TAYLOR'S UNIVERSITY

Wisdom • Integrity • Excellence

APPLICATION OF OPEN-AIR MODEL (R-PACKAGE) TO ANALYZE AIR POLLUTION DATA IN DELHI, INDIA

Faysal Hossain[1], Kang Jia Cheng[1], Atul Dsouza Victor Francis[1], Rainaf Akif[1], Abdelkader Youssef Karim Ahmed[1]

[1] Computer Science, School of Liberal Arts and Sciences, Taylor's University Lakeside Campus, Subang Jaya, Selangor

*Corresponding authors' emails: fsfoysal15@gmail.com, jckangsfw@gmail.com, atuldsouza33@gmail.com, rainafakif125@gmail.com, youssefabdelkader7@gmail.com

ABSTRACT

Anthropogenic activities such as rapid industrialization development, unrestrained transportation, excessive use of fossil fuels, and rising global power consumption contribute to the worsening air quality globally. Thousands of industries, gas stations, gasoline pumps, and automobiles release toxic compounds that can impact public health. A comprehensive database for air pollution levels from November 25, 2020, to January 24, 2023, was examined in the sparsely populated city of Delhi, India. This research explores the various ways the open-source openair model can be used to analyze data on urban air quality. Several monitoring stations across different parts of Delhi gathered data for each pollutant, which is later processed and imported by the openair model in the Comma-separated value (CSV) format. The input data consists of date-time and pollutant. The TheilSen, timeVariation, and scatterPlot functions for trend analysis, temporal variations, and linear correlation analysis, respectively, are used for this research analysis. Discussion ensued about the results of these functions. In conclusion, the openair model can analyze long-time air quality data.

KEYWORDS

- Statistical software R,
- Urban air quality,
- Openair,
- Open source,
- Air pollution,
- Delhi

1. INTRODUCTION

1.1 Background

Delhi is among one of the fastest-growing economic centers of South Asia, and in 2022, it witnessed a total of 79.18 hundred thousand motor vehicles on the road. This number poses significant environmental and health issues in Delhi. Carbon monoxide (CO), nitrogen oxides (NO_x), and particulate matter (PM) are the primary pollutants released from the transportation sector. Additionally, secondary air pollutants like ozone (O_3) have seen a striking increase in levels.

India is one of the most air-polluted countries in the world. The World Health Organization (WHO) estimates that 99% of people worldwide continue to breathe unhealthy air that exceeds WHO air quality standards. For instance, pollutants with diameters equal to or less than ten micrometers (PM_{10}) or 2.5 micrometers ($PM_{2.5}$) are capable of penetrating deep into the lungs and entering the bloodstream, causing cardiovascular, cerebrovascular (stroke), and respiratory effects. Moreover, these pollutants are primarily produced by human activities related to the combustion of fossil fuels. Additionally, NO_2 can cause respiratory conditions, particularly asthma, which can cause hospital admissions, ER visits, and respiratory symptoms like coughing, wheezing, or difficulty breathing.

The air tends to change from time to time. Correspondingly, research on air quality typically only describes a phenomenon at the time of conducting (in terms of location and study period), leading to limited data for the present air quality. Data on good air quality are measured down to the tiniest scale; for instance, the PM_{10} hourly concentration for Delhi, leading to a more accurate interpretation of the data. This small measurement scale, however, tends to increase the amount of data and lengthen the analysis process. As a result, this analysis requires a tool or application to examine the massive amounts of air quality data, but even Microsoft Excel has its limitations in this regard.

Big data analysis tools are available using computer programming languages, which can more accurately predict spatial and temporal pollutants and interpret ambient air quality. Examples of computer programming languages for scientific computation, including air quality modeling, include FORTRAN, C++, and R. R is a language dedicated to the statistical calculation that provides more in-depth analysis. Also, R has developed an openair model or openair package to analyze air quality data; many functions are designed specifically for performing air quality monitoring analysis.

1.2 Purpose

Many of Delhi's air quality monitoring stations collect large amounts of data on the city's air quality. To better understand the air quality issues, more in-depth analysis is critical to the pollutant concentrations and meteorological data. This study will look at different applications

of the open-source openair model, including trend analysis, temporal variation (pollutant fluctuations over time), and correlations of meteorological factors.

