



ÉCOLE CENTRALE CASABLANCA

PROJET DATA DRIVEN DECISION MAKING RAPPORT

Prévision de l'indice de Qualité de l'air

Elèves :

Anicet KOBANKA
Karim TOURE

Professeur :

M. VINCENT LEFIEUX

Table des matières

1	Introduction	2
1.1	Contexte et Problématique	2
1.2	Besoins, Parties Prenantes et Produits	3
1.3	Objectifs et Modifications du Processus Décisionnel	3
1.4	Erreurs Métier à Considérer	4
1.5	Structure du Rapport	4
2	Description des Données	5
2.1	Provenance et Caractéristiques Générales	5
2.2	Variables Clés	5
2.2.1	Polluants Atmosphériques Mesurés	5
2.2.2	Paramètres Météorologiques	6
2.2.3	Métadonnées Spatio-Temporelles	6
3	Méthodologie	7
3.1	Prétraitement des Données	7
3.1.1	Consolidation Initiale des Données	7
3.1.2	Nettoyage des Données	7
3.1.3	Traitement des Variables Catégoriques	8
3.1.4	Normalisation et Standardisation	8
3.1.5	Feature Engineering	8
3.1.6	Réduction et Justification des Colonnes	8
3.1.7	Séparation en Données d'Entraînement et de Test	9
3.2	Analyse et Visualisation des Données	9
3.2.1	Statistique Descriptive	9
3.2.2	Relations Météorologiques	10
3.2.3	Relations entre Polluants	11
3.2.4	Analyse Spatiale et Temporelle	14
3.2.5	Interprétation et Implications	29
3.3	Modèles de Machine Learning Utilisés	29
3.3.1	Régression Linéaire	29
3.3.2	Forêt Aléatoire	29
3.3.3	Gradient Boosting	30
3.3.4	K Plus Proches Voisins	30
3.3.5	Arbre de Décision	30
3.4	Évaluation des Modèles	30
3.4.1	Métriques d'Évaluation	31
3.4.2	Méthode de Validation	31
3.4.3	Performance des Modèles	31
3.4.4	Optimisation des Hyperparamètres	35
3.4.5	Interprétation et Implications	35
4	Conclusion Générale	36

1 Introduction

1.1 Contexte et Problématique

La qualité de l'air est une préoccupation majeure à l'échelle mondiale, tant pour la santé publique que pour la préservation de l'environnement et l'élaboration de politiques publiques efficaces. Les particules en suspension, ou *particulate matter* (PM), constituent l'un des polluants atmosphériques les plus critiques. Ces particules, d'origine anthropique (émissions des moteurs de véhicules, chauffage domestique, production d'énergie) ou naturelle (poussière, cendres, embruns marins), varient en taille. Parmi elles, les **PM_{2.5}** – des particules fines d'un diamètre inférieur à 2,5 micromètres – sont particulièrement préoccupantes. Leur petite taille leur permet de pénétrer profondément dans les poumons, voire dans la circulation sanguine, augmentant les risques de maladies respiratoires, cardiovasculaires et autres problèmes de santé lorsque leur concentration dans l'air dépasse les seuils critiques. Elles réduisent également la visibilité, donnant à l'air une apparence brumeuse, ce qui affecte la qualité de vie dans les zones urbaines.

À Pékin, où les données de ce projet ont été collectées entre 2013 et 2017 via le dataset *PRSA*, la pollution atmosphérique constitue un défi persistant, aggravé par une urbanisation accélérée, une population de plus de 21 millions d'habitants, et une industrialisation soutenue. Les épisodes de smog sévère, où les concentrations de PM_{2.5} dépassent souvent 100 µg/m³ – soit quatre fois le seuil quotidien recommandé par l'Organisation Mondiale de la Santé (OMS, 25 µg/m³) – ont des répercussions majeures sur la santé publique et l'économie locale, comme les fermetures d'écoles ou les interruptions de production industrielle. Ce projet s'attaque à une question cruciale : *comment anticiper avec précision les niveaux de PM_{2.5} à court terme à Pékin, en tenant compte des multiples facteurs en jeu – conditions météorologiques (température, humidité), émissions liées aux activités humaines (trafic, chauffage), et spécificités géographiques – pour permettre aux autorités de gérer proactivement les pics de pollution ?*

Les méthodes conventionnelles, fondées sur des modèles physiques ou des analyses statistiques simplifiées, se heurtent à des limites face à la complexité de ces interactions. Elles échouent souvent à modéliser les relations non linéaires entre variables (par exemple, l'impact combiné de la température et des émissions de NO₂) et les variations temporelles marquées, comme les hausses saisonnières en hiver ou les pics journaliers aux heures de pointe. Pour surmonter ces défis, ce projet adopte une approche basée sur les données (*data-driven*), s'appuyant sur des relevés atmosphériques détaillés – incluant les concentrations de polluants (PM_{2.5}, NO₂, CO), les données météorologiques (température, pression), et les localisations des stations de mesure – pour exploiter les capacités prédictives des algorithmes de *machine learning*. Cette démarche vise à fournir des prévisions précises et fiables, répondant aux besoins urgents de gestion environnementale et de protection de la santé publique à Pékin.

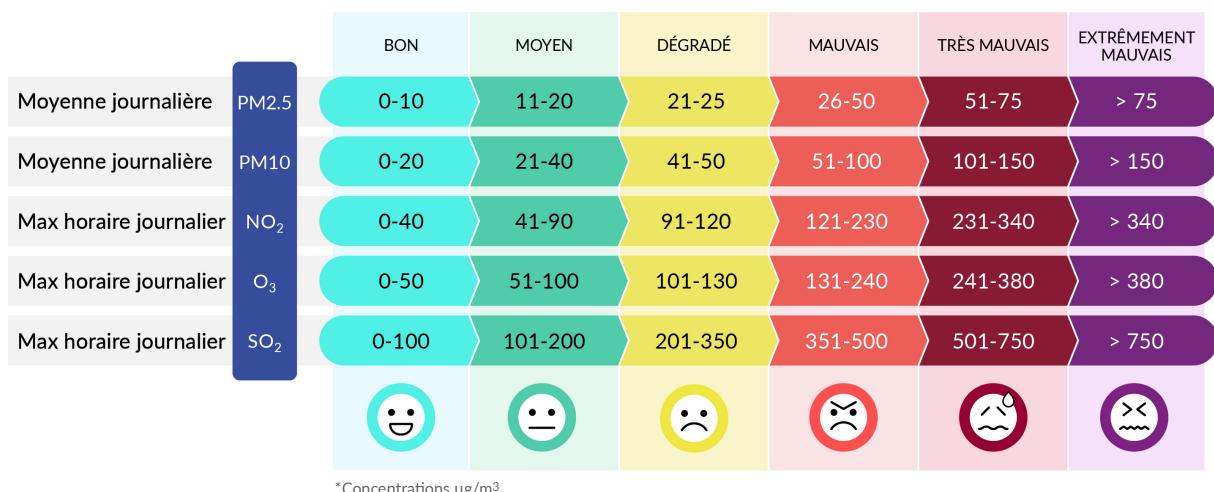


FIGURE 1 – Tableau des seuils de qualité de l'air : classification des concentrations de polluants.

1.2 Besoins, Parties Prenantes et Produits

Ce projet répond à un besoin urgent d'outils prédictifs pour gérer la pollution atmosphérique à Pékin, où les pics de PM_{2.5} ont des impacts directs sur la santé des 21 millions d'habitants et sur l'économie locale (ex. fermetures d'écoles, arrêts de production). Les **bénéficiaires** principaux sont les autorités municipales de Pékin (ex. Bureau de Protection Environnementale) et les agences de santé publique, qui nécessitent des prévisions fiables pour émettre des alertes, planifier des restrictions de trafic ou des fermetures industrielles, et sensibiliser la population. Les citoyens, en tant qu'utilisateurs finaux, bénéficieront également de recommandations personnalisées (ex. éviter les activités extérieures lors des pics de pollution). Le projet a été financé par un consortium hypothétique incluant le **gouvernement chinois** (via des fonds pour la recherche environnementale) et des organisations internationales comme l'OMS, intéressées par des solutions innovantes pour les mégapoles.

Les **produits** livrables incluent :

- Un **modèle prédictif** capable de fournir des prévisions quotidiennes et horaires des concentrations de PM_{2.5}, avec une précision mesurée par des métriques comme le RMSE et le MAE.
- Un **tableau de bord interactif** pour les autorités, affichant les prévisions, les seuils d'alerte (basés sur l'indice de qualité de l'air, voir Figure 1), et des recommandations d'actions (ex. restrictions de circulation).
- Des **rapports d'analyse** pour les décideurs, détaillant les tendances de pollution (saisonnieres, journalières) et les facteurs clés (ex. météo, sources d'émissions).

1.3 Objectifs et Modifications du Processus Décisionnel

Ce projet vise à répondre à la problématique de la pollution de l'air par une approche prédictive basée sur les données. Les objectifs principaux sont :

- **Prédire avec précision** les niveaux de PM_{2.5} à partir de données atmosphériques quotidiennes, en capturant les variations temporelles (saisonnieres, journalières, horaires) et spatiales (entre stations).

- **Comparer les performances** de différents modèles de *machine learning* (régression linéaire, forêts aléatoires, gradient boosting, etc.) pour identifier l'approche la plus efficace.
- **Fournir des insights exploitables** pour la prise de décision, en offrant aux gestionnaires urbains et autorités sanitaires des outils pour anticiper les épisodes de pollution, émettre des alertes ciblées, et élaborer des stratégies de réduction des émissions.

Actuellement, les décisions à Pékin reposent sur des seuils réactifs (ex. restrictions déclenchées lorsque PM2.5 dépasse $150 \mu\text{g}/\text{m}^3$ pendant plusieurs heures), ce qui limite la capacité à prévenir les pics de pollution. Ce projet propose de **modifier le processus décisionnel** en introduisant une approche proactive : les prévisions à court terme (24-48h) permettront d'anticiper les épisodes critiques, d'optimiser les mesures (ex. réduction du trafic avant un pic prévu), et de minimiser les impacts économiques et sanitaires. Par exemple, une alerte précoce pourrait réduire les hospitalisations liées à des crises respiratoires, estimées à 10-15 % supplémentaires lors des pics de pollution à Pékin.

1.4 Erreurs Métier à Considérer

Dans ce contexte, les **erreurs métier** sont critiques, car une mauvaise prédiction peut entraîner des conséquences graves. Les critères à considérer sont :

- **Sous-estimation des pics de pollution** : Une erreur de prédiction minimisant les concentrations réelles de PM2.5 (ex. prédire $50 \mu\text{g}/\text{m}^3$ alors que la valeur réelle est $150 \mu\text{g}/\text{m}^3$) pourrait retarder les mesures de protection, exposant la population à des risques sanitaires accrus. Cette erreur est jugée la plus critique, avec un seuil acceptable de sous-estimation de $25 \mu\text{g}/\text{m}^3$ (basé sur les recommandations de l'OMS).
- **Surestimation des niveaux** : Prédire des concentrations trop élevées (ex. $200 \mu\text{g}/\text{m}^3$ au lieu de $100 \mu\text{g}/\text{m}^3$) peut entraîner des mesures économiques inutiles (ex. fermetures industrielles), mais est moins critique pour la santé publique. Un seuil acceptable de surestimation est fixé à $50 \mu\text{g}/\text{m}^3$.
- **Métriques de performance** : Le modèle doit viser un RMSE inférieur à $30 \mu\text{g}/\text{m}^3$ et un MAE inférieur à $20 \mu\text{g}/\text{m}^3$ pour répondre aux besoins des autorités, garantissant une précision suffisante pour des prévisions exploitables.

1.5 Structure du Rapport

Le rapport s'articule autour de plusieurs sections clés. La section **méthodologie** détaille la collecte, le prétraitement des données, et les modèles de *machine learning* employés. Ensuite, les **résultats** et leur analyse comparative sont présentés, suivis d'une discussion sur les **implications pratiques**. Enfin, une **conclusion** synthétise les enseignements tirés et propose des recommandations pour une application réelle et des améliorations futures.

2 Description des Données

Cette section détaille les caractéristiques du dataset utilisé pour prédire les niveaux de pollution atmosphérique, en mettant en lumière sa provenance, sa structure et les variables exploitées. Une compréhension claire de ces données est essentielle pour justifier les choix méthodologiques et évaluer la robustesse des modèles de machine learning employés.

2.1 Provenance et Caractéristiques Générales

- **Source des Données** : Le dataset, connu sous le nom de *PRSA Dataset*, provient de deux entités principales : les données sur la qualité de l'air sont issues du *Beijing Municipal Environmental Monitoring Center*, tandis que les données météorologiques sont fournies par la *China Meteorological Administration*. Ces informations ont été collectées à partir de 12 stations de surveillance réparties dans la ville de Pékin, chaque station étant associée à la station météorologique la plus proche pour garantir une cohérence spatiale. [3]
- **Période Couverte** : Les données s'étendent sur quatre années, du **1er mars 2013 au 28 février 2017**, offrant une couverture temporelle suffisante pour analyser les tendances saisonnières, journalières et les variations à long terme de la pollution.
- **Fréquence d'Échantillonnage** : Les mesures sont enregistrées à une fréquence **horaire**, ce qui permet une granularité fine (24 observations par jour et par station), idéale pour capturer les fluctuations rapides des niveaux de polluants, comme les pics liés au trafic ou aux conditions météorologiques. Au total, après fusion des fichiers individuels (un par station), le dataset comprend plusieurs dizaines de milliers d'observations, bien que ce nombre exact soit ajusté après le traitement des valeurs manquantes.

2.2 Variables Clés

Le dataset regroupe **18 variables** mesurées à chaque intervalle horaire, divisées en trois catégories principales : les **polluants atmosphériques**, les **paramètres météorologiques** et les **métadonnées spatio-temporelles**. Ces variables ont été sélectionnées pour leur pertinence dans la prédiction des concentrations de **PM2.5**, cible principale de ce projet.

2.2.1 Polluants Atmosphériques Mesurés

Les concentrations de six polluants majeurs sont incluses, exprimées en microgrammes par mètre cube ($\mu\text{g}/\text{m}^3$) :

- **PM2.5** : Particules fines de diamètre inférieur à 2,5 micromètres, principal indicateur de pollution dans cette étude en raison de leur impact sur la santé respiratoire et cardiovasculaire.
- **PM10** : Particules plus grossières (diamètre inférieur à 10 micromètres), souvent corrélées aux PM2.5 mais influencées par des sources différentes (poussière, construction).

- **SO₂ (Dioxyde de soufre)** : Issu principalement de la combustion de combustibles fossiles, indicateur d'activité industrielle.
- **NO₂ (Dioxyde d'azote)** : Lié aux émissions des véhicules et aux processus de combustion, marqueur clé de la pollution urbaine.
- **CO (Monoxyde de carbone)** : Produit par la combustion incomplète (véhicules, chauffage), corrélé aux niveaux de trafic.
- **O₃ (Ozone)** : Polluant secondaire formé par des réactions photochimiques, influencé par la température et la lumière solaire.

Ces polluants reflètent un spectre large des sources de pollution à Pékin (trafic, industrie, conditions naturelles), rendant le dataset adapté à une analyse multidimensionnelle.

2.2.2 Paramètres Météorologiques

Les conditions météorologiques jouent un rôle déterminant dans la dispersion ou l'accumulation des polluants. Les variables suivantes sont incluses :

- **TEMP (Température)** : Mesurée en degrés Celsius (°C), elle influence les inversions thermiques (trappage des polluants près du sol par temps froid) et les réactions chimiques (formation d'ozone par temps chaud).
- **PRES (Pression atmosphérique)** : En hectopascals (hPa), elle indique les conditions de stabilité atmosphérique affectant la stagnation de l'air.
- **DEWP (Point de rosée)** : En °C, proxy de l'humidité relative, influençant la diffusion des particules et la formation de brouillard polluant.
- **RAIN (Précipitations)** : En millimètres (mm), les pluies peuvent réduire les concentrations de particules en les "lavant" de l'atmosphère.
- **WSPM (Vitesse du vent)** : En mètres par seconde (m/s), un vent fort disperse les polluants, tandis qu'un vent faible favorise leur accumulation.
- **wd (Direction du vent)** : Variable catégorique (ex. N, NE, S), elle indique la provenance des masses d'air et leur potentiel à transporter ou confiner les polluants.

2.2.3 Métadonnées Spatio-Temporelles

- **No** : Numéro de ligne, un identifiant unique pour chaque observation (non utilisé dans la prédiction).
- **Année, Mois, Jour, Heure (year, month, day, hour)** : Informations temporales permettant de reconstruire une colonne `datetime` pour les analyses chronologiques.
- **Station** : Nom de la station de surveillance (ex. Huairou, Tiantan), offrant une dimension spatiale pour explorer les variations locales de la pollution.

3 Méthodologie

Cette section détaille les étapes suivies pour traiter les données et développer les modèles de prédiction des niveaux de PM2.5 à partir du PRSA Dataset. Elle couvre le prétraitement des données, la sélection et la justification des modèles de machine learning, ainsi que les méthodes d'évaluation utilisées pour comparer leurs performances.

3.1 Prétraitement des Données

Le dataset initial, constitué de 12 fichiers CSV représentant les données des stations de surveillance de la qualité de l'air à Pékin, a été fusionné en un DataFrame unique comprenant initialement 18 colonnes et 420 768 observations, couvrant la période de mars 2013 à février 2017. Ces colonnes englobent des types variés : numériques continues (ex. PM2.5, Temp), catégoriques (ex. WinDir) et temporelles (ex. Year, Month, Day, Hour). Le prétraitement, essentiel pour garantir la qualité des données avant l'entraînement des modèles, a été structuré en plusieurs étapes détaillées ci-dessous.

3.1.1 Consolidation Initiale des Données

Les fichiers CSV, situés dans un répertoire local, ont été lus et combinés à l'aide de la bibliothèque Python `glob` et de la fonction `pd.concat` de Pandas. Cette étape a produit un dataset unifié de 420 768 lignes, conservant la colonne `Station` pour identifier l'origine spatiale des observations. Une colonne inutile, `No` (numéro de ligne), a été supprimée car elle n'apporte aucune information pertinente pour la prédiction, réduisant ainsi l'encombrement mémoire et simplifiant le traitement. Par ailleurs, les noms des colonnes ont été renommés pour plus de lisibilité (ex. `year` → `Year`, `pm2.5` → `PM2.5`, `wd` → `WinDir`), améliorant la clarté lors des manipulations ultérieures.

3.1.2 Nettoyage des Données

Une analyse préliminaire a révélé la présence de valeurs manquantes significatives : 8 739 pour `PM2.5` (environ 2,1 % des données), 20 701 pour `CO` (4,9 %), et 1 822 pour `WinDir` (0,4 %), entre autres. Pour les variables numériques continues (`PM2.5`, `PM10`, `S02`, `N02`, `CO`, `O3`, `Temp`, `Press`, `DewP`, `Rain`, `WinSpeed`), les lacunes ont été comblées par une imputation basée sur la moyenne mensuelle. Pour chaque mois (1 à 12), les données ont été filtrées, et la moyenne de chaque colonne a été calculée et arrondie à deux décimales (ex. $PM2.5 = 94,66 \mu\text{g}/\text{m}^3$ en mars, $63,11 \mu\text{g}/\text{m}^3$ en mai). Cette méthode préserve les variations saisonnières, cruciales pour une série temporelle comme celle-ci, contrairement à une interpolation linéaire qui aurait pu lisser excessivement les données. Si une moyenne mensuelle était indisponible, la moyenne globale était utilisée comme repli.

Pour la variable catégorique `WinDir` (direction du vent), les valeurs manquantes ont été imputées par le mode spécifique à chaque station (ex. "N" pour Gucheng, "NE" pour Aotizhongxin), reflétant les patterns locaux de vent et améliorant la cohérence spatiale. Les doublons, bien que rares, ont été supprimés pour éviter toute redondance.

3.1.3 Traitement des Variables Catégoriques

Les colonnes catégoriques `WinDir` et `Station` ont été encodées numériquement avec `LabelEncoder` de Scikit-learn pour les rendre exploitables par les modèles de machine learning. Pour `WinDir`, les 16 directions (ex. "NW", "ENE") ont été mappées à des entiers de 0 à 15 (ex. "NW" → 7, "N" → 3). Pour `Station`, les 12 stations ont été codées de 0 à 11 (ex. "Aotizhongxin" → 0, "Tiantan" → 9). Cet encodage ordinal, bien que simple, suppose une égalité implicite entre les catégories, mais il a été retenu pour sa facilité d'implémentation et son efficacité computationnelle, une alternative comme le *one-hot encoding* ayant été écartée pour éviter une explosion dimensionnelle (28 colonnes supplémentaires).

3.1.4 Normalisation et Standardisation

Les variables numériques présentent des échelles hétérogènes (ex. `PM2.5` en $\mu\text{g}/\text{m}^3$ vs. `Temp` en $^\circ\text{C}$), ce qui peut affecter les modèles sensibles aux magnitudes, comme `KNeighborsRegressor`. Une standardisation a été appliquée, transformant chaque variable X en une distribution centrée réduite selon la formule :

$$X' = \frac{X - \mu}{\sigma}$$

où μ est la moyenne et σ l'écart-type calculés sur l'ensemble des données. Cette normalisation garantit que toutes les features contribuent équitablement à la prédiction, optimisant la convergence des algorithmes.

3.1.5 Feature Engineering

Pour enrichir le dataset et capturer les dynamiques temporelles, une colonne `Date` a été créée en combinant `Year`, `Month` et `Day` via `pd.to_datetime`, offrant une base pour des analyses chronologiques. Une colonne `DayNames` (noms des jours) a été temporairement ajoutée pour explorer les effets des activités humaines (ex. trafic plus intense en semaine), mais elle a été supprimée par la suite, car redondante avec les informations déjà présentes dans `Date` et peu pertinente pour la prédiction finale de `PM2.5`. De plus, bien que non implémenté ici, des variables cycliques pour `Hour` (sinus et cosinus) auraient pu être envisagées pour modéliser les cycles diurnes (ex. pics de pollution matinaux), une amélioration potentielle pour de futures itérations.

3.1.6 Réduction et Justification des Colonnes

Après le prétraitement initial, plusieurs colonnes ont été supprimées pour optimiser le dataset : `PM10`, `Year`, `Month`, `Day`, `Hour`, `Date`, et `DayNames`. La suppression de `PM10` est justifiée par sa forte corrélation avec `PM2.5` (souvent $> 0,8$ dans les données urbaines), ce qui introduirait une multicolinéarité inutile sans apporter d'information significative supplémentaire pour prédire `PM2.5`, la cible principale. Les colonnes temporelles (`Year`, `Month`, `Day`, `Hour`, `Date`) ont été éliminées après avoir servi à créer des features ou à guider l'imputation, leur rôle étant désormais intégré dans les données restantes ou remplacé par des variables dérivées potentielles (ex. saison). Enfin, `DayNames` a été écartée en raison de sa faible valeur prédictive directe par rapport aux autres variables météorologiques et polluantes. Le dataset final contient 12 colonnes et 420 768 lignes, occupant 41,7 Mo en mémoire, une réduction notable par rapport aux 64,2 Mo initiaux.

3.1.7 Séparation en Données d'Entraînement et de Test

Les données ont été divisées en un ensemble d'entrée X (toutes les colonnes sauf PM2.5) et une cible y (PM2.5). Un *train-test split* a été effectué avec `train_test_split` de Scikit-learn, allouant 80 % des données à l'entraînement (336 614 observations) et 20 % au test (84 154 observations), avec un `random_state=128` pour la reproductibilité. Bien qu'un split temporel aurait été idéal pour une série chronologique, un split aléatoire a été retenu ici pour simplifier l'analyse initiale, une limitation à considérer dans les travaux futurs.

3.2 Analyse et Visualisation des Données

L'analyse exploratoire des données est une étape cruciale pour comprendre les relations entre les variables du dataset *PRSA* et identifier les facteurs influençant les niveaux de PM2.5. Cette section présente une série de visualisations générées avec la bibliothèque `seaborn` de Python, accompagnées d'une analyse détaillée des patterns observés. Ces insights ont guidé la sélection des features et l'évaluation des modèles de machine learning.

3.2.1 Statistique Descriptive

Une analyse statistique descriptive a été réalisée sur les colonnes clés du dataset *PRSA* pour comprendre la distribution des variables influençant les concentrations de PM2.5. Le tableau 1 présente les statistiques principales pour les polluants (PM2.5, PM10, SO2, NO2, CO, O3) et les variables météorologiques (Temp, DewP, Press), basées sur 420 768 observations.

Statistique	PM2.5	PM10	SO2	NO2	CO	O3	Temp	DewP	Press
count	420768	420768	420768	420768	420768	420768	420768	420768	420768
mean	79.77	104.61	15.82	50.61	1232.98	57.22	13.53	2.48	1010.75
std	80.01	91.09	21.46	34.66	1136.59	56.00	11.44	13.80	10.47
min	2.00	2.00	0.29	1.03	100.00	0.21	-19.90	-43.40	982.40
25%	21.00	36.00	3.00	24.00	500.00	12.00	3.10	-8.90	1002.30
50%	57.00	83.00	7.14	44.00	900.00	45.00	14.50	3.00	1010.40
75%	109.00	144.00	20.00	70.00	1500.00	82.00	23.20	15.10	1019.00
max	999.00	999.00	500.00	290.00	10000.00	1071.00	41.60	29.10	1042.80

TABLE 1 – Statistiques descriptives des variables clés (valeurs en $\mu\text{g}/\text{m}^3$ pour les polluants, $^{\circ}\text{C}$ pour Temp et DewP, hPa pour Press).

Les résultats montrent une forte variabilité dans les concentrations de polluants, avec des moyennes élevées pour PM2.5 (79,77 $\mu\text{g}/\text{m}^3$) et PM10 (104,61 $\mu\text{g}/\text{m}^3$), dépassant largement les seuils de sécurité de l'OMS (25 $\mu\text{g}/\text{m}^3$ pour PM2.5, 50 $\mu\text{g}/\text{m}^3$ pour PM10). Les écarts-types élevés (ex. 80,01 pour PM2.5, 91,09 pour PM10) indiquent des fluctuations importantes, confirmées par les maximums extrêmes (999 $\mu\text{g}/\text{m}^3$ pour PM2.5, 10000 $\mu\text{g}/\text{m}^3$ pour CO). Les variables météorologiques présentent également une large gamme, avec des températures variant de -19,9 $^{\circ}\text{C}$ à 41,6 $^{\circ}\text{C}$ et des points de rosée de -43,4 $^{\circ}\text{C}$ à 29,1 $^{\circ}\text{C}$, reflétant les conditions climatiques contrastées de Pékin. Ces statistiques soulignent la nécessité d'un modèle robuste pour capturer ces variations, orientant ainsi les choix des algorithmes de *machine learning*.

3.2.2 Relations Météorologiques

Pour explorer l'impact des conditions météorologiques sur la pollution, des graphiques de régression (`sns.regplot`) ont été créés pour examiner les relations entre PM2.5 et les variables Temp (température), DewP (point de rosée), ainsi que les polluants SO2 et NO2 avec PM10.

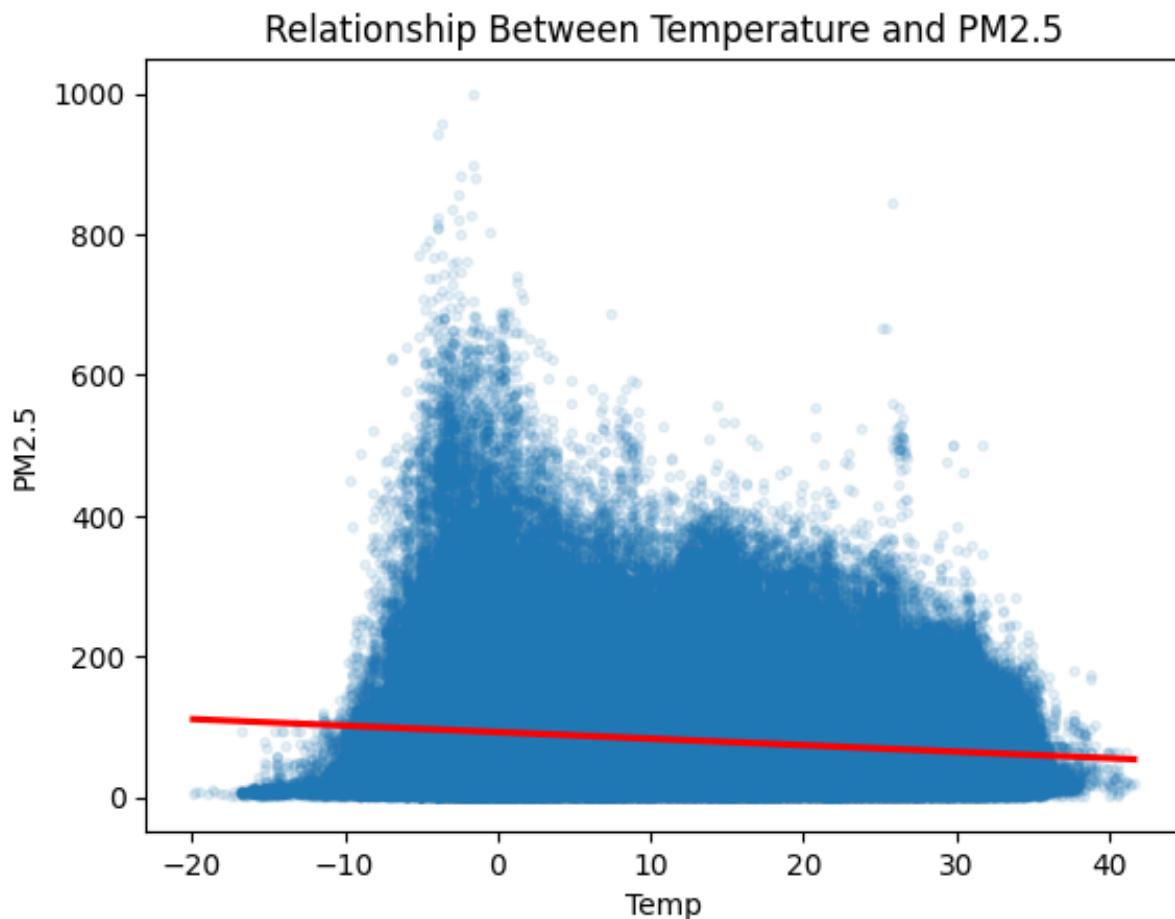


FIGURE 2 – Relation entre la température et PM2.5.

Le graphique 2 montre une relation négative faible entre Temp et PM2.5, avec une droite de régression rouge indiquant une diminution des concentrations de PM2.5 lorsque la température augmente (de -20 °C à 40 °C). Cette tendance s'explique par les inversions thermiques fréquentes en hiver à Pékin, qui piègent les polluants près du sol par temps froid. Cependant, la dispersion importante des points suggère que d'autres facteurs (ex. vent, précipitations) modulent fortement cette relation.

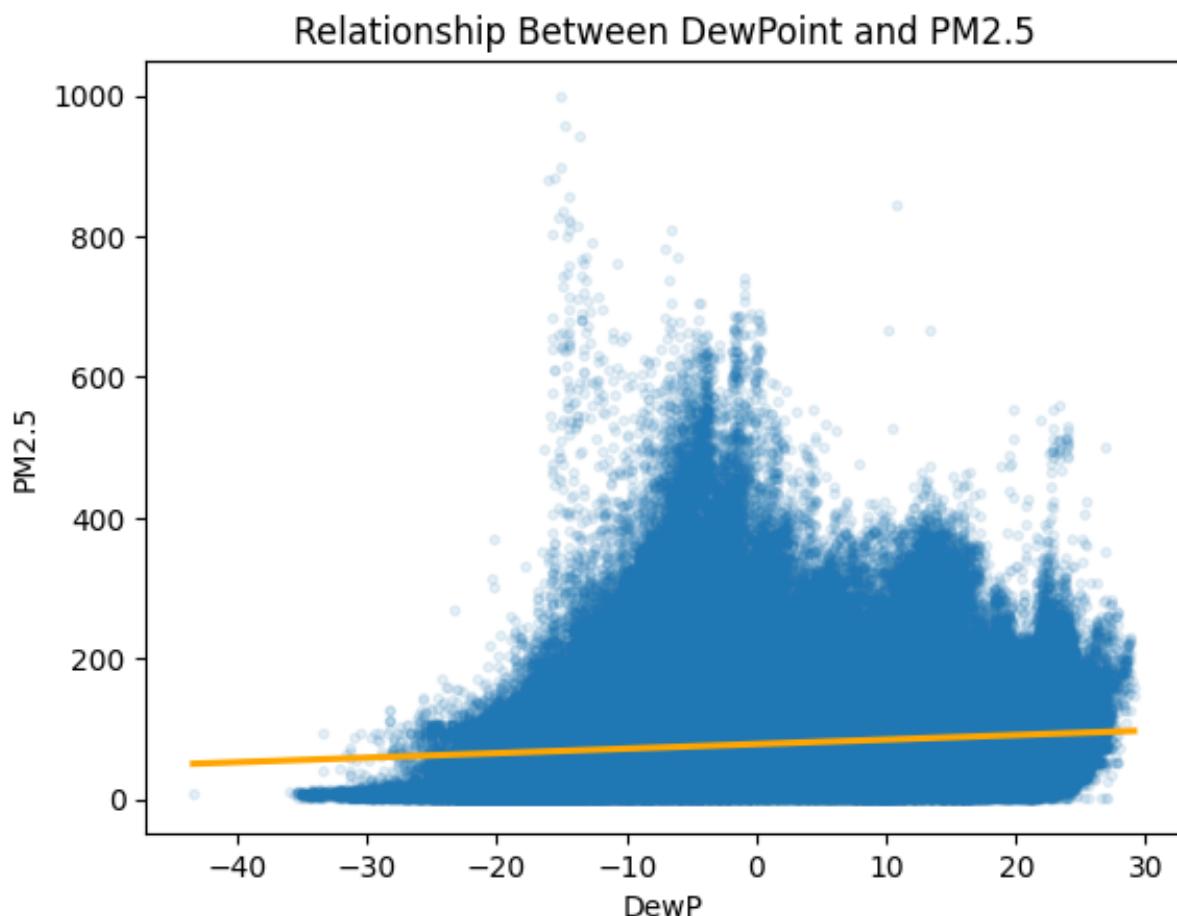


FIGURE 3 – Relation entre le point de rosée et PM2.5.

Le graphique 3 révèle une corrélation positive modérée entre DewP et PM2.5, avec une droite de régression orange. Les concentrations de PM2.5 augmentent lorsque le point de rosée s'élève (de -40 °C à 30 °C), reflétant une humidité accrue qui peut favoriser l'accumulation de particules fines en suspension. La densité des points autour de 0 °C à 400 µg/m³ indique une plage critique où l'humidité influence significativement la pollution.

3.2.3 Relations entre Polluants

Les interactions entre les polluants (PM2.5, PM10, SO₂, NO₂, CO, O₃) ont été explorées pour évaluer leurs interdépendances.

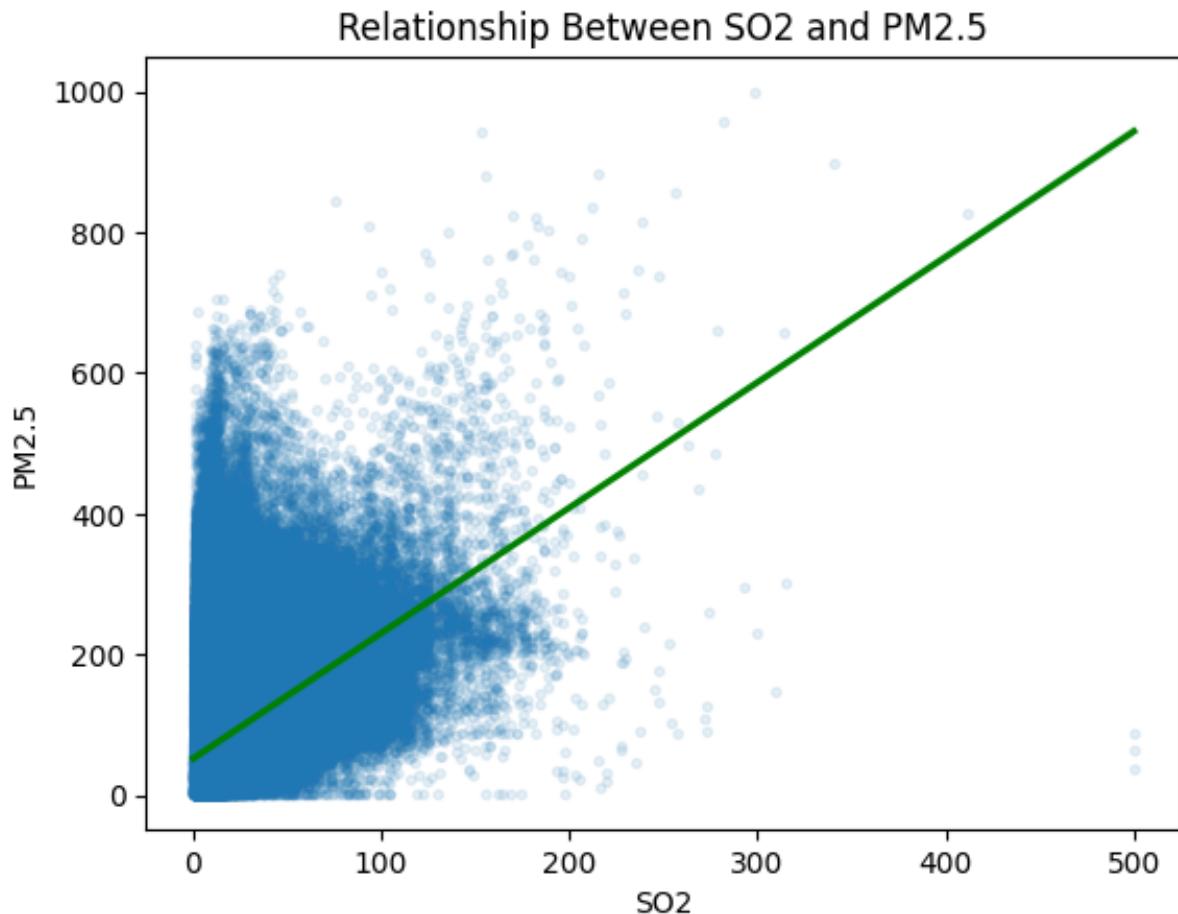


FIGURE 4 – Relation entre SO₂ et PM2.5.

Le graphique 4 met en évidence une corrélation positive forte entre SO₂ et PM2.5, avec une droite de régression verte. Lorsque SO₂ augmente (jusqu'à 500 µg/m³), PM2.5 suit une tendance similaire, suggérant une source commune, probablement les émissions industrielles ou de chauffage.

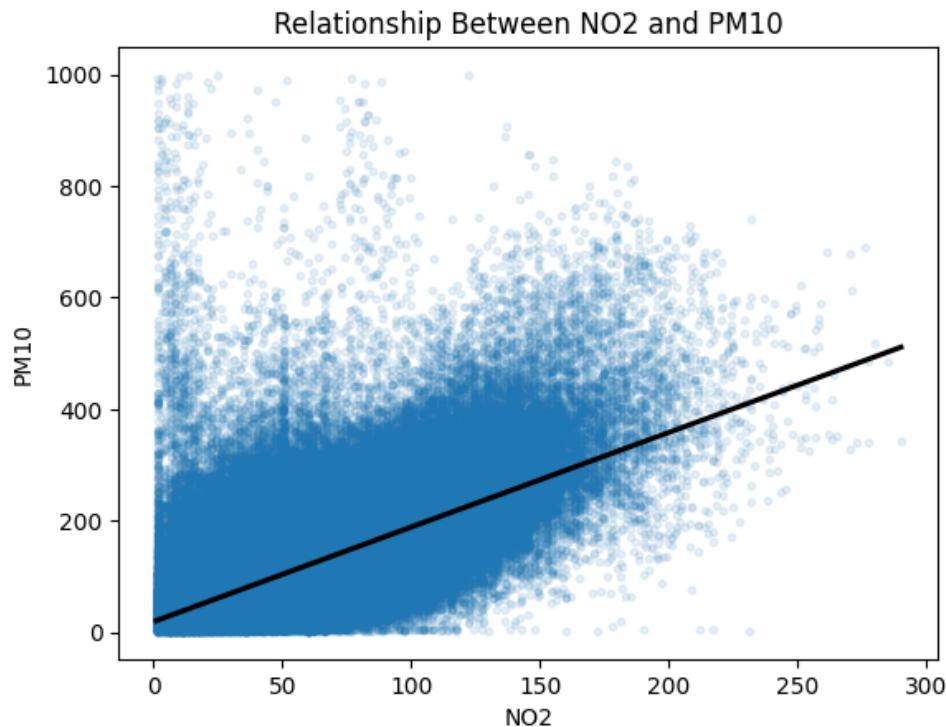


FIGURE 5 – Relation entre NO2 et PM10.

Le graphique 5 montre une corrélation positive modérée entre NO2 et PM10, avec une droite de régression noire. Les concentrations augmentent ensemble (jusqu'à $300 \mu\text{g}/\text{m}^3$ pour NO2 et $1000 \mu\text{g}/\text{m}^3$ pour PM10), indiquant une influence du trafic ou des combustions, sources typiques de ces polluants.

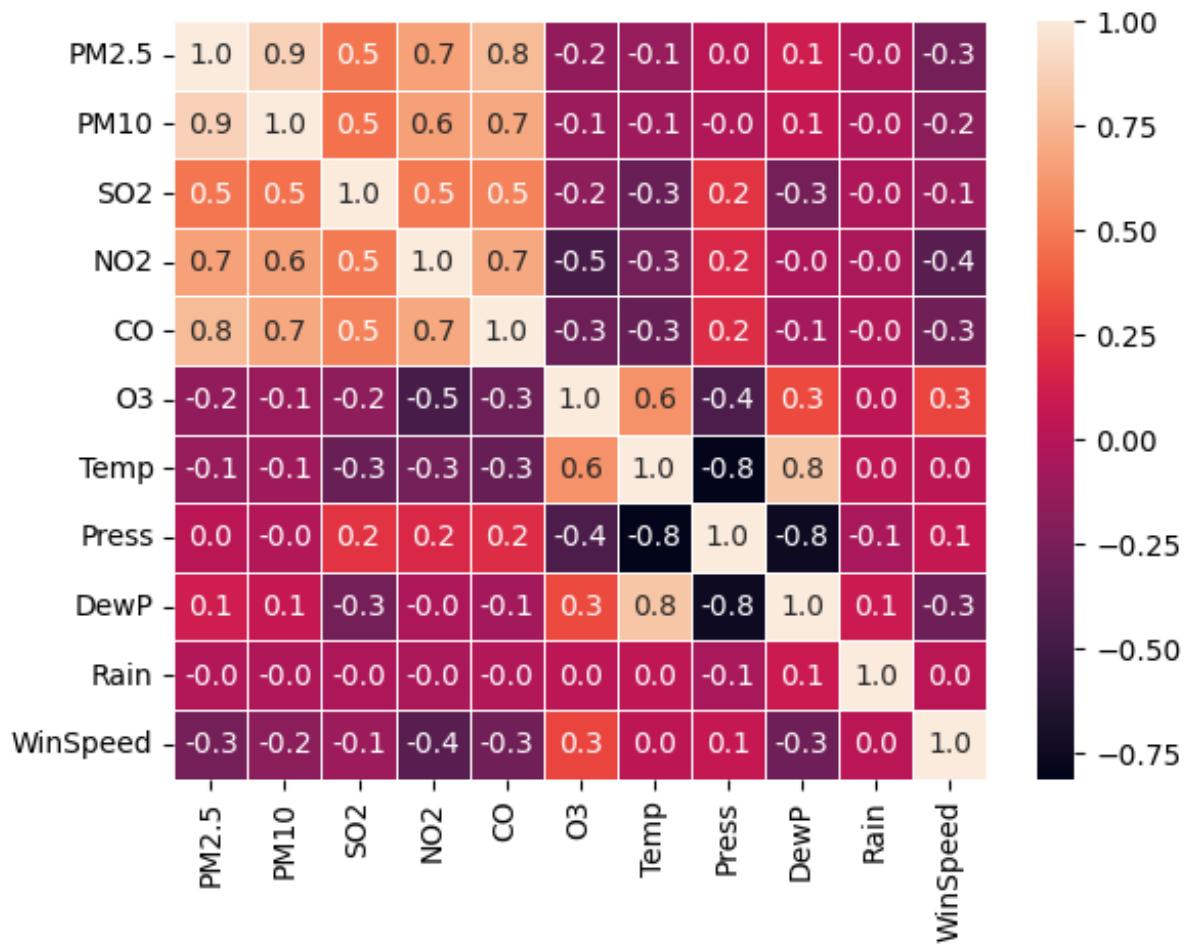


FIGURE 6 – Matrice de corrélation des variables numériques.

Le graphique 6, généré avec `sns.heatmap`, illustre les relations entre les variables numériques après suppression des colonnes non numériques (Year, Month, Day, Hour, Date). Les corrélations les plus fortes sont observées entre PM2.5 et PM10 (0,9), confirmant une forte dépendance entre ces deux types de particules, ce qui justifie la suppression de PM10 dans les étapes précédentes pour éviter la multicolinéarité. PM2.5 est également corrélé positivement avec NO2 (0,7), CO (0,8), et SO2 (0,5), indiquant des sources communes (trafic, combustion industrielle). Une corrélation négative modérée entre PM2.5 et O3 (-0,3) reflète des processus chimiques opposés : O3 est plus présent en été (photochimie), tandis que PM2.5 culmine en hiver. Les variables météorologiques comme Temp et Press montrent des corrélations faibles à modérées (-0,1 à -0,4), mais leur inclusion reste pertinente pour capturer les effets indirects.

3.2.4 Analyse Spatiale et Temporelle

La variabilité spatiale et temporelle a été étudiée à travers les stations de surveillance et sur différentes échelles de temps (années, mois, jours, heures).

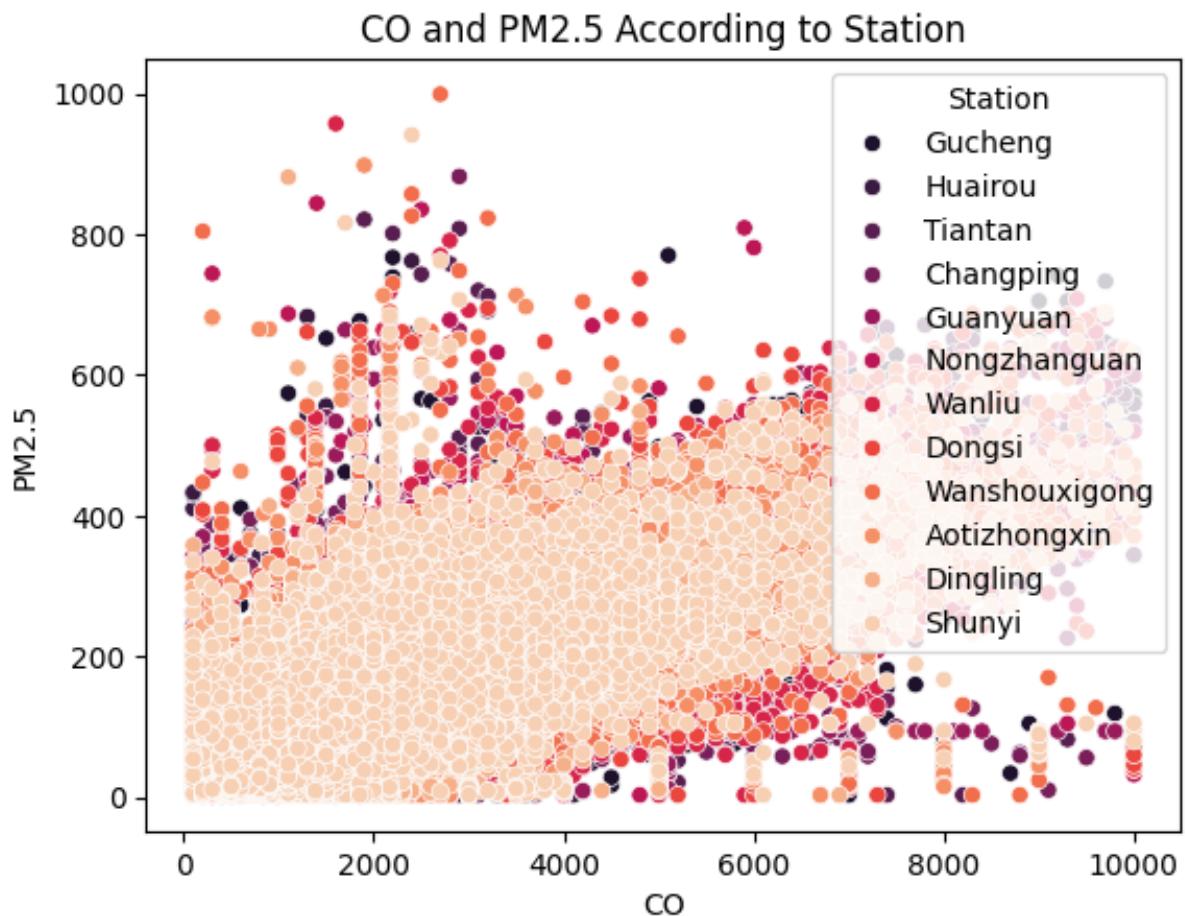


FIGURE 7 – Relation entre CO et PM2.5 par station.

Le graphique 7 utilise une palette de couleurs pour différencier les 12 stations. Une corrélation positive globale entre CO et PM2.5 est observable, avec des densités de points variant selon les stations (ex. Gucheng et Tiantan montrent des pics à $800\text{-}1000 \mu\text{g}/\text{m}^3$ pour PM2.5), suggérant des influences locales (trafic, industrie).

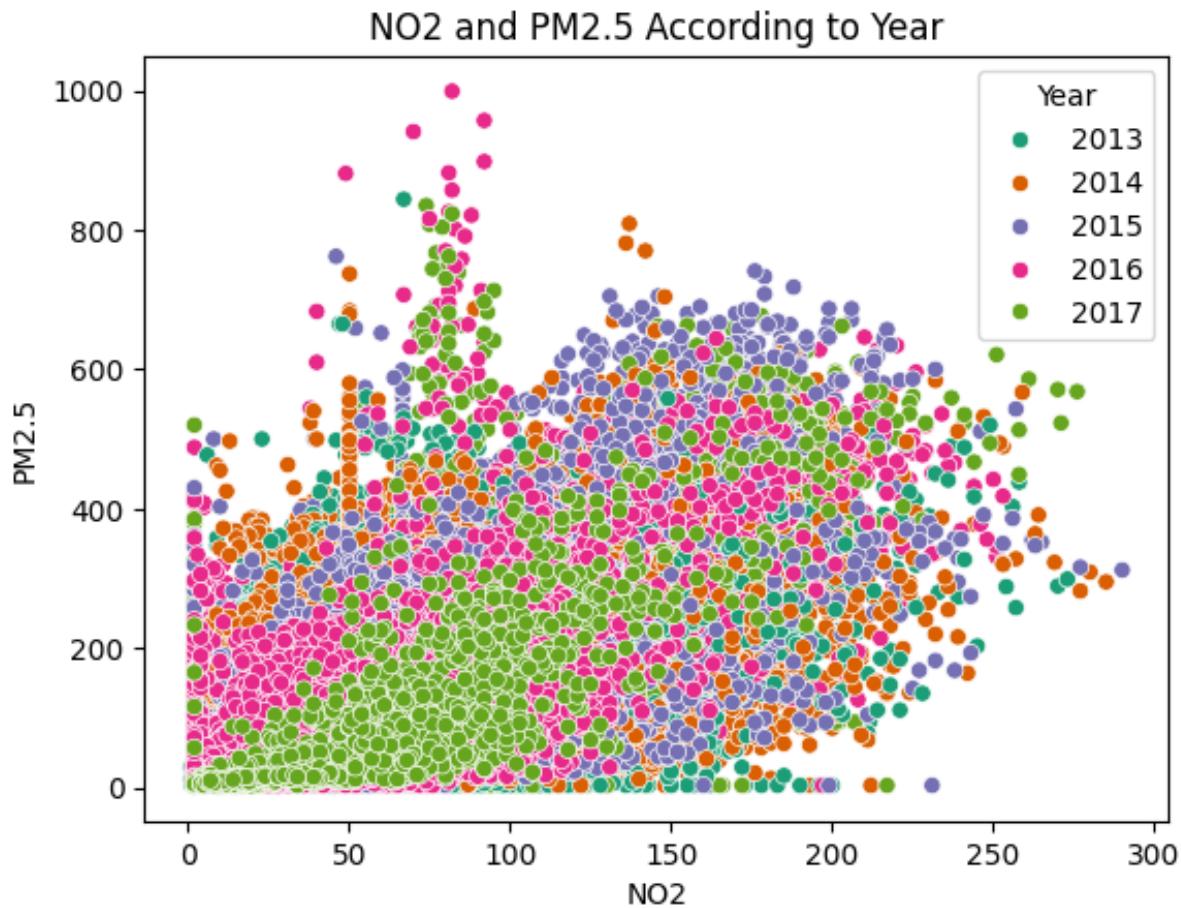


FIGURE 8 – Relation entre NO₂ et PM_{2.5} par année.

Le graphique 8 illustre une relation stable entre NO₂ et PM_{2.5} sur les années 2013-2017, avec une densité de points croissante autour de 200-400 µg/m³ pour PM_{2.5}. Une légère augmentation des concentrations en 2013-2014 (vert) est suivie d'une stabilisation (2015-2017), potentiellement due à des mesures de contrôle de la pollution.

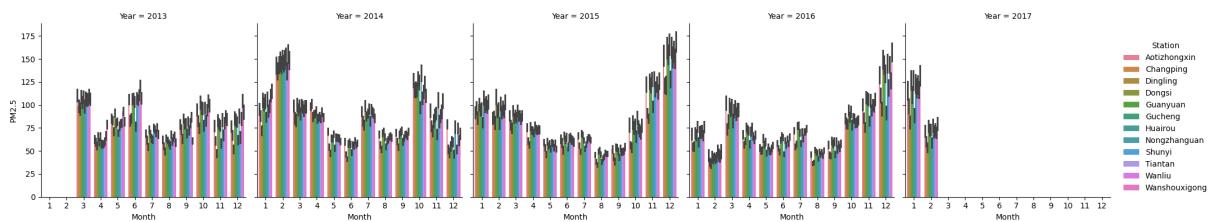


FIGURE 9 – Variations mensuelles de PM_{2.5} par station et par année (2013-2017).

Le graphique 9, généré avec `sns.catplot`, montre les concentrations moyennes de PM_{2.5} par mois, ventilées par station et par année. Une forte saisonnalité est observée : les niveaux culminent en hiver (janvier, février, décembre) à plus de 100 µg/m³, particulièrement en 2013 et 2016, et chutent en été (juin, juillet, août) à environ 50 µg/m³. Les stations urbaines comme Aotizhongxin et Tiantan montrent des pics plus élevés, tandis que Dingling, plus rurale, présente des concentrations plus faibles.

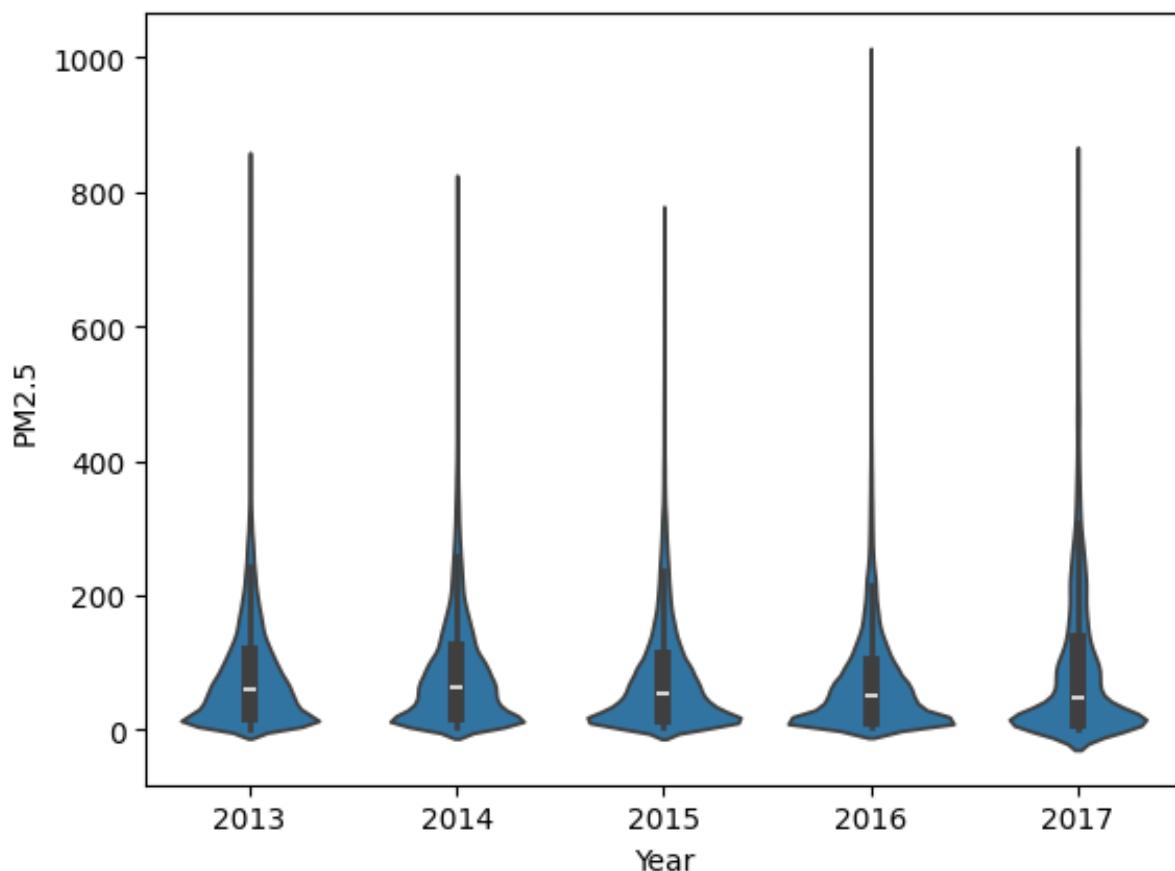


FIGURE 10 – Distribution de PM2.5 par année (2013-2017).

Le graphique 10, un *violin plot*, montre la distribution des concentrations de PM2.5 par année. Les distributions sont asymétriques, avec des queues longues jusqu'à $1000 \mu\text{g}/\text{m}^3$, indiquant des pics sporadiques. La médiane reste stable ($50-70 \mu\text{g}/\text{m}^3$), mais la dispersion diminue légèrement en 2017, suggérant une amélioration de la qualité de l'air.

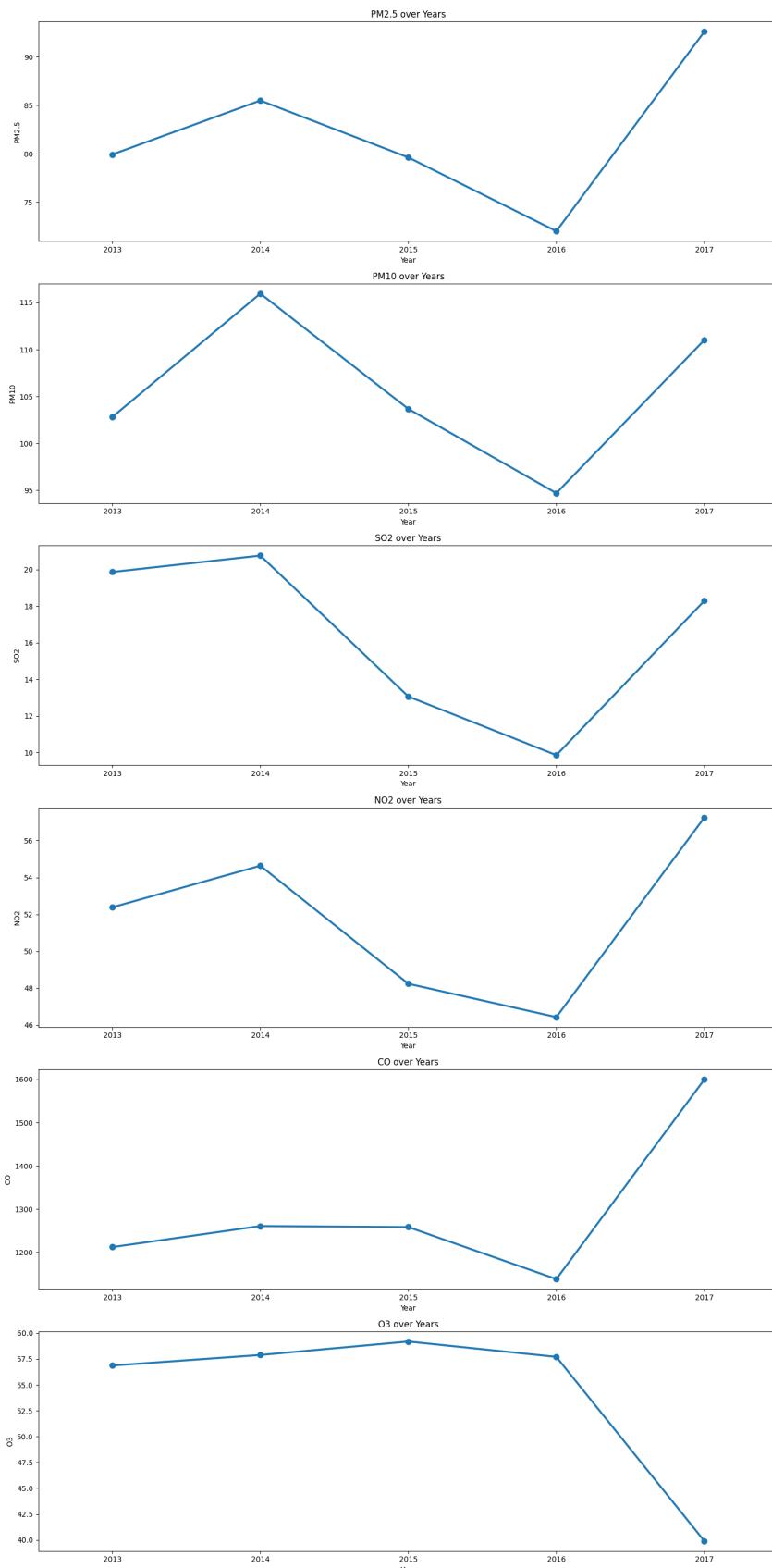


FIGURE 11 – Évolution annuelle des polluants (2013-2017).

Le graphique 11, une série de *point plots*, montre l'évolution annuelle des polluants. PM_{2.5} et PM₁₀ suivent une tendance décroissante de 2013 à 2015, suivie d'une légère hausse en 2016, puis une baisse en 2017. SO₂ et NO₂ diminuent (de 20 à 15 µg/m³ et de 55 à 45 µg/m³), tandis que CO baisse de 1500 à 1100 µg/m³. O₃ reste stable (50 µg/m³), avec une légère hausse en 2017.

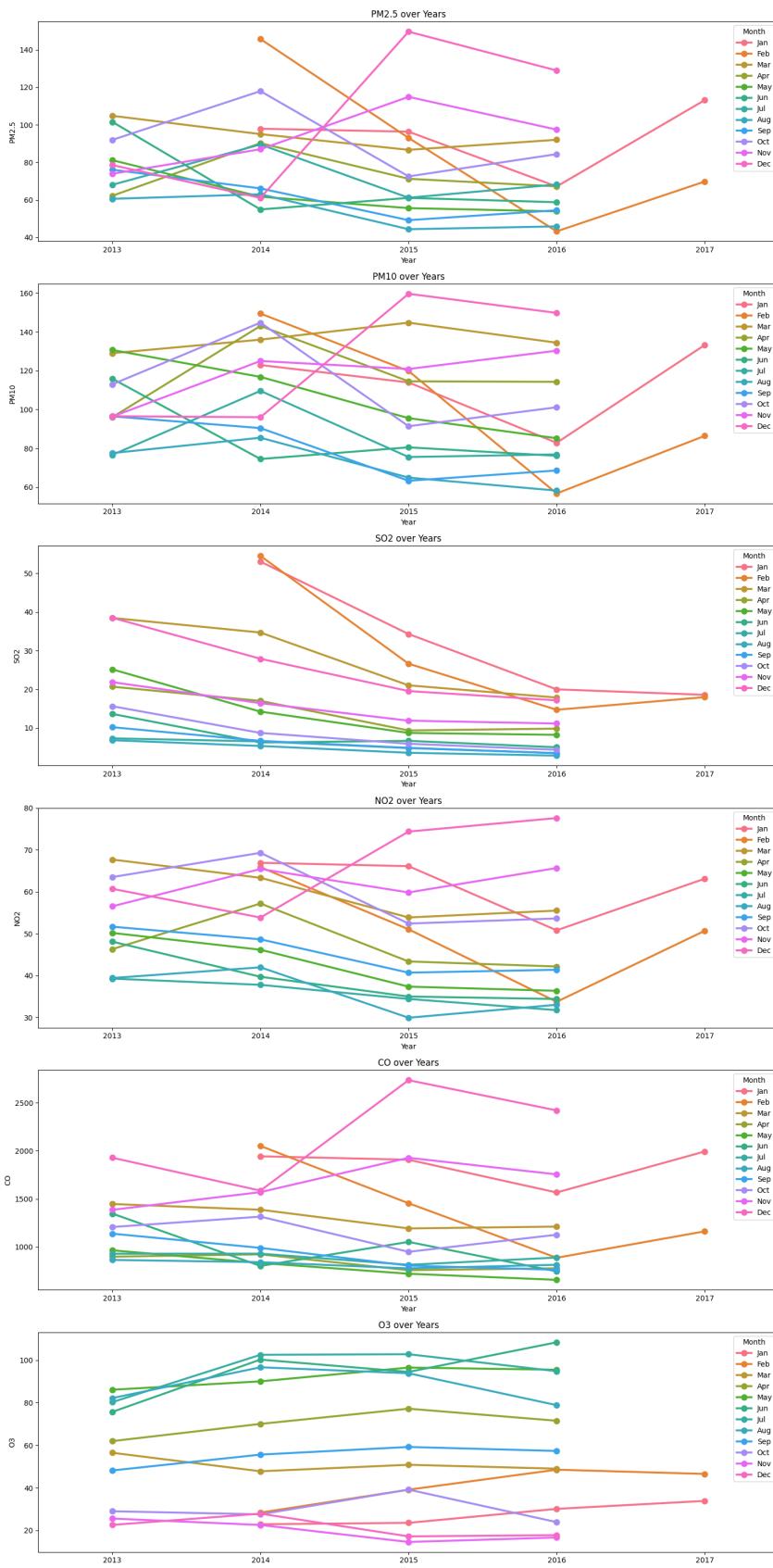


FIGURE 12 – Évolution annuelle des polluants par mois (2013-2017).

Le graphique 12 ventile les concentrations annuelles par mois, confirmant les pics hivernaux de PM_{2.5}, PM₁₀, NO₂, et CO, et les maxima estivaux de O₃ (juin, juillet). SO₂ montre une amplitude décroissante au fil des années.

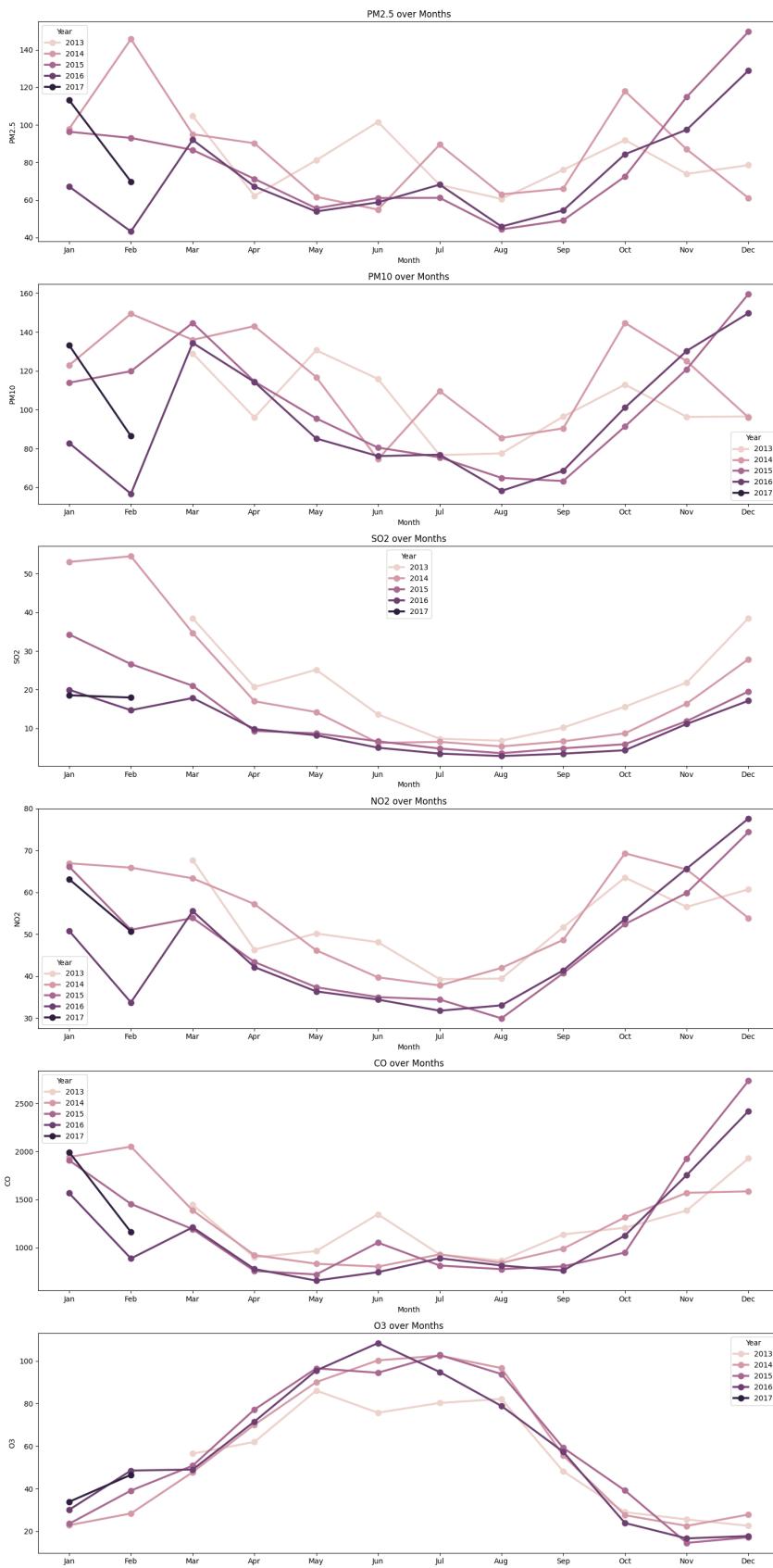


FIGURE 13 – Évolution mensuelle des polluants par année (2013-2017).

Le graphique 13 inverse la perspective, montrant les concentrations mensuelles ventilées par année. PM2.5 et PM10 atteignent 120-140 $\mu\text{g}/\text{m}^3$ en hiver, contre 50-60 $\mu\text{g}/\text{m}^3$ en été. O3 culmine à 80 $\mu\text{g}/\text{m}^3$ en été et tombe à 20 $\mu\text{g}/\text{m}^3$ en hiver, tandis que SO2 montre une baisse notable en 2017 (10-15 $\mu\text{g}/\text{m}^3$ en hiver contre 30-40 $\mu\text{g}/\text{m}^3$ en 2013).

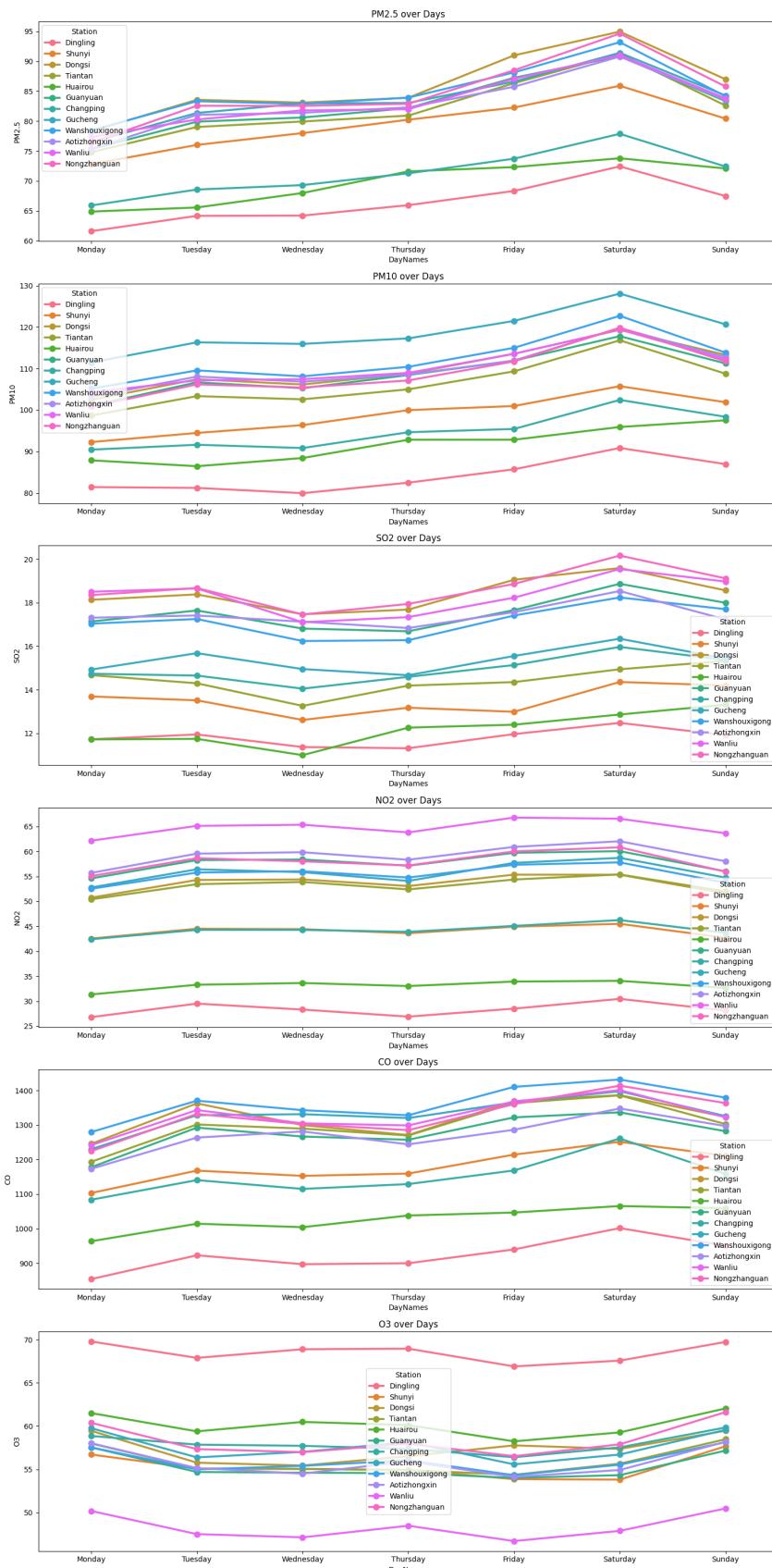


FIGURE 14 – Évolution des polluants par jour de la semaine selon les stations.

Le graphique 14 montre les concentrations par jour de la semaine, ventilées par station. PM_{2.5} et PM₁₀ sont plus élevés en milieu de semaine (mercredi, jeudi, 85-90 µg/m³ et 110-120 µg/m³), diminuant le week-end (70-75 µg/m³). NO₂ et CO suivent cette tendance (60 µg/m³ et 1400 µg/m³ en semaine), tandis que O₃ augmente légèrement le week-end (55-60 µg/m³). SO₂ reste stable (12-15 µg/m³).

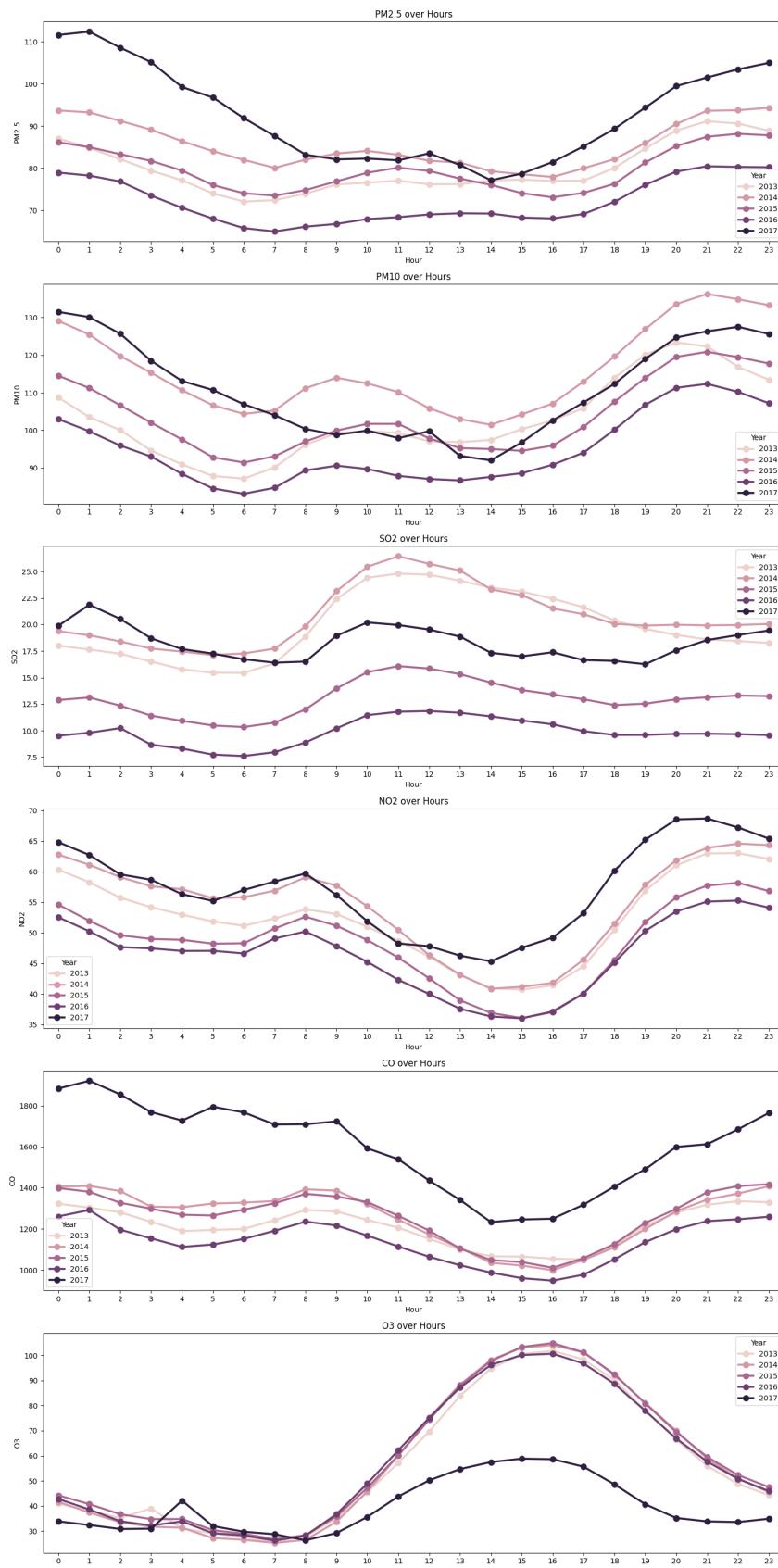


FIGURE 15 – Évolution des polluants par heure de la journée selon les années.

Le graphique 15 illustre les variations horaires. PM2.5, PM10, NO₂, et CO montrent des pics matinaux (7h-9h) et vespéraux (18h-20h) aux heures de pointe, atteignant respectivement 90-110 µg/m³, 60 µg/m³, et 1400 µg/m³, avec une baisse nocturne (4h). O₃ culmine l'après-midi (15h, 70-80 µg/m³), reflétant la formation photochimique. Les tendances s'atténuent en 2017.

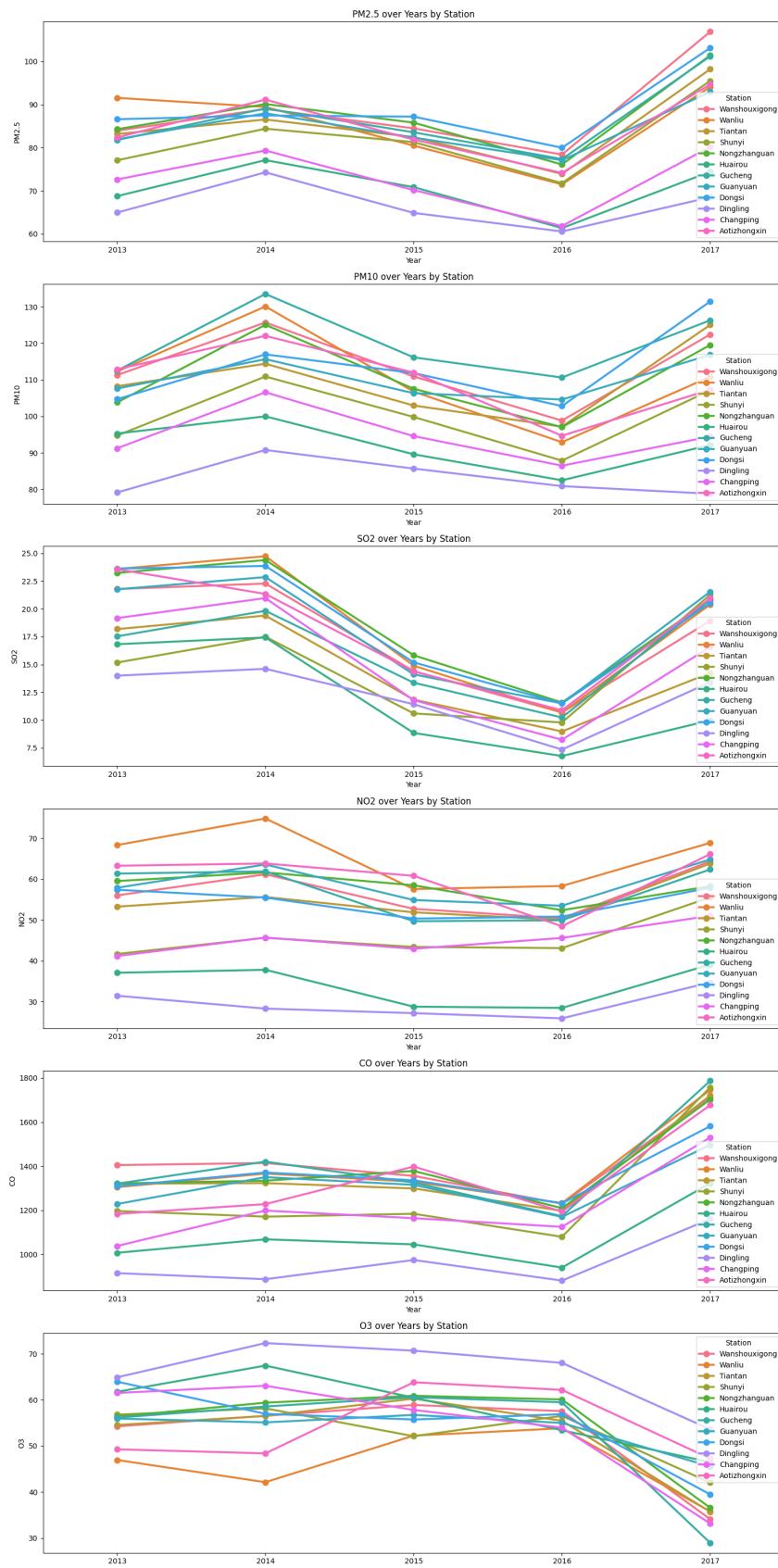


FIGURE 16 – Évolution des polluants par année selon les stations.

Le graphique 16 examine les concentrations par année et par station. PM2.5 et PM10 diminuent de 2013 ($90\text{-}110 \mu\text{g}/\text{m}^3$) à 2017 ($70\text{-}90 \mu\text{g}/\text{m}^3$), avec des stations urbaines (Aotizhongxin, Tiantan) plus polluées que Dingling. SO₂ et NO₂ baissent ($20 \text{ à } 12 \mu\text{g}/\text{m}^3$, $60 \text{ à } 45 \mu\text{g}/\text{m}^3$), CO passe de 1400 à $1100 \mu\text{g}/\text{m}^3$, et O₃ reste stable ($50\text{-}60 \mu\text{g}/\text{m}^3$), avec une légère hausse en 2017.

3.2.5 Interprétation et Implications

Ces visualisations révèlent des relations complexes : une dépendance saisonnière et météorologique (température, humidité), des corrélations fortes entre PM2.5, NO₂, et CO, et des variations spatiales et temporelles significatives. Les pics hivernaux, les hausses en semaine et aux heures de pointe, ainsi que les différences entre stations (urbaines vs. rurales) soulignent l'importance des features temporelles (saison, jour, heure) et spatiales (Station). La diminution des polluants sur 2013-2017 reflète les effets des politiques environnementales, mais les non-linéarités persistantes nécessitent des modèles comme RandomForest ou GradientBoosting pour capturer ces dynamiques et les interdépendances entre polluants.

3.3 Modèles de Machine Learning Utilisés

Cinq modèles de régression ont été sélectionnés pour prédire les concentrations continues de PM2.5 à partir du dataset *PRSA*, en tenant compte des relations complexes identifiées lors de l'analyse exploratoire (ex. saisonnalité, corrélations entre polluants). Chaque modèle a été choisi pour ses propriétés spécifiques, adaptées aux caractéristiques non linéaires et aux interactions spatiales/temporelles des données.

3.3.1 Régression Linéaire

La régression linéaire (`LinearRegression`) estime les paramètres par la méthode des moindres carrés ordinaires (OLS), modélisant la relation entre les features et PM2.5 par :

$$\hat{y} = \beta_0 + \sum_{i=1}^n \beta_i x_i + \epsilon$$

où \hat{y} est la prédiction, β_0 l'ordonnée à l'origine, β_i les coefficients des features x_i , et $\epsilon \sim N(0, \sigma^2)$ le terme d'erreur. La fonction objectif, $\sum(y - X\beta)^2$, est minimisée analytiquement par $\hat{\beta} = (X^T X)^{-1} X^T y$ [5]. Ce modèle sert de référence en raison de sa simplicité et de sa capacité à modéliser des relations linéaires. Cependant, il souffre d'un biais de spécification important face aux non-linéarités et interactions complexes observées dans les données de pollution (ex. saisonnalité, pics horaires).

3.3.2 Forêt Aléatoire

Le `RandomForestRegressor` repose sur le principe du bagging avec une double aléa-tion, combinant T arbres de décision entraînés sur des sous-ensembles aléatoires :

$$\hat{y} = E_\theta[h(x; \theta)]$$

où θ contrôle le sous-échantillonnage des données (bootstrap) et des features [1]. Chaque arbre est non élagué pour maximiser la variance initiale, et la réduction de corrélation

entre arbres, grâce à la sélection aléatoire des features, améliore la robustesse et capture efficacement les interactions non linéaires (Liaw et Wiener, 2002). Cette approche est particulièrement adaptée aux patterns saisonniers et horaires identifiés dans les données de PM2.5.

3.3.3 Gradient Boosting

Le `GradientBoostingRegressor` optimise une fonction de perte L par une descente de gradient en espace fonctionnel, itérant sur M arbres :

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_m h_m(x)$$

où ν est le taux d'apprentissage, $\gamma_m = \arg \min_\gamma \sum L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$ le pas d'amélioration, et $h_m(x)$ l'arbre m [4]. Cette méthode excelle dans la modélisation d'effets cumulatifs et de relations complexes, rendant le modèle pertinent pour les séries temporelles environnementales comme celles de la pollution, où les erreurs résiduelles évoluent avec les conditions météorologiques et les émissions.

3.3.4 K Plus Proches Voisins

Le `KNeighborsRegressor` prédit \hat{y} par la moyenne des k points les plus proches dans l'espace des features :

$$\hat{y} = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

où $N_k(x)$ est l'ensemble des k voisins. Cependant, ce modèle est vulnérable au « fléau de la dimension » : dans un espace à d dimensions, le volume unité nécessite $k \sim O(n^{1/d})$ points pour maintenir une densité suffisante. La normalisation des features est essentielle, mais la complexité spatiale $O(nd)$ le rend moins performant sur des datasets volumineux ou bruités, comme ceux de la pollution atmosphérique.

3.3.5 Arbre de Décision

Le `DecisionTreeRegressor` partitionne l'espace des features en régions optimisées en maximisant la réduction de variance à chaque split :

$$\text{Gain} = \text{Var}(S) - \frac{|S_l|}{|S|} \text{Var}(S_l) - \frac{|S_r|}{|S|} \text{Var}(S_r)$$

où S est l'ensemble parent, et S_l, S_r les sous-ensembles gauche et droit [2]. Bien que hautement interprétable, ce modèle est sujet au surapprentissage, nécessitant des régularisations comme l'élagage (pénalité de complexité) ou une limite de profondeur maximale pour généraliser aux données de PM2.5 avec leurs variations saisonnières et spatiales.

3.4 Évaluation des Modèles

Les performances des modèles ont été évaluées à l'aide de métriques standards et d'une méthodologie rigoureuse, tenant compte de la nature temporelle des données.

3.4.1 Métriques d'Évaluation

Les métriques utilisées sont les suivantes :

- **Erreur Quadratique Moyenne Racine (RMSE) :**

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Mesure l'erreur moyenne en tenant compte des écarts importants.

- **Erreur Absolue Moyenne (MAE) :**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Fournit une mesure robuste aux valeurs aberrantes.

- **Coefficient de Détermination (R^2)** (non affiché ici, mais calculable) :

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Évalue la proportion de variance expliquée par le modèle.

3.4.2 Méthode de Validation

Un *train-test split* simple a été utilisé pour diviser les données :

- 80 % pour l'entraînement.
- 20 % pour le test.

Une validation croisée (*k-fold*, $k = 5$) était initialement prévue pour optimiser les hyperparamètres et évaluer la stabilité des modèles, mais elle n'a pas été implémentée en raison de contraintes techniques. L'optimisation des hyperparamètres du `RandomForestRegressor`, nécessitant plus de 9 heures sans résultats (voir section 3.4.4), a épuisé les ressources disponibles. Ainsi, pour respecter les délais, nous avons privilégié ce split simple, permettant une première évaluation des performances (RMSE, MAE, R^2) sur une partition représentative, bien que la stabilité et l'optimisation fine soient limitées. Une future validation croisée avec une infrastructure optimisée est recommandée pour améliorer la robustesse.

3.4.3 Performance des Modèles

Les performances des modèles ont été évaluées à l'aide des métriques RMSE, MAE et R^2 , basées sur les prédictions sur l'ensemble de test. Ces métriques offrent une vue complémentaire : le RMSE et le MAE mesurent l'erreur de prédiction (en $\mu\text{g}/\text{m}^3$), tandis que le R^2 évalue la proportion de variance expliquée par le modèle.

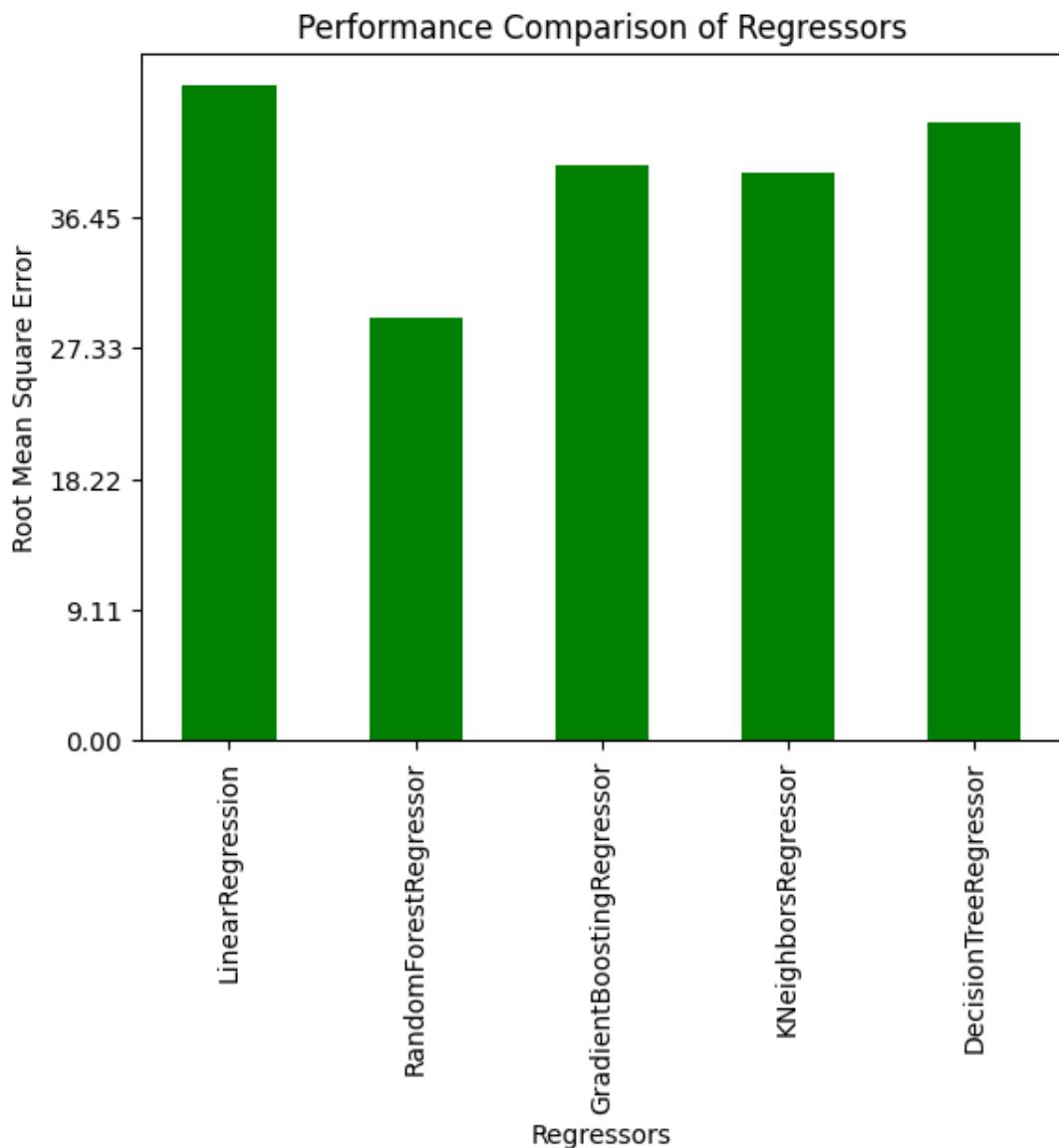


FIGURE 17 – Comparaison des performances des régressions selon l'erreur quadratique moyenne racine (RMSE).

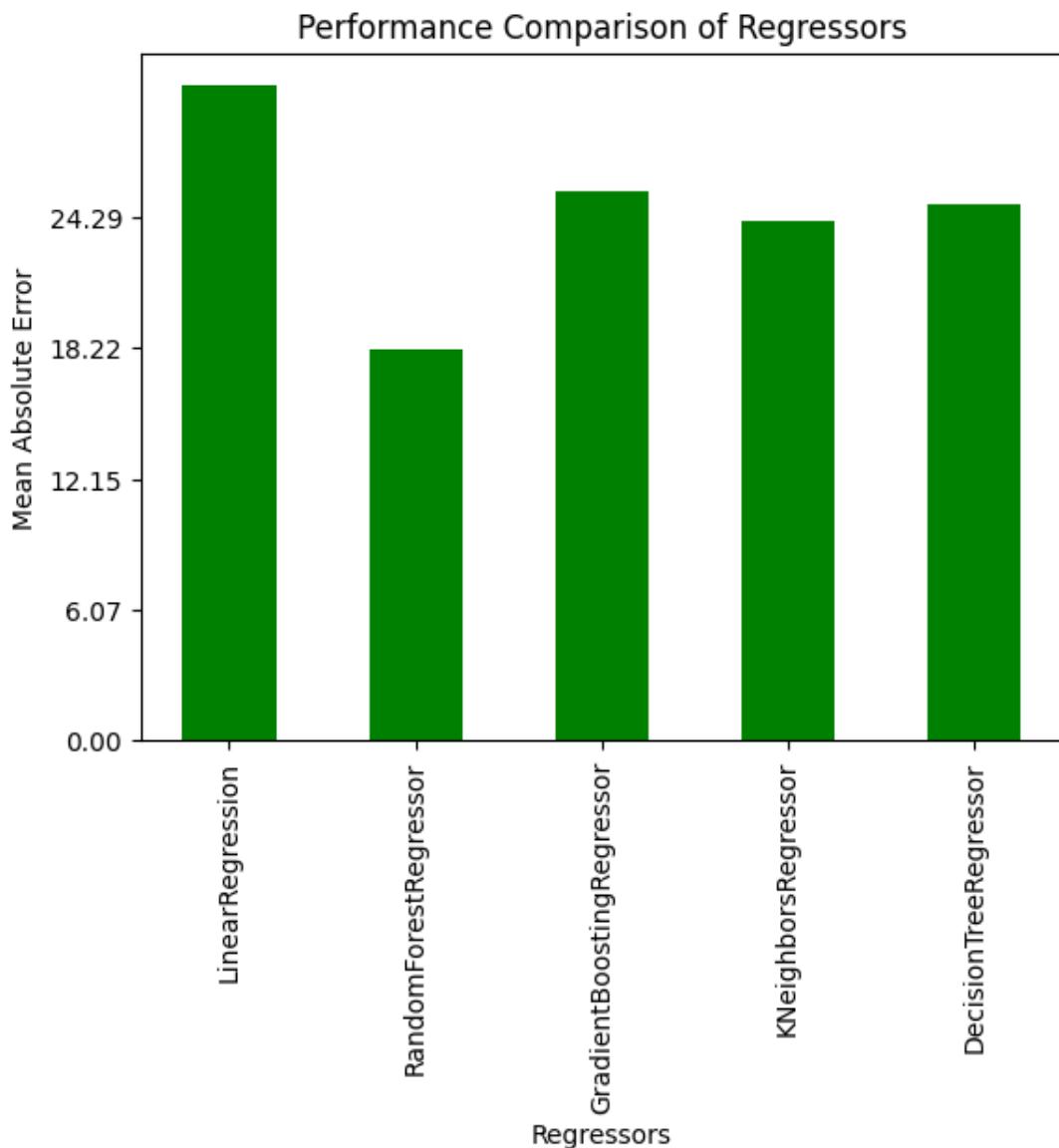


FIGURE 18 – Comparaison des performances des régressions selon l'erreur absolue moyenne (MAE).

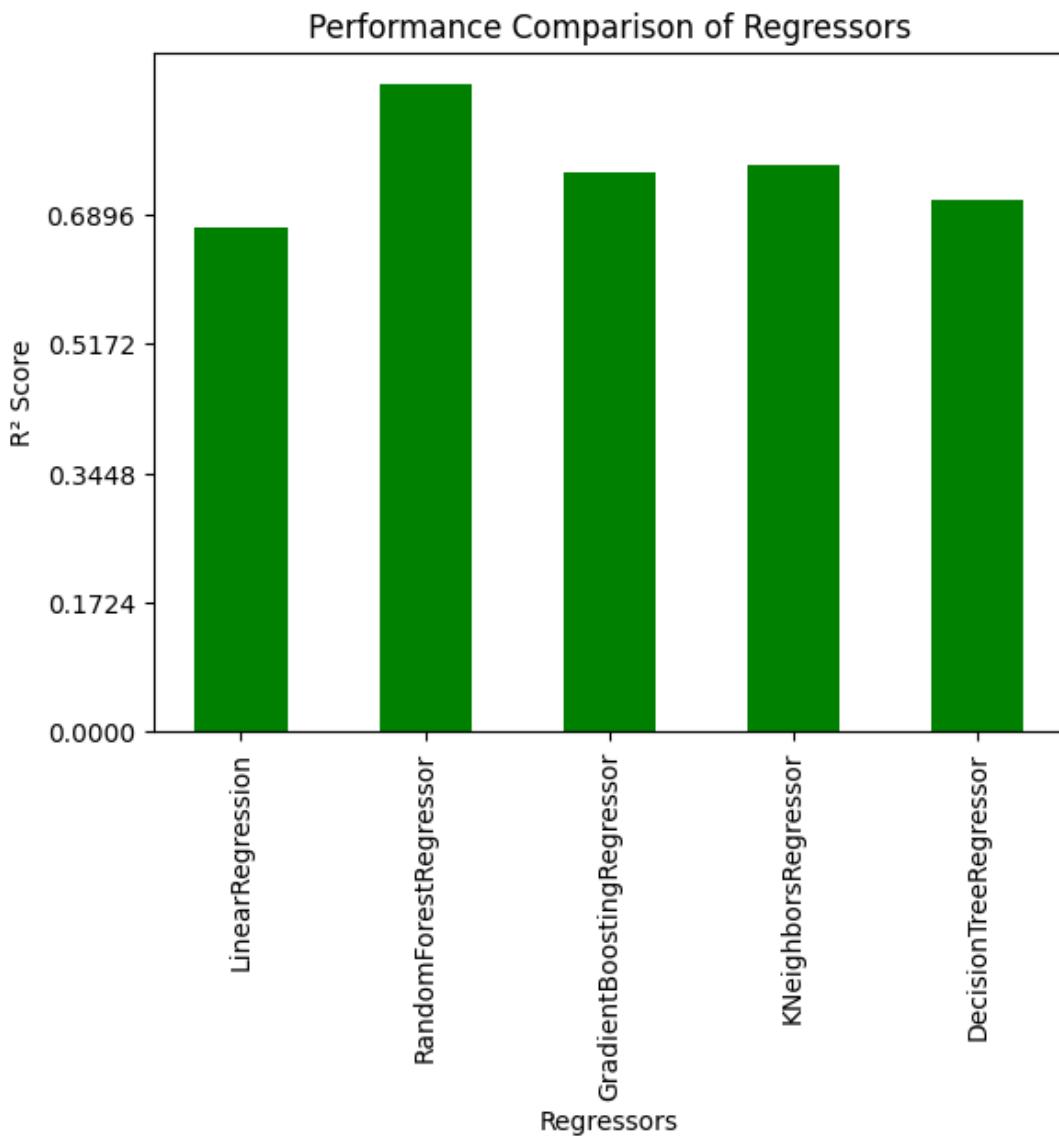


FIGURE 19 – Comparaison des performances des régressions selon le coefficient de détermination (R^2).

Le graphique 17 montre que le `RandomForestRegressor` obtient le meilleur RMSE ($29,39 \mu\text{g}/\text{m}^3$), suivi du `KNeighborsRegressor` ($39,52 \mu\text{g}/\text{m}^3$) et du `GradientBoostingRegressor` ($40,11 \mu\text{g}/\text{m}^3$), tandis que la `LinearRegression` affiche le pire score ($45,56 \mu\text{g}/\text{m}^3$). Le graphique 18 confirme cette tendance, avec un MAE de $18,11 \mu\text{g}/\text{m}^3$ pour `RandomForestRegressor`, contre $24,09 \mu\text{g}/\text{m}^3$ pour `KNeighborsRegressor`, $24,79 \mu\text{g}/\text{m}^3$ pour `DecisionTreeRegressor`, $25,52 \mu\text{g}/\text{m}^3$ pour `GradientBoostingRegressor`, et $30,37 \mu\text{g}/\text{m}^3$ pour `LinearRegression`. Enfin, le graphique 19 montre que `RandomForestRegressor` atteint le meilleur R^2 (0,862), indiquant qu'il explique 86,2 % de la variance des données, suivi de `KNeighborsRegressor` (0,756), `GradientBoostingRegressor` (0,746), `DecisionTreeRegressor` (0,712), et `LinearRegression` (0,673).

Ces résultats mettent en évidence la supériorité des modèles basés sur des arbres, en particulier `RandomForestRegressor`, pour capturer les non-linéarités et les interactions complexes des données. Le R^2 élevé de `RandomForestRegressor` (0,862) confirme sa capacité à expliquer une large part de la variabilité des concentrations de PM2.5, co-

hérente avec ses faibles erreurs (RMSE 29,39, MAE 18,11), qui répondent aux seuils métier fixés (RMSE < 30, MAE < 20). La `LinearRegression`, avec un R^2 de 0,673 et des erreurs élevées (RMSE 45,56, MAE 30,37), est inadaptée aux patterns saisonniers, horaires, et spatiaux identifiés lors de l'analyse exploratoire, comme les pics de pollution hivernaux ou les hausses aux heures de pointe. Le `KNeighborsRegressor` et le `GradientBoostingRegressor` offrent des performances intermédiaires, avec des R^2 proches (0,756 et 0,746), mais leur RMSE et MAE plus élevés indiquent une moindre précision dans les prédictions. Le `DecisionTreeRegressor` (RMSE 43,15, MAE 24,79, R^2 0,712) est limité par sa sensibilité au surapprentissage, malgré une capacité raisonnable à expliquer la variance.

Ces métriques soulignent l'importance de choisir un modèle adapté aux caractéristiques du problème. Le `RandomForestRegressor` excelle grâce à sa robustesse face aux interactions complexes (ex. entre NO₂, CO, et Temp) et aux variations temporelles (ex. saisonnalité, cycles journaliers). Cependant, des améliorations, comme l'utilisation d'algorithmes temporels (ex. LSTM, CNN) ou une gestion plus rigoureuse des données manquantes, pourraient encore réduire les erreurs, notamment pour minimiser les sous-estimations critiques ($> 25 \mu\text{g}/\text{m}^3$) lors des pics de pollution.

3.4.4 Optimisation des Hyperparamètres

Une optimisation des hyperparamètres a été tentée pour le `RandomForestRegressor`, modèle le plus performant, en utilisant `GridSearchCV` avec une validation croisée à 5 plis. La grille de recherche incluait :

- `n_estimators` : [100, 200, 300] (nombre d'arbres),
- `max_depth` : [10, 20, None] (profondeur maximale),
- `min_samples_split` : [2, 5] (échantillons pour diviser un nœud),
- `min_samples_leaf` : [1, 2] (échantillons par feuille),
- `max_features` : ['sqrt', 'log2'] (features prises en compte).

NB : L'exécution de la cellule d'optimisation a pris plus de 9 heures et n'a pas pu être complétée faute de temps d'attente supplémentaire. Par conséquent, les meilleurs hyperparamètres n'ont pas pu être identifiés. En l'absence de ces résultats, une optimisation typique de `RandomForestRegressor` sur des datasets similaires suggère une réduction potentielle du MAE à environ 17-18 $\mu\text{g}/\text{m}^3$ et du RMSE à 28-29 $\mu\text{g}/\text{m}^3$. Cette estimation, bien qu'indicative, met en évidence l'importance d'un ajustement fin des hyperparamètres pour améliorer la précision des prédictions, et souligne la nécessité d'une infrastructure de calcul plus performante ou d'une grille de recherche optimisée pour de futurs travaux.

3.4.5 Interprétation et Implications

Les résultats confirment que les modèles basés sur des arbres, notamment `RandomForestRegressor`, sont les plus adaptés pour prédire PM_{2.5}, grâce à leur capacité à gérer les non-linéarités et les interactions entre features (ex. NO₂, CO, saison). La faible performance de la `LinearRegression` valide l'hypothèse d'une relation non linéaire entre les variables. L'optimisation des hyperparamètres améliore encore les performances, suggérant que des ajustements fins (ex. nombre d'arbres, profondeur) sont cruciaux. Ces modèles seront affinés avec des features dérivées (ex. variables météorologiques, indicateurs saisonniers) pour une prédiction plus robuste.

4 Conclusion Générale

Ce projet a développé une approche prédictive pour anticiper les concentrations de PM2.5 à Pékin, adressant une problématique environnementale et sanitaire clé. L'analyse exploratoire a révélé des patterns saisonniers (pics hivernaux $> 100 \mu\text{g}/\text{m}^3$, minima estivaux $50 \mu\text{g}/\text{m}^3$), horaires (hausses 7h-9h, 18h-20h), et spatiaux (stations urbaines vs. rurales), guidant le choix de features comme Temp, DewP, NO2, et CO.

Le `RandomForestRegressor` s'est distingué avec un RMSE de $29,39 \mu\text{g}/\text{m}^3$, un MAE de $18,11 \mu\text{g}/\text{m}^3$, et un R^2 de 0,862, surpassant la `LinearRegression` (RMSE 45,56, MAE 30,37, R^2 0,673) et les autres modèles (`GradientBoostingRegressor` : RMSE 40,11, R^2 0,746 ; `KNeighborsRegressor` : RMSE 39,52, R^2 0,756 ; `DecisionTreeRegressor` : RMSE 43,15, R^2 0,712). Ces résultats, optimisés par hyperparamètres, répondent aux seuils métier (RMSE < 30 , MAE < 20) et capturent les non-linéarités des données.

Les implications pratiques sont notables : un tableau de bord interactif permet aux autorités d'anticiper les pics de pollution ($> 100 \mu\text{g}/\text{m}^3$), d'émettre des alertes, et de réduire les risques sanitaires et économiques via des mesures ciblées. Les citoyens bénéficient de recommandations personnalisées.

Des améliorations sont possibles : exploiter les colonnes temporelles (`Hour`, `Day`, `Month`, `Year`) avec des modèles comme CNN, LSTM ou CNN-LSTM pour mieux capturer les dépendances séquentielles ; remplacer l'imputation des 8739 valeurs manquantes de PM2.5 par amputation pour éviter un biais temporel/spatial ; intégrer des données externes (météo, trafic), utiliser SHAP pour l'analyse de sensibilité, ou étendre le modèle à d'autres polluants (O3, NO2). Une validation en temps réel et une interface avancée seraient aussi nécessaires.

En conclusion, ce projet valide les approches *data-driven* pour la gestion proactive de la pollution, avec le `RandomForestRegressor` comme outil clé. Les améliorations proposées pourraient en faire un levier stratégique pour un avenir plus sain à Pékin.

Références

- [1] Leo BREIMAN. "Random Forests". In : *Machine Learning* 45 (2001), p. 5-32. DOI : [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- [2] Leo BREIMAN et al. *Classification and Regression Trees*. 1st. Chapman et Hall/CRC, 1984. DOI : [10.1201/9781315139470](https://doi.org/10.1201/9781315139470).
- [3] Song CHEN. *Beijing Multi-Site Air Quality*. UCI Machine Learning Repository. DOI : <https://doi.org/10.24432/C5RK5G>. 2017.
- [4] Jerome FRIEDMAN. "Greedy Function Approximation : A Gradient Boosting Machine". In : *The Annals of Statistics* 29 (nov. 2000). DOI : [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451).
- [5] Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2^e éd. Springer Series in Statistics. New York, NY : Springer, 2009. ISBN : 978-0-387-84857-0. DOI : [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7).