TW-SIR: Time-window based SIR for COVID-19 forecasts

Zhifang Liao
School of Computer Science and Engineering, Central South University, Changsha, 410075, China
Peng Lan
School of Computer Science and Engineering, Central South University, Changsha, 410075, China
Zhingning Liao (leoleo.db@gmail.com)
Nuffield health research group, Nuffield Health, Ashley Avenue, Epsom, Surrey KT18 5AL, UK
Yan Zhang
Department of Computing, School of Computing, Engineering and Built Environment, Glasgow Caledonian University,Glasgow G4 OBA,UK
Shengzong Liu (lshz179@163.com)
Department of Information Management,Hunan University of Finance and Economics, Changsha 410075,China

Research Article

Keywords: COVID-19, SIR Model, Time window, Basic reproduction number, Exponential growth rate

Posted Date: September 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-75220/v1

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2020.
See the published version at https://doi.org/10.1038/s41598-020-80007-8.

TW-SIR: Time-window based SIR for COVID-19 forecasts

Zhifang Liao[1], Peng Lan[1], Zhingning Liao[2,*], Yan Zhang[3], and Shengzong Liu[4,*]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410075, China; zfliao@csu.edu.cn; lanpeng5@qq.com
[2]Nuffield health research group, Nuffield Health, Ashley Avenue, Epsom, Surrey KT18 5AL, UK
[3]Department of Computing, School of Computing, Engineering and Built Environment, Glasgow Caledonian University,Glasgow G4 OBA,UK; yan.zhang@gcu.ac.uk
[4]Department of Information Management,Hunan University of Finance and Economics, Changsha 410075,China
[*]Correspondeing author; leoleo.db@gmail.com; lshz179@163.com

ABSTRACT

Since the outbreak of COVID-19, many COVID-19 research studies have proposed different models for predicting trend of COVID-19. Among them, the prediction model based on mathematical epidemiology (SIR) is the most widely used, but most of these models are adapted in special situations based on various assumptions. In order to reflect the real-time trend of the epidemic in the process of infection for different areas, different policies and different epidemic diseases, a general adapted time- window based SIR model is proposed, which is characterized by introducing a time window mechanism for dynamic data analysis and using machine learning method predicts the Basic reproduction number R0 and the exponential growth rate of the epidemic. Multiple data sets of epidemic diseases are analyzed, and the numerical results showed that the framework can effectively measure the real-time changes of the parameters during the epidemic, and error rate of predicting the number of COVID-19 infections in a single day is within 5%.
Keywords: COVID-19; SIR Model; Time window; Basic reproduction number; Exponential growth rate

1 Introduction

Since the outbreak of COVID-19, the epidemic has spread rapidly in many countries and regions in the world, The World Health Organization declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on January 30, 2020. According to data released by Johns Hopkins University, on June 28, 2020, confirmed cases of COVID-19 have been detected in 185 countries and regions around the world, of which 9,953,038 have been confirmed and 498,178 have died. In order to reduce the impact of COVID-19, forecasting trend of COVID-19, such as COCID-19 peak and stage of its spread, is of great significance for the government to formulate prevention and control strategies, take timely measures, and allocate medical resources. There have been many studies to predict the development trend of the epidemic in various countries and regions. These studies can be roughly divided into three categories: ①statistical models② AI-based predictions③mathematical epidemic models.

Statistical model methods estimate main epidemic parameters through case reports and other data statistics, including the basic reproduction number(R0), the incubation periodserial interval and generation time etc., then use mathematical models such as exponential growth to predict the epidemic curve[1–5]. This method is suitable for roughly estimating the epidemic in the early stage of the epidemic. With the development of the epidemic, these epidemic parameters are constantly changing in different countries and regions, which leads to the kind of prediction is only does not reflect the actual situation of epidemic.

The AI-based prediction method is an emerging method for predicting COVID-19, which is used to predict how COVID-19 propagates over time and space. Hu et al.(2020) used a modified stacked Auto-Encoder for modelling the transmission dynamics of the epidemics to real-time forecasting the confirmed cases of COVID-19 in China[6]. Yang et al.(2020) divided into the data of SARS outbreak in 2003 with three days as input, and used the long and short-term memory network model(LSTM) for training to predict the new coronavirus outbreak in China mainland[7]. Friston et al.(2020) developed a dynamic causal model of COVID-19 based population dynamics, and this model leveraged Bayesian model comparison[8]. Although the accuracy of AI-based method is very good and the prediction curve can be fitted well, there are still two problems with AI-based methods. The first one is that the prediction method cannot be trained well to achieve the desired effect, because of lacking of the training data, special at the beginning of pandemic. The other one is overfitting in this kind of methods then it may therefore fail to fit additional data or predict future observations reliably. Hence, established mathematical epidemiological models were used to

track and forecast in most studies so far.

There are two kind of typical mathematical epidemiological models including SIR and SEIR (Susceptible, Exposed, Infected, and Removed) and so on. Several the studies are modified on these two mathematical epidemiological models to adapt to specific requirements and analyze the transmission dynamics of COVID-19. The modifications of the model are divided into several types: adding a new state or modifying the model parameters on the basis of the original model, integrating additional external data into the model, adding the effects of non-pharmaceutical interventions on the model, etc. Liu etc al.(2020) added infected individuals who did not report symptoms based on the SIR model, and used the case data reported in China early to predict the cumulative number of reported cases. The main feature of the model is to model the timing of the government 's main public policies[11]. Peng L et al. (2020) and others proposed a generalized SEIR model, re-formulated a new isolation state and considered the effects of preventive measures, and analyzed the epidemic situation in 5 different regions of China[12]. However, due to the limitations of detection methods and diagnostic criteria, unreported cases and exposed cases are difficult to be estimated and accurate numbers of these cases are difficult to be obtained. These number are regarded as hidden variables in the research process. Sun H et al.(2020) developed a time-varying coefficient vSIR model to reflect the changes of model parameters due to the government intervention[13]. Chen et al.(2020) developed a time-dependent SIR model for COVID-19 with undetectable infected persons and used the two finite impulse response filters to track and predict the numbers of infected persons and recovered people in China[14]. As far as the results are concerned, the prediction error is very small, but the training of the model is based on the fact that the data is sufficient, and it is not suitable for the early stage of prediction of the epidemic. Fanelli D et al.(2020) used the SIRD model with the death status to predict the epidemic trends in China, Italy and France, and found that the time evolution of COVID-19 has a certain degree of universality and has little connection with geographical changes[15], but this research study is only based on a simple quantitative model to evaluate the effect of strict epidemic prevention. In addition, other studies have modified the SEIR modelsuch as considering the population migration data[7], analyzing the proportion of infected passengers on evacuation flights[16], and so on. Although these methods of modifying epidemiological models can be used to assess the spread of epidemics and the impact of government intervention strategies, these models require the introduction of additional parameters and depend on many assumptions. At the same time, studies have shown that the increase in the number of unknown parameters in a complex model needs to be estimated by model fitting, which will lead to higher uncertainty in model predictions. Therefore, simple models may be more reliable than complex models in the process of model selection[17].

In the traditional SIR model, there are two key parameters that reflect the characteristics of the epidemic: infection rate of the pathogen β and recovery rate γ. The infection rate β indicates that each susceptible population randomly infects β people every day; the recovery rate γ indicates that the infected person recovers or dies with the probability of γ. These two parameters are constant in the traditional SIR model. When applied it to the real world, they are often not able to measure and predict the trend of epidemics. Therefore, many studies have regarded them as functions that change over time[13,14].

However, due to differences in epidemic prevention and control measures in various countries and regions and with the evolution of the epidemic, the manually selected function is not suitable for real-time changes of these parameters. That different policies of prevention and control measures are adopted by different countries and regions during the epidemic will be leading to different results, in order to reflect this change of the key parameters in SIR model, a time window-based SIR prediction model (TW-SIR) is proposed, which can capture, track and predict dynamic changes of epidemic parameters in real time. The model first divides historical data based on time windows, uses the fourth-order Runge-Kutta method to numerically solve the SIR model, and uses machine learning methods to predict R0 and the exponential growth rate of the epidemic to solve β and γ. The model is applied to COVID-19 data in multiple countries and the experimental results show that the TW-SIR model can effectively measure the real-time changes in the epidemic infection process, and can effectively predict the peak and the end of the epidemic.

The rest of this paper is organized as follows: in the second section, we propose the TW-SIR model. In the third section, we conducted some numerical experiments and analyzed the experimental results to illustrate the effectiveness of our model. Then, in Section 4, we made some discussions and suggestions. Finally, the fifth section is a summary of the paper.

2 Methods

2.1 SIR epidemic model

The susceptibility-infection-recovery (SIR) model[18] is one of the simplest and commonly used epidemic models. The model consists of three compartments: S: The number of susceptible individuals, I: The number of infectious individuals, R for the number of removed (and immune) or deceased individuals. The SIR epidemic model can be expressed by following set of ordinary differential equations (ODE):

$$\frac{dS(t)}{dt} = -\frac{\beta I(t)S(t)}{N} \tag{1}$$

$$\frac{dI(t)}{dt} = \frac{\beta I(t)S(t)}{N} - \gamma I(t) \tag{2}$$

$$\frac{dR(t)}{dt} = \gamma I(t) \tag{3}$$

$$N = S(t) + I(t) + R(t) \tag{4}$$

Among them, $S(t)$, $I(t)$ and $R(t)$ respectively represent the functions of S, I and R related to time t, and their sum satisfies formula (4); N represents the total number of populations; β represents the probability of infection rate, Which means that each susceptible population randomly infects β people every day. The recovery rate γ indicates that the infected person recovers or dies with the probability of γ.

Although the SIR model is simple, the analysis and use of it in many studies generally show that it can capture the trend and overall characteristics of the epidemic. In the traditional SIR model, β and γ are parameters that reflect the characteristics of the epidemic, and they are constants. However, if the parameters are constant, it is often impossible to measure and predict the development trend of epidemics when applied to the real world. Therefore, many studies have regarded them as functions that change over time and used formulas to derive them. Considering that during the development of the epidemic, the parameters in the SIR model are changing in real time for different countries and regions. In order to reflect these changes in the parameters of the epidemic model, in this article we propose the TW-SIR prediction model, which can capture, track and predict the dynamic changes of the epidemic parameters in real time. We will introduce this model in detail in the next section.

2.2 Time-Window SIR model

In order to represent the changes of parameters in the SIR model, we propose a time window-based SIR model (TW-SIR) which splits historical data into a time window segment. The purpose of this method is to capture R0 and real-time changes in the growth rate of the epidemic index. The SIR-TW model is based on the assessment of the changes in the epidemiological parameters of historical data every day through a time window and solves the problem that the formula derivation method cannot be measured in real time. The TW-SIR model is mainly composed of three parts: model solution, parameter evaluation and parameter prediction. Among them, the model solution uses the input data to solve the SIR model; the parameter evaluation is mainly to divide the input data into a time window to evaluate the parameter values of each day; the parameter prediction uses the existing parameter values obtained by the parameter evaluation to predict value of parameters in the future. Figure 1 shows the main workflow of the model. The input is the data of number of infected people per day, splits this data into a time window segment, the parameters of each day are evaluated through the data in the time window of the model, and then the estimated parameter values obtained are used to predict the parameters in the future, and finally used to substitute into the model Solve to get the prediction result.

In Fig. 1, ①is the input historical data. ②③④are main part of the model, which are parameter evaluation, parameter prediction and model solution. ⑤is the final output prediction result. Aim of TW-SIR is to evaluate the changes in the parameters of the epidemic in order to predict the development trend of the epidemic. In the remainder of this section, we will describe the contents of each part in detail.