2. DATA ANALYSIS

An openair model can monitor air quality data while handling atmospheric conditions, such as wind speed and direction. An openair model can assess model performance and analyze pollutant characteristics, source emissions, and trend estimates. The advantage of the openair model is the ability to manipulate or interpolate data, analyze statistical data, and produce and display high-quality graphics. Trend estimates analysis is the main topic of this study.

To ensure the package's availability, the Openair model should be downloaded first in R software on the official website https://cran.r-project.org/bin/windows/base/. Once downloaded, the openair model package can be activated in R software by typing "library (openair)." The air quality data to be analyzed can be input from computer files or imported from monitoring stations across Delhi. After that, the datasets are processed and imported in the openair model (software R) and are presented in comma-separated value (CSV) format, one of the Microsoft Excel extension files.

The Openair model is an air quality modeling that has a function to stimulate the mathematical formula into the computer program. This model is a tool for statistically analyzing semi-empirical mathematical relationships between air pollutant concentration and other factors that may affect it. Some fundamental analyses in the openair model include linear regression, decision-making with *p-values*, and coefficient of determination.

2.1 Linear Regression

Linear regression is a statistical method for developing a relationship model between the dependent and independent variables. The model's coefficient represents the assumed parameter value for the actual condition. However, the regression model coefficients are an average value that may occur in the variable Y (dependent variable) corresponding to a value of X (independent variable). There are two types of regression coefficients: intercept (point intersection with the Y axis) and slope (line gradient). In statistics, the slope value is the average increase or decrease in variable Y for each unit increase in variable X.

2.2 Decision-making with the *p-value*

Statistics use sample data to infer the overall condition of the population. As a result, the potential for error in making a population decision is also relatively high. Nonetheless, the statistical concept seeks to minimize error as much as possible. A test criterion is required to determine whether H_0 is rejected or accepted. The *p-value* is the most used test criterion in a computer program.

P-value provides two pieces of information at once: the reason for rejecting the null hypothesis (H_0) and the probability of occurrence mentioned in H_0 (assuming H_0 is accurate). The definition of *p-value* is the slimmest level of meaning at which the result of a statistical test can still be meaningful. Furthermore, it can also be interpreted as the magnitude of the possibility of making a mistake when deciding whether to reject H_0. Generally, the *p-value* is compared to the significance level (α).

2.3 Coefficient of Determination

The coefficient of determination (R^2) is the amount of diversity (information) in the Y variable that the regression model can provide. R^2 has a value ranging from 0 to 1. When the value of R^2 is multiplied by 100%, the percentage of diversity (information) within Y (dependent variable) that is influenced by X (independent variable) is calculated. The higher the R^2 value, the better the regression model.

3. RESULTS AND DISCUSSION

Numerous benefits exist for the open-air model, including:

1) It is available for download from its official website as free software.
2) It can run on many operating systems, including Windows, Mac OS, and Linux.
3) It was created with effective and dependable air quality data analysis.
4) The system provides statistical and data analysis capabilities.
5) Impressive output of graphics.

Despite its many advantages, R does have some disadvantages, such as:

1) R can be challenging to learn because it requires precise typing, much like the programming languages C++ and FORTRAN.
2) The system only has a screen that resembles a DOS prompt where users enter commands; there is no graphical user interface. Compared to the interactive computing experience of today, it seems incredibly antiquated.
3) Unlike other software, there is no option for help or support. R does, however, offer online support to users who have used it successfully.

Many statistical functions are available in the open-air model air quality analysis, as shown in Table 1.

Table 1 Main *openair* analysis functions

No.	*Function*	Purpose

1	*calcFno2*	Estimate primary NO_2 emissions ratio from monitoring data
2	*calendarPlot*	Calendar-type view of mean values
3	*conditionalEval*	Extensions to conditionalQuantile
4	*conditionalQuantile*	Quantile comparisons for model evaluation
5	*kernelExceed*	bivariate kernel density estimates for exceedance statistics
6	*linearRelation*	explore linear relationships between variables in time
7	*TheilSen*	Calculate Theil-Sen slope estimates and uncertainties
8	*modStats*	Calculate a range of model evaluation statistics
9	*percentileRose*	Percentiles by wind direction
10	*polarAnnulus*	Polar annulus plot for temporal variations by wind direction
11	*polarCluster*	Cluster analysis of bi-variate polar plots for feature extraction
12	*polarFreq*	Alternative to wind rose/pollution rose
13	*polarPlot*	Bi-variate polar plot
14	*pollutionRose*	Pollution rose
15	*scatterPlot*	Traditional scatter plots with enhanced options
16	*smoothTrend*	Smooth trend estimates
17	*summaryPlot*	summary view of a data frame
18	*TaylorDiagram*	model evaluation plot
19	*timePlot*	Time-series plotting
20	*timeProp*	Time-series plotting with categories as stacked bar chart
21	*timeVariation*	Diurnal, day of week and monthly variations
22	*trajCluster*	HYSPLIT back trajectory cluster analysis
23	*trajPlot*	HYSPLIT back trajectory plots - points of lines
24	*trajLevel*	HYSPLIT back trajectory plots - binned or smoothed
25	*trendLevel*	flexible level plots or ‘heat maps’
26	*windRose*	Traditional wind rose

Based on the outcomes of air quality studies in Delhi that used openair model application, this study only discusses three functions of the openair model: TheilSen for trend analysis, timeVariation for temporal variations, and scatterPlot for linear correlation analysis.

3.1 TheilSen Function

This function helps understand concentration changes (trend) over time and for comparison with an air quality standard. It produces a slope value as the trend percentage and concentration value in the unit per period change. The positive slope of the linear regression line represents the value of the increasing trend, and the negative slope represents the value of the decreasing trend.

TheilSen Function analyzes pollutant concentration trends using linear regression and the Man-Kendall method, with a 95% confidence interval and a 5% significance level (α). The confidence interval and significance level will vary to account for the study's constraints. Whether the trend concentration change tends to be significant or not over the time that occurs is also observed. In this instance, an alternative hypothesis (H_a) formulates correspondingly as the trend changes by pollutant concentration.

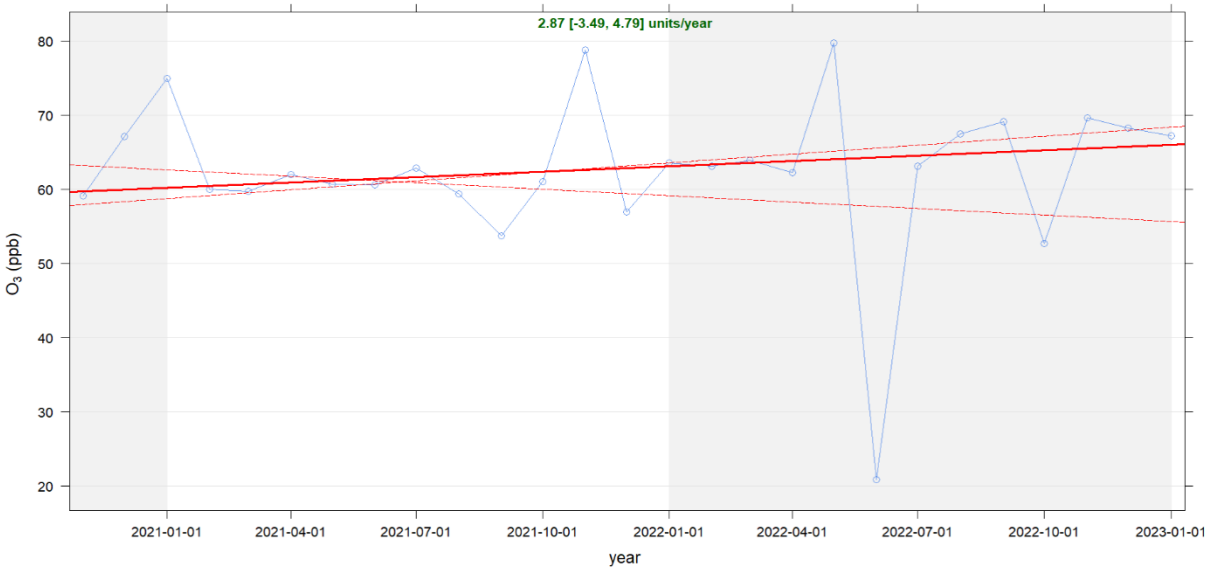

Figure 1 Trend of Ozone (O_3) concentration to ambient air of Delhi from 2021 to 2023

Trends in ozone at Delhi. The plot shows the deseasonalized monthly mean concentrations of O_3. The solid red line shows the trend estimate, and the dashed red lines show the 95% confidence intervals for the trend based on resampling methods. The overall trend (as shown at the top-left) is 2.87 (ppb) per year; the 95 % confidence intervals in the slope from -3.49 to 4.79 ppb/year. In this situation, Delhi must take additional steps to reduce O_3 emissions.

