Université de Technologie de Troyes

ISI – NF21 : Business & Data Understanding

Respiratory Risk Analytics

Pollution Atmosphérique et Maladies Respiratoires

Rapport d'Analyse de Données

Méthodologie CRISP-DM

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

3 décembre 2025

Table des matières

1 Introduction

La pollution atmosphérique constitue aujourd'hui l'un des principaux déterminants environnementaux de santé. Ses effets sur les pathologies respiratoires sont largement documentés, mais leur ampleur et leur dynamique varient selon les territoires, les périodes et les types de polluants.

Le projet **Respiratory Risk Analytics** s'inscrit dans cette perspective. Il vise à mieux comprendre les interactions entre les émissions atmosphériques et l'évolution des maladies respiratoires à l'échelle mondiale. En s'appuyant sur des bases de données reconnues et actualisées, le projet cherche à fournir des éléments factuels pour aider les décideurs à identifier les zones les plus sensibles et à prioriser les actions de réduction de la pollution.

Ce rapport couvre les deux premières phases de la méthodologie CRISP-DM :

— **Business Understanding** : compréhension du contexte métier, des enjeux et des objectifs du projet
— **Data Understanding** : exploration et analyse des données disponibles

1.1 Objectifs de l'étude

— Comprendre et quantifier la relation entre la pollution atmosphérique et l'incidence des maladies respiratoires
— Identifier les zones prioritaires d'intervention
— 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

Première partie
Business Understanding

2 Contexte Général

Ce projet s'inscrit dans le champ de la santé publique et de l'environnement, deux domaines où les interactions entre qualité de l'air et santé humaine constituent des préoccupations majeures. L'augmentation de la fréquence des épisodes de pollution et la progression des maladies respiratoires en font un enjeu prioritaire pour les autorités sanitaires nationales et internationales. Les politiques publiques s'appuient désormais fortement sur les données pour orienter les stratégies territoriales de prévention et d'intervention.

Plusieurs acteurs sont impliqués dans cette dynamique : les collectivités territoriales, les agences de santé publique, les ministères en charge de la santé et de l'environnement, ainsi que les organisations européennes comme la Commission Européenne. Les citoyens, directement exposés aux risques liés à la pollution atmosphérique, représentent la première population bénéficiaire du projet. Dans ce contexte institutionnel riche, le recours à une analyse *data-driven* constitue un outil d'aide à la décision essentiel.

3 Question Métier et Objectifs

L'objectif global du projet Respiratory Risk Analytics consiste à comprendre et quantifier la relation entre la pollution atmosphérique et l'incidence des maladies respiratoires dans le monde. Cette question métier répond à un besoin décisionnel réel : identifier les zones prioritaires d'intervention et mesurer l'impact potentiel ou réel des politiques publiques de réduction de la pollution.

La question centrale formulée est la suivante : « **Quelle est la relation entre la pollution atmosphérique et l'incidence des maladies respiratoires à l'échelle mondiale ?** ».

Plusieurs sous-questions guident cette problématique :

- **Diagnostiquer** : identifier les zones où les émissions et les maladies respiratoires sont les plus élevées
- **Expliquer** : analyser les liens temporels entre variations de pollution et variations de santé
- **Prédire** : estimer l'évolution potentielle des maladies à partir des niveaux de pollution
- **Recommander** : proposer des actions prioritaires

4 Enjeux du Projet

Les enjeux sont multiples et concernent plusieurs dimensions :

4.1 Enjeux environnementaux

La réduction des émissions de polluants constitue un impératif lié à la fois aux réglementations européennes et à la protection de la biodiversité.

4.2 Enjeux sanitaires

La pollution atmosphérique est reconnue comme un facteur aggravant majeur pour des pathologies telles que l'asthme, la bronchopneumopathie chronique obstructive (BPCO) et les infections respiratoires aiguës.

4.3 Enjeux économiques

Une diminution de l'incidence des maladies respiratoires permettrait une réduction des dépenses hospitalières, des coûts de prise en charge et des pertes de productivité liées aux arrêts de travail.

4.4 Enjeux sociaux

Les enjeux sociaux touchent directement la qualité de vie des populations, notamment les plus vulnérables.

Ce projet présente ainsi une forte valeur ajoutée, car il vise à éclairer la décision publique et à cibler les zones où les actions de réduction de la pollution atmosphérique auraient l'effet le plus significatif.

5 Objectifs Métier

Le projet se fixe plusieurs objectifs métier :

1. **Identification des zones prioritaires** : grâce à une démarche de data mining, identifier les zones géographiques où les marges de réduction de la pollution sont les plus fortes et où les interventions publiques peuvent avoir un impact significatif.
2. **Réduction de l'incidence** : proposer des solutions permettant de réduire l'incidence des maladies respiratoires de 5% dans les zones les plus sensibles, et de diminuer de 10% les émissions des polluants les plus dangereux pour la santé.
3. **Identification des polluants critiques** : identifier les cinq polluants ayant l'impact le plus important sur la santé respiratoire à l'échelle mondiale.

6 Jeux de Données

Pour répondre à ces objectifs, deux principaux jeux de données externes ont été identifiés.

6.1 Données EDGAR

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

Les données se présentent sous forme de fichiers Excel contenant environ 6 000 lignes par polluant. Les variables incluent :

— L'année (numérique)
— Le pays (catégoriel)
— Le secteur d'émission (catégoriel)
— Le type de polluant (catégoriel)
— La quantité d'émissions en gigagrammes (numérique)

Ces données permettent de suivre l'évolution des polluants majeurs (PM2.5, PM10, NOx, SO2, etc.) dans le temps.

6.2 Données IHME : Global Burden of Disease

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 les maladies respiratoires.

