Data and Metadata. 2025; 4:862
doi: 10.56294/dm2025862

ORIGINAL

Comparison of Time Series Regression, Support Vector Regression, Hybrid, and Ensemble Method to Forecast PM2.5

Comparación de Los Métodos de Regresión de Series Temporales, Regresión de Vector de Soporte, Híbrido y Conjunto Para Predecir las PM2.5

Elly Pusporani[1], Ghisella Asy Sifa[1], Nurin Faizun[1], Pressylia Aluisina Putri Widyangga[1], Adma Novita Sari[1], M. Fariz Fadillah Mardianto[1], Sediono[1]

[1]Departement of Mathematics, Faculty of Science and Technology, Universitas Airlangga. Surabaya, Indonesia.

Cite as: Pusporani E, Asy Sifa G, Faizun N, Putri Widyangga PA, Novita Sari A, Fadillah Mardianto MF, et al. Comparison of Time Series Regression, Support Vector Regression, Hybrid, and Ensemble Method to Forecast PM2.5. Data and Metadata. 2025; 4:862. https://doi.org/10.56294/dm2025862

Submitted: 08-06-2024 **Revised:** 05-10-2024 **Accepted:** 25-03-2025 **Published:** 26-03-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Elly Pusporani

ABSTRACT

Introduction: $PM_{2.5}$ pollution poses significant health risks, particularly in Jakarta, where levels often exceed safety thresholds. Accurate forecasting models are essential for air quality management and mitigation strategies.
Method: this study compares four forecasting models: Time Series Regression (TSR), Support Vector Regression (SVR), a hybrid TSR-SVR model, and an ensemble approach. The dataset consists of 9119 hourly $PM_{2.5}$ observations from January 1, 2023, to January 15, 2024. Missing values were imputed using historical hourly trends. Model performance was evaluated using Root Mean Squared Error (RMSE).
Results: the hybrid TSR-SVR model achieved the lowest RMSE (6,829), outperforming TSR (7,595), SVR (7,477), and the ensemble method (7,486). The hybrid approach effectively captures both linear and nonlinear patterns in $PM_{2.5}$ fluctuations, making it the most accurate model.
Conclusions: integrating statistical and machine learning models improves $PM_{2.5}$ forecasting accuracy, aiding policymakers in pollution control efforts. Future studies should explore additional external factors to enhance prediction performance.

Keywords: $PM_{2.5}$ Forecasting; Time Series Regression; Support Vector Regression; Hybrid Model; Ensemble Learning.

RESUMEN

Introducción: la contaminación por $PM_{2.5}$ representa un grave riesgo para la salud, especialmente en Yakarta, donde los niveles suelen superar los umbrales de seguridad. Modelos de pronóstico precisos son esenciales para la gestión de la calidad del aire.
Método: este estudio compara cuatro modelos de pronóstico: Regresión de Series Temporales (TSR), Regresión de Vectores de Soporte (SVR), un modelo híbrido TSR-SVR y un enfoque en conjunto. Se utilizaron 9119 observaciones horarias de $PM_{2.5}$ entre el 1 de enero de 2023 y el 15 de enero de 2024. Se imputaron valores faltantes mediante tendencias horarias históricas. La evaluación del modelo se realizó con Error Cuadrático Medio (RMSE).
Resultados: el modelo híbrido TSR-SVR obtuvo el RMSE más bajo (6,829), superando a TSR (7,595), SVR (7,477) y el método en conjunto (7,486). Capturó eficazmente patrones lineales y no lineales, logrando la mayor precisión.
Conclusiones: la integración de modelos estadísticos y de aprendizaje automático mejora la predicción del

$PM_{2.5}$, apoyando estrategias de control de la contaminación. Investigaciones futuras deberían considerar factores externos adicionales para optimizar la precisión.

Palabras clave: Pronóstico de $PM_{2.5}$; Regresión de Series Temporales; Regresión de Vectores de Soporte; Modelo Híbrido; Aprendizaje en Conjunto.

INTRODUCTION

Air pollution is a global issue affecting almost every part of the world. 2024th Indonesia's Minister of Environment and Forestry, Siti Nurbaya, stated that air pollution is caused by prolonged droughts and the concentration of pollutants.[1] According to the Regulation of the Minister of Environment and Forestry of the Republic of Indonesia Number P.14/MENLHK/SETJEN/KUM.1/7/2020 on the Air Pollution Standard Index (ISPU), the main pollutants include Particulate Matter (PM_{10}), Particulate Matter ($PM_{2.5}$), Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Ozone (O3), and Hydrocarbons (HC).[2] Among these pollutants, $PM_{2.5}$, and HC were newly added compared to the previous regulations. In addition to the inclusion of new pollutants, the frequency of ISPU information dissemination to the public has also increased. Specifically, ISPU calculations for $PM_{2.5}$ are now reported hourly for 24 h, as $PM_{2.5}$ is considered the most significant air pollutant affecting human health.[3] There are variations in $PM_{2.5}$ air quality standards worldwide. In Indonesia, the acceptable standard for $PM_{2.5}$ concentration is below 15µg/m^3. However, on November 8, 2023, Jakarta's air quality was categorized as unhealthy, with an Air Quality Index (AQI) of 153 and a $PM_{2.5}$ concentration of 58,7µg/m^3, worsening over the following two days. These measurements indicated unhealthy air conditions that exceeded the permissible thresholds.

