Licence de Mathématiques
3M248 - Projet

Emmanuel Barillot - 3370161

6 mai 2018

1 Introduction

Notre jeu de données provient des bases de données de l'INSEE. Il se compose d'un fichier Excel qui contient des informations sur les 36000 (environ) communes françaises. Il a été publié le 12 décembre 2014 à l'adresse : https://www.data.gouv.fr/s/resources/data-insee-sur-les-communes/20141212-105948/MDB-INSEE-V2.xls
Nous nous proposons d'analyser ces données de façon à chercher des liens entre les caractéristiques démographiques, les caractéristiques économiques des départements français, et le niveau de vie mesuré par le niveau moyen des salaires. Nous voulons notamment déterminer l'importance du caractère rural ou non d'un département selon ces trois aspects.

2 Présentation des données

Le fichier brut contient plus de 36000 lignes et 100 colonnes. Certaines variables ne nous sont pas utiles comme le code de la région. Chaque ligne est identifiée par le code postal de la commune.
Il y a une variable qualitative que nous allons exploiter : l'aspect ruralité d'une commune. Certaines variables quantitatives sont relatives au bassin de vie, dont la définition n'est pas donnée dans ce jeu de données. Certaines variables semblent redondantes, nous n'avons gardé que celles qui nous semblaient les plus pertinentes, sans plus d'explications de la part de l'INSEE sur ces "doublons" apparents.
Chaque ligne contient le code du département auquel elle appartient. Nous allons concentrer notre étude sur les 100 département présents dans le fichier, en procédant par agrégation des variables quantitatives.
Il nous a fallu prendre des précautions pour l'agrégation des données : certaines variables sont en réalité relatives aux communes, d'autres aux départements, d'autres aux bassins de vie et d'autres encore aux régions. Certaines sont des pourcentages et d'autres des quantités.
Nous procédons d'abord à une transformation simple des données brutes :

- suppression des variables non retenues
- agrégation des données par département
- séparation des variables en deux familles : caractéristiques démographiques et caractéristiques économiques.

Puis nous calculons le nombre de communes rurales par département en comptant le nombre de communes de moins de 10000 habitants par département. Nous obtenons finalement les jeux de données suivants, avec leur variables respectives :

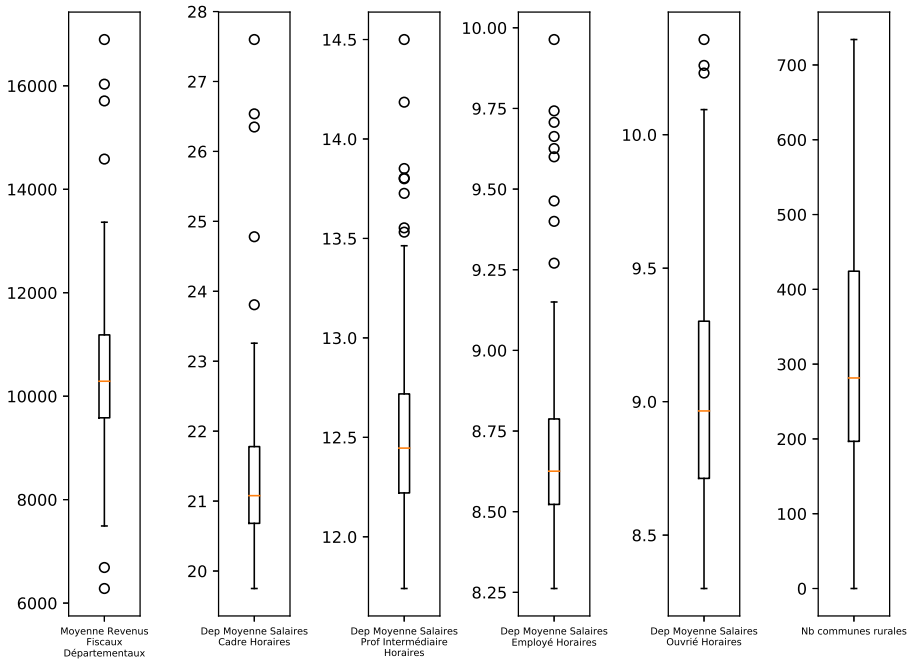
- variables relatives aux salaires :
  - Moyenne Revenus Fiscaux Départementaux,
  - Moyenne Salaires Cadre Horaires,
  - Moyenne Salaires Prof Intermédiaire Horaires,
  - Moyenne Salaires Employé Horaires,

    - Moyenne Salaires Ouvrié Horaires (avec la faute d'orthographe, dans le fichier),
- variables relatives à la population d'entreprises et d'organismes :
    - Nb Education, santé, action sociale,
    - Nb Entreprises,
    - Nb Création Entreprises,
- variables relatives à la population :
    - Population,
    - Evolution Population,
    - Nb Ménages,
    - Nb propriétaire,
    - Nb Etudiants.

2.1 Exploration

2.1.1 Salaires

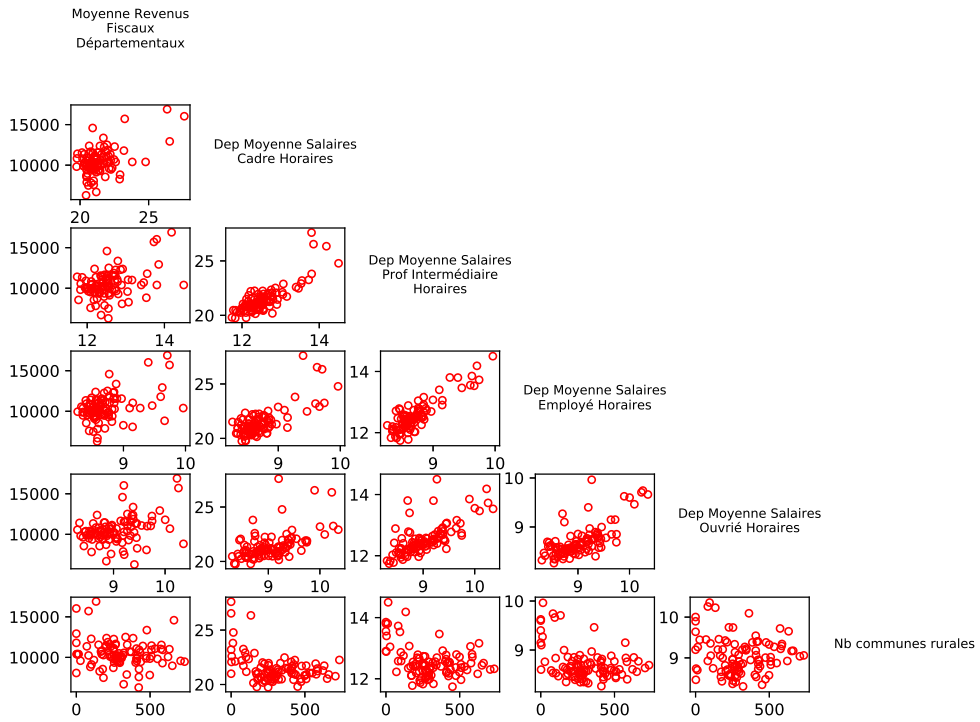
Une description des variables de type "salaire" dans des boxplots :



La dispersion des salaires horaires des cadres semblent plus importante que pour les autres catégories. Il y a des départements dans lesquels les salaires des cadres sont bien plus élevés qu'ailleurs. Le phénomène est moins marqué pour les autres catégories.

