MECHANICAL AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING

UIC

TIME SERIES ANALYSIS AND FORECASTING OF WIND ENERGY GENERATION IN USA

PROJECT REPORT

Submitted by

SUBRAMANIAN RAMASAMY **658455855**

ASHWIN KUMAR V **6505507960**

In partial fulfillment of the requirements for

the course under the guidance of

Prof. Dr. Lin Li

IE 594- TIME SERIES ANALYSIS AND FORECASTING

Date: 11/30/2020

Table of Contents

Table of Figures

ABSTRACT

This project applies a class of models known as autoregressive moving average (ARMA) models. Wind Energy Generation is taken as the dataset which has a monthly time interval 12 for the period of 2001 to 2019. The training data is separated and the appropriate fit is found out for the raw dataset. The raw data is plotted and polynomial trend is calculated. The essential polynomial trend is obtained by reducing the model size to the smallest appropriate model. With the obtained residuals the stationary data is obtained and it is plotted. With this stationary data the adequate ARMA Model is found by following the Modeling Procedure. The F-test is taken and the adequate ARMA model is chosen. ARMA model is very powerful in forecasting a future event related to the information that is periodically recorded with time. The deterministic part and the adequate ARMA model is joint optimized to do the forecasting. The model is further forecasted and the update to the predicted values are also shown in this project. There is a great demand for the Wind forecasting techniques in order to maximize profits , economic scheduling and also for proper planning and execution. This project highly satisfies those needs.

1. INTRODUCTION

1.1 Growth in Wind Energy

The continual use of wind power plants made an strong impact of wind on several aspects of power system. The wind source of power is intermittent in nature. It is due to this reason of non-steady characteristics of the wind power generation means that the efficient power system operation will depend in part on the ability to forecast available wind power.

By 2023, it will be predicted that renewable energy sources (RES) will meet more than 70% of global electricity generation growth, led by solar and wind. As wind has been a source of clean energy, its production cost has been cheaper, and sustained evolution of wind energy has been taking part in energy transition around the globe

Wind power was the largest U.S. renewable energy provider in 2019. In 2019, 9.1 GW of new wind power was brought online, representing 39% of new utility-scale power additions. Adding this, operating wind power capacity in the United States now stands at over 105 GW, making it the largest renewable energy provider in the country, supplying more than 7% of the nation's electricity in 2019.

1.2 Economic Rise

The United States is home to one of the largest and fastest-growing wind markets in the world. To maintain the market , the Energy Department invests in wind research and development projects, to advance technology innovations, create job opportunities and boost economic growth. The newly released Wind Powers America Annual Report 2019 reveals that U.S. wind energy supports a record of 120,000 American jobs, 530 domestic factories and $1.6 billion a year in revenue for states and communities that host wind farms. The US Wind industry added 6,309 MW of new wind capacity in the first nine months of 2020 which is a record. There are

now over 60,000 Wind turbines with a combined capacity of 111,808 MW operating across 41 states. US wind power has more than tripled over the last decade , and today is the largest source of renewable electricity in the country.

Wind forecasting thus becomes a essential part in the Renewable energy Generation. This project aims at developing adequate model for the Wind energy and analyzing the model for the forecasting which will make it easy for the grid operators to schedule the economically efficient generation to meet the demand of electrical customers.

1.3 Data Acquisition

For this project we used Wind Energy Generation data collected from "US Energy Information Administration (EIA)" [1].The modeling technique is fitted to this data set.

2. OVERVIEW OF THE ANALYSIS

The raw dataset is taken initially and then it has been split into two sets. One with the training data and other with the testing data. The training data is used for finding the corresponding deterministic trend and the adequate ARMA model and the testing data is used for validating the predicted results from the ARMA model. Then, with the training data, we will be curve-fitting the best trend for our Non-stationary dataset and then take the residuals to convert the non-stationary data into a stationary data. Then with the help of these stationary residuals, we find the adequate ARMA model and then we jointly optimize this adequate ARMA model with the deterministic part. Then we predict the future data and finally forecast it. The above explanation can be put into the steps that we carried on to create our model.

There are 7 steps involved in creating our entire model:

1. Plot the raw data and obtain and model the deterministic trend along with the sine-cosine periodicity.
2. Then, calculate the residuals from the actual data and fitted model. These residuals are being used for converting the non-stationary data into a stationary data.
3. Then, follow the ARMA modeling procedure i.e., (2n, 2n-1) approach to find the adequate ARMA model for the residuals.
4. After finding the adequate ARMA model, calculate the characteristic roots to check whether there is seasonality present in the model. If so, find the ‘parsimonious’ model and compare the parsimonious model with the adequate ARMA model to check the significance of the model.
5. Then, jointly optimize the parameters by combining the step 2 and step 4 to find the integrated model.
6. Predict the test data and validate with the raw data.
7. Update the predicted data for a given period and the forecast it to the future.

2.1 Evaluating the Raw Data

The raw dataset that we have taken in our project is the Wind energy generation in USA from 2001 to 2020. The data has been recorded in terms of months starting from January 2001 till 2020 August.

The raw data is plotted for training set as well as for the full data set and the plots are displayed below. We use 90% of the data for training the model and the rest 10% of the data for testing. From the dataset that we have plotted, using R software, we try to create a model by applying polynomial trend. We begin our model by specifying the time trend order of 1 and sine-cosine functions with periodicity of 3, 4, 6 and 12. With that, we gradually eliminate the parameters by taking F-tests and p-values thereby, curve-fitting the appropriate model for our dataset.

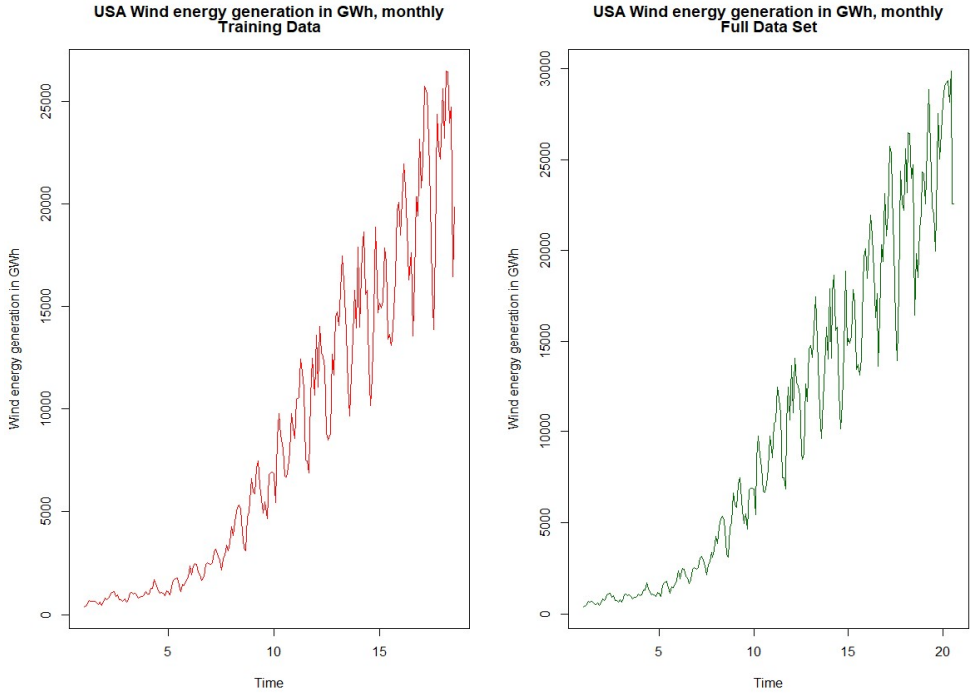

Fig2.1

Further using the R programming software the built in season trend separation function called 'stl()' (Seasonal Trend by Loess) is employed. This is done to represent the portion of data of each components. Once the function is applied to the data that we have, the following plots are shown. The portion represented by data, seasonal, trend and the remainder is obtained in the graphical representation and also the percentage is also displayed.

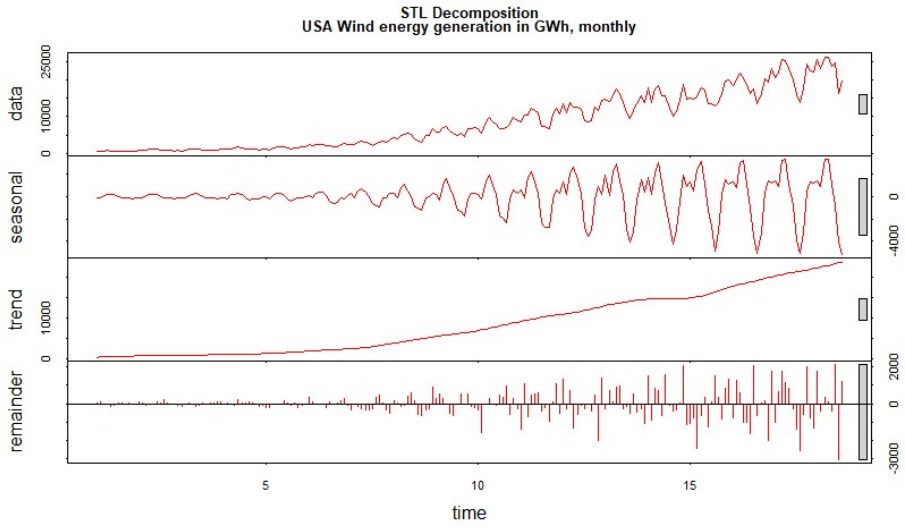

Fig.2.2

From the plots, it can be seen that our dataset mostly contributes for the trending part and then there are some proportion of seasonality present in our data. The trend part of our data seems to have a good fit of polynomial trend. So, we went with fitting the polynomial trend for detrending the data.

2.2 Obtaining the deterministic trend

In order to find out the deterministic polynomial trend, initially we create a time trend index along with the sine-cosine functions. In our case, the index starts at 0 and each month is 1/12 of a unit. So the time axis will be in terms of years. Say for example, for a data in December 2008, the corresponding time-axis value will be 8. With that index, we increase the time trend order and sine-cosine functions with periodicity of 3, 4, 6 and 12. In each case the sine-cosine functions are included. With that, we gradually eliminate the parameters by taking F-tests and p-values thereby, curve-fitting the appropriate model for our dataset. F test is employed of each order of the polynomial model and the adequacy is compared with the next reduced model. This is continued till the point when the p value of the F test crosses the 5%. During this case the previous order is chosen as the deterministic polynomial trend order. Based on the above explanation, our deterministic model will be of the following form.

$$Y_t = \sum_{j=0}^{J} \beta_j * time^j + \sum_{i=1}^{4} \delta_{0,i} \sin\left(\frac{2 * \pi * time}{12} * i\right) + \delta_{1,i} \cos\left(\frac{2 * \pi * time}{12} * i\right)$$

Where, time is the time unit of our data. In our case, the p-value of the trend at order 2 crosses 5% (p-Value = 0.4872). Hence, we fix the polynomial order to be 2 along with the sine-cosine 3, 4, 6, 12 periodicity. This is represented in Table 2.1 below

Polynomial order	F-statistic	p-Value
1	-	-
2	**74.13075**	**0.0000**
3	0.48452	0.4872

Table 2.1 : F-statistic and p-values for the polynomial order

After calculating the deterministic polynomial trend, in order to fit the curve properly the parameters of the polynomial model is inspected and upon inspection of their respective p-values, we gradually eliminated the sine and the cosine periodicity which has the highest p-value and re-ran the model in order to find the adequate trend that will be a better fit to the raw data. This process of eliminating the parameter and re-running is done until we get the smallest adequate model. In our model, the parameters are reduced to polynomial trend of order 2 and cosine function with periodicity of 12. The final smallest adequate polynomial trend model is fitted to the raw training data and the plot is shown below.

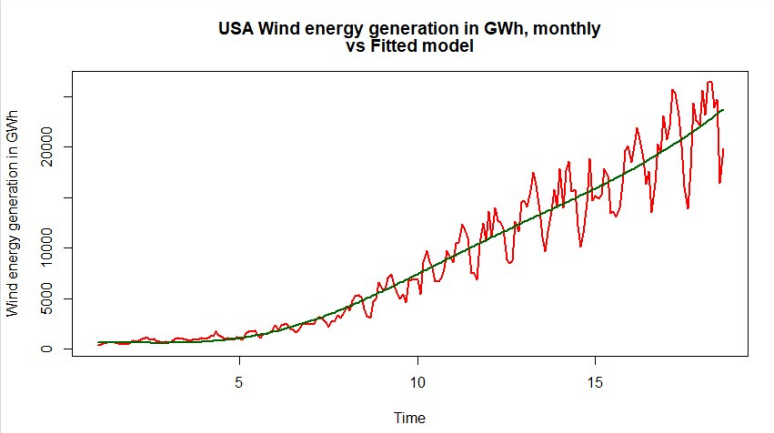

Fig.2.3

The model is then used to determine the residuals which is stationary in nature. These residuals plots are obtained from the difference between the raw data and fitted curve and it is plotted. The obtained plot is stochastic stationary residuals. Using these residuals the adequate ARMA model is selected. As you see in the below residuals plot, you can notice that although the mean is constant over time, the variance of residuals is not. Hence, we tried applying the deterministic and stochastic modeling by creating a separate dataset only for last 10 years of data instead of using entire 20-year dataset so that the residuals have constant mean and constant variance. The resulting ARMA model came to be (2,1), and moreover the predicted and forecasted results were too poor[1]. Hence we proceeded with entire 20-year dataset as such below without any changes for further analysis.

1- The predicted and forecasting plots of that ARMA(2,1) are shown in the Appendix 2 below.

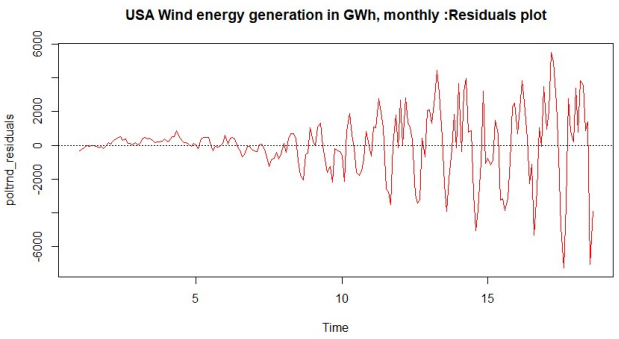

Fig.2.4

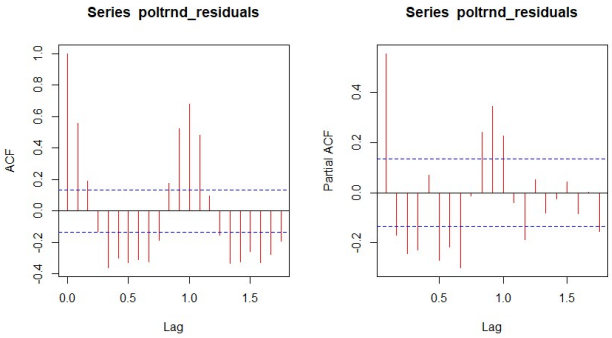

Fig.2.5

From the above ACF and PACF plots, it can be seen that there exist a strong correlation between the data points for several lags. This indicates that we might expect a higher order ARMA model.

2.3 Modeling Procedure

Using the stochastic portion of the deterministic trend model, we find the adequate ARMA model for the residuals by following the ARMA modeling procedure. The basic flowchart for the modeling procedure is as shown below in the figure Fig. 2.6. This methodology is followed in this project.

According to the modeling procedure, the order (p, q) of the ARMA model is (2n, 2n-1) where the value of n is started from 1 and the incremented accordingly. So if n = 1, then the ARMA model is ARMA(2, 1), if n = 2, then it is ARMA(4, 3) and so on. Then by calculating the RSS (Residual Sum of Squares) values, the F-statistic test is taken to compare the current ARMA order to the prior ARMA order to check its significance. If it is significant, then we move to the higher ARMA model. The F-test is as follows:

$$F = \frac{A_1 - A_0}{s} \div \frac{A_0}{N-r} \sim \mathrm{F(s,\ N-r)}$$

A_0 – sum of squares of unrestricted model

A_1- sum of squares of restricted model

N – Number of observations

s – Number of restricted parameters

r – Number of total unrestricted parameters

F(s, N-r): F-distribution with s and N-r degrees of freedom

The above procedure is carried on till we find the F-test to be insignificant. In this case, we will proceed with the previous model we obtained. Then, we will check whether the confidence interval of the AR and MA parameters include zero. If it includes, then we reduce the MA parameters till we obtain the adequate model at that level of significance. If not, then we will

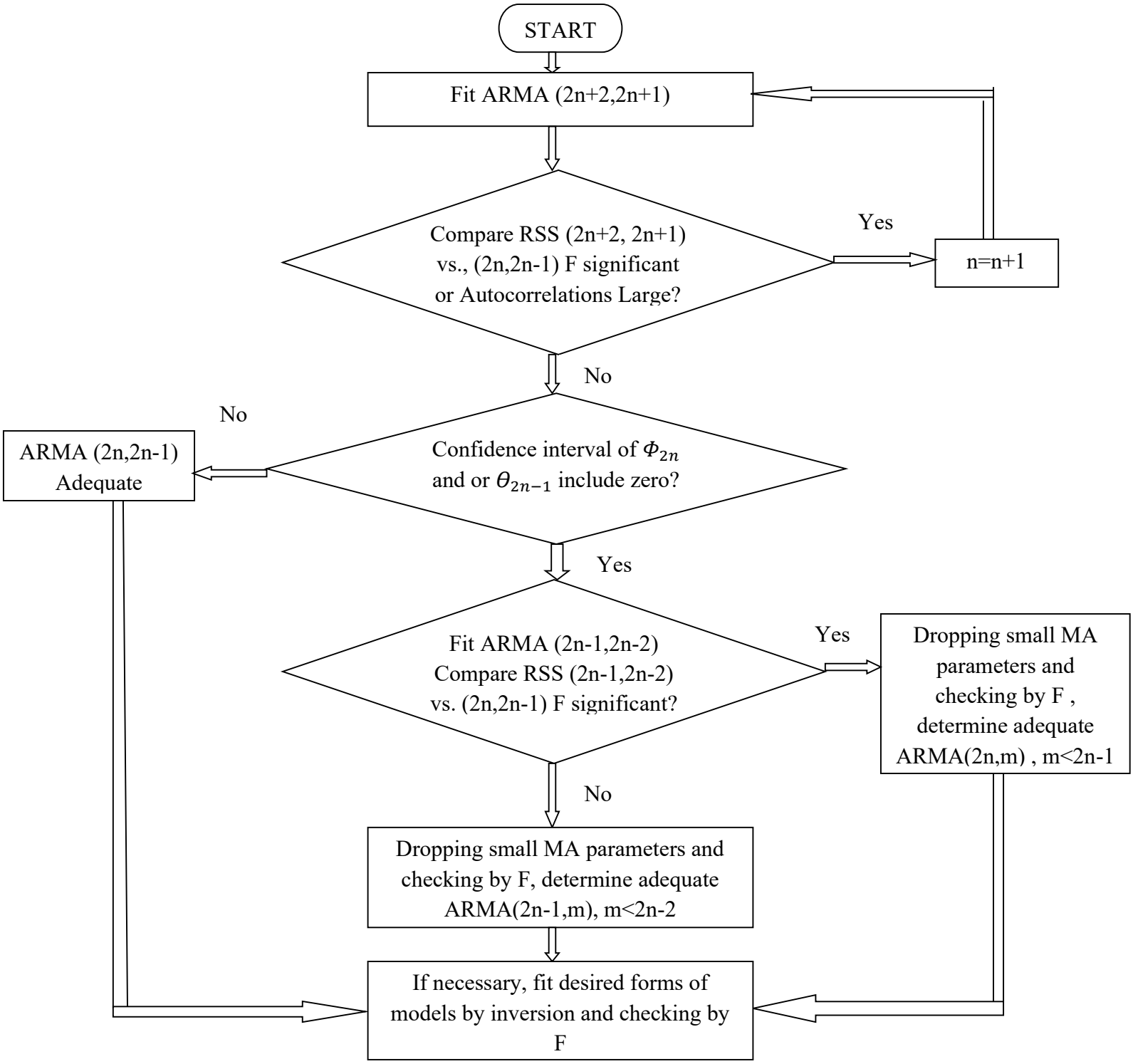
START
Fit ARMA (2n+2,2n+1)
Compare RSS (2n+2, 2n+1)
vs., (2n,2n-1) F significant
or Autocorrelations Large?
Yes
n=n+1
No
Confidence interval of Φ_{2n}
and or θ_{2n-1} include zero?
No
ARMA (2n,2n-1)
Adequate
Yes
Fit ARMA (2n-1,2n-2)
Compare RSS (2n-1,2n-2)
vs. (2n,2n-1) F significant?
Yes
Dropping small MA
parameters and
checking by F ,
determine adequate
ARMA(2n,m) , m<2n-1
No
Dropping small MA parameters and
checking by F, determine adequate
ARMA(2n-1,m), m<2n-2
If necessary, fit desired forms of
models by inversion and checking by
F

Fig. 2.6

finalize that particular (2n, 2n-1) model. In our case, the same modeling procedure is followed till we reach ARMA(16,15) where we take F-test to compare ARMA(14, 13) and ARMA(16, 15) and F-test is insignificant, i.e. F < F(s, N-r) so we get ARMA(14, 13). Also, the confidence interval of $\varphi_{14}$and θ_{13} does not include zero. Hence we finalize that ARMA(14, 13) is our adequate model and we proceed with it for the next steps.

