Forecasting of total daily solar energy generation using ARIMA: A case study

Conference Paper · January 2019

DOI: 10.1109/CCWC.2019.8666481

4 authors:

Sharif Atique
Texas Tech University
10 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Subrina Noureen
Texas Tech University
13 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Vishwajit Roy
Texas Tech University
11 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Stephen Bayne
Texas Tech University
196 PUBLICATIONS **1,298** CITATIONS
SEE PROFILE

Forecasting of total daily solar energy generation using ARIMA: A case study

Sharif Atique
Dept. of Electrical and Computer Engineering
Texas Tech University
Lubbock, TX, USA
taufique.atique@ttu.edu

Subrina Noureen
Dept. of Electrical and Computer Engineering
Texas Tech University
Lubbock, TX, USA
subrina.noureen@ttu.edu

Vishwajit Roy
Dept. of Electrical and Computer Engineering
Texas Tech University
Lubbock, TX, USA
vishwajit.roy@ttu.edu

Vinitha Subburaj
Dept. of Computer Science
Texas Tech University
Canyon, TX, USA
vsubburaj@wtamu.edu

Stephen Bayne
Dept. of Electrical and Computer Engineering
Texas Tech University
Lubbock, TX, USA
stephen.bayne@ttu.edu

Joshua Macfie
Electrical Engineer
Group NIRE
Lubbock, TX, USA
joshua.macfie@groupnire.com

***Abstract*—In this paper, a well known statistical modeling method named ARIMA has been used to forecast the total daily solar energy generated by a solar panel located in a research facility. The beauty of the ARIMA model lies in its simplicity and it can only be applied to stationary time series. So our time series data, which is seasonal and non-stationary, is transformed into a stationary one for applying the ARIMA model. The model is developed using sophisticated statistical techniques. The optimum model is chosen and validated using Akaike information criterion (AIC) and residual sum of squares (SSE). Error analysis is done to demonstrate the efficiency of the proposed method. The accuracy of the developed model can be further increased, which is subject to future research.**

***Keywords*—Forecasting, solar, ARIMA, time series, stationarity, generation**

I. Introduction

The electricity demand in the world is always increasing. However, the traditional source of fossil fuel is limited and they leave significant carbon footprint. These factors, along with the technological advancements, have driven the increasing usage of distributed renewable energy resources [**?**]. Penetration of renewable energy resources are only going to increase as the grid is becoming smarter. Among all the renewable resources, photovoltaic (PV) based solar energy is the most promising [**?**]. However, like any other renewable resources, solar energy is inherently uncertain as it is heavily dependent on solar irradiance and other environmental factors like humidity, temperature and geographic location [**?**]. As a result, forecasting plays a significant role in PV based systems for operation and planning purposes [**?**]. Accurate solar energy generation forecast can help with mitigating the uncertainty and result in better demand side management [**?**]. As the amount of solar power generation is uncertain, it can be modeled as a stochastic time series model [**?**].

Time series forecasting method is particularly useful when there is little knowledge about the effects of explanatory variable on the output. In a time series model, a dependent variable or output is dependent only on its past values. After a model has been established, it is then used to predict the future values [**?**]. Time series methods are well studied in the forecasting area and it is continually being improved. One of the most well studied models of this arena is ARIMA, acronym for Autoregressive Integrated Moving Average. The reasons behind the popularity is simplicity of implementation and use of the famous Box-Jenkins methodology [**?**]. This simplicity is a result of the linear correlation assumption between time series values of the past and present. This is a major drawback of ARIMA model even though it can model various types of time series data. As complex real world time series data is not always linear, ARIMA models might not be the best solution. In cases like this, there are other statistical models that can be used to incorporate the non-linearity. In spite of all this, ARIMA model is good benchmark for solar energy forecasting.

In this paper, ARIMA models, both the seasonal and non-seasonal variations, have been studied to predict the daily total solar energy generation of a 10kW solar panel. This solar panel is installed in the rooftop of Group Nire building in the Reese Reseach Center located in Lubbock, TX. The time series data is transformed to a stationary one, analyzed for determining the model parameters and validated using various criteria like Akaike Information Criterion (AIC) and sum of square of residuals (SSE). Finally, the performance of the model is judged by necessary error analysis.

