International Journal of Advanced Scientific Research
ISSN: 2456-0421; Impact Factor: RJIF 5.32
Received: 04-06-2020; Accepted: 05-06-2020; Published: 13-06-2020
www.allscientificjournal.com
Volume 5; Issue 1; 2020; Page No. 24-28

Coronavirus (COVID-19) forecasting in India: Application of ARIMA and periodic regression models

Bharath Prasad Cholanayakanahalli Thyagaraju[1], Shinduja Rajamani[2], Dharshan Hebbuluse Veerabhadrappa[3], Sharanagouda Patil[4], Parimal Roy[5], Srikantiah Chandrashekar[6], Kuralayanapalya Puttahonnappa Suresh[7], Raghavendra G Amachawadi[8]

[1-3, 7] Department of Spatial Epidemiology, ICAR- National Institute of Veterinary Epidemiology and Disease Informatics, Bengaluru, Karnataka, India

[4] Department of Virology, ICAR- National Institute of Veterinary Epidemiology and Disease Informatics, Bengaluru, Karnataka, India

[5] Director, ICAR- National Institute of Veterinary Epidemiology and Disease Informatics, Bengaluru, Karnataka, India

[6] Chanre Rheumatology and Immunology Centre and Research, Bengaluru, Karnataka, India

[8] Department of Clinical Sciences, College of Veterinary Medicine, Kansas State University, Manhattan, KS, United States of America

Abstract
Coronavirus disease, COVID-19 is the deadliest pandemic, which has affected most of the countries worldwide. Disease outbreak analysis has become a priority for the Government to take healthcare measures in reducing the impact of this pandemic. In this study, we attempt to analyse the disease outbreak data collected from 4th March 2020 to 26th May 2020 in India. Auto Regressive Integrated Moving Average (ARIMA) and Periodic Regression models were employed to predict the epidemiological trend of the incidence and probable number of new cases for the next ninety days for COVID-19 in India. The total number of probable daily new cases would be increased in the future as predicted by both ARIMA and Periodic regression models. Both ARIMA and Periodic regression models are best fitted to the observed data on daily incidence of COVID-19 in India. Incidence of COVID-19 expected to increase in next ninety days allowing to employ the stringent infection control measures such as public awareness and social distancing for effective mitigation and spread of disease in India.

Keywords: ARIMA Model, Autocorrelation, COVID-19, Disease forecast, Periodic regression, Prediction

1. Introduction

The novel corona virus (severe acute respiratory syndrome 2 [SARS-CoV-2]), COVID-19, has continued to spread around the world and turned into Pandemic [1, 2]. Currently, COVID-19 has affected more than 210 countries globally according to WHO reports [1]. Due to rapid spreading potential and the absence of vaccines and drugs, the contagious COVID-19 disease has devastated the normal life across the globe. At present, corona virus has infected more than half a million population, and killed more than 3 lakh people, and forced more than 3 billion to stay in their homes [3]. The pandemic spread of COVID-19's situation is not different in India, as the numbers still seem to be on an ascent, with no clear signs of decline, even as we are near the fourth phase of lockdown. There are now 1,45,553 confirmed cases in the country with reported 4,175 deaths in India [3]. The infection rate of COVID-19 in India is reported to be 3.85%, significantly lower than the worst affected countries across the globe. The outbreak has been declared as Pandemic in more than a dozen states and union territories of India, the government has taken extraordinary steps to contain the epidemics of COVID-19, and the effects are already apparent. India has conducted over 1.85 million tests (data available during the preparation of this research work), out of which 4 per cent of the samples have tested COVID-19 positive. Due to the lack of knowledge about this virus, the COVID-19 pandemic placed tremendous catastrophe on everyone around the world. In order to prevent further transmission, India began thermal screening and strong preventive measures for passengers arriving from different countries, however, the number of infected cases are consistently increasing around the world, even after undergoing different preventive measures.