Parameters	ARMA(2,1)	ARMA(4,3)	ARMA(6,5)	ARMA(8,7)	ARMA(10,9)	ARMA(12,11)	ARMA(14,13)	ARMA(16,15)
φ_1	1.388±0.125	-0.17±0.119	1.723±0.1928	0.4911±0.191	0.0764±0.0821	-0.2086±0.3067	-0.3266±0.0527	0.1732±0.08957
φ_2	-0.639±0.10	0.579±0.125	-1.227±0.461	-0.063±0.238	0.2413±0.037	-0.0544±0.0519	0.2621±0.00092	0.1455±0.18855
φ_3		0.289±0.102	-0.393±0.616	-0.295±0.231	-0.332±0.042	-0.2072±0.1035	0.1334±0.0306	0.2425±0.1913
φ_4		-0.741±0.104	0.9058±0.555	-0.697±0.174	-0.884±0.0397	-0.1768±0.3020	-0.1853±0.0374	0.0596±0.14171
φ_5			-0.398±0.381	0.345±0.1842	0.2832±0.0786	-0.3302±0.0412	-0.0816±0.0153	-0.2944±0.1729
φ_6			-0.237±0.165	0.0313±0.216	-0.066±0.0480	-0.1299±0.0902	-0.1088±0.0010	-0.2469±0.1258
φ_7				-0.543±0.203	-0.3672±0.062	-0.1589±0.2054	0.3018±0.0143	0.2292± -
φ_8				-0.292±0.165	-0.437±0.0648	-0.1196±0.0798	-0.0396±0.0178	-0.2503± -
φ_9					-0.3017±0.053	-0.3913±0.1509	-0.469±0.01333	-0.2560±0.1397
φ_{10}					0.1754±0.0894	-0.3145±0.1593	0.1334±0.0135	-0.1919± -
φ_{11}						0.3704±0.10153	0.6287±0.0218	0.4913± -
φ_{12}						0.4542±0.147	0.3963±0.0218	0.2292±0.0958
φ_{13}							0.4011±0.0570	0.1576±0.1078
φ_{14}							-0.2306±0.0416	-0.6977± -
φ_{15}								-0.1061± -
φ_{16}								-0.1001±0.0784
θ_1	-0.798±0.11	0.844±0.1123	-1.646±0.157	-0.344±0.165	0.0970±0.0531	0.3986±0.2621	0.7121±0.0298	-0.1230±0.0692
θ_2		-0.143±0.172	1.2077±0.341	0.1353±0.183	-0.093±0.0654	0.4256±0.1711	0.1935±0.04822	0.0767±0.24147
θ_3		-0.761±0.112	0.4832±0.458	0.3558±0.159	0.4981±0.0813	0.5139±0.17052	0.1990±0.0376	-0.2412±0.2207
θ_4			-1.198±0.353	0.6933±0.098	0.8684±0.0439	0.3±0.29498	0.1528±0.04567	-0.4501±0.1425
θ_5			0.7904±0.153	-0.307±0.166	-0.294±0.0638	0.4963±0.1019	0.1411±0.02587	0.2835±0.18698
θ_6				-0.152±0.183	0.1560±0.0488	0.1564±0.3116	0.1862±0.03783	0.1557±0.22756
θ_7				0.7913±0.163	0.5336±0.0874	0.4783±0.2540	-0.3058±0.0502	0.0115± -
θ_8					0.2961±0.0941	0.0540±0.1354	0.0085±0.03842	0.3363± -
θ_9					0.4008±0.0000	0.6353±0.2468	0.7694±0.0406	0.2908±0.11152
θ_{10}						0.6069±0.2482	-0.0546±0.0831	0.3255±0.1987
θ_{11}						-0.4886±0.2964	-0.6310±0.0580	-0.4172±0.0000
θ_{12}							-0.1468±0.1117	0.0369±0.11897
θ_{13}							-0.5248±0.0784	-0.0214±0.1940
θ_{14}								0.6092± -
θ_{15}								0.3658± -
RSS	439334250.7	362983538.6	212009671.1	193070186.5	158704157.67	142359694.271	122046632.6166	148082990.2414
F-stat	-	10.7275*	35.60542	4.806722	10.39399	5.396118	7.656097	-7.912023

* - F-test between ARMA(2,1) and ARMA(4,3). Same applies for higher order ARMA models.

Table 2.2: Parameters of various ARMA models for our data

The table above shows the parameters and its values for each ARMA model that we modeled using the above procedure. The values are tabulated from left to right according to the procedure that we followed. The corresponding ARMA order Green's function plot is as follows:

The formula for green function G_{13} is shown.

$G_{13}= \Phi_1 G_{12}+\Phi_2 G_{11} + \Phi_3 G_{10} + \Phi_4 G_9 + \Phi_5 G_8 + \Phi_6 G_7+\Phi_7 G_6 + \Phi_8 G_5 + \Phi_9 G_4+\Phi_{10} G_3+\Phi_{11} G_2 + \Phi_{12} G_1 + \Phi_{13} G_0-\theta_{13}$

Plots from G_1 to G_{13} is calculated in a similar manner and plotted

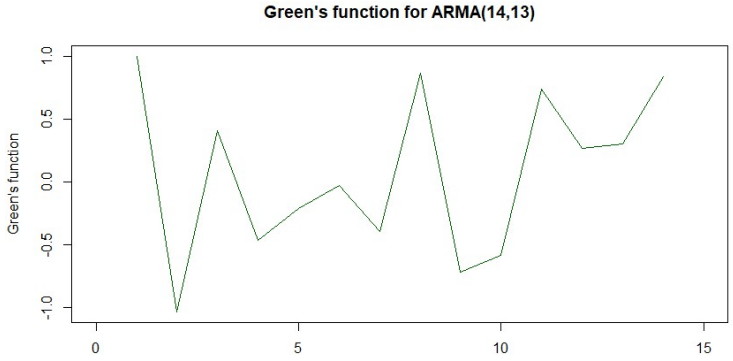

Fig 2.7

From the figure, it can be seen that the Green's function oscillates between -1 and +1. Hence it can be said that the system is stable but not asymptotically stable.

2.4 Check with the parsimonious ARMA model

Now that we have our adequate ARMA(14, 13) model, the characteristic roots of autoregressive parts (AR) of the model is found. Those 14 characteristic roots are displayed below.

Roots	**Value**	$\|\lambda_{ij}\|$	φ_1	φ_2
λ_1,λ_2	0.672052±0.698733i	0.969	1.3441	-0.9399
λ_3,λ_4	-0.977438±0.48004i	1.089	-1.955	-1.1858
λ_5,λ_6	-0.544534±0.90936i	1.059	-1.089	-1.1234
λ_7,λ_8	0.216387±0.882987i	0.909	0.4327	-0.8265
λ_9	-1.06	1.06	1.67	2.893
λ_{10}	2.73	2.73		
$\lambda_{11},\lambda_{12}$	0.955248±0.251775i	0.988	1.9104	-0.977
$\lambda_{13},\lambda_{14}$	-0.288087±1.186014i	1.221	-0.5762	-1.4900

Table 2.3: Characteristic roots of ARMA(14, 13) model

The real roots indicate that we might have stochastic trend and complex roots indicate that we might have stochastic seasonality.

