Time Series Forecasting of COVID-19 Infections in India

Sarthak Vishnoi (2016CS10336), Arundhati Dixit (2016ME10824)
April 30, 2020

Abstract

Predicting values requires history and there is still uncertainty because it is not necessary that the future replicates from the past. Despite that, we work on forecasts in order to anticipate dynamics and to get estimates in a situation where being overcautious and wrong is better than negligence caused underestimation. We collect the time series data of COVID-19 infections in India, analyze the data with an aim to establish the important characters of the underlying time series process and attempt to formulate time-series models and validate the same. We have formulated and analysed multiple models for short term and long term forecasts, with integral policy implications. We look at a number of forecasting methods to identify strengths and shortcomings of each model, and how the results from a committee of mathematical modelling can help pave way for policy planning.
Key words: Coronavirus, COVID-19, Time Series, India, District wise, Regression Model, Compartmental Model, Holt Exponential Smoothing, Forecasting

1 Introduction

Coronavirus disease 2019 (COVID-19) is a highly contagious disease, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This disease was first identified in Wuhan of China in December, 2019. Since then it has spread globally, resulting in COVID-19 pandemic. According to the Ministry of Health and Family Welfare, the 2019–20 coronavirus pandemic was reported in India on January 30, 2020. During the initial days, the frequency of COVID-19 infections was relatively low. However, a steep rise (+256%) was observed on March 4, 2020. As on April 30 2020, a total of 33,255 cases have been reported in India. Till this date, India is in Stage II of virus spread, indicating only local transmission is taking place in the country. Nevertheless, there is a tremendous fear that India may soon reach Stage III where the disease transmission would take place community wise. The last and worst stage of virus spread is Stage IV where the disease takes the shape of an epidemic as it happened in China. Therefore, almost everybody in the country is eager to know if India will remain in Stage II in the coming months or India will carry on to Stage III or even Stage IV. In this study, we aim to develop forecasting models to predict the number of cases India may anticipate in the days to come.

We first establish the regression models using VAR (Vector Auto Regression), ARIMA (Auto Regressive Integrated Moving Average) and OLS (Ordinary Least Square) estimators for the new cases reported in India, and use the fit model to forecast values for upcoming days. First, the national level analysis is done. Further, we add temperature and humidity to city level models to capture their effect, if any. The next model that we look at is Holt Exponential Smoothing, wherein we forecast number of cases, deaths and recoveries. We also demonstrate the forecasting exercise by updating the model every ten days and presenting forecasts. Lastly, we develop iterations of SEIR (Susceptible, Exposed, Infected, Recovered) model and highlight the importance of the Reproduction number, R_o in assessing the threat that COVID-19 poses. First, we formulate the deterministic SEIR with cases considering the average effects of lockdown to predict peak number of cases and the day when India could see the peak number of cases. Next, we explore a stochastic SEIR model which can accommodate R_o varying on a day to day basis. Lastly, we relax the assumption of homogeneous interaction in the population by incorporating age specific contact matrices for a more accurate model and to look at disease dynamics in population of different ages.

We draw key results and insights from the activities described above. We compare and contrast the forecasting abilities of these models, and how each of them have a direct impact on policy planning.

2 Data and Implementation

All the data compiled and condensed by us can be found on `https://github.com/SarthakVishnoi01/TSF-Challenge`. The link also has all the codes that have been used to generate results presented in the submission. All the methods are implemented using RStudio version 1.2.5033, along with R version 3.5.1. The data used is till 25th April 2020 (1) and forecasting starts from 26th April. We have cited data and modelling sources, if any, with every model step and description.

3 Modelling

3.1 Regression for cases at National Level

We analyse the number of daily cases which are being reported in the country (1) and fit models which best explain the trends in the time series and forecast for the upcoming days. Regression model needs to be updated after a few days because of the increasing value attached to every new data point, so we will use this model only for short term predictions. We have included results for 1-day forecasting and 4-day forecasting exercise.

3.1.1 Variables

1. θ_t: The number of COVID-19 cases reported on the t^{th} day
2. ϕ_t: The number of samples tested on t^{th} day
3. L_t: The binary variable indicating whether the t^{th} day was after Lockdown or not
4. η: The date when structural break happens in the data.
5. $\epsilon_t, u_t, e_t, v_t$: The error terms in the model

3.1.2 Plots for Daily Trends

We begin our analysis by plotting the Daily Confirmed Cases for COVID-19 on each day, starting on March 20.

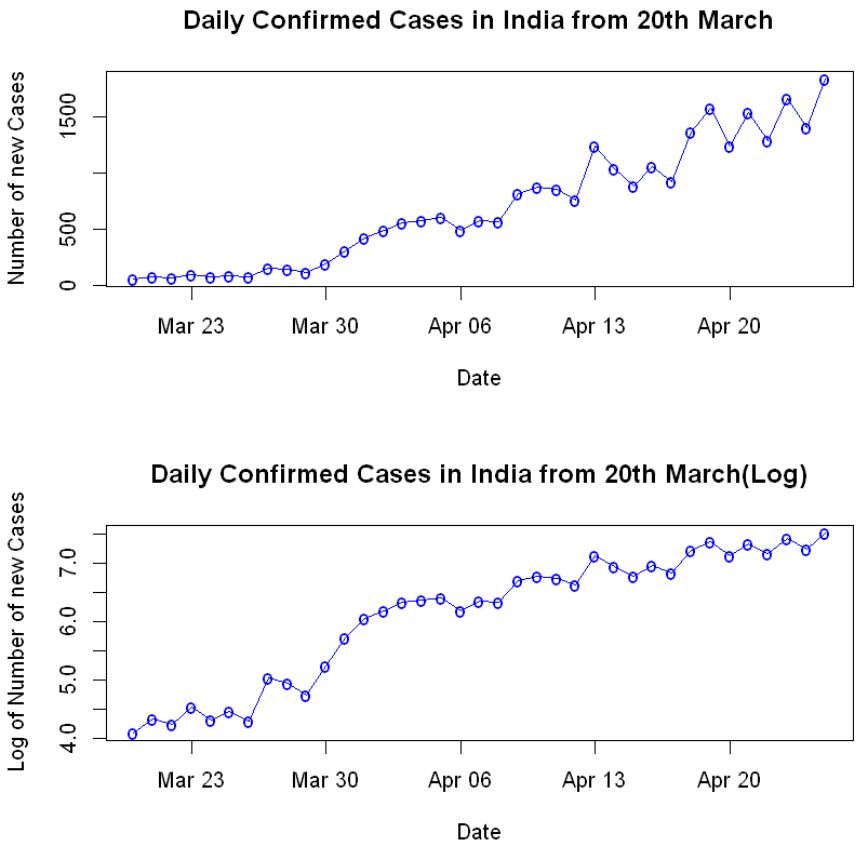

Figure 1: a) Trend of number of confirmed cases each day starting from 20th March b) Trend of log transformation of number of confirmed cases each day starting from 20th March.

We observe that in Figure 1, the second plot ($log(\theta_t)$) initially varies linearly with time, and after some time, the trend changes to quadratic. This date of change, as previously mentioned, is termed η and it helps us to identify the date when a flattening of curve begins in our data (12). The method used to identify such date will be discussed in the appropriate model below. The number of confirmed COVID-19 cases depends on the number of reports, and hence on the number of COVID-19 tests performed in the country. Figure 2 shows the time series for the same (23).

