

Analyse des principaux pays du monde en matière de développement. Aide à la décision pour les organisations de solidarité internationale.

FEVER Quentin - BERNARD Ambroise
FISE 2023 UV ODATA
IMT NORD EUROPE



Objectifs du projet

Le but de ce projet est d'analyser au travers plusieurs indices socio-économiques, une liste de 167 pays afin de déterminer lesquels pourront bénéficier d'une aide humanitaire. Il s'agit donc ici de faire une classification non supervisée des données observées par le biais de cinq méthodes de clustering :

- La méthode des K-Means
- La classification ascendante hiérarchique (CAH)
- Le modèle de mélange des Gaussiennes
- DBSCAN
- Le partitionnement spectral

Pour cela, nous allons faire une étude comparative de ces méthodes avant de les appliquer à notre jeu de données. Ensuite, grâce à notre partition des données, nous choisirons une liste d'environ dix pays auxquels nous accorderons l'aide en question.

Pour commencer, nous allons donc étudier ces cinq méthodes de clustering sur différents jeux de données "témoins" afin d'en déterminer les performances selon le jeu de données et de trouver la plus adaptée à notre problème.

Les différentes méthodes utilisées non vues en cours.

DBSCAN

L'algorithme DBSCAN utilise deux paramètres : la distance epsilon et le nombre minimum de points devant se trouver dans un rayon epsilon pour que ces points soient considérés comme un cluster.

Si le nombre minimum de points est atteint dans le rayon epsilon d'un point, celui-ci est choisi comme centre du nouveau cluster.

Avantages : il identifie les valeurs aberrantes et les classe comme bruit, il est capable de trouver des clusters de taille et forme arbitraire.

Inconvénients : il est moins performant quand la dimension des données augmente

Spectral Clustering

Le clustering spectral partitionne les données en utilisant les vecteurs propres de la matrice de similarité des données.

Avantages : cette méthode forme des clusters non convexes et de forme quelconque ce qui rend plus pertinent le partitionnement selon la forme du nuage de données.

Inconvénient : sensible au bruit

I) Etude préalable : Comparaison des méthodes de clustering sur des données simulées

Nous avons donc étudié les cinq méthodes sur trois jeux de données : *jain*, *aggregation* et *pathbased*.

Paramètres d'initialisation :

Pour ces trois jeux, nous avons choisi de partitionner les données respectivement en 2, 7 et 3 clusters car c'est de cette manière qu'ils étaient classés.

K-means : les paramètres ont été les mêmes pour les trois jeux, à savoir “*init = ‘k-means++’*” car cela permet de faire converger l'algorithme plus rapidement et “*n_init = 10*” car c'est la valeur par défaut et que l'augmentation de cette valeur ne rendait pas de meilleurs résultats.

Cette méthode n'a pas été très efficace pour le partitionnement. En effet, nous n'avons jamais réussi à obtenir le partitionnement réel, quel que soit le jeu de données. En particulier pour *jain*, à peine un tiers des points a été affecté correctement.

CAH : Après plusieurs tests nous avons fixé le paramètre *t* respectivement à 18, 8 et 20 (*t* est inversement proportionnel au nombre de clusters)
Ensuite, pour les trois jeux, nous avons la distance de Ward comme méthode de linkage car c'est théoriquement la plus performante étant donné qu'elle prend en compte le nombre de points dans le cluster dans sa formule.

Cette méthode a été plutôt efficace étant donné qu'elle a donné les clusters réels pour deux des trois jeux. Seul *pathbased* n'a pas été partitionné convenablement : un tiers des points n'a pas été affecté au bon cluster.

Gaussian Mixture : Après plusieurs tests de performance, nous avons choisi de fixer le paramètre *covariance_type* à “full” étant donné qu'il nous donnait de meilleurs résultats. De la même manière que précédemment, nous avons fixé *n_init* à 10.

Cette méthode fait partie des plus efficaces étant donné qu'elle a fourni le bon partitionnement pour chacun des trois jeux de données.

DBSCAN : En prenant les paramètres par défaut, nous avons trouvé les clusters réels pour deux des trois jeux de données. Pour *aggregation*, nous avons dû baisser la valeur d'*epsilon* à 0.3 pour obtenir les mêmes résultats. Pour chaque jeu, nous avons fixé la variable *algorithm* à "auto" car cela nous donnait des résultats très performants.

Cette méthode fait aussi partie des plus performantes. Elle a fourni elle aussi le bon partitionnement pour chacun des trois jeux de données.

Spectral clustering : Pour chacun des jeux, nous avons pris comme paramètres : *n_init = 10*, *affinity = 'rbf'* et *n_neighbors=10*. En effet, à force de tests, nous avons constaté que ces paramètres nous donnaient les meilleurs résultats.

Cette méthode a été plutôt performante. Elle nous a donné le bon partitionnement pour deux jeux. Pour *pathbased*, environ 60% des points ont été attribués à son cluster réel.

	K-means	CAH	Gaus. Mix.	DBSCAN	Spec. clust.
jain	0.324	1	1	1	1
aggregation	0.774	1	1	1	1
pathbased	0.462	0.677	1	1	0.613

Figure 1 : Tableau des valeurs de l'indice de RAND en fonction du jeu de donnée et de la méthode de clustering utilisée

