Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
https://science.utm.my/procscimath
Volume 21 (2024) 95-104

Forecasting Monthly PM_{10} Concentration in Petaling Jaya using Dynamic Regression Model

Izzah Atirah Zambri, Ani Shabri*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: ani@utm.my

Abstract
In recent years, the problem of atmospheric pollution has received increasing attention. Air pollution is one of the serious environmental issues in urban areas. The high concentrations of particulate matter can seriously impact human health, agricultural and ecosystems. Time series analysis and forecasting has become a major tool in many applications in air pollution and environmental management fields. Accurate air quality prediction can help governments and individuals make proper decisions to cope with potential air pollution. Among the most effective approaches for analyzing time series data is the Box-Jenkins (BJ) methodology or autoregressive integrated moving average (ARIMA) models. In this study, the average monthly particulate matter, PM_{10} data taken from the Petaling Jaya, Selangor monitoring station for the period 2003 to 2022 with a total of 240 readings was used in three models which is Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) and a combination of regression and ARIMA refers to the dynamic regression model. The aim of this study was to determine the best model to forecast PM_{10} concentration in Petaling Jaya. The lowest Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) values were used as the model selection criteria. Among the three models, the ARIMA model appears to be the best model as it has the lowest RMSE and MAPE values. ARIMA $(0,1,2)$ $(0,1,1)_{12}$ was used to forecast the PM_{10} concentration from January 2022 to December 2025.
Keywords: air quality status; particulate matter; dynamic regression; forecasting; ARIMA

Introduction

The rapid urbanization, industrialization and population growth in both developed and developing countries have contributed to the increase in pollution problems. Air pollution is the most prevalent type of pollution in the world and poses a significant impact on global public health, agricultural crops and ecosystem. It is predominantly caused by natural activities such as volcano eruptions and human-based factors, for example open burning, industrial processes and fuel burning vehicle. In Malaysia, emission from motor vehicles, open burning and factories are the main reasons for air pollution to increase especially in urban areas.

Particulate matter especially PM_{10}, consistently records the highest Air Pollution Index (API) values compared to other air pollutants, particularly in industrial and urban areas [1]. Particulate matter is a combination of solid and liquid particles present in the air. The characteristics of PM_{10} in the atmosphere are influenced by meteorological conditions such as ambient temperature ($°C$), wind speed (ms^{-1}) and relative humidity and gaseous pollutants which control the dispersion, formation and transportation of PM_{10} [2]. Meteorological parameters are one of the important factors influencing urban quality.

According to the Department of Statistics Malaysia (DOSM), the total population of Malaysia was 32.7 million in year 2022, compared to 32.6 million in year 2021. In 2022, the three states with the highest population composition were Selangor followed by Johor and Sabah with 21.6%, 12.3% and 10.4% respectively. In term of urban population, it increased to 77.0% in year 2021 compared to 70.9% in year 2010. As Kuala Lumpur and Putrajaya reached of urbanization level, Selangor and Penang also are among the states that recorded highest urbanization [3]. Petaling Jaya is in ninth largest city in Malaysia with a population of 520,698 people.

Due to the accuracy of existing time series forecasting techniques, time series models play a crucial role in many decision-making processes. Air quality forecasting is highly reliable and effective for implementing control measures and can also be recommended as a preventive action for upcoming regulations. According to numerous time series forecasting studied, the predictive performance of combined models has been improved. Combined models are used when a single model alone may not be sufficient to capture all the characteristics of the time series data. To achieve more accurate results, hybrid models can be adopted. The purpose of these models is to reduce the risk of using an unsuitable model by combining several models together [4].

This research aims to (1) predict the PM_{10} concentration over the next 36 months using three different models which is Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) and Dynamic Regression Model (2) find out the accuracy of the models results based on the RMSE and MAPE values. The amount of MAPE and RMSE values is obtained by comparing the forecasting results of PM_{10} concentration and testing data in January 2022-December 2022. The smaller the MAPE and RMSE values, the better the forecasting results will be.

