

# Projet ML - Clustering

Anthony LEZIN

4/19/2021

## I - Généralités et préparation des données

### A - Intoduction

Le but de l'analyse exploratoire est de caractériser les départements français du point de vue de leur composition en population générale et en population active. L'analyse doit être mise en oeuvre en deux étapes.

Je crée donc deux ensembles de données agrégées au niveau du département. Dans ces ensembles de données, chaque département métropolitain français doit être décrit par des statistiques résumant la composition de la population.

Le premier ensemble de données correspond aux départements tels que caractérisés par leur population générale, tandis que dans le second ensemble de données, les départements sont décrits en utilisant la population active.

Dans un deuxième temps, les deux ensembles de données seront analysés à l'aide d'algorithmes de clustering et de visualisation dans le but d'évaluer si les départements français sont homogènes en termes de composition de la population ou, au contraire, séparés en différents groupes.

## B - Sélection des covariables et construction du Dataset pour le clustering

Afin de décrire les départements, je ne considère que des variables numériques ou numérisé (comme préconisé dans l'intitulé). J'intègre le taux de femmes par département, des indicateurs démographiques tels la population, le nb\_d'habitants, l'âge moyen.

J'aimerai intégrer dans mon étude le niveau d'étude moyen, car je pense qu'il est susceptible de significativement séparer des clusters. Je crée une variable supplémentaire "edu" permettant de numériser le niveau d'étude de la façon suivante :

- ses valeurs représentent le nombre approximatif d'années d'études à partir de l'année de CP (année 0).
- lorsque ce nombre n'est pas renseigné où lorsqu'il s'étend sur une plage de plusieurs années, la valeur est fixée au nombre d'année médian (par exemple 2.5 est attribué au nombre d'année d'un individu ayant arrêté sa scolarité entre le CP et le CM2 non révolu)
- Lorsqu'une examen diplômant achève l'année, une majoration de 0.1 est attribué. Ainsi, pour un individu ayant terminé son cycle scolaire en terminale, son nombre d'année d'étude est 12.1 (5 années de primaire, 4 de collège, 3 de lycée et 0.1 du diplôme).

```
##  
## csp_1_1 csp_1_2 csp_1_3 csp_2_1 csp_2_2 csp_2_3 csp_3_1 csp_3_3 csp_3_4 csp_3_5  
##      526      52     148    1651    1444     362     999     348     239     370  
## csp_3_7 csp_3_8 csp_4_2 csp_4_3 csp_4_4 csp_4_5 csp_4_6 csp_4_7 csp_4_8 csp_5_2  
##      173     183     441     731      35     288     955     411     209     643  
## csp_5_3 csp_5_4 csp_5_5 csp_5_6 csp_6_2 csp_6_3 csp_6_4 csp_6_5 csp_6_7 csp_6_8  
##      273     601     680     872     313     887     245     185     733     288  
## csp_6_9 csp_7_1 csp_7_2 csp_7_4 csp_7_5 csp_7_7 csp_7_8 csp_8_1 csp_8_4 csp_8_5  
##      372    1324    2105    2393    5654    8664    6790     815    8406    6604  
## csp_8_6  
##      1780  
## [1] 80.93767
```

81 % des individus possédant un salaire manquant possède un statut professionnel appartenant aux séries des csp-7 ("anciens quelque chose"), aux csp-8 (chômeurs, militaires, élèves, étudiants, personnes sans activité), aux csp-1 (agriculteurs) et csp\_2 (professions libérales). N'ayant aucune donnée permettant de leur définir un salaire, je prend le parti de leur attribuer un salaire égal à 0.

Pour les autres csp, malgré la salaire manquant, je possède quelques informations employeurs. Je décide donc de leur affecter la moyenne des salaires de leur csp correspondante. Cela permet d'ajouter environ 12 000 individus à la catégorie des individus avec un salaire positif.

Je résume chaque département par des statistiques résumant la composition de la population telles :

- le pourcentage de ses habitants féminins
- son nombre d'habitants
- la répartition des diplômes des habitants
- le salaire moyen
- le niveau d'étude numérisé

Je crée donc 2 jeux de données :

- un premier ensemble de données correspondant aux départements tels que caractérisés par leur population générale

Dep	nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
01	490890	420	9.668333	11509.802	55.23810	50.20952	189
02	275062	143	9.512587	9869.189	45.45455	50.24476	90
03	248102	196	9.583673	11452.520	48.46939	48.07143	95
04	89641	41	10.180488	9733.098	51.21951	50.68293	25
05	75602	37	9.718919	8116.676	56.75676	56.72973	20
06	1082574	2041	9.976335	10461.586	55.21803	50.91622	111

- un second ensemble de données où seule la population active est prise en compte

Dep	nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
01	412586	216	10.014352	22380.17	51.38889	42.69444	130
02	233540	75	10.194667	18817.25	42.66667	38.68000	59
03	214110	108	9.374074	20784.20	54.62963	42.12963	66
04	74105	23	10.221739	17350.30	47.82609	39.30435	19
05	61422	16	10.031250	18769.81	62.50000	39.31250	9
06	1067261	968	10.488533	22057.95	54.13223	43.02789	88

## B - Étude descriptive simple

### 1 - Population globale

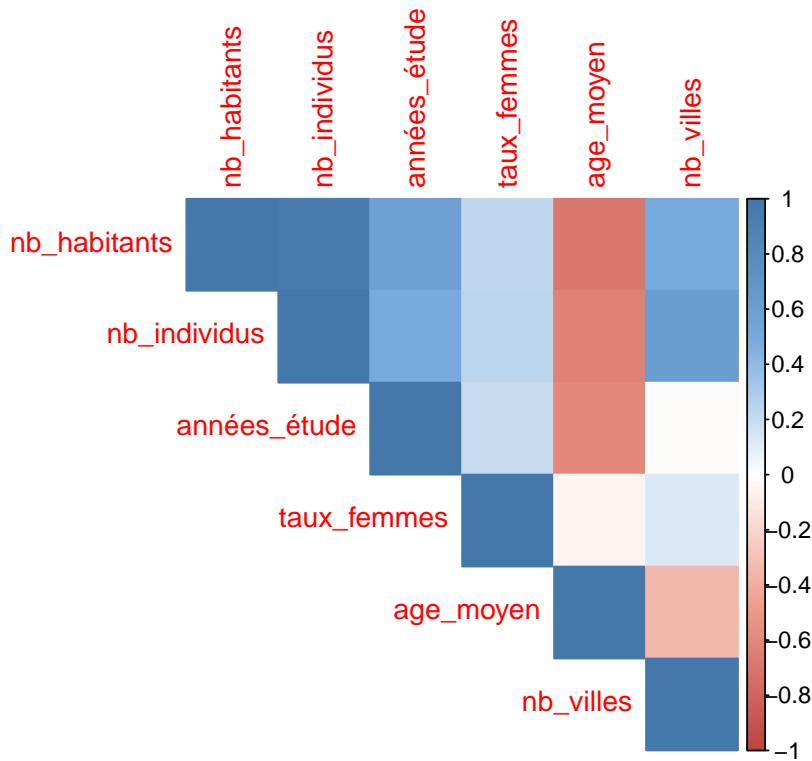
J'effectue une carte de France descriptive des indicateurs que je trouve intéressant.

Le nombre pertinent de catégories est l'objet des parties ultérieures. Pour l'exemple, je visualise des cartes avec des variables comportant 5 catégories.

Dep	nb_habitants	nb_individuals	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes	NAME_2	REG
01	490890	420	9.668333	11509.802	55.23810	50.20952	189	Ain	84
02	275062	143	9.512587	9869.189	45.45455	50.24476	90	Aisne	32
03	248102	196	9.583673	11452.520	48.46939	48.07143	95	Allier	84
04	89641	41	10.180488	9733.098	51.21951	50.68293	25	Alpes-de-Haute-Provence	93
05	75602	37	9.718919	8116.676	56.75676	56.72973	20	Hautes-Alpes	93
06	1082574	2041	9.976335	10461.586	55.21803	50.91622	111	Alpes-Maritimes	93

En théorie, la proportion d'actifs par département a été repectée dans l'échantillonage proposé, mais dans le doute, j'ôte le salaire moyen de l'étude de la population de manière globale.

Observons ce qu'un corrélogramme nous indique.

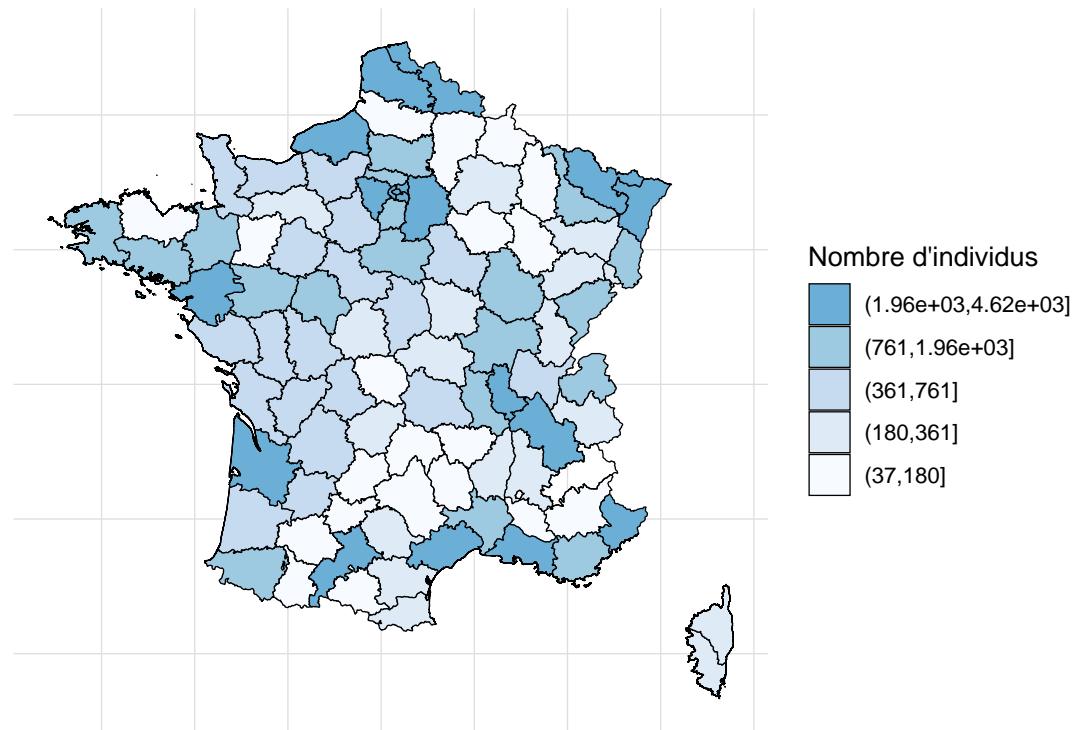


En mettant à la marge les corrélations triviales, on observe une corrélation positive entre le niveau d'études et les indicateurs démographiques. Ainsi, il est plus élevé dans les zones densément peuplées.

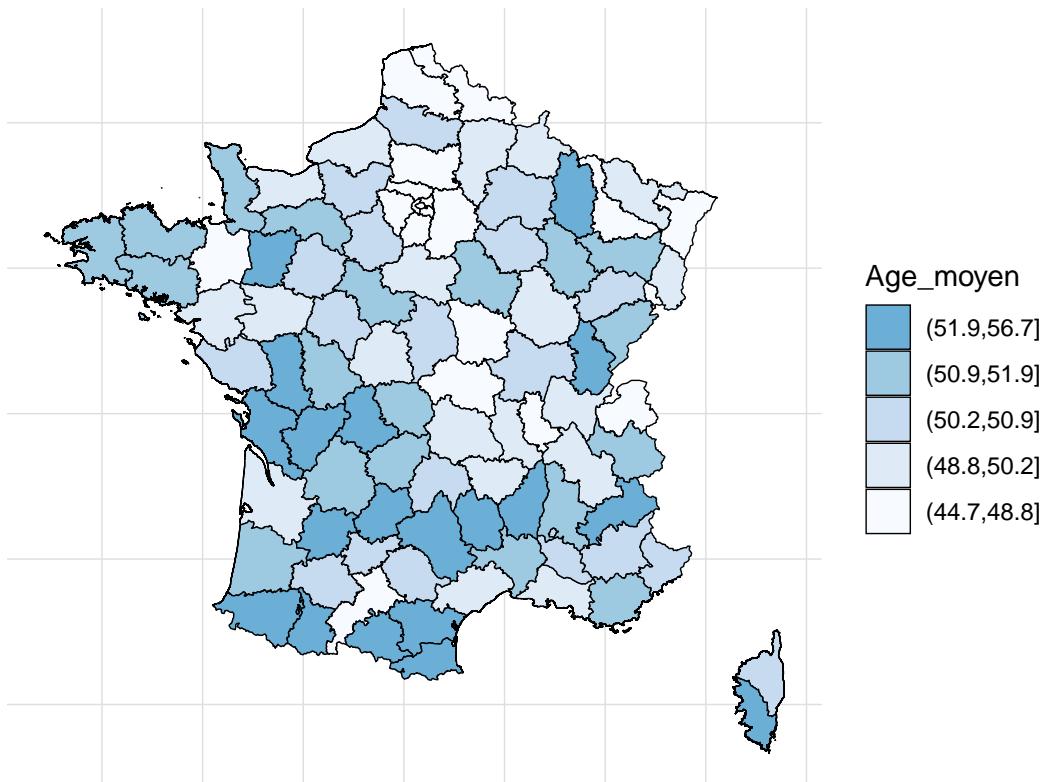
L'âge moyen, lui, est fortement négativement corrélé avec le nombre d'habitants/individuals, ce qui semble indiquer que les zones les plus densément peuplées sont également les zones "les plus jeunes".

