Air Quality and Pollution Assessment Data Analysis

Sandra ROCHE

March 31, 2025

1 Introduction

World Health Organization (WHO) defines air pollution as contamination of the indoor or outdoor environment by any chemical, physical, or biological agent that modifies the natural characteristics of the atmosphere. Household combustion devices, motor vehicles, industrial facilities, and forest fires are common sources of air pollution. Pollutants of major public health concern include particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulfur dioxide.

Indoor and outdoor air pollution causes respiratory and other diseases and is one of the major sources of morbidity and mortality. WHO data show that almost all of the global population (99%) breathe air that exceeds WHO guideline limits and contains high levels of pollutants, with low- and middle-income countries suffering from the highest exposures.

The combined or joint effects of ambient (outdoor) and household air pollution exposure cause about 7 million premature deaths every year, increased mortality from stroke, heart disease, chronic obstructive pulmonary disease, lung cancer and acute respiratory infections.

Does this dataset confirm WHO information? Does it include all pollutants highlighted by the WHO? Are pollutants values closed, above or under WHO recommended threshold?

2 Methodology

2.1 Dataset

The dataset contains 5000 entries corresponding to air quality assessment across various regions in the world. It includes critical parameters and demographic factors that may impact pollution level.

The dataset is providing by the Kaggle website https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment which contains a lot of datasets on multiple thematic.

2.2 Variables

Variables including in the studied dataset are explained in the Table 1.

Unit for PM2.5, PM10 and NO2 is µg/m^3 and the one for SO2 and CO is mg/m^3 meaning that the quantity is not considered at the same level.

Variables	Unit	Explanation
Temperature	°C	Average temperature of the region
Humidity	%	Relative humidity recorded in the region
PM2.5	µg/m^3	Fine particulate matter level
PM10	µg/m^3	Coarse particulate matter level
NO2	µg/m^3	Nitrogen dioxide
SO2	mg/m^3	Sulfur dioxide
CO	mg/m^3	Carbon monoxide
Proximity to industrial areas	km	Distance to the nearest industrial zone
Population	People/km^2	Number of people per square kilometer in the region

Table 1: Variables explanation.

Air quality is considered:
- "good" when the air is clean with low pollution levels,
- "moderate" when air quality is acceptable but with some pollutants present,
- "poor" when there are noticeable pollutions that may cause issues for sensitive people,
- "hazardous" when air is highly polluted leading to serious health risks for the population.

2.3 WHO guidelines

The recommended values for each pollutant highlighted by the WHO are summarized in the Table 2.

Pollutant	Guideline value	Averaging time
PM2.5	15 µg/m^3	24-hour
PM10	45 µg/m^3	24-hour
NO2	25 µg/m^3	24-hour
SO2	40 mg/m^3	24-hour
CO	4 mg/m^3	24-hour
Formaldehyde	0.1 mg/m^3	30-minute
Polycyclic aromatic hydrocarbons	8,7x10-5 per ng/m^3	Not defined
Radon	100 Bq/m^3	Not defined
Lead	0.5 µg/m^3	Annual

Table 2: WHO guideline value for each pollutant.

WHO website for Table 2 references:
https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants

2.4 Method

First, dataset is imported from Kaggle. At the beginning, data are analyzed from a quality point of view including search of missing values, duplicate data and typographic issue. Then, some variables distributions are observed.

Next, an analysis is performed to explore the dataset and try to find a link between air quality and pollutants. Moreover, data are compared to the WHO guideline limits.

Finally, limitations and discussion are proposed at the end of the report.

3 Results

The dataset is clean since no missing value, no duplicated data and no typo issue (N=5000).

Distribution of the data linked to air quality is analyzed and visualized in the Figure 1. The air quality is qualified as "good" for 30% of the data, "moderate" for 30%, "poor" for 20% and "hazardous" for 10%, meaning that majority of the air quality is acceptable in this dataset (70%).

