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

**Étude de la pollution en France**

Guillaume Canat

Jean-Bastien Sarda

|  |
| --- |
| To do list   * Ajouter légende * Ajouter tailler des sites de mesure * Visualisation du df\_tot * Birch au lieu de random forest * Ajout partie dbscan * Reprendre les algo sans les données GPS * Prédiction : SVM, reg logistique, SGD |

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 :

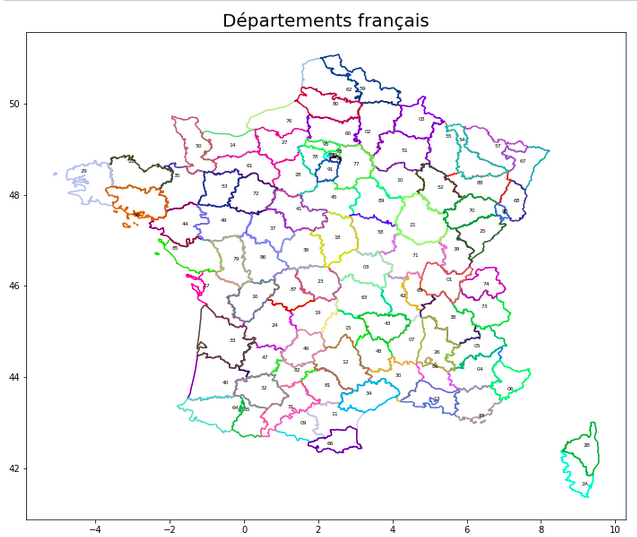
1. faire apparaitre visuellement les cumuls ou moyennes de pollution sur tous les sites sur une période donnée,
2. par clustering, déterminer des nappes de pollution sur une période donnée (k-means, random forest)
3. mesurer la croissance ou décroissance de pollution dans le temps dans des époques particulières (confinements, déconfinements, installations d’usines),
4. par apprentissage, tenter de déterminer les évolutions à venir de la pollution et comparaison avec les valeurs mesurées.

# Recueil, nettoyage et compilation des données

## Données sur les sites de mesures

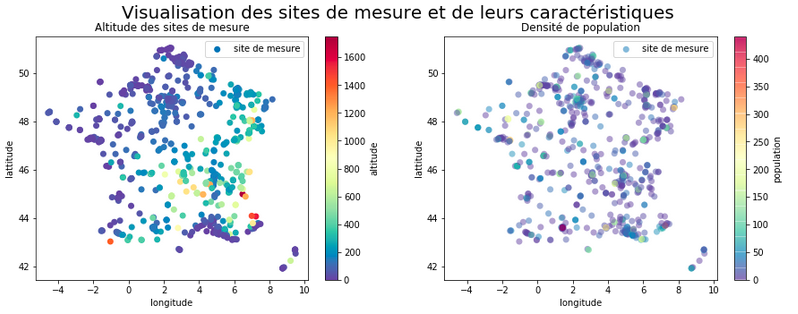
Les données géographiques concernant les sites de mesures sont compilées dans le fichier *fr-2020-d-lcsqa-ineris-20210412.xml*. Les paramètres qui s’y trouvent concernent le code du site, son nom, son altitude et ses coordonnées géographies en format latitude-longitude.

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.

