European Proceedings of
Social and Behavioural Sciences
EpSBS
www.europeanproceedings.com
e-ISSN: 2357-1330

DOI: 10.15405/epsbs.2020.03.03.92

ICMR 2019
8th International Conference on Multidisciplinary Research

PREDICTION OF PM_{10} CONCENTRATIONS USING LOGISTIC REGRESSION ANALYSIS: CASE STUDY IN JERANTUT

Ahmad Shukri Yahaya (a), Hazrul Abdul Hamid (b)*, Mohamad Hazlami Abdul Hamid (a)
*Corresponding author

(a) School of Civil Engineering, Universiti Sains Malaysia, Nibong Tebal, Penang, Malaysia, ceshukri@usm.my
(b) School of Distance Education, Universiti Sains Malaysia, Penang, Malaysia, hazrul@usm.my

Abstract

Particulate matter (PM_{10}) can cause several serious negative health effects to humans when it is present in the environment. Thus, it is important for us to forecast its concentration levels in the environment so that we can reduce the risk of exposure towards particulate matter. Secondary data on the concentration of PM_{10}, sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ground level ozone (O_3), carbon monoxide (CO) along with temperature and relative humidity at Jerantut monitoring stations between 2010 to 2012 obtained from Department of Environment. The main objective of this study is to describe the relationship between PM_{10} with other gases and weather conditions by using correlation. It also aims to determine the best prediction categories. Furthermore, this research aims to find a model for predicting the concentration of PM_{10} using logistic regression. PM_{10} and O_3 at Jerantut monitoring station were found to have a strong positive correlation. The best logistic regression model was obtained at Jerantut station in 2010 with an R^2 value of 0.565. The best prediction category for Jerantut monitoring stations was shown to be healthy with a correct percentage of more than 85% obtained from the analysis of the overall and annual results between 2010 to 2012.

Keywords: Air pollution, Particulate matter, prediction model.

https://doi.org/10.15405/epsbs.2020.03.03.92
Corresponding Author: Hazrul Abdul Hamid
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Air pollution basically refers to the contamination of the indoor or outdoor environment by any types of agent that modifies the natural characteristics of the atmosphere (World Health Organization, 2014). Hanapi and Din (2012) have described that the cause of air pollution may come from many sources such as waste products, construction work, factory emissions and vehicles. Pollutants at ground level are caused by human activities and natural events. Acson International in their Healthy Air Booklet stated that the main sources of air pollution in Malaysia are industrial fuel burning, motor vehicles, domestic fuel burning, power stations as well as the burning of industrial and municipal waste (Acson Malaysia Sales and Service Sdn Bhd, 2012). Malaysia, Air Pollution Index (API) is currently use as an indicator to measure the air quality (Hanapi & Din, 2012). According to the Department of Environment Malaysia (2013), API is calculated based on five major types of air pollutants at air pollution monitoring stations belong to Department of Environment Malaysia. These include PM_{10}, sulphur dioxide, nitrogen dioxide, ground level ozone and carbon monoxide. The indications of API value below 50 classified as good, 51-100 classified as moderate, 201-300 classified as unhealthy, more than 300 is classified as hazardous whereas an API value above 500 is classified as an emergency.

PM known as particulate matter or fine dust, it is a complex mixture of liquid droplets with extremely small particles. In addition, it is made up of several components including organic chemicals, acids, dust particles, soil and metals. PM with an aerodynamic diameter of less than $10\mu m$ (PM_{10}) is one of the major air pollutants in Malaysia and in most of cities in Southeast Asia (Afroz, Hassan, & Ibrahim, 2003). The yearly average ambient concentration levels of PM_{10} between 1999 and 2013 in Malaysia were generally within the Malaysian Ambient Air Quality Guidelines (MAAQG) with a value of less than $50\mu g/m^3$. The highest level of concentration was $50\mu g/m^3$ which was recorded in 2002 whereas the lowest level of concentration was $39\mu g/m^3$ which was recorded in 2010. The concentration of PM at a specific location depends on many factors such as local and regional particulate matter sources as well as geographical situation and meteorological conditions (Titos, Lyamani, Pandilfi, Alastuey, & Alados-Arboledas, 2014). The main source of air pollutants, especially PM is traffic exhaust emissions (Bycenkiene, Plauskaite, Dudoitis, & Ulevicius, 2014).

