Mapping Air Pollution Sources with Sequential Transformer Chaining: A Case Study in South Asia

Hafiz Muhammad Abubakar[1], **Raahim Arbaz**[1], **Hasnain Ahmad**[1], **Mubasher Nazir**[2], **Usman Nazir**[1]
[1] Center for AI Research (CAIR), School of Computer and IT (SCIT), Beaconhouse National University
{f2021-641, f2021-556, f2021-597, usman.nazir}@bnu.edu.pk
[2] Legal Advisor, Solve Agri Pak Private Limited
mubashar.nazir@solveagripak.com

Abstract

This study presents a comprehensive framework for detecting pollution sources, specifically factory and brick kiln chimneys, in major South Asian cities using a combination of remote sensing data and advanced deep learning techniques. We first identify hotspots of Acute Respiratory Infections (ARI) by correlating health data with air pollutant concentrations, including sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO). For these identified hotspots, both low-resolution and high-resolution satellite imagery are acquired. Our approach employs a sequential process, beginning with a Vision Transformer model that utilizes high resolution satellite imagery and a broad range of text inputs with a lower confidence threshold to initially filter the data. This is followed by the application of the Remote CLIP model, which is run twice in succession using satellite imagery paired with different text inputs to refine the detection further. This sequential tranformer chaining filter out 99% of irrelevant data from high-resolution imagery. The final step involves manual annotation on the remaining 1% of the data, ensuring high accuracy and minimizing errors. Additionally, a novel multispectral chimney index is developed for detecting chimneys in low-resolution imagery. The study introduces a unique, annotated chimney detection dataset capturing diverse chimney types, which improves detection accuracy. The results provide actionable insights for public health interventions and support regulatory measures aimed at achieving the United Nations' Sustainable Development Goal 3 on health and well-being. We plan to make the dataset and code publicly available following the acceptance of this paper.

1 Introduction

According to the World Health Organization (WHO), approximately 7 million people die annually from respiratory infections, with over 90% of these deaths occurring in low- and middle-income countries (23). This high mortality rate is primarily due to exposure to fine particulate matter and other pollutants that penetrate deep into the lungs and cardiovascular system, adversely affecting respiratory health, particularly in children (21). This situation represents a significant challenge to achieving the United Nations Sustainable Development Goal (SDG) 3, which is aimed at ensuring good health and well-being.

Air pollution, a major contributor to these health issues, is caused by pollutants such as particulate matter, carbon monoxide, ozone, nitrogen dioxide, and sulfur dioxide (24). These pollutants often originate from industrial activities, including emissions from coal-fired power plants and other factories (12; 14; 13). To address these concerns, the United Nations Environment Programme has recommended 25 measures to reduce air pollution (Coalition). Furthermore, regulations such as the National

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2024.

Environmental Quality Standards (NEQS) from the Government of Pakistan and the National Clean Air Programme (NCAP) from the Government of India aim to control and reduce emissions of particulate matter and carbon monoxide (NEQS, Government of Pakistan; NCAP, Government of India).

In response to these critical issues, this study presents a comprehensive framework for detecting pollution sources, specifically focusing on factory and brick kiln chimneys across major South Asian cities. We start by identifying Acute Respiratory Infection (ARI) hotspots by correlating health data with high concentrations of pollutants including sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO). We acquire both low-resolution and high-resolution satellite imagery for these hotspots (see Appendix for details).

Our approach employs a sequential transformer chaining method to process this data. Initially, a Vision Transformer model (5) filters high-resolution imagery using a broad range of text inputs with a lower confidence threshold. This is followed by the Remote CLIP model (11), applied twice with different text inputs to further refine detection. This method filters out 99% of irrelevant data. The final step involves manual annotation of the remaining 1% of the data to ensure high accuracy (see Fig. 1).

Additionally, we develop a novel multispectral chimney index for detecting chimneys in low-resolution imagery and introduce a unique, annotated dataset that captures diverse chimney types (see Fig 1). This framework provides actionable insights for public health interventions and supports regulatory measures aimed at achieving SDG 3 on health and well-being. We plan to make the dataset and code publicly available following the acceptance of this paper.

