Université de Technologie de Troyes

ISI – NF21 : Data Understanding

Pollution Atmosphérique et Maladies Respiratoires

Rapport d'Exploration de Données

Méthodologie CRISP-DM

Sources de données :	EDGAR (Émissions) GBD 2023 (Maladies)
Période :	1980 – 2022
Couverture :	197 pays

29 novembre 2025

Table des matières

1 Introduction

Ce rapport présente l'exploration des données dans le cadre de l'étude des liens entre la pollution atmosphérique et les maladies respiratoires. L'objectif est de comprendre les caractéristiques des données avant toute modélisation, conformément à la méthodologie CRISP-DM.

1.1 Objectifs de l'étude

— Analyser les tendances temporelles des émissions de polluants atmosphériques
— Étudier la distribution géographique des maladies respiratoires
— Identifier les corrélations entre polluants et pathologies
— Préparer les données pour la phase de modélisation

1.2 Sources de données

EDGAR (Emissions Database for Global Atmospheric Research) : Base de données de la Commission Européenne contenant les émissions de 9 polluants atmosphériques par pays, année et secteur d'activité (1970-2022).

GBD 2023 (Global Burden of Disease) : Étude de l'Institute for Health Metrics and Evaluation (IHME) fournissant les taux de mortalité standardisés par âge pour 5 maladies respiratoires (1990-2023).

2 Description des Données

2.1 Vue d'ensemble

Table 1 – Caractéristiques des jeux de données

Caractéristique	EDGAR	GBD 2023
Nombre de pays	215	204
Période	1970–2022	1990–2023
Variables	9 polluants	5 maladies
Granularité	Pays, année, secteur	Pays, année, sexe
Unité	Kilotonnes (kt)	Taux pour 100 000 hab.

2.2 Polluants étudiés (EDGAR)

1. **PM2.5** : Particules fines ($< 2.5\mu m$)
2. **PM10** : Particules ($< 10\mu m$)
3. **NOx** : Oxydes d'azote
4. **SO2** : Dioxyde de soufre
5. **CO** : Monoxyde de carbone
6. **NH3** : Ammoniac
7. **NMVOC** : Composés organiques volatils non méthaniques
8. **BC** : Carbone noir
9. **OC** : Carbone organique

2.3 Maladies respiratoires (GBD)

1. Cancer de la trachée, des bronches et des poumons
2. Maladie pulmonaire obstructive chronique (MPOC)
3. Asthme
4. Pneumoconioses
5. Maladies pulmonaires interstitielles

3 Qualité des Données

3.1 Valeurs manquantes

L'analyse des valeurs manquantes révèle une excellente complétude des données. Le taux de valeurs manquantes global est inférieur à 0.01%, principalement concentré dans les correspondances de noms de pays entre les deux sources.

