



MINISTRY OF ECONOMY
DEPARTMENT OF STATISTICS MALAYSIA

Performance Evaluation of SARIMA Model for Solar Radiation Forecasting

BEABBYLINE BINTI MUNTASIR

EDUCATION UNIVERSITY OF SULTAN IDRIS


**11th MALAYSIA
STATISTICS CONFERENCE**
"Data and Artificial Intelligence: Empowering the Future"

**19th September
2024**

Organized by:



OUTLINE

-  1. INTRODUCTION
-  2. PROBLEM STATEMENT
-  3. RESEARCH OBJECTIVES
-  4. RESEARCH QUESTIONS
-  5. METHODOLOGY
-  6. RESULTS
-  7. DISCUSSION AND CONCLUSION
-  8. REFERENCES

1. INTRODUCTION

- The forecasted electricity demand per capita is expected to continue increasing due to population growth.
- 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.

2. PROBLEM STATEMENT

- Electricity demand and generation in Malaysia indicates that both will significantly expand year by year (Azman, 2021).
- Even though the empirical findings by Raihan, 2022 show that the coefficient of economic growth in Malaysia is positive and significant, Malaysia must plan well in order maintain the 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 is one of the model that allows for precise modelling of these variations (Al-Rousan, 2021).
- This study aims to evaluate the performance of SARIMA model for solar radiation forecasting in Ipoh.

3. RESEARCH OBJECTIVES

- a) To find the best imputation method for solar radiation data in Ipoh
- b) To find the optimum SARIMA model to forecast solar radiation in Ipoh
- c) To evaluate the performance of the optimum SARIMA model in predicting solar radiation in Ipoh

4. RESEARCH QUESTIONS

- a) What is the best imputation method for solar radiation data in Ipoh?
- b) What is the optimum SARIMA model to forecast solar radiation in Ipoh?
- c) How well does the optimum SARIMA model perform in predicting solar radiation in Ipoh?

5. METHODOLOGY

DAILY SOLAR RADIATION DATA IN IPOH PERAK FROM 1ST JANUARY 1980 UNTIL 31ST DECEMBER 2021

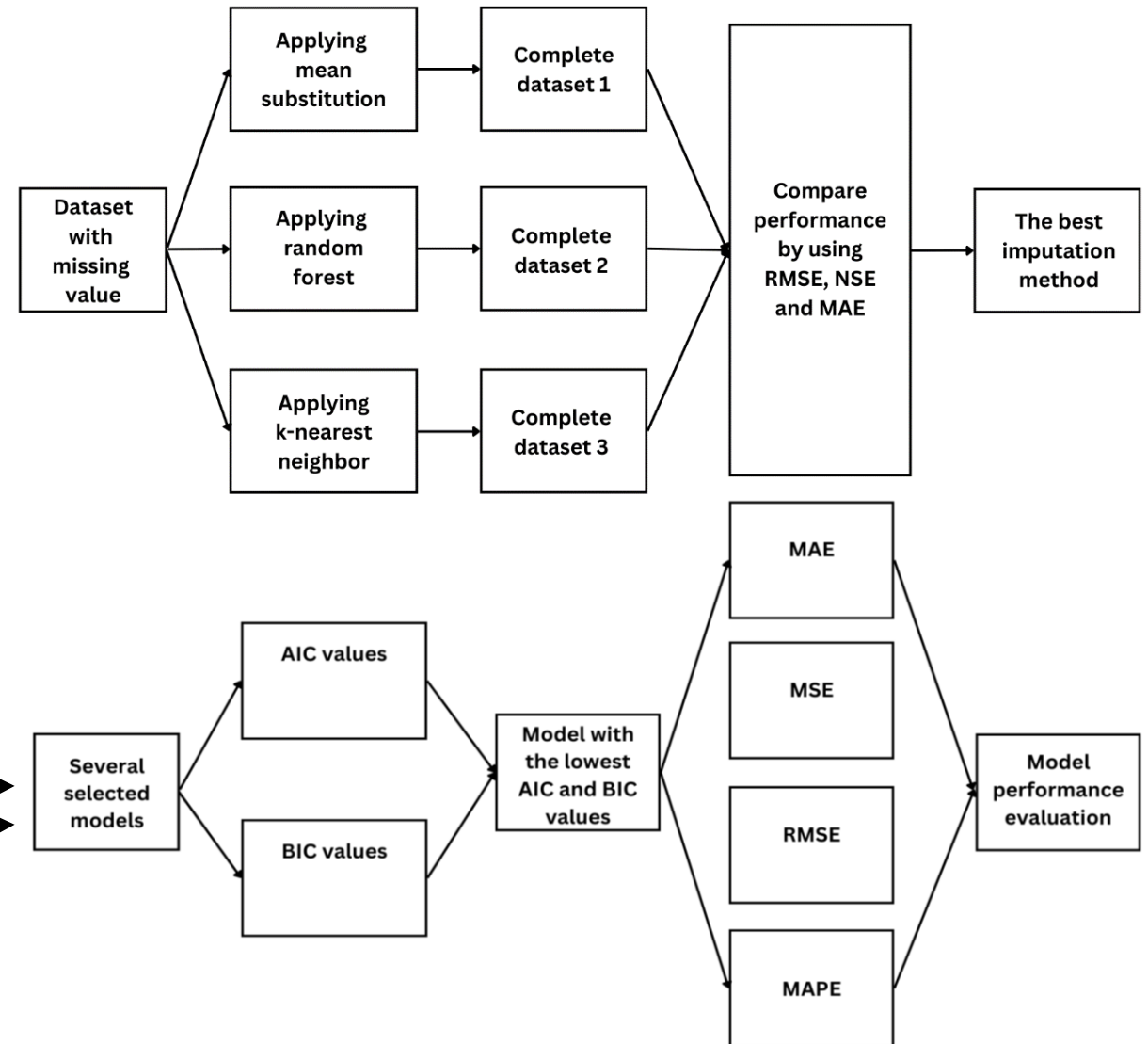
DATA PREPROCESSING

- Imputation method
- Data conversion (from daily data to weekly data)
- Data partition (data partitioned into training and testing sets, 52 data points for testing set)
- Data normalization (Box- cox transformation)
- Seasonal and trend decomposition

SARIMA MODELLING

- Ensuring data stationary (ACF and PACF graph)
- Parameter Selection (ACF and PACF graph)
- Model Selection (p values of the model coefficients and Ljung-Box Q test)

Model Performance Evaluation



6. RESULTS

DATA PREPROCESSING

- It was found that the dataset contained 973 missing values, accounting for 6.34% of the total data.
- After evaluating imputation methods using root mean square error (MSE), nash-sutcliffe efficiency (NSE), and mean absolute error (MAE), it was determined that random forest (RF) method performed most effectively in handling these missing values.

Imputation Method	Mean Substitution	Random Forest	K-Nearest Neighbors
RMSE	2.847	2.592	2.881
NSE	0.398	0.516	0.399
MAE	2.286	2.021	2.214

Table 6.1: Imputation results of weekly solar radiation data in Ipoh

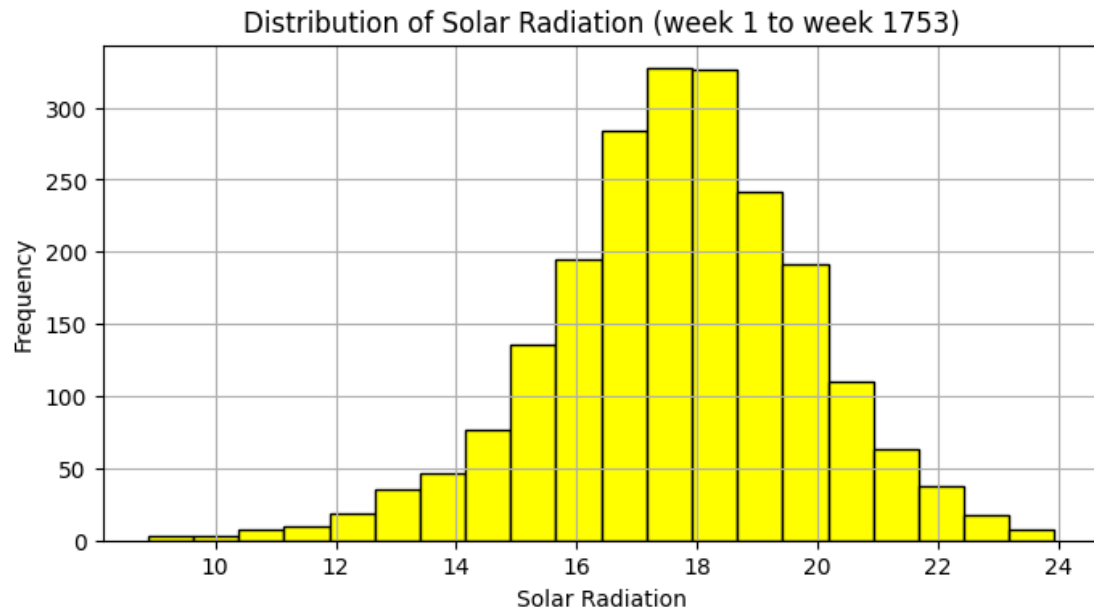
6. RESULTS

DATA PREPROCESSING

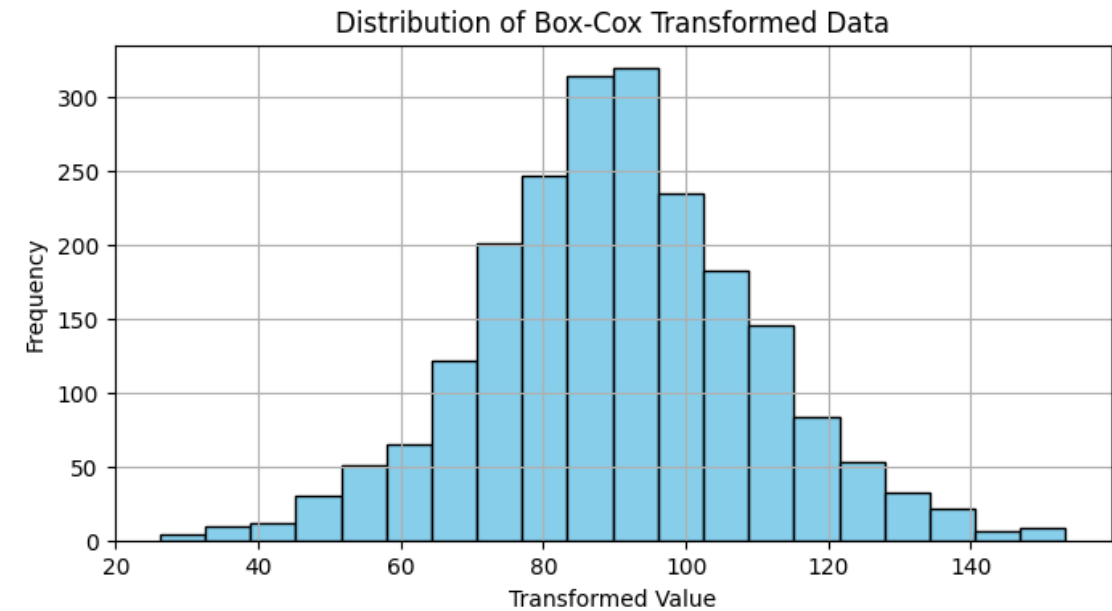
- Due to the large dataset comprising 15,341 daily entries, the data was aggregated into weekly data, resulting in 2,191 weekly data points.
- Four data points were removed due to incomplete weeks.
- Since the data starts on January 1, 1980, which is a Tuesday, each data point after data conversion refer to weekly data that start with Tuesday until Monday.
- Therefore, the first forecast value will refer to the weekly average forecast value, which starts on Tuesday and ends on Monday.
- After that, the data was partitioned into training and testing sets to evaluate the performance of the predictive models. In this study, the test set consists of 52 data points, representing one year (52 weeks).

6. RESULTS

DATA PREPROCESSING



(a)



(b)

Figure 6.1: (a) Distribution before; (b) Distribution after Box-Cox transformation for weekly solar radiation training dataset

- By applying Box- Cox transformation, the skewness in the data is reduced.