II. ARIMA Modeling

A. General Formulation

The value of a dependent variable is expressed as a linear relationship between past values of the dependent variable and random errors in ARIMA model. However, a time series can only be modeled as a ARIMA process if it is stationary. As strong stationarity is somewhat complex to demonstrate [**?**], in this paper we would assume stationarity if the time series is weakly stationary. In general terms, a weakly stationary time series has constant statistical properties, namely mean and variance [**?**]. Transformation operations like differencing, logging and deflating [**?**] are performed on a non-stationary

978-1-7281-0554-3/19/$31.00 ©2019 IEEE

time series to make it stationary. There are two different variations of ARIMA models: non-seasonal and seasonal. If there is seasonality in the time series data, then seasonal ARIMA model is used. Otherwise, the non-seasonal ARIMA model is used for the general cases.

The non-seasonal ARIMA is modeled in the following way [**?**]:

$$\hat{y_t} = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}, \quad (1)$$

where $\hat{y_t}$ is the d^{th} difference of a non-stationary time series **Y**. The order of autoregressive lag terms, differencing and moving average lag terms are represented by p,d and q, respectively. The autoregressive and moving average parameters are expressed with ϕ and θ terms, respectively. Finally, μ is a constant.

Depending on the values of p,d and q, an ARIMA process can undertake the form of purely moving average (MA), purely autoregressive (AR) or autoregressive moving average (ARMA) processes.

Presence of a periodic pattern in the time series is called seasonality. Seasonality in a time series is expressed by its span, **S**. For example, monthly solar energy generation has higher values in summer months, so $S = 12$ in this case. ARIMA models can be used to forecast seasonal time series data, just like non-seasonal time series data.

A multiplicative model, including both the non-seasonal and seasonal fluctuations, is used to represent seasonal ARIMA model. Seasonal ARIMA is generally expressed in the following way [**?**]:

$$ARIMA(p,d,q) * (P,D,Q)_S, \quad (2)$$

where

- P = seasonal AR order
- Q = seasonal MA order
- D = seasonal differencing order
- S = span of pattern in seasonality

A more formal representation of seasonal ARIMA model is as follows:

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - \Phi_1 B^S - \cdots - \Phi_P B^{PS})(1 - B)^d (1 - B^S)^D y_t = (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B^S + \cdots + \Theta_Q B^{QS})\epsilon_t, \quad (3)$$

where B is the backshift operator, whose operation is governed by 4:

$$B^m y_t = y_{t-m} \quad (4)$$

B. Model Parameter Selection

The first step in the modeling process is checking for the stationarity of the time series. A rough estimate of stationarity can be graphically obtained by plotting the partial auto correlation function (PACF)and auto correlation function (ACF) plots of the time series. ACF measures the correlation of a time series value with other values of the same time series at different lags. PACF also measures the correlation between a value of a time series and another value at different lag. However, PACF ignores the other values at different lags while calculating the correlation for a particular lag value [**?**]. If the ACF doesn't display any significant value after a few lags or the PACF contains a sharp cutoff after the initial value [**?**], then we have a stationary time series on our hand. However, most real life problems are not as straightforward and stationary.

After the initial estimation, a more methodical approach, named Augmented Dickey Fuller (ADF) test, is executed to confirm stationarity [**?**], [**?**], [**?**]. ADF is also known as unit root test. If there is no unit root of the characteristic equation, then the time series is stationary. Otherwise, the time series in non-stationary.

The general equation for testing stationarity using the ADF test is as follows:

$$\partial Y_r = \mu + \beta t + \rho Y_{t-1} + \partial_1 Y_{t-1} + \cdots + \partial_p Y_{t-p} + e_t. \quad (5)$$

Here, β represents the trend. Moreover, e_t represents a sequence of independent normal random variables of zero mean and unit variance. Then hypothesis is formulated in the following way [**?**]:

$$NullHypothesis : H_0 :| \rho |= 0(Non - stationarity)$$
$$AlternativeHypothesis : H_1 :| \rho |\neq 0(Stationarity)$$

Rejection or acceptance of the null hypothesis is dictated by the p-value. A confidence level of 95% is assumed in this work. If $p \geq 0.05$, the time series is non-stationary (null hypothesis is true). Otherwise, the time series is stationary (null hypothesis is rejected)

C. Model Selection and Validation

After the initial checking of stationarity, differencing operation is performed in case the time series is non-stationary. If the initial time series is stationary, then the order of differencing, $d = 0$. Differencing would be performed as long as the time series isn't transformed into a stationary one. In this work, other transformation techniques are not studied. After each differencing operation, the stationarity can be checked using the ACF and PACF plots or ADF test or both. Finally, the PACF and ACF plots of the derived stationary time series would determine the p and q parameters. p and q generally correspond to significant terms in PACF and ACF plots, respectively. However, they might not always be the optimum model parameters. The seasonal parameters can also be determined from the ACF and PACF plots.

