Comparison of the Susceptible-Infected-Recovered Model and Grey Forecasting Model on Predicting the Number of Different Categories in COVID-19

Runhan Yang

Abstract

Humans have been fighting against the virus, the first batch of inhabitants of this planet, for a long time. Since COVID-19 hit Wuhan, China, three years ago, research on the dynamic of this pandemic has developed rapidly all over the world. Many classical models, like the SIR model, simulating epidemics have breakthroughs and perform much better. Machine learning, a newly developed approach, has an impressive classification, forecast, and clustering performance. Many researchers have applied machine learning algorithms to analyze the dynamic of COVID-19.

This work uses the modified SIR model and grey forecasting model, representing the machine learning algorithm, to predict the tendency of COVID-19 in the following year in China, the UK, and the US. Generally, both models give a good simulation of different countries, which fits the present national policies. However, the result shows that neither of the two models is perfect, and each algorithm has drawbacks and limitations from their basic logic after further analysis. Fortunately, the combination of the better result from the two models offers a much more reliable prediction than any single one.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

Runhan Yang
December 19, 2022

Acknowledgements

I would like to express my sincere gratitude to the lecturer of this course, Gongqiu Zhang, who delivered the whole lecture this semester. With the systematic knowledge he offered in class, I devised the idea to optimise the previous model and accomplished this project.

I would also like to thank the technical and support staff in this course, especially our teaching assistant, Fangchen Yu. He answered my questions on time and gave useful suggestions for my study.

I would like to show my deep appreciation to my friends and family. Thanks to Mengqi Su, who gave me her best support to debug my main code. Thanks to Kai Yu, who accompanied me online to finish the program. Thanks to my mother and father for the excellent care and encouragement to me during that period.

I wish to extend my special thanks to my bed because I came up with the main idea when I was dreaming.

Contents

Chapter 1

Introduction

As World Health Organization (WTO) reported in March 2021, COVID-19 has become one of the most terrible pandemics on this planet, which has spread all over the world[1]. This pandemic brings severe challenges to a broader range of industries than just medical institutions, economics, and pubic policies[2, 3, 4]. Billions of people in the world must stay home to prevent the spread of the virus under challenging situations required by the national government. In the meantime, due to the deficiency of complementary medicine and vaccine for COVID-19, government officials have made an effort to control the pandemic, including forcing people to work or study online and forbidding travel[5].

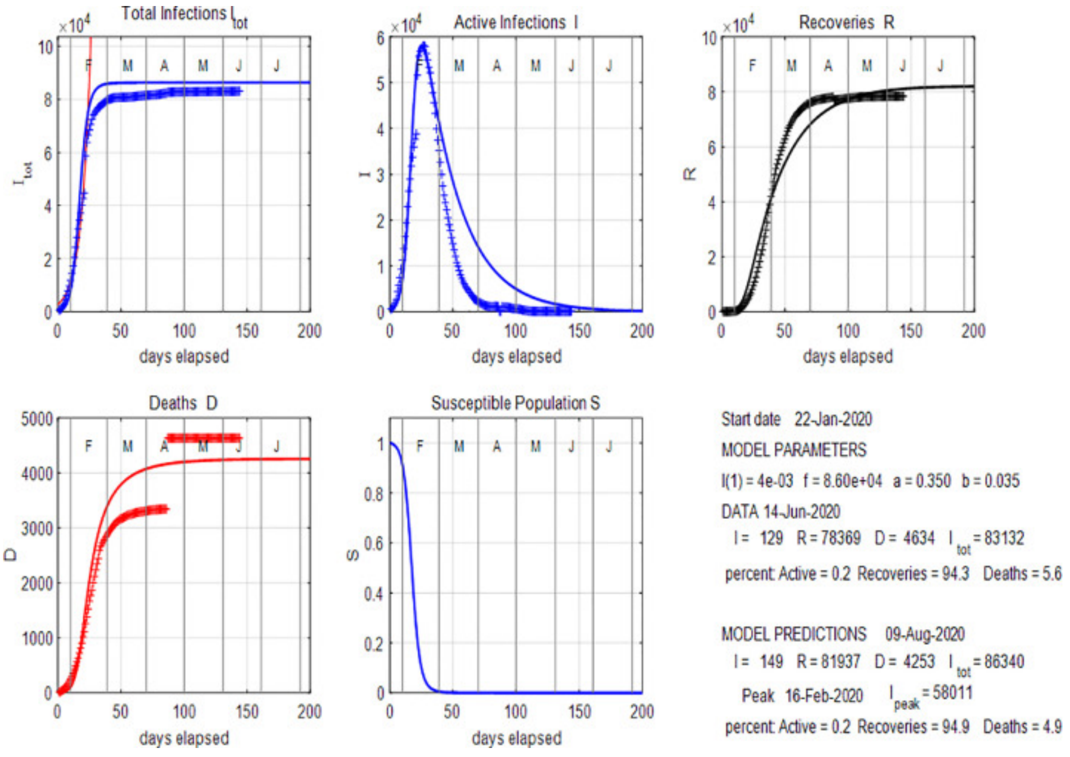
Against this background, it is important for policymakers to have a general perspective of the spread of COVID-19, especially for the number of infected, recovered, and dead people daily, to take action against the pandemic more quickly and accurately. There are three different approaches to predicting the dynamic, compartmental models, statistical methods, and ML-based methods[6, 7]. Compartmental models subdivide a population into mutually exclusive categories, with dynamical equations explaining the transitions among categories. The classical Susceptible-Infected-Removed(SIR) model simulates the system's dynamic with the help of differential equations, which was established in 1927 by Kermack-Mckendrick[8, 9]. Statistical methods extract general statistics from the data to fit mathematical models that explain the evolution of the epidemic, which will not be covered in this paper.

However, since this traditional model was designed about a hundred years ago, it has its drawback in prediction and limitation of input data. Machine learning models are widely used in almost every field of computer science. They have a terrific performance on forecasting, classification clustering, and decreasing the dimension, which can all be applied in health care[10, 11, 12]. In this paper, the grey forecasting model, one of the machine learning algorithms, is used to predict the future number of infected, recovered, and death.

Chapter 2

Literature review

A survey conducted in 2020 has augmented the classic SIR model with the ability to accommodate surges in the number of susceptible people, supported by the recorded data from China, America, and many other countries and regions to predict the spread of COVID-19. The model prediction could be fitted to the published data reasonably well, with some fitting better than others[13]. The model constructed by this work also reveals the personality of prediction from different countries. For example, in reality, the number of infected decreased more sharply than predicted by the model, which should be attributed to the successful policy of the Chinese government.



Start date 22-Jan-2020

MODEL PARAMETERS

I(1) = 4e-03 f = 8.60e+04 a = 0.350 b = 0.035

DATA 14-Jun-2020

I = 129 R = 78369 D = 4634 I_{tot} = 83132

percent: Active = 0.2 Recoveries = 94.3 Deaths = 5.6

MODEL PREDICTIONS 09-Aug-2020

I = 149 R = 81937 D = 4253 I_{tot} = 86340

Peak 16-Feb-2020 I_{peak} = 58011

percent: Active = 0.2 Recoveries = 94.9 Deaths = 4.9

Figure 2.1: China: Model predictions for the period from 22 January to 9 August 2020 with data from January to June, 2020. The data show a discrete jump in death D in mid-April.

The above survey mentions that due to the present government policy, there would be some errors between the prediction and real-world data. For example, strict isolation approaches will cause a decrease in the infection rate. Due to this drawback, a better model involving a machine learning algorithm, SIMLR model, is implemented to predict the dynamic of COVID-19[6]. As Xin et al. stated, a machine learning algorithm consists of four main steps [14]:
1. extract the features;
2. select the appropriate machine learning algorithm;
3. train the model and then select the model with the best performance by evaluating different adjusting parameters;
4. classify or predict unknown data using the trained model.

The main idea of combining compartmental model and machine learning algorithms is to replace the fixed given parameter with a time-varying parameter[15, 16]. The infected and removed rates are updated daily according to the characteristics of the previous data instead of using a fixed given parameter to simulate the whole process. SIMLR applies the gradient descent method the optimize the iterated rate, which will react correctly to the reality[6]. The below figure shows the comparison of the classical SIR model and SIMLR model.

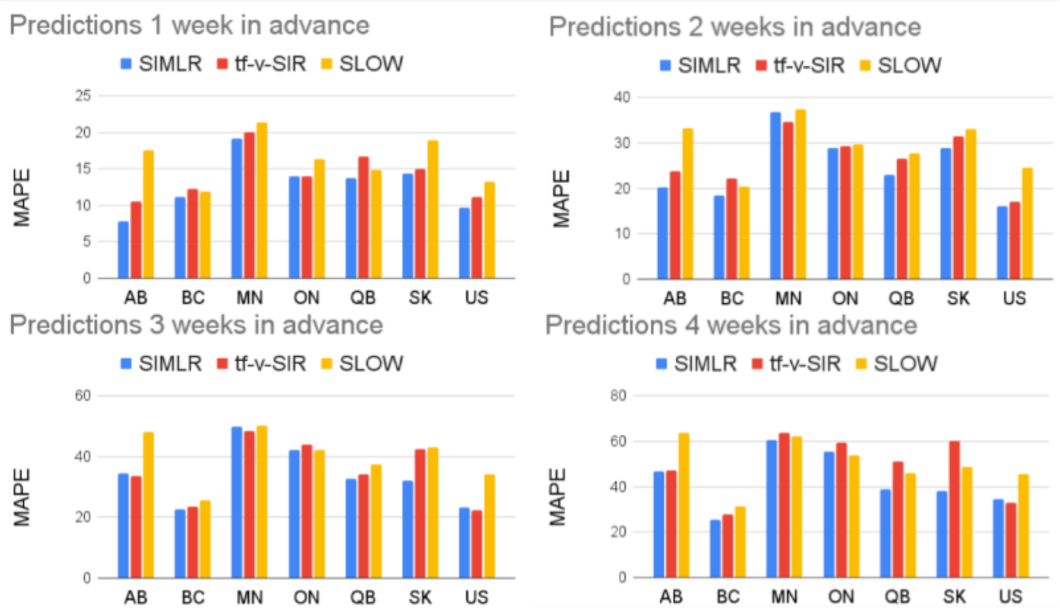


Figure 2.2: Comparison of SIMLR, SIR model with time-varying parameters, and SLOW.

There are many other approaches to predicting the dynamic of COVID-19. Instead of using the gradient descent method to optimize the parameter, a better algorithm is implemented to finish the forecast.

Chapter 3

Methodology

3.1 Susceptible-Infected-Recovered Model

The classic SIR model is an epidemiological model that computes the spread of a disease in a closed population over time. The model's name involves three classes of people: those who are susceptible to the disease, those who have been infected, and those who have already recovered with immunity.

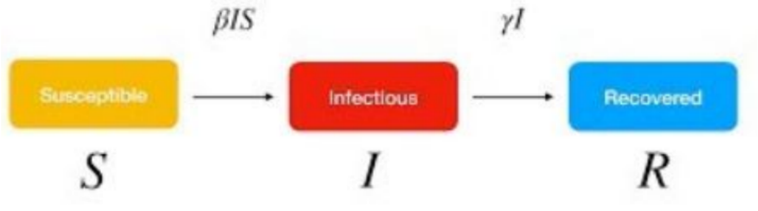


Figure 3.1: SIR model diagram

- S represents the number of people susceptible to being infected at time t.
- I represents the number of infected people in time t.
- R represents the number of removed persons at time t.
- β is the infection rate. On average, the probability that a randomly infected person infects a random susceptible person in time Δt is $\beta\Delta t$.
- γ is the removed rate, On average, the probability that a randomly infected person infects a random susceptible person in time Δt is $\gamma\Delta t$.

As analysed above, the differential equations describing this model can be written as below.

$$\begin{cases} \frac{dS}{dt} = -\beta IS \\ \frac{dI}{dt} = \beta IS - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \tag{3.1}$$

The infection and remove of one person are stochastic processes. Applying the continuous-time Markov Chain, the needed time of the newly infected person or removed person conforms to the exponential distribution, with the infected rate βIS and removed rate γI, respectively.

For more details of SIR model, please check[17].

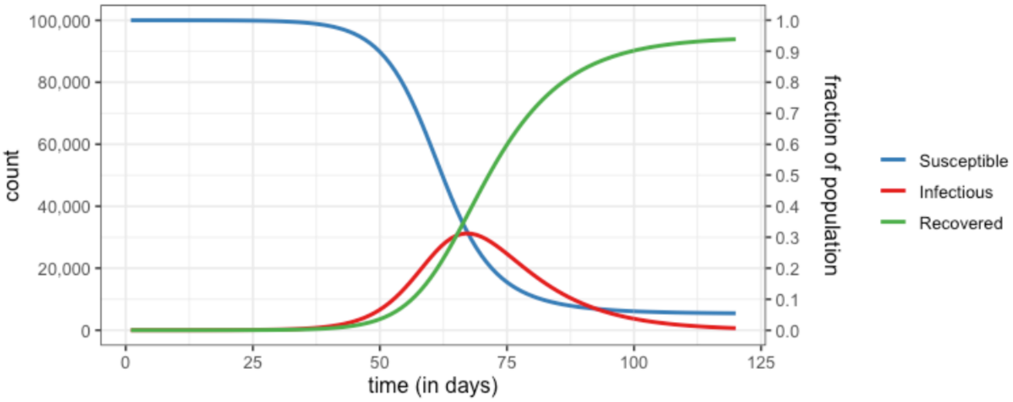


Figure 3.2: Sample of SIR model, with β=0.3 and =0.1.

3.2 Grey Forecasting Model

With more details, there are four steps to implement a machine learning algorithm:

1) Choose and prepare a training data set.
Training data represents the data that the machine learning application will ingest to tune model parameters. Training data is sometimes labeled, tagged to call out classifications or expected values the machine learning model is asked to predict. Labeled data should be divided into a training subset and a testing subset. The former is used to train the model, and the latter to evaluate the effectiveness of the model and try to optimize it. Other training data may be unlabelled, so the model must extract features and assign clusters autonomously.

2) Select an algorithm and apply it to the training data set.
The type of chosen machine learning algorithm will primarily depend on a few aspects. One is the size of the training data set. Another is the nature of the problem

the model seeks to solve. For prediction or classification cases, regression algorithms like logistic regression are often used.

3) Train the algorithm to build the model.
   Training the algorithm is the process of tuning model variables and parameters to more accurately predict the appropriate results. This process is usually iterative and uses a variety of optimization methods depending upon the chosen model.

4) Use and improve the model.
   The last step is to feed new data to the model as a means of improving its effectiveness and accuracy over time. Where the new information will come from depends on the nature of the problem to be solved. For instance, a machine learning model for depend on time, the prediction will be on road conditions, objects on the road, and traffic laws.

The Grey forecasting model, one of the most important machine learning algorithms for prediction, significantly affects the modeling and analysis of systems with short times, little data, and incomplete information. Grey prediction is to identify the degree of difference between the development trends of system factors, that is, to conduct association analysis, and to generate and process the original data to find the rules of system changes, generate data sequences with strong regularity, and then establish the corresponding differential equation model to predict. Regression algorithms like logistic regression are often used for prediction or classification cases.

In this paper, the simplest model, GM (1,1), is implemented to simulate the spread of COVID-19. Here, GM (1,1) means grey model in first order differential equation with one parameter. The basic form of GM (1,1) is shown below.

$$\frac{\mathrm{d}x^{(1)}(t)}{\mathrm{d}t} + ax^{(1)}(t) = b. \tag{3.2}$$

where a and b are parameters needed to be determined by the least square method.

The solution of the above first-order linear differential equation is

$$\hat{x}^{(1)}(t) = \frac{b}{a} + \left[x^{(0)}(1) - \frac{b}{a}\right] e^{-ak}. \tag{3.3}$$

We use the least square method and linear algebra to formulate this part of the problem and get the iterated formula.

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) = \frac{b}{a} + (1-e^{a})\left[x^{(0)}(1) - \frac{b}{a}\right] e^{-ak}, k = 1, 2, 3, \cdots, n \quad (3.4)$$

and

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}}\right) e^{-\hat{a}k} + \frac{\hat{b}}{\hat{a}}, \quad (3.5)$$

where represents the predicted value and x represents the true value.

For details of this algorithm, please check[18].

3.3 Motivation for the New Approach

The traditional SIR model has two main drawbacks. One is that the parameters are independent of time. Another is that this model is too simple to gain a more precise prediction.

It is easy to notice that the infected and removed rate of the SIR model is fixed without any time relationship. However, as time passes by, the rate of infection and removal is highly likely to be changed since the government will take actions to prevent the spread of the virus; for example, the infection rate at the very first stage of the pandemic must be higher than that after government issues the isolation policy. The removed rate must have a high jump after the vaccine has been developed successfully. If the parameters of the two rates depend on time, the prediction will possibly be more accurate.

On the other hand, the basic logic of the SIR model is not rigorous, just simplifying the dynamic of the spread of an epidemic system. The removed people have two sub-classes: those who have recovered from the disease and those who have died. Significantly, the recovered also can be infected again, while the dead will not. To simulate this process, more sub-classes are needed since the rate of people who have ever been infected should be lower than those who have never been infected empirically. Moreover, individuals divided into the removed group can also attain immunity by vaccine instead of recovering from the disease.

For the machine learning part, as mentioned before, the basic mathematical theory for machine learning algorithms is hard to optimize. Besides, the machine learning algorithm's performance depends on the training and testing data and the strategy it applies.

The optimized algorithm only has higher accuracy or faster processing speed in the specific problem and thus cannot fit more general situations. To improve the accuracy of prediction from the database, a better method with better basic logic is needed. In short, the machine learning method determines the performance of the data. The grey forecasting model is famous for its preciseness of prediction provided a small-size data set. So, choose this algorithm to represent the machine learning model.

The data used in this project is from https://github.com/Lewuathe/COVID19-SIR. The website offers data on daily confirmed, deaths, and recovered from more than one hundred countries. This project focuses on China, the UK, and the US to represent the general cases worldwide. Significantly, none of the two models performs well in every aspect of this prediction. However, combining the result of the two models gives a good forecast.

Chapter 4

Analysis and Result

4.1 Initialization

Empirically, the rate of infection, death, and removal should be parameters concerning time. However, the parameters of the classical SIR model are fixed. Since the aim of this project is to develop a model that predicts the daily confirmed, death, and recovered numbers in the next year based only on the population data up to that day, the fixed parameters will cause bad performance. It is reasonable to adjust the parameter as time passes by. This SIR model uses the least square linear equation to predict in every iteration instead of fixed rates. According to several experiments of testing the best size of the given data, five-day data perform well in the balance of accuracy and running time. Thus, in the models, both algorithms use data from five days to predict the number of confirmed, death and recovered in the sixth day. Especially, I choose the statistical data from China, US, and United Kingdom to finish the whole procedure. In the SIR model, since the number of susceptible is required, considering each country in reality, set the default population.

4.2 Result of SIR Model

The prediction of the modified SIR model does a good job in general. It reveals the dynamic of the pandemic reliably. Below is the result of different categories from different countries. According to the policies of different countries, each tendency can be explained very well and fit reality.

Significantly, the last figure showed the comparison between predicted and real data in China.

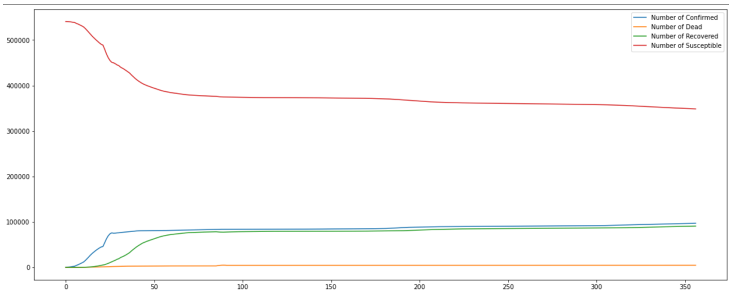


Figure 4.1: China: SIR Model predictions for the period from 20 January 2020 to 13 January 2021.

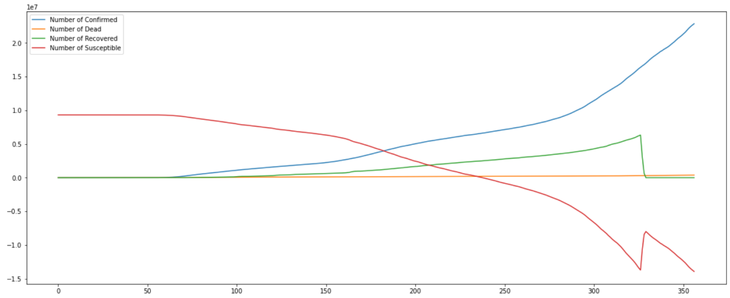


Figure 4.2: US: SIR Model predictions for the period from 20 January 2020 to 13 January 2021.

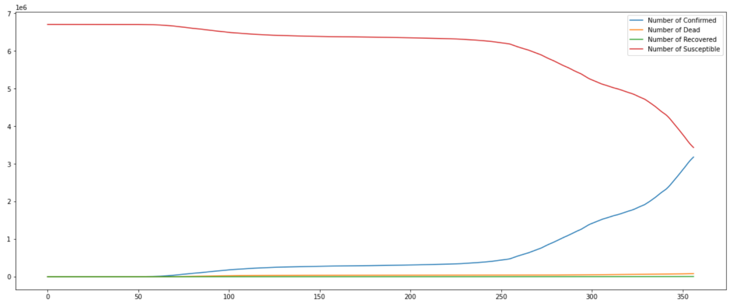


Figure 4.3: UK: SIR Model predictions for the period from 20 January 2020 to 13 January 2021.

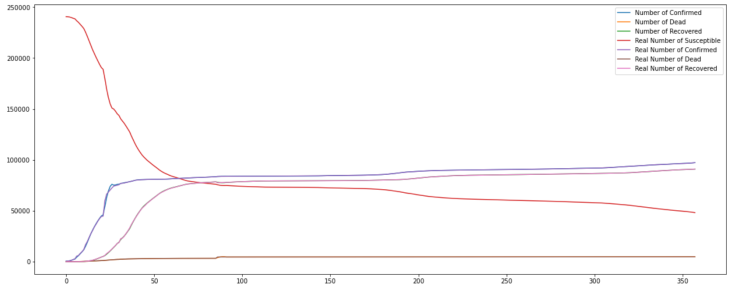


Figure 4.4: China: SIR Model predictions for the period from 20 January 2020 to 13 January 2021, with real-world data.

4.3 Result of Grey Forecasting Model

Due to the limitation of this model, it needs better data sets than the least square linear equation to finish the prediction, which will be analyzed in the next section. Unfortunately, the prediction of the US failed. The result for UK and China is shown below.

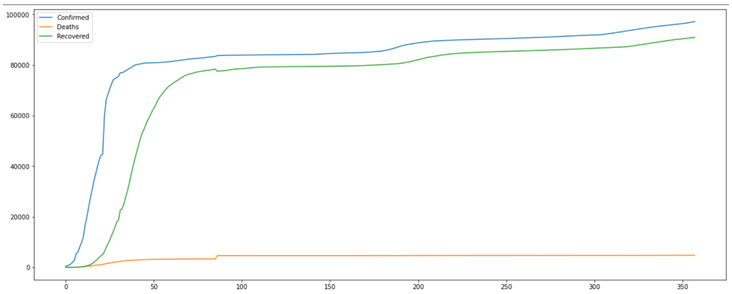


Figure 4.5: China: Grey Forecasting Model predictions for the period from 20 January 2020 to 13 January 2021.

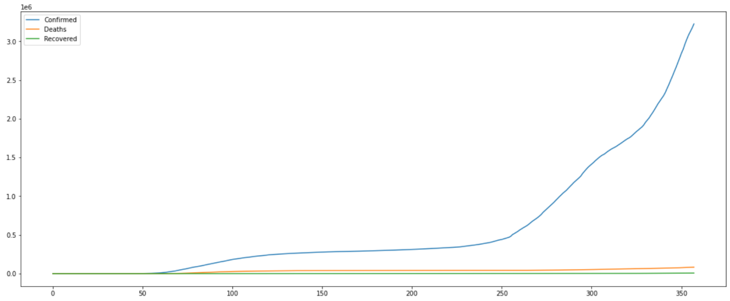


Figure 4.6: UK: Grey Forecasting Model predictions for the period from 20 January 2020 to 13 January 2021.

4.4 Accuracy, Limitation, and Running Time

To evaluate the accuracy of the prediction, define the relative residual $\varepsilon(k)$ which is calculated by

$$\varepsilon(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, \quad k = 2, \ldots, n. \tag{4.1}$$

If $\varepsilon(k)$ is smaller than 0.2, then we say the prediction meets the general requirements; if $\varepsilon(k)$ is smaller than 0.1, we say the prediction meets high requirements. In general, the performance of the grey forecasting model is much better than the SIR model. The detailed $\varepsilon(k)$ is shown in the below table.

	SIR Model	Grey Forecasting model
Confirmed	0.94986	0.17652
Deaths	0.00956	17.01038
Recovered	0.94473	2.04298e-5

Table 4.1: The Relative Loss of Different Categories of China from Two Models

From the above table and criteria, it is easy to conclude that the results of the deaths from the SIR model, the confirmed and recovered from the grey forecasting model are reliable. These three prediction offers excellent performance in forecasting the dynamic of COVID-19 in China.

Note that the main logic of the modified SIR model is to use the least square linear equation to predict the sixth day; it is a fitting model instead of making a prediction. Thus, it is natural for this SIR model to perform worse than the grey forecasting model, which is specially used for prediction from small-size data sets. However, the grey forecasting model becomes unreliable when predicting the everyday death number, despite the impressive performance of predicting the confirmed and recovered. This involves the drawback of the grey forecasting model. In the main logic of its prediction, it needs to get the coefficient matrix inverse. Though the coefficient matrix is positive definite, that is, the inverse of the matrix must exist, due to the limitation of the computer, if the data set is terrible enough, the computer will recognize a minimal number as zero, which will cause the inaccuracy after several matrices' multiplication, or return an error when doing the inverse of the matrix. Specifically, the data set of the US belongs to the case where a minimal number appears, and the computer recognizes a column as a zero column, which fails to get the matrix inverse.

SIR model 1. The parameters of classical SIR are fixed without any time relationship, which will cause inaccurate simulation in the long term. 2. The dynamic of the model builds a bit simple. It ignores many other cases, which simplifies the reality indeed but causes inaccuracy as well. 3. SIR model always needs another algorithm to apply the time-related parameters. The performance of the modified SIR depends on the chosen algorithm.
Grey forecasting model 1. It needs a pre-test for given data to ensure that the data is valid to transform and predict. 2. In the first-order linear differential equation, it uses the least square solution to calculate parameters, which involves matrix inverse. If minimal entries appear in the same row or column, the computer is likely to fail in matrix inverse.

Table 4.2: Drawbacks and Limitation of SIR and grey forecasting model

By running the models repeatedly, the average running time of the grey forecasting is much faster than that of the SIR model. Specifically, the running time depends on the data set. The average running time for China is about 22 seconds while the running time for the US is about 33 seconds according to the result of the SIR model.

	SIR Model	Grey Forecasting Model
China	21.8s	7.6s
the UK	23.6	8.3s
the US	33.5	-

Table 4.3: The Average Running Time from Two Models and Three Data Sets

Chapter 5

Conclusions and Future Directions

In this work, the writer applies two different models to predict the dynamic of COVID-19 for three different countries in the world. The modified SIR model must have a prediction while the grey forecasting model needs further requirements for the data set. Once the grey forecasting model completes the prediction, its accuracy and efficiency are both better than the SIR model.

