("1", "3"),repel=TRUE)
```

```r
ggord(lda, factor(dep$groupe), ylim = c(-10, 10),xlim=c(-15,15),c("2", "3")
    ,repel=TRUE)

ggord(lda, factor(dep$groupe), ylim = c(-10, 10),xlim=c(-15,15),c("1", "4")
    ,repel=TRUE)

#classification de l'algorithme
p1 <-predict(lda, dep[,4:18])
class <-predict(lda, dep[,4:18])$class
TAble<-cbind(dep[,3:18],p1$posterior,p1$x,class)
View(TAble)

TAble<-cbind(p1$posterior,p1$x,class,dep[,c(3,2)])

View(TAble)


# AT

pays.dist = dist(dep1[,3:17])
pays.hclust = hclust(pays.dist,method = "ward.D2")
plot(pays.hclust,labels=dep1$depart1,main='Dendogramme')
inertie <-sort(pays.hclust$height, decreasing = TRUE)
plot(inertie[1:10], type = "s", xlab = "Nombre de classes", ylab = "Inertie
    ",lwd=2);grid()
k <-4


k <-4
abline(v=k,col="red",lty=3)
points(k,inertie[k],pch=16,cex=2,col="red")
plot(pays.hclust,labels=dep1$depart1,main='Dendogramme',cex=0.7)
rect.hclust(pays.hclust,k=4)
groups.3 = cutree(pays.hclust,4)
pays.hclust$merge
dep1$depart1[groups.3 == 1]
dep1$depart1[groups.3 == 2]
dep1$depart1[groups.3 == 3]
dep1$depart1[groups.3 == 4]
a <- aggregate(dep1[,3:17],list(groups.3),mean)

a <- round(a,3)
attach(a)

locator()
text(37.42724,611.6893,"1",col="red",cex=2)
text(10.05612,1059.848,"4",col="red",cex=2)
text(73.74316,1109.644,"3",col="red",cex=2)
text(86.44479,985.155,"2",col="red",cex=2)
legend("topleft",legend = c("1 -> Groupe 1","2 -> Groupe 2","3 -> Groupe 3"
    ,"4 -> Groupe 4"),col = "red",text.col="red")
```