

## Table des matières

<b>1</b>	<b>Description des données</b>	<b>1</b>
<b>2</b>	<b>Méthode de travail</b>	<b>2</b>
2.1	Filtrage des données . . . . .	2
2.1.1	Réduction des dimensions . . . . .	2
2.1.2	Valeurs manquantes . . . . .	3
2.1.3	Outliers . . . . .	3
2.2	Détermination du nombre de <i>clusters</i> . . . . .	6
2.3	Clustering . . . . .	6
2.3.1	Hierarchical Clustering . . . . .	6
2.3.2	K-Mean . . . . .	6
2.3.3	Fuzzy C-Mean . . . . .	6
<b>3</b>	<b>Secteurs d'activité des différents pays</b>	<b>7</b>
3.1	Sélection des données . . . . .	7
3.2	Clustering . . . . .	7
3.2.1	Clustering hiérarchique . . . . .	7
3.2.2	Clustering K-Means . . . . .	7
3.3	Interprétation . . . . .	7
3.3.1	Arbre de décision . . . . .	7
<b>4</b>	<b>Création d'un module de vérification du K-Mean</b>	<b>8</b>

## 1 Description des données

Le jeu de données provient du site <http://worldbank.org>. Il concerne 210 pays et fournit des informations comme :

- Le pourcentage du **Revenue National Brute (RNB)** généré par l'agriculture, l'industrie ou les services.
- La quantité de produits importés et exportés.
- Des informations démographiques et géographiques comme la densité et l'âge de la population ou la surface du pays.
- Des informations sanitaires comme la quantité de personnes infectées par le **Virus d'immunodéficience Humaine (VIH)**.
- Des centaines d'autres indicateurs.

Par ailleurs, ce jeu de donnée semble être un très bon entraînement pour nous exercer au *datamining*. En effet, en plus d'exhiber des données très concrètes pour nous, il est évident que des corrélations sont présentes.

Enfin, il pourra être très intéressant d'observer ces corrélations d'abord à l'échelle mondiale, puis à l'échelle de l'Europe par exemple.

## 2 Méthode de travail

Cette partie va tenter de décrire très précisément la démarche de travail que nous avons établi au début de ce TP, notamment pendant les deux séances préparatoires. En plus de vous indiquer quelles sont les étapes que nous avons suivies pour arriver à nos résultats, elles constitueront une base de travail solide que nous pourrions améliorer tout au long des deux séances de *datamining*. Enfin, elle seront à nos yeux un indicateur important quant à la rigueur de notre travail et pourra, après correction nous être d'une grande aide.

### 2.1 Filtrage des données

Contrairement à ce que nous avons cru avant de commencer, il faut très souvent élaguer son *dataset* afin de donner plus de sens aux valeurs. Étudions les différentes phases et techniques associées à ce pré-traitement.

#### 2.1.1 Réduction des dimensions

Une des problématique majeure du *datamining* s'appelle la *malédiction de la dimensionnalité*. Cette « malédiction » provient du fait que plus le nombre de dimensions (i.e. le nombre de colonnes) du jeu de donnée croît, moins la notions de distance<sup>1</sup> n'a de sens. En effet, les distances ont tendance à se réduire proportionnellement avec le nombre de dimensions. Or, les algorithmes de *clustering* utilisent cette distance comme critère principal de choix à l'appartenance à tel ou tel *cluster*. Les résultats risquent donc fortement de perdre en sens et en précision.

Voyons alors comment nous pouvons, dans le cadre d'un jeu de donnée contenant beaucoup de colonne, palier à ce défi technique.

#### Choix des colonnes

Dans un premier temps, il est fondamental de ne sélectionner que les colonnes qui nous intéressent.

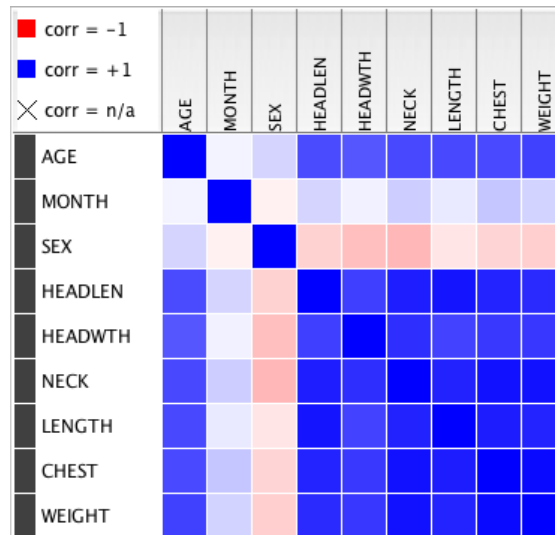
Dans le cadre d'une étude de *datamining* réalisée par des professionnels, l'expérience dans le domaine analysé joue beaucoup et permettra de faire un choix cohérent et judicieux des dimensions à conserver. Cependant, il est fréquent que les données ne soient pas dans le domaine d'expérience du chercheur ou pire, très abstraites. Dans ce cas, il faut faire appel à d'autres techniques.

