A Report on

Time Series Analysis and Forecasting on Weather Data

Prepared in partial fulfillment of

MATH F432: Applied Statistical Methods

Submitted to

Dr. Sumanta Pasari, Assistant Professor

(Dept. of Mathematics)

On

November 22, 2019

By

Anuj Hydrabadi	**2017A8PS0420P**
Kartik Agrawal	**2017A4PS0443P**
Divyam Sharma	**2017A3PS0102P**
Nitin Vinayak Agrawal	**2017A4PS0415P**
Nayan Nilesh Bala	**2017A3PS0190P**
Debabrata Chaudhury	**2017A3PS0211P**
Aryamick Singh	**2017A3PS0389P**
Rohan Tibrewal	**2017A3PS0322P**

Table of Contents

Contents **Page No.**

INTRODUCTION

This is a story from the year 2011, set up in the historical town of Patan in Gujarat, lived 2 boys, Narendra and Mukesh. Mukesh was a daydreamer, always being in his imaginary world where he has all the luxuries of life, where he is miles away from the truth, that he lived as an underprivileged. Narendra, contrary to that, realized very early that their condition was what they had and he couldn't change it with the power of his imagination. This realization motivated him to work for the betterment of society at a very young age. While Mukesh went to the city to realise his dreams of name and fame, Narendra devoted himself completely in the service of his people. A major problem that his people faced was a reliable supply of electricity. To combat this problem, he called an engineer from the Ministry of Renewable Energy and tasked him with finding solutions to their energy requirements issues. He came up with a brilliant idea of going for Solar and Wind energy plants. Narendra, being new to this field inquired from the engineer about 'Renewable Energy'.

The engineer explained as follows: The 20th century saw an exponential increase in the usage of natural resources, mainly the fossilized fuels and gases to cater to our exorbitant energy requirements fueled by world wars and rapid urbanization. This led to a rapid depletion of these resources leading to the rise of renewable energy, largely to mitigate the impacts of climate change. Some of its types are geothermal, solar and wind energy, which are everlasting. Geothermal energy is harnessed from the heat derived from within the surface of the earth. This energy is characterised with only a few sites potentially able to cater to our requirements. Solar energy can be harnessed by two methods, Photovoltaic Cells (PV cells) and Solar Concentrators and is much more user-friendly. These utilities cost a premium for installation but don't require much maintenance afterwards. Wind Energy can be harnessed using windmills, which incorporate a wide variety of turbines. Then, he explained to him what India has achieved till date and what are its future ambitions. India has been heavily investing in a variety of renewable energy technologies. The total renewable energy capacity installed in the country crossed the 50 GW mark at the end of 2016 India is at the cusp of a solar revolution. The government is ambitious, having set a target of 100 GW by 2022 The National Institute of Solar Energy in India has determined about 750 GW of solar power potential in India. At present, the installed capacity is 30 GW. India has the lowest capital cost per MW globally for installation of solar plants. Karnataka, Telangana, and Rajasthan are states with the highest installed solar capacity. Today, India is the 4th largest wind market globally, with total installations having crossed the 31GW mark at the end of March 2017. India's wind power installations accounted for a 6.6 per cent share of the global market in 2016. The National Institute for Wind Energy's (NIWE) latest estimate for India's wind power potential is 302 GW at 100 meters. The major wind power states are Tamil Nadu, Gujarat, Karnataka, Maharashtra and Rajasthan. Narendra was very impressed with these new insights.