Disease modelling and analysis of incidences will help in projecting the disease probability and guides in methods to control the disease and to develop appropriate preventive healthcare measures. Mathematical approaches are widely used to infer critical epidemiological transitions and parameters of COVID-19. Methods such as epidemic curve fitting, surveillance data during the early transmission, and other epidemic models are frequently applied to predict COVID-19 pandemic across the world [4]=. A The Auto-Regressive Integrated Moving Average Model (ARIMA) for predicting the spreading of COVID-19 disease trajectories of COVID-19 for next two months was employed. Periodic regression is a type of curve that relates few variable to time and is repeated at fixed time intervals.

An effective management of emerging and re-emerging diseases needs multidisciplinary approach involving effective surveillance, rapid reporting, collection and

transport of clinical materials for diagnosis of the etiological agents. And also, strengthening of basic research, epidemiological modelling, prediction and development of forecast models, development of novel vaccine candidates and suitable adjuvant, etc. are also must to contain this pandemic [5]. There exist a large number of evidences where machine learning algorithms have proven to give efficient predictions in healthcare [6, 7, 8]. Nsoesie in his study has provided a systematic review of approaches used to forecast the dynamics of influenza pandemic. They have reviewed research papers based on deterministic mass action models, regression models, prediction rules, Bayesian network, SEIR, ARIMA forecasting models. Recent studies on COVID-19 include only exploratory analysis of the available limited data [9]. Effective and well-tested vaccines against COVID-19 has not been developed and hence a key part in managing this pandemic is to decrease the epidemic peak, also known as flattening the epidemic curve. The role of data scientists and data mining researchers is to integrate the related data and technology to understand the virus and its characteristics, which can help in taking right decisions and concrete plan of actions. It will lead to a bigger picture in taking aggressive measures in developing infrastructure, facilities, vaccines and restraining similar epidemics in future. In the present study, we evaluated the rate of disease spread and to develop a stochastic model using ARIMA and periodic regression approach to predict the daily probable new cases for the next ninety days in India (August 19, 2020). The comparison of different models to predict exact probable number of new cases for COVID-19 outbreaks in future.

2. Material and Methods

2.1 Data source

The time series data on the number of COVID cases reported was extracted from cloud source database from the official website www.covid19india.org.[3] The data for model development were updated to May 26, 2020.

2.2 Model Development and statistical analysis

2.2.1 ARIMA modelling in R

ARIMA models afford an alternative approach to time series forecasting. The two excessively used approaches in time series forecasting are Exponential smoothing and ARIMA.

ARIMA models describe the autocorrelations in the data with short-term fluctuations, meanwhile Exponential smoothing models describes the trend and seasonality in the data. Parameters of the ARIMA model were estimated by autocorrelation function (ACF) graph and partial auto correlation (PACF) correlogram. To determine the prevalence, ARIMA (2, 0, 2) was selected as the best ARIMA model of COVID-2019.[10]

It is important to make the time series stationary; the most commonly used techniques are:

1. Detrending: Here, we simply remove the trend component from the time series. For instance, the equation of my time series is:

x(t) = (mean + trend * t) + error

the parentheses will be removed to the build model.

2. Differencing is the commonly used to remove non-stationarity. The differences of the terms are modelled and not the actual term.

Given by, x(t) – x(t-1) = ARMA (p, q)

The differencing is the Integration part in AR(I)MA. The three parameters are:

p:AR, d: I, q:MA

3. Seasonality can easily be incorporated in the ARIMA model directly.

Auto-Regressive Integrated Moving Average Model (ARIMA): Moving average (MA) is the present value of series is defined as linear combination of past errors. So, assuming the errors to be independently distributed with the normal distribution. Order q is defined as:

$yt=c+\varepsilon t+\theta 1\ yt\text{-}1+\theta 2\ yt\text{-}2+\ldots\ldots\ldots\ldots\ldots+\theta q\ yt\text{-}q$ (eq. i)

where;

εt = white noise, yt-1 and yt-2 = lags. the ACF plot gives the value of order q in the MA process; at this lag instance ACF crosses the upper confidence interval for the first time. We combined differencing with Moving Average (MA) and auto-regression (AR) models and combined model can be expressed as:

$y't=c+\phi 1\ y't\text{-}1+\phi 2\ y't\text{-}2+\ldots+\phi\ p\ y't\text{-}p+\theta 1\ yt\text{-}1+\theta 2\ yt\text{-}2+\ldots\ldots+\theta q\ yt\text{-}q+\varepsilon t$ (eq. ii)

where, y′t is the differenced series. Right-hand side are the "predictors" that include both lagged values of yt and errors. This is an ARIMA (p, d, q) model, where,

p=order of the autoregressive part;

d=degree of first differencing involved;

q=order of the moving average part.

