Analysis of Time-Series Trends and ARIMA models to Forecast COVID – 19 cases

Dr.L.V.Nandakishore[1], Dr.S.Aruna[2]

[1]Department of Mathematics, Dr. M.G.R. Educational and Research Institute, Chennai, India.

[2]Department of Computer Science, Agurchand Manmull Jain College, Chennai, India.

Abstract

COVID-19 a novel corona virus originated from Wuhan China. It turned into a pandemic resulting in a large number of deaths and loss of livelihood. It is vital to determine the manner in which the number of cases propagates so that future pandemics can be tackled scientifically. However the pandemic can be controlled systematically using efficient health care systems. It is difficult to predict the pandemic propagation over a large period of time due to various factors. In this paper an analysis is made for short periods using statistical tools like predicting the probability curve, probability density function. Forecasting of Covid-19 cases is done using time series trend analysis and ARIMA models. The test of hypothesis for difference of means and standard deviations of the actual and forecasted values with 99% CI showed no significant difference between them.

Keywords: COVID-19, Probability distribution function, Time Series, Trend analysis, ARIMA, and Hypothesis testing.

1. Introduction

The pandemic of COVID-19 originated in Wuhan, China and has caused a heavy loss in lives, lockdowns and loss of livelihood etc. Data sets are available for this pandemic in the official website of Johns Hopkins University. Data set for India is considered for statistical analysis for this pandemic to predict the propagation of the disease and control the same scientifically. This must be modeled scientifically to assist policy makers and healthcare community to be prepared for future consequences to help control the problem.

2. Methodology

2.1 Dataset

For this analysis Data set for India for a period of first April 2020 to fifteenth June 2020) was obtained from the official website of Johns Hopkins University (https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html). From this the 1st April 2020 to 31st May is data is analyzed statistically to predict the number of cases for the period from June 1 to June 15, 2020 and compared with the actual data.

2.2 Model development

The data for India for the above period is considered. A probability distribution is fitted to the data, which is a best fit based on Kolmogorov Smirnov ranking test. Time series trend analysis is used to find the parameters of various models like MAPE, MAD and MSD values. An efficient model is the one which has the lowest value for the above measures. For forecasting a time series, ARIMA modeling is an efficient method. ARIMA procedure analyzes and forecasts equally spaced univariate time series data, transfer function data, and intervention data using the Auto Regressive Integrated Moving-Average (ARIMA) or autoregressive moving-average (ARIMA) model. An ARIMA model forecasts future values as a linear combination of Auto regressive terms, moving average terms and a constant.

The model for forecasting future confirmed Corona cases is given below,

Forecasting equation Y_t (Predict) = AR terms – MA terms + Constant

$$= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots\ldots\ldots \phi_p y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots \theta_q e_{t-q} + \mu \quad (1)$$