Literature Review

The presence of globalized development has elevated the risk of air pollution in Malaysia, Air pollution is defined as a condition where gaseous pollutants are present in the atmosphere at concentration above their normal ambient levels. In Malaysia, particulate matter is one of the atmospheric pollutants caused by automobile exhaust and power plants and it can be formed in the atmosphere through reactions with gaseous emissions [13].

A study conducted by [5] in Ahvaz using nonlinear autoregressive (NAR) and multi-layer perceptron (MLP) models to predict respiratory mortality. The MLP, an artificial neural network consists of three layers (input, hidden, output). They considered various inputs such as particulate matter (PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3) and nitrogen dioxide, as well as temperature (T) and relative humidity (RH). The output layer represented mortality or morbidity of respiratory diseases. The authors found that CO had a greater impact on mortality and morbidity compared to other pollutants, suggesting it as a key predictor of air pollution's effects on respiratory health. Meanwhile, [6] investigate the relationship between PM_{10} concentration during the summer monsoon dry seasons with local meteorological parameters, synoptic weather conditions and hotspot number at six stations in Klang Valley using a simple multiple linear regression (MLR). The result showed that local meteorological factors, particularly local surface temperature, humidity and wind speed together with foreign hotspot numbers and synoptic weather conditions significantly correlated to PM_{10} concentration.

[7] characterized the pattern of $PM_{2.5}$ concentration at seven stations, including Alor Setar, Shah Alam, Pasir Gudang, Ipoh, Kuantan, Kuala Terengganu and Miri using seven indicator parameters (carbon monoxide, ozone, sulphur dioxide, nitrogen dioxide, humidity, temperature and wind speed). $PM_{2.5}$ concentration were predicted for each monitoring stations using multiple linear regression (MLR) and artificial neural networks (ANN). The predictive accuracy of MLR and ANN was measured using the coefficient of determination (R^2) while errors were calculated using the sum of square error (SSE) and mean square error (MSE). The higher R^2 value and smaller SSE and MSE values for all stations by using ANN indicate that this method is better at predicting $PM_{2.5}$ concentration than MLR. This research demonstrated that the ANN model performs better, reducing the deviation of the model and increasing the precision of the $PM_{2.5}$ model forecast.

Furthermore, [8] compared the results of the ARIMA model, Recurrent Neural Network's Long Short Term Memory algorithm (LSTM) and Facebook's Prophet algorithm with $PM_{2.5}$ concentration data, the experimentation showed that the standard ARIMA model of order (3,1,1) provided the best fit for predicting the observation with a low RMSE compared to the other models studied. The error values for each model were calculated based on the residual difference between the observed and predicte values. Therefore, the ARIMA (3,1,1) model was used t forecast the future value of $PM_{2.5}$ data from January 2022 to June 2022 (6 months). Similarly, [9] used an ARIMA model to forecast the ozone, O_3 concentration from 2000 to 2010. They proposed that by combining both seasonal and non-seasonal

models, ARIMA (1,0,00(0,1,1) would successfully predict the long term of O_3 concentration in Klang Valley. The selected model was found to be the best forecast for future surface O_3. The result showed that the O_3 concentration increased steadily in Klang Valley until 2020.

In a research by [14], different methods for predicting PM_{10} concentration in the Sfax Southern Suburbs were investigated. They tested three models namely a multilayer perceptron (MLP) network, an ARIMAX model that introduced external variables and a novel hybrid model combining both (ARIMAX-ANN). Their main goal was to determine which model worked best for forecasting the maximum 24-hour PM_{10} concentration. The analysed data consist of hourly and daily time series of PM_{10} and meteorological data from 2005-2009 and they compared the performance of these models. They hybrid model is better than the other two models since it introduces autoregressive, linear and nonlinear time series patterns. Therefore, the combined ARIMAX-ANN model can be used as an efficient tool for forecasting the maximum 24-hour PM_{10} concentration.

Methodology

Multiple Linear Regression

Multiple Linear Regression (MLR) model is globally and widely used over many years as a method for air pollution forecasting, which can help to attempt the uncertainty of the future simply by relying on past and current data for decision making. The fundamental basis of this model represents the relationship between the dependent variable and several independent variables such as meteorological factors and gaseous pollutants [7].

