Mathematical Sciences and Informatics Journal
Vol. 5, No. 2, Nov. 2024, pp. 1-13
http://www.mijuitm.com.my DOI : 10.24191/mij.v5i2.951

Forecasting the Air Pollution Index: A Case Study in Shah Alam, Selangor

Nur Hafiraniza Bakhtiar
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
2020627664@isiswa.uitm.edu.my

Isnewati Ab Malek
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
isnewati@uitm.edu.my

Haslinda Ab Malek
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
haslinda8311@uitm.edu.my

Siti Sarah Januri
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
sarahjanuri@uitm.edu.my

Jaida Najihah Jamidin
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
jaida5698@uitm.edu.my

Article Info

Article history:

Received Apr 09, 2024
Revised Aug 05, 2024
Accepted Sept 8, 2024

Keywords:

Air Pollution Index
Forecast
ARIMA
Univariate Technique
Department of Environment

ABSTRACT

The World Health Organization (WHO) defines air pollution as any chemical, physical, or biological agent that tampers with the atmosphere's natural characteristics and contaminates either the indoor or outdoor environment**.** The evaluation of air pollution can be done by using air pollution prediction. When air pollution levels are high, it can notify and warn the public while assisting the management of many different chemical compounds through policy. The objective of this study is to find the best forecasting model for the air pollution index (API). This study also attempts to predict the monthly mean concentration of the API in Shah Alam for 2023 by using the time series model. To achieve the objectives, the Box-Jenkins Methodology and Univariate Techniques were used. This study examines the API using Holt's Method, Double Exponential Smoothing Technique, and ARIMA models. Based on the smallest value of root mean squared error (RMSE) and mean absolute error (MAE), it shows that the most adequate model for the API for this period is the ARIMA model. Air quality forecasting is reliable and effective in controlling the composition of air pollution. With the ability to forecast the mean concentration of the Air Pollution Index, these findings could aid the Department of Environment in analyzing the substances that contribute to air pollution. Additionally, this information could help reduce the incidence of air pollution-related diseases among Malaysians.

Corresponding Author:
Isnewati Ab Malek
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
Email: isnewati@uitm.edu.my

1. Introduction

The Air Pollution Index (API) is employed in Malaysia to assess air quality. The API system indicated that air pollution consists of five main elements, which are ozone (O3), carbon dioxide (CO2), nitrogen dioxide (NO2), sulphur dioxide (SO2), and particulate matter (PM10). Malaysia was expected to become an industrialized nation in 2020, highlighting the urgent need to address the air quality issue that has been found to exacerbate existing respiratory health conditions [1]. Both natural and human activities contribute to air pollution in the country. Some of the natural causes of air pollution include soil dust, forests, and sea surface emissions. Commonly, these sources contribute to the level of air pollution. However, man-made sources such as the burning of fossil fuels, transportation emissions, and deforestation also contribute significantly to air pollution, apart from causing global climate change [2]. The air quality tended to be worsened due to the high level of man-made sources.

The Malaysian Department of the Environment (DOE) set up the Recommended Malaysian Air Quality Guidelines (RMG) in 1989 to control air pollutants. Subsequently, in 1993 the Malaysian Air Quality Index (MAQI) was introduced to inform the public about air quality conditions. In 1996, they revamped the system and adopted the air pollution index (API), which is modelled after the United States pollutant standard index (PSI) [3]. The API serves as an effective tool for assessing air quality. The API status indicator is categorized into five levels, including good, moderate, unhealthy, very unhealthy, and hazardous, as outlined in Table 1. These categories serve as benchmarks for air quality management and aid in interpreting data for decision-making processes.

Table 1. API Status

API Value	Status
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very Unhealthy
>300	Hazardous

In this study, the API was utilized to determine the best model for forecasting the API from January 2023 to December 2023. Numerous studies showed major cities with high seasonal heating demands, heavy industry, and high vehicular traffic volumes, or with all these three, as the worst air pollution [4].

In 2022, 7.9 million people were living in Selangor. According to the Selangor State Structure Plan 2020, the state's population was anticipated to grow to 9 million people by the year 2035. Around 36,592.52 hectares of land had been designated for development, representing 80% of the total area. This indicates that the state struggled with air quality issues over the years due to population growth and development. Shah Alam is particularly susceptible to air pollution because of its densely populated surrounding areas, significant industrial and commercial developments, and heavy traffic [5]. These features make Shah Alam more vulnerable to air pollution [6]. Therefore, this study aims to identify the best forecasting model for the API and predict the API's monthly mean concentration in Shah Alam for 2023.

2. Literature Review

Air pollution is the presence of harmful substances in the atmosphere that exceed a certain concentration and cause negative effects on both, the ecological system and human life. With the growing concern for the environment, many researchers have conducted numerous studies, with air pollution forecasting being of utmost importance [7]. An accurate forecasting is a foundation for taking any effective pollution control measures, making it a crucial task.

Accurate AQI forecasting is crucial for safeguarding human health, protecting the environment, and supporting economic stability. It enables authorities to make decisions, facilitates sustainable urban planning, and aids in the management of pollution control strategies [8]. Recognizing the significance of the Air Pollution Index (API) and air quality forecasting, this study aims to predict the monthly mean API by using a time series model.

