A Report
On

Statistical Analysis and Forecasting of Wind Energy (Inter-State)

By (Group 4)

Names	**ID No.**
V. Aravindan	2017B4A70849P
Dhruv Patel	2017B4A70583P
Bhoomi Sawant	2017A7PS0001P
Abhinava Arasada	2017A7PS0028P
Mohit Kumar Jangir	2017B3PS1217P
Abdul Kadir Khimani	2017B4A70696P

MATH F432
Applied Statistical Methods

Table of Contents

Introduction

Renewable energy is produced from the sources that are constantly being replenished, such as sunlight, water, and wind. This means that we can use them without having to worry about them running out. Moreover, renewable energy sources are much more environmentally friendly than fossil fuels because they release very few chemicals, like carbon dioxide, that can potentially harm the environment.

Wind power is one of the most efficient alternative energy sources. There has been a significant development in wind turbine technology over the last decade with many new companies entering this sector. Wind turbines have become larger, efficiency and availability have improved and wind farm concepts have become popular. The economics of wind energy is already strong, despite the relative immaturity of the industry. The downward trend in wind energy costs is predicted to continue. As the world market in wind turbines continues to boom, wind turbine prices will continue to fall. India now ranks as a "wind superpower" having a net potential of about 45000 MW only from 13 identified states.

We intend to identify the energy producing potential for four Indian states(Andhra Pradesh, Tamil Nadu, Rajasthan, Madhya Pradesh) by doing a time-series analysis of wind speed by evaluating various models on monthly and daily data.

1. Understanding various terminologies

1. **Diffuse Horizontal Irradiance (DHI)** is termed as amount of radiation received per unit area by a surface (not subjected to any kind of shadow) that does not arrive directly from the sun, but has been scattered by small obstructions(molecules and particles) in the atmosphere. Therefore , from the understanding of the above definition, this variable would not be useful in analysis of wind energy.
2. **Direct Normal Irradiance (DNI)** refers to the amount of solar radiation that a surface receives per unit area that is always held perpendicular to the incoming straight line rays from the direction of the sun at its current position in the sky. This is also a term related to solar energy and thus not useful for wind energy forecast.
3. **Global horizontal irradiance** is also a term related to solar energy.GHI is the total amount of shortwave radiation received from above by a surface horizontal to the ground. This value includes both DNI and DHI. It is given as **GHI = DNI*cos(Θ) + DHI**
4. **Dew point** is the atmospheric temperature below which water droplets begin to condense and formation of dew starts. Dew point data doesn’t correlate with wind speed data and thus dew point data is also not of significance for wind energy forecasting.

5. **Temperature** is the measure of hotness or coldness. This parameter is directly influenced by incoming solar radiations.Temperature physically influences pressure and thus wind speed, which is evident from the heat map generated from the data.
6. **Pressure** is an important parameter which affects wind speed directly.Technically it is defined as force exerted by air on the surface per unit area. Wind flows from high pressure to low pressure.Higher the pressure gradient, higher is the wind speed(e.g. coastal areas). Also, development of pressure gradient depends on the temperature as discussed above. Higher temperature in an area,lower will be the pressure.
7. **Relative Humidity** is defined as the amount of water vapour present in air expressed as a percentage of the amount needed for saturation at the same temperature. More is the relative humidity, the air will become heavier and hence wind speed will decrease. Thus, wind energy output will decrease. However correlation between relative humidity and wind speed was very weak to conclude any correlation from the heat map.
8. **Wind Speed** is simply the rate at which air is moving in a particular area.Clearly wind speed is directly related to wind energy output. Higher the wind speed, higher is the wind energy output. Different states have different wind speed distributions over the year. Coastal area enjoys maximum wind energy output.

2. Data Preprocessing

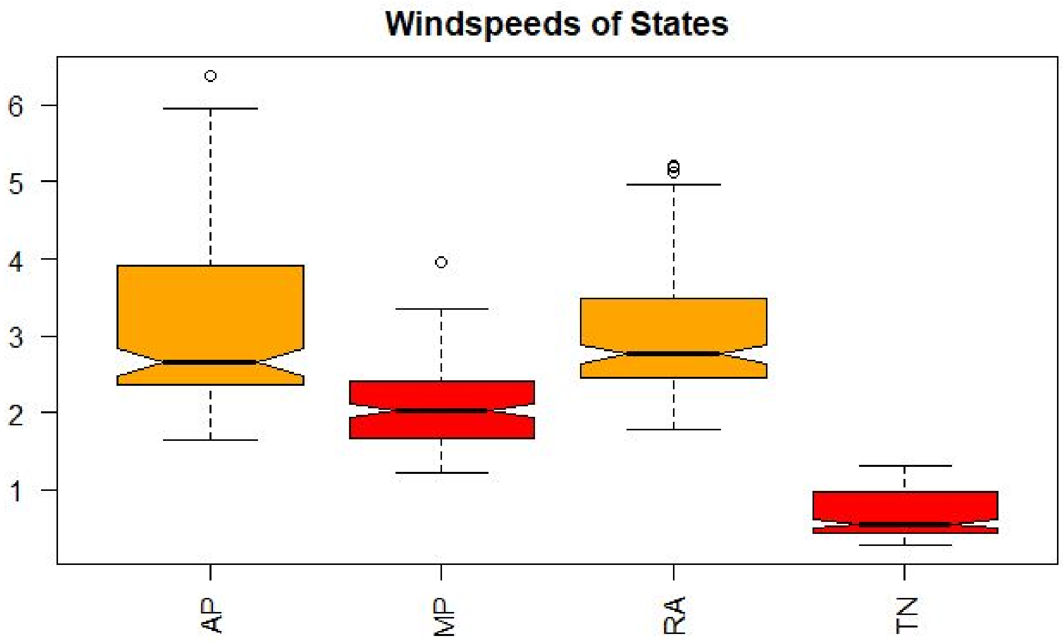
Our goal is to perform time-series analysis on wind speed data. As seen from the correlation heatmap, most of the other variables are not strongly correlated to wind speed. We clean our data by combining the yearly files from 2000-2014 for each state into a single file. We also remove all the other variables and keep only four variables of interest for each state: Year, Month, Hour, Wind speed. We perform further data reduction by converting the hourly data to :

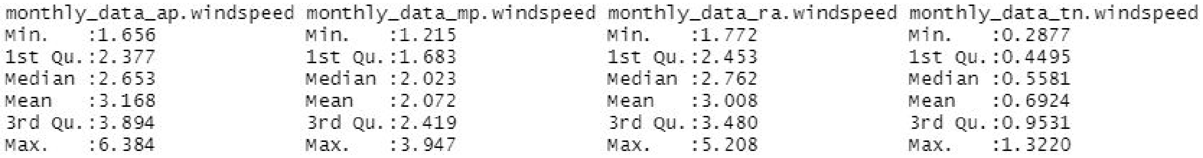
1. Daily data(averaging for 24 hours).
2. Monthly data(averaging for 30 days).

We shall now examine this data for each state individually and perform hourly and monthly time series analysis.

We have used the State Codes for naming in figures: Rajasthan(RA), Tamil Nadu(TN), Andhra Pradesh(AP), Madhya Pradesh(MP).

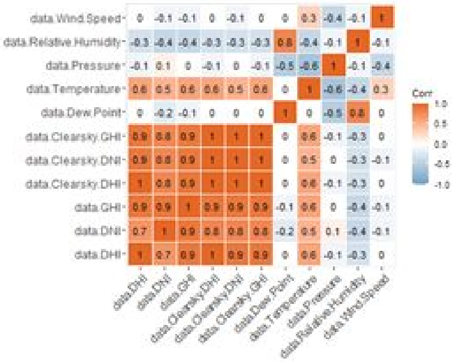
3. Descriptive Statistics

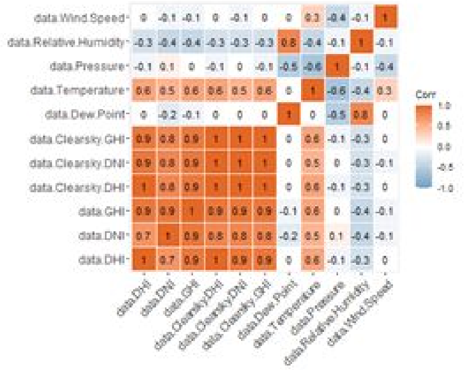


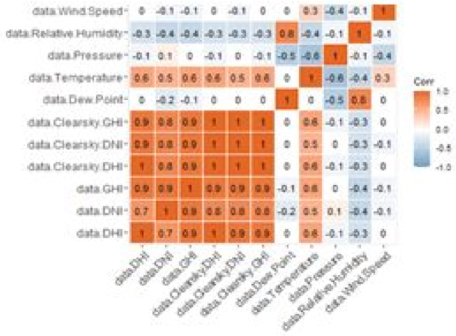


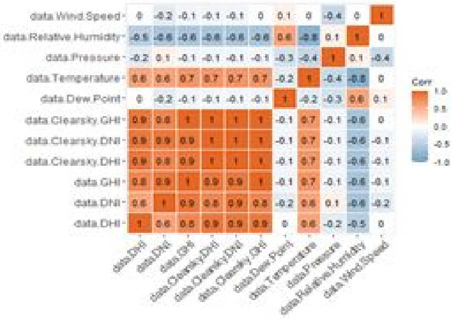
```
monthly_data_ap.windspeed monthly_data_mp.windspeed monthly_data_ra.windspeed monthly_data_tn.windspeed
Min.   :1.656             Min.   :1.215             Min.   :1.772             Min.   :0.2877           
1st Qu.:2.377             1st Qu.:1.683             1st Qu.:2.453             1st Qu.:0.4495           
Median :2.653             Median :2.023             Median :2.762             Median :0.5581           
Mean   :3.168             Mean   :2.072             Mean   :3.008             Mean   :0.6924           
3rd Qu.:3.894             3rd Qu.:2.419             3rd Qu.:3.480             3rd Qu.:0.9531           
Max.   :6.384             Max.   :3.947             Max.   :5.208             Max.   :1.3220           
```

It can be seen that Andhra Pradesh has the highest mean wind speed, followed by Rajasthan, Madhya Pradesh and Tamil Nadu. Outliers can be observed in Andhra Pradesh, Madhya Pradesh and Rajasthan, with Rajasthan showing a higher quantity. The wind speed seems the strongest in Andhra Pradesh and Rajasthan and we would expect it to have larger amounts of wind energy production, than Madhya Pradesh and Tamil Nadu.







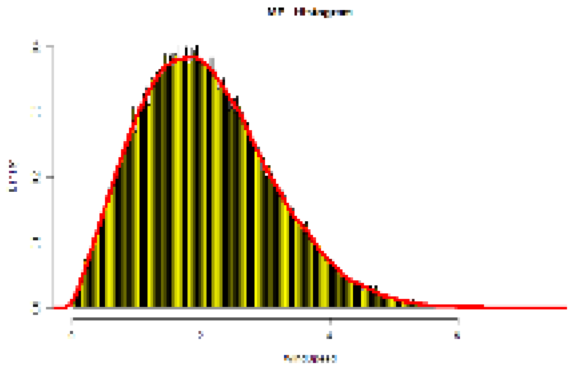


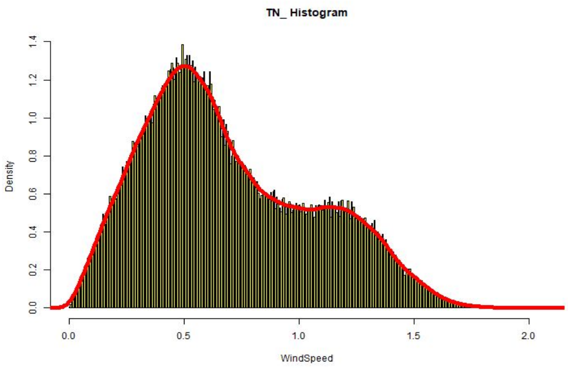
It can be observed that wind speed has a positive correlation of 0.3 with Temperature in states except Andhra Pradesh and a negative correlation of -0.4 with Pressure in all the four states. **It is also observed that there was 0 correlation between wind speed and temperature in Andhra Pradesh**. There also seems to be a mildly weak negative correlation(-0.1) clearskyDNI, DNI, GHI and relative humidity. Andhra Pradesh also shows a weak negative correlation(-0.1) in clearskyDHI.