Multi-collinearity assumption will be verified by Variable of Inflation (VIF) accompanied with the regression output, where as long as the average VIF under 10 the conducted regression should be fine, there is no multicollinearity between the independent variables [8]. The VIF is given by:

$$VIF_i = \frac{1}{1-R_i^2} \quad (1)$$

where:

VIF_i : the variance inflation factor associated with i-th predictor

R_i^2 : the multiple coefficients of determination in a regression of the i-th predictor on all other predictors

The Durbin-Watson (D-W) test was used to determine the autocorrelation ability of PM_{10} concentration from previous day to predict PM_{10} concentration in the current data. The range values of the test must be between 0 and 4 to show that the residuals are uncorrelated. The DW equation is given by:

$$DW = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \quad (2)$$

where:

n : the number of observations

$\varepsilon_i = y_i - \bar{y}_i$ (y_i = observed values and $\bar{y}_i$ is the predicted value)

3.2. Autoregressive Integrated Moving Average (ARIMA)

The stages involved in building an ARIMA model

1) Model Identification
2) Model Estimation
3) Diagnostic Checking
4) Forecasting

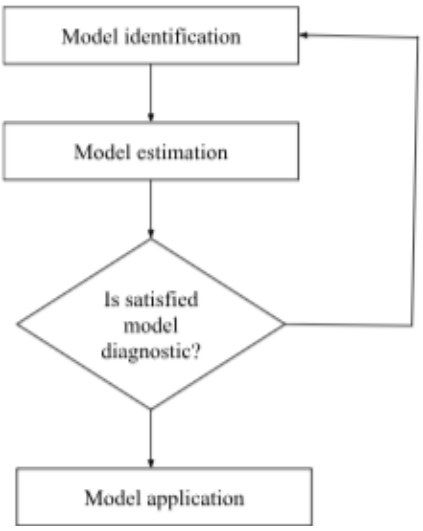

Figure 1 Main Stage of Building an ARIMA model

Model Identification

Determine whether the series is stationary or not by considering the graph of Autocorrelation Function (ACF). If a graph of ACF of the time series values either cuts off fairly quickly or dies down fairly quickly, then the time series value should be considering stationary. If a graph of ACF dies down extremely slowly, then the time series values should be considered non-stationary. If the series is not stationary, it can often be converted to a stationary series by differencing. Differencing is done until a plot of the data indicates the series varies about a fixed level, and the graph of ACF either cuts off fairly quickly or dies down fairly quickly.

Model	ACF	PACF
AR (p)	Dies down	Cut off after lag q
MA (q)	Cut off after lag p	Dies down
ARMA (p,q)	Dies down	Dies down

Table 1 The theory of ACF and PACF

Model Estimation

The estimation of AR parameters is very crucial in time series analysis for the adequate information about the model. Maximum likehood methods, ordinary lease squares (OLS), and method of moments are some of the extensively used techniques for parameter estimation in time series analysis.

Model Diagnostic Checking

Diagnostics checking in time series model is similar to the regression analysis which included testing the parameters and residuals tests. Parameters testing by using the t-test is to check and retain only those estimate parameters $\hat{\alpha}(L)$ and $\hat{\beta}(L)$ whose t- ratios are significantly greater than a predetermined critical value (that is, $|t| > 2$ at 5% significance level). Then, the residual tests are carried out using the Akaike Information Criterion (AIC) test and the Ljung-Box test or also known as Q statistics.

Dynamic Regression

The dynamic regression model describes the dynamic relationship that link the input series (X_t) with the output series (Y_t). Hence the effect of the input series on the output series is shown by the transfer function, and this effect is distributed over subsequent periods. The ARIMAX model is sometimes called a conversion function model and is expressed mathematically as shown in the following formulas:

$$Y_t = (w_0 + w_1 B + \cdots + w_k B^k) X_t + \varepsilon_t \quad (3)$$

where:

Y_t : Output series (dependent variable)

X_t: Input series (independent variable)

$w_0, w_1, \ldots, w_k$: Conversion function weights k

ε_t : White noise, which is a time series that includes other effects on the series (Y_t), it a series independent of the series (X_t)

Measure of accuracy

In this study, model performance is evaluated by measuring the value of Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The equation of MSE, RMSE and MAPE is shown below:

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2}{N}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{\sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{N} \times 100\%$$

where:

$y_i =$ actual observation of ith

$\hat{y}_i =$ predicted observation of ith

The smallest values of MSE, RMSE and MAPE are chosen as the best model to be used in forecasting.

