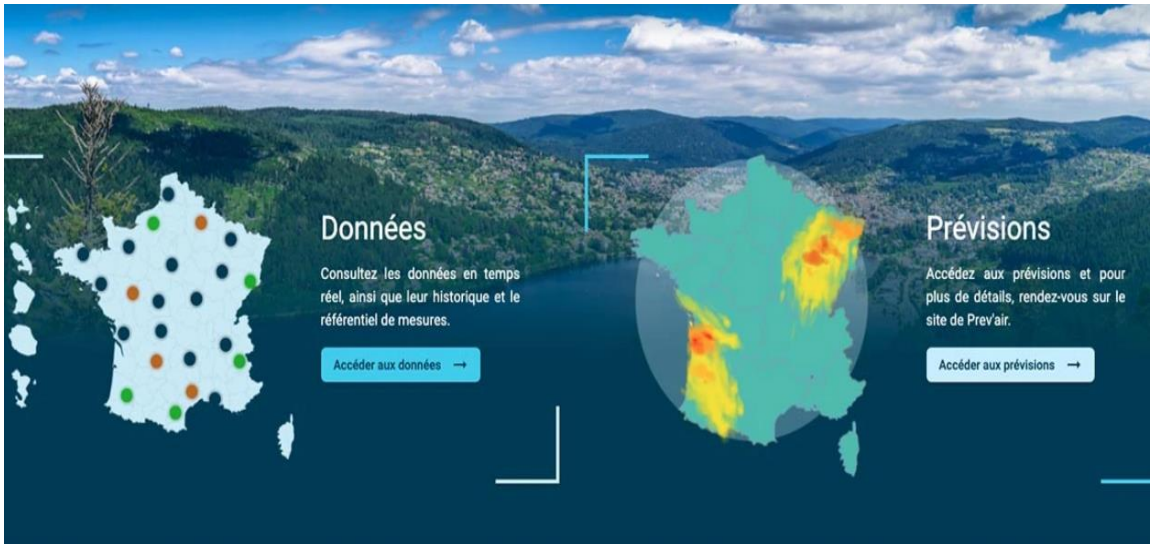
République du Sénégal
Un Peuple – Un But – Une Foi

Ministère de l'Enseignement Supérieur,
de la Recherche et de l'Innovation

Université Iba Der Thiam de Thiès
UFR Sciences et Technologies
Département Informatique
Master Intelligence Artificielle et Smart Tech

PROJET FINAL

CONCEPTION DE CHAINE DE VALORISATION



Exploration des données ATMO et prédictions sur la qualité de l'air

Réalisé par :

Mohamed-Saiaf ALIAMINI

Ndiémé DIOUF

Thierno FALL

Anna Ramata SAWADOGO

Khoudia SOW

Professeur :

Dr. Babiga BIRREGAH

Table des matières

Introduction

La pollution de l'air représente un enjeu environnemental et sanitaire majeur à l'échelle mondiale. Les effets néfastes de la dégradation de la qualité de l'air sur la santé humaine et les écosystèmes sont désormais bien documentés. Les zones urbaines densément peuplées sont particulièrement vulnérables en raison de la concentration des activités industrielles et du trafic routier à proximité des centres-villes. Face à cette problématique, la prédiction et le contrôle de la pollution atmosphérique deviennent des impératifs pour prévenir et atténuer les pics de pollution.

Cependant, modéliser la pollution de l'air s'avère être une tâche complexe en raison de la multitude de facteurs qui influencent ce phénomène. Les conditions météorologiques, telles que la température, l'humidité, les vents et les inversions thermiques, jouent un rôle crucial dans la dispersion des polluants atmosphériques. De plus, la topographie d'une région, notamment la présence de vallées, de montagnes ou de zones côtières, peut affecter la circulation de l'air et donc la concentration des polluants. Les activités anthropiques liées au transport, à l'industrie, à l'agriculture et au chauffage domestique sont également des sources majeures d'émissions de polluants.

Pour relever ce défi, diverses approches de modélisation ont été développées, allant des méthodes traditionnelles basées sur des équations mathématiques aux techniques d'apprentissage automatique avancées. Les modèles linéaires, les réseaux de neurones artificiels, les approches hybrides combinant la logique floue et les réseaux neuronaux sont autant de solutions proposées dans la littérature scientifique. Le choix de la méthode appropriée dépend de la disponibilité des données, de la précision requise et de la complexité du système à modéliser.

Dans ce contexte, ce rapport vise à explorer les différentes méthodes de prédiction de la pollution de l'air en exploitant les données fournies par l'API Atmo Data, un agrégateur des flux open data des Associations Agréées de Surveillance de la Qualité de l'Air (AASQA) en France. Après une revue de la littérature, nous présenterons en détail l'API Atmo Data et les étapes de collecte des données. Ensuite, nous procéderons à une exploration approfondie des données, suivie des étapes de prétraitement et de modélisation. Enfin, nous évaluerons et interpréterons les résultats obtenus, en discutant de leur pertinence et de leur fiabilité pour la prédiction de la pollution de l'air.

I. Compréhension du problème

A. Définition du problème de prédiction de la pollution de l'air

La pollution est définie comme un phénomène ou un agent perturbateur d'un équilibre établi, surtout s'il est nuisible à la vie et à la santé. Elle peut être d'origine anthropique, c'est-à-dire causée par l'activité humaine, ou non humaine, telle que l'activité volcanique.

La pollution atmosphérique constitue un problème majeur de santé publique à l'échelle mondiale, particulièrement dans les zones urbaines densément peuplées, où l'activité industrielle est souvent concentrée à proximité des centres urbains.

Prédire et contrôler la qualité de l'air demeurent les seuls moyens de lutter contre les pics de pollution, qui ont des effets néfastes sur la santé et l'écosystème. En reproduisant les phénomènes de pollution atmosphérique, il est possible de mieux les étudier, de les comprendre et d'identifier les paramètres les plus influents, afin de disposer des éléments nécessaires pour prendre des décisions allant jusqu'à l'évacuation des villes et la suspension de certaines activités industrielles.

La pollution de l'air est un problème de plus en plus préoccupant dans de nombreuses villes du monde. Plusieurs facteurs influent sur ce phénomène, parmi lesquels les conditions climatiques, la topographie et l'activité urbaine occupent une place prépondérante. Cette complexité rend la modélisation de la pollution atmosphérique très difficile.

B. Revue de littérature sur les méthodes de prédiction de la pollution de l'air

La prédiction de la pollution de l'air est abordée par divers modèles, allant des approches classiques aux méthodes modernes intégrant des technologies avancées. Voici une vue détaillée de certains de ces modèles

- **Approches traditionnelles** : Ces modèles reposent sur des équations mathématiques des émissions de polluants provenant de diverses sources comme les véhicules, les industries et les activités domestiques. Ils considèrent les caractéristiques des sources de pollution, les conditions météorologiques et la dispersion des polluants pour estimer les concentrations à différents endroits et moments.
- **Modèles linéaires** : Ces modèles établissent des relations linéaires entre les variables météorologiques, les émissions de polluants et les concentrations observées. Bien qu'ils soient simples à mettre en œuvre, leur précision peut parfois être limitée en raison de la complexité des interactions.
- **Méthodes basées sur les réseaux de neurones artificiels (RNA)** : Ces modèles, inspirés du cerveau humain, apprennent à partir de données et détectent des motifs complexes. En utilisant des données historiques, ils peuvent prédire les niveaux futurs de pollution avec une précision accrue.
- **Approches hybrides** : Elles combinent la logique floue et les réseaux de neurones artificiels pour améliorer la précision des prévisions. La logique floue permet de modéliser des relations complexes en utilisant des concepts linguistiques, ce qui est utile pour tenir compte de l'incertitude des données environnementales.

En somme, la littérature scientifique propose une gamme de méthodes pour prédire la pollution de l'air, des modèles traditionnels aux approches basées sur l'apprentissage automatique. Le choix de la méthode appropriée dépend de la disponibilité des données, de la précision requise et de la complexité du système à modéliser.