Department of Occupational Safety and Health Malaysia (2014) stated that PM_{10} can negatively affect human health if the API value exceeds 100. Environment Statistics Time Series Malaysia (2013) summarised that unhealthy events caused by transboundary heavy particulate matter were recorded between 2002 to 2013 with a maximum of 3 days recorded in 2005. By referring to heavy particulate matter pollution reported by the New Straits Times (2014), the standard operating procedure for schools to close is when the API exceeds 200 where the air quality is at a “very unhealthy” level. However, the number of days for the school to be closed depends on the duration of the high particulate event. This causes uncertainty to the public because they would not be able to know the duration of closure. It would make it difficult for them to plan or schedule outdoor social activities. In reduced the difficulties, the appropriate method of prediction.

https://doi.org/10.15405/epsbs.2020.03.03.92
Corresponding Author: Hazrul Abdul Hamid
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

2. Problem Statement

According to Pascal et al. (2014), exposure to PM_{10} has been consistently associated with serious health outcomes, resulting in an increase in mortality and hospital admissions predominantly related to cardiovascular and respiratory disease. There are many significant studies have linked PM_{10} to a series of significant health problems, including aggravated asthma, increase in respiratory symptoms like coughing and difficult breathing, chronic bronchitis, decreased lung function, and premature death. One of the unhealthy events in Malaysia is the presence of heavy particulate matter caused by uncontrolled forest fires originating from the Indonesian province of Sumatra during the burning season (Norela, Saidah, & Mahmud, 2013). Forest fires are normally used for land preparation and forest clearance by people involved in farming. Unfortunately, this could develop into uncontrollable wildfires. This situation usually happens between June and November coinciding with drier weather conditions (Salinas et al., 2013). Due to these issues, there are need to provide an early warning to those who may be effected. Short term prediction is quite relevant to provide the information about PM_{10} concentration.

3. Research Questions

Is logistic regression suitable for prediction of PM_{10} concentration?

4. Purpose of the Study

The main objective in this study is to describe the relationship between PM_{10} with other gases and weather conditions by using correlation. It also aims to determine the best prediction categories. Furthermore, this research aims to develop a model for predicting the concentration of PM_{10} using logistic regression.

5. Research Methods

From a set of variables that can be continuous, discrete, dichotomous or a mixture of these variables, we can use a method to predict a discrete outcome. This method is known as logistic regression. Logistic regression can be used to answer the same questions as discriminant analysis. However, the difference between logistic regression and discriminant analysis is that it has no assumption about the distribution of independent variables. The application of logistic analysis is predicting the success or failure of a new product, determining what category of a credit risk a person will fall into and predicting whether a firm will be successful or otherwise.

In statistical analysis, the main objectives of logistic regression are to correctly predict categories of outcome for individual cases as well as to establish a relationship between the outcome and the independent variables.

The main purpose of logistic regression in statistical analysis to correctly predict categories of outcome for individual cases. A model must create that includes useful and related independent variables in order accomplish this purpose. Beside that, logistic regression also purposely to measure the relationship between categorical dependent variable and independent variables.

https://doi.org/10.15405/epsbs.2020.03.03.92
Corresponding Author: Hazrul Abdul Hamid
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Logistic regression does not require the assumption of normality. However, the sample size must be large enough, at least 100 observations and a ratio of 20 observations for each independent variable. For this distribution, a log transformation needed along to create link with a normal regression equation. The log transformation or known as logistic regression of $\boldsymbol{p}$ also called as *logit*($\boldsymbol{p}$) defined as:

$$\boldsymbol{logit(p) = \log_e\left(\frac{p}{1-p}\right) = \ln\left(\frac{p}{1-p}\right)} \quad (1)$$

