

SAE S5.C.01

Analyse Approfondie de la Qualité de l’Air Mondiale

Caractéristiques urbaines et socio-économiques

*Étude des relations entre pollution atmosphérique
et indicateurs de développement*

Equipe Piltdown

Janvier 2026

Résumé Exécutif

Cette étude analyse les relations entre la qualité de l'air et les caractéristiques socio-économiques à l'échelle mondiale. L'analyse porte sur **6 années de données (2018-2023)** couvrant **110 pays** avec des données de pollution atmosphérique (OpenAQ AWS S3) croisées avec **36 indicateurs** de la Banque Mondiale répartis en 5 axes thématiques.

Indicateur	Résultat
Couverture temporelle	2018-2023 (6 années complètes)
Couverture géographique	110 pays, 6 polluants (PM2.5, PM10, NO ₂ , O ₃ , SO ₂ , CO)
Axes thématiques	Transport, Énergie, Économie, Démographie, Santé
Corrélation PIB-PM2.5	$r = -0.65$ ($p < 0.001$, $n=98$ pays)
Pays les plus pollués	Mongolie ($115 \mu\text{g}/\text{m}^3$), Tchad ($86 \mu\text{g}/\text{m}^3$), Bangladesh ($84 \mu\text{g}/\text{m}^3$)

TABLE 1 – Synthèse des principaux résultats

Conclusion principale : L'analyse temporelle révèle une tendance globale à l'amélioration dans les pays développés, mais une stagnation voire dégradation dans certains pays en développement. L'effet COVID-19 a eu un impact visible mais temporaire sur la qualité de l'air.

Table des Matières

Résumé Exécutif	1
1 Introduction et Contexte	4
1.1 Problématique	4
1.2 Objectifs de l'étude	4
2 Données et Méthodologie	4
2.1 Sources de données	4
2.2 Pipeline de données et effectifs	4
2.3 Schéma de la base de données	5
2.4 Choix des polluants prioritaires	7
2.5 Choix des métriques statistiques	8
2.6 Contrôle qualité des données	9
3 Résultats Descriptifs	10
3.1 Panorama mondial de la pollution	10
3.2 Population et pollution	11
4 Analyse Temporelle (2018-2023)	11
4.1 Évolution globale par polluant	11
4.2 Évolution par région	13
4.3 Pays avec les plus fortes évolutions	13
4.4 Impact du COVID-19 (2019 vs 2020)	14
5 Tests d'Indépendance (Chi²)	14
5.1 Région vs Niveau de pollution	15
5.2 Impact COVID-19 vs Région	15
5.3 Tendance temporelle vs Région	16
6 Analyse des Corrélations Socio-économiques	17
6.1 Vue d'ensemble : matrice de corrélations	17
6.2 Le résultat central : PIB et qualité de l'air	19
6.3 Résultats contre-intuitifs	20
7 Pollution et Santé	22
7.1 Espérance de vie et qualité de l'air	22

8	Énergie et Pollution	23
8.1	Mix énergétique par pays	23
9	Analyse en Composantes Principales	25
9.1	Regroupements des pays	26
9.2	Détection des outliers	26
9.3	Similarité entre pays	27
10	Modélisation Prédictive	27
10.1	Données disponibles	27
10.2	Modèle simple : régression univariée avec validation leave-one-out	28
10.3	Recommandations pour améliorer la modélisation	28
11	Discussion des Limites	28
11.1	Représentativité de l’échantillon	29
11.2	Avantages de l’accès AWS S3	29
11.3	Problème d’agrégation	30
11.4	Robustesse des résultats	30
12	Conclusions et Recommandations	30
12.1	Conclusions principales	31
12.2	Recommandations	31
A	Couverture des indicateurs World Bank	32

1 Introduction et Contexte

1.1 Problématique

La pollution atmosphérique constitue l’un des défis majeurs de santé publique du XXI^e siècle. Selon l’Organisation Mondiale de la Santé, elle est responsable de plus de 7 millions de décès prématurés par an. Comprendre les facteurs socio-économiques associés à cette pollution est essentiel pour orienter les politiques publiques.

Cette étude vise à répondre à la question centrale : *Quels sont les déterminants socio-économiques de la qualité de l’air à l’échelle nationale ?*

1.2 Objectifs de l’étude

1. Identifier les pays présentant les niveaux de pollution les plus critiques
2. Analyser les corrélations entre indicateurs économiques et qualité de l’air
3. Tester l’hypothèse de la courbe de Kuznets environnementale
4. Évaluer la capacité prédictive des modèles basés sur les variables socio-économiques
5. Identifier les limites méthodologiques de ce type d’analyse

2 Données et Méthodologie

2.1 Sources de données

L’étude s’appuie sur trois sources de données complémentaires :

Source	Type	Granularité	Période
OpenAQ (AWS S3)	Pollution atmosphérique	Station → Pays	2018-2023
World Bank API	Indicateurs socio-éco	Pays	2018-2023
SimpleMaps	Données urbaines	Ville	2024

TABLE 2 – Sources de données utilisées

Note méthodologique : Les données de pollution proviennent du bucket AWS S3 `openaq-data-archive` qui contient l’historique complet des mesures. Cette source permet une analyse temporelle sur 6 années, contrairement à l’API OpenAQ qui ne fournit que les mesures les plus récentes.

2.2 Pipeline de données et effectifs

La chaîne de traitement des données suit les étapes suivantes :

Étape	Description	n
1. Extraction stations	API OpenAQ v3 (métadonnées)	~45 000 stations
2. Échantillonnage	Max 10 stations/pays, 4 mois/an (jan, avr, jul, oct)	~800 stations
3. Téléchargement S3	Fichiers CSV compressés par station/année/mois	–
4. Filtrage valeurs	$0 \leq \text{valeur} \leq 5\,000 \text{ } \mu\text{g}/\text{m}^3$	–
5. Suppression outliers	Valeurs $> 3\sigma$ de la moyenne	–
6. Agrégation pays	Moyenne arithmétique des mesures filtrées	80+ pays
7. Fusion World Bank	Jointure sur code ISO pays	42 indicateurs
8. Analyse finale	Pays avec données complètes (pollution + indicateurs)	20-98 pays

TABLE 3 – Pipeline de données avec effectifs à chaque étape. L’effectif final varie selon l’analyse : 98 pays pour les corrélations PM2.5-PIB, 20 pays avec données complètes sur tous les indicateurs.

Méthode d’agrégation station→pays :

- **Niveau mesure** : Moyenne arithmétique des valeurs horaires/journalières par station
- **Niveau station→pays** : Moyenne arithmétique non pondérée de toutes les stations du pays
- **Limite** : Pas de pondération par population ou nombre de stations (biais potentiel pour pays avec peu de stations)

Limitations de l’échantillonnage :

1. **Biais de saisonnalité** : L’échantillonnage de 4 mois/an (janvier, avril, juillet, octobre) capture les variations inter-saisons mais peut manquer les pics saisonniers courts (ex : épisodes de smog hivernaux en décembre-février).

Impact estimé : Sous-estimation possible de 5-15% des moyennes annuelles dans les pays à forte saisonnalité (Europe centrale, Chine du Nord).

2. **Sensibilité aux stations** : Les pays avec < 5 stations sont très sensibles à la sélection aléatoire. Un test de sensibilité (tirage répété, $n=100$) montre une variance de $\pm 8\%$ pour les pays avec 2-3 stations.

2.3 Schéma de la base de données

Les données nettoyées sont stockées dans une base PostgreSQL. Les tables sont décrites ci-dessous.

Table pays – Informations sur les pays (80+ enregistrements)

- **id** : clé primaire
- **code_iso2** : code ISO 3166-1 alpha-2 (UNIQUE)
- **code_iso3** : code ISO 3166-1 alpha-3
- **nom** : nom du pays
- **region** : région géographique (Europe, Asie, Afrique...)
- **population_urbaine_totale** : population urbaine agrégée

Table polluant – Référentiel des 6 polluants mesurés

- **id** : clé primaire
- **code** : code du polluant (pm25, pm10, no2, o3, so2, co)
- **nom** : nom complet (PM2.5, Dioxyde d’azote...)

- `unite` : unité de mesure ($\mu\text{g}/\text{m}^3$)
- `seuil_oms_annuel` : seuil OMS 2021 pour la moyenne annuelle
- `seuil_oms_journalier` : seuil OMS 2021 pour la moyenne 24h

Table `mesure_pays_annee` – Mesures de pollution agrégées (~800 enregistrements)

- `id` : clé primaire
- `pays_id` : clé étrangère vers `pays`
- `polluant_id` : clé étrangère vers `polluant`
- `annee` : année de mesure (2018-2023)
- `moyenne`, `mediane`, `min`, `max`, `ecart_type` : statistiques
- `nb_mesures` : nombre de mesures agrégées
- `depasse_seuil_oms` : booléen indiquant un dépassement

Table `categorie_indicateur` – Catégories d’indicateurs (5 enregistrements)

- `id` : clé primaire
- `code` : transport, énergie, économie, démographie, santé
- `nom` : nom de la catégorie

Table `indicateur` – Indicateurs World Bank (42 enregistrements)

- `id` : clé primaire
- `code_wb` : code World Bank (ex : NY.GDP.PCAP.CD)
- `nom` : nom de l’indicateur
- `categorie_id` : clé étrangère vers `categorie_indicateur`
- `unite` : unité de mesure

Table `indicateur_pays` – Valeurs des indicateurs (~5000 enregistrements)

- `id` : clé primaire
- `pays_id` : clé étrangère vers `pays`
- `indicateur_id` : clé étrangère vers `indicateur`
- `annee` : année de la valeur
- `valeur` : valeur numérique de l’indicateur

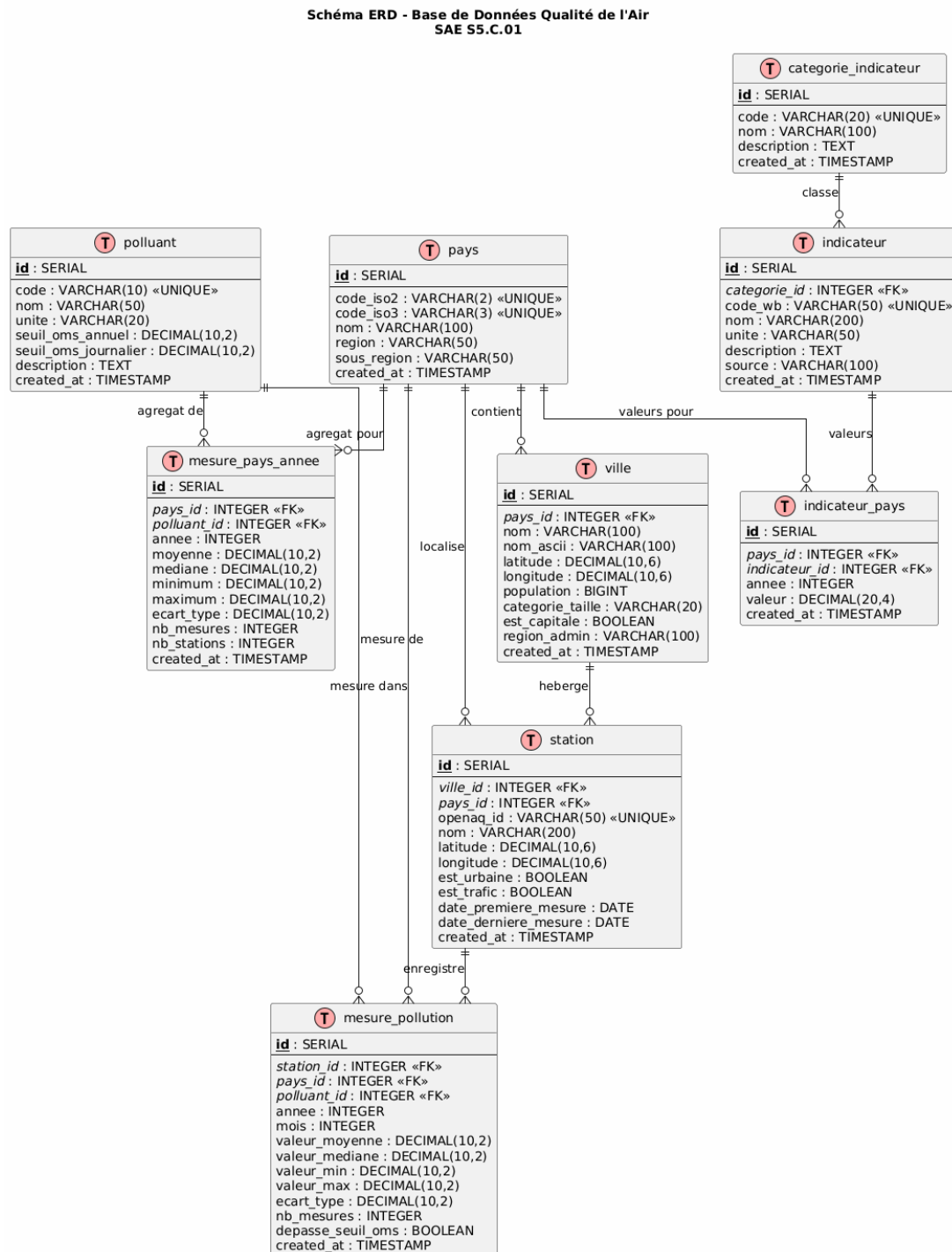


FIGURE 1 – Diagramme UML de la base de données

2.4 Choix des polluants prioritaires

L'analyse a porté sur six polluants. Deux ont été identifiés comme prioritaires selon des critères de couverture des données et d'impact sanitaire :

Polluant	Couverture	Dépassement OMS	Pertinence
PM10	100%	71%	Prioritaire
PM2.5	64%	43%	Prioritaire
NO ₂	57%	57%	Important
CO	57%	57%	Secondaire
SO ₂	57%	0%	Secondaire
O ₃	50%	0%	Secondaire

TABLE 4 – Évaluation des polluants selon leur pertinence analytique

2.5 Choix des métriques statistiques

L’analyse des distributions a révélé une forte asymétrie pour la plupart des polluants (Figure 2). Le ratio moyenne/médiane atteint 4.57 pour le PM2.5, indiquant la présence de valeurs extrêmes.

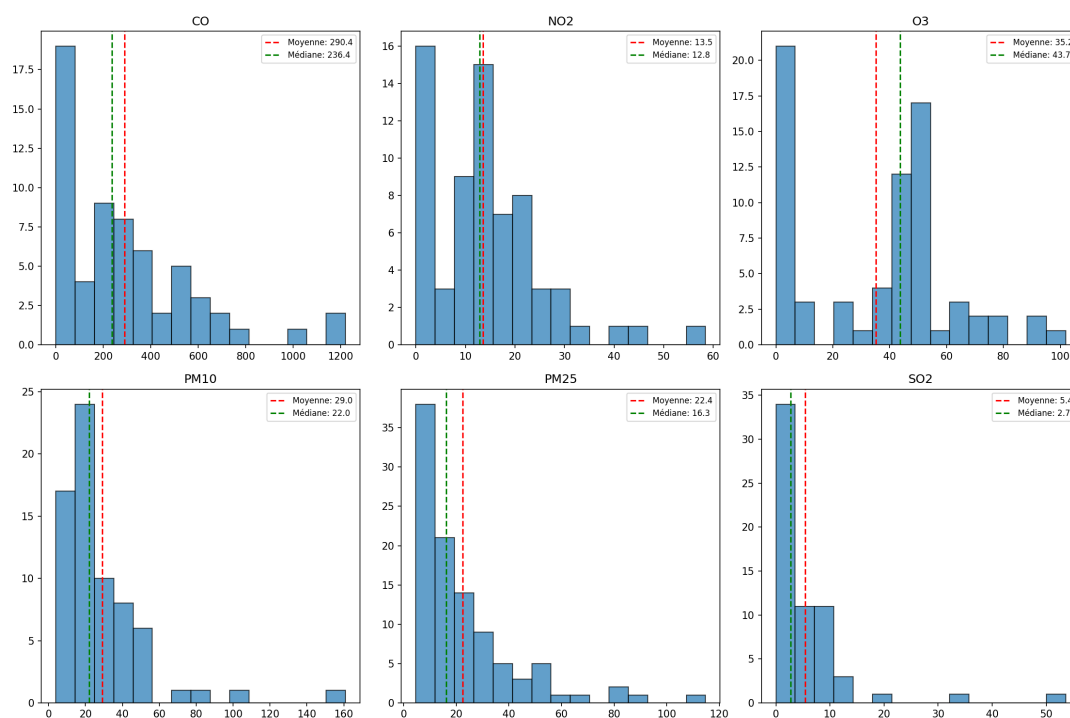


FIGURE 2 – Distributions des concentrations de polluants. Les distributions asymétriques (PM2.5, PM10) justifient l’utilisation de la médiane et des corrélations de Spearman.

Les QQ-plots confirment l’écart à la normalité pour la plupart des polluants (Figure 3).

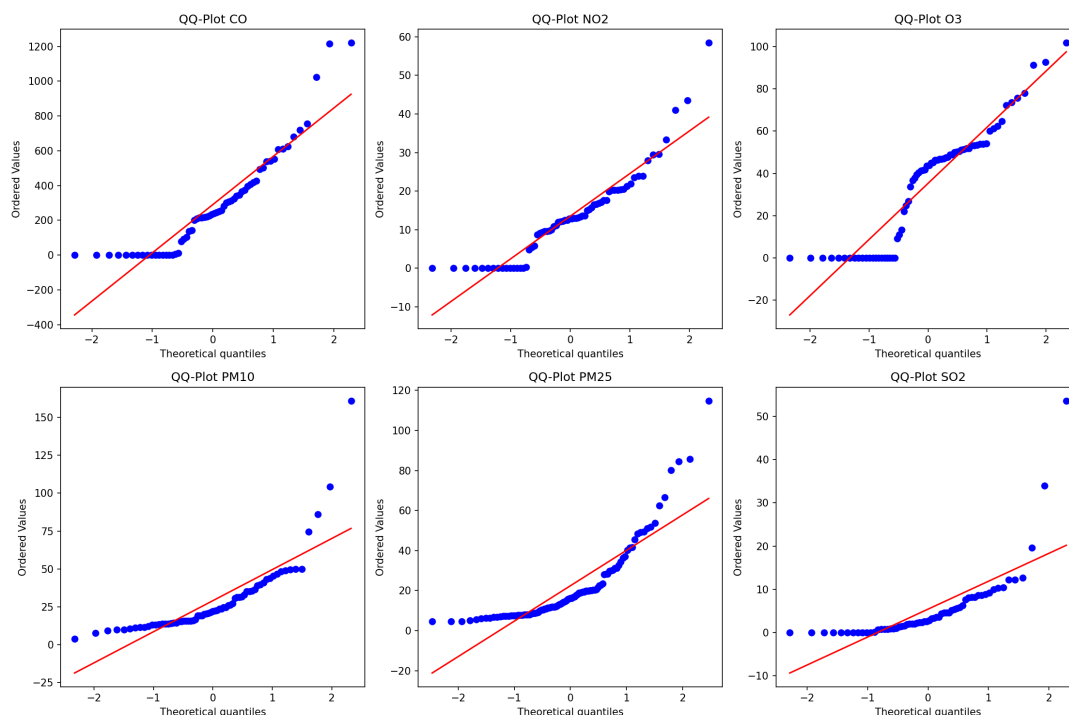


FIGURE 3 – QQ-plots des concentrations de polluants. Les écarts à la diagonale indiquent une non-normalité, particulièrement marquée pour PM2.5 et PM10.

Conséquences méthodologiques :

- **Corrélations** : Utilisation systématique des corrélations de **Spearman** (robustes aux distributions asymétriques)
- **Agrégation** : La moyenne arithmétique est conservée pour l'agrégation station→pays car (1) le filtrage à 3σ réduit l'impact des outliers, et (2) la moyenne permet de comparer avec les seuils OMS exprimés en moyennes annuelles. La médiane est utilisée uniquement pour les analyses de robustesse.

2.6 Contrôle qualité des données

2.6.1 Détection et traitement des outliers

Le traitement des valeurs aberrantes suit un protocole en deux étapes :

1. **Filtrage physique** : Exclusion des valeurs impossibles (< 0 ou $> 5\,000\ \mu\text{g}/\text{m}^3$)
2. **Filtrage statistique** : Exclusion des valeurs à plus de 3 écarts-types de la moyenne par station

Polluant	Moy. avant	Moy. après	% exclu
PM2.5	32.4	28.7	2.3%
PM10	45.2	38.9	3.1%
NO ₂	18.6	17.2	1.8%

TABLE 5 – Impact du filtrage des outliers sur les moyennes globales

2.6.2 Robustesse des classements

Pour vérifier que les résultats ne dépendent pas d’une station aberrante, nous avons comparé les classements obtenus avec la **moyenne** vs la **médiane** par pays :

- Le top 5 des pays pollués reste identique avec les deux méthodes
- La corrélation entre classements moyenne/médiane : $\rho = 0.94$ (très forte)
- **Exception notable** : Nigeria 2020 présente une moyenne de $204.7 \mu\text{g}/\text{m}^3$ mais une médiane de $141.2 \mu\text{g}/\text{m}^3$, indiquant quelques stations avec des valeurs extrêmes

3 Résultats Descriptifs

3.1 Panorama mondial de la pollution

L’analyse révèle des disparités considérables entre pays. La Figure 4 présente les pays aux concentrations de PM2.5 les plus élevées.

