ARIMA modelling of predicting COVID-19 infections

W.Regis Anne[a], S.Carolin Jeeva[b]

[a]Department of Computer Applications, Sri Krishna College of Engineering and Technology,Tamil Nadu, India

[b]Department of Computer Applications, Karunya University,Tamil Nadu, India

[a]regisanne@skcet.ac.in,[b] carolin@karunya.edu

Abstract

The World Health Organization (WHO) Director-General, Dr. Tedros Adhanom Ghebreyesus on March 11, 2020 declared the novel coronavirus (COVID-19) outbreak a global pandemic [4] the reason being the number of cases outside China increased 13-fold and the number of countries with cases increased threefold. In this paper a time series model to predict short-term prediction of the transmission of the exponentially growing COVID-19 time series is modelled and studied. Auto Regressive Integrated Moving Average (ARIMA) model prediction is performed on the number of cumulative cases over a time period and is validated over Akaike information criterion (AIC) statistics.

1.Introduction

The exponentially growing COVID-19 time series data can be modelled and studied using Auto Regressive Integrated Moving Average (ARIMA) model[1] This model can be used to do short term prediction[2]. The data is taken from the Johns Hopkins university (https://gisanddata. maps.arcgis.com/apps/opsdashboard/ index.html) are useful because they can provide a forecast for COVID-2019 pandemic to effectively control the spread of this highly infectious disease in India. Depending on the predictions, the government officials should adapt aggressive interventions to control this exponential growth [3] of this rapid infectious disease and curtail the COVID-19 pandemic. The updation of Johns Hopkins university on a daily basis is considered and the data for India till 14 th April 2020 is considered for this analysis and a time-series database was created in Excel. Exploratory data analysis of the data was performed to predict on a short term prediction of confirmed cases of COVID-19 in India for the next 10 days is predicted effectively.

Keywords

Covid-19, ARIMA Model, Time Series, Short term prediction

2 . Methodology

ARIMA forecasts on its previous past values and there are three distinct integers (p, d, q) that are used to parametrize ARIMA models. The three parameters account for seasonality, trend, and noise in datasets are denoted with the notation ARIMA(p, d, q). In the model, ***p*** is the auto-regressive part of the model and incorporates the effect of past values in the model. ***d*** is the integrated part of the mode and incorporates the amount of differencing to apply to the time series. The parameter ***q*** is the moving average parameter that allows to set the error of the proposed model as a linear combination of the error values observed at previous time points in the past. Our goal is to that optimizes the metric of interest. The experiment is carried out in R Programming. The Fig 5. Plots the cumulative cases from the John Hopkins University dataset for India.

