International Journal of Advances in Data and Information Systems
Vol. 7, No. 2, August 2026, pp. 761~771
ISSN: 2721-3056, DOI: 10.25008/ijadis.v7i2.1602

Enhancing Prediction of Satellite-Based NO_2 Concentration in ASEAN Using Random Forest Models with Spatial Coordinates and Temporal Lag Features

Nabil Naufal[1], Anik Djuraidah[2], Yenni Angraini[3]
[1,2,3]Study Program of Statistics and Data Science, Bogor Agricultural University, Indonesia

Article Info

Article history:
Received Apr 30, 2026
Revised June 18, 2026
Accepted June 25, 2026

Keywords:
Random Forest
NO_2 Concentration
Satelite Remote Sensing
Spatio-temporal Features
ASEAN

ABSTRACT

Nitrogen dioxide (NO_2) is a critical indicator of anthropogenic emissions, making its monitoring essential for environmental management in the rapidly urbanizing ASEAN region. While satellite imagery provides the necessary high spatial coverage to overcome the limitations of sparse ground-based stations, traditional machine learning models applied to these datasets often overlook the inherent spatial heterogeneity and temporal persistence of air pollutants. This study aims to quantify the marginal contributions of geographic coordinates and temporal lag features to the prediction accuracy of satellite-derived NO_2 concentrations using Random Forest (RF) models. By evaluating four progressive RF configurations across 1,500 locations from July 2018 to December 2024, the study identifies the optimal feature combination for regional air quality modeling. The results demonstrated that the RF model incorporating both geographic coordinates and a 12-month lag variable achieved the best performance, yielding an R^2 of 0.832 and an RMSE of 5.42×10^6. Feature importance analysis revealed that the 12-month lag of NO_2, nighttime lights, and location were the most influential predictors, highlighting the strong annual seasonality and the impact of economic activities on pollution levels. These findings provide a robust, data-driven framework for regional air quality monitoring and policy formulation in developing tropical regions.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Nabil Naufal,
Study Program of Statistics and Data Science,
Bogor Agricultural University,
Raya Dramaga Street, IPB University Campus, Dramaga, Bogor 16680, West Java, Indonesia.
Email: mrnabilnaufal@apps.ipb.ac.id

1. INTRODUCTION

Nitrogen dioxide (NO_2) is one of the most significant atmospheric pollutants in the context of rapid economic development, functioning as a direct indicator of anthropogenic activity driven by fossil fuel combustion from the transportation and industrial sectors [1]. The concentration of NO_2 in the troposphere is highly sensitive to human activities, a relationship made strikingly clear during the COVID-19 pandemic, when mobility restrictions led to an observed reduction in NO_2 concentrations of 27–34% in major urban centers worldwide [2]. Beyond its role as an economic indicator, NO_2 poses substantial public health risks, contributing to respiratory diseases, premature mortality, and increased cardiovascular morbidity [3].

The ASEAN region is currently experiencing one of the most rapid phases of urbanization and economic transformation globally, with direct implications for air quality. Long-term satellite observations have recorded consistent annual increases in NO_2 concentrations in ASEAN metropolitan areas, ranging from 0.8% to 7.7% [4]. This trend is closely tied to the expansion of

motor vehicle ownership and industrial activity accompanying rapid population growth in urban centers across the region [5]. The case of Yangon, Myanmar, where tropospheric NO_2 increased sharply from 0.18 ppb to 2.29 ppb during its economic transition period, exemplifies the acute sensitivity of this pollutant to macro-level developmental shifts [5]. Collectively, these trends position NO_2 monitoring as a critical component of environmental governance and public health protection in the region.

Effective air quality monitoring at regional scale requires data that are spatially consistent and temporally comprehensive. Ground-based monitoring stations provide high measurement accuracy but suffer from sparse and geographically uneven deployment, particularly in developing nations across ASEAN [6]. The Sentinel-5P satellite, carrying the TROPOMI (TROPOspheric Monitoring Instrument) sensor, has fundamentally transformed this landscape by providing global NO_2 column density data at a spatial resolution of 3.5×3.5 km², enabling pollution monitoring at sub-urban scale across continuous geographic extents [7], [8]. Importantly, the use of satellite-derived grid-based data reduces aggregation bias that commonly arises when using administrative boundary-based measurements, offering a more spatially objective analytical framework [9].

