



Projet SD701 : Exploration de grands volumes de données

Étude de la pollution en France

Guillaume CANAT & Jean-Bastien SARDA
MS BGD 2021-2022

Repository du projet :
<https://github.com/jbSarda/SD701>

12 décembre 2021

Table des matières

1	Introduction	1
2	Recueil, nettoyage et compilation des données	1
2.1	Données sur les sites de mesures	1
2.2	Mesures de pollution	2
2.3	Données supplémentaires	3
3	Visualisation brute des données	5
3.1	Visualisation sur carte	5
3.2	Dataset compilé	6
4	Clustering	6
4.1	K-means	6
4.2	Clustering hiérarchique (BIRCH)	9
4.3	DBScan	10
4.4	Clustering : conclusions	10
5	Prédictions sur 2021	11
6	Conclusions	13

1 Introduction

La disponibilité sur le site data.gouv.fr de mesures par heure de différents types de polluants dans l'air sur 784 sites en France métropolitaine depuis 2018 permet d'obtenir un dataset final conséquent et d'envisager des études d'exploration de données.

Les objectifs fixés dans cette étude s'établissent comme suit :

- faire apparaître visuellement les cumuls ou moyennes de pollution sur tous les sites sur une période donnée,
- mesurer la croissance ou décroissance de pollution dans le temps dans des époques particulières (confinements, déconfinements, installations d'usines),
- par clustering, déterminer des nappes de pollution sur une période donnée (KMeans, BIRCH...)
- par apprentissage, tenter de déterminer les évolutions à venir de la pollution et comparaison avec les valeurs mesurées.

2 Recueil, nettoyage et compilation des données

2.1 Données sur les sites de mesures

Les données géographiques concernant les sites de mesures sont compilées dans le fichier `fr-2020-d-lcsqa-ineris-20210412.xml`. Les paramètres qui s'y trouvent concernent le code du site, son nom, son altitude et ses coordonnées géographiques en format latitude-longitude. Ces données ne sont pas suffisantes pour caractériser un site.

Il est souhaitable également de posséder le ou les polygones du département d'appartenance pour une meilleure visualisation des données sur carte.

Ainsi, à partir d'un fichier des communes françaises, nous avons procédé à un calcul de la distance la plus faible du site de mesure à une commune pour identifier le code postal du site et ainsi obtenir le polygone adapté (disponible sur le site [france-geojson](https://france-geojson.fr)).

La densité démographique a également été répertoriée pour chaque site de mesure à partir du fichier des communes françaises. Une visualisation du résultat est proposée en Figure 1.

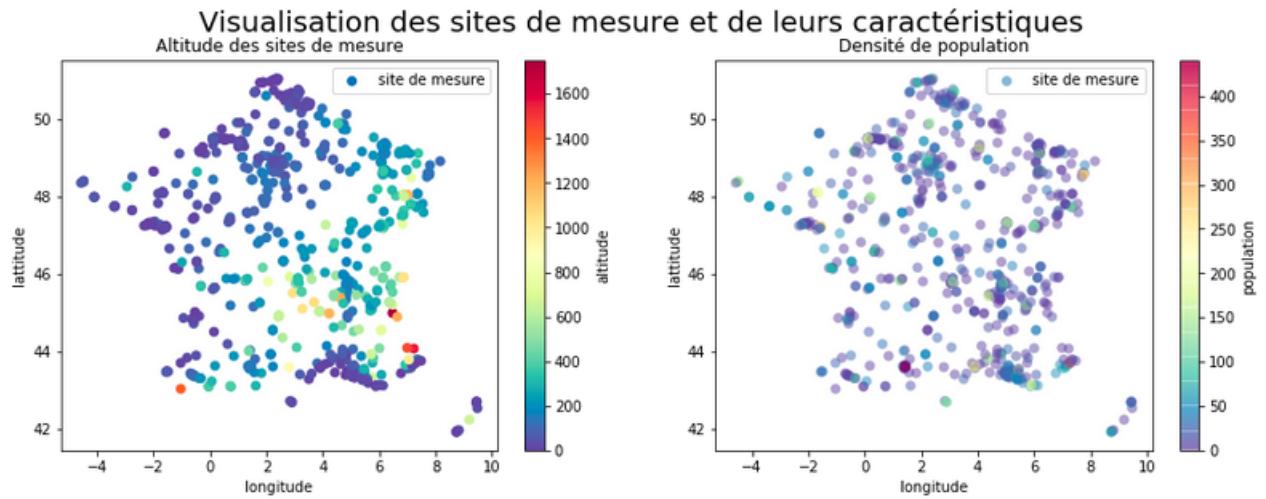


FIGURE 1 – Localisation des sites de mesures, distingués par altitude puis population

2.2 Mesures de pollution

Source des données : [concentration de polluants atmosphériques sur data.gouv.fr](#)

Les données de pollution recueillies s'échelonnent sur une plage temporelle allant de septembre 2018 à mars 2021. Les données ont été collectées à partir de fichiers .xml qui sont générés plusieurs fois par jour et sont déposés sur le site. La récupération et le nettoyage ont demandé un travail conséquent dans les phases de parsing. Un script a été créé et lancé sur les machines de tp de l'école pour récupérer de manière fractionnée un dataset par année avant la fusion finale (la concaténation des données dans un seul tableau faisait crasher les machines).

Obtenant environ 45 000 lignes par jour (environ une mesure par site et par type de polluant), le dataframe complet est monté à environ 50 millions de lignes (30 Go de données) nous avons opté pour une mesure journalière par site.