A. Model solution

The function of model solution is to numerically solve SIR model equations to facilitate subsequent parameter evaluation. After the model parameters are determined in the SIR model, the model can be solved. Because SIR model equations are coupled nonlinear ordinary differential equations, it is difficult to find analytical solutions to the equations. Although it is possible to derive the analytical solution of the equation in implicit form, the solution process is complicated and practical applications have limitations[19]. Compared with analytical solutions, methods such as numerical solutions are more commonly used in such research problems, and these methods are more effective. In this paper, numerical solution method, namely Runge-Kutta method, is used to numerically solve the SIR model. The Runge-Kutta method is a high-precision single-step algorithm, and its classic method is the fourth-order Runge-Kutta method (RK4). RK4 divides the time interval between t and t +1 into four subintervals and solves ordinary differential equations by calculating the slope values of these subintervals points and weighting them as the average slope. For the three states of the SIR model, we use RK4 to modify the differential equations in (1), (2) and (3) into discrete differential equations:

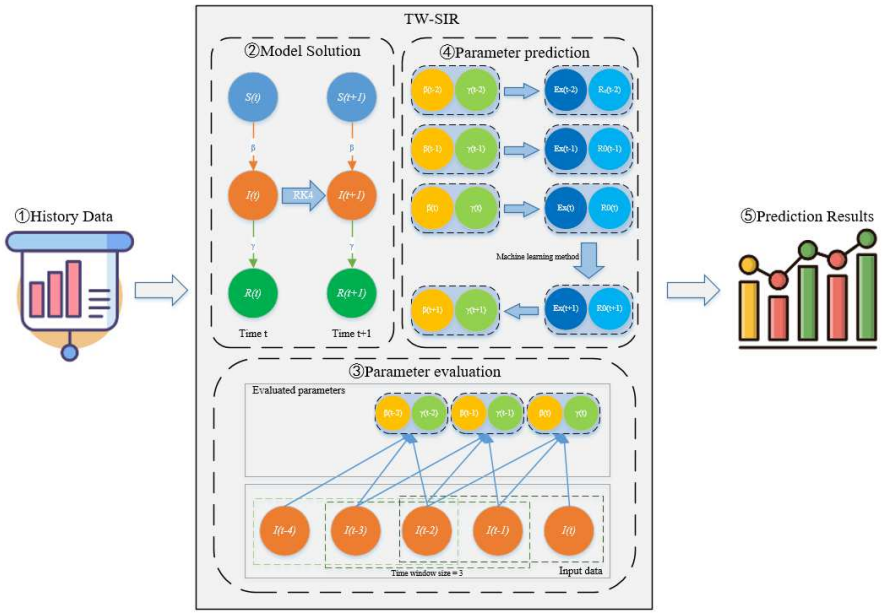


Figure 1. The composition of the TW-SIR model

$$S(t+1) = S(t) + \frac{h(S_1^{'} + 2S_2^{'} + 2S_3^{'} + S_4^{'})}{6} \tag{5}$$

$$I(t+1) = I(t) + \frac{h(I_1^{'} + 2I_2^{'} + 2I_3^{'} + I_4^{'})}{6} \tag{6}$$

$$R(t+1) = R(t) + \frac{h(R_1^{'} + 2R_2^{'} + 2R_3^{'} + R_4^{'})}{6} \tag{7}$$

Where h is the step-size, $S_i^{'}$, $I_i^{'}$ and $R_i^{'}(i = 1,2,3,4)$ respectively indicate the slopes of the four subintervals in the interval $[t, t+1]$ of $S(t)$, $I(t)$ and $R(t)$, which can be calculated by formulas (8), (9), (10) and (11).

$$\begin{cases} S_1^{'} = -\frac{\beta I(t) S(t)}{N} \\ I_1^{'} = \frac{\beta I(t) S(t)}{N} - \gamma I(t) \\ R_1^{'} = \gamma I(t) \end{cases} \tag{8}$$

$$\begin{cases} S_2^{'} = -\frac{\beta (S(t) + \frac{hS_1^{'}}{2})(I(t) + \frac{hI_1^{'}}{2})}{N} \\ I_2^{'} = \frac{\beta (S(t) + \frac{hS_1^{'}}{2})(I(t) + \frac{hI_1^{'}}{2})}{N} - \gamma (I(t) + \frac{hI_1^{'}}{2}) \\ R_2^{'} = \gamma (I(t) + \frac{hI_1^{'}}{2}) \end{cases} \tag{9}$$

$$\begin{cases} S_3^{'} = -\frac{\beta (S(t) + \frac{hS_2^{'}}{2})(I(t) + \frac{hI_2^{'}}{2})}{N} \\ I_3^{'} = \frac{\beta (S(t) + \frac{hS_2^{'}}{2})(I(t) + \frac{hI_2^{'}}{2})}{N} - \gamma (I(t) + \frac{hI_2^{'}}{2}) \\ R_3^{'} = \gamma (I(t) + \frac{hI_2^{'}}{2}) \end{cases} \tag{10}$$

$$\begin{cases} S_4^{'} = -\frac{\beta (S(t) + hS_3^{'})(I(t) + hI_3^{'})}{N} \\ I_4^{'} = \frac{\beta (S(t) + hS_3^{'})(I(t) + hI_3^{'})}{N} - \gamma (I(t) + hI_3^{'}) \\ R_4^{'} = \gamma (I(t) + hI_3^{'}) \end{cases} \tag{11}$$

Through the above equations, β and γ are substitute into the SIR model to solve the model numerically. The three functions of $S(t)$, $I(t)$ and $R(t)$ satisfy formula (4).

B. Parameter evaluation

The parameter evaluation part is mainly to characterize the change of the infection coefficient β and the cure coefficient γ over time in the historical data, so as to facilitate subsequent parameter prediction. Firstly, the historical data is divided according to the size of the time window, then an initial values of the model parameters are set within the time window, and then traverse the search for the model parameters, and finally get the best model parameters for each day through evaluation. A time-dependent β(t) and γ(t) functions are used to instead of β and γ in the SIR model, which can be obtained

$$\frac{dS(t)}{dt} = -\frac{\beta(t)I(t)S(t)}{N} \tag{12}$$

$$\frac{dI(t)}{dt} = \frac{\beta(t)I(t)S(t)}{N} - \gamma(t)I(t) \tag{13}$$

$$\frac{dR(t)}{dt} = \gamma(t)I(t) \tag{14}$$

Where $\beta(t)$ and $\gamma(t)$ are functions with time t as an independent variable rather than constants. Due to the government action on infection prevention and control for COVID-19 and awareness of the population on COVID-19, $\beta(t)$ and $\gamma(t)$ change in real time. In order to measure this change, the time series data set is divided into time windows of size W, and then use the optimal parameter solution in the time window as the evaluation value. For historical data at time t, its time window is $\{w_t, 0 \le t \le T-1\}$, we can get the following formula:

$$\beta(t) = opt\{\beta_{w_t}\}, w_t = [t-w+1, t] \tag{15}$$

$$\gamma(t) = opt\{\gamma_{w_t}\}, w_t = [t-w+1, t] \tag{16}$$

Among them, $\beta_(w_t)$and $\gamma_(w_t)$represent a certain parameter solution in the SIR model at time t in the historical data with a time window size of w. In order to obtain the optimal solution $opt\{\beta_(w_t)\}$ under the time window w, two steps are applied in calculating it through search: firstly, determine the initial values of the model parameters and secondly perform traversal search on the model parameters. The first is the determination of the initial values of the model parameters. In the early stages of the epidemic, the proportion of the number of infected and cured population in the population is negligible. We can regard the susceptible number $S(t)$ and the total population N as approximately equal, so the differential equation (2) can be written as the formula (17):

$$\frac{dI(t)}{dt} = (\beta - \gamma)I(t) \tag{17}$$

Then we can get the analytical solution of the model through the above formula, as shown in formula (18):

$$I(t) = e^{(\beta-\gamma)t} \tag{18}$$

Where, the number of infected people is an exponential function that changes over time, and then the least squares method is used to retrospectively fit the actual data of the epidemic to obtain the initial values $\beta_0$and γ_0 of the parameters. The initial value obtained can evaluate the characteristics of the early stage of the epidemic, but a simple exponential growth model cannot fully reflect the full picture of the epidemic and a more accurate estimation needed. Therefore, based on the initial values, total number of confirmed COVID-19 cases and model numerical solution methods are used to traverse the model parameters.

Given the data within a specified time window $\{C(t), R(t), D(t), 0 \le t \le T-1\}$($C(t)$,$R(t)$ and $D(t)$ are respectively the cumulative number of COVID-19 cases, cumulative number of cured COVID-19 cases, and cumulative number of death cases per day), formula (19) is used to calculate the actual daily number of infections $I(t)$:

$$I(t) = C(t) - R(t) - D(t) \tag{19}$$

After getting the daily actual number of infected people, we use the RK4 method to find the numerical solution of the model, which is the predicted number of infected people $\hat{I}(t)$. In order to evaluate the parameters β and γ, the following formula is used to calculate the MSE (mean squared error) of the predicted result:

$$MSE(\beta,\gamma) = \frac{1}{T}\sum_{t}^{T}(\hat{I}(t) - I(t))^2 \tag{20}$$

In order to get the optimal size of time window, the size of time window is set from 3 to 30 to be tested and the accumulated forecast error is used to evaluate the accuracy and effectiveness of the forecast under each time window. $Error_w$ is the accumulated prediction error under the time window w, and the formula is shown as following.

$$Error_w = \frac{1}{T-W}\sum_{t=W-1}^{T}\frac{\left|\hat{I}(t) - I(t)\right|}{I(t)} \tag{21}$$

In the process of searching for model parameters, it takes too much time if a grid search is applied and it is easy to fall into the local optimum. To overcome this problem, in this article an optimized search method is used. First, we assume that the value of β is greater than the value of γ in the early stage of the epidemic, because this is necessary to ensure that the epidemic infection continues[20], namely the value of R_0 is greater than 1 and estimate the initial parameter values β_0 and γ_0 using formulas (17), (18). Based on the initial values β_0 and γ_0, we set the size of search step and the size of search interval. Then RK4 is used to solve the model by using formula (6). Finally, the MSE for β_{w_t} and γ_{w_t} are calculated and the β_{w_t} and γ_{w_t} with minimize of MSE are as Optimal parameters. The detailed steps of our parameter evaluation based on time window are shown in Algorithm 1.

Algorithm 1 Parameters evaluation based on time window.

Input: $\{I(t), 0 \le t \le T-1\}$ using (19);W.
Output: $\{\hat{I}(t), W-1 \le t \le T-1\}$, $\{\beta(t), W-1 \le t \le T-1\}$, and $\{\gamma(t), W-1 \le t \le T-1\}$.

while $W-1 \le t \le T-1$ **do**
  estimate the initial parameter values β_0 and γ_0 using (17), (18);
  while $\beta_{w_t} and \gamma_{w_t}$ in interval **do**
    calculate the MSE using (20);
    $\beta(t) = opt\{\beta_{w_t}\}$;
    $\gamma(t) = opt\{\gamma_{w_t}\}$;
  end while
end while
return $\{\hat{I}(t), W-1 \le t \le T-1\}$, $\{\beta(t), W-1 \le t \le T-1\}$, $\{\gamma(t), W-1 \le t \le T-1\}$;

After getting $\beta(t)$,$\gamma(t)$ $\{\beta(t), \gamma(t), w-1 \le t \le T-1\}$, machine learning methods can be applied to predict the time change of the infection coefficient and the cure coefficient and predict the future development trend of the epidemic.

