A REPORT

ON

AIR POLLUTION CORRELATION WITH URBANISATION

BY

Rishi Bharadwaj 2022A7PS2001H

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

(September,2024)

Acknowledgements

I would like to express my gratitude to Prof. Manik Gupta for giving me the opportunity to contribute to this project and for her mentorship and guidance.

I would also like to thank Mr. Ravi Bhushan for his assistance and weekly follow-ups which gave me clear ideas on how to proceed with this project.

**Table Of Contents**

1. **Cover** ................................................................................. 1
2. **Acknowledgments** ........................................................................... 2
3. **Table of Contents** .......................................................... 5
4. **Abstract** ..................................................................... 6
5. **Main Text** .......................................................................... 7
6. 7.1 Intent Detection Using Sentence Similarity................... 7
7. 7.2 Transformer Architecture Used in Embedding Models……………….8
8. 7.3 Choosing an Appropriate Embedding Model ............................................... 9
9. 7.4 Implementation of Intent Detection Model .................................................. 10
10. 7.5 Results after Deployment .................................................... 11
11. **Conclusions**................................................................ 12
12. **References** ..................................................................... 13
13. **Glossary** ……………………………………………...14

# Abstract

Air pollution related studies have become a vital part of climate science due to pollutions impact on the environment and on human well-being. The effect of air pollution on human health has been well documented and researched. Various pollutants like Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Carbon Dioxide (CO2), Sulphur dioxide (SO2) etc. when inhaled can cause severe damage to the lungs and other organs over time. They can also cause breathing issues like asthma and contribute to allergies.

On a larger scale these pollutants also contribute to global warming and phenomena such as acid rain. As climate change is becoming a larger and larger issue, tracking what factors contribute to air pollution has become more and more vital. One of the largest factors contributing to climate change is human activity. As our population has increased and our cities have rapidly developed, our use of fossil fuels for energy has also skyrocketed. For this project, the region of interest is Hyderabad, a growing metropolitan city in India. As it has grown and expanded over the previous few decades, air pollution has become a significant issue, caused by several factors including increasing population density, industrial activity, and vehicular traffic.

This project aims to prove a direct correlation between air pollution and urbanisation. The region of study is Hyderabad, a growing metropolitan city in India. The region includes urban, peri-urban and rural areas. The study uses data from Sentinel 5-P satellite’s TROPOMI sensors for air quality. For data on urbanisation data Normalized Difference Vegetation Index (NDVI) and Nighttime Land Surface Temperature (NLST) data from the MODIS (Moderate Resolution Imaging Spectroradiometer) sensor were used as well as nighttime lights (NTL) data obtained from the Visible Infrared Imaging Radiometer Suite (VIIRS). Using these datasets, it is possible to show a direct correlation between air pollution and increasing urbanisation.

# Literature Survey

Remote sensing and machine learning using air pollution data has become quite widespread. Many studies have been done using machine learning models to predict future levels of pollutants using different methods. These include classical methods such as Kth Nearest Neighbours (KNN) and Support Vector Regression (SVR). SVR is a supervised machine learning algorithm that is being used for both classification and regression problems. In this algorithm, a data point is plotted in n-dimensional space after which it is classified by finding the plane that will differentiate the classes very well. The hyperplane which is considered is a linear separator of any dimension (line, plane, hyperplane). The training points are used in the decision function and are called support vectors (Simu et al., 2020) [1]. Deep learning techniques such as Artificial Neural Networks (Maleki et al., 2019) [2] have also been used. A neural network utilizes artificial neurons, which are the smallest units of data processing. They are arranged in the form of layers such that each layer learns to give a particular output. The final layer gives the output prediction. More recently, models such as LSTMs (Seng et al., 2021) [3] as well as combinations of CNNs and LSTMs have also been put forward (Qin et al., 2019) [4]. Long Short-Term Memory (LSTM) models are a type of machine learning model that contains a memory cell, which can maintain its state over time (Greff et al., 2017) [5]. This makes them particularly effective on time series data such as air pollution. Graph Neural Networks have also been used to model the spatial component of such datasets with more accuracy (Terroso-Saenz et al., 2024) [6], (Dun et al., 2022) [7]. Graph Neural Networks (GNNs) are a type of neural network that model both the characteristic features of data as well as its structural relationships. This lets them represent the spatial component of datasets like air pollution for a particular region.

Many studies have been done on the relationships between air pollution and increasing urbanisation in growing cities (Wang et al., 2020) [8], (Wang et al., 2018) [9]. These look into various factors including economic development, developing infrastructure, transport, increasing population, etc.

With the introduction of Sentinel 5-P’s TROPOMI sensor, high-resolution global air pollution data has become easily accessible. A variety of different studies have used this data. Some examples include analyzing data to assess vulnerability in certain areas (Hassaan et al., 2023) [10], conducting statistical analysis to identify the effect of anomalies such as lockdowns during the COVID-19 pandemic (Gadakh et al., 2022) [11], temporal zoning of air pollution for health management (Safarianzengir et al., 2020) [12], and exploring relationships between air pollutants and geographical and demographic data (Kaplan and Avdan, 2020) [13]. Some studies also combine Sentinel data with newer machine learning techniques such as transformers to predict future air pollution levels (Khirwar and Narang, 2024) [14].

Work has also been done on combining data from ground stations with satellite data to make more accurate predictions (Duan et al., 2024) [15]. This includes various interpolation methods such as kriging interpolation. Similar studies have analyzed pollution spatiotemporally using methods such as Moran's I (Wang et al., 2019) [16] and Local Indicators of Spatial Association (LISA) (dos Santos et al., 2024) [17], which are special methods used to identify spatial autocorrelation.

This project exploring the relationship between air pollution and urbanization in Hyderabad builds upon previous work segregating urban, peri-urban, and rural areas (Bhushan et al., 2024) [18]. It uses some of the same indicators such as NDVI, NLST, and NTL to differentiate between rural and urban areas and uses different correlation metrics to establish a relationship between air pollution and urbanization.

# Datasets

This study used a variety of different data sets for air pollution data and for indicators of urbanisation.

Sentinel 5-P datasets [19]:

Data from Sentinel 5-P’s TROPOMI sensor was used for air quality. Data for a variety of different pollutants was collected:

* Nitrogen Dioxide (NO2)
* Formaldehyde (HCHO)
* Carbon Monoxide (CO)
* Ozone (O3)
* Sulphur Dioxide (SO2)

Sentinel 5-P data is available from April-2018 to the present. The maximum resolution available is “5.5 x 3.5" km, a spatial resolution of 5.5 km in the satellite flight direction and 3.5 km in the perpendicular direction at nadir. Data released before 6 August 2019 had a resolution in the flight direction up to 7 km. To maintain uniformity, the lower resolution was used. All the data was in the form of L-2 products. It was obtained through the Sentinel Hub API.

Normalized Difference Vegetation Index (NDVI):

The Normalized Difference Vegetation Index (NDVI) is a widely used remote sensing index that measures the density and health of vegetation. It is calculated using the difference between near-infrared (which vegetation strongly reflects) and visible red light (which vegetation absorbs) from satellite imagery. NDVI values range from -1 to +1, where higher values indicate healthy, dense vegetation, while lower values signify sparse or stressed plant cover, barren land, or water bodies. Urban areas generally have less vegetation as compared to rural areas, so NDVI can be used as an indicator of urbanisation.[20]

This data was collected from the MODIS (Moderate Resolution Imaging Spectroradiometer) sensor. The available resolution was 250m. It was obtained through the Google Earth Engine API.

Nighttime Land Surface Temperature (NLST):

The Nighttime Land Surface Temperature (NLST) is a metric used to estimate the temperature of the Earth's surface using satellite data. By standardizing land surface temperature measurements, NLST helps assess thermal conditions across different regions and time periods. NLST is generally greater in regions with more concrete and human made structures. Hence it can be used as an indicator of urbanisation. [21]

This data was collected from the MODIS (Moderate Resolution Imaging Spectroradiometer) sensor.

The available resolution was 1000m. It was obtained through the Google Earth Engine API.

Nighttime Light (NTL):

Nighttime Lights (NTL) data refers to satellite observations of artificial lighting on the Earth's surface during nighttime. NTL captures the intensity and distribution of lights, which lets it act as an indicator for population density. As populations increase, number of lights generally do as well. [22]

This data was taken from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor.

The available resolution was 15 arc second (~500m at the Equator). It was obtained through the Earth Observation group website.[22]

# Data Preparation and Basic Analysis

Data was collected from a variety of different sources and in different resolutions. The time period under study was June 2018 to December 2023. All satellite datasets were available globally. The first step in processing the data was cropping it to our region of interest, the Hyderabad greater metropolitan area. This was done by specifying a bounding box using latitude and longitude coordinates ([78.00405826, 16.93264351, 79.04971836, 17.90150706]).

For uniform analysis, all data was aligned to the same temporal scale. The chosen scale was monthly. For Sentinel, the data was available daily. This was averaged down to monthly data. For NDVI, it was available for every 16 days. This was also averaged down to monthly data. NLST data was available daily. This was averaged down to monthly. NTL data was available as monthly data. This was not changed.

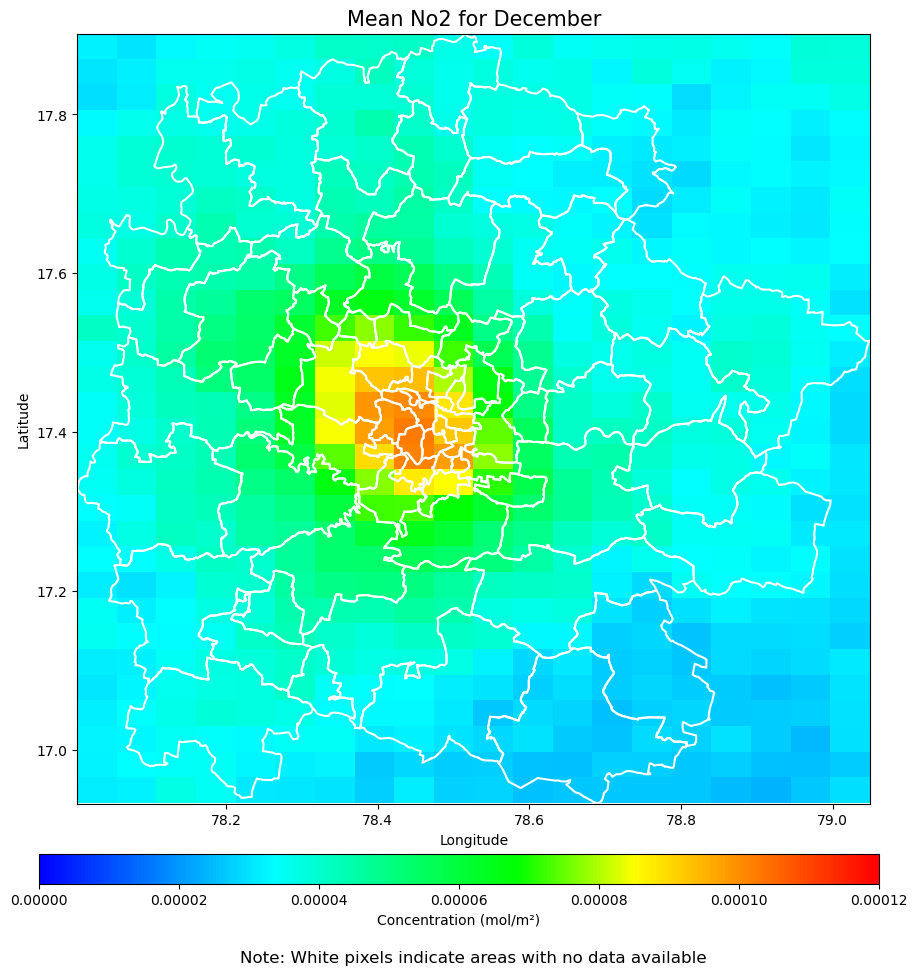
Sentinel data, NLST data, NTL data and NDVI data each had different resolutions. To compare them directly they had to be scaled up or down to the same resolution. As the region of interest was sufficiently large, and to prevent artifacts, down scaling to the lowest resolution was chosen. This was 7.5 x 3.5 km of the Sentinel data. Downscaling was done through average resampling using the rasterio python library.

In average resampling pixels from the original image are grouped to form a coarser grid. For each pixel in the new coarser grid the values of all the finer resolution pixels that fall within its bounds are averaged. This process involves computing the mean value of the original higher resolution data and assigning that mean to the corresponding coarser pixel. This process preserves the overall distribution of the original data. It also results in loss of some detail but it was necessary to maintain the same resolution across all datasets.

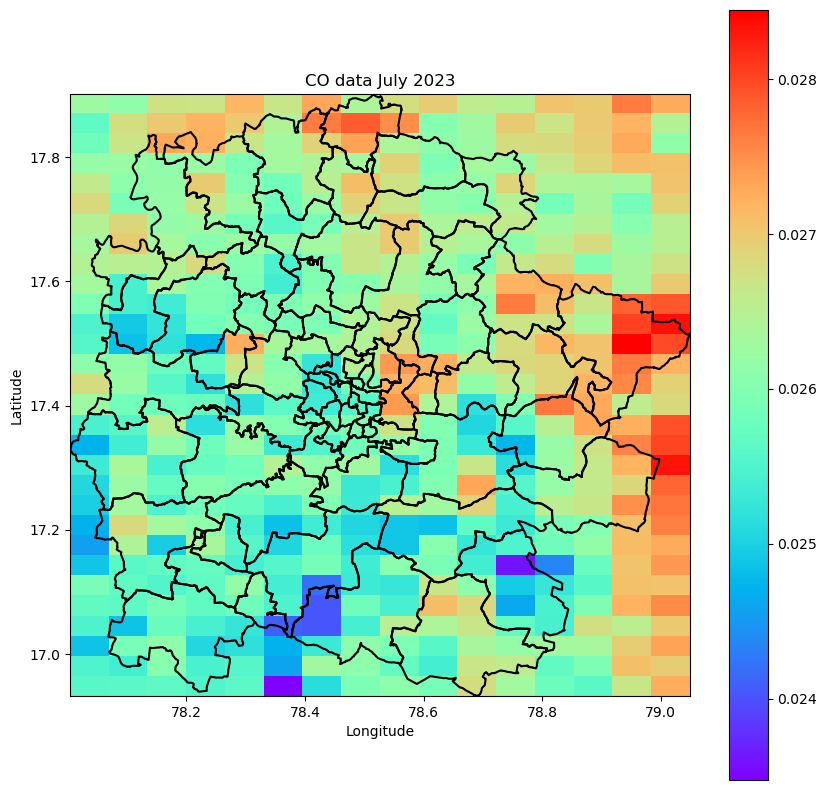
After scaling, the data for the region of interest was in the form of a 2d array or matrix with 16 columns and 30 rows. The columns represent longitude and the rows represent latitude. This gives a total of 40 datapoints for each month for each dataset.

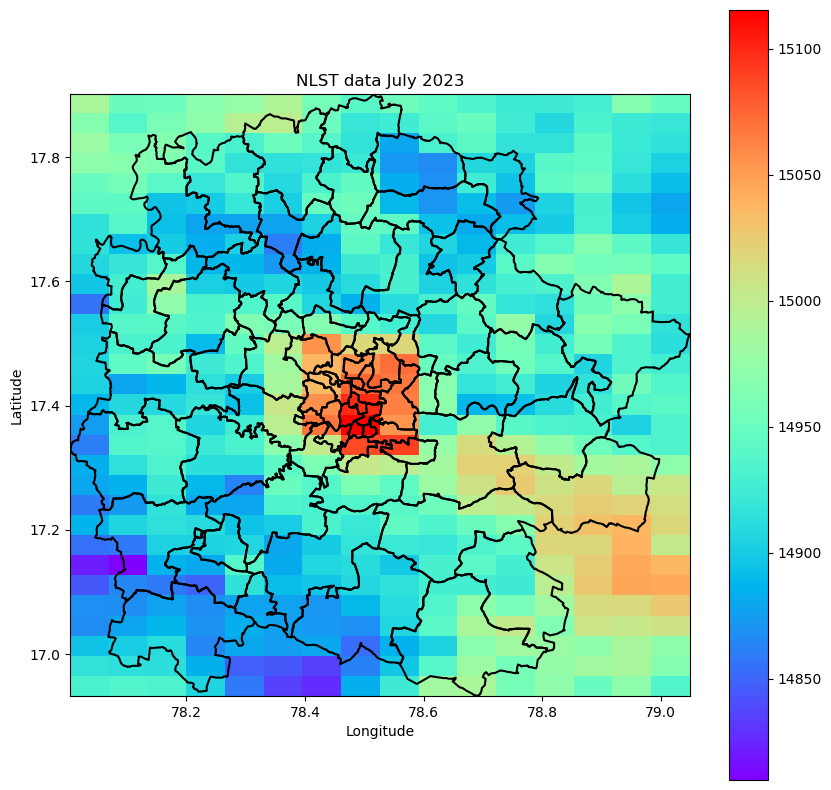
After the different datasets were cropped and scaled to the same spatial and temporal resolutions, they were visualised using heatmaps. These heatmaps give us a rough idea of the patterns of the different pollutants and indicators. It helps identify clusters and compare datasets. These heatmaps were plotted using matplotlib and they were overlaid on a map of the region of interest, Hyderabad for easier comparison.

Pollutant (NO2):

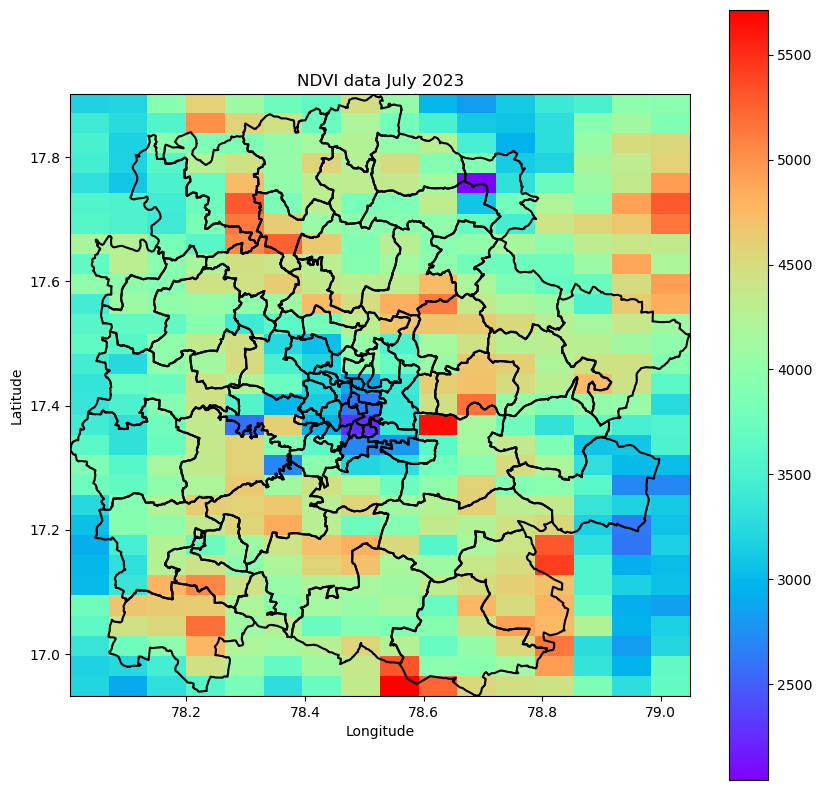


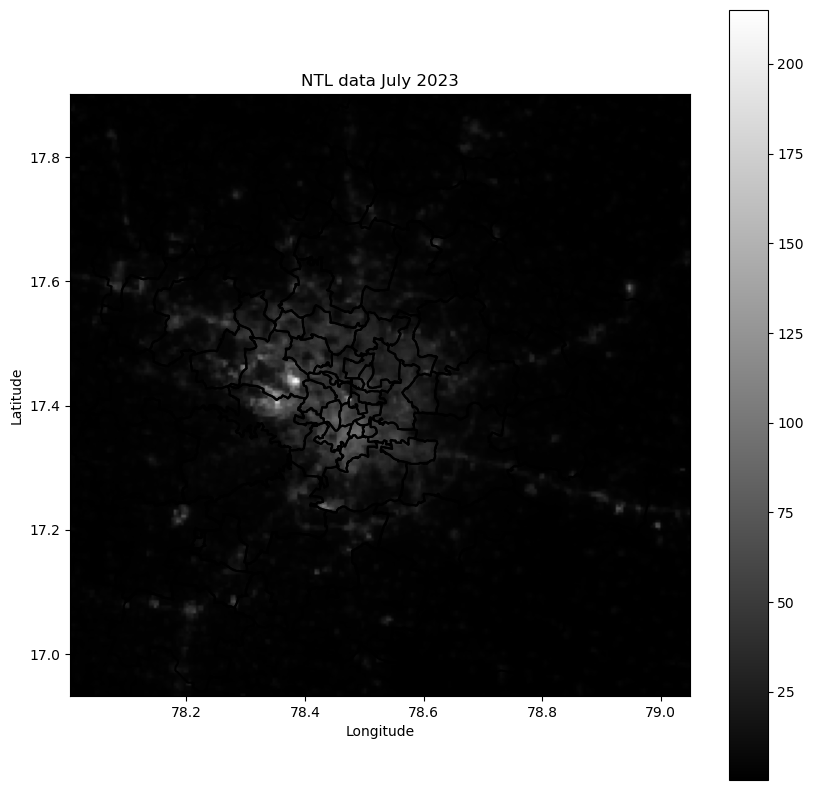
Pollutant (CO):



NLST: 

