**Projet SD701 – Exploration de grands volumes de données**

**Étude de la pollution en France**

*Date de rendu : 17 décembre 2021*

*Guillaume Canat*

*Jean-Bastien Sarda*

|  |
| --- |
| 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 :   1. faire apparaitre visuellement les cumuls ou moyennes de pollution sur tous les sites sur une période donnée, 2. mesurer la croissance ou décroissance de pollution dans le temps dans des époques particulières (confinements, déconfinements, installations d’usines), 3. par clustering, déterminer des nappes de pollution sur une période donnée (k-means, random forest) 4. par apprentissage, tenter de déterminer les évolutions à venir de la pollution et comparaison avec les valeurs mesurées.   Enfin le code est disponible [sur ce répertoire Github](https://github.com/jbSarda/SD701). |

Table des matières

[I. Recueil, nettoyage et compilation des données 2](#_Toc89942640)

[1.1 Données sur les sites de mesures 2](#_Toc89942641)

[1.2 Mesures de pollution 3](#_Toc89942642)

[1.3 Données supplémentaires 3](#_Toc89942643)

[II. Visualisation brute des données 5](#_Toc89942644)

[2.1 Visualisation sur carte 5](#_Toc89942645)

[2.2 Dataset compilé 6](#_Toc89942646)

[III. Clustering 7](#_Toc89942647)

[3.1 Kmeans 7](#_Toc89942648)

[3.2 Arbres de décision 9](#_Toc89942649)

[3.3 DBScan 10](#_Toc89942650)

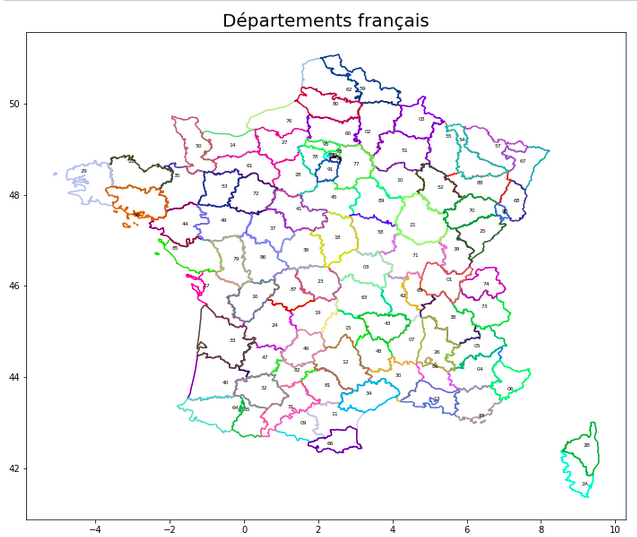
[3.4 Conclusion 11](#_Toc89942651)

[IV. Prédiction sur 2021 11](#_Toc89942652)

# Recueil, nettoyage et compilation des données

## 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 géographies 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 *https://france-geojson.gregoiredavid.fr/*).

La densité démographique a également été répertoriée pour chaque site de mesure à partir du fichier des communes françaises.

Figure 1: carte des départements de France

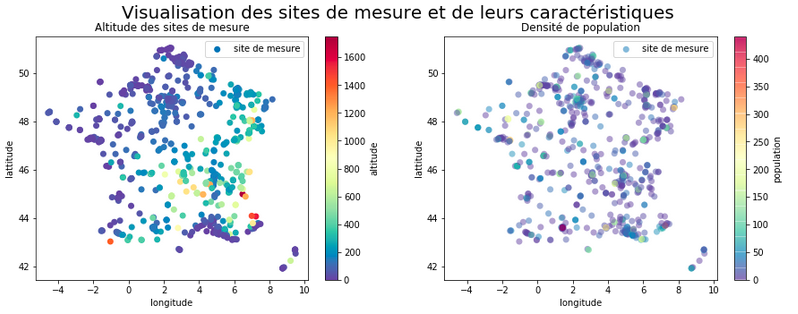


Figure 2: localisation des sites de mesures, distingués par altitude puis population

## Mesures de pollution

Site : *https://files.data.gouv.fr/lcsqa/concentrations-de-polluants-atmospheriques-reglementes/old/*

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 Gb de données) nous avons opté pour une mesure journalière par site.

