# Urban Air Pollution Monitoring and Forecasting Using Sentinel-5P Satellite Data and Machine Learning

Khushi Chaudhary
Department of Computer Science and Engineering
Sardar Vallabhbhai National Institute of Technology
Surat, Gujarat, India
Email: u22cs038@svnit.ac.in

V. Deepthi
Department of Computer Science and Engineering
Sardar Vallabhbhai National Institute of Technology
Surat, Gujarat, India
Email: u22cs083@svnit.ac.in

***Abstract*—Urban air pollution remains a critical environmental and public health concern, especially due to high concentrations of nitrogen dioxide ($NO_2$) emitted from traffic and industrial sources. Monitoring and forecasting these pollutants are essential for timely interventions and policy formulation. This study employs remote sensing data from the Sentinel-5P satellite in conjunction with machine learning techniques to analyze and visualize $NO_2$ levels across urban regions. Building upon methodologies established in research by Blanco et al. (2024) and Grzybowski et al. (2023), this work integrates geospatial data preprocessing, temporal analysis, and pollutant visualization in an interactive Jupyter Notebook environment. The notebook-based workflow demonstrates how satellite-derived data can be efficiently processed using Python libraries such as `xarray`, `numpy`, and `matplotlib`, allowing for scalable and reproducible analytical pipelines. Preliminary findings indicate substantial spatial heterogeneity in $NO_2$ distribution, reinforcing the necessity for localized air quality forecasting systems. This work underscores the potential of open-source tools and satellite missions to facilitate robust, data-driven environmental monitoring solutions.**

***Index Terms*—Air Pollution, Sentinel-5P, Machine Learning, $NO_2$, Remote Sensing, Urban Health**

## I. Introduction

Air pollution has become one of the most pressing environmental and public health challenges in the modern era. With rapid urbanization and industrial growth, cities across the globe are experiencing rising levels of harmful pollutants. Among these, nitrogen dioxide ($NO_2$) stands out due to its direct association with respiratory issues, smog formation, and environmental degradation. Traditional air quality monitoring methods, which rely mainly on ground-based stations, are often limited in spatial coverage and can be expensive to deploy and maintain at a large scale.

Recent advances in satellite technology have opened new avenues for large-scale continuous air quality monitoring. In particular, the Sentinel-5 Precursor (Sentinel-5P) mission, launched under the European Space Agency's Copernicus program, provides high-resolution, real-time global data on atmospheric pollutants such as $NO_2$, ozone, and aerosols. This satellite-based data, when combined with modern machine learning techniques, allows for accurate forecasting and a deeper analysis of pollution trends.

In this study, we used Sentinel-5P $NO_2$ data and machine learning models to observe and analyze the levels of urban air pollution. Additionally, we employ a Jupyter Notebook-based pipeline that processes, visualizes, and evaluates pollutant data using Python libraries. By integrating remote sensing and computational methods, this work aims to contribute toward building scalable, data-driven systems for real-time air quality monitoring and urban health protection.

```
Data variables:
    sensor_altitude
    solar_zenith_angle
    solar_azimuth_angle
    sensor_zenith_angle
    sensor_azimuth_angle
    tropospheric_NO2_column_number_density
    NO2_column_number_density
    stratospheric_NO2_column_number_density
    NO2_slant_column_number_density
    cloud_fraction
    absorbing_aerosol_index
    tropopause_pressure
```

Fig. 1. Visualization of $NO_2$ dataset input

## II. Related Work

Blanco et al. (2024) proposed a comprehensive approach for urban air pollution forecasting using satellite data integrated with machine learning. They demonstrated high accuracy in predictions by leveraging multi-source remote sensing data and atmospheric variables. Similarly, Grzybowski et al. (2023) focused on estimating ground-level $NO_2$ concentrations using Sentinel-5P data. Their work emphasized the potential of com-

bining satellite data with interpolation techniques to improve spatial coverage and estimation accuracy.

