INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

CR TP2 ML

Antonin Laborde-Tastet
Johann Courand
5AE SIEC

SOMMAIRE

I - Introduction

L'apprentissage non supervisé, et plus particulièrement le clustering, constitue une approche essentielle dans l'analyse de données non étiquetées, en regroupant ces données en fonction de leurs similarités et différences. Parmi les méthodes couramment utilisées, le K-means et le clustering agglomératif se démarquent comme des choix populaires. Ce rapport vise à mettre en œuvre ces méthodes sur des jeux de données en deux dimensions, avec pour objectif d'effectuer une analyse comparative afin de discerner les avantages spécifiques et les limitations inhérentes à chaque approche.

Dans le cadre de notre étude visant à déterminer le nombre optimal de clusters, nous ferons appel à des métriques d'évaluation essentielles, notamment le coefficient de silhouette, l'indice de Davies-Bouldin, et l'indice de Calinski-Harabasz. Ces métriques jouent un rôle crucial dans l'évaluation de la qualité des résultats obtenus par la méthode K-means.

Coefficient de Silhouette :

Valeur : Varie de -1 à 1. Une valeur proche de 1 indique un regroupement optimal.

Interprétation : Évaluant la proximité des instances à l'intérieur d'un cluster et leur éloignement des autres, un coefficient élevé suggère une bonne séparation des clusters, tandis qu'une valeur négative indique une mauvaise attribution des instances.

Indice de Davies-Bouldin :

Valeur : Plus basse est meilleure.

Interprétation : Mesurant la compacité intra-cluster et la séparation inter-cluster, une valeur faible indique des clusters bien définis et distincts, facilitant l'identification des groupes au sein des données.

Indice de Calinski-Harabasz :

Valeur : Plus élevée est meilleure.

Interprétation : Évaluant la séparation entre les clusters et leur compacité, une valeur plus élevée indique des clusters bien séparés et compacts, reflétant une organisation claire et distincte des groupes au sein des données.

II - Points forts et points faibles identifiés pour les différentes méthodes de clustering étudiées.

A. Méthode K-Means.

Principe de la Méthode K-means :

La méthode K-means est une technique d'apprentissage non supervisé qui vise à regrouper des données en clusters en fonction de leur proximité les unes par rapport aux autres. Le processus itératif comprend l'initialisation aléatoire de centres de gravité, l'assignation des points aux clusters les plus proches, la mise à jour des centres de gravité, et la répétition de ces étapes jusqu'à convergence.

Analyse du jeu de données "twodiamonds.arff" :

Le jeu de données "twodiamonds.arff" présente des clusters distincts, ce qui en fait un candidat approprié pour évaluer l'algorithme K-means. Visuellement, nous pouvons aisément identifier deux clusters clairement séparés dans le Figure 1.

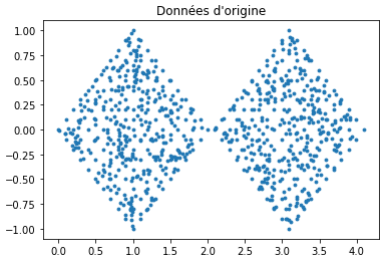


Figure 1 : Dataset “twodiamonds.arff” d'origine

Pour déterminer le nombre optimal de clusters (K), nous avons utilisé trois métriques d'évaluation : le coefficient de silhouette, l'indice de Davies-Bouldin et l'indice de Calinski-Harabasz. Ces métriques sont représentées dans la Figure 2.

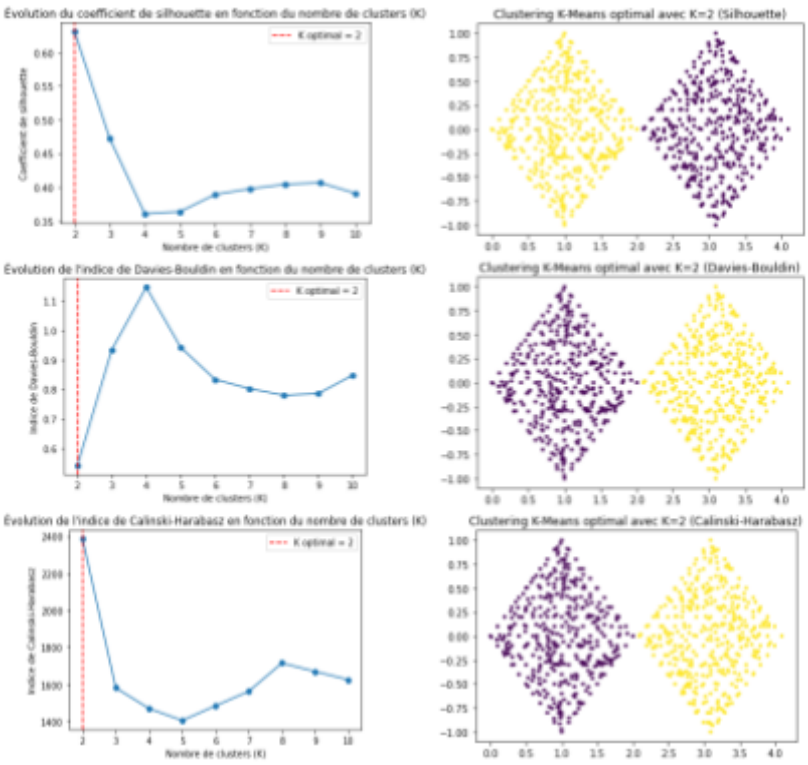


Figure 2 : Analyse "twodiamonds.arff" selon les trois critères

En analysant les courbes, on constate que chaque métrique converge vers la même conclusion : le nombre optimal de clusters est égal à 2. Cela signifie que l'algorithme K-means identifie correctement les deux clusters présents dans le jeu de données "twodiamonds.arff". Cette performance s'explique par la nature bien définie et séparée des clusters, ce qui facilite la tâche de K-means dans ce scénario.

Ainsi, cette analyse confirme l'efficacité de l'algorithme K-means sur le jeu de données "twodiamonds.arff" et valide visuellement la pertinence du choix du nombre de clusters.

Analyse du jeu de données "zelnik1.arff" :

Le jeu de données "zelnik1.arff" présente des clusters interconnectés, ce qui en fait un candidat peu approprié pour évaluer l'algorithme K-means. Visuellement, nous pouvons aisément identifier trois groupes distincts dans la Figure 3.

