Prediction of COVID-19 pandemic based on data-driven model

Qiyang Sun

Hampton Roads Academy, United States

Keywords: COVID-19 prediction; data-driven; machine learning; SEIR; LSTM Networks

Abstract: In order to predict the COVID-19 outbreak, several epidemiological models are used around the world to predict the number of infections and mortality. An accurate predictive model is essential to take appropriate action. According to the latest population migration data in Wuhan, China around January 23, 2020. A data-driven epidemic and prediction method is proposed, and COVID-19 epidemiological data is imported into the Susceptible Exposure Infection Elimination (SEIR) model to derive the epidemic curve. The logistic regression method is used to predict the spread of the virus over time. For further comparison and verification, the LSTM time series model is established to study the trend of virus spread and predict the spread of COVID-19. The epidemic in China should reach its peak in late February and gradually decline by the end of April. The machine learning prediction method adopted in this paper confirms this result and has certain reference significance.

1. Introduction

The new pneumonia began in Wuhan, Hubei Province, China in December 2019. It has infected 114,350 cases in 107 countries/regions on March 10, 2020, resulting in 4,023 deaths. The virus that causes pneumonia has been named the new coronavirus 2019 (2019-ncov) and was renamed COVID-19 according to an article published by the World Health Organization (WHO) on February 11. It is highly contagious, has caused tens of thousands of cases and caused great panic around the world. Even if prevention and control measures such as home quarantine and wearing masks are adopted across the country, the number of confirmed cases is still on the rise. At this stage there are still many unclear and awaiting investigations [1].

To contain the epidemic, China implemented an unprecedented intervention strategy on January 23, 2020 (6). The city was quarantined, national holidays were extended, travel was strictly restricted and public gatherings were introduced, public places were closed and strict temperature monitoring was implemented nationwide. These control measures have caused great impacts to disrupt China's social and economic structure and the world. Therefore, it is important to evaluate the effectiveness of these control measures to benefit the development of the epidemic. According to the latest population migration data in Wuhan, China around January 23, 2020. The susceptible exposure and infection elimination (SEIR) model is used to predict the progress of the epidemic, combined with the LSTM time series model to study the trend of virus transmission and predict the spread of COVID-19 [2][6].

2. Data Sources and data analysis

The COVID-19 epidemic data used in this article comes from the People’s Daily, CCTV News, Hubei Daily and other websites. Including epidemic prevention and control measures, the number of cases, the number of people entering and leaving Hong Kong daily, air and road traffic travel information, etc.

Based on the existing literature, this paper gives a heat map of the regional distribution of COVID-19 in China as of March 9, as shown in Figure 1. Different colors indicate the number of confirmed cases in different provinces. Among them, the darker the color, the more serious the spread of the epidemic. The infected areas are mainly distributed in Hubei Province and spread to surrounding provinces. Wuhan is the capital of Hubei Province, with 49,965 infections as of March 9. Dark colors

show more cases of infection. In order to control the spread of the epidemic, the Chinese government has taken measures such as closing the city, isolating people returning from Wuhan, or contacting people from Wuhan to cut off the source of infection [3].

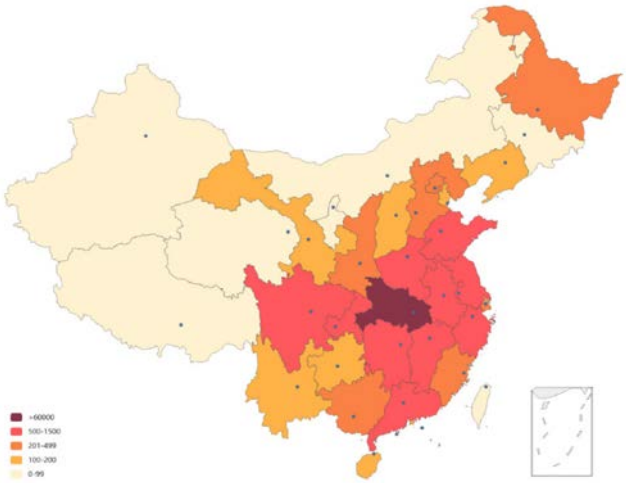

Fig 1. China epidemic of COVID-19

3. Data-driven model

3.1 SEIR model

The basic mathematical model of infectious diseases studies the spread speed, space range, transmission route, dynamic mechanism and other issues of infectious diseases to guide the effective prevention and control of infectious diseases. S, E, I, and R are the four types of people in the epidemic range of infectious diseases: susceptible, exposed, infected, and recovered. Susceptible persons represent those who have not gotten the disease, but lack immunity and are susceptible to infection after contact with infected persons [4]. An exposed person represents a person who has been in contact with an infected person but is temporarily unable to infect other people. It is suitable for infectious diseases with a long incubation period. An infected person represents a person who has contracted an infectious disease and can transmit it to a class S member, turning it into a class E or class I member. Recovered persons represent people who have been isolated or who have become immune due to illness. If the immunization period is limited, members of the R class can revert to the S class. The SEIR infectious disease model is established by defining the four types of population, and the model is established as follows in combination with the data. The basic model of SEIR is as follows.

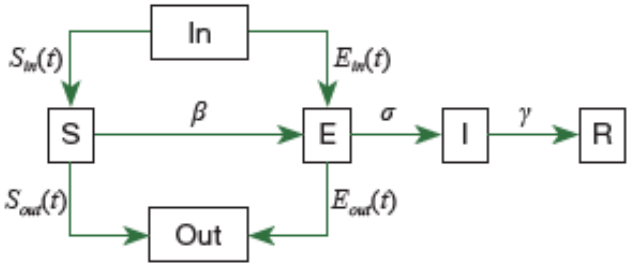

Fig.2 The basic model of SEIR

Step1: By establishing differential equations:

$$\frac{dS}{dt} = -\frac{r_1\beta IS}{N} - \frac{r_2\beta_2 ES}{N} \tag{1}$$

$$\frac{dE}{dt} = \frac{r_1\beta_1 IS}{N} + \frac{r_2\beta_2 ES}{N} - \alpha E \tag{2}$$

$$\frac{dI}{dt} = \alpha E - \gamma I \tag{3}$$

$$\frac{dR}{dt} = \gamma I \tag{4}$$

Step2: Transformation iteration.

$$\frac{dS}{dt} = -\frac{r_1 \beta I S}{N} - \frac{r_2 \beta_2 E S}{N} \tag{5}$$

Step3: Integrate the left and right sides.

$$\int_{n-1}^{n} \frac{dS}{dt} dt = \int_{n-1}^{n} \left(-\frac{r_1 \beta I S}{N} - \frac{r_2 \beta_2 E S}{N} \right) dt \tag{6}$$

Step4: Use rectangular formula.

$$\left(\int_a^b f(x) dx \approx (b-a) f(a) \right) \tag{7}$$

$$S_n - S_{n-1} = -\frac{r_1 \beta_1 I_{n-1} S_{n-1}}{N} - \frac{r_2 \beta_2 E_{n-1} S_{n-1}}{N} \tag{8}$$

Step5: Results collation

$$S_n = S_{n-1} - \frac{r_1 \beta_1 I_{n-1} S_{n-1}}{N} - \frac{r_2 \beta_2 E_{n-1} S_{n-1}}{N} \tag{9}$$

$$E_n = E_{n-1} + \frac{r_1 \beta_1 I_{n-1} S'_{n-1}}{N} + \frac{r_2 \beta_2 E_{n-1} \$_{n-1}}{N} - \alpha E_{n-1} \tag{10}$$

$$I_n = I_{n-1} + \alpha E_{n-1} - \gamma I_{n-1} \tag{11}$$

$$R_n = R_{n-1} + \gamma I_{n-1} \tag{12}$$

The explanation and meaning of each variable parameter are shown in Table 1.

Table 1. Meaning of each variable parameter

Variable	Meaning
S	Number of susceptible
E	Number of lurkers
I	Number of people infected
R	Number of recovered people
r_1	The average number of people in contact with each infected person per day
r_2	The average number of people lurking in daily contact
β_1	Probability of susceptible person being infected by infected person
β_2	Probability of susceptible persons being infected by latent persons
α	Probability of a latent person turning into an infected person
γ	Probability of Recovery
N	Total people

Before applying the SEIR model, its parameters need to be estimated. The parameters are β, σ, and γ, where β is the person who comes into contact with the infected person every day (k) and the probability of transmission at exposure (b) (β= kb) and σ are the incubation rates of individuals with symptoms (average duration of incubation is 1/σ). γ is the average recovery rate or death rate of the infected population. Using epidemiological data from Hubei, we simulated the tilted SEIR model to determine the probability of transmission (b) which is used to derive β and the probability of recovery or death (γ).

3.2 Long-Short-Term-Memory (LSTM) model

Long short-term memory network (LSTM) is a kind of time recurrent neural network, which is designed on the basis of recurrent neural network and belongs to a kind of time recurrent neural network (RNN). LSTM is suitable for processing and predicting important events with very long intervals and delays in time series [4].

The memory block(Fig.3) in the figure below mainly contains three gates (forget gate, input gate, output gate) and one memory cell (cell). The horizontal line above the box is called the cell state, and it is like a conveyor belt that can control the transfer of information to the next moment. In other words, the state of the top line of the cell represents long-term memory, while the bottom line represents working memory or short-term memory. The most important thing in LSTM is Forget gate, followed by Input gate, and last is Output gate. You can see that there are four small yellow boxes in the middle cell. Each small yellow box represents a feedforward network layer, and num_units is the number of hidden neurons in this layer. The activation function of 1, 2, 4 is sigmoid, and the activation function of the third is tanh.

The internal structure of LSTM is shown in Figure 4.

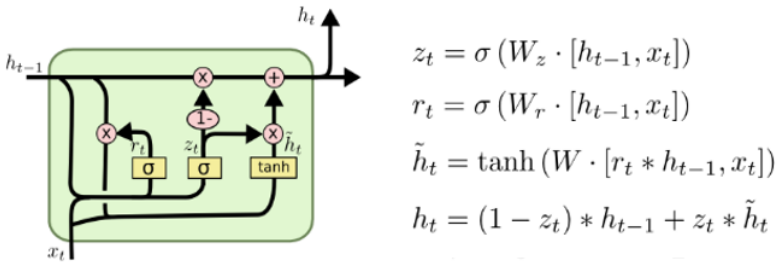

Fig.3 Memory block of LSTM

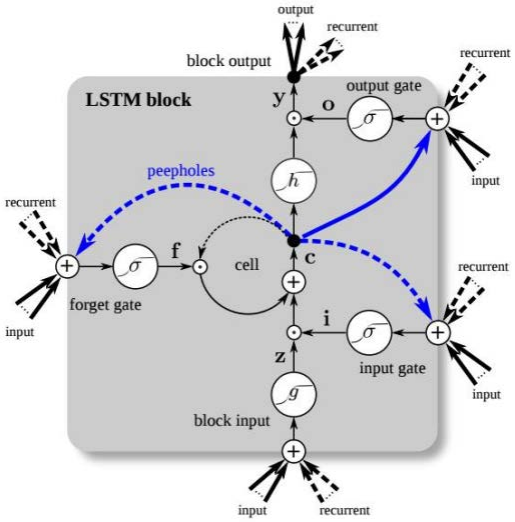

Fig.4 The internal structure of LSTM

3.3 Logistic regression model

The dependent variable of logistic regression can be two-category or multi-category, but two-category is more commonly used and easier to interpret. Therefore, the most commonly used in

practice is the two-category Logistic regression. Logistic regression is mainly used in epidemiology [5].

The first is to look for risk factors for a certain disease. The second is to predict the probability of a certain disease or a certain situation based on the model under different independent variables. The third is that it is actually somewhat similar to prediction. It is also based on the model to determine the probability that a person belongs to a certain disease or a certain situation, that is, to see how likely the person is to belong to a certain disease.

The graph of the relationship between the probability and the independent variable of the binary classification problem is often an S-shaped curve, which is realized by the Sigmoid function.

$$g(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

Construct the prediction function as:

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}} \tag{14}$$

After constructing the loss function Cost function and J function as follows, they are derived based on the maximum likelihood estimation. The detailed formula can refer to the relevant literature.

4. Epidemic prediction based on data-driven model

To evaluate the difference between the predicted and actual values of the case and find the gradient descent to reduce the gap, the loss function of this model was set to mean squared error (MSE) according to the following equation:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \tag{15}$$

Due to the small data set in this paper, a simpler network structure is adopted to prevent over-fitting by using LSTM neural network and fully connected layers. For neural network parameter selection,

The model chooses the ADAM optimizer, uses a training round designed for 1000 rounds, the batch size is 1, and the loss function is selected in the above MSE.

The LSTM network structure used in this article. The input is fixed time step data. The model uses three days of new input as input, and the input dimension is (3,1). The hidden layer receives input data from the input layer and enters the middle layer of the LSTM unit, which is set to 20.

The LSTM model is based on RNN training. RNN contains COVID-19 epidemiological parameters and 2003 SARS epidemic statistics, such as spread and latency. Probability of recovery or death and contact number. The LSTM model predicts that new infections will reach a peak on February 4th, as shown in Figure 5. In addition, this paper also compares the predictions of the SEIR model and the LSTM model with the actual data, as shown in Figure 6. Both SEIR and LSTM models predict that between February 4 and 7, the number of daily infections will reach a peak of 4,000. The SEIR model also predicted several smaller new peaks of infection in mid-to-late February.

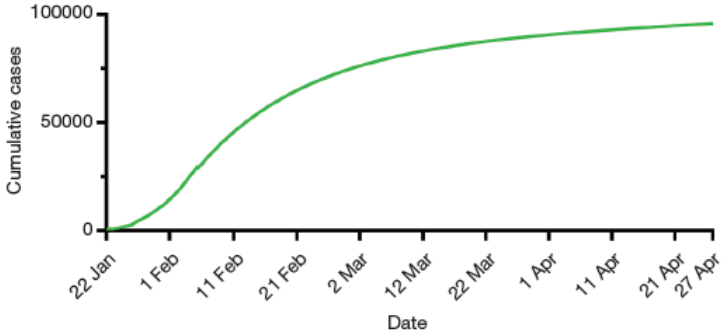

Fig.5 LSTM-predicted cumulative number of COVID-19 cases in China

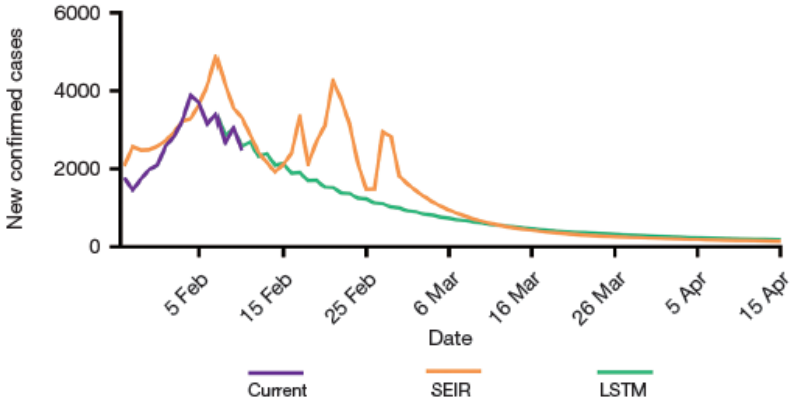

Fig.6 COVID-19 cases according actual data (purple), SEIR-model (orange) and LSTM model (green)

Despite the rapid spread of COVID-19, we are aware that the transmission process occurs in a limited number of people. By implementing various containment measures, the transmission rate drops above a certain threshold value. This process is fully in line with the logistic curve. In this article, the dates are divided into two levels according to the different months. The first phase is from January 13th to 31st. Using the logic model, the average error is -1.6%, which confirms the reliability of the model. The confirmed cases in other provinces and Hubei were selected for comparison, and it turned out to coincide in early January, while the number of confirmed cases in Hubei Province rose at the end of the same month. This trend is due to the province's lack of drug resources and data are limited to the early stages of infection. Later, medical resources from other provinces supported Hubei Province to improve diagnostic efficiency and provide more accurate information. Our model is a good representation of this January infection process, as shown in Figure 7.

The virus has been spreading at a rate of r = 1.0 since February. During this incubation period, the number of infections rose rapidly. In particular, on February 12, the number of infections increased by 14,840 due to changed diagnostic criteria, bringing the total number of cases to over 60,000. In such a grave situation, Hubei Province blocked townships and roads. Political and medical circles have worked hard to continue containment efforts. Compared to Hubei Province, the disease situation in other provinces and cities across the country remained relatively stable in February: the confirmed cases have generally not increased since February 16, 2020. Our model shows that the number of confirmed cases in Hubei Province continued to decline at the end of February with a relative error of -9.31% and the critical point occurred on March 1st.

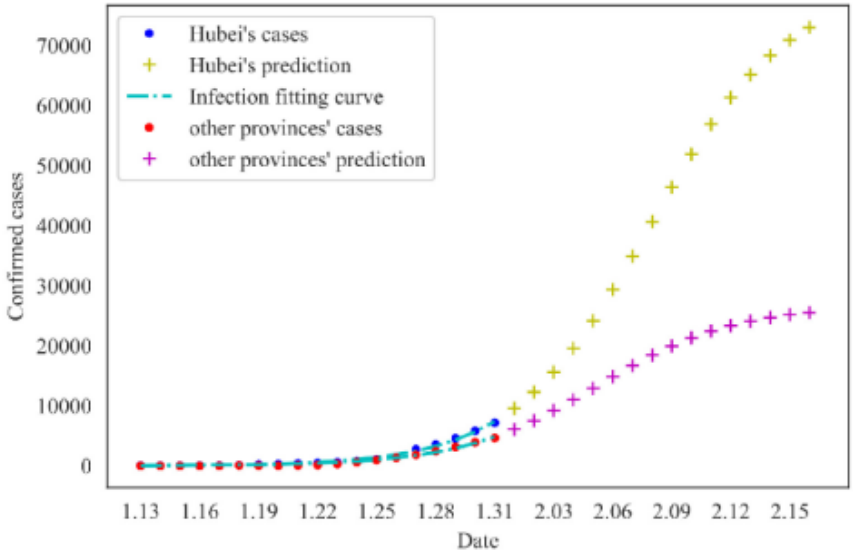

Fig.7 Real cases, fitting curves and prediction for the confirmed cases in January

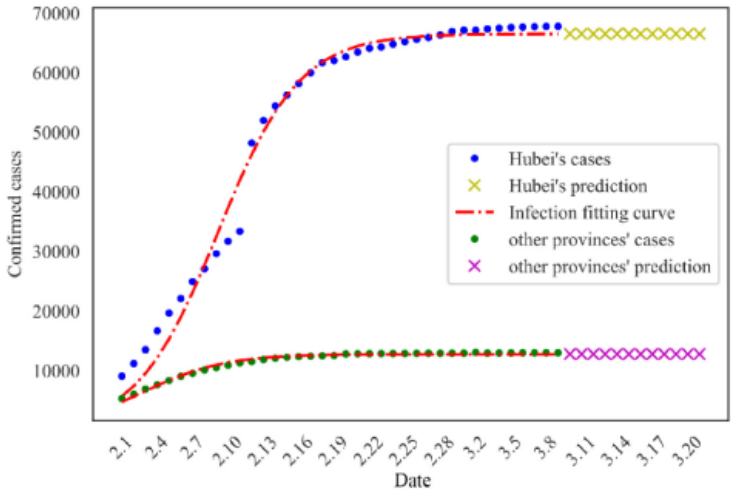

Fig.8 Real cases, fitting curves and prediction for the confirmed cases in February to March

5. Conclusion

In this article, epidemiological data on COVID-19 was imported into the Susceptible Exposure and Elimination of Infection (SEIR) model to derive the epidemic curve. The logistic regression method is used to predict the spread of the virus over time, and for further comparison and verification, the LSTM time series model is used to study the spread of the virus and predict the spread of COVID-19.

Based on the dynamic SEIR model, the peak and extent of the COVID-19 epidemic can be effectively predicted. The control measures put in place on January 23rd are expected to reduce the scale of the COVID-19 epidemic in China, and the policy of strict surveillance and early detection should be maintained until the end of April 2020. The bottom line shows that the epidemic can be effectively combated from March 1, 2020, which corresponds to the real situation. The spread of COVID-19 in other provinces and cities in China is relatively weak, and strict isolation may have contributed to this outcome. This article can provide valuable reference information for healthcare workers and health decision makers as they experience the new coronavirus disease outbreak.

References

[1] Lipsitch M, Cohen T, Cooper B, Transmission dynamics and control of severe acute respiratory syndrome[J]. Sci 2003(300): 1966-1970.

[2] Read JM, Bridgen JRE, Cummings DAT, Ho A, Jewell CP, Novel coronavirus 2019-nCoV: Eearly estimation of epidemiological parameters and epidemic predictions[J]. medRxiv ,2020(1): 20018549.

[3] Huang Rui,Liu Miao,Ding Yongmei. Spatial-temporal distribution of COVID-19 in China and its prediction: A data-driven modeling analysis [J]. Journal of infection in developing countries,2020,14(3).

[4] Yang Zifeng,Zeng Zhiqi,Wang Ke,. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions [J]. Journal of thoracic disease,2020,12(3).

[5] Singh Ram Kumar,Rani Meenu,Bhagavathula Akshaya Srikanth,. Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model[J]. JMIR public health and surveillance,2020,6(2).

[6] Zhao S, Lin QY, Ran JJ, Salihu S. Musa, Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak[J]. Int J Infect Dis. 2020(92): 214-217.