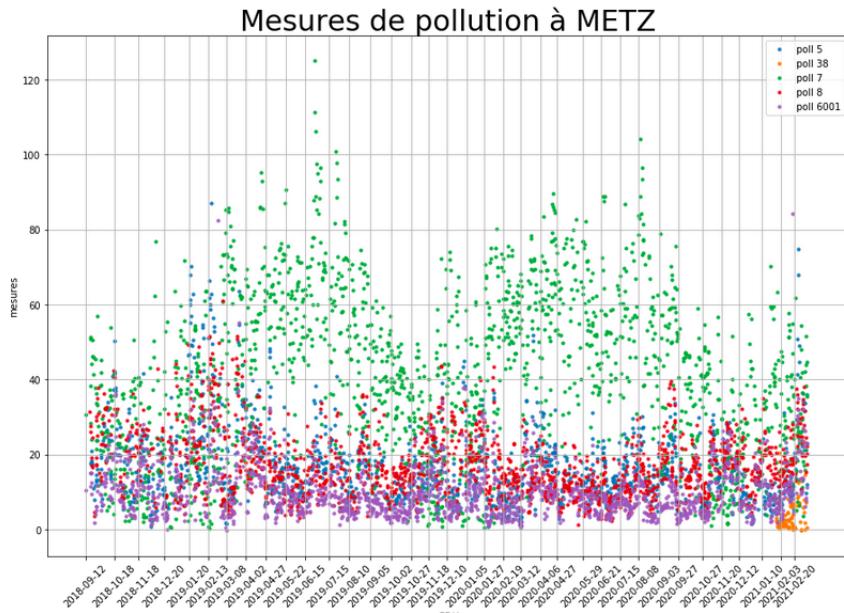


FIGURE 2 – Représentation des mesures de pollution à Metz, pour différents types de polluants

Pour un site pris au hasard (Figure 2), on constate que tous les polluants ne sont pas mesurés et certains ne sont mesurés qu'à partir de certaines périodes. Ce constat nous pousse alors à standardiser les mesures et les regrouper sur une valeur moyenne. Ce choix assumé peut comporter un biais : il est possible par exemple qu'un polluant soit plus concentré dans certaines zones (urbain VS rural) par rapport aux autres, et ainsi faire augmenter significativement la moyenne de pollution, engendrant alors des disparités ou à l'inverse lissant les disparités.

Nous ferons par la suite l'hypothèse que ce biais affecte peu le modèle.

2.3 Données supplémentaires

Pour pouvoir entraîner convenablement notre modèle, nous devons trouver des “features” permettant d’expliquer davantage l’évolution de la pollution sur les sites de mesure, outre l’aspect géographique. Ainsi, deux types de données supplémentaires ont été intégrés :

1. Les données météorologiques adaptées à la granularité temporelle définie (données journalières) et aux coordonnées géographiques des sites. Deux méthodes ont été testées, à savoir le parsing des sites météorologiques (difficulté d’obtenir des données propres à chaque site) et la librairie meteostat. Cette dernière offre avantageusement la possibilité d’obtenir la météo journalière en prenant en paramètre la date et les coordonnées géographiques. La fonction meteostat.daily retourne alors un DataFrame tel que présenté en figure 3.

```

1 Daily(Point(random.uniform(41,51), random.uniform(-5,10) , random.uniform(0, 1000)),
2      datetime.datetime(2021, 1,1),
3      datetime.datetime(2021,1,5)).fetch()

tavg tmin tmax prcp snow wdir wspd wpgt pres tsun
time
2021-01-02 -5.5 -9.1 -3.2 0.0 NaN 292.0 5.3 NaN 1014.1 NaN
2021-01-03 -4.6 -6.3 -4.0 0.9 NaN 32.2 14.0 NaN 1014.3 NaN
2021-01-04 -4.6 -5.7 -3.8 1.0 NaN 51.8 11.6 NaN 1015.2 NaN
2021-01-05 -4.6 -5.4 -4.0 0.1 NaN 42.1 14.0 NaN 1015.8 NaN

# légende Metéostat
pd.read_html('https://dev.meteostat.net/python/daily.html#api',header=0,
keep_default_na=False)[0]

Column Description Type
0 station The Meteostat ID of the weather station (only ... String
1 time The date Datetime64
2 tavg The average air temperature in °C Float64
3 tmin The minimum air temperature in °C Float64
4 tmax The maximum air temperature in °C Float64
5 prcp The daily precipitation total in mm Float64
6 snow The snow depth in mm Float64
7 wdir The average wind direction in degrees (°) Float64
8 wspd The average wind speed in km/h Float64
9 wpgt The peak wind gust in km/h Float64
10 pres The average sea-level air pressure in hPa Float64
11 tsun The daily sunshine total in minutes (m) Float64

```

FIGURE 3 – Informations retournées par l’API meteostat

2. Les données de consommation énergétique ont été choisies pour mesurer l’activité humaine. Toutefois, nous n’avons pas trouver de jeu de données descendant à un niveau de granularité spatio-temporel de l’ordre du département/jour. Nous avons alors intégré la mesure du pic journalier de consommation énergétique calculé sur l’ensemble du territoire national (Figure 4).