3.2 Time Variation Function

When using the timeVariation function, it is possible to visualize the temporal variation of pollutant concentration in a line graph. The output of this function is an image made up of four-line graphs based on different time scales, including hourly, daily, hourly, and monthly time. Based on a 95% confidence interval, this function analyzes data. The advantage of this function

is that it can plot more than one pollutant simultaneously.

Knowing a pollutant's temporal variation over time allows one to predict when its concentration will be at its lowest or highest, on which days of the week, or in which months throughout the years.

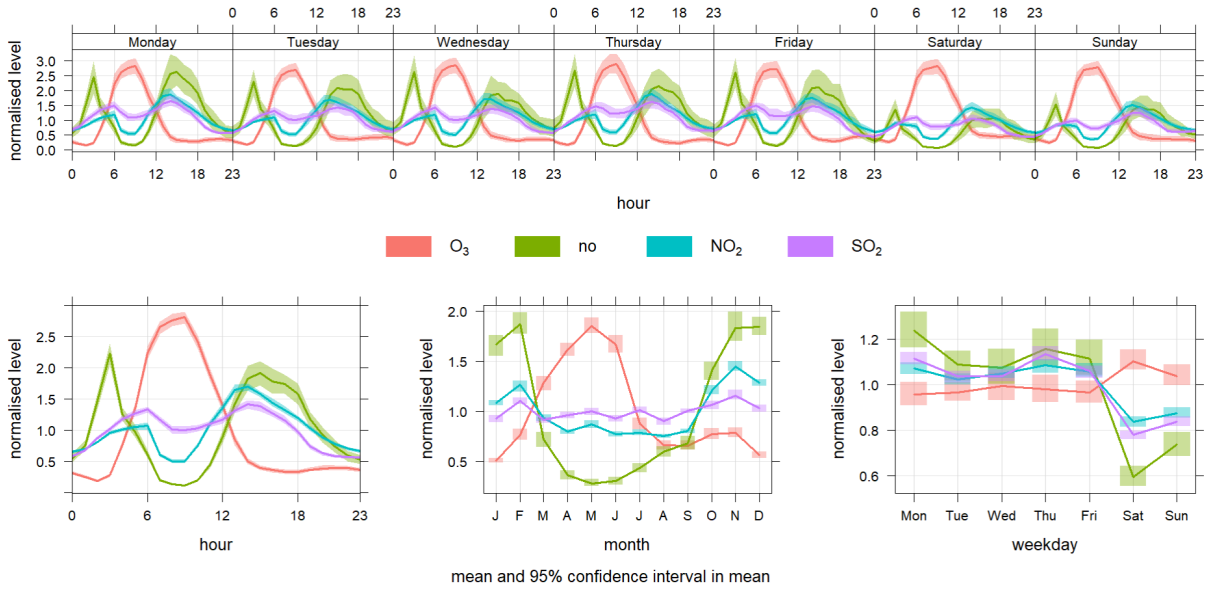

FIGURE 2 An example plot that shows the normalized concentrations of O_3, NO, NO_2, and SO_2 concentrations in Delhi using the timeVariation function

This plot suggests a peak during the morning rush hour from 7:00 to 9:00. The difference line also shows a more significant difference in pollutant emissions between weekdays and weekends. Also, given that the number of cars at this location is roughly constant throughout the day, the variation could result from the emissions of other vehicle types.

3.3 ScatterPlot Function

Linear regression helps to determine whether a relationship exists between the dependent variable (pollutant concentration) and the independent variable (meteorological factor). It will produce the coefficient of determination (R^2), which can help to interpret the correlation/relationship result. If a relationship does exist between two variables (meteorological factors and pollutant concentration), the value of the relationship will either be positive or negative.

Positive values can be identified by the slope being an upwardly sloped curve (positive slope); negative values can be determined by the gradient being a downwardly sloped curve (negative slope). A positive relationship's direction indicates a proportional relationship, such as an increase in the condition/value of a meteorological factor followed by an increase in pollutant concentration. A negative relationship shows a reversed relationship that increases the condition/value of the meteorological facet, followed by a decrease in pollutant concentration.

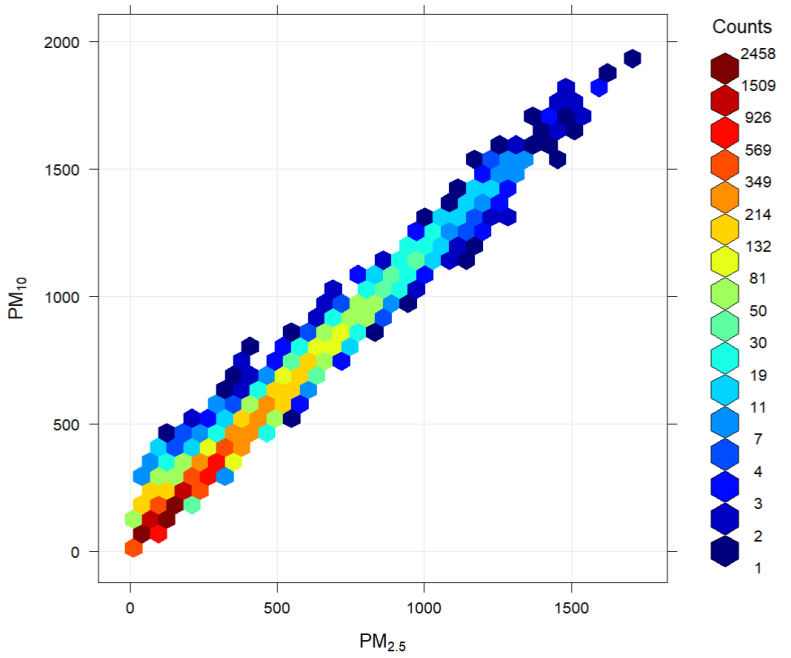

FIGURE 3 Scatter plot of hourly $PM_{2.5}$ vs. PM_{10} in Delhi using hexagonal binning. The number of occurrences in each bin is color-coded (but not on a linear scale). It is now possible to see where most of the data lies in a clearer correlation picture between $PM_{2.5}$ and PM_{10}.

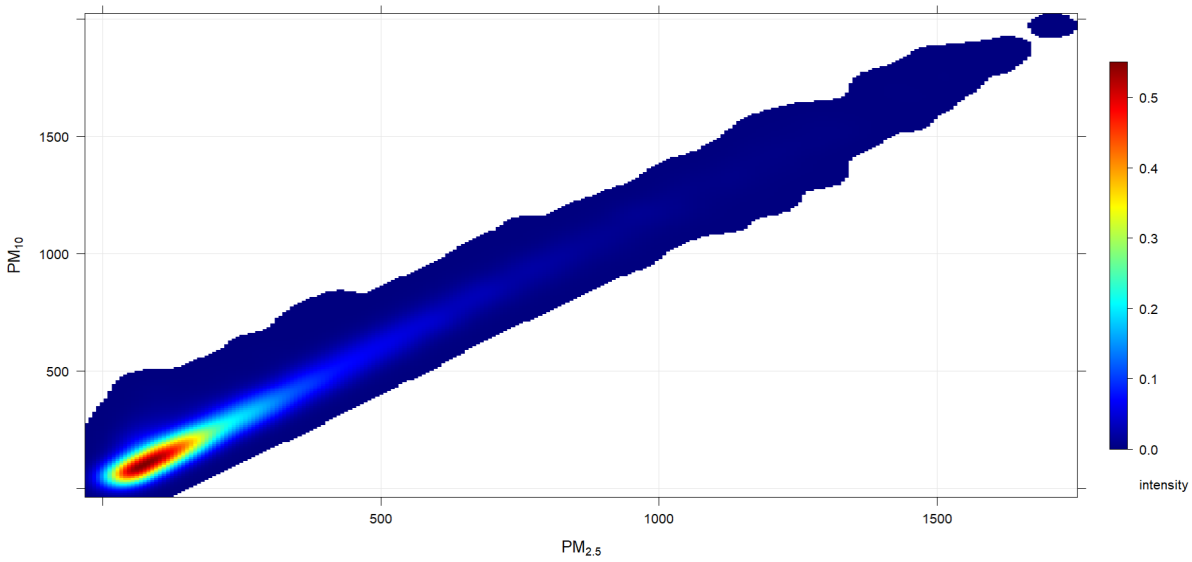

FIGURE 4 A scatter plot of hourly $PM_{2.5}$ vs. PM_{10} levels in Delhi using a kernel density estimate shows the locations for most of the points. The "intensity" measures how many dots of $PM_{2.5}$ and PM_{10} concentration exist in a unit area.

4. CONCLUSION

The Openair model is a tool to analyze, interpret, and comprehend air pollution data to improve

air quality management. This model uses statistical analyses designed for air quality modelings, such as linear regression, *p-value* decision-making, and coefficient determination, all of which correlate to various functions. Each function in the openair model serves a specific utility purpose. This research only looks at three model functions: TheilSen for analyzing pollution concentration trends, timeVariation for analyzing pollutant temporal variations, and scatterPlot for analyzing the linear correlation between two variables. The results from the three functions show deteriorating air quality from 2020 to 2023. As air pollutants and monitoring periods lengthen, the use of open-air models in Delhi will soon increase to conclude more accurate air quality data.

5. REFERENCES

35% drop in total vehicles on Delhi roads since ban on Overage Automobiles: Economic Survey. The Economic Times. (n.d.). Retrieved April 21, 2023, from https://economictimes.indiatimes.com/news/india/35-drop-in-total-vehicles-on-delhi-roads-since-ban-on-overage-automobiles-economic-survey/articleshow/98824753.cms

Bevans, R. (2022, November 15). *Linear regression in R: A step-by-step guide & examples*. Scribbr. Retrieved April 22, 2023, from https://www.scribbr.com/statistics/linear-regression-in-r/

David, C. (n.d.). The openair manual open-source tools for analysing air pollution data. Retrieved April 21, 2023, from https://davidcarslaw.com/files/openairmanual.pdf

Frost, J. (2023, March 21). *How to interpret p-values and coefficients in regression analysis*. Statistics By Jim. Retrieved April 22, 2023, from https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Garzón, J. P., Huertas, J. I., Magaña, M., E. Huertas, M., Cárdenas, B., Watanabe, T., Maeda, T., Wakamatsu, S., & Blanco, S. (2015, August 22). *Volatile organic compounds in the atmosphere of Mexico City*. Atmospheric Environment. Retrieved April 22, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S1352231015302594

Gaur, M., Singh, R., & Shukla, A. (2016, September 9). *Variability in the levels of BTEX at a pollution hotspot in New Delhi, India*. Journal of Environmental Protection. Retrieved April 21, 2023, from https://www.scirp.org/journal/paperinformation.aspx?paperid=70493

Gour, A. A., Singh, S. K., Tyagi, S. K., & Mandal, A. (2015, January 12). *Variation in parameters of ambient air quality in National Capital Territory (NCT) of Delhi (India)*. Atmospheric and Climate Sciences. Retrieved April 21, 2023, from https://www.scirp.org/journal/paperinformation.aspx?paperid=53097

Intan, A., Hernani, Y., Endro , S., & Dodo, G. (n.d.). *Application of open air model (R package) to analyze air pollution data*. Retrieved April 21, 2023, from https://www.researchgate.net/publication/322500987_APPLICATION_OF_OPEN_AIR_MODEL_R_PACKAGE_TO_ANALYZE_AIR_POLLUTION_DATA

R-4.3.0 for windows. Download R-4.3.0 for Windows. The R-project for statistical computing. (n.d.). Retrieved April 22, 2023, from https://cran.r-project.org/bin/windows/base/

Sirohiwal, D. (2023, January 24). *Air Quality Data of Delhi, India*. Kaggle. Retrieved April 21, 2023, from https://www.kaggle.com/datasets/deepaksirohiwal/delhi-air-quality

Turney, S. (2022, September 14). *Coefficient of determination (R^2): Calculation & interpretation*. Scribbr. Retrieved April 22, 2023, from https://www.scribbr.com/statistics/coefficient-of-determination/

World Health Organization. (n.d.). *Billions of people still breathe unhealthy air: New who data*. World Health Organization. Retrieved April 22, 2023, from https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data

World's most polluted countries in 2022 - PM2.5 ranking. IQAir. (n.d.). Retrieved April 22, 2023, from https://www.iqair.com/world-most-polluted-countries