Une description en termes de corrélations visuelles entre couples de variables (valeurs brutes) :

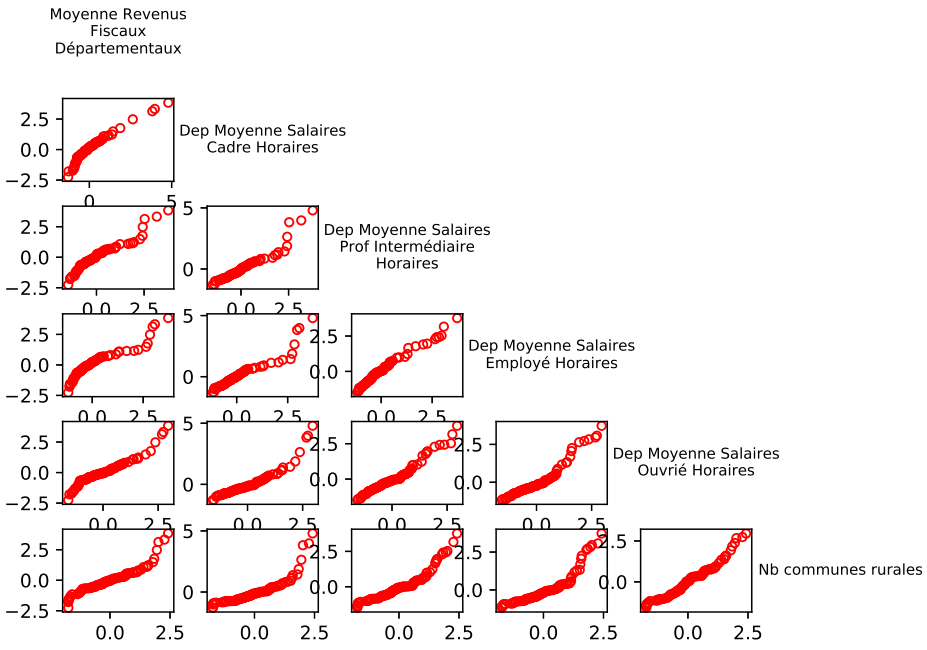
Plots variables salaires



Il apparait déjà que beaucoup de variables semblent grossièrement corrélées linéairement. la variable "Nb communes rurales" a un comportement différent : elle ne semble pas avoir de relation de linéarité évidente avec les autres.

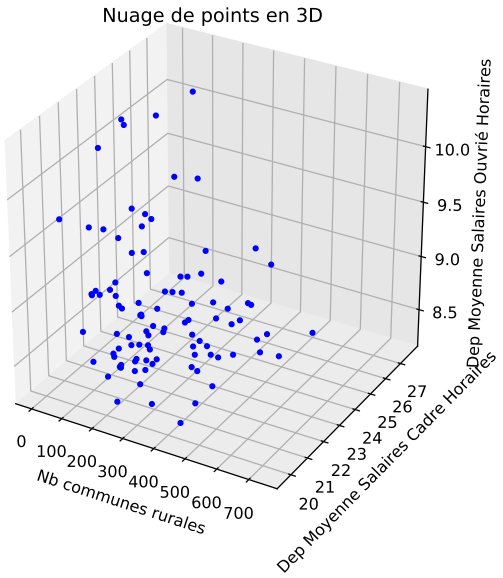
Une description en termes de QQ-plots :

QQ-plots variables salaires



Les QQ-plots montrent que certaines relations entre variables sont linéaires visuellement, mais que d'autres ont une forme sigmoïde ou en épaulement.

Nous avons choisi une représentation 3D particulière en ne retenant que 3 variables :

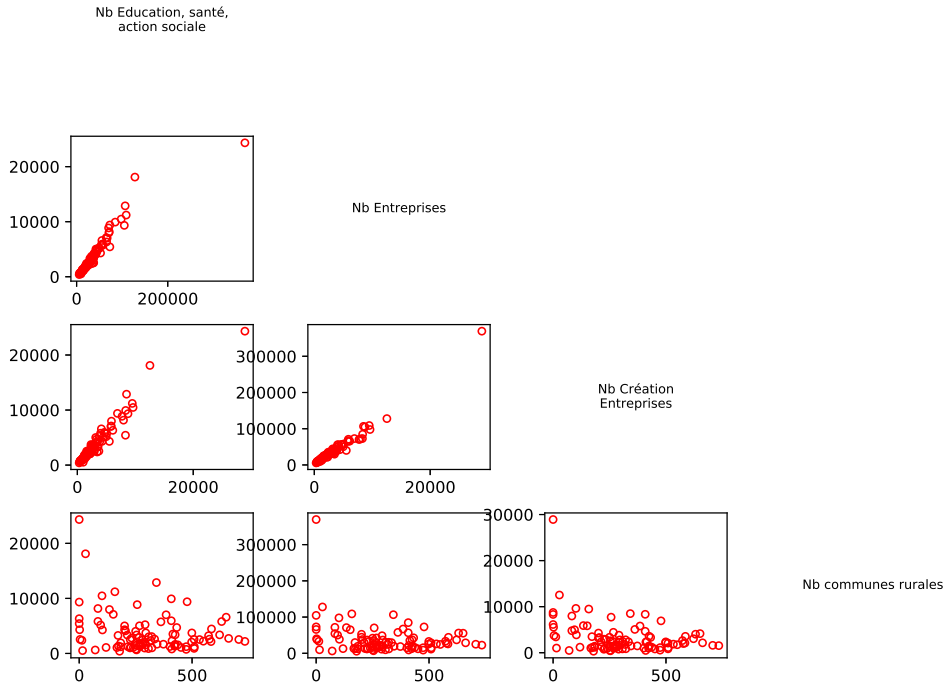


On voit par exemple, que que les points se regroupent plutôt vers les valeurs basses des salaires moyens et que les salaires les plus élevés se trouvent plutôt dans les départements les moins ruraux (peu de communes rurales dans le département). La rotation interactive permet de mieux voir.