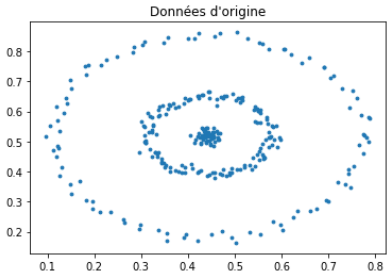


Figure 3 : Dataset "zelnik1.arff" d'origine

Comme précédemment, nous avons testé notre algorithme avec les trois métriques d'évaluation.

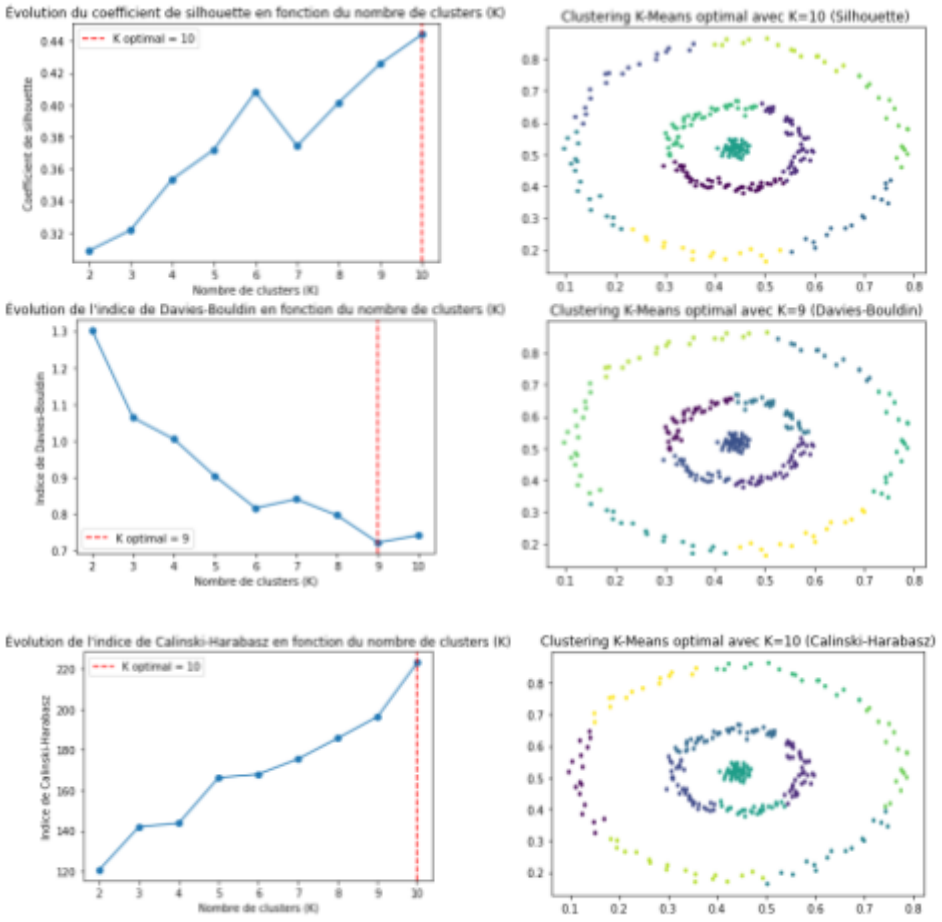


Figure 4 : Analyse "zelnik1.arff" selon les trois critères

En analysant les courbes, on constate que chaque métrique ne converge pas. Le nombre optimal de clusters est de 10 (valeur maximale). Cela indique que l'algorithme K-means ne parvient pas à identifier correctement les 3 clusters présents dans le jeu de données "zelnik1.arff".

Cette performance s'explique par le fait que l'algorithme K-means n'est pas adapté à ce genre de dataset, où les clusters sont interconnectés, rendant la séparation claire difficile. La complexité de la structure des données entrave l'efficacité de K-means dans ce contexte spécifique

Avantage de la Méthode K-means :

L'avantage majeur de la méthode K-means réside dans sa simplicité et sa rapidité d'exécution. Elle est efficace pour identifier des clusters homogènes dans des ensembles de données, facilitant ainsi la segmentation de données complexes en groupes distincts.

Inconvénient de la Méthode K-means :

Un inconvénient notable de K-means est sa sensibilité à l'initialisation des centres de gravité. Les résultats peuvent varier en fonction des positions initiales, et la méthode peut avoir des difficultés à identifier des clusters de formes complexes ou irrégulières.

B. Étude de la méthode de clustering agglomératif.

Principe de la Méthode de Clustering Agglomératif :

La méthode de clustering agglomératif diffère de K-means en regroupant les données de manière ascendante, débutant avec chaque point comme un cluster distinct, puis fusionnant progressivement les clusters les plus similaires jusqu'à la formation d'un cluster global. Ce processus hiérarchique génère un dendrogramme illustrant les relations entre les clusters.

Variation du paramètre "distance_threshold" :

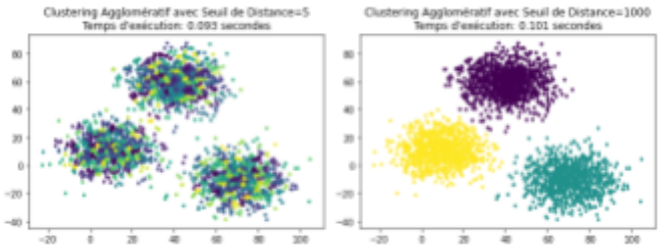


Figure 6 : Clustering agglomératif avec différent treshold

Nous avons exploré la variation du paramètre "distance_threshold" en laissant le paramètre "n_clusters" à None. Pour le jeu de données actuel, nous avons observé une amélioration significative des résultats à mesure que le seuil de distance augmente. Après une analyse approfondie, nous avons décidé de fixer le seuil à 1000 pour obtenir les performances optimales.

Variation du paramètre "n_clusters" et analyse :

Dans cette partie, nous avons utilisé le jeu de données "clara.arff".

