11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara
19th September 2024

Performance Evaluation of SARIMA Model for Solar Radiation Forecasting

Beabbyline Muntasir[1]; Shazlyn Milleana Shaharudin[1*]; Nur Haizum Abd Rahman[2]; Mou Leong Tan[3]

[1] Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia, d097047@siswa.upsi.edu.my; shazlyn@fsmt.upsi.edu.my
[2] Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, 26300 Gambang, Pahang, Malaysia, haizum@umpsa.edu.my
[3] GeoInformatic Unit, Geography Section, School of Humanities, Universiti Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia, mouleong@usm.my

Abstract:

Solar radiation forecasting is vital for optimizing renewable energy systems and addressing environmental issues. Accurate forecasts enable better energy production planning and economic decision-making, ensuring cost savings and enhanced energy efficiency. This study aims to evaluate the performance of Seasonal Autoregressive Integrated Moving Average (SARIMA) model for forecasting solar radiation in Ipoh, Perak. Daily solar radiation in Ipoh, obtained from Meteorological Department of Perak, were analyzed. The methodology involved data preprocessing, which included handling missing values, addressing data skewness, and analyzing patterns, as well as SARIMA modeling, which included ensuring data stationarity, selecting possible models, and validating selected models' performance. Python programming was used to select models based on model coefficient p-values and the Ljung-Box Q test p-values, and selected models' comparison based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Performance evaluation was done using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). The optimal SARIMA model identified was SARIMA $(6, 1, 0)(2, 2, 0)_{52}$, with MAE of 2.81, MSE of 12.09, RMSE of 3.48, and MAPE of 15.74%.

Keywords:

Prediction, Seasonal, Modelling

1. Introduction:

The forecasted electricity demand per capita is expected to continue increasing due to population growth. Electricity demand and generation in Malaysia indicates that both will significantly expand year by year (Azman, 2021). This surge in energy demand underscores the need for efficient and sustainable energy solutions but at the same time,

the environmental impact of energy production remains a significant concern. Traditional energy sources, especially those involving fuel combustion, contribute to air pollution and greenhouse gas emissions. Addressing these environmental issues requires a shift towards more cleaner energy sources, and solar energy is one of the great renewable energy sources.

The empirical findings by Raihan, 2022 show that the coefficient of economic growth in Malaysia is positive and significant. However, Malaysia must plan well in order maintain a consistent economic growth and regain its historical high growth momentum. In this context, accurate solar radiation forecasting becomes crucial. Precise predictions can optimize the placement and efficiency of solar panels, thus reducing the overall costs of solar technology installation and operation. Since the solar radiation data often exhibits seasonal and trend components, SARIMA model allows for precise modelling of these variations (Al-Rousan, 2021). This study not only aims to evaluate the performance of SARIMA model for solar radiation forecasting in Ipoh but also provides valuable information for future researchers to supports economic and environmental goals by optimizing energy use and reducing pollution.

2. Methodology:

Daily solar radiation measurements that obtained from the Meteorological Department of Perak covering an extensive period spanning multiple decades. This long-term dataset is crucial for capturing seasonal and long-term trends in solar radiation, which are essential for accurate SARIMA modelling. To manage the computational complexity inherent in such an extensive dataset, the daily measurements were aggregated into weekly intervals. This aggregation simplifies the data without significantly losing important information, making the SARIMA modelling process more efficient. Incomplete weeks were excluded from the analysis to ensure consistency and reliability of the data. Each week was structured to contain seven data points, ensuring uniformity in the dataset.

The methodology then proceeds with data preprocessing, which includes handling missing values, addressing data skewness, and analysing patterns, and SARIMA modelling, including model selection and validation. For addressing missing data, three distinct methods which are mean substitution (MS), random forest (RF), and k-nearest neighbors (KNN) was outlined. MS involves replacing missing data for a variable with the mean of the available non-missing data for that variable. RF method combines the predictions of multiple decision trees, using bagging (bootstrap aggregation) to aggregate predictions from various random predictors. The KNN method imputes missing values by identifying a set number of instances that are most similar to the instance of interest. The performance of these imputation method evaluated using RMSE, NSE, and MAE.