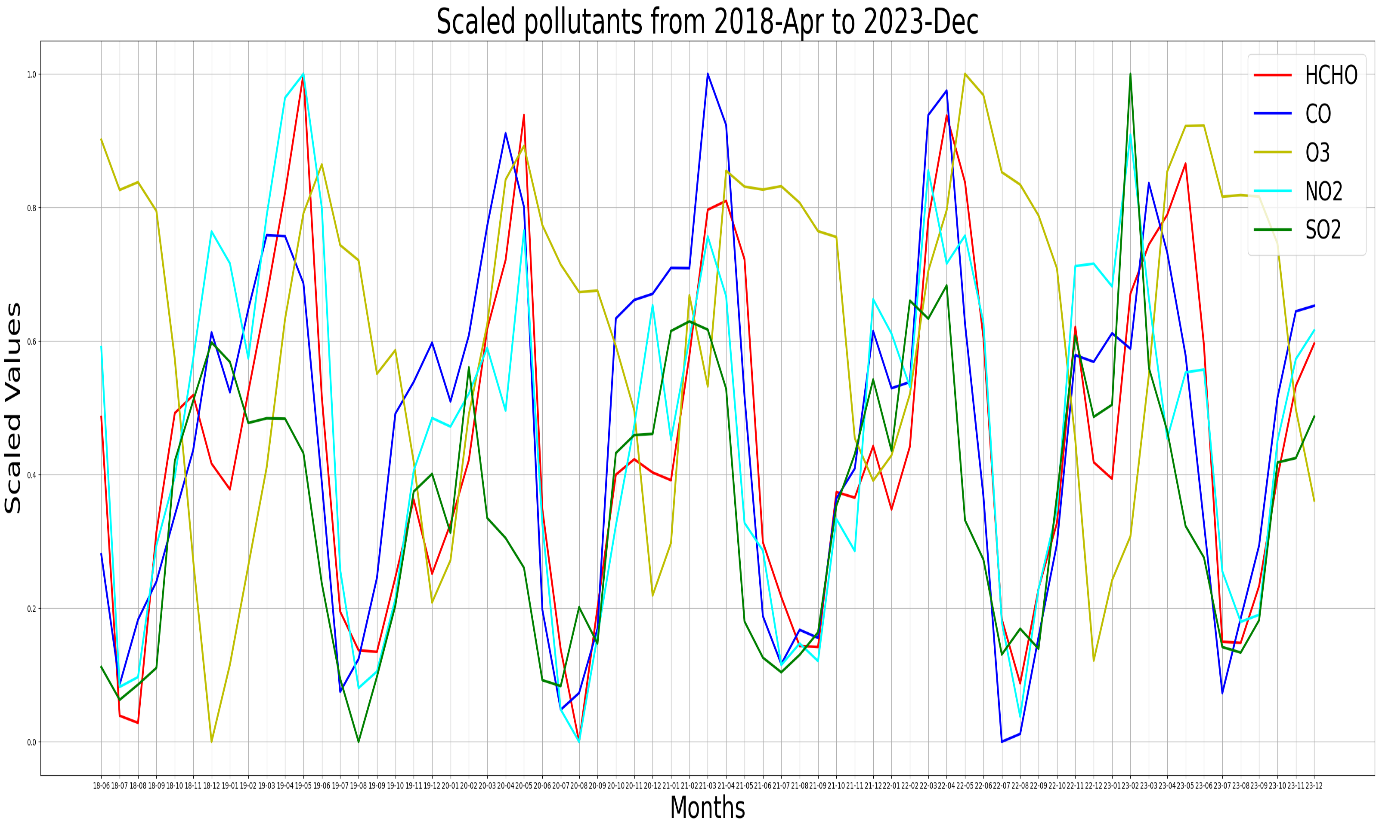
NDVI:



NTL: 

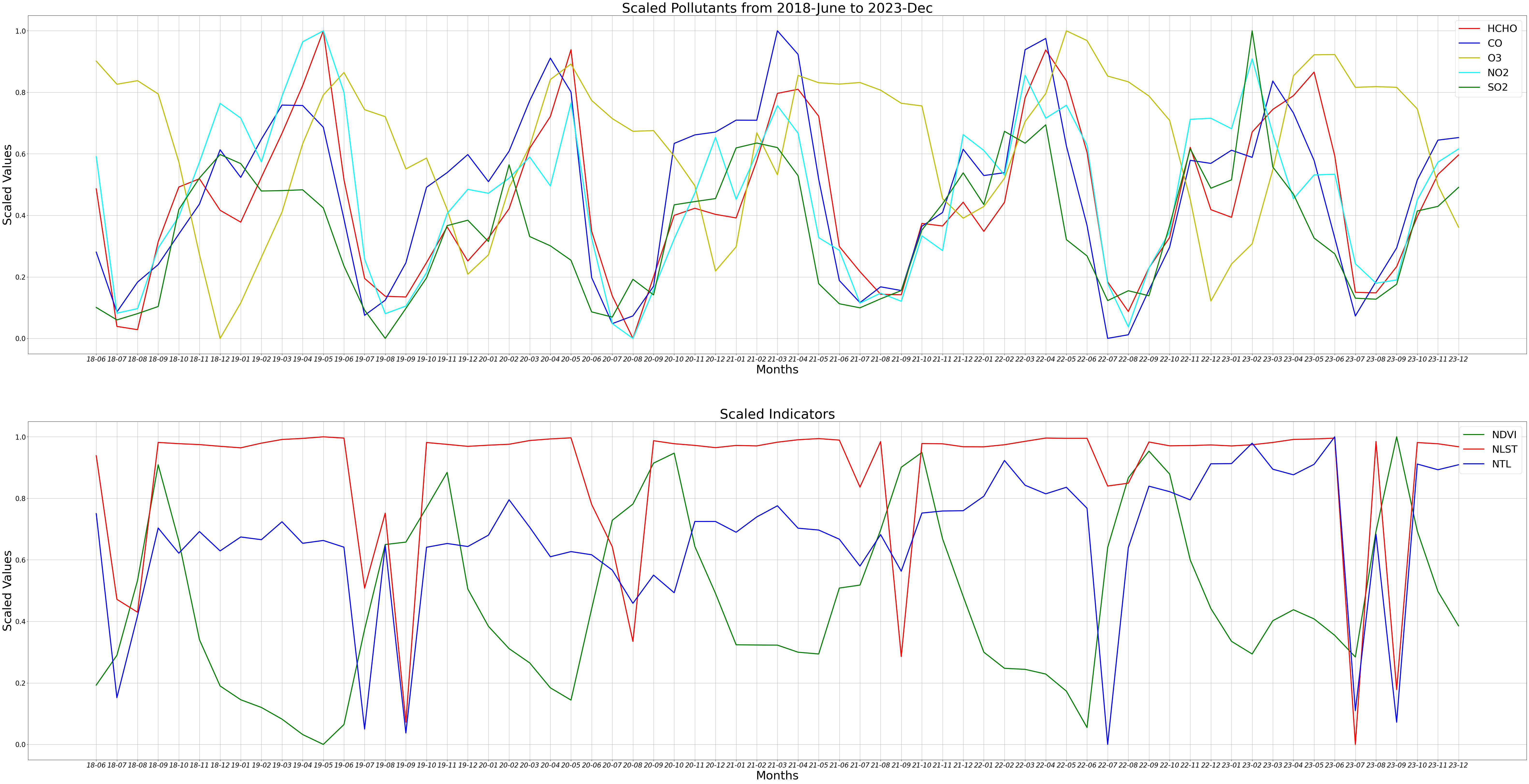
Each pollutant and indicator had a different scale and unit. To make them directly comparable, they were all scaled to values between 0 and 1. This was done by using min-max normalisation. Each value in the data had the following transformation applied:

After scaling, the average values for each month were plotted to help identify seasonal and yearly trends in both pollutants and indicators. Taking monthly averages gave a total of 67 datapoints to be plotted.



It is clear from the above graph that there is a seasonal trend in air pollution. Pollution increases on average during the winter months and falls during summer. O3 alone is slightly displaced from the other pollutants.

Similar line graphs were also made with the urbanisation indicators.



# Correlation Measures

After completing the basic data scaling and plotting, the next step was to analyze the relationship between the pollutants and environmental indicators using various correlation methods. The data after processing was available in the form of two-dimensional arrays. Each element within these 2D arrays represented a specific location, defined by its latitude and longitude, for a particular month. To make comparison directly possible across the variables, these 2D arrays were flattened into 1D arrays. This involved placing all the rows in order sequentially to get a 1-D array. This transformation allowed for the application of correlation techniques without altering the spatial relationships between the data points. As the same transformation was applied to each dataset, there was no change in the order of datapoints between sets. For instance, when comparing CO concentrations with NLST values, the i-th element in both arrays corresponded to the same geographical location and time period, ensuring that comparisons were valid and spatially consistent. After this transformation correlation could be directly calculated between the datasets.

## Spearman Correlation

Spearman correlation measures the strength and direction of a monotonic relationship between two ranked variables. Unlike Pearson correlation, Spearman does not assume that the relationship is linear or that the data is normally distributed. Instead, it assesses whether the variables tend to increase or decrease together, regardless of whether the change follows a straight line. It works by ranking the data points and then computing the Pearson correlation between the ranks. This makes it more appropriate for data that does not is not necessarily linear, but is still correlated.

The formula for Spearman's rank correlation coefficient ρ is:

Where:

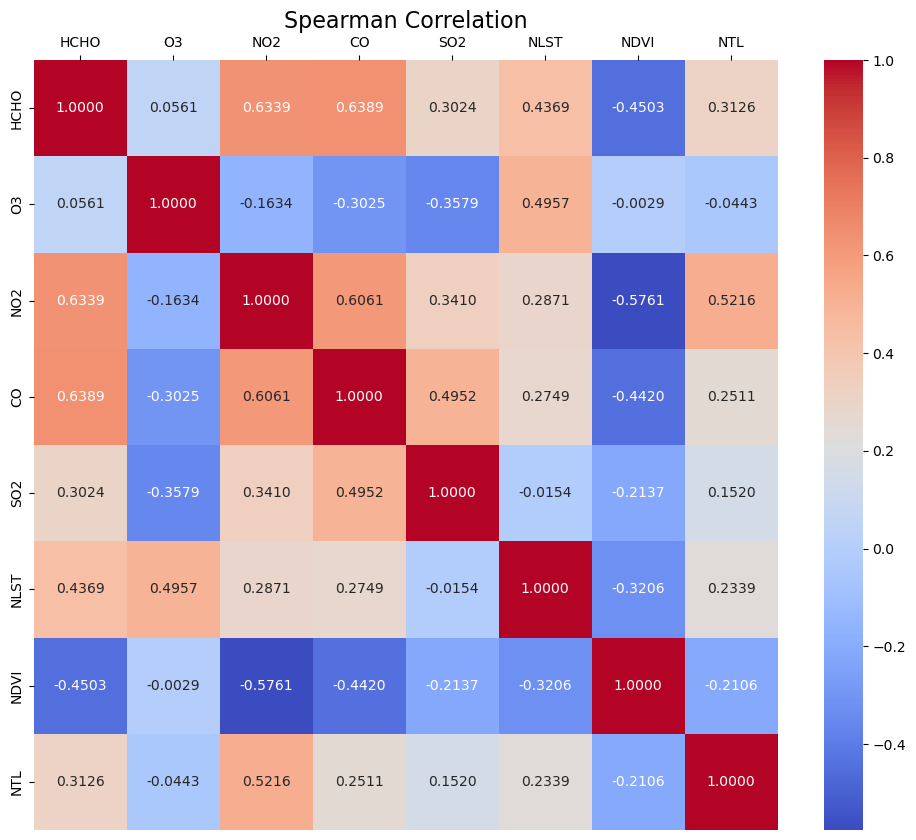
* ​ is the difference between the ranks of corresponding values of the two variables.
* is the number of data points.

Initially spearman correlation was calculated for each month individually, for each pair of pollutants and indicators. The average correlation over these individual months is given below.

Average Spearman correlation taken for each month individually:

|  |  |  |  |
| --- | --- | --- | --- |
| Pollutants | NLST | NDVI | NTL |
| CO | 0.341601 | -0.056234 | 0.025260 |
| NO2 | 0.148313 | -0.233280 | 0.547565 |
| O3 | 0.248842 | 0.070109 | -0.216006 |
| SO2 | 0.004142 | 0.003618 | 0.014581 |
| HCHO | 0.102634 | -0.090301 | 0.188029 |

It was also measured for a combined array for all 67 months (approximately 32000 values).



From the heatmap it is clear that most pollutants are positively correlated with NLST and NTL, and negatively correlated with NDVI.

## Pearson Correlation

Pearson correlation is a statistical measure that evaluates the linear relationship between two continuous variables. It quantifies how strongly the variables are related by producing a value between -1 and +1, where +1 indicates a perfect positive linear relationship, -1 signifies a perfect negative linear relationship, and 0 means no linear correlation. Pearson correlation assumes that the relationship between the variables is linear and that the data is normally distributed.

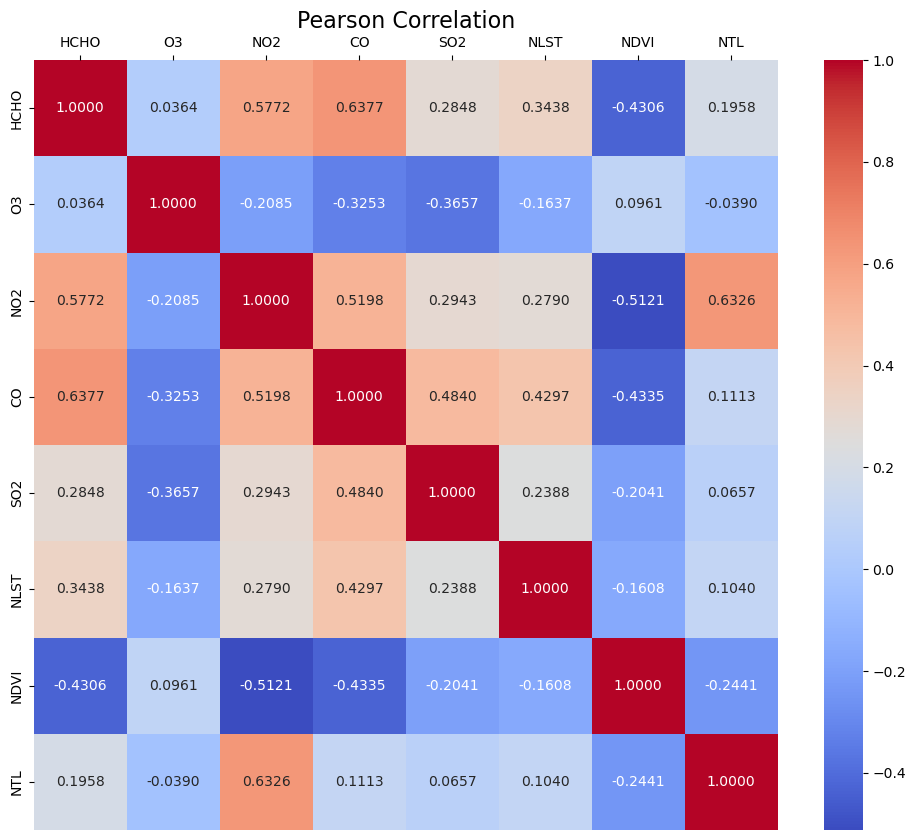
The formula for Pearson correlation coefficient r is:

Where:

* ​ and are the individual data points.
* and are the means of their respective datasets.
* The summation is over all data points.

Pearson correlation was also calculated on the combined array described above, consisting of about 32000 datapoints.

The results were similar, showing the same relationship between the pollutants and indicators.



## Multivariate Linear Regression

Multivariate Linear Regression is a machine learning technique that helps estimate a dependent variable in terms of separate independent variables. It assumes the relationship between the dependent and independent variables is linear.

For our purposes, the air pollutants were taken as the dependent variables and the indicators were considered independent variables.

The dataset used was the monthly spatial average of each pollutant and indicator, giving a total of 67 datapoints for each. All the datapoints were scaled between 0 and 1 to make them more uniform.

Using regression, the coefficients for NDVI, NLST and NTL are estimated to give a linear relationship between them and the pollutants.

Results:

HCHO = -0.4678 NDVI + 0.2586 NLST + 0.2632 NTL

O3 = 0.0562 NDVI + -0.0852 NLST + -0.1459 NTL

NO2 = -0.5805 NDVI + 0.2148 NLST + 0.2997 NTL

CO = -0.3974 NDVI + 0.2693 NLST + 0.3056 NTL

SO2 = -0.2130 NDVI + 0.0532 NLST + 0.4218 NTL

The regression was also performed on the average of all the pollutants, represented here as average air pollution.

Average Air pollution = -0.3205 NDVI + 0.1421 NLST + 0.2288 NTL

This also demonstrates a negative correlation between the air pollutants and NDVI as well as a positive correlation with NLST and NTL. This agrees with the results found by Pearson and Spearman correlation above.

# Citations

1. S. Simu et al., "Air Pollution Prediction using Machine Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2020, pp. 231-236. doi: 10.1109/IBSSC51096.2020.9332184.
2. H. Maleki, A. Sorooshian, G. Goudarzi, et al., "Air pollution prediction by using an artificial neural network model," Clean Techn. Environ. Policy, vol. 21, pp. 1341–1352, 2019. doi: <https://doi.org/10.1007/s10098-019-01709-w>.
3. D. Seng, Q. Zhang, X. Zhang, G. Chen, and X. Chen, "Spatiotemporal prediction of air quality based on LSTM neural network," Alexandria Engineering Journal, vol. 60, no. 2, pp. 2021-2032, 2021. doi: <https://doi.org/10.1016/j.aej.2020.12.009>.
4. D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao, and B. Zhang, "A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration," IEEE Access, vol. 7, pp. 20050-20059, 2019. doi: 10.1109/ACCESS.2019.2897028.
5. K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222-2232, Oct. 2017. doi: 10.1109/TNNLS.2016.2582924.
6. F. Terroso-Saenz, J. Morales-García, and A. Muñoz, "Nationwide Air Pollution Forecasting with Heterogeneous Graph Neural Networks," ACM Trans. Intell. Syst. Technol., vol. 15, no. 1, Art. 18, pp. 1-19, Feb. 2024. doi: <https://doi.org/10.1145/3637492>.
7. A. Dun, Y. Yang, and F. Lei, "Dynamic graph convolution neural network based on spatial-temporal correlation for air quality prediction," Ecological Informatics, vol. 70, p. 101736, 2022. doi: <https://doi.org/10.1016/j.ecoinf.2022.101736>.
8. S. Wang, S. Gao, S. Li, and K. Feng, "Strategizing the relation between urbanization and air pollution: Empirical evidence from global countries," Journal of Cleaner Production, vol. 243, p. 118615, 2020. [Online]. Available: <https://doi.org/10.1016/j.jclepro.2019.118615>.
9. S. Wang, X. Liu, X. Yang, B. Zou, and J. Wang, "Spatial variations of PM2.5 in Chinese cities for the joint impacts of human activities and natural conditions: A global and local regression perspective," Journal of Cleaner Production, vol. 203, pp. 143-152, 2018. doi: <https://doi.org/10.1016/j.jclepro.2018.08.249>.
10. M. A. Hassaan, S. M. Abdallah, E. S. A. Shalaby et al., "Assessing vulnerability of densely populated areas to air pollution using Sentinel-5P imageries: a case study of the Nile Delta, Egypt," Sci. Rep., vol. 13, p. 17406, 2023. doi: <https://doi.org/10.1038/s41598-023-44186-4>.
11. A. Gadakh, P. Shah, and C. Patel, "Characterising the pollution concentration in highly urbanized area: An application of remote sensing," Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLVIII-4/W5-2022, pp. 45–52, 2022. doi: <https://doi.org/10.5194/isprs-archives-XLVIII-4-W5-2022-45-2022>.
12. V. Safarianzengir, B. Sobhani, M. H. Yazdani et al., "Monitoring, analysis and spatial and temporal zoning of air pollution (carbon monoxide) using Sentinel-5 satellite data for health management in Iran, located in the Middle East," Air Qual. Atmos. Health, vol. 13, pp. 709–719, 2020. doi: <https://doi.org/10.1007/s11869-020-00827-5>.
13. G. Kaplan and Z. Yigit Avdan, "Space-borne air pollution observation from Sentinel-5p Tropomi: relationship between pollutants, geographical and demographic data," IJEG, vol. 5, no. 3, pp. 130–137, 2020. doi: 10.26833/ijeg.644089.
14. M. Khirwar and A. Narang, "GeoFormer: A vision and sequence transformer-based approach for greenhouse gas monitoring," arXiv, vol. abs/2402.07164, 2024. [Online]. Available: <https://arxiv.org/abs/2402.07164>.
15. L. Duan, Z. Jiang, and D. Carlson, "Augmenting Ground-Level PM2.5 Prediction via Kriging-Based Pseudo-Label Generation," arXiv preprint arXiv:2401.08061, 2024.
16. Y. Wang, J. Wang, M. Zhang, and L. Shi, "Spatial Correlation Analysis of Energy Consumption and Air Pollution in Beijing-Tianjin-Hebei Region," Energy Procedia, vol. 158, pp. 4280-4285, 2019. doi: <https://doi.org/10.1016/j.egypro.2019.01.797>.
17. D. de A. dos Santos, T. R. Lopes, F. M. Damaceno, and S. N. Duarte, "Evaluation of deforestation, climate change and CO2 emissions in the Amazon biome using the Moran Index," Journal of South American Earth Sciences, vol. 143, p. 105010, 2024. doi: <https://doi.org/10.1016/j.jsames.2024.105010>.
18. R. Bhushan, S. Hooda, H. Vidhani, M. Gupta, L. Suresh, and T. Clune, "Supervised Model for Peri-Urban Area Demarcation in Hyderabad, India," IEEE Geoscience and Remote Sensing Letters, 2024. doi: 10.1109/LGRS.2024.3359632.
19. "Copernicus Sentinel-5P," \*European Space Agency\*. [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/copernicus/sentinel-5p. [Accessed: Sep. 29, 2024].
20. "MOD13 - Vegetation Indices," \*NASA MODIS\*. [Online]. Available: https://modis.gsfc.nasa.gov/data/dataprod/mod13.php. [Accessed: Sep. 29, 2024].
21. "MOD11 - Land Surface Temperature and Emissivity," \*NASA MODIS\*. [Online]. Available: https://modis.gsfc.nasa.gov/data/dataprod/mod11.php. [Accessed: Sep. 29, 2024].