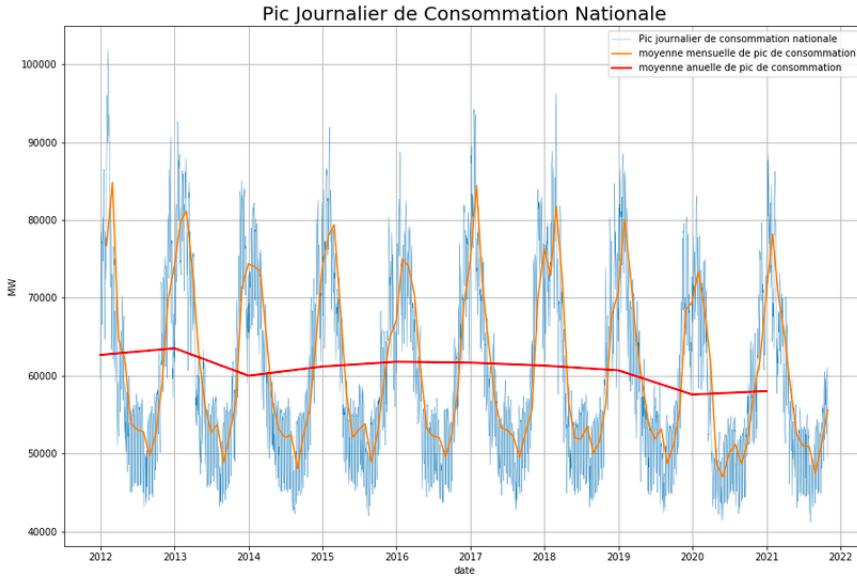


FIGURE 4 – Evolution du pic de consommation électrique journalier de 2012 à 2022

Les jours de la semaine ont également été intégrés au dataset complet sous forme de coordonnées de point ($\cos(2i\pi/7), \sin(2i\pi/7)$) pour i allant de 0 à 7. Cette représentation du jour permet d’intégrer la métrique au calcul du cluster.

3 Visualisation brute des données

3.1 Visualisation sur carte

Une fonction de visualisation automatique, prenant en entrée une date de début, une date de fin, les codes des départements concernés et optionnellement le label des sites de mesure permet de visualiser rapidement les moyennes de pollution en contrôlant le cadre spatio-temporel.

Les représentations à l'échelle nationale sur des périodes allant du mois à l'année (Figure 5) font apparaître que les mesures les plus élevées se concentrent dans et autour des zones urbaines. D'autre part, on constate la présence de la fameuse “diagonale du vide” allant du Sud-Ouest au Nord Est (trainée bleue), probablement plus apparente lors du clustering.

Fonction : `graph_poll(date_start, date_end, lis_dep, lab = False)`

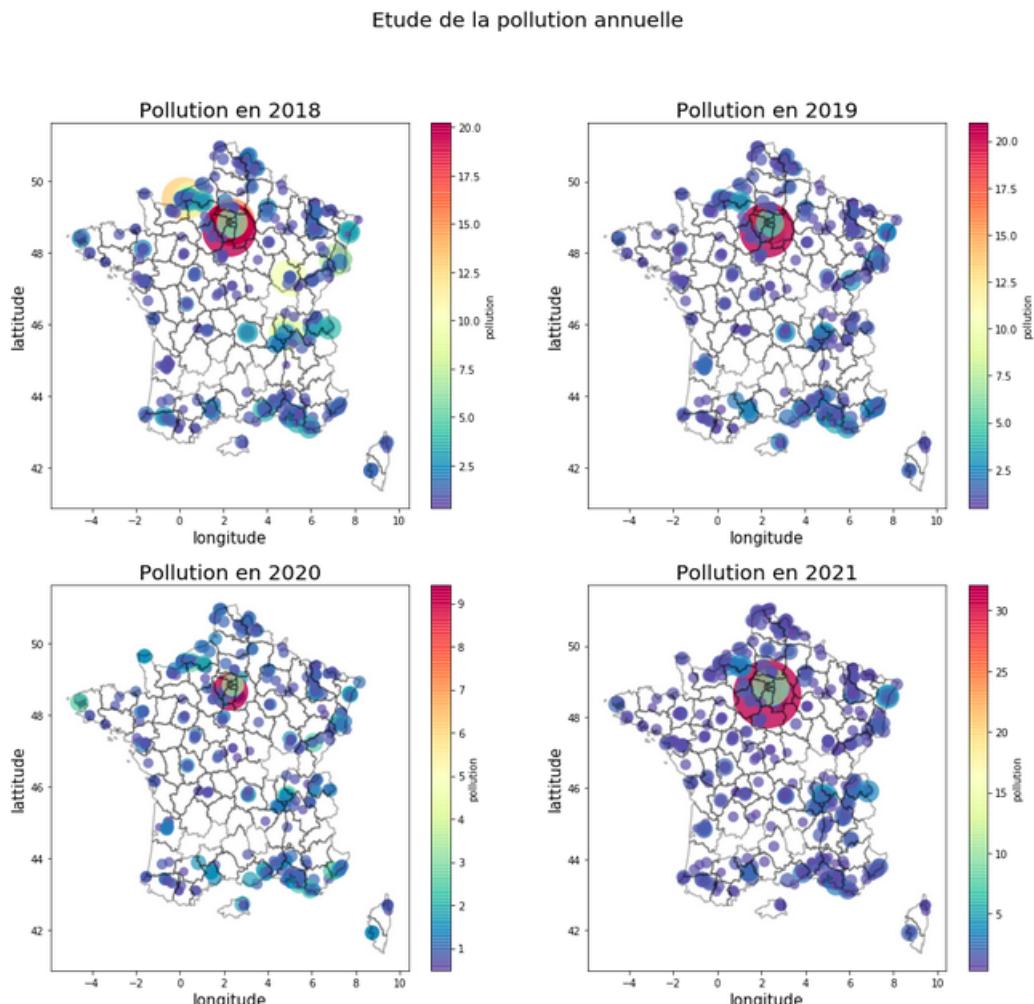


FIGURE 5 – Représentation de la pollution annuelle moyenne sur les sites de mesure en France par année

On peut prendre comme exemple l’Île de France, dont l’étude de la pollution moyenne révèle une concentration des valeurs les plus hautes à Paris, et une dispersion de la pollution à mesure que l’on s’éloigne de la capitale.

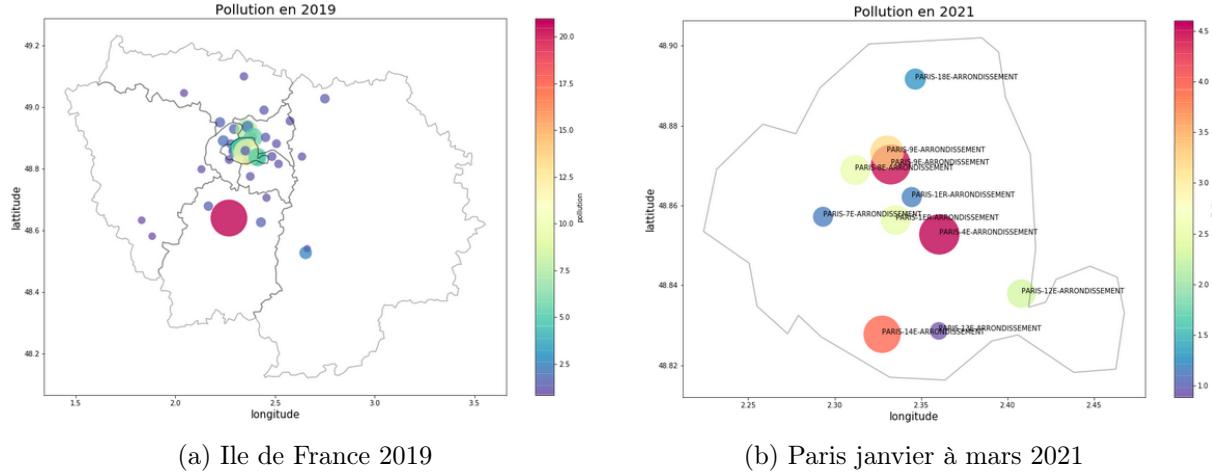


FIGURE 6 – Etude de la pollution moyenne en Ile de France

En faisant un focus sur Paris, on constate que quelle que soit la période de temps données, l’air de la capitale concentre logiquement bien plus de polluant que dans les campagnes (Figure 6).

3.2 Dataset compilé

Au final, le dataset de travail contient 23 colonnes, 1 695 084 observations enregistrées entre le 23 septembre 2018 et le 06 mars 2021. Un aperçu est présenté en Figure 7.

GDH	code site	PicJourConsoNat	type_poll	verif	valid	mesure	mesure_max	mesure_min	lat	long	alt	Population	Département	Code	tavg	tmin	tmax	prcp	pres	Jx	Jy	lat_cr	long_cr
0	2018-09-23	FR02033	-1.355028	1	3	1	8.821807	3.004590	7.766743	43.339893	5.058827	-0.624859	-0.092434	13	1.890986	1.876290	1.870339	-0.407749	0.135226	0.913773	-1.184771	-1.47736	
1	2018-09-25	FR02033	-0.590960	1	3	1	-0.215870	-0.048974	-0.159242	43.339893	5.058827	-0.624859	-0.092434	13	1.026681	0.877713	1.076525	-0.407749	0.984498	0.913773	1.047821	-1.47736	
2	2018-10-02	FR02033	-0.326680	1	2	1	0.633907	0.917723	-0.159242	43.339893	5.058827	-0.624859	-0.092434	13	0.495967	0.661355	0.487567	-0.407749	0.024931	0.913773	1.047821	-1.47736	
3	2018-10-03	FR02033	-0.377194	1	2	1	-0.426315	-0.345628	-0.159242	43.339893	5.058827	-0.624859	-0.092434	13	0.617273	0.028923	0.986901	-0.407749	0.355816	-0.275024	1.323823	-1.47736	
4	2018-10-04	FR02033	-0.492852	1	3	1	-0.321854	-0.212014	-0.159242	43.339893	5.058827	-0.624859	-0.092434	13	0.890211	0.395068	1.191756	-0.407749	0.433023	-1.228366	0.551022	-1.47736	

FIGURE 7 – Premières lignes du dataset final utilisé pour entraîner les algorithmes

4 Clustering

4.1 K-means

On souhaite entraîner le modèle en utilisant la moyenne des mesures de pollution sur la période complète. Dans un premier temps, les données standardisées sont les données météorologiques, l’altitude, et la densité de population. La consommation énergétique et les coordonnées des jours de la semaine ne sont pas des métriques spécifiques aux sites mais dépendent uniquement de la date ; elles ne sont donc pas prises en compte

dans le clustering. Le graphe obtenu sur la France entière révèle un découpage par région (Figure 8).

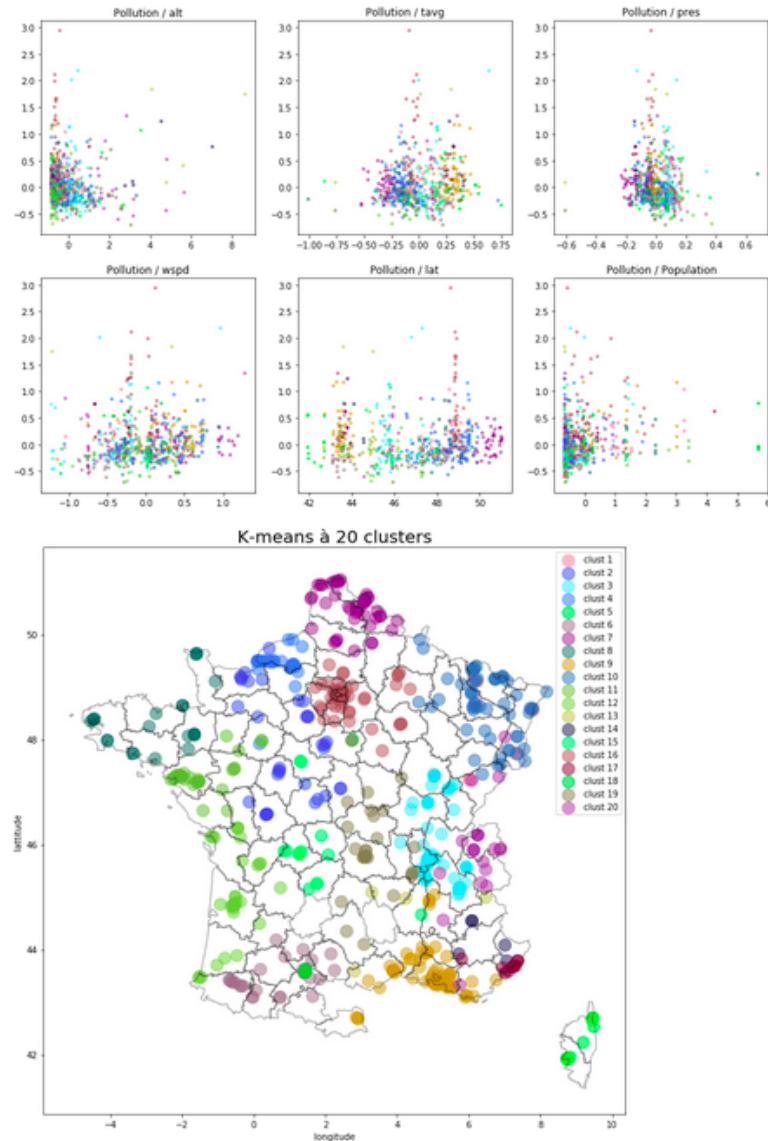
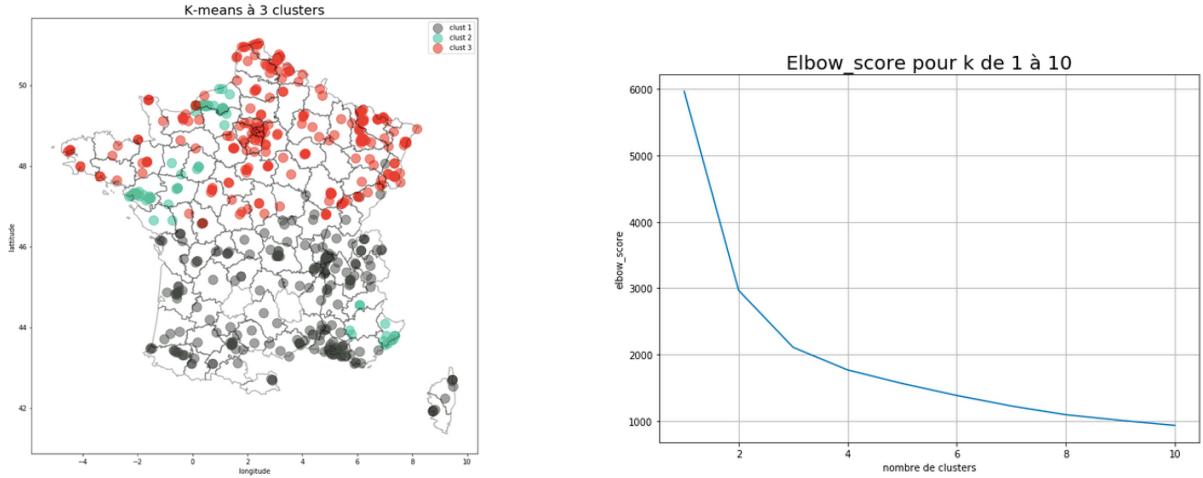


FIGURE 8 – Premier K-means - essai de clustering sur la France entière sur 20 clusters

Il est probable que les paramètres de coordonnées géographique, non standardisées, ont un poids trop important dans le calcul de distance pour faire apparaître des clusters regroupant des villes éloignées mais qui possèdent les mêmes caractéristiques de pollution (Paris et Lyon par exemple).

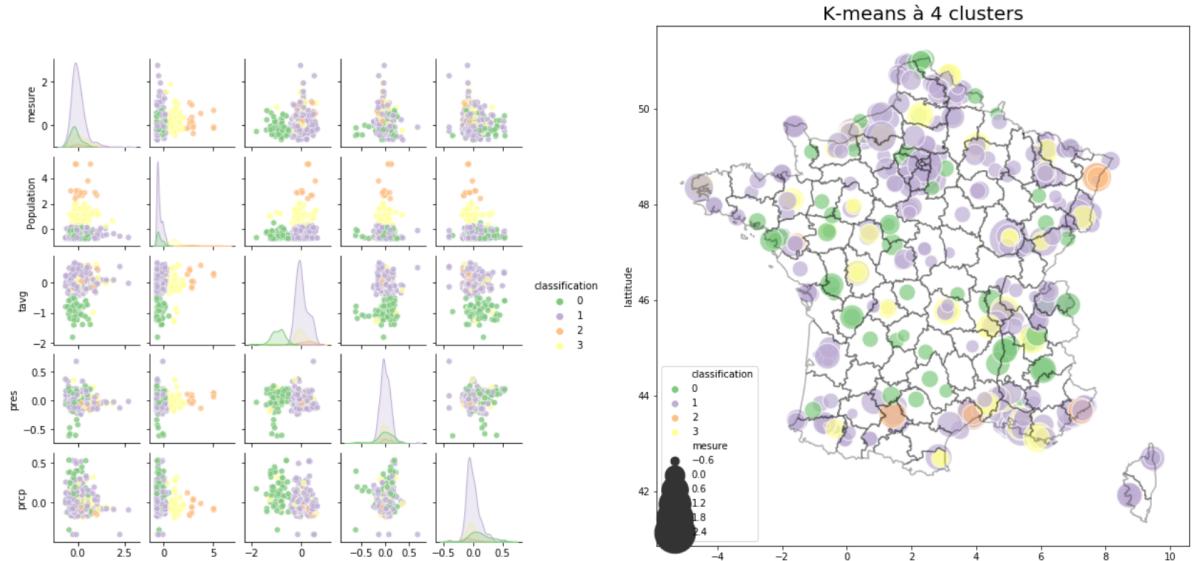


(a) K-means sur la France entière avec 3 clusters

(b) Evolution de l'elbow-score en fonction du nombre de clusters

FIGURE 9 – Recherche du nombre optimal de clusters

La prise en compte des coordonnées géographiques standardisées permet de réduire le découpage géographique (Figure 9a). Néanmoins, les coordonnées apportent une importance régionale trop importante, alors que l'objectif du clustering est bien d'identifier des zones de pollution identiques tant par les mesures que par la signature géographique et météorologique. Les coordonnées lat-long sont alors retirées, et cette action permet d'obtenir les résultats présentés en Figure 10 où le diamètre des points est exponentiellement proportionnel à la mesure moyenne totale de pollution sur le site concerné et les labels des clusters sont triés par ordre croissant de moyenne de pollution par cluster.



(a) Scatter plot des variables colorées selon le cluster

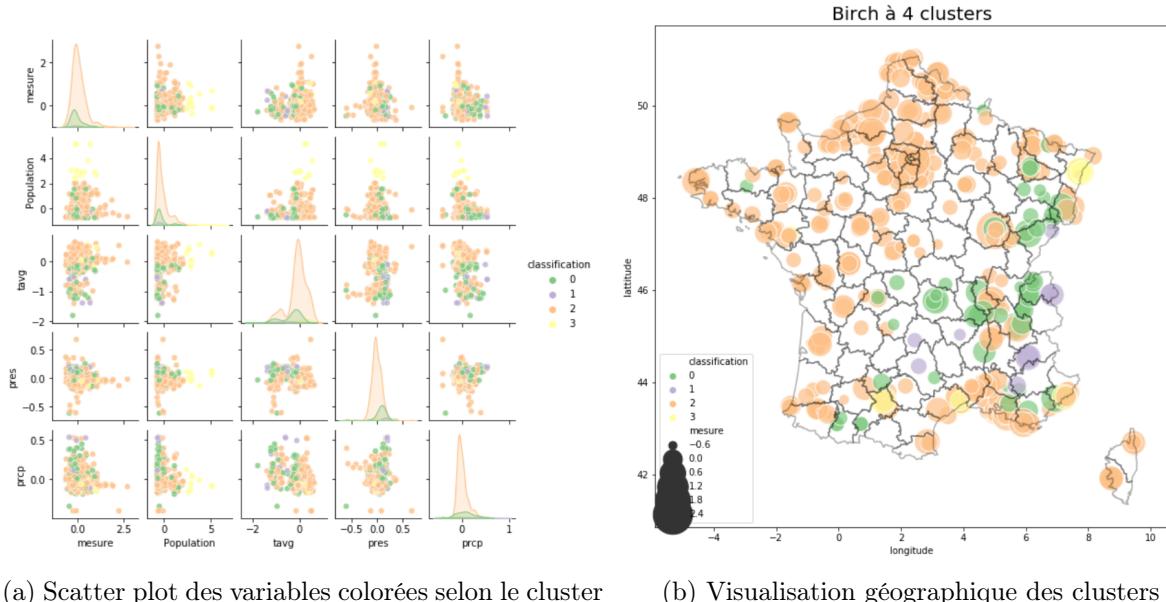
(b) Visualisation géographique des clusters

FIGURE 10 – Clustering K-means sans coordonnées lat-long.

Enfin, aux vues de la courbe d'elbow score tracée (Figure 9b), on peut penser que le nombre optimal de cluster est compris entre 3 et 4.