Results and discussion

Introduction

Petaling Jaya, an industrial area in the state of Selangor is located to the west of Kuala Lumpur, the capital city of Malaysia. Selangor is Malaysia's most populous state as well as state with the largest economy in terms of gross domestic product The machinery and equipment industry have played an important role in the economic development of Petaling Jaya. The location of Air Quality Monitoring Station (AQMS) for Petaling Jaya is precisely located at Sekolah Kebangsaan Bandar Utama, Petaling Jaya.

Time series mapping of PM_{10} concentration of Petaling Jaya 1 January 2003 - 31 December 2021, as shown in Figure 2. From the time series chart, it can be seen that in the past year, PM_{10} concentration in Petaling Jaya fluctuated greatly where there were two abnormal peaks, and the PM_{10} value was not always in a constant value near the fluctuation.

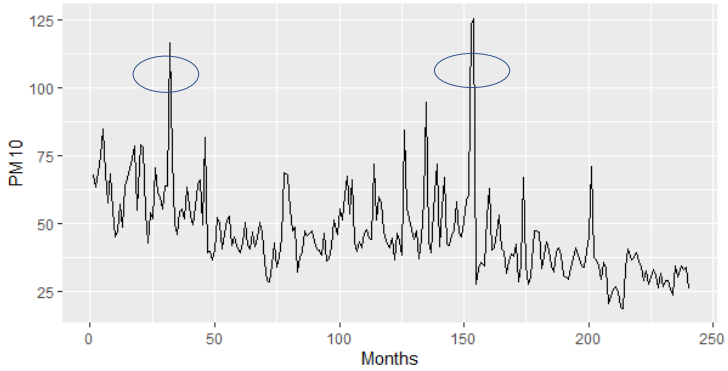

Figure 2 Time series of PM_{10} in Petaling Jaya

Multiple Linear Regression

The multiple linear regression model is used for studying the relationship between the concentration of particulate matter (PM_{10}) and four basic meteorological variables. The statistical software for data processing R Studio was used to perform the necessary analyses and calculations. Multiple linear regression model as follow:

$$Y = -5362.9095 + 8.2012X_1 + 12.2171X_2 + 0.8347X_3 + 5.0595X_4$$

The multiple linear regression model for Petaling Jaya was obtained with R^2 of 0.1145, meaning the developed model is able to explain 11.45% of the variance in the data. The range of the Variance Inflation Factor (VIF) for the independent variables in the MLR model was 1.266-2.522. The model is deemed to be no multicollinearity problem as the VIF values are all below 10. The Durbin Watson (DW) statistic indicates that the model does not have any first-order autocorrelation problem with a value of 0.88889, which is still within 0-4. According to the MLR model, variables such as temperature, wind speed, relative humidity and atmospheric pressure have positive influences.

Based on the equation of multiple linear regression model, the PM_{10} concentration for the year 2022 was forecasted for Petaling Jaya monitoring station. The time series plot of actual values and predicted values are shown in Figure 3.

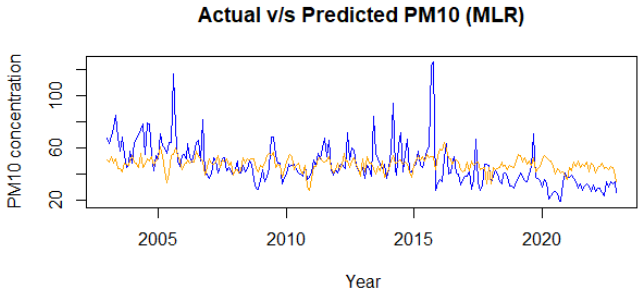

Figure 3 Fitting result of Multiple Linear Regression

Autoregressive Integrated Moving Average

Figure 4 shows the ACF are slowly decrease indicate the presence of a trend in the data. Hence the seasonal factors were considered for the ARIMA models in this study. Based on the result of ACF and PACF plot, it clearly indicated that the series is not stationary. Therefore, the monthly PM_{10} data of Petaling Jaya are first differenced to eliminate the linear trend and then differenced to eliminate the seasonal periodicity in 12 lags.

