

Projet réalisé par



RAPPORT
SCIENCES DES DONNÉES 4

Collaborateurs : Anthony Combes-Aguera, Mohamed Rekhis Chaouki, Ayoub Akkouh et Wassim Harraga



Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Soumis comme contribution partielle pour le cours Sciences des Données 4

Table des matières

1 Introduction et Contexte du Projet PureOxy	3
1.1 Historique et Objectifs Initiaux	3
1.2 Version du Semestre Précédent	3
1.3 Évolution et Nouveautés du Semestre Actuel	3
2 Ajout des Données Prédictives	5
2.1 Détails de la création des données prédictives	5
2.2 Intégration des Données Prédictives et Historiques sur le Site	6
3 Refonte de la Base de Données et Nouveau Modèle de Données	7
3.1 Présentation des Sources de Données	7
3.2 Nouvelles Tables et Modifications	7
3.2.1 Table <code>all_years_cleaned_daily</code>	7
3.2.2 Table <code>prediction_cities</code>	8
3.2.3 Table <code>donnees_villes</code>	8
3.2.4 Table <code>moy_pollution_villes</code>	9
3.3 Schéma Relationnel et Avantages de la Nouvelle Structure	9
3.4 Présentation du Modèle Physique et du Modèle Conceptuel (MCD)	9
4 Amélioration des Interfaces Utilisateurs et des Fonctionnalités	11
4.1 Évolution de la Page Détails	11
4.2 Ajout de 2 nouvelles pages : Classement et Compare	13
4.3 Refonte du Système de Commentaires et de Favoris	15
4.4 Refonte de la Page d'Accueil et du Menu de Navigation	16
4.5 Refonte du Menu de Navigation	16
5 Intégration et Optimisations Fonctionnelles	17
5.1 Optimisations de Performances	17
5.2 Intégration des Données Prédictives	17
5.3 Optimisation de la Carte Interactive	17
5.4 Intégration du Chatbot IA Mistral	18
6 Défis Techniques et Solutions Apportées	19
6.1 Optimisation des Requêtes et de l'Interface	19
6.2 Principaux Défis Techniques	19
7 Exemple d'utilisation de PureOxy : Analyse comparative des données Historique et Normalisée (Habitants, km²)	20
7.1 Sélection des Indicateurs et Méthodologie	20
7.2 Résultats et Observations Clés	20
7.3 Interprétation Générale et Apport de PureOxy	23
7.4 Conclusion : Une Plateforme Complète d'Aide à la Décision	23

8	Détails Techniques et Architecture du Code	24
9	Conclusion	25

Introduction et Contexte du Projet PureOxy

Dans un contexte où la qualité de l'air représente à la fois un enjeu de santé publique et un défi environnemental majeur, le projet **PureOxy** a été conçu afin de fournir un outil de visualisation, d'analyse et de prédition des niveaux de pollution atmosphérique dans les villes françaises. Ce rapport retrace, de manière chronologique et détaillée, l'évolution du projet depuis la version du semestre précédent jusqu'aux nombreuses améliorations et optimisations apportées lors du semestre actuel.

1.1 Historique et Objectifs Initiaux

Au semestre précédent, la plateforme **PureOxy** proposait déjà une carte interactive accompagnée de graphiques permettant de visualiser la pollution atmosphérique. Les objectifs étaient les suivants :

- **Visualisation des données environnementales** : Une carte interactive affichant des indicateurs de pollution par ville.
- **Création d'outils informatifs** : Graphiques simples et statistiques pour observer l'évolution de la pollution.
- **Sensibilisation** : Informer les citoyens et les décideurs sur la qualité de l'air.

1.2 Version du Semestre Précédent

La version initiale de **PureOxy** comportait :

- Une **carte interactive** permettant de visualiser les données de pollution pour diverses villes françaises.
- Des **graphes statistiques** fournissant des représentations simples des niveaux de pollution.
- Une base de données fusionnant les mesures de pollution (extraites de jeux de données publics tels que OpenAQ) avec des informations géographiques.

Bien que fonctionnelle, cette version offrait une vision limitée et ne permettait pas d'intégrer des fonctionnalités de prédition ou d'analyse avancée.

1.3 Évolution et Nouveautés du Semestre Actuel

Face aux nouvelles exigences du cours de *Sciences des Données 4*, le semestre actuel a vu l'intégration de multiples fonctionnalités de data science et l'amélioration significative de l'existant. Parmi les innovations majeures, on peut citer :

- **Création de la page Classement** : Ce qui était initialement pensé comme un podium sur la page d'accueil (les trois villes les plus polluées) a évolué vers une page dédiée regroupant l'ensemble du classement par polluant, incluant un filtrage et un tri précis.
- **Transformation de l'onglet Comparaison en page Compare** : L'outil de comparaison intégré dans la page détails a été extrait pour devenir une page *Compare* dédiée, permettant une analyse comparative approfondie entre un nombre illimité de villes.
- **Intégration de données prédictives** : Préparation et ajout de données journalières sur 5 ans pour plus de 275 villes, permettant de générer des prévisions sur une période allant de janvier 2025 à janvier 2026.

- **Refonte de la base de données** : Création de nouvelles tables pour optimiser le chargement (notamment `all_years_cleaned_daily`, `prediction_cities`, `donnees_villes` et `moy_pollution_villes`).
- **Optimisations de l'interface utilisateur** : Refonte de la page d'accueil, de la page détails (avec système d'onglets et filtres dynamiques) et amélioration générale du menu de navigation.
- **Intégration d'un chatbot IA** : Développement et intégration d'un chatbot basé sur l'IA Mistral, accessible depuis toutes les pages du site.

Ajout des Données Prédictives

2.1 Détails de la création des données prédictives

Pour répondre à la demande d'intégration de fonctionnalités de data science, l'équipe a entrepris les travaux suivants, menés par Ayoub en collaboration avec Wassim :

Pour la conception, l'entraînement et l'évaluation des différents modèles de machine learning, nous avons principalement travaillé avec le langage Python. Les librairies pandas et NumPy ont ainsi servi à la manipulation et à l'analyse de données (nettoyage, création de nouvelles variables, etc.), tandis que sklearn s'est avéré indispensable pour la sélection et l'entraînement de modèles classiques (régression linéaire, arbres de décision, RandomForest, etc.).

De plus, pour exploiter la puissance et la flexibilité des algorithmes de gradient boosting, nous avons eu recours à XGBoost, qui propose un compromis intéressant entre rapidité, performance et capacité de traitement de gros volumes de données.

1. Préparation et Nettoyage des Données : Un jeu de données regroupant des mesures journalières sur 5 ans pour plus de 275 villes a été sélectionné. Les étapes incluaient :

- Conversion des dates au format `datetime`.
- Tri par ville, polluant et date.
- Suppression des valeurs manquantes.
- Harmonisation des noms de villes via un algorithme de *fuzzy matching* pour corriger la segmentation en secteurs.
- Encodage one-hot pour les variables catégorielles.

2. Feature Engineering : Afin de mieux prendre en compte l'aspect séquentiel et cyclique des données, j'ai enrichi le jeu de données avec deux séries de variables permettant de modéliser le comportement temporel. D'abord, les variables `lag_1` et `lag_2` reproduisent la valeur mesurée un jour et deux jours auparavant. Elles permettent au modèle d'identifier plus facilement les relations d'autocorrélation : si un niveau de pollution est élevé un jour donné, il est susceptible de rester élevé (ou de baisser/moduler) le jour suivant, et ainsi de suite. De ce fait, l'algorithme apprend à repérer des tendances à court terme et à adapter ses prédictions.

En parallèle, l'introduction d'informations calendaires (comme le jour de la semaine et le mois) sert à capturer les habitudes et cycles récurrents (activité quotidienne, différences entre les jours travaillés et les week-ends, saisonnalité, etc.). Par exemple, un pic de pollution peut survenir de façon plus marquée en hiver (mois) ou en milieu de semaine (jour de la semaine). Le modèle peut donc assimiler ces schémas temporels et prévoir plus finement les fluctuations de la variable cible.

3. Modélisation : Afin de produire des prévisions fiables pour l'année 2024, nous avons mis en place un protocole de modélisation rigoureux. Plusieurs algorithmes de machine learning (régression linéaire, RandomForestRegressor, XGBoost) ont été entraînés sur les données couvrant la période 2022–2023, intégrant ainsi les tendances récentes et la dynamique saisonnière observées. Pour évaluer leurs performances, nous avons utilisé différentes métriques : le RMSE (erreur quadratique moyenne), le R² (coefficient de détermination) et le MAE (erreur absolue moyenne). Ces indicateurs offrent une vision complémentaire : le RMSE et le MAE mesurent l'ampleur des erreurs de prédiction, tandis que le R² reflète la proportion de variance expliquée par le modèle.

Au cours du processus de sélection, une attention particulière a été portée à la robustesse et à la capacité de généralisation des modèles. Les résultats obtenus sur les données de validation ont permis d'identifier l'algorithme le plus performant, c'est-à-dire celui dont les prédictions s'approchent le plus des mesures réelles de 2024. Ce modèle (RandomForestRegressor) sera ensuite retenu comme modèle de prédiction final et déployé pour alimenter les projections à plus long terme.

4. **Prévision Multi-Step** : Pour anticiper les valeurs jour par jour sur toute l'année 2024, nous avons adopté une stratégie de prévision multi-step. Concrètement, à partir de la dernière mesure réelle disponible fin 2023, un processus itératif a été mis en place : la valeur prédite pour le premier jour de 2024 a servi de base (en tant que lag 1 et lag 2 selon le décalage) pour la prévision du jour suivant, et ainsi de suite. Autrement dit, chaque nouvelle prévision, une fois calculée, est réinjectée dans le modèle pour générer la prévision suivante. Cette mise à jour dynamique des variables de retard permet de simuler plus fidèlement l'évolution du phénomène au fil des jours, même en l'absence de données réelles pour 2024. Bien que cette approche puisse entraîner une propagation des erreurs (les écarts s'accumulant potentiellement jour après jour), elle reste un moyen efficace d'obtenir des projections quotidiennes cohérentes sur une période prolongée.
5. **Optimisation et Résolution des Problèmes** : Les processus de prévision, notamment l'approche multi-step, ont parfois généré des récurrences ou répétitions dans les prédictions, nuisant à la dynamique globale du modèle. Pour y remédier, nous avons ajusté plusieurs paramètres : affinement des lags, révision des hyperparamètres (par exemple, la profondeur des arbres pour XGBoost ou la taille des forêts aléatoires) et tests de différentes configurations de variables explicatives. Ces optimisations visaient à améliorer la réactivité et la cohérence des prédictions sur plusieurs jours consécutifs, tout en préservant la robustesse et la performance des modèles. Bien que ces ajustements se soient parfois traduits par des temps d'exécution plus longs, ils ont permis de maintenir des résultats plus fiables et plus proches des tendances observées.

2.2 Intégration des Données Prédictives et Historiques sur le Site

Après la génération des prévisions, de nouvelles données ont été intégrées sur la plateforme :

- Mise en place d'un **onglet Prédiction** dans la page détails, affichant les prévisions pour une période allant de janvier 2025 à janvier 2026.
- Ajout de **2 ans de données journalières historiques** pour chaque ville, remplaçant les anciennes mesures limitées (une moyenne de 1,7 données historiques journalières par ville pour plus 5475 données historiques journalières par villes actuellement).

Ces ajouts ont considérablement augmenté le volume de données et, dans un premier temps, provoqué de gros ralentissements lors du chargement des pages, notamment sur la page détails et la carte interactive.

Refonte de la Base de Données et Nouveau Modèle de Données

3.1 Présentation des Sources de Données

Les nouvelles données intégrées dans PureOxy proviennent de sources publiques reconnues et variées, garantissant la fiabilité et l'actualité des informations.

Données géographiques : Les informations concernant les communes et villes de France ont été obtenues via la plateforme *DataGouv*. Ces données regroupent la liste complète des communes françaises avec leurs codes d'identification et d'autres informations géographiques essentielles.

- **Sources utilisées :** INSEE, geo.api.gouv.fr, Ministère de l'Éducation, La Poste.
- Ces données permettent de compléter les informations géographiques de la table `donnees_villes` et d'assurer une adéquation avec les mesures environnementales.

Données journalières des concentrations de polluants : Les données « temps réel » de mesure des concentrations de polluants atmosphériques réglementés sont issues de la plateforme *DataGouv* et proviennent du Laboratoire Central de Surveillance de la Qualité de l'Air (LCSQA).

- **Source :** Données fournies par les 18 Associations Agréées de Surveillance de la Qualité de l'Air (AASQA) et transmises au LCSQA pour être intégrées dans la base nationale *Geod'air*.
- Ces données sont essentielles pour alimenter les tables historiques (`all_years_cleaned_daily`), et serviront à créer les données prédictives (`prediction_cities`), assurant une analyse précise de la qualité de l'air en France.

Afin d'optimiser les temps de chargement et d'assurer une gestion efficace du grand volume de données, une refonte complète de la base de données a été réalisée.

3.2 Nouvelles Tables et Modifications

Plusieurs tables ont été créées ou modifiées afin de séparer clairement les données historiques, les données prédictives et les informations géographiques.

3.2.1 Table `all_years_cleaned_daily`

Utilisation : Exclusivement pour l'onglet historique.

Contenu : Données journalières historiques pour 276 villes sur 2 ans.

Variable	Description
year	Année de la mesure
jour	Date de la mesure (format YYYY-MM-DD, ex : 2025-01-18)
ville	Nom de la ville (format harmonisé)
polluant	Nom du polluant mesuré
unite_de_mesure	Unité de mesure (toujours µg/m³)
valeur_journaliere	Concentration mesurée historique du polluant
id_ville	Identifiant unique de la ville, commun à toutes les tables

TABLE 1 – Variables de la table `all_years_cleaned_daily`

3.2.2 Table prediction_cities

Utilisation : Exclusivement pour l'onglet prédictions.

Contenu : Données journalières prédictives pour 276 villes sur 1 an.

Variable	Description
jour	Date de la mesure prédictive (format YYYY-MM-DD)
ville	Nom de la ville (format harmonisé)
polluant	Nom du polluant concerné
valeur_predite	Concentration prédictive du polluant
id_ville	Identifiant unique de la ville

TABLE 2 – Variables de la table prediction_cities

3.2.3 Table donnees_villes

Utilisation : Ancienne table principale pollution_villes ne contenant plus que des données géographiques.

Contenu : Informations géographiques et complémentaires (population, superficie, densité, descriptif).

Variable	Description
id_ville	Identifiant unique de la ville
ville	Nom de la ville
postal_code	Code postal de la ville
latitude	Latitude géographique
longitude	Longitude géographique
departement	Nom du département
region	Nom de la région
population	Chiffre de population
superficie_km2	Superficie en km ²
densite	Densité de la ville
grille_densite_texte	Descriptif catégorique (ex : Grands centres urbains, Ceintures urbaines, etc.)

TABLE 3 – Variables de la table donnees_villes

3.2.4 Table moy_pollution_villes

Utilisation : Permet d'optimiser les temps de chargement via le calcul préalable des moyennes.

Contenu : Diverses moyennes calculées pour chaque ville, chaque mois et par polluant.

Variable	Description
id_ville	Identifiant de la ville.
polluant	Nom du polluant.
avg_value	Moyenne globale historique.
avg_par_habitant	Moyenne ajustée par habitant.
avg_par_km2	Moyenne ajustée par km ² .
de moy_janv2023 à moy_janv2025	Moyennes mensuelles les données historiques.
de moy_predic_janv2025 à moy_predic_janv2026	Moyennes mensuelles pour les données prédictives.

TABLE 4 – Variables de la table moy_pollution_villes

3.3 Schéma Relationnel et Avantages de la Nouvelle Structure

La refonte de la base de données a permis :

- Une séparation claire des données historiques et prédictives.
- Une optimisation du calcul des moyennes via la table moy_pollution_villes.
- Une amélioration notable des temps de chargement, notamment sur la carte interactive et la page détails.

3.4 Présentation du Modèle Physique et du Modèle Conceptuel (MCD)

Pour accompagner la refonte de la base de données, nous avons également revu le modèle conceptuel et physique afin d'optimiser la gestion et l'accès aux données. La nouvelle version du modèle (version 2) présente une meilleure séparation entre les données historiques, les données prédictives et les informations géographiques.

Modèle Physique : Le schéma physique a été repensé pour intégrer les tables all_years_cleaned_daily, prediction_cities, donnees_villes et moy_pollution_villes. Cette refonte permet notamment une optimisation du calcul des moyennes et une amélioration significative des temps de chargement.

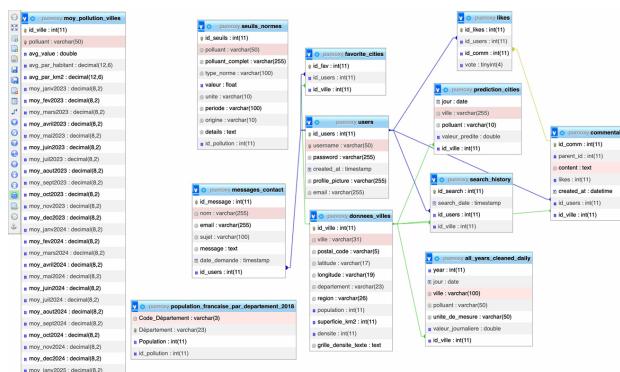


FIGURE 1 – Schéma physique (Modèle physique) de la nouvelle structure de la base de données.

Modèle Conceptuel (MCD) : Le MCD a été mis à jour pour refléter les nouvelles entités et relations, assurant une cohérence et une intégration optimale entre les différentes sources de données.

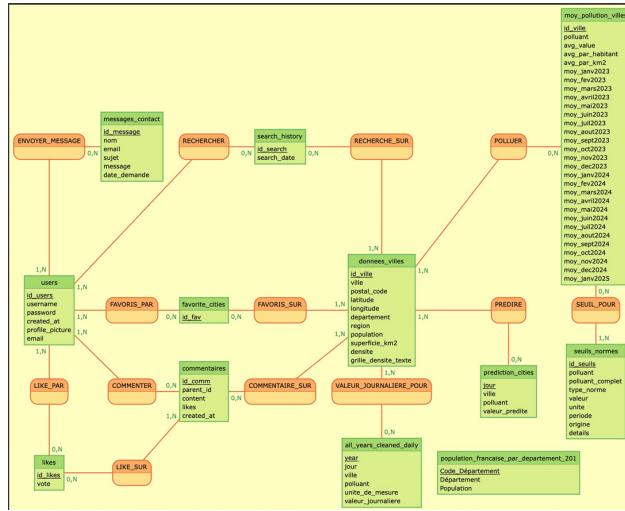


FIGURE 2 – Modèle Conceptuel de Données (MCD) correspondant à la nouvelle structure.

Amélioration des Interfaces Utilisateurs et des Fonctionnalités

La refonte du projet ne s'est pas limitée aux données ; l'expérience utilisateur a été repensée.

4.1 Évolution de la Page Détails

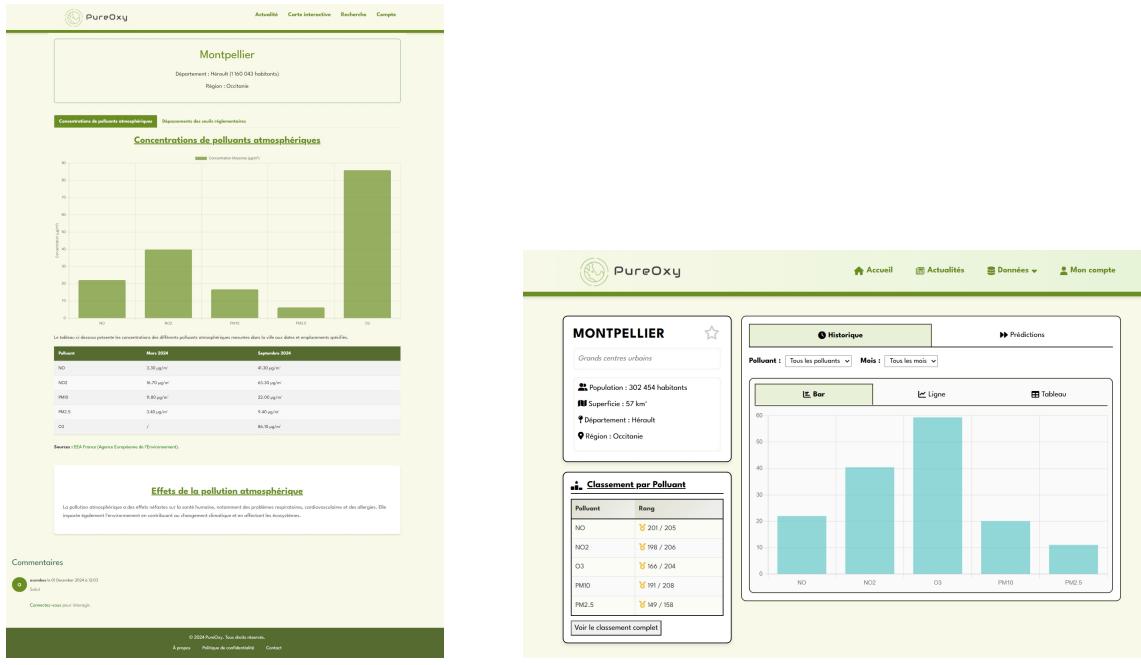


FIGURE 3 – Comparaison des différentes version de la page Détails.

La page détails a subi une refonte majeure pour répondre aux attentes de clarté et d'interactivité. Cette page se découpe en 2 parties :

- **Première Partie** : Offre à l'utilisateur un contexte permettant une meilleure compréhension des données.
 - **Carte de présentation de la ville améliorée** : La carte de présentation de la ville est composée d'informations telles que le descriptif de la densité (usage de la variable `grille_densite_texte`), la population, la superficie, le département et la région.
 - **Classement par polluant** : Le classement par polluant affiche le rang de la ville. Il est trié par ordre croissant. Dans la capture d'écran, le classement de Montpellier montre que la ville est la 4ème ville la plus pollué en NO, 8ème ville la plus pollué en NO₂, etc.. Ce classement se base uniquement sur les villes disponibles dans notre base de données. Un bouton permet d'accéder à la page complète de classement des villes.

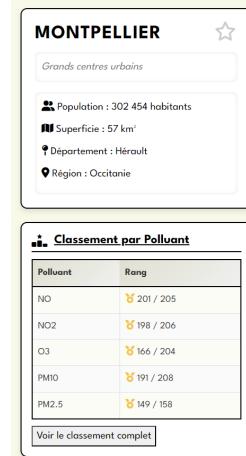


FIGURE 4 – Capture d’écran de la carte de présentation de la ville Montpellier.

- **Seconde Partie** : Donne à l’utilisateur une visualisation des données.
 - **Système d’onglets** : Deux onglets principaux : Historique et Prédictions. Ces 2 onglets sont composés des sous-onglets Bar, Ligne et Tableau pour une visualisation détaillée.

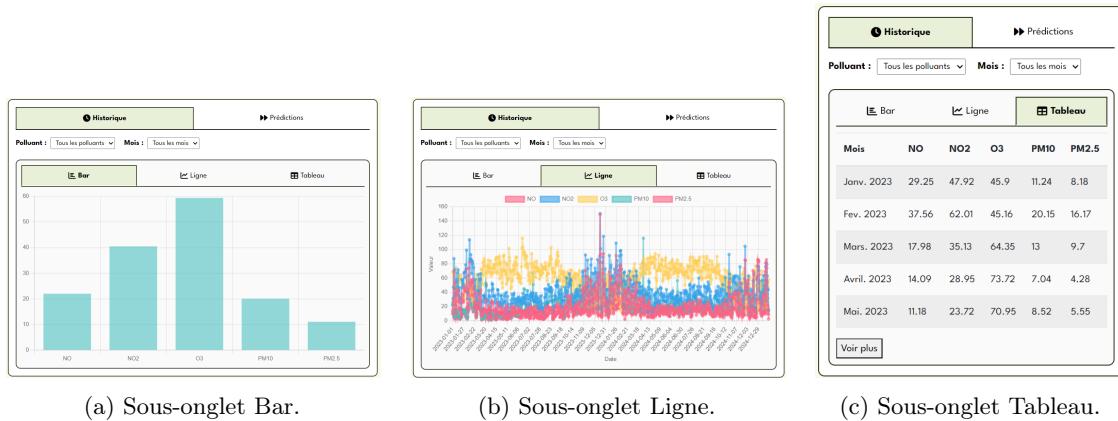


FIGURE 5 – Captures d’écran des sous onglets de la page Détails.

- **Filtres dynamiques** : Des filtres par polluant et par mois permettent de restreindre l’affichage des données, avec des plages différencierées pour les données historiques (de janvier 2023 à janvier 2025) et prédictives (de janvier 2025 à janvier 2026).

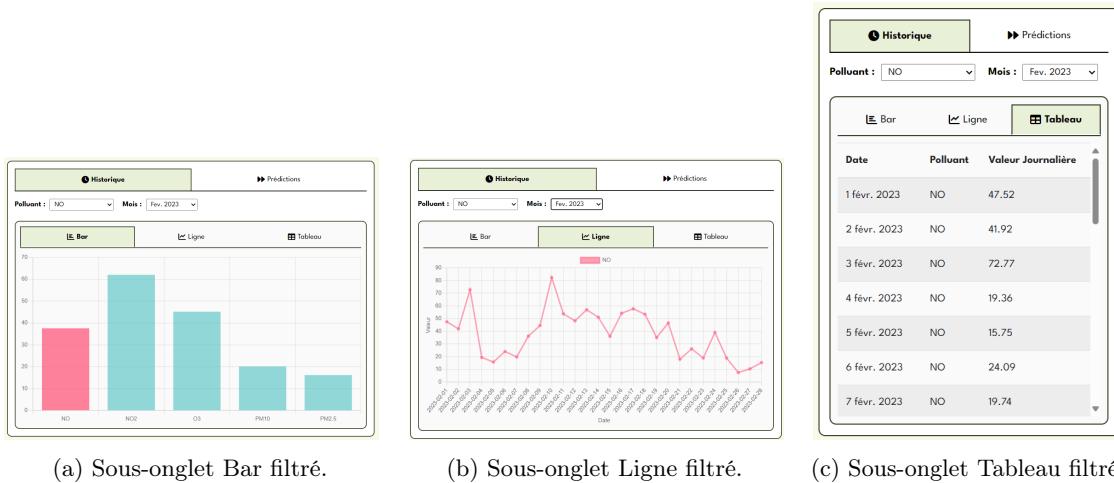


FIGURE 6 – Captures d’écran des sous onglets de la page Détails filtré.

4.2 Ajout de 2 nouvelles pages : Classement et Compare

- **Page Classement :** Au départ, l’idée d’un classement a été réalisé sous la forme d’un podium situé dans la page d’accueil. Au fil du semestre, ce podium a évolué en une page classement complète. Cette page permet :
 - Un filtrage permettant de sélectionner le polluant souhaité.
 - Le descriptif des polluants et des dépassements de seuils.



FIGURE 7 – Capture d’écran de la première partie de la page classement.

- L'affichage d'un classement des villes pour chaque polluant.
- L'utilisation des options de pagination et de tri (croissant, décroissant, alphabétique).
- La redirection vers la page details dédié pour chaque ville.

Rang / Ville	Moy. ($\mu\text{g}/\text{m}^3$)	Moy. par hab	Moy. par km 2
1. Saint-Michel-l'Observatoire	28.06	0.0229	1.0023
2. Le Petit-Quevilly	25.61	0.0012	6.4021
3. Bordeaux	24.10	0.0001	0.4820
4. Pantin	23.86	0.0004	4.7717
5. Montpellier	22.01	0.0001	0.3862

FIGURE 8 – Capture d’écran d’un extrait de la seconde partie de la page classement.

- **Page Compare** : Précédemment intégrée dans la page Détails via un onglet "Comparer", cette fonctionnalité a été extraite dans une nouvelle page pour offrir une expérience plus complète et intuitive. Voici les principales caractéristiques de cette page :
 - **Choix du type de données et filtrage temporel** : L'utilisateur peut sélectionner le type de données à afficher parmi :
 - Historique : permet de visualiser les moyennes des polluants sur des périodes définies (via un sélecteur de mois dynamique).
 - Prédiction : propose des valeurs prévisionnelles, également filtrables par mois.
 - Moyenne par habitants et Moyenne par superficie : pour des comparaisons standardisées sans dimension temporelle.
 - **Filtrage par polluant** : Un menu déroulant permet de restreindre l'analyse à un polluant spécifique (parmi NO, NO₂, O₃, PM10, PM2.5).
 - **Recherche et ajout de villes individuelles** : Un champ de recherche interactif proposant des suggestions permet à l'utilisateur d'ajouter facilement des villes spécifiques à la liste de comparaison.

FIGURE 9 – Capture d'écran de l'interface de filtrage de la page Compare.

- **Sélection de groupes de villes** : En plus de la sélection individuelle, il est possible de constituer la comparaison à partir de groupes de villes basés sur différents critères :
 - Département, région et densité – via une requête AJAX qui charge dynamiquement les valeurs distinctes depuis la base de données.
 - Superficie et population – avec des options prédéfinies (paliers) pour faciliter la sélection. Pour ces groupes, des boutons tels que « Ajouter tout le département » ou « Ajouter toute la région » simulent l'ajout de l'ensemble des villes correspondant au critère sélectionné.
- **Récupération et affichage des données** : Une fois les villes ou groupes sélectionnés, l'utilisateur déclenche la comparaison en cliquant sur le bouton « Comparer ». Une requête AJAX est alors envoyée à get_compare_data.php, qui :
 - Interroge la base de données en fonction des filtres appliqués (type de données, mois, polluant).
 - Agrège les valeurs (avec un traitement particulier pour exclure les valeurs négatives ou certains polluants comme « C₆H₆ » et « SO₂ »).
 - Retourne les résultats sous forme d'un graphique interactif (réalisé avec Chart.js) et d'un tableau récapitulatif.

Chaque entrée du tableau propose également une redirection vers la page détaillée d'une ville ou d'un groupe afin d'obtenir plus d'informations.

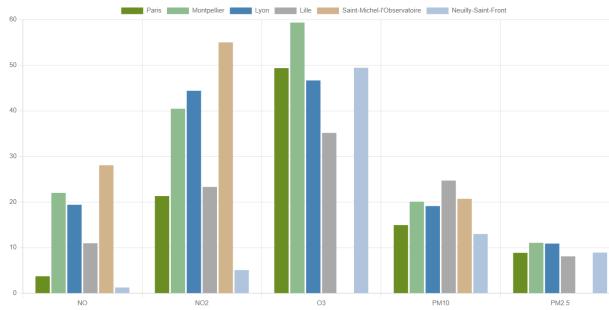


FIGURE 10 – Capture d’écran d’un exemple de graphique en barre de la page Compare.

4.3 Refonte du Système de Commentaires et de Favoris

Le système de commentaires et de favoris a été entièrement repensé pour améliorer l’interaction utilisateur :

- **Commentaires :**

- Intégrés directement dans la page détails (via `details.php` et `commentaires.js`) et plus dans (`commentaires.php`) qui a été supprimé.
- La table des commentaires utilise désormais l’identifiant de la ville (`id_ville`) au lieu de l’URL complète.
- Un système de vote (like/dislike) a été implémenté via une variable `vote` (1 pour like, -1 pour dislike).

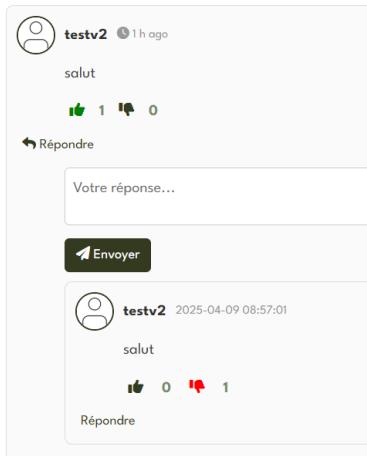


FIGURE 11 – Capture d’écran d’un extrait du nouvel espace commentaire.

- **Favoris :**

- Les utilisateurs connectés peuvent ajouter une ville à leurs favoris via une icône en forme d’étoile et plus en forme de cœur.
- Une animation de particules étoilées renforce visuellement l’ajout.
- Le code JavaScript initial (`messagesAjax.js`) a été divisé en modules (`favoris.js` et `historique.js`) pour une meilleure maintenabilité.

4.4 Refonte de la Page d'Accueil et du Menu de Navigation

Une nouvelle page d'accueil a été développée avec une approche minimaliste :



FIGURE 12 – Capture d'écran de la nouvelle page d'accueil.

Redirection claire vers les modules principaux : Carte interactive, Recherche et Classement.

4.5 Refonte du Menu de Navigation

- Le menu comporte désormais quatre onglets – *Accueil*, *Actualité*, *Données* et *Mon Compte* – facilitant l'accès à l'ensemble des fonctionnalités.

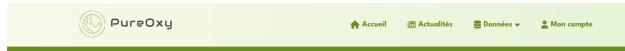


FIGURE 13 – Capture d'écran du nouveau header de PureOxy.

- L'onglet *Données* affiche quatre sous-onglets permettant la visualisation des données.
- Pour les utilisateurs connectés, l'onglet *Mon Compte* affiche le nom de l'utilisateur et offre deux sous-onglets : *Tableau de Bord* et *Déconnexion*.

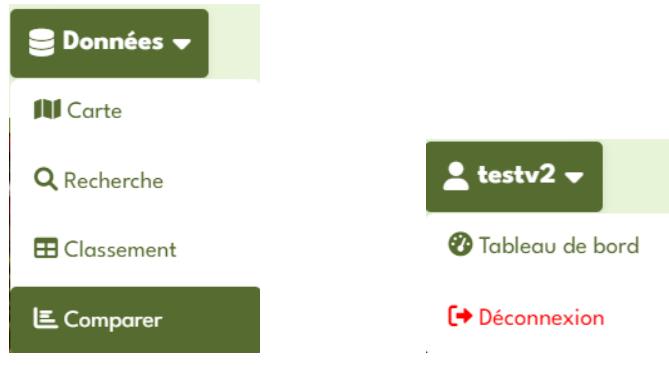


FIGURE 14 – Capture d'écran des menus déroulants du header.

Intégration et Optimisations Fonctionnelles

Ce chapitre présente l'intégration des fonctionnalités innovantes et les optimisations techniques mises en œuvre afin d'améliorer la performance globale du projet.

5.1 Optimisations de Performances

Face à l'augmentation considérable du volume de données et aux calculs complexes (par exemple, le calcul des moyennes pour 276 villes sur plusieurs années et pour divers polluants), plusieurs optimisations ont été réalisées :

- **Refonte de la base de données** : Création de la table `moy_pollution_villes` permettant de pré-calculer et stocker les moyennes.
- **Optimisation des requêtes SQL** : Utilisation d'index et optimisation des jointures pour réduire le temps de réponse.
- **Division et refonte du code JavaScript** : Modularisation en fichiers distincts (`favoris.js`, `historique.js`, `suggestions.js`) afin d'améliorer la maintenabilité et accélérer l'exécution.
- **Mise en cache des données statiques** : Réduction de la charge serveur lors des rechargements.

5.2 Intégration des Données Prédictives

Pour enrichir l'information proposée aux utilisateurs, le site intègre désormais à la fois les données historiques et les prévisions :

- Un onglet **Prédictions** permet de consulter les valeurs prédictes pour chaque ville.
- Les données prédictives proviennent de modèles de machine learning évalués via des indicateurs tels que le RMSE, le R² et le MAE.

5.3 Optimisation de la Carte Interactive

Pour offrir une visualisation précise et personnalisée, deux filtres complémentaires ont été ajoutés :

- **Filtre par Polluant** : Affiche uniquement les mesures pour le polluant sélectionné, avec une coloration des marqueurs en fonction du pourcentage en rapport au seuil du polluant.
- **Filtre par Mois** : Permet d'afficher les mesures pour un mois précis (de janvier 2023 à janvier 2025).

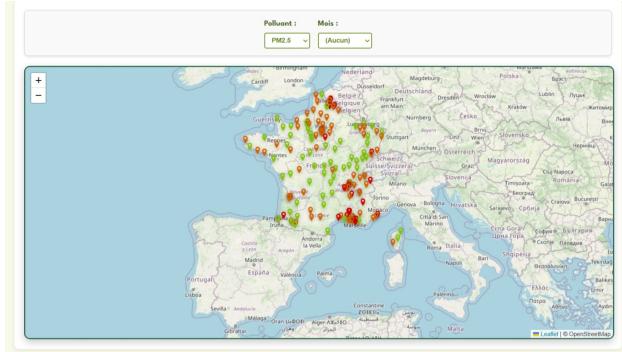


FIGURE 15 – Capture d'écran de nouveau système de filtrage de la carte interactive.

5.4 Intégration du Chatbot IA Mistral

Une innovation majeure du projet est l'intégration d'un chatbot reposant sur l'IA Mistral :

- **Implémentation Technique** : Le chatbot est intégré dans le header du site et s'affiche sous forme d'une bulle interactive en bas à droite de chaque page. Il a été développé en utilisant Rasa, un framework open source de dialogue, combiné avec le LLM Mistral, exécuté localement via Ollama pour garantir autonomie et rapidité.
- **Connexion au modèle Mistral** : Lorsqu'un utilisateur saisit une question, celle-ci est transmise à un serveur d'actions personnalisé. Ce dernier envoie la requête à Mistral, qui génère une réponse. Cette réponse est ensuite renvoyée dynamiquement à l'interface via fetch API (JavaScript)
- **Fonctionnalités et Contraintes** : Le chatbot a été configuré pour ne répondre qu'aux questions en lien avec l'environnement, la pollution de l'air, les polluants. Un système de filtrage thématique empêche volontairement les réponses aux questions hors sujet (ex : cuisine, sport...), afin d'assurer la cohérence et la pertinence de l'assistant.

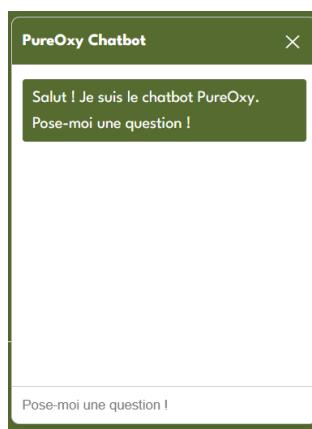


FIGURE 16 – Capture d'écran de l'interface du Chatbot.

Défis Techniques et Solutions Apportées

6.1 Optimisation des Requêtes et de l'Interface

Pour répondre aux exigences de performance, plusieurs améliorations complémentaires ont été apportées :

- **Réduction du coût de calcul :** Le recalculation initial des moyennes sur la carte interactive impliquait plus de $365 \text{ (jours)} \times 3 \text{ (années)} \times 4 \text{ (polluants)} \times 276 \text{ (villes)}$ calculs, soit 1 208 880 calculs. La table `moy_pollution_villes` permet aujourd’hui de réduire considérablement ce coût.
- **Optimisation des Modèles de Prédition :** Malgré des temps d’exécution élevés pour certains modèles (comme RandomForest et la régression linéaire), leur précision a justifié leur utilisation après ajustement notamment dans la mise à jour des variables de retard (`lag_1` et `lag_2`).

6.2 Principaux Défis Techniques

Plusieurs difficultés ont été rencontrées et résolues dans le cadre du développement :

1. **Nettoyage et Harmonisation des Données :** La coexistence de noms de villes complets et segmentés en secteurs a nécessité la création d’un fichier CSV récapitulatif et l’utilisation d’un algorithme de *fuzzy matching*.
2. **Mise à Jour des Variables de Retard :** Une première implémentation entraînait des répétitions dans les prévisions. La logique a été corrigée pour éviter ces duplications.
3. **Division des Scripts JavaScript :** La transformation d’un fichier monolithique (`messages-Ajax.js`) en modules dédiés (`favoris.js`, `historique.js`, etc.) a amélioré tant la lisibilité du code que les performances.
4. **Intégration du Chatbot :** l’une des principales difficultés techniques concernait l’impossibilité d’interroger directement la base de données MySQL hébergée sur InfinityFree via le chatbot. L’hébergement InfinityFree bloque les requêtes entrantes JSON et n’autorise pas les connexions distantes à sa base de données MySQL. Cela empêchait notre action Rasa de récupérer les données de pollution via des requêtes.
5. **Déploiement du site en ligne :** InfinityFree ne permettait pas l’importation directe de notre base de données via un fichier `.sql`, en raison de restrictions liées à la taille, nous avons donc contourné ce problème en utilisant un script d’importation PHP personnalisé, qui nous a servi à insérer les données dans la base hébergée ligne par ligne.

Exemple d'utilisation de PureOxy : Analyse comparative des données Historique et Normalisée (Habitants, km²)

Ce chapitre présente une étude approfondie réalisée grâce à PureOxy, en s'appuyant exclusivement sur trois types de données :

1. **Historique** (concentrations moyennes en µg/m³),
2. **Moyennes normalisées par habitant,**
3. **Moyennes normalisées par km².**

Nous comparons six villes emblématiques aux profils contrastés : **Paris, Montpellier, Lyon, Lille** (grands centres urbains) et deux localités beaucoup plus petites (**Saint-Michel-l'Observatoire, Neuilly-Saint-Front**) à caractère rural. L'objectif est de montrer dans quelle mesure la répartition géographique, la densité et la superficie influencent la qualité de l'air.

7.1 Sélection des Indicateurs et Méthodologie

- **Polluants** : NO (monoxyde d'azote), NO₂ (dioxyde d'azote), O₃ (ozone), PM10 et PM2.5 (particules fines).
- **Extraction des Données** :
 - *Historique* : Concentrations moyennes annuelles (µg/m³) pour chaque ville.
 - *Moyennes par habitant* : Poids relatif de chaque polluant rapporté à la population (exprimé en fraction de µg/m³ par habitant).
 - *Moyennes par km²* : Concentrations rapportées à la superficie de la commune (exprimé en fraction de µg/m³ par km²).

Pour faire cette analyse, nous nous sommes principalement appuyés sur la page Compare et sa génération de graphiques permettant de juxtaposer les différentes moyennes, afin de mettre en évidence les disparités selon les profils urbains/ruraux.

7.2 Résultats et Observations Clés

1. Données Historiques (Concentrations Brutes)

Polluant	1. Paris	2. Montpellier	3. Lyon	4. Lille	5. Saint-Michel-l'Observatoire	6. Neuilly-Saint-Front
NO	3.74	22.01	19.42	10.98	28.06	1.27
NO ₂	21.31	40.46	44.4	23.32	55.02	5.09
O ₃	49.36	59.36	46.67	35.15	-	49.43
PM10	14.95	20.09	19.12	24.71	20.71	13
PM2.5	8.87	11.07	10.9	8.1	-	8.93

FIGURE 17 – Capture d'écran du tableau affichant les moyennes historiques.

- **Forte disparité NO et NO₂** : Saint-Michel-l'Observatoire (*bourg rural*) présente des niveaux surprenants de NO (28.06 µg/m³) et NO₂ (55.02 µg/m³), supérieurs à ceux de Lyon ou Paris. Cette situation peut s'expliquer par une moindre couverture des infrastructures de contrôle ou

par la concentration d'émissions locales (agricoles, industrielles ou routières) dans un périmètre restreint.

- **O3 plus élevé à Montpellier (59.36 µg/m³)** : Les conditions météorologiques méditerranéennes, associées à l'ensoleillement, peuvent favoriser la formation d'ozone.
- **PM10 significatives à Lille (24.71 µg/m³)** : En tant que grand centre industriel et fortement urbanisé, Lille présente une valeur notable pour les particules PM10.

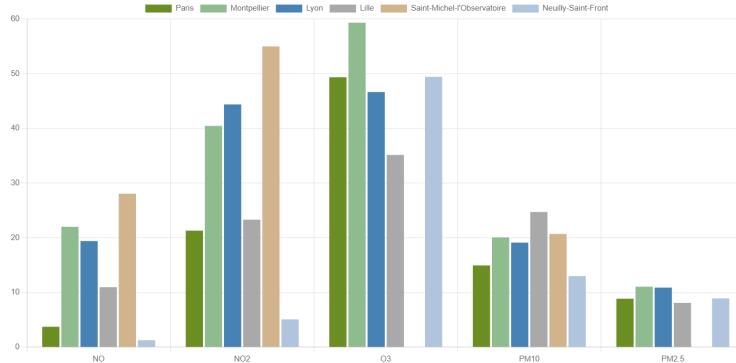


FIGURE 18 – Capture d'écran du graphique comparant les moyennes historiques.

2. Normalisation par Habitants

Polluant	1. Paris	2. Montpellier	3. Lyon	4. Lille	5. Saint-Michel-l'Observatoire	6. Neuilly-Saint-Front
NO	0.000002	0.000073	0.000037	0.000046	0.022890	0.000628
NO ₂	0.000010	0.000134	0.000085	0.000098	0.044874	0.002522
O ₃	0.000023	0.000196	0.000089	0.000149	-	0.024495
PM10	0.000007	0.000066	0.000037	0.000104	0.016896	0.006443
PM2.5	0.000004	0.000037	0.000021	0.000034	-	0.004427

FIGURE 19 – Capture d'écran du tableau affichant les moyennes par habitants.

- **Charge polluante élevée dans les bourgs ruraux** : Saint-Michel-l'Observatoire affiche des valeurs par habitant extrêmement supérieures à celles des grandes villes (par exemple, 0.022890 vs 0.000002 à Paris pour NO). Cela signifie qu'en considérant la population, la pollution par habitant est bien plus pénalisante pour les résidents des zones rurales où les habitants sont peu nombreux mais subissent l'essentiel des rejets locaux.
- **Paris, Lyon, Montpellier** : Le fait que leurs moyennes par habitant soient si faibles (de l'ordre de 0.0000xx) révèle l'effet de la forte population, qui dilue la charge polluante globale lorsqu'on la ramène par habitant.

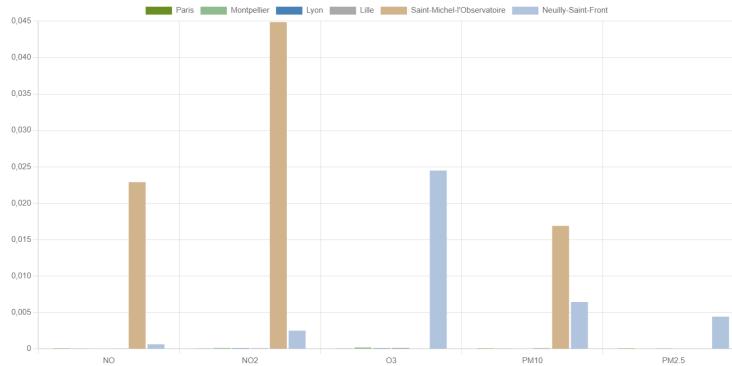


FIGURE 20 – Capture d’écran du graphique comparant les moyennes par habitants.

3. Normalisation par km²

Polluant	1. Paris	2. Montpellier	3. Lyon	4. Lille	5. Saint-Michel-l'Observatoire	6. Neuilly-Saint-Front
NO	0.035627	0.386160	0.404651	0.313837	1.002266	0.070359
NO2	0.202978	0.709847	0.925053	0.666156	1.964857	0.282713
O3	0.470142	1.041488	0.972209	1.004365	-	2.746210
PM10	0.142339	0.352505	0.398338	0.706109	0.739819	0.722372
PM2.5	0.084518	0.194222	0.227050	0.231432	-	0.496284

FIGURE 21 – Capture d’écran du tableau affichant les moyennes par superficie.

- **Concentration extrême dans certaines communes rurales :** Si l’on rapporte la pollution à la superficie, Saint-Michel-l’Observatoire peut atteindre 1.002266 pour NO et 1.964857 pour NO₂, bien au-dessus des grandes agglomérations. Cette donnée illustre qu’une faible superficie, combinée à des sources locales, peut générer une concentration importante sur la surface disponible.
- **Neuilly-Saint-Front avec un O₃ notable (2.746210) :** Alors que Paris ne présente que 0.470142 pour l’ozone/km², Neuilly-Saint-Front, malgré sa taille réduite, atteint des valeurs très élevées.

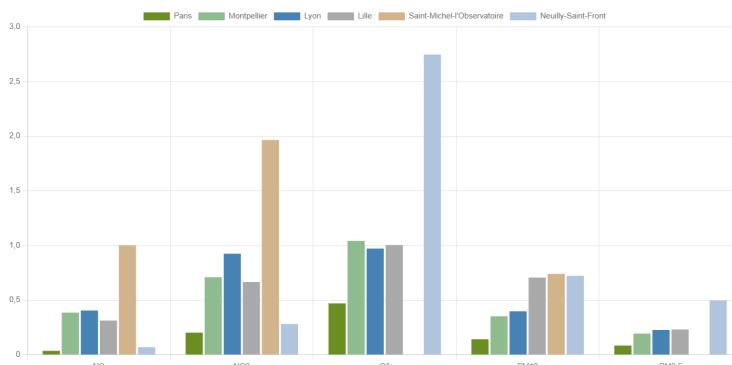


FIGURE 22 – Capture d’écran du graphique comparant les moyennes par superficie.

7.3 Interprétation Générale et Apport de PureOxy

Les comparaisons effectuées à partir de la page Compare de PureOxy permettent de tirer plusieurs enseignements :

1. **Rôle des Politiques Locales et de la Dispersion :** Les grandes villes (Paris, Lyon, Montpellier, Lille) bénéficient souvent de politiques environnementales structurées, d'infrastructures de contrôle et d'une dispersion liée au trafic ou au tissu urbain. À l'inverse, certaines communes rurales peuvent subir un cumul local de rejets qui, une fois normalisés, se révèlent beaucoup plus problématiques pour les habitants.
2. **Intérêt d'une Approche Multicritère :** Regarder uniquement la concentration brute (historique) peut être trompeur. Les normalisations par habitant et par km^2 offrent deux grilles de lecture complémentaires :
 - *Par habitant* : illustre la charge subie par la population locale.
 - *Par km^2* : révèle la concentration spatiale du polluant.PureOxy simplifie grandement ces analyses, grâce à l'extraction automatique de toutes ces valeurs et leur visualisation sur la même plateforme.
3. **Influence de la Géographie et de la Démographie :** Saint-Michel-l'Observatoire représente un cas typique d'un bourg rural ayant des indicateurs élevés en valeur brute comme en valeurs normalisées, suggérant des émissions concentrées et moins régulées. Neuilly-Saint-Front, avec un O_3/km^2 particulièrement élevé, confirme que l'effet de la densité et de la superficie peut créer des situations critiques, même en zone rurale.

7.4 Conclusion : Une Plateforme Complète d'Aide à la Décision

Cette comparaison de six villes, réalisées avec PureOxy, met en lumière la puissance des analyses historiques et normalisées pour comprendre l'influence de la répartition géographique, de la densité et de la superficie sur les niveaux de pollution. En intégrant ces trois types de données, PureOxy fournit aux décideurs et aux chercheurs une vision à la fois globale et granulaire, rendant possible l'identification rapide des points critiques et la mise en œuvre d'actions ciblées pour améliorer la qualité de l'air.

Détails Techniques et Architecture du Code

Pour garantir robustesse, sécurité et maintenabilité, le code de PureOxy repose sur des choix technologiques précis :

- **Back-end en PHP :**

- Une classe `Database` gère la connexion à MySQL via l'extension MySQLi et utilise systématiquement des requêtes préparées pour sécuriser les accès.
- L'architecture est modulaire : des dossiers dédiés regroupent la connexion (dossier `bd`), l'en-tête et le pied de page (dossier `includes`), les pages principales (dossier `pages`) et les fonctionnalités spécifiques (dossier `fonctionnalites`).

- **Front-end en HTML/CSS/JavaScript :**

- L'interface utilisateur s'appuie sur `Leaflet` pour la carte interactive et `Chart.js` pour les graphiques.
- Le code JavaScript est divisé en modules spécialisés (`carte.js`, `classement.js`, `chatbot.js`, `favoris.js`, `historique.js`, `suggestions.js`) pour faciliter la maintenance.
- La communication asynchrone avec le serveur s'appuie sur `fetch()` et AJAX pour l'actualisation dynamique des contenus.

- **Sécurité et Optimisation :**

- L'utilisation systématique de requêtes préparées renforce la sécurité du back-end.
- La division du JavaScript en modules et la mise en cache des données statiques assurent des performances accrues.

- **Intégration du Chatbot IA :**

- Intégré via une API REST (inspirée par Rasa), le chatbot est accessible depuis le header et répond uniquement aux questions relatives aux fonctionnalités du site.

Conclusion

La refonte du projet **PureOxy** représente une avancée majeure par rapport à la version initiale. Grâce à l'intégration de fonctionnalités avancées en data science, la refonte de la base de données et l'amélioration des interfaces, la plateforme offre aujourd'hui une expérience utilisateur à la fois riche et performante pour la surveillance et l'analyse des niveaux de pollution atmosphérique en France.

Les principaux apports de ce semestre sont :

- La transformation du concept initial de podium en une page **Classement** complète, avec des filtres et des options de tri, permettant d'identifier clairement la position de chaque ville dans le classement des polluants.
- La mise en place d'une page **Compare** dédiée, permettant une comparaison approfondie entre un nombre illimité de villes, enrichie par de nombreux filtres.
- L'intégration des données prédictives et historiques, offrant une vision complète des tendances et permettant d'analyser l'impact des caractéristiques géographiques sur les niveaux de pollution.
- Une refonte technique de la base de données qui réduit significativement les temps de chargement et améliore la réactivité du site.
- L'amélioration des interfaces utilisateurs avec des filtres dynamiques, des onglets interactifs et une navigation repensée.
- L'implémentation d'un chatbot IA basé sur Mistral, qui renforce l'interactivité en fournissant une aide contextuelle en temps réel.

En particulier, l'analyse des corrélations géographiques montre que, par exemple, des villes comme Paris, Lyon ou Montpellier, bien qu'appartenant à la même catégorie des grands centres urbains, présentent des niveaux de pollution très différents. La normalisation par km^2 et par habitant permet ainsi d'affiner l'interprétation des données et d'identifier les zones où la pollution est particulièrement concentrée. Cette approche aide les décideurs et les citoyens à mieux comprendre l'impact des facteurs géographiques sur la qualité de l'air.

Ainsi, PureOxy s'affirme non seulement comme un outil de visualisation, mais également comme une plateforme d'aide à la décision, capable de répondre à la problématique posée et de soutenir des politiques environnementales éclairées.