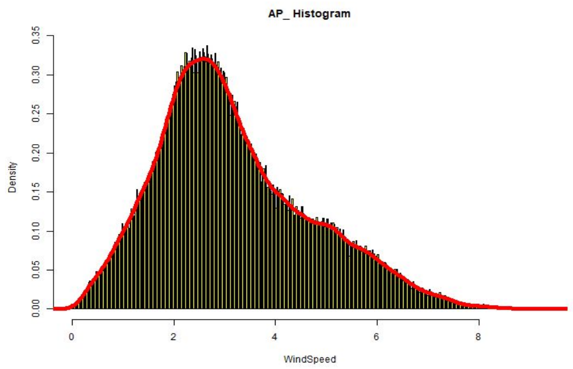
Before we jump to Time Series Analysis, we shall first examine the wind speed plots, distributions, and deviation from normality(if any) for each state.

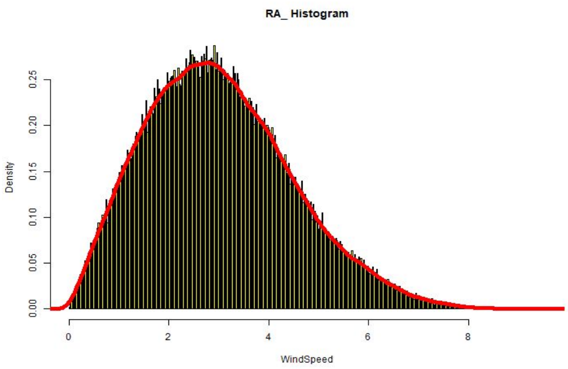
4. Histogram plots

We plot and examine the histogram plot for all the states. Histogram visual inspection reveals a slight skew towards the right. The distribution also differs a bit across all states. Following figures show a histogram for each state.







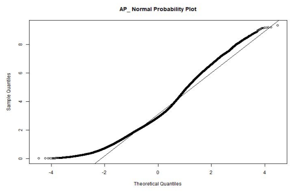


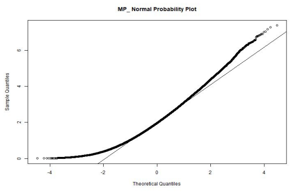
5. Checking for Normality

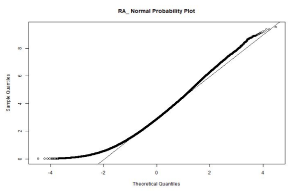
For checking Normality for each state, we perform a one-sample Kolmogorov-Smirnov(KS) test. We also plot the normal probability plots for visual inspection.

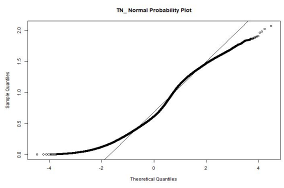
5.1 Normal Probability Plots(QQ plots):

It is the inverse of the standard normal cumulative versus the ordered observations. For a normal distribution of the data, the points in the QQ plot fall along a straight line. Deviations from this line indicate various types of non-normality. Stragglers at either end of the QQ-plot point to outliers. The curvature at both ends of the plot is an indicator of long or short distribution tails. Convex, concave, or a curvature indicates a lack of symmetry. Gaps, plateaus, or segmentation in the plot indicate phenomena that need further scrutiny.









Each state plot shows deviation from the straight-line giving an indication that data may not be normal. We shall perform statistical tests to confirm this.

5.2 Kolmogorov-Smirnov(KS) Test:

The Kolmogorov-Smirnov Goodness of Fit Test (K-S test) compares the given data with a known distribution and lets one know if they have the same distribution. Even though the test is nonparametric - it doesn't assume any underlying distribution - it is commonly used as a test for normality on the given data. It's also used to check the assumption of normality in the Analysis of Variance. In simple words, the test compares a specified hypothetical probability distribution (e.g. the normal distribution) to the distribution generated by given data — the empirical distribution function(EDF). (Glen)

The hypothesis for the test are:

- Null hypothesis (H_0): Data comes from the specified distribution ($P = P_0$).
- Alternate Hypothesis (H_1): At least one value does not match the specified distribution (Data doesn't come from a specified distribution)($P \neq P_0$).

P is the distribution of your sample (i.e. the EDF) and P_0 is a specified distribution. For our purposes, we specified distribution is Normal.

The K-S test statistic measures the largest distance between the EDF $F_{data}(x)$ and the theoretical function $F_0(x)$, measured in a vertical direction.

The **test statistic** is given by:

$$D = \sup_x |F_0(x) - F_{data}(x)|$$

where (for a two-tailed test):

- $F_0(x)$ = The cdf of the hypothesized distribution.
- $F_{data}(x)$ = The empirical distribution function of given data.

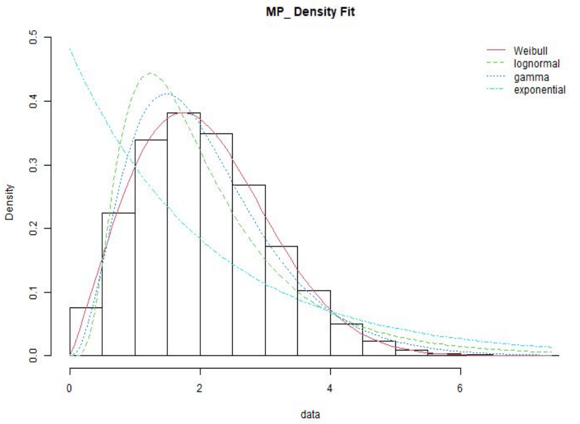
For a one-tailed test, one must remove the absolute values from the formula. If D is greater than the critical value, the null hypothesis is rejected. On the basis of the p-value, reject if p-value $\leq \alpha$.

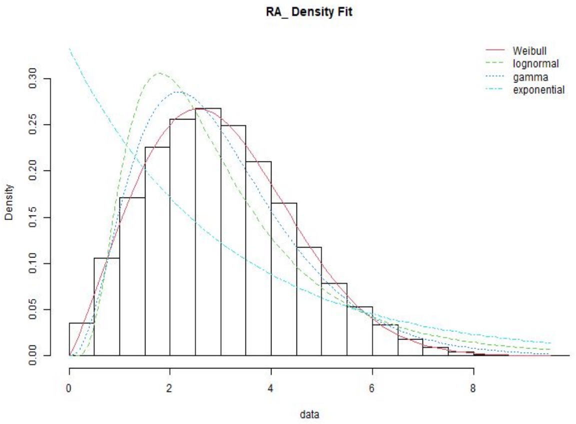
We perform the K-S test on our data. Results can be found in the **appendix**.

For all states, we can see that p-value < 0.05, hence we reject the null hypothesis. Thus, we conclude that all distributions are significantly different from Normal Distribution.

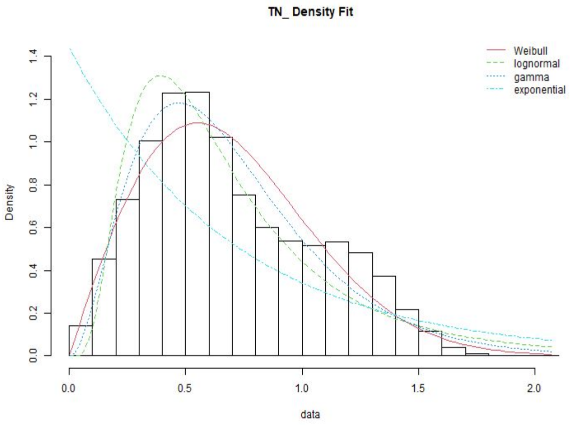
6. Distribution Fitting

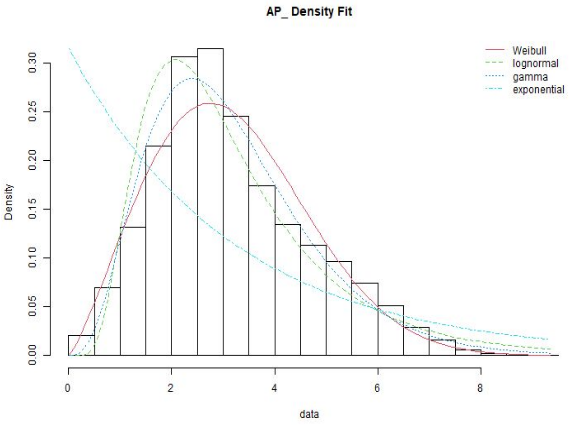
We try to fit different known distributions to the Wind Speed Dataset of each state. The Weibull and Lognormal functions are commonly used for fitting the measured wind speed probability distribution. (Chauhan and Saini)





For Rajasthan and Madhya Pradesh, it seems the Weibull Distribution fits well from visual inspection. It also has the least AIC scores of all distributions. But the KS Test fails for both. Results for the K-S Test and AIC Scores can be found in the **appendix**.





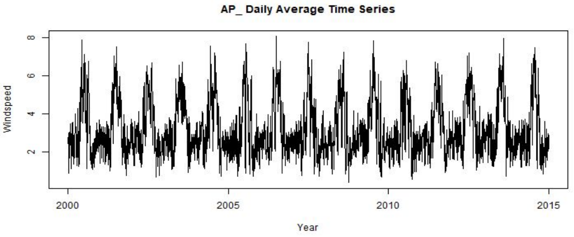
Similarly, for Tamil Nadu and Andhra Pradesh, it seems the Weibull Distribution fits well from visual inspection. It also has the least AIC scores of all distributions. But the KS Test fails for both. These results for the KS Test and AIC Scores can be found in the **appendix**.

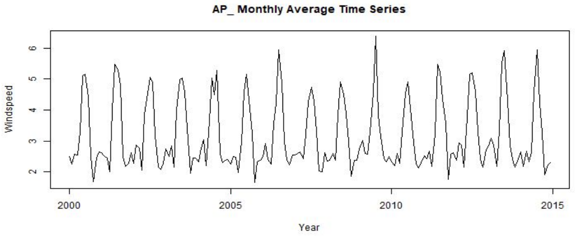
For large datasets like wind speed, we can't really be sure about the standard distributions that fit the data. Fitting distributions to such data is a complex process. Specifically, for Wind Speed data, research papers have tried to fit Weibull distributions and it generally works well, but not in our case. So we shall proceed towards a time series analysis of Wind Speed Data.

7. Time Series Analysis

We perform Time Series Analysis for each of the states using monthly and daily data. We start with Moving Average(MA) models and progressively move towards Seasonal Auto-Regressive Integrated Moving average(SARIMA) models. SARIMA is not applied to daily data because of software limitations.

7.1 Time Series Data Exploration





We plot the daily and monthly plots for each of the states. Visually, we can see that data exhibits clear seasonality and no trend. Seasonality is a type of non-stationarity. Before fitting a model to our dataset, we need a stationary time series. We shall first check for trend stationarity using statistical tests and then proceed to make the data stationary.

7.2 Stationarity Checks

We check for trend stationarity of our data. We can see a clear seasonal pattern, hence we perform statistical tests only for trend stationarity checking.

7.2.1 ADF Test

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not. It belongs to a category of test known as the ‘unit root’ test. Technically it is the augmented version of the DIckey Fuller test. The ADF test expands the Dickey-Fuller test equation to include high order regressive processes in the model:

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2}.. + \phi_p \Delta Y_{t-p} + e_t$$

The rest of the equation remains the same while only an extra differencing term is added, as it adds more thoroughness to the test. While the null hypothesis is still the same as the Dickey Fuller test. Here, in order to test the null hypothesis, which assumes the presence of unit root, the p-value should be less than the significance level which will reject the null hypothesis and would infer that the series is stationary.

