Statistical Analysis and Forecasting of Solar Energy - Inter States

(APPLIED STATISTICAL METHODS)

GROUP 9 - FRIEDMAN GROUP

Kalit Inani
Ayush Singh
Mohul Maheshwari
Himanshu Pandey
Ayush Upadhyay
Sobat Singh
Karanvir Singh Sidana
Himanshu Verma

November 27, 2020

Birla Institute of Technology and Science, Pilani

1. INTRODUCTION

Solar energy is the radiant light and heat energy from the Sun that is mobilized and utilized using a multitude of progressive technologies like photovoltaics, thermal power plants, concentrated solar power systems, artificial photosynthesis, etc. It is a renewable energy resource, which is widely regarded as the most pollution-free, abundant and safest resource of energy available. The development of cost-effective and inexhaustible solar technology resources can help increase sustainability, mitigate pollution, reduce the dependence on fossil fuels and control global warming. With the focus on renewable energy resources due to environmental reasons being the need of the hour, analysing and forecasting solar energy is of great importance for developing systems that harness this energy more effectively. This report focuses on the inter state analysis of solar energy data, using various statistical methods. The methodology followed in the report can be broadly divided into the following sections-

1- Analysis of Solar Energy parameters and their correlation
2- Determining a suitable distribution fit for the data
3- Time series analysis and decomposition
4- Stationarity analysis
5- Finding a suitable model for forecasting
6- Spatial analysis of the results for Inter-states

2. DESCRIPTIVE ANALYSIS

In the available data, we saw that there are various parameters that can affect the solar radiation reaching the surface of the earth. So let us define these parameters so that we can better understand the problem and get an idea on how these variables affect each other and also the solar radiation value. Some important terminologies are discussed below:

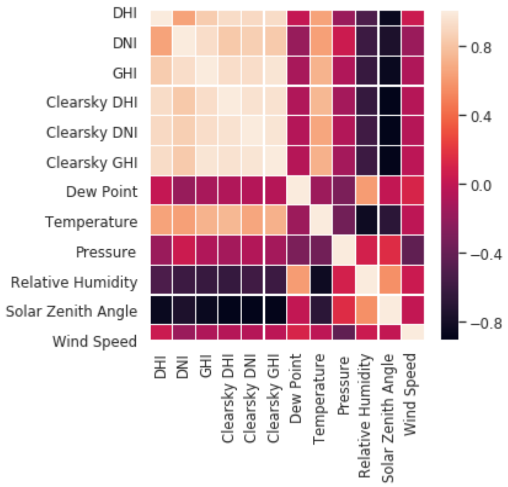
1. **Diffuse horizontal irradiance(DHI)**: The amount of sun's radiation reaching the terrestrial surface per unit area which is scattered by the atmosphere is known as Diffuse Horizontal Irradiance. It does not include the radiation that comes on the direct path from the sun.
2. **Direct Normal Irradiance(DNI)**: As the name suggests, the radiations received per unit area from the sun to the earth's surface that are normal to the surface is known as Direct Normal Irradiance. Typically, we can maximize the amount of irradiance annually received by a surface by keeping it normal to incoming radiation.
3. **Solar-Zenith Angle**: It is defined as the angle between the vertical and the sun's ray.
4. **Global Horizontal Irradiance(GHI)**: The total amount of short wave radiation received per unit area from above by the horizontal earth's surface is known as GHI or Global Horizontal Irradiance. Its value includes both the DHI and DNI components for a particular value of the Solar-Zenith Angle. Mathematically it can be expressed as:

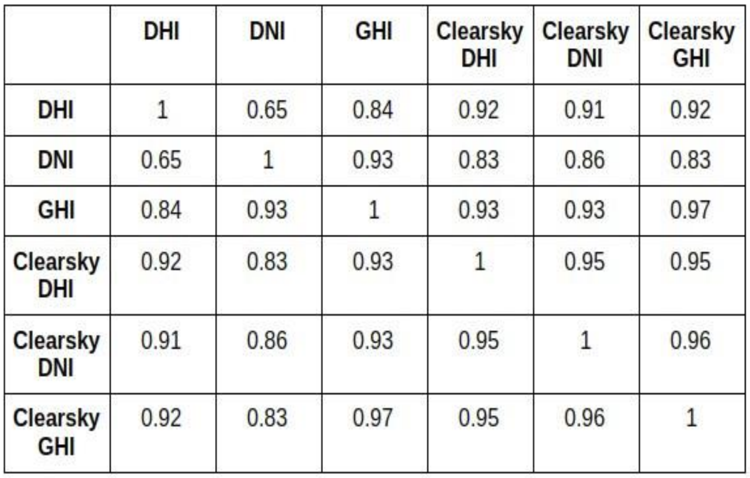
$$\mathbf{GHI = DHI + DNI \cdot \cos(\theta)} \quad \text{(where } \theta \text{ is the solar zenith angle)}$$

The data also includes several other variables like Clearsky DHI, Clearsky DNI, Clearsky GHI, Dew Point, Temperature, Pressure, Relative Humidity, Snow Depth and Wind Speed. The explanation of all these variables is beyond the scope of the report.

3. DESCRIPTIVE STATISTICS

We will first try to understand how all the variables are correlated with each other using a heatmap in order to get a bigger picture of our dataset. Please note that here we have only used data of the state of Andhra Pradesh from the year 2000 to 2014. It is expected that the data of the other states will also show similar behavior for the various parameters.





	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI
DHI	1	0.65	0.84	0.92	0.91	0.92
DNI	0.65	1	0.93	0.83	0.86	0.83
GHI	0.84	0.93	1	0.93	0.93	0.97
Clearsky DHI	0.92	0.83	0.93	1	0.95	0.95
Clearsky DNI	0.91	0.86	0.93	0.95	1	0.96
Clearsky GHI	0.92	0.83	0.97	0.95	0.96	1

Fig 3.1:Correlation heatmap *Table 3.1:Correlation Table*

- From the Heatmap it's clearly evident that the **GHI, DHI and DNI**, **Clearsky GHI**, **Clearsky DHI**, **Clearsky DNI,** are highly positively correlated. Each of the pairs formed from above variables have a correlation value of 0.8 or more(except DNI & DHI, which have correlation value of 0.65). Since the above variables are highly correlated, therefore from the above observations we can safely choose one variable for the purpose of our Analysis. We will choose GHI here, since it incorporates both the diffusive as well as direct component of irradiation along with the solar-zenith angle.
- Following are the correlation values of GHI with the remaining variables:

S. No	Variable	Correlation value
1	Solar Zenith Angle	-0.86
2	Relative Humidity	-0.63
3	Temperature	0.72
4	Pressure	0.07
5	Dew Point	-0.11
6	Wind Speed	0.08

Table 3.2: Correlation of GHI with other variables

- **Solar Zenith Angle, Relative Humidity** are negatively correlated while the temperature is positively correlated with GHI and the correlation values are quite significant
- **Wind Speed, Dew Point** and **Pressure** have a very low value of correlation and we can assume that any change in the value of these will not alter the value of GHI significantly.

4. GHI DATA ANALYSIS

In order to test whether these datasets follow our specific distributions, our datasets GHI and wind speed, we used a software jupyter notebook along with python libraries to find which model best fits the data.

- From the below PP plot it is clear that the distribution does not follow normal distribution.
- Next we use Fitter library to get graphical representation and AIC values for different distributions.

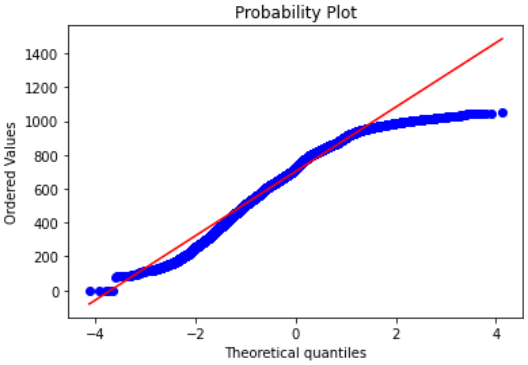


Figure 4.1 : PP plot for normal distribution

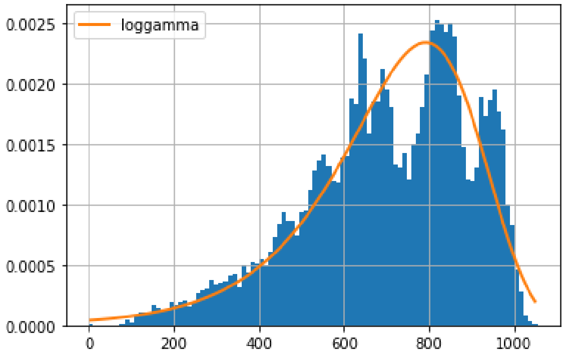


Figure 4.2 : Fitting log gamma distribution

AIC (Akaike Information Criterion):

A good model is the one that has minimum AIC among all the other models. **AIC = 2K - 2ln(L')** where L is the likelihood, and k is number of estimated parameters. AIC criterion was used because it compares both - how well the model fits the data and how complex the model is. Therefore, AIC balances the fitting and complexity of the model and therefore is a balanced and much better criterion than others

KS Test: The K-S statistic reported is alpha, where alpha is the reject level for the hypothesis that the fitted curve is the same as the empirical curve**.** K-S should be a high value (Max =1.0) when the fit is good and a low value (Min = 0.0) when the fit is not good.

AIC Values

Beta	1529.87
Genextreme	1520.84
dweibull	1580.27
loggamma	**1513.45**
dgamma	1568.45

KS Test

normal	.99863
loggamma	1.0
gamma	.99864
expon	.99875
weibull	.83239

Comparing the different AIC values and KS score we see log gamma distribution (AIC-1513, KS - 1.0) is lower than all the other distributions hence we can conclude our distribution follows log gamma distribution.

5. TIME SERIES ANALYSIS AND DECOMPOSITION

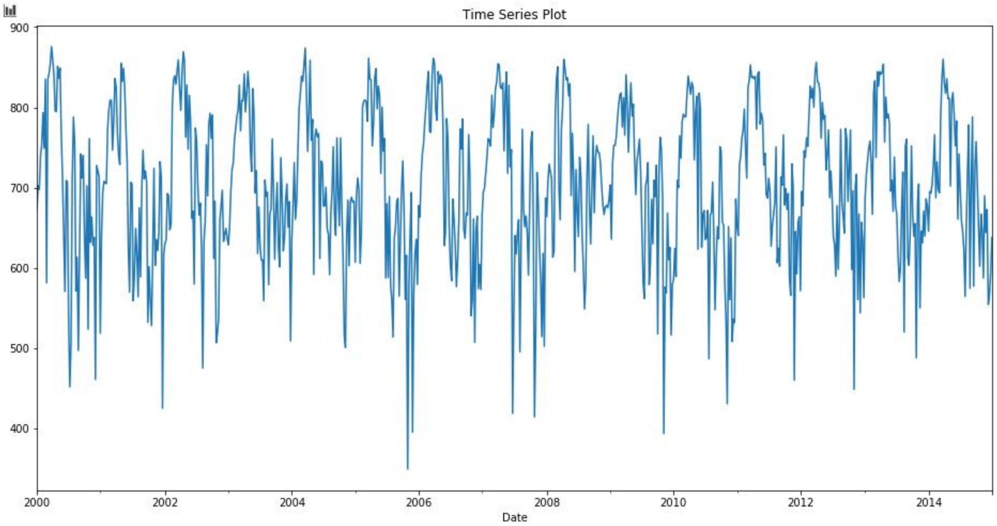


Figure 5.1 : Time series plot

The dataset was processed and only the first day of the week is taken in account. Analysing from the time series plot, it can be inferred that there is seasonality present. For the decomposition of this time series, additive decomposition is relevant as the magnitude of the seasonal variations do not vary with the level of the time series. Mathematically, the decomposition can be represented as :

$$\mathbf{Y_t = S_t + T_t + R_t}$$

Where Y_t is the observed data, S_t is the seasonal component, T_t is the trend component and R_t is the residual component. Observing the time series decomposition in Fig, the trend does not show any pattern. There is a clear pattern in the seasonal component, but the residual component is random. Thus, it can be concluded that there is no trend in the data, and there exists seasonality over an annual period.

