Rapport du Projet
Statistique Multidimensionnelle

Groupe Info n°1
JOSSE Thomas, HUYLENBROECK Florent, DELFOSSE Charly

Année Académique 2018-2019
Bachelier en Sciences Informatiques

Faculté des Sciences, UMons

7 juin 2019

Résumé

Ce rapport est écrit dans le cadre du cours de `Statistique Multidimensionnelle` dispensé par M. *Michel VOUÉ*. Ce projet consistait en l'application de l'`Analyse en Composantes Principales` vues au cours sur un cas concret, ainsi que la découverte d'une technique d'analyse non abordée au cours à savoir, l'`Analyse Discriminante Linéaire` ou ADL.

Table des matières

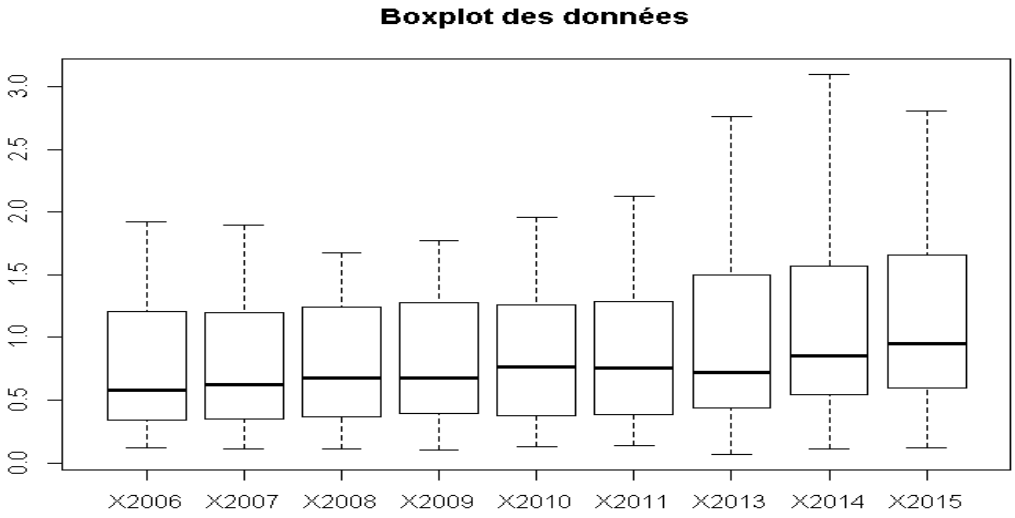
1 Question 1 : ACP sur un cas concret

1.1 Introduction

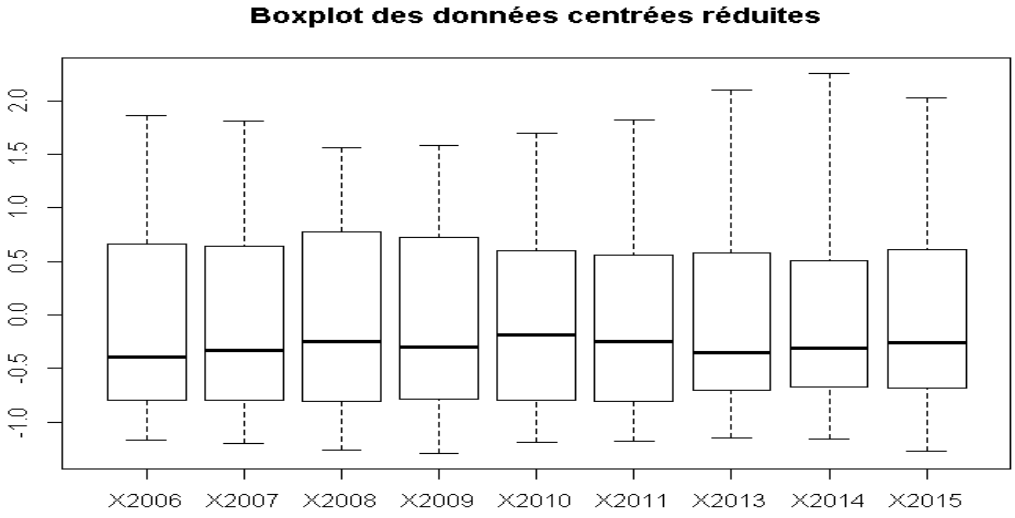
Pour cette question nous avions plusieurs sous-points à analyser : tout d'abord effectuer notre ACP des données à proprement dites, puis prédire à l'aide de notre analyse si le profil des données était stable et ensuite grouper les Provinces/Régions qui avaient des comportements similaires avec une méthode de classification hiérarchique ascendante.

1.2 ACP

D'abord, les données ont été extraites et analysées, un boxplot a été effectué avant et après centrage et réduction des données.

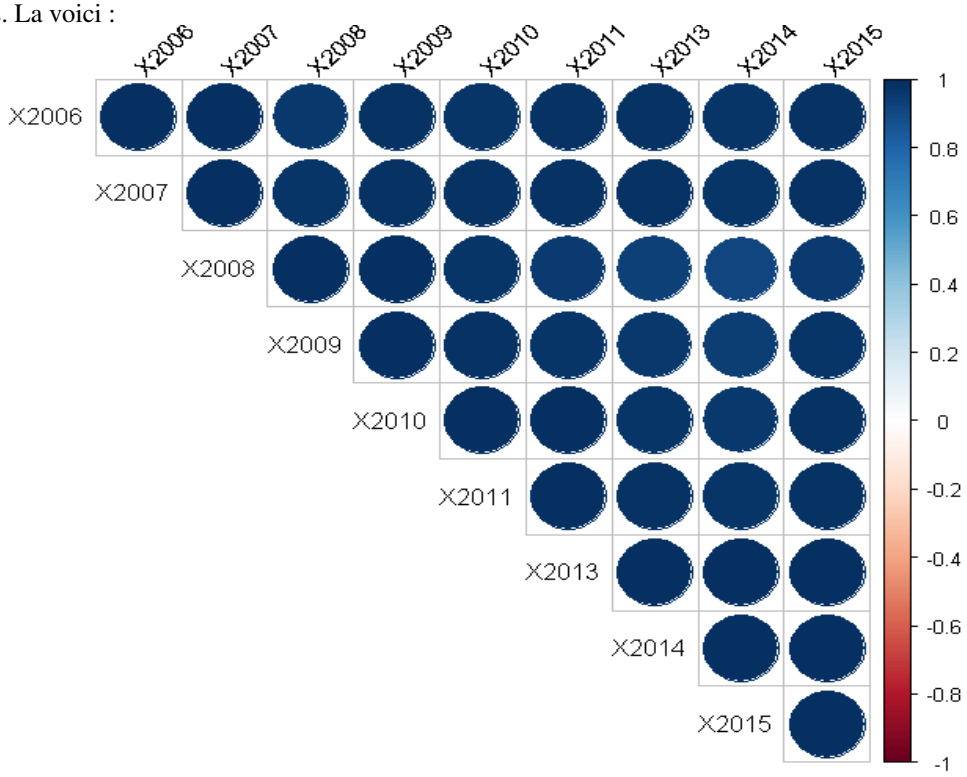


On remarque que la moyenne et les quantiles ont tendance à augmenter au fil des années.



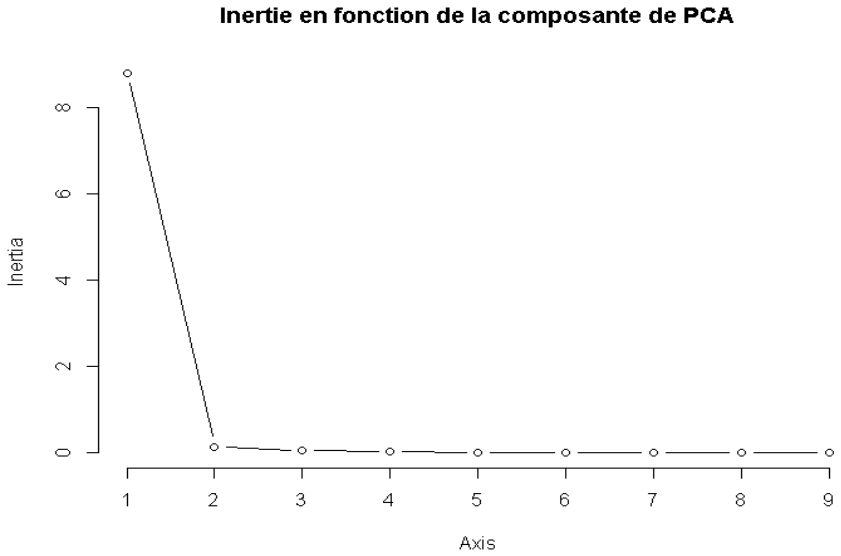
Une fois les données centrées et réduites, on remarque que la moyenne et les quantiles sont plutôt constants au fil des ans ce qui est normal vu que les données ont été réduites.

On a ensuite cherché à connaitre la correlation entre les variables, pour se faire, la matrice de correlation a été calculées. La voici :



Ce graphique représente la moitié de cette matrice, l'autre moitié est inutile car la matrice de correlation est égale à sa transposée. On remarque que toutes les correlations sont très proches de 1.Ceci s'explique par le fait que les données représentent l'évolution du pourcentage de chercheurs employés dans les différentes régions de Belgique au fil des ans. En effet, les valeurs de ce pourcentage d'une année à une autre sont fortement liées/correlées, la valeur de celui-ci à une année est basée sur sa valeur à l'année précédente et n'évolue pas drastiquement. La plus petite correlation est entre l'année 2008 et 2014, celle-ci vaut 0.915, ceci s'explique par le fait qu'il y a surement eu une inversion de tendance entre ces 2 années, que certaines villes ont eu un pourcentage plus élevé et l'inverse pour d'autres, par exemple le brabant wallon est passé de 1.55% en 2008 à 3.1% en 2014, la majorité des autres villes n'ont pas eu une telle augmentation.

On a ensuite déterminé les valeurs propres de la matrice de correlation, les voici : TABLEAU EIGENVALUES On remarque la première est bien plus grande que les autres, en réalité le premier axe représente 97.8% du pourcentage de l'inertie, le 2e ne représente lui que 1.44%. Pour connaitre le nombre de facteurs à retenir, il y a plusieurs critères possibles. Selon le critère de Kaiser, il faudrait retenir celle qui ont une valeur propre plus grande que 1, dans notre cas, il ne faudrait donc retenir que la première. Un autre critère dit qu'il faut garder la valeur avant le "coude" c'est à dire l'éboulement dans le graphe de l'inertie que voici :



On remarque que le coude s'effectue entre la première et la deuxième valeur, selon ce critère il faudrait donc aussi ne garder que le premier facteur.

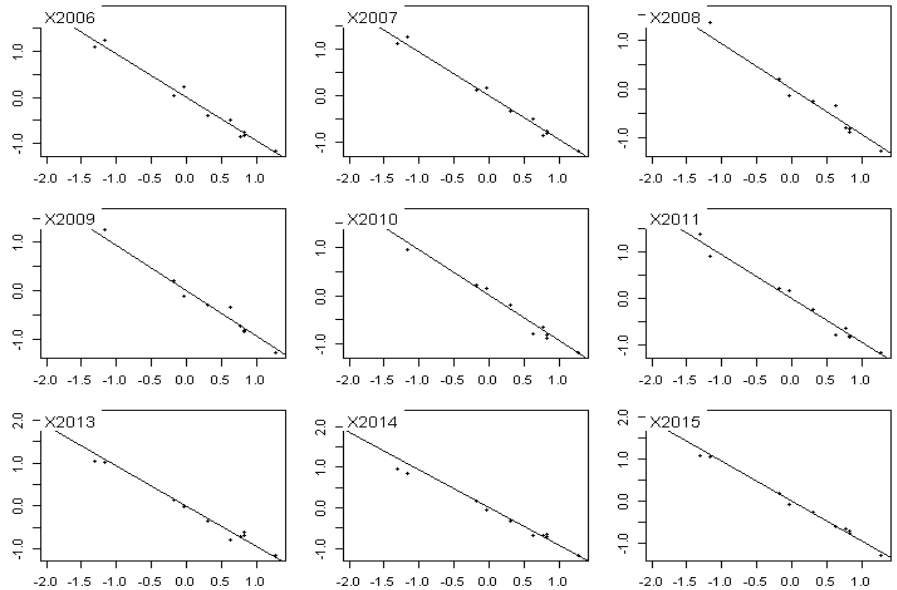
Pour appliquer l'analyse par composantes principales, il a donc été décidé de ne garder qu'une composante.

On obtient donc la seule composante principale : $Z_1 = -33... - 33...$

Voici la droite de corrélation : DROITE DE CORRELATION

Les coefficients de cette composante sont quasi identiques, ceci est le résultat du fait que toutes les variables sont fortement correlées.

Sur le graphique ci-dessous, on observe la ligne de la premiere composante pour chaque variable, on remarque que celle-ci symbolise bien la direction de la plus grande variance dans les données, on voit même que les données suivent un modèle linéaire.



1.3 Le profil est-il stable ?

1.4 Groupement des régions

2 Question 2 : `Analyse Discriminante Linéaire`

2.1 Utilité et