4.2 Clustering hiérarchique (BIRCH)

L'algorithme BIRCH est testé pour le clustering des sites de mesure en France. En reprenant la conclusion tirée par K-means, on ne prend pas en compte les coordonnées lat-long, même centrées réduites. L'arbre de décision donne des résultats intéressants (Figure 11). Nous avons également essayé de faire le clustering en ayant au préalable réduit la dimension des variables explicatives par PCA mais cela ne semble pas apporter de différence significative.



(a) Scatter plot des variables colorées selon le cluster (b) Visualisation géographique des clusters

FIGURE 11 – Clustering BIRCH à 4 clusters

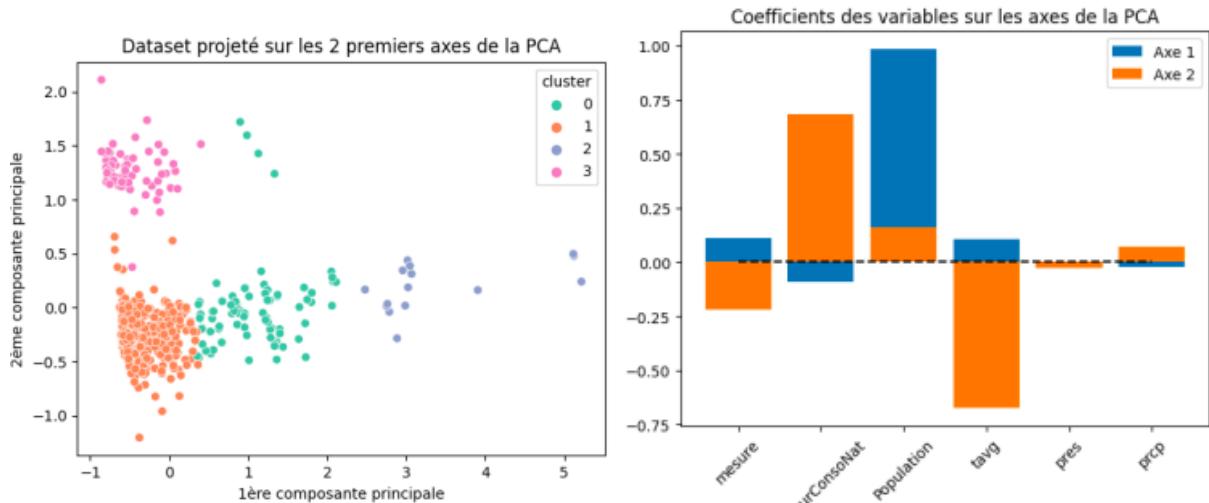


FIGURE 12 – Visualisation du clustering grâce à la PCA

Toutefois, la PCA offre la possibilité de visualiser le résultat du clustering en 2 dimensions ce qui peut aider à l'interprétation de celui-ci (Figure 12) :

- 0 : les villes de taille moyenne,
- 1 : les villes de petite taille, au climat méditerranéen et à la consommation énergétique faible,
- 2 : les grandes agglomérations,
- 3 : les villes de petite taille, au climat continental et à consommation énergétique élevée.

4.3 DBScan

Enfin un clustering par algorithme DBScan est testé. Néanmoins, la diagonale de la scatter matrice présentée en Figure 13 démontre que cet algorithme ne permet pas de distinguer de clusters avec des frontières distinctes au sein de chaque variable. Cet algorithme nous semble peu pertinent pour notre étude.

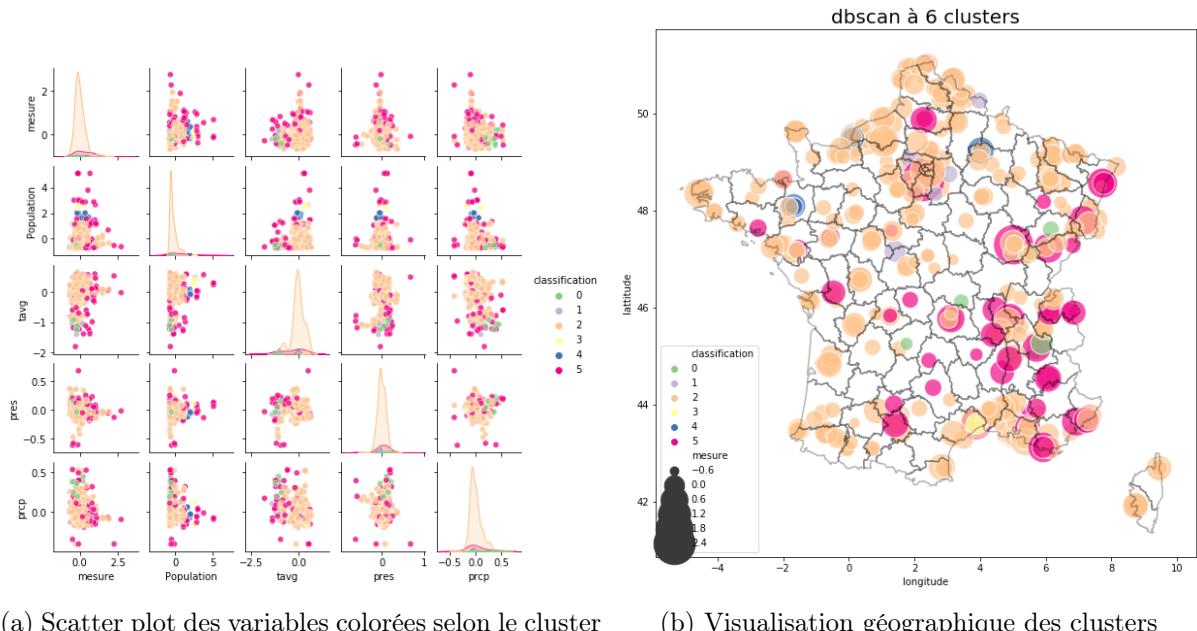


FIGURE 13 – Clustering DBSCAN à 6 clusters

4.4 Clustering : conclusions

Une fois retiré les variables latitude et longitude, les clustering par K-Means et BIRCH s'avèrent assez convaincants. Cependant, l'hétérogénéité des résultats selon les méthodes et variables explicatives retenues nous laisse penser qu'il serait nécessaire de collecter d'autres données et/ou variables pour accroître la pertinence des modèles.

