Piece-wise linear regression: A new approach to predict COVID-19 spreading

Apurbalal Senapati¹, Soumen Maji² and Arunendu Mondal³

¹ Department of Computer Science & Engineering, Central Institute of Technology Kokrajhar, Assam-783370, India.
² Department of Civil Engineering, Central Institute of Technology Kokrajhar, Assam-783370, India.
³ Department of Chemistry, Central Institute of Technology Kokrajhar, Assam-783370, India.

Email: a.senapati@cit.ac.in

Abstract. The coronavirus disease 2019 (COVID-19) pandemic is the most rapidly evolving global emergency since March 2020 and one of the most exercised topics in all aspects of the world. So far there are numerous articles that have been published related to COVID-19 in various disciplines of science and social context. Since from the very beginning, researchers have been trying to address some fundamental questions like how long it will sustain when it will reach the peak point of spreading, what will be the population of infections, cure, or death in the future. To address such issues researchers have been used several mathematical models from the very beginning around the world. The goal of such predictions is to take strategic control of the disease. In most of the cases, the predictions have deviated from the real data. In this paper, a mathematical model has been used which is not explored earlier in the COVID-19 predictions. The contribution of the work is to present a variant of the linear regression model is the piece-wise linear regression, which performs relatively better compared to the other existing models. In our study, the COVID-19 data set of several states of India has been used.

Keywords: Piece-wise linear regression, COVID-19, Mathematical model, Prediction, Machine learning

1. Introduction

The COVID-19 pandemic gradually deteriorated day by day since the beginning of the year 2020 around the world. Several countries like Vietnam, New Zealand, South Korea, Cuba, Germany, France, etc. survived successfully and the spread of the disease is declining now. However, on the other hand, in countries like the USA, India, Brazil, Russia, etc. the disease is spreading very fast. At present in India, the daily positive cases are identified as more than 80,000 on average. Almost around the world imposed the lockdown as a common strategy to control the spreading of the disease up to some extent. The different countries change their lockdown strategies from time to time for their socio-economical situations and that also affects the spreading of a pandemic. Fig. 1 shows the trend of spreading positive COVID-19 cases in some of the selected states in India whereas Fig. 2. shows a comparison on the daily averaged Covid-19 positive cases during lockdown and unlock-1.
Fig. 1: Trend of spreading positive COVID-19 cases in Gujarat, Maharashtra, Tamil Nadu, Delhi, and Uttar Pradesh from 24th March to 30th June 2020.

Fig. 2. Daily averaged Covid-19 positive cases shown during lockdown and unlock-1 in most affected states like Maharashtra, Gujarat, Tamil Nadu, Delhi, and Uttar Pradesh.

2. Existing predicted models:
   History shows the outbreaks of infectious diseases in several civilizations is quite past [1], [2]. Some standard mathematical models were established for infectious disease [3], [4]. A recent one was Middle East Respiratory Syndrome (MERS) (2012), avian influenza (H7N9), and Ebola [5]. Bernoulli et al. [6] used a mathematical model to analyze the mortality of smallpox in England and later on different types [7] of models come. Since the last few months from a computational point of view, several data-driven
statistical models have been adopted to predict the infection, recovery, death, time interval of the pandemic, etc. of the COVID-19. So many models like the non-linear regression framework are used to address the cumulative and daily death rate [8]. A data-driven technique is also used where the mobile GPS tracker is used to find out the social distancing behavior [9]. Yang et al. [10] used an AI-based SEIR model to predict the trend of pandemic COVID-19. Apart from these, many researchers across the globe have used different statistical models to forecast COVID-19 related outcomes and their interpretations [11], [12] and [13]. Some of these predictions are good to some extent but most of them are far from reality [7], [8] and [14]. Although attempts have been made [15] to improve the existing model with the recent availability of the updated data and external policies on the predictions as well [15], [16] and [17]. Interestingly, Luo [18] used the heuristic approach to address the problem and discussed the challenges of such kind of models, and explained their complexity. These reports do not conclude that all such models or analyses are reliable, rather these should be considered as predictions based on the available data, and these predictions are highly dependent on the reliability of these data. So the accuracy of the models depends on the accuracy of the collected data although there will be an error and which will not be the weakness of the models, but a part of the model. Since there are several reports already predicted based on the data available in the past, hence now it is the time to verify their predictions. Few such results are discussed here.

Fig. 3. Prediction of Covid-19 pandemic in India (Luo 2020)

Luo [18] has used the "SIR epidemic model" using the data from different countries to predict the key dates of transition during the coronavirus lifecycle across the globe. For India, they have predicted that 97% of the pandemic will be ended by the 3rd week of May 2020 and at the end of May 2020, it will over 99% with their declaration that, it may contain errors. Fig.3 shows their predictions with data as of dated 06-May-2020.

Xu [19] tried to model the COVID-19 pandemic by the Farr Law. It says that the epidemics like AIDS, SARS, Ebola, etc. tend to rise and fall in a roughly symmetrical pattern or bell-shaped curve. But, the practical scenario does not follow exactly the symmetric pattern rather it is skewed on one side in the case of COVID-19.

On the other hand, Ben-Israel [20] concluded, based on simple statistical analysis that, the spread of COVID-19 follows a particular pattern of repetition in a curve. He also says that the growth and decline in the new cases follow the same trend across countries around the world. Particularly, he stressed that
it reaches a peak after about 40 days and flattened to almost zero after 70 days and it is independent of the other external factors. But, the actual scenario is far from this conclusion.

Some researchers [21], [22] attempted to predict the impact of lockdown in the spreading of COVID-19 in India and did similar studies for the other countries. Schueller et. al. [22] considered the first 21 days of the lockdown in March 2020 and also considered several parameters like incubation period, the infected population has no symptoms or mild symptoms, the age distribution of the populations, etc. They predicted, with moderate and hard lockdown may reduce the transmission of the disease. But this prediction is difficult to verify. Because based on the clinical expert it is obvious that the lockdown reduces the spreading of infection but the actual measure is quite difficult. Singh [21] tried to understand the spreading characteristic of COVID-19 in India taking into account the physical distancing and lockdown. They have used differential based mathematical modeling. Singh [21] used the modified logistic growth model in his prediction using the truncated information on novel corona confirmed cases in India. There are several conditions applied in his model and predicted that at the end of April 2020 increasing growth will start to decline and there will be no newer cases at the end of July 2020. Whereas a similar study of Mandal et al. [23] has shown that, the impact of social distancing and that can reduce the maximum cases up to 62 percent.

Chatterjee et al. [24] predicted the COVID-19 epidemic in India based on a stochastic model. Their study shows that uninterrupted epidemic in India could result a minimum of 364 million positive cases with the death of 1.56 million and the highest acceleration of the pandemic by mid-July. To verify the prediction yet has to observe.

Ghosh et al. [25] performed the state-wise analysis in India. They built three growth models (exponential model, the logistic model, and the Susceptible Infectious Susceptible (SIS) model) to predict infected people in the states in the next coming 30 days. They used the parameter daily infection-rate (DIR) values for each state. This model furnished the three predictions for the states, they have interpreted the results combining from the models rather than individually. They claimed that over 126 thousand people will be affected in COVID-19 by May 31, 2020. They also observed people are not following the lockdown because of the essential work in several places and results spreading the infection.

Menon et al. [26] described several models in their ongoing projects. They have incorporated features like asymptomatic, symptomatic infections, the effect of migration, quarantine, etc. and have also addressed several subjective type questions and the predictions of the number of deaths, the number of infections, etc.

Senapati et al. [27] tried to overcome the limitations of the linear regression model using the piecewise regression model. They have done their experiment from the data of some selected states of India. The results are compared with the other regression model and show their results bid others.

3. **Linear regression model and its limitation**

Linear Regression is a supervised machine learning algorithm used for a regression task. The regression model is used for prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. There are variations of regression models, they differ based on – the kind of relationship between the dependent and independent variables, they also considered the number of independent variables.

The linear regression performs the task to predict a dependent variable (y, considered as an output) based on a given independent variable (x, considered as an input). So, the regression technique finds out a linear relationship between input and output. Geometrically this is interpreted as a best fitted straight line. Mathematically it can be represented as in equation (1)

\[ y = mx + c \]  

(1)
Where the unknown values \( m \) and \( c \) are to be estimated by the gradient descent technique from the given data i.e. from the training data is given as in equation (2)

\[
\{(x_i, y_i) : i = 1, 2, ..., n\}
\]  

(2)

### 3.1 Limitation of the model

Since the linear regression model is based on the principle of the relationship of the input and the output data which is linear but in real-life data, the linear relationship does not hold. Sometimes only partitions of the data hold the linearity (Fig. 4) in that case, a single line is not enough to fit the whole data. On the other hand, the outlier data makes huge effects on the regression.

### 3.2 Proposed model

Since we are dealing with the COVID-19 data and based on our observation is that the COVID-19 data does not follow the linear property. Rather it follows the linear property for a short period and then it changes its direction. In this case, it is not well suited for the linear regression and if still it is used the linear regression the predictions are far from the actual. To overcome this problem several researchers, use polynomial regressions (Gupta et al. 2020). In our study, the piecewise regression model [28] has been used.

From the COVID-19 data, it has been observed that it follows the linearity within a small interval (window). Then it shifts its directions, again follows the linearity, again shifts its trends, and so on. Under this condition, one line simply is not enough to fit the entire data set, then the concept of piecewise linear regression comes to overcome such limitation. When the data set follows different linear trends over the different partitions of data, then we should model the regression function in pieces. Each linear regression is corresponding to a partition i.e., the pieces, and whether the pieces are connected or not connected depends, on the data and the problem. The connecting points are known as the breakpoints, i.e., the locations where the slope changes. Fig 4 shows the actual data, and Fig 5 shows the piecewise correlation regression lines.

The mathematical representation of the piecewise linear regression, consider the lines in Fig. 5 represented by equation (3)

\[
y = \begin{cases} 
  m_1 x + c_1 & \text{when, } x \leq p \\
  m_2 x + c_2 & \text{when, } x > p 
\end{cases}
\]  

(3)
The point at $x = p$ is the joining point of two lines, i.e., a breakpoint. Find out the breakpoints is also a challenging problem. In our case, we have chosen heuristically, and it is considered a seventh length window, i.e., increasing rate follows the linearity for seven days, and then it changes.

4. System evaluation

The system is evaluated in two different aspects. First, its output is compared with the actual output and it is also compared with the other existing systems.

To check the performance of the piece-wise regression model past COVID-19 positive confirm cases have been used. The system generated data have been compared with the actual data. Particularly in this evaluation the COVID-19 positive data (29) of the state Maharashtra, the most spreading state in India have been used from 1st, 6th, 12th, 18th, 24th, and 30th August 2020. Table 1 shows the comparison. The column Actual is the actual positive confirm cases and the column Predicted represents prediction by the piece-wise regression model.

| Date       | Actual | Predicted |
|------------|--------|-----------|
| 01-08-2020 | 42,2118| 41,9238   |
| 06-08-2020 | 46,8265| 46,7689   |
| 12-08-2020 | 53,5601| 53,6944   |
| 18-08-2020 | 59,5865| 60,8557   |
| 24-08-2020 | 67,1942| 68,4708   |
| 30-08-2020 | 74,7995| 75,9658   |

Now the system is compared with the other systems. The comparison is done by the SEIR model and a Polynomial regression model [30]. The comparison is only meaningful when it is considering the same data for the same dates. Hence, we have considered the same data set for the same date. Data have been used in the positive confirm cases from 17th May 2020 to 25th May 2020. Fig 6. shows the graphical representation of the comparison. The result (Fig 6) shows that our system (piece-wise linear regression) is very close to the actual value compared to the other existing system.
5. Conclusions
In this study, a piece-wise linear regression model has been used for the COVID-19 predictions which are not done earlier. A comparative study shows that it performs better than the other existing model and hence can be used for future prediction in better accuracy. From the mathematical point of view, the model performs better in local data i.e. in a small partition. Hence there needs further investigation about the optimal partition to improve further.

6. References
[1] Siro Igino, T 2007 The biblical plague of the Philistines now has a name, tularemia. Med Hypotheses. 69:1144e1146. https://doi.org/10.1016/j.mehy.2007.02.036.
[2] Acuna-Soto R, Stahle DW, Therrell MD, Gomez Chavez S, Cleaveland MK 2005 Drought, epidemic disease, and the fall of classic period cultures in Mesoamerica (AD 750-950). Hemorrhagic fevers as a cause of massive population loss. Med Hypotheses. 65(2):405-9. doi: 10.1016/j.mehy.2005.02.025. PMID: 15922121.
[3] William Ogilvy, K., McKendrick A. G. (1927). A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A115700–721 http://doi.org/10.1098/rspla.1927.0118.
[4] Anderson RM, May RM 1992 Infectious Diseases of Humans: Dynamics and Control. Oxford, New York: Oxford University Press; 1992.
[5] Emergencies preparedness, response (WHO), Disease outbreaks by year 2020 Available at : https://www.who.int/csr/don/archive/year/en/ (Accessed April 23, 2020)
[6] Bernoulli D, Blower S 2004 An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. Rev Med Virol. 14:275e288. https://doi.org/10.1002/rmv.443.
[7] Siettos CI, Russo L 2013 Mathematical modeling of infectious disease dynamics. Virulence. 4:295e306. https://doi.org/10.4161/viru.24041.
[8] Christopher JL M 2020 IHME COVID-19 health service utilization forecasting team. Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries medRxiv 2020.04.21.20074732; doi: https://doi.org/10.1101/2020.04.21.20074732
[9] Woody S, Tec MG, Dahan M., Gaither K, Lachmann M., Fox S, Meyers LA, G Scott J 2020 Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones medRxiv 2020.04.16.20068163; doi: https://doi.org/10.1101/2020.04.16.20068163
[10] Yang Z, Zeng Z, Wang K, Wong S S, Liang W, Zanin M, Liu P, Cao X, Gao Z, Mai Z, Liang J, Liu X, Li S, Li Y, Ye F, Guan W, Yang Y, Li F, Luo S, Xie Y, … He J 2020 Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of thoracic disease 12(3), 165–174. https://doi.org/10.21037/jtd.2020.02.64
[11] COVID-19 Daily Epidemic Forecasting by the Institute of Global Health 2020 https://renku-lab.shinyapps.io/COVID-19-Epidemic-Forecasting (Accessed on 25 May 2020).
[12] COVID-19 Confirmed and Forecasted Case Data 2020 https://covid-19.bsvgateway.org/. (Accessed on May 25, 2020).
[13] COVID-19 Projections 2020 https://covid19.healthdata.org/projections. (Accessed on May 25, 2020).
[14] Jewell NP, Lewnard JA, Jewell BL 2020 Caution Warranted: Using the Institute for Health Metrics and Evaluation Model for Predicting the Course of the COVID-19 Pandemic. Ann Intern Med. DOI: https://doi.org/10.7326/M20-1565.
[15] Marchant R, Samia NI, Rosen O, Tanner MA. Cripps S 2020 Learning as We Go: An Examination of the Statistical Accuracy of COVID19 Daily Death Count Predictions. medRxiv 2020.04.11; doi: https://doi.org/10.1101/2020.04.11.20062257
[16] Grey S, Mac Askill A 2020 Special Report: Johnson listened to his scientists about coronavirus - but they were slow to sound the alarm. Reuters. A. https://reut.rs/2Rk3sa7. (Accessed on May 25, 2020).
[17] Resnick B 2020 The White House projects 100,000 to 200,000 Covid-19 deaths. Vox. https://www.vox.com/science-and-health/2020/3/31/21202188/us-deaths-coronavirus-trump-white-house-presser-modeling-100000. (Accessed on May 25, 2020).

[18] Luo J 2020 Predictive Monitoring of COVID-19. SUTD Data-Driven Innovation Lab. https://ddi.sutd.edu.sg/when-will-covid-19-end/.

[19] Xu J, Cheng Y, Yuan X, Li W V, Zhang L 2020 Trends and prediction in daily incidence of novel coronavirus infection in China, Hubei Province and Wuhan City: an application of Farr’s law. American Journal of Translational Research, 12(4), 1355.

[20] Ben-Israel I 2020 The end of exponential growth: The decline in the spread of coronavirus. The Times of Israel. https://www.timesofisrael.com/the-end-of-exponential-growth-the-decline-in-the-spread-of-coronavirus/. (Accessed on May 25, 2020).

[21] Singh B P 2020 Forecasting Novel Corona Positive Cases in India using Truncated Information: A Mathematical Approach. medRxiv. Doi: https://doi.org/10.1101/2020.04.29.20085175.

[22] Schueller E, Klein E, Tseng K, Kapoor G, Joshi J, Sriram A, Laxminarayan R 2020 COVID-19 in India: Potential Impact of the Lockdown and Other Longer Term Policies. Retrieved from Washington D.C. https://cddep.org/publications/covid-19-india-potentialimpact-of-the-lockdown-and-other-longer-term-policies/.

[23] Mandal S, Bhatnagar , Arinaminpathy N, Agarwal A, Chowdhury A, Murhekar M, Sarkar S 2020 Prudent public health intervention strategies to control the coronavirus disease 2019 transmission in India: A mathematical model-based approach. Indian Journal of Medical Research, 151(2), 190.

[24] Chatterjee K, Chatterjee Ko, Kumar A, Shankar S 2020 Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model, Medical Journal Armed Forces India. Doi: https://doi.org/10.1016/j.mjafi.2020.03.022.

[25] Ghosh P, Ghosh R, Chakraborty B 2020 COVID-19 in India: State-wise Analysis and Prediction. medRxiv preprint. Doi: https://doi.org/10.1101/2020.04.24.2007792.

[26] Menon A, Rajendran N K, Chandrachud A, Setlur G 2020 Modelling and simulation of COVID-19 propagation in a large population with specific reference to India. medRxiv preprint. Doi: https://doi.org/10.1101/2020.04.30.20086306.

[27] Senapati A, Nag A, Mondal A, Maji S 2020 A novel framework for COVID-19 case prediction through piecewise regression in India. International Journal of Information Technology. Doi: https://doi.org/10.1007/s41870-020-00552-3

[28] McZgee V E, Carleton W T 1970 Piecewise regression. Journal of the American Statistical Association, 65(331), 1109-1124.

[29] SRK Novel Corona Virus Disease 2019 in India 2020 covid_19_india.csv, Version 205, https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_19_india.csv. (Accessed on Sept 28, 2020).

[30] Gupta R, Pandey G, Chaudhary P 2020 Machine Learning Models for Government to Predict COVID-19 Outbreak. Doi: https://doi.org/10.1145/3411761