MEDWIN PUBLISHERS
Committed to Create Value for researchers

Journal of Infectious Diseases & Travel Medicine
ISSN: 2640-2653

# Biostatistical Analysis of the Novel Coronavirus

Zhao B[1]* and Cao J[2]

[1]School of Science, Hubei University of Technology, China

[2]School of Information and Mathematics, Yangtze University, China

***Corresponding author:** Dr. Bin Zhao, School of Science, Hubei University of Technology, Wuhan, Hubei, China, Tel: +86 130 2851 7572; Email: zhaobin835@nwsuaf.edu.cn

**Research Article**

**Volume 4 Special Issue 1**

**Received Date:** May 18, 2020

**Published Date:** June 29, 2020

DOI: 10.23880/jidtm-16000S1-002

## Abstract

**Background:** Since the first appearance of novel coronavirus (COVID-19) in Wuhan in December 2019, it has quickly swept the world and become a major security accident facing humanity nowadays. While threatening people's lives, the economies of various countries have also been severely damaged because of the epidemic. Because of the epidemic, it leading to the closure of a large number of companies, employment is becoming more difficult and people's lives have been greatly affected. So to the Hubei Province, where the COVID-19 first broke out, and the United States, the most severely affected area, we establish time series models to analyze the spread of the new coronavirus and short-term forecasts. This will help countries better understand the development trend of the epidemic, and make better preparations, timely intervention and treatment to prevent the further spread of the virus.

**Methods:** The data that collected from Hubei Province from 20 January, 2020 to 28 April, 2020 includes the cumulative confirmed diagnoses, death and cure. We use Excel to organize the data first, and then use SPSS to establish time series models and statistical analysis. Because there is no problem of missing data, so we define the day as the time variable, make time series graphs and observe the overall change rule. We remove the outliers, then use the SPSS expert modeler to automatically find the best fitting model for each dependent sequence, and predict by designating the independent variable and setting the width of the confidence interval to 95%. ACF and PACF graphs of residuals and Q test are used to determine whether the residual is a white noise sequence and whether the model is an appropriate model. The Holt model is used for the cumulative confirmed diagnoses in Hubei Province, and the ARIMA (1,2,0) model is used for cumulative cures and deaths in Hubei Province. Because the outbreak in the United States is later than China, we collect data from 29 February, 2020 to 28 April, 2020, which also includes the cumulative confirmed diagnoses, deaths and cures. The ARIMA (2,2,6) model is used for cumulative diagnoses in U.S., the ARIMA (0,2,0) model is used for cumulative deaths in U.S., and the ARIMA (0,2,1) model is used for cumulative cures in U.S.

**Findings:** From our modeling of the data, the time series diagrams of the real the fitted data almost overlap, so the fitting effect of the Holt model and the ARIMA model we use is very suitable. We compare the predicted values with the real values of the same period and found that the epidemic situation in Hubei Province has basically ended after May, but the epidemic situation in the United States has become more severe after May, so the Holt model and the ARIMA model are also very appropriate in predicting the epidemic situation in short-term.

**Interpretation:** Because the Chinese government has always put the safety of people's lives in the first place, when the epidemic broke out, it decisively closed the city of Hubei Province. One side is in trouble, all sides support; they concentrate all resources of whole country to save Hubei Province at the expense of the economy only in order to save more people. Now we can clearly see that the epidemic has been controlled in China and the whole country is developing in a good direction. In contrast, the epidemic in the United States, because of the government's lack of control, unwillingness to sacrifice the economy, premature return to work, and failure to call on people to wear masks, will lead to the epidemic in the United States has been going in a bad direction.

**Keywords:** COVID-19; Time Series Analysis; Holt Model; ARIMA Model

**Abbreviations:** COVID-19: Coronavirus Disease; MERS: Middle East Respiratory Syndrome; SARS: Severe Acute Respiratory Syndrome; ARIMA: Autoregressive Integrated Moving Average.

## Introduction

As a serious respiratory infectious disease, the COVID-19 [1-5] has been spreading around the world since January, 2020, which seriously threatens people's lives and normal lives. It has spread in China since December 2019 and has been basically controlled in May. However, during this period, the COVID-19 had a great impact on people's lives and national economic development. COVID-19 is a large family of viruses known to cause colds and more serious diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The COVID-19 is a new strain of coronavirus that has never been found in humans before.

Common signs of people infected with COVID-19 include respiratory symptoms, fever, cough, shortness of breath, difficulty breathing and so on. In more severe cases, infection can cause pneumonia, severe acute respiratory syndrome, kidney failure, and even death. Unfortunately, there is currently no specific treatment for diseases caused by COVID-19. However, many symptoms can be managed, so it needs to be treated according to the patient's clinical situation. In addition, auxiliary care for infected persons may be very effective.

In the global fight against the epidemic in the process, we can see that China is the best country to do so. It has controlled the spread of the novel coronavirus in the most efficient way, providing other countries with valuable Chinese experience and Chinese pattern in the fight against the epidemic. By establishing COVID-19 spread models between China and the United States, we can clearly see the different results of different policies and measures on the control of the epidemic. This is also the significance of our model. Through accurate data and rational analysis, we can provide the future development trend and control strategy research for the world in the fight against the epidemic [6].

## Methods

### Data

The data of Hubei Province is derived from the Health Commission of Hubei Province on its official platform from 20 January, 2020. The Hubei Province data collected in this paper includes cumulative deaths, cumulative cures and cumulative diagnoses from 20 January, 2020 to 28 April, 2020. The data of the United States comes from the domestic data platform, *the Real-time Big Data Report of the Epidemic*. The data collected in this paper includes cumulative deaths, cumulative cures and cumulative number of diagnoses from 29 February, 2020 to 28 April, 2020.

### The Model

Through the collected data, we conduct a time series analysis of the novel coronavirus [7-9]. Because there is no data missing, we import the data into SPSS software, define the day as a time variable, remove the outliers, and make a time series graph. The most suitable fitting models are automatically found by the expert modeler, which include the cumulative deaths, cumulative cures and cumulative diagnoses.

### TS Model-Based Method for Estimation in Hubei Province

Based on the given data of cumulative number of diagnoses in Hubei Province, we use the expert modeler to process the data and obtain the Holt model to describe it. The corresponding equation set shown below.

$$\begin{cases} S_t = \propto X_t + (1-\propto)(S_{t-1} + T_{t-1}) \\ T_t = \gamma(S_t - S_{t-1}) + (1-\gamma)T_{t-1} \\ \hat{X}_t(m) = S_t + mT_t \end{cases} \quad \text{Equation set 1}$$

The related mathematical symbols used above are listed in the following Table 1.

| Classes | Meanings for each classes |
|---|---|
| $t$ | Current period |
| $X_t$ | Actual observations in period $t$ |
| $S_t$ | Estimated level at period $t$ |
| $T_t$ | Predicted trend at period $t$ |
| $\hat{X}_t(m)$ | Estimated value before period $m$ |
| α | Horizontal smoothing parameter |
| β | Trend smoothing parameter |

Finally, we find that α =1, β= 0.1.

**Table 1:** Mathematical symbols used in equation set ①.

In order to determine that the selected Holt model can correctly describe the cumulative number of diagnoses, we use white noise [10] for residual test. As can be seen from Figure 1, the ACF and PACF graphs of the residuals, the autocorrelation coefficients and partial correlation coefficients of all lag orders are not significantly different from 0; From Table 2, it can also be seen that the P value obtained by performing the Q test on the residual is 1, which is, we cannot reject the null hypothesis, confirming that the residual is a white noise sequence.

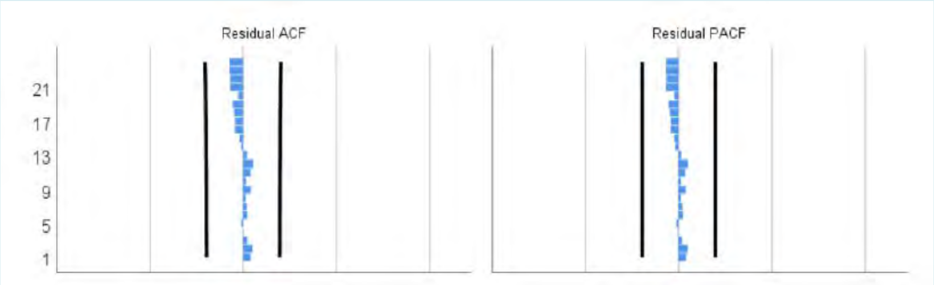

**Figure1:** The ACF and PACF graphs of the residuals in cumulative number of diagnoses in Hubei Province case.

| **Model Statistics** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Model Fit Statistics** | | | | **Ljung-Box Q(18)** | | |
| Model | Number of Predicators | Stationary R-squared | R-squared | Normalized BIC | Statistics | DF | Sig. | Number of Outliers |
| The Cumulative confirmed | 0 | 0.423 | 0.994 | 15.254 | 2.273 | 16 | 1.000 | 0 |

**Table 2:** The Q test on the residual of the case of cumulative number of diagnoses in Hubei Province case.

From the analysis above, we can conclude that the Holt model can describe the cumulative number of diagnoses well. Similarly, the expert modeler evaluates that the cumulative number of cures conforms the ARIMA (1, 2, 0) model. The related equations are written followed.

$$(1-\sum_{i=1}^{P} \propto_i L^i)(1-L)^2 y_t = \propto_0 + (1+\sum_{i=1}^{q} \beta_i L^i)\varepsilon_t$$

$$(1-\propto_1)(1-L)^2 y_t = \propto_0 + \varepsilon_t$$

Equations set 2 &3

We also use white noise to test this model for residuals. As Figure 2 below shows, the ACF and PACF graphs of the residuals can be seen that the autocorrelation coefficients and partial correlation coefficients of all lag orders are not significantly different from 0. Therefore, the ARIMA (1, 2, 0) model can well describe the cumulative number of cured people.

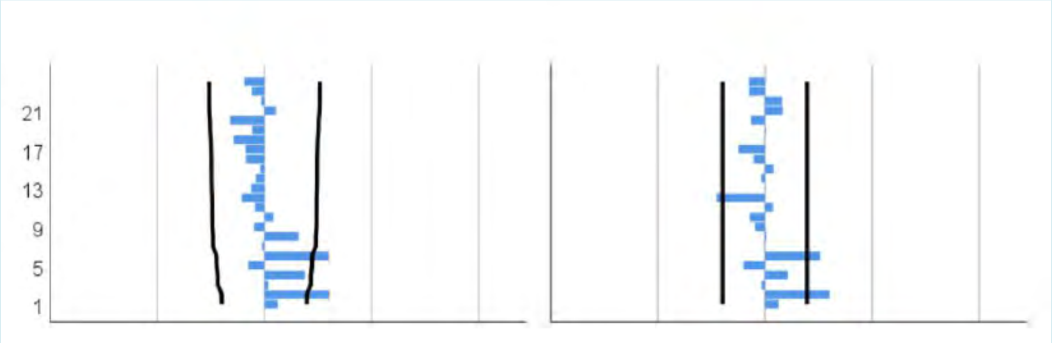

**Figure 2:** The ACF and PACF graphs of the residuals in cumulative number of cures in Hubei Province case.

The related mathematical symbols used above are listed in the following Table 3.

| Classes | Meanings for each classes |
|---|---|
| p | Number of autoregressive items |
| q | Moving average number of items |
| L | Lag operator |
| $\varepsilon_t$ | Because { $\varepsilon_t$ } is a white noise sequence, E ( $\varepsilon_t$ )=0 |
| $1-\sum_{i=1}^{p}\propto_i L_i$ | AR(p) model |
| $1+\sum_{i=1}^{q}\beta_i L^i$ | MA(q) model |
| $(1-L)^2$ | 2nd order difference |
| $y_t$ | Number of people on day t |

**Table 3:** Mathematical symbols used in equations ②③.

Similar to the cumulative number of people cured in Hubei Province, the cumulative number of deaths in Hubei Province also conforms to the ARIMA (1, 2, 0) model. The corresponding equations and the related symbol meanings are the same as equations ②.

Next, we use white noise for the residual test. From Figure 3, the ACF and PACF graphs of the residuals can be seen that the autocorrelation coefficients and partial correlation coefficients of all lag orders are not significantly different from 0. We can find out that the ARIMA (1, 2, 0) model can also describe the cumulative death toll well.

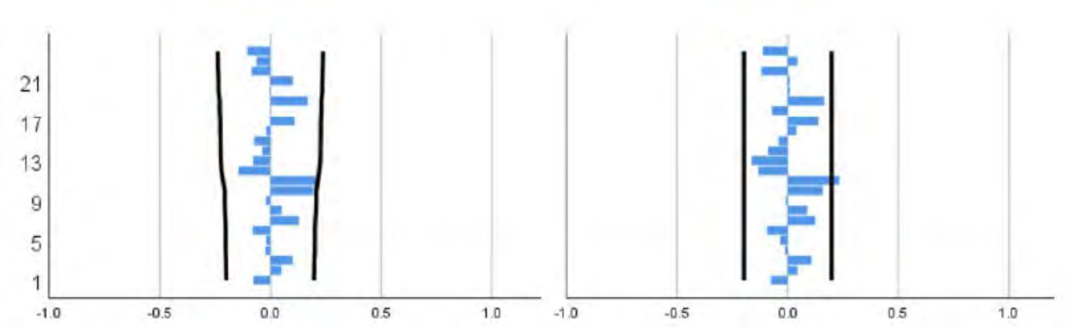

**Figure 3:** The ACF and PACF graphs of the residuals in cumulative number of deaths in Hubei Province case.

### TS Model-based method for estimation in the United States

Based on the given data of America, we use the expert modeler to process the data and we find that all of the cumulative number of diagnoses, deaths and cures conform the ARIMA model. However, the parameters setting of each group of them are not identical. After processing, it is found that the cumulative diagnosis in the United States applies the ARIMA (2, 2, 6) model. The corresponding equations are the same with equations ②.

Next, we performed a residual test on the model based on white noise. As can be seen from Figure 4, the ACF and PACF graphs of the residuals, the autocorrelation coefficients and partial correlation coefficients of all lag orders are not significantly different from 0. From Table 4, it can also be seen that the P value obtained from the Q test of the residual is 0.304, that is, we cannot reject the null hypothesis, and think that the residual is a white noise sequence. Therefore, the ARIMA (2,2,6) model can well describe the cumulative number of diagnoses.

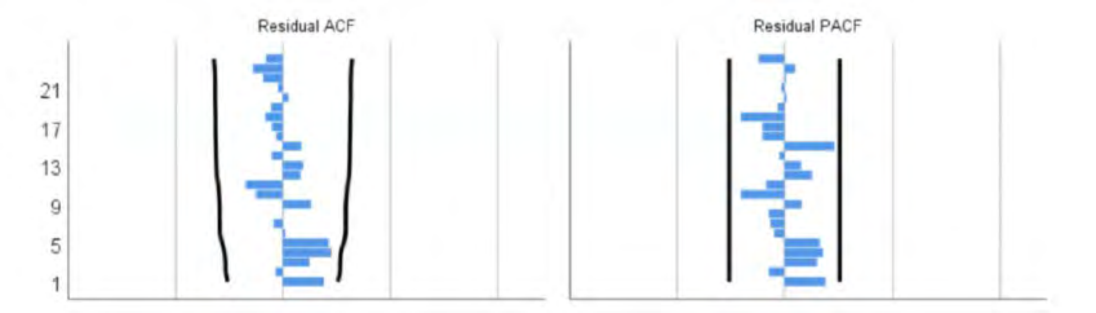

**Figure 4:** The ACF and PACF graphs of the residuals in cumulative number of diagnoses in U.S. case.

| Model Statistics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Model Fit Statistics | | | Ljung-Box Q(18) | | | |
| **Model** | **Number of Predicators** | **Stationary R-squared** | **R-squared** | **Normalized BIC** | **Statistics** | **DF** | **Sig.** | **Number of Outliers** |
| The Cumulative confirmed | 0 | 0.793 | 0.999 | 19.470 | 17.247 | 15 | .304 | 3 |

**Table 4:** The Q test on the residual of the case of cumulative number of diagnoses in U.S. case The cumulative number of cures conforms ARIMA (0, 2, 0) model, which is equals to 2nd order difference equation. The related equation set is similar to the equations ②.

As usual, we should use white noise to perform a residual test. As the Figure 5 shows below, the ACF and PACF graphs of the residuals can be seen that the autocorrelation coefficients and partial correlation coefficients of all lag orders are not significantly different from 0.

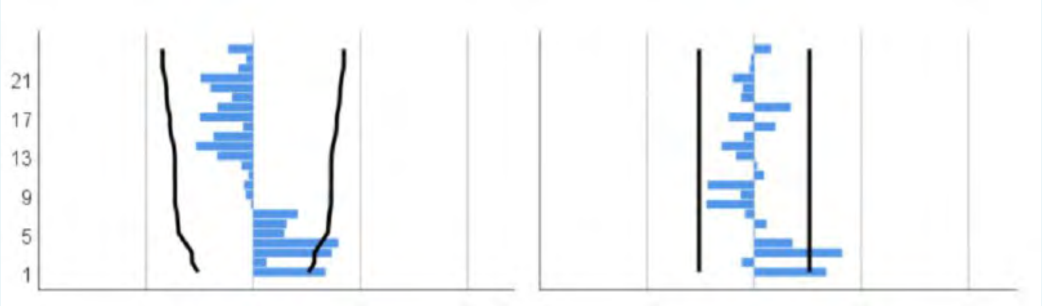

**Figure 5:** The ACF and PACF graphs of the residuals in cumulative number of cures in U.S. case.

At last, we use expert modeler to process the data of cumulative deaths in U.S., then it is found that it conforms the ARIMA (0, 2, 1) model. The related equation set still conforms with the equations ②. Then we perform a white noise residual test. As can be seen from Figure 6, the ACF and PACF graphs of the residuals, the autocorrelation coefficients and partial correlation coefficients of all lag orders are not significantly different from 0.

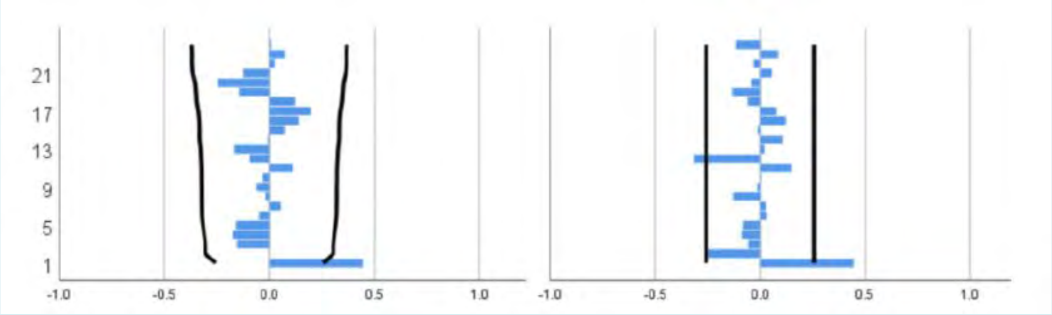

**Figure 6:** The ACF and PACF graphs of the residuals in cumulative number of deaths in U.S. case.

## Results

### The TS Model-Based Method Results in Hubei Province

We set the width of the confidence interval to 95%, and then use the Holt and ARIMA model to fit and predict the cumulative number of people diagnosed, cumulatively cured and cumulatively died in Hubei Province respectively. The obtained results are shown in the following Figures 7-9.

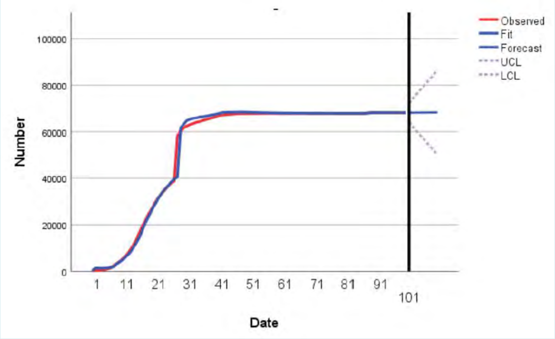

**Figure 7:** The cumulative number of diagnoses in Hubei Province.

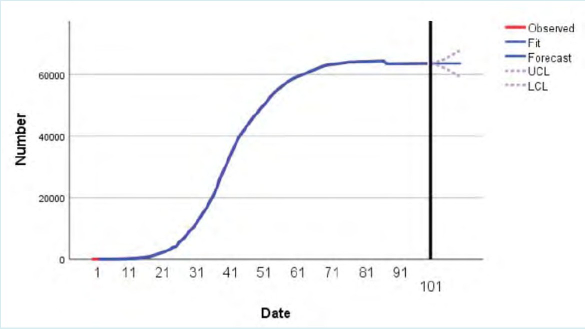

**Figure 8:** The cumulative number of cures in Hubei Province.

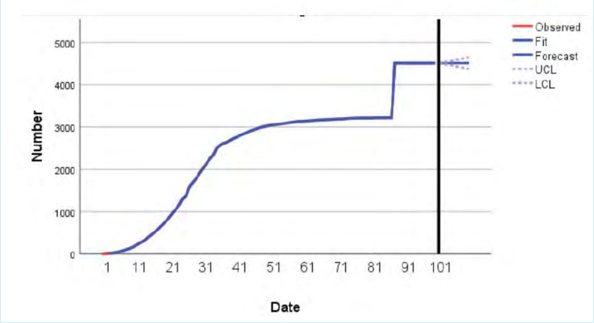

**Figure 9:** The cumulative number of deaths in Hubei Province.

It can be seen from Figures 7-9 that the timing charts of the real data and the fitted data almost overlap, and the Holt model and the ARIMA model fit the original data very well. At the same time, after 28 April, the epidemic situation in Hubei Province has been controlled, and the cumulative number of diagnoses will basically not increase dramatically. The Holt model and ARIMA model can also well predict the cumulative diagnoses, cumulative cures and cumulative deaths. The Table 5 listed below contains the predicted values of cumulative number of diagnoses, cures and deaths in Day 101 to 105 by using the equation ② and the Figures 7-9.

| Day | Cumulative diagnoses | Cumulative cures | Cumulative deaths |
|---|---|---|---|
| 101 | 68140 | 63616 | 4512 |
| 102 | 68152 | 63616 | 4512 |
| 103 | 68164 | 63616 | 4512 |
| 104 | 68176 | 63616 | 4512 |
| 105 | 68188 | 63616 | 4512 |

**Table 5:** Short-term predicted values in Hubei Province.

### The TS Model-based method results in the United States

Our analysis of the U.S. epidemic situation also set the width of the confidence interval to 95%. The results obtained by fitting and predicting the number of cumulative diagnoses, cumulative cures and cumulative deaths using the ARIMA model are shown in the following Figures 10-12.

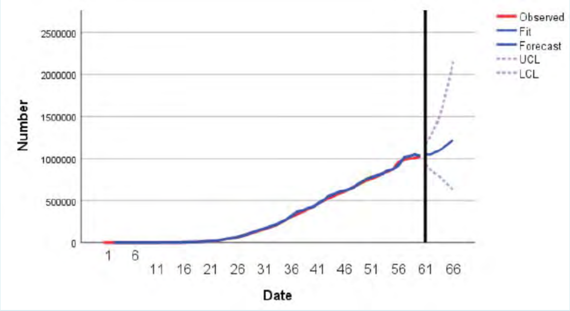

**Figure 10:** The cumulative number of diagnoses in U.S.

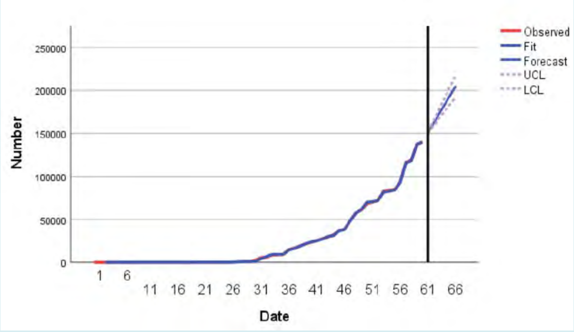

**Figure 11:** The cumulative number of cures in U.S.

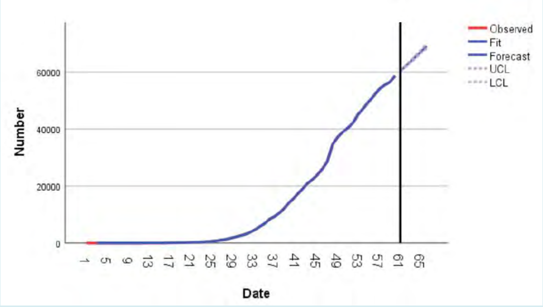

**Figure 12:** The cumulative number of deaths in U.S.

It can be seen from Figures 10-12 that the timing charts of the observed data and the fitted data almost overlap, and the ARIMA model fits the original data very well. At the same time, after 28 April, the cumulative number of diagnoses, cumulative deaths and cumulative cures will continue to increase substantially in the short term, which is related to the policies adopted by the US government to combat the epidemic. Not only has the epidemic situation not been controlled, but the situation has become more severe. This shows that the ARIMA model can also predict the cumulative diagnoses, cumulative deaths and cumulative cures. The specific predicted values of cumulative number of diagnoses, cures and deaths in Day 61 to 65 are listed in the Table 6 below.

| Day | Cumulative diagnoses | Cumulative cures | Cumulative deaths |
|---|---|---|---|
| 61 | 1049197 | 151006 | 60312 |
| 62 | 1051063 | 161873 | 62006 |
| 63 | 1080926 | 172741 | 63722 |
| 64 | 1122025 | 183609 | 65458 |
| 65 | 1169266 | 194477 | 67217 |

**Table 6:** Short-term predicted values in U.S.

## Discussion

When the virus breaks out, everyone lives in panic, worrying about the safety of themselves and their families, and also worrying about the safety of the country. We use SPSS [11] to accurately get the models we need, such as Holt model and ARIMA model, and then use these models to fit the sequences, and estimate the model parameters based on the sequence values. Finally, we perform residual tests on the models with white noise to check whether the model is applicable. By analyzing the COVID-19 through time series, the development process, direction and trend of the epidemic are obtained. And we predict how the epidemic situation might develop in the future as well. This will give us some guidance in our lives, such as what measures should be taken to intervene in the development of the epidemic in order to save more lives.

Not only can time series be used for the analysis of infectious diseases, but also can be used in many disciplines in society such as measurement [12] and economics [13]. The data sequence and data size in the time series both contain information about the objective world, its changes and represent dynamic processes. Therefore, the main purpose of time series analysis is to understand the dynamic system under consideration, predict future events and control future events through intervention [14,15].

## Limitation

When establishing the model, we regard the data with large fluctuations as outliers. In fact, there are many more complex models that can catch these outliers. At the same time, when we make predictions, the ARIMA model is only suitable for short-term prediction. Over a certain period of time, the predicted value will not change any more, due to the essence of the model. So when solving this problem, we can assume the predicted values as the observed values, thus the long-term prediction would be possible, but the difference between the truly observed data might be more and more larger.

Since the epidemic is only predicted in the short term, it can be seen from the analysis chart of epidemic situation in the United States that the cumulative number of people who are cured, died and diagnosed, cases are all moving in an increasing direction. However, in practice, the number of people in these categories should reach a stable value in the end.

## References

1. Prevention and control of new coronavirus infection pneumonia epidemic. Health Commission of Hubei Province.
2. Epidemic real-time big data report. Coronavirus disease.
3. Holt model.
4. ARIMA Model.
5. 2019 New Coronavirus.
6. Ladi Wang (2005) Study on infectious disease model and control strategy. Beijing: China science and technology press.
7. Xuan Zhou (2019) Epidemiological characteristic and time series analysis of hand foot and mouth disease in Wuzhou city from 2014 to 2018. Guangxi: Guangxi medical university.
8. Yu G, Zhang J (2005) The application of time series analysis in the study of epidemic Disease. J Liaoning University, 2005(2).
9. Cryer JD, Kung-Sik C (2009) Time Series Analysis with Applications in R. Beijing: Machinery Industry Press.
10. White noise (Physics concept).
11. Statistical Product and Service Solutions (spss).
12. Sun T (2013) An Application of Time Series Analysis and Its Application in Measurement. Surveying and Spatial Geographic Information 36(3): 12-13.
13. Lan G (1994) Application of time series analysis in economy. Beijing: China Statistics Press.
14. Zhou Y (2015) Time series analysis and application. Beijing: Higher Education Press.
15. He S (2007) Applied time series analysis. Beijing: Peking University Press.