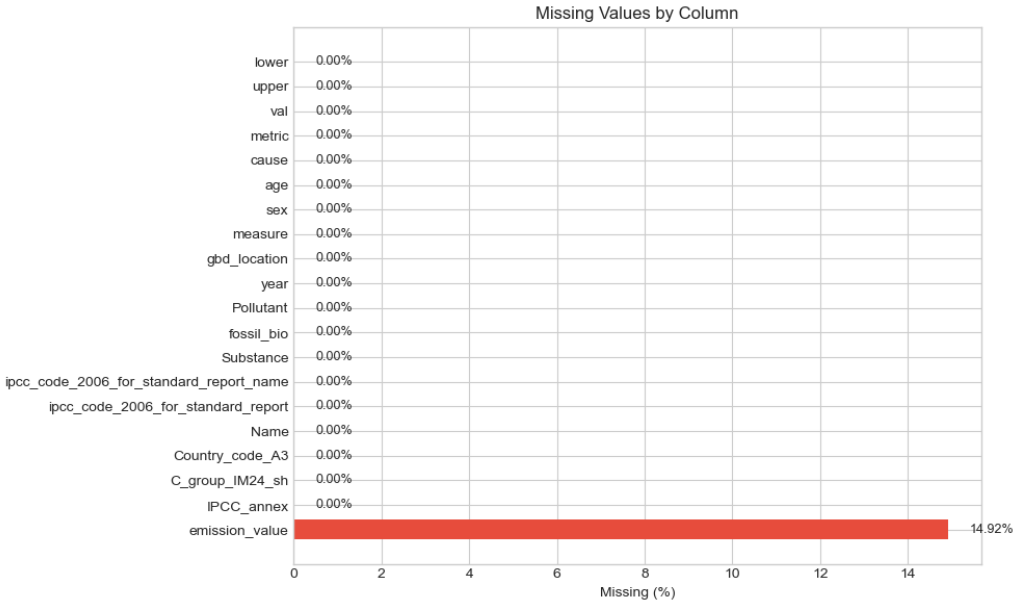

FIGURE 1 – Distribution des valeurs manquantes par variable

3.2 Jointure des données

La jointure des deux sources de données (EDGAR et GBD) a été réalisée sur :

— Le code ISO à 3 lettres des pays (197 pays communs)
— L'année (période commune : 1990–2022)

Le dataset final contient **38 millions d'observations** après jointure complète, et **63 445 observations** après agrégation par pays-année.

4 Analyse Exploratoire

4.1 Distribution des émissions par polluant

La figure 2 montre la distribution des émissions pour chaque polluant. On observe une forte asymétrie positive (skewness) pour tous les polluants, avec quelques pays émettant des quantités significativement supérieures à la moyenne mondiale.

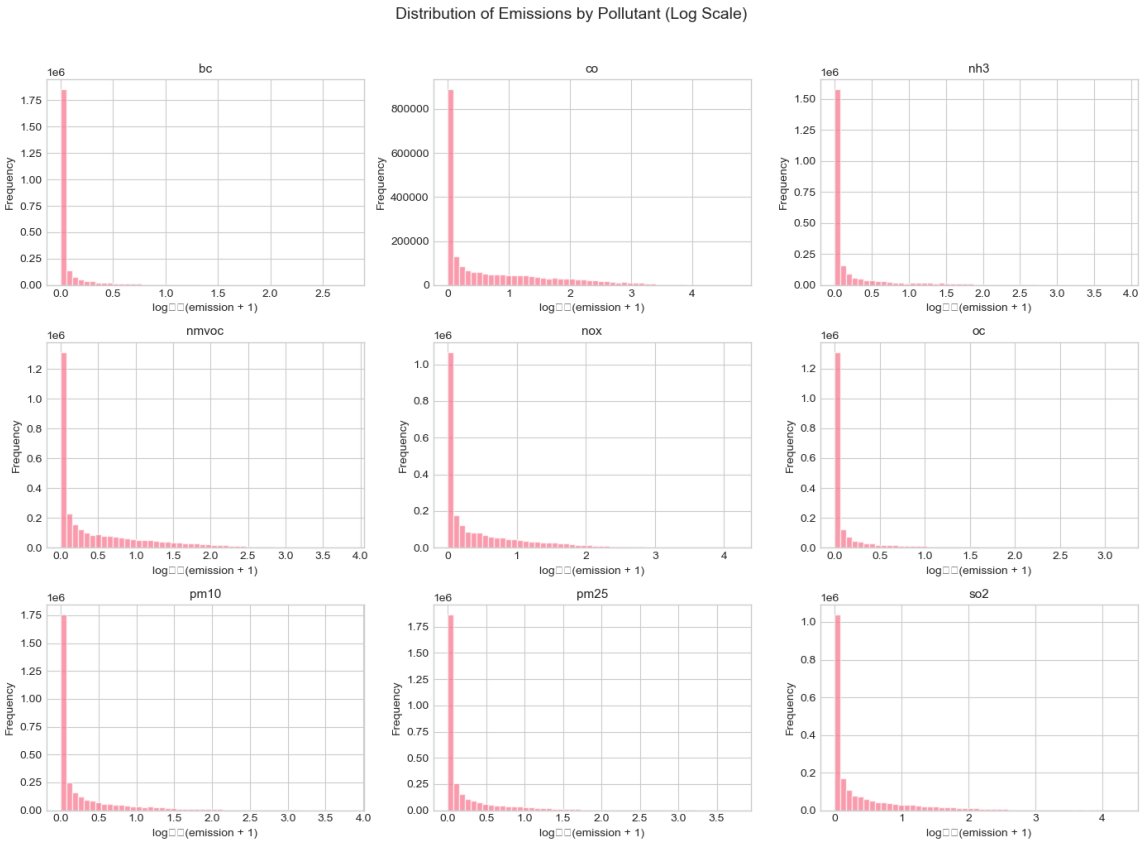

FIGURE 2 – Distribution des émissions par polluant (échelle logarithmique)

4.2 Comparaison hommes/femmes

L'analyse par sexe révèle des différences significatives dans les taux de mortalité. Les hommes présentent des taux plus élevés pour la majorité des maladies respiratoires, notamment pour le cancer du poumon et les MPOC.