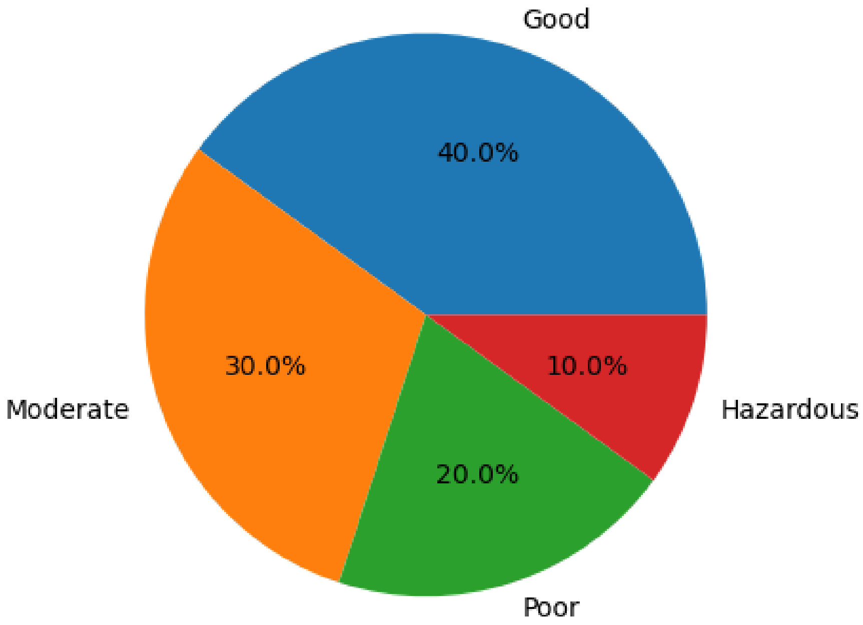

Figure 1: Data distribibion per air quality category.

Next, global data distribution is analyzed for each pollutant and result can be visualized in the Figure 2. Data distribution for PM2.5 and PM10 pollutants is more dispersed with outlier data compared to NO2, SO2 and CO pollutants data.

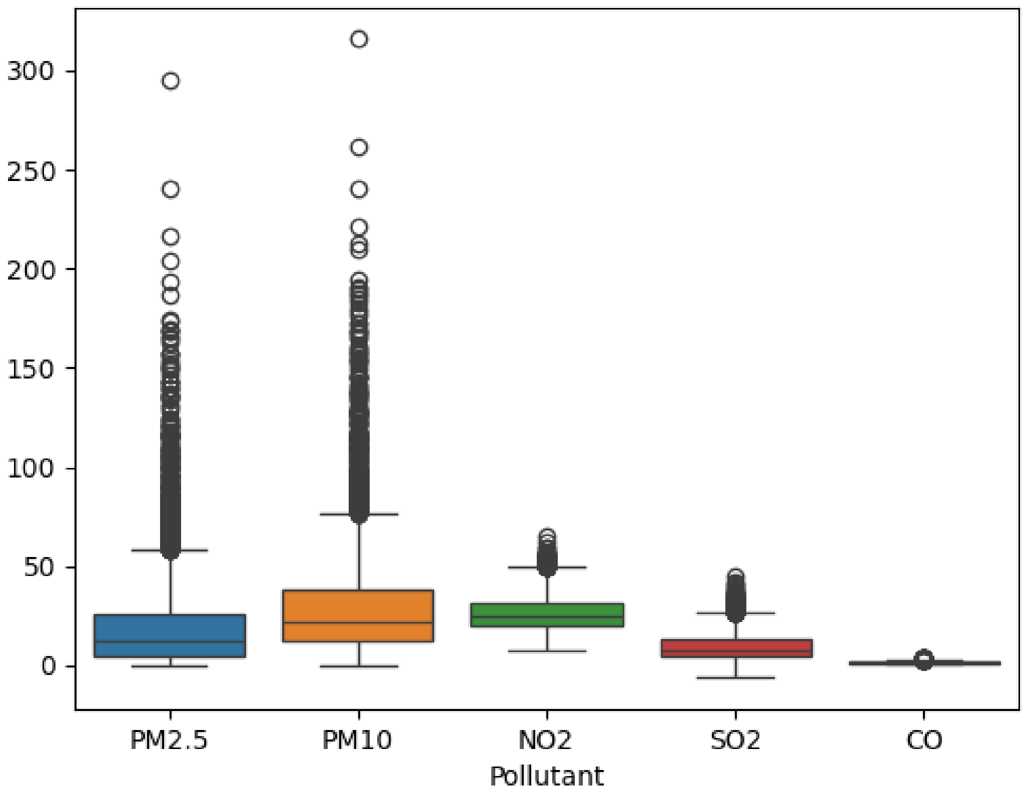

Figure 2: Data distribution for each pollutant containing in the dataset.

Then, data averages are compared to the WHO threshold for each pollutant including in this dataset and results can be visualized in the Figure 3. In details:

- The average for the PM2.5 pollutant is a little above the WHO threshold (guideline value at 15 µg/m3 for 24h and data average at 20)
- The average for the PM10 pollutant is under the WHO threshold (guideline value at 45 µg/m3 for 24h and data average at 30)
- The average for the NO2 is a little above the WHO threshold (guideline value at 25 µg/m3 for 24h and data average at 26)
- The average for the SO2 pollutant is under the WHO threshold (guideline value at 40 mg/m3 for 24h and data average at 10)
- The average for the CO pollutant is under the WHO threshold (guideline value at 4 mg/m3 for 24h and data average at 1.5)

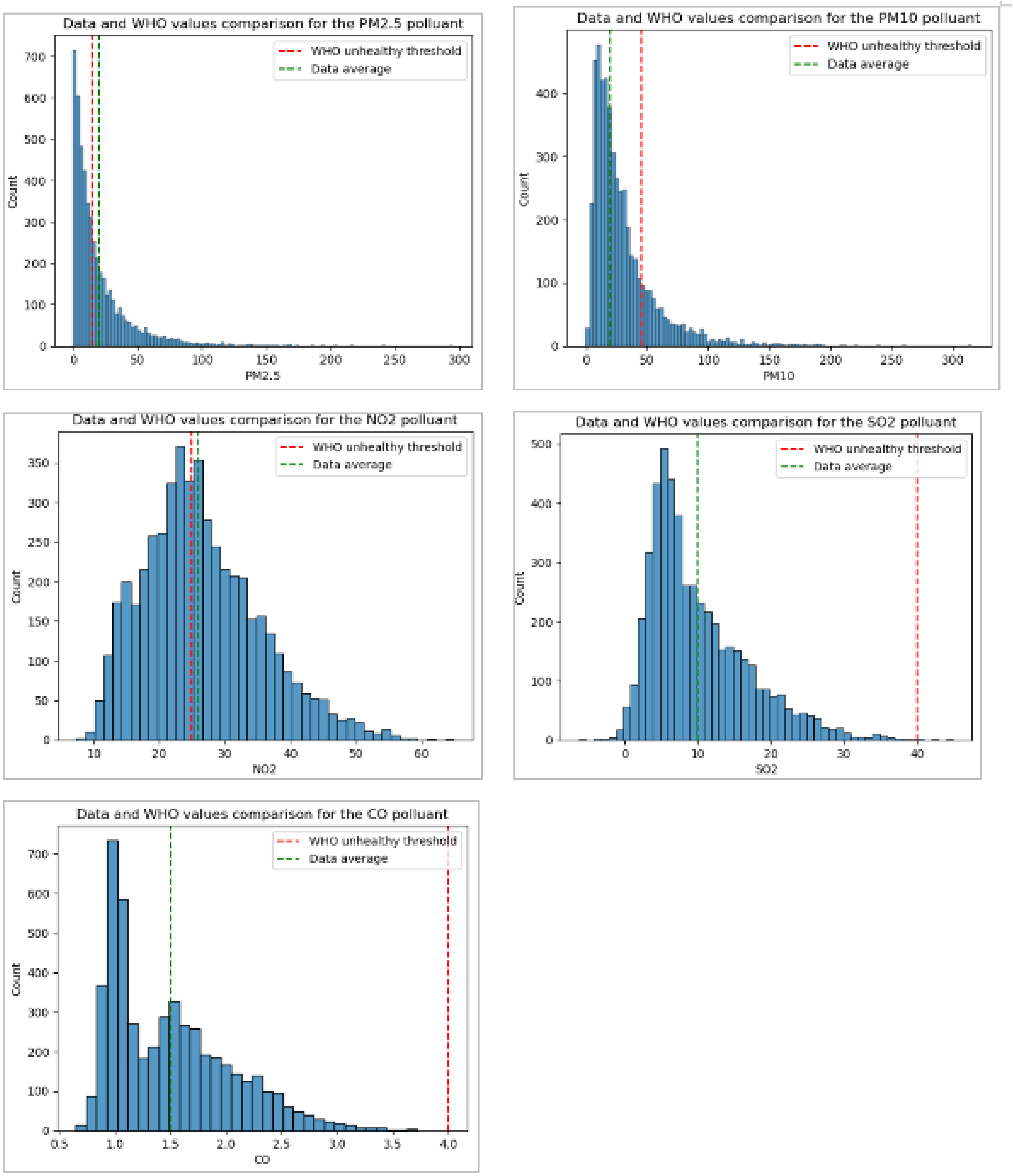