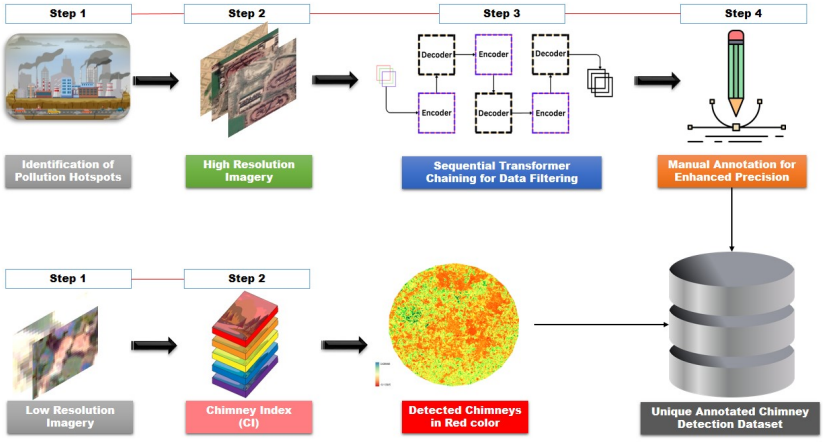


Figure 1: Overview of the Methodology for Pollution Source Detection

2 Methodology

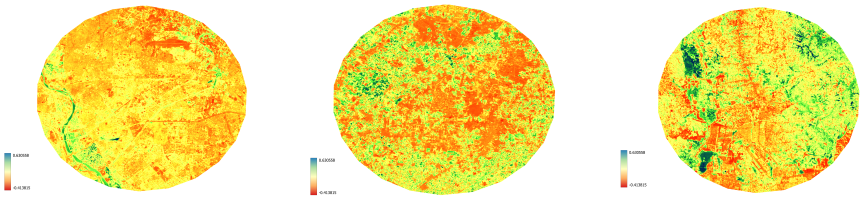
2.1 Chimneys detection using novel multi-spectral Chimney Index (CI)

This methodology enhances the detection of chimneys and industrial sites using a multi-index approach based on Sentinel-2 satellite imagery. The process introduces a novel Chimney Index (CI), which integrates multiple spectral indices to improve the accuracy of identifying industrial structures.

$$CI = (\omega_1 \times (1 - NDVI)) + (\omega_2 \times BI) + (\omega_3 \times BUI) \quad (1)$$

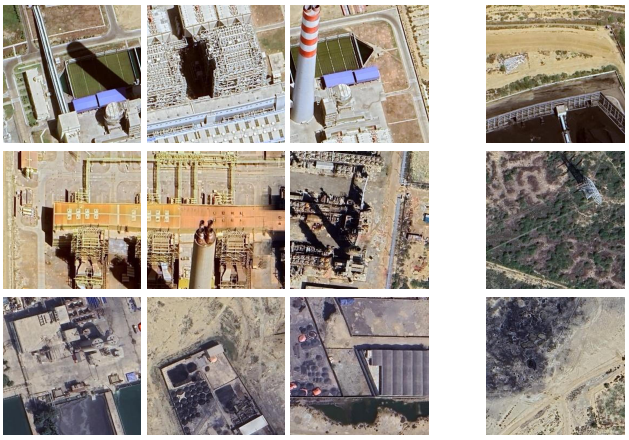
where NDVI is the Normalized Difference Vegetation Index, BI is the Burn Index, and BUI is the Built-Up Index. The weights $\omega_1, \omega_2, \omega_3$ are assigned to each index to balance their contributions according to their relevance for detecting chimneys and industrial sites.

The Chimney Index is calculated by combining three key spectral indices: the Normalized Difference Vegetation Index (NDVI), the Burn Index (BI), and the Built-up Index (BUI). This composite index leverages the unique spectral signatures associated with industrial areas, particularly those containing chimneys. Low NDVI values indicate areas with minimal vegetation, characteristic of industrial



(a) (b) (c)

Figure 2: Chimneys detection using novel multi-spectral Chimney Index (CI) (a) Delhi Industrial Estate, India; (b) Sundar Industrial Estate, Pakistan; and (c) Tongi Industrial Estate, Bangladesh. Red colors indicate areas with a high Chimney Index, reflecting regions with significant industrial activity.



