Journal of Nonlinear Modeling and Analysis
Volume 3, Number 4, December 2021, 547–559
http://jnma.ca; http://jnma.ijournal.cn
DOI:10.12150/jnma.2021.547

Regional Prediction of COVID-19 in the United States Based on the Difference Equation Model*

Ceyu Lei[1] and Xiaoling Han[1,†]

Abstract The novel coronavirus pneumonia 2019 (COVID-19) has swept the globe in just a few months with negative social and psychological consequences for public health. So far, the United States has been one of the countries most affected by the epidemic. In this study, 51 states in the United States are divided into 10 state clusters according to relevant factors, and a difference equation model with spatio-temporal dynamic characteristics is established to predict the transmission dynamics of COVID-19 in the 10 state clusters and obtain data on regional aggregation levels (the United States). The study showed that the Pearson Correlation Coefficient between the actual data and the predicted data in the 10 state clusters is between 0.6 and 0.96 (mean R^2=0.8448), and the mean absolute error (MAE) of the newly confirmed cases in each cluster is between 300 and 1650 (mean MAE=878) and the average forecasting error rate (AFER) of the total confirmed cases in each cluster is between 0.9% and 3% (mean AFER=1.57%). These results show that the difference equation model can well predict the changes in the recent confirmed cases of infectious diseases such as COVID-19.

Keywords COVID-19, Prediction, Difference equation, Modeling, Mean absolute error.

MSC(2010) 35R02, 97M10.

1. Introduction

Since the first case appeared in Wuhan, China in December 2019, the COVID-19 has aroused people's attention. Since March 2020, the COVID-19 has rapidly spread around the world. As of December 2020, more than 200 countries and regions have been affected by the COVID-19. Among them, the United States, India, Brazil, Russia and France are the five countries affected by the pandemic most. In particular, the United States has become the center of the global pandemic with about 100,000 newly confirmed cases every day. The COVID-19 is disrupting the lives of people around the world in a variety of ways, and has a negative impact on global economic development. However, in the process of epidemic prevention and control, the prediction of COVID-19 is particularly important.

†the corresponding author.
Email address: 714327480@qq.com (C. Lei), hanxiaoling@nwnu.edu.cn (X. Han)
[1]Department of Mathematics, Northwest Normal University, Lanzhou, Gansu 730070, China
*The authors were supported by National Natural Science Foundation of China (No. 11561063) and Natural Science Foundation of Gansu Province (No. 20JR10RA086).

With the development of the dynamic theory of infectious diseases, the methods and theories for studying infectious diseases are becoming more abundant, such as SI model, SIR model, SEIR model, partial differential equation model, machine learning technology and other methods. In the process of the continuous spread of COVID-19, many scholars in the world have used different methods to obtain abundant and classic research results [1, 2, 4–13, 15–20, 23, 27]. Tang et al,. [21] established a random discrete epidemic model with case input, analyzed the effectiveness of China's Shaanxi Province epidemic prevention policy and conducted a predictive analysis of multiple outbreaks caused by economic recovery. Wang et al,. [24] predicted the epidemic in Arizona, the United States by establishing a partial differential equation model with temporal and spatial factors, and analyzed the impact of human preventive measures on the reduction of COVID-19 cases. O. Torrealba-Rodriguez et al,. [22] studied the outbreak in Mexico by using the Gompertz, Logistic and artificial neural network models, the results showed that these models had good predictive performance and the R^2 of each model was 0.9998, 0.9996 and 0.9999 respectively.

This work aims to explore the spread of COVID-19 between state clusters and to make further predictions of the changes in the epidemic. Specifically, we divided the United States as a whole into 10 different state clusters according to the relevant regulations of the United States, and abstractly divided the spread of COVID-19 into two processes: local spread and global spread. On this basis, we established a difference equation model with time and space factors is used to describe the spatiotemporal dynamic propagation process of COVID-19 and predict the development of the epidemic. The results show that the model not only has good predictive ability, but also provides a policy introduction to a certain extent, and provide a more scientific theoretical basis for controlling the development of the epidemic.

The rest of this article is structured as follows: In Section 2, the data sources are introduced in details, and the established model is described in details. Section 3 gives experiments and results analysis. Finally, the results of our research are discussed in Section 4.

2. Methods

2.1. Data set

We use the motif clustering algorithm in [3] to divide the United States into several clusters, and then use the difference equation model to predict the confirmed cases of each cluster of COVID-19. To make modeling and reporting more feasible and easier for the public to understand the situation in the United States according to the definition of the U.S. Department of Health and Human Services (HHS), 51 states in the United States are divided into 10 regions [25,26], as shown in Figure 1. The COVID-19 data repository adopted in this study is obtained from the World Health Organization (WHO) website (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports). In this study, the number of confirmed cases and the number of deaths in 51 states in the United States from 10 March, 2020 to 24 November, 2020 for a total of 260 days are used. We calculated the relevant data of 10 clusters based on the collected data of each state, as is shown in Figure 2.

The states contained in each cluster are as follows:

- Cluster 1 = [Massachusetts, Connecticut, New Hampshire, Maine, Rhode Island, Vermont];
- Cluster 2 = [New York, New Jersey];
- Cluster 3 = [Pennsylvania, Washington District of Columbia (Washington, D. C.), Maryland, Virginia, Delaware, West Virginia];
- Cluster 4 = [Florida, North Carolina, Tennessee, Georgia, Alabama, Mississippi, South Carolina, Kentucky];
- Cluster 5 = [Illinois, Michigan, Ohio, Indiana, Minnesota, Wisconsin];
- Cluster 6 = [Texas, Louisiana, New Mexico, Oklahoma, Arkansas];
- Cluster 7 = [Missouri, Iowa, Nebraska, Kansas];
- Cluster 8 = [Utah, Colorado, South Dakota, North Dakota, Wyoming, Montana];
- Cluster 9 = [California, Nevada, Arizona, Hawaii];
- Cluster 10 = [Alaska, Idaho, Oregon, Washington].

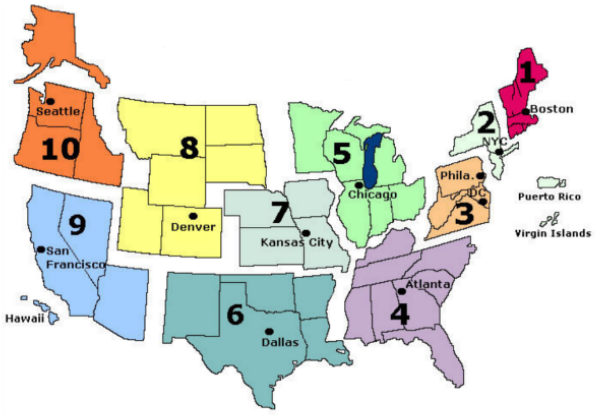


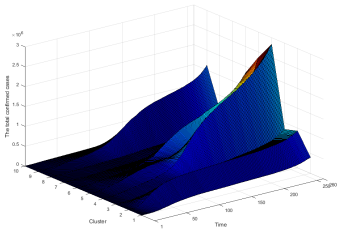
Figure 1. National Center for Chronic Disease Prevention and Health Promotion regions 1-10 represent 10 different CDC regions in the United States.

2.2. Embedding

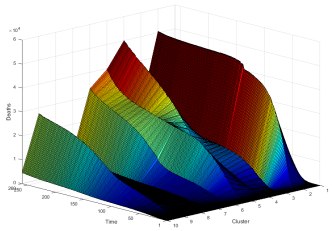
In order to use the difference equation to model the spread of COVID-19, the corresponding graph must be embedded in the Euclidean space [26], as is shown in Figure 3. We embed 10 clusters on the x-axis from east to west, where $x = i(i = 1, 2, \cdots, 10)$ represents the position of cluster i on the x-axis. When we embed the entire study area into the x-axis, the x-axis will be divided into three parts. We define the middle part as $x = 1, 2, \cdots, 10$, the right side as $x = 0$, and the left side as $x = 11$. That is, the right and left are the parts outside the study area.

2.3. Data preprocessing

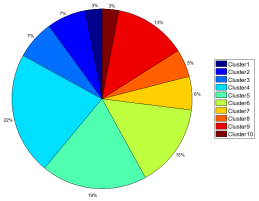
The data collected are preprocessed to improve the prediction accuracy of the model. When the level difference between the data in the collected original data set is large, if the original data value is directly used for analysis, the role of the data with higher value in the comprehensive analysis will be highlighted, and the role of the data with



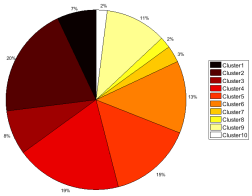
(a) Total confirmed cases per cluster



(b) Proportion of confirmed cases



(c) Number of deaths per cluster



(d) Percentage of deaths

Figure 2. Changes in total confirmed cases and deaths in each cluster, and their proportions

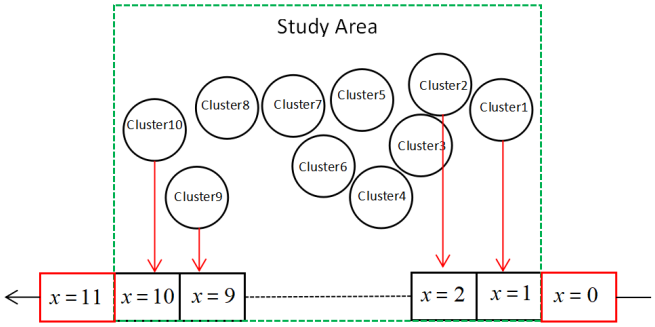


Figure 3. Embedding of clusters into the x-axis

lower value will be relatively weakened. Therefore, in order to ensure the reliability of the results, it is necessary to standardize the original index data.

Furthermore, to accelerate the convergence of the model and reduce the impact of outliers, the features in the data records are normalized as follows:

$$f_i^* = \frac{f_i}{10^j}$$

where f_i is raw data, f_i^* is standardized data and j is the smallest integer such that $max|f_i^*| < 1$. In this study, we assume $j = 7$.

2.4. Modeling analysis

As is shown in Figure 4, the source of infection in each state cluster has the greatest impact on the cluster (local process), and different state clusters also influence

each other through factors such as population flow or food flow (global process). Therefore, this paper proposes a difference equation model with time and space factors to describe the dynamic spread of infectious diseases. For a state cluster, the virus carriers in the cluster move in the cluster and spread the virus to people inside the cluster, which can be regarded as local process. This can be considered global process when a carrier brings the virus into another cluster through population movements, or when an item carrying the virus enters a cluster and sickens an individual in the cluster. Local process reflects the diffusion of the internal and underlying network structure of the cluster, and is directly related to the cluster. Global process is the spread of viruses between clusters caused by other factors such as population, air, and the movement of goods (usually manifested as more or less random walking). This method will extend our analysis of the results of difference equation modeling.

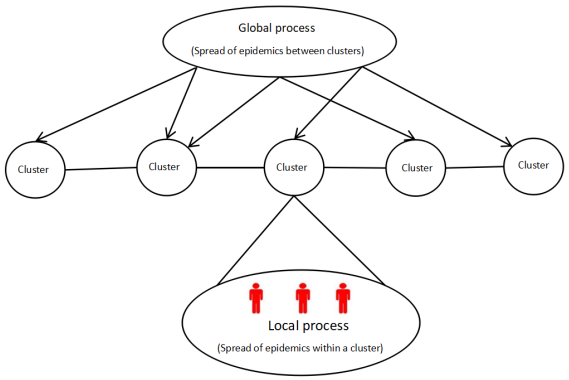


Figure 4. Embedding of ten regions into the x-axis and two spreading processes

The following is the description of the difference equation model:

$$\triangle u_t(x,t) = D(x)[\nabla(\triangle u_x(x,t))] + r(t)u(x,t)[1 - \frac{u(x,t)}{K}],$$

$$u(x,1) = \varphi(x), \quad 1 \le x \le 10,$$

$$\nabla u_x(1,t) = \triangle u_x(10,t) = 0, \quad t > 1,$$

where

• $u(x,t)$: representing the total number of confirmed cases in the area x at time t.

• $\triangle u_t(x,t)$: representing the first-order forward difference of $u(x,t)$ with respect to time t i.e. $\triangle u_t(x,t) = u(x,t+1) - u(x,t)$. It describes the amount of change in the confirmed cases of cluster x in unit time t.

• $D(x)[\nabla(\triangle u_x(x,t))]$: representing the regional spreading (global process) of infectious disease between different clusters. Where $\nabla(\triangle u_x(x,t)) = u(x+1,t) - 2u(x,t) + u(x-1,t)$.

(1) $D(x)$: representing the spread of infectious diseases between different regions, so a piecewise function is used to represent $D(x)$, the value of each segment needs to be determined according to the actual situation.

• $r(t)u(x,t)[1-\frac{u(x,t)}{K}]$: representing the spread process (local process) within a cluster. This mathematical expression has been used to describe and predict the dynamics of various populations, such as the growth of bacteria and tumors [14].

(1) $r(t)$: the growth rate with time t in local process. Therefore, the form $r(t)$ can be expressed as $r(t) = A + e^{-B(t-C)}, A, B, C$ are parameters and $A, B, C > 0$. Their optimal value will be determined by the actual data collected by us.

(2) K: the carrying capacity (the maximum possible volume of u at a given location x).

• $u(x,1) = \varphi(x)$: The initial function (number of confirmed cases at time $t = 1$ to be $\varphi(x)$, which specifies that the initial function has to be always ≥ 0).

• $\nabla u_x(1,t) = \triangle u_x(10,t) = 0$: is the Neumann boundary condition. For simplicity, we assume no infectious disease spread across the boundaries at $x = 1, 10$. That is, there is no movement of infected persons between the area corresponding to $x = 0$ and $x = 11$ and the study area, which is an ideal condition.

2.5. Evaluation indices

2.5.1 Error evaluation indices

In order to comprehensively evaluate the performance of the prediction model, we use the following two evaluation indicators to estimate the prediction error: the mean absolute error (MAE) and the average forecasting error rate (AFER). These indicators are calculated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|x(t) - y(t)|,$$

$$\text{AFER} = \frac{1}{n}\sum_{t=1}^{n}\frac{|x(t) - y(t)|}{x(t)} \times 100\%,$$

where $x(t)$ is the actual number of confirmed cases, $y(t)$ is the predict number of confirmed cases and n is the number of samples in the data set.

2.5.2 Pearson's linear correlation coefficient

The Pearson correlation coefficient is a measure of the degree of correlation between two variables X and Y. It is a value between 1 and -1, and a coefficient value of 1 means that X and Y can be well described by the equation of a line, and all the data points fall well on a line, and Y increases as X increases. A coefficient value of -1 means that all data points fall on the line, and Y decreases as X increases. A coefficient of 0 means that there is no linear relationship between the two variables. It can be calculated as follows:

$$R^2 = \frac{\sum_{t=1}^{n}(x(t) - \overline{x}(t))(y(t) - \overline{y}(t))}{\sqrt{\sum_{t=1}^{n}(x(t) - \overline{x}(t))^2} \cdot \sqrt{\sum_{t=1}^{n}(y(t) - \overline{y}(t))^2}}$$

where $x(t)$ is the actual number of newly confirmed cases per day, $\overline{x}(t)$ is the actual average number of newly confirmed cases, $y(t)$ is the forecast of the number of newly confirmed cases per day, $\overline{y}(t)$ is the predicted average number of newly confirmed cases and n is the number of samples in the data set.

3. Results

Using the difference model proposed in this paper, the real-time prediction effect of the model was verified using the cumulative 260-day consecutive the total confirmed cases and the newly confirmed cases in the United States since 10 March, 2020.

3.1. Predicted results based on the difference equation model

After collecting COVID-19 data from throughout the United states, we conclude the prediction process as follows: First, we normalize the data, reduce the experimental error and calculate the newly confirmed and total confirmed COVID-19 cases per cluster. Then, the 5-day training data set is applied to predict the confirmed cases of COVID-19 the next day. we use Days 1-5, 2-6, 3-7,...as training data and predict the next day 6, 7, 8,...correspondingly and record the prediction accuracy for all 10 regions on the Day 6, 7, 8,.... Specifically, we give the detailed prediction process for Day 6 as an instance. The first day's data are used to build the initial function. Next, 1-5 days' data is used to calculate the parameters in the model through the Lsqcurvefit function in MATLAB. Finally, we use the obtained parameters to predict the data on Day 6.

Figure 5 shows the prediction results of the total confirmed cases in the 1-10 regions from 10 March, 2020 to 24 November, 2020. By observing the image, we find that difference equation model can predict confirmed cases effectively. The forecast curve (red line) is highly similar to the actual curve (black line), and the two curve almost overlap. The green line is the 95% confidence interval, and we can also find that all the prediction curve are within in the 95% confidence interval. Therefore, the difference equation model in this article accurately estimates the alteration trend of the total confirmed cases, and the predicted trend is basically consistent with the actual trend of change.

3.2. Correlation and mean absolute error

Figure 6 shows the actual number of newly confirmed cases and the predicted number of newly confirmed cases in 10 state clusters for 259 days. The actual number of newly confirmed cases on the first day is equal to the total number of confirmed cases on the second day minus the total number of confirmed cases on the first day, and the corresponding projected number of newly confirmed cases is equal to the projected total number of confirmed cases on the second day minus the predicted total number of confirmed cases on the first day. By observing the images, it is found that the number of newly confirmed cases each day estimated by the difference equation model is very close to the actual number of newly confirmed cases.

Figure 7 shows a cross-validation diagram of actual and predicted newly confirmed cases. As can be seen from Figure 7, all data points are clustered near the linear regression equation, and the minimum and maximum Pearson correlation coefficients of the 10 clusters can be calculated to be 0.6057 and 0.9573. Hence, we can get a high degree of similarity between the actual value and the predicted value. In addition, Table 1 shows the actual confirmed cases and predicted confirmed case performance indicators for each day from 10 solstice March 2020 to 24 November, 2020 for the 10 clusters. In the chart, MAE-T represents the mean absolute error of total confirmed cases and MAE-N represents the mean absolute error of newly

confirmed cases. According to this table, we find that the minimum and maximum MAE-T values are 332 and 3556, and the average MAE-T values of the 10 clusters are 1327. The minimum value and maximum value of MAE-N are 331 and 1648 respectively, and the average MAE-N value is 878. The minimum AFER value is 0.91%, the maximum value is 2.67%, and the average AFER value is 1.57%. In addition, we found an average R^2 value of 0.8448 for the 10 clusters. These results demonstrate the success of the difference equation model in predicting the number of COVID-19 cases per day in the recent past.

In the overall evaluation of the research results, the above results can be obtained. The difference equation model with temporal and spatial characteristics can accurately predict the spread of infectious diseases when they come.

Table 1. Evaluation indicators between actual confirmed cases and predicted confirmed cases in each cluster from 10 March to 24 November 2020.

Regions	Days	MAE-T	MAE-N	AFER	R^2	Linear regression equation
Cluster 1	260	529	558	1.06%	0.6057	$y = 0.7956x + 333.1$
Cluster 2	260	965	560	0.91%	0.9337	$y = 9678x + 100.61$
Cluster 3	260	488	448	1.57%	0.9361	$y = 0.9941x + 31.35$
Cluster 4	260	3556	1648	2.67%	0.869	$y = 0.9297x + 761.5$
Cluster 5	260	2112	1376	1.67%	0.9485	$y = 1.0006x + 64.749$
Cluster 6	260	2239	1418	2.65%	0.8159	$y = 0.8855x + 782.66$
Cluster 7	260	852	1157	1.20%	0.6618	$y = 0.8325x + 532.08$
Cluster 8	260	580	336	1.05%	0.9573	$y = 1.0034x + 15.749$
Cluster 9	260	1622	955	1.47%	0.8981	$y = 0.9267x + 428.72$
Cluster 10	260	332	331	1.46%	0.8226	$y = 0.9107x + 122.48$
Mean		1327	878	1.57%	0.8448	

4. Conclusions

As can be seen from Figure 2(a,b), clusters 4, 5, 6 and 9 have the largest number of confirmed cases, accounting for 56% of the total confirmed cases in the United States, and it can be clearly found that the growth rate of confirmed cases in these 4 clusters is significantly higher than the other 6 clusters. From Figure 2(c,d), it can be seen that the death toll of the five clusters 2, 4, 5, 6 and 9 is significantly higher than that of the other clusters, accounting for 78% of the total deaths. Therefore, in terms of epidemic prevention and control, the government should focus on the 4, 5 and 6 city clusters.

This paper establishes a difference equation model with temporal and spatial characteristics to predict the spread of COVID-19. The results of the model show that there is a good fit between the actual total confirmed cases and the predicted

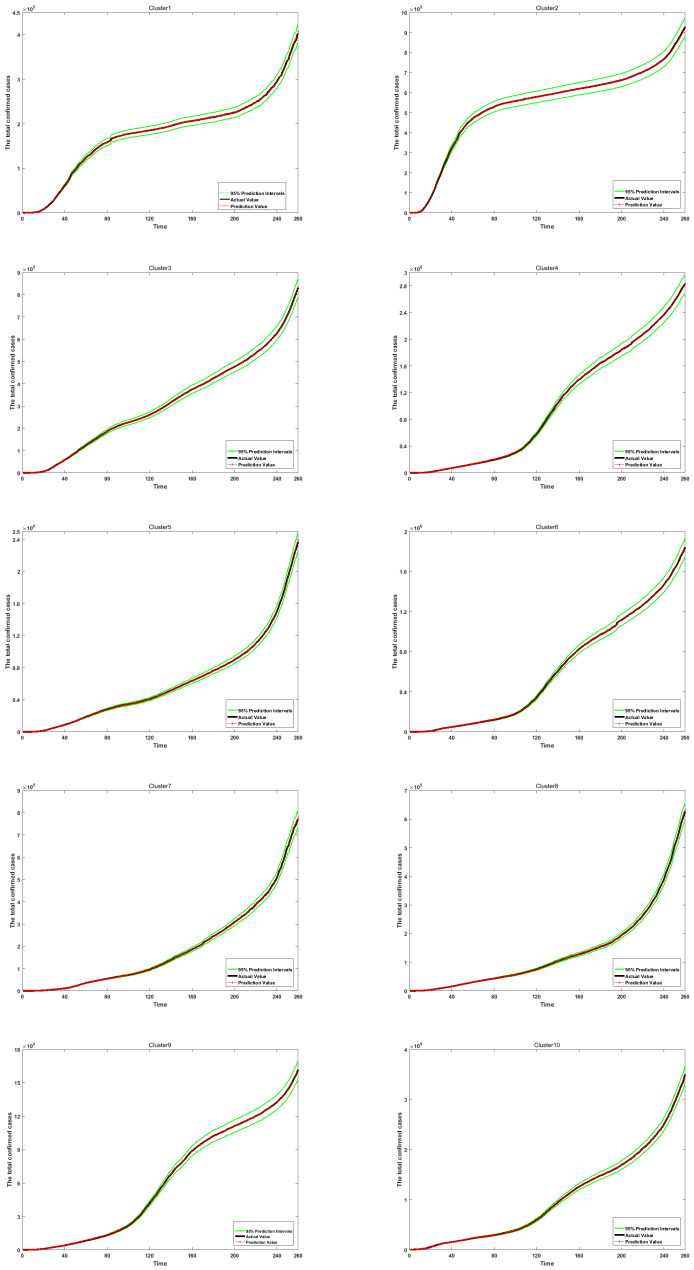


Figure 5. The prediction curve of the total confirmed cases in 10 different clusters from 10 March, 2020 to 24 November, 2020. The black line represents the actual value, the red line represents the predicted value and the green line represents the 95 confidence interval line.

total confirmed case data. In addition, the Pearson correlation coefficients between the actual newly confirmed cases and the predicted newly confirmed cases in the 10 clusters are: 0.6057, 0.9337, 0.9361, 0.869, 0.9485, 0.8159, 0.6618, 0.9573, 0.8981 and 0.8226 respectively, the average R^2 value is 0.8448, and the mean absolute error of the total confirmed cases and newly confirmed cases in 10 clusters is 1327 and 878.

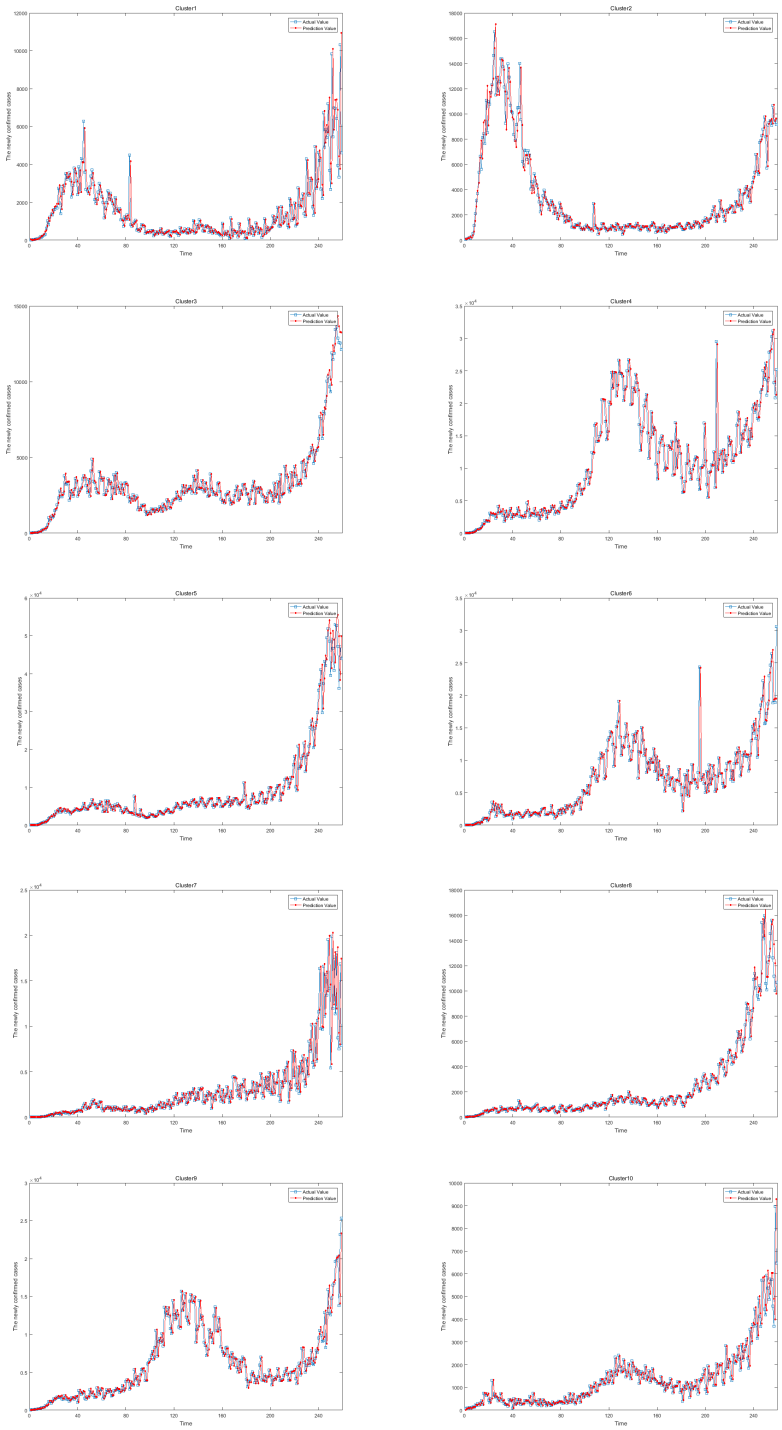


Figure 6. The predicted value of newly confirmed cases in 10 different clusters from 10 March, 2020 to 24 November, 2020. The blue line represents the actual value, the red line represents the predicted value.

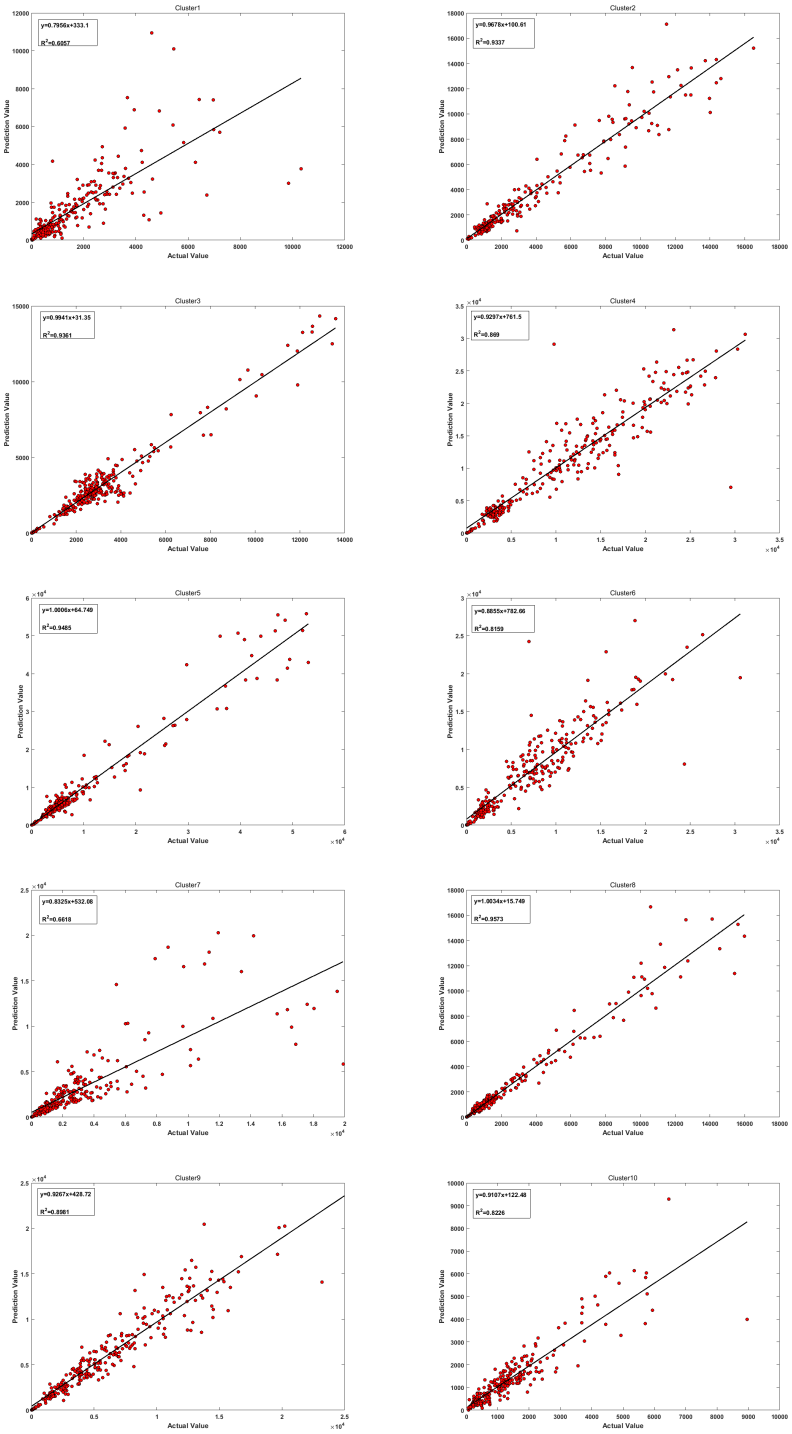


Figure 7. The cross-validation graph of the actual newly-confirmed cases and the predicted value of newly-confirmed cases in 10 clusters in 259 days, and the corresponding linear regression equation.

Therefore, the actual value is highly similar to the predicted value. The prediction results of this study show that the outbreak in the United States is still in the stage of large-scale outbreak, and our quantitative analysis can estimate the progress of the epidemic in different regions of the United States, which will help scientific research and policy makers to understand the different dynamics and situations in different regions of the United States in a simple and transparent way. However, due to different time periods and different policies implemented by the government, the forecast results of the model at the later stage may not be consistent with the actual results.

References

[1] J. Arino and S. Portet, *A simple model for COVID-19*, Infectious Disease Modelling, 2020, 5, 309–315.

[2] G. Barmparis and G. Tsironis, *Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach*, Chaos, Soliton & Fractals, 2020, 135, Article ID 109842.

[3] A. Benson, D. Gleich and C. Leskovec, *Higher-order organization of complex networks*, Science, 2016, 353(6295), 163–166.

[4] O. Bjrnstad, K. Shea, M. Krzywinski and N. Altman, *Modeling infectious epidemics*, Nature Methods, 2020, 17(5), 455–466.

[5] F. A. M. Cássaro and L. F. Pires, *Can we predict the occurrence of COVID-19 cases? Considerations using a simple model of growth*, Science of The Total Environment, 2020, 728, Article ID 138834.

[6] S. Choi, H. Seo and M. Yoo, *A multi-stage SIR model for rumor spreading*, Discrete Continuous Dynamical Systems-B, 2020, 25(6), 2351–2372.

[7] M. Ekum and A. Ogunsanya, *Application of hierarchical polynomial regression models to predict transmission of COVID-19 at global level*, International Journal of Clinical Biostatistics and Biometrics, 2020, 6(1), 1–18.

[8] A. Gorbalenya, et al., *The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2*, Nature Microbiology, 2020, 5(4), 536–544.

[9] L. Li, Z. Yang, et al., *Propagation analysis and prediction of the COVID-19*, Infectious Disease Modelling, 2020, 5(5), 282–292.

[10] K. Liang, *Mathematical model of infection kinetics and its analysis for COVID-19, SARS and MERS*, Infection, Genetics and Evolution, 2020, 82, Article ID 104306.

[11] C. Lin, A. K. H. Lau, et al., *A mechanismbased parameterisation scheme to investigate the association between transmission rate of COVID-19 and meteorological factors on plains in China*, Science of The Total Environment, 2020, 737, Article ID 140348.

[12] Y. Lin, W. Chi, Y. Lin and C. Lai, *The Spatiotemporal Estimation of the Risk and the International Transmission of COVID-19: A Global Perspective*, Scientific Reports, 2020, 10(1), Article ID 20021.

[13] B. Maier and D. Brockmann, *Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China*, Science, 2020, 368(6492), 742–746.

[14] L. Nátr and J. Murray, *Mathematical Biology. I. An Introduction*, Photosynthetica, 2020, 40, 414.

[15] D. Nguyen, G. Yin and C. Zhu, *Long-Term Analysis of a Stochastic SIRS Model with General Incidence Rates*, SIAM Journal of Applied Mathematics, 2020, 80(2), 814–838.

[16] O. Otunuga and M. Ogunsolu, *Qualitative analysis of a stochastic SEITR epidemic model with multiple stages of infection and treatment*, Infectious Disease Modelling, 2020, 5, 61–90.

[17] B. Peter, T. Tamas, V. Zsolt, D. Attila, A. Ferenc and R. Gergely, *Risk Assessment of Novel Coronavirus COVID-19 Outbreaks Outside China*, Journal of Clinical Medicine, 2020, 9(2), 571–583.

[18] F. Petropoulos and S. Makridakis, *Forecasting the novel coronavirus COVID-19*, PLoS One, 2020, 15(3), 231–236.

[19] J. Segars, et al., *Prior and novel coronaviruses, Coronavirus Disease 2019 (COVID-19), and human reproduction: what is known*, Fertil Steril, 2020, 113(6), 1140–1149.

[20] B. Tang, N. Bragazzi, Q. Li, S. Tang, Y. Xiao and J. Wu, *An updated estimation of the risk of transmission of the novel coronavirus (COVID-19)*, Infectious Disease Modelling, 2020, 5(5), 248–255.

[21] S. Tang, B. Tang, et al., *Stochastic Discrete Epidemic Modeling of COVID-19 Transmission in the Province of Shaanxi Incorporating Public Health Intervention and Case Importation*, MedRxiv, 2020.
DOI: 10.1101/2020.02.25.20027615

[22] O. Torrealba-Rodriguez, R. Conde-Gutierrez and A. Hernandez-Javier, *Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models*, Chaos, Solitons & Fractals, 2020, 138, Article ID 109946.

[23] S. Tuli, S. Tuli, R. Tuli and S. Gill, *Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing*, Internet Things, 2020, 11, Article ID 100222.

[24] H. Wang and Y. Nao, *Using a Partial Differential Equation with Google Mobility Data to Predict COVID-19 in Arizona*, Mathematical Biosciences and Engineering, 2020, 17(5), 4891–4904.

[25] H. Wang, K. Xu, Y. Kang, H. Wang, F. Wang and A. Avram, *Regional Influenza Prediction with Sampling Twitter Data and PDE Model*, International Journal of Environmental Research and Public Health, 2020, 17(3), 678–690.

[26] H. Wang, F. Wang and K. Xu, *Modeling Information Diffusion in Online Social Networks with Partial Differential Equations*, Springer, New York, 2020.
DOI: 10.1007/978-3-030-38852-2

[27] Y. Yang, L. Zou, T. Zhang and Y. Xu, *Dynamical analysis of a diffusive SIRS model with general incidence rate*, Discrete Continuous Dynamical Systems-B, 2020, 25(7), 2433–2451.