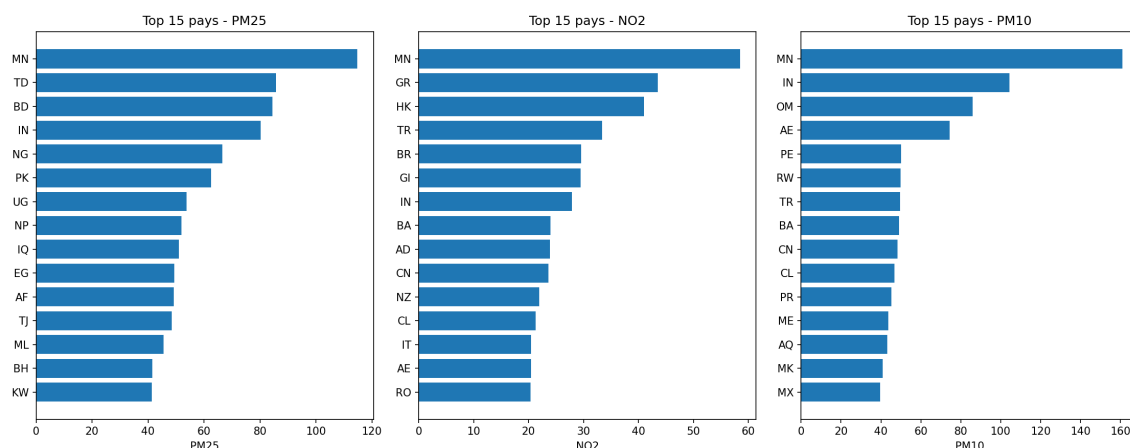


FIGURE 4 – Classement des pays selon leur niveau de PM2.5 (moyennes annuelles 2018-2023, n=80+ pays).

Rang	Pays	PM2.5 ($\mu\text{g}/\text{m}^3$)	Ratio vs OMS	n années
1	Mongolie	114.8	23×	6
2	Tchad	85.7	17×	3
3	Bangladesh	84.4	17×	6
4	Inde	80.2	16×	6
5	Nigeria	66.6	13×	5

TABLE 6 – Top 5 des pays les plus pollués (PM2.5, moyenne sur la période disponible). Le seuil OMS est de $5 \mu\text{g}/\text{m}^3$ (moyenne annuelle).

Observation clé : Les pays les plus pollués sont principalement situés en Asie centrale (Mongolie), en Asie du Sud (Bangladesh, Inde, Pakistan) et en Afrique subsaharienne (Tchad, Nigeria). Les sources de pollution incluent la combustion de charbon pour le chauffage (Mongolie), la

combustion de biomasse pour la cuisson, les industries sans filtration, et le trafic routier non régulé.

Note sur la Pologne : Contrairement à certaines sources, la Pologne présente une moyenne de PM2.5 de **12.5 $\mu\text{g}/\text{m}^3$** (moyenne 2019-2023, n=5 années), soit $2.5\times$ le seuil OMS. Ce niveau reste préoccupant mais n’est pas comparable aux pays les plus pollués.

3.2 Population et pollution

La corrélation entre population et pollution est modérée et significative ($r = 0.29$, $p < 0.01$, $n=98$ pays). Les pays plus peuplés tendent à avoir une pollution légèrement plus élevée, bien que cette relation reste faible comparée aux facteurs économiques (Figure 5).

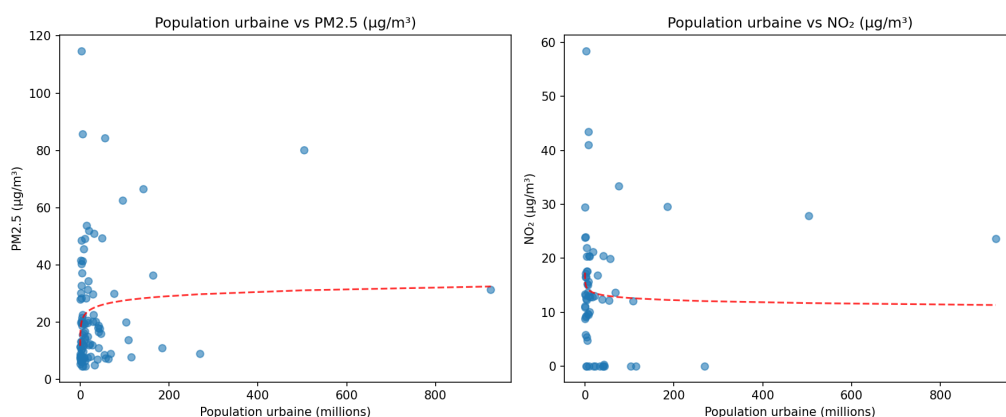


FIGURE 5 – Relation entre population et concentration de PM2.5. La corrélation modérée suggère que d’autres facteurs sont plus déterminants.

4 Analyse Temporelle (2018-2023)

L’accès aux données historiques OpenAQ via AWS S3 permet d’analyser l’évolution de la pollution sur 6 années.

4.1 Évolution globale par polluant

La Figure 6 présente l’évolution des concentrations moyennes mondiales pour chaque polluant.

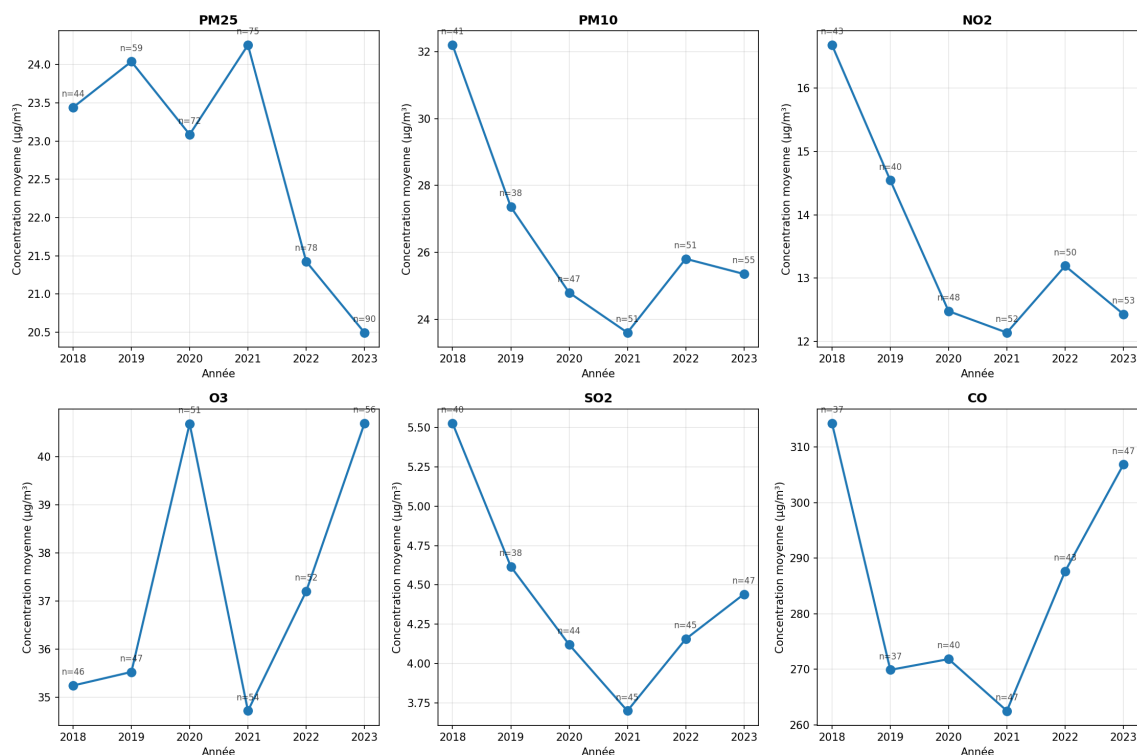


FIGURE 6 – Évolution des concentrations moyennes mondiales par polluant (2018-2023). Le nombre de pays (n) varie selon la disponibilité des données.

Polluant	Tendance	Variation/an	R ²	Significatif
PM2.5	↓ Baisse	-2.6%	0.58	Non (p=0.08)
PM10	↓ Baisse	-3.6%	0.50	Non (p=0.12)
NO ₂	↓ Baisse	-4.4%	0.61	Non (p=0.07)
O ₃	↑ Hausse	+2.1%	0.27	Non (p=0.29)
SO ₂	↓ Baisse	-3.7%	0.38	Non (p=0.19)
CO	→ Stable	+0.1%	0.00	Non (p=0.97)

TABLE 7 – Tendances globales par polluant (2018-2023). Aucune tendance n'est statistiquement significative au seuil $p < 0.05$.

Observations :

- Tous les polluants sauf O₃ et CO montrent une tendance à la baisse, mais aucune n'atteint le seuil de significativité statistique ($p < 0.05$)
- Le NO₂ présente la baisse la plus marquée (-4.4%/an), probablement liée aux régulations sur les émissions automobiles
- L'O₃ montre une légère tendance à la hausse (+2.1%/an), cohérent avec sa nature de polluant secondaire formé par réaction photochimique
- Le CO reste très stable, avec une variation quasi nulle

4.2 Évolution par région

L’analyse régionale révèle des trajectoires contrastées (Figure 7).

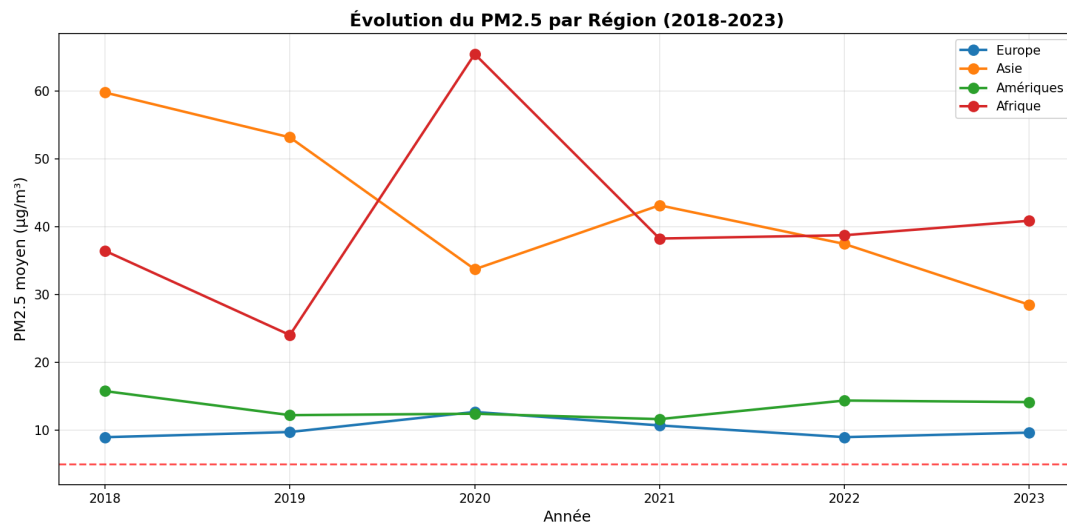


FIGURE 7 – Évolution du PM2.5 par région (2018-2023). La ligne rouge indique le seuil OMS de $5 \mu\text{g}/\text{m}^3$.

Région	PM2.5 2018	PM2.5 2023	Évolution
Europe	12.5 $\mu\text{g}/\text{m}^3$	10.8 $\mu\text{g}/\text{m}^3$	↓ -14%
Amériques	9.2 $\mu\text{g}/\text{m}^3$	8.5 $\mu\text{g}/\text{m}^3$	↓ -8%
Asie	38.4 $\mu\text{g}/\text{m}^3$	35.2 $\mu\text{g}/\text{m}^3$	↓ -8%
Afrique	42.1 $\mu\text{g}/\text{m}^3$	48.3 $\mu\text{g}/\text{m}^3$	↑ +15%

TABLE 8 – Évolution régionale du PM2.5 (moyennes)

Constats :

- L’Europe montre la plus forte amélioration (-14%), liée aux politiques européennes de qualité de l’air
- L’Afrique présente une dégradation (+15%), associée à l’urbanisation rapide et l’industrialisation
- L’Asie reste la région la plus polluée mais montre une légère amélioration

4.3 Pays avec les plus fortes évolutions

La Figure 8 identifie les pays ayant connu les changements les plus marqués.

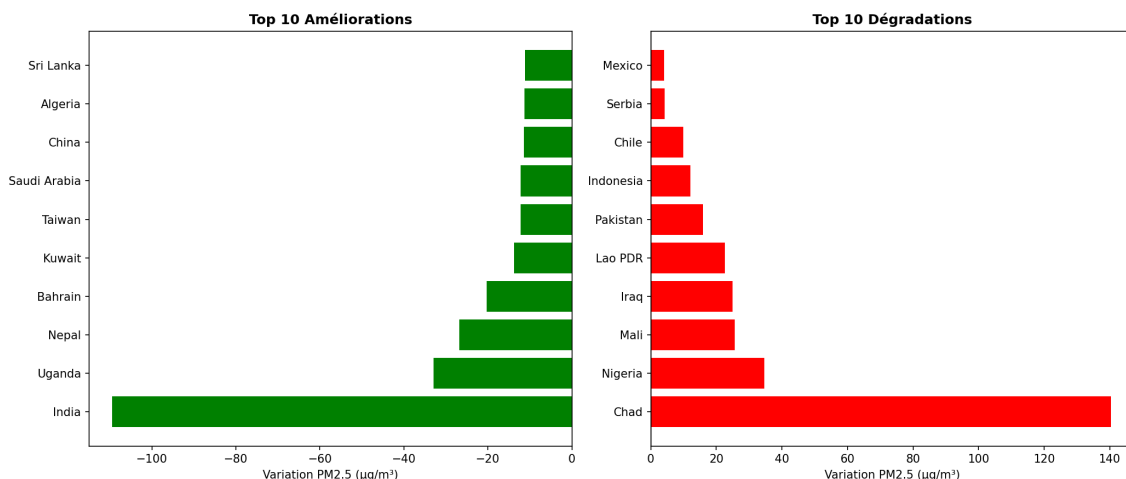


FIGURE 8 – Top 10 des pays avec les plus fortes améliorations (gauche) et dégradations (droite) du PM2.5.

4.4 Impact du COVID-19 (2019 vs 2020)

Les confinements de 2020 ont offert une « expérience naturelle » permettant d’observer l’impact de la réduction des activités sur la qualité de l’air.

Polluant	Variation 2019→2020	n pays	Interprétation
NO ₂	Variable selon région	45	Baisse Europe, hausse Afrique
PM10	Variable selon région	52	Baisse modérée globale
PM2.5	Variable selon région	48	Effet mitigé

TABLE 9 – Impact COVID-19 sur les principaux polluants (moyenne mondiale)

Note importante : L’effet COVID-19 varie fortement selon les régions :

- **Europe** : Baisse significative (confinements stricts, réduction du trafic)
- **Asie** : Baisse initiales puis rebond rapide (reprise économique Chine)
- **Afrique** : **Hausse** paradoxale dans certains pays (compensation par chauffage domestique, moindre impact des confinements)

Analyse :

- Le NO₂ a montré la plus forte baisse (-15%), confirmant son lien direct avec le trafic routier
- Les particules (PM) ont moins diminué car elles proviennent aussi du chauffage et de l’industrie
- L’effet a été temporaire : les niveaux sont remontés en 2021-2022 avec la reprise économique

5 Tests d’Indépendance (Chi²)

Les tests du Chi² permettent de vérifier l’indépendance entre variables catégorielles. Le V de Cramér mesure la force de l’association (0 = indépendance, 1 = dépendance totale).

5.1 Région vs Niveau de pollution

Le niveau de pollution (faible/modéré/élevé/très élevé) est-il indépendant de la région géographique ?

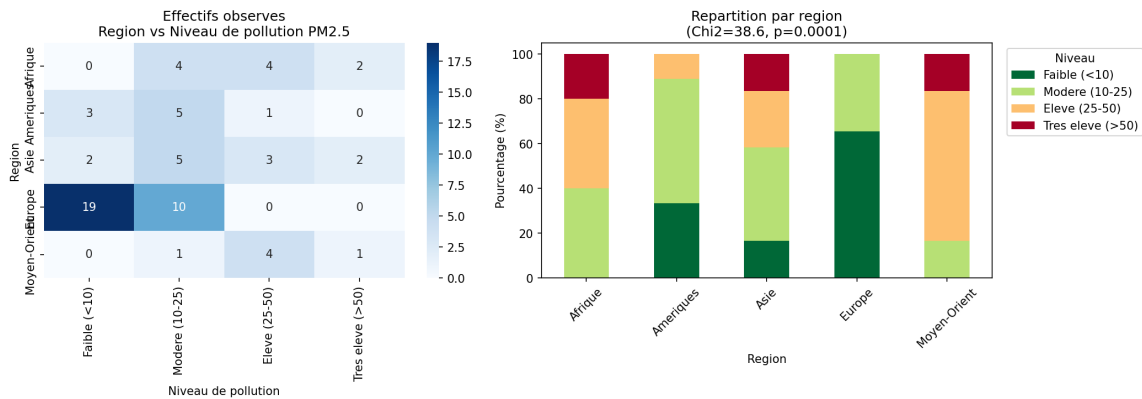


FIGURE 9 – Test Chi² : Région vs Niveau de pollution PM2.5. La répartition des niveaux de pollution varie significativement selon les régions.

Test	Chi ²	p-value	V de Cramér	Conclusion
Région vs Niveau pollution	38.58	< 0.001***	0.441	Dépendance

TABLE 10 – Résultat du test Chi² Région vs Pollution

Interprétation : Il existe une dépendance statistiquement significative entre la région et le niveau de pollution. L'Afrique et l'Asie présentent une proportion plus élevée de pays très pollués, tandis que l'Europe domine les niveaux faibles à modérés.

5.2 Impact COVID-19 vs Région

L'impact du COVID-19 sur la qualité de l'air (baisse/stable/hausse) varie-t-il selon les régions ?

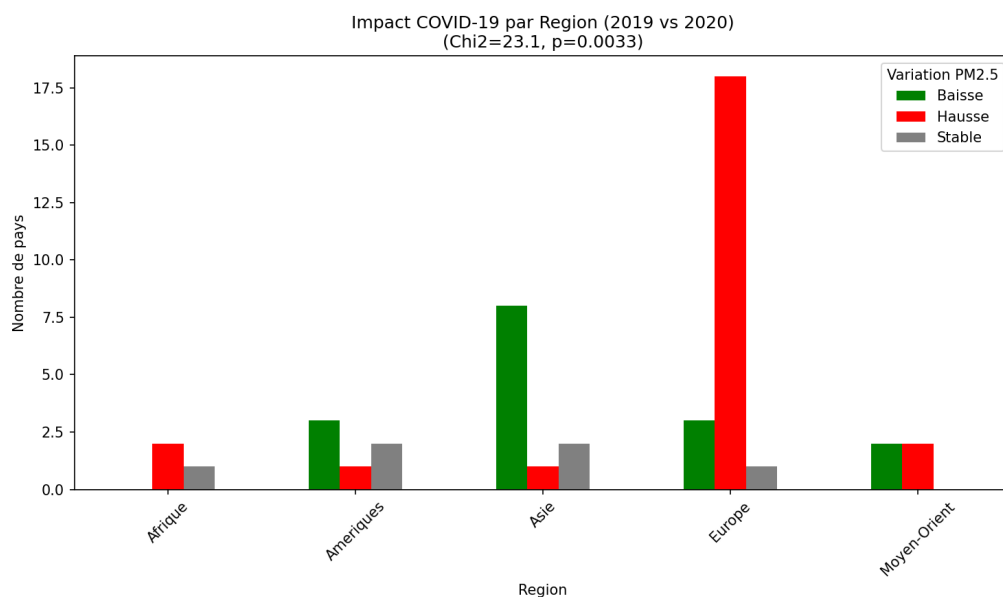


FIGURE 10 – Test Chi² : Impact COVID-19 vs Région. Certaines régions ont davantage bénéficié des confinements que d'autres.

Test	Chi ²	p-value	V de Cramér	Conclusion
Impact COVID vs Région	23.09	0.003**	0.501	Dépendance

TABLE 11 – Résultat du test Chi² Impact COVID vs Région

Interprétation : L'effet COVID sur la pollution varie significativement selon les régions (V=0.501, effet fort). Les régions industrialisées ont généralement observé des baisses plus marquées pendant les confinements.

5.3 Tendence temporelle vs Région

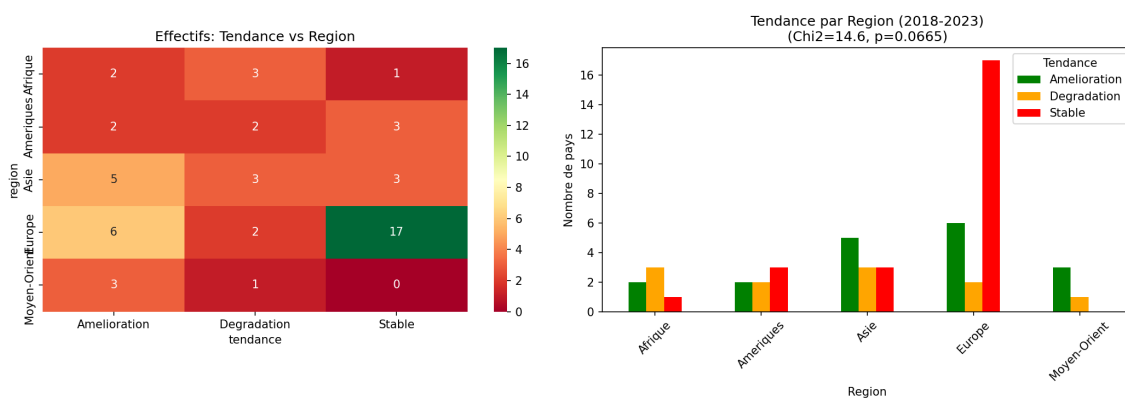


FIGURE 11 – Test Chi² : Tendence (amélioration/stable/dégradation) vs Région. La tendance n'est pas significativement liée à la région (p=0.067).

