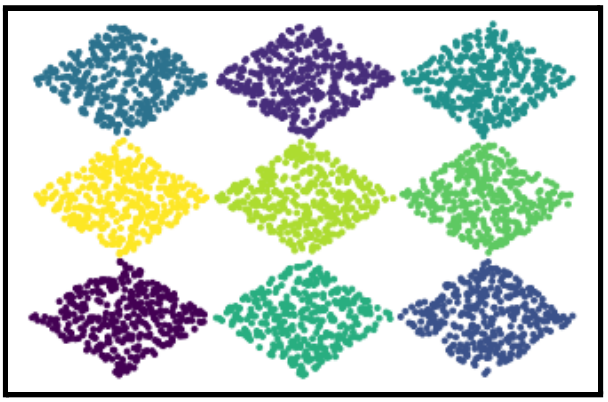
TP Clustering



Git : https://github.com/ALievre/BigData.git

Agathe Lièvre
Assia Nguyen

5ISS

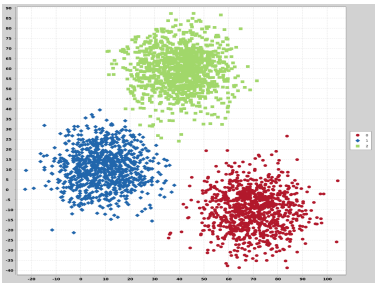
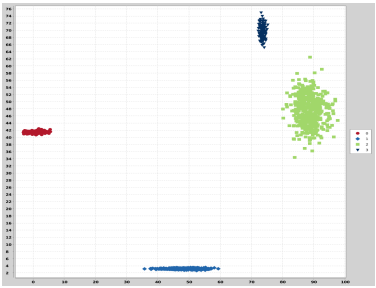
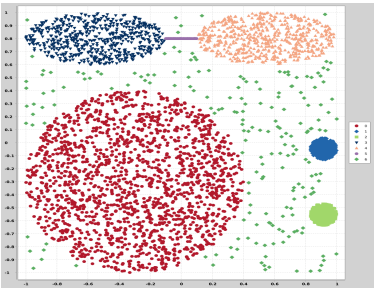
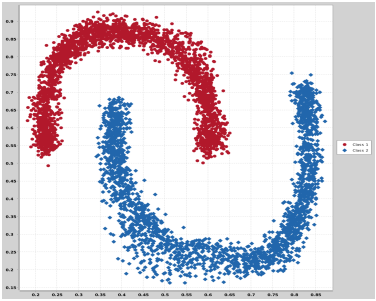
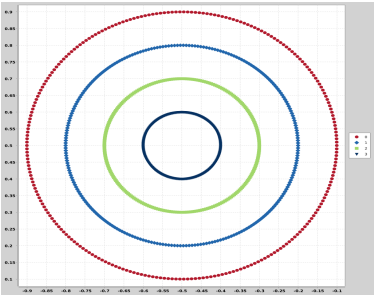
2021/2022

I. Introduction

Ce TP a pour but de tester et évaluer différentes méthodes de clustering fournies par *scikit-learn*. Nous allons donc les implémenter sur différents data sets et ainsi en déduire leurs avantages et leurs inconvénients. Nous voulons pouvoir analyser quelles méthodes et métriques sont les plus adaptées pour les data sets à la fois fictifs et réels.

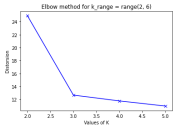
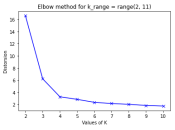
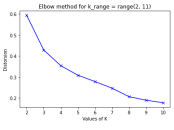
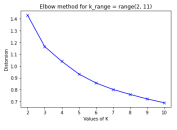
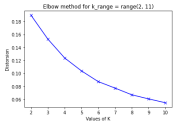
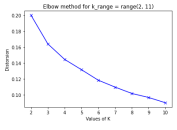
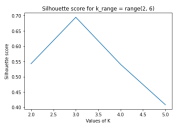
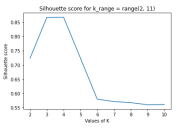
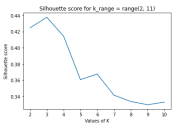
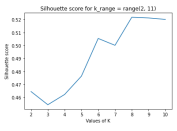
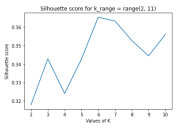
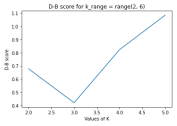
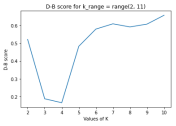
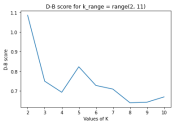
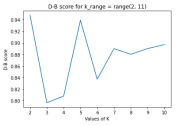
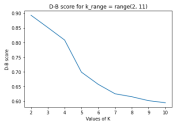
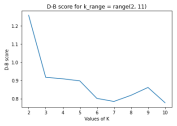
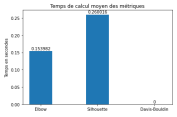
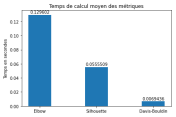
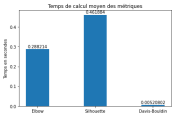
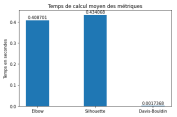
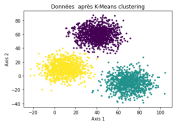
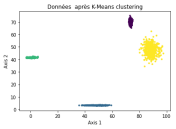
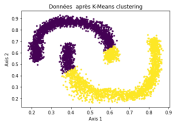
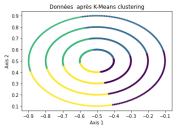
II. Analyse des méthodes de clustering

Pour faire l'analyse des différentes méthodes de clustering, nous avons sélectionné 6 data sets qui nous paraissaient assez différents et variés. Les voici ci-dessous avec leur caractéristiques principales:

xclara  - convexe - densité variable	**2d-4c**  - convexe - bien séparé - densité variable	**cure-t2-4k**  - convexe - densité variable - très bruité
engytime - convexe - mal séparé - densité variable	**banana**  - non convexe - densité variable	**dartboard1**  - non convexe - bien séparé - non bruité

A. Clustering k-Means

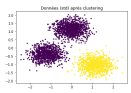
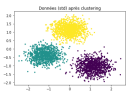
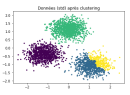
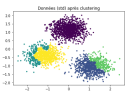
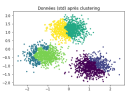
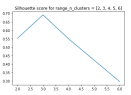
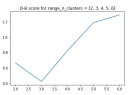
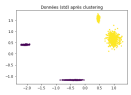
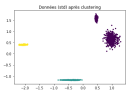
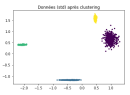
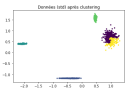
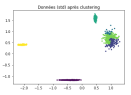
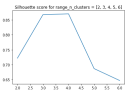
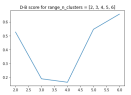
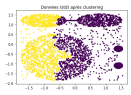
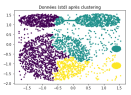
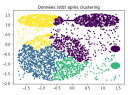
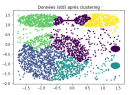
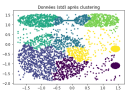
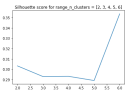
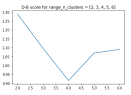
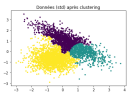
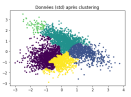
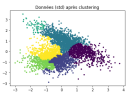
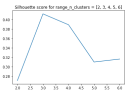
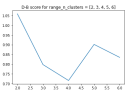
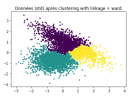
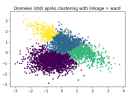
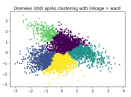
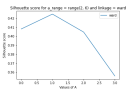
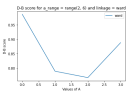
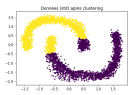
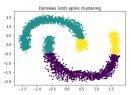
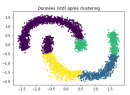
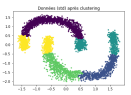
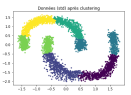
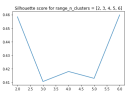
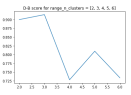
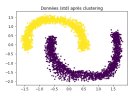
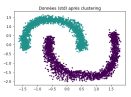
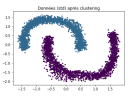
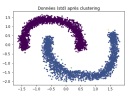
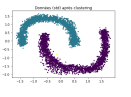
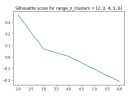
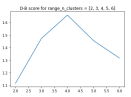
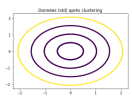
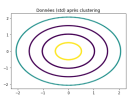
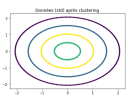
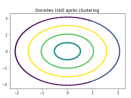
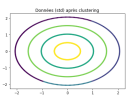
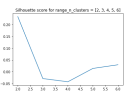
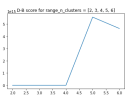
Nous allons commencer par évaluer la méthode de clustering k-Means.

xclara	2d-4c	cure-t2-4k	engytime	banana	dartboard1
 3 clusters	 3 ou 4	 Indéterminé	 3 clusters	 Indéterminé	 Indéterminé
 3 clusters	 4 clusters	8 clusters	 3 clusters	 8 clusters	 6 clusters
 3 clusters	 4 clusters	 8 clusters	 3 clusters	 Indéterminé	 7 clusters
					
 En réalité : 3 clusters	 En réalité : 4 clusters	En réalité : 6 clusters	En réalité : 2 clusters	 En réalité : 2 clusters	 En réalité : 4 clusters

Nous pouvons donc définir ses points forts et ses points faibles:

Points forts	Points faibles
Fonctionne avec les clusters convexes et de densité variable	Bruité et non convexe

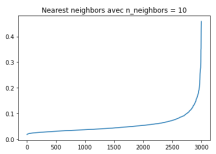
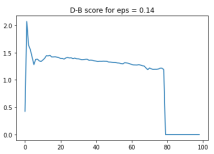
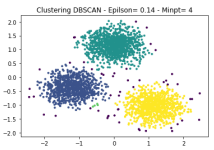
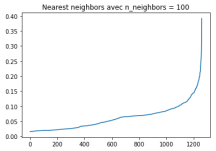
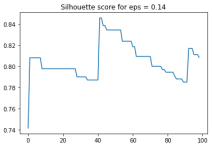
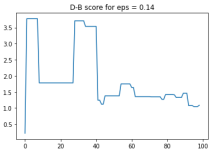
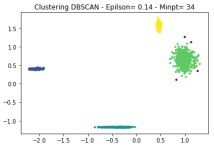
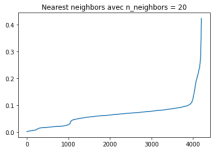
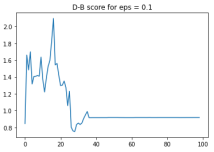
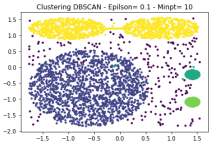
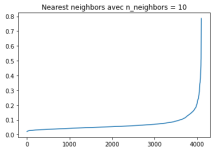
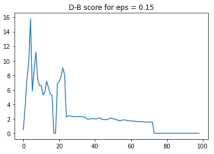
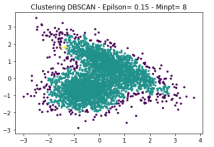
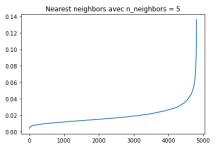
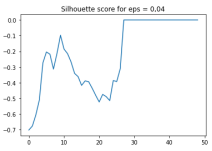
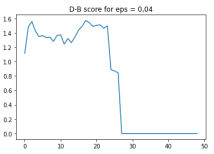
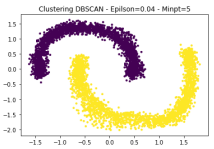
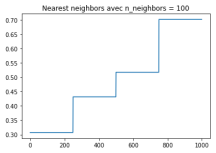
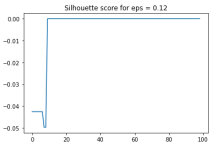
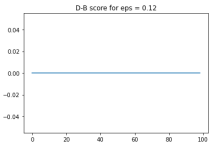
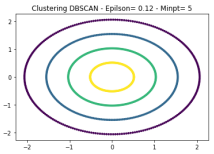
B. Clustering agglomératif

xclara ‘complete’							
2d-4c ‘complete’							
cure-t2-4k ‘complete’							
engytime ‘complete’							
‘ward’							
banana ‘complete’							
‘single’							
dart board1 ‘complete’							

Nous pouvons donc définir ses points forts et ses points faibles:

	Points forts	Points faibles
‘complete’	Fonctionne avec les clusters convexes et de densité variable	Bruité et non convexe
‘single’	Non convexe, bien séparé, même densité ou densité variable	Convexe, bruité
‘ward’	Mal séparé	

C. Clustering DBSCAN

xclara noise = 51				
2d-4c noise = 4				
cure-t2-4k noise = 199				
engytime noise = 305				
banana				
dartboard1				

Nous pouvons donc définir ses points forts et ses points faibles:

Points forts	Points faibles
Non convexe et non bruité	Bruit, variation densité et mal séparé

D. Récapitulatif

	xclara	2d-4c	cure-t2-4k	engytime	banana	dartboard1
K-Means	OK	OK	X	X	X	X
Agglomératif	OK	OK	X	X	X	X
DBSCAN	~~	OK	~X	X	OK	OK

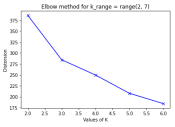
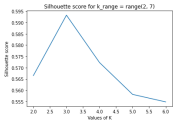
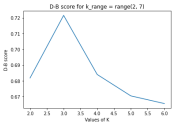
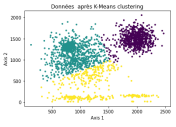
III. Analyse comparative sur des données mystères

Dans cette partie, nous allons utiliser les méthodes de clustering que nous avons découvertes dans la partie précédente pour analyser des sets de données mystères. Nous avons également représenté pour le k-Means de chacun des data sets le temps de calcul.

A. Datasets fictifs

1. N1

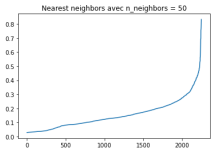
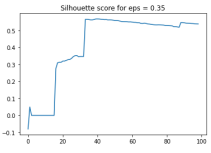
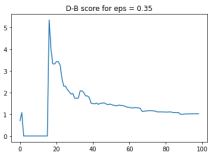
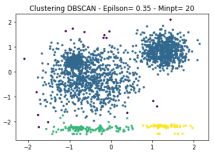
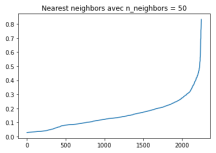
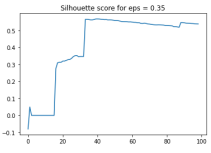
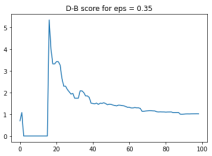
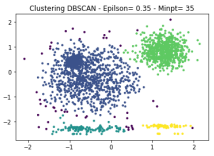
a) Clustering k-Means

Range	Méthode du coude	Silhouette score	Davis-Bouldin score	Temps de calcul	Après clustering
[2;6]	 3 clusters	 3 clusters	 3 clusters		

b) Clustering agglomératif

En annexe, vous pouvez voir les tests que nous avons effectués en faisant varier le linkage et le nombre de clusters. Nous avons remarqué que ‘ward’ avec n_clusters = 3 et ‘average’ avec n_clusters = 3 ou 4 donnaient des résultats se rapprochant de ce que nous avons obtenu avec la méthode de clustering K-Means.

c) Clustering DBSCAN

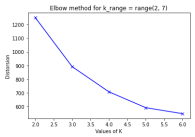
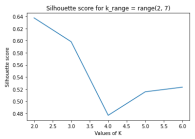
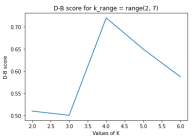
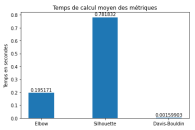
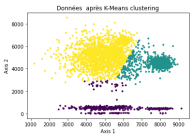
n_cluster = 3				
n_cluster = 4				

d) Comparaison des méthodes

Même si les trois méthodes nous donnent bien le même nombre de clusters (3), nous observons visuellement que la méthode agglomérative donne de meilleurs résultats.

2. N2

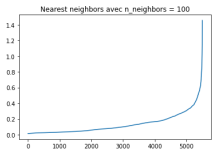
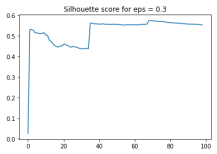
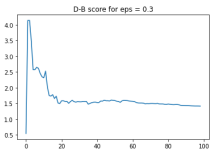
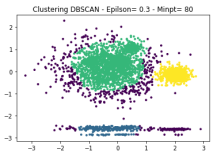
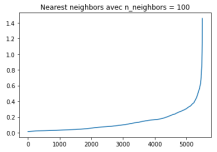
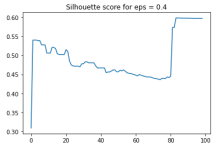
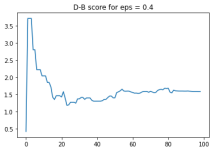
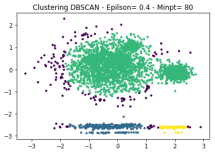
a) Clustering k-Means

Range	Méthode du coude	Silhouette score	Davis-Bouldin score	Temps de calcul	Après K-Means clustering
[2;6]		 2/3 clusters	 3 clusters		

b) Clustering agglomératif

En annexe, vous pouvez voir les tests que nous avons effectués en faisant varier le linkage et le nombre de clusters. Nous avons remarqué que ‘ward’ avec n_clusters = 3 donnait des résultats se rapprochant de ce que nous avons obtenu avec la méthode de clustering K-Means.

c) Clustering DBSCAN

n_cluster = 3				
n_cluster = 3				

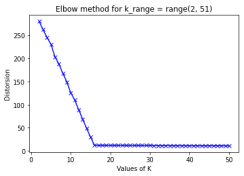
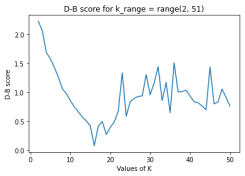
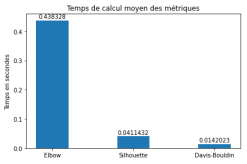
d) Comparaison des méthodes

Comme pour N1, même si les trois méthodes nous donnent bien le même nombre de clusters (3), nous observons visuellement que la méthode agglomérative donne de meilleurs résultats.

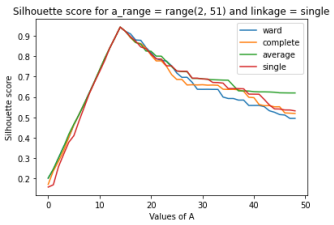
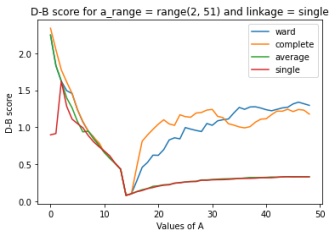
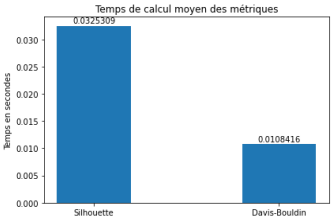
3. D32

Ce set possède 32 dimensions. Il est donc impossible de le représenter en 2D. Nous devons donc nous baser sur des métriques pour l'analyser.

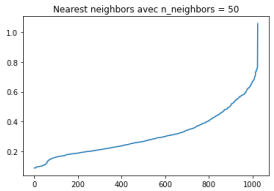
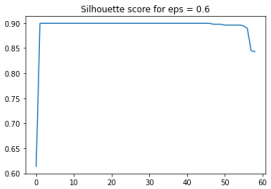
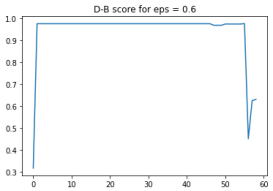
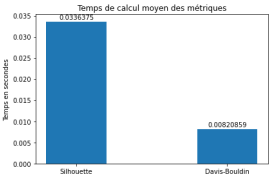
a) Clustering k-Means

Range	Méthode du coude	Silhouette score	Davis-Bouldin score	Temps de calcul
[2;50]				

b) Clustering agglomératif

Range	Silhouette score	Davis-Bouldin score	Temps de calcul
[2;50]			

c) Clustering DBSCAN

Nearest neighbors (=50)	Silhouette score (epsilon = 0.6)	Davis-Bouldin score (epsilon = 0.6)	Temps de calcul
			
`Estimated number of clusters: 16` `Estimated number of noise points: 28` avec eps = 0.6 et min_sample = 55			

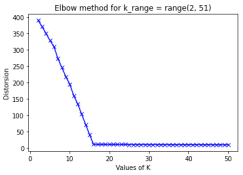
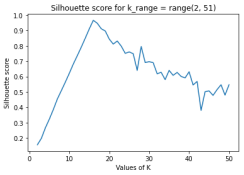
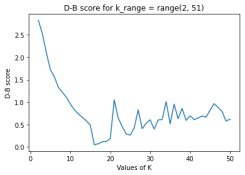
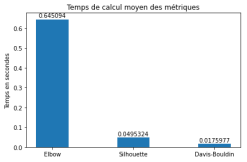
d) Comparaison des méthodes

Les trois métriques donnent le même ordre de grandeur de clusters (autour de 15).

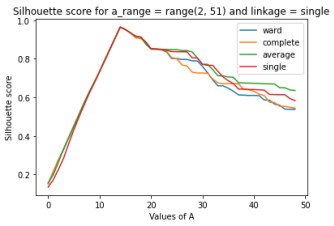
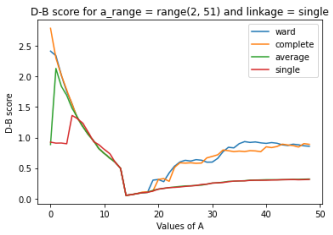
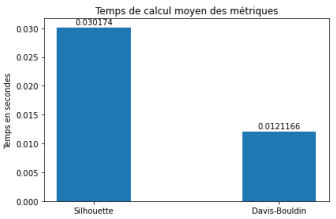
4. D64

Ce set possède 64 dimensions. Il est donc impossible de le représenter en 2D.

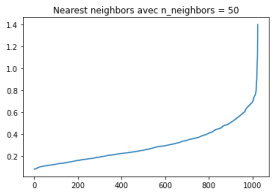
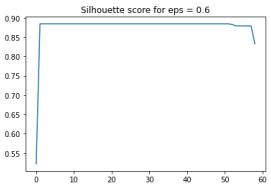
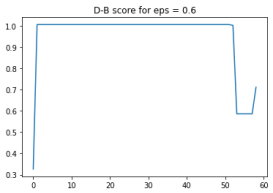
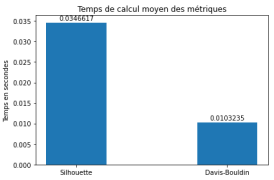
a) Clustering k-Means

Range	Méthode du coude	Silhouette score	Davis-Bouldin score	Temps de calcul
[2;50]				

b) Clustering agglomératif

Range	Silhouette score	Davis-Bouldin score	Temps de calcul
[2;50]			

c) Clustering DBSCAN

Nearest neighbors (=50)	Silhouette score (epsilon = 0.6)	Davis-Bouldin score (epsilon = 0.6)	Temps de calcul
			

```
Estimated number of clusters: 15
Estimated number of noise points: 99
```

avec eps = 0.6 and min_sample = 55

d) Comparaison des méthodes

Les trois métriques donnent le même ordre de grandeur de clusters (autour de 15).

5. W2

Vu l'énorme quantité de données de ce set, nous sommes obligées de l'échantillonner.

a) Echantillonnage

Nous avons effectué plusieurs échantillonnages que vous pouvez voir en annexe :

1. un avec un set de 5558 données
2. un avec un set de 2112 données
3. un avec un set de 1056 données
4. et un avec un set de 704 données.

Nous utiliserons dans la suite les échantillons 1, 3 et 4.

b) Clustering k-Means

En annexe, vous pouvez voir la méthode de clustering K-Means sur les trois différents échantillonnages. Grâce à trois différentes métriques (la méthode du coude, la note de silhouette et la note de Davis-Bouldin), nous trouvons que les trois indiquent la présence de 4 clusters.

c) Clustering agglomératif

Les tests sur les trois échantillons se trouvent en annexe. Le linkage qui semble donner le même résultat que la méthode K-Means est 'ward'. Ce type de linkage donne 4 clusters.

d) Clustering DBSCAN

De même, en annexe se trouvent les résultats des tests de la méthode DBSCAN sur les trois échantillons. Nous avons remarqué que cette méthode n'est pas concluante sur ce data set car, en effet, les clusters ne sont pas très bien séparés.

e) Comparaison des méthodes

Les méthodes K-Means et agglomérative donnent des résultats équivalents. ?ous constatons que la méthode DBSCAN ne fonctionne pas sur ce data set.

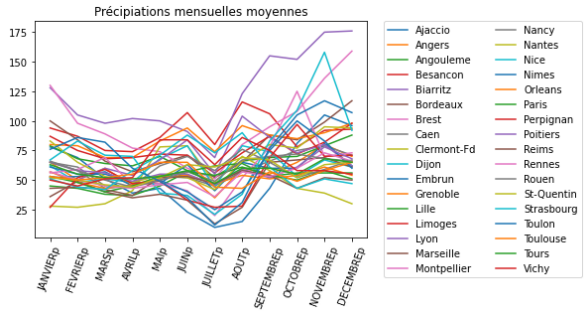
B. Données réelles

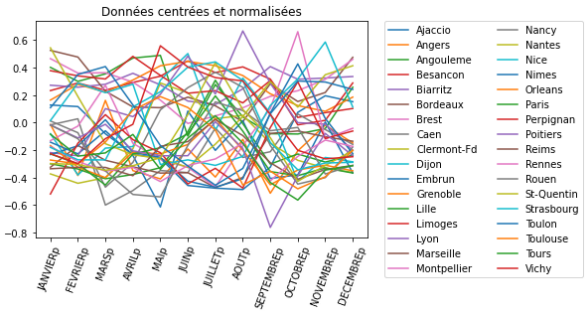
1. Analyse en composantes principales

Nous avons découpé les données réelles en deux sets : précipitations et nombre de jours de pluie.

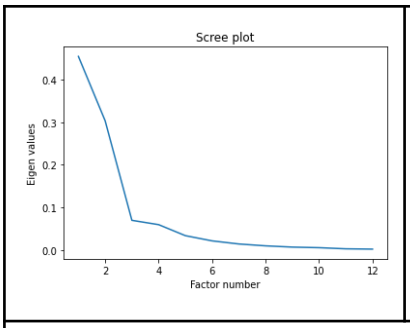
a) Données sur les précipitations

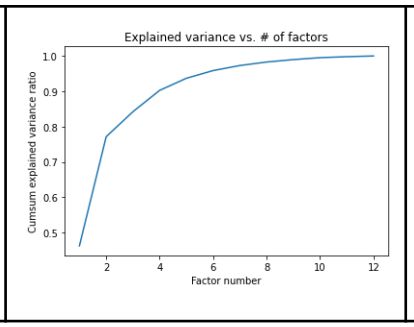
Nous avons d'abord représenté les données telles qu'elles sont fournies. Ensuite, nous les avons centrées et normalisées.

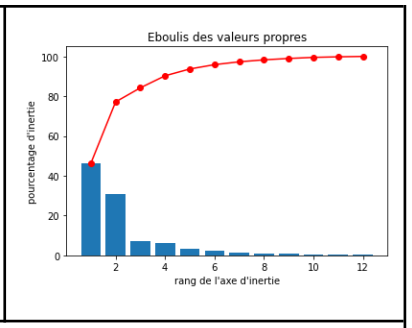




Enfin, nous avons réduit les axes grâce à une analyse en composantes principales.







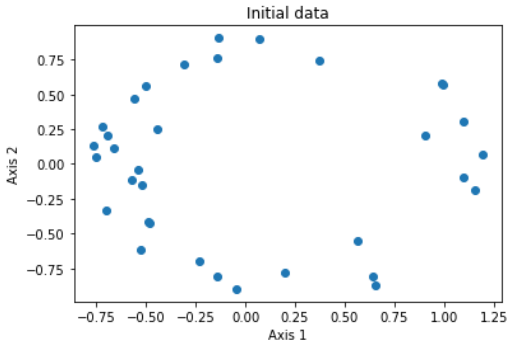

```
ACP : recherche d'axe de projection
[46.25371026 30.87075887  7.08536531  6.05313073  3.42264137  2.17598589
  1.43325074  0.98804254  0.68728739  0.548556    0.27770547  0.20356541]
```

Axes avec leur % d'inertie

En faisant le total d'inertie des 7 premiers axes, nous obtenons 97.26% d'inertie (et avec le 8e, nous avons un total de 98.25% d'inertie). Cependant, en regardant le premier graphe, nous voyons un coude au 3e axe. Il est donc possible de garder seulement les 3 premiers. Nous avons donc 84.21% d'inertie, ce qui semble suffisant.

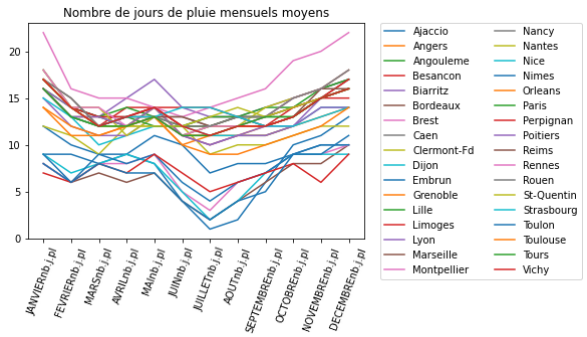
Pour être sûres de nous, nous avons effectué le test des bâtons brisés. Il montre que seulement les deux premiers axes sont valides. Nous ne pouvons donc garder que ceux-ci pour avoir une représentation dans le plan. (total d'inertie de 77.12%)

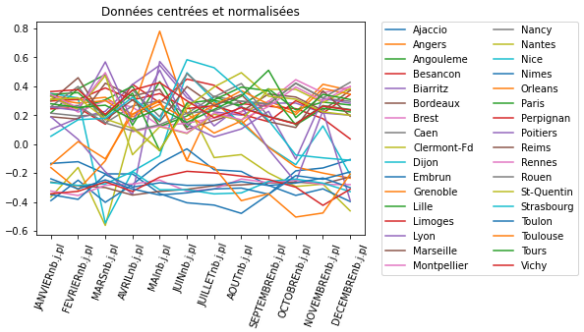
	Val.Propre	Seuils
0	0.454422	3.103211
1	0.303292	2.103211
2	0.069611	1.603211
3	0.059469	1.269877
4	0.033626	1.019877
5	0.021378	0.819877
6	0.014081	0.653211
7	0.009707	0.510354
8	0.006752	0.385354
9	0.005389	0.274242
10	0.002728	0.174242
11	0.002000	0.083333



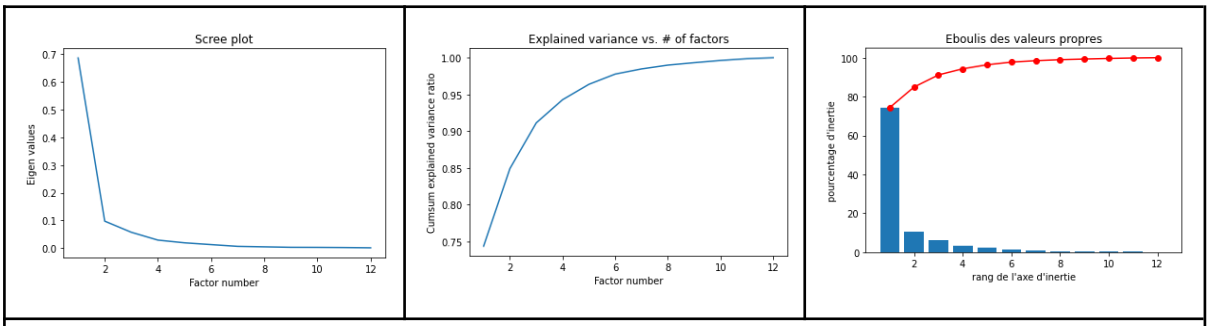
b) Données sur le nombre de jours de pluie

Nous avons d'abord représenté les données telles qu'elles sont fournies. Ensuite, nous les avons centrées et normalisées.





Enfin, nous avons réduit les axes grâce à une analyse en composantes principales.



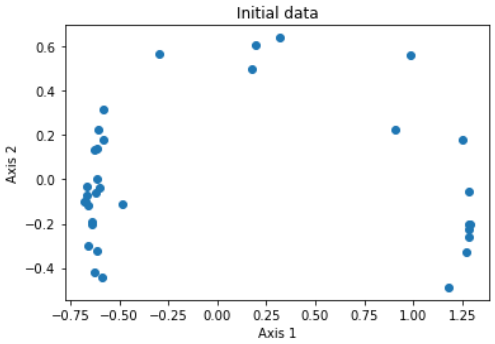

```
ACP : recherche d'axe de projection
[74.35650817 10.56335883  6.21691833  3.15176389  2.09390442  1.3945993
  0.70384472  0.51610006  0.32724072  0.30170474  0.23911081  0.13494602]
```

Axes avec leur % d'inertie

En faisant le total d'inertie des 6 premiers axes, nous obtenons 97.77% d'inertie (et avec le 7e, nous avons un total de 98.47% d'inertie). Ici, nous voyons un coude au 2e axe. Il est donc possible de garder seulement les 2 premiers. Nous avons donc 84.92% d'inertie, ce qui semble suffisant.

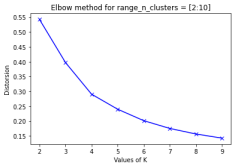
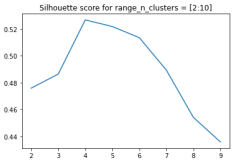
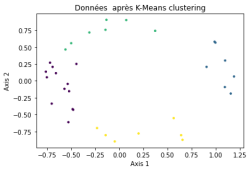
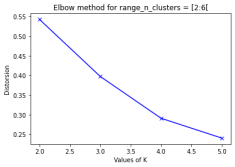
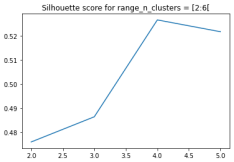
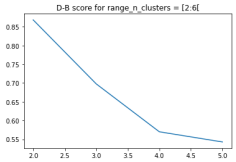
Pour être sûres de nous, nous avons effectué le test des bâtons brisés. Il confirme que seulement les deux premiers axes sont valides.

	Val.Propre	Seuils
0	0.687007	3.103211
1	0.097599	2.103211
2	0.057440	1.603211
3	0.029120	1.269877
4	0.019346	1.019877
5	0.012885	0.819877
6	0.006503	0.653211
7	0.004768	0.510354
8	0.003023	0.385354
9	0.002788	0.274242
10	0.002209	0.174242
11	0.001247	0.083333

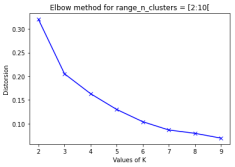
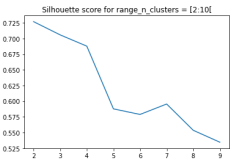
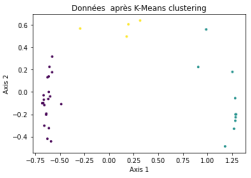
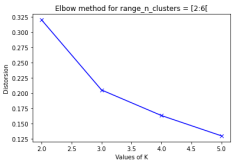
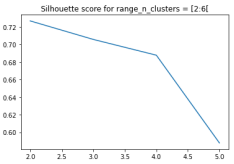
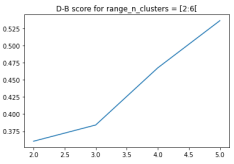


2. Clustering k-Means

a) Données sur les précipitations

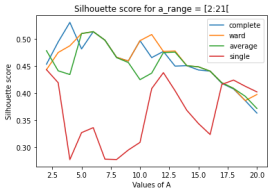
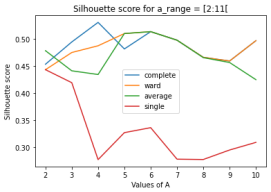
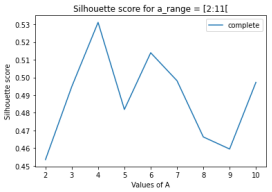
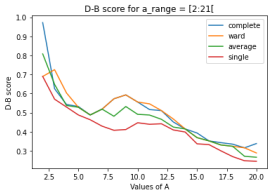
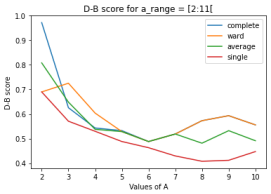
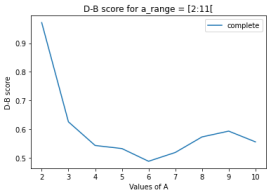
Range	Méthode du coude	Silhouette score	Davis-Bouldin score	
[2;9]				Nous remarquons qu'il y a sûrement 4 clusters. 
[2;5]				

b) Données sur le nombre de jours de pluie

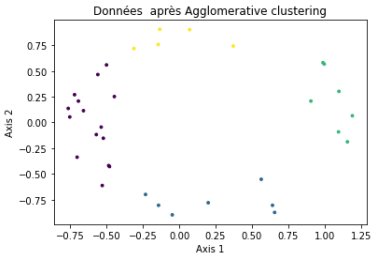
Range	Méthode du coude	Silhouette score	Davis-Bouldin score	
[2;9]				Nous remarquons qu'il doit y avoir 3 clusters. 
[2;5]				

3. Clustering agglomératif

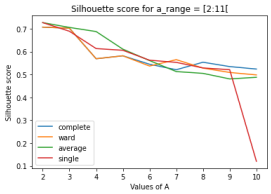
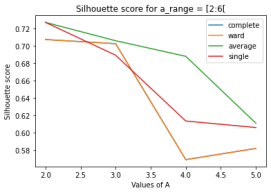
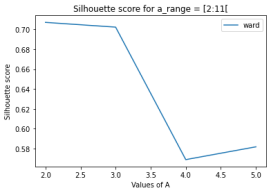
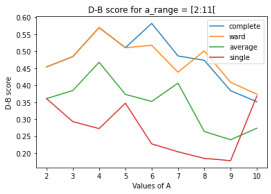
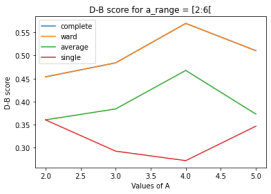
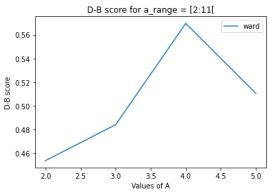
a) Données sur les précipitations

Range	[2;20]	[2;10]	[2;10]
Silhouette score			
Davis-Bouldin score			

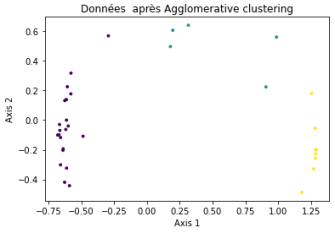
Le linkage ‘complete’ semble donner le résultat le plus proche de celui trouvé avec la méthode K-Means.



b) Données sur le nombre de jours de pluie

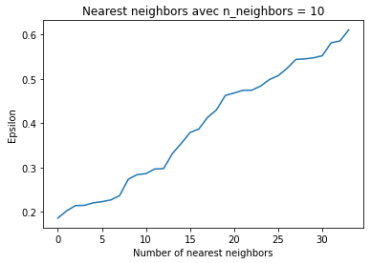
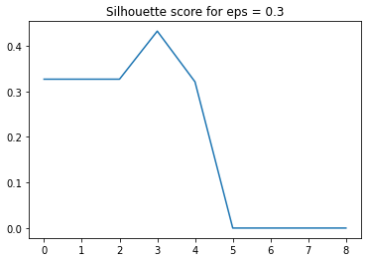
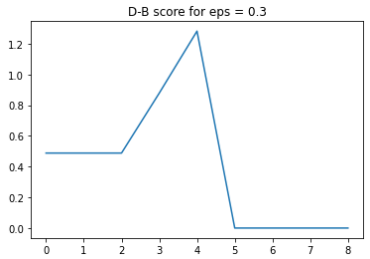
Range	[2;10]	[2;5]	[2;10]
Silhouette score			
Davis-Bouldin score			

Le linkage ‘ward’ semble donner le résultat le plus proche de celui trouvé avec la méthode K-Means.



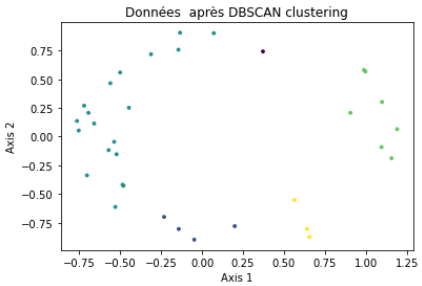
4. Clustering DBSCAN

a) Données sur les précipitations

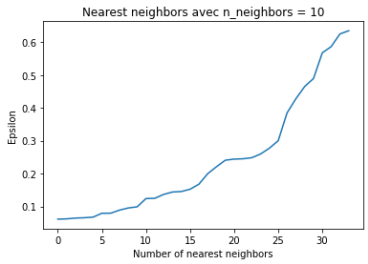
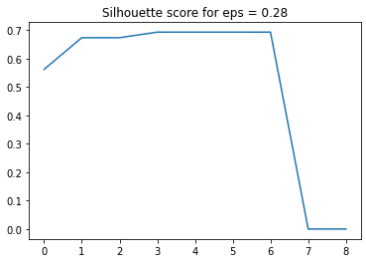
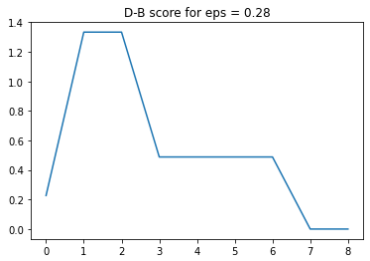
Nearest neighbors (= 10)	Silhouette score (epsilon = 0.3)	Davis-Bouldin score (epsilon = 0.3)
		

Nous voyons ici que eps = 0.3 et que n_sample = 2 ou 3.

Pour n_samples = 2 ou 3, nous trouvons le même résultat :

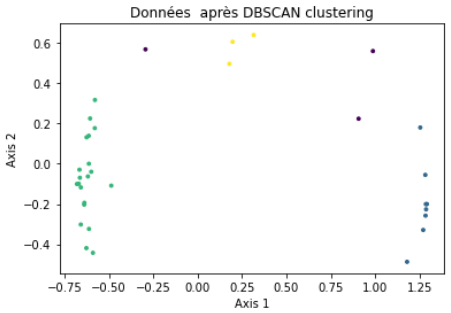


b) Données sur le nombre de jours de pluie

Nearest neighbors (= 10)	Silhouette score (epsilon = 0.28)	Davis-Bouldin score (epsilon = 0.28)
		

Nous voyons ici que eps = 0.28 et que n_sample = [3,5].

Pour n_samples = 3, nous trouvons le meilleur résultat :



5. Analyses des résultats

a) Données sur les précipitations

En annexe, vous pouvez voir deux cercles des corrélations ainsi que le nuage des données. Sur le premier cercle des correlation, nous voyons que :

- F1 est corrélé avec : Amplitude annuelle des températures
- F2 est corrélé avec : Latitude, insolation annuelle

Sur le deuxième cercle des correlation, nous constatons que tous les attributs sont corrélés positivement et nous pouvons interpréter :

- F1, amplitude des précipitations, opposition été/hiver
- F2, la plupart des villes ont des précipitations faibles ou fortes toute l'année, précipitations moyennes

Sur le nuage des exemples, nous pouvons faire les analyses suivantes :

- Il y a des villes où les précipitations sont fortes toute l'année (Biarritz, Bordeaux, Brest…)
- Il y a des villes où les précipitations sont faibles toute l'année (Reims, Orléans…)
- Il y a des villes où les précipitations sont fortes en hiver (Grenoble, Besançon…)
- Il y a des villes où les précipitations sont faibles en hiver (Clermont-Ferrand…)
- Il y a des villes où les précipitations sont faibles en été (Marseille…)

b) Données sur le nombre de jours de pluie

De même, en annexe, vous pouvez voir deux cercles des corrélations ainsi que le nuage des données. Sur le premier cercle des correlation, nous voyons que :

- F1 est corrélé avec : Amplitude annuelle des températures
- F2 est corrélé avec : Nombre annuel de jours de pluie, insolation annuelle

Sur le deuxième cercle des correlation, nous constatons que tous les attributs sont corrélés positivement et nous pouvons interpréter :

- F1, amplitude du nombre de jours de pluie, opposition été/hiver
- F2, la plupart des villes ont un nombre de jours de pluie élevé ou bas toute l'année, nombre de jours de pluie moyens

Sur le nuage des exemples, nous pouvons faire les analyses suivantes :

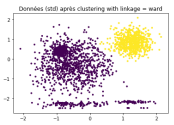
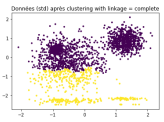
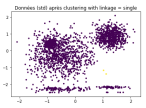
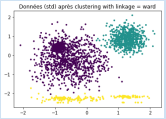
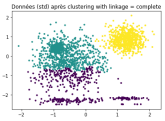
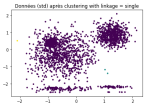
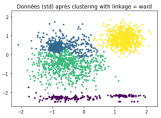
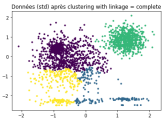
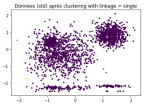
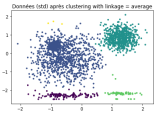
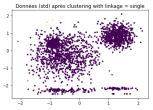
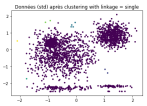
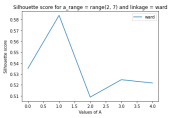
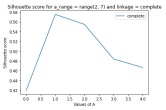
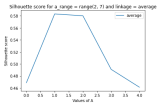
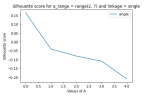
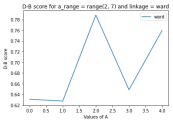
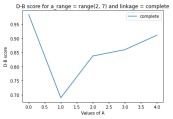
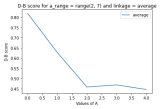
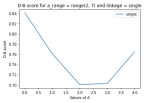
- Il y a des villes où le nombre de jours de pluie est élevé toute l'année (Perpignan…)
- Il y a des villes où le nombre de jours de pluie est bas toute l'année (Paris, Caen…)
- Il y a des villes où le nombre de jours de pluie est élevé en hiver (Clermont-Ferrand…)
- Il y a des villes où le nombre de jours de pluie est bas en hiver (Ajaccio…)
- Il y a des villes où le nombre de jours de pluie est bas en été (Nantes, Rennes…)

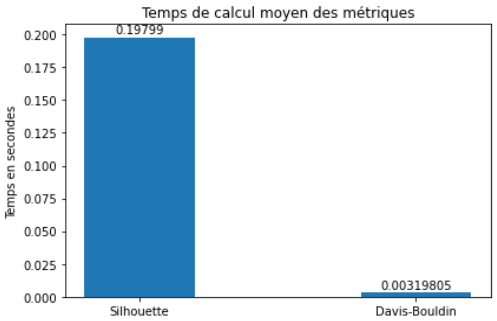
IV. Conclusion

Ce TP nous a permis de découvrir quelques méthodes de clustering et de les analyser pour voir leurs forces et leurs faiblesses. Nous avons pu mettre en oeuvre nos nouvelles connaissances en les testant sur des données mystères et même sur des données réelles.

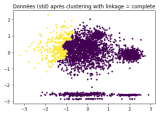
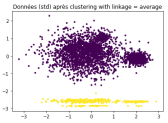
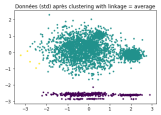
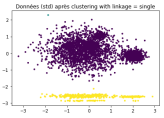
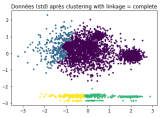
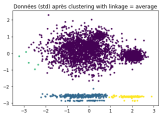
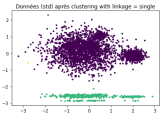
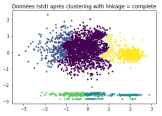
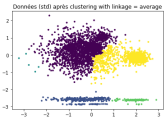
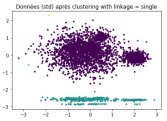
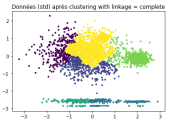
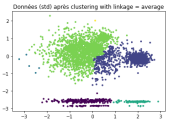
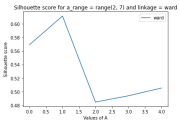
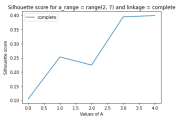
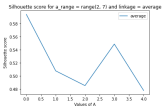
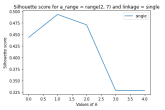
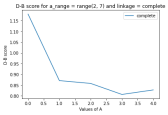
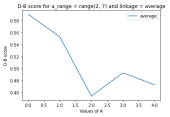
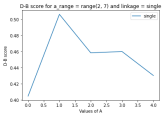
V. Annexes

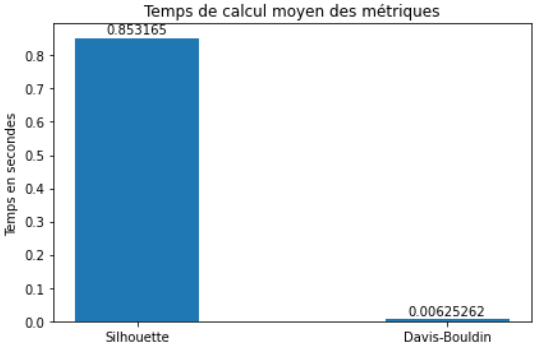
A. N1, clustering agglomératif

	‘ward’	‘complete’	‘average’	‘single’
n_cluster = 2				
n_cluster = 3				
n_cluster = 4				
n_cluster = 5				
n_cluster = 6				
Silhouette score				
Davis-Bouldin score				

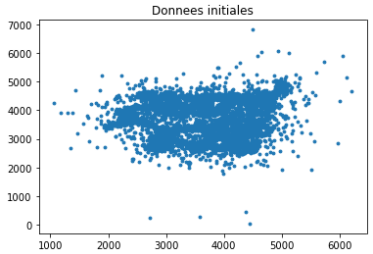
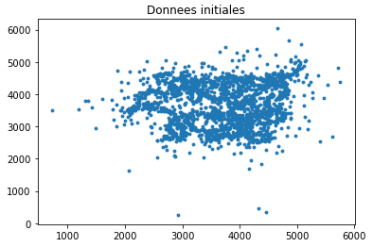
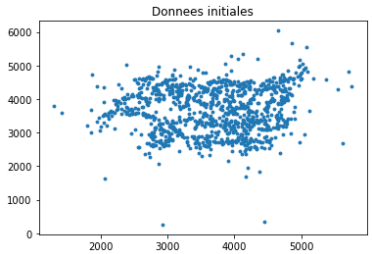
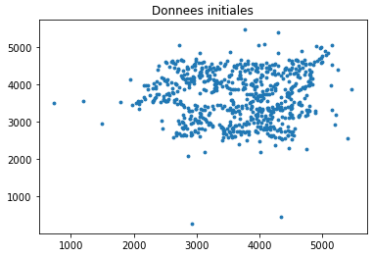


B. N2, clustering agglomératif

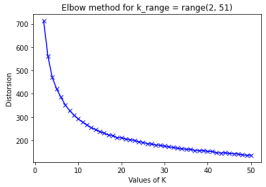
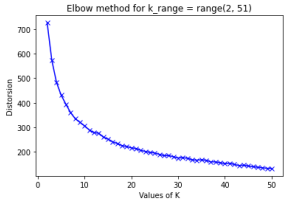
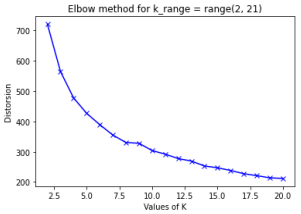
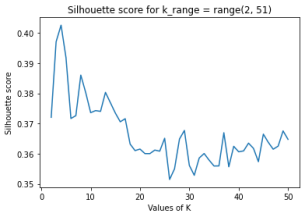
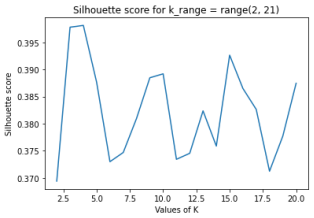
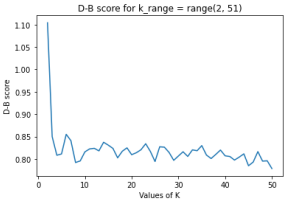
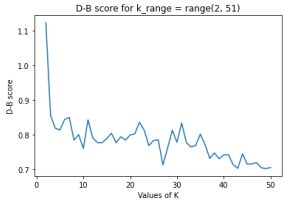
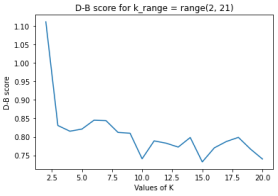
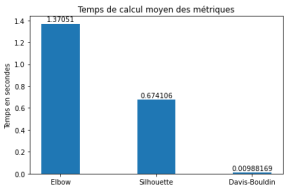
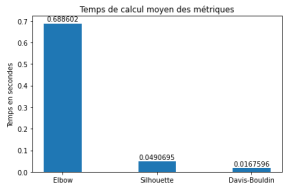
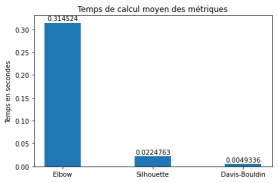
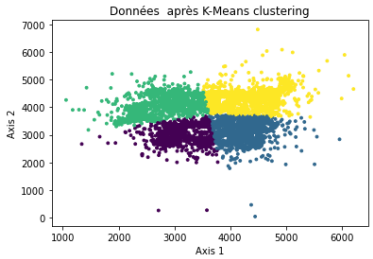
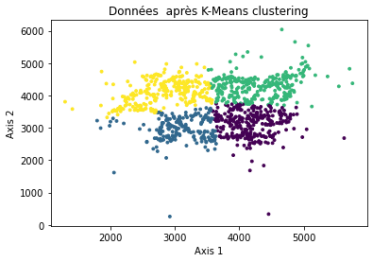
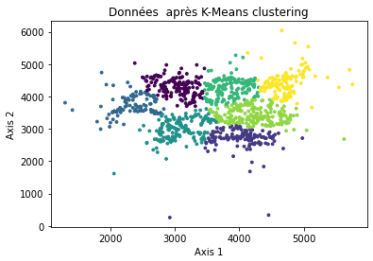
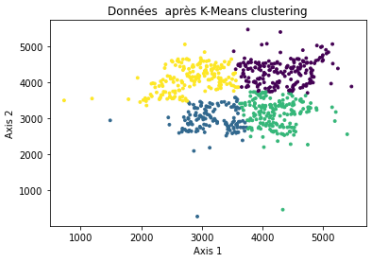
	‘ward’	‘complete’	‘average’	‘single’
n_cluster = 2				
n_cluster = 3				
n_cluster = 4				
n_cluster = 5				
n_cluster = 6				
Silhouette score				
Davis-Bouldin score				



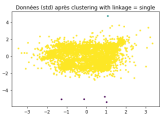
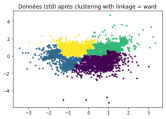
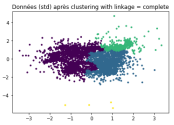
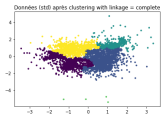
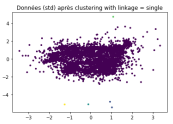
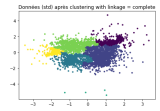
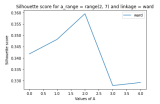
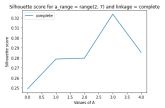
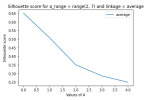
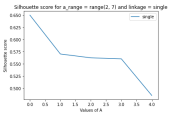
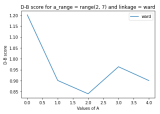
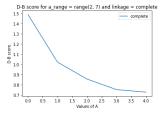
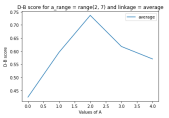
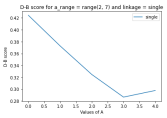
C. W2, échantillonnage

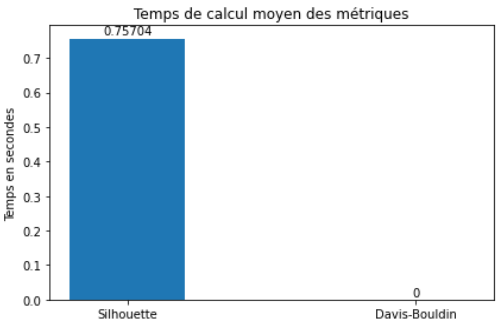
Nombre de données	Représentation	Code
5558		datanp = arr[1::19,:]
2112		datanp = arr[1::50,:]
1056		datanp = arr[1::100,:]
704		datanp = arr[1::150,:]

D. W2, clustering K-Means

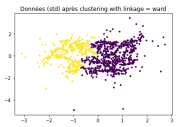
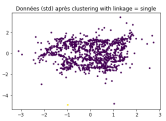
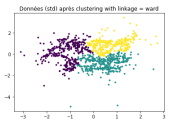
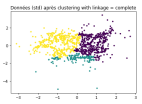
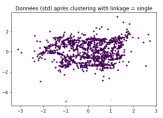
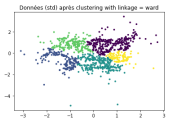
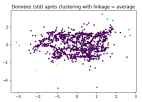
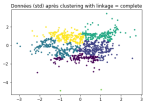
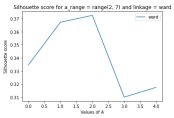
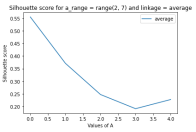
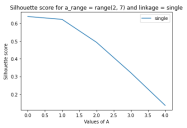
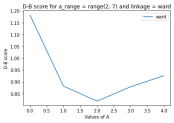
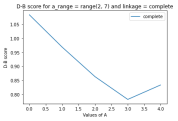
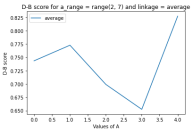
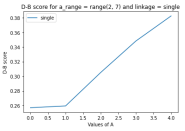
5558	1056	704
		
		
		
		
4 clusters : 	4 clusters :  7 clusters : 	4 clusters : 

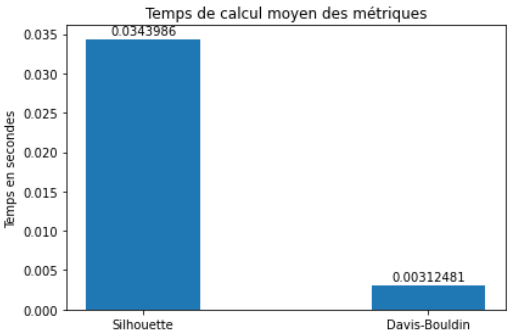
E. W2, clustering agglomératif, 5558 données

5558 données	‘ward’	‘complete’	‘average’	‘single’
n_cluster = 2				
n_cluster = 3				
n_cluster = 4				
n_cluster = 5				
n_cluster = 6				
Silhouette score				
Davis-Bouldin score				

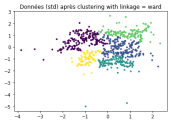
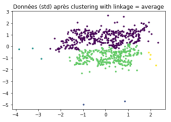
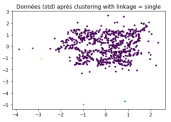
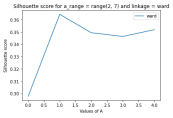
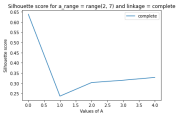
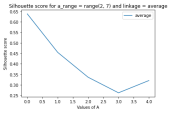
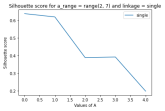
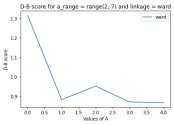
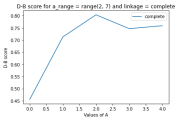
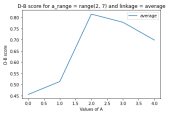
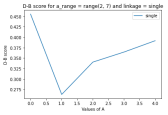


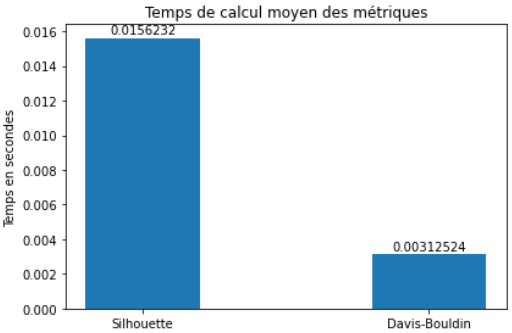
F. W2, clustering agglomératif, 1056 données

1056 données	‘ward’	‘complete’	‘average’	‘single’
n_cluster = 2				
n_cluster = 3				
n_cluster = 4				
n_cluster = 5				
n_cluster = 6				
Silhouette score				
Davis-Bouldin score				

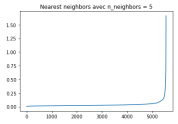
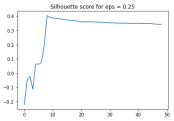
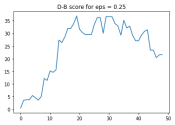
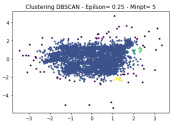
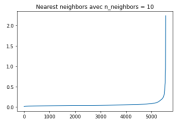
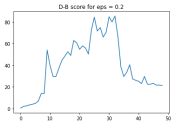
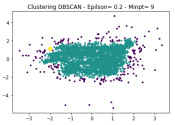
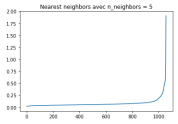
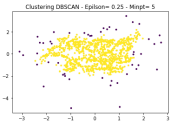
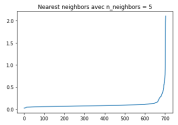
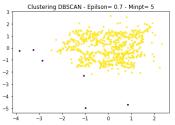


G. W2, clustering agglomératif, 704 données

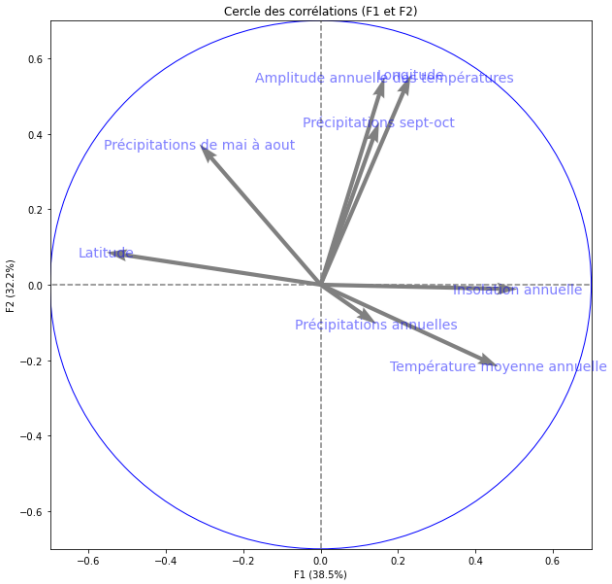
704 données	‘ward’	‘complete’	‘average’	‘single’
n_cluster = 2				
n_cluster = 3				
n_cluster = 4				
n_cluster = 5				
n_cluster = 6				
Silhouette score				
Davis-Bouldin score				

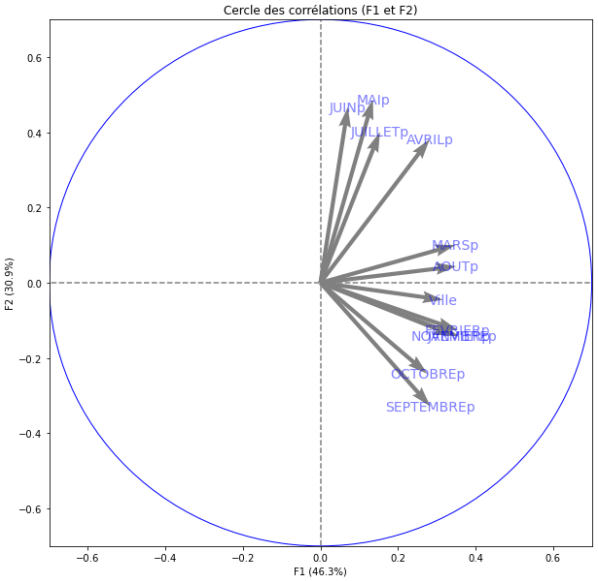


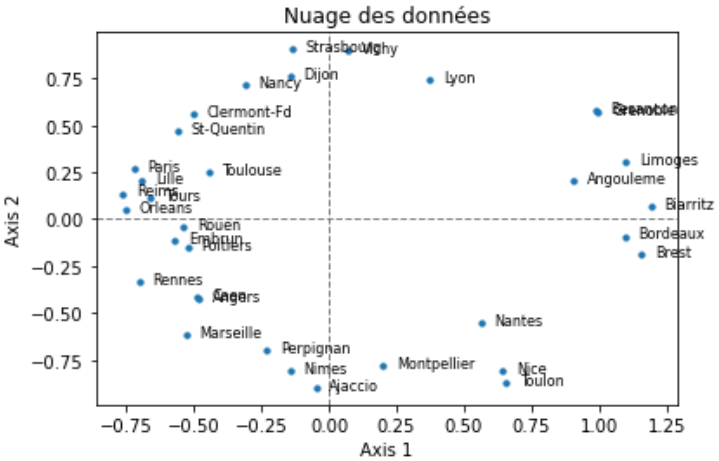
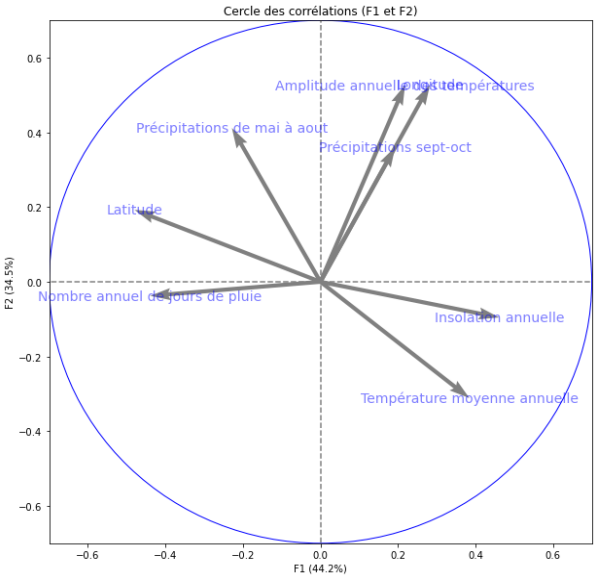
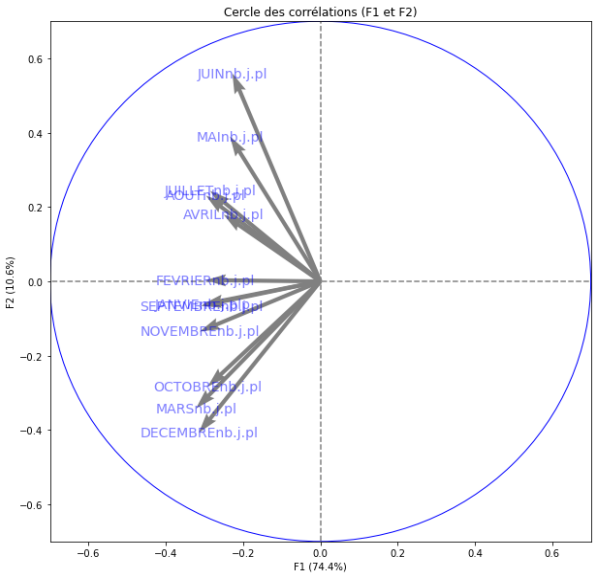
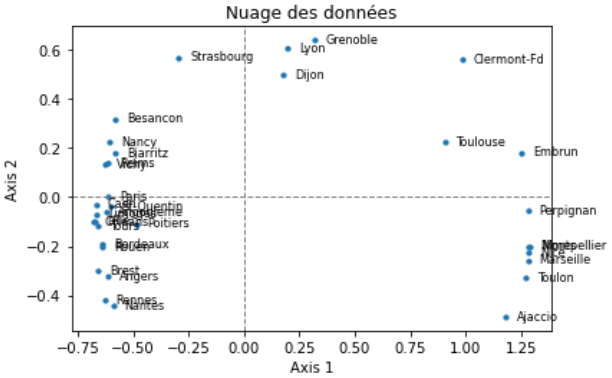
H. W2, clustering DBSCAN

5558					
					
1056		Pas de métriques car il n'y a qu'un cluster			
704					

I. Précipitations, analyses







J. Nombre de jours de pluie, analyses