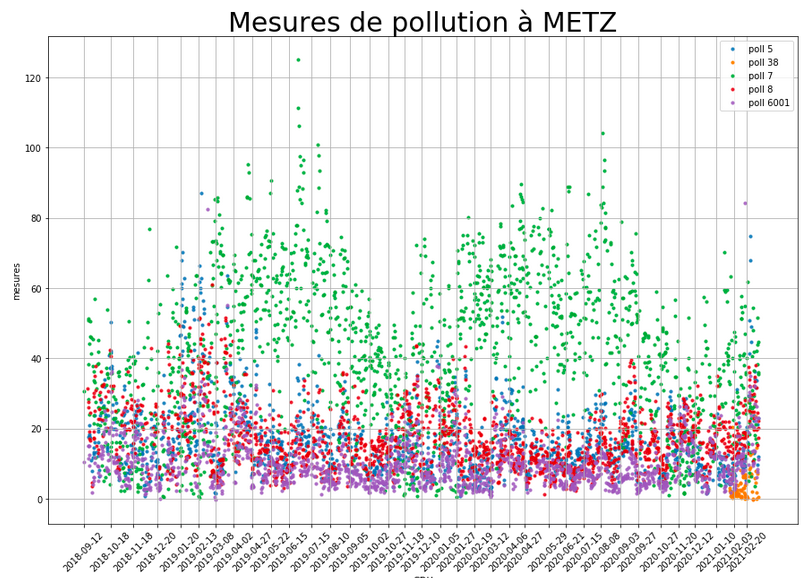
Pour un site pris au hasard, on constate que tous les polluants ne sont pas mesurés et certains ne sont mesurés qu'à partir de certaines périodes. Ce constat nous pousse alors à standardiser les mesures et les regrouper sur une valeur moyenne. Ce choix assumé peut comporter un biais : il est possible 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.

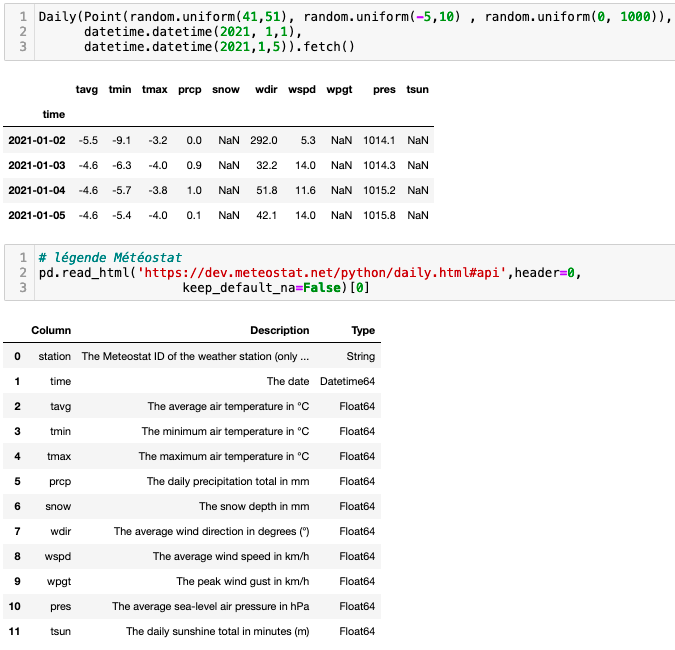
Figure 3: représentation des mesures de pollution à Mets, pour différents types de polluants

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

## Données supplémentaires

Pour pouvoir entrainement convenablement notre modèle, nous devons trouver des métriques 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 :

* 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 comme suit :



* 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.

