Received: July 16, 2025. Revised: September 26, 2025.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Spatial Estimation of PM2.5 Using Random Forest with Temporally Adjusted Portable Sensors, Land Use, and AOD in a Data-Scarce Urban Area

Retno Tri Wahyuni[1,2]* Dirman Hanafi[1] M. Razali Tomari[1] Dadang Syarif Sihabudin Sahid[3]

[1]*Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia*
[2]*Industrial Technology Department, Politeknik Caltex Riau, Pekanbaru, Indonesia*
[3]*Information Technology Department, Politeknik Caltex Riau, Pekanbaru, Indonesia*
* Corresponding author's Email: retnotri@pcr.ac.id

Abstract: This study proposes a spatial modeling approach to estimate annual PM2.5 concentrations in Pekanbaru, Indonesia, a city with limited monitoring data. The dataset spans January–December 2024, a period characterized by relatively stable background conditions, free from severe forest and peatland fires in Riau. A Random Forest model was trained using temporally adjusted low-cost sensor measurements, land-use data, and MODIS-derived AOD. Hyperparameters were tuned with LOLO-based RandomizedSearchCV; robustness was checked with spatially buffered LOLO. Spatial performance was stable to 3 km (MAE ≈2.0–2.1 μg/m³; R^2 ≈0.55–0.57) and declined at 4–5 km, indicating controlled overfitting but weaker generalization at larger separations. SHAP identified AOD as dominant; peatlands tended to raise, and vegetation to lower, PM2.5. Citywide predictions ranged 24.52–34.86 μg/m³, placing all areas in the "Moderate" category under Indonesia's ISPU. A Ridge baseline scored slightly higher, but overlapping CIs make differences non-definitive; RF is retained for nonlinear capture and interpretability via SHAP. This framework supports air quality mapping; future work will expand sites, add meteorology, address missing AOD, and test on independent data.

Keywords: PM2.5, Random forest, Temporally adjusted, Land use, AOD.

1. Introduction

Air pollution caused by fine particulate matter (PM2.5) has emerged as one of the most critical environmental issues affecting global public health. Exposure to PM2.5 over long periods contributes to increased risks of respiratory and cardiovascular disease, and has been associated with premature death [1-3]. To address these health concerns, accurate spatial estimation of PM2.5 is needed to guide mitigation strategies and ensure policies are grounded in data.

However, in many regions, especially in developing countries such as Indonesia, the limited number and uneven distribution of fixed monitoring stations remain a significant obstacle to obtaining representative spatial data. Pekanbaru City, as an example of an urban area prone to land fires and transportation emissions, has only a few permanent monitoring stations.

As an alternative solution, low-cost sensors are increasingly used due to their flexibility and ability to cover areas not monitored by official systems. The use of these low-cost sensors can be implemented continuously or in the form of sampling. When deployed in short sampling campaigns, these sensors provide short-term snapshots that may not represent long-term conditions. Therefore, we apply a temporal adjustment using long-term data from fixed monitoring stations to align short-term measurements with annual-scale conditions for spatial modeling.

With the rapid development of environmental monitoring technologies, machine learning has become an important tool for improving air quality modeling. Random Forest (RF) is especially popular because it captures complex nonlinear associations, is relatively unaffected by multicollinearity, and can integrate heterogeneous predictor data. A literature review conducted by the authors on various spatial

PM2.5 modeling studies based on machine learning indicates that RF is one of the most commonly used and widely employed approaches [4]. Previous studies have successfully combined RF with diverse data sources, including satellite-derived aerosol information, meteorological variables, and land use data, to generate fine-scale PM2.5 maps [5-7]. However, most of these studies have been conducted in data-rich regions and rely on dense ground-based monitoring networks as well as high-resolution meteorological datasets.

In response to data-scarcity challenges, this study proposes a framework for spatial PM2.5 estimation in data-poor contexts, with a case study focused on Pekanbaru, Riau, Indonesia. The proposed framework comprises five components: (i) temporally adjusts short-term measurements from portable sensors to produce spatially consistent annual targets; (ii) uses a predictor set consisting of land use and Aerosol Optical Depth (AOD), so it remains applicable when meteorological data are limited; (iii) enforces strict spatial validation via Leave-One-Location-Out and spatially buffered LOLO to limit near-neighbor bias and prevent spatial target leakage; (iv) quantifies uncertainty with nonparametric bootstrap 95% confidence intervals on spatial predictions; and (v) provides local interpretability with SHAP, revealing the direction and magnitude of each predictor's effect at each location. For context, we also compare Random Forest with linear baselines (Ridge) under the same spatial cross-validation protocol. Taken together, these elements yield a simple, reproducible, and transferable framework for other data-scarce cities.

The remainder of this paper is organized as follows: Section 2 describes the study area. Section 3 presents the data collection, while Section 4 explains the temporal adjustment. Section 5 mentions the modeling setup. Section 6 discusses the implications of the findings, limitations of the study, and comparisons with similar works. Finally, Section 7 presents the conclusions and recommendations for future research.

2. Study area

As the largest economic center in the eastern region of Sumatra, Pekanbaru City has undergone significant urbanization and industrial expansion [8-9], which has contributed to increasingly complex air quality issues. The city is divided into 15 sub-districts, each with distinct characteristics in terms of economic activities, residential areas, and infrastructure. In this study, the area of Pekanbaru was divided into small spatial units using a 5 km × 5 km grid system as part of the methodological design. All data used in the modeling process, including prediction outputs, represent the conditions within each grid cell. Figure 1 illustrates the division of Pekanbaru into 42 grids, of which 36 fall within the city's official administrative boundaries and are used for spatial analysis. Grid IDs were then used as spatial identifiers in the modeling and analysis process.

Previous studies have employed various grid resolutions, ranging from 500 meters [10] to 10 km [11], where finer grids enable more detailed spatial analysis but require greater computational cost and data availability. In this study, a 5 km × 5 km grid resolution was selected as a compromise between spatial detail and resource efficiency.

Figure 1 also displays 21 PM2.5 sampling points, including 18 portable sensors (P1–P18) and 3 fixed monitors (FS1–FS3) managed by the government or industry. The placement of portable sensors was based on the land use characteristics of each grid. Compared to the authors' previous publication [12], the observation point locations differ due to the use of a more recent base map. This study adopts the updated 15 district division, whereas the earlier study used the previous 13 district boundaries.

3. Data collection

3.1 Land use data collection

The composition of land use was obtained from the 2024 Land Use Map of Pekanbaru City. The area of Pekanbaru City is classified into 11 land use categories. The area of each land use type within each grid is presented in Figure 2.

Residential-dominated grids are generally concentrated in the central part of Pekanbaru City, while plantation land use is more prevalent in peripheral or outlying grids. In addition to these main categories, some grids also reflect the presence of secondary forests, shrublands, and water bodies, although their proportions are relatively small. Land use types such as Air and Sea Transport Infrastructure, Shrubs/Wetlands, and Ponds appear only in a few specific grids, indicating a limited spatial distribution.

3.2 PM2.5 data collection

PM2.5 was measured with low-cost sensors calibrated to a commercial IQAir unit at 18 sites across Pekanbaru representing diverse land uses. Sampling spanned three consecutive days between August–September 2024, a fire-free period in Riau, ensuring stable conditions. This study also used hourly 2024 PM2.5 from three fixed Pekanbaru

stations (public open data) to temporally adjust the portable-sensor measurements.

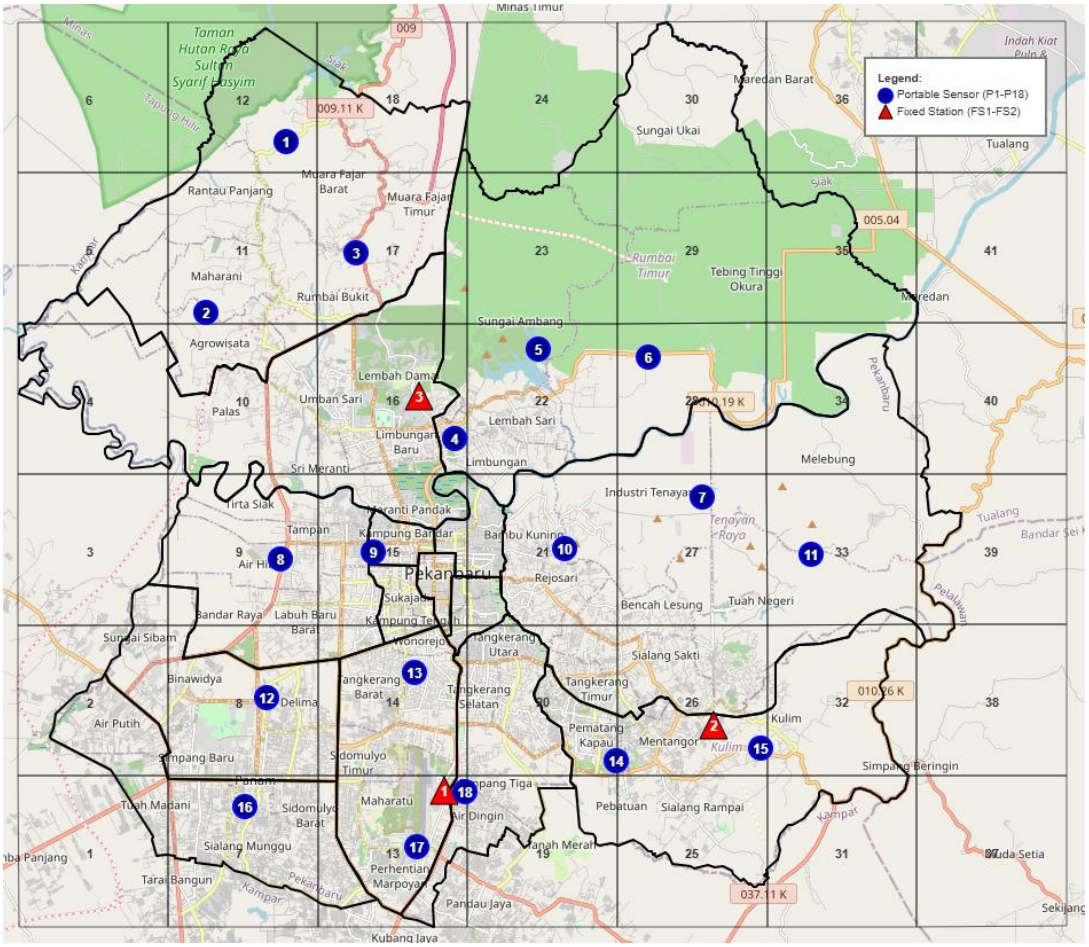

Figure 1 Grid Division of Pekanbaru and PM2.5 Sampling Locations

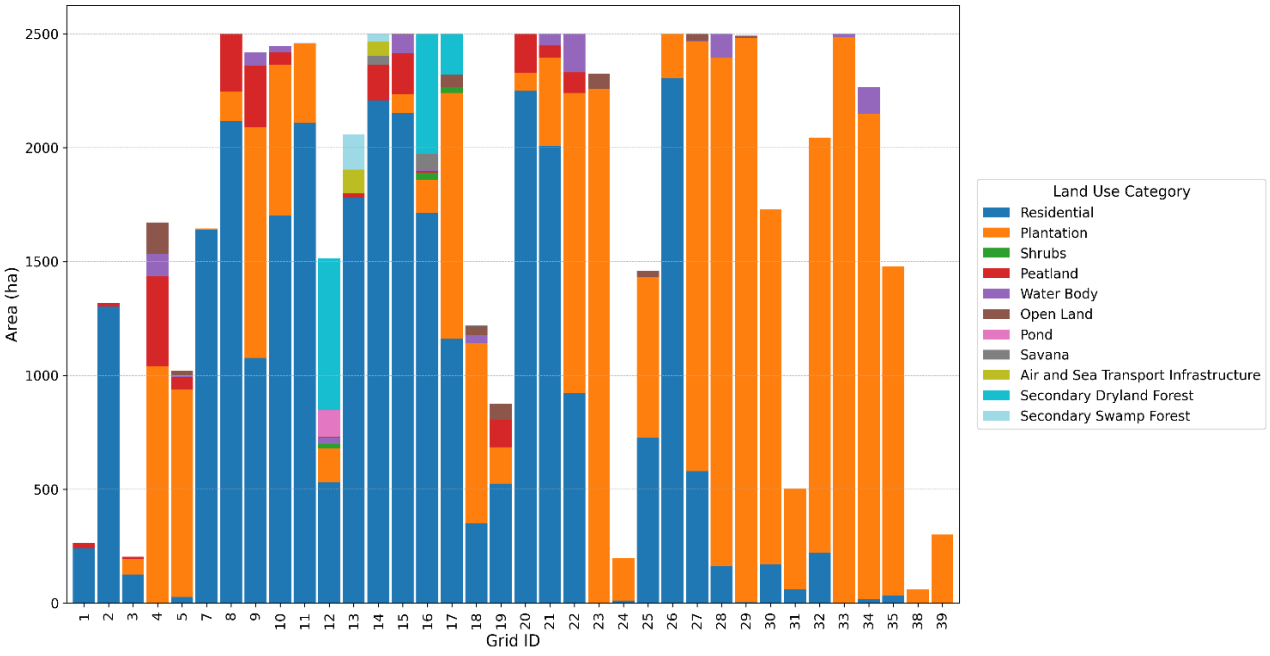

Figure 2 Land Use Composition per Grid

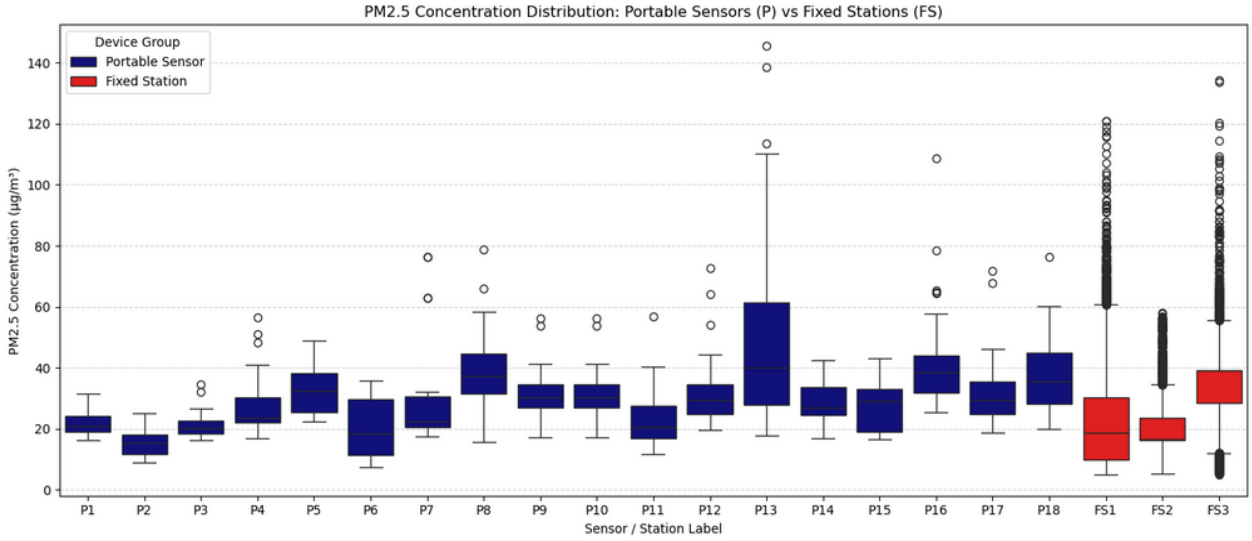

Figure 3. Boxplot of PM2.5 Sensor Data

Figure 3 presents the sensor measurement data for all locations, where a boxplot visualization is used to depict the variation, median, and the presence of extreme values at each observation point. The boxplot illustrates clear differences between portable sensors (P1–P18) and fixed stations (FS1–FS3). Most portable sensors recorded relatively consistent concentrations with medians around 20–40 µg/m³, several sensors exhibited wider spreads. By comparison, the fixed stations, especially FS2 and FS3, exhibited greater variability and a higher number of outliers, indicating both sensitivity to extreme fluctuations and the impact of site-specific conditions. These contrasts highlight the importance of applying temporal adjustment so that portable sensor data align with fixed station records, allowing short-term measurements to be reliably used in spatial modeling.

3.3 AOD data collection

AOD is an optical indicator commonly used to estimate the concentration of fine particulate matter in the atmosphere, including PM2.5. Although AOD and PM2.5 generally exhibit a positive correlation, the relationship is complex due to the influence of atmospheric factors (such as humidity and aerosol layer height) and surface characteristics. AOD measurement instruments are available at various spatial resolutions depending on the satellite type, such as TROPOMI on Sentinel-5P (approximately 3.5 km), MODIS on Terra and Aqua (1–10 km), VIIRS on Suomi NPP (approximately 750 m), and geostationary satellites like Himawari-8 and GOES-16 (4–5 km), which also offer high temporal resolutions of 10 to 15 minutes [13].

In this study, AOD was obtained from the MODIS MCD19A2_GRANULES version 061 collection via the Google Earth Engine (GEE) platform, which provides daily products from MODIS sensors onboard the Terra and Aqua satellites. The primary band used was *Optical_Depth_047* (0.47 µm), a wavelength commonly applied in PM2.5 estimation studies. The data were analyzed for the period from January 1 to December 31, 2024, then temporally aggregated into representative annual AOD values. Figure 4 illustrates the spatial distribution of annual mean AOD in 2024 as a colored raster clipped to the administrative boundary of Pekanbaru City, with a gradient ranging from yellow (low AOD) to dark purple (high AOD).

A 5 × 5 km grid was used to extract zonal features such as average AOD and land use data. Three grid cells (14, 35, and 47) were excluded due to limited spatial coverage and unrepresentative AOD values. Figure 5 presents the visualization of average AOD values per grid, using standardized grid ID.

Based on Figure 5, the highest AOD concentrations are observed in the southern and southwestern parts of Pekanbaru City, likely correlating with intensive anthropogenic activities such as biomass burning, industrial operations, and heavy traffic. In contrast, the northern and eastern areas show lower AOD values, indicating relatively better air quality, which aligns with the dominance of vegetated land cover in those regions. This condition is also consistent with the land use composition described in Figure 2, where the southern and southwestern areas are predominantly residential, while the northern and eastern areas are primarily plantation zones.

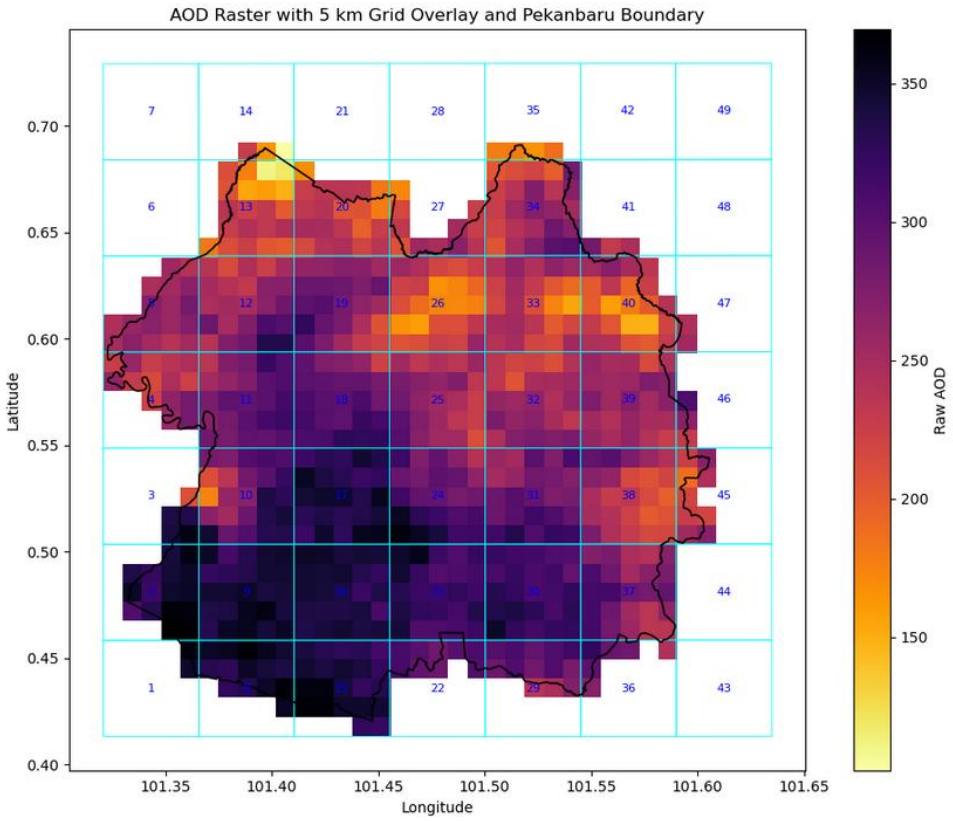

Figure 4. Annual Mean AOD Image from MODIS Satellite for Pekanbaru City in 2024

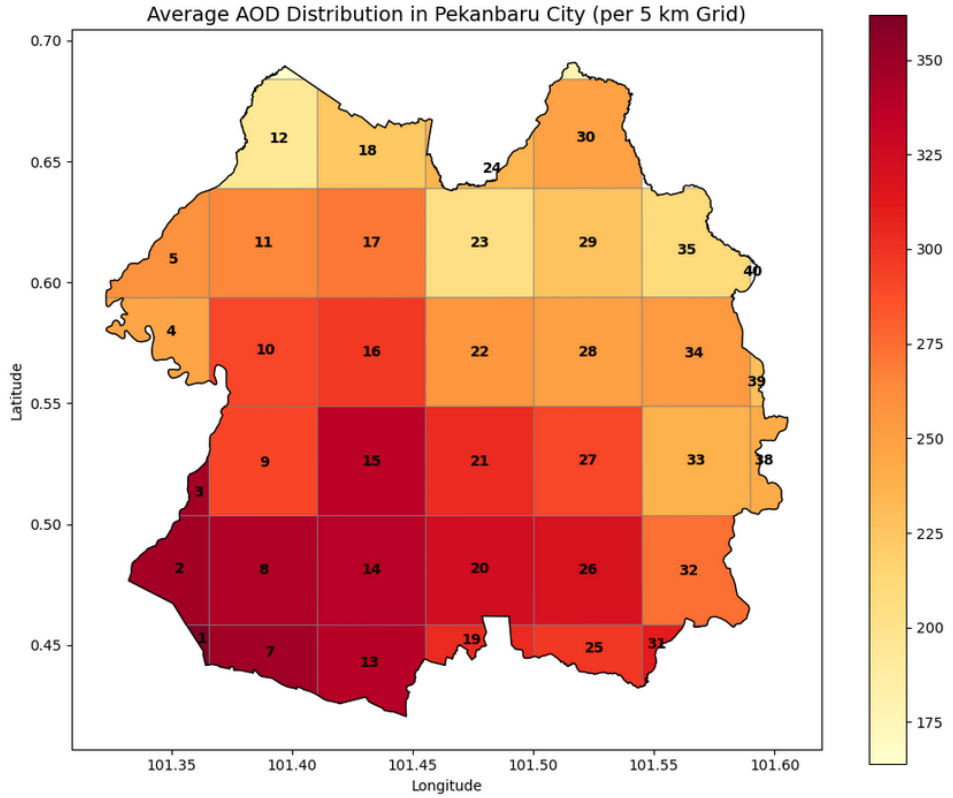

Figure 5. Distribution of Average AOD in Pekanbaru City Based on 5 km Grid Resolution

4. Temporal adjustment

The short-term portable sensor data were temporally adjusted against data from fixed monitoring stations to ensure comparable distributions. Several previous studies have shown that temporal adjustment is a crucial step in improving the estimation of long-term concentrations from irregular or short-duration data collections, such as those from portable sensors. A comprehensive study [14] compared five temporal adjustment approaches, including mean-scaled, median-scaled, log-transformed scaling, and naïve adjustment, and demonstrated that the effectiveness of each method varies considerably depending on sample size and measurement campaign design. These findings highlight that no single adjustment method is universally applicable, and its effectiveness must be evaluated within the specific context.

In line with this, the present study employed a three parameter linear optimization approach for temporal adjustment, involving a scaling factor (c) and two offset parameters (a and b), which were determined by minimizing the absolute error between the reference data (from fixed stations) and the target data (from portable sensors) during the overlapping period. The parameter optimization was conducted using the Nelder–Mead simplex algorithm, selected for its robustness in handling non-differentiable and nonlinear objective functions. For each sensor–station pair, the optimization started from an initial guess of [1.0, 0.0, 0.0], representing the scaling factor (c) and two offset parameters (a and b). The objective function minimized the mean absolute error (MAE) between the temporally adjusted reference data and the target portable sensor measurements during the overlapping observation period. All optimizations converged successfully with no failures or abnormal termination status. This approach shares methodological similarities with the mean-scaled and median-scaled adjustment techniques, which also rely on the ratio between short-term observations and long-term values from the reference station (mean or median). However, the method used in this study is more flexible as it does not depend on fixed statistical ratios, but instead explicitly searches for the optimal correction parameters. Therefore, it can be considered a generalized form of the mean/median-scaled adjustment, capable of capturing both scale differences between measurement devices.

The correction was applied in the following Eq. (1).

$$PM_{2.5}^{Final} = \left(cPM_{2.5}^{Ref} + a\right) + b \qquad (1)$$

Where,

c: scaling factor;

a: initial offset;

b: final offset (corrective bias);

$PM_{2.5}^{Ref}$: reference data from fixed stations.

The pairing between portable sensors and fixed stations for temporal adjustment was based on two main criteria: similarity in land use characteristics and, more importantly, temporal correlation during the overlapping observation period. While spatial proximity is commonly used in sensor pairing, it was not adopted as a primary criterion due to the limited number of available reference stations. In several cases, the nearest reference station did not share similar land use or exhibit coherent temporal patterns, which could lead to mismatched adjustment baselines. In such cases, stations with stronger temporal correlations to the target sensor were prioritized, even if their land use characteristics differed. This high temporal correlation was also considered a reasonable surrogate for comparable meteorological influences during the overlap period, acknowledging that direct meteorological data were unavailable in this study. For clarity, we denote the correlation between target and reference during the overlapping period as r_{ovr}, while the correlation after temporal adjustment is referred to as r_{adj}. The resulting sensor pairs along with their corresponding r_{ovr} and r_{adj} values are summarized in Table 1.

Based on Table 1, the initial correlation between portable sensors and reference stations during the overlapping period (r_{ovr}) ranged from 0.40 to 0.81, indicating temporal relationships from moderate to strong. After applying temporal adjustment using a three-parameter linear optimization approach, all pairs exhibited a substantial increase in correlation, with r_{adj} values calculated over the full one-year period following the adjustment reaching 0.98–1.00. The main variation among sensors was primarily determined by the scaling factor (c), which ranged from 0.72 to 1.92, while the offset parameters (a and b) were relatively small, contributing less to the correction. These findings indicate that inter-sensor differences were mainly related to measurement scale discrepancies, and the applied correction method proved effective in aligning the temporal patterns of portable sensors with reference data over the long term.

5. Data preparation and modeling setup

We used a Random Forest Regressor (RF), an ensemble of decision trees trained on random subsets

Table 1. Point Pairings and Evaluation of Temporal Adjustment Results

Target	r_{ovr}	Ref	c	a	b	r_{adj}
P1	0.46	FS3	0.7761	0.0004	0.0003	1.00
P2	0.59	FS3	0.7327	0.0002	0.0005	1.00
P3	0.45	FS3	0.7259	0.0003	0.0004	1.00
P4	0.81	FS3	0.7658	0.0003	0.0003	1.00
P5	0.57	FS3	0.8139	0.0002	0.0002	1.00
P6	0.68	FS2	1.2899	0.0003	-0.0006	1.00
P7	0.40	FS3	0.8193	0.0002	0.0003	1.00
P8	0.56	FS3	0.8982	0.0002	0.0002	1.00
P9	0.54	FS2	1.9249	0.0006	-0.0022	1.00
P10	0.50	FS3	1.6050	0.0003	-0.0013	0.98
P11	0.63	FS3	0.7609	0.0003	0.0003	1.00
P12	0.48	FS2	1.0364	0.0001	0.0001	1.00
P13	0.54	FS1	1.5558	0.0004	-0.0013	0.99
P14	0.77	FS3	0.8651	0.0002	0.0001	1.00
P15	0.58	FS3	0.7782	0.0004	0.0003	1.00
P16	0.42	FS1	1.3467	0.0005	-0.0009	0.99
P17	0.427	FS1	1.1054	0.0000	-0.0000	1.00
P18	0.66	FS2	1.6624	0.0003	-0.0015	1.00

of samples and features. RF is well suited to heterogeneous environmental data because it captures nonlinearities and mitigates overfitting through aggregation. Given the modest sample size (18 locations) relative to the number of predictors (12), we explicitly controlled model complexity via hyperparameter tuning and spatial cross-validation.

The prediction target is the PM2.5 concentration, and spatial predictors were compiled at a 5 × 5 km grid resolution and consist of:

a) Land use composition (11 classes) from Pekanbaru's 2024 official land-use map: Residential, Plantation, Shrubs, Peatland, Water Body, Open Land, Pond, Savanna, Transport Infrastructure, Secondary Dryland Forest, and Secondary Swamp Forest. For each grid, we calculated the areal proportion of every class.
b) Satellite-based AOD, taken as the average raster value per grid.

All features were checked for consistency, and we avoided redundancy where possible. To prevent target leakage in spatial validation, raw latitude/longitude were not used as features; coordinates were only used to define location groups and spatial buffers.

We tuned RF hyperparameters (e.g., number of trees, maximum depth, minimum samples per split/leaf, and max_features) using RandomizedSearchCV with LOLO-CV and MAE as the optimization metric. The best configuration was then refit on the full training set. To assess spatial generalizability, we used LOLO-CV, holding out one location at a time, and a spatially buffered LOLO-CV that excludes training samples within a specified radius (0–5 km) around the test location to reduce near-neighbor dependence. Performance was summarized using Mean Absolute Error (MAE) and R^2, with 95% bootstrap confidence intervals. We report 95% CIs using a non-parametric pairs bootstrap on pooled LOLO held out predictions (B=2,000 resamples; seeds fixed 42–47); fold-level CIs for MAE were obtained by bootstrapping folds (B=2,000). These choices quantify both accuracy and explained variance while reflecting uncertainty due to the small number of locations. As a benchmark, we also trained and evaluated linear baselines (Ridge) under the same LOLO-CV protocol, with inner LOLO tuning of the regularization parameter, so their results are directly comparable and are reported in the Results section. Finally, we used SHAP on the final RF to quantify each predictor's contribution and direction of effect and to screen for spurious drivers.

6. Results and discussion

6.1 Model parameter optimization

Using RandomizedSearchCV with a Leave-One-Location-Out (LOLO) CV scheme and MAE as the objective, the best configuration was bootstrap =

True, max_depth = None, max_features = 0.5, min_samples_leaf = 2, min_samples_split = 3, n_estimators = 351, achieving an inner - CV MAE of 2.067 μg/m³. We then refit the model with these fixed hyperparameters on the full training data and used the same configuration for all subsequent evaluations (standard LOLO and spatially buffered LOLO) to avoid double-dipping and obtain conservative generalization estimates. Regularization is provided by bagging and feature subsampling, reinforced by the min_samples_* constraints; additionally, raw coordinates were excluded from the feature set during validation to prevent spatial target leakage.

6.2 Feature importance interpretation

To interpret the contribution of each predictor to PM2.5 concentration, SHAP values were computed from the final Random Forest model. SHAP provides both the magnitude of feature importance and the direction of each predictor's effect on individual predictions, offering a comprehensive and interpretable approach. This method has also been widely applied in spatial PM2.5 modeling to uncover localized influences of various environmental factors. For example, Li et al. used SHAP in combination with Random Forest to identify the relative importance of topography, land cover, and meteorological parameters in predicting PM2.5 concentrations across Zhejiang Province, China [15]. These examples support the suitability of SHAP for interpreting complex relationships in spatial air quality models such as ours.

Figure 6 presents the SHAP summary plot, which highlights the average impact of each predictor across all predictions. The results indicate that AOD is the most influential variable, consistently contributing positive values to PM2.5 predictions. This is consistent with the well-established relationship between atmospheric aerosol loading and surface-level particulate matter concentrations.

Beyond AOD, land cover features, such as peatlands and plantation areas, also show a substantial influence. Peatlands generally contributed positively to PM2.5 levels, likely due to fire-related emissions, while plantations exhibited negative SHAP values, suggesting that vegetated areas are associated with lower pollution concentrations. Other spatial features including residential areas, transport infrastructure, and water bodies had more modest SHAP values but contributed complementary information to improve spatial predictions.

Given AOD's dominant role in the model, it is important to acknowledge known limitations in satellite-based AOD retrievals. These products are only available under clear-sky and low-humidity conditions and are often missing or unreliable in the presence of clouds, shadows, or high moisture levels [16]. In such cases, PM2.5 may be underrepresented in the model due to unavailable aerosol input. Moreover, spatial heterogeneity in topography and meteorology weaken the AOD PM2.5 relationship in some areas [17], and low temporal sampling during pollution peaks can result in bias [18].

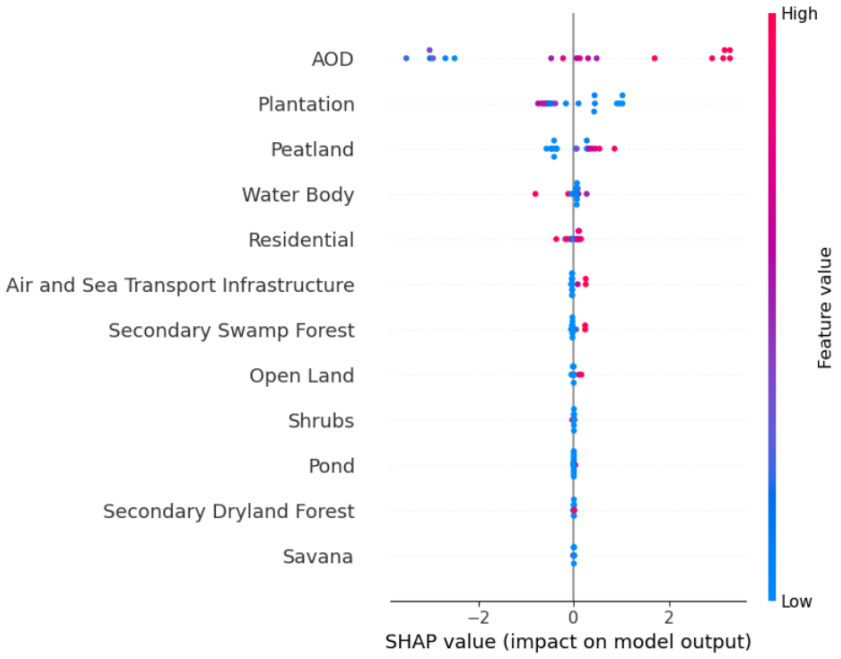

Figure 6. SHAP Summary of Feature Importance

6.3 Model evaluation

Spatial model validation requires specific approaches to minimize bias from spatial autocorrelation, i.e., the similarity of values among geographically adjacent locations. In this context, spatial-based cross-validation has become a widely adopted strategy in air quality modeling. A common method is Leave-One-Location-Out Cross-Validation (LOLO-CV), which evaluates model generalization across spatial locations and has been effectively applied in PM2.5 studies at various scales [19]. Other researchers have explored alternative schemes; for instance, [5] used sample, time, and spatial based ten-fold cross-validation to calibrate high-resolution (1 km) PM2.5 concentrations across China using Random Forest. A refinement is the iterated Spatial Leave-One-Out Cross-Validation (iSLOOCV) [20], which applies multiple minimum distances between training and test data to generate more stable error estimates.

Based on the above approach, we applied LOLO-CV and spatially buffered LOLO-CV to assess the robustness of a Random Forest model for estimating PM2.5 in Pekanbaru. From an overfitting perspective, the gap between training metrics (MAE 1.14 μg/m³; R^2 0.86) and spatial LOLO-CV (MAE 2.07 μg/m³; R^2 0.57) indicates a generalization gap that is reasonable for this data size rather than evidence of short-distance target leakage. Performance at 0–2 km buffers is essentially unchanged (MAE ≈ 2.07 μg/m³; $R^2 \approx 0.57$) and remains very similar at 3 km (MAE = 2.06 μg/m³; $R^2 = 0.55$), indicating that excluding neighbors up to 3 km does not materially affect accuracy. Degradation appears at larger buffers (4–5 km: MAE = 2.20 μg/m³ / $R^2 = 0.51$; MAE = 2.43 μg/m³ / $R^2 = 0.44$; see Table 2), consistent with reduced nearby information for spatial generalization. Overlapping 95% confidence intervals up to 3 km (e.g., 0 km: MAE 1.38–2.84; R^2 0.23–0.76; 3 km: MAE 1.33–2.86; R^2 0.22–0.75) support the conclusion that performance differences at these distances are not statistically significant, whereas the shifts at 4–5 km indicate genuine degradation.

Our spatial-validation performance ($R^2 \approx 0.55$–0.57 for buffers ≤3 km) is comparable to results in other data-scarce regions. For example, Adong et al. reported $R^2 = 0.62$ for a Random Forest combining MODIS MAIAC AOD and low-cost sensors in Kampala, Uganda [21], while Arowosegbe et al. obtained a spatial-CV $R^2 = 0.48$ for PM10 in South Africa using an RF/GBM/SVR ensemble [22]. These studies, like ours, highlight common challenges in data-poor contexts, including sparse ground monitoring and missing AOD retrievals. Unlike ensemble setups, our single Random Forest framework integrates temporally adjusted low-cost sensor data, satellite AOD, and land-use predictors, providing a simpler yet effective solution with competitive accuracy.

As a benchmark to contextualize the RF results, we also trained simple linear baselines (Ridge) using the same features and the same LOLO-CV protocol, with the regularization parameter α tuned by inner LOLO-CV. Although the Ridge baseline shows slightly higher LOLO performance (MAE 1.743 μg/m³ [95% CI: 1.090–2.403]; R^2 0.670 [0.383–0.818]) than RF (MAE 2.067 μg/m³ [1.383–2.840]; R^2 0.568 [0.23–0.76]), the overlapping confidence intervals indicate that the difference is not statistically definitive. We retain RF as the primary model because (i) it captures nonlinear relationships and interactions between AOD and land-use composition that are difficult to model linearly; (ii) it is more robust to multicollinearity; and (iii) it offers local interpretability via SHAP (direction and magnitude of each predictor's effect at each location). Moreover, our goal is to produce spatial maps that are locally interpretable and easy to replicate; RF satisfies these needs within a single, simple framework, while linear baselines serve as comparators to validate the robustness of our findings.

Overall, Random Forest captures local spatial variability of PM2.5 at 5×5 km resolution, but its generalization weakens as spatial separation increases. The combination of LOLO-CV, spatial buffering, and hyperparameter tuning indicates that overfitting risk is reasonably controlled, although the limited number of sites still constrains performance at highly isolated locations.

6.4 Model implementation

The developed model was subsequently used to predict PM2.5 concentrations in other grid areas without measurement data across the entire Pekanbaru City. The prediction results for all grids are presented in Table 3.

To assess air quality levels based on these values, the classification system of Indonesia's Air Pollution Standard Index (ISPU), as regulated by the Indonesian Ministry of Environment and Forestry Regulation No. 14 of 2020, was used as the main reference, alongside the U.S. Environmental Protection Agency (EPA) Air Quality Index (AQI) as an international benchmark.

Table 2. Validation Model Using LOLO-CV and Spatially Buffered LOLO-CV

Buffer Radius (km)	MAE (μg/m³)	CI MAE	R^2	CI R^2
0 (LOLO-CV)	2.06	1.38 – 2.84	0.57	0.23 – 0.76
1	2.06	1.38 – 2.84	0.57	0.23 – 0.76
2	2.06	1.38 – 2.84	0.57	0.23 – 0.76
3	2.06	1.33 – 2.86	0.55	0.22 – 0.75
4	2.20	1.47 – 2.99	0.51	0.19 – 0.70
5	2.43	1.68 – 3.25	0.44	0.03 – 0.63

Table 3. Predicted PM2.5 Concentration for All Grids

Grid_ID	Predicted (μg/m3)	Grid_ID	Predicted (μg/m3)	Grid_ID	Predicted (μg/m3)	Grid_ID	Predicted (μg/m3)
1	33.72	11	25.70	20	30.49	29	26.00
2	33.81	12	26.78	21	29.89	30	25.52
3	33.29	13	34.86	22	24.52	31	28.61
4	26.22	14	34.44	23	26.19	32	25.60
5	26.04	15	32.35	24	26.12	33	26.02
7	33.00	16	29.93	25	28.99	34	25.29
8	33.84	17	25.66	26	28.15	35	25.85
9	30.75	18	26.34	27	28.32	38	27.08
10	28.71	19	30.15	28	25.37	39	25.98

The estimated PM2.5 concentrations across 36 grid areas in Pekanbaru City ranged from 24.52 to 34.86 μg/m³. According to the ISPU, PM2.5 concentrations between 15.6 and 55.4 μg/m³ fall into the "Moderate" category. Thus, all grids in this study are classified as "Moderate". The lowest predicted concentration was recorded in Grid 22 (24.52 μg/m³), which, while close to the lower threshold of the "Moderate", does not yet qualify for the "Good" category (0–15.5 μg/m³). On the other hand, the highest predicted values were found in Grid 13 (34.86 μg/m³) and Grid 14 (34.44 μg/m³), both nearing the upper boundary of the "Moderate". These areas warrant close attention, as a slight increase in PM2.5 levels could push them into the "Unhealthy" category (>55.4 μg/m³).

In general, the "Moderate" category indicates that air quality is still acceptable for most of the population, but it may pose health risks to sensitive groups such as children, the elderly, and individuals with respiratory illnesses. Therefore, although the predictions suggest relatively moderate air quality, vigilance is still required in areas where PM2.5 levels approach the upper threshold of this category.

When evaluated using the EPA's AQI standard, PM2.5 concentrations between 12.1 and 35.4 μg/m³ are also classified as "Moderate." Accordingly, all grid areas fall within the same category under the EPA classification, reinforcing the consistency of results between the two systems.

6.5 Visualization and spatial analysis

The estimated PM2.5 concentrations for each grid were subsequently visualized in a map, as shown in Figure 7. Based on the analysis of the relationship between PM2.5 concentrations and land cover types, it was observed that grids dominated by plantation areas such as Grids 4, 5, 18, 22, 23, 24, and 27–35, as well as Grid 39 generally exhibited lower predicted PM2.5 values, ranging between 24.52–28.99 μg/m³. This finding aligns with the assumption that large vegetated areas, such as plantations, have the potential to absorb pollutants or are typically located farther from dense emission sources, thereby reducing PM2.5 concentrations.

Conversely, grids predominantly covered by residential land use, such as Grids 1–3, 7–10, 13–21, and others, tended to show higher predicted PM2.5 concentrations. Some of these grids, including Grid 13 (34.86 μg/m³), Grid 14 (34.44 μg/m³), and Grid 2 (33.81 μg/m³), were even close to the upper threshold of the "moderate" category, indicating higher particulate accumulation in densely populated residential areas.

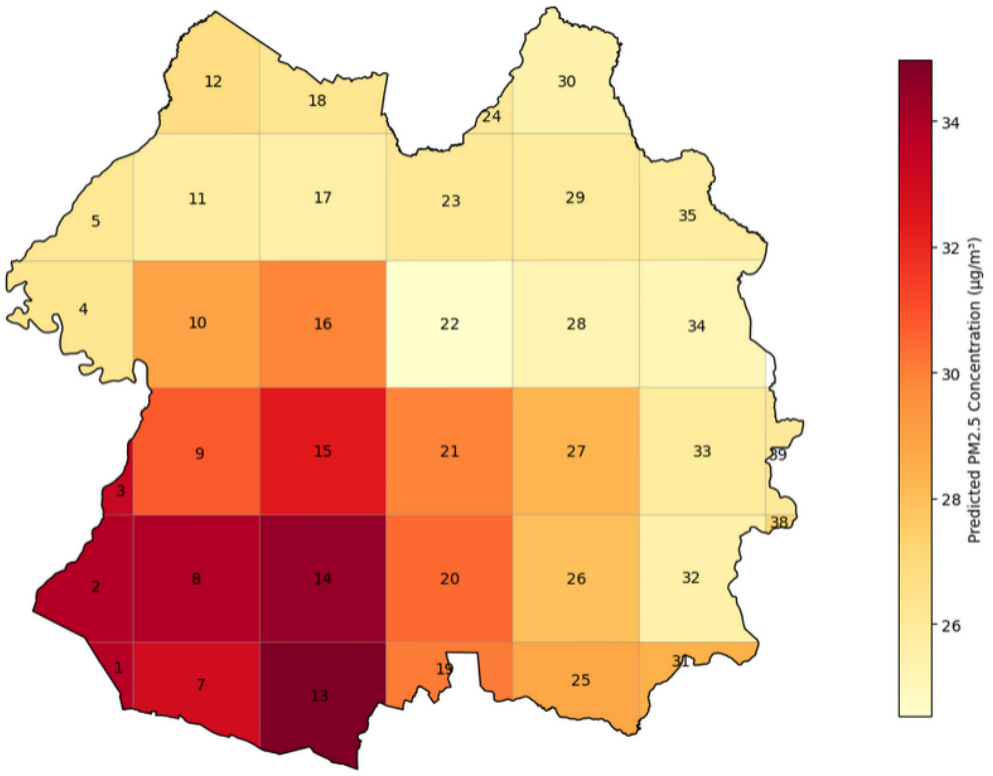

Figure 7. Visualization of Predicted PM2.5 Concentration in Pekanbaru Map

This is attributed to increased anthropogenic activity, vehicular emissions, and sparse vegetation. Interestingly, Grid 12, which is dominated by Secondary Dryland Forest, also recorded a relatively low PM2.5 value (26.78 µg/m³), suggesting that areas with secondary forest cover may also contribute positively to air quality.

Overall, the combination of land cover types significantly influenced the predicted PM2.5 concentrations. Vegetative land types such as plantations and secondary forests were strongly associated with lower PM2.5 levels, while residential areas correlated with higher predicted concentrations. These findings offer practical insights for air quality management and urban planning in Pekanbaru. For example, grid areas identified with consistently high predicted PM2.5 concentrations, particularly those dominated by residential and peatland land use, can be prioritized for targeted emission control measures. These may include strengthening peatland fire prevention programs, promoting green buffers in dense residential zones, or integrating air quality criteria into spatial zoning policies.

7. Conclusion

This study successfully developed a spatial predictive model to map PM2.5 concentrations in Pekanbaru (2024) using a Random Forest (RF) trained on temporally adjusted low-cost sensor data, land-use composition (11 classes), and satellite-derived AOD. Hyperparameters were tuned with RandomizedSearchCV under a Leave-One-Location-Out (LOLO) scheme, and spatial generalization was assessed with LOLO and spatially buffered LOLO to avoid near-neighbor bias and target leakage. Model performance under spatial validation was stable up to 3 km buffers (MAE ≈ 2.0–2.1 µg/m³; $R^2 \approx 0.55$–0.57) and declined at larger buffers (4–5 km), indicating that overfitting risk was reasonably controlled while generalization weakens as nearby information diminishes, an expected outcome in a data-scarce setting with only 18 sites. Although a Ridge baseline scored slightly higher in point estimates, the overlapping confidence intervals make the difference non-definitive; RF is retained because it captures nonlinearities and interactions and provides local interpretability via SHAP.

SHAP analysis shows AOD as the most influential predictor (generally increasing PM2.5), followed by positive contributions from peatlands and mitigating effects from vegetated land (e.g., plantations). Residential and transport infrastructure provide additional, smaller signals consistent with anthropogenic emissions. These patterns align with domain expectations and support the substantive interpretability of the RF outputs.

City-wide predictions for 36 grids ranged from 24.52 to 34.86 μg/m³, placing all areas in the "Moderate" category under both Indonesia's ISPU and the U.S. EPA AQI. Grids dominated by plantations and secondary forest tended to have lower predicted PM2.5 levels, whereas densely residential areas had higher levels, making them more useful for prioritizing mitigation measures such as strengthening peat-fire prevention, adding green buffers in residential zones, or integrating air-quality criteria into zoning.

Overall, the RF framework delivers a simple yet effective and interpretable tool for spatial PM2.5 estimation in data-limited urban contexts. Future work should expand the number of sites and monitoring duration, incorporate meteorological variables, and develop strategies for handling missing AOD retrievals. It should also test these models on independent datasets and map predictive uncertainty steps to further improve robustness and generalizability to more isolated locations.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Individual contributions for this research: Conceptualization, RTW; methodology, RTW; software, RTW; validation, DH, MRT, and DDS; formal analysis, RTW; investigation, RTW; resources, RTW; data curation, RTW; writing—original draft preparation, RTW; writing—review and editing, RTW; visualization, RTW; supervision, DH, MRT, DDS.

Acknowledgments

We would like to express our gratitude to Politeknik Caltex Riau and Universiti Tun Hussein Onn Malaysia for their academic and financial support during the research process.

References

[1] C. Liu *et al.*, "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities", *New England Journal of Medicine*, Vol. 381, No. 8, pp. 705–715, 2019, doi: 10.1056/nejmoa1817364.

[2] R. B. Hayes *et al.*, "PM2.5 air pollution and cause-specific cardiovascular disease mortality", *Int J Epidemiol*, Vol. 49, No. 1, pp. 25–35, Feb. 2020, doi: 10.1093/ije/dyz114.

[3] V. C. Pun, F. Kazemiparkouhi, J. Manjourides, and H. H. Suh, "Long-Term PM2.5 Exposure and Respiratory, Cancer, and Cardiovascular Mortality in Older US Adults", *Am J Epidemiol*, Vol. 186, No. 8, pp. 961–969, 2017, doi: 10.1093/aje/kwx166.

[4] R. T. Wahyuni, D. Hanafi, M. R. Tomari, and D. S. S. Sahid, "Research trends in spatial modeling of PM2.5 concentration using machine learning: a bibliometric review", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 37, No. 2, pp. 1317–1327, 2025, doi: 10.11591/ijeecs.v37.i2.pp1317-1327.

[5] B. Guo *et al.*, "Estimating PM2.5 concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017", *Science of the Total Environment*, Vol. 778, pp. 1–14, 2021, doi: 10.1016/j.scitotenv.2021.146288.

[6] Y. Liu, G. Cao, N. Zhao, K. Mulligan, and X. Ye, "Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach", *Environmental Pollution*, Vol. 235, pp. 272–282, 2018, doi: 10.1016/j.envpol.2017.12.070.

[7] K. Goyal and S. Goyal, "Predicting PM2.5 Air Quality Using Random Forest Regression Enhanced with Polynomial Features", *International IEEE Conference proceedings, IS*, pp. 1–6, 2024, doi: 10.1109/IS61756.2024.10705219.

[8] W. Mulyana, N. Ardhyarini, and H. Pratiwi, "Urban Analysis Report 2020: Pekanbaru", *CRIC Project, co-funded by the European Union*, 2020.

[9] W. Hidayat, "Dynamics of Spatial Transformation in Pekanbaru City During the Era of Regional Autonomy", *International Journal of Architecture and Urbanism*, Vol. 8, No. 1, pp. 125–130, 2024, doi: 10.32734/ijau.v8i1.15120.

[10] J. Xiong, R. Yao, W. Wang, W. Yu, and B. Li, "A spatial-and-temporal-based method for rapid particle concentration estimations in an urban environment", *J Clean Prod*, Vol. 256, pp. 1–7, 2020, doi: 10.1016/j.jclepro.2020.120331.

[11] G. Geng *et al.*, "Tracking Air Pollution in China: Near Real-Time PM2.5Retrievals from Multisource Data Fusion", *Environ Sci Technol*, Vol. 55, No. 17, pp. 12106–12115, 2021, doi: 10.1021/acs.est.1c01863.

[12] R. T. Wahyuni, J. N. Sari, K. D. K. Wardhani, T. Arfan, D. S. S. Sahid, and S. A. Zainuddin, "Spatial Analysis of PM2.5 Data from Low-

Cost Sensor Related to Economic Activities in Pekanbaru City", *International Journal of Sustainable Development and Planning*, Vol. 20, No. 1, pp. 1–12, 2025, doi: 10.18280/ijsdp.200101.

[13] Q. Cui, F. Zhang, S. Fu, X. Wei, Y. Ma, and K. Wu, "High Spatiotemporal Resolution PM2.5 Concentration Estimation with Machine Learning Algorithm: A Case Study for Wildfire in California", *Remote Sens (Basel)*, Vol. 14, No. 7, pp. 1–17, 2022, doi: 10.3390/rs14071635.

[14] K. Chastko and M. Adams, "Assessing the accuracy of long-term air pollution estimates produced with temporally adjusted short-term observations from unstructured sampling", *J Environ Manage*, Vol. 240, pp. 249–258, 2019, doi: 10.1016/j.jenvman.2019.03.108.

[15] X. Li *et al.*, "Factors Underlying Spatiotemporal Variations in Atmospheric PM2.5 Concentrations in Zhejiang Province, China", *Remote Sens (Basel)*, Vol. 13, pp. 1–20, 2021.

[16] S. Park *et al.*, "Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models", *Science of the Total Environment*, Vol. 713, pp. 1–13, 2020, doi: 10.1016/j.scitotenv.2020.136516.

[17] Q. He, M. Wang, and S. H. L. Yim, "The spatiotemporal relationship between PM2.5 and aerosol optical depth in China: Influencing factors and implications for satellite PM2.5 estimations using MAIAC aerosol optical depth", *Atmos Chem Phys*, Vol. 21, No. 24, pp. 18375–18391, 2021, doi: 10.5194/acp-21-18375-2021.

[18] R. Li, X. Mei, L. Chen, Z. Wang, Y. Jing, and L. Wei, "Influence of spatial resolution and retrieval frequency on applicability of satellite-predicted pm2.5 in northern China", *Remote Sens (Basel)*, Vol. 12, No. 4, pp. 1–14, 2020, doi: 10.3390/rs12040736.

[19] C. Brokamp, "A high resolution spatiotemporal fine particulate matter exposure assessment model for the contiguous United States", *Environmental Advances*, vol. 7, pp. 1–9, 2022, doi: 10.1016/j.envadv.2021.100155.

[20] A. Stock and A. Subramaniam, "Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing", *GIsci Remote Sens*, Vol. 59, No. 1, pp. 1281–1300, 2022, doi: 10.1080/15481603.2022.2107113.

[21] P. Adong, A. Pakrashi, and S. Dev, "Using Random Forest to Estimate Low-Cost Sensor PM25 from MODIS MAIAC AOD in Kampala", In: *Proc. of 2024 17th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2024*, Shanghai, pp. 1–5, 2024, doi: 10.1109/CISP-BMEI64163.2024.10906267.

[22] O. O. Arowosegbe *et al.*, "Ensemble averaging using remote sensing data to model spatiotemporal PM10 concentrations in sparsely monitored South Africa", *Environmental Pollution*, Vol. 310, pp. 1–10, 2022, doi: 10.1016/j.envpol.2022.119883.