The augmented Dickey-Fuller (ADF) test is a formal statistical test done to ensure stationarity. In ARIMA modeling using R the univariate data is converted into time series data format. The graph follows an overall upward trend with some outliers in terms of sudden lower values. The Augmented Dickey Fuller Test (ADF) is unit root test for stationarity. Since the data is not stationary, the data is differenced and is computed by the differences between consecutive observations.

```
Augmented Dickey-Fuller Test

data:  tsdata1
Dickey-Fuller = -4.4522, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Since the p-value after differencing is 0.01 and is less than 0.05 the null hypothesis is rejected and the data do es not have a unit root and is stationary. The time series after the data is removed of its non stationarity is giv en in the Fig 1.

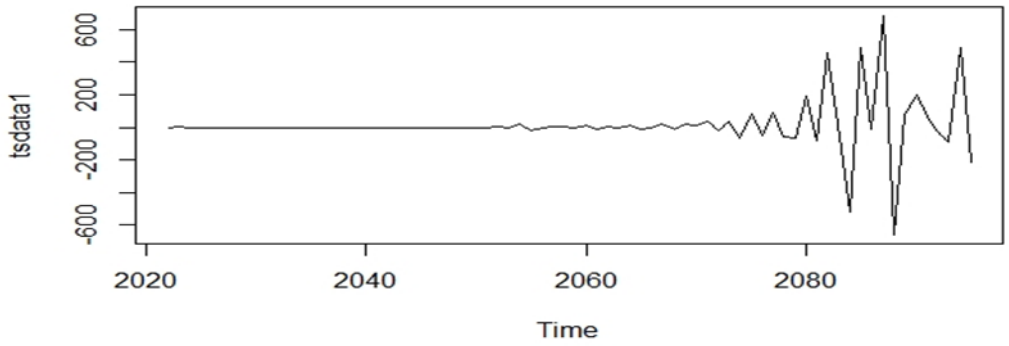

Fig 1. Non Stationarity removed Time series

The autocorrelation function gives the autocorrelation at all possible lags. The autocorrelation at lag 0 is included by default which always takes the value 1 that represents the correlation between the data and themselves. The ACF and the PACF is given in Fig.2

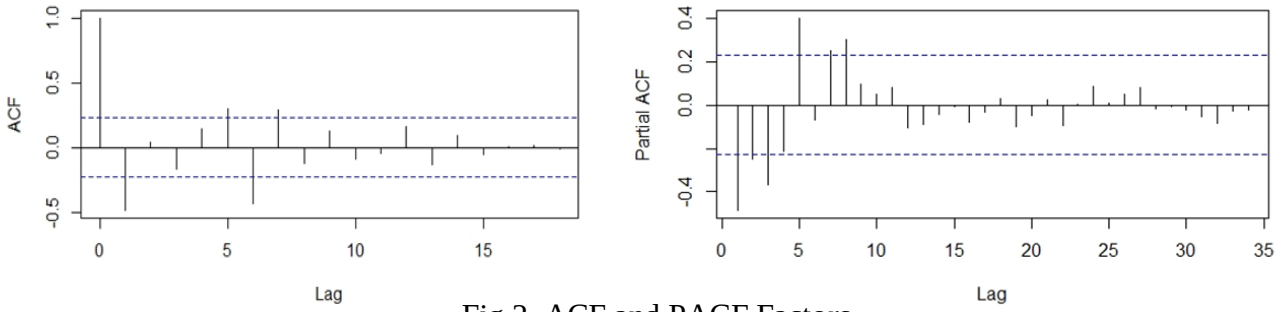

Fig 2. ACF and PACF Factors

The best fit model is selected based on Akaike Information Criterion (AIC) value of the model. The idea is to choose a model with minimum AIC and BIC values.The best model is ARIMA(1,2,2) with the AIC value of 932 and the the BIC value of 941is fitted using the auto.arima() function.

```
ARIMA(1,2,2)

Coefficients:
         ar1      ma1     ma2
      0.8294  -1.7021  0.8534
s.e.  0.1894   0.1144  0.0606

