UNIVERSITI TEKNOLOGI MARA

MATHEMATICS AND STATISTICS

UNDERGRADUATE RESEARCH PROCEEDINGS 2025

UiTM CAWANGAN NEGERI SEMBILAN

Performance of Box Jenkins and Artificial Neural Networks Models on Air Pollution Index in Klang

Nor Akmal Md Noh [1,*], Nur Aqilah Aina Abd Latif [1]
[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Negeri Sembilan Seremban Campus, 70300 Negeri Sembilan*
[*]norak029@uitm.edu.my

Abstract

This study focuses on air pollution levels in Klang Valley, a region frequently experiencing poor air quality due to industrial activity, haze, traffic and urbanization. The performance of Box Jenkins (ARIMA) and Artificial Neural Network (ANN) models was compared using data collected from Malaysia's Department of Environment (DOE) over five years (2019-2023). The ARIMA model analyzed Air Pollution Index (API) as the variable, while the ANN model utilized API as the output variable and six pollutants sulphur dioxide (SO_2), particulate matter (PM_{10}), ($PM_{2.5}$), ozone (O_3), nitrogen dioxide (NO_2) and carbon monoxide (CO) as input variables. Under Box Jenkins procedure, six ARIMA models were tested, with ARIMA (4,1,1) selected as the best based on AIC and BIC. As an ANN model, networks with varying hidden nodes were evaluated, and the model with five hidden nodes was identified as the best, achieving highest R^2 and lowest RMSE. Comparisons between the best ARIMA and ANN models showed that ANN outperformed ARIMA, with lower RMSE (5.5502) and MAE (4.1799). In conclusion, both ARIMA and ANN models effectively predicted API, however an ANN model offering slightly better performance based on the root mean square error.

Keywords*:* ARIMA, ANN, Air Pollution

Introduction

Pollution is a major global environmental issue, with air pollution being the leading cause of death. It consists of gases and particles, and its sources vary by location and time [1]. Air pollution is a growing concern for public health, especially for vulnerable populations such as children and the elderly [2]. In China, heavy industry and coal combustion contribute to air pollution, while in Australia, emissions from coal power stations exceed health safety standards [3,4].

The World Health Organization (WHO) reports that 92% of the global population faces poor air quality, and air pollution related deaths have increased [5]. Studies show that people in more polluted areas face higher mortality risks. Air pollution also negatively affects pregnancy and child development, leading to issues like low birth weight and neurodevelopmental disorders [6].

In Malaysia, air pollution is caused by rapid urbanization, industrialization and motor vehicle emissions [7]. Power plants contribute 85% of the pollution, with transboundary pollution from Indonesia's biomass burning also playing a significant role [8]. The Air Pollution Index (API) of Malaysia measures pollution using six pollutants, including sulphur dioxide (SO_2),

particulate matter (PM_{10}), ($PM_{2.5}$), ozone (O_3), nitrogen dioxide (NO_2) and carbon monoxide (CO) [9].

Klang Valley, a major focus of air pollution research in Malaysia, saw a significant increase in unhealthy days from 2022 to 2023 as shown in Figure 1 compared to other regions; Southern Regions, East Coast, Sabah, Federal Territory Labuan and Sarawak [9]. The region faces severe air quality issues due to industrial activity, traffic congestion and construction along with transboundary haze from forest fires in Indonesia [10,11]. This study aims to analyze the API trend in Klang Station, which has the worst air quality in Malaysia.

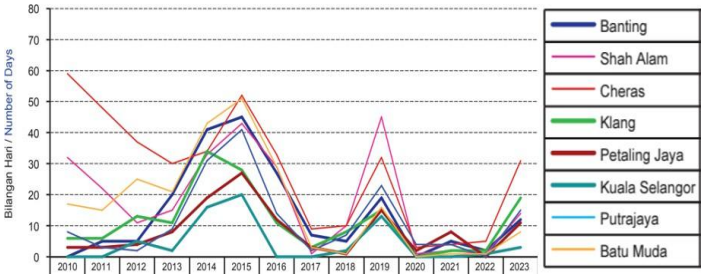


Figure 1: Number of Unhealthy API Days in Klang Valley in Year 2023
Sources: DOE, 2024.

