JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 3, July 2026, pp. 829-843

Comparative Evaluation of Eigenvector Selection in Eigenvector Spatial Filtering using a Gradient Boosting Machine for $PM_{2.5}$ Concentration Prediction

Putri Nisrina Az-Zahra[1*], Anik Djuraidah[1], Erfiani[1]
[1]Statistics and Data Science, IPB University, Indonesia
putrinisazzahra@apps.ipb.ac.id

Article History:
Received : 28-03-2026
Revised : 16-04-2026
Accepted : 02-05-2026
Online : 01-07-2026

Keywords:
Eigenvector Spatial Filtering;
Gradient Boosting Machine;
SHAP;
PM2.5.

ABSTRACT

Spatial dependence remains a critical issue in spatial data analysis. To address this issue, various eigenvector selection methods within the Eigenvector Spatial Filtering (ESF) framework have been proposed. However, these methods often do not provide explicit information regarding the individual contribution of each spatial component, limiting model interpretability, particularly when dealing with a large number of candidate eigenvectors and complex models. In addition, ESF has limitations in capturing nonlinear relationships and complex interactions inherent in spatial data, while its integration with advanced feature selection methods within machine learning frameworks remains underexplored. This quantitative empirical study aims to evaluate different eigenvector selection methods within ESF integrated with a Gradient Boosting Machine (GBM) model for predicting $PM_{2.5}$ concentrations in DKI Jakarta. Data were collected from 100 monitoring stations across five administrative regions for the first half of 2025. Spatial eigenvectors were derived from a spatial weights matrix and selected using four methods: positive eigenvalues, Moran's Index significance, LASSO regression, and SHAP values obtained from the GBM model. Model performance was assessed using both 10-fold random cross-validation and spatial blocked cross-validation to evaluate predictive accuracy and spatial generalization. The results showed that adding spatial eigenvectors significantly improved the model performance compared to models without spatial components. Under 10-fold cross-validation, the SHAP-based selection method achieved the highest predictive accuracy (R^2 = 0.619), effectively capturing spatial dependence and nonlinear relationships. The SHAP method demonstrated robustness by selecting stable and consistent spatial components across different regions. These findings highlight the methodological advantage of integrating ESF with machine learning and SHAP-based feature selection, offering a more interpretable and robust framework for spatial modelling. Practically, the improved prediction of $PM_{2.5}$ concentrations can support more accurate air quality assessments and inform environmental management strategies in urban areas.

https://doi.org/10.31764/jtam.v10i3.38883

A. INTRODUCTION

Spatial dependence refers to the pattern of interrelationships between locations in spatial data, where events at one location are related to those at nearby locations. When the pattern of spatial dependence shows a systematic relationship between the value of a variable at one location and the value of the same variable at surrounding locations, this phenomenon is called spatial autocorrelation, which, if not properly addressed, can lead to bias in the statistical models. To address this issue, the eigenvector spatial filtering (ESF) method has shown

promising results in handling the problem compared to other spatial regression models, and the effectiveness of this approach has been demonstrated in studies where ESF was able to significantly reduce spatial autocorrelation (McCord et al., 2020; Sun et al., 2023).

ESF utilizes the eigenvectors of a spatial weight matrix (SWM) to capture spatial dependence patterns in the data and incorporates them into the regression model as additional independent variables, thereby improving the quality of the prediction models (Zhang et al., 2018). Linear combinations of these eigenvectors filter out positive spatial autocorrelation, allowing the modelling process to proceed as if each observation is independent. In addition, the ESF approach can deliver better performance with a lighter computational load by constructing eigenvectors using Moran's index (Murakami et al., 2017). The ESF approach, combined with the eigenvector spatial filtering–variance component (ESF-VC) model, has also been demonstrated to enhance the accuracy of regression parameter estimates and effectively address spatial dependence issues (Mahkya et al., 2024).

The selection of optimal feature vectors is one aspect that must be considered in the implementation of ESF. This stage is important in modelling to avoid overfitting, enhance the model interpretability, and reduce the computational load. Approaches used include employing all feature vectors with positive eigenvalues (Murakami & Griffith, 2019), selection based on the Moran's Index (Griffith & Chun, 2019), stepwise procedures, and the Least Absolute Shrinkage and Selection Operator (LASSO) (Seya et al., 2015). Although effective for partial selection, these feature vector selection methods do not provide explicit information regarding the individual contribution of each feature vector to the model's prediction. This poses a challenge in interpretation, especially when the number of candidate feature vectors is very large, and the model applied is complex. In addition, the ESF has limitations in capturing nonlinear relationships and complex interactions inherent in spatial data. Although machine learning models can model nonlinearities, the integration of ESF with advanced feature selection methods to optimize spatial component selection remains underexplored.

To address these limitations, the SHAP method can be employed for feature selection, in this case for feature vectors, with a strong foundation in game theory (Marcilio & Eler, 2020). SHAP quantifies the contribution of each feature vector to the model output using Shapley values, providing a consistent and interpretable measure of importance. Feature vectors with higher SHAP values can be prioritized, whereas those with lower contributions can be excluded. This approach enhances both the predictive performance and interpretability of the spatial structure. Given these considerations, integrating the ESF with machine learning approaches offers a promising direction, as these methods can capture complex and nonlinear patterns, particularly in large-scale geospatial datasets (Liu et al., 2022). Previous studies have integrated the ESF with various statistical and machine learning models to enhance the modelling performance (Ahmadi et al., 2024; Islam et al., 2022; Wang et al., 2023). Among these approaches, the Gradient Boosting Machine (GBM) is particularly effective, as it constructs models iteratively by combining multiple weak learners in the form of decision trees to reduce prediction errors, enabling flexible modelling of nonlinear relationships, and interactions among variables (Singh et al., 2021).

This study introduces SHAP values derived from a GBM as a novel eigenvector selection method within the ESF framework. This approach enhances the ability to capture both spatial

dependence and nonlinear relationships in air quality prediction. Poor air quality generally occurs in developing countries with high population densities, particularly in urban areas (Shaddick et al., 2020). The province of DKI Jakarta, as the capital city of Indonesia, faces increasingly critical air pollution issues because of its high population density, rapid urbanization, and intense use of transportation. These conditions often cause $PM_{2.5}$ concentrations to exceed safe limits (Kusumaningtyas et al., 2021). Several previous studies have shown that $PM_{2.5}$ concentrations have a significant spatial distribution pattern (Sotoudeheian & Arhami, 2021; Xu et al., 2021), leading to complex spatial and nonlinear patterns that are difficult for traditional models to capture, while SHAP improves eigenvector selection by making it more interpretable and effective for handling spatial heterogeneity.

This study aims to evaluate various eigenvector selection methods within the ESF framework, including positive eigenvalues, Moran's Index, LASSO, and SHAP, combined with a GBM to predict $PM_{2.5}$ air quality in the DKI Jakarta region. Beyond methodological contributions, this study also supports the Sustainable Development Goals (SDGs). Improved $PM_{2.5}$ prediction can contribute to better health risk assessment and pollution management in line with SDG 3 on good health and well-being, provide insights for enhancing urban environmental quality as emphasized in SDG 11 on sustainable cities and communities, and support climate-related mitigation efforts consistent with SDG 13 on climate action.

B. METHODS

This quantitative empirical study evaluates eigenvector selection methods within the ESF integrated with a GBM. The dataset consists of monthly $PM_{2.5}$ concentrations ($\mu g/m^3$) for the first half of 2025 in the Province of DKI Jakarta, along with several explanatory variables, including relative humidity (%), surface temperature (K), wind speed (m/s), air pressure (Pa), rainfall (mm/day), aerosol optical depth (10^{-3}), normalized difference vegetation index (%), industrial density (industry/km^2), and distance to the road network (m). Environmental variables at 1 km resolution were interpolated using inverse distance weighting (IDW), after aggregating duplicate ERA5 grid values into unique observations at their central locations. All data processing and analyses were conducted using Python.

Data were obtained from air quality monitoring stations, Google Earth Engine, and OpenStreetMap, followed by preprocessing and exploratory analysis. SWM was constructed to characterize spatial relationships among locations, and spatial dependence was subsequently assessed. Eigen decomposition was then performed within the ESF framework to derive spatial eigenvectors. Feature selection was conducted using positive eigenvalues, Moran's Index, LASSO, and SHAP. In the SHAP-based approach, a GBM model was trained to compute SHAP values, which quantify the contribution of each eigenvector. The selected eigenvectors were then integrated with the explanatory variables to construct the final dataset. Subsequently, the data were partitioned using random 10-fold cross-validation and spatial block cross-validation. Model development was performed using GBM with different eigenvector selection strategies, and model performance was evaluated using appropriate metrics. A flowchart of the research procedure is presented in Figure 1.

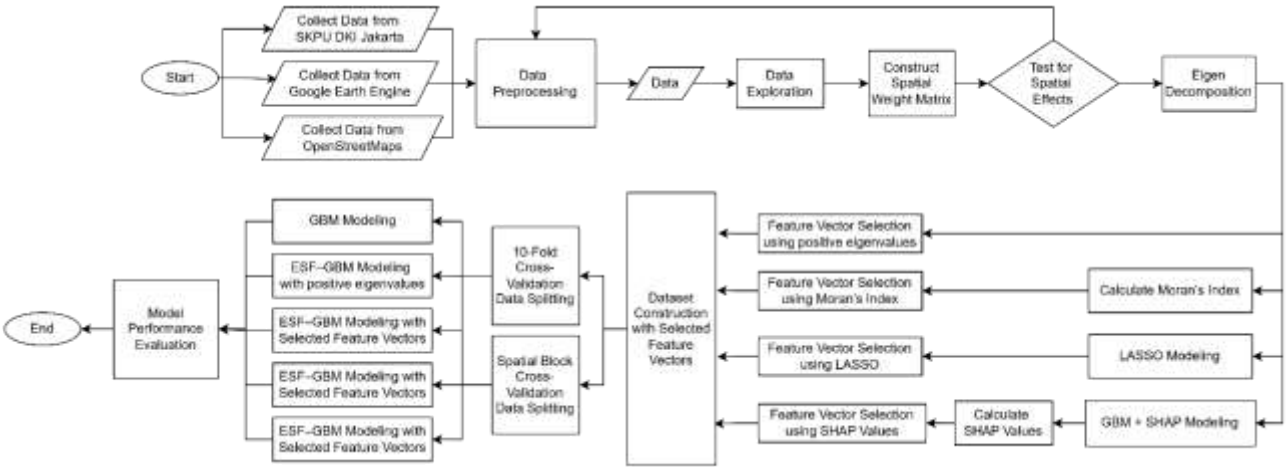


Figure 1. Flowchart of the research methodology

1. $PM_{2.5}$ Concentration

Particulate matter ($PM_{2.5}$) refers to fine particles with a diameter of less than 2.5 μm that are suspended in the air. The extremely small size of these particles allows them to penetrate the respiratory tract all the way to the alveoli, potentially causing adverse health effects (Crinnion, 2017). Short-term exposure to $PM_{2.5}$ has been linked to increased respiratory symptoms and lung diseases such as asthma. Furthermore, long-term exposure is known to increase the risk of cardiopulmonary diseases, stroke, and even death. These particles consist of a complex mixture of solid and liquid components and originate from two main sources. Primary particles are directly released into the air through emissions. Secondary particles are formed in the atmosphere through chemical reactions of precursor gases. The physical and chemical characteristics of $PM_{2.5}$ greatly depend on the source, geographical location, and atmospheric conditions, causing its concentrations to vary spatially and temporally.

2. Eigenvector Spatial Filtering (ESF)

The eigenvector spatial filtering (ESF) approach is based on Moran's coefficient (MC) eigenvectors obtained from the spatial weights matrix. The eigenvectors are extracted through eigen decomposition, as follows (Murakami & Griffith, 2019):

$$\boldsymbol{MWM} = \boldsymbol{E}^{*}\boldsymbol{\Lambda}^{*}\boldsymbol{E}^{*\prime} \tag{1}$$

with $\boldsymbol{M} = (\boldsymbol{I} - \mathbf{1}\mathbf{1}'/n)$ as the centering matrix and $\boldsymbol{W}$ as the symmetric spatial weight matrix of $n\,x\,n$ size with zeros on the diagonal. $\boldsymbol{E}^{*} = [\boldsymbol{e_1}, \dots, \boldsymbol{e_n}]$ is the eigenvector matmodellinggis a diagonal matrix whose elements are the square roots of the eigenvalues $\{\lambda_1, \dots, \lambda_N\}$. These eigenvectors can be interpreted in the context of the Moran's coefficient, which is a diagnostic statistic for measuring spatial autocorrelation. The Moran's coefficient for the random variable vector $\boldsymbol{y}$ is defined as follows.

$$\mathrm{MC}[\boldsymbol{y}] = \frac{n}{\mathbf{1}\prime\boldsymbol{W}\mathbf{1}}\frac{\boldsymbol{y}\prime\boldsymbol{MWMy}}{\boldsymbol{y}\prime\boldsymbol{My}} \tag{2}$$

Moran's coefficient M$C[\boldsymbol{y}]$ is positive if there is positive spatial autocorrelation and negative if there is negative spatial autocorrelation. The value of Moran's coefficient for the resulting feature vector is given in the following equation.

$$MC[\boldsymbol{e_l}] = \frac{n}{\mathbf{1'W1}} \frac{\boldsymbol{e'_l MWMe_l}}{\boldsymbol{e'_l Me_l}} = \frac{n}{\mathbf{1'W1}} \frac{\boldsymbol{e'_l E^* \Lambda^* E^* e_l}}{\boldsymbol{e'_l Me_l}} = \frac{n}{\mathbf{1'W1}} \lambda_l \tag{3}$$

Equation (3) indicates that the eigenvectors corresponding to positive eigenvalues signify the presence of positive spatial autocorrelation. The eigenvectors with higher MC are ordered from largest to smallest and form a matrix $\boldsymbol{E}^*$ of size $n\ x\ k$, with $k < n$. The eigenvector $\boldsymbol{e_1}$, which has the largest eigenvalue, also has the highest positive MC. Meanwhile, the eigenvector $\boldsymbol{e_2}$, which has the second largest eigenvalue, also has the second highest positive MC, and so on. These eigenvectors are arranged in such a way that each one is uncorrelated and orthogonal to the previous vector. The last eigenvector, $\boldsymbol{e_n}$, has the largest negative MC, reflecting the maximum level of negative spatial autocorrelation. If the order of these eigenvectors is visualized in map form, a pattern of gradual change in spatial autocorrelation becomes apparent.

3. Gradient Boosting Machine (GBM)

A gradient boosting machine (GBM) is a technique that combines several weak learners to form a stronger and more accurate model (Friedman, 2001). GBM works by building models sequentially, where each new model attempts to correct the errors made by the previous model by minimizing the loss function using a gradient descent approach. Mathematically, the gradient boosting tree (GBT) model can be represented as the sum of regression trees.

$$F(x) = \sum_{i=1}^{n} f_i(x) \tag{4}$$

with $f_i(x)$ is a decision tree. This model is built incrementally by adding new decision trees $f_{n+1}(x)$ based on the following optimization equation.

$$\underset{f_{n+1}}{\arg min} \sum_{t=1}^{N} L(y_t, F_n(x_t) + f_{n+1}(x_t)) \tag{5}$$

where *L(.)* is a differentiable loss function, and this model can capture nonlinear relationships. This optimization is solved using the steepest descent method, where the model iteratively updates the prediction function by adding new decision trees that reduce prediction errors based on the gradient of the loss function.

4. Shapley Additive exPlanations (SHAP)

The Shāpley additive explanations (SHAP) method explains how each feature contributes to the outcome of a model's prediction. This method originates from game theory concepts and calculates the extent to which each feature influences the model's output in an additive and consistent manner (Lundberg & Lee, 2017). Mathematically, the Shapley value is expressed as follows:

$$\emptyset_j(v) = \sum_{S \subseteq M\{j\}} \frac{|S|!(M-|S|-1)!}{M!}[v(S \cup \{j\}) - v(S)], \quad j = 1, \dots, M \tag{6}$$

with M being the total number of variables, S is a subset of other variables of size $|S|$, is the $v(S \cup \{j\})$ model's prediction when variable j is included, and $v(S)$ is the model's prediction when variable j is not included. One of the advantages of SHAP is its ability to provide an interpretation for each variable's contribution to the model's prediction for every individual observation. To obtain a global interpretation, the contribution of the j-th variable can be averaged across all observations as a measure of global feature importance, namely:

$$FI_j = \frac{1}{n}\sum_{i=1}^{n} |\emptyset_j^{(i)}| \tag{7}$$

C. RESULT AND DISCUSSION

1. Data Exploration

$PM_{2.5}$ concentration data obtained from the air quality monitoring system in the DKI Jakarta Province for the period from January 2025 to June 2025 consisted of 439 rows of data aggregated monthly. In total, there were 100 monitoring points spread across the five administrative regions of DKI Jakarta, namely, 23 points in West Jakarta, 12 points in Central Jakarta, 27 points in South Jakarta, 32 points in East Jakarta, and six points in North Jakarta. The PM2.5 concentration distribution was categorized based on the air quality index (AQI), as shown in Figure 2. Overall, most observation points are marked in red, indicating poor air quality in DKI Jakarta throughout the first six months of 2025. The highest average $PM_{2.5}$ concentration was recorded in June (82.93 μg/m^3), followed by May (76.07 μg/m^3), February (59.89 μg/m^3), and April (59.63 μg/m^3), all of which fall into the unhealthy category. Meanwhile, March (49.92 μg/m^3) and January (49.19 μg/m^3) were categorized as unhealthy for sensitive groups. The minimum PM2.5 concentration value from January to December 2025 was 4.09 μg/m^3, which falls under the good AQI category, while the maximum value was 134.44 μg/m^3, which falls under the very unhealthy AQI category.

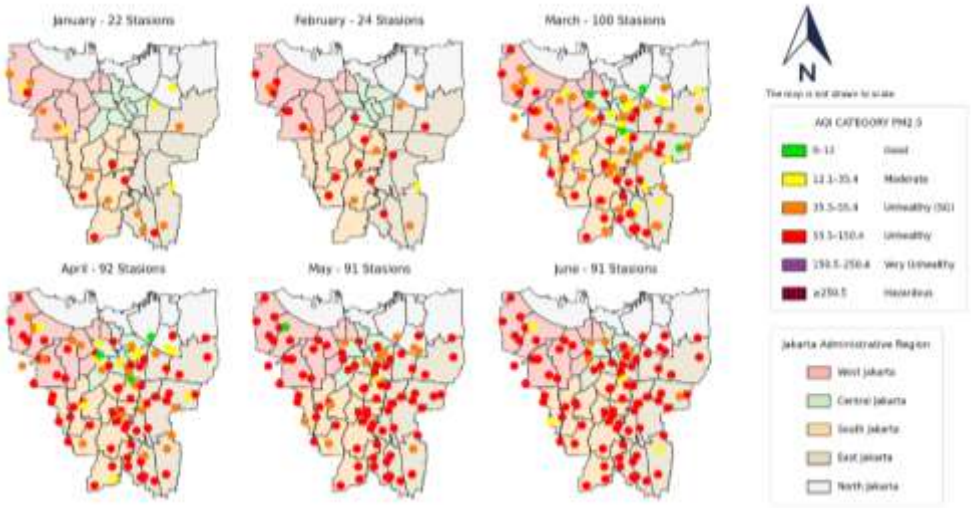


Figure 2. PM2.5 Distribution in DKI Jakarta from January to June 2025

The temporal distribution of $PM_{2.5}$ concentrations indicates an increasing trend from January to June. The presence of extreme values suggests unstable air quality conditions during certain periods. This pattern implies that $PM_{2.5}$ concentrations not only increase over time but also become more variable, which may affect the model stability and increase the prediction uncertainty, particularly for extreme values. Consequently, it is essential to have a model that can adeptly manage such variations and extremes.

Table 1. Descriptive Statistics of the Response Variable and Explanatory Variables

Variables	Description	Minimum	Maximum	Mean	Median
Y	PM2.5 Concentration	4.093	134.440	65.020	63.662
X_1	Relative Humidity	0.286	100.000	43.741	35.785
X_2	Surface Temperature	299.066	300.941	300.194	300.202
X_3	Wind Speed	0.193	2.681	0.864	0.841
X_4	Air Pressure	99,853.64	100,615.54	100,444.58	100,455.54
X_5	Rainfall	4.987	13.250	7.280	7.152
X_6	AOD	141.333	876.200	415.099	405.333
X_7	NDVI	0.041	0.532	0.273	0.271
X_8	Industrial Density	0.319	3.826	0.625	0.319
X_9	Road Network Distance	0.232	88.950	15.041	9.708

Table 1 shows the descriptive statistics for the variables included in the analysis. An average relative humidity of 43.71% indicates that the air is not overly humid. The average surface temperature was approximately 27°C, with relatively little variation throughout the observation period. An average wind speed of 0.864 m/s indicates weak wind conditions, resulting in a tendency for $PM_{2.5}$ concentrations to accumulate in the atmosphere. The average air pressure of 100,444 Pa indicates relatively stable atmospheric conditions. The average rainfall of 7.28 mm suggests moderate rainfall intensity, which has the potential to reduce $PM_{2.5}$ concentrations. The average AOD value of 0.415 indicates the presence of relatively high atmospheric aerosols and pollutants. The average NDVI value indicates the dominance of built-up areas with relatively low to moderate vegetation cover. Industrial density had an average value of 0.625 with an uneven distribution of industries. In addition, the road network distance was 15.04 m on average and 9.71 m at the median, indicating that most observation points were located near major roads, which are sources of transportation emissions.

The correlation plot in Figure 3 shows the relationship between $PM_{2.5}$ concentration as the response variable and the explanatory variables. The highest correlation is with Rainfall, which $r = -0.403$ displays a negative correlation, as rainfall intensity increases in certain areas and times, the $PM_{2.5}$ concentration in those areas and times decreases. A similar pattern is observed with Wind Speed, which $r = -0.302$ also shows that as wind speed increases, $PM_{2.5}$ concentration decreases, because pollutant particles are more likely to be carried away and dispersed to other areas. In addition, NDVI, AOD, relative humidity, and industrial density have positive correlations with $PM_{2.5}$ concentration, with respective values of $r = 0.256$, $r = 0.229$, $r = 0.155$, and $r = 0.107$. Meanwhile, road network density, air pressure, and surface temperature have correlation values close to zero, indicating that the linear relationship between these variables and $PM_{2.5}$ concentration is very weak at locations and during the periods studied. Multicollinearity testing using the variance inflation factor (VIF) was

conducted and showed that all explanatory variables have VIF values less than 5, indicating that there are no multicollinearity issues among the explanatory variables.

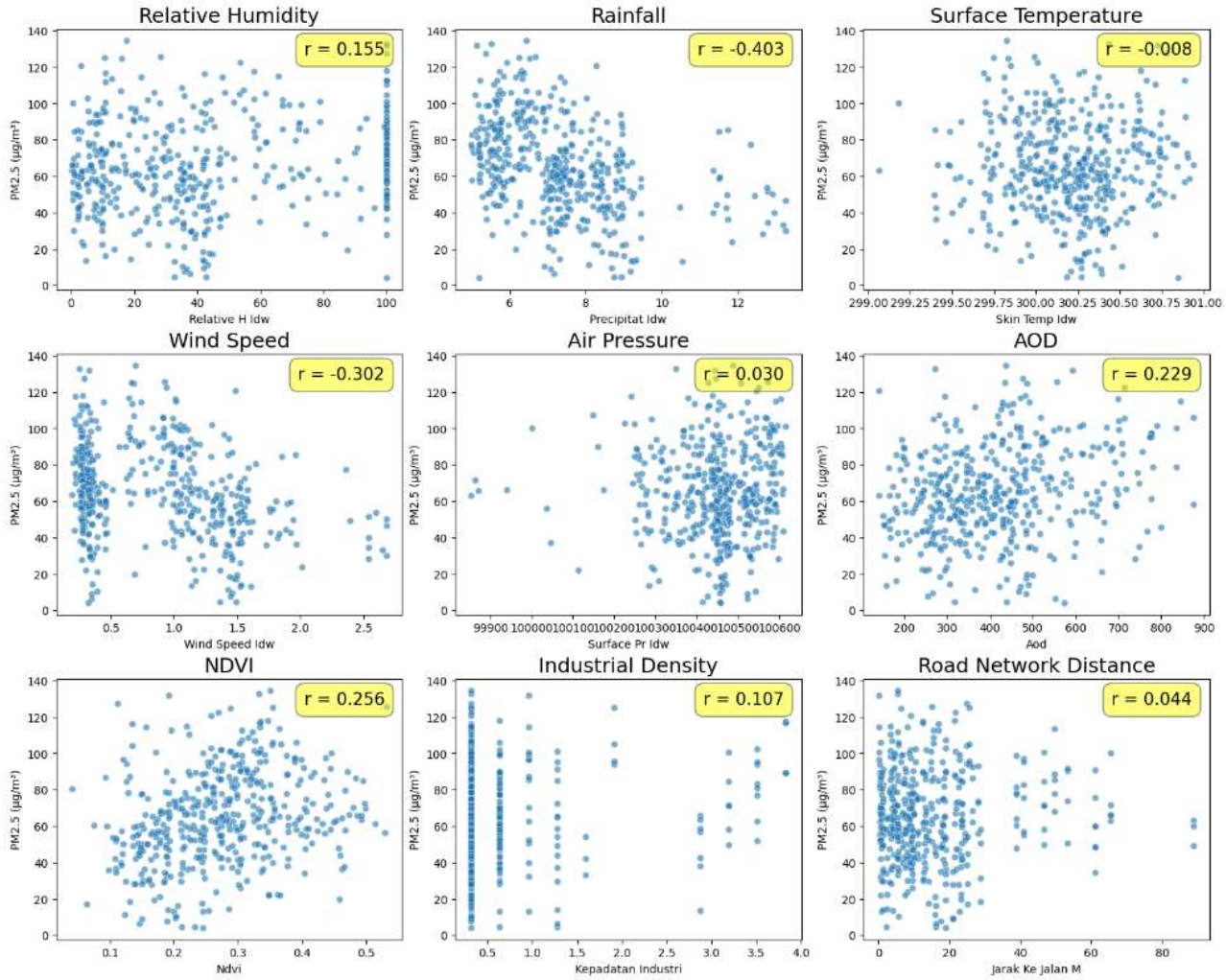


Figure 3. Correlation Between the Response Variable and the Independent Variable

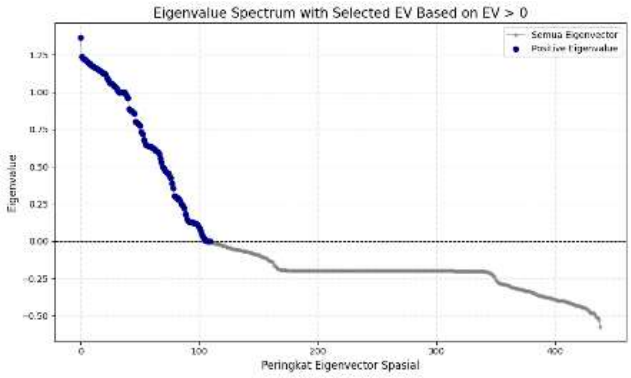
2. Eigenvector Selection Methods

The spatial weighting matrix based on the k-Nearest-Neighbor method with $k = 5$ was used to represent spatial relationships among observation locations. The Moran's Index value of 0.254 indicates that there are similar characteristics in $PM_{2.5}$ concentrations at nearby locations. The resulting p-value of 0.001 demonstrated a positive spatial autocorrelation among the observation point regions. Further spatial dependency testing with the LM test confirmed that $PM_{2.5}$ concentrations were spatially correlated across regions in DKI Jakarta. This suggests that air pollution levels in one location are influenced by those in nearby areas, highlighting the importance of incorporating spatial effects into the modelling process.

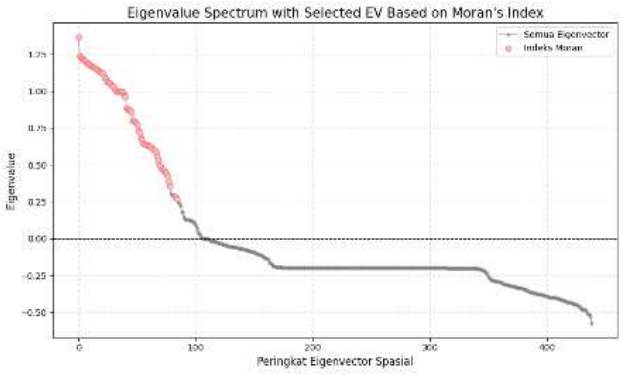
The ESF process was performed by decomposing the spatial weight matrix, resulting in 439 pairs of eigenvalues and eigenvectors. The selection of eigenvectors is necessary to prevent excessive model complexity due to a large number of spatial components. The first selection method was based on positive eigenvalues, resulting in 110 selected eigenvectors being obtained. Next, the selection method was based on testing Moran's index, which showed that

81 selected eigenvectors had a significant Moran's index. Another selection method compared in this study is based on LASSO, which retained 114 eigenvectors with nonzero coefficients. Furthermore, this study proposed the use of the SHAP method, constructed using the GBM model, as an eigenvector selection method. The boosting parameters were tuned on the dataset without the addition of eigenvectors using two dataset splitting validation scenarios, namely, 10-fold cross-validation and spatial blocked cross-validation with four folds.

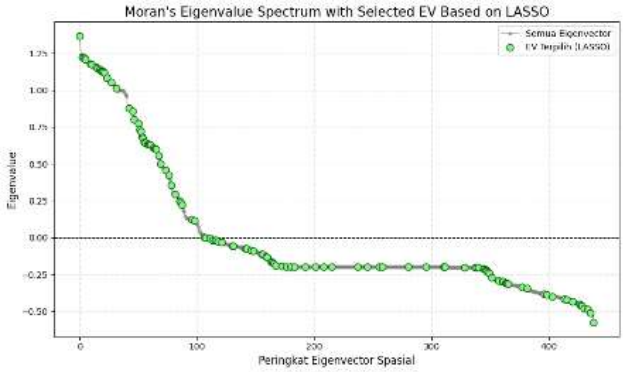
In the modelling using GBM for both validation scenarios, the model was trained on each fold, and SHAP values were calculated for all independent variables, including the 439 eigenvectors that were previously added to the dataset as additional independent variables. The contribution value of each eigenvector was calculated using the mean absolute SHAP value of the training data. A contribution threshold greater than 0.05 was set as the tool for eigenvector selection based on the SHAP method. The selection results showed that in the 10-fold cross-validation scenario, 221 eigenvectors met the selection criteria, whereas in the spatial blocked cross-validation scenario, 100 eigenvectors were selected. The smaller number of selected eigenvectors in the spatially blocked cross-validation scenario indicates that only spatial components that are stable and consistent across locations are selected, whereas in the 10-fold cross-validation scenario, more eigenvectors are selected because the model is still influenced by spatial proximity.



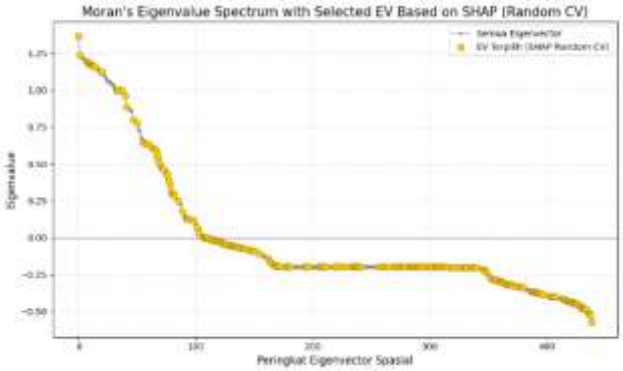
(1)



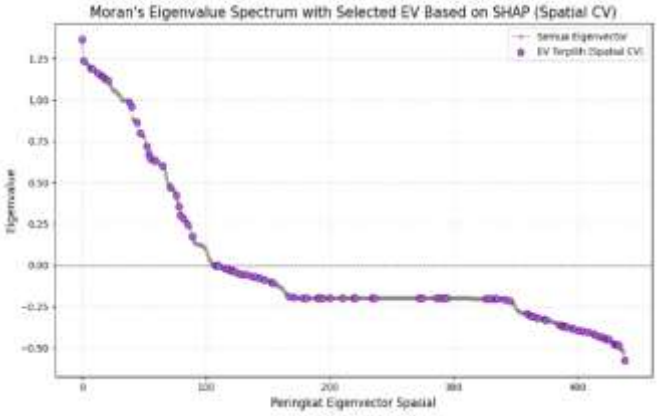
(2)



(3)



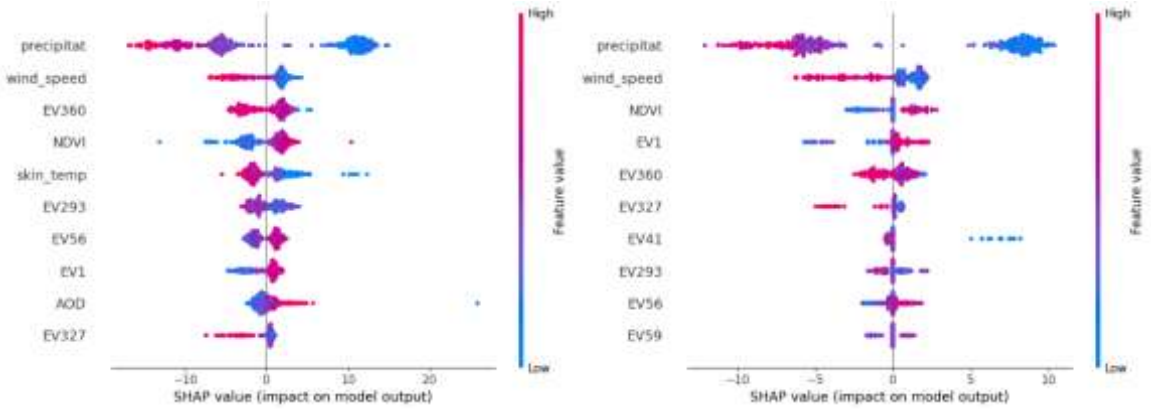
(4)



(5)

Figure 4. Spectrum of Eigenvalues with Selected Eigenvectors for Each Selection Method: (1) Positive Eigenvalue, (2) Moran's Index, (3) LASSO, (4) SHAP - Random CV, (5) SHAP - Spatial CV

Figure 4 shows the spectra of eigenvalues from the eigenvectors selected by each selection method. In Plot (1), which is selection based on the positive eigenvalue criterion, all selected eigenvectors have positive eigenvalues, in accordance with the initial criterion that only components representing positive spatial autocorrelation are retained. In Plot (2), selection based on the Moran's index at a 5% significance level also results in eigenvectors with positive eigenvalues. However, the number of selected eigenvectors is smaller than that in Plot (1) because eigenvectors with eigenvalues close to zero tend to be insignificant and are thus not selected. Next, Plot (3), which uses the LASSO selection method, shows that the selected eigenvectors are not limited to only positive eigenvalues but also include negative eigenvalues. This indicates that LASSO selects spatial components based on their contribution to the predictive model, regardless of the direction of spatial autocorrelation represented by the eigenvalues. Plot (4), with selection based on SHAP using the random cross-validation scenario with 10 folds, results in a relatively larger number of eigenvectors, with the eigenvalue distribution spread across both positive and negative sides. This shows that the predictive contribution-based approach can retain spatial components from various spectra of autocorrelation structures. Meanwhile, Plot (5), which is selection based on SHAP with spatial blocked cross-validation, shows a narrower spectrum of eigenvalues than LASSO and SHAP random cross-validation. The number of selected eigenvectors is smaller, indicating that only truly stable spatial components are retained in the model.



(1) (2)

Figure 5. Feature Contribution Based on SHAP in the GBM Model: (1) 10-Fold Cross-Validation and (2) Spatial Blocked Cross-Validation

A summary of feature contributions based on SHAP values is shown in Figure 5, where each point represents one observation for a particular variable. The Rainfall and Wind Speed variables had the largest contributions to the prediction model in both data-splitting scenarios. Low values of Rainfall and Wind Speed tended to increase the predictions, whereas high values tended to decrease them. Based on the results from the 10-fold cross-validation scenario, other independent variables that also had relatively large contributions were the NDVI, surface temperature, and AOD. For spatial components, the eigenvectors (EV) with the highest contributions included EV360, EV293, EV56, EV1, and EV327. Meanwhile, in the spatial blocked cross-validation scenario, the number of variables with significant contributions was more limited. In addition to the NDVI, the main contributions from the spatial components were shown by EV1, followed by EV360, EV327, EV41, EV293, EV56, and EV59. This indicates that, based on spatial validation, only some spatial components that are stable and consistent across regions continue to play an important role in the model.

The value of each observation in an eigenvector represents the relative position of that location within an underlying spatial pattern, where neighboring locations tend to have similar eigenvector values. Eigenvectors such as EV1 generally capture broader and more consistent spatial structures, reflecting large-scale spatial variation across regions. In contrast, eigenvectors such as EV360, EV293, and EV56 tend to represent more localized spatial patterns, capturing variations specific to certain areas. These differences indicate that each eigenvector contributes to the model by representing spatial variability at different scales. Eigenvectors associated with larger-scale patterns tend to provide more stable and generalizable information, while those representing localized patterns capture finer spatial heterogeneity but may be less consistent across regions. This highlights the importance of selecting appropriate eigenvectors, as the choice of spatial components directly influences how spatial structure is represented and learned by the model.

3. Model Evaluation

$PM_{2.5}$ concentration prediction was performed using the previously selected eigenvectors for each selection and data splitting method. The performance of the predictive model was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE), as presented in Table 2.

Table 2. Model Performance Evaluation

CV Method	EV Selection Method	Number of Selected EV	R^2	MAE	RMSE
10-*fold cross-validation*	Without Eigenvector	0	0.287	16.609	21.103
	Positive Eigenvalue	110	0.607	11.364	15.541
	Moran Index	81	0.582	11.934	16.044
	LASSO	114	0.594	11.689	15.808
	SHAP	**221**	**0.619**	**11.213**	**15.387**
Spatial blocked cross-validarion	Without Eigenvector	0	0.180	17.376	22.363
	Positive Eigenvalue	110	0.153	17.903	22.684
	Moran Index	81	0.110	18.456	23.262
	LASSO	114	0.096	18.439	23.515
	SHAP	**100**	**0.181**	**17.652**	**22.310**

In data separation based on 10-fold cross-validation, the model without the addition of eigenvectors as independent variables showed relatively low performance, with an R^2 value of 0.287 and MAE of approximately 16%. After spatial components were added through eigenvectors, the model performance improved significantly. The highest accuracy was achieved using the SHAP-based eigenvector selection method, with an R^2 value of 0.619 and a total of 221 selected eigenvectors. This was followed by selection based on positive eigenvalues ($R^2 = 0.607$), LASSO ($R^2 = 0.594$), and the Moran index ($R^2 = 0.582$). The MAE values for these four methods were relatively similar, at approximately 11%.

Based on spatial blocked cross-validation, all methods showed a decrease in model performance compared to 10-fold cross-validation. The R^2 values ranged from 0.096 to 0.181, with a MAE of approximately 17–18%. The decline in performance during spatial blocked cross-validation occurs because the model is required to predict the $PM_{2.5}$ concentration at locations that were completely excluded from the training phase. Unlike random 10-fold cross-validation, where the model can exploit spatial proximity by learning from nearby points, spatial validation requires the model to make predictions in entirely new locations without support from nearby training data. Consequently, the model cannot rely on the spatial similarity between neighboring points, making predictions more difficult and leading to lower performance. This implies that the predictive ability of the model is more limited when applied to entirely new spatial areas, highlighting the importance of spatial representativeness in training data for accurate environmental modeling.

The results of this study are consistent with those of previous studies on the application of ESF and machine learning for $PM_{2.5}$ concentration prediction. These findings support earlier studies showing that incorporating spatial components through eigenvectors improves model performance by effectively capturing spatial dependence (Zhang et al., 2018). In line with studies employing machine learning approaches, such as gradient boosting models for

spatiotemporal $PM_{2.5}$ prediction, this study also demonstrated the capability of GBM to capture nonlinear spatial patterns (Wang et al., 2023). Moreover, while previous studies have utilized SHAP primarily as a post hoc interpretability tool within machine learning frameworks, this study extends its application by employing SHAP for eigenvector selection, resulting in a more interpretable and robust identification of spatial features.

The finding that SHAP-based selection produces stable and consistent spatial components supports previous work highlighting the effectiveness of SHAP in interpreting spatial effects within machine learning models. This consistency reinforces the validity and reliability of SHAP as an interpretability tool, as demonstrated in earlier studies applying SHAP with models such as eXtreme Gradient Boosting (Li, 2022). Overall, this study not only supports existing findings but also extends them by integrating ESF with advanced machine learning and SHAP-based feature selection, thereby improving both predictive accuracy and spatial generalization.

D. CONCLUSION AND SUGGESTIONS

This study aimed to evaluate eigenvector selection methods within the ESF framework integrated with a GBM model for $PM_{2.5}$ concentration predictions. The results show that incorporating spatial components significantly improved the model performance, confirming the importance of accounting for spatial dependence. Among the evaluated methods, the SHAP-based selection achieved the highest predictive accuracy, demonstrating its effectiveness in capturing both spatial patterns and nonlinear relationships. These improvements can support more accurate air quality assessments and contribute to better environmental management strategies in urban areas.

Future research should focus on improving model performance in spatially separated areas by incorporating spatiotemporal modelling approaches, such as temporal lag variables and higher-resolution environmental data, to better capture dynamic patterns. In addition, further work is required to optimize the SHAP-based eigenvector selection by defining threshold criteria based on the distribution of SHAP values, allowing a more systematic identification of relevant spatial components. Combining SHAP with other selection methods, such as LASSO or Moran's Index, may further enhance selection efficiency and model interpretability.

ACKNOWLEDGEMENT

The authors would like to acknowledge the IPB Graduate School and Statistics & Data Science Study Program at IPB University for their guidance and support throughout this study. Special thanks are extended to the PPID Provinsi DKI Jakarta for providing the pollution data essential for this study. Additionally, we gratefully acknowledge the contributors of open-access datasets for providing data that supported this analysis.

REFERENCES

Ahmadi, M., Shafapourtehrany, M., Özener, H., Yilmaz, O. M., Kalantar, B., & Shabani, F. (2024). Eigenvector spatial filtering enhancing natural hazards vulnerability assessment in a susceptible urban environment: A case study of Izmir earthquake in Turkey. *Environmental Technology & Innovation*, *35*(May), 103666. https://doi.org/10.1016/j.eti.2024.103666

Crinnion, W. (2017). Particulate Matter Is a Surprisingly Common Contributor to Disease. *Integrative Medicine (Encinitas, Calif.)*, *16*(4), 8–12. http://www.ncbi.nlm.nih.gov/pubmed/30881250

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Griffith, D. A., & Chun, Y. (2019). Implementing Moran eigenvector spatial filtering for massively large georeferenced datasets. *International Journal of Geographical Information Science*, *33*(9), 1703–1717. https://doi.org/10.1080/13658816.2019.1593421

Islam, M. D., Li, B., Islam, K. S., Ahasan, R., Mia, Md. R., & Haque, M. E. (2022). Airbnb rental price modeling based on Latent Dirichlet Allocation and MESF-XGBoost composite model. *Machine Learning with Applications*, *7*, 100208. https://doi.org/10.1016/j.mlwa.2021.100208

Kusumaningtyas, S. D. A., Khoir, A. N., Fibriantika, E., & Heriyanto, E. (2021). Effect of meteorological parameter to variability of Particulate Matter (PM) concentration in urban Jakarta city, Indonesia. *IOP Conference Series: Earth and Environmental Science*, *724*(1). https://doi.org/10.1088/1755-1315/724/1/012050

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, *96*. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Liu, X., Kounadi, O., & Zurita-Milla, R. (2022). Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features. *ISPRS International Journal of Geo-Information*, *11*(4), 242. https://doi.org/10.3390/ijgi11040242

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mahkya, D. Al, Djuraidah, A., Wigena, A. H., & Sartono, B. (2024). Rainfall modeling with CMIP6-DCPP outputs and local characteristic information using eigenvector spatial filtering varying coefficient (ESF-VC). *Journal of Agrometeorology*, *26*(3), 311–317. https://doi.org/10.54386/jam.v26i3.2599

Marcilio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347. https://doi.org/10.1109/SIBGRAPI51738.2020.00053

McCord, M. J., McCord, J., Davis, P. T., Haran, M., & Bidanset, P. (2020). House price estimation using an eigenvector spatial filtering approach. *International Journal of Housing Markets and Analysis*, *13*(5), 845–867. https://doi.org/10.1108/IJHMA-09-2019-0097

Murakami, D., & Griffith, D. A. (2019). Eigenvector Spatial Filtering for Large Data Sets: Fixed and Random Effects Approaches. *Geographical Analysis*, *51*(1), 23–49. https://doi.org/10.1111/gean.12156

Murakami, D., Yoshida, T., Seya, H., Griffith, D. A., & Yamagata, Y. (2017). A Moran coefficient-based mixed effects approach to investigate spatially varying relationships. *Spatial Statistics*, *19*, 68–89. https://doi.org/10.1016/j.spasta.2016.12.001

Seya, H., Murakami, D., Tsutsumi, M., & Yamagata, Y. (2015). Application of LASSO to the Eigenvector Selection Problem in Eigenvector-based Spatial Filtering. *Geographical Analysis*, *47*(3), 284–299. https://doi.org/10.1111/gean.12054

Singh, U., Rizwan, M., Alaraj, M., & Alsaidan, I. (2021). A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments. *Energies*, *14*(16), 5196. https://doi.org/10.3390/en14165196

Sotoudeheian, S., & Arhami, M. (2021). Estimating ground-level PM2.5 concentrations by developing and optimizing machine learning and statistical models using 3 km MODIS AODs: case study of Tehran, Iran. *Journal of Environmental Health Science and Engineering*, *19*(1), 1–21. https://doi.org/10.1007/s40201-020-00509-5

Sun, W., Murakami, D., Hu, X., Li, Z., & Kidd, A. N. (2023). Supply – Demand Imbalance in School Land : An Eigenvector Spatial Filtering Approach. *Sustainability*, *15*(17), 12935. https://doi.org/10.3390/su151712935

Wang, Z., Wu, X., & Wu, Y. (2023). A spatiotemporal XGBoost model for PM2.5 concentration prediction and its application in Shanghai. *Heliyon*, *9*(12), e22569. https://doi.org/10.1016/j.heliyon.2023.e22569

Xu, J., Liu, Z., Yin, L., Liu, Y., Tian, J., Gu, Y., Zheng, W., Yang, B., & Liu, S. (2021). Grey Correlation Analysis of Haze Impact Factor PM2.5. *Atmosphere*, *12*(11), 1513. https://doi.org/10.3390/atmos12111513

Zhang, J., Li, B., Chen, Y., Chen, M., Fang, T., & Liu, Y. (2018). Eigenvector Spatial Filtering Regression Modeling of Ground PM2.5 Concentrations Using Remotely Sensed Data. *International Journal of Environmental Research and Public Health*, *15*(6), 1228. https://doi.org/10.3390/ijerph15061228