According to Xing et al.[4], $PM_{2.5}$ has significant health effects and increases mortality rates due to respiratory, cardiovascular, and pulmonary diseases. Continuous exposure to $PM_{2.5}$ poses serious health risks, with every 10µg/m^3 increase in $PM_{2.5}$, raising the overall mortality risk by 6 % and cardiovascular-related mortality by 11 %.[5] These findings are further supported by a study conducted by Hayes et al.[6], which states that long-term exposure to $PM_{2.5}$ air pollution is associated with higher mortality rates from ischemic heart disease and stroke. The increasing severity of air pollution, particularly that of $PM_{2.5}$, highlights the urgent need for mitigation strategies and public awareness to reduce health risks. Addressing air pollution requires coordinated efforts, including stricter regulations, sustainable urban planning, and technological innovations, to improve air quality and public health outcomes.

Previous Study

Given the serious health risks posed by $PM_{2.5}$, predicting $PM_{2.5}$ pollution levels have become increasingly important, particularly in Jakarta, where pollution levels often exceed the minimum threshold. Most air pollution forecasting studies have traditionally relied on daily or monthly data, as seen in the research by Hasnain, et al.[7]. However, several international studies have utilized hourly observation data, such as those conducted by Drewil et al.[8] in India, Cordova et al.[9] in Lima-Peru, and Kothandaraman et al.[10] in New Delhi. In contrast, studies in Indonesia, particularly in Jakarta, have predominantly used daily data, as shown in the research by Handhayani[11]. Daily forecasting methods provide only a general overview of air pollution conditions for a given day, whereas pollution levels fluctuate continuously throughout the day. This highlights the need for $PM_{2.5}$ forecasting based on hourly data to capture real-time variations and provide more accurate air quality assessments. Hourly forecasting can help policymakers and the public take timely preventive measures to mitigate the adverse effects of air pollution. The increasing severity of air pollution, particularly $PM_{2.5}$ levels, highlights the urgent need for mitigation strategies and public awareness to reduce health risks. Addressing air pollution requires coordinated efforts, including stricter regulations, sustainable urban planning, and technological innovations to improve air quality and public health outcomes.

Overview and Research Objective

This study utilized classical methods, machine learning, ensemble, and hybrid approaches for forecasting $PM_{2.5}$. These methods were selected based on research by Makridakis et al.[12] in the M4-Competition, which concluded that combining multiple methods generally yields a higher accuracy than using a single method. This finding is consistent with previous forecasting competitions and empirical studies by Smith et al.[13]. However, combining these methods does not always guarantee better forecasts than a single model. Therefore, classical and machine-learning methods are used to represent single models, whereas ensemble and hybrid methods are used to represent combination models. The primary objective of this study is to provide a comprehensive understanding of $PM_{2.5}$ pollution in Jakarta and identify the best forecasting model for hourly $PM_{2.5}$. Additionally, this study aims to compare forecasting methods using classical models, machine learning, ensembles, and hybrid techniques. This comparison will help determine the most accurate model for predicting $PM_{2.5}$ levels in Jakarta. This study is crucial as it will provide valuable insights into the factors influencing $PM_{2.5}$ levels in Jakarta. These insights can be used to develop effective policies and strategies for reducing air pollution. The findings of this study can also contribute to improving the early warning system for air pollution in Jakarta, enabling the public to take precautionary measures against harmful $PM_{2.5}$ exposure. Ultimately, this study is expected to make a significant contribution to addressing air pollution problems in Jakarta.

METHOD

Data and Research Variables

This study uses hourly secondary data obtained from the website https://www.airnow.gov/ during the period January 1, 2023 at 01:00 WIB until January 15, 2024 at 23:00 WIB, with a total of 9119 observations. Data from January 1, 2023, to January 8, 2024, were used as training data, while data from January 9, 2024, to January 15, 2024, were used as testing data. The variables used in this study are $PM_{2.5}$ in Central Jakarta with measurement unit ug/m^3.

Unfortunately, many values were missing from the data. This is because of various issues with the data series, such as missing, suspect, or invalid data provided by the data source. In total, 689 problematic observations were observed. Suspect and invalid data indicated that the data at that time could not be used for analysis. Therefore, suspect and invalid data were removed so that they became missing values. Unfortunately, time series forecast method need the data that don't have any missing value. The most common method to do it is filled the missing data with mean of the data. However, there are obstacles in applying this method, as the $PM_{2.5}$ data shows a pattern that resembles seasonal data. In addition, some problems occurred in certain weeks with more than hundred missing data. This causes the imputation results to appear as a straight line and does not reflect other data fluctuations. To overcome these problems, this study proposes replacing missing values with the average of the same hours and days in the periods before and after the value. As an illustration, if the data is missing on Monday, 11-12-2023, at 09:00 PM, then the data will be filled with the average of the same hour and day in the period before and after using this equation:

$$\frac{\text{data[monday, 04−12−23 09.00PM]+data[monday, 18−12−23 09.00PM]}}{2} \qquad (1)$$

Step Analysis

This study aims to predict seasonal time series data using the Time Series Regression (TSR), Support Vector Regression (SVR), hybrid TSR-SVR, and Ensemble method. The research process involved the following steps:

1. Data Preprocessing: the raw data obtained from this source contained missing values and invalid records. Data preprocessing is carried out by applying an imputation technique in which missing values are replaced with the average of the same hours and days in the period before and after.[14] This ensured that the data structure remained consistent during the analysis.
2. Identifying Significant Lags Using Partial Autocorrelation Function (PACF): PACF helps determine the order of the lagged variables to be included in the TSR model.[15]
3. Developing the Time Series Regression (TSR) Model: using the identified significant lag(s), the TSR model was constructed to establish a linear relationship between the current $PM_{2.5}$ level and its past values.[16] The regression model is expressed as:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_k Y_{t-k} + \epsilon_t \qquad (2)$$

Where Y_t is the $PM_{2.5}$ concentration at time t, Y_{t-1} is the lag-1, Y_{t-k} is the lag-k value, β_0, β_1,..., β_k are regression coefficients, and ϵ_t is the error term. The model was validated using the t-test, and its Root Mean Squared Error (RMSE) was computed as:[17]

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{n}} \qquad (3)$$

Where Y_t was the actual value, $\hat{Y}_t$ was the predicted value, and n was the number of observations.