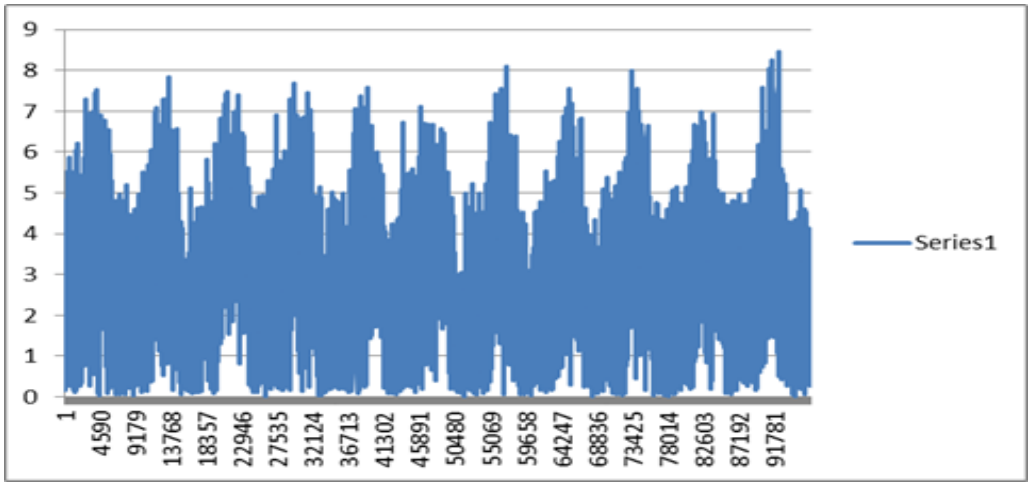
The engineer began his work by collecting meteorological data of that area for the past 11 years. On seeing new terms like GHI, DHI, and DNI, Narendra inquired about their meaning. DNI is direct normal irradiance which is the amount of solar radiation received per unit area by a surface that always helps perpendicular to the rays that come in a straight line from the direction of the sun at its current position in the sky. Diffuse Horizontal Irradiance (DHI) is the amount of

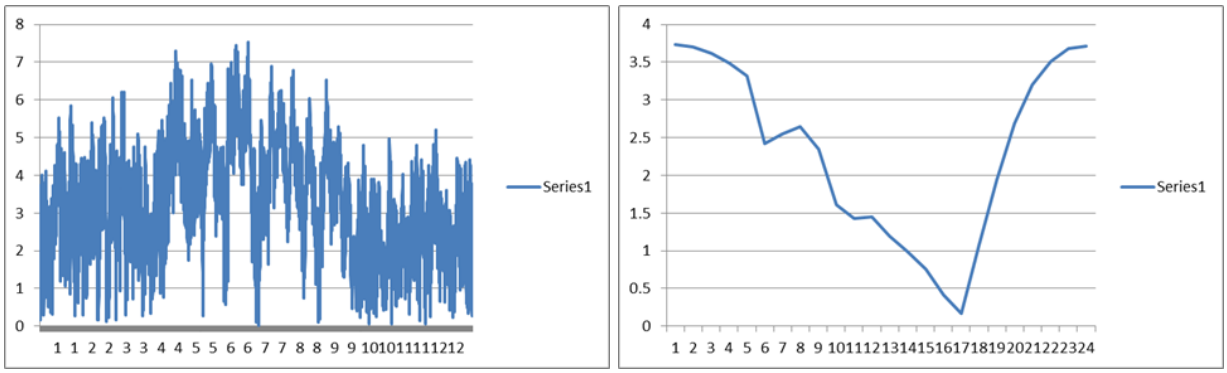
radiation received per unit area by a surface (not subject to any shade or shadow) that does not arrive on a direct path from the sun, but has been scattered by molecules and particles in the atmosphere and comes equally from all directions. Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation received from above by a surface horizontal to the ground. This value is of particular interest to photovoltaic installations and includes both Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI). Dew point is a point at which the air is unable to hold any more moisture. The solar zenith angle is the angle between the zenith and the centre of the Sun's disc. Precipitable water is the depth of water in a column of the atmosphere if all the water in that column were precipitated as rain. The parameters which influence **solar energy** are **DHI, DNI, humidity** and **solar zenith angle**. Meanwhile, **wind energy** is influenced by **temperature, pressure, precipitation and wind speed**.

