Analysis of COVID-19 Cases in India using SEIR, ARIMA and LSTM Models

Dr. Souvik Sengupta
Department of Computer Science and Engineering, Aliah University

Abstract

After one year from the start of the COVID-19 pandemic in India, the country is now having a steady decay in the number of daily new cases and active cases. Although the vaccination process is about to start from mid of January 2021, it would not affect the number of daily cases at least for the next three to four months for obvious reasons like phase-wise implementation and six to eight weeks time span required from the first dosage to develop the immunity. Therefore, the prime question is now, where would we reach at the end of the first quarter of 2021, and what could be the number of new cases and active cases before the vaccination immunity starts working. This paper analyzes the growth and decay pattern of Indian COVID-19 cases with help of SEIR epidemical modeling, ARIMA statistical modeling, and time series analysis by LSTM. The models learn the parameter and hyper-parameter values that are best suited for describing the pattern for the COVID-19 pandemic in India. Then it tries to predict the numbers for India by the end of March, 2021. It is forecasted that the number of new cases would come down near 5000 per day, active cases near 40,000 and the total number of infected may reach 11.1 million if the current pattern is followed.

Keywords: **COVID-19 prediction, Time series analysis, LSTM, Epidemic model, SEIR, ARIMA**

INTRODUCTION

Forecasting the number of possible infected people in the next few months can be useful in assisting the policymakers in designing better strategies for vaccination and in taking productive measures. At the initial stages of the pandemic, most of the measures were taken based on assumptions and previous experience on the spread of other contagious diseases. Now, since the vaccination is about to start, it is very important to understand that to what extent we have managed to bring down the number of daily cases without the vaccination. There could be multiple influencing factors like public awareness, lockdowns, government initiatives, weakening of the virus strain, and even hard immunity. However, the forecast for the next three months is very important in the strategic planning of the vaccination procedure and also in contemplating further relaxations in social life.

The first COVID-19 case in India was reported on January 27th, 2020. By the end of March 2020, there were 2545 total cases, with the number of the daily new case being 425. India stopped its International flights from 10th March and imposed a complete lockdown all over the country from 25th March 2020. Since then the country has followed 4 consecutive lockdowns between 25th March and 7th July 2020. This lockdown imposed a ban on public gatherings, public commute, multiplex and cinema halls, shopping malls, bars and restaurants, and shops and markets other than essential commodities. However, according to media reports, the strictness of the lockdown has been between low to moderate. While lockdown was observed more rigidly in urban areas, rural areas have been more reluctant. The number of new cases has increased exponentially from the start of April to the mid of September 2020 and is decreasing since then. It is obvious from the case history of all such viral pandemics that it takes a longer time in diminishing phase than the growing phrase.

The objective of this paper is to analyze the pattern that has been followed in the rise and decay of the COVID-19 daily new and active cases in India between 1st April to 31st Dec 2020, and then forecasting the probable number of daily new and active cases by the end of the first quarter of 2021, i.e. 31st March. The proposed architecture of this work employs three components: one for statistical analysis of the time-series data of daily new cases using Auto Regressive Integrated Moving Average (ARIMA) model, second for estimating the unconstrained growth and decay of active infected cases based on Susceptible Exposed Infectious Recovered (SEIR) epidemical model, and the third one for learning the time-series pattern of active cases using Long Short Term Memory (LSTM) model. The models are then used for predicting the future cases in India from 1st January to 31st March 2021. The rest of the paper is divided as followed: section II describes the review works, section III describes the overall methodology, section IV analyzes the results, and section V concludes this work.

REVIEW WORKS

More than four hundred thousand research papers/articles have been published in 2020 since COVID-19 has spread worldwide [13]. Scientific researchers from all disciplines are trying their best to contribute in their own way. Machine Learning and data science have been used extensively to understand, analyze and recognize the pattern of the spread of the disease, to identify any latent pattern and key factors in the spreading of the disease, and to predict the future growth and decay of the pandemic. We review some of the recent works on analysis and prediction of COVID-19 cases mostly related to India.

Singh et al. [1] forecasted the COVID-19 epidemic in India with mitigated social distancing. This work introduces a mathematical model of the infection spread in a population considering social contact between ages. It uses the social contact structure as described by Prem et al. [11] and then the impacts of social distancing measures like workplace non-attendance, school closure, and lockdown are investigated. The authors suggested that sustained periods of lockdown with periodic relaxation will reduce the number of cases to a satisfactory level. Gupta et. al[2] investigated the importance of lockdown in six social components i) restaurants and cafes ii) Grocery markets and food shops iii) community parks and gardens iv) public transports v) private and government offices and vi) residential places. This work uses exponential and polynomial regression for predicting the number of future infections and claims a significant drop could be possible with considering the first five of these categories. Mortality predictions have been done through binary classification using a decision tree model and recorded an accuracy of 60%. Gupta et al. [19] used the SIRD compartment model for investigating the progress and prediction of COVID-19 in India. Behavioral changes due to changes in key parameters because of lockdown, social distancing, and other non-pharmaceutical interventions are considered, and dynamic behavior of the parameters and R0 has been considered. The model forecasted that after 350 days from the onset of the pandemic, more than 273,586 people would be infected with a total number of infected people more than 10.7 million. Das [6] used an extension of the SIR model known as Susceptible- Infectious-Quarantined-Recovered (SIQR) along with a statistical machine learning (SML) model for the prediction of the infected population. This work combines two approaches, one to understand the severity of the ground situation and the second for the prediction. With a polynomial regression model, this paper predicted around a total of 66,224 cases by May 01, 2020, in India. In a similar work, Pandey et al. [7] used SEIR (Susceptible, Exposed, Infectious, Recovered) model with polynomial regression to predict the number of cases in the next two weeks in India based on data available up to 30th March 2020. The model predicted around six thousand cases within mid of April 2020.

For time series modeling of COVID-19 data, the ARIMA model has been the most popular among the researchers. Tandon et. al [3] proposed ARIMA based model for the prediction of future cases. Autocorrelation function (ACF) graph and partial autocorrelation (PACF) graphs are used to determine the initial parameters. These models are then used to test for variance in normality and stationary of the time-series data. ARIMA (2,2,2) appeared to be the best fit model with respect to scores of Mean absolute percentage error (MAPE), Mean absolute deviation (MAD), and Mean squared deviation (MSD). The model forecasted that the infection cases are expected to greatly rise in mid of May and may start to decline after that. Chakraborty et al. [4] proposed a two-folded approach- first for generating real- time forecasts of the future COVID-19 cases in multiple countries and second for predicting fatality rate by considering different demographic and disease characteristics. It uses a hybrid model of ARIMA and a wavelet-based forecasting model. It also employs a decision tree regression model for predicting the risk associated with the fatality rate. In another work, Deb et al.[5] proposed a time series model to analyze the trend and pattern of the COVID-19 outbreak. The authors claim that a time-dependent quadratic trend successfully captures the incidence pattern of the disease. This work uses the ARIMA model to identify if the trend changes after any point. This work estimated the average contagious rate in India to be 1.42

Tomar et. al [10] have given data-driven estimation using long short-term memory (LSTM) and predicted the number of COVID-19 cases in India. It also analyzed the effect of preventive measures like social isolation and lockdown on the spread of the disease. The training data for this work contains case details up to 4th Apr 2020, and the prediction is made for the next 90 days. In another work, Tiwari [8] proposed a model for the prediction of infected number based on simple Ordinary Differential Equation (ODE). This model predicted that infection in India would hit the peak with daily 22 thousand active cases during the last week of April followed by a decline in active cases. In [9] Tiwari et al. used the pattern of china with help of ML models to predict the COVID-19 outbreak in India. The predictive model is built using WEKA to predict the day-wise number of confirmed cases, recovered cases, and death cases. The model is trained with data from China and predicts the case of India, assuming that the trend of the pandemic would be the same in both countries except for a time lag. However, almost all of the above the mentioned works failed to predict the outburst of the disease that India observed. Acknowledging the famous "All models are wrong, but some are useful" theory, we try to find out the thriving and failing parts of the above discussed models. It is found that models that relied on the data of lockdown and social distancing (in whatever form that is available) the prediction failed more than the models that considered only infected, recovered, and death records. It is mainly due to the unreliability of the available lockdown and social distancing data, and also small changes in its number have a large reciprocate effect on the predictions.

METHODOLOGY

Figure1 represents the architecture of the proposed methodology. The dataset is collected from the WHO Corona Virus Disease (COVID-19) website [14]. It contains country- wise information about number of new cases, cumulative cases, new deaths, and cumulative deaths on each day. We selected a time period of 1st April to 31st December 2020 for this study. We explore three different models - SEIR model, ARIMA model, and LSTM model, independently on this dataset for identifying the pattern of daily new and active cases in India. The SEIR compartmental model gives an idea on the spread of the disease in an unconstraint environment. The ARIMA model analyzes the rise and fall of the daily new cases purely from statistical perspective. Finally, a stacked LSTM model is trained with the time-series data with a time-stamp of 7 days, to match with the actual data. Then the trained LSTM model is used to make forecasting on the number of future cases in India in next three months. For each of the cases the model learns the parameters those are best suited to match the actual data.

A. SEIR Epidemiological model

This work uses the classical Susceptible Exposed Infected Recovered (SEIR) model, which is a compartmental model ideal for modeling epidemic cases like COVID-19 [15]. It splits the population into four compartments S, E, I, and R and defines the transition rates among them. People who are at the equal risk of getting infected are considered susceptible (S). Susceptible individuals are who come in contact with the virus and become exposed (E) but not yet infectious. Infectious people (I) is the fraction of individuals who are capable of transmitting the disease. Recovered (R) are recovered or deceased individuals. This model makes some basic assumptions: i) the total number of population (N) remains constant during the time period $N= S+E+I+R$ ii) recovered people do not become susceptible again iii) vaccination has not started to effect the spread, so everybody is equally susceptible, and iv) Quarantine or isolation of infected people is not effective (or failed badly).

The rate of change in each category is as follows:
$dS/dt = - \beta.SI$
$dE/dt = \beta.SI - \sigma E$ $dI/dt = \sigma E - \gamma I$ $dR/dt = \gamma I$
Where, β= rate of contact and γ is the rate of recovery which is equal to (1/ number of days required to recover for an infected person). The rate of change of susceptible people (dS/dt) is always negative as the number of susceptible people reduces with more people getting infected. If S0 is the number of initial susceptible population, then at anytime (t) the number of susceptible (St) is always less than S0. The rate dS/dt is same as the rate in which number of new exposed people increases minus the rate of people becoming infectious σ. Where, $1/\sigma$ is the mean latent period for the disease. As infected people recover at a constant rate (dR/dt), the effective rate of increase of infection (dI/dt) is equal to σE - dR/dt.
Epidemiological model defines a parameter R0 as the basic reproduction number which denotes the contagious ability of the disease. It is defined as the average number of people who can get infected from a single infected individual. It is formulated as β/γ. If the value of $R0 < 1$, it means
that the disease would not spread. If the value is $R0 = 1$, it signifies the spread is stable or endemic. If the value of $R0 > 1$ it means the spread would be pandemic. In this paper we found it to be 1.77 on average for COVID-19 spread in India.

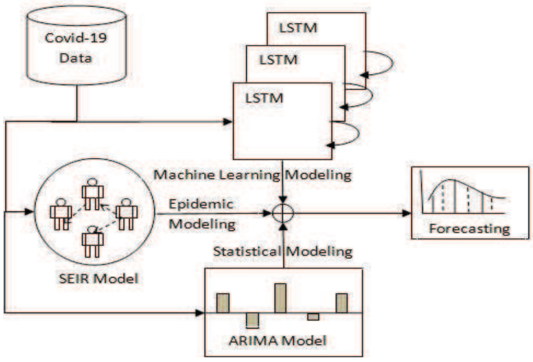


Figure1: Proposed architecture of prediction model

B. ARIMA model

It is a combination of two models namely Auto Regressive (AR) and Moving Average (MA) models. The ‘I’ stands for integration of the two models. An ARIMA model is represented with three parameters p, d and q, where p is the autoregressive lag, d is the order of differencing, and q is the moving average. ARIMA model is formally expressed as [12]

$$y_t = \theta_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

Where, the predicted value y_t = Constant + Linear combination Lags of y_{t-i} (upto p lags) + Linear Combination of Lagged forecast errors ε_{t-j} (upto q lags)

ARIMA model requires data to be stationary. A stationary time series has the mean and variance constant over time. If the data is non stationary i.e., the data has trend or seasonality, then we need to transform it into stationary series using differencing. The integration parameter *d* refers to the number of times differencing is required to get the data stationary. We can test stationarity with Augmented Dickey-Fuller test (ADCF). The p value of ADCF test should be much less than 1 for stationary data. Since the data for India had first increasing and then decreasing trend we performed differencing of order 2 to make the data stationary (Figure 2(a) and 2(b)) and obtained p-value of 0.002. For estimation of ARIMA parameters p and q we use the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots (Figure 2(c) and 2(d)).

C. LSTM model

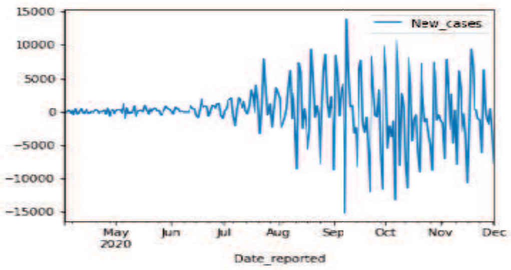
LSTM is a type of Recurrent Neural Network (RNN) that allows the network to retain long term dependencies at a given time from many previous time-steps. Time series data over progressive time frames of 7 days are prepared from the dataset to feed into LSTM model and the 8th day data is taken as the target value. This work uses a stacked LSTM model of 50 cells, and learns the best hyper parameters for matching the Indian scenario. The combination of selecting activation function as relu', optimizer as 'adam', loss function as 'mse', and recurrent activation function as "sigmoid" has given the best performance in cross validation. The model is trained on the daily new cases. from 1st April to 17th October and tested on data from 18th October to 31st Dec 2020. Finally it makes prediction on the daily cases for next three months, i.e up to 31st March 2021.

RESULT AND ANALYSIS

Figure 3(a) and (b) depict the SEIR modeling for COVID-19 cases in India. The recovery period is taken as 12 days which implies $\gamma= 1/12$. We consider the social contact rate=2.5 which makes the effective contact rate β=0.148. The basic reproduction number $R_{0,}$ which is the ratio of β and γ, is calculated to be 1.77. The selections of parameters are based on the result having best match with the current (up to 31st Dec 2020) record of the daily cases in India. The result predicts that the number of infected people (active cases) in India can come down to near 40,000 by the end of March 2021, if current trend continues. The SEIR model gives a simplistic idea of what would be the numbers in India in an unconstrained environment. However, it does not consider protecting measures like lockdown, social distancing, and use of mask and sanitizer which work as resisting factors against the spread of the disease.

The original non-stationary data for daily cases in India is differenced to make it stationary. We used ACF and PACF graph on this stationary data for an early estimation of the parameters (p,d,q) of the ARIMA model (Figure 2(c), 2(d)). Then the best combination (5,2,1) was found by applying grid- search method on the early estimated values. Figure 4 depict the ARIMA prediction on training data and future data respectively. Since ARIMA is solely based on previous time stamps, it is apparent from the result that although the model has learned the pattern successfully but the magnitude of the rise and fall in the daily new cases are not predicted properly.

For LSTM model, we split the dataset of 274 days data into train_data (200 days – 1st April to 17th October) and test_data (74 days - 8th October to 31st December). We prepare the input time series data with one week of previous time-steps for each record with the batch size of 50. Thus the input shape of the data becomes (50,7,1) which is fed to a stacked LSTM model. Figure 6(b) and 6(c) show training and testing performance of LSTM model on the COVID-19 daily new cases in India. Figure 6(d) shows the forecasted cases for future three months. The model predicts that the daily new cases would come to 5000 by the end of March 2021.



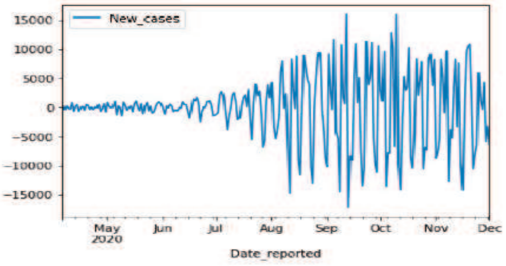


Figure 2(a): 1-Differenced data with p-value:0.

Figure 2(b): 2-Differenced data with p-value 0.002

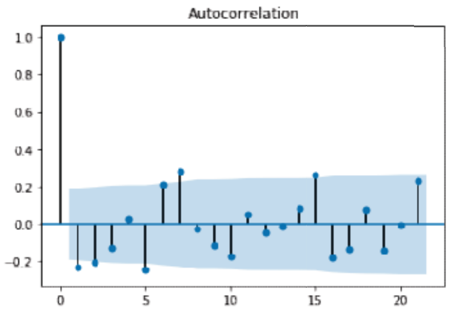


Figure 2(c): ACF graph after differencing(2)

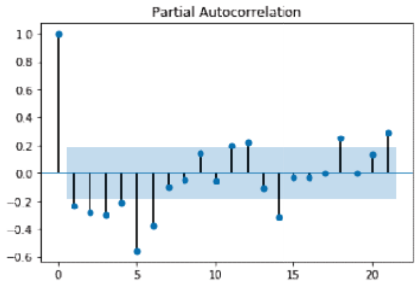


Figure 2(d): PACF graph after differencing(2)

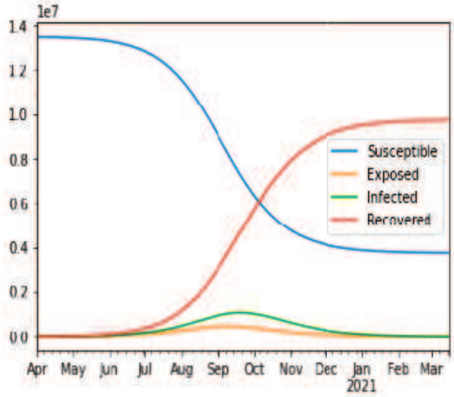


Figure 3(a): SEIR model for Indian cases

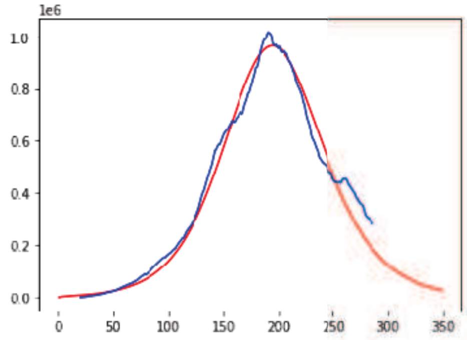


Figure 3(b): SEIR model predicting Active cases

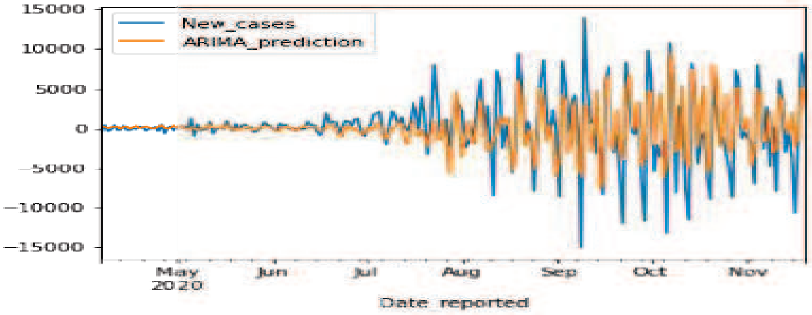


Figure 4: ARIMA prediction

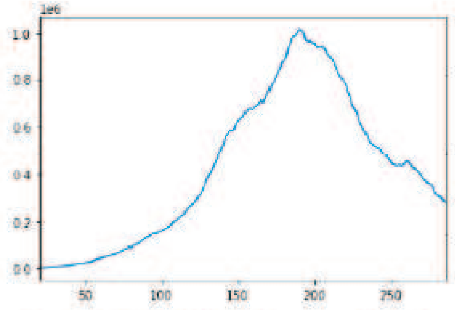


Figure 6(a): Daily Active cases(Total)

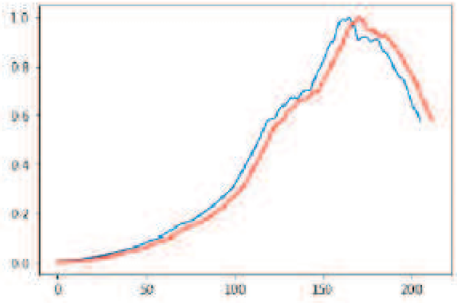


Figure 6(b): LSTM training data

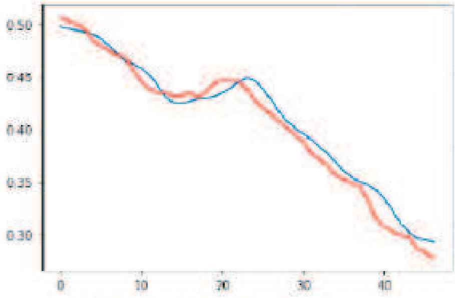


Figure 6(c): LSTM test data

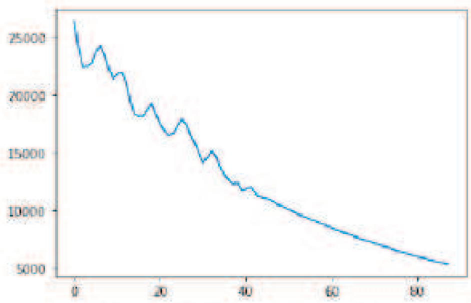


Figure 6(d): LSTM Forecasting

CONCLUSION

This paper presents an analysis of the COVID-19 spread in India. Three different modeling techniques namely SEIR, ARIMA and LSTM are used independently to find out any latent pattern in the spread of the disease between 1st April and 31st December 2020, and then a forecasting for next three months is made. The SEIR epidemical model learns the conversion rates of susceptible - exposed - infectious – recovered that fit best for our case. The value of basic reproduction number (R_0) is found to be 1.77 for India. The SEIR model predicts that the number of daily active cases would be around 40,000 and the total number of infected cases may reach up to 11.1 million by the end of March 2021. The LSTM model used a time frame of seven days to predict the next number of new cases. It is forecasted that the number of daily new cases would be around 5000 by the end of March 2021. The ARIMA model also confirms to this prediction. However, the limitation of this work is that it did not consider issues like social distancing, quarantine, and isolation which largely restrict the spread of the disease if maintained properly. However, at present there is no reliable data available on these attributes but the proposed model has the capability of adopting it when it will be available in future.

References

[1] Singh, R., & Adhikari, R. (2020). Age-structured impact of social distancing on the COVID-19 epidemic in India. arXiv preprint arXiv:2003.12055.

[2] Gupta, R., Pal, S. K., & Pandey, G. (2020). A Comprehensive Analysis of COVID-19 Outbreak situation in India. medRxiv.

[3] Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA based timeseries analysis to forecast near future. arXiv preprint arXiv:2004.07859.

[4] Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel corona virus (COVID-19) cases: A data-driven analysis. Chaos, Solitons & Fractals, 109850.

[5] Deb, S., & Majumdar, M. (2020). A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. arXiv preprint arXiv:2003.10655.

[6] Das, S. (2020). Prediction of COVID-19 disease progression in India: Under the effect of national lockdown. arXiv preprint arXiv:2004.03147.

[7] Pandey, G., Chaudhary, P., Gupta, R., & Pal, S. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv preprint arXiv:2004.00958.

[8] Tiwari, A. (2020). Modeling and analysis of COVID-19 epidemic in India. medRxiv.

[9] Tiwari, S., Kumar, S., & Guleria, K. (2020). Outbreak Trends of Coronavirus Disease–2019 in India: A Prediction. Disaster medicine and public health preparedness, 1-6.

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

[11] K. Prem, A. R. Cook, and M. Jit, "Projecting social contact matrices in 152 countries using contact surveys and demographic data," PLoS Comp. Bio 13, e1005697 (2017).

[12] Sowell, F. (1992). Modeling long-run behavior with the fractional ARIMA model. Journal of monetary economics, 29(2), 277-302.

[13] URL: https://www.kaggle.com/allen-institute-forai/ CORD-19-research-challenge

[14] URL: https://COVID19.who.int/

[15] Li, M. Y., Smith, H. L., & Wang, L. (2001). Global dynamics of an SEIR epidemic model with vertical transmission. SIAM Journal on Applied Mathematics, 62(1), 58-69.