Machine learning approaches, particularly Random Forest (RF), have been increasingly adopted for environmental prediction tasks due to their ability to model complex nonlinear relationships and handle high-dimensional input features without distributional assumptions [10], [11]. The fundamental capability of RF to handle spatial data complexities was demonstrated in [12], showing that the algorithm effectively captures nonlinear patterns and complex variable interactions, making it highly advantageous for spatial modeling. However, existing applications of RF to satellite-based NO_2 estimation commonly treat each observation as spatially and temporally independent, failing to account for two fundamental properties of atmospheric pollutants. First, NO_2 exhibits strong spatial autocorrelation, where concentrations at nearby locations tend to be more similar than those farther apart, reflecting shared emission sources and atmospheric dispersion patterns [13]. Second, NO_2 demonstrates pronounced temporal persistence, where current concentrations are strongly influenced by those from preceding months due to the accumulation and slow dissipation of atmospheric pollutants and their drivers [14]. Studies incorporating lag features into machine learning models for air quality prediction have demonstrated RMSE reductions of 25–39%, underscoring the importance of temporal memory in these systems [15]. Despite these findings, no study to date has systematically compared the relative and combined contributions of spatial coordinate features and temporal lag variables within a unified RF framework applied to large-scale satellite data across the ASEAN region.

This study addresses this gap by constructing and evaluating four progressive variants of the Random Forest model. The specific objectives are: (1) to quantify the marginal contribution of geographic coordinates and temporal lag features to prediction accuracy for satellite-derived NO2 concentrations; (2) to identify the optimal feature configuration among the four model variants; and (3) to determine the relative importance of each feature type through feature importance analysis

2. RESEARCH METHOD

This study employs a quantitative, comparative modeling framework to evaluate the contribution of spatial and temporal features to Random Forest prediction accuracy for satellite-derived NO_2 concentrations across ASEAN. The workflow proceeds through three stages: (1) exploratory data analysis, (2) construction of four RF model variants, and (3) model evaluation and feature importance analysis. All analyses were conducted in R version 4.x using the ranger and spdep packages.

2.1. Data and Study Area

This study covers the entire geographic extent of the ten ASEAN member states (Brunei Darussalam, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, and Vietnam). The analysis is based on 1,500 sampling points, selected to ensure adequate spatial coverage of both urban and rural environments across diverse land use types within the region.

All data used in this study are satellite-derived secondary products acquired and processed through the Google Earth Engine (GEE) cloud computing platform. Table 1 provides a summary of

all variables, their sources, and their native spatial resolutions. The observation period spans 78 months from July 2018 to December 2024, with July 2018 representing the beginning of stable operational data provision from the Sentinel-5P TROPOMI instrument. Monthly temporal aggregation was applied to all products to synchronize their differing orbital cycles, reduce data volume, and minimize the proportion of missing values attributable to cloud cover in the tropical environment.

Table 1. Summary of satellite-based data sources used in this study

Code	Variable	Operational Definition & Source	Native Resolution
Y	NO2 Concentration	Tropospheric NO_2 Column Density (mol/m²). Sentinel-5P / TROPOMI	3.5 × 3.5 km²
X1	Economic Activity	Nighttime Lights radiance (nW/cm²/sr). Proxy for electricity consumption and economic activity. NOAA VIIRS (VNP46A1)	500 m²
X2	Vegetation Health	Normalized Difference Vegetation Index (NDVI). Indicator of green land cover and carbon absorption capacity. MODIS (MOD13Q1)	250 m²
X3	Population Density	Gridded population density (persons/km²). WorldPop	100 m²
S	Spatial Coordinates	Longitude and Latitude of the centroid of each 10 km × 10 km grid cell	—

A grid resolution of 10 km × 10 km was adopted as the common spatial unit for all analyses. This resolution was chosen to match the effective spatial scale of the primary response variable, as the native resolution of Sentinel-5P TROPOMI (~3.5 km²) represents the finest meaningful scale for NO_2 column density. Resampling other products to a finer resolution would artificially inflate spatial precision without adding substantive information. Furthermore, NO_2 concentrations represent atmospheric column averages that naturally smooth over small distances due to wind-driven dispersion, rendering sub-10 km spatial variation analytically unreliable. A 10 km grid therefore represents the optimal balance between spatial detail and physical representativeness.

2.2. Data Prepocessing and Exploratory Data Analysis

Optical satellite sensors, including both the Sentinel-5P TROPOMI and MODIS instruments, are subject to data gaps arising from cloud cover, instrument calibration periods, and satellite overpass limitations. This is particularly acute in the tropical ASEAN region, where persistent cloud cover can render entire months of data unavailable for specific grid cells. A two-stage hybrid imputation procedure was applied to address this challenge.