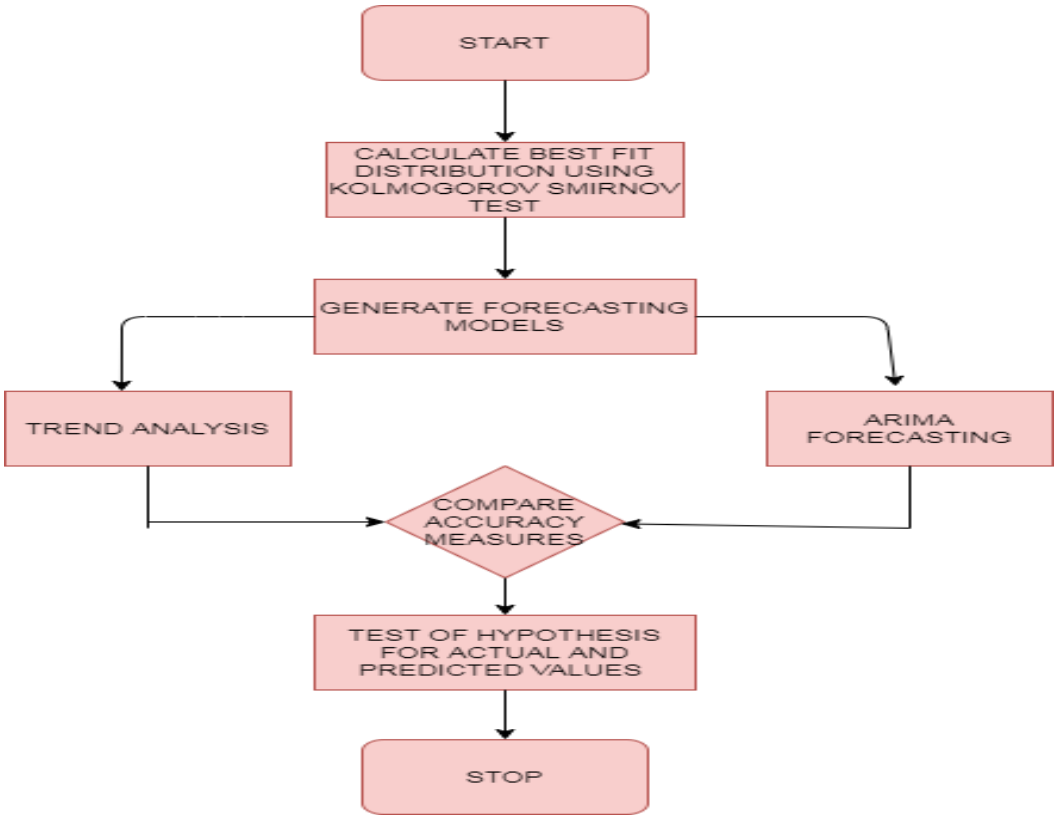

Figure 1. Flow chart for Model Development

2.3 Measures of Accuracy

The accuracy of fitted time series values as a percentage is given by the Mean Absolute Percentage Error (MAPE).The accuracy of fitted time series values is given by Mean Absolute Deviation (MAD). Mean Squared Deviation (MSD) is applied to determine the overall deviation of data set from the mean.

3. Statistical Analysis

3.1 Probability Distribution

The given data is input into Math wave software to predict the best fit distribution and its parameters.
It was found that the Johnson SB distribution was the best fit by Kolmogorov Smirnov and χ^2 ranking test (Table 1).

Table 1. Ranking for Best Fit Distribution

Distribution	Kolmogorov Smirnov		Anderson Darling		χ^2 test	
	Statistic	Rank	Statistic	Rank	Statistic	Rank
Johnson SB	0.0273	1	0.05606	1	0.10843	1
Beta	0.05472	2	0.7337	2	1.0656	2

The probability distribution function for Johnson SB distribution is given by

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}z(1-z)} e^{\frac{1}{2}\left(\gamma+\delta\ln\left(\frac{z}{1-z}\right)^2\right)}, \quad z = \frac{x-\xi}{\lambda} \qquad (2)$$

Z is a standard normal random variable, γ and δ is the shape parameters; λ is a scale parameter and ξ is a location parameter. The parameters for this data set are given in Table 2. A prediction can be

attempted for future cases applying the parameters obtained. Figure 2 gives the probability distribution function.