Le taux de femmes semble peut influer sur les différents indicateurs.

représentation du nombre d'individus par département

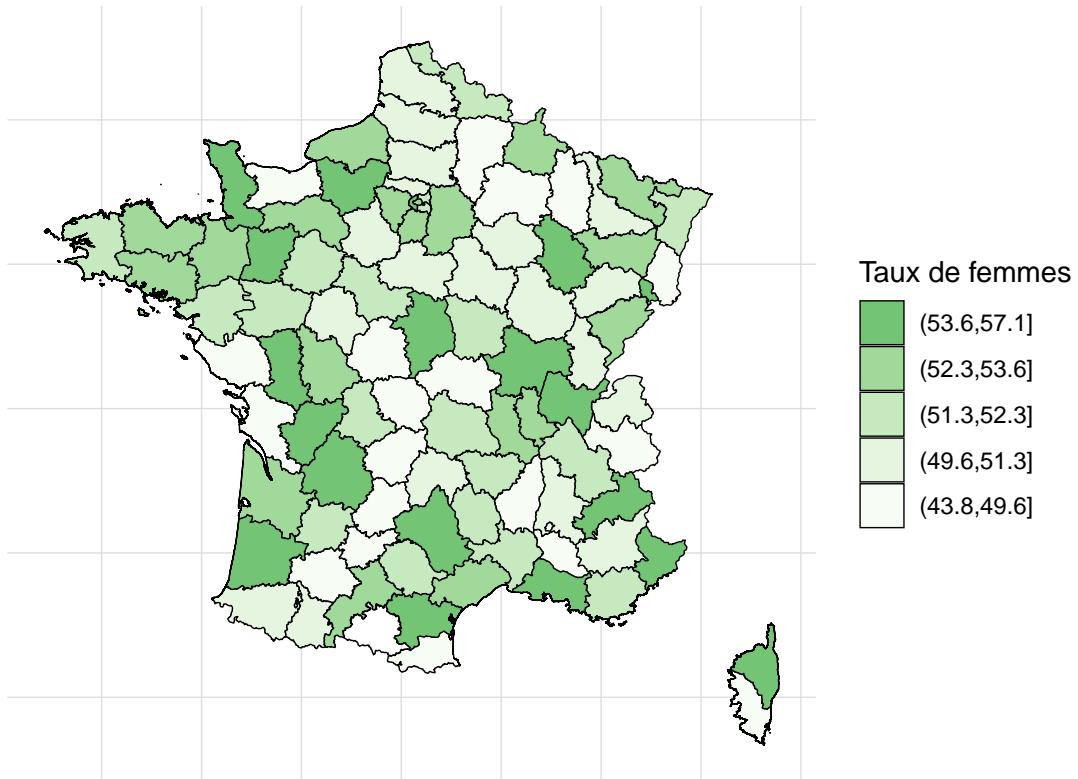


représentation de l'âge moyen par département



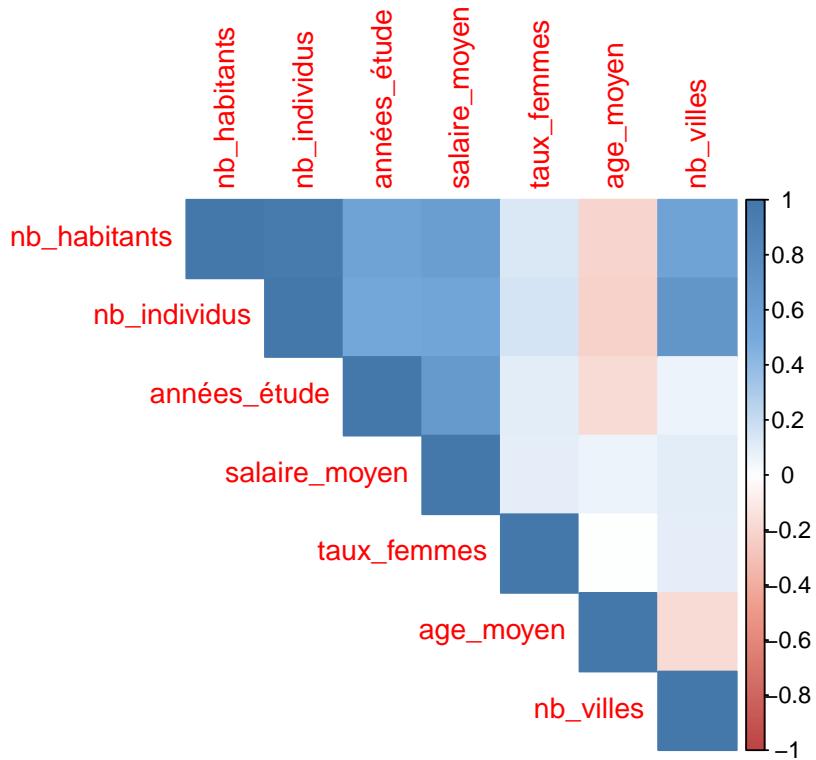
L'intensité chromatique est bien inversée. La cartographie confirme le 2nd point.

## représentation du taux de femmes par département



A ce stade, je n'observe pas d'éléments particuliers sur l'indicateur du taux de femmes par département.

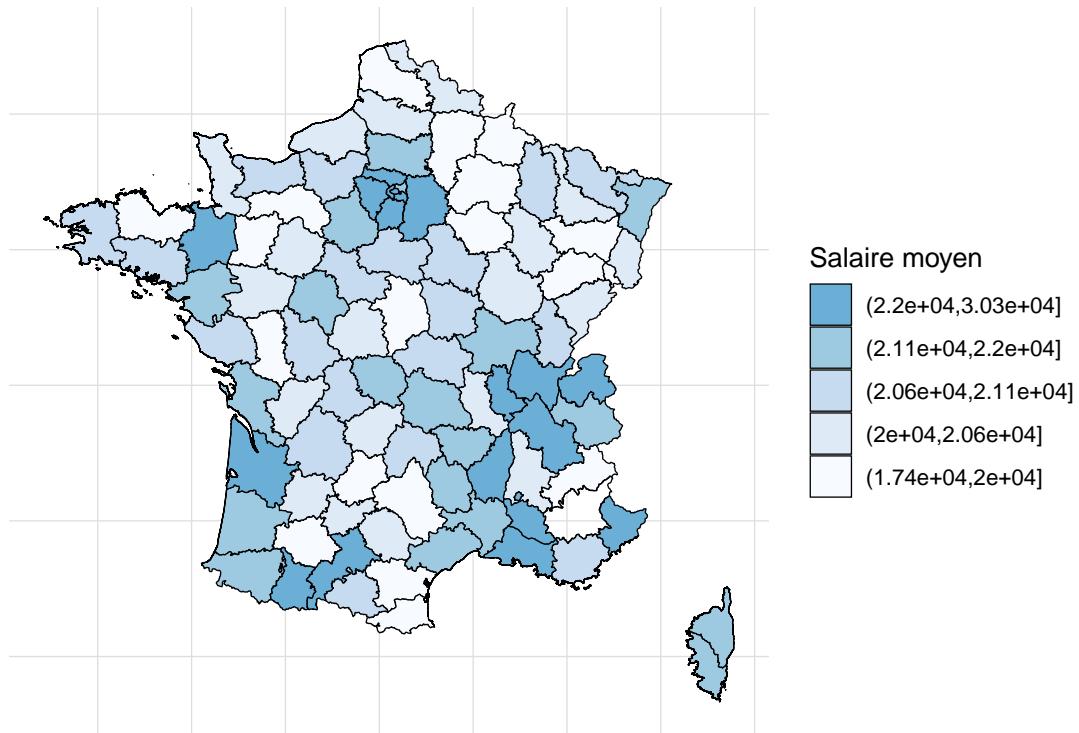
## 2 - population active



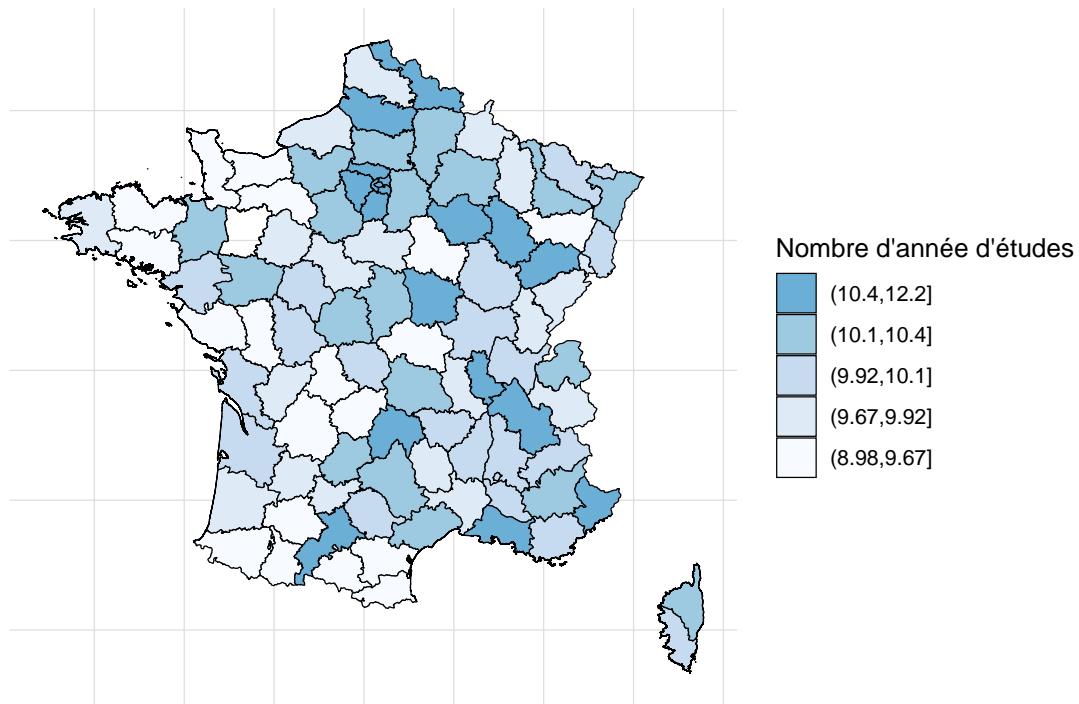
On retrouve un peu les mêmes tendances que précédemment :

le taux de femmes semble toujours peut influer sur les différents indicateurs et l'âge moyen semble toujours corrélé négativement avec le nombre d'habitants (même si cela semble un peu moins marqué). La nouveauté réside sur le salaire moyen. Il est positivement correlé avec le niveau d'études et les indicateurs démographiques.

## représentation des salaires moyens par département



## Nombre d'année d'études par département



Les deux dernières cartes sont assez similaires. Elles mettent toutes les deux en avant une concentration des salaires et des niveaux d'étude les plus élevés autour des agglomérations les plus denses (Île-de-France, région lyonnaise, département de l'Ille et Vilaine (en Bretagne) et le sud-Est).

Ces régions peuvent être opposées à une zone communément appelée “la diagonale du vide” où les salaires,

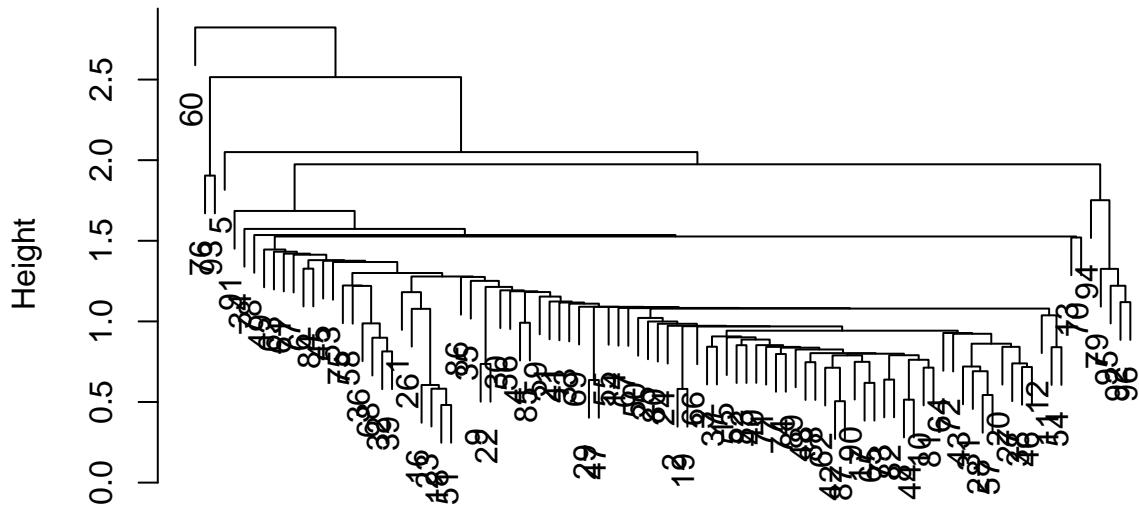
tout comme les niveaux d'étude semblent moins élevés.

## II. Classification Ascendante Hiérarchique dans la population globale

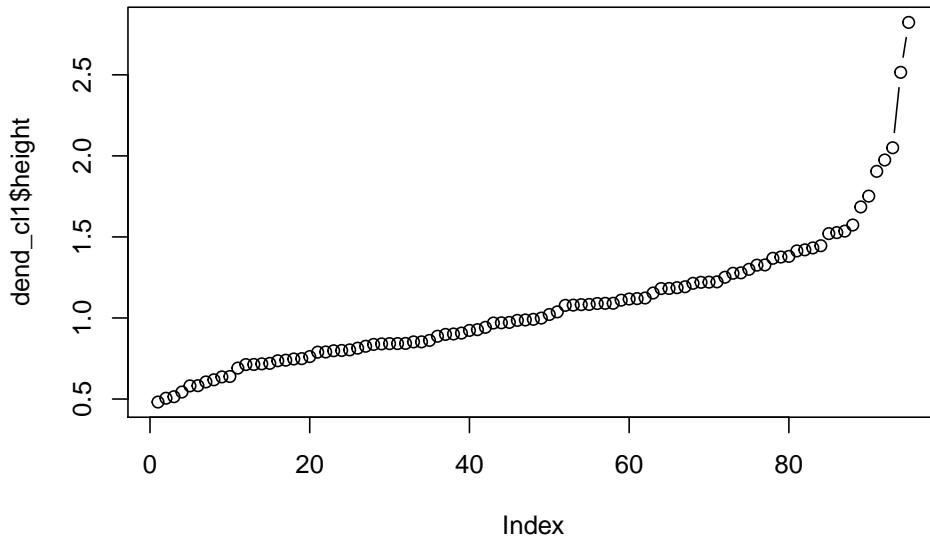
Je normalise les données afin d'uniformiser l'échelle de grandeur des nombres pour éviter que les performances du modèle soit dirigées dans la direction des nombres les plus grands.

nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
-0.2171	-0.5593	-0.0074	0.5778	1.3276	-0.0786	0.1286
-0.6211	-0.8085	-0.3588	-0.3816	-2.1199	-0.0617	-0.7238
-0.6715	-0.7608	-0.1984	0.5443	-1.0575	-1.1035	-0.6807
-0.9681	-0.9002	1.1483	-0.4612	-0.0884	0.1483	-1.2834
-0.9944	-0.9038	0.1068	-1.4065	1.8628	3.0469	-1.3265
0.8902	0.8990	0.6876	-0.0352	1.3206	0.2601	-0.5430

**Cluster Dendrogram**

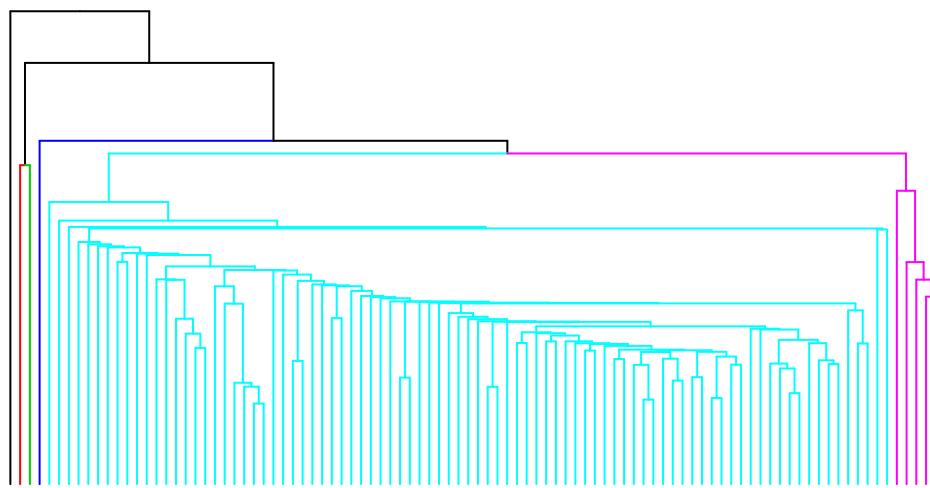


```
dist(train_cl_sans_dep, method = "euclidean")
hclust (*, "single")
```



La règle du coude engagerait à choisir 3 ou 6 clusters.

```
##  1   2   3   4   5   6
##  1   1   1   1  87   5
```



La décomposition dans chacune des configurations est excrécable.

Par exemple, avec 6 clusters, 85 départements ne semblent former qu'un seul cluster et 4 clusters ne contiennent qu'un seul département...

### Quels sont les départements atypiques ?

nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
-0.2171	-0.5593	-0.0074	0.5778	1.3276	-0.0786	0.1286
-0.6211	-0.8085	-0.3588	-0.3816	-2.1199	-0.0617	-0.7238
-0.6715	-0.7608	-0.1984	0.5443	-1.0575	-1.1035	-0.6807
-0.9681	-0.9002	1.1483	-0.4612	-0.0884	0.1483	-1.2834

Ces départements isolés ont comme caractéristique commune d'être **moins peuplés que la moyenne**.