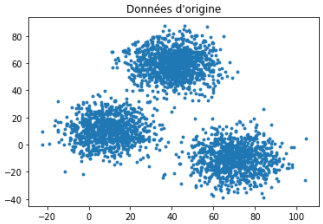


Figure 7 : Dataset "clara.arff" d'origine

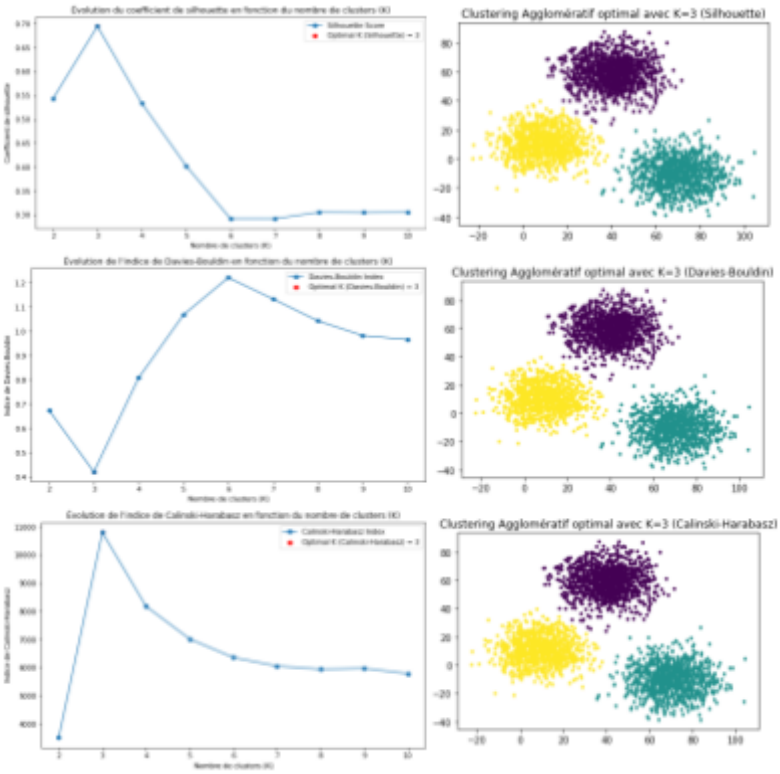


Figure 8 : Analyse "clara.arff" selon les trois critères

En poursuivant avec le jeu de données "clara.arff" et en utilisant l'algorithme Clustering Agglomératif avec le type de lien "single", nous avons obtenu des résultats prometteurs. L'analyse des courbes des trois métriques (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index) a montré une convergence vers la même conclusion : le nombre optimal de clusters est égal à 3.

Cette constatation démontre que l'algorithme Clustering Agglomératif, avec le type de lien "single", parvient à identifier correctement les trois clusters présents dans le jeu de données "clara.arff". Cette méthode semble particulièrement adaptée à ce type de dataset, caractérisé par une certaine séparation entre les clusters et un choix judicieux du type de lien.

Ainsi, notre analyse renforce la pertinence de l'algorithme Clustering Agglomératif sur le jeu de données "clara.arff". Les résultats obtenus valident visuellement le choix du nombre de clusters, renforçant ainsi la fiabilité de notre approche.

Avantages de la Méthode de Clustering Agglomératif :

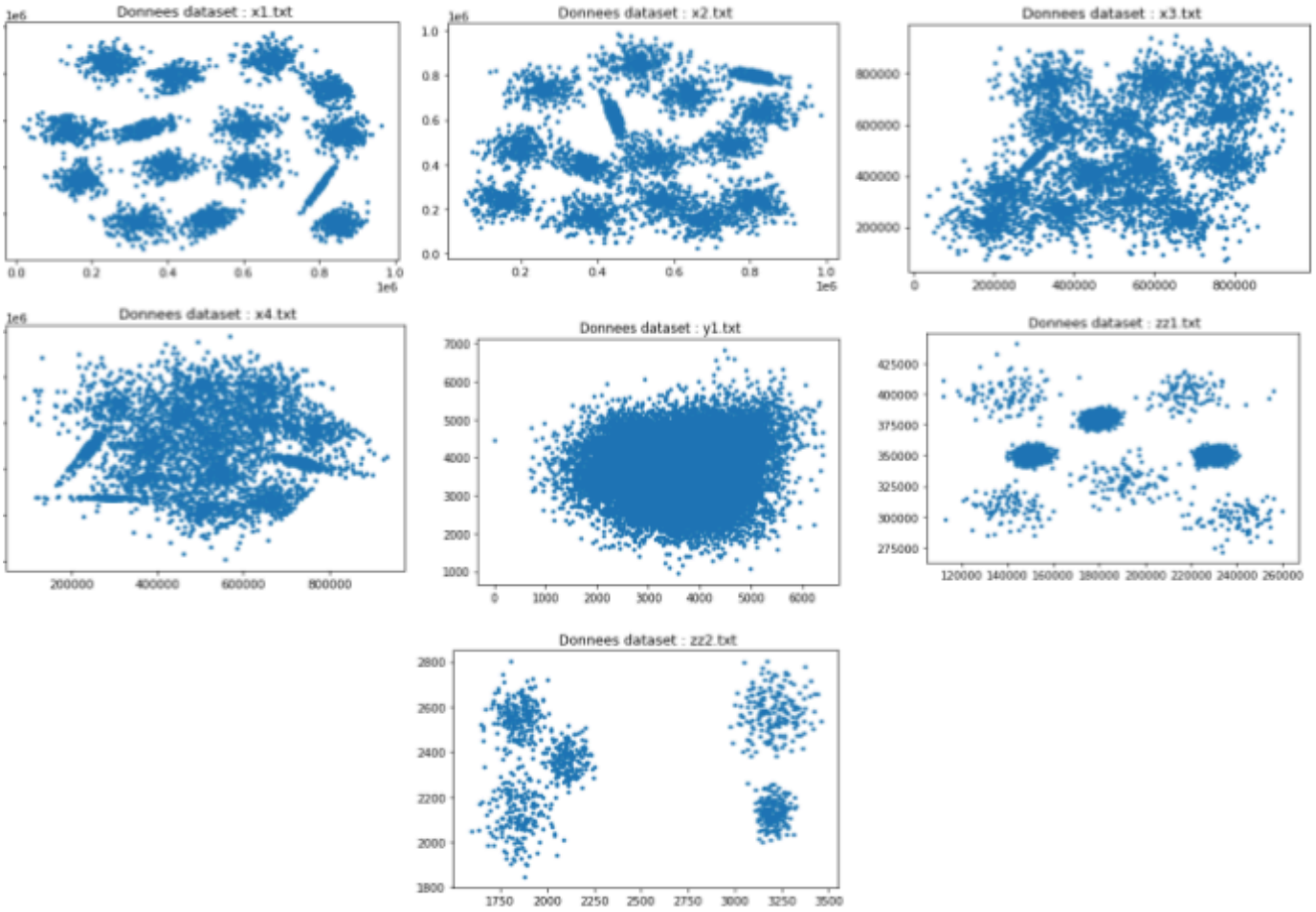
La méthode agglomérative offre une flexibilité accrue par rapport à K-means, adaptée à la détection de clusters de formes variées. Son approche hiérarchique fournit une vue détaillée des relations entre les clusters. Elle est généralement plus interprétable et ne nécessite pas d'initialisation aléatoire des centres de gravité.

Inconvénients de la Méthode de Clustering Agglomératif :

Cependant, la méthode agglomérative est sensible aux outliers et peut être relativement lente pour des datasets de grande dimension. La construction d'un dendrogramme peut être coûteuse en termes de temps et de ressources computationnelles.

III - Etude et Analyse comparative de méthodes de clustering sur de nouvelles données.

C. Jeux de données



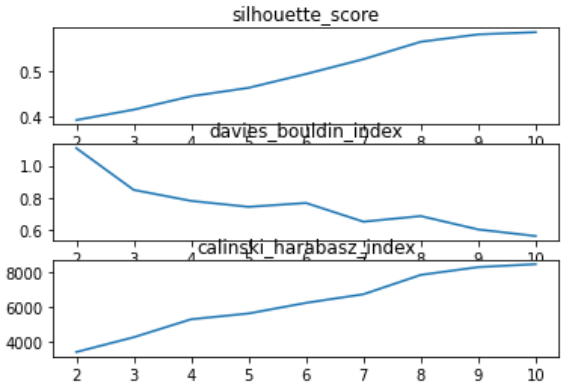
D. Clustering k-Means

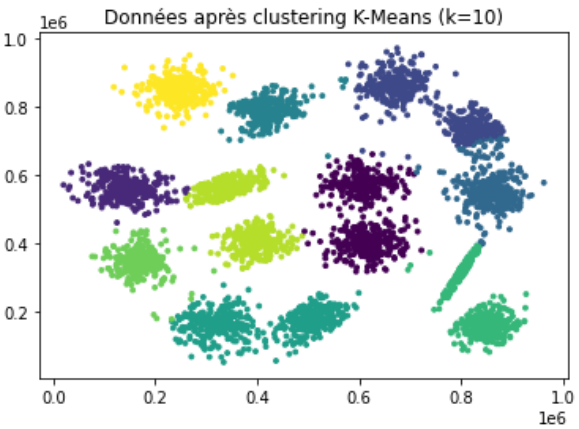
On applique dans cette section l'algorithme de clustering ‘k-means’ au 7 datasets fournis. On utilise les métriques d’évaluations: “Silhouette”, “Davies-Bouldin” et “Calinski-Harabasz” présentées dans la partie précédente pour déterminer le nombre de cluster optimal pour chaque dataset.

Figure 9 : Plot des métriques et clustering optimal pour chaque dataset avec “K-means” Clustering