Les fichiers CSV contiennent environ 30 000 lignes et décrivent les taux normalisés d'incidence et de mortalité pour 100 000 habitants. Les variables incluent :

— L'année (numérique)
— Le pays (catégoriel)
— Le sexe (catégoriel)
— Les taux d'incidence et de décès (numériques)

Ces deux jeux de données sont complémentaires et permettent une analyse spatio-temporelle robuste. La littérature scientifique, notamment les travaux de Santé Publique France, confirme l'existence d'un lien entre pollution atmosphérique et santé respiratoire, justifiant pleinement leur mise en relation.

7 Périmètre Analytique

Le périmètre analytique du projet est défini comme suit :

Table 1 – Périmètre du projet

Dimension	Périmètre
Couverture géographique	Ensemble des pays du monde (197 pays)
Période	1980–2022
Polluants	PM2.5, PM10, NOx, SO2, CO, NH3, NMVOC, BC, OC
Maladies	Asthme, BPCO, cancer du poumon, pneumoconioses, maladies pulmonaires interstitielles

Hors périmètre : Les analyses à l'échelle intra-urbaine, ou celles portant sur les déterminants sociaux individuels, ne sont pas incluses en raison du manque de données disponibles.

8 Parties Prenantes

8.1 Commanditaire principal

L'**Agence Régionale de Santé (ARS)** constitue une partie prenante centrale dans ce projet, à la fois comme utilisatrice directe des analyses produites et comme actrice opérationnelle des décisions qui pourront en découler. En tant qu'autorité sanitaire de proximité, l'ARS est chargée de mettre en œuvre au niveau régional les politiques nationales liées à la prévention, à la surveillance épidémiologique et à la protection de la santé des populations.

Dans le contexte de la pollution atmosphérique, l'ARS joue un rôle d'interface entre les données scientifiques, les acteurs locaux et les décisions publiques. Les résultats du projet pourront être mobilisés pour ajuster les plans régionaux santé-environnement, cibler les territoires les plus vulnérables et renforcer les dispositifs de prévention.

8.2 Équipe projet

L'équipe projet est composée d'analystes et de spécialistes des données chargés de transformer les données brutes (EDGAR, IHME) en indicateurs intelligibles.

8.3 Acteurs secondaires

Les résultats du projet sont également destinés à des acteurs secondaires :

- **Santé Publique France** : intégration des indicateurs dans la surveillance sanitaire nationale
- **Observatoires régionaux de la qualité de l'air (ATMO)** : relais de l'information à l'échelle locale
- **Collectivités territoriales** (métropoles, régions) : déploiement des actions concrètes de réduction des émissions
- **Organismes de recherche** spécialisés en climat, pollution ou santé publique

9 Risques et Hypothèses

9.1 Risques identifiés

Plusieurs risques concernent la qualité et la comparabilité des données :

- Les méthodes de mesure diffèrent selon les pays
- Certaines séries peuvent être incomplètes
- Des facteurs socio-économiques non observés peuvent introduire des biais
- La granularité géographique est limitée à l'échelle des pays

9.2 Hypothèses

Les hypothèses principales du projet reposent sur :

- La fiabilité des données EDGAR et IHME
- La comparabilité des observations dans le temps et entre pays
- L'existence d'un lien statistique entre pollution atmosphérique et maladies respiratoires

Deuxième partie
Data Understanding

10 Indicateurs et Chaînes de Valorisation

Dans le cadre de la méthodologie CRISP-DM, cette section présente les indicateurs clés et les chaînes de valorisation identifiées pour l'exploration des données sur la pollution atmosphérique et les maladies respiratoires.

10.1 Indicateurs clés

TABLE 2 – Indicateurs identifiés pour l'étude

Indicateur	Description	Unité
Émissions totales par polluant	Somme des émissions annuelles par type de polluant	Kilotonnes (kt)
Taux de mortalité standardisé	Décès pour 100 000 habitants ajusté par âge	Taux / 100 000 hab.
Corrélation polluant-maladie	Coefficient de Pearson entre émissions et mortalité	[-1, 1]
Part sectorielle	Pourcentage des émissions par secteur d'activité	%
Tendance temporelle	Variation annuelle moyenne des émissions/mortalité	% / an

10.2 Chaînes de valorisation

Trois chaînes de valorisation principales ont été identifiées pour cette phase d'exploration.

10.2.1 Chaîne 1 – Analyse des corrélations polluants-maladies

Objectif : Identifier les liens statistiques entre les émissions de polluants atmosphériques et les taux de mortalité par maladie respiratoire.

TABLE 3 – Chaîne 1 – Corrélations polluants-maladies

Tâche générique (TG)	Tâches spécifiques (TS)	Outils
TG11 – Collecte et Intégration	Télécharger les données EDGAR (émissions) et GBD (mortalité), extraire les fichiers Excel/CSV	Python (requests), Excel, CSV
TG12 – Préparation et Mise en qualité	Harmoniser les noms de pays (ISO 3166), traiter les valeurs manquantes, uniformiser les unités	Polars, Pandas
TG13 – Traitement et Analyse	Calculer les matrices de corrélation, identifier les associations significatives	Python (scipy, numpy), Omniscope
TG14 – Restitution et Visualisation	Générer les heatmaps de corrélation, scatter plots polluant vs maladie	Matplotlib, Seaborn, Omniscope

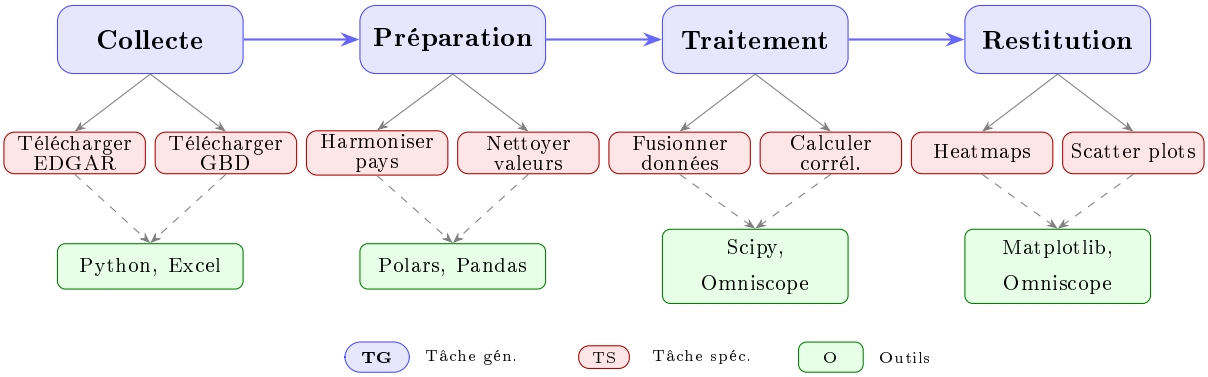

FIGURE 1 – Schéma de la chaîne 1 – Corrélations polluants-maladies

10.2.2 Chaîne 2 – Analyse géographique et temporelle

Objectif : Comprendre la distribution spatiale et l'évolution temporelle des émissions et de la mortalité respiratoire.

TABLE 4 – Chaîne 2 – Analyse géographique et temporelle

Tâche générique (TG)	**Tâches spécifiques (TS)**	**Outils**
TG21 – Collecte et Intégration	Joindre les données EDGAR et GBD sur code ISO pays et année	Polars (join)
TG22 – Préparation et Mise en qualité	Agréger par pays-année, filtrer la période commune (1980-2022), 197 pays	Polars (group_by, filter)
TG23 – Traitement et Analyse	Calculer les classements (top émetteurs, top mortalité), tendances temporelles	Python (numpy, pandas), Omniscope
TG24 – Restitution et Visualisation	Graphiques bar charts par pays, courbes d'évolution temporelle	Matplotlib, Seaborn, Omniscope

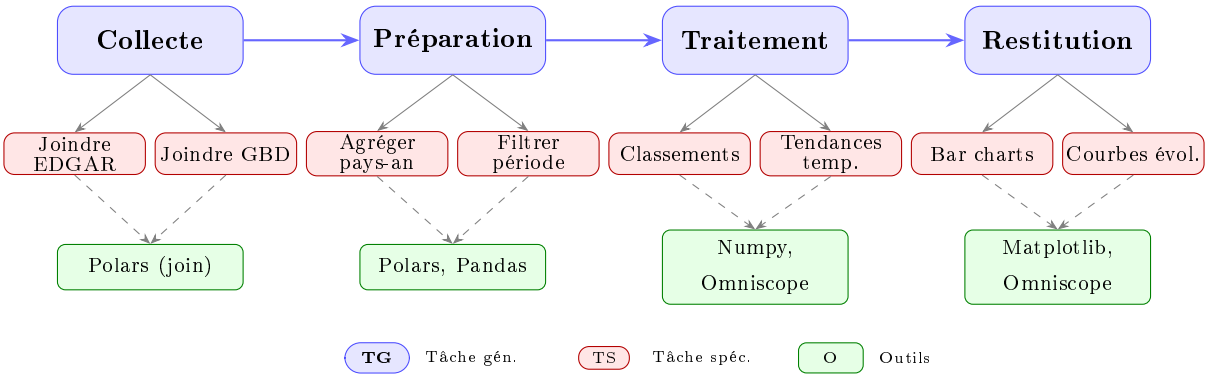

FIGURE 2 – Schéma de la chaîne 2 – Analyse géographique et temporelle

10.2.3 Chaîne 3 – Analyse sectorielle des émissions

Objectif : Identifier les secteurs d'activité les plus polluants pour orienter les analyses et recommandations.

TABLE 5 – Chaîne 3 – Analyse sectorielle

Tâche générique (TG)	**Tâches spécifiques (TS)**	**Outils**
TG31 – Collecte et Intégration	Extraire les données EDGAR avec granularité sectorielle (IPCC categories)	Python, Excel
TG32 – Préparation et Mise en qualité	Regrouper les secteurs par catégorie IPCC, normaliser les noms	Polars, Pandas, Omniscope
TG33 – Traitement et Analyse	Calculer les émissions totales par secteur et polluant, identifier les associations secteur-polluant	Python (numpy), Omniscope
TG34 – Restitution et Visualisation	Heatmap secteur-polluant, bar charts des top secteurs	Matplotlib, Seaborn, Omniscope

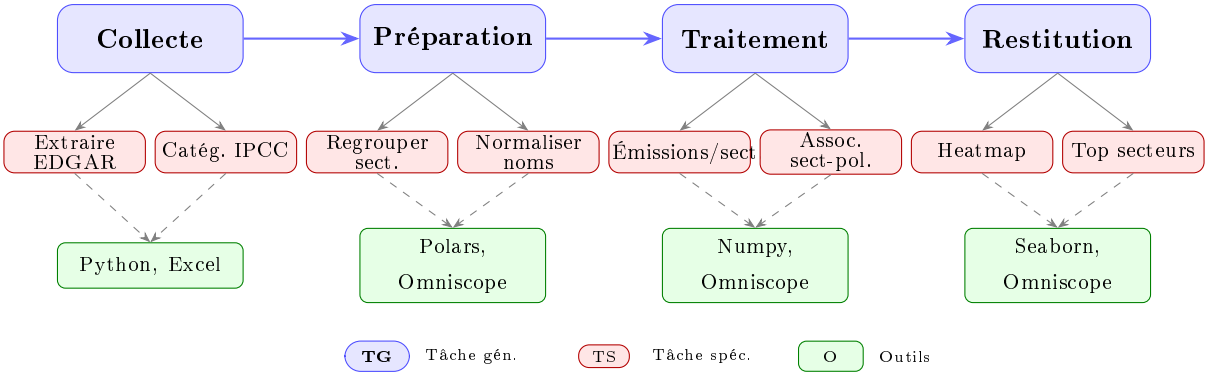

FIGURE 3 – Schéma de la chaîne 3 – Analyse sectorielle des émissions

11 Description des Données

11.1 Vue d'ensemble

Table 6 – Caractéristiques des jeux de données

Caractéristique	EDGAR	GBD 2023
Nombre de pays	215	204
Période	1970–2022	1980–2023
Variables	9 polluants	5 maladies
Granularité	Pays, année, secteur	Pays, année, sexe
Unité	Kilotonnes (kt)	Taux pour 100 000 hab.