One of the most effective statistical techniques for forecasting from time-series data is the Autoregressive Integrated Moving Average (ARIMA) model, and often referred to as the Box-Jenkins model. ARIMA models are employed to identify the model that best fits the historical data in a time series. This forecasting model has been extensively used across various sectors without limitation to air pollutant time series [9]. Another type of time series model is univariate time series analysis, where a single variable varies over equal time increments in the data. The time increments can be daily, hourly, monthly, or yearly. Univariate modelling, known as a projective approach to forecasting, generates predicted values based on data from previous observations [10]. Like the Box-Jenkins methodology, univariate modelling has been widely used for forecasting in various industries, not just limited to air pollution time series.

In a study which focused on the Thiruvananthapuram District of Kerala, India, the ARIMA and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) methods were employed to forecast air quality indices [11]. The study utilized monthly air quality data from 2012 to 2015 for nitrogen dioxide (NO2), sulphur dioxide (SO2), suspended particulate matter (SPM), and respirable particulate matter (RSPM), collected at four sites in Thiruvananthapuram District. The result indicated that all stations' air quality index readings between 2012 and 2015 fell within the satisfactory (51-100) AQI range. The researchers found that the ARIMA models outperformed SARIMA models in terms of forecasting accuracy.

Another study aimed to compare the performance of artificial neural networks (ANN) and Arima models for a better forecasting the air pollution data in Malaysia [12]. The outcomes highlight the fact that compared to ARIMA, the ANN provided the lowest forecasting error to predict API in Klang. As such, the ANN may be regarded as a reliable predictive method to generate data for the general public regarding the status of air quality at a particular time.

Additionally, a study utilized a time series technique with autoregressive integrated moving average (ARIMA) modelling to predict the maximum daily surface ozone (O3) concentration [13]. The study focused on surface O3 data that was collected at the airport in Brunei Darussalam between July 1998 and March 1999. The fitted ARIMA model had an order of (1,0,1), and it was found that the maximum O3 concentrations predicted by the model closely matched the observed values. The model's effectiveness was evaluated with several widely used statistical metrics.

In a study conducted by [14], it was about the air quality index (AQI) in Miyun County, Beijing, China. The original AQI data were found to be non-stationary during the model construction process. However, the first-order differencing data of the original AQI data were stationary. After comparing various models, the ARIMA (3,1,3) model was selected as the final model for fitting the ARIMA model. The least squares approach was used to represent the data in the Holt exponential smoothing model fitting. In terms of capturing trends and minimizing mean squared error (MSE), Holt modelling outperformed ARIMA modelling in these two model fittings. Therefore, based on this data, the Holt model is preferred for predicting future AQI values.

Air pollution levels reflect the environment's health, and declining air quality immediately affects public health. Air quality forecasting, monitoring, and early warning systems are essential preventive measures for sustainable smart cities, environmental sustainability, and pollution control management [2]. Forecasting air quality is effective in protecting public health by giving early warnings about dangerous air pollutants [7]. Moreover, it aids the Department of Environment (DOE) in planning future monitoring programs and identifying significant air quality changes that could harm the environment and human health. API forecasting enables real-time processing and analysis of air quality data at the network's edge, providing decision-makers with timely and accurate information to address air pollution effectively [15].

Hence, it is clear that modelling and forecasting the Air Pollution Index (API) can be beneficial for various organizations. Therefore, the objective of this research is to develop a model using API time series data from Shah Alam, Selangor. This model aims to provide estimation of future API values, which can be valuable for understanding and managing air quality in the city.

3. Methodology

3.1 Description of Data

This study used secondary data that was extracted from Malaysia's Department of Environment (DOE). The monthly mean concentration of Air Pollution Index (API) of all five main elements which were the particulate matter with a diameter of 10 micrometres or less (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and carbon dioxide (CO_2) from January 2012

until December 2022 was obtained from Malaysia's Department of Environment (DOE) website. A total of 132 months of data on the API value in Shah Alam from January 2012 to December 2022 was applied to forecast the future of Shah Alam's API value.

3.2 Box-Jenkins Methodology

To achieve the objectives of this study, Box-Jenkins were used. The Box-Jenkins method comprises three primary models: autoregressive (AR), integrated (I), and moving average (MA). The AR and MA models are appropriate for analyzing stationary time series patterns. The combination of AR and MA models results in ARMA models. In cases where the data is non-stationary, the *I* model is used to transform the dataset into a stationary form, allowing for the creation of ARIMA models [16].

i. Autoregressive Integrated Moving Average (ARIMA) Model

To better comprehend or predict future values of time series, one could utilize an autoregressive integrated moving average model, or ARIMA model. The Box-Jenkins (1970) ARIMA model was a type of regression analysis that was capable in evaluating the strength of the dependent variable to its independent variable. It had three major processes which were an autoregressive (AR) of order *p*, differencing of degree *d* to render the time-series stationary, and moving average (MA) of order *q*. It was abbreviated as ARIMA (*p*,*d*,*q*) [17]. A simple model case ARIMA (1,1,1) was as shown below,

$$w_t = \mu + \emptyset_1 w_{t-1} - \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (1)$$

where $w_t = y_t - y$ serves the first difference of the series and was considered to be stationary. In this scenario, the values of *p*=1, *d*=1 and *q*=1. The equation could also be written as,

$$(1 - \emptyset_1 B) w_t = \mu + (1 - \theta_1 B) \varepsilon_t \quad (2)$$

The ARIMA model was applied by using the Box-Jenkins framework, which assumes stationarity in the data series. Stationarity means that the statistical properties of the data, such as the mean and variance, remain constant over time. This assumption implies that all instances of the process, regardless of when they occur, exhibit the same statistical characteristics [18].

Differentiation is a method used to transform non-stationary time series data into stationary form. This process involves some calculations of the differences between consecutive observations, which helps stabilize the mean of the series by removing trends and seasonality. Specifically, the first difference is computed as $w_t = y_t - y_{t-1}$, where w_t represents the difference between the current API value, y_t and the previous API value, y_{t-1}. If the first difference indicates that the data series is non-stationary, thus, this transformation is applied to achieve stationarity.

ii. Stages in ARIMA Model Development

The Box-Jenkins modelling approach had four main stages: Model Identification, Model Estimation and Validation, and Model Application as presented in Figure 1.

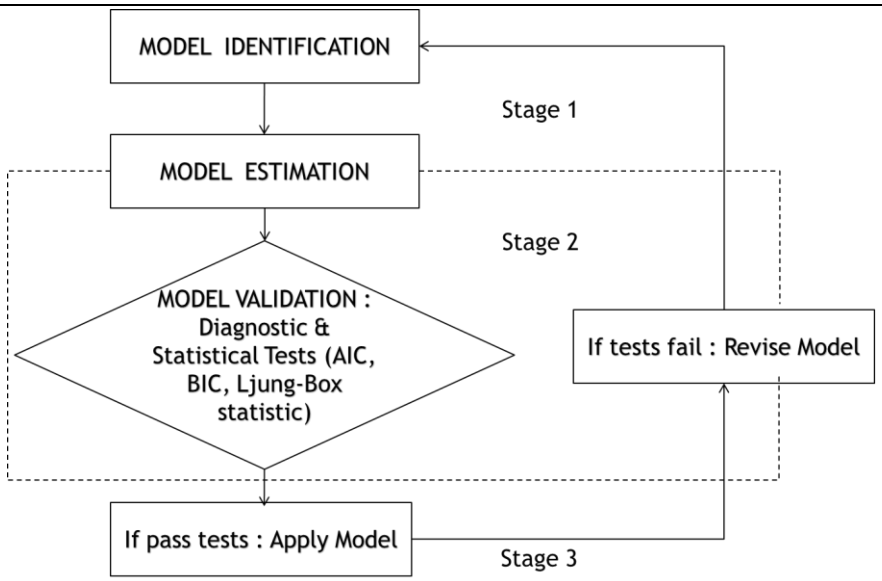

Figure 1. The Stages in ARIMA Model Development

In the Box-Jenkins technique, model identification is the initial step, which involves determining the most suitable class of models for the dataset. Once the data is made stationary, the parameters of the ARIMA model need to be determined. The ARIMA model employed three coefficients, *p*, *d*, and *q*, where *p* was the number of autoregressive terms, *q* denoted the number of moving average terms, and *d* specifies the order of differencing needed for stability. To find the autoregressive (*p*) and moving average (*q*) parameters, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are used. The number of spikes in the PACF indicates p, whereas the number of spikes in the ACF determines q [19].

The next step is model estimation and validation. The ARIMA model was implemented using R-software, which was also employed for forecasting the monthly mean concentration of the Air Pollution Index (API) in Shah Alam, Malaysia. To validate the ARIMA models, statistical measurements such as the Ljung-Box Statistic, Akaike's Information Criteria (AIC), and Bayesian Information Criterion (BIC) were utilized.

The serial correlation of the residuals was assessed by using the Ljung-Box statistic, which helps in determining the adequacy of the model and the randomness of the residuals [20]. The hypothesis of the test is as follows:

H_0 : The errors are random (errors are white noise)

H_1 : The errors are non-random (errors are not white noise)

If the probability value is less than 0.05, the null hypothesis would be rejected, indicating that the model is mis-specified or inadequate. Conversely, if the probability value is more than 0.05, suggesting that the model is adequate.

The AIC and BIC are commonly used to evaluate the fitness of ARIMA. AIC is utilized to analyze various potential models and determine the one that best fits the data. On the other hand, BIC aims to balance model complexity and goodness of fit to produce the most accurate out-of-sample forecast. When both values are low, both criteria show that a model is the best ARIMA model. One similarity between the AIC and BIC was that the model was considered as the best ARIMA model when both values were low [21]. The formula of AIC and BIC are as follows:

$$AIC = e^{\frac{2k}{T}} \frac{\sum_{t=1}^{T} e_t^2}{T} \tag{3}$$

$$BIC = T^{\frac{k}{T}} \frac{\sum_{t=1}^{T} e_t^2}{T} \tag{4}$$

where *T* is the number of observations and *k* is the estimated model's total number of parameters, including the constant.

Once all test conditions are met and the model's fitness is confirmed, it can be applied to obtain forecast values. These values can be represented using confidence intervals or single-value estimates. Confidence interval estimation provides a valuable stochastic measure of the certainty and uncertainty associated with the forecasted values. However, if the test conditions are not satisfied, the model needs to be revised.

3.3 Univariate Techniques

The term univariate solely refers to the forecasts that relies on a sample of time series data for the air pollution index, without any consideration about the influences of other variables such as air temperature and wind direction. In this section, the application of Holt's method and Double Exponential Smoothing method was explained in this study. Consequently, the notations which are utilized in this research are denoted by y_t and *t*, representing the API and month, respectively.

i. Holt's Method

This technique is commonly used to handle data with linear trends because it offers more flexibility in tracking trends and slopes at different rates, besides smoothing them directly using multiple constants. In Holt's method, a final forecast was created by combining three primary equations: an exponential smoothing equation, a trend smoothing equation, and a forecast equation.

The exponentially smoothed series,

$$S_t = \alpha(y_t) + (1-\alpha)(S_{t-1} + T_{t-1}) \quad (5)$$

The trend estimate,

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \quad (6)$$

Forecasts *m* period into the future,

$$F_{t+m} = S_t + T_t \times m \quad (7)$$

where, S_t = smoothed value, S_{t-1} = the smoothed value for the previous period, y_t = smoothed constant for trend estimate ranges from 0 to 1, T_t = trend estimate, T_{t-1} = the trend estimated for the previous period, m = period to be forecast into future, which is 12 months, F_{t+m} = the forecast for the m periods into the future.

ii. Double Exponential Smoothing

Double exponential smoothing is also known as Brown's method. It is useful for a series that has many traits of a linear trend. The primary benefit of double exponential smoothing utilization over single exponential smoothing is the ability to produce multiple forecasts for the future. Generally, four main equations were involved. The notations are used to illustrate this technique as follows:

Let S_t represent the exponentially smoothed value of y_t at time t, and S_t' denote the double exponentially smoothed value of y_t at time t.

The single exponentially smoothed of API,

$$S_t = \alpha(y_t) + (1-\alpha)S_{t-1} \quad (8)$$

The double exponentially smoothed value of API,

$$S_t' = \alpha(S_t) + (1-\alpha)S_{t-1}' \quad (9)$$

The difference between both of exponentially smoothed values,

$$\alpha_t = 2S_t - S_t' \quad (10)$$

The adjustment factor which was forecast for one-year steps ahead,

$$F_{t+m} = a_t + b_t m \quad (11)$$

where F_{t+m} forecast the API at period m made in period t, for $m = 1,2,3,\ldots,12$ and $t = 1,2,3,\ldots,132$.

The accuracy prediction of the model, which was provided by most statistical software, has been compared using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

$$RMSE = \sqrt{\frac{\sum_t^n e_t^2}{n}} \qquad MAE = \frac{1}{n}\sum_{t=1}^{n}|(y_t - \hat{y}_t)|$$

which $e_t = y_t - \hat{y}$, where y_t is the actual value at time t and $\hat{y}$, was the fitted value at time t.

4. Results and Discussion

4.1 Trend of Air Pollution Index (API)

As illustrated in Figure 2, a drawing of time series plot is the initial stage in time series analysis and gives a rough knowledge of the time behaviour of the series. The original series trend seems to have slightly decreased. However, this needs to be verified and proven by using descriptive analysis and trend modelling.

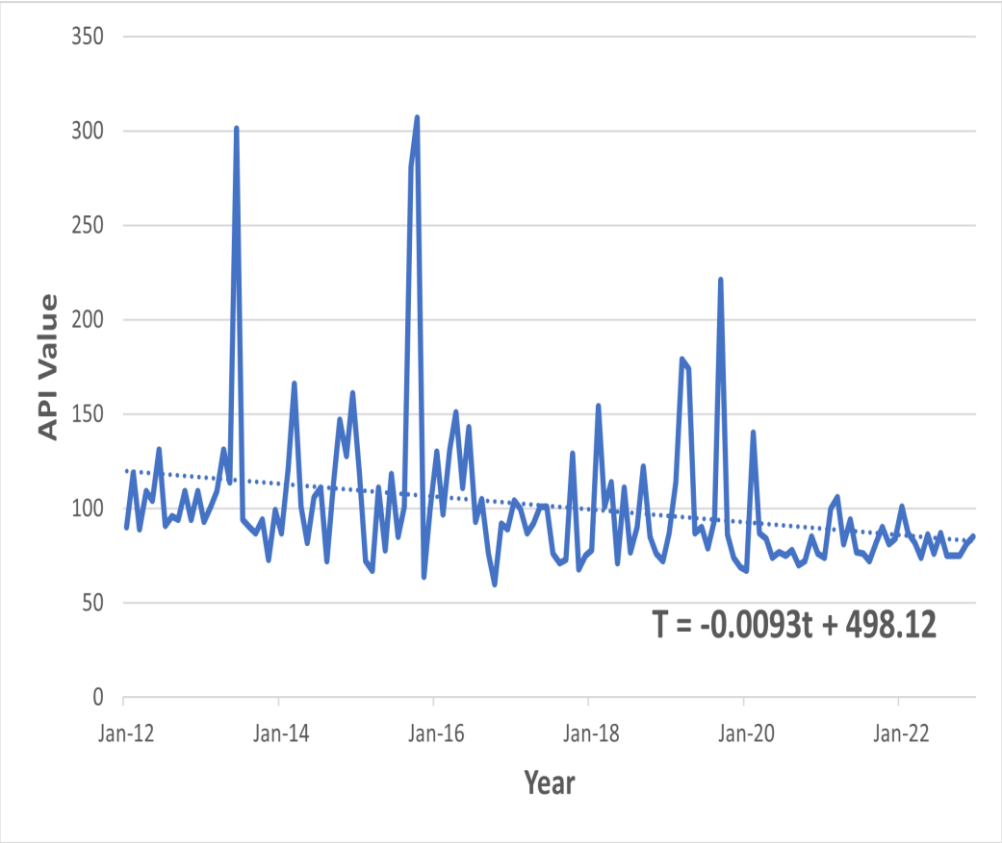

Figure 2. Air Pollution Index (API) Trend Line Graph

Figure 2 illustrates the monthly time series of the Air Pollution Index (API) in Shah Alam, Malaysia, from January 2012 to December 2022. The monthly API fluctuates throughout 132 months with the trend line, T = -0.0093t + 498.12. This equation indicates that as time increases, the API trend decreases by 0.0093. Analysis indicates that the air monitoring station at TTDI Shah Alam in Selangor recorded the highest API reading of 302 in June 2013. API levels exceeding 300 are considered hazardous. In that month also, the monitoring station at Muar experienced its worst, with 663 (emergency level) air quality readings due to haze episodes. Additionally, other locations recorded unhealthy, very unhealthy, and hazardous API levels, indicating widespread poor air quality [22].

Air pollution is expected to be worse during the dry season (June-September) than the wet season (November-March) because there is less rain. More rain during the wet season is expected to improve air quality, as it washes out and reduces the concentration of air pollutants in the

atmosphere [23]. However, in 2015 the highest peak in API records was observed in October, with a hazardous reading of 307. The trend of API good days showed a peak during November–December, which was during the wet season (Figure 2). Meanwhile, the average trend of API unhealthy days also showed the peak during the wet season, which was in February. A downtrend component is evident from the end of 2019 to the end of 2022, likely due to the temporary closure of factories in Shah Alam during the COVID-19 pandemic. Additionally, a turning point occurs from the middle of 2015 to the beginning of 2016, where the API value transitions from a downward trend to an upward trend.

4.2 Box-Jenkins Methodology

In the Box-Jenkins methodology, it is assumed that the characteristics of the initial data series are known. The fundamental assumption is that the data series is stationary. If the series is not stationary, differentiation is necessary to achieve stationarity before proceeding to the next stages.

Figure 3 revealed a stationary and a constant mean in the series. The fluctuation of the line around zero demonstrates its stationary nature. Since the probability value was less than 0.05, the Augmented Dickey-Fuller Test (ADF) revealed that the series was stationary (ADF test statistic = -4.3392, p-value = 0.01). Thus, the series does not require in differencing, and it is said to be integrated of order zero $[I(0)]$. The model obtained is represented in general term as ARIMA (p,d,q) with d=0.

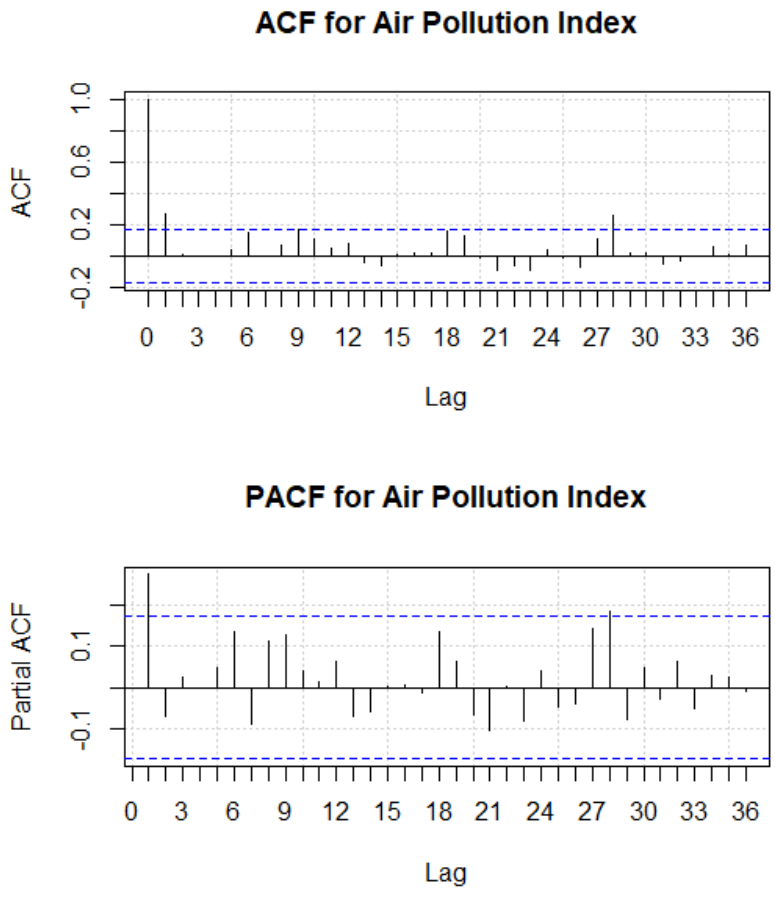

Figure 3. ACF and PACF of Air Pollution Index (API)

Significant spikes were observed respectively in the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for the parameters q and p. The confidence limits in this study were determined to be (-0.174078, +0.174078), although these limits can be varied based on the parameters of the ARIMA models. Identifying the right model formulation was rather difficult

because of the nature of the economic or business data series. Hence, several models to be the best possible formulations were identified and estimated.

In Figure 3 (ACF), two significant spikes were observed at lag 1 and 9 that determine MA(q=2). Meanwhile, Figure 5 (PACF), showed only one significant spike at lag 1, indicating AR(p=1). However, these observations do not guarantee the correct values of p and q for the model. Hence, several model formulations would be estimated in ensuring that a well-specified model was not missed out. The five specified models were ARIMA (2,0,2), ARIMA (1,0,2), ARIMA (1,0,1), ARIMA (1,0,3) and ARIMA (0,1,2).

AIC and BIC were used to assess and select the best ARIMA models, with the lowest values indicating the preferred models. To determine whether a serial autocorrelation occurs in a time series or vice versa, the value of the p-value for each model for the Ljung-Box test statistic was examined. The null hypothesis of the Ljung-Box Test assumes that the results exhibit white noise properties, which indicate no significant serial autocorrelation and satisfy the stationary condition. A model was considered to have no serial autocorrelation if the p-value of the Ljung-Box Test was greater than 0.05. The best model was selected by considering the absence of serial autocorrelation along with other relevant factors. Table 2 displays the AIC, BIC, and p-value of the Ljung-Box test for the selected model.

