Sains Malaysiana 52(12)(2023): 3603-3618
http://doi.org/10.17576/jsm-2023-5212-20

Quantifying Haze Effect using Air Pollution Index Data

(Pengukuran Kesan Jerebu menggunakan Data Indeks Pencemaran Udara)

Razik Ridzuan Mohd Tajuddin* & Nurulkamal Masseran

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

Received: 13 July 2023/Accepted: 7 December 2023

ABSTRACT

Malaysia has been misfortunate with intermittent haze episodes since 1997 which affect the airquality tremendously. In Malaysia, an instrument named as air pollution index (API) is utilizedin determining the quality of air, which is influenced by the presence of haze. API values arecalculated by considering the concentration of harmful particles in haze. So, any haze episodeheavily affects the API values and can be considered as a determining factor. Since Malaysiais prone to haze, it is crucial to identify and quantify the haze effect on the API values.Therefore, four models – an autoregressive integrated moving average (ARIMA), regressionmodel with ARIMA errors (ARIMAX), time series regression and Prophet models areemployed. It is found that ARIMAX (4,0,1) with non-zero mean is the best model in describingthe API data with presence of haze as external regressor based on the smallest adequacy anderror measures for training and test datasets. In conclusion, the effect of haze is significant indescribing the API values and thus, proper health management is required during haze episodes.

Keywords: ARIMAX; haze effect; regression with ARIMA errors

ABSTRAK

Malaysia mengalami nasib malang dengan episod jerebu yang berterusan sejak tahun 1997 yang memberi kesan yang besar terhadap kualiti udara. Di Malaysia, terdapat satu pengukur yang dikenali sebagai indeks pencemaran udara (IPU) yang digunakan untuk menentukan kualiti udara yang dipengaruhi oleh kehadiran jerebu. Nilai IPU dihitung berdasarkan kepekatan zarah berbahaya dalam jerebu. Oleh itu, apa-apa episod jerebu akan memberi kesan yang besar kepada nilai IPU dan boleh dianggap sebagai satu faktor penentu. Memandangkan Malaysia cenderung untuk mengalami jerebu, adalah penting untuk mengenal pasti dan mengukur kesan jerebu terhadap nilai IPU. Oleh itu, empat model – purata bergerak terintegrasi auto regresif (ARIMA), regresi dengan ralat ARIMA (ARIMAX), regresi siri masa dan model Prophet digunakan. Didapati bahawa ARIMAX (4,0,1) dengan min bukan sifar merupakan model terbaik dalam menerangkan data IPU dengan kehadiran jerebu sebagai regresor luaran berdasarkan ukuran kecukupan serta ralat terkecil untuk set data latihan dan set data ujian. Kesimpulannya, kesan jerebu adalah signifikan dalam menerangkan nilai IPU dan oleh yang demikian, pengurusan kesihatan yang betul diperlukan sepanjang jerebu berlaku.

Kata kunci: ARIMAX; kesan jerebu; regresi dengan ralat ARIMA

Introduction

Malaysia has been suffering from haze since 1997 sourcing from domestic and neighboring countries, especially due to forest fires which was intensified by El Niño and improper management in Indonesia, which then affected several countries in Southeast Asia (Glover & Jessup 2006). Haze contains harmful particles such as sulphur dioxide, nitrogen dioxide, sulphates, nitric acid, ozone, nitrates and sulphuric acid (Liu et al. 2016). Concentrations of some of the harmful substances has already been considered in developing air pollution index (API) in Malaysia. Haze can be started by human factors such as industrializations and open burnings. Besides human factors, natural disasters which include volcanic eruptions and wildfires can also contribute to haze. Haze can cause adverse effects to humans such as lung and cardiovascular diseases (Isaifan 2023; Liu et al. 2016) as well as respiratory diseases (Mohd Nadzir et al. 2021). These effects indirectly may contribute to high hospital admissions (Albahar et al. 2022; Priyankara et al. 2021; Zhang et al. 2014).

In Malaysia, the Malaysian Department of Environment is responsible in collecting data pertinent to the development of API as well as strategizing in reducing pollution nationwide. Prior to 2017, the API values are calculated based on five influential observed pollutants – sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), particulate matter less than 10 microns in size (PM_{10}) and ozone (O_3). The PM_{10} is measured in micrograms per cubic meter (µg/m^3) whereas the remaining pollutants are measured in the parts per million (ppm) unit mass of a containment. Most of the time, the PM_{10} and O_3 values dominate the API due to their high value (Al-Dhurafi et al. 2018) and in overall, the average PM$_{10}$ value in Malaysia exceeds the standard set by Recommended Malaysian Ambient Air Quality Guideline (RMAAQG) (Rahim et al. 2023). The PM_{10} is mainly contributed by dust, waste burnings, wildfires and industrial sources (California Air Resources Board) which further contribute to the haze episodes.

Usually, air quality data are investigated using time series analysis (Abdulali & Masseran, 2021; Gourav et al. 2020; Liu & Yuan 2023; Liu et al. 2018), machine learning (Bakar et al. 2022; Leong et al. 2020; Mun et al. 2022) or stochastic analysis (Alyousifi, Masseran & Ibrahim 2018; Alyousifi et al. 2020). However, recent studies have examined the air quality data especially the API in the context of intensity, duration, and severity (Ismail & Masseran 2023; Masseran 2022, 2021; Masseran & Safari 2020a, 2020b). In this paper, we attempt to quantify the effect of haze on the API values using time series analysis with addition of an exogenous variable. Two ways of modelling API with haze effect are time series regression (TSR) with haze dummy variable and regression with autoregressive integrated moving average (ARIMA) errors or also known as ARIMAX. The ARIMAX model is an extension to the commonly used time series model which is the ARIMA model. However, it is well-known that ARIMA-based models have greater flexibility over traditional time series regression. One advantage of an ARIMAX model over an ARIMA model is that an ARIMAX model can estimate the effect of exogenous variables. One can also consider seasonal ARIMAX models (SARIMAX), but it is inefficient when dealing with daily API datasets and hence, not included in this study. SARIMAX models work better when the season is weekly or monthly. However, converting the daily data into weekly data by taking an average of every seven API values will distort the original API values data which exceeds 200. Therefore, it is not advisable to use other than daily API data to investigate the effect of haze.

In this study, the presence of haze acts as an exogenous variable. The usage of ARIMAX in describing air quality data is not new. However, the exogenous variable used may vary from research to research. Liu et al. (2018) has considered particulate matter less than 2.5 microns in size ($PM_{2.5}$), O_3 and NO_2 as the exogenous variables in explaining the air quality in Hong Kong. In Malaysia, the concentration of $PM_{2.5}$ has only been collected since 2017 and our dataset ends at 30th December 2016. Furthermore, each component directly involves in the calculation of API, therefore, it is not advised to include any of the components as the external regressors. Taşpınar (2015) on the other hand, used air temperature and residential natural gas consumption as the exogenous variables in explaining PM_{10} and SO_2. In this paper, we aim to quantify the effect of haze episodes on the API values. The paper is organized as follows. Next section explains about the air pollution data and haze episodes. Subsequent section describes pre-analysis such as separating the dataset into training and test datasets as well as the stationarity tests. It also discusses the ARIMA and ARIMAX modelling. In the section that follows, we discuss the results of pre-analysis and model fittings along with model adequacy measures. Last section concludes the study.

AIR POLLUTION DATA

Figure 1 shows the flowchart for calculating API values (Department of Environment 1997; Masseran 2022). From Figure 1, it can be noted that the API are calculated by taking the maximum values between five standardized indices, one from each observed pollutant.