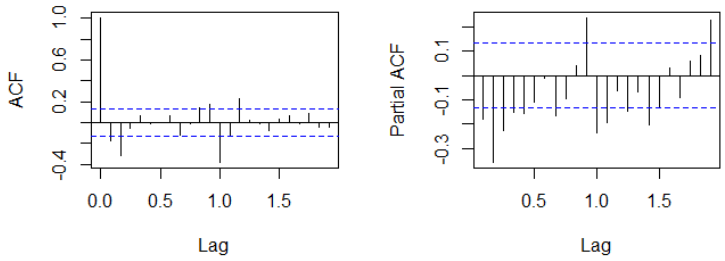

Figure 4 ACF and PACF after the first-order 12 lags differencing

After carefully examining, ACF and PACF several models were identified for test to ensure that a well specified model is not missed, the best model that satisfied statistical requirement will be chosen. The ARIMA (0,1,2) $(0,1,1)_{12}$ models were significant (p>0.05) indicating that residuals appeared to be uncorrelated and the errors were white noise. The ARIMA (0,1,2) $(0,1,1)_{12}$ have the lowest value of AIC, RMSE and MAPE. ARIMA (0,1,2) $(0,1,1)_{12}$ be selected as the best model.

Model	ARIMA $(0,1,1)\ (0,1,0)_{12}$	ARIMA $(1,1,1)\ (0,1,0)_{12}$	ARIMA $(0,1,0)\ (0,1,1)_{12}$	ARIMA $(0,1,2)\ (0,1,1)_{12}$
Ljung Box	9.558e-13	4.043e-07	0.01163	0.5653
AIC	1832.69	1799.12	1786.89	1715.14
RMSE	16.3136	15.0468	14.3232	11.9349
MAPE	23.2951	21.7447	20.6879	17.5059

Table 2 Model Selection

Based on the best fit ARIMA model, the PM_{10} concentration for the year 2022 was forecasted for Petaling Jaya monitoring station. The time series plot of actual values and predicted values are shown in Figure 5. From the graph, the model is validated since the predicted PM_{10} fluctuates around the fit.

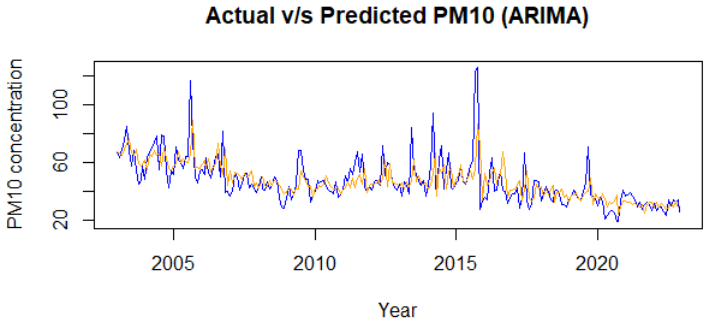

Figure 5 Fitting result of ARIMA model

Dynamic Regression

The time series for the meteorological variables used in the dynamic regression analysis are illustrated in Figure 6. It is difficult to discern relationship between the variables from this without statistical analysis.

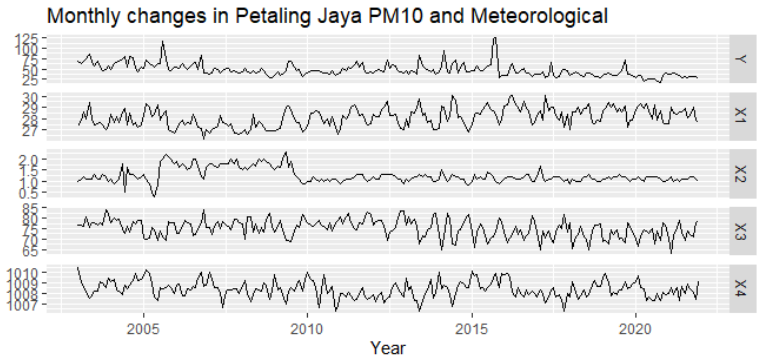

Figure 6 PM_{10} concentration and Meteorological variables

The histogram and autocorrelation function (ACF) plots of the residuals presented in Figure 7 show how well the estimated model performed on the stationary series.