#### Correlation

Il peut être très intéressant d'utiliser une matrice de corrélation afin de mettre en valeur les colonnes qui pourraient nous être utiles, et surtout celles qui nous sont inutiles.

---

1. Quelque soit la fonction de distance utilisée (Manhattan, Euclidienne, Sebestyen, etc.)



La lecture de cette matrice est très simple. Plus les couleurs (rouge ou bleu) sont foncées, plus la corrélation (*i.e.* la facilité à déduire les données d'une colonne à partir d'une autre) entre les colonnes est forte.

Il est évident que les diagonales soient de la couleur la plus foncée possible car une colonne est toujours corrélée avec elle même.

Dans cet exemple, il est facile de déduire de la matrice que le sexe et le mois de mesure n'ont finalement pas beaucoup d'impacte sur les autres paramètres.

Nous pouvons donc les ignorer dans notre *clustering*.

## PCA

La **Primary Component Analysis (PCA)** est une méthode de réduction automatique des dimensions. Elle a l'avantage de préserver *mathématiquement* le plus d'information possible. Pour cela, elle essaie de chercher la meilleure combinaison linéaire des différentes colonnes afin d'atteindre, au choix : *un nombre fixé de dimension* ou *une conservation de X% de l'information*.

Bien que pratique, elle est (tout du moins à notre niveau d'expertise en *datamining*) relativement dangereuse car les combinaisons linéaires qui résultent d'une **PCA** perdent beaucoup de leur sémantique. Aussi, il est assez compliqué d'obtenir la formule de sortie (*i.e.* l'application linéaire qui a été définie par la **PCA**) afin de l'interpréter.

### 2.1.2 Valeurs manquantes

Certains *dataset* contiennent des lignes n'exhibant pas toutes les colonnes. Il y a donc « des cases vides ». Ce problème anodin peut pourtant poser des problèmes et il faut adopter une des stratégies suivantes :

- Supprimer l'ensemble des lignes où au moins une valeur manque : C'est la solution la plus propre mais le jeu de donnée risque de se réduire drastiquement.
- Remplacer la valeur manquante : Soit par une moyenne des autres valeurs, soit par une valeur « label » (*e.g.* **missing**), soit même par une valeur aléatoire.

La meilleure solution reste toute de même de bien sélectionner ses colonnes, et de supprimer les lignes contenant une valeur manquante. Il faut cependant trouver un équilibre entre l'exactitude des valeurs utilisées et la taille du *dataset*.

### 2.1.3 Outliers

Un *outlier* est une valeur atypique du jeu de donnée risquant de biaiser l'analyse que nous pourrions en faire.

Cependant, elles peuvent aussi aider à traiter un problème dans un cas plus général. C'est pourquoi leur élimination (ou conversation) est problématique et fait appel aussi bien au bon sens, qu'à une parfaite compréhension des données ainsi qu'à l'expérience.

### Recherche « à la main »

La première méthode de recherche a l'avantage d'être simple. Elle consiste simplement à effectuer une recherche visuelle des *outliers* via un *scatter plot* pour les jeux de données de dimension inférieur à 2 et via un *parallel coordinates* sinon. L'humain étant naturellement doué fort pour détecter les valeurs extrêmes, les *outliers* devraient être assez simple à exclure.

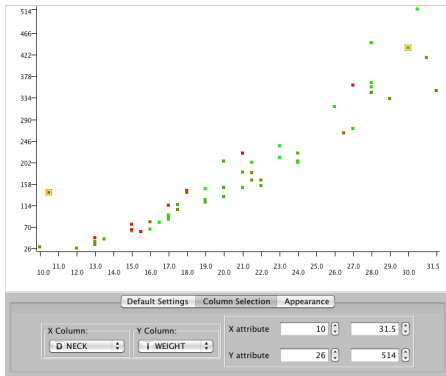


FIGURE 1 – Un exemple de diagramme « Scatter plot »

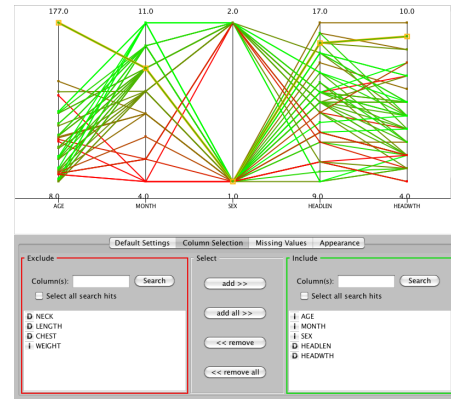


FIGURE 2 – Un exemple de diagramme « ParallelCoordinates »

### Clustering hiérarchique

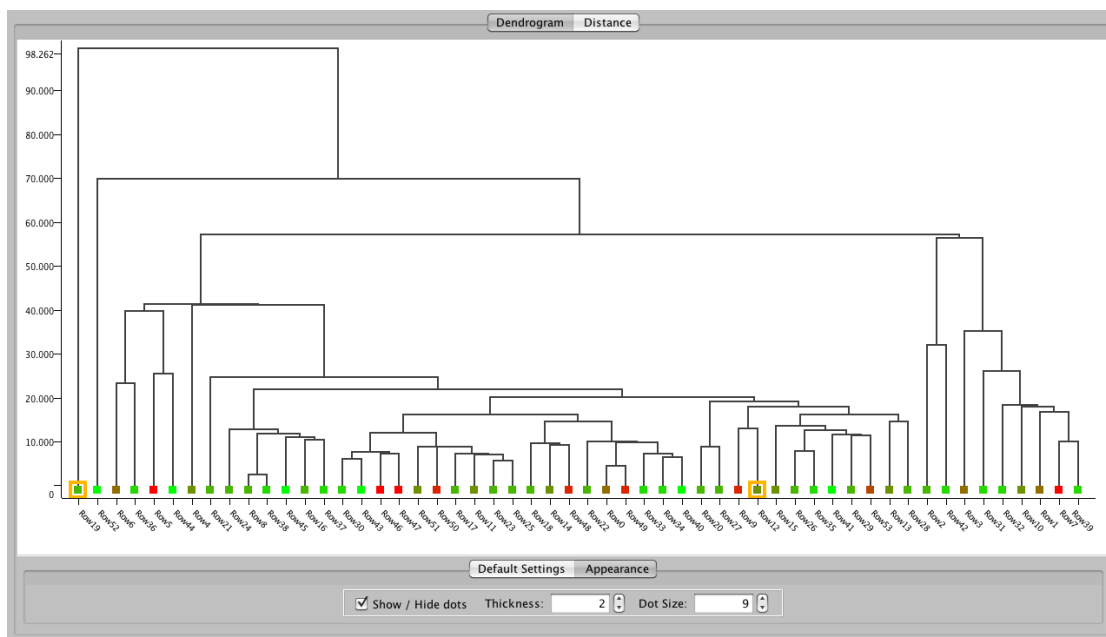
Une méthode efficace et qui me semble un peu plus rigoureuse consiste à utiliser un *clustering* hiérarchique afin d'en extraire les *outliers*.

La méthode est très simple :

*Chaque point ou cluster est progressivement absorbé par le cluster le plus proche.*

En effet, le dernier point ou petit groupe de point qui se fait « coller » en dernier représente un *outlier*.

Sur la figure 4, il est évident que le point le plus à gauche est un *outlier*. En effet, il s'est collé en dernier, et ce, à une distance presque égale au rayon du cluster contenant tous les autres points.

FIGURE 3 – Exemple de dendrogramme extrait d'un *clustering* hiérarchique

### Box plot

C'est encore un moyen très rapide de démasquer les outliers car ils font figurer sur les même diagramme, la médian, les quartiles et les décile. Nous avons donc toutes les données statistiques nécessaire afin d'évaluer l'appartenance d'une point à notre futur jeu de données.