The imputed data using the best imputation method then partitioned into training and testing sets to evaluate the performance of the predictive models. In this study, the test set consisted of 52 data points, representing one year (52 weeks). Data normalization before applying the SARIMA model was crucial because it helped stabilize the variance and improve the model's performance (Al-Rousan, 2021). SARIMA models assume that the time series data is stationary, meaning that its statistical properties do not change over time. Box-Cox Transformation, a statistical technique, was used to transform non-normally distributed data into a form that more closely approximated a normal distribution.

Seasonal and trend decomposition was then performed to analyse time series data by breaking it down into its fundamental components, which are trend, seasonal, and residual

components. This decomposition helps in better understand the underlying patterns in the data and can inform the modelling process, which are the value of seasonal period in SARIMA model. The trend component of a time series represents the long-term progression of the data, indicating the overall direction in which the data is moving over time, whether upward, downward, or stable. The seasonal component captures the repeating patterns or cycles that occur at regular intervals, reflecting consistent seasonal variations. Meanwhile, the residual component, also known as the noise component in the data that cannot be explained by the trend or seasonal components.

Ensuring data stationarity is crucial for SARIMA modelling as it's met the model assumptions (Haddad, 2019). To assess stationarity, the Autocorrelation Function (ACF) graph is used. A stationary series shows a rapid decline in spikes after a few lags, while a slow decay suggests non-stationarity due to trends or seasonality. Regular and seasonal differencing are applied to address these issues. The number of regular differencing (d) and seasonal differencing (D) applied are the values for d and D in SARIMA model. After these adjustments, the ACF is replotted to confirm stationarity. This approach also ensures that the SARIMA model effectively captures both non-seasonal and seasonal components of the data.

Selecting parameters for SARIMA models involves analysing the ACF and PACF plots to determine both non-seasonal and seasonal components (Al, Rousan, 2021). For non-seasonal parameters, the PACF plot is used to identify the number of autoregressive terms (p). This is indicated by the lag where the spikes cuts off or drops sharply. The ACF plot helps determine the moving average terms (q), with an abrupt cut-off after a certain lag suggesting the number of MA terms. For seasonal parameters, the ACF and PACF are analysed at lags that correspond to multiples of the seasonal cycle length (s). The ACF plot reveals seasonal moving average terms (Q) by showing an abrupt cut-off at these lags, while the PACF plot indicates seasonal autoregressive terms (P) by showing where seasonal partial autocorrelations cut off sharply.

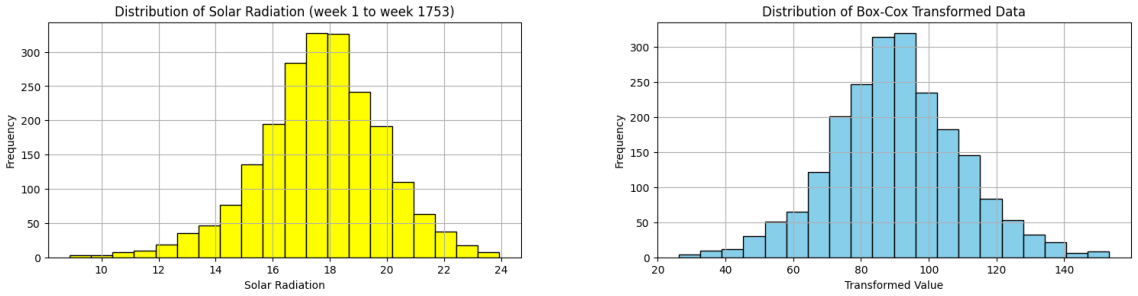
Model selection involves evaluating p-values of model coefficients and the Ljung-Box Q test. Coefficients with p-values below 0.05 are considered statistically significant, and the Ljung-Box Q test ensures residuals are white noise. Models meeting these criteria are compared using the AIC and BIC. The model with the lowest AIC and BIC values is preferred for its optimal fit and minimal complexity. For model validation, several metrics are used which are MAE, measures average error magnitude, MSE assesses squared differences, and RMSE provides error in data units. MAPE expresses accuracy as a percentage for easy comparison. Residuals' ACF and PACF are checked to ensure they are white noise, validating the model's reliability and effectiveness.

3. Result:

The data preprocessing process began with examination of missing values. It was found that the dataset contained 973 missing values, accounting for 6.34% of the total data. After evaluating imputation methods using RMSE, NSE, and MAE, it was determined that RF performed most effectively in handling these missing values.

Imputation Method	Mean Substitution	Random Forest	KNN
RMSE	2.847	2.592	2.881
NSE	0.398	0.516	0.399
MAE	2.286	2.021	2.214

Table 3.1: Imputation method of weekly solar radiation data in Ipoh



(a) (b)

Figure 3.1: (a) Distribution before; (b) Distribution after Box-Cox transformation for weekly solar radiation training dataset

Figures 3.1 illustrate the effects of addressing skewness in the weekly solar radiation data for the training dataset. It's demonstrating the distribution of the original transformed weekly training solar radiation data, highlighting a slight right skew. By applying Box- Cox transformation, the skewness in the data is reduced.

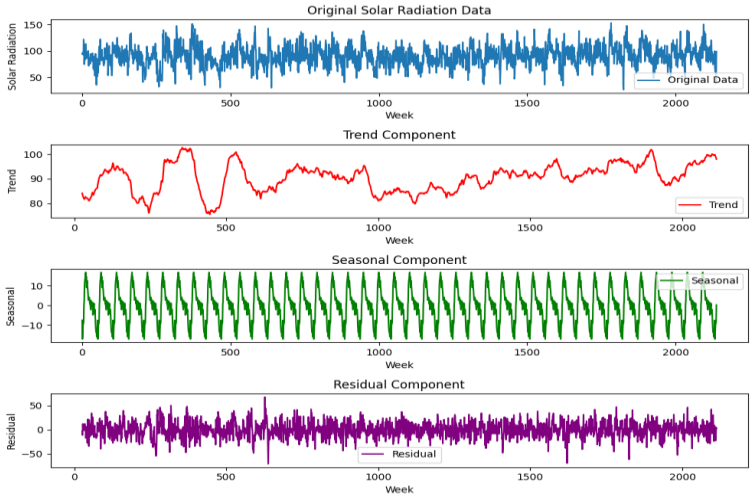
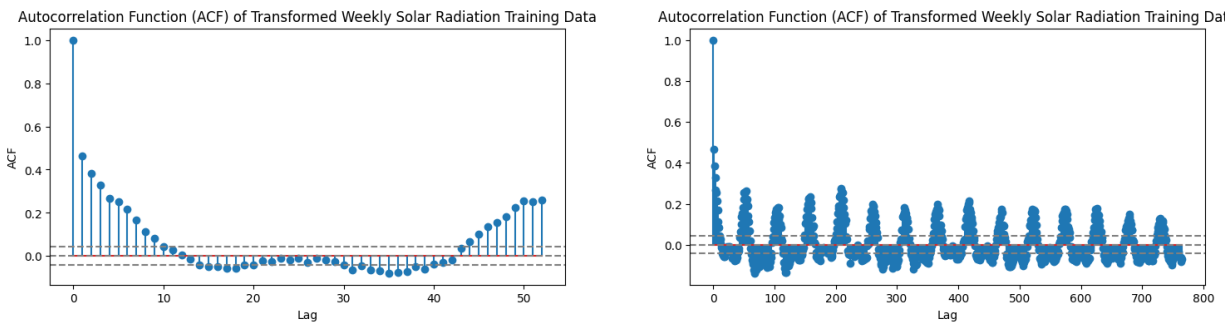


Figure 3.2: Seasonal and trend decomposition of weekly transformed solar radiation training data

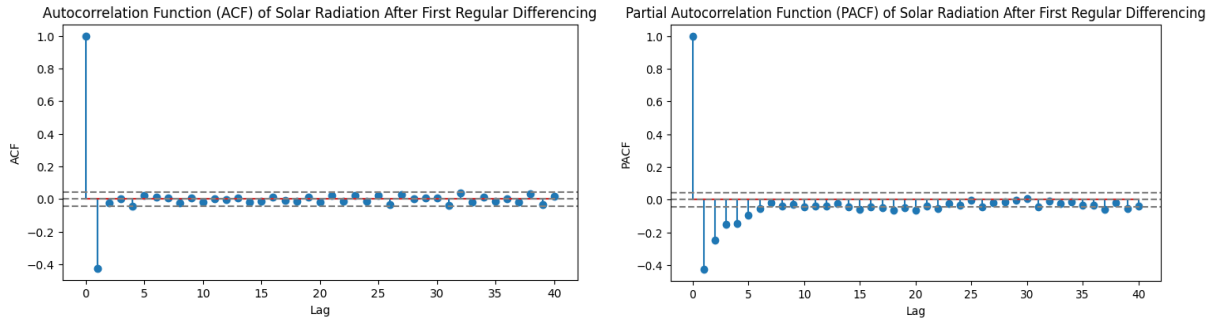
Figure 3.2 illustrates that the transformed weekly solar radiation data in the training set exhibit seasonal and trend components, with a repeating cycle every 52 weeks. This cycle corresponds to yearly seasons, indicating an annual pattern in solar radiation. Thus, the seasonal period chosen for the SARIMA model in subsequent analyses will be 52 weeks.



(a) (b)

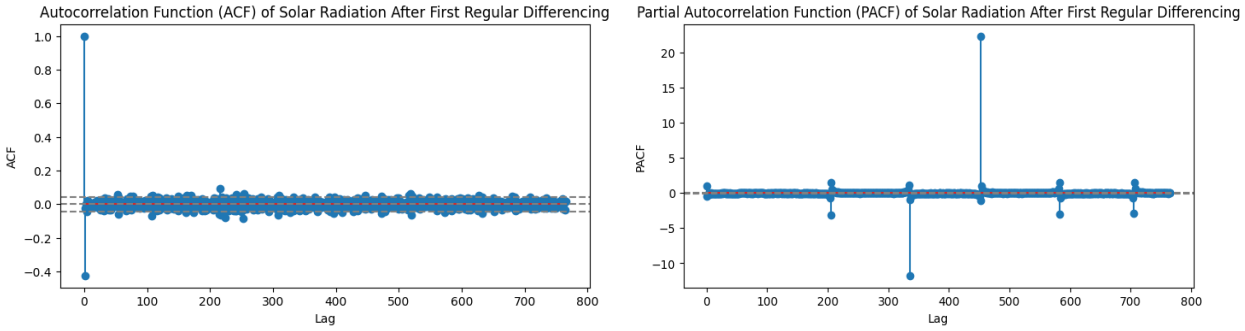
Figure 3.3: (a) ACF graph up to 52 lags; (b) ACF graph up to 765 lags of transformed weekly solar radiation training data

Based on Figure 3.3 and 3.4, the ACF plot reveals the correlation of the time series with its own lagged values. The presence of initial spikes followed by a gradual decay and repeated pattern every 52 lags, strongly indicates that the data is non-stationary.



(a) (b)

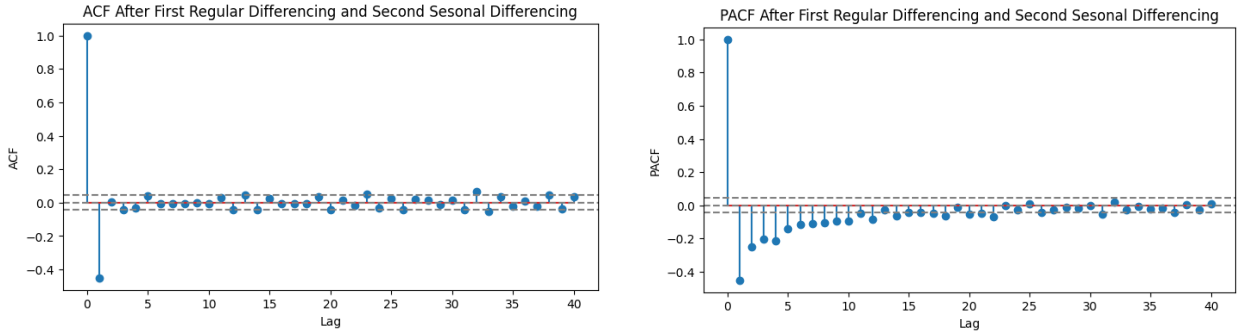
Figure 3.4: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing up to 40 lags



(a) (b)

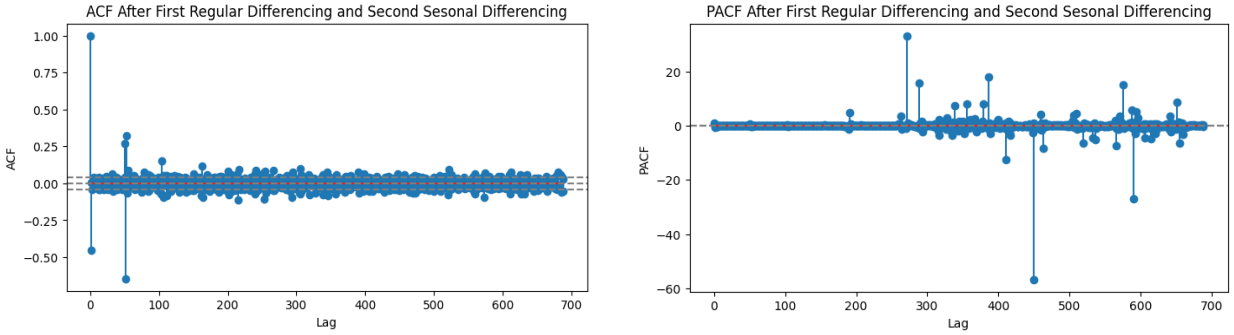
Figure 3.5: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing up to 765 lags

Based on figure 3.4, after the first regular differencing, the ACF graph shows a rapid decline, indicating reduced non-stationarity. The PACF graph show few spikes at lower lags. However, figure 3.5 shows that there are still few significant spikes every 52 lags for both ACF and PACF graph, suggesting some remaining patterns.



(a) (b)

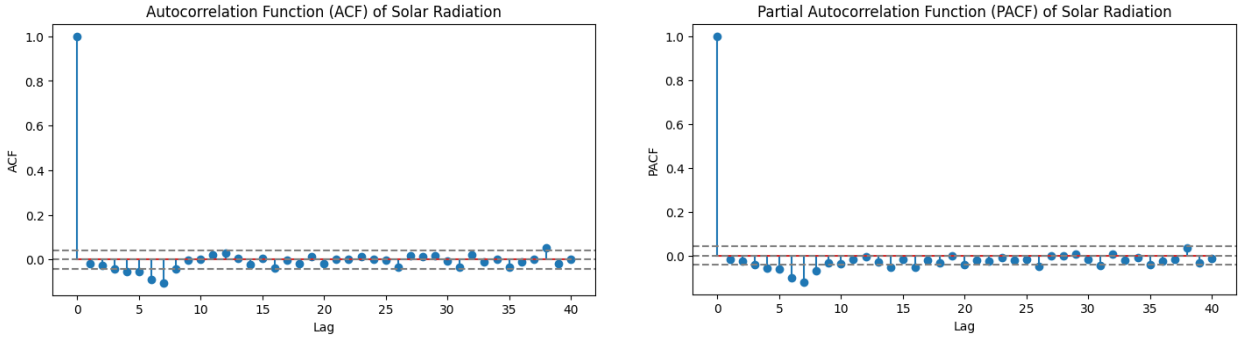
Figure 3.6: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing and second seasonal differencing up to 40 lags



(a) (b)

Figure 3.7: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing and second seasonal differencing up to 688 lags

Based on Figure 3.7, after applying the second seasonal differencing, the ACF shows that the spikes cut off after 104 lags, equivalent to two cycles of 52 lags. The PACF in Figure 3.7 shows that the spikes cut off after 156 lags, corresponding to three cycles of 52 lags. Therefore, the SARIMA model parameters suggested by the graphs in figures 3.6 and 3.7 are SARIMA(10,1,1)$(3,2,2)_{52}$. By using Python programming, among all the models considered, SARIMA(6,1,0)$(2,2,0)_{52}$ was the only model that met both criteria which are all of the p-values of model coefficient is 0.00 (statistically significant) and p-values of Ljung-Box Q Test yielded a high p-value of 0.66, indicating no significant autocorrelation in the residuals.



(a) (b)

Figure 3.8: (a) ACF; (b) PACF graph of SARIMA(6,1,0)$(2,2,0)_{52}$ residual

The ACF and PACF plot in figure 3.8 shows the autocorrelation of the SARIMA(6,1,0)$(2,2,0)_{52}$ residuals at different lags. The majority of the residuals fall within the 95% confidence bounds (dashed lines).

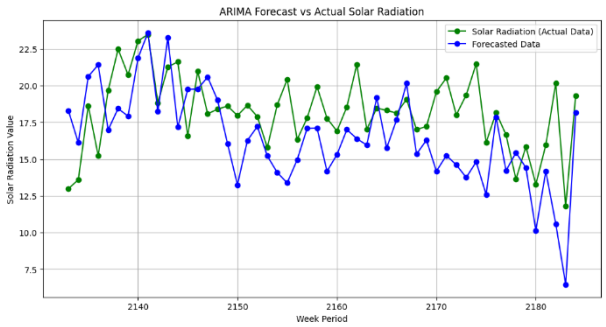


Figure 3.9: Comparison between testing data and forecasted data using SARIMA(6,1,0)$(2,2,0)_{52}$

The ACF plot in figure 3.9 shows the autocorrelation of the residuals at different lags. The majority of the residuals fall within the 95% confidence bounds (dashed lines), indicating that there is no significant autocorrelation in the residuals. The evaluation performance obtained for SARIMA(6,1,0)$(2,2,0)_{52}$ are MAE of 2.809, MSE of 12.09, RMSE of 3.477, and a MAPE of 15.74%.

4. Discussion and Conclusion:

A detailed analysis of weekly solar radiation data in Ipoh underscores the importance of accurate forecasting for optimizing solar energy systems. The dataset revealed that 6.34% of the data was missing. To address this, three imputation methods were compared which are MS, KNN and RF. RF emerged as the superior method, with an RMSE of 2.592, NSE of 0.516, and MAE of 2.021, outperforming MS and KNN. To correct the slight right skew in the data, the Box-Cox transformation was applied, resulting in a more symmetrical distribution suitable for advanced modelling. Seasonal and trend decomposition revealed a significant annual cycle, repeating every 52 weeks, highlighting strong seasonal influences on solar radiation in Ipoh. This finding was crucial for configuring the SARIMA model to accurately capture these patterns. Analysing the ACF and PACF plots helped identify non-stationarity, which was mitigated through first and second seasonal differencing. The final (6,1,0)$(2,2,0)_{52}$ model was selected for its statistical significance and absence of significant autocorrelation in the residuals, as confirmed by the Ljung-Box Q Test. The SARIMA model's performance, with an MAE of 2.809, MSE of 12.09, RMSE of 3.477, and MAPE of 15.74%, demonstrated its precision and reliability in forecasting solar radiation. Accurate forecasting is crucial for optimizing the placement and operation of solar panels, thereby minimizing installation and operational costs and making solar technology more economically viable.

Acknowledgement:

This research was funded by the Universiti Sains Malaysia, Research University Team (RUTeam) Grant Scheme with Project No. 1001/PHUMANITI/8580014.

References:

1. Azman, A. H., Tukimat, N. N. A., Malek, M. A., & Che, R. F. (2021, October). Analysis of Malaysia electricity demand and generation by 2040. In *IOP Conference Series: Earth and Environmental Science* (Vol. 880, No. 1, p. 012050). IOP Publishing.

2. AL-Rousan, N., & Al-Najjar, H. (2021). A comparative assessment of time series forecasting using NARX and SARIMA to predict hourly, daily, and monthly global solar radiation based on short-term dataset. Arabian Journal for Science and Engineering, 46(9), 8827-8848

3. Haddad, M., Nicod, J., Mainassara, Y. B., Rabehasaina, L., Al Masry, Z., & Péra, M. (2019, September). Wind and solar forecasting for renewable energy system using sarima-based model. In International conference on time series and forecasting.

4. Raihan, A., & Tuspekova, A. (2022). Toward a sustainable environment: Nexus between economic growth, renewable energy use, forested area, and carbon emissions in Malaysia. *Resources, Conservation & Recycling Advances*, *15*, 200096.