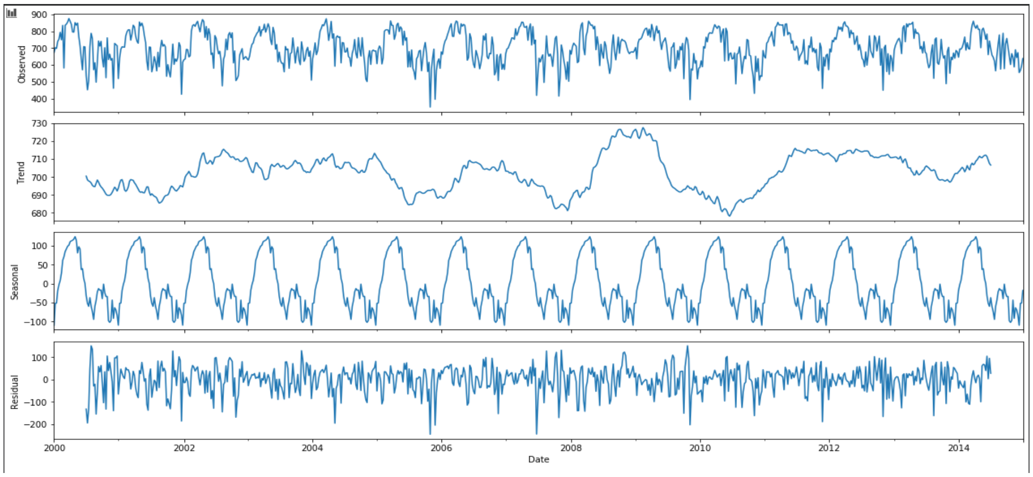


Figure 5.2: Time series decomposition

6. FORECASTING

6.1 Forecasting Models

1. Auto Regressive (AR) Model

An autoregressive model forecasts the value of a variable based upon its past values. Mathematically, a model with order p could be expressed as :

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Here, an **order p** refers to the number of past values of variables used to forecast. The ε_t handles the randomness in data (also known as, white noise). In our forecasting, in order to ensure simplicity, we may restrict the order p between 0 and 2.

2. Moving Average (MA) Model

In contrast to the AR model, a moving average model forecasts the value of a variable based upon the past forecast errors. Mathematically, a model with order q could be expressed as :

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Here, an **order q** refers to the number of past forecast errors accounted for the current forecast. In our forecasting, in order to ensure simplicity, we may restrict the order q between 0 and 2.

3. Auto Regressive Moving Average (ARMA) Model

An ARMA model basically combines the AR and MA model. It provides a greater accuracy than individual AR or MA models. Mathematically, it could be expressed as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

4. Auto Regressive Integrated Moving Average (ARIMA) Model

The models discussed above, require the data to be strictly stationary, that is, the mean, variance, and covariance is constant over a period of time. If the data is not stationary, it could be transformed using differencing. An ARIMA model, in addition to the features of the ARMA model, also accounts for the differencing operation. It is, thus, characterized by three parameters: **p, d, q**. The parameter 'd' specifies the number of differencing required to make the data stationary. The parameters **p and q** have the same meaning as mentioned in the AR and MA models' section above.

5. Seasonal Auto Regressive Integrated Moving Average (SARIMA) Model

If the data has a seasonal component, ARIMA would not be able to perform well. So, we use an extension of ARIMA that helps in modelling the seasonal component. The model adds some new hyperparameters for the autoregression, differencing, moving average for the seasonal component, along with the period of seasonality.

$$\textit{Seasonal ARIMA} = \textit{ARIMA}(p,d,q) * (P,D,Q)_S$$

6.2 Analysis

Augmented Dickey Fuller (ADF) test

The Augmented Dickey-Fuller Test was performed to check for trends in the time series. A time series whose properties are not dependent on the time at which the series is observed is called a stationary time series. Thus, time series with trends or seasonality are not stationary.

