GeoAI ground-level NO2 estimation

Jiayou Shen
jiayoush@usc.edu

Archana Bhatia
archanam@usc.edu

Abstract

Our project focuses on the development of robust machine learning models, which will be used for NO2 ground-level concentration predictions based on remote sensing data from Google Earth Engine and in-situ measurements from air quality monitoring stations. The main goal is to develop a model that can predict NO2 concentration with high spatial resolution across areas and temporal resolution across different meteorological conditions, seasons, and locations. Key elements of the approach include the use of convolutional neural networks (CNN) and long short-term memory (LSTM) to capture both spatial and temporal patterns in the data, while addressing challenges related to missing values and varying spatial resolutions across datasets. Model performance shall be evaluated based on Pearson correlation, RMSE, and MAE criteria because it gives a high predictive accuracy and generalization is guaranteed. The specific application of this work will be for the accurate estimation of NO2 levels to facilitate environmental monitoring and air quality management, thereby improving our understanding of pollution dynamics in urban settings.

1 What is the problem?

The detection and prediction of ground-level nitrogen dioxide (NO_2) is critically important for several reasons, particularly in relation to public health and effective policy-making. NO_2 is a harmful air pollutant, primarily produced by the combustion of fossil fuels in vehicles, power plants, and industrial facilities. Exposure to elevated levels of NO_2 results in various diseases, such as asthma, bronchitis, and even increased risks of cardiovascular diseases[1]. Therefore, predicting ground-level NO_2 concentrations is essential to safeguarding public health and preventing health issues in communities.

However, Nitrogen dioxide prediction faces several challenges. Firstly, the temporal variability of ground-level NO_2 concentration is significant due to factors like seasons, aerosol concentration and meteorological conditions like temperature and humidity. These non-linear variations will make it difficult to construct the model. Besides, due to the limitations in detection conditions and levels, some NO_2-related data in the dataset may be missing or contain errors. Therefore, the robustness of the model is required to handle these problems.

Obviously, predicting ground-level NO_2 concentrations would benefit various groups. For example, public health authorities and policymakers would gain valuable information, allowing them to implement timely interventions to protect communities from harmful exposure. Environmental organizations would have more precise data to raise awareness and advocate for cleaner technologies. Additionally, people with respiratory conditions would benefit from access to real-time air quality predictions, enabling them to take preventive measures, such as avoiding outdoor activities when pollution levels are high.

2 How is it currently approached?

Numerous studies have explored machine learning and deep learning methods to predict ground-level nitrogen dioxide concentrations. Long et al.[2], using data from the China National Environmental Monitoring Center, applied four tree-based algorithms: decision trees (DT), gradient boost decision tree (GBDT), random forest (RF), and extra-trees (ET). Their findings revealed that the extra-trees (ET) algorithm outperformed the other methods in predicting NO_2 concentrations across China, demonstrating superior accuracy. Chi et al.[3] conducted a study utilizing NO_2 concentration data from six different regions in China, implementing the XGBoost model. This research demonstrated that the XGBoost approach provided high-quality estimations of surface NO_2 exposure. Additionally, Li and Wu[4] used data from NASA's Goddard Earth Sciences Data and Information Services Center and the Ministry of Ecology and Environment of China, employing a full residual network to estimate missing satellite-derived NO_2 data. This deep learning model achieved higher accuracy than traditional machine learning methods when predicting surface NO_2 levels. These studies highlight the effectiveness of advanced machine learning and deep learning models in improving the accuracy of surface NO_2 concentration predictions.

However, given the inherent temporal variability of ground-level NO_2 concentration data, current prediction models are limited in their ability to fully capture the dynamic nature of NO_2 distribution over time. While tree-based models, such as extra-trees and XGBoost, and deep learning approaches, such as full residual networks, have demonstrated strong performance, they often lack the ability to effectively model time-dependent patterns and trends in air quality data. NO_2 concentrations can fluctuate due to a variety of factors, including meteorological conditions, human activities, and seasonal changes, all of which exhibit complex temporal dynamics. Therefore, our approaches will consider the temporal and spatial properties to fit the characteristics of our dataset.

3 How do you plan to approach it?

After studying the given data and the existing models for the task at hand, we will proceed with data preprocessing in which we will deal with missing and redundant data. Some parameters like LST (Land Surface Temperature) have 46% missing values, while NO2_trop (Tropospheric NO_2) has 41% missing values. It would be crucial to handle this missing data through interpolation, imputation, or using models that can deal with incomplete data. The spatial resolution for different parameters varies, with Precipitation data at 5566 meters, and other data at 1133 meters or 1000 meters. The spatial data will likely be interpolated from different resolutions before feeding it to the models.

Furthermore, model selection will require critical consideration of both spatial and temporal parameters. Convolutional Neural Networks (CNNs) prove to be useful tools for handling spatial data. We can use CNNs to extract spatial features from remote sensing inputs, which can then be correlated with NO_2 levels. On the other hand, Long Short-Term Memory (LSTM) networks can model the time-series aspect of NO_2 data, incorporating the historical data to improve predictions. We can also combine the LSTM layers with spatial features extracted by CNNs to create a spatio-temporal model. However, given the spatial variability of the data (different resolutions and missing data), we will have to first pre-process the spatial data to handle different resolutions, then use LSTM for temporal forecasting. In order to capture both spatial and temporal variations together, we will use 3D CNNs that are capable of learning from both dimensions. Spatio-Temporal Graph Convolutional Networks (ST-GCNs) are also designed to handle spatio-temporal data in a more structured way, especially if the data has an underlying graph structure (e.g., locations connected by weather patterns or pollution transport).

For baseline comparison, we will use XGBoost combined with imputation techniques. XGBoost can be used to handle missing data and model both temporal and spatial interactions. This method can provide a benchmark for more complex models.

The evaluation metrics that will be used are Pearson correlation®, Root mean square error (RMSE), and Mean Absolute Error (MAE).

Based on the complexity of the provided dataset, the only obstacle we can foresee is to build adaptable models that are capable of estimating NO2 levels in different weather conditions, locations, and seasons, while maintaining high accuracy.

The competition ends on Nov 11, 2024, so we have approximately 7 weeks to complete the project. Following is a rough project timeline based on the deadline.

- Week 1 (Sept 22 - Sept 28): Literature review
- Week 2 (Sept 29 - Oct 5): Data-preprocessing
- Week 3 (Oct 6 - Oct 12): Feature Selection
- Week 4 (Oct 13 - Oct 19): Model development and implementation
- Week 5 (Oct 20 - Oct 26): Comparison of model results
- Week 6 (Oct 27 - Nov 2): Hyper-parameter tuning
- Week 7 (Nov 3 - Nov 9): Testing/Running experiments

References

[1] Gurjar, B. R., Jain, A., Sharma, A., Agarwal, A., Gupta, P., Nagpure, A. S., & Lelieveld, J. (2010). Human health risks in mega-cities due to air pollution. *Atmospheric Environment*, 44(36), 4606-4613.

[2] Long, S., Wei, X., Zhang, F., Zhang, R., Xu, J., Wu, K., Li, Q., & Li, W. (2022). Estimating daily ground-level NO2 concentrations over China based on TROPOMI observations and machine learning approach. *Atmospheric Environment*.

[3]Chi, Y., Fan, M., Zhao, C., Yang, Y., Fan, H., Yang, X., ... & Tao, J. (2022). Machine learning-based estimation of ground-level NO2 concentrations over China. *Science of The Total Environment*, 807, 150721.

[4] Li, L., & Wu, J. (2021). Spatiotemporal estimation of satellite-borne and ground-level NO2 using full residual deep networks. *Remote Sensing of Environment*, 254, 112257.