In order to do the test in R adf.test function is used and then the data is passed on to perform the test. ADF test passes for all states showing that the data is trend stationary. Results can be found in the **appendix.**

7.2.2 KPSS Test

Like the ADF test, KPSS test is also commonly known to test the stationarity of the given time series data. The KPSS test, short for Kwiatkowski-Phillips-Schmidt-Shin (KPSS), is a type of Unit root test that tests for the stationarity of a given series around a deterministic trend. It is a common misconception that these tests can be used interchangeably with the ADF test.

The output of the KPSS test contain 4 main components:

1. The KPSS statistic
2. p-value
3. Number of lags used by the test
4. Critical values

Where, the number of lags reported is the number of lags of the series that was actually used by the model equation of the kpss test. Rest definition remains the same. A major difference between KPSS and ADF tests is the capability of the KPSS test to check for stationarity in the ‘presence of a deterministic trend’.

KPSS test passes for all states indicating level stationarity. Results can be found in the **appendix**.

7.3 Time Series Decomposition

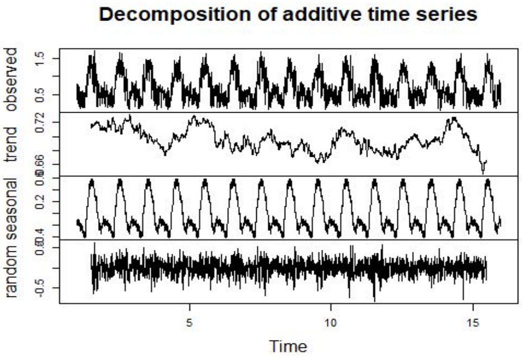
Time series decomposition is a method of breaking down a time series into sub patterns like trend, seasonality and noise components. Trend denotes the overall increasing or decreasing pattern. Seasonality component refers to the patterns that repeat over a fixed period of time. Random components are also called noise, irregular or remainder, that is, the residuals of the original time series after the seasonal and trend components are removed. The mathematical representation of this is:

$$Y_t = f(S_t, T_t, E_t)$$

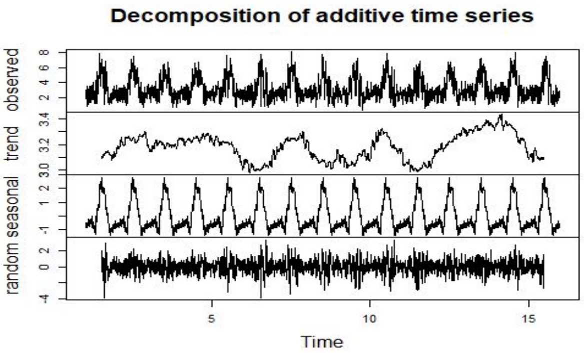
where Y_t is the actual data, that is the time series value at period t, S_t is seasonal, S_t Ts trend, E_t is irregular component at period t.

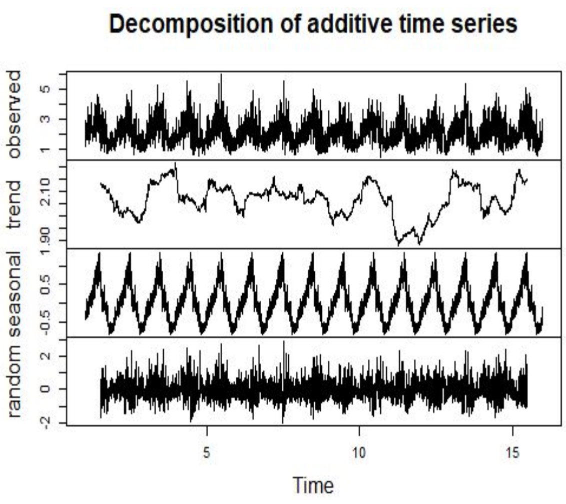
The decomposition can be done using additive or multiplicative models. Since there was no overall trend and the magnitude of seasonal variation was not varying much with the level of series, we used additive models. The additive model can be represented by $Y_t = S_t + T_t + E_t$ and the multiplicative model is given by $Y_t = S_t \text{ x } T_t \text{ x } E_t$. The results for additive decomposition for daily data for all four states is given the figures below. As we can see from the results, there is a clear seasonal pattern. We have used the “decompose” function of R to obtain the following plots.

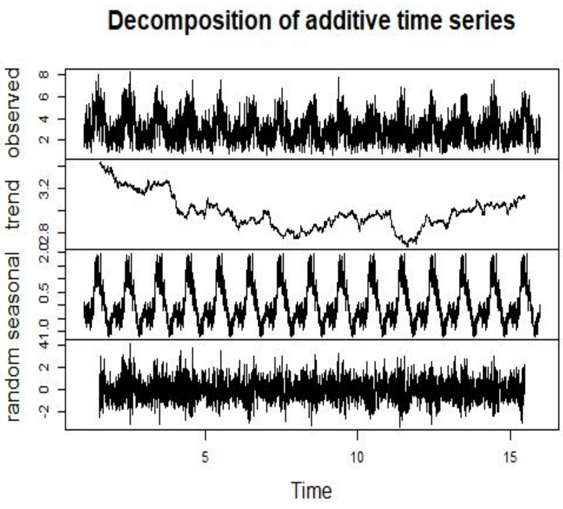
Andhra Pradesh



Madhya Pradesh





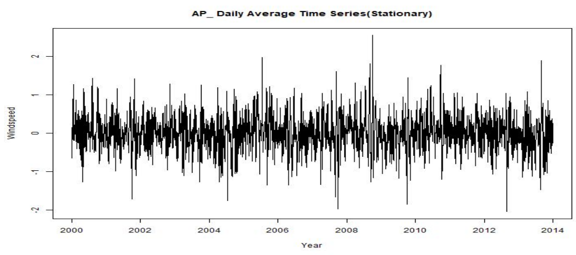


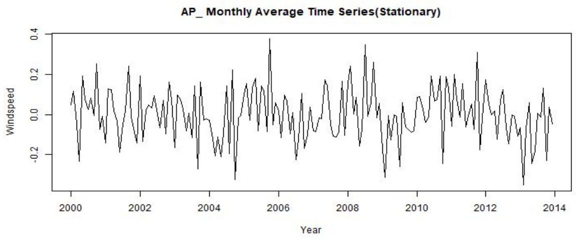
Rajasthan

Tamil Nadu

7.4 Stationary Data

We use a lag differencing of 12 to remove seasonal non-stationarity for monthly data and 365 for daily data. We do not need any first differencing as the data is trend stationary.





The data is now ready to be used for Time Series Models. As the data has strong seasonality, we do not expect AR, MA, ARMA, ARIMA to give good results. We expect SARIMA to work the best and give proper results. We shall experimentally move towards SARIMA.

Before we proceed to apply time series models, we need to understand a few terms. We shall briefly discuss them.

7.5 ACF and PACF

Autocorrelation is defined as the correlation of a variable with itself at different time lags. The sample auto-covariance function (ACVF), is defined as

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} \left(x_t - \bar{x} \right) \left(x_{t+k} - \bar{x} \right)$$

The sample autocorrelation function (ACF) is defined as

$$r_k = \frac{c_k}{c_0} = \mathrm{Cor}(x_t, x_{t+k})$$

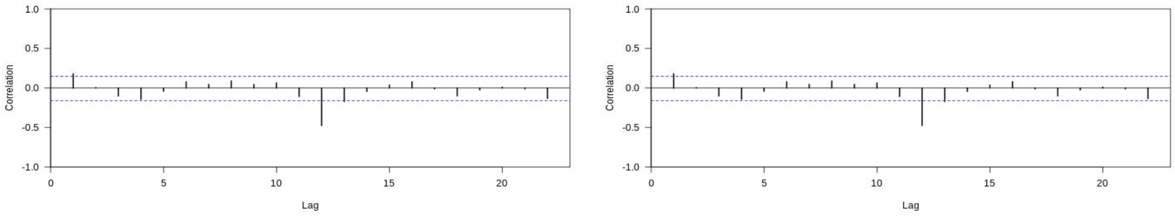
The ACF at lag 0, i.e. r_0, equals 1 by default. ACF for an MA(q) process abruptly cuts at 0 after lag > q.

Partial Autocorrelation function (PACF) measures the linear correlation of a series $\{x_t\}$ and a lagged version of itself $\{x_{t+k}\}$ with the linear dependence of $\{x_{t-1}, x_{t-2}, \ldots x_{t-(k-1)}\}$ removed. PACF is defined as

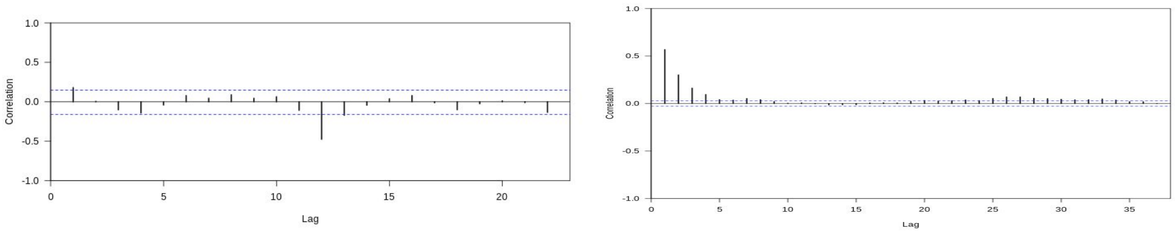
$$\begin{aligned} f_k &= Cor\left(x_1, x_0\right) = r_1 \quad , if\ k = 1 \\ &= Cor\left(x_k - x_k^{k-1}, x_0 - x_0^{k-1}\right) \ , if\ k > 1 \end{aligned}$$

ACF for an AR(p) process abruptly cuts at 0 after lag > p.

We plot the ACF and PACF plots for our stationary monthly and daily data. Below shown results are for the state of Madhya Pradesh. Other states' ACF and PACF plots can be found in the **appendix**.



ACF Monthly PACF



ACF Daily PACF

7.6 Time Series Models

7.6.1 Auto-Regressive(AR(p)) Model

In a general multiple regression model, we attempt to predict a single dependent variable of interest using a linear combination of predictors. In an AR model, we attempt to forecast the variable of interest using a linear combination of previously known values of itself. Hence, the term autoregression is used, indicating that it is a regression of the variable against itself.

An **autoregressive model of order p** can be written as

$$y_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \dots + \Phi_p y_{t-p} + \varepsilon_t$$

where ε_t is white noise which has a mean of 0 and is generally assumed to be a normal distribution of the form $\varepsilon_t \sim N(0, p)$. AR(p) indicates an autoregressive model of order p. (“Forecasting: Principles and Practice”)

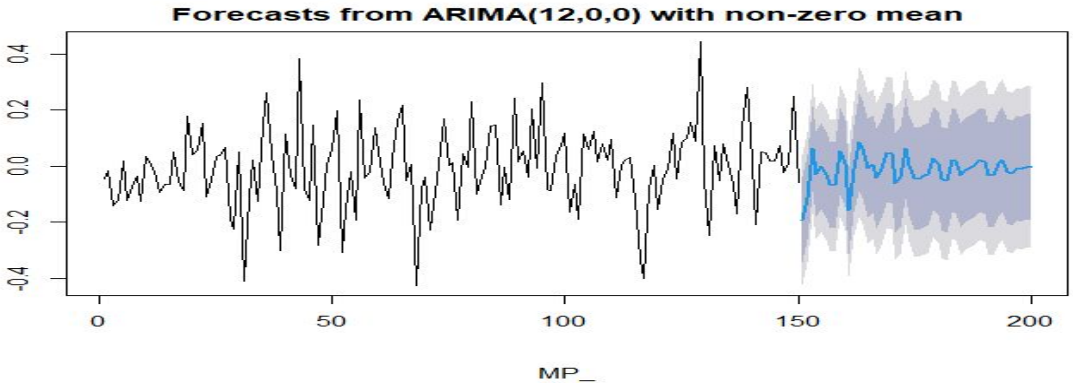
These models see a lot of usage in the field of Ecology as they are remarkably flexible at handling a wide range of different time-series patterns. Changing the parameters Φ_1, $\Phi_{2,}$ etc. will result in different time-series patterns. The variance of the error term ε_t will only change the scale of the series, not the patterns.

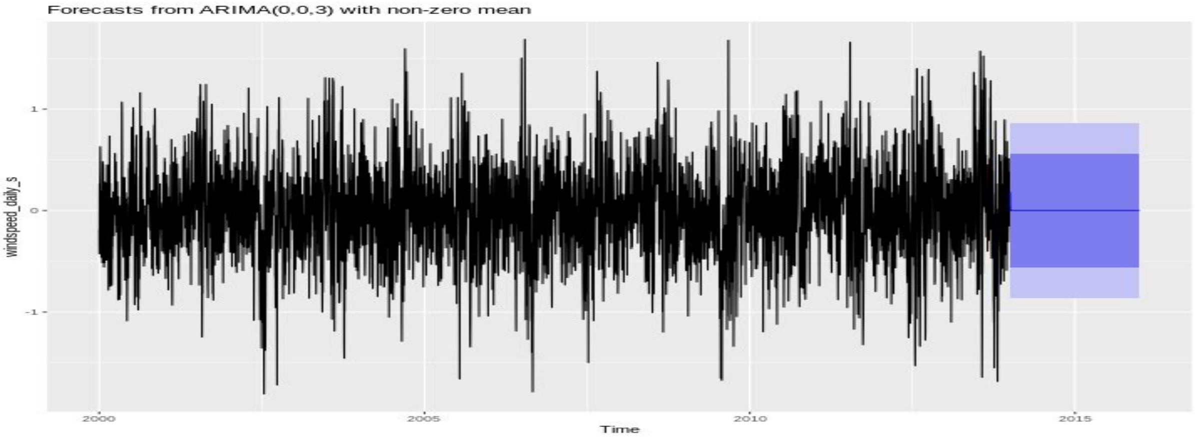
We generally use autoregressive models on stationary data, due to which some constraints on the values of the parameters are required.

- For an AR(1) model: $-1 < \Phi1 < 1$.
- For an AR(2) model: $-1 < \Phi_2 < 1$, $\Phi_1 + \Phi_2 < 1$, $\Phi_2 - \Phi_1 < 1$.

As p increases, the restrictions only continue to be more and more complicated. However, we need not concern ourselves with this complexity in R, as it takes care of these restrictions when estimating a model.

We consider p values like 1, 2, and 3 and select the best model on the basis of the least AICc scores for daily and monthly data. More information on the AICc score can be found in the **appendix**. We did not spend much time searching for p values as the model wouldn't work well on our seasonal data.





The daily forecast for Madhya Pradesh is a straight line and the monthly forecast gives some predictions which are not at all accurate. We do not calculate accuracy for this model as it is not useful at all. Forecasts for other states are also similar and can be found in the **appendix**.

7.6.2 Moving Average(MA(q)) Model

Together with the autoregressive (AR) model, the moving-average model is a special case and key component of the more general ARMA and ARIMA models of time series, which have a more complicated stochastic structure. Contrary to the AR model, the finite MA model is always stationary.

Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model. ("Applied Time Series Analysis for Fisheries and Environmental Sciences")

$$\mathbf{Y_t = c + \Phi_1\varepsilon_{t-1} + \Phi_2\varepsilon_{t-2} + \ldots + \Phi_q \varepsilon_{t-q} + \varepsilon_t}$$

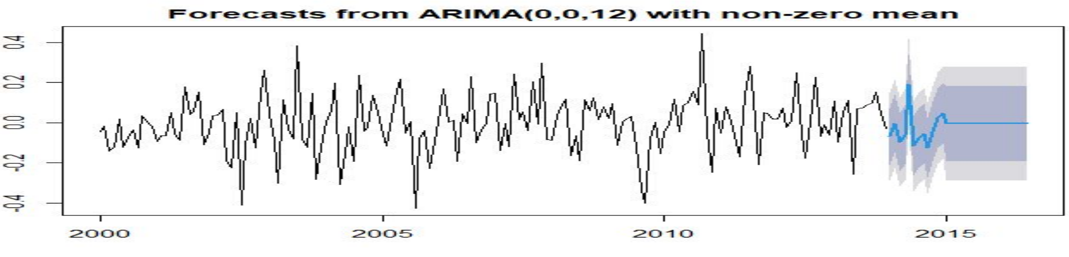
where ε_t and all the other error terms are white noise with a mean of 0 and is generally assumed to be a normal distribution of the form $\boldsymbol{\varepsilon_t \sim N(0, q)}$. All the random error terms are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with the mean at zero and a standard deviation decided by the parameter q.

We refer to this model as an **MA(q) model**, a moving average model of order p. Of course, we do not observe the values of $\boldsymbol{\varepsilon_t}$, so it is not really a regression in the usual sense.

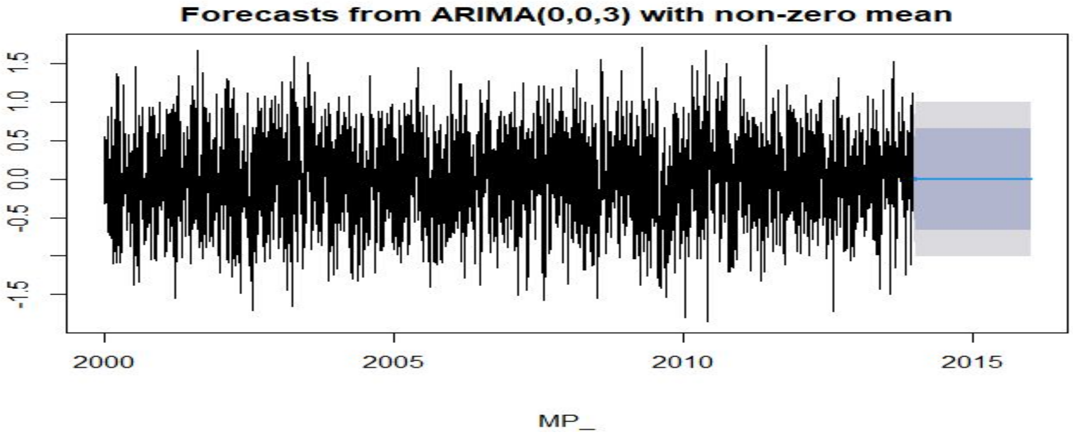
Quite similar to AR models, changing the parameters Φ_1, $\Phi_{2,}$ etc. will result in varying time-series patterns. Additionally, the variance of the error term ε_t will only change the scale of the series, not the patterns.

Each value of $\mathbf{Y_t}$ can be thought of as a weighted moving average of the past few forecast errors. However, moving average models should not be confused with the moving average smoothing. A moving average model is used for forecasting future values while moving average smoothing is used for estimating the trend-cycle of past values.

We consider q values as 1, 2, 3,.. 12, and select the best model on the basis of the least AICc scores for daily and monthly data. ACF and PACF plots show a high value at lag 12 hence we try to fit the MA(12) to our data. More information on AICc score can be found in the **appendix**. We did not spend much time searching for q values as the model wouldn't work well on our seasonal data. Given below are monthly and daily forecasts for Madhya Pradesh.



For monthly forecasts, there is some prediction and then the plot falls off to the straight line.



The daily forecast for Madhya Pradesh is a straight line. We do not calculate accuracy for this model as it is not useful at all. Forecasts for other states are also similar and can be found in the **appendix**.

7.6.3 Auto-Regressive Moving Average (ARMA(p,q)) Model

An autoregressive moving average model, ARMA comes from merging two models the autoregressive, AR and the moving average MA. ARMA is used to describe time series in terms of two polynomials, one for the autoregression and other for the moving average. ARMA is a stationary model. We have used lag differencing to make the time series stationary before applying ARMA. In the ARMA(p,q) model, p is the order of the autoregressive polynomial and q is the order of the moving average polynomial. The equation is given as:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

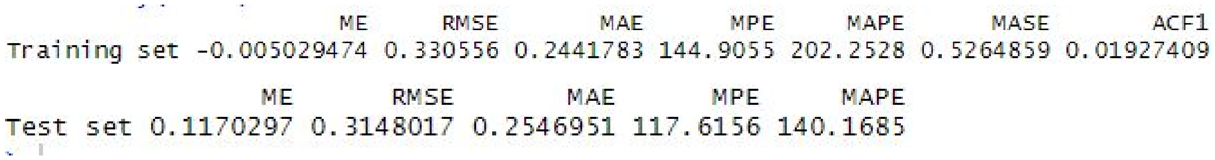
where φ's are the autoregressive model's parameters, θ's are the moving average model's parameters, ε's are the error terms (white noise) and c is a constant.

7.6.4 Auto-Regressive Integrated Moving Average (ARIMA(p,q,d)) Model

ARIMA stands for auto regressive integrated moving average. An ARIMA model is characterized by 3 parameters p, q and d. p is the order of the AR term, q is the order of the MA term and d is the number of non-seasonal differencing required to make the time series stationary.

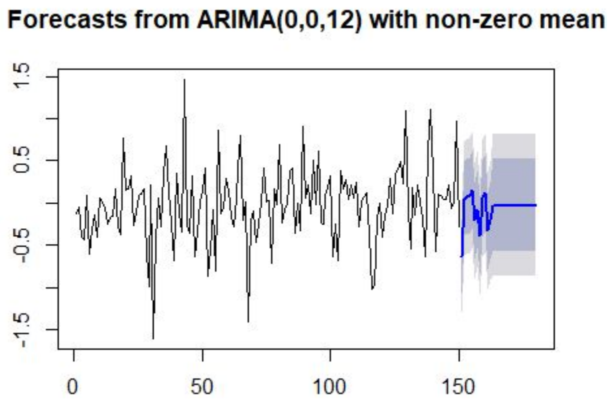
Since our data was stationary with respect to trend, d parameter was zero and same results were obtained for ARMA and ARIMA. We tried different values of p and q and chose the parameter values based on the lowest AIC values. The Akaike Information Criteria (AIC) is a measure used to compare statistical models. It is a statistic that incorporates both goodness of fit and complexity of the model. A lower AIC model is considered better. The results were obtained from the ARMA/ARIMA model for monthly data of four states. 150 months were used for training and 30 months for testing. The accuracy of model with respect to training and testing set, forecast plot and the plot showing actual and predicted values for last 30 months is shown in the figures below:

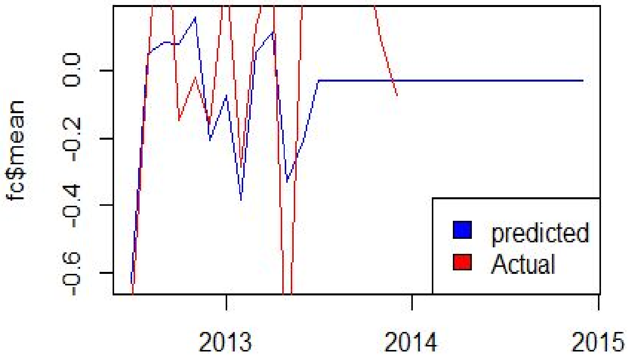
1. Rajasthan



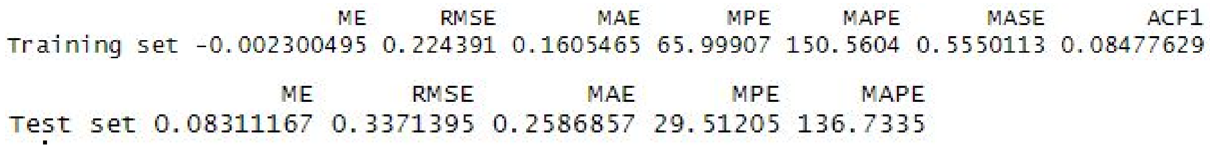
```
                      ME      RMSE       MAE      MPE     MAPE      MASE       ACF1
Training set -0.005029474 0.330556 0.2441783 144.9055 202.2528 0.5264859 0.01927409
```

```
                ME      RMSE       MAE      MPE     MAPE
Test set 0.1170297 0.3148017 0.2546951 117.6156 140.1685
```



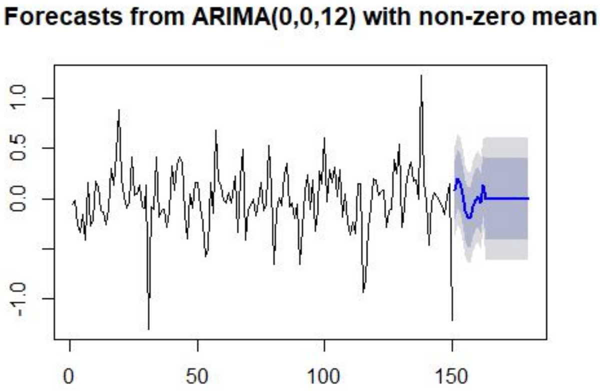


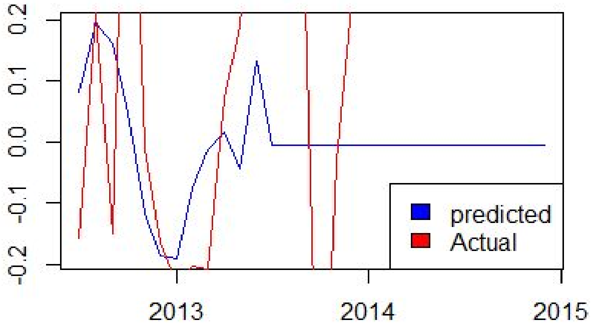
2. Madhya Pradesh



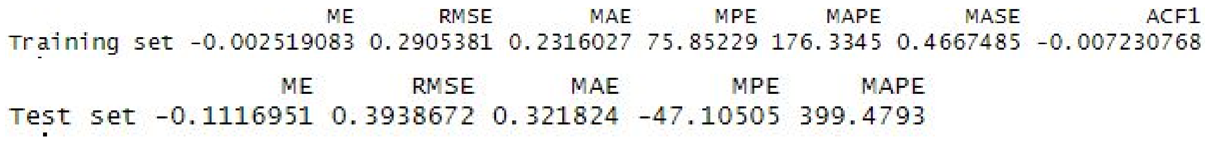
```
                      ME     RMSE       MAE      MPE     MAPE      MASE       ACF1
Training set -0.002300495 0.224391 0.1605465 65.99907 150.5604 0.5550113 0.08477629
```

```
                 ME      RMSE       MAE      MPE     MAPE
Test set 0.08311167 0.3371395 0.2586857 29.51205 136.7335
```



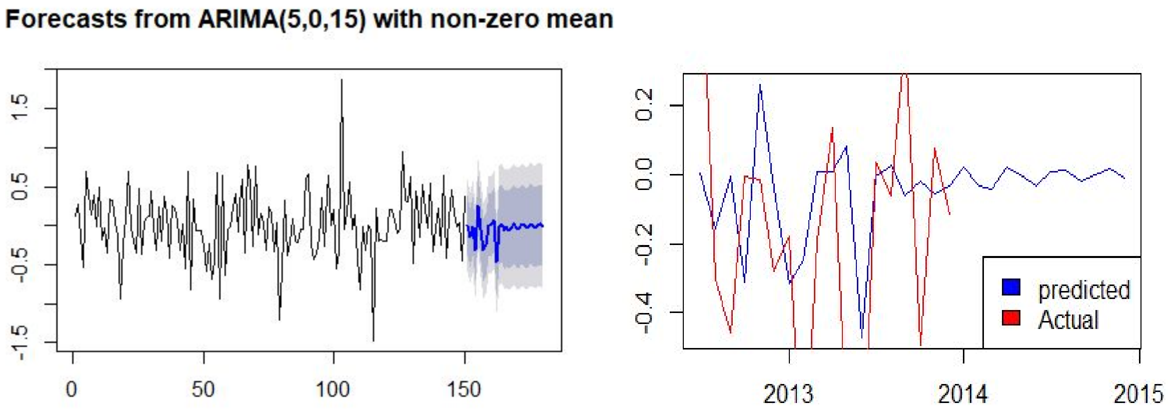


3. Andhra Pradesh

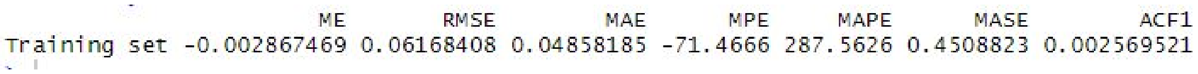


```
                      ME      RMSE       MAE      MPE     MAPE      MASE         ACF1
Training set -0.002519083 0.2905381 0.2316027 75.85229 176.3345 0.4667485 -0.007230768
```

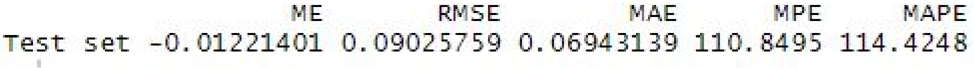
```
                 ME      RMSE      MAE       MPE     MAPE
Test set -0.1116951 0.3938672 0.321824 -47.10505 399.4793
```



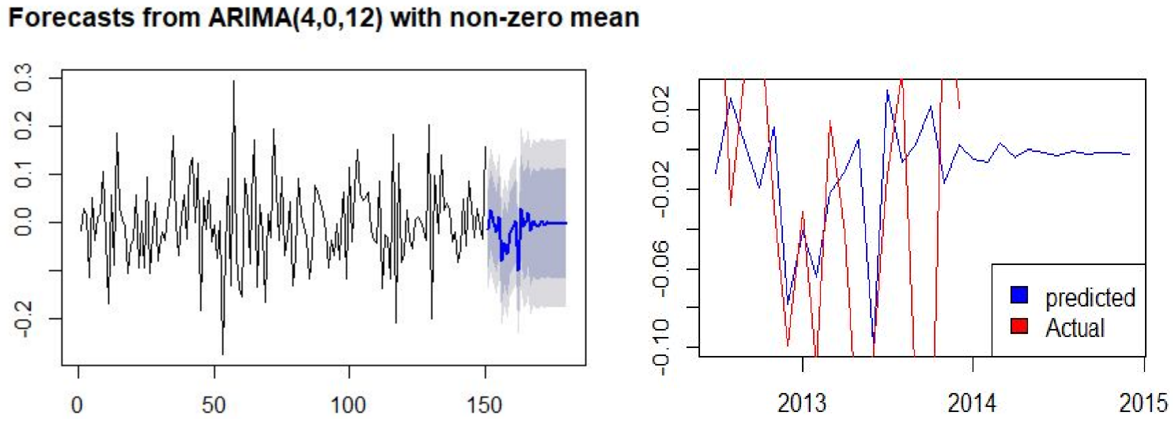
4. Tamil Nadu



```
                    ME       RMSE        MAE      MPE     MAPE      MASE        ACF1
Training set -0.002867469 0.06168408 0.04858185 -71.4666 287.5626 0.4508823 0.002569521
```



```
                 ME       RMSE        MAE      MPE     MAPE
Test set -0.01221401 0.09025759 0.06943139 110.8495 114.4248
```



7.6.5 Seasonal Auto-Regressive Integrated Moving Average (SARIMA) Model

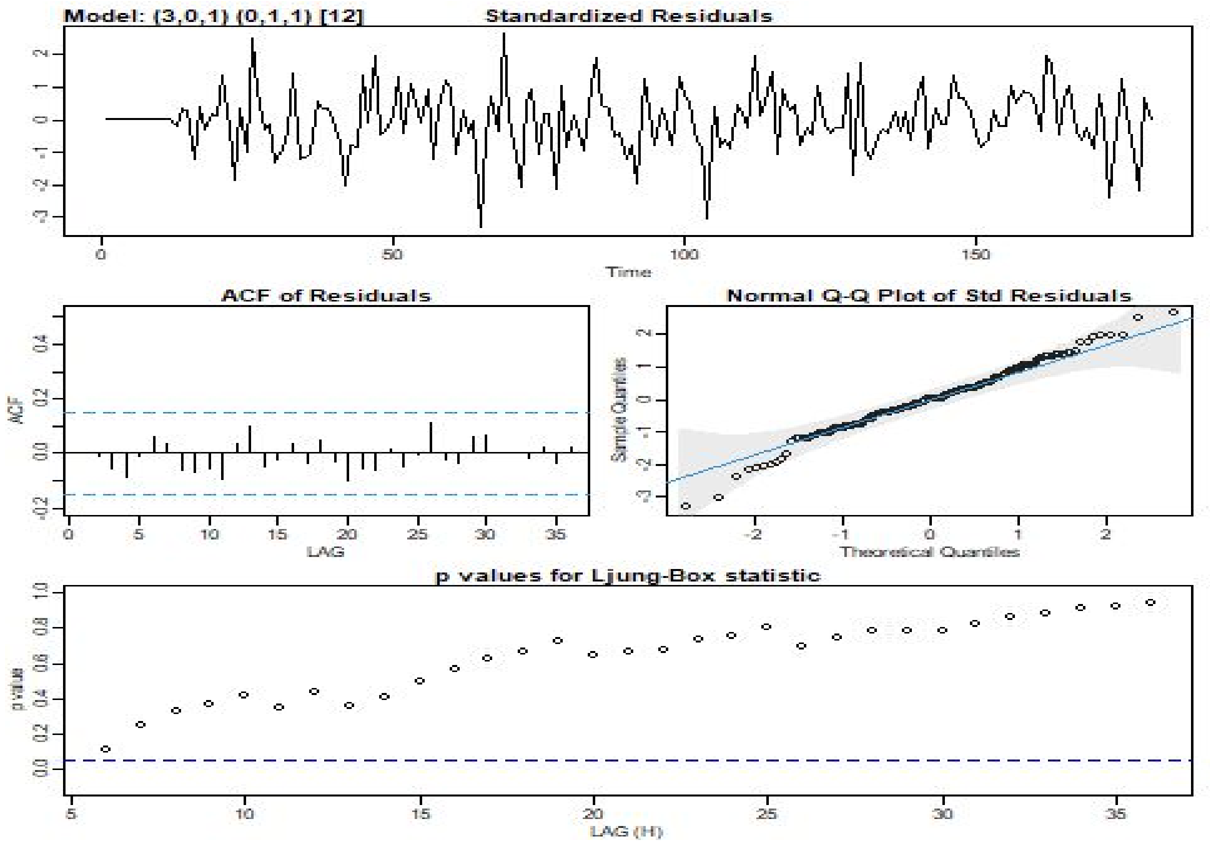
Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

It adds three new parameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

$SARIMA(p, d, q, P,D,Q)_S$ has two parts: Non-seasonal part (p, d, q) and seasonal parts $(P,D,Q)_S$.

1. p – order of non-seasonal AR terms
2. d – order of non-seasonal differencing
3. q – order of non-seasonal MA terms
4. P – order of seasonal AR (i.e., SAR) terms
5. D – order of seasonal differencing (i.e., power of $(1 - B\ S\ )$)

6. Q – order of seasonal MA (i.e., SMA) terms



Box-Ljung test is defined as:

H_0 :- The data are independently distributed.
H_a :- The data are not independently distributed; they exhibit serial correlation.

Test Statistic :- $Q = n(n+2)\sum_{k=1}^{m} \frac{\widehat{r_k^2}}{n-k}$

Where $\hat{r}_k$ is the estimated autocorrelation of the series at lag k, and m is the number of lags being tested.

Significance level : **α**

Critical region : The Box-Ljung test rejects the null hypothesis (indicating that the model has a significant lack of fit) if $Q > {}^{2}_{1-,h}$

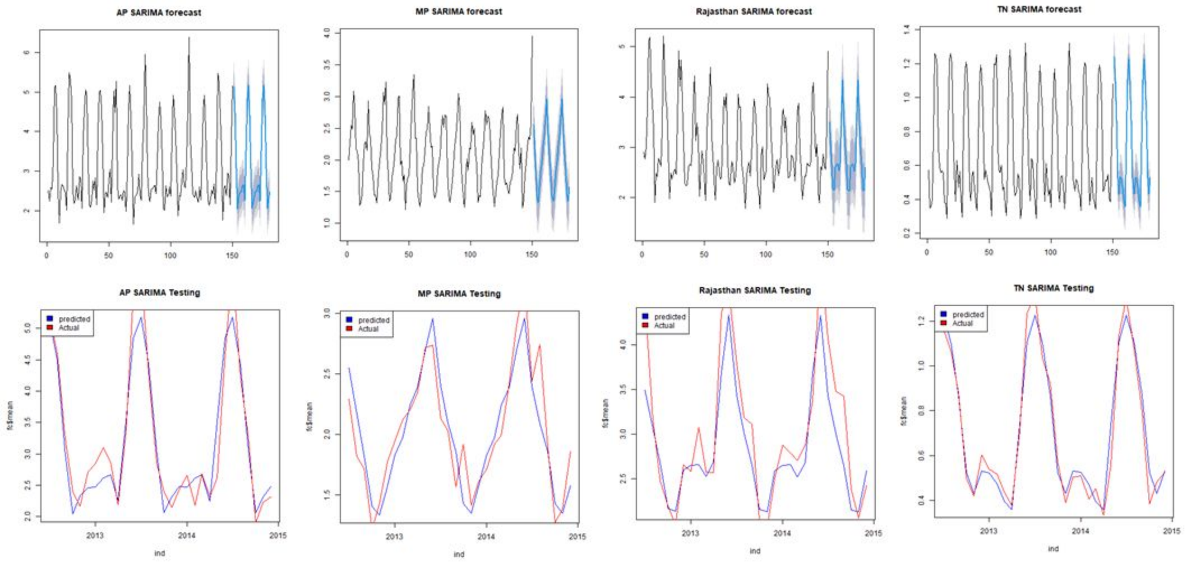
Where $Q > {}^{2}_{1-,h}$ is the chi-square distribution table value with h degrees of freedom and significance level **α**

Because the test is applied to residuals, the degrees of freedom must account for the estimated model parameters so that h=m−p−q, where p and q indicate the number of parameters from the ARMA(p,q) model fit to the data.

For a significance level of 0.05, we can observe that the p-values for all the lags seem to be greater than the significance level. Hence, we fail to reject the Null hypothesis and conclude that our model has no significant auto-correlation for all the lags.

Here, we present the results obtained for the SARIMA model, the model accuracy

Results for fitting a SARIMA (3,0,1,0,1,1) model with a period=12



State	Training	Testing
Andhra Pradesh	7.013385	8.622273
Madhya Pradesh	7.522623	8.625677
Tamil Nadu	8.266464	8.470952
Rajasthan	8.456506	8.583876

8. Conclusions

Out of all the model we used for time series analysis namely: AR, MA, ARMA, ARIMA and SARIMA, we observed that SARIMA was the best model which can model the data in the best possible way, this was apparent from the clear seasonality in the data and is affirmed by the excellent forecasting results obtained from out model. Now that the wind speed for all the states have been forecasted, we now decide its relationship with wind-energy as it is evident that higher wind speed would be directly proportional to higher wind-energy, we would advise more windmills and energy farms being set up in Rajasthan and Madhya Pradesh to harness more of this.

9. Bibliography

“Applied Time Series Analysis for Fisheries and Environmental Sciences.” *Applied Time Series Analysis for Fisheries and Environmental Sciences*, E. E. Holmes, M. D. Scheuerell, and E. J. Ward, 2020, https://nwfsc-timeseries.github.io/atsa-labs/. Accessed 07 11 2020.

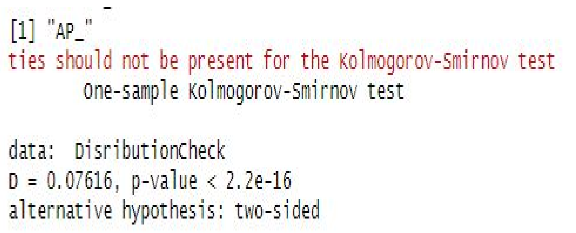
Chauhan, Anurag, and R.P. Saini. “Statistical Analysis of Wind Speed Data Using Weibull Distribution Parameters.” *Proceedings of 2014 1st International Conference on Non Conventional Energy (ICONCE 2014)*, vol. 1, no. 1, 2014.

“Forecasting: Principles and Practice.” *Forecasting: Principles and Practice*, Rob J Hyndman and George Athanasopoulos, 2018, https://otexts.com/fpp2/. Accessed 12 11 2020.

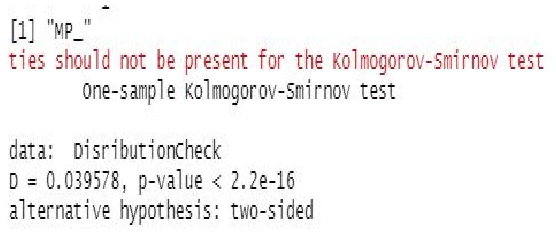
Glen, Stephanie. “Kolmogorov-Smirnov Goodness of Fit Test.” *Kolmogorov-Smirnov Goodness of Fit Test*, 2016, https://www.statisticshowto.com/kolmogorov-smirnov-test/. Accessed 20 11 2020.

Appendix

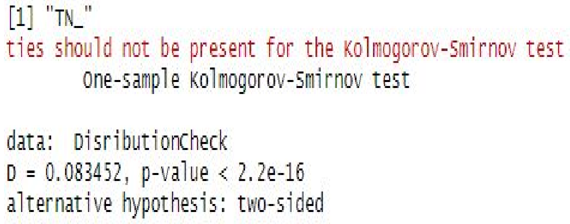
Results for KS Test(Checking for Normality)



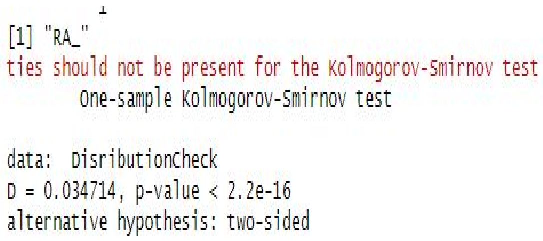
```
[1] "AP_"
ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  DisributionCheck
D = 0.07616, p-value < 2.2e-16
alternative hypothesis: two-sided
```



```
[1] "MP_"
ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  DisributionCheck
D = 0.039578, p-value < 2.2e-16
alternative hypothesis: two-sided
```

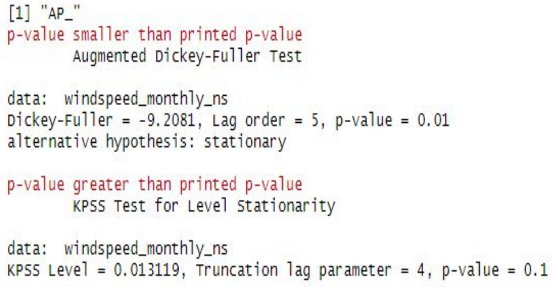


```
[1] "TN_"
ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  DisributionCheck
D = 0.083452, p-value < 2.2e-16
alternative hypothesis: two-sided
```

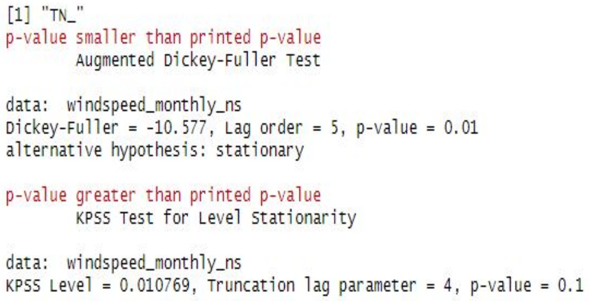


```
[1] "RA_"
ties should not be present for the Kolmogorov-Smirnov test
        One-sample Kolmogorov-Smirnov test

data:  DisributionCheck
D = 0.034714, p-value < 2.2e-16
alternative hypothesis: two-sided
```

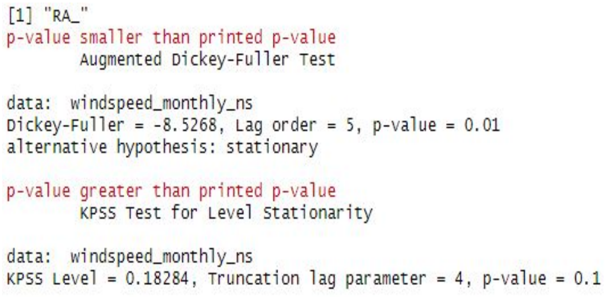
Results for ADF Test



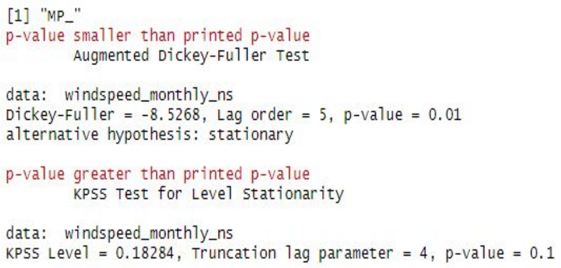
```
[1] "AP_"
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  windspeed_monthly_ns
Dickey-Fuller = -9.2081, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

p-value greater than printed p-value
        KPSS Test for Level Stationarity

data:  windspeed_monthly_ns
KPSS Level = 0.013119, Truncation lag parameter = 4, p-value = 0.1
```



```
[1] "TN_"
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  windspeed_monthly_ns
Dickey-Fuller = -10.577, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

p-value greater than printed p-value
        KPSS Test for Level Stationarity

data:  windspeed_monthly_ns
KPSS Level = 0.010769, Truncation lag parameter = 4, p-value = 0.1
```



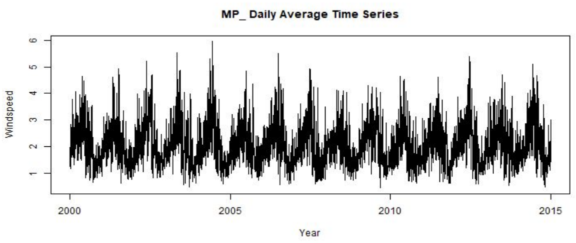
```
[1] "RA_"
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  windspeed_monthly_ns
Dickey-Fuller = -8.5268, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

p-value greater than printed p-value
        KPSS Test for Level Stationarity

data:  windspeed_monthly_ns
KPSS Level = 0.18284, Truncation lag parameter = 4, p-value = 0.1
```

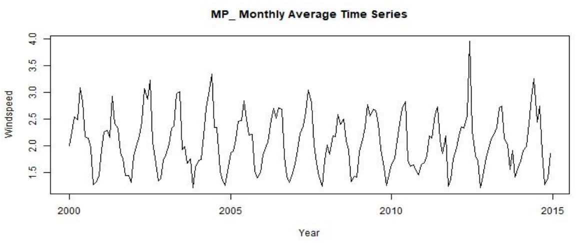


```
[1] "MP_"
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  windspeed_monthly_ns
Dickey-Fuller = -8.5268, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

p-value greater than printed p-value
        KPSS Test for Level Stationarity

data:  windspeed_monthly_ns
KPSS Level = 0.18284, Truncation lag parameter = 4, p-value = 0.1
```