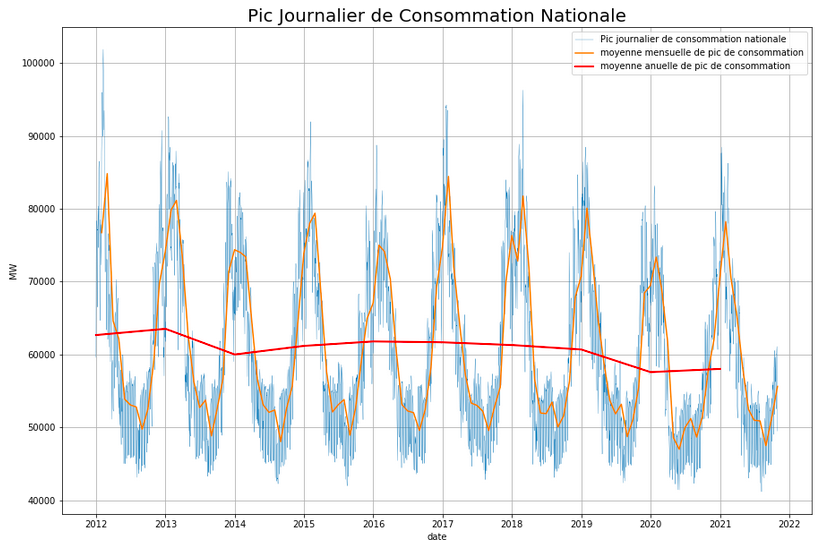


Figure 4: évolution du pic de consommation électrique journalier de 2012 à 2022

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

# Visualisation brute des données

## 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 font apparaître que les mesures les plusélevées 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), possiblement plus apparente lors du clustering.

Fonction : *graph\_poll(date\_start, date\_end, lis\_dep, lab = False)*



Figure 5: représentation de le pollution annuelle moyenne sur les sites de mesure en France par année

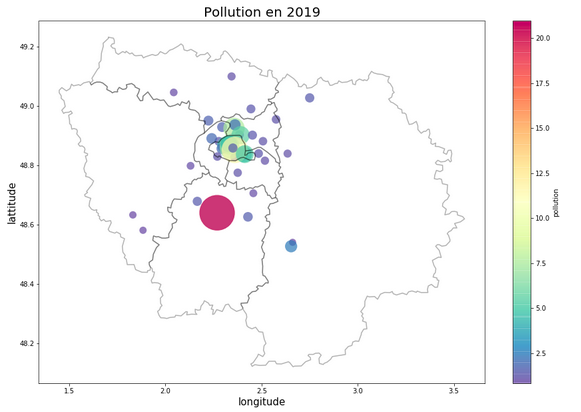
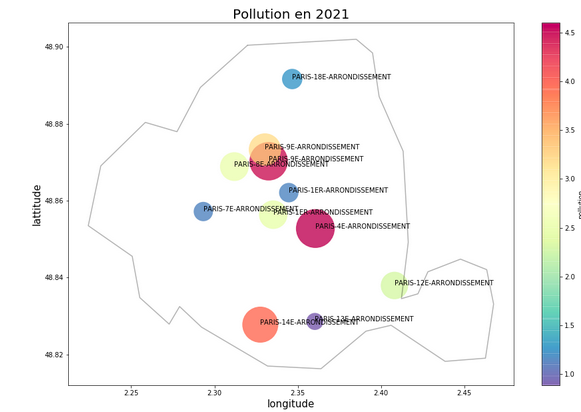


Figure 6: étude de la pollution moyenne en Ile de France en 2019

On peut prendre comme exemple l’Ile 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 7: étude de la pollution moyenne de janvier à mars 2021 à Paris



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 :

## 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.

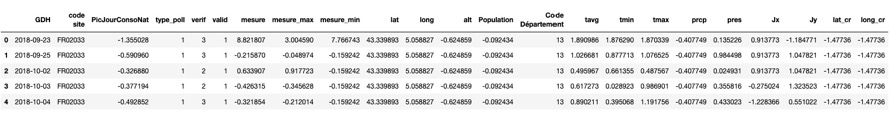


Figure 8: premières lignes du dataset final utilisé pour entrainer les algorithmes

# Clustering

## Kmeans

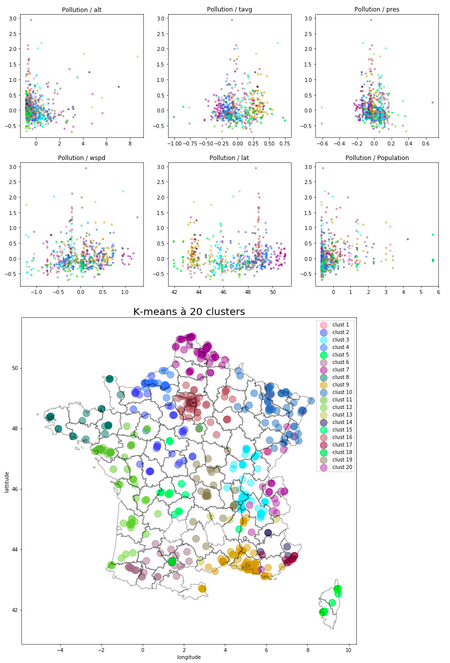


Figure 9: premier K-means - essai de clustering sur la France entière sur 20 clusters

On souhaite entrainer 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.