2. Descriptive Analysis:

2.1. Wind Speed Analysis:

The following graph shows the hourly variations of wind speed across the 11 years ranging from 2000-2010. The Y-axis represents the wind speed (in m/s) and X-axis is in Sr.No. of datapoint (which is taken at an interval of 1 hr at 30 min i.e. half-mark of every hour everyday). We can clearly see a periodic behaviour in the wind speed data every year. This may be attributed to the nearly fixed weather patterns at the location Charanka.

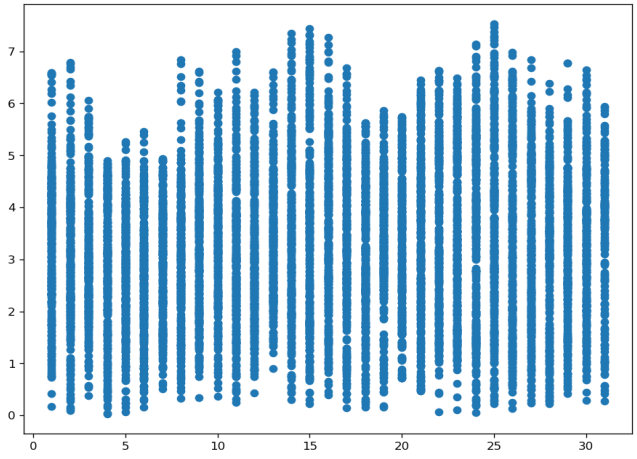




The **left graph** represents the wind speed over a year in 12 months (of year:2000). The peaks can be seen in the months of June (month 6) & July (month 7). This is due to the fact that the **monsoon** season arrives in India during that period. The low pressure created during summers attracts the monsoon clouds and fast winds (South-West trade winds) which increases the overall wind speed across the Indian Subcontinent.

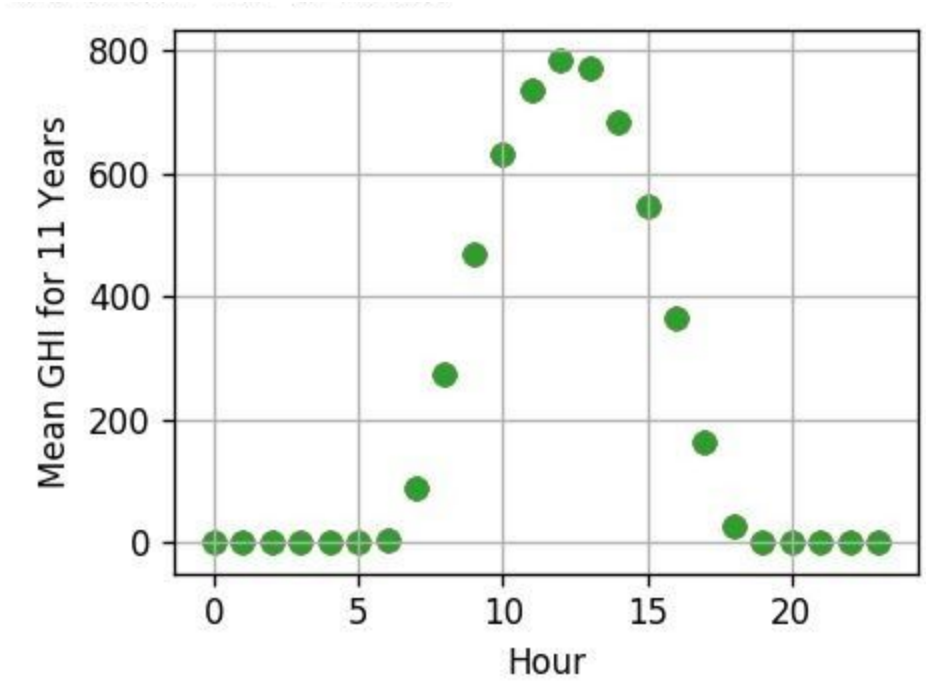
The **right graph** depicts the variations of wind speed recorded over a day (here January 1, 2000). So, we can infer that the maximum speed is reached during night hours (after sunset) and minima is at around 5 PM. This phenomenon can be attributed to the concept of sea breeze (from Arabian Sea) over Gujarat during the night.

