IJCS INTERNATIONAL JOURNAL OF CONSERVATION SCIENCE
ISSN: 2067-533X Volume 12, Issue 1, January-March 2021: 217-224 www.ijcs.ro

THREE-DAYS AHEAD PREDICTION OF DAILY MAXIMUM CONCENTRATIONS OF PM_{10} USING DECISION TREE APPROACH

Wan Nur SHAZIAYANI[1] , Firhad Dhiyafique HARUN[3] , Ahmad Zia UL-SAUFIE[2,4*], Norshuhada SAMSUDIN[1], Norazian Mohamed NOOR[4]

[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, Malaysia.
[3]Faculty of Chemical Engineering, Universiti Teknologi MARA, 13500 Pulau Pinang, Malaysia.
[4]Sustainable Environment Research Group (SERG), Centre of Excellence Geopolymer and Green Technology (CEGeoGTech), School of Environmental Engineering, Universiti Malaysia Perlis (UniMAP), Kompleks Pusat Pengajian Jejawi 3,02600 Arau, Perlis, Malaysia.

Abstract

The air pollution in Malaysia is always fluctuated throughout the year. This is corresponding to the growth of industrial area, emission from vehicle and transboundary haze. These episodes have indirectly had a major role on Malaysia's air quality. Therefore, this analysis aims to analyzed the pattern and predict the three days ahead of maximum daily PM_{10} concentration. The station is located at Jerantut, Pahang as it is significant to the centre of Peninsular Malaysia. As for the trend, air pollution maximum daily monitoring records from 2004 until 2017 were used in analysing the statistical data analysis. This research has selected eight parameters which are known as SO_2, CO, PM_{10}, NO_2, O_3, temperature (T), relative humidity (RH), and wind speed (WS). Decision Tree (DT) was used to predict PM_{10} concentrations. The results obtained for performance indicators such as Root Mean Squared Error (RMSE) are 10.164, 13.853, and 13.281 respectively for day 1, day 2 and day 3. As for Coefficient of Determination (R^2), the results obtained are 0.767, 0.530, and 0.510 respectively for day 1, day 2 and day 3. Others error measurement for this study are Absolute Error (AE), Relative Error Lenient (REL) and Squared Error (SE). (AE = 5.893, REL = 11.95%, SE =108.640) and next 2-day (AE = 8.268, REL = 16.21%, SE =197.565) and next 3-day (AE = 9.052, REL = 17.92%, SE =177.699). Assessment of model performance indicated that DT method can be used for short term PM_{10} concentration prediction for three-days ahead.

Keywords*: PM_{10}, Decision Tree, Root Mean Squared Error, Absolute Error, Performance Indcators*

Introduction

Air pollutions are becoming a serious major factor in human difficulty in breathing. It affects millions of people around the world. This observation is supported by World Health Organization (WHO). It is estimated that 7 million people die every year because of the effect of air pollution [1]. It is widely recognised that implementation of pollution reduction strategies can bring significant health benefits. For example, the Environmental Protection Agency has suggested that measures to reduce pollution from diesel engines may result in 12 000 fewer deaths and prevent 15 000 heart attacks and 8 900 hospital admissions per year in the United States [2]. This is because it allows the person predicts more accurately. The aim of air quality

* Corresponding author: ahmadzia101@uitm.edu.my

monitoring is to evaluate the quality of the ambient air, and if there is a drastic change in the level of air quality, the public should be advised to prepare for the situation [3].

The Decision Tree (DT) has two key elements, such as the problem statement represented by the tree root, and a set of implications or solutions represented by the tree branches and it can expand all the possibilities of a problem statement to any length [4]. A main distinction between real trees and DT is that the root at the top is usually an inverted tree. DT is an alternative method use to predict PM_{10}. Hence, it is significant to use DT methods as a prediction of PM_{10}. A user can visualize each step, which also much easier making a decision based on regression than most other methods. Since most irrelevant data is filtered from each level, less data needs to be worked on to go further into the tree [5].

The aim of this study was to examine the performance of the Decision Tree method in predicting future concentration levels of PM_{10} in Jerantut Pahang, Malaysia for three days ahead. Furthermore, this study also investigates the attribute of each component towards the prediction of PM_{10} concentration.