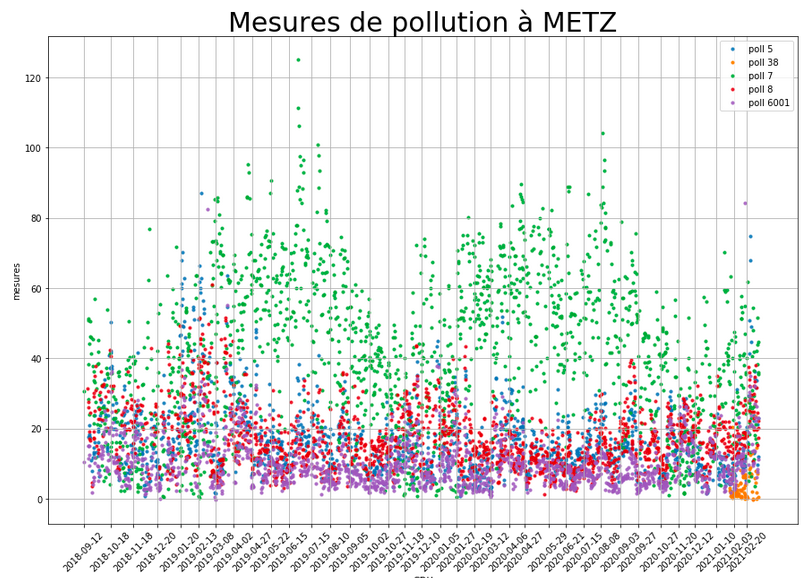


## Mesures de pollution

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.

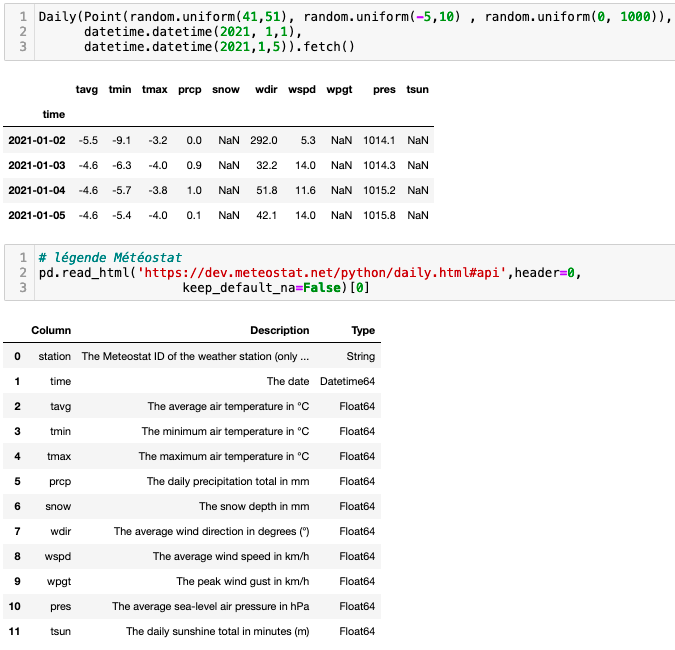
Pour un site pris au hasard, on constate que tous les polluants ne sont pas mesurés et certains ne sont mesurés qu'à partir de certaines périodes. Ce constat nous pousse alors à standardiser les mesures et les regrouper sur une valeur moyenne. Ce choix assumé peut comporter un biais : il est possible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

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

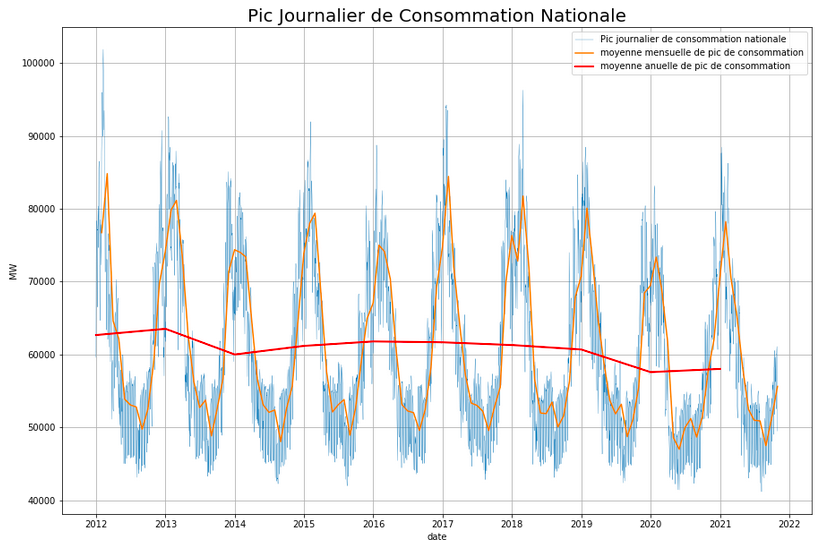
## Données supplémentaires

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.



* 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

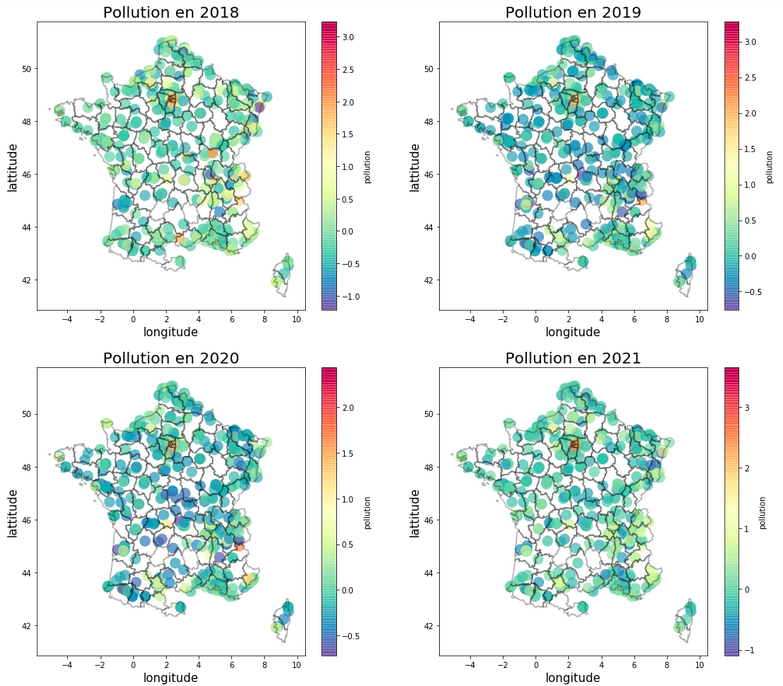
## Visualisation par tracé

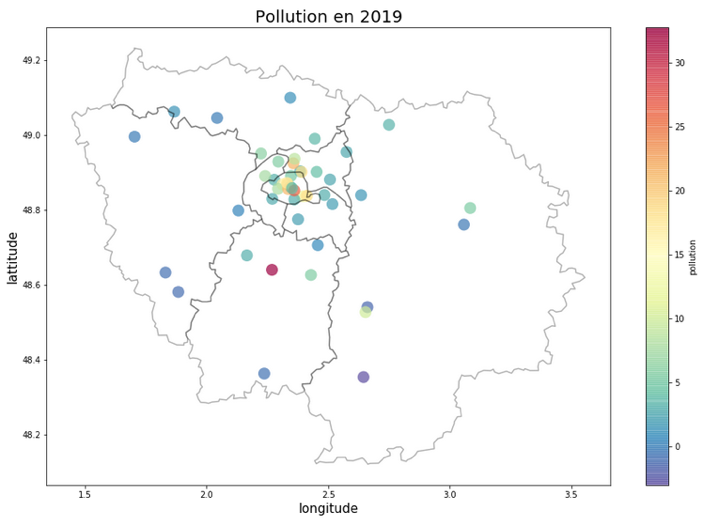
## Visualisation sur carte

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.