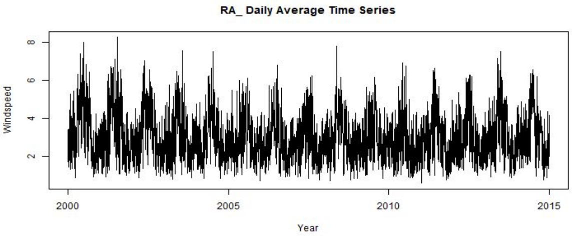
Time Series(Non Stationary) Plots for all states

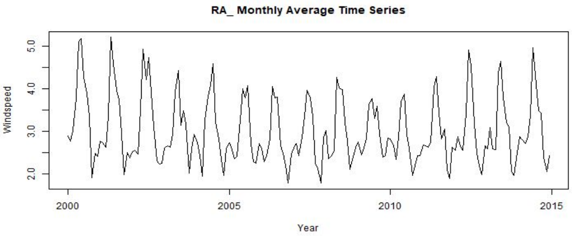
Madhya Pradesh:



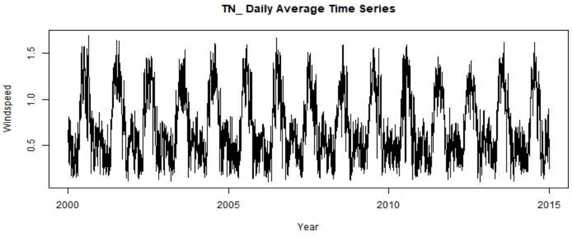


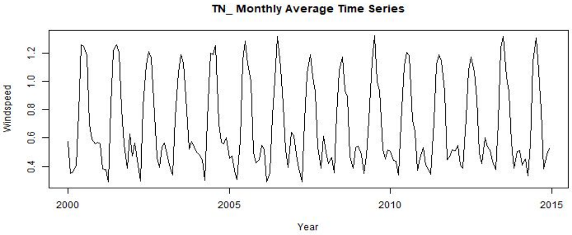
Rajasthan:





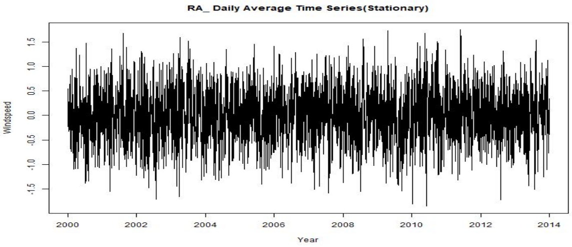
Tamil Nadu:

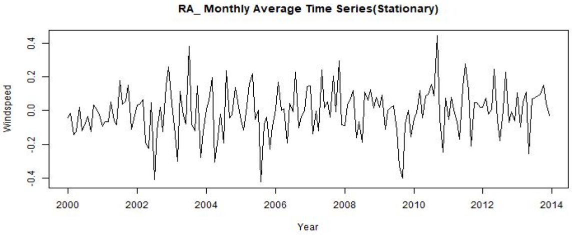




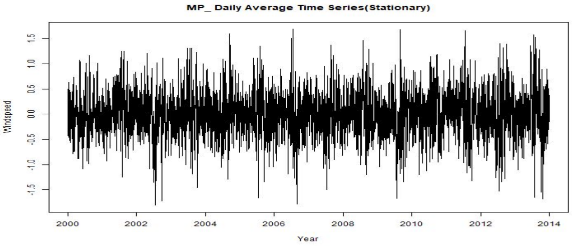
Stationary Time Series Plots for remaining states

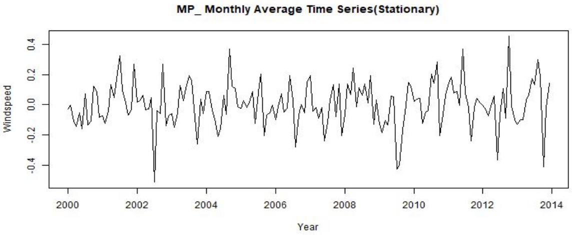
Rajasthan:



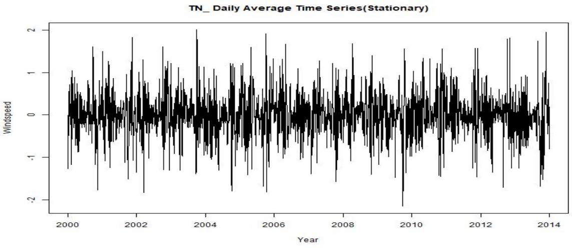


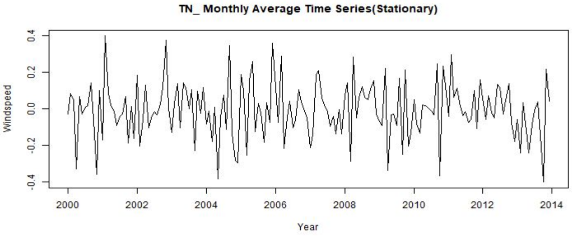
Madhya Pradesh:





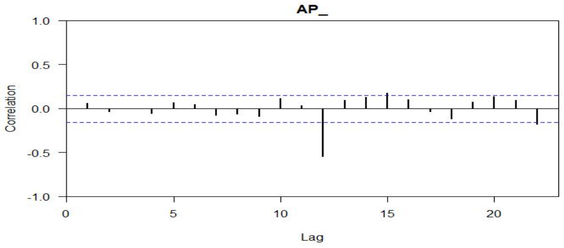
Tamil Nadu:

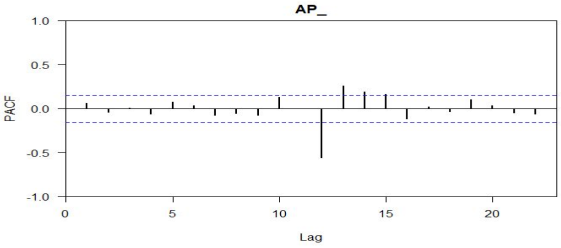




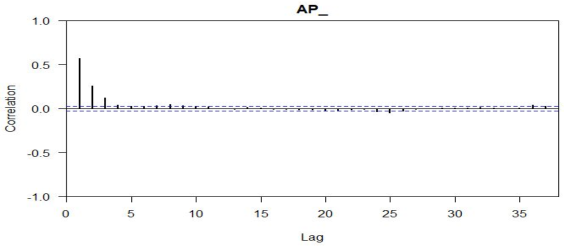
ACF and PACF plots for Stationary Data

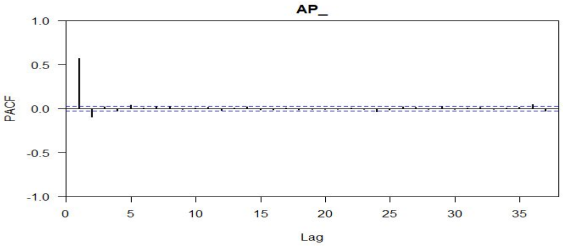
Andhra Pradesh:





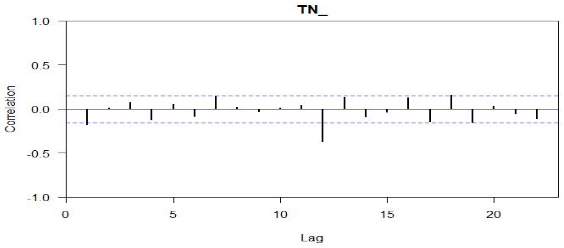
ACF Monthly PACF

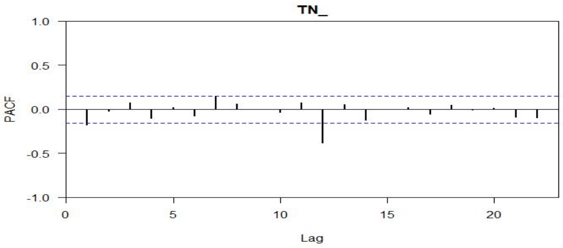




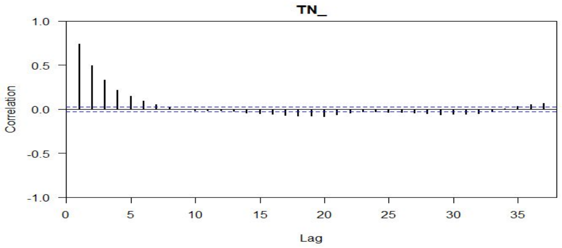
ACF Daily PACF

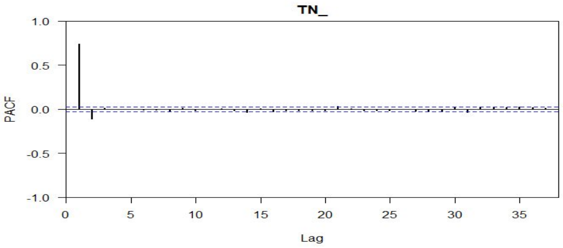
Tamil Nadu:





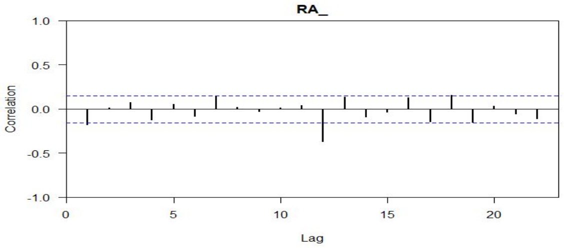
ACF Monthly PACF

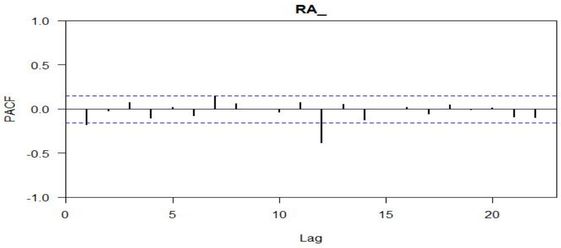




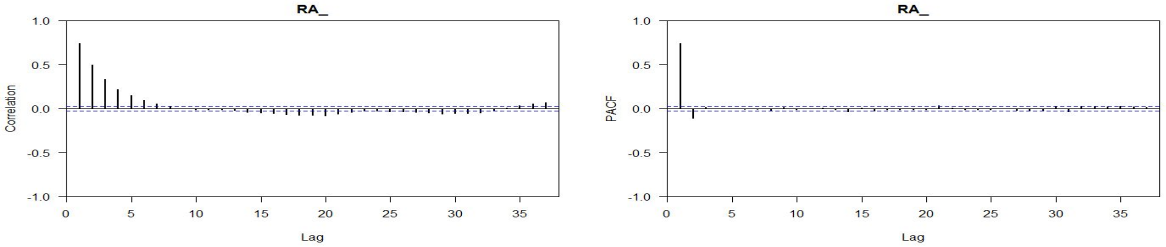
ACF Daily PACF

Rajasthan:





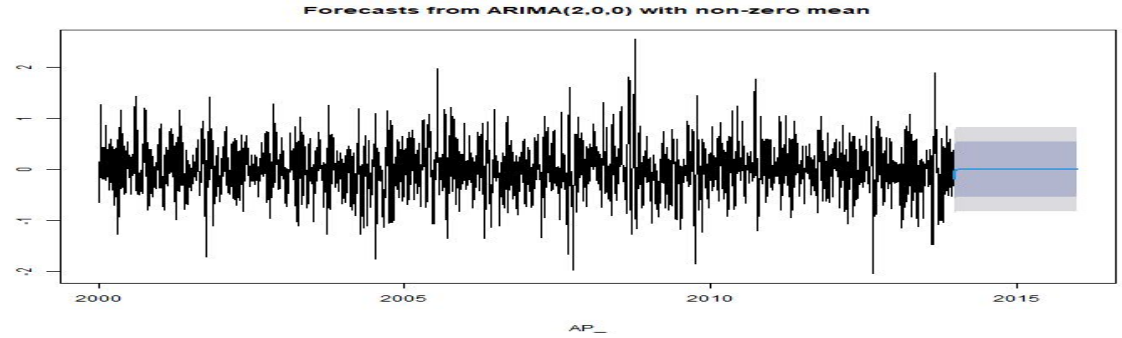
ACF Monthly PACF

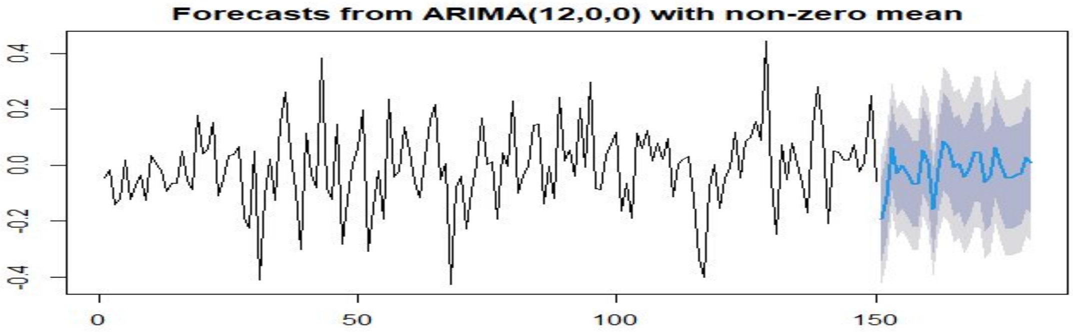


ACF Daily PACF

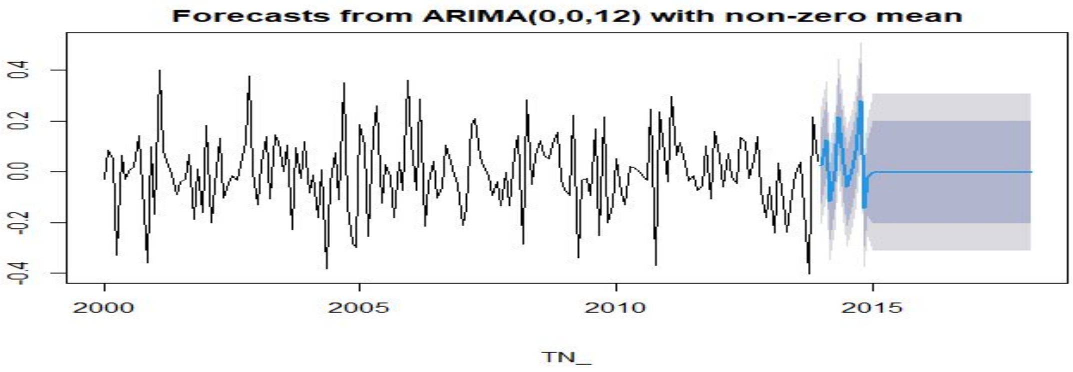
AR and MA Forecasts

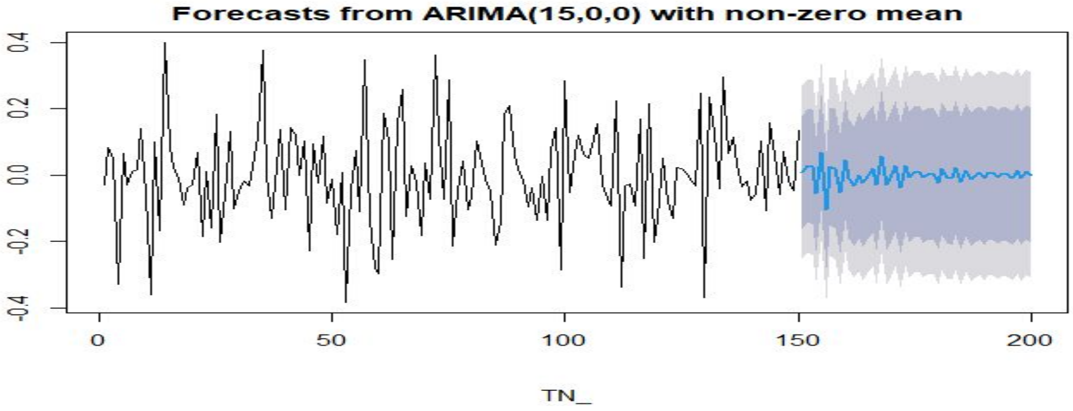
Andhra Pradesh:

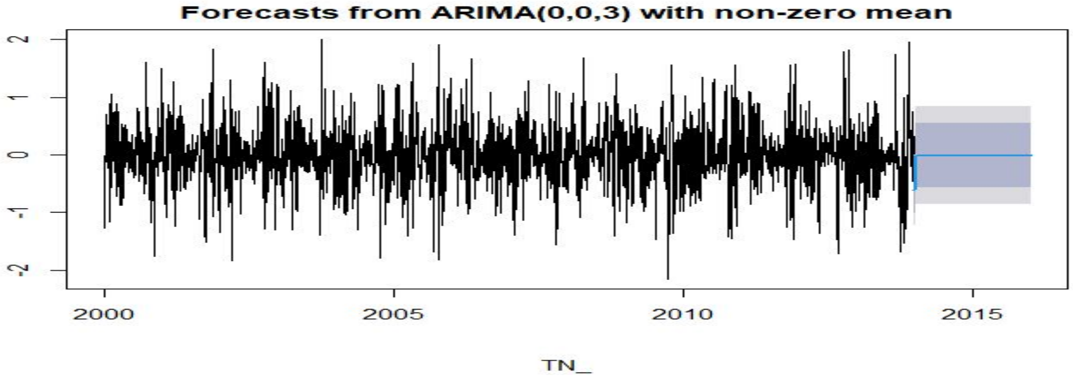


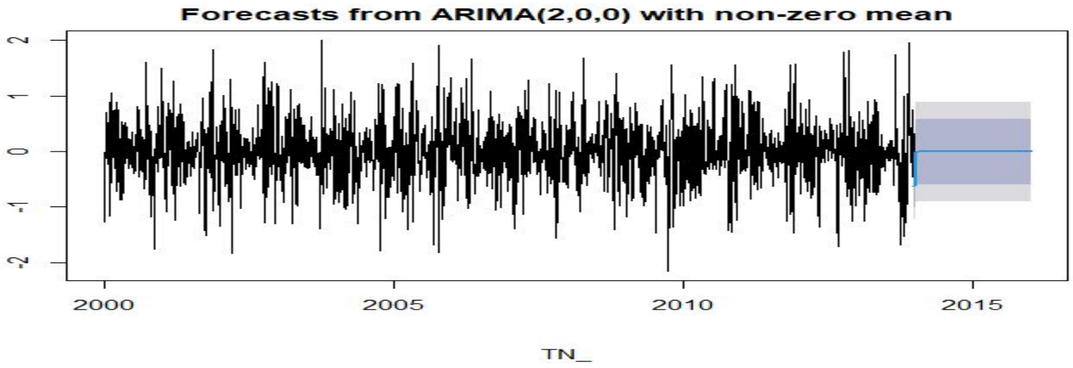


Tamil Nadu:

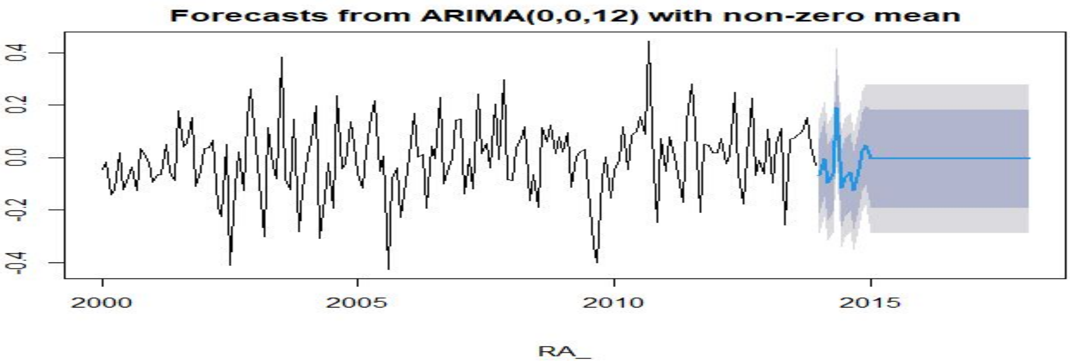

Forecasts from ARIMA(0,0,12) with non-zero mean
0.4
0.2
0.0
-0.2
-0.4
2000
2005
2010
2015
TN_

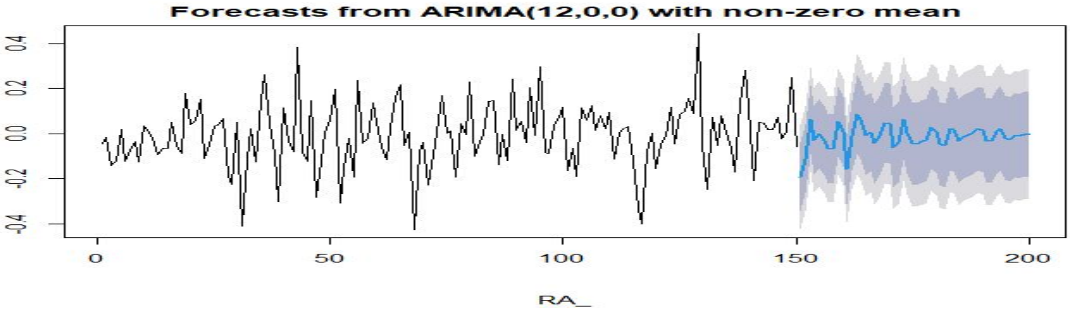

Forecasts from ARIMA(15,0,0) with non-zero mean
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
TN_

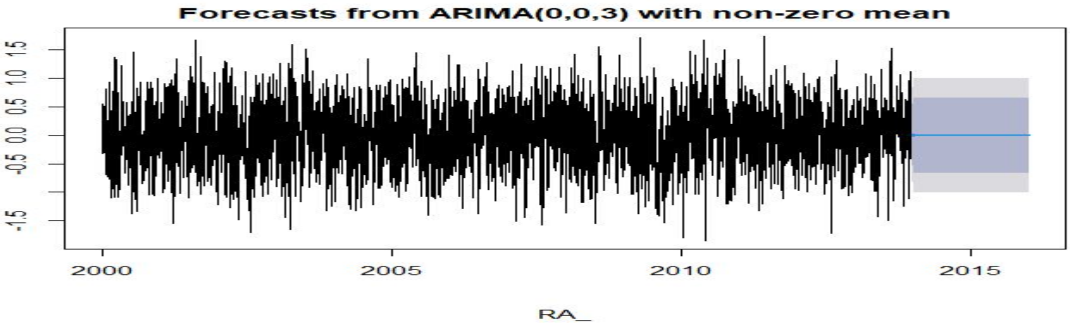

Forecasts from ARIMA(0,0,3) with non-zero mean
2
1
0
-1
-2
2000
2005
2010
2015
TN_


Forecasts from ARIMA(2,0,0) with non-zero mean
2
1
0
-1
-2
2000
2005
2010
2015
TN_

Rajasthan:


Forecasts from ARIMA(0,0,12) with non-zero mean
0.4
0.2
0.0
-0.2
-0.4
2000
2005
2010
2015
RA_


Forecasts from ARIMA(12,0,0) with non-zero mean
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
RA_


Forecasts from ARIMA(0,0,3) with non-zero mean
1.5
1.0
0.5
0.0
-0.5
-1.5
2000
2005
2010
2015
RA_