**AQI prediction and analysis of Delhi using machine learning**

|  |  |  |
| --- | --- | --- |
| *Khushi Yaduka*  *SOT, Computer Science & Engineering*  *Pandit Deendayal Energy University*  *Gandhinagar, India khushi.yce21@sot.pdpu.ac.in* | *Divya Kapse*  *SOT, Computer Science & Engineering*  *Pandit Deendayal Energy University*  *Gandhinagar, India divya.kce21@sot.pdpu.ac.in*  *Samir Patel*  *SOT, Computer Science & Engineering*  *Pandit Deendayal Energy University*  *Gandhinagar, India samir.patel@pdpu.ac.in* | *Dhanvi Bhimani*  *SOT, Computer Science & Engineering*  *Pandit Deendayal Energy University*  *Gandhinagar, India dhanvi.bce21@sot.pdpu.ac.in* |

**Abstract***—* Air pollution in Delhi, the national capital of India, has been a grave concern for decades, with persistent violations of air quality standards and adverse health implications for its residents. This work presents a comprehensive analysis of Delhi's air quality over the past seven years (2017-2023), with a particular focus on the Air Quality Index (AQI) as a holistic indicator of air pollution levels. Through a systematic evaluation of research studies, monitoring data, and policy interventions, the work explores the spatiotemporal variations in air pollutant concentrations, identifies the major contributing sources, and examines the role of meteorological factors in exacerbating pollution episodes. Additionally, the work delves into the application of statistical and machine learning techniques for air quality analysis and prediction, highlighting their potential in informing air quality management strategies. Furthermore, the work scrutinizes the impact of the COVID-19 lockdowns on Delhi's air quality, offering insights into the relationship between anthropogenic activities and pollution levels. Finally, this work critically assesses the effectiveness of existing policies and proposes a comprehensive framework for sustainable air quality management in Delhi, incorporating regulatory measures, public engagement, and regional cooperation.

***Keywords***: - ***Air Quality Index (AQI), COVID-19, Spatiotemporal variations, Air pollution, Delhi, Respiratory diseases***

1. **Introduction**

Air pollution has emerged as a pressing environmental and public health concern in urban centres worldwide, with Delhi, the national capital of India, being one of the most severely affected cities. Despite various policy interventions and judicial directives, Delhi's air quality has remained a persistent challenge, with alarming levels of particulate matter (PM2.5 and PM10) and gaseous pollutants like nitrogen oxides (NOx), sulphur dioxide (SO2), carbon monoxide (CO), and ozone (O3) consistently exceeding national and international standards. [2] [9] [4]

India is one of the countries with the highest exposure to particulate matter (PM) 2.5 in the world. The average level of PM 2.5 was 89.9 mg/m³ in 2017, with pollution highest in Delhi, followed by Uttar Pradesh, Bihar and Haryana (Balakrishnan et al., 2019) [7]. Air quality is monitored at more than 450 stations across the country by the Central Pollution Control Board (CPCB) under the National Air Quality Monitoring Program (NAMP) to assess concentrations in industrial, residential, transport and environmentally sensitive areas [8] [13]. Total suspended particulate matter was estimated by a gravimetric method using a high-flow respirable dust sampler for 12 to 24 hours. [1]

The deteriorating air quality in Delhi can be attributed to a confluence of factors, including rapid urbanization, population growth, and the associated increase in vehicular emissions, industrial activities, and construction activities. Additionally, the open burning of biomass, including crop residue burning in the neighbouring states, contributes significantly to the pollution load. Meteorological conditions, such as low wind speeds, temperature inversions, and low boundary layer heights, further exacerbate the situation by inhibiting the dispersion of pollutants, particularly during the winter months. [11]

In recent years, there has been a surge in research efforts aimed at analysing and predicting air quality in Delhi, driven by the urgent need for effective air quality management strategies. Numerous studies have employed statistical and machine learning techniques, such as regression models, time series analysis, and artificial neural networks, to analyse historical air quality data and meteorological parameters, with the goal of predicting future air quality indices (AQIs) and informing decision-making processes. [11]

Furthermore, the COVID-19 pandemic and the subsequent lockdowns in 2020 and 2021 provided a unique opportunity to observe the impact of reduced anthropogenic activities on Delhi's air quality. While temporary improvements were observed during the strict lockdown phases, the resurgence of pollution levels after the easing of restrictions highlighted the need for sustained and comprehensive measures to address the root causes of air pollution in the city. [3] [6]

This review paper aims to synthesise the existing body of research on air quality analysis and prediction in Delhi, with a particular emphasis on the AQI as a comprehensive indicator of air pollution levels. By critically evaluating the spatiotemporal variations in air pollutant concentrations, identifying the major contributing sources, and examining the role of meteorological factors, the paper seeks to provide a holistic understanding of the air quality dynamics in Delhi [5] [3]. Furthermore, the review delves into applying statistical and machine learning techniques for air quality analysis and prediction, highlighting their potential in informing air quality management strategies. Additionally, the paper scrutinises the impact of the COVID-19 lockdowns on Delhi's air quality, offering insights into the relationship between anthropogenic activities and pollution levels [3]. Finally, the review critically assesses the effectiveness of existing policies and proposes a comprehensive framework for sustainable air quality management in Delhi, incorporating regulatory measures, public engagement, and regional cooperation. [8] [9]

1. **Literature review**

Air pollution in Delhi, India is a serious threat to public health. The World Health Organization (WHO) has specified RR values ​​and the corresponding base incidence for various air pollutants as well as the types of diseases associated with these values [14]. Delhi (or the National Capital Territory of Delhi), is jointly administered by the central and state governments. It accommodates nearly 167.5 lakh people (2011 Census of India) [10]. Predicting the Air Quality Index (AQI) can be a crucial tool for mitigating these risks by allowing people to take preventive measures. This review examines existing research on AQI prediction in Delhi, focusing on the effectiveness of various techniques and potential areas for improvement.

***Current Prediction Techniques:***

Existing research on AQI prediction in Delhi explores various techniques. One common approach utilizes statistical modelling. Kumar and Goyal (2011) investigated Principal Component Regression (PCR) to forecast daily AQI based on the previous day's AQI and meteorological variables like temperature, wind speed and direction, humidity, and rainfall.[9] While this method performed well in winter, its accuracy decreased in other seasons. This highlights the need for models that account for seasonal variations in air quality. These models achieve moderate accuracy but might not capture the complex relationships between various factors influencing air quality.

***Future Research Directions:***

Several promising avenues exist for future research on AQI prediction in Delhi. Machine learning algorithms like Artificial Neural Networks (ANNs) hold significant potential. Studies by Pal and Mather (2003) and Singh et al. (2017) explored the use of ANNs for AQI prediction. These algorithms can learn complex non-linear relationships between data points, potentially leading to more accurate predictions compared to traditional statistical models.

Another crucial area for improvement lies in integrating real-time data streams. Existing models might not fully utilize the wealth of information available from air quality monitoring stations. Future research should explore methods for incorporating real-time data like pollutant concentrations and satellite imagery into prediction models. This real-time data can allow for more up-to-date and accurate forecasts.

Developing multi-pollutant models is another key area for advancement. Current models often focus on individual pollutants, while AQI is a function of the combined effects of various pollutants like PM2.5, PM10, NO2, and SO2. Multi-pollutant models can provide a more comprehensive picture of air quality and improve prediction accuracy. [2] [5]

Finally, there is a need for spatiotemporal prediction models. Current models might not provide location-specific information. Future research should explore methods that consider spatial variations in air quality and predict AQI for different areas within Delhi. This would provide residents with more actionable information to protect their health based on their specific location.[3]

***Dataset Parameters Description:***

Particulate Matter (PM2.5 and PM10): Particulate matter consists of various solids and liquids, including carbon, complex organic compounds, sulphates, nitrates, mineral dust, and water suspended in the atmosphere. PM exists in various sizes, with some particles visible to the naked eye, such as dust, soot, dirt, or smoke. However, the most harmful particles are smaller and categorized as PM10 and PM2.5.[12]

Nitrogen Oxides (NO2): Nitrogen oxides encompass a group of seven gases and compounds comprising nitrogen and oxygen, collectively referred to as NO2 gases. Among these, nitrogen dioxide (NO2) is the most prevalent and hazardous.[12]

Sulphur Dioxide (SO2): Sulphur dioxide, a colourless gas with a pungent Odor akin to a recently lit match, forms when sulphur-containing fuels like coal and oil are burned, contributing to air pollution.[12]

Carbon Monoxide (CO): Carbon monoxide, a colourless and highly toxic gas, can transform into a liquid under pressure. It is generated by the combustion of gasoline, natural gas, charcoal, wood, and other fuels.[12]

Ozone (O3): Ground-level ozone, a colourless and highly irritating gas, forms just above the Earth's surface. It is categorized as a "secondary" pollutant because it arises from the reaction of primary pollutants—nitrogen oxides (NOx) and volatile organic compounds (VOCs)—in sunlight and stagnant air.[12]

1. **AQI Calculation**

To quantify the Air Quality Index (AQI) for individual pollutants, a standardized formula is employed, as prescribed by regulatory agencies such as the Central Pollution Control Board (CPCB) and the Environmental Protection Agency (EPA). The formula, outlined below, calculates the index of a pollutant (Ip) based on its concentration and predefined breakpoints:

Where:

- (Ip) = Index of pollutant (p)

- (Cp) = Truncated concentration of pollutant (p)

- (BPHI) = Concentration breakpoint greater than or equal to (Cp)

- (BPLO) = Concentration breakpoint less than or equal to (Cp)

- (IHi) = AQI value corresponding to (BPHI)

- (ILO) = AQI value corresponding to (BPLO)

*AQI Categories:*

Table 1 provides a comprehensive breakdown of Air Quality Index (AQI) categories, corresponding pollutant concentrations, and health breakpoints. Each AQI category is defined by a specific range, and the corresponding concentration ranges for key pollutants, including PM10, PM2.5, NO2, O3, CO, and SO2 are provided for different periods (e.g., 24-hour and 8-hour averages).

**Good (0-50):** Represents excellent air quality with minimal health risks.

**Satisfactory (51-100):** Indicates acceptable air quality, posing low health risks for the general population.

**Moderately Polluted (101-200):** This signifies air quality that may cause slight discomfort to sensitive individuals and those with respiratory conditions.

**Poor (201-300):** Indicates air quality that may have adverse health effects, particularly for sensitive groups and individuals with existing health conditions.

**Very Poor (301-400):** Represents significantly degraded air quality, posing serious health risks to the general population and substantial risks to vulnerable individuals.

**Severe (401-500):** Indicates hazardous air quality, with severe health impacts for all individuals.

*Description of Pollutants and Health Breakpoints:*

**PM10 and PM2.5**: Fine particulate matter, categorized by particle size, with higher concentrations indicating increased health risks, especially for respiratory and cardiovascular systems.

**NO2:** Nitrogen dioxide, a gas primarily emitted from vehicles and industrial sources, is associated with respiratory issues and inflammation.

**O3:** Ozone, a reactive gas formed by chemical reactions between nitrogen oxides and volatile organic compounds, with high levels causing respiratory problems and aggravating existing conditions.

**CO:** Carbon monoxide, a colourless and odourless gas emitted from combustion processes, with high concentrations leading to reduced oxygen transport in the bloodstream and adverse health effects.

**SO2:** Sulphur dioxide, a gas emitted from burning fossil fuels, contributes to respiratory issues, particularly in individuals with asthma or other lung diseases.

**Table 1 : Aqi Category, Pollutants and Health Breakpoints**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AQI**  **Category**  **(Range)** | **PM10**  **24-hr** | **PM2.5**  **24-hr** | **NO2**  **24-hr** | **O3**  **8-hr** | **CO**  **8-hr**  **(mg/m3)** | **SO2**  **24-hr** | **NH3**  **24-hr** | **Pb**  **24-hr** |
| Good  (0-50) | 0-50 | 0-30 | 0-40 | 0-50 | 0-1.0 | 0-40 | 0-200 | 0-0.5 |
| Satisfactory (51-100) | 51-100 | 31-60 | 41-80 | 51-100 | 1.1-2.0 | 41-80 | 201-400 | 0.5-1.0 |
| Moderately  Polluted  (101-200) | 101-250 | 61-90 | 81-180 | 101-168 | 2.1-10 | 81-380 | 401-800 | 1.1-2.0 |
| Poor  (201-300) | 251-350 | 91-120 | 181-280 | 169-208 | 10-17 | 381-800 | 801-1200 | 2.1-3.0 |
| Very poor (301-400) | 351-430 | 121-250 | 281-400 | 209-748 | 17-34 | 801-1600 | 1200-1800 | 3.1-3.5 |
| Severe  (401-500) | 430 + | 250 + | 400 + | 748 + | 34 + | 1600 + | 1800 + | 3.5 + |

1. **Data Preprocessing And Method**

*Data Collection and Compilation:*

Historical air quality data was meticulously collected from various authoritative sources, with a focus on capturing a comprehensive view of air quality trends over a period of seven years, from 2017 to 2023.

The primary source of data was the Central Pollution Control Board (CPCB), which serves as the central regulatory body responsible for monitoring and controlling pollution levels in India, including air quality.

Data obtained from the CPCB encompassed a wide range of pollutants, including particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulphur dioxide (SO2), carbon monoxide (CO), ozone (O3), and others, measured at multiple monitoring stations across Delhi.

*Data Cleaning and Quality Assurance:*

The dataset underwent rigorous cleaning to rectify missing values, outliers, and inconsistencies. Imputation or removal addressed missing values, while outliers were detected and evaluated for validity before correction or removal. Quality assurance measures, including cross-referencing and consistency checks, ensured dataset reliability. Detailed documentation-maintained transparency and reproducibility. This meticulous process ensured the dataset's suitability for analysis, interpretation, and modelling, providing a solid foundation for understanding Delhi's air quality dynamics. By enhancing accuracy and consistency, the cleaned dataset facilitates evidence-based decision-making and policy formulation to combat air pollution effectively, benefiting diverse stakeholders.

*Statistical Analysis with Visualization:*

Descriptive statistics, encompassing mean, median, and standard deviation, were computed to delineate the central tendency and dispersion of pollutant levels across monitoring stations and seasons, offering insights into their typical concentrations and variability. Visualizations such as histograms and box plots were employed to depict the distribution of pollutant concentrations, aiding in the identification of data patterns and outliers. Additionally, correlation analysis was conducted to examine relationships between pollutant concentrations, meteorological parameters, and other relevant variables. Heatmaps and scatter plots were utilized to visualize correlations, facilitating the identification of influential factors impacting air quality. These analytical and visualization techniques provided a comprehensive understanding of the dataset's characteristics and relationships, informing subsequent analysis and modelling efforts. Moreover, a dynamic dashboard incorporating interactive charts and graphs was developed to provide stakeholders with a user-friendly interface for exploring and analysing air quality data, enhancing accessibility and decision-making processes.

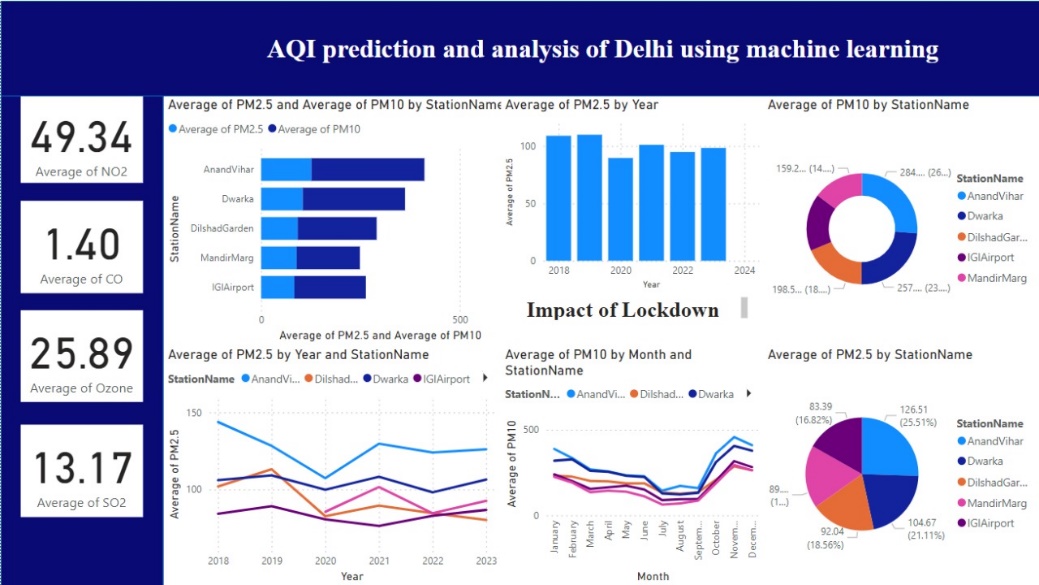


Figure 1: Dashboard of AQI prediction and analysis of Delhi using machine learning

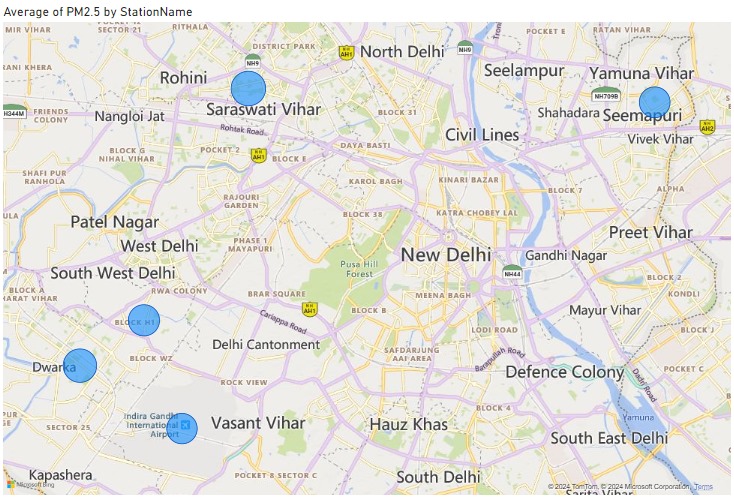


Figure 2: Average of PM2.5 by Station Name

*Machine Learning Modelling (Random Forest Regressor):*

A Random Forest Regressor model was chosen due to its capability to handle intricate relationships and nonlinearities present in the air quality dataset. This ensemble learning algorithm constructs multiple decision trees and combines their predictions to generate a robust model. The historical air quality data, enriched with meteorological variables such as temperature, wind speed, and humidity, served as inputs for training the Random Forest Regressor model. Hyperparameter tuning was conducted to optimize the model's performance. This involved fine-tuning parameters like the number of trees in the forest and the maximum depth of each tree to strike a balance between model complexity and generalization. By leveraging Random Forest Regression, the model aimed to accurately predict Air Quality Index (AQI) variations, providing valuable insights for air quality management strategies and decision-making processes.

*Model Evaluation and Prediction:*

Using the trained Random Forest Regressor model, predictions of Air Quality Index (AQI) variations for future time periods were generated. These predictions offer valuable insights into expected changes in air pollution levels, allowing for proactive planning and implementation of air quality management strategies. By forecasting AQI trends, stakeholders can anticipate potential pollution hotspots, identify vulnerable populations, and allocate resources efficiently to mitigate adverse health and environmental impacts. Additionally, these predictions serve as early warning indicators, enabling timely interventions and public advisories to safeguard community health and well-being. Overall, the predictive capabilities of the model empower decision-makers with actionable information to address air pollution challenges effectively.

1. **Results And Discussion**

*Descriptive Statistics:*

Descriptive statistics, including metrics like mean, median, and standard deviation, play a vital role in understanding pollutant levels across multiple stations. They provide valuable insights into the central tendencies and variability of the data distribution, facilitating effective decision-making and informed policy interventions. Moreover, in pollutant-level analysis, the accuracy of predictive models, such as random forest models, is paramount. Achieving an impressive accuracy rate of 99 per cent, these models offer reliable forecasts, empowering stakeholders to make timely, data-driven decisions to mitigate environmental risks and safeguard public health.

Table 2: Results [ Accuracy with RMSE]

|  |  |
| --- | --- |
| **Machine Learning Model** | **Accuracy with RMSE** |
| Random Forest | 0.01 |
| SVM | 7.17 |
| KNN | 5.66 |

In the realm of regression analysis, metrics such as RMSE, MAE, and R² serve as critical tools for evaluating model performance. RMSE, reflecting the average disparity between predicted and observed values, is pivotal, with lower scores indicative of heightened precision. Additionally, MAE offers insights into the average absolute deviation, while R² elucidates the model's explanatory power. Notably, the best-performing model, identified as the Random Forest model with an RMSE of 0.005567007450193569, underscores the efficacy of the chosen approach.

*Exploratory Data Analysis (EDA):*

Exploratory Data Analysis (EDA) serves as a foundational step in understanding the dataset's characteristics and exploring potential relationships between variables. Through techniques like histograms and box plots, EDA visually represents the distribution and spread of pollutant concentrations. These visualizations provide initial insights into data patterns, identifying potential outliers, skewness, or multimodal distributions. Moreover, EDA facilitates hypothesis generation by revealing associations between variables, guiding subsequent analytical steps. By visually inspecting data trends and relationships, researchers can make informed decisions about further analyses, model selection, and feature engineering, enhancing the overall effectiveness of air quality analysis and prediction efforts.

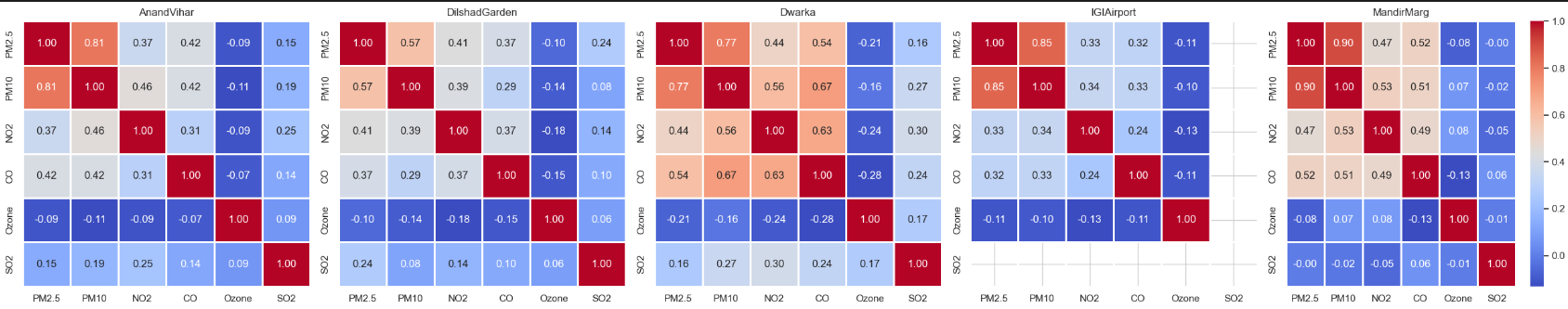


Figure 3: Exploratory Data Analysis (EDA) between different parameters

*Prediction of AQI:*

The study utilized a Random Forest Regressor for predicting Air Quality Index (AQI) variations with satisfactory accuracy. Evaluation metrics like MSE, RMSE, and R-squared validated model performance, while cross-validation ensured robustness and generalization.

1. **Conclusion**

Our study reveals the persistent challenge of air pollution in Delhi, driven by urbanization and seasonal factors. Statistical and machine learning techniques, particularly Artificial Neural Networks, demonstrated potential in accurately analysing and predicting air quality. Despite improvements, challenges persist, especially during winter months. The study underscores the importance of continuous monitoring and robust analysis to understand and address the complexities of air quality dynamics in the city.

Future research should prioritize integrating real-time data and refining prediction models to enhance accuracy. Multi-pollutant models and spatiotemporal approaches offer promising avenues for comprehensive air quality analysis. These advancements are crucial for informing evidence-based policies and interventions to mitigate air pollution effectively in Delhi and beyond.

Collaborative efforts between policymakers, researchers, and communities are essential for sustainable improvements in Delhi's air quality. Stricter regulations, public awareness campaigns, and regional cooperation are imperative to address the root causes of air pollution. By implementing holistic strategies informed by robust data analysis, we can safeguard public health and promote environmental sustainability in the national capital and beyond

**References**

1. Pande JN, Bhatta N, Biswas D, Pandey RM, Ahluwalia G, Siddaramaiah NH, Khilnani GC. Outdoor air pollution and emergency room visits at a hospital in Delhi. Indian J Chest Dis Allied Sci. 2002 Jan-Mar;44(1):13-9. PMID: 11845928.
2. Seema, A. & Sood, Aditya & Bhalla, Simran & Bajpai, Shantam. (2018). Impact of Rising Air Pollution in New Delhi: An Empirical Study. International Journal of Mechanical Engineering and Technology. 9.
3. Shankar, S., Gadi, R. Variation in Air Quality over Delhi Region: A Comparative Study for 2019 and 2020. Aerosol Sci Eng 6, 278–295 (2022). <https://doi.org/10.1007/s41810-022-00144-7>.
4. Nidhi Sharma, Shweta Taneja, Vaishali Sagar, Arshita Bhatt. Forecasting air pollution load in Delhi using data analysis tools. Procedia Computer Science, Volume 132. 2018. Pages 1077-1085, ISSN 1877-0509. <https://doi.org/10.1016/j.procs.2018.05.023>.
5. Dutta A, Jinsart W. Air pollution in Delhi, India: It's status and association with respiratory diseases. PLoS One. 2022 Sep 20;17(9): e0274444. doi: 10.1371/journal.pone.0274444. PMID: 36126064; PMCID: PMC9488831.
6. Talukdar, S.; Tripathi, S.N.; Lalchandani, V.; Rupakheti, M.; Bhowmik, H.S.; Shukla, A.K.; Murari, V.; Sahu, R.; Jain, V.; Tripathi, N.; et al. Air Pollution in New Delhi during Late Winter: An Overview of a Group of Campaign Studies Focusing on Composition and Sources. Atmosphere 2021, 12, 1432. <https://doi.org/10.3390/atmos12111432>.
7. Chowdhury, Soumi & Pohit, Sanjib & Singh, Rishabh. (2023). Health and Economic Impact of Air Pollution in Delhi HEALTH AND ECONOMIC IMPACT OF AIR POLLUTION IN DELHI NCAER Working Paper.
8. De Vito, Laura & Chatterton, Tim & Namdeo, Anil & SM, Shiva Nagendra & Gulia, Sunil & GOYAL, SANJIV & Bell, Margaret & Goodman, Paul & Longhurst, J. & Hayes, Enda & Kumar, Rakesh & Sethi, Virendra & RAMADURAI, SENGUPTA & MAJUMDER, SHOBAN & Menon, Jyothi & TURAMARI, MALLIKARJUN & Barnes, Jo. (2018). Air pollution in Delhi: A review of past and current policy approaches. WIT Transactions on Ecology and the Environment. 230. 441-451. 10.2495/AIR180411.
9. Anikender Kumar, Pramila Goyal, forecasting of air quality in Delhi using principal component regression technique, Atmospheric Pollution Research, Volume 2, Issue 4,2011, Pages 436-444, ISSN 1309-1042. <https://doi.org/10.5094/APR.2011.050>.
10. Suliankatchi, Rizwan & Nongkynrih, Baridalyne & Gupta, Sanjeev. (2013). Air pollution in Delhi: Its Magnitude and Effects on Health. Indian Journal of Community Medicine. 38. 4-8.
11. Guttikunda, S.K.; Dammalapati, S.K.; Pradhan, G.; Krishna, B.; Jethva, H.T.; Jawahar, P. What Is Polluting Delhi’s Air? A Review from 1990 to 2022. Sustainability 2023, 15, 4209. <https://doi.org/10.3390/su15054209>
12. Vamshi, B Nitesh. (2021). Air Quality analysis. Journal of Data Analysis.
13. Mohan, Manju & Kandya, Anurag. (2007). An Analysis of the Annual and Seasonal Trends of Air Quality Index of Delhi. Environmental monitoring and assessment. 131. 267-77. 10.1007/s10661-006-9474-4.
14. Nagpure, Ajay & Gurjar, Bhola & Marterl, Jc. (2014). Human health risks in National Capital Territory of Delhi due to air pollution. Atmospheric Pollution Research. 5. 10.5094/APR.2014.043.