The ARIMA model provide weight to past few values and error values to corrects it model prediction, so it better than other basic regression and exponential methods.

2.2.2 Periodic Regression

In the analysis of periodicity in the time series data, the important determinants were the length of the cycle or fundamental period, its amplitude or the range from the minimal to the maximal response and angular point in time during the periodic cycle when the response is maximal [11]. These parameters were easily estimated by using many statistical software's. A time series data or an outbreak Y_t (t =1,..., N) observed at equal intervals of time was expressed as $Y_t = \hat{Y}_t + \varepsilon_t$, where is unobserved fixed value at time t and $\{\varepsilon_t\}$ is a sequence of random errors identically and independently distributed with expectation 0 and variance σ^2.

To determine variability of the disease outbreaks data whether it has periodic components, the series was approximated by finite Fourier series of the form, if the number of data was even,

if the number of data was odd: N= 2n–1

$\hat{Y}_t = A_0 + 2^{\sum_{m=1}^{n-1}(Am\cos 2\pi mf1t\ +\ Bm\sin 2\pi mf1t)+ An\cos 2\pi mf\ t1}$,

or if the number of data is odd: $N= 2n\text{-}1$

$$\hat{Y}_t = A_0 + 2\sum_{m=1}^{n-1}(Am\cos 2\pi mf1t\ +\ Bm\sin 2\pi mf1t)$$

Here $R_{m=}\sqrt{(A_m^2 + B_m^2)}$ is the amplitude, and $\varphi_m = arctg\ (B_m/A_m)$ is the phase of the i^{th} component. The function $\hat{Y}_t$ is a linear combination of sinus and cosinus functions with frequencies proportional to fundamental frequencies $f_1 = 1\ N$, so it is linear multiple regression with sinus and cosinus functions as regressors. Since $\frac{1}{N}\sum_{t=1}^{N} y_t^2 = R_0^2 + 2\sum_{m=1}^{n-1} R_m^2 + R_n^2$ contribution of i-

th harmonic component to the mean of the total sum of squares of time series is equal to $R_i^{2.}$ The harmonic components which describe the series is obtained by the mean of the total sum of squares.

With additional assumption that errors are normally distributed, the estimated periodic regression model may be tested by *F*-test and particular estimates of parameters with *t*-test. We computed the prediction of daily probable new cases for the next ninety days using ARIMA and periodic regression models. The goodness of fit can be performed for any model by using F-test and t-test respectively.

All statistical analyses and forecast model development were performed using R software version 3.6.3.

4. Results & Discussion

The results of ARIMA and Periodic regression model for prediction of COVID-19 outbreaks in India for next ninety days from May 25, 2020 till August 19, 2020 is represented in Figs. 1. and 2. respectively. The basic steps of methodology used in describing the incidence of COVID-19 in India by ARIMA are elimination of possible cyclic and seasonal behaviour to improve forecast using sequence stationary forecast. Differentiation was used to remove the seasonal behaviour and autocorrelation function was used to identify seasonal and/or cyclical component present in data. The autocorrelation and partial autocorrelation were used to identify the order of model. Non-linear least square was used for parametric estimation. Chi-squared method was used to test whether incidence series has a white-noise, if residual sequence is not a white-noise sequence, then there is no useful information extraction and indication for further improvement in the model [12].

