REPUBLIQUE DE CÔTE D'IVOIRE

INTERNATIONAL DATA SCIENCE INSTITUTE

Année académique : 2023 – 2024

RAPPORT DU PROJET TUTORE 2

Etude sur la qualité/pollution de l'air dans de grandes villes à travers le monde

Présenté par :

KESSIE CHRISTELLE ARMANDE

KOUADJA PRINCE K.H.M

SILUE ISMAEL

Master 1 Data Science - Big Data et Intelligence Artificielle

SOMMAIRE

INTRODUCTION

La qualité de l'air est un enjeu majeur de santé publique et d'environnement, particulièrement dans le contexte actuel de l'urbanisation rapide, de l'industrialisation et des changements climatiques. Les polluants atmosphériques ont des effets significatifs sur la santé humaine et les écosystèmes. Analyser ces polluants est crucial pour comprendre leur évolution, les relations qui les lient, les tendances et leurs impacts afin d'élaborer des stratégies efficaces de gestion de la qualité de l'air.

Ce projet se concentre sur les données de pollution de l'air de 2019 à 2023 dans le monde à travers 32 villes du monde, une période marquée par des événements mondiaux tels que la pandémie de COVID-19 et des réformes réglementaires importantes.

L'avènement du Big Data offre des opportunités inédites pour explorer et valoriser ces vastes ensembles de données en utilisant des techniques avancées de traitement et d'analyse des données telles que les statistiques descriptive, la régression linéaire, le clustering…

Ce projet vise donc à approfondir et mettre en pratique sur un thème d'actualité, toutes les techniques et méthodes de structuration et d'analyse des données acquises durant notre première année de master en Data science, big Data et intelligence artificielle.

Dans ce rapport, nous nous altèrerons à faire sortir les différentes étapes et les points clés de notre analyse sur la qualité de l'air dans le monde

I. COMPREHENSION DU PROJET

1. Problématique

La qualité de l'air est un enjeu majeur pour la santé publique et l'environnement. La pollution de l'air est responsable de nombreuses maladies respiratoires et cardiovasculaires, ainsi que de la dégradation des écosystèmes.

Les questions qui se posent sont donc :

- Quelles sont les polluants atmosphériques utilisés pour analyser la qualité de l'air ?
- Quelles dont les tendances et patterns ?
- Comment la qualité de l'air influence t'elle la santé humaine et animal ainsi que le climat.

2. Objectifs

Ce projet a pour objectif :

- Identification des polluants atmosphériques
- Évaluer les tendances temporelles des niveaux de pollution de l'air sur plusieurs années
- Déterminer les relations entre les différents polluants de l'air
- Examiner les relations de dépendance entre les conditions météorologiques et les polluants
- Classifier les pays de nos données en fonction de leur niveau de pollution
- Analyser la relation entre les niveaux de pollution de l'air et les incidences de maladies respiratoires et cardiovasculaires dans les populations

II. NETTOYAGE DES DONNEES

1. Les données

Plusieurs indicateurs permettent d'analyser la qualité de l'air. Pour la réalisation de ce projet, les indicateurs choisis sont notamment les particules fines (PM2.5 et PM10), le dioxyde de soufre (SO2), le dioxyde d'azote (NO2), l'ozone (O3) et le monoxyde de carbone (CO) généralement utilisés pour étudier la qualité de l'air car plus significatifs dû à leur impact important sur la santé et l'environnement. A ceux là s'ajout l'Indice de Qualité de l'Air (IQA) qui est un champs calculé et qui nous permettra de juger de la qualité de l'air.

a) Explication des indicateurs

- **PM2.5** : Ces particules sont particulièrement dangereuses car elles peuvent pénétrer profondément (diamètre < 2 ,5 micromètres) dans les poumons et même entrer dans la circulation sanguine. Elles proviennent principalement de la combustion des combustibles fossiles, des incendies de végétation et des émissions industrielles.
- **PM10** : Bien que moins pénétrantes (diamètre ≤ 10 micromètres) que les PM2.5, elles peuvent néanmoins causer des problèmes respiratoires et cardiovasculaires. Elles sont produites par des sources similaires à celles des PM2.5, ainsi que par les poussières de construction et les émissions des véhicules.