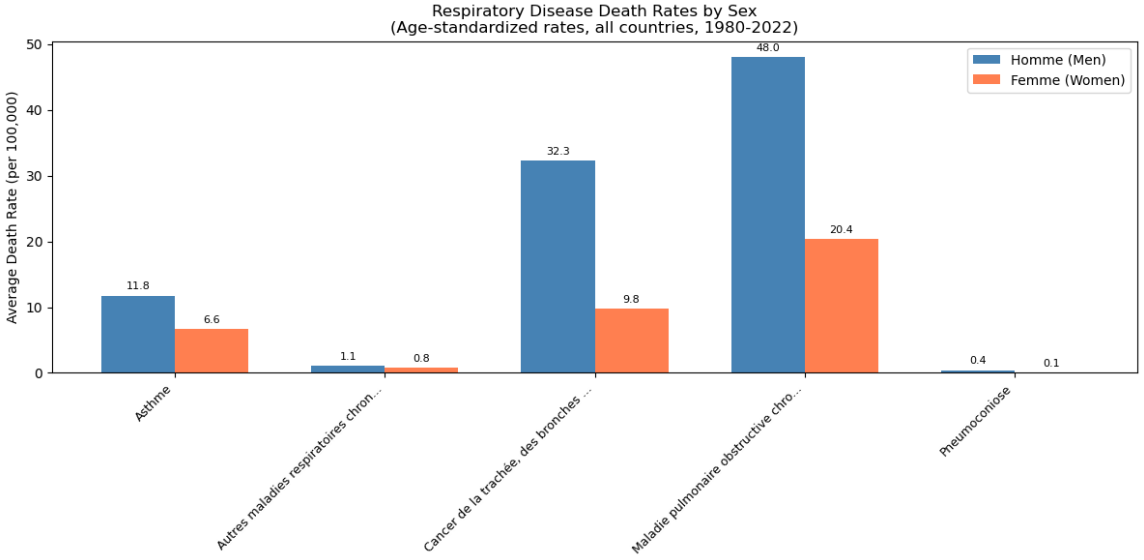

FIGURE 3 – Comparaison des taux de mortalité par sexe

5 Évolutions Temporelles

5.1 Tendances des émissions (1980–2022)

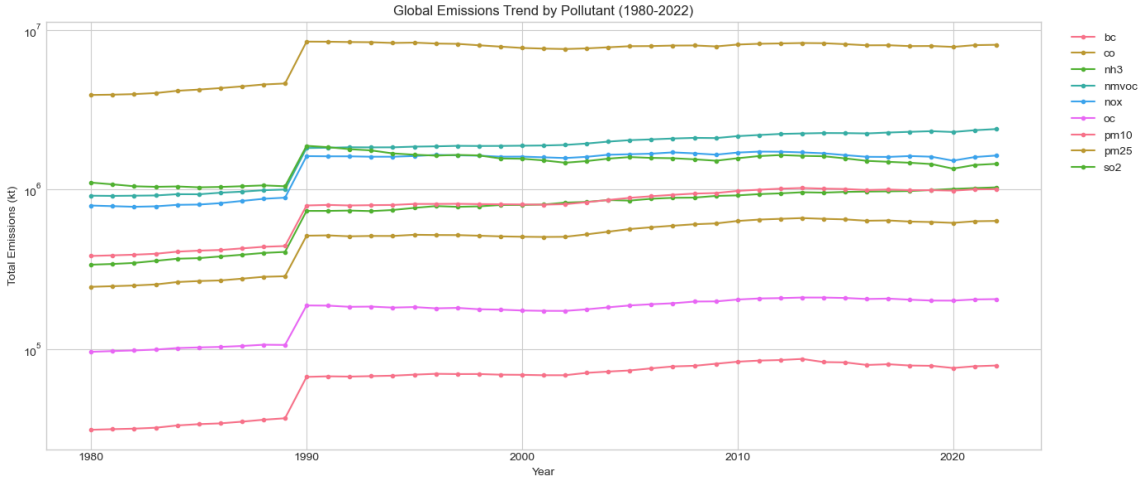

FIGURE 4 – Évolution temporelle des émissions mondiales par polluant

Observations clés :