Time series models, like Autoregressive Integrated Moving Average (ARIMA), effectively consider serial linear correlation among observed incidence, whereas, periodic regression can satisfactorily describe the non-linear cyclic trend. To predict the outbreak of COVID-19, ARIMA (2, 0, 2) was selected as the best fit model at lowest AIC value of 1118.65 (log likelihood = -605.93), the coefficients of autoregressive for lag1 and lag2 are 0.609 and 0.367 respectively. The coefficient of second order moving average was estimated at -0.564 because it absorbs the short-term fluctuations, which was fine-tuned to eliminate all residual autocorrelation. The number of daily cases reported were analysed for trends of the disease and predicted the future disease outbreaks in future. In the present study, we have taken data till the 26 May, 2020 extracted from covidindia.org to describe the ARIMA curve.

The Periodic regression that relates incidence of disease to the time and is repeated at fixed time interval and is useful for any kind of data which fluctuates in regular intervals [13]=. The three parameters like length of cycle, amplitude or the range from minimal to maximum response, and phase angle are angular point of the cycle when the response is maximum were analysed in periodic regression (estimate x=62.33(sd=3.961, t=15.713), c1=725.975 (sd=72.33, t=10.027), s1= -242.275(sd=100.30, t=-2.420), R^2=0.9561, adj. R^2=0.9542). The *F*-test was used for testing the goodness of fit of any model (Fig. 2.).

We then calculated the probable number of possible daily new cases for the next ninety days based on existing data using ARIMA and Periodic regression models. As represented in Fig. 1 & 2, the probable daily new cases with 95% confidence interval (CI) at five day's interval (between May, 25 and August 19, 2020) will be 7,110, 7,589, 8,097, 8,832, 9,077, 9,256, 9,377, 9,449, 9,479, 9,473, 9,438, 9,378, 9,297, 9,199, 9,088, 8,967 and 8,837 for ARIMA model. The total number of probable daily new cases according to ARIMA model would gradually increases at early stage of prediction and later on it flattens at the end of ninety days of prediction cycle (Fig. 1.). The predicted number of daily news cases for periodic regression model at five day's interval (between May, 25 and August 19, 2020) will be 4,960, 5,268, 5,487, 5,625, 5,705, 5,754, 5,804, 5,886, 6,028, 6,250, 6,564, 6,968, 7,452, 7,994, 8,568, 9,140, 9,680 and 10,161 respectively. The total number of probable daily new cases would be increased in the future as predicted by Periodic regression model (Fig. 2.).

We then compared two models (ARIMA & PR) to check the pattern of disease prediction for COVID-19 outbreak in India, shown in Fig. 3. The comparison of these two models shows that PR and ARIMA model intercept at some point of time, PR model enters into geometric progression and displays increase in number of new cases during early August, 2020. But ARIMA graph display exponential increase during early stage of prediction and then the graph fattens gradually showing decrease in number of cases, during mid-August, 2020. The ARIMA model (2,0,2) absorbs the short-term fluctuations, which was fine-tuned to eliminate all residual autocorrelation, while the Periodic regression attempted to capture the cyclic behaviour of COVID-19 incidence.

Our current situation, demands predication of disease pattern and trend to take appropriate healthcare interventions for the future. The analysis predicts very alarming outcomes, which defines to worsen the conditions in India, especially Maharashtra, Tamil Nadu, Gujarat, Delhi, Rajasthan, MP and UP states. The pandemic has also exacerbated and brought to the forefront the systemic and deeply entrenched economic and societal inequalities that are among the root causes of human trafficking. Based on the predictions resulted our study, public health officials shall tailor their aggressive interventions to grasp the power exponential cases, and introduce rapid infection control measures at hospital levels as well as at community level to curtail the COVID-19 pandemic and was similar to study analysed at global level and extracted data upon Machine Learning approach using Artificial intelligence techniques for top 10 countries [2].

The COVID-19 outbreaks globally present a significant challenge for modellers, with the available resources, it has become challenge for government and civil societies to mitigate the impact of health crisis. In our study, we adopted two models namely ARIMA and Periodic regression models to predict the real-time predictions of daily new cases in India. In a study by Chakraborty and Ghosh, two alarmingly important models relevant to ongoing COVID-19 pandemic were considered. They proposed a hybrid ARIMA-WBF model that can explain the nonlinear and nonstationary behaviour present in the univariate time series datasets of COVID-19 cases [14].