4. Support Vector Regression (SVR) Analysis: the SVR model was implemented with different kernel functions (Radial Basis Function (RBF) and sigmoid). SVR solves the following optimization problem:[18]

$$\min_{\omega, \xi} \frac{1}{2}|\omega|^2 + C\sum_{i=1}^{n}\xi_t \qquad (4)$$

Subject to $|Y_t-(\omega\cdot X_t+b)|\leq\in+\xi_t$. Parameter tuning is performed to optimize C,γ, and ξ values.

5. Hybrid TSR-SVR Model: a hybrid approach was applied, in which the residual errors from the TSR model were further modeled using SVR. The residuals are given by:[19]

$$e_t = Y_t - \hat{Y_t}^{TSR} \qquad (5)$$

The value of e_t will be modelled by SVR model so it will get the value of $\hat{e}_t^{SVR}$. The result of SVR model is then added to the forecast of TSR model and the final prediction is:[20]

$$\hat{Y}_t = \hat{Y_t}^{TSR} + \hat{e_t}^{SVR} \qquad (6)$$

6. Ensemble TSR-SVR Model: the ensemble method combines the predictions of TSR and SVR using weighted averaging:[21]

$$\hat{Y}_t = \omega_{1t}\hat{Y_t}^{TSR} + \omega_{2t}\hat{Y_t}^{SVR} \qquad (7)$$

Where ω_{1t} and $\omega_{t2}=1-\omega_{1t}$ are the assigned weights based on the squared residual of TSR and SVR models. The calculation of the weight based on variance covariance method is as follow:

$$\omega_{1t} = \omega_{2t} = 0.5$$

Fixed Window Variance-Covariance (FWVC):

$$\omega_{1t} = \frac{\sum_{T=t-24}^{\tau-1}\epsilon_{2T}^2}{\sum_{T=t-24}^{\tau-1}\epsilon_{1T}^2+\sum_{T=t-24}^{\tau-1}\epsilon_{2T}^2} \qquad (8)$$

Where τ was the last time of training data.
Seasonal Variance-Covariance (SVC):

$$\omega_{1t} = \frac{\epsilon_{2,t-24}^2}{\epsilon_{1,t-24}^2+\epsilon_{2,t-24}^2} \qquad (9)$$

Shifting Window Variance-Covariance (SWVC) methods.

$$\omega_{1t} = \frac{\sum_{T=t-24}^{t-1}\epsilon_{2T}^2}{\sum_{T=t-24}^{t-1}\epsilon_{1T}^2+\sum_{T=t-24}^{t-1}\epsilon_{2T}^2} \qquad (10)$$

Where ϵ_{jT}^2 was the error of forecast for jth individual model at t.

7. Model Evaluation and Comparison: the performance of all models (TSR, SVR, Hybrid TSR-SVR, and Ensemble) is evaluated using RMSE as written in equation (3). A comparative analysis was conducted to determine the best-performing model for $PM_{2.5}$.

RESULTS

Time Series Regression (TSR) Analysis

Based on the method analysis, the data that will be used as the based is the data that has been imputed with the calculated value so now the data didn't have any missing value in the series. The first method that will be used was Time Series Regression. Time-series regression analysis begins by identifying a significant lag. One method used to identify significant lags is Partial Autocorrelation (PACF). The PACF results are shown in figure 1.

In this case, the PACF results have not been able to identify all significant lags; therefore, the concept of parsimony is used. The first step is to model only using lag 1 of the data. In the analysis using Time Series Regression, it is necessary to test whether there is an influence given by lag 1, which is then represented by the notation y_{t-1} to y_t. Significance analysis was carried out using the t-test, and the results are presented in table 1.

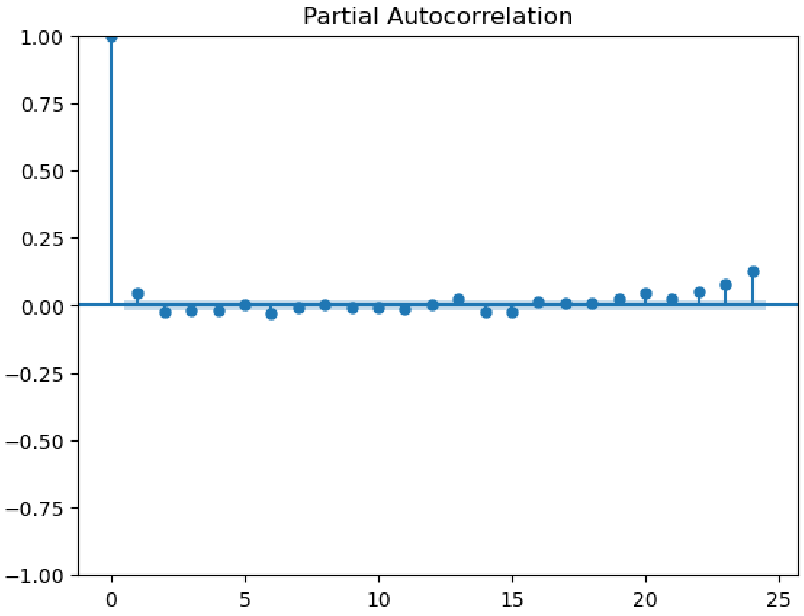

Figure 1. PACF Data

Table 1. t-Test Results

	Coefficient	t	P-Value	Decision
β_0	4,6331	21,633	0,000	Significant
β_1	0,8758	171,745	0,000	Significant