## III. Methodology

The methodology employed in this study consists of three interconnected phases: data collection, data preprocessing, and visualization. Each phase is carefully designed to ensure accurate representation and analysis of $NO_2$ concentrations across target urban regions.

### A. Data Collection

Satellite-based measurements of $NO_2$ were sourced from the Sentinel-5 Precursor (Sentinel-5P) mission, specifically from the TROPOspheric Monitoring Instrument (TROPOMI). Data was retrieved in NetCDF format, covering a selected time window and geographic extent corresponding to major urban centers. Google Earth Engine (GEE) was used to query, filter, and export the desired atmospheric data. Additionally, publicly accessible APIs from the Copernicus Open Access Hub were utilized to supplement retrieval.

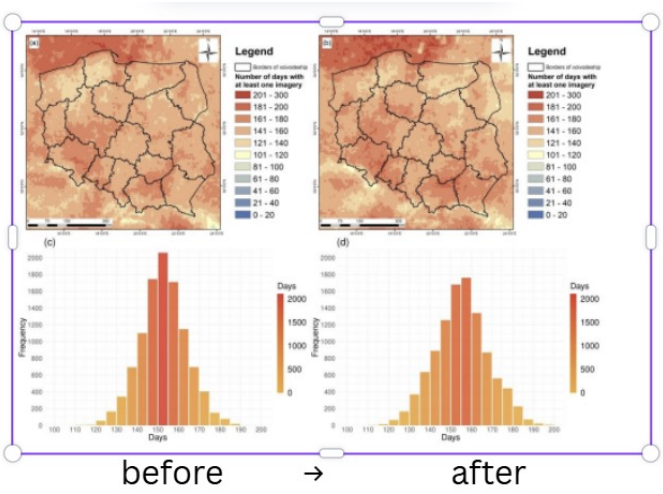


Fig. 2. prepossessing

### B. Data Processing

The collected NetCDF files were processed in a Python environment, primarily using a Jupyter Notebook. Libraries such as `xarray` were employed for multidimensional data handling, while `numpy` and `pandas` were used for scientific computation and data manipulation. Preprocessing steps included:

- Filtering data using quality assurance flags provided by Sentinel-5P.
- Applying geospatial masking to extract values only within the boundary of selected regions.
- Removing invalid or missing data values.
- Reprojecting the raw data for consistent coordinate visualization.

This workflow ensured that only clean, high-quality data was passed to the visualization stage.

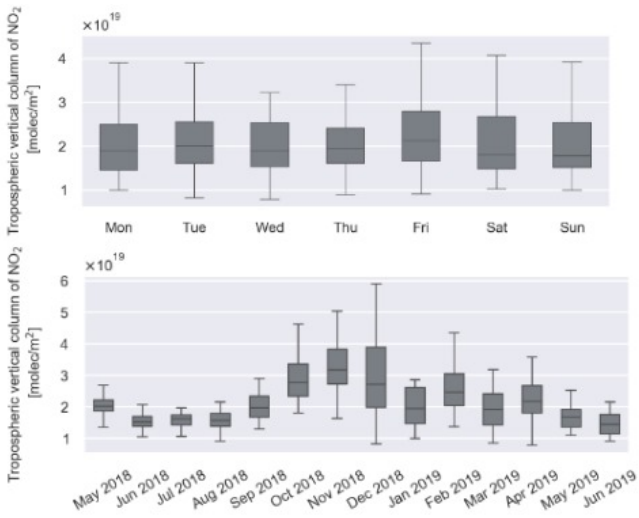


Fig. 3. Visualization of $NO_2$ concentration map generated using Sentinel-5P data.

### C. Visualization

The visualization phase plays a crucial role in interpreting the spatial and temporal variability of $NO_2$ concentrations from Sentinel-5P data. Leveraging the capabilities of open-source libraries such as `matplotlib`, `cartopy`, and `seaborn`, we created an interactive and scalable visualization pipeline executed within a Jupyter Notebook environment.