Logit($\boldsymbol{p}$) is the log base e of the $\boldsymbol{p}$. From Equation 1, value $\boldsymbol{p}$ must in range between 0 and 1, then *logit* ($\boldsymbol{p}$) will scale from negative infinity and positive infinity. The graph of *logit*($\boldsymbol{p}$) symmetrical at $\boldsymbol{p = 0.5}$. From Equation (1), the logistic regression equation form:

$$\boldsymbol{logit(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_3 X_3 + \cdots + \beta_k X_k} \quad (2)$$

Equation (2) show the logistic equation form behaviour of linear fit model. This model uses maximum likehood in criterion for find the best fit rather than least square deviation. The value of $\boldsymbol{p}$ can calculated by following formula:

$$\boldsymbol{p = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots}}{1 + e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots}}} \quad (3)$$

where p is the probability of the parameter of interest, which is the probability of the concentration of PM_{10}, $\boldsymbol{e}$ is value of natural logarithm (approximate 2.178…), $\boldsymbol{\alpha}$ is the value of constant coefficient and β is the coefficient for independent variables (temperature, relative humidity, NO_2, SO_2, O_3 and CO. There are three possible outcomes of PM_{10} level for the logistic regression model which are healthy (Y=1), moderate (Y=2) and unhealthy (Y=3). These variables of PM_{10} are grouped according to the relationship between PM_{10} concentration and Air Pollution Index in Malaysia as shown in Table 1.

Table 01. Air Pollution Index, description of air quality and the relationship with PM_{10} values. (Source: Ul-Saufie, 2012)

API	**Description**	**Concentration of PM_{10} ($\mu g/m^3$)**
0 < API < 50	Healthy	$0 < PM_{10} \leq 75$
50 < API < 100	Moderate	$75 < PM_{10} \leq 150$
API > 100	Unhealthy	$PM_{10} > 150$

60% of the training data was used to obtain the logistic regression model. Another 40% of the data was used for validation purposes. When the percentage correct prediction of the training data is the same or higher than the validation data, the model is considered as good and suitable to used for prediction.

5.1. Data and Area of Study

In this research, the secondary data used was recorded between 2010 to 2012. This data set consists of the data on air pollutants such as PM_{10}, CO, SO_2, NO_2 and O_3 with the meteorological data of temperature and relative humidity. The secondary data was obtained from the Air Quality Division of the Department

https://doi.org/10.15405/epsbs.2020.03.03.92
Corresponding Author: Hazrul Abdul Hamid
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

of Environment Malaysia. The data was collected and monitored by Alam Sekitar Malaysia Sdn. Bhd. (ASMA), which is the authorized agency for DoE (Azid et al., 2014). The data was subjected to standard quality control processes and quality assurance procedures which followed the standard quality outlines by the United States Environment Protection Agency (USEPA) (Latif et al., 2014).

6. Findings

Based on the descriptive statistics provided in Table 02, the reading of PM10 concentration does not exceed the hourly Malaysia Ambient Air Quality Guidelines (MAAQG) which is $150\mu g/m^3$. The highest average of PM_{10} concentrations recorded in 2011 ($39.09\mu g/m^3$) while in 2010 ($37.00\mu g/m^3$) and 2012 ($37.49\mu g/m^3$). These averages are lower than $\boldsymbol{50\mu g/m^3}$ (daily MAAQG), give indication the PM_{10} concentration in Jerantut area still meet the standard set by DoE. For the standard deviation and coefficient of variance, the lowest value show in 2011 rather than 2010 and 2012. The value of kurtosis (-.10) and skewness (0.77) in 2010 lowest from this three consecutive years state that the pattern of distribution of data in 2010 close to the normal distribution.

Table 02. Descriptive Statistics for PM_{10} at Jerantut Station

Parameter/Year	2010	2011	2012
Minimum ($\mu g/m^3$)	14.000	16.000	17.0000
Maximum ($\mu g/m^3$)	82.0000	92.0000	104.0000
Mean ($\mu g/m^3$)	37.8800	39.0900	37.4900
Std. Deviation ($\mu g/m^3$)	14.66	13.5500	15.1200
Kurtosis	-0.1000	1.1600	1.5000
Skewness	.7700	.9700	1.1900
Coefficient of Variation	.3900	.3500	.4000

6.1. Correlation between PM_{10}, other Gaseous and Meteorological Parameters

The Pearson correlation analysis was used to study the correlation between gaseous (SO_2, NO_2, O_3 and CO_2), PM_{10} and meteorological parameters. The correlation between other gaseous, PM_{10} and meteorological parameters for Jerantut monitoring stations is shown in Table 03.