Figure 7 Residual diagnostics for dynamic regression

The PM_{10} concentration for the year 2022 was forecasted for Petaling Jaya monitoring station using dynamic regression. The time series plot of actual values and predicted values are shown in Figure 8.

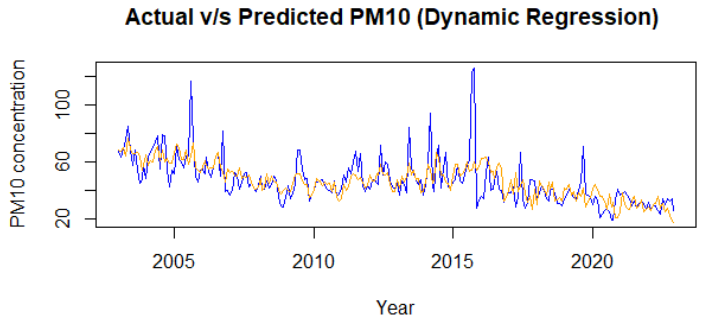

Figure 7 Fitting results of dynamic regression

Conclusion

In order to verify which model is the best model, root mean squared error (RMSE) and mean absolute percentage error (MAPE) are selected to test the prediction effect of the model.

Fitting Model	RMSE	MAPE
MLR	14.44832	47.86419
ARIMA	3.679155	9.857326
Dynamic Regression	5.472322	13.765

Table 3 Model Evaluation Comparison

As can be seen from Table 3 above, ARIMA model seem to be more accurate than dynamic regression. ARIMA showed lowest RMSE and MAPE values. From the best fitted model, the results concluded the ARIMA showed better result compared to Multiple Linear Regression and Dynamic Regression Model. ARIMA $(0,1,2)\ (0,1,1)_{12}$ turn out to be well-suited model to predict the monthly PM_{10} concentration data from January 2022 to December 2025 in Petaling Jaya. Forecasted values of January 2022 to December 2022 were used to compare the observed and forecasted values. The forecasted value for PM_{10} concentration shows significant result. The prediction series of PM_{10} concentration based on the selected model, showed consistent decreasing trend till year 2025.

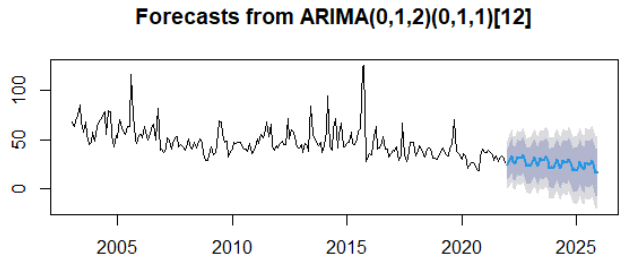

Figure 8 Forecasts from ARIMA $(0,1,2)\ (0,1,1)_{12}$

Conclusion

The key question that was intended to be answered was, whether a multiple linear regression (MLR) or an ARIMA model or a dynamic regression model is the best model to predict the PM_{10} concentration in Petaling Jaya. The ARIMA model surpassed Multiple Linear Regression and dynamic regression across all performance metrics, showing superior alignment with historical data and enhanced predictive accuracy. Due to the non-normal distribution of meteorological data, the dynamic regression model may struggle to make accurate predictions. In such cases, ARIMA models provide a better alternative, as they are robust to the non-normality of the data, making them particularly suitable for analyzing and predicting time series data affected by such distribution. ARIMA model are capable of capturing the temporal dependencies and patterns present in the data, thereby offering improved forecasting accuracy. Incorporating advanced modelling techniques like Artificial Neural Networks (ANN) holds promise for improving prediction accuracy. Air quality is influence by various factors, including weather conditions, transportation. The full use of this information may lead to higher accuracy in predicting air pollution.