6. RESULTS

DATA PREPROCESSING

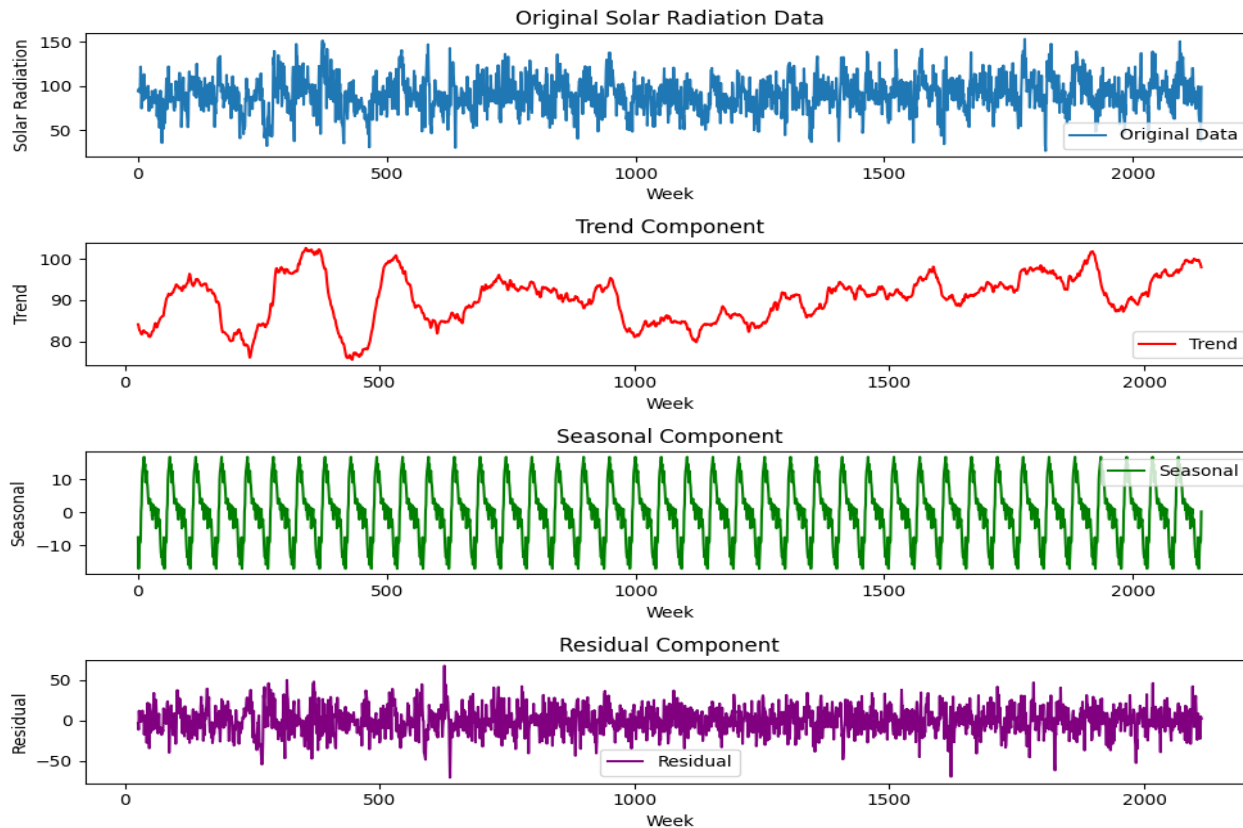
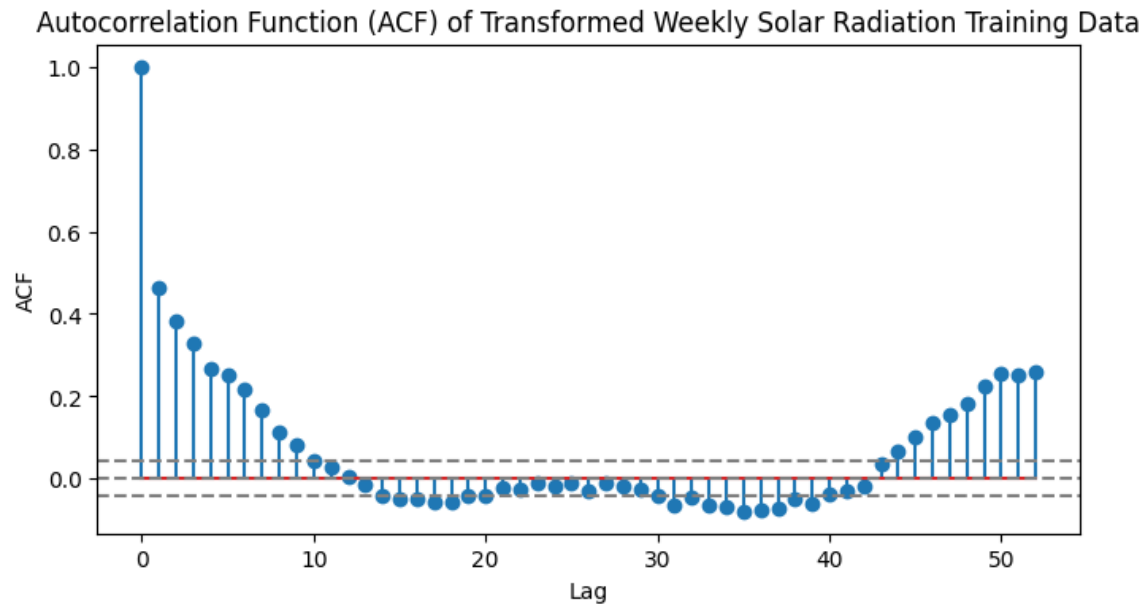


Figure 6.2: Seasonal and trend decomposition of weekly transformed solar radiation training data

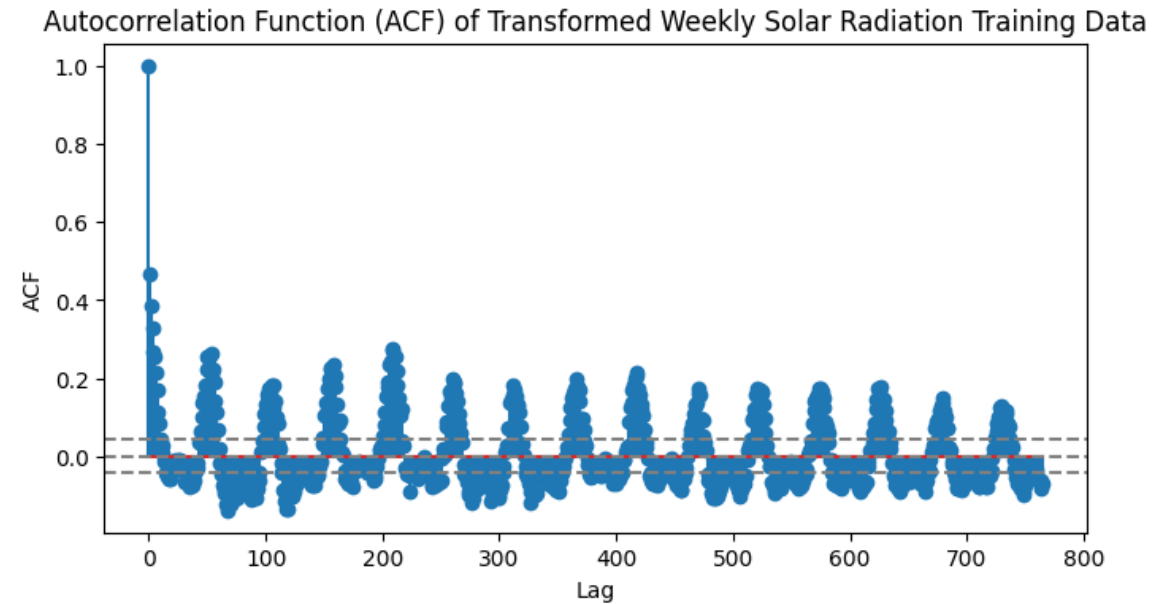
- Transformed weekly solar radiation data in the training set exhibit seasonal and trend components, with a repeating cycle every 52 weeks.
- This cycle corresponds to yearly seasons, indicating an annual pattern in solar radiation.
- Thus, the seasonal period chosen for the SARIMA model in subsequent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

6. RESULTS

SARIMA MODELLING



(a)



(b)

Figure 6.3: (a) ACF graph up to 52 lags; (b) ACF graph up to 765 lags of transformed weekly solar radiation training data

- The presence of initial spikes followed by a gradual decay and repeated pattern every around 52 lags, strongly indicates that the data is non-stationary.