Air pollution related deaths are rising annually. According to the Department of Statistics Malaysia (DOSM), pneumonia was the second leading cause of death in 2023, accounting for 13.3% of deaths, with a 2.2% increase from 2022. Pneumonia, an acute lung infection, inflames the tiny air sacs in the lungs, making it hard to breathe [12]. Children are particularly vulnerable, as they may struggle to fight off infections, leading to long term lung damage and respiratory issues [12].

In Malaysia, areas like Klang, with high API levels, are especially concerning. Klang's industrial activities, dense urbanization, and traffic emissions contribute to severe air pollution, linked to respiratory diseases like pneumonia. Klang's location is in the Klang Valley, surrounded by hills, trapped pollutants, and worsening air quality. Given these concerns, this study aims to develop a statistical model using Box Jenkins and ANN methods to identify the best-performing parameters and compare their effectiveness.

Methodology

The data used in this research were obtained from Malaysia's Department of Environment (DOE), covering the daily mean API and six major pollutants from 1 January 2019 to 31 December 2023. For the ARIMA model, API was used as the univariate input. In the ANN model, API is the output variable, with inputs including SO_2, PM_{10}, $PM_{2.5}$, O_3, NO_2 and CO, as listed in Table 1. Before undergoing analysis, the dataset underwent a data cleaning process and was then partitioned into 70% estimation and 30% evaluation. ARIMA model uses RStudio while ANN model will use JMP PRO17 to perform the result.

Table 1: List of Variables.

Variables	Unit
Sulphur Dioxide, SO_2	parts per million (ppm)
Particulate Matter, PM_{10}	microgram per cubic meter ($\mu g/m^3$)
Particulate Matter, $PM_{2.5}$	microgram per cubic meter ($\mu g/m^3$)
Ozone, O_3	parts per million (ppm)
Nitrogen Dioxide, NO_2	parts per million (ppm)
Carbon Monoxide, CO	parts per million (ppm)

Box Jenkins

The Box Jenkins methodology is a prominent approach for time series analysis, focusing on ARIMA models that combine Autoregressive (AR), Integrated (I) and Moving Average (MA) components. It includes three stages which are model identification, estimation and validation. The process begins with ensuring stationarity, as the Box Jenkins model assumes the data series is stationary. If this assumption is not met, steps must be taken to make the series stationary.

Stationarity is typically checked using the Unit Root Test, with the Augmented Dickey Fuller (ADF) Test being the most common due to its stable critical values, consistent performance across small and large sample sizes, and reliable power properties. The ADF Test operates under the null hypothesis (H_0) that the series has a unit root and is nonstationary, while the alternative hypothesis (H_1) assumes the series is stationary. A p-value less than 0.05 indicates stationarity.

If the data is not stationary, differencing is applied repeatedly until stationarity is achieved. ACF and PACF plots are then used to determine the model order (*p, d, q*). During estimation, models are evaluated for fit using metrics like AIC, BIC, and the Ljung Box test to ensure residuals are white noise. The selected model is then analyzed and refined to capture the underlying patterns and relationships in the data effectively [13].

Model Identification

The ACF and PACF plots help determine the Box-Jenkins model. An AR model is suitable when the ACF decays exponentially and the PACF has spikes, with the order (p) determined by significant PACF spikes. An MA model is preferred when the PACF decays and the ACF has spikes, with the order (q) based on notable ACF spikes. For irregular patterns in both plots, an ARMA model is used, with the order (p, q) corresponding to the spikes in the PACF and ACF, respectively [13]. The ARIMA model is expressed by equation 1 [13].

$$Y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

where Y_t is Klang daily API, μ is constant term, $\phi_1, \phi_2, \ldots, \phi_p$ is an estimated parameters values. $\theta_1, \theta_2, \ldots, \theta_p$ is an estimated moving average parameters values, p is the autoregressive ordered, q is moving average ordered and ε_t refers to error term.

Model Estimation and Validations

Estimating models and conducting diagnostic tests are crucial for selecting the best forecasting model. Key objectives are ensuring the fitted values resemble actual values and using the fewest parameters for a good fit. Three common statistical measures for ARIMA models are the Ljung Box Test, Akaike's Information Criteria (AIC), and Bayesian Information Criteria (BIC). Firstly, Ljung Box Test checks if residuals are white noise. A p-value of the test less than 0.05 suggests model misspecification [13]. The null hypothesis of the Ljung Box Test is the errors are white noise.

Secondly, AIC evaluates model fit, with lower values indicating better fit. The model with the fewest parameters is preferred [13]. The equation 2 represents the AIC equation where k is the number of parameters in the model.

$$AIC = e^{\frac{2k}{T}} \left(\frac{\sum_{t=1}^{T} e_t^2}{T} \right) \tag{2}$$

Thirdly, BIC used to choose models by balancing model complexity and goodness of fit. The model with the lowest BIC is preferred, with a stronger penalty for more parameters [13]. The equation 3 represents the BIC where k denotes the number of parameters in the model, T refers to the number of observations.

$$BIC = T^{\frac{k}{T}} \left(\frac{\sum_{t=1}^{T} e_t^2}{T} \right) \tag{3}$$

Model Application

Once the test criteria are met and the model's fit is confirmed, it can generate forecast values. These forecasts are presented as single values. The model formulation and estimation process are repeated, refining it each time, until the best model is identified. The next step is to establish a system for monitoring forecasts. If new information shows the model is no longer accurate, it must be revised and updated.

Artificial Neural Network

An Artificial Neural Network (ANN) maps input variables (input layer) to output variables (output layer) through a learning process. It uses basic processors, called neurons, to identify relationships within data. Inspired by the biological structure of the human brain, ANNs are designed to learn, generalize, and make decisions [14].

ANNs are feed-forward neural network systems where neurons process inputs, recognize patterns, and perform nonlinear classification tasks [15]. Mathematically, an ANN can be represented as shown in equation 4 [16].

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \quad \text{and} \quad y_k = \phi(u_k + b_k) \tag{4}$$

Where bias, b_k, influences the net input of the activation function by either increasing or decreasing its value. The inputs are $x_1, x_2, \cdots, x_m$ and the weights of neuron k are $w_{k1}, w_{k2}, \cdots, w_{km}$. The linear combiner output resulting from the input signals is u_k. The activation function is denoted by $\phi(\bullet)$, and the output signal of the neuron y_k.

The ANN used was a multilayer network consisting of an input layer, an output layer, and a hidden layer in between. The hidden layer uses a transfer activation function to introduce non-linearity into the network. The activation of $a_j^{(l)}$ (the j^{th} neuron in the l^{th} layer) is related to the neurons in the (1-1)th layer by the equation 5. The equation 5, where $a_j^{(l-1)}$ is the k^{th} neuron in the (1-1)th layer, n_{l-1} is the total number of neuron in the (1-1)th layer, $w_{jk}^{(l)}$ is the weight for connection from the k^{th} neuron in the (1-1)th layer to the j^{th} neuron in the l^{th} layer $b_j^{(l)}$ is the bias of the j^{th} neuron in the l^{th}, f(*) is the activation function.

$$a_j^{(l)} = f\left(\sum_{k=1}^{n_{l-1}} w_{jk}^{(l)} a_k^{(l-1)} + b_j^{(l)} \right) \tag{5}$$

The input layer defines all the attribute values used as inputs for ANN model. In this study, the input values were SO_2, PM_{10}, $PM_{2.5}$, O_3, NO_2, and CO. The hidden layer consists of neurons that process the input data and pass the results to the output layer, with weights assigned to the input neurons to influence this process. The output layer contains neurons that represent the model's output, which in this study was the API.

The effectiveness of a model is evaluated using statistical measures that assess how well the observed outcomes are replicated. Two common indicators are the coefficient of determination, and the root mean square error [14]. Firstly, the coefficient of determination, R^2 is a widely used measure that indicates how well a model fits the data. An R^2 value closer to 1 suggests a better fit, as it means the model explains a greater proportion of the data's variability [17]. The equation 6 represent R^2 where $\hat{y}_i$ is the predicted i^{th} value, y_i is the i^{th} actual value, $\bar{y}$ is the mean of the actual value, m is the total number of observations.

$$R^2 = 1 - \left(\frac{\sum_{i-1}^{m} \left(y_i - y_i \right)^2}{\sum_{i-1}^{m} \left(y_i - \overline{y_i} \right)^2} \right) \tag{6}$$

Lastly, Root Mean Square Error, RMSE measures how well a model fits the data by quantifying the average magnitude of the prediction errors [18]. The equation 7 represents RMSE where n is the number of observations, y_i represents the actual value, $\hat{y}_i$ represents the predicted value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{m}\left(y_i - y_i\right)^2} \tag{7}$$

Best Model Presentation

The performance of a model is evaluated using error measures, with lower values indicating better accuracy [13]. This study compared error measures, specifically Root Mean Square Error, (RMSE) and Mean Absolute Error (MAE), between the best ARIMA and ANN models. The RMSE evaluates performance by measuring the difference between predicted and observed values, with larger errors having a greater impact due to squaring [11]. A smaller RMSE indicates better model performance. The equation is shown in equation 7.

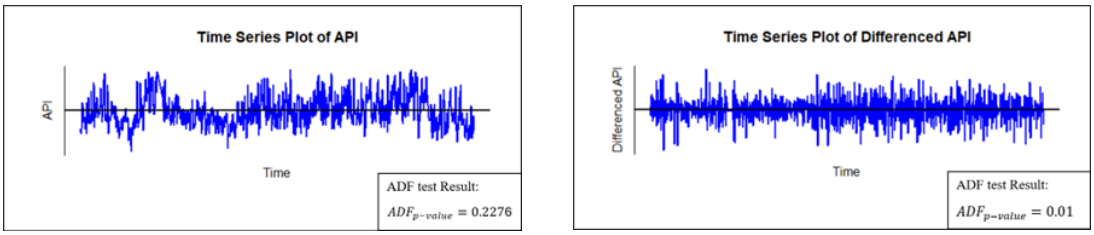
While, the MAE is less affected by large errors, making it stable, especially when actual values are near zero. A lower MAE indicates better model performance by showing closer average predictions to actual values [19]. The equation 8 represent MAE where n is the number of observations, y_i represents the actual value, $\hat{y}_i$ represents the predicted value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{m}\left(y_i - y_i\right)^2} \tag{8}$$

Result and Analysis

Box Jenkins Methodology

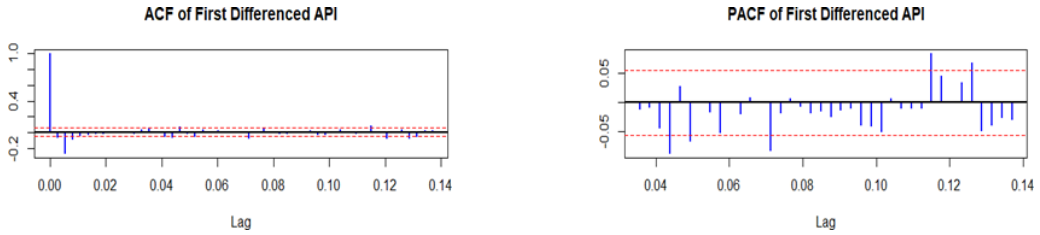
This study models daily API values using the ARIMA method in RStudio. Missing values and outliers were imputed. The Augmented Dickey Fuller (ADF) test is applied to ensure the time series data is stationary.



(a) before differencing (b) first differencing of API

Figure 2: API in Klang Valey.

Figure 2 illustrates the time series plots of API in Klang and the ADF test results. The initial plot in Figure 2(a) shows irregular fluctuations with the mean and variance of error values is not constant, and the ADF test showing p-value of 0.2276, indicating non-stationarity of the API series. After applying first differencing technique, the revised plot in Figure 2(b) displays consistent fluctuations with stable mean and variance. The ADF test confirms stationarity with a p-value of 0.01, supporting the rejection of the null hypothesis.



(a) ACF (b) PACF

Figure 3: API in Klang Valey.

Based on the ACF and PACF diagrams in Figure 3, significant spikes suggest potential ARIMA parameters. The ACF plot (Figure 3(a)) shows one significant spike at lag 2, suggesting an AR parameter of 1. Meanwhile, based on the PACF diagram (Figure 3(b)) gives five significant spikes (lags 4, 6, 11, 27 and 30), indicating an MA parameter of 5 were resulting in a potential ARIMA (5,1,1) model. Though, to identify the best model under ARIMA model [13], five other ARIMA models with parameters (p,d,q) = (4,1,1), (6,1,1), (5,1,2), (4,1,2) and (6,1,2) are tested. The Ljung Box test is used to assess each model for residual white noise, with the null hypothesis indicating no autocorrelation.

Table 2: Result of ARIMA Model.

Model	p-value	AIC	BIC
ARIMA (5,1,1)	0.2561	8292.351	8328.417
ARIMA (4,1,1)	0.2409	8290.440	8321.353
ARIMA (6,1,1)	0.2881	8294.070	8335.288
ARIMA (5,1,2)	0.2431	8294.387	8335.605
ARIMA (4,1,2)	0.2736	8292.210	8328.276
ARIMA (6,1,2)	0.2619	8296.337	8342.707

Table 2 shows the summary of the proposed ARIMA model result, which evaluates whether the errors in ARIMA models meet the white noise criteria. The p-values of all ARIMA models are greater than 0.05, leading to the rejection of the null hypothesis. Therefore, all six ARIMA models are well-specified and adequate, meeting the requirements for optimality. By evaluating the AIC and BIC, the ARIMA (4,1,1) model is identified as the best fit, with the lowest AIC and BIC values of 8290.44 and 8321.353, respectively. This indicates that the ARIMA (4,1,1) model strikes the best balance between model complexity and goodness of fit, outperforming the other candidate models. The ARIMA (4,1,1) model can be written as in equation 9.

$$\Delta_{y_t} = \phi_1 \Delta_{y_{t-1}} + \phi_2 \Delta_{y_{t-2}} + \phi_3 \Delta_{y_{t-3}} + \phi_4 \Delta_{y_{t-4}} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (9)$$

In the ARIMA (4,1,1) model, the observed value at time t is denoted as y_t, and the first difference of y_t is $\Delta y_t = y_t - y_{t-1}$. The model consists of four autoregressive (AR) parameters, ϕ_1, ϕ_2, ϕ_3, ϕ_4, which represent the influence of the previous four values on the current one. The differencing order, $d = 1$, is applied to make the data stationary by removing trends or patterns. The moving average (MA) parameter θ_1 accounts for the impact of the previous error term, which helps improve predictions by considering past prediction errors.

ANN Methodology

This study uses ANN to model API as the target variable, with SO_2, PM_{10}, $PM_{2.5}$, O_3, NO_2, and CO as inputs. Data screening and normalization were performed in JMP PRO 17 using imputation and the scale-and-offset method.

The ANN model comprises three layers: the input layer (receiving data), the hidden layer (assigning weights), and the output layer (target variable). To prevent overfitting, five networks with one to first hidden nodes were constructed. Model performance was evaluated using the coefficient of determination (R^2) and root mean square error (RMSE). The best-fit model had the highest R^2 and lowest RMSE, subject to model parsimony. Table 3 presents the ANN model results.

Table 3: R^2 and RMSE Result.

Model	Hidden Nodes	AIC	BIC
1	1	0.6102	5.8829
2	2	0.6165	5.8347
3	3	0.6184	5.8209
4	4	0.6306	5.7269
5	5	0.6354	5.6894

Model 5, with five hidden nodes, delivers the best performance, achieving the highest R^2 value of 0.6487 and the lowest RMSE of 5.7284 as shown in Table 3. This indicates that the model effectively captures relationships in the data, explains variability, and provides accurate predictions. The model also shows low error rates on both training and testing datasets, indicating a balance between learning from the data and generalizing to unseen data.

This suggests that Model 5 avoids overfitting, where a model performs well on training data but poorly on new data, and underfitting, where it fails to learn the patterns in the training data. The five hidden nodes represent processing units in the hidden layer that work together to identify patterns and relationships between the input variables and the target, API. These nodes, with their respective weights and biases, help the model generalize well, making Model 5 suitable for forecasting API values in Klang. Figure 4 illustrates the network structure of Model 5.

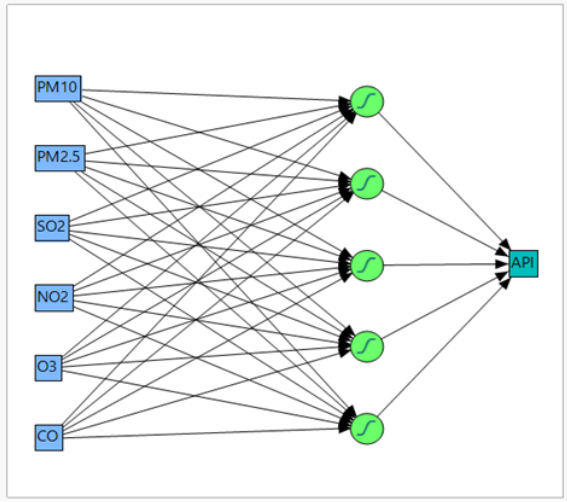


Figure 4: Model 5 Network.

The Model Performance

Table 4 shows that ANN Model 5, with five hidden nodes, outperforms ARIMA (4,1,1) based on error measures. ANN achieves an RMSE of 5.5502 and an MAE of 4.1799, compared to ARIMA's RMSE of 14.35176 and MAE of 11.28785. This demonstrates that ANN provides more accurate predictions and is better at capturing patterns in the data.

Table 4: R^2 and RMSE Result.

Model	RMSE	MAE
ARIMA (4,1,1)	14.35176	11.28785
ANN Model 5	5.5502	4.1799

Conclusion and Recommendation

The study applied both the Box-Jenkins and ANN methods to forecast API values. The Box-Jenkins approach identified ARIMA (4,1,1) as the best model based on its ability to balance accuracy and model complexity, demonstrated by the lowest AIC and BIC values. On the other hand, the ANN method determined that Model 5, with five hidden nodes, provided the best performance. This model achieved the highest R^2 and the lowest RMSE, indicating its effectiveness in capturing data patterns and maintaining low error rates. Overall, the ANN method showed superior performance compared to the Box-Jenkins method due to its lower error measurements and better adaptability to the data.

Future research could benefit from exploring additional modeling techniques beyond Box-Jenkins and ANN, such as linear regression, logistic regression, exponential smoothing, or Holt's method. Incorporating a wider range of methods could uncover unique patterns in the data and provide more insights. Additionally, using larger and more recent datasets could enhance the model's reliability and its ability to generalize across various environmental

conditions. This would also help researchers evaluate the model's performance over different time periods, improving its effectiveness for forecasting air quality.

Lastly, future studies should focus on real-time forecasting of API values, which was not part of the current research. Utilizing real-time air pollution data could lead to more accurate predictions and allow for better monitoring of daily air quality. Providing timely warnings about poor air conditions could help mitigate the health risks associated with air pollution and improve public health outcomes.

Acknowledgments

The researcher would like to sincerely thank everyone who helped make this study a success. Special thanks go to the management of Universiti Teknologi MARA Seremban Campus for their support and help in making this research possible.

References

[1] Susanto, A. D. (2020). Air pollution and human health. *Medical Journal of Indone- sia*, 29, 8–10. https://doi.org/https://doi.org/10.13181/mji.com.20457.

[2] Karliansyah, M. (2024). View of air pollution impacts on human health and policies to reduce air pollution *Medical Journal of Indonesia* https://mji.ui.ac.id/journal/ in- dex.php/mji/article/view/4579/1723Chen, J., & Mu, X. (2021). Review on environmental treatment of heavy pollution industry. https: // doi.org / https: //doi.org /10.1051 /e3sconf / 202132901045

[3] Roden, J. (2020). Air pollution issues in australia. *Medical Research Archives* 8. https: //doi.org/https://doi.org/10.18103/mra.v8i10.2262

[4] WHO. (2016) WHO releases country estimates on air pollution exposure and health impact

[5] WHO. (2023). Air Pollution

[6] Chang, F., & Ashfold, M. J. (2020). Public perceptions of air pollution and its health impacts in greater kuala lumpur. *IOP Conference Series: Earth and Environmental Science* 489. https://doi.org/https://doi.org/10.1088/1755-1315/489/1/012027

[7] Sentian, J., Herman, F., Yih, C. Y., & Hian Wui, J. C. (2019). Long-term air pol- lution trend analysis in malaysia. *International Journal of Environmental Impacts: Management, Mitigation and Recovery, 2*. DOI=https://doi.org/10.2495/ei-v2-n4- 309-324

[8] DOE. (2024). Environmental quality report 2023. Department of Environment Malaysia.

[9] Sahak, N., Asmat, A., & Yahaya, N. Z. (2022). Spatio-temporal air pollutant characterization for urban areas. *Journal of Geoscience and Environment Protection, 10(01)*, 218–237. https://doi.org/https://doi.org/10.4236/gep.2022.101015

[10] Albashir Abdulali, B. A., & Masseran, N. (2021). Artificial Neural Network (ANN) and ARIMA Models for Better Forecast of the Air Pollution Data in Malaysia.

[11] WHO. (2021). Pneumonia. https://www.who.int/news-room/fact-sheets/detail/pneumonia

[12] Lazim, M. A. (2011). *Introductory business forecasting : A practical approach* (3rd ed). UiTM Press.

[13] Mousavian, A., Mahmodabady, H. Z., & Jadval Ghadam, A. G. (2015). Prediction of the pollutants concentration using artificial neu- ral network (ann). *Environment Conservation Journal, 16(SE)*, 171–180. https://doi.org/https://doi.org/10.36953/ecj.2015. se1619

[14] Sobri, N. M., Yaacob, W. F. W., Ismail, N. A., Malik, M. A. A., Rahman, R. A., Baser, N. A., & Sukhairi, S. A. M. (2021). Predicting particulate matter (pm2.5) in malaysia using multiple linear regression and artificial neural network. *Jour- nal of Physics: Conference Series.* https://doi.org/https://doi.org/10.1088/1742- 6596/2084/1/ 012010

[15] Ku Yusof, K. M. K., Azid, A., Abdullah Sani, M. S., Samsudin, M. S., Muham- mad Amin, S. N. S., Abd Rani, N. L., & Jamalani, M. A. (2019). The evalua- tion on artificial neural networks (ann) and multiple linear regressions (mlr) mod- els over particulate matter (pm10) variability during haze and non-haze episodes: A decade case study. *Malaysian Journal of Fundamental and Applied Sciences.* https://doi.org/https: //doi.org/ 10.11113/ mjfas.v15n2.1004

[16] Zulkepli, N. E. S. (2021). A study on air pollution index in sabah and sarawak using principal component analysis and artificial neural network. http : / / ir. uitm . edu . my/id/ eprint/46511/1/46511.pdf

[17] Maltare, N. N., & Vahora, S. (2023). Air quality index prediction using machine learning for ahmedabad city. 7, 100093–100093. https://doi.org/https://doi.org/10.1016/j.dche.2023.10 0093

[18] Liu, T., & You, S. (2022). Analysis and forecast of beijing's air quality index based on arima model and neural network model. Atmosphere, 13. https : // doi . org / https: //doi.org/ 10.3390/atmos13040512