Statistical Analysis and Forecasting of Wind Energy

Under the guidance of Prof. Sumanta Pasari

Department of Mathematics

By

Bhavya Maroo (Group Leader)	**2018B4A70905P**
Aman Maheshwari	**2018B4A70906P**
Nandita Suresh Kamath	**2018B4A20868P**
Sparsh Agarwal	**2018B4A40936P**
Akshat Lal	**2018B4A70051P**
Sanskriti Jindal	**2018B4A40486P**
Umang Jain	**2018B4A80885P**
Samarth Joshi	**2018B4A20873P**

Prepared in partial fulfillment of the

Applied Statistical Methods (MATH F432)

Birla Institute of Technology and Science, Pilani

(December 2021)

Table of Content

Introduction

Renewable energy refers to the form of energy that is naturally obtained from the environment and from sources that can be replenished naturally. There are many advantages of renewable energy including less emission of waste in the environment, less maintenance cost, more economical than using fossil fuels and they do not deplete, so better prospects for the future. Various renewable energy sources include wind energy, solar energy, geothermal energy, biomass, and hydropower.

This report will now focus on wind energy. Wind energy has become one of the most economical renewable energy sources in recent times. In India, the growth of the wind industry has resulted in a strong ecosystem, project operation capabilities, and manufacturing base of about 10,000 MW per annum. The country currently has the fourth-highest wind installed capacity in the world with a total installed capacity of 39.25 GW (as of 31st March 2021) and has generated around 60.149 billion Units during 2020-21.

The purpose of this assignment is to analyze and forecast wind speed for the states Rajasthan, Tamil Nadu, Andhra Pradesh, and Madhya Pradesh.

Dataset

The data provided to us contains the hourly data for the year 2000-2014 containing various factors that may be associated with renewable energy production. The data is associated with four locations from the state of Rajasthan, MP, Andhra Pradesh, and TN. For each row, the various factors(columns) given are Year, Month, Day, Hour, Minute, DHI, DNI, GHI, Clearsky DHI, Clearsky DNI, Clearsky GHI, Dew Point, Temperature, Pressure, Relative Humidity, Solar Zenith Angle, Snow Depth, and, Wind Speed.

The following is a brief explanation of terminologies and their connection to wind energy:

Diffuse Horizontal Irradiance (DHI): Measures the light reaching the earth's surface after being scattered by clouds and particles in the atmosphere.

Direct Normal Irradiance (DNI): The amount of light received by the surface that is always held perpendicular to the sun.

Global Horizontal Irradiance (GHI): It represents the total amount of shortwave radiation received from above by a surface that is horizontal (parallel) to the ground.

Global Horizontal Irradiance (GHI) = Direct Normal Irradiance (DNI)* cos (solar zenith angle) + Diffused Horizontal Irradiance (DHI)

Humidity: It is a measure of the amount of water vapor in the air. The humid air implies lower density resulting in lower power from a wind turbine.

Dew Point:: It is the temperature to which the air must be cooled to become saturated.

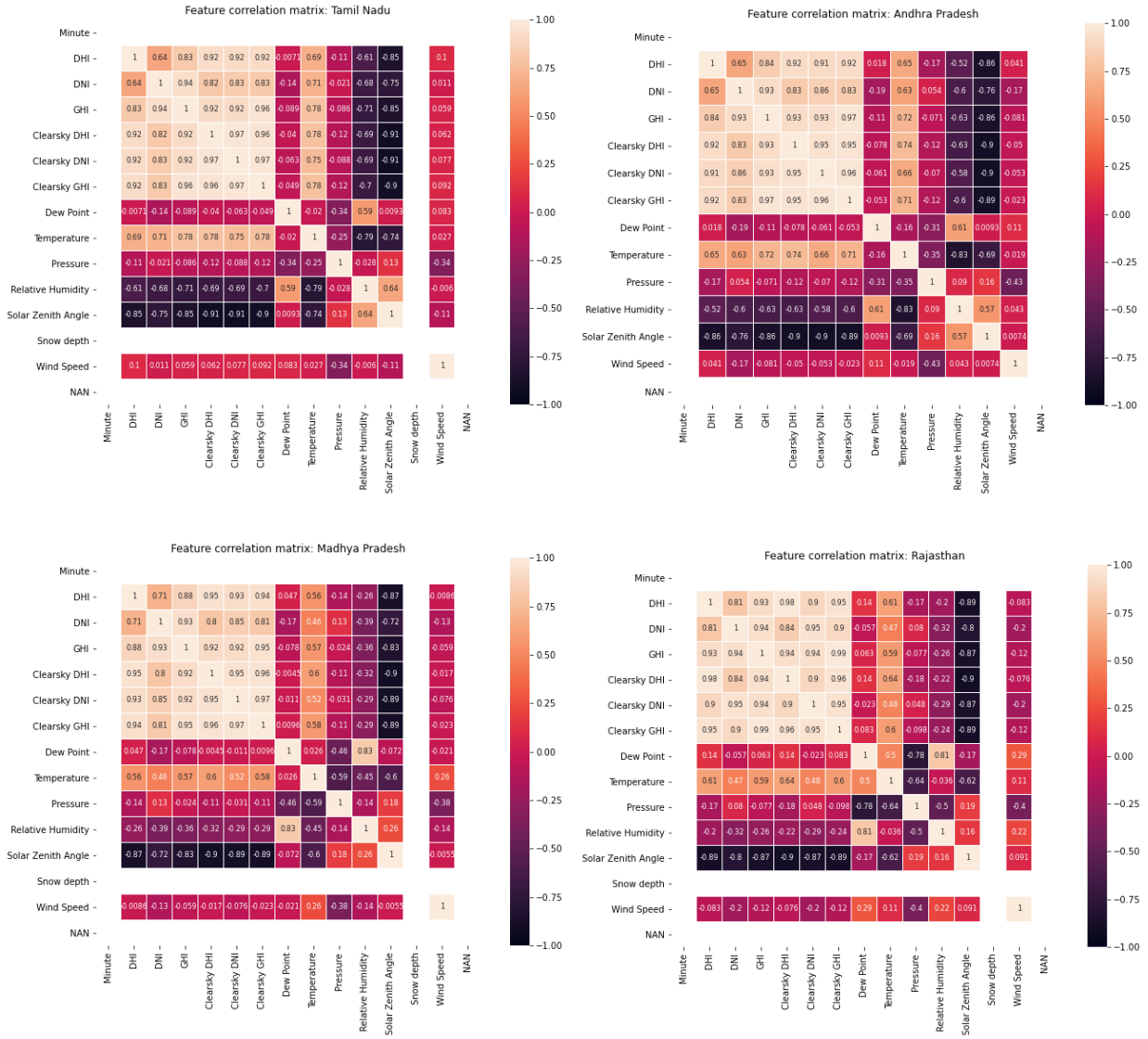
Temperature: When the temperature decreases, density increases, and thus wind energy generated increases.

Wind Speed: More the wind speed, the more is the wind energy generated.

Pressure: Air moves from high pressure (low temperature) to low pressure (high temperature).This difference in pressure results in strong winds.

Of all the terms we see that the terms which are more relevant to Wind energy are humidity, dew point, Wind speed, temperature, pressure, and the terms DHI, GHI, DNI are not that relevant for wind energy but may be useful for solar energy analysis. The main attribute used for analysis and forecasting is wind speed, as it is directly related to Wind energy.

Feature Correlation



We obtain feature correlation matrices for the attributes of the dataset. From the heatmap we observe some trends for the attributes relevant to wind energy in correlation with wind speed. In case of pressure, we have negative correlation. This is expected as an increase in pressure. Gases move from high-pressure areas to

low-pressure areas. The bigger the difference between the pressures, the faster the air will move from the high to the low pressure. For dew points, we have a low positive correlation. In the case of temperature and relative humidity, we have a low positive/ negative correlation, indicating that they do not have a direct relation with wind speed, because they are more related to pressure.

Normality Tests

The extent to which we can analyze the time series data depends on whether the data is normally distributed, as most statistical models operate on the assumption of normality of the data. Thus, we first decided to check the normality of the given Wind Speed data through two tests:

- Shapiro-Wilk Test
- D'Agostino's K-squared test

The test statistic computed differs, but the hypotheses remain the same across the two tests:

➢ H0: Data is normally distributed

➢ H1: Data is not normally distributed

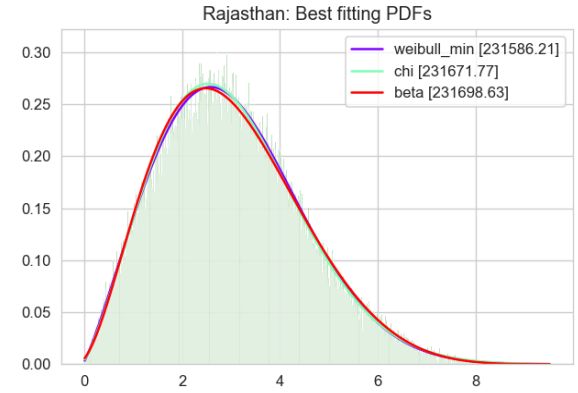
Results

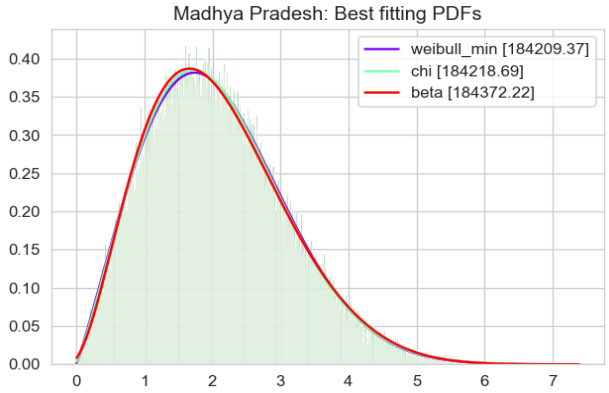
The results obtained for the data are as shown. Since both statistics give us an alternative hypothesis, we conclude that the data is not normal.

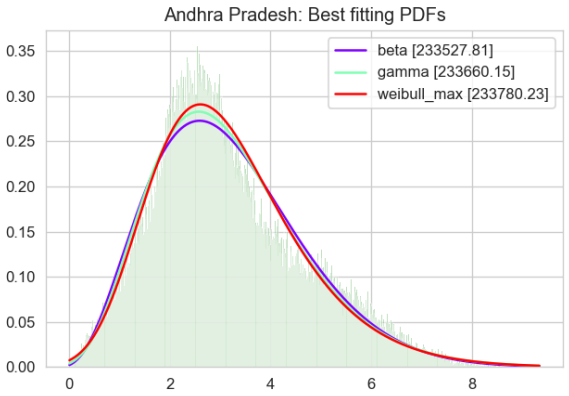
```
Rajasthan:
Shapiro-Wilk Statistic=0.983, p=0.000
Sample does not look Gaussian (reject H0)

D'Agostino's K^2 Statistic=4618.005, p=0.000
Sample does not look Gaussian (reject H0)



Madhya Pradesh:
Shapiro-Wilk Statistic=0.978, p=0.000
Sample does not look Gaussian (reject H0)

D'Agostino's K^2 Statistic=6340.722, p=0.000
Sample does not look Gaussian (reject H0)



Tamil Nadu:
Shapiro-Wilk Statistic=0.962, p=0.000
Sample does not look Gaussian (reject H0)

D'Agostino's K^2 Statistic=9026.001, p=0.000
Sample does not look Gaussian (reject H0)



Andhra Pradesh:
Shapiro-Wilk Statistic=0.968, p=0.000
Sample does not look Gaussian (reject H0)

D'Agostino's K^2 Statistic=7710.597, p=0.000
Sample does not look Gaussian (reject H0)
```

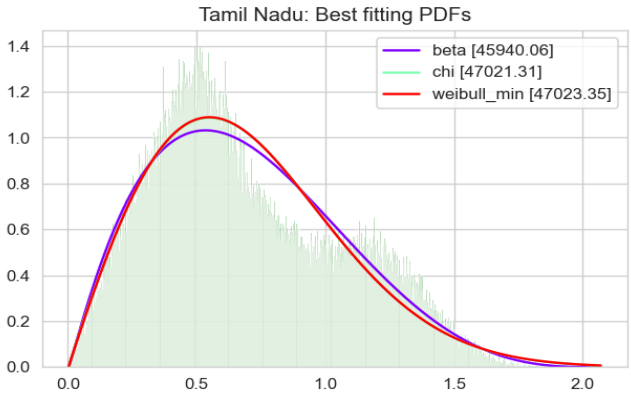
Best Fit Distributions

We check well-known distributions to find the best fit for our wind speed data. The following distributions were considered: Beta, Gamma, Rayleigh, Normal, Logistic, Weibull, Lognormal, Chi-Squared, and Exponential. We observe that these fit the data much better than the normal distribution and plot the best 3 distributions.



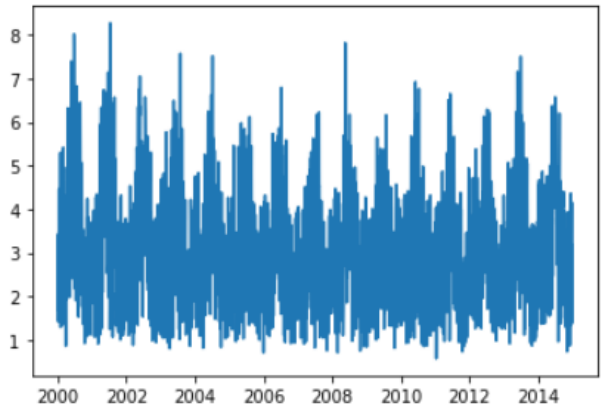




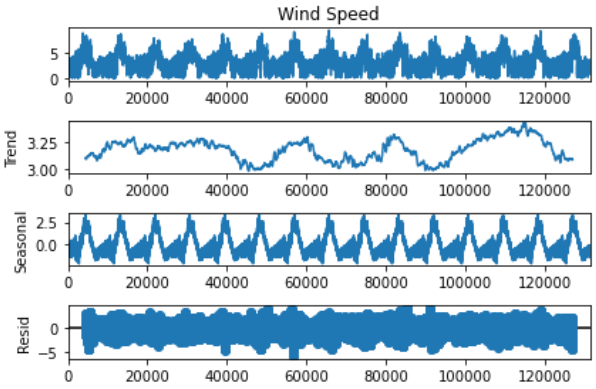


Time Series Decomposition

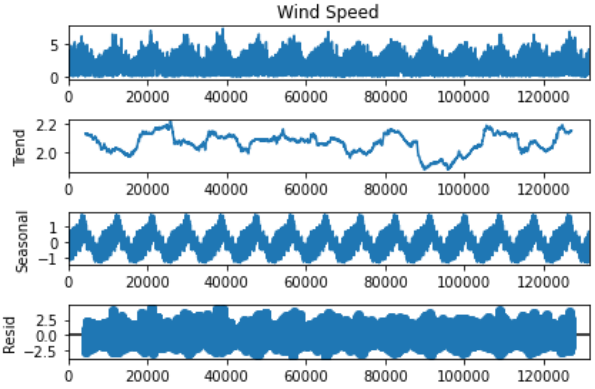
A plot of the time series clearly shows a seasonality component, in a real sense as well, as they vary according to real seasons. The time series were decomposed into Trend, Seasonality, and Residual components.



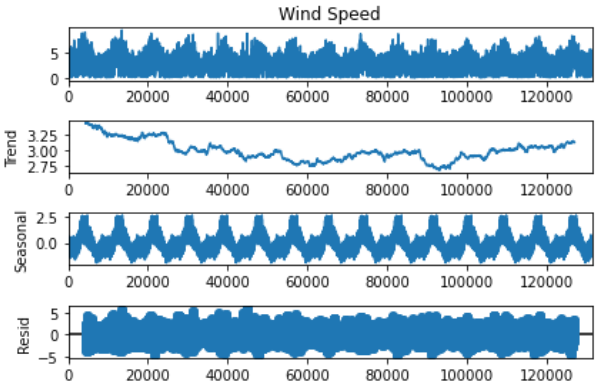
We decomposed the time series into trend, seasonality, and residual components using the additive model, as shown below for the states. The observations follow.



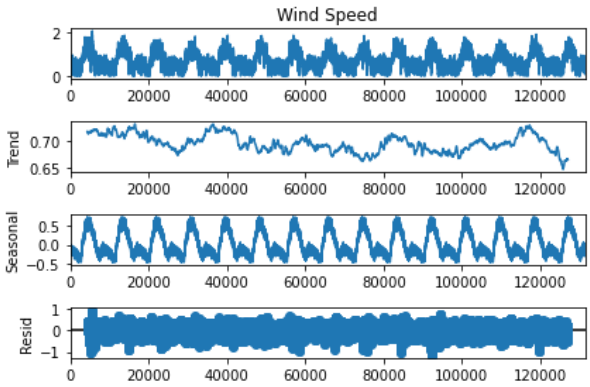
(Andhra Pradesh)



(Madhya Pradesh)



(Rajasthan)



(Tamil Nadu)

The uniform seasonality, no uniform variation in the trend component, and low residuals gave us justification for the choice of the additive decomposition model. Moreover, the variations in Wind Speed for each of the states can be explained through natural seasonal wind patterns in each location. For instance, the observed seasonal peak around May-June each year in the data for Rajasthan can be explained by the Loo, which hits a peak around that time as well. Finally, the data looks to be roughly stationary for the time series.

Stationarity Tests

The Augmented-Dickey Fuller Test (ADF) and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests were used to validate the data's presumed stationarity (at a 1% level of significance). The following are the specifics for the states.

ADF Test

The following are the hypotheses for this test:

H0: unit root present

H1: no unit root present, data is stationary

As indicated, the test statistic's result is more negative than the 1 percent critical value. As a result, the null hypothesis is rejected, and we infer that this is a stationary series.

```
Andhra PradeshTest Stat: -13.419314697388518
p-value: 4.2204640597964625e-25
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587


Madhya PradeshTest Stat: -21.875115048848766
p-value: 0.0
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587


RajasthanTest Stat: -22.48771245276804
p-value: 0.0
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587


Tamil NaduTest Stat: -11.656812627313302
p-value: 1.9763893073538965e-21
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587
```

KPSS Test

The following are the hypotheses for this test:

H0: There is no unit root, and the data is stationary.

H1: unit root is present

As seen, the p-value is more than the alpha, indicating that we cannot reject the null hypothesis, resulting in the conclusion that the data is stationary.

```
Andhra PradeshTest Stat: 0.15758708105037406
p-value: 0.1
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347


Madhya PradeshTest Stat: 0.3601461711453458
p-value: 0.09433354692010956
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347


RajasthanTest Stat: 1.9474099510337564
p-value: 0.01
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347


Tamil NaduTest Stat: 0.1728249842137999
p-value: 0.1
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347
```

Conclusion

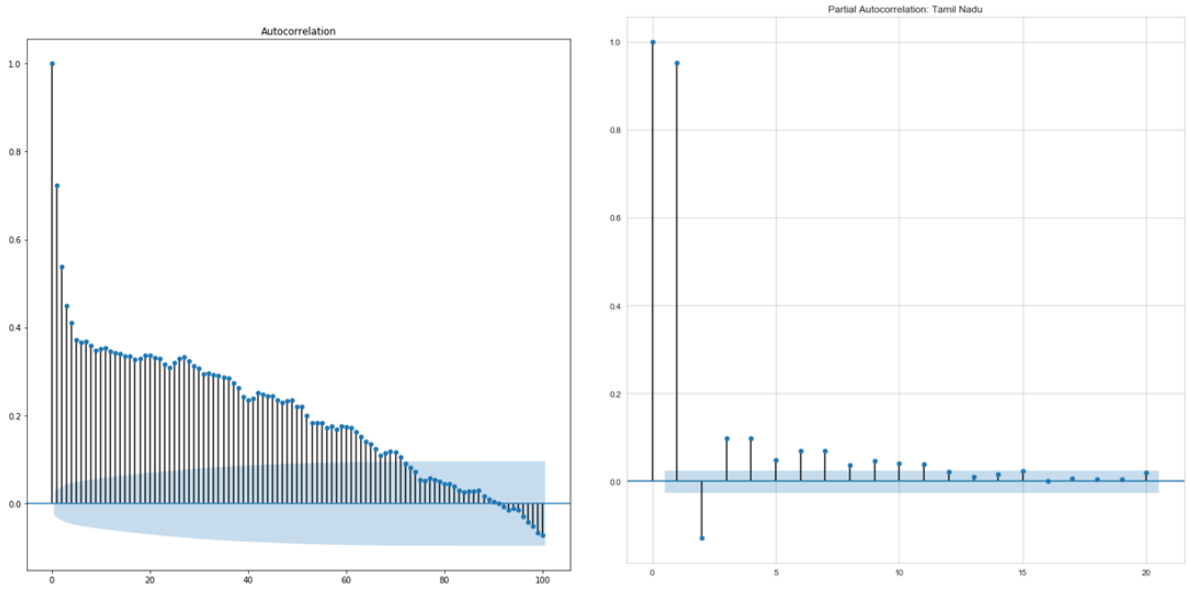
We discovered that the data was stationary for all four states using the two tests. As a result, we proceed under the premise that the Wind Speed data time series for all four states is stationary .

Model Fitting and Forecasting on Wind Speed data

Here, we present the application of AR, MA, ARMA, ARIMA, and SARIMA methods and apply them to the given dataset. The first four models are used to predict daily and weekly data, while the SARIMA model was only used to predict weekly and monthly data due to a lack of assessing power. Data was compiled as a measure over time (daily/weekly/monthly). This section will analyze the data section of the single region (Rajasthan). Results for other regions can be found at the end.

Parameter Estimation

We plotted the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots for the data – the plots for the daily Wind Speed data of Rajasthan are as shown below.



Counting the number of significant lag values before the data entered the confidence zone (shown in blue) provided us with estimates for the parameters of the MA and AR models.

The parameters were obtained through grid search for the other models, keeping the computational time in mind. The final models for daily data of Rajasthan were found as:

- AR (16)
- MA (62) – but MA (20) was used for fitting due to assessing power constraints

● ARMA (9, 10)
● ARIMA (9, 1, 10)

The SARIMA model for monthly data was found as:
● SARIMA ([2, 0, 3], [2, 0, 1], 12)

Model Fitting

To select the best model for out-of-sample forecasting, all of the models were fitted to the data, and MAPE values were calculated using the residuals. To justify the noise being 'white,' residual ACF graphs are also presented for the ultimate best-fitting models. The following are the fittings:

AR Model:

This model is a multiple regression model that estimates the target variable's values based on a linear combination of its previous values. The following equation can be used to describe it: X is the forecasted variable, is the autoregressive operator (polynomial of order p), B is the backshift operator, and w is the white noise:

$$\phi(B)X_t=w_t$$

We fit the model with an order of 16 and obtained a MAPE value of 22.728% for Rajasthan. Low p-values were found in the computed parameters, indicating a satisfactory match. The residual ACF plot also revealed mostly insignificant values, implying that the residuals are white noise.

MA Model:

This model is essentially a regression model based on previous residuals/mistakes/noise, and it forecasts the new value using moving averages of previous errors. The following statement can be used to represent it: where θ is the moving average operator and is the value of the variable (polynomial of order q).

$$X_t=\theta(B)w_t$$

Due to computing resource constraints, we fitted the model with an order of 20 and obtained a MAPE value of 23.424 percent for Rajasthan. Low p-values were found in the computed parameters, indicating a satisfactory match.

The residual ACF plot, on the other hand, revealed some statistically significant values, showing that this model is significantly worse than the AR model in terms of fit.

ARMA Model:

For a stationary time series, this model essentially combines the previous two models, doing regression like fitting both ways. It can be expressed using the notation described previously:

$$\phi(B)X_t=\theta(B)w_t$$

We fitted the model with order (9, 10) and got a MAPE of 22.874% for Rajasthan and low p-values for the feature weights, indicating a strong match. With only a few significant values, the residual ACF plot was also the best of all the models. This approach would be the most appropriate for daily data (although still not very good as we can see from the feature weight p-values and the MAPE score).

ARIMA Model:

This is an ARMA model modification that also works with non-stationary time series data. It uses differencing between present and historical values (up to a specific order d) as the new values for the ARMA model to transform time series data to stationary. It's written like this:

$$\phi(B)(1-B)^d X_t=\theta(B)w_t$$

We fitted the model with order (9, 1, 10) and achieved a MAPE of 23.551 percent for Rajasthan, as well as low p-values for the feature weights, indicating a strong fit – these, however, were more significant than the ARMA model. The residual ACF plot also had very few significant values, indicating that the model's white noise assumption was correct. Because this model did not outperform the ARMA model, the time series was confirmed to be stationary. However, due to computational resource constraints, a true comparison would require the use of the optimal ordering as determined by the ACF and PACF plots before, which could not be done.

SARIMA Model:

After differencing the data with a latency equal to the seasonality of the data, this model uses different regression-like models (approximated as yearly for our case). (p, d, q) is the ARIMA order, (P, D, Q) is the seasonal order, and m is the seasonality. Model expression:

$$\phi(B_m)\phi(B)\ (1-B_m)^D\ (1-B)^d\ X_t=\theta(B_m)\theta(B)w_t$$

Because the seasonality of 365/52 was too computationally expensive for grid search, we fit the model using monthly data rather than daily/weekly data. Later, using a more simple pre-determined SARIMA model, we would do a weekly analysis. For monthly data, we discovered the following parameters: ([2, 0, 3], [2, 0, 1], 12). With a MAPE of only 10.714 percent for Rajasthan and good results for both the residuals and residual ACF plot, fitting this model provided us incredibly good results.

Below are the MAPE values obtained for forecasting using different models and for different states:

State/Model	AR	MA	ARMA	ARIMA	SARIMA
AP	15.973	16.629	15.997	16.090	8.234
MP	23.592	23.946	23.676	23.633	10.210
RJ	22.728	23.424	22.874	23.551	10.714
TN	16.692	17.718	16.823	16.607	8.516

Forecast Validation

The dataset was split into test and training data and then performed rolling forecasting to validate our model. The models were computationally expensive, since computation time was an issue the splits were done based on that.

Train : Test Splits	AR, MA, ARMA, ARIMA	SARIMA
Daily data	90:10	-
Weekly data	90:10	95:5
Monthly data	-	70:30

Therefore, we have found the optimum parameters for the smaller space. We used MSE (Mean Squared Error) and MAPE (Mean Absolute Percent Error) as metrics. The forecast results for each Rajasthan wind speed data model are as follows.

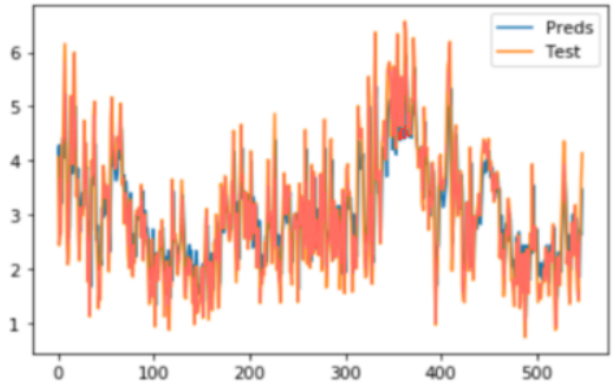
AR Model:

There were 9 parameters in daily and 10 in weekly data.

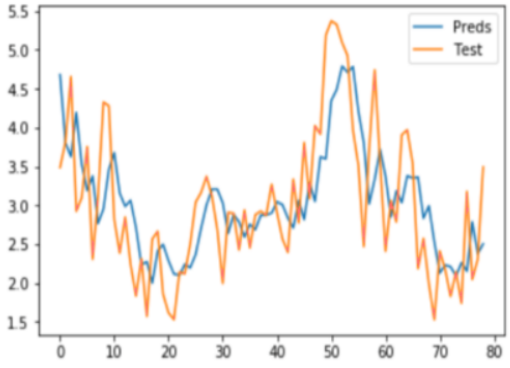
Metrics: Daily -> MSE- 0.5162; MAPE- 22.729%

Weekly -> MSE-0.479 ; MAPE-20.298%

```
Mean Absolute Percentage Error (Rajasthan): 22.72879148073194 %
Mean Square Error (Rajasthan): 0.516221049117893
```




```
Mean Absolute Percentage Error (Rajasthan): 20.298803690700883
Mean Square Error (Rajasthan): 0.4789770582422554
```



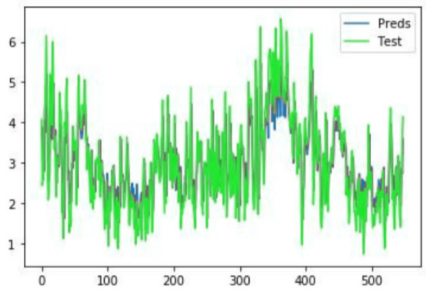
MA Model:

There were 11 parameters in both daily and weekly data.

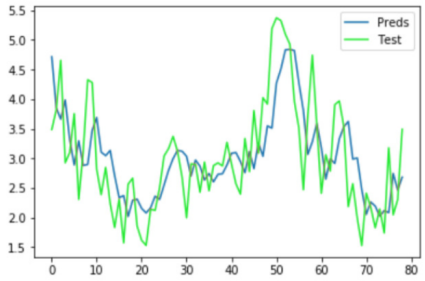
Metrics: Daily -> MSE- 0.545;MAPE- 23.425%

Weekly -> MSE-0.485 ; MAPE-20.199%

```
Mean Absolute Percentage Error (Rajasthan): 23.42486830037479 %
Mean Square Error (Rajasthan): 0.5450029090314916
```




```
Mean Absolute Percentage Error (Rajasthan): 20.19860488128218 %
Mean Square Error (Rajasthan): 0.4852124811023596
```



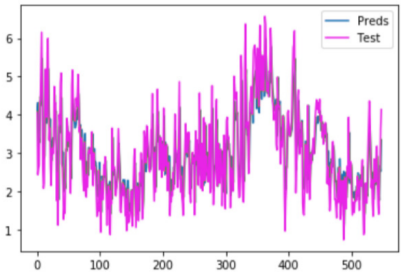
ARMA Model:

There were 2,3 parameters in daily and 2,4 in weekly data.

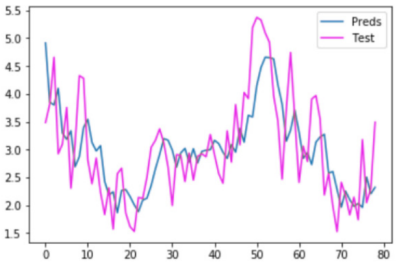
Metrics: Daily -> MSE- 0.519;MAPE- 22.874%

Weekly -> MSE-0.4179; MAPE-19.495%

```
Mean Absolute Percentage Error (Rajasthan): 22.87413662095137 %
Mean Square Error (Rajasthan): 0.5191950322659858
```




```
Mean Absolute Percentage Error (Rajasthan): 19.495407904757872 %
Mean Square Error (Rajasthan): 0.4719975909768209
```



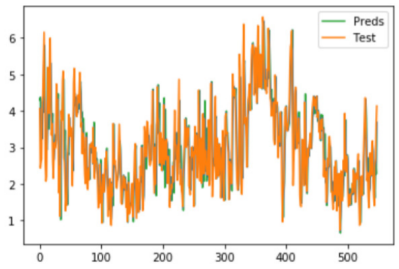
ARIMA Model:

There were 2,2,3 parameters in daily and 2,2,4 in weekly data.

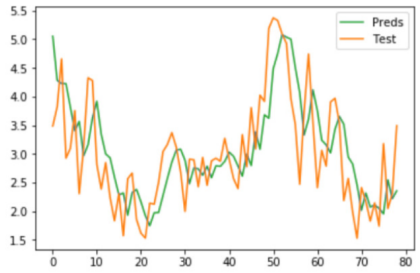
Metrics: Daily -> MSE- 0.603;MAPE- 23.551%

Weekly -> MSE-0.505; MAPE-20.749%

```
Mean Absolute Percentage Error (Rajasthan): 23.551036184300177 %
Mean Square Error (Rajasthan): 0.6031709469152163
```




```
Mean Absolute Percentage Error (Rajasthan): 20.749269432561693 %
Mean Square Error (Rajasthan): 0.505006280939738
```



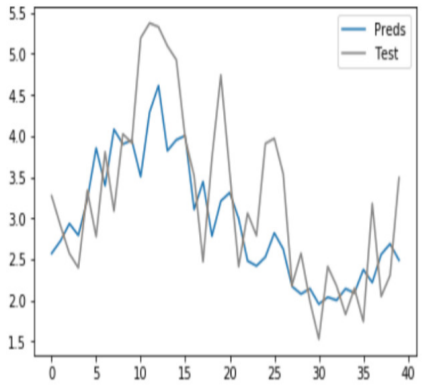
SARIMA Model:

There were ([1,0,1],[1,0,1],52) parameters in weekly and ([2,0,3],[2,0,1],12) in monthly data.

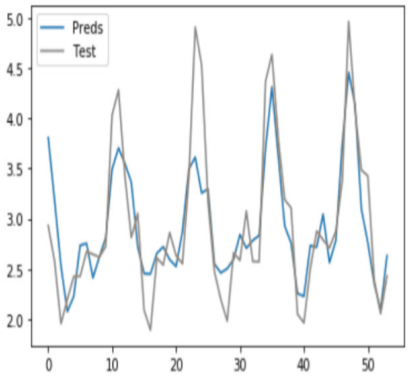
Metrics: Weekly -> MSE- 0.577;MAPE- 18.959%

Monthly -> MSE-0.175; MAPE-10.714%

```
Mean Absolute Percentage Error (Rajasthan): 18.958795025742248 %
Mean Square Error (Rajasthan): 0.577061044379608
```




```
Mean Absolute Percentage Error (Rajasthan): 10.714485781394401 %
Mean Square Error (Rajasthan): 0.17517053075415187
```



Conclusion

The best model accuracies that we found are given below:

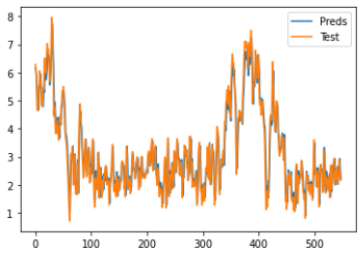
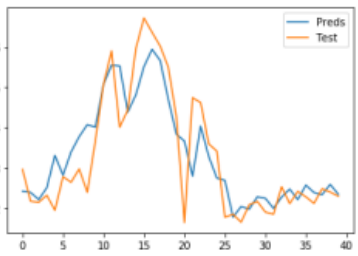
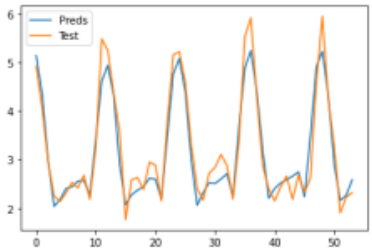
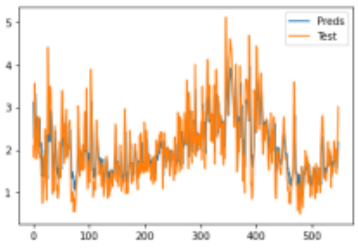
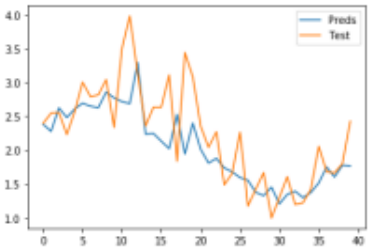
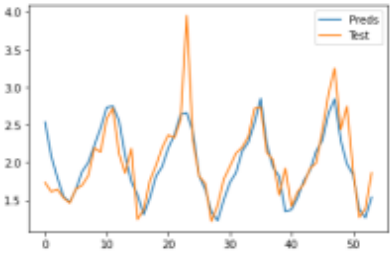
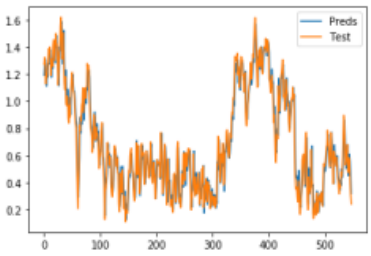
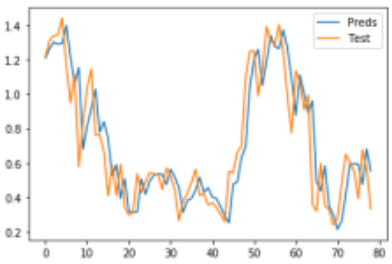
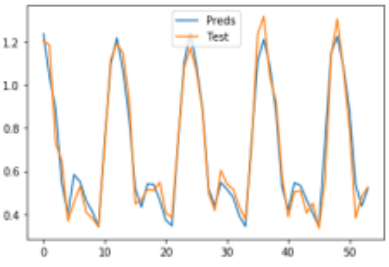
State/Period	Daily	Weekly	Monthly
Andhra Pradesh	84.02%	80.89%	91.77%
Madhya Pradesh	76.41%	84.62%	89.79%
Rajasthan	77.27%	81.05%	89.29%
Tamil Nadu	83.64%	76.06%	84.70%

Following are the conclusions drawn after performing this analysis on each state:

- For daily data, ARMA > ARIMA > AR > MA
- For weekly data, SARIMA > ARMA > ARIMA > AR > MA
- For monthly data, the forecasting model which gave us the best results for the data was SARIMA since the data has some kind of seasonality associated with it. This model was seen to be an exceptionally good fit for all states.

Due to high computational power requirements, we forecasted using simpler models. More complex models would have resulted in better accuracy. So if we observe the above table, the discrepancy in the weekly data for Andhra Pradesh and Tamil Nadu (daily data is more accurate than weekly data which is generally not the case in other observations) can be because of this computation power constraint. The calculated confidence in the models gave a few significant p-values for estimated feature weights. Thus models weren't great fits. However, they provide reasonably good accuracies considering the simplicity of models and computational resources they take.

Best Forecasts

Daily	Weekly	Monthly
		
Andhra Pradesh		
		
Madhya Pradesh		
		
Tamil Nadu		