The Augmented Dickey–Fuller (ADF) statistic is used in the test, which is a negative number. The more negative the test statistic, the stronger is the rejection of the null hypothesis that there is a unit root at some level of confidence. More specifically, the ADF test helps to determine whether there is a trend in the time series.

Hypothesis of the Dickey-Fuller Test:

1. Null Hypothesis (H0): If failed to be rejected, it suggests that the time series has a unit root, meaning it is nonstationary. It has a time dependent structure.
2. Alternate Hypothesis (H1): If the null hypothesis (H0) is rejected, it implies that the time series does not have a unit root, which means it is stationary. It does not have a time-dependent structure.

TEST RESULTS

Running the Augmented Dickey-Fuller test on our dataset prints the test statistic value of -13.55, which is less than the critical value at 1% (-3.43). Thus, there is enough evidence to reject the null hypothesis at a significance level of less than 1%. Therefore, through the ADF we can conclude that there is **no trend** in our dataset.

AutoCorrelation (ACF) and Partial AutoCorrelation Function (PACF) plots

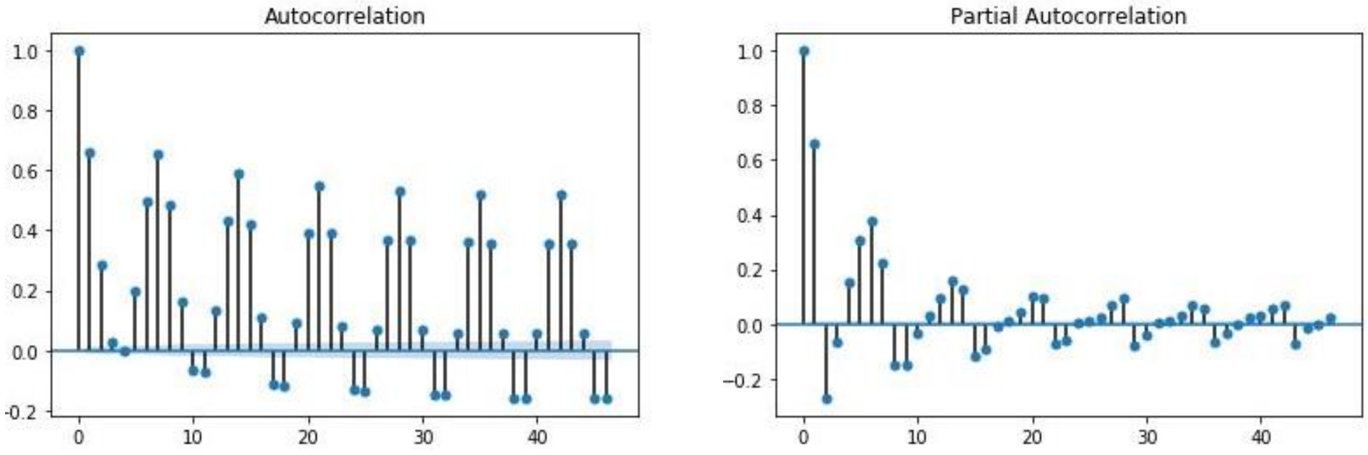


Figure 6.1: ACF and PACF plots

The ACF and PACF plots help in gauging the stationarity in the dataset as a whole taking into account the trend, seasonality, cyclicity, residuality. From the oscillating nature of above plots, it is evident that the time series has seasonality in it. This result is consistent with the results of Section 5 (Time series analysis and decomposition) of the report, so we can conclude that the time series is **non-stationary**.