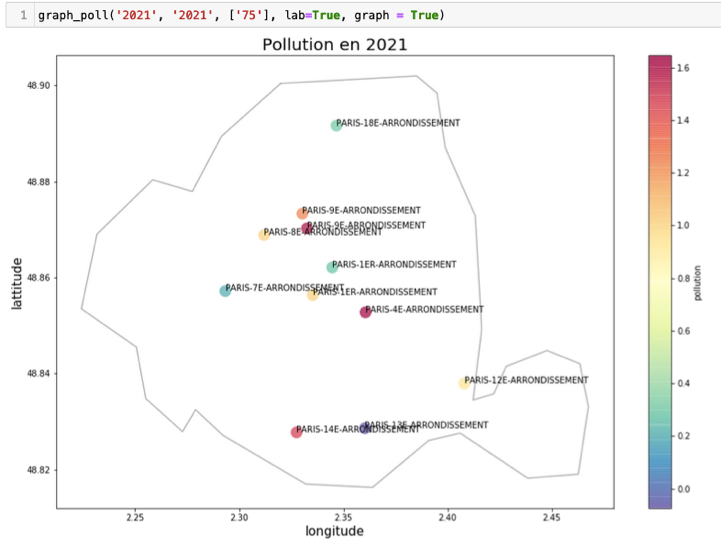
Les représentations à l’échelle nationale sur des périodes allant du mois à l’année font apparaître que les mesures les plusélevées se concentrent dans et autour des zones urbaines. D’autre part, on constate la présence de la fameuse « diagonale du vide » allant du Sud Ouest au Nord Est (trainée bleue), possiblement plus apparente lors du clustering.

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





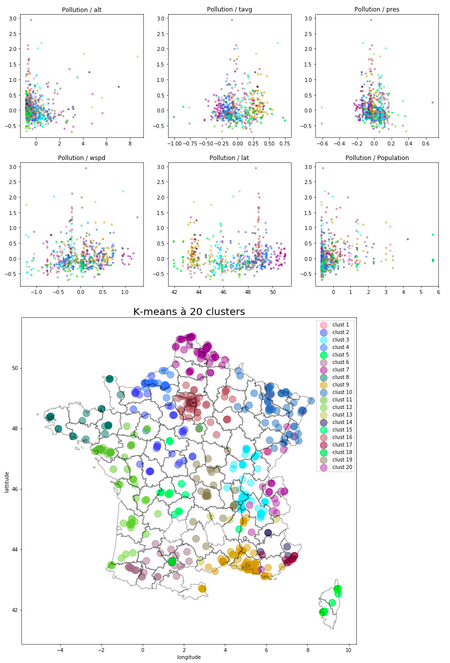
On peut prendre comme exemple l’Ile de France, dont l’étude de la pollution moyenne révèle une concentration des valeurs les plus hautes à Paris, et une dispersion de la pollution à mesure que l’on s’éloigne de la capitale.



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 :

