Applied Statistical Methods Assignment 2

Statistical Analysis and Forecasting of Solar Energy (Inter-States) : Problem 3

Fisher Group

Aryan Devrani : 2019A8PS0408P
Mayank Jain : 2019A7PS0141P
Abhiraj E : 2019A7PS0050P
Sarthak Choudhary : 2019A7PS0112P
Racchit Jain : 2019A7PS0145P
Suchismita Tripathy : 2019A7PS0554P
Mihir Kumar : 2019A3PS0218P
Arshdeep Taneja : 2019A4PS0469P

7th December 2021

Contents

1 Introduction

Renewable energy is energy obtained from naturally replenishing sources. They have much fewer adverse effects on the environment, as they release fewer chemicals and greenhouse emissions and do not need to be constantly mined/depleted. They are virtually inexhaustible but give limited energy per unit of time.

Solar energy is the most abundantly available and clean form of renewable energy, and has been harnessed since the use of solar ovens. Solar power is converted into electrical energy using photovoltaic devices or solar cells but harnessing solar power depends on a number of factors which affect the efficiency of the process, since the amount of sunlight that reaches the earth's surface is relatively small, and also depends on a variety of conditions including location, temperature, wind etc. Due to this, and the increasing importance of using renewable energy, the analysis and forecasting of solar energy is vital to perfecting processes for harnessing it,

In this report, we work on inter-state solar energy data, and further focus on GHI (among a list of other factors including DHI, DNI, wind speed, dew point etc.).

2 Methodology

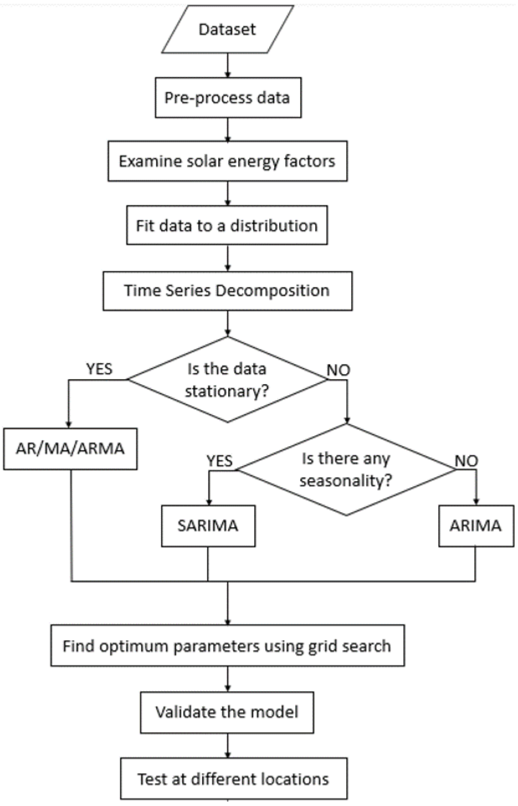

Dataset
Pre-process data
Examine solar energy factors
Fit data to a distribution
Time Series Decomposition
YES
Is the data stationary?
NO
AR/MA/ARMA
YES
Is there any seasonality?
NO
SARIMA
ARIMA
Find optimum parameters using grid search
Validate the model
Test at different locations

Figure 1: Methodology

3 Descriptive Analysis

From the obtained data, it is evident that there are a number of variables that can influence the amount of solar radiation that reaches the earth's surface. Let us establish these factors so that we can better comprehend the situation and see how these variables affect each other (correlation) as well the solar radiation value The following are some key terminology to be aware of:

1. **DNI (Direct Normal Irradiance) -** 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.

2. **DHI (Diffuse Horizontal Irradiance) -** 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.

3. **Solar Zenith Angle**(θ) **-** It is defined as the angle between the vertical and the sun's ray.

4. **Global Horizontal Irradiance(GHI)** The total amount of solar radiation received per unit area on the horizontal surface of Earth. Mathematically, it can be expressed as:

$$GHI = DHI + DNI.\cos(\theta) \tag{1}$$

There are several other variables included in the dataset such as 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.

4 Descriptive Statistics

It is important to understand how all the variables in the dataset are correlated with each other and the strength of their correlation. We first took the data of Andhra Pradesh from the year 2000 to 2014 and created a seaborn heatmap for it. The following correlation matrix was obtained:

	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI
DHI	1	0.22	0.62	0.79	0.76	0.81
DNI	0.22	1	0.85	0.63	0.69	0.63
GHI	0.62	0.85	1	0.86	0.86	0.92
Clearsky DHI	0.79	0.63	0.86	1	0.87	0.91
Clearsky DNI	0.76	0.69	0.86	0.87	1	0.93
Clearsky GHI	0.81	0.63	0.92	0.91	0.93	1

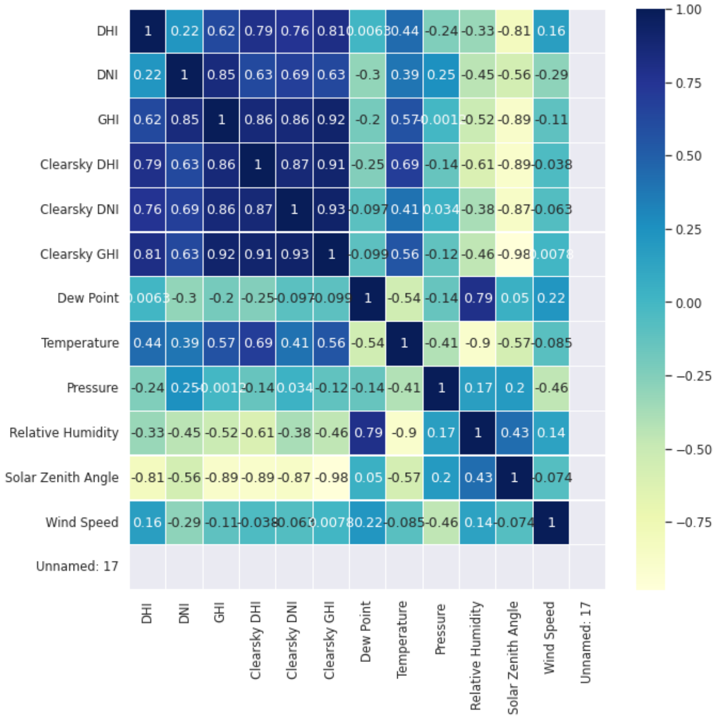


Figure 2: Correlation Heatmap

1. From the heatmap, it is clearly evident that GHI, DHI, DNI, Clearsky GHI, Clearsky DHI and Clearsky DNI share a highly positive correlation with each other.
2. These correlation values confirm the Solar Irradiance equation which relates GHI to DHI, DNI and the Solar Zenith Angle. The same conclusion was drawn from heatmaps for the other states.

We now select GHI for the purpose of Data Analysis and Estimation of Distribution Fit. Following are the correlation values of of GHI with rest of the variables:

Variable	Correlation with GHI
Solar Zenith Angle(θ)	-0.89
Relative Humidity	-0.52
Temperature	0.57
Pressure	0.001
Dew Point	-0.2
Wind Speed	-0.11

1. Solar Zenith Angle and Relative Humidity are negatively correlated with GHI and Temperature is positively correlated. These correlation values are significant (>0.5) and ideally should not be ignored.
2. Rest of the variables such as Pressure, Dew Point and Windspeed show very low correlation values and will have no significant effect on GHI.

5 GHI Data Analysis

We now focus purely on the GHI Data and check for Normality in its distribution. Two different methods were used for this:

1. **P-P and Q-Q Plots:** The non-linearity or deviation of the blue lines (cdf-cdf and quantile-quantile) from the straight line shows that the distribution is not Normal.

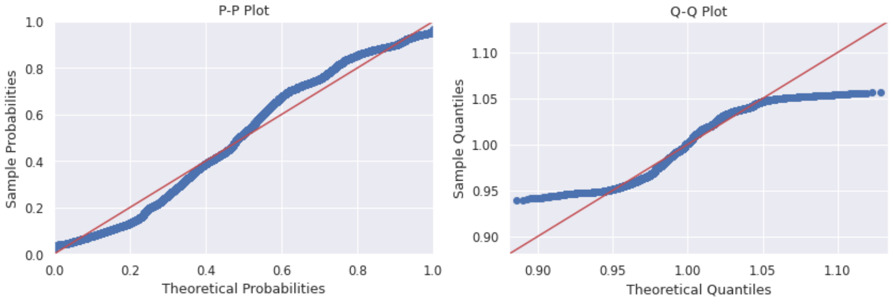


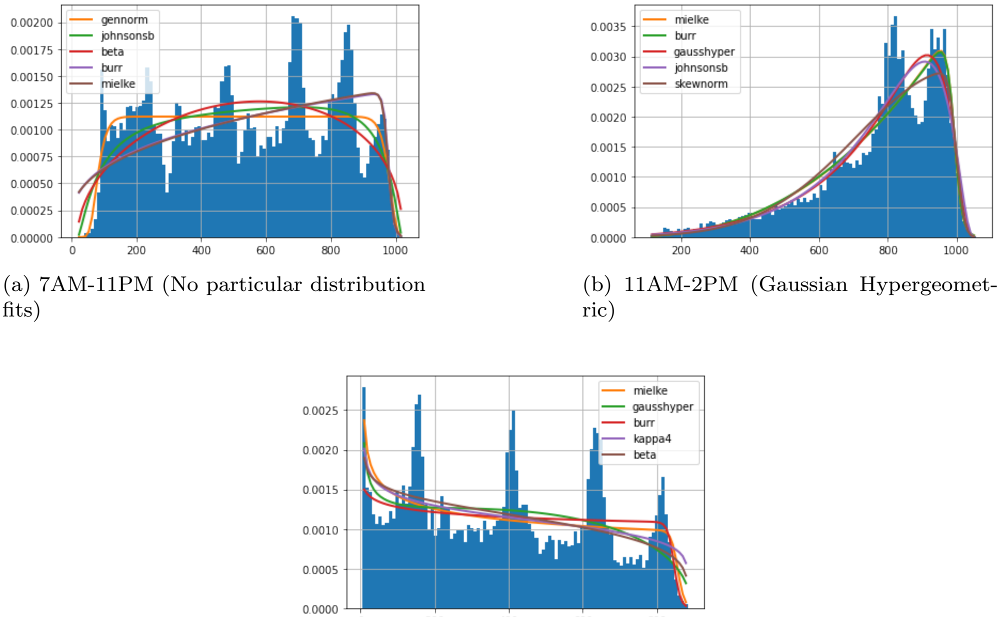
Figure 3: p-p and q-q Plots for the aggregated data

2. **D'Agostino K^2 Test :**

$$H_0\text{: Distribution is normal}$$
$$H_a\text{: Distribution is not normal}$$

We conducted a 95% significance D'Agostino K^2 Test for all the four states. A p-value lesser than α=0.05 leads to rejecting the null hypothesis. Hence, it was confirmed that the GHI Distribution is not normal.

The next step is to find the best distribution fit for this data. To do this, we made use of the distfit() Python package, which tests for 89 univariate distributions based on the Akaike Information Criterion (AIC). A lower AIC value is considered to be a better fit as it also means a lower SSE. We split the hourly data on a given Day into 3 time brackets: 7am-11am, 11-am-2pm and 2pm-5pm, but none of the time brackets seemed to fit any particular distribution, except 11am-2pm which resembled a Gaussian Hypergeometric distribution.



(a) 7AM-11PM (No particular distribution fits)

(b) 11AM-2PM (Gaussian Hypergeometric)

(c) 2PM-5PM (No particular distribution fits)

Figure 4: Distributions Fitting for various time brackets

Nonetheless, we then aggregated the whole 7am-5pm dataset and obtained the following results:

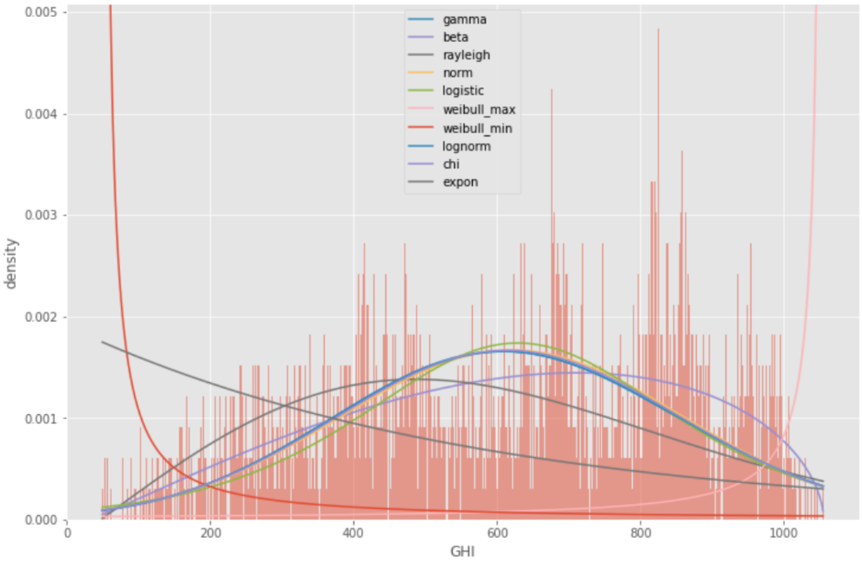
	SSE	AIC
Beta	8.34e^{-6}	1404.62
Logistic	1.38e^{-5}	1416.39
Lognorm	1.42e^{-5}	1434.64
weibull-max	1.57e^{-5}	1433.22
chi	1.74e^{-5}	1434.73

Beta distribution produced the lowest AIC value and lowest SSE, hence it is the best estimate for GHI distribution. The following parameters were obtained for it: α= 2.0866 β=1.3782. where the equations for beta-distribution are

$$f(x) = \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} \qquad a \leq x \leq b; p, q > 0, \tag{2}$$

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \tag{3}$$

Figure 2: Distribution Fits for Andhra Pradesh GHI Data



6 Time Series Analysis and Decomposition

The aim of this section is to analyze the time-series of Global Horizontal Irradiance (GHI) data. However, instead of analyzing the hourly GHI Time-Series, we have adjusted the data to take into account the aggregate of GHI values over every week. We shall first start by ensuring stationarity of the data using a combination of two tests, namely Augmented Dickey-Fuller (ADF) test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. Once ensured, we shall then proceed to analyzing the time-series itself by performing Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) on the daily-adjusted before decomposing it into its Trend, Seasonality and Residual Components.
For simplicity, the procedure was performed for a single state - Madhya Pradesh.

6.1 Box Jenkins Method

Box Jenkins method starts with the assumption that the time series can be modelled using ARMA if stationary and ARIMA if non-stationary. It consists of 3 main steps:

1. **Identification:** In this step we assess whether the time series data is stationary or not. If the data is stationary then we move forward to configure the AR metric p and MA metric Q. If not we first make the time data stationary by differencing it and then continue with finding the metrics, once the unit root test of confirming stationary data is passed, otherwise keep differencing the data till stationarity is achieved. PACF (Partial Autocorrelation Function) and ACF (Autocorrelation Function) is used to find the p and q metrics.

2. **Estimation:** In this step we use the data to train the parameters of the model. It involves using numerical methods to minimize the losses and error terms.

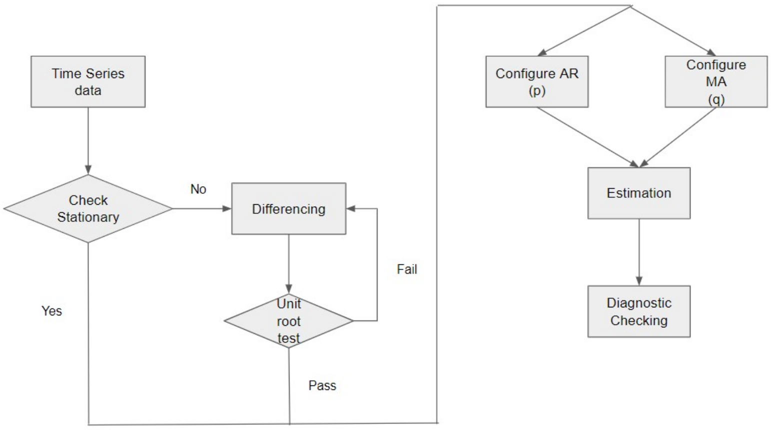


Figure 5: Box-Jenkins Method

3. **Diagnostic Checking:** This step is used to determine whether the model is a good fit or not. The two main areas investigated over here are overfitting and residual errors. Overfitting needs to be taken care of as it negatively impacts the ability of the model to generalise. Therefore overfitting on training data needs to be avoided as it captures the noise in the set and reduces the robustness of the model.Residual errors help us identify if there is more information left in the data which can be exploited. This can be done by checking the serial correlation in ACF and PACF graphs of residual errors.

We have used the above mentioned method to fit the data on various models and do time series forecasting.

6.2 Time-Series Aggregation

The Global Horizontal Irradiance (GHI) data with us presents hourly variations. However, this detail can be ignored and the data can be simplified by reducing it down to daily variations. Hence, before we proceed to the Analysis and Decomposition procedure, we sum hourly GHI data points over the span of a day. This not only helps in understanding the daily and monthly variations easily, but also reduces the size of the data, thereby reducing computations.

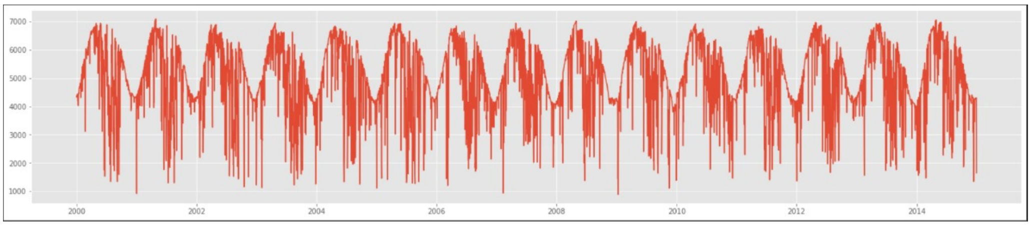


Figure 6: Adjusted Time-Series Plot

6.3 Stationarity Tests

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most forecasting and analysis methods assume the time-series to be atleast approximately stationary, hence a stationarity test becomes a must before proceeding forward.

We shall perform two tests in succession, however the order of performing the tests does not matter. It is generally considered a good practice to apply both ADF and KPSS tests for this purpose.

6.3.1 Augmented Dickey-Fuller (ADF) Test

The ADF procedure tests the following set of hypotheses:

$$H_o : \text{Time-series has unit root; } H_a : \text{Time-series has no unit root.}$$

The value of the test-statistic found from the procedure is -13.906 and the critical value at 99% Confidence is -3.4304. Using the Critical Value Approach, since the test-statistic is more negative than the critical-value, we can reject the null hypothesis. Hence, we can conclude that the time-series is stationary.

6.3.2 Kwiatkowski–Phillips–Schmidt–Shin (KPSS) Test

To validate the results obtained from the ADF Test, we proceed ahead with the KPSS Test. Here, the procedure tests the following set of hypotheses:

$$H_o : \text{Time-series is stationary; } H_a : \text{Time-series has unit root.}$$

The p-value so obtained comes out to be 0.1. For 99% Confidence, alpha is 0.01. Since p-value is greater than alpha, we cannot reject the null hypothesis. Hence, as pointed out by the ADF Test as well, we may safely conclude that the time-series is stationary, as suggested by ADF Test as well.

6.4 Analysis and Decomposition

We shall now analyze the time-series by performing Partial Auto-correlation Function (PACF) and Auto-correlation Function (ACF) on the time-series and plotting the respective graphs.

6.4.1 Auto-Correlation Function (ACF)

Correlation is the summary of the strength of relationship between two variables. Auto-Correlation is a measure of how strongly correlated the values of the time-series are to their predecessor data points. ACF values well above the Confidence bound suggest high auto-correlation amongst the Time-Series data points. It is important to note that ACF does take into account intermediate lags too.

6.4.2 Partial Auto-Correlation Function (PACF)

Partial Autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationship of intervening observations removed. PACF removes the indirect correlation between the current step and the lag. This gives us a good metric for the value of p for our Autoregression (AR) model.

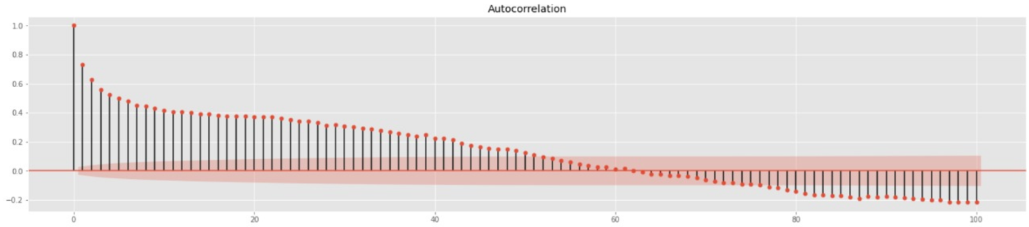


Figure 7: Auto-Correlation Function

Similar to ACF, the PACF procedure removes the effect of the intermediate lags, and only calculates the auto-correlation values between the two separated points and nothing else in between. PACF gives us a simple metric for a p-value for Autoregression Models or AR.

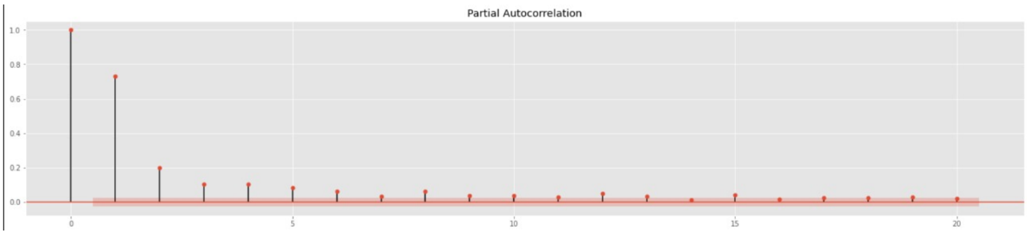


Figure 8: Partial Auto-Correlation Function

6.4.3 Time-Series Decomposition

Additive models are seen to be more useful when the seasonal variation of the time-series is relatively constant over time, while multiplicative models are preferred when this seasonal variation itself changes over time. Here, we have decomposed the time-series into three systematic or recurring components: Trend, Seasonality, and Residual, using the Additive Decomposition Model.

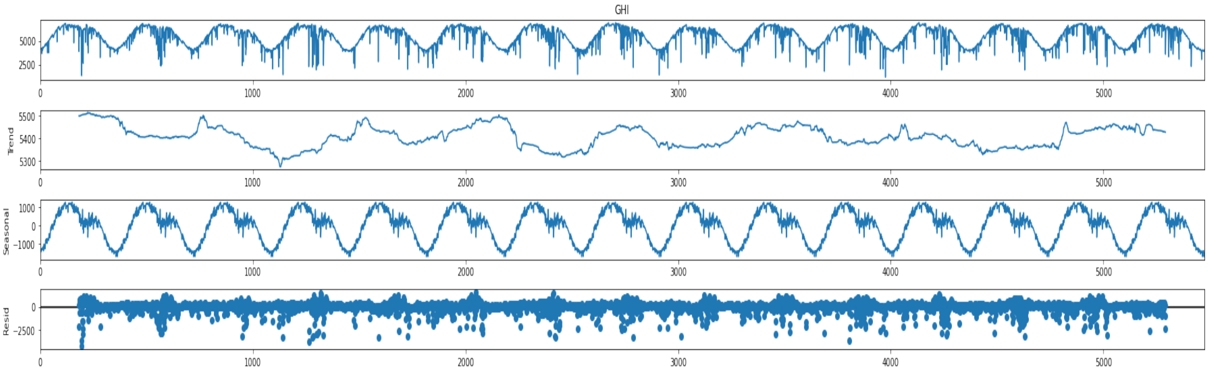


Figure 9: Time-Series Decomposition Plot

It can be observed how the time-series has no prominent upward or downward trend. Furthermore, the seasonality shows recurrence or repetition of behaviour annually since a unit pattern can be seen repeating over nearly 365 days which also makes logical sense as sunlight or GHI time-series should repeat itself on an annual basis.

7 Forecasting

7.1 AR

AutoRegressive (AR) model is a multiple regression model where we forecast the variable of interest using a linear combination of past values of the same variable in the time series. An AR model of order p, referred to as AR (p), can be written as:

$$\theta_p(B)x_t = (1 - \alpha_1 B - \alpha_2 B^2 - ... - \alpha_p B^p)x_t = w_t$$

where B is the backshift operator and w_t is white noise. Although the PACF plot recommends that the value of p should be around 17, on running a grid search for the best fit to AR, we got a value of p equal to 12. We ran an AR model with p as 12 over the entire time series and got an MAPE of 6.83%. On running the above stated model, we get the following ACF plot for the residuals.

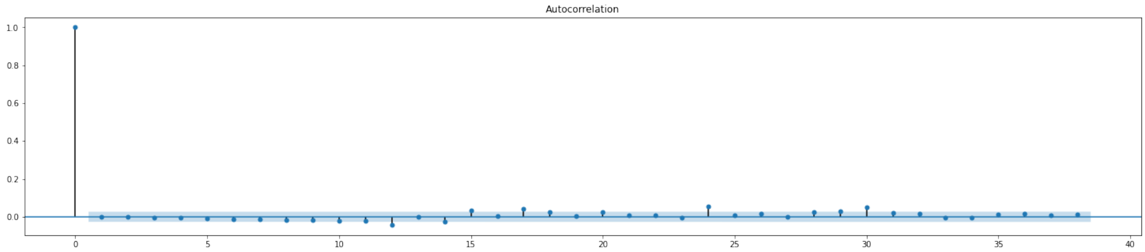


Figure 10: ACF Plot - AR (8)

Note that an ideal ACF plot for residuals should have all the points within the blue region, indicating that the correlations are statistically zero (the residuals can be called white noise). Clearly there are a few significant points in our plot, showing that AR might not be the best model for this time series. But for the sake of completeness, we have gone ahead with an AR analysis anyway.

Although it is not the ideal value, the AR model was run with a p value of 8 to avoid excessively long runtimes.. It gave us an MAPE of 9.45% on the entire test set. Similarly to the process described above, weekly and monthly rolling forecasts were done, and we got MAPEs of 11.58% and 13.23% respectively. As expected, the MAPE values increased as we increased the span of each forecast.

7.2 MA

As opposed to AR which uses past values of forecast variable in time series for regression analysis, the Moving Average(MA) model uses past forecast errors in a regression-like model. An MA(q) process can be expressed as follows

$$x_t = (1 - \beta_1 B - \beta_2 B^2 - ... - \beta_p B^p)w_t = \phi_q(B)w_t$$

where ϕ_q is a polynomial of order q and B is the backshift operator. Here, each value of x_t can be thought of as a weighted moving average of the past q forecast errors. The ACF plotted earlier suggests that the value of q should be around 60 for MA. However, this is computationally expensive and leads to very large training times. So instead, we use the value of q as 30. Running the MA model over the entire time series with this q yields a Mean Absolute Percentage Error (MAPE) value of 7.28%. Also, the following ACF plot for residuals is obtained.

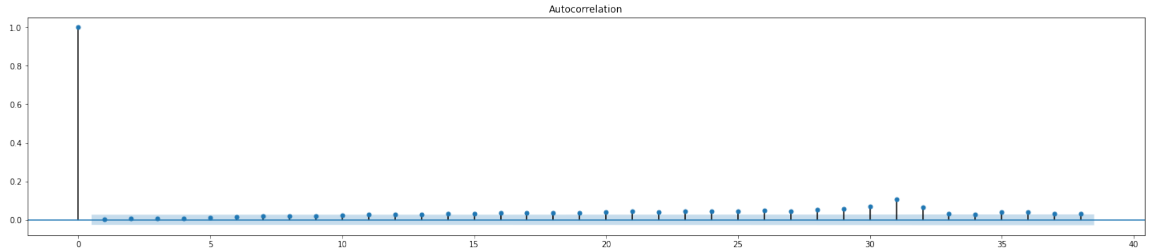


Figure 11: Residuals ACF Plot - MA (30)

As expected, we can clearly see in the above plot that there are significant autocorrelations (outside blue region) for lags after 30. Now, although the value of q as 30 provides a good fitting model, it is almost impossible for us to train the model with such a high value of q. So instead, we use MA (6) for rolling forecast. For daily forecasting, MA (6) fits decently and gives a MAPE value of 8.79%, marginally better than AR. However, for weekly and monthly predictions, MAPE values are significantly larger. MAPE for weekly forecast is 14.25% and for monthly forecast is 16.97%. Clearly, MA model doesn't perform very well at forecasting for this time series.

7.3 ARMA

AutoRegressive Moving Average or ARMA, as the name suggests, is a combination of the previous two models, AR and MA. It is used to describe a stationary time series in terms of two component polynomials, one of which corresponds to autoregression and the other to moving averages. The model has two parameters, p and q, for Auto-regression and Moving Averages respectively. ARMA (p,q) model is as follows

$$\theta_p(B)x_t = \phi_q(B)w_t$$

where θ_p is a polynomial of order p and ϕ_q is a polynomial of order q. Due to computational constraints, we did a grid search for the hyper-parameters p and q each in the range of 1 to 20 and got the best values as 12 and 10 respectively. The model turns out to be a decent fit for the time series, as the p-values for most of the 22 coefficients are less than 0.05; this argument is further strengthened by the ACF plot of the residuals (please see the plot in the appendix for reference). The MAPE for the same was 6.78%. As discussed before, owing to heavy computational costs, the grid search for the best parameters for the rolling forecast was done in a much smaller space. This resulted in the values of p and q as 3 and 1. The daily rolling forecast gave an MAPE of 7.06% while the weekly and monthly forecasts gave MAPEs of 8.63% and 10.23% respectively.

7.4 ARIMA

AutoRegressive Integrated Moving Average or ARIMA is a generalization of ARMA. Unlike ARMA, which is only applicable in stationary models, ARIMA can be applied on non-stationary models as well. It does this by using differencing to convert the non stationary model into a stationary one.

Mathematically, d order differencing is of the form

$$(1-B)^d x_t$$

where B is the backshift operator. A series is integrated of order d if differencing by order d results in white noise. ARIMA(p, d, q) performs ARMA(p, q) on data that has been integrated by order d, which is the meaning of the added I - for integration. This model can be succinctly expressed by the equation

$$\theta_p(B)(1-B)^d x_t = \phi_q(B) w_t$$

where θ_p and ϕ_q are polynomials of order p and q respectively. Though the data is already confirmed to be stationary by our previous tests, we still performed ARIMA for testing purposes. We performed a grid search for parameters with $d \geq 1$ which best fit our model. We found that an order of (12, 1, 10) fit our model best (MAPE 6.76%), and resulted in insignificant correlation in the residuals.

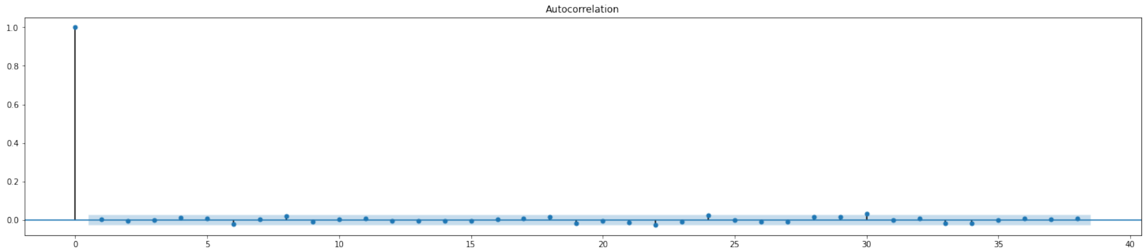


Figure 12: Residuals ACF Plot - ARIMA(12, 1, 10)

However, this model is too computationally intensive to use on a rolling forecast. Instead, we used an order of (3, 1, 1) for our rolling forecast, giving us an MAPE of 6.97%, the best of all the forecasting methods employed so far. This model also performed exceedingly well on weekly and monthly predictions, giving an MAPE of 8.49% and 10% respectively.

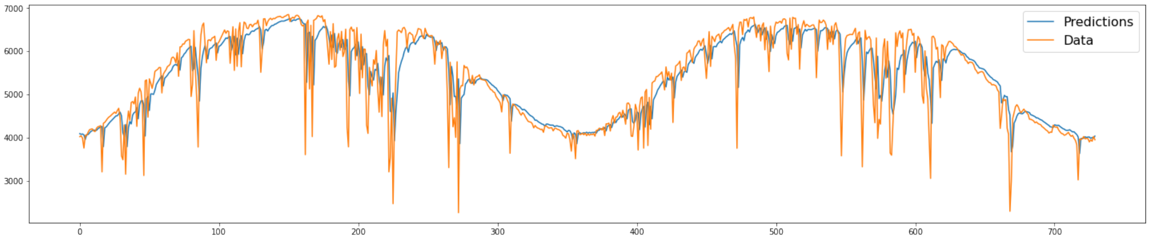


Figure 13: Daily Predictions - ARIMA(12, 1, 10)

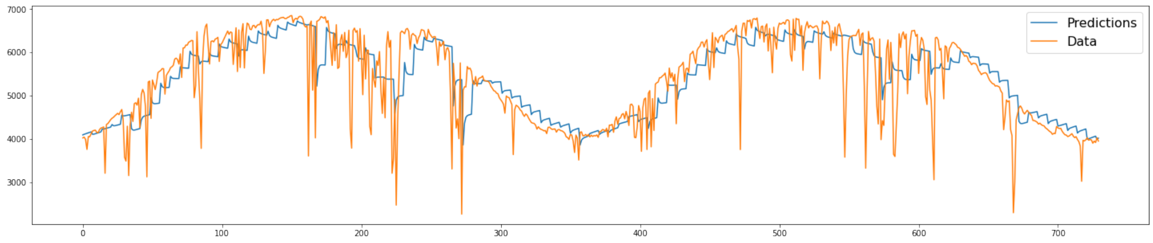


Figure 14: Weekly Predictions - ARIMA(12, 1, 10)

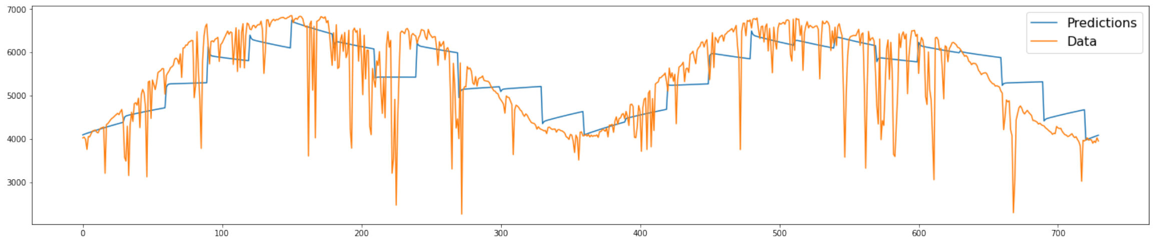


Figure 15: Monthly Predictions - ARIMA(12, 1, 10)

7.5 SARIMA

Seasonal ARIMA model uses differencing at a lag equal to the seasonality to remove additive seasonal effects. It also introduces autoregressive and moving average terms at the same lag, resulting in the order being expressed as (p, d, q)(P, D, Q) s , where s is the seasonality. This can be expressed mathematically as

$$\Theta_P(B^s)\theta_p(B)(1-B^s)^D(1-B)^d x_t = \Phi_Q(B^s)\phi_q(B)w_t$$

where Θ_P, θ_p, Φ_Q, ϕ_q are polynomials of orders P, p, Q and q respectively. The seasonality in SARIMA appears as an exponent in this model and therefore a daily seasonality of 365 is highly computationally expensive. So instead, we aggregate the data into monthly data (by summing over the entire month), thus reducing the seasonality to 12. Through a grid search we obtain the optimal parameters for our SARIMA model. It was found that a non-seasonal order of (3, 0, 3) and seasonal order of (2, 0, 1) worked best for our data. The results are shown in the graph below, which gives us a MAPE of only 2%.

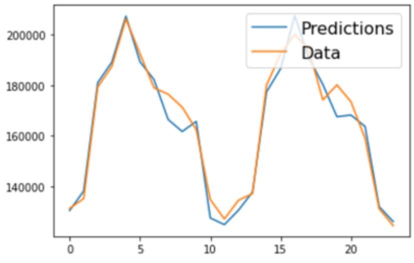


Figure 16: SARIMA Predictions Plot

8 Conclusion

The following conclusions were made from analysing the given data :

1. GHI is the most important factor for analysing the solar energy data.
2. The GHI Data not follow a Normal distribution. The best fitting estimate for it turned out to be the Beta Distribution.
3. Time Series Analysis of the data showed Annual Seasonality, without any upward/downward Trend.
4. ACF and PACF plots proved that series is stationary. Regardless, ARIMA and SARIMA were tested out and their MAPEs were compared with AR/MA/ARMA.
5. Although the data has seasonality in it and SARIMA should perform better, processing limitations didn't allow us to test for ideal parameters in SARIMA. Hence it was concluded that ARIMA producded the best results out of all forecasting models tested.
6. In ARIMA, Daily prediction produced the least MAPE, closely followed by Weekly and lastly, Monthly predictions.

Selection of Model and Analysis for Other States:

The ARIMA Model was also tested for Time Series forecasting on other states. The results (MAPE values) of their fits are tabulated below:

State	MAPE
Madhya Pradesh	10.14%
Tamil Nadu	15.47%
Andhra Pradesh	11.45%
Rajasthan	6.97%

9 Appendix

9.1 Time-Series Analysis and Decomposition

Please find the Time-Series Aggregation, ACF, PACF and Decomposition plots for all the remaining three states here. Plots for Madhya Pradesh have been shown in Section 6 itself.

9.1.1 Rajasthan

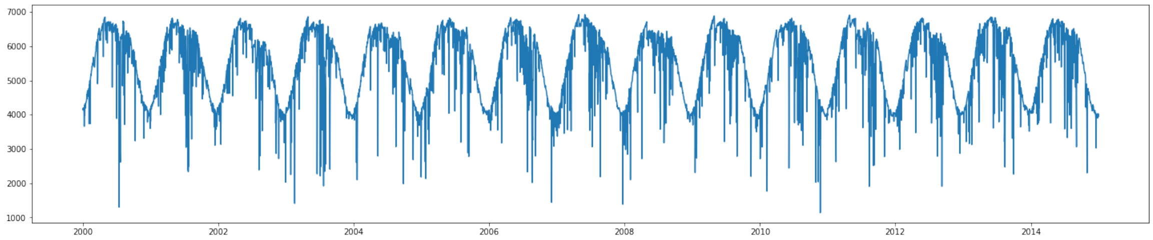


Figure 17: Time Series Aggregation; Rajasthan

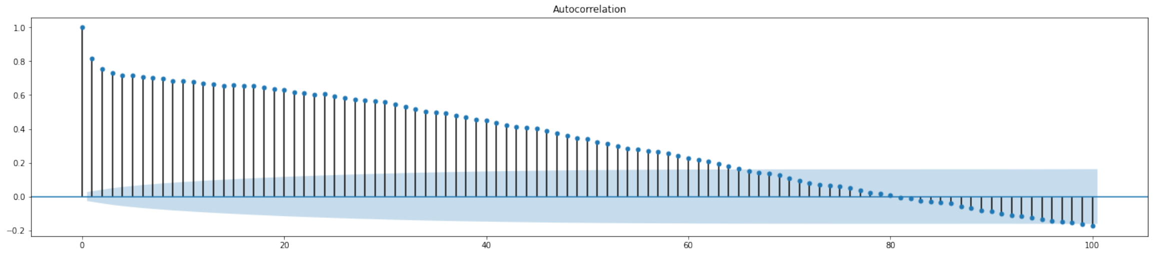


Figure 18: Autocorrelation Function; Rajasthan

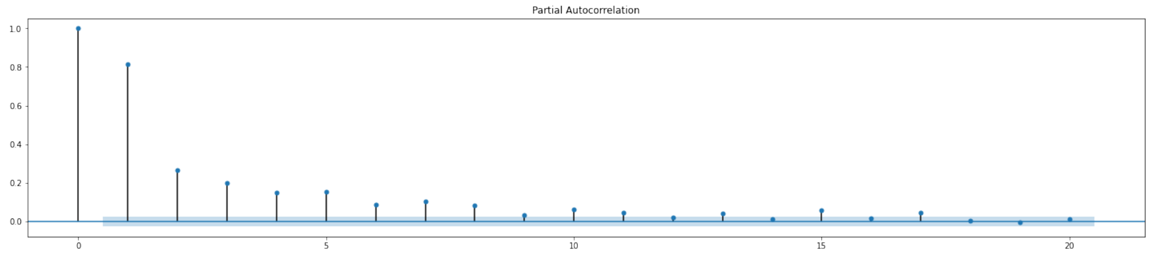


Figure 19: Partial Autocorrelation Function; Rajasthan

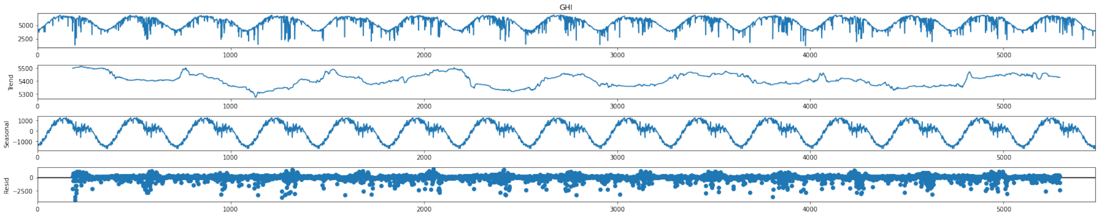


Figure 20: Time Series Decomposition; Rajasthan

9.1.2 Andhra Pradesh

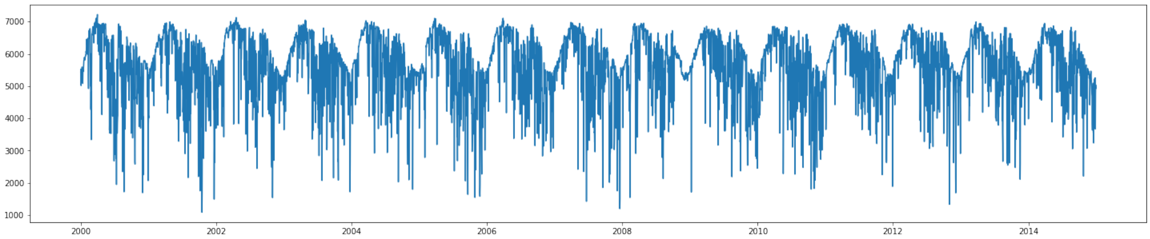


Figure 21: Time Series Aggregation; Andhra Pradesh

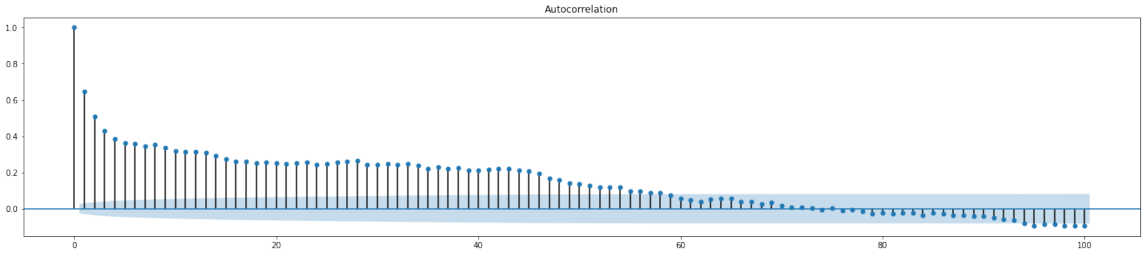


Figure 22: Autocorrelation Function; Andhra Pradesh

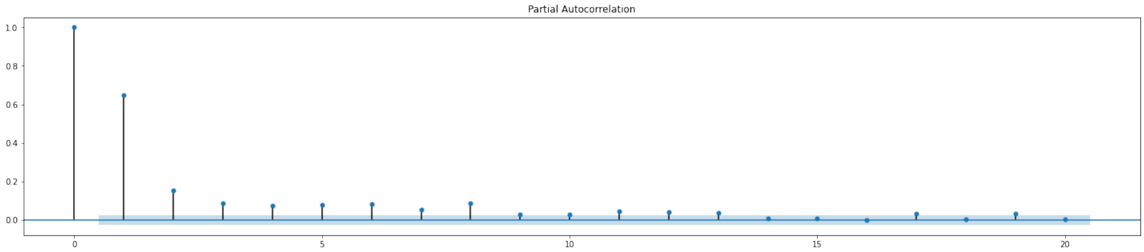


Figure 23: Partial Autocorrelation Function; Andhra Pradesh

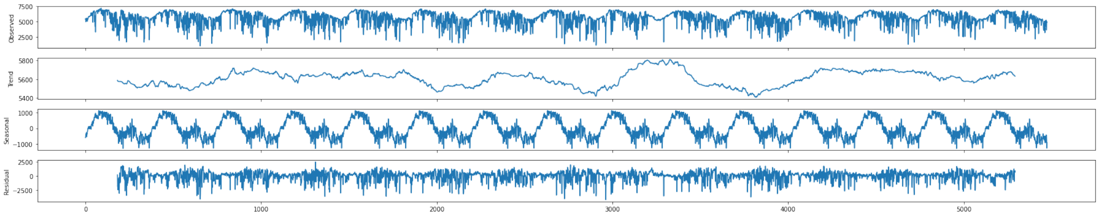


Figure 24: Time Series Decomposition; Andhra Pradesh

9.1.3 Tamil Nadu

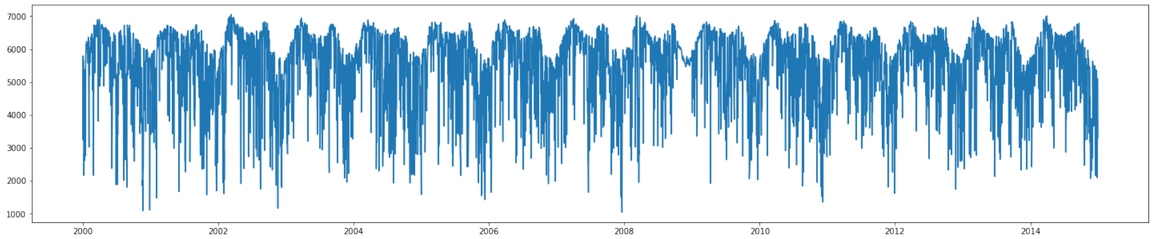


Figure 25: Time Series Aggregation; Tamil Nadu

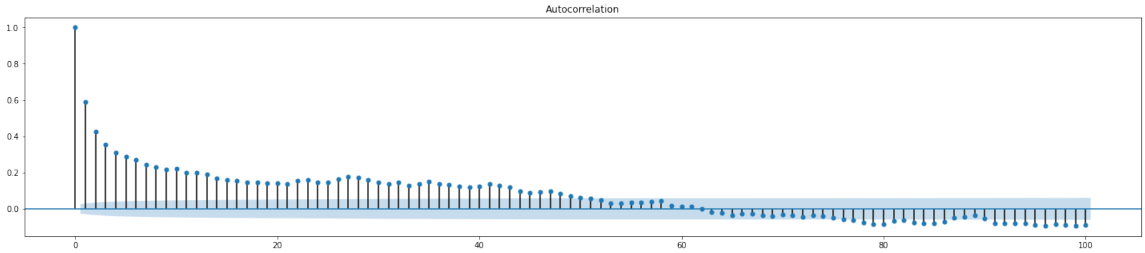


Figure 26: Autocorrelation Function; Tamil Nadu

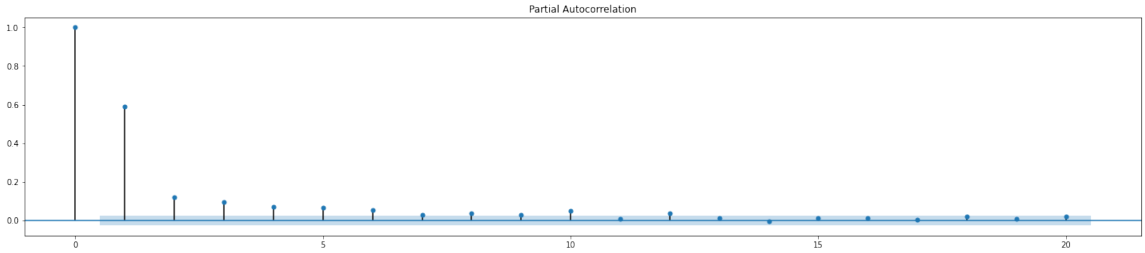


Figure 27: Partial Autocorrelation Function; Tamil Nadu

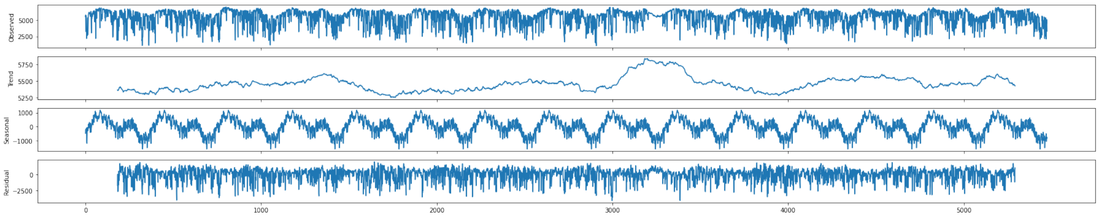


Figure 28: Time Series Decomposition; Tamil Nadu

10 References

[1] D.R. Anderson, D.J. Sweeney, T.A. Williams, J.D. Camm, and J.J. Cochran. *Statistics for Business & Economics, Revised.* Cengage Learning, 2017.

[2] Ralph B. D'Agostino and Albert Belanger. *A Suggestion for Using Powerful and Informative Tests of Normality.* The American Statistician 44.4, 1990.

[3] Skipper Seabold and Josef Perktold. *statsmodels: Econometric and statistical modeling with Python.* 9th Python in Science Conference, 2010.

[4] S. Vashishtha. Differentiate between the dni,dhi and ghi. `https://firstgreenconsulting.wordpress.com/2012/04/26/differentiate-between-the-dni-dhi-and-ghi/`.