6.3 Model Parameter Search

The AR, MA, ARMA models cannot be directly applied to the data because it is non stationary, and these models do not perform well if there is seasonality in the data. The parameters for non-seasonal ARIMA: p,d,q and for seasonal part : P, D, Q were limited to take values between 0 and 2. This was done to minimize the complexity of the model. To find the optimum values of the parameters, the grid search method was adopted. An AIC (Akaike information criterion) value for each model was calculated, the lower is the AIC value, the better is the model.AIC criterion was used because it compares both - how well the model fits the data and how complex the model is. Therefore, AIC balances the fitting and complexity of the model and therefore is a balanced and much better criterion than others. Further, the dataset was split into training (from 2000 to 2012) and testing (from 2013 to 2014)

6.3.1 ARIMA Forecasting

The results for the best parameter combinations obtained using grid search for the ARIMA model(over daily data for 2000-2012) are as follows-

p	d	q	AIC
1	1	0	66420.62
1	1	1	65536.99
2	1	0	66206.43
2	1	1	65526.97

Table 6.1: Performance comparison of various ARIMA models

The much higher AIC values for ARIMA model (as compared to SARIMA) is primarily due to the fact that ARIMA is unable to capture the seasonal component in the data accurately. For this purpose SARIMA is better. Nonetheless, as evident from the above table , the parameter combination of (p=2 , d=1 , q=1) had the least AIC value and was thus considered for our ARIMA model

6.3.2 SARIMA Forecasting

From the time series decomposition, we could see that there is a seasonal component present in the data. So, let us check how the SARIMA model(over weekly data for 2000-2012 for ease of training model) performs on the data. The dataset was processed and only the first day of every week was taken into account. So, the appropriate value of the period of seasonality came out to be 52.

p	d	q	P	D	Q	S	AIC
0	0	0	1	1	1	52	7650.94
1	0	1	1	1	1	52	7614.42
1	1	1	0	1	1	52	7615.24
1	1	1	1	1	1	52	7612.55

Table 6.2: Performance comparison of various SARIMA models

Thus, we find the model ARIMA(1,1,1)(1,1,1,52) yields us the minimum AIC value. Upon validation, this model gave us a mean absolute percentage error (MAPE) of **7.098%** for the location in Andhra Pradesh.

7. RESULTS

7.1 ARIMA Forecasting

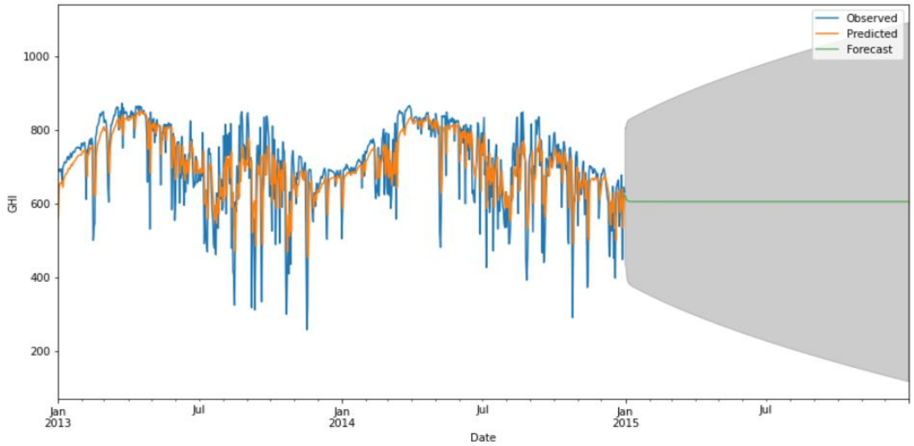


Fig 7.1.1: Andhra Pradesh with MAPE : 9.67%

As evident from above ARIMA is unable to capture seasonality while forecasting future values and gives a straight line while predicting future values.

7.2 SARIMA Forecasting

The SARIMA model, thus trained, was most suitable for solar energy forecasting. The model was run on different locations provided and their respective accuracies were obtained. The solar park in Rajasthan gave the best accuracy, while the location in Madhya Pradesh had a comparatively low accuracy. Further, we obtained a forecast for a complete year of 2015 at each location, which can be seen by the green line in each of the plots. The grey shaded area denotes a 95% confidence interval for the forecast.