From the above table, it can be seen that for $\lambda_{11,12}$, we have φ_2 approximately equal to -1 which indicates that we have stochastic seasonality. Out of calculating the time period T of our seasonality, we get T = 20.9 months. This time period T can be calculated as follows:

$$cos\omega = \cos\left(\frac{2\pi}{T}\right) = \frac{\varphi_1}{2} = 0.9552 \text{ , T = 20.9 months}$$

Hence, we find our parsimonious ARMA model to be,

$(1 - 1.9104B + B^2)(1 - \varphi_1 B - \varphi_2 B^2 - .... - \varphi_{12} B^{12})X_t = (1 - \theta_1 B - \theta_2 B^2 - .... - \theta_{13} B^{13})a_t$

Which is an ARMA(12, 13) model. After calculating the RSS values of ARMA(12, 13) model, F-statistic test is taken between ARMA(12, 13) model and ARMA(14, 13) model to test the significance of the higher model.

By taking the F-test, it is found that the F-value is significant (40.72719) and hence, we can conclude that the model ARMA(14, 13) is adequate and the final model.

2.5 Jointly optimizing the integrated model

Using the appropriate trend polynomial order (j = 2), the appropriate cosine order (i=1, only includes $\delta_{1,1}$) and the adequate ARMA model of order (14, 13), we jointly optimized the integrated model. The integrated model in our case is of the following form:

$$Y_t = \beta_2 * time^2 + \delta_{1,1}\cos\left(\frac{2*\pi*time}{12}\right) + X_t$$

Where, X_t = $-0.2067X_{t-1} + 0.8060X_{t-2} + 0.3180X_{t-3} - 0.7952X_{t-4} - 0.1200X_{t-5} + 0.0664X_{t-6} + 0.3512X_{t-7} - 0.3558X_{t-8} - 0.6139X_{t-9} + 0.1104X_{t-10} + 0.9144X_{t-11} - 0.0182X_{t-12} - 0.3243X_{t-1} - 0.2341X_{t-14} + a_t + 0.426a_{t-1} - 0.6004a_{t-2} - 0.2913a_{t-3} + 0.5698a_{t-4} + 0.0241a_{t-5} - 0.0456a_{t-6} - 0.2006a_{t-7} + 0.3428a_{t-8} + 0.6865a_{t-9} - 0.1082a_{t-10} - 0.8093a_{t-11} + 0.2777a_{t-12} + 0.541a_{t-13}$

Thus, the jointly optimized integrated model parameters along with the initial estimates are tabulated. Initial estimates are the values obtained separately from deterministic trend and ARMA models.

Parameters	Initial estimate	Joint estimate	Standard Error	p-value
φ_1	-0.3266	-0.2067	0.0723	7.164982e-03
φ_2	0.2621	0.8060	0.0564	5.361050e-31
φ_3	0.1334	0.3180	0.1256	1.682857e-02
φ_4	-0.1853	-0.7952	0.0978	1.949134e-13
φ_5	-0.0816	-0.1200	0.1285	2.572840e-01
φ_6	-0.1088	0.0664	0.1313	3.503744e-01
φ_7	0.3018	0.3512	0.1137	3.714213e-03
φ_8	-0.0396	-0.3558	0.1227	6.370756e-03
φ_9	-0.4690	-0.6139	0.1302	1.054040e-05
φ_{10}	0.1334	0.1104	0.1202	2.610391e-01
φ_{11}	0.6287	0.0914	0.1063	1.102665e-14
φ_{12}	0.3963	-0.0182	0.1090	3.928512e-01
φ_{13}	0.4011	-0.3243	0.0967	1.646794e-03
φ_{14}	-0.2306	-0.2341	0.0905	1.469878e-02
θ_1	0.7121	0.426	0.011	1.476481e-89
θ_2	0.1935	-0.6004	-	-
θ_3	0.1990	-0.2913	0.0835	1.073249e-03
θ_4	0.1528	0.5698	0.0942	2.095208e-08
θ_5	0.1411	0.0241	0.1034	3.876330e-01
θ_6	0.1862	-0.0456	0.0784	3.361541e-01
θ_7	-0.3058	-0.2006	0.0791	1.664046e-02
θ_8	0.0085	0.3428	0.0767	2.881884e-05
θ_9	0.7694	0.6865	0.1011	4.185858e-10
θ_{10}	-0.0546	-0.1082	0.0789	1.555177e-01
θ_{11}	-0.6310	-0.8093	0.0872	1.607009e-16
θ_{12}	-0.1468	0.2777	0.0202	2.483367e-29
θ_{13}	-0.5248	0.541	0.085	4.054952e-09
β_2	70.9283	71.2733	0.9655	1.656400e-137
$\delta_{1,1}$	751.4846	726.3613	142.8146	2.085971e-06
RSS	122046632.617	165296271.413		

Table 2.4: Joint estimate parameters of the integrated model

Thus, the jointly optimized integrated model parameters along with the initial estimates are tabulated. Initial estimates are the values obtained separately from deterministic trend and ARMA models.

2.6 Diagnostics on the integrated model residuals

We note that the although most of the initial estimates are close to the jointly optimized estimates, some values of the initial estimate and joint estimate parameters deviate to a larger extent. We believe this is due to the magnitude of each data in the dataset and its corresponding high residual values.

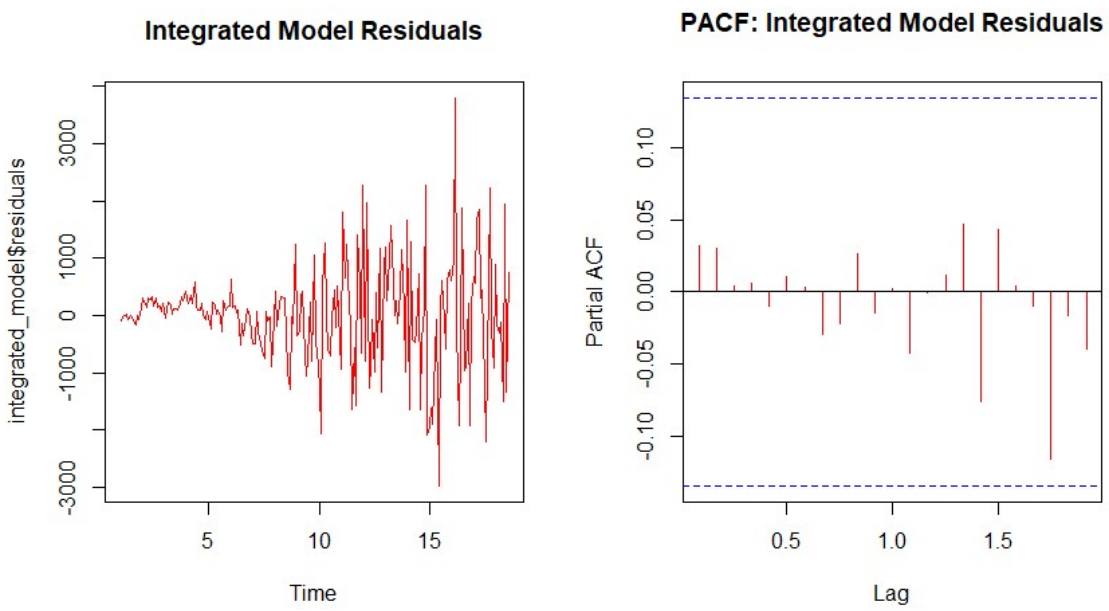

Fig 2.8

The plot for the integrated model residuals and its Partial autocorrelation function chart is displayed in the figure above. From the above integrated model residuals vs time plot, it seems that there is heteroskedasticity present in the residuals. And from the partial autocorrelation chart on the right, it seems that the residuals are uncorrelated i.e., independent. To further validate this characteristic of residuals, we perform the Box-Pierce, Box-Ljung tests to asses serial correlation in residuals. The definition for these tests are given in the Appendix 1 Section 1.1. For these two tests, we get the p-value of 1, which means that the residuals are serial independent i.e., uncorrelated. In case of Box-Pierce and Box-Ljung tests, both are testing for the null hypothesis H_0 : residuals are independent. With these high p-values, we fail to reject the null hypothesis and conclude that the residuals are independent. This can also be confirmed with the partial autocorrelation chart above.

There is another test to check the normality of the residuals and that is called Shapiro-Wilk test. Below is the histogram plot of the residuals that is being displayed.

From the histogram plot below, it can be seen that the residuals seems to follow a pattern of F-distribution, which is the distribution that is being used for our entire model. This can be validated with the Shapiro-Wilk test. The definition of this test is given in the Appendix 1 Section 1.2. For this test, the p-value was found to be 1.342e-06, which is almost 0, indicating that we reject H_0: residuals are not normally distributed. This test validates our assumption that our model residuals are F-distributed

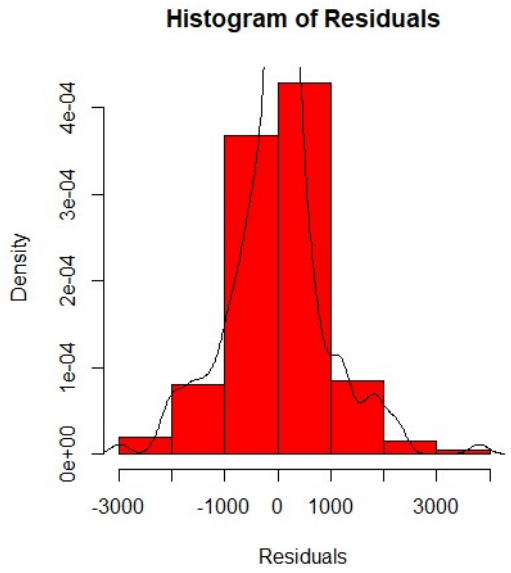

Fig 2.9

At last, the one thing that we have to validate is the heteroskedasticity of the residuals. For this, the Breusch-Pagan test for heteroskedasticity within residuals of the 'trend model' is performed. Since Breusch-Pagan test is a test for errors in regression, we use our trend model i.e., regressors of the residuals to perform the test. Upon performing the test, the p-value is 1.675e-14 which is almost 0, indicating that we reject H_0: there is heteroskedasticity within the residuals. This is clearly evident from the visual plot of the integrated residuals vs time above.

The presence of heteroskedasticity is entirely fine for our model prediction because we are using gls() 'generalized least squares' function in R for calculating the trend model. This function already assumes heteroskedasticity in the errors irrespective of whether the variance of the residuals is linearly dependent of the regressors (explanatory variables) or not. Hence, this function is efficient for calculating the trend fit model and the residuals.

2.7 Prediction and Forecasting

With the integrated model that we obtained in the previous steps, we can predict the future data from the training set that we used. This future data is predicted for the period of the test data (10% of the data) and then the predicted results are validated with the original raw test data. These results are plotted below as follows:

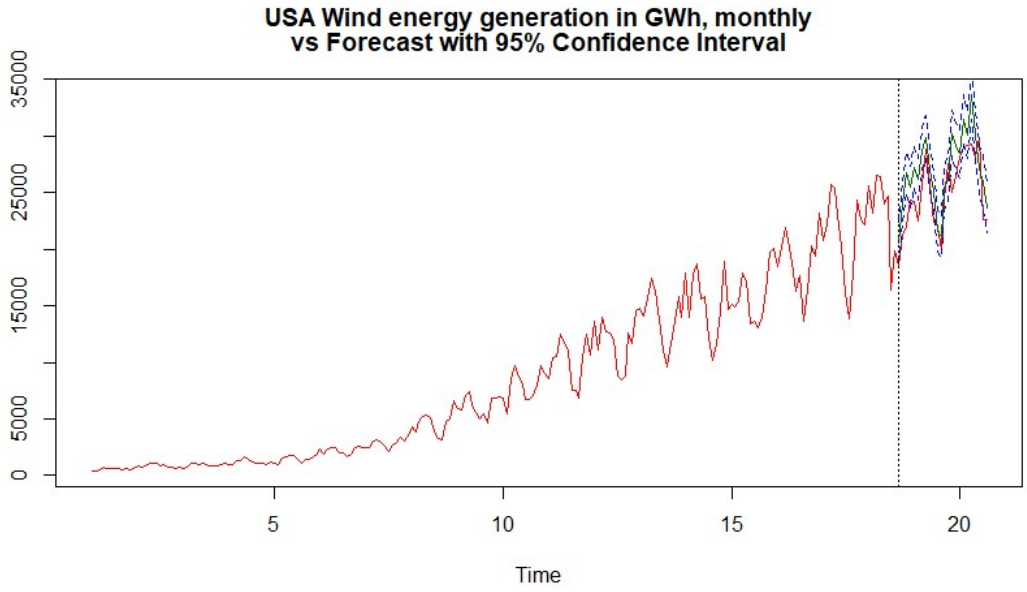

Fig 2.10

From the above plot, it can be seen that our model predicts the testing data to some good accuracy, which shows the robustness of our model. This shows that ARMA model is very powerful in predicting and forecasting a future event related to the information that is periodically recorded with time. The dark green line indicates the predicted data values and the dotted blue line indicates the 95% confidence interval. To visualize the predicted values clearly, we plot the close-up plot on the prediction period and the graph is displayed below.

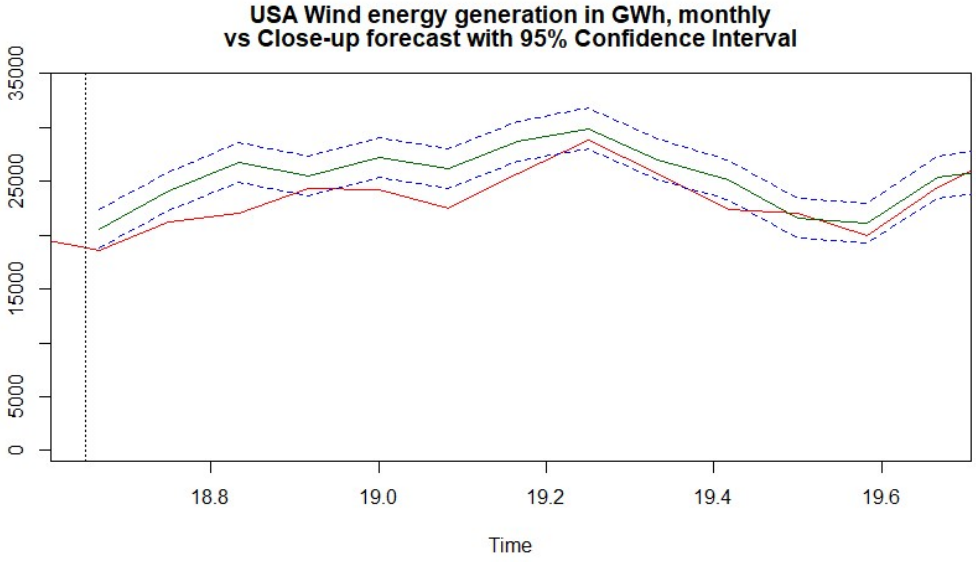

Fig 2.11

As we see from the above plot, the predicted values are close to the actual values. Initially in the first half of the prediction period, the predicted values lies on top of the actual values. But as the period progresses, the predicted as well as actual values come closer to a point where they both lie within the confidence interval. Hence, this model is pretty good for future forecasting.

We carried a forecast for a period of next 24 months i.e., 2 years and plotted the results. The results are displayed below:

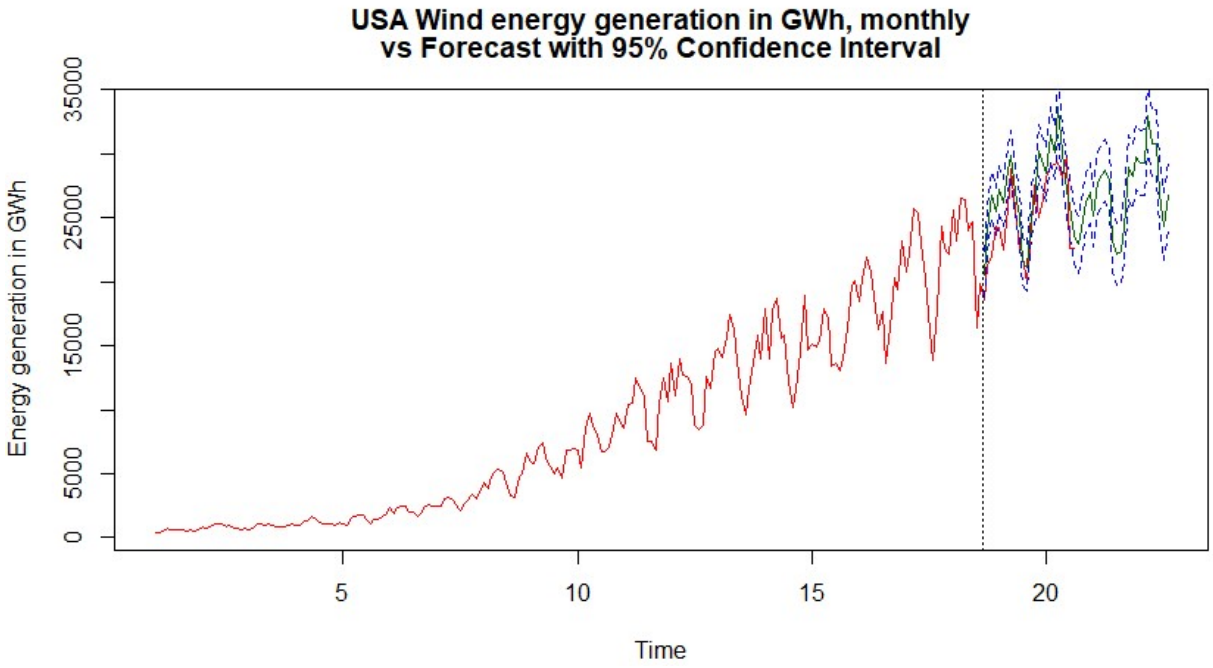

Fig 2.12

It can be seen from the above plot that the forecast gives the information about how the production of wind energy is growing up in the next 2 years. As a forecasting robustness test, we decided to play with the quantity of training data and testing data respectively. This is to ensure the right balance between the fitted model and accurate forecast in the future. This is because not only fitting the model is important, but also the future forecast in important and that is our ultimate goal. We played with the quantity of the data to be tested by increasing the training data set and decreasing the testing data set. We expected a decline in the MSE (Mean Square Error) value as we increase the training data and decreasing the testing data. To ensure this, we considered the training data sizes of 90%, 93%, and 95% of the full dataset. The results are displayed below in the table. It was curious to see that the MSE value decreased from 90% to 93% of the data but the MSE value increased when we go from 93% to 95% of our data. Hence, this shows that we can update the data during the period of 1-1.5 years repetitively so that the forecasts results turn out to be as much accurate as possible.

Training %	90%	93%	95%
MSE	6918056.999	3876220.435	4687219.101

2.8 Updating the forecast

We proceeded to update the predicted results for a period of one year (last 12 data points) to see whether we obtain better results. Updating the forecast will be of form:

$$\widehat{X_{t+1}}(l) = \widehat{X_t}(l+1) + G_l a_{t+1}$$

$a_{t+1} = X_{t+1} - \widehat{X_t}(1)$, where a_{t+1} is the error between actual true value and predicted value,

where, $\widehat{X_{t+1}}(l)$ – New updated forecast,

$\widehat{X_t}(l+1)$ – Old forecast

G_l - Green's Function

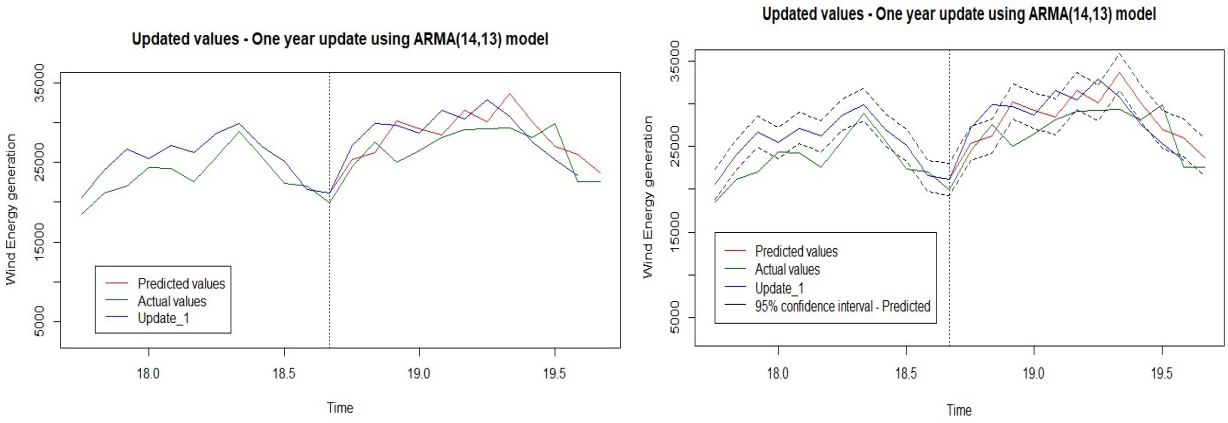

Fig 2.13

From the plot it can be seen that the values are updated in the yearly fashion. That is, the predicted values are compared with the actual values after one year (12 data points) data of predicted values and updated it accordingly using the Green's function. The plots above on both the sides are the same plot with left one being without the confidence interval and the right one with the 95% confidence interval of the predicted values.

3. CONCLUSION

The obtained predicted and forecasted results of our deterministic and stochastic modeling are fairly meeting our expectations. As the year progresses, as per our forecast, the energy production also increases. This can be attributed to the fact that renewable energy demand is rising owing to the fact of depletion of fossil fuel resources. Moreover, our results are cross-checked with the official EIA forecasts. They have predicted, in 2019, that according to the January STEO (Short-Term Energy Outlook), wind generation will grow by 12% and 14% during the next two years[2]. Our prediction results tells that there will be an increase of 11.12% during the years 2019 - 2021. Although the percent increase of our forecast varies a little with the

official EIA forecasts, it is clearly evident that there will be an increase in the wind energy generation for the next two years. Hence, we can say that our model is robust in predicting the results.

This work can be further extended to analysis of wind energy generation contributed from each state in US. With that data, we can be able to infer the economical method of production of wind energy by analyzing the states that have the kind of geography to produce more wind energy and thereby, setting up more wind farms in those places.

3.1 Impact of COVID-19 on wind energy generation

Although COVID-19 has caused severe impact on other renewable energy production, US wind farms commissioning is on track to reach record levels in 2020 despite disruptions caused by the coronavirus pandemic, according to the new data from the American Wind Energy Association (AWEA) [7]. Hence, our predicted results might be close to the realistic data that we obtain in the upcoming future.

APPENDIX 1: Definition of selected statistics

1.1: Box-Pierce and Box-Ljung tests:

The Box-Pierce test is that the distribution of $Q(r) = n\sum_{k=1}^{m} r_k^2$, where $r_k^2 = \frac{\sum_{t=k+1}^{n} a_t a_{t-k}}{\sum_{t=1}^{n} a_t^2}$ could be approximated by that of χ_m^2[3]. Basically, this statistic calculates the sum of autocorrelation errors for lag n multiplied by n. The Box-Ljung test is that it is the modified test based on the criterion $Q(\hat{r}) = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}\hat{r}_k^2$[3], where $\hat{r}$ is the estimate of r and r_k^2 is same as from the Box-Pierce test.

1.2: Shapiro-Wilk test:

Shapiro-Wilk test is a test of normality in frequentist statistics. The Shapiro-Wilk test tests the null hypothesis that a sample x_1,….,x_n came from a normally distributed population. The test statistic is,

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$ [4]

Where, $x_{(i)}$ is the *i*th order statistic,

$(a_1,\ldots,a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$, and $m = (a_1, \ldots., a_n)^T$ is a vectorand is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and finally, V is the covariance matrix of those normal order statistics [4]. If the p-value is insignificant, then we reject the null hypothesis.

1.3: Breusch-Pagan test:

This test is used to test for heteroskedasticity in a linear regression model [5]. For computational purposes, let $\mathbf{Z}$ be the *n* x *P* matrix of observations on (1, $\boldsymbol{z}_i$), and let **g** be the vector of observations of $g_i = \frac{e_i^2}{e'e/n} - 1$, then LM = $\frac{1}{2}[\boldsymbol{g'Z(Z'Z)^{-1}Z'g}]$. Under the null hypothesis of homoskedasticity, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in $\boldsymbol{z_i}$ [6].

APPENDIX 2: Predicted and forecasting plots of the ARMA(2,1) model

As we discussed in the section 2.2 above, for considering constant mean and constant variance of the residuals value, we created a separate dataset of the last 10 years data alone instead of the entire 20-year dataset. The predicted and future forecast plots are shown below.

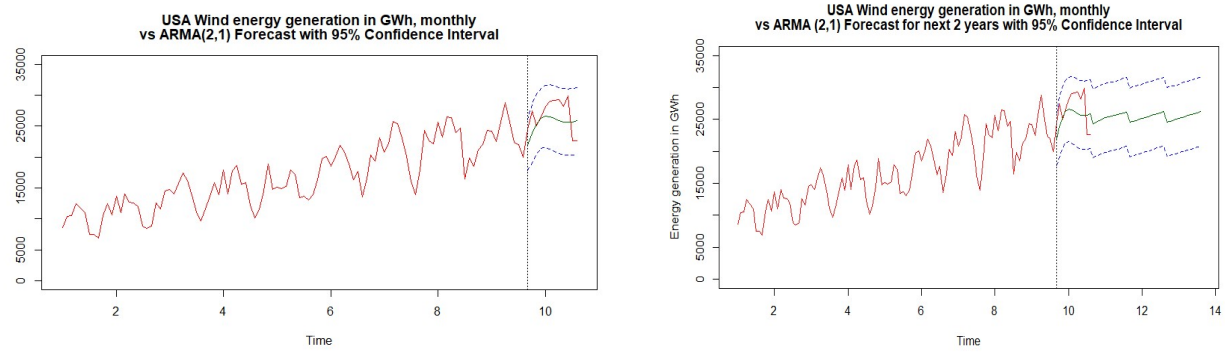

Fig. 1

Moreover, the Mean-Squared Error (MSE) value of the actual and predicted values turned out to be poorer (MSE for 93% of training data for ARMA(2,1) = 7910173.12) than the analysis that we made in this report so far (MSE for 93% of training data for ARMA(14,13) = 3876220.435). Thus, we proceeded with ARMA(14, 13) modeling with the entire 20-year dataset and moreover, the obtained results were fairly good.

4. REFERENCES

1. Online:https://www.eia.gov/electricity/data/browser/#/topic/0?agg=1,0,2&fuel=008&geo=vvvvvvvvvvvvo&sec=o3g&linechart=ELEC.GEN.WND-US-99.M~ELEC.GEN.WND-IA-99.M~ELEC.GEN.WND-TX-99.M&columnchart=ELEC.GEN.WND-US-99.M~ELEC.GEN.WND-IA-99.M~ELEC.GEN.WND-TX-99.M&map=ELEC
2. Online:https://www.eia.gov/todayinenergy/detail.php?id=38053#:~:text=According%20to%20the%20January%20STEO,very%20little%20growth%20in%202020.&text=An%20additional%208%20GW%20of,to%20come%20online%20in%202020.
3. G.M. Ljung, G.E.P. Box, “On a measure of lack of fit in time series models”, Biometrika (1978), 65, 2, pp. 297-303.
4. Shapiro-Wilk test, Wikipedia online: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
5. Breusch-Pagan test, Wikipedia online: https://en.wikipedia.org/wiki/Breusch%E2%80%93Pagan_test
6. William H. Greene, Econometric Analysis, seventh edition (p. 316)
7. Online: https://www.windpowermonthly.com/article/1698788/us-wind-power-on-track-record-year-despite-covid-19