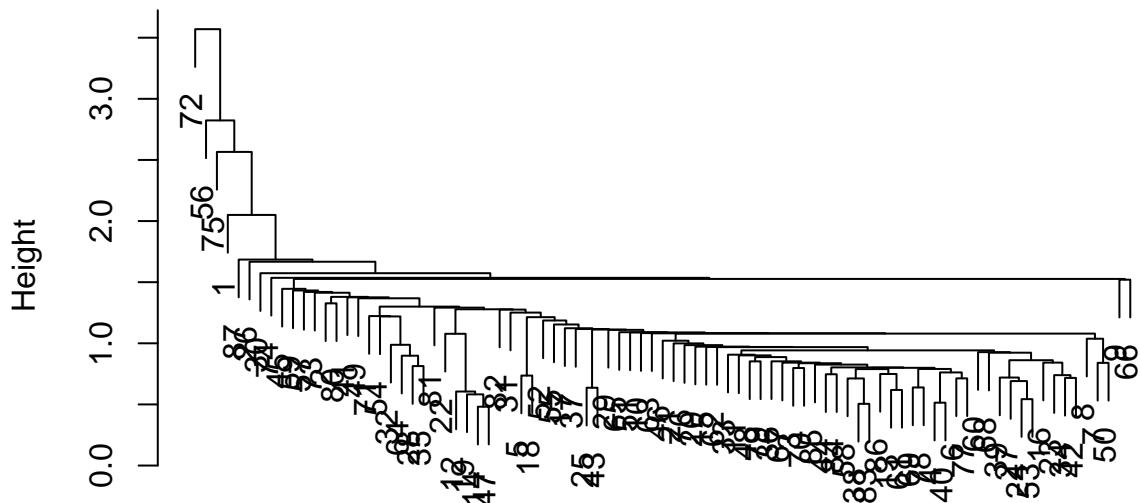
Effectuons une anova pour identifier d'éventuelles variables discriminantes.

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## nb_habitants   1 1.66  1.6596  4.521 0.0363 *
## nb_individus   1 0.00  0.0002  0.001 0.9818
## années_étude   1 0.17  0.1697  0.462 0.4984
## salaire_moyen  1 0.00  0.0000  0.000 1.0000
## taux_femmes    1 0.04  0.0364  0.099 0.7536
## age_moyen      1 0.09  0.0913  0.249 0.6192
## nb_villes       1 0.48  0.4804  1.309 0.2557
## Residuals      88 32.30  0.3671
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

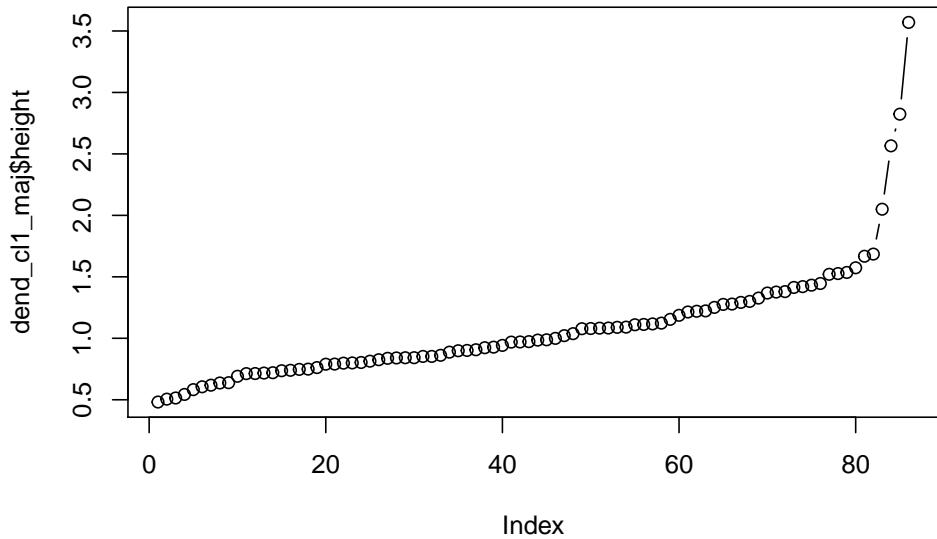
Mon hypothèse est confirmée par une analyse de la variance :  
mis à part le nombre d'habitants , aucune covariable ne semble significative dans les clusters formés.

J'effectue un clustering sur la classe majoritaire (en l'occurrence ici : le cluster n°5) afin de voir si la classification hiérarchique permet d'identifier des éléments intéressants.

## Cluster Dendrogram

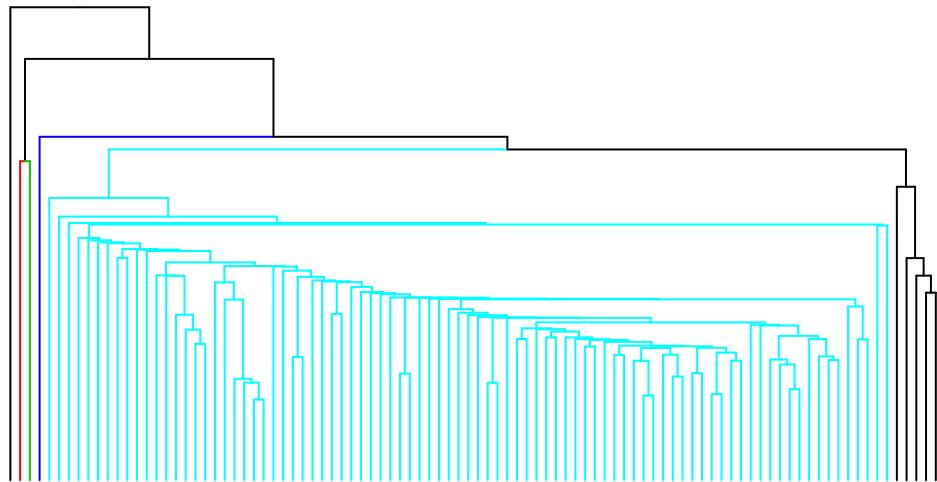


```
dist(train_cl_sans_dep[which(clusters == 5), ], method = "euclidean")
hclust (*, "single")
```



La règle du coude engagerait cette fois-ci à choisir 3, 5 clusters. Voyons ce qui se passe avec 5 clusters.

```
##  1  2  3  4  5
## 1  1  1  1 83
```

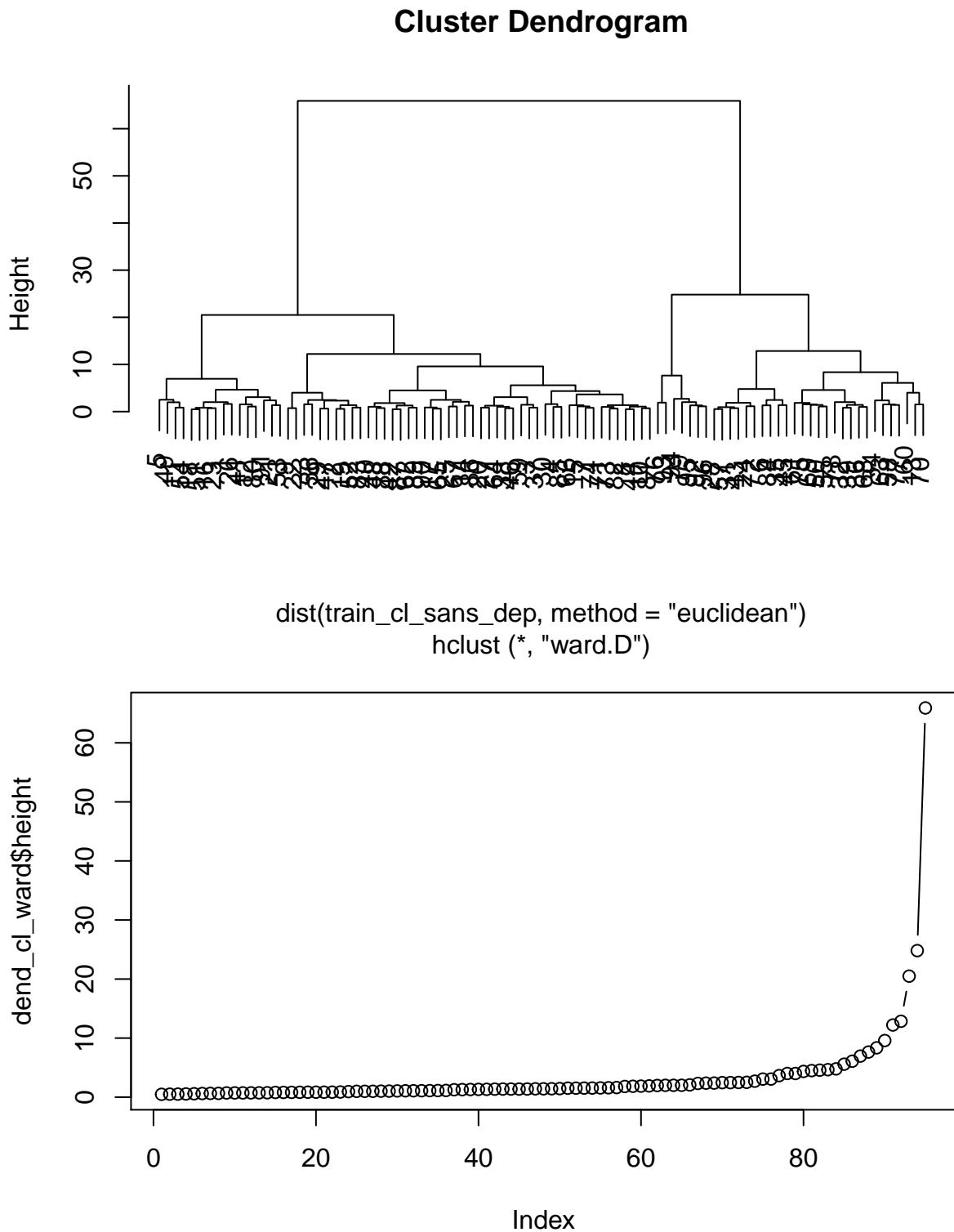


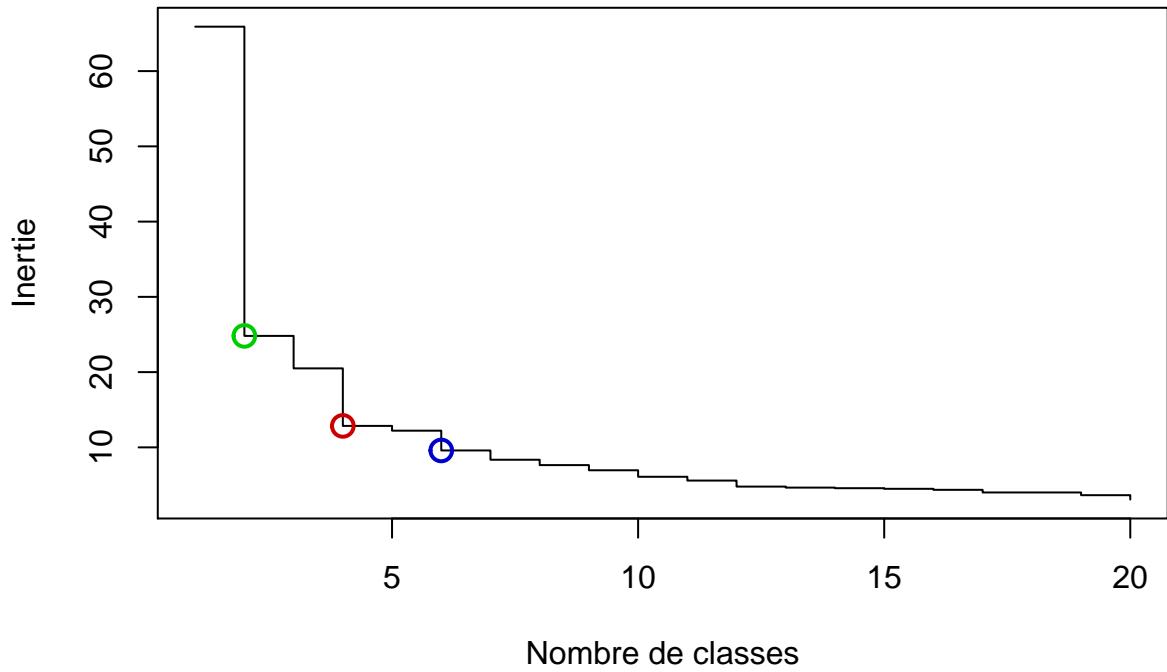
```
##                               Df Sum Sq Mean Sq F value    Pr(>F)
## nb_habitants      1  0.107  0.107   0.368  0.545699
## nb_individus      1  0.000  0.000   0.000  0.986906
## années_étude      1  0.472  0.472   1.617  0.207298
## salaire_moyen     1  0.561  0.561   1.923  0.169455
## taux_femmes       1  1.109  1.109   3.802  0.054732 .
## age_moyen         1  3.418  3.418  11.717  0.000984 ***
## nb_villes          1  0.138  0.138   0.475  0.492858
## Residuals        79 23.045  0.292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On est ramené à la situation précédente avec un clustering extrêmement déséquilibré. Cette fois-ci, seul l'âge moyen semble discriminant. Bref, cette méthode ne semble pas adaptée pour un clustering adéquat.

### III. Clustering hiérarchique avec la méthode de Ward dans la population globale

#### A. Clustering

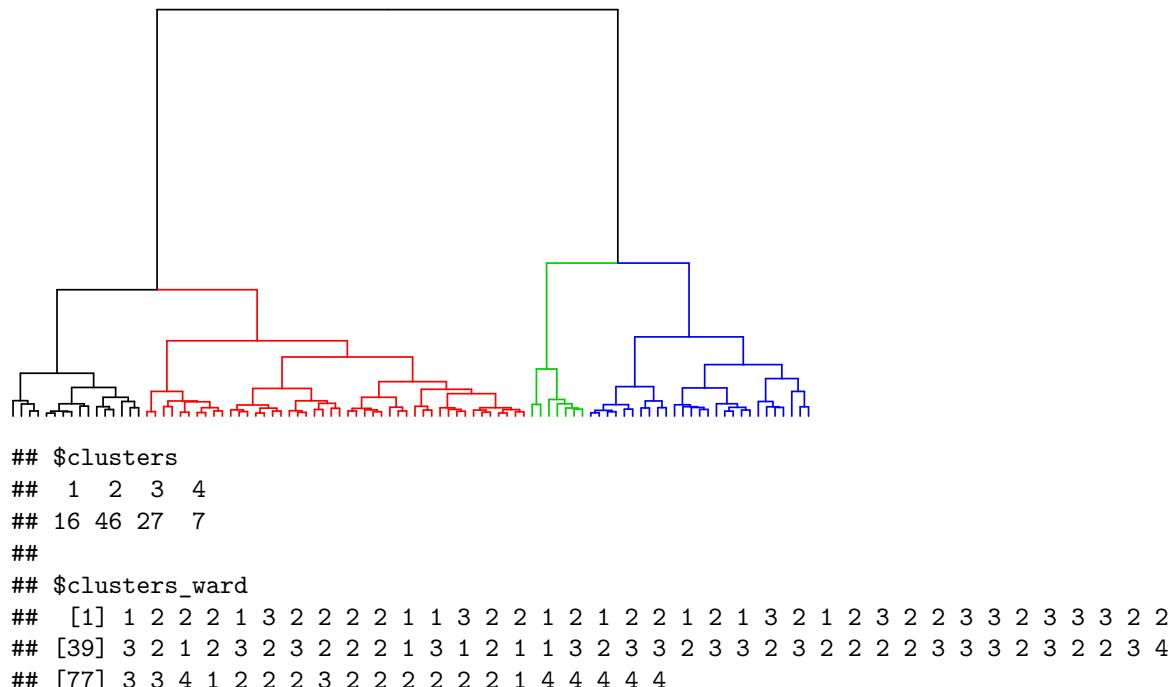




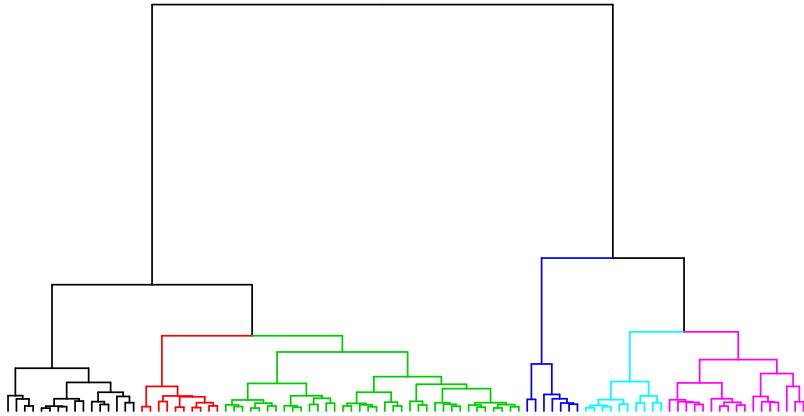
L'étude de l'inertie par rapport aux nombres de classes semblent confirmer les choix raisonnables de 4 et/ou 6 clusters ( $k = 2$  me semble trop faible).

Observons les dendogrammes associés, ainsi que leurs clusters respectifs.

Ainsi pour 4 clusters



Ainsi pour 6 clusters



```

## $clusters
##  1  2  3  4  5  6
## 16 10 36 10 17  7
##
## $clusters_ward
## [1] 1 2 3 3 1 4 3 3 2 3 1 1 5 3 3 1 3 1 2 3 1 2 1 4 3 1 3 4 2 3 4 5 2 5 4 5 3 3
## [39] 5 3 1 3 4 3 4 3 2 3 1 5 1 2 1 1 5 2 4 5 3 5 5 3 5 3 3 3 5 5 5 3 4 3 3 5 6
## [77] 5 5 6 1 3 3 2 4 3 3 3 3 3 1 6 6 6 6 6

```

C'est beaucoup mieux avec la méthode de Ward. Les “sauts” provoqués par le nombre de clusters semblent équilibrer la décomposition.

```

##              Df Sum Sq Mean Sq F value    Pr(>F)
## nb_habitants   1 37.53  37.53 200.022 < 2e-16 ***
## nb_individus   1  0.14   0.14   0.732  0.39452
## années_étude   1  2.07   2.07  11.041  0.00130 **
## salaire_moyen  1  1.48   1.48   7.878  0.00616 **
## taux_femmes    1  5.91   5.91  31.508 2.28e-07 ***
## age_moyen      1  0.77   0.77   4.092  0.04612 *
## nb_villes       1  0.08   0.08   0.424  0.51679
## Residuals     88 16.51   0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##              Df Sum Sq Mean Sq F value    Pr(>F)
## nb_habitants   1 128.90 128.90 222.942 < 2e-16 ***
## nb_individus   1   1.52   1.52   2.629  0.10850
## années_étude   1   5.67   5.67   9.800  0.00237 **
## salaire_moyen  1   4.87   4.87   8.419  0.00469 **
## taux_femmes    1  11.84  11.84  20.486 1.87e-05 ***
## age_moyen      1   5.11   5.11   8.840  0.00380 **
## nb_villes       1   0.69   0.69   1.199  0.27660
## Residuals     88 50.88   0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

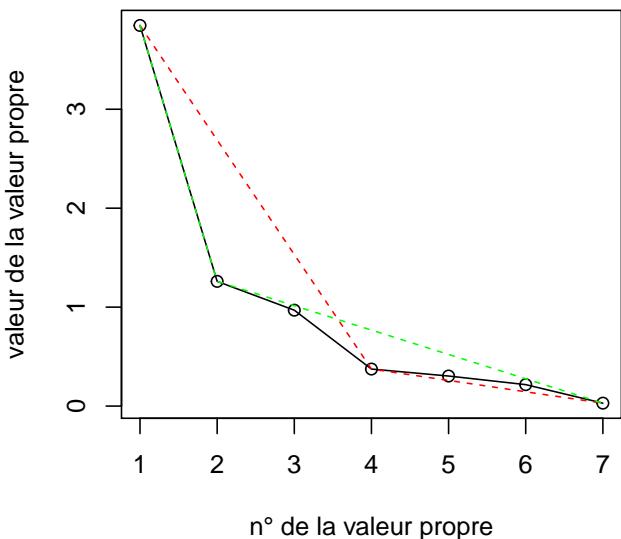
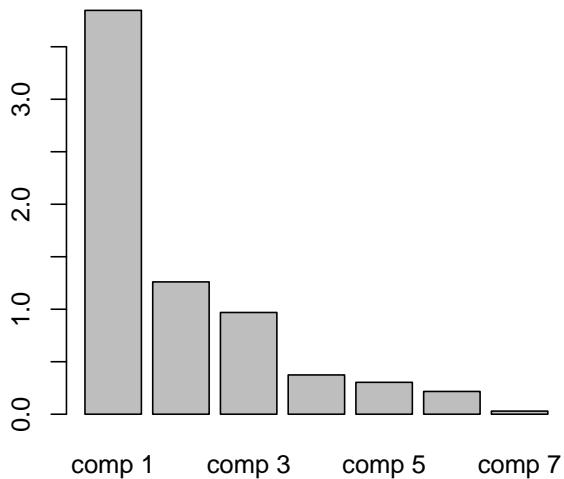
Dans une classification à 4 ou 6 clusters, le taux de femmes, le nombre d'année d'études, le salaire moyen, le nombre d'habitants et l'âge moyen s'avèrent être des covariables extrêmement ou sensiblement discriminantes.

## B. ACP avec la méthode de Ward

### 1. Pourcentage de variance expliquée

Observons le pourcentage de variance expliquée.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.8465086	54.9501236	54.95012
comp 2	1.2603377	18.0048243	72.95495
comp 3	0.9689248	13.8417823	86.79673
comp 4	0.3741028	5.3443257	92.14106
comp 5	0.3037483	4.3392608	96.48032
comp 6	0.2168315	3.0975931	99.57791
comp 7	0.0295463	0.4220904	100.00000



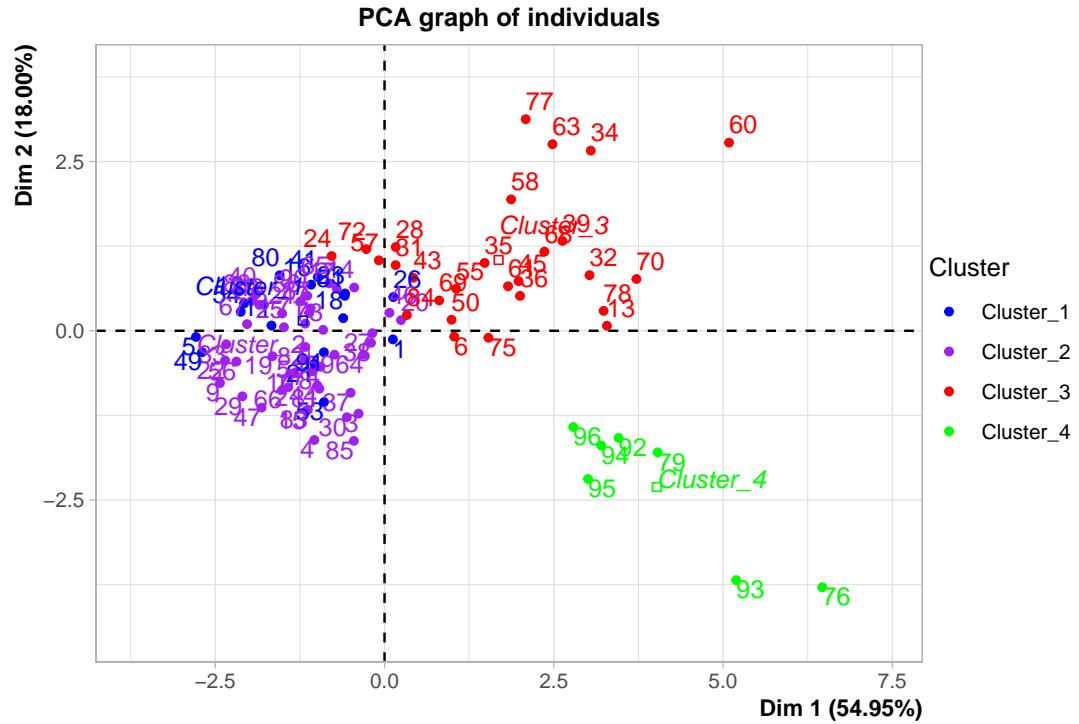
En ACP normée, l'inertie moyenne  $I/p$  vaut 1, je ne devrais retenir que les axes associés à des valeurs propres "quasi" supérieures à 1, mais je préfère conserver le nombre de groupes fourni par le clustering.

La proportion d'inertie expliquée par les 2 premiers axes est de 73 %, 87% pour les 3 premiers, 92% pour les 4 premiers et 96% pour les 5 premiers.

## 2. Description des axes selon les individus

Je pars des 4 clusters originels définis par la méthode de ward.

J'observe les individus obtenus par l'ACP en “séparant” les éléments selon leurs classes.



Le 1er plan de l'ACP semble convenablement départager les individus en fonction des clusters retenus.

L'axe 1 semble opposer les individus de la classe 3 et 4 (positivement corrélés avec cet axe) avec ceux des classes 1 et 2 (négativement corrélés).

L'axe 2, lui, oppose parfaitement les individus des clusters 3 et 4.

En revanche, les clusters 1 et 2 sont mal séparés par cet axe.

J'étudie les points possédant une contribution maximale sur chacun des axes sélectionnés en terme de  $\cos^2$  et non significative sur les autres.

	32	Dim.1	Dim.2	Dim.3	Dim.4
32	0.92	0.07	0.00	0.00	
28	0.01	0.80	0.01	0.04	
18	0.09	0.01	0.79	0.09	
84	0.05	0.02	0.08	0.76	

Voici les individus sélectionnés selon les clusters allant de 1 à 4

nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
1.2895	1.7492	0.8254	0.9884	0.4259	-1.0988	1.4545
0.5596	0.5860	-0.7141	-0.4131	0.3040	0.3196	0.5935
-0.6320	-0.5044	-0.2005	-0.3347	1.7895	0.1927	-0.1727
0.8246	0.7722	0.2693	-0.4303	0.2237	0.6808	-0.3708

L'ACP apporte de nouveaux éléments d'interprétation.

Il semble que :

- le cluster 1 caractérise les départements denses en nombre de villes, en nombre d'habitants avec une population plutôt jeune (bien en)dessous de la moyenne). Ces individus ont un taux de femmes, un salaire et un niveau d'étude au-dessus de la moyenne.
- le cluster 2 caractérise les départements dont les individus ont un salaire et un niveau d'étude en-dessous de la moyenne en conservant les autres indicateurs au-dessus
- le cluster 3 caractérise les départements dont les individus ont un salaire et un niveau d'étude fortement en-dessous de la moyenne. Ils résident dans un département plus faiblement peuplé comportant un taux de femmes assez bas
- le cluster 4 semble plus délicat à interpréter. Néanmoins, c'est seulement dans ce cluster que le salaires sont au-dessous de la moyenne alors que le niveau d'études est au-dessus,. Même phénomène pour le nombre d'habitants plutôt important par rapport au nombre de villes.

## IV. Clustering avec K-means dans la population globale

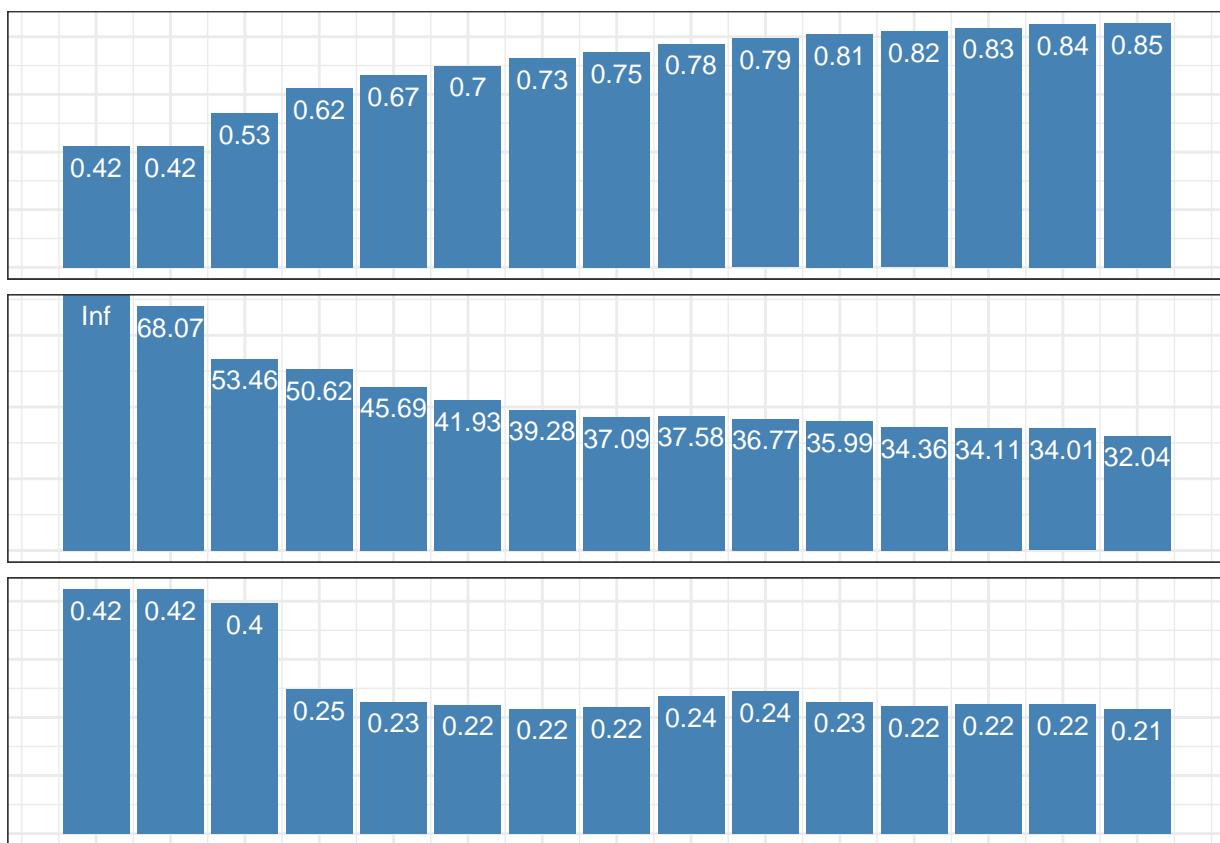
### A. Clustering

#### 1. Choix du nombre de classes

On émet l'hypothèse que les données ont une distribution gaussienne avec des variances isotropes (sans direction privilégiée). Sous cette hypothèse, la méthode des k-means peut se révéler intéressante pour, par exemple, effectuer de la réduction de dimensions. Sa construction permet de regrouper les données en petits clusters que l'on résume par le centre des classes obtenues.

Le k-means produit des classes parfaitement sphériques. Le nombre de clusters a été déterminé par le biais d'une autre classification.

Utilisons des indicateurs pour déterminer la valeur optimale de  $k$ .



Les indicateurs sont loin d'être d'accord et leur lecture s'avère peu aisée.

Il semblerait que 9 clusters fournissent une bonne valeur pour le pourcentage de variance expliquée. L'indice silhouette fournit un maximum à 2 clusters (ce qui semble trop peu au vu de la nature des données) et ne cesse de décroître. Enfin, l'indice de Calinski-Harabasz propose, quant à lui, 2 ou 3 clusters.

Compte-tenu des résultats précédemment obtenus, j'effectue un k-means avec 5 clusters et une initialisation aléatoire.

## 2. k-means à 5 classes

```

## K-means clustering with 5 clusters of sizes 6, 2, 33, 37, 18
##
## Cluster means:
##   nb_habitants nb_individus années_étude salaire_moyen taux_femmes age_moyen
## 1    1.6337597    1.1885457    1.6152302    1.7300299    0.3068410 -1.7375022
## 2    2.4773547    1.7469088    4.0564782    4.1478846    0.9925356 -1.9609048
## 3   -0.4035508   -0.3832919   -0.3888939   -0.4811326    0.8331923  0.6368505
## 4   -0.5821758   -0.6272933   -0.3047906   -0.2780723   -0.9134440  0.2698959
## 5    1.1166896    1.4018552    0.3503562    0.4161166    0.1375537 -0.9252995
##   nb_villes
## 1 -0.4812550
## 2 -1.2575768
## 3 -0.2176108
## 4 -0.4189305
## 5  1.5602371
##
## Clustering vector:
## [1] 3 4 4 4 3 3 4 3 4 4 3 3 1 4 4 3 4 3 4 4 3 4 3 3 4 3 4 3 4 3 3 5 4 5 5 5 4 4
## [39] 5 4 3 3 3 4 5 4 4 3 3 5 3 4 3 3 5 4 3 5 4 5 5 3 5 4 4 4 4 5 5 5 4 3 3 4 5 2
## [77] 5 5 1 3 4 4 4 3 4 4 3 3 3 4 3 1 2 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 14.20452 1.81351 83.18974 75.73190 56.48574
## (between_SS / total_SS = 65.2 %)
##
## Available components:
##
## [1] "cluster"         "centers"          "totss"            "withinss"
## [5] "tot.withinss"    "betweenss"        "size"             "iter"
## [9] "ifault"           "inicial.centers"

```

Effectuons une analyse de la variance afin de déterminer les variables les plus discriminantes

```

##              Df Sum Sq Mean Sq F value    Pr(>F)
## nb_habitants  1  0.37  0.366   0.768  0.3832
## nb_individus  1 26.71 26.710  55.988 5.21e-11 ***
## années_étude  1  1.95  1.951   4.090  0.0462 *
## salaire_moyen 1  0.25  0.247   0.517  0.4739
## taux_femmes   1 13.96 13.964  29.271 5.36e-07 ***
## age_moyen     1  1.98  1.984   4.160  0.0444 *
## nb_villes     1 11.54 11.537  24.183 4.04e-06 ***
## Residuals    88 41.98  0.477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Une analyse de la variance du clustering k-means relève que les covariables discriminantes sont plutôt de nature démographique : nombre d'habitants, d'individus, de villes, l'âge moyen et la proportion de femmes.

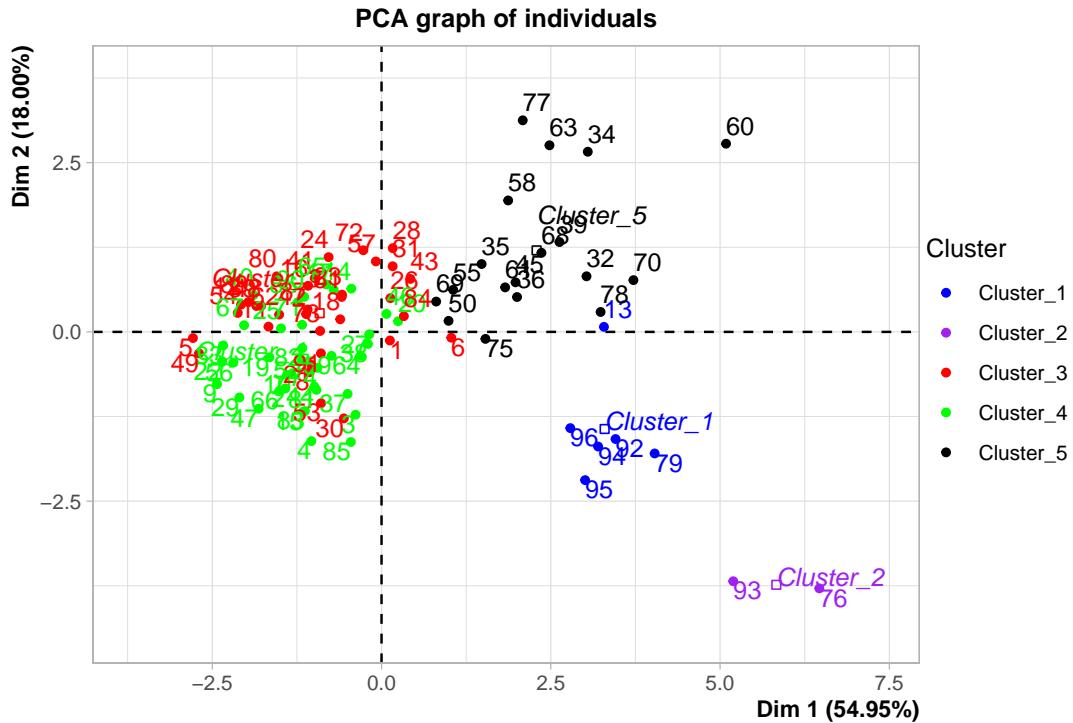
Voici la classification obtenue

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
13	76	1	2	32
79	93	5	3	34
92	NA	6	4	35
94	NA	8	7	36
95	NA	11	9	39
96	NA	12	10	45
NA	NA	16	14	50
NA	NA	18	15	55
NA	NA	21	17	58
NA	NA	23	19	60
NA	NA	24	20	61
NA	NA	26	22	63
NA	NA	28	25	68
NA	NA	30	27	69
NA	NA	31	29	70
NA	NA	41	33	75
NA	NA	42	37	77
NA	NA	43	38	78
NA	NA	48	40	NA
NA	NA	49	44	NA
NA	NA	51	46	NA
NA	NA	53	47	NA
NA	NA	54	52	NA
NA	NA	57	56	NA
NA	NA	62	59	NA
NA	NA	72	64	NA
NA	NA	73	65	NA
NA	NA	80	66	NA
NA	NA	84	67	NA
NA	NA	87	71	NA
NA	NA	88	74	NA
NA	NA	89	81	NA
NA	NA	91	82	NA
NA	NA	NA	83	NA
NA	NA	NA	85	NA
NA	NA	NA	86	NA
NA	NA	NA	90	NA

## B. ACP avec k-means

Je pars des 5 clusters originels définis par la méthode des k-means.

J'observe les individus obtenus par l'ACP en “séparant” les éléments selon leurs classes.



### Interprétation des axes

Le 1er plan de l'ACP semble convenablement départager les individus en fonction des clusters retenus.

L'axe 1 semble opposer les individus des classes 1,2 et 5 (positivement corrélés avec cet axe) avec ceux des classes 3 et 4 (négativement corrélés avec cet axe).

L'axe 2 oppose parfaitement les individus des clusters 1,2 avec ceux de la classe 5 et 4. En revanche, les clusters 3 et 4 sont mal séparés par cet axe.

### Différence d'affectation entre Ward et EM

Des petites différences de sélection se retrouvent sur des individus situés proche des axes de l'ACP (surtout l'axe 1). Par exemple, les individus (1,6 30) sont dans la même classe avec Ward, mais dans 3 classes séparées avec kmeans. On observe le même phénomène pour les individus 13 et 78.

En étudiant les points possédant une contribution maximale sur les axes sélectionnés en terme de  $\cos^2$  et non significative sur les autres, j'obtiens :

	32	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
32	0.92	0.07	0.00	0.00	0.00	
28	0.01	0.80	0.01	0.04	0.10	
18	0.09	0.01	0.79	0.09	0.01	
84	0.05	0.02	0.08	0.76	0.05	
27	0.05	0.00	0.18	0.03	0.44	

Voici les individus sélectionnés selon les clusters allant de 1 à 5. Ce sont les mêmes que dans la méthode de Ward. L'interprétation des clusters est donc la même pour les 4 premiers.

nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
1.2895	1.7492	0.8254	0.9884	0.4259	-1.0988	1.4545
0.5596	0.5860	-0.7141	-0.4131	0.3040	0.3196	0.5935
-0.9387	-0.8300	-0.8385	-1.5178	-1.8507	0.2479	-0.8357
0.8246	0.7722	0.2693	-0.4303	0.2237	0.6808	-0.3708
-0.4235	-0.2984	0.3326	-0.0631	-0.3693	-0.0359	0.4041

Le cluster 5 conserve l'opposition du cluster 4 en terme de niveau d'étude/salaires, mais est un peu à l'opposé au niveau démographique : nombre de villes plus élevé que la moyenne, mais moins d'individus et moins d'habitants. (pour rappel, dans le cluster 4, les salaires étaient au-dessous de la moyenne alors que le niveau d'études est au-dessus. On observait le même phénomène pour le nombre d'habitants plutôt important par rapport au nombre de villes)

Une spécificité des kmeans est que par construction, il produit des clusters "parfairement sphériques", ce qui n'est pas toujours adapté à la structure des données. De plus, kmeans est un clustering crisp, c'est-à-dire qu'une observation est affectée à un et un seul cluster. Ce qui peut poser problème pour les individus situés "à la frontière" de plusieurs clusters.

Pour palier à ces caractéristiques de construction, nous allons tenter d'implémenter un algorithme EM (Expectation Maximization) dans un modèle bayésien.

## V. Modèles bayésiens (Expectation-Maximization algorithm) dans la population globale

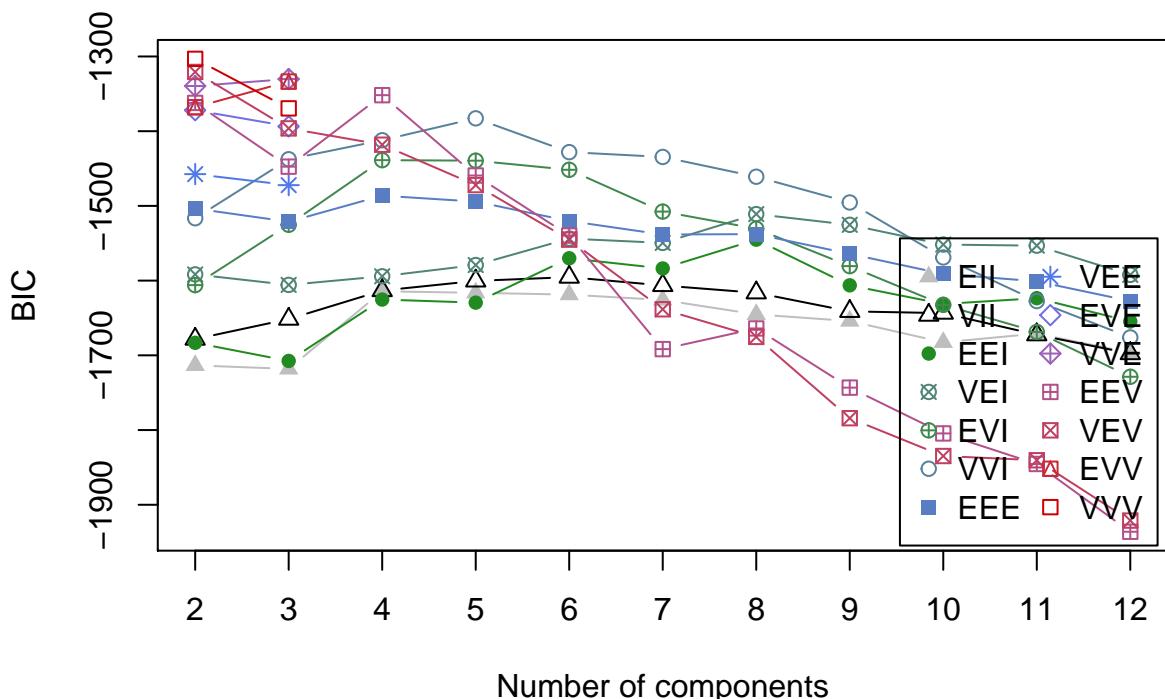
### A. Clustering

On va chercher à utiliser un algorithme de mélange de modèle appliquée au clustering (Gaussian Mixture Model).

L'idée principale est que : par rapport aux autres modèles type k-means, une classe ne possède pas l'exclusivité sur un individu, mais plutôt une probabilité d'appartenance. Ce point important se retrouve également à l'étape de mise à jour des paramètres ("step M") où la moyenne pondérée par les probabilités d'appartenance se retrouve "plus lissée" que dans les autres modèles précédemment

Je recherche le nombre de clusters adéquat proposé par l'algorithme EM. Pour cela, je choisis arbitrairement "un nombre minimal de classes non trivial" en proposant une décomposition avec au moins 4 clusters.

```
## fitting ...
## |
```



Selon le critère de pénalisation BIC, les 3 modèles les plus performants sont :

- VVV (ellipsoidal, varying volume, shape, and orientation) et VEV (ellipsoidal, equal shape) à 2 clusters
- VVE à 3 clusters (ellipsoidal, equal orientation)

Je choisis le 3ème modèle, soit celui possédant 3 clusters

```
## fitting ...
```

```

##   |
## [1] 1 2 2 2 1 3 2 2 2 2 1 1 3 1 2 1 1 1 2 2 1 2 1 1 1 2 3 2 2 3 3 2 3 3 3 2 1
## [39] 3 1 1 1 3 2 3 1 2 1 1 3 1 1 2 1 3 2 3 3 2 3 3 1 3 1 1 2 1 3 3 3 2 1 1 1 3 3
## [77] 3 3 3 1 2 1 2 3 2 1 1 1 2 2 3 3 3 3 3