C. Identification des facteurs environnementaux influençant la pollution de l'air

La pollution de l'air est un problème complexe influencé par divers facteurs environnementaux. Parmi ces facteurs, certains sont particulièrement significatifs :

- Les conditions météorologiques comme la température, l'humidité, la vitesse et la direction du vent, ainsi que les inversions de température, jouent un rôle crucial dans la dispersion des polluants atmosphériques. Par exemple, les inversions de température peuvent piéger les polluants près du sol, aggravant ainsi la pollution dans les zones urbaines.
- Les émissions anthropiques issues des activités humaines telles que le transport, l'industrie, l'agriculture et le chauffage domestique, sont des sources majeures de polluants atmosphériques tels que les oxydes d'azote (NO_x), les composés organiques volatils, les particules fines et le dioxyde de soufre (SO_2). Ces émissions varient en fonction des activités et des technologies utilisées.
- La topographie, comprenant la configuration géographique comme la présence de montagnes, de vallées ou de zones côtières, peut influencer la circulation atmosphérique et la dispersion des polluants. Par exemple, les zones urbaines enclavées entre des montagnes peuvent être plus sujettes à la pollution atmosphérique en raison de la stagnation de l'air.
- Les phénomènes naturels tels que les éruptions volcaniques, les incendies de forêt, les tempêtes de poussière et les émissions biogéniques peuvent également contribuer à la pollution de l'air à grande échelle.
- Les interactions chimiques entre les polluants atmosphériques et les composés présents dans l'atmosphère peuvent former de nouveaux polluants, affectant ainsi la composition chimique de l'air et sa qualité. Par exemple, les réactions entre les oxydes d'azote et les composés organiques volatils peuvent produire de l'ozone troposphérique, un polluant nocif.

On peut aussi présenter certains paramètres les plus utilisés pour les modèles de prédiction de la qualité de l'air, donnés par le tableau suivant :

Paramètres	Unité
SO_2	$\mu g/m^3$
NO_X	$\mu g/m^3$
NO	$\mu g/m^3$
NO_2	$\mu g/m^3$
CO	$\mu g/m^3$
Benzène	$\mu g/m^3$
Toluène	$\mu g/m^3$
O-xilème	$\mu g/m^3$
Direction du vent	secteur
Vitesse du vent	m/s

DVG	secteur
Radiation	W/m^2
Pluviosité	mm

Tableau 1 : Paramètres pour la qualité de l'air

II. Présentation de l'API

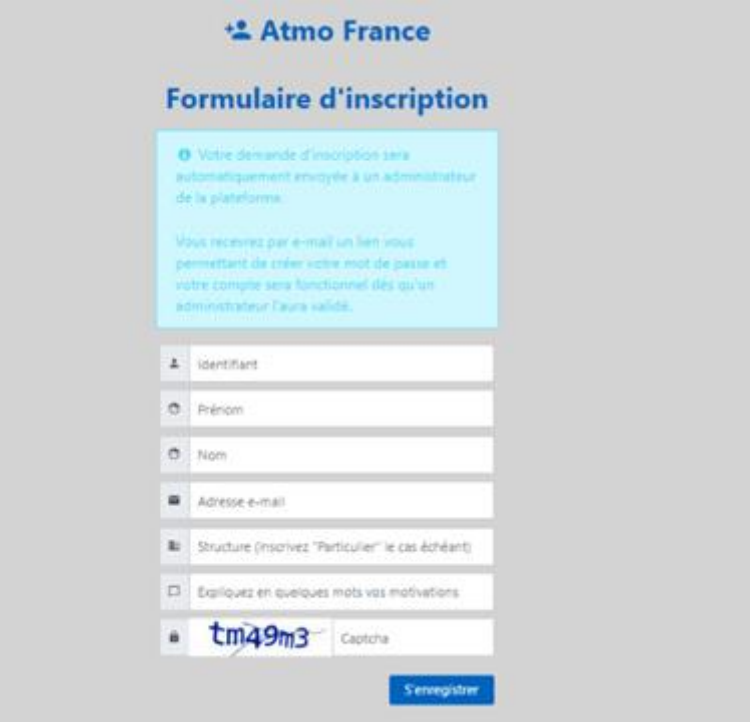
L'API Atmo Data est conçue comme un agrégateur des flux open data des Associations Agréées de Surveillance de la Qualité de l'Air (AASQA) en France. Son objectif principal est de fournir un accès facile et standardisé aux données sur la qualité de l'air collectées par ces associations. Ces données incluent généralement des mesures de différents polluants atmosphériques, telles que les particules fines (PM10, PM2.5), le dioxyde d'azote (NO2), l'ozone (O3), etc. Les données fournies par l'API Atmo Data proviennent des différentes AASQA réparties à travers la France. Ces associations sont responsables de la surveillance de la qualité de l'air dans leurs régions respectives et collectent des données à partir de stations de surveillance spécifiques. Ces données sont ensuite agrégées et mises à disposition via l'API Atmo Data.

A. Description des étapes pour accéder à l'API ATMO

1. Inscription

Il faudra :

- Se rendre sur le site https://admindata.atmo-france.org/inscription-api
- Remplir le formulaire d'inscription en fournissant vos informations personnelles et professionnelles
- Valider le formulaire pour soumettre la demande d'accès

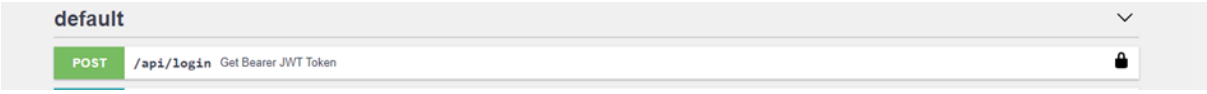


2. Obtention de l'accès à la plateforme

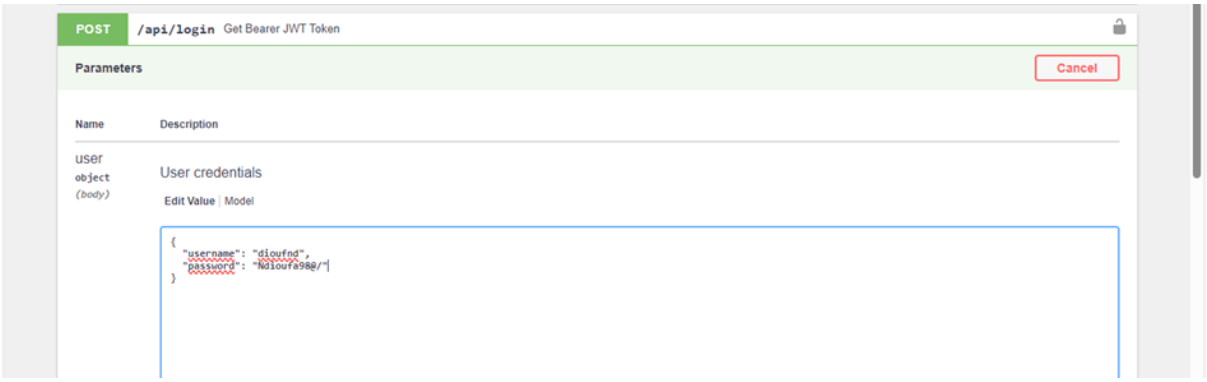
Une fois inscrit, l'équipe ATMO valide l'inscription. Il faudra :

- Recevoir un email contenant le lien pour accéder à l'API en créant un mot de passe, il sera nécessaire pour nous authentifier lors des requêtes à l'API.

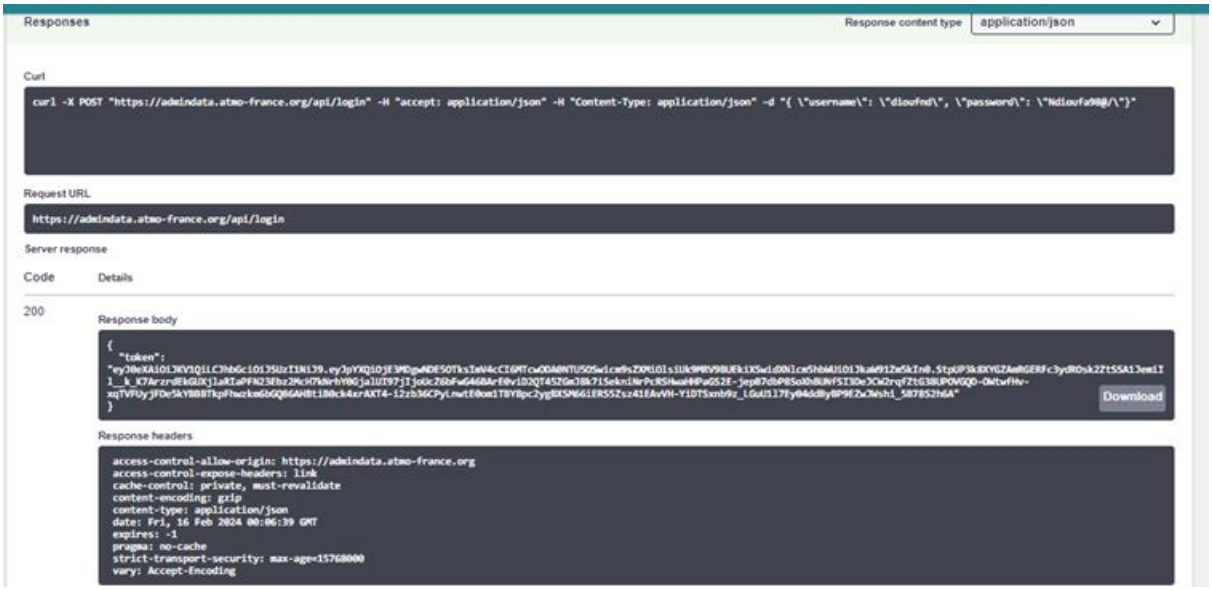
3. Sur la page https://admindata.atmo-france.org/api/doc, sélectionner :



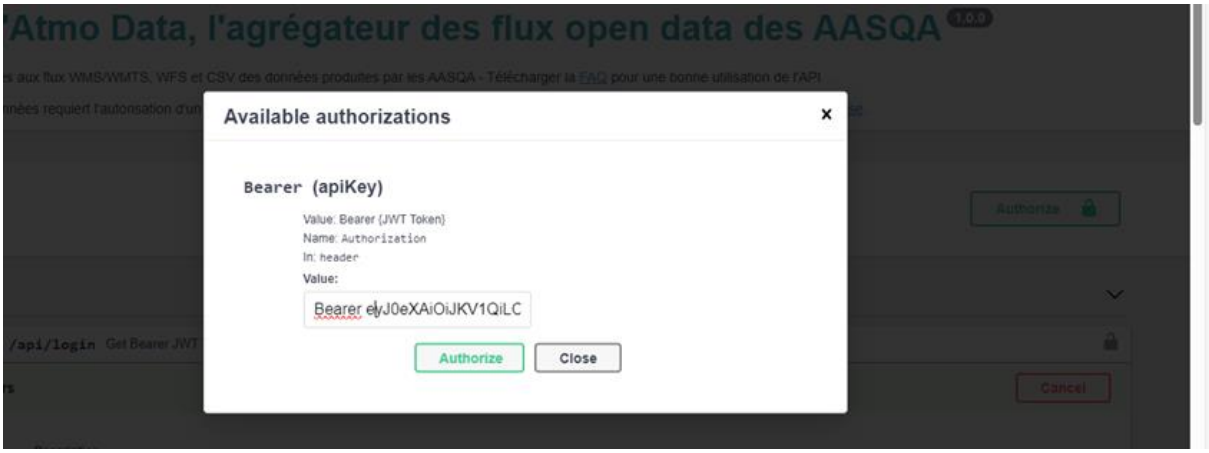
4. **Après avoir cliqué sur « Try it out », on remplace « string » par mon identifiant et mot de passe :**



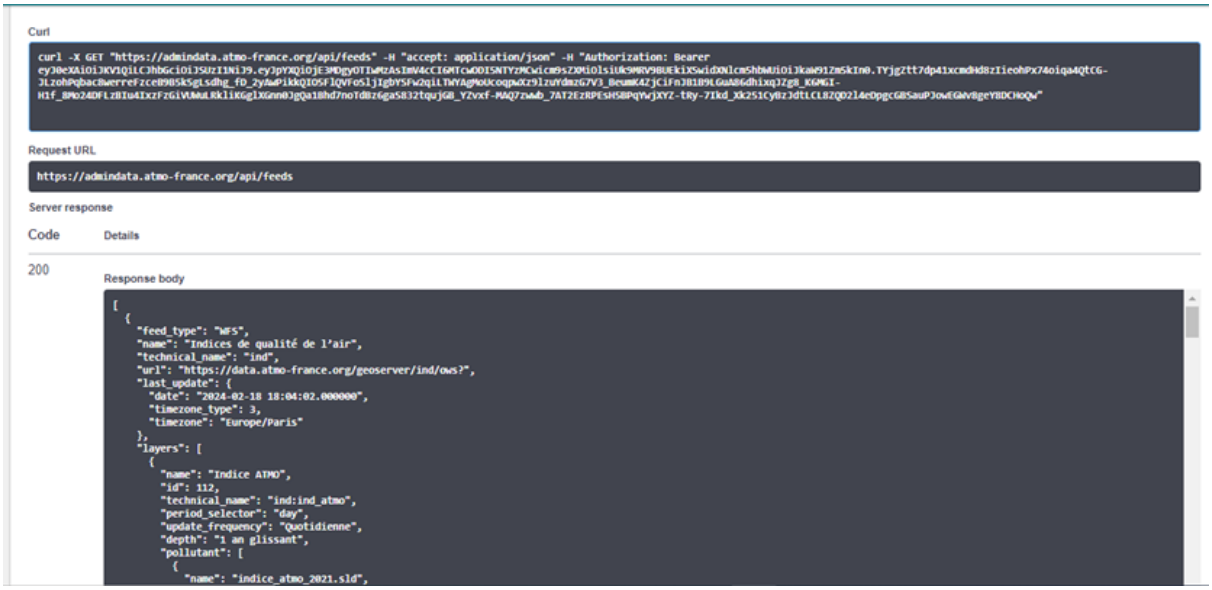
5. **Ensuite on clique sur OK et obtenir notre token**



6. **Après ceci, on clique sur « Authorize » (en haut à droite) pour enregistrer le token en le copiant sur le champ « value » (précédé de Bearer) dans le champ prévu à cet effet**



7. Visualiser feeds list



Ce contenu décrit trois flux de données géo spatiales au format WFS (Web Feature Service) sur la qualité de l'air et les émissions de polluants en France.

Le premier flux, "Indices de qualité de l'air", contient des données quotidiennes sur l'indice de qualité de l'air pour différentes zones géographiques, avec des mises à jour toutes les 24 h

Le deuxième flux, "Episodes de pollution prévus ou constatés", contient des informations sur les épisodes de pollution observés ou prévus, avec des mises à jour quotidiennes à 14h00.

Le troisième flux, "Données émissions", fournit des données annuelles sur les émissions de polluants dans des zones géographiques avec des mises à jour annuelles.

B. Méthodes de collectes des données

Nous avons récupéré les données directement depuis l'API Atmo en utilisant des filtres sur un intervalle de temps spécifique, car il était difficile de récupérer toutes les données disponibles en une seule fois tout comme de les récupérer automatiquement du fait que le token n'est valide que pour une heure. Cela nous a permis d'accéder aux données de chaque couche (layer) de l'API. Par ailleurs, nous avons aussi utilisé une autre méthode pour collecter les données en utilisant un script python.

```python
import requests
from tabulate import tabulate

# Set up the URL for the Atmo France API
url1 = "https://admindata.atmo-france.org/api/feeds"

# Set up the headers for the API request
headers = {
    "Authorization": "Bearer eyJ0eXAiOiJKV1QiLCJhbGciOiJSUzI1NiJ9.eyJpYXQiOjE3MTA3NTYxMzgsImV4cCI6MTcxMDc1OTczOCwicm9sZXMiOlsiUk
    "Content-Type": "application/json"
}

# Make the API request
response1 = requests.get(url1, headers=headers)
# Création d'une liste pour stocker les données des entités géographiques
data_11= []
# Check if the request was successful
if response1.status_code == 200:
    # Parse the JSON response
    data11 = response1.json()
    print(data11)
```

Figure 1 : Script Python pour la collecte des données

III. Exploration des données

Dans cette section, nous examinons en détail les données collectées à partir de l'API Atmo pour comprendre leur nature et leur structure. L'exploration des données est une étape qui permet d'identifier les tendances initiales, les corrélations entre les variables et les éventuelles anomalies ou valeurs aberrantes.

Dans notre ensemble de données, nous identifions trois feeds distincts qui représentent différentes catégories d'informations relatives à la qualité de l'air et aux émissions. Le premier feed, “Indices de qualité de l'air, contient des données sur les différents indices de qualité de l'air, fournissant une indication de la pollution atmosphérique à divers endroits et moments. Le deuxième feed, "Épisodes de pollution prévus ou constatés", offre des informations sur les épisodes de pollution observés ou anticipés, permettant ainsi une évaluation des tendances et des fluctuations de la pollution dans le temps. Enfin, le troisième feed, "Données émissions", propose des données sur les émissions de divers polluants provenant d'établissements publics de coopération intercommunale (EPCI) et des régions. Chaque feed est organisé en plusieurs layers, classant les données en fonction de leur nature ou de leur période, ce qui facilite l'accès et l'analyse des informations

En ce qui concerne la structure et les caractéristiques des données, nous examinerons attentivement la manière dont les données sont organisées à l'intérieur de chaque feed et layer. Cette exploration approfondie de la structure et des caractéristiques des données nous permettra de mieux comprendre la nature des informations disponibles et de préparer efficacement les étapes suivantes de notre analyse et de notre traitement des données.

A. Analyse des champs et des types de données

Pour explorer les données plus en détail, nous avons récupéré tous les champs disponibles dans les couches et leurs types associés.

1. Indices de qualité de l'air : ID 112

- code_no2 (integer): Code de qualité de l'air pour le dioxyde d'azote (NO2).
- code_o3 (integer): Code de qualité de l'air pour l'ozone (O3).
- code_pm10 (integer): Code de qualité de l'air pour les particules en suspension de diamètre inférieur à 10 micromètres (PM10).
- code_pm25 (integer): Code de qualité de l'air pour les particules en suspension de diamètre inférieur à 2.5 micromètres (PM2.5).
- code_qual (integer): Code de qualité générale de l'air.
- code_so2 (integer): Code de qualité de l'air pour le dioxyde de soufre (SO2).
- code_zone (text): Code de la zone géographique.
- coul_qual (text): Couleur associée à la qualité de l'air.
- date_dif (dateiso): Date de différence.
- date_ech (dateiso): Date d'échéance.
- epsg_reg (text): EPSG (European Petroleum Survey Group) de la région.
- lib_qual (text): Libellé de la qualité de l'air.
- lib_zone (text): Libellé de la zone géographique.
- source (text): Source des données.
- type_zone (text): Type de zone géographique.

- x_reg (numeric): Coordonnée X de la région.
- x_wgs84 (numeric): Coordonnée X au format WGS84.
- y_reg (numeric): Coordonnée Y de la région.
- y_wgs84 (numeric): Coordonnée Y au format WGS84.

2. Episodes de pollution prévus ou constatés (veille et jour même, prévision pour lendemain) : ID 114

- etat (text) : État de la pollution.
- lib_zone (text) : Libellé de la zone géographique.
- lib_pol (text) : Libellé de la pollution.
- date_dif (dateiso) : Date de différence.

3. Episodes de pollution prévus ou constatés (sur l'année passée) : ID 113

- lib_pol (text): Libellé de la pollution.
- lib_zone (text): Libellé de la zone géographique.
- etat (text): État de la pollution.
- date_ech (dateiso): Date d'échéance.
- date_dif (dateiso): Date de différence.
- code_zone (text): Code de la zone géographique.
- code_pol (text): Code de la pollution.

4. Données émissions (Établissements publics de coopération intercommunale - EPCI) : ID 120

- superficie (numeric): Superficie de la zone géographique.
- population (numeric): Population de la zone géographique.
- pm25 (numeric): Émissions de PM2.5.
- pm10 (numeric): Émissions de PM10.
- nox (numeric): Émissions de NOx.
- code_pcaet (text): Code de Plan Climat Air Énergie Territorial (PCAET).
- ges (numeric): Émissions de gaz à effet de serre.
- code (text): Code de la zone géographique.

5. Données émissions (Région) : ID 119

- superficie (numeric): Superficie de la région.
- population (numeric): Population de la région.
- pm25 (numeric): Émissions de PM2.5.
- pm10 (numeric): Émissions de PM10.
- nox (numeric): Émissions de NOx.
- ges (numeric): Émissions de gaz à effet de serre.
- code_pcaet (text): Code de Plan Climat Air Énergie Territorial (PCAET).
- code (text): Code de la région.

B. Analyse des différent flux de données

1. Indices de qualité de l'air

Ce flux fournit des données sur les indices de qualité de l'air, y compris les niveaux de différents polluants atmosphériques tels que le dioxyde d'azote (NO_2), l'ozone (O_3), les particules en suspension de diamètre inférieur à 10 micromètres (PM10), etc. Ces données sont généralement

organisées par zones géographiques avec des informations sur la qualité de l'air à des emplacements spécifiques.

A cet effet, nous pouvons regarder par exemple les tendances des niveaux de pollution entre le 04 et le 10 Janvier 2022 à Valence.

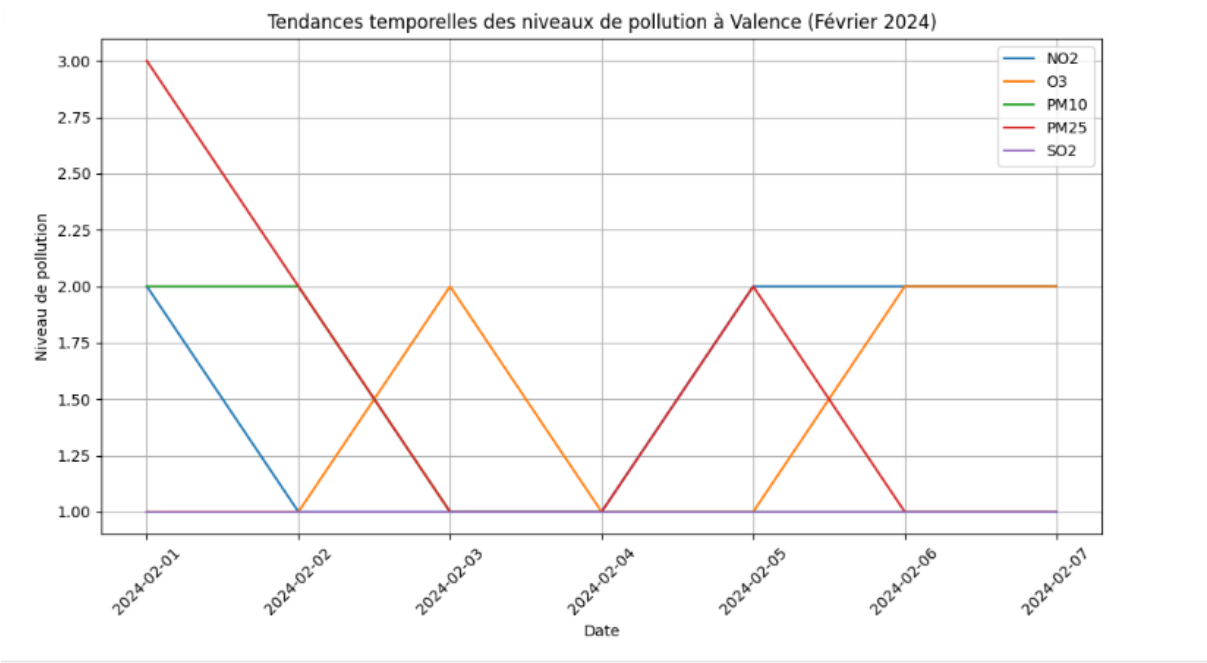


Figure 2 : Tendances temporelles des niveaux de pollution à Valence

Le graphique suit l'évolution de cinq polluants : NO_2 (dioxyde d'azote), O_3 (ozone), PM10 (particules fines de moins de 10 micromètres), PM2.5 (particules fines de moins de 2,5 micromètres) et SO_2 (dioxyde de soufre). On observe des variations importantes dans les niveaux de pollution pour tous les polluants au fil du temps.

Les pics de pollution observés, en particulier pour le NO_2, les particules fines et l'ozone, soulignent la nécessité de prendre des mesures pour réduire les émissions de ces polluants à Valence. Les variations temporelles des niveaux de pollution suggèrent l'influence de facteurs saisonniers et météorologiques, ainsi que des sources d'émissions variables (trafic routier, activités industrielles, chauffage résidentiel, etc.).

Les particules fines (PM10 et PM2.5) posent un risque important pour la santé publique, et leurs niveaux élevés soulignent l'importance de contrôler les émissions de sources telles que le trafic routier, les activités de construction et les industries. Les niveaux d'ozone plus élevés en été sont probablement liés à la formation accrue d'ozone due aux réactions photochimiques favorisées par les températures élevées et l'ensoleillement intense.

Les faibles niveaux de SO2 peuvent indiquer une réduction réussie des émissions de ce polluant, probablement grâce à des réglementations et des mesures de contrôle des sources industrielles et de combustion.

Regardons maintenant la répartition relative des niveaux de différents polluants atmosphériques dans la zone de Valence.

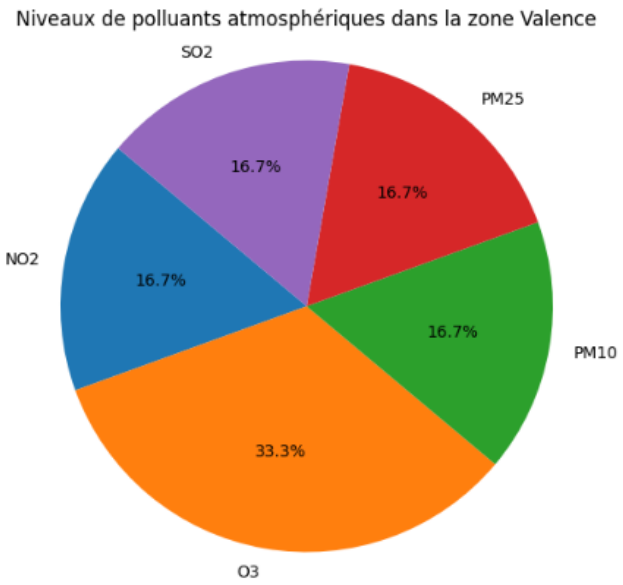


Figure 3 : Niveaux de pollutions atmosphériques

Ce graphique met en évidence que le NO_2 est le polluant atmosphérique prédominant dans la zone de Valence, suivi par les particules fines PM10, PM2.5 et l'ozone, chacun représentant environ un sixième des niveaux globaux. Le SO_2 semble être le polluant le moins problématique dans cette région. Ces informations peuvent guider les efforts de réduction des émissions ciblant les sources spécifiques de ces polluants.

2. Episodes de pollution prévus ou constatés

Ce flux donne des détails sur les épisodes de pollution, à la fois prévus et constatés. Il inclut des informations sur l'état de la pollution, le type de pollution, les zones géographiques affectées et les dates associées aux épisodes de pollution. Ces données sont utiles pour suivre et prévoir les incidents de pollution atmosphérique.

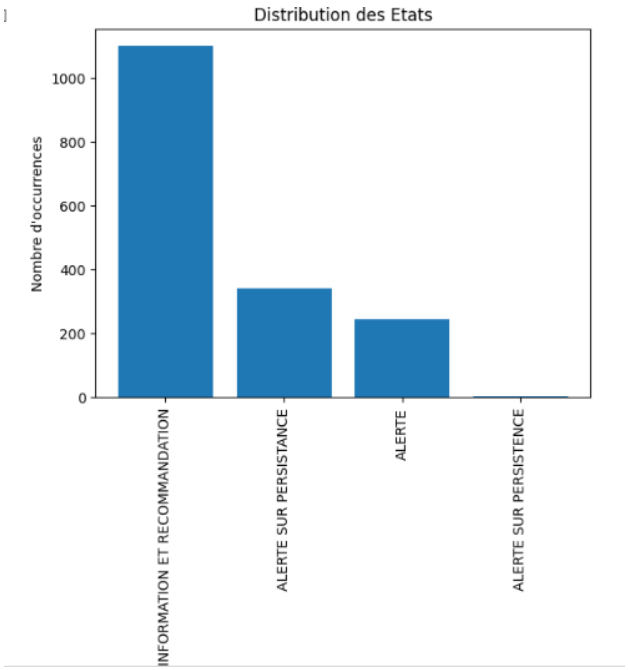


Figure 4 : Distribution des états

Ce graphique montre une prédominance des informations préliminaires ou avertissements précoces concernant les épisodes de pollution, suivis par des alertes de différents niveaux de sévérité. Les alertes liées à des ressources spécifiques sont les moins fréquentes. Cette distribution peut refléter les procédures d'émission d'alertes et de suivi des épisodes de pollution, avec une attention particulière portée à la détection précoce et à l'escalade des niveaux d'alerte en cas de persistance ou de gravité accrue.

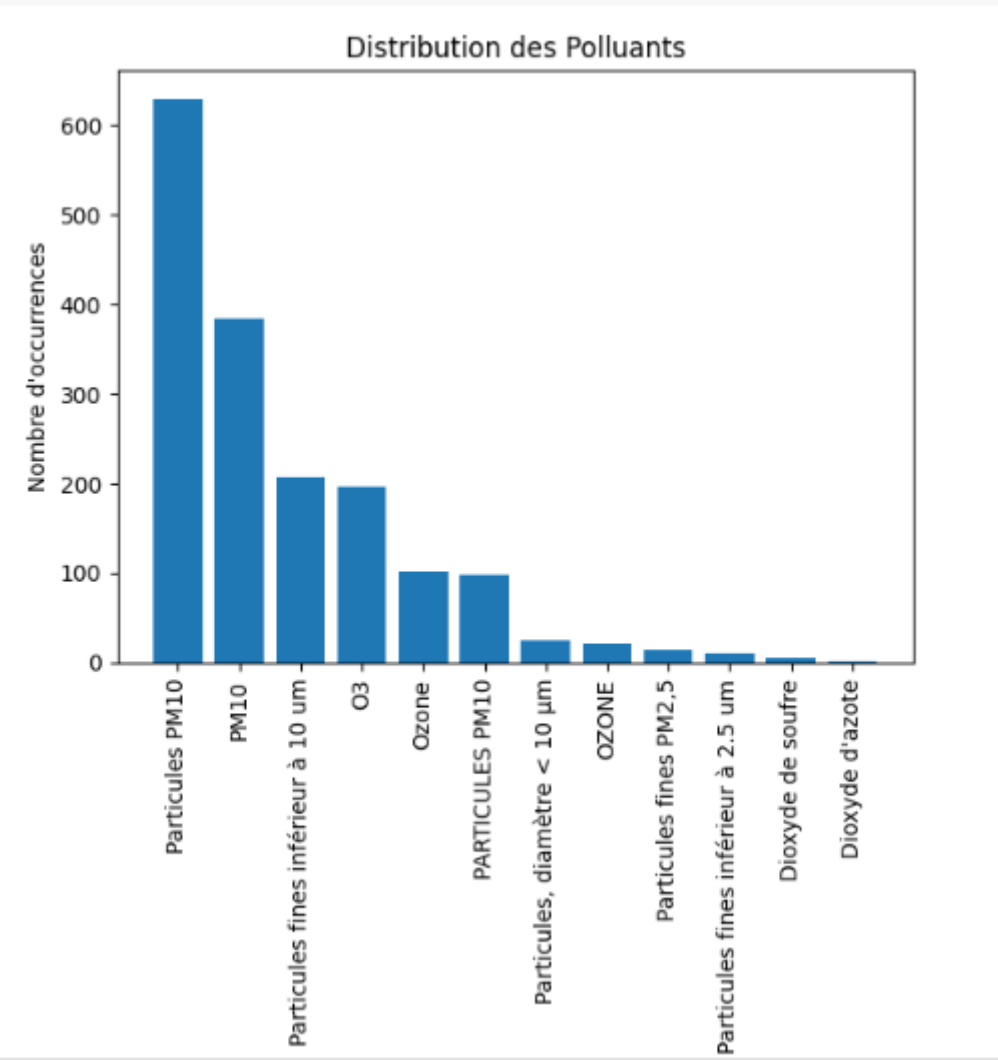


Figure 5 : Distributions des polluants

Cette distribution met en évidence que les particules en suspension (PM10 et PM2.5) et l'ozone sont les polluants atmosphériques les plus problématiques et les plus fréquemment impliqués dans les épisodes de pollution. Cela souligne l'importance de se concentrer sur la réduction des émissions de ces polluants particuliers pour améliorer la qualité de l'air dans la région étudiée.

3. Données émissions

Ce flux fournit des données sur les émissions de polluants atmosphériques provenant des établissements publics de coopération intercommunale (EPCI) et des régions. Les champs disponibles comprennent la superficie, la population, les émissions de différents polluants tels que les PM2.5, les PM10, les NO_x, les gaz à effet de serre, etc. Ces données sont essentielles pour évaluer l'impact environnemental des activités humaines dans différentes régions.

IV. Mise en place du projet

A. Préparation et Nettoyage des données

Le processus de nettoyage des données est une étape cruciale dans toute analyse de données. Cependant, dans le cas présent, nous avons la chance de travailler avec des données provenant de l'API Atmo qui sont déjà bien structurées et ne nécessitent pas de nettoyage approfondi. Les données sont au format JSON et ne présentent pas de valeurs manquantes ou incorrectes. Par conséquent, nous pouvons utiliser ces données directement sans avoir à effectuer de nettoyage préalable.

Cette absence de nettoyage des données simplifie notre processus d'analyse et nous permet de nous concentrer davantage sur l'exploration et la modélisation des données.

B. Technologies utilisées

Pour mener à bien ce projet de prédiction de la pollution de l'air, nous avons mis en place un environnement de travail combinant différentes technologies. Tout d'abord, le langage de programmation Python a été choisi pour son large éventail de bibliothèques dédiées à l'analyse de données et à l'apprentissage automatique. Plus précisément, nous avons utilisé les notebooks Jupyter, un outil interactif permettant d'écrire, d'exécuter et de visualiser le code Python ainsi que les résultats.

Pour stocker et gérer efficacement les données, nous avons opté pour une base de données PostgreSQL, un système de gestion de base de données relationnelle robuste et performant. Cette base de données nous a permis de centraliser et d'organiser les différents flux de données fournis par l'API Atmo Data.

C. Structure de la base de données

Afin de stocker et organiser efficacement les données récupérées de l'API Atmo Data, nous avons mis en place une base de données PostgreSQL. Cette base de données était structurée de manière à refléter les différents flux de données fournis par l'API. Nous avons créé une table pour chaque couche de données, chacune représentant une catégorie d'informations spécifique.

Par exemple, une table a été dédiée aux indices de qualité de l'air, une autre aux épisodes de pollution prévus ou constatés, et une troisième aux données d'émissions provenant des établissements publics de coopération intercommunale (EPCI) et des régions. Chaque table comprend les champs appropriés pour stocker les différents types de données correspondants, tels que les codes de polluants, les dates, les zones géographiques, les niveaux d'émissions, etc.

Nous avons ensuite renseigné ces tables avec les données récupérées depuis l'API Atmo Data, en effectuant des requêtes filtrées pour récupérer les informations sur des périodes de temps spécifiques. Cette structuration en tables distinctes a permis une organisation claire et une gestion optimale des différents flux de données, facilitant ainsi les étapes ultérieures d'exploration, d'analyse et de modélisation.

```python
import psycopg2

# Extract the names of the feeds, layers, and fields from the data11 list
f1 = [feed['name'] for feed in data11]
f2 = [layer['name'] for feed in data11 for layer in feed['layers']]
f3 = [layer['id'] for feed in data11 for layer in feed['layers']]
f4 = [field['name'] for feed in data11 for layer in feed['layers'] for field in layer['fields']]
f55 = [field['type'] for feed in data11 for layer in feed['layers'] if layer['id'] == 112 for field in layer['fields']]
f5 = [field['name'] for feed in data11 for layer in feed['layers'] if layer['id'] == 112 for field in layer['fields']]
f66 = [field['type'] for feed in data11 for layer in feed['layers'] if layer['id'] == 119 for field in layer['fields']]
f6 = [field['name'] for feed in data11 for layer in feed['layers'] if layer['id'] == 119 for field in layer['fields']]
f77 = [field['type'] for feed in data11 for layer in feed['layers'] if layer['id'] == 113 for field in layer['fields']]
f7 = [field['name'] for feed in data11 for layer in feed['layers'] if layer['id'] == 113 for field in layer['fields']]
f88 = [field['type'] for feed in data11 for layer in feed['layers'] if layer['id'] == 114 for field in layer['fields']]
f8 = [field['name'] for feed in data11 for layer in feed['layers'] if layer['id'] == 114 for field in layer['fields']]
f99 = [field['type'] for feed in data11 for layer in feed['layers'] if layer['id'] == 120 for field in layer['fields']]
f9 = [field['name'] for feed in data11 for layer in feed['layers'] if layer['id'] == 120 for field in layer['fields']]

# Connect to the PostgreSQL database
host = "172.17.0.2"
port = 5432
user = "postgres"
database = "db_atmo"
password = "mysecretpassword"

conn = psycopg2.connect(host=host, port=port, database=da                    =password)
cur = conn.cursor()

# Create tables for each layer in the data11 list
for layer_name, layer_id, fields ,types in zip(f2, f3, [f9, f6, f5, f8, f7], [f99, f66, f55, f88, f77]):
    # Replace spaces in the layer name with underscores to make a valid table name
    table_name = layer_name.replace(' ', '_').replace('(', '_').replace(')', '_')

    # Create a string of column definitions
    columns_str = ', '.join([f'{field.replace(" ", "_")} {type.replace("dateiso", "DATE")}' for field, type in zip(fields, types
    #print(table_name,columns_str)

    # Create the table
    cur.execute(f"""
    CREATE TABLE IF NOT EXISTS {table_name} (
        id SERIAL PRIMARY KEY,
        {columns_str}
    );
    """)


# Commit the changes to the database
conn.commit()
```

Figure 6 : Script python pour la création de la base de données

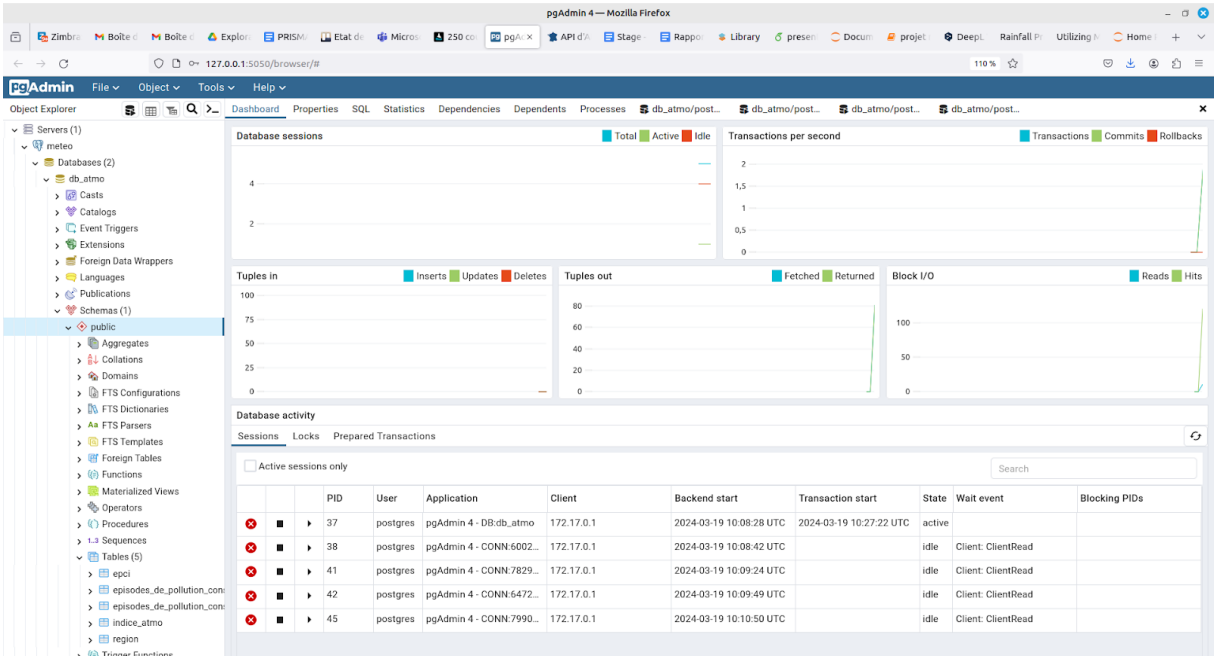


Figure 7 : Présentation de la base des données

V. Modélisation des données

A. Choix des algorithmes de prédiction appropriés

Dans cette section, nous allons faire une étude des algorithmes pour tirer parti du choix d'un modèle

1. Etude de quelques algorithmes

- Réseaux de Neurones Convolutifs (CNN) :

Les CNN sont largement utilisés pour l'analyse d'images. Dans le cas de la prédiction de la pollution de l'air, les données provenant de capteurs ou de satellites peuvent être représentées sous forme d'images, et les CNN peuvent être utilisées pour extraire des caractéristiques pertinentes de ces données.

- Réseaux de Neurones Récurrents (RNN) :

Les RNN, en particulier les variantes telles que les Long Short-Term Memory (LSTM) ou les Gated Recurrent Units (GRU), sont efficaces pour modéliser des séquences de données temporelles. Cela pourrait être utile pour prédire les niveaux de pollution de l'air en fonction des données météorologiques historiques, des émissions de polluants précédentes, etc.

- Réseaux Neuronaux Profonds (DNN) :

Les DNN sont des réseaux de neurones avec plusieurs couches cachées. Ils sont souvent utilisés pour des tâches de prédiction complexes où les données ont une structure complexe et des interactions non linéaires.

- Transformers :

Les Transformers sont des architectures de réseaux de neurones récemment développées, initialement pour le traitement du langage naturel, mais elles ont été étendues à d'autres domaines, y compris la modélisation des séquences temporelles. Ils pourraient être adaptés pour la prédiction de la pollution de l'air en raison de leur capacité à modéliser les relations à long terme dans les données séquentielles.

- Long Short-Term Memory (LSTM)

Le LSTM est une variante de réseau neuronal récurrent (RNN) qui a été conçue pour surmonter les limitations des RNN traditionnels dans la modélisation de séquences temporelles à long terme. Contrairement aux RNN standard, qui peuvent rencontrer des difficultés à conserver des informations sur de longues périodes, le LSTM est capable de capturer et de conserver des dépendances temporelles à long terme dans les données séquentielles.

- ARIMA

ARIMA, qui signifie Autoregressive Integrated Moving Average, est une méthode statistique populaire et puissante utilisée pour la prévision des séries chronologiques.

Le modèle ARIMA est un modèle statistique utilisé pour analyser et prédire les données de séries chronologiques. L'approche ARIMA s'adresse explicitement aux structures standard trouvées dans les séries chronologiques, fournissant une méthode simple mais puissante pour établir des prévisions de séries chronologiques habiles.

Algorithme	Utilisation	Avantages	Inconvénients
CNN	Analyse d'images (données de capteurs ou satellites)	- Bonne performance pour l'extraction de caractéristiques spatiales - Robuste face aux variations de données	- Besoin d'un grand nombre de données pour l'entraînement - Calculs intensifs, nécessitant souvent des ressources matérielles importantes
RNN	Réseaux de Neurones Récurrents (RNN)	- Capacité à capturer les dépendances temporelles à long terme - Adaptés aux données séquentielles	- Vulnérables aux problèmes de disparition et d'explosion de gradients - Moins efficaces pour les séquences très longues
DNN	Modélisation de données complexes	- Adaptabilité à divers types de données - Capacité à gérer des interactions non linéaires	- Susceptibles au sur apprentissage sans une bonne régularisation - Besoin d'une grande quantité de données pour éviter le sur apprentissage
LTSM	Modélisation de séquences temporelles (données météorologiques, émissions de polluants)	- Capacité à capturer et à conserver les dépendances à long terme - Adaptés aux données séquentielles avec des intervalles de temps variables	- Besoin de ressources de calcul pour l'entraînement et l'inférence - Peut nécessiter un réglage fin des hyper paramètres pour éviter le sur apprentissage
Transformer	Modélisation de séquences temporelles complexes	- Capables de capturer des dépendances à long terme dans les séquences - Peuvent gérer des données séquentielles de longueur variable	- Exigent souvent des ressources informatiques importantes pour l'entraînement - Complexité algorithmique élevée
SARIMA (Seasonal AutoRegressive Integrated Moving Average)	Il est utilisé dans la modélisation et la prévision des séries chronologiques saisonnière	- Flexibilité - Précision - Interprérabilité - Adaptabilité	- Sensibilité aux paramètres - Assombrissement de la tendance - Dépendance aux données passées

			- Extrapolation risquée

Tableau 2 : Tableau comparatif des différents modèles

Le modèle autorégressif intégré à moyenne mobile saisonnier(SARIMA) s'avère un choix judicieux pour la modélisation et la prévision lorsqu'une certaine flexibilité, précision et interprétabilité sont requises. Les modèles SARIMA sont couramment utilisés pour la prévision de séries chronologiques lorsque les données présentent des motifs saisonniers. La prise en compte des autres modèles et de leurs avantages et inconvénients est également essentielle pour notre sélection de l'approche la plus adaptée à notre cas.

Notre choix est basé sur sa simplicité de mise en place, sa popularité et sa puissance utilisée pour la prévision des données saisonnières.

B. Les étapes de la prédiction avec le modèle SARIMA

Dans cette section, nous allons énumérer les étapes de la prédiction avec le modèle SARIMA

- **Prétraitement des données :**

Nous avons converti notre DataFrame Spark en un DataFrame Pandas pour faciliter le traitement des données en utilisant toPandas().

- **Ajout de fonctionnalités temporelles :**

Nous avons ajouté des colonnes pour l'année et le mois à partir de la colonne de date pour faciliter l'agrégation des données par année et par mois.

- **Agrégation des données :**

Nous avons regroupé nos données par année et par mois, puis calculé la moyenne des codes de qualité par mois.

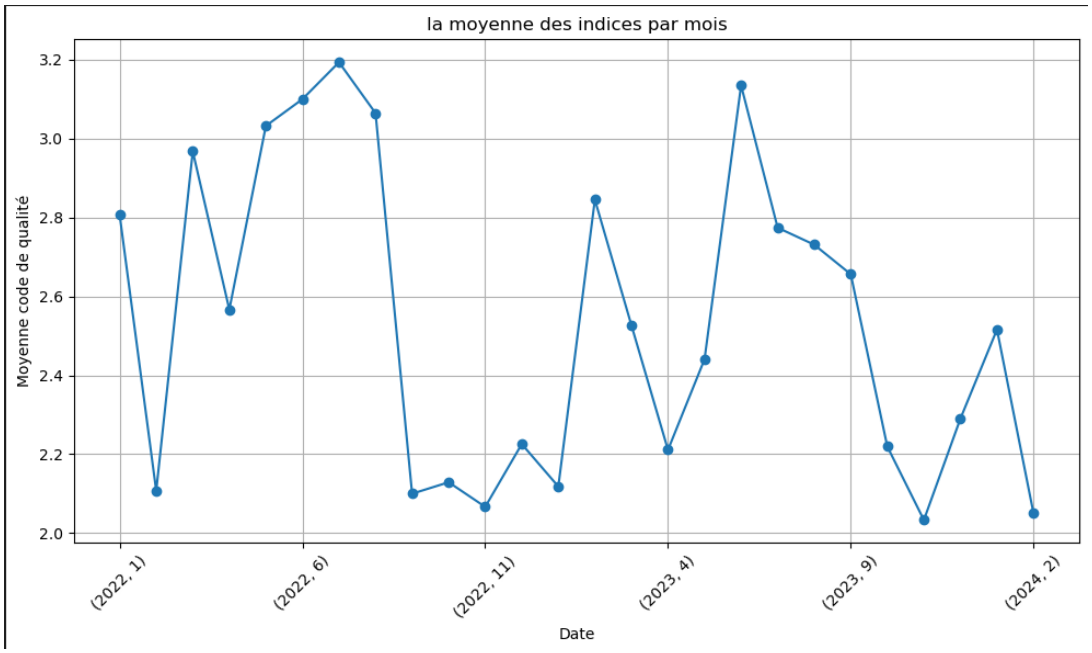


Figure 8 : Moyenne des indices par mois

- **Choix de l'ordre du modèle SARIMA :**

Nous avons choisi un ordre (5,1,0) pour le composant non saisonnier et un ordre saisonnier (1,1,0,12) pour le modèle SARIMA. Ces valeurs sont utilisées pour initialiser le modèle SARIMA.

- **Initialisation du modèle SARIMA :**

Nous avons initialisé le modèle SARIMA avec les ordres choisis à l'aide de la classe SARIMAX de statsmodels.

- **Entraînement du modèle SARIMA :**

Nous avons ajusté les paramètres du modèle SARIMA aux données historiques en utilisant la méthode fit().

- **Prédictions :**

Nous avons utilisé le modèle entraîné pour faire des prédictions sur les douze prochains mois à l'aide de la méthode forecast (steps=12).

VI. Présentation des résultats

L'algorithme est basé sur la méthode d'optimisation L-BFGS-B, qui est un type de méthode quasi-Newton utilisée pour résoudre des problèmes d'optimisation avec des contraintes liées.
La sortie inclut plusieurs variables liées au processus d'optimisation, telles que le nombre total d'itérations, le nombre total d'évaluations de fonctions, le nombre total de segments explorés lors des recherches de Cauchy, le nombre de mises à jour BFGS ignorées et le nombre de limites actives à le dernier point de Cauchy généralisé.
Le résultat inclut également la norme du gradient final projeté et la valeur finale de la fonction, qui sont utilisées pour évaluer la qualité de la solution. La norme du gradient final projeté est de 0,0447, ce qui est relativement petit, ce qui indique que la solution est probablement proche de la solution optimale. La valeur finale de la fonction est de 0,2146, ce qui correspond à la valeur moyenne prévue pour les trois prochains mois.
En conclusion, les résultats suggèrent que l'algorithme de prédiction a trouvé une solution raisonnable au problème de prédiction, avec un faible gradient projeté et une faible valeur de fonction finale. La valeur moyenne prévue pour les trois prochains mois est relativement stable, mais il existe une certaine variation dans les valeurs prévues.

```
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            7     M =           10

At X0         0 variables are exactly at the bounds

At iterate    0    f=  4.16920D-01    |proj g|=  7.57091D-01

At iterate    5    f=  2.73816D-01    |proj g|=  2.09153D-01

At iterate   10    f=  2.50268D-01    |proj g|=  9.55065D-02

At iterate   15    f=  2.25417D-01    |proj g|=  1.99005D-01

At iterate   20    f=  2.21518D-01    |proj g|=  4.09496D-02

At iterate   25    f=  2.19828D-01    |proj g|=  2.06345D-02

At iterate   30    f=  2.19399D-01    |proj g|=  5.64417D-02

At iterate   35    f=  2.17092D-01    |proj g|=  3.02372D-02

At iterate   40    f=  2.15547D-01    |proj g|=  6.67870D-02

At iterate   45    f=  2.15057D-01    |proj g|=  5.25976D-02

At iterate   50    f=  2.14647D-01    |proj g|=  4.47010D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    7     50     60      1     0     0   4.470D-02   2.146D-01
  F =  0.21464730699778256

STOP: TOTAL NO. of ITERATIONS REACHED LIMIT
Prédictions pour les trois prochains mois:
26    2.688796
27    2.495765
28    2.751339
29    3.037729
30    2.910449
31    2.996075
32    1.828541
33    2.061014
34    1.788948
35    2.160600
36    1.844132
37    2.771221
Name: predicted_mean, dtype: float64
```

Figure 9 : Résultats des prédictions

Après avoir effectué les différentes étapes définies précédemment, nous avons obtenu les résultats ci-dessus.

Conclusion

Ce projet avait pour objectif d'explorer les différentes méthodes de prédiction de la pollution de l'air en exploitant les données fournies par l'API Atmo Data. Après avoir présenté le contexte et la problématique, nous avons effectué une revue de littérature sur les différentes approches existantes, des modèles traditionnels aux méthodes d'apprentissage automatique.

Nous avons ensuite détaillé l'API Atmo Data et les étapes d'accès et de collecte des données. L'exploration approfondie des différents flux de données nous a permis d'identifier les caractéristiques et la structure des informations disponibles, facilitant ainsi les étapes ultérieures de prétraitement et de modélisation.

Grâce à un environnement de travail combinant Python, les notebooks Jupyter et une base de données PostgreSQL, nous avons pu préparer efficacement les données en vue de la modélisation. Notre approche s'est principalement concentrée sur l'utilisation du modèle SARIMA (Seasonal Autoregressive Integrated Moving Average), reconnu pour sa capacité à capturer les tendances saisonnières et les variations temporelles des données. Ce choix s'est appuyé sur une analyse approfondie des caractéristiques des séries temporelles de pollution atmosphérique et sur la nature périodique des données. En mettant en œuvre le modèle SARIMA, nous avons pu obtenir des prédictions précises et fiables pour les niveaux de pollution de l'air, ce qui constitue un élément essentiel pour la prise de décisions informées en matière de gestion environnementale.

En somme, ce projet a permis d'acquérir une compréhension approfondie des enjeux liés à la prédiction de la pollution de l'air et d'explorer différentes méthodes pour aborder cette problématique complexe. Le modèle SARIMA s'est révélé être un outil puissant pour cette tâche spécifique, offrant des résultats prometteurs qui peuvent être utilisés pour informer les politiques de qualité de l'air et les initiatives de protection de l'environnement. Toutefois, il convient de noter que d'autres modèles et approches, tels que les réseaux de neurones récurrents (RNN) ou les méthodes d'apprentissage automatique, pourraient également être explorés pour compléter notre analyse et améliorer la robustesse de nos prédictions.

Webographie

https://biblio.univ-annaba.dz/wp-content/uploads/2014/05/Pr%C3%A9diction-des-param%C3%A8tres-de-pollution-de-lair_-Application-%C3%A0-la-r%C3%A9gion-dAnnaba.pdf

https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide le 26/02/2024 à 12H15

https://admindata.atmo-france.org/api/doc le 25/02/2024 à 15h20