



Rapport

Project 2: Unsupervised Analysis of the Human Development Report

Auteur:

Siéwé Kouéta ANICET : 00364245

Professeur:

Benoît FRÉNEY



UNIVERSITÉ LIBRE DE BRUXELLES

ULB

November 30, 2019

1 task 1

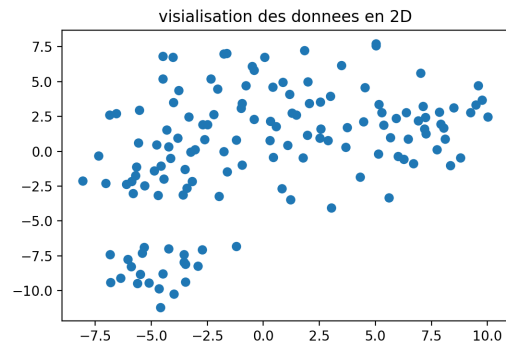


Figure 1: repartition des données (*doc/Graphes/visialisation_des_donne_en_2D.png*)

la figure 1 nous montre une visualisation des données du Dataset `/Data/hdr_data.dat`, via un diagramme de dispersion bidimensionnel. nous avons un jeu de données constitué de 138 points (instances).

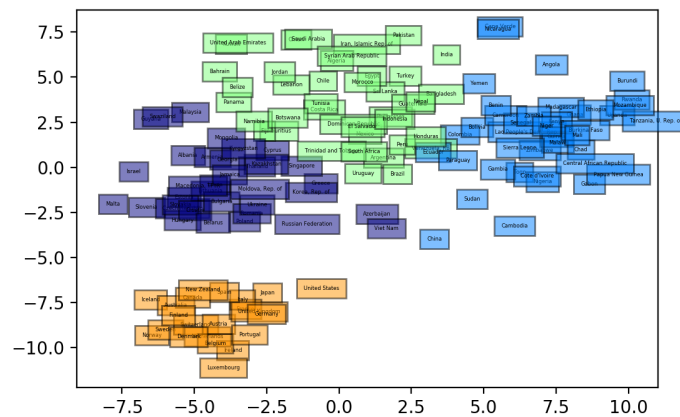


Figure 2: K-means, avec $k=4$ (*doc/Graphes/visialisation_des_donne_pour_k_fixer_a_4.png*)

Après avoir essayé et visualiser plusieurs nombre de cluster, on peut estimer que la meilleur valeur de k est **quatre(4)** car c'est le regroupement (voir figure 2) ou : d'une part les couleurs entre les différents cluster se chevauche le moins et d'autre part ou les cluster ont été repartir en fonction des disposition des point dans l'espace 2D (voir figure 1). Voir également en **Annexe_1** une methode pour determiner la meilleur valeur optimal de cluster.

les valeur des centroïdes pour ces 4 cluster sont :

- | **Latvia** -> centroïdes associer au Cluter N*: 1 color=(navy)
- | **Niger** -> centroïdes associer au Cluter N*: 2 color=(dodgerblue)
- | **Tunisia** -> centroïdes associer au Cluter N*: 3 color=(lightgreen)
- | **Austria** -> centroïdes associer au Cluter N*: 4 color=(darkorange)

2 task 2

NOTE 1 : les réponses sur cette partie se base sur les données stockés dans les fichier *.txt* se trouvant dans */doc/Graphes/ *.txt*. Voir **Annex_2** pour savoir comment lire et interpreter les informations du Tableau_2 de chaque fichier *.txt*

NOTE 2: lors de l'analyse des différent cluster , Les caractéristiques énumérer sont uniquement celle que j'ai jugé pertinente de liste en fonction des pays de chaque cluster.

pas assez de clusters (k=3)

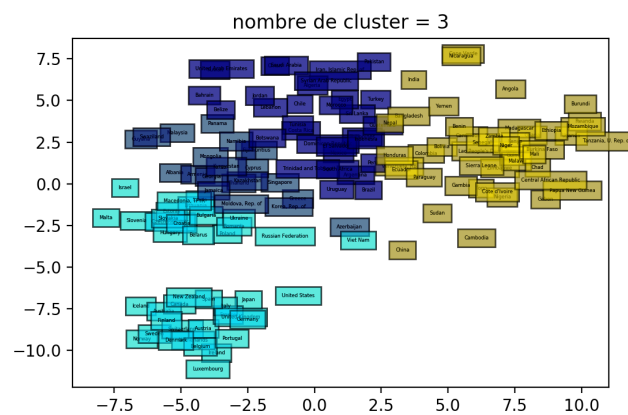


Figure 3: 3_Means (*doc/Graphes/3_Means.png*)

Voir Toute les informations sur ces cluster dans **doc/Graphes/data_3_Means.txt**. Dans chacun des cluster, les noms des pays les plus proche des centres sont les suivant :

- | **Botswana** -> centroïdes associer au Cluster N*: 1 color=(navy)
- | **New Zealand** -> centroïdes associer au Cluster N*: 2 color=(deepskyblue)
- | **Senegal** -> centroïdes associer au Cluster N*: 3 color=(gold)

Le regroupement k=3 a été choisi comme celui qui contient pas assez de cluster, car on peut facilement voir la transition entre les cluster. On voit clairement sur l'image (voir figure 3) un chevauchement (une sorte de transition/passage) entre le cluster N*1 et le N*2. On voit donc très distinctement la portion où notre Algorithme a du mal à assigner certains pays à un des cluster. on constate que :

- **le Cluster N*1** (color="navy") regroupe majoritairement les pays à faible revenu. ces pays sont caractérisés par un faible indice de consommation, un taux moyen d'exportations de biens et services, un PIB très faible, un faible taux de pénétration internet, un fort pourcentage de leur population réfugié à l'extérieur, une faible population, ainsi qu'une faible aide publique au développement.
- **le Cluster N*2** (color="deepskyblue") regroupe majoritairement les pays d'Europe Centrale. ces pays sont caractérisés par un faible taux de croissance annuel de leur population, un faible taux de consommation, un PIB élevé, une croissance moyenne du PIB par habitant, un taux élevé de bébé en sous-poids, un taux élevé de réfugié et une faible population.
- **le Cluster N*3** (color="gold") regroupe majoritairement les pays sous développer.

assez de clusters (k=5)

Voir Toute les informations sur ces cluster dans **doc/Graphes/data_5_Means.txt**. Dans chacun des cluster, les noms des pays les plus proches des centres sont les suivants :

- | **Chile** -> centroïdes associer au Cluster N*: 1 color=(navy)

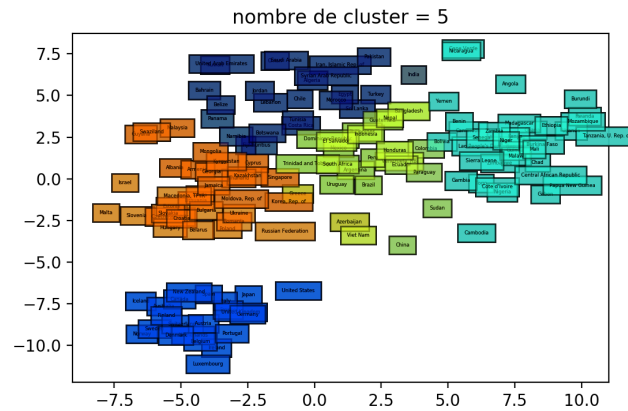


Figure 4: 5_Means (*doc/Graphes/5_Means.png*)

| **Austria** -> centroïdes associer au Cluster N*: 2 color=(dodgerblue)

| **Niger** -> centroïdes associer au Cluster N*: 3 color=(turquoise)

| **Peru** -> centroïdes associer au Cluster N*: 4 color=(greenyellow)

| **Latvia** -> centroïdes associer au Cluster N*: 5 color=(orangered)

Le regroupement $k=5$ a été choisi comme celui qui contient assez de cluster parce qu'il est celui qui fait une distinction claire des couleurs des pays appartenant aux différents cluster (voir figure 4). dans ce regroupement, les couleurs (ou cluster) de chaque pays ne se chevauche pas et chaque bloc est facilement identifiable. on constate que :

-le **Cluster N*1** (color="navy") regroupe majoritairement les pays à faible revenu. ces pays sont caractériser par une faible croissance économique et un PIB très bas. une faible population et l'aide publique au développement est très bas, un faible taux Investissements étrangers et une faible consommation énergie.

-le **Cluster N*2** (color="dodgerblue") regroupe majoritairement les pays développer, caractériser par un taux de consommation élever , une faible croissance de leur population, un fort taux d'investissement étranger et de l'aide publique au développement.

-le **Cluster N*3** (color="turquoise") regroupe majoritairement les pays d'Afrique . ces pays sont caractériser par un PIB et un accès aux soins médicaux très faible, une faible pénétration d'internet et un taux d'investissement étranger relativement bas.

-le **Cluster N*4** (color="greenyellow") regroupe majoritairement les pays a forte production. ces pays ont un faible PIB, un faible taux d'investissement étranger.

-le **Cluster N*5** (color="orangered") regroupe majoritairement les pays d'Europe de l'Est. ces pays sont caractériser par un faible PIB, un faible taux de consommation, un faible taux d'activité économique des femmes, un faible taux de femmes scolarisé, un faible taux d'utilisation d'internet, un faible taux d'investissement étranger et une faible population

trop de clusters (k=7)

Voir Toute les informations sur ces cluster dans **doc/Graphes/data_7_Means.txt**.

Dans chacun des cluster, les noms des pays les plus proche des centres sont les suivant :

| **Austria** -> centroïdes associer au Cluster N*: 1 color=(navy)

| **Argentina** -> centroïdes associer au Cluster N*: 2 color=(blue)

| **Croatia** -> centroïdes associer au Cluster N*: 3 color=(deepskyblue)

| **Guinea** -> centroïdes associer au Cluster N*: 4 color=(turquoise)

| **Jordan** -> centroïdes associer au Cluster N*: 5 color=(greenyellow)

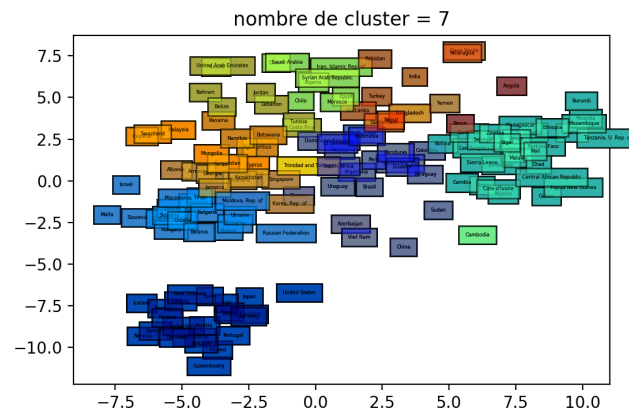


Figure 5: 7_Means (*doc/Graphes/7_Means.png*)

Kyrgyzstan -> centroïdes associer au Cluster N*: 6 color=(orange)

India -> centroïdes associer au Cluster N*: 7 color=(orangered)

Le regroupement $k=7$ a été choisi comme celui qui contient trop de cluster car à partir de ce nombre de cluster, on distingue de moins en moins bien les différents cluster sur notre figure 5. beaucoup de cluster s'entrecroise. on constate que :

- le **Cluster N*1** (color="navy") regroupe majoritairement les pays développés. ces pays sont caractérisés par un faible taux annuel de croissance de leur population, une faible variation de l'indice de prix, une forte émission de CO₂, une forte croissance du PIB, un fort taux d'Investissements étrangers, un faible taux de réfugiés par pays d'origine, une faible population et enfin par une faible réception d'aide publique au développement.

- le **Cluster N*2** (color="blue") regroupe majoritairement les pays en voie de développement. des pays sont caractérisés par un faible PIB, un faible taux d'utilisation d'internet, un faible taux d'Investissements étrangers, et une forte Aide publique au développement.

- le **Cluster N*3** (color="deepskyblue") regroupe majoritairement les pays d'Europe de l'Est. ces pays sont caractérisés par un faible PIB, un taux moyen d'activité économique des femmes, un faible taux d'investissement étrangers, et une faible Aide publique au développement.

- le **Cluster N*4** (color="turquoise") regroupe majoritairement les pays d'Afrique. ces pays sont caractérisés par un faible indice de consommation, une faible émission de CO₂, une faible Consommation d'électricité par habitant et un PIB également faible, un taux d'accès à internet et au système de santé très faible, un faible taux d'Investissements étrangers et une faible population militaire.

- le **Cluster N*5** (color="greenyellow") regroupe majoritairement les pays Musulmans. ces pays sont caractérisés par un faible indice de consommation et PIB, un faible accès à internet et au système de santé, un faible taux d'Investissements étrangers et une faible population.

- le **Cluster N*6** (color="orange") regroupe majoritairement les pays moins développés. ces pays sont caractérisés par un faible indice de consommation et PIB, un faible accès à internet, une faible population, une faible consommation de carburant et une faible Aide publique au développement.

- le **Cluster N*7** (color="orangered") regroupe majoritairement les pays sous-développés. ces pays sont caractérisés par une faible émission de CO₂, une faible consommation d'électricité, un faible accès à internet et au système de santé, un faible taux d'Investissements étrangers, une faible population et une faible Aide publique au développement.

Annexe

Annexe 1

on peut aussi utiliser la methode du **Coude** pour trouver la meilleur valeur optimal de cluster. on constate dans le graphique de la figure 6, que le *coude* est situé à $k = 4$, ce qui montre que **4 cluster** est effectivement un bon choix pour cet ensemble de données.

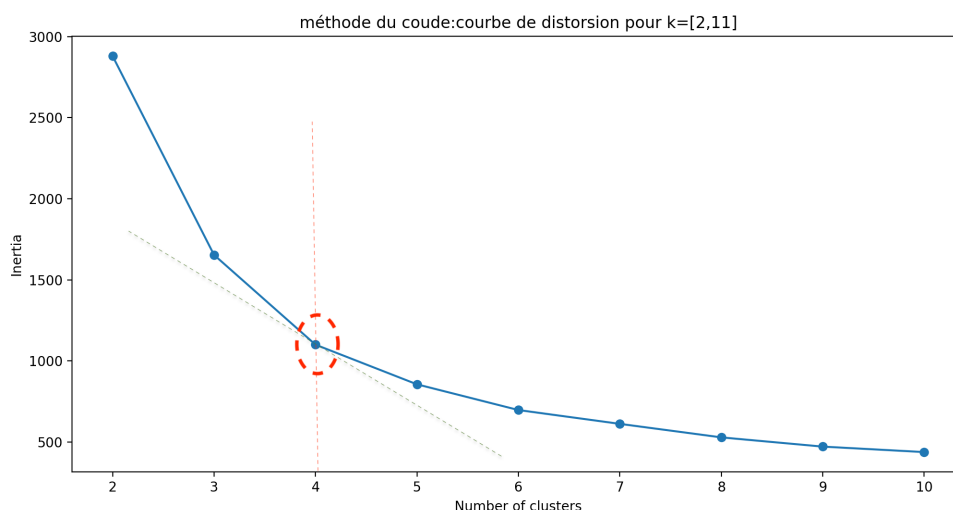


Figure 6: courbe de distorsion (*doc/Graphes/methode_du_coude_courbe_de_distorsion.png*)

Annexe 2

indicator_names	Nombre de pays/cluster ayant des valeur de caracteristique proche		
	Cluster_ID		
	1	2	3
Pop growth	13 [-1.8; 4.62]	7 [-0.87; 1.66]	20 [-1.29; -0.36]
Pop growth 2004	10 [-2.28; 1.55]	19 [-1.07; 2.56]	19 [-1.07; 0.04]
Price index	43 [-0.42; 10.11]	55 [-0.36; 1.9]	22 [-0.42; -0.33]

A : est l'identifiant du cluster. ce cluster peut être identifier par sa couleur associer dans le tableau_1, sur sa figure associer.

B: Nombre de pays du cluster "A" qui ont approximativement la même valeur pour une caractéristique choisie (*indicator_names*). cette approximation dépend du paramètre **precision=0.5** qui permet de définir une plage de valeur comprise entre [Default_value-precision ; Default_value+precision].

C: la plus petite valeur observer pour une caractéristique donnée, dans le cluster "A"

D: la plus grande valeur observer pour une caractéristique donnée, dans le cluster "A"