## classif_em
## 1 2 3
## 36 28 32

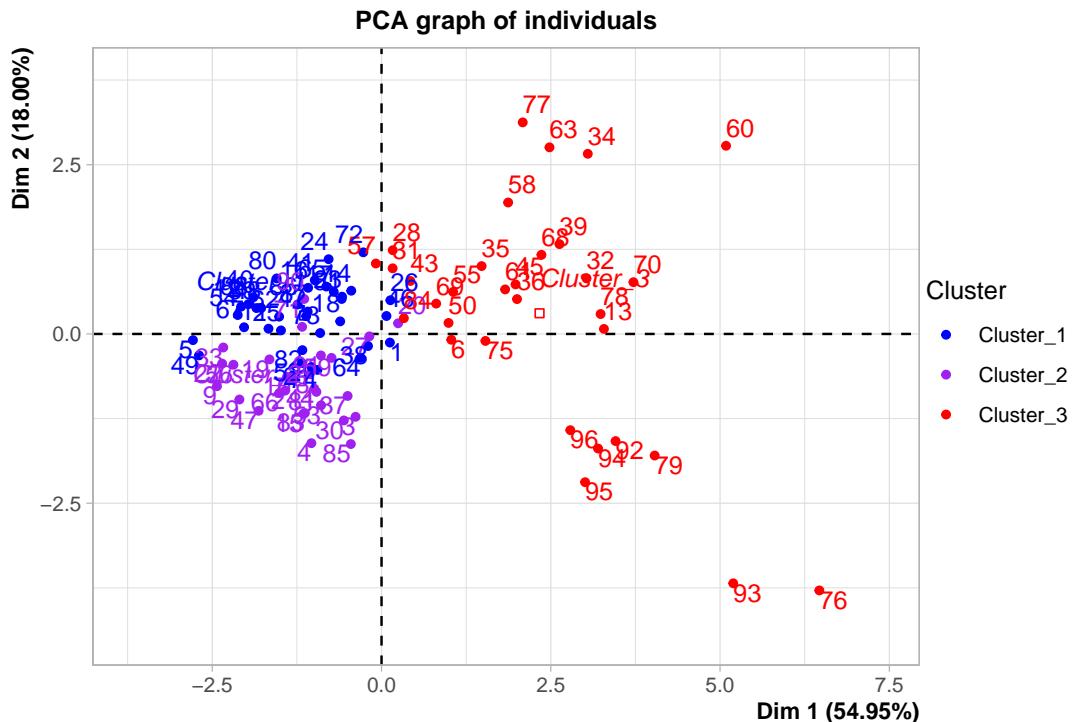
```

Voici la décomposition par clusters

	Cluster 1	Cluster 2	Cluster 3
1	2	6	
5	3	13	
11	4	28	
12	7	31	
14	8	32	
16	9	34	
17	10	35	
18	15	36	
21	19	39	
23	20	43	
24	22	45	
25	27	50	
26	29	55	
38	30	57	
40	33	58	
41	37	60	
42	44	61	
46	47	63	
48	53	68	
49	56	69	
51	59	70	
52	66	75	
54	71	76	
62	81	77	
64	83	78	
65	85	79	
67	90	84	
72	91	92	
73	NA	93	
74	NA	94	
80	NA	95	
82	NA	96	
86	NA	NA	
87	NA	NA	
88	NA	NA	
89	NA	NA	

## B. ACP avec EM

J'observe les individus obtenus par l'ACP en “séparant” les éléments selon leurs classes.



### Interprétation des axes

Dans le premier plan de l'ACP, l'axe 1 semble opposer les individus de la classe 3 (positivement corrélés avec cet axe) avec ceux des classes 1 et 2 (négativement corrélés avec cet axe).

L'axe 2 oppose plutôt les individus des clusters 1 et 2.

### Différence d'affectation entre EM et Kmeans

On observe le même type de différence d'affectation pour certains individus entre les algorithmes Kmeans et EM. Ainsi, les individus (9, 53, 30) qui semblaient “un peu perdus dans une classe étrangère” avec la méthode Ward se retrouvent de la classe de leurs plus proches voisins avec EM. De même que dans la section précédente, certains points proches des axes (les individus (5,49) pour l'axe 1, (28,81) pour l'axe 2) semblent être attribués à 2 familles de classe différente.

On observe le même type de différence d'affectation pour certains individus entre les algorithmes Kmeans et EM. Ainsi, les individus (9, 53, 30) qui semblaient “un peu perdus dans une classe étrangère” avec la méthode Ward se retrouvent de la classe de leurs plus proches voisins avec EM. De même que dans la section précédente, certains points proches des axes (les individus (5,49) pour l'axe 1, (28,81) pour l'axe 2) semblent être attribués à 2 familles de classe différente.

En étudiant les points possédant une contribution maximale sur les axes sélectionnés en terme de  $\cos^2$  et non significative sur les autres, j'obtiens :

	32	Dim.1	Dim.2	Dim.3
32	0.92	0.07	0.00	
28	0.01	0.80	0.01	
18	0.09	0.01	0.79	

Voici les individus sélectionnés selon les clusters allant de 1 à 3.

Ce sont les mêmes que dans la méthode de Ward. L'interprétation des clusters est encore une fois la même.

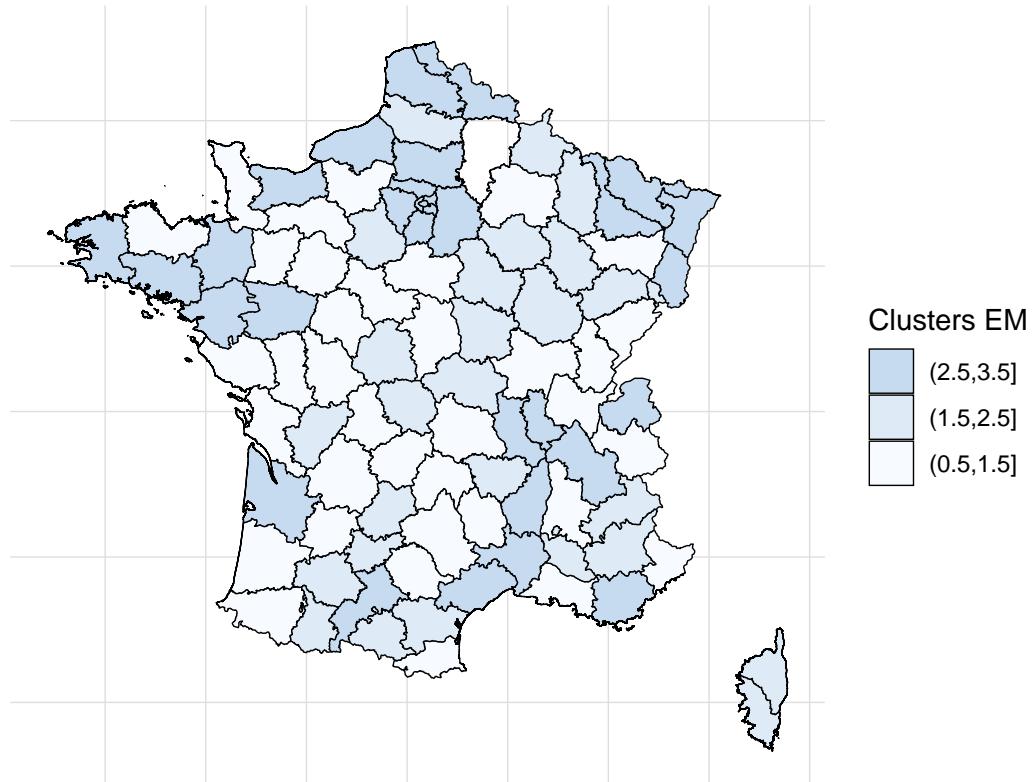
Pour rappel, voici les individus caractéristiques des clusters en terme de  $R^2$ .

nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
1.2895	1.7492	0.8254	0.9884	0.4259	-1.0988	1.4545
0.5596	0.5860	-0.7141	-0.4131	0.3040	0.3196	0.5935
-0.6320	-0.5044	-0.2005	-0.3347	1.7895	0.1927	-0.1727

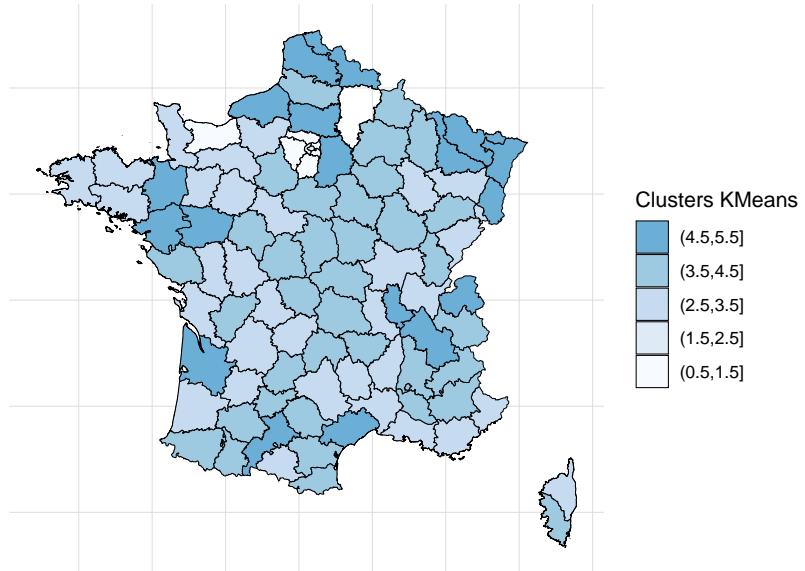
## V. Cartographie des différentes méthodes de clustering dans la population globale

J'effectue une carte descriptive avec le nombre de clusters définis par chacun des algorithmes : 3 clusters définis par l'algorithme EM, 5 pour K-means et 4 pour Ward

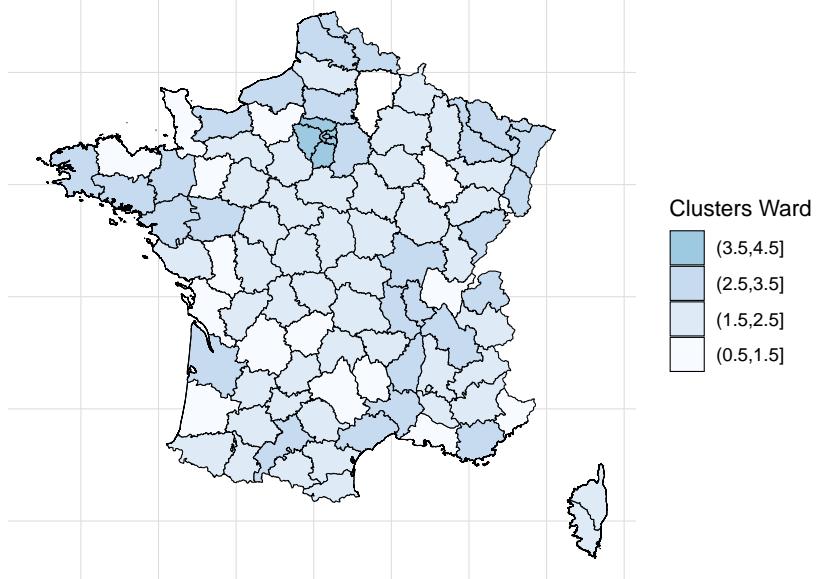
représentation des départements selon les clusters EM



représentation des départements selon les clusters Kmeans



représentation des départements selon les clusters Ward



Dans l'ensemble, on retrouve un nombre important de similitudes de la carte du “Nombre d’individus par quantile” dans la carte de Ward. Par exemple, l’Ile-de-France privée de la Seine-et-Marne est un cluster à elle tout seule. cela reste cohérent avec les covariables les plus discriminantes sélectionnées par l’ANOVA. La carte K-means, avec beaucoup de covariables discriminantes dont le taux de femmes présente des similitudes avec la carte du “taux de femmes par quantile”.

## VI. Étude de la population active

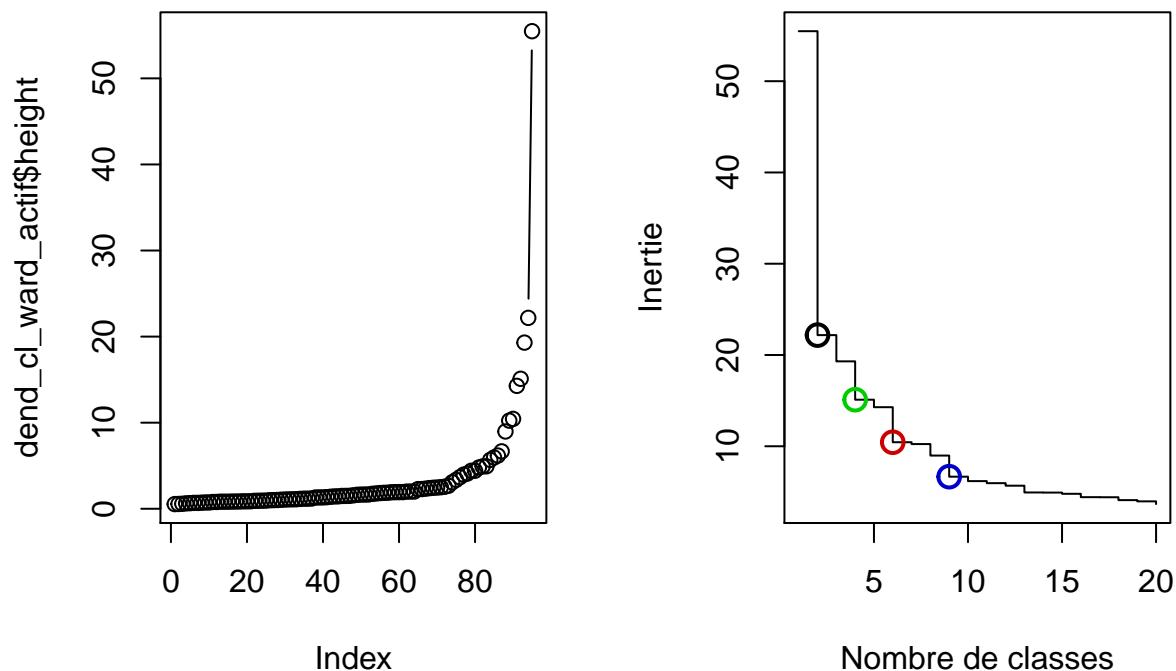
### A. Introduction

Je vais désormais m'attacher à étudier les différents clusterings associés à la population globale. Les principaux critères d'analyse, ainsi que l'architecture de présentation sont amplement détaillés dans l'étude de la population globale. Je propose donc de présenter et commenter succinctement les résultats que j'estime les plus importants ou significatifs.

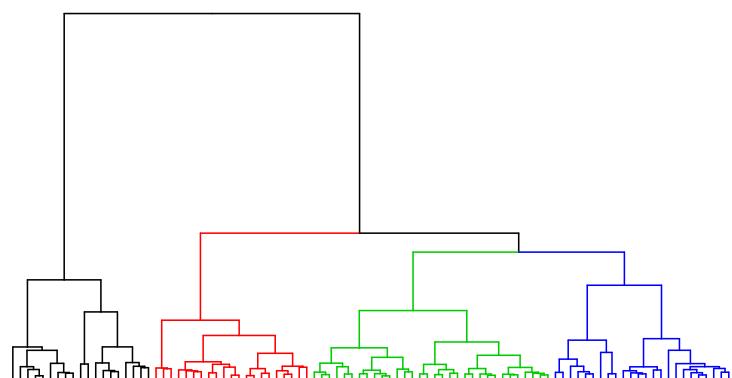
### B. Clustering

je délaisse la CAH qui comme précédemment décompose relativement mal les classes.

#### 1. Méthode de Ward

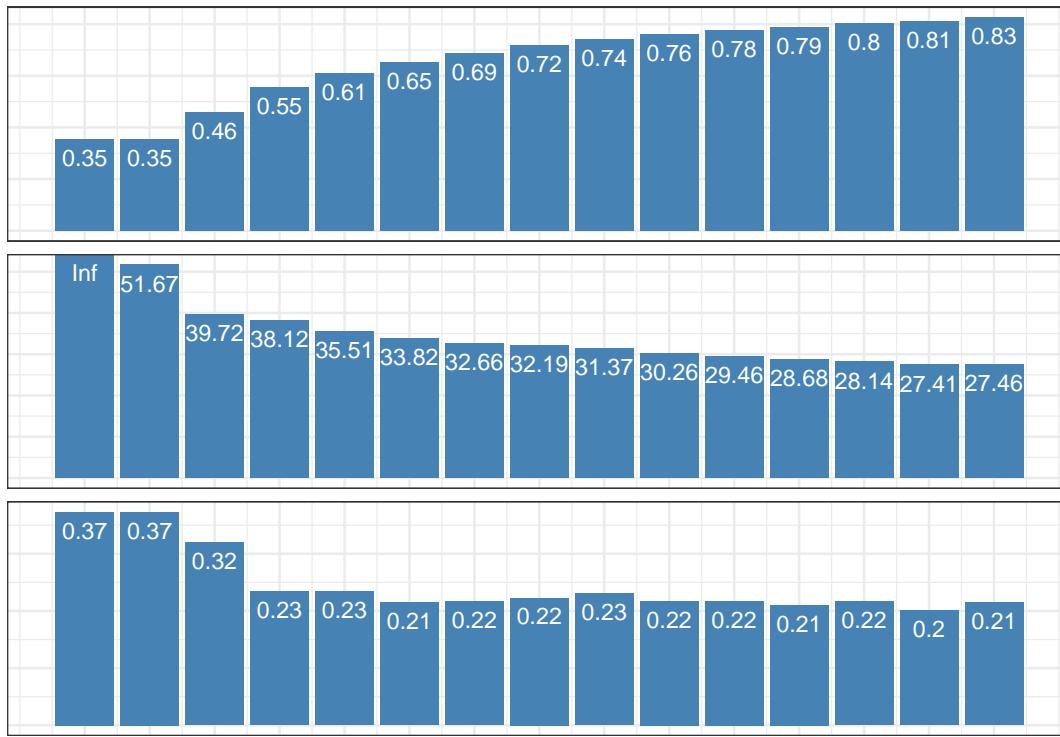


La méthode de Ward propose raisonnablement 4 ou 6 clusters avec un clustering toujours aussi équilibré.



## 2. Méthode de Kmeans

Observons les résultats fournis par les indicateurs de sélection du nombre optimal de cluster pour le kmeans.



Là encore, les indicateurs ne sont pas du tout d'accord

Il semblerait que 9 clusters fournissent une bonne valeur pour le pourcentage de variance expliquée : \* 2 pour les indices silhouette \* l'indice de Calinski-Harabasz propose, quant à lui, 2 ou 3 clusters

Notons qu'à partir de 4 clusters, le dernier indice ne varie quasi plus.

Afin "d'équilibrer un peu les indicateurs", j'effectue un k-means avec 4 clusters et une initialisation aléatoire.

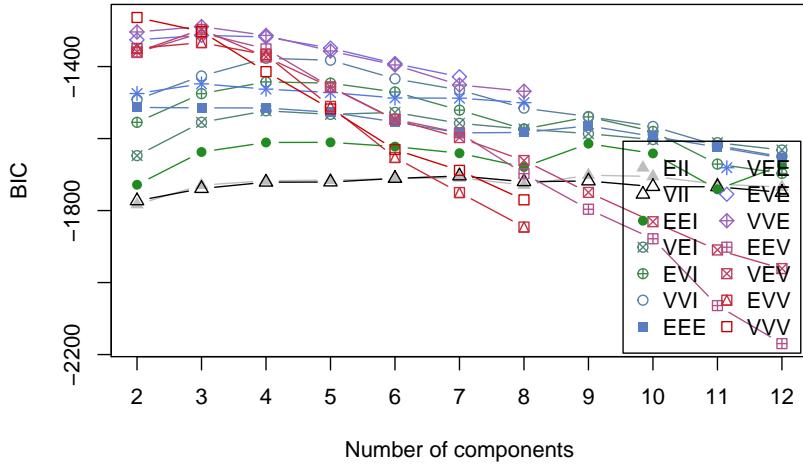
```
## [1] 5 34 21 36