Table 03. Correlation air pollutants and meteorological parameters in Jerantut

Parameter	Temperature	Humidity	SO_2	NO_2	O_3	CO
PM_{10}	0.204	-0.138	0.004	0.394	0.614	0.476

From the Table 03, a strong correlation of 0.614 between PM_{10} and O_3 while the correlation between PM_{10} and SO_2 was weak as indicated by values of 0.004. The strong correlation between PM_{10} and O_3 at Jerantut station indicated that an increase in the concentration of O_3 will increase the concentration of PM_{10}. There was no negative correlation recorded between PM_{10} and other gaseous parameters.

A positive significant correlation between PM_{10} and temperature is expected as higher temperature leads to high evaporation and resuspension of particles in ambient air. Furthermore, the negative correlation between relative humidity and PM_{10} was also expected. This is because humidity and rainfall would reduce the number of particulate matter in the air because of the wash-out process (Mahiyuddin et al., 2013). High

https://doi.org/10.15405/epsbs.2020.03.03.92
Corresponding Author: Hazrul Abdul Hamid
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

temperature tends to cause lower humidity level and hot weather, which in turn promotes local and regional biomass burning that subsequently increases the quantity of particles in air (Latif et al., 2014).

6.2. Logistic Regression Analysis

The logistic regression analysis was conducted to determine the best fitting model describing the relationship between dependent variables which include healthy, moderate or unhealthy and a set of independent explanatory variables which include temperature, relative humidity, SO_2, NO_2, O_3 and CO. The value of R^2 and the percentage of the correct prediction of group classification were also calculated between 2010 to 2012 to find the best fit model.

The overall and yearly regression model and R^2 values between 2010 to 2012 at Jerantut Station are shown in Table 04. The results showed the overall R^2 value of the model was 0.956. The highest R^2 values for each year obtained for the model in 2010, 2011 and 2012 were 0.565, 0.349 and 0.296 respectively.

Table 04. Logistic regression model and R^2 values at Jerantut station

Year	Logistic Model Function	R^2 value (Nagelkerke)
2010	$Y_1 = 89.46 + 0.340T – 1.050RH – 190.76NO_2 – 332.11O_3 + 13.71CO$	0.565
2011	$Y_1 = -34.90 + 0.264T + 0.190RH – 42.99NO_2 – 48.63O_3 + 14.86CO$	0.296
2012	$Y_1 = 108.30 – 0.318T – 1.128RH – 35.257NO_2 + 54.051O_3 + 4.817CO$	0.349
Overall	$Y_1 = 3.589 + 0.124T – 0.128RH – 8.385NO_2 – 8.672O_3 + 4.040CO$	0.110

Note: Y_1 = Model for the healthy group compared with the unhealthy group

Table 05 and Table 06 show the results of the overall percentage correct group classification of training data and validation data, respectively. The results obtained show that training data obtained a percentage correct classification of 97.2% while validation data obtained a percentage correct classification of 97.0%. This indicated that the model was good because training data had a higher percentage correct prediction value compared to validation data. The healthy group obtained 100% in terms of correct prediction. However, the percentage of prediction for the moderate group was 0.0% due to the small number of PM_{10} data that in the moderate category.