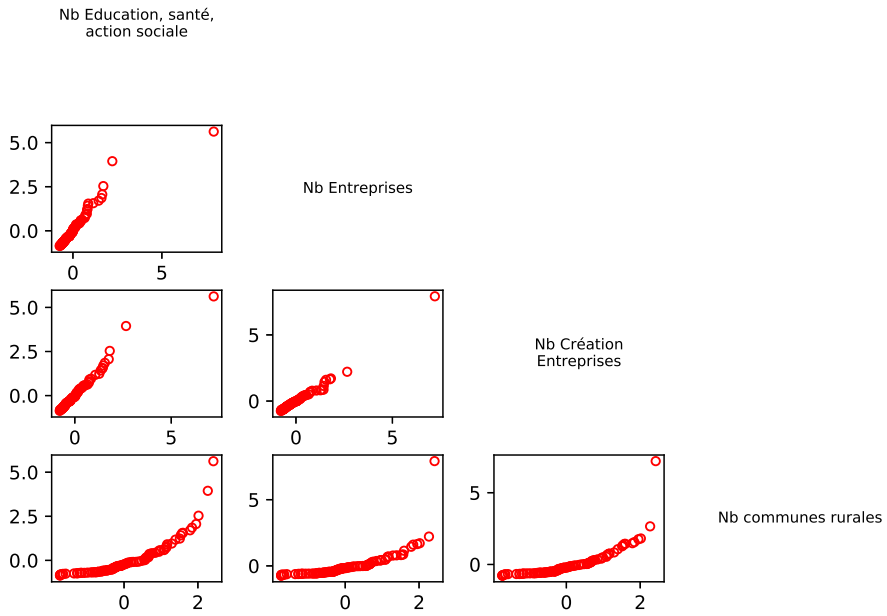
2.1.2 Entreprises

Nous adoptons un cheminement exploratoire analogue aux variables "Salaires" pour les variables "entreprises", à l'exception des boxplots qui nous apportent peu d'informations.

Plots variables entreprises

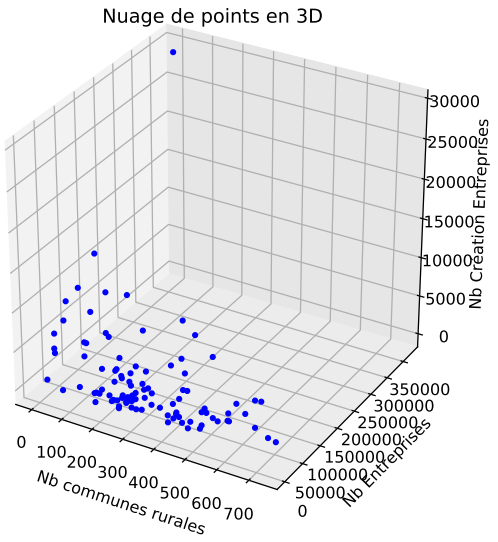


QQ-plots variables entreprises



Les deux graphiques précédents montrent que les relations entre variables semblent linéaires, sauf avec la variable "Nb communes rurales". Nous interprétons cela comme : moins un département est rural et plus l'activité économique est favorisée. Ou encore, plus un département se développe et plus il sera facile qu'il se développe. A l'inverse, plus un

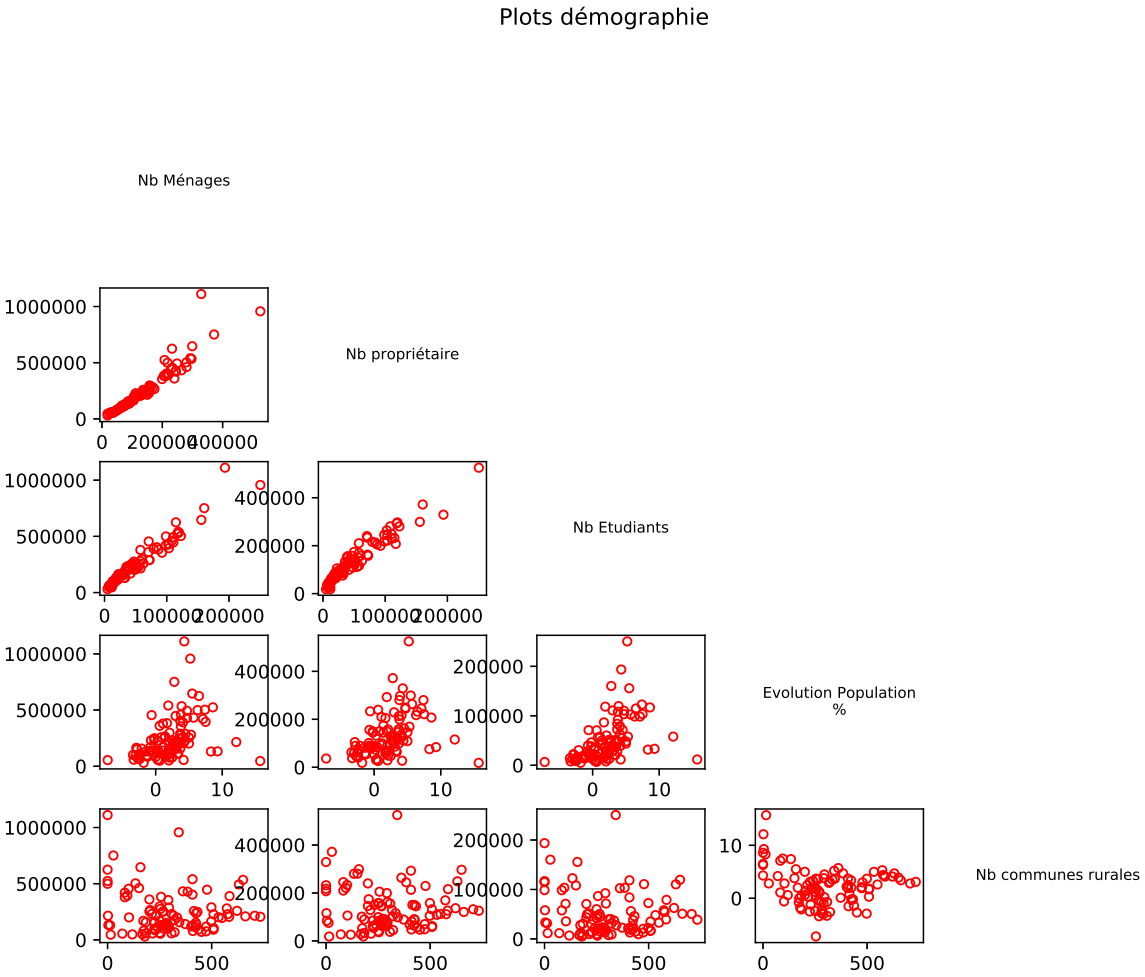
département est rural et plus il a de "chance" de le rester.



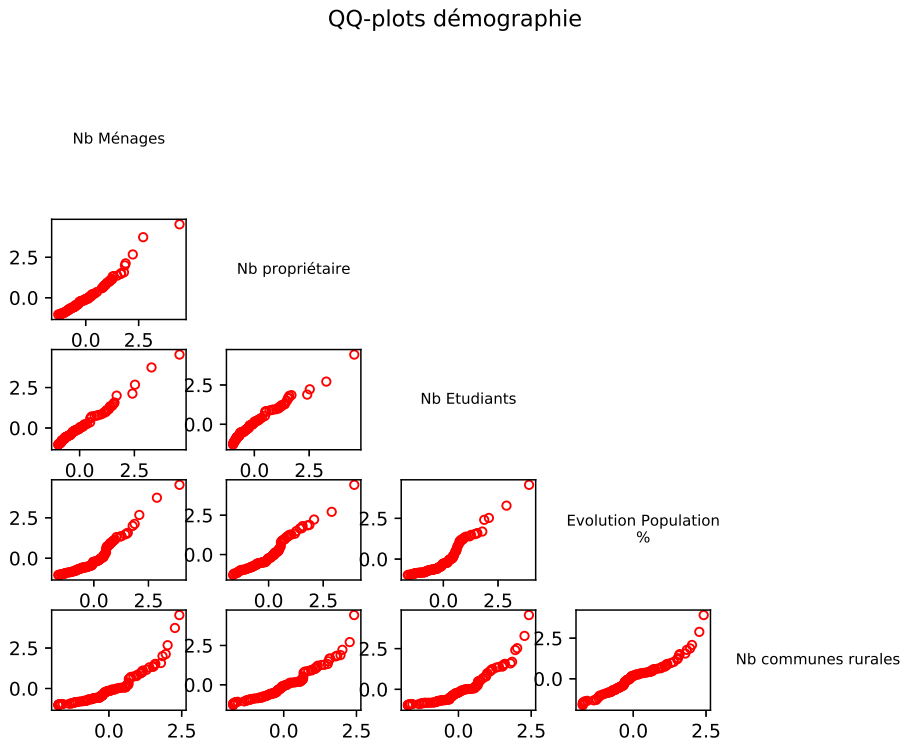
La représentation ne le montre pas bien, mais en utilisant la rotation interactive, on voit clairement que les points sont regroupés dans un plan incliné.
On voit aussi qu'un département se distingue des autres : le département 75, connu pour n'être pas rural, avec beaucoup d'entreprises installées et beaucoup de créations d'entreprises.

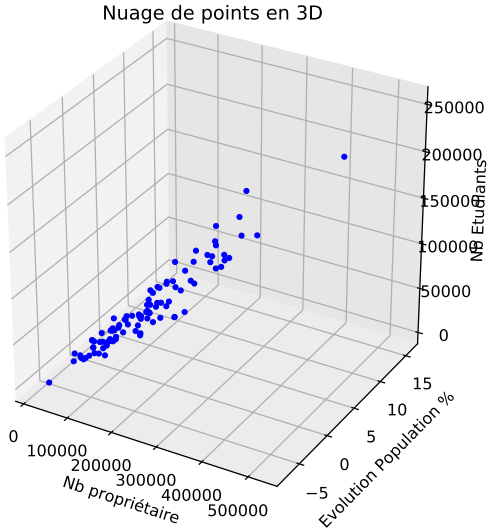
2.1.3 Population

Même graphiques descriptifs que précédemment, pour les variables du type "population".



Les variables "Nb de ménages", Nb de propriétaires" et "Nb d'étudiants" semblent corrélées linéairement. L'évolution de la population semble aussi proportionnelle à la population elle même, par accroissement et sans doute par attractivité. La variable "ruralité"



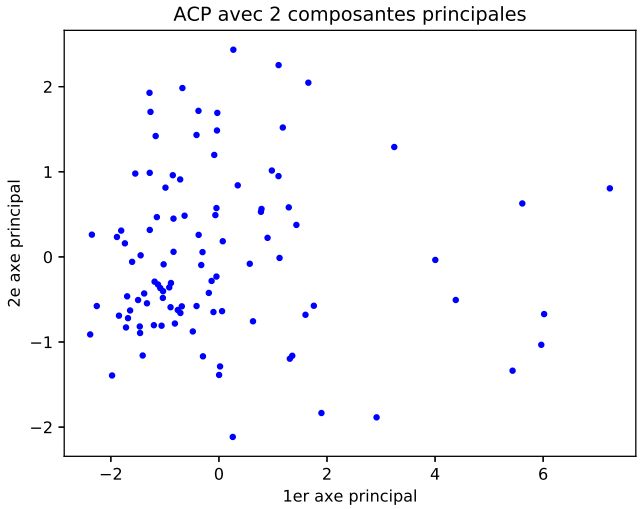


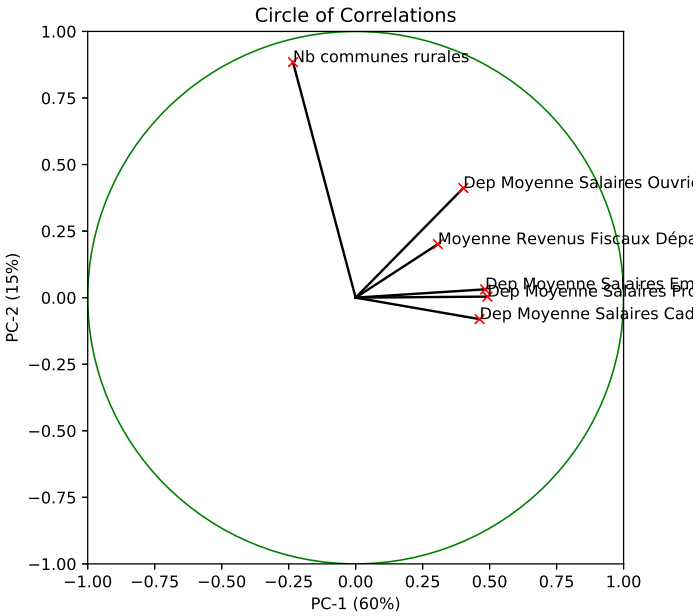
Sous cet angle, les variables indiquées semblent se regrouper le long d'une droite.

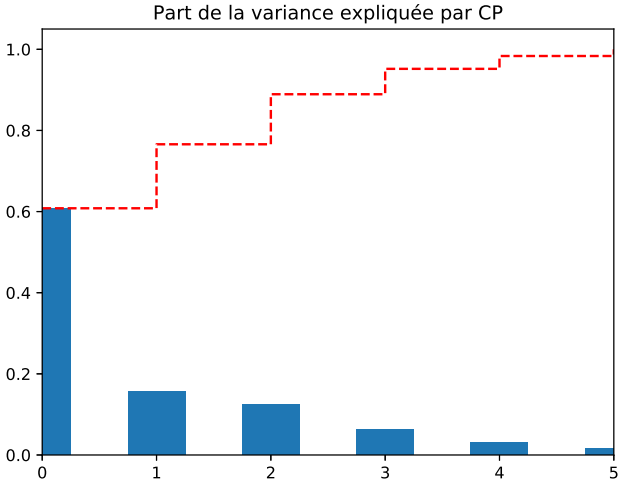
2.2 Analyse multivariée

2.2.1 Salaires

Le calcul d'une ACP donne les résultats suivants :



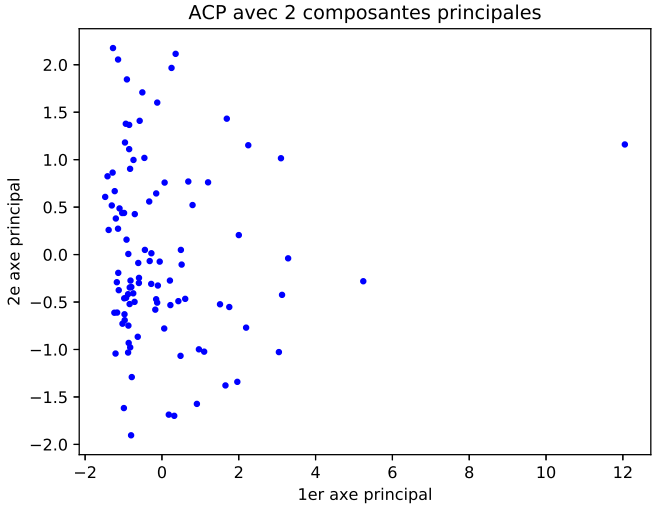


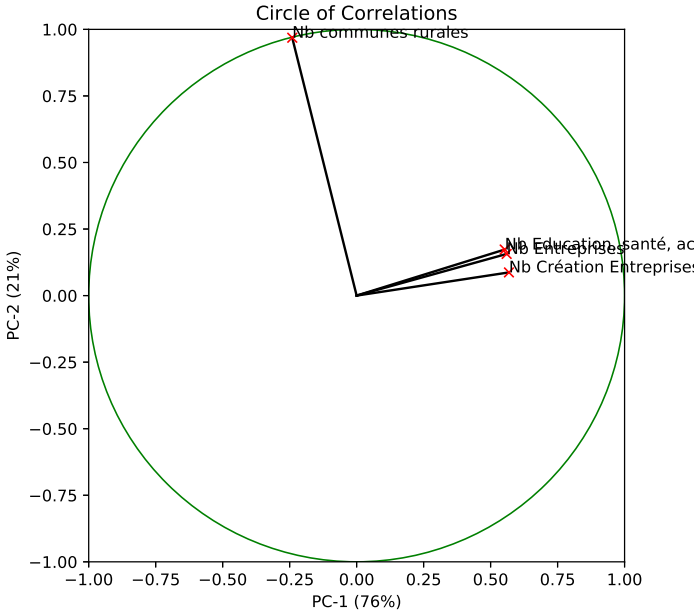


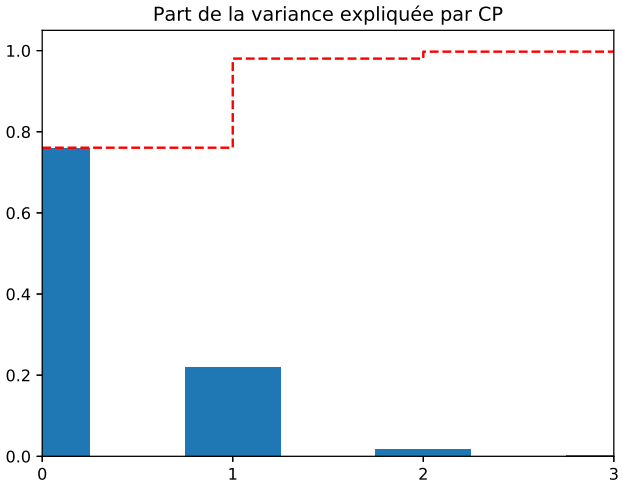
Il faut 3 composantes pour approcher 90% de l'inertie totale. Les individus ne sont donc pas "bien" regroupés autour d'un plan.

L'ACP montre que la variable "Nb de communes rurales" est la variable la plus corrélée aux CP. Cela confirme son rôle prédominant pour expliquer la distribution des salaires.

2.2.2 Entreprises

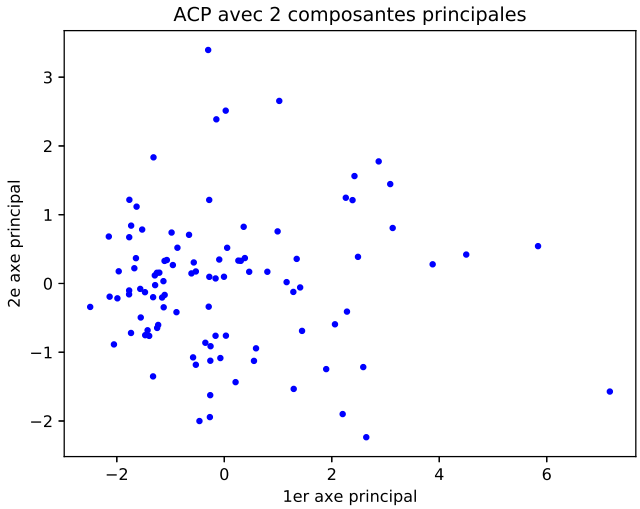


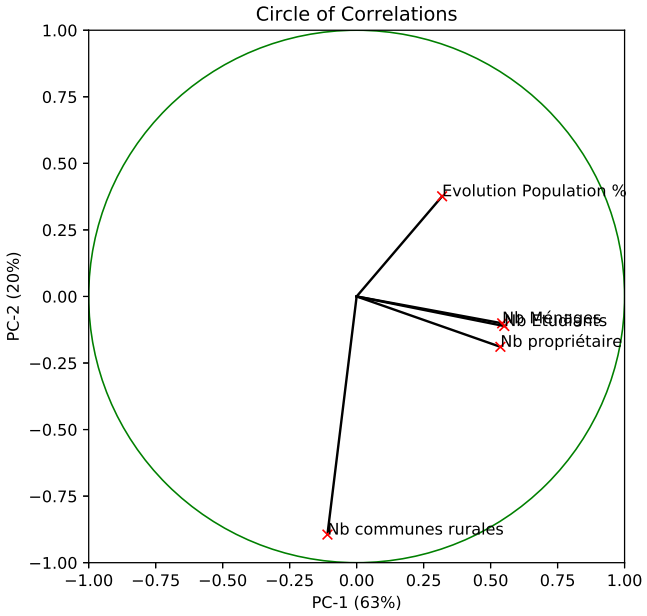


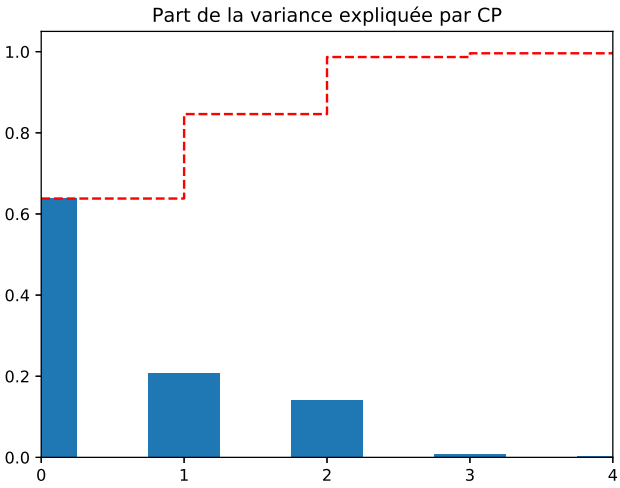


Il suffit de 2 CP pour dépasser 90% de l'inertie totale. L'ACP confirme le regroupement des individus (départements) selon un plan, comme la projection 3D semblait le suggérer précédemment.
La variable "Nb de communes rurales" est, comme dans le cas des salaires, la variable la plus corrélée aux CP. Cela confirme son rôle prédominant pour expliquer la distribution des entreprises et services.
Sur la projection à 2 CP, nous retrouvons le département 75, très à l'écart des autres sur le plan économique.

2.2.3 Population



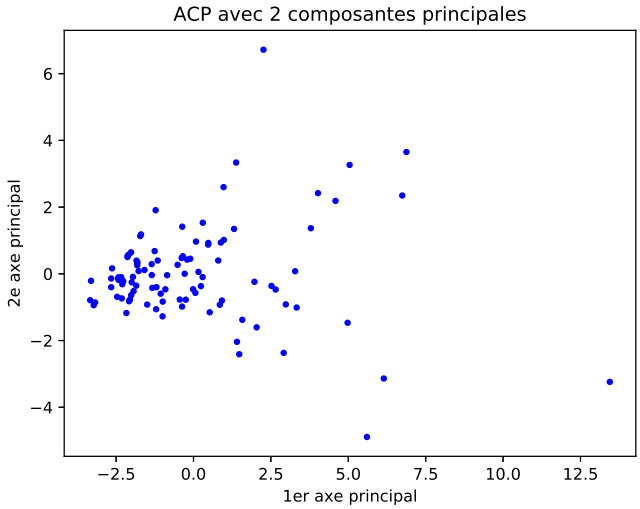


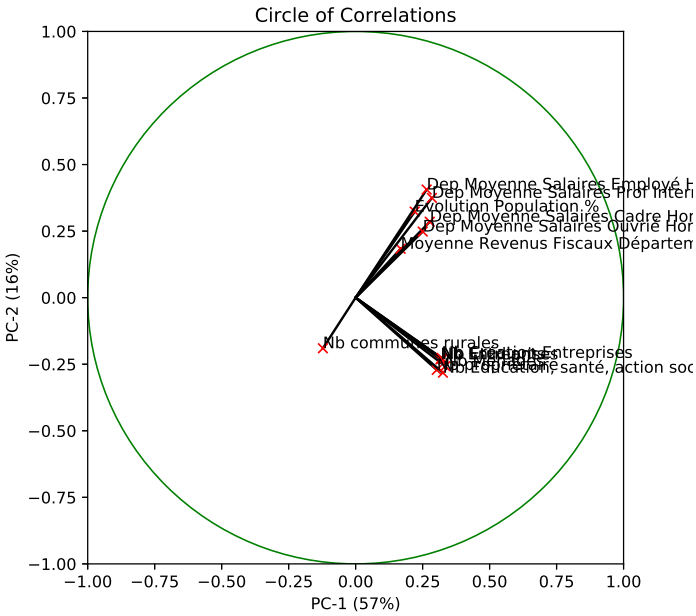


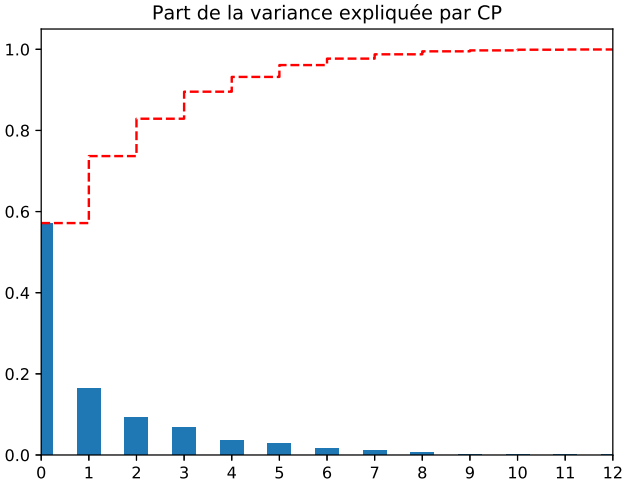
Il faut 2 CP pour approcher 90% de l'inertie totale. Le regroupement des points le long d'une droite, noté précédemment sur la projection 3d, était une illusion.
La variable "Nb de communes rurales" est encore la variable la plus corrélée aux CP.

2.2.4 Toutes variables réunies

Nous avons calculé une ACP en regroupant toutes les variables disponibles sur les départements (une vingtaine de variables), de façon à voir si la ruralité ressortait toujours comme une variable essentielle pour expliquer la variance.







La variable "Nb de communes rurales" ressort bien mais n'est plus la plus corrélée aux 2 CP. Y-aurait-il un effet de "dilution" quand on ajoute un grand nombre de variables ?

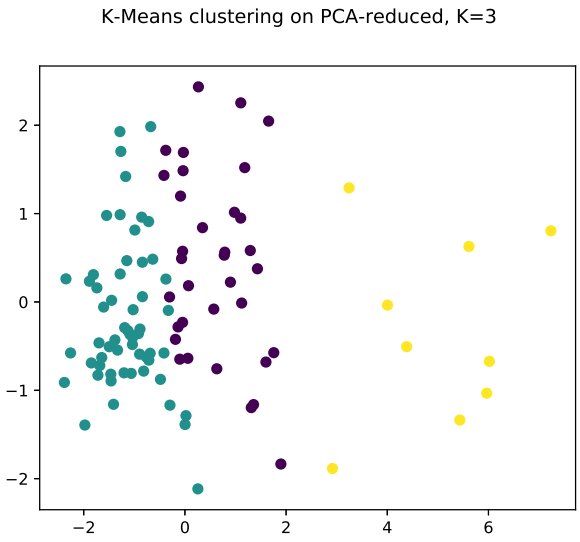
Globalement, il faut 4 CP pour approcher les 90% de la variance totale.
Sur le cercle des corrélations, on voit que les variables se regroupent en 3 ensembles, relativement à leur corrélation aux 2 premières CP. Relativement aux 2 premières CP, les variables d'origine, ne varient que de trois façon différentes. Cependant, nous ne voyons pas bien comment interpréter cette observation, ni comment l'utiliser pour lancer d'autres analyses.

2.3 Classification

Nous avons utilisé deux méthodes de classification, pour chaque ensemble de variables : K-means et regroupement hiérarchique ascendant. Pour le K-means, nous avons demandé chaque fois 3 groupes.
Pour le regroupement hiérarchique ascendant, nous avons ajusté le seuil de façon à obtenir 2 groupes.

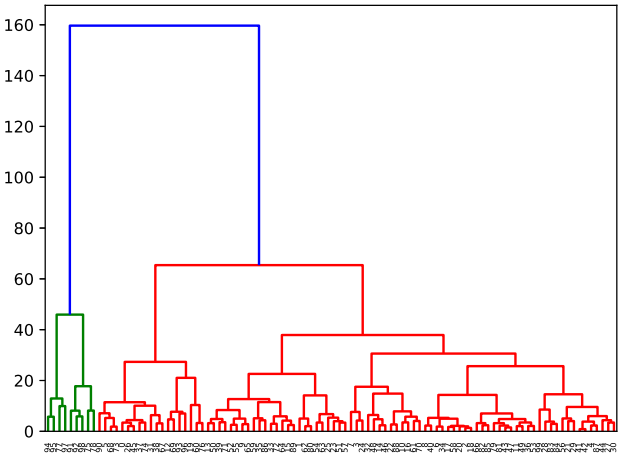
2.3.1 Salaires

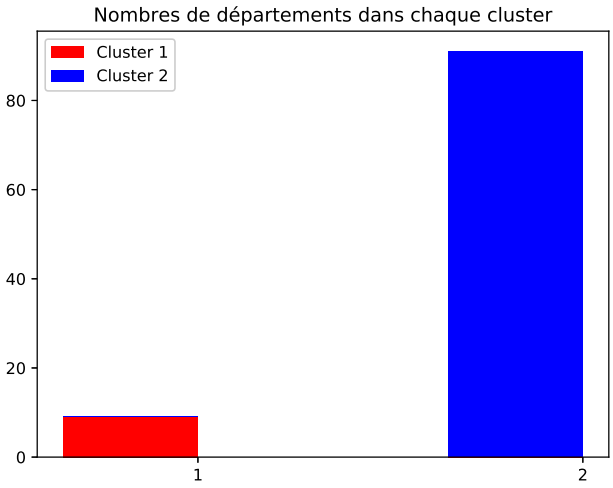
K-means



Visuellement, nous avons 3 groupes, mais le groupe jaune se détache plus nettement. C'est le groupe des départements où les salaires sont les plus élevés : 75, 77, 78, 91, 92, 94, 95, 972, 973.

Regroupement hiérarchique ascendant

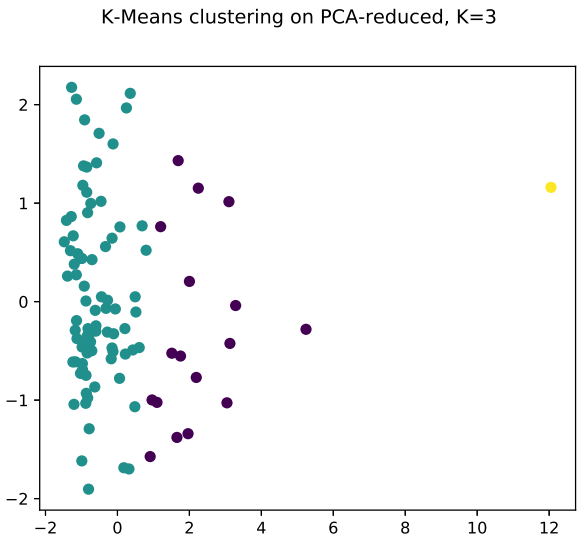




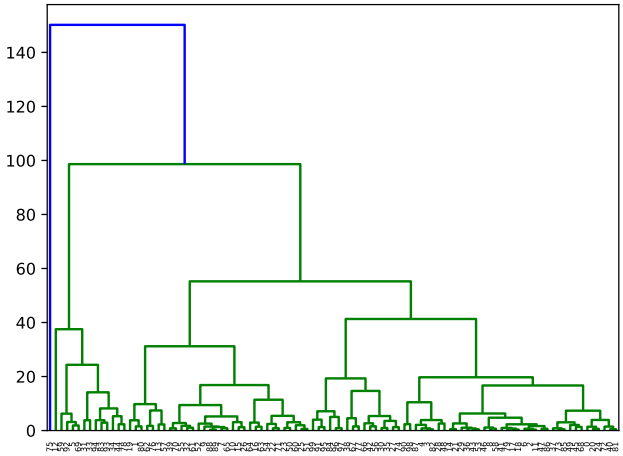
Cette méthode de classification nous donne deux groupes dont celui qui correspond aux salaires les plus élevés : 75, 77, 78, 91, 92, 94, 95, 972, 973. Cette liste est identique à la précédente. Les deux méthodes de classification utilisées donnent le même résultat sur ce groupe de variables.

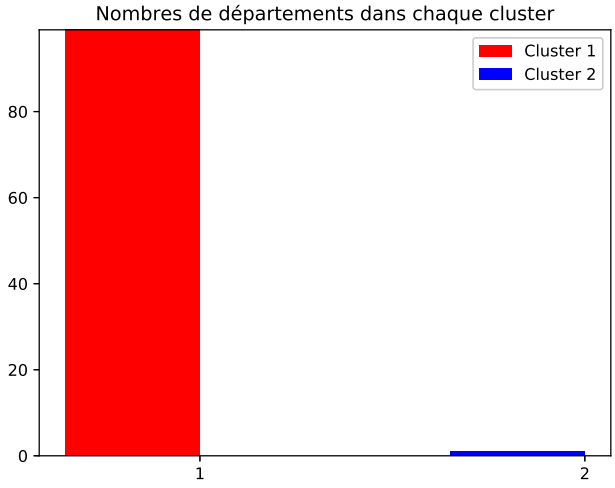
2.3.2 Entreprises

K-means



Regroupement hiérarchique ascendant

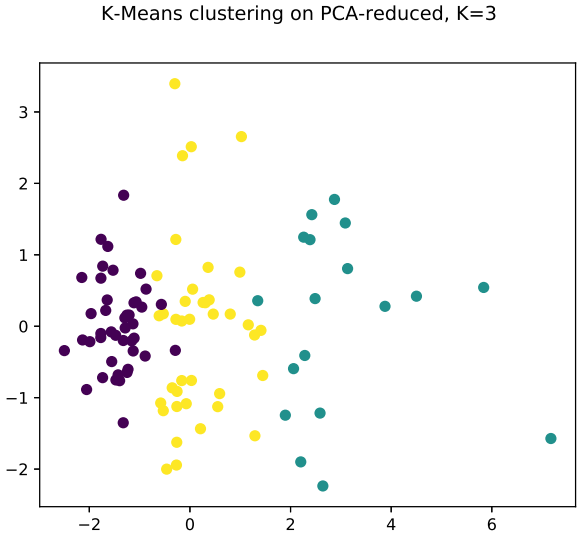




Pour ce groupe de variables, nous retrouvons le département 75 détaché de tous les autres. Il nous faudrait approfondir ce point, de façon à vérifier s'il ne s'agit pas d'une aberration dans les données, ce que nous n'avons pas fait par manque de temps (et d'énergie).

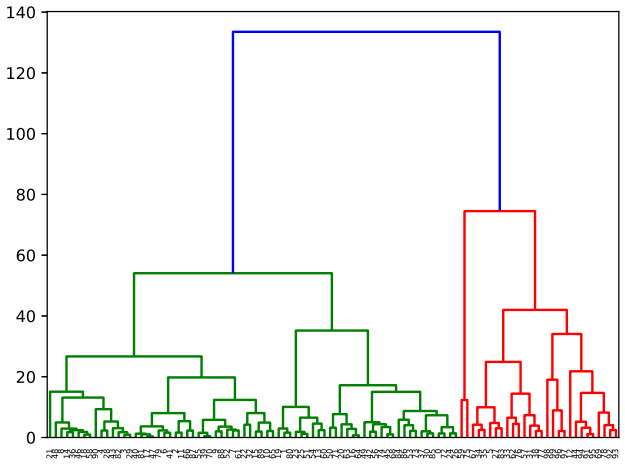
2.3.3 Population

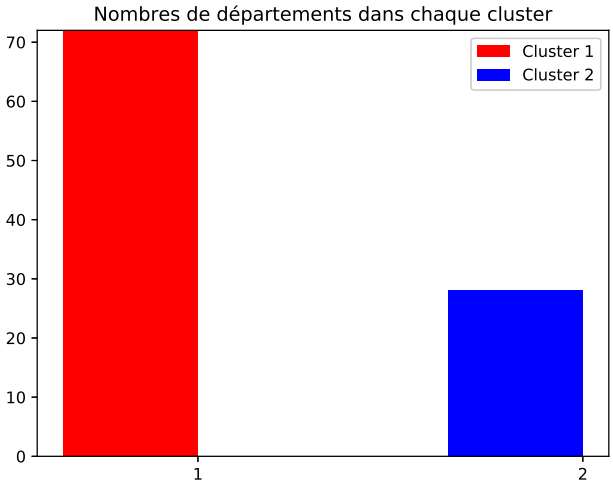
K-means



Les trois groupes ne sont pas nettement détachés. Visuellement, il y a plustôt une multitude de petits groupes.

Regroupement hiérarchique ascendant





Du point de vue des variables liées à la population, la méthode de regroupement hiérarchique ascendant semble montrer que deux groupes se détachent : ceux qui ont une dynamique forte et les autres. Cette méthode sépare plus nettement les groupes. Cependant, les département du cluster 2 correspondent à ceux du groupe vert obtenu avec le K-means.

3 Conclusion

Ce travail est une étude succincte de quelques variables relatives aux départements français. Nous avons fait des choix dans les variables à exploiter. Nous avons souhaité les grouper en trois familles, relatives à trois aspects des départements qui nous paraissaient pertinents : économique, démographiques et niveau des salaires Nous avons pu montrer que l'aspect rural d'un département expliquait pour beaucoup les différences économiques, démographiques et de niveau des salaires entre les départements. C'est un lieu commun qui est confirmé par les statistiques.
Il serait intéressant de construire une étude analogue sur d'autres pays d'Europe, de façon à voir si la ruralité joue un rôle prépondérant sur la démographie, le niveau de vie ou la dynamique économique, comme elle le joue en France, pays de forte centralisation.

[Le travail purement technique (programmation Python) a pris la majeure partie du temps limité que je peux consacrer à mes études hors temps professionnel et familial. Je suis bien conscient des nombreuses insuffisances des analyses et de la "légèreté" de mes interprétations.]