In the first stage, linear temporal interpolation was applied to grid cells with isolated missing values, defined as gaps of up to two consecutive months bounded by valid observations on both sides. Under this approach, the missing value for grid cell i at time t was estimated as the arithmetic mean of the observations at t−1 and t+1 for the same cell [16]. This method is appropriate when the proportion of missing data is small and temporal variation between adjacent periods is smooth, conditions that are generally satisfied for monthly NO_2 data outside of exceptional meteorological events [17].

In the second stage, spatial neighborhood substitution was applied to grid cells exhibiting longer consecutive gaps of three or more months, for which temporal interpolation would introduce excessive smoothing bias. Missing values were replaced with the spatial mean of all valid observations among the immediate neighbors of the target cell for the same calendar month [18]. This approach exploits the strong spatial autocorrelation characteristic of atmospheric pollutants, whereby neighboring cells share similar emission source profiles and dispersion conditions.

Prior to modeling, missing values resulting from cloud cover in the optical satellite sensors were addressed using a two-stage hybrid imputation procedure, combining linear temporal interpolation and spatial neighborhood substitution. Following data completeness, temporal characteristics were examined using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify temporal persistence in NO2 concentrations. The determination of lag features in this study is based on the autocorrelation patterns observed in the ACF and PACF plots. Similar to the methodology applied by Maharani and Angraini [19] in modeling the Rupiah exchange rate using the ARIMA-GARCH approach, where ACF and PACF plots were utilized to assess data stationarity and identify significant lags for determining the ARIMA model order. Adopting this principle, lags that exhibit autocorrelation values beyond the

significance threshold in the ACF plot are considered to retain critical information regarding current values. Consequently, these lags (specifically lag-12) were constructed as lag features and subsequently employed as predictor variables in the Random Forest model.

2.3. Random Forest

Random Forest is a nonparametric ensemble learning algorithm that constructs B decision trees on bootstrap samples of the training data, with each tree built using a randomly selected subset of m predictors at each split node [20]. The final prediction is the average across all B trees:

$$\hat{Y}(x) = \left(\frac{1}{B}\right)\sum_{b=1}^{I} f_b(x) \quad (1)$$

where $f_b(x)$ is the prediction of the b-th tree.

This study utilizes Random Forest to model satellite-based NO2 concentrations by considering spatial information (longitude and latitude coordinates) and temporal aspects (lag features). The application of Random Forest in the context of spatial data aligns with Pradana, Djuraidah, and Soleh [12], who modeled land-use changes in Kubu Raya Regency, West Kalimantan, using Logistic Regression, Random Forest, and Additive Logistic Regression. In their study, Random Forest successfully captured nonlinear patterns and spatial variable interactions to classify land-use changes. Building on these findings, this study adopts Random Forest as the primary model to map and predict NO_2 concentrations across spatial grids, utilizing coordinates and NO_2 lag values derived from ACF analysis.

To isolate and quantify the marginal contribution of spatial coordinates and temporal lag features, four model variants were constructed, each representing a distinct feature configuration. All four models were trained using identical default hyperparameters (number of trees B = 500, mtry = p/3 where p is the number of predictors, minimum node size = 5) to ensure that observed performance differences are attributable solely to differences in feature composition rather than to model tuning. This controlled design directly addresses the primary research question of feature contribution. Table 2 presents the complete specification of each model variant.

Table 2. Specification of four Random Forest model variants

Model	X1, X2, X3	Longitude & Latitude	NO2 Lag-12
RF_base	✓	✗	✗
RF_lag12	✓	✗	✓
RF_coords	✓	✓	✗
RF_coords_lag12	✓	✓	✓

2.5. Data Partitioning

The full dataset was partitioned chronologically into a training set (covering July 2018 to Desember 2023) and a test set (covering January 2024 to December 2024). Chronological splitting is essential for spatiotemporal data to prevent data leakage, as random partitioning would allow the model to learn from future observations during training, producing artificially inflated performance estimates that would not generalize to prospective forecasting scenarios [21].

2.6. Model Evaluation Metrics

Model performance on the held-out test set was evaluated using three complementary metrics. Root Mean Square Error (RMSE) quantifies prediction accuracy in the original units of the response variable, assigning greater penalty to larger errors [22]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

The coefficient of determination (R^2) measures the proportion of variance in the response variable explained by the model [23]:

$$R^2 = 1 - \left[\frac{SS_{res}}{SS_{tot}}\right] \quad (3)$$

Mean Absolute Percentage Error (MAPE) expresses prediction accuracy in percentage terms, facilitating comparison across different scales [24]:

$$MAPE = \left(\frac{100\%}{n}\right)\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (4)$$

MAPE values are interpreted according to standard criteria: values below 10% indicate excellent forecasting accuracy, 10–20% indicate good accuracy, 20–50% indicate acceptable accuracy, and values above 50% indicate poor accuracy [25]. Table 3 summarizes these criteria.

