Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220054

Modeling of Covid-19 Transmission Using Machine Learning

Zhuang TIAN[a], Yu CAO[a,1], Xuting ZHENG[b] and Jingping ZHANG[b]
[a] *Liaoning Petrochemical University, China*
[b] *The First Hospital of China Medical University, China*

Abstract. A susceptible-infected-susceptible (SIS) model with a nonlinear infection rate, a forecast model based on autoregressive integrated moving average (ARIMA), and a forecast model based on long short-term memory (LSTM) artificial neural networks were developed using the COVID-19 epidemic data from four countries (China, Italy, the United Kingdom, Germany, France, and Poland) to simulate and forecast the epidemic trends in these countries. The models were compared in terms of forecast errors, and the LSTM model was found to forecast virus transmission very well.

Keywords. SIS model, COVID-19, LSTM, ARIMA, epidemic forecast.

1. Introduction

The virus that causes COVID-19 has spread rapidly since the outbreak of the epidemic in Wuhan at the end of December 2019. As of November 21, 2020, there was a cumulative total of 57,839,814 individuals with confirmed COVID-19 and a cumulative total of 1,373,300 deaths from COVID-19 worldwide; this disease continues to severely threaten human lives around the globe. Therefore, it is necessary to study virus transmission models so that the epidemic trends can be understood and forecast in a timely manner and early preventive and control measures can be taken.

In 1927, to reflect the virus transmission process, Kermark and McKendrick proposed the compartment model. Classical compartment models include the SI, SIS, and SIR models. To better reflect the development of infectious diseases, various improvements in the classical compartment model have been made over nearly a century. Fan et al. used the SEIR model to simulate the spread of the epidemic in Wuhan and forecast and analyzed the inflection points. They well forecast the development trend of the epidemic and proposed prevention and control measures; in addition, they also noted that the model is unable to judge the length of the incubation period, and therefore, it is difficult to count the number of people in the incubation period[1]. Chang et al. established a model with a nonlinear infection rate to investigate the impact of media coverage on the infection rate in an infectious disease model; they analyzed the local stability of each equilibrium point and the global stability of the model validate the model and demonstrate the important influence of media coverage on disease transmission [2]. However, it is difficult to determine the threshold of the nonlinear function when the

[1] Corresponding Author, College of Computer and Communication Engineering, Liaoning Petrochemical University, Shenyang; E-mail: Yucao_lnshu@163.com.

model is used in actual situations. Chen et al. developed a time delay dynamic transmission model that considered that COVID-19 infections occurred in the incubation period and noted the importance of isolation for epidemic prevention and control; however, the model has a drawback: the infection rate in the incubation period cannot be determined [3]. Esmaeilzadeh et al. used an autoregressive integrated moving average (ARIMA) model to fit and analyze data from Iran to derive the influence of crowd gatherings on virus transmission and suggested prevention and control measures to curb the spread of the epidemic; however, they noted that the model is susceptible to the influence of outlier data and only has good performance in the very short term [4]. Based on preprocessing of epidemic data, Sheng et al. applied a logistic model for the free transmission stage to compare and analyze the epidemic data five days in advance and later with the actual data[5,6], demonstrating the importance of taking timely epidemic prevention measures. Luo investigated deep learning-based techniques for infectious disease forecasts[7~9]; the forecast of hand, foot and mouth disease in Xiamen was studied in detail using linear and nonlinear relations in modeling; and the forecast accuracy of an ARIMA model was compared with that of the long short-term memory (LSTM) model with different step sizes, achieving good results [10].

In the present study, we used global COVID-19 epidemic data released by the Johns Hopkins University Center for Systems Science and Engineering to establish an SIS model with a nonlinear infection rate[11], an ARIMA model[12], and an LSTM model[13] for forecasting the epidemic trends in four countries (China, Italy, the United Kingdom/UK, Germany, France, and Poland). The forecast accuracies of the models were quantified to determine the method that can best reflect the development trends of the epidemic, so as to provide guidance for future epidemic prevention and control.

2. Data Description

This study was conducted based on epidemic data from four countries (China, Italy, the UK, Germany, France, and Poland) derived from the global COVID-19 epidemic data and statistics website published by the Johns Hopkins University Center for Systems Science and Engineering[14]. The data source started on January 22, 2020, and continues to date, with data updated every morning and evening. The data sources contain three data files: time-series statistical data of the cumulative numbers of confirmed cases, deaths, and recovered cases during the COVID-19 pandemic[15]. Thus far, the epidemic has been largely controlled in the four countries, and this study focuses on forecasting during the early stage of the epidemic. For this reason, the data from January 22 to April 11, 2020, were selected for China and, to facilitate the study, the data from March 1 to June 30, 2020, were selected for the other three countries, where the COVID-19 epidemic outbreak occurred relatively late.

3. Introduction to the Modeling Methods

3.1. Sis Model with a Nonlinear Infection Rate

Based on the COVID-19 transmission mechanism, this study uses an SIS model as the baseline model[16]. This model divides the population into two compartments, and the population is transferred between the compartments with a certain probability. The transmission mechanism is shown in the following figure 1:

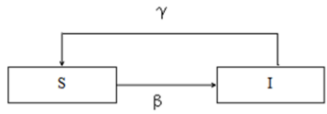

Figure 1. Transmission mechanism, SIS model.

where S is the healthy state and I is the infected state. The S population contacts the I population and is infected with a probability β, becoming the I population; the I population recovers with a probability γ, becoming the S population[17]. The SIS model is expressed as follows:

$$\begin{cases} \frac{dS}{dt} = -\beta * S * I / N + \gamma I \\ \frac{dI}{dt} = \beta * S * I / N - \gamma I \end{cases} \tag{1}$$

In the above expression, N = S + I. Because the transmission of infectious diseases can be affected by a variety of factors, the infection rate β in the SIS model is often nonlinear[18]. Under human intervention, the infection rate tends to decrease gradually. Wang considered the deterministic infectious disease model with an infection rate of $\frac{\beta SI}{\varphi(I)}$:

$$\begin{cases} \frac{dS}{dt} = -\beta(I) * S * I / N + \gamma I \\ \frac{dI}{dt} = \beta(I) * S * I / N - \gamma I \\ \beta(I) = \frac{\beta}{1 + mI} \end{cases} \tag{2}$$

where m is the influencing factor and is greater than zero. Clearly, the function decreases monotonically in the domain of definition, and hence, it is consistent with the development of the epidemic in the presence of government intervention or increased public awareness of epidemic prevention.

3.2. Arima Model

Proposed by Box and Jenkins, ARIMA models are widely used for time-series data forecasting. These models are suitable for stationary time-series data, and nonstationary time-series data can be made stationary by methods such as log transformation and differentiation[19]. The ARIMA model has three parameters, i.e., p, d, and q:

autoregressive order, differential order, and moving average order, respectively. The model is mathematically expressed as follows:

$$\left(1-\sum_{i=1}^{p}\Phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \tag{3}$$

In determining the three parameters of the ARIMA model, the present study introduced the Akaike information criterion (AIC) to avoid the subjectivity of judgment using autocorrelation and partial autocorrelation plots. The AIC was proposed by Akaike in 1973, with a full name of the minimum information criterion, which is a standard measure of the goodness of fit of a statistical model. In the present study, the AIC value was used as a criterion for model optimization, and the optimal ARIMA parameters were obtained through iteration.

3.3. Lstm Model

The LSTM network was proposed by Schmidhuber in 1997 to overcome the gradient disappearance and gradient explosion problems of traditional recurrent neural networks (RNNs). It is built based on RNNs, with the addition of a memory module, which effectively solves the long-term dependency problem[20].

LSTM stores historical information through the state of memory units, each of which contains three "gate" structures, i.e., input gate, forget gate, and output gate, to control whether the state of the memory unit is to retain or discard the information. The structure of the LSTM memory unit is shown in the following figure 2.

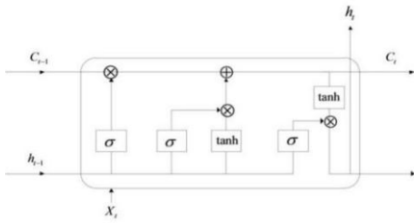

Figure 2. The structure of the LSTM model.

The input, gating, and output are realized in a memory unit module through a "gate" structure, and precisely because of this special structure, the LSTM model is very effective for forecasting time-series data.

4. Modeling of Covid-19 Transmission

In the modeling process for China, the data from January 22 to March 1, 2020, were used for model fitting or training, and the data from March 2 to April 11, 2020, were used for model testing. In the modeling process for the other three countries, due to the relatively slow development of the epidemic, the data from March 1 to May 29, 2020, were used for model fitting or training, and the data from May 30 to June 30, 2020, were used for model testing. Then, the forecast values were compared with the actual values to assess the goodness of the model forecast.

4.1. Mechanism Modeling

The training data for China and the other three countries were each imported into the improved SIS model, and the error between the fitted values of the model and the actual values was used as the loss function, which was minimized using the minimize function in the lmfit package (which encapsulates scipy.optimize to make it more user friendly) of Python to obtain the optimal model and parameters for each country. The optimal model parameters for the four countries are shown in table 1.

Table 1. Optimal parameters for the improved SIS model.

country	m	β	γ
China	0.00809260	0.34279596	0.03046488
Italy	0.02488288	0.24021533	0.02850514
the UK	0.06820503	0.38080172	0.00433425
Germany	0.07834901	0.54591313	0.06554033

To more intuitively show the forecast results of the improved SIS model, the results were visualized using plotly package in Python, which is a very powerful open-source data visualization framework that can create interactive graphs in web form. In this study, the graph_objects module was used to visualize the forecast results. The results are shown in the following figures 3-6.

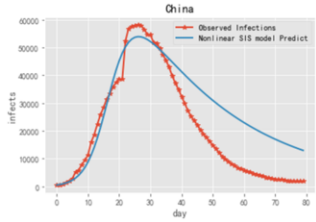

Figure 3. Forecast of the epidemic in China using the improved SIS model.

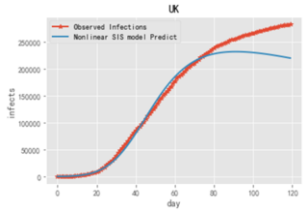

Figure 5. Forecast of the epidemic in the UK using the improved SIS model.

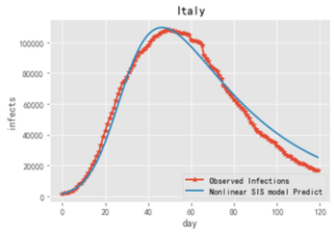

Figure 4. Forecast of the epidemic in Italy using the improved SIS model .

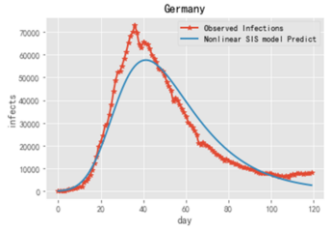

Figure 6. Forecast of the epidemic in Germany using the improved SIS model.

As seen in the above figures, the number of daily infections is in general not very accurately forecast by the improved SIS model, except for a few countries, for which the forecast is barely acceptable.

4.2. Machine Learning Models

The training data were introduced into the ARIMA model, the parameter p was set to 0-4, and q and d were each set to 0-2, resulting in 32 different models. The optimal

model was obtained by minimizing the AIC of the model and the training data. The optimal parameters (p, d, q) for the ARIMA model were as follows: (3,1,1) for China, (3,1,0) for Italy, and (3,1,1) for the UK. The forecast results obtained using the optimal model are provided in the following figures 7-10:

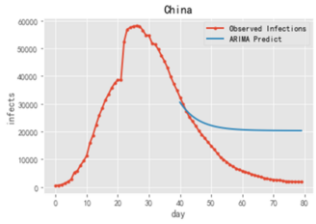

Figure 7. Forecast of the epidemic in China using the ARIMA model.

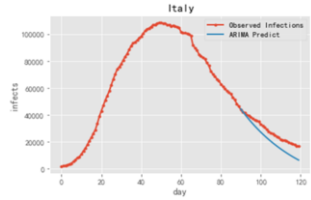

Figure 8. Forecast of the epidemic in Italy using the ARIMA model.

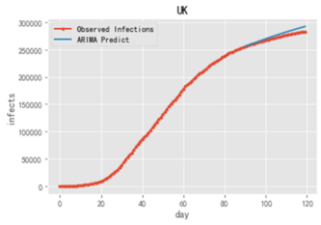

Figure 9. Forecast of the epidemic in the UK using the ARIMA model.

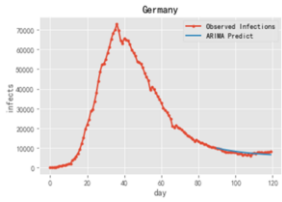

Figure 10. Forecast of the epidemic in Germany using the ARIMA model.

In the above figures, the red and blue curves show the actual and forecast numbers of infections in a country, respectively. When the fluctuation of a curve is not large, i.e., relatively smooth, the forecast results by the ARIMA model are fairly good; however, the forecast results are poor for cases with large fluctuations in data.

Similarly, the training data were imported into the LSTM model, and the LSTM network was implemented using the Keras framework in TensorFlow. A total of 100 epochs were trained, and the root mean square error (RMSE) was used as the loss function. After 100 training sessions, the loss decreased to 1e-4. The forecast results for the LSTM model are provided in the following figures 11-14:

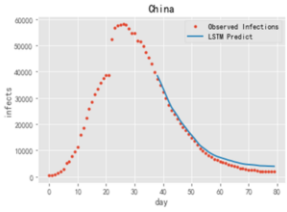

Figure 11. Forecast of the epidemic in China using the LSTM model.

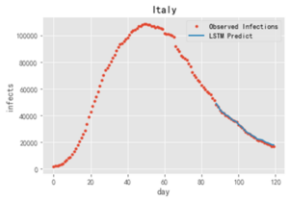

Figure 12. Forecast of the epidemic in Italy using the LSTM model.

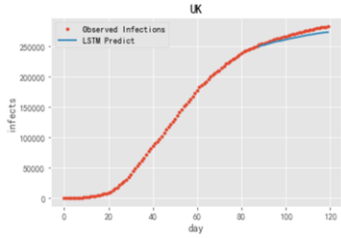

Figure 13. Forecast of the epidemic in the UK using the LSTM model.

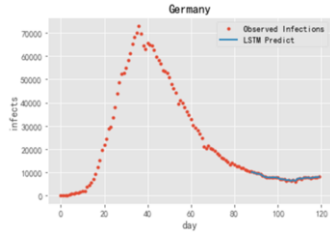

Figure 14. Forecast of the epidemic in Germany using the LSTM model.

In Figures 11-14, the yellow smooth curve shows the forecast number of infections using the LSTM model, and the blue curve is the actual number of infections. As seen, the LSTM model performs well in forecasting the epidemic for the four countries, a result that is attributed to the memory function of the LSTM network. In this study, after a number of training sessions, three was the optimal value of the parameter look_back.

5. Comparison of Simulated Forecasts

To facilitate the analysis of the forecast performance of the three methods, the forecast results obtained for the four countries using the three models in this study are provided in the following figures 15-18.

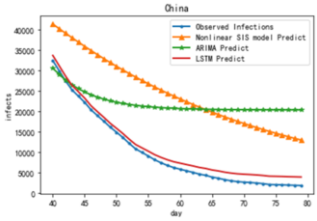

Figure 15. Comparison of forecasts by the three models for the epidemic in China.

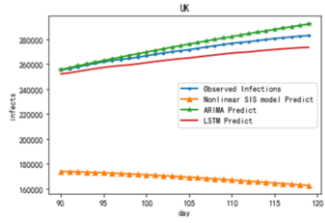

Figure 17. Comparison of forecasts by the three models for the epidemic in the UK.

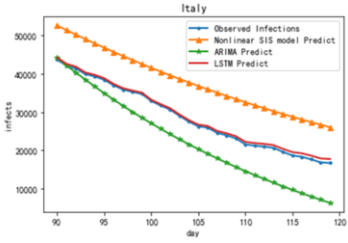

Figure 16. Comparison of forecasts by the three models for the epidemic in Italy.

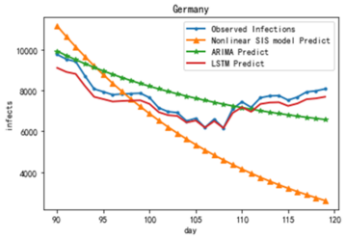

Figure 18. Comparison of forecasts by the three models for the epidemic in Germany.

As seen in the above figures, the LSTM model has the best forecast performance among the three models.

To quantify the difference between forecast results, we used the mean absolute error (MAE) between the forecast and actual values as the reference standard. The MAE is defined as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m} | y_i - x_i | \tag{4}$$

where y_i is the forecast value and x_i is the actual value. The smaller the MAE is, the smaller the error between the forecast and actual value, i.e., the better the forecast performance. The MAEs for the forecast results by different models for each of the four countries are provided in table 2.

Table 2. MAEs for the forecast results by different models.

Country	Improved SIS model	ARIMA model	LSTM model
China	14789	12187	1125
Italy	10911	6099	567
the UK	42895	4236	6281
Germany	2126	782	116

As seen in table 2, the LSTM model significantly improved the forecast accuracy, and hence, the corresponding results are more in line with the actual development of the epidemic.

6. Conclusion

Based on data for the COVID-19 epidemic in China, Italy, the UK and Germany obtained from the Johns Hopkins University Center for Systems Science and Engineering, this study discussed and established an SIS model with a nonlinear infection rate, an ARIMA model, and an LSTM model and used them to forecast the development trends of the COVID-19 epidemic. The MAEs between the simulated forecast values and the actual values were calculated as the basis for model comparison. The LSTM model had the best forecast performance among the three models. The forecast results for the LSTM model based on the available data had high accuracy. In addition, the LSTM model has a built-in function for preventing overfitting, which indeed served the purpose. Therefore, the LSTM model can well forecast the future transmission trends of the COVID-19 epidemic, which has far-reaching significance for the prevention and control measures in response to the development of the epidemic.

Acknowledgments

Project funded by the Key Research and Development Program of Liaoning Province:Research on Scientific Early Warning, Prevention, and Control of the COVID-19 Epidemic (2020JH2/10300040) and the Liaoning Provincial Department of Education:Research on Mining and Modeling of Factors Associated with Epidemic Transmission Using Machine Learning (L2020031).

References

[1] Fan Ruguo, Wang Yibo, Luo Ming, et al. A novel coronavirus pneumonia transmission model and inflection point prediction based on SEIR [J]. Journal of UESTC,2020,49(03):369-374
[2] Chang Yuting, Niu chenjie, Yang Jian. An epidemic model with nonlinear infection rate under the influence of media reports[J]. Journal of Luoyang Normal University, 2019, 38(11):6-103
[3] Yu Chen, Jin Cheng, Yu Jiang, et al. A time delay dynamical model for outbreak of 2019-nCoV and the parameter identification. 2020, 28(2):243-250
[4] Nayereh Esmaeilzadeh, Mohammadtaghi Shakeri, Mostafa Esmaeilzadeh, Vahid Rahmanian.ARIMA models forecasting the SARS-COV-2 in the Islamic Republic of Iran[J].Asian Pacific Journal of Tropical Medicine,2020,13(11):521-524
[5] Sheng Huaxiong, Wu Lin, Xiao Changliang. Modeling and prediction of novel coronavirus pneumonia epidemic situation [J]. Journal of system simulation,2020,32(05):759-766
[6] Cao Yu, Jing Yuanwei. Establishment and stability analysis of SIRS model with nonlinear infection rate [J]. Control theory and application, 2013,30 (02): 229-232
[7] Zhou Yanli, Pu Guiping. A study on SIS epidemic vaccination model with nonlinear infection rate [J]. Journal of Shanghai University of technology, 2019,41 (04): 339-343
[8] COVID-19/SARS-CoV-2 News from Preprints; COVID 19: Real-time Forecasts of Confirmed Cases, Active Cases, and Health Infrastructure Requirements for India and its Majorly Affected States using the ARIMA model.[J]. Medical Letter on the CDC & FDA,2020
[9] Shen pingxu, Wen Chenglin, Sun Xiaohui, et al. Multi step prediction of LSTM network based on variable correlation analysis [J]. Power science and Engineering.
[10] Luo Jianxiang. Research on infectious disease prediction technology based on deep learning [D]. Jimei University,2020
[11] Liu Xiaodong, Wei Haiping, Cao Yu, et al. Modeling of scarlet fever transmission process [J]. Computer simulation, 2020, 37 (08):171-176 + 375
[12] Pradeep Mishra,Chellai Fatih,Deepa Rawat,Saswati Sahu,Sagar Anand Pandey,M. Ray,Anurag Dubey,Olawale Monsur Sanusi. Trajectory of COVID-19 Data in India: Investigation and Project Using Artificial Neural Network, Fuzzy Time Series and ARIMA Models[J]. Annual Research & Review in Biology,2020
[13] Magesh S.,Niveditha V.R.,Rajakumar P.S.,Radha RamMohan S.,Natrayan L.. Pervasive computing in the context of COVID-19 prediction with AI-based algorithms[J]. International Journal of Pervasive Computing and Communications,2020,16(5)
[14] Götz Thomas,Heidrich Peter. Early stage COVID-19 disease dynamics in Germany: models and parameter identification.[J]. Journal of mathematics in industry,2020,10(1)
[15] Liu Yiping. Sir and SIRS models under the influence of media reports [D]. Nanjing Normal University, 2007
[16] Zhu Yongfang. Analysis of SEIQR epidemic model with time delay and isolation [D]. Anhui University, 2012
[17] Li Haiyan, Wei Yuming, Peng Huaqin. SIS epidemic model with nonlinear incidence and Ornstein Uhlenbeck process [J]. Journal of Mianyang Normal University, 2019, 38 (11):5-13
[18] Shao Junjie , Yu Shixiong , Gao Jingjing, et al. Comparative analysis of early transmission characteristics of COVID-19 epidemic in Shandong Province of China and South Korea based on SEIR model [J / OL]. Journal of central China Normal University (NATURAL SCIENCE EDITION): 1-8 [2020-11-21]
[19] Cao Yu, Jing Yuanwei, Yuan Feng, et al. Analysis of SIRS model with nonlinear infection rate on complex networks [J]. Journal of Northeast University (NATURAL SCIENCE EDITION), 2012,33 (01): 17-20
[20] Gautam Yogesh. Transfer Learning for COVID-19 cases and deaths forecast using LSTM network[J]. ISA Transactions,2021(prepublish)