Table 05. Overall percentage correct of group classification of training data at Jerantut station

Observed	Predicted		
	Healthy	Moderate	Percentage Correct, %
Heathy	546	0	100.0
Moderate	16	0	0.0
Overall Percentage, %	97.2		

https://doi.org/10.15405/epsbs.2020.03.03.92
Corresponding Author: Hazrul Abdul Hamid
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Table 06. Overall percentage correct of group classification of validation data for at Jerantut station

Observed	Predicted		
	Healthy	**Moderate**	**Percentage Correct, %**
Heathy	324	0	100.0
Moderate	10	0	0.0
Overall Percentage, %	97.0		

Table 07 and Table 08 show the results of the percentage correct group classification of training data and validation data in 2010, respectively. The results showed that the training data obtained a percentage correct classification of 97.3% whereas the validation data obtained a percentage correct classification of 94.5%. This indicated that the model was good because the training data had a higher percentage correct prediction value compared to the validation data. The healthy group obtained 100% in terms of correct prediction for the training data and 98.1% for the validation data. The percentage of prediction for the moderate group was 25.0% for the validation data and 0.0% for the training data.

Table 07. Percentage correct of group classification of training data in 2010

Observed	Predicted		
	Healthy	**Moderate**	**Percentage Correct, %**
Heathy	107	0	100.0
Moderate	3	1	25.0
Overall Percentage, %	97.3		

Table 08. Percentage correct of group classification of validation data in 2010

Observed	Predicted		
	Healthy	**Moderate**	**Percentage Correct, %**
Heathy	52	1	98.1
Moderate	2	0	0.0
Overall Percentage, %	94.5		

Table 09 and Table 10 show the results of the percentage correct group classification of training data and validation data in 2011 at, respectively. The results showed that the training data obtained a percentage correct classification of 96.6% whereas the validation data obtained a percentage correct classification of 96.2%. This indicated that the model was good because the training data had a higher percentage correct prediction value compared to validation data. The healthy group obtained 99.6% in terms of correct prediction for the training data and 98.4% for the validation data. The percentage of prediction for the moderate group was 40.0% for the validation data and 0.0% for the training data.

https://doi.org/10.15405/epsbs.2020.03.03.92
Corresponding Author: Hazrul Abdul Hamid
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Table 09. Percentage correct of group classification of training data in 2011

Observed	Predicted		
	Healthy	Moderate	Percentage Correct, %
Healthy	227	1	99.6
Moderate	1	0	0.0
Overall Percentage, %	99.6		

Table 10. Percentage correct of group classification of validation data in 2011

Observed	Predicted		
	Healthy	Moderate	Percentage Correct, %
Healthy	123	2	98.4
Moderate	3	2	40.0
Overall Percentage, %	96.2		

Table 11 and Table 12 show the results of the percentage correct group classification of training data and validation data in 2012 at Jerantut station, respectively. The results obtained showed that the training data obtained a percentage correct classification of 98.6% whereas the validation data obtained a percentage correct classification of 95.9%. This indicated that the model was good because the training data had a higher percentage correct prediction value compared to the validation data. In terms of correct prediction, the healthy group scored 100.0% for the training data and 99.3% for the validation data. The percentage of prediction for the moderate group was 0.0% for both training and validation data.

Table 11. Percentage correct of group classification of training data in 2012

Observed	Predicted		
	Healthy	Moderate	Percentage Correct, %
Heathy	216	0	100.0
Moderate	3	0	0.0
Overall Percentage, %	98.6		

Table 12. Percentage correct of group classification of validation data in 2012

Observed	Predicted		
	Healthy	Moderate	Percentage Correct, %
Heathy	140	1	99.3
Moderate	5	0	0.0
Overall Percentage, %	95.9		

7. Conclusion

From the secondary data obtained from the DoE which was analysed via descriptive statistics and correlation, the result shows that the level of maximum concentration of PM_{10} at Jerantut station was under the limit based on the Malaysian Ambient Air Quality Guidelines (MAAQG) from 2010 to 2012. The correlation analysis between PM_{10} and other gases and meteorological parameters at Jerantut station showed a strong correlation value of 0.614 between PM_{10} and O_3. The result of the logistic regression analysis had a classification percentage of more than 90% for training and validation data every year. Moreover, the best

logistic regression at Jerantut station in 2010 was an R^2 value of 0.565. The best prediction of percentage correct obtained was more than 85% which is considered healthy for the overall and yearly analysis.

Acknowledgments

Special thanks to Universiti Sains Malaysia for the funding with a Short-term Grant (PJJAUH/6315089).

References

Acson Malaysia Sales and Service Sdn Bhd (2012). Air Pollution and Its Sources. *Healthy Air Booklet*, Available from: http://www.acson.com.my

Afroz, R., Hassan, M. N., & Ibrahim, N. A. (2003). Review of Air Pollution and Health Impacts in Malaysia. *Environmental Research, 93*(2), 71-77.

Azid, A., Juahir, H., Toriman, M. E., Endut, A., Kamarudin, M. K. A., Rahman, M. N. A., & Yunus, K. (2014). Source Apportionment of Air Pollution: A Case Study in Malaysia. *Jurnal Teknologi, 72*(1), 83-88.

Bycenkiene, S., Plauskaite, K., Dudoitis, V., & Ulevicius, V. (2014). Urban Background Levels of Particle Number Concentration and Sources in Vilnius, Lithuania. *Atmospheric Research, 143*, 279-292.

Department of Environment Malaysia (2013). Ministry of Natural Resources and Environment, http://apims.doe.gov.my/apims/General%20Info%20of%20Air%20Pollutant%20Index.pdf

Department of Occupational Safety and Health (2014). *Guidelines for the Protection of EmployeesAgainst the Effects of Haze at Workplace*, Available from: http://www.dosh.gov.my/index.php?option=com_content&view=article&id=856:guidelines for-the-protection-of-employees-against-the-effects-of-haze-at-workplaces&catid=491:guidelines&Itemid=1199&lang=en

Environment Statistics Time Series Malaysia (2013). Available from: https://www.statistics.gov.my/dosm/uploads/files/3_Time%20Series/Malaysia%20Time%20Series%202013/19Alam_Sekitar.pdf

Hanapi, N., & Din, S. A. M. (2012). A Study on the Airbone Particulates Matter in Selected Museums of Peninsular Malaysia. *Procedia-Social and Behavioural Sciences, 50,* 602-613.

Latif, M. T., Dominick, D., Ahamad, F., Khan, M. F., Juneng, L., Hamzah, F. M., & Nadzir, M. S. M. (2014). Long Term Assessment of Air Quality from a Background Station on theMalaysian Peninsula. *Science of the Total Environment, 482,* 336-348.

Mahiyuddin, W.R.W., Sahani, M., Aripin, R., Latif, M.T., Thach, T.Q., and Wong, C.M. (2013). Short-term Effects of Daily Air Pollution on Mortality. *Atmospheric Environment, 65,* 69-79.

New Straits Times (2014). *Haze: Schools in Kuala Langat District also ordered to close*, 14 March. Available from: http://www2.nst.com.my/7-day-news/wednesday/haze-schools-in-kuala- langat-district-also-ordered-to-close1.512403

Norela, S., Saidah, M. S., & Mahmud, M. (2013). Chemical Composition of the Haze in Malaysia2005. *Atmospheric Environment, 77,* 1005-1010.

Pascal, M., Falq, G., Wagner, V., Chatignoux, E., Corso, M., Blanchard, M., & Larrieu, S. (2014). Short-term Impacts of Particulate Matter (PM_{10}, $PM_{10-2.5}$, $PM_{2.5}$) on Mortality in Nine French Cities. *Atmospheric Environment, 95,* 174-184.

Salinas, S. V., Chew, B. N., Miettinen, J., Campbell, J. R., Welton, E. J., Reid, J. S., & Liew, S. C. (2013). Physical and Optical Characteristics of the October 2010 Haze Event over Singapore: A Photometric and Lidar Analysis. *Atmospheric Research, 122,* 555-570.

Titos, G., Lyamani, H., Pandilfi, M., Alastuey, A., & Alados-Arboledas, L. (2014). Identification of Fine(PM_1) and coarse (PM_{10-1}) Sources of Particulate Matter in an Urban Environment. *Atmospheric Environment, 89,* 593-602.

Ul-Saufie, A. Z. (2012). *PM_{10} Concentration Short Term Prediction Using Regression Artificial Neural Network and Hybrid Models* (Doctoral Dissertation). Universiti Sains Malaysia.

World Health Organization (2014). Available from: http://www.who.int/topics/air_pollution/en/