Enhancing PM_{10} Forecasting in Malaysian Industrial Area: A Hybrid Model with Feature Selection Using Weight by Deviation Method

Izzati Amani Mohd Jafri[1,2], Norazian Mohamed Noor[1,2,*], Nur Alis Addiena A Rahim[1,2]

[1]Faculty of Civil Engineering & Technology, Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia
[2]Geopolymer & Green Technology, Centre of Excellence (CEGeoGTech),
Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia

ABSTRACT

PM_{10}, a critical air pollutant, is of particular concern in industrial settings due to its adverse health and environmental impacts. This study focuses on enhancing PM_{10} forecasting in a Malaysian industrial area through the development of a hybrid model. A hybrid model for next-day PM_{10} concentrations in peninsular Malaysia, specifically at Bukit Rambai and Larkin was developed. Hourly data during haze events in 1997, 2005, 2013, and 2015, encompassing air pollutant concentrations (PM_{10}, NO_x, NO_2, SO_2, CO, O_3) and meteorological parameters (RH, T, WS), were utilized. A hybrid model (D-QR) was constructed by combining the weight by deviation filter feature selection technique with Quantile Regression (QR) to reduce the number of input variables. Performance evaluation employing indicators such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Index of Agreement (IA) demonstrated the D-QR effectiveness of the model across all areas. Both stations exhibited stable performance, with Bukit Rambai recording an RMSE of 10.58 and MAE of 6.53, while Larkin displayed slightly higher values (RMSE: 14.67, MAE: 8.22). This strategy proved successful in reducing model complexity and enhancing predictive capacity.

Keywords: Air quality forecasting, particulate matter, hybrid modelling, feature selection, weight by deviation.

1. INTRODUCTION

In recent years, Malaysia has experienced deteriorating air quality due to rapid urbanization, industrial growth, and increased industry demands [1]. Local anthropogenic sources such as vehicle emissions and peatland fires are significant contributors to air pollution [2]. Additionally, transboundary haze, mainly from Indonesian wildfires during the dry season and southwest monsoon, has emerged as a major cause of air pollution in Malaysia [3]. Numerous studies have demonstrated the adverse effects of air pollution on respiratory and circulatory systems [4].

Recognizing the continuous challenge of air pollution, numerous studies have sought to predict air pollutant concentrations. One of the most known method among researchers was statistical methods for its simplicity and reliability in handling linear distributions [5]. However, the predominant focus on overall mean PM_{10} levels may be inappropriate, especially in extreme conditions [6].

To address this need, this research article introduces a novel approach for enhancing PM_{10} forecasting in Malaysian industrial areas by employing a hybrid model and utilizing the feature selection method. This approach aims to improve the accuracy and reliability of PM_{10} predictions, contributing to better air quality management.

[*] Corresponding authors: norazian@unimap.edu.my

2. EXPERIMENTAL PROCEDURE

Continuous hourly data of air pollutants which is Particulate Matter (PM_{10}), Nitrogen Oxide (NO_x), Sulphur dioxides (SO_2), Surface Ozone (O_3), Nitrogen Dioxides (NO_2), Carbon Monoxide (CO) and meteorological parameters namely Temperature (T), Windspeed (WS) and Relative Humidity (RH) were obtained from Department of Environment (DOE), Malaysia. As shown in table 1, two stations located in peninsular Malaysia namely Bukit Rambai (Melaka) and Larkin (Johor) was chosen as location study. These stations reside at the west coast of peninsular Malaysia and prone to the transboundary smoke from the Sumatera regions. The data selected for this study were from the years when Malaysia experienced historic haze episode (1997, 2005, 2013, and 2015).

Table 1 Location of study area

Study Area	Latitude(N)	Longitude(E)
Sek. Men. Keb. Bukit Rambai, Melaka	02°12.789’	102°14.364’
IPG Temenggong Ibrahim, Larkin, Johor Bharu	01°28.225’	103°53.637’

Before conducting the analysis, missing observations for all air pollutant parameters were initially filled in. The handling of these missing data involved employing the Linear Interpolation (LI) method with IBM SPSS Software Version 26.

2.1 Feature Selection by using Deviation Weighting

Feature selection is the data pre-processing step which is to reduce the number of input variables when developing a predictive model [7][8]. In this study, a filter method feature selection technique which is weight by deviation was selected. The technique involves calculating the deviation of each data point from the mean or median of the data. The deviation is a measure of how far a data point is from the center of the distribution [9]. Data points with higher deviations are assigned higher weights, while data points with lower deviations are assigned lower weights.

2.2 Selection of The Best Percentile

Quantile Regression (QR) model was develop by using SPSS version 26 to generates a set of coefficients and equations at nine percentiles, specifically at the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th. Given a random variable y with right continuous distribution, , $F_y = P_r(\, Y \leq y)$. The quantile regression $Q(\tau)$ with $\tau\epsilon(0,1)$ is defined as follows [10]:

$$Q(\tau) = inf\{y \colon F(y) \geq \tau\} \tag{1}$$

The quantile also formulated as the solution to minimize problem:

$$\hat{Q}_y(\tau) = arg\,\underset{a}{min}\left\{\textstyle\sum_{i:y_i \geq a} \tau|y_i - a| + \sum_{i:y_i<a}(1-\tau)|y_i - a|\right\} = arg\,\underset{a}{min}\textstyle\sum_i \rho\tau(y_i - a) \tag{2}$$

From equation 2, the quantile regression coefficients are obtained by solving with respect to

$$\hat{\beta}(\tau) = arg\,\underset{\beta(\tau)\epsilon R^k}{min}\left\{\textstyle\sum_{i:y_i \geq \acute{x}\beta(\tau)} \tau\,|y_i - x_l\beta(\acute{\tau})| + \sum_{i:y_i<\acute{x}\beta(\tau)}(1-\tau)|y_i - x_l\acute{\beta}(\tau)|\right\} \tag{3}$$

where i is equal to n observations; τ = specified percentile value (0.1,0.2,0.3…,0.9); y_i = dependent variable (predicted PM_{10} level); x_l are the explanatory variables (air pollutants and weather parameters); β is the y-intercept with a dependency on the τ (constant term); $\hat{\beta}$ are the slope coefficients for each explanatory variable with a dependency on the τ.

Consider the result of the model performance evaluation, the percentile that show minimum error and high accuracy was chosen to be used in developing of the hybrid model. Three performance indicators were used to evaluate the performance of the model, which is Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Index of Agreement (IA). The performance indicator formulae was shown as follows [11]:

$$RMSE = \frac{1}{N}\sum_{i=1}^{N} |P_i - O_i| \tag{4}$$

$$MAE = \sum_{i=1}^{n} \frac{Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i} \tag{5}$$

$$IA = 1 - [\frac{\sum_{i=N}^{N}(P_i - O_i)^2}{\sum_{i=N}^{N}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \tag{6}$$

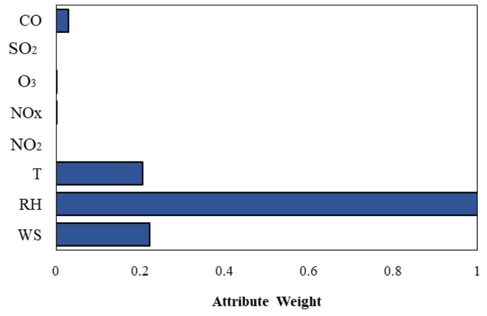
3. RESULTS AND DISCUSSION

Table 2 presents the summary of the best percentile chosen for each monitoring station in the predictions for the next day PM_{10} concentration. The selected percentile for all stations was 0.6. The selected percentile fell in the intermediate range enables the model to achieve equilibrium between capturing the central tendency of air quality conditions and accommodating a certain degree of extreme or exceptional occurrences [12]. The percentile of 0.6, which exhibits the lowest error was determined to be the ideal percentile for implementation in hybrid modelling.

Table 2 Summary of best selected percentile

Monitoring Station	Best Percentile
Bukit Rambai	0.6
Larkin	0.6

Weighting by deviation reveals that RH exhibits the highest weighting as evidence in Figure 1. Both stations show similar ranking order with the same first four parameter which is RH > WS > T > CO. It was expected since weighting by deviation assigns weights to data points based on their deviation from the mean or median of the data. Features that exhibit higher deviation from the mean or central tendency are assigned higher weights, indicating their greater impact on the overall variability of the data [13].



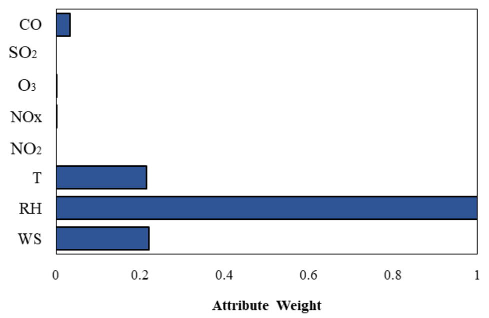


Figure 1. The rank of input parameters.

Table 3 shows the performance measure of best Deviation-Quantile Regression (D-QR) model for the next day PM_{10} concentration in Bukit Rambai and Larkin. The models at both stations demonstrate a stable level of performance, showing an RMSE value of 10.58 and MAE of 6.53 at Bukit Rambai. Hence, Larkin exhibits slightly higher values with an RMSE of 14.67 and MAE of 8.22. It was also noted that both stations exhibit excellent performance, as indicated by the IA values surpassing 0.9.

Table 3 The performance measure of the best Deviation-QR model for the next day PM_{10} concentration

Location	RMSE	MAE	IA
Bukit Rambai	10.58	6.53	0.98
Larkin	14.67	8.22	0.93

The findings of the study indicate that D-QR models have demonstrated superior performance in predicting outcomes with the different number of input parameters. As shown in Table 4, D-QR model at Larkin (D-3) exhibits a lower number of selected input parameters compared to Bukit Rambai (D-5) which is H, WS, T and H, WS, T, CO, O_3, respectively.

The difference in the number of selected parameters highlights the significance of customizing predictive models to match the unique characteristics of each monitoring station. The D-QR model highlights its adaptability, demonstrating how it can optimize performance by adjusting input parameters based on the specific characteristics of the monitoring site [14].

Table 4 The summary of the best model and selected input parameters

Location	Model	Input Parameter
Bukit Rambai	D-5	H, WS, T, CO, O_3
Larkin	D-3	H, WS, T

Lastly, feature selection method was proven as an effective and efficient method in to exclude the features from the data set that are deemed to be of the least significance [15]. This method was worthy to be used in data pre-processing to achieve an efficient data reduction.

4. CONCLUSION

In conclusion, D-QR model was a suitable and effective predictive method to predict the next day PM_{10} concentration. It was determined that the percentile of 0.6 is the ideal choice for implementation in hybrid modeling. The analysis of weighting by deviation identified RH as the variable with the highest weight. Performance evaluation of the models at Bukit Rambai and Larkin stations revealed stability, with slightly higher error values observed at Larkin. However, both stations demonstrated excellent overall performance, surpassing an IA value of 0.9. The findings highlighted the superior predictive performance of D-QR models across different numbers of input parameters, emphasizing their adaptability to specific characteristics, underscores its suitability for predictive modeling in environmental applications.

ACKNOWLEDGEMENT

The author extends their appreciation to the Ministry of Higher Education Malaysia for their generous support through the FRGS/1/2020/TK0/UNIMAP/02/53 grant and also to the Department of Environment, Malaysia (DOE) for the valuable contribution of the air quality dataset.

REFERENCES

[1] Samsuri, A., Marzuki, I., Si Yuen, F., Mahfoodh, A., & Ahmed, A. N. (2016). Evaluation for long term PM10 concentration forecasting using multi linear regression (MLR) and principal component regression (PCR) models. *EnvironmentAsia*, 9(2), 101–110.

[2] Department of Environment M, Malaysia Environmental Quality Report 2016. 2016.

[3] Abdullah, S., Napi, N. N. L. M., Ahmed, A. N., Mansor, W. N. W., Mansor, A. A., Ismail, M., ... & Ramly, Z. T. A. (2020). Development of multiple linear regression for particulate matter (PM_{10}) forecasting during episodic transboundary haze event in Malaysia. *Atmospher*e, 11(3), 289.

[4] Sahani, M., Zainon, N. A., Mahiyuddin, W. R. W., Latif, M. T., Hod, R., Khan, M. F., Tahir, N. M. and Chan, C. C. (2014). A case-crossover analysis of forest fire haze events and mortality in Malaysia. *Atmospheric Environment*, 96, 257-265.

[5] Bai, L., Liu, Z., & Wang, J. (2022). Novel hybrid extreme learning machine and multi-objective optimization algorithm for air pollution prediction. *Applied Mathematical Modelling*, 106, 177-198.

[6] Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N., & Hamid, H. A. (2012). Performance of multiple linear regression model for long-term PM_{10} concentration prediction based on gaseous and meteorological parameters. *Journal of applied sciences*, 12(14), 1488-1494.

[7] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in biology and medicine*, 112, 103375.

[8] Ul-Saufie, A. Z., Hamzan, N. H., Zahari, Z., Shaziayani, W. N., Noor, N. M., Zainol, M. R. R. M. A., Sandu, A. V., Deak, G. and Vizureanu, P. (2022). Improving air pollution prediction modelling using wrapper feature selection. *Sustainability*, 14(18), 11403.

[9] Ahmat, H., Yahaya, A. S., & Ramli, N. A. (2015). PM_{10} analysis for three industrialized areas using extreme value. *Sains Malaysiana*, 44(2), 175-185.

[10] Yong, N. K., & Awang, N. (2017). Quantile regression for analysing PM_{10} concentrations in Petaling Jaya. *Malaysian Journal of Fundamental and Applied Sciences*, 13(2), 86-90.

[11] Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2015, January). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. *In Materials Science Forum* (Vol. 803, pp. 278-281). Trans Tech Publications Ltd.

[12] Sayegh, A. S., Munir, S., & Habeebullah, T. M. (2014). Comparing the performance of statistical models for predicting PM_{10} concentrations. *Aerosol and Air Quality Research*, 14(3), 653-665.

[13] Marinov, E., Petrova-Antonova, D., & Malinov, S. (2022). Time Series Forecasting of Air Quality: A Case Study of Sofia City. *Atmosphere*, 13(5), 788.

[14] Yousefpour, A., Ibrahim, R., Abdull Hamed, H. N., & Hajmohammadi, M. S. (2014). Feature reduction using standard deviation with different subsets selection in sentiment analysis. In Intelligent Information and Database Systems: 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, April 7-9, 2014, Proceedings, Part II 6 (pp. 33-41). Springer International Publishing.

[15] Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008, September). On the relationship between feature selection and classification accuracy. In New challenges for feature selection in data mining and knowledge discovery (pp. 90-105). PMLR.