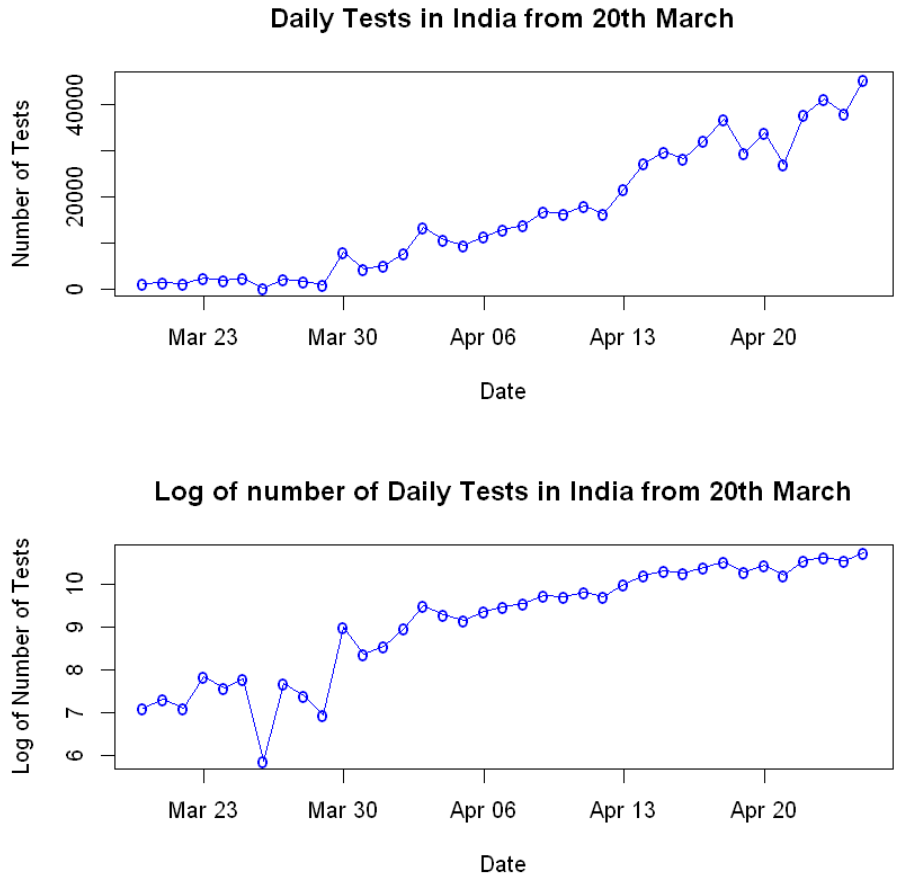

Figure 2: a) Trend of number of samples tested daily with time from 20th March, the date from when there was a significant increase in number of cases. b) Trend of log transformation of number of samples tested daily with time after 20th March.

In the time series graphs for number of samples tested, we fail to see a significant quadratic relationship with time, which we saw in the case of daily confirmed cases. Hence, for regression we will use the linear time trend without the quadratic term.

3.1.3 Auto correlation and Cross correlation plots

After studying the above time series graphs, we conclude that there is some amount of auto-correlation in both the time series. Hence to quantify whether these correlations are significant, we look at the auto-correlation plots for both the series and also the cross-correlation plot to identify whether the number of confirmed cases and number of samples tested are correlated as hypothesised in the previous section.

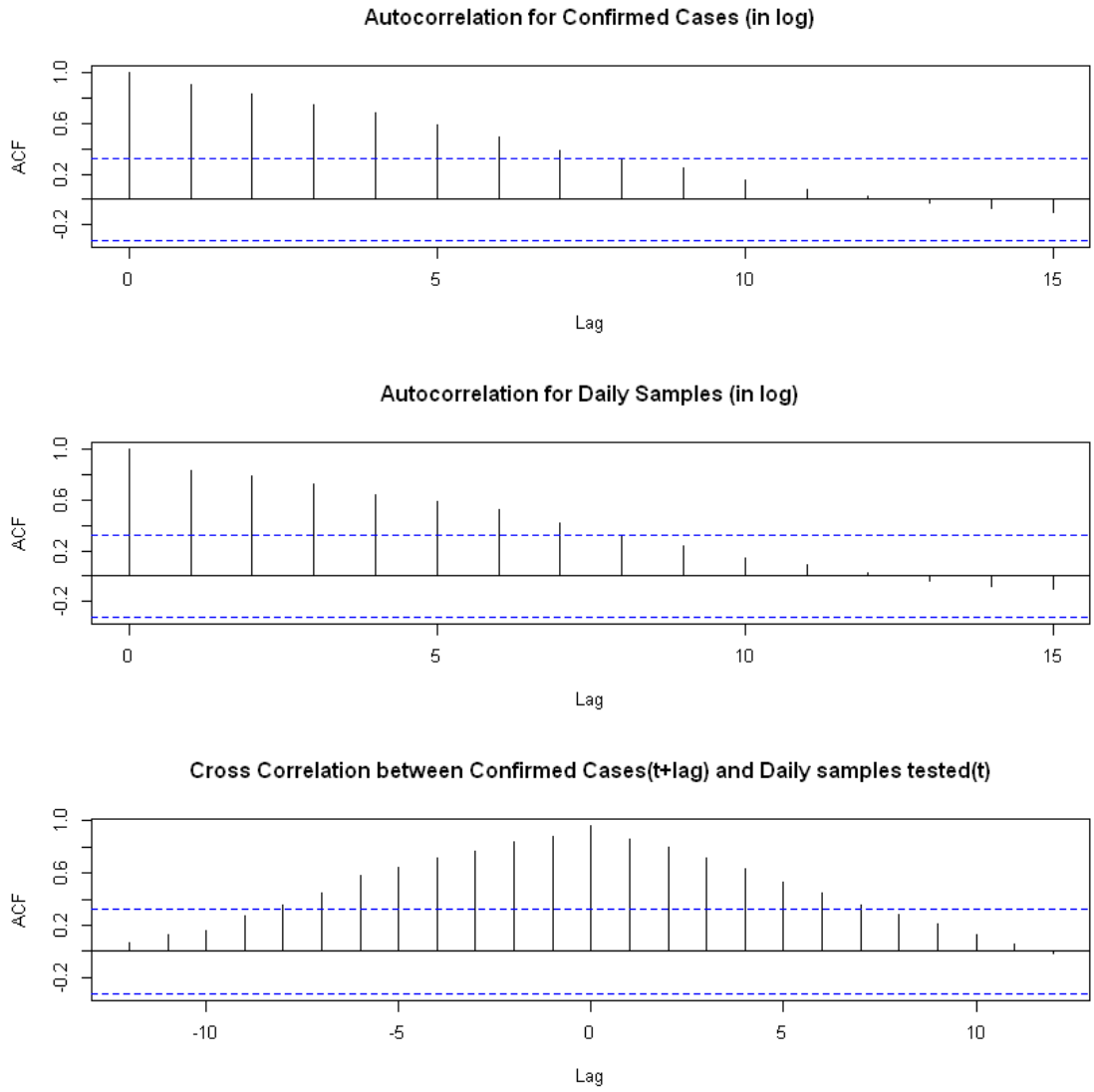

Figure 3: a) Auto correlation for log transformation of number of new cases daily b) Auto correlation for log transformation of number of samples tested daily. c) Cross correlation between log transformation of number of new cases daily and log transformation of number of samples tested daily.

In the first plot we see that the number of confirmed cases is highly auto correlated with its lag values and the same is true for number of samples tested daily. One of the main reasons for this might be a positive time trend which is present in both these series. The third graph shows that there is a cross-correlation between the two series at both positive and negative lags. A possible interpretation for the same may be that higher number of confirmed cases today will lead to higher number of samples tested tomorrow which in turn will lead to higher number of cases day after tomorrow. Or a simplistic way to look at it would be that since both of these are positive correlated with time, correlation without causation exists. In the models described below, we will check whether these correlations exist only because of time trend or inherently.

3.1.4 Models

The models discussed below are built using the techniques VAR (Vector Auto Regression), ARIMA (Auto Regressive Integrated Moving Average) and the OLS (Ordinary Least Square) estimators. Henceforth we describe the four models which have been formulated to explain the underlying time series of the number of new cases of COVID-19 in India. Post modelling, we find that some of these models potentially provide an accurate forecast of actual number of cases, while the others give a good estimate of the upper and lower bounds for the actual number.

1. **Model 1**
   We start with a fundamental model where we detrend the data for $\log(\theta_t)$ and $\log(\phi_t)$. In this model we don't assume any structural break in the time series for $\log(\theta_t)$. The two data generating processes which we estimate in this model are:

$$\log(\theta_t) = \beta_0 + \beta_1 * t + \beta_2 * t^2 + \gamma * L_t + \epsilon_t$$
$$\log(\phi_t) = \beta_0' + \beta_1' * t + \gamma' * L_t + e_t$$

   The OLS estimates for the coefficients are given below.

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β_0	3.665	0.121	30.137	0 ***
β_1	0.224	0.029	9.801	0 ***
β_2	-0.003	0.001	-6.004	0 ***
γ	-0.363	0.193	-1.885	0.068 .

Table 1
OLS estimates for $\log(\theta_t)$

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β_0'	7.058	0.232	30.372	0 ***
β_1'	0.113	0.010	11.480	0 ***
γ'	-0.086	0.308	-0.277	0.784

Table 2
OLS estimates for $\log(\phi_t)$

   We observe that all the variables are quite significant for the first equation. We also observe that γ is negative thus we can say that lockdown has helped to bring down the potential growth in the number of COVID-19 cases in India. For the second data generating process γ' does not come out to be significant, hence we cannot form any opinion on the effect of lockdown on the number of samples tested.

   The auto correlation and cross correlation plots for the residuals for this model are shown in Figure 4. We observe that no auto correlation is present in the residuals for both the series, but there is some cross correlation between the residuals for the second lag. Hence, we perform Vector Auto Regression on these residuals.

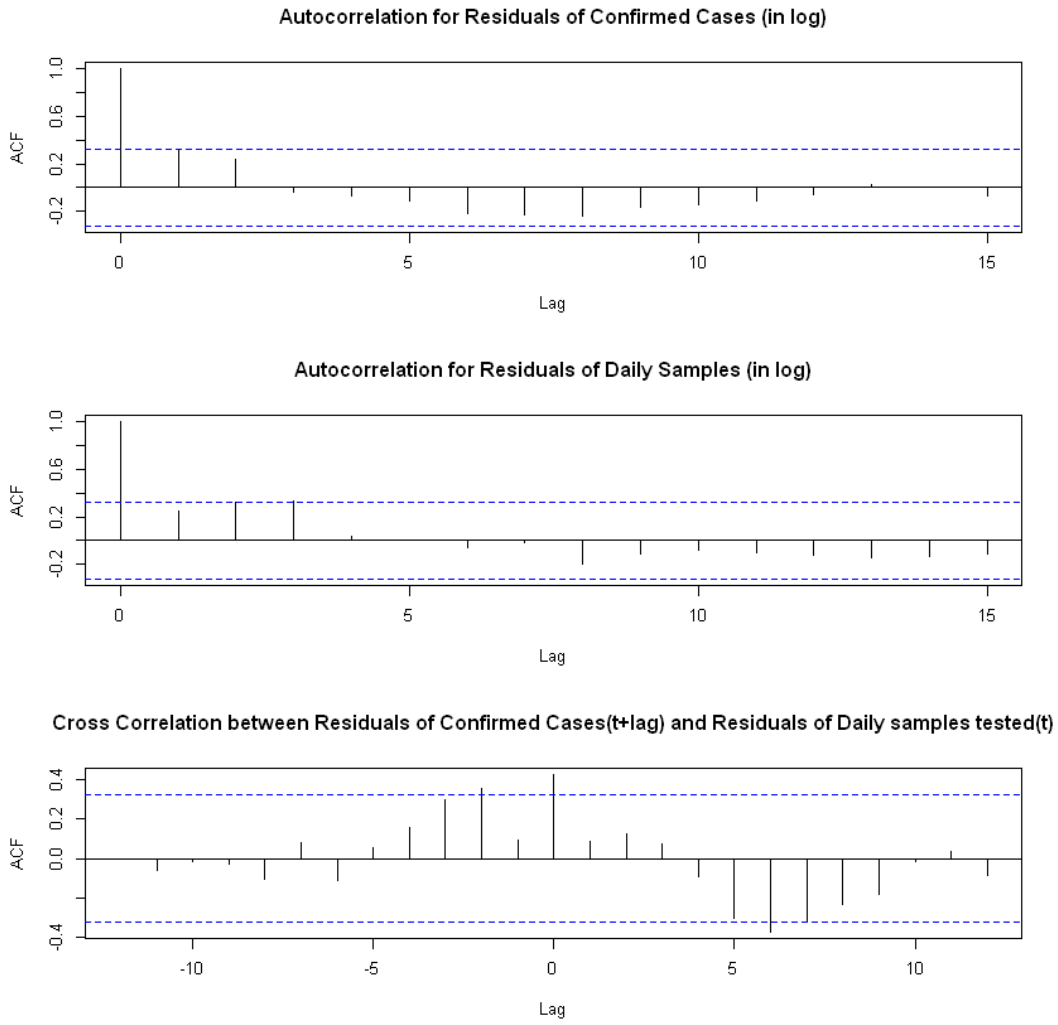

Figure 4: a) Auto correlation for residuals of first estimation b) Auto correlation for residuals of second estimation c) Cross correlation between residuals of both the estimations

After de-trending both the data series, we use Vector Auto Regression(VAR) with a lag of 2, as that is where we had a significant cross correlation. Thus, the data generating processes which we estimate are:

$$\epsilon_t = \lambda_0 + \lambda_1 * \epsilon_{t-1} + \lambda_2 * e_{t-1} + \lambda_3 * \epsilon_{t-2} + \lambda_4 * e_{t-2} + u_t$$

$$e_t = \lambda'_0 + \lambda'_1 * \epsilon_{t-1} + \lambda'_2 * e_{t-1} + \lambda'_3 * \epsilon_{t-2} + \lambda'_4 * e_{t-2} + v_t$$

The VAR estimates are given below

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
λ_0	-0.009	0.038	-0.266	0.823
λ_1	0.289	0.209	1.382	0.177
λ_2	-0.030	0.089	-0.345	0.733
λ_3	0.125	0.205	0.611	0.546
λ_4	0.027	0.089	0.305	0.763

Table 3
VAR estimates for (ϵ_t)

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
λ'_0	-0.008	0.082	-0.091	0.928
λ'_1	-0.383	0.456	-0.839	0.408
λ'_2	0.276	0.193	1.431	0.163
λ'_3	0.754	0.449	1.683	0.103
λ'_4	0.138	0.194	0.708	0.485

Table 4
VAR estimates for (e_t)

We observe that none of the coefficients are significant. λ_3' is significant at 90% significance level. From the graphs, we anticipated significant correlation, but in regression we don't get a significant coefficient for this term.

2. **Model 2**
This model is similar to the first model with the exception of a structural break being introduced while detrending the data. The structural break gives us a different value of coefficients for time trends before and after the break and helps quantify difference in these coefficients, thus indicating significant difference in trends before and after a structural break. Formally, the data generating process is described as follows:

$$\log(\theta_t) = \beta_0 + \beta_{1,j} * t + \beta_{2,j} * t^2 + \gamma * L_t + \epsilon_t$$

$$\log(\phi_t) = \beta_0' + \beta_{1,j}' * t + \gamma' * L_t + e_t$$

with

$$\beta_{i,j} = \begin{cases} \beta_{i,1} & t \leq \eta \\ \beta_{i,2} & t > \eta \end{cases}$$

$$\beta_{i,j}' = \begin{cases} \beta_{i,1}' & t \leq \eta' \\ \beta_{i,2}' & t > \eta' \end{cases}$$

Here η and η' are the dates when the structural break is assumed in $\log(\theta_t)$ and $\log(\phi_t)$ respectively. On observing the graphs, we observe that the trend changes around April 1, which is confirmed by the OLS regressions which results in $\eta = 1/4/20$ and $\eta' = 31/3/2020$. The best OLS model was chosen among a pool of OLS models by comparing the AIC values for each. The results of the OLS are given below.

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β_0	4.313	0.173	24.868	0 ***
$\beta_{1,1}$	-0.041	0.070	-0.588	0.560
$\beta_{2,1}$	0.010	0.006	1.829	0.077 .
$\beta_{1,2}$	0.140	0.023	6.006	0 ***
$\beta_{2,2}$	-0.002	0.0004	-3.325	0 ***
γ	0.073	0.176	0.417	0.680

Table 5
VAR estimates for $\log(\theta_t)$

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β_0'	7.620	0.187	40.677	0 ***
$\beta_{1,1}'$	-0.074	0.031	-2.337	0.025 *
$\beta_{1,2}'$	0.083	0.008	9.802	0 ***
γ'	0.193	0.220	0.875	0.3878

Table 6
VAR estimates for $\log(\phi_t)$

We see that Lockdown does not remain a significant variable now, which can be explained by the fact that now we have assumed a structural break which captures this information and explains that the true effects of Lockdown were observed in data from April 1 (η). Also, for the first equation, which pertains to number of confirmed cases daily, we have a negative value for $\beta_{2,2}$ (coefficient of t^2) after the break, hence the curve starts to flatten in this model after this date.

Let us also see whether any kind of auto correlation or cross correlation exists between the residuals of these two models.

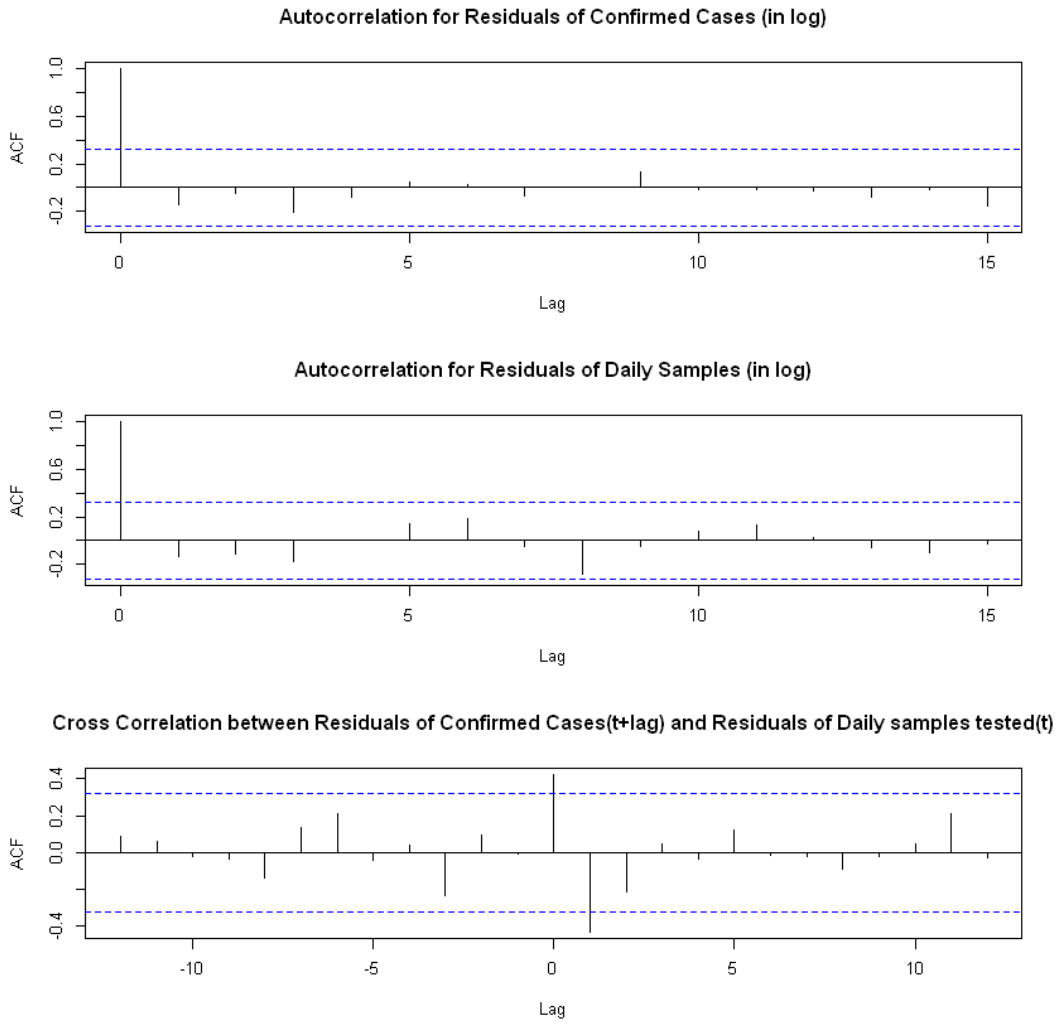

Figure 5: a) Auto correlation for residuals of first estimation b) Auto correlation for residuals of second estimation c) Cross correlation between residuals of both the estimations

No significant auto-correlation among the residuals of the model is observed, but there is some amount of cross-correlation between the two residuals. To extract this information from the residuals, we proceed with Vector Auto Regression (VAR). Only a single lag is taken, as that is when a significant cross-correlation exists. The assumed data generating processes are:

$$\epsilon_t = \lambda_0 + \lambda_1 * \epsilon_{t-1} + \lambda_2 * e_{t-1} + u_t$$

$$e_t = \lambda_0' + \lambda_1' * \epsilon_{t-1} + \lambda_4' * e_{t-1} + v_t$$

The VAR results are as follows:

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
λ_0	0.007	0.025	0.263	0.794
λ_1	-0.217	0.080	-2.717	0.010 *
λ_2	0.007	0.025	0.263	0.794

Table 7
VAR estimates for (ϵ_t)

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
λ_0'	0.013	0.058	0.228	0.821
λ_1'	0.127	0.406	0.312	0.757
λ_2'	-0.157	0.188	-0.835	0.410

Table 8
VAR estimates for (e_t)

For the first equation we see that there is a significant and negative coefficient for e_{t-1} which is also seen in the cross correlation plot above. The causality is hard to establish but a possible explanation can be that a high number of confirmed cases on a day will the lead the medical authorities to act more efficiently which in turn will lead to more number of samples being tested the next day.

3. **Model 3**
In the above models, it is seen that even though there is a high cross correlation between $log(\theta_t)$ and $log(\phi_t)$, it disappears after de-trending both the series. Hence, in this model we will not de-trend the series, but directly apply the VAR model, upto a lag of 2. The data generating processes assumed here are

$$\log(\theta_t) = \beta_0 + \beta_1 * \log(\theta_{t-1}) + \beta_2 * \log(\phi_{t-1}) + \beta_3 * \log(\theta_{t-2}) + \beta_4 * \log(\phi_{t-2}) + \epsilon_t$$

$$\log(\phi_t) = \beta'_0 + \beta'_1 * \log(\theta_{t-1}) + \beta'_2 * \log(\phi_{t-1}) + \beta'_3 * \log(\theta_{t-2}) + \beta'_4 * \log(\phi_{t-2}) + e_t$$

The estimated coefficients using VAR are

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β_0	0.595	0.430	1.384	0.176
β_1	0.689	0.214	3.218	0.003 **
β_2	-0.071	0.111	-0.635	0.530
β_3	0.344	0.220	1.565	0.128
β_4	-0.003	0.111	-0.029	0.977

Table 9
VAR estimates for $\log(\theta_t)$

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β'_0	2.899	0.782	3.706	0.0008 ***
β'_1	0.352	0.390	0.903	0.373
β'_2	-0.071	0.202	-0.350	0.729
β'_3	0.877	0.400	2.192	0.036 *
β'_4	-0.051	0.202	-0.253	0.802

Table 10
VAR estimates for $\log(\phi_t)$

There is high auto-correlation in $\log(\theta_t)$, which is also evident from the auto-correlation plot presented above. The regression for $\log(\phi_t)$ produces a highly significant constant term and a significant coefficient for $\log(\theta_{t-2})$ which can also be seen in the cross-correlation plot described above.

4. **Model 4**
The three models described above have assumed that there is some degree of cross-correlation between $\log(\theta_t)$ and $\log(\phi_t)$ as it is also evident from the cross-correlation plot, but in all the three models, we don't have a highly significant coefficient to describe this relation, hence this model does not consider the time series for $\log(\phi_t)$. The best ARIMA fit on the time series for $\log(\theta_t)$ is identified without de-trending this data, because when the residuals were plotted after de-trending the data in Model-1, no auto correlation was found and the residuals behaved like white noise. ARIMA on these residuals will result in an ARIMA(0,0,0) model. The auto-ARIMA function gives ARIMA(1,1,0) model with some drifts the best fit by comparing AIC values of different models fitted on the data. It is observed from the auto correlation plot that the series is highly auto-correlated. According to the ARIMA model fitted, the underlying data generating equation is:

$$(\log(\theta_t) - \log(\theta_{t-1})) = \beta_0 + \beta_1 * (\log(\theta_{t-1}) - \log(\theta_{t-2})) + \gamma * L_t + \epsilon_t$$

The estimated coefficients along with their standard errors are:

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β_0	0.096	0.029	3.328	0.002 **
β_1	-0.358	0.161	2.223	0.033 *
γ	-0.135	0.234	0.577	0.568

Table 11
ARIMA(1,1,0) estimates for $\log(\theta_t)$

From the coefficient table above, we find that β_0 and β_1 are significant. This means that the current value of $\log(\theta_t)$ depends highly on $\log(\theta_{t-1})$ and $\log(\theta_{t-2})$. We do not get a significant γ which means that this model does not take into account the effect of Lockdown significantly.

5. **Model 5**
   This model is a linear combination of the four models described above. The underlying data generating equation is:

$$\log(\theta_t) = \beta_0 + \beta_1 * Model_{1,t} + \beta_2 * Model_{2,t} + \beta_3 * Model_{3,t} + \beta_4 * Model_{4,t} + \epsilon_t$$

   Here, $Model_{i,t}$ represents the fitted value for $\log(\theta_t)$ by Model-i.

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β_0	-0.149	0.296	-0.505	0.617
β_1	0.092	0.256	0.361	0.721
β_2	0.9374	0.160	5.589	0 ***
β_3	0.320	0.786	0.408	0.686
β_4	-0.326	0.740	-0.441	0.662

Table 12
OLS estimates for $\log(\theta_t)$

   β_2 in this model is highly significant. Thus a linear combination of the four models is highly biased towards the results from the second model.

3.1.5 Forecasts

We compare the forecasts done by the five models described above in this section. We have prepared forecasts spanning over a period of 4 days, and 4 forecasts over a period of 1 day for each model. The results for these are presented below:

Model	26th April 2020	27th April 2020	28th April 2020	29th April 2020
Model-1	**1556**	**1498**	1413	1367
Model-2	1707	1640	1672	**1687**
Model-3	1764	1894	**1934**	2012
Model-4	1903	2141	2340	2584
Model-5	*1601*	*1596*	1575	1553
Actual	1607	1561	1902	1702

Table 13: Forecasts and Actual data (1), (23). The forecasts are done in block of four days.

Model-5, which comprises of the linear combination of all the models, gives the closest forecasts for the given data. After Model-5, Model-1 gives the second best forecasts for $\log(\theta_t)$. It is also observed that Model-3 and Model-4 give us a good upper bound for the number of confirmed COVID-19 cases in India.

Model	26th April 2020	27th April 2020	28th April 2020	29th April 2020
Model-1	**1556**	**1569**	1511	**1658**
Model-2	1707	1774	1709	1805
Model-3	1764	1768	1620	1842
Model-4	1903	1918	**1783**	2007
Model-5	*1531*	*1627*	1616	*1698*
Actual	1607	1561	1902	1702

Table 14: Forecasts vs Actual data (1), (23). The forecasts are done in block of one day.

The table lists the forecasts based on providing the data on confirmed COVID-19 cases in India on a daily basis to the models and forecasts for next day. We see that Model-1 outperforms the other four models, while Model-4

again gives an upper bound. Such models with short forecast range can be used for policy implementations at state level and also to decide the action plan on the daily and weekly basis.

3.2 Regression to find the effects of Temperature and Relative Humidity

Some of the early news reports {(9) & (10)} speculated that an increase in Temperature and Relative Humidity might lead to a decrease in number of COVID-19 cases in the area. In this section we will validate the genuity of this speculation. To study these effects, we have collected the data for five cities with a high number of COVID-19 cases: Mumbai, Indore, Jaipur, Ahmedabad and Delhi.

3.2.1 Variables

The variables used in above the models are retained, and the new variables that are added are:

1. $temp_t$: Temperature at the place on the t^{th} day
2. RH_t: Relative Humidity at the place on the t^{th} day.

3.2.2 The Effects Model

We have collected the data on Temperature and Relative Humidity (3) to perform regression on the number of new cases which are reported daily. The regression is performed on the data starting from March 25 to April 25. The assumed data generating process is
$\log(\theta_t) = \beta_0 + \beta_1 * t + \beta_2 * t^2 + \beta_3 * temp_{t-1} + \beta_4 * RH_{t-1} + \epsilon_t$

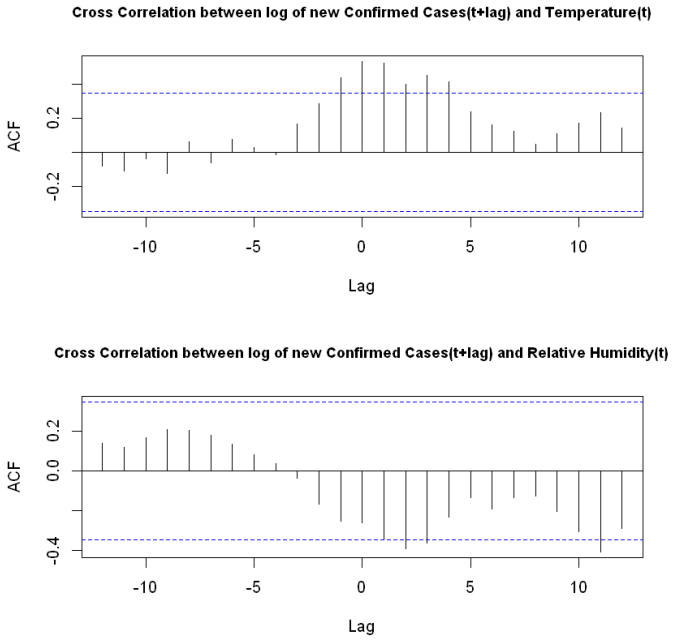

Figure 6: Cross correlation between daily new cases and Relative Humidity and Temperature, for Indore

We have specifically used temperature and Relative Humidity as these were most often cited by various news articles. These variables have been incorporated with a lag since the auto correlation was significant with the first lag. There is significant negative cross correlation between the number of cases and relative humidity and a positive cross correlation between the number of cases and temperature. The positive cross correlation might also be due to the time trend in number of cases and in temperature, since we had been progressing from winters towards the

warmer months of April and May. The following table lists whether the coefficients for the regression performed turned out to be significant or not. Out of the total ten coefficients, only two turned out to be significant. Hence,

City	Coefficient for Temperature	Coefficient for Relative Humidity
Mumbai	**Negative Significance**	Insignificant
Indore	Insignificant	Insignificant
Delhi	Insignificant	Insignificant
Ahmedabad	Insignificant	**Negative Significance**
Jaipur	Insignificant	Insignificant

Table 15: Significance (at 90% level) for the coefficients for Temperature and Relative Humidity for five cities.

with the given data, we cannot claim that there is any observed trend of decrease in number of confirmed COVID-19 cases either with Temperature or with Relatuve Humidity.

3.2.3 The Forecasting Model

We forecast the future number of cases using the same data and the models which we have fitted above. We use temperature and relative humidity for $(t-1)^{th}$ day to forecast the total number of cases on t^{th} day, and for forecasting in general, the model needs to be updated everyday with daily data for temperature and relative humidity. This type of daily forecasting can help in micro level policy implementation, specifically relevant at district level. The data generating process assumed in this section is the same as above.

Using the data till April 25 and running the model everyday after including new data, the forecasts for the next three days and the actual values are as follows:

City	26th April 2020		27th April 2020		28th April 2020	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
Mumbai	358	307	371	488	391	614
Indore	91	46	31	72	165	76
Delhi	293	203	190	150	206	247
Ahmedabad	178	361	197	277	164	340
Jaipur	16	29	25	26	26	13

Table 16: Forecasts for next three days for the cities of Mumbai, Indore, Delhi, Ahmedabad and Jaipur. The actual data is taken from daily news bulletin of various states and twitter accounts of relevant authorities (5), (6), (8), (7), (4).

On some days, the actual numbers and the forecast values were quite similar whereas on the other days, these numbers were off by a significant margin. This kind of modelling can help us get a rough estimate for the number of cases the next day with the caution that the numbers generated by the model should not be anticipated blindly by the relevant authorities.

3.3 Exponential Smoothing

Like regression, exponential smoothing (14) is also a live forecast technique, meaning that the data is updated every few days to produce forecast for a short period of time. Simple exponential smoothing assumes no underlying trend or seasonality. Holt exponential smoothing permits for trend fitting, while Holt-Winters exponential smoothing accommodates a seasonality component too (16). We have a clear time trend in our data albeit no seasonality, so we implement Holt smoothing, with parameters α and β.
Forecast equation: $\hat{y}_{t+h|t} = l_t + hb_t$
Level equation: $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$
Trend Equation: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$
where l_t is the estimate of level of series at time t, b_t is an estimate of trend (slope) of series at time t, α is the smoothing parameter of the level, $0 < \alpha < 1$, β is the smoothing parameter of trend, $0 < \beta < 1$, and h-step-ahead forecast is equal to the last estimated level plus h times the last estimated trend value (15).

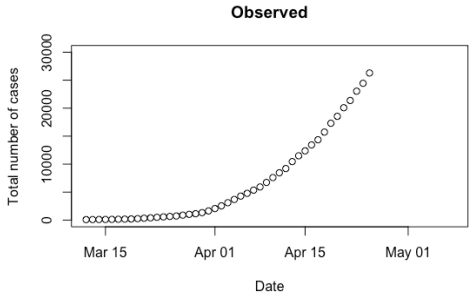

Figure 7: Observed cases of COVID-19 in India till April 25

3.3.1 Forecast for the next ten days

The table in 8 lists forecasts for days beginning from April 26, while values post March 12 have been used in all modelling exercises. Time is discretised to 1 unit = 1 day, which explains actual and forecast plot line gap.

```
Model Information:
Holt's method

Call:
 holt(y = tbu)

  Smoothing parameters:
    alpha = 0.7864
    beta  = 0.7127

  Initial states:
    l = 26.1889
    b = 26.9172

  sigma:  161.751

     AIC     AICc      BIC
634.8560 636.3944 643.8893

Error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE
Training set 52.8937 154.3948 101.0834 1.588033 4.888669 0.1697583
                   ACF1
Training set -0.131710Z

Forecasts:
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
46       27936.22 27728.93 28143.51 27619.19 28253.25
47       29659.57 29286.02 30033.12 29088.28 30230.86
48       31382.92 30791.52 31974.32 30478.45 32287.39
49       33106.27 32259.34 33953.19 31811.01 34401.53
50       34829.62 33695.70 35963.53 33095.44 36563.79
51       36552.97 35104.22 38001.71 34337.30 38768.63
52       38276.32 36487.40 40065.24 35540.40 41012.23
53       39999.67 37847.10 42152.23 36707.60 43291.73
54       41723.02 39184.83 44261.21 37841.19 45604.84
55       43446.37 40501.79 46390.94 38943.03 47949.70
```

Figure 8: Summary and forecast

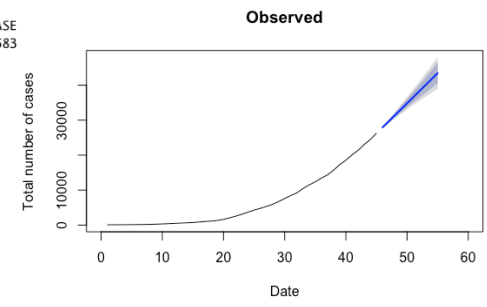

Figure 9: Graph depicting forecast values

3.3.2 Forecasting exercise

Since this method gives good results only for short term estimates, we estimate the number of cases for upcoming days by using past data beginning from March 12, and we keep updating the estimates as days pass, as depicted in Figure 10.

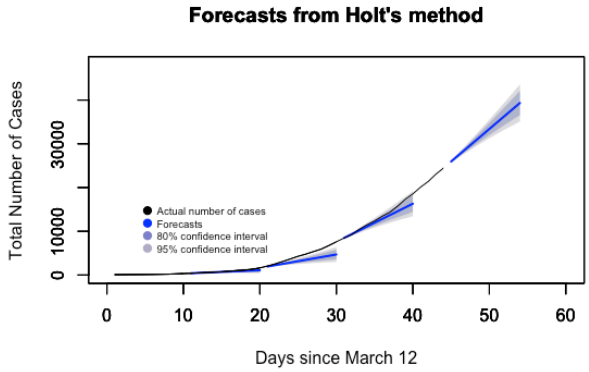

Figure 10: Forecast updated every ten days and actual numbers have been depicted here

Date	Forecast	Actual	Date	Forecast	Actual
March 23	410	497	March 28	790	1024
March 24	486	571	March 29	866	1139
March 25	532	657	March 30	942	1329
March 26	638	730	March 31	1018	1635
March 27	714	883	April 1	1094	2059

Table 17: Forecasting done using data from March 12 to March 22

Date	Forecast	Actual	Date	Forecast	Actual
April 2	1941	2545	April 7	3471	5351
April 3	2247	3105	April 8	3777	5916
April 4	2553	3684	April 9	4083	6729
April 5	2859	4289	April 10	4388	7600
April 6	3165	4778	April 11	4694	8454

Table 18: Forecasting done using data from March 12 to April 1

Date	Forecast	Actual	Date	Forecast	Actual
April 12	8471	9212	April 17	12826	14355
April 13	9342	10455	April 18	13697	15725
April 14	10213	11490	April 19	14568	17304
April 15	11084	12372	April 20	15439	18543
April 16	11955	13433	April 21	16310	20080

Table 19: Forecasting done using data from March 12 to April 11

This forecasting exercise has to be repeated every few days, say once a week, in order to keep logs up to date and results relevant, as we have demonstrated above.

3.4 SEIR

3.4.1 Deterministic SEIR

We implement a deterministic Susceptible Exposed Infectious Recovered (SEIR) compartmental model based on progression of the disease, epidemiological status of individuals and intervention measures of disease transmission (17). The basis for compartmental modelling (11) was laid by Kermack and McKendrick in 1927, and the SEIR model is described in Figure 11.

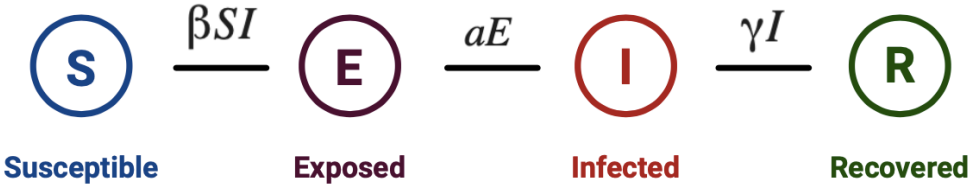

Figure 11: The compartmental model

$$\frac{dS}{dt} = \frac{-\beta IS}{N}$$
$$\frac{dE}{dt} = \frac{\beta IS}{N} - aE$$
$$\frac{dI}{dt} = aE - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$
$$S + E + I + R = N$$

Further compartmentalising of each of the S, E, I and R is possible, as demonstrated in age-structured SEIR. In that case, number of compartments = number of differential equations, and number of parameters accordingly increase. Here, N represents the total population. In all our models, we have assumed N to be 70% of the total Indian population, given that there have been no cases in some states and UTs.
An integral assumption in this deterministic model is that the total population is constant, that is the natural birth rate and death rate in the given time frame is negligible (or almost similar).
An important concept here is that of **reproduction number**, $R_o = \frac{\beta}{\gamma}$. This number captures the dynamics of disease spread in the population. A high value of reproduction number means that every person is infecting a larger number of individuals, and the spread of disease is less controlled. On the contrary, a low R_o indicates that the spread is controlled and fewer people will be affected in the population. Estimates of R_o vary from country to country, in fact town to town and are a good way of looking at impact and extent of any policy implementation like lockdown.
Constant parameters in the model (13):

$$a = \frac{1}{latent\_period} = \frac{1}{5.2}$$

$$\gamma = \frac{1}{infectious\_period} = \frac{1}{2.9}$$

- **Case 1**: The current trend is followed
  If we plot the daily total number of cases versus time, and try to obtain the R_o value from this fit, we get $R_o = 2.98$. The fit is illustrated in Figure 12, while Figure 13 shows the long term evolution of disease. According to this, India will observe peak number of cases near to June 10th, with the peak total number of infections being over 1 million, the total death count standing at nearly 0.79 million, assuming 3.04% death rate (2).

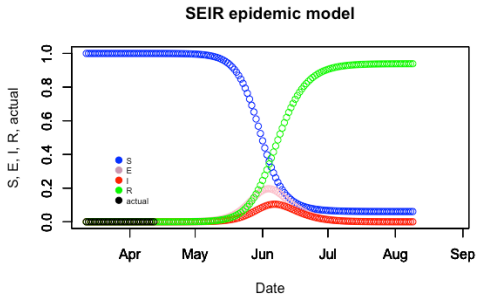

Figure 12: The SEIR model

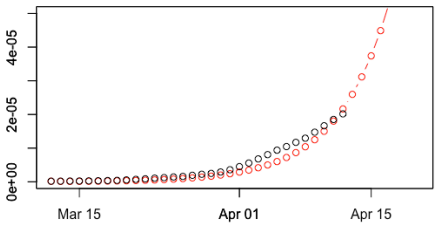

Figure 13: Exploded view with current cases

We have some clear reasons to question this model, which have been listed below:

1. The government has issued the lockdown directive which has reduced the reproduction number over time in other countries significantly.
2. Initial spurt in cases due to travellers and specific gatherings like Tablighi Jamaat are unlikely to be recurrent in the near future.

- **Case 2**: Worst case scenario
  The Reproduction number, R_o is equal to 3.68, which is the WHO estimate of reproduction number (13) in case of no lockdown or policy implementation. In this case, the peak is observed towards the end of May and the peak number of cases of people infected on the peak day comes out to be 12.5 million, with total death count being 0.82 million. This helps us appreciate the effect of lockdown apparent in the analysis done in Case 1, and the importance of policy implication in saving valuable lives.

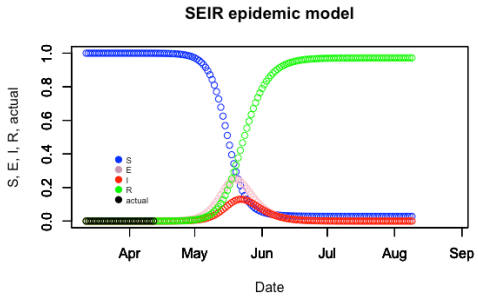

Figure 14: Disease evolution in case of no lockdown

- **Case 3**: Current lockdown scenario, extended
  The initial R_o was assumed to be 2.98. According to studies, this has been brought down to 1.55 because of prevalent lockdown (18). Say this persists. The peak number of cases on a day will be reduced to over 9,000, but it will be in early September, as we can see from the graph. Again, this is not a viable solution, since closing down the entire nation for so long will mean that all economic, agricultural, supply chain, academic, and professional activities are brought to a standstill, and the economy will crash to the worst that nation has seen. Also, survival will be hard because resource availability will also be halted, and a huge proportion of the population relies on short term salary and lacks the savings required to live through such a period.

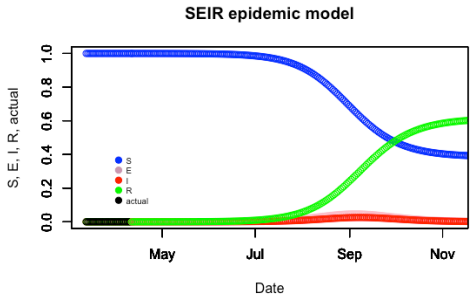

Figure 15: Disease evolution if lockdown persists

- **Case 4**: We are able to bring down the value of R_o to below 1, an estimate number being 0.73
  During the lockdown, if the R_o is brought to below 1, the number of cases can only decrease and the situation would have been completely controlled.

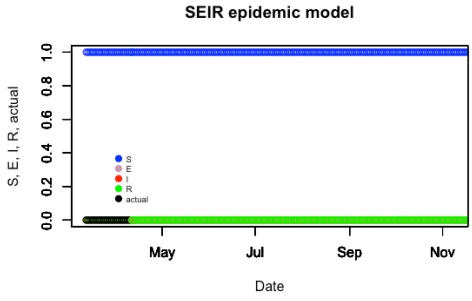

Figure 16: The peak in case of Ro<1 is too small to be observable, and overall the situation is controlled

The trouble with this is that it is next to impossible to achieve this value of R_o, given the socio-demographic distribution and population density of India. Should this have been achievable, we would have much less to worry about.

- **Case 5**: Lockdown ends on May 3
  The average value of R_o prior to lockdown is taken to be 2.2 (18), 1.55 during lockdown, and 3.68 post lockdown. We observe that in this case the peak is delayed to June 20th, but the number of daily cases is still very high, peaking at 125.4 million.

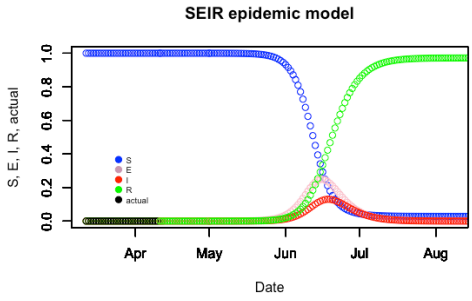

Figure 17: Disease evolution if lockdown ends on May 3

- **Case 6**: Implementing partial lockdown
  Below shown are some graphs where we tweak R_o in chunks as shown in adjoining graphs, where least value of R_o represents lockdown and maximum represents normalcy. Middle one represents partial lockdown. Using the code, the government can plot and plan policy implications.

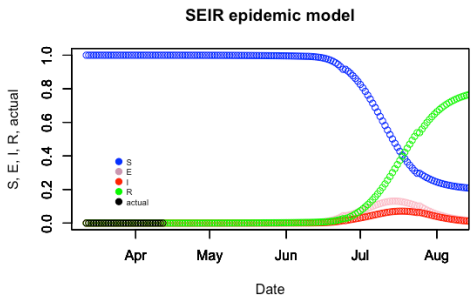

Figure 18: Partial lockdown, case 1

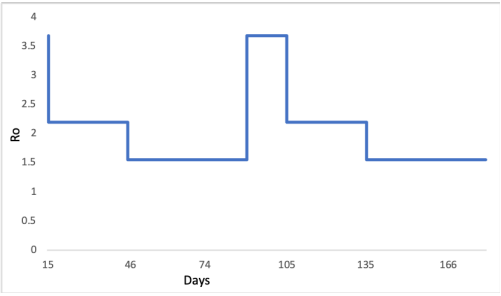

Figure 19: Variation in Ro causes flattened peak in July itself

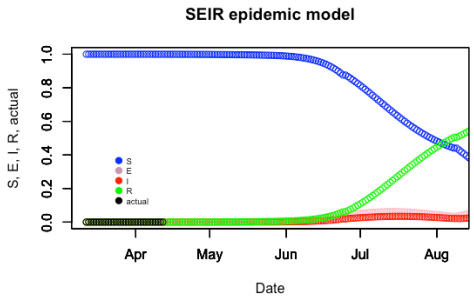

Figure 20: Partial lockdown, case 2

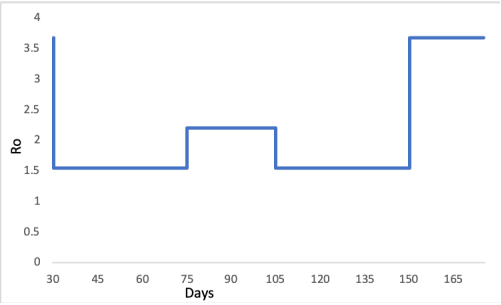

Figure 21: Variation in Ro causes curve flattening but period of policy implementation extends beyond August

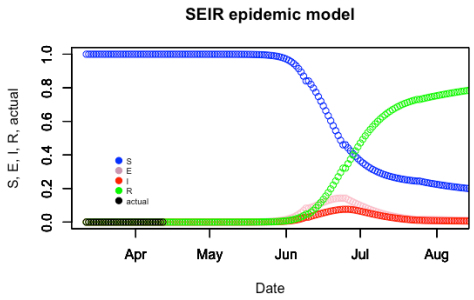

Figure 22: Partial lockdown, case 3

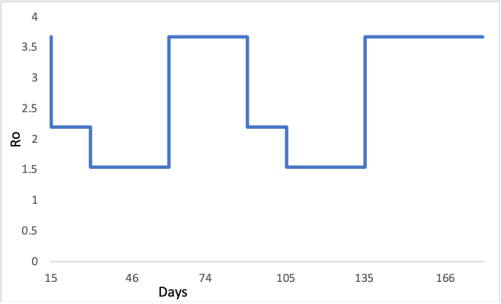

Figure 23: Variation in Ro causes curve flattening and number of cases peak in July

We have demonstrated three very generic possible cases of alternating between normalcy and lockdown, punctuated by partial lockdown. We understand that such a move will help save lives in addition to being much less detrimental on the economy. More cases can be explored using our code, which can be found on the before mentioned github link. Formulating and studying such cases would help the government better anticipate the COVID-19 dynamics in days to come.

3.4.2 Stochastic SEIR modelling

In this model, we relax the assumption that R_o is a constant for a given chunk of days. The R_o values have been discretised on a day to day basis, making observation of the disease dynamics more realistic. We take this in a case wise manner, as depicted below. This is what the scenario will look like if lockdown till May 3 is followed by partial lockdown and then normalcy for a month.

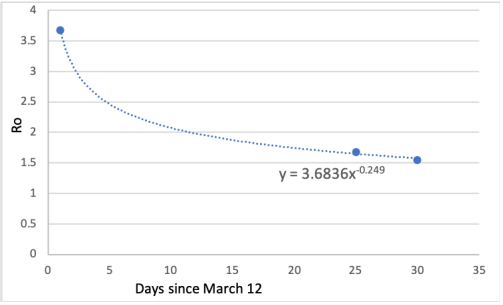

Figure 24: We plot Ro values for India, which dropped from 3.68 to 1.68, and further to 1.55 and fit an exponential equation to it and use this equation to generate a series of Ro values for 40 days of lockdown as it is the case now

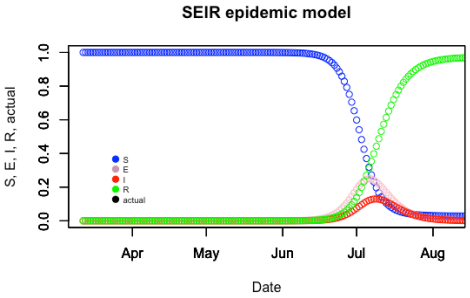

Figure 25: Further this is followed by partial lockdown for a month and then and normalcy. Dates and numbers change different from what it was in deterministic case

3.4.3 Age-Structured SEIR

The structuring of SEIR model can be done in several ways by including age based or geo spatial, or a combination of multiple. We have picked and demonstrated age structured compartmental modelling in this paper.
The Indian demographic (19) and contact pattern varies greatly, and assuming homogeneity in the same is inaccurate,

Figure 26: Population pyramid of India

to say in the least. However, this too can be addressed within the framework of SEIR. Taking 8.5% of the Indian population to be senior citizens denoted by subscript j, and i subscript referring to everyone else, we redefine SEIR as follows (20):

$$\frac{dS_j}{dt} = -\lambda_j S_j$$

$$\frac{dS_i}{dt} = -\lambda_j S_i$$

$$\frac{dE_j}{dt} = \lambda_j S_j - aE_j$$

$$\frac{dE_i}{dt} = \lambda_j S_i - aE_i$$
$$\frac{dI_j}{dt} = aE_j - \gamma I_j$$
$$\frac{dI_i}{dt} = aE_a - \gamma I_i$$
$$\frac{dR_j}{dt} = \gamma I_j$$
$$\frac{dR_i}{dt} = \gamma I_i$$

λ denotes the age specific force of infection, and

1. $\lambda_j = \beta_{jj} I_j + \beta_{ji} I_i$
2. $\lambda_i = \beta_{ij} I_j + \beta_{ii} I_i$

Rest of the parameters are the same as before, only the new contact matrix data is taken (21).

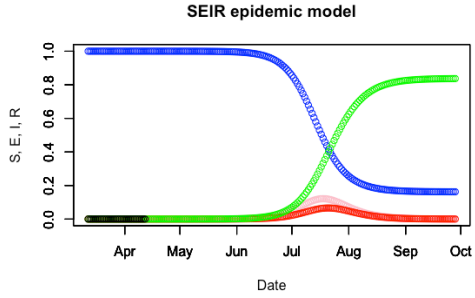

Figure 27: This is plotted for constant Ro = 2.98 adjusted as per contact matrices, based on no lockdown and complete normalcy

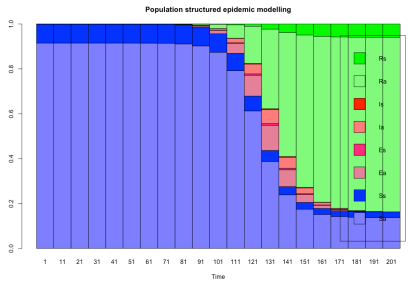

Figure 28: The bar plot represents how disease spreads and affects both sections of population

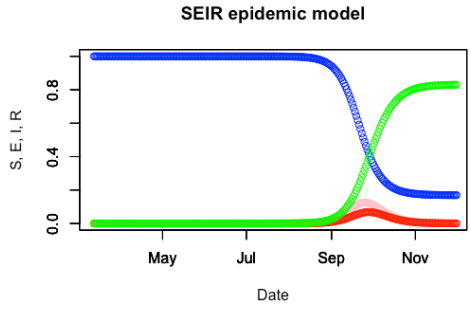

Figure 29: This is plotted for the current scenario, assuming partial lockdown after May 3 and normalcy post June

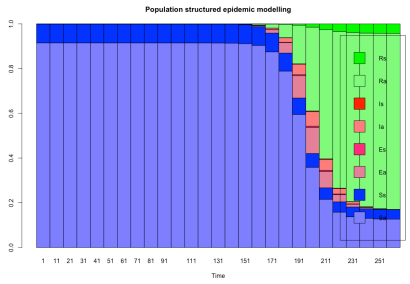

Figure 30: The bar plot represents how disease spreads and affects both sections of population

This analysis helps us not only plot with more accuracy, but also helps us keep count of the more susceptible aged population. More such cases can be constructed using the code which is available at the github repository mentioned earlier.

4 Results

We list results in the form of identifying the key characteristics and findings of every model's underlying time series, that is the number of COVID-19 cases, instead of the forecasts of each model itself since that has already been listed in the above sections, and quantitative results would be outdated in a few days with more data at hand, while the models themselves capture the process and are the key to continuous tracking of cases.

1. **Regression modelling**
    - We realise that short term modelling is the key to efficient planning of testing and healthcare in the country. The current trend is increasing cases everyday, and so the key is to capture the growth trend geo-spatially so that resources can be mitigated accordingly.
    - The time variable in modelling is always significant, which again highlights how, with every passing day, it is important to establish the models that capture time trend in extrapolating for near future forecast.
    - By introducing structural break into the model, which effectively means that we check whether the trend before and after a certain event is significantly different (in this case, lockdown is the event), we are able to study the extent of effect of a lockdown. This helps us better understand the ground reality of implementation of a lockdown/partial lockdown, and whether or not this helps break the increased growth rate of number of cases in general. This helps pin down testing statistics and its effect formally.
    - After identifying suitable models, the key is to stay updated because every time we run the exercise, we get updated and more accurate results for the near future.
    - Modelling an upper bound for forecasts in this case is fairly accurate and hence can give a good conservative estimate of the number of cases that are expected each day.
2. We look at the effect of temperature and humidity by district wise analysis of data, and overall we find that it would be an unrealistic consolation to believe that an increase in the before mentioned parameters would lead to a decrease in the number of cases, since there is no such trend observed in the data as of now.
3. **Exponential Smoothing**
    - We see that short term forecasts are closer to actual numbers, and also that while the model gives an underestimation, the real number never lies outside the 95% confidence interval (closer to the forecast upper limit).
    - Since this model heavily depends on past data, every day's data adds value to it and its forecast for the upcoming days is extremely relevant, as it has been described in regression.
    - While regression modelled above depends on series' past values up to a lag and the time variable, exponential smoothing uses all of the past data to forecast future values. In this case, ever day's data adds value to the forecasting and hence new cases that are predicted are specific to the place where prediction is done and the entire trend is captured in the model in two parts, level trend as well as time trend.
    - This model coupled with regression gives us a short term interval prediction and gives us an idea about the required level of testing, tracking and healthcare facilities everyday.
4. **SEIR modelling**
    - In all the models discussed in the paper, there is a trade off between number of cases and number of days to peak, and we cannot afford to compromise on either, which makes such modelling extremely important. While it is hard to estimate the accuracy or establish the most accurate model because of lack of past data and increased uncertainty in the future, which depends on the societal response and the disease evolution itself (with the disease strain itself varying among populations), this modelling helps us get an idea of the possible disease evolution in the purview of different scenarios, as discussed.
    - The only solutions for such case might be to bring down the susceptible population (S) enough so that the peak is not insurmountable, or reduce R_o to less than 1 so that maximum number of cases is capped and spread of the pandemic contained.
    - A possible way to achieve this would be by well planned and implemented periods of lockdown, partial

lockdown and normalcy. At the same time, it is important to model or plan accurately so that things don't worsen. For instance, if we open to normalcy on May 3, things will be the same, and in fact worse, as they were in in March, with some untracked cases and more and more people getting exposed and affected. So, the lockdown would have nothing but delayed an inevitable peak.

- A good way to reduce R_o value would be increased testing, efficient tracking and quarantining of existing cases, which will help reduce peak without compromising on the time period of lockdown.
- Also important is to realise that this is a marathon battle until we are either able to attain herd immunity, wait out the curve, contain the spread, develop a vaccine, or succumb to losses of lives. So the preparation and response should be befitting.

5 Conclusion

Time is the key variable in modelling short term forecasts like regression and exponential smoothing, and it gives best and most relevant results if the study is done on a smaller scale, say district wise. But since the data available is scarce, large scale analysis on the national level gives us a better idea about the trend of the emerging number of cases. There is no significant trend of any variable other than time, and temperature and humidity have insignificant effect on the number of cases in general so far. The SEIR model captures disease evolution for a length of period of time, and can be improved by introducing geo-spatial contact matrices along with age. Short term forecasts help in efficient planning for testing and healthcare facilities, while long term forecasts are the key to implementation of policies like lockdown and making resources available to the general population, a large proportion of which cannot work from home, and depends on daily to monthly earnings and cannot afford going out of work. By studying both, we have tried to capture the entire picture of the pandemic. Increased testing will help identify more and more cases, but at the same time, instances of asymptotic cases, estimated to be at around 80% of the total existing ones (22), are a cause of worry as they can only be identified by contact tracing and without any symptoms, they are disease carriers. Use of technology to tackle problems, such as that done by the initiative by the Government in launching the Aarogya Setu app will go a long way in improving conditions. In addition to increased case locking, it is important in such a scenario for the unaffected population too to take precautionary steps to ensure that they are protected from any possible form of transmission. Meanwhile, the nation must brace itself for a long and hard battle, for every case and death is more than just a statistic and it is crucial to come together to prevent losses detrimental to our nation.

References

[1] https://www.covid19india.org

[2] https://www.worldometers.info/coronavirus/country/india/

[3] https://www.worldweatheronline.com/developer/

[4] https://twitter.com/PIBJaipur

[5] https://twitter.com/Maha_MEDD

[6] http://mphealthresponse.nhmmp.gov.in/covid/

[7] https://gujcovid19.gujarat.gov.in/

[8] https://twitter.com/CMODelhi

[9] https://www.indiaspend.com/early-studies-claim-heat-humidity-will-slow-down-covid-19-spread-some-experts-disagree/

[10] https://economictimes.indiatimes.com/industry/healthcare/biotech/healthcare/studies-saying-warm-weather-slows-covid-19-not-conclusive-report/articleshow/75081946.cms?from=mdr

[11] https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology

[12] Soudeep Deb and Manidipa Majumdar: A time series method to analyze incidence pattern and estimate reproduction number of COVID-19 doi: https://arxiv.org/pdf/2003.10655.pdf

[13] Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction. [Submitted]. Bull World Health Organ. E-pub: 19 March 2020. doi: http://dx.doi.org/10.2471/BLT.20.255695

[14] https://en.wikipedia.org/wiki/Exponential_smoothing

[15] https://otexts.com/fpp2/holt.html

[16] https://johannesmehlem.com/blog/exponential-smoothing-time-series-forecasting-r/

[17] Srijana: SEIR model. https://rpubs.com/srijana/110753

[18] https://theprint.in/science/covid-19-fight-is-a-test-match-not-a-t20-heres-what-india-needs-to-do-to-win/406207/

[19] https://www.populationpyramid.net/india/2019/

[20] Aaron A King, Helen J Weating: Age Structured Models. Avaiable on https://ms.mcmaster.ca/~bolker/eeid/2011_eco/waifw.pdf

[21] K. Prem, A. R. Cook, and M. Jit: Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comp. Bio 13, e1005697(2017). https://doi.org/10.1371/journal.pcbi.1005697

[22] https://www.ndtv.com/india-news/coronavirus-80-per-cent-cases-asymptomatic-matter-of-concern-medical-research-body-icmrs-scientist-t-2214799

[23] https://icmr.nic.in/content/covid-19