From table 1, it can be seen that the data at the previous time (lag 1) affect the current data. From the table, the regression model equation was obtained as follows:

$$\hat{y}_t = 4{,}6331 + 0{,}8758\ y_{t-1} \qquad (11)$$

From the regression model, the R² value is 76,7 %, which means that 76,7 % of the characteristics of y_t can be explained by y_{t-1} the rest is explained by other variables or lags. To prove that its better to only uses lag 1 of the data for the subsequent analysis this study actually try to add more lag to the predictor valiable but the result show less than 0,1 % increased in the R² value.

This meant that the addel lag doesn't have any significant contribution to make the model had better result. This is the basis why for the subsequent analysis, only lag 1 is used.

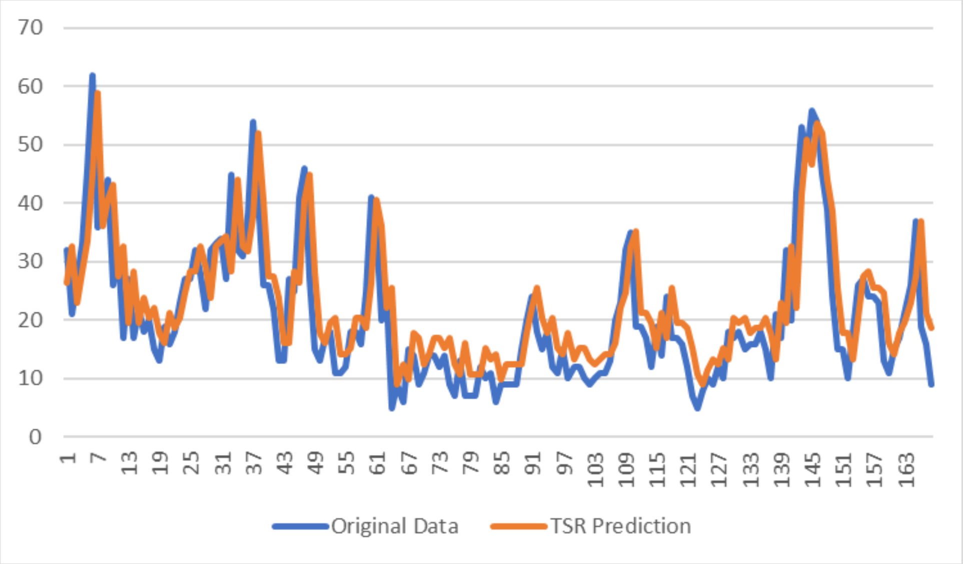

Figure 2. Time Series Plot between Original Testing Data and Prediction Results Using TSR Method

https://doi.org/10.56294/dm2025862

The model obtained is then used to obtain prediction results on testing data containing 168 data. Based on the prediction results, the RMSE value of 7,595 is much better than the standard deviation of the testing data, which is 11,660, which shows that the prediction with this method is actually good. A comparison of the prediction results and the original testing data on the time-series regression method is presented in figure 2.

Support Vector Regression (SVR) Analysis

In this method, the lag used to predict the $PM_{2.5}$ data is the same as that used in the Time Series Regression method. However, the analysis process is quite different from that of the TSR method. In the SVR method, data are processed using kernel functions and parameter tuning. In this study, two kernel functions were used: the Radial Basis Function (RBF) and sigmoid with tuned parameters, as presented in table 2.

Table 2. Parameter Tuning in the SVR Method

Parameter	Value
Gamma	[‘scale’, ‘auto’]
C	[0,03125, 0,0625, 0,125, 0,25, 0,5, 1, 2, 4, 8, 16, 32]
Epsilon	[0,01, 0,02, 0,03, 0,04, 0,05, 0,06, 0,07, 0,08, 0,09, 0,10, , 1,00]

Because the computation is very large, the running for each kernel is separated, the results are compared, and the analysis is performed by sampling as many parameters as 1000 samples for each kernel. The best results for the RBF kernel were gamma=scale, epsilon=0,23, and C=32, whereas the best results for the sigmoid kernel were gamma=auto, epsilon=0,01, and C=0,0625. Subsequently, these parameters were run and the best parameters were kernel=rbf, gamma=scale, epsilon=0,23, and C=32. The parameters obtained were then used to predict the testing data, and an RMSE value of 7,477 was obtained. Comparison of prediction results and original testing data on the Support Vector Regression method can be seen in figure 3.