Table 2. Parameters for the distribution obtained

γ	δ	Λ	ξ
0.77518	0.52552	2.2204E+5	811.58

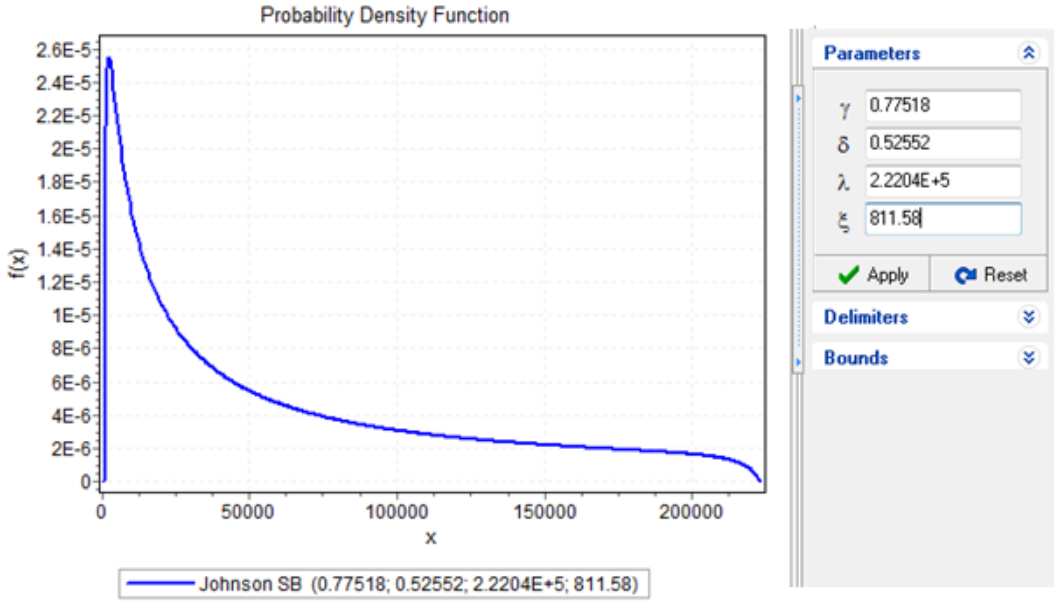

Figure 2. Probability distribution function for the COVID data

3.2. Time series trend analysis

Table 3. Measures of Accuracy and Formulae

MEASURE	MAPE	MAD	MSD
Formula	$\frac{\sum_{t=1}^{n}\left\lvert\frac{y_a - y_p}{y_a}\right\rvert}{n} * 100$	$\frac{\sum_{t=1}^{n}\lvert y_a - y_p\rvert}{n}$	$\frac{\sum_{t=1}^{n}\lvert y_a - y_p\rvert^2}{n}$

Y_a is actual value at time t

Y_p is forecasted value

n is number of observations

Table 4. Measures of Model Accuracy

It is evident from the table above that MSD, MAPE and MAD have the least value for Double exponential Smoothing. (Table 4.).

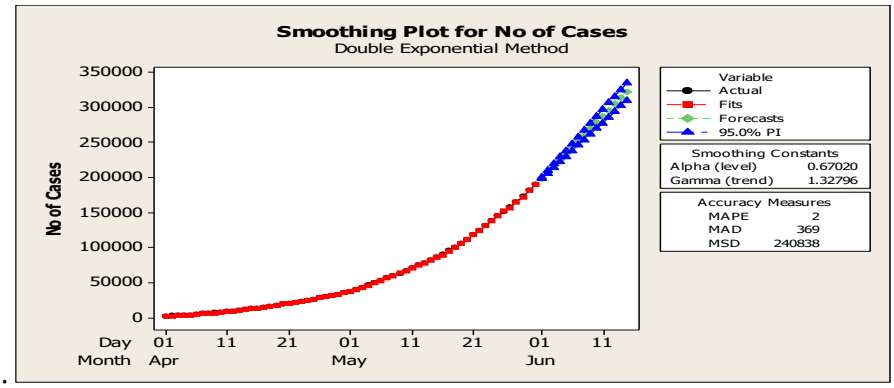

Figure 3. Plot for Double Exponential Method

3.2. ARIMA Method

The data for the months of April and May were normalized using the Box Cox plot method.

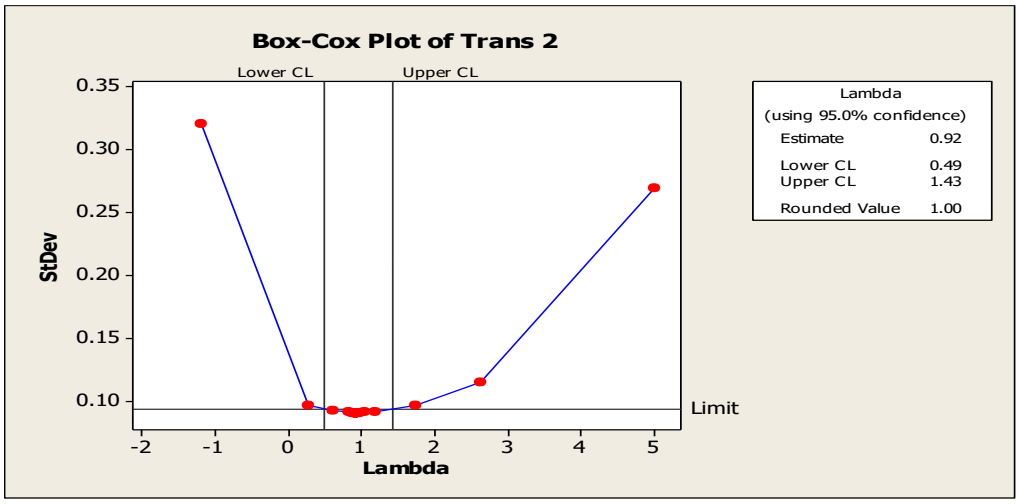

Figure 3. Box –Cox Plot

In Box-Cox plot a rounded value of λ =1 was obtained after two transformations as shown in figure 3. Table 5 shows the measures of ARIMA. It was found that the types AR (2) and MA (2) gave the least values for measure of accuracy.