C. Parameter prediction

Parameter prediction is to predict the subsequent model parameters based on the changes over time of the model parameters obtained from the previous part of the parameter evaluation. In this section, the polynomial regression algorithm widely used in machine learning is applied to track and predict $\beta(t)$ and $\gamma(t)$. It is difficult to accurately directly predict $\beta(t)$ and $\gamma(t)$ because of value fluctuations. Therefore, this paper proposes a new prediction method, using the method of predicting the R_0 and exponential growth rate $Ex(t)$ to calculate them, which their changing curve is easier to predict in the development of the epidemic. The Basic reproduction number R_0 also reflects the development of the epidemic. It can also be regarded as a function over time $R_0(t)$ which can be obtained by using formula (22):

$$R_0(t) = \frac{\beta(t)}{\gamma(t)} \tag{22}$$

In order to get $\beta(t)$ and $\gamma(t)$, we define an exponential growth rate index $Ex(t)$ according to the exponential growth model of formula (18), which is shown in the following formula:

$$Ex(t) = \beta(t) - \gamma(t) \tag{23}$$

Where the predicted basic reproduction number is $\hat{R_0}(t)$, and the predicted exponential growth rate is $\hat{Ex}(t)$. Through polynomial regression, they can be written in the following form:

$$\hat{R_0}(t) = a_0 + a_1 R_0(t-1) + a_2(R_0(t-1))^2 + \cdots + a_n(R_0(t-1))^n = \sum_{i=0}^{n} a_i(R_0(t-1))^i \tag{24}$$

$$\hat{Ex}(t) = b_0 + b_1 Ex(t-1) + b_2(Ex(t-1))^2 + \cdots + b_m(Ex(t-1))^m = \sum_{j=0}^{m} b_j(R_0(t-1))^j \tag{25}$$

n and m are the order of $\hat{R_0}(t)$ and$\hat{Ex}(t)$ polynomials $(n, m \geq 2), a_i (i = 0, 1, \ldots, n)$ and $b_j (j = 0, 1, \ldots, m)$ are the coefficients of these two polynomial functions. In order to determine the coefficient and order of the polynomial function, the most widely used least squares method (OLS) to evaluate the prediction results. At the same time, in order to ensure that the model is under-fitting and reflect the real-time changes of the epidemic, Time window method mentioned in the previous section is used to solve the following optimization problems:

$$min \sum_{t=T-W}^{T-1} (\hat{R_0}(t) - R_0(t))^2 \tag{26}$$

$$min \sum_{t=T-W}^{T-1} (\hat{Ex}(t) - Ex(t))^2 \tag{27}$$

W is the size of the time window. The coefficients and orders of the polynomial can be obtained by solving the objective optimization function, such as $a_i, i = 0, 1, \ldots, n$, and $b_j, j = 0, 1, \ldots, m$. After obtained these coefficients, $\hat{R_0}(t)$ and $\hat{Ex}(t)$at time $t = T$ can be obtained through the formula (24, 25), and then the predicted infection rate $\hat{\beta}(t)$ and the predicted recovery rate$\hat{\gamma}(t)$ can be calculated by using formula (28) and (29), namely:

$$\hat{\beta}(t) = \frac{\hat{Ex}(t)\hat{R_0}(t)}{1 - \hat{R_0}(t)} \tag{28}$$

$$\hat{\gamma}(t) = \frac{\hat{Ex}(t)}{1 - \hat{R_0}(t)} \tag{29}$$

Now we have got $\hat{\beta}(t)$ and $\hat{\gamma}(t)$, and then through the model solution method in Section A, the number of infections $\{\hat{I}(t), t > T\}$ in the subsequent epidemic can be predicted.

3 Results

3.1 Data sources

In this paper, we gathered epidemiological data from Johns Hopkins University[21]. The data include the daily cumulative number of confirmed cases, cumulative death cases, and cumulative cured cases of various countries from January 23, 2020 up to now. Taking China as an example, Table 1 shows the details of the data we used. In this article, we use the data of 7 countries including China, South Korea, France, Spain, Italy, Germany and Brazil as our data set. In addition, in order to verify that our method is applicable to different epidemics, we also gathered the SARS epidemic data of Beijing, China from April 20, 2003 to June 23, 2003 from the website of the Ministry of Health of China[22], and the format of the data is the same as in Table 1. Table 2 shows the COVID-19 data for China, South Korea, France, Spain, Italy, Germany, and Brazil, and the time frame of the 2003 Beijing SARS data.

Date	Confirmed	Deaths	Recovered
2020/01/27	2877	131	58
2020/01/27	5509	133	101
...	...	...	...
2020/07/01	84816	4641	79650
2020/07/02	84830	4641	79665

Table 1. COVID-19 data in China.

3.2 Parameter Setup

(1) Determination of window value W

Different time window sizes are used in the experiment, which scope is from 3 to 30. Figure 2 shows the cumulative forecast error of China under different time windows calculated according to formula (21). It can be found that there is a time window that minimizes the cumulative forecast error, that is, $W = 7$.

For every country in the data set, the respective optimal time window size is shown in Table 3.

Country or province	Date	Type of epidemic
China	2020/01/27–2020/07/02	COVID-19
Korea South	2020/02/20–2020/07/02	COVID-19
Italy	2020/02/26–2020/07/02	COVID-19
Spain	2020/02/26–2020/07/02	COVID-19
Brazil	2020/02/26–2020/07/02	COVID-19
Germany	2020/02/27–2020/07/02	COVID-19
France	2020/02/28–2020/07/02	COVID-19
Beijing province in China	2003/04/20–2003/04/20	SARS

Table 2. Data set description.

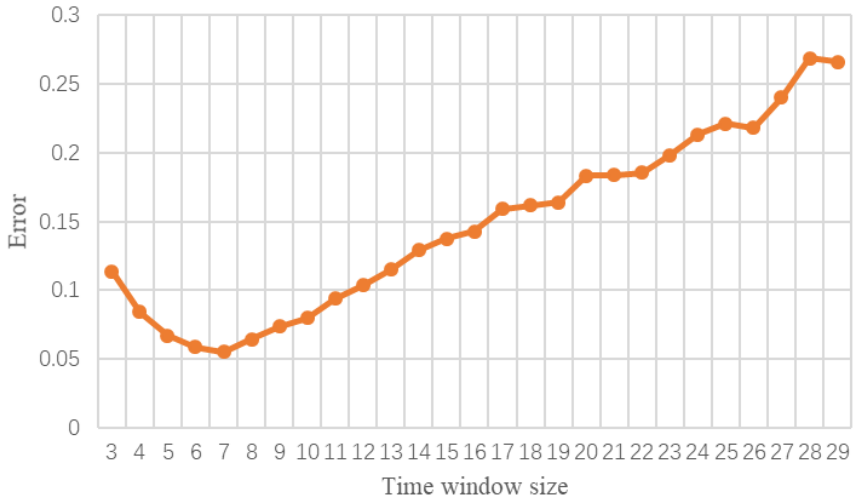


Figure 2. Changes in prediction error when the time window size is 3 to 29

(2) Parameter evaluation

After determining the appropriate time window size, Algorithm 1 is used to evaluate the model parameters. When using polynomial regression to predict the parameters $\hat{\beta}(t)$ and $\hat{\gamma}(t)$, we set initial order of the polynomial to 2, that is, $n = m = 2$. Because $\beta(t)$ and $\gamma(t)$ are non-negative, if their value is less than 0 in the regression calculation, we set them to 0. The stopping condition in the model solution process is $I(t) \leq 0$. Finally, we use model solution methods to predict the development trend of the epidemic.

3.3 Experiment and Result Analysis

In order to systematically evaluate and explain the scientific and effectiveness of the TW-SIR model, the following three problems is trying to answer: RQ1: Compared with the formula derivation method, how does the TW-SIR model perform in measuring the R0 in the process of epidemic? Is the proposed index of exponential growth rate useful? RQ2: How effective is the prediction of the TW-SIR model in epidemic COVID-19? RQ3: For each country, what is situation of the epidemic?

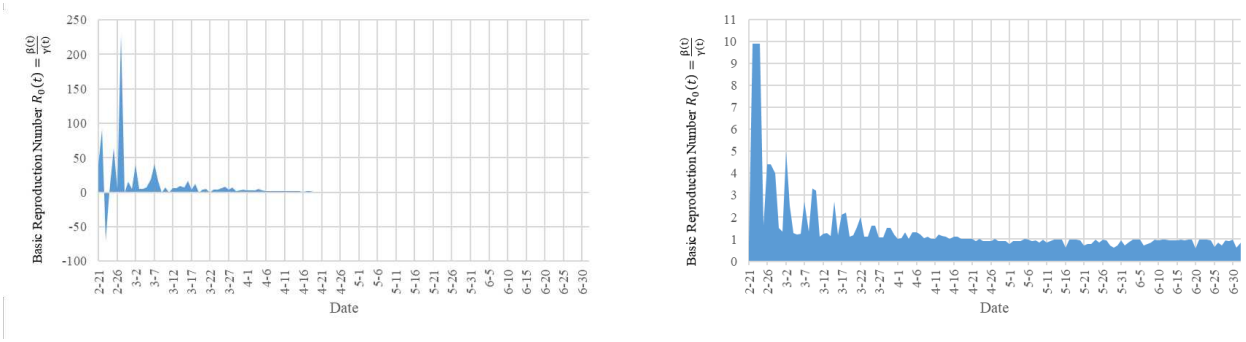
(1) RQ1 experiment results

In the epidemic model, a very important question is when the epidemic will end. As we known, if the R_0 in the population is greater than 1, the infection will spread exponentially. If R_0 is less than 1, the infection will spread only slowly, and it will eventually die out. In the TW-SIR prediction model, $R_0(t)$ is a time-dependent function. If $R_0(t) > 1$, the epidemic will spread quickly and infect a certain percentage of the total population N. On the contrary, if $R_0(t) < 1$, the epidemic will eventually be brought under control and end. Therefore, by observing the changes in $R_0(t)$ and predicting the future $\hat{R_0}(t)$, the development trend of the epidemic and whether the control measures of the epidemic are effective can be known. At the same time, in this paper, a new indicator exponential growth rate $Ex(t)$ is used, that is, the difference between $\beta(t)$ and $\gamma(t)$, to measure the exponential growth trend of the epidemic, which also reflects the changing trend of the epidemic. When $Ex(t) > 0$, it means that the infection speed of the epidemic is faster than the cure. On the contrary, the number of people infected by the epidemic

country	China	Korea South	Italy	Spain	Brazil	Germany	France	Beijing
Optimal time window size	7	7	4	4	6	6	5	7

Table 3. The optimal time window size of each country or region in the data set.

is gradually cured and the epidemic is gradually coming to an end. Here, our proposed TW-SIR prediction model is applied to data of COVID-19 in Italy from January 27 to July 2, 2020 to measure $R_0(t)$. We compare TW-SIR prediction model with the measurement method based on formula derivation proposed in [14]. Figure 3(a) shows the result of using the data to measure $R_0(t)$ method in the literature [14], and Fig. 3(b) is the result of using the TW-SIR model to measure $R_0(t)$. Data are from February 21, 2020 in the two figures. $R_0(t)$ in Fig. 3(a) has reached two hundred, and there are negative values, which is obviously not true. We can also see from Fig. 3(b) that the value of the $R_0(t)$ is much smaller and more in line with the actual situation. In addition, in Fig 3(b), it can be seen that there is a turning point of $R_0(t) < 1$ on April 19, 2020, that is, the epidemic situation in Italy reaches its peak at this moment. After April 19, 2020, $R_0(t)$ remains at a level less than 1, which means that the number of infected people $I(t)$ will decrease and will lead to the end of the Italian epidemic. TW-SIR model can accurately measure the time when $R_0(t) < 1$ and the measured value is close to the actual situation. At the same time, our results are similar to those measured in most literatures[23], which shows the effectiveness of TW-SIR model to measure $R_0(t)$.



(a) (b)

Figure 3. Basic reproduction number $R_0(t)$ in Italy

Similarly, Fig. 4 shows the results of TW-SIR model and formula derivation method in measuring the exponential growth rate $Ex(t)$. The exponential growth rate $Ex(t)$ calculated by the two methods can reflect the development and changes of the epidemic, and the overall trend is roughly the same, and both can measure the peak time of the epidemic. However, the $Ex(t)$ value calculated based TW-SIR model includes the value calculated based on the formula derivation method, which can more clearly reflect the change of the exponential growth rate.

Figure 4 The result of the exponential growth rate $Ex(t)$ in Italy from February 21 to July 2, 2020. (The dark green curve represents the measurement result of our proposed TW-SIR prediction model, and the light green curve is the formula-based method used in [14].)

In order to verify the accuracy of TW-SIR model in different country, data of COVID-19 in China is used to verify the accuracy of TW-SIR prediction model. Figure 5(a) shows the results of using historical data to measure $R_0(t)$ in China in the literature [14], and Fig. 5(b) is the result of using the TW-SIR model.

Figure 6 shows the results of TW-SIR model and formula derivation method in measuring China's exponential growth rate $Ex(t)$. This is similar to the previous Fig. 5. The exponential growth rate $Ex(t)$ calculated by our method can better reflect the development and changes of the epidemic.

In order to show that our method is applicable to different epidemics, Fig. 7 and Fig. 8 respectively show the change curves of R_0 and exponential growth rate Ex of the SARS in Beijing, China in 2003. Compared with the spread of COVID-19 in China, R_0 of SARS in the early stage of infection is about half of R_0 of COVID-19, and the exponential growth rate is about a quarter of that of COVID-19. This is consistent with the actual situation[24], indicating that COVID-19 spread more violently than the SARS in 2003.

(2) RQ2 experiment results

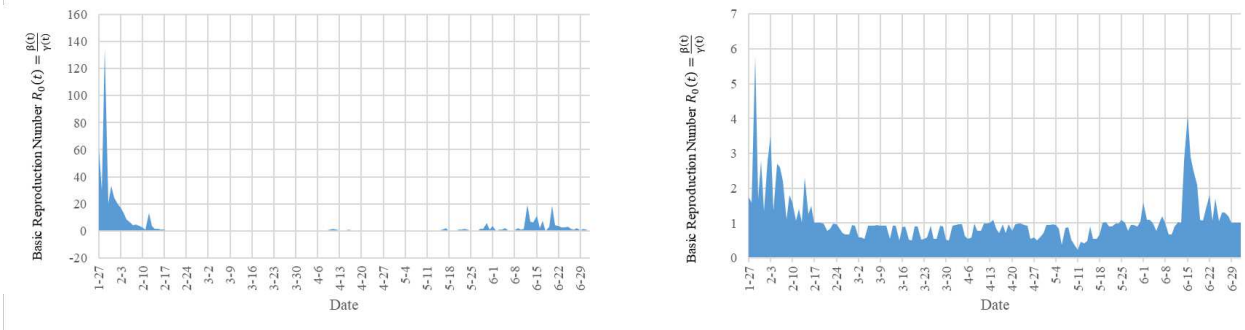
Figure 9 shows the measured $R_0(t)$ and the predicted $(\hat{R}_0(t)$ in Italy by using SIR-TW model. The blue curve is the measured $R_0(t)$, from February 26, 2020 to July 2, 2020. The gray curve is the predicted $(\hat{R}_0(t)$ from June 1, 2020 to July 2, 2020. The

Figure 4. The result of the exponential growth rate $Ex(t)$ in Italy from February 21 to July 2, 2020.The dark green curve represents the measurement result of our proposed TW-SIR prediction model, and the light green curve is the formula-based method used in [14].

red dotted line is the threshold value representing the $(\hat{R}_0(t) = 1$. We can see that R_0 in Italy was almost the same as R_0 in China in the early stages of the epidemic. From the figure that R_0 is a turning point around April 19, which means a peak of the epidemic. Compared with China, Italy has a relatively long time to enter the peak, which may be caused by different prevention and control strategies.

In Fig. 10, we show the exponential growth rate $Ex(t)$ measured by Italy and the predicted exponential growth rate $\hat{Ex}(t)$. The green curve is the measured exponential growth rate $Ex(t)$, from February 26, 2020 to July 2, 2020. The yellow curve is the predicted exponential growth rate $\hat{Ex}(t)$, from June 1, 2020 to July 2, 2020. From this graph, we can see that the exponential growth rate of the Italian epidemic has approached zero, which means that the peak of the number of infected persons in the Italian epidemic will come and the epidemic will be faded. Fig. 9 and Fig. 10 show that TW-SIR model accurately predicted the changes of $(\hat{R}_0(t)$ and $\hat{Ex}(t)$ from June 1 onwards, which shows that our parameter prediction method is effective.

In order to show the accuracy of our model, we show the prediction results of our model for the next day (single-day forecast) in Fig. 11. The orange curve in the figure represents the actual number of infections $I(t)$ in Italy, and the blue curve



(a) (b)

Figure 5. $R_0(t)$ in China.

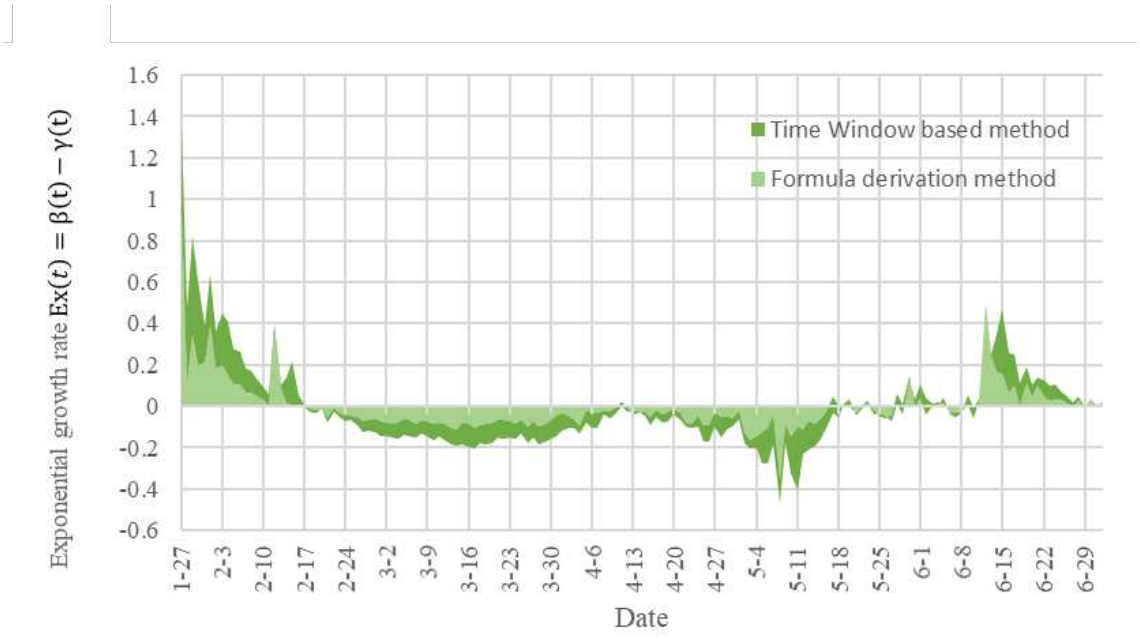


Figure 6. China's exponential growth rate $Ex(t)$ from January 27 to July 2, 2020.The dark green curve represents the measurement result of our proposed TW-SIR prediction model, and the light green curve is the formula-based method used in [14].

represents the predicted number of infections $\hat{I}(t)$. The figure shows that the predicted curve is very close to the actual data curve.

We further tested the accuracy of our prediction and calculated the error of the single-day prediction of the number of infected people, as shown in Fig. 12. The error rate of the predicted number of infected people is all within 5%, which shows that our model can accurately predict the number of infected people next day.

Judging from the results of applying TW-SIR model to the data of epidemic in China and Italy, the model can effectively measure the real-time changes of parameters during the development of the epidemic, including the Basic reproduction number of the epidemic and the exponential growth rate of the development of the epidemic, as well as the development trend of the epidemic follow up and forecast.

(3) RQ3 experiment results

Figure 13 shows the changes of $R_0(t)$ in some countries in the data set. Among them, the outbreak time of Spain, South Korea, Germany and France was almost the same, all of which started in mid-to-late February. As of July 2nd, their $R_0(t)$ is close to 1, which shows that the peak of the epidemic has passed, and the follow-up is just waiting for the end of the epidemic. Compared

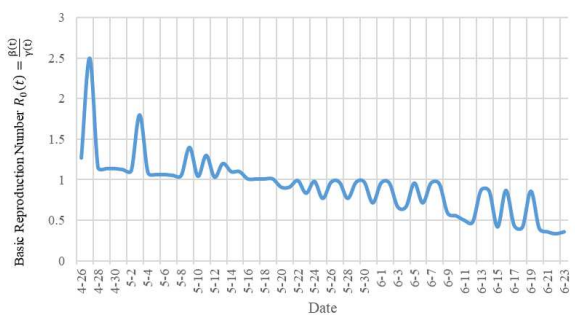


Figure 7

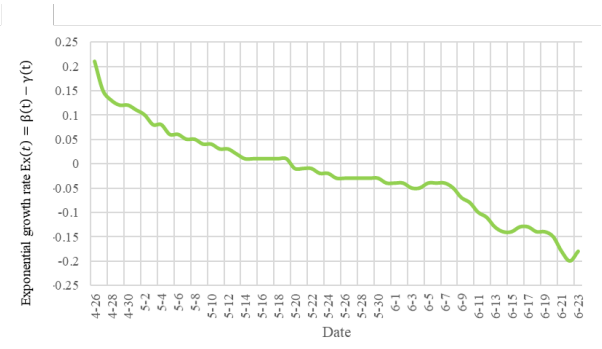


Figure 8

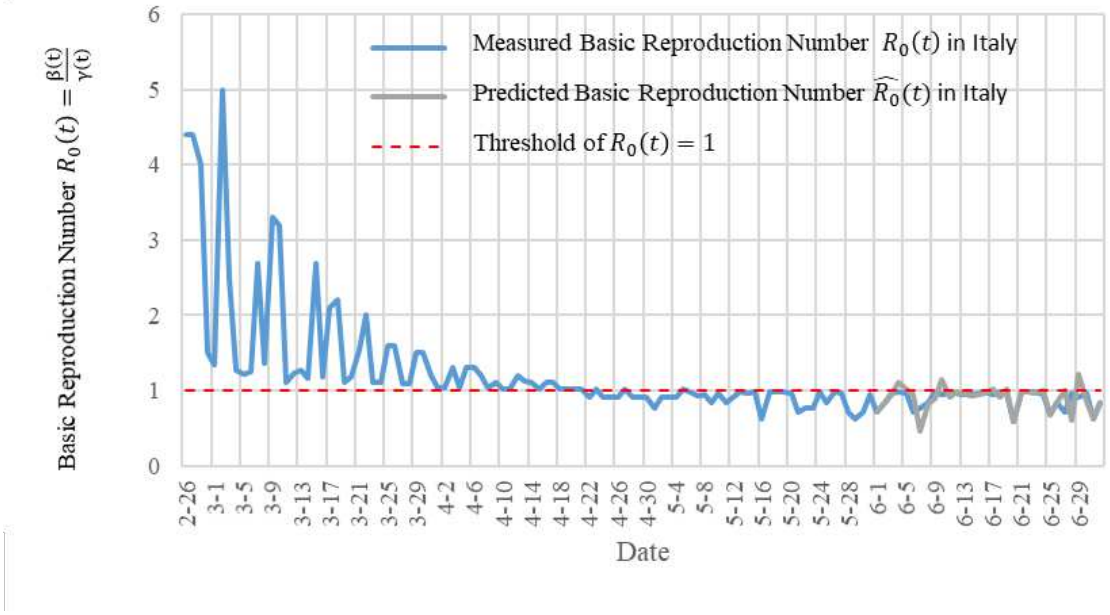


Figure 9. $R_0(t)$ and the predicted $(\hat{R}_0(t)$ in Italy measured by the TW-SIR prediction model.

with other countries, the outbreak time in Brazil was late, and it had not reached its peak as of July 2. Figure 14 shows the changes in the exponential growth rate $Ex(t)$ in some country in the data set. From here, we can see the development trend of the epidemic.

4 Conclusion

With the outbreak of the epidemic in other countries and regions, COVID-19 has swept the world. In order to forecasting trend of COVID-19, , an TW-SIR model is proposed in this paper, which is able to reflect the real-time trend of the epidemic in the process of infection for different areas, different policies and different epidemic diseases. A novel data expression - exponential growth rate - $Ex(t)$ of the epidemic is used to be better for people understanding the trend of epidemic. Machine learning methods are applied to predict the basic number of infections $R0$ and the exponential growth rate of the epidemic. We obtained COVID-19 data from the Johns Hopkins University and also used the 2003 SARS data in Beijing, China for verification. The numerical results Analysis shows that the model can effectively measure the real-time changes of parameters during the spread of epidemics, including the basic number of infections $R_0(t)$ and exponential growth rate $Ex(t)$, which reflects the exponential growth rate of epidemics and the model can be applied in different epidemics such as SARS. In general, the measurement of these parameters is of great significance for understanding the spread of COVID-19 and guiding the designation of control strategies and measures.

References

1. Zhao, S. *et al.* 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. *Int. J. Infect. Dis.* **92**, 214 – 217 (2020).
2. Sanche, S. *et al.* The novel coronavirus, 2019-ncov, is highly contagious and more infectious than initially estimated. *arXiv: Populations Evol.* (2020).
3. Pike, W. T. & Saini, V. An international comparison of the second derivative of covid-19 deaths after implementation of social distancing measures. *medRxiv* (2020).
4. Li, L. *et al.* Propagation analysis and prediction of the covid-19. *ArXiv* **abs/2003.06846** (2020).
5. Tang, B. *et al.* An updated estimation of the risk of transmission of the novel coronavirus (2019-ncov). *Infect. Dis. Model.* **5**, 248 – 255 (2020).
6. Hu, Z., Ge, Q., Li, S., Jin, L. & Xiong, M. Artificial intelligence forecasting of covid-19 in china. *arXiv: Other Quant. Biol.* (2020).

7. Yang, Z. *et al.* Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *J. thoracic disease* **12 3**, 165–174 (2020).
8. Friston, K. J. *et al.* Dynamic causal modelling of covid-19. *Wellcome open research* **5**, 89 (2020).
9. Naudé, W. Artificial intelligence vs covid-19: limitations, constraints and pitfalls. *Ai Soc.* 1 – 5 (2020).
10. Naudé, W. Artificial intelligence against covid-19: An early review (2020).
11. hua Liu, Z., magal, P., Seydi, O. & Webb, G. Predicting the cumulative number of cases for the covid-19 epidemic in china from early data (2020).
12. Peng, L., Yang, W., Zhang, D., Zhuge, C. & Hong, L. Epidemic analysis of covid-19 in china by dynamical modeling. *arXiv: Populations Evol.* (2020).
13. Sun, H. *et al.* Tracking and predicting covid-19 epidemic in china mainland. *medRxiv* (2020).
14. Chen, Y., Lu, P.-E. & Chang, C. A time-dependent sir model for covid-19. *ArXiv* **abs/2003.00122** (2020).
15. Fanelli, D. & Piazza, F. Analysis and forecast of covid-19 spreading in china, italy and france. *Chaos, Solitons, Fractals* **134**, 109761 – 109761 (2020).
16. Kucharski, A. *et al.* Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The Lancet. Infect. Dis.* **20**, 553 – 558 (2020).
17. Roda, W. C., Varughese, M., Han, D. & Li, M. Why is it difficult to accurately predict the covid-19 epidemic? *Infect. Dis. Model.* **5**, 271 – 281 (2020).
18. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. royal society london. Ser. A, Containing papers a mathematical physical character* **115**, 700–721 (1927).
19. Harko, T., Lobo, F. S. N. & Mak, M. K. Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Appl. Math. Comput.* **236**, 184–194 (2014).
20. Brauer, F. *et al.* Mathematical epidemiology (lecture notes in mathematics / mathematical biosciences subseries) (2008).
21. University, J. H. "covid-19" (2020). Available from: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.
22. Ministry of health of the people's republic of china (2003). Available from: https://web.archive.org/web/20030801083745/http://www.moh.gov.cn/zhgl/yqfb/index.htm.
23. Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of covid-19 is higher compared to sars coronavirus. *J. Travel. Medicine* **27** (2020).
24. Organization, W. H. Consensus document on the epidemiology of severe acute respiratory syndrome (sars) (2003). Available from: https://doi.org/10.1007/s12110-009-9068-2.

Acknowledgements

Funding: The works that are described in this paper are supported by NSF 61802120, Hunan Provincial Key Laboratory of Finance & Economics Big Data Science and Technology (Hunan University of Finance and Economics) 2017TP1025 and HNNSF 2019JJ50018, The scientific research project of Hunan Provincial Education Department No.: 18B480.

Author contributions statement

Zhifang Liao conceived the experiments, Peng Lan and Zhingning Liao conducted the experiments, Yan Zhang and Shengzong Liu analysed the results. All authors reviewed the manuscript.

Competing interests

The author(s) declare no competing interests.

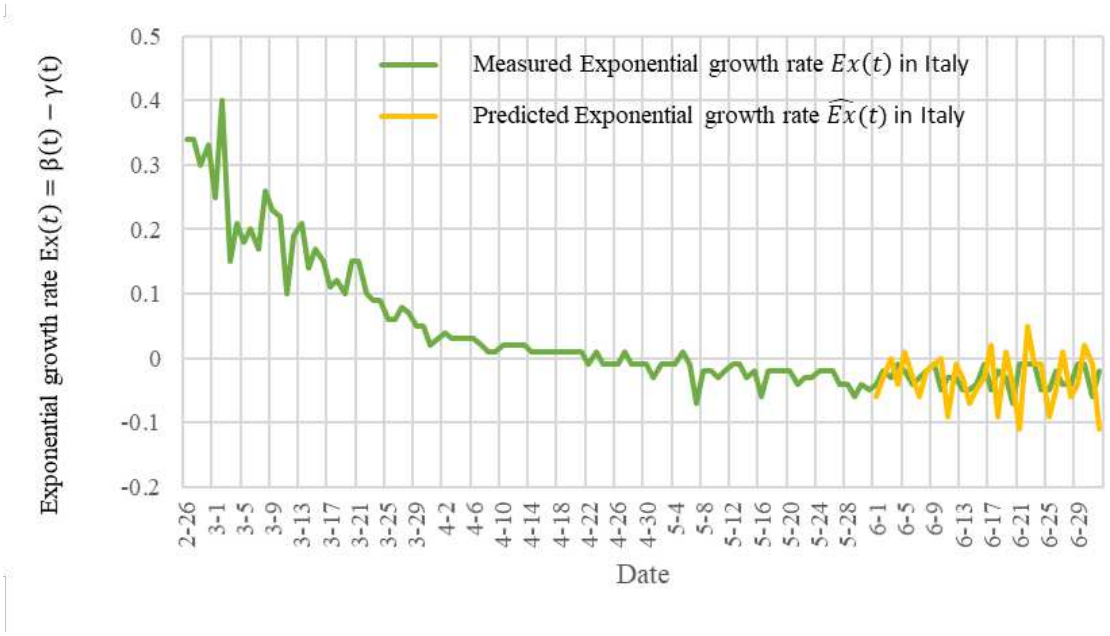


Figure 10. The basic number of infections $Ex(t)$ and the predicted basic number of infections $\hat{E}x(t)$ of COVID-19 in Italy measured by the TW-SIR prediction model.

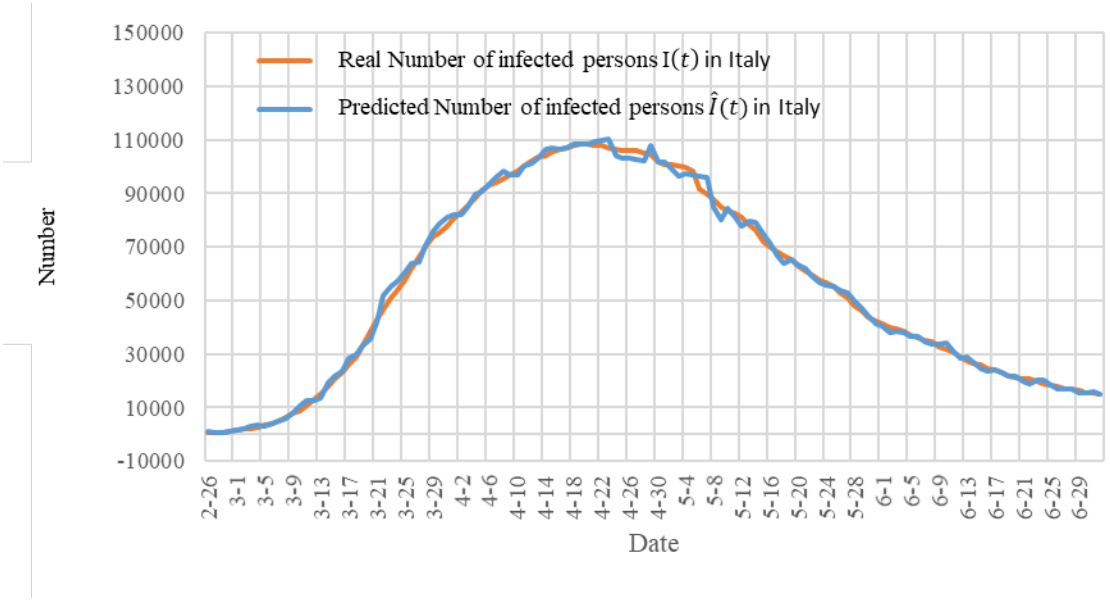


Figure 11. A single-day forecast of the number of infections in Italy. The orange curve represents the actual number of infections $I(t)$ in Italy, and the blue curve represents the predicted number of infections $\hat{I}(t)$.

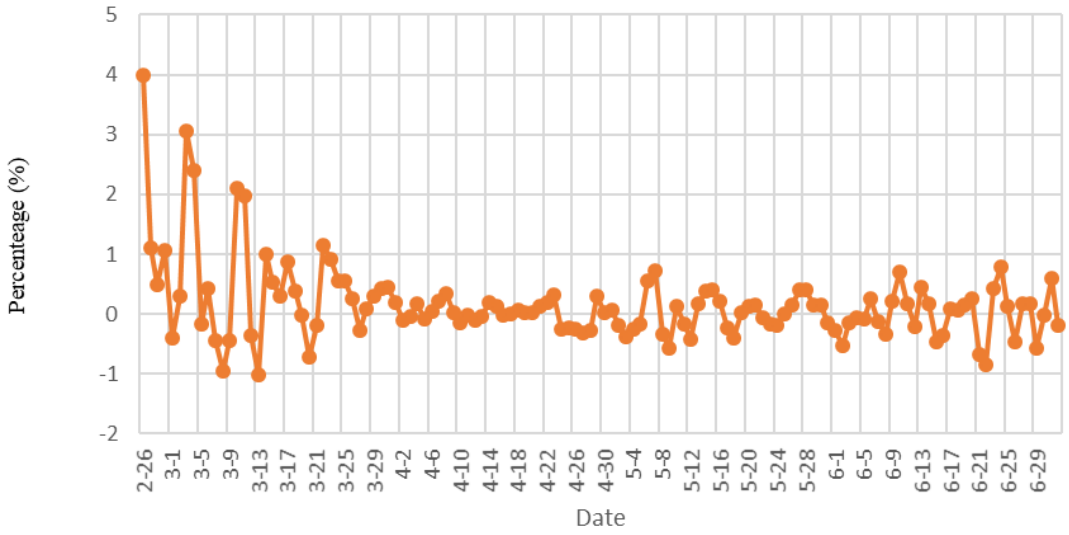


Figure 12. The forecast error of the single-day forecast of the number of infections in Italy.

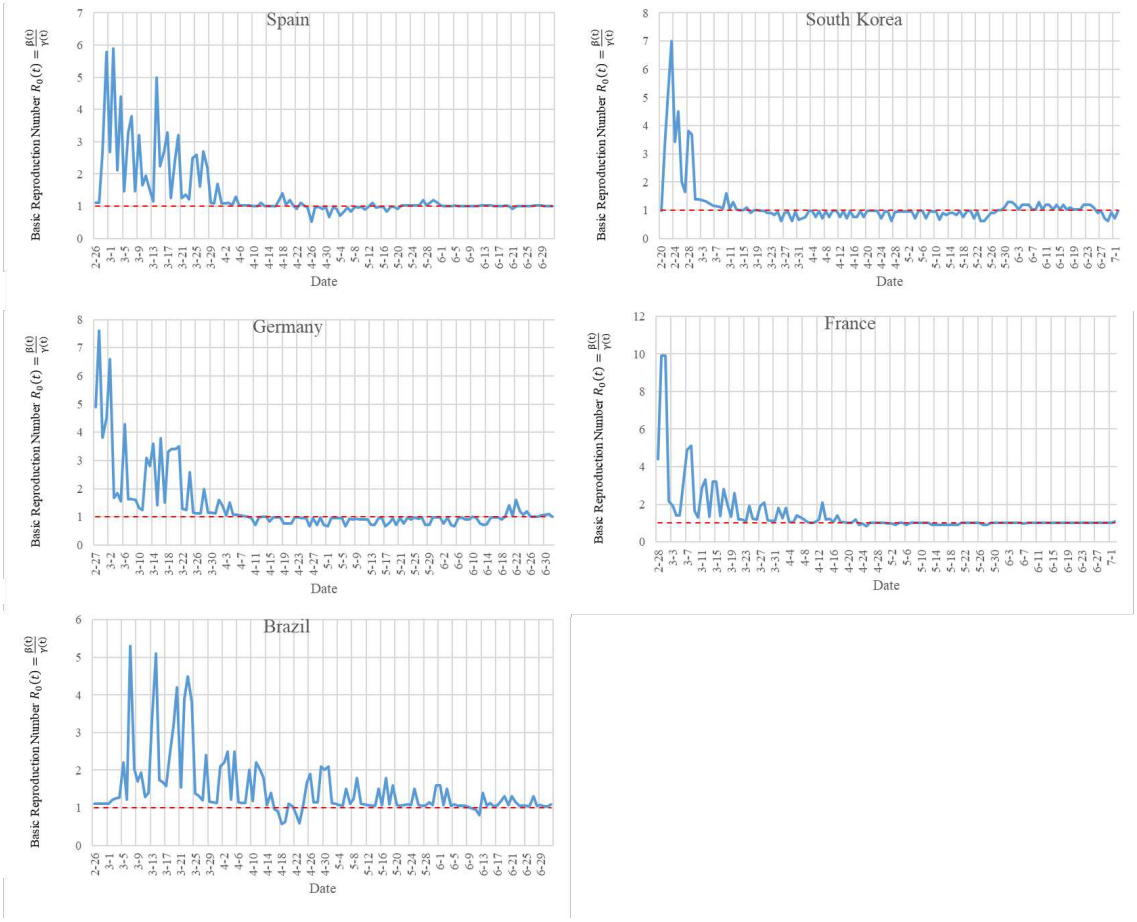


Figure 13. Changes of $R_0(t)$ of some countries.

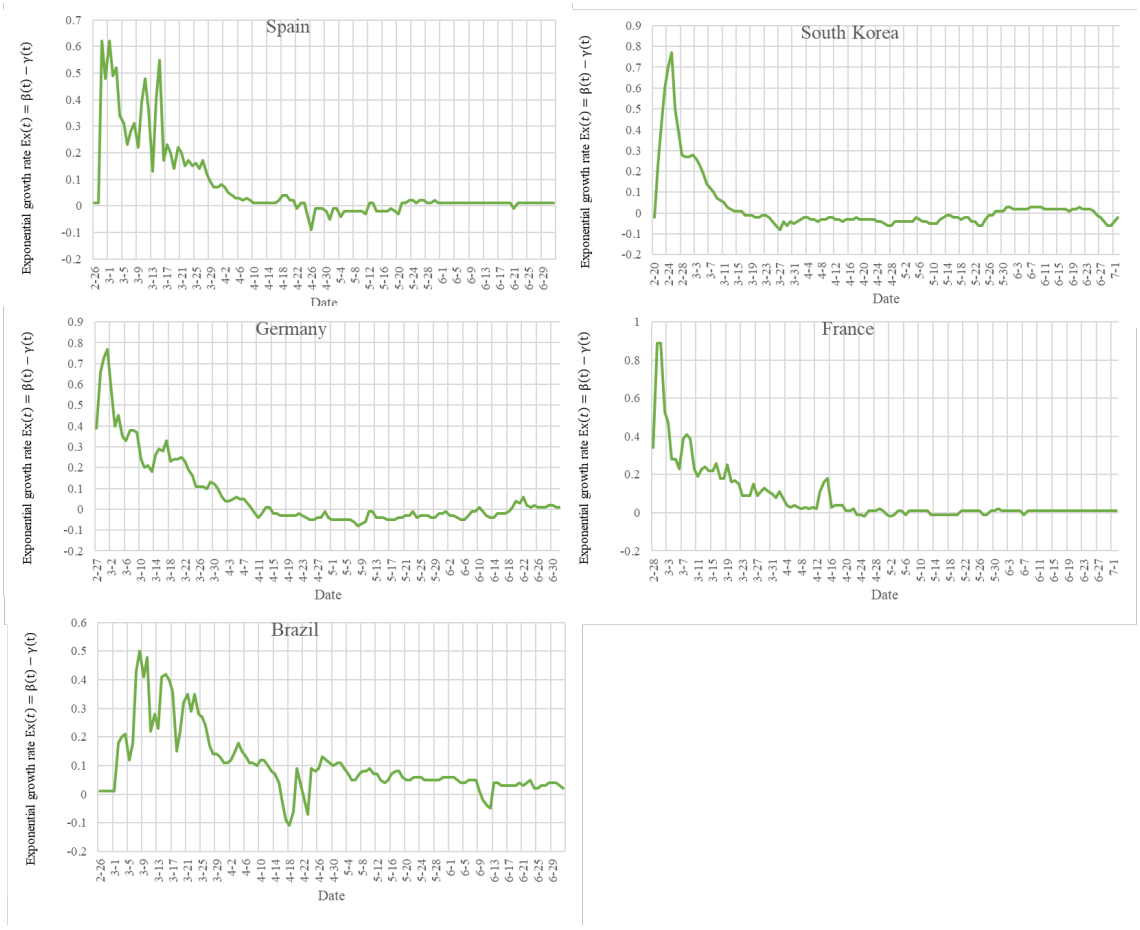


Figure 14. Changes in the exponential growth rate $Ex(t)$ of some countries.

Figures

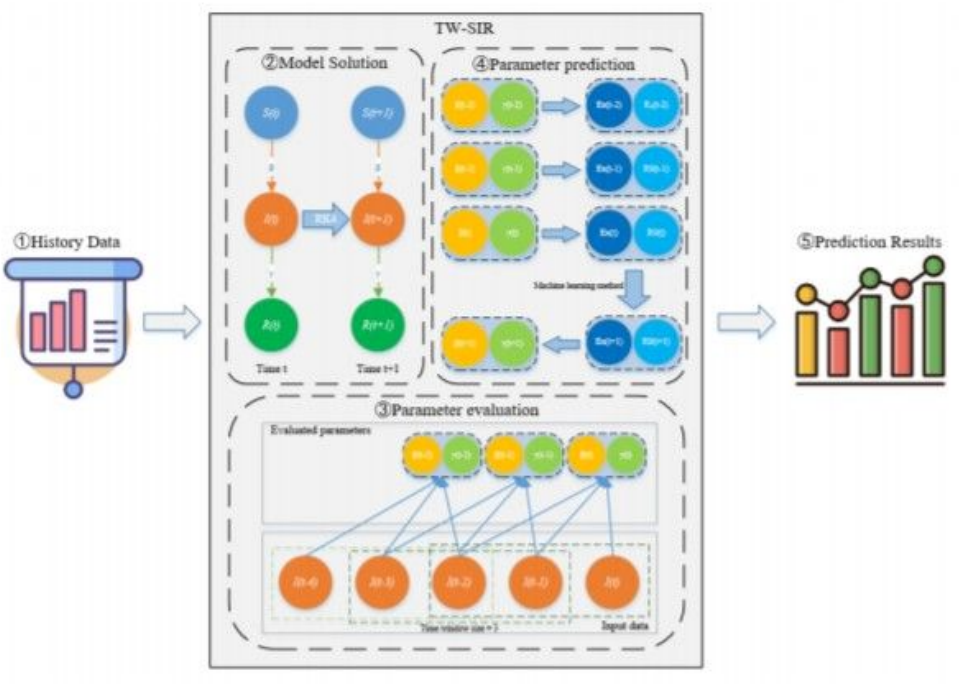


Figure 1

The composition of the TW-SIR model

Figure 2

Changes in prediction error when the time window size is 3 to 29

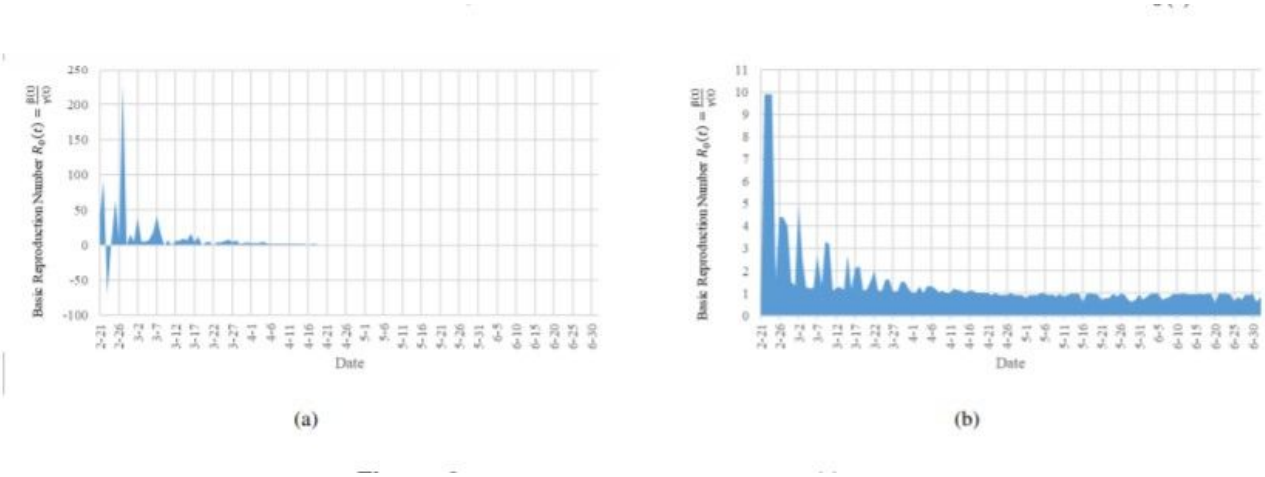


Figure 3

Basic reproduction number R0(t) in Italy

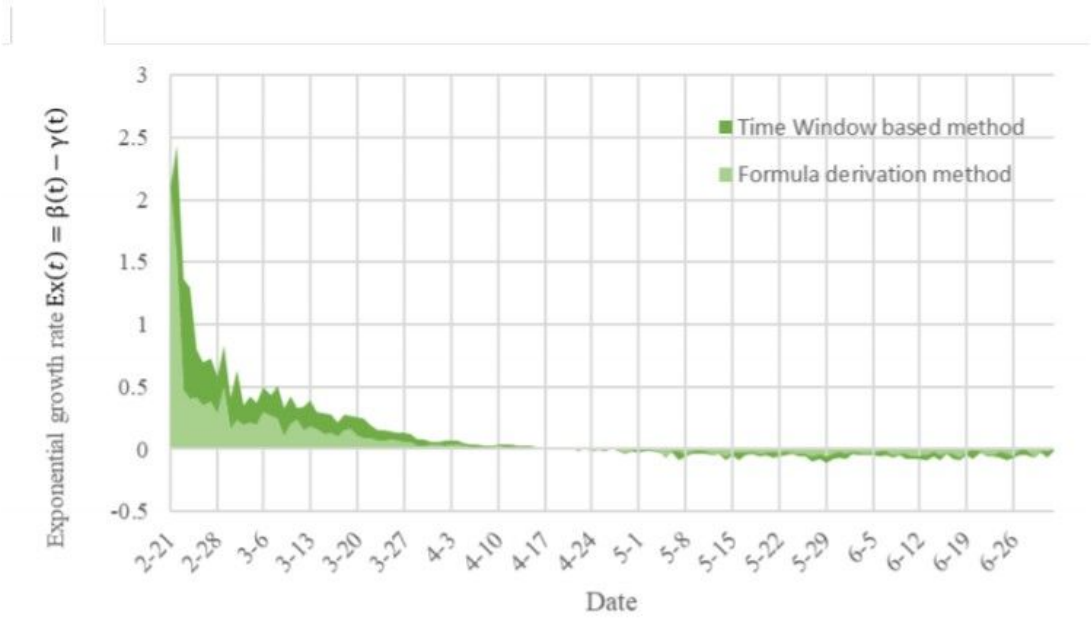


Figure 4

The result of the exponential growth rate Ex(t) in Italy from February 21 to July 2, 2020.The dark green curve represents the measurement result of our proposed TW-SIR prediction model, and the light green curve is the formula-based method used in [14].

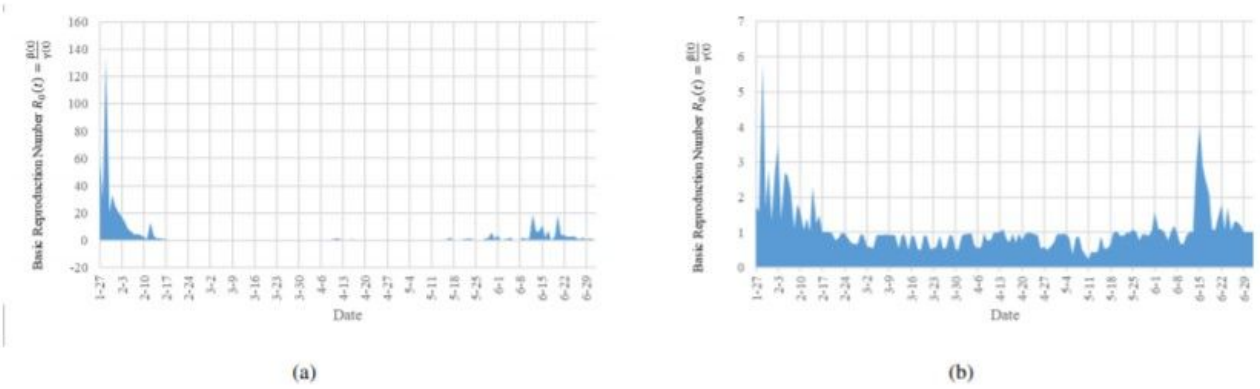


Figure 5

R0(t) in China.

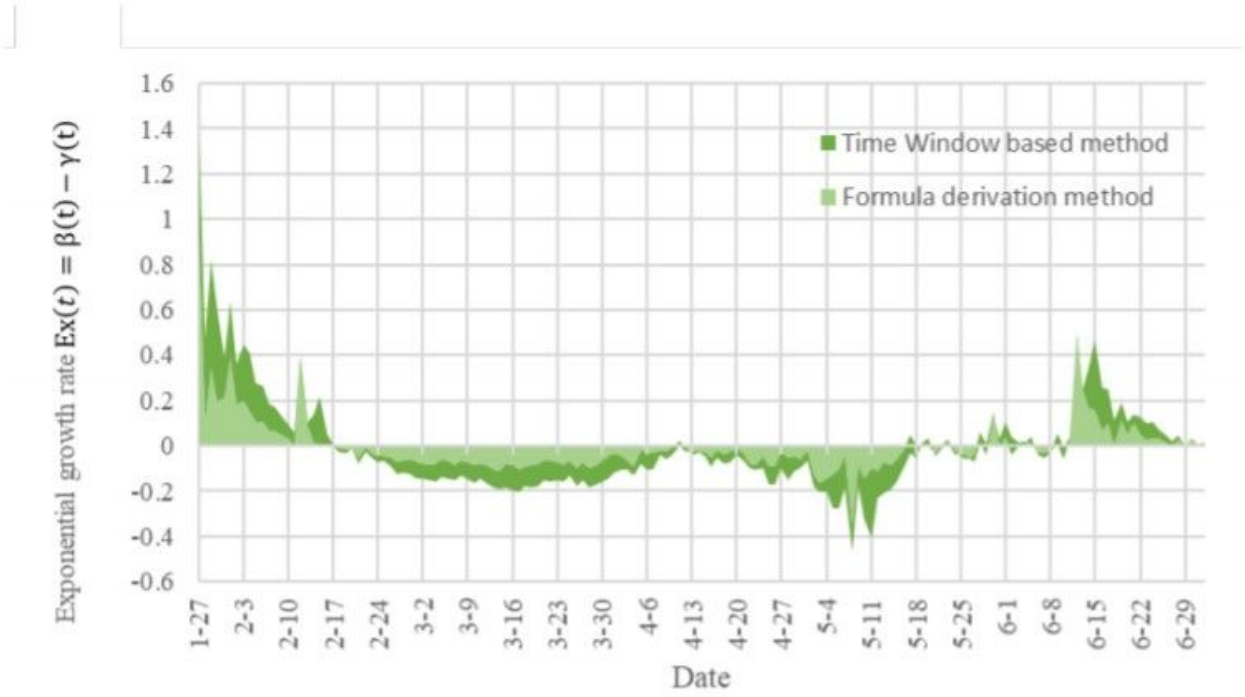


Figure 6

China's exponential growth rate Ex(t) from January 27 to July 2, 2020.The dark green curve represents the measurement result of our proposed TW-SIR prediction model, and the light green curve is the formula-based method used in [14].

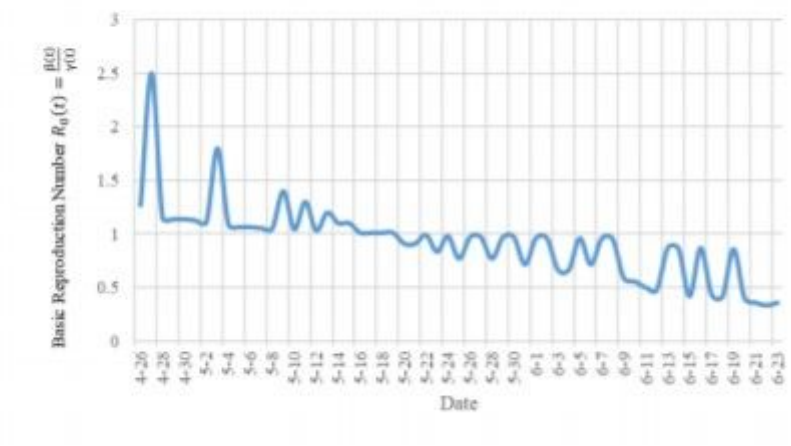


Figure 7

Legend not provided in this version

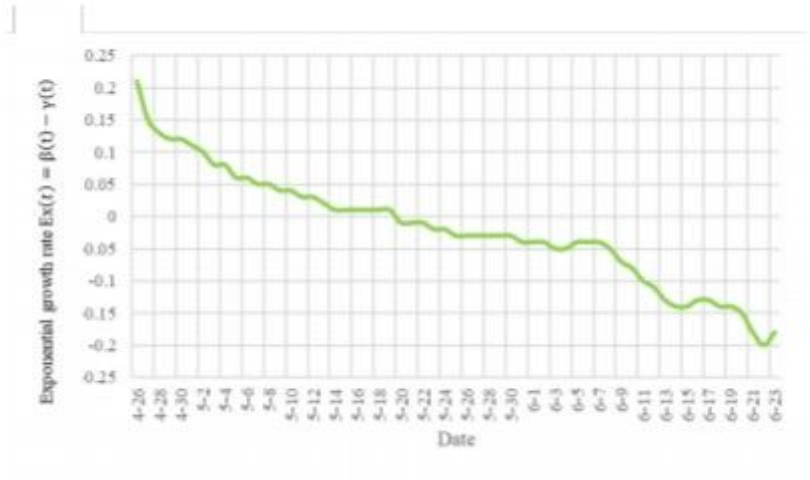


Figure 8

Legend not provided in this version

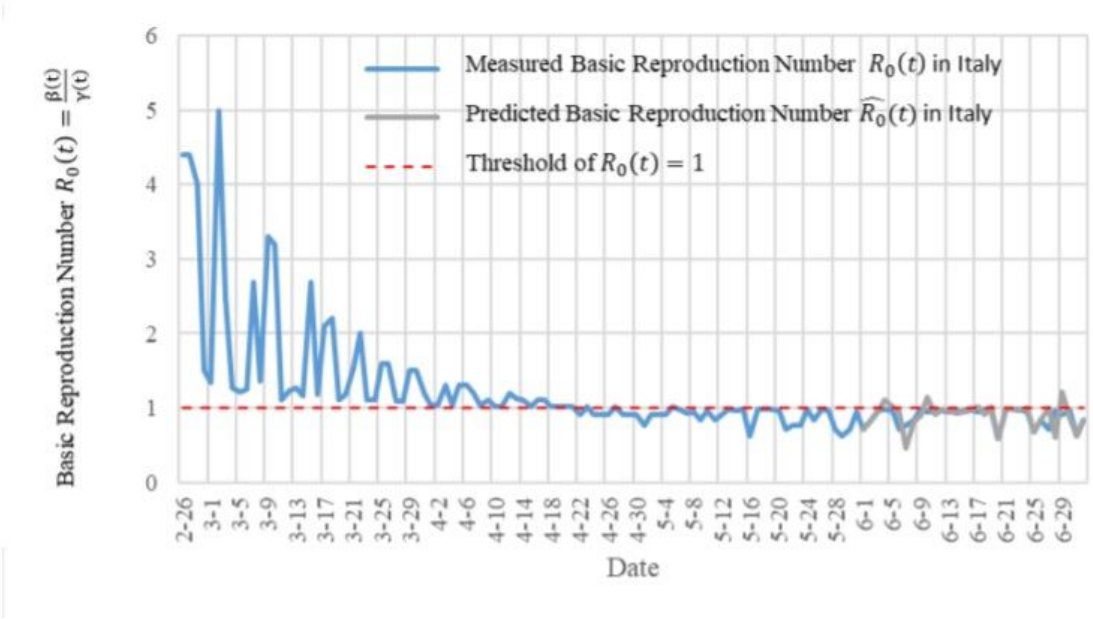


Figure 9

R0(t) and the predicted (R0(t) in Italy measured by the TW-SIR prediction model.

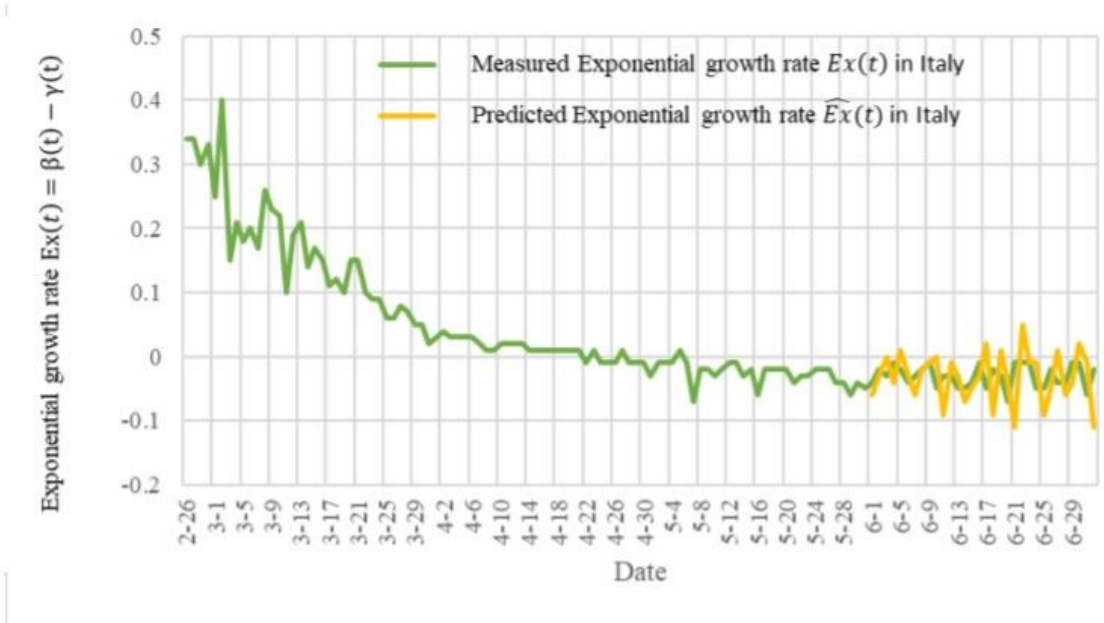


Figure 10

The basic number of infections Ex(t) and the predicted basic number of infections Exˆ (t) of COVID-19 in Italy measured by the TW-SIR prediction model.

