J. Inst. Eng. India Ser. B
https://doi.org/10.1007/s40031-021-00585-7

ORIGINAL CONTRIBUTION

Time Series Analysis of COVID-19 Data to Study the Effect of Lockdown and Unlock in India

Saswat Singh[1] · Chandreyee Chowdhury[1] · Ayan Kumar Panja[1] · Sarmistha Neogy[1]

Received: 15 August 2020 / Accepted: 24 March 2021

Abstract The ongoing COVID-19 pandemic has caused worldwide socioeconomic unrest, forcing governments to introduce extreme measures to reduce its spread. Being able to accurately forecast the effect of unlocking in India would allow governments to alter their policies accordingly and plan ahead. The study investigated prediction forecasts using the ARIMA model on the COVID-19 data on the lockdown period and the unlock period. In this work, we have considered not only the number of positive COVID cases but also considered the number of tests carried out. The time series data sample was collected till June 2020, and the prediction and analysis are done for August 2020. The model developed and the forecasted results align very closely with the actual number of cases, and some important inferences have been drawn through the experimentation.

Keywords COVID-19 · Time series · ARIMA · Trend · Forecast

Introduction

COVID-19 is an infectious disease caused by the Coronavirus, biologically known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease was first spotted in the capital of China's Hubei Province, Wuhan district, in December 2019 and has spread throughout the world and became a global pandemic. The COVID-19 cases are growing ever since it broke into India. Patients are suffering from respiratory failure from acute respiratory distress syndrome (ARDS) which is the major cause of death. The healthcare personals are one of the majorly affected individuals during this pandemic. A study of the characteristics of people affected in the healthcare domain was published by CDC COVID-19 response team in [1]. Various researches [2–4] are carried out to identify features of cases with novel Coronavirus disease (COVID-19). Significant progress has also been made in the field of vaccine development [5]. After the 74 days of lockdown in India, the government has issued Unlock 1 on 8 June 2020, where various sectors of the businesses were allowed to run in a restricted manner. India is among a group of most risked countries where relaxing lockdowns could lead to a spike in new infections. According to various analysis carried out, it has been observed that one of the two cases can crop up due to the effects of unlocking. The first case can be if the number of cases declines and the growth curve for number of cases flattens a positive feedback loop will dive in. The second case can be that the reopening of the economy can accelerate the number of cases which in turn will cease mobility.

Analyzing the outcome of unlock condition should be the need of the hour research domain. In this work, we have explored various time series models for the prediction of positive cases. Then, we employed Auto-Regressive Integrated Moving Average (ARIMA) model for analyzing the unlocking effect and predicting the incidence of 2019-nCov disease. The main objective of the paper is to identify the effect of unlocking in India by doing a comparative study on the forecasting using the unlock and the lockdown COVID-19 data for both positive cases and the total number of tests conducted.

✉ Chandreyee Chowdhury
chandreyee.chowdhury@jadavpuruniversity.in

[1] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

Published online: 08 April 2021

Related Work

Few recent papers could be found on the analysis of infection spreading pattern in different countries [6]. There have been a lot of previous studies that employed machine learning and statistical approaches to capture the patterns and trends of a number of varying events related to infectious diseases [7]. Works can also be found on application of SIR models [8] for contact tracing and prediction of positive number of cases. However, a common assumption of such works is to treat the density of the Indian population to be homogeneous and not considering the number of tests conducted.

Time series analysis is popularly used to forecast different diseases such as SARS, Ebola, pandemic influenza and dengue. Tandon et al. [9] have performed a study by using ARIMA (2,2,2) model to predict the future number of COVID-19 cases in India and have drawn conclusion that cases in India will increase exponentially and predicted that the cases decline from their anticipated forecast. They had reflected the importance of social distancing and sanitization in effort to decrease human exposure to the virus. The motive of their study was to help the government and medical workforce to be prepared for the situation. The drawback of their study was that they had not taken into consideration the effect of number of COVID-19 test that resulted in positive cases and were unable to reflect to the point of asymptomatic cases and its effect on the spread of the virus.