Below graph (left) represents a scatter plot containing the wind speed on a period of one month (January, 2000). After every fortnight, we can see an increase in the speed. This will be attributed to the new moon and full moon phenomena (high & low tides).



2.2. GHI Analysis

The following plot shows the mean GHI for every hour of the day for the entire period of 11 years.



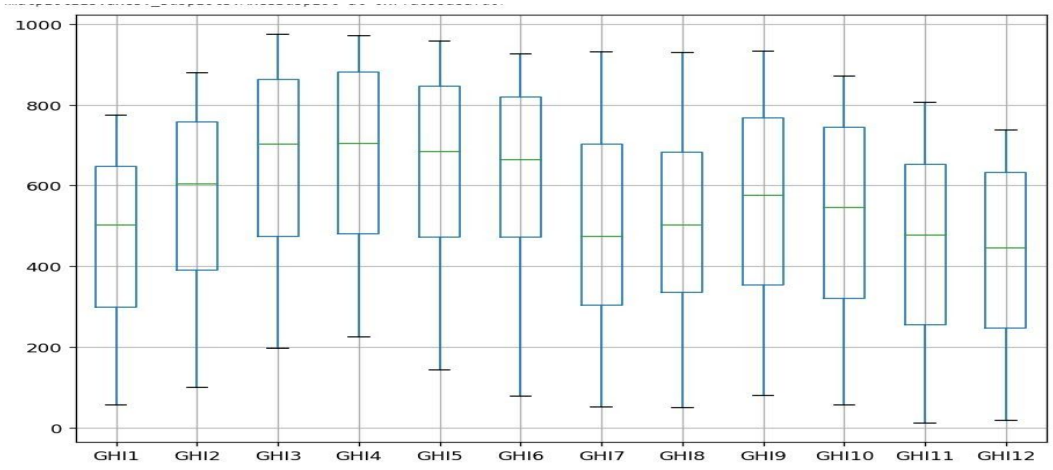
Kurtosis: It is the measure of the heavy-tailed ness of the data when compared to a normal distribution of the same mean and standard deviation

The value comes out to be -1.2297852906976423.

The negative value implies that our distribution is flatter than a normal distribution of the same mean and standard deviation.

Skew: It is a measure of the asymmetry of our distribution. The skew comes out to be -0.08121768534127373. The negative value of skew implies that the left tail of our data is thicker than the right tail.

Box whisker plot: It shows the five-number summary (minimum, lower quartile, median, upper quartile, maximum) of the data for each of the 12 months of the year 2000.



The following plot shows the GHI values at six different hours (hours 10 to 15) of the day where we receive peak sunlight.