- **SO2**, produit principalement par la combustion des combustibles fossiles contenant du soufre, comme le charbon et le pétrole, peut provoquer des irritations des voies respiratoires et des yeux, ainsi que contribuer à la formation des pluies acides.
- **NO2**, résultant principalement des émissions des véhicules à moteur, des centrales électriques et des procédés industriels, peut causer des irritations des poumons, aggraver les maladies respiratoires et participer à la formation de l'ozone troposphérique et des particules fines.
- **O3**, est un gaz formé par des réactions photochimiques entre les oxydes d'azote (NOx) et les composés organiques volatils (COV) en présence de lumière solaire, il peut irriter les voies respiratoires, réduire la fonction pulmonaire et aggraver les maladies pulmonaires comme l'asthme.
- **CO**, produit principalement par la combustion incomplète de carburants fossiles dans les véhicules, les chaudières, et autres sources. Il peut perturber l'apport en oxygène du corps, entraînant des symptômes comme des maux de tête, des étourdissements, et dans les cas extrêmes, la mort.
- **IQA, (voir Annexe pour la formule de calcul)** est un indice qui cumule quatre polluants réglementés (NO_2, O_3, PM2.5, PM10), Ce nouvel indice prend en compte les effets cumulatifs des différents polluants, permettant de mettre en évidence les zones à exposition multiple.il est basé sur les lignes directrices de l'OMS.

En plus des cinq polluants mentionnés ci-dessus, la température, les variables suivantes sont également prises en compte :

- La température et l'humidié pour analyser l'impact des conditions météorologique sur la qualité de l'air
- Le pays, la ville, la date de prélèvement des données
- La médian comme mesures représentative pour chaque indicateur.

b) Grandes villes choisies

Nous avons choisi 32 pays avec leur capitale dans les 5 continents du monde (Afrique, Asie, Amérique, Océanie, Europe) pour notre étude.

c) Choix des dates

Les données vont de 2019 à 2023.Analyser les données de 2019 à 2023 permet de saisir les tendances récentes et actuelles de la qualité de l'air. Cette période inclut les années les plus récentes, offrant une perspective actuelle sur l'évolution des niveaux de pollution. Cette période couvre des événements mondiaux significatifs, notamment la pandémie de COVID-19 ainsi que de nouvelles réglementations et politiques environnementales visant à réduire la pollution de l'air, qui ont eu des impacts majeurs sur les niveaux de pollution.

2. Inspection des données

L'inspection des données est la première étape cruciale pour comprendre la qualité et la structure des données collectées. Elle permet de détecter les anomalies, les valeurs manquantes et les erreurs de saisie.

Cette étape nous a permis de détecter des valeurs manquantes dans notre jeu de données dû à une absence de prélèvement de certains indicateurs dans certains pays.

	Column	MissingPercentage
1	Country	0.000
2	CO	28.125
3	Humidity	0.000
4	NO2	31.250
5	O3	34.375
6	PM10	28.125
7	PM25	0.000
8	SO2	34.375
9	Temperature	0.000

Tableau 1: Taux des valeurs manquantes

Nos données manquantes sont du types MNAR (Missing Not At Random), pour gérer ses valeurs manquantes, nous avons utilisé la méthode d'imputation multiple.

