Covid-19 Pandemic Spread as Growth Factor using Forecasting and SIR Models

Gudimella Sai Vennela$^{1,2}$ and Priyanka Kumar$^{1,3}$

$^1$Dept of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India

$^2$cb.en.p2aid19011@cb.students.amrita.edu

$^3$k_priyanka@cb.amrita.edu

Abstract. The novel Coronavirus outbreak began in Wuhan which is located in China in December 2019 and after the outbreak immediately the virus has spread across the globe. Millions of people across the globe are infected and many more have died. World Health Organization (WHO) declared this virus as global emergency. In this work, main goal is to analyse the spread of virus in India that to particularly for Guntur District within India in order to settle the situation quickly. In order to analyse the spread of virus three types of forecasting methods have been compared and the SIR model is used to predict the number of Susceptible, Infected and Recovered persons for the Guntur district which is located in Andhra State.

1. Introduction

The Covid-19 pandemic was identified on December 31st, 2019 in Wuhan city, Hubei province, China reported that the group of people are affected by the virus known as SARS-Cov. This virus spreads through many ways but in the present-day scenario person-to-person contact is the main way which causes the virus to spread widely across the nations. Mainly it starts with the cold and fever and gradually the virus results in the failure of respiratory system which ultimately leads to person’s death [1][2][3][4][5]. At the initial stage of pandemic, the two big nations Italy, US are badly affected by virus [6][7][8]. In India reportedly the viruses have its spread from 30th January 2020.Ever since outbreak in India the virus is showing the adverse effects leading to third largest country in the world with the highest recorded confirmed cases after US and Italy. On an average in India daily 70,000 active cases are recorded. In order to support the prevention of the disease and aid in the healthcare service preparation, we have conducted this study to examine the finest model for the prediction of confirmed COVID-19 infection cases and to employ that model for forecasting future COVID-19 infection cases in India. As per the model forecast, the confirmed cases are expected to greatly rise in the coming days. The time series analysis shows an exponential enhancement in the infected cases [13].

There are many Forecasting models [12] out of which three models have been used for this analysis. This paper mainly aims to compare the results in terms of RMSE error for the three forecasting methods like Holt’s Winter’s, ARIMA and Auto-ARIMA models. Apart from this in order to predict the infection and recovery rate in terms of population, Compartment or Disease model namely SIR Model [9][10] is used for the analysis of Guntur District. In previous work [11] they have predicted the high incidence states using SIR Model but they did not mention regarding infection, recovery rate and reproduction number which are very important.so in this work we have given the
results for the infection, recovery rate and reproduction number by taking the population size as 48.9 lakhs (i.e.; population in Guntur District).

This paper consists of four sections in total. Section 2 is based on the methodology; section 3 is on discussion of results and inferences from models. Finally, in section 4 this work culminates with conclusion and future scope.

2. Methodology
In this section a detailed description of all the Forecasting Models and SIR Model have been provided.

2.1. Dataset

| Date       | State                  | District   | Confirmed | Recovered | Deceased | Other | Tested  |
|------------|------------------------|------------|-----------|-----------|----------|-------|---------|
| 0 2020-04-26 | Andaman and Nicobar Islands | Unknown   | 33        | 11        | 0        | 0     | 2679.0  |
| 1 2020-04-26 | Andhra Pradesh | Anantapur | 53        | 14        | 4        | 0     | NaN     |
| 2 2020-04-26 | Andhra Pradesh | Chittoor  | 73        | 13        | 0        | 0     | NaN     |
| 3 2020-04-26 | Andhra Pradesh | East Godavari | 39   | 12        | 0        | 0     | NaN     |
| 4 2020-04-26 | Andhra Pradesh | Guntur     | 214       | 29        | 8        | 0     | NaN     |

**Figure 1.** Dataset sample with districts in India

Dataset districts.csv used in this work is taken from api.India.org. This dataset contains district wise Confirmed, Recovered and Deceased cases from April 26 to August 30 as shown in Figure 1. The dataset has the day level information of these cases. And the column level information is shown in the below Table.1.

**Table 1.** Column level information in the dataset

| Columns | Description |
|---------|-------------|
| Date    | Date of observation in the format of YY-MM-DD |
| State   | State of observation |
| District | Within the States there are districts of observation |
| Confirmed | Till the date (30th August) cumulative number of confirmed cases |
| Recovered | Till the date (30th August) cumulative number of confirmed cases |
| Deceased | Till the date (30th August) cumulative number of confirmed cases |
| Tested  | Number of Testing happen in states |

2.2. Proposed Work

2.2.1. Pre-Processing
The flowchart mentioned in Figure 2, gives the information of the entire work where we began with the pre-processing stage. In this stage we will check whether the dataset contains any null values or the columns which are not useful for our analysis.

**Figure 2.** Flowchart for Proposed Work
2.2.2. Exploratory Data Analysis

Exploratory data analysis is well known as EDA is one of the important steps in data analysis. It gives the clear analysis of patterns, summarize the main characteristics of data with the visual methods. In original dataset there are no active cases mentioned so i created the column active cases as Active = Confirmed-Recovered-Deaths.

Note: All the forecasting graphs are implemented using the plotly-dark library because if we use the plotly-white library the variations are not clearly visible for prediction. There are so many plots for performing EDA, due to the space constraint only three plots are included.

Now if we look further, Maharashtra is the one of the highly affected state with the confirmed, recovered, deceased, active cases. Ever since the outbreak Maharashtra is only the state that ranks high till the date. One of the reasons why the cases are not slowing down in this state is because of urbanization. Table.2 shows the top five highly affected shows the due to pandemic. And Statewide corona cases are shown in the Figure 3. Figure 4 shows the recovery and mortality rate and corresponding Table.3 show the values of mean and median recovery and mortality rate from all the states.

| State            | Confirmed  | Recovered | Active    | Deceased |
|------------------|------------|-----------|-----------|----------|
| Maharashtra      | 32991815   | 20678398  | 11132556  | 1180861  |
| Tamil Nadu       | 17804787   | 13421194  | 4103482   | 280111   |
| Andhra Pradesh   | 11380858   | 8253849   | 3794615   | 109680   |

Figure 3. Statewide corona cases
Table 3. Mean, Median Mortality and Recovery rate

|                      | Value                      |
|----------------------|----------------------------|
| Mean Recovery Rate   | 54.255911481607264         |
| Mean Mortality Rate  | 1.427615748355606          |
| Median Recovery Rate | 52.89984554347497          |
| Median Mortality Rate| 1.2717691962974982         |

2.2.3. Calculation of Growth Factor

Growth factor gives the information of the new cases like whether the cases are increasing or decreasing. The main assumption is if the growth factor is greater than 1 then each day there is significant increase in the number of new cases. If the growth factor is less than 1 then the number of new cases slowly getting down and the situation is under control. Growth Factor is calculated by taking today’s new cases divided by yesterday’s new cases. Whereas if the growth factor for the recovered cases is high, which gives the intuition like when compared to deaths and confirmed cases the number of persons recovered are high especially at the end of April for Guntur district shown in Figure 5.
2.2.4. **Forecasting Methods**

Basically, for every Forecasting model there are two main assumptions. (1) Historical information is available. (2) Past patterns continue in the future. Based on these two assumptions three types of Forecasting models have been discussed.

2.2.4.1. **Holt’s Winter’s Model**

Holt’s winter’s model is also known as Triple exponential smoothing. In this model there are three main components 1) Level or Intercept 2) Slope or Trend and 3) Seasonality. The third component seasonality tells us the repeated variation around the Intercept and Trend. For example, if we say length of season is seven then it indicates weekly seasonality.

\[
\text{Predict}_{\text{max}}(x) = I(x - 1) + T(x-1) + S(x - t) + m.d(t - T) \quad (1)
\]

\[
\text{Predict}_{\text{min}}(x) = I(x - 1) + T(x-1) + S(x - t) - m.d(t - T) \quad (2)
\]

\[
d(t) = \gamma(\text{predicted} - \text{actual}) + (1-\gamma)d(t - T) \quad (3)
\]

Where I is intercept, T is trend, S is seasonality, m is arbitrary constant, d is predicted deviation, T is length of the season, gamma is the seasonal change in the smoothing factor. There are mainly two variations in the Holt’s winter model based on the nature of the seasonality component. One is the Additive method which is useful when in the seasonality is constant. Second one is based on the multiplicative method which is useful when there is change in the seasonality component. Here, in this work data is undergoing seasonal changes, so we are going to the multiplicative method.

2.2.4.2. **ARIMA Model**

In the ARIMA model the first two letters AR stands for Auto regression which means present values are dependent on the past or previous values with some lag. The maximum lag is indicated by p. The last two letters MA stands for moving average which means the present error values depend on the past or the previous error values with some lag. The maximum lag is indicated by q. In this model I stands for order of integration of non-seasonal difference. This difference term is indicated by d.

\[
y = \text{present\_values}(Y_t) - \text{past\_values}(Y_{t-1}) \quad (4)
\]

2.2.4.3. **Auto-ARIMA Model**

The main difference between the Arima and Auto-Arima model is there is no need of calculating p.q.d values instead these values are estimated by using the two statistical models known as AIC and BIC. Two statistical models are used in the Auto ARIMA model, are AIC (Akaike information criteria) and BIC (Bayesian information criteria). AIC which is an estimation of constant plus the relative distance
between the true probability of the unknown data and the model’s equipped probability function, if the model has the lower AIC then that model is closer to the true values. In BIC the approximation of a function of the posterior likelihood of a model is being valid, so a lower BIC means that a model is assumed to be more likely to be the true model.

\[
\text{AIC} = n \ln(\text{sum_of_the_squared_error}_n) + 2k \tag{5}
\]

\[
\text{BIC} = n \ln(\text{sum_of_the_squared_error}_n) + k \ln(n) \tag{6}
\]

2.2.5. SIR Model

The main idea behind this SIR model is let’s say take the population of Guntur district (48.9 lakhs) and divide the population into 3 categories (1) susceptible s(t) (2) Infected i(t) (3) recovered r(t) shown in Figure 6. Susceptibility means the people who are all capable of getting infection or the people who are capable of infecting everybody. In this context infection is nothing but the coronavirus. Second category is of Infected people, when a person gets infected, he has to leave the category of Susceptibility. And the third category is those who are all infected they are going hopefully into the recovered status at some time.

\[
s(t) + i(t) + r(t) = \text{total_population} \tag{7}
\]

We all know that at the beginning of the pandemic there are no susceptible, no infected and no recovered people. So, the initial conditions can be taken as follows

\[
s(0) = S_0, i(0) = i_0, r(0) = r_0 \tag{8}
\]

Here our main concern is to find the rate of change with respect to time, like how the number of susceptible, infections and recoveries change as the time changes. We can clearly know that all these three variables are dependent on the time. If more people are interacting then there is more chance for s and i getting multiplied resulting in more infections. Therefore, the time dependent equations for the susceptible, infected and recovered are as follows

\[
\begin{align*}
\frac{ds}{dt} &= a(s \ast i), \\
\frac{di}{dt} &= a(s \ast i) \ast (b \ast i), \\
\frac{dr}{dt} &= (b \ast i)
\end{align*} \tag{9}
\]

In equation (9) a is known as infection rate which means the rate of transmission of virus between the susceptible and the infected person and b is known as the recovery rate. And for any SIR model the basic reproduction number R is calculated as shown in equation (10). This R0 reproduction number tells us in a total population, people who are vulnerable to the infections for which the expected number of cases are directly generated from the one case.

\[
R_0 = a/b \tag{10}
\]

![SIR Model](image)

**Figure 6. SIR Model**
3. Analysis and Results
This section elucidates model architecture and the quality metrics which is used to analyse the best suitable models from the Forecasting (as mentioned in section 2.2.4). On an average for the estimated period of time results are predicted for the SIR model.

3.1. Model Architecture

3.1.1. Forecasting Models
In order to get the results using Forecasting models first we split the dataset into training and validation sets from the last 95 periods from the test data. However, the variations in seasonality are increasing so here seasonality is taken as multiplicative variation. The Fig.7,8,9 shows the predicted values for the Confirmed, Recovered and Deceased cases for the Guntur district in India using Holt’s winter's, Arima model and Auto-Arima model. The Table.4 shows the comparative analysis of all the three models. Out of all the three models Auto-Arima model shows the good predicted results which are nearly equal to the actual values, because we are using two variations namely AIC, BIC as mentioned in 2.2.4.3.

| Model          | Confirmed (actual) | Confirmed (predicted) | Recovered (actual) | Recovered (predicted) | Deaths (actual) | Deaths (predicted) |
|----------------|--------------------|-----------------------|--------------------|-----------------------|----------------|-------------------|
| Holt’s Winter’s| 35761              | 40610                 | 27799              | 31320                 | 376            | 396               |
| ARIMA          | 35761              | 37195                 | 27799              | 30588                 | 376            | 390               |
| Auto-ARIMA     | 36761              | 36799                 | 27799              | 30299                 | 376            | 382               |

Table 4. Comparison between Forecasting Models
Figure 7. Prediction of results for Holt’s Winter’s Model
**Figure 8.** Prediction of results for ARIMA Model
3.1.2. Quality Metrics

The metrics used in this work is root-mean-square-error (RMSE). It takes the difference between the actual values and predicted values by the model. Figure 10 Shows the RMSE values for the three forecasting models.

![Figure 9](image1.png)

**Figure 9.** Prediction of results for Auto-ARIMA Model

| Models/Different Cases | Confirmed | Recovered | Deceased |
|------------------------|-----------|-----------|----------|
| Holt’s Winter’s Model  | 7030.11   | 1993.55   | 11.679   |
| ARIMA Model            | 616.299   | 1589.20   | 9.168    |
| Auto-ARIMA Model       | 468.583   | 1394.44   | 8.994    |

![Figure 10](image2.png)

**Figure 10.** Error comparison between Forecasting Models
3.1.3. SIR Model

As discussed in section 2.2.5 a number of cases will be increased if there is direct contact between the susceptible and infected. For this analysis the cumulative confirmed, recovered and deceased cases for Guntur district are taken. The data is trained on this cumulative case for 131 days and the results for the infection rate $a$, recovery rate $b$, reproductive number $R_0$, confirmed, recovered, infected cases for the next day and total confirmed cases in Guntur district are shown in the Table 5. The actual infection rate, predicted infection rate and actual recovery, predicted recovery rates, finally the rate of change of cases with respect to time for SIR model is shown in Figure 11.

![Graph](image)

**Figure 11. Predictions for Guntur District using SIR model**

**Table 5. SIR Model**

| Parameter                        | Value   |
|----------------------------------|---------|
| Infection Rate of SIR Model      | 0.08769 |
| Recovery Rate of SIR Model       | 0.05598 |
| Reproduction Number              | 1.56643 |
| Confirmed cases on 132 day       | 28144   |
| Recovered cases on 132 day       | 18453   |
| Infected People                  | 9692    |
| Total Confirmed cases in Guntur district | 68418 |
4. Conclusion and Future Scope

In this work, three Forecasting Models and the disease model SIR have been analysed. The data is trained from 26 April to 30 August especially on Guntur district. Among the three Forecasting models (as mentioned in section 2.2.4) the predicted results for the Auto-ARIMA model are pretty much similar to the actual trained data. The quality metrics used for this analysis is RMSE. Form the SIR model it has been observed that reproduction number (as mentioned in Table 5) is greater than one, which means there will be significant increase in the number of infected persons, it’s like alert that number of infections will grow.

However, there are so many compartment models for the disease modelling problems. Hence, this work can be further extended to the other compartment models like SIS, SIRS, Age-Structured models (age is one of the important factors to be considered for the disease spread).

References
[1] J T Wu, K Leung and G M Leung 2020 Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, china: a modelling study The Lancet 395 (10225) (2020) p 689-697
[2] World Health organization, coronavirus disease (COVID-19) outbreak. https://www.who.int/emergencies/diseases/novel-coronavirus-2019,Google Scholar
[3] Novel coronavirus (covid-19) cases, provided by JHU CSSE https://github.com/CSSEGISandData/COVID-19.
[4] B Tang, X Wang, Q Li and N L Bragazzi, S. Tang, Y. Xiao, J. Wu Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions, Journal of clinical medicine, 9 (2) (2020), p. 462.
[5] M.U. Kraemer, C.H. Yang, B. Gutierrez, C.H. Wu, B. Klein, D.M. Pigott and J.S. Brownstein, The effect of human mobility and control measures on the COVID-19 epidemic in china, Science, 368 (6490) (2020), pp. 493-497.
[6] P Song, L Wang, Y Zhou, J He, B Zhu, F Wang and M. Eisenberg An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in CineMedia. Epub (2020), p. 3.
[7] J Wangping, H Ke, S Yang, C Wenzhe, W Shengshu, Y Shanshan and L Miao Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with hunan, china. Frontiers in Medicine, 7 (2020), p. 169.
[8] R Dandekar and G Barbastathis Quantifying the effect of quarantine control in covid-19 infectious spread using machine learning,medRxiv (2020), 10.1101/2020.04.03.20052084, Google Scholar.
[9] David Smith and Lang Moore: The SIR Model for Spread of Disease - The Differential Equation Model, Google Scholar.
[10] Ian cooper, Argha Mondal, Chris and G.Antonopoulos A SIR model assumption for the spread of COVID-19 in different communities, Chaos, Solitons Fractals, Volume 139, October 2020, 110057.
[11] B. Malavika and S. Marimuthu: Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models, Clinical Epidemiology and Global Health, Available online 27 June 2020.
[12] Classical Time Series Forecasting Methods in Python (Cheat Sheet) https://machinelearningmastery.com/time-series-forecasting-methods-in-pythoncheat-sheet/,Google Scholar.
[13] Hiteshi Tandon Coronavirus (COVID-19) ARIMA based time-series analysis to forecast near future Department of Chemistry, Manipal University Jaipur, Jaipur 303007, India