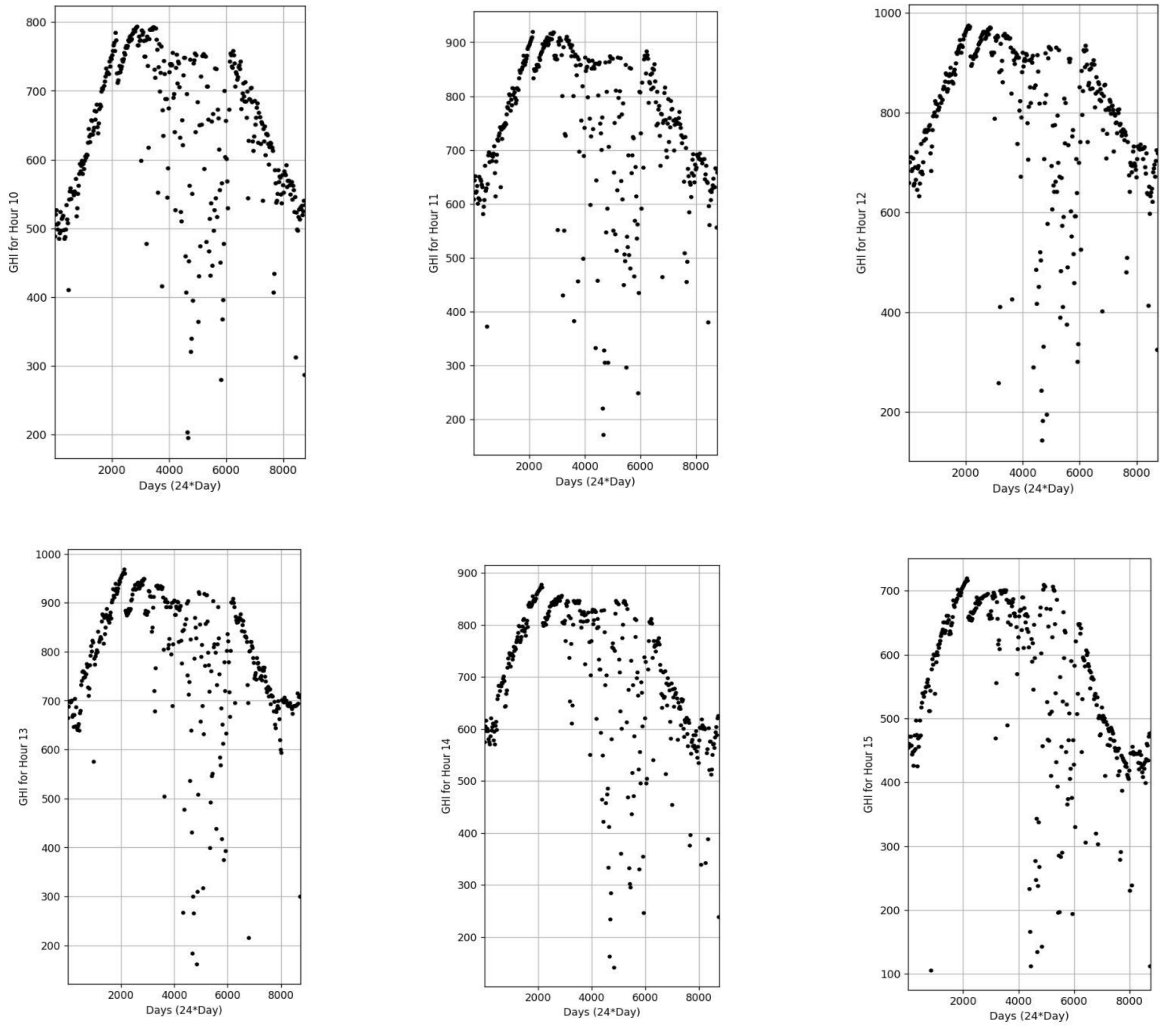


Fig. Comparing GHI Values for Hours 10-15 in the Year 2000

Distribution of the data: The data has been tested for various distributions (normal, exponential, lognormal, gamma, beta and Weibull) using the Kolmogorov-Smirnov test for the null hypothesis that the data belongs to that distribution. The p-value for each of the cases turns out to be approximately 0. Hence our data does not follow any particular distribution.