We also made a prediction of daily new cases and the probable size of the disease outbreak for the next three months using two best fitted models. According to the prediction models the number of daily new cases may reach 6,385 to 9,292 by the July end, suggesting more people would be infected in future. Fortunately, the India has taken strict measures like enforced quarantine, lockdown, curfews,

travel restrictions etc., to control the spread of infection at the early stage of pandemic, but due to relaxation of lockdown may trigger increase in transmission and spread. Improper social distancing, air travel, travel using public transport, travel of migrants from different states and overseas, social and religious congregations may impact adversely on spread of disease all over the country within a very short span of time. Quarantine of the suspected cases must be imposed to control spread of disease. If these are not followed, it would lead to natural process of disease transmission and the number of infected cases might be doubled than actual in the future days. Social distancing and lockdown practised in different countries have failed to curtail its spread due to violations and misunderstanding, varied perceptions among people on the ways it should be practiced [15, 16, 17]. Our data-driven analytics mainly aims the importance of the transmission of disease in the population at this stage and results of prediction helps the policy makers and planners to improve the risk management and risk governance by prioritising the risk management efforts.

Infectiousness of COVID-19 pandemic in the world reported 5,601,871 infections with 6.5% deaths in the world population. But India reported 145,456 infections with 2.9% deaths almost less than other country of the world [18]. In the early stages of disease outbreak, took almost 59 days to reach 10,000 infections, but in later days it took 7 to 8 days to add 10,000 new cases. From this study, the current trend shows 10,000 new cases in 2 days. If same condition persists, more new cases are predicted to be added as forecasted by different models. The spread may occur all over the country within very short span of time.

Time series models, like Autoregressive Integrated Moving Average (ARIMA), effectively consider serial linear autocorrelation among observed incidence. Whereas, periodic regression satisfactorily describes the non-linear cyclic trend. However, have some limitations like no explicit seasonal indices, hard to interpret coefficients or explain "how the model works", and it does not support any volatility and any in between changes in the prediction periods.[2] Thus, the periodic regression and ARIMA analysis of COVID-19 diseases will help in knowing the trends, prediction of future outbreaks and assist in planning the preventive measures, allocation of scarce resources effectively. Good management practices like stringent biosecurity measures, strict sanitation and hygiene practices, isolation and quarantine of diseased individuals, and trade restrictions are necessary for successful operation of control programmes.

We have used only time series data for confirmed and death cases in this study. It was observed that the initial death cases due to COVID-19 infection was reported primarily in elderly population, this may be due their age factor and weak immune system which allows rapid viral infection [19, 20]. The forecast analysis of COVID-19 dynamics showed a different angle for India, and it looks scarier than imagined. Interestingly, with more positive recovery the situation may remain stable. Thus, it can slow down the surge of COVID-19 pandemic during the proceeding months depends upon various administrative intervention and public awareness about the COVID-19 pandemic spreading.

The current trend shows that there will be an arithmetic progression in the next few days if stringent control measures are taken by the government and there are less likely chances of sliding into the geometric progression in the long run as per results of periodic regression. Hospital and medical facility extension and enhancement work should be continued at a very rapid pace to prepare the country for exponential growth, if it occurs. However, with current interventions and preparations, the Government of India is looking forward to flatten the epidemic curve.

5. Conclusion

Our study shows the results of predictive number of COVID-19 cases using both ARIMA model and Periodic regression model and shows that current trend of increase is linear for the next few months. Also, the death rate (2.9%) is comparatively lower with respect to number of positive cases, which is very promising. There is a likely chance of exponential increase in the cases, due to relaxation of lockdown, as population starts mixing, increased social and religious congregations. However, during 50 plus days of lockdown period, India has given a buffer time for the hospitals and healthcare organisations to prepare themselves with appropriate preparedness if the pandemic explodes in future. Both, ARIMA and Periodic regression models are best fitted to the observed data, the slight difference in future prediction may arises due to intrinsic characteristics of each models and parametrization, the ARIMA models are best suited for short term prediction, while the Periodic regression can be used to long term prediction with more non-linear cyclic trend is present in daily incidence. Use of covariates such as demography, environmental geographic layout of the country, dynamics of population and governance, the model prediction rate can be further improved. Our findings are more helpful for policy makers and planners to improve the risk management and risk governance by prioritising the risk management efforts.