The final step before forecasting is selection of the optimum ARIMA model. The following criteria are commonly used to estimate the goodness of fit for the developed models:

1) Akaike Information Criterion (AIC)
2) Corrected Akaike Information Criterion (AICc)
3) Bayesian Information Criterion (BIC)
4) Residual sum of squares (SSE)

1) AIC: The formulation of AIC ([**?**], [**?**]) is as follows:

$$AIC = -2log(maximumlikelihood) + 2k, \quad (6)$$

where k is independently adjusted number of parameters.

2) AICc: The formulation of AICc [**?**] is as follows:

$$AICc = -2log(maximumlikelihood) + \frac{n+k}{n-k-2}, \quad (7)$$

where n is total number of data points.

3) BIC: The formulation of BIC ([**?**]) is as follows:

$$BIC = -2log(maximumlikelihood) + \frac{klogn}{n}, \quad (8)$$

where k and n are the same as defined in AIC and AICc.

The preferred model is the one that minimizes all these criteria. In this work, AIC and SSE have been used for optimum model selection.

III. Data Preparation

We have used total daily solar energy generation (in kWh) as our dependent variable. The data has been collected for a complete year (6^{th} November, 2017 - 5^{th} November, 2018) from the 10kW solar plant located in the rooftop of the Group Nire building in Reese Research Center, Lubbock, TX. The data was initially stored in .csv format. The data was read and plotted as a time series (1) using the prominent statistical software R.

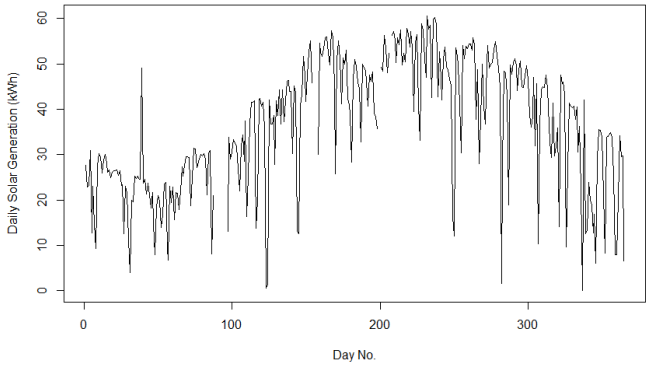

Fig. 1: Daily solar energy generation data before cleaning

As evident from 1, there are a few missing data points, 15 to be exact, as the solar panel wasn't functional on those days. So, this data was processed for filling up the missing data using the $tsclean()$ function in R and plotted in 2. This time series is eventually utilized for the analysis in this paper.

IV. Analysis

An initial assumption about stationarity of our data set can be made just by looking at 2. Two trends, one upward and one downward, can be guessed from this figure. However, decision about stationarity is made after plotting of the necessary autocorrelation functions and performing ADF test.

The ACF and PACF of the cleaned time series are plotted in 3 and 4, respectively.

In 3, ACF of the time series data has been plotted. It is evident from this figure that the ACF doesn't become insignificant after a few lags. In fact, the ACF remains significant

