SEIR-Based Model for India COVID-19 Prediction

Tan Meng [1, †], Peidong Ye [2, †], Yifei Zhao [3, *, †] Jie Zheng [4, †], Xiaoxue Zuo [5, †]

[1] College of Veterinary Medicine, China Agriculture University, Yantai, China, 264670

[2] Software Engineering, Xiamen University, Xiamen, China, 361102

[3] Business School, Seoul National University, Seoul, Korea, 08826

[4] School of Computer Science and Information Technology, Beijing Jiaotong University, Weihai, China, 264401

[5] School of Physical Science, University of Science and Technology of China, Hefei, China, 230026

* Corresponding Author Email: 18722039@bjtu.edu.cn

[†] These authors contributed equally.

Abstract. Up to August 19, 2021, there have been 207,784,507 confirmed cases and 4,370,424 confirmed deaths worldwide because of the COVID-19. Facing such a difficult situation of the epidemic situation of COVID-19, all the countries must struggle together and help each other. Just as the Peterson Institute for International Economics says in their report: The pandemic is not under control anywhere unless controlled everywhere. Therefore, it is crucial to pay more attention to the developing countries with fewer vaccines of COVID-19 and insufficient medical resources. Considering that India is one of the largest developing countries and the epidemic situation there is very serious, we decided to choose India as our research object. We propose a modified SEIR model, added the vaccination rate and the stringency of isolation measures - two crucial factors of the development of the COVID-19. It can be a helpful reference to comparative research and policies of governments and help us better judge the epidemic situation in developing countries.

Keywords: SEIR model, multiple linear regression, loss function

1. Introduction

The emergence of the SARS-CoV-2 strain of the human coronavirus has thrown the world into the midst of a new pandemic. In the human body, the virus causes COVID-19, a disease characterized by shortness of breath, fever, and pneumonia. Moreover, owing to its close genomic similarities to SARS-CoV, the virus that causes the disease SARS, SARS-CoV2 has the same way of transmitting through the inhalation of droplets and interaction with contaminated [1]. Due to the strong transmission ability of SARS-CoV2, there have been 207,784,507 confirmed cases and 4,370,424 confirmed deaths worldwide, which can be a massive threat to people's health [2]. Besides, it is also an unprecedented global societal and economic disruptive impact for all the world. In order to slow the spread of COVID-19, unprecedented measures are taken by many countries - major cities and even entire nations implemented lockdowns, restrictions on travel and gatherings, and closures of businesses and schools, which have stifled financial and economic confidence, sparking fears of a global recession [3].

Facing the current severe COVID-19, the whole world must work together, just as the Peterson Institute for International Economics (PIIE) says in their report: The pandemic is not under control anywhere unless controlled everywhere. Compared with the developed countries, the developing countries have fewer vaccines, worse medical treatments, and less effective government responses. Therefore, the COVID-19 pandemic in developing countries was not controlled well. The situation is very grave and appears to be deteriorating. And then a larger pool of infected people in countries where the pandemic remains uncontrolled provides a more extensive "laboratory" for viral variants vying for genetic dominance, which made the COVID-19 harder to defeat for the whole world. So how to help many developing countries control the epidemic is the key to whether the world can overcome the COVID-19 epidemic. Among most developing countries, India is one of the most

populous developing countries in the world. The epidemic is severe (According to the data of WHO, until August 19, 2021, 32320898 COVID-19 cases have been confirmed in India, and the cumulative number of deaths has reached 433063[4]). It can well represent the majority of developing countries lacking vaccines and good medical conditions. Therefore, we decide to choose India as our research object and design a new mathematical model to make some predictions for the development of the COVID-19 pandemic situation in India and hope it can help people better understand the situation of the pandemic in the developing countries and defeat the global pandemic of COVID-19 earlier.

There were many other outbreaks and transmission of diseases such as dengue fever, malaria, influenza, pestilence, and HIV/AIDS in the past. How to build a proper epidemiological model for these epidemics is a challenging task. At present, in epidemiology, the population risk of infection can be divided into four subclasses, S (susceptible), E (exposed), I (infected), and R (recovered), according to the different health statuses of the population [5]. Depending on the above population division, the model can be divided into four categories, SI [6], SIS [7], SIR [8], and SEIR [9]. Because the four subclasses (S, E, I, and R) all exit in the pandemic of COVID-19 and patients with COVID-19 have a specific immunity to COVID-19, SEIR is more suitable for the prediction of the COVID-19 epidemic situation. Some researchers used the SEIR model to finish many works. For example, He et al. designed an SEIR-based model combined with the PSO algorithm. They illustrated the main trends of the epidemic evolution reasonably [10]. Suwardi Annas et al. [11], Gaurav Pandey et al. [12] structured the SEIR model for the pandemic of COVID-19 in India and Indonesia.

However, with the continuous development of the COVID-19 epidemic and the world's efforts, some key factors have changed the COVID-19 pandemic, such as mass vaccination of the COVID-19 vaccine. Therefore, to better predict the pandemic situation in India and help control the global pandemic of COVID-19, some changes are made in this paper. We add some parameters representing vaccines and the efforts of the Indian government to the SEIR model. Original data set would be divided into the training set and verifying set, which is used for training the models and testing their accuracy. The training set would be pre-processed, such as normalized, at the beginning. Next, the data set would be used in the iteration of multiple linear regression and finally produce the model's parameter.

2. Method

Based on the SEIR model, we use two autoregressive models, two autoregressive distributed lag models, and linear regression to predict different parts of the spread of COVID-19. To show the impacts of quarantine policy and vaccination rate, we divide the time into three stages: spreading, quarantine and vaccination. We use the former day's data to predict this day's numbers, as shown in Figure 1.

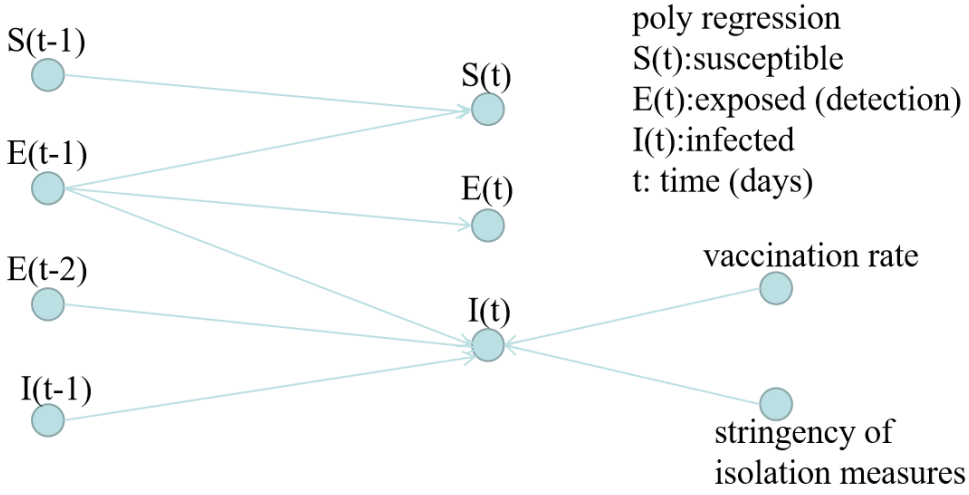

Figure 1. Relationships and key influencing factors in the models.

Because we do not find the data about the exposure in India, we replace it with the number of detections. However, unlike the number of the exposed, detection is limited by the number of instruments. It does not necessarily relate to the susceptible. The number of detections is an autoregressive model:

$$E(t) = \theta_{E0} + \theta_{E1}E(t-1) \tag{1}$$

Where $\theta_{E0}$and $\theta_{E1}$are both learnable parameters.

A. The first stage: the infected people can infect others freely because the government has not realized the virus's infectivity and does not act. The number of susceptible is an autoregressive distributed lag model.

$$S(t) = \theta_{S0} + \theta_{S1}S(t-1) + \theta_{S2}E(t-1) \tag{2}$$

Whereθ_{S0}, θ_{S1} and $\theta_{S2}$are all learnable parameters. And the number of the infected is also an autoregressive distributed lag model

$$I(t) = \theta_{I0} + \theta_{I1}I(t-1) + \theta_{I2}[E(t-1) - E(t-2) \tag{3}$$

Whereθ_{I0}, $\theta_{I1}$and θ_{I2} are all learnable parameters.

B. The second stage: the government takes isolation measures to control the spread. There are many different measures, so we use an overall stringency to replace them, affecting isolation's control effect. The number of the infected is:

$$I(t) = \theta_{I0} + \theta_{I1}I(t-1) + \frac{\theta_{I2}[E(t-1)-E(t-2)]}{[1+\mu(t-1)]} \tag{4}$$

Where $\mu(t-1)$ is the stringency of isolation measures predicted by the autoregressive model?

$$\mu(t) = 2\mu(t-1) - \mu(t-2) \tag{5}$$

C. The third stage: the vaccine is developed, and people are gradually vaccinated. With the increase of vaccination rates, people who contact the virus carriers will decline. Because of the scatter spot of the number of vaccinations, we use linear regression to predict its growth trend and discard the early part of data that is too small to influence the influence to fit the last curve better. The number of the infected is:

$$I(t) = \theta_{I0} + \theta_{I1}I(\text{t-1}) + \theta_{\text{I2}}(1-\varphi(t-1))[E(t-1) - \frac{E(t-2)]}{[1+\mu(t-1)} \tag{6}$$

Whereθ_{I0}, θ_{I1} and $\theta_{\text{I2}}$are all learnable parameters, and $\varphi(t-1)$ is vaccination rate predicted by the linear regression:

$$\varphi(t) = \theta_0 + \theta_1 t \tag{7}$$

Besides, there is a turning point between March and April in 2021 which means an epidemic resulting from mass crowd gatherings. So, there should be a sudden change in a parameter, which lasts for a while. This happened at the beginning of the third stage and may fall back in the latter part.

3. Experimental results and analysis

3.1. Data description and pre-processing

The experiment is aimed to research the number of infected in three different stages, based on the models. The data about the detection of COVID-19 in India, from Kaggle, is used to train the machine-learning model, which predicts the number of infected in the future. The original data set has been filtrated into three new sets, including features demanded in the equation. For instance, there are features like the date, the number of tests, the number of diagnosed patients in the tests, the number of recovered patients, the number of deaths, the increasing days in this period, the number of susceptible already exist, the number of vaccinators, and the intercept which can improve the

efficiency of calculation. In addition, there are some other features in the data set of infected, including the number of vaccinators in the previous day, the number of diagnosed patients in the previous day, the number rate of vaccinators in the previous day. Furthermore, there are five features in the model for vaccinators, date, and the type of vaccine, the number of people who got one shot of the vaccine, the number of people who got two shots of vaccine, and the number of tests. These features are all used in the model training, which can be considered the fundamental factors that influence the spread of COVID-19.

The original data has been processed before training, plenty of useless features have been deleted. In addition, invalid data has also been rejected. The normalization has been processed in this stage, which is aimed to eliminate dimensional effects and optimize the implementation of analysis. We set the mean as 0 and the standard deviation as 1 to follow the normal distribution. The process implements the equation below.

$$X = \frac{X' - m}{std} \tag{8}$$

X Is the data required processing, m is the mean of X and STD is the standard deviation. Apart from that, according to the feature Time, which implies the date, 1/3 of the data has been chosen for verification. The rest of the data set has been fed into model calculations and fit results of training. In addition, feature scaling is implemented because there is a difference in magnitudes between different features. This method makes the influence between features significant. On the other hand, scaling can avoid overflow when data values are enormous, especially in calculating loss.

3.2. Experimental results

In the experiment, multiple linear regression has been used for the training model predicting the number of the infected. Linear regression has an equation in the following form:

$$y = aX + b \tag{9}$$

In the situation of multiple linear regression, X implies the matrix of features. Three models are constructed to fit the changing curves of infected, susceptible, and vaccinator, based on the equations proposed in the method section, after 300,000 times of iteration.

A. Susceptible

This model implies the number of susceptible, which is based on the equation (2). Therefore, the result can be formed as the linear function of the number of susceptible in the previous day and the number of detections decided by equation (1). The result of the linear regression is:

$$y = 40011869.3 + 0.9780042\ x1 - 0.02179016\ x2 \tag{10}$$

$$x1 = S(t-1) \tag{11}$$

$$x2 = E(t-1) \tag{12}$$

$x1$ Is the number of susceptible in the previous day, and $x2$ is the number of detection in the previous day. Therefore, it is indicated that the number of susceptible in the previous day is the most fundamental factor affecting the susceptible. At the same time, the detection has a slight influence on the result. The result is visualized in Figure 2.

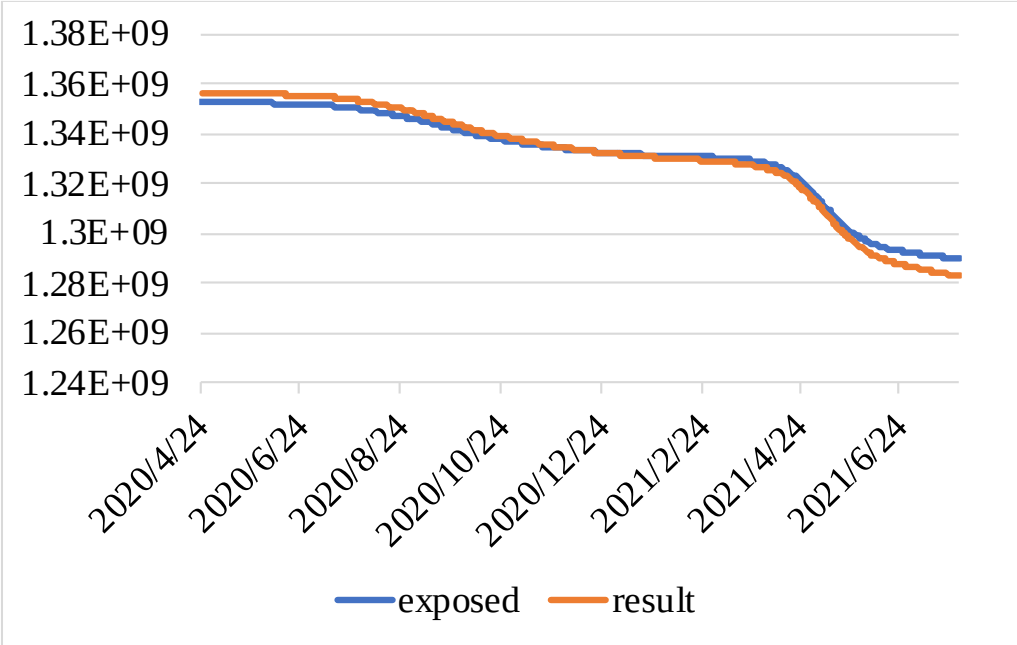

Figure 2. Comparison of Experimental and Real Value of Susceptible.

According to Figure 2, the number of susceptible remained stable initially. Then it started to decline slightly. However, the value started to decrease rapidly in about 2021/3/24. In addition, the value of the experiment tends that higher than the real value in the end, which demands some improvement in future research if the conditions change significantly. The loss function, which evaluates the performance of the model according to the difference between experimental and real value, is shown as follow, which is in the form of the equation:

$$L(y, y\hat{}) = (y + y\hat{})^2 \tag{13}$$

The loss function is the mean of $L(y, y\hat{})$, and the learning rate of the optimizer is 0.005. Figures 3, visualizing the loss of susceptibility, show that the loss is reduced during the iteration and finally minimized.

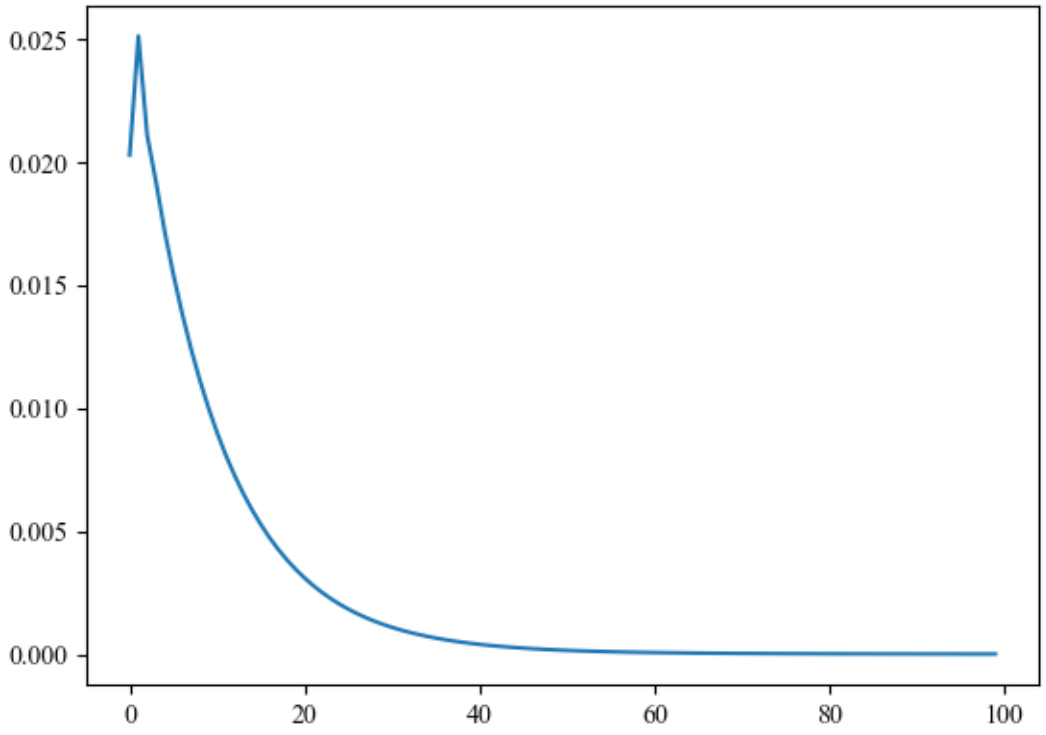

Figure 3. Loss Function of Susceptible.

The fluctuation sometimes occurs in Figure 3 because the initial parameter is generated randomly and the minimum could be skipped. In this case, the learning rate decreases until it rises up, which is divided by 10 or 2 in each iteration time.

B. Infected

According to equation (6), the number of infected affected by the number of infected in the previous day, the number of vaccinators, and the number of suspected patients. The result is

$$y = -0.0163194 + 1.0024213\ x1 - 0.01295365\ x2 \tag{14}$$

$$x1 = \mathrm{I}(\mathrm{t}-1) \tag{15}$$

$$x2 = \frac{\big(1-\phi(\mathrm{t}-1)\big)\big(\mathrm{E}(\mathrm{t}-1)-\mathrm{E}(\mathrm{t}-2)\big)}{1+\mu(\mathrm{t}-1)} \tag{16}$$

Where $x1$ means the number of infected in the previous day. In the equation of$x2$, $\phi(t-1)$ means the number of vaccination rates in the previous day, $E(t-1)$ and $E(t-2)$ imply the number of suspected patients in the previous day and two days before, $\mu(t-1)$ is the stringency of isolation measures. Th is result can conclude that the number of infected affected most efficiently, and the conditions including vaccination rate, isolation measures, and increase of the suspected inhibit the growth of infected. The resulting figure is shown in Figure 4.

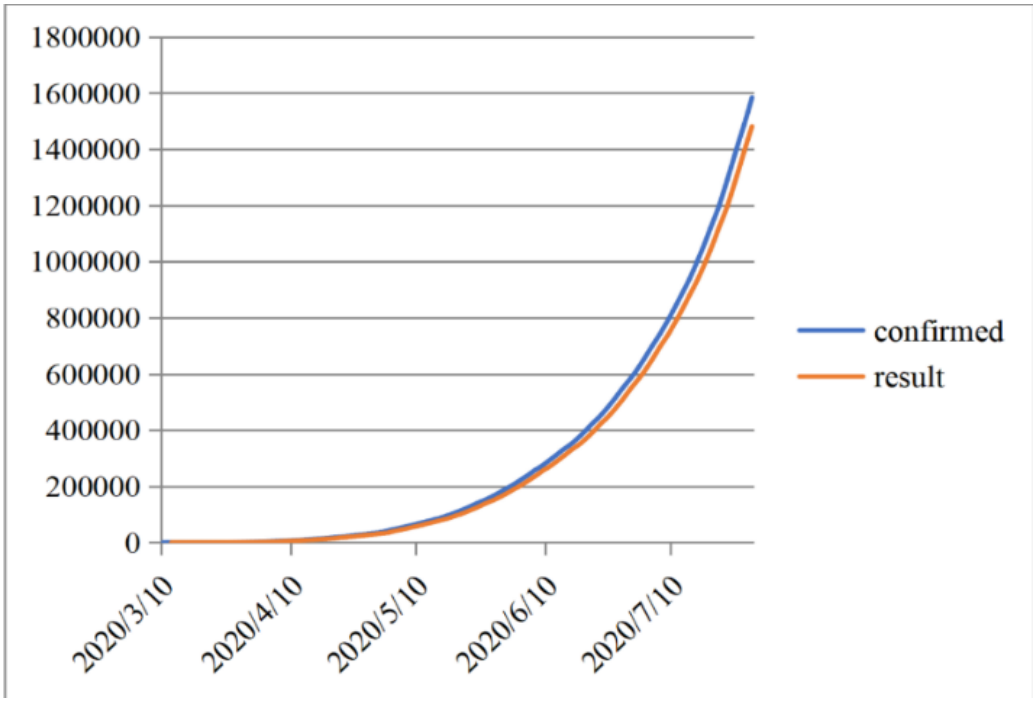

Figure 4. Comparisonof Experimental and Real Value of Infected.

It shows that the number of infected is keeping increasing, and the speed is getting higher. Therefore, the effect of the inhabiting condition is little in the case of the large amount the diagnosed patients. The loss function of this model is similar to Model A. However, the learning rate of this model is 0.01, and it drops faster without fluctuation.

C. Vaccinator

The final model is about the vaccinator, according to equation (7). The result is

$$y = (-4490.447 + 269.04877\ x) * 10000 \tag{17}$$

Where x implies the increase of date as a consequence of reducing the order of magnitudes before training for convenience, the result was multiplied by 10000. The result of this model is shown in Figure 5.

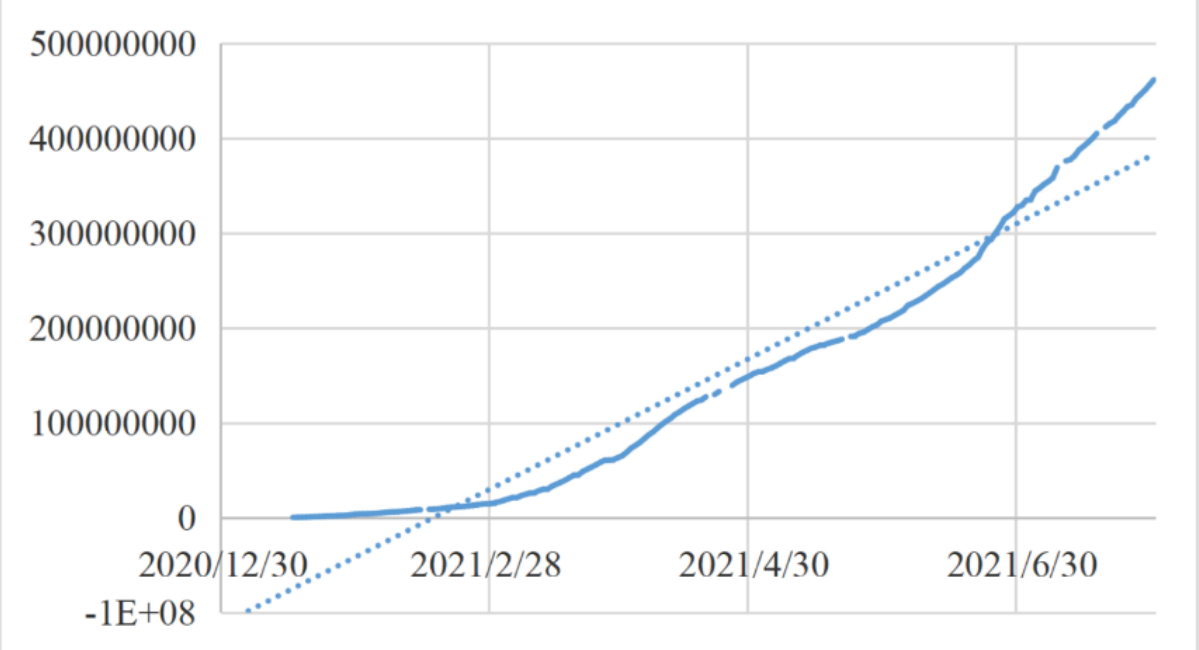

Figure 5. Comparison of Experimental and Real Value of Vaccination.

It shows that the experimental fit the value well in the middle period. The real value increased moderately initially and started to rise at high speed recently, caused by the development of vaccine production and change of policies. Suppose it keeps a tendency of increase in the future. In that case, the model could be modified by ignoring some data recorded in early 2021.

There are three data set used for model training, which was chosen about 1/3 data for verification. Specifically, in model A, data from January 22 to April 23 in 2020 are used for verification, while the rest are used for training. In model B, the data between January 30 and March 10 are used for verification, and data between March 10 and July 29are used for training. In model C, data from January 15 to February 8 in 2021 is used for verification, and the rest until July 30 are used for training. The models fit the real value well. It can predict the number of susceptible, infected, and vaccinators accurately. This conclusion is quite what we expected.

4. Conclusion

In conclusion, due to the pandemic of COVID-19 and its variation, the global economy and safety suffer an enormous blow, especially in developing countries, such as India, which is the object of the project because it is a typical case that has value to research. Based on the SEIR model, which is a commonly used model about infectious diseases, two autoregressive models and two autoregressive distributed lag models are built for predicting the spread of COVID-19. In the next step, the three models are trained by multiple linear regression. The original data sets are filtrated and pre-processed before training, including normalization and feature scaling methods, to optimize the result's analysis. There are three models built for susceptible, infected, and vaccinator, marked as A, B, C, respectively. During the training process, loss function, which indicates the model's performance by measuring the difference between experimental and real value, is minimized by gradient descent. The unique situation that the loss rises occurred in model A is settled by reducing the learning rate in each iteration, which took 0.005 as the initial value. The results have been analyzed and drawn in the figure after verification with 1/3 data, which indicates that they fit the real value well. Furthermore, the models could be modified and retrained if conditions change in the future. Therefore, the models can predict the number of susceptible, infected, and vaccinators accurately, which is a helpful reference to comparative research and policies of governments.

References

[1] Atzrodt, C.L., Maknojia, I., McCarthy, R.D.P., Oldfield, T.M., Po, J., Ta, K.T.L., Stepp, H.E. and Clements, T.P. (2020), A Guide to COVID-19: a global pandemic caused by the novel coronavirus SARS-CoV-2. FEBS J, 287: 3633-3650. https://doi.org/10.1111/febs.15375

[2] Coronavirus disease (COVID-19) (who.int)

[3] elso JK, Milne GJ, Kelly H. Simulation suggests that rapid activation of social distancing can arrest epidemic development due to a novel strain of influenza. BMC Public Health. 2009 April 29; 9: 117. doi: 10.1186/1471-2458-9-117. PMID: 19400970; PMCID: PMC2680828.

[4] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=IN

[5] Osman, M. A.-R., Adu, I., & Yang, C. (2017). A Simple SEIR Mathematical Model of Malaria Transmission. Asian Research Journal of Mathematics, 7(3), 1-22. https://doi.org/10.9734/ARJOM/2017/37471

[6] Allen LJ. Some discrete-time SI, SIR, and SIS epidemic models. Math Biosci. 1994 Nov; 124(1):83-105. doi: 10.1016/0025-5564(94)90025-6. PMID: 7827425.

[7] J. Liu, P. E. Paré, E. Du and Z. Sun, "A Networked SIS Disease Dynamics Model with a Waterborne Pathogen," 2019 American Control Conference (ACC), 2019, pp. 2735-2740, doi: 10.23919/ACC.2019.8815082.

[8] Brugnano L, Iavernaro F, Zanzottera P. A multiregional extension of the SIR model, with application to the COVID-19 spread in Italy. Math Methods Appl Sci. 2020 Nov 23:10.1002/mma.7039. doi: 10.1002/mma.7039. Epub ahead of print. PMID: 33362323; PMCID: PMC7753330.

[9] Osman, M. A.-R., Adu, I., & Yang, C. (2017). A Simple SEIR Mathematical Model of Malaria Transmission. Asian Research Journal of Mathematics, 7(3), 1-22. https://doi.org/10.9734/ARJOM/2017/37471

[10] He S, Peng Y, Sun K. SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dyn. 2020 Jun 18:1-14. doi: 10.1007/s11071-020-05743-y. Epub ahead of print. PMID: 32836803; PMCID: PMC7301771.

[11] Rustan R, Handayani L. the outbreak's modeling of coronavirus (COVID-19) using the modified seir model In Indonesia [J]. SPEKTRA Jurnal Fisika dan Aplikasinya, 2020, 5(1):61-68.

[12] Pandey G, Chaudhary P, Gupta R, et al. SEIR and Regression Model based COVID-19 outbreak predictions in India. 2020.