— Pics d'émissions en 1990 pour tous les polluants
— Tous les polluants semblent très corrélés
— L'Oxide de Carbone (CO) est le polluant le plus présent

5.2 Tendances des décès (1990–2022)

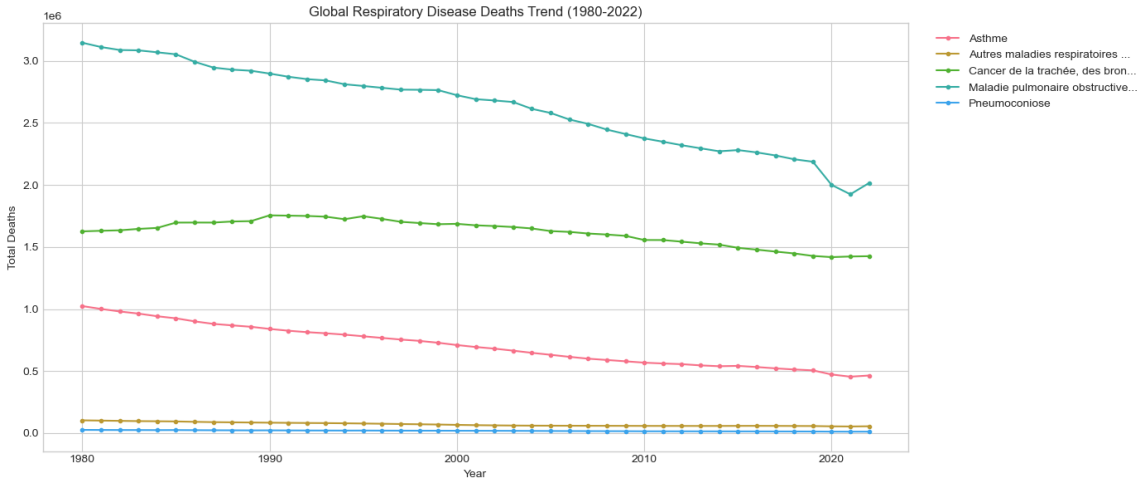

FIGURE 5 – Évolution temporelle des taux de mortalité par maladie

Observations clés :

— Diminution globale des taux de mortalité pour les MPOC
— Stabilité relative du cancer du poumon
— Baisse significative de l'asthme mortel

6 Analyse Géographique

6.1 Pays les plus émetteurs

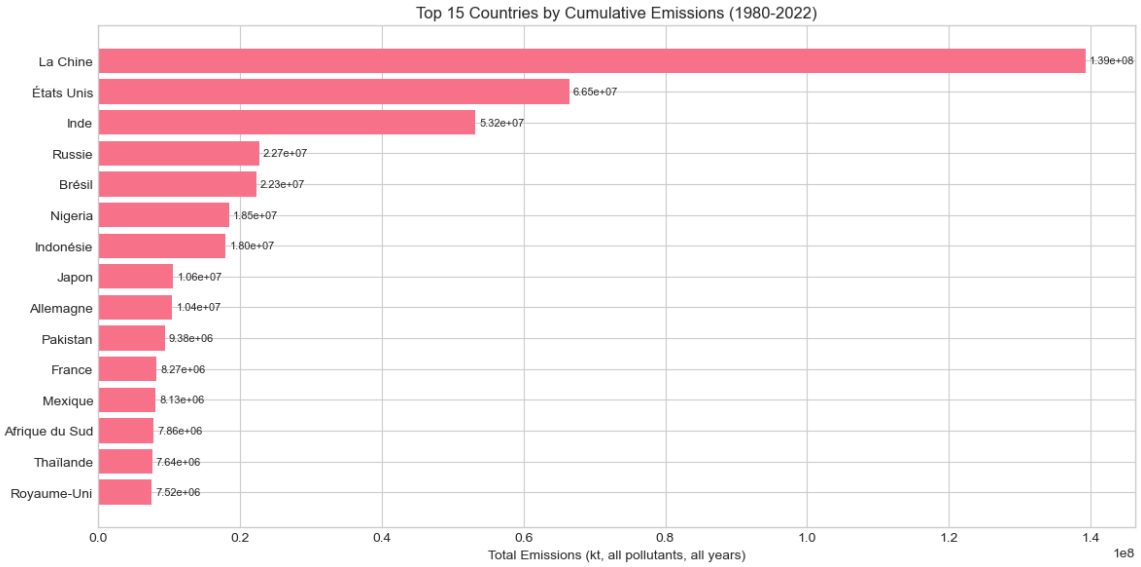

FIGURE 6 – Top 15 des pays émetteurs (toutes émissions confondues)

Les plus grands émetteurs sont la Chine, les États-Unis, l'Inde et la Russie, reflétant leur activité industrielle et leur population.

6.2 Pays avec les taux de mortalité les plus élevés

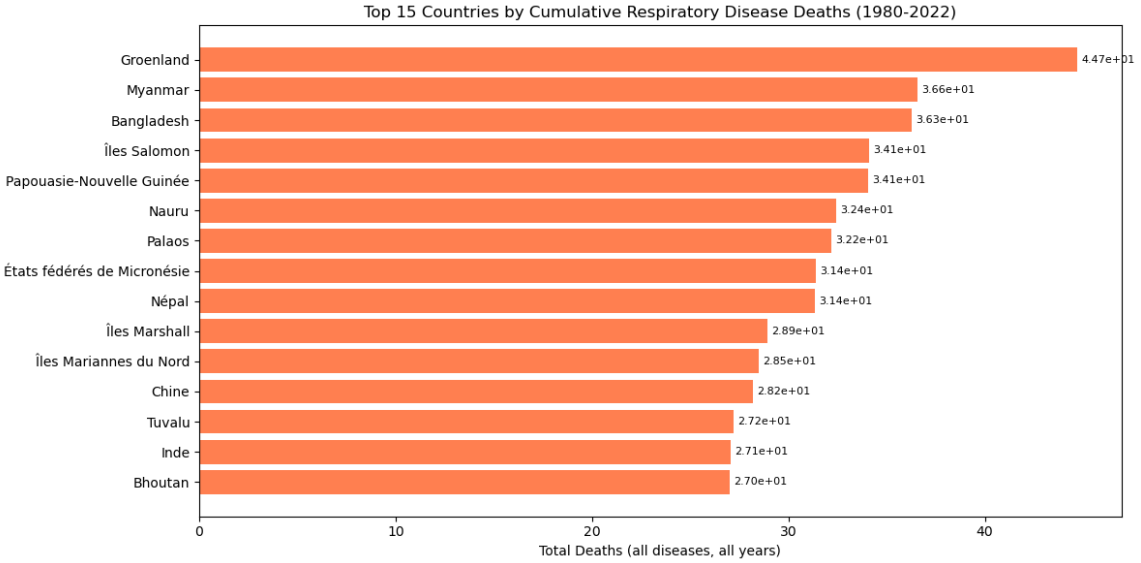

FIGURE 7 – Top 15 des pays par taux de mortalité respiratoire

7 Analyse Sectorielle

7.1 Secteurs d'activité

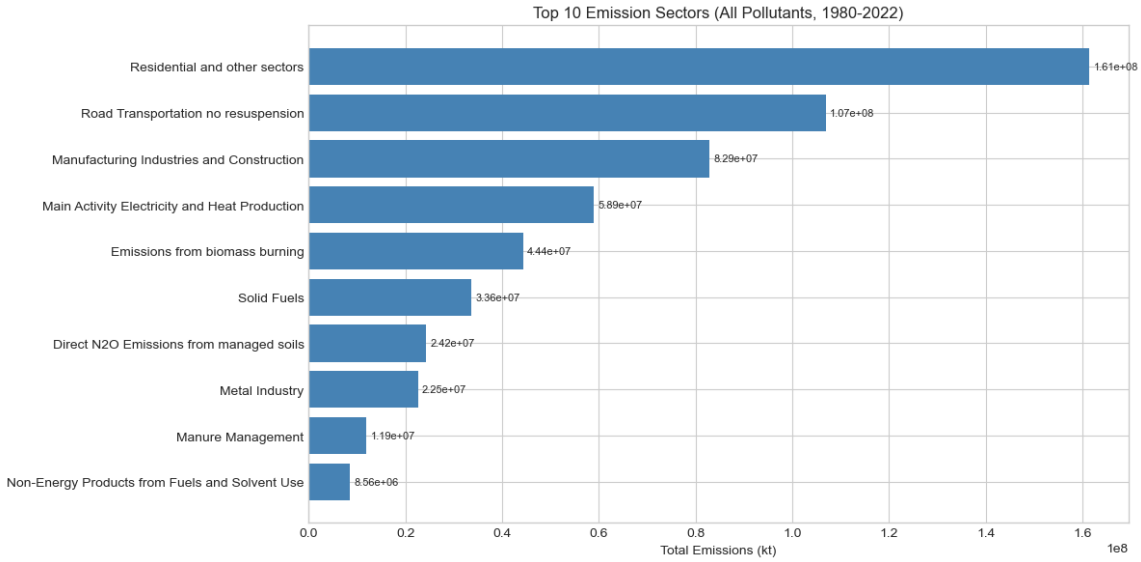

FIGURE 8 – Top 10 des secteurs d'activité par émissions totales

Les secteurs dominants sont :

— Transport routier
— Industrie manufacturière
— Production d'énergie
— Agriculture

7.2 Relation secteur-polluant

FIGURE 9 – Heatmap des émissions par secteur et polluant

Cette heatmap révèle les associations secteur-polluant : le transport routier est associé aux NOx et CO, tandis que l'agriculture domine les émissions de NH3.

8 Analyse des Corrélations

8.1 Corrélations polluants-maladies

La matrice de corrélation entre polluants et maladies respiratoires constitue le cœur de cette étude exploratoire.

FIGURE 10 – Matrice de corrélation : polluants vs maladies respiratoires

Corrélations notables :

— PM2.5 et PM10 montrent des corrélations positives modérées avec toutes les maladies
— NH3 et OC présentent les corrélations les plus fortes avec le cancer du poumon
— les MPOC est corrélée à la plupart des polluants

8.2 Matrice de corrélation complète

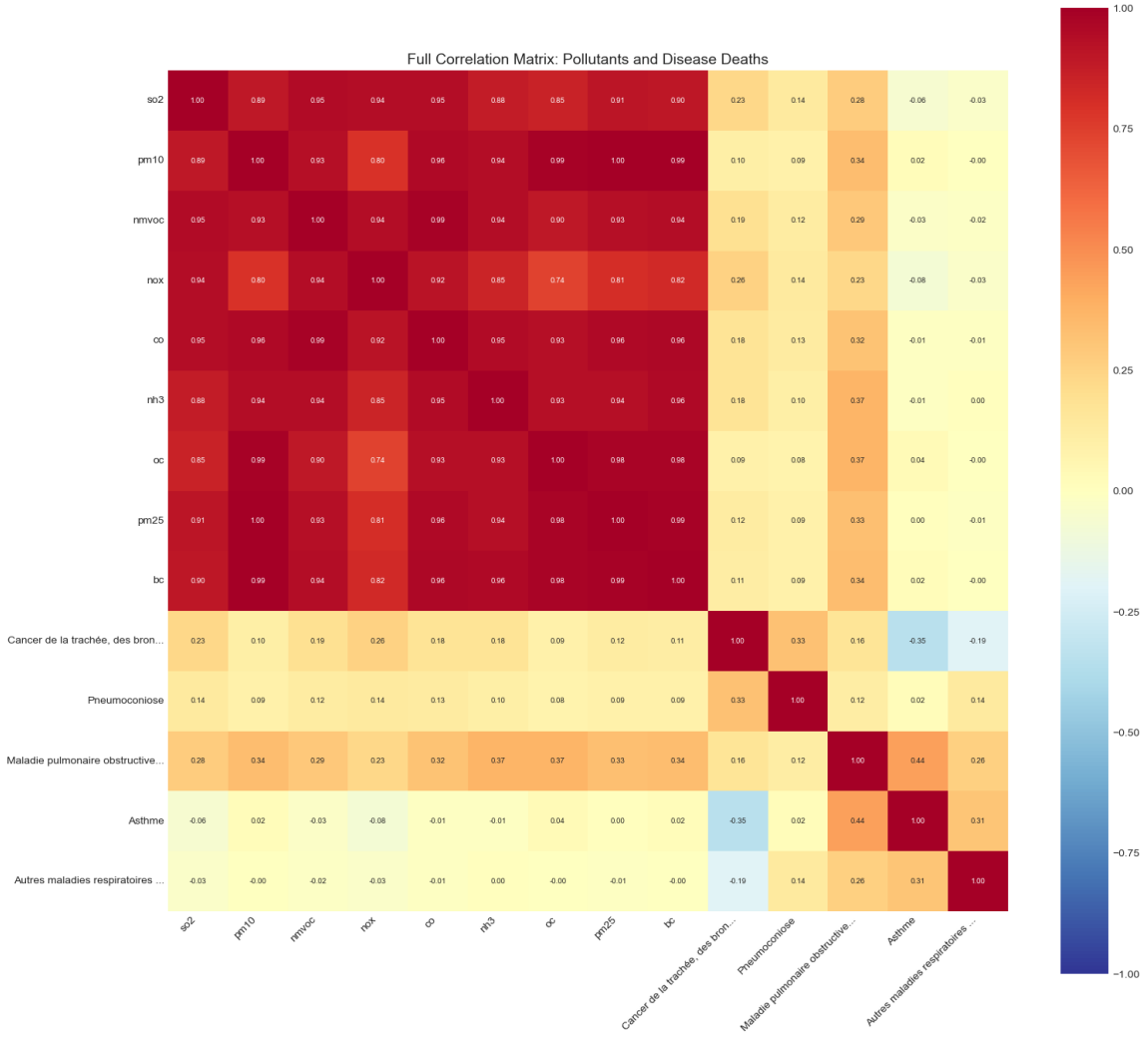

FIGURE 11 – Matrice de corrélation complète (polluants et maladies)

Observations :

— Forte multicolinéarité entre polluants ($r > 0.8$ pour la plupart)
— Les maladies sont également corrélées entre elles
— Ces corrélations suggèrent des facteurs communs (développement économique, urbanisation)
— les MPOC semblent plus corrélées aux polluants que les autres maladies (notamment NMVOC, OC, PM10 et PM25)

8.3 Relations détaillées (scatter plots)

FIGURE 12 – Scatter plots : polluants clés vs maladies (échelle log)

9 Synthèse et Conclusions

9.1 Qualité des données

Table 2 – Résumé de la qualité des données

Critère	Évaluation
Complétude	Excellente ($> 99.99\%$)
Couverture géographique	197 pays
Couverture temporelle	43 ans (1980–2022)
Cohérence des unités	Vérifiée
Outliers	$< 2\%$ (à traiter)

9.2 Principales conclusions

1. **Corrélations significatives** : Des corrélations positives modérées existent entre les polluants atmosphériques et les maladies respiratoires, particulièrement pour les particules fines (PM2.5, PM10), les NVMOC, l'Oxide de Carbone et le cancer du poumon ainsi que les maladies pulmonaires obstructives chroniques (MPOC) .
2. **Multicolinéarité** : Les polluants sont fortement corrélés entre eux, ce qui nécessitera une attention particulière lors de la modélisation (régularisation, réduction de dimension).
3. **Disparités géographiques** : Les pays en développement présentent des taux d'émission croissants tandis que les pays développés montrent des tendances à la baisse.
4. **Différences par sexe** : Les hommes sont plus touchés par les maladies respiratoires, avec des taux de mortalité 2 à 3 fois supérieurs pour le cancer du poumon.
5. **Évolution temporelle** : Malgré l'augmentation des émissions globales, les taux de mortalité standardisés tendent à diminuer, suggérant l'impact positif des avancées médicales.

9.3 Recommandations pour la modélisation

— Appliquer une transformation logarithmique aux émissions
— Considérer une analyse en composantes principales (ACP) pour réduire la multicolinéarité
— Inclure des variables de contrôle (PIB, urbanisation, accès aux soins)
— Utiliser des modèles de panel pour exploiter la dimension temporelle
— Tester des modèles avec décalage temporel (lag) entre exposition et maladie

Références

— EDGAR v8.0 : https://edgar.jrc.ec.europa.eu/
— GBD 2023 : https://ghdx.healthdata.org/gbd-2023
— Méthodologie CRISP-DM : https://moodle.utt.fr/pluginfile.php/13371/mod_resource/content/1/CRISP-DM.pdf