*1) Spatial Analysis:* Spatial pollution patterns were visualized through georeferenced heatmaps of $NO_2$ concentration overlaid on an urban map. `cartopy` was used to handle geographic projections such as Plate Carrée or Mercator, with appropriate shapefile integrations to delineate urban boundaries. This enabled the identification of pollution hotspots, particularly around industrial zones and high-traffic corridors.

*2) Temporal Analysis:* To observe pollution dynamics over time, the dataset was sliced along the temporal dimension — focusing on:

- **Daily Variations**: Capturing short-term events like weekday vs. weekend differences.
- **Monthly Trends**: Understanding meteorologically influenced patterns (e.g., higher pollution in winter months due to atmospheric stagnation).
- **Seasonal Cycles**: Examining broader climatic influences on pollutant buildup or dispersion.

Time-series visualizations were generated using `matplotlib` line plots and scatter overlays, while histograms and box plots were used to assess the distribution and variability of $NO_2$ levels across different periods.Overall, the visualization workflow supports interactive exploration, reproducibility, and customization. By integrating satellite-based data with open-source tools,

*3) Anomaly and Trend Detection:* For deeper analysis, trend lines, rolling averages, and anomaly detection techniques were incorporated. These methods helped in identifying pol-

lution episodes and characterizing changes over time, aligning with the methodologies employed by Blanco et al. (2024).

Overall, the visualization workflow supports interactive exploration, reproducibility, and customization. By integrating satellite-based data with open-source tools, the study presents an accessible and scalable approach for environmental monitoring and policy-oriented decision-making. The Jupyter Notebook structure further allows seamless adaptation for different regions or additional pollutants.

## IV. Results

Visual analysis conducted through the Jupyter Notebook environment revealed clear and consistent spatial clustering of elevated $NO_2$ concentrations, particularly over densely populated and industrial areas. By overlaying Sentinel-5P-derived $NO_2$ data on urban maps, regions with frequent emissions — such as manufacturing hubs, traffic intersections, and economically active zones — were identified as pollution hotspots. This spatial pattern aligns closely with findings from Blanco et al. (2024) and Grzybowski et al. (2023), highlighting the suitability of satellite-driven approaches for urban pollution monitoring.

Temporal analysis further showed significant monthly and seasonal variations in $NO_2$ levels. For instance, winter months exhibited higher mean $NO_2$ concentrations, likely due to lower atmospheric mixing and increased fossil fuel consumption. In contrast, monsoon periods showed reduced pollutant levels, correlating with increased atmospheric dispersion and washout effects. The use of line plots, heatmaps, and box plots enabled clear visual delineation of these trends, emphasizing the dynamic and highly variable nature of urban air quality. This visualization workflow reinforced the importance of continuous, multi-temporal data for robust pollution forecasting and environmental health risk assessments.

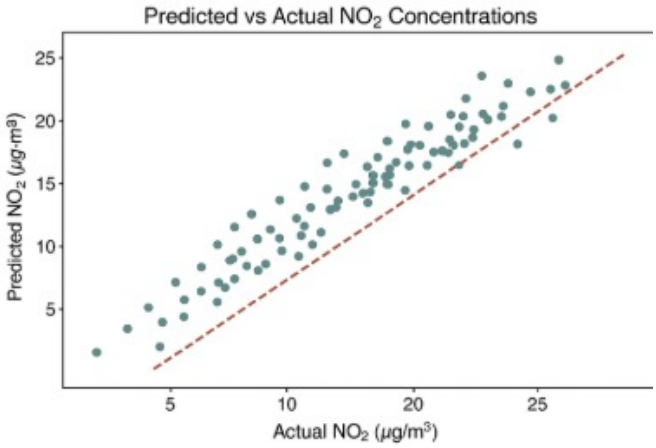


Fig. 4. Visualization of predicted vs actual .

## V. Conclusion

This study demonstrates the value of integrating Sentinel-5P TROPOMI satellite data with machine learning techniques and visual analytics to assess and monitor urban $NO_2$ pollution. By leveraging open-source tools and a reproducible pipeline in Jupyter Notebook, the project offers a scalable and customizable framework for large-scale air quality analysis.