# Clustering

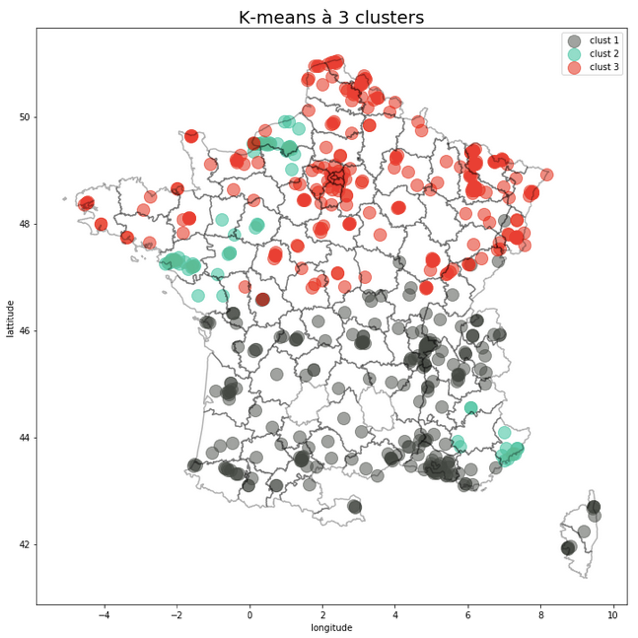
## Kmeans



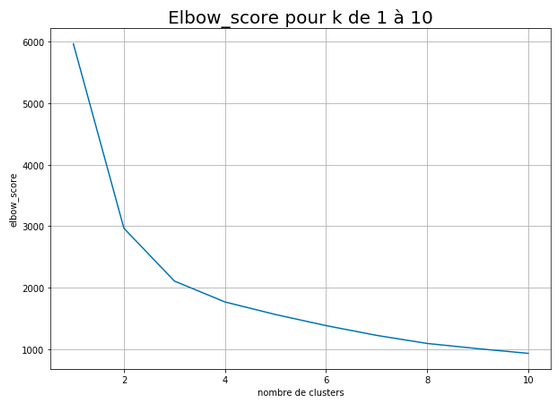
On souhaite entrainer le modèle en utilisant la moyenne des mesures de pollution sur la période complète. Dans un premier temps, les données standardisées sont les données météorologiques, l’altitude, et la densité de population. La consommation énergétique et les coordonnées des jours de la semaine ne sont pas des métriques spécifiques aux sites mairs 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 (ci-dessous, on impose 4 clusters).

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)



La prise en compte des coordonnées géographiques standardisées permet de réduire le découpage géographique. Pour autant, K-means ne semble pas adapté à la recherche de cluster.



Aux vues de la courbe d’elbow\_score tracée ci-dessus, on peut affirmer que le nombre optimal de cluster est compris entre 3 et 4.

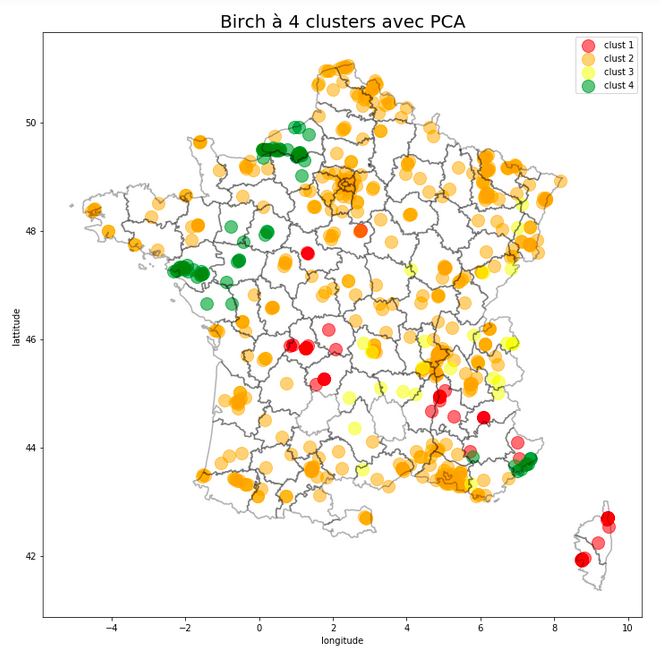
## Random Forest



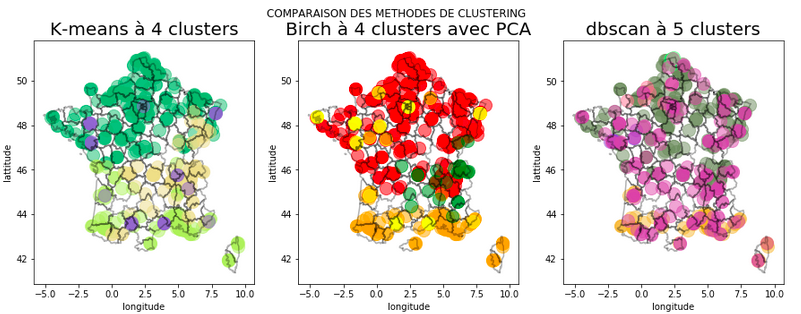
L’algorithme de Birch est testé pour la classification des sites de mesure en France.

