

Compte-rendu du projet fouille de données

BACHATENE Chafik

RODIÈRE Yoann
4IF2

Table des matières

1	Introduction	1
1.1	Contexte	1
1.2	Problématique	1
2	Détermination du champ d'étude	1
2.1	Identification des attributs pertinents	1
2.1.1	Social	1
2.1.2	Santé	2
2.1.3	Armée	3
2.1.4	Démographie	4
2.1.5	Economie	5
2.1.6	Bilan des choix d'attributs	7
2.2	Introduction d'un nouvel attribut	8
2.3	Elimination des outliers	9
2.4	Discretisation de la dimension « liberté »	11
3	Classification non supervisée	13
4	Classification supervisée	16
5	Conclusion	17

1 Introduction

1.1 Contexte

Cette étude s'inscrit dans le cadre du projet Fouille de données. L'objectif est de mettre en oeuvre notre capacité à adopter une démarche efficace et intéressante pour analyser un jeu de données concernant un certain nombre de pays. Nous avons, pour cela, à notre disposition trois jeux de données sous forme de fichiers CSV. Ces trois fichiers contiennent de nombreuses informations sur un certain nombre de pays. Nous avons restreint notre étude au fichier `countries2007_noMissing1.csv` qui contient 88 entrées.

En ce qui concerne les outils utilisés, un logiciel de fouille de données présenté en séance nous permet d'effectuer différentes explorations : il s'agit de KNIME¹.

1.2 Problématique

Les mouvements révolutionnaires de libération que connaît le monde arabe nous ont poussé à étudier de plus près l'impact des libertés civiles sur les situations sociale, économique et militaire de certains pays, ou du moins la corrélation entre ces deux groupes de caractéristiques. Après l'introduction d'un nouvel attribut au jeu de données que nous possédons, nous nous poserons la question de savoir s'il y a un lien entre ce nouvel indicateur (indice de liberté) et les données démographique, militaire, économique, sociale, ou autres, d'un pays. On s'intéressera par la suite à retrouver cet indice de liberté à partir d'un ensemble de données définis durant l'étude précédente. En d'autres termes, nous essaierons de corroborer les différents résultats de la première partie, s'il y a lieu, en retrouvant l'indice de liberté de certains états et en le comparant aux données de notre source d'informations initiale.

1. <http://www.knime.org/>

2 Détermination du champ d'étude

2.1 Identification des attributs pertinents

Cette partie tentera de justifier les choix d'attributs pour leur rapport *possible* avec la liberté civile des peuples. Dans un premier temps, nous commencerons par constituer des groupes d'attributs sous forme de catégories (social, santé, armée, démographie, économie) en justifiant ce choix d'attributs à l'aide de leurs éventuels liens avec la liberté civile des peuples. Après quoi, nous passerons à l'étude des corrélations des différents attributs d'un même groupe pour finalement aboutir au choix définitif.

2.1.1 Social

Adolescent fertility rate Il est possible de supposer — à la limite — qu'un état totalitaire aura des particularités du point de vue de la fécondité des adolescents. En effet, on peut imaginer que le taux d'accès aux études supérieures plus bas influence

Worker's remittances and compensation of employees Reflet direct du niveau de vie des habitants, cet attribut serait dépendant de l'indice de liberté et s'illustrerait par des salaires et des primes très bas au sein d'une dictature militaire.

Internet users Les aspects de censure liés aux dictatures se traduiraient par un usage très limité d'internet. Aussi le nombre d'internautes au sein d'une dictature se réduisant aux membres du régime, on s'attend à trouver une relation proportionnelle entre l'indice de liberté et le nombre d'internautes.

Mobile cellular suscriptions Le fort contrôle des moyens de communication de la part d'un régime totalitaire pourrait induire une utilisation de la téléphonie mobile très limitée.

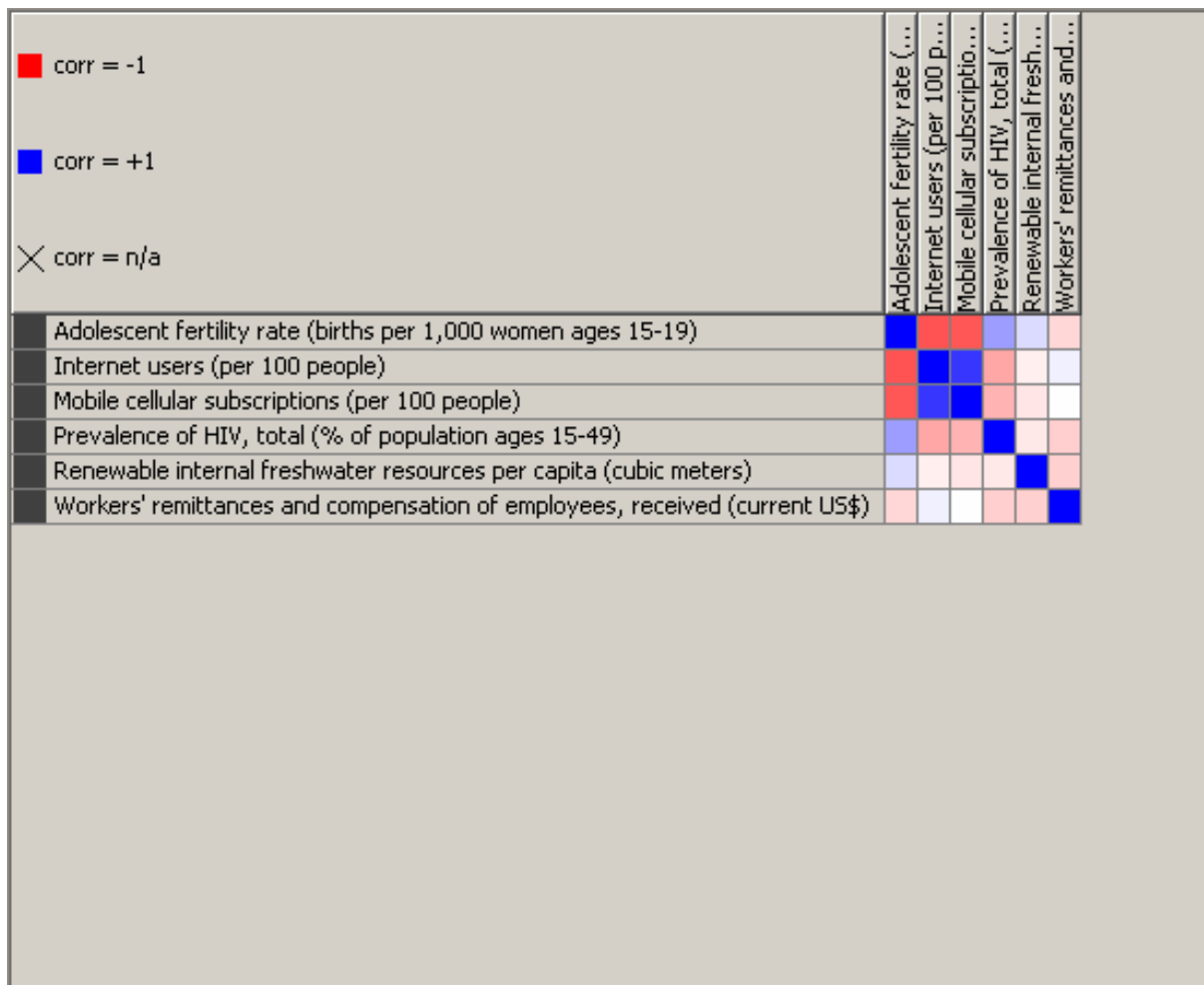


FIGURE 1 – Corrélations linéaires des différents attributs choisis dans la catégorie *Social*

Les attributs ayant trait aux nouvelles technologies de l'information sont linéairement corrélés, pour des raisons évidentes. De manière plus surprenante, la fécondité des adolescents semble se rapprocher d'une fonction linéaire décroissante du taux d'accès à Internet... Sans nous livrer à des conclusions trop hâtives, nous ne considérerons cependant que le taux d'accès Internet en tant qu'attribut représentant les trois.

2.1.2 Santé

Immunization, measles Cet attribut est étroitement lié aux impacts de la liberté sur les aspects économiques d'un pays. Un pays totalitaire accorderait peu d'importance aux achats (imports?) de vaccins contrairement aux investissements de l'armement militaire.

Life expectancy at birth, total On pourrait croire que les libertés d'un pays influencent l'espérance de vie de sa population. Cette hypothèse serait une conséquence des autres attributs qui lieraient liberté aux aspects économiques et sociaux d'un pays. En d'autres termes, moins un pays est libre plus sa situation sociale et sanitaire se dégrade, plus l'espérance de vie décroît.

Mortality rate, under-5 TODO

Prevalence of HIV, total Il est envisageable que le manque de liberté influence l'accès aux soins, par exemple en supposant qu'un pays sous pression politique soit peu développé.

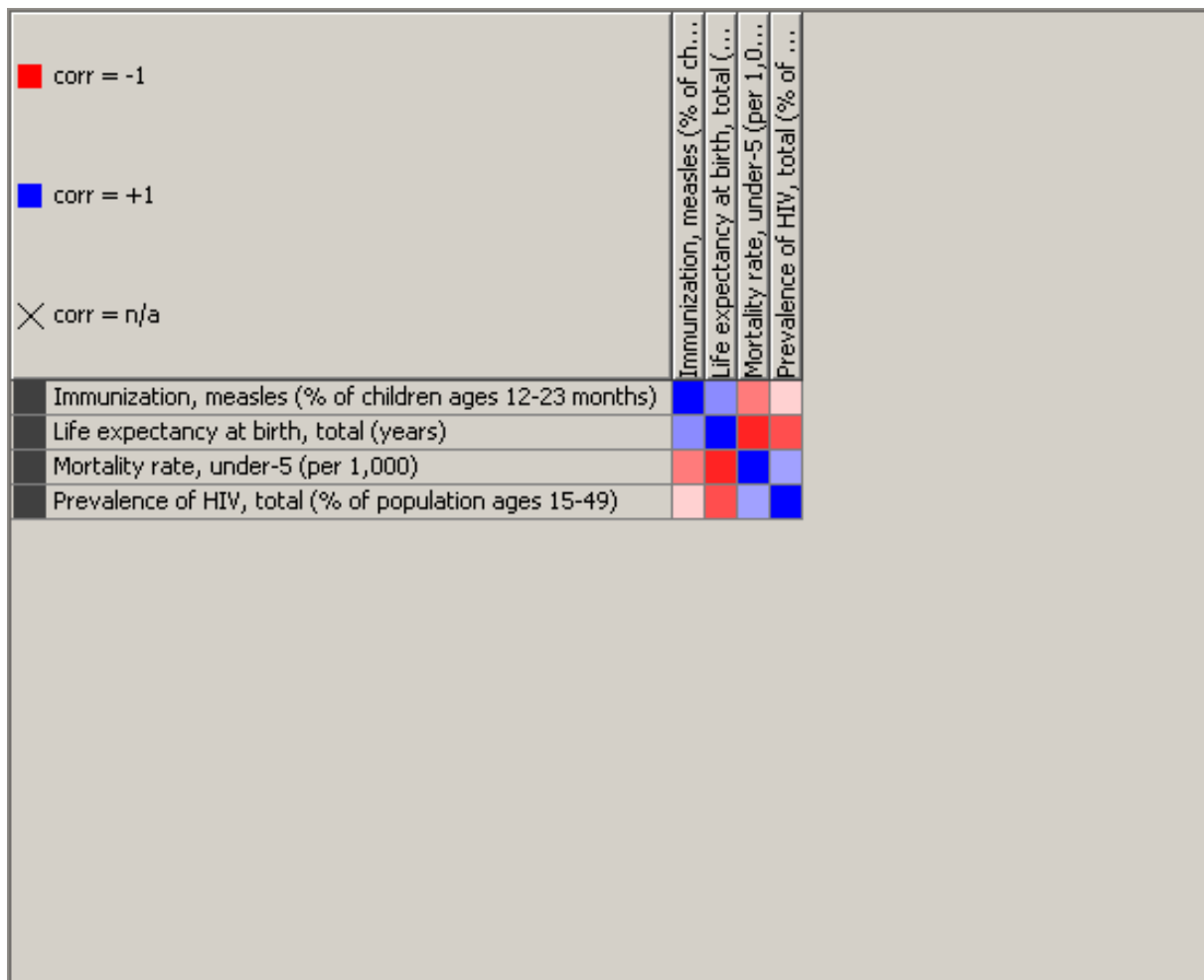


FIGURE 2 – Corrélations linéaires des différents attributs choisis dans la catégorie *Santé*

Comme tous les attributs sont corrélés entre eux (on peut trouver au maximum deux attributs ayant un coefficient de corrélation linéaire relativement faible), il est difficile de continuer l'étude sur ce seul groupe d'attributs. Nous verrons par la suite comment remédier à ce problème.

2.1.3 Armée

Adolescent fertility rate Cf plus haut.

Military expenditure On prévoit de trouver un lien très fort entre les dépenses militaires et les libertés civiles d'un pays. Un pays totalitaire pourrait avoir de très importantes dépenses militaires et inversement.

Fertility rate, total On s'attend à trouver un lien entre le taux de fécondité et la liberté. Un indice de liberté bas se traduirait peut-être par un taux de fécondité bas également et inversement.

Life expectancy at birth, total Cf plus haut.

Surface area Un grand pays en termes de surface pourrait être difficilement contrôlable par une dictature militaire. Il posséderait donc selon toute probabilité une liberté civile plus importante qu'un pays plus petit en surface.

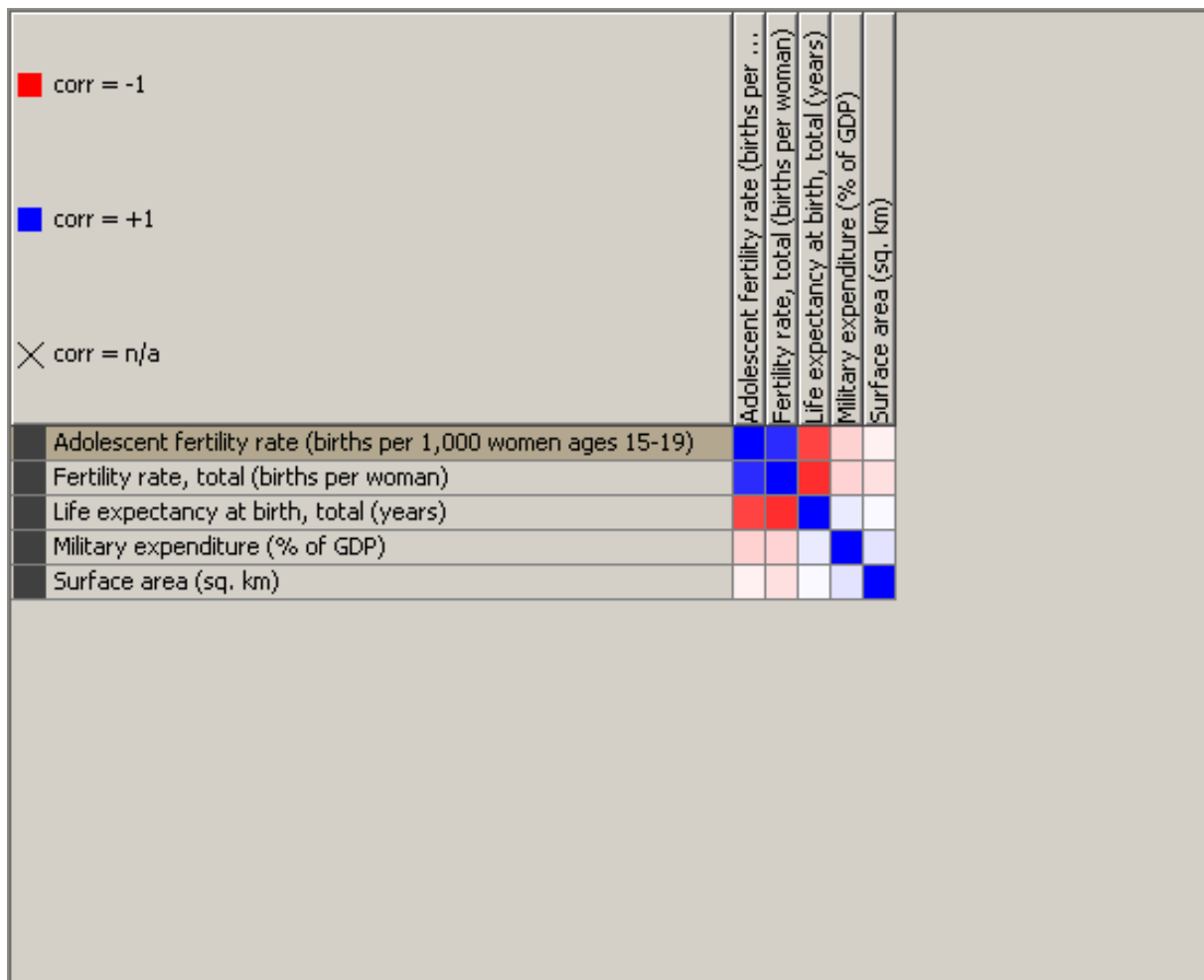


FIGURE 3 – Corrélation linéaire des différents attributs choisis dans la catégorie *Armée*

Sans surprise, le taux de fécondité des adolescentes est fortement lié à celui des femmes en général. De même, l'espérance de vie est fortement liée à ces taux de fécondité. On ne retiendra donc que l'espérance de vie.

2.1.4 Démographie

Fertility rate Cf. plus haut.

Adolescent fertility Cf. plus haut.

Population, total On part de la même hypothèse que celle concernant la surface : plus la population est importante plus elle peut être incontrôlable, donc libre (en excluant des pays singuliers tels que la Chine).

Population growth Tout comme le taux de mortalité, la croissance serait inversement proportionnelle à l'indice de liberté d'un pays.

Life expectancy at birth, total Cf. plus haut.

Surface area Cf. plus haut.

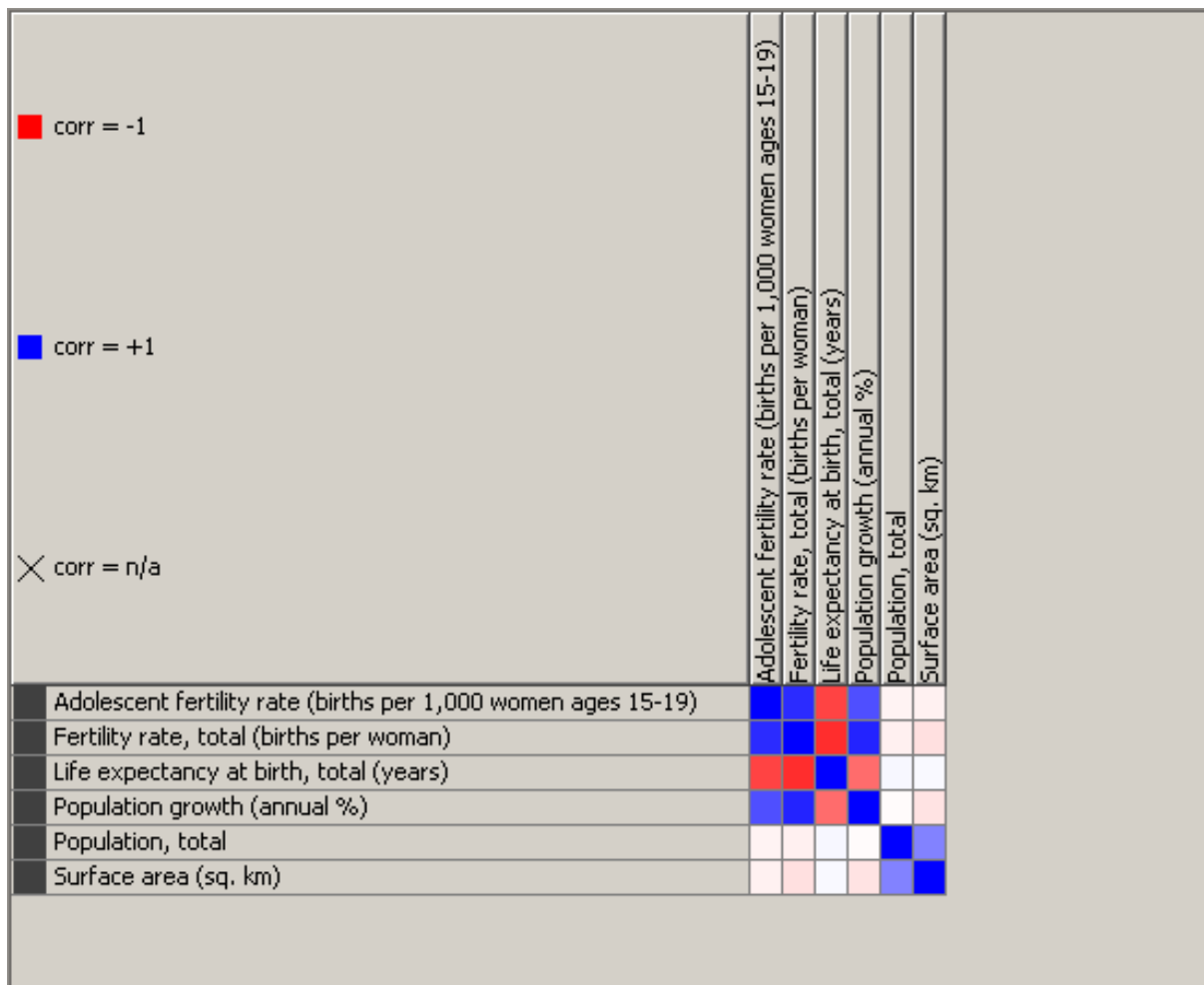


FIGURE 4 – Corrélation linéaire des différents attributs choisis dans la catégorie *Démographie*

Comme dans le domaine de la santé, les attributs sont corrélés dans des proportions déraisonnables. Ce problème est traité plus bas.

2.1.5 Economie

Agriculture, value added Parmi les impacts économiques d'une absence de liberté, on pourrait trouver une très forte participation de l'agriculture dans l'économie d'une dictature militaire.

Exports of goods and services On s'attend à trouver une relation inversement proportionnelle entre la capacité d'une dictature à exporter des services et des biens et l'indice de liberté de celle-ci.

Foreign direct investment, net inflows Idem que pour les exportations.

GNI per capita, PPP Le revenu national brut pourrait dépendre de l'indice de liberté. Un indice de liberté bas pourrait induire une baisse du revenu national brut.

Imports of goods and services Une dictature militaire, exporterait peu mais importerait de manière importante. notamment les matières première et les produits élémentaires.

Industry, value added Au vu des précédentes hypothèses, on pourrait s'attendre à une faible valeur ajoutée industrielle pour une dictature militaire.

Inflation, GDP deflator On suppose que l'inflation est fortement lié à l'indice de liberté si bien qu'une dictature militaire connaîtrait une inflation très importante.

Time required to start a business Un indice de liberté bas représente un obstacle aux jeunes entrepreneurs. Aussi, on s'attend à trouver des temps relativement hauts afin de démarrer une nouvelle entreprise.

Workers remittances and compensation of employees Cf. plus haut.

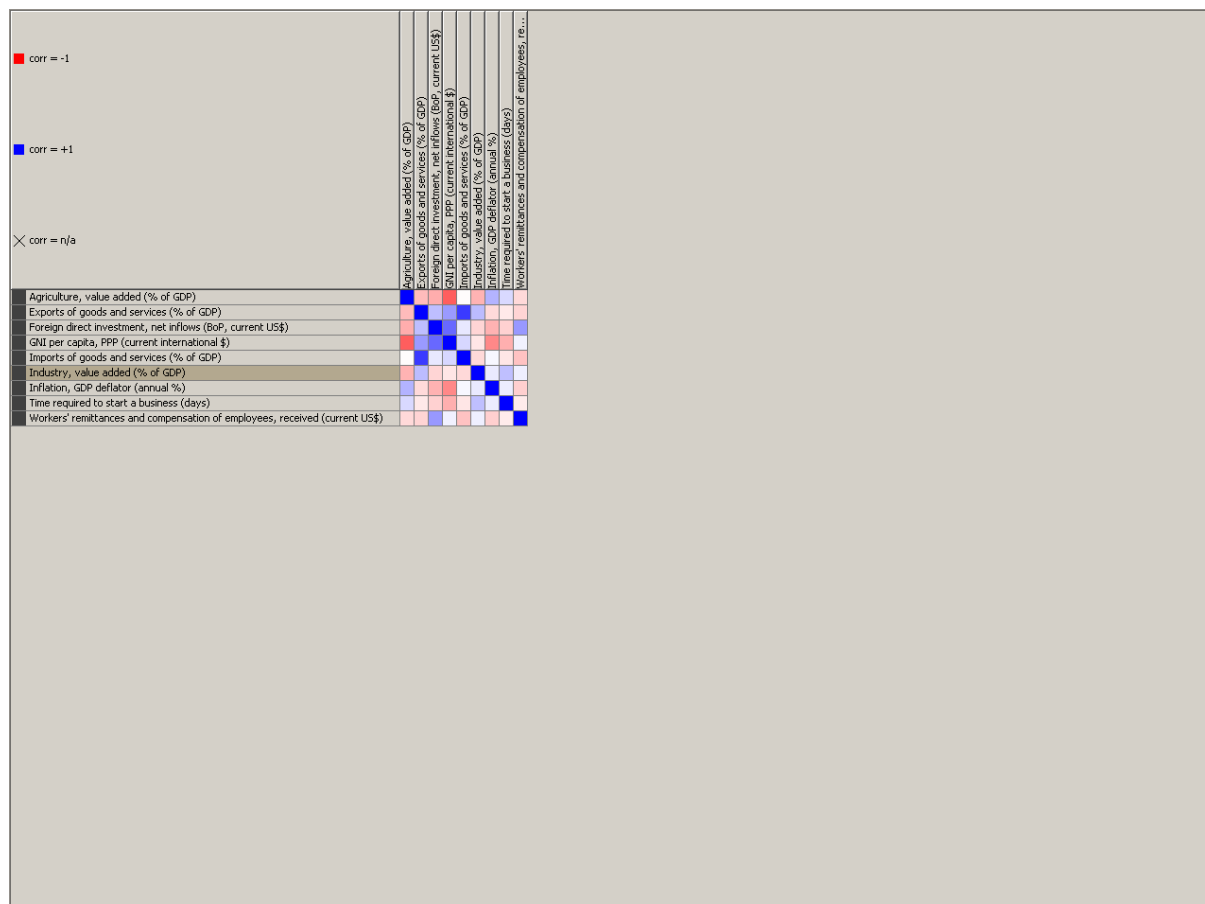


FIGURE 5 – Corrélacion linéaire des différents attributs choisis dans la catégorie *Economie*

Cette matrice montre l'importance de la purification des attributs, qui aboutit à la nouvelle matrice suivante.

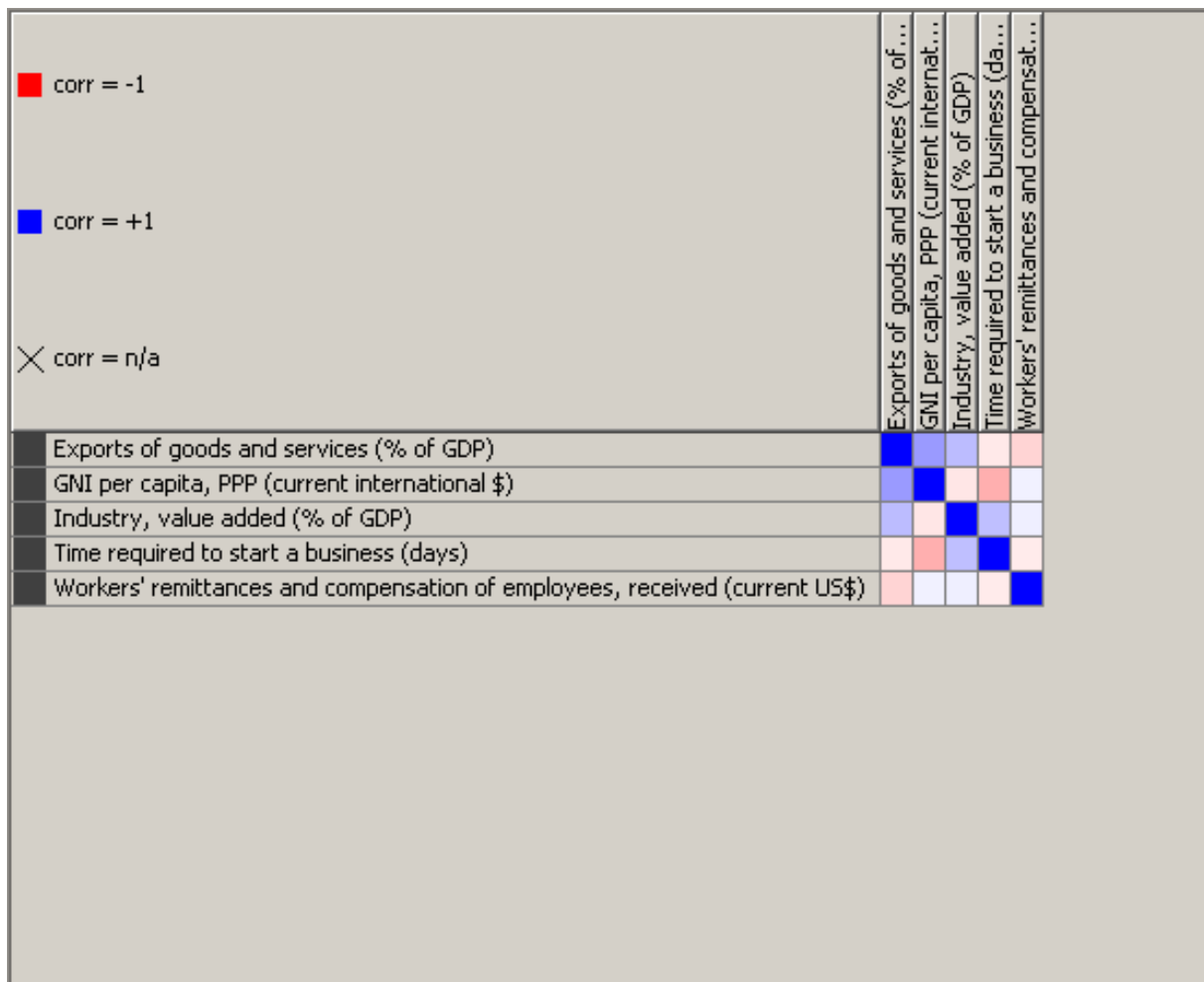


FIGURE 6 – Corrélation linéaire des différents attributs conservés dans la catégorie *Economie*

2.1.6 Bilan des choix d'attributs

Au terme de cette étude des colonnes, nous disposons de 3 jeux d'attributs acceptables — à la limite. Deux jeux ont dû être éliminés faute d'attributs non corrélés en assez grand nombre. Ces données étant quelque peu limitées, nous tenterons d'utiliser un jeu d'attributs de sémantique hétérogène : constitué du résultat de la PCA de chaque jeu d'attributs, il nous permettra de tenter une autre approche : on essaiera de déterminer les pays libres socialement (ou non) en se basant sur les indices produits par les PCA dans différents domaines : « social », « santé », « armée », « démographie » et « économie ».

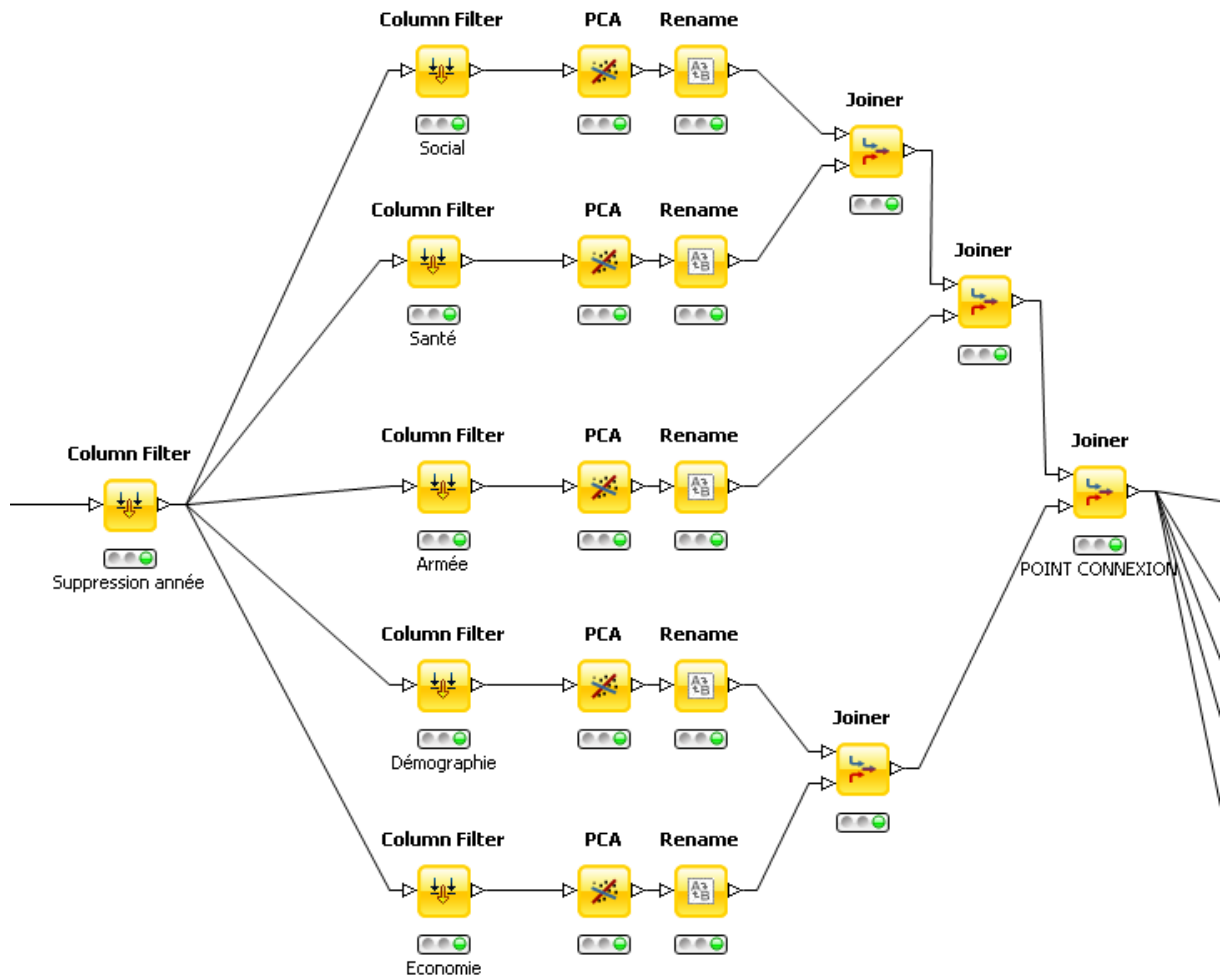


FIGURE 7 – Méthode de construction de l'ensemble d'attributs final

2.2 Introduction d'un nouvel attribut

Notre démarche étant d'étudier le jeu de données fourni associé à l'introduction d'un nouvel attribut qui est l'indice de liberté, nous avons choisi comme source d'un tel indicateur le site internet <http://perspective.usherbrooke.ca/> qui fournit par pays et par année un indice compris entre 1 et 7 (7 représentant une liberté civile quasi inexistante).

L'organisation ayant recueilli ces informations est FreedomHouse et le site les mettant à notre disposition est celui de l'université de Sherbrooke.

Le site internet cité précédemment ne nous fournit pas les données directement exploitables sous Knime. Les informations sont disponibles sous forme de tableaux représentés sur des pages web. Aussi, nous avons réalisé un script en bash qui permet d'extraire du site l'indice de liberté civile des pays présents dans notre jeu de données et de l'intégrer à ce dernier. Il suffira ensuite de corriger les entrées erronées et inexistantes.

```

1  #!/bin/bash
2
3  export IFS=$'\n'
4
5  SRC_FILE="countries2007_all.csv"
6  TMP_FILE="countries2007_liberty.csv.tmp"
7  FINAL_FILE="countries2007_liberty.csv"
8
9  cut -d ',' -f 2,3 "${SRC_FILE}" | sed 's/^(.*$)/\1,/g' > "${TMP_FILE}"
10
11 for LINE in $(cat "${TMP_FILE}")
12 do
13     COUNTRY_CODE=$(echo "$LINE" | cut -d ',' -f 2 | sed 's/"(.*$)"/\1/g')

```

```

14     LIBERTY_INDEX=$(curl "http://perspective.usherbrooke.ca/bilan/tend/${COUNTRY_CODE}/
    fr/PF.CIV.RIGH.POL.IN.html" 2>/dev/null | grep "2007\*\"?</td><td class='
    tableauBarreDroite'>.*</td></tr>" | sed 's,<tr><td class=[^>]*>[^<]*</td><td
15     echo $LINE$LIBERTY_INDEX
16 done > ${FINAL_FILE}

```

2.3 Elimination des outliers

En procédant tout d'abord à une analyse sur chaque dimension, on élimine tout d'abord les premiers et derniers déciles dans chacune d'elles.

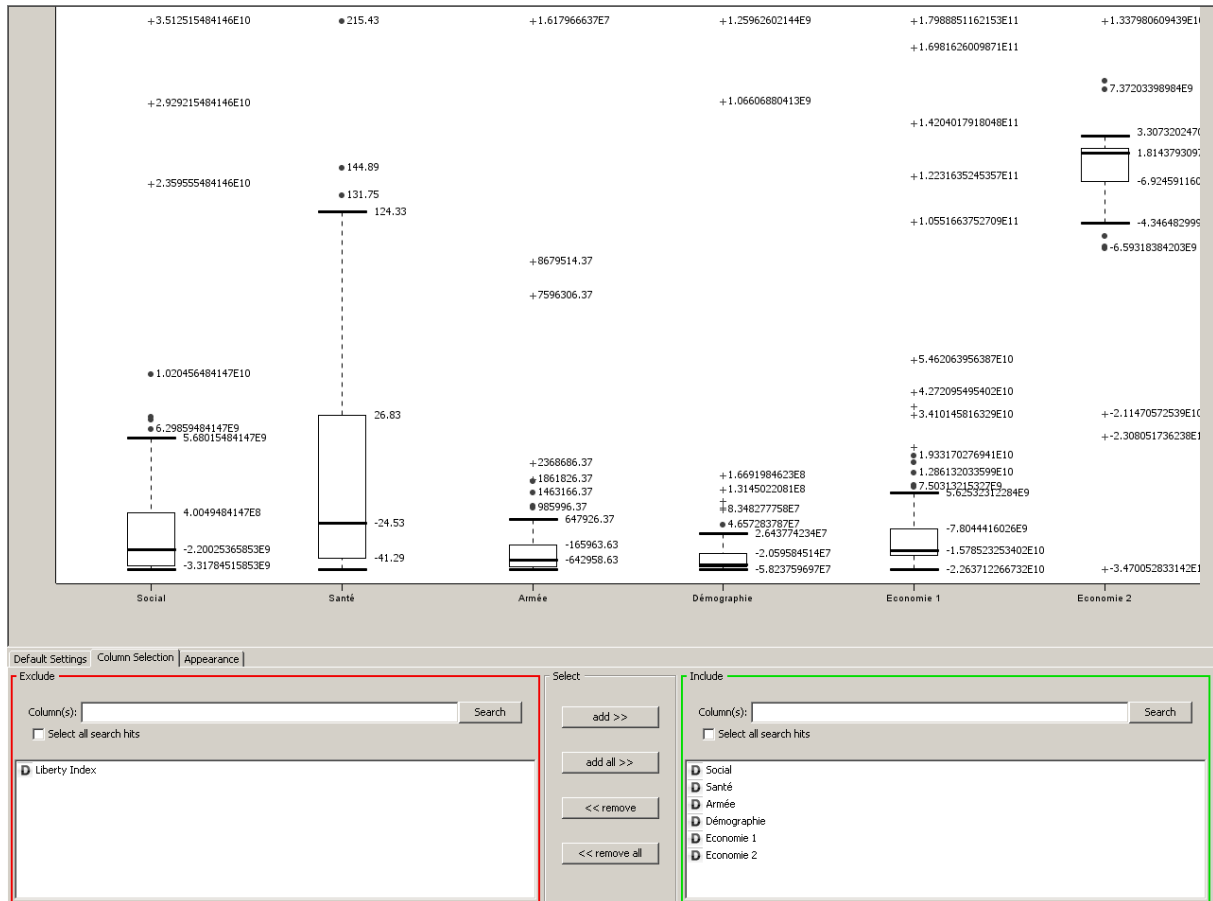


FIGURE 8 – Répartition des pays selon chacune des dimensions obtenues au terme de la PCA

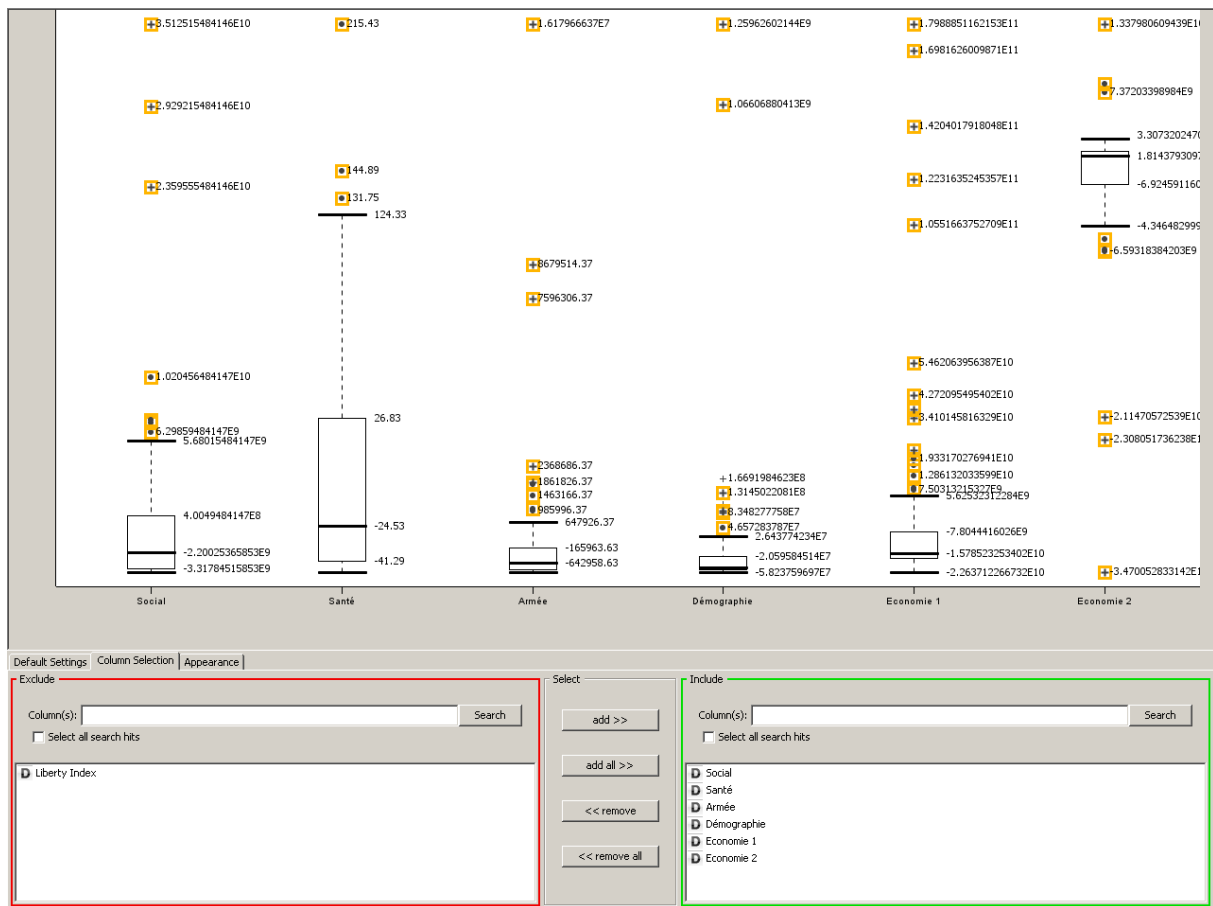


FIGURE 9 – Choix des outliers selon chacune des dimensions obtenues au terme de la PCA

La tentative d'éliminer plus d'outliers via une analyse en deux dimensions échoue, puisque tous les outliers visibles ont déjà été détectés grâce à l'analyse 1D.



FIGURE 10 – Choix des outliers selon chaque couple parmi les dimensions obtenues au terme de la PCA

2.4 Discretisation de la dimension « liberté »

On tente ici de mettre en place un attribut de « classe », un libellé déterminé par l'indice de liberté. Dans un premier temps, il faut déterminer le nombre de classes à créer. Le noeud « Hierarchical clustering » nous y aide : un optimum de 5 clusters se lit directement sur le graphe suivant.

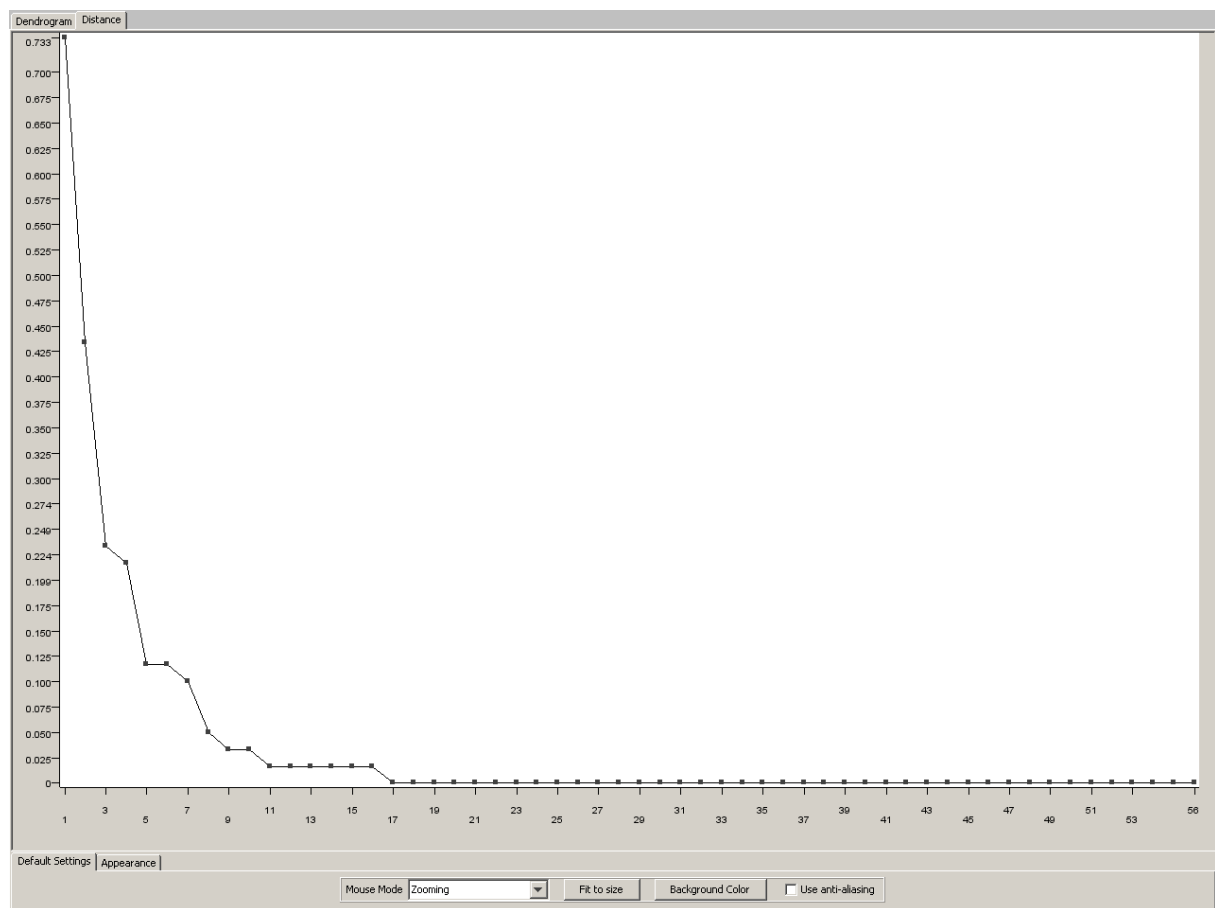


FIGURE 11 – Distance entre clusters en fonction du nombre de clusters demandé (clusters hiérarchiques sur l'indice de liberté)

Cette propriété est flagrante sur le dendrogramme.

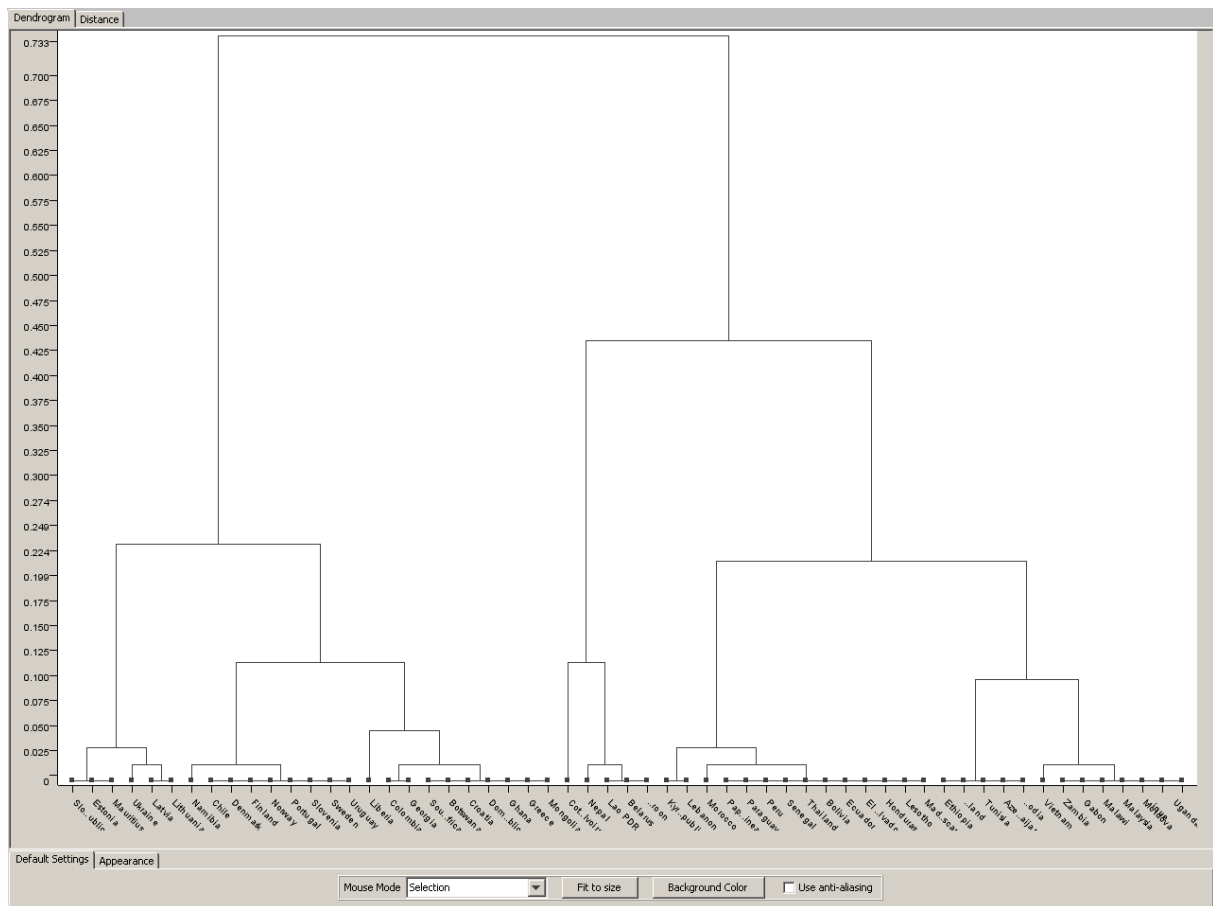


FIGURE 12 – Dendrogramme du clustering hiérarchique sur l'indice de liberté

On utilise le même noeud pour nous fournir ces classes. Ceci fait, nous pouvons commencer l'étude proprement dite.

3 Classification non supervisée

On tente de réaliser trois types de clustering différents sur les attributs choisis. Ceci en espérant que ces clusters correspondront aux classes de liberté, ce qui est mesuré grâce au noeud « Entropy Scorer ».

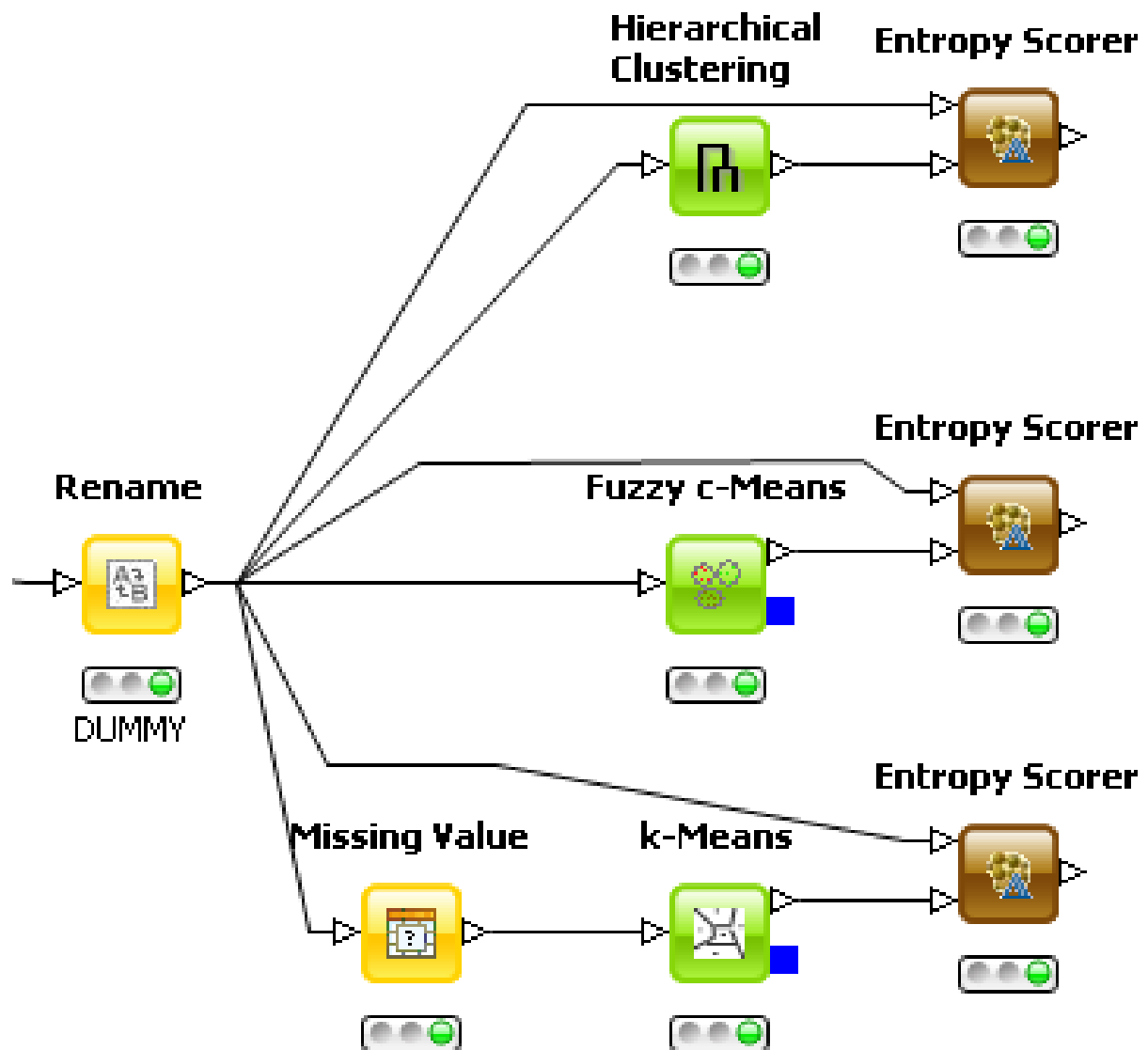


FIGURE 13 – Structure des noeuds réalisant la classification non supervisée dans KNIME

Dans un premier temps, on inspecte une mesure de distances en fonction du nombre de clusters hiérarchiques :

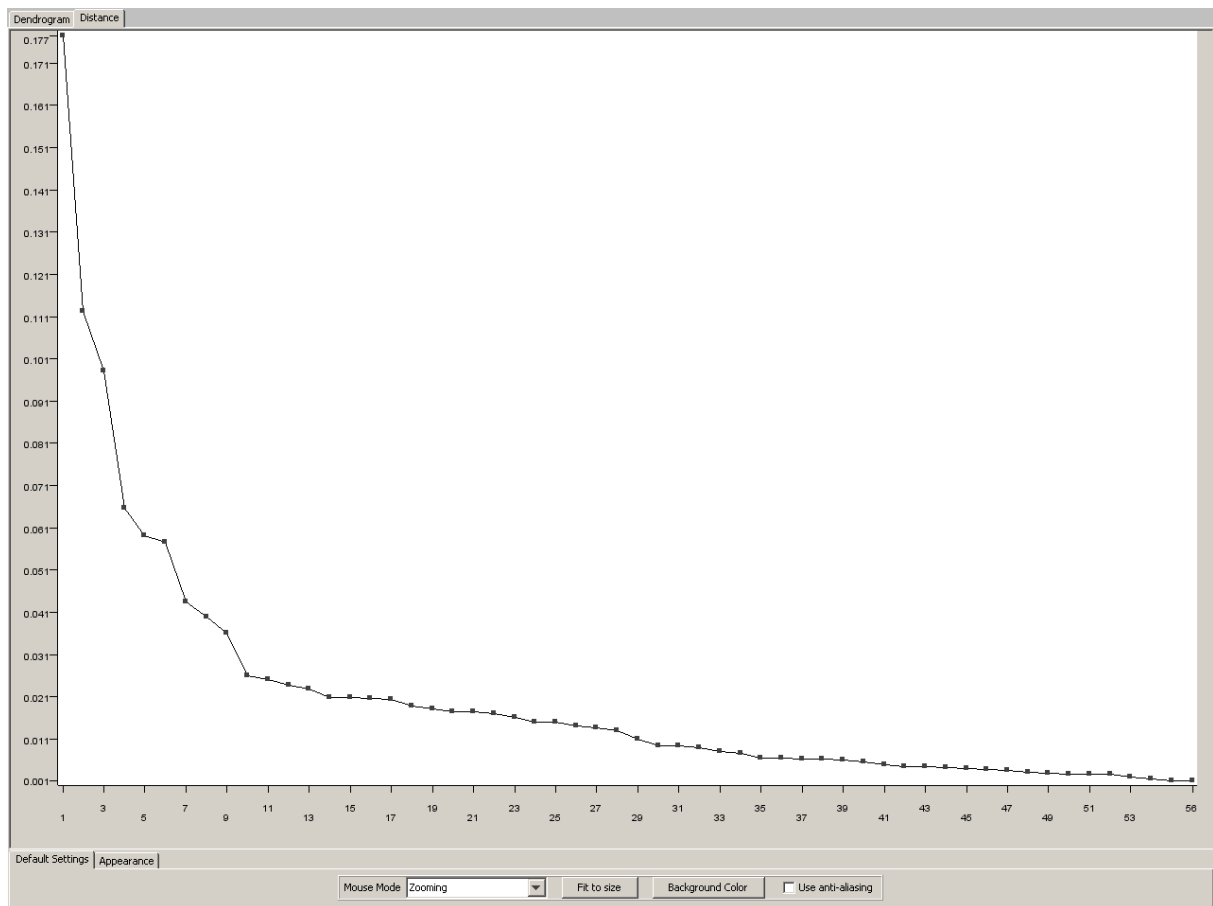
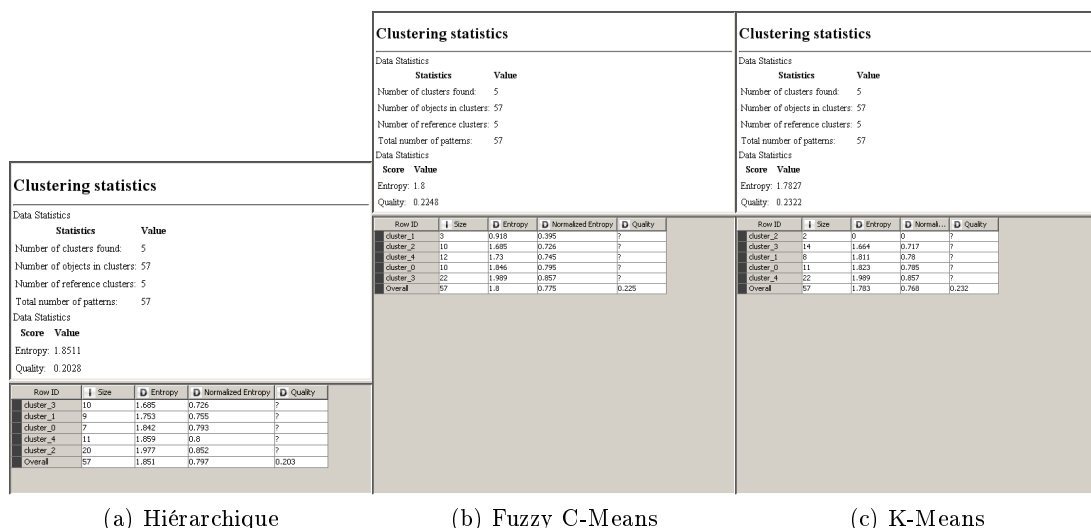


FIGURE 14 – Distance entre clusters en fonction du nombre de clusters demandé (clustering hiérarchique sur les attributs choisis)

Ce diagramme nous inciterait à choisir 4 clusters plutôt que les 5 choisis plus haut pour l'indice de liberté. Cependant, il ne permet pas de se prononcer sur l'adéquation des données aux classes de liberté.



(a) Hiérarchique

(b) Fuzzy C-Means

(c) K-Means

FIGURE 15 – Mesure de l'entropie des différents clusterings avec les classes de références

C'est un échec. L'entropie est clairement énorme et nous montre que le jeu d'attributs choisi ne constitue pas une base intéressante pour déterminer l'indice de liberté : son clustering n'aboutit à rien de semblable aux classes de liberté.

En étudiant sur les jeux d'attributs « acceptables à la limite » vus plus haut (*Social*, *Armée* et *Economie*), on aboutit au même type de résultats.

4 Classification supervisée

Le principe est cette fois de réaliser un arbre décisionnel qui synthétisera la relation entre classe de liberté et les autres attributs. Compte tenus des résultat obtenus en classification non supervisée, il est probable que cette tentative n'aboutisse pas.

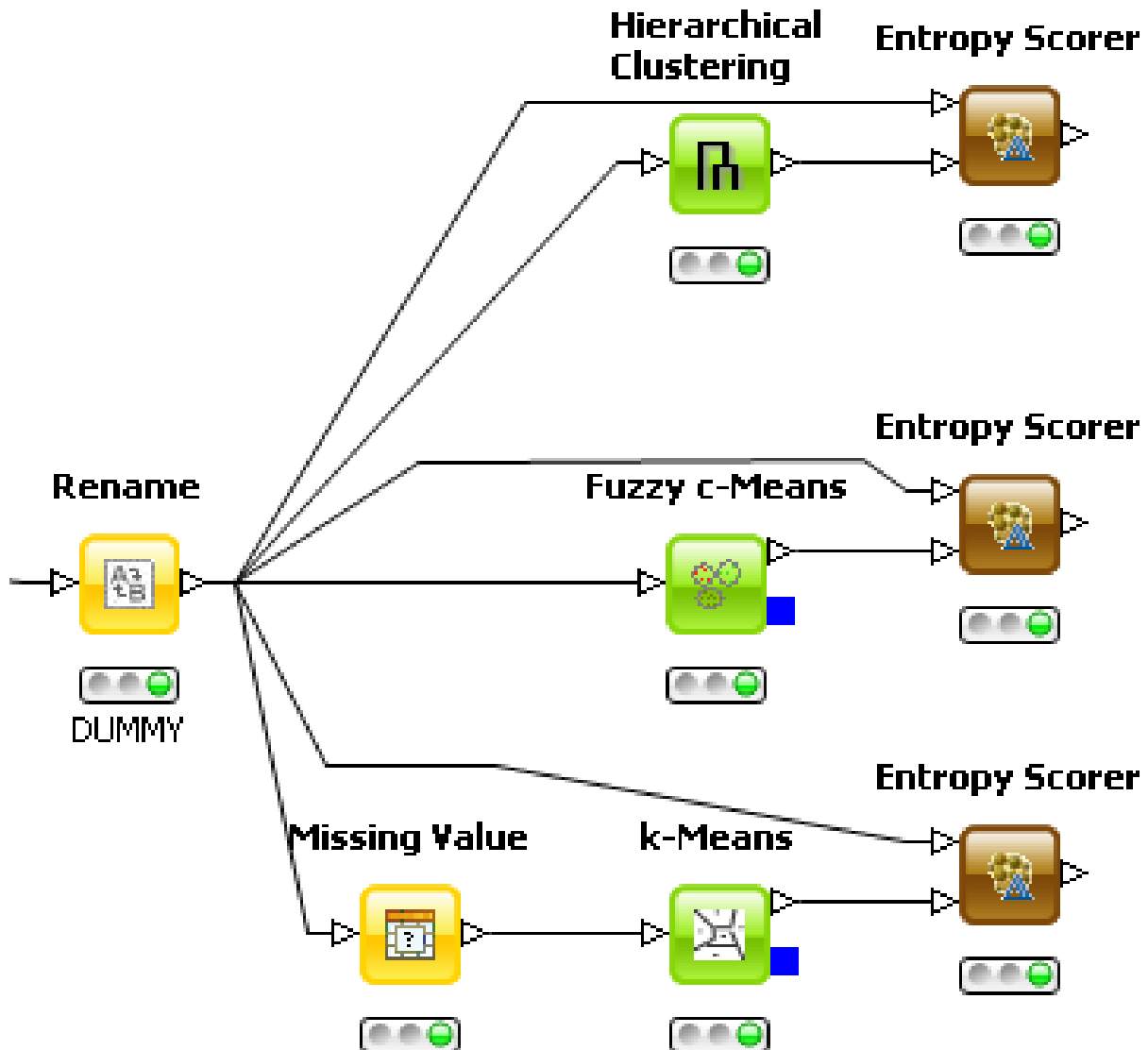


FIGURE 16 – Structure des noeuds réalisant la classification supervisée dans KNIME

On obtient les statistiques d'erreurs suivantes.

Row ID	D Error in %	Size of Test Set	Error Count
fold 0	50	6	3
fold 1	50	6	3
fold 2	33.333	6	2
fold 3	80	5	4
fold 4	100	6	6
fold 5	83.333	6	5
fold 6	60	5	3
fold 7	66.667	6	4
fold 8	83.333	6	5
fold 9	60	5	3

(a) J48

Row ID	D Error in %	Size of Test Set	Error Count
fold 0	66.667	6	4
fold 1	66.667	6	4
fold 2	100	6	6
fold 3	20	5	1
fold 4	33.333	6	2
fold 5	66.667	6	4
fold 6	80	5	4
fold 7	100	6	6
fold 8	83.333	6	5
fold 9	80	5	4

(b) JRip

Row ID	D Error in %	Size of Test Set	Error Count
fold 0	83.333	6	5
fold 1	50	6	3
fold 2	100	6	6
fold 3	60	5	3
fold 4	50	6	3
fold 5	66.667	6	4
fold 6	80	5	4
fold 7	66.667	6	4
fold 8	83.333	6	5
fold 9	100	5	5

(c) KNN

FIGURE 17 – Mesure du taux d’erreur des arbres décisionnels obtenus via 3 méthodes

Encore une fois, on constate de très mauvais résultats. Les tests aboutissent à de nombreuses erreurs (souvent plus de 50%), ce qui, pour un ensemble de 5 clusters, est à peine meilleur qu’un choix aléatoire.

5 Conclusion

TODO

D’un point de vue plus personnel, cette étude nous a appris combien le choix des attributs de départ est important, mais difficile et chronophage. Il nous a semblé que c’est là que réside tout le défi d’une fouille de donnée propre et fructueuse : nous y avons passé beaucoup de temps, mais paradoxalement c’est là que tous les problèmes de l’étude semblent résider.