Quantitative Analysis and Prediction of the Worldwide COVID-19 Pandemic

Xiaohui Pan*
School of Artificial Intelligence and Law, Shanghai University of Political Science and Law
Email: panxiaohui@shupl.edu.cn

Abstract. At present, the novel coronavirus disease 2019 (COVID-19) has spread rapidly worldwide. As COVID-19 impacts societies, governments, communities, and individuals, we want to provide information for governments and people to understand the situation better. In this paper, we perform quantitative analysis and prediction of the global COVID-19 pandemic. First, we summarize worldwide confirmed, recovered, death, and active cases. Second, we select eight representative countries and analyze their confirmed and death cases with considering their populations. Third, we exploit the AR model in machine learning to predict the cases of these representative countries for the next 30 days. Our analysis and prediction show that some countries, including the United States, Spain, and Brazil, are in and will still be in the severe COVID-19 outbreak. This work can provide reference information for the prevention and control of COVID-19.

1. Introduction
For many years, infectious diseases have been a significant issue for the medical and health systems of all countries in the world. There are currently 64 major infectious diseases in the world, involving 82 countries and regions. The novel coronavirus disease 2019 (COVID-19) refers to pneumonia caused by the infection of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). At present, the main modes of COVID-19 transmission are respiratory transmission and contact transmission. There is evidence that there is a fecal-oral transmission, but further confirmation is needed. COVID-19 is highly contagious, its clinical manifestations are similar to those of a cold, and severe cases can cause death.

The COVID-19 pandemic is causing great harm to our life. According to the report by the world health organization [1], there have been confirmed cases of COVID-19 in 218 countries, areas, or territories, including 42,540,739 confirmed cases and 1,160,650 confirmed deaths as of 27 October 2020. There are 17 countries that have more than 10,000 deaths by COVID-19. By adopting active medical treatment and preventive and control measures, the epidemic situation in China has been effectively controlled in April, and normal economic and social order is being restored in an orderly manner. However, the epidemic situation is still spreading rapidly in countries and regions other than China, and the most affected countries by the epidemic are in North America, Asia, and Europe. By 27 October 2020, the United States, India, and Brazil have the highest number of confirmed cases of 8777432, 77990322, and 5439641.

Since the outbreak of the COVID-19 epidemic, many studies have analyzed the development of the epidemic based on infectious disease dynamics models or machine learning models and related data, interpreted the epidemic trends in different countries, and made relevant recommendations for
epidemic prevention and control. Zhan et al. [2] used the COVID-19 data of Chinese cities to establish a virus spread spectrum and evaluated the development of the epidemic situation in countries such as Italy and South Korea. Zhuang et al. [3] used the maximum likelihood method to estimate the basic COVID-19 in Italy and South Korea and predicted the infection rate in the early stage. Li et al. [4] estimated the basic infection numbers before and after South Korea's prevention and control measures were taken and analyzed the Italian epidemic. Zheng et al. [5] used the state transition matrix model to analyze the development of the epidemic in South Korea and Italy. There are also studies to estimate the basic infection coefficient of COVID-19 [6-11]. However, most of these studies were carried out before March 2020, earlier than the time when the epidemic broke out on a global scale. Therefore, it is mandatory to conduct quantitative analysis and prediction of the worldwide COVID-19 pandemic with the latest data so as to provide information for the prevention and control of COVID-19.

The remainder of the paper is organized as follows. In Section 2, we introduce the methodology, including the dataset and experiment methods. Section 3 analyzes the global COVID-19 pandemic on the total numbers and the daily increased numbers of confirmed, death, and recovered cases. Section 4 provides an in-depth analysis of the COVID-19 pandemic in eight representative countries. In Section 5, we train an AR (Auto Regression) model with the latest data and use it to predict the future number of cases for representative countries. We conclude this paper in Section 6.

2. Methodology

2.1. Dataset
We use the COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [13]. This data repository contains the number of confirmed, recovered, and death cases of all countries and regions in the world. It is updated every day, from 22 January 2020 till now. We use the data from 22 January 2020 to 27 October 2020 in our experiments.

We perform data preprocessing to the COVID-19 dataset [12]. First, we merge the confirmed, recovered, and death cases of the provinces/states of each country to get the total number of confirmed, recovered, and death cases of each country. Then, we make data cleaning to convert the date from string to DateTime and replace the missing value. Finally, we conduct data aggregation to calculate active cases as active = confirmed – deaths – recovered.

2.2. Experiment Method
We perform both analysis and prediction of the global COVID-19 pandemic:

1) COVID-19 analysis: we perform two groups of analysis: (a) global COVID-19 cases; (b) COVID-19 cases of selected countries. The analysis of the global COVID-19 cases gives us a full view of the worldwide pandemic. In contrast, the analysis of selected countries shows us a detailed, in-depth view of the pandemic in eight representative countries.

2) COVID-19 prediction: we use the historical data to train an AR (Auto Regression) model, then use the trained model to predict COVID-19 cases for the future 30 days.
3. Overview of the Global COVID-19 pandemic

Figure 1. World total confirmed, recovered, death, and active cases.

Fig. 1 shows the world's total confirmed, recovered, death, and active cases from February to October, this year. As of 27 October, there are 43,950,352 confirmed, 26,681,686 recovered, 1,166,732 death, and 16,101,934 active cases. From February to October, the number of confirmed cases increases rapidly, especially after July. Moreover, the number of active cases also increases rapidly, especially in October. This demonstrates that the world is facing a more and more serious situation on COVID-19.

Figure 2. World daily increase in confirmed, death, and recovered cases.

Fig. 2 shows the world's daily increase in confirmed, death, and recovered cases. We can see steady growth in new confirmed cases. Before July, the daily increase in confirmed cases is below 200k. However, the daily increase in new confirmed cases is doubled in October, which means the COVID pandemic is getting worse.

We also analyze counties with the highest number of confirmed cases and deaths cases. The top 10 countries of confirmed cases are United States, India, Brazil, Russia, France, Spain, Argentina, Colombia, United Kingdom, and Mexico with 8,777,432, 7,990,322, 5,439,641, 15,371,42, 12,44,242, 11,167,38, 11,166,09, 10,332,18, 9,206,64, and 9,012,68 confirmed cases. The top 10 countries of death cases are United States, Brazil, India, Mexico, United Kingdom, Italy, France, Spain, Peru, and Iran with 2,266,91, 1,579,46, 1,200,10, 8,981,4, 4,545,5, 37,700, 35,582, 35,298, 34,197, and 33,299 death cases.

4. Analysis of the COVID-19 Pandemic in Selected Countries

We select eight representative countries from different continents and different severity of the epidemic for detailed analysis. The location and number of cases of these countries are shown in Table 1. We perform an in-depth analysis of three aspects: the absolute value of confirmed or death cases, population, the number of days to become the current situation of the epidemic.
Table 1. Location and cases of the eight representative countries.

| Countries     | Location       | Confirmed   | Deaths   | Recovered | Active    |
|---------------|----------------|-------------|----------|-----------|-----------|
| United States | North America  | 8777432     | 226691   | 3487666   | 5063075   |
| Brazil        | South America  | 5439641     | 157946   | 4914733   | 726064    |
| Spain         | Europe         | 1116738     | 35298    | 0         | 1081440   |
| United Kingdom| Europe         | 920664      | 45455    | 2018      | 873191    |
| France        | Europe         | 1244242     | 35582    | 117373    | 1091287   |
| Italy         | Europe         | 564778      | 37700    | 271988    | 255090    |
| China         | Asia           | 91222       | 4739     | 82226     | 59032     |
| Japan         | Asia           | 98146       | 1726     | 89465     | 17240     |

Fig. 3(a) shows the COVID-19 confirmed cases per million people of the representative countries. China and Japan have a relatively low value of 104 and 106.5, respectively. Italy has a middle value of around 109. The United States, Brazil, France, the United Kingdom, and Spain have a high value of around 1010, which means a high infection rate in these countries.

Fig. 3(b) shows the COVID-19 death cases per million people of the representative countries. China and Japan have a relatively low value of 101.2 and 102.6, respectively. The United States, Brazil, France, the United Kingdom, Italy, and Spain have a high value of around 106.5, which means that there is a high death rate in these countries.

5. Prediction of the COVID-19 Pandemic in Selected Countries

5.1. The AR (Auto Regression) Model

The AR (Auto Regression) model is a machine learning model for analyzing a series of data and making future predictions. The AR model is often applied to stationary time series that have no systematic change in the mean and variance and have no periodic change. The AR model uses the correlation between the historical value and the future value (autocorrelation) to establish a regression equation to achieve the purpose of prediction.

In general, the formula definition of the p-order autoregressive process is as follows:

\[ y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t, \]  

where \( y_t \) is the current value, \( \mu \) is a constant, \( p \) is the order of the autoregressive process, \( \gamma_i \) is the coefficient of autocorrelation, \( \epsilon_t \) is the error (i.e., white noise). The expanded form of (1) is as follows:

\[ x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + \mu_t. \]
If the random disturbance term is a white noise (i.e., $\mu = 0$ and $\mu_t = \epsilon_t$), then we have:

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + \epsilon_t.$$ (3)

It can be seen from equation (3) that the current value is predicted by the historical value, and $p$ is an order in the autoregressive model, indicating that how many historical values are used to predict the current value.

5.2. Training the AR Model for Prediction

We partition the COVID-19 dataset [13] into a training set and a test set. The training set contains data from 22 January 2020 to 18 October 2020, which is used to train the AR model. The test set contains data from 19 October 2020 to 27 October 2020, which is used to test the accuracy of the trained AR model. We train an AR model for each of the eight countries and then used the trained model to predict the confirmed cases from 19 October 2020 to 27 October 2020 for each country. Finally, we calculate the root mean square error (RMSE) between the test set and the predicted value, as shown in the following Table 2.

| Country         | United States | Brazil       | Spain        | United Kingdom | France       | Italy        | China       | Japan       |
|-----------------|---------------|--------------|--------------|----------------|--------------|--------------|-------------|-------------|
| RMSE            | 82107.59      | 40928.39     | 33076.86     | 28729.18       | 93891.20     | 59369.43     | 56.42       | 296.44      |

5.3. Predicting the Future Confirmed Cases with the Trained AR Model

Finally, we use the trained AR model to predict the future confirmed cases for the eight countries. The prediction is made for the future 30 days from 28 October 2020 to 26 November 2020, as shown in Fig. 4. In the figure, the blue line indicates the current (i.e., historical) data for confirmed cases, while the red line shows the prediction of the future 30 days. We also list the numbers from 22 November 2020 to 26 November 2020 in Table 3. From Fig. 4 and Table 3, we can see that the curves of all countries except China grow upward sharply, which means that the COVID-19 confirmed cases would keep growing rapidly in these countries. According to the prediction, the confirmed cases in the United States will exceed 10 million on 19 November 2020.

Figure 4. Prediction of confirmed cases for the future 30 days (from October 28 to November 26, 2020)
Table 3. Prediction Results of confirmed cases from November 22 to November 26, 2020.

| Prediction | United States | Brazil | Spain | United Kingdom | France | Italy | China | Japan |
|------------|---------------|--------|-------|----------------|--------|-------|-------|-------|
| 11-22      | 10187576      | 5896195 | 1363853 | 1343706       | 1578764 | 585200 | 90912  | 113036|
| 11-23      | 10248359      | 5916633 | 1376179 | 1362195       | 1599492 | 590149 | 90890  | 113632|
| 11-24      | 10309397      | 5937098 | 1388545 | 1380744       | 1620289 | 595082 | 90866  | 114231|
| 11-25      | 10370700      | 5957523 | 1400951 | 1399354       | 1641153 | 599997 | 90841  | 114832|
| 11-26      | 10432236      | 5977936 | 1413398 | 1418025       | 1662087 | 604894 | 90816  | 115434|

Based on the detailed analysis in Section 4 and the prediction in Section 5, we classify these countries into three severity levels of the COVID-19 pandemic, as follows.

1. High severity level: United States, Brazil, Spain, United Kingdom, France. These countries have a high number of confirmed cases and deaths per million people. In addition, they have a trend of rapid increase in future cases.

2. Medium severity level: Italy, Japan. These countries have a medium number of confirmed cases and deaths per million people.

3. Low severity level: China. This country has a low number of confirmed cases and deaths per million people. In addition, they have a trend of very slow increase in future cases.

6. Conclusions
In this paper, we performed quantitative analysis and prediction of the global COVID-19 pandemic based on the latest data from Johns Hopkins University [13]. We summarized worldwide confirmed, recovered, death, and active cases and analyzed the confirmed and death cases with considering the population for eight representative countries. Finally, we trained an AR model using historical data and used it to predict the cases of these representative countries for the next 30 days. Our analysis and prediction show that some countries, including the United States, Spain, and Brazil, are in and will still be in the Severe COVID-19 outbreak. Governments and people should cooperate and take all the necessary actions to prevent and control COVID-19 spreading.

7. Acknowledgments
This work was supported in part by the Special Research Project on Epidemic Prevention and Control of Shanghai University of Political Science and Law under Grant 2020YQFK12, and in part by the Philosophy and Social Science Planning Project of Shanghai under Grant 2019BGL028.

8. References
[1] World health organization 2020 novel coronavirus 2019 report https://www.who.int/emergencies/diseases/novel-coronavirus-2019
[2] Zhan C, Chi K T, Lai Z, Hao T and Su J 2020 Prediction of COVID-19 spreading profiles in South Korea, Italy and Iran by data-driven coding PLoS ONE 15(7) pp 1–17
[3] Zhuang Z, Zhao S, Lin Q, Cao P, Lou Y, Yang L, Yang S, He D and Xiao L 2020 Preliminary estimating the reproduction number of the coronavirus disease (COVID-19) outbreak in Republic of Korea and Italy by 5 March 2020 International Journal of Infectious Diseases 95 pp 308–10
[4] Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, Wang D, Chen G, Zhang J, Peng H and Shao Y 2020 Propagation analysis and prediction of the COVID-19 Infectious Disease Modelling 5 pp 282–92
[5] Zheng Z, Wu K, Yao Z, Zheng X, Zheng J and Chen J 2020 The prediction for development of COVID-19 in global major epidemic areas through empirical trends in China by utilizing state transition matrix model BMC Infectious Diseases 20(1) pp 1–12
[6] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung K, Lau E, Wong J, Xing X,
Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhao S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam T, Wu J, Gao G, Cowling B, Yang B, Leung G and Feng Z 2020 Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia *New England Journal of Medicine* **382** pp 1199–207

[7] Tang B, Wang X, Li Q, Bragazzi L, Tang S, Xiao Y and Wu J 2020 Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions *Journal of clinical medicine* **9**(2) 462

[8] Wu J T, Leung K and Leung G M 2020 Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study *The Lancet* **395**(10225) pp 689–697

[9] Zhou W, Xiao Y and Heffernan J M 2019 Optimal media reporting intensity on mitigating spread of an emerging infectious disease *Plos one* **14**(3)

[10] Backer J A, Klinkenberg D and Wallinga J 2020 Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020 *Eurosurveillance* **25**(5) 2000062

[11] Wang W, Tang J and Wei F 2020 Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China *J. of medical virology* **92**(4) pp 441–447

[12] Dong E, Du H and Gardner L 2020 An interactive web-based dashboard to track COVID-19 in real time *The Lancet infectious diseases* **20**(5) pp 533–534