Table 3. MAPE interpretation criteria

MAPE Value (%)	Criteria
< 10	Excellent
10 – 20	Good
20 – 50	Acceptable
> 50	Poor

3. RESULTS AND DISCUSSION

3.1. Spatiotemporal Characteristics of NO2 Concentration

The spatial distribution of NO_2 concentrations across the ASEAN region exhibits pronounced heterogeneity and strong spatial autocorrelation. As illustrated in Figure 1 (representing the mean NO_2 column density for the year 2024), tropospheric air pollution does not disperse randomly across the landscape. Instead, the highest concentration levels indicated by the dense red clusters are distinctly aggregated within major metropolitan and highly industrialized hubs. Based on the spatial mapping, these primary emission hotspots are strictly localized around Hanoi and its surrounding industrial zones in Vietnam, the Bangkok metropolitan region in Thailand, the Kuala Lumpur conurbation in Malaysia, Singapore, the highly urbanized corridor of Java Island in Indonesia, and Metropolitan Manila in the Philippines. This significant spatial clustering confirms that localized geographic contexts, particularly the density of anthropogenic activities and fossil fuel combustion, play a deterministic role in emission levels.

Furthermore, the spatial map reveals a striking macro-regional disparity: mainland Southeast Asia (encompassing Myanmar, Thailand, Laos, Cambodia, and Vietnam) exhibits a noticeably wider spread of high NO2 concentrations (red/orange hues) compared to the maritime regions, such as Kalimantan, Sulawesi, Papua in Indonesia, and the broader Philippine archipelago, which are dominated by low-concentration baseline minimums (blue hues). This stark contrast is driven by several intersecting factors. First, mainland Southeast Asia experiences massive, recurring agricultural biomass burning cycles during the dry season, releasing substantial NO2 emissions over vast contiguous landmasses. Second, the interconnected land topography of the mainland facilitates transboundary pollution accumulation. Conversely, the maritime regions outside of Java are predominantly covered by vast tropical rainforests with significantly lower industrial and population densities. Additionally, these maritime islands experience higher year-round precipitation, which acts as a natural cleansing mechanism through the *wet scavenging* effect, effectively washing NO_2 pollutants from the atmospheric column.

This structural geographic disparity from highly industrialized, fire-prone mainland corridors to remote, high-rainfall maritime forests serves as the primary justification for incorporating geographic coordinates (Longitude and Latitude) as explicit predictor variables. By embedding these spatial anchors, the Random Forest model is equipped to learn and capture these complex, unmeasured location-specific variations that cannot be fully explained by dynamic environmental predictors alone.

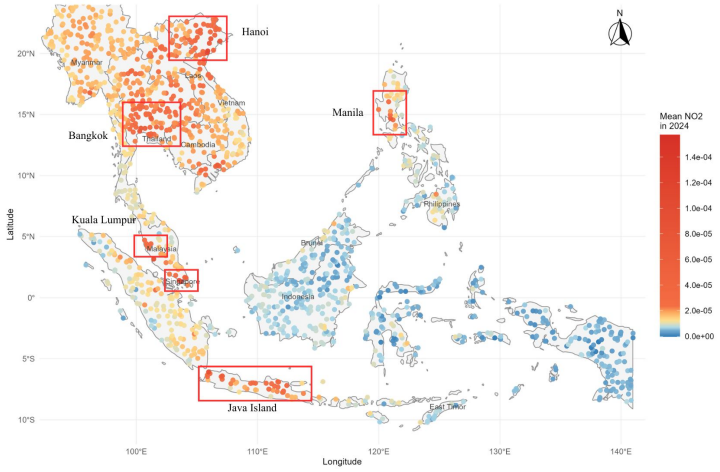
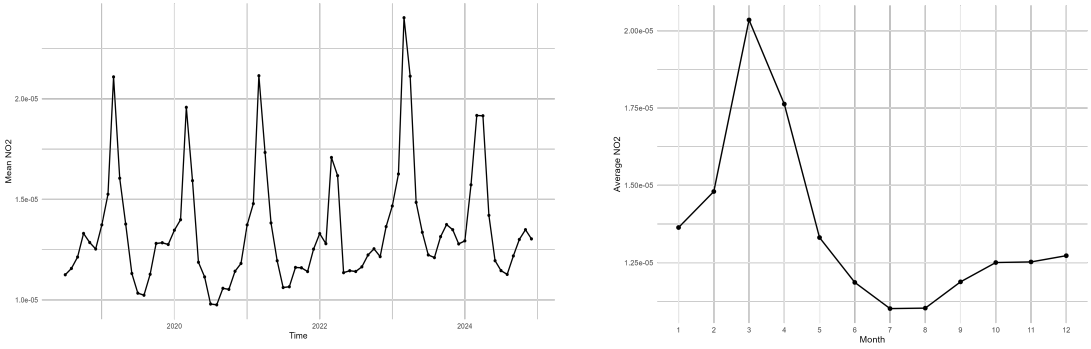


Figure 1. Spatial distribution of mean tropospheric NO_2 column density across the ASEAN region in 2024

Temporally, NO_2 concentrations demonstrate a highly consistent and cyclical seasonal pattern throughout the observation period. Figure 2b depicts the aggregated monthly seasonal behavior, revealing a distinct unimodal annual cycle. Concentrations consistently peak around March, driven by the convergence of the regional dry season, annual agricultural biomass burning cycles, and intensified industrial activities. Following this peak, pollution levels decline sharply, reaching their lowest point during the wet monsoon months (July and August) due to the wet scavenging effect of heavy precipitation. This predictable annual recurrence serves as the primary justification for the extraction and inclusion of the 12-month lag variable (lag-12) derived from the ACF analysis. By embedding this temporal memory, the model is equipped with a highly accurate historical baseline that perfectly encapsulates the recurring seasonal and climatic conditions of the exact same month in the preceding year.



(a) (b)

Figure 2. (a) Monthly mean NO_2 concentration over the 78-month observation period (July 2018 – December 2024); (b) Aggregated annual seasonal cycle highlighting peak accumulations in March

3.2. Model Performance of Random Forest

Table 4 presents the complete performance metrics for all four RF model variants evaluated on the held-out test set. Figure 6 provides a visual comparison of RMSE, MAPE, and R^2 values across models to facilitate interpretation of marginal feature contributions.

Table 4. Predictive performance of four Random Forest model variants on the test set

Model	RMSE ($\times 10^{-6}$ mol/m²)	MAPE (%)	R^2
RF_base	9.546	1.316	0.480
RF_lag12	5.670	0.754	0.817
RF_coords	6.606	0.656	0.751
RF_coords_lag12	**5.423**	**0.608**	**0.832***

**Best-performing model based on R^2, MAPE and RMSE*

The results reveal a clear and consistent pattern of performance improvement as spatial and temporal information is progressively incorporated into the model. The baseline RF model (RF_base), which utilizes only the three environmental predictors (nighttime lights, NDVI, and population density) without any location or lag information, achieves an R^2 of 0.480 and a MAPE of 1.316%. While the MAPE value technically falls within the 'Excellent' category (<10%), the R^2 of 0.480 indicates that the baseline model explains less than half of the total variance in NO2 concentrations, highlighting the substantial unexplained variation attributable to spatial and temporal factors not captured by the three environmental predictors alone.

The addition of geographic coordinates alone (RF_coords) produces a marked improvement, raising R^2 to 0.751 (+56.5% relative to baseline) and reducing RMSE by 30.8%. This improvement demonstrates that the three environmental predictors, despite their direct physical relevance to NO_2 emissions, are insufficient to capture the full spatial structure of NO_2 concentrations across the heterogeneous ASEAN landscape. Longitude and latitude serve as implicit proxies for a range of spatially structured factors not included as explicit predictors, including prevailing wind patterns, proximity to coastlines, elevation gradients, and country-specific emission control regulations, all of which vary systematically across the region.

The most substantial single-feature improvement is observed when the 12-month lag is added. RF_lag12 achieves an R^2 of 0.817 (+70.3% relative to baseline), surpassing even the coordinates-only model despite adding only a single lagged predictor. This finding is particularly noteworthy as it quantitatively establishes that temporal persistence, specifically the annual cycle of NO_2 variation, is a more powerful predictor of current NO_2 concentrations than the combined spatial information from longitude and latitude. The physical interpretation is straightforward: the NO2 concentration 12 months prior encapsulates the full seasonal and emission cycle context of a given location, serving as a compressed representation of the recurring annual patterns of monsoon dynamics, agricultural burning seasons, and industrial output cycles that dominate NO_2 variation in the ASEAN context.

