Prediction of Particulate Matter (PM_{10}) during High Particulate Event in Peninsular Malaysia using Novel Hybrid Model

Izzati Amani Mohd Jafri[1,2], *Norazian* Mohamed Noor[1,2*], *Nur Alis Addiena* A Rahim[1,2], *Ahmad Zia* Ul Saufie[2,3], *György Deak* Habil[4]

[1]Faculty of Civil Engineering & Technology, Universiti Malaysia Perlis, Jejawi 02600, Perlis, Malaysia
[2]Sustainable Environment Research Group (SERG), Centre of Excellence Geopolymer and Green Technology (CEGeoGTech), Universiti Malaysia Perlis, Jejawi 02600, Perlis, Malaysia
[3]Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), Shah Alam 40450, Selangor, Malaysia
[4]National Institute for Research and Development in Environmental Protection INCDPM, Splaiul Independentei 294, 060031 Bucharest, Romania

Abstract. High Particulate Events (HPE) contributes to the deterioration of air quality, as the fine particles present can be inhaled, leading to respiratory diseases and other health problem. Knowing the adverse effects of air pollution episodes to human health, it is crucial to create suitable models that can effectively and accurately predict air pollution concentration. This study proposed a hybrid model for forecasting the next day PM_{10} concentration in peninsular Malaysia namely Shah Alam, Nilai, Bukit Rambai and Larkin. Hourly air pollutant concentration (PM_{10}, NO_x, NO_2, SO_2, CO, O_3) and meteorological parameters (RH, T, WS) during the HPE events in 1997, 2005, 2013 and 2015 were used. Support Vector Machine (SVM) and Quantile Regression (QR) was combined to construct a hybrid models (SVM-QR) to reduce the number of input variables. Performance indicators such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Index of Agreement (d_2) were used to evaluate the performance of the predictive models. SVM-QR model resulted good performance in all areas. SVM-3 was selected as the best model at Bukit Rambai (MAE=5.72, RMSE=9.71) and Shah Alam (MAE=11.89, RMSE=22.66), while SVM-1 as the best model at Larkin and Nilai with the value (MAE=7.22, RMSE=13.38) and (MAE=6.88, RMSE=11.84), respectively. This strategy was proven to help reducing the complexity of the model and enhance the predictive capacity of the model.

* Corresponding author: norazian@unimap.edu.my

1 Introduction

Extremely high concentration of Particulate Matter (PM) in ambient air or known as high particulate event (HPE) usually occurred during haze episodes. These severe episodes occurs when the concentrations of PM far exceeded the Malaysian Ambient Air Quality Standard (MAAQS) for PM_{10} concentration which was 150 $\mu g/m^3$ for 24-hour average [1].

Harrison et al. [2] reported that peat land fire is a significant contributor to high particulate events in Southeast Asia. The increasing severity of air pollution conditions can be attributed to the potent impact of wildfire pollution originating from neighbouring countries, coupled with the influence of the southwest monsoon, which typically intensifies biomass burning during the dry and hot weather conditions prevalent during this season [3,4].

The escalation of air pollution has resulted in significant health consequences which includes higher mortality rates, respiratory and cardiovascular diseases. Numerous academic studies have demonstrated the impact of air pollution on the respiratory and circulatory systems [5–7]. Since air pollution is a problem that has been occurred over decades, there are a lot of researches that has been conducted to predict air pollutant concentration.

One of the most common method used is statistical approach was linear regression as it was foreknown due to its simplicity and reliability especially dealing with linear distribution. However, most of the studies focuses on overall mean of PM_{10} levels that is not appropriate to be used during extreme condition. Linear regression may not provide accurate predictions in some complex situations such as non-linear data and extreme values data [6].

When all devastating effects of air pollutants considered, it is crucial to create suitable models to predict air pollution levels in order to determine future concentrations or to locate pollutant sources. Therefore, this study proposed a hybrid model for forecasting of PM_{10} concentration in selected areas in peninsular Malaysia during HPE. Feature selection process was proposed to reduce the input variables before developing the hybrid models.

2 Methodology

2.1 Data Acquistition

Four stations located in peninsular Malaysia namely Larkin (Johor), Bukit Rambai (Melaka), Nilai (Negeri Sembilan) and Shah Alam (Selangor) were chosen as location study. Table 1 describe the background of each of the monitoring stations. Larkin, Bukit Rambai and Nilai was located at the southern region of peninsular Malaysia while Shah Alam was located at near the central peninsular of Malaysia. It also known as urban-industrial area with residential and commercial areas surrounded by busy motorways [8]. All of these stations were prone to the transboundary smoke from the Sumatera regions as these stations resides at the west coast of peninsular Malaysia.

Continuous hourly data of air pollutants which is Particulate Matter (PM10), Nitrogen Oxide (NOx), Sulphur dioxides (SO2), Surface Ozone (O3), Nitrogen Dioxides (NO2), Carbon Monoxide (CO) and meteorological parameters namely Temperature (T), Windspeed (WS) and Relative Humidity (RH) were obtained from Department of Environment (DOE), Malaysia. These data was in the year that Malaysia experienced historic HPE (1997, 2005, 2013 and 2015) were chosen in this study.

Table 1. Specific location of the monitoring stations and background

Monitoring Station	Latitude (N)	Longitude (E)	Study Area
Taman Semarak (Phase II), Nilai, Negeri Sembilan	02°49.246’	101°48.877’	Industrial
Sek. Men. Keb. Bukit Rambai, Melaka	02°12.789’	102°14.364’	Industrial
IPG Temenggong Ibrahim, Larkin, Johor Bharu	01°28.225’	103°53.637’	Industrial
Sek. Keb. TTDI Jaya, Shah Alam, Selangor	03°06.286’	101°33.367’	Urban

2.2 Data Pre-processing

Firstly, the missing observation of all air pollutant parameters were first fill-in before the analysis were done. These missing data will be treated by using Linear Interpolation (LI) method using IBM SPSS Software Version 26. It is important to fill in the missing data before any analysis because the success of the modelling depends on the quality of the dataset [9,10]. A random selection of 80% of the data was used to develop the model, and the remaining 20% of the data was used to evaluate the model’s accuracy.

2.3 Development of Hybrid Model

2.3.1 Feature Selection by using SVM weighting

Feature selection is the process of reducing the number of input variables when developing a predictive model [11]. This method used in data pre-processing to achieve efficient data reduction [12]. In this study, a filter based feature selection method which is Support vector machine (SVM) was selected. SVM are a set of supervised learning methods used for classification, regression and outlier detection [13]. This selection method was conducted by using SVM based operators in RapidMiner Studio version 9.10. This operator uses the coefficients of the normal vector of a linear SVM as attribute weights and calculates the relevance of the attributes by computing for each attribute of the input dataset and the weight with respect to the class attribute [14]. This weight by SVM operation works as a filter process and ranked the air pollutants parameters before it can be used for modeling.

2.3.2 Quantile Regression

Quantile Regression (QR) will generates a set of coefficients and equations at all quantiles which can examine the entire distribution of the variable of interest rather than a single measure of the central tendency of its distribution[15]. Therefore, QR method is able to provide a more holistic picture of the effects of predictors at various PM_{10} distributions. 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 [16]:

$$Q(\tau) = inf\{y: F(y) \geq \tau\} \qquad (1)$$

The quantile also formulated as the solution to minimize problem:

$$\hat{Q}_y(\tau) = arg\ \min_a \{\sum_{i:y_i \geq a} \tau |y_i - a| + \sum_{i:y_i < a} (1-\tau)|y_i - a|\} = arg\ \min_a \sum_i \rho\tau(y_i - a) \quad (2)$$

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

$$\hat{\beta}(\tau) = arg \min_{\beta(\tau) \epsilon R^k} \{\sum_{i:y_i \geq \acute{x}\beta(\tau)} \tau\, |y_i - x_l\beta(\acute{t})| + \sum_{i:y_i < \acute{x}\beta(\tau)} (1-\tau)|y_i - x_l\acute{\beta}(\tau)|\} \quad (3)$$

where i is equal to n observations; τ = specified percentile value (0.1,0.2,0.3…,0.9); y_i = dependent variable (predicted PM10 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.

This research study used SPSS version 26 software to develop models based on nine percentiles, specifically at the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th percentiles. The performance across different percentiles was studied to understand how well the models fit the data at each specific quantile. Consider the result of the model performance evaluation, the percentile that show minimum error and high accuracy was chosen as the best percentile to be used in developing the hybrid model. Then, a hybrid were developed by combining the filter-based feature selection methods called Support Vector Machine (SVM) with Quantile Regression (QR).

2.4 Performance Indicator

In order to evaluate the performance of the regression model, several performance indicators such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Index of Agreement (d_2), and Prediction Accuracy (d_2)) was used to describe the goodness of fit for the predicting models of PM10 concentration. The performance indicator formulae was shown as in table 2[10]:

Table 2. The formula of performance indicator

Performance Indicator	Formula
Mean Absolute Error (MAE)	$NAE = \sum_{i=1}^{n} \frac{Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i}$ (4)
Root Mean Squared Error (RMSE)	$RMSE = \frac{1}{N}\sum_{i=1}^{N} \lvert P_i - O_i \rvert$ (5)
Index of Agreement (d_2)	$d_2 = 1 - [\frac{\sum_{i=N}^{N}(P_i - O_i)^2}{\sum_{i=N}^{N}(\lvert P_i - \overline{O} \rvert + \lvert O_i - \overline{O} \rvert)^2}$ (6)

3 Result and Discussion

3.1 Optimizing the Number of Predictors (SVM Weight)

The SVM weights were ranked according to their attribute weights; the higher the attribute weight, the more significant the variable was for the purpose of developing a hybrid models. Figure 1(a) to 1(d) illustrate the SVM weight for all study areas. CO was selected as the most significant parameter for all monitoring stations. CO which mainly released by motor vehicle and machinery that used diesel fuel seem to have the strongest correlation with PM10 concentration. Since all locations in this study are classified as industrial and or urban area,

it may contribute to these results. Furthermore, the seasonal fires from Indonesia can also be the main contribution since it was during haze [17].

The weighting by SVM order for Bukit Rambai starting to differ after the second rank as it was followed by O_3 > NO_x > T > RH > SO_2 > NO_2 and WS. Larkin and Nilai has the same parameter rank from first to third order which is CO, SO_2 and NO_2 respectively and begin to differ after. For Larkin, the parameters that ranked from the 4th to 6th were NO_x > RH > WS > T and O_3 ; meanwhile, Nilai O_3 > WS > RH > NO_x and T. Shah Alam showed the ascending order of weight by SVM parameter as follows ; CO > NO_2 > NO_x > RH > T > O_3 > WS > SO_2 .

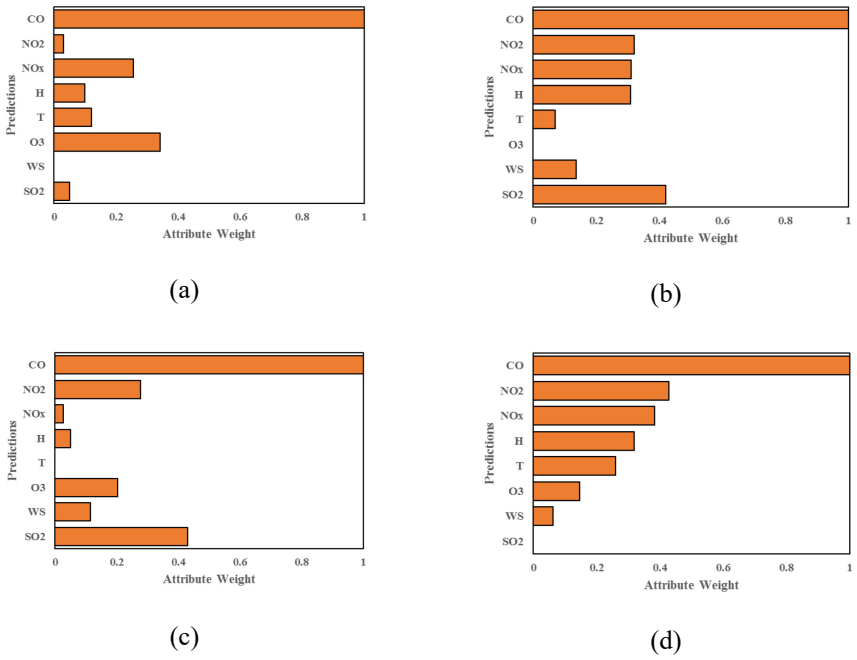

Fig 1. Attribute weight by SVM for (a) Bukit Rambai, (b) Larkin, (c) Nilai and (d) Shah Alam.

3.2 Quantile Selection

In order to select the best quantile to be modified with SVM features, the performance of QR method was evaluated. Table 4 shows the summary of best selected percentile of PM_{10} concentration prediction for the next-day at the 4 study areas. Table 4 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 fell within the range of 0.5 and 0.6. The percentile selected for best prediction performance at Bukit Rambai, Nilai and Larkin was 0.6, except for Shah Alam which achieved a percentile of 0.5. The percentile that exhibits the lowest error was determined to be the ideal percentile for implementation in hybrid modelling. Opting for percentiles in the intermediate range of 0.5 to 0.6 enables the model to achieve equilibrium between capturing the central tendency of air quality conditions, which represents typical or average values, and accommodating a certain degree of extreme or exceptional occurrences [18].

Table 3. Summary of best selected percentile

Monitoring Station	Best Selected Percentile
Bukit Rambai	0.6
Nilai	0.6
Larkin	0.6
Shah Alam	0.5

3.3 Hybrid Model

Table 4 shows the summary of the best prediction model of for the next day PM10 concentration according to model number and optimum input parameter. The result of SVM-QR model selected input parameter is vary depending on the location. SVM-1 model with one input parameter which is CO was the best model for Larkin and Nilai. Hence, Bukit Rambai and Shah Alam best next day prediction model was SVM-3 with 3 selected input parameter which is CO, O_3, NO_x and CO, NO_2, NO_x , respectively. It also can be concluded that CO has the strongest correlation with PM_{10} concentration especially during HPE.

Table 4. The summary of the best prediction model according to model number and input parameter

Location	Model Number	Input Parameter
Bukit Rambai	SVM-3	CO, O_3, NO_x
Larkin	SVM-1	CO
Nilai	SVM-1	CO
Shah Alam	SVM-3	CO, NO_2, NO_x

Overall, hybrid model exhibits exceptional performance across all stations. Table 5 shows the performance measures of hybrid model for next day PM10 concentration at all study areas. Bukit Rambai has the best performance with 3 selected parameters (SVM-3) with the value of MAE=5.72, RMSE=9.71 and $\boldsymbol{d_2}$=0.99. Similar with Bukit Rambai, SVM-3 model also show the best performance at Shah Alam.

Table 5. Performance measure of hybrid model for next day PM10 concentration at all study areas

Area	Method		MAE	RMSE	d_2
Bukit Rambai	MLR		9.05	14.05	0.97
	QR	0.6	6.58	10.65	0.98
	SVM-3		**5.72**	**9.76**	**0.99**
Larkin	MLR		16.6	21.86	0.85
	QR	0.6	8.11	14.73	0.93
	SVM-1		**7.22**	**13.38**	**0.95**
Nilai	MLR		12.55	18.67	0.94
	QR	0.6	11.31	16.91	0.95
	SVM-1		**6.88**	**11.84**	**0.98**
Shah Alam	MLR		11.92	18.81	0.95
	QR	0.5	14.33	25.08	0.89
	SVM-3		**11.89**	**22.66**	**0.92**

Larkin and Nilai recorded the SVM-1 as best model (with only 1 optimum input parameter) with the value (MAE=7.22, RMSE=13.38, R^2=0.85, IA=0.95) and (MAE=6.88, RMSE=11.84, R^2=0.95, IA=0.98), respectively. The performance of hybrid model was compared with two others traditional regression method which is QR and QR model exhibits least error in comparison to MLR methods at all monitoring stations except for Shah Alam. However, QR does not surpasses the performance of hybrid model. Overall, these results imply that the SVM–QR model was the most accurate predictive model to predict PM10 concentration during HPE. The performance indicators results for SVM-QR for all stations calculated less error and greater accuracy by compared to MLR and QR method. The SVM feature selection approach was an excellence method to optimize the number of selected parameters in predicting the PM_{10} concentration.

4 Conclusion

In conclusion, SVM-QR is an excellent alternative method for predicting PM_{10} concentration. This model saves training time by reducing the feature size given in the data representation, and prevents learning from noise, also known as overfitting, to improve accuracy. The proposed model can accurately predict maximum daily air pollution episodes for the next days; it can be used as an early warning tool in giving air quality information to local authorities to formulate air quality improvement strategies.

Acknowledgement

Author would like to thank the Ministry of Higher Education Malaysia for the FRGS/1/2020/TK0/UNIMAP/02/53 and the Department of Environment, Malaysia (DOE) for providing the air quality dataset.

References

1. A. Z. Mohd Zahid, N. N. A. Abdul Malik, and J. Kassim, MATEC Web Conf. **250**, 1 (2018)
2. M. E. Harrison, S. E. Page, and S. H. Limin, Biol. | **56**, (2009)
3. W. N. Shaziayani, A. Z. Ul-Saufie, H. Ahmat, and D. Al-Jumeily, Air Qual. Atmos. Heal. **14**, 1647 (2021)
4. M. Radzi Bin Abas, D. R. Oros, and B. R. T. Simoneit, Chemosphere **55**, 1089 (2004)
5. W. R. Wan Mahiyuddin, M. Sahani, R. Aripin, M. T. Latif, T. Q. Thach, and C. M. Wong, Atmos. Environ. **65**, 69 (2013)
6. B. J. Lee, B. Kim, and K. Lee, Toxicol. Res. **30**, 71 (2014)
7. C. H. Linaker, D. Coggon, S. T. Holgate, J. Clough, L. Josephs, A. J. Chauhan, and H. M. Inskip, Thorax **55**, 930 (2000)
8. S. Abdullah, N. N. L. M. Napi, A. N. Ahmed, W. N. W. Mansor, A. A. Mansor, M. Ismail, A. M. Abdullah, and Z. T. A. Ramly, Atmosphere (Basel). **11**, (2020)
9. E. Marinov, D. Petrova-Antonova, and S. Malinov, Atmos. 2022, Vol. 13, Page 788 **13**, 788 (2022)
10. M. N. Norazian, Y. A. Shukri, R. N. Azam, and A. M. M. Al Bakri, ScienceAsia **34**, 341 (2008)
11. B. Remeseiro and V. Bolon-Canedo, Comput. Biol. Med. **112**, 103375 (2019)
12. A. Z. Ul-Saufie, N. H. Hamzan, Z. Zahari, W. N. Shaziayani, N. M. Noor, M. R. R. M. A. Zainol, A. V. Sandu, G. Deak, and P. Vizureanu, Sustain. **14**, 1 (2022)

13. K. Yan and D. Zhang, Sensors Actuators, B Chem. **212**, 353 (2015)
14. H. T. Shahraiyni and S. Sodoudi, Atmosphere (Basel). **7**, 10 (2016)
15. R. Koenker and G. Bassett, Econometrica **46**, 33 (1978)
16. J. C. M. Pires, F. G. Martins, S. I. V. Sousa, M. C. M. Alvim-Ferraz, and M. C. Pereira, Am. J. Environ. Sci. **4**, 445 (2008)
17. V. Huijnen, M. J. Wooster, J. W. Kaiser, D. L. A. Gaveau, J. Flemming, M. Parrington, A. Inness, D. Murdiyarso, B. Main, and M. Van Weele, Sci. Rep. **6**, (2016)
18. A. S. Sayegh, S. Munir, and T. M. Habeebullah, Aerosol Air Qual. Res. **14**, 653 (2014)