##   nb_habitants nb_individus années_étude salaire_moyen taux_femmes    age_moyen
## 1     1.8917249     1.5153971     2.7258439      3.1245167    0.41554455 -0.18179103
## 2     -0.6012553    -0.6431759     -0.4005775     -0.3431695   -0.99860740 -0.06529041
## 3     1.2165078     1.3898429     0.4422560      0.3069643   0.03988223 -0.36615404
## 4     -0.4045168    -0.4137696     -0.2582489     -0.2889187   0.86215005  0.30050178
##   nb_villes
## 1 -0.4456116
## 2 -0.5594515
## 3  1.3944641
## 4 -0.2231760

## [1] 4 2 4 2 4 4 2 4 2 2 4 4 3 2 2 4 4 4 2 2 4 2 4 4 2 4 4 2 2 4 3 2 3 3 2 2 2
## [39] 3 2 4 4 4 2 3 2 2 4 4 3 4 4 4 2 3 2 4 3 4 3 3 4 3 4 2 4 2 3 3 3 4 4 2 2 3 1
## [77] 3 3 1 4 2 2 2 4 2 2 2 4 1 1 3 1 3
```

### 3. Algorithme EM

```
## fitting ...
## |
```



Selon le critère de pénalisation BIC, les 3 modèles les plus performants sont :

- VVV (ellipsoidal, varying volume, shape, and orientation) à 2 clusters
- VVE (ellipsoidal, equal orientation) et VEV (ellipsoidal, equal shape) à 3 clusters

Je choisis le modèle avec un type de variance VVE à 3 clusters.

```
## fitting ...
## |
```

### C. Analyse de la variance

Voici les résultats des différentes ANOVA effectuées selon les clustering

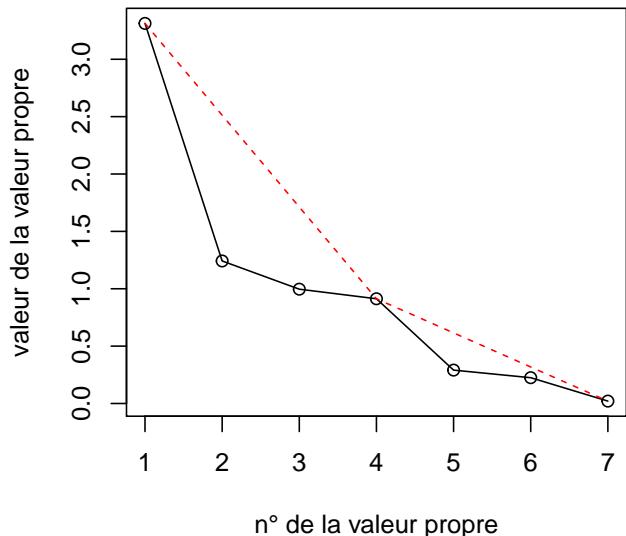
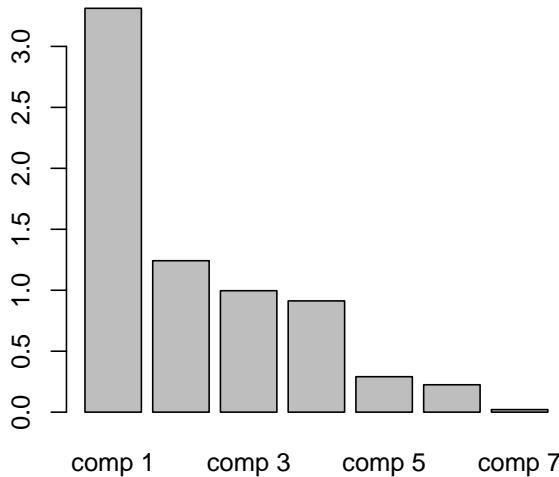
	Ward	K-Means	EM
nb_habitants	6.16e-09 ***	0.02452 *	1.42e-08 ***
nb_individus	0.13560	0.00314 **	0.1833
années_étude	0.44391	0.00362 **	0.3455
salaire_moyen	0.64698	0.00116 **	0.5790
taux_femmes	0.67261	2e-16 ***	0.7036
age_moyen	0.00217 **	0.00415 **	0.2961
nb_villes	0.16997	0.00624 **	0.0062 **

- avec la méthode de Ward, seuls le nombre d'habitants et l'âge moyen s'avèrent être des covariables discriminantes pour une classification à 4 clusters
- avec la méthode de K-Means, toutes les covariables du dataset sont discriminantes pour une classification à 4 clusters également !
- enfin, les variables discriminantes fournies par l'algorithme EM sont les mêmes que celles fournies par la méthode de Ward.

## D. ACP pour les actifs

### 1. Pourcentage de variance expliquée.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.3123816	47.3197375	47.31974
comp 2	1.2421392	17.7448455	65.06458
comp 3	0.9962360	14.2319435	79.29653
comp 4	0.9126064	13.0372337	92.33376
comp 5	0.2909087	4.1558379	96.48960
comp 6	0.2248470	3.2121000	99.70170
comp 7	0.0208811	0.2983021	100.00000

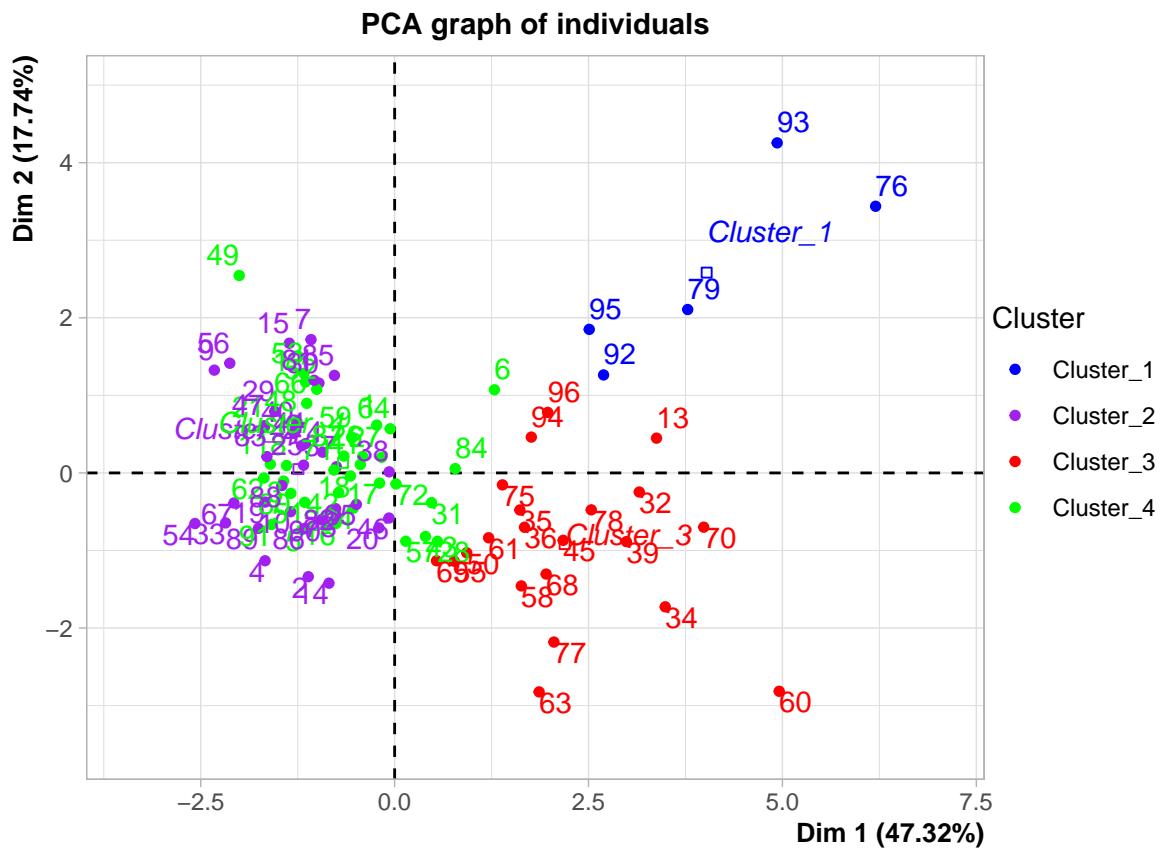
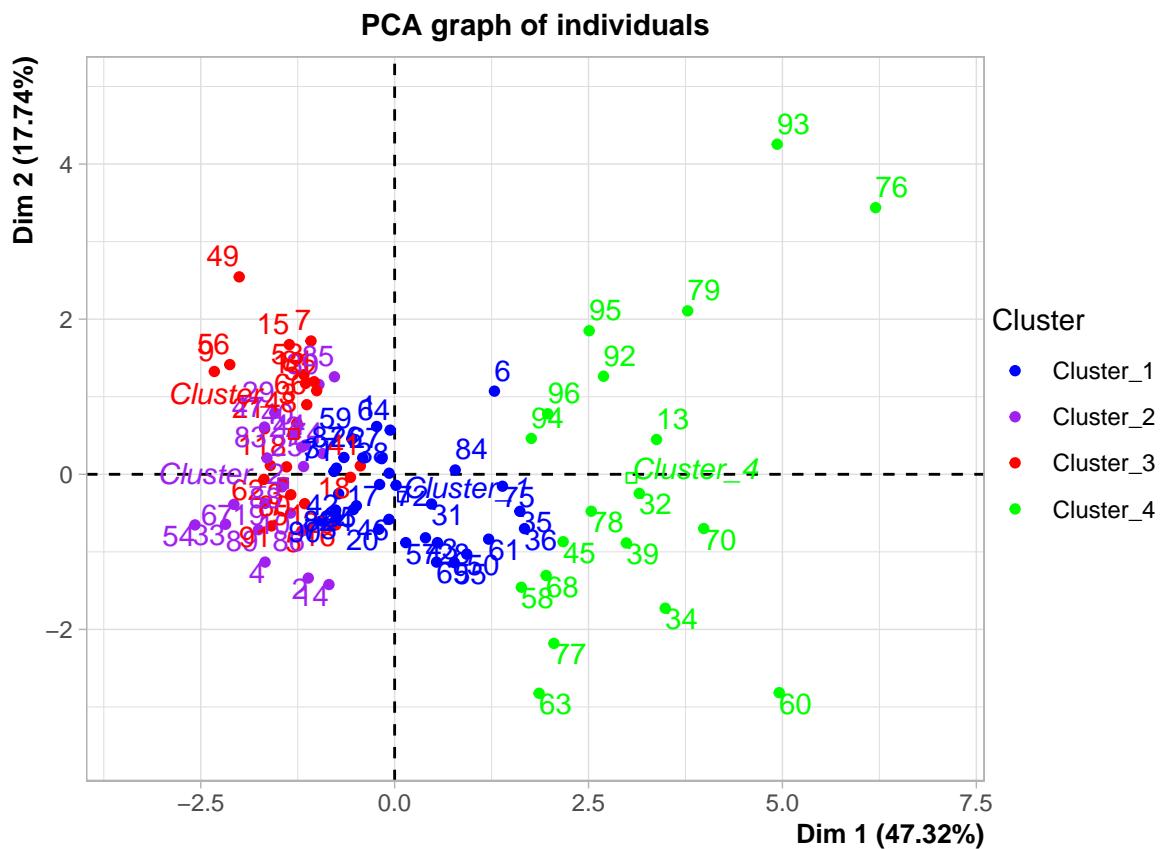


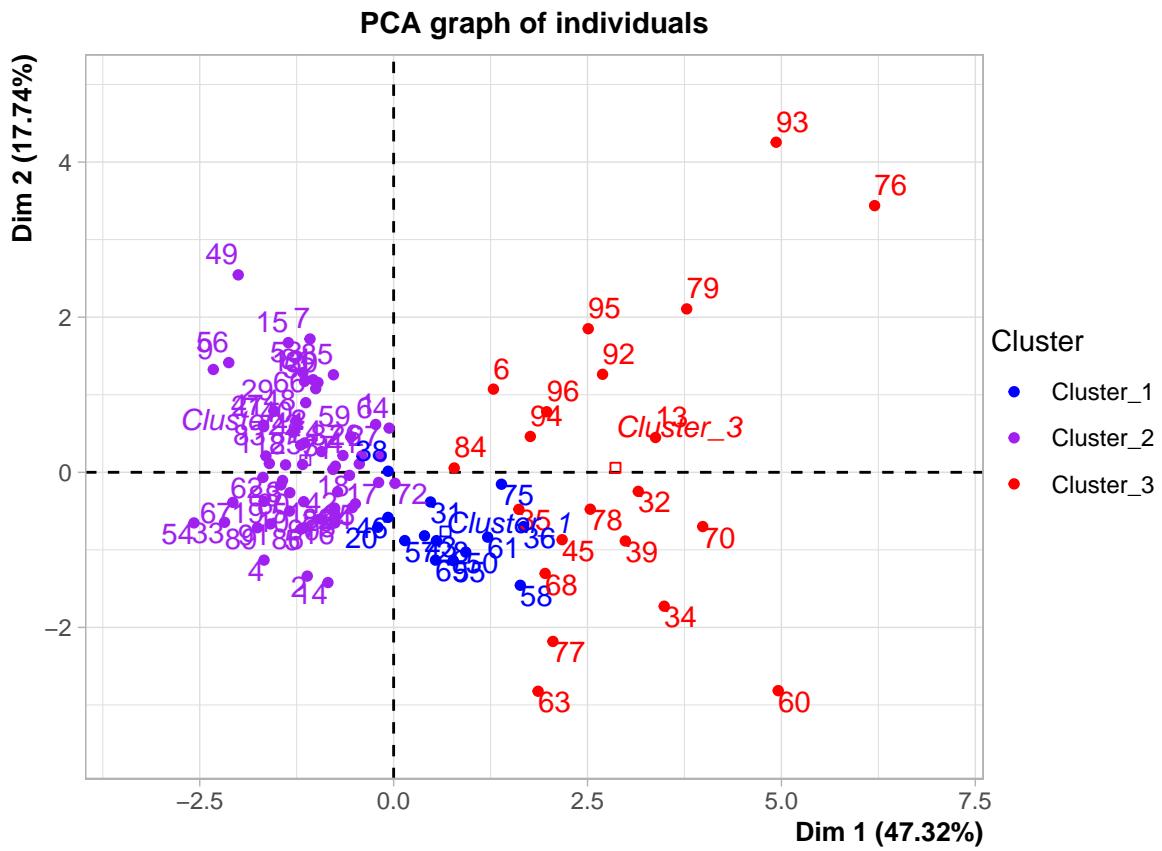
La structure est un peu différente de celle observée pour la population globale.

Les 2 premiers axes explique 65 % de la variance. La règle du coude m'enclinerait à choisir 4 axes où cette fois 92 % de la variance est expliquée.

### 2. Description des axes selon les individus

Le point de départ est le clustering de chacun des modèles étudiés. Les graphiques suivants représentent respectivement : la méthode de Ward, puis K\_Means avec 4 clusters, enfin la méthode EM avec 3 clusters.





Avec la **méthode Ward**, l'axe 1 semble opposer les individus de la classe 4 (positivement corrélés avec cet axe) avec ceux des classes 2 et 3 (négativement corrélés).

L'axe 2, lui, est difficilement interprétable.

Avec la **méthode Kmeans**, l'axe 1 semble opposer les individus de la classe 1 et 3 (positivement corrélés avec cet axe) avec ceux (négativement corrélés) des classes 2 et 4 dans une certaine mesure.

L'axe 2 oppose principalement les individus des classes 1 et 3.

Avec la **méthode EM**, l'axe 1 semble opposer les individus de la classe 1 et 3 (positivement corrélés avec cet axe) avec ceux (négativement corrélés) de la classe 2.

Sur l'axe 2, tous les individus de la classe ont des valeurs inférieures à la moyenne.

Le 1er axe de l'ACP présenté sur chaque graphique met en évidence des classes assez différentes même avec le même nombre de clusters (cf. Ward contre Kmeans). Par exemple, les départements 77 et 92 sont dans le même cluster avec Ward, contrairement à Kmeans. De même, la répartition EM avec son nombre de classes inférieur ne regroupe pas 2 classes en une nouvelle, mais à plutôt tendance à casse les 2 structures précédentes en créant sa propre agrégation.

J'étudie les points possédant une contribution maximale sur chacun des axes sélectionnés en terme de  $\cos^2$  et non significative sur les autres.

	32	Dim.1	Dim.2	Dim.3	Dim.4
32	0.96	0.01	0.00	0.00	
43	0.14	0.60	0.21	0.01	
44	0.21	0.02	0.74	0.00	
5	0.07	0.03	0.07	0.83	

Voici les individus sélectionnés selon les clusters allant de 1 à 4

nb_habitants	nb_individus	années_étude	salaire_moyen	taux_femmes	age_moyen	nb_villes
1.2832	1.6836	1.1048	1.4219	0.1559	-0.5201	1.5057
0.2221	0.4979	-0.2919	-0.3602	0.4705	-0.2518	0.5567
-0.8138	-0.7971	0.1325	0.1148	-1.9137	-0.7161	-0.8401
-0.9361	-0.8841	-0.0230	-1.2351	3.0115	-2.4093	-1.2453

L'ACP apporte les éléments d'interprétation suivants.

Il semble que :

- tout comme pour la population globale, le cluster 1 possède la même interprétation. Il caractérise les départements denses en nombre de villes, en nombre d'habitants avec une population plutôt jeune (bien en-dessous de la moyenne). Ces individus ont un taux de femmes, un salaire et un niveau d'étude au-dessus de la moyenne ()
- le cluster 2,3 sont plus délicats à interpréter car les individus choisis possèdent entre 23 et 35 % de leur  $R^2$  sur les autres axes. Néanmoins, ils sont clairement opposés (par rapport à la moyenne) sur l'ensemble des covariables excepté le taux de femmes. Ainsi, le cluster 2 pourrait représenter les départements modérément peuplés plutôt "féminins", "plus jeunes" et dont les individus ont un salaire et un niveau d'étude en-dessous de la moyenne. Encore plus féminin, le cluster 3 pourrait représenter son opposé sur les autres indicateurs.
- le cluster 4 semble caractériser les départements "très jeunes extrêmement féminins" possédant un salaire moyen assez faible.