Table 2. Summary of the Estimated Model

Models	AIC	BIC	Box-Ljung (p-value)
ARIMA (2,0,2)	1032.819	1048.390	0.9980
ARIMA (1,0,2)	1030.926	1043.902	0.9978
ARIMA (1,0,1)	1029.539	1039.920	0.9558
ARIMA (1,0,3)	1032.858	1048.429	0.9983
ARIMA (0,1,2)	1331.503	1340.128	0.9370

Based on the results in Table 2, the p-values of the Ljung-Box test for the selected models were all above 0.05. This indicates that the errors for each model can be adequately represented by white noise, leading to the acceptance of the null hypothesis and suggesting no serial correlation in the model. The best model, ARIMA (1,0,1), was determined from Table 2 as it exhibited respectively the lowest AIC and BIC values of 1029.539 and 1039.920. Additionally, as the Principle of Parsimony favors choosing the simplest model, thus, ARIMA (1,0,1) was deemed the most appropriate selection. Therefore, ARIMA (1,0,1) is recommended as the most accurate model for predicting the Air Pollution Index (API) in Shah Alam for 2023.

4.3 Univariate Techniques

The API data was analyzed using univariate techniques, specifically Holt's method and the double exponential smoothing technique. The goal was to determine the best model, and the results were compared, as shown in Table 3.

Table 3. Error Measure of Holt's Method and Double Exponential Smoothing

Model / Error Measures	RMSE	MAE
Holt's Method	9.9964	8.2513
Double Exponential Smoothing	10.4605	8.2588

The optimal parameters for both models were determined using Excel software. In Holt's method, the parameter values for alpha and beta were respectively found to be 0.5237 and 0.01. Meanwhile, the optimal parameter for alpha was determined to be 0.2473 in the double exponential smoothing technique.

Based on the error measures using these parameter values, Holt's method was deemed more effective for forecasting the API compared to the double exponential smoothing technique. This conclusion was reached by comparing the evaluation components of RMSE and MAE, as shown in Table 3. Therefore, Holt's method is recommended as the best model among the univariate techniques for forecasting the API.

4.4 Determining the Best Model

The reliability of a model is not determined by the existence of a single good forecast. A good forecast model consistently produces good forecast values. Forecasters commonly assess a series of errors that have been produced over time or across time series before making a judgement about the model's quality. A model is therefore deemed superior to others if it meets a set of criteria. It is more customary to measure the error with the smallest magnitude. To determine the most effective method for forecasting the API, the performance of the Box-Jenkins model and univariate techniques were compared, as detailed in Table 4.

Table 4. Best Model Selection

Model	RMSE	MAE
ARIMA (1,0,1)	8.3481	6.7439
Holt's Method	9.9964	8.2513

The evaluated final models were ARIMA (1,0,1) and Holt's method. The error measures for ARIMA (1,0,1) were lower than those for Holt's method, as indicated in Table 4. Specifically, an analysis between the RMSE and MAE values from the evaluation section showed that ARIMA (1,0,1) had much lower values, with 8.3481 and 6.7439, respectively. Therefore, ARIMA (1,0,1) was selected as the best model for forecasting the API in Shah Alam.

4.5 Forecast the Air Pollution Index (API) in Shah Alam

Since ARIMA (1,0,1) model was chosen as the best model for forecasting the API in Shah Alam, it was used to forecast the data for the next 12 months. The future trend forecast is depicted in Figure 4, which appears to be satisfactory at 95 percent confidence interval.

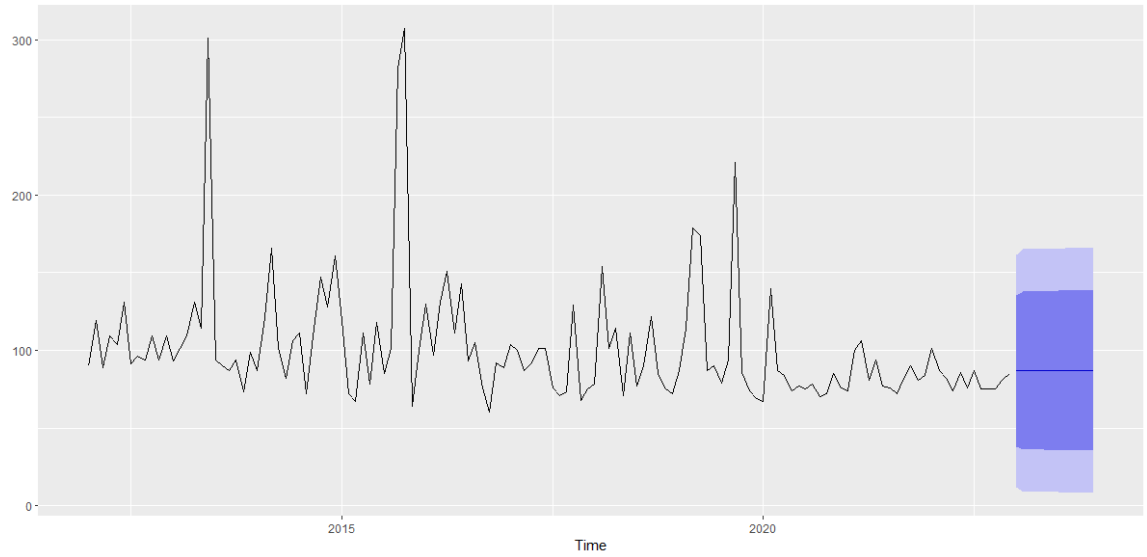

Figure 4. Forecast of API in Shah Alam from January 2023 to December 2023 with ARIMA(1,0,1)