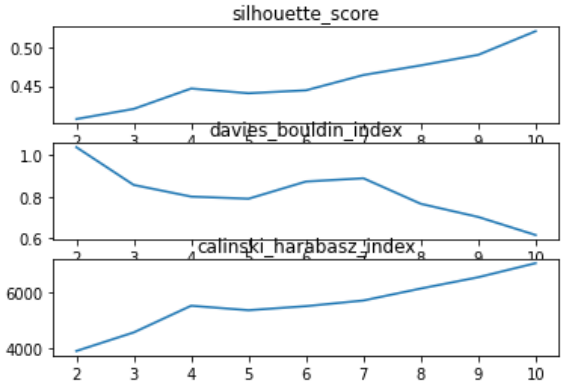
Dataset	Metrics	Best inference

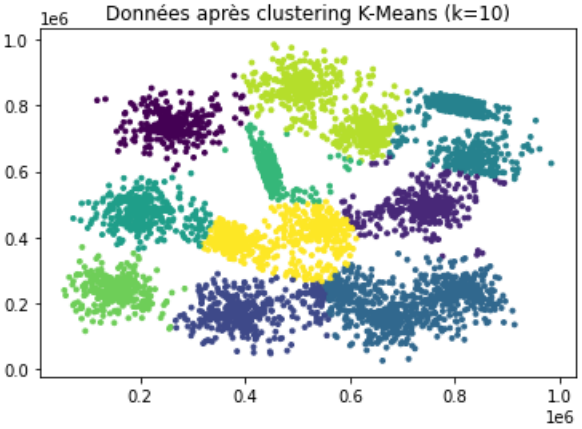
x1 S


silhouette_score
davies_bouldin_index
calinski_harabasz_index

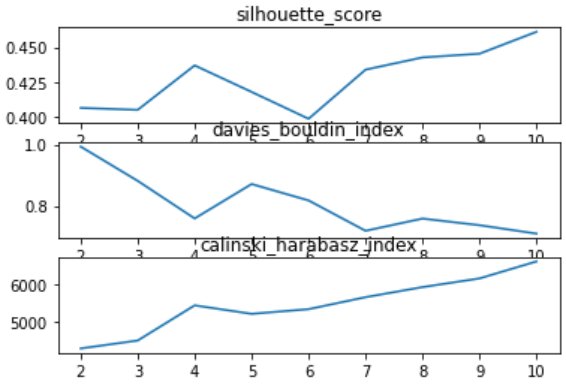

Données après clustering K-Means (k=10)
1e6

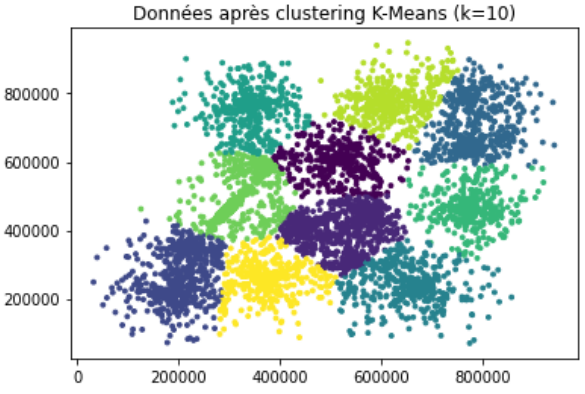
x2


silhouette_score
davies_bouldin_index
calinski_harabasz_index

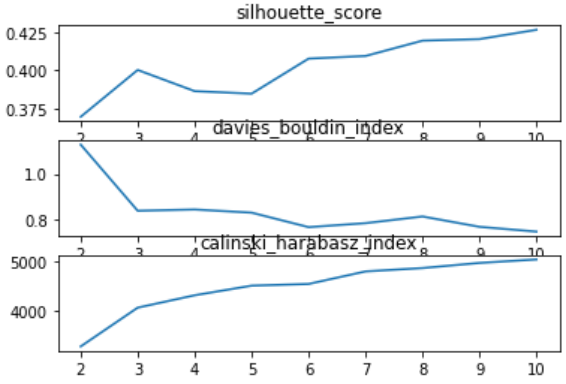

Données après clustering K-Means (k=10)
1e6

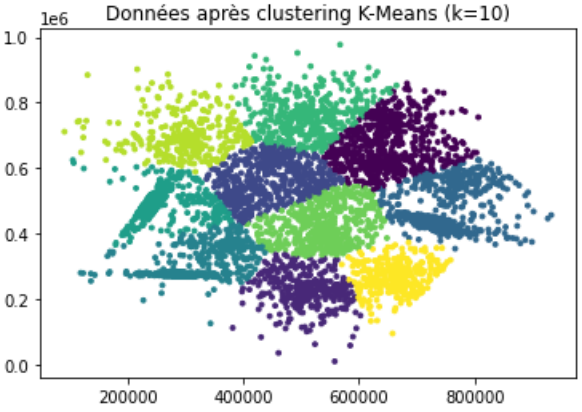
x3


silhouette_score
davies_bouldin_index
calinski_harabasz_index

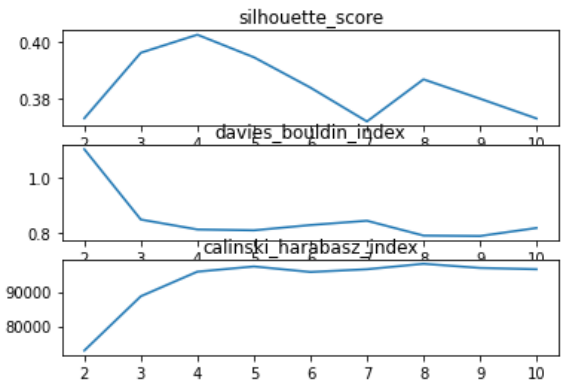

Données après clustering K-Means (k=10)

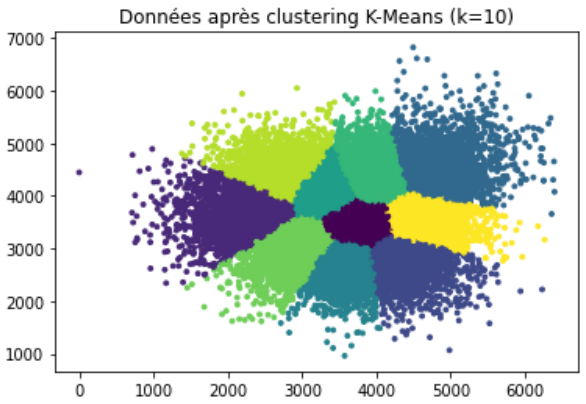
x4


silhouette_score
davies_bouldin_index
calinski_harabasz_index

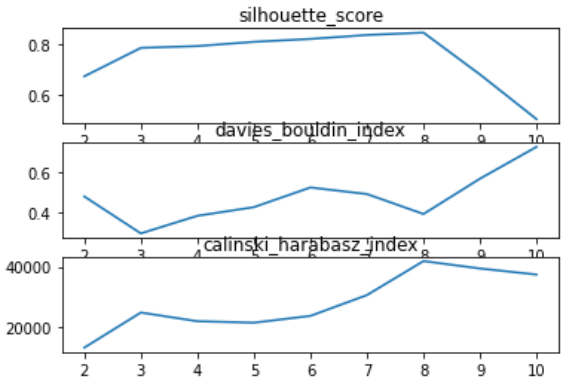

Données après clustering K-Means (k=10)
1e6

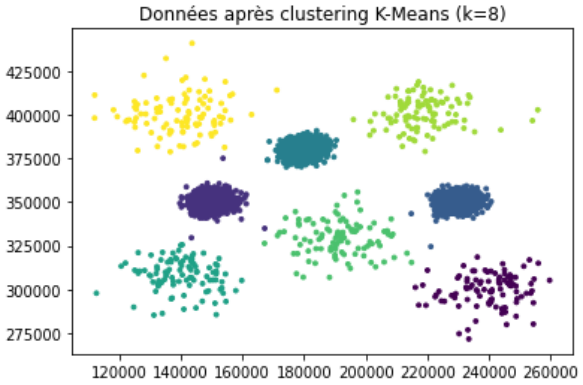
y1



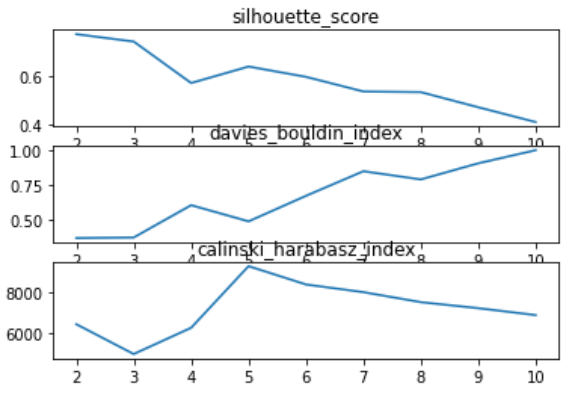


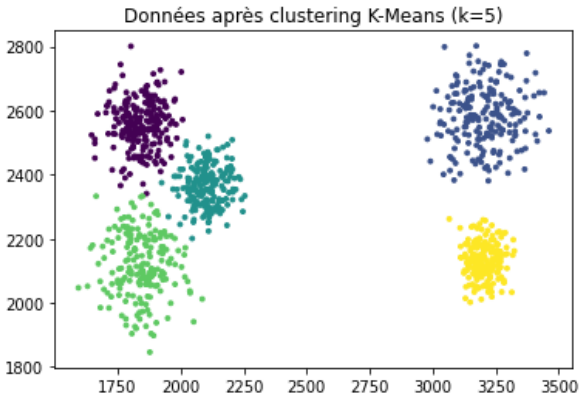
zz1





zz2





On voit sur les datasets **x1, zz1 et zz2** que le **k-means** clustering donne des clusters très proches de l'identification humaine, avec des **clusters distincts**. C'est un résultat cohérent, en effet les datasets x1, zz1 et zz2 possèdent des clusters **homogènes**, de **taille semblable** et **relativement sphériques** ce qui correspond au mieux au fonctionnement de cet algorithme de clustering. Les 3 métriques sont unanimes pour désigner le nombre optimal de cluster ce qui montre bien la cohérence du résultat obtenu. Cependant on peut noter que les clusters **zz1 et zz2 ont des concentrations variables**, ce qui est bien toléré par l'**aglometrative clustering**. On pourrait donc avoir de meilleurs résultats avec cette méthode.

Pour les datasets **x2, x3** l'algorithme **k-means performe moins bien**. Les clusters correspondent moins à l'intuition humaine et sont **moins distincts**. Les métriques convergent toutes vers un même nombre de clusters mais les résultats moins bons peuvent s'expliquer par la forme des clusters qui sont ici moins homogènes et de **formes variées**.

Finalement, pour les datasets **x4,y1**, on retrouve le pire résultat de clustering. **K-means ne parvient pas créer de clusters distincts**. Les 3 métriques d'évaluation ne donnent pas clairement de résultat optimal et la frontière des clusters ne semble pas suivre de logique. Il s'agit du cas où K-means performe le moins. Dans un cas de **clusters de taille et forme très variées, non distincts** et l'autre pas de réels clusters présent (ou un seul du moins).

E. Clustering agglomératif

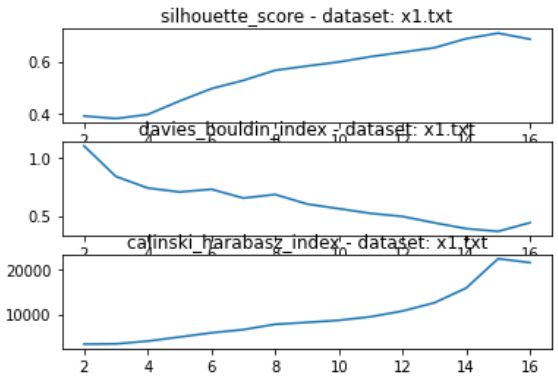
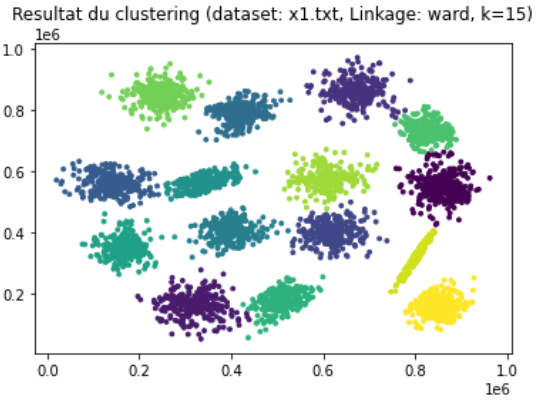
On applique dans cette section l'algorithme de clustering 'agglomerative' au 7 datasets fournis. On utilise les métriques d'évaluations: "Silhouette", "Davies-Bouldin" et "Calinski-Harabasz" présentées dans la partie précédente pour déterminer le nombre de cluster ainsi que le méthode de liaison (linkage method) optimal pour chaque dataset.

Les 4 linkage method que nous utiliserons sont:

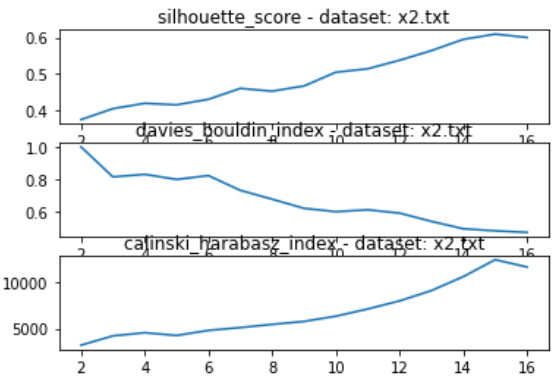
- 'ward' minimise la variance des clusters fusionnés.
- 'average' utilise la moyenne des distances de chaque observation des deux clusters.
- 'complete' utilise les distances maximales entre toutes les observations des deux clusters.
- 'single' utilise le minimum des distances entre toutes les observations des deux clusters.

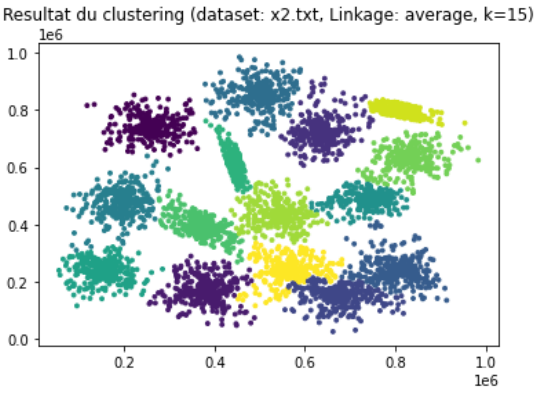
Selon : https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Figure 10 : Plot des métriques et clustering optimal pour chaque dataset avec "Agglomerative" Clustering

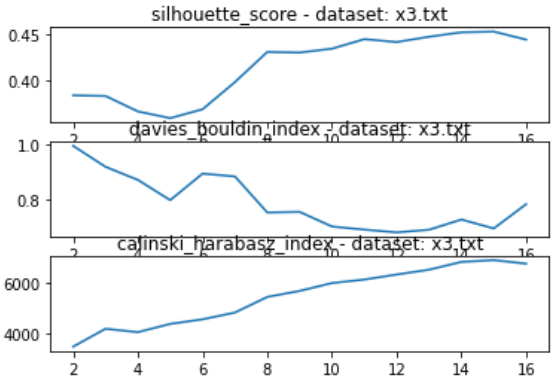
Dataset	Linkage method	Metrics	Best inference
x1	Ward		

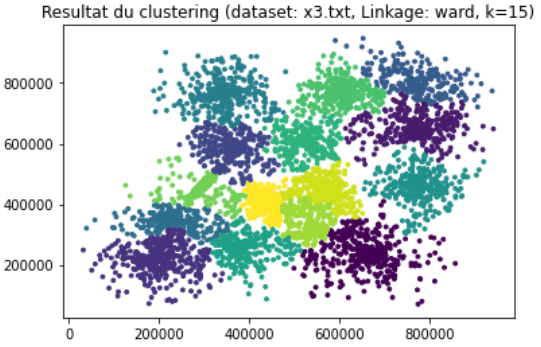
x2 Average



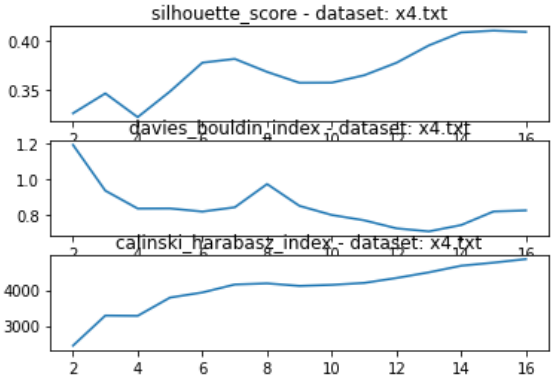


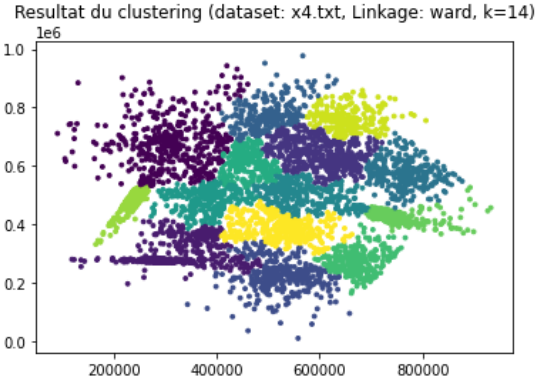
x3 Ward



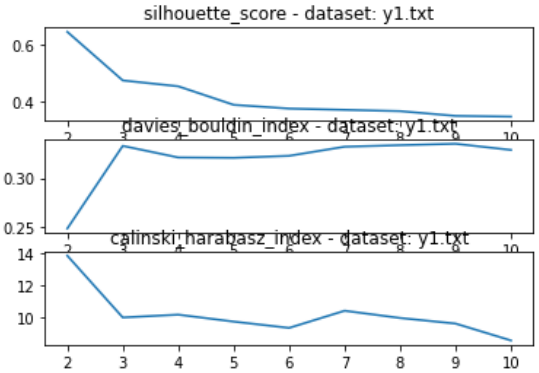


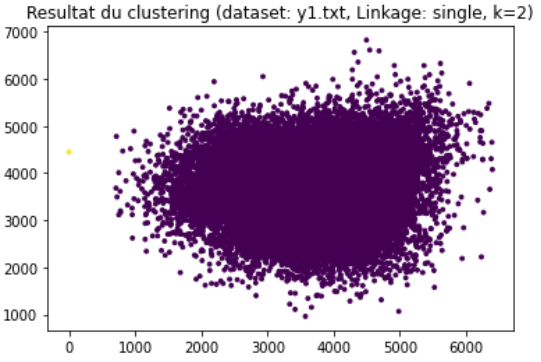
x4 Ward



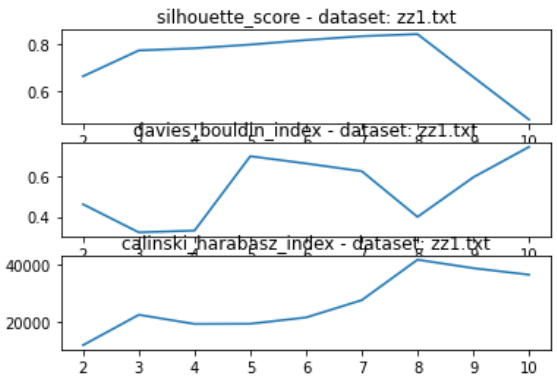


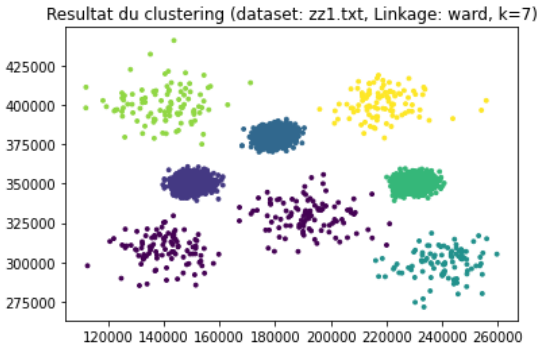
y1 single



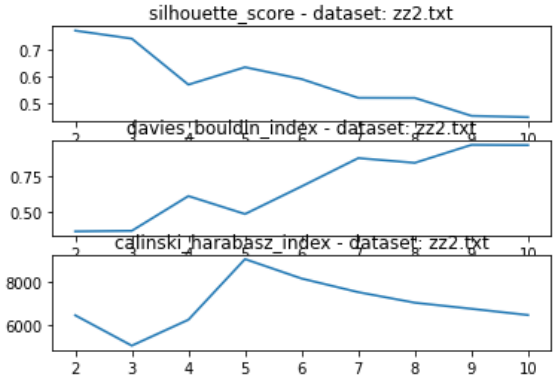


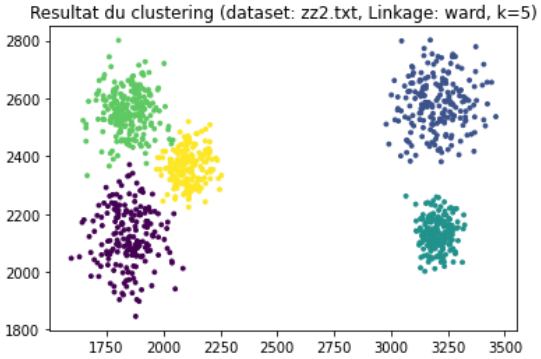
zz1 Ward - Average





zz2 Ward - complete





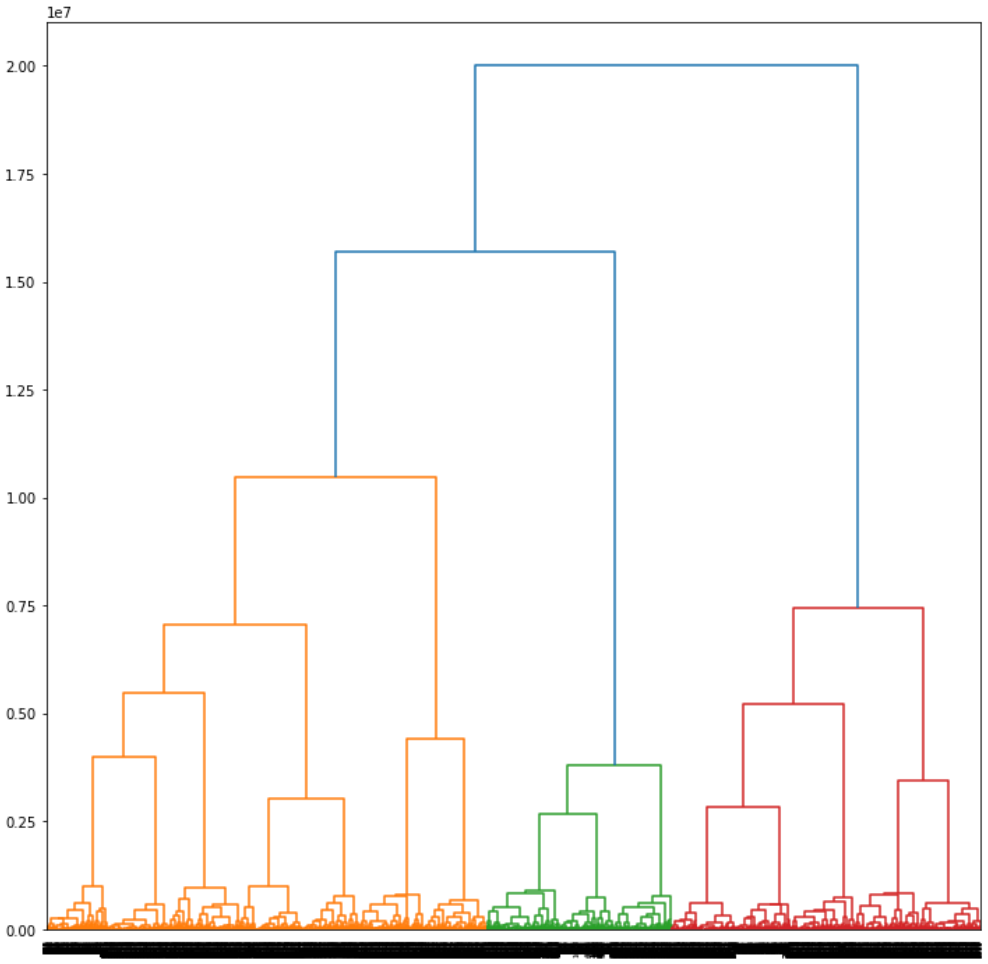
Pour les datasets **x2 et zz1**, on voit que le clustering **agglomératif performe bien mieux que le k-means**. Les clusters étant de f**ormes, tailles et concentrations variées**, les résultats sont donc plus réalistes et les clusters plus distincts. De plus, la méthode de liaison **ward** et adaptée à ce type de configuration, la **minimisation de la variance intra-cluster**s peut donner lieux à des clusters compacts mais aussi à des clusters de tailles différentes.

Pour le **dataset y1**, le temps et les **ressources de calcul nécessaires sont trop importants**. On ne peut obtenir de résultat que pour la méthode de liaison 'single'. Étant donné qu'il s'agit d'un seul cluster, le résultat est donc convenable. Sans avoir pu tester les autres méthodes de liaison.

```
MemoryError: Unable to allocate 41.5 GiB for an array with shape
(5575627200,) and data type float64
```

Pour les datasets **x1, x3,x4 et zz2**, la méthode **agglomérative performe de manière équivalente à la méthode k-means**. Proposant cependant **plus de modularité dans les résultats vis-à-vis de la taille et formes** des clusters avec la linkage method mais avec en **contrepartie un temps de calcul bien plus important.** Cependant l'algorithme **agglomératif a pour avantage d'aider à trouver facilement le nombre de clusters optimal** en traçant par exemple un **dendrogramme**.

Figure 11 : Dendrogramme de x2 avec linkage "ward"



IV - Conclusion.

Pour conclure, les algorithmes de clustering ‘k-means’ et ‘agglomerative’ ont donné des résultats satisfaisants pour les 7 datasets fournis. Avec cependant quelques points négatifs.

Concernant ‘k-means’, on peut noter le manque d’adaptabilité pour des datasets dont la forme, taille et concentration varient. En effet, l’algorithme donne de bons résultats pour des clusters homogènes, de taille semblable et relativement sphériques. On peut également noter que k-means est plus sensible aux outliers, qui ‘tirent’ le centroïdes dans leur direction et altèrent la formation des clusters. De plus, il faut indiquer à l’algorithme le nombre de clusters désirés ce qui peut s’avérer problématique dans un cas réel ou on ne pourrait pas visualiser les valeurs.

Concernant ‘Agglomerative’, l’algorithme fait preuve d’une meilleure capacité d’adaptation concernant la taille, forme des clusters, et performe également bien avec des clusters compacts. Les différentes méthodes de liaisons permettent également de prendre en compte les spécificités des différents cas. De plus, il n'est pas nécessaire de fournir un nombre de clusters prédéfini en entrée et le nombre optimal peut être trouvé grâce au dendrogramme. Malheureusement, cette méthode nécessite un temps et une capacité de calcul bien plus importante que ‘k-means’ ce qui est problématique dans le cas de très grands datasets (comme y1).

Annexe

Lien github : https://github.com/AntoninLT/ML_unsupervised_Laborde-Tastet_Courand