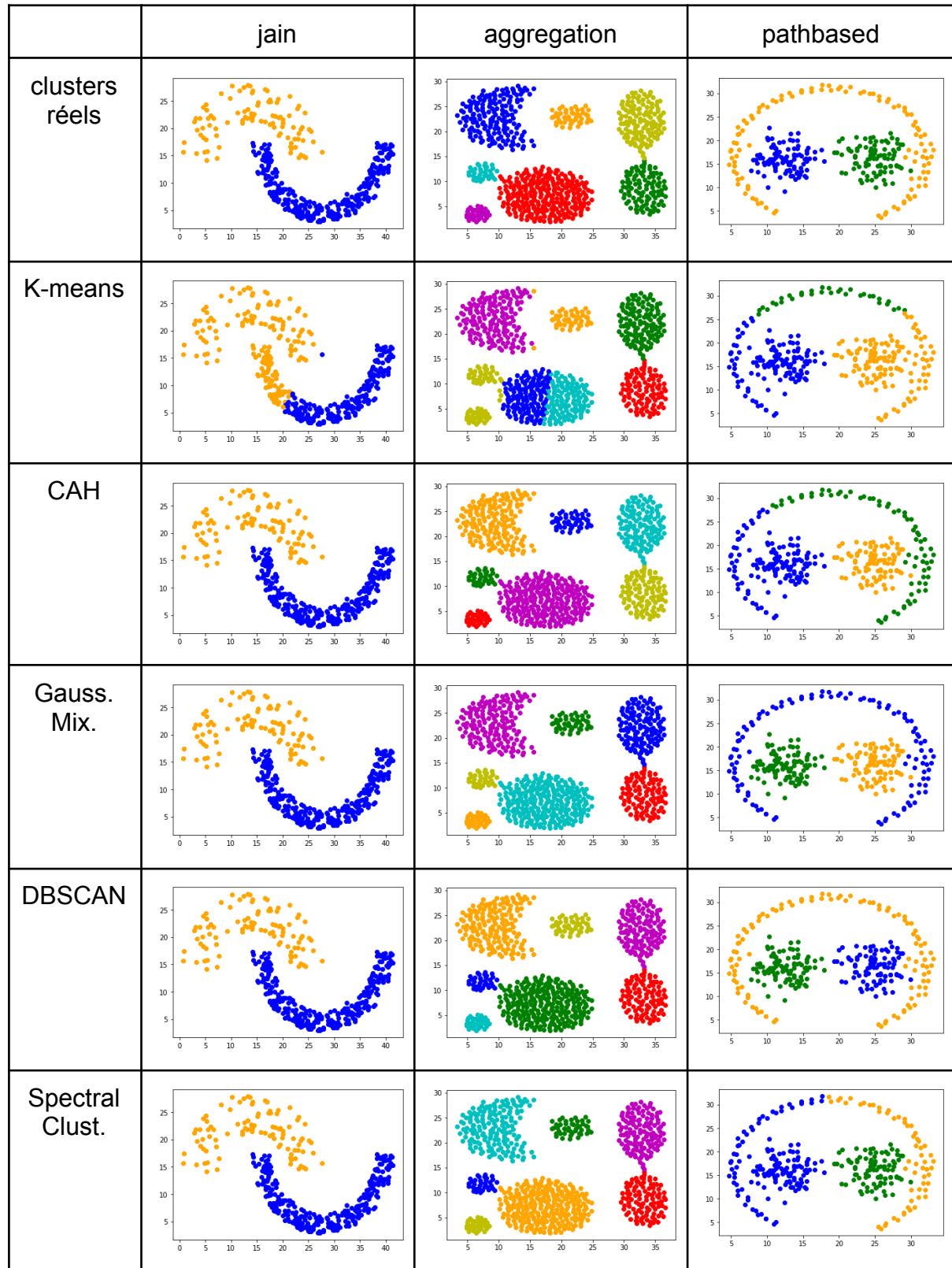


Figure 2 : Représentation graphique des clusters en fonction du jeu de données et de la méthode de clustering utilisée

Pour conclure, les deux méthodes les plus performantes ont été le mélange des Gaussiennes et le DBSCAN. En effet, ces deux méthodes nous ont donné dans chacun des cas le bon partitionnement des données. Elles sont adaptées quelle que soit la forme des clusters. En particulier, la méthode du mélange des gaussiennes est intéressante étant donné qu'elle n'a pas nécessité l'adaptation d'une variable au problème posé (contrairement à *epsilon* dans le DBSCAN par exemple).

Ensuite, les méthodes du CAH et du clustering spectral ont été relativement intéressantes dans le sens où elles ont trouvé le bon partitionnement deux fois sur trois. En revanche, pour le jeu *pathbased*, toutes deux ont été relativement inefficaces car elles n'affectent qu'environ deux tiers des points au bon cluster.

Enfin, la méthode des K-means, a été la moins performante des cinq méthodes. Elle n'a jamais trouvé le bon partitionnement des données. Elle n'a affecté pour *jain*, *aggregation* et *pathbased* respectivement que 32, 77 et 46% des points au bon cluster. La méthode des K-Means n'est donc pas adaptée lorsque les clusters ne sont pas distincts, elle reste limitée.

II) Partitionnement des principaux pays du monde en fonction de leur développement

A partir des 5 méthodes décrites précédemment, nous allons les mettre en œuvre pour classer les principaux pays du monde en fonction de leur développement à partir de données socio-économiques et de santé sur chaque pays.

L'intérêt de cette classification va être de guider les choix d'affectation de l'aide humanitaire internationale en identifiant les pays dans lesquels il est nécessaire d'intervenir en priorité, nous préciserons également les différents domaines d'activités sur lesquels intervenir en priorité.

Examinons d'abord le jeu de données que nous allons exploiter :

Caractéristiques :

- Il est constitué de neuf variables
- (country,child_mortality,exports,health,imports,income,inflation,life_expectatio n,total_fertility,GDP)
- Appart le nom des différents pays, ce sont toutes des données numériques.
- Présence de données manquantes et aberrantes.

1	country,child_mortality,exports,health,imports,income,inflation,life_expectation,total_fertility,GDP
2	Afghanistan,90,2,10,7.58,44,9,1610,9.44,56,2,5,82,553
3	Albania,16,6,28,6.55,48,6,9930,4,49,76,3,1,65,4090
4	Algeria,27,3,38,4,4,17,31,4,12900,16,1,76,5,2,89,4460
5	Angola,119,62,3,2,85,42,9,5900,22,4,60,1,6,16,3530
6	Antigua and Barbuda,10,3,45,5,6,03,58,9,19100,1,44,76,8,2,13,12200
7	Argentina,14,5,18,9,8,1,16,18700,20,9,75,8,2,37,10300
8	Armenia,18,1,20,8,4,4,45,3,6700,7,77,73,3,1,69,3220
9	Australia,4,8,19,8,8,73,20,9,41400,1,16,82,1,93,1000000
0	Austria,4,3,51,3,11,47,8,43200,0,873,80,5,1,44,46900
1	Azerbaijan,39,2,54,3,5,88,20,7,16000,13,8,69,1,1,92,5840
2	Bahamas,13,8,35,7,89,43,7,22900,-0,393,73,8,1,86,28000
3	Bahrain,8,6,69,5,4,97,50,9,41100,7,44,76,2,16,20700
4	Bangladesh,49,4,16,3,52,21,8,2440,7,14,0,2,33,758
5	Barbados,14,2,39,5,7,97,48,7,15300,0,321,76,7,1,78,16000
6	Belarus,5,5,51,4,5,61,64,5,16200,15,1,70,4,1,49,6030
7	Belgium,4,5,76,4,10,7,74,7,41100,1,88,80,1,86,44400
8	Belize,18,8,58,2,5,2,57,5,7880,1,14,71,4,2,71,4340
9	Benin,111,23,8,4,1,37,2,1820,0,885,61,8,5,36,758
0	Bhutan,42,7,42,5,5,2,70,7,6420,5,99,72,1,2,38,2180
1	Bolivia,46,6,41,2,4,84,34,3,5410,8,78,71,6,3,2,1980
2	Bosnia and Herzegovina,6,9,29,7,11,1,51,3,9720,1,4,76,8,1,31,4610
3	Botswana,52,5,43,6,8,3,51,3,13300,8,92,57,1,2,88,6350
4	Brazil,19,8,8,10,7,9,01,11,8,14500,8,41,74,2,1,8,11200

Figure 3 : jeu de données initial

1) Etudes des corrélations

Après avoir préparé et nettoyé les données, une première comparaison des données peut être réalisée en comparant les différentes corrélations des variables.