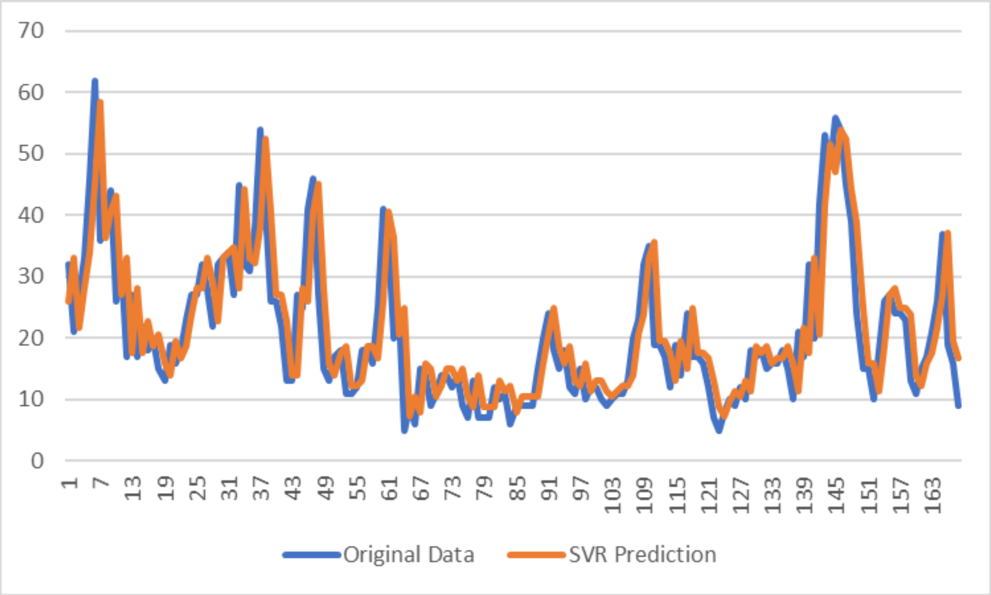

Figure 3. Time Series Plot between Original Testing Data and Prediction Results Using SVR Method

Hybrid Time Series Regression (TSR) - Support Vector Regression (SVR) Analysis

The process of analyzing the hybrid method involves taking the residual data of the first method and then modeling it using the second method, after which the residual prediction results are added to the prediction results of the first method. In this study, the first method was the TSR method with the training data residual data presented in figure 4.

As shown in figure 4, the residual data were around point 0. This is in line with the concept of modeling, which assumes that the error that occurs is close to 0. These data were then used in the SVR model analysis. In this analysis, the running concept carried out on the residual data from the TSR analysis is the same as the analysis concept on the original data, namely when running, each kernel is separated and then the results are compared. In addition, the analysis is carried out by sampling a combination of parameters of 1000 samples for each kernel. The parameters to be sampled are listed in table 3.

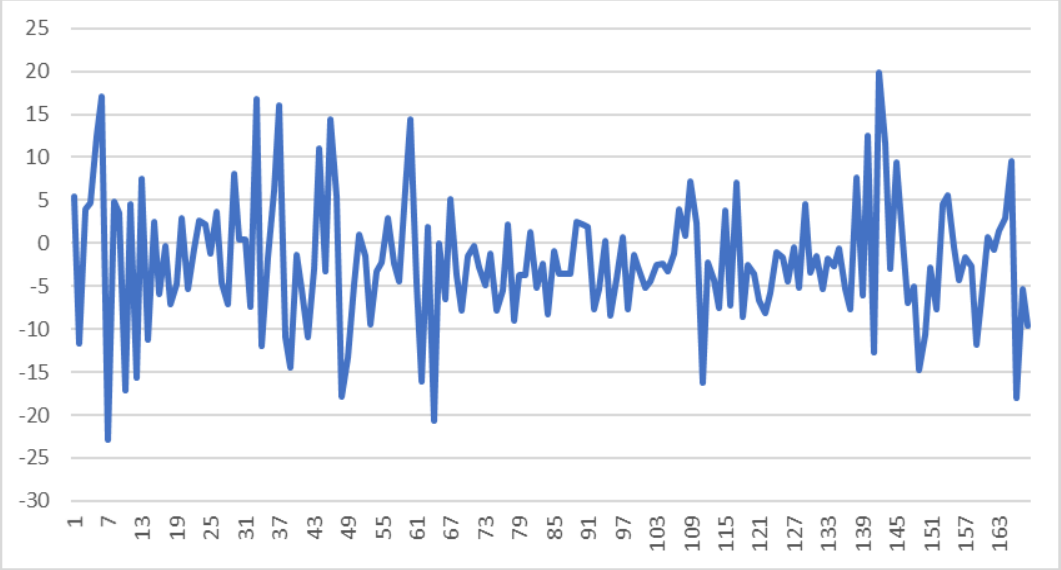

Figure 4. Time Series Plot of Residual Training Data Prediction results with TSR

Table 3. Parameters sampled in the TSR Method

Parameter	Value
Gamma	[‘scale’, ‘auto’]
C	[0,03125, 0,0625, 0,125, 0,25, 0,5, 1, 2, 4, 8, 16, 32]
Epsilon	[0,01, 0,02, 0,03, 0,04, 0,05, 0,06, 0,07, 0,08, 0,09, 0,10, ... , 1,00]

The results of the analysis showed that the best parameters for the RBF kernel were gamma=scale, epsilon=0,02, and C= 0,0625. When running the program for the sigmoid kernel function, it turns out that the best result for the sigmoid kernel is the same as that for the rbf kernel, namely gamma=scale, epsilon=0,02, and C= 0,0625. This causes the next step to simply compare the best kernel for the same parameters. The results of the analysis show that, in this case, the sigmoid kernel is better than the RBF kernel. The next step was to predict the test data using the best parameters. Based on the prediction results, the RMSE value was 6,892. A comparison of the prediction results and the original testing data in the time-series regression method is presented in figure 5.