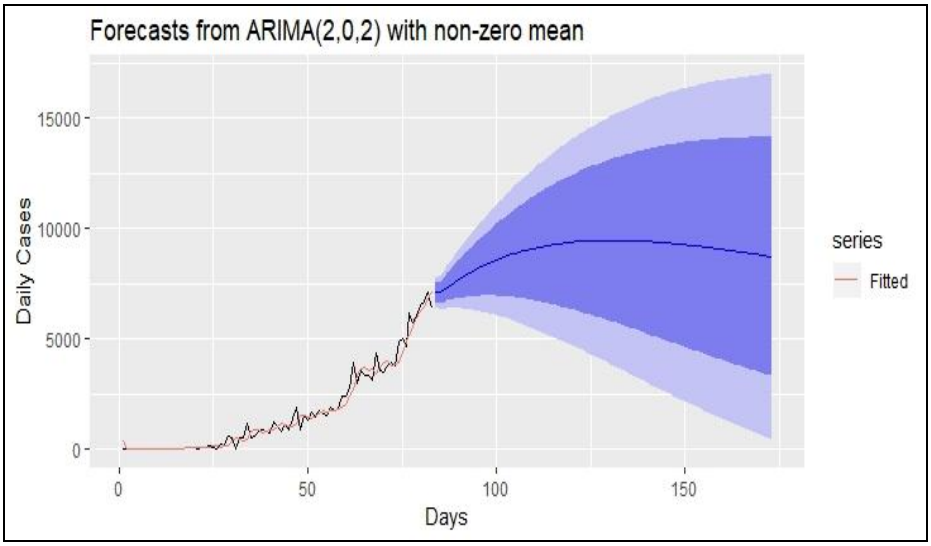

Fig 1: Prediction of COVID-19 outbreak using Auto-Regressive Integrated Moving Average (ARIMA) Model in India for next ninety days.

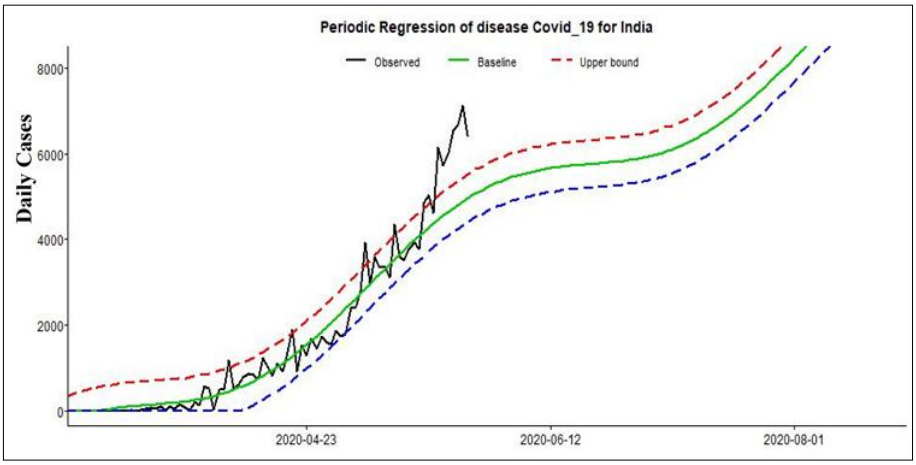

Fig 2: Prediction of COVID-19 outbreak using Periodic regression in India for next ninety days. Green line shows the baseline values and red dotted lines shows the upper/threshold level and blue dotted line shows lower range with 95% CI.

