Using Mathematical Modeling and Neuron Network to Predict the Dynamics of COVID-19 in China

Qinchen Ruan

The Department of Quantitative Theory & Methods, Emory University, 201 Dowman Dr, Atlanta, GA 30322, United States

qruan3@emory.edu

Abstract. The COVID-19 pandemic spreading from Wuhan in 2019 had a severe continuous impact globally. Even now when several vaccinations were approved by WHO and accepted widely, there are still millions of new confirmed cases daily. To provide insights for governments to make prompt and effective response with the smallest social and economic cost, numerous studies have proposed to predict COVID-19 development trend, in which mathematical modeling, such as SEIR, and neuron networks, such as LSTM, were utilized and modified widely. Among the reviewed papers, population migrations and quarantine policies were popularly considered as influential factors, while nature factors were rarely mentioned. The construction focuses of SEIR, and LSTM were parameter selection and dynamics, and solving overfitting from the data shortage and over-complicated structure respectively. The expansion of applicable environments and increase of prediction accuracy still seems to be necessary. Though this review is limited to the studies based on Chinese datasets, research from other countries may benefit from the analysis strategy.

Keywords: COVID-19, mathematical modeling, neuron network, SEIR, LSTM

1. Introduction

At the end of 2019, the first case of COVID-19, the official name from World Health Organization (WHO) [1] was discovered in Wuhan, Hubei province, China. Due to its high infectivity and transmission, the virus spread rapidly globally, and virus has caused 1,965,515 death cases cumulatively at the end of 2020 [2]. Even now when several vaccinations were approved by WHO and accepted widely, there are still millions of new confirmed cases daily. To prevent further transmission, governments chose complete or partial block down, quarantine and other social restrictions as health policies [3, 4]. Although the polities suppressed the pandemic effectively, the various mental health issues [5], economic challenges [6], and food allocation and affordability problems [7] caused by the polities show that these restrictions cannot be implemented in a long time on a large scale. Thus, to carry out effective restrictions in a short time on a small scale and minimize the potential issues mentioned, COVID-19 dynamic prediction has become an important research topic globally.

After WHO issued the global public health emergency, researchers have built models to forecast COVID-19 case number, the development trend, or the outbreak peak time and size to understand the virus propagation and suggest implications to authorities to prevent and control the transmission. Mathematical modeling has long been used for epidemiology and public health policy to predict and plan in both long and short term for restricting the spreading of a disease [8][9]. In the COVID-19pandemic, Susceptible-Exposed-Infected-Removed (SEIR) and related models were employed frequently to study the development trend and thus will be discussed in detail. On the other hand, neuron network has become a popular tool for prediction in the recent decades, which uses computer algorithms to analyze giant complex data structures, learn patterns in the data and create prediction models. The long-term short-term memory system (LSTM) model has been popular for epidemic dynamic prediction and thus will be studied closely in the paper. Since China was the country where the pandemic first erupted and took strict policies to solve the epidemic, the paper will focus on the studies analyzing Chinese pandemic.

2. Data sourcing

The statistics used by the papers are listed in Table 1. Most of the data were related to pandemic dynamics, such as the number of confirmed cases, and were used for parameter estimation for mathematical models, training for neural network models and testing for both categories. The statistics were mainly from China as the focus of this review paper, but statistics from other countries were also included for testing. Some authors used the data in country unit, while others cited the data in province, or even smaller city unit. Besides the pandemic dynamics related data, other statistics such as population migration related data were also incorporated in the dataset, which were considered as important factors effecting the pandemic development trend and integrated into the model.

Table 1. Summary of datasets used in COVID-19 studies (continue).

Author	Country / Province / City	Period	Variable
Choujun Zhan et al. [10]	China	Jan 24 - Feb 16, 2020	COVID-19 epidemic data: confirmed, recovered, death
		Jan 1 - Feb 13, 2020	Migration data
Joseph T Wu et al. [11]	China	Jan - Feb, 2019	Flight number to Wuhan globally
		Jan 6 - Mar 7, 2019	Domestical passenger volume from Wuhan
		Chunyun 2020	Domestic passenger volumes from and to Wuhan
		Dec 25, 2019 - Jan 19, 2020	Domestic confirmed case number from Wuhan in other cities
Zifeng Yang et al. [13]	China	Jan 11 - Feb 8, 2020	COVID-19 epidemic data
		2019	Migration data
		April and June, 2003	SARS epidemic data
Zengyun Hu et al. [15]	Guangzhou	Jan 27 - Feb 20, 2020	COVID-19 epidemic data: confirmed, cured, death
		2018	population size
		Jan 26 - Feb 20, 2020	Domestic population from and to Guangzhou
Bingjie Yan et al. [20]	Tianjin, Hong Kong, Hubei, Zhejiang, Henan, South, Korea, Italy	Feb - Apr, 2020	COVID-19 epidemic data: confirmed, death, recovered
			lockdown period
Fenglin Liu et al. [18]	China	Jan - Mar, 2020	Domestic population from and to Wuhan
Nanning Zheng et al. [19]	Wuhan, Beijing, Shanghai	Jan 23 - Feb 24, 2020	COVID-19 news and reports
			COVID-19 epidemic data: confirmed
Chiou-Jye Huang et al. [21]	Wuhan, Huanggang, Xiaogan, Ezhou, Yichang, Wenzhou, Shenzhen	Jan 23 - Mar 2, 2020	COVID-19 epidemic data: confirmed
Yan Hao et al. [22]	Wuhan	Jan 23 – Apr 6, 2020	COVID-19 epidemic data: confirmed, death, cured
	United States	Mar 23 – Jan 5, 2020	COVID-19 epidemic data: confirmed, death
Simon James Fong et al. [23]	Wuhan	Jan 21 – Feb 3, 2020	COVID-19 epidemic data: confirmed, cured, suspected, critical

3. Preprocessing

Preparing for the later matrix computation, Zhan et al. [10] processed the migration population size in and out of each city into the migration matrix in which the entry at position (i, j) indicated the number of citizens from city i to j. Wu et al. [11] excluded the data of Hong Kong when estimating the transmissibility of Wuhan using the data from 2019 considering social unrest in Hong Kong typical in 2020. Yan et al. [20] processed the online data to get the statistics about diagnoses as input of the model and standardized the parameters by MinMaxScalar. For the input of NLP module in their model, Zheng et al. [19] sorted the COVID-19 related news by date, city, filtered out case reports and foreign news, and extracted the titles and main content of news. The data were processed by a pretrained model of the BERT language model before input into NLP module. The titles and main content were input into the module separately to avoid overfitting and improve training efficiency.

4. Mathematical modeling

Mathematical modeling has long been employed to depict the dynamics of infectious diseases and understand the epidemic growth patterns. The ordinary mathematical models were refined to capture the characteristics of the specific pathogen and the social context.

4.1. Susceptible-exposed-infected-removed (SEIR) model

As an epidemiological model, SEIR has been widely used to study the epidemic spreading. In the model, individuals of a population will be categorized into one of the four stages of epidemic spreading, and proportion of individuals changed from one stage to another as time passed will be represented by some parameters.

Zhan et al. [10], considering the unusually high volume and frequency of population flow between cities in the onset of the pandemic in China, modified the classic SEIR model by integrating the daily intercity migration data to forecast the dynamics of COVID-19 in China. In the modified SEIR model, both the difference between category E and R, and the net flow of infections into the city were counted as parts of the daily increase of infected cases. The parameters were estimated by constrained nonlinear programming, in which the data from Jan 24 to Feb 13, 2020, were used for fitting. In the result, the time of the predicted outbreak would be between February to March of 2020, and the size would reach about 0.8 in Wuhan, less than 0.1 in Hubei province and less than 0.01 for the rest of China in percentage of populations.

Another study by Wu et al. [11] also considered the population flow as a main factor to estimate epidemic spreading extent but focused on the export of infected individuals from Wuhan domestically and internationally. The transmissibility of Wuhan was estimated based on the transportation related data in the similar time from 2019, and the outbreak sized was estimated based on the confirmed case number exported from Wuhan reported outside of China. The results showed that in the baseline scenario where R0 was set as 2.68, Wuhan exported 461, 113, 98, 111 and 80 infected cases to Chongqing, Beijing, Shanghai, Guangzhou, and Shenzhen provinces, respectively.

Prem et al. [12], concerned of the effectivity of physical distance control policies, applied age structured SEIR model to forecast the effect of population mixing on the virus progression. Prem et al. [12] assumed that the social mixing pattern was different for individuals in different groups and position and thus with different probabilities to be exposed to coronavirus. For the effect of location on social mixing patterns, three scenarios were considered, namely usual social mixing, Lunar New Year holiday and relax intervention patterns. The simulation showed that policies limiting social mixing were effective to reduce the size and delay the peak of the pandemic, though the effect varied in different age groups. The simulation also suggested that the staggered return to work starting at April would maximize the effect of these measures.

Like Prem et al. [12]'s concern, Yang et al. [13] wandered whether the severe policies taken in China, such as the quarantine of whole cities, causing unignorable social and economic disruption, limited the epidemic effectively. Like Zhan et al. [10]'s study, Yang et al. [13] integrated population migration into the classic SEIR model by stating two additional parameters, the move-in and move-out. Considering the unique characteristics of coronavirus, E was associated with asymptomatic but infectious individuals while I was associated with symptomatic and infectious individuals. The incubation time was set to be 7 days, the midpoint of reported incubation period. The epidemic data from Hubei province were used to model the skewed SEIR model to determine other constants in the model. In the result, a good fit was shown between the predictions and the reported data. Also, the severe policies were important to limit the epidemic size.

4.2. Other models

Hu et al. [15] employed Susceptible-Exposed-Infectious-Removed-Quarantined (SEIRQ) model with seven compartments for the whole population. Four of the compartments were the same as those in SEIR model, and the rest three were quarantined susceptible, quarantined exposed and quarantined infectious. Guangzhou province was chosen as the analysis example to study the effects of population control strategies on the dynamic of COVID-19 considering its large gross domestic product compared to other provinces and expected giant inflow of workers in the future. The values of determinant coefficients, AE, DISO, RE showed the high accuracy of the model.

Time-dependent susceptible-infected-recovered (SIR) model was proposed by Chen et al. [16] with transition rate and recover rate as functions of time to forecast the infected and recovered number during some period. The two time-sensitive parameters were estimated by ridge regression. For evaluation, they obtained prediction curves highly aligned with the real curves and the prediction errors within 3% for the confirmed cases.

Wang et al. [17] tried to integrate the time-sensitive quarantine policies to the SIR model by two main methods. Like Chen et al.'s study, the focus of the first method was the time-sensitive transmission rate which was achieved by adding a transmission rate modifier varying with different quarantine protocols or time. The second approach was like Hu et al. [15]'s consideration, which extended the three compartments of SIR model into four by adding a quarantine compartment. The parameters were estimated by Markov Chain Monte Carlo algorithm. Shown in the results, the integrated quarantine factor improved both estimation and prediction.

5. Neuron network

NN [14] are simplified models imitating the human intelligence. Their structure are layers consisting of the basic units, neurons. The input data pass through an input layer, hidden layers, and output layer. Then, the networks generate predictions for all observations and adjust weights to improve the predictions. The process is repeated until some stop criteria are achieved.

5.1. Long-term short-term memory system (LSTM)

As an improved recurrent neural network method, though RNN frequently uses for prediction, the vanishing gradient or exploding gradient problems and the storage limited to short-term memory are still the main shortcomings of RNN. The presence of gate functions in the structure of LSTM enables the model to solve the problems of long-term dependencies. The gate functions represent the states of the four gates and build interactions between gates.

Yang et al. [13] constructed a LSTM trained by SARS data, incorporated the parameters of coronavirus, and optimized by the Adam optimizer with 500 iterations. The structure was kept simple to prevent overfitting from the small training dataset. The results of the confirmed case number from the model fit remarkably with the real statistics. The model predicted an outbreak in February with 4000 daily infections, which aligned with the prediction from SEIR model.

Liu et al. [18] compared the predictions with modified Susceptible-Exposed-Infected-Recovered-Dead (SEIRD) dynamic model, Geographically Weighted Regression (GWR) model and an LSTM. The LSTM with an input layer, an LSTM layer, a fully connected layer, and an output was trained by the data from four provinces in China and optimized by Adam optimizer with MSE as the loss function. To incorporate the effect of migrants from Wuhan to the provinces in the training dataset, the cumulative migrants from Wuhan and the incidence were also considered in the model. The results from LSTM fit with the real situation and got a MAPE smaller than the other two models.

Zheng et al. [19], realized the inability of susceptible-infected (SI) model to capture the change of policies and emergency conditions, used LSTM and natural language processing (NLP) module to reduce the deviation of prediction. LSTM updated the parameters corresponding to different pandemic policies, and NLP module counted people's awareness of epidemic prevention affected by news. The proposed model made an improved prediction of confirmed case numbers with the smallest mean absolute and mean absolute percentage error than the SI model, improved SI model and improved SI model with LSTM network.

Yan et al. [20], targeted to outdo the prediction of mathematical equations and population prediction models, constructed a modified LSTM to predict the positive cases. Concerned of the biased fitting for cities with a large case number, the traditional LSTM was improved to adjust its parameters according to different epidemic stages judged by the standard deviation of n days before. The prediction results from the proposed LSTM were compared with those of ordinary LSTM, logistic and hill equation algorithms by goodness of fit and deviation rate. The proposed LSTM had a better goodness of fit than the ordinary LSTM, and a smaller deviation rate within 2% then logistic and hill equation algorithms, showing the improvement made by Yan et al. [20].

5.2. Other models

Through Convolutional Neural Network (CNN) method, Huang et al. [21] forecasted the positive number in part area of China. CNN is a feedforward neural network with a relatively small weight number leading to easy training and effective characteristic extraction. The proposed deep CNN model included CNN and a dropout layer to prevent overfitting due to the small sample size. Huang et al. [21] compared the results of the deep CNN model with other neural networks, on the criteria of MAE and RMSE and found that CNN had the smallest errors and the best performance among the networks.

Fong et al. [23], aimed at solving the data deficiency at the start of a pandemic, proposed Group of Optimized and Multisource Selection (GROOSE) method in which five models were constructed to complete for the best prediction. Polynomial Neural Network with Corrective Feedback (PNN+cf) was used in group one. PNN was an evolutionary neural network that increased its powers of polynomial coefficients until the best fitting with the data was reached. By comparing the RMSE of each group, group one with PNN+cf had the smallest error and the best performance. For PNN+cf model, the input as a combination of suspected, confirmed, and critical cases gave the most correct prediction.

6. Evaluation

Hu et al. [15] also used a range of parameters, such as determinant coefficients, AE, DISO, RE, to show the high accuracy of SEIRQ model, and its forecasting performance was supported by the absolute values of RE. Chen et al. [16], Wang et al. [17] drew their predictions and observed data on the same graph to do the comparison and found out the predicted epidemic peak times and sizes were close to the real situations. Accuracy was also tested among models. Different performance of SEIRD model, LSTM and GWR by their MAPE with observed statistics are compared in Liu et al. [18]. Zheng et al. [19] showed the outstanding prediction accuracy of the hybrid model of improved SI model, LSTM and NLP module by the smallest MAE and PMAE of the model among other hybrid models. Huang et al. [21] concluded the high feasibility of the proposed deep CNN model by comparing its MAE and RMAE with other neural network models. Fong et al. [23] selected PNN+cf model from the other model groups for prediction due to its smallest RMSE. Yan et al. [20] employed goodness of fit and deviation rate to show the improvement of the modified LSTM.

7. Discussion

Though research had been done to make mathematical models fit with the situation of the coronavirus pandemic, the model could only depict some basic aspects of the reality. Because of this limitation, researchers made assumptions to simplify real situations before using the model. The modified SEIR model from Zhan et al. [10] was built on four assumptions, and the parameters were assumed to be unchanged in the short period for computational simplicity. Chen et al. [16] and Wang et al. [17] made lists of assumptions and limitations of their SIR models. The static parameters were another aspect of the mathematical models concerned by mathematicians, which could not self-adjust corresponding to the changing pandemic situations. To overcome this disadvantage, Wang et al. [17] added a transmission rate modifier or a quarantine compartment to the original SIR model. On the other hand, Zheng et al. [19] solved the problem by integrating LSTM and NLP module to the SEIRD dynamic model.

For LSTM, its prediction performance outdid some other models. Liu et al. [18] showed that MAPE of LSTM was smaller than those of SEIRD and GWR models. Yan et al. [20] compared the deviation rates of the proposed LSTM, logistic and hill equation algorithms and found that the proposed LSTM had the highest accuracy. However, LSTM may be still not the most suitable model for the prediction. Huang et al. [21] chose a CNN model due to its lower MAE and RMSE than LSTM. Hao et al. [22] selected an Elman neural network model since the model predicted the deaths and cumulative cured cases better than LSTM. Also, neural networks were unstable to predict the aperiodic epidemic data, described in Hao et al. [22]'s study, and likely to overfit due to data deficiency, mentioned in Huang et al. [21]'s study, and excessive structure complication, stated in Fong et al. [23]'s paper.

In this situation, the strategies to hold the assumptions in real application or release the limits seem to be necessary. Other algorithms and models should be explored or integrated with existing models to improve prediction accuracy. The number of days the model can predict ahead need to be extended to give more time for epidemic outbreak prepare [24].

8. Conclusion

In this paper, the studies that using mathematical modeling and neural network methods to predict COVID-19 dynamic were analyzed, combining with different factors including pandemic development trend, the proposed new sights, and prediction accuracy. Population migrations, quarantine policies and people's awareness were popularly considered as influential factors, while nature factors, such as temperature, were rarely mentioned. In the two method categories, SEIR and LSTM are widely used models, and each showed high prediction accuracy. The consideration for parameter selection and the adjustment of static parameters are two main focuses for SEIR model construction. LSTM was modified to solve overfitting from the data shortage and over-complicated structure. Model hybrid are a popular idea to overcome these problems. These studies may provide better understanding for the pandemic development and reference for governments to propose pandemic policies.

Reference

[1] Huang, Chaolin, et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." *The lancet* 395.10223 (2020): 497-506.

[2] WHO Coronavirus (COVID-19) Dashboard. 2020. https://covid19.who.int/?mapFilter=deaths

[3] World Economic Forum. 2020. https://www.weforum.org/agenda/2020/03/todays-coronavirus-updates/

[4] Brauner JM, Mindermann S, Sharma M, Johnston D, Salvatier J, "Inferring the effectiveness of government interventions against COVID-19." *Science* 371.6531 (2021): eabd9338.

[5] Sameer, A. S., et al. "Assessment of mental health and various coping strategies among general population living under imposed COVID-lockdown across world: a cross-sectional study." *Ethics, Medicine and Public Health* 15 (2020): 100571.

[6] Atalan, Abdulkadir. "Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective." *Annals of medicine and surgery* 56 (2020): 38-42.

[7] Batlle-Bayer, Laura, et al. "Environmental and nutritional impacts of dietary changes in Spain during the COVID-19 lockdown." *Science of The Total Environment* 748 (2020): 141410.

[8] Brauer, Fred, Carlos Castillo-Chavez, and Zhilan Feng, "Introduction: A Prelude to Mathematical Epidemiology," *Mathematical Models in Epidemiology*. Springer, New York, NY, 2019. 3-19.

[9] G. Chowell, L. Sattenspiel, S. Bansal and C. Viboud, “Mathematical models to characterize early epidemic growth: A review,” *Physics of Life Reviews*, vol. 18, pp. 66-97 (2016).

[10] C. Zhan, C.K. Tse, Y. Fu, Z. Lai and H. Zhang, “Modeling and prediction of the 2019 coronavirus disease spreading in China incorporating human migration data,” *Plos one*, 15(10), p.e0241171 (2020).

[11] J. T. Wu, K. Leung and G.M. Leung, “Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study,” *Lancet*, 395:689-97, S0140-6736(20)30260-9 (2020).

[12] K. Prem, Y. Liu, T.W. Russell, A.J. Kucharski, R.M. Eggo and N. Davies, “The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study,” *Lancet Public Health*, 5:e261-70, S2468-2667(20)30073-6 (2020).

[13] Z. Yang, Z. Zeng, K. Wang, S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong and J. He, “Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions,” *J Thorac Dis*, 12(3):165-174 (2020).

[14] Ahmed, Abdulmalek, et al. "New artificial neural networks model for predicting rate of penetration in deep shale formation." *Sustainability* 11.22 (2019): 6527.

[15] Z. Hu, Q. Cui, J. Han, X. Wang, W. E.I. Sha and Z. Teng, "Evaluation and prediction of the COVID-19 variations at different input population and quarantine strategies, a case study in Guangdong province, China,” International *Journal of Infectious Diseases*, vol. 95, pp. 231-240 (2020).

[16] Y. Chen, P. Lu and C.Chang, “A Time-Dependent SIR Model for COVID-19 With Undetectable Infected Persons,” *IEEE Transactions on Network Science and Engineering,* vol. 7, no. 4, pp. 3279-3294 (2020).

[17] L. Wang, Y. Zhou, J. He, B. Zhu, F. Wang, L. Tang, M. Kleinsasser, D. Barker, M.C. Eisenberg and P. X.K. Song, “An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China,” *Journal of Data Science*, 18(3), 409-432 (2020).

[18] F. Liu, J. Wang, J. Liu, Y. Li, D. Liu, J. Tong, Z. Li, D. Yu, Y. Fan, X. Bi, X. Zhang and S. Mo, “Predicting and analyzing the COVID-19 epidemic in China: Based on SEIRD, LSTM and GWR models,” *Plos one*, 15(8), p.e0238280 (2020).

[19] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma, Q. Yuan, S. Zhang, D. Zhang, F. Ye and J. Xin, “Predicting COVID-19 in China Using Hybrid AI Model,” *IEEE Transactions on Cybernetics,* vol. 50, no. 7, pp. 2891-2904 (2020).

[20] B. Yan, X. Tang, B. Liu, J. Wang, Y. Zhou, G. Zheng, Q. Zou, Y. Lu, W. Tu and N. Xiong, “An Improved Method for the Fitting and Prediction of the Number of COVID-19 Confirmed Cases Based on LSTM,” *arXiv* (2020).

[21] C. Huang, Y. Chen, Y. Ma and P. Kuo, “Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China,” *medRxiv* (2020).

[22] Y. Hao, T. Xu, H. Hu, P. Wang and Y. Bai, “Prediction and analysis of Corona Virus Disease 2019,” *Plos one*, 15(10), p.e0239960 (2020).

[23] S. J. Fong, G. Li, N. Dey, R. G. Crespo and E. Herrera-Viedma, “Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak,” *arXiv*, vol. 6, no. 1, pp. 132-140 (2020).

[24] P. Wang, X. Zheng, G. Ai, D. Liua and B. Zhua, “Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran,” *Chaos, Solitons & Fractals*, vol. 140 (2020).