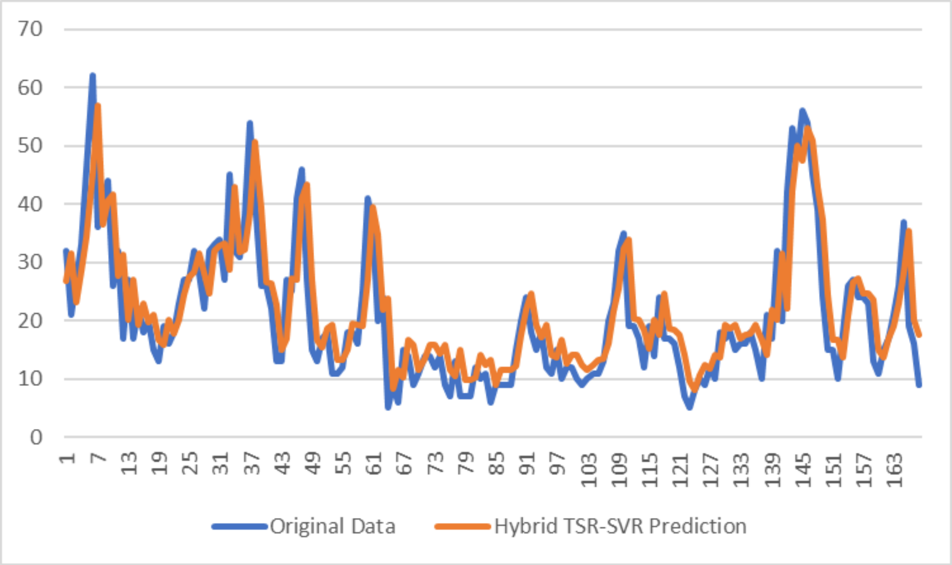

Figure 5. Time Series Plot between Original Testing Data and Prediction Results Using the TSR-SVR hybrid Method

Ensemble Time Series Regression (TSR) - Support Vector Regression (SVR) Analysis

Unlike the analysis using the hybrid method, the ensemble method combines the prediction results of two methods by weighting the prediction results used. In this study, four weights were used: average, shifting window variance covariance (SWVC), seasonal variance covariance (SVC), and fixed window variance covariance

https://doi.org/10.56294/dm2025862

(FWVC). In the averaging weighting, both prediction results, namely the TSR and SVR model predictions, are given the same weight of 0,5. The FWVC weight is different for the TSR and SVR prediction results, but does not change for each observation. In this case, the weights for the TSR prediction and SVR were 0,49742924352667917 and 0,5025707564733208, respectively. It can be seen that the prediction with SVR has a greater weight than that with TSR, which is in accordance with the RMSE results of the SVR model, which are smaller than those of the TSR model. As for weighting with SVC and SWFC, the obtained weight changes according to the data to be predicted. The weights of the two methods are listed in table 4.

Table 4. Weight Value of SVC and SWVC

Date and Time	SVC Weight		SWVC Weight	
	TSR	SVR	TSR	SVR
09/01/2024 00:00	0,429762	0,570238	0,497429	0,502571
09/01/2024 01:00	0,566226	0,433774	0,498073	0,501927
09/01/2024 02:00	0,331535	0,668465	0,496988	0,503012
09/01/2024 03:00	0,612595	0,387405	0,497271	0,502729
09/01/2024 04:00	0,54938	0,45062	0,497262	0,502738
09/01/2024 05:00	0,504993	0,495007	0,497005	0,502995
09/01/2024 23:00	0,362882	0,637118	0,497676	0,502324
10/01/2024 00:00	0,59411	0,40589	0,498273	0,501727
15/01/2024 23:00	0,508566	0,491434	0,493020	0,506980

The prediction results for these weights are shown in figure 6.