Il apparait que l’algorithme de clasification soit très dépendant de la latitude, comme représenté ci-contre.

Ce découpage peut être atténué en appliquant une PCA sur les neuf variables étudiées (ci-dessous, avec un PCA à 3 composantes principales).

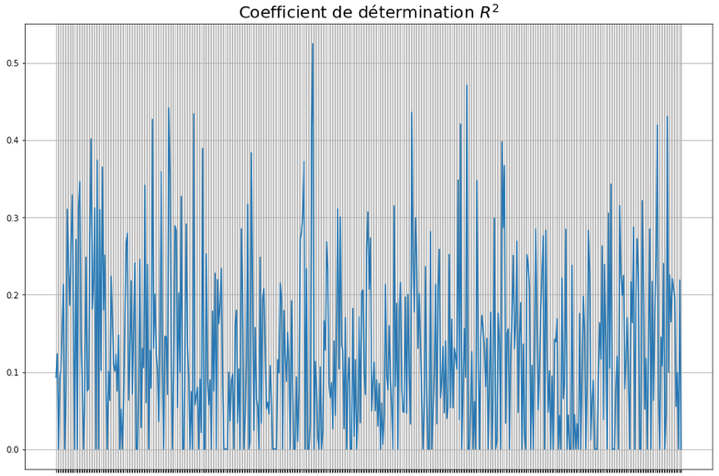


Conclusion : le clustering retourne des résultats différents selon a méthode utilisée. L’algorithme du K-means est le plus crédible.

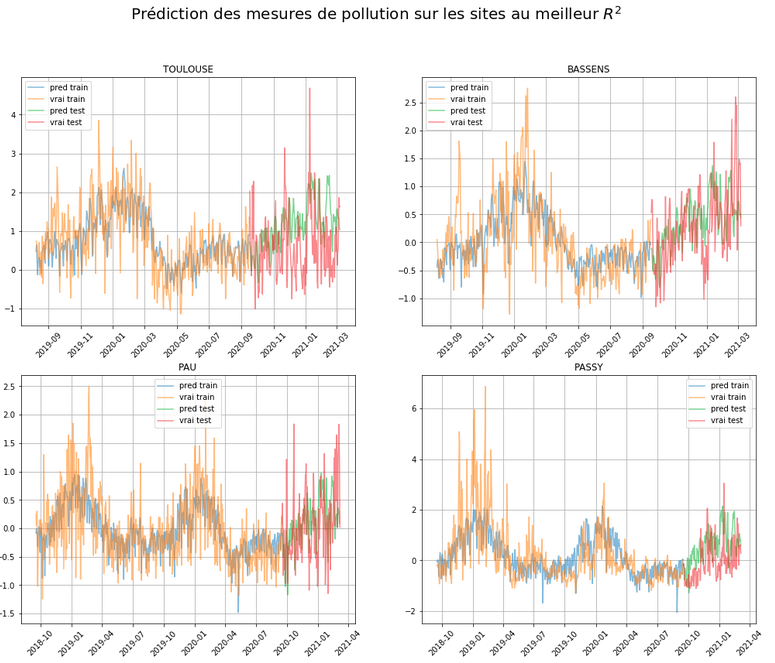


# Prédiction sur 2021

On réalise des opérations de régression linéaire par site, en ayant auparavant supprimé toutes les lignes contenant des NaN dans le dataset global, et en ayant fait une moyenne des mesures de pollution centrées réduites enregistrées sur chaque site quotidiennement.



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 validation s’étalant sur une période de 6 mois.



Les résultats obtenus divergent selon le site de mesure, comme le montre le graphe ci-contre représentant la valeur du coefficient de détermination sur chaque site.

Toutefois, la visualisation des prédictions sur les sites dont le coefficient de détermination est le maximal donne des résultats montrant que les prédictions sont moins dispersées que la réalité. Elles suivent toutefois la tendance mensuelle sans parvenir à coller à la précision journalière.