```
## [1] 0.09506183
```

Le correlation entre le salaire moyen et le taux de femmes dans le département de la population active est quasi-nul, laissant peut-être entendre que la tendance présente dans le cluster 4 n'est pas systématique...

## VII. Cartographie des des différentes méthodes de clustering dans la population active

J'effectue une carte descriptive avec le nombre de clusters définis par chacun des algorithmes : 3 clusters définis par l'algorithme EM, 4 pour K-means et Ward

```
##      Dep EM KMeans Ward
## 1    01  2     4   1
## 2    03  2     2   2
## 3    04  2     4   3
## 4    05  2     2   2
## 5    06  2     4   3
## 6    07  3     4   1
## 7    08  2     2   3
## 8    09  2     4   3
## 9    10  2     2   3
## 10   11  2     2   2
## 11   12  2     4   3
## 12   13  2     4   3
## 13   14  3     3   4
## 14   15  2     2   2
## 15   16  2     2   3
## 16   17  2     4   3
## 17   18  2     4   1
## 18   19  2     4   3
## 19   20  2     2   2
## 20   21  1     2   1
## 21   22  2     4   3
## 22   23  2     2   2
## 23   24  2     4   3
## 24   25  2     4   1
## 25   26  2     2   2
## 26   27  2     4   1
## 27   28  2     4   1
## 28   29  1     4   1
## 29   2A  2     2   2
## 30   2B  2     2   2
## 31   30  1     4   1
## 32   31  3     3   4
## 33   32  2     2   2
## 34   33  3     3   4
## 35   34  3     3   1
## 36   35  1     3   1
## 37   36  2     2   1
## 38   37  1     2   1
## 39   38  3     3   4
## 40   39  2     2   2
## 41   40  2     4   3
## 42   41  2     4   1
## 43   42  1     4   1
## 44   43  2     2   2
## 45   44  3     3   4
## 46   45  1     2   1
## 47   46  2     2   2
```

```

## 48 47 2      4      3
## 49 48 2      4      3
## 50 49 1      3      1
## 51 50 2      4      3
## 52 51 2      4      3
## 53 52 2      4      3
## 54 53 2      2      2
## 55 54 1      3      1
## 56 55 2      2      3
## 57 56 1      4      1
## 58 57 1      3      4
## 59 58 2      4      1
## 60 59 3      3      4
## 61 60 1      3      1
## 62 61 2      4      3
## 63 62 3      3      4
## 64 63 2      4      1
## 65 64 2      2      1
## 66 65 2      4      3
## 67 66 2      2      2
## 68 67 3      3      4
## 69 68 1      3      1
## 70 69 3      3      4
## 71 70 2      4      1
## 72 71 2      4      1
## 73 72 2      2      1
## 74 73 2      2      2
## 75 74 1      3      1
## 76 75 3      1      4
## 77 76 3      3      4
## 78 77 3      3      4
## 79 78 3      1      4
## 80 79 2      4      3
## 81 80 2      2      3
## 82 81 2      2      1
## 83 82 2      2      2
## 84 83 3      4      1
## 85 84 2      2      2
## 86 85 2      2      2
## 87 86 2      4      1
## 88 87 2      2      2
## 89 88 2      2      2
## 90 89 2      2      1
## 91 90 2      4      3
## 92 91 3      1      4
## 93 92 3      1      4
## 94 93 3      3      4
## 95 94 3      1      4
## 96 95 3      3      4

## # A tibble: 96 x 13
##       Dep   nb_habitants nb_individus années_étude salaire_moyen taux_femmes
##       <fct>        <dbl>        <int>        <dbl>        <dbl>        <dbl>
## 1 01          412586         216       10.0      22380.      51.4

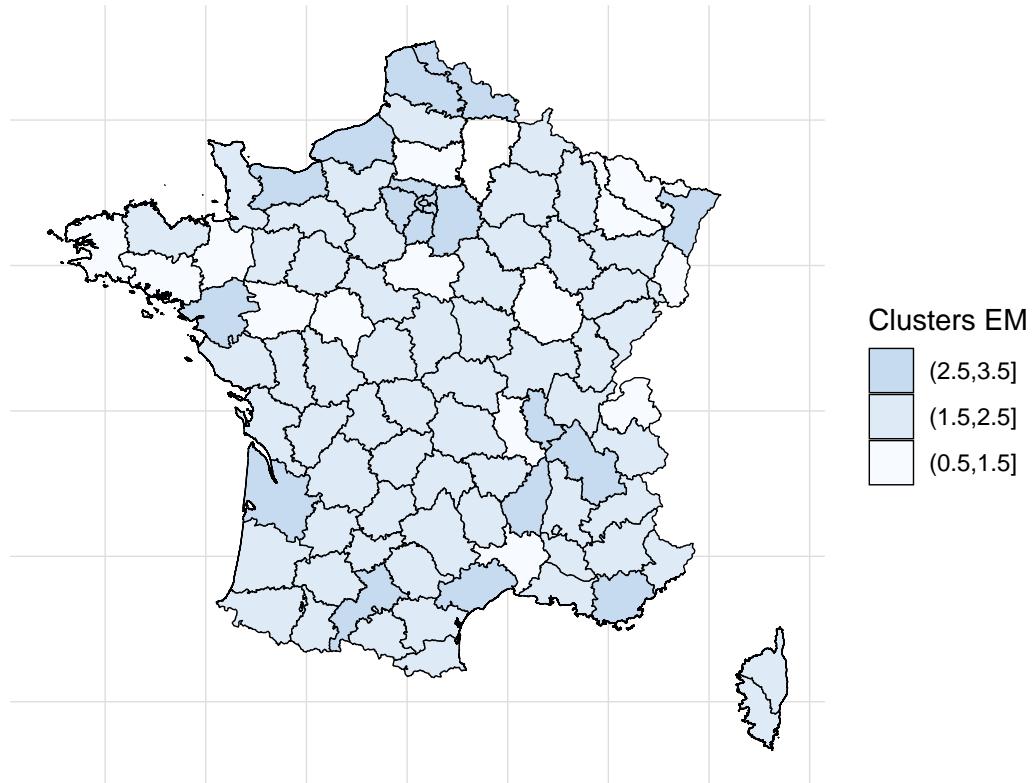
```

```

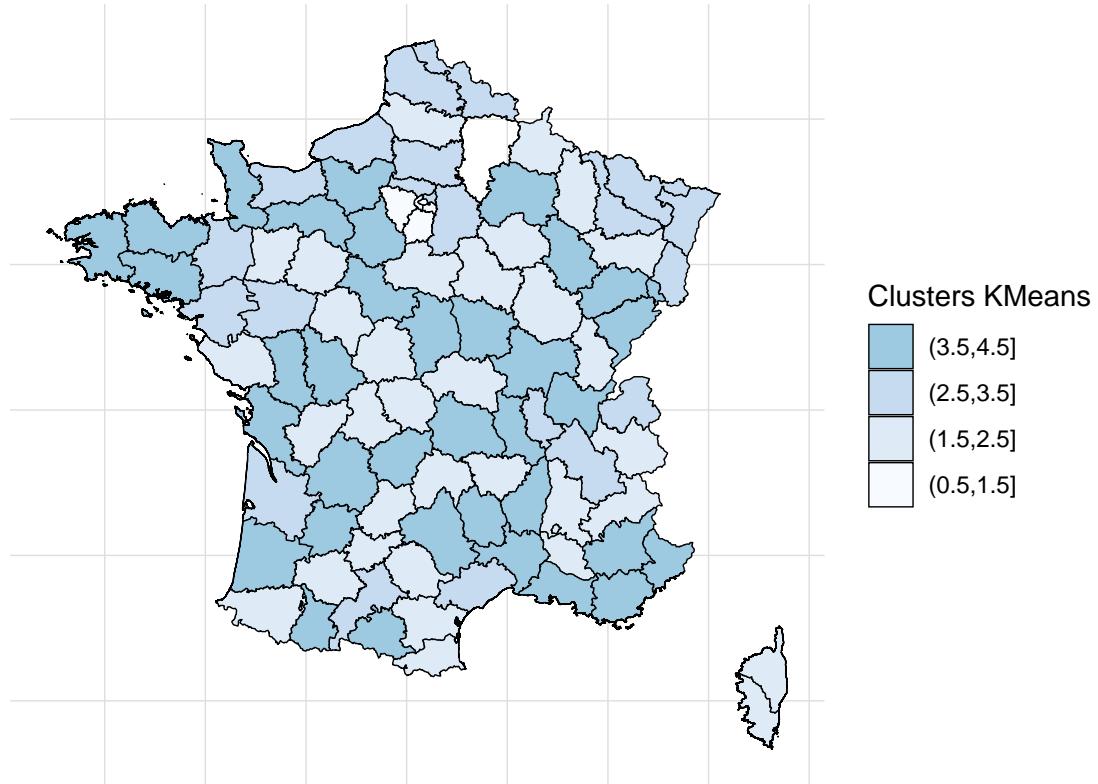
## 2 02      233540      75      10.2      18817.      42.7
## 3 03      214110     108      9.37      20784.      54.6
## 4 04      74105       23      10.2      17350.      47.8
## 5 05      61422       16      10.0      18770.      62.5
## 6 06     1067261      968      10.5      22058.      54.1
## 7 07     160906      109      10.1      22428.      45.9
## 8 08     133855       76      9.84      19932.      53.9
## 9 09     80519        50      9.51      20982.      40
## 10 10    171950      88      10.4      18458.      43.2
## # ... with 86 more rows, and 7 more variables: age_moyen <dbl>,
## #   nb_villes <int>, NAME_2 <fct>, REG <int>, EM <fct>, KMeans <fct>,
## #   Ward <fct>

```

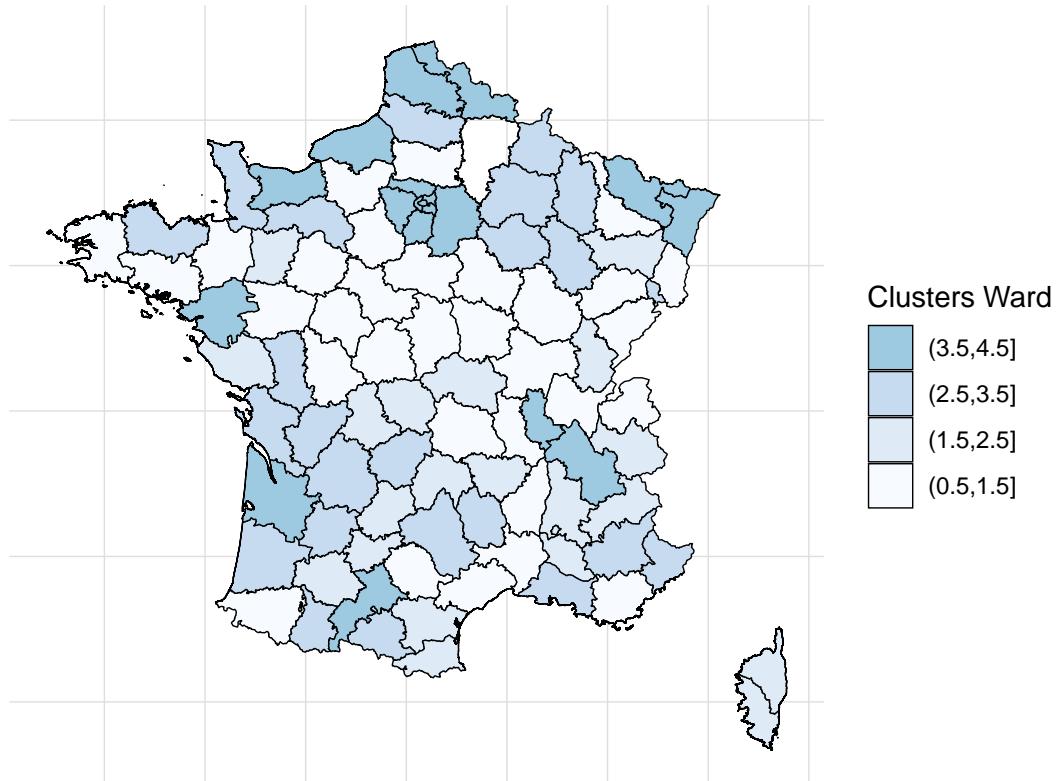
### représentation des départements selon les clusters EM



représentation des départements selon les clusters Kmeans



représentation des départements selon les clusters Ward



Dans l'ensemble, on retrouve un nombre important de similitude de la carte du “Nombre d’individus par quantile” dans les cartes EM et Ward, cela reste cohérent avec les covariables les plus discriminantes sélectionnées par l’ANOVA.

La carte K-means, avec toutes ses covariables discriminantes est moins aisée à interpréter.