The combination of geographic coordinates with the 12-month lag (RF_coords_lag12) yields the best overall performance, achieving RMSE of 5.423×10^{-6} mol/m², MAPE of 0.608%, and R^2 of 0.832 (+73.4% relative to baseline). This result confirms that spatial and temporal information are complementary rather than redundant: coordinates capture location-specific structural differences between regions that persist across time, while the 12-month lag captures the recurring temporal dynamics at each location. Notably, the marginal gain from adding coordinates to the lag-12 model (R^2: 0.817 to 0.832, +1.8%) is considerably smaller than the standalone contribution of coordinates (+56.5%) or lag-12 (+70.3%). This suggests that once the annual temporal structure is accounted for, much of the remaining spatial variation is of second order in magnitude.

3.3. Feature Importance

The underlying drivers of NO_2 concentrations across the ASEAN region were evaluated using the feature importance metric derived from the Increase in Node Purity (IncNodePurity). As illustrated in Figure 3, the Random Forest model identified a distinct hierarchy of predictors, where temporal memory and spatial anchors played a more decisive role than traditional environmental indicators.

The absolute dominance of the 12-month lag (NO2_lag12) underscores the critical importance of temporal persistence in atmospheric pollution modeling. Logically, the NO_2 concentration from exactly one year prior serves as a comprehensive summary of a location's environmental capacity. This lag captures the identical seasonal context including recurring annual biomass burning cycles, monsoon-driven dispersion patterns, and industrial rhythms that define a specific month. Essentially, the past environment acts as the most robust baseline for predicting current pollution levels, as it encapsulates the structural periodicity inherent in the tropical ASEAN atmosphere.

Nighttime lights (NightLights) emerged as the second most influential predictor, functioning as a high-fidelity proxy for human anthropogenic activity. Areas characterized by intense radiance at night typically correspond to dense urban centers, 24-hour industrial zones, and

major transportation hubs. These locations are the primary epicenters of fossil fuel combustion. From a causal perspective, the strong importance of nighttime lights confirms that where economic intensity is concentrated, elevated NO_2 emissions inevitably follow. This relationship allows the model to distinguish between high-emission urban clusters and low-emission rural peripheries.

Interestingly, the geographic coordinates (Longitude and Latitude) demonstrated higher predictive power than traditional indicators such as population density and NDVI. These features act as spatial anchors that encode "hidden contextual information" unique to a specific geographic point. For instance, a coordinate may implicitly represent a coastal city with distinct sea-breeze dispersion patterns or a valley topography where pollutants are naturally trapped. By incorporating Lon and Lat, the Random Forest model "learns" the unique geographic footprint of each ASEAN sub-region. This spatial awareness enables the model to resolve regional disparities and localized emission behaviors that might otherwise remain obscured even when other environmental indicators appear similar.

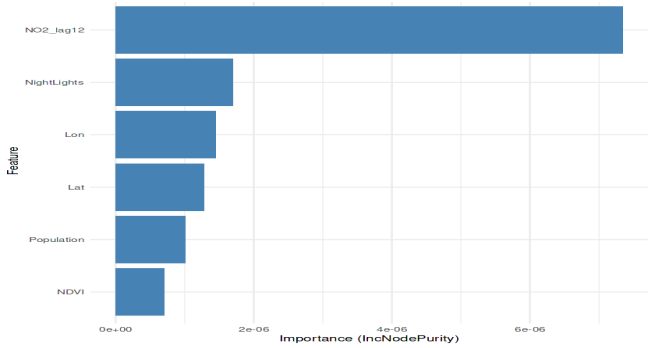


Figure 3. Feature importance ranking of the optimized Random Forest model (RF_coords_lag12) based on the Increase in Node Purity metric

3.5. Discussion

The findings of this study are consistent with and extend the existing literature on spatio-temporal machine learning for air quality estimation. The dominant role of the 12-month lag feature corroborates the findings of Zhalehdoost and Taleai [15], who reported RMSE reductions of 25–39% upon incorporating lag features into machine learning models for urban air pollution prediction. The present study extends this finding to a regional satellite-based context at the scale of an entire continent, demonstrating that the benefit of temporal lag features is not limited to dense urban monitoring networks but generalizes to sparse satellite-derived observations across heterogeneous tropical environments.

The finding that geographic coordinates provide substantial predictive value independently of the environmental predictors is consistent with the theoretical motivation for spatially explicit modeling approaches such as Geographically Weighted Random Forest (GWRF), as discussed by Georganos *et al*. [26], and demonstrated empirically in the ASEAN context by Suprayogi *et al*. [27]. When coordinates are included as explicit features in the global RF model, they allow the ensemble to implicitly learn spatial gradients and regional clusters that reflect unmeasured location-specific drivers of NO_2 concentration. This result underscores the importance of spatial awareness in environmental machine learning, even in the absence of a fully localized modeling framework.