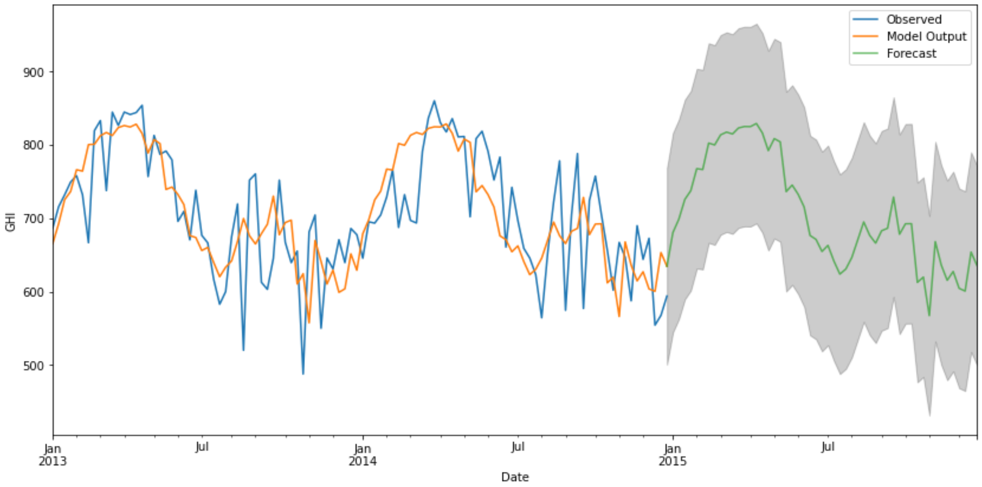


Figure 7.2.1: Andhra Pradesh

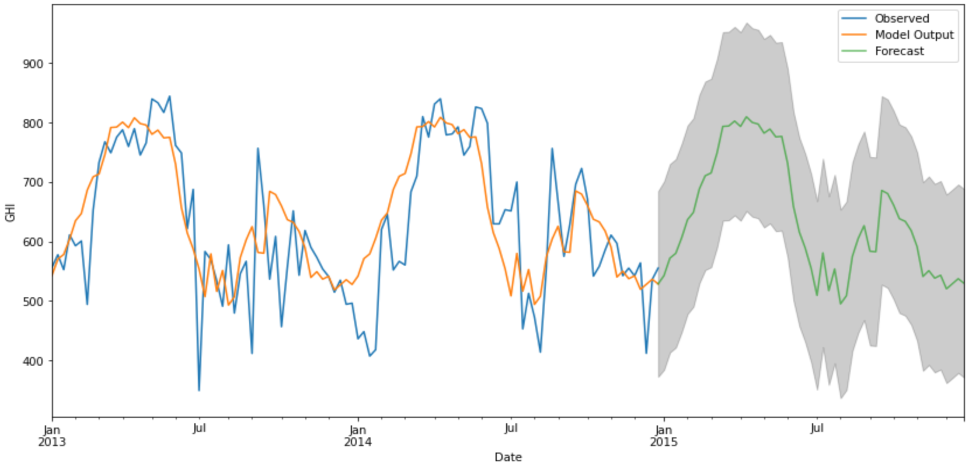


Figure 7.2.2: Madhya Pradesh

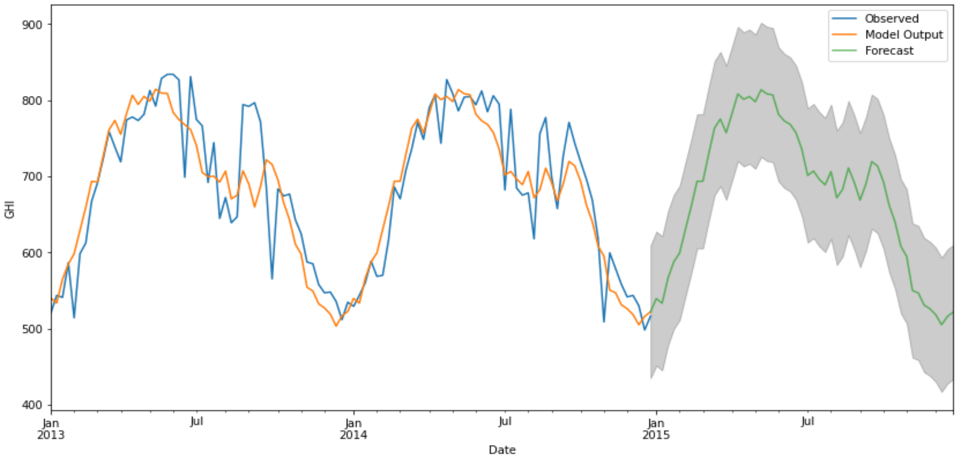


Figure 7.2.3: Rajasthan

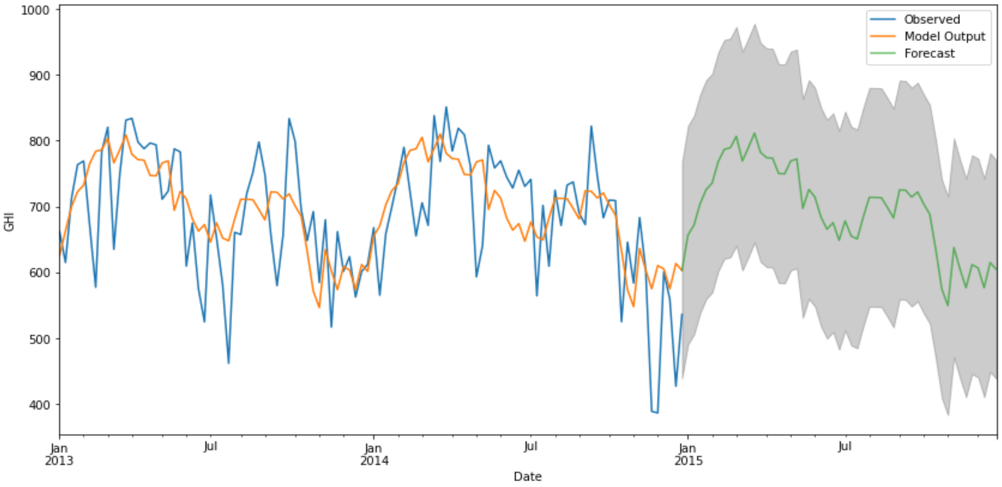


Figure 7.2.4: Tamil Nadu

State	MAPE	MSE	RMSE
Andhra Pradesh	7.098%	3718.458	60.97
Madhya Pradesh	10.392%	6226.84	78.91
Rajasthan	4.768%	1779.18	42.18
Tamil Nadu	9.808%	6065.07	77.87

Table 7.1: Forecasting errors for states

8. CONCLUSIONS

- GHI is the most important factor for forecasting solar energy
- Log gamma distribution best describes the GHI data
- Time series decomposition shows presence of seasonal components in GHI data
- ACF and PACF plots prove that series is non stationary (because of seasonality in data)
- Parameters for ARIMA and SARIMA models were obtained using grid search algorithm
- SARIMA model is best suited for forecasting the GHI values
- The solar park in Rajasthan gave the best accuracy, while the MP had a comparatively low accuracy.

9. REFERENCES

1. Atique, Sharif & Noureen, Subrina & Roy, Vishwajit & Bayne, Stephen. (2019). Forecasting of total daily solar energy generation using ARIMA: A case study. 10.1109/CCWC.2019.8666481.

2. Fuller, W. A. (1976). Introduction to Statistical Time Series. New York: John Wiley and Sons. ISBN 0-471-28715-6

3. Glen, S. (2016). ADF – Augmented Dickey Fuller Test [statisticshowto]

4. Kumar, V. (2017). Test of Kolmogorov-Smirov for the Log-Gamma Distribution. [rdocumentation]

5. Perktold, J., Seabold S., Taylor J. (2009). Official Documentation of statsmodels Python library. [statsmodels]

6. Prabhkaran, S. (2018). ARIMA model- Complete Guide to Time Series Forecasting for Python. [machinelearningplus]

7. Salvi, J. (2019). Significance of ACF and PACF plots in Time Series Analysis. [towardsdatascience]

8. Vasishtha, S. (2012). Differentiate between the DHI, DNI and GHI. [firstgreen]

All the code for this Assignment can be found at – Code Repository Friedman Group