Since the classical SIR model only offers the basic structure of the dynamic of an epidemic, the choice of the parameter-generating algorithm is pretty flexible. In reality, considering the privacy of data and the transform limitation of speed, federated learning is the best choice to be applied to the SIR model[19]. It even preserves the convergence on non-IID data[20]. Also, people made the effort to use a new approach to deal with the data in the grey forecasting model[21].

Bibliography

[1] "Wto response to the pandemic, trade and food security take centre stage at mc12." [Online]. Available: https://www.wto.org/english/news_e/news22_e/mc12_13jun22_e.htm

[2] R. Padhan and K. Prabheesh, "The economics of covid-19 pandemic: A survey," *Economic Analysis and Policy*, vol. 70, pp. 220–237, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0313592621000321

[3] D. Blumenthal, E. J. Fowler, M. Abrams, and S. R. Collins, "Covid-19 — implications for the health care system," *New England Journal of Medicine*, vol. 383, no. 15, pp. 1483–1488, 2020, pMID: 32706956. [Online]. Available: https://doi.org/10.1056/NEJMsb2021088

[4] T.-A. T. Le, K. Vodden, J. Wu, and G. Atiwesh, "Policy responses to the covid-19 pandemic in vietnam," *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/2/559

[5] T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar *et al.*, "A global panel database of pandemic policies (oxford covid-19 government response tracker)," *Nature human behaviour*, vol. 5, no. 4, pp. 529–538, 2021.

[6] R. Vega, L. Flores, and R. Greiner, "Simlr: Machine learning inside the sir model for covid-19 forecasting," *Forecasting*, vol. 4, no. 1, pp. 72–94, 2022. [Online]. Available: https://www.mdpi.com/2571-9394/4/1/5

[7] S. Arik, C.-L. Li, J. Yoon, R. Sinha, A. Epshteyn, L. Le, V. Menon, S. Singh, L. Zhang, M. Nikoltchev *et al.*, "Interpretable sequence learning for covid-19 forecasting," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 807–18 818, 2020.

[8] W. Kermack and A. McKendrick, "Contributions to the mathematical theory of

epidemics–i. 1927," *Bulletin of mathematical biology*, vol. 53, no. 1-2, p. 33—55, 1991. [Online]. Available: https://doi.org/10.1007/bf02464423

[9] F. Brauer, "The kermack–mckendrick epidemic model revisited," *Mathematical Biosciences*, vol. 198, no. 2, pp. 119–131, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0025556405001331

[10] C. Mair, G. Kadoda, M. Lefley, K. Phalp, C. Schofield, M. Shepperd, and S. Webster, "An investigation of machine learning based prediction systems," *Journal of Systems and Software*, vol. 53, no. 1, pp. 23–29, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121200000054

[11] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037014000464

[12] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107 562–107 582, 2020.

[13] I. Cooper, A. Mondal, and C. G. Antonopoulos, "A sir model assumption for the spread of covid-19 in different communities," *Chaos, Solitons Fractals*, vol. 139, p. 110057, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077920304549

[14] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *Ieee access*, vol. 6, pp. 35 365–35 381, 2018.

[15] Z. Liao, P. Lan, Z. Liao, Y. Zhang, and S. Liu, "Tw-sir: Time-window based sir for covid-19 forecasts," *Scientific Reports*, vol. 10, no. 1, 2020.

[16] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, "Data-based analysis, modelling and forecasting of the covid-19 outbreak," *PLOS ONE*, vol. 15, no. 3, 2020.

[17] H. H. Weiss, "The sir model and the foundations of public health," *Materials matematics*, pp. 0001–17, 2013.

[18] N.-m. Xie and S.-f. Liu, "Discrete grey forecasting model and its optimization," *Applied mathematical modelling*, vol. 33, no. 2, pp. 1173–1186, 2009.

[19] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[20] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of fedavg on non-iid data,” *arXiv preprint arXiv:1907.02189*, 2019.

[21] B. Cayir Ervural and B. Ervural, “Improvement of grey prediction models and their usage for energy demand forecasting,” *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 4, pp. 2679–2688, 2018.

This SIR model uses