Figure 3: Pollutants data compared to WHO threshold. The WHO threshold is represented by the broken red line and average data by the broken green line.

Another analysis is performed on population density and proximity to industrial areas are analyzed and results can be visualized in the Figure 4. To summarize, the air quality is acceptable when the population density is low and industrial areas are faraway which is not surprising.

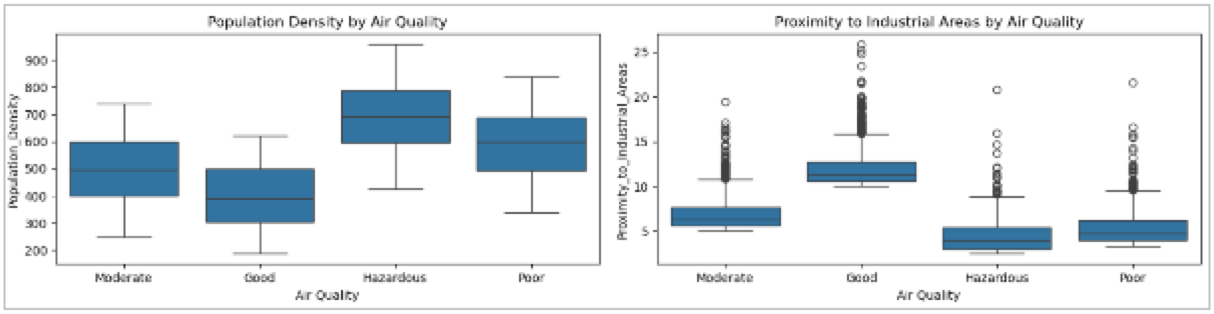

Figure 4: Impact of population density and proximity to industrial areas on air quality.

Next analysis is performed on temperature and humidity and results can be visualized in the Figure 5. These 2 parameters seem to have a low impact on air quality even if the lowest temperature and humidity are found in the “good” category and highest temperature and humidity in the “hazardous” category as shown in the Figure 5.