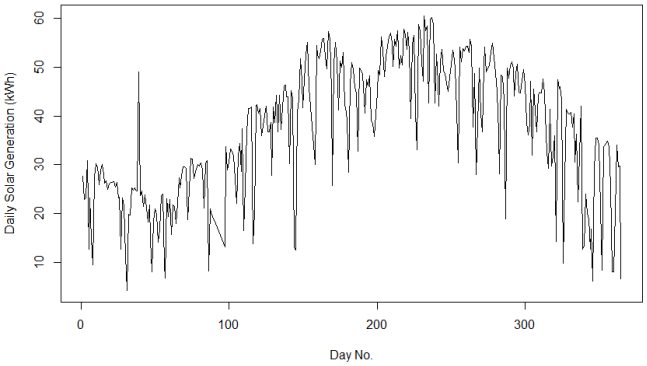

Fig. 2: Daily solar energy generation data after cleaning

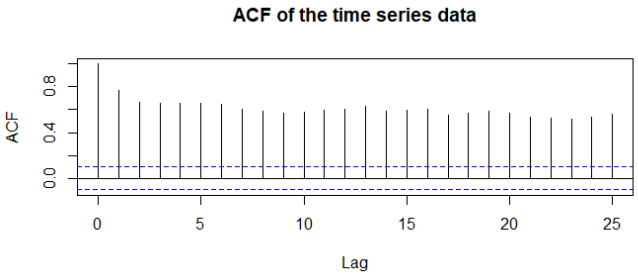

Fig. 3: ACF plot of the original time series data

even after 50 lags (5). There is also some periodicity in the ACF plot, which implies seasonality. In the PACF plot in 4, there is not sharp cutoff. So the graphical test confirm non-stationarity of the time series. However, the ADF test still need to be performed to be certain about the conclusion on the non-stationarity of the time series. The result of the ADF test is summarized in 6.

As we can see from 6, the p-value is 0.6639. So the null hypothesis can't be rejected and the time series can be declared as non-stationary.

If a time series has inherent trend and seasonality, then the time series is always non-stationary as they imply systemic variation in mean and variance. The seasonal component, trend

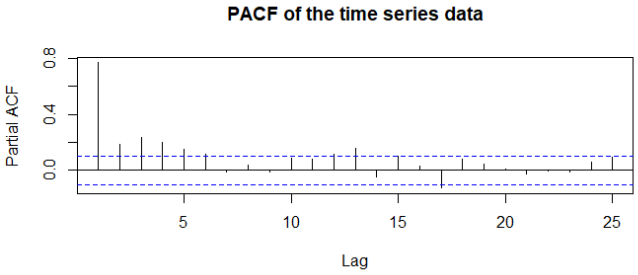

Fig. 4: PACF plot of the original time series data

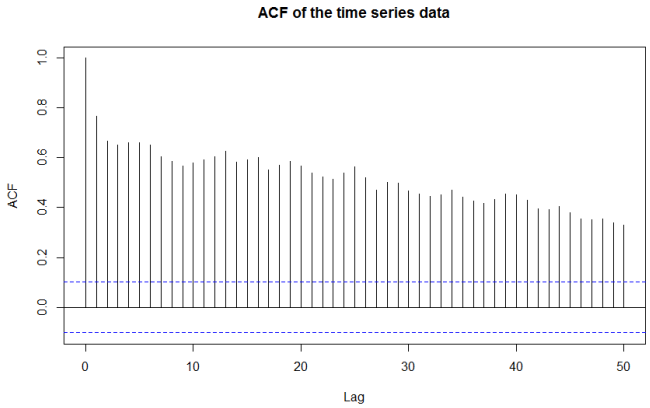

Fig. 5: ACF plot of the original time series data upto 50 lags

```
        Augmented Dickey-Fuller Test

data:  data.sum.ts.clean
Dickey-Fuller = -1.7941, Lag order = 7, p-value = 0.6639
alternative hypothesis: stationary
```

Fig. 6: ADF test of the original time series data

component and the residuals of the time series is plotted in 7. We can see from this figure that both the trend and seasonality are present in our data set. So necessary transformation techniques need to be carried out in order to make our data a stationary one. As previously mentioned, only the differencing operation is covered in this paper.

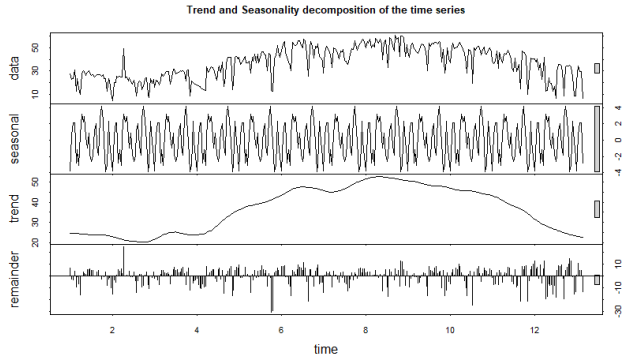

Fig. 7: Trend and seasonality decomposition of the time series

A first order differencing operation is performed on the time series and the differenced time series is plotted in 8. The differenced time series looks stationary in the first look as no systemic variation in mean and variance is readily evident. This claim is further enhanced by both graphical and mathematical analyses.

