SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Projet analyse de données MAIN 4

Étude statistique de Seshat

Achille Baucher
Laurent Dang-Vu

3 Juin 2020

Contents

1 Présentation des données

Nous avons choisi d'utiliser la base Seshat [1], qui regroupe une grande quantitié d'information concernant diverses civilisations anciennes. Elle est complétée régulièrement par de nombreux spécialistes, archéologues, anthropologues, etc. venant du monde entier. En regroupant toutes ces informations dans une même base, il devient possible de vérifier des hypothèses concernant l'histoire et les organisations humaines. La base principae a été découpée et modelée par des experts afin de faciliter l'analyse de donnée sur certains sujets spécifiques. Les 3 jeux de données ainsi obtenus sont accessible sur le site [2].

1.1 Modification des individus

Comme nous nous sommes d'abord intéressés à analyse non-temporelle, nous avons décidée de regrouper les variables **NGA** (Nom du groupe culturel) et **time** (date) au sein d'une même variable, que nous avons mis en index (Exemple : le groupe *Konya Plain* à la date *1700* devient *Konya Plain (1700)*). Cela nous permet de considérer un individu comme un groupe politique à une période précise. Un même groupe 100 ans plus tard sera donc condidéré comme un individu différent. Cela nous permet de ne pas avoir à faire de moyenne sur une période, ce qui nous paraît très imprudent, et d'augmenter l'effectif de la base. Nous avons effectué ce regroupement avec la librairie Pandas de Python, que nous maitrisons mieux.

1.2 La base Axial Age

1.2.1 Présentation

Nous avons choisi de nous occuper dans les parties 2 et 3 de la base dénomée **Axial age**, spécialement adaptée pour faciliter notre analyse puisqu'elle ne dispose que de variables quantitatives. Celles-ci ont été calculées par des spécialistes à partir de données qualitatives, telles que la présence ou non d'un attribut, pour les convertir en scores dans différents domaines. Par exemple, la présence de textes sacrés apportera des points dans la catégorie **writing** et celle d'un système d'irrigation dans la catégorie **infrastructures**. Certaines variables comme la tailles du territoire et la population de la capitale, qui sont directement quantitatives, ont été renormalisés par un procédé dont nous attendons la réponse de la part des créateurs.

1.2.2 Variables utilisées

- **Polpop** : Polity Population, la population de l'entité politique observée.
- **PolTerr** : Polity territory, la taille du teritoire de l'entité politique.
- **CapPop** : Capital population, la population de la capitale de l'entité politique.
- **levels** : Niveaux de hiérarchies (religieux, militaires, politiques)
- **government** Score du gouvernement (bureaucrates, bâtiments spécialisés, ...)
- **infrastr** : Score des infrastructures (bâtiments publics, pont ...)
- **writing** : Score des textes (littérature religieuse ou scientifique, courrier ...)
- **money** : Score du système monétaire (systèmes de dettes et de crédit, marchés ...)

Si la base originale dispose de nombreuses autres variables concernants les rituels, les aspects moraux etc. , celle-ci se concentre sur les différentes facettes de ce qu'on considère comme un niveau général de "dévelopement".

1.3 La base Morale

Cette base ne contient que des valeurs ternaires, présent (1), absent (0), ou inconnu (0.5). Elle concerne des sujets plus en rapport avec la présence de religions, de morales ou de lois. Les individus, c'est à dire les **NGA** et leurs **time**, sont presque identiques à ce ceux de la précédente. Cependant, nous avons du la réduire pour que tous les individus de cette base soient inclus dans la première, en raison de l'utilisation que nous en faisons dans la partie 5. Le nom des variables est ici beaucoup plus explicite, sur la Figure 1.

Figure 1: Les variables de la base Morale

1.4 Vue d'ensemble

Nous avons commencé par observer les variables de la base Axiale à l'aide d'un BoxPlot afin de visualiser les différences.

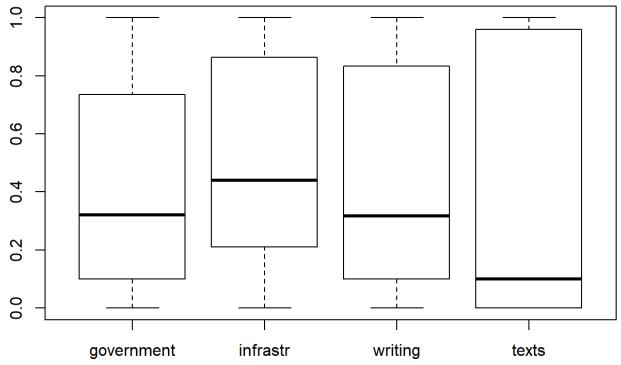


Figure 2: Les petites variables

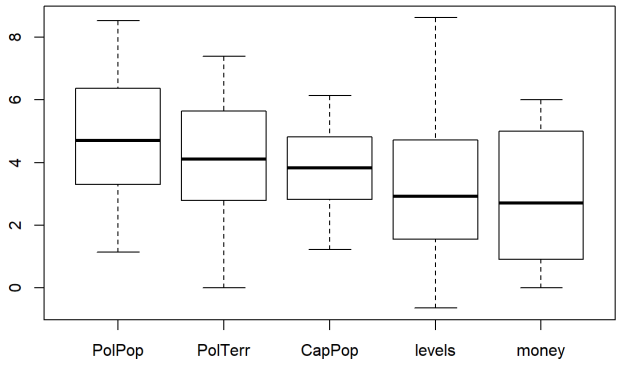


Figure 3: Les variables plus grandes

On peut voir que la moitié des variables est comprises entre 0 et 1 (Figure 2), tandis que l'autre l'est entre 0 et 10 (Figure 3). Il faut noter que ces variables ont été renormalisées par des spécialistes pour que toutes les variables s'apparentent à peu près à un score de même envergure.

2 Analyse en Composantes Principales

Nous avons débuté notre étude par une ACP afin de voir s'il existait des ensembles de variables différenciant plus effficacement les groupes, ce qui pourrait permettre d'identifier des critères, des conditions, et de classifier les individus. Cette idée a été aussi développée dans un article [3] dont nous nous sommes inspirés pour nous orienter et pour les interprétations.

2.1 Explication

Nous avons commencé par représenter sur la Figure 4 la part d'inertie que chacune des nouvelles composantes explique dans l'ACP.

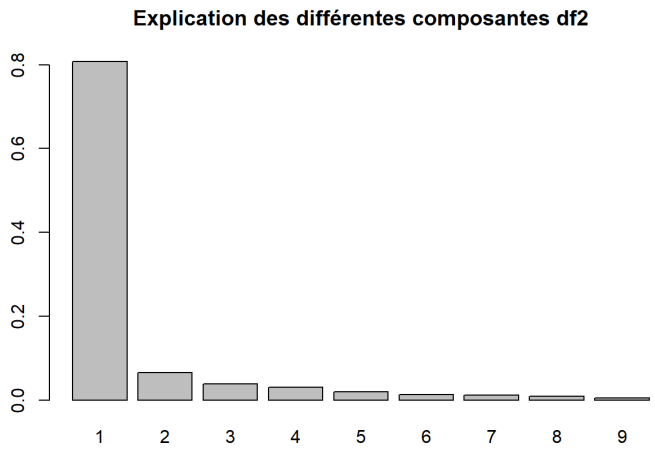


Figure 4: Explication des nouvelles composantes

La Figure 4 montre que l'ACP peut apporter des résultats intéressants, car la part d'inertie expliquée par les différentes composantes est loin d'être régulière. En effet, la première composante explique plus de 80% de l'inertie. Cela signifie qu'on peut très bien différentier les groupes grâce à seulement une seule composante, que nous étudierons plus en détail dans la partie 2.3.

2.2 Projection

Nous avons projeté les individus sur les deux axes principaux de l'ACP sur la Figure 6.

L'affichage des noms rendant le graphique illisible, et ceux-ci étant trop nombreux pour interpréter quelque chose, nous avons décidé de ne pas les afficher. En revanche, cette projection nous a permis de voir qu'on pouvait distinguer à peu près deux groupes, l'un à droite et l'autre à gauche, et peut-être aussi les points dispersés du milieu. Nous avons donc procédé à une classification dans la partie 3.

2.3 Contributions

L'analyse des contributions de la première composante apparaît comme cruciale puisqu'elle nous indiquera selon quelles variables les différents indivius sont le plus différentiables. Nous l'avons donc représentée sur la Figure 5.

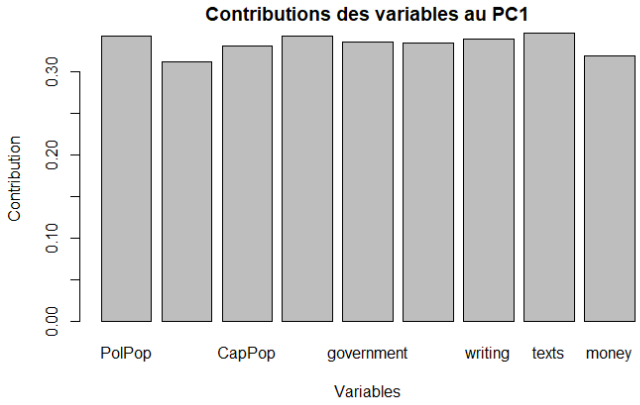


Figure 5: Contribution des variables au PC1

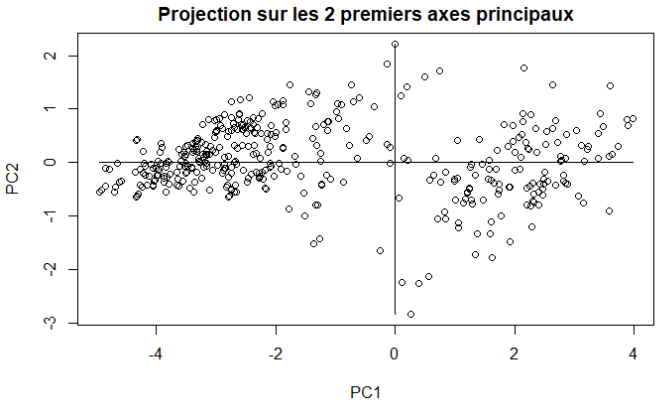


Figure 6: Projection des individus sur les deux axes principaux

On obtient le résultat particulier d'une égalité quasiment parfaite, et positive, entre toutes les variables. Cela signifie que la principale différence entre les individus se joue de la même manière sur toutes les variables, et donc que les variables choisies sont assez auto-corrélées, dans le sens où avoir un bon score dans une d'entre elles signifie en avoir aussi un dans les autres. Vu que les catégories choisies concernent toutes un aspect de développement, la composante principale représente en quelque sorte un score de développement général.

Nous avons aussi représenté les contributions à la deuxième composante sur la Figure 30 en annexe, mais il faut toutefois avoir à l'esprit qu'elle n'explique que 5% de l'inertie.

2.4 Corrélations

Nous avons utilisé le package *factoMineR* pour obtenir le cercle des corrélations. On retrouve pour la première composante principale une corrélation orientée dans le même sens pour toutes les variables, tandis qu'on a une opposition entre les variables de tailles (population,territoires, niveaux de hiérarchies) et les autres dans la deuxième composante. Si l'explication de la deuxième composante n'était pas si faible, et surtout par rapport à la première, on pourrait interpréter qu certains groupes sont plus spécialisés dans l'expansion des territoires et de la population, tandis que d'autres dans le développement des autres critères, mais il faut rester très prudent.

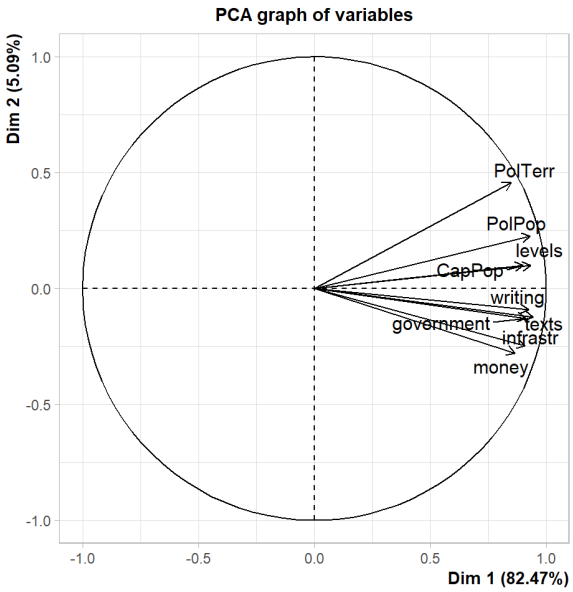


Figure 7: Corrélations entre les différentes variables pour les deux premiers axes de l'ACP

3 Classification

L'analyse en composante principale nous offre à distringuer 2 groupes, l'un au dessus et l'autre en dessous de la moyenne du PC1, et peutêtre un troisième au milieu. Nous avons donc essayé de classifier les individus pour tenter de comprendre ce qui les différencie, et si vraiment on peut considérer un groupe au milieu.

3.1 K-moyennes

Nous avons commencé par utiliser l'algorithme de k-moyennes de R sur toutes les variables de la base, pour 2 puis 3 groupes. Nous avons ensuite affiché les résultats en représentant les individus sur les deux composantes principales de la partie précédente, sur les Figures 8 et 9. Nous remarquons qu'il est assez difficile d'identifier si il ya plutôt 2 ou 3 groupes.

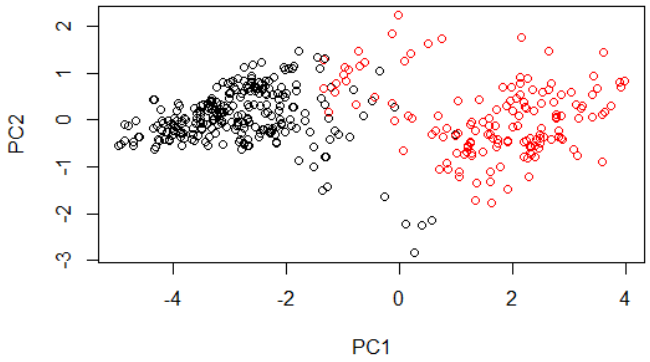


Figure 8: k-means avec k=2

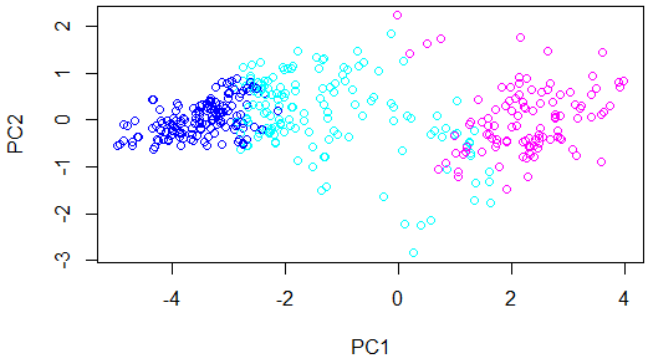


Figure 9: k-means avec k=3

3.2 Dendogramme

Afin de pouvoir choisir entre 2 et 3 groupes, nous avons efectué une Classification Ascendante Hiérarchique (fonction hclust). Nous avons représenté le dendogramme obtenu sur la Figure 10 et l'évolution de l'inertie correspondante sur la Figure 11.

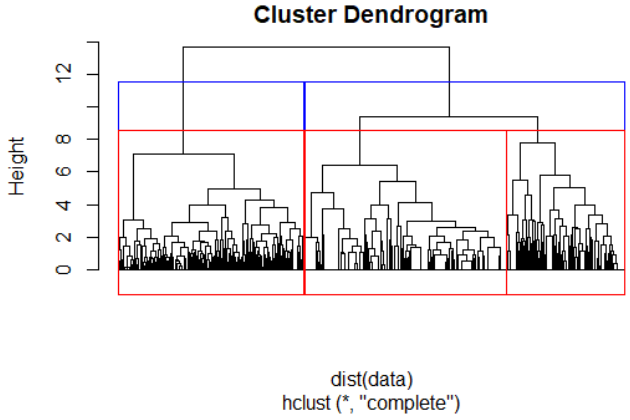


Figure 10: Dendogramme

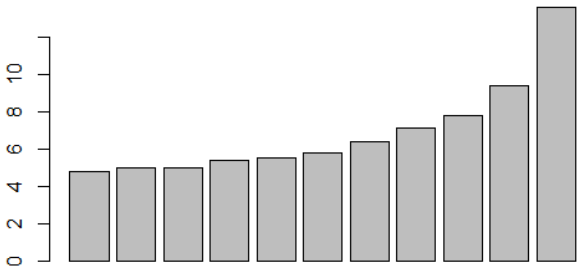


Figure 11: Inertie intra-classe

Dans les carrés rouges correspondant aux 3 groupes, on remarque une très forte densité sur les groupements des côtés, qui est beaucoup plus faible au milieu. Le carré du mileu ne laisse pas vraiment penser qu'il appartient au groupe de droite, comme le suggère l'étape suivante.

On voit que le un saut radicalement plus grand se situe au passage de 2 à 1 groupe, ce qui laisse plutôt penser que deux groupes suffisent pour expliquer les données.

3.3 Comparaisons

Nous avons finalemnt choisi de comparer les groupes obtenus avec la classification avec k=2. Nous avons pour cela affiché les variables des deux groupes dans un même boxplot sur les Figures 12 et 13.

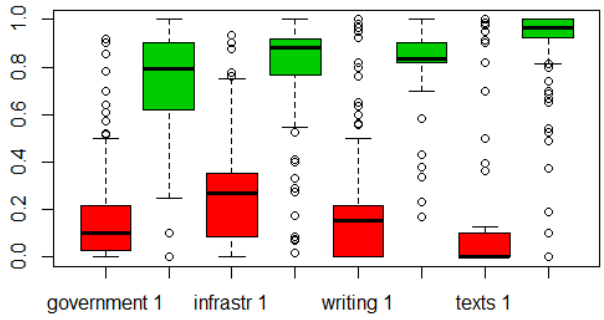


Figure 12: Comparaison des deux groupes

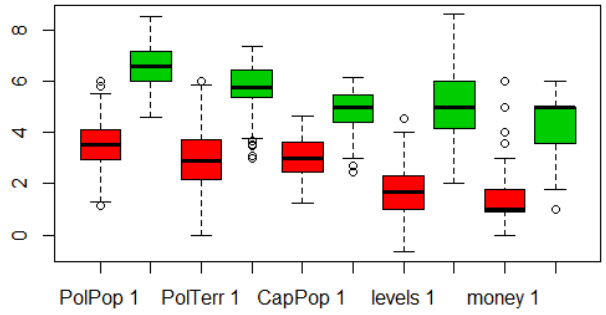


Figure 13: Comparaison des deux groupes

On remarque qu'un groupe a toutes ses variables largement au dessus de l'autre, comme nous pouvions le prévoir au regard des contributions de l'analyse en composantes principales.

4 Classification sur trois époques

Nous avons également essayé de réaliser la même étude à des époques différentes. Cela dans le but de les comparer les unes avec les autres et avec l'analyse globale précédente. C'est-à-dire que nous avons réparti les 864 observations de manière équitable (qui vont de - 9600 à 1900) sur trois différentes périodes :

- De -9600 à -2100 : Préhistoire (288 observations)
- De -2000 à 400 : Antiquité (296 observations)
- De 500 à 1900 : Moyen-Âge et période moderne (280 observations)

On nommera leur sous-jeu de données associé par la suite respectivement par df1, df2 et df3.

Nous nous sommes restreints à trois périodes pour que chaque sous-jeu de données ait assez d'observations pour que son analyse ait du sens. Cela explique la troisième période qui aurait pu elle-même être découpée pour avoir par exemple d'un côté le Moyen-Âge et de l'autre la période moderne.

Une autre limite du jeu de données est que certains pays n'ont que des observations de date appartenant à la troisième. Il est donc difficile d'étudier l'évolution de pays précis. Pour cette raison, nous nous contenterons de répéter l'analyse précédente : classifier les pays. Mais cette fois nous comparerons les résultats des trois sous-jeu de données.

4.1 ACP

4.1.1 Explication

Nous avons représenté sur la Figure 14 la part d'inertie que chacune des nouvelles composantes explique dans l'ACP, ici pour df2. On observe à peu près la même chose chez les autres.

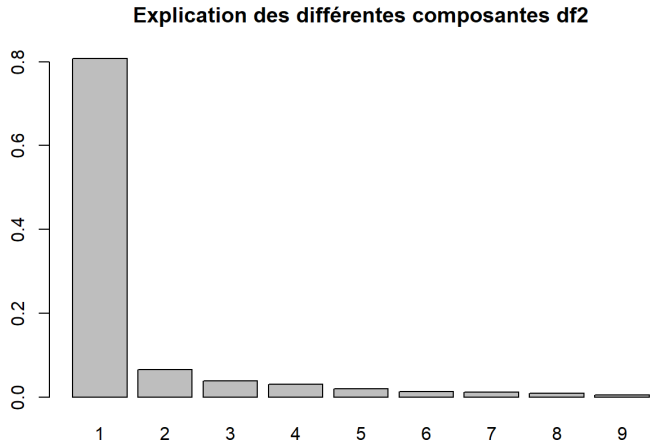


Figure 14: Explication des nouvelles composantes

Encore une fois, la première composante explique plus de 80% de l'inertie. On peut donc encore très bien différencier les groupes grâce à seulement une seule composante, comme nous le verrons plus tard.

4.1.2 Projection

Nous avons projeté les individus sur les deux axes principaux de l'ACP sur les Figures 15, 16 et 31. Cette fois-ci, nous avons choisi les graphiques de FactoMineR comme ils sont maintenant lisibles étant donné que les jeux de donnée sont à chaque fois plus petit. On peut donc facilement choisir de ne représenter que les indidividus bien représentés avec le critère du cos supérieur à 0.8.

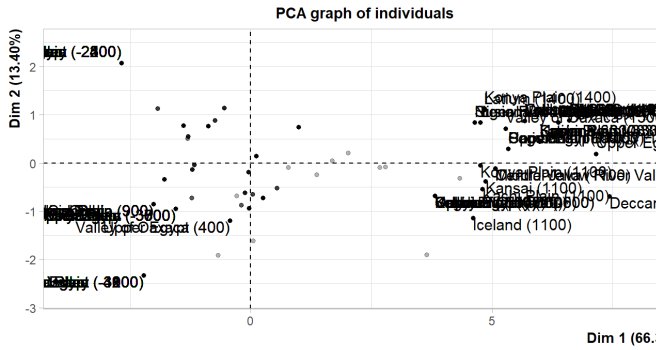


Figure 15: Projection pour df1

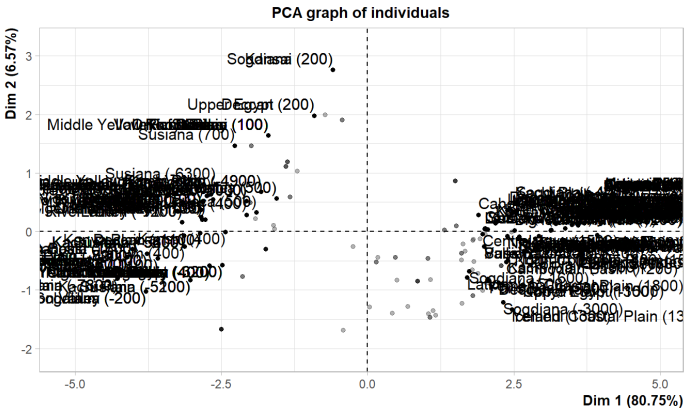


Figure 16: Projection pour df2

On voit que pour les trois jeu de données, on peut distinguer 2 groupes.

4.1.3 Contributions

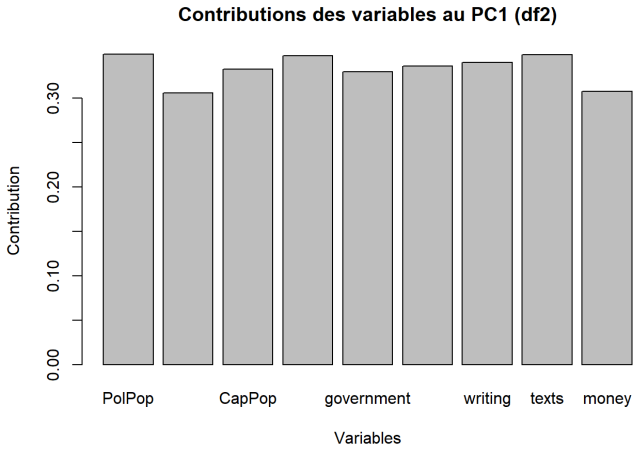


Figure 17: Contribution des variables au PC1

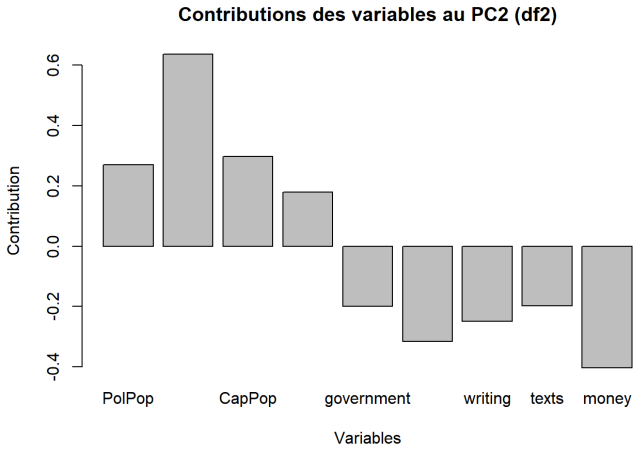


Figure 18: Contribution des variables au PC2

Le résultat de la Figre 17 pour les autres époques est semblable. On retrouve la même caractéristique que dans l'analyse globale : chaque variable contribue positivement autant à la première composante. Par conséquent, aucune n'a une influence positive prédominante par

rapport aux autres. On retrouve auss le fait que toutes les variables sont corrélées dans le même sens. Le critère prédominant de séparation des pays en deux groupes semble au final être le même à chaque époque, signifiant qu'il n'y a pas de changements majeurs sur cet aspects au fil des époques.

Nous avons de nouveau également représenté les contributions à la deuxième composante sur la Figure 18. On observe à nouveau à chaque époque le même résultat qu'au point de vue global : les variables ne sont pas corrélées et certaines prennent des valeurs négatives. Cependant, les variables négatives diffèrent selon l'époque même si l'on remarque une tendance.

4.1.4 Corrélations

Il y a toujours une corrélation orientée dans le même sens (du côté droit du cercle) pour toutes les variables. Toutefois, ce qui change d'une époque à une autre, c'est la corrélation des variables et le fait qu'elle soit plus ou moins accentuée. En effet, d'une époque à l'autre, certaines variables sont corrélées selon la deuxième composante tantôt négativement à d'autres, tantôt positivement. On peut essayer d'en déduire qu'à chaque époque, les caractéristiques d'un pays sont reliées différemment.

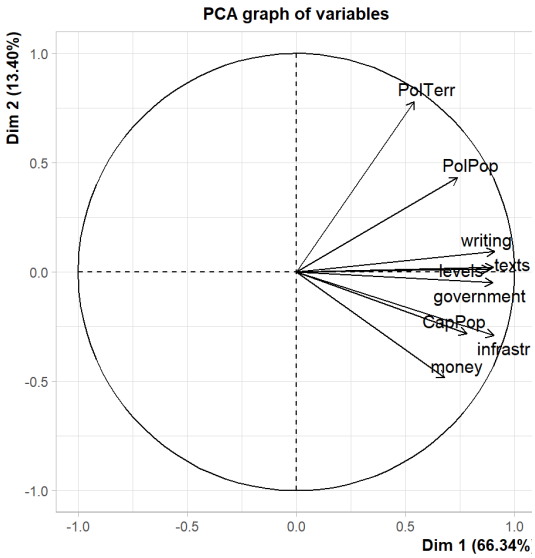


Figure 19: Corrélations pour df1

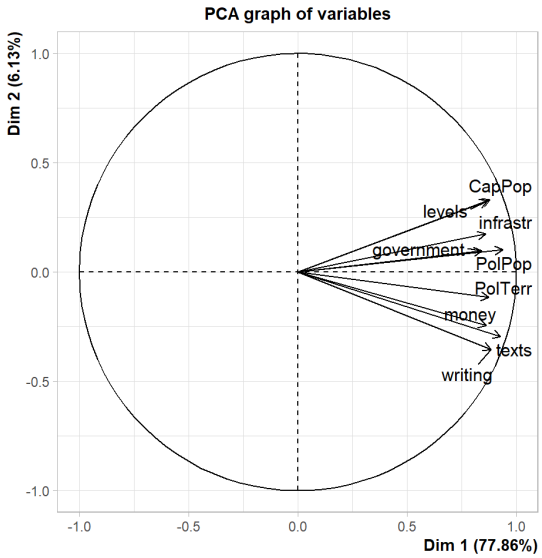


Figure 20: Corrélations pour df3

En conclusion, l'analyse en composante principale nous incite encore une fois à distinguer 2 groupes, l'un au dessus et l'autre en dessous de la moyenne du PC1, et cela à chaque époque. L'absence d'un troisième groupe semble plus claire que dans l'analyse à toute époque confondue.

4.2 K-moyennes

L'algorithme de k-moyennes de R (les Figures 21, 33 et 22) sur toutes les variables de la base pour 2 groupes dans la nouvelle base de l'ACP semble confirmer notre répartition en 2 groupes pour df2 et df3. Cependant, le résultat est plus nuancé pour df1.

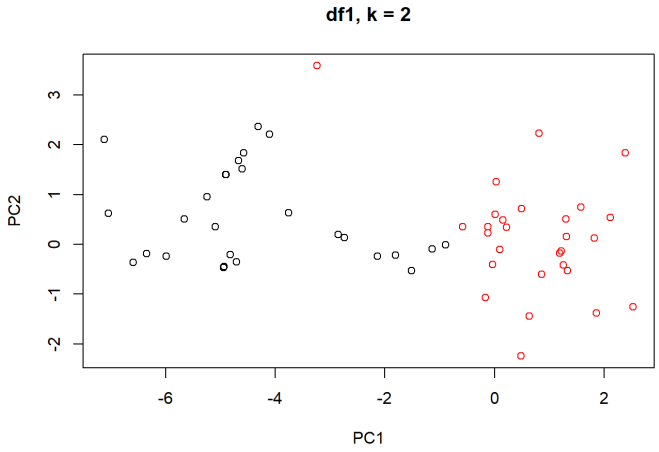


Figure 21: k-means avec k=2 pour df1

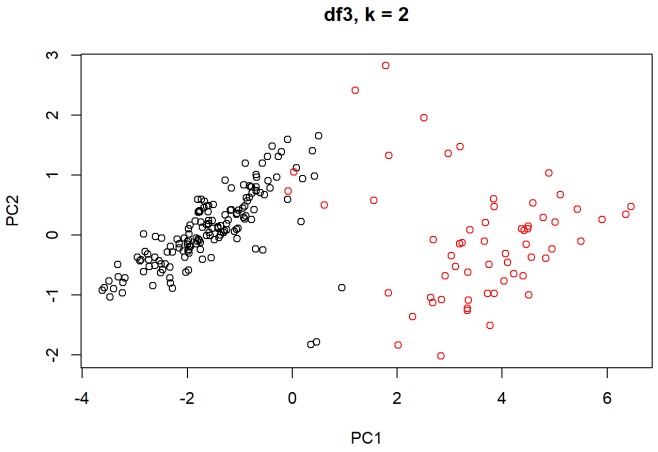


Figure 22: k-means avec k=2 pour df3

4.3 Dendogramme

Le résultat est semblable pour les 3 époques, nous affichons le dendogramme du CAH pour df2 sur la Figure 23 et l'évolution de l'inertie correspondante sur la Figure 24.

Dans les carrés rouges correspondant aux 2 groupes, on remarque une très forte densité sur les 4 sous-groupes ce qui donnerait à penser qu'il y a 4 groupes contrairement à ce que nous verrons au graphique d'inertie.

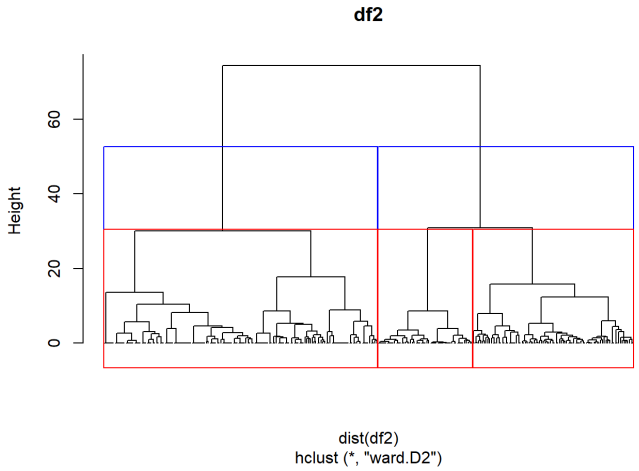


Figure 23: Dendogramme pour df2

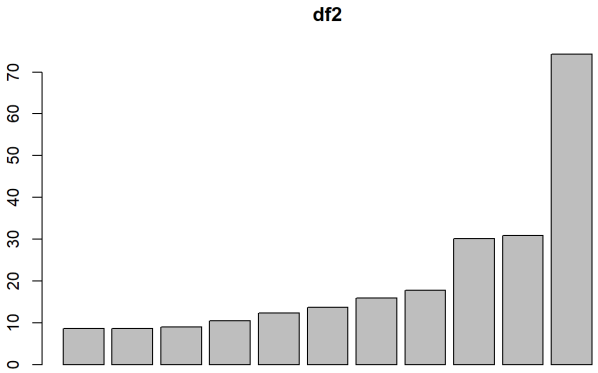


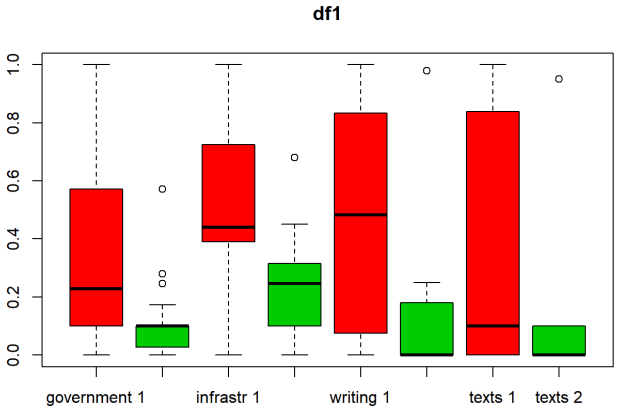
Figure 24: Inertie intra-classe pour df2

De même qu'au niveau global, le saut radical au passage de 2 à 1 groupe justifie le fait que 2 groupes suffisent pour expliquer les données. Le CAH confirme donc Kmeans.

4.4 Comparaisons

Nous avons encore affiché les variables des deux groupes dans un même boxplot pour chaque époque. En ce qui concerne les boxplot des variables représentées sur la figure 26, le résultat

est le même pour chaque époque. Cependant, on a une différence sur la figure 25.



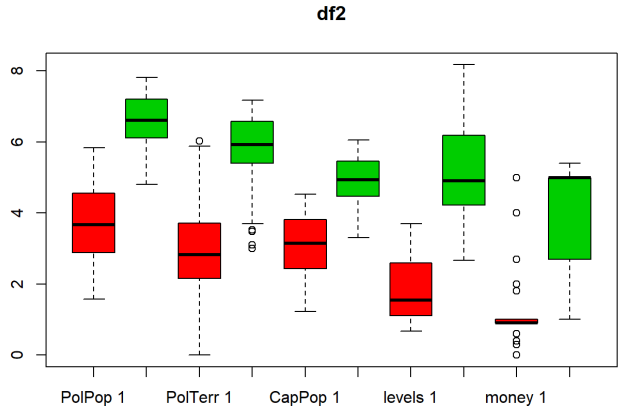


Figure 25: Comparaison pour df1

Figure 26: Comparaison pour df2

On remarque qu'à nouveau que pour df2 et df3, les valeurs des variables d'un groupe sont supérieures à celles de l'autre. Exception faite pour df1 pour certaines variables.

5 Explication avec la base Morale

Nous avons donc pu classifier dans la partie 3 les individus en 2 principales catégories. Cela nous permet à présent d'essayer à présent une autre analyse, qui concernera la deuxième base de données. L'objectif de cette partie est d'observer dans quelles mesures certaines variables de la base *Morale* peuvent expliquer l'appartenance d'une entité dans une des deux classes obtenues précédemment. Nous pourrions connaître alors quelles spécificité sont plus propres à un des groupes qu'à un autre, et savoir si la classification effectuée avec des variables qui consernent le "développement" a aussi un sens pour d'autres types de variables.

Nous avons donc commencé par effectuer une régression linéaire sur cette base pour expliquer la classification précédente.

5.1 Pertinence du modèle

Nous avons obtenu un résultat $< 2.2e{-}16$ au test du Chi2 sur l'Analyse de la variance comparant le modèle sans variables et le modèle avec toutes les variables. Cela signfie qu'il y a au moins une variable significative, et donc qu'un lien existe effectivement entre d'un côté les groupes distingués avec la base Axial et de l'autre les variables de la base Morale.

Cherchons à présent quelles sont les variables les plus pertinentes pour distinguer deux groupes. En observant les contributions individuelles sur la Figure 27, on remarque que c'est la présence d'une **applicabilité générale de la loi** qui est la plus déterminante pour l'appartenance au groupe 1, suivi par **l'absence de punitions morales**.

```
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.9801     0.4426  -2.214 0.026803 *
## X1._Moralistic_punishment         -2.2834     0.8484  -2.691 0.007117 **
## X2._Moralizing_norms              -0.5112     0.6930  -0.738 0.460695
## X3._Promotion_of_prosociality      1.4368     0.5674   2.532 0.011331 *
## X4._Omniscient_supernatural_beings 2.4379     1.2431   1.961 0.049853 *
## X5._Rulers_not_gods               -0.4511     0.6884  -0.655 0.512308
## X6._Equating_elites_and_commoners -0.2055     0.8437  -0.244 0.807553
## X7._Equating_rulers_and_commoners  2.1845     0.9698   2.253 0.024285 *
## X8._Formal_legal_code              0.9356     0.6282   1.489 0.136373
## X9._General_applicability_of_law   2.4217     0.7191   3.368 0.000758 ***
## X10._Constraint_on_executive      -1.5355     0.7283  -2.108 0.035000 *
## X11._Full.time_bureaucrats         0.6503     0.5234   1.243 0.214009
## X12._Impeachment                   1.0425     0.8235   1.266 0.205531
```

Figure 27: Contributions des différentes variables au modèle de régression linéaire

5.2 Arbre CART

La construction d'un arbre CART nous permet ici de mieux se représenter les résultats obtenus.

On y remarque (Figure 28) en effet que le premier critère est la présence d'une **applicabilité générale de la loi**. Cependant, la **punition morale** n'est pas prise en compte, mais c'est la présence d'une **promotion des comportements pro-sociaux** qui est plus déterminante une fois considérée la première variable.

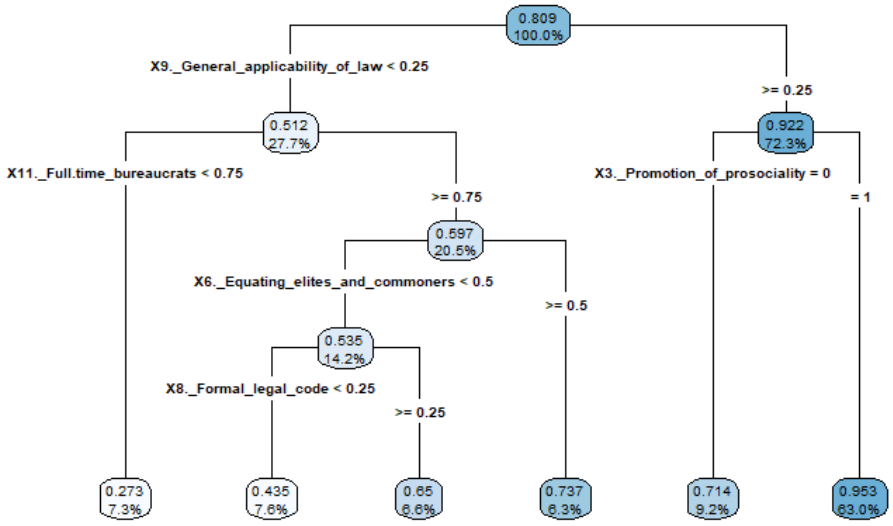


Figure 28: Arbre CART sur la base Morale pour les groupes identifiés

```
##                         (Intercept)          X1._Moralistic_punishment
##                           0.3752810                          0.1019414
##                X2._Moralizing_norms       X3._Promotion_of_prosociality
##                           0.5997592                          4.2074037
## X4._Omniscient_supernatural_beings                X5._Rulers_not_gods
##                          11.4491608                          0.6369282
##  X6._Equating_elites_and_commoners  X7._Equating_rulers_and_commoners
##                           0.8142326                          8.8862247
##               X8._Formal_legal_code   X9._General_applicability_of_law
##                           2.5488605                         11.2649718
##       X10._Constraint_on_executive         X11._Full.time_bureaucrats
##                           0.2153397                          1.9161609
##                   X12._Impeachment
##                           2.8362184
```

Figure 29: Odds-Ratios

5.3 Analyse des Odds-Ratio

On voit sur la Figure 29 que les variables les plus significatives pour l'appartenance au groupe 1 sont la présence d'une **applicabilité générale de la loi** ainsi que de la présence **d'êtres omnicients supernaturels**, à l'opposé de la présence de **punitions morales** et de **contraintes sur l'exécutif**.

6 Conclusion

L'étude a finalement constitué en une application d'outils statistiques grâce auxquels nous avons pu extraire des informations remarquables. Cependant, ces résultats sont trop complexes à interpréter pour nous, et cette tâche revient donc à des experts dans le domaine de l'évolution culturelle.

7 Annexes

7.1 Figures supplémentaires

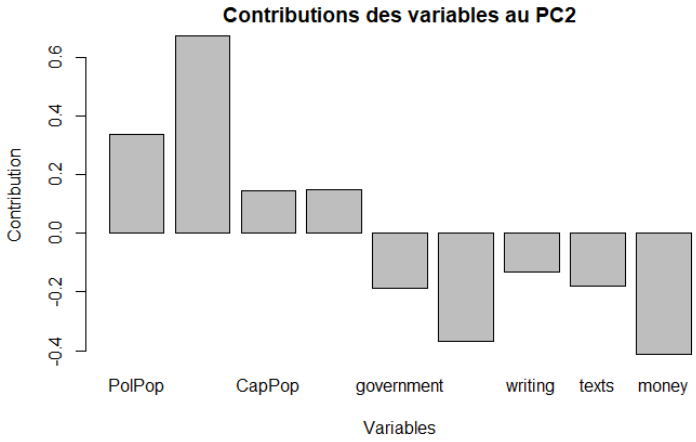


Figure 30: Contribution des variables au PC2

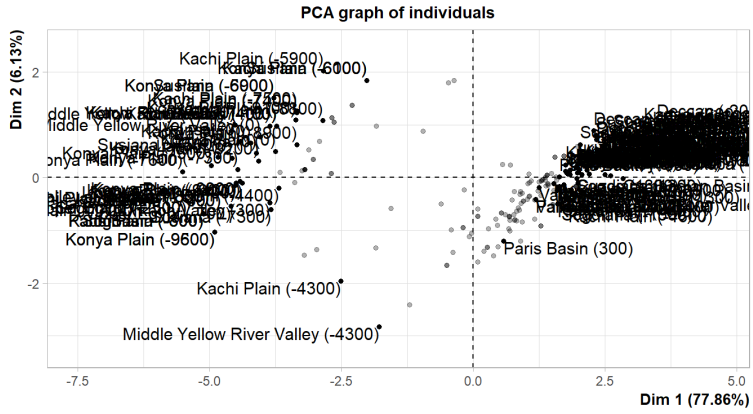


Figure 31: Projection des individus sur les deux axes principaux pour df3

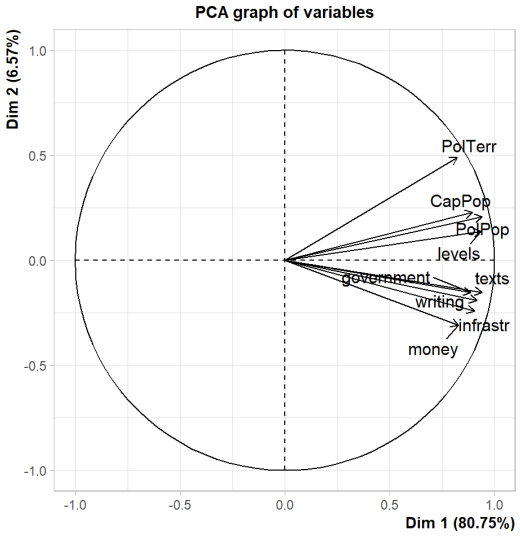


Figure 32: Corrélations entre les différentes variables pour les deux premiers axes de l'ACP pour df2

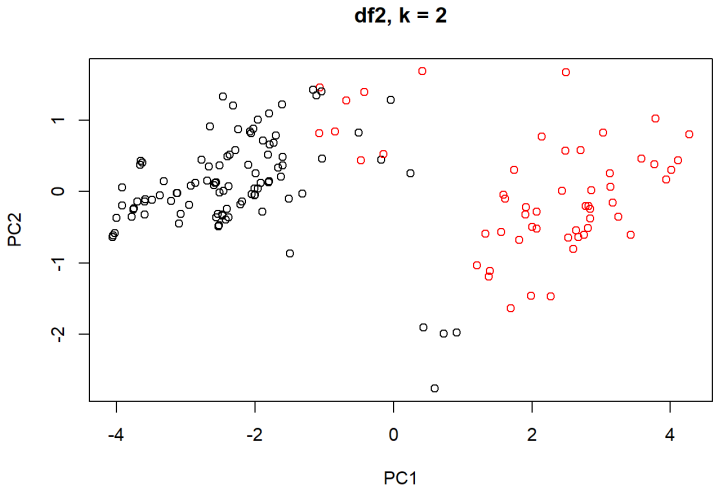


Figure 33: Classification par k-means avec k=2, représentés sur les 2 composantes principales pour df2

7.2 Références

[1] Seshat : `http://seshatdatabank.info/`

[2] Bases de données : `http://seshatdatabank.info/datasets/`

[3] Article de référence : `https://www.pnas.org/content/115/2/E144`