Detected Chimneys Missed Chimneys

Figure 3: Chimneys detection using sequential transformer chaining mechanism: Illustrative examples on South Asian industrial regions.

zones. High BI values suggest the presence of burned or heat-affected areas, which can be indicative of industrial activity. High BUI values identify built-up urban areas, where industrial structures are likely to be located. By integrating these indices, the Chimney Index provides a more robust and accurate method for detecting industrial sites and chimneys. This approach allows for efficient large-scale screening of satellite imagery, significantly enhancing the ability to identify and map industrial structures across extensive urban landscapes in South Asia that mainly produce pollutants causing respiratory infections.

2.2 Chimneys detection using sequential transformer chaining mechanism

This methodology began with collecting a large dataset of high-resolution satellite images from densely populated urban areas. The dataset then underwent a series of increasingly refined filtering stages. Initially, a base multimodal Vision Transformer model (**?**) was used to perform a broad classification with multiple text prompts and a low confidence threshold, ensuring a wide inclusion of potential industrial structures. The resulting subset of images was then subjected to secondary filtering using Remote CLIP (11), which specifically targeted chimneys with a higher confidence threshold to enhance precision. A tertiary filtering process followed, again utilizing Remote CLIP to further isolate images containing smokestacks. The filtered dataset was meticulously reviewed manually to verify accuracy. This multi-stage approach, which combines advanced deep learning techniques with thorough manual curation, resulted in a highly specific and accurate dataset of industrial structures.

3 Results and Discussion

In this study, we evaluated the effectiveness of two key methodologies (see Fig. 2 & 3) for detecting industrial structures such as chimneys: the Chimney Index and the Sequential Transformer Chaining

mechanism. The overall accuracy of these methods was assessed based on their performance in identifying chimneys within three major industrial estates: Delhi, Sundar, and Tongi.

3.1 Chimney Index Accuracy

The Chimney Index, which utilizes combined criteria of NDVI, Burn Index (BI), and Built-up Index (BUI), was applied with a threshold value of 0.4. The detection rates for the three industrial estates were as follows:

- Patparganj Industrial Area, Delhi (India): Detection rate of approximately 79.17%.
- Sundar Industrial Estate, Lahore (Pakistan): Detection rate of 68.42%.
- Tongi Industrial Estate, Dhaka (Bangladesh): Detection rate of 70%.

These detection rates reflect the ability of the Chimney Index to identify industrial areas with a significant presence of chimneys.

3.2 Sequential Transformer Chaining Accuracy

The Sequential Transformer Chaining mechanism, which involves a multi-stage process starting with a Vision Transformer model and followed by Remote CLIP models, was evaluated for its accuracy in refining chimney detection. The overall performance of this method was assessed as follows:

- Initial Vision Transformer Filtering: Captured a broad range of potential industrial structures with a precision of of 0.45 and Recall of 0.90. The accuracy is lower because we set a low confidence threshold to ensure that the Vision Transformer captures all potential chimneys, which increases the likelihood of false positives. These false positives are subsequently filtered out by the Remote CLIP model.
- Secondary Remote CLIP Filtering: Improved detection precision with a confidence threshold, resulting in an accuracy of 85% for chimneys.
- Tertiary Remote CLIP Refinement: Further refined the dataset, achieving an accuracy of 85% for smokestacks.
- Final Manual Annotation: Verified the remaining 1% of data, resulting in a final accuracy of 95% for detecting chimneys and smokestacks.

4 Conclusion

This study introduces a robust framework for detecting pollution sources, specifically factory and brick kiln chimneys, in major South Asian cities by combining remote sensing data with advanced deep learning techniques. Key contributions include the identification of pollution hotspots through ARI and air pollutant correlation, a sequential transformer chaining process for precise data filtering, and the development of a novel multispectral chimney index. Additionally, the creation of a unique, annotated dataset significantly enhances detection accuracy. This framework provides valuable insights for environmental monitoring and public health interventions, with the dataset and methods to be made publicly available for further research.

References

[1] Anderson, H. R. et al. (2018). Air pollution and health: A review. *Environmental Health Perspectives*.

[2] Boyd, D. S. et al. (2018). Using satellite data to inform health interventions: a systematic review. *Remote Sensing in Public Health*.

[Coalition] Coalition, C. . C. A. Air pollution measures for asia and the pacific. [UN Air Quality Measures for Asia].

[4] Cohen, A. J. et al. (2017). The global burden of disease due to outdoor air pollution. *Environmental Health Perspectives*.

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[6] Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al. (2012). Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36.

[7] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

[8] Hansen, M., Potapov, P., Moore, R., Hancher, M., Turubanova, S., Tyukavina, A., Thau, D., Stehman, S., Goetz, S., Loveland, T., et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853.

[9] Huang, L. et al. (2019). Social disparities in air pollution exposure: A systematic review. *Environmental Research Letters*.

[10] Jensen, J. R. (2005). *Introductory Digital Image Processing: A Remote Sensing Perspective*. Prentice Hall.

[11] Liu, F., Chen, D., Guan, Z.-R., Zhou, X., Zhu, J., and Zhou, J. (2023). Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16.

[12] Mahlangeni, N., Kapwata, T., Laban, T., and Wright, C. Y. (2024). Health risks of exposure to air pollution in areas where coal-fired power plants are located: protocol for a scoping review. *BMJ open*, 14(3):e084074.

[13] Mittal, M. L., Sharma, C., and Singh, R. (2012). Estimates of emissions from coal fired thermal power plants in india. In *2012 International emission inventory conference*, pages 13–16.

[14] Munsif, R., Zubair, M., Aziz, A., and Zafar, M. N. (2021). Industrial air emission pollution: potential sources and sustainable mitigation. In *Environmental Emissions*. IntechOpen.

[NCAP, Government of India] NCAP, Government of India. National Clean Air Programme (NCAP). [India National Clean Air Programme].

[NEQS, Government of Pakistan] NEQS, Government of Pakistan. National Environmental Quality Standards (NEQS). [Pakistan Air Quality Standards].

[17] Newell, K., Kartsonaki, C., Lam, K. B. H., and Kurmi, O. P. (2017). Cardiorespiratory health effects of particulate ambient air pollution exposure in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Planetary Health*, 1(9):e368–e380.

[18] Norton, A. et al. (2020). Satellite monitoring of air quality: Impacts on public health. *Environmental Monitoring and Assessment*.

[19] Patel, S. et al. (2017). Modeling the impact of air pollution on public health with machine learning techniques. *Journal of Environmental Management*.

[20] Ricotta, C., Avena, G., and Palma, A. D. (1999). Mapping and monitoring net primary productivity with avhrr ndvi time-series: statistical equivalence of cumulative vegetation indices. *ISPRS Journal of Photogrammetry & Remote Sensing*, 54:325–331.

[21] Tabaku, A., Bejtja, G., Bala, S., Toci, E., and Resuli, J. (2011). Effects of air pollution on children's pulmonary health. *Atmospheric Environment*, 45(40):7540–7545.

[22] Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150.

[23] World Health Organization (WHO) (2016). *World Health Statistics 2016 [OP]: Monitoring Health for the Sustainable Development Goals (SDGs)*. World Health Organization.

[24] World Health Organization (WHO) (2022). Ambient (outdoor) air pollution. Accessed: 2024-06-06.

[25] Zhang, L. et al. (2019). Associations between air pollution and respiratory health outcomes: A meta-analysis. *Environmental Pollution*.

[26] Zhang, X., Hu, Y., Zhuang, D., Oi, Y., and Ma, X. (2009). Ndvi spatial pattern and its differentiation on the mongolian plateau. *Journal of Geographical Sciences*, 19:403–415.

A Literature Review

Previously, it has been well documented that air pollutants like particulate matter (PM2.5) have been a major cause of premature deaths with especially high burden on South Asian Countries due to their severe effects on cardiorespiratory system(17). Not only does these air pollutants deteriorate the health and environment, but also cause economic burden on countries.The National Environment Agency (NEA) Air Quality Monitoring from Singapore and among others in the Asia Pacific region have taken impactful measures towards the mitigation of these deadly air pollutants To mitigate the effects of these air pollutants-in this study, we identify illegal gaseous emissions in different cities of South Asia using gaseous emissions data from Sentinel-5P satellite and heat signature values.

Recent advancements in remote sensing technologies and health data analytics have provided a multidimensional approach to evaluating respiratory health, particularly through the integration of satellite-derived air quality metrics. The relationship between air pollution and respiratory diseases has been extensively documented, with studies highlighting the detrimental effects of pollutants such as particulate matter (PM), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2) on public health (25).

Satellite remote sensing offers a unique capability to monitor air quality at a large scale, providing critical data that can complement ground-based measurements. For instance, the use of Moderate Resolution Imaging Spectroradiometer (MODIS) and Sentinel-5P data has been instrumental in capturing temporal variations in air pollutants and their spatial distribution across urban and rural landscapes (4). These datasets enable researchers to identify pollution hotspots and correlate them with health outcomes in exposed populations (18).

Integrating satellite-derived air quality metrics with health data allows for more comprehensive epidemiological studies. Recent research has demonstrated that exposure to elevated levels of airborne pollutants is associated with increased incidence of respiratory conditions such as asthma and chronic obstructive pulmonary disease (COPD) (1). By employing machine learning algorithms, researchers can analyze complex datasets to reveal patterns and predict health outcomes based on air quality metrics (19).

The socio-economic disparities in exposure to air pollution and subsequent health effects underscore the need for targeted public health interventions. Studies have shown that vulnerable populations, including children and the elderly, are disproportionately affected by poor air quality (9). Utilizing health data from sources like the Demographic and Health Surveys (DHS) alongside satellite data can facilitate the identification of these at-risk groups, thereby informing policy decisions aimed at improving air quality and respiratory health (2).

The integration of satellite-derived air quality metrics with health data presents a promising avenue for understanding and mitigating the impacts of air pollution on respiratory health. Future research should focus on refining data collection methods and enhancing the predictive capabilities of models that incorporate these multidimensional datasets, ultimately contributing to improved public health outcomes.

Remote sensing has evolved significantly over the past few decades, leveraging satellite imagery to monitor and detect various environmental and anthropogenic changes. The use of satellite imagery for detection purposes spans multiple disciplines, including environmental science, urban planning, and public health. This literature review aims to provide an overview of the current remote sensing techniques using satellite imagery for detection purposes, with a focus on methodologies, applications, and advancements in the field.

Multispectral and Hyperspectral Imaging
Multispectral sensors capture data at several specific wavelength bands, while hyperspectral sensors

acquire data across a continuous spectrum of wavelengths. These techniques are crucial for identifying and analyzing different materials and land cover types based on their spectral signatures. For instance, multispectral and hyperspectral imaging are widely used in agriculture to monitor crop health, detect diseases, and manage irrigation (10).

Overview of Normalized Difference Vegetation Index (NDVI)
The Normalized Difference Vegetation Index (NDVI) is a widely utilized method for assessing crop health in agriculture. Due to the spatial variability in soil properties, different areas within a field may need varying amounts of nitrogen to optimize yield (20). NDVI data points can be used in conjunction with geostatistical methods to create spatial continuity surfaces, which help in developing precision agriculture strategies (26).

Remotely sensed images are defined by their spatial and spectral resolutions. Spatial resolution refers to the pixel arrangement in relation to each other in an image, while spectral resolution indicates the variation within pixels based on different wavelengths. The NDVI of a multispectral remote sensing image depends on these resolutions. Differences in reflectivity across various spectral bands are crucial for identifying features in remotely sensed images. Vegetation can be identified using remote sensing data because of its distinct absorption in the red and blue parts of the visible spectrum, its higher green reflectance, and its very strong reflectance in the near-infrared spectrum (20; 26). Among various indices used to highlight vegetation areas on remote sensing images, NDVI is the most common and widely adopted.

The NDVI is a widely used index to assess vegetation health and cover. It is calculated using the red and near-infrared bands of multispectral imagery. NDVI has been extensively used to monitor agricultural lands, forest cover, and urban green spaces. This index helps in identifying areas of healthy vegetation and those under stress due to various factors such as drought or disease (22).

Environmental Monitoring
Remote sensing provides key mechanisms whereby deforestation, desertification and biodiversity can be monitored. A good example is the use of Landsat imagery to map deforestation in Amazon rainforest and give essential data for conservation initiatives. Monitoring land cover change through time can be helpful in the management and conservation of natural resources (8).

National governmental and non-governmental organizations who have utilized similar change detection methods at the landscape scale illustrate how Landsat, along with other satellites used for continuous monitoring can serve to process our understanding of deforestation dynamics, illegal logging activities as well as assess the performance of conservation policies. This kind of real-time data is crucial for governments and other organizations in order to make quick interventions or mitigate the scars from deforestation. Land cover changes happening around the land may have some human activities warrants for an environmental regulation that are prompted by satellite data collecting.

Public Health
Satellite imagery is increasingly used in public health to monitor environmental factors affecting health. NDVI and other indices have been employed to study the correlation between green spaces and respiratory health in urban areas. The presence of green spaces has been linked to lower levels of air pollutants and improved respiratory health outcomes.

High-Resolution Satellites
The launch of high-resolution satellites such as Sentinel-2, WorldView-3, and PlanetScope has revolutionized remote sensing by providing finer spatial resolution. This allows for more detailed analysis of small-scale features and phenomena (6). High-resolution imagery is particularly valuable in urban planning, environmental monitoring, and disaster response .

Cloud Computing and Big Data
The advent of cloud computing platforms like Google Earth Engine has facilitated the processing and analysis of large volumes of satellite imagery. These platforms enable researchers to perform complex analyses at a global scale without the need for significant local computing resources (7). Cloud computing has democratized access to remote sensing data, allowing for more collaborative and large-scale studies.

B Study Region

Since Acquiring satellite images for the entire study region would've been a monumental task we settled on one city from each country where the concentration of our study gasses were the highest. We ended up selecting chimneys from Lahore , Delhi and Dhaka cities from three South Asian countries. We assumed that these are highly dense areas and their impact can be used as base for any other region in these countries.

Algorithm 1: Detection of Integrating DHS Data for Geo-Coded Acute Respiratory Infection Data

Input : DHS datasets ← [ARI]
Output : Geo-Coded Loactions of ARI

Function `Main`($datasets$)*:*
    DHSKRData ← Load Children's Health Data;
    DHSSpatialData ← Load Geo-coded Health Data;
    ARI ← `ProcessDHSData`(DHSKRData, DHSSpatialData);
    Locations ← `ExtractLocations`(`MergeData`(*DHS-Data,DHSSpecial-Data*));
    return Locations;

Function `ProcessDHSData`(DHSKRData, DHSSpatialData)
    `ApplySPSS`(DHSKRData);
    ARI ← `MergeData`(DHSKRData, DHSSpatialData);
    return ARI;

Function `ApplySPSS`(DHSKRData)
    Apply SPSS script from GitHub repository to calculate ARI;

Function `MergeData`(DHSKRData, DHSSpatialData)
    Merge ARI data with Geo-coded data using cluster ID;
    return *Geo-coded ARI Data*;

C Respiratory Infection Detection

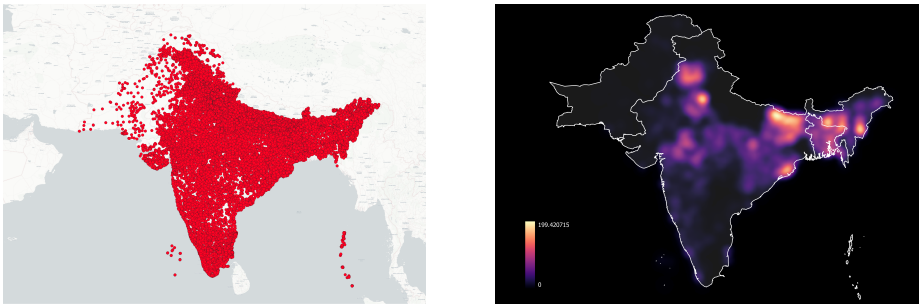
To address the sources of air pollution, particularly from brick kilns and factory chimneys, we developed a systematic approach utilizing various geo-coded datasets. The algorithm begins with the loading of essential datasets, which include the Children's Health Data from the Demographic and Health Surveys (DHS) and geo-coded health data. The Children's Re-code dataset, which provides critical health indicators, is processed alongside the spatial data to calculate the prevalence of Acute Respiratory Infections (ARI) within the population. This processing involves applying specific scripts to ensure the data's integrity and merging the ARI data with geographic information using cluster IDs, thereby creating a geo-coded ARI dataset.

To retrieve data for Acute Respiratory Diseases (ARD) in children, we began by accessing the Children's Health Data from the Demographic and Health Surveys (DHS). This dataset includes vital health indicators relevant to child health across various countries. Specifically, we utilized the Children's Re-code dataset, which contains detailed information on health outcomes and demographic characteristics.

The first step involved selecting the appropriate phase of the DHS. We focused on Phase 71 for Bangladesh, India, and Pakistan to ensure consistency across the data. In instances where Phase 71 was unavailable, we opted for the closest available survey to maintain the integrity of our analysis.

Next, the datasets were retrieved in SPSS format from the DHS. We employed specific statistical scripts to process the Children's Re-code data, which allowed us to apply necessary transformations and calculations for the ARI indicators. Subsequently, we merged the processed ARI data with the geo-coded health data using cluster IDs. This integration was crucial, as it provided spatial context to the health outcomes, resulting in a comprehensive geo-coded ARI dataset.

In Figure 4 (a), the raw data represents all survey locations, providing a comprehensive view of where health assessments have been conducted across the South Asia regions. This dataset serves as the foundation for our analysis. In contrast, figure 4 (b) highlights the filtered ARI data, where locations with elevated rates of Acute Respiratory Infections are depicted as red clusters. These clusters indicate areas with significant public health concerns, allowing for targeted interventions of respiratory diseases among children in these regions.



(a) DHS Survey Locations (b) Locations For Acute Respiratory Infections

Figure 4: (a): Illustration of the geographic distribution of the survey locations from the Demographic and Health Surveys (DHS); (b): The filtered data for Acute Respiratory Infections (ARI).

D Detection of Gaseous Indicators Using Sentinel-5P

The Sentinel-5 Precursor (Sentinel-5P) mission, launched by the European Space Agency (ESA), is dedicated to monitoring atmospheric composition. The mission's TROPOMI (Tropospheric Monitoring Instrument) is designed to collect data essential for assessing air quality and tracking various atmospheric pollutants. This section outlines the methodology for retrieving and analyzing atmospheric concentrations of nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO) using Sentinel-5P data.

Algorithm 2: Calculation of Gases Indicators (CO, NO2, SO2) in South Asia using Sentinel-5P

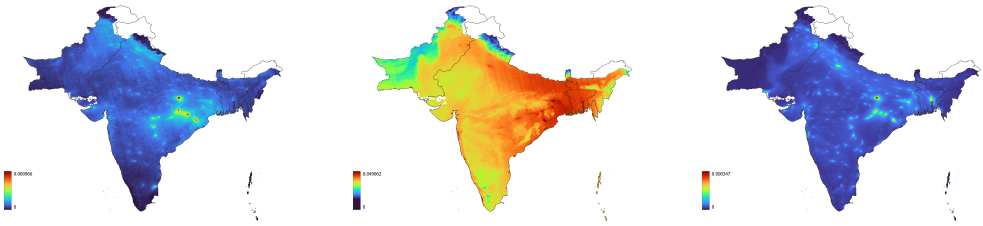
Input : Geo-coded datasets ← [Sentinel-5P (CO, NO2, SO2)], [Area of Interest (South Asia)], [Time Period (Start Date, End Date)]
Output : Mean concentrations of CO, NO2, and SO2 in South Asia

```
Function Main(SouthAsiaAOI, StartDate, EndDate)
    COData ← LoadData(COPERNICUS/S5P/OFFL/L3_CO, SouthAsiaAOI, StartDate, EndDate);
    NO2Data ← LoadData(COPERNICUS/S5P/OFFL/L3_NO2, SouthAsiaAOI, StartDate, EndDate);
    SO2Data ← LoadData(COPERNICUS/S5P/OFFL/L3_SO2, SouthAsiaAOI, StartDate, EndDate);
    COMean ← CalculateMean(COData, SouthAsiaAOI);
    NO2Mean ← CalculateMean( NO2Data, SouthAsiaAOI);
    SO2Mean ← CalculateMean( SO2Data, SouthAsiaAOI);
    return COMean, NO2Mean, SO2Mean;
Function LoadData( Collection, AOI, StartDate, EndDate)
    Data ← .filterDate(StartDate, EndDate);
    Data ← .filterBounds();
    return Data;
Function CalculateMean( Data, AOI)
    MeanData ← .mean();
    MeanData ← .clip();
    return MeanData;
```

ROPOMI instrument onboard of the Sentinel-5P satellite forCO, NO_2, and SO_2 concentrations specifications A multispectral sensor, TROPOMI observes the reflection at important wavelength bands for studying atmospheric gases and aerosols with a relatively high resolution of 0.01 arc degrees Data are extracted from online repositories and portals, filtered temporally and spatially to necessary periods or the study region. The application of TROPOMI's data quality flags prevent some reliability issues with systematic errors.

The methodology for emissions from gas indicators (CO, NO_2, and SO_2) in South Asia is discussed below. Step1: The geo-coded datasets are extracted from Sentinel-5P for the selected locations and dates. Datasets for each gas are loaded separately, filtered by time and location in South Asia. This function, LoadData (filterType), is used to filter the datasets and then load them. Finally, the CalculateMean function is used to calculate the average concentrations of each gas which clips it down in both location (to South Asia) and time across all grid cells using dataset. This gives us a representative idea of pollutant concentration levels during the Defined Time Period The final output - the estimated air pollutant concentrations for CO, NO_2, and SO_2 mean concentration values will help in quantifying their spatio-temporal distribution - supports informed decision-making practices

regarding environmental policies as well as strategies aiming to improve ambient air quality across South Asia.



(a) SO_2 (b) CO (c) NO_2

Figure 5: Spatial distribution of air pollutants across the study area.

The results of the data analysis in Figure: 5 are visualized using various mapping and graphing tools, including interactive maps and heatmaps that effectively display the spatial distribution of NO_2, SO_2, and CO. Complementing the spatial analyses, graphs are employed to illustrate variations in the concentrations of these pollutants. These graphs provide insights into how pollutant levels fluctuate, helping to identify increased pollution areas.

E Detection of Pollutant Emission Sources using NDVI

For detecting the sources of these gaseous emissions in the atmosphere we used a technique known as The Normalized Difference Vegetation Index (NDVI). NDVI is a measure used in remote sensing to assess the presence and condition of vegetation. It is calculated using the difference between

Algorithm 3: Detection of Industrial Zones in South Asia using NDVI

Input : Geo-coded datasets ← [Sentinel-2 Imagery], [Area of Interest (South Asia)]
Output : Identified Industrial Zones in South Asia

```
Function Main(SouthAsiaAOI, 2017-01-01, 2024-06-30)
    Images ← DownloadImages(SouthAsiaAOI, 2017-01-01, 2024-06-30);
    NDVIImages ← CalculateNDVI(Images);
    IndustrialZones ← IdentifyIndustrialZones(NDVIImages);
    StoreResults(IndustrialZones);
    return IndustrialZones;
Function DownloadImages(SouthAsiaAOI, 2017-01-01, 2024-06-30)
    ← Sentinel-2.filterDate(2017-01-01, 2024-06-30);
    ← .filterBounds(SouthAsiaAOI);
    return;
Function CalculateNDVI(Images)
    NDVIImages ← {};
    for each in Images do
        ← .select(NIR);
        ← .select(Red);
        ← .subtract().divide(.add()).rename('NDVI');
        NDVIImages.append();
    return NDVIImages;
Function IdentifyIndustrialZones(NDVIImages)
    IndustrialZones ← {};
    for each in NDVIImages do
        ← .lt(0.2);
        IndustrialZones.append();
    return IndustrialZones;
```

the near-infrared (NIR) and red (visible) light reflected by vegetation. The multi spectral bands from Copernicus Sentinel-2 satellite were utilized to calculate the NDVI. To get better results in the imagery, cloud masking technique was used along with temporal data from only built up areas so that the industrial area can be easily filtered out. The temporal data taken was from 2017 till 2024. The cloud masking technique helps us select the best clear images taken from the satellite that do not have clouds.

The NDVI is calculated as :

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \tag{2}$$

The NDVI is calculated with the resolution of 10 metre per pixel which means that every pixel in the image showing the NDVI index is actually covering an area of 10 metres as seen from the satellite.