

PREDICTING NITROGEN DIOXIDE POLLUTANT LEVEL FOR USA USING SATELLITE DATA

Tahmina Akter Anondi, Subhajit Chakrabarty
Department of Computer Science
Louisiana State University Shreveport, Shreveport, LA

Corresponding Author: Tahmina Akter Anondi
Email: tahmina.akter.anondi@gmail.com
doi: 10.34107/UDUK9890436

ABSTRACT

Accurate and timely forecasts of ground-level nitrogen dioxide (NO_2) are essential for both air quality management and public health protection. Sentinel-5P TROPOMI provides daily global tropospheric NO_2 column retrievals, but conversion to surface concentrations requires additional modeling. Previous research has demonstrated feasibility at the regional and national scale in Europe and Asia, typically reporting $R^2 \approx 0.6–0.8$, but often limited in geography, predictor range, or validation design. We develop a reproducible workflow in Google Earth Engine to merge Sentinel- 5P near real-time NO_2 with daily meteorological and vegetation predictors (land surface temperature, relative humidity, wind components, MODIS-derived NDVI/EVI) at EPA monitors in the United States (2019–2023). The resulting dataset included ~880,000 station-day observations. We benchmarked models systematically from linear regression and ridge regression to ensemble trees (Random Forest, HistGradientBoosting, XGBoost), neural networks (MLP), and stacked ensembles. The evaluation was done under several regimes: random hold-out, leave-one-location-out (LOLO), leave- one-year-out (LOYO), and spatio-temporal validation. Random hold-out yielded excellent results (R^2 as high as 0.85 for stacked ensembles). LOYO also showed good temporal generalization ($R^2 \approx 0.82$ for MLP), whereas LOLO showed poor spatial transferability ($R^2 \approx 0.22–0.42$). Spatio-temporal validation confirmed medium accuracy robustness ($R^2 \approx 0.75$). Compared to prior Sentinel-5P studies, this work contributes national-scale U.S. coverage, an open and reproducible GEE workflow, heterogeneous predictors, and explicit spatial and temporal cross-validation. The findings indicate that while very good accuracy is possible at the level of known sites and across years, spatial generalization remains the main limitation. This reproducible framework provides a scalable template for extension to other pollutants ($PM_{2.5}$, O_3) and domains, serving both epidemiological research and regulatory applications.

Keywords: Nitrogen dioxide, Prediction, Satellite, Remote sensing, Google Earth Engine

INTRODUCTION

Ambient air pollution is a significant environmental health risk, causing respiratory and cardiovascular disease worldwide. Nitrogen dioxide (NO_2), emitted mainly from fossil-fuel combustion, is linked to asthma, heart disease, and death. NO_2 is also a precursor to ground-level ozone and particulate matter, leading to ecosystem damage and additional health impacts. Detailed mapping of NO_2 concentrations is crucial for both policy and public health, especially in cities where NO_2 concentrations are usually highest. Ground-based monitoring of NO_2, however, is sparse. US regulatory networks have just a few hundred operating NO_2 monitors for the entire nation, with most clustered in cities. This sparse, uneven coverage leaves large rural and suburban gaps that are unmonitored, making it hard to pick up on fine-scale pollution patterns. Also, fixed monitors sample at a point and cannot easily capture spatial heterogeneity within a city. As a result, ground networks alone under-represent population exposure in most regions. Satellite