ISRAA UNIVERSITY JOURNAL OF APPLIED SCIENCE

Volume 6 :Issue1, October, 2022 | ISSN: 2523-0522

ORIGINAL ARTICLE

Evaluating the Performance of Random Forest and Multiple Linear Regression for Higher Observed PM_{10} Concentrations

Nurhafizah Ahmad[1], Ahmad Zia Ul-Saufie[2,*], Wan Nur Shaziayani[1] , Aida Wati Zainan Abidin[2], Nur Elis Sharmila Zulazmi[3], Suheir M. Harb[4]

Received: 30/04/2022 **Accepted:**19/06/2022 **Published:**01/10/2022

OPEN ACCESS

Doi:
https://doi.org/10.52865/WHPM9019

Competing Interests: The authors declare that this manuscript was approved by all authors in its form and that no competing interest exists.

Affiliation and Correspondence:

1 Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,13500 Pulau Pinang, MALAYSIA.

2 Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,40450 Shah Alam, Selangor, MALAYSIA.

3 Faculty of Chemical Engineering, Universiti Teknologi MARA,13500 Pulau Pinang, MALAYSIA.

4 Palestine Technical College, Dier Elbalah, Gaza Strip, PALESTINE.

*Corresponding Author. Email: ahmadzia101@uitm.edu.my

ABSTRACT:

Background: Air pollution is notable for its direct impact on human health. Hence, the ability to accurately predict air pollution concentrations is vital to raising public awareness of this issue and for better understanding of air quality management.

Aim: Therefore, the aim of this research is to predict PM_{10} concentrations in Malaysia, specifically on Langkawi Island using random forest and multiple linear regression.

Method: The predictive analytics were based on air pollution hourly data from 2003 until 2017. The eight parameters chosen in this study were PM_{10}, NO_2, O_3, CO, SO_2, Relative Humidity (RH), Temperature (T), and Wind Speed (WS). The findings revealed that PM_{10}, SO_2, NO_2, CO, and O_3 hourly trends at Langkawi Island were below the recommended Malaysian Ambient Air Quality Guidelines (MAAQG) standard. Multiple linear regression (MLR) and random forest (RF) were used for modelling and compared based on their prediction accuracy.

Result: The values of RMSE, NAE, IA, PA and R^2 for MLR were 8.0698, 0.1368, 0.8584, 0.7737 and 0.5984 respectively while the values of RMSE, NAE, IA, PA and R^2 for RF were 6.674038, 0.107664, 0.911974, 0.852570 and 0.726681 correspondingly. From the results, the RF method was chosen as a better model than MLR since both; the error measures and the accuracy measures results are close to 1. Nevertheless, the PM_{10} models (RF and MLR) are unable to take into account the higher observed concentrations.

Keywords: Air pollution, modelling, Random Forest, Multiple Linear Regression, Data mining

Introduction

For many years, the South East Asia region has experienced ongoing episodes of air pollution (Wen et al. 2016). The worst air pollution episode in 2015 has been reported as a result of the uncontrolled agriculture burning and activities in Indonesia which cause the transboundary haze pollution detected in Malaysia (DOE 2018). These events have demonstrated a high propensity for adverse effects on human respiratory difficulties and the environment (Shaziayani et al., 2021).

Air pollution can be caused by a wide range of both natural and man-made sources, such as windblown dust, volcanic eruptions, power plants, factories, vehicles and open burning of forests or any kind of garbage burning. As a matter of fact, the rapid growth of power plants in Malaysia over the past few years have contributed significantly to the reduction in air quality in Malaysia (Azid et al., 2013). This is due to the continuous release of high emissions of waste gas streams from power plants containing organic, inorganic, dust and chemicals pollutants during operational hours into the atmosphere.

In the face of increasingly serious environmental pollution problems, a significant amount of research has been conducted with predictions of air pollution being of primary importance (Bai et al., 2018). Certainly, accurate and precise estimates of air pollution are essential for effective pollution control measures. Artificial neural networks (ANN), MLR, and Principal component regression (PCR) analysis models are among the most popular methods of modelling that have been widely used for PM_{10} forecasting (Stadlober et al. 2008; Cai et al. 2009; Hrust et al. 2009; Slini et al.2006; Ul-Saufie et al. 2012; Abdullah et al. 2017; Yunus et al. 2017, Elbayoumi, 2018).

Multiple linear regression (MLR), though is commonly used to predict PM_{10} concentration, its results (R^2) were found out to have a low accuracy as compared to other techniques (Shahraiyni et al., 2016). According to Sayegh et al., (2014). MLR has several limitation. Due to the linear representation of non-linear structures, it has problems with precision and does not capture exceptional values. Several authors have specifically clarified the shortcomings of MLR inability to expand the response to non-central locations of explanatory variables, as well as it inability to fulfil model assumptions, such as homoscedasticity (Hao and Naiman 2007 and Ul- Saufie et al. 2012)

In recent years, a new model called random forest (RF) has been developed and has been outperform other model that existed (Ponggi et al. 2011; Ivanov et al. 2018 ; Brokamp et al. 2017; Kaminska 2018; Shamsoddini 2017; Hu et al. 2017). Yuen et al., (2018) considered RF method as the most precise compared to other methods (Yuen et al. 2018). Pan (2018) compared five data mining models in Tianjin, China and the results revealed that RF is more effective method with R^2= 0.9426 compared to MLR with R^2=0.8922.