Grâce à la matrice des corrélations, on observe que les variables les plus corrélées positivement sont le revenu (income) et le PIB (GDP), assez logique car ce sont toutes les deux des variables faisant état de la santé économique d'un pays. D'autre part les données les plus corrélées négativement sont la mortalité infantile et l'espérance de vie car plus l'espérance de vie augmente plus la mortalité infantile diminue. Enfin les variables les moins corrélées sont l'espérance de vie avec les importations et la santé avec les importations.

2) l'analyse en composantes principales

Nous allons maintenant réaliser une analyse en composantes principales (ACP) afin de mieux comprendre les corrélations entre les variables, regrouper les pays similaires et de les visualiser dans un plan (réduction de dimension).

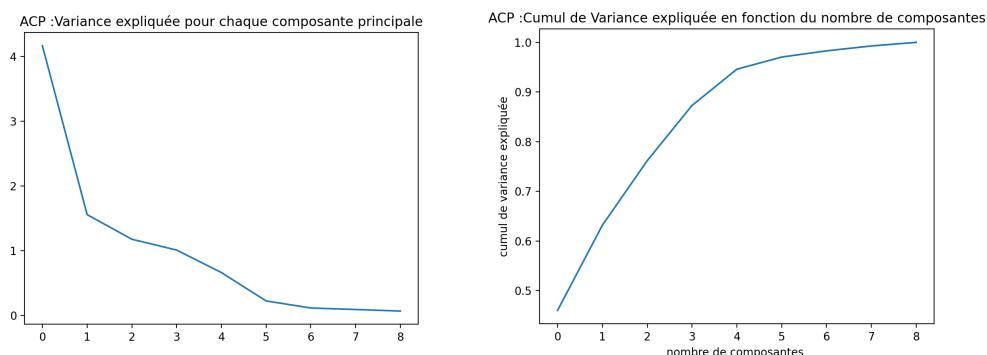


Figure 4 et 5 : pour l'aide au à la décision du nombre de composantes principales à conserver

La première figure nous permet d'appliquer la méthode du coude dans le choix du nombre de composantes à conserver. Ici on pourrait conserver jusqu'à 5 composantes principales

La seconde figure nous montre que conserver 4 composantes suffit pour conserver plus de 90% de la variance expliquée.

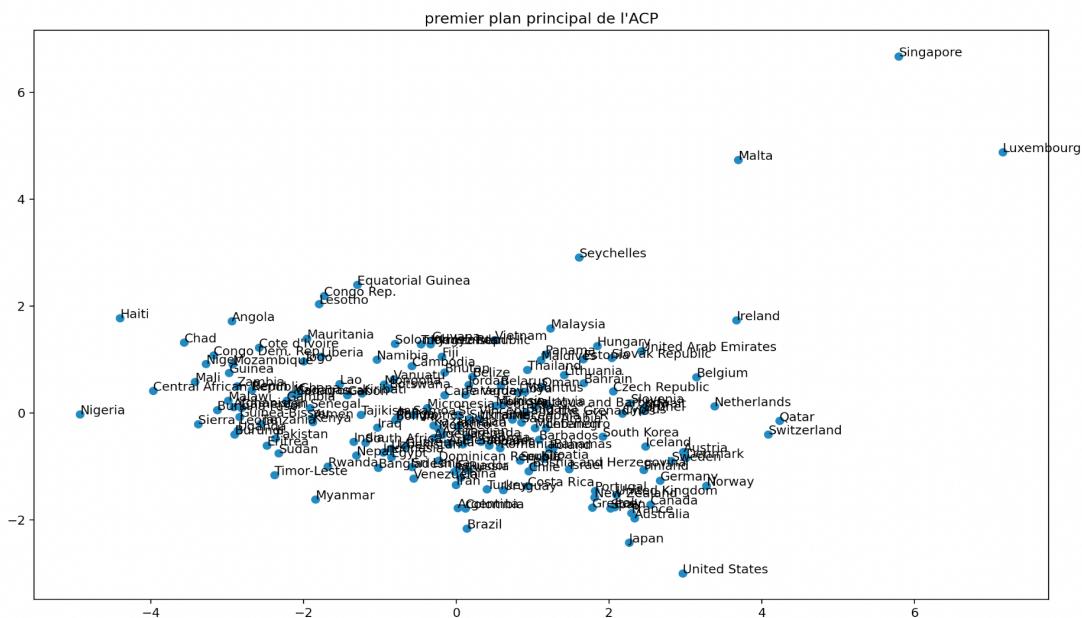
Enfin, l'étude des valeurs propres nous apporte également une aide à la décision.

Liste des valeurs propres de l'ACP :

[4.16539287 1.55531965 1.17524955 1.00790753 0.66079931 0.22192275
0.11263361 0.08941066 0.06558094]

D'après la règle de Kaiser, on remarque qu'il y a 4 valeurs propres supérieures à 1, on peut donc conserver 4 composantes principales selon cette règle.

Bilan : On décide de conserver 4 composantes principales et nous pouvons observer les individus dans les deux plans principaux