As we see from 9, ACF becomes insignificant after 2 lags, if we neglect the sparse significant ACFs at higher lags. Partial autocorrelation functions of the differenced time series mostly become insignificant after 5 lags (10), if the sparse significant values at higher lags are ignored. These two plots indicate possible stationarity of the differenced time series,

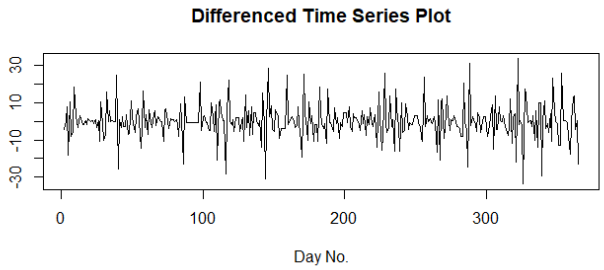

Fig. 8: Plot of the time series data after differencing

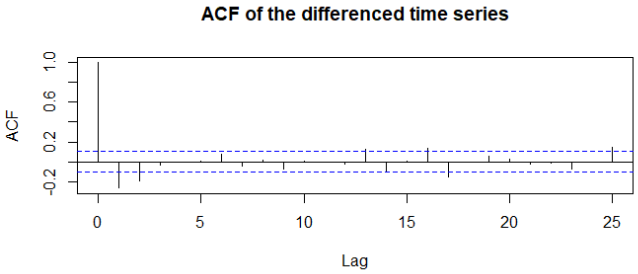

Fig. 9: ACF plot of the time series data after differencing

which is confirmed using ADF test subsequently. Moreover, these significant lags from the ACF and PACF plots help us with initial assumption of AR and MA orders in the ARIMA model.

Finally, the ADF test is performed on the differenced time series data and the result is summarized in 11. The calculated p-value is 0.01, which confirms the rejection of null hypothesis and subsequent stationarity. So our desired stationary time series is produced with first order differencing operation of the actual time series data that was collected and cleaned.

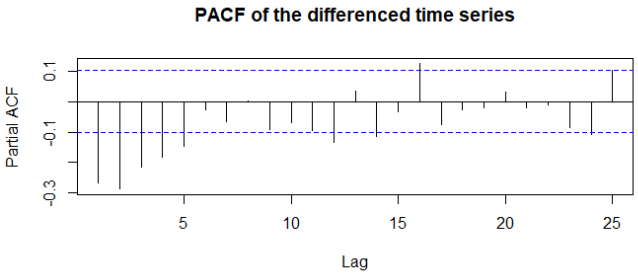

Fig. 10: PACF plot of the time series data after differencing

```
        Augmented Dickey-Fuller Test

data:  c.ts
Dickey-Fuller = -9.1105, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Fig. 11: ADF test for the differenced time series data

	AIC	AICc	BIC
Seasonal	2507	2507.16	2526.48
Non-seasonal	2513.78	2513.89	2529.37

TABLE I: Performance comparison of seasonal and non-seasonal model using auto.arima() routine

TABLE II: Performance comparison of ARIMA models

p	d	q	P	D	Q	S	AIC	SSE	p-value
0	1	1	1	0	1	30	2500.358	18002.47	0.004
0	1	2	0	0	1	30	2508.908	20450.01	0.99
0	1	2	1	0	1	30	2480.266	16983.22	0.99
1	1	1	1	0	1	30	2481.902	17067.95	0.89
1	1	2	1	0	1	30	2482.236	17076.8	0.99

V. Model Validation

The span of seasonality is not apparent from figures 9 and 10. So, an initial auto.arima() routine is applied on the time series to obtain the optimum seasonal periodicity. The auto.arima() routine yielded the model ARIMA$(0,1,2)(0,0,2)_{30}$. Other periodicity was randomly chosen and tested against 30. However, periodicity of 30 performed better than the other values in terms of minimized AIC. The auto.arima() routine is also performed without the seasonality option. However, the seasonal model outperforms the non-seasonal one in terms of all the information criteria, which should be the case as our time series has seasonality in it. So the further analyses in this work will be solely focused on the seasonal model. The result is summarized in I.

The approximate non-seasonal AR and MA orders are resembled by significant terms in PACF and ACF plots, respectively. So, the AR and MA orders in this work should be approximately 5 and 2, as evidenced by 10 and 9. However, higher orders in the model bring increased cost and complexity. In order to ensure simplicity and reduced cost, the non-seasonal AR and MA orders are limited to 2 and seasonal AR and MA orders are limited to 1. So the constraints are:

$$0 \leq p, q \leq 2 \tag{9}$$

$$0 \leq P, Q \leq 1 \tag{10}$$

Arima models are simulated based on these constraints and the results of the 5 best models are summarized in II. Results from all the models are not presented due to space constraint.

The p-value, obtained by performing Ljung-Box test [**?**] on the residuals, is needed to reject the hypothesis that there is no autocorrelation among the model residuals. So p-value needs to be more than 0.05 for a 95% significance level. From II, it is obvious that ARIMA$(0,1,2)(1,0,1)_{30}$ model outperforms all the other models both in terms of AIC and SSE. The residual analysis of this model is presented in 12. It is obvious that there is no significant correlation between the residuals and the residuals mostly follow the normal distribution, except for the lower tail, which proves the rationale of our model.

So, our model equation has the form:

$$(1-\Phi_1 B^{30})(1-B)y_t = \mu + (1+\theta_1 B+\theta_2 B^2)(1+\Theta_1 B^{30})\epsilon_t. \tag{11}$$

The relevant parameter values are summarized in III. Only

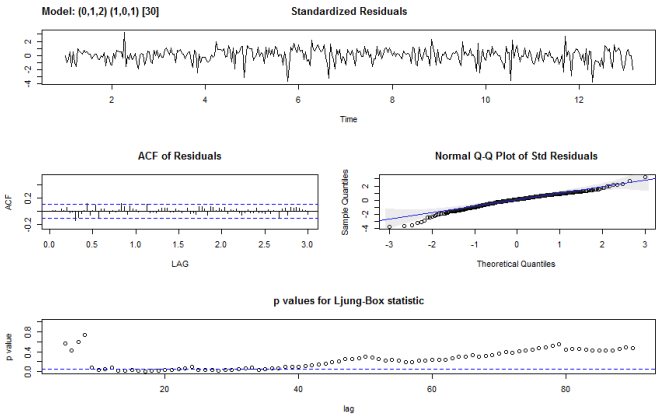

Fig. 12: Residual analysis of ARIMA$(0,1,2)(1,0,1)_{30}$ Model

TABLE III: Model parameter values

Parameter	Estimated value	p-value
Φ_1	0.6543	0
θ_1	-0.5319	0
θ_2	-0.2578	0
Θ_1	-1	0
μ	0.0201	0.6116

the constant μ has an associated p-value of greater than 0.05, which makes it insignificant. So the constant value has been disregarded in the final simplified forecasting equation in 12.

$$y_t = y_{t-1} + \Phi_1 y_{t-30} - \Phi_1 y_{t-31} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \Theta_1\epsilon_{t-30} + \theta_2\Theta_1\epsilon_{t-31} + \theta_2\Theta_1\epsilon_{t-32}. \tag{12}$$

Finally, equation 12 is used to forecast the value of total daily sonar energy generation for a particular day. In this work, the forecasted values for the last 30 days in our dataset have been compared with the actual values and demonstrated in 13.

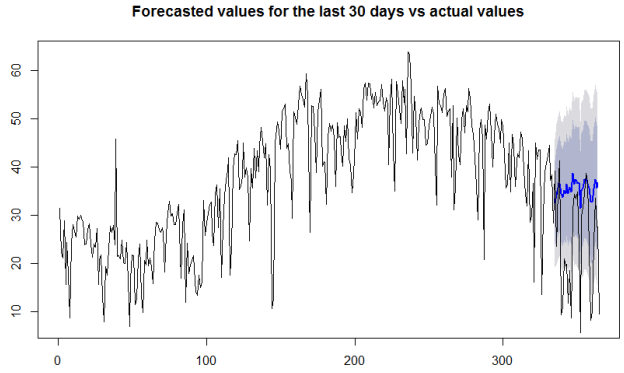

Fig. 13: Forecasted vs actual values for the last 30 days

After forecasting the values, the accuracy of the model is tested using R. The mean absolute percentage error (MAPE) for our model is 17.70%.

VI. Conclusion and Future Work

In this work, a model has been established and tested for predicting the total daily solar energy generation of a research facility using the popular and simple statistical time series method called ARIMA. The model is developed using techniques like differencing, ACF, PACF and ADF test. The model is validated using AIC and SSE. The MAPE is slightly high, however this doesn't necessarily indicate an issue in the modeling process. There might be certain factors affecting the accuracy of the developed model which should be further investigated. Upon inspecting the original time series, visible fluctuations in the last 30 day period can be spotted. This volatility might be tackled better using something like moving average of solar outputs, instead of daily data. Better smoothing techniques would be studied in future research. Although a stationary time series was obtained using a differencing operation. The variance still displayed possible heteroscedasticity. So forecasting of the time series data using models like generalized autoregressive conditional heteroscedasticity (GARCH) and autoregressive conditional heteroscedasticity (ARCH) would also be studied in future. Finally, this work can be used as a good building block for further research into forecasting of renewable energy generation.

References

[1] I. Khan, H. Zhu, J. Yao, and D. Khan, "Photovoltaic power forecasting based on elman neural network software engineering method," in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Nov 2017, pp. 747–750.

[2] I. Majumder, M. K. Behera, and N. Nayak, "Solar power forecasting using a hybrid emd-elm method," in *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, April 2017, pp. 1–6.

[3] M. Z. Hassan, M. E. K. Ali, A. B. M. S. Ali, and J. Kumar, "Forecasting day-ahead solar radiation using machine learning approach," in *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, Dec 2017, pp. 252–258.

[4] V. P. Singh, V. Vijay, M. S. Bhatt, and D. K. Chaturvedi, "Generalized neural network methodology for short term solar power forecasting," in *2013 13th International Conference on Environment and Electrical Engineering (EEEIC)*, Nov 2013, pp. 58–62.

[5] J. Wu and C. K. Chan, "The prediction of monthly average solar radiation with tdnn and arima," in *2012 11th International Conference on Machine Learning and Applications*, vol. 2, Dec 2012, pp. 469–474.

[6] G. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159 – 175, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231201007020

[7] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco, CA, 1970.

[8] D. E. Myers, "To be or not to be... stationary? that is the question," *Mathematical Geology*, vol. 21, no. 3, pp. 347–362, Apr 1989. [Online]. Available: https://doi.org/10.1007/BF00893695

[9] M. Poulos and S. Papavlasopoulos, "Automatic stationary detection of time series using auto-correlation coefficients and lvq — neural network," in *IISA 2013*, July 2013, pp. 1–4.

[10] "Introduction to arima: nonseasonal models," https://people.duke.edu/ rnau/411arim.htm, accessed: 2018-10-08.

[11] R. Shumway and D. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, ser. Springer Texts in Statistics. Springer New York, 2010. [Online]. Available: https://books.google.com/books?id=dbS5IQ8P5gYC

[12] J. H. F. Flores, P. M. Engel, and R. C. Pinto, "Autocorrelation and partial autocorrelation functions to improve neural networks models on univariate time series forecasting," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, June 2012, pp. 1–8.

[13] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Australia: OTexts, 2018.

[14] D. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the Americal Statistical Association*, vol. 74, no. 366, pp. 427–431, 1979.

[15] S. Halim, I. N. Bisono, Melissa, and C. Thia, "Automatic seasonal auto regressive moving average models and unit root test detection," in *2007 IEEE International Conference on Industrial Engineering and Engineering Management*, Dec 2007, pp. 1129–1133.

[16] J. Wu and C. K. Chan, "The prediction of monthly average solar radiation with tdnn and arima," in *2012 11th International Conference on Machine Learning and Applications*, vol. 2, Dec 2012, pp. 469–474.

[17] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, December 1974.

[18] S. Halim, I. N. Bisono, Melissa, and C. Thia, "Automatic seasonal auto regressive moving average models and unit root test detection," in *2007 IEEE International Conference on Industrial Engineering and Engineering Management*, Dec 2007, pp. 1129–1133.

[19] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 03 1978. [Online]. Available: https://doi.org/10.1214/aos/1176344136

[20] G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *Journal of the American Statistical Association*, vol. 65, no. 332, pp. 1509–1526, 1970.