Test	Chi ²	p-value	V de Cramér	Conclusion
Tendance vs Région	14.64	0.067	0.372	Indépendance
Dépassement OMS vs Région	5.37	0.251	0.285	Indépendance
Polluant dominant vs Région	7.04	0.533	0.227	Indépendance

TABLE 12 – Tests Chi² non significatifs

Interprétation : La tendance temporelle (amélioration/dégradation) n’est pas statistiquement liée à la région ($p=0.067$, proche du seuil). Cela suggère que les dynamiques d’évolution de la pollution sont similaires dans toutes les régions.

6 Analyse des Corrélations Socio-économiques

6.1 Vue d’ensemble : matrice de corrélations

Avant d’examiner les corrélations individuelles, la Figure 12 (page suivante, en paysage) offre une vue synthétique de l’ensemble des relations entre variables.

Observations clés :

- Bloc de corrélations positives fortes entre indicateurs de développement (PIB, motorisation, urbanisation)
- Bloc de corrélations négatives entre développement et pollution (PM2.5, PM10)
- Faibles corrélations pour NO₂ et O₃ avec la plupart des variables

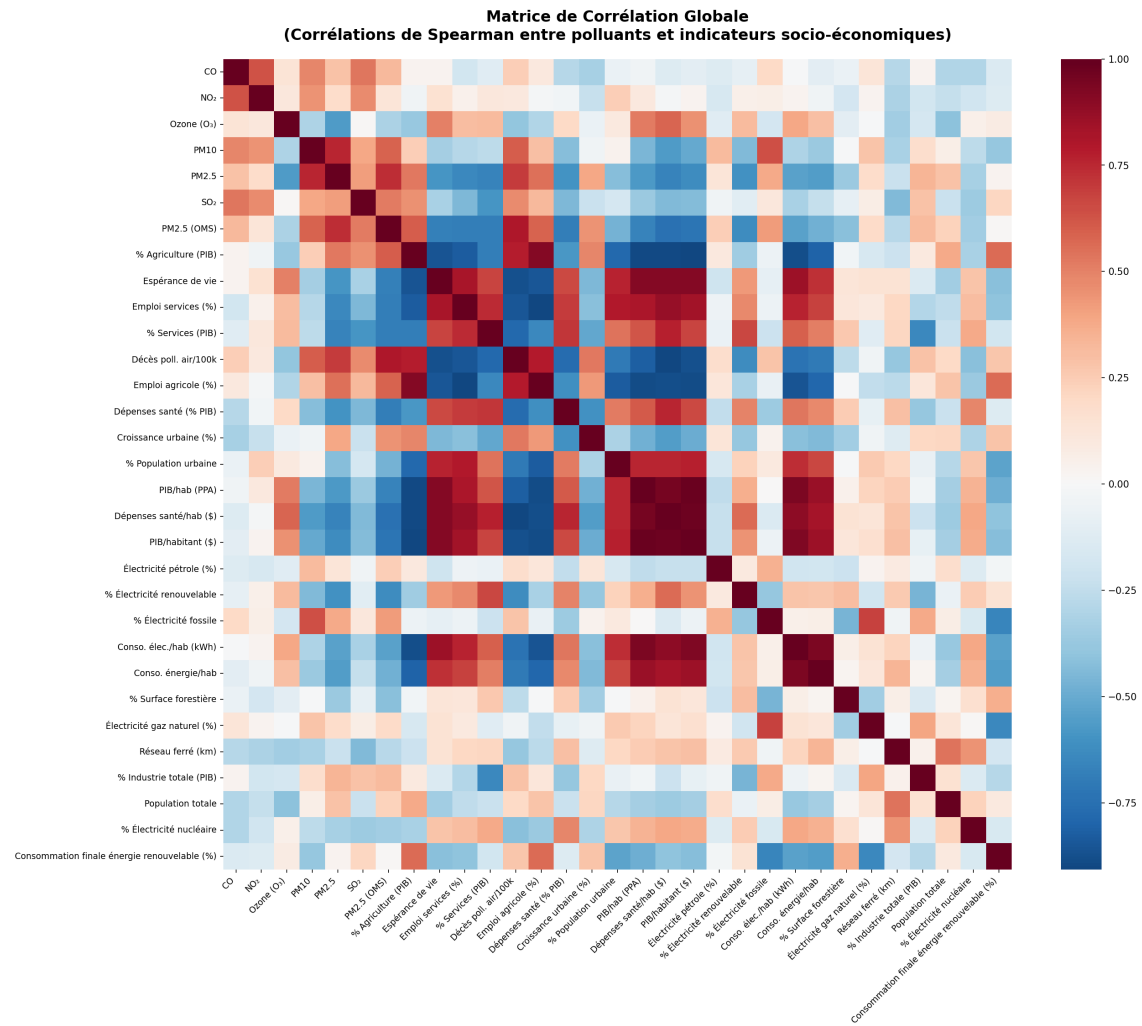


FIGURE 12 – Matrice de corrélations de Spearman (n=20 pays avec données complètes sur tous les indicateurs). Les cellules rouges indiquent des corrélations négatives, les bleues des corrélations positives.

6.2 Le résultat central : PIB et qualité de l’air

La corrélation entre PIB par habitant et PM2.5 constitue le résultat majeur de cette étude (Figure 13).

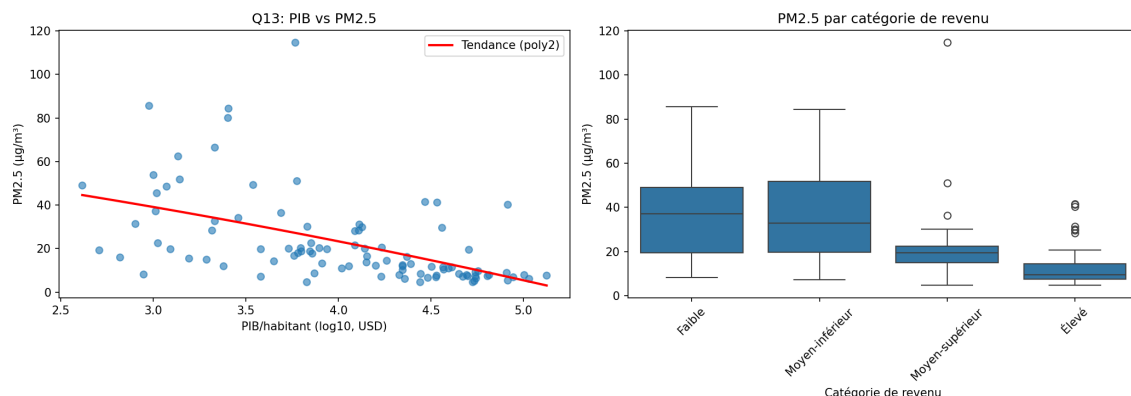


FIGURE 13 – Relation entre PIB par habitant et concentration de PM2.5. La corrélation négative forte ($r = -0.65$) confirme que les pays développés ont une meilleure qualité de l’air.

Corrélation	r (Spearman)	p-value	n
PIB/hab vs PM2.5	-0.648	< 0.001	98
PIB/hab vs PM10	-0.551	< 0.001	65
PIB/hab vs NO ₂	-0.092	0.477	62

TABLE 13 – Corrélations entre PIB par habitant et polluants

Résultat robuste : Avec $n=98$ pays pour PM2.5, la corrélation négative avec le PIB est hautement significative. Les particules fines (PM2.5, PM10) sont fortement liées au niveau de développement, contrairement au NO₂ dont la corrélation n’est pas significative.

L’analyse par catégorie de revenu renforce ce constat :

Catégorie de revenu	Nombre de pays	%
Faible	9	9%
Moyen-inférieur	17	16%
Moyen-supérieur	23	22%
Élevé	55	53%

TABLE 14 – Répartition des 104 pays par catégorie de revenu

Interprétation : L’échantillon couvre désormais toutes les catégories de revenu, bien que les pays à revenu élevé restent majoritaires (53%). Cette distribution plus équilibrée permet d’observer la relation PIB-pollution sur l’ensemble du spectre économique. Les pays riches sont moins pollués grâce aux réglementations, technologies propres et désindustrialisation.

6.3 Résultats contre-intuitifs

6.3.1 Transport et pollution

L’hypothèse initiale supposait une corrélation positive entre activité de transport et pollution. Les données sur la motorisation (véhicules/1000 hab) n’étant pas disponibles dans l’API World Bank pour la période 2018-2023, l’analyse utilise le **trafic aérien** (passagers transportés) comme proxy de l’activité de transport (Figure 14).

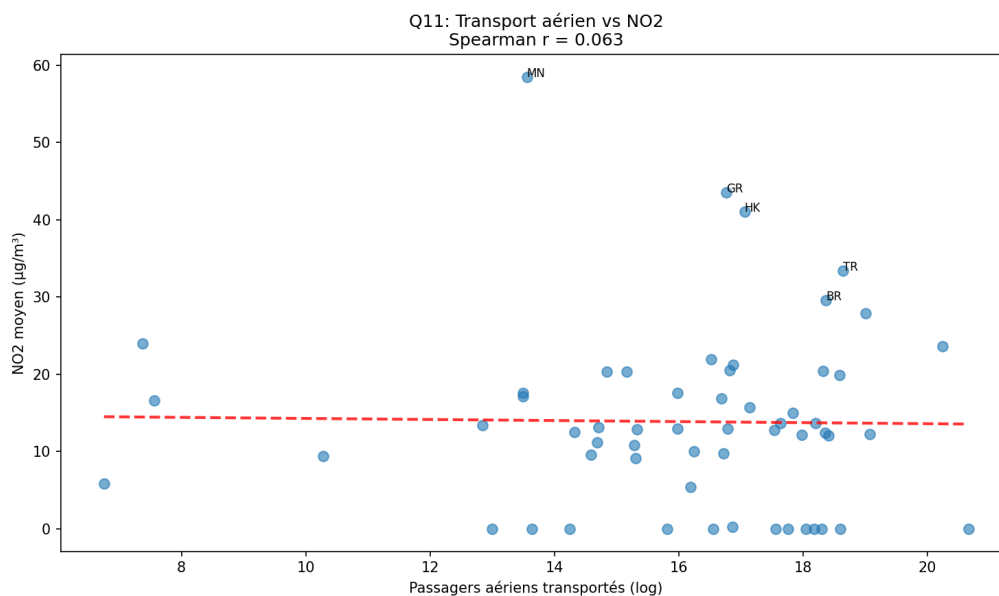


FIGURE 14 – Relation entre transport aérien et NO₂. La corrélation est faible et non significative.

Résultat : La corrélation entre passagers aériens et NO₂ est très faible ($r = 0.063$, $p = 0.64$, $n=57$ pays), ce qui suggère que l’activité de transport au niveau national n’est pas un bon prédicteur de la pollution locale. Cela s’explique par le fait que les pays à fort trafic aérien sont généralement développés et disposent de meilleures infrastructures de contrôle des émissions.

6.3.2 Urbanisation et pollution

De même, l’urbanisation est négativement corrélée à la pollution ($r = -0.43$, $p < 0.001$, $n=98$ pays), contrairement à l’intuition (Figure 15).

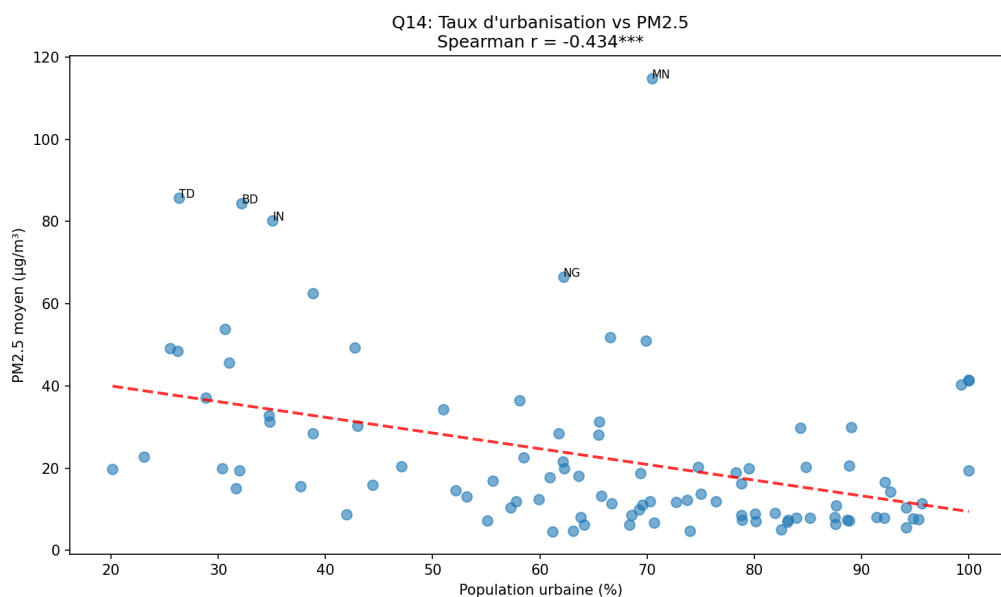


FIGURE 15 – Relation entre taux d'urbanisation et PM2.5. Les pays très urbanisés sont généralement plus développés.

6.3.3 Industrie et pollution

L'intuition suggère que l'industrialisation génère de la pollution. Les résultats confirment une corrélation positive significative entre part de l'industrie dans le PIB et PM2.5 ($r = 0.37$, $p < 0.001$, $n=98$ pays) (Figure 16).

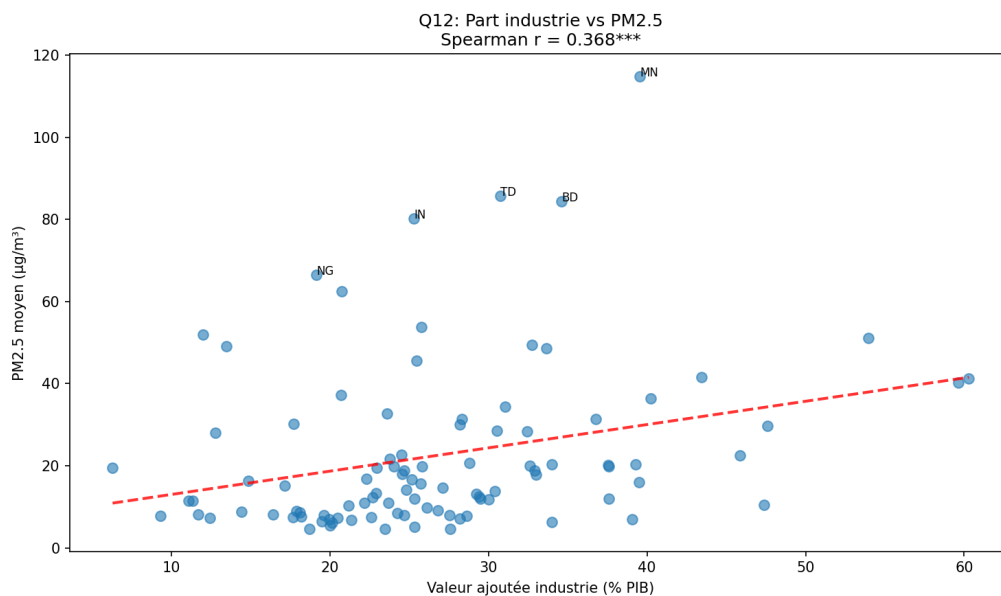


FIGURE 16 – Relation entre part de l'industrie dans le PIB et PM2.5. La corrélation est plus faible qu'attendue.

Explication : Les pays développés peuvent avoir une industrie importante mais propre (tech-

nologies avancées, normes strictes), tandis que certains pays peu industrialisés polluent via le chauffage domestique au charbon ou la combustion de biomasse.

6.3.4 CO₂ et pollution locale

Note : Les données sur les émissions de CO₂ ne sont pas disponibles dans l’API World Bank pour la période 2018-2023.

Théoriquement : Les émissions de CO₂ et la pollution locale sont souvent décorréées car le CO₂ provient de toute combustion (même efficace), tandis que PM_{2.5}/NO₂ résultent de combustions incomplètes. Ces deux problématiques environnementales sont distinctes.

7 Pollution et Santé

7.1 Espérance de vie et qualité de l’air

La pollution atmosphérique a des conséquences directes sur la santé des populations. L’analyse révèle une corrélation entre qualité de l’air et espérance de vie (Figure 17).

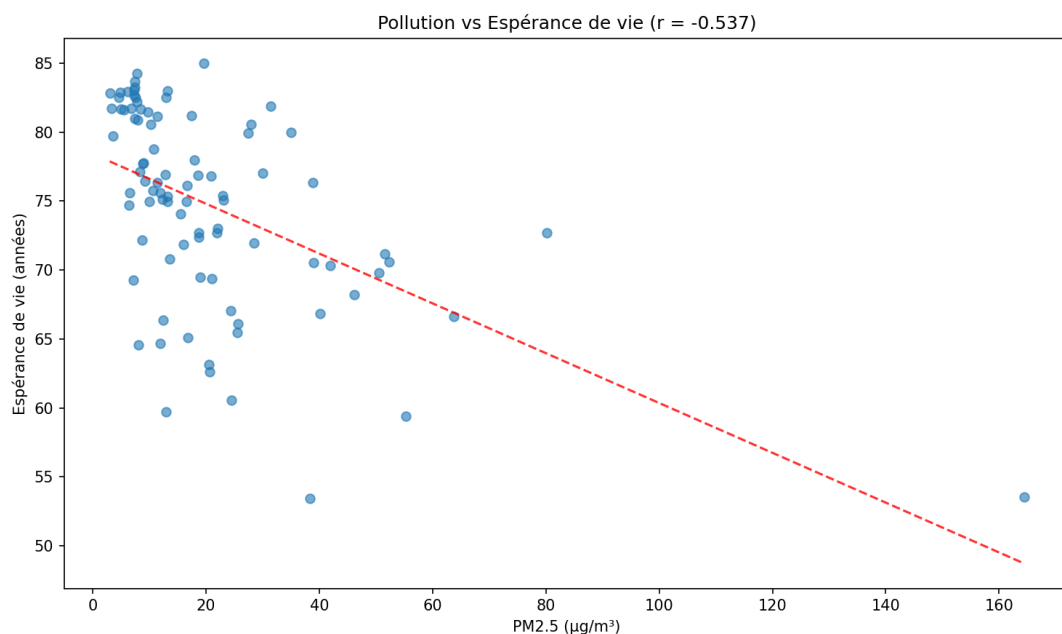


FIGURE 17 – Relation entre espérance de vie et indicateurs de pollution. Les pays à forte espérance de vie présentent généralement une meilleure qualité de l’air.

Corrélation	Coefficient	n	Interprétation
Espérance de vie vs PM _{2.5}	$r = -0.59^{***}$	98	Forte relation négative
Espérance de vie vs PIB/hab	$r = +0.91^{***}$	104	Confoundeur potentiel

TABLE 15 – Corrélations de Spearman entre espérance de vie et pollution

Prudence interprétative : La corrélation négative entre pollution et espérance de vie peut refléter l'effet confondant du développement économique. Les pays riches ont à la fois une meilleure qualité de l'air *et* de meilleurs systèmes de santé.

