Method Article

Application of autoregressive moving average model in the prediction of COVID-19 of China

【 Abstract 】 **Objective** To establish ARIMA model through time series analysis to understand the occurrence law of newly confirmed cases of novel coronavirus pneumonia and provide references for taking epidemic prevention and control measures. **Methods** The cumulative confirmed and cured cases of COVID-19 are collected through the official website of the National Health Commission, and the number of newly confirmed and cured cases per week are sorted out. weanalyze the time series of newly diagnosed and cured COVID-19 cases every week from April 12, 2020 to December 5, 2021 by IBM SPSS 25.0 software. The model is established through model identification, parameter estimation and model fitting. **Results** The number of reported cases of COVID-19 has no obvious seasonal characteristics. The ARIMA (2,1,1) model well fittedthe time series, R^2 = 0.542/0.617. Through the residual white noise test, all parameters of the model have statistical significance, Ljung box q = 9.095/9.651, P> 0.05. We predict the cases and cures in the four weeks after December 5, 2021 by ARIMA(2,1,1). The measured values in the first week and the second week are within the predicted 95% CI range. **Discussion** The epidemiological characteristics of COVID-19 need a

longer time series for validation and analysis. ARIMA model can predict the incidence of COVID-19 in a short term, and the model should be constantly revised according to the actual situation.

[Key words] COVID-19; Time series analysis; ARIMA; Infectious Diseases; Forecasting.

1. INTRODUCTION

Corona virus disease (COVID-19)is an acute respiratory infectious disease which first appeared in Wuhan, China. The symptoms of COVID-19aresimilar to the influenza, which can be fever, cough, shortness of breath, pneumonia and breathing difficulties [1]. The transmission speed of COVID-19 is very fast because of its strong infectivity and population susceptibility[2]. Most SARS-CoV-2 infected patients present with mild respiratory symptoms,however, some can representvarious fatal complications including organ failure, septic shock, pulmonary oedema, severe pneumonia, and Acute Respiratory Distress Syndrome (ARDS) [3]. The seriously ill patients need to be hospitalized, or even use respirators. Since the original outbreak, COVID-19 has spread to 221 countries and territories [4]. On January30, 2020 the World Health Organization declareCOVID-19 to be a Public Health Emergency of International Concern (PHEIC) posing a high risk to countries with vulnerable health systems [5]. The effective methods to prevent the spread of the virus are early detection, isolation, prompt treatment, and the implementation of a robust system to trace contacts [6]. On the other hand, vaccination can also protect susceptible populations.The constant

variation of virus strains brings new challenges to epidemic prevention and control. The World Health Organization (WHO) has classified COVID-19 variant strains into two categories according toits risk level: worrying variant strainsand noteworthy variant strains. At present, worrying variant strainspose the greatest impact on the epidemic and threat to the world, including alpha, beta, gamma and delta. WHO issued a statementthat worrying variant strains spread faster and the risk of reinfection is higher[7].We can formulate public health policies and allocate health resources better by means of timely and accuratelyprediction of the epidemic trend of COVID-19.

Many methods can be used to predict the epidemic of infectious diseases, which could broadly divided into three categories: statistical modeling, mathematical pandemic modeling and deep learning methods [8]. The prediction methods for infectious diseases include Artificial Neural Networks (ANN), Radial Basis Function (RBF), Time Delay Neural Networks (TDNN),The Auto Regressive Integrated Moving Average (ARIMA), and Susceptible-Infectious-Recovered-Susceptible (SIRS) [9~10]. Developing mathematical models can provide key information for policy makers to take preventive measures. ARIMAis a moving average autoregressive techniquewhichcreated by box and Jenkins in the 1970s [11].ARIMA can describe the changes of time series by using mathematical methods. ARIMA was used to forecast COVID-19 duration, infections, and deaths[12]. Alabdulrazzaq H etal[13] verified the accuracy of COVID-19 propagation prediction based on ARIMA.In this

study, we fit ARIMA model for the time series of confirmedand cured cases respectively.

