Analyzing the trend in COVID-19 data: The structural break approach

Nityananda Sarkar and Kushal Banik Chowdhury[®]

Submission Time: 2 March 2021 | **Acceptance Time:** 8 January 2023

ABSTRACT

In this paper, we have considered three important variables concerning COVID-19 *viz.*, (i) the number of daily new cases, (ii) the number of daily total cases, and (iii) the number of daily deaths, and proposed a modelling procedure, so that the nature of trend in these series could be studied appropriately and then used for identifying the current phase of the pandemic including the phase of containment, if happening /happened, in any country. The proposed modelling procedure gives due consideration to structural breaks in the series. The data from four countries, Brazil, India, Italy and the UK, have been used to study the efficacy of the proposed model. Regarding the phase of infection in these countries, we have found, using data till 19 May 2020, that both Brazil and India are in the increasing phase with infections rising up and further up, but Italy and the UK are in decreasing/containing phase suggesting that these two countries are expected to be free of this pandemic in due course of time provided their respective trend continues. The forecast performance of this model has also established its superiority, as compared to two other standard trend models *viz.*, polynomial and exponential trend models.

Key words: *COVID-19, Structural breaks, Non-stationarity, Forecasting*.

JEL code: C22

1. INTRODUCTION

Novel coronavirus originated in Wuhan, China, and then travelled across boundaries in lightning speed in this closely connected globalised world. The entire world is now under an unprecedented situation facing a number of serious challenges. The most immediate challenge from medical point of view is to find out an effective treatment protocol based on available medicines (since no new proper medicines have yet been found) so that coronavirus afflicted patients recover and death due to this disease is as few as possible. The ultimate goal, of course, is to find a proper vaccine against this disease. From the public health and government point of view, the challenge is to contain the spread of this dreaded virus most effectively.

This disease, called COVID-19, has been straining health – care system worldwide. Even some of the most developed economies like the USA, the UK, Spain, Italy, and France having best health-care infrastructure have failed in providing adequate and effective medical care to their citizens suffering from this disease, resulting in not only very large number of infections but also very high deaths in these countries. The situation is obviously very alarming in the developing and poor countries. Since this disease is highly infectious, the World Health Organisation (WHO) and public health experts are all along emphasising on effective intervention by the governments in enforcing strict measures to control its spread. Interventions by the governments have focussed on, *inter alia*, measures like partial /complete lockdown, more testing, social distancing, tracing of probable cases and quarantining them. Monitoring of these interventions is being made regularly and accordingly changes in the nature as well as

[®] Nityananda Sarkar, Indian Statistical Institute, Kolkata, West Bengal, India, 700108, (email: tanya@isical.ac.in). Kushal Banik Chowdhury, Indian Statistical institute, North-East Centre, Tezpur, Assam, India 784501, (email: kush.kolkata@gmail.com).

duration of such interventions are being done so that the spread of this dreaded virus could be contained effectively.

While the governments all over the world are acting very seriously to deal with this pandemic, the required intensity, duration and urgency of responses by the governments crucially depend on the likely magnitude and extent of spread of this disease in the coming days. For all these purposes, reliable forecasts of important variables related to this pandemic are very important. To that end, the data generating process underlying such time series needs to be obtained and studied. Among all relevant variables the forecasts of which would help governments in assessing the requirements for health infrastructure and initiating steps to provide adequate medical care, and also in formulating their intervention policies, three variables are very important. These are: (i) the number of daily new COVID-19 cases (DNC), (ii) the number of daily total COVID-19 cases (DTC), and (iii) the number of daily deaths due to COVID-19 (DD). The importance of the first two time series is obvious and well-recognised. The third one, *viz*., the number of daily deaths due to COVID-19, is also very useful and relevant since containment of any infection means that not only the DNC figures decrease over time, but the number of daily deaths also decreases indicating the effectiveness of the medical treatment being used for curing the afflicted patients.

The primary focus of this work is to find appropriate models underlying the data generating processes (DGP) of these three variables by using the tools of time series analysis. The modelling approach proposed here involves studying the nature of trend, deterministic/stochastic or both, along with consideration to structural breaks/changes in the time series[1]. The issue of structural breaks in time series is very crucial since finding of breaks would mean that statistically significant changes (increase/decrease) in the underlying DGP of these series have happened during the period under study. In case a significant change, say, increase, is found at some point of time and the increasing trend continues, it means that the disease is actually on the way up significantly till another change occurs, which may also be a further significant rise (or fall) in the series. Our approach where due consideration to structural changes in the time series is given is important for this pandemic since epidemiologists, virologists and medical researchers have been repeatedly cautioning about the different stages of transmission of this infection. These stages are characterised in terms of source/origin being known/ traceable/ unknown, extent of spread of infection, clustering of infection etc.

However, from the of view of health management concerning COVID-19, it is important to keep in mind that effective steps in controlling the spread of this pandemic requires identifying when significant increases in the time series of DNC or DD have happened, and also when it is likely to stabilise or start decreasing indicating likely containment of the disease in due course of time. Finding these breaks or change points would enable the governments to decide on when to enforce stricter measures and mobilise health –care infrastructure with top most priority, and also when to start relaxing these measures partially/gradually. We call the time periods identified by these break points in the time series as phases. These phases may be broadly categorised as ‘increasing’ where further and further increases in the successive breaks points happen, followed by a ‘stable’ phase which may or may not exist, and then finally the ‘decreasing or containing phase’ where also there could be more than one time point with significant decrease. In this paper, we propose a procedure, based on application of some existing methodologies of modern time series analysis, so as to be able to test for and then estimate such brake/change points in a given time series, thus enabling us identification of the

[1] ‘Breaks’ and ‘changes’ would be used interchangeably.

underlying phases of this pandemic, and finally, concluding about the likely containment, if happening /happened, of this pandemic**.** It may be pointed out that our phases need not necessarily coincide with the different stages of transmission of this disease as the virologists and epidemiologists describe. In our proposed modelling approach, we basically adopt a trend model with consideration to existence of structural breaks in the time series and also use the general concept of trend which include stochastic trend apart from deterministic trend. As we have stated in the next section on literature review, there are some works which use pure mathematical models, and few other works where essentially statistical modelling is done. Our approach, as stated above, is somewhat different in the sense that we incorporate, in our model, a distinctive property of time series called 'structural break' which is often found in many time series of moderate size and above, to understand and study some important COVID-19 variables.

The paper has a secondary purpose, namely, to compare the forecasting performance of some conventional trend models *vis a vis* the model proposed based on our modelling approach. While in the conventional time series analysis it is assumed that trend is entirely deterministic in nature and a specific trend curve (for example, linear/quadratic/exponential) is fitted, it is well-known that this traditional trend analysis has many limitations and hence conclusions based on such trend analysis may be inappropriate and/or misleading. It may be pointed out that sometimes it is being mentioned by some experts that the spread of this infection in respect of daily number of total cases is exponential in nature. It is quite possible that this is indeed so. But it is also possible that this is not the most appropriate description of trend in the time series of daily total cases. We fit, as an alternative to our proposed model, the conventional trend models like polynomials (in time) of suitable degree and exponential trend model. In these models, trend is assumed to be completely deterministic in nature and there is no consideration to any structural break in the time series. We then obtain forecasts based on these two trend models, and finally compare them with those obtained from our proposed model by applying the standard forecast performance criteria.

We have applied the proposed modelling procedure to the time series data of four countries. These countries are: Brazil, India, Italy, and the UK. The choice of the countries has been made keeping in view the developmental status of the countries – the first two being emerging countries and the last two developed. Also, the choice is guided by the purpose to study the efficacy of the proposed model[2] regardless of the phase of infection prevalent in a country.

The remaining sections of the paper are formatted as follows. The literature review is made in the next section. The methodology applied is discussed in Section 3. Empirical findings of the proposed model are discussed in Section 4. Forecast performances of the proposed model and two other standard trend models are presented in Section 5. The paper ends with some concluding remarks in Section 6.

2. LITERATURE REVIEW

Epidemics have a long history in human civilisation. To understand the dynamics of an epidemic / pandemic and also to predict its spread, developing appropriate models is very important. In such developments, mathematical models have always taken precedence in providing deeper understanding of the transmission mechanisms of a disease outbreak. One of

[2] For sake of convenience of expression, by 'proposed model' we mean the model based on the procedure of data analysis proposed in this paper.

the very well-known early models explaining human-to-human transmission is known as Susceptible-Infectious-Removed (SIR) epidemic model which was proposed by Kermack-Mckendrick in 1927. Following their seminal work, SIR epidemic model has been used extensively in understanding the spread of different viruses. Subsequently, SIR model has been extended in many directions leading to other well-known models including what is known as Susceptible-Exposed-Infectious-Removed model (SEIR). Several researchers (Wu et al., 2020; Calafiore et al., 2020; Kucharski et al., 2020; Simha et al., 2020; Anastassopoulou et al., 2020; Nesteruk, 2020; Nabi, 2020; Fanelli and Piazza, 2020 and Mandal et al., 2020) have applied the SEIR model and its many extensions to understand the spread of COVID-19.

Besides developing models applying mathematical approach, scientists and researchers are also using statistical models to analyze, predict and understand the spread of COVID-19. For instance, Zhang et al. (2020) have applied segmented Poisson model to predict the turning point and duration of outbreaks of coronavirus in some western countries. Using a long short-term memory (LSTM) model, Tomar and Gupta (2020) have predicted the total number of COVID-19 cases in India for a 30-day ahead prediction window. Yonar et al. (2020) have applied auto-regressive integrated moving average (ARIMA) model and the Brown-Holt linear exponential smoothing method to estimate and forecast the number of COVID-19 cases in the G8 countries. Rafiq et al. (2020) have employed state-space model to evaluate the COVID-19 situation in India. Prediction of infected cases in Italy has been made using ARIMA model Chintalapudi et al. (2020). Ribeiro et al. (2020) have applied stochastic regression models to forecast COVID-19 cases in ten most affected states of Brazil. A Gaussian mixture model has been applied by Singhal et al. (2020) to model and predict the COVID-19 pandemic. In another recent work, Chakraborty and Ghosh (2020) have proposed a hybrid ARIMA-Wavelet transformed model to forecast COVID-19 cases for some countries.

3. METHODOLOGY

This study, as discussed in Section 1, finds time series models with due consideration to trend, both deterministic and stochastic, and structural breaks/changes in the series, for three variables *viz.*, (i) number of daily new COVID-19 cases, (ii) number of daily total cases, and (iii) number of daily deaths. We have already stated that there are few studies that have applied time series techniques for predicting the future values of the important variables related to COVID-19. Generally speaking, our work differs from other such works in the sense that, unlike others, we have applied a newly developed time series technique to examine the breaks in the trend function of some of the COVID-19 variables. As mentioned in ‘Introduction’, the time series model with trend break(s) can provide valuable information about the data generating process of the variables and hence improves future prediction of the same. The nature of spread of the disease, in so far as observed, has episodes of changes, to start with increases followed by decreases, suggests that ‘structural break analysis’ would be useful in analyzing these three important COVID-19 variables.

As of now, it is not known, and, in fact, it is very unlikely, that the spread of this infection has any periodic and/or cyclical behaviour. Hence the other two components of a time series *viz.*, seasonality and cyclical fluctuations, are not considered here.

It is well-known that trend is the long- run smooth movement of a time series. Modern time series analysis considers trend to be both deterministic as well as, what is called stochastic, in nature. To test for non-stationarity of a time series in the sense of having a stochastic trend, most often the well-known unit root test, called the augmented Dickey-Fuller (ADF) (1979) test, is used. However, one major limitation of the ADF test is that the deterministic trend function in the estimating equation is assumed to be constant all throughout the sample period. In his seminal work, Perron (1989) showed that the autocorrelation process of a random walk model is almost the same as that of a deterministic trend model with a break in the trend function. And hence the ADF test may misleadingly conclude that there is a unit root when, in fact, there is no unit root, and the true data generating process is stationary with a structural change in the deterministic trend function. Perron (1989) proposed a unit root test under three different types of deterministic trend function where the (single) change point was assumed to be known *a priori*. This latter assumption being restrictive, Zivot and Andrews (2002), and Vogelsang and Perron (1998) relaxed this, and the change/break point was assumed unknown and estimated endogenously.

Kim and Perron (2009) have pointed out that all such studies which assume the break date as unknown do not allow for the possibility of a trend break under the null hypothesis of unit root; those tests consider break under the alternative of stationarity only and thus the proposed test statistics are inferior in terms of size and power. To overcome this limitation, Kim and Perron (2009) have developed a new test on the line of Perron's (1989) original formulation of trend break being allowed under both the null and alternative hypotheses, but the (single) break date is now assumed to be unknown. Carrion-i-Silvestre et al. (2009) generalized the Kim-Perron test of unit roots by allowing multiple structural breaks under both the null and alternative hypotheses.

Now, prior to applying the unit root test, knowing if a structural break is present in a given time series is very crucial. However, tests for structural breaks in terms of intercept and / or slope of a deterministic trend function suggest that the performance of these tests depend on whether the time series under study is stationary or non-stationary having unit roots. To deal with the above 'circularity' problem, Perron and Yabu (2009) proposed a novel test for structural change in the trend function of a univariate time series, which can be performed without any prior knowledge on whether the noise component is stationary or non-stationary containing unit roots. Later, the Perron-Yabu test has been generalised by Kejriwal and Perron Kejriwal and Perron (2010). In their study, Kejriwal and Perron (2010) have designed a test to detect multiple structural breaks in the deterministic trend function without the requirement of any prior condition of stationarity or non-stationarity of noise term. Since the Kejriwal-Perron test is very general in its approach insofar as the assumption on noise is concerned, we have performed this test to detect the presence of multiple structural breaks in the trend function of a time series, and then applied the test proposed by Carrion-i-Silvestre et al. (2009) to test the null hypothesis of unit roots against the alternative of stationarity allowing for presence of multiple breaks in the deterministic trend function under both the hypotheses.
We give below a brief description of the Kejriwal-Perron method of testing for multiple structural breaks followed by the unit root tests proposed by Carrion-i-Silvestre et al. (2009).

3.1. Sequential break tests to detect multiple structural breaks

The sequential testing method of Kejriwal and Perron (2010) assumes that the time series variable y_t is generated in the following manner

$$y_t = x_t'\beta + u_t,$$
$$u_t = \alpha u_{t-1} + \varepsilon_t, t = 2, \dots, T$$
$$u_1 = \varepsilon_1,$$

where x_t is an $(r \times 1)$ vector of deterministic components which, in our study, is the deterministic trend of the time series y_t, β is a $(r \times 1)$ vector of unknown parameters and u_t

is the random error. The parameter $\alpha \in (-1, 1]$, which implies that u_t can be stationary or can have unit roots, and $\varepsilon_t \sim iid\ (0, \sigma^2)$. The sequential testing procedure of identifying l breaks against the alternative of $(l+1)$ breaks is defined as follows. First, it estimates the l break dates $\widehat{T}_1, \dots, \widehat{T}_l$ as global minimizers of the sum of squared residuals (SSR) from the model of y_t with l breaks which are estimated by the ordinary least squares (OLS) method: $\left(\widehat{T}_1, \dots, \widehat{T}_l\right) = \text{argmin}_{(\widehat{T}_1, \dots, \widehat{T}_l)} SSR(T_1, \dots, T_l)$. The estimated break dates are obtained by using dynamic programming algorithm proposed by Bai and Perron (2003). Next, the procedure searches for an additional break in each of the $(l+1)$ intervals *viz.*, $I_1 = \left[0,\ \widehat{T}_1\right], I_2 = [\widehat{T}_1,\ \widehat{T}_2], \dots, I_{l+1} = [\widehat{T}_l,\ \widehat{T}]$. In order to construct the test at the interval I_i (where $i = 1,2, \dots, (l+1)$), Kejriwal-Perron considered the regression $y_t = x_t'^{(i)} \beta^{(i)} + u_t^{(i)}$, where $x_t^{(i)}$ is a set of deterministic variables representing structural breaks. In this exercise, we assume that the break occurs both at the intercept and slope coefficients of the trend function. Therefore, $x_t^{(i)} = (1, \text{F}(t > BD),\ t - \widehat{T}_{i-1},\ (t - BD)\text{F}(t > BD))'$ where BD is the estimated break date, and $\text{F}(t > BD)$ is an indicator function that takes the value one for $t > BD$ and zero otherwise. The residuals from this regression, denoted $\hat{u}_t^{(i)}$, are then used to compute the ordinary least squares (OLS) estimate of α of the regression $\hat{u}_t^{(i)} = \alpha \hat{u}_{t-1}^{(i)} + \sum_{j=1}^{k} a_j^* \Delta\, \hat{u}_{t-j}^{(i)} + \varepsilon_t$ where Δ is the difference operator and k is the optimal lag chosen by Bayesian Information Criterion (BIC). This estimate of α is further used to obtain the super-efficient estimate of α, denoted by $\hat{\alpha}_s^{(i)}$, by following the equation $\widehat{\alpha_s} = \begin{cases} \hat{\alpha} \ if \ T^{\delta}|\hat{\alpha} - 1| > d \\ 1 \ if \ T^{\delta}|\hat{\alpha} - 1| \leq d \end{cases}$. Perron and Yabu (2009) have obtained that the values $d = 1$ and $\delta = 0.5$ lead the best finite sample results. Using these information, Kejriwal-Perron defined the transformed variables as follows: $y_t^{*i} = y_t - \widehat{\alpha_s} y_{t-1}$ and $x_t^{*i} = x_t - \widehat{\alpha_s} x_{t-1}$ for every $t \in I_i$, and specify the feasible Generalized Least Squares (GLS) regression as $y_t^{*i} = x_t^{*i'} \beta^i + u_t^i$. Depending on the feasible regression, a Wald test statistic (W_{FS}^{τ}) is computed for every permissible break dates. Thereafter exp-functional is defined as $\text{Exp}W_{FS}^i = \log\left[(\widehat{T}_\iota - \widehat{T_{\iota-1}})^{-1} \sum_{\tau \in I_i} \exp(W_{FS}^{\tau}/2)\right]$. Given the exp-functionals $\{\text{Exp}W_{FS}^i\}_{i=1,\dots,(l+1)}$, Kejriwal-Perron defines the sequential test as

$$F_T(l+1|l) = max_{1 \leq i \leq l+1} \{\text{Exp}W_{FS}^i\}.$$

The conclusion will be in favour of $l+1$ breaks if $F_T(l+1|l)$ is found to be larger than the critical values reported in Kejriwal and Perron (2010).

In our analysis, we first apply the test to detect if there exists one break in the deterministic trend function or not. Upon rejection, we test if there are two breaks, and so on, until the test fails to reject the existence of $l+1$ breaks.

3.2. Unit root tests under multiple structural breaks

As stated earlier, Carrion-i-Silvestre et al. (2009) proposed a set of unit root tests that allows multiple structural breaks in trend function under both the null and alternative hypotheses.

The unit roots tests proposed by Carrion-i-Silvestre et al. (2009) are better than the existing tests in several aspects. Firstly, Carrion-i-Silvestre et al. (2009) have allowed multiple breaks in the trend function i.e., more than one changes in both the intercept and slope coefficients in their model. Secondly, the quasi-generalised least squares (GLS) detrending procedure has been used to formulate the tests, which ensures the tests to have local asymptotic power functions close to the local asymptotic Gaussian power function. Carrion-i-Silvestre et al. also argued that the quasi-GLS based approach offer improvement over commonly used alternative tests in small samples. Lastly, Carrion-i-Silvestre et al. (2009) have considered a variety of unit root tests, in particular the class of *M*-tests, that were introduced by Stock (1999) and Ng and Perron (2001).

Carrion-i-Silvestre et al. (2009) advocated five different test statistics to test for the null hypothesis of unit root under multiple structural breaks *viz*., $\mathrm{P}_{\mathrm{T}}^{gls}$, $\mathrm{MP}_{\mathrm{T}}^{gls}$, $\mathrm{MZ}_{\alpha}^{gls}$, MSB^{gls} and MZ_{t}^{gls}. Here, $\mathrm{P}_{\mathrm{T}}^{gls}$ stands for Gaussian point optimal statistic, $\mathrm{MP}_{\mathrm{T}}^{gls}$ represents modified feasible point optimal statistic, and $\mathrm{MZ}_{\alpha}^{gls}$, MSB^{gls} and MZ_{t}^{gls} are M-type test statistics based on Ng and Perron (2001) computed using GLS-detrending methods (see, Carrion-i-Silvestre et al. (2009) for details). The asymptotic critical values for all the aforementioned test statistics are obtained using bootstrap procedure. Therefore, the rejection of the null hypothesis suggests that the time series is free from unit roots and having multiple structural breaks in the deterministic trend function.

3.3. Proposed modelling approach

In the modelling approach proposed here, the following steps are carried out to find the data generating process of each of the three variables under study.

STEP 1: First, the sequential testing procedure of Kejriwal and Perron (2010) is applied to test the presence of multiple structural breaks in the deterministic trend function of a variable and then to find the estimated break points.

STEP 2: Based on the finding on the number of breaks from STEP 1, the tests suggested by Carrion-i-Silvestre et al. (2009) are performed to test the null hypothesis of unit roots against the alternative of 'no unit roots' with deterministic trend breaks being present under both the hypotheses. Rejection of the null hypothesis implies that the variable is stationary with breaks in its trend function.

STEP 3: Based on the findings that the variable under study is stationary and have trend breaks, we proceed to model the underlying trend in the series by the following equation.

$$y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{m} \gamma_j DU_{jt} + \sum_{j=1}^{m} \delta_j DT_{jt} + \varepsilon_t \qquad (3.1)$$

where it is assumed that the time series model in (3.1) is linear piece-wise i.e., it is linear in each of the sub-periods characterised by the break points. This is quite a common assumption to start with. This has the advantage that the nature of the relationship over the entire time period under study thus becomes a nonlinear function of time provided, of course, there is at least one structural break with different slope and/or intercept parameter in the deterministic trend function. The parameters β_0 and β_1 in model (3.1) are intercept and trend coefficients irrespective of breaks, respectively. In this equation, γ_j and δ_j denote the coefficients of intercept and trend dummy variables, respectively, for j^{th} break point, m is the total number of deterministic trend breaks obtained from the Kejriwal-Perron test. The dummy variables are defined as

$$DU_{jt} = \begin{cases} 1 & \text{if } t > j\text{th break date} \\ 0 & \text{if } t \leq j\text{th break date} \end{cases} \text{ and } DT_{jt} = \begin{cases} (t - j\text{th break date}) & \text{if } t > j\text{th break date} \\ 0 & \text{if } t \leq j\text{th break date} \end{cases}.$$

The actual (total) slope of this trend model at the mth break point is $\beta_1 + \sum_{j=1}^{m} \delta_j$. In case it is found that the actual slope at the last break point is negative irrespective of the preceding ones, it means the time series is in the phase of a declining trend. And hence in case of DNC, for instance, it suggests that the phase towards containment of this infection has started from the last break point. It could then be concluded that in due course of time the disease is expected to be contained provided that this declining trend continues.

This model in (3.1) is estimated by the method of least squares, and the residuals of the estimated model are also obtained. Thereafter the Ljung-Box test (Ljung and Box, 1978) on the residuals is done to test whether the errors are autocorrelated or not. If the residuals are found to be autocorrelated, we proceed to STEP 4 for further modelling of the autocorrelation, if any, in the stationary residuals. In case residuals are found to be white noise, obviously no further analysis is done.

STEP 4: The correlogram analysis of Box and Jenkins is applied to find the appropriate stationary model (autoregressive (AR) / moving average (MA) /ARMA) for the residuals. Furthermore, we check if there is any significant break in this stationary model which essentially means significant change(s) in the underlying autocorrelation process of these residuals, by applying the well-known test due to Bai and Perron (2003).

4. EMPIRICAL FINDINGS

We have chosen four (Brazil, India, Italy, and the UK) from amongst the highly- affected COVID-19 countries up to the time when the study was undertaken in 2020. This choice has been primarily from consideration of economic development since public health measures and other steps required to be taken in such pandemic situations depends largely on the health infrastructure and availability of fund. To that end, we selected two developed (Italy and the UK) and two developing nations (Brazil and India). This would enable us to examine how the spread and the effects of COVID-19 differs in these two groups of countries characterized by their development status *viz.*, developed and emerging. Further, as stated in Section 1, the choice is guided by the purpose to study the efficacy of the proposed model[3] regardless of the phase of infection prevalent in a country. It appears from the plots that these four countries have somewhat different phases of infection.

It may also be mentioned here that from visual inspection of the plots of the three time series i.e., number of daily new COVID-19 cases, number of daily total COVID-19 cases, and number of daily deaths due to COVID-19, it looks like that the data generating processes of these three series for each of the chosen four countries are quite different, and hence the choice of these three variables for this study. Also, these are very important variables to study from consideration of containing the spread of this pandemic.

The data for Brazil and India have been taken from the official website of CEIC (https://www.ceicdata.com/en). As for the other two countries in our study, namely, Italy and the UK, data have been obtained from Worldometer

[3] For sake of convenience of expression, by ‘proposed model’ we mean the model based on the procedure of data analysis proposed in this paper.

(https://www.worldometers.info/coronavirus/). These websites collect the data on COVID-19 of many countries including these four from the official websites of the Health Ministry of respective countries. Data have been taken from the available first time point itself for any of these countries unless the number of new cases/number of deaths is very low and hardly changing. For instance, in case of India, the starting dates, although available from 20 February 2020 for the number of daily total COVID-19 cases and the number of daily deaths are taken to be 4 March 2020 and 15 March 2020, respectively because of this reason. To be specific, the starting dates for the time series on daily new cases for Brazil, India, Italy and the UK are 10 March, 5 March, 2 March and 9 March 2020, respectively; on daily total cases the starting dates are 9 March, 4 March, 1 March, 8 March 2020, respectively, for these countries; and on daily deaths the starting dates are 18 March, 15 March, 2 March and 9 March 2020, respectively. However, the last time point considered for this study is 19 May 2020, and this is same for all three series for all the four countries.

We first present the plots of each of the three series under study, namely, number of daily new COVID-19 cases, number of daily total COVID-19 cases, and number of daily deaths due to COVID-19. Henceforth, these three variables would be referred to in their abbreviations as DNC, DTC and DD, respectively. These plots for the four countries, Brazil, India, Italy and the UK are given in Figures 4.1 through 4.12. We discuss the nature of the plots along with the breaks founds in these time series (*vide* Table 4.1 below) in the next paragraph.

Figure 4.1 DNC (Brazil)

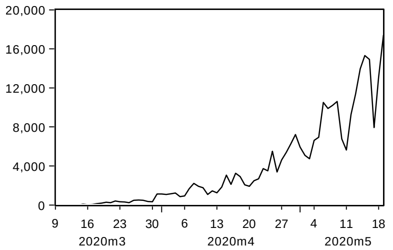


Figure 4.2 DTC (Brazil)

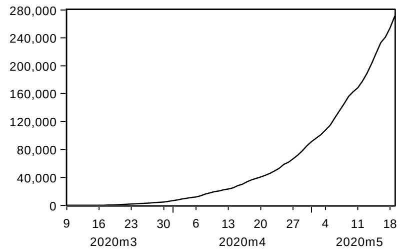


Figure 4.3 DD (Brazil)

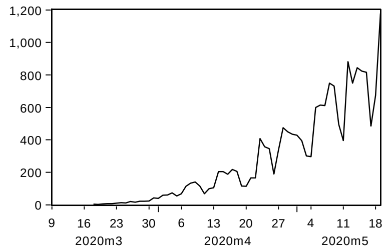


Figure 4.4 DNC (India)

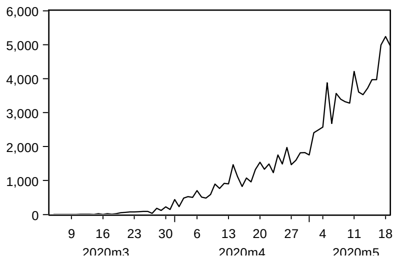


Figure 4.5 DTC (India)

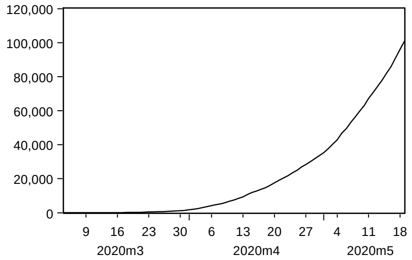


Figure 4.6 DD (India)

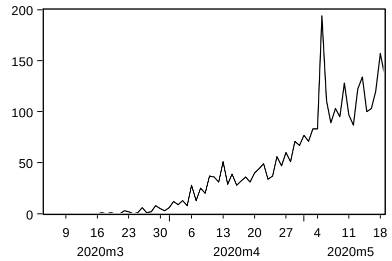


Figure 4.7 DNC (Italy)

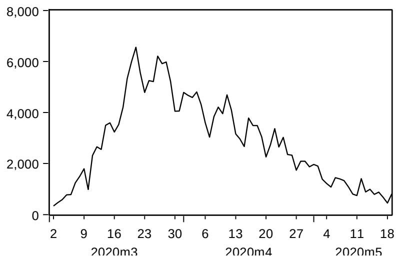


Figure 4.8 DTC (Italy)

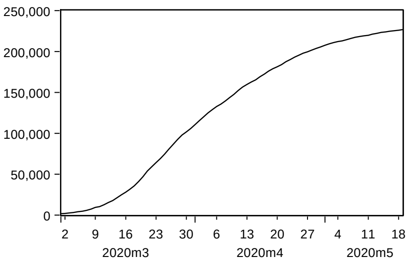


Figure 4.9 DD (Italy)

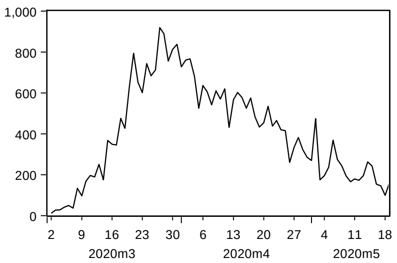


Figure 4.10 DNC(the UK)

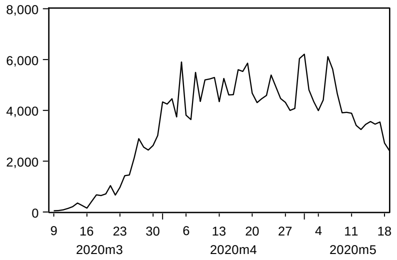


Figure 4.11 DTC(the UK)

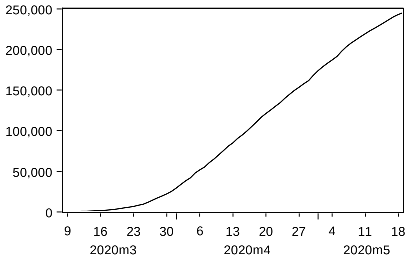
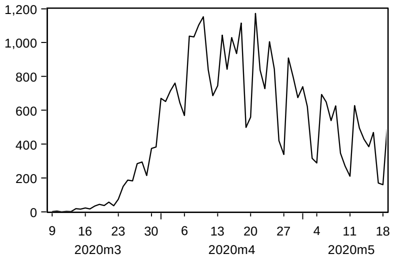


Figure 4.12 DD(the UK)



4.1. Findings of the Kejriwal-Perron test for determining breaks

We have applied the sequential testing procedure of Kejriwal and Perron (2010) to detect the number of trend breaks in DNC, DTC and DD series. In their paper, Kejriwal and Perron (2010) discussed that the maximum number of breaks, denoted as l, should be decided with regard to the available sample size. Otherwise, sequential procedures for detecting trend breaks will be based on successively smaller data sub-samples (as more breaks are allowed) thereby leading to low power and/or size distortions. It is therefore important to allow for a sufficient number of observations in each segment and choose the maximum number of permissible breaks accordingly. Further, the codes available for actual computations allow for a maximum of two breaks. Thus, we carried out this test with $l = 2$. The test statistics values *viz*., $ExpW$ and $ExpW(2|1)$, are reported in Table 4.1. It may be stated here that if existence of two structural breaks are found in any time series by this test, we performed the well-known Bai-Perron test for trended series to find if there are more than two breaks in the series.

Panel A: Daily New Cases (DNC)			
	$ExpW$	$ExpW(2\|1)$	Break dates
Brazil	105.63*	13.97*	[April 19, May 8]
India	14.35*	82.95*	[March 26, May 1]
Italy	701.44*	7.72*	[March 18, May 2]
The UK	16.37*	16.34*	[March 31, April 29]
Panel B: Daily Total Cases (DTC)			
	$ExpW$	$ExpW(2\|1)$	Break dates
Brazil	91.66*	43.24*	[April 13, May 3]
India	134.88*	73.07*	[April 8, May 2]
Italy	12.27*	28.41*	[March 12, April 17]
The UK	91.87*	297.49*	[March 28, May 7]
Panel C: Daily Deaths (DD)			
	$ExpW$	$ExpW(2\|1)$	Break dates
Brazil	18.03*	6.23*	[April 22, May 4]
India	15.05*	17.65*	[April 26, May 6]
Italy	251.67*	18.50*	[March 19, March 31]
The UK	82.92*	59.21*	[March 21, April 6]

Table 4.1 Results on sequential break tests of Kejriwal-Perron.
* indicates significance at 1% level of significance.

The results presented in Table 4.1 above show that all the series have two breaks in their respective trend functions. In case of daily new cases (DNC) for Brazil and India, the first break in the series has happened on April 19 and March 26, respectively, whereas the second break date is May 8 for Brazil and May 1 for India. It is worth noting from Figures 4.1 and 4.4 that the DNC series of Brazil and India started increasing slowly with very few cases for some initial days, the number rose steadily after their respective first break point and finally the rise became very sharp and this increasing trend continues even after their second breaks. But, in case of Italy (*vide* Figure 4.7), after the continuing rising phase till around the first break on March 18, the number of DNC started falling with some fluctuations from then onwards. As for the UK, the plot of DNC series in Figure 4.10 clearly shows that it first had a continuing rising pattern till it had the first break on March 31, then it stabilised, also called 'flattening of curve', till April 29 when the second break in trend was found, and since then it is showing a decreasing pattern. It thus appears from these plots that while Brazil and India have similar trend in DNC, being still in the continuing rising phase, in contrast Italy and the UK are in the path towards containing the pandemic and in that sense these are also similar.

The plots of daily total cases (DTC) for the four countries, given in Figures 4.2, 4.5, 4.8, and 4.11, indicate that as in case of DNC, its trend pattern appears similar in case of Brazil and India. Both indicate that the nature of trend function of DTC is convex. The trend behaviour exhibited in case of DTC for Italy and the UK is also somewhat similar although being different from that of Brazil and India. Here the trend function, barring the initial few time points where it appears convex, appears mildly concave in nature for Italy. For the UK, after the initial convex nature, the trend looks almost linear. As given in Table 4.1 above, there are two breaks in trend of DTC for all the four countries. It is only expected since, by definition, DTC is nothing but the cumulated values of DNC. Hence the significant changes in the trend of DNC series are likely to be found in the trend of DTC series as well. However, the estimated break dates are now different from those for DNC although, as expectedly, the two corresponding break dates for DNC and DTC are quite close.

As regards the daily deaths (DD) series, it is found from the Kejriwal-Perron test that all the four countries have two trend breaks in their respective DD series. The break dates occur almost at the same time points for Brazil and India. The two estimated break dates for Brazil are April 22 and May 4 while the same for India are April 26 and May 6. The similarity in COVID-19 situation in these two emerging countries appears similar even in case of daily death figures. From the plots in Figures 4.3 and 4.6, a rising pattern is observed in the DD series of these two countries. This means that with the surging DNC and hence DTC, the existing health care system in these two countries are under severe strain resulting in continued rise in daily deaths. However, unlike Brazil and India, the number of daily deaths rose comparatively rather sharply in Italy and the UK during the whole month of March. However, from the first week of April a declining pattern of trend in DD series is visible in these two countries, although with some fluctuations which is more prounced in case of the UK (*vide* Figures 4.9 and 4.12 for Italy and the UK, respectively). It may be seen from these two plots that the decreasing trend started from around the second break date, *viz.*, March 31 and April 6 for Italy and the UK, respectively. We conclude by stating that the estimated break dates obtained by applying the Kejriwal-Perron test is very close to what is noted from the actual plots of daily death figures. This also holds for the time series of DNC as well as DTC, although for the later it is not so clearly visible because these observations are large in magnitude but the scale in the vertical axis representing this is highly compacted. This shows the high level of performance of this test.

We conclude this section by stating that because of computational limitations as already mentioned earlier, the maximum number of breaks could not be taken to be more than two for this test. However, since we have found existence of two breaks in each time series by this test which does not require any assumption on stationary or non-stationary having unit roots for the time series under study, unlike the other tests for structural breaks, we further carried out the well-known Bai-Perron test to find if there are more than two breaks in trend in any of the series. This test showed no further breaks and thus confirmed existence of only two breaks in each of three series for all the four countries.

4.2. Findings on the Carrion-i-Silvestre et al. test on stationarity /non-stationarity

Based on the finding that all the time series under study *viz*., DNC, DTC and DD, have two breaks each in their deterministic trend functions for all the four countries, we have applied the tests suggested by Carrion-i-Silvestre et al. (2009) to conclude on the nature of stationarity / non-stationarity of the individual series. The null hypothesis here is unit roots (i.e., non-stationarity) with two breaks in deterministic trend and it is tested against the alternative of stationarity with two breaks in deterministic trend. The values of the five test statistics *viz*., $\mathrm{P}_{\mathrm{T}}^{gls}$, $\mathrm{MP}_{\mathrm{T}}^{gls}$, $\mathrm{MZ}_{\alpha}^{gls}$, MSB^{gls} and MZ_{t}^{gls}, are reported in Table 4.2. It is noted from this table that the DNC and DD series are both stationary with trend breaks for all the countries. On the other hand, all the four DTC series are found to have unit roots with two trend breaks.

	MZ_{α}^{gls}	MSB^{gls}	MZ_{t}^{gls}	P_{T}^{gls}	MP_{T}^{gls}	Decision
			Panel A: Daily New Cases (DNC)			
Brazil	-31.63*	0.12*	-3.93*	6.26*	6.26*	Stationary with trend breaks
India	-37.31*	0.12*	-4.31*	5.89*	6.01*	Stationary with trend breaks
Italy	-33.47*	0.12*	-4.09*	6.65*	6.39*	Stationary with trend breaks
The UK	-30.15*	0.13*	-3.88*	8.17	7.62*	Stationary with trend breaks
			Panel B: Daily Total Cases (DTC)			
Brazil	-5.21	0.28	-1.44	36.92	36.23	Unit root with trend breaks
India	-4.16	0.29	-1.20	42.64	42.20	Unit root with trend breaks
Italy	-1.95	0.42	-0.83	88.56	88.02	Unit root with trend breaks
The UK	-7.53	0.25	-1.92	31.50	27.52	Unit root with trend breaks
			Panel C: Daily Deaths (DD)			
Brazil	-28.19*	0.12*	-3.59	8.13	8.12	Stationary with trend breaks
India	-31.65*	0.13*	-3.97*	5.56*	5.60*	Stationary with trend breaks
Italy	-38.72*	0.11*	-4.38*	5.58*	5.59*	Stationary with trend breaks
The UK	-34.79*	0.12*	-4.11*	6.66*	6.71*	Stationary with trend breaks

Table 4.2 Results of the test due to Carrion-i-Silvestre et al (2009).
* denotes the rejection of the null hypothesis of unit roots at the 5% level of significance.

4.3. Model estimation

Given the outcomes of the two tests that each of DNC and DD series follows a stationary process with two deterministic trend breaks, we estimate the trend model as specified in equation (3.1) in Section 3.3., by the method of least squares for these two series for all the four countries. The estimated values of the parameters of this model along with their significance or otherwise are presented in Table 4.3. Further, the plots of DNC along with their respective fitted values are given in Figures 4.13 through 4.16, and the same for DD in Figures 4.17 through 4.20.

Panel A: Brazil		
	DNC	DD
β_0	-490.21	-28.47
β_1	68.21*	5.90*
γ_1	-726.08	182.16**
γ_2	-2217.80**	187.02**
δ_1	335.21*	-5.62
δ_2	338.99**	18.92***

Table 4.3 contd.

Table 4.3 (contd. from previous page)

Panel B: India		
	DNC	DD
β_0	-21.00	-9.07
β_1	4.46	1.33*
γ_1	-76.89	-12.24
γ_2	481.87*	-40.17*
δ_1	47.59*	8.13*
δ_2	84.62*	-6.07*
Panel C: Italy		
	DNC	DD
β_0	-204.89	-68.35**
β_1	245.07*	26.87*
γ_1	2072.91*	218.18*
γ_2	-347.09	-142.46*
δ_1	-339.68*	-8.64
δ_2	49.30**	-30.50*
Panel D: The UK		
	DNC	DD
β_0	-593.90**	-11.38
β_1	139.63*	4.35
γ_1	2044.36*	-65.19
γ_2	1047.53*	314.40*
δ_1	-136.43*	43.46*
δ_2	-157.70*	-65.42*

Table 4.3 Estimates of time series models for DNC and DD.
*, ** and *** indicate significance at 1%, 5% and 10% level of significance, respectively.

4.3.1. Daily number of new cases (DNC)

We note from Panel A and Panel B of Table 4.3 referring to Brazil and India respectively, that the estimates of the two slope coefficients at the first and second break points, δ_1 and δ_2 , are both significant and positive for both the countries. These estimates are 335.21 and 338.99 for Brazil, and 47.59 and 84.62 for India. The estimate of β_1 is also positive being 68.21 and 4.46 for Brazil and India, respectively. This means that the slope all throughout is positive for DNC of both Brazil and India. Further, the estimated actual slope value ($i.e.,\ \beta_1 + \sum_{j=1}^{m} \delta_j$) is higher after the first break (403.42 for Brazil and 52.05 for India) and further higher (742.41 for Brazil and 129.93 for India) after the second break in both the countries. Thus, we can summarise, based on our analysis, the behaviour of trend of DNC as being in the increasing phase all throughout but with higher increases after the first and second breaks on 19 April 2020 and 8 May 2020 for Brazil, and 26 March 2020 and 1 May 2020 for India. Thus it shows that both these countries are still in the phase where the disease is actually on the way up and up. While the trend behaviour is similar in these two countries, one point worth noting is that in case of Brazil, the two estimated slope coefficients at the break points are almost the same, but in case of India the second one is almost double that of the first. This suggests that in the rising phase continuing after the first break, the additional rise in the slope at the second break point is sharper for India. This is evident in the plot of fitted DNC for India in Figure 4.14 which shows steeper slope at the second break on 1 May 2020. Hence, it may be concluded that India has taken a little time to pick up probably because of imposition of lockdown at the very early stage of infection, but thereafter the infection is rising hugely, and this is likely to continue for some time because of the observed higher slope values. We can also, based on our modelling approach, conclude that the trend is overall nonlinear but piece-wise i.e., in the sub-periods characterised by the two break points, it is linear for both Brazil and India.

Figure 4.13 Plot of DNC in Brazil and its fitted values

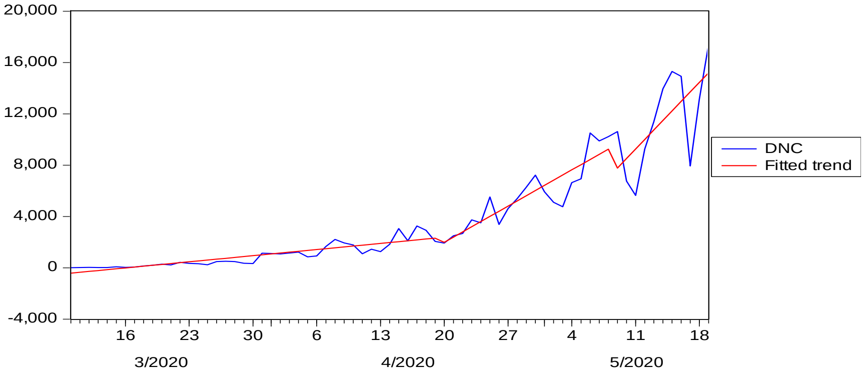
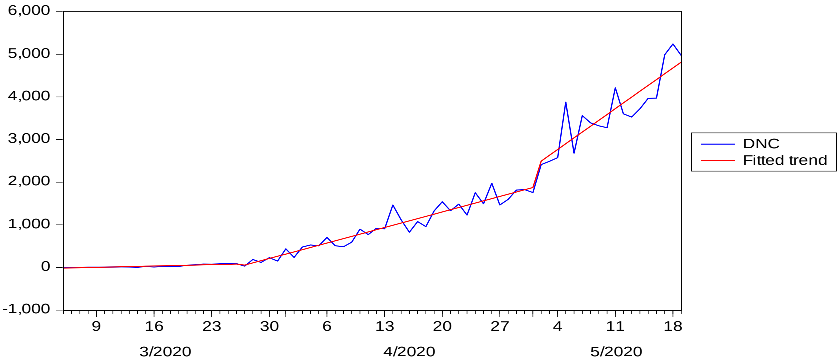
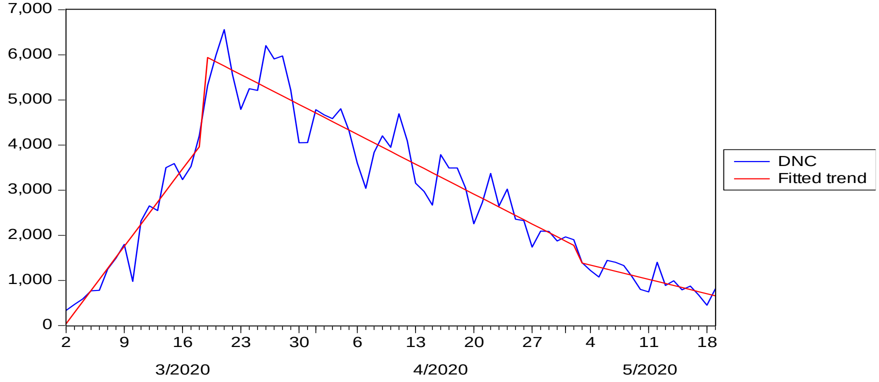


Figure 4.14 Plot of DNC in India and its fitted trend values



In case of Italy, the two estimated slopes at the two (estimated) break points, 18 March 2020 and 2 May 2020, are significant with values -339.68 and 49.30, respectively. β_1 is also significant with its estimate being 245.07. Although the additional slope at the second break point is positive, it is to be noted that the estimate of the actual (total) slope of the fitted trend curve at the second break point is - 45.31 ($i.e.,\ \beta_1 + \sum_{j=1}^{2} \delta_j = -45.31$) which is negative. In fact, the estimate of actual slope between the first and second break points is also negative being -94.61 ($i.e.,\ \beta_1 + \delta_1 = -94.61$). Thus the overall pattern of trend in DNC for Italy, as also can be seen from the plot of fitted values of DNC in Figure 4.15, is that starting with a sharply increasing trend till the occurrence of first break on 18 March 2020, the trend started decreasing right from the first break and this declining pattern continued after the second break also. Thus the conclusion about the nature of this pandemic in terms of DNC for Italy, based on our modelling approach with data up to 19 May 2020, is that there are basically two phases of this infection in Italy, and these phases are: first increasing and then decreasing, and that the pandemic is expected to be under control and is on the way towards containment provided this decreasing trend continues.

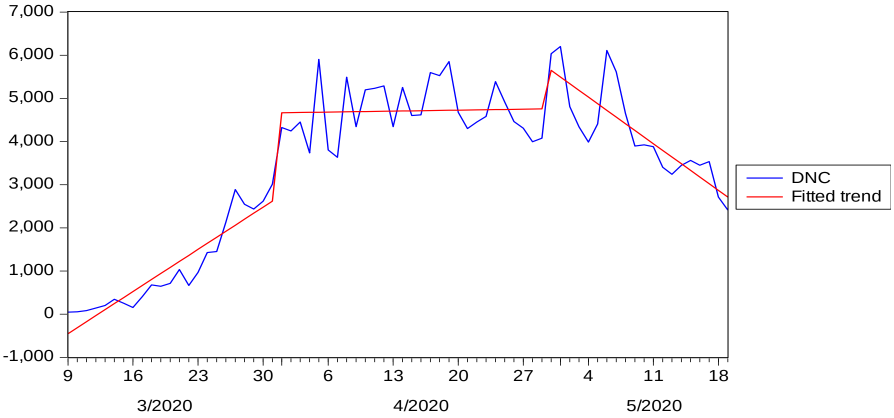
Figure 4.15 Plot of DNC in Italy and its fitted trend values



Finally, we note from Panel D of Table 4.3 that the parameter β_1 has a significant positive estimate of 139.63, which means that the trend in DNC in the UK is sharply rising till the occurrence of first break. We further note that both the slope coefficients at the two (estimated) break points are negative being -136.43 and -157.70. Thus it can be inferred that, based on data till 19 May 2020, the estimated actual slope $(i.e., \ \beta_1 + \sum_{j=1}^{2} \delta_j)$ after the second break on 29 April 2020 is negative, and hence the trend is decreasing after the second break point. Moreover, the estimated actual slope between the first and second break points is very small, although positive, being merely 3.2 $(i.e., \ \beta_1 + \delta_1 = 3.2)$, meaning that during this time period the slope is almost zero and hence there is hardly any trend in the time series in this period. This means that unlike in Italy, there is a phase in the behaviour of trend in the UK which is often called 'flattening of curve', preceding the phase of decreasing trend. This can be seen clearly from the plots of the fitted values of DNC for the UK (*vide* Figure 4.16). We find that the series is first rising and this continuously rising nature of DNC happens till around the first break date of 31 March 2020 and then in the next phase of infection the daily number has a tendency to remain more or less the same till the second break date of 29 April 2020 and then it starts falling after the second break date of 29 April 2020.

Thus, our proposed time series modelling approach with observations on number of daily new cases in the UK shows that there are three phases of this pandemic in the UK. These phases for this series are found as: increasing phase to the phase of no trend and finally the phase of decreasing trend. The overall nature of trend is nonlinear although it is linear in the three phases of this pandemic characterised by two statistically significant changes in the series. Given these findings, it may be concluded that the pandemic is moving in the likely path towards containment in the UK, much like the same in Italy, provided this declining trend continues.

Figure 4.16 Plot of DNC in the UK and its fitted trend values



4.3.2. Daily total cases (DTC)

We have noted from Tables 4.1 and 4.2 that the trend model for the series of daily total cases has a unit root with two breaks in deterministic trend for all the four countries. It may also be noted that, by definition, DNC series is nothing but the first difference values of DTC, i.e., $\mathrm{DNC}_t = \mathrm{DTC}_t - \mathrm{DTC}_{t-1}$. Thus DTC is an integrated process (of order 1) of DNC**.** Hence, once the unit root is removed by taking the first difference, the series reduces to the DNC series. Since we have discussed about the modelling of DNC, we will not be reporting anything further on DTC model except noting that the estimated DTC model for any country would typically be: $\text{Estimated DTC}_t = \text{Estimated DTC}_{t-1} + \text{Estimated DNC}_t$

4.3.3. Daily deaths (DD)

We now discuss the estimated time series models explaining trend underlying the DD series of the four countries. Starting with Brazil, we note from Panel: A of Table 4.3 that the estimated slope coefficients at the two break points, 22 April 2020, and 4 May 2020, are significant with values -5.62 and 18.92, and that the estimate of β_1 is 5.90. Looking into the estimates of the actual slopes, we can conclude that the daily deaths in Brazil started with slow increase till the first break point where the slope is almost zero 0.28 ($i.e.,\ \beta_1 + \delta_1 = 0.28$), and then finally again increasing with the slope being 19.20 ($i.e.,\ \beta_1 + \sum_{j=1}^{2} \delta_j = 19.20$) at the second break point. Figure 4.17 clearly shows that the slope continues to rise even at the second break point and beyond, and hence it can be concluded that daily deaths is in the increasing phase. Since the slope at the last break point is yet to have negative value, it would mean that by our modelling approach and using data till 19 May 2020, nothing can be concluded on likely change in trend towards containment of the disease in terms of daily deaths in Brazil.

As regards India, it is obvious to note from the estimates of β_1 (1.33) and also δ_1 and δ_2 (8.13 and -6.07, respectively), all of which are significant**,** that the conclusion for India in the same as that of Brazil. Figure 4.18 also shows clearly the rising pattern all throughout since the slope $.e.,\ \beta_1 + \sum_{j=1}^{2} \delta_j$, is positive in the whole time period although at the second break point on 6 May 2020, the additional slope is negative.

Hence we may conclude that insofar as Brazil and India are concerned, not only the DNC series, but also the DD series figures show no sign of any declining trend. In fact, the rising nature of both the series is quite alarming. Thus, it is a very serious challenge to both the Brazil and Indian governments to take effective intervention measures and improve health-care infrastructure sufficiently in order that the pandemic could be contained in the near future.

Figure 4.17 Plot of DD of Brazil and the fitted trend values

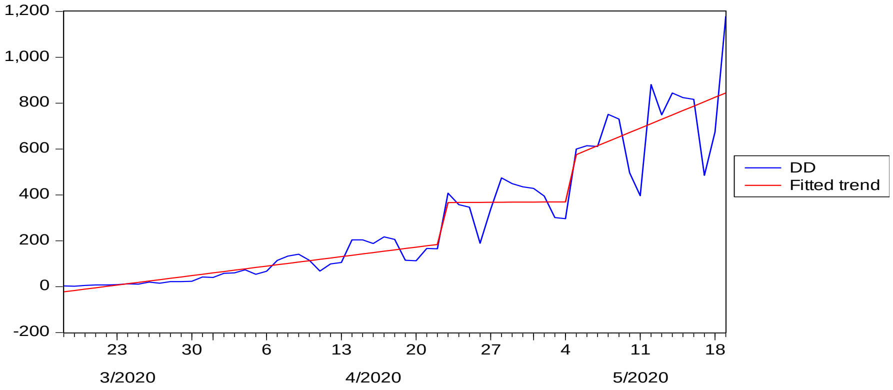
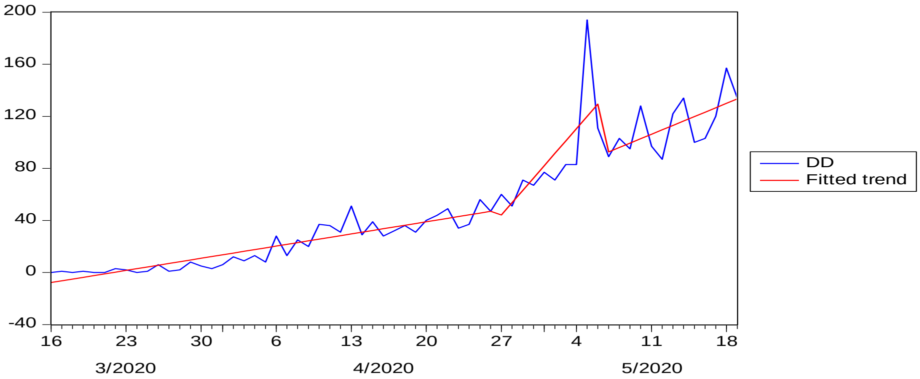
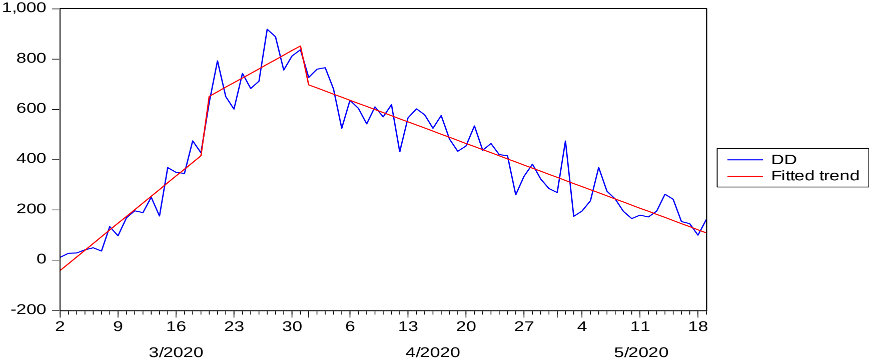


Figure 4.18 Plot of DD of India and the fitted trend values



Insofar as daily deaths in Italy are concerned, we find from the estimated results as shown in Panel: C of Table 4.3, a downward trend in the DD series. The estimates of the three slope coefficients β_1, δ_1 and δ_2 are found to be 26.87, -8.64 and -30.50, respectively. It is thus found that slope in the trend model for DD is positive till the first break point. Thereafter it has still remained positive, being 22.23$(i.e., \ \beta_1 + \delta_1 = 22.23)$, till the second break occurs. But from the second break it is negative with the value being -8.27$(i.e., \ \beta_1 + \sum_{j=1}^{2} \delta_j = -8.27)$. Thus we can conclude that our estimated trend model for daily deaths in Italy clearly shows a significant downward trend from the second break point of 31 March 2020. This is also evident for the plot (*vide* Figure 4.19) of fitted trend values of daily deaths. Thus, we note that in case of daily deaths also, Italy is in declining or containing phase. The overall nonlinearity of this trend is also clearly obvious and visible from the plot.

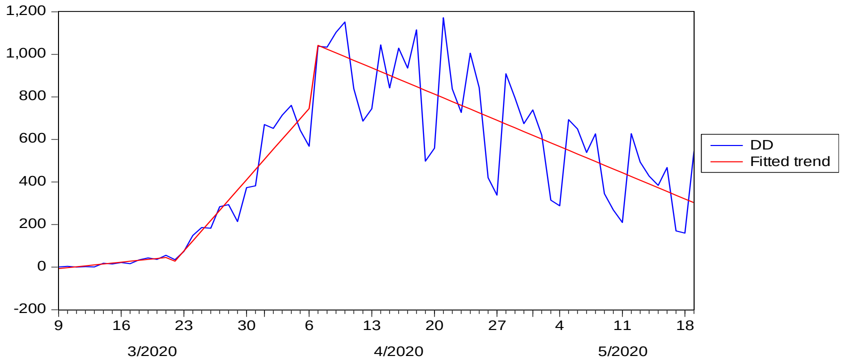
Figure 4.19 Plot of DD of Italy and the fitted trend values



Finally, in case of the UK, we find that both the slope dummies at the two break points are significant (*vide* Panel: D of Table 4.3). These estimated coefficients are noted to be 43.46 and

-65.42. The estimate of β_1 is also significant and positive (4.35). It is thus noted that starting with a very slowly rising trend, daily deaths in the UK picked up substantially after the first break on 21 March 2020 till the second break on 6 April 2020. Since then the DD series in the UK is maintaining a declining trend. Thus, by our modelling approach, we can conclude that daily deaths are now in the decreasing phase in the UK. The same is clearly visible from Figure 4.20 as well.

Figure 4.20 Plot of DD of the UK and the fitted trend values



Based on the findings on trend in the time series on daily deaths as well as daily new cases in Italy and the UK, we may conclude that the trend in both the series are clearly declining since their last observed (here second) breaks in Italy as well as in the UK. The overall conclusion that can be drawn about these two countries on their management of this pandemic is, therefore, encouraging in the midst of very gloomy situation in most of the countries all over the world. This speaks volume about the COVID-19 management in these two countries. Early reports suggested lack of seriousness in interventions by the two governments in terms of lockdown, testing and tracking of probable cases in the initial days. It was also reported that the treatment received by the afflicted patients were not good enough despite both the countries having very good in overall health-care infrastructure. However, it goes to the credit of the two governments, and the doctors, nurses and other health- care personnel along with public–health officials that they put together the best of services, and this has resulted in this stupendous success in almost containing this dreaded disease both in terms of DNC and DD. Since the nature of transmission of this virus, as virologists and epidemiologists opine, is such that once the infection has reached its peak, DNC starts remaining more or less the same and then decreasing or decreasing right from the peak point, it is very unlikely to suddenly start rising up once again unless there is a resurgence. Hence, it is expected that complete containment of this pandemic in Italy and the UK would happen sooner than later.

4.3.4. Study of residuals of trend models for DNC and DD

We obtained the residuals of the estimated trend models for the DNC and DD series of each of the four countries. Thereafter the Ljung-Box test, denoted by $Q(.)$, was applied to test if residuals of any of the series is white noise. It is noted from Table 4.4 that the residuals of the fitted models for DNC are white noise for India and the UK, and also the residuals of DD model for India and Italy are white noise. Significant autocorrelation has been observed in the residuals of the remaining series.

Panel A: Brazil		
	DNC	DD
$Q(1)$	3.78	0.43
$Q(5)$	27.93*	15.13*
$Q(10)$	41.44*	40.74*
Panel B: India		
	DNC	DD
$Q(1)$	0.51	2.29
$Q(5)$	8.73	4.61
$Q(10)$	16.10	13.63
Panel C: Italy		
	DNC	DD
$Q(1)$	5.44	0.03
$Q(5)$	36.75*	4.70
$Q(10)$	75.41*	13.61
Panel D: The UK		
	DNC	DD
$Q(1)$	4.43	0.53
$Q(5)$	11.40	23.89*
$Q(10)$	20.27	85.99*

Table 4.4 Results of the Ljung-Box Test.

* indicates significance at 1% level of significance. $\mathrm{Q}(.)$ denotes the Ljung-Box test statistic.

Now, as stated in STEP 4 in Section (3.3), we first applied the well-known Bai-Perron test for detecting the presence of break in the autocorrelation structure in the residuals in any of these remaining four series viz., DNC (Brazil), DD (Brazil), DNC (Italy), and DD (the UK), and the results showed no such breaks in these stationary series[4]. Then we employed the Box-Jenkins procedure to model the stationary residuals of these four series. Through the analysis of autocorrelation function (ACF) and partial autocorrelation function (PACF), we found that the residuals of DNC and DD series for Brazil follow an AR (3) and AR (5) processes, respectively. Similarly, the DNC series for Italy and DD series for the UK follow an AR (4) and AR (7) processes, respectively. The estimates of these models, the general form of which is given in equation (2), are reported in Table 4.5 below.

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \cdots + \phi_5 x_{t-p} + u_t \qquad (2)$$

where p is the optimal lag order of the AR model. Further testing of the residuals of all these stationary models by the Ljung-Box test showed that all these residuals have become white noise.

[4] For brevity of space, results of the Bai-Perron test are not given.

	DNC (Brazil)	DD (Brazil)	DNC (Italy)	DD (The UK)
ϕ_0	12.28	-2.72	-6.73	0.27
ϕ_1	0.25**	0.02	0.16	-0.09
ϕ_2	-0.32*	-0.69*	-0.29**	-0.34**
ϕ_3	-0.33*	-0.35**	-0.14	-0.22***
ϕ_4	--	-0.21	-0.39*	-0.15
ϕ_5	--	-0.59*	--	-0.30**
ϕ_6	--	--	--	-0.07
ϕ_7	--	--	--	0.48*
$Q(1)$	0.61	0.51	0.37	0.23
$Q(5)$	6.19	3.66	5.01	2.90
$Q(10)$	13.44	20.03	11.29	18.72

Table 4.5 Estimates of the AR models for the residuals of DNC and DD series.
*, ** and *** indicate significance at 1%, 5% and 10% level of significance, respectively. $Q(.)$ denotes the Ljung-Box test statistic.

5. FORECAST PERFORMANCE OF THE PROPOSED MODEL

It has been argued in Section 1 as to why the proposed method of modelling is useful and appropriate in explaining the underlying data generating process of the three time series of DNC, DTC, and DD for a country. The findings presented in the preceding section have clearly established how closely the fitted trend models explain the observed behaviour of these time series. In Section 1, it was also stated that we would study the forecast performance of our model and compare it with those for a few standard trend models by using the usual forecast performance criteria. For the purpose of comparison, we have chosen two other standard trend models *viz.*, polynomial of suitable degree and exponential trend model both of which are being used by researchers for finding trend as well as for forecasting future values of variables concerning this pandemic.

To that end, we have estimated these two models by standard methods of estimation and then obtained both in-sample and out-of-sample forecasts values of the three variables, DNC, DTC and DD, at daily level. For out-of-sample forecasts, the hold-out period has been taken to be 20 May – 4 June 2020. In case of forecasting of a time series variable, a general guideline is to take the last 15% /20% /25% of the total sample as hold-out sample depending on the actual number of observations available, and the remaining ones as in-sample observations. Given that our total sample size is rather moderate, we have taken 15% to be the size of hold-out sample, which is expected to yield reliable results for the hold-out period.

The performance of the forecasts across the three models have been compared by two most frequently used criteria, root mean squared error (RMSE) and mean absolute error (MAE). The results on the forecast performances are reported in Table 5.1. It is worth noting from this table that in terms of in-sample forecast performance, the trend model based on our proposed approach has the least value by both the RMSE and MAE criteria, indicated by * and **, respectively in the table, for all the three variables and for all the four countries. This clearly establishes the superior in-sample forecast performance of our model over the exponential trend and polynomial trend models. As pointed out earlier, the first difference of DTC is DNC, and hence the RMSE and MAE values for in-sample forecasts of DTC are the same as those of DNC.

As regards comparison among these models in terms of out-of-sample forecast performance where the hold-out sample has been taken to be 20 May - 4 June 2020, we note from the relevant entries in Table 5.1 that both the RMSE and MAE values are minimum (among the three models) for all the three series for both Italy and the UK. Further, barring DNC for Brazil and India and DD for India, the out-of-sample performance of our model is far superior than in the other nine cases. Even in those three cases where our model has higher RMSE and MAE values than the other two, the margin is smaller. We can thus conclude that considering all the three series and all the four countries together, our trend model performs better than the other two standard trend models in terms of in-sample forecasts in all cases and also in terms of out-of-sample forecasts in almost all cases.

In-sample				Out-of-sample		
Time series	Proposed model	Polynomial trend of suitable degree	Exponential trend model	Proposed model	Polynomial trend of suitable degree	Exponential trend model
Panel A: Brazil						
In-sample				Out-of-sample		
DTC	1111.47*	1480.34	98010.60	6712.72*	6842.92	1923269
	[731.71**]	[1034.04]	[41123.1]	[5685.12]**	[5685.03**]	[1662669]
DNC	1111.47*	1306.93	3156.10	5900.33	5691.72*	44022.03
	[731.71**]	[787.81]	[1499.31]	[4736.43]	[4537.54**]	[38886.55]
DD	70.72*	99.10	205.83	237.16*	307.70	2792.40
	[53.10**]	[64.37]	[109.63]	[184.39**]	[198.02]	[2499.50]
Panel B: India						
In-sample				Out-of-sample		
DTC	224.96*	363.24	37529.05	10489.04*	50141.82	778206.0
	[142.92**]	[262.36]	[15237.62]	[8727.77**]	[37234.39]	[672441.0]
DNC	224.96*	250.20	2154.61	1379.52	543.69*	33774.14
	[142.92**]	[163.29]	[977.61]	[1258.84]	[486.33**]	[29915.91]
DD	14.09*	16.52	33.95	35.54	25.82*	356.21
	[9.23]	[8.77**]	[18.43]	[23.84]	[18.11**]	[327.91]
Panel C: Italy						
In-sample				Out-of-sample		
DTC	347.82*	1450.50	94947.07	1699.10*	13355.04	701828.7
	[261.81**]	[1050.76]	[62386.31]	[1317.43**]	[10455.44]	[668142.2]
DNC	347.82*	470.72	1676.09	263.89*	9427.79	984.50
	[261.81**]	[351.94]	[1360.92]	[220.01*]	[6342.73]	[977.44]
DD	61.27*	70.46	265.34	107.09*	1094.52	314.26
	[46.92**]	[53.41]	[221.64]	[96.55**]	[801.68]	[311.56]
Panel D: The UK						
In-sample				Out-of-sample		
DTC	545.47*	801.98	139810.2	4811.32*	79015.18	1726358
	[447.98**]	[635.96]	[70692.24]	[3728.76**]	[58122.58]	[1573669]
DNC	545.47*	572.41	2496.59	898.67*	8165.13	12354.78
	[447.98**]	[457.16]	[1888.44]	[804.34*]	[6543.28]	[12004.97]
DD	99.00*	163.02	511.58	190.45*	1377.26	2219.68
	[72.23**]	[128.63]	[390.83]	[160.80**]	[1126.48]	[2134.41]

Table 5.1 RMSE and MAE values for in-sample and out-of-sample forecasts. In each entry, RMSE value is given first and MAE value is given within parentheses [] below the RMSE value. The minimum RMSE is indicated by *, whereas ** represents the minimum MAE.

6. CONCLUDING REMARKS

COVID-19 is an unprecedented disease with the virus having an extraordinary capability to spread very rapidly cutting across geographical boundaries of countries. To deal with this pandemic, governments all over the world are concerned with managing COVID-19 effectively by way of imposing several intervention measures to contain the disease and also doing their best in terms of providing adequate health – care facilities to the coronavirus afflicted patients so that number of deaths is minimum and recovery is very high. Reliable forecasts of important and relevant variables are absolutely essential for effective management of this huge crisis. It is also very useful if it is possible to have an idea about the phase of this infection at any point of time and the likelihood of the disease moving towards control/containment. From the modelling point of view, virologists, epidemiologists and medical researchers are working to develop bio-mathematical models to understand the dynamics of the pandemic, while other scientists and researchers are trying to obtain appropriate statistical models to analyse and predict its spread. Keeping these in mind, in this paper, we have proposed a modelling procedure based on application of tools of time series analysis, which would serve these purposes.

To be specific, the modelling approach proposed here allows for structural breaks/changes in the underlying time series and also considers trend in the time series being deterministic and/or stochastic. This modelling procedure enables identification of different phases of infection — increasing, remaining more or less the same, and decreasing/containing. The actual (total) slope of the trend function at the last (estimated) break point being negative is indication of the pandemic moving towards control/containment provided the same trend continues. This can obviously be checked with the estimates of slope coefficients at the estimated break points being available after the model has been estimated. Three important variables concerning COVID-19 have been considered in this paper. These are: (i) the number of daily new cases (DNC), (ii) the number of daily total cases (DTC), and (iii) the number of daily deaths (DD). This model has been applied to four countries *viz*., Brazil, India, Italy and the UK. The forecasting performance of the proposed trend model has also been studied *vis a vis* two other standard trend models which are often used in studying trend behaviour in case of this as well as other such infections.

We have found existence of two structural breaks in each of the three time series for all the four countries. Also, both DNC and DD series of each country have been found to be stationary with two deterministic trend breaks, while DTC series for all the four countries have been found to have unit roots with two brakes in deterministic trend. As regards the findings on the phase of infection in these countries using data till 19 May 2020, we have found that both Brazil and India are in increasing phase with infection rising up and further up, but Italy and the UK are in decreasing/containing phase suggesting that these two countries are expected to be free of this pandemic in due course of time provided their respective trend continues. The forecast performance of this model also clearly establishes its superiority, in almost all cases, as compared to two standard trend models *viz*., polynomial and exponential trend models for both in-sample and out-of-sample forecasts.

Acknowledgement

The authors sincerely thank the anonymous reviewer for very useful comments on their paper. Errors, if any, are the sole responsibility of theirs. The authors also thank Mohitosh Kejriwal for providing them with the GAUSS codes of the Kejriwal and Perron (2010) test.

REFERENCES

Anastassopoulou, C., L. Russo, A. Tsakris and C. Siettos (2020). Data based analysis, modelling and forecasting of the COVID-19 outbreak. *J. Mach. Learn. Res*, 15, 1593-1623.

Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1), 1-22.

Calafiore, G.C., C. Novara and C. Possieri (2020). A time-varying SIRD model for the COVID-19 contagion in Italy. *Annual Reviews in Control*, 50, 361-372.

Carrion-i-Silvestre, J.L., D. Kim and P. Perron (2009). GLS-based unit root tests with multiple structural breaks under both the null and the alternative hypotheses. *Econometric Theory*, 25(6), 1754-1792.

Chakraborty, T. and I. Ghosh (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals*, 135, 109850.

Chintalapudi, N., G. Battineni and F. Amenta. (2020). COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. *Journal of Microbiology, Immunology and Infection*, 53(3), 396-403.

Dickey, D.A. and W.A. Fuller, (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a), 427-431.

Fanelli, D. and F. Piazza (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals*, *134*, 109761.

Kejriwal, M. and P. Perron (2010). A sequential procedure to determine the number of breaks in trend with an integrated or stationary noise component. *Journal of Time Series Analysis*, 31(5), 305-328.

Kim, D. and P. Perron (2009). Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses. *Journal of Econometrics*, 148(1), 1-13.

Kucharski, A.J., T.W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, ... and S. Flasche (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(5), 553-558.

Ljung, G.M. and G.E. Box (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297-303.

Mandal, M., S. Jana , S.K. Nandi, A. Khatua, S. Adak and T.K. Kar (2020). A model based study on the dynamics of COVID-19: Prediction and control. *Chaos, Solitons & Fractals*, 136, 109889.

Nabi, K.N. (2020). Forecasting COVID-19 pandemic: A data-driven analysis. *Chaos, Solitons & Fractals*, 139, 110046.

Nesteruk, I. (2020). Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China. *MedRxiv*, 2020-02.

Ng, S. and P. Perron (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, 69(6), 1519-1554.

Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica: Journal of the Econometric Society*, 1361-1401.

Perron, P. and T. Yabu. (2009). Testing for shifts in trend with an integrated or stationary noise component. *Journal of Business & Economic Statistics*, 27(3), 369-396.

Rafiq, D., S.A. Suhail and M.A. Bazaz (2020). Evaluation and prediction of COVID-19 in India: A case study of worst hit states. *Chaos, Solitons & Fractals*, 139, 110014.

Ribeiro, M.H.D.M., R.G. da Silva, V.C. Mariani and L. dos Santos Coelho. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fractals*, 135, 109853.

Simha, A., R.V. Prasad and S. Narayana (2020). A simple stochastic sir model for covid 19 infection dynamics for karnataka: Learning from europe. *arXiv preprint arXiv:2003.11920*.

Singhal, A., P. Singh, B. Lall and S.D. Joshi (2020). Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. *Chaos, Solitons & Fractals*, 138, 110023.

Stock, J.H. (1999). A class of tests for integration and cointegration. *Cointegration, Causality and Forecasting. A Festschrift in Honour of Clive WJ Granger*, 137, 167.

Tomar, A. and N. Gupta (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, 728, 138762.

Vogelsang, T.J. and P. Perron (1998). Additional tests for a unit root allowing for a break in the trend function at an unknown time. *International Economic Review*, 1073-1100.

Wu, J.T., 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), 689-697.

Yonar, H., A. Yonar, M.A. Tekindal and M. Tekindal (2020). Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. *EJMO*, 4(2), 160-165.

Zhang, X., R. Ma and L. Wang (2020). Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. *Chaos, Solitons & Fractals*, *135*, 109829.

Zivot, E. and D.W.K. Andrews (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, 20(1), 25-44.