6. RESULTS

SARIMA MODELLING

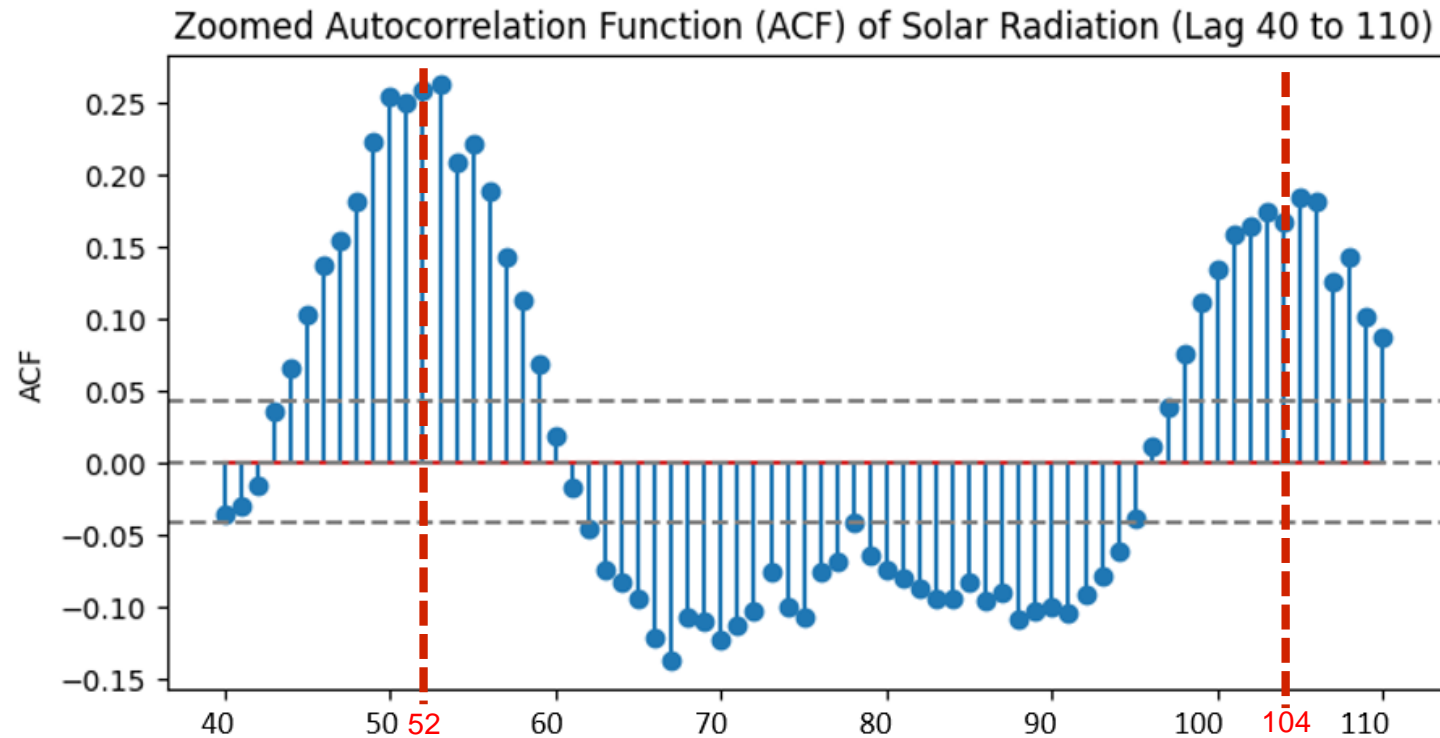


Figure 6.4: Zoomed ACF graph up to 110 lags

- repeated pattern every around 52 lags

6. RESULTS

SARIMA MODELLING

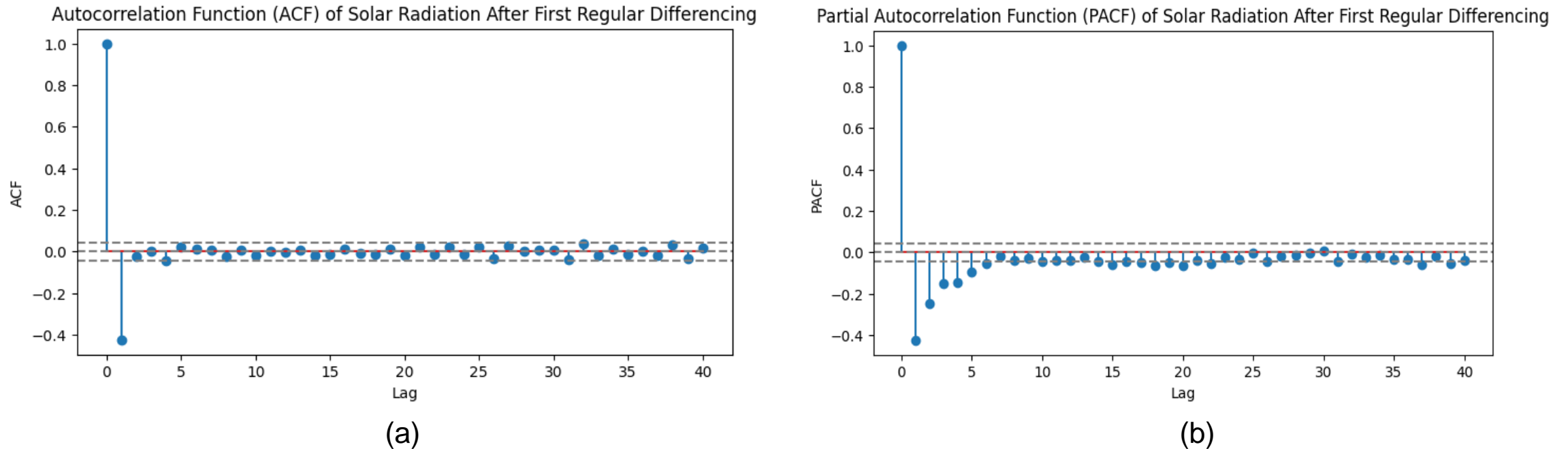


Figure 6.5: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing up to 40 lags

- The ACF graph shows a rapid decline, indicating reduced non-stationarity.
- The PACF graph show few spikes at lower lags.

6. RESULTS

SARIMA MODELLING

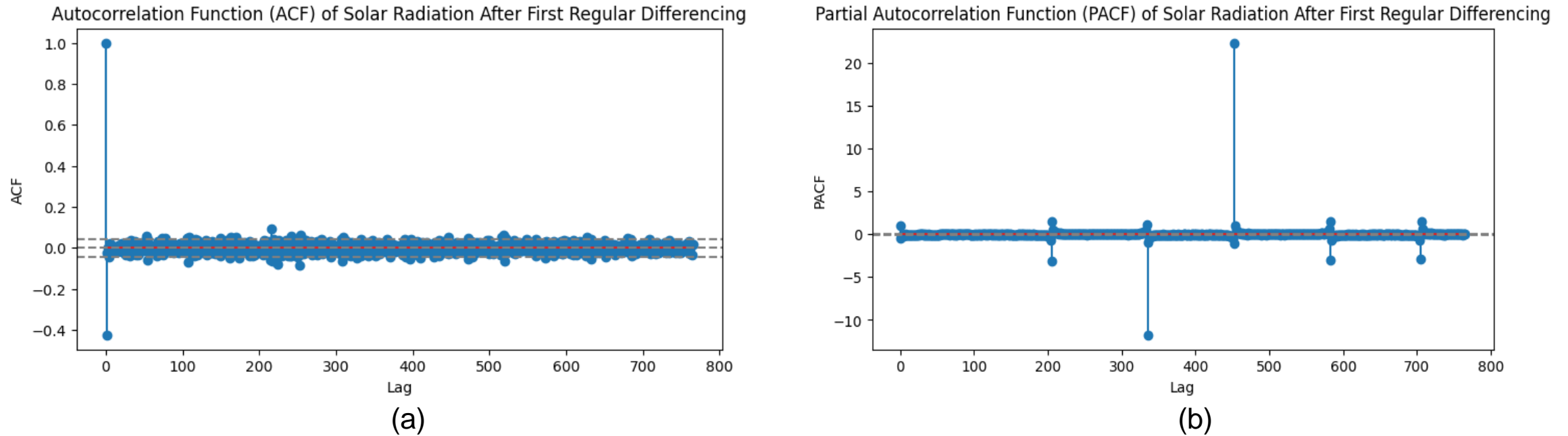


Figure 6.6: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing up to 765 lags

- However, figure 6.6 (b) shows that there are still few significant repeated spikes for PACF graph, suggesting some remaining patterns.

6. RESULTS

SARIMA MODELLING

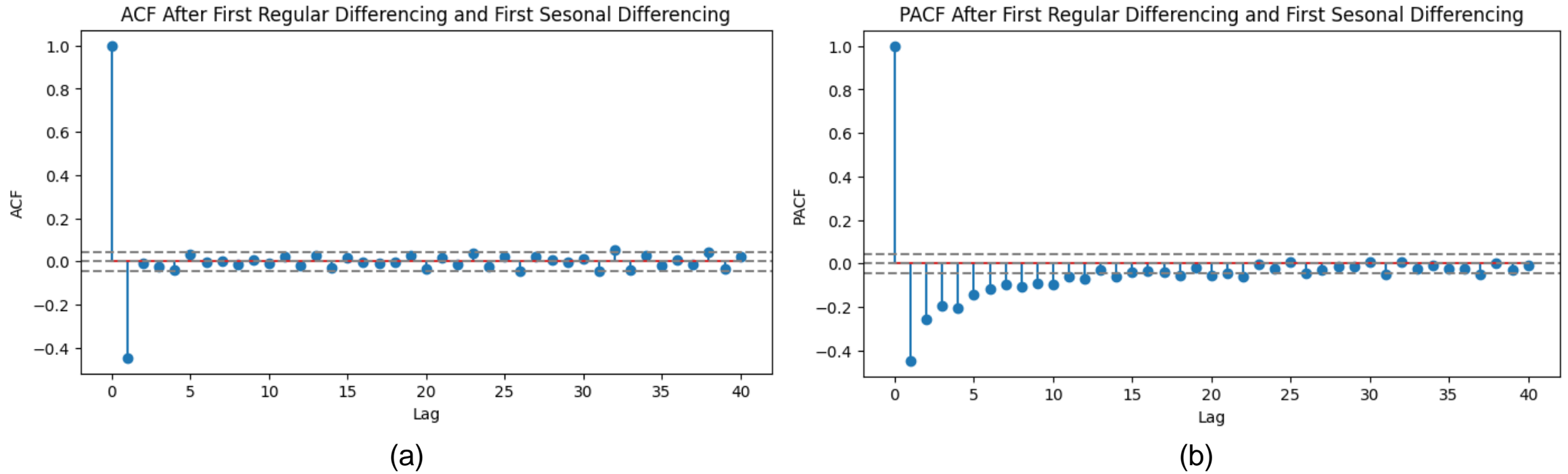


Figure 6.7: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing and first seasonal differencing up to 40 lags

6. RESULTS

SARIMA MODELLING

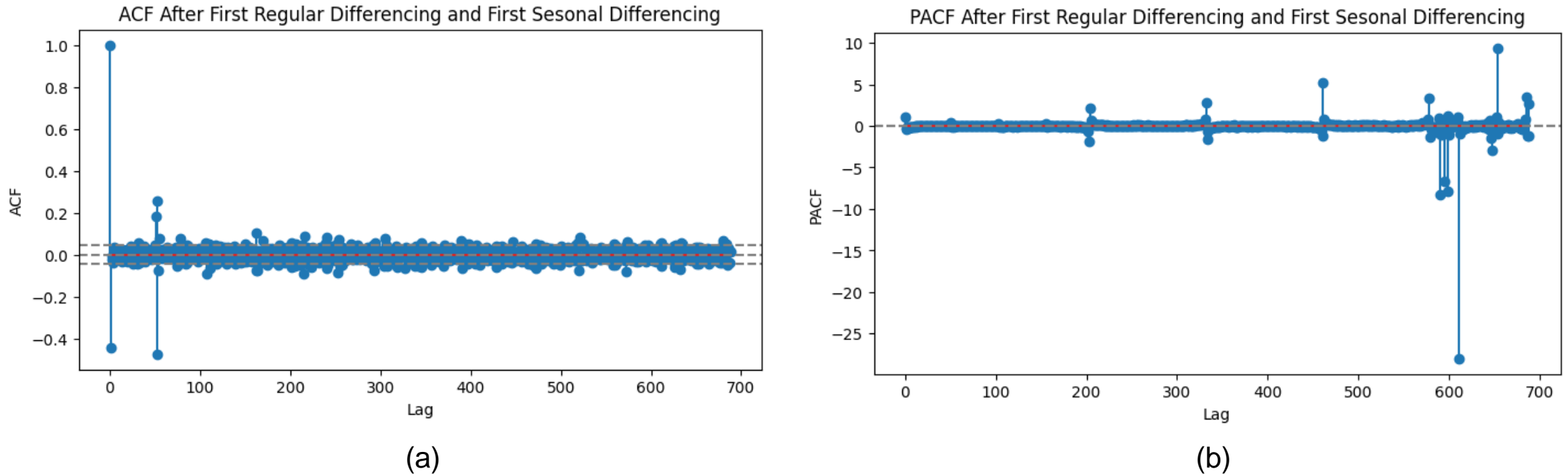


Figure 6.8: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing and first seasonal differencing up to 688 lags

- there are still few significant repeated spikes for PACF graph, suggesting some remaining patterns.

6. RESULTS

SARIMA MODELLING

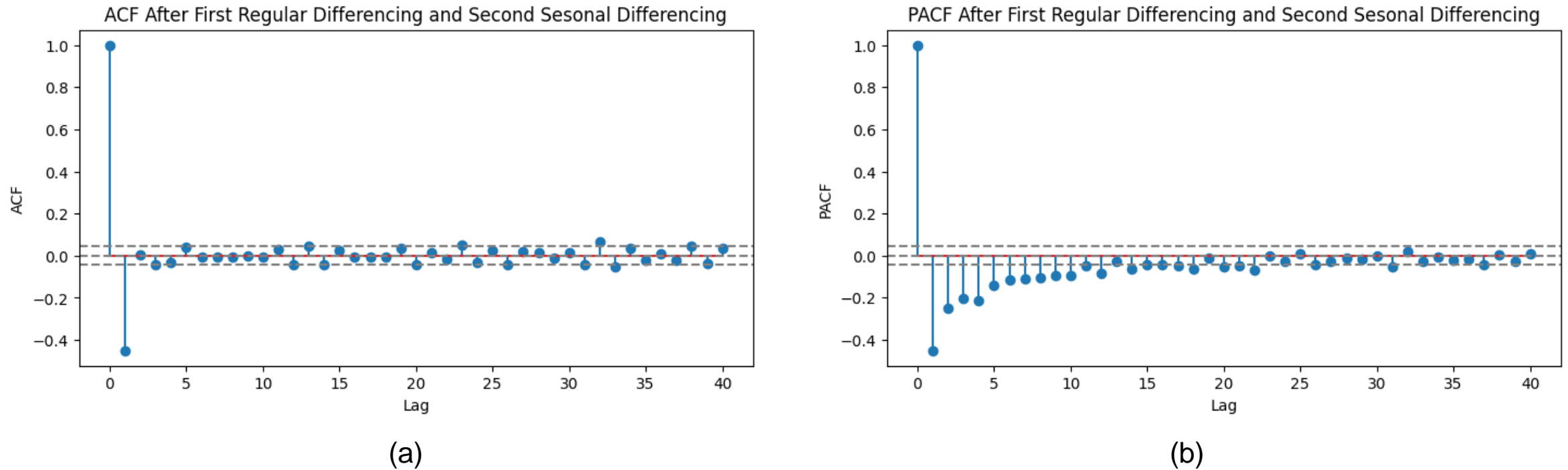
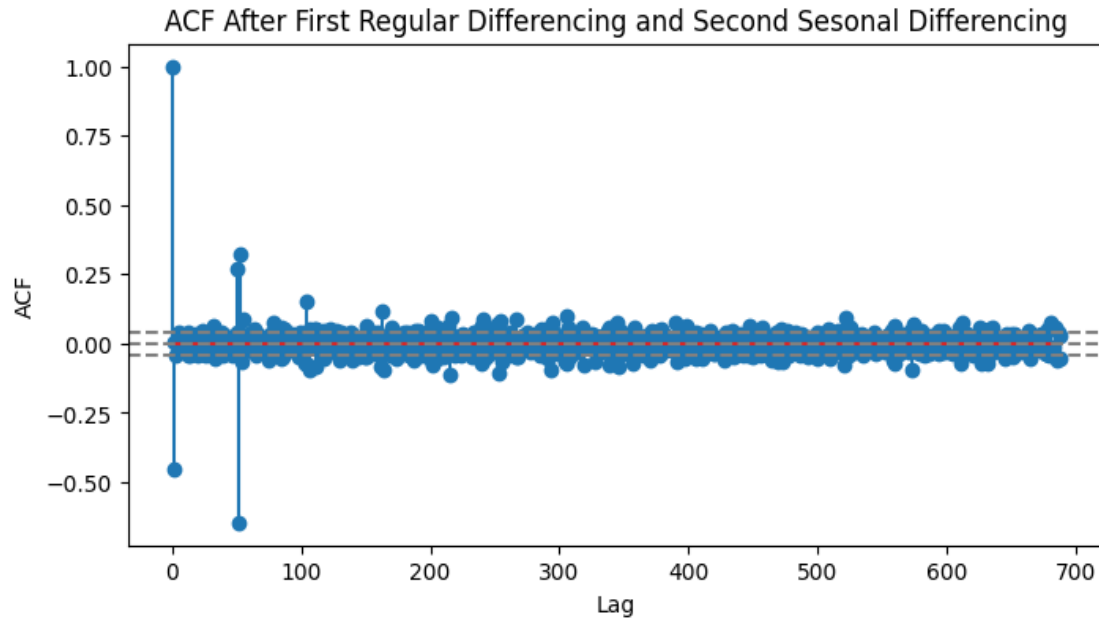


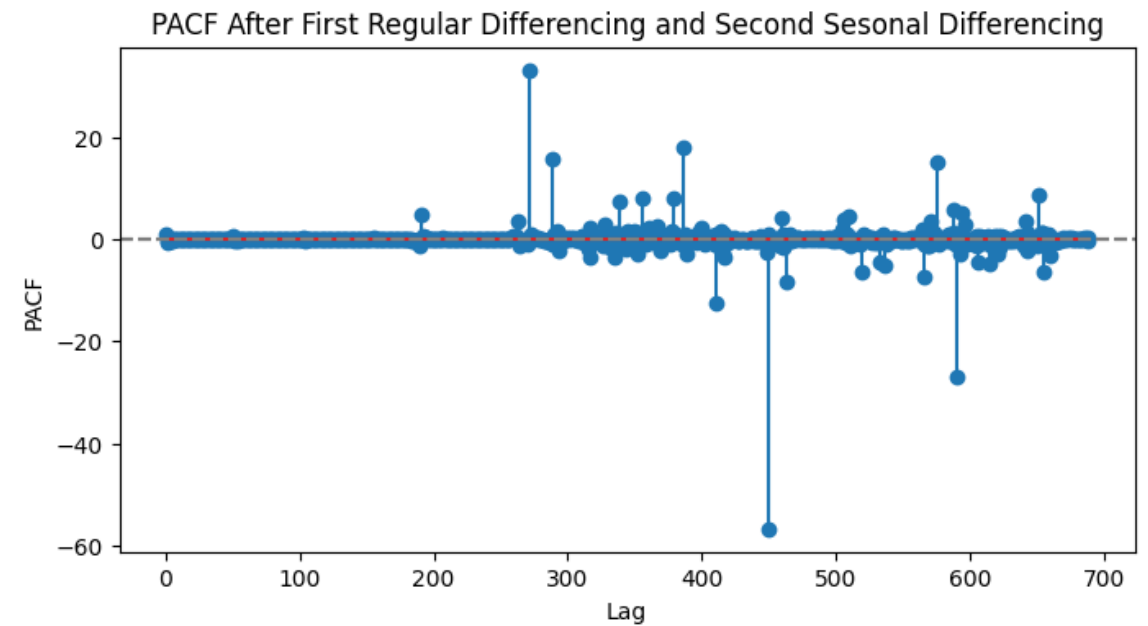
Figure 6.9: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing and second seasonal differencing up to 40 lags

6. RESULTS

SARIMA MODELLING



(a)



(b)

Figure 6.10: (a) ACF; (b) PACF graph of weekly transformed solar radiation training data after first regular differencing and second seasonal differencing up to 688 lags

6. RESULTS

SARIMA MODELLING

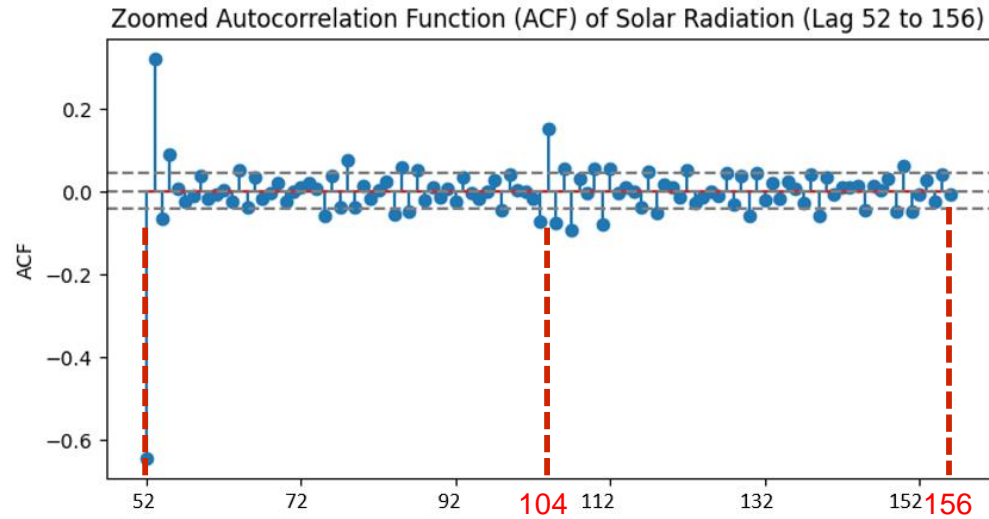


Figure 6.11: Zoomed ACF graph of weekly transformed solar radiation training data after first regular differencing and second seasonal differencing from lag 52 to 156

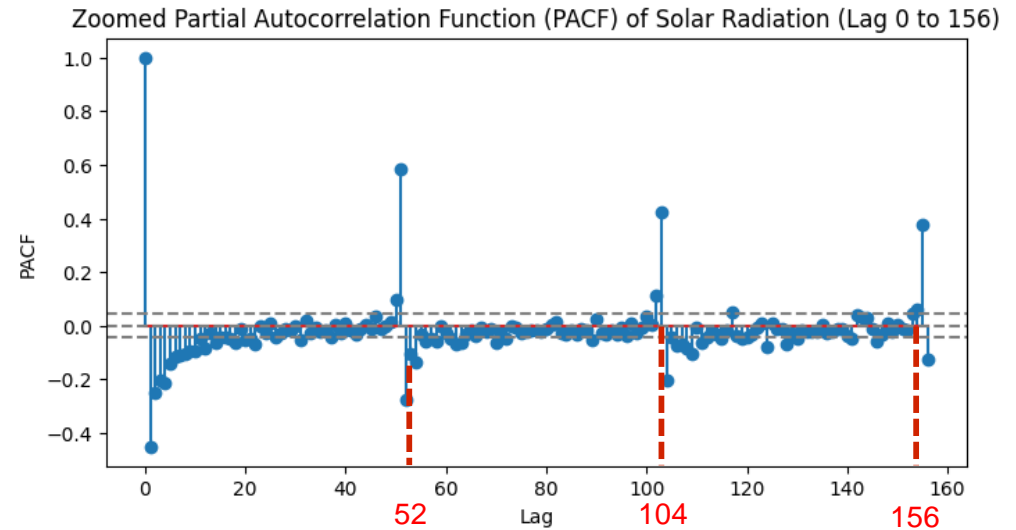


Figure 6.12: Zoomed PACF graph of weekly transformed solar radiation training data after first regular differencing and second seasonal differencing from lag 0 to 156

- The ACF shows that the spikes cut off after 104 lags, equivalent to two cycles of 52 lags.
- The PACF 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 are $SARIMA(10,1,1)(3,2,2)_{52}$.

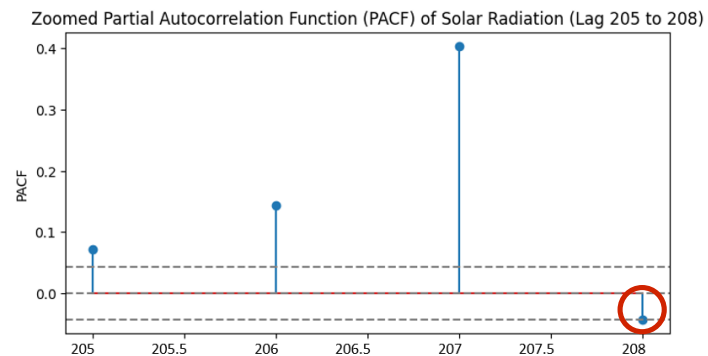


Figure 6.13: Zoomed PACF graph of weekly transformed solar radiation training data after first regular differencing and second seasonal differencing from lag 205 to 208

6. RESULTS

SARIMA MODELLING

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.

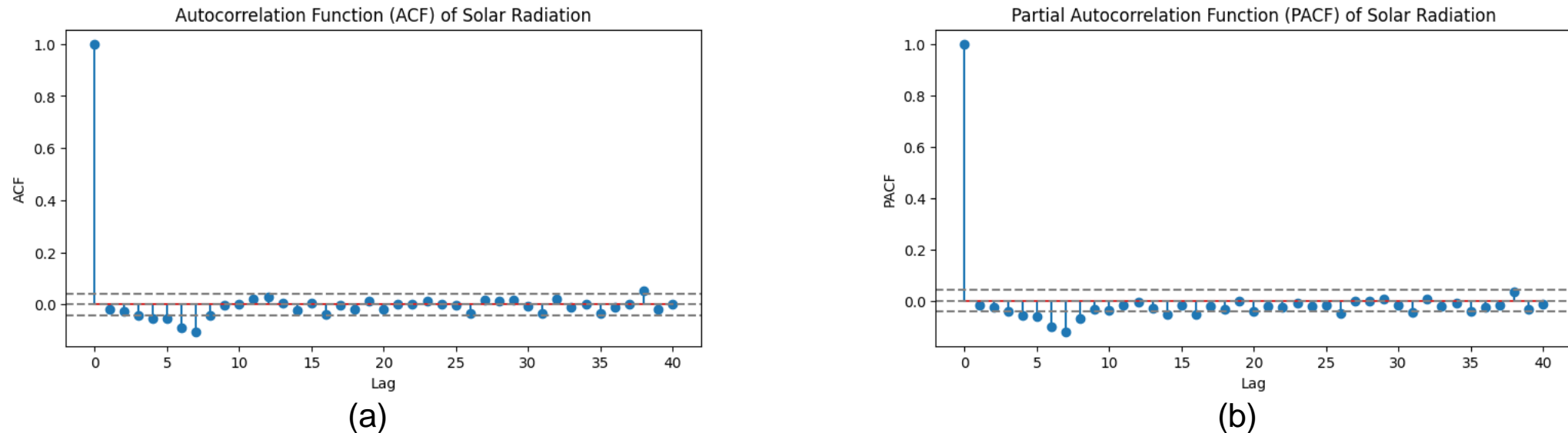


Figure 6.14: (a) ACF; (b) PACF graph of $SARIMA(6,1,0)(2,2,0)_{52}$ residual

- The majority of the residuals fall within the 95% confidence bounds (dashed lines), indicating that there is no significant autocorrelation in the residuals.

6. RESULTS

MODEL PERFORMANCE EVALUATION

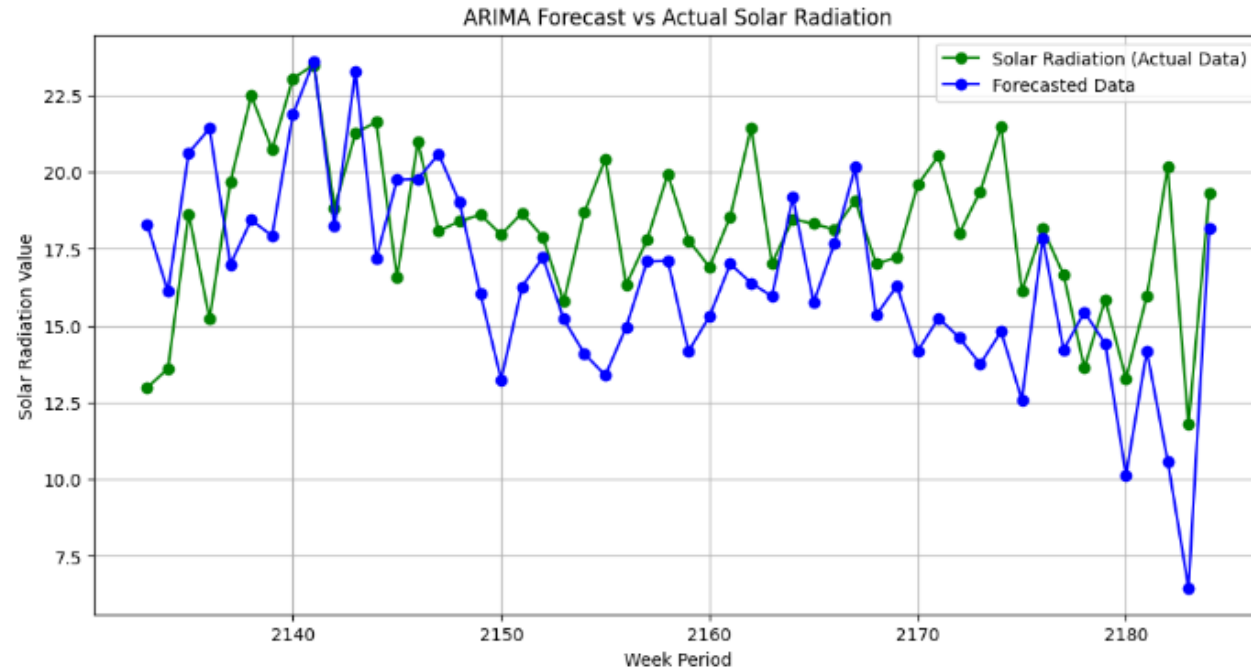


Figure 6.15: Comparison between testing data and forecasted data using SARIMA(6,1,0)(2,2,0)₅₂

- 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%.

7. DISCUSSION AND CONCLUSION

- The dataset revealed that 6.34% of the data was missing. To address this, three imputation methods were compared which are mean substitution, k-nearest neighbors and random forest. Random forest emerged as the superior method, with an RMSE of 2.592, NSE of 0.516, and MAE of 2.021, outperforming mean substitution and k-nearest neighbors method.
- 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 results demonstrate that the SARIMA $(6,1,0)(2,2,0)_{52}$ model achieved an MAE of 2.809, MSE of 12.09, RMSE of 3.477, and MAPE of 15.74%.
- 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.
- These findings may provide a useful reference for other researchers to compare with their forecasting models. By considering this approach, they could potentially reduce the time needed to achieve more accurate forecasts, helping to improve the precision of solar radiation predictions or applying similar methods to other areas such as climate modeling.

8. 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
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.

Thank you

**11th MALAYSIA
STATISTICS CONFERENCE**
"Data and Artificial Intelligence: Empowering the Future"

**19th September
2024**

Organized by:

