Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 2, 2021, page 128-137

# Hybrid Model in Forecasting Monthly Maximum Air Pollution Index of Pasir Gudang, Johor

**[1]Liew Bih Hua and [2]Nur Arina Bazilah Kamisan**

[1,2]Department of Mathematical Sciences
Faculty of Science, Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia.

e-mail: [1]bihhua@graduate.utm.my, [2]nurarinabazilah@utm.my

**Abstract** As a developing country with rapid urbanization and industrial growth, the air quality in Malaysia is decreasing, contributing to the increase of lung diseases. Air Pollution Index (API) system presents the ambient air quality with its respective impact on human health. In this study, Box-Jenkins and the regression approach will be applied to analyse and forecast the future monthly maximum API in Pasir Gudang, Johor. A hybrid model is developed to improve the accuracy of the forecast. The results show the hybrid model is the best model to forecast API since it obtained the smallest error for MAPE and MAE.

**Keywords** air pollution index (API); forecasting performance evaluation; hybrid; regression; SARIMA

## 1 Introduction

Air pollution has been and remains to be a significant health hazard worldwide during the process of economic development [1]. Malaysia is a developing country where we are witnessing rapid urbanization and industrial growth. It cannot be denied that our air quality decreases, although this will economically benefit the development process [2].

The common pollutants are particulate matter, ozone, nitrogen dioxide, carbon monoxide, and sulphur dioxide [1]. These pollutants are released from the combustion of fossil fuels for energy generation of vehicles and electricity, emissions from factories and industries, agriculture activities with the use of pesticide, and others human activities such as open burning [3]. According to World Health Organization [4], ambient air pollution caused 3.0 million deaths in 2012. This shows that monitoring air pollution and forecasting the future air pollution index is necessary to systematically track the adoption of a policy that reduces air pollution.

The Air Pollution Index (API) is Malaysia's indicator of ambient air quality. Six-level categories represent a different range value of the index corresponding to the health impacts [5]. The index values sub-range and their respective status and health effect are shown in Table 1.

Table 1: Impact on health at different API levels

| API | Status | Health Effect |
|---|---|---|
| 0-50 | Good | Low pollution without any bad health effects. |
| 51-100 | Moderate | Moderate pollution with no negative health effects. |

| | | |
|---|---|---|
| 101-200 | Unhealthy | Worsen the health condition for older people, pregnant women, infants, and people with complications of the heart and lungs. |
| 201-300 | Very Unhealthy | Worsen the health condition of people with heart and lung complications and poor tolerance of physical activities. Affect public health. |
| >300 | Hazardous | Hazardous people at high risk and to public health. |

To continuously monitor the API reading, air quality monitoring stations are placed in strategic locations, such as industrial, urban, sub-urban, and rural areas [6]. Particulate matter with the size of 2.5 micron ($PM_{2.5}$), particulate matter with the size of 10 micron ($PM_{10}$), ozone ($O_3$), nitrogen dioxide ($NO_2$), carbon monoxide ($CO$) and sulphur dioxide ($SO_2$) are the parameters that will be measured.

In this study, the univariate API data of Pasir Gudang, Johor is obtained from the Department of Environment (DOE). SARIMA and seasonal regression models are chosen to fit the API data then another hybrid model will be created to improve the accuracy of the forecast. The performance of these three models will be compared. The time period of the collected monthly maximum API data is from January 2012 to December 2019. To determine the accuracy as well as whether the model is the best to be used to forecast the future API, Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) will be applied as the performance index of the forecast. To visualize the performance of the models, the graphical plot or time series plot will be made to compare the obtained result from each forecasting model with the actual data.

## 2 Literature Review

Box-Jenkins is widely used in the studies of air quality. The SARIMA model is able to capture the seasonality variation in the air quality index data of Dhaka and Sylhet divisions in Bangladesh while estimating the weekly air quality [7]. SARIMA and Fuzzy time series were proposed to observe the monthly API trend from 3 different stations in Malaysia. The result shows SARIMA could perform better than the Fuzzy Time Series forecasting method [8]. Alin Pohoata and Emil Lungu [9] evaluated the air quality in Ploiesti City, Romania with several statistical methods and the ARIMA model. They conclude that the ARIMA model produces a satisfactory forecasted result for the data corresponding to $NO_2$, $NO_x$ and $O_3$ with their relative errors of less than 10%. Based on the study of predicting the future air pollution level in Blagoevgrad, Bulgaria, the selected SARIMA model exhibits strong fitting performance in terms of the observed air pollutants [10].