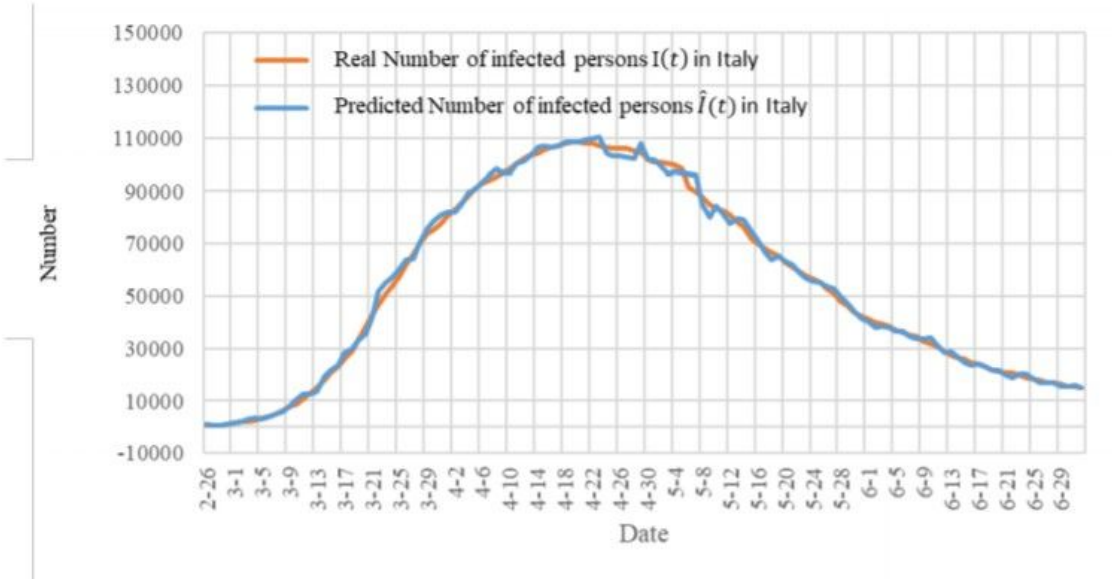


Figure 11

A single-day forecast of the number of infections in Italy. The orange curve represents the actual number of infections I(t) in Italy, and the blue curve represents the predicted number of infections ˆI(t).

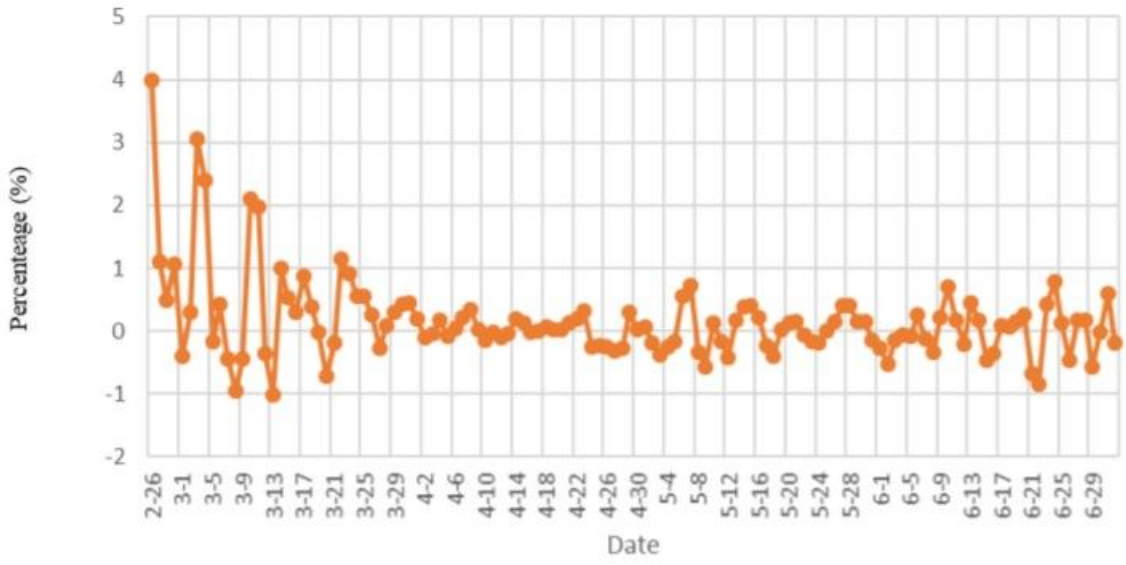


Figure 12

The forecast error of the single-day forecast of the number of infections in Italy.

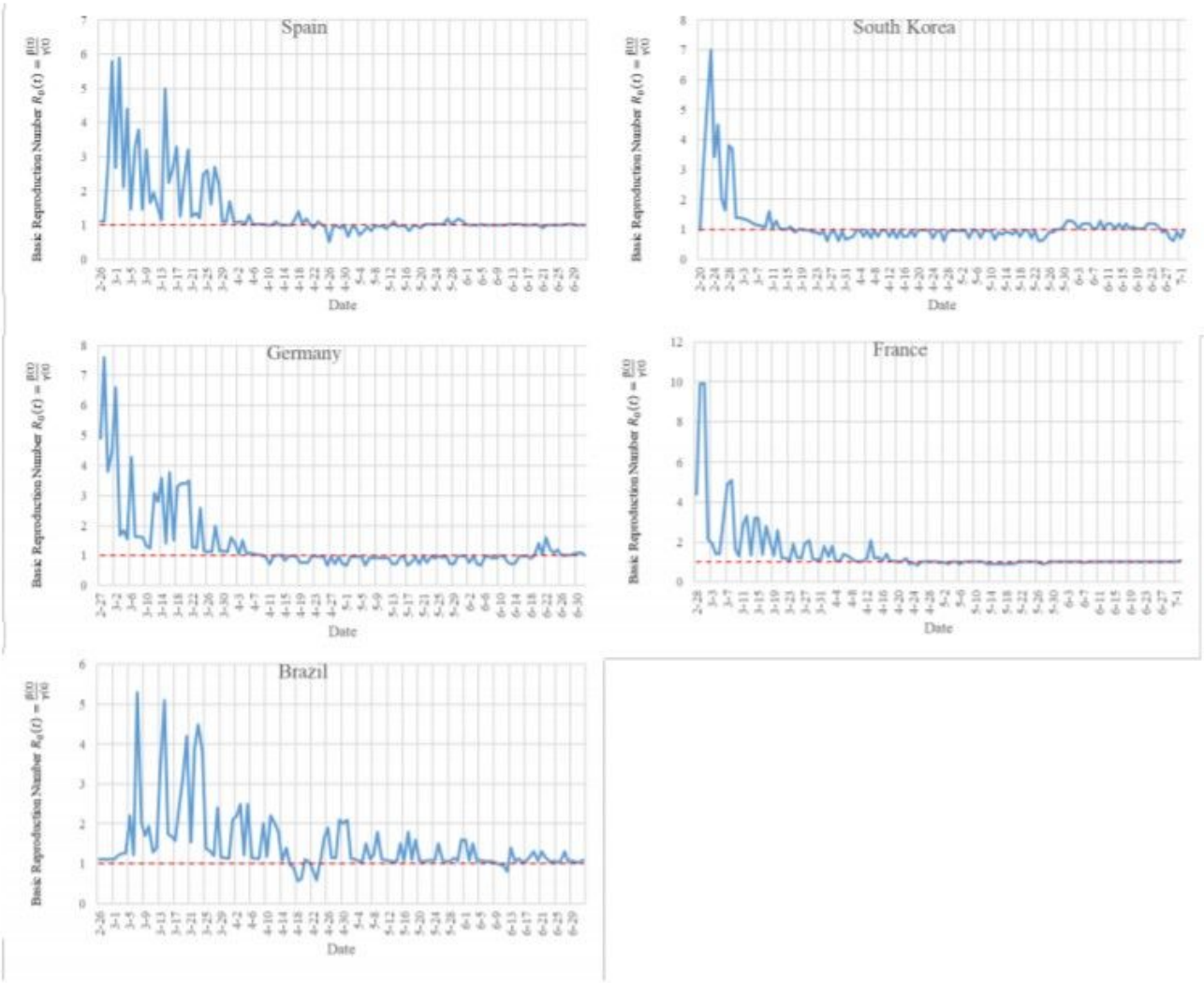


Figure 13

Changes of R0(t) of some countries.

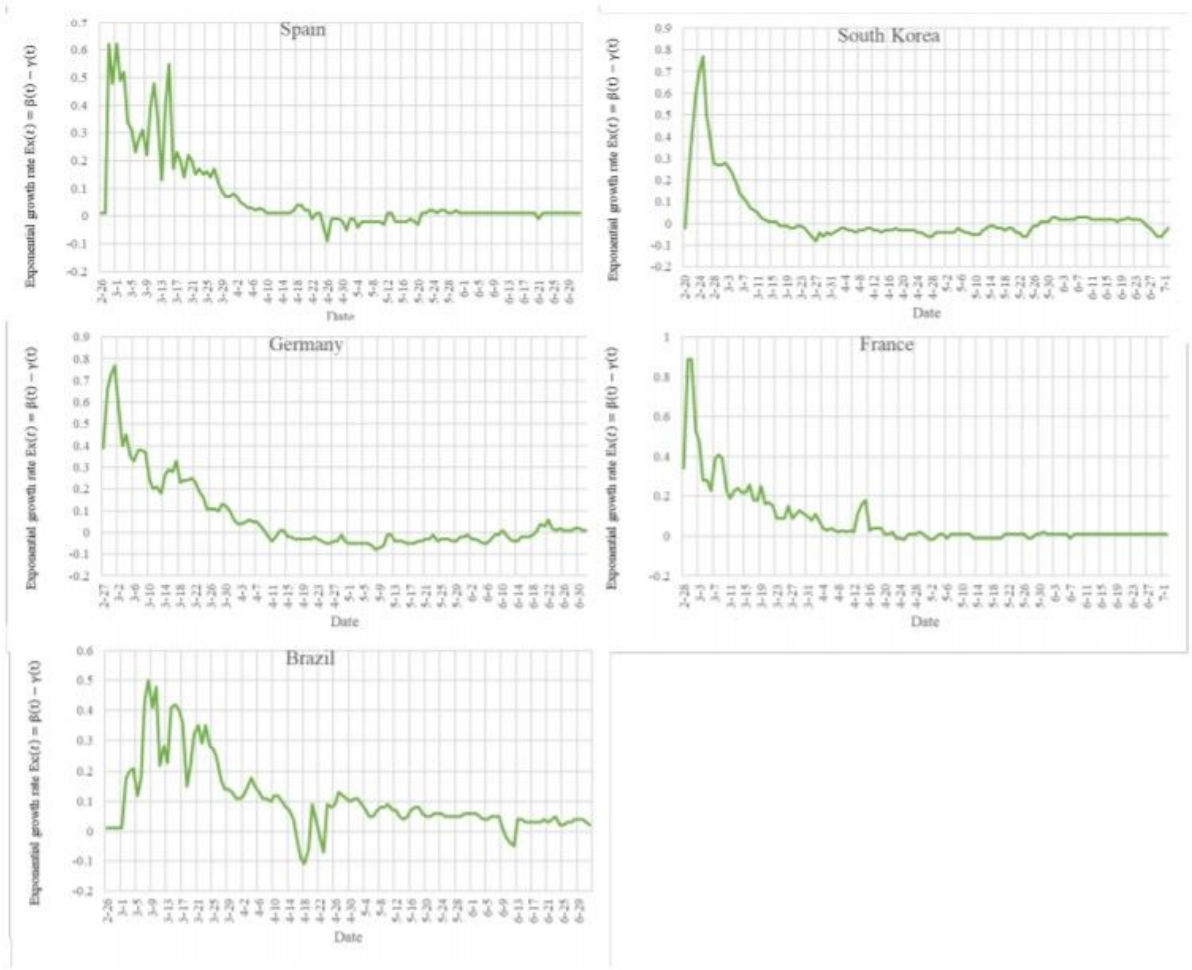


Figure 14

Changes in the exponential growth rate Ex(t) of some countries.