A REPORT ON

Forecasting and Analysis of Renewable Energy

BY

Group 12 (Nightingale Group)

Akshit Khanna	2017A7PS0023P
Jhaveri Ayush Rajesh	2017A7PS0215P
Vitthal Bhandari	2017A7PS0136P
Shivankur Sharma	2017B4A71013P
Jyoti Ranjan Pagoda	2017B4PS0506P
Mohit Kulhari	2017B4TS1206P
Nishchay Pareek	2017B4TS1204P

Prepared in partial fulfilment of the course

Applied Statistical Methods - MATH F432

SUBMITTED TO

Dr. Sumanta Pasari

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

1. INTRODUCTION

Renewable energy is energy produced from sources that do not deplete or can be replenished within a human's life time. The most common examples include wind, solar, geothermal, biomass, and hydropower. This is in contrast to non-renewable sources such as fossil fuels.
Renewable energy accounts for 13.5% of the world's total energy supply, and 22% of the world's electricity.
For the past two years, the Indian Government has taken several initiatives such as introduction of the concept of solar parks, organizing RE-Invest 2015—a global investors' meet, earmarking of Rs.38,000 crore (Euros 4 billion) for a Green Energy Corridor, eight-fold increase in clean environment cess from Rs.50 per tonne to Rs.400 per tonne (Euro 0.62 to Euros 5 per tonne) and losses to be levied for solar and wind power, compulsory procurement of 100 per cent power from waste to energy plants, and Renewable Generation Obligations on new thermal and lignite plants, etc.

Renewable Energy Targets: The Indian Government has increased the target of renewable energy capacity to 175 GW by the year 2022 which includes 100 GW from solar, 60 GW from wind, 10 GW from bio-power and 5 GW from small hydro-power.

<table>
<tr><th>Source</th><th>Total Installed Capacity (MW)</th><th>2022 Target (MW)</th></tr>
<tr><td>Wind Power</td><td>34,046</td><td>60,000</td></tr>
<tr><td>Solar Power</td><td>21,651</td><td>1,00,000</td></tr>
<tr><td>Biomass Power (Biomass & Gasification and Bagasse Cogeneration)</td><td>8,701</td><td rowspan="2">10,000</td></tr>
<tr><td>Waste-to-Power</td><td>138</td></tr>
<tr><td>Small Hydro Power</td><td>4,486</td><td>5,000</td></tr>
<tr><td>TOTAL</td><td>69,022</td><td>1,75,000</td></tr>
</table>

Table I: Installed grid interactive renewable power capacity in India (excluding large hydropower) as of 31 March 2018 (RES MNRE)

Wind Power: Although India is a relative newcomer to the wind industry compared with Denmark or the US, domestic policy support for wind power has led India to become the country with the fourth largest installed wind power capacity in the world. As of 30 June 2018, the installed capacity of wind power in India was 34,293 MW. Wind power accounts for 10% of India's total installed power capacity. India has set a target to generate 60,000 MW of electricity from wind power by 2022. MNRE announced a new wind-solar hybrid policy in May 2018.

2. Descriptive Analysis and Statistics

2.1. Descriptive Analysis

Before diving into the core part of analysis we will define the variables associated with it that corresponds to wind and solar energy.

Direct Normal Irradiance is measured at the surface of the Earth at a given location with a surface element perpendicular to the Sun. It excludes diffuse solar radiation (radiation that is scattered or reflected by atmospheric components). Direct irradiance is equal to the extraterrestrial irradiance above the atmosphere minus the atmospheric losses due to absorption and scattering. The irradiance above the atmosphere also varies with time of year (because the distance to the sun varies), although this effect is generally less significant compared to the effect of losses on DNI.

Diffuse Horizontal Irradiance: (DHI), or *Diffuse Sky Radiation* is the radiation at the Earth's surface from light scattered by the atmosphere. It is measured on a horizontal surface with radiation coming from all points in the sky excluding *circumsolar radiation* (radiation coming from the sun disk). There would be almost no DHI in the absence of atmosphere.

Global Horizontal Irradiance :(GHI) is the total irradiance from the sun on a horizontal surface on Earth. It is the sum of direct irradiance (after accounting for the solar zenith angle of the sun z) and diffuse horizontal irradiance.

Global Normal Irradiance (GNI) is the total irradiance from the sun at the surface of Earth at a given location with a surface element perpendicular to the Sun.

Global Horizontal (GHI) = Direct Normal (DNI) X cos(θ) + Diffuse Horizontal (DHI)

These are all the terms relating to SOLAR ENERGY ,with these parameters in knowledge, one must take care while taking a preconceived notion of statistics values with the data.

Also parameters such as relative humidity,dew point and pressure could affect the solar energy analysis so are taken into account.

Parameters such as wind direction ,temperature ,dew point, relative humidity,snow depth,precipitable water were taken into account with coinciding parameters such as DHI,DNI and GHI. With these definitions and classification of parameters into solar energy and wind energy we are headed towards the descriptive statistics part.

2.2. Descriptive statistics

The statistics we used are correlation coefficients of all variables with both wind energy and solar energy. The prominent variables are identified and the variance they account for is calculated.

Variables	WIND SPEED	GLOBAL HORIZONTAL IRRADIANCE(GHI)
1.DHI	-0.1227	**0.91594**
2.CL-DHI	-0.1177	**0.93369**
3.CL-DNI	-0.1943	**0.9414**
4.Dew point	**0.2468**	NA
5.Pressure	**-0.4152**	0.04501
6.Relative humidity	**0.218**	**-0.3235**
7.Precipitable water	**-0.2615**	NA
8.Solar zenith angle	0.1433	**-0.8579**
9.Snow depth	NaN	NA
10.DNI	**-0.2862**	**0.93369**
11.Temperature	0.0203	**0.6628**

Table II : Correlation Coefficients with Wind Speed and GHI

Taking all the prominent variables into account on the basis of maximum correlation with wind speed we ultimately are left with **DNI, DEW POINT, PRESSURE, RELATIVE HUMIDITY, PRECIPITABLE WATER**, accounting for about **70%** of the total variance.

The variables such as **DHI, GHI, CLEAR SKY DHI, TEMPERATURE, SOLAR ZENITH ANGLE, SNOW DEPTH, WIND DIRECTION** showed minimal correlation with wind speed in this data.

Taking all the prominent variables into account on the basis of maximum correlation with **GHI**, we are ultimately are left with **DHI,DNI,CLEAR-SKY DHI,CLEAR-SKY DNI, RELATIVE HUMIDITY,SOLAR-ZENITH ANGLE,TEMPERATURE**, accounting for about **85%** of the total variance.

The variables such as **PRESSURE** show minimal correlation with GHI due to lower relative values.

3. GHI and Wind-Speed Analysis

Now we will find out that which probability distributions our datasets GHI and Wind Speed follows for the year 2000:
In order to test whether these datasets follows our specific distributions or not , we will take help from SPC software of excel which provides **p-p plo**t and **AIC** values.
Before testing our datasets we will proceed by defining the above terms:

3.1. P-P(Probability-Probability) plot

The P-P plots the empirical cumulative distribution function(CDF) values (based on the data) against the theoretical CDF values (based on the specific distribution). If the P-P plot is close to a straight line, then the specific distribution fits the data.

3.2. AIC(Akaike Information Criterion)

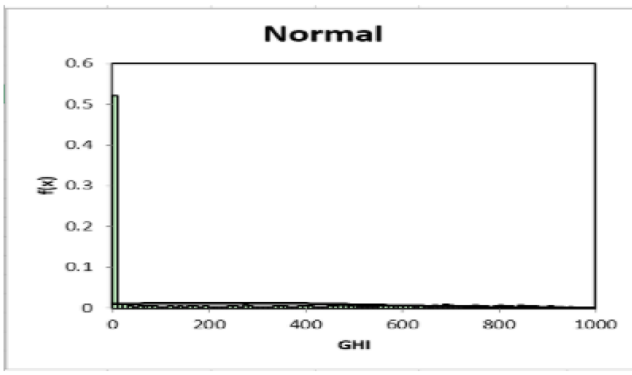
The main objective of AIC is to provide a means for model selection. AIC estimates the quality of each model, relative to each of the other models. Suppose that we have a statistical model of some data. Let K be the number of estimated parameters in the model . Let L' be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following:

$$AIC = 2K - 2\ln(L')$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus AIC rewards goodness of fit(as assessed by the likelihood function).

3.3. GHI Data Analysis

First we will start by plotting a histogram of our data set and will try to find out which distribution fits our data set.



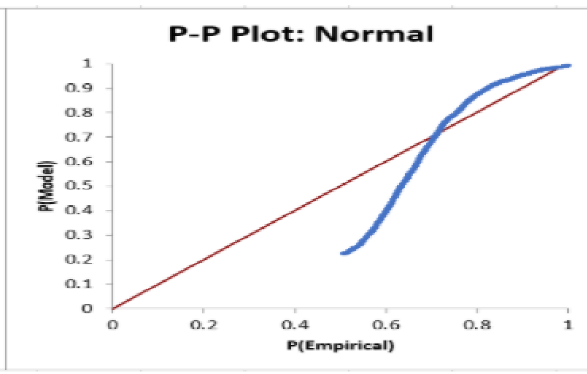


Fig 3.1: P-P plot of GHI dataset.

From the above plot, it is clear that our data for GHI dataset does not follow normal probability distribution, as the P-P plot clearly indicates that the cumulative distribution function of GHI dataset is not matching with the theoretical CDF of the normal distribution.
Hence we will look into different probability distribution models and their respective AIC values.
The table provides the suitables models for our datasets in an increasing AIC values;

Distribution Fitting Summary for GHI

Descriptive Statistics

Count	Mean	StDev	Median	Min	Max	Skew	Kurt
8760	233.2	308.8	0.000	0.000	975.0	0.936	-0.668

Distribution	Location	Shape	Scale	Threshold	Log-Likelihood	AD	p Value	LRT	AIC
LogNormal - Three Parameter	0.502		5.253	-0.0100	-31356	1231.0	*	*	62718
LogLogistic - Three Parameter	0.401		3.384	-0.0100	-32012	1100.2	<0.005	*	64030
Exponential - Two Parameter			233.2	0.000	-56518	3267.3	<0.001	*	113040
Normal	233.2		308.7		-62647	963.2	*		125298
Smallest Extreme Value	963.0		1.388		-4617717	-8615.2	>0.25		9235439
Largest Extreme Value	0.0000000678		0.000000100		############	-3319.4	>0.25		############

The following distributions could not be fitted since the data had negative values:
Exponential
Gamma
LogLogistic
LogNormal
Weibull

The following distributions did not converge on a solution:
Gamma - Three Parameter
Logistic
Weibull - Three Parameter

Table III: Distribution fitting summary table for GHI

From the table below we can compare the AIC values of these likely distributions:

S.NO	PROBABILITY DISTRIBUTION	AIC VALUE
1	logNormal-Three Parameter	62718
2	log Logistic - Three Parameter	64030
3	Exponential-Two Parameter	113040
4	Normal	125298

Table IV: AIC values for possible distribution models for GIC dataset

Since the AIC value of logNormal distribution(62718) is minimum from all other distributions. Hence we can conclude that the data for GHI dataset comes from logNormal probability distribution and the parameters of our distribution are:

$$P(x; m, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-[\log(x/m)]^2}{2\sigma^2}\right]$$

where **m=0.502 and σ=5.253.**

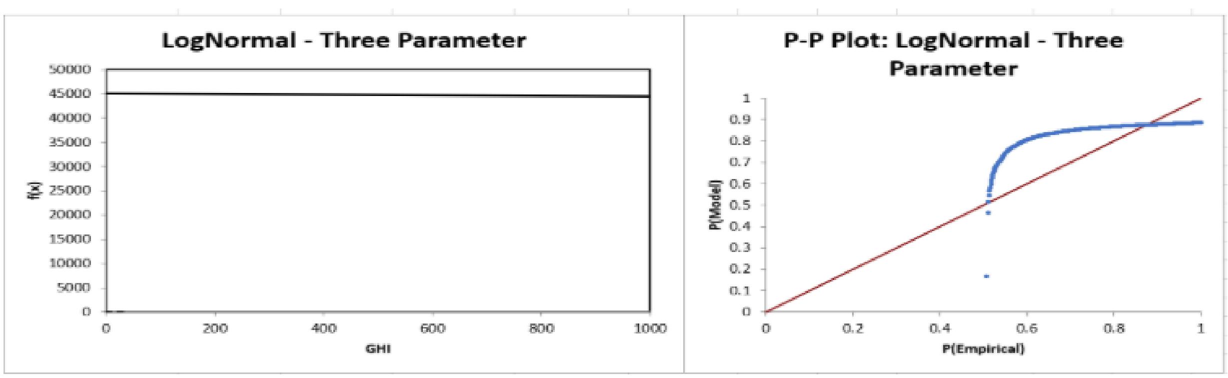


Fig 3.2: P-P plot of dataset winspeed and Weibull distribution

From the P-P plot in Fig 3.2, we can conclude that our CDF for dataset and model(logNormal) converged better than any other model.

3.4. Wind Speed Data Analysis:

For the wind speed dataset we will apply the same steps as we did for GHI dataset. In order to check whether our dataset follows Normal distribution or not, we will plot the histogram graph and the P-P plot.

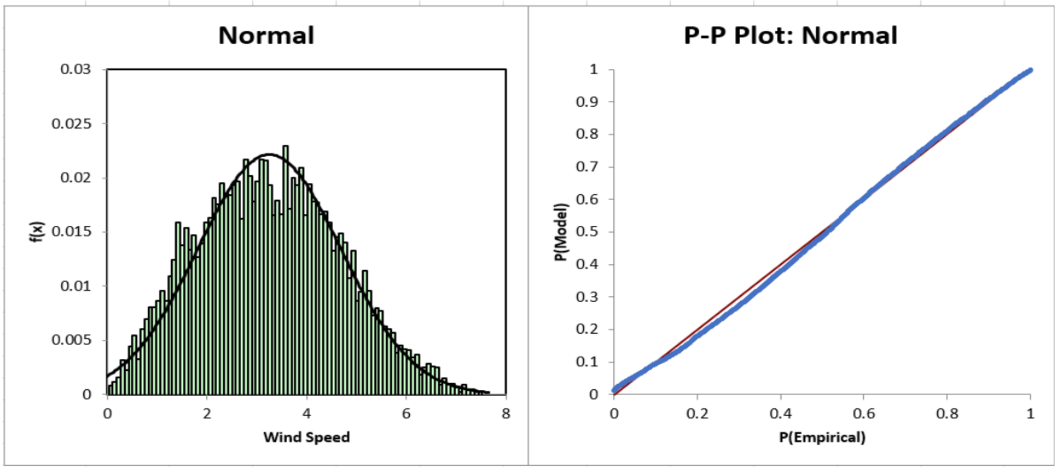


Fig 3.3: P-P plot for Normal distribution and wind speed dataset.

From Fig 3.3, we can initially assume that our dataset follows Normal distribution but we can check for other distributions. After using the SPC software for excel for our dataset we have different distribution models with their respective AIC values , the table below provides us the different possible models for our datasets . From the table, we can compare the AIC values of these likely distributions:

S.NO.	PROBABILITY DISTRIBUTION	AIC VALUE
1	Weibull-Three Parameter	30956
2	Weibull	31010
3	Gamma-Three Parameter	31156
4	logNormal - Three Parameter	31229
5	Normal	31725

Table V: AIC values for possible distributions for wind speed dataset

Since the AIC values of all other distributions are very similar to Normal distribution, we can conclude that the data for Wind-Speed dataset comes from a normal distribution.

4. Time Series Decomposition of Wind-Speed Data

The hourly data has been resampled to a weekly frequency time series for better data visualization purposes and reduced computation costs in forecasting.

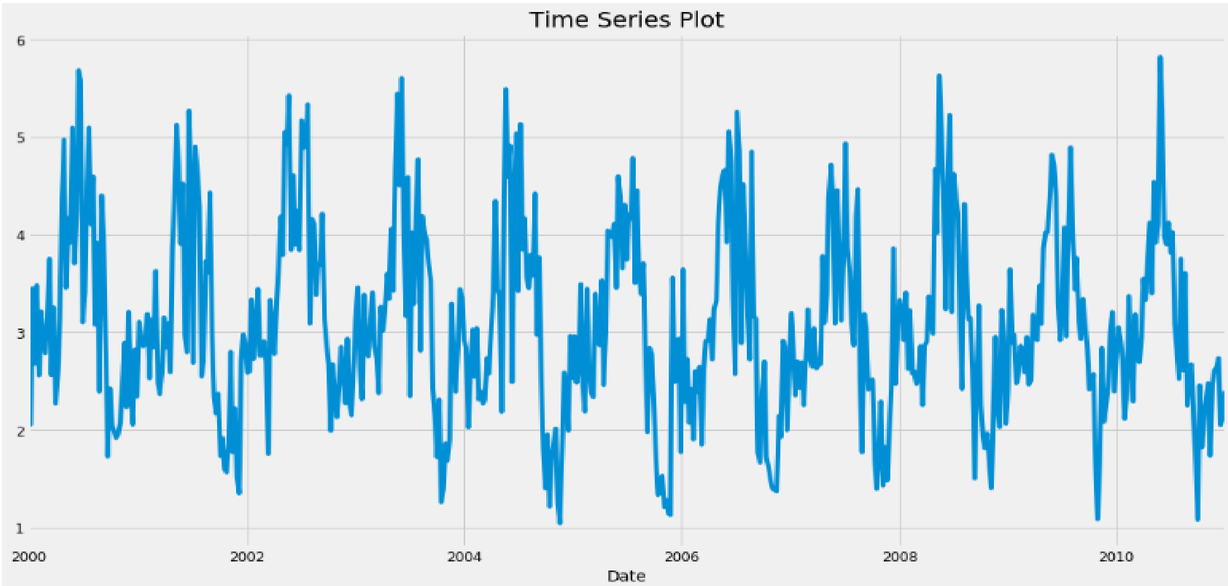


Fig 4.1: Time series plot of wind-speed data

Analysing from the time series plot in Fig 4.1, it can be inferred that there is seasonality present with a horizontal trend. But in order to validate this inference, the time series needs to be decomposed into various components.

The additive decomposition is the most appropriate as the magnitude of seasonal variations do not vary with the level of the time series. The equation for additive decomposition is

$$y_t = S_t + T_t + R_t$$

where y_t is the data, S_t is the seasonal component, T_t is the trend component and R_t is the residual component.

We proceed by decomposing the time series data using the statsmodels library of python.

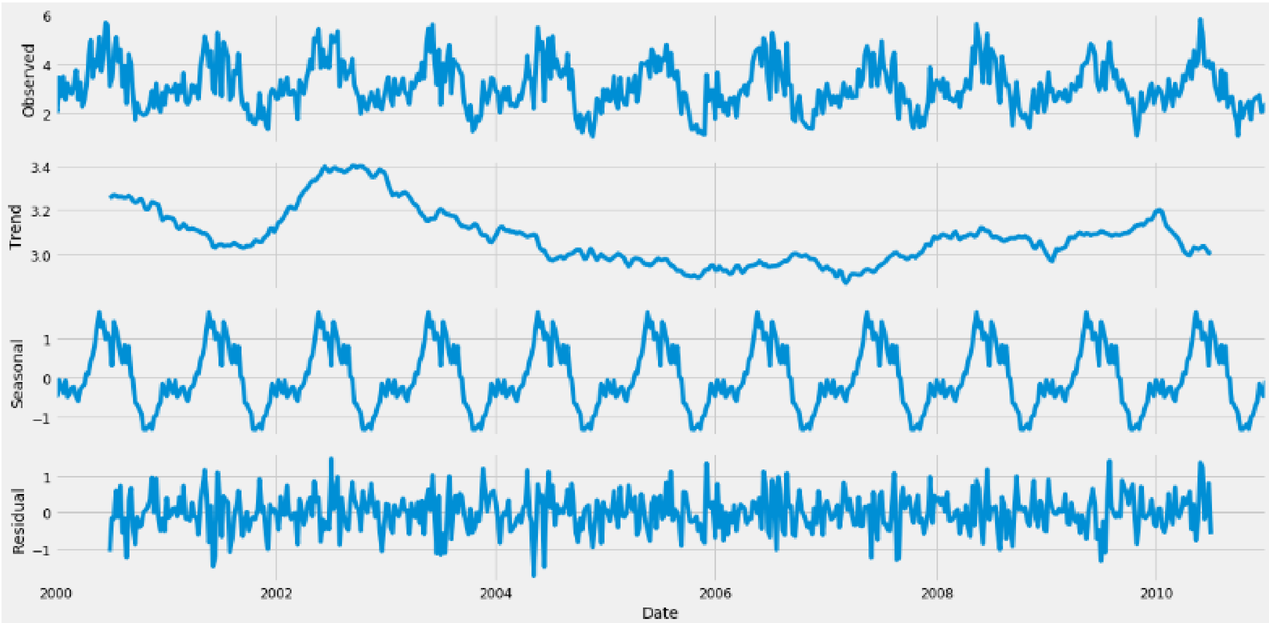


Fig 4.2: Additive time series decomposition of wind-speed data

Observing from the time series decomposition in Fig 4.2, the trend does not show any pattern, there is a clear pattern in the seasonal component and the residual component is random. Thus, it can be concluded that there is no trend in the data (horizontal trend) and there exists seasonality over an annual period.

5. Forecasting of Wind-Speed Data

The dataset used for forecasting is resampled to weekly frequency for reduced computation costs in forecasting models. The weekly dataset has a yearly seasonal period as concluded from the decomposition used above.
We proceed for forecasting by observing several time series forecasting methods.

5.1. ARIMA Forecasting

Auto Regressive Integrated Moving Average (ARIMA) is a class of models that explains a given time series based on its own past values and the lagged forecast errors, so that equation can be used to forecast future values.

The autoregressive (AR) part of the model indicates the variable is regressed on its past values. The integrated (I) part for differencing process makes the series stationary. The moving average (MA) part indicates the regression error is a linear combination of error terms. All these parts need to be tuned to best fit the data.

Any **non-seasonal time series** that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q where:

p : the order of the AR term
q : the order of the MA term
d : the number of differencing required to make the time series stationary.

ARIMA model in words:

Predicted Yt = Constant (After d differencing) + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)

5.2. SARIMA Forecasting

If a time series, has seasonal patterns, then you need to add seasonal terms to the ARIMA model and it becomes SARIMA, short for 'Seasonal ARIMA'.

This model has 4 more parameters P,D,Q and M.

- P,D and Q are autoregressive, differencing, and moving average terms for the seasonal part.
- M is used to refer to the number of periods in a season.

5.3. Forecasting Used

Since our data has prominent seasonal component and apart from it the time series is stationary, therefore we use SARIMA model to forecast Wind-Speed data.

Assumptions for SARIMA -

- There must be a consistent seasonal variation.
- The time series involved must be weakly stationary or can be integrated to form a stationary series apart from the seasonal component.
- The error terms are assumed to be independent and identically distributed variables sampled from a normal distribution with zero mean.

The parameters for SARIMA are selected by using the method of grid search by testing the AIC value to get the optimal parameters. AIC is used to estimate the quality of relative models and estimates goodness of fit. The lower the AIC value the better the fit is.

The weekly dataset has 52 periods in a season as a season consists of a single year. The data used for training is from 2000 to 2010 Wind-Speed data.

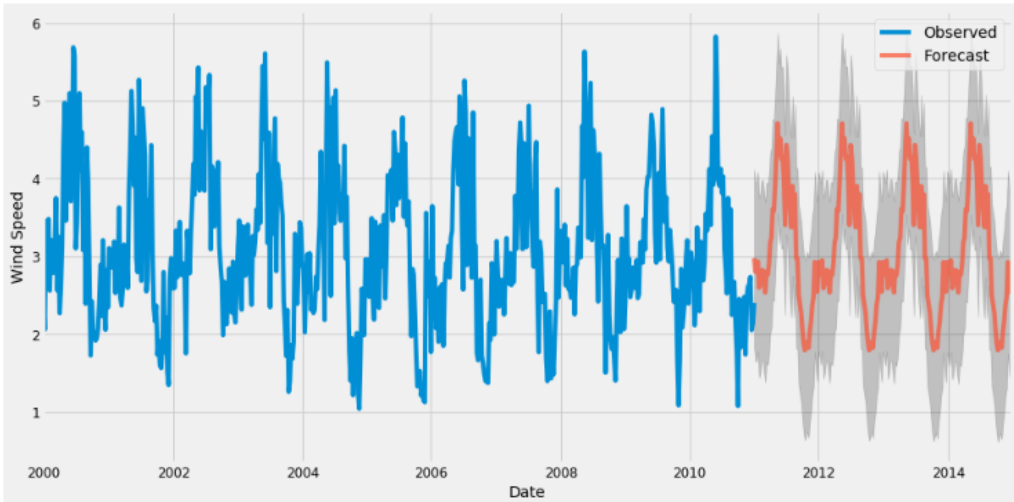


Fig 4.1: Forecast of SARIMA model for years 2011-14

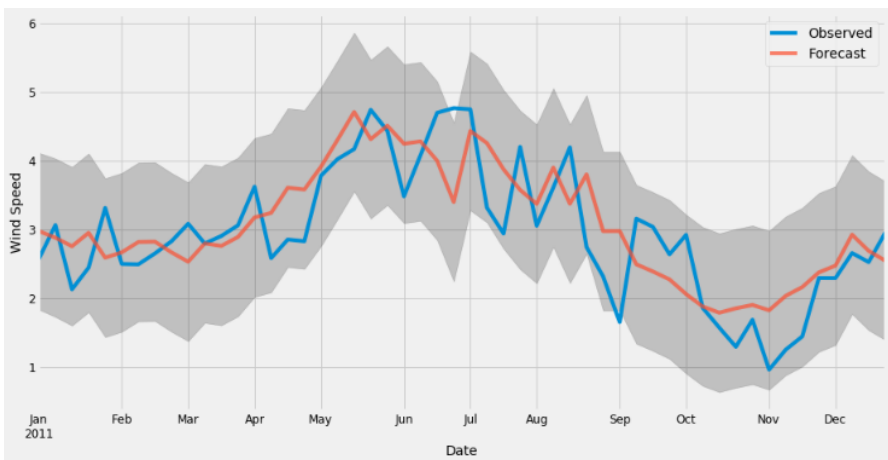


Fig 4.2: Comparison of the forecast and actual data of 2011 with confidence intervals

Root Mean Square Error (RMSE)	0.59
Mean Absolute Error (MAE)	0.49
Mean Absolute Percentage Error (MAPE)	19.53

Table VI: Forecast Evaluation Metrics

With a Mean Absolute Percentage Error of 19.53%, we can conclude that the SARIMA forecasting model has a has a reasonable performance.

```
ARIMA(1, 0, 0)x(0, 1, 0, 52)52 - AIC:1223.7302586299938
ARIMA(1, 0, 0)x(0, 1, 1, 52)52 - AIC:872.3168369023771
ARIMA(1, 0, 0)x(1, 0, 0, 52)52 - AIC:1203.793556807318
ARIMA(1, 0, 0)x(1, 0, 1, 52)52 - AIC:998.7098038695106
ARIMA(1, 0, 0)x(1, 1, 0, 52)52 - AIC:967.0445823681426
ARIMA(1, 0, 0)x(1, 1, 1, 52)52 - AIC:901.8985341569075
ARIMA(1, 0, 1)x(0, 0, 0, 52)52 - AIC:1222.7991887677113
ARIMA(1, 0, 1)x(0, 0, 1, 52)52 - AIC:1076.3055599133327
ARIMA(1, 0, 1)x(0, 1, 0, 52)52 - AIC:1219.9824150878615
ARIMA(1, 0, 1)x(0, 1, 1, 52)52 - AIC:869.8777096437917
ARIMA(1, 0, 1)x(1, 0, 0, 52)52 - AIC:1070.884101906287
```

REFERENCE::

https://www.spcforexcel.com/
https://en.wikipedia.org/wiki/Likelihood_function
https://en.wikipedia.org/wiki/P%E2%80%93P_plot
https://en.wikipedia.org/wiki/Akaike_information_criterion
https://en.wikipedia.org/wiki/Log-normal_distribution