Acknowledgement

The researcher would like to thank all people who have supported the research and Air Quality Division, Malaysian Department of Environment (DOE) for the air quality data and the Malaysian Meteorological Department for the meteorological data.

References

[1] Samsuri, A., Marzuki, I., Si Yuen, F., Mahfoodh, A., & Ahmed, A. N. (2016). Evaluation for long term PM10 concentration forecasting using multi linear regression (MLR) and principal component regression (PCR) models. *EnvironmentAsia*, *9*(2), 101–110

[2] Abdullah, S., Ismail, M., Samat, N. N. A., & Ahmed, A. N. (2018). Modelling particulate matter (PM10) concentration in industrialized area: A comparative study of linear and nonlinear algorithms. *ARPN J. Eng. Appl. Sci*, *13*(20), 8227–8235.

[3] Jamil, N. I., Amit, N., & Yusof, N. M. (2020). Model Evaluation on Air Pollutant Index (API) in Petaling Jaya, Malaysia. *International Journal of Advanced Science and Technology*, *29*(5s), 1959–1966.

[4] V, N., & N, A. (2017). Time series analysis to forecast air quality indices in Thiruvananthapuram district, Kerala, India. *International Journal of Engineering Research and Applications*, *07*(06), 66–84.

[5] Abdullah, Samsuri, Napi, N. N., Ahmed, A. N., Mansor, W. N., Mansor, A. A., Ismail, M., Abdullah, A. M., & Ramly, Z. T. (2020). Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia. *Atmosphere*, *11*(3), 289

[6] Juneng, L., Latif, M. T., & Tangang, F. (2011). Factors influencing the variations of PM10 aerosol dust in Klang Valley, Malaysia during the Summer. *Atmospheric Environment*, *45*(26), 4370–4378.

[7] Sobri, N. M., Yaacob, W. F., Ismail, N. A., Malik, M. A., Rahman, R. Ab., Baser, N. A., & Sukhairi, S. A. (2021). Predicting particulate matter (PM2.5) in Malaysia using multiple linear regression and artificial neural network. *Journal of Physics: Conference Series*, *2084*(1), 012010

[8] Suresh, S., Sindhumol, M. R., Ramadurai, M., Kalvinithi, D., & Sangeetha, M. (2023). Forecasting particulate matter emissions using time series models. *Nature Environment and Pollution Technology*, *22*(1), 221–228.

[9] Abdullah, Samsuri, Napi, N. N., Ahmed, A. N., Mansor, W. N., Mansor, A. A., Ismail, M., Abdullah, A. M., & Ramly, Z. T. (2020). Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia. *Atmosphere*, *11*(3), 289

[10] Abdullah, Samsuri, Ismail, M., & Fong, S. Y. (2017). Multiple linear regression (MLR) models for long term PM10 concentration forecasting during different monsoon seasons. . *Journal of Sustainability Science and Management*, *12*(1), 60–69.

[11] Mahiyuddin, W. R., Jamil, N. I., Seman, Z., Ahmad, N. I., Abdullah, N. A., Latif, M. T., & Sahani, M. (2018). Forecasting ozone concentrations using Box-Jenkins Arima modeling in Malaysia. *American Journal of Environmental Sciences*, *14*(3), 118–128.

[12] Khojasteh, D. N., Goudarzi, G., Taghizadeh-Mehrjardi, R., Asumadu-Sakyi, A. B., & Fehresti-Sani, M. (2021). Long-term effects of outdoor air pollution on mortality and morbidity–prediction using nonlinear autoregressive and Artificial Neural Networks models. *Atmospheric Pollution Research*, *12*(2), 46–56.

[13] Jalaludin, J., Wan Mansor, W. N., Abidin, N. A., Suhaimi, N. F., & Chao, H.-R. (2023). The impact of air quality and meteorology on covid-19 cases at Kuala Lumpur and Selangor, Malaysia and prediction using machine learning. *Atmosphere*, *14*(6), 973.

[14] Samsuri, A., Marzuki, I., Si Yuen, F., Mahfoodh, A., & Ahmed, A. N. (2016). Evaluation for long term PM10 concentration forecasting using multi linear regression (MLR) and principal component regression (PCR) models. *EnvironmentAsia*, *9*(2), 101–110.