RESEARCH ON COVID-19 EPIDEMIC BASED ON ARIMA MODEL

Li Zhihao, Wang Junpei, Zhang Xiaoliang, Ning Huijun*

Department of Engineering Mechanics, Civil Engineering,
Henan University of Science and Technology, Luoyang, China

*Corresponding Author: Email: ninghuijun@haust.edu.cn

Abstract: Since the outbreak of the novel coronavirus, the epidemic has received extensive attention all over the world. In this article, an evaluation system was established to analyze the epidemic prevention and control situation of some countries. And the ARIMA model was built to predict the epidemic situation in a short period of time. Then taking the United States as an example, the predicted values of the number of newly diagnosed cases, death rate, and cure rate in 10 days were obtained, which were then compared with the actual data. It is shown from the results that the ARIMA model can be used to predict the epidemic and provide a decision-making basis for the current world’s epidemic prevention and control.

1. Introduction
The COVID-19 is an acute infectious pneumonia. The symptoms of the patient mainly include fever, dry cough, fatigue and dyspnea. It has not been completely defeated by humans as of today. And it has now spread all over the world and has had a huge impact on people’s lives, property, and national development worldwide. All countries have done their best to control the spread of the virus, both in terms of human and financial resources. However, due to the uneven development level in different countries, shortcomings in the prevention and control of the epidemic were exposed. Some countries such as China and the United States have successfully developed anti-coronavirus vaccine, which greatly reduced the death rate of the infected. Meanwhile, it provided effective protection for susceptible people. But the situation in many countries is still grim. The death number in India is still increasing. At present, most predictions of the new crown epidemic are based on the use of the infectious disease transmission dynamics models such as SEIR model [1], SIR model [2] and statistical models such as time series analysis [3]. The infectious disease transmission dynamics model requires the understanding of various model parameters and their future numbering trends, which are difficult to obtain. While time series models only need the historical sequence of the number of cases to construct the prediction model of the number of cases.

The autoregressive integrated moving average model (ARIMA) in the time series model has the characteristics of strong short-term predictability and simplicity, and is widely used in the forecast of infectious diseases.
2. Research Methods

2.1 Theoretical Basis

Autoregressive Integrated Moving Average model (ARIMA) is widely used in all aspects of influenza prediction. The model with the following structure is an autoregressive summation moving average model [4].

\[
\phi(B)\nabla^d x_t = \theta(B)\varepsilon_t, \\
E(\varepsilon_t) = 0, \text{Var}(\varepsilon_t) = \sigma^2, E(\varepsilon_t\varepsilon_s), s \neq t \\
E(x_t\varepsilon_t) = 0, \forall s < t
\]

\(\nabla^d = (1-B)^d\), \(\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p\) is autoregressive coefficient polynomial of smoothly reversible ARIMA \((p, d, q)\) model. \(\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q\) is moving smooth coefficient polynomial of smoothly reversible ARIMA \((p, d, q)\) model. “AR” is Autoregression; “p” is autoregressive coefficient, can be estimated by autocorrelation graph; “MA” is moving average; q is number of moving average terms, can be estimated with partial autocorrelation graph; “d” is the order of difference made when the time series becomes stationary.

2.2 Establish an ARIMA Model (Take the United States as an Example)

2.2.1 The ARIMA Model Modeling Process

(1) Model Recognition

The basic form of the ARIMA model is ARIMA \((p, d, q)\), “p” is autoregressive coefficient, “d” is the order of difference made when the time series becomes stationary, “q” is number of moving average terms. First perform stationarity test on the time series, if the sequence is not stationary, use methods such as difference and logarithm to make the sequence stationary. Then use the autocorrelation function (ACF) graph and partial autocorrelation function (PACF) graph recognize and rank the model. For the order of P and Q, usually try from low-level to high-level, check the fit of each model and compare, generally more than 2 levels are rare.

(2) Estimation of Model Parameters and Test of Goodness of Fit

Perform statistical tests on the parameters in the model, determine whether it is statistically significant. Use Ljung-Box statistics to test whether the residual sequence is a white noise residual. Determine whether the model fully extracts all trend information of the original sequence. At the same time, use the identified Smooth R^2, Bayesian Information Criterion (BIC) and other different models to compare the goodness of fit and select the best model.

(3) Predict Application

Use the established model to make predictions. This study uses IBM software for data processing and modeling analysis. The modeling process is shown in Figure 1.
2. 3 Model Recognition

The time series of the number of new cases in the United States from January 1 to April 2, 2021 is shown in Figure 2. These data are obviously not stationary series. Therefore, the sequence needs to be smoothed and transformed. The sequence time span is 3 months. At the same time, the sequence itself does not show periodic changes and there is no need to consider seasonal factors. Therefore, a general difference is performed on the sequence to eliminate the trend influence. Figure 3 shows the sequence after a difference using SPSS software. It can be seen that the sequence is evenly distributed on the upper and lower sides of the 0 value. Therefore, the series can be regarded as a stationary series. ACF shows obvious tailing. PACF shows censoring (Figure 4). Use maximum likelihood estimation method to estimate the parameters of the model. Combining Akaike information criterion (AIC) and Bayesian Information Criterion (BIC), autocorrelation function graph, partial autocorrelation function graph, etc., the optimal model is ARIMA (1, 1, 2).
2.4 Model Parameter Estimation and Model Checking

The null hypothesis for model testing is that the coefficients of all parameters are 0, and model fitting statistics are used to compare the pros and cons of the fitting effects between models. The goodness of fit statistics given by SPSS have Smooth $R^2$, p value of the Q statistic, especially the larger the stable $R^2$ value, the better. ARIMA (1, 1, 2) model parameter test results, goodness of fit statistics and white noise test statistics of residual series are shown in Table 1. The fitting and prediction results of the ARIMA (1, 1, 2) model on the number of newly confirmed COVID-19 cases in the United States are shown in Figure 5. Combined with Figure 6 residual ACF and residual PACF lag diagram, we can see that most of the results are within the confidence interval, so the fitting effect is good. According to Table 1, it can be concluded that the p value in the Q test after removing an outlier value is 0.136>0.05. It shows that using the ARIMA (1, 1, 2) model to simulate the number of newly diagnosed people is very effective.

| Model          | Parameter Estimate | Smooth $R^2$ | Ljung-Box Q | P       |
|----------------|--------------------|--------------|-------------|---------|
| ARIMA (1, 1, 2)| Parameter Estimate | Smooth $R^2$ | Ljung-Box Q | P       |
| Constant       | -1839.05           | 0.442        | 16          | 0.136   |
| AR (1)         | -0.822             | -10.687      |             |         |
| MA (2)         | 0.523              | 4.314        |             |         |
2. 5 Forecast and Effect Evaluation

Select the optimal model of ARIMA (1, 1, 2,) as shown in Table 1. Use the calculation formula of the predicted value in the ARIMA (p, d, q) prediction model [3]. According to the parameters given by SPSS, the calculation formula of predicted data is obtained 2. The predicted data and actual data of the number of newly diagnosed people in the United States from April 3 to April 12 are calculated from Formula 2 as shown in Table 2. According to Table 2, the overall dynamics predicted by the model are basically consistent with the actual situation, the model makes a good prediction of the changing trend of the number of cases in the short term in the future.

\[(1 + 0.822t)\Delta y_t = -1839.045 + \varepsilon_t + 0.532\varepsilon_{t-2}\]  
(2)

Table 2 Comparison of Actual and Predicted Number of Newly Diagnosed People in the U.S. from April 3-12, 2021

| Date   | Actual Value | Predictive Value | UCL      | LCL      | The Absolute Value of the Forecast Error |
|--------|--------------|------------------|----------|----------|-----------------------------------------|
| April 3 | 69986        | 61033            | 114957   | 7110     | 8953                                    |
| April 4 | 68501        | 61270            | 116045   | 6494     | 7231                                    |
| April 5 | 37205        | 47726            | 115322   | 130      | 10521                                   |
| April 6 | 70232        | 67287            | 115940   | -1366    | 2945                                    |
| April 7 | 69875        | 64298            | 115255   | -6660    | 5577                                    |
| April 8 | 70805        | 63404            | 115545   | -8737    | 7401                                    |
| April 9 | 80161        | 70788            | 114905   | -13329   | 9373                                    |
| April 10| 85368        | 79587            | 114956   | -15782   | 5781                                    |
| April 11| 67208        | 57224            | 114348   | -19900   | 9984                                    |
| April 12| 48611        | 45815            | 114222   | -22591   | 2796                                    |

Note: ULC is the upper limit of the interval at the 95% confidence level, LCL is the lower limit of the interval with the confidence level below 95%.

Select the optimal model of ARIMA(0, 1, 0,) as shown in Table 3. Use the calculation formula of the predicted value in the ARIMA(p, d, q) prediction model [3]. According to the parameters of SPSS, the calculation formula (3) of predicted data is obtained (q=1>0. 05, white noise residual test passed). Using formula 3 to calculate the predicted data and actual data of the COVID-19 cure rate in the United States from April 3 to April 12 are shown in Table 4. According to Figure 7, the predicted value of the cure rate in the United States from April 3-12, 2021, is gradually increasing. The overall dynamics predicted by the model are basically consistent with the actual situation. The model makes a good prediction of the change trend of the cure rate in the short term in the future.

\[\Delta y_t = 0.002 + \varepsilon_t\]  
(3)

Table 3 Parameter Estimation and Model Test Results of the ARIMA (0, 1, 0) Model

| Model          | Parameter Estimation | Model Checking |
|----------------|----------------------|----------------|
| ARIMA (0, 1, 0)| Parameter Value      | t        | Smooth R² | Ljung-Box Q | P |
|                | Constant             | 0.002   | 3.962    | -2.22E-16  | 18| 1  |
Select the optimal model of ARIMA (1, 1, 7,) as shown in Table 5. Use the calculation formula of the predicted value in the ARIMA(p, d, q) prediction model [3]. According to the parameters of SPSS, the calculation formula (4) of predicted data is obtained. Using formula 3 to calculate the predicted data and actual data of the COVID-19 mortality rate in the United States from April 3 to April 12 are shown in Table 6. According to Figure 8, the mortality rate in the United States from April 3-12, 2021, has gradually stabilized at 1.81%, which is still relatively high compared to the world average. But it can be seen from the forecast chart that the mortality rate is still on a downward trend. Therefore, the overall situation of the United States for epidemic prevention and control is gradually becoming more comprehensive.

$$(1 + 0.635t)\Delta y_t = e_t - 0.325e_{t-7}$$ (4)

Table 5 Parameter Estimation and Model Test Results of the ARIMA (1, 1, 7) Model

| Model     | Parameter | Estimated Value | t    | Smooth R² | Ljung-Box Q | P     |
|-----------|-----------|-----------------|------|-----------|-------------|-------|
| ARIMA (1, 1, 7) | Constant | 0               | 0.701 | 16        | 0.143       |
|           | AR (1)   | 0.635           | 8.894 |           |             |       |
|           | MA (7)   | -0.328          | -3.613 |           |             |       |

Table 6 Comparison of actual and predicted mortality rates in the U.S. from April 3 to April 12, 2021

| Date     | Actual Value | Predictive Value | Prediction error |
|----------|--------------|------------------|------------------|
| April 3  | 1.813%       | 1.813%           | 0                |
| April 4  | 1.812%       | 1.813%           | 0.001%           |
| April 5  | 1.810%       | 1.812%           | 0.002%           |
| April 6  | 1.808%       | 1.812%           | 0.004%           |
| April 7  | 1.807%       | 1.812%           | 0.005%           |
| April 8  | 1.811%       | 1.811%           | 0                |
| April 9  | 1.809%       | 1.811%           | 0.002%           |
| April 10 | 1.808%       | 1.811%           | 0.003%           |
| April 11 | 1.806%       | 1.811%           | 0.005%           |
| April 12 | 1.804%       | 1.811%           | 0.007%           |
3. Conclusion
The ARIMA model is one of the most commonly used time series models. It is a model built by transforming a non-stationary time series into a stationary time series and then redressing the lag value of the series and the present value and lag value of the random error term. Not only the dependence on the time series is considered but also the interference of random fluctuations. Therefore, the prediction accuracy of short-term trends is relatively high, which is often used in the fields of epidemic trend prediction of infectious diseases and non-communicable diseases. In this article, the number of new confirmed cases, cure rate, and mortality in the United States from April 3 to April 12, 2021 were used to establish prediction models of new cases, the cure rate prediction, and death. It is shown in the results that the ARIMA (1, 1, 2) optimal model was closer to the actual observed value for the number of new cases. And the average error of the predicted value of the total number of new cases for 10 days was 7056.2. The selected ARIMA (0,1,0) optimal model compared the predicted value of the cure rate with the actual observed value, and the error was only 0.628%. The selected ARIMA (1,1,7) optimal model compared the predicted mortality rate with the actual observation value, the error was only 0.0029%. And the predicted change trend accurately reflects the actual development trend of the epidemic.

Short-term forecasts can be made very well by the ARIMA model. Compared with the cumbersome degree of the traditional SEIR model, the ARIMA is more accurate for forecasting results and the forecasting method is simpler. And the ARIMA model requires too high for residual white noise testing to get the appropriate data, resulting in fewer predictable indicators.

Acknowledgments
This article is supported by the Doctoral Research Fund of Henan University of Science and Technology (13480023)

References
[1] KENTARO IWATA, CHISATO MIYAKOSHI. A simulation on potential secondary spread of novel coronavirus in an exported country using a stochastic epidemic SEIR model [J]. Journal of Clinical Medicine, 2020(9):944
[2] BUSHRA ZAREIE, AMIN ROSHANI, MOHAMMAD ALIMANSOURNIA, et al. A model for COVID-19 prediction in Iran based on China parameters [J]. Arch Iran Med, 2020, 23(4):244
[3] SINGH RK, RANI M, BHAGAVATHULA AS, et al. Prediction of the COVID-19 Pandemic Integrated Moving Average (ARIMA) model [J]. JMIR Public Health Surveille, 2020, 6(2): e19115
[4] Wang Yan. Time Series Analysis-Analysis is Based on R [M]. China Renmin University Press, 2015
[5] Wen Liang, Huang Qingzhen, Wang Zhigang, et al. Analysis of the Results of Using the ARIMA Model to Predict the Development Trend of the COVID-19 in Pakistan [J]. Journal of Preventive Medicine of Chinese People's Liberation Army, 2020, 38(8):96-99