A trivial indicator of the presence of haze is the high value API because SO$_2$, NO$_2$ and SO$_3$ are contributed by haze. API is a positive real number which takes on from 0 to ∞. Malaysian Department of Environment (2019) has provided a simple guidance on the API levels, as presented in Table 1.

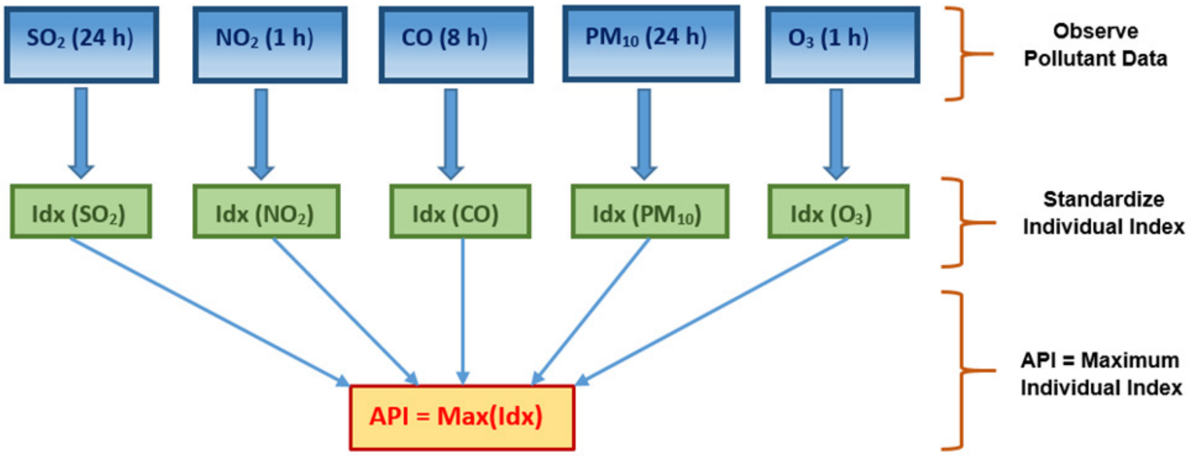

FIGURE 1. Flowchart of API calculations

TABLE 1. API levels and its descriptions

API	Descriptions
0 – 50	Good
51 – 100	Moderate
101 – 200	Unhealthy
201 – 300	Very unhealthy
301 – 500	Hazardous
> 500	Emergency

Source: Department of Environment (2019)

The Malaysian Department of Environment (1997) has also outlined the formulae used to calculate the standardized individual index for each pollutant as follows:

$$Idx(SO_2) = \begin{cases} SO_2 \times 2500 & ; \quad \text{if } SO_2 < 0.04\ ppm \\ 100 + \left[(SO_2 - 0.04) \times 384.61\right] & ; \text{ if } 0.04 \le SO_2 < 0.30\ ppm \\ 200 + \left[(SO_2 - 0.30) \times 3233.333\right] & ; \text{ if } 0.30 \le SO_2 < 0.60\ ppm \\ 300 + \left[(SO_2 - 0.60) \times 500\right] & ; \quad \text{if } SO_2 \ge 0.60\ ppm \end{cases},$$

$$Idx(NO_2) = \begin{cases} NO_2 \times 588.23529 & ; \quad \text{if } NO_2 < 0.17\ ppm \\ 100 + \left[(NO_2 - 0.17) \times 232.56\right] & ; \text{ if } 0.17 \le NO_2 < 0.60\ ppm \\ 200 + \left[(NO_2 - 0.60) \times 166.667\right] & ; \text{ if } 0.60 \le NO_2 < 1.20\ ppm \\ 300 + \left[(NO_2 - 1.20) \times 250\right] & ; \quad \text{if } NO_2 \ge 1.20\ ppm \end{cases},$$

$$Idx(CO) = \begin{cases} CO \times 11.11111 & ; \quad \text{if } CO < 9\ ppm \\ 100 + \left[(CO - 9) \times 16.66667\right] & ; \text{ if } 9 \le CO < 15\ ppm \\ 200 + \left[(CO - 15) \times 6.66667\right] & ; \text{ if } 15 \le CO < 30\ ppm \\ 300 + \left[(CO - 30) \times 10\right] & ; \quad \text{if } CO \ge 30\ ppm \end{cases},$$

$$Idx(PM_{10}) = \begin{cases} PM_{10} & ; \quad \text{if } PM_{10} < 50\mu g/m^3 \\ 50+[(PM_{10}-50)\times 0.5] & ; \quad \text{if } 50 \le PM_{10} < 350\mu g/m^3 \\ 200+[(PM_{10}-350)\times 1.4286] & ; \quad \text{if } 350 \le PM_{10} < 420\mu g/m^3 \\ 300+[(PM_{10}-420)\times 1.25] & ; \quad \text{if } 420 \le PM_{10} < 500\mu g/m^3 \\ 400+[PM_{10}-500] & ; \quad \text{if } PM_{10} \ge 500\mu g/m^3 \end{cases}$$

$$Idx(O_3) = \begin{cases} O_3 \times 1000 & ; \quad \text{if } O_3 < 0.2\ ppm \\ 200+[(O_3-0.2)\times 500] & ; \quad \text{if } 0.2 \le O_3 < 0.4\ ppm. \\ 300+[(O_3-0.4)\times 1000] & ; \quad \text{if } O_3 \ge 0.4\ ppm \end{cases}$$

From 2017, the concentration of particulate matters less than 2.5 microns in size ($PM_{2.5}$) is also considered in developing API values (Department of Environment 2019). For this study, Klang city in Peninsular Malaysia is selected because of its dense population and an elevated economic activities. Hence, the Klang city suffers from frequent unhealthy air pollution (Masseran & Safari 2020a). Figure 2 shows a time series plot for the daily API values in Klang city from January 1st, 1997, until December 30th, 2016. The green dotted line shows the cutoff between moderate and unhealthy API. The blue dotted line shows the cutoff between unhealthy and very unhealthy API. The orange dotted line shows the cutoff between very unhealthy and hazardous API. Finally, the red dotted line shows the cutoff between hazardous and emergency API. As seen from Figure 2, Klang city does frequently have unhealthy API and occasionally, the API skyrockets over 300. Between 1997 and 2016, Klang city is fortunate to not experience emergency level API. Even though Klang city is very prone to air pollution, its API values has never exceeded the emergency level. The Malaysian Department of Environment (2021) has provided a comprehensive chronology of haze episodes in Malaysia from 1997 to 2016, which is tabulated in Table 2. Figure 3 shows the incorporation of the haze episodes and the API values. From Figure 3, it is evident that the occurrence of haze does affect the API values, especially when the API values exceed 200. Therefore, it is crucial to include the haze effect when modelling the API values.

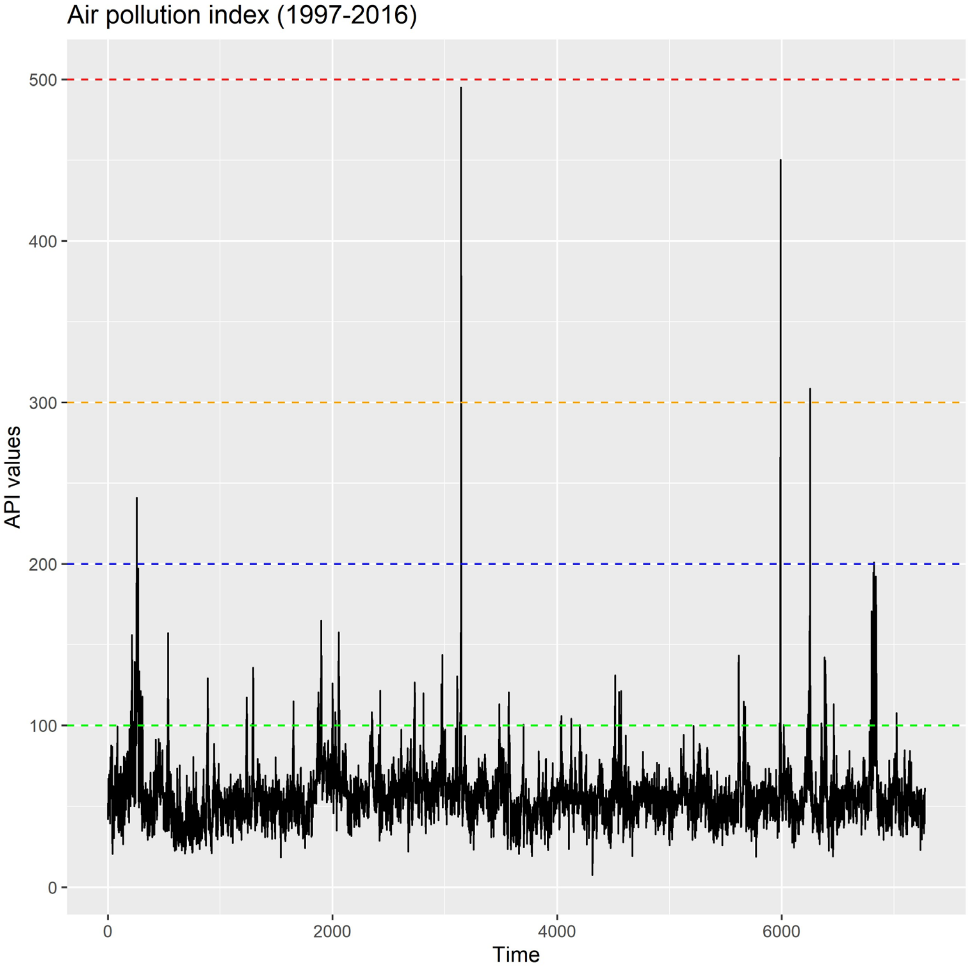

FIGURE 2. Time series plot for API of Klang, Malaysia

TABLE 2. Haze period from 1997 to 2016

Date	Occurrence of haze	Length of days	Cumulative length of days
01/01/1997 – 31/08/1997	No	243	243
01/09/1997 – 30/11/1997	Yes	91	334
01/12/1997 – 31/07/2005	No	2800	3134
01/08/2005 – 13/08/2005	Yes	13	3147
14/08/2005 – 16/07/2006	No	337	3484
17/07/2006 – 19/07/2006	Yes	3	3487
20/07/2006 – 20/09/2006	No	63	3550
21/09/2006 – 10/10/2006	Yes	20	3570
11/10/2006 – 30/04/2011	No	1633	5203
01/05/2011 – 30/09/2011	Yes	153	5356
01/11/2011 – 31/05/2012	No	244	5600
01/06/2012 – 31/08/2012	Yes	92	5692
01/09/2012 – 14/06/2013	No	287	5979
15/06/2013 – 27/06/2013	Yes	13	5992
28/06/2013 – 31/01/2014	No	218	6210
01/02/2014 – 31/03/2014	Yes	59	6269
01/04/2014 – 21/06/2014	No	82	6351
22/06/2014 – 24/07/2014	Yes	33	6384
25/07/2014 – 16/09/2014	No	54	6438
17/09/2014 – 12/10/2014	Yes	26	6464
13/10/2014 – 31/07/2015	No	292	6756
01/08/2015 – 30/09/2015	Yes	61	6817
01/10/2015 – 30/12/2016	No	457	7274

METHODOLOGY

PRE-ANALYSIS

Prior to modelling the API time series data, the data from 01/01/1997 to 30/12/2016 (7274 data points) is separated into training set (7264 data points) and test set data (10 data points). Since ARIMA-based models are great for short-term forecasting, it is reasonable to only select the latest 10 data points for the forecasting. The stationarity of the training set data is investigated using unit root tests such as Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. In ADF test, the hypotheses are:

H_0: The time series is not stationary

H_0: The time series is stationary

In KPSS test, the hypotheses are:

H_0: The time series is stationary

H_0: The time series is not stationary

Using these above tests, a time series is said to be stationary if the null hypothesis in the ADF test is rejected but fail to reject the null hypothesis in the KPSS test.

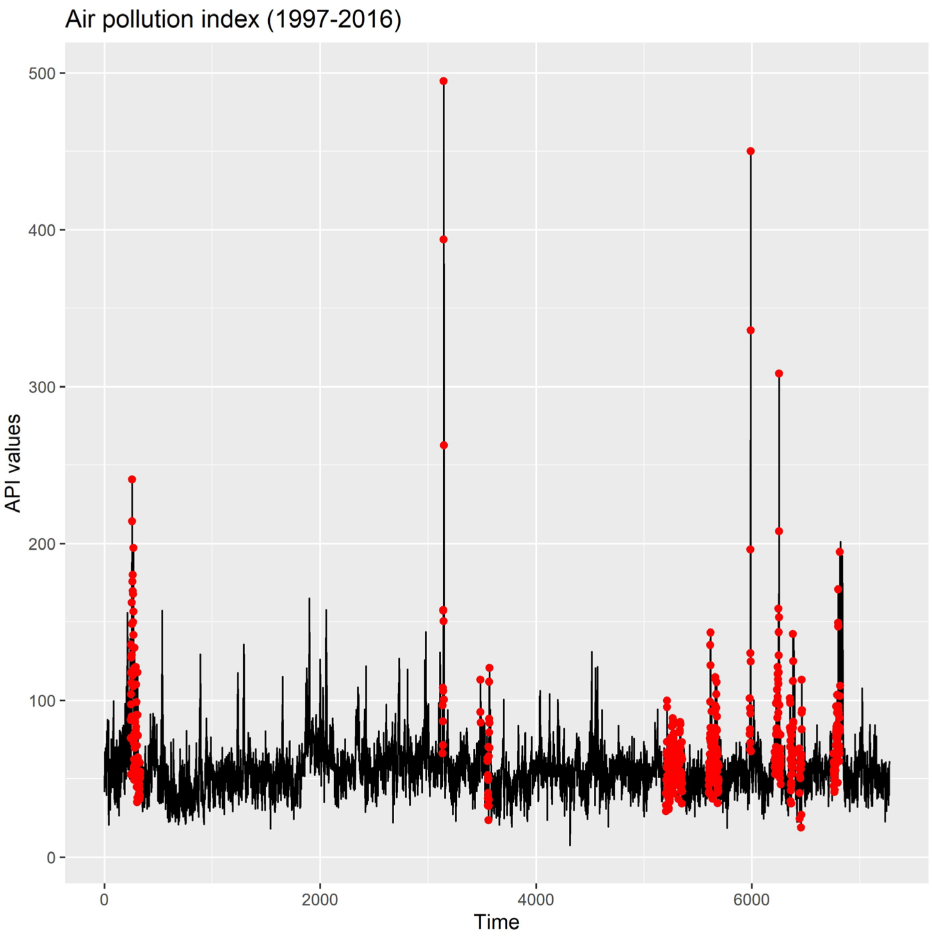

FIGURE 3. Time series plot for API of Klang, Malaysia with haze episodes (in red dots)

ARIMA AND ARIMAX MODELLING

ARIMA model

The common time series analysis involves autoregressive integrated moving average (ARIMA) model with orders *p,d,q*. An ARIMA (*p,d,q.*) model which can be written as:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \;,\; \varepsilon_t \sim \text{NID}\left(0, \sigma^2\right). \tag{1}$$

where *d* refers to the order of integration or differencing.

Regression with ARIMA errors or ARIMAX model

One way of quantifying the effect of haze using the API time series is by employing regression with ARIMA errors. In general, a regression model with ARIMA errors or hereon as ARIMAX model for a time series y_t explained by *k* predictors $x_{i,t}$ for i = 1, 2, ..., *k* can be written as:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k x_{k,t} + \eta_t \;; \\ \eta_t &= \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \ldots + \phi_p \eta_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \\ &\quad + \ldots + \theta_q \varepsilon_{t-q} \\ &\quad \varepsilon_t \sim \text{NID}\left(0, \sigma^2\right). \end{aligned} \tag{2}$$

An ARIMAX model, in general can be written as:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k x_{k,t} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots \\ &\quad + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \;; \\ &\quad \varepsilon_t \sim \text{NID}\left(0, \sigma^2\right). \end{aligned} \tag{3}$$

The regression model with ARIMA errors and ARIMAX model may be mathematically equivalent, but ARIMAX model has slight difficulty in terms of interpretation (Hyndman 2022). Furthermore, in regression with ARIMA errors, regression is conducted first, and the residuals are modelled using ARIMA procedures. On the other hand, in ARIMAX model, the exogenous variables are fitted along with the ARIMA components. Hyndman (2022) mentioned that the β_i coefficients should be interpreted conditional to lagged y_t values.

However, it is common to refer one with the other and hence, we have opted to do the same. In this study, the form in (2) is used and referred as ARIMAX using the 'forecast' package (Hyndman et al. 2020). For this study, only one predictor which is the presence of haze, h_t will be considered and h_t is a dummy variable with indicator function which indicates one if haze presents at time t and zero vice-versa. The general formula for the model can be written as:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 h_t + \eta_t \;; \\ \eta_t &= \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \ldots + \phi_p \eta_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \\ &\quad + \ldots + \theta_q \varepsilon_{t-q} \\ &\quad \varepsilon_t \sim \text{NID}\left(0, \sigma^2\right). \end{aligned} \tag{4}$$

The η_t can be considered the as the remaining API values unexplained by the haze episodes.

TSR model

Time series regression is another technique to quantify the haze effect in API data. General TSR model with haze dummy and polynomial trend can be written as:

$$y_t = \beta_0 + \sum_{p=1}^{k} \beta_p t^p + \beta_{k+1} h_t + \varepsilon_t \;. \tag{5}$$

If the API is found to be stationary, a simple TSR model with haze dummy and no trend will be considered for modelling the API data and quantifying haze effect. The model can be written as:

$$y_t = \beta_0 + \beta_1 h_t + \varepsilon_t \;. \tag{6}$$

Prophet model

Prophet model is an open-source framework by Facebook which takes into account linear additive between growth, seasonality and influence of holidays (Taylor & Letham 2018). A general additive Prophet model is defined as:

$$y_t = g\left(t\right) + s\left(t\right) + v\left(t\right) + \varepsilon_t \;, \tag{7}$$

where $g(t)$, $s(t)$ and $v(t)$ refer to growth, seasonality, and holidays influence components, respectively. The 'prophet' package (Prophet 2022) is used for fitting the Prophet model. The growth, referring to the trend can be either linear, logistics or flat. In the case of stationary data, flat trend seems more plausible than linear or logistics. The seasonality component, which can take on daily, weekly and yearly seasonality, incorporates Fourier series for more flexibility periodic effects (Taylor & Letham 2018). In this study, we do not introduce any holiday influence component, but an additional haze effect is considered. So, one may deduce that $v(t) = h_t$.

Comparison between ARIMA, ARIMAX, TSR and Prophet models

The best fitted model from each modelling approach is obtained and compared in terms of Akaike's Information Criterion (*AIC*), corrected *AIC* (*AICc*) and Bayesian Information Criterion (*BIC*). The formulae for *AIC* (Akaike, 1974), AIC_C (Sugiura 1978) and *BIC* (Schwarz 1978) are respectively given a:

$$\begin{aligned} &AIC = -2\ln L + 2k \ ; \\ &AIC_C = AIC + \frac{2k(k+1)}{n-k-1} \ ; \\ &BIC = -2\ln L + k\ln n \ , \end{aligned} \tag{8}$$

where ln *L* is the log-likelihood values; *n* is the length of training data; and *k* is the number of estimated parameters. The four models are also compared by calculating error measurements for in-sample and out-sample using the root mean squared error values, *RMSE*, the mean absolute error values, *MAE*, the mean absolute percentage error values, *MAPE* and the weighted mean absolute percentage error values, *WMAPE*. The formulae for error measurements are given, respectively, as:

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(y_t - \hat{y}_t\right)^2} \ ;$$

$$MAE = \frac{1}{m}\sum_{t=1}^{m}\left|y_t - \hat{y}_t\right| \ ;$$

$$MAE = \frac{1}{m}\sum_{t=1}^{m}\left|y_t - \hat{y}_t\right| \ ;$$

$$WMAPE = 100\% \frac{\sum_{t=1}^{m}\left|y_t - \hat{y}_t\right|}{\sum_{t=1}^{m} y_t} \ ;$$

$$RMSLE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left[\ln\left(y_t + 1\right) - \ln\left(\hat{y}_t + 1\right)\right]^2};$$

where *m* can be either 7264 (in-sample) or 10 (out-sample); $\hat{y}_t$ is the estimated y_t at time $t \le m$. The best model among ARIMA, ARIMAX, TSR, and Prophet models is selected based on the smaller values of *RMSE,MAE,MAPE,WMAPE* and *RMSLE*.

RESULTS

PRE-ANALYSIS

The stationarity of the training API data is tested using ADF and KPSS tests. The results from the unit root tests are presented in Table 3. From Table 3, it is clear that the training dataset is stationary.

TABLE 3. Results from unit root tests

Unit root tests	p-value	Decision	Conclusion
ADF	0.0100	Reject null hypothesis	The training API data is stationary
KPSS	0.1000	Fail to reject null hypothesis	

MODELLING

The *auto.arima*(┤) function from the library 'forecast' by Hyndman et al. (2020) in R software (R Core Team 2022) is used to obtain the best model for the ARIMA and for the ARIMAX models. For the ARIMAX model, additional information as '*xreg*' will be passed into the *auto.arima*(┤) function. The '*xreg*' argument will allow us to add the presence of haze as the exogenous variable in the original ARIMA model, yielding an ARIMAX model. For ARFIMA and ARFIMAX models, the arfima(┤) function from the same library 'forecast' will be used. For all models, the arguments '*approximation*' and '*stepwise*' are set to 'FALSE' to allow comprehensive model fittings.

ARIMA model

Table 4 summarizes the model fittings with its associated AIC_C values. The ARIMA (4,0,1) with non-zero mean model is found to be the best model in describing the training API data. The fitted model can be written as:

$$y_t = 56.8634 + 1.7964 y_{t-1} - 1.0145 y_{t-2} + 0.3027 y_{t-3}$$

$$- 0.0914 y_{t-4} + \varepsilon_t - 0.9352 \varepsilon_{t-1} \ ,$$

$$\varepsilon_t \sim \text{NID}\left(0, 141.37\right).$$

TABLE 4. The AIC_C for various ARIMA models

Model	AIC_c	Model	AIC_c
ARIMA (0,0,0) with zero mean	80173.17	ARIMA (1,0,4) with non-zero mean	56615.35
ARIMA (0,0,0) with non-zero mean	64174.55	ARIMA (2,0,0) with zero mean	57577.54
ARIMA (0,0,1) with zero mean	71923.67	ARIMA (2,0,0) with non-zero mean	56824.47
ARIMA (0,0,1) with non-zero mean	59302.49	ARIMA (2,0,1) with zero mean	57525.47
ARIMA (0,0,2) with zero mean	67048.87	ARIMA (2,0,1) with non-zero mean	56755.67
ARIMA (0,0,2) with non-zero mean	57878.53	ARIMA (2,0,2) with zero mean	-
ARIMA (0,0,3) with zero mean	64392.38	ARIMA (2,0,2) with non-zero mean	56603.54
ARIMA (0,0,3) with non-zero mean	57410.85	ARIMA (2,0,3) with zero mean	-
ARIMA (0,0,4) with zero mean	62896.79	ARIMA (2,0,3) with non-zero mean	56598.49
ARIMA (0,0,4) with non-zero mean	57177.33	ARIMA (3,0,0) with zero mean	57193.59
ARIMA (0,0,5) with zero mean	61649.67	ARIMA (3,0,0) with non-zero mean	56676.12
ARIMA (0,0,5) with non-zero mean	56991.76	ARIMA (3,0,1) with zero mean	-
ARIMA (1,0,0) with zero mean	57576.59	ARIMA (3,0,1) with non-zero mean	56625.15
ARIMA (1,0,0) with non-zero mean	56878.54	ARIMA (3,0,2) with zero mean	-
ARIMA (1,0,1) with zero mean	57576.61	ARIMA (3,0,2) with non-zero mean	56602.26
ARIMA (1,0,1) with non-zero mean	56798.24	ARIMA (4,0,0) with zero mean	57096.33
ARIMA (1,0,2) with zero mean	-	ARIMA (4,0,0) with non-zero mean	56658.95
ARIMA (1,0,2) with non-zero mean	56676.19	ARIMA (4,0,1) with zero mean	-
ARIMA (1,0,3) with zero mean	-	**ARIMA (4,0,1) with non-zero mean**	**56590.60**
ARIMA (1,0,3) with non-zero mean	56622.92	ARIMA (5,0,0) with zero mean	-
ARIMA (1,0,4) with zero mean	-	ARIMA (5,0,0) with non-zero mean	56645.99

* The best ARIMA model is written in bold

ARIMAX model

Table 5 summarizes the model fittings with its associated AIC_C values. The ARIMAX (4,0,1) with non-zero mean model is found to be the best model in describing the training API data. The fitted model can be written as:

$$y_t = 55.8270 + 13.0460 h_t + \eta_t \ ;$$

$$\eta_t = 1.7983\eta_{t-1} - 1.0163\eta_{t-2} + 0.3087\eta_{t-3} - 0.0964\eta_{t-4}$$

$$+ \varepsilon_t - 0.9467\varepsilon_{t-1}.$$

$$\varepsilon_t \sim \text{NID}\left(0, 140.67\right).$$

TABLE 5. The AIC_C for various ARIMAX models

Model	AIC_c	Model	AIC_c
ARIMAX (0,0,0) with zero mean	79246.76	ARIMAX (1,0,4) with non-zero mean	56584.31
ARIMAX (0,0,0) with non-zero mean	63660.79	ARIMAX (2,0,0) with zero mean	57576.73
ARIMAX (0,0,1) with zero mean	71208.53	ARIMAX (2,0,0) with non-zero mean	5677.18
ARIMAX (0,0,1) with non-zero mean	58972.80	ARIMAX (2,0,1) with zero mean	57524.83
ARIMAX (0,0,2) with zero mean	66601.86	ARIMAX (2,0,1) with non-zero mean	57607.57
ARIMAX (0,0,2) with non-zero mean	57677.93	ARIMAX (2,0,2) with zero mean	-
ARIMAX (0,0,3) with zero mean	64147.94	ARIMAX (2,0,2) with non-zero mean	56573.35
ARIMAX (0,0,3) with non-zero mean	57270.30	ARIMAX (2,0,3) with zero mean	-
ARIMAX (0,0,4) with zero mean	62743.92	ARIMAX (2,0,3) with non-zero mean	56563.95
ARIMAX (0,0,4) with non-zero mean	57061.86	ARIMAX (3,0,0) with zero mean	57184.85
ARIMAX (0,0,5) with zero mean	61525.22	ARIMAX (3,0,0) with non-zero mean	56633.13
ARIMAX (0,0,5) with non-zero mean	61525.22	ARIMAX (3,0,1) with zero mean	-
ARIMAX (1,0,0) with zero mean	57576.01	ARIMAX (3,0,1) with non-zero mean	56595.14
ARIMAX (1,0,0) with non-zero mean	65836.34	ARIMAX (3,0,2) with zero mean	-
ARIMAX (1,0,1) with zero mean	57575.57	ARIMAX (3,0,2) with non-zero mean	56568.98
ARIMAX (1,0,1) with non-zero mean	56748.75	ARIMAX (4,0,0) with zero mean	57087.62
ARIMAX (1,0,2) with zero mean	-	ARIMAX (4,0,0) with non-zero mean	56619.52
ARIMAX (1,0,2) with non-zero mean	56635.48	ARIMAX (4,0,1) with zero mean	-
ARIMAX (1,0,3) with zero mean	-	**ARIMAX (4,0,1) with non-zero mean**	56555.42
ARIMAX (1,0,3) with non-zero mean	56589.48	ARIMAX (5,0,0) with zero mean	-
ARIMAX (1,0,4) with zero mean	-	ARIMAX (5,0,0) with non-zero mean	56609.59

* The best ARIMAX model is written in bold

TSR model

For TSR model, ordinary least square (OLS) technique is utilized which estimates the parameters **β** by minimizing the squared error values. Since the API data is stationary, model in (6) with **β** = (β_0, β_1) will be considered for modelling. The fitted model is written as:

$$\hat{y}_t = 55.3549 + 19.6122h_t.$$

Despite both estimated β_is being significant for i = 0,1, the resulting R-square value showed that there is only 6.8% variation in the data is explained by the fitted TSR model. A vast proportion of the data are still unexplainable by the fitted TSR model.

Prophet model

In Prophet model, the growth rate is set as flat, and the seasonality is set to 'auto' for the API data. Since the API data is a daily-type time series data, the daily seasonality in 'prophet' package (Prophet 2022) is automatically disabled. To consider the haze effect, a dummy variable is considered in place of holiday influence. Furthermore, by considering Table 2, a total of 23 changepoints are considered and included in the fitting command in R software (R Core Team 2022). These changepoints basically shows the possible structural change switching from absence to presence of haze.

Figure 4 shows the components from the Prophet model. Based on Figure 4, the flat trend can be seen

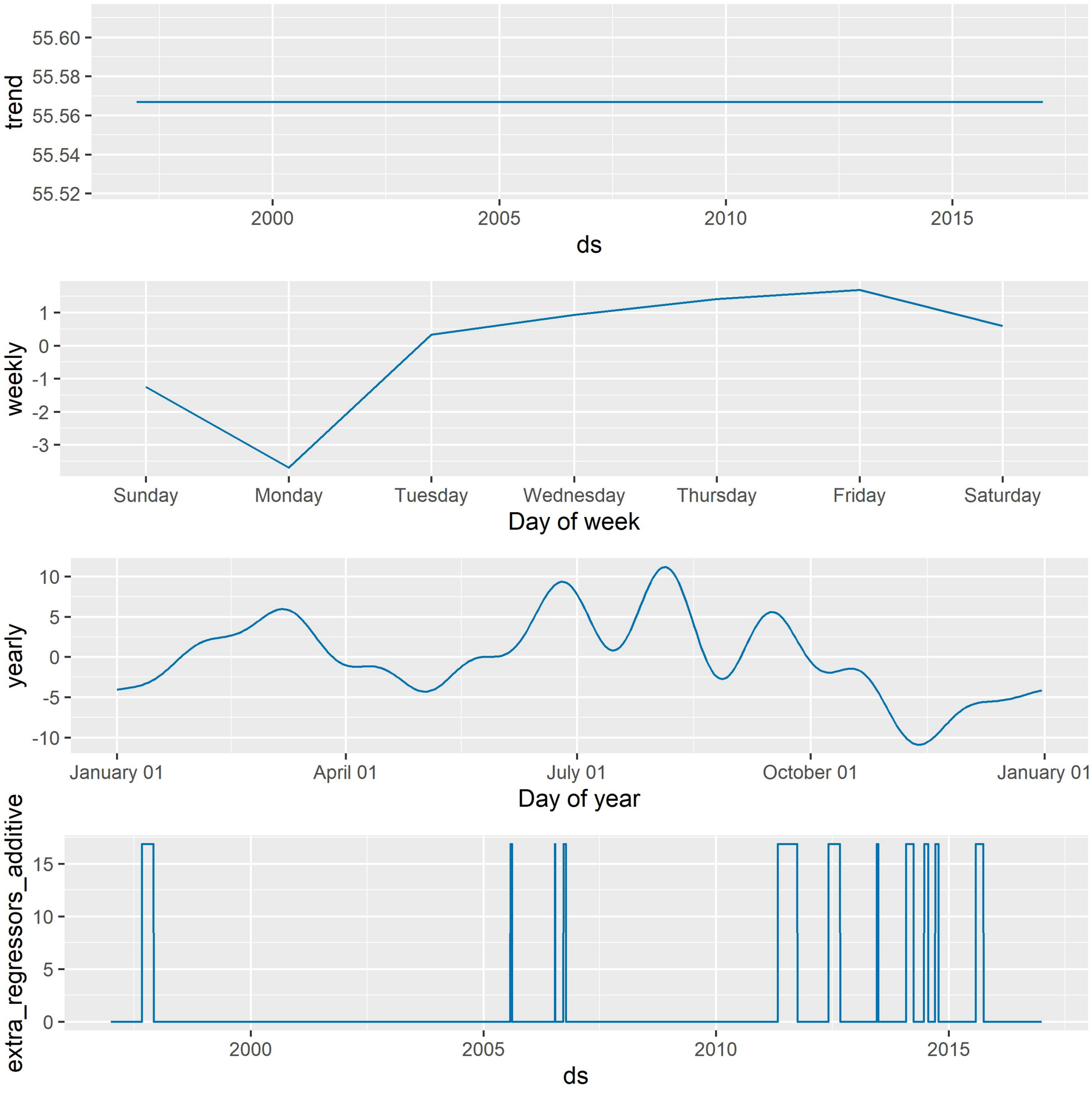

FIGURE 4. Prophet component plot

throughout years but there is a small dip on Mondays and a surge on Tuesdays. Besides that, yearly seasonality shows that high values of API occurs in late Junes, late Augusts, and late Septembers. The effect of the external regressor, which is the presence of haze, can also be noticed in Figure 4. Despite being easy to conduct analysis, the likelihood-based values such as *AIC, AIC_c* and *BIC* are not obtainable. Furthermore, the full model cannot be extracted from the results of the analysis as well.

Comparison between ARIMA, ARIMAX, TSR and Prophet models

Table 6 compares the *AIC, AIC_C*, and *BIC* values from the best ARIMA, ARIMAX, TSR, and Prophet models. The ARIMAX (4,0,1) with non-zero mean model provides a smaller *AIC*, *AIC_C*, and BIC values and hence may be chosen as the better model. Prophet model, despite being new and simpler model than any ARIMA-based models, the log-likelihood values as well as the *AIC, AIC_C*, and *BIC* values cannot be obtained.

Therefore, further comparison between the four best models is required so that an objective decision can be made on the final best model in describing the API data. Despite ARIMAX (4,0,1) with non-zero mean yields smaller *AIC_C* compared to that found from ARIMA (4,0,1) with non-zero mean, it is customary to measure the adequacy and the accuracy of the models using the training data (in-sample) and test data (out-sample). Table 7 shows the results of adequacy measures for in-sample and out-sample data. From Table 7, ARIMAX model gives smallest values of *RMSE,MSE,WMAPE,* and RMSLE for in-sample data and the smallest values of *MAE,MAPE,WMAPE* and *RMSLE* for out-sample data. ARIMA model, on the other hand, gives the smallest *MAPE* for in-sample data and the smallest values of *RMSE* for out-sample data. Both TSR and Prophet models do not perform admirably and adequately in describing the API data. However, Prophet model did perform slightly better than TSR model.

Ultimately, it can be concluded that the ARIMAX model ultimately describes the APIdata adequately and provides the best model fitting followed by ARIMA model, Prophet modeland finally, TSR model. The final best fitted ARIMAX model is given as:

$$y_t = 55.8270 + 13.0460 h_t + \eta_t \ ;$$

$$\eta_t = 1.7983\eta_{t-1} - 1.0163\eta_{t-2} + 0.3087\eta_{t-3} - 0.0964\eta_{t-4}$$

$$+ \varepsilon_t - 0.9467\varepsilon_{t-1} \ ;$$

$$\varepsilon_t \sim \mathrm{NID}(0, 140.67).$$

TABLE 6. The comparisons of the results between the best ARIMA, ARIMAX, TSR, and Prophet models

Model	ARIMA	ARIMAX	TSR	Prophet
Order (*p,d,q*)	(4,0,1)	(4,0,1)	-	-
Mean	Non-zero	Non-zero	-	-
Log-likelihood	-28288.29	**-28269.70**	-31819.5	-
AIC	56590.58	**56555.40**	63645.02	-
AIC_c	56590.60	**56555.42**	63645.02	-
BIC	56638.82	**56610.53**	63665.69	-
*σ*2	141.37	**140.67**	373.65	-

*The better model is written in bold

TABLE 7. The adequacy measures for ARIMA, ARIMAX, TSR and Prophet models

In-sample (training)				
Type of model	ARIMA	ARIMAX	TSR	Prophet
RMSE	11.8852	**11.8548**	19.3479	18.6475
MAE	7.5153	**7.5093**	11.9779	11.5594
MAPE (%)	**13.7236**	13.7286	22.3111	21.3567
WMAPE (%)	13.2130	**13.2026**	21.0590	20.3232
RMSLE	0.1736	**0.1735**	0.2779	0.2647
Out-sample (test)				
RMSE	**7.2882**	7.2987	11.5899	8.6147
MAE	6.4725	**6.4691**	10.0711	7.6788
MAPE (%)	13.4905	**13.4708**	24.0111	17.7980
WMAPE (%)	13.7457	**13.7383**	21.3879	16.3074
RMSLE	0.1462	**0.1461**	0.2398	0.1837

*The smaller error measurement is written in bold

Figure 5 shows the training API dataset and the fitted values from the best ARIMAX model above. From Figure 5, it can be seen that the ARIMAX model still slightly underestimates the API values even during the haze period, especially when the API values are tremendously high. This suggests that there may be some other hidden factors, other than haze, which may have contributed to the high spike in the API values, which warrants further investigation in future.

The presence of haze impose a total of 13.0460 effect on the API values. In general, one can describe that if the haze is present currently, then the current API value will increase by 13.0460. The value of 13.0460 may seem low but it affects the API significantly when the lagged η_ts are considered together. Figure 6 shows the superimposed plot between Figure 3 and Figure 5. From Figure 6 and as explained for Figure 5, even with the presence of haze, the ARIMAX model still underestimate the true API values. This suggests that when the haze is present, the true API values will be much greater than what this ARIMAX model can explain. This serves as a wake-up call for the pertinent authorities such as the Malaysian Department of Environment to equip themselves and public with proper knowledge regarding the dos and don’ts during haze period as well as preparing the distribution of certified protection gears such as N95 to the public.

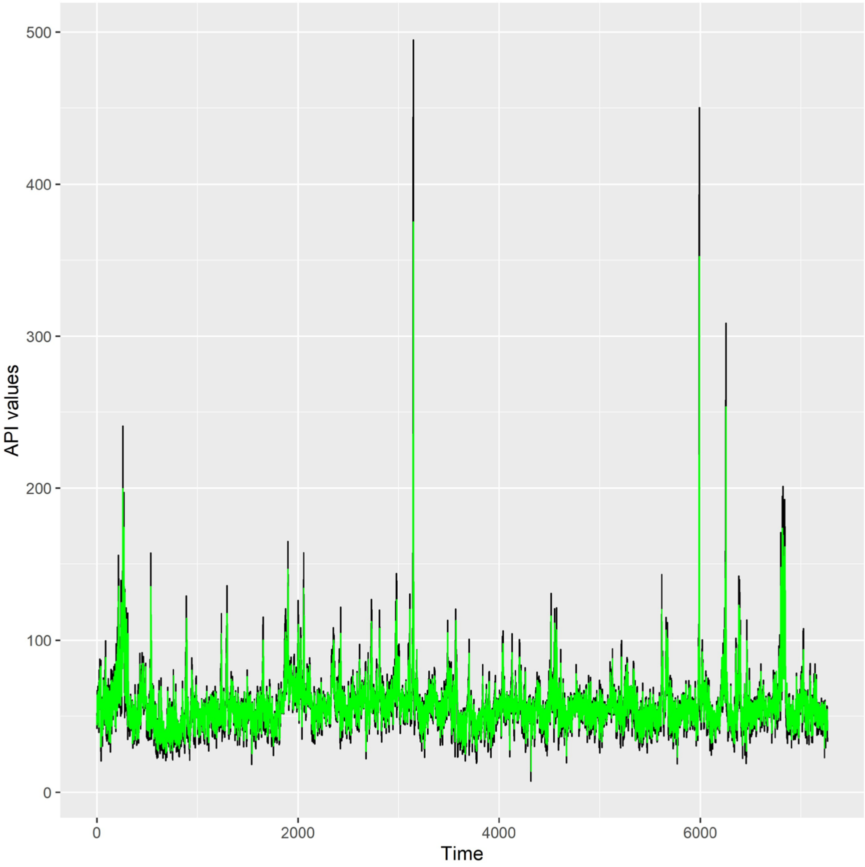

FIGURE 5. Time series plot for API of Klang, Malaysia (in black) and its fitted values from ARIMAX (4,0,1) with non-zero mean (in green)

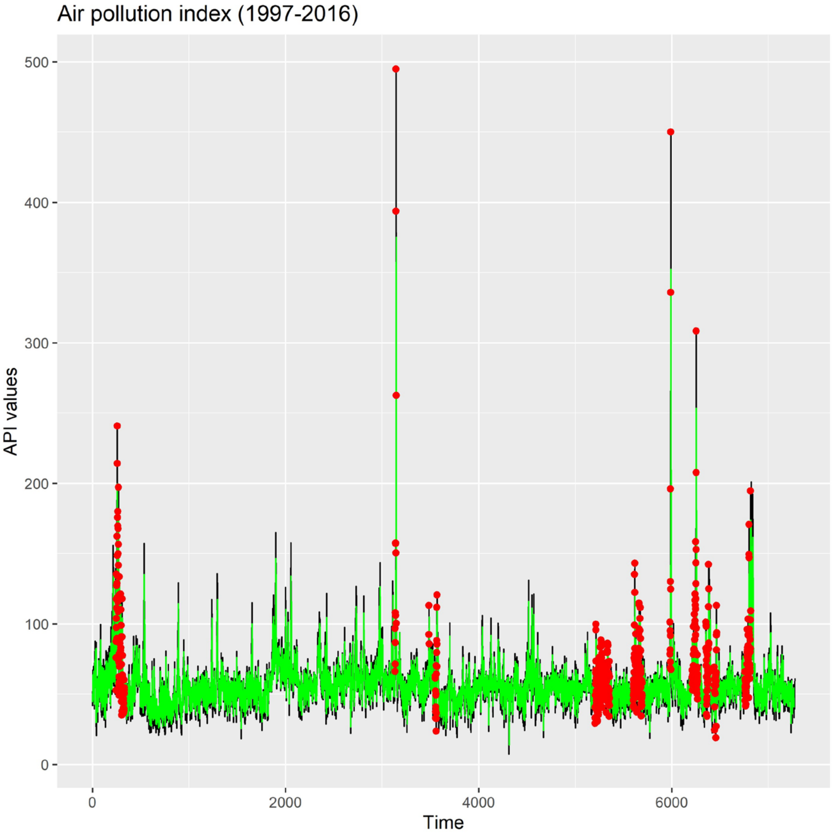

FIGURE 6. Time series plot for API of Klang, Malaysia (in black), its fitted values from ARIMAX (4,0,1) with non-zero mean (in green) and haze presence (in red)

Conclusions

Haze has been a contributing factor to the quality of air in Malaysia, which is determined using air pollution index (API) values. Concentrations of harmful substances in haze are used in quantifying the API values. In Malaysia, haze happens biennially from 2011 (Table 2 & Figure 3). The haze can occur due to man-made or natural catastrophes. The effect of haze episodes was investigated using the API values from Klang City, a city with high population density and vibrant economic activities. Model fittings from ARIMA and ARIMAX procedures showed that the ARIMAX model with order (4,0,1) with non-zero mean is the best model in describing the API values with the presence of haze. The haze is found to affect the API values by 13.0460. The values may seem small at glance, but it is advised to note that the η_t refers to the remaining API values unexplained by the haze occurrences. Both η_t and haze episodes concurrently affect the API values. The presence of haze described by the ARIMAX model indicates that the API values will skyrocket than what the model can predict.

Factors such as the transportation, rapid industrialization and open burnings may affect the haze and it may be fruitful to study the effect of each factor to air quality data rather than haze as a collective factor. One way of studying the effect of these factors is by employing a new ARIMAX model with each factor serves as individual exogenous variable. By doing such, we can identify the most influential factors in contributing to the high values of API. Besides that, it is advised to conduct spatiotemporal analysis on the API values of whole Malaysia to further determine the polluted cities and the polluting cities. By knowing the hotspots of the polluted cities and the polluting cities as well as the influential factors, the Malaysian Department of Environment may take proper course of action through policies or regulations to mitigate the damages to the environment. Besides policies or regulations, general awareness campaigns for the public can be implemented to further sow knowledge and importance of environment for the current and future generations.

All in all, haze does affect humans and we should always take precautionary measures to prevent the haze by not becoming a contributing factor, as well as to survive the haze by wearing protective and quality masks.

Acknowledgments

The authors acknowledge the Dana Impak Perdana grant number DIP-2022-002 and Ganjaran Penerbitan code GP-K020446 funded by Universiti Kebangsaan Malaysia.

References

Abdulali, B.A.A. & Masseran, N. 2021. Artificial Neural Network (ANN) and Arima Models for better forecast of the air pollution data in Malaysia. *Scholars Journal of Physics, Mathematics and Statistics* 10: 184-196.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6): 716-723.

Al-Dhurafi, N.A., Masseran, N., Zamzuri, Z.H. & Razali, A.M. 2018. Modeling unhealthy air pollution index using a peaks-over-threshold method. *Environmental Engineering Science* 35(2): 101-110.

Albahar, S., Li, J., Al-Zoughool, M., Al-Hemoud, A., Gasana, J., Aldashti, H. & Alahmad, B. 2022. Air pollution and respiratory hospital admissions in Kuwait: The epidemiological applicability of predicted $PM_{2.5}$ in arid regions. *International Journal of Environmental Research and Public Health* 19(10): 5998.

Alyousifi, Y., Masseran, N. & Ibrahim, K. 2018. Modeling the stochastic dependence of air pollution index data. *Stochastic Environmental Research and Risk Assessment* 32: 1603-1611.

Alyousifi, Y., Othman, M., Sokkalingam, R., Faye, I. & Silva, P.C. 2020. Predicting daily air pollution index based on fuzzy time series markov chain model. *Symmetry* 12(2): 293.

Bakar, M.A.A., Ariff, N.M., Bakar, S.A. & Ramyah, G. 2022. Peramalan kualiti udara menggunakan kaedah pembelajaran mendalam Rangkaian Perlingkaran Temporal (TCN). *Sains Malaysiana* 51(11): 3785-3793.

California Air Resources Board. *Inhalable Particulate Matter and Health ($PM_{2.5}$ and PM_{10})*. https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health#:~:text=PM10%20also%20includes%20dust%20from,pollen%20and%20fragments%20of%20bacteria. Accessed 13 July 2023.

Department of Environment. 2021. *Kronologi Episod Jerebu di Malaysia*. https://www.doe.gov.my/2021/10/04/kronologi-episod-jerebu-di-malaysia-2/ Accessed 13 July 2023

Department of Environment. 2019. *Air Pollutant Index (API) Calculation.* http://apims.doe.gov.my/pdf/API_Calculation.pdf Accessed 13 July 2023.

Department of Environment. 1997. *A Guide to Air Pollutant Index in Malaysia (API)*. https://aqicn.org/images/aqi-scales/malaysia-api-guide.pdf Accessed on 10 July 2023.

Glover, D. & Jessup, T. 2006. *Indonesia's Fires and Haze: The Cost of Catastrophe*. ISEAS, IDRC.

Gourav, Rekhi, J.K., Nagrath, P. & Jain, R. 2020. Forecasting air quality of Delhi using ARIMA model. *Advances in Data Sciences, Security and Applications*. Lecture Notes in Electrical Engineering, Vol. 612, edited by Jain, V., Chaudhary, G., Taplamacioglu, M. & Agarwal, M. Singapore: Springer

Hyndman, R.J. 2022. *The ARIMAX model muddle*. https://robjhyndman.com/hyndsight/arimax/

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L. & O'Hara-Wild, M. 2020. *Package 'forecast'*. https://Cran.r-Project.Org/Web/Packages/Forecast/Forecast.pdf

Isaifan, R.J. 2023. Air pollution burden of disease over highly populated states in the Middle East. *Frontiers in Public Health* 10: 1002707.

Ismail, M.S. & Masseran, N. 2023. Modeling the characteristics of unhealthy air pollution events using bivariate copulas. *Symmetry* 15(4): 907.

Leong, W., Kelani, R. & Ahmad, Z. 2020. Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering* 8(3): 103208.

Liu, J-B. & Yuan, X-Y. 2023. Prediction of the air quality index of Hefei based on an improved ARIMA model. *AIMS Mathematics* 8(8): 18717-18733.

Liu, S-K., Cai, S., Chen, Y., Xiao, B., Chen, P. & Xiang, X-D. 2016. The effect of pollutional haze on pulmonary function. *Journal of Thoracic Disease* 8(1): E41.

Liu, T., Lau, A.K., Sandbrink, K. & Fung, J.C. 2018. Time series forecasting of air quality based on regional numerical modeling in Hong Kong. *Journal of Geophysical Research: Atmospheres* 123(8): 4175-4196.

Masseran, N. 2022. Power-law behaviors of the severity levels of unhealthy air pollution events. *Natural Hazards* 112(2): 1749-1766.

Masseran, N. 2021. Power-law behaviors of the duration size of unhealthy air pollution events. *Stochastic Environmental Research and Risk Assessment* 35: 1499-1508.

Masseran, N. & Safari, M.A.M. 2020a. Intensity–duration–frequency approach for risk assessment of air pollution events. *Journal of Environmental Management* 264: 110429.

Masseran, N. & Safari, M.A.M. 2020b. Risk assessment of extreme air pollution based on partial duration series: IDF approach. *Stochastic Environmental Research and Risk Assessment* 34: 545-559.

Mohd Nadzir, M.S., Mohd Nor, M.Z., Mohd Nor, M.F.F., A Wahab, M.I., Ali, S.H.M., Otuyo, M.K., Abu Bakar, M.A., Saw, L.H., Majumdar, S. & Ooi, M.C.G. 2021. Risk assessment and air quality study during different phases of COVID-19 lockdown in an urban area of Klang Valley, Malaysia. *Sustainability* 13(21): 12217.

Mun, C., Abd Rahman, N.H. & Ilias, I.S.C. 2022. Performance of Levenberg-Marquardt neural network algorithm in air quality forecasting. *Sains Malaysiana* 51(8): 2645-2654.

Priyankara, S., Senarathna, M., Jayaratne, R., Morawska, L., Abeysundara, S., Weerasooriya, R., Knibbs, L.D., Dharmage, S.C., Yasaratne, D. & Bowatte, G. 2021. Ambient $PM_{2.5}$ and PM_{10} exposure and respiratory disease hospitalization in Kandy, Sri Lanka. *International Journal of Environmental Research and Public Health* 18(18): 9617.

Prophet. 2022. *Automatic Forecasting Procedure*. https://github.com/facebook/prophet

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Rahim, N.A.A.A., Noor, N.M., Jafri, I.A.M., Ul-Saufie, A.Z., Ramli, N., Seman, N.A.A., Kamarudzaman, A.N., Zainol, M.R.R.M.A., Victor, S.A. & Deak, G. 2023. Variability of PM_{10} level with gaseous pollutants and meteorological parameters during episodic haze event in Malaysia: Domestic or solely transboundary factor? *Heliyon* 9(6): e17472.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2): 461-464.

Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by Akaike's. *Communications in Statistics-theory and Methods* 7(1): 13-26.

Taşpınar, F. 2015. Time series models for air pollution modelling considering the shift to natural gas in a Turkish city. *CLEAN–Soil, Air, Water* 43(7): 980-988.

Taylor, S.J. & Letham, B. 2018. Forecasting at scale. *The American Statistician* 72(1): 37-45.

Zhang, Z., Wang, J., Chen, L., Chen, X., Sun, G., Zhong, N., Kan, H. & Lu, W. 2014. Impact of haze and air pollution-related hazards on hospital admissions in Guangzhou, China. *Environmental Science and Pollution Research* 21: 4236-4244.

*Corresponding author; email: rrmt@ukm.edu.my