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

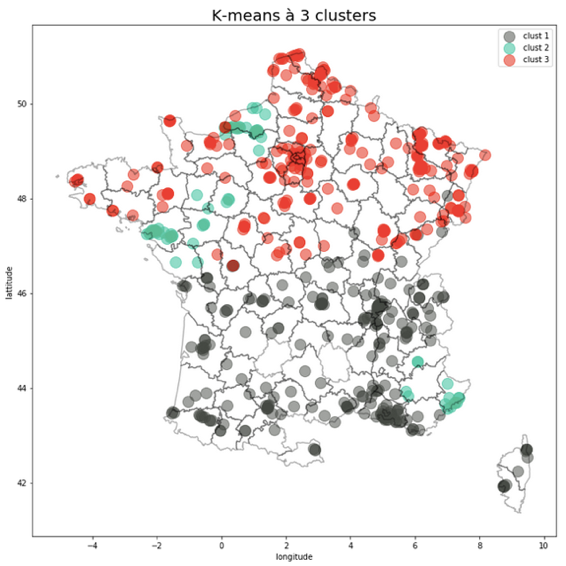


Figure 10: K-means sur la France entière avec 3 clusters

La prise en compte des coordonnées géographiques standardisées permet de réduire le découpage géographique (Figure 10). 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 le résultat présenté en Figure 11.

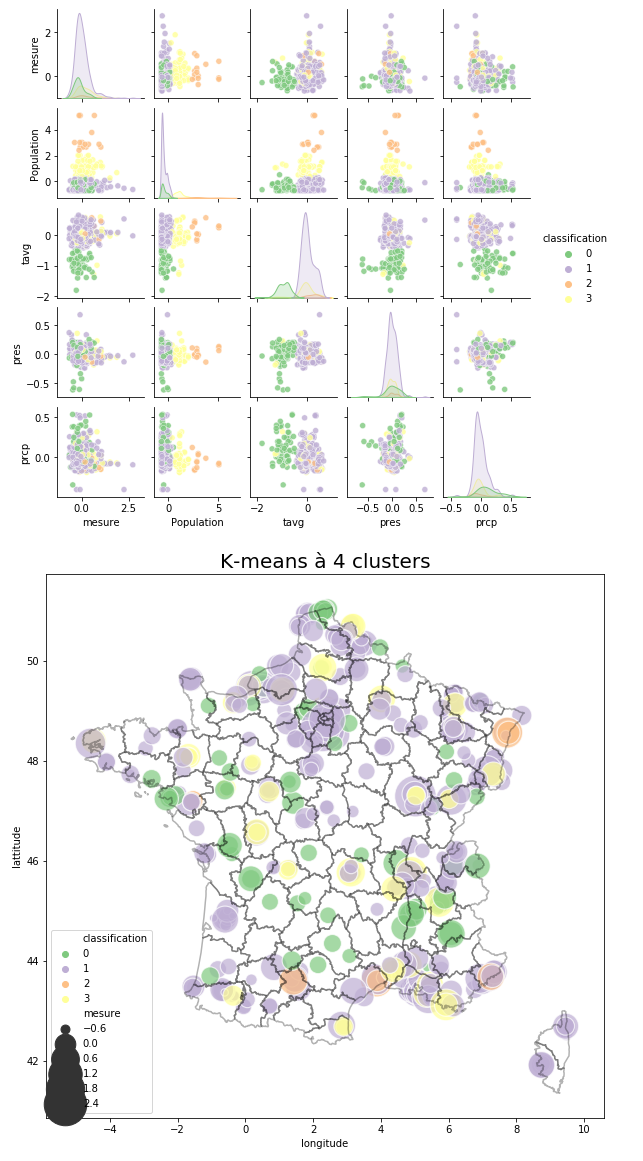


Figure 11: clustering par K-means en retirant les métriques de coordonnées lat-long. le diamètre des points est exponentiellement proportionnel à la mesure moyenne totale de pollution sur le site concerné. Enfin les labels des clusters sont triés par ordre croissant de moyenne de pollution par cluster.