Karatzas et al. [11] developed a regression model based on the principles of ensemble modelling to assess its ability to predict air quality levels at various locations. The overall findings show that the linear regression model has the best result since it employs the ensemble principles, although the dimensionality of functional space is restricted. The seasonal regression model is utilized when the data consist of seasonality factors [12]. The input of dummy variables in a seasonal regression model is used to predict the seasonal effect, improving the quality of the forecast [13]. Principal component regression is the combination of principal component analysis and multiple linear regression applied to forecast the daily air quality in Delhi, India. The model is later combined with the ARIMA model forming another hybrid model that produced a more accurate forecast result than the single models [14].

Since the trend and seasonality in the data are simple to apply, exponential smoothing is used to estimate the main atmospheric pollutants [15]. To forecast air pollution in Delhi, India, the

ARIMA and exponential smoothing models are potentially forecasting models among others where the results of both models produce good prediction and smaller errors [16]. ARIMA and exponential smoothing models are applied to assess the future quality index of nitrogen dioxide in Madrid City. The results show that the exponential smoothing model has higher accuracy, overcoming the ARIMA model [17].

## 3 Methodology

### 3.1 Data Pre-processing

Data pre-processing is an essential process in adjusting the historical data into readable data that can improve forecasting quality. In this study, Johnson transformation is used to transform the original data and this can be done by Minitab software. Only one distribution will be selected out of the three different systems of Johnson family distribution which is: bounded system ($S_B$), lognormal system ($S_L$) and unbounded system ($S_U$). The Johnson family with its respective transformation function are shown below:

$$S_B: Y = \theta + \eta * Log\left(\frac{y_t - \kappa}{\lambda + \kappa - \chi}\right) \quad (1)$$

$$S_L: Y = \theta + \eta * Log\left(\frac{y_t - \kappa}{\lambda}\right) \quad (2)$$

$$S_U: Y = \theta + \eta * \sinh^{-1}\left(\frac{y_t - \kappa}{\lambda}\right) \quad (3)$$

$$\sinh^{-1}(z) = Log\left(z + \sqrt{1 + z^2}\right) \quad (4)$$

where $Y$ is the transformation function, $y_t$ is the original value of API, $\theta$ is shape 1 parameter, $\eta$ is the shape 2 parameter, $\kappa$ is the location parameter, $\lambda$ is the scale parameter.

### 3.2 Box-Jenkins Approach

ARIMA extension that includes the seasonal components is called Seasonal Autoregressive Integrated Moving Average (SARIMA). SARIMA model can be written as SARIMA($p,d,q$)($P,D,Q$)$_s$. The SARIMA model is represented as:

$$\phi_p(B)\Phi_P(B^S)\nabla^d\nabla_S^D y_t = \theta_q(B)\Theta_Q(B^S)a_t \quad (5)$$

where

$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$

$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}$

$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$

$\Theta_Q(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots - \Theta_Q B^{QS}$

$\phi_p(B)$ is the non-seasonal autoregressive operator of order $p$, $\theta_q(B)$ is the non-seasonal moving average operator of order $q$,$\Phi_P(B^S)$ and $\Theta_Q(B^S)$ are the seasonal autoregressive and moving

average operator of order $P$ and order $Q$, $\nabla^d$ is the degree of non-seasonal differencing, $\nabla_s^D$ is the degree of seasonal differencing, $B$ is backward shift operator, $a_t$ is a white noise process with a mean equal to 0 and a variance equal to $\sigma_a^2$.

### 3.3 Regression Approach

When the regression approach is applied to the data with seasonality, the additional seasonal dummy variables will be included to adjust the seasonal variation and a seasonal regression model is formed. For the seasonal regression model is given by,

$$Y_t = \delta + \omega t + \sum_{i=1}^{s} \gamma_i S_{i,t} + \epsilon_t \tag{6}$$

where $Y_t$ is the dependent variable, $\delta + \omega t$ is the trend in the data, $\sum_{i=1}^{s} \gamma_i S_{i,t}$ is the seasonal additive component and $\epsilon_t$ is the residual at *t*.

### 3.4 Hybrid Methodology

In this study, the hybrid model combines the Box-Jenkins approach with the exponential smoothing approach. The SARIMA model will first be used to forecast the out-sample data while the SES model will be applied to predict the out-sample residuals of the SARIMA model. The in-sample residuals can be calculated as

$$r_t = y_t - F_t \tag{7}$$

Then, the in-sample residuals will be checked whether there are outliers before they are applied into the SES model to forecast the out-sample residuals. The equation of this SES model is expressed as follow:

$$\hat{r}_{t+1} = \alpha r_t + (1 - \alpha)\hat{r}_t \tag{8}$$

where $\hat{r}_{t+1}$ is the forecasted value of residual for next period of $t$, $\hat{r}_t$ is the forecasted value in period $t$, $\alpha$ is the smoothing constant and $r_t$ is the actual value of residual in period $t$. Finally, the out-samples of the hybrid model can be obtained by adding the out-sample forecast of the SARIMA model to the out-samples residuals from the SES model. The equation of the proposed hybrid model as

$$\hat{y}_t = \hat{F}_t + \hat{r}_t \tag{9}$$

where $\hat{F}_t$ is the forecasted value from SARIMA at time $t$ and $\hat{r}_t$ is the out-sample residual at time $t$ from the SES model.

### 3.5 Forecasting Performance Evaluation

The best forecasting model will obtain the smallest error. To evaluate the forecasting performance, Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used and their formulas are shown as below:

$$MAPE = \frac{100\%}{N}\sum_{t=1}^{n}\left|\frac{(y_t - \hat{y}_t)}{y_t}\right| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2} \tag{11}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_t - \hat{y}_t| \tag{12}$$

where $y_t$ is the actual value at time $t$, $\hat{y}_t$ is the predicted value at time $t$ and $N$ is the number of the evaluation period.

## 4 Result and Discussion

### 4.1 Overview

To construct the forecasting models, the training data used is the monthly maximum API from January 2012 to December 2018. The testing data used to evaluate the performance of different forecasting models is the monthly maximum API from January 2019 to December 2019. The time series plot of in-sample API data is shown in Figure 1.

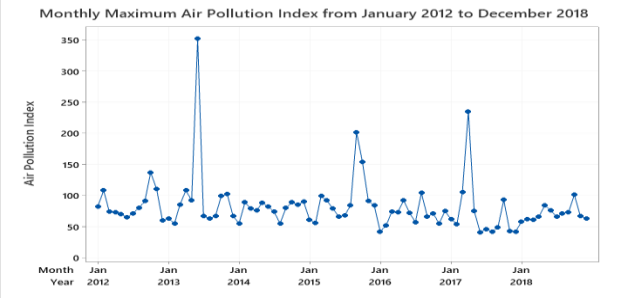


Figure 1: Time series plot of monthly maximum API from January 2012 to December 2018

### 4.2 Data Pre-processing

Johnson transformation will be applied to transform the data to follow a normal distribution before we proceed to analysis. The transformation function equals to

$$Y = -0.625617 + 0.924235 \times \sinh^{-1}\left(\frac{y_t - 62.9142}{14.1960}\right) \tag{15}$$

After the Johnson transformation, all in-sample data will be rescaled as shown in the time series plot in Figure 2.

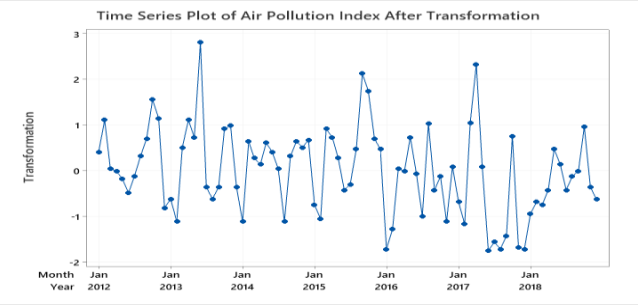


Figure 2: Time series plot of API after transformation

### 4.3 Box-Jenkins Approach

SARIMA model is applied due to the existence of seasonality in the in-sample API. Figure 3 shows the ACF plot of the monthly maximum API.

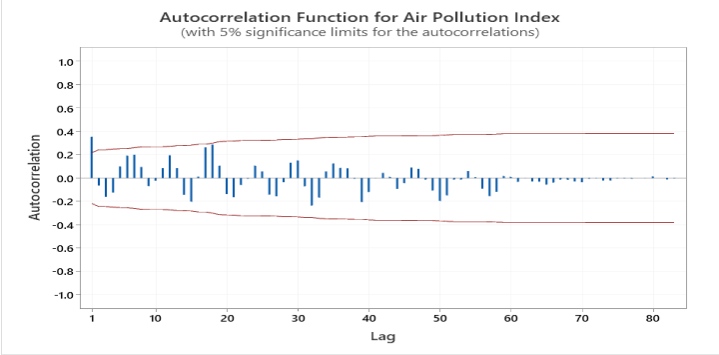


Figure 3: ACF plot of the monthly maximum API

To remove the seasonality, differencing is applied. After the first differencing followed by the seasonal differencing of order 12, the ACF and PACF plots are shown in Figure 4 and Figure 5. The possible SARIMA models and their respective AIC are shown in Table 2.

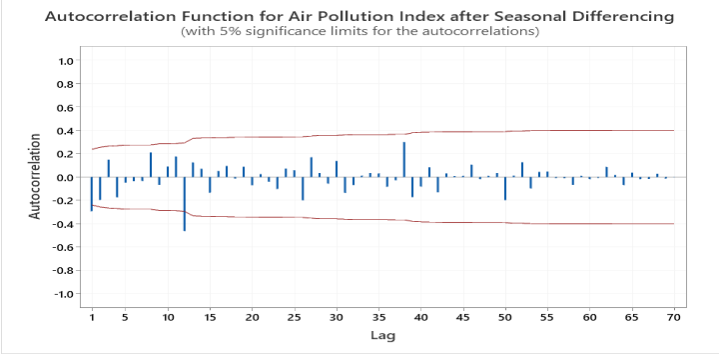


Figure 4: ACF plot of the monthly maximum API after differencing

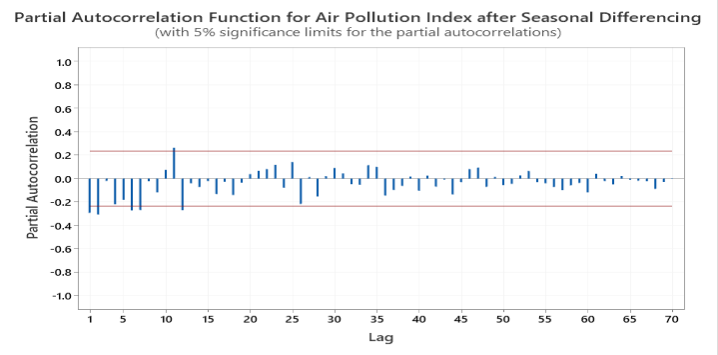


Figure 5: PACF plot of the monthly maximum API after differencing

Table 2: Possible SARIMA models

| $SARIMA\ (p,d,q)(P,D,Q)_s$ | AIC |
|---|---|
| $(2,1,1)(1,1,1)_{12}$ | 12.21495599 |
| $(1,1,1)(1,1,1)_{12}$ | 10.28996654 |
| $(0,1,1)(0,1,1)_{12}$ | 6.156207939 |

From Table 2, the best SARIMA model is SARIMA$(0,1,1)(0,1,1)_{12}$ which obtain the smallest AIC value. Therefore, the forecasting will be carried out using SARIMA$(0,1,1)(0,1,1)_{12}$ model.

### 4.4 Regression Approach

Due to the seasonal factors, the data is analysed using the seasonal regression model. To capture the seasonality in the regression model, 11 dummy variables are created for the monthly data. The dummy variables are added into the simple linear regression model forming the seasonal regression model. The equation for the seasonal regression model can be expressed as:

$$\begin{aligned} Y_t = {} & 0.2361 - 0.0117t - 0.5786S_1 - 0.3010S_2 + 0.5153S_3 + 0.7809S_4 \\ & + 0.6323S_5 + 0.3413S_6 - 0.2661S_7 + 0.0270S_8 + 0.4203S_9 \\ & + 1.2214S_{10} + 0.3390S_{11} \end{aligned} \quad (16)$$

### 4.5 Hybrid Methodology

The hybrid model for this study will combine the SARIMA$(0,1,1)(0,1,1)_{12}$ model with another potential forecasting model to improve the forecasting performance of the model. Generally, a hybrid model is able to produce the prediction with higher forecast accuracy and reliability [18]. A hybrid model will then be developed by applying the residuals of SARIMA$(0,1,1)(0,1,1)_{12}$ model into another forecasting model. The time series plot for the in-sample residuals of the SARIMA$(0,1,1)(0,1,1)_{12}$ model is shown in Figure 6.

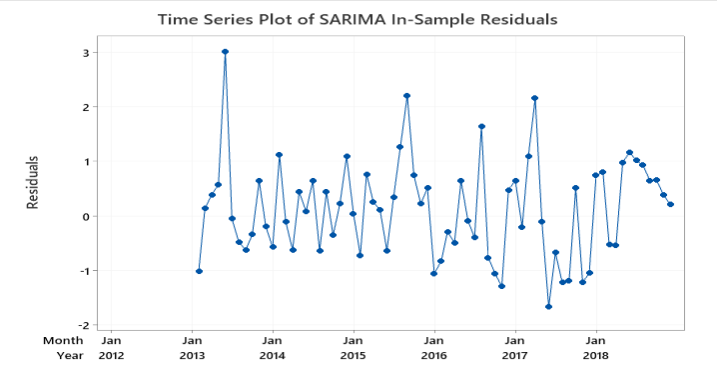


Figure 6: Time series plot of in-sample residuals of SARIMA model

The time series plot of in-sample residuals of SARIMA$(0,1,1)(0,1,1)_{12}$ model does not show trend and seasonality. Thus, the in-sample residuals will then be fitted into single exponential smoothing (SES) model and forecast the out-sample residuals from January 2019 to December 2019. Based on the Minitab output, the optimal value of the only smoothing parameter, $\alpha$ is 0.0159 and the equation of the model can be expressed as:

$$\hat{r}_{t+1} = 0.0159\, r_t + (1 - 0.0159\,)\hat{r}_t \tag{17}$$

To obtain the predicted values of API for the year 2019 by the hybrid model, the out-sample API by SARIMA$(0,1,1)(0,1,1)_{12}$ will be added with the forecasted out-sample residuals by the single exponential smoothing model.

$$\hat{y}_{84+n} = \hat{F}_{84+m} + \hat{r}_{84+n} \tag{18}$$

Figure 7 plot the actual and forecasted values of monthly maximum API from the three different forecasting models.

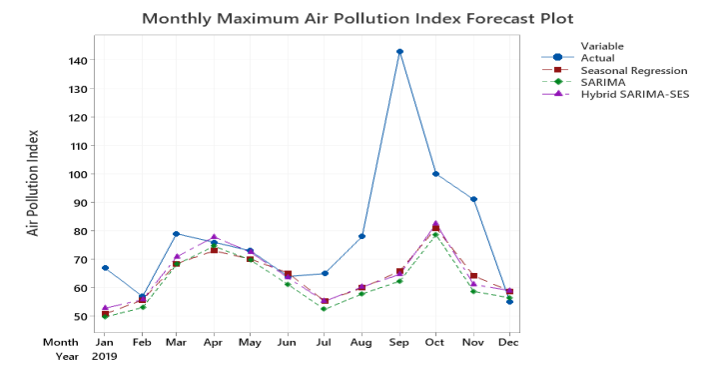


Figure 7: Comparison of the actual and forecasted API from different models

MAPE, RMSE and MAE are applied to evaluate the accuracy or forecasting performance of each of the forecasting models. The best forecasting model will obtain the lowest value for the three performance evaluations, especially for MAPE. The result of the accuracy evaluation for each model is summarised in Table 3.

Table 3: Accuracy evaluation for all forecasting models

| Forecasting Model | MAPE (%) | RMSE | MAE |
|---|---|---|---|
| Seasonal Regression | 16.3672 | **25.5671** | 15.7680 |
| SARIMA$(0,1,1)(0,1,1)_{12}$ | 18.1533 | 27.4017 | 17.3108 |
| **Hybrid SARIMA-SES** | **15.5938** | 25.8380 | **15.2663** |

Overall, these three forecasting models can produce good forecasts for the monthly maximum API of Pasir Gudang because their MAPE are between 10% to 20%. The out-sample data forecasted by the hybrid model has the lowest error or highest forecasting accuracy where its MAPE and MAE are lowest. In contrast, SARIMA$(0,1,1)(0,1,1)_{12}$ obtained the highest value for all the performance evaluation test where this means it has the lowest forecast accuracy. Hence, the hybrid SARIMA-SES model can be considered as the best model to forecast API of Pasir Gudang followed by seasonal regression and SARIMA$(0,1,1)(0,1,1)_{12}$.

## 5 Conclusion

In this study, the maximum monthly API from January 2012 to December 2018 is used in forecasting the future API from January 2019 to December 2019 and a comparison is made to the actual value to determine the performance of each time series model. The time series models in this study are seasonal regression, SARIMA and a hybrid model that combines SARIMA and SES.

In conclusion, the hybrid SARIMA-SES model obtains the lowest MAPE and MAE. This proved that the hybrid model has successfully become the best model to be utilized as the forecasting model for future monthly maximum API since it can outperform the single models. A hybrid model can achieve a better forecasting result because it considers the residual in forecasting to generate forecast data with closer value to the actual data. This can enhance forecasting efficiency.

As the recommendation for future study, the research can be developed by forecasting the specific pollutant sub-indexes, especially the most frequent pollutant with the highest index value, since it always presents in the highest concentration. Besides, further research or study on hybrid models using more advanced models in a hybrid model is highly recommended because the advanced models can be applied to non-linear and dynamic data.

## 6 References

[1] Chen, B., & Kan, H. (2008) 'Air pollution and population health: a global challenge', *Environmental Health and Preventive Medicine.* 13(2): 94-101.

[2] Chin, Y. S. J., De Pretto, L., Thuppil, V., & Ashfold, M. J. (2019) 'Public awareness and support for environmental protection-A focus on air pollution in peninsular Malaysia', *PLoS ONE.* 14(3).

[3] Adamkiewicz, G., Liddie, J., & Gaffin, J. M. (2020) 'The Respiratory Risks of Ambient/Outdoor Air Pollution', *Clinics in Chest Medicine.* 41(4): 809-824.

[4] World Health, O. (2016) *World health statistics 2016: monitoring health for the SDGs, sustainable development goals*. Geneva: World Health Organization.

[5] Koo, J. W., Wong, S. W., Selvachandran, G., Long, H. V., & Son, L. H. (2020) 'Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models', *Air Quality, Atmosphere and Health.* 13(1): 77-88.

[6] Zakaria, N. N., Sokkalingam, R., Daud, H., & Othman, M. (2019) 'Forecasting air pollution index in Klang by markov chain model', *International Journal of Engineering and Advanced Technology.* 8(6 Special Issue 3): 635-639.

[7] Islam, M. M., Sharmin, M., & Ahmed, F. (2020) 'Predicting air quality of Dhaka and Sylhet divisions in Bangladesh: a time series modeling approach', *Air Quality, Atmosphere & Health.* 13(5): 607-615.

[8] ABD RAHMAN, N. H., Lee, M. H., Suhartono, & Latif, M. (2016) 'Evaluation performance of time series approach for forecasting air pollution index in Johor, Malaysia', *Sains Malaysiana.* 45(11): 1625-1633.

[9] Pohoata, A., & Lungu, E. (2017) 'A Complex Analysis Employing ARIMA Model and Statistical Methods on Air Pollutants Recorded in Ploiesti, Romania', *Revista de Chimie.* 68: 818-823.

[10] Gocheva-Ilieva, S. G., Ivanov, A. V., Voynikova, D. S., & Boyadzhiev, D. T. (2014) 'Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach', *Stochastic Environmental Research and Risk Assessment.* 28(4): 1045-1060.

[11] Karatzas, K., Katsifarakis, N., Orlowski, C., & Sarzyński, A. (2018) 'Revisiting urban air quality forecasting: a regression approach', *Vietnam Journal of Computer Science.* 5(2): 177-184.

[12] Caruana, A. (2001) 'Steps in forecasting with seasonal regression: A case study from the carbonated soft drink market', *Journal of Product & Brand Management.* 10: 94-102.

[13] Ostertagova, E., & Ostertag, O. (2015) 'Regression Analysis and Seasonal Adjustment of Time Series', *Journal of Automation and Control.* 3: 118-121.

[14] Kumar, A., & Goyal, P. (2011) 'Forecasting of daily air quality index in Delhi', *Science of The Total Environment.* 409(24): 5517-5523.

[15] Roy, S., Biswas, S. P., Mahata, S., & Bose, R. (2018, 12-13 Oct. 2018). *Time Series Forecasting using Exponential Smoothing to Predict the Major Atmospheric Pollutants.* Paper presented at the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).

[16] Taufik, M. R., Rosanti, E., Eka Prasetya, T. A., & Wijayanti Septiarini, T. (2020) 'Prediction algorithms to forecast air pollution in Delhi India on a decade', *Journal of Physics: Conference Series.* 1511: 012052.

[17] Setiawan, I. (2020) 'Time series air quality forecasting with R Language and R Studio', *Journal of Physics: Conference Series.* 1450: 012064.

[18] Du, P., Wang, J., Yang, W., & Niu, T. (2019) 'A novel hybrid model for short-term wind power forecasting', *Applied Soft Computing.* 80: 93-106.