Several limitations of this study warrant explicit acknowledgment. First, the use of default RF hyperparameters, while methodologically justified for the purpose of isolating feature contributions, means that the absolute performance figures reported here do not represent the upper bound of predictive accuracy achievable with this dataset. Systematic hyperparameter optimization through spatial cross-validation could yield further improvements, particularly for the best-performing models. Second, the 10 km grid resolution, while appropriate for the scale of the primary NO_2 data, may inadequately resolve sub-urban spatial heterogeneity in densely populated cities such as Jakarta, Bangkok, and Manila, where NO_2 gradients can be extreme over distances of one to two kilometers. Third, the imputation procedure, while validated against distributional

diagnostics, introduces a degree of data uncertainty particularly in regions with high cloud cover prevalence such as equatorial Sumatra and Borneo. Future work should incorporate imputation uncertainty into model evaluation through multiple imputation or ensemble-based sensitivity analysis.

4. CONCLUSION

This study demonstrates that the systematic incorporation of geographic coordinates and temporal lag features substantially enhances the predictive accuracy of Random Forest models for satellite-derived NO_2 concentration estimation across ASEAN. Among the four model variants evaluated, the RF_coords_lag12 model, combining environmental predictors with geographic coordinates and a 12-month temporal lag, achieves the best performance, with RMSE of 5.423 × 10^{-6} mol/m², MAPE of 0.608%, and R^2 of 0.832. This represents a 73.4% improvement in explained variance over the baseline RF model, confirming the critical importance of both spatial and temporal context in atmospheric pollutant modeling. The 12-month lag emerges as the single most impactful feature, capturing the dominant annual seasonality of NO_2 driven by monsoon dynamics and recurring emission cycles across the ASEAN region. The scalable, satellite-only methodology presented here offers a practical and transferable framework for NO_2 monitoring in tropical developing regions where ground-based monitoring infrastructure remains limited. Future research directions include the application of Geographically Weighted Random Forest to capture spatial non-stationarity in feature relationships, the integration of meteorological covariates to further disentangle emission and dispersion dynamics, and the development of near-real-time prediction pipelines leveraging the operational continuity of the Sentinel-5P mission.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the European Space Agency (ESA) for providing open access to Sentinel-5P TROPOMI data, NASA and USGS for MODIS and VIIRS products, and the WorldPop project for gridded population estimates. All satellite data acquisition and preprocessing were conducted through the Google Earth Engine cloud computing platform.

REFERENCES

[1] M. Li *et al.*, "MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP," *Atmos. Chem. Phys.*, vol. 17, no. 2, pp. 935–963, Jan. 2017, doi: 10.5194/acp-17-935-2017.

[2] K. D. Kanniah, N. A. F. Kamarul Zaman, D. G. Kaskaoutis, and M. T. Latif, "COVID-19's impact on the atmospheric environment in the Southeast Asia region," *Science of The Total Environment*, vol. 736, p. 139658, Sep. 2020, doi: 10.1016/j.scitotenv.2020.139658.

[3] K. Vohra *et al.*, "Rapid rise in premature mortality due to anthropogenic air pollution in fast-growing tropical cities from 2005 to 2018," *Sci. Adv.*, vol. 8, no. 14, Apr. 2022, doi: 10.1126/sciadv.abm4435.

[4] M. Rendana, M. H. Dahlan, F. Alia, D. Bahrin, R. Pebrianto, and P. K. Wardhani, "Impact of vegetation cover and land surface temperature on the seasonal tropospheric NO2 level variation from satellite observation," *Advances in Space Research*, vol. 74, no. 10, pp. 4760–4772, Nov. 2024, doi: 10.1016/j.asr.2024.07.073.

[5] P. Fan *et al.*, "Urbanization, economic development, and environmental changes in transitional economies in the global south: a case of Yangon," *Ecol. Process.*, vol. 11, no. 1, p. 65, Nov. 2022, doi: 10.1186/s13717-022-00409-6.

[6] T. Holloway *et al.*, "Satellite Monitoring for Air Quality and Health," *Annu. Rev. Biomed. Data Sci.*, vol. 4, no. 1, pp. 417–447, Jul. 2021, doi: 10.1146/annurev-biodatasci-110920-093120.

[7] F. Tack *et al.*, "Assessment of the TROPOMI tropospheric NO $_2$ product based on airborne APEX observations," Jun. 29, 2020. doi: 10.5194/amt-2020-148.