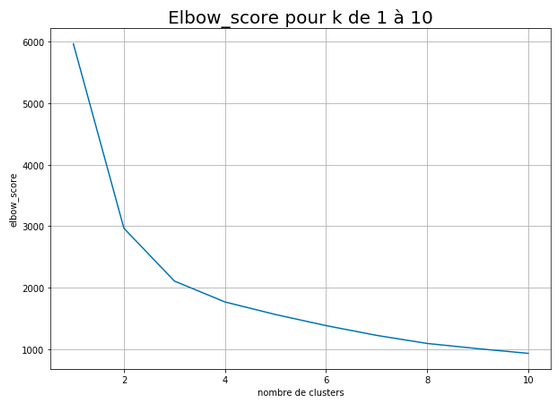


Figure 12: étude du elbow\_score pour déterminer la valeur optimale du nombre de clusters (3, 4)

Enfin, aux vues de la courbe d’elbow score tracée ci-contre, on peut penser que le nombre optimal de cluster est compris entre 3 et 4.

## Clustering hiérarchique (BIRCH)

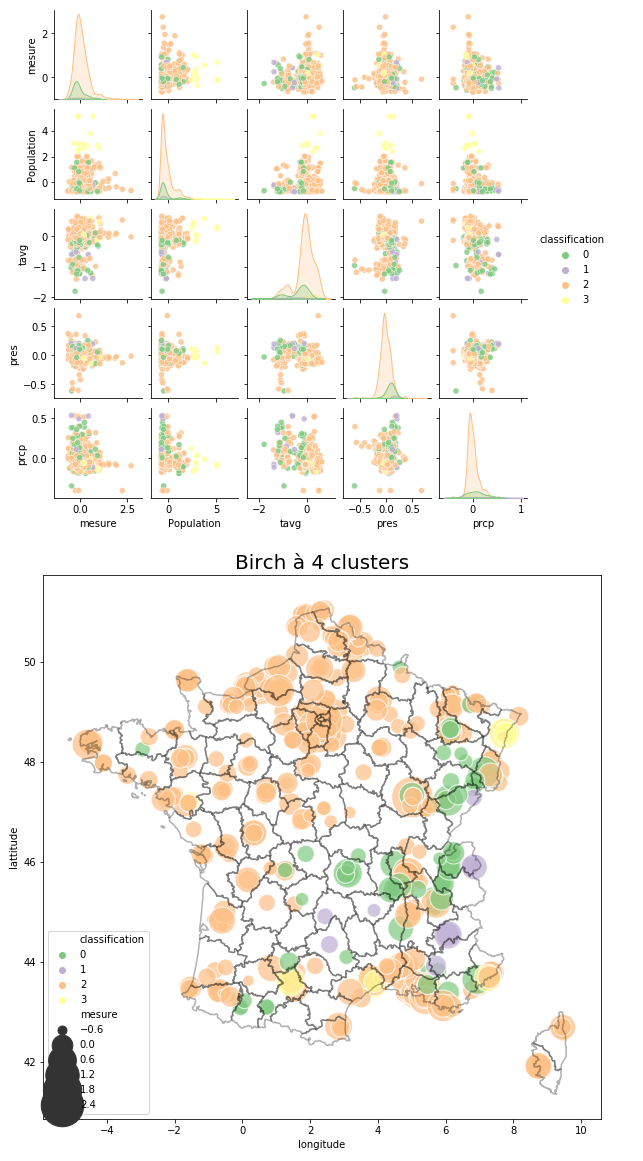


Figure 13: clustering par algorithme de Birch

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 intéressant (Figure 13). 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 (Figure 14).

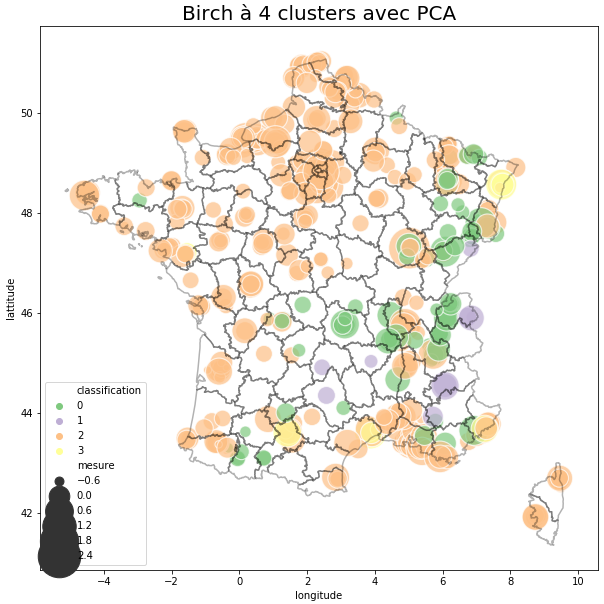


Figure 14: clustering par algorithme de Birch avec PCA à 3 composantes auparavant

