MAJOR PROJECT REPORT

ON

"Covid-19 Time Series Forecasting"

Submitted in partial fulfilment of the

requirement for the award of the degree of

Bachelor of Computer Application

Submitted To

Dr. Neha Goel

Assistant Professor

Submitted By

Vineet Gupta

14517702018

BCA 6-C

(Batch: 2018-2021)

Vivekananda Institute of Professional Studies

(Affiliated to Guru Gobind Singh Indraprastha University)

CERTIFICATE

This is to certify that Vineet Gupta of BCA 6th Semester from Vivekananda Institute of Professional Studies, Delhi has presented this MAJOR PROJECT REPORT entitled “Covid-19 Time Series Forecasting” in partial fulfilment of the requirements for the award of the degree of Bachelor of Computer Applications under our supervision and guidance.

Dr. Neha Goel

Assistant Professor

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my college that gave me the golden opportunity to do this wonderful project on the topic “Covid-19 Time Series Forecasting” which also helped me in doing a lot of research and I came to know about so many new things. I am thankful to them.

Secondly, I would also like to thank Coursera.org and Datacamp.com that offered all the necessary courses which helped me to learn so many new things.

The project has offered me a great opportunity to grow and develop. It has propelled me to be able to overcome challenges and develop my career. I learnt extensively about data manipulation, visualization, tidy data, and time series. The program has enhanced my analytic thinking and skills as well as improved my professional skills and ability to work in a multicultural environment. Working on this was not only an honor and privilege but a lifelong experience that will forever shape my professional life.

Vineet Gupta

Table of Contents

Introduction

Objective:

The objective of this project is to analyze the time-series of SARS-CoV-2/Covid-19 newly infected cases and study the trends to forecast expected future trends in India.

About COVID-19:

Corona viruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and severe acute respiratory syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19.

COVID-19 is the infectious disease caused by the most recently discovered corona virus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019.

Summary:

Data is retrieved from a website which contains global data. Global data is cleaned, and new data has the variables Date and Total number of Confirmed Cases in India.

We use this data to predict number of infected for further 2 weeks.

Platform/Software

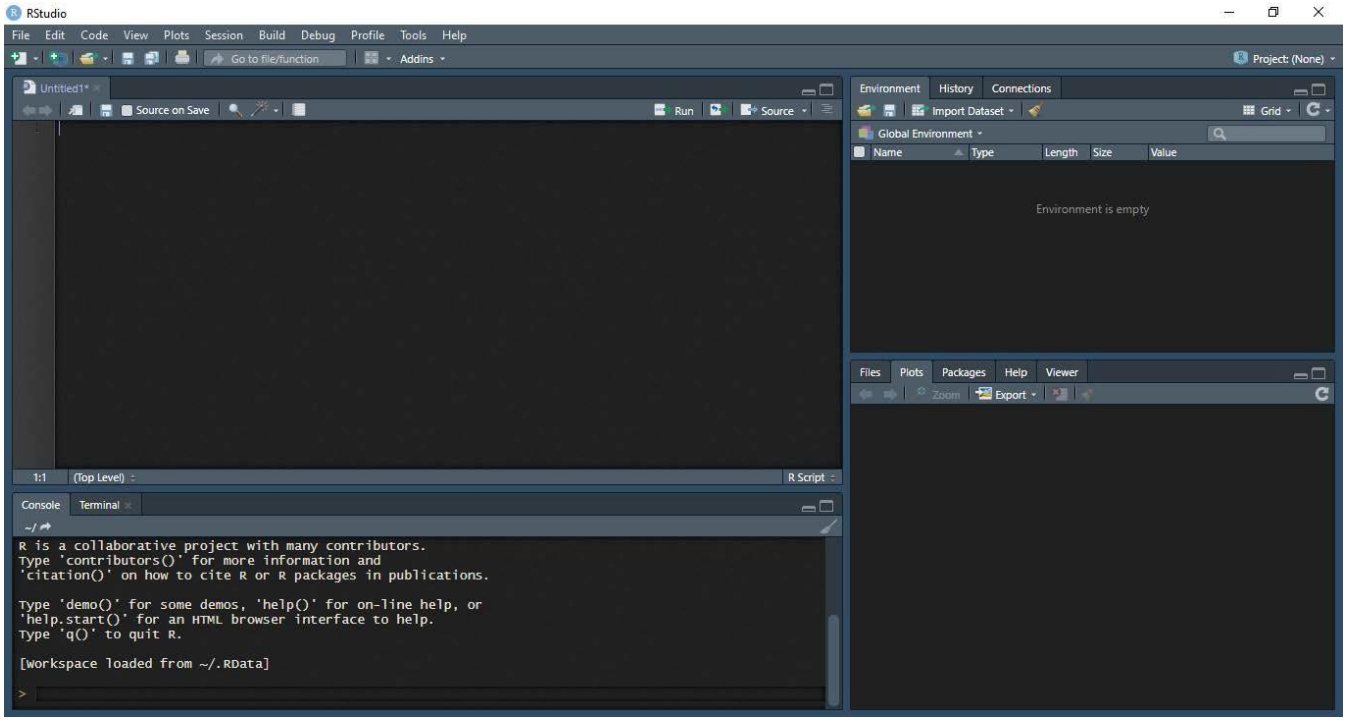
R Programming

R is a language and environment for statistical computing and graphics.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

RStudio

RStudio is dedicated to sustainable investment in free and open-source software for data science, to help people understand and improve the world through data.



Theory

Basic time series models:

1. White Noise (WN)
2. Random Walk (RW)
3. **Autoregression (AR)**
4. **Moving Average (MA)**

Basic assumptions:

1. Consecutive observations are equally spaced.
2. Apply a discrete-time observation index.
3. This may only hold approximately.

Why ts() objects?

1. Improved plotting.
2. Access to time index information.
3. Model estimation and forecasting.

ARIMA (Autoregressive Integrated Moving Average):

An autoregressive integrated moving average model is a generalization of an autoregressive moving average model. Both models are fitted to time series data. It basically combines the two main types of time series models.

Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

Sources of data

OurWorldInData.rg
Link: https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv

Trainings and Libraries

1. **Data Manipulation (library: dplyr)**

Data manipulation involves modifying data to make it easier to read and to be more organized. We manipulate data for analysis and visualization. It is also used with the term 'data exploration' which involves organizing data using available sets of variables.
At times, the data collection process done by machines involves a lot of errors and inaccuracies in reading. Data manipulation is also used to remove these inaccuracies and make data more accurate and precise.

2. **Data visualization (library: ggplot2)**

Data is visualized using R includes data plotting with R's open source library ggplot2. After an introduction to base graphics, we look at several R plotting examples, from simple graphs such as scatterplots to plotting correlation matrices. This includes using R plot colors effectively and creating and saving complex plots in R.

R supports four different graphics systems:
base graphics, grid graphics and lattice graphics

3. **Time Series Modelling (library: stats)**

Time Series is a sequence of data in chronological order. Data is commonly recorded sequentially, over time. Modelling involves the estimation of parameters of time series, minimizing AIC (Akaike Information Criteria), also keeping in mind that number of parameters are not excessive, otherwise it would result in more complex model and over-modelled data.

4. **Time Series Forecasting (library: forecast)**

Time Series Forecasting involves estimation of future values using the parameters estimated while modelling and providing a confidence interval (80% and 95%) for decision making.

Project Documentation

COVID-19 TIME SERIES FORECASTING (India)

Objective: The objective of this project is to analyse the time-series of SARS-CoV-2/Covid-19 confirmed and death cases and study the trends to forecast expected future trends in India.

The following libraries were used:

```
library(dplyr) #data manipulation

## Warning: package 'dplyr' was built under R version 4.0.5

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(forecast) #forecast()

## Warning: package 'forecast' was built under R version 4.0.5

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

library(ggplot2)
```

Raw data is imported and manipulated as per requirements. It is ensured that the datatype in the imported file is interpreted exactly as what it was before, or at least what we want it to be interpreted as.

```
##working with data
#importing data
data.world<-read.csv("https://raw.githubusercontent.com/owid/covid-19-
data/master/public/data/owid-covid-data.csv",sep = ",",header = T)
#str(data.world)
```

Raw data is cleaned and only the required variables are retrieved.

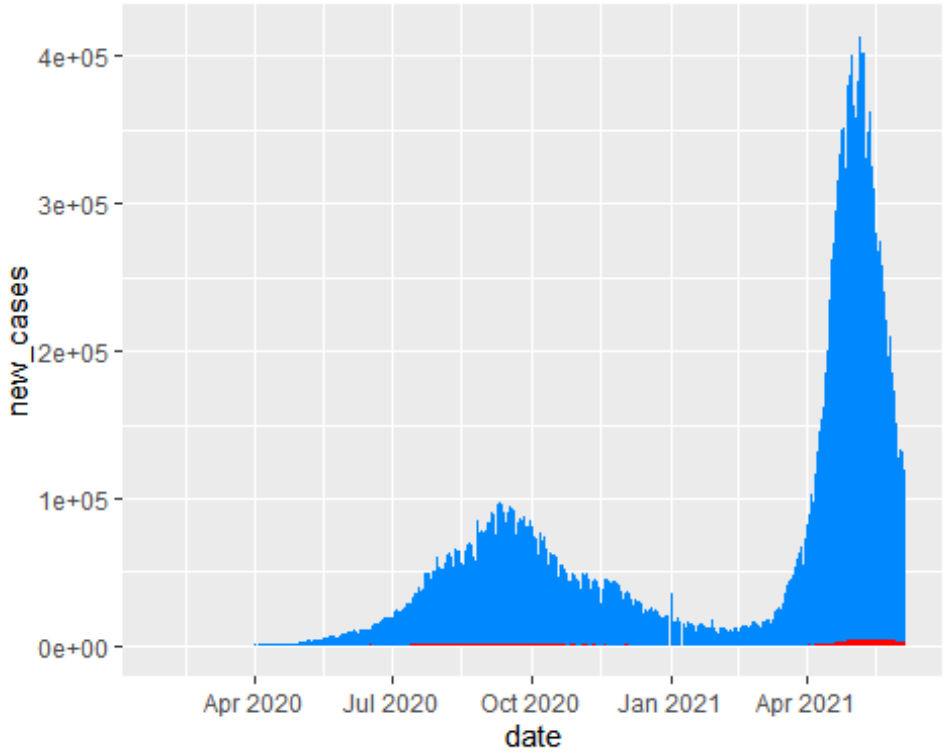
```
#preprocessing data
{
data.india<-data.world %>%
  filter(location=="India") %>%
  select(date,new_cases,new_deaths)
data.india$date<-as.Date(data.india$date)
n<-nrow(data.india)
```

```
str(data.india)
}
```

```
## 'data.frame':    493 obs. of  3 variables:
##  $ date      : Date, format: "2020-01-30" "2020-01-31" ...
##  $ new_cases : num  1 0 0 1 1 0 0 0 0 0 ...
##  $ new_deaths: num  NA NA NA NA NA NA NA NA NA NA ...
```

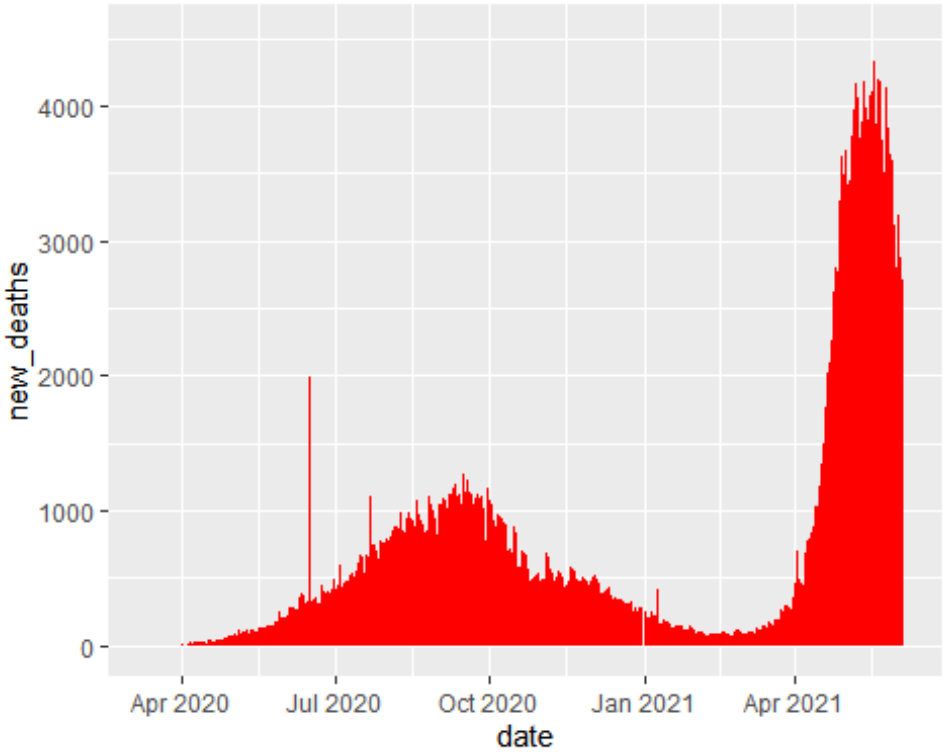
```
ggplot(data.india)+
  geom_area(aes(x=date,y=new_cases),fill="#0088FF")+
  geom_area(aes(x=date,y=new_deaths),fill="#FF0000",na.rm = TRUE)
```

```
## Warning: Removed 41 rows containing missing values (position_stack).
```



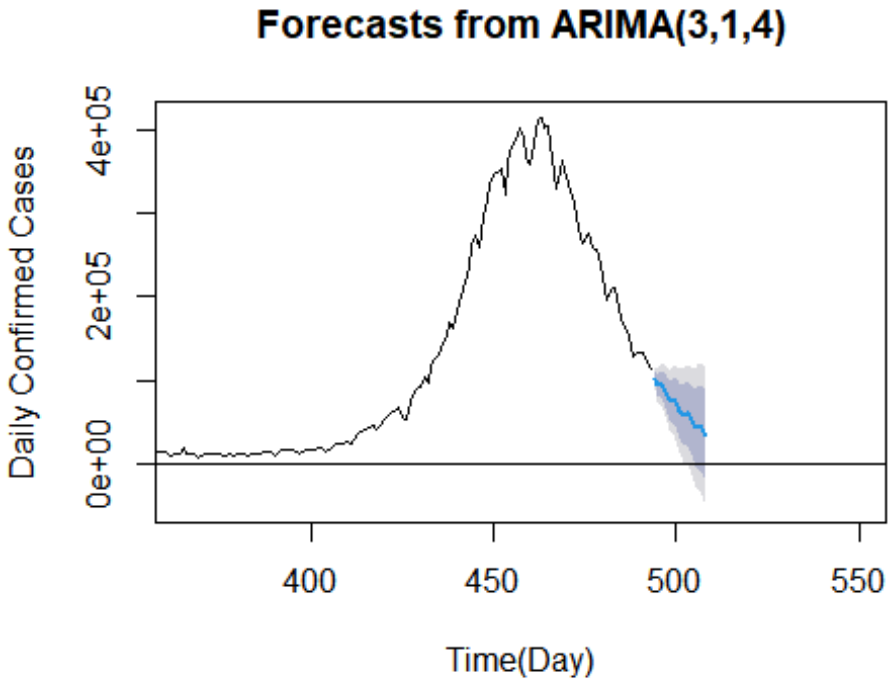

```
ggplot(data=data.india)+
  geom_area(aes(x=date,y=new_deaths),fill="#FF0000",na.rm = TRUE)
```

```
## Warning: Removed 41 rows containing missing values (position_stack).
```



1. The first model represents ARIMA model for daily new cases.

```
model.1<-auto.arima(data.india$new_cases); model.1

## Series: data.india$new_cases
## ARIMA(3,1,4)
##
## Coefficients:
##          ar1      ar2     ar3      ma1     ma2      ma3     ma4
##       0.4759  -0.5008  0.9384  -0.3648  0.4571  -0.8281  0.0733
## s.e.  0.0226   0.0219  0.0220   0.0522  0.0343   0.0350  0.0482
##
## sigma^2 estimated as 47107802:  log likelihood=-5042.15
## AIC=10100.3   AICc=10100.6   BIC=10133.89

model.1.forecast<-forecast(data.india$new_cases,model = model.1,h = 15)
head(model.1.forecast$mean)

## Time Series:
## Start = 494
## End = 499
## Frequency = 1
## [1] 102752.46  94384.96  95358.08  88545.12  76963.36  75777.48

{
plot(model.1.forecast,xlab="Time(Day)",ylab="Daily Confirmed
Cases",xlim=c(365,550),
     main = cat("Forecasted New Cases till ",end_date))
abline(h=0)
}
```

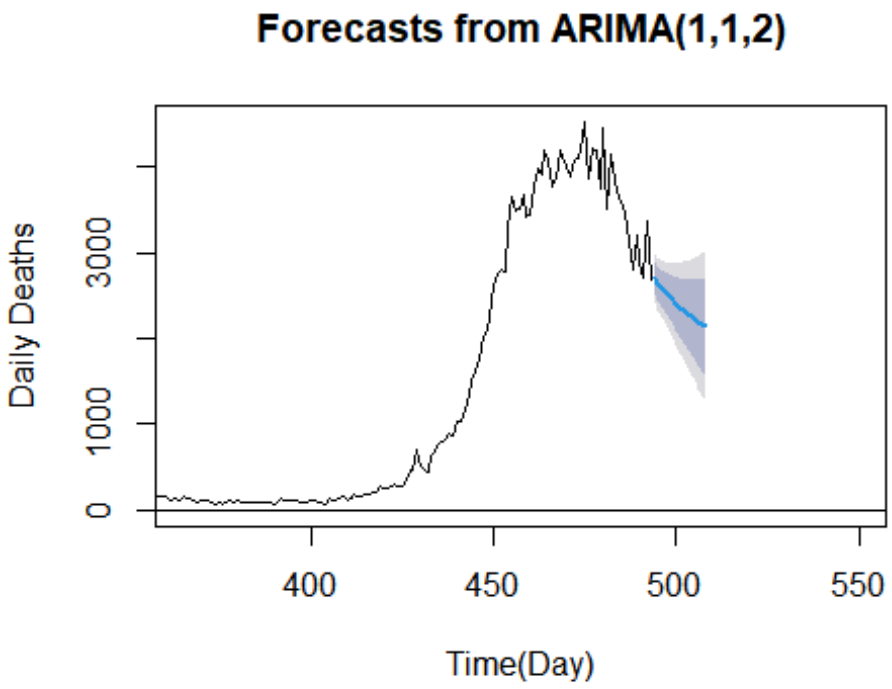


The model suggested is ARIMA(3,1,4); 3 parameter for Autoregression, 1 parameter for difference and 4 parameters for Moving Average.

2. The second model represents ARIMA model for daily new deaths.

```
model.2<-auto.arima(data.india$new_deaths); model.2

## Series: data.india$new_deaths
## ARIMA(1,1,2)
##
## Coefficients:
##          ar1      ma1     ma2
##       0.9391  -1.5916  0.6879
## s.e.  0.0262   0.0387  0.0377
##
## sigma^2 estimated as 21562:  log likelihood=-2889.25
## AIC=5786.5   AICc=5786.59   BIC=5802.94

model.2.forecast<-forecast(data.india$new_deaths,model = model.2,h = 15)
head(model.2.forecast$mean)

## Time Series:
## Start = 494
## End = 499
## Frequency = 1
## [1] 2708.569 2648.853 2592.773 2540.109 2490.651 2444.206

{
plot(model.2.forecast,xlab="Time(Day)",ylab="Daily Deaths",xlim=c(365,550),
     main = cat("Forecasted New Deaths till ")
```

```
abline(h=0)
}
```



The model suggested is ARIMA(1,1,2); 1 parameter for Autoregression, 1 parameter for difference and 2 parameters for Moving Average.

```
#end of documentation
```

CONCLUSION

1. Project Conclusion

- A case study on the spread of the novel Coronavirus (Covid-19), in which we observe the trends of daily newly infected cases through time-series.
- Daily cases and deaths were modelled through Autoregressive Integrated Moving Average modelling for further forecasting.
- The predicted numbers are displayed along with their confidence interval at 80% and 95% confidence level.

2. Limitations

The model suggested does not take into the account the external factors that are responsible for the trends.

External factors involve:

- government policies
- effect of vaccination
- the fact that the number of newly infected data cannot exceed the number of tests conducted, which basically means we might not even have real data.

This suggests that the time-series forecasting can be used only for the immediate future and when the conditions are identical and independent.

3. Future scope of the project

- A live data feeding can be added using data scrapping, and the predicted visuals can be displayed on a website.
- The timeline of the external events can be added at their respective times to observe how these external factors, if they do, change the trend.

SUNY The State University of New York

COURSE CERTIFICATE

Jan 4, 2021

Vineet Gupta

has successfully completed

Practical Time Series Analysis

an online non-credit course authorized by The State University of New York and offered through Coursera

Tural Sadigov
Lecturer
Applied Mathematics

William Thistleton
Lecturer
Applied Mathematics

Verify at coursera.org/verify/MSZKK7YNCN8E

Coursera has confirmed the identity of this individual and their participation in the course.