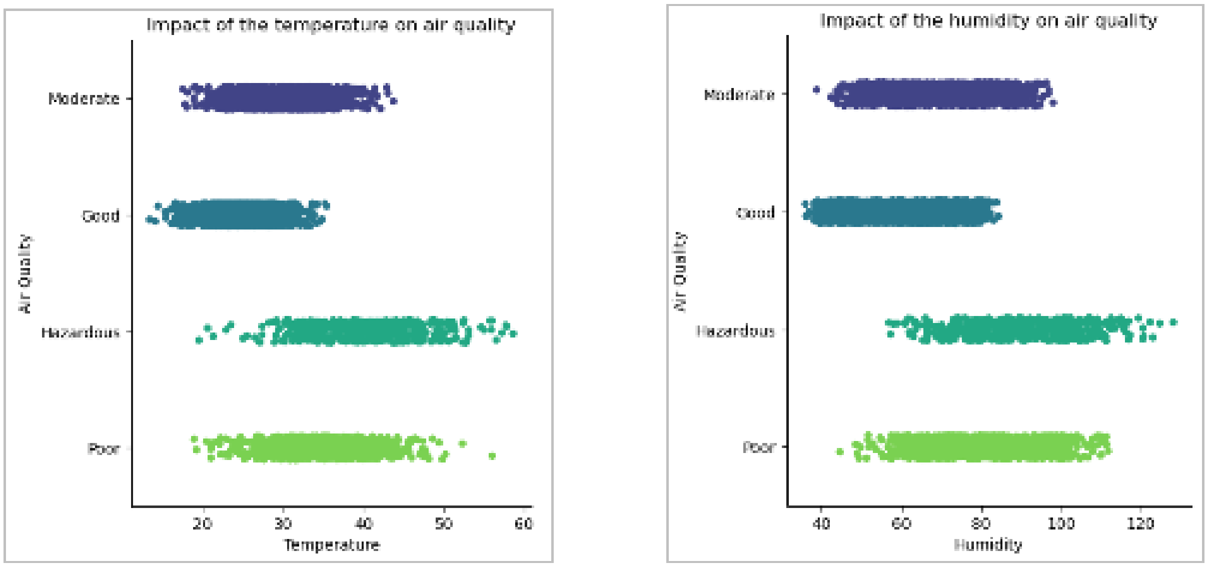

Figure 5: Temperature and humidity distribution data per air quality category.

Last analysis is performed on CO, NO2 and SO2 pollutants which seem to be linked (correlation coefficient at 0.7 between each variable, data not shown). As shown in the Figure 6, these parameters seem to be linear linked and to have a direct impact on the air quality. Indeed, acceptable air quality corresponding to the lowest CO, NO2 and SO2 levels and highest levels to bad air quality.

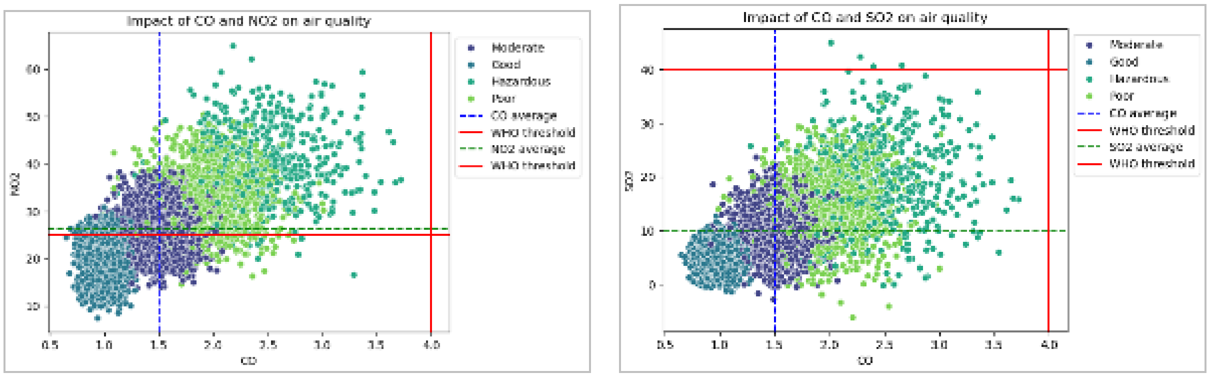

Figure 6: CO, NO2 and SO2 distribution data per air quality category.

4 Discussion

Regarding the WHO list of pollutants, the studied dataset in the present report misses some of them such as:

- Formaldehyde
- Polycyclic aromatic hydrocarbons
- Radon
- Lead

Moreover, no indication on the regions of the world where values were measured is included in this dataset. Thus, it's not possible to confirm WHO information such as low- and middle-income countries suffering from the highest exposures.

To conclude, air quality is closely linked to the earth's climate and ecosystems globally. Many of the drivers of air pollution (i.e. combustion of fossil fuels) are also sources of greenhouse gas emissions. Policies to reduce air pollution, therefore, offer a win-win strategy for both climate and health, lowering the burden of disease attributable to air pollution, as well as contributing to the near- and long-term mitigation of climate change.