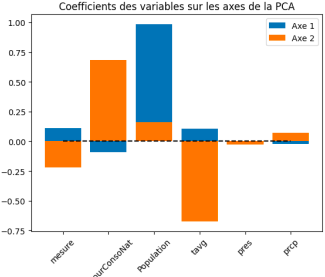
Toutefois, la PCA offre la possibilité de visualiser le résultat du clustering en 2 dimension ce qui peut aider à l’interpretation de celui-ci (Figure 15) :

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.



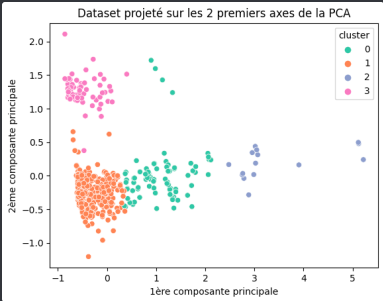
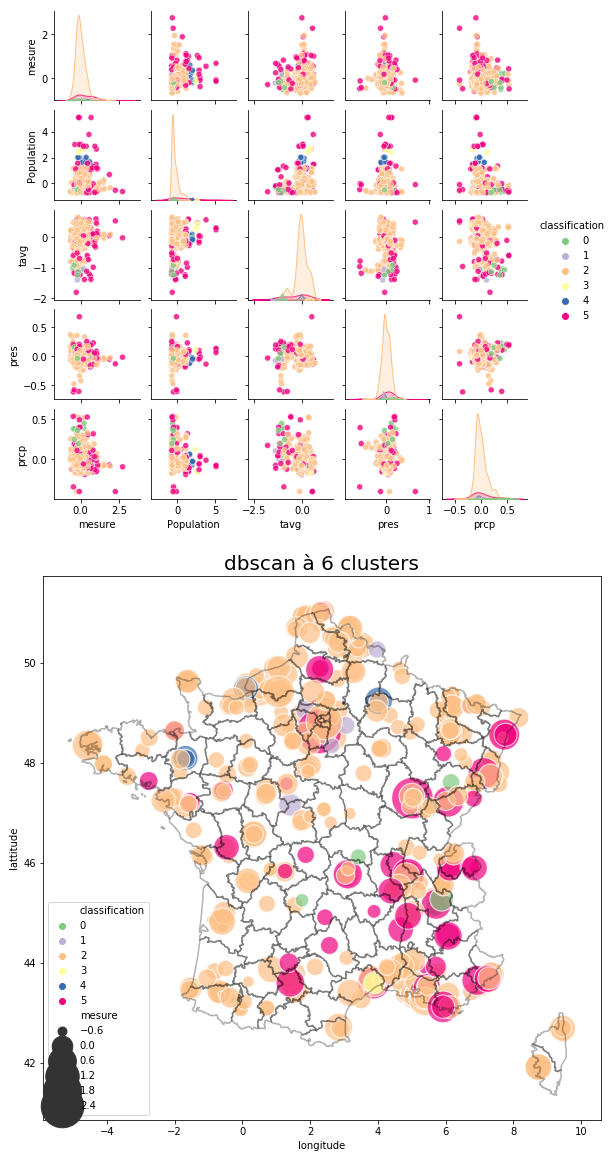


Figure 15: visualisation du clustering par K-means grâce à la PCA

## DBScan

Enfin un clustering par algorithme DBScan est testé. Néanmoins, la diagonale de la scatter matrice présentée ci-dessous 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.



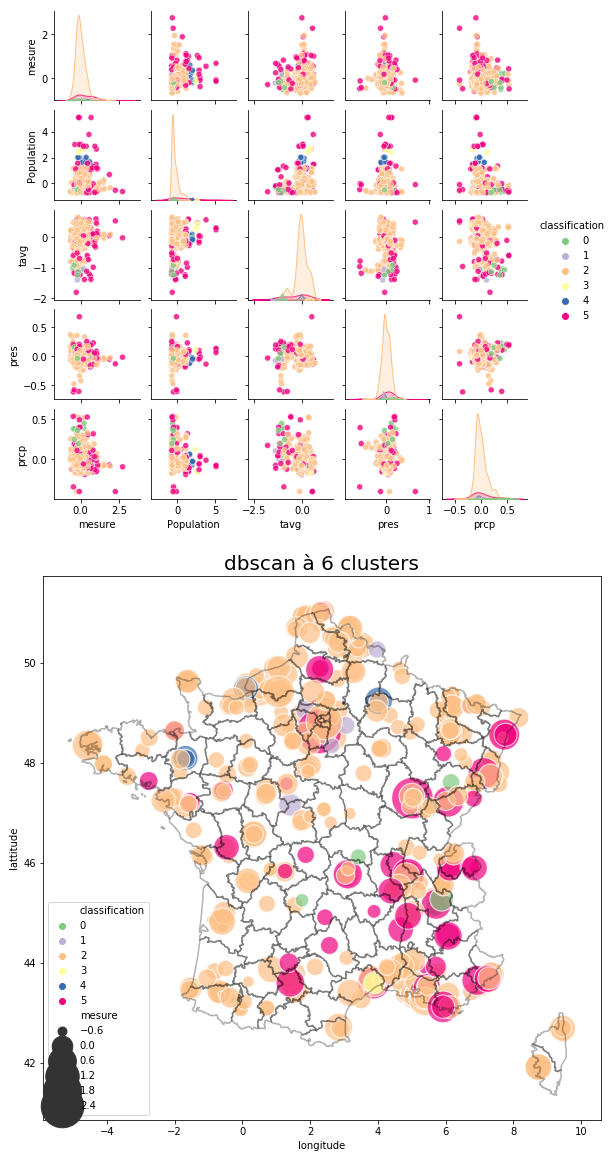


Figure 16: clustering par DBScan

## Conclusion

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/variables pour accroitre la pertinence des modèles.

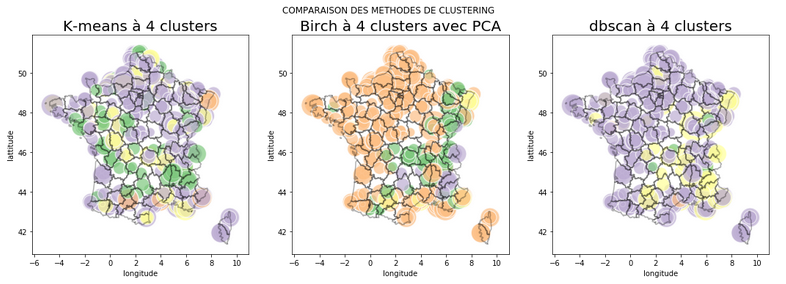
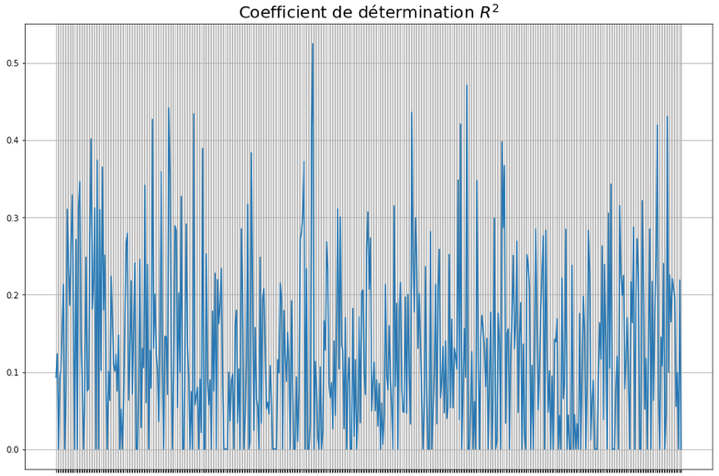


Figure 17: comparaison des algorithmes de clustering testés

# Prédiction sur 2021

On va chercher à prédire la mesure moyenne de pollution par jour et par site. Autrement dit on entraine 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’entrainement 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 ci-contre (Figure 18) représentant la valeur du coefficient de détermination sur chaque site.