Another similar study was done by Tyagi et al. [10] where they concluded to have predicted the number of COVID-19 cases till the end of month June and predicted the requirement of ICU beds, ventilators and isolation beds in India using ARIMA(1,3,1) model. The main objective of their work was to predict the medical requirement (ICU 10% of the active cases, BEDS, ventilators for 5% of the active cases and isolation bed). So, in this paper the aim is for studying the effect of lockdown and unlock data using time series prediction models considering the number of tests conducted corresponding to the number of positive cases.

Methodology Description

Dataset Details

The number of confirmed cases, recovered cases and death cases of COVID-19 infection are collected for India from the official website of prelegislative research from 12 March 2020 to 2 July 2020 (https://prsindia.org/COVID-19/cases?search-box=700023). These data are used to build predictive model for analysis.

Description of the Time Series Models Under Consideration

Time series is a series of data points indexed (or listed or graphed) in time order. Therefore, the data are organized by relatively deterministic timestamps and may be compared to random sample data containing additional information that can be extracted. Over the past several decades, a lot of effort and research output has been produced towards the development and improvement of time series forecasting models. In this work, seven different forecasting methods have been tested which are presented and analyzed on the accuracy of the prediction for accurate selection of model.

TBAT

TBATS [11] is an acronym for key features of the model:

- T: Trigonometric seasonality
- B: Box–Cox transformation
- A: ARIMA errors
- T: Trend
- S: Seasonal components

TBATS model has the capability to deal with complex seasonality (e.g., non-integer seasonality, non-nested seasonality and large-period seasonality) with no seasonality constraints, making it possible to create detailed long-term forecasts. TBATS makes it easy for users to handle data with multiple seasonal patterns. This model is preferable when the seasonality changes over time.

In case of COVID-19 data, various models had to choose such as Box–Cox transformation or with or without trend or non-seasonal model and various amounts of harmonics used to model seasonal effects or with and without ARMA (P,Q) process used to model residuals. Firstly, the data plots are looked into for any seasonal pattern, but the data did not show any seasonal pattern. To further investigate for any traces of seasonality in the COVID-19 data, we redefined the data and fit it into TBATS model taking two lengths of seasonality period, one of a fortnight and one over a month.

Based on our findings, the trend has been modeled with [0,0] seasonal harmonics. TBATS method is very generic. Under the hood, it builds and evaluates many model candidates. This results in making the processing slower. However, TBAT is usually more useful when one needs to train models for numerous parallel time series.

Prophet

Prophet [12] is a procedure for forecasting time series data based on an additive model where nonlinear trends are fit with yearly, weekly and daily seasonality, plus holiday effects. Prophet is robust to missing data and shifts in the trend and typically handles outliers well.

Prophet requires time series data to have a minimum of two columns—the time stamp and the values. With just a few lines, Prophet can make a forecast model every bit as sophisticated as the ARIMA model. It has been forecasted till the month of August (frequency considered weekly) by using Prophet. The method uses an easily decomposable time series model consisting of three main components: trend, seasonality and holidays. To forecast trend, a piecewise linear model is used because reaching a saturation point in the near future is not feasible until a vaccine is developed. Since the growth rate in India has been varying over 2–3% over the period of lockdown, it is taken to be constant for that period. Seasonality is taken as false. Holidays, as implemented in Prophet, can be thought of as unnatural events when the trend will deviate from the baseline but return once the event is over as in the case of different phases of lockdown and unlock in India.

Auto-regressive Integrated Moving Average (ARIMA)

In every time series analysis, the forecast is solely based on the past values of the series called lags. A simple overview of a model that depends on one lag or one variable in the series is depicted in equation 1.

$$Y_t = \omega + \phi \times Y_{t-1} + e_t \tag{1}$$

Here, the predicted value, Y_t, depends on the previous prediction Y_{t-1} and the error e_t calculated as the difference between the predicted and actual outcome. ϕ is the slope coefficient, and ω is the nonzero mean. The ARIMA model [13] is one of the models which directly doesn't depend on the lags or variables of the times series but depend on the error lags which is estimated by subtracting the actual outcome from the forecasted outcome. ARIMA models assume a linear correlation between the time series values and attempt to exploit these linear dependencies in observations, in order to extract local patterns, while removing high-frequency noise from the data.