3. Time series analysis of GHI data:

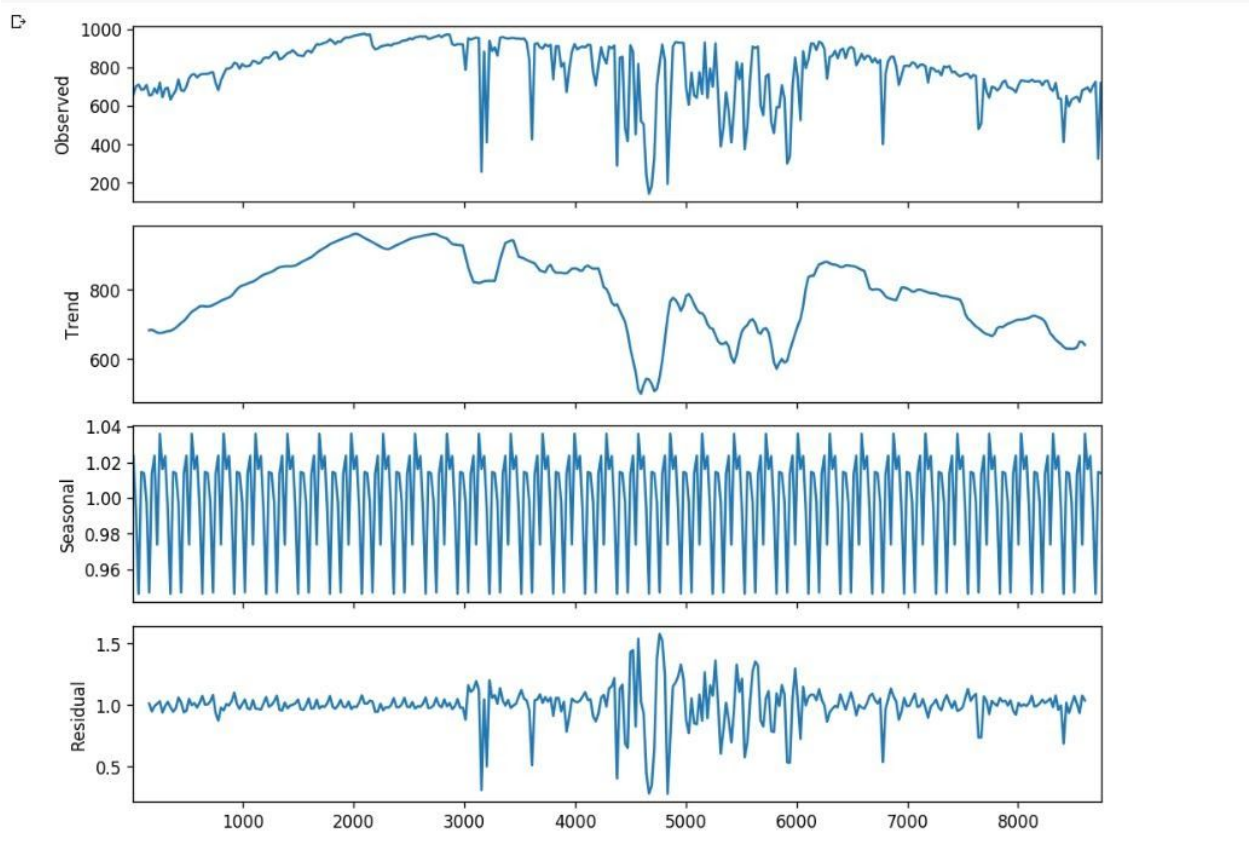
ADF test for stationarity of time-series data: The Augmented Dickey-Fuller test is the test to determine if our time-series data is stationary. We test the null hypothesis that the data is non-stationary, against the alternative hypothesis, that the data is stationary.

The test statistic obtained is -12.438007603191737.

The corresponding p-value is 3.8086422673487954e-23 which is approximately equal to zero.

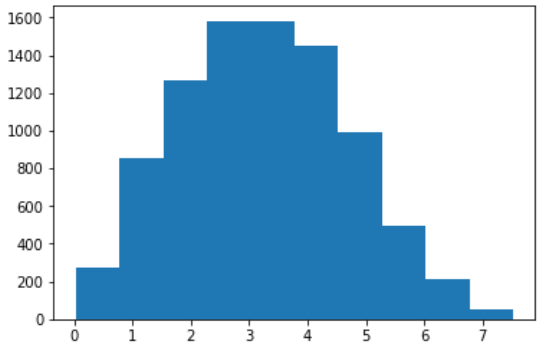
Hence, we reject the hypothesis that the data is non-stationary. The data is stationary.

We now proceed to the separation of the various components of the time series data, shown in the figure below.



4. Inferential Analysis:

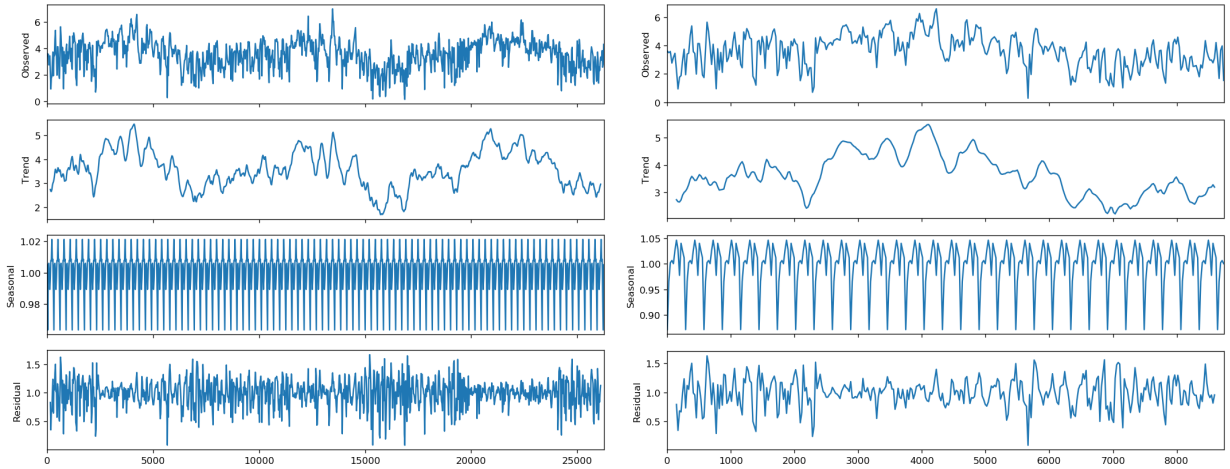
The histogram of wind speed for the year 2000 is as shown below.



Our **first objective** is to find the distribution of wind speed for the 2000 data. We have applied KS-test for various important distributions and according to the p-values obtained we finally selected that distribution to fit our model. After that step, we used Maximum Likelihood Estimation (MLE) for estimating the values of our distribution. (AIC wasn't used as the no of parameters for our model is not large so MLE provides good results). All the process was implemented in Python and re-verified in MATLAB. The Python code for the same is attached in Appendix.

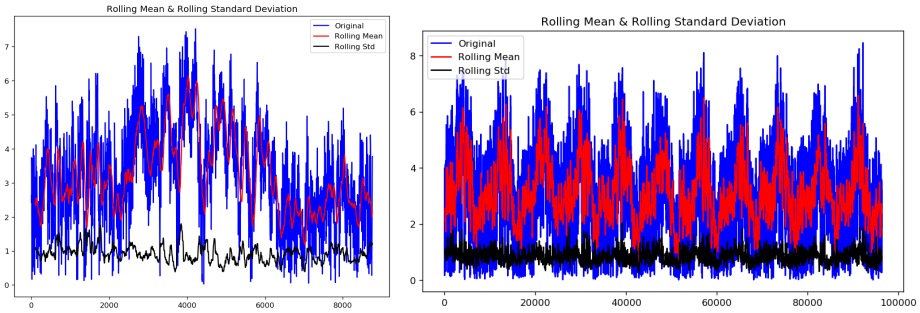
Our best fit distribution was found to be **Weibull-min** distribution with shape, scale and location parameter to be found as (after MLE) 0.879, 0.025 & 1.664 respectively.

Second objective was to decompose the time series into trend, seasonal and residual components



The left graph shows the decomposition of 3 years time series data (the complete 11 years data was getting too dense due to many data points so was avoided). The right graph shows the yearly data (year 2000) decomposed into the required 3 components.
We chose **multiplicative model** due to the fact that the wind speed changes with the weather. Thus, the seasonal value changes over the year and it's not independent of time.
Trend was not found in our data as can be viewed in the graph. Also, we have carried out the Add-Fuller test for the same (code attached in appendix) where p-value obtained was in the order of 10^(-8).

There's **seasonality** in the data as can be clearly seen in the graph as the wind speed changes **periodically** over the day across the complete year.
Residual plot is symmetric about zero axis which indicates normality and is expected to be.
Third objective was to provide a method for forecasting the wind speed.



Forecasting the decomposed seasonal and trend patterns obtained separately to predict the wind speed with the multiplicative model for next month. The above graphs depict that our time-series is of stationary nature as the overall mean averages out over the period of the entire year. The left graph shows the moving average for a month and and the right one shows the time series data over a period of 11 years. So, the red curve can be seen to follow a periodic pattern which averages out over the entire year and this cycle repeats every year with slight changes.

Hence, after all this analysis, Narendra found the prospective of Renewable Energy appealing and decided to implement it for helping his fellow town people.

APPENDIX

In [68]:
```
import scipy.stats as st
def get_best_distribution(data):
    dist_names = ["norm","weibull_min","weibull_max","gamma","expon","lognorm","beta"]
    dist_results = []
    params = {}
    for dist_name in dist_names:
        dist = getattr(st, dist_name)
        param = dist.fit(data)

        params[dist_name] = param
        # Applying the Kolmogorov-Smirnov test
        D, p = st.kstest(data, dist_name, args=param)
        print("p value for "+dist_name+" = "+str(p))
        dist_results.append((dist_name, p))

    # select the best fitted distribution
    best_dist, best_p = (min(dist_results, key=lambda item: item[1]))
    # store the name of the best fit and its p value

    print("Best fitting distribution: "+str(best_dist))
    print("Best p value: "+ str(best_p))
    print("Parameters for the best fit: "+ str(params[best_dist]))

    return best_dist, best_p, params[best_dist]
```

In [69]:
```
get_best_distribution(data)
```

```
p value for norm = 4.957589316492145e-06
p value for weibull_min = 0.0
p value for weibull_max = 0.000883030470997937
p value for gamma = 4.883460563172084e-05
p value for expon = 0.0
p value for lognorm = 0.00011279407003596572
p value for beta = 0.11324311056186273
Best fitting distribution: weibull_min
Best p value: 0.0
Parameters for the best fit: (0.8794880570070567, 0.025221297999999996, 1.6638144563016062)
```

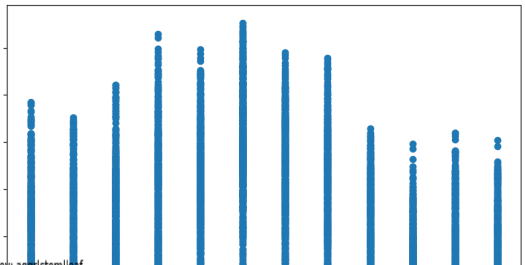
Out[69]:
```
('weibull_min',
 0.0,
 (0.8794880570070567, 0.025221297999999996, 1.6638144563016062))
```

In [8]:
```
from statsmodels.tsa.stattools import adfuller

results=adfuller(data)
```

In [9]:
```
results
```

Out[9]:
```
(-6.479946390654602,
 1.3025584483729644e-08,
 37,
 8722,
 {'1%': -3.431099968539641,
  '5%': -2.86187143613454,
  '10%': -2.5669464184887825},
 -8461.72454175441)
```



```
7.525559425
  900   7 2235
  896   6 00001111222333344555666778888899
  863   5 00000000001111111111111222222222222233333333333334444444455555555555666666666777788888888899999
  761   4 000000000000000000011111111111111111222222222222233333333333334444444444444444455555555555555566666666666666666677777777777777788888888888889999999999999999999
  576   3 0000000000000000011111111111111111111111222222222222222222233333333333333333344444444444444444445555555555556666666666666777777777777777777788888888888888888888899999999999999
  379   2 000000000000000001111111111111111112222222222222222222223333333333333333333344444444455555555555555555566666666666666666666777777777777777777777888888888888999999999999999
  182   1 00000000111111111122222222333333333333333444444444444444444555555555555555566666666666666666667777777777777778888888888888999999999999999
   41   0 12224555555666666777777778888888899999999
0.025221298
```