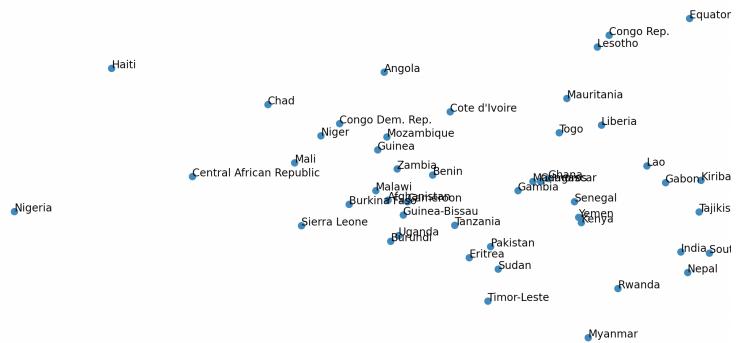


Figure 6 et 7 : Représentation des pays dans le premier plan principal

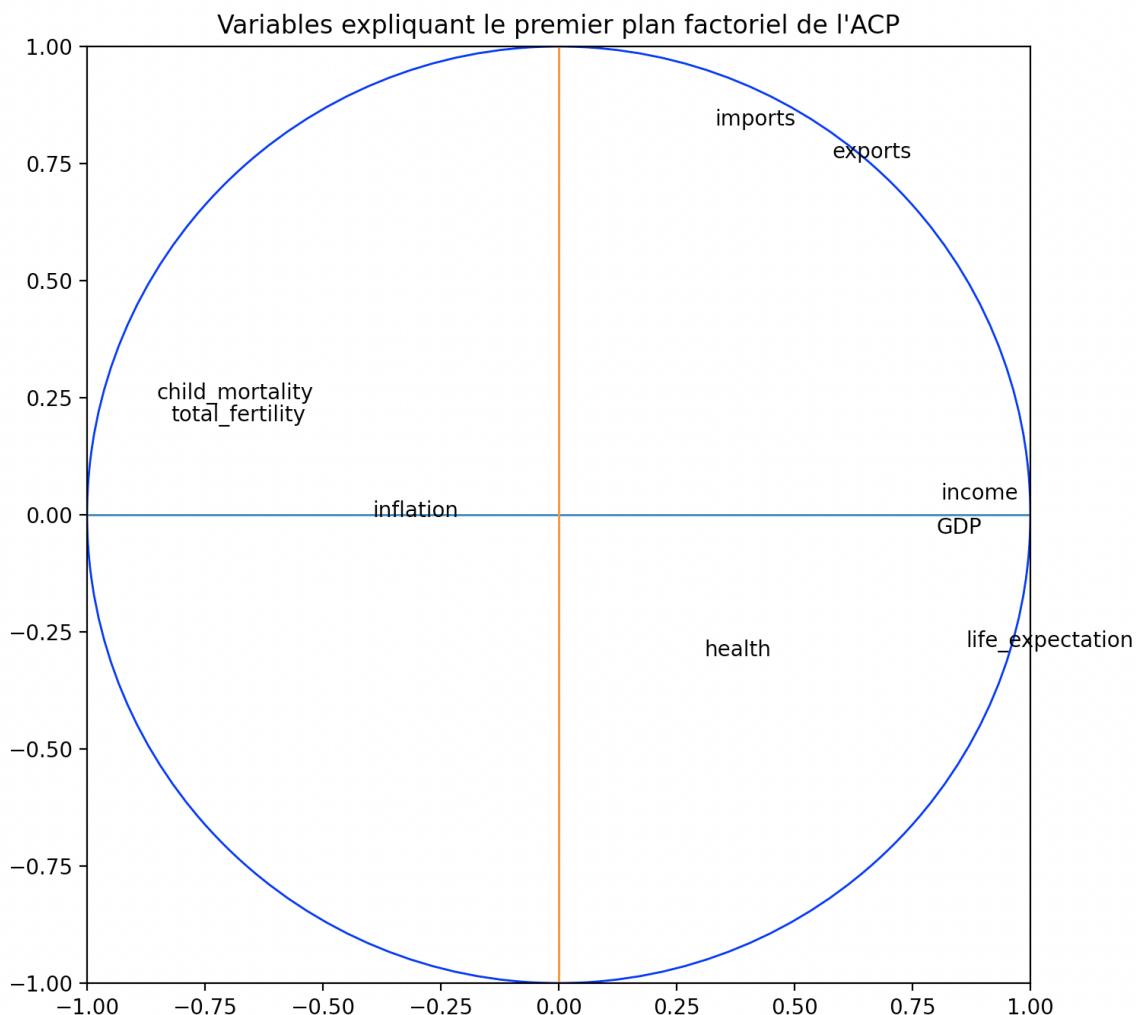


Figure 8 : Cercle des corrélations du premier plan factoriel

Explication de l'ACP pour le premier plan factoriel :

Axe 1:

Sur le cercle des corrélations, on remarque que l'axe 1 est déterminé par le revenu, le PIB, l'espérance de vie, soit les variables financières pour la droite de l'axe. la mortalité infantile et la fertilité déterminent la gauche de l'axe

On remarque que beaucoup de pays d'Afrique avec également Haïti et Sierra Leone souffrent le plus de la mortalité infantile et qui ont les développements économiques les plus faibles.

Ce sont également des pays qui traditionnellement font plus d'enfants que les pays "développés du nord".

A l'inverse, les pays les plus à droite de l'axe sont ceux avec la meilleure santé économique, il n'est pas étonnant d'y retrouver la Suisse, le Qatar, le Luxembourg ou Singapour.

Axe 2 :

l'axe 2 quant à lui est déterminé par les importations/exportations en haut selon le cercle des corrélations et aucune véritable variable en bas.

En haut de l'axe 2, on retrouve les pays qui dépendent le plus de l'importation (et dans une moindre mesure l'exportation),

c'est pour cela que l'on retrouve les îles (Seychelles, Irlande) et les pays qui n'ont pas d'accès directs à la mer comme le Luxembourg.

Explication de l'ACP pour le second plan factoriel :