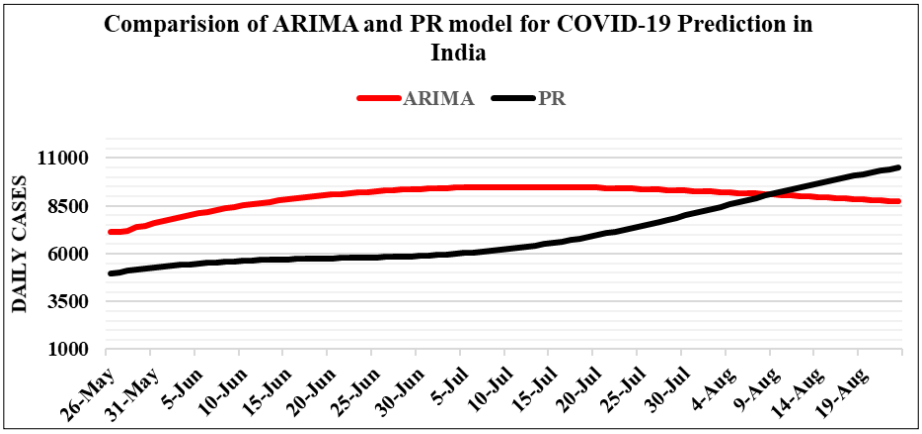

Fig 3: Comparison of ARIMA (Black) & Periodic regression (Red) models for COVID-19 prediction in India.

6. Acknowledgment
None.

7. References
1. World Health Organization. Report of the WHO–world Joint Mission on Coronavirus Disease (COVID-19), 2019.
2. Singh RK, Rani M, Bhagavathula AS, *et al*. Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model. JMIR Pub Health Surv. 2020; 6(2): e19115.
3. Data on number of Covid-19 cases reported in India. Available from: https://www.covid19india.org/.
4. Ping Y. Distribution theory, stochastic processes and infectious disease modelling. In: Fred B, Pauline van den D, Wu J, editors. Mathematical epidemiology. Springer, Berlin: Heidelberg, 2008, p. 229-93.

5. Krishnamoorthy P, Kurli R, Patil SS, Roy P, Suresh KP. Trends and future prediction of livestock diseases outbreaks by periodic regression analysis. Indian J Anim Sci. 2019; 89(4):369-76.
6. Ye QH, Qin LX, Forgues M, *et al*. Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med. 2003; 9(4):416-23.
7. Mai MV, Krauthammer M. Controlling testing volume for respiratory viruses using machine learning and text mining. In AMIA Annual Symposium Proceedings (Vol. 2016, p. 1910). American Medical Informatics Association, 2016.
8. Purcaro G, Rees CA, Wieland-Alter WF, *et al*. Volatile fingerprinting of human respiratory viruses from cell culture. J Breath Res. 2018; 12(2):026015.
9. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influ Other Resp Vir 2. 2014; 8(3):309-16.
10. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief, 2020, 105340.
11. Bliss CI. Statistics in Biology. McGraw-Hill Book Company, New York, USA, 1970.
12. Wang S, Feng J, Liu G. Application of seasonal time series model in the precipitation forecast. Math Comp Model. 2013; 58(3-4):677-83.
13. Čobanović K, Lozanov-Crvenković Z, Nikolić-Đorić E. 2006. Periodic Regression.
14. Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos Soliton Fract, 2020, 109850.
15. Singh BR, Gandharva R. Are BCG Vaccination, Population Density, Median Age and Poverty Important Determinants of COVID-19 Pandemic Spread, Morbidity and Mortality? DOI: 10.13140/RG.2.2.21116.49282
16. Maharaj S, Kleczkowski A. Controlling epidemic spread by social distancing: Do it well or not at all. BMC Pub Health, 2012, 12:679.
17. Fast SM, González MC, Markuzon N. Cost-effective control of infectious disease outbreaks accounting for societal reaction. PLoS ONE, 2015; 10:e0136059.
18. World meter. Coronavirus (COVID-19) mortality rate. Last updated May 25, 2020. https://www.worldometers.info/coronavirus/coronavirusdeath-rate/#who-03-03-20.
19. Fine PE. The interval between successive cases of an infectious disease. Am J Epi, 2003; 158:1039-47.
20. Nishiura H, Kobayashi T, Yang Y, *et al*. The rate of under ascertainment of novel coronavirus (2019-nCoV) infection: Estimation using Japanese passenger's data on evacuation flights. J Clin Med, 2020, 9:E419.