[8] Z. A. Hanami *et al.*, "Spatial–Temporal Changes in Air Pollutants in Four Provinces of Sumatra Island, Indonesia: Insights from Sentinel-5P Satellite Imagery," *Urban Science*, vol. 9, no. 2, p. 42, Feb. 2025, doi: 10.3390/urbansci9020042.

[9] D. Stratoulias, N. Nuthammachot, R. Dejchanchaiwong, P. Tekasakul, and G. R. Carmichael, "Recent Developments in Satellite Remote Sensing for Air Pollution Surveillance in Support of Sustainable Development Goals," *Remote Sens. (Basel).*, vol. 16, no. 16, p. 2932, Aug. 2024, doi: 10.3390/rs16162932.

[10] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[11] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 6, p. e5518, Aug. 2018, doi: 10.7717/peerj.5518.

[12] A. N. Pradana, A. Djuraidah, and A. M. Soleh, "Land Use Change Modelling Using Logistic Regression, Random Forest and Additive Logistic Regression in Kubu Raya Regency, West Kalimantan," *Forum Geografi*, vol. 37, no. 2, Dec. 2023, doi: 10.23917/forgeo.v37i2.23270.

[13] V. van Zoest, F. B. Osei, G. Hoek, and A. Stein, "Spatio-temporal regression kriging for modelling urban NO_2 concentrations," *International Journal of Geographical Information Science*, vol. 34, no. 5, pp. 851–865, May 2020, doi: 10.1080/13658816.2019.1667501.

[14] A. Srivastava, A. K. Thakur, A. Rai, R. N. Yadava, R. D. Garg, and P. K. Garg, "Temporal and Spatial Variations of Nitrogen Dioxide Concentrations in Kolkata, India," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. X-G-2025, pp. 855–862, Jul. 2025, doi: 10.5194/isprs-annals-X-G-2025-855-2025.

[15] A. Zhalehdoost and M. Taleai, "Unravelling the importance of spatial and temporal resolutions in modeling urban air pollution using a machine learning approach," *Sci. Rep.*, vol. 15, no. 1, p. 27708, Jul. 2025, doi: 10.1038/s41598-025-13639-3.

[16] J. Fan, P. Zhang, J. Chen, B. Li, L. Han, and Y. Zhou, "Quantitative Estimation of Missing Value Interpolation Methods for Suomi-NPP VIIRS/DNB Nighttime Light Monthly Composite Images," *IEEE Access*, vol. 8, pp. 199266–199288, 2020, doi: 10.1109/ACCESS.2020.3035408.

[17] S. Cheng, P. Peng, and F. Lu, "A lightweight ensemble spatiotemporal interpolation model for geospatial data," *International Journal of Geographical Information Science*, vol. 34, no. 9, pp. 1849–1872, Sep. 2020, doi: 10.1080/13658816.2020.1725016.

[18] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: The analysis of spatially varying relationships*. Wiley, 2002.

[19] N. S. Maharani *et al.*, "APLIKASI MODEL ARIMA GARCH DALAM PERAMALAN DATA NILAI TUKAR RUPIAH TERHADAP DOLAR TAHUN 2017-2022," *Jurnal Matematika Sains dan Teknologi*, vol. 24, no. 1, pp. 37–50, May 2023, doi: 10.33830/jmst.v24i1.4875.2023.

[20] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[21] A. Vamsikrishna and E. V. Gijo, "New Techniques to Perform Cross-Validation for Time Series Models," *Operations Research Forum*, vol. 5, no. 2, p. 51, Jun. 2024, doi: 10.1007/s43069-024-00334-8.

[22] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.

[23] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis Fifth Edition*, Fifth. New Jersey: John Wiley & Sons, Inc., 2012.

[24] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016, doi: 10.1016/j.neucom.2015.12.114.

[25] C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," *Journal of the Operational Research Society*, vol. 66, no. 8, pp. 1352–1362, Aug. 2015, doi: 10.1057/jors.2014.103.

[26] S. Georganos *et al.*, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto Int.*, vol. 36, no. 2, pp. 121–136, Jan. 2021, doi: 10.1080/10106049.2019.1595177.

[27] M. A. Suprayogi, B. Sartono, and K. A. Notodiputro, "GEOGRAPHICALLY WEIGHTED MACHINE LEARNING MODEL FOR ADDRESSING SPATIAL HETEROGENEITY OF PUBLIC HEALTH DEVELOPMENT INDEX IN JAVA ISLAND," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 4, pp. 2577–2588, Oct. 2024, doi: 10.30598/barekengvol18iss4pp2577-2588.