3.3. Final Estimates of Parameters

Table 5. Measures of ARIMA

Type	Coeff	SE Coeff	T	P
AR (1)	-0.6955	0.5058	-1.37	0.175
AR (2)	-0.4642	0.2596	-1.79	0.079
MR (1)	-0.1489	0.5146	-0.29	0.773
MA (2)	-0.3163	0.2671	-1.18	0.242
Constant	306.82	90.88	3.38	0.001

Table 6. Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

Lag	12	24	36	48
Chi-Square	2.7	21.4	29.3	47.3
DF	7	19	31	43
p-Value	0.909	0.314	0.556	0.301

Analysis of the Modified Box-Pierce (Ljung – Box) $\chi 2$ statistic for Differencing: Two regular differences. Table 6 shows that different values of lag (12, 24, 36 and 48), the p –value is greater than the significant level of 0.05 when compared to Chi-Square and DF values.

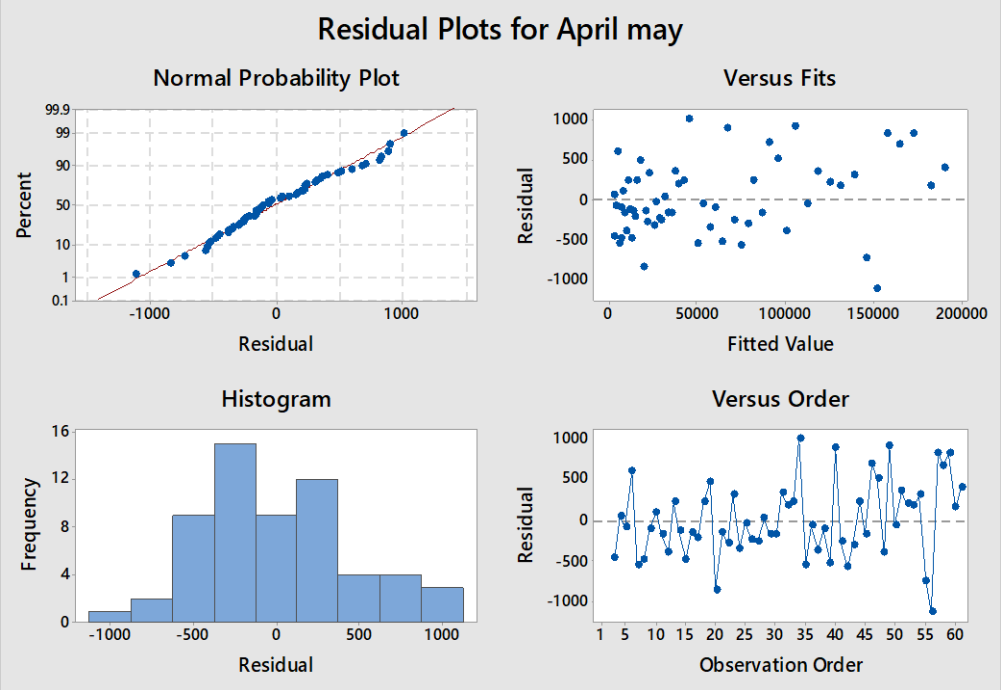

Figure 4. Residual Plot

From figure 4. it is observed that there is not much deviation of residuals from the straight line indicating normalcy. The histogram shows no outliers and shows near normal curve. The graph of fitted values and residuals shows that the residuals show a random pattern and constant variance. The versus order plot doesn't display any non random pattern implying that the data can be used to predict time related events. It shows the non correlation of residuals.

Table 7. Forecasts from Period I June to 15 June at 95% Limits

June 2020	Actual	Predicted	Lower Limit	Upper Limit
1	198370	199387	198453	200321
2	207191	208392	206744	210040
3	216824	217548	214958	220137
4	226713	226800	223041	230559
5	236184	236222	231227	241217
6	246622	245788	239433	252142
7	257486	255482	247643	263321

8	265928	265327	255919	274734
9	276146	275314	264242	286385
10	286605	285439	272611	298267
11	297535	295709	281044	310374
12	308993	306121	289538	322703
13	320922	316674	298095	335252
14	332424	327369	306722	348016
15	343091	338207	315419	360995

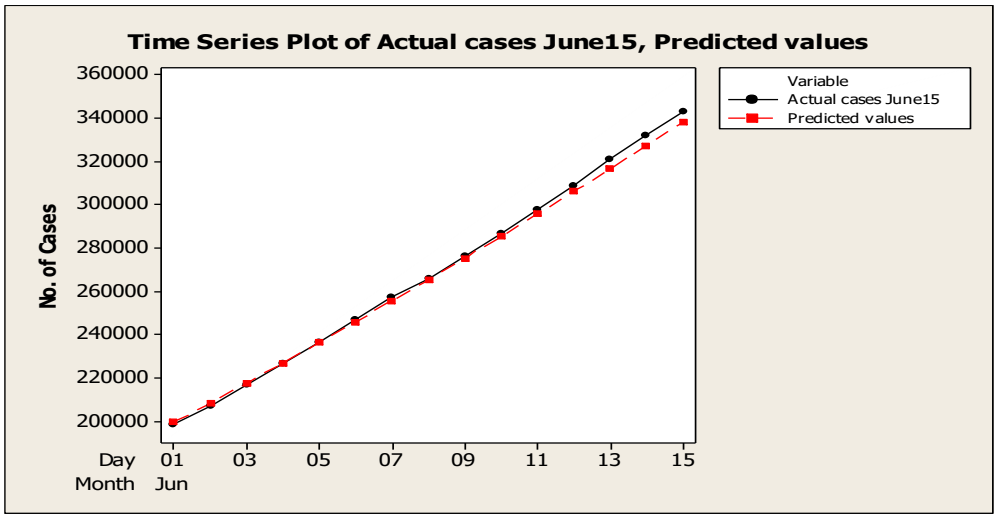

Figure 5. Time Series Plot for Actual and Predicted Value

3.4. Test of Hypothesis

Test of hypothesis was conducted for the 2 samples actual and forecasted. It was found that there was no significant difference in means of the two. If it is not possible to compare the two values of predicted and actual for a large prediction, this method can be applied to infer that the means of actual and predicted values are not significantly different. An average prediction may be sufficient.

The two samples consisting of actual and predicted values are tested for significant difference between means and standard deviations using test of hypothesis.

H_o: No significant difference between two means.
H_1: The two means differ significantly.

The null hypothesis is accepted. It is evident from Figure 6 and Figure 7.that there is no significant difference in the two means and standard deviations at 99% confidence intervals.

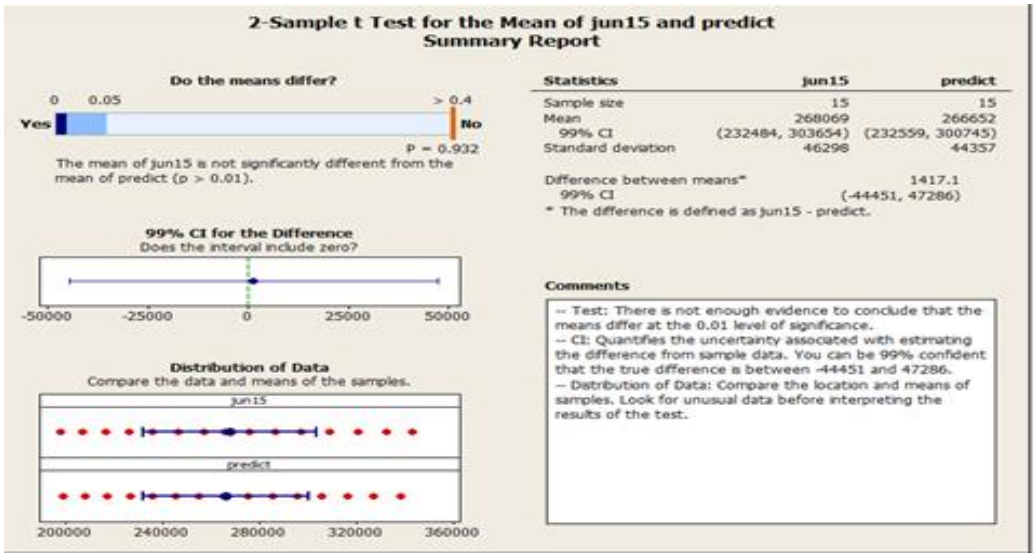

Figure 6. Hypothesis Test for Mean

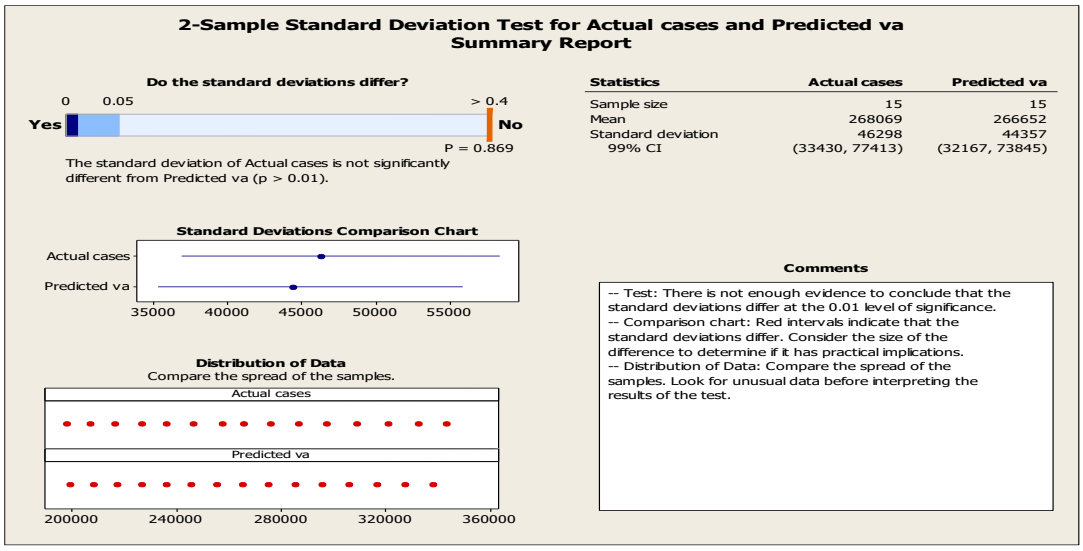

Figure 7. Hypothesis Test for Standard Deviation.

Conclusion

Data related to number of cases of COVID-19 was taken for India for the months of April and May for time series trend analysis and ARIMA forecasting analysis from the official website of Johns Hopkins University. A probability distribution was fitted and the best fit parameters found using Kolmogorov Smirnov ranking method. Time series trend analysis and ARIMA methods were applied to forecast data for the first half of the month of June using the best fit parameters. The actual and predicted values were compared using test of hypothesis for significant difference in mean and standard deviation and found that there is no significant difference.

References

[1] Jamal Fattah , Latifa Ezzine1, Zineb Aman, Haj El Moussami, and Abdeslam "Forecasting of demand using ARIMA model", Journal of Engineering Business Management Volume 10: (**2018**), pp. 1–9

[2] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi,1and Charles Korede Ayo "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction", Hindawi Publishing Corporation Journal of Applied Mathematics Volume (**2014**), pp. 1-7.

[3] Dr. Jiban Chandra Paul, Md. Shahidul Hoque, Mohammad Morshedur Rahman "Selection of Best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square Pharmaceutical Ltd." Global Journal of Management and Business Research Finance Volume 13 Issue 3 Version 1.0 Year (**2013**).

[4] Sudeshna Ghosh Scottish Church College, India "Forecasting Cotton Exports in India using the ARIMA model" Amity Journal of Economics 2(2), (**2017**), ADMAA, pp 36-52.

[5] Andrew T. Jebb, Louis Tay, "Introduction to Time Series Analysis for Organizational Research",Organizational Research Methods, Sage,(**2017**).

[6] Mohammed Moyazzem Hossain1, Faruq Fbdulla, Imran Parvez "Time series analysis of onion production in Bangladesh" Innovare Journal of Agricultural Sciences, Vol 5, Issue 1, (**2017**),pp. 1-4

[7] O'Brien, Chris, van Riper, Charles, and Myers, Donald E, " Making reliable decisions in the study of wildlife diseases: using hypothesis tests, statistical power, and observed effects", Journal of Wildlife Diseases, 45(3) : pp.700-712