Seasonal Analysis and Forecasting of Solar Panel Production Using SARIMA

Joseph Tivnan  
*Wentworth Institute of Technology*

*Abstract*—This report investigates the use of SARIMA modeling to analyze and forecast solar panel production data. The analysis covers the period from January 2019 to October 2024, with a focus on seasonal trends in energy output. Stationarity was ensured through seasonal differencing, and the Augmented Dickey-Fuller test was applied to the series. A grid search was conducted to optimize model parameters, followed by the generation of 12-month forecasts. The results demonstrate the model's ability to capture seasonal fluctuations and its potential for future energy production predictions.

Keywords—SARIMA, solar energy, forecasting, seasonal modeling, time series analysis

# Introduction

The increasing global demand for renewable energy has led to significant interest in accurately predicting energy production from solar panels. Solar energy production is inherently affected by seasonal patterns, which can complicate predictions. This study investigates the use of **Seasonal Autoregressive Integrated Moving Average (SARIMA)** models to analyze and forecast monthly solar panel production data. The dataset spans from January 2019 to October 2024 and includes data from the Wentworth Institute of Technology’s solar energy system.

Understanding solar energy production patterns is critical for optimizing energy output, improving energy storage systems, and integrating solar power into the grid. Accurate forecasting can also aid in reducing reliance on non-renewable energy sources. This research aims to assess the effectiveness of SARIMA in predicting solar energy output, provide recommendations for energy production planning, and contribute to the body of knowledge in renewable energy forecasting.

# Datasets

## Source of dataset

The dataset used in this analysis was obtained from the Wentworth Institute of Technology’s solar energy system, specifically from their sustainability department. The dataset includes monthly solar panel production data, spanning from January 2019 to October 2024. This data was collected and provided by the facilities management team at Wentworth after the original subscription to their ATRIUS health system expired. The dataset is credible, as it comes directly from the institution’s solar panel monitoring system. The data was likely generated through automated sensors monitoring the performance of the solar panels, recording the energy output in watt-hours (Wh) on a monthly basis. Missing values were imputed based on historical trends, ensuring continuity for analysis.

## Character of the datasets

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Description** | **Unit** |
| **Date** | **Date of the recorded data point** | **Month/Day/Year** |
| **Production (Wh)** | **Monthly energy production** | **Watt-hours (Wh)** |

The dataset used for this analysis consists of monthly solar panel production data, spanning from January 2019 to October 2024. The data was originally collected by the Wentworth Institute of Technology’s solar panel system and includes the above columns. The dataset was cleaned to handle missing values that occurred during specific periods (e.g., January to March 2019, July and August 2022). These missing values were imputed by using the average of the respective months for other years.

# Methodology

For time series modeling, it is crucial to ensure that the data is stationary, meaning that its statistical properties (such as mean and variance) do not change over time. Non-stationary data can lead to unreliable model predictions. To check for stationarity, we used the Augmented Dickey-Fuller (ADF) test, which tests the null hypothesis that a unit root is present, indicating non-stationarity. The ADF test was applied to both the original and differenced series. For the original series, the test returned a high p-value, indicating non-stationarity, prompting us to apply seasonal differencing with a lag of 12 and regular differencing with lag 1. After differencing, the ADF test on the transformed series showed a p-value closer to zero, suggesting that the data had become stationary and was suitable for model fitting.

## Manual Parameter Estimation

The parameters for the SARIMA model were selected based on the analysis of the ACF and PACF plots. The AR(1) term was chosen due to a significant spike at lag 1 in the PACF plot, indicating the presence of a short-term autoregressive effect. Similarly, the MA(1) term was selected because the ACF plot showed a significant spike at lag 1, suggesting that the residuals from the previous time period still influence the current value, warranting the inclusion of a non-seasonal moving average term. As for the seasonal components, both the ACF and PACF plots showed no significant seasonal spikes at lag 12, leading to the conclusion that no seasonal AR or MA terms were necessary, resulting in P = 0 and Q = 0. By carefully analyzing the autocorrelation patterns, the model parameters were determined to appropriately capture both the short-term dependencies and the absence of significant seasonal effects.

(1)

Where is the observed value at time t after differencing, is the error value at time t, is the AR(1) term, and is the MA(1) term.

## Grid Search Parameter Estimation

For the grid search model, the parameters were determined by exploring all possible combinations of the non-seasonal and seasonal components within a specified range. A nested grid search was performed for the AR (p), MA (q), seasonal AR (P), seasonal MA (Q), and differencing orders (d, D) with values ranging from 0 to 1. This method involved testing 64 different combinations of parameters, evaluating each model's performance based on the AIC (Akaike Information Criterion). The combination that resulted in the lowest AIC was selected as the best model. This process ensured that the chosen parameters best captured the underlying patterns in the time series, including both short-term dependencies and seasonal effects, while minimizing overfitting.

Where is the observed value at time t after differencing, is the error value at time t, is the Seasonal AR(1) term, and is the MA(1) term, and term is the Seasonal MA(1) term

# Results

When comparing the manual model (SARIMA(1, 1, 1)(0, 1, 0)[12]) and the grid search model (SARIMA(0, 1, 1)(1, 1, 1)[12]), both models performed well in forecasting solar panel production over the next 12 months. The manual model has a slightly higher AIC of 1749.23, indicating a reasonable fit but also suggesting there might be room for improvement in capturing the seasonal patterns. The grid search model, on the other hand, had a lower AIC of 1744.91, which signifies a better fit to the data and suggests that the additional seasonal AR and MA components are helping improve the model’s accuracy.

While the grid search model performs better in terms of AIC, it is also more complex due to the inclusion of seasonal components (AR and MA terms), which can capture more intricate seasonal patterns in the data. In contrast, the manual model, which includes only one AR and one MA term without seasonal components, is simpler and may be preferred if model simplicity is a key consideration. Ultimately, the choice of model depends on the balance between model complexity and forecasting accuracy, with the grid search model providing better performance at the cost of added complexity.

## Manual Model Forecast

The manual model was built using the SARIMA(1, 1, 1)(0, 1, 0)[12] structure, with regular differencing of order 1 and seasonal differencing with a period of 12 months. The model parameters were chosen based on the analysis of the ACF and PACF plots, which indicated the inclusion of one AR term and one MA term. The model’s forecasted values for the next 12 months provide insights into the expected solar energy production. The predicted values are relatively stable with some fluctuation, reflecting the seasonal effects captured by the model. The AIC for this model is 1749.23, indicating a reasonable fit to the data, with good predictive accuracy for the short-term forecast.

A diagram of a forecast

Description automatically generated

## Grid Search Foreast

The grid search model was selected by evaluating 64 different parameter combinations, using the AIC criterion. The best parameters were found to be ARIMA(0, 1, 1)(1, 1, 1)[12], which includes both seasonal AR and MA terms in addition to the regular AR and MA terms. This model captures both short-term dependencies and seasonal variations in the data. The forecasted values from the grid search model reflect a more dynamic seasonal pattern compared to the manual model, with some noticeable peaks and valleys over the forecast period. The AIC for this model is 1744.91, indicating a slightly better fit compared to the manual model, suggesting that the seasonal components further improve the predictive capability of the model.

A graph of a graph showing the time and the time

Description automatically generated with medium confidence

# Discussion

While the SARIMA models provided valuable insights into forecasting solar panel production, there were several limitations in the current analysis. One key limitation was the imputation of missing data, specifically for the periods between January and March 2019 and July and August 2022. Imputing missing values using the mean of their respective months may have introduced biases or inaccuracies, potentially affecting the model's performance. These imputations may not fully capture the natural variability of the data during these periods, which could have reduced the model's accuracy.

Additionally, the ADF test returned a p-value that was not as low as expected, indicating that the series was not fully stationary even after applying both seasonal and regular differencing. This suggests that there may still be underlying trends or structural breaks in the data, which were not captured by the model, leading to less reliable forecasting results. Future work could involve testing alternative methods for handling missing data and further improving stationarity testing to ensure a more accurate and stable model.

In terms of improvements, exploring different imputation methods for missing data or refining the differencing approach could improve model accuracy. Additionally, expanding the dataset to cover longer periods or using more granular data could help the model capture more complex trends and improve forecasting performance.

# Conclusion

This analysis aimed to forecast solar panel production using SARIMA models, with a focus on capturing both short-term dependencies and seasonal variations. The manual model provided a solid baseline, while the grid search model showed improved accuracy by incorporating seasonal components. Despite the overall success of the models, limitations such as missing data imputation and the failure to achieve a fully stationary series highlight areas for future refinement. The results demonstrate the importance of model selection and parameter tuning, especially when dealing with time series data with seasonal components. Moving forward, improving the imputation method and testing alternative differencing strategies could further enhance the predictive power of the model. Ultimately, this study contributes to the growing field of renewable energy forecasting by showcasing the potential of time series models to predict solar energy production.