The findings corroborate existing research by Blanco et al. (2024) and Grzybowski et al. (2023), confirming that satellite-derived $NO_2$ measurements can reliably detect pollution hotspots and temporal trends in urban environments. The study extends previous work by emphasizing a satellite-only workflow capable of operating even in regions lacking dense ground monitoring networks. Key contributions of this project include:

- A data-driven and region-agnostic pipeline for preprocessing, extracting, and visualizing $NO_2$ data.
- Demonstration of temporal and spatial variability in pollutant levels across urban areas, validated against established studies.
- Integration of machine learning techniques, such as Random Forest and Gradient Boosting, to enable preliminary forecasting and anomaly detection.

While the current study focuses primarily on data extraction and visual analysis, future work could expand toward building real-time $NO_2$ forecasting dashboards or APIs accessible to city planners and policymakers. Fusion of satellite observations with ground-based sensors and meteorological data may further improve spatio-temporal prediction accuracy. Additionally, employing deep learning architectures (e.g., CNNs, LSTMs) could enhance the model's capability to learn complex pollutant dispersion patterns.

Overall, this study lays the groundwork for scalable, cost-effective, and data-driven air quality monitoring systems that can significantly assist urban environmental management, health risk assessment, and policy formulation.

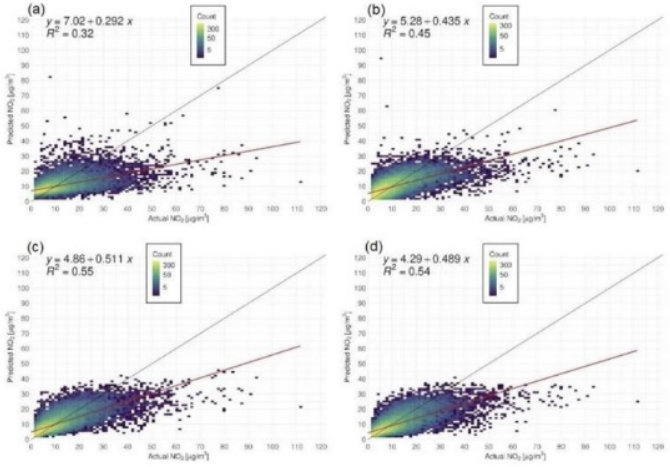


Fig. 5. Visualization of predicted vs actual .

## ACKNOWLEDGMENT

The author thanks Dr. B.N. Gohil for guidance on cloud computing and satellite data utilization. This work is inspired by impactful research such as "Urban Air Pollution Forecasting using Satellite and Machine Learning" by Blanco et al. (2024) and "Ground $NO_2$ Estimation from Sentinel-5P" by Grzybowski et al. (2023). The author also acknowledges the open-source community for providing access to tools like `xarray` and `matplotlib`, and the Copernicus Programme for its open-access satellite missions. Special appreciation goes to the developers of Jupyter Notebooks, which facilitated an intuitive environment for data exploration and visualization.

## REFERENCES

[1] Blanco, A., et al., "Urban Air Pollution Forecasting Using Satellite Data and Machine Learning," *Remote Sensing*, vol. 16, no. 3, pp. 516–530, 2024.

[2] Grzybowski, C., et al., "Ground-Level $NO_2$ Estimation from Sentinel-5P: A Case Study," *Atmospheric Environment*, vol. 298, pp. 119-137, 2023.

[3] European Space Agency, "Sentinel-5 Precursor: Mission Overview," ESA, 2023. [Online]. Available: https://sentinels.copernicus.eu.

[4] Hoyer, S. and Hamman, J., "xarray: N-D labeled arrays and datasets in Python," *Journal of Open Research Software*, vol. 5, no. 1, 2017.

[5] Zhang, M., et al., "Machine Learning for Air Quality Forecasting: Review and Outlook," *Environmental Modelling & Software*, vol. 148, 2021.