Figure 18: coefficient de détermination calculé pour chaque site. Cette visualisation permet de déterminer les bornes inférieure et supérieure

Toutefois, la visualisation des prédictions sur les sites dont le coefficient de détermination est le plus élevé (Figure 19) 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.

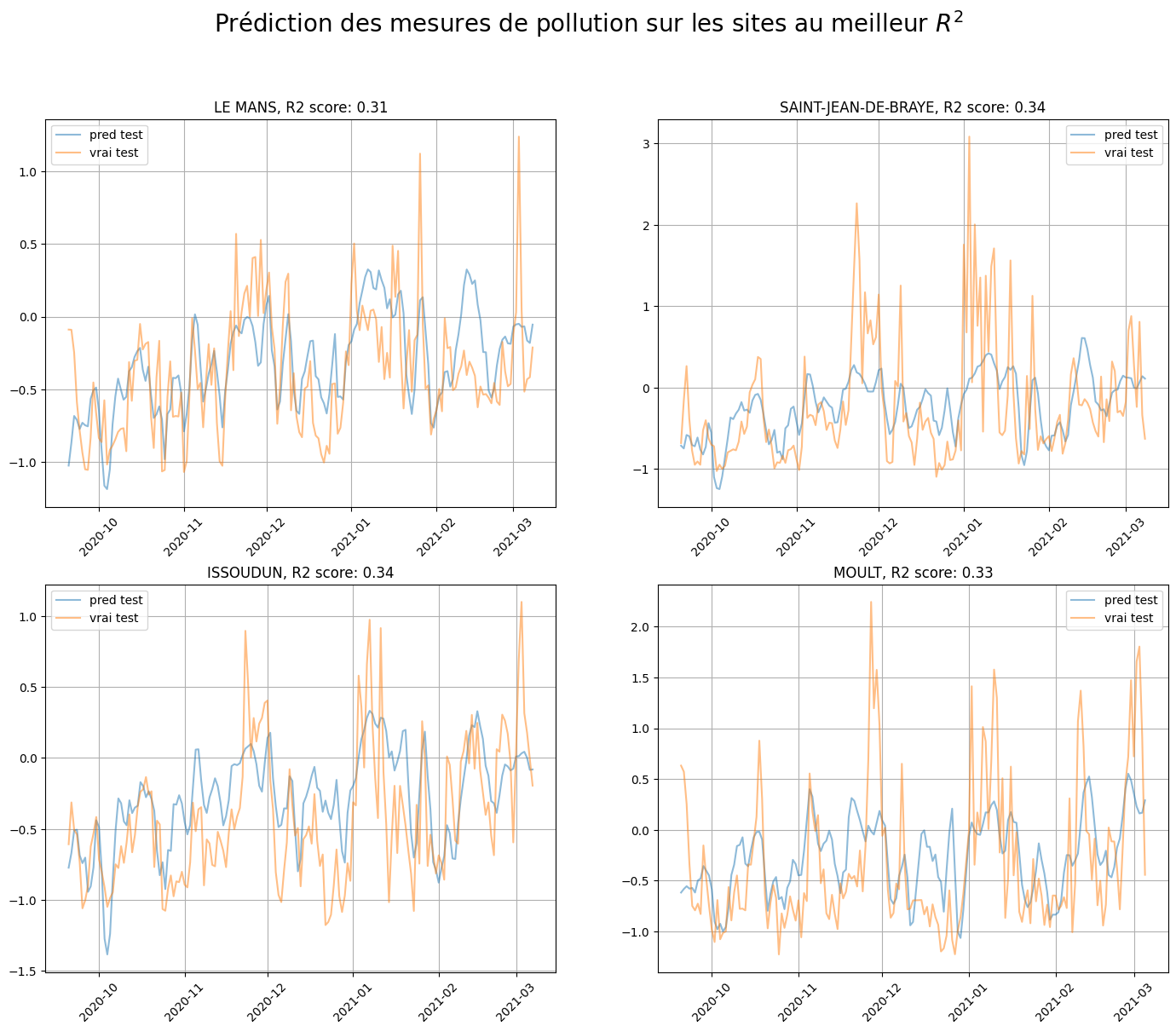


Figure 20: Prédictions sur le test set des modèles de régression linéaires

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



Figure 21: Prédictions sur le test set des modèles de type SVM

# Conclusion

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.