III. METHODES D'ANALYSE UTISISES

1. Séries temporelles

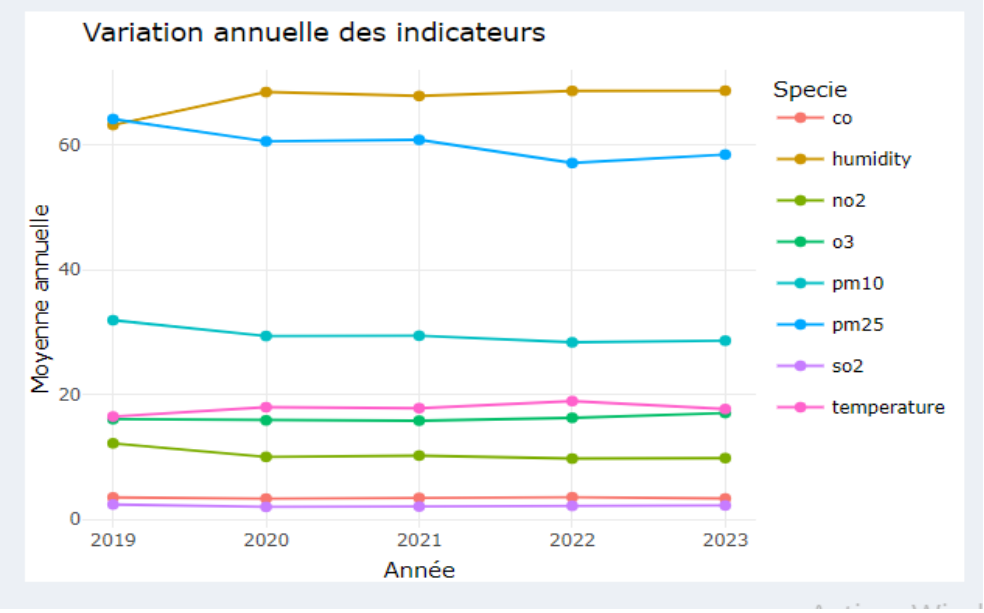


Figure 1 : Séries temporelles

Ces graphiques montrent l'évolution des concentrations des polluants (PM2.5, PM10, SO2, NO2, O3, CO) au fil du temps.

On constate de 2019 à 2020 une baisse rapide des particules fines (PM2,5 et PM10) et le NO2, par contre on observe une augmentation nette de la température et de l'humidité. A cette même période le SO2 et le CO restent constants.

De 2020 à 2023 les indicateurs restent pratiquement constants sauf la température qui rencontre une baisse de 2022 à 2023, le PM2,5 et le NO2 qui eux par contre sont en hausse.

On observe également que le SO2 et le CO ont des valeurs particulièrement pochent au fil du temps.

Quelles sont les relations entre les différents indicateurs ?

2. Régression linéaire

La régression linéaire est une méthode statistique utilisée pour examiner la relation entre une variable dépendante (par exemple, la concentration d'un polluant) et une ou plusieurs variables indépendantes (par exemple, température, humidité, autre polluant). Elle permet de quantifier l'impact des variables indépendantes sur la variable dépendante.

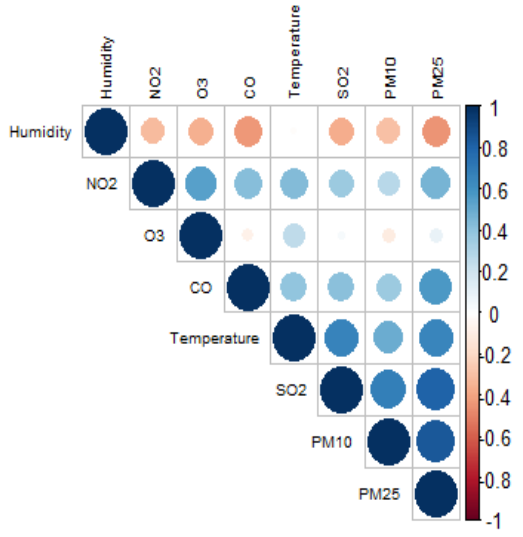


Figure 3: Matrice de corrélation entre les indicateurs

```
> #Modèle pour PM25
> model_PM25 <- with(imputed_data, lm(PM25 ~ Humidity + CO + NO2 + O3 + PM10 + SO2 + Temperature))
> pooled_results_PM25 <- pool(model_PM25)
> summary(pooled_results_PM25)
         term   estimate  std.error  statistic        df      p.value
1 (Intercept) 52.1105727 56.6402268  0.9200276  3.826229 0.4118186906
2    Humidity -0.6671837  0.7291206 -0.9150526  3.680176 0.4161051857
3          CO  1.6821627  2.6649288  0.6312224  2.821767 0.5752999118
4         NO2  0.6013904  1.5416951  0.3900839  3.777803 0.7174616807
5          O3 -0.2822882  1.4240708 -0.1982262  2.795444 0.8563975125
6        PM10  0.8299107  0.1836222  4.5196648 15.536032 0.0003742948
7         SO2 -2.6772111  5.8324124 -0.4590229  3.198947 0.6756257144
8 Temperature  1.4138712  1.2576840  1.1241864  3.751245 0.3276804919
> #Modèle pour SO2
> model_SO2 <- with(imputed_data, lm(SO2 ~ Humidity + CO + NO2 + O3 + PM10 + PM25 + Temperature))
> pooled_results_SO2 <- pool(model_SO2)
> summary(pooled_results_SO2)
         term    estimate  std.error  statistic        df    p.value
1 (Intercept)  5.37207484 2.32868915  2.3069094 14.099610 0.03674486
2    Humidity -0.06610715 0.03271634 -2.0206155  9.445619 0.07256557
3          CO -0.09324176 0.12933635 -0.7209246  6.486629 0.49610788
4         NO2  0.05334409 0.05979619  0.8920985 20.950341 0.38247239
5          O3 -0.09300823 0.04888388 -1.9026360 14.577458 0.07702973
6        PM10  0.02619358 0.02925627  0.8953151  3.222423 0.43242439
7        PM25 -0.01174832 0.02917740 -0.4026512  2.659743 0.71734591
8 Temperature  0.13373465 0.05680096  2.3544434  7.987911 0.04640478
> model_SO2_av <- with(imputed_data, lm(SO2 ~  Temperature))
> pooled_results_SO2_av <- pool(model_SO2_av)
> summary(pooled_results_SO2_av)
         term   estimate  std.error  statistic        df    p.value
1 (Intercept) -0.1726839 0.81587894 -0.2116539 10.758946 0.83633298
2 Temperature  0.1295570 0.04543535  2.8514578  8.499089 0.02015718
```

Figure 2: Résultat des régressions linéaires

De la régression linéaire, on a comme interprétation que le PM10 et le PM2,5 sont liés ainsi que le SO3 et la température.

3. Clustering

Le clustering est une méthode d'apprentissage non supervisé qui regroupe des observations similaires en clusters (groupes). Cette technique a été utilisée pour identifier les classes que nous pouvons construire avec nos données à l'aide de l'algorithme K-means.

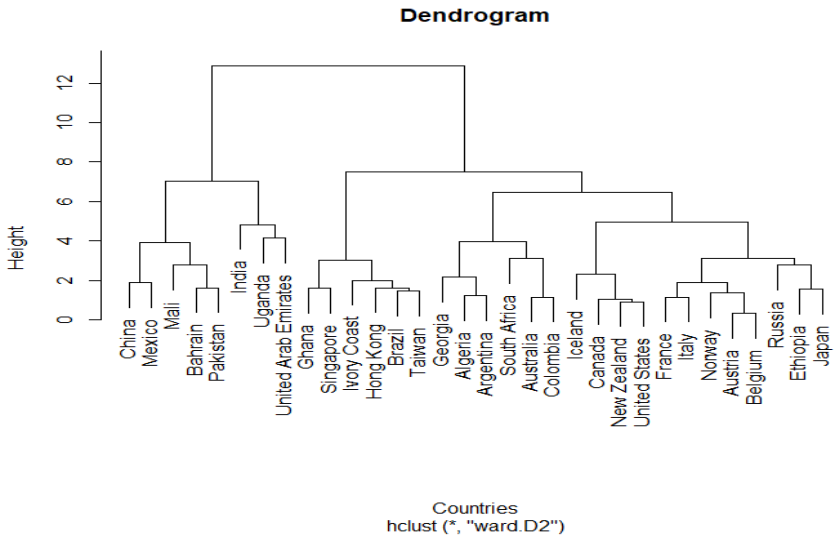


Figure 4 : Dendrogramme

De ce dendrogramme résulte qu'on peut diviser les groupes en 5 classes. Voir Annexe(figure 1 et figure2)

Les caractéristiques de ces différentes classes sont représentées ci-dessous.

```
> catdes_result[["quanti"]]
$`1`
     v.test Mean in category Overall mean sd in category Overall sd     p.value
NO2 -2.046099         7.425739     10.42398       2.937674   4.317174 0.040746607
O3  -3.422985         8.973743     16.72908       2.704833   6.675064 0.000619376

$`2`
       v.test Mean in category Overall mean sd in category Overall sd      p.value
PM10  4.268170       102.096948     31.571374    18.82583696  23.754188 1.970835e-05
PM25  3.242362       138.873434     63.084753    15.61505224  33.603006 1.185435e-03
SO2   2.382543         5.151252      2.362143     0.06358471   1.682906 1.719351e-02
O3   -2.221325         6.414955     16.729077     2.11914673   6.675064 2.632898e-02

$`3`
         v.test Mean in category Overall mean sd in category Overall sd    p.value
Humidity 2.320579         75.36484      67.09344       8.588738   10.50129 0.02030957

$`4`
                v.test Mean in category Overall mean sd in category Overall sd      p.value
NO2         -2.004941         8.1181252    10.423979      4.3504269   4.317174 0.0449693063
PM10        -2.637738        14.8796040    31.571374      3.9613999  23.754188 0.0083460953
PM25        -3.330506        33.2708529    63.084753      9.3283780  33.603006 0.0008668842
SO2         -3.380427         0.8466216     2.362143      0.4260117   1.682906 0.0007237331
Temperature -3.621550        11.7474526    18.064204      3.3476739   6.547400 0.0002928437
CO          -3.865969         0.5478561     3.511407      0.8457498   2.877554 0.0001106488

$`5`
                v.test Mean in category Overall mean sd in category Overall sd      p.value
SO2          3.107895         4.317637     2.362143       1.285394   1.682906 1.884252e-03
O3           3.051683        24.345046    16.729077       2.364996   6.675064 2.275619e-03
PM25         2.998800       100.759981    63.084753      22.047816  33.603006 2.710448e-03
NO2          2.483252        14.432190    10.423979       1.638683   4.317174 1.301888e-02
Temperature  1.977157        22.904147    18.064204       6.129668   6.547400 4.802388e-02
Humidity    -4.021674        51.303534    67.093442       7.294548  10.501286 5.778588e-05
```

Figure 5 : Caractéristiques des différentes classes

IV. IMPACT DES POLUANTS

1. Sur la santé

- Humaine

Maladies respiratoires et cardiovasculaires accrues Une exposition prolongée aux particules fines (PM2.5 et PM10), au dioxyde de soufre (SO2), au dioxyde d'azote (NO2), à l'ozone (O3) et au monoxyde de carbone (CO) augmente le risque de développer des maladies respiratoires chroniques comme l'asthme, la bronchite chronique et les infections pulmonaires. De plus, ces polluants peuvent aggraver les maladies cardiovasculaires préexistantes et augmenter les risques de crises cardiaques et d'accidents vasculaires cérébraux.

- Animale

P**roblèmes respiratoires et dermatologiques** : Les animaux domestiques et sauvages peuvent subir des effets similaires à ceux des humains en cas d'exposition prolongée à des niveaux élevés de pollution atmosphérique. Cela peut se manifester par des problèmes respiratoires chroniques et des maladies dermatologiques dues à l'inhalation de particules nocives et à l'exposition aux gaz polluants.

Impact sur la reproduction et la fertilité : Certaines études suggèrent que la pollution de l'air peut également affecter la reproduction et la fertilité chez les animaux, en réduisant la qualité du sperme et en augmentant les complications pendant la gestation.

2. Sur l'environnement

- Végétation

Réduction de la biodiversité végétale : Les polluants atmosphériques peuvent endommager la végétation en affectant la photosynthèse, en augmentant la sensibilité aux maladies et en réduisant la résilience des plantes face au stress environnemental. Cela peut entraîner une diminution de la diversité végétale et une altération des écosystèmes locaux.

- Climat

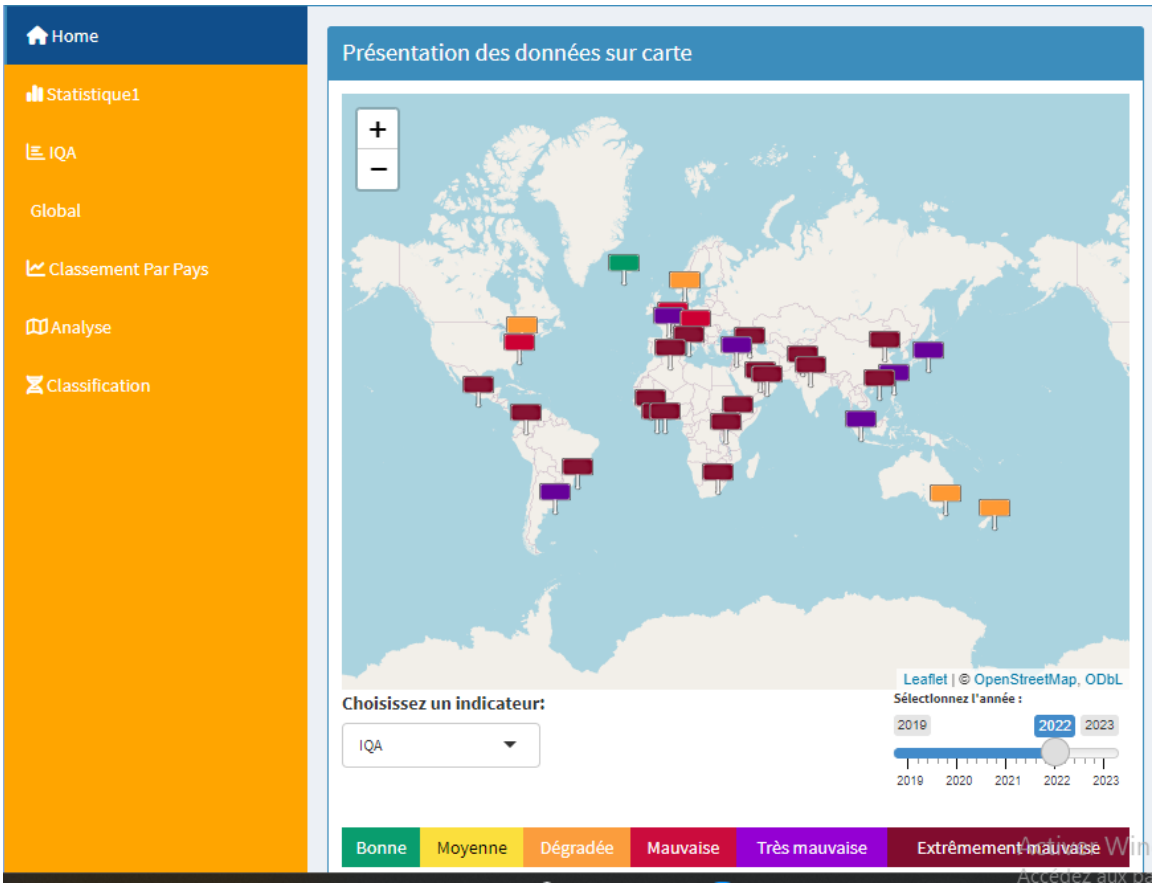
Effets sur le climat et le réchauffement planétaire : Certains polluants comme les gaz à effet de serre contribuent au réchauffement climatique en piégeant la chaleur dans l'atmosphère (effet de serre). En outre, la formation de smog et d'ozone troposphérique peut modifier les régimes climatiques locaux et régionaux, affectant la météorologie et les précipitations.

Quelques illustrations

En Chine, où la pollution de l'air est souvent sévère dans les grandes villes comme Pékin, les taux élevés de particules fines ont été associés à une augmentation significative des admissions à l'hôpital pour des problèmes respiratoires aigus et des maladies cardiovasculaires.

Les incendies de forêt massifs de 2019 à 2020 qui ont ravagé l'Australie ont produit d'énormes quantités de fumée contenant des particules fines et des polluants toxiques. Ces conditions ont provoqué une détérioration significative de la qualité de l'air dans les zones touchées, augmentant les risques pour la santé respiratoire des populations locales et exacerbant les conditions préexistantes comme l'asthme et les maladies pulmonaires

V. VISUALISATION DES RESULTATS



TABLEAUX ET FIGURES

CONCLUSION

L'analyse de la qualité de l'air de 2019 à 2023 a permis de mettre en lumière des tendances cruciales et des anomalies significatives influencées par des événements mondiaux comme la pandémie de COVID-19 et les changements réglementaires. Les méthodes utilisées, telles que les séries temporelles, la régression linéaire et le clustering, ont fourni des insights précieux sur les variations des niveaux de polluants et leurs impacts sur la santé et l'environnement.

Cependant, ce projet n'a pas été sans ses défis. L'un des principaux obstacles rencontrés a été la qualité et la disponibilité des données, notamment dans les régions où les systèmes de surveillance de la qualité de l'air ne sont pas aussi développés ou accessibles. Cela a parfois limité la précision des analyses et la capacité à faire des comparaisons globales cohérentes.

En termes de perspectives, il est crucial d'améliorer la collecte et la normalisation des données sur la qualité de l'air à l'échelle mondiale, afin de mieux comprendre les tendances à long terme et d'évaluer l'efficacité des politiques de réduction de la pollution. De plus, l'intégration de techniques d'intelligence artificielle et de modélisation prédictive pourrait ouvrir de nouvelles possibilités pour anticiper et atténuer les effets de la pollution atmosphérique sur la santé publique et les écosystèmes.

ANNEXE

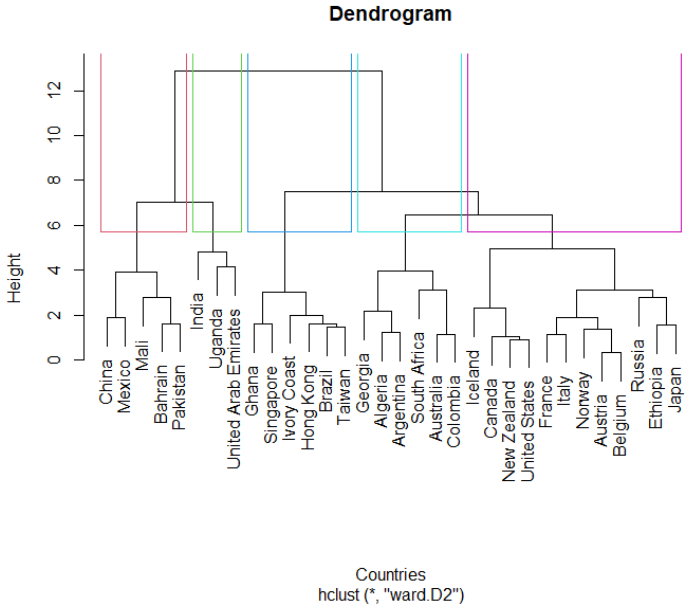


Figure 1

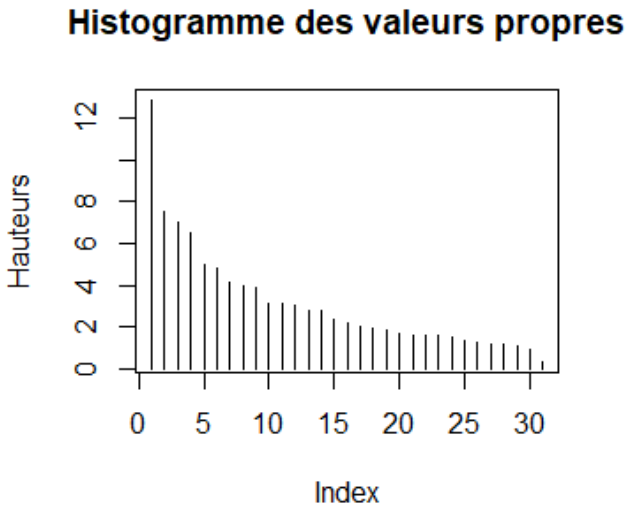


Figure 2

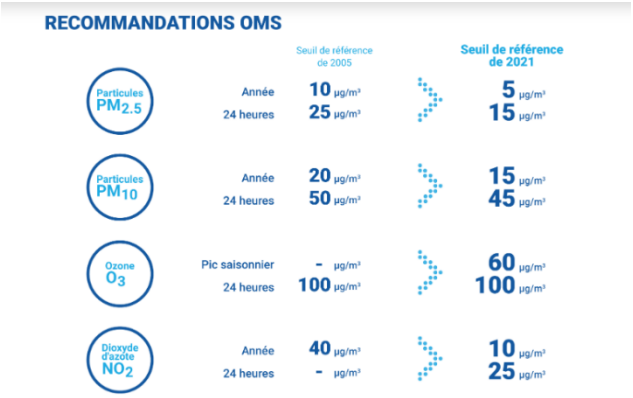


Figure 3

Règles de calcul du IQA : Un sous-indice Ip (moyenne annuelle) est calculé pour chaque polluant « p » (NO_2, O_3, PM10, PM2.5), en se référant aux Lignes Directrices OMS (LD).

Pour une concentration [P] du polluant « p » : **Ip = [P]/LDp**

L'IQA est ensuite calculé de la manière suivante :

IQA = max (IPM10 ; IPM2.5) + INO_2 + IO_3

Remarque :LDp est la ligne directrice de l'OMS pour le polluant p.(figure3)