2. LITERATURE REVIEW

ARIMA is a combination of Autoregressive Model (AR) and Moving Average model (MA) [14]. The prediction and evaluation technology of time series has been relatively perfect, and its prediction results are relatively clear.In many cases, the prediction results obtained by time series method are more reliable than those obtained by traditional modeling. In the medical field, it has been proved that this method can getidealized results in predicting the incidence or death of some diseases [15]. The researchers can predictthe incidence rate of Hepatitis A(HAV) [16], Severe Acute Respiratory Syndrome (SARS) [17], Hemorrhagic Fever with Renal Syndrome (HFRS) [18] and Hantavirus pulmonary syndrome(HPS) [19] using ARIMA.

David etal[20] used ARIMAtoconfirmedrehabilitated and dead COVID-19 cases in Nigeria from February 27, 2020 to July 16, 2020. The results showed that ARIMA (2, 1, 4), ARIMA (2, 1, 2) and ARIMA (2, 1, 3) models were selected as the best candidates,which could predict the confirmed, rehabilitated and dead covid-19 cases in Nigeria.

A comparative study on SIR model, linear regression, logic function and ARIMA model to predict covid-19 cases found that ARIMA model performed better than all three in the prediction cases[21]. Comparatively speaking, SIR model cannot accurately predict the surge for the early stage of infectious disease, the linear regression cannot be used for a long

time when data are nonlinear, and logistic regression is more suitable for short-term prediction.

3. MATERIALS AND METHODS

3.1 Data source

Data source:The notification module which comes fromofficial website of the National Health Commission of novel coronavirus epidemic prevention and control. Collect the number of newly diagnosed and cured COVID-19 cases in China from April 12, 2020 to December 5, 2021.

3.2 Methods

Time series analysis was used to analyze theconfirmedand cured cases of COVID-19 in China from April 12, 2020 to December 5, 2021and ARIMA model was established. The established model was used to predict the number of newly confirmed cases and cured cases within 4 weeks after December 5, 2021, so as to test the prediction effect of the model.

ARIMA modelconsiders that there is a certain relationship between the state of things at a certain time in the future and the historical data of the past and the present. When the observation value is a stationary time series, the model can be written as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \cdots + \alpha_q \varepsilon_{t-q}$$

The basic form can be written as ARIMA (p, d, q), where p is the order of autoregression, d is the order of difference, and qis the order of moving average.

3.3 Modeling steps

Time series analysis and modeling mainly includes four steps: ①

Check the stationarity of time series: if it is an unstable series, make the series stable through difference and then analyze it; Identification and preliminary order determination of the model: by drawing and observing the Autocorrelation Coefficient (ACF) map and Partial Autocorrelation Coefficient (PACF) map, the model is preliminarily identified and the order is determined; Parameter estimation and test: use nonlinear least square method to estimate model parameters, use stable R^2, standardized Bayesian information criterion (BIC) and other information, and determine the best model according to the BIC minimum principle [10]. After the parameters are determined, the residual sequence of the original data and the fitting data is subjected to white noise detection with the Ljung box Q statistic to test whether the model can fully extract the trend information of the original sequence. If P > 0.05, it is considered to pass the white noise test; Validation and prediction of the model: use the selected model to predict the number of newly confirmed cases per week from April 14, 2020 to December 5, 2021, and compare it with the actual observation value to verify the fitting effect of the model; the model was used to predict the number of confirmed cases and cured cases in the four weeks after December 5, 2021, and compared with the actual observation value.

3.4Statistical analysis

The collected data were preprocessed with Excel, and the number of newly confirmed cases per week was calculated according to the cumulative number of confirmed cases per day to form a time series in weeks. Then IBM SPSS 25.0 statistical software was used to process and

model the data.

4. RESULTS AND DISCUSSION

4.1 Trend characteristics of time series

The time series diagram (Fig. 1) of weekly diagnosed cases from April 12, 2020 to December 5, 2021 shows that there was a significant increase in January 2021. The maximum number of newly confirmed cases in a week was 800. In addition, we observed that this time series was non-stationary and no seasonal trend.The time series diagram (Fig. 2) of weekly cured cases from April 12, 2020 to December 5, 2021 shows that its change trend was closely related to the time series of confirmed case, However, it would be about 2 weeks late.

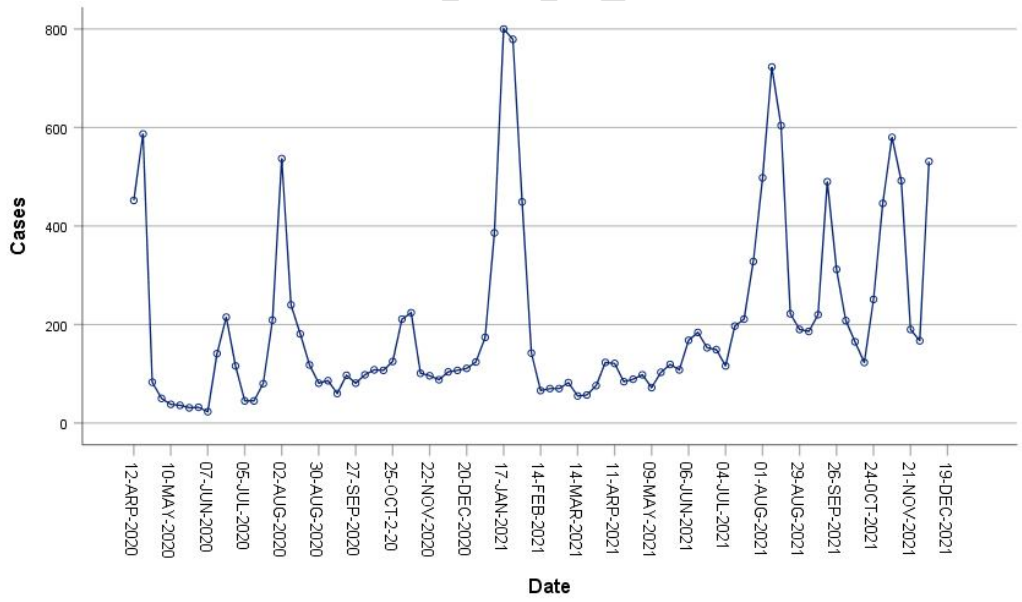

Fig. 1 Time series of confirmed cases of COVID-19 every week from April 12, 2020 to December 5, 2021

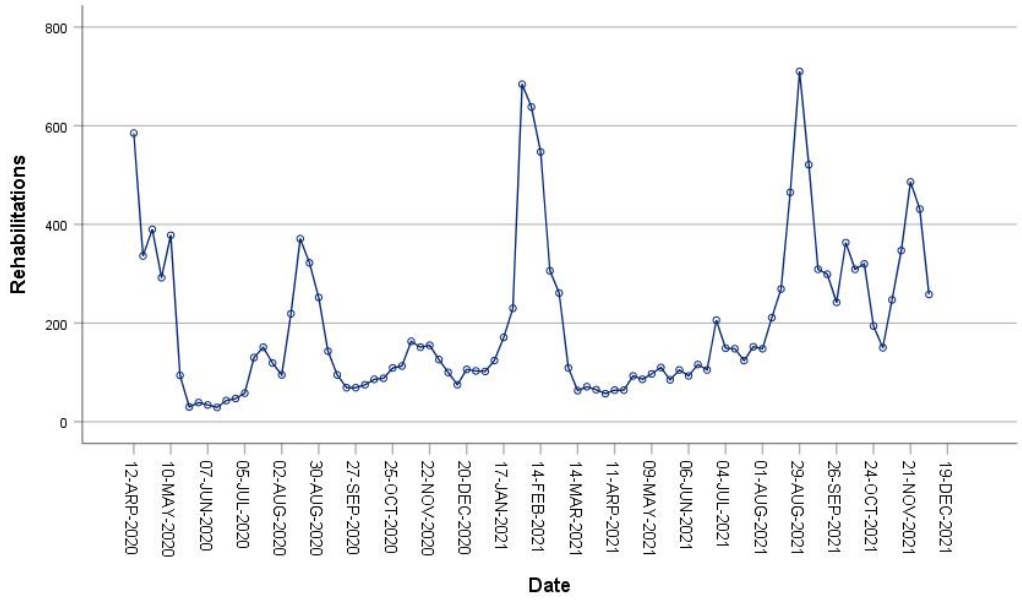

Fig. 2 Time series of COVID-19 cured cases every week from April 12, 2020 to December 5, 2021

4.2 Model identification and construction

4.2.1 The characteristic description of the stationarity trend of the data showed that the two time series were non-stationary series. In order to eliminate the influence of the trend, observe the difference of the time series of diagnosedand cured cases and observe the ACF and PACF charts, as shown in Fig. 3, Fig. 4, Fig. 5, Fig. 6. The model form was preliminarily determined as ARIMA (p, 1, q), and p and q were the auto-regressive and moving average orders respectively. According to the significant lag orders of ACF and PACF charts and previous experience, 9 models were preliminarily selected as ARIMA (2,1,2), ARIMA (2,1,1), ARIMA (2,1,0), ARIMA (1,1,2), ARIMA (1,1,1), ARIMA (1,1,0), ARIMA (0,1,2) , ARIMA (0,1,1), ARIMA (0,1,0)to screen and determine the optimal model.

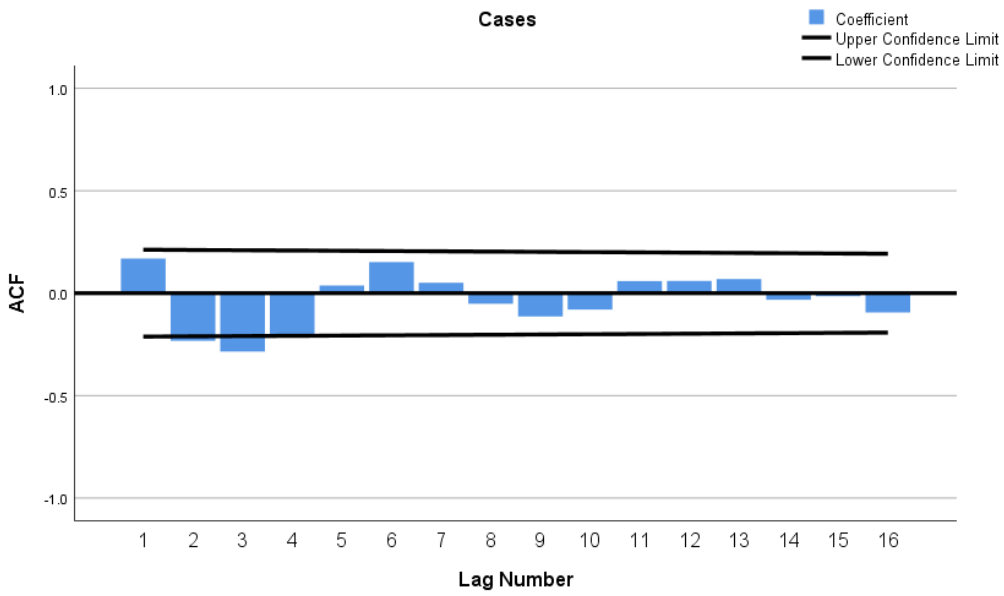

Fig. 3 ACF diagram of differential sequence of confirmed cases every week

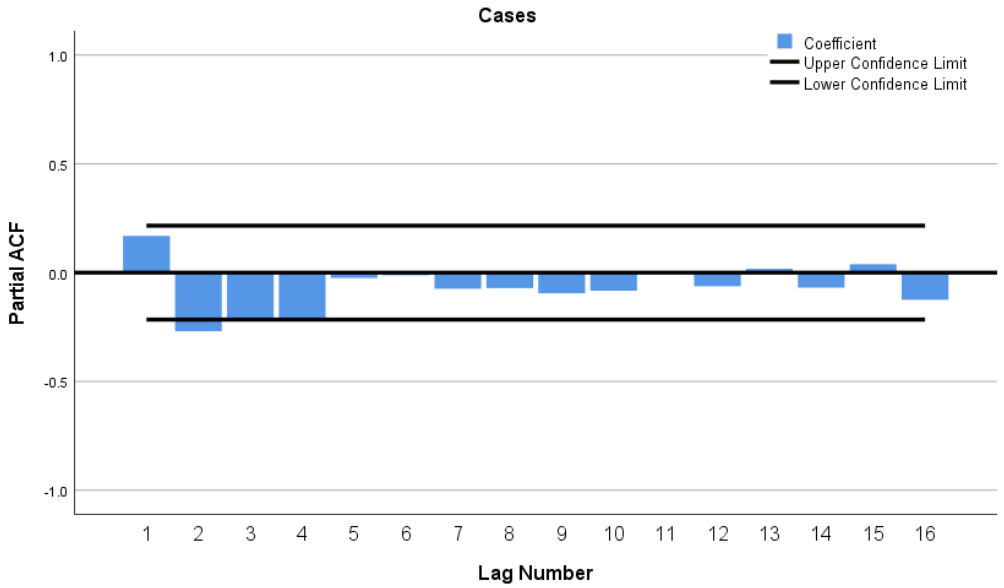

Fig. 4 PACF diagram of differential sequence of confirmed cases every week

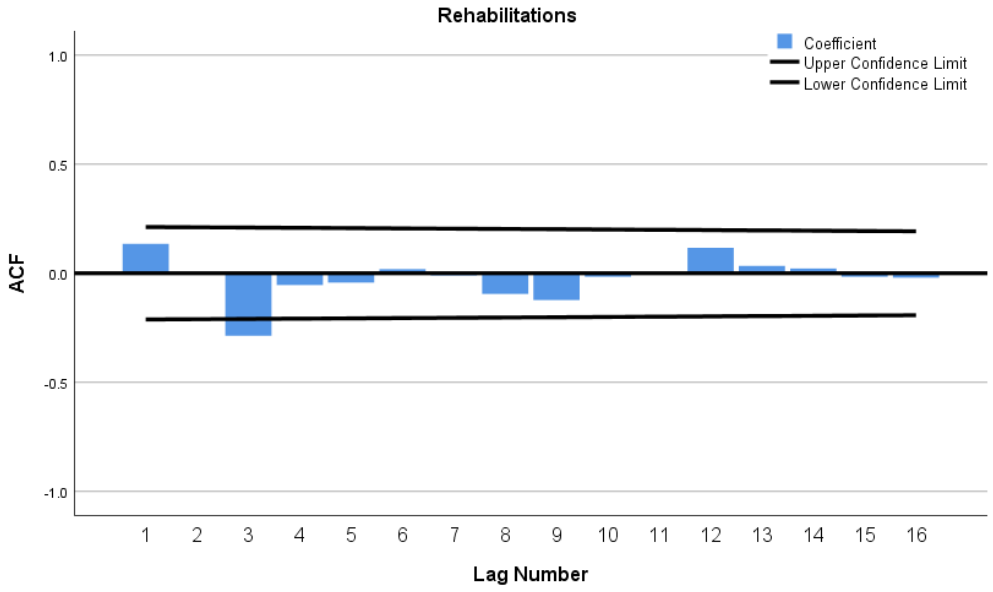

Fig. 5