11.2 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

11.3 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

12 Qualité des Données

12.1 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.

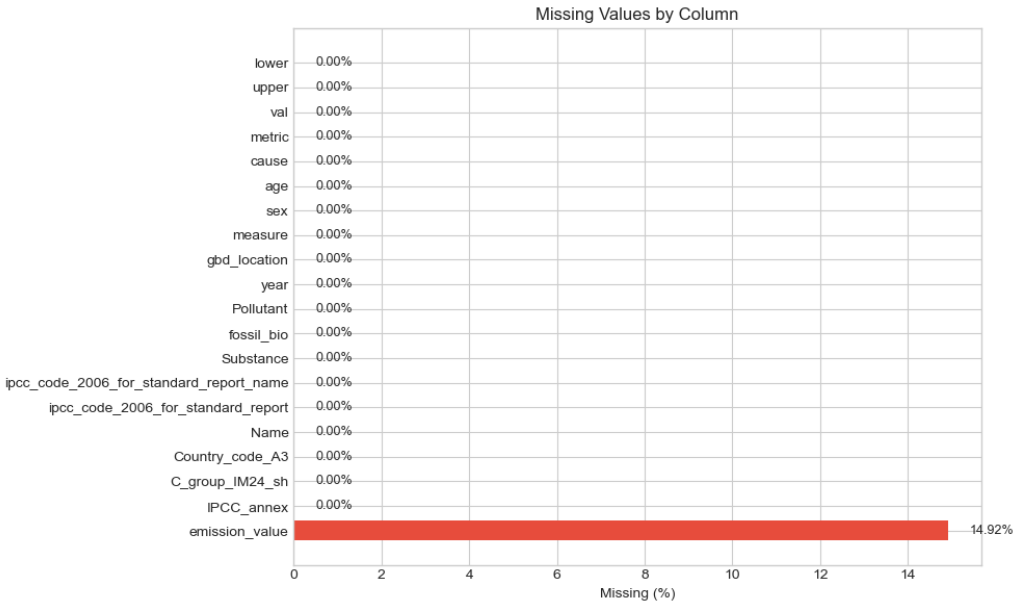

FIGURE 4 – Distribution des valeurs manquantes par variable

12.2 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 : 1980–2022)

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

13 Analyse Exploratoire

13.1 Distribution des émissions par polluant

La figure 5 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.

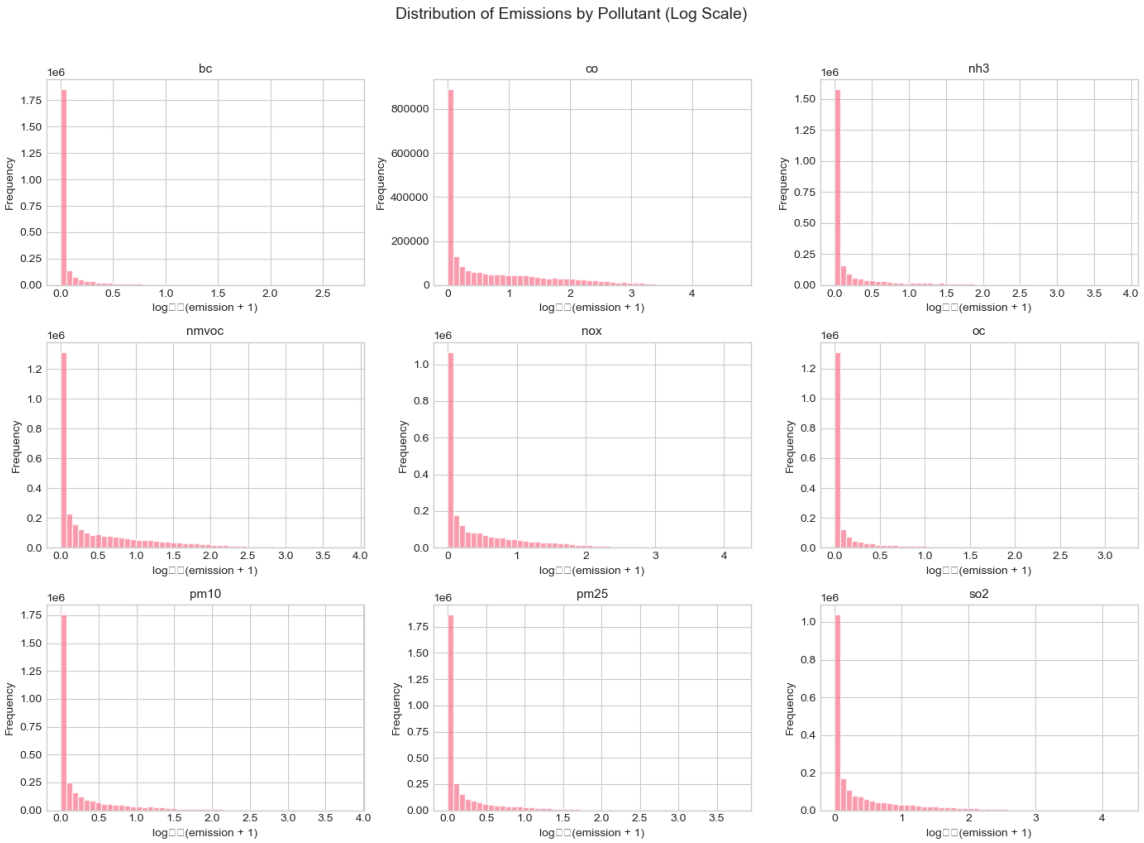

FIGURE 5 – Distribution des émissions par polluant (échelle logarithmique)

13.2 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.

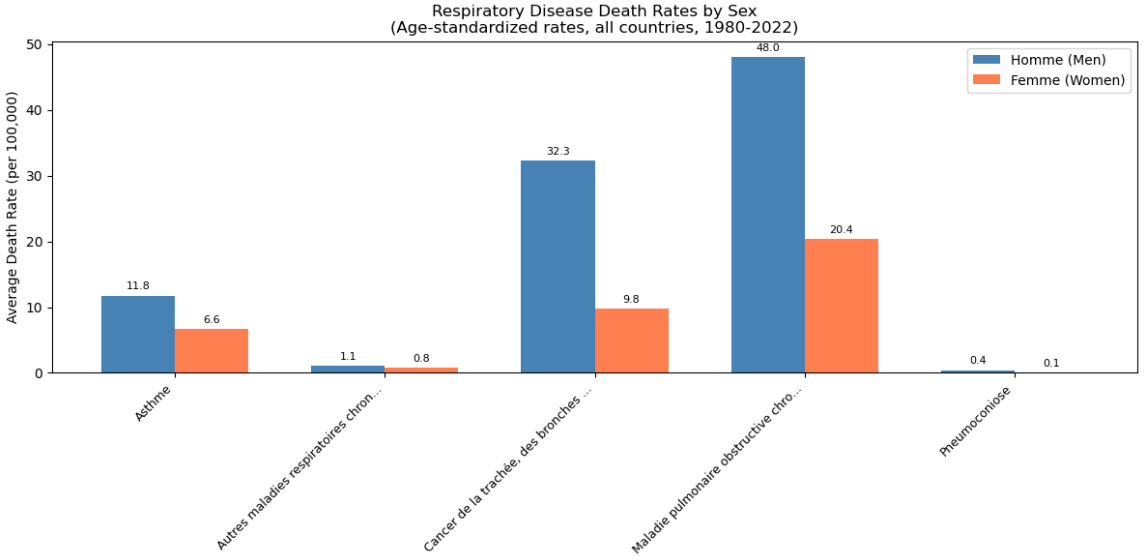

FIGURE 6 – Comparaison des taux de mortalité par sexe

14 Évolutions Temporelles

14.1 Tendances des émissions (1980–2022)

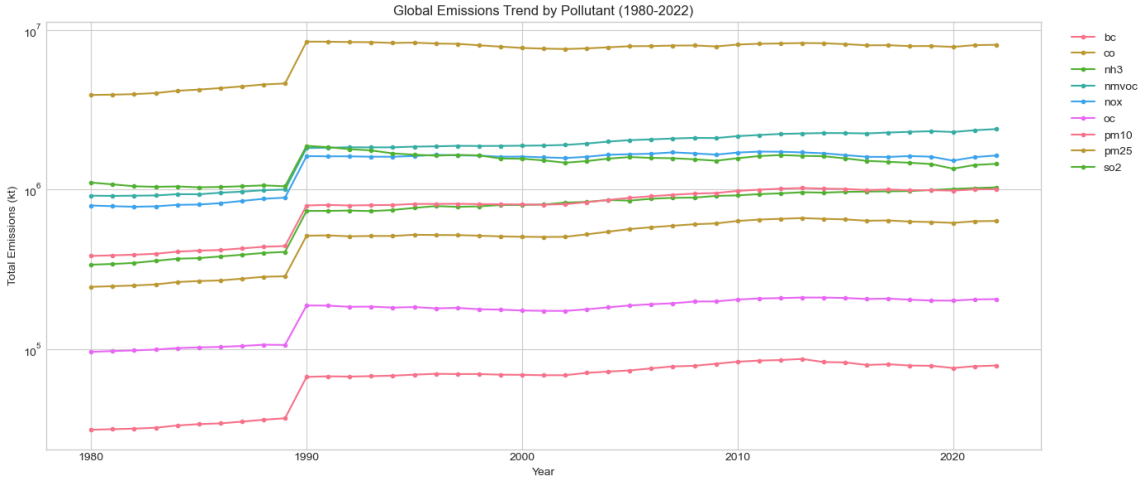

FIGURE 7 – Évolution temporelle des émissions mondiales par polluant

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

14.2 Tendances des décès (1980–2022)

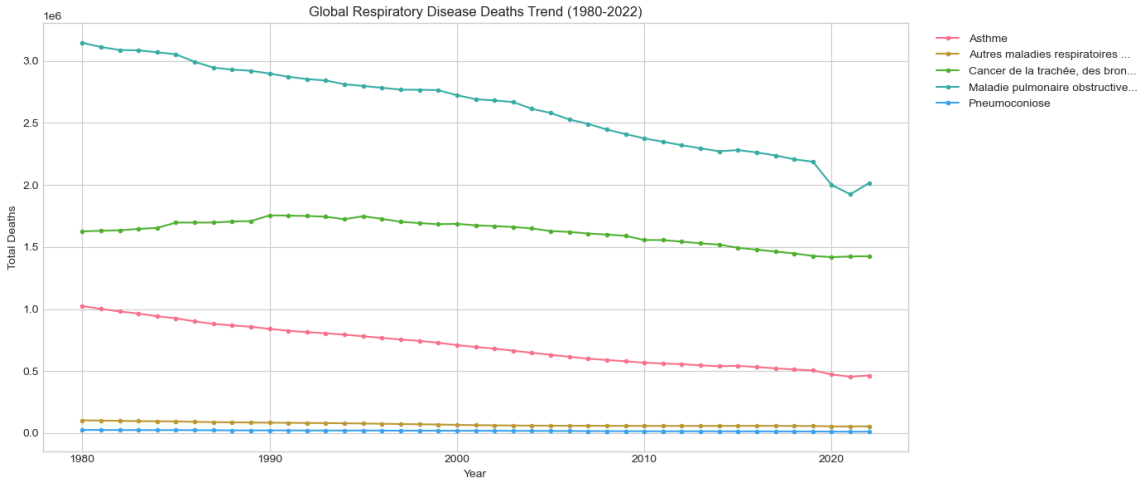

FIGURE 8 – Évolution temporelle des taux de mortalité par maladie

Observations clés :

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

15 Analyse Géographique

15.1 Pays les plus émetteurs

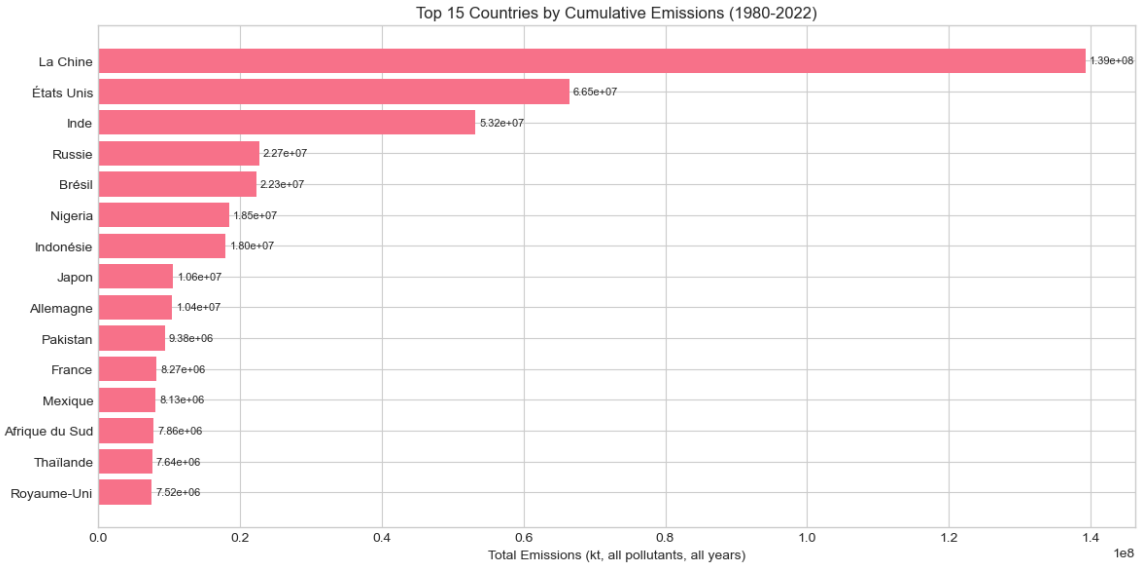

FIGURE 9 – Top 15 des pays émetteurs (toutes émissions confondues)

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

15.2 Pays avec les taux de mortalité les plus élevés

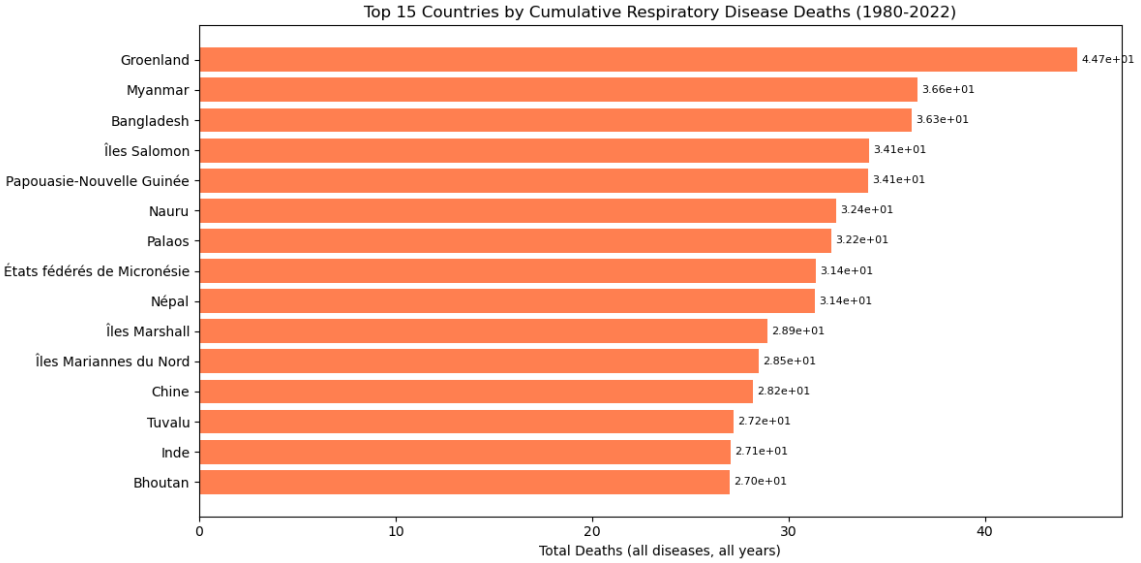

FIGURE 10 – Top 15 des pays par taux de mortalité respiratoire

15.3 Cartographie des maladies respiratoires

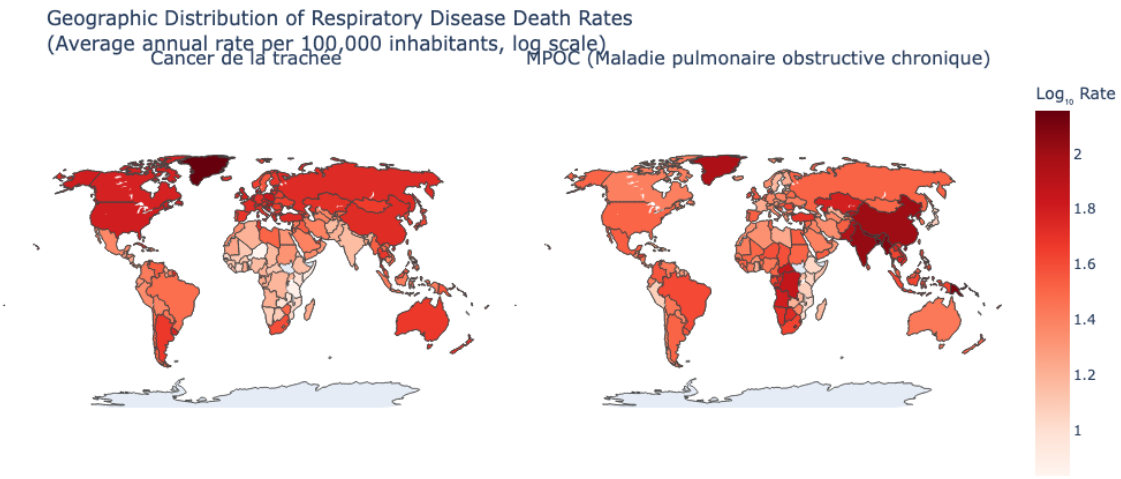

FIGURE 11 – Distribution géographique des taux de mortalité (cancer de la trachée et MPOC)

Observations :

— Le cancer de la trachée présente des taux élevés en Europe de l'Est, Amérique du Nord et Océanie
— La MPOC touche particulièrement l'Asie du Sud-Est (Chine, Inde) et certains pays africains
— Les pays développés montrent des patterns différents selon la maladie

15.4 Cartographie des émissions de polluants

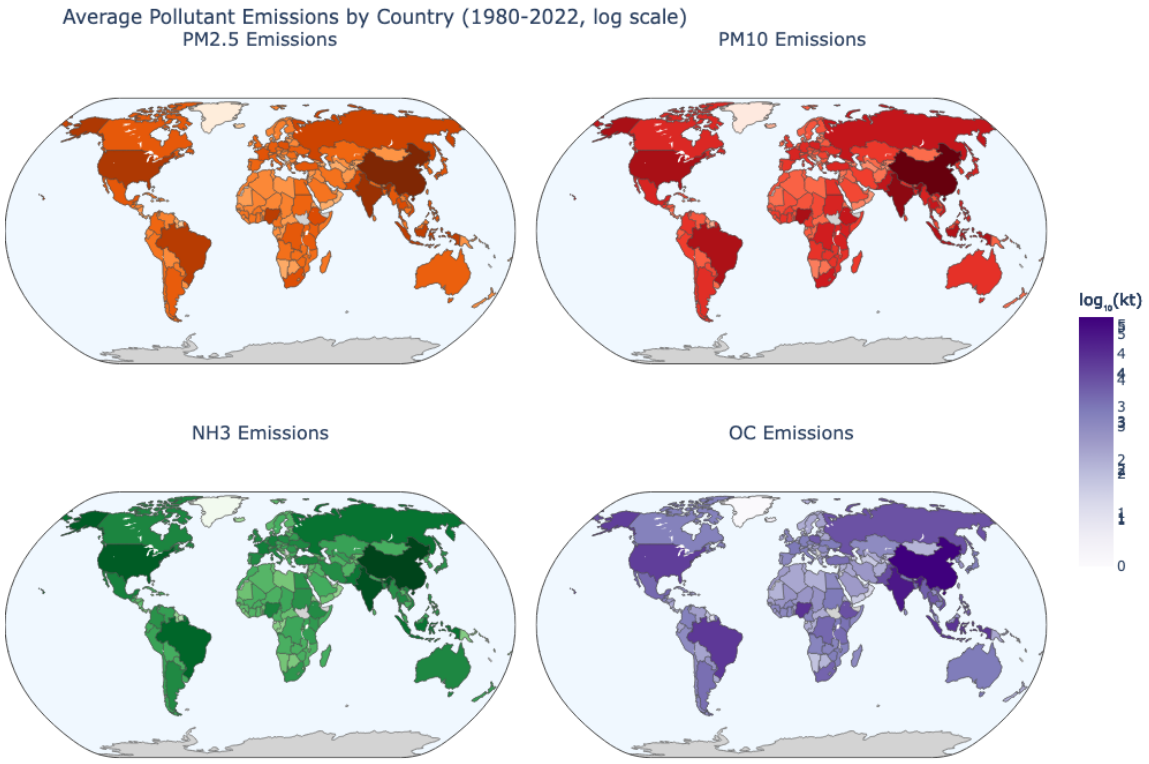

FIGURE 12 – Distribution géographique des émissions (PM2.5, PM10, NH3, OC)

Observations :

— La Chine et les États-Unis dominent pour tous les polluants
— Les particules fines (PM2.5, PM10) montrent des répartitions similaires
— L'ammoniac (NH3) est particulièrement élevé dans les pays à forte activité agricole (Inde, Chine, Brésil, États-Unis, Russie)
— Le carbone organique (OC) suit une distribution proche de celle des particules fines

16 Analyse Sectorielle

16.1 Secteurs d'activité

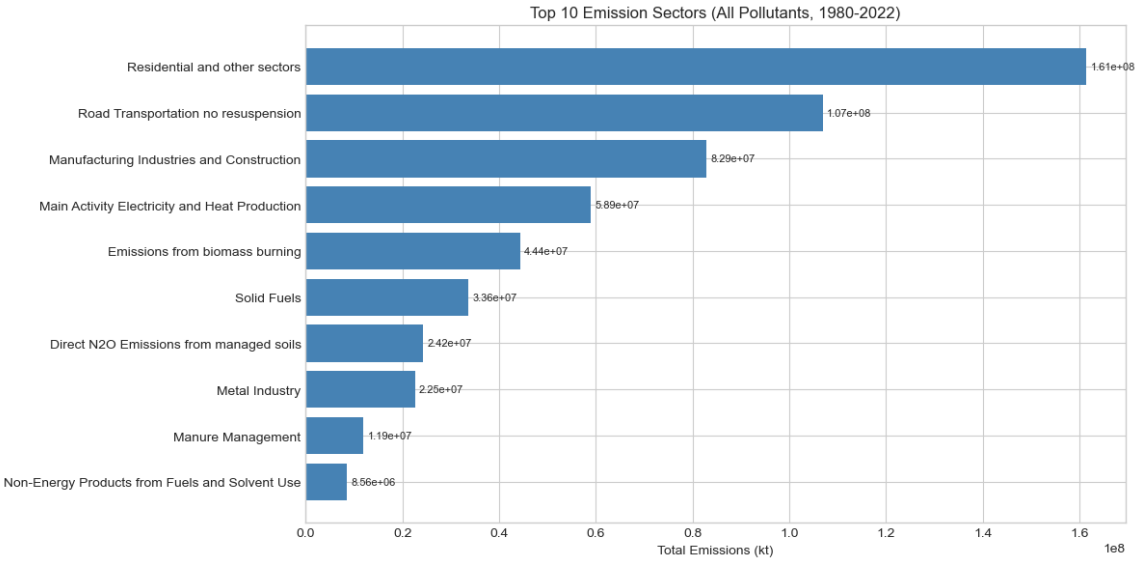

FIGURE 13 – Top 10 des secteurs d'activité par émissions totales

Les secteurs dominants sont :

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

16.2 Relation secteur-polluant

FIGURE 14 – 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.

17 Analyse des Corrélations

17.1 Corrélations polluants-maladies

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

FIGURE 15 – 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

17.2 Matrice de corrélation complète

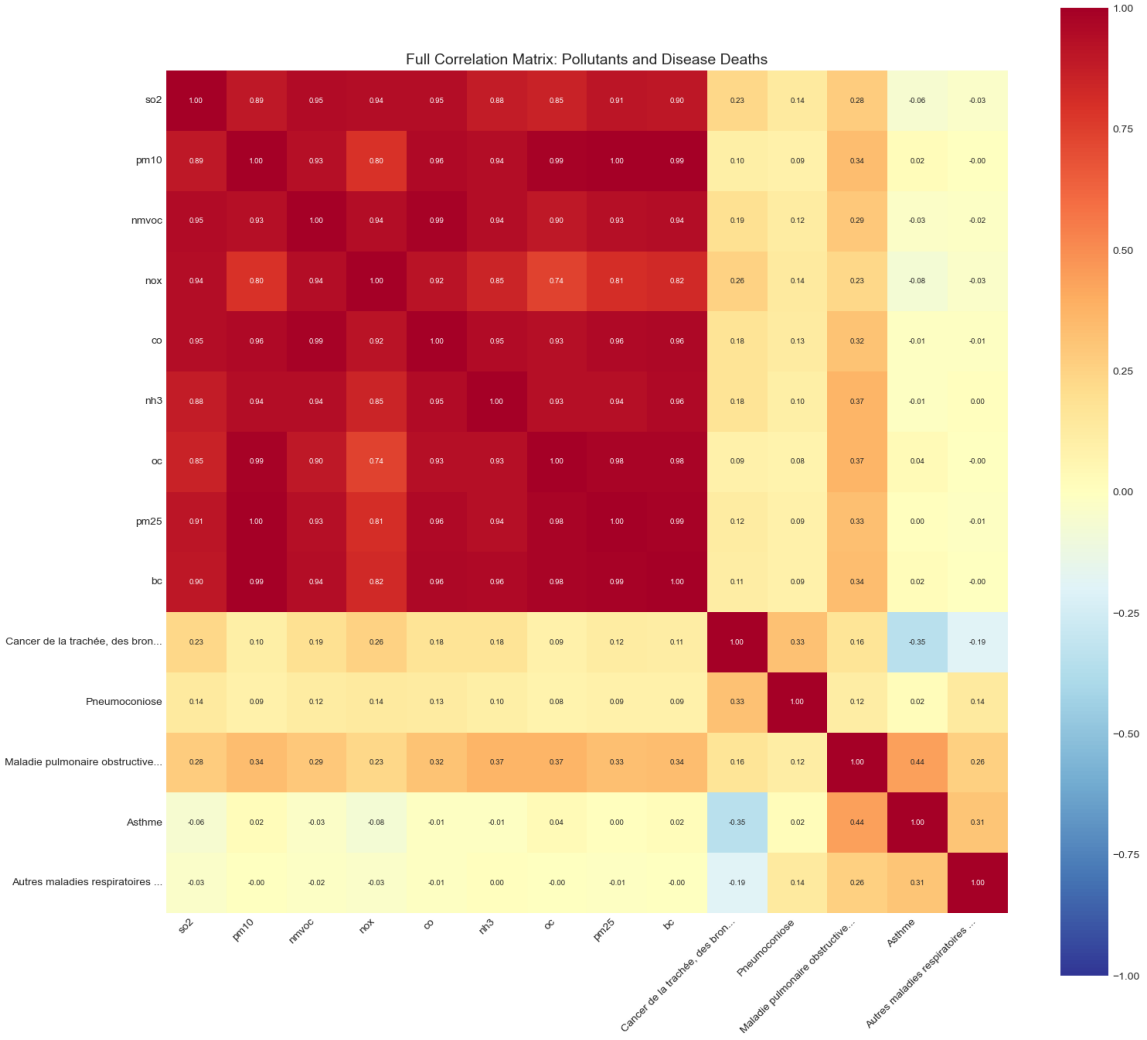

FIGURE 16 – Matrice de corrélation complète (polluants et maladies)

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)

17.3 Relations détaillées (scatter plots)

FIGURE 17 – Scatter plots : polluants clés vs maladies (échelle log)

18 Synthèse et Conclusions

18.1 Qualité des données

Table 7 – 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)

18.2 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.

18.3 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