Moving Average

Moving averages [14] are a simple and common type of smoothing used in time series analysis and time series forecasting. Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series moving average of order m; where $m = 2k + 1$. The estimate of the trend cycle at time t is obtained by averaging values of the time series within k periods of t. The average eliminates some of the randomness in the data, leaving a smooth trend cycle component.

With respect to the COVID-19 data, it helps to figure out the trend in the data. We take the period of 5–7 days as the number of tests keeps increasing and trend keeps changing with respect to time.

Neural Basis Expansion Analysis (N-BEATS)

N-Beats [15] is a times series forecasting model that employs a deep neural architecture consisting of forward and backward residual links along with a deep stack of fully connected layers. Generally, the model operates in the same way as traditional decomposition techniques, such as the seasonality-trend-level approach. More specifically, the architecture is comprised of two stacks: the trend stack, followed by the seasonality stack, each consisting of several blocks connected using residual connections. From the lookback period, the model learns the behavior and predicts future points in the forecast Period. This input is passed through layers of stacks. We set the forecast period to 34 data points and lookback period of 88 data points.

Single Exponential Method

Simple exponential smoothing is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha (α), also called the smoothing factor. This parameter controls the rate at which the influence of the observations at prior time steps decay exponentially. α is often set to a value between 0 and 1. This method was used for the analysis as the raw data itself do not show any clear trend or seasonal pattern.

Double Exponential Method

Double exponential smoothing [16] is an extension to exponential smoothing that explicitly adds support for trends in the univariate time series. Apart from α, another factor, β, is added to control the decay of the influence of the change in trend. The method supports trends that change in both additive (linear trend) and multiplicative (exponential trend) way.

Double exponential method had an upper edge compared to single exponential method because a change in trend is observed in the COVID-19 data. A change from linear trend to exponential trend to linear trend was observed.

Model Selection

All the models are executed on the COVID-19 dataset, and the errors are listed in Table 1. It has been observed from the root mean square error (RMSE) measures that the ARIMA model [17, 18] has the best forecasting accuracy. So, we have used ARIMA-based model to develop relationship between the lockdown and the unlock data and perform forecasting on the time series data.

Model Development

ARIMA model [19] is represented in the form of P, D, Q. Here, P stands for the order of auto-regression, D signifies the degree of trend difference, and Q is the order of moving average. Here, ARIMA model is applied to the time series data of confirmed COVID-19 cases in India. Auto-correlation function (ACF) graph and partial auto-correlation (PACF) graph are used to find the initial number of ARIMA models. These ARIMA models are then tested for variance in normality and stationarity.

Two tests are conducted to check for the stationarity of the series, Dickey–Fuller test and rolling statistics. In order to get a stationary series, a mathematical transformation in the series is applied as follows.

$$Y_t = log(X); \quad Y_t = Y_t - Y_{t-m}; \tag{2}$$

Here, X is the representation of the data, and m is lag differences. Dickey–Fuller and rolling statistics test gave the result that showed that the transformed data are stationary. The next step is to check for the ACF and PACF plots to get the value of P and Q.

1. P: The lag value where the PACF chart crosses the upper confidence interval for the first time (P = 2 for lockdown and P = 1 for unlock)
2. Q: The lag value where the ACF chart crosses the upper confidence interval for the first time (Q = 2 for lockdown and Q = 1 for unlock)

It has been observed that forecasting of positive cases considering only number of positive cases increased the error deviation. Hence, both the number of positive cases and the number of tests conducted are considered for future prediction and analysis. With these two considerable lags, our prediction model is developed. The number of forecasted test data for a single day was used to develop a chart for the number of positive COVID-19 cases with respect to the number of tests conducted. An average of the per day positivity rate (percentage of positive cases w.r.t the total tests conducted) was considered. It has been mapped the predicted value of positive cases by the ARIMA model based on the positive cases already recorded and the predicted chart by the ARIMA model based on the reported number of tests carried out per day and the average positivity rate. In order to get the total number of cases for the very day d, an average of the predicted chart was considered and the estimated positive cases which was added with the number of cases recorded till $(d - 1)$th day.

Results and Observation

A detailed overview of the ARIMA model used along with the parameters for lockdown and unlock is listed in Tables 2 and 3, respectively.

In Fig. 1, a line plot of the residual errors is presented. It suggests that there may still be some trend information not captured by the model.

In Fig. 2, a density plot of the residual error values was plotted. It can be observed from the plot that the residual error is distributed in a Gaussian manner which is centered around 0. This indicates that the model can be used for analysis of the COVID-19 dataset under consideration. The forecast results are presented in Fig. 3 along with the original plot to compare the two.

In Fig. 3, it can be observed that the predicted data plot for both the unlock and the lockdown data aligns with the

Table 1 RMSE values corresponding to the performance of different models on the COVID-19 data

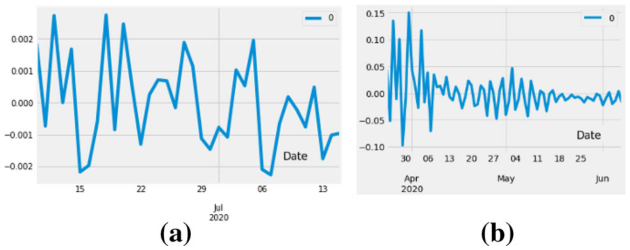
Model name	RMSE
TBAT	0.005642
Prophet	0.098641
ARIMA	0.004897
Moving average	0.018425
N-BEATS	0.251691
Single exponential method	0.022187
Double exponential method	0.020894

Table 2 ARIMA model result on lockdown data

Dep. variable:	D. Confirmed cases	**No. observations:**	76
Model:	ARIMA(5, 1, 4)	Log Likelihood:	147.201
Method:	css-mle	S.D. of innovations:	0.030
Date:	Wed, 15 Jul 2020	AIC:	− 272.402
Time:	13:03:25	BIC:	− 246.764
Sample period:	24-3-2020 to 07-06-2020	HQIC:	− 262.156

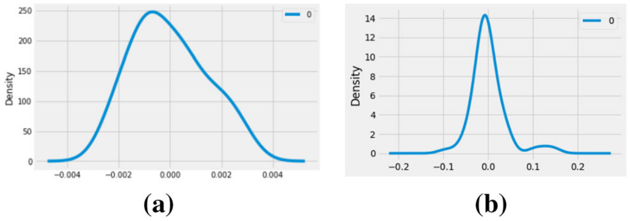
Table 3 ARIMA model result on unlock data

Dep. variable:	D. Confirmed cases	No. of observations:	36
Model:	ARIMA(1, 1, 1)	Log Likelihood:	185.340
Method:	css-mle	S.D. of innovations:	0.001
Date:	Wed, 15 Jul 2020	AIC:	− 362.679
Time:	12:51:35	BIC:	− 356.345
Sample Period:	06-10-2020 to 07-15-2020	HQIC:	− 360.468



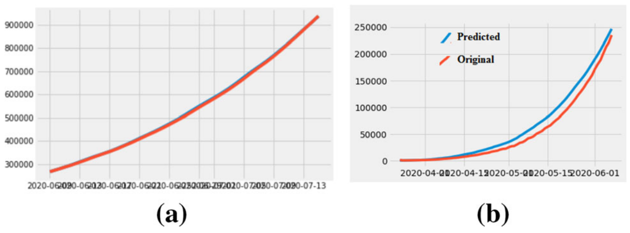
(a) (b)

Fig. 1 **a** Error residual graph for unlock. **b** Error residual graph for lockdown



(a) (b)

Fig. 2 **a** Density plot error for unlock. **b** Density plot error for lockdown



(a) (b)

Fig. 3 **a** Prediction v/s original positive cases for unlock period. **b** Prediction v/s original positive cases during lockdown in India

actual result. The two ARIMA models are also used for forecasts as shown in Fig. 4. For the lockdown data, a graph with 95% confidence interval has been observed that signifies a sharp increase in COVID-19 positive cases. The slope of the graph signifies the magnitude of the increase in cases per day. If the predicted graph is compared with the plotted graph of the number of COVID-19 cases post-lockdown, the same linear plot as the ARIMA (5,1,4) model is predicted.

From the above plots, the following observations could be made.

1. The unlock period is not the only factor for the increase in COVID-19 cases. There are numerous other lags in the system. The analysis on the lockdown model predicted the same trend in positive COVID-19 cases. This brings up some important questions: (a) Did the spread already occur in the community during lockdown? Was the lockdown effectively carried out to stop the virus spread? This question arises as some of the study suggests that among the total Coronavirus tests conducted so far, 69% were asymptomatic cases and 31% were symptomatic cases. So, another question also arises on the fact that the first cases detected in India were on January 30; hence, could it be possible that the virus had spread before the very date?
2. If unlock period were the prime factor in increase of COVID-19 cases, then we would have witnessed noises and spikes in the unlock graph on the number of cases. The only difference was that the slope of Figs. 4 and 5 is that the slope of graph in Fig. 5 is more steeper than that of Fig. 4a. So, the steeper slopes state

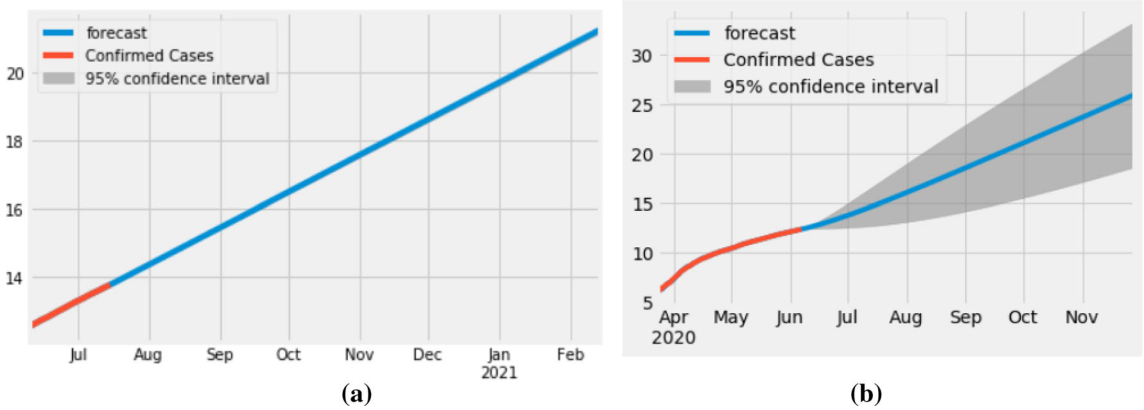
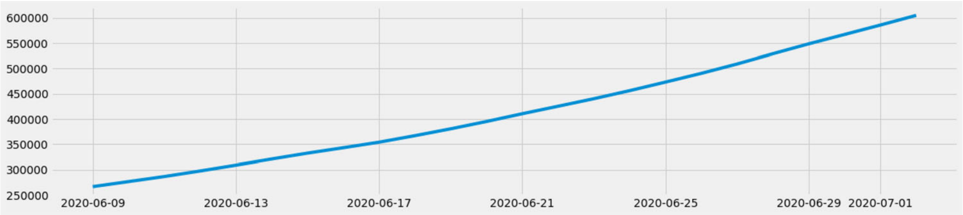


Fig. 4 **a** Forecasted result for unlock. **b** Forecasted result for lockdown

Fig. 5 Number of COVID-19 positive cases predicted for the unlock period



that the spread of COVID-19 is fast and the cases will not slow over time as per the forecasted result and the change in slope is the result of increase in number of COVID-19 tests which contribute in more cases detected per day .

With the number of tests increasing per day, an increase in the number of cases could be observed. According to the data, around 10–12% of total tests were positive COVID-19 cases (Fig6). So, in the proposed model, both predictions based on number of positive cases and number of tests conducted are merged to reduce the errors.

According to the proposed model, 518,082 total tests and 50,234 positive cases were predicted on 31 July 2020. The very model has been extended to predict till August 14 as shown in Fig. 7. It can be observed that the predicted and the actual values align and the average positivity rate stays within 7–11% both during the lockdown and the unlock.

The analysis made on both the lockdown data and the unlock data does not guarantee community spread in India.

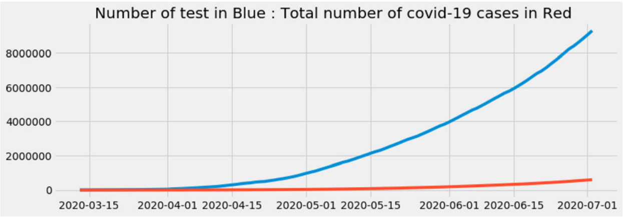


Fig. 6 Number of tests conducted 8. Number of positive COVID-19 cases

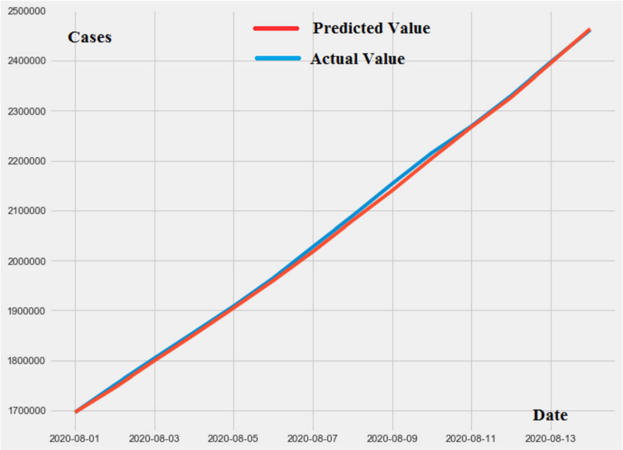


Fig. 7 Number of positive cases predicted by the proposed model vs Actual cases, from 1 August to 14 August 2020

However, region-wise community spread can be a possibility. So, the government must organize mass testing in regions where many COVID-19 patients are found.

Conclusion

This study investigates the applicability of different time series models to analyze COVID-19 data for India during lockdown and unlock. An ARIMA-based prediction model is developed that makes prediction taking into account the number of positive cases, number of tests conducted and the average positivity rate. In the present circumstances, the proposed model could be valuable in anticipating future cases of infection if the pattern of virus spread did not change abnormally. The analysis showed that the increase in number of cases per day that shoot up after lockdown was not an abnormal trend. The predicted graph based on the lockdown data had showed a significant increase in number of COVID-cases. With the ARIMA model, the forecasts are produced on the prior values of the time series and the error lags which actually helps the model to adjust its prediction values from sudden change in trends. A short-term forecasting was made in the time series and the outcome in the very near months, the ARIMA proved to be the most suitable model unlike N-Beats which requires a large amount of previously estimated timestamp samples.

The future modifications to further improve the predictive accuracy of the models will include the creation of ensembles of the presented models that would combine the best of many worlds in order to reduce the overall error as well as the adoption of multivariate time series modeling that take into account other factors that are either directly or indirectly related to the spread of the pandemic. Another future ambition would be to use some form of transfer learning in order to bring learning's from one country to another in order to know the majority parameters for the actual cause of the spread.

References

1. D. Acemoglu, V. Chernozhukov, I. Werning, M.D. Whinston, A multi-risk sir model with optimally targeted lockdown. Tech. rep., National Bureau of Economic Research (2020)
2. G.E. Box, D.A. Pierce, Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J. Am. Stat. Assoc. **65**(332), 1509–1526 (1970)
3. J. Brożyna, G. Mentel, B. Szetela, W. Strielkowski, Multi-seasonality in the tbats model using demand for electric energy as a case study. Econ. Computa. Econ. Cybern. Stud. Res. **52**(1), 229–246 (2018)
4. T.C. COVID et al., Characteristics of health care personnel with covid-19-united states, February 12–April 9, 2020. MMWR Morb. Mortal. Wkly. Rep. 2020 **69**(15), 477–481 (2020)
5. A.S. Fauci, H.C. Lane, R.R. Redfield, Covid-19-navigating the uncharted. N. Engl. J. Med. **382**(13), 1268–1269 (2020). https://doi.org/10.1056/NEJMe2002387
6. S.C. Hillmer, G.C. Tiao, An arima-model-based approach to seasonal adjustment. J. Am. Stat. Assoc. **77**(377), 63–70 (1982)
7. K. Kalpakis, D. Gada, V. Puttagunta, Distance measures for effective clustering of arima time-series. In: Proceedings 2001 IEEE international conference on data mining, pp. 273–280. IEEE (2001)
8. J.J. LaViola, Double exponential smoothing: an alternative to Kalman filter-based predictive tracking. Proc. Worksh. Virt. Environ. **2003**, 199–206 (2003)
9. T.T. Le, Z. Andreadakis, A. Kumar, R.G. Roman, S. Tollefsen, M. Saville, S. Mayhew, The covid-19 vaccine development landscape. Nat. Rev. Drug Discov. **19**(5), 305–306 (2020)
10. P. Mehta, D.F. McAuley, M. Brown, E. Sanchez, R.S. Tattersall, J.J. Manson, H.A.S. Collaboration et al., Covid-19: consider cytokine storm syndromes and immunosuppression. Lancet (London, England) **395**(10229), 1033 (2020)
11. H. Nishiura, H. Oshitani, T. Kobayashi, T. Saito, T. Sunagawa, T. Matsui, T. Wakita, M. COVID, M. Suzuki, Closed environments facilitate secondary transmission of coronavirus disease 2019 (covid-19) (2020). https://doi.org/10.1101/2020.02.28.20029272
12. C.P.E.R.E. Novel et al., The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china. Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi **41**(2), 145 (2020)
13. B.N. Oreshkin, D. Carpov, N. Chapados, Y. Bengio, N-beats: Neural basis expansion analysis for interpretable time series forecasting (2019)
14. R.C. Sato, Disease management with arima model in time series. Einstein (Sao Paulo) **11**(1), 128–131 (2013)
15. G.R. Shinde, A.B. Kalamkar, P.N. Mahalle, N. Dey, J. Chaki, A.E. Hassanien, Forecasting models for coronavirus disease (covid-19): a survey of the state-of-the-art. SN Comput. Sci. **1**(4), 1–15 (2020)
16. H. Tandon, P. Ranjan, T. Chakraborty, V. Suhag, Coronavirus (covid-19): Arima based time-series analysis to forecast near future. arXiv:2004.07859 (2020)
17. R. Tyagi, L.K. Dwivedi, A. Sanzgiri, Estimation of effective reproduction numbers for covid-19 using real-time bayesian method for india and its states (2020). https://doi.org/10.21203/rs.3.rs-45937/v1
18. N. Zhao, Y. Liu, J.K. Vanos, G. Cao, Day-of-week and seasonal patterns of pm2. 5 concentrations over the united states: time-series analyses using the prophet procedure. Atmos. Environ. **192**, 116–127 (2018)
19. J. Contreras, R. Espinola, F.J. Nogales, A.J. Conejo, Arima models to predict next-day electricity prices. IEEE Trans. Power Syst. **18**(3), 1014–1020 (2003)