Methodology

Data Preparation

Data preparation is the first process of working with a predictive model for PM_{10} concentration use. It includes three main steps, i.e. step 1 data acquisitions which is maximum daily data of air quality monitoring in Jerantut, Pahang Malaysia was acquired from Department of Environment (DOE) Malaysia. The duration of data collection were from 2004 until 2017. The location of sampling stations are chosen at air monitoring station ((N03° 58.238', E102° 20.863') which is at Jerantut, Pahang. This monitoring station is in fact a background station established by the Malaysian Department of Environment and located at the Malaysian Meterological Deparment at Batu Embun, Jerantut, Pahang in the middle of the Malaysian Peninsular. Step 2 involved data exploration and data observation for descriptive information and looking at abnormal data including extreme values and missing values. The third step is data cleaning for treating missing and outlier values. The air pollutant parameters used in this study were PM_{10}, CO, SO_2, NO_2, O_3 and several meteorogical factors which known as relative humidity, wind speed and temperature. According to [6] DOE developed national standards for air quality for each of these air pollutants to protect the public health.

Data Pre-Processing

Data pre-processing is the preparation of the data before developing the models. This process also known as data management. It involves three steps i.e. step 1 is feature selection. The selection in this research is based on the previous study. Correlation analysis to determine the structure in the relationships between parameters Step 2 is feature scaling. This step will be used to standardize the values of inputs and output data to avoid noise from different unit scales. Step 3 is partitioning the dataset to 80% of data is segregated into training while the rest (20%) is used for validation and testing.

Model Development

The DT model is a nonparametric approach which can be used for quantitative and qualitative predictors. A type DT stands for absolute and indirect associations between various independent variables and a dependent variable, with a tree structure and with a reciprocal classification of data [7]. The DT has a major advantage compared to others techniques of modelling that create a model which can reflect interpretable rules, or statements of logic. The capability of clarification for axis-producing trees essential function is parallel decision surfaces [8].

Model Evaluation

This study used the performance indicator to evaluate the performance models. There are five performance indicators which are the Coefficient of Determination (R^2), Root Mean

Squared Error (RMSE), Absolute Error (AE), Relative Error Lenient (REL) and Squared Error (SE) between decision tree predicted data and the observed data are used for evaluation of the quality of develop model.

Results and discussion

The statistics of collective data that varies from Wind Speed (WS), Temperature (T), Relative Humidity (RH), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Ozone (O_3), Carbon Monoxide (CO) and Particulate Matter under 10 micrometers (PM_{10}) throughout the whole set of data from 2004 until 2017 are shown in Table 1.

Table 1. Descriptive statistics for daily average gaseous and meteorological parameters in Jerantut, Pahang

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Ws	115101	1	17	3.59	2.044	0.722	0.007
Temp	100384	19	41	26.48	3.649	0.733	0.008
Humidity	107980	24	100	83.63	15.715	-0.890	0.007
SO_2	102774	0	0	0.00	0.001	1.199	0.008
NO_2	107158	0	0	0.00	0.001	2.117	0.007
O_3	108508	0	0	0.01	0.012	1.154	0.007
CO	112998	0	3	0.31	0.156	2.794	0.007
PM_{10}	114890	9	298	38.55	17.864	2.956	0.007
Valid N (listwise)	76892						

The box plot data in Figure 1 shows that in the year of 2015 was the highest number of PM_{10} concentration. The [9] reported that in 2015, from August to September Malaysia experienced air pollution problems during the Southwest Monsoon due to large scale land and forest fires in Sumatra and Kalimantan, Indonesia. On 15 September 2015, 34 areas in the country for the first time in the history of Malaysia since 1997 reported unhealthy air quality status. Because of the API reading to 200, all schools in the states of Putrajaya, Kuala Lumpur, Selangor, Negeri Sembilan and Melaka were closed on September 15, 2015 while all schools in the divisions of Kuching and Samarahan, Sarawak were closed on September 18, 2015. On 14 September 2015, the highest API reading was 211 (very unhealthy), in Banting, Selangor [10].This shows that the Air Pollution Index reading during the year of 2015 is correlated with the PM_{10} concentration [11].

The [9] also found that cross-border emissions made Malaysia experience extreme haze episodes between 15 and 27 June 2013. A large number of areas in Peninsular Malaysia were severely affected. As far as Peninsular Malaysia is concerned, the August 2005 haze episode was considered to be more extreme than the previous episode in 1997 when the entire part of Klang Valley and its surrounding areas were badly affected by smoke haze.

The monthly PM_{10} concentrations data in Jerantut, Pahang are shown in Figure 2. The value of PM_{10} concentration is fluctuate between the range of 38. There are months that spike which is from June until September. This may be due to Wind direction that affect from outside Malaysia. The southwest monsoon winds that approaching Malaysia is in between May until September. As the the festive of open burning activities which are likely to occur during May until September from neighbour country [12].

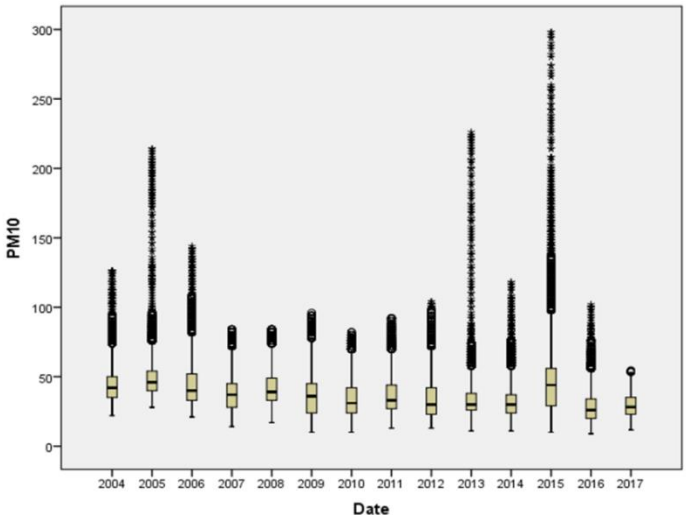

Fig. 1. Box plot for PM_{10} concentrations vs Year

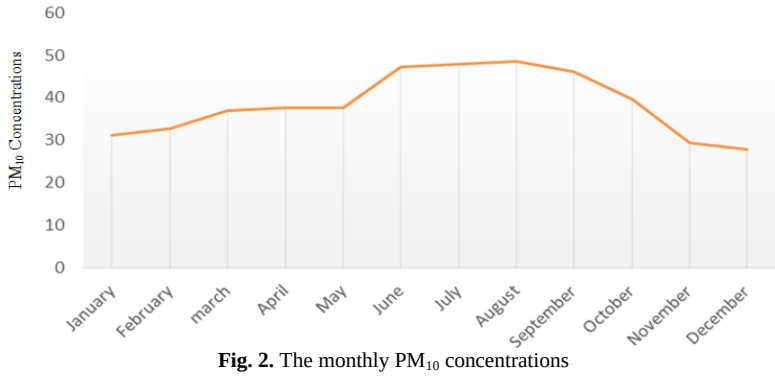

Fig. 2. The monthly PM_{10} concentrations

The results of the attribute of each component towards the prediction of PM_{10} concentrations within the whole 13 years are given in Figure 3. It shows that PM_{10} has the highest weightage and holds the majority component regarding the prediction of the next day of PM_{10} concentrations.

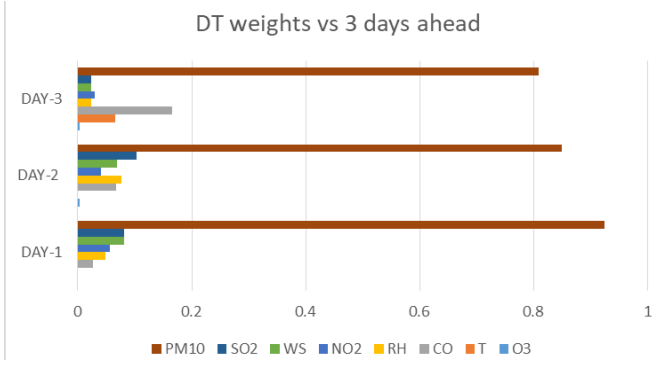

Fig. 3. The Decision tree weights for three days ahead

The predictions value and true value are being scattered among the charts in Figures 4.

