Research on COVID-19 Global Forecast based on SIR Model-ML regression

He Wei, Zeng Qingtao, Lu Likun, Li Yeli
School of Information Engineering, Beijing Institute Of Graphic Communication, Beijing 102600, China
871613503@qq.com, zengqingtao@bigc.edu.cn, lklu@bigc.edu.cn, liyl@bigc.edu.cn

Abstract. Under the background of the global COVID-19 pandemic, the global COVID-19 epidemic spread was modeled and analyzed, and its future trend was predicted. The dataset covers 163 countries and almost 2 full months from 2020, which is enough data to get some clues about the pandemic. The epidemic situation was predicted by the fusion of SIR model and ML regression, and the results showed that the model analysis was basically consistent with the real performance of the epidemic development. Finally, the development stage and trend of the epidemic situation are evaluated to provide a basis for the government to formulate relevant epidemic prevention policies.

1. Introduction
Infectious diseases not only seriously threaten people's lives, but also seriously hinder the development of human society, such as smallpox in ancient Greece, plague and cholera in ancient Rome and Europe, SARS virus atypical pneumonia in China in 2003 and Ebola in Africa in 2014 [4]. The COVID-19 [3] epidemic has spread all over the world, and as far as the current situation is concerned, there is still a long way to go before the epidemic is completely brought under control. The sudden global pandemic of COVID-19 has spread to 217 countries (regions) by the end of June 2020. the total number of confirmed cases worldwide has exceeded 10 million, with more than 500,000 deaths and a mortality rate of about 5%. The global COVID-19 epidemic continues to spread, and the number of new diagnoses every day is still about 100000. All countries have implemented large-scale quarantine and city closure measures to prevent and control the spread of the epidemic.

Although the SIR model [2] has a good estimation of the basic reproductive number and the early trend of the epidemic, the dynamics-based prediction model still has limitations in practical application: first, the warehouse model can not accurately estimate the spread of the virus in the open flow environment; second, this kind of model lacks the introduction of the influence of external environmental stress in the evaluation of relevant parameters. Third, the constant assumption of disease transmission ability and cure probability is not consistent with the actual situation. Therefore, this kind of model can not make a long-term and accurate analysis of the epidemic trend.

In summary, this paper combines the accuracy of SIR model to describe the propagation mechanism and the effectiveness of ML regressions to estimate external factors, and constructs a SIR-ML regressions algorithm model for real-time prediction based on the fusion of SIR model and ML regressions. In this model, the final state of transmission is further divided into cure state and death state, and the time-varying law of probability from infection state to these two states is predicted for more accurate prediction. Experiments show that the SIR-ML regressions algorithm model can achieve accurate real-time prediction.
2. SIR Model

2.1 SIR Model

This paper studies and collects the data dissemination of four countries. COVID-19 has a very high transmission speed between people. Based on the epidemic data, mathematical models can be used to effectively mine various epidemiological parameters of COVID-19 and extract features from these data. In this study, the traditional warehouse SIR model is improved, and the Gauss function is used to simulate the changes of daily infection rate and daily isolation rate. According to the actual epidemic data of four countries and regions, the evolution of COVID-19 epidemic infection, prevention and control is deduced, and the basic reproduction number and effective reproduction number of COVID-19 are inferred. So I'm going to use one of the most famous epidemiological models: SIR.

1. Susceptible (S). These people themselves are not infected with the disease, but when they encounter people who carry the disease, they are easily infected.

2. Infected (I). Has been infected with the disease.

3. Recovered/Deceased (R). People who are infected with the disease have two conditions, either lucky to survive and develop immunity to it, or die.

![Fig.1 SIR model](image)

SIR Model equations

\[
\frac{dS}{dt} = -\frac{\beta SI}{N}
\]

\[
\frac{dI}{dt} = \frac{\beta S}{N} - \gamma I
\]

\[
\frac{dR}{dt} = \gamma I
\]

Where \( \beta \) (infection rate) indicates the probability that the susceptible person will be infected with the infectious disease after contact with the infected person, \( \gamma \) (Cure rate) it indicates the probability that the susceptible person will recover after a period of time.

2.2 Implementing the SIR model

As a result of artificial isolation and prevention and control measures, incubation carriers are actually divided into isolated and unisolated parts. The SEIR model[2] only describes the process of infectious diseases from the perspective of infectious disease pathology, but can not better describe the isolation measures in the actual process of the development of infectious diseases and the transformation relationship of all kinds of people under novel coronavirus system. In view of this shortcoming, the model is improved: the incubation period is subdivided into isolated part and unisolated part. Solve the system of differential equations describing the model.
In Europe as a whole, Italy and Spain have the worst epidemic and the highest infection rates. But Britain is a little different. Britain, which has a strait across the European continent, makes the mobility of the population of other countries in the UK not so obvious. Let's continue to study the situation in Singapore, and China is an Asian country. First, let's take a look at the spread of confirmed infections in four countries:

![Fig.2 SIR Model](image)

![Fig.3 Infection of confirmed cases in four countries](image)

3. ML regressions

3.1 ARIMA Model

In the time series (Time-series Approach) analysis method proposed by Box and Jenkins of Wisconsin University in the United States, the summation autoregressive moving average model (Autoregressive Integrated Moving Average Model, ARIMA) is mainly used to analyze non-stationary time series without seasonal variation. ARIMA model includes autoregressive (Auto Regressive, AR) model,
moving average (Moving Average, MA) model, or autoregressive moving average model ((Auto Regressive Moving Average, ARMA) model.

ARIMA (p, d, q) The structure of the model is as follows. Where p is the number of autoregressive terms, Q is the number of moving average terms, and d is the number of differences made when the time series becomes stationary.

\[
\begin{align*}
\Phi(B)\nabla^d x_t &= \Theta(B)\epsilon_t \\
E(\epsilon_t) &= 0, \quad Var(\epsilon_t) = \sigma^2, E(\epsilon_s\epsilon_t) = 0, s \neq t \\
E(\epsilon_s\epsilon_t) &= 0, \forall s < t \\
\nabla^d &= (1-B)^d \\
\Phi(B) &= 1-\phi_1 B - \cdots - \phi_p B^p \\
\Theta(B) &= 1-\theta_1 B - \cdots - \theta_q B^q
\end{align*}
\]

Fig.4 Predicted Values for the next 15 Days

The accuracy of any prediction model depends on reliable data. COVID-19 incubation period fluctuates in a large range, and the infectious ability of asymptomatic infection is still uncertain, so we have a simple linear regression.
4. Conclusion
The ARIMA model has a high degree of fit for the epidemic trend of novel coronavirus, and can be used for the short-term prediction of COVID-19 monitoring data series, which can provide reference for the prevention and control effect and development trend of the epidemic, and provide a certain policy basis for epidemic intervention decision-making in the next few days. Models tailored specifically for epidemic spreading (i.e. SIR and its versions) are designed to reproduce a certain
phenomenology, in order to understand the underlying mechanics of a contagion process. They might eventually help to find some features or parameters that are particularly important for the model’s fitting, but by no means they should be confused with scientific epidemic models.

According to the evaluation of the epidemic situation in various places, the number of people diagnosed by simulation is in good agreement with the actual number of people diagnosed; the development trend of the epidemic situation in different areas is predicted; and the basic reproduction number of the epidemic situation is calculated. There is still a deviation between the simulation results and the reality, and the reasons are that it is difficult to estimate the parameters such as the conversion rate and screening rate of suspected patients, and does not take into account the stage of the development of the epidemic and other factors. Therefore, to make a more accurate analysis and prediction of the development trend of the COVID-19 epidemic, we need more data as support, and dynamically adjust the values of the parameters in the model in stages.

Acknowledgments
Beijing science and technology innovation service capability construction project (PXM2016_014223_000025)

References
[1] Kermack W O Mckendrick A G. Contributions to the Mathematical Theory of Epidemics[J]. Proceedings of the Royal Society, 1927, 115(05):700-721.
[2] Ma Z, Jin Z. The stability of an SIR epidemic model with time delays[J]. Mathematical Biosciences and Engineering(Online), 2005, 3(1):101-109.
[3] Guan W J, Ni Z Y, Hu Y, et al. Clinical characteristics of 2019 novel coronavirus infection in China[J]. New England Journal of Medicine, 2020.
[4] Huang Desheng, Guan Peng, Zhou Baosen. Study on the fit of SIR model to the epidemic law of SARS in Beijing[J]. Chinese Journal of Disease Control & Prevention, 2004, 8(5):398-401.
[5] COOKE K L, DRIESSCHE V D P. Analysis of an seir epidemic model with two delays[J]. Journal of Mathematical Biology, 1996, 35(2):240-260.