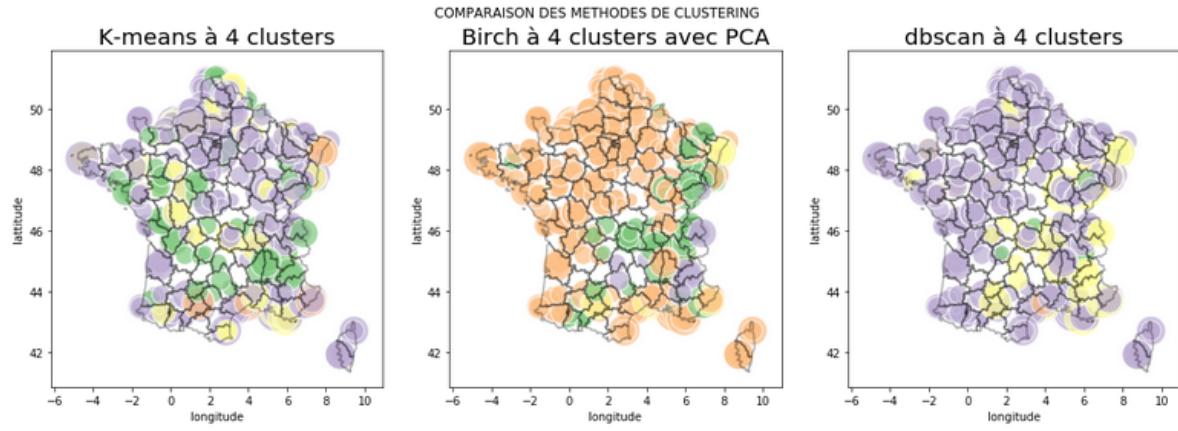


FIGURE 14 – Comparaison des algorithmes de clustering testés

5 Prédictions sur 2021

On va chercher à prédire la mesure moyenne de pollution par jour et par site. Autrement dit on entraîne un modèle par site en ayant fait une moyenne des mesures de pollution sur chaque site quotidiennement. On débute avec un modèle très simple : la régression linéaire.

Pour un site retenu, le dataset est découpé en un jeu d’entraînement s’étalant sur une période de deux ans et un jeu de test s’étalant sur une période de 6 mois.

Les résultats obtenus divergent selon le site de mesure, comme le montre le graphe présenté en Figure 15 qui illustre la valeur du coefficient de détermination sur chaque site. Cette visualisation permet de déterminer les bornes inférieure et supérieure.

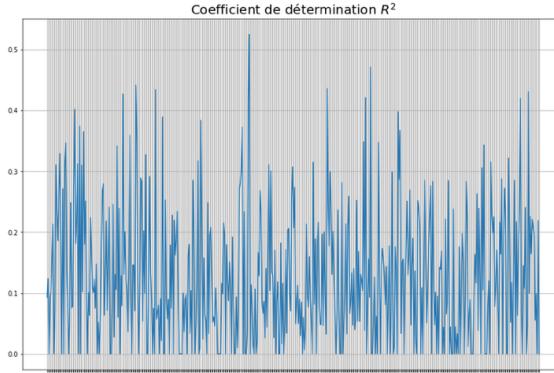


FIGURE 15 – Coefficients de détermination calculés pour chaque site.

Toutefois, la visualisation des prédictions sur les sites dont le coefficient de détermination est le plus élevé (Figure 16) montre que les prédictions sont moins dispersées que la réalité. Elles suivent la tendance mensuelle sans parvenir à coller à la précision journalière. A noter que nous avons ajouté les moyennes glissantes sur une fenêtre de 3 jours dans les features. Nous avons également essayé différentes tailles de fenêtre pour ces moyennes glissantes, sans résultat probant.



FIGURE 16 – Prédiction sur le test set des modèles de régression linéaires.

Enfin, nous avons essayé un autre type de modèle, le Support Vector Machine dont les résultats (Figure 17) sont assez similaires à la régression linéaire.



FIGURE 17 – Prédiction sur le test set des modèles SVM.

6 Conclusions

Cette étude a permis d’appréhender la difficulté à compiler un volume conséquent de données présentes sous différents formats et différents types (séries temporelles, métriques locales ou globales, etc), et d’y appliquer des algorithmes de régression à titre de prédiction.

D’autre part, l’étude de clustering sur ce type de données a posé la question de l’interprétabilité des résultats obtenus et donc la pertinence de l’application de ce type de méthode.

Enfin, la prédiction de la mesure de pollution, si elle s’avère satisfaisante en ce sens qu’elle suit l’évolution globale réelle, pourrait par la suite alimenter un modèle global prenant en entrée de quelconques coordonnées géographiques et une date, et renvoyant une mesure prédite.