sigma^2 estimated as 15786:  log likelihood=-462.22
AIC=932.43   AICc=933.01   BIC=941.65
```

The forecast for the next 10 days is given below,

Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
77	12500.63	12339.61	12661.65	12254.38	12746.89
78	13545.02	13302.38	13787.66	13173.94	13916.11
79	14614.92	14285.47	14944.38	14111.06	15118.79
80	15705.98	15272.63	16139.34	15043.23	16368.74
81	16814.59	16254.71	17374.48	15958.32	17670.87
82	17937.76	17226.03	18649.49	16849.26	19026.26
83	19073.00	18183.10	19962.89	17712.02	20433.98
84	20218.25	19123.78	21312.72	18544.41	21892.09
85	21371.81	20046.78	22696.83	19345.36	23398.26
86	22532.25	20951.37	24113.13	20114.50	24950.00

The forecast of the Confirmed Cases for the next 10 days, that is, until 24th April 2020. The Forecast reaches 22532 on the 86th day, that is, by 24th April 2020 if proper social distancing and other measures are not followed. The Figure plot shows the predicted cases. The blue line represents the forecast and the silver shade around it represents the confidence interval.

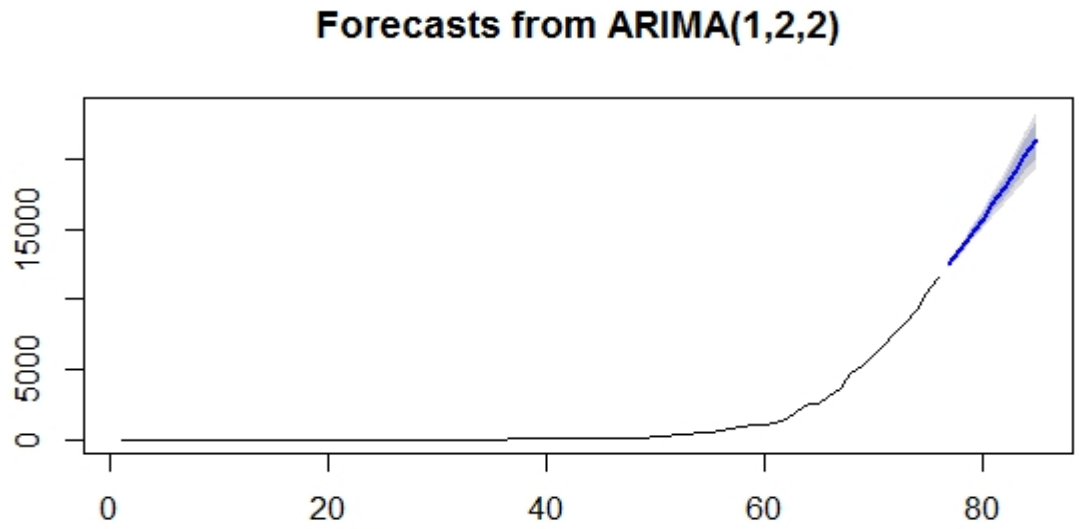

Fig 3. ARIMA Forecast of confirmed cases of COVID-19 in India

This Ljung-Box Q test assess the overall randomness based on a number of lags, and is therefore a portmanteau test. It is applied to the residuals of a fitted ARIMA model, not the original series, and in such applications the hypothesis actually being tested is that the residuals from the ARIMA model have no autocorrelation.

```
Box-Ljung test

data:  forecast2$residuals
X-squared = 0.052651, df = 1, p-value = 0.8185
```

The different metrics to evaluate the model is given in the following Table 1. That shows significant performance values.

S.No	Metrics	Value
1.	Mean Error	17.72206
2.	Root Mean Square Error	121.4391
3.	Mean Absolute Error	47.42588
4.	Mean Percentage Error	7.768734
5.	Mean Absolute Scaled Error	0.3096762
6.	Autocorrelation of errors at lag 1	-0.02580961

Table 1. Performance Metrics of ARIMA Model

This model assume that the population of COVID-19 and the infected but not yet isolated population have the same contact rate. The factors like lock down, social distancing, wearing of masks and usage of sanitizers are not considered while modelling and predicting the confirmed cases of COVID-19.Also the continuous release of the epidemic data there might be changes in the spread of of COVID-19 among the population.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

References

[1] D.Benvenuto, M. Giovanetti, L.Vassallo, S. Angeletti, M. Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in brief*. 105340 (2020). https://doi.org/10.1016/j.dib.2020.105340

[2] S.Deb, M. Manidipa. A time series method to analyze incidence pattern and estimate reproduction number of covid-19, *arXiv preprint arXiv:2003.10655* (2020).

[3] R. Gupta, P. Saibal Kumar ,Trend Analysis and Forecasting of COVID-19 outbreak in India, *medRxiv* (2020).

[4]. World Health Organization (WHO). Novel coronavirus (COVID-19) situation [Internet]. Geneva: WHO; 2020 [cited 2020 Mar 18]. Available from: https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd.