Analysis and Prediction of Epidemic Infectious Diseases Based on SEIDR Model – A Case Study of COVID-19 in New York

Zhigang Xu (imdr@163.com)
Hubei University of Technology https://orcid.org/0000-0002-6304-8568
Huiling Xia
Hubei University of Technology

Research

Keywords: SEIDR model, COVID-19, disease-related mortality, latent morbidity rate, turning point

Posted Date: January 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-156792/v1

Analysis and Prediction of Epidemic Infectious Diseases Based on SEIDR Model -- A Case Study of COVID-19 in New York

Zhigang Xu[1]· Huiling Xia[1]

Abstract

In the epidemic prevention and control of infectious diseases, improper prevention and control can easily lead to a large-scale epidemic. However, the epidemic of diseases follows certain rules, so it is very necessary to simulate the spread of infectious diseases, which can provide reference for the formulation of prevention and control measures. This paper proposes a SEIDR model for analyzing and predicting epidemic infectious diseases. Taking the development situation of COVID-19 in New York City as an example, firstly, the SEIDR model proposed in this paper was compared with the traditional SIR model, and it was found that the SEIDR model was better than the SIR model. Then the SEIDR model and the L-BFGS optimization method were used to fit the early transmission data of COVID-19 in New York City, and important parameters such as infection rate, latent morbidity rate, disease-related mortality and recovery rate were obtained. Moreover, the value of basic regeneration number R_0 between 4.0 and 4.6 proved that the situation of COVID-19 in New York City was relatively serious. Finally, these parameters were used to predict the future development of COVID-19 in New York City, and the turning point of COVID-19 in New York City was found. However, even if the turning point be reached, the development trend of COVID-19 will not be controlled in the short term. Data verification shows that the SEIDR model established in this paper can effectively provide a scientific quantitative index for governments in the prevention and control of COVID-19 and other epidemic infectious diseases.

Keywords SEIDR model· COVID-19· disease-related mortality· latent morbidity rate· turning point

1 INTRODUCTION

Epidemiology [1][3][10] is a discipline that studies the distribution and determinants of diseases and health conditions in a specific population. It mainly studies measures and strategies for preventing and treating diseases and promoting health. Traditionally, the research objects of epidemiology [25][29][30] are various infectious diseases. Infectious diseases, according to the infectious strength of their infectivity, the speed of transmission, the difficulty of the route of transmission, the extent of the population's vulnerability to infection can be divided into the following three categories: ①High infection, fast transmission, easing to achieve transmission, and susceptible population, such as cholera, plague, etc.; ②infectiousness, transmission speed, transmission route, and population susceptibility is second to the first category of infectious diseases, For example: AIDS, rabies, measles, schistosomiasis, malaria, etc.; ③According to the scope of the possible occurrence and epidemic of the infectious disease, the infectious diseases monitored and managed by the disease monitoring center and laboratories are less harmful than the first category of infectious diseases ang the second category of infectious diseases, such as common influenza, leprosy, infectious diarrhea and infectious diarrheal diseases other than bacterial ones.

Coronaviruses belong to the second category of epidemic infectious diseases mentioned above. Coronaviruses belong to the order of nesting viruses, the family of coronaviruses, and the genus coronaviruses in the systematic classification. They are a large group of viruses that exist in nature. Coronavirus only infects vertebrates, such as humans, genus, cats, pigs, chickens, cattle, dogs, and poultry.

The novel coronavirus pneumonia (Corona Virus Disease 2019, COVID-19 referred to as "New Coronary Pneumonia", named by the World Health Organization as "Coronavirus Disease 2019") is the seventh coronavirus known to infect humans. Self-isolation, reducing

[1] School of Computer Science, Hubei University of Technology, 28 Nanli Road, Wuhan 430068, Hubei, China, imdr@163.com

interpersonal communication, wearing masks, reducing going out utilizing transport, washing hands frequently, and maintaining an optimistic mood to improve self-

the rate of disease transmission. At the same time, with effective treatment, COVID-19 has been already under control in China. However, the number of confirmed cases worldwide has exceeded 63 million so far, and the cumulative deaths have exceeded 1.46 million. The United States, India, Brazil, France, and Russia are among the countries hardest hit. In the United States, there has even been a daily increase of 200,000 confirmed cases. COVID-19 has caused varying degrees of impact on all countries, including the economy and people's livelihood. Therefore, it is urgent to analyze and predict COVID-19. Only by understanding the development trend of COVID-19 can we make a more effective response policy.

The United States is currently the country with the largest number of confirmed COVID-19 cases in the world, so this paper takes the COVID-19 data from New York State as an example to analyze and predict COVID-19. Since the data has a certain lag, there will be slight errors. The classic models for the spread of this infectious disease include SI model, SIS model, SIR model, SIRS model, SEIR model, etc. The idea of these models is to divide the population into susceptible, infected, recovered, and exposed and other groups, and use the transmission mechanism from one group to another group or multiple groups to establish a mathematical model to explain the spread of COVID-19. These models and their methods are used to study infectious diseases. Such as AIDS, common flu, acute infectious diseases and malignant infectious diseases with incubation period and so on. In the on-site prevention and control of infectious diseases, if the prevention and control measures are not taken properly, it is very likely to cause a large-scale outbreak of the epidemic. Therefore, domestic and foreign countries pay more attention to the research and development of on-site prevention and control technologies and systems for infectious disease epidemics. Moreover, according to news reports, there may be a second new type of coronavirus pneumonia at the end of 2020, so it is particularly necessary to study the mathematical model of the COVID-19 infectious disease.

immunity can minimize the cross-infection rate. These effective preventions, controls, and governances have helped to block the crazy spread of the virus and reduce

In the mathematical model of epidemic virus, Ma Yanli, Chu Zhengqing and others studied a dynamic model of SI infectious diseases with saturated contact rate, and obtained an important threshold to measure whether the infectious disease is prevalent, that is, the basic reproduction value. Using the SIJR model, Professors Liu Chang and Ding Hongguang et al. from Fudan University analyzed the key factors such as the spread rate of the SARS [6][27][28] virus and the basic reproduction value in 2003, and found the development trend of SARS in 2003. Ni Shunjiang et al. established a dynamic model of infectious diseases based on the complex network theory. Compared with traditional epidemiological studies, this model can better describe the transmission process of infectious diseases in real heterogeneous large-scale social networks in the analysis of uniformly mixed populations. Zhu Ling et al. proposed the gradual behavior of a logistic SIR infectious disease model in which the natural mortality rate is randomly disturbed by external noise, The solution of the model surrounds the solution of the deterministic model for a long time to produce random oscillations, and the random oscillation is used as the standard to judge Whether the disease will be epidemic. Chen Qizhi applied the random compartment model to atypical pneumonia, estimated the parameters of the model, and compared the changes of the parameters in Beijing and Hong Kong. Yan Yue, Chen Yu et al. proposed a type of infectious disease dynamic model of time-delay dynamic systems to simulate the development of the epidemic situation and analyze the prevention and control measures of governments at all levels. David, Pejman et al. introduced time series on the basis of SEIR model, mainly aiming at the transmission dynamics from regional synchronous fluctuations to complex spatial incoherent childhood infectious diseases, and took measles as an example to simulate its transmission. In the simulation system, the Los Almos National Laboratory in the United States developed an urban EpiSimS(epidemic simulation system) based on the social transportation network model, which is used to simulate the spread of infectious diseases and provides reference for the formulation of prevention

and control measures. The Virginia Tech team established the FluCaster system to evaluate and predict influenza, and it can predict the probability of influenza infection by entering information such as personal location, time, and preventive measures. In addition, The Virus Tracker simulation platform built by the Institute of Biocomplexity at Virginia Tech was used to simulate the path and process of virus transmission, and to demonstrate the impact of vaccination on the spread of virus epidemics. The establishment of these system platforms provides necessary data support for the formulation of on-site control measures for infectious diseases. In this regard, Chinese scholars have also carried out related research work. For example, Yao Wei and others have established an on-site investigation support system for infectious disease emergencies, which provides nearly 200 types of infectious disease information that can be queried intelligently. Yang Qing et al. established an evolutionary model of infectious disease emergencies based on cellular automata, they established an evolutionary inference model by analyzing the characteristics of infectious disease emergencies and simulated the 2009 influenza A epidemic.

The above papers [13][16][22][23] are all papers doing retrospective studies on the characteristics of epidemic infectious diseases, and a small part of the papers or the content of the papers only predict the number of people diagnosed with a certain infectious disease in the future. This kind of prediction only includes a specific group of all groups in infectious diseases, which is still not specific enough for the overall trend, and does not take into account the differences in the natural state. It is not accurate enough to predict future trends through calculations. Inspired by these papers, this paper considers the impact of natural population factors on infectious diseases, that is natural birth rate and natural mortality rate. And compared with the traditional SIR model, this paper considers the SIR model of natural birth rate and natural mortality rate, the SEIR model with incubation period. Based on this, this paper proposes a SEIDR model for analyzing and predicting epidemic infectious diseases. Taking the situation of COVID-19 in New York City as an example, it verifies the development trend of COVID-19, and explains the superiority of the SEIDR model compared to the SIR model. The model can also be applied to the analysis and prediction of COVID-19 in other regions to evaluate the effectiveness of prevention and control measures of COVID-19 in that region, and provide quantitative indicators for the government's control of COVID-19 in the future.

2 RELATED WORK

2.1 Basic characteristics of COVID-19

COVID-19 [2][5][12][14] ,is mainly transmitted through respiratory droplets and close contact. In a relatively closed environment, if exposed to an environment full of high concentrations of aerosols for a long time, it may also be transmitted through aerosols. The main source of infection is the patients infected with COVID-19. Of course, asymptomatic infected persons may also become the source of infection. There is a certain incubation period after the patient is infected, so it is not easy to be diagnosed immediately. The degree of infection of susceptible varies from person to person, and will be affected by body immunity and age. Since COVID-19 has a certain incubation period, it is necessary to add the factor of the exposed in the traditional SIR model.

2.2 The traditional infectious disease model --SIR model

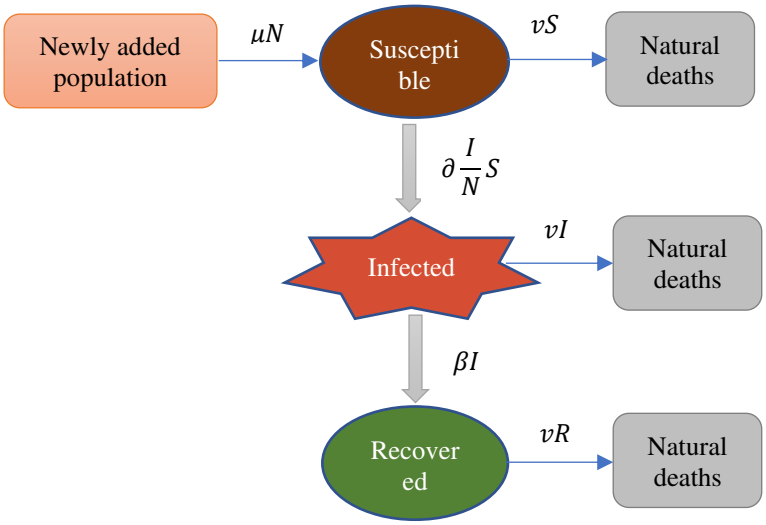
The situation of infectious diseases described by the traditional SIR model[7][9][21][24] can be described briefly: developing rapidly, containing antibodies after recovery and not being re-infected. SIR model and the corresponding differential Eq can be drawn, as shown in Fig. 1 and Eq (1).

In the Eq set (1), S represents the susceptible, I represent the infected, R represents the recovered. S, I, and R are all functions of time t. Taking into account the natural birth rate and the death rate, the total population of the region is

$$(\mu - v)N + N = S + I + R + (\mu - v)I + (\mu - v)R,$$

μ represents the natural birth rate, v represents the (except for deaths due to COVID-19). Newborn babies are susceptible by default. ∂ represents the probability that a susceptible person will be infected after contacting with an infected person. It is not that a susceptible person will be

infected as long as he has contacted an infected person, he just may be infected. Here, it is considered that every susceptible person has the same probability of contacting an infected person, that is ∂ represents the average number of people infected due to contact with susceptible people S per unit time; since the probability that a susceptible person can contact with an infected person is I/N the number of newly infected susceptible persons per day is $\partial IS/N$. In general, the recovery of diseases has a certain recovery cycle. For COVID-19, this paper assumes that β represents the rate from the infected person to the removed person, so the number of people removed is βI; because some diseases are fatal, the dead will not infect susceptible people after being properly handled, and will not be infected by infected people at the same time, it has no effect on the changes of the number of other groups, so in terms of the probability of infection, its significance is the same as that of restorers. Therefore, the removed people R at this time includes the number of deaths and the number of restorers. Each of these groups will have natural deaths, so changes of the number of people in each group will result in the corresponding number of natural deaths be subtracted.



descent, coordinate descent, Newton method, and quasi-Newton method) can solve the problem well, but they also have their own shortcomings.

- Since gradient descent is based on the gradient of the objective function, the convergence rate of the algorithm is linear, so when encountering ill-conditioned problems or when the problem scale is relatively large, the convergence rate is extremely slow. Therefore, the algorithm can be said to be almost unusable;

Compared with gradient descent, although the gradient of objective function doesn't need to be calculated, the convergence speed of coordinate gradient descent is still very slow, which results in a very limited range of use.

Newton's method is based on the Hessian matrix of the objective function. It has a faster convergence speed and fewer iterations, especially the velocity near the optimal value is quadratic. However, if the Hessian matrix of the objective function is too dense, the calculation for each iteration will increase. Because the inverse of the Hessian matrix of the objective function needs to be calculated in each iteration, so when the problem scale is large, it will cause a huge calculation, and the memory space is also needed to store the matrix;

2.3 L-BFGS optimization method

When the scale of the optimization problem is relatively small, the basic optimization algorithms (including gradient

Fig. 1 SIR Model

$$\begin{cases} \dfrac{dS}{dt} = \mu N - \partial \dfrac{S}{N} I - vS \\ \dfrac{dI}{dt} = \partial \dfrac{S}{N} I - vI - \beta I \\ \dfrac{dR}{dt} = \beta I - vR \end{cases} \tag{1}$$

The quasi-Newton method introduces the approximate matrix of the Hessian matrix on the basis of the Newton method, which avoids the problem of calculating the inverse of the Hessian matrix in each iteration. The convergence speed of the quasi-Newton method is slightly lower than that of the Newton method, which is between gradient descent and Newton's method. Similar to Newton's method, when the scale of the optimization problem of the quasi-Newton method is large, the approximate matrix will be very dense, and the calculation and storage of the matrix will consume a lot of memory overhead. Moreover, Newton's method cannot guarantee that the Hessian matrix of the objective function is always positive definite in each iteration. If the Hessian matrix is not positive definite, it will directly lead to the deviation of the optimization direction, and the Newton's method will be useless; The reason why quasi-Newton method is better than Newton method is that quasi-Newton method uses the inverse of Hessian matrix to instead of Hessian matrix. Although each iteration cannot guarantee that it gets the optimal direction, the approximation matrix must be positive definite, that is, the algorithm can always be Searching in the direction of the optimal value.

It can be seen from the above that when the scale of the optimization problem is large, none of the above algorithms can solve the problem well. However, in practical application, many problems are ill-conditioned, which will lead to the failure of the gradient-based algorithm. Even if the algorithm is iterated thousands of times, it will not necessarily get a good convergence result; The large amount of data will consume a lot of memory to calculate and save the matrix when using quasi-Newton method and Newton method. So in general, these basic optimization algorithms (including gradient descent, coordinate descent method, Newton method, and quasi-Newton method) are not applicable when solving large-scale optimization problems, so other optimization algorithms should be considered.

The full name of L-BFGS [4][11][15][18] is Limit-memory BGFS. BFGS is the abbreviation of the names of the four inventors: Broyden, Fletcher, Goldfarb, and Shanno. L-BFGS is still considered the best quasi-Newton algorithm for more than 40 years since it was invented. L-BFGS is the best method to solve unconstrained nonlinear programming problems, which has fast convergence speed and low memory overhead. Compared with the traditional stochastic gradient descent (SGD), L-BFGS converges much faster than SGD if the same amount of data is processed.

The iteration of BFGS is as follows:

$$B_{i+1} = V_i^T B_i V_i + \rho_i s_i s_i^T \tag{2}$$

Among them,

$$t_i = \nabla f\left(X_{i+1}\right) - \nabla f\left(X_i\right) \tag{3}$$

$$s_i = X_{i+1} - X_i \tag{4}$$

$$\rho_i = \frac{1}{t_i^T s_i} \tag{5}$$

$$V_i = I - \rho_i t_i s_i^T \tag{6}$$

Substituting Eq (3), (4), (5) and (6) into Eq (2), we can get

$$B_{i+1} = B_i + \frac{1}{s_i^T}\left(1 + \frac{t_i^T B_i t_i}{s_i^T t_i}\right) s_i s_i^T - \frac{1}{s_i^T}\left(s_i t_i^T B_i + B_i t_i s_i^T\right) \tag{7}$$

In BFGS, suppose we need to use an N*N matrix B, suppose N is 50000, and only use the double type to store the matrix, then the required memory size can be calculated as follows:

$$\frac{The\ number\ of\ bytes\ in\ matrix\ B}{The\ number\ of\ bytes\ in\ 1\text{GB}} = \frac{50000*50000*8}{1024*1024*1024} \approx 18.66GB$$

The data volume of 18.66GB is quite large. Since matrix B is symmetric, the memory usage can be reduced

by half, but this is only the memory occupied by 50,000 pieces of data. What's more, in actual work, 50,000 pieces of data can only be regarded as small and medium-sized data, so the optimization of BFGS is very necessary to reduce the memory overhead in the iterative process.

L-BFGS[19][20][26] is an optimized method relative to BFGS, which has been greatly improved. The basic idea of the L-BFGS algorithm is: instead of saving the matrix B, it saves and uses the curvature information s_i and $t_i$of the last m iterations to construct an approximate matrix of the Hessian matrix. That is, when each iteration is completed, the oldest curvature information will be discarded, and the latest curvature information will be saved, so as to ensure that the saved curvature information comes from the most recent m iterations. We can refer to Formula (8):

From formula (2), it can be deduced that:

$$B_i = V_{i-1}^T B_{i-1} V_{i-1} + \rho_{i-1} s_{i-1} s_{i-1}^T \tag{8}$$

Substituting formula (8) into formula (2), then:

$$\begin{aligned} B_{i+1} &= V_i^T B_i V_i + \rho_i s_i s_i^T \\ &= V_i^T \left(V_{i-1}^T B_{i-1} V_{i-1} + \rho_{i-1} s_{i-1} s_{i-1}^T \right) V_i + \rho_i s_i s_i^T \\ &= V_i^T V_{i-1}^T B_{i-1} V_{i-1} V_i + V_i^T \rho_{i-1} s_{i-1} s_{i-1}^T V_i + \rho_i s_i s_i^T \end{aligned} \tag{9}$$

Joint formula (2)-(9), after m iterations, we can deduce that:

$$B_i = \left(V_{i-1}^T V_{i-2}^T \cdots V_{i-m}^T \right) B_i^0 \left(V_{i-m}^T V^T \right.$$
$$+ \rho_{i-m} \left(V_{i-1}^T V_{i-2}^T \cdots V_i^T \right.$$
$$+ \cdots \rho_{i-1} s_{i-1} s_{i-1}^T$$

3 MODEL ANALYSIS

3.1 Problem Description

As we all know, the United States has the most people infected with COVID-19 in the world. In order to have a more comprehensive understanding of the impact of COVID-19, as New York is the largest city in the United States, this paper takes New York City in the United States as an example to analyze and predict the situation of infection, prevention and control. Fig.2 shows the population distribution diagram.

Suppose that the circle represents New York City, where infected people, susceptible people, exposed people with COVID-19, recovered people, and people died of illness are distributed in different areas. The unit of analysis and prediction in this paper is in days. Since the average number of natural birth rate and natural mortality rate per day is very small, it is not shown in our hypothetical circle. The question mark in the figure represents the outcome we need to predict, that is, over time, how will the numbers of people who are infected, susceptible, exposed and recovered with COVID-19, and those who died due to illness be distributed int the future.

3.2 Proposed SEIDR Model

The basic reproduction value can measure the severity of an infectious disease. The basic reproduction [illegible] average number of people that [illegible] environment where there [illegible] tervention. In short, [illegible] ct on average.

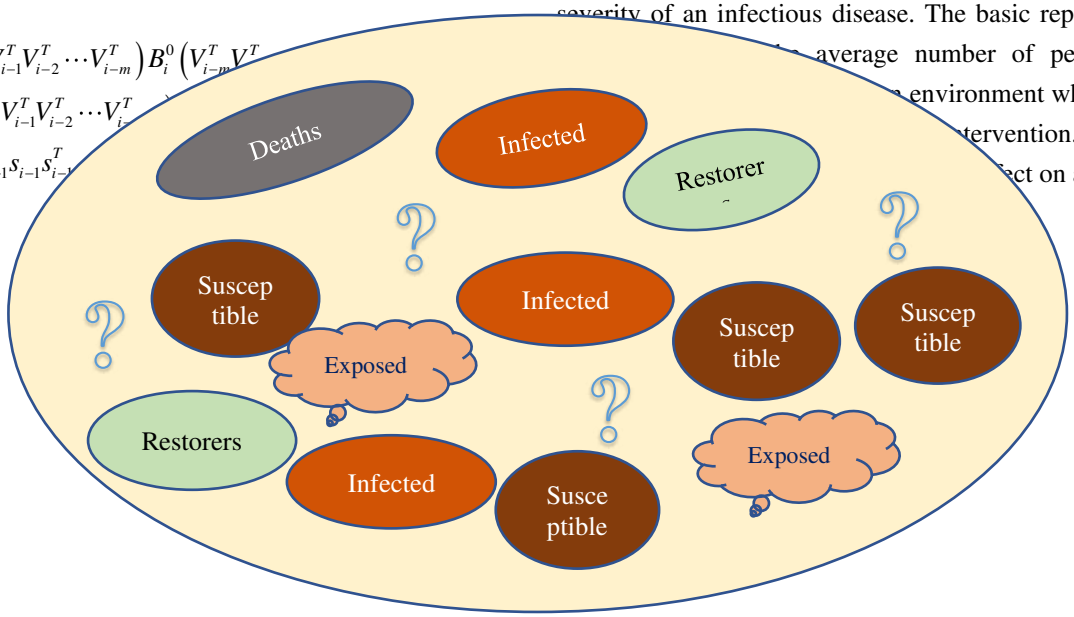


Fig. 2 Population Distribution Map

Under normal circumstances, $R_0 > 1$,otherwise the disease will not spread and will be eliminated in evolution. According to the data queried, the larger the value of R_0 , the more serious the situation at the peak of the epidemic.

This paper defines [illegible] COVID-19 in New York City as $R_0 = \frac{\mu N}{\varepsilon + \beta}$. Table 1 lists the values of R0 of several types of infectious diseases for comparison:

Table1 Infectious diseases and their basic regeneration values

[illegible]	R_0
Influenz[illegible]09	1.4-3.6
[illegible]vere acute respiratory syndrome SARS	2.2-3.6
COVID-19	1.4-5.5

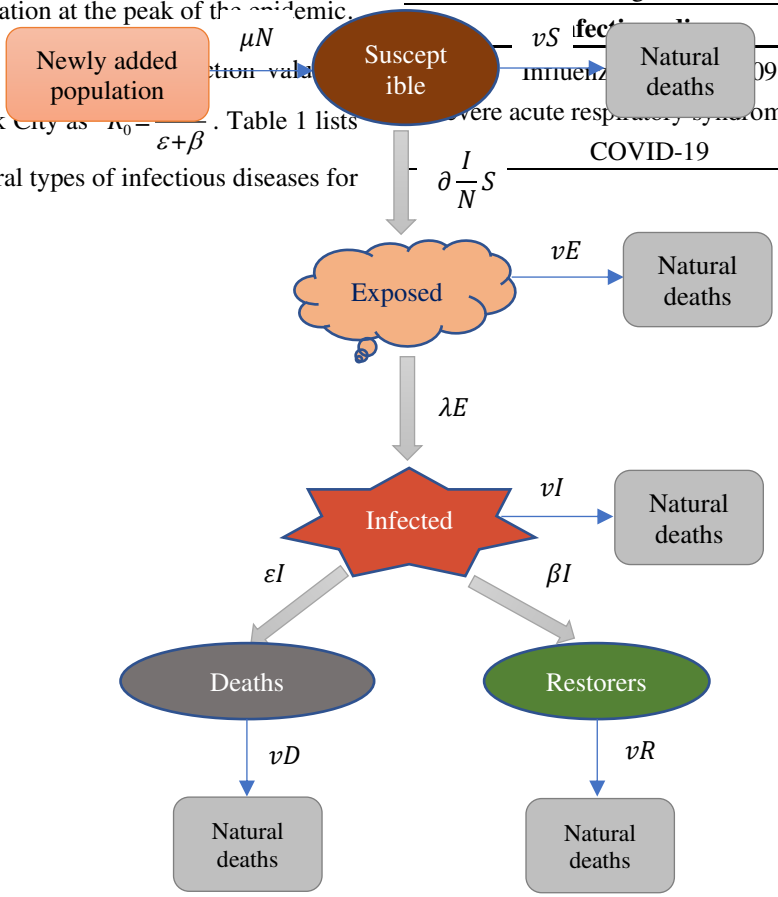


Fig. 3 SEIDR Model

$$\begin{cases} \dfrac{dS}{dt} = \mu N - \partial \dfrac{S}{N} I - vS \\ \dfrac{dE}{dt} = \partial \dfrac{S}{N} I - vE - \lambda E \\ \dfrac{dI}{dt} = \lambda E - vI - \beta I - \varepsilon I \\ \dfrac{dD}{dt} = \varepsilon I - vD \\ \dfrac{dR}{dt} = \beta I - vR \end{cases} \tag{11}$$

In the traditional SIR model of infectious diseases, only three groups: susceptible persons, infected persons, and removed persons are considered. For details, please refer to Fig1 and the system of Eq (1). It can't measure the exposed people with viruses well who have been infected but haven't been found. In order to better analyze and predict the situation of COVID-19 in New York City, the following assumptions are made:

- ∂ - The spread at this time is in a closed system, that is, the entire New York City, and everyone has the same probability of being infected;
- λ - Since COVID-19 has an incubation period, the incubation period must be considered for transmission. Therefore, introducing λ to represent the latent morbidity rate of COVID-19. It can be seen that the number of patients in the incubation period is λE ;
- ε - Because part of people infected with COVID-19 will be cured or die at any time. At this time, in order to distinguish the rate of being died due to illness and the rate of recovering due to treatment, subdividing the removed people of traditional algorithms into D who died of illness and R the who recovered R. At this time, the rate of being died due to illness ε is introduced, and the number of deaths due to disease is εI ;
- β - It's important to note that the β is not the removal rate, but the recovery rate.

This paper analyzes and forecasts in days. For example, the infection rate ∂ can be understood as follows: the probability that a susceptible person is in contact with an infected person today and he will be found to be a suspected or confirmed case tomorrow; Similarly, the fitting of the latent morbidity rate λ, disease-related mortality ε, the recovery rate β are also measured in days.

Under the above assumptions, the traditional SIR is modified, and the SEIDR model is obtained, as shown in Fig3 and Eq (11).

In this paper, since the values of each group are all numerical data that change with time t, the loss function is set as the sum of the mean square error of the predicted value and the true value of each group every day. Since the total population is $(1+\mu-v)N$, when select a day as the unit, the total population changes very little, and the number of susceptible persons can be calculated by the number of other groups, so the loss function defined at this time need to subtract the mean square error of the predicted value and the true value of the number of infected people S, that is:

$$Loss_{count} = \min_{\partial,\lambda,\varepsilon,\beta} \frac{1}{len(E(t))} \begin{Bmatrix} \sum_{t=1}^{T} (loss\left(E(t), \hat{E(t)}\right) \\ +loss\left(I(t), \hat{I(t)}\right) \\ +loss\left(D(t), \hat{D(t)}\right) \\ +loss\left(R(t), \hat{R(t)}\right)) \end{Bmatrix} \tag{12}$$

Among them,

$$loss\left(E(t), \hat{E(t)}\right) = \left(E(t) - \hat{E(t)}\right)^2$$

$$loss\left(I(t), \hat{I(t)}\right) = \left(I(t) - \hat{I(t)}\right)^2$$

$$loss\left(D(t), \hat{D(t)}\right) = \left(D(t) - \hat{D(t)}\right)^2$$

$$loss\left(R(t), \hat{R(t)}\right) = \left(R(t) - \hat{R(t)}\right)^2$$

Note: In Eq(12), the loss value of each group should be divided by the corresponding length of data set. But for simplicity, it is simplified into the form of Eq (12).

Training steps of SEIDR model

① Bring $\partial_{count}, \lambda_{count}, \varepsilon_{count}, \beta_{count}$ into the differential Eq(11) to solve, get the predicted number of infected people $\hat{I}$, the predicted number of latent people $\hat{E}$, the number of deaths due to illness $\hat{D}$ and the number of recovered people $\hat{R}$.

$\hat{I} = \{I(0), I(1), \text{L}, I(T)\}$
$\hat{E} = \{E(0), E(1), \text{L}, E(T)\}$
$\hat{D} = \{D(0), D(1), \text{L}, D(T)\}$
$\hat{R} = \{R(0), R(1), \text{L}, R(T)\}$

② Calculate the value of the loss function based on the predicted value and the true value $loss_{count}$.

③ if $|loss_{count} - loss_{count-1}| < \delta$;(Among them: $count > 1$).

④Use the L-BFGS algorithm to optimize the loss function, update the value of the parameter $\partial, \lambda, \varepsilon, \beta$, and get $\hat{\partial}, \hat{\lambda}, \hat{\varepsilon}, \hat{\beta}$

⑤$count \leftarrow count + 1$;

$\partial_{count+1} \leftarrow \partial_{count}$ $\lambda_{count+1} \leftarrow \lambda_{count}$,

$\varepsilon_{count+1} \leftarrow \varepsilon_{count}$ $\beta_{count+1} \leftarrow \beta_{count}$.

3.3 Model Analysis of SEIDR

Under normal circumstances, when COVID-19 first occurs, that is, when t = 0, the initial distribution of each group should satisfy: S(0)≈N(0), E(0) = 0, I(t) = 0, D(t) = 0, R(t) = 0. However, due to the influence of some factors that can't be controlled by human, the initial value of the model may be changed slightly, and as far as possible, we should try to use the most primitive data of New York City that is closer to the situation of COVID-19.

The $S(t)$, $E(t)$, $I(t)$, $D(t)$, and $R(t)$ of COVID-19 are all continuous numerical functions with respect to time t, so they are all differentiable functions. Among them, we need to analyze the situation of COVID-19 and make predictions for future situation.

According to the system of Eq(11), it can be seen that:

$$\begin{aligned}&\frac{dS}{dt}+\frac{dE}{dt}+\frac{dI}{dt}+\frac{dD}{dt}+\frac{dR}{dt}\\&=\mu N-v(S+E+I+D+R)\\&=(\mu-v)N\end{aligned}$$

which shows that the total population of each group in the region changes in accordance with the laws of nature, that is, the total population of a region is always equal to the result of multiplying the difference between the natural birth rate and the natural mortality rate and the total population.

The $I(t)$ and $E(t)$ are the focus of the analysis. It can predict when COVID-19 will reach its peak in a certain area and when it will usher in a turning point.

From the Eq $\frac{dI}{dt} = \lambda E - vI - \beta I - \varepsilon I$ in the system of Eq (11), we can find that when $I'(t) = 0$, there is:

$$I(t) = \frac{\lambda}{v+\beta+\varepsilon} E(t) \tag{13}$$

According to equation (13), we found that to measure the trend of COVID-19 turning point, we need to first solve the $E(t)$.

Similarly: let $\frac{dE}{dt} = \partial \frac{S}{N} I - vE - \lambda E = 0$. There is:

$$E(t) = \frac{\partial S(t) I(t)}{N(v+\lambda)} \tag{14}$$

Substituting equation (14) into equation (13), and we have:

$$I(t) = \frac{\lambda}{v+\beta+\varepsilon} * \frac{\partial S(t) I(t)}{N(v+\lambda)} \tag{15}$$

Solving equation (15), then:

$$S = \frac{(v+\beta+\varepsilon)(v+\lambda)N}{\lambda\partial} = N\left(1+\frac{v}{\lambda}\right)\frac{v+\beta+\varepsilon}{\partial} \tag{16}$$

In summary, when equation (16) is satisfied, the number of people infected with COVID-19 will reach a peak, and then the number of people infected will drop, that is, COVID-19 will reach a turning point at this time. Observing equation (16) carefully, it is found that the number of susceptible people, that is S, can be increased by controlling the latent morbidity rate λ, the infection rate ∂, the recovery rate β, disease-related mortality ε and the natural mortality rate v. This paper does not list the natural mortality v as research objective, so the best case is to

reduce the infection rate ∂ and the latent morbidity rate λ and increase the recovery rate β and the mortality rate ε due to illness as soon as possible, so that the value of S in equation (15) can be as large as possible to make the turning point come early.

In addition, according to the equation $\frac{dD}{dt} = \varepsilon I - vD$ in the Eq (11), combining with Euler approximation method $\frac{dD}{dt} \approx \frac{\Delta D}{\Delta t}$, we can get:

$$\frac{\Delta D}{\Delta t} = \varepsilon I(t) - vD(t) \tag{17}$$

Similarly, we can get:

$$\frac{\Delta R}{\Delta t} = \beta I(t) - vR(t) \tag{18}$$

This paper uses days as the unit of time, so the change of each group will be relatively small. The number of natural deaths per day is extremely small, not enough to affect the COVID-19 trend. Therefore, in the analysis of the trend of the number of deaths due to disease in equation (16), the influence of natural death factors is not considered for the time being.

Since the recovery rate β, disease-related mortality ε satisfy $\beta \in (0,1)$, $\varepsilon \in (0,1)$, so when the effects of natural death factors are temporarily ignored, both $\frac{\Delta D}{\Delta t}$ and $\frac{\Delta R}{\Delta t}$ change with the change of $I(t)$, so the change of $I(t)$ corresponds to the change of $D(t)$ and $R(t)$, it can be found that the curve of $I(t)$ will be smoother than that of $D(t)$ and $R(t)$.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Introduction of experimental data

Data is indispensable for the analysis and prediction of COVID-19, and the authenticity of the data source is extremely important. Based on past experience, we can know that the conclusions analyzed based on the wrong data may differ a lot from the fact, and the lack of real data will also cause the model to have deviation in the prediction process. In order to analyze and predict the trend of COVID-19 more accurately, data of COVID-19 must have timeliness, completeness and accuracy.

The data's time span of COVID-19 in New York City used in this paper is from 2020-04-01 to 2020-10-26. The data source includes two parts. One part comes from NetEase Internet. The data source is updated in real time. We use the Uniform Resource Locator (URL) obtains the location and access method of the resource from the Internet, send requests to the address and save the data needed; the other part comes from the official data released by the New York Department of Health. In particular, because it takes a certain time for the health department to report data, the data we use (including the number of confirmed cases, the number of deaths, etc.) may have a certain lag, but the lag is not strong.

Hardware: HP, Windows 10, intel (i7-8700).

Software: Python3.6, Pandas, Numpy, Scipy, Matplotlib.

4.2 Experimental process and results

4.2.1 Check out trends in COVID-19

In order to understand the real situation of COVID-19 in New York City, at first, we visually display the distribution of various groups in New York City, as shown in Fig. 4:

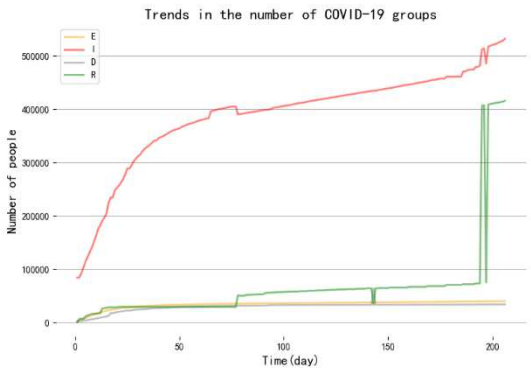


Fig. 4 Trends in the number of COVID-19 groups

From Fig.4, it can be found that the number of infected people has been on the rise; while the number of

exposed people has been similar to the number of people died of illness, and the rise is very slow; the number of recovered people has been increasing, but it suddenly exploded around the 195th day, then leveled off again, for reasons that are not yet clear. As can be seen from Fig.4, the trend of the number of people died of illness is relatively stable compared to that of infected people, and the number is much smaller than the number of infected people, this can indicate that the mortality rate of COVID-19 is not so high; But the rate of recovery is still lower than the rate of infection, indicating that the situation of COVID-19 in New York City remains a serious threat to people's lives.

4.2.2 Prove that the proposed SEIDR model is superior to SIR model

After processing and analyzing the real data source of COVID-19 in New York City, we used the traditional SIR model and the SEIDR model we proposed to conduct experiments, and used the same optimization method (that is, the L-BFGS method) to fit the parameters, and then used the fitted parameters to predict the future situation of COVID-19. The SIR model and the SEIDR model used the same calculation method of loss value, and the SIR model calculates the sum of loss value of infected people I and removed people R. The fitting results of the SIR model and the SEIDR model are shown in Table 2:

Table 2 The SIR model and SEIDR model parameters fitting table

Parameter	SIR Model	SEIDR Model
∂	0.0163	0.2216
λ		1e-08
ε		0.0356
β	0.0043	0.0131
R_0	3.7742	4.5489
$loss$	77423133	21470469

Note: The absence of values in Table 2 indicates that there is no corresponding value for the model

It can be seen from Table 1 that when the same data set and the same optimization method are used, the loss of fitting parameters using the SEIDR model is reduced by 72.27% compared with fitting parameters using the SIR model, indicating that using the SEIDR model to analysis the situation of COVID-19 in New York City is more effective.

What needs to be pointed out here is that since the loss value is the sum of the mean square error of the real value and the predicted value of several groups, it may seem a bit large, but this is just the reference standard which we use to measure how the SEIDR model is better than the SIR model.

4.2.3 Fitting parameters

Select different training sets from the same data source to do 12 comparative experiments, and respectively fit the four parameters: infection rate, latent morbidity, mortality due to disease, and recovery rate. The fitting results are shown in Table 3:

In these 12 experiments, the change trajectory of basic regeneration number R_0 is shown in Fig. 5:

It can be seen from Fig.5 that the range of R_0 of COVID-19 has always been between 4.0 and 4.6. Refer to the range of R_0 of COVID-19 in Table 1, which indicates that the situation of COVID-19 in New York City is still relatively serious, it is urgent to strengthen epidemic prevention measures. The value of R_0 in Fig. 5 has always been stable, but the value of R_0 in the fourth experiment has changed greatly, indicating that the training set selected in the fourth experiment may have some special data. This may have something to do with the sharp rise in the number of restorers in Fig4, which causes a sudden increase in recovery rate. Then according to the $R_0 = \frac{\partial}{\varepsilon + \beta}$,

the value of the basic regeneration number R_0 [8][17] can be reduced.

Therefore, we calculated the average value of the other 11 groups excluding the fourth group, and got $\overline{\partial} = 0.2215, \overline{\lambda} = 1e\text{-}08, \overline{\varepsilon} = 0.0356, \overline{\beta} = 0.0131, \overline{R_0} = 4.5505$.

Table 3 Parameter fitting table

	∂	λ	ε	β	R_0
1	0.2252	1e-08	0.0364	0.0134	4.5194
2	0.2245	1e-08	0.0363	0.0134	4.5236
3	0.2239	1e-08	0.0361	0.0133	4.5289
4	0.2399	0.0005	0.0386	0.0202	4.0812
5	0.2229	1e-08	0.0359	0.0132	4.537
6	0.2224	1e-08	0.0358	0.0131	4.5418
7	0.2216	1e-8	0.0356	0.0131	4.5489
8	0.221	1e-08	0.0355	0.0130	4.5548
9	0.2205	1e-08	0.0354	0.0130	4.5591
10	0.2185	1e-08	0.035	0.0128	4.5779
11	0.2182	1e-08	0.0349	0.0127	4.5805
12	0.2179	1e-08	0.0348	0.0127	4.5831

Fig.5 Variation trend of Basic regeneration R_0

4.2.4 Forecast the future development trend of COVID-19

Based on the parameters obtained above, predicting the situation of each group of COVID-19 in New York City in the next two years and the next five years respectively. Since S can be obtained from E, I, D and R, so it is not shown here. The prediction results of the four groups are shown in Fig 6 and Fig 7:

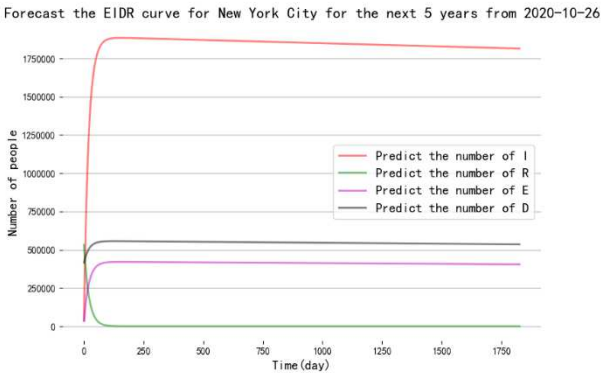
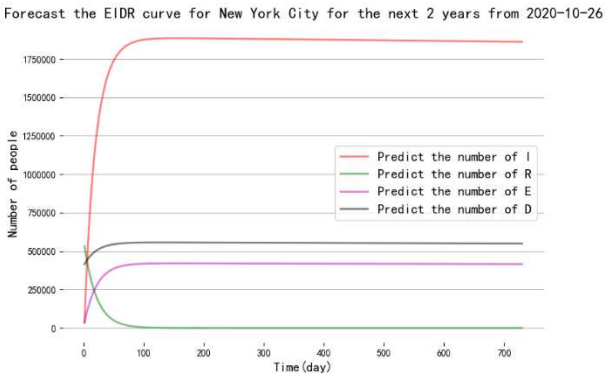


Fig 6 New York City's EIDR curve in the next 2 years from 2020-10-26

Fig 7 New York City's EIDR curve in the next 5 years from 2020-10-26

Based on Fig 6 and Fig 7, we can see that the predicted peak of the number of infected people of COVID-19 will be reached approximately 100 days after October 26, 2020, i.e. on February 15, 2021, the number of confirmed cases of COVID-19 will reach its peak. At its peak, the cumulative number of patients will reach approximately 2 million, accounting for approximately 24% of the total population of New York City. In the same way, February 15, 2021 will reach the turning point for COVID-19 in New York City. However, even if the turning point is reached, it can be seen from Fig 6 and Fig 7 that the number of infected people will decline extremely slowly, which requires the government to introduce more effective policies to reduce the infection rate ∂ , latent morbidity rate λ , and disease-related mortality ε ,and increase recovery rate β . Only in this way can COVID-19 be effectively



controlled.

It needs to be pointed out that the arrival of the aforementioned peak means that if New York City has been unable to introduce more effective epidemic prevention policies and the medical situation cannot be improved faster, that is, this peak is what is expected if the situation of COVID-19 in New York City continues to maintain such infection rate, latent morbidity rate, disease-related mortality and recovery rate.

5 CONCLUSION

This paper proposes a SEIDR model for analyzing and predicting epidemic infectious diseases. Compared with the traditional SIR model, the SEIDR model proposed in this paper can simulate the situation of COVID-19 in New York City more accurately. Through the SEIDR model, the infection rate ∂ , latent morbidity rate λ , disease-related mortality ε and the recovery rate β of COVID-19 in this

region are effectively fitted, and the basic reproduction value R_0, which can measure the severity of infectious diseases, is obtained. It also predicted the development trend of the epidemic in New York City. According to the projections, the situation of COVID-19 in New York City will reach a turning point on February 15, 2021 with the same level of prevention and control, but it will not be controlled completely in the short term, and this need to strengthen epidemic prevention and improve medical standards. The predictions in this paper are based on the situation of maintaining the local epidemic status quo. That is, if the government has better treatment policies, the future epidemic situation will be better than the results predicted by this paper.

However, this paper does not take into account the re-infection of the recovered people after recovery, nor does it take into account the isolation of the infected people and the exposed people. If these factors are taken into account, the COVID-19 situation can be better simulated and the prediction results will be more accurate. Since there is no data in this field at present, so it is difficult to study, and we will continue to study in this field if there is an opportunity in the future.

6 DECLARATIONS

Ethics approval and consent to participate. The dataset in this paper comes from a public data source, which can be used for scientific research and does not violate ethical issues

Consent for publication. The data's time span of COVID-19 in New York City used in this paper is from 2020-04-01 to 2020-10-26. The data source includes two parts. One part comes from NetEase Internet. The data source is updated in real time. We use the Uniform Resource Locator (URL) obtains the location and access method of the resource from the Internet, send requests to the address and save the data needed; the other part comes from the official data released by the New York Department of Health. In particular, because it takes a certain time for the health department to report data, the data we use (including the number of confirmed cases, the number of deaths, etc.) may have a certain lag, but the lag is not strong.

Availability of data and material. Since this paper predicts the trend of COVID-19 in New York, there is no data available to support the results in this paper. Moreover, this paper extends to the entire epidemiology with Covid-19, and the SEIDR model proposed in this paper can be used to fit the important parameters of the disease for subsequent studies.

Competing interests. The authors declare that they have no competing interests.

Funding. Not applicable for the moment.

Authors' contributions. We propose a SEIDR model for analyzing and predicting epidemic infectious diseases. And we fit the early transmission data of COVID-19 in New York City, obtains important parameters, such as infection rate, latent morbidity rate, disease-related mortality and recovery rate, and we use the value of basic regeneration number R_0 proved that the situation of COVID-19 in New York City was relatively serious. And we predict the future development of COVID-19 in New York City.

Acknowledgements. Not applicable for the moment.

Authors' information. Zhigang Xu, School of Computer Science, Hubei University of Technology, 28 Nanli Road, Wuhan 430068, Hubei, China, imdr@163.com. Huiling Xia, School of Computer Science, Hubei University of Technology, 28 Nanli Road, Wuhan 430068, Hubei, China.

7 REFERENCE

[1] Pierpoint Lauren A,Collins Christy. Epidemiology of Sport-Related Concussion.[J]. Clinics in sports medicine,2021,40(1).

[2] Nikolopoulos Konstantinos,Punia Sushil,Schäfers Andreas,Tsinopoulos Christos,Vasilakis Chrysovalantis. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions[J]. European Journal of Operational Research,2021,290(1).

[3] Bakre Abhijeet A.,Jones Les P.,Kyriakis Constantinos S.,Hanson Jarod M.,Bobbitt Davis E.,Bennett Hailey K.,Todd Kyle V.,Orr Burks Nichole,Murray Jackelyn,Zhang Ming,Steinhauer David A.,Byrd Leotis Lauren,Cummings Richard

D.,Fent Joseph,Coffey Terry,Tripp Ralph A.. Molecular epidemiology and glycomics of swine influenza viruses circulating in commercial swine farms in the southeastern and midwest United States[J]. Veterinary Microbiology,2020,251.

[4] Sun Zhangpeng,Yao Wenqi,Lu Tiao. Optimization Modeling and Simulating of the Stationary Wigner Inflow Boundary Value Problem[J]. Journal of Scientific Computing,2020,85(1).

[5] Qingyan Chen. Can we migrate COVID-19 spreading risk?[J]. Frontiers of Environmental Science & Engineering,2020,15(3).

[6] Zahra Belhadjer,Mathilde Méot,Fanny Bajolle,Diala Khraiche,Antoine Legendre,Samya Abakka,Johanne Auriau,Marion Grimaud,Mehdi Oualha,Maurice Beghetti,Julie Wacker,Caroline Ovaert,Sebastien Hascoet,Maëlle Selegny,Sophie Malekzadeh-Milani,Alice Maltret,Gilles Bosser,Nathan Giroux,Laurent Bonnemains,Jeanne Bordet,Sylvie Di Filippo,Pierre Mauran,Sylvie Falcon-Eicher,Jean-Benoît Thambo,Bruno Lefort,Pamela Moceri,Lucile Houyel,Sylvain Renolleau,Damien Bonnet. Acute heart failure in multisystem inflammatory syndrome in children (MIS-C) in the context of global SARS-CoV-2 pandemic[J]. Circulation,2020.

[7] Sunil Kumar,Ali Ahmadian,Ranbir Kumar,Devendra Kumar,Jagdev Singh,Dumitru Baleanu,Mehdi Salimi. An Efficient Numerical Method for Fractional SIR Epidemic Model of Infectious Disease by Using Bernstein Wavelets[J]. Mathematics,2020,8(4).

[8] Yu Chen,Jin Cheng,Yu Jiang,Keji Liu. A time delay dynamical model for outbreak of 2019-nCoV and the parameter identification[J]. Journal of Inverse and Ill-posed Problems,2020,28(2).

[9] Linhe Zhu,Xuewei Wang,Huihui Zhang,Shuling Shen,Yimin Li,Yudong Zhou. Dynamics analysis and optimal control strategy for a SIRS epidemic model with two discrete time delays[J]. Physica Scripta,2020,95(3).

[10] Yuan Gao,Qiyong Liu. The Epidemic Dynamics of 2019 Novel Coronavirus (2019-nCoV) Infections in China by 28 January[J]. SSRN,2020.

[11] Albert S. Berahas,Martin Takáč. A robust multi-batch L-BFGS method for machine learning[J]. Optimization Methods and Software,2020,35(1).

[12] Yue Y, Yu C, Keji L, et al. Modeling and prediction for the trend of outbreak of NCP based on a time-delay dynamic system[J]. Scientia Sinica Mathematica, 2020.

[13] Wencai Zhao,Jinlei Liu,Mengnan Chi,Feifei Bian. Dynamics analysis of stochastic epidemic models with standard incidence[J]. Advances in Difference Equations,2019,2019(1).

[14] Salata Cristiano,Calistri Arianna,Parolin Cristina,Palù Giorgio. Coronaviruses: a paradigm of new emerging zoonotic diseases.[J]. Pathogens and disease,2019,77(9).

[15] Chang Daqing,Sun Shiliang,Zhang Changshui. An Accelerated Linearly Convergent Stochastic L-BFGS Algorithm.[J]. IEEE transactions on neural networks and learning systems,2019,30(11).

[16] Hu Hai Jun,Yuan Xu Pu,Huang Li Hong,Huang Chuang Xia. Global dynamics of an SIRS model with demographics and transfer from infectious to susceptible on heterogeneous networks.[J]. Mathematical biosciences and engineering : MBE,2019,16(5).

[17] Zhishuang Wang,Quantong Guo,Shiwen Sun,Chengyi Xia. The impact of awareness diffusion on SIR-like epidemics in multiplex networks[J]. Applied Mathematics and Computation,2019,349.

[18] Hasan Badem,Alper Basturk,Abdullah Caliskan,Mehmet Emin Yuksel. A new hybrid optimization method combining artificial bee colony and limited-memory BFGS algorithms for efficient numerical optimization[J]. Applied Soft Computing Journal,2018,70.

[19] Rongrong Wu,He Huang,Xusheng Qian,Tingwen Huang. A L-BFGS Based Learning Algorithm for Complex-Valued Feedforward Neural Networks[J]. Neural Processing Letters,2018,47(3).

[20] Begoña Cantó,Carmen Coll,Elena Sánchez. Estimation of parameters in a structured SIR model[J]. Advances in Difference Equations,2017,2017(1).

[21] Jinhui Li,Zhidong Teng,Guangqing Wang,Long Zhang,Cheng Hu. Stability and bifurcation analysis of an SIR epidemic model with logistic growth and saturated treatment[J]. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena,2017,99.

[22] Yanju Xiao, Weipeng Zhang, Guifeng Deng, Zhehua Liu. STABILITY AND BOGDANOV-TAKENS BIFURCATION OF AN SIS EPIDEMIC MODEL WITH SATURATED TREATMENTFUNCTION[J]. Inventi Impact Engineering Mathematics,2015,2015.

[23] Mimi Yan,Ruiqing Shi. Analysis of an SI Epidemic Model with Nonlinear Incidence Rate in an Environmentally-driven Infectious Disease[J]. Journal of Advances in Mathematics and Computer Science,2014.

[24] Guihua Li,Gaofeng Li. Bifurcation Analysis of an SIR Epidemic Model with the Contact Transmission Function[J]. Abstract and Applied Analysis,2014,2014.

[25] Faryad Darabi Sahneh,Caterina Scoglio,Piet Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks[J]. IEEE/ACM Transactions on Networking (TON),2013,21(5).

[26] José Luis Morales,Jorge Nocedal. Remark on “algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization”[J].

ACM Transactions on Mathematical Software (TOMS),2011,38(1).
[27] John W. Glasser,Nathaniel Hupert,Mary M. McCauley,Richard Hatchett. Modeling and public health emergency responses: Lessons from SARS[J]. Epidemics,2011,3(1).
[28] Zhang Juan,Lou Jie,Ma Zhien,Wu Jianhong. A compartmental model for the analysis of SARS transmission patterns and outbreak control measures in China.[J]. Applied mathematics and computation,2005,162(2).
[29] David J. D. Earn,Pejman Rohani,Benjamin M. Bolker,Bryan T. Grenfell. A Simple Model for Complex Dynamical Transitions in Epidemics[J]. Science,2000,287(5453).
[30] Capasso Vincenzo,Serio Gabriella. A generalization of the Kermack-McKendrick deterministic epidemic model[J]. Elsevier,1978,42(1-2).4.06535 (2019).