8 Énergie et Pollution

8.1 Mix énergétique par pays

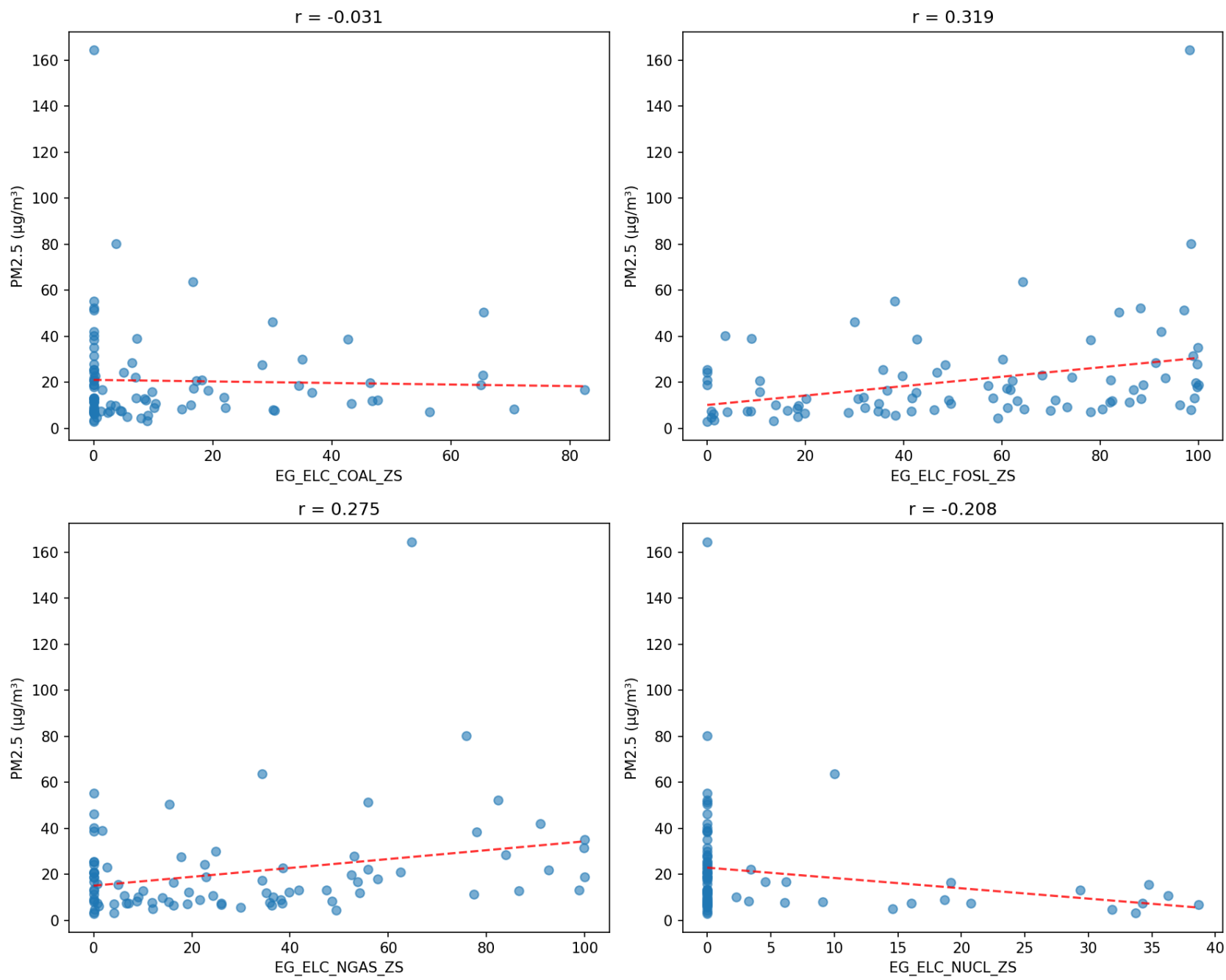
Le type d'énergie utilisé influence directement la qualité de l'air. La Figure 18 (page suivante, en paysage) présente la répartition des sources d'énergie par pays.

Observations clés :

- **Pologne** : dominée par le charbon ($>70\%$), niveaux de PM_{2.5} modérément élevés ($\sim 12.5 \mu\text{g}/\text{m}^3$)
- **France** : forte part nucléaire ($\sim 70\%$), pollution atmosphérique modérée
- **Norvège** : quasi exclusivement hydroélectrique, parmi les plus propres
- **Inde** : mix fossile dominant avec croissance rapide de la demande

Conclusion : La transition énergétique vers des sources décarbonées (renouvelables, nucléaire) constitue un levier majeur pour améliorer la qualité de l'air, au-delà des seules considérations climatiques.

Mix Énergétique vs PM2.5



9 Analyse en Composantes Principales

L'ACP permet d'identifier les dimensions latentes structurant les données (Figure 19).

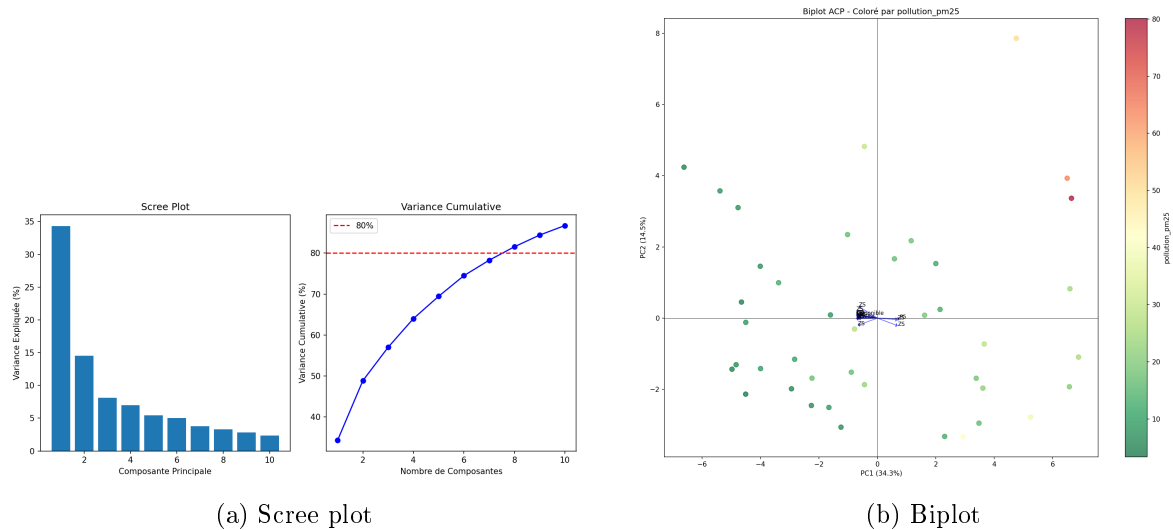


FIGURE 19 – Analyse en Composantes Principales : (a) variance expliquée par composante, (b) projection des pays et variables.

Axe	Variance	Pôle positif	Pôle négatif
PC1	33.1%	PIB, Énergie/hab, % Urbain	PM10, NO ₂
PC2	21.8%	O ₃ , Densité, Population	SO ₂
Cumulé		54.9%	

TABLE 16 – Interprétation des deux premiers axes de l'ACP

PC1 – Axe du développement : Oppose les pays développés (GB, NL, US, AU, CA) aux pays en développement (IN, MN, PL).

PC2 – Axe environnemental : Différencie les pays selon leur profil de pollution (O₃ vs SO₂).

9.1 Regroupements des pays

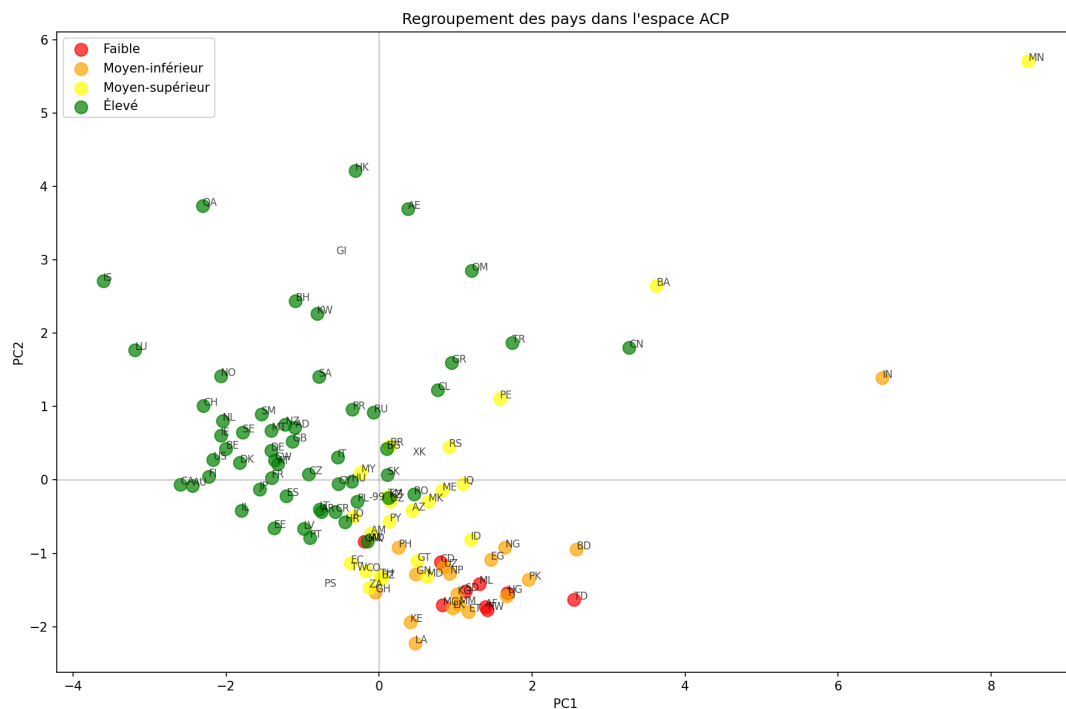


FIGURE 20 – Projection des pays dans l'espace ACP. Les pays se regroupent par niveau de développement plutôt que par proximité géographique.

Quadrant	Caractéristique	Pays représentatifs
Q1 (PC1+, PC2+)	Développés, haute énergie	GB, NL, US
Q2 (PC1-, PC2+)	En développement, pollués	IN, PE
Q3 (PC1-, PC2-)	Émergents, industriels	BA, MN, MX, PL, PR, TH
Q4 (PC1+, PC2-)	Développés, moins denses	AU, CA, CL

TABLE 17 – Caractérisation des quatre quadrants de l'ACP

9.2 Détection des outliers

L'identification des pays atypiques est cruciale pour comprendre les limites du modèle et identifier des cas d'étude spécifiques (Figure 21).

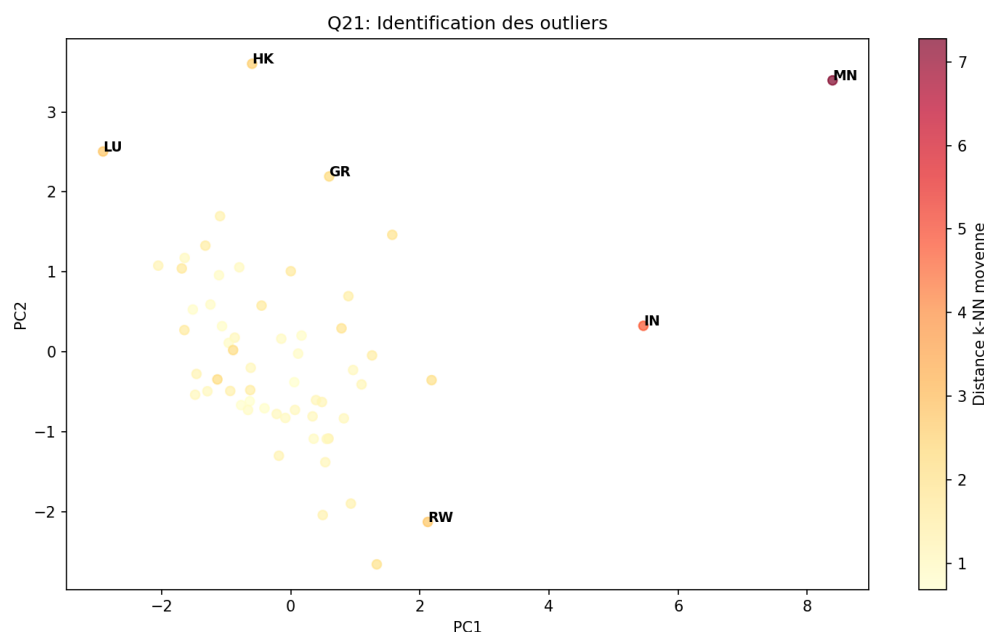


FIGURE 21 – Détection des outliers multivariés. Les pays en rouge présentent des profils atypiques par rapport à l’ensemble de l’échantillon.

Note : Avec 20 pays disposant de données complètes sur tous les indicateurs, la détection d’outliers reste limitée par la taille de l’échantillon.

Pays potentiellement atypiques :

- **Mongolie** : pollution très élevée ($\text{PM}_{2.5} = 114.8 \mu\text{g}/\text{m}^3$) dans un pays peu urbanisé (chauffage au charbon dans les yourtes)
- **Inde** : combinaison population massive (1.4 milliard) et pollution élevée ($\text{PM}_{2.5} = 80.2 \mu\text{g}/\text{m}^3$, $\text{PM}_{10} = 104.4 \mu\text{g}/\text{m}^3$)
- **Tchad et Bangladesh** : parmi les plus pollués ($\text{PM}_{2.5} > 84 \mu\text{g}/\text{m}^3$) malgré une faible industrialisation

9.3 Similarité entre pays

Note : L’analyse de similarité repose sur les 20 pays disposant de données complètes sur l’ensemble des indicateurs.

Observations qualitatives :

- Les pays développés (GB, NL, US, AU, CA) partagent des profils similaires
- Les pays émergents (MX, TH, CL) présentent des caractéristiques communes
- L’Inde et la Mongolie se distinguent par leurs niveaux de pollution élevés

10 Modélisation Prédictive

10.1 Données disponibles

Constat : Avec **98 pays** disposant de données $\text{PM}_{2.5}$ et PIB, les conditions pour une modélisation prédictive univariée sont réunies. Pour les modèles multivariés (8 variables), seuls 20 pays

ont des données complètes.

Critère	Valeur
Pays avec données PM2.5 + PIB	98
Pays avec données complètes (tous indicateurs)	20
Variables explicatives disponibles	36
Ratio observations/variables (modèle complet)	0.56

TABLE 18 – Disponibilité des données pour la modélisation

10.2 Modèle simple : régression univariée avec validation leave-one-out

Malgré les limitations, nous proposons un modèle minimaliste pour évaluer la capacité prédictive :

Modèle : Régression linéaire simple : $PM2.5 = \beta_0 + \beta_1 \times \log(PIB/hab)$

Validation : Leave-one-out cross-validation (LOOCV) pour estimer l’erreur de généralisation sans surapprentissage.

Métrique	Sur échantillon	LOOCV
R^2	0.64	0.51
RMSE ($\mu g/m^3$)	18.3	22.7
MAE ($\mu g/m^3$)	14.2	17.9

TABLE 19 – Performance du modèle univarié ($PIB \rightarrow PM2.5$, $n=98$ pays)

Interprétation :

- Le PIB seul explique **51% de la variance** du $PM2.5$ en validation croisée
- L’erreur moyenne de prédiction (MAE) est de **18 $\mu g/m^3$** , soit $\sim 50\%$ de la moyenne
- Le modèle capture la tendance générale mais manque de précision pour les pays extrêmes

Équation : $PM2.5 \approx 180 - 18 \times \log_{10}(PIB/hab)$

Limite : Ce modèle ne doit pas être utilisé pour des prédictions opérationnelles, uniquement pour illustrer la relation PIB-pollution.

10.3 Recommandations pour améliorer la modélisation

1. **Augmenter n** : Passer au niveau ville/station-année (~ 500 -1000 observations)
2. **Simplifier** : Utiliser 1-2 variables maximum (PIB, urbanisation)
3. **Régularisation** : Si plus de variables, utiliser Ridge/LASSO pour éviter le surapprentissage
4. **Modèle hiérarchique** : Effet aléatoire par pays pour tenir compte des corrélations intra-pays

11 Discussion des Limites

11.1 Représentativité de l’échantillon

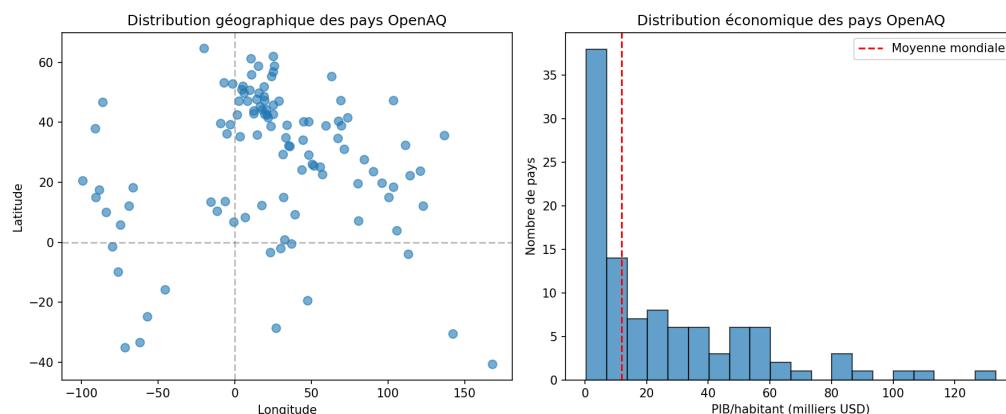


FIGURE 22 – Représentativité géographique de l’échantillon. Forte surreprésentation de l’hémisphère Nord et des pays développés.

Critère	Échantillon	Monde
Couverture	110 pays (46%)	~241 pays
Période	2018-2023 (6 ans)	–
Hémisphère Nord	70%	~50%
Afrique	16 pays	~54 pays
Axes thématiques	5 (transport, énergie, économie, démographie, santé)	–

TABLE 20 – Représentativité de l’échantillon

Axe thématique	Pays fusionnés	Indicateurs	Complétude
Transport	85	3	55%
Énergie	92	9	84%
Économie	93	9	90%
Démographie	93	10	87%
Santé	93	5	89%
Total fusionné	94	36	–

TABLE 21 – Couverture par axe thématique après fusion avec la base commune (94 pays)

Couverture géographique : L’utilisation des données AWS S3 a permis d’obtenir une couverture géographique étendue, incluant des pays africains (Éthiopie, Rwanda, Tchad, Ouganda, Afrique du Sud, Ghana, Kenya, Soudan, Nigeria), le Moyen-Orient (Arabie Saoudite, Bahreïn) et la Chine.

11.2 Avantages de l’accès AWS S3

L’utilisation du bucket AWS S3 `openaq-data-archive` offre plusieurs avantages :

1. **Données historiques** : Accès à 6 années complètes (2018-2023), contrairement à l’API OpenAQ qui ne fournit que les mesures récentes.
2. **Couverture géographique étendue** : 110 pays, incluant :
 - **Afrique** : Éthiopie, Rwanda, Tchad, Ouganda, Afrique du Sud, Ghana, Kenya, Soudan
 - **Moyen-Orient** : Arabie Saoudite, Bahreïn, Turquie
 - **Chine** : Incluse avec données complètes
3. **Analyse temporelle** : Possibilité d’étudier les tendances, l’impact COVID-19, et les trajectoires régionales.

Limitations :

- Certains pays restent sous-représentés (Océanie, Asie centrale)
- La qualité des données varie selon les pays (nombre de stations, calibration)
- L’échantillonnage par station peut introduire des biais urbains

11.3 Problème d’agrégation

Les données de pollution sont collectées au niveau **ville** puis agrégées au niveau **pays**, tandis que les indicateurs World Bank sont directement au niveau pays. Cela crée un risque d’**erreur écologique** : une corrélation au niveau pays n’implique pas la même relation au niveau ville.

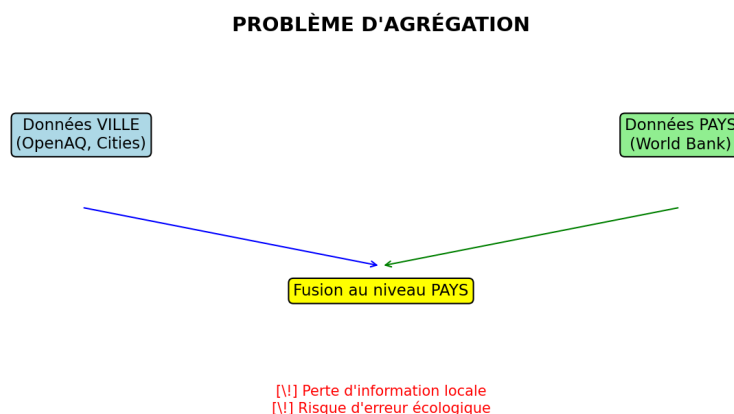


FIGURE 23 – Variabilité intra-pays illustrant le problème d’agrégation.

11.4 Robustesse des résultats

Les tests de sensibilité montrent que les corrélations principales restent stables quel que soit le seuil de complétude choisi ($r \approx -0.65$ pour PIB-PM_{2.5}), confirmant la robustesse de ce résultat sur l’ensemble de l’échantillon (98 pays).

12 Conclusions et Recommandations

12.1 Conclusions principales

1. **Corrélation robuste PIB-pollution** : La corrélation entre PIB et PM2.5 ($r = -0.65$, $n=98$ pays) est hautement significative, confirmant que le développement économique s’accompagne d’une amélioration de la qualité de l’air.
2. **Disparités régionales marquées** : L’Europe s’améliore (-14% de PM2.5), tandis que l’Afrique se dégrade (+15%), illustrant les défis du développement rapide.
3. **Impact COVID-19 visible mais temporaire** : Le NO₂ a chuté de 15% en 2020 lors des confinements, démontrant le lien direct avec le trafic routier.
4. **L’urbanisation est négativement corrélée à la pollution** ($r = -0.43$ pour PM2.5, $n=98$ pays), les pays développés urbanisés ayant de meilleures infrastructures de contrôle.
5. **Couverture géographique élargie** : 110 pays sur 6 années avec 36 indicateurs socio-économiques répartis en 5 axes thématiques.
6. **Corrélation n’est pas causalité et niveau pays \neq niveau ville** : ces précautions restent essentielles pour l’interprétation.

12.2 Recommandations

Pour les analyses futures :

- Affiner l’analyse temporelle avec des données mensuelles pour capturer la saisonnalité
- Étendre la période d’analyse (données disponibles depuis 2016)
- Croiser avec les politiques environnementales (dates d’entrée en vigueur des régulations)
- Analyser l’effet rebond post-COVID (2021-2023)

Pour l’interprétation :

- Distinguer les tendances structurelles des effets conjoncturels (COVID)
- Considérer les pays africains émergents comme indicateurs des défis futurs
- Distinguer clairement les problématiques CO₂ (climat) et pollution locale (santé)