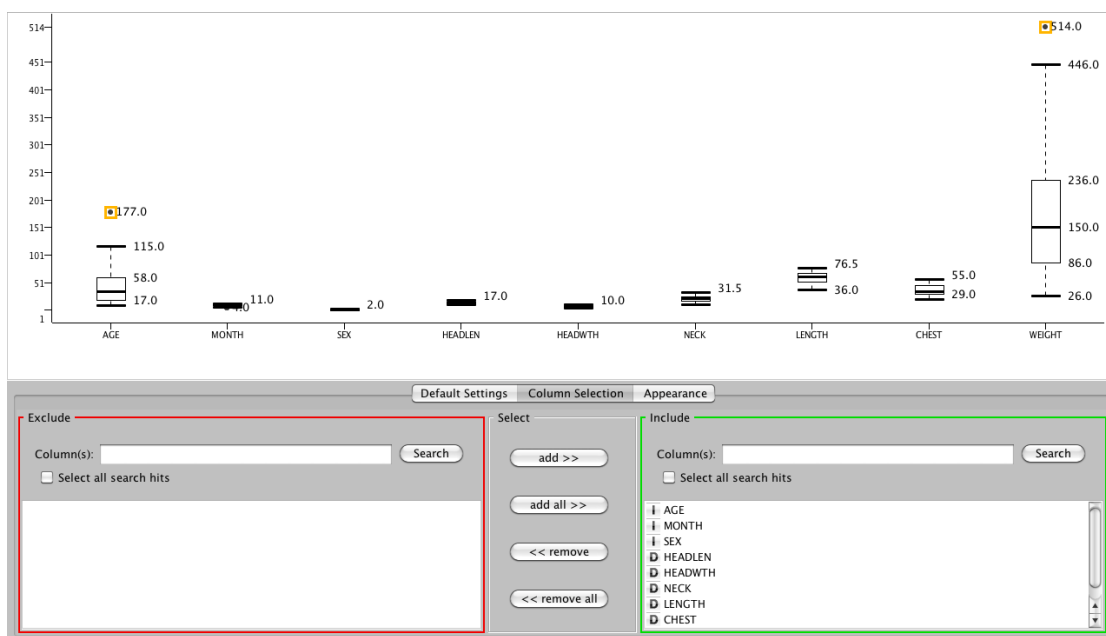


FIGURE 4 – Exemple de boîte à moustache

### Méthodes statistiques

Une des dernières méthodes (que nous n'utilisons pas sur Knime) est une approche statistique des données. Il est prouvé que de nombreux phénomènes suivent une loi

normale. De plus, nous connaissons très bien cette loi et des mathématiciens ont pu modéliser l'atypique afin de fournir des méthode automatique d'exclusion des *outliers*. Nous ne rentrerons pas plus dans le détail mais parmi ces méthodes, nous pouvons citer le critère de PEIRCE, de CHAUVENET ou encore le test de GRUBB.

## 2.2 Détermination du nombre de clusters

Intro ou le  
nombre est  
connu

### *Clustering* hiérarchique

SINGLE :  
Bras

## 2.3 Clustering

COMPLETE  
et AV-  
ERAGE :  
Paquets,  
et mieux  
adapté à  
*K-Means*

### Normalisation

Distance :  
Utilisation  
des gaps  
plus que de  
la dérivé.

### 2.3.1 Hierarchical Clustering

### 2.3.2 K-Mean

### 2.3.3 Fuzzy C-Mean

Faire la  
liste et les  
spécificité  
des  
différentes  
méthodes de  
*clustering*

Elle est  
parfois inu-  
tile (dans  
le cas de  
valeur en  
pourcent-  
ages e.g.

### 3 Secteurs d'activité des différents pays

Pour cette première étude de cas, nous avons choisi d'utiliser les données afin de caractériser l'activité d'un pays en fonction de :

- La taille du pays
- La population
- Le **Produit Intérieur Brute (PIB)**
- Le **RNB**
- Les exports industriels

Notre démarche dans cette analyse est simple : Après avoir établi un *clustering* sur le pourcentage du **PIB** dû à l'agriculture, l'industrie et les services, nous avons créé un arbre de décision prenant pour critère de sélection les données citées ci-dessus.

#### 3.1 Sélection des données

Après avoir sélectionné les colonnes qui nous intéressent ( *i.e.* **Agriculture, value added (% of GDP), Industry, value added (% of GDP)** et **Service, value added (% of GDP)**), nous obtenons la matrice suivante :

Le diagramme de est :

Enfin le diagramme *parallel coordinates* :

#### 3.2 Clustering

##### 3.2.1 Clustering hiérarchique

Le choix du type de *clustering* hiérarchique est, *a priori* délicat. Voici les résultats obtenus avec un *clustering* hiérarchique :

**SINGLE**

**COMPLETE**

##### 3.2.2 Clustering K-Means

**Vérification de la stabilité du *clustering***

#### 3.3 Interprétation

##### 3.3.1 Arbre de décision

Ajouter le  
nom de la  
colonne cor-  
respondante

Ajouter  
l'environ-  
nement  
subimages  
pour une  
meilleur  
intégration

boite à  
chaussure ??

Guinée :  
95% de son  
PIB est  
industriel

Inclure le  
tableau de  
vérification  
de l'entropie

## 4 Création d'un module de vérification du K-Mean



## Glossaire

**DATASET**

Jeu de donnée [Page(s) [3](#)]

## Liste des accronymes

**PCA**

Primary Component Analysis [Page(s) [3](#)]

**PIB**

Produit Intérieur Brute [Page(s) [7](#)]

**RNB**

Revenue National Brute [Page(s) [1](#), [7](#)]

**VIH**

Virus d'immunodéficience Humaine [Page(s) [1](#)]

**Fin**