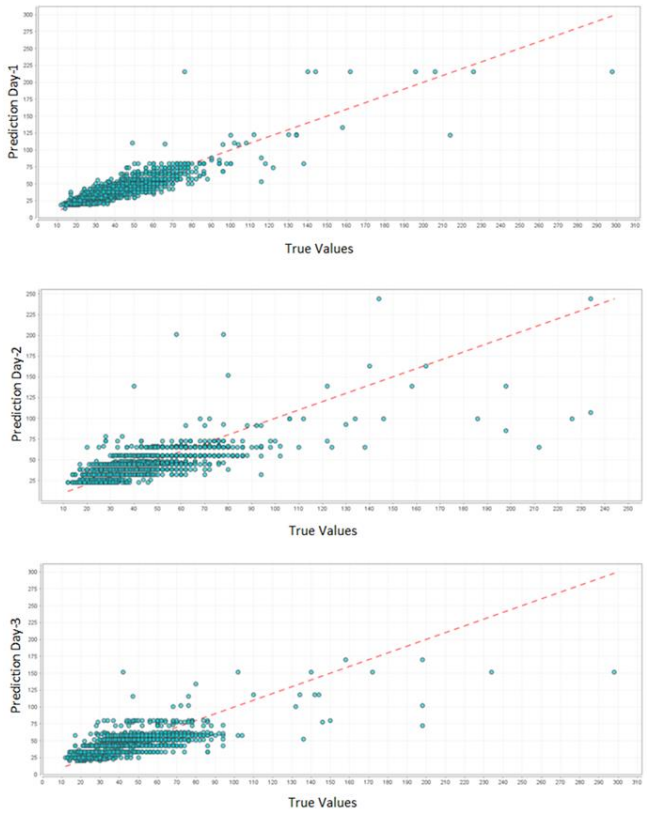

Fig. 4. Decision tree prediction chart

The results show that it is a linear graph which is made the results of the prediction is almost accurate following the mathematical theory. The correlation of this values is scattered and hits differently between 3 days. This is can be impacted to the value of squared error. The lesser the scattered points towards linear line, the lower the value of the squared error.

Table 2 shows for day 1, the RMSE results are 10.164. For absolute error, the value gains are 5.893 with the performance error is approaching to zero. This showed that the error is acceptable as it comply with basic statistic philosophy which is error gain must be approaching to zero. As for relative error lenient, it showed that 11.95%, 16.21%, and 17.92% respectively for day 1, day 2 and day 3. The predicting value for day 1, day 2, and day 3 is almost and approaching the true values. The R^2 of the data obtained is 0.767 which the number is approaching to 1. This showed that the the accuracy of the model for first day is better than second and third day.

Performance indicators were used to compare between the models to indicate the models' performance as shown in Tables 3. RMSE was used to find the error of the model, where a value closer to 0 indicated a better model. R^2 was used to check the accuracy of the model result, where a higher accuracy is given by a value closer to 1. The R^2 of PCR and MLR models show strong correlations between predicted and observed concentrations. In addition, the values of the RMSE are smaller for DT model compared with other models. It showed that this DT model can be used to predict PM_{10} concentrations since it improved the accuracy of the models (R^2) and reduced error (RMSE).

Table 2. Criteria and the results gain after performing DT

Criterion	Day- 1	Day-2	Day-3
Root mean squared error (RMSE)	10.164	13.853	13.281
Absolute error (AE)	5.893	8.268	9.052
Relative error lenient (REL)	11.95%	16.21%	17.92%
Squared error (SE)	108.640	197.565	177.699
R^2	0.767	0.530	0.510

Table 3. Criteria and the results gain from other methods

Authors	Year	Method	Accuracy
[13]	2017	Multiple Linear Regression (MLR)	RMSE=27.99-35.71 R^2=0.58-0.72
[14]	2018	Multiple Linear Regression (MLR)	RMSE=21.63 R^2=0.609
[15]	2018	Principle Component Regression (PCR)	RMSE=11.65-19.56 R^2=0.39-0.66
[16]	2018	Principle Component Regression (PCR)	RMSE=11.39-35.71 R^2=0.45-0.72
[17]	2013	Feed-Forward Artificial Neural Network Model (FFANN)	RMSE=11.16-11.49 R^2=0.36-0.40
[18]	2013	Principle Component Regression (PCR)	RMSE=11.34-18.27 R^2=0.60-0.76

Conclusions

DT is an alternative method to predict for PM_{10} concentrations in Jerantut, Malaysia. It can be predicted within 3 days advanced. The result showed that the accuracy of the model for first day is better than second and third day. In this prediction, PM_{10} concentration is the parameter that contributes most to obtaining the good accuracy.

The quality and reliability of the developed models were evaluated via performance indicators (RMSE, AE, REL, SE and R^2). Assessment of model performance indicated that DT model has successfully predicted PM_{10} concentrations.

Acknowledgments

The authors would like to extend their appreciation to the Universiti Teknologi MARA and Department of Environmental Malaysia (DoE) for providing air quality monitoring data. The research was funded by 600-IRMI/FRGS 5/3 (289/2019).

References

[1] W. Roberts, *Air pollution and skin disorders.* **International Journal of Women's Dermatology**, **7**(1), 2021, pp. 91–97.

[2] P.A.M. Overgaauw, C.M. Vinke, M.AE. Van Hagen, L.J.A. Lipman, *A One Health Perspective on the Human-Companion Animal Relationship with Emphasis on Zoonotic Aspects,* **International Journal of Environmental Research Public Health, 17**(11), 2020. pp. 1-30.

[3] F.F. Sukatis, N.M. Noor, N.A. Zakaria, A.Z Ul-Saufie, A. Suwardi, *Estimation Of Missing Values In Air Pollution Dataset By Using Various Imputation Methods*, **International Journal of Conservation Science**, **10**(4), 2019, pp. 791–804.

[4] J.R. Quinlan, *Simplifying Decision Trees***, International Journal of Human-Computer Studies, 51(**2), 1999, pp. 497-510 .

[5] C. De Stefano, G. Lando, C. Malegori, P. Oliveri, S. Sammartano, *Prediction of water solubility and Setschenow coefficients by tree-based regression strategies*, **Journal of Molecular Liquids, 282,** 2019, pp. 401-406.

[6] M.M. Kamal, J. Rozita, S R. L. A. Shauri, P*rediction of ambient air quality based on neural network technique. **SCOReD 2006,* Proceedings of 2006 4th Student Conference on Research and Development "Towards Enhancing Research Excellence in the Region",** 2006, pp. 115–119.

[7] E. Boroujeni, M.S. Shamsabadi, H. Shirani, Z. Mosleh, M.B. Bodaghabadi, M.H. Salehi, *Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations*, **Catena, 191**, 2020.

[8] G.K.F. Tso, K.K.W. Yau, *Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks*, **Energy, 32**(9), 2007, pp. 1761–1768.

[9] DOE. *Kronologi episode jerebu di Malaysia*, **Department of Environment, Malaysia,** 2015**.**

[10] M. Awang, A.B. Jaafar, A.M. Abdullah, M. Ismail, M.N. Hassan, R. Abdullah, S. Johan, H. Noor, *Air quality in Malaysia: Impacts, management issues and future challenges*, **Respirology**, **5,** 2000, pp. 183–196.

[11] A.Z. Ul-Saufie, A.S. Yahaya, A. Ramli, H.A. Hamid, *Future PM_{10} Concentration Prediction Using Quantile Regression Models*, **Ipcbee**, **37**. 2012, pp. 15–19.

[12] S.Z. Azmi, M.T. Latif, A.S. Ismail, L. Juneng, A.A. Jemain, *Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia, Air Qual*, **Atmospheric Healt, 3**, 2010, pp. 53–64.

[13] S. Abdullah, M. Ismail, S.Y. Fong, *Multiple Linear Regression (MLR) models for long term PM_{10} concentration forecasting during different monsoon seasons*, **Journal of Sustainability Science and Management, 12**(1), 2017, pp. 60–69.

[14] W.N. Shaziayani, A.Z. Ul-saufie, H, Ahmat, *A 24-Hour Forecasting of PM_{10} Concentration in Urban Area*, **Proceedings of the International Conference on Mathematics, Engineering and Industrial Applications 2018 (ICOMEIA 2018) Book Series: AIP Conference Proceedings,** 2018**, 2013** .

[15] M. Ismail, S. Abdullah, A.D, Jaafar, T.A.E. Ibrahim, M.S.M. Shukor, *Statistical modeling approaches for PM10 forecasting at industrial areas of Malaysia*, **Advances in Civil Engineering and Science Technology, Book Series: AIP Conference Proceedings, 2020,** 2018, doi 10.1063/1.5062670.

[16] S.Y. Fong, S. Abdullah, M. Ismail, *Forecasting of Particulate Matter (PM_{10}) Concentration Based on Gaseous Pollutants and Meteorological Factors for Different Monsoons of Urban Coastal Area in Terengganu*, **Journal of Sustainability Science and Management, 74**(5), 2018, pp. 637–642.

[17] A. Azid, H. Juahir, M.T. Latif, S.M. Zain, M.R. Osman, *Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia***, Journal of Environmental Protection, 4,** 2013, pp. 1-10.

[18] A.Z. Ul-Saufie, A.S. Yahaya, N.A. Ramli, N. Rosaida, H.A. Hamid, *Future daily PM_{10} concentrations prediction by combining regression models and feedforward*

backpropagation models with principle component analysis (PCA), **Atmospheric Environment**, **77**, 2013, pp. 621–630.

Received: September 02, 2020
Accepted: February 24, 2021