Prediction of PM_{10} Level During High Particulate Event in Malaysia Using Modified Model

Nur Alis Addiena A Rahim[1,2], *Norazian* Mohamed Noor[1,2], *Izzati Amani* Mohd Jafri[1,2], *Ahmad Zia* Ul Saufie[2,3] and *Boboc* Madalina[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 Bucharest (INCDPM), 294 Splaiul Independentei Street, 6th District, 060031 Bucharest, Romania

Abstract. Particulate matter (PM_{10}) is one of the key indicator of air quality index (API) during high particulate event (HPE). PM_{10} can cause adverse effect on human health and environment; hence, it is important to develop a reliable and accurate predictive model to be used as forecasting tool to alarm the citizen especially during HPE. This study aims to develop a modified Quantile Regression (QR) model to forecast the PM_{10} concentration during HPE in Malaysia. The performances of three predictive models namely Multiple Linear Regression (MLR), Quantile Regression (QR) and a modified QR models i.e. combination of QR with Relief-based were compared. The hourly dataset of PM_{10} concentration with other gaseous pollutants and weather parameters at Klang from the year with severe haze event in Malaysia (1997, 2005, 2013 and 2015) were obtained from Department of Environment (DOE) Malaysia. Three performance measures namely Mean Absolute Error (MAE), Normalised Absolute Error (NAE) and Root Mean Squared Error (RMSE) were calculated to evaluate the accuracy of the predictive models. This study found that the Relief-QR model showed the best performance compared to MLR and QR models. The prediction of future PM_{10} concentration is very important because it can aid the local authorities to implement precautionary measures to limit the impact of air pollution.

1 Introduction

Malaysia experienced air pollution issue for over a decade as a result of high particulate event (HPE) from its neighbouring country, Indonesia. Consequently, the occurrence of HPE is not exceptional in Malaysia as it was first recorded back in the year 1982 when regional haze from biomass burning disrupted daily life in Malaysia [1]. Since then, several episodes of HPE have been reported whereby the concentrations of particulate matter (PM) with an aerodynamic diameter of less than 10 μm (PM_{10}) concentrations greatly exceeded the Malaysian Air Quality Guideline for PM_{10} concentration (150 μgm^{-3} for a 24-hour average) at one or more locations across Malaysia. In most years, the Malaysian air quality was influenced by the occurrence of dense HPE episodes. A research on air quality in Kuala Lumpur by [2] found that the smoke haze was associated with high levels of suspended micro particulate matter, but with relatively low levels of other gaseous pollutants such as carbon monoxide, nitrogen dioxide, sulphur dioxide, and ozone. Series of HPEs were recorded in peninsular Malaysia, Sabah, and Sarawak in 1991, in 1994, and during September and October of 1997, resulted from the significant amounts of particle matter that have been transported by south-westerly winds from neighbouring country due to uncontrolled biomass burning activities. This is common at some poorly managed disposal sites and results in smoke and fly ash problems. The large-scale forest and plantation fires, mainly in southern Sumatra and central Kalimantan, both in neighbouring Indonesia have contributed to the cause of the occurrence in 1997. Department of Environment (DOE) Malaysia reported the HPE episodes in Malaysia which can be highlighted with severe incidents recorded in the year 2005, 2013 and 2015 [3]. The crisis has also affected not just Malaysia but other neighbouring countries such as Singapore and Brunei. Health problems such as respiratory, cardiovascular diseases and increase mortality rate has long been linked to the long-term exposure to PM_{10} [4], [5]. The prediction of PM_{10} can provide a good insight allowing the government and authorities to plan appropriate proper mitigation actions in order to minimize the health issues arising due to exposure to PM_{10}.

Over the past decade, several studies have been conducted to predict air quality. However, the majority of these studies were restricted to utilizing a statistical approach. For example, a study by [6]–[8] forecasted PM_{10} level in the East Coast peninsular Malaysia during various monsoon was conducted to analyse its variation during usual condition of ambient atmosphere by developing a multiple linear regression (MLR) model, based on various site backgrounds. A study on the distribution of the ozone in Athens via quantile regression (QR) was conducted by [9]. The results of the study exhibited that the influence of independent variables vary over the quantile distributions of ozone and the nonlinear relationship between ozone and the independent variables was delineated by using QR. The number of inputs were not optimized in most of the studies. Therefore, this study aims to compare different approaches for selecting significant input variables before selecting the best one to predict the PM_{10} concentration.

Various researches implemented models to forecast air pollutants during usual condition but there are lack of studies that predict air pollutant especially PM_{10} specifically during HPE. This study focuses on developing hybrid model to forecast the PM_{10} concentrations specifically during HPE occurrence in Malaysia, by combining QR approach with Relief-based method. The development of single MLR and QR models, along with a hybrid model combining QR and Relief-based operator, was aimed at exploring different methodologies to forecast PM_{10} concentrations during haze episodes. However, the models exhibited a degree of bias due to the variation in weighting strategies and model complexity. The observed bias underscores the importance of acknowledging potential divergences in modelling approaches. The model developed will be very beneficial for local authority to take precautionary measures to avoid or minimize their exposure to unhealthy PM_{10} levels and introduce necessary actions aimed at improving air quality.

2 Methodology

2.1 Dataset

The hourly datasets at Klang that sited in the west coast region of peninsular Malaysia was obtained from Department of Environment (DOE) Malaysia. The dataset consisted of PM_{10}

concentration, gaseous pollutants such as nitrogen oxides (NOx), suhfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO). The meteorological parameters such as wind speed, temperature and humidity were also included in the dataset. The hourly data were taken from year 1997. 2005, 2013 and 2015 where severe haze was recorded in Malaysia.

2.2 Feature Selection

In this study, the process of feature selection, which involves reducing the number of input variables during the development of a predictive model, was employed using the filter method. It picks and retains only the most significant features from the dataset. Relief-Based Algorithm (RBA) was utilized in this study. RBA is a group of algorithms that select the most informative features from high-dimensional data sets based on their ability to distinguish between different classes [10]. The primary principle of "Relief" is to assess the quality of features by evaluating how effectively their values distinguish between cases of the same and different classes that are in close proximity to each other. Relief assesses the applicability of features by sampling examples and comparing current feature's value for the nearest example of the same and of a different class. Relevant parameters were selected by using RBA approach prior to modelling of PM_{10}. The datasets were evaluated by weight by Relief using RapidMiner software by computing the attribute weights for each parameter involved. The weights computed were normalized into the interval between 0 and 1 if the normalize weights parameter is set to true.

2.3 Prediction Model

In this study, PM_{10} concentrations for the next-day (PM_{10+24}), next-two-days (PM_{10+48}) and next-three-days (PM_{10+72}) were forecasted. The hourly data of the PM_{10} concentrations, gaseous pollutants and weather parameters were distributed into training and testing dataset. The training dataset was used to develop the prediction model, while the testing dataset was used in the model validation process. The training dataset consists of 80 percent of the data meanwhile 20 percent of the data was used for validation purposed. Three predictive models were developed which include Multiple Linear Regression (MLR), Quantile Regression (QR) and hybrid model (Relief-QR).

MLR is a widely used forecasting approach that predicts the outcome of a dependent variable by fitting a linear equation to observed data, considering the values of two or more independent variables. It is among the most commonly employed methods for making predictions in various fields.

QR was used to develop a model to predict the PM_{10} concentration at each study area. It is an extension of median regression that includes assessing the value of the parameter vector β from the range of acceptable vectors that reduces the mean loss function. The relation between a set of independent variables and specific percentiles of a dependent variable, is modelled using quantile regression. A series of coefficients and equations at several quantiles were produced using this approach. Consequently, a clear picture of how predictors affect PM_{10} concentrations at each quantile will be shown. This study adopted 9 quantiles (0.1 to 0.9 with an increment of 0.1) and thus 9 equations were generated. The quantile that exhibited best performance were selected to develop the hybrid model.

The hybrid model was developed by combining two models. QR models were combined with Relief-based algorithm to forecast the PM_{10+24}, PM_{10+48} and PM_{10+72}. It is expected that the hybrid model able to improve the accuracy and reduce the error of prediction model. Fig.1 illustrates the procedures involved in obtaining the best prediction model.

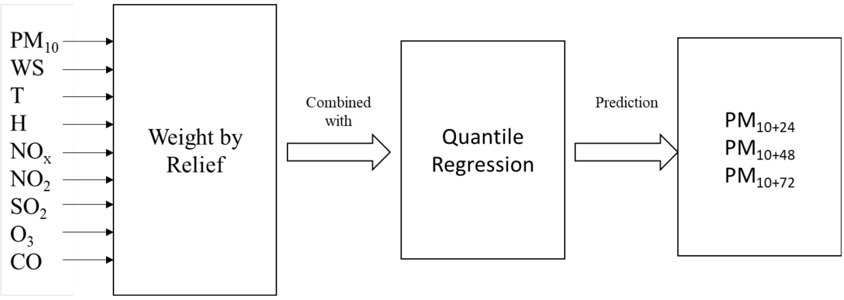

Fig. 1. Illustration of Relief-QR model.

2.4 Performance Indicator

Performance indicators (PI) based on the model's error such as Root Mean Squared Error (RMSE), Normalized Absolute Error (NAE) and Mean Absolute Error (MAE) were used to evaluate the prediction model for the PM_{10} concentration at each study location. The best method in forecasting PM_{10} concentration were chosen based on the least values of error for each of the PI.

3 Result and Discussion

Table 1 shows the performance measure for $PM_{10+24,}$ PM_{10+48} and PM_{10+72} in Klang, Malaysia. The Relief-QR prediction model gives a good performance in predicting PM_{10} level for three consecutive days during HPE. The model was compared with the MLR as well as QR. The proposed model achieved the least error compared to MLR and QR models, in terms of MAE, RMSE and MAE.

Table 1. Performance measures of prediction models in Klang.

Time	Method		MAE	NAE	RMSE
PM_{10+24}	MLR		27.85	0.36	42.85
	QR (p=0.4)		17.38	0.23	28.44
	Relief-QR (p=0.4)	1	13.71	0.18	25.44
		2	13.67	0.18	25.37
		3	**13.65**	**0.18**	**25.32**
		4	13.70	0.18	25.40
		5	13.69	0.18	25.39
		6	13.69	0.18	25.41
		7	13.67	0.18	25.68
		8	17.38	0.23	28.44

PM_{10+48}	MLR		38.21	0.50	58.95
	QR (p=0.4)		22.72	0.30	39.35
	Relief-QR (p=0.4)	1	18.83	0.25	35.35
		2	18.84	0.25	35.36
		3	**18.80**	**0.25**	**35.35**
		4	18.83	0.25	35.41
		5	18.83	0.25	35.40
		6	18.80	0.25	35.40
		7	19.37	0.25	36.67
		8	22.72	0.30	39.35
PM_{10+72}	MLR		37.73	0.49	56.52
	QR (p=0.4)		24.31	0.32	43.56
	Relief-QR (p=0.4)	1	21.51	0.28	40.58
		2	**21.48**	**0.28**	**40.53**
		3	21.51	0.28	40.60
		4	21.52	0.28	40.61
		5	21.51	0.28	40.60
		6	21.49	0.28	40.61
		7	22.50	0.29	42.53
		8	24.31	0.32	43.56

Referring to the Relief-QR model, the numbers from 1 to 8 were denoting to the parameters selected in Klang from the weight by relief method as shown in Table 2. It was detected that only CO, O_3 and SO_2 were the significant parameters in developing the best predictive model in Klang.

Table 2. Parameters selected from weight by relief approach in Klang.

Method	**Parameters**
1	CO
2	CO, O_3
3	CO, O_3, SO_2
4	CO, O_3, SO_2, NO_2

5	CO, O_3, SO_2, NO_2, NO_X
6	CO, O_3, SO_2, NO_2, NO_X, T
7	CO, O_3, SO_2, NO_2, NO_X, T, WS
8	CO, O_3, SO_2, NO_2, NO_X, T, WS, H

Fig. 2 to Fig. 4 illustrate the accuracy of all the prediction models for the next-day, the next-two-day and the next-three-day of PM_{10} level in Klang. Obviously, the proposed hybrid method reduced the calculated error for the prediction of PM_{10} concentration for the three consecutive days compared to MLR and QR at p=0.4. Hence, the proposed method can be considered as the most accurate predictive model for estimating PM_{10} level during HPE or haze event.

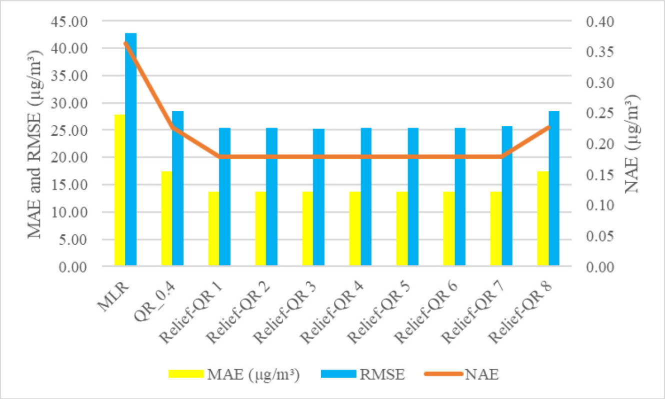

Fig. 2. Performance measure of prediction models in Klang for PM_{10+24}.

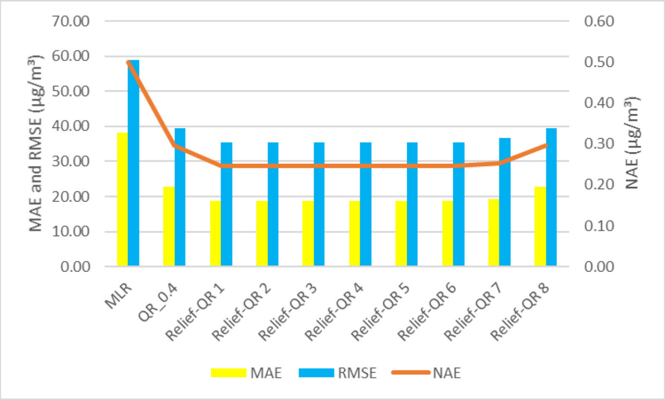

Fig. 3. Performance measure of prediction models in Klang for PM_{10+48}.

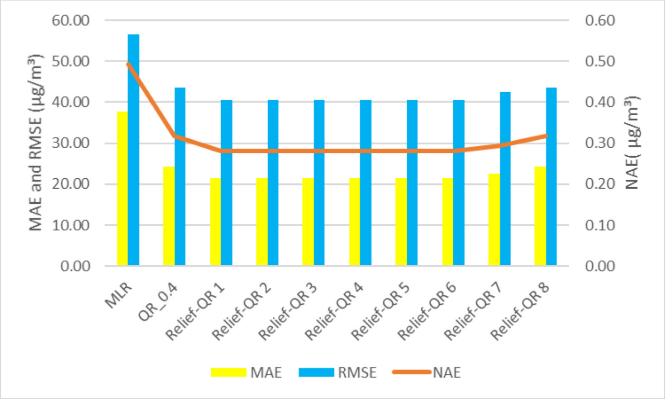

Fig. 4. Performance measure of prediction models in Klang for PM_{10+72}.

4 Conclusion

The hourly air quality parameters in Klang that is situated in the west coast of peninsular Malaysia during severe haze event in 1997, 2005, 2013 and 2015 were investigated. The goal of this study is develop a modified Quantile Regression (QR) model to forecast the PM_{10} concentration during HPE in Malaysia. The performance of Relief-QR model to predict the next-day (PM_{10+24}), the next-two-day (PM_{10+48}) and the next-three-day (PM_{10+72}) of PM_{10} level were assessed. Significant parameters were chosen to develop PM_{10} predictive model using feature selection i.e. Relief-based method. These models were compared with QR and MLR. MAE, RMSE and NAE were used to evaluate the performances of the models. It was proven that, Relief QR model at p=0.4, where CO, O_3, SO_2 were selected as the significant parameters, showed the best performance for the prediction of PM_{10} level in Klang for the next-day to the next-two-day. Meanwhile, Relief-QR with p=0.4, where CO and O_3 were selected as the significant parameters, was chosen as the best model in Klang for the next-three-day of PM_{10} prediction. Thus, it was verified that Relief-QR method can be one of the reliable method for predicting air quality specifically during HPE.

References

1. D. Glover, T. Jessup, International Development Research Centre, *Indonesia's Fires and Haze: The Cost of Catastrophe*, Institute of Southeast Asian Studies (2006)
2. M. Bin Awang et al., Respirology, *Air quality in Malaysia: Impacts, management issues and future challenges,* **5**, 183–196 (2000)
3. Department of Environment, Environmental Quality Report 2018 (2018)
4. H. J. Jahn, A. Schneider, S. Breitner, R. Eißner, M. Wendisch, A. Krämer, Int. J. Hyg. Environ. Health, *Particulate matter pollution in the megacities of the Pearl River Delta, China - A systematic literature review and health risk assessment*, **214**, 281–295 (2011)
5. L. Wen Zhang et al., Environ. Int, *Long-term exposure to high particulate matter pollution and cardiovascular mortality: A 12-year cohort study in four cities in northern China*, **62**, 41–47 (2014)
6. S. Abdullah, M. Ismail, S. Y. Fong, A. N. Ahmed, EnvironmentAsia, *Evaluation for Long Term PM10 Concentration Forecasting using Multi Linear Regression (MLR) and Principal Component Regression (PCR) Models*, **9**,101–110 (2016)

7. S. Abdullah, M. Ismail, S. Y. Fong, J. Sustain. Sci. Manag., *Multiple Linear Regression (MLR) Models for Long Term PM10 Concentration Forecasting During Different Monsoon Season*, **12**, 60–69 (2017)
8. S. Abdullah, M. Ismail, A. N. Ahmed, A. M. Abdullah, Atmos., *Forecasting Particulate Matter Concentration Using Linear and Non-Linear Approaches for Air Quality Decision Support*, **10** (2019)
9. D. Baur, M. Saisana, N. Schulze, Atmos. Environ., *Modelling the effects of meteorological variables on ozone concentration - A quantile regression approach*, **38**, 4689–4699 (2004)
10. R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, J. H. Moore, J. Biomed. Inform., *Relief-based feature selection: Introduction and review*, **85**, 189–203 (2018)