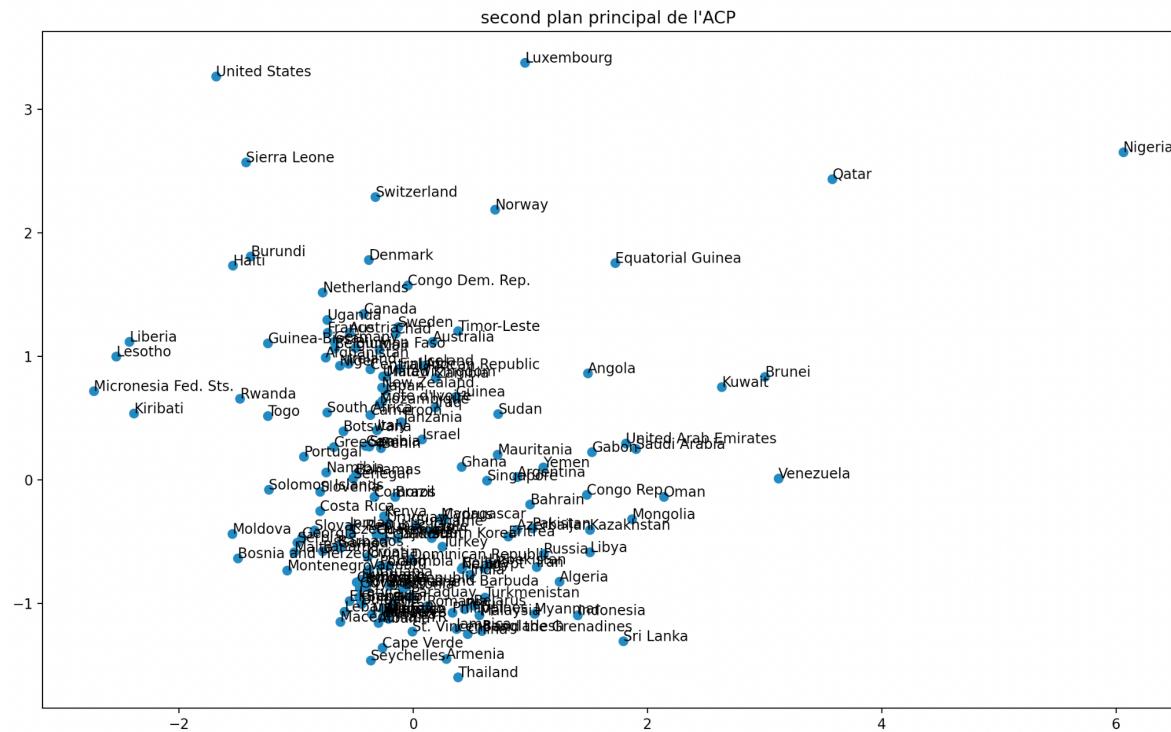


Figure 9 : Second plan principal de l'ACP

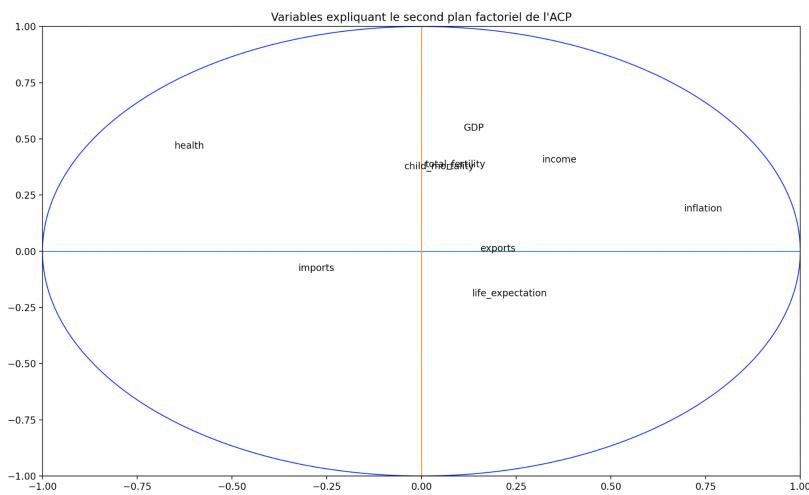


Figure 10 : Cercle des corrélations du second plan factoriel de l'ACP

Les axes 3 et 4 sont plus difficiles à interpréter que les deux précédents car il est plus difficile de déterminer les variables qui contribuent le plus aux axes. On peut noter pour l'axe 4 qui serait expliqué par le PIB, que les Etats-Unis(première puissance économique mondiale) sont placés tout en haut de l'axe, ce qui semble cohérent.

Conclusion de l'ACP :

L'axe 1 est le plus déterminant pour prioriser les pays nécessitant l'aide internationale, qui sont ceux les plus à gauche de l'axe 1 comportant les puissances économiques et les systèmes de santé les plus faibles..

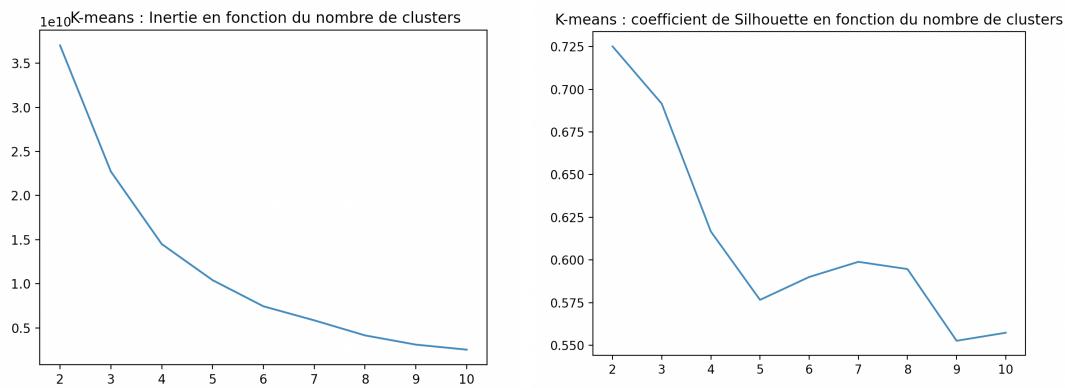
L'aide mondiale devrait se porter sur les infrastructures de santé pour limiter la mortalité infantile, et les aider également à se développer économiquement via des investissements directs à l'étranger(IDE) par exemple de pays riches.

3) Applications de différents algorithmes de clustering pour identifier le groupe de pays le plus démuni.

Nous allons utiliser les 5 méthodes étudiées en première partie sur notre jeu de données regroupant les données socio-économiques des différents pays du monde.

Algorithme K-means :

Il faut au préalable déterminer le nombre de clusters idéals pour réaliser le clustering, on applique alors la méthode du coude et celle du coefficient de silhouette pour déterminer le bon nombre de clusters.



Figures 11 et 12 : Détermination du nombres de clusters K

Première figure :

En traçant l'inertie en fonction du nombre de clusters, on déduit que l'on peut conserver idéalement entre 6 et 7 clusters.

Seconde figure :

La seconde figure montre qu'il peut-être plus intéressant de réaliser le clustering avec 7 ou 8 clusters plutôt que 6 en regard de la valeur du coefficient de silhouette pour chaque cluster.

De plus, pour partitionner les différents pays, il peut être intéressant de conserver un maximum de clusters pour isoler davantage les pays nécessitant le plus l'aide internationale.

On décide donc de conserver 8 clusters.

Résultats de l'algorithme K-means:

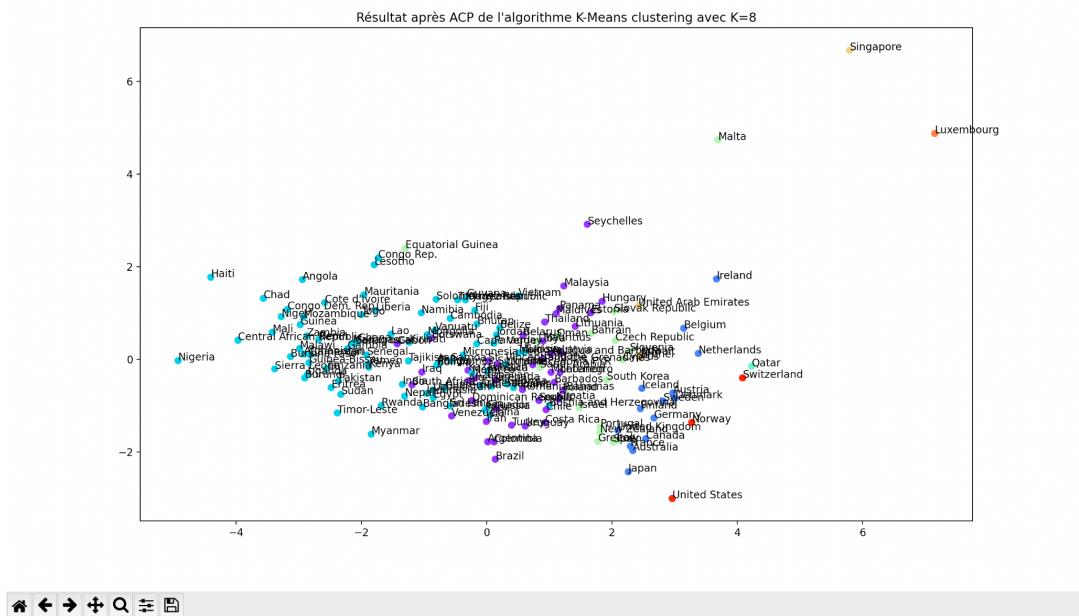


Figure 13 : Résultat du clustering réalisé pour K=8
coefficient de silhouette : 0,59.

On identifie clairement les différents clusters, on retrouve des points isolés qui constituent à eux seuls des clusters. De plus, le groupe des pays les plus démunis est assez conséquent (bleu ciel).