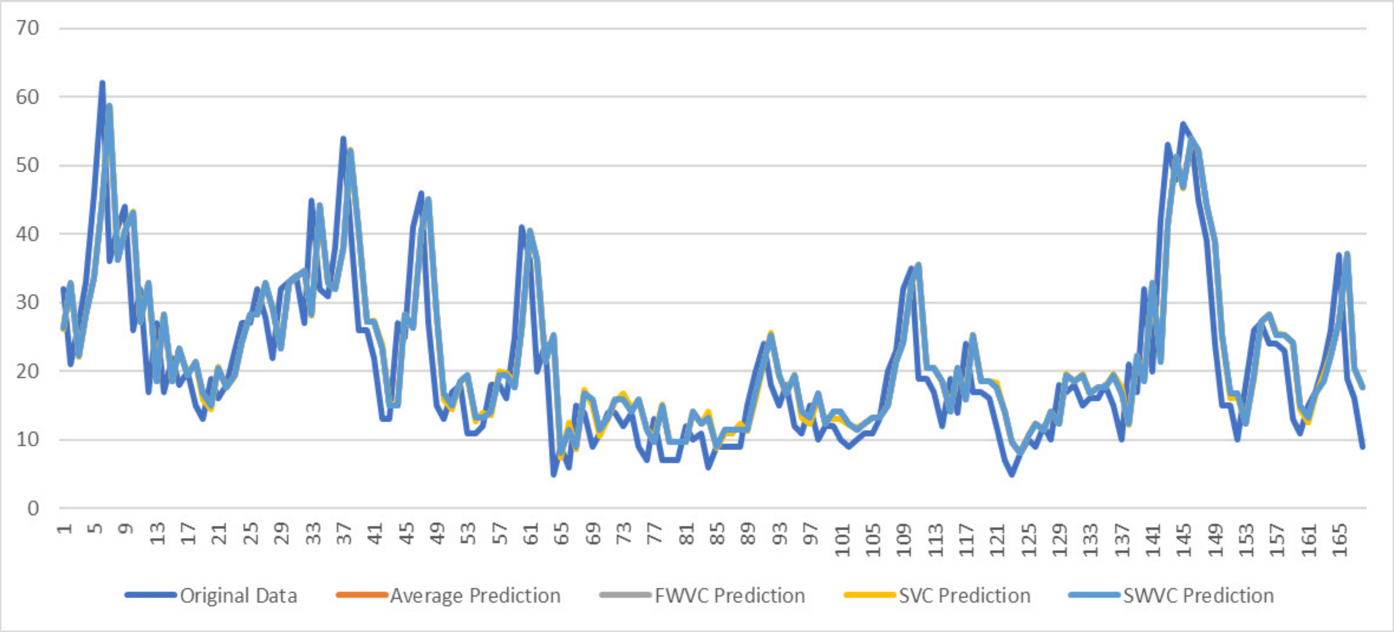

Figure 6. Time Series Plot between Original Testing Data and Prediction Results Using TSR-SVR Ensemble Method

Based on figure 6, it can be seen that the prediction results of the four methods are not much different, resulting in overlapping prediction results. Because of this, the best ensemble method can be determined by using the RMSE value presented in table 5.

Table 5. RMSE Value Based on Ensemble Method

Ensemble Method	RMSE
Average	7,497
FWVC	7,497
SVC	7,486
SWVC	7,497

Based on the RMSE results, the Seasonal Variance-Covariance method has the best results for predicting $PM_{2.5}$, among the four ensemble methods that have been tried with an RMSE of 7,486.

Model Comparison

To determine the best model to be used for forecasting, a comparison of prediction results on testing data on the four TSR, SVR, hybrid, and ensemble methods was performed using the RMSE value, the results of which are presented in table 6.

Table 6. RMSE Value Based on Each Method

Methods	RMSE
TSR	7,595
SVR	7,477
hybrid TSR-SVR	6,829
Ensemble (SVC)	7,486

The table above shows that the hybrid model that combines the TSR and SVR methods is the best model for predicting $PM_{2.5}$, compared to other methods. It should be noted that the ensemble method that combines the TSR and SVR prediction results by weighting the prediction results also produces an RMSE value that is smaller than methods that only use TSR or SVR methods. However, the improvement in the prediction results with the ensemble method is not better than that with the hybrid method. In addition, the prediction results with the hybrid method are suitable for use because the prediction data pattern follows the original data pattern, which can be seen in figure 4.

CONCLUSIONS

In conclusion, this study demonstrated that a hybrid model combining Time Series Regression (TSR) and Support Vector Regression (SVR) is the most effective method for predicting $PM_{2.5}$ concentrations in Jakarta. The analysis showed that the TSR-SVR hybrid model outperformed the individual TSR and SVR models as well as ensemble methods that integrate both predictions through various weighting techniques. Specifically, the hybrid approach resulted in a lower Root Mean Squared Error (RMSE), indicating its superior predictive accuracy for capturing the temporal patterns and seasonal fluctuations inherent in $PM_{2.5}$. The findings highlight the advantages of leveraging both statistical and machine learning techniques to improve air pollution forecasting. Although TSR effectively captures linear relationships and time-dependent structures, SVR enhances the model's ability to account for nonlinear patterns and complex interactions within the data. These results suggest that incorporating robust imputation strategies can further enhance the accuracy of air pollution predictions. These insights have valuable implications for policymakers, environmental agencies, and researchers working on air quality management. Future research could explore additional machine learning techniques, refine the hybrid modeling framework, and incorporate external factors, such as meteorological conditions and traffic density, to further improve air quality predictions.

BIBLIOGRAPHIC REFERENCES

1. J. E. Goldstein, "The Volumetric Political Forest: Territory, Satellite Fire Mapping, and Indonesia's Burning Peatland," Antipode, vol. 52, no. 4, pp. 1060–1082, Jul. 2020, doi: 10.1111/anti.12576.

2. Y. Sri Susilo et al., "Valuation of the Economic Impact of Air Pollution to Promote Public Welfare in Jakarta," Journal of Business and Information Systems, vol. 6, no. 2, 2024, doi: 10.36067/jbis.v6i2.260.

3. K. Aji Tritamtama, F. E. S. Sembiring, A. Choiruddin, and H. Patria, "Analysis of Air Pollution (SO2) at Some Point of Congestion in DKI Jakarta," Disease Prevention and Public Health Journal, vol. 17, no. 1, pp. 82–92, Feb. 2023, doi: 10.12928/dpphj.v17i1.6147.

4. Y. F. Xing, Y. H. Xu, M. H. Shi, and Y. X. Lian, "The Impact of PM2.5 on The Human Respiratory System," J Thorac Dis, vol. 8, no. 1, pp. E69–E74, 2016, doi: 10.3978/j.issn.2072-1439.2016.01.19.

5. Kunovac, "Maternal Engineered Nanomaterial Inhalation Exposure: Cardiac Maternal Engineered Nanomaterial Inhalation Exposure: Cardiac Molecular Reprogramming in Progeny through Epigenetic and Molecular Reprogramming in Progeny through Epigenetic and Epitranscriptomic Mechanisms Epitranscriptomic Mechanisms," West Virginia University, Morgantown, 2021. Online. Available: https://researchrepository.wvu.edu/etd/10255

6. 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.

7. Hasnain et al., "Time Series Analysis and Forecasting of Air Pollutants Based on Prophet Forecasting Model in Jiangsu Province, China," Front Environ Sci, vol. 10, Jul. 2022, doi: 10.3389/fenvs.2022.945628.

8. G. I. Drewil and R. J. Al-Bahadili, "Air Pollution Prediction Using LSTM Deep Learning and Metaheuristics Algorithms," Measurement: Sensors, vol. 24, Dec. 2022, doi: 10.1016/j.measen.2022.100546.

9. C. H. Cordova, M. N. L. Portocarrero, R. Salas, R. Torres, P. C. Rodrigues, and J. L. López-Gonzales, "Air Quality Assessment and Pollution Forecasting Using Artificial Neural Networks in Metropolitan Lima-Peru," Sci Rep, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-03650-9.

10. D. Kothandaraman et al., "Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning," 2022, Hindawi Limited. doi: 10.1155/2022/5086622.

11. T. Handhayani, "An Integrated Analysis of Air Pollution and Meteorological Conditions in Jakarta," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-32817-9.

12. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods," Int J Forecast, vol. 36, no. 1, pp. 54–74, Jan. 2020, doi: 10.1016/j.ijforecast.2019.04.014.

13. J. Smith and K. F. Wallis, "A Simple Explanation of The Forecast Combination Puzzle," Oxf Bull Econ Stat, vol. 71, no. 3, pp. 331–355, Jun. 2009, doi: 10.1111/j.1468-0084.2008.00541.x.

14. B. Cho et al., "Effective Missing Value Imputation Methods for Building Monitoring Data," in IEEE International Conference on Big Data, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 2866–2875. doi: 10.1109/BigData50022.2020.9378230.

15. W. Wibowo, B. S. S. Ulama, T. H. Siagian, T. Purwa, and R. N. Wilantari, "Impact of earthquakes on the number of airline passenger arrivals and departures: A case Study of West Nusa Tenggara Province, Indonesia," Regional Statistics, vol. 11, no. 3, pp. 133–157, 2021, doi: 10.15196/RS110302.

16. L. Wang et al., "Satellite-Based Assessment of The Long-Term Efficacy of PM2.5 Pollution Control Policies Across The Taiwan Strait," Remote Sens Environ, vol. 251, p. 1, Dec. 2020, doi: 10.1016/j.rse.2020.112067.

17. D. Alita, A. D. Putra, and D. Darwis, "Analysis of Classic assumption test and multiple linear regression Coefficient Test for Employee Structural Office Recommendation," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 3, pp. 295–306, Jul. 2021, doi: 10.22146/ijccs.65586.

18. L. L. Li, Z. Y. Cen, M. L. Tseng, Q. Shen, and M. H. Ali, "Improving Short-Term Wind Power Prediction Using Hybrid Improved Cuckoo Search Arithmetic - Support Vector Regression Machine," J Clean Prod, vol. 279, pp. 1–15, Jan. 2021, doi: 10.1016/j.jclepro.2020.123739.

19. C. D. Aju, A. L. Achu, M. P. Mohammed, M. C. Raicy, G. Gopinath, and R. Reghunath, "Groundwater Quality Prediction and Risk Assessment in Kerala, India: A Machine-Learning Approach," J Environ Manage, vol. 370, p. 1, Nov. 2024, doi: 10.1016/j.jenvman.2024.122616.

20. W. Y. Duan, Y. Han, L. M. Huang, B. B. Zhao, and M. H. Wang, "A Hybrid EMD-SVR Model for The Short-Term Prediction of Significant Wave Height," Ocean Engineering, vol. 124, pp. 54–73, Sep. 2016, doi: 10.1016/j.oceaneng.2016.05.049.

21. C. Go, Y. J. Kwak, S. Kwag, S. Eem, S. Lee, and B. S. Ju, "On Developing Accurate Prediction Models for Residual Tensile Strength of GFRP Bars Under Alkaline-Concrete Environment Using a Combined Ensemble Machine Learning Methods," Case Studies in Construction Materials, vol. 18, p. 16, Jul. 2023, doi: 10.1016/j.cscm.2023.e02157.

FINANCING

This publication is funded by Universitas Airlangga (Unair) through Airlangga Research Fund (ARF) with the number of research contract 1586/UN3.FST/PT.01.03/2024.

CONFLICT OF INTEREST

There is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Elly Pusporani.
Data curation: Elly Pusporani, M. Fariz Fadillah Mardianto.
Formal analysis: Nurin Faizun, Ghisella Asy Sifa.
Research: Elly Pusporani, Pressylia Aluisina Putri Widyangga.
Methodology: M. Fariz Fadillah Mardianto.
Project management: Sediono.
Resources: Adma Novita Sari.
Software: Elly Pusporani.
Supervision: Elly Pusporani.
Validation: Sediono.
Display: Ghisella Asy Sifa.
Drafting - original draft: Elly Pusporani, Ghisella Asy Sifa.
Writing - proofreading and editing: Ghisella Asy Sifa.