Table 5. Forecasts of API Values for 2023 using ARIMA (1,0,1)

Month	Forecast	95% Confidence Interval
January	82.5010	(43.0386, 121.9634)
February	81.7968	(40.5118, 123.0818)
March	81.6648	(40.0629, 123.2667)
April	81.6400	(39.9345, 123.3454)
May	81.6354	(39.8692, 123.4016)
June	81.6345	(39.8171, 123.4519)
July	81.6344	(39.7680, 123.5008)
August	81.6343	(39.7195, 123.5491)
September	81.6343	(39.6713, 123.5973)
October	81.6343	(39.6231, 123.6455)
November	81.6343	(39.5750, 123.6936)
December	81.6343	(39.5269, 123.7416)

Figure 4 depicts an indication of graph in which the positive outcomes will be under control for the next 12 months or until December 2023. The horizontal axis stands for the number of months from January 2012 onward. The graph also showcases the forecasted monthly Air Pollution Index (API) based on evaluated data up to December 2022. Table 5 displays the specific forecasted values derived from ARIMA (1,0,1). The trend line indicates a consistent decrease, averaging 0.025 percent, suggesting a progressive decline in the API value each month until December 2023. The results imply that the API readings from January to December 2023 will fall within the moderate (51-100) air quality index range by referring to Table 1. Furthermore, all forecasted API values are within 95% confidence intervals.

5. Conclusion

Five ARIMA models were proposed based on the simplest model by estimation value: ARIMA (2,0,2), ARIMA (1,0,2), ARIMA (1,0,1), ARIMA (1,0,3), and ARIMA (0,1,2). Through observation, ARIMA (1,0,1) was found to have the smallest values of AIC and BIC, making it the best model for the Box-Jenkins Methodology. Additionally, Holt's method was identified as the best model for Univariate Techniques, as it exhibited the lowest evaluation values of RMSE and MAE compared to Double Exponential Smoothing. The first goal of the study was to evaluate the best model among the Box-Jenkins Methodology and Univariate Techniques. The results showed that ARIMA (1,0,1) had the lowest value across all error measurements.

The second goal of this study was to forecast the Air Pollution Index in Shah Alam one year ahead, from January 2023 to December 2023 by using the best model. The forecasting was done by using ARIMA (1,0,1) model, and the results revealed that the trend line of anticipated values remained constant at a range of 81.63 and 82.50, which indicates a satisfactory category because the readings fell within the moderate reading (51-100) of API status (Table 1).

For future research, it is recommended to examine how deep learning, or the combination of neural networks and data mining approaches, could enhance the effectiveness of forecasting models. Creating hybrid models that integrate the strengths of different approaches could lead to more accurate forecasts of the air quality index in Shah Alam [24].

Acknowledgements

The authors are grateful to Universiti Teknologi MARA (UiTM), Seremban branch, and Malaysia's Department of Environment (DOE) for providing the data. They also extend their gratitude to other researchers for their valuable ideas and discussions.

Conflict of Interest

The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1] K. Morrissey *et al.*, “The effects of air quality on hospital admissions for chronic respiratory diseases in Petaling Jaya, Malaysia, 2013–2015,” *Atmosphere*, vol. 12, no. 8, p. 1060, Aug. 2021. doi:10.3390/atmos12081060.

[2] S. Subramaniam *et al.*, “Artificial intelligence technologies for forecasting air pollution and human health: A narrative review,” *Sustainability*, vol. 14, no. 16, pp. 9951, Aug. 2022, doi:10.3390/su14169951.

[3] Department of Environment Malaysia (DOE), *A Guide to Air Pollutant Index (API) in Malaysia*. Kuala Lumpur, Malaysia: Department of Environment, 2000.

[4] J. Shen, D. Valagolam, and S. McCalla, “Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea,” *PeerJ.*, vol. 8, pp. e9961, Sep. 2020, doi: 10.7717/peerj.9961.

[5] W. N. W. M. Din and N. Shaadan, “Application of functional time series model in forecasting monthly diurnal maximum API curves: A comparison between multi- step ahead and iterative one-step ahead approach,” *Mal. J. Fund. Appl. Sci.*, vol. 18, pp. 124-137, 2022.

[6] N. Shaadan and W. N. W. M. Din, “Application of functional time series model in forecasting monthly diurnal API curves: A comparison between multi- step ahead and iterative one-step ahead approach,” *Mal. J. Fund. Appl. Sci.*, vol. 18, no. 1, pp. 124-137, Feb. 2022, doi: 10.11113/mjfas.v18n1.2435.

[7] K. S. Wong, Y. J. Chew, S. Y. Ooi, and Y. H. Pang, “Toward forecasting future day air pollutant index in Malaysia,” *The Journal of Supercomputing*, vol. 77, no. 5, pp. 4813–4830, May 2021, doi:10.1007/s11227-020-03463-z.

[8] C. B. Pande *et al.*, “Daily scale air quality index forecasting using bidirectional recurrent neural networks: Case study of Delhi, India,” *Environ. Pollut.*, vol. 351, p. 124040, Jun. 2024, doi: 10.1016/j.envpol.2024.124040.

[9] R. Das, A. I. Middya, and S. Roy, “High granular and short-term time series forecasting of PM 2.5 air pollutant - a comparative review,” *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 1253-1287, Feb. 2022, doi: 10.1007/s10462-021-09991-1.

[10] J. K. Sethi and M. Mittal, “Analysis of air quality using univariate and multivariate time series models,” in *Proc. of the Confluence 2020 -10th Int. Conf. on Cloud Computing, Data Science and Engineering*, 2020, doi: 10.1109/confluence47617.2020.9058303.

[11] V. Naveen and N. Anu, “Time series analysis to forecast air quality indices in Thiruvananthapuram District, Kerala, India,” *Int. J. Eng. Res. Appl.*, vol. 07, no. 06, pp. 66-84, Jun. 2017, doi: 10.9790/9622-0706036684.

[12] B. A. A. Abdulali and N. Masseran, “Artificial Neural Network (ANN) and ARIMA models for better forecast of the air pollution data in Malaysia,” *Sch. J. Phys. Math. Stat.*, vol. 8, no. 10, pp. 184–196, Dec. 2021, doi:10.36347/sjpms.2021.v08i10.001.

[13] K. Kumar, A. K. Yadav, M. P. Singh, H. Hassan, and V. K. Jain, “Forecasting daily maximum surface ozone concentrations in Brunei Darussalam—an ARIMA Modeling approach,” *J. Air Waste Manag. Assoc.*, vol. 54, no. 7, pp. 809-814, Jul. 2004, doi: 10.1080/10473289.2004.10470949.

[14] J. Zhu, “Comparison of ARIMA model and exponential smoothing model on 2014 air quality index in Yanqing County, Beijing, China,” *Appl. Comput. Math.*, vol. 4, no. 6, p. 456, Jan. 2015, doi: 10.11648/j.acm.20150406.19.

[15] M. Ansari and M. Alam, “An intelligent IoT-Cloud-Based air pollution forecasting model using univariate Time-Series analysis,” *Arab. J. Sci. Eng.,* vol. 49, no. 3, pp. 3135–3162, May 2023, doi: 10.1007/s13369-023-07876-9.

[16] S. Nurman, M. Nusrang, and N. Sudarmin, “Analysis of rice production forecast in Maros District using the Box-Jenkins method with the ARIMA model,” *ARRUS J. Math. App. Sci.*, vol. 2, no. 1, pp. 36-48, Feb. 2022, doi: 10.35877/mathscience731.

[17] R. Thapa, S. Devkota, S. Subedi, and B. Jamshidi, “Forecasting area, production and productivity of vegetable crops in Nepal using the Box-Jenkins ARIMA model,” *Turk. J. Agric. - Food Sci. Technol.*, vol. 10, no. 2, 2022, doi: 10.24925/turjaf.v10i2.174-181.4618.

[18] M. A. A. Bakar, N. M. Ariff, M. S. M. Nadzir, O. L. Wen, and F. N. A. Suris, “Prediction of multivariate air quality time series data using Long Short-Term Memory Network,” *Mal. J. Fund. Appl. Sci.*, vol. 18, no. 1, pp. 52–59, Feb. 2022, doi:10.11113/mjfas.v18n1.2393.

[19] Q. Zhou, Z. Chen, Z. Cai, and Z. Xia, “Prediction of the best portfolio for Bitcoin and gold based on the ARIMA model,” *Front. Bus. Econ. Manag.*, vol. 4, no. 3, Aug. 2022, doi: 10.54097/fbem.v4i3.1284.
[20] D. Adedia, S. Nanga, S. K. Appiah, A. Lotsi, and D. A. Abaye, “Box-Jenkins’ methodology in predicting maternal mortality records from a public health facility in Ghana,” *Open J. Appl Sci.*, vol. 08, no. 06, pp. 189-202, Jan. 2018, doi: 10.4236/ojapps.2018.86016.
[21] M. A. Lazim, *Introductory Business Forecasting: A Practical Approach*. Shah Alam, Selangor: UiTM Press, 2011.
[22] N. L. A. Rani, A. Azid, S. I. Khalit, H. Juahir, and M. S. Samsudin, “Air pollution index trend analysis in Malaysia, 2010-15,” *Pol. J. Environ. Stud.*, vol. 27, no. 2, pp. 801–807, Jan. 2018, doi:10.15244/pjoes/75964.
[23] O. L. H. Leh, S. Ahmad, K. Aiyub, Y. M. Jani, and T. K. Hwa, "Urban air environmental health indicators for Kuala Lumpur city," *Sains Malaysiana*, vol. 41, no. 2, pp. 179-191, 2012.
[24] J. W. Koo, S. W. Wong, G. Selvachandran, H. V. Long, and L. H. Son, “Prediction of air pollution index in Kuala Lumpur using fuzzy time series and statistical models,” *Air Qual. Atmos. Health*, vol. 13, no. 1, pp. 77–88, 2020, doi:10.1007/s11869-019-00772-y.

Biography of all authors

Picture	Biography	Authorship contribution
	Nur Hafiraniza binti Bakhtiar is a full-time student studying for a Bachelor of Science (Hons.) in Statistics at UiTM Seremban Campus and graduated in 2023. Her expertise includes Statistical Modelling and Forecasting.	Data analysis and thesis writing
	Isnewati Ab Malek is a Senior Lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her area of specialization includes time series analysis and statistical modelling.	Writing – review and editing, supervision
	Haslinda Ab Malek is a Senior Lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her expertise lies in Environmental Statistics and Statistical Modelling.	Writing – review and editing, validation
	Siti Sarah Januri is a Senior Lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her expertise lies in Operations Research and Stochastic Modelling.	Review and editing
	Jaida Najihah Jamidin is a lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her experties lies in Regression Analysis and Statistical Modelling.	Review and formatting