Algorithme Clustering hiérarchique ascendant :

On applique désormais le clustering hiérarchique ascendant à notre jeu de données. La première étape est de visualiser le dendrogramme pour savoir où couper l'arbre afin d'obtenir le bon nombre de clusters qui limitent la distance intra-classe. En choisissant $t=6$, on limite la distance intra classe et l'on obtient un nombre de clusters plus conséquent que pour l'algorithme K-means (15).

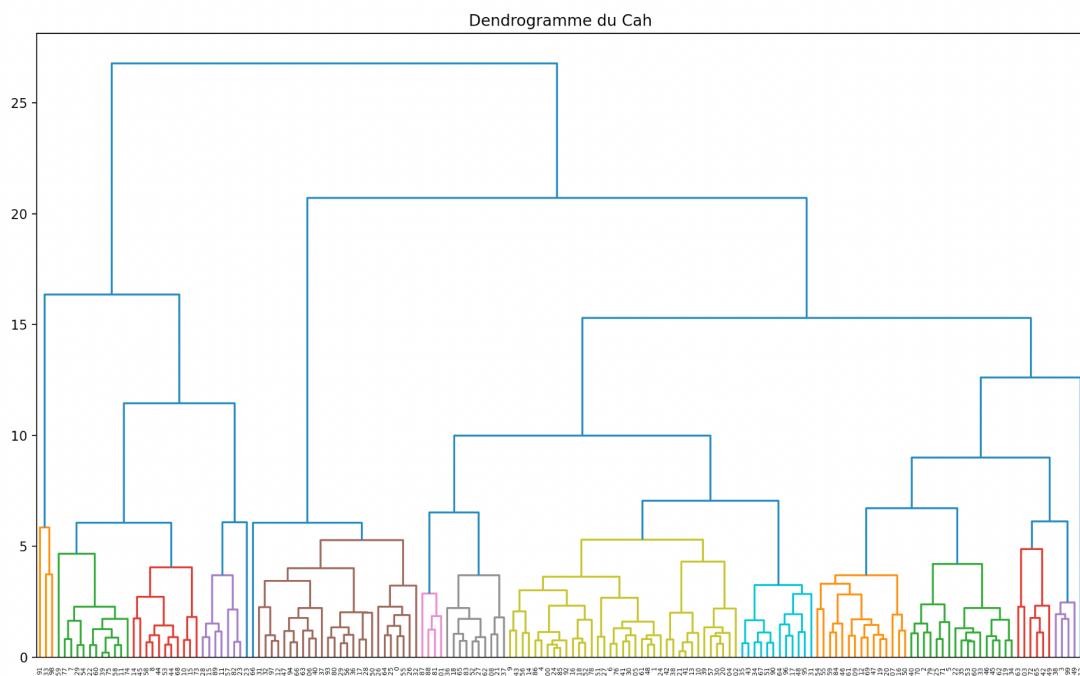
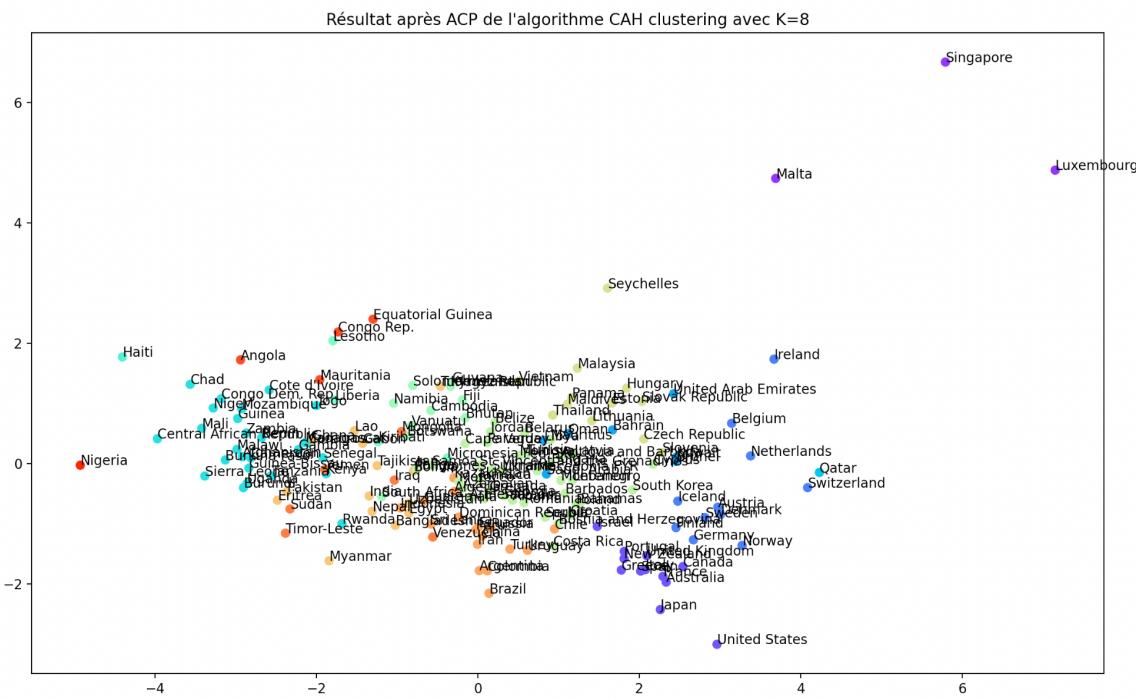


Figure 14 : Le Dendrogramme

Résultat de l'algorithme de clustering hiérarchique ascendant :



*Figure 15 : Résultat du clustering hiérarchique ascendant
coefficient de silhouette : 0,20*

L'augmentation du nombre de clusters par rapport à l'algorithme K-means nous permet d'identifier un groupe de pays les plus en difficultés plus restreint (bleu ciel + Angola et Nigeria) mais une dégradation du coefficient de silhouette car des clusters interfèrent entre eux.

L'algorithme DBSCAN :

Pour appliquer l'algorithme DBSCAN de façon optimale, il est nécessaire de déterminer la distance Epsilon pour chaque observation. On applique la méthode de recherche des plus proches voisins pour évaluer Epsilon.

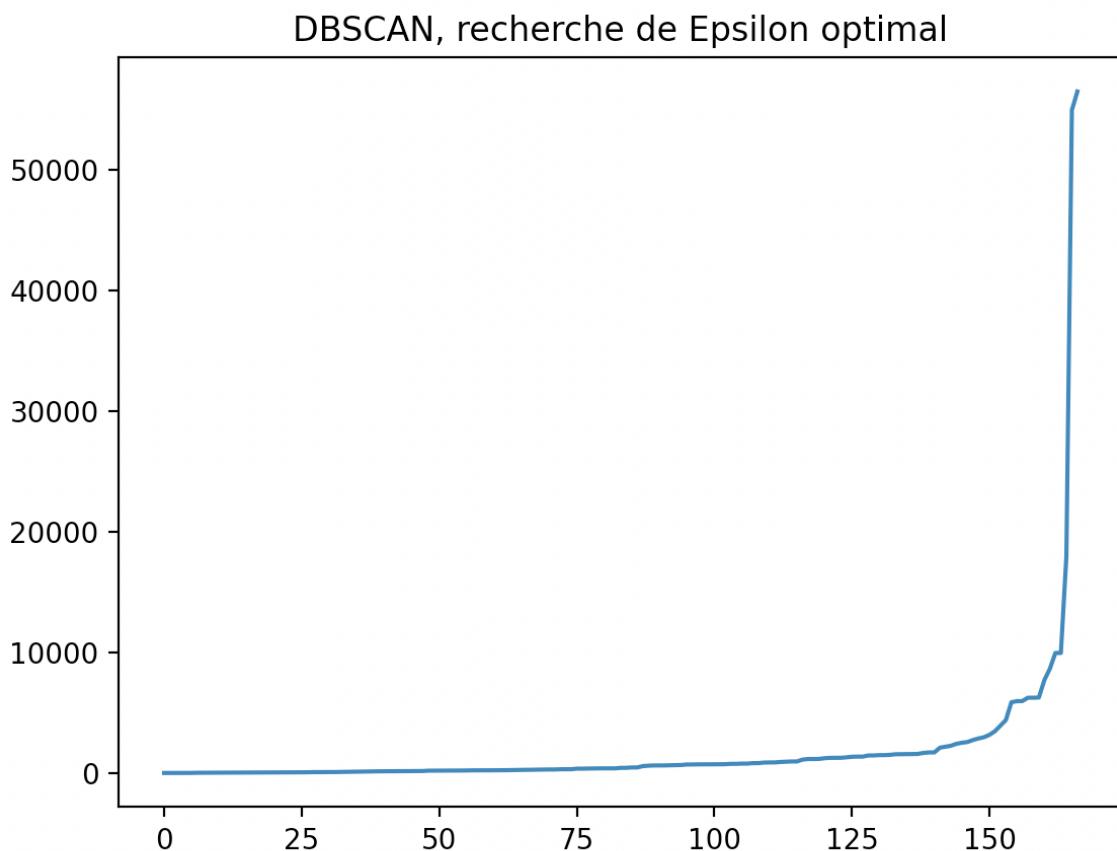
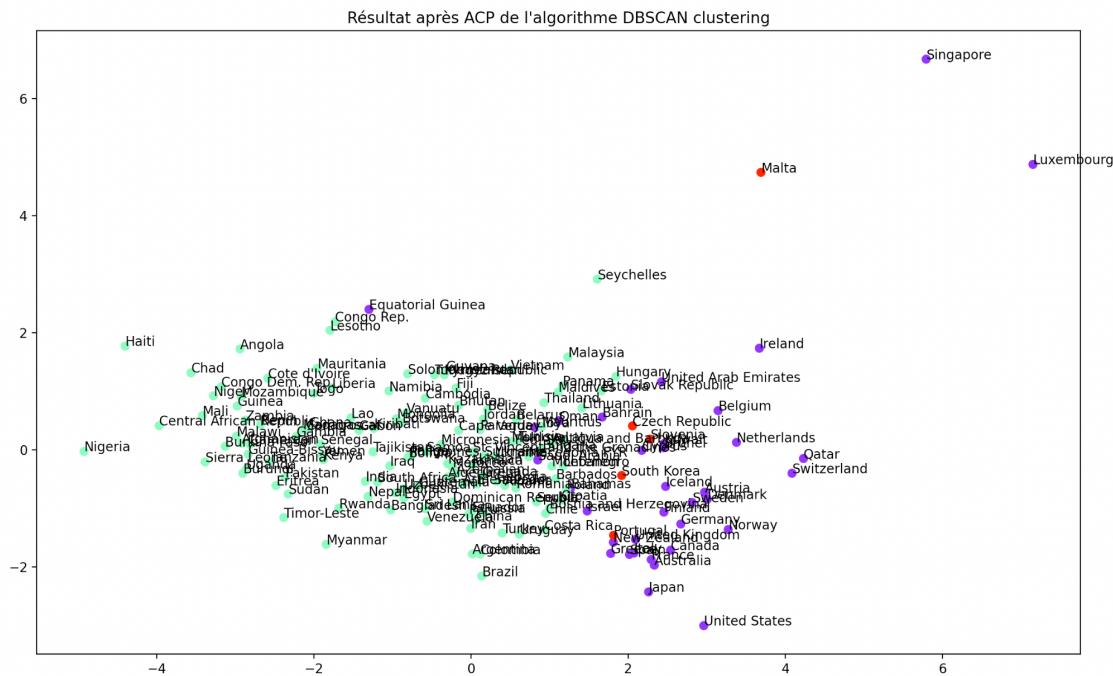


Figure 16 : voisins les plus proches de chaque observation ainsi que les distances.

Nous allons choisir un ϵ de tel sorte que 90% des observations aient une distance au proche voisin inférieure à ϵ .

ϵ compris entre 3000 et 6000 semble être un bon choix, après différents tests dans cette intervalle de valeurs, on a décidé de conserver $\epsilon=3200$ pour la suite.

Résultat de l'algorithme DBSCAN :



*Figure 17: Résultat de l'algorithme DBSCAN
coefficient de silhouette : 0.47*

Cet algorithme offre un coefficient de silhouette meilleur que le clustering hiérarchique ascendant. En résultat, on retrouve un grand cluster regroupant les pays les plus démunis selon les critères de l'axe 1. C'est l'inconvénient de cet algorithme, on ne peut pas déterminer au préalable le nombre de clusters voulus.

Algorithme du modèle de mélange de Gaussiennes :

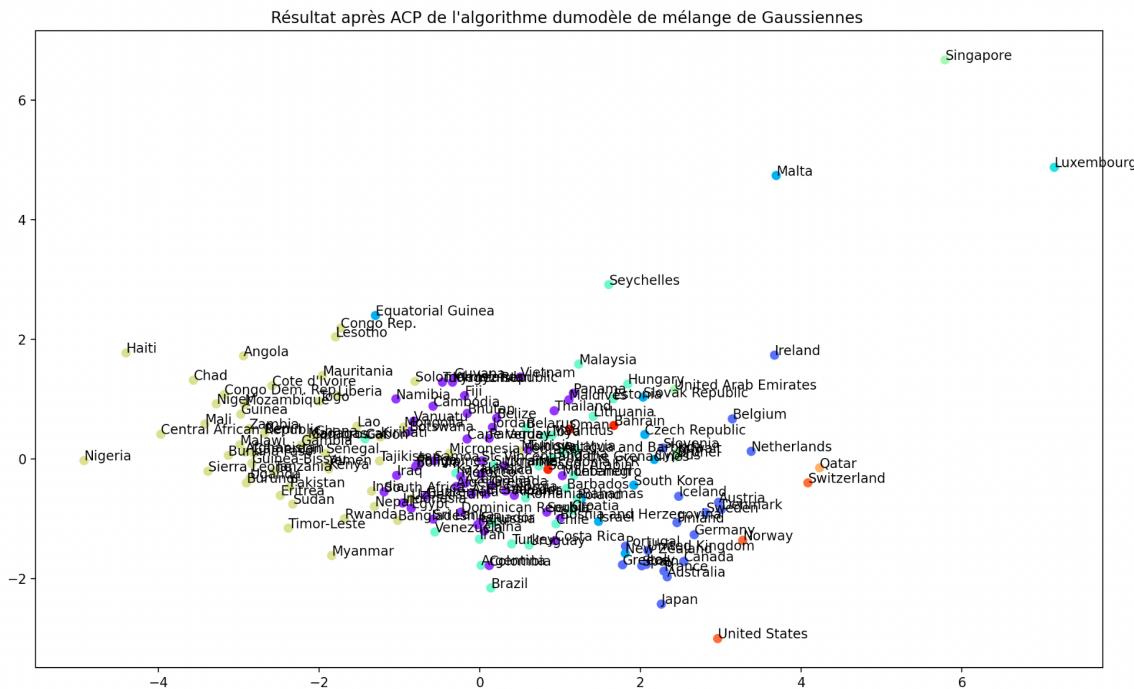


Figure 18: Résultat de l'algorithme du modèle de mélange de Gaussiennes
coefficient de silhouette : 0,43

Cet algorithme permet d'obtenir un cluster de pays démunis (en jaune) plus réduit que pour le DBSCAN avec un coefficient de silhouette meilleur que pour le clustering hiérarchique ascendant et l'algorithme K-means. C'est donc l'algorithme le plus intéressant pour le moment.

Spectral clustering :

On détermine grâce au coefficient de silhouette le nombre de clusters optimal pour appliquer le Spectral clustering à notre jeu de données

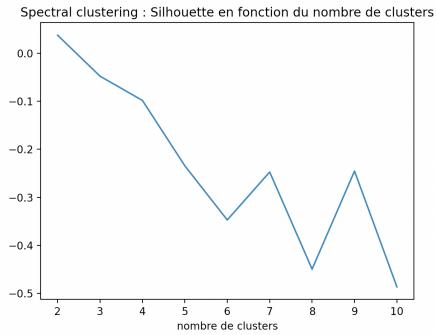


Figure 19: Détermination du nombre de clusters optimaux grâce au coefficient de silhouette

Afin de conserver un nombre de clusters suffisant, on choisit de conserver 9 clusters malgré un coefficient de silhouette négatif.

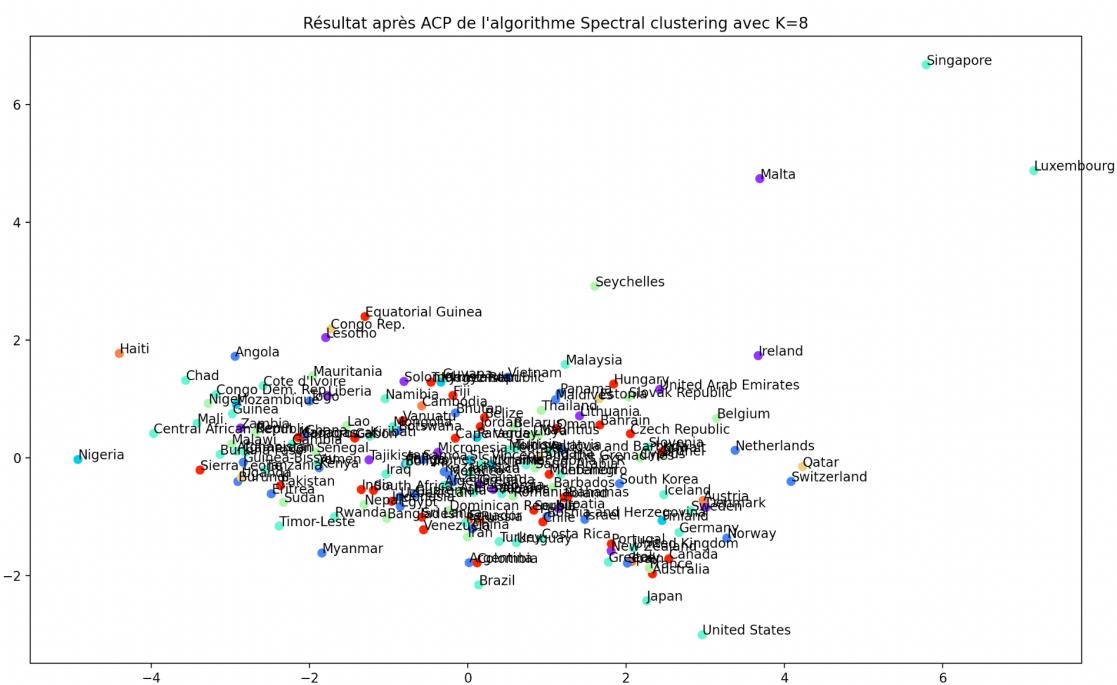


Figure 20 : Résultat de l'algorithme du Spectral Clustering
coefficient de silhouette : -0,38

Cet algorithme offre le pire résultat sur notre jeu de données, on ne distingue pas les différents clusters qui impliquent un coefficient de silhouette négatif. Cet algorithme ne nous permet pas de conclure à notre problématique initiale d'aide aux pays en besoin.

Analyse globale des différents algorithmes appliqués à notre jeu de données :

Au sens du coefficient de silhouette, le classement en terme de performance est le suivant par ordre décroissant :

- 1° DBSCAN le plus performant coefficient de silhouette) mais peu exploitable car le cluster de pays en difficulté est bien trop important.
- 2° modèle de mélange de Gaussiennes, le plus convaincant au regard du coefficient de silhouette et des clusters obtenus.
- 3° K-means, résultat proche du modèle de mélange de Gaussiennes avec un cluster des pays en difficulté assez conséquent.
- 4° CAH, résultat proche de K-means et du modèle de mélange de Gaussiennes
- 5° Spectral clustering le moins efficace, ne forme pas de clusters.

Conclusion : Aide à la décision pour les organisations de solidarité internationales

Grâce aux différents algorithmes mis en place dans cette étude, nous pouvons dresser une liste des pays qui ont le plus besoin de l'aide internationale. On s'appuie notamment pour cela principalement sur la base de l'ACP et du mélange gaussien (confirmé par K-means et Cah).

Par ordre d'importance décroissant:

Premier groupe : HAITI, NIGERIA, TCHAD, CENTRAFRIQUE, MALI, SIERRA LEONE, BURKINA FASO, NIGER, CONGO, BURKINA FASO.

2ème groupe (dans le même cluster) : MOZAMBIQUE, GUINÉE, ZAMBIE, MALAWI, AFGHANISTAN, CAMEROUN, TANZANIE, OUGANDA, BURUNDI, BÉNIN, GAMBIE.

Ces pays ont le plus besoin de l'aide internationale. Grâce à l'ACP, on sait qu'il faut privilégier des actions pour lutter contre la mortalité infantile et donc améliorer le système de santé d'une part mais aussi les aider à se développer économiquement via des investissements des ONG voire des pays les plus développés pour que ces pays en difficulté deviennent des pays en voie de développement.