Despite extensive research on the comparison of different modelling techniques (e.g., Kukkonen et al., 2003; Paschalidu et al., 2011; Ul-Saufie et al., 2011), no relevant research has been conducted to compare the performance of the MLR and RF models in predicting PM_{10} concentration for different levels of air pollution. The goal of this paper is therefore to compare the performance of the MLR and RF models in determining the best approach for modelling PM_{10} concentrations in three air quality categories: PM_{10} concentrations <50µg/m^3 (good), PM_{10} concentrations <100µg/m^3 (moderate) and PM_{10} concentrations ≥100µg/m^3 (unhealthy). This is the first study focused on differential air pollution/efficiency categories, and the findings would help improve air quality management. Currently, the Department of Environment (DOE), Malaysia only provide real-time data on the concentration of air pollution at a particular day and time. On that account, the aim of this study is to come out with a model that later could be used as a DOE prediction method. The ultimate goal would be a forecasting system of air pollution index in Malaysia that could predict accurately the state of air quality in advance.

Methodology

Data Acquisition

Data for air pollution control was obtained from the Department of Environment (DOE), Malaysia from 2003 and 2017. Among the air pollutants available to be used in the analysis were particulate matter (PM), carbon monoxide (CO), sulphur dioxide (SO2), nitrogen dioxide (NO2) and some of the variables such as wind speed (WS), relative humidity (RH) and temperature (T).

Langkawi Island, a small island off the north of Peninsular Malaysia, has been selected as the case study. Langkawi Island is anchored at 6.35° north latitude and 99.80° east longitude with a land area of 478.5 sq km and considered as one of Malaysia's most desirable tourism destinations. According to the official portal of Langkawi Municipal Council (2018), more than three million tourists visited Langkawi in 2015. This figure is has increased tremendeously from just over 2.4 million in 2010. In fact, the numbers is expected to increase for upcoming years. Fig. 1 shows the position of Langkawi Island in the Malaysia map.

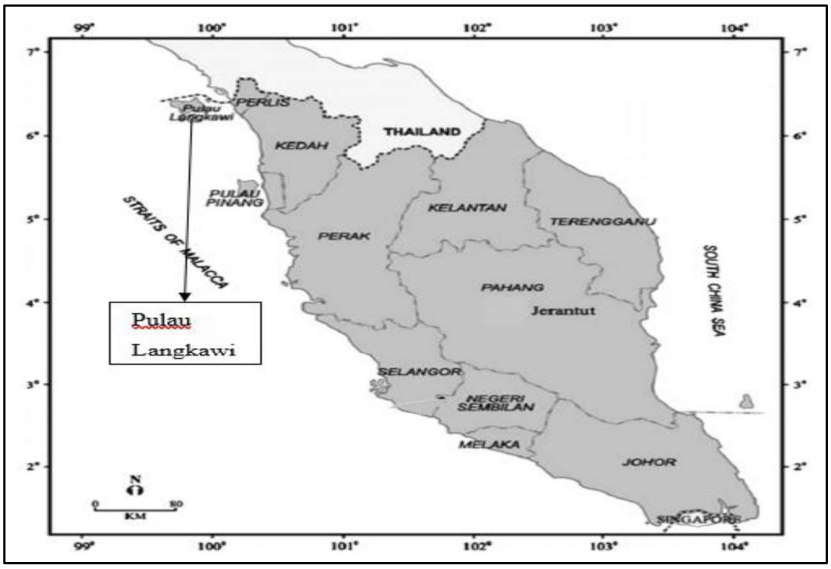

Figure. 1 Map of Langkawi Island [Ul-Saufie et. al., (2012)]

Parameters Selection

Based on the literature review of previous studies on factors affecting PM_{10} concentration, this study would utilize SO_2, CO, O_3, NO_2, relative humidity, temperature and wind speed as the parameters. Table 1 lists various researches on PM_{10} models using different parameters. Naturally, rain is the best way to reduce air pollution as it traps all pollutants and prevent them from spreading into the air. Since hourly rainfall data is unavaliable, relative humidity is used to replace rainfall parameters. Ul-Saufie et al., (2011) stated that relative humidity which is used to calculate the total percentage of water vapor occurring in gas mixtures and air could affect the concentration of PM_{10} when its value exceeded 55%.

Temperature and air pollution are closely associated. It is reported that temperature have the greatest effect on PM_{10}, with the most important components in PM_{10} are SO_2 and NO_2 (Ul-Saufie et al., 2011). This is so true especially when fires rapidly erupt and spread in dry or hot areas. However, during wet conditions where it is difficult to start a fire, a significant quantities of greenhouse gases and acid precipitation are released hence could also affect the environment (Shahraiyni & Sodoudi 2016).

Wind speed and wind directions are correlated with PM_{10} can cause change in PM_{10} concentrations. At a low wind speed, the concentration of PM_{10} would increase and transport the air pollutants through dispersion and re-suspension (Awang et al. (2000). Dispersion transport the pollutants horizontally; from one area to another while re-suspension would transport the pollutants vertically from lower ground to a higher level.

Table 1 Recent studies on PM_{10} variables forecasting in Malaysia

Author	**PM_{10}**	**O_3**	**SO_2**	**NO_2**	**CO**	**RH**	**T**	**WS**	**Others**
Sultan Alsultan et al. [13]	√	√	√	√	√				
Amnorzahira Amir [14]	√	√	√	√	√				Wind direction (WD)
Ahmad Zia Ul-Saufie et al. [15]	√		√	√	√		√	√	
Zainal Ahmad et al. [16]						√	√	√	WD
Samsuri Abdullah et al. [17]	√					√	√	√	Rainfall, mean sea level pressure
Ahmad Zia Ul-Saufie et al. [10]	√		√	√	√	√	√	√	
This study	√	√	√	√	√	√	√	√	

Besides PM_{10}, in terms of chemical parameters, the ground level O_3 was the other pollutant. Chemical reaction between Volatile Organic Compounds (VOCs) and nitrogen oxides (NOx) contribute on the formation of O_3 pollutants specifically in the presence of sunlight. The main sources of VOCs and NOx releases are motor vehicles and factories (DOE 2015). Table 1 summarized the variables used in forecating the PM_{10} in a recent studies conducted in Malaysia and the variables used in this study in order to close the gap

Multiple Linear Regression (MLR)

Multiple linear regression (MLR) determines the relationship between a dependent variable and numerous independent variables and it is widely practiced by researchers as the on of the air pollutants modelling methods. The general equation of MLR model can be written as Equation 1 (Ul-Saufie et al., 2012),

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon \qquad (1)$$

Where,

Y = Next 24 hours of PM_{10} concentration (µg/m^3)

x_1 = PM_{10} concentration (µg/m^3)

x_2 = CO concentration (ppm)

x_3 = O_3 concentration (ppm)

x_4 = NO_2 concentration (ppm)

x_5 = SO_2 concentration (ppm)

x_6 = Relative humidity (%)

x_7 = Temperature (^{o}C)

x_8 = Wind speed (m/s)

Random Forest (RF)

Random forest is an extension of a decision tree approach that uses an ensemble of trees to make a decision. A binary recursive categorizing algorithm is used in decision trees to generate pure nodes by splitting observations into two homologous classes (Grange et al., 2018). To achieve node purity, the splitting will be repeated due to the algorithm's recursive existence, and the entire sequence of splits, each with an explicitly called node or branch, is referred to as a tree. The recursive algorithm tends to have high accuracy in classifying the input data if the trees are allowed to expand to their full depth. The term "greedy" refers to this type of algorithm. This greedy behaviour can result in incredibly deep trees with only two observations tested at the final split. It's rarely used on new data that hasn't been used to train the model before. As a result, decision trees might be prone to overfitting (Kotsiantis 2013). This problem might be lessen by bagging decision trees from a training set. Numerous decision trees created from the bagged data will be used in unison when the model is utilized for prediction (Grange et al., 2018).

Bagging is the process of randomly sampling observations and replacing them with observations from the training collection, as well as sampling explanatory variables (Breiman 1996). The bagging results are referred to as out of bag (OOB), and they will often be lacking some input data points. If the method is repeated, a unique tree grown from OOB data may not contain any observations and variables used by other trees. For random forest models, it is common to surround hundreds of trees with OOB data, resulting in a forest that contains several decorrelated trees that have been trained on different subsets of the training set. Every single tree can be used for predictions to form a single prediction, and the mean of the predictions will be used in regression methods (Grange et al., 2018). Random forests are widely regarded as the best machine learning approach because they generate predictive models that generalize well and are more accurate (Wang et al., 2018).

Fig. 2 depicts how the dataset will be divided into two parts: 80 percent for model creation (training set) and 20% for model validation (testing set). For model validation, the data was divided into four sets, the entire data set, the other three sets were according to the Air Pollution Index (API) as summarized in Table 2. The types of air pollutants monitored in APIs are particulate matter less than or equal to 10 μm (PM_{10}), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and ozone (O_3), which is dangerous to human health and the environment. When the API value is between 50 and 100, it is considered to be good and moderate, which means that the ambient air is healthy and safe for people. If the API exceeds 100, it is known to be an unhealthy state for humans and the environment. When the API was around 300 and above the area was considered hazardous that posed a risk to human health. According to Shaziayani et al., (2021) PM_{10} and O_3 are the main causes of unhealthy days in Malaysia.

Table 2 Malaysian Ambient Air Quality Guidelines (MAAQG)

POLLUTANTS	AVERAGING TIME	AMBIENT AIR QUALITY STANDAR		
		IT-1 (2015) μgm^{-3}	IT-2 (2018) μgm^{-3}	Standard (2020) μgm^{-3}
PM10	1Year	50	45	40
	24 Hour	150	120	100
PM2.5	1Year	35	25	15
	24 Hour	75	50	35
SO2	1Hour	350	300	250
	24 Hour	105	90	80
NO2	1Hour	320	300	280
	24 Hour	75	75	70
O3	1Hour	200	200	180
	8 Hour	120	120	100
CO	1Hour	35	35	30
	8 Hour	10	10	10

Source: Department of Environment, 2018

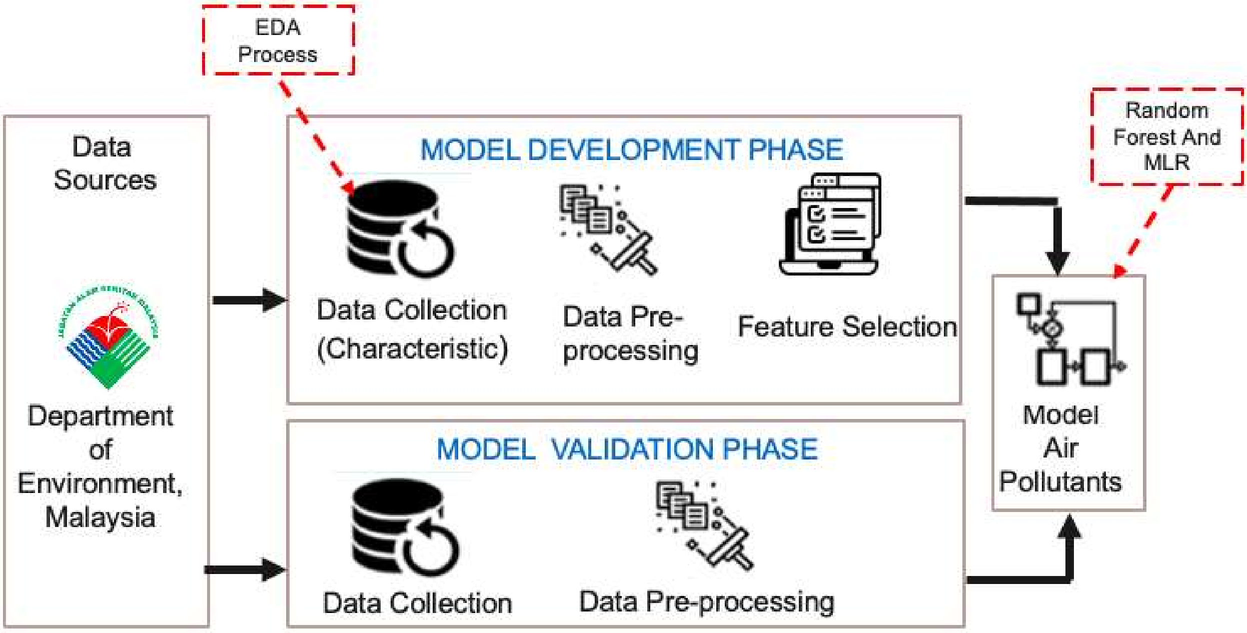

Figure 2 Workflow in predicting PM10 concentration using MLR and RF

Results and discussion

Long Term Record of Air Quality Data

The characteristics of the data on air quality in Langkawi Island are summarised in Table 3. Interestingly, relative to others, PM_{10} was discovered to be the most dominant air pollutant. The average PM_{10} concentration during the study period was 39.259±16.016µg/m^3. After PM_{10}, CO was found to be the predominant air pollutant (0.525 ppm), followed by O_3 (0.02 ppm), NO_2 (0.005 ppm) and SO_2 (0.001 ppm). The concentration of CO, SO_2, NO_2 were decreased year by year and did not exceed the limit by MAAQG due to the use of better fuel for motor vehicles (DOE 2015).

The data exhibited a highly skewed distribution based on skewness value for PM_{10} (5.51), CO (1.24), NO_2 (2.27) and SO_2 (2.02) which make extreme events that promote increased PM_{10} concentrations. The maximum levels in Langkawi Island exceeded the PM_{10} concentration threshold (150 µg/m^3) recorded at 386.94 µg/m^3.

The concentrations of PM_{10} showed exceedance of the standards due to extreme haze episodes. The August 2005 haze episode for Peninsular Malaysia, is considered more serious than 1997. The entire Klang Valley and surrounding areas were severely affected through a cloud of smoke (DOE 2015). In the middle of July, mid-August and late September to October 2006, Malaysia experienced mild to moderate haze episodes. A short duration of extreme haze episode was encountered out of 15 to 27 June 2013 due to transboundary emissions. The most impacted states were Johor, Melaka and Negeri Sembilan in Peninsular Malaysia. The haze resulted from forest fires in Sumatra and Kalimantan, Indonesia, causing Malaysia to experience a decline in air quality from August to September 2015 (DOE 2015).

Table 3 Overall data on air quality at Langkawi Island.

Parameters	Mean	Standard deviation	Max	Skewness
PM_{10} (µg/m^3)	39.259	16.016	386.94	5.506
CO (ppm)	0.525	0.207	3.2050	1.238
NO_2 (ppm)	0.005	0.004	0.0677	2.269
SO_2 (ppm)	0.001	0.001	0.0118	2.020
O_3 (ppm)	0.02	0.014	0.147	0.949

Bivariate Correlation Analysis

The Pearson correlation matrices of the variables for Langkawi Island are shown in Table 4. PM_{10} concentrations were positively correlated with CO, SO_2, NO_2, Ozone, temperature and wind speed and negatively correlated with RH. The highest positive correlation was obtained between PM_{10} and PM_{10T24} concentrations (0.772), as shown in the table. Moreover, due to the south-western monsoon season, the T and WS influenced the distribution of PM_{10} concentration within the vicinity. It was shown that there is a positive and strong link between CO and PM_{10}, which could be related to combustion processes, notably those emanating from motor vehicles. CO was shown to be associated with the concentration of PM_{10}, which is attributable to the creation of secondary pollutants, according to Dominick et al. (2012). Furthermore, Shukla and Sharma (2008) found that main precursor pollutants gases including SO_2, CO, NO, and (NH_3) can promote the generation of secondary particles PM_{10} in the atmosphere in the presence of moderate temperatures and high relative humidity.

Table 4 Pearson correlation between two parameters.

Parameters	**WS**	**O_3**	**CO**	**SO_2**	**PM_{10}**	**NO_2**	**T**	**RH**	**PM_{10T24}**
WS	1	0.534	-0.232	0.027	0.014	-0.125	0.445	-0.565	0.026
O_3		1	-0.135	-0.059	0.228	-0.096	0.620	-0.723	0.239
CO			1	0.088	0.249	0.155	-0.174	0.214	0.191
SO_2				1	0.083	0.058	0.004	0.101	0.070
PM_{10}					1	0.048	0.100	-0.159	0.772
NO_2						1	-0.055	0.048	0.042
T							1	-0.864	0.099
RH								1	-0.165
PM_{10T24}									1

Conversely, PM_{10} and RH concentrations were negatively correlated (-0.159). The inverse relationship between PM_{10} concentration and RH is due to the fact that high humidity is generally related to the number of rain events, which decreases the number of particles in the atmosphere due to the wash-out processes of atmospheric aerosols (Azmi et al., 2010).

PM_{10} Concentration Modelling.

Multiple Linear Regression (MLR)

The MLR model was developed for the next 24 hours by using SPSS software. Table 4 shows the model acquired for PM_{10} concentration predictions based on air pollutants and meteorological parameters. The value obtained for the Durbin-Watson is 2.008 which means that the model did not have any autocorrelation problem for the next day.

Table 4 Model summary of PM_{10} using MLR method.

Model	Durbin-Watson
$PM_{10+24hour} = 21.426 + 0.744\ PM_{10} + 76.154\ O_3 + 0.284\ CO + 250.382\ SO_2 + 7.727\ NO_2 – 0.12\ WS – 0.262\ T – 0.066\ RH$	2.008

In this analysis, the RMSE, NAE, IA and R^2 components were selected to measure the performance of the model. In order to assess model accuracy, RMSE and NAE are considered relevant indexes in which the model is noted to be highly accurate when its values are close to zero, whereas the IA and R^2 values are closest to 1 (Yuen et al., 2018). Table 5 summarizes quantitatively the performance of the model in terms of RMSE, NAE, IA and R^2 for this research by using the MLR method. The value of R^2 for overall data was 0.598416 and RMSE value was 8.069841. The four goodness of fit measures showed that the residual distributions were approximately normal, with zero means and no detectable serial and the correlation coefficients of the regressions were all highly statistically significant ($P< 0.01$).

The result of MLR for three categories which are PM_{10} concentration <50μg/m^3, 50μg/m^3≤PM_{10} concentration <100μg/m^3 and PM_{10} concentration ≥100μg/m^3 presented in table 5. These three categories represent good, moderate, and unhealthy air quality respectively. The results show that the PM_{10} concentration<50μg/m3 had a lower value of RMSE at 5.962807 compared to 50μg/m^3≤PM_{10} concentration<100μg/m^3 and PM_{10} concentration≥100μg/m3 at 12.120965 and 92.465131 respectively. Furthermore, the value of R^2 for overall data was 0.598416 where the accuracy measures should be close to 1 to get a good model of PM_{10} concentration prediction model. This figure is close to the value of R^2 for the PM_{10} concentration<50μg/m^3 which is 0.548347. However, the accuracy value for the three categories showed a decrease in the value of R^2. Moreover, comparison of the overall model with the categorical models for PM_{10} revealed that the former provided better explanation than the latter.

Table 5 Results of performances indicator by using MLR method.

Performances indicator	MLR			
	Overall	PM_{10} concentration $<50\mu g/m^3$	$50\mu g/m^3 \leq PM_{10}$ concentration$<100\mu g/m^3$	PM_{10} concentration $\geq 100\mu g/m^3$
RMSE	8.069841	5.962807	12.120965	92.465131
NAE	0.136767	0.126743	0.158208	0.570136
IA	0.858433	0.844791	0.622770	0.369767
R^2	0.598416	0.548347	0.241491	0.101908

Random Forest (RF)

The RF model was developed for the next 24 hour by using Rstudio software. Table 6 shows the results of the performance indicator by using the RF method. The value of R^2 for overall data was 0.726681 and RMSE value was 6.674038. The result of RMSE value for three AIP categories models 5.076431, 9.078162 and 86.366315 respectively. These figures showed an increment in error measures especially for high PM_{10} concentration. Based on the result of performance indicator, value of RMSE and NAE showed an increasing trend, whereas the value of IA and R^2 showed decreasing trend value.

Figure 3 shows the comparison results of the performance indicator in terms of IA and R^2 using MLR and RF method. For overall data, it can be observed that the accuracy value of R^2 and IA for RF which are 0.726681 and 0.911974 respectively is higher than MLR as it is much closer to 1. In other words, if the value of the accuracy is greater than 0.8 it is indicating that the predicted values are highly accurate and this model is good as the value is closer to 1. This finding supports evidence from previous observations (Xuefei et al., (2017); Brokamp et al. (2017); Poggi & Portier (2011)). The results revealed that the RF method gives the best ultimate accuracy figures for overall data better than MLR in predicting PM_{10} concentration in Langkawi Malaysia.

Table 6 Results of performances indicator by using RF method.

Performances indicator	RF			
	Overall	PM_{10} concentration $<50\mu g/m^3$	$50\mu g/m^3 \leq PM_{10}$ concentration$<100\mu g/m^3$	PM_{10} concentration $\geq 100\mu g/m^3$
RMSE	6.674038	5.076431	9.078162	86.366315
NAE	0.107664	0.104366	0.110127	0.513527
IA	0.911974	0.898547	0.747708	0.364826
R^2	0.726681	0.674064	0.437509	0.164461

One further interesting observation is revealed when comparing the performance indicator in Fig. 3. The results showed that PM_{10} concentration$<50\mu g/m^3$ (good category) yielded the highest IA (0.898547) and R^2 value (0.674064) for RF method compared to the moderate and unhealthy category. Furthermore, it can be clearly seen that the accuracy value (IA and R^2) is declining for the moderate and unhealthy category, then contributes an equivalent trend for the MLR method. Based on the results from accuracy measures, it can be clarified that both RF and MLR methods provided better PM_{10} concentration forecasting capability for PM_{10} concentration$<50\mu g/m^3$. These imply that both RF and MLR methods are not suitable for data with extreme value.

Some previous field studies for PM_{10} prediction in the world are summarized in Table 7. It is evident that daily predicted PM_{10} concentration using RF is better than MLR. Furthermore, the PM_{10} models (RF and MLR) are unable to take into account the higher observed concentrations.

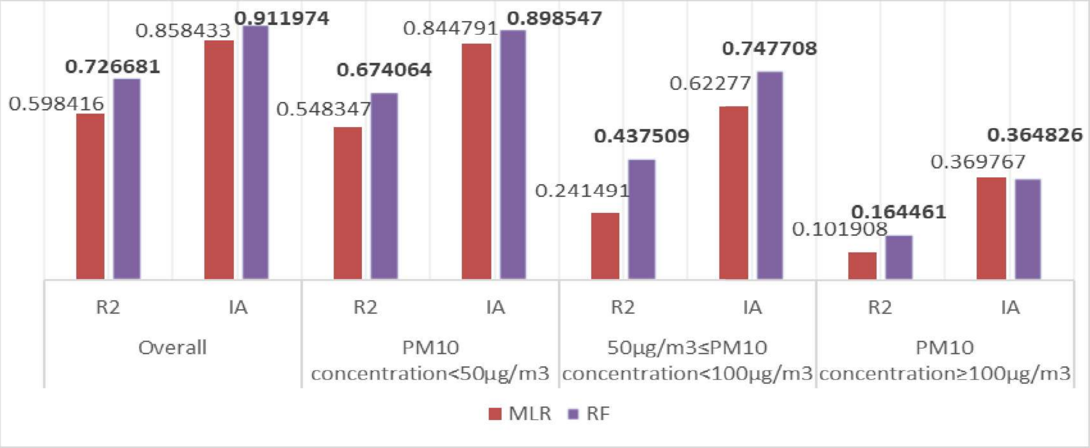

Figure 3 Results of performances indicator by using MLR and RF method

Table 7 PM_{10} pollution modelling in worldwide.

Author	**Time Series**	**Places**	**Method**	**PM_{10} parameter**	**Results**
Stadlober et al. (2008)	2001-2006	Italy	Multiple Linear Regression (MLR)	Daily	R^2 = 0.55
Cai et al. (2009)	2007	China	Multi-Layer Perceptron (MLP)	Hourly	MAPE=12.9% MAE=15.5 RMSE=20.1 R= 0.961
Cai et al. (2009)	2007	China	Multiple Linear Regression (MLR)	Hourly	R= 0.971
Ul-Saufie et al. (2012)	2004-2007	Pulau Pinang, Malaysia	MLR and Artificial Neural Network (ANN)	Hourly	R^2 = 0.942 R= 0.946
Abdullah et al. (2017)	2005-2011	Terengganu, Malaysia	Multiple Linear Regression (MLR)	Monthly	R^2 = 0.68, 0.58, 0.57, 0.63
Ponggi et al. (2011)	2011	France	Random Forest	Daily	R= 0.78 RMSE= 6.34 IA= 0.86
Brokamp et al. (2017)	2017	Ohio	Random Forest	Hourly	R^2= 0.7
Shamsoddini (2017)	2017	Tehran	Random Forest	Daily	R^2 = 0.42 RMSE= 19.85
Hu et al. (2017)	2017	United states	Random Forest	Daily	R^2= 0.80 RMSE= 2.83
Kaminska (2018)	2018	Wroclaw	Random Forest	Hourly	R^2 = 0.444
Ivanov et al. (2018)	2018	Bulgaria	Random Forest	Daily	R^2= 0.932 RMSE= 9.6162 MAPE= 0.1946

Conclusion

In conclusion, the average air pollutant concentrations on Langkawi Island have not yet exceeded the standard MAAQG limit. In Langkawi Island, PM_{10} turned out to be the dominant type of air pollutants, followed by CO, O_3, NO_2 and SO_2 where hot weather conditions contributed to such PM_{10} concentration emission levels. Moreover, due to the south-western monsoon season, the wind speed and temperature influenced the distribution of PM_{10} concentration within the vicinity. The results of such evaluations are essential for protecting human health as well as the environment.

Next, the model created in this project is much better than other methods, thus, the second objective was achieved. RF method has been chosen as a good model than MLR because the error measures is close to 0 while the accuracy measures results are close to 1 where the value of RMSE, NAE, IA, PA and R^2 is 6.674038, 0.107664, 0.911974, 0.852570 and 0.726681 respectively. This showed that the model produced was good and will be much more accurate if using random forests method. However, the results showed an in error measure increment in error measures e. Based on the result of the performance indicator, the value of RMSE and NAE showed an increasing trend, whereas the value of IA and R^2 showed decreasing trend value especially for extreme value. For further analysis, it can be proposed that a researcher needs to concentrate on predicting PM_{10} concentrations for haze events. It may assist policymakers in the relevant field to prepare adequate steps to curb the occurrence of extreme concentrations of PM_{10} and ultimately reduce the impact on human health.

ACKNOWLEDGEMENT

Thank you to Universiti Teknologi MARA for their support and also thanks to the Department of Environment Malaysia for providing air quality monitoring data.

References

Abdullah S, Ismail M, Fong SY (2017). Multiple Linear Regression (MLR) models for long term Pm10 concentration forecasting during different monsoon seasons. J. Sustain. Sci. Manag. 12:60–69.

Abdullah S, Ismail M, Samat NNA, Ahmed AN (2018). Modelling Particulate Matter (PM10) Concentration in Industrialized Area: A Comparative Study of Linear and Nonlinear Algorithms. ARPN J. Eng. Appl. Sci. 13:8227–8235.

Afroz R, Hassan MN, Ibrahim NA (2003). Review of air pollution and health impacts in Malaysia. Environmental Research 92:71–77. doi:10.1016/S0013-9351(02)00059-2

Awang MB (2000). Air quality in Malaysia: impacts, management issues and future challenges. J Respirology 5:183–96.

Azid A, Juahir H, Latif MT, Zain SM, Osman MR (2013). Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia. Journal of Environmental Protection 4:1–10. doi:10.4236/jep.2013.412a1001

Azmi, SZ, Latif MT, Ismail AS, Juneng L, Jemain AA (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. Air Quality, Atmosphere and Health 3:53–64. doi:10.1007/s11869-009-0051-1

Bai L, Wang J, Ma X, Lu H (2018). Air Pollution Forecasts: An Overview. Int J Environ Res Public Health 15:780. doi: 10.3390/ijerph15040780.

Breiman, L. (1996). Bagging predictions. Machine Learning 24:123–140.

Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P (2017). Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmos. Environ. 151:1–11.

Cai M, Yin Y, Xie M (2009). Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. Transportation Research Part D. Transport and Environment 14:32–41. doi:10.1016/j.trd.2008.10.004

Maher Elbayoumi, Suheir Harb (2018). Prediction of Hourly Indoor Carbon Monoxide Concentrations by Using Multivariate Methods with Sensitivity Analysis Technique, Israa University Journal of Applied Science, no.2 :43-54

DOE. (2015). Department of Environment, Malaysia. https://www.doe.gov.my/portalv1/en/info-umum/info-kualiti-udara/kronologi-episod-jerebu-di-malaysia/319123

DOE (2018) Department of Environment, Malaysia. Malaysia Environmental Quality Report 2018. Kuala Lumpur: Ministry of Energy, Science, Technology, Environment and Climate Change, Malaysia.

Dominick, D, Juahir H, Latif MT, Zain SM, Aris AZ (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmospheric Environment 60:172-181

Grange SK, Carslaw DC, Lewis AC, Boleti E, Hueglin C (2018). Random forest meteorological normalisation models for Swiss PM 10 trend analysis. Atmos. Chem. Phys 18:6223–6239, 2018.

Hrust L, Klaić ZB, Križan J, Antonić O, Hercog P (2009). Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. Atmos. Environ 43:5588–5596.

Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickl, MJ, Liu Y (2017), Estimating $PM_{2.5}$ Concentrations in the Conterminous United States Using the Random Forest Approach, Environ. Sci. Technol 51:6936–6944.

Ivanov A, Voynikova D, Stoimenova M, Gocheva-Ilieva S, Iliev I, (2018). Random forests models of particulate matter PM10: A case study.

Kamińska JA (2018), The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław, J. Environ. Manage 217:164–174.

Kotsiantis SB (2013). Decision trees: A recent overview. Artif. Intell. Rev. 39:261–283.

Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, Niska H, Dorling S, Chatterton T, Foxall R, Cawley G (2003). Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment 37:4539–4550. doi:10.1016/S1352-2310(03)00583-1

Ku Yusof KMK, Azid A, Samsudin MS, Jamalani MA (2017). An overview of transboundary haze studies: The underlying causes and regional disputes on Southeast Asia Region. Malaysian. Journal of Fundamental and Applied Sciences 13:747–753. doi:10.11113/mjfas.v0n0.719

Pan B, (2018). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conference Series. Earth and Environmental Science 113. doi:10.1088/1755-1315/113/1/012127

Paschalidou AK, Karakitsios S, Kleanthous S, Kassomenos PA (2011). Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: Implications to local environmental management. Environmental Science and Pollution Research 18:316-327.

Poggi JM and Portier B (2011). PM10 forecasting using clusterwise regression. Atmos. Environ 45:7005–7014.

Rossita MY, Masud MH, (2017). Predicting Hourly PM10 Concentration in Seberang Perai and Petaling Jaya Using Log-Normal Linear Model. International Journal of Management and Applied Science (IJMAS) 3:103-108.

Sayegh AS, Munir S, Habeebullah TM (2014). Comparing the Performance of Statistical Models for Predicting PM10 Concentrations. Aerosol and Air Quality Research. 14:653-665. doi:10.4209/aaqr.2013.07.0259.

Shahraiyni HT, Sodoudi S (2016). Statistical Modeling Approaches for PM10 Prediction in Urban Areas; A Review of 21st-Century Studies. Atmosphere. 7. doi:10.3390/atmos7020015

Shamsoddini A, Aboodi MR, Karami J, (2017). Tehran Air Pollutants Prediction Based On Random Forest Feature Selection Method. Tehran's Joint ISPRS Conferences of GI Research, SMPR and EOEC 41.

Shaziayani, WN, Harun FD, Ul-Saufie AZ (2021). Three-Days Ahead Prediction Of Daily Maximum. International Journal Of Conservation Science 12:217–224.

Shaziayani WN, Ul-Saufie AZ, Yusoff SAM, Ahmat H, Libasin Z, (2021). Evaluation of boosted regression tree for the prediction of the maximum 24-hour concentration of particulate matter. International Journal of Environmental Science and Development, 12:126–130. doi:10.18178/IJESD.2021.12.4.1329

Shukla SP, Sharma M, (2008). Source apportionment of atmospheric PM_{10} in Kanpur, India. Environmental Engineering Science 25:849-862.

Slini T, Kaprara A, Karatzas K, Moussiopoulos N, (2006). PM10 forecasting for Thessaloniki, Greece. Environ. Model. Softw 21:559–565.

Stadlober E, Hörmann S, Pfeiler B, (2008). Quality and performance of a PM10 daily forecasting model. Atmospheric Environment, 42:1098-1109.

Tourism Official Portal of Langkawi Municipal Council (MPLBP) (2018). LANGKAWI MUNICIPAL COUNCIL http://www.mplbp.gov.my/en/citizens/services/tourism/page/0/1.

Ul-Saufie AZ, Yahaya AS, Ramli A, Hamid HA, (2011). Comparison Between Multiple Linear Regression And Feed forward Back propagation Neural Network Models For Predicting PM 10 Concentration Level Based On Gaseous And Meteorological Parameters. International Journal of Applied Science and Technology 1:42–49.

Ul-Saufie AZ, Yahaya AS, Ramli A, Hamid HA, (2012). Future PM10 Concentration Prediction Using Quantile Regression Models. Ipcbee 37:15–19.

Ul-Saufie AZ, Yahaya AS, Ramli A, Hamid HA (2012). Performance of multiple linear regression model for long-term PM_{10} concentration prediction based on gaseous and meteorological parameters. Journal of Applied Sciences 12:1488–1494.

Wang Z. Wang Y, Zeng R, Srinivasan RS, Ahrentzen S (2018). Random Forest Based Hourly Building Energy Prediction. Energy Build. 171:11–25.

Wen YS, Fauzan A, Nabila N, Sulaiman Z (2016). Transboundary Air Pollution in Malaysia : Impact and Perspective on Haze. Nova Journal of Engineering and Applied Sciences 5:1–11.

Yuen FS, Abdullah S, Ismail M (2018). Forecasting of Particulate Matter (PM10) Concentration based on Gaseous Pollutants and Meteorological Factors for Different Monsoons of Urban Coastal Area in Terengganu. J. Sustain. Sci. Manag. 13:3–17.