Rapport Projet Statistique multidimensionnelle - Groupe 2

Gustavo MAGAÑA LÓPEZ, Simon MATTENS, Joachim SNEESSENS

11 juin 2019

Introduction

Ce projet est constitué de deux questions. La première consiste à produire l'analyse et le commentaire du fichier de données et la deuxième à décrire une méthode d'analyse non vue au cours.

Pour ce faire, nous avons utilisé le logiciel R.

Question 1

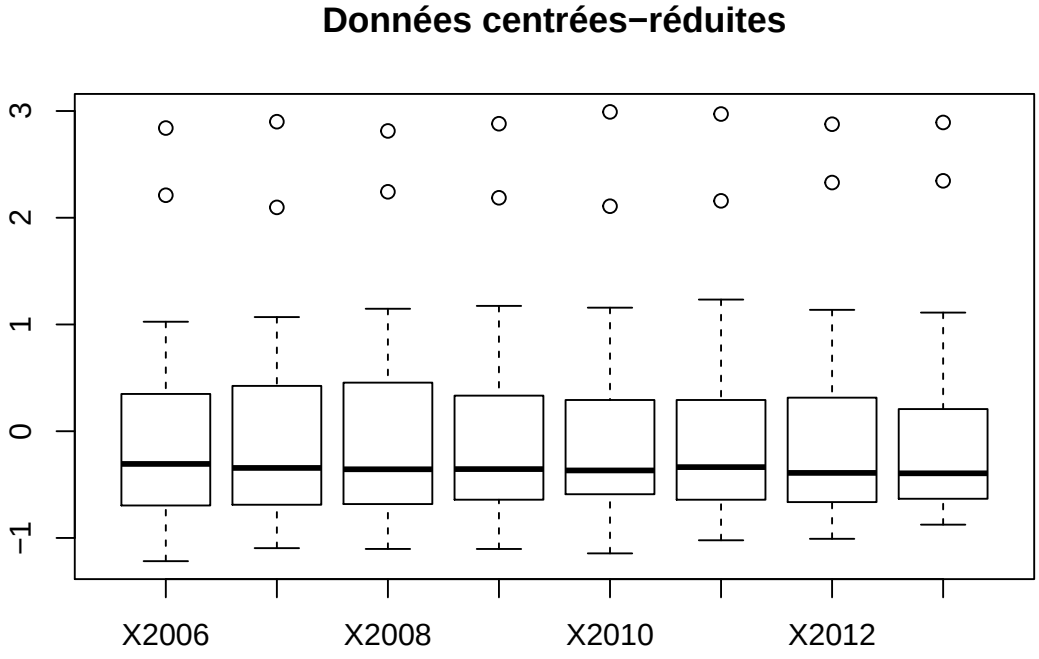
Le fichier à analyser est le fichier "pourcentage_chercheurs_france.xslx".

Le fichier contient une table dont chaque case donne la proportion de chercheurs dans la région spécifiée par la ligne pour une certaine année spécifiée par la colonne. Le but était d'effectuer une analyse en composantes principales des données, de dire si le profil est stable dans le temps et enfin de dégager des groupements de régions en fonction de leur comportement à l'aide d'une classification hiérarchique ascendante.

ACP

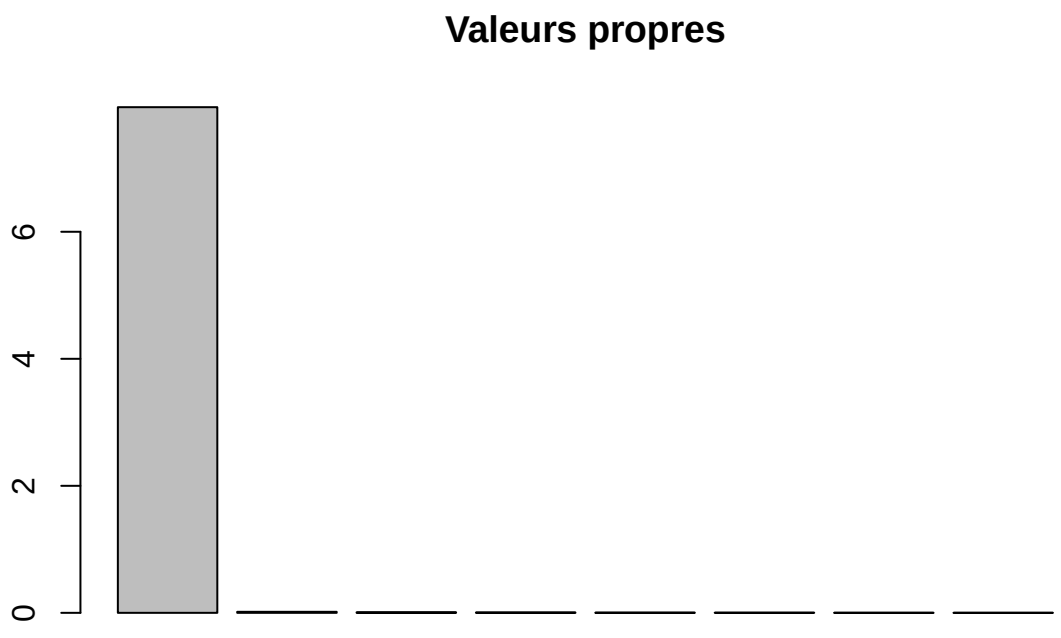
Le but de l'ACP est de réduire la dimension des données initiales (ici 8 car 8 années différentes) en remplaçant ces 8 variables initiales par n facteurs appropriés (avec $n < 8$). Avant de commencer l'ACP nous avons retiré la première colonne contenant les noms des différentes régions de France. Ci-dessous un boxplot des données non centrées non réduites.

Voici le boxplot obtenu après centrage et réduction des données.



Nous pouvons remarquer que pour chaque année, les médianes sont inférieures aux moyennes ce qui signifie que, pour chaque année, la répartition des individus au sein de celle-ci est assez déséquilibrée.

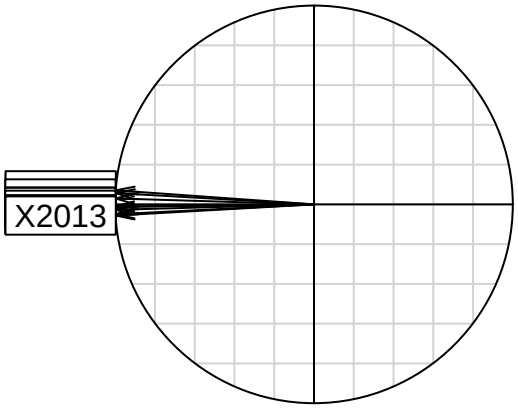
De manière globale, on constate que ce déséquilibre se retrouve pour chacune des variables, et que ces dernières ont des distributions relativement similaires entre elles.




```
## [1] 0.9952309807 0.0019689972 0.0010415277 0.0007368368 0.0004357696
## [6] 0.0003108272 0.0001456718 0.0001293890
```

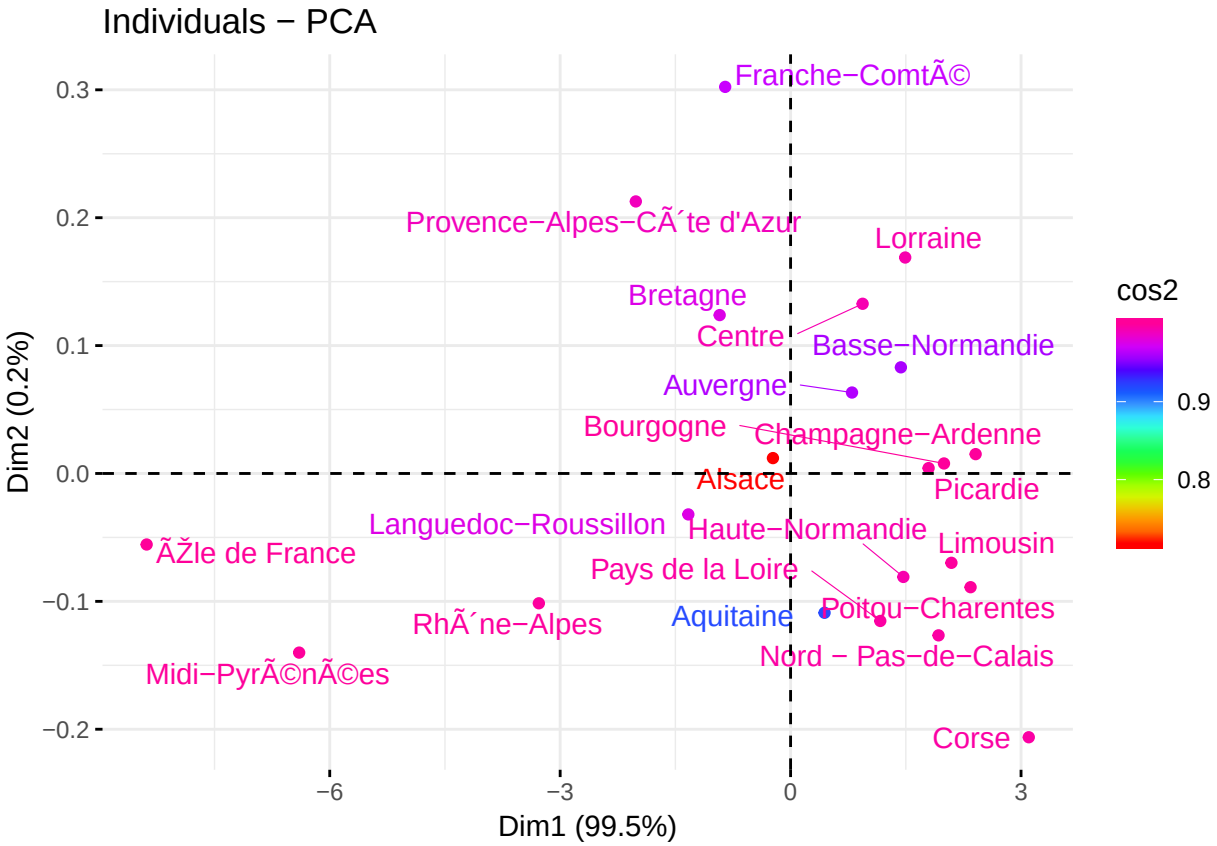
Cette étape a consisté dans un premier temps à calculer la matrice des corrélations de la matrice centrée-réduite, pour dans un second temps en calculer les valeurs propres. Vu l'allure du graphe des valeurs propres nous allons garder deux axes comme la visualisation ne peut pas se faire correctement si on ne garde qu'un seul axe.

Cercle des corrélations



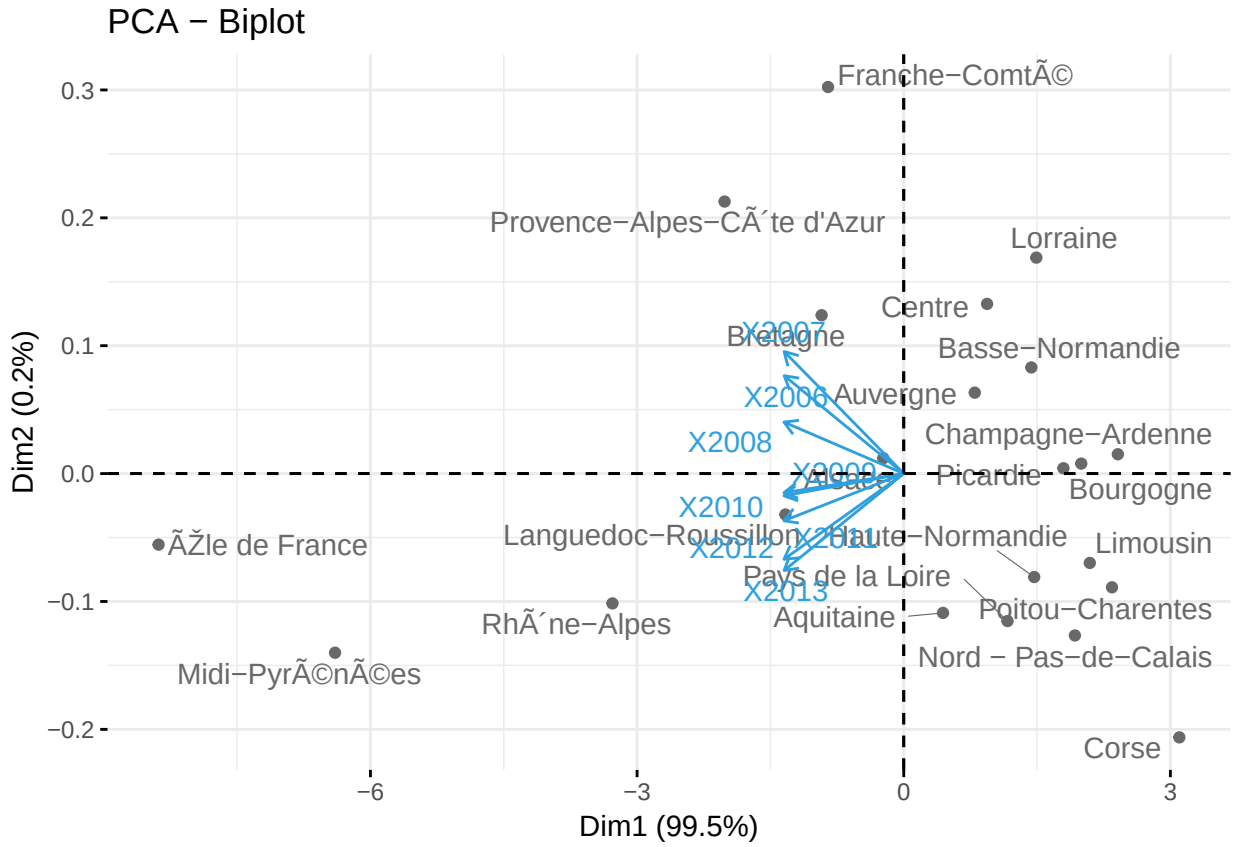
Grâce au cercle des corrélations, nous pouvons voir que toutes les variables (années) sont très fortement corrélées entre elles. Nous voyons aussi qu'elles sont - négativement - très fortement corrélées avec la première composante principale. Concernant les corrélations entre les variables et la seconde composante principale, nous constatons qu'elles sont toutes très faibles (vecteurs quasiment perpendiculaires), certaines négatives et d'autres positives.

ACP : Représentations complémentaires



Visualisation de l'ACP où les individus sont des points, colorés en fonction de la valeur du cosinus carré,

c'est-à-dire la qualité de leur représentation dans l'ACP.



Le biplot nous permet d'avoir la représentation simultanée des variables et individus utilisés pour l'ACP. Précisons que les pourcentages exprimés entre parenthèses sur les titres des axes correspondent à la part de la variance totale expliquée par l'axe en question.

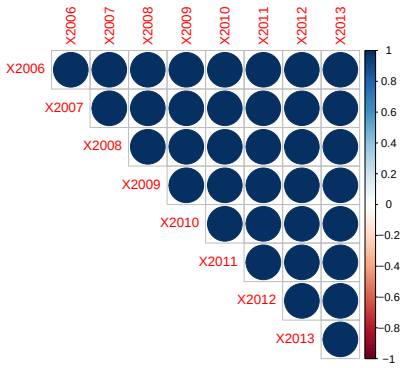
Dans un biplot, la direction des vecteurs correspondants aux variables nous permet de voir approximativement la corrélation que chaque variable a avec chaque composante principale. D'après le biplot, toutes les variables sont fortement et négativement corrélées avec la première composante principale et faiblement corrélées avec la seconde, ce qui confirme les observations faites sur le cercle des corrélations.

Si on ajoute à ce faitla variation minimale expliquée par la seconde composante, nous avons que plus les régions sont à gauche dans le biplot, plus haute est la valeur qu'elles ont prises pour toutes les années.

Stabilité temporelle du profil

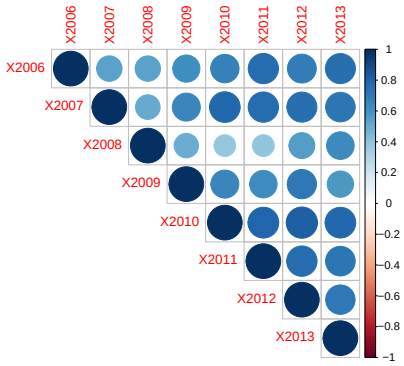
Pour faire cette partie de l'analyse nous avons utilisé plusieurs outils. Dans un premier temps, la matrice des corrélations nous permet de voir les corrélations existantes entre chaque couple de variables. Au moyen de la fonction **corrplot()** on obtient une matrice contenant des cercles colorés et une échelle de telle façon qu'un coup d'oeil suffit à avoir une vision générale des corrélations. Chaque colonne que l'on a prise comme variable n'est autre qu'un vecteur contenant la valeur que chaque région a obtenue pour une année spécifique : le profil de l'année. S'il existe une corrélation positive assez forte parmi les colonnes, on peut dire que le profil, c'est-à-dire la relation entre régions, est stable dans le temps.

```
chercheurs.corr <- cor(chercheurs.scaled)
corrplot(chercheurs.corr, type="upper")
```



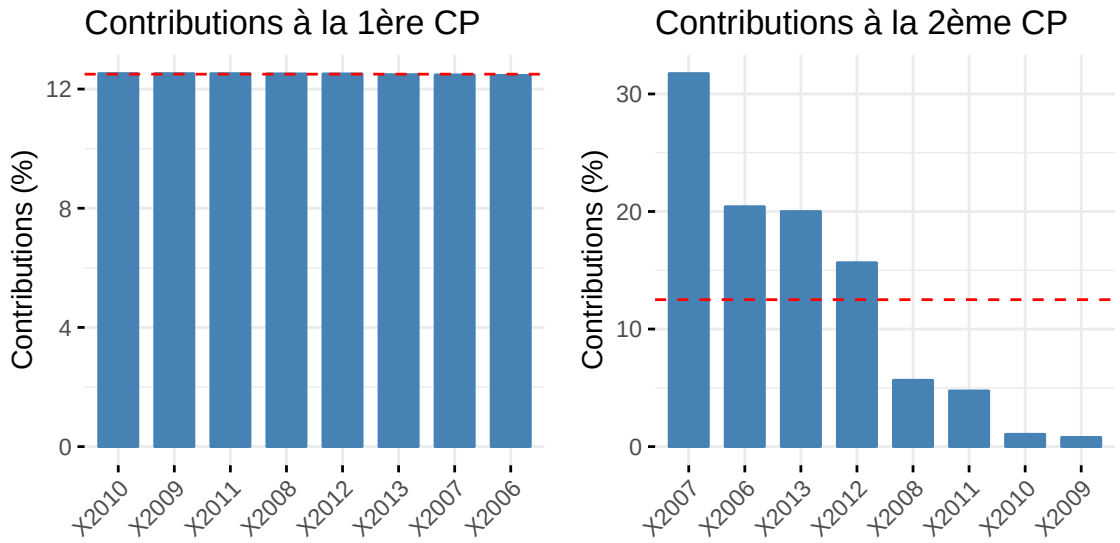
Ces corrélations nous montrent que **le profil est stable dans le temps**. Pour justifier ce raisonement il suffit de penser au cas contraire : Un profil qui n'est pas stable dans le temps. Disons que nous avons deux années où une région a augmenté ou diminué d'une façon différente aux autres. La matrice de corrélations ne ressemblerait plus à celle que l'on a dans la figure précédente. Pour simuler cette situation, nous créons un **corrplot** avec la matrice que l'on avait plus un peu de bruit.

```
foo <- apply(chercheurs, 2, function(x){ x + rnorm(n = length(x), mean = 1, sd = 0.3) })
foo.scaled <- scale(foo)
corrplot(cor(foo.scaled), type="upper")
```



Nous pouvons également ajouter à la justification de notre affirmation de la stabilité temporelle du profil l'analyse des contributions des variables aux composantes principales. Il faut premièrement rappeler ce que l'on avait obtenu grâce à l'ACP : deux composantes représentant respectivement le 99,5% et 0,2% de la variance totale. Nous présentons les contributions des variables à chacune des composantes principales. Il faut tenir en compte que ce qui se passe sur la deuxième composante principale n'est pas très significatif étant donné qu'elle ne représente que 0,2% de la variance totale.

Contributions des variables :



Nous pouvons voir que chaque année contribue également à la première composante principale. Si le profil changeait considérablement pour un année, sa variation serait différente à celles des autres, ce qui nous donnerait une autre distribution.

Pour justifier encore notre affirmation, nous affichons les variations temporelles pour chaque région (**var.regions**) et les variations régionales pour chaque année (**var.annees**). On constate que la variance entre les individus pour une année donnée est beaucoup plus importante que la variance entre les différentes années pour une région donnée.

Analyse des variances :

```
var.regions <- apply(chercheurs, 1, function(x){ var(as.numeric(x)) })
var.annees <- apply(chercheurs, 2, function(x){ var(as.numeric(x)) })
var.regions
```

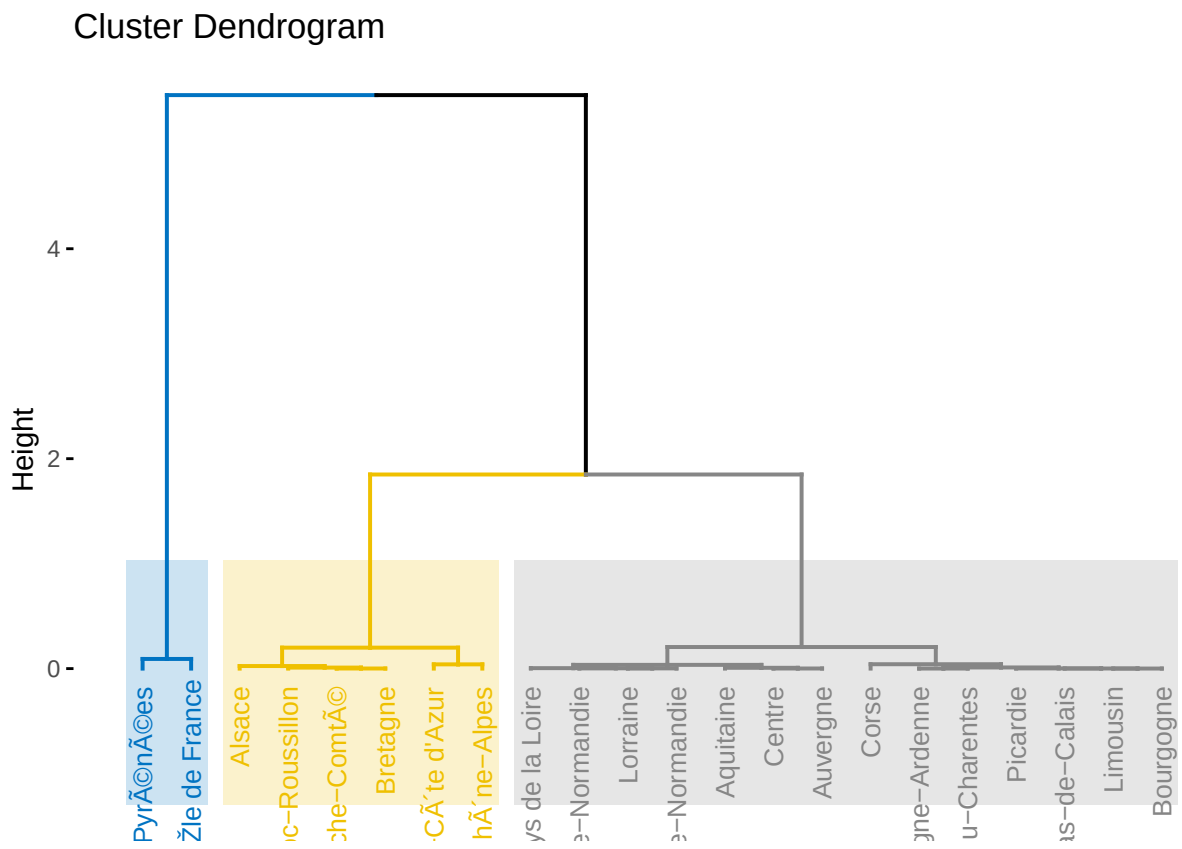
```
##               ÃŽle de France           Champagne-Ardenne
##                  0.0222732098                0.0003245613
##                     Picardie              Haute-Normandie
##                  0.0006487429                0.0022070629
##                       Centre              Basse-Normandie
##                  0.0001741784                0.0017569907
##                    Bourgogne         Nord - Pas-de-Calais
##                  0.0003543286                0.0015341743
##                     Lorraine                      Alsace
##                  0.0003038341                0.0022767984
##              Franche-ComtÃ©             Pays de la Loire
##                  0.0011147021                0.0021390541
```

```
##                     Bretagne              Poitou-Charentes
##                  0.0017517371                   0.0009552098
##                    Aquitaine               Midi-PyrÃ©nÃ©es
##                  0.0027433055                   0.0198914170
##                    Limousin                  RhÃ´ne-Alpes
##                  0.0011718784                   0.0087598714
##                    Auvergne           Languedoc-Roussillon
##                  0.0015082998                   0.0042686571
## Provence-Alpes-CÃ´te d'Azur                         Corse
##                  0.0029634941                   0.0026920484
```

```
var.annees
```

```
##      X2006     X2007     X2008     X2009     X2010     X2011     X2012
## 0.1265820 0.1416750 0.1434138 0.1557430 0.1744529 0.1775602 0.1926834
##      X2013
## 0.2019562
```

Classification Hiérarchique Ascendante



Ici nous voyons comment une classification hiérarchique ascendante nous fournit des groupes contenant des profils similaires qui n'est rien d'autre que la moyenne de la valeur qu'elles prennent au fil des années.

```
rev(sort(apply(chercheurs, 1, mean)))[1:2]
```

```
##   ÃŽle de France Midi-PyrÃ©nÃ©es
##         1.811212         1.534563
```

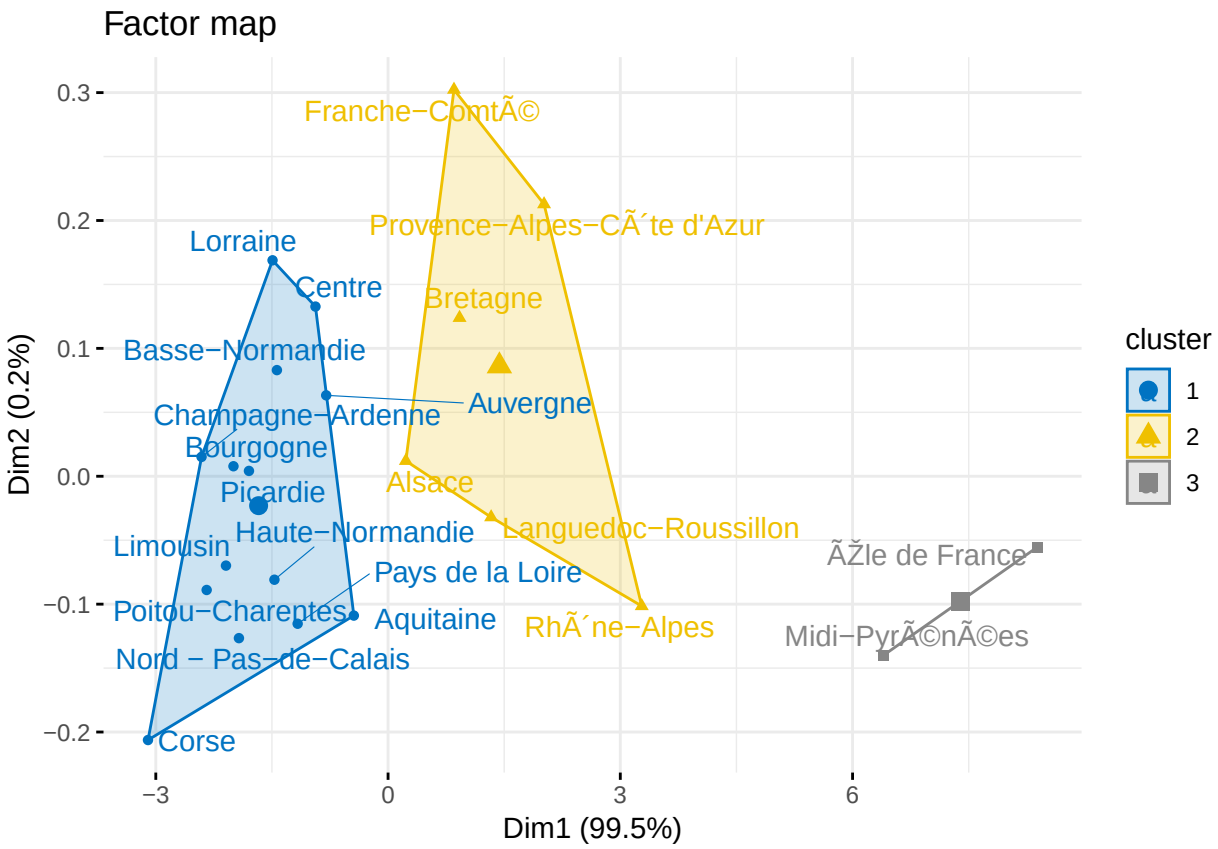
```
rev(sort(apply(chercheurs, 1, mean)))[3:8]
```

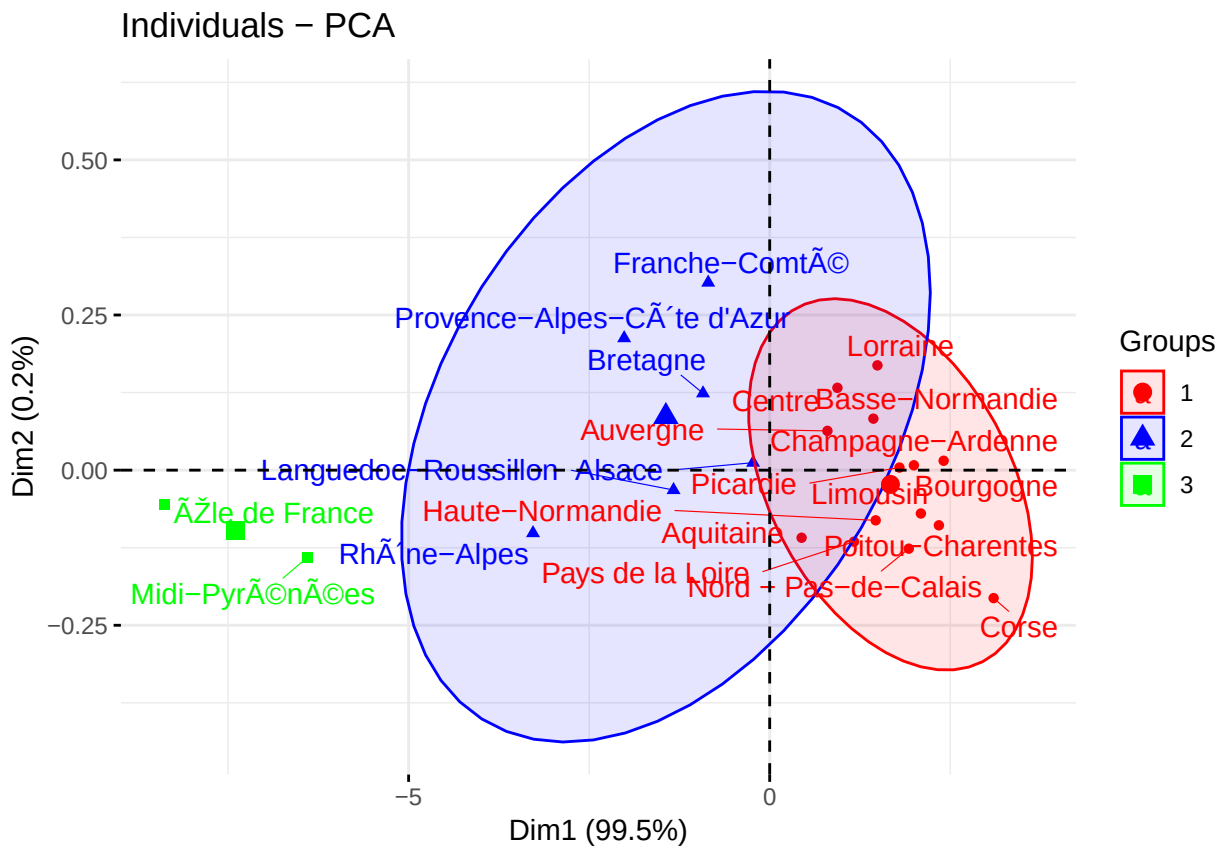
```
##                RhÃ´ne-Alpes  Provence-Alpes-CÃ´te d'Azur
##                   1.0985000                    0.9195375
##        Languedoc-Roussillon                    Bretagne
##                   0.8259000                    0.7678500
##              Franche-ComtÃ©                      Alsace
##                   0.7561250                    0.6720875
```

```
rev(sort(apply(chercheurs, 1, mean)))[9:22]
```

```
##            Aquitaine              Auvergne                Centre
##            0.5791375            0.5281625             0.5078125
##     Pays de la Loire       Basse-Normandie       Haute-Normandie
##            0.4781625            0.4385250             0.4363000
##             Lorraine              Picardie  Nord - Pas-de-Calais
##            0.4299375            0.3894000             0.3724000
##            Bourgogne              Limousin      Poitou-Charentes
##            0.3612500            0.3486875             0.3139125
##    Champagne-Ardenne                 Corse
##            0.3036875            0.2097375
```

Cette classification réaffirme notre hypothèse : le profil est stable dans le temps et la variation se trouve plutôt parmi les individus. Le premier groupe que l'on trouve sont les deux régions qui ont pris la valeur plus haute à travers les années. Dans la page suivante on trouve deux plots qui montrent les groupes créés par la fonction **HCPC()**, un en forme de "Factor Map", l'autre superposé à la visualisation de l'ACP.


Factor map
Franche−ComtÃ©
Provence−Alpes−CÃ´te d'Azur
Lorraine
Centre
Bretagne
Basse−Normandie
Champagne−Ardenne
Auvergne
Bourgogne
Picardie
Alsace
Haute−Normandie
Languedoc−Roussillon
Limousin
ÃŽle de France
Pays de la Loire
Poitou−Charentes
Aquitaine
RhÃ´ne−Alpes
Midi−PyrÃ©nÃ©es
Nord − Pas−de−Calais
Corse
Dim2 (0.2%)
Dim1 (99.5%)
0.3
0.2
0.1
0.0
−0.1
−0.2
−3
0
3
6
cluster
1
2
3



Question 2 - Réseaux de neurones

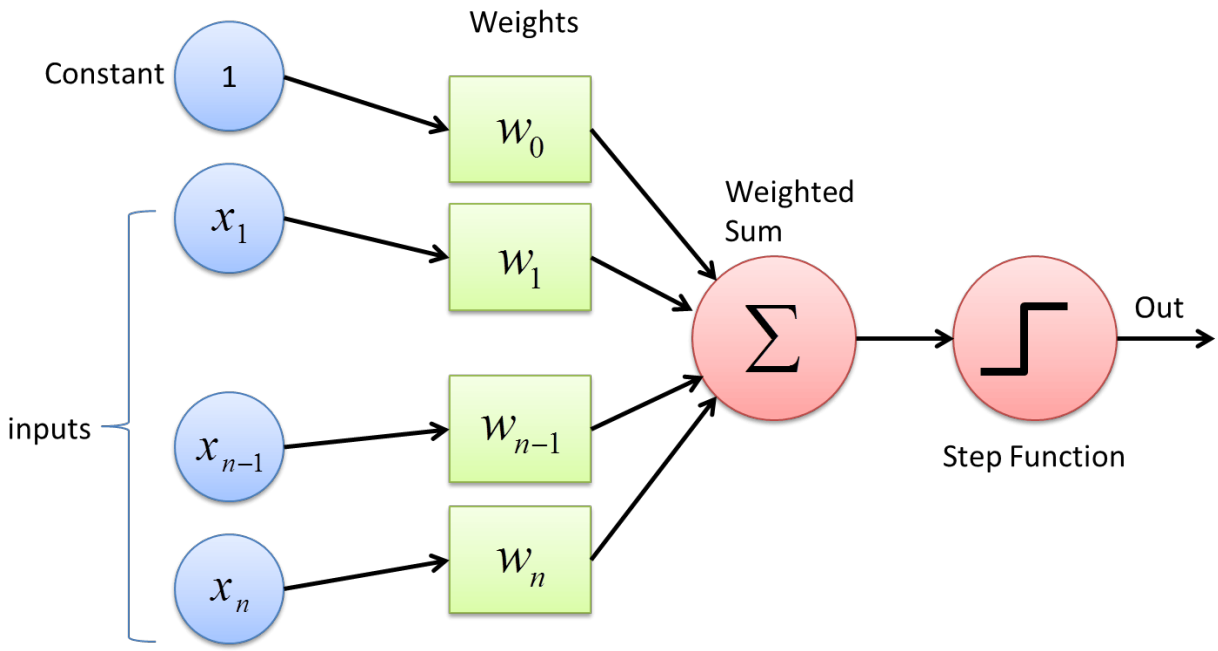
L'analyse par réseaux de neurones est une méthode de classification par apprentissage supervisé. Les groupes sont fixés et on dispose d'exemples d'invidus pour chaque groupe.

Un neurone à p entrées est une fonction $f : R^{p+1} \times R^p \to R$ définie par:

- une fonction de transfert $g : R \to R$
- un vecteur de poids $W \in R^{p+1}, W = (w_0, ..., w_p, w_{p+1})$
- $\forall X \in R^p, f(W, X) = g(w_0 + \sum_{i=1}^{p+1} w_i x_i)$ avec $X = (x_1, ..., x_p)$.

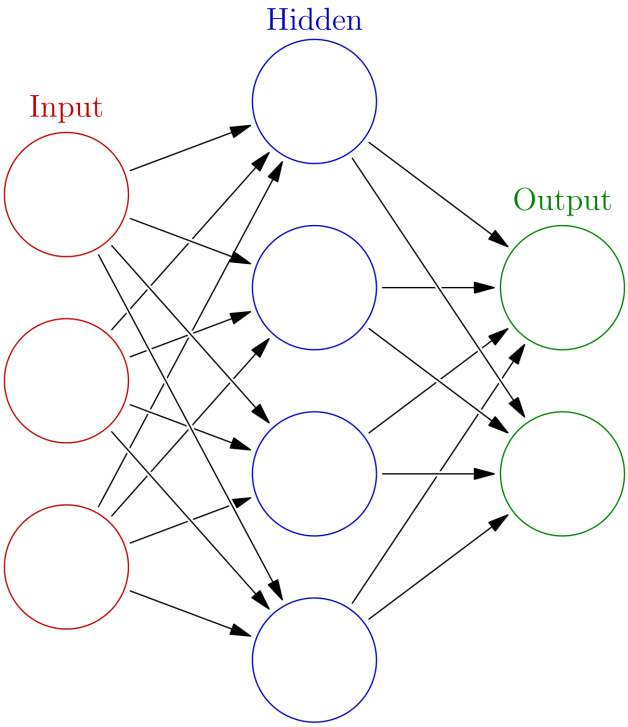
Les fonctions de transfert les plus couramment utilisées sont les les fonctions linéaires ou sigmoïdes ($\frac{1}{1+e^{-x}}$).

Ci-dessous une représentation graphique d'un neurone. Les noms de variables correspondent à la défition énoncée précedemment, et la *step function* du graphique correspond à la fonction de transfert g dans la définition.



Une couche de neurones est une juxtaposition de neurones ne communiquant pas entre eux et possédant les mêmes entrées.

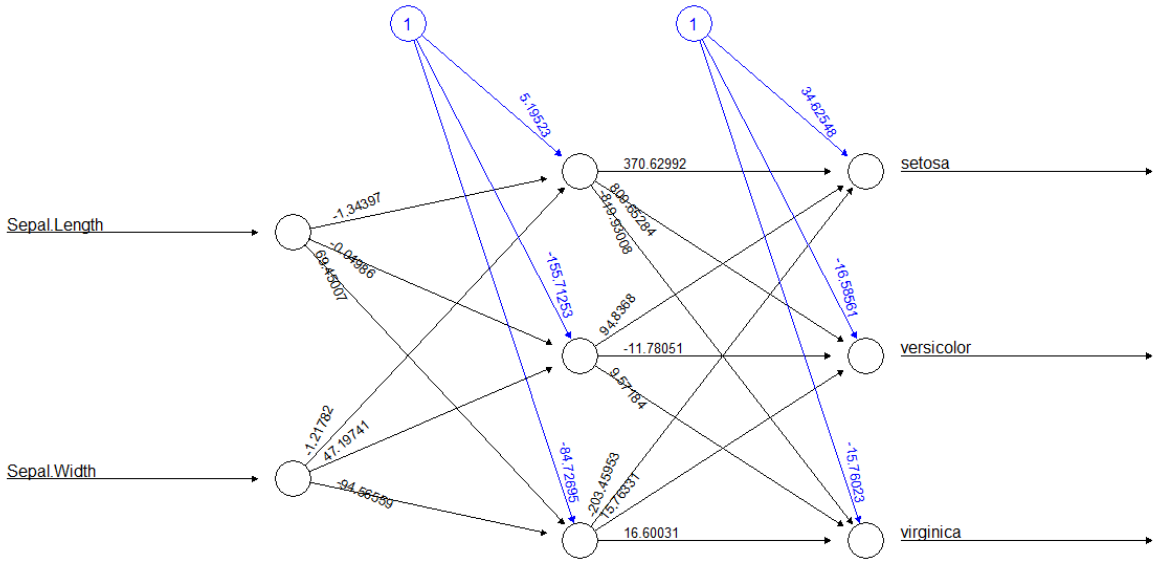
Un réseau de neurones est une superposition de couches connectées les unes autres. Dans le cadre d'un problème de classification, le réseau de neurones a p entrées et q sorties, où p est le nombre de variables du problème et q le nombre de classes du problème. Les vecteurs W des différents neurones du réseau sont initialisés aléatoirement. Ensuite on prend un ensemble d'individus qu'on fait passer dans le réseau pour obtenir des prédictions. L'étape suivante consiste à comparer les prédictions avec les valeurs attendues et la calculer la *LossFunction*, pour ensuite mettre à jour les différents paramètres - les vecteurs W - en propageant l'information par "backpropagation". La phase d'apprentissage consiste à répéter ces différentes étapes jusqu'à ce qu'un modèle satisfaisant soit obtenu.



Ce schéma représente un réseau de neurones comprenant trois couches, la première couche étant la couche des entrées et la dernière étant la couche des sorties. Les couches intermédiaires sont dites *couches cachées*. La plupart du temps, des réseaux de neurones à trois couches comme celui du schéma suffisent à résoudre la majorité des problèmes. Augmenter le nombre de neurones sur la couche cachée rend les modèles plus fexibles, donc un nombre de neurones sur cette couche trop élevé entrainerait de l'overfitting.

En R, la librairie *neuralnet* permet d'utliser les réseaux de neurones. L'exemple utilisé s'appuie sur le jeu de données *iris*. Après avoir ajouté trois colonnes *setosa*,*versicolor* et *virginica* contenant des booléens (True dans la colonne *setosa* si la variable Species de cet individu vaut setosa), le jeu de données a été séparé en deux parties. La première servant de training set et la seconde de testing set.

Ci-dessous un plot du résulat de l'utilisation de la fonction neuralnet pour déterminer les variables *setosa*, *versicolor* et *virginica* en fonction des variables *Sepal.Length* et *Sepal.Width*. Ici le nombre de neurones sur la couche cachée a été fixé à trois.



Error: 49.860195 Steps: 39076

Correspondant au réseau de neurones de la figure précédente, voici la table comparant les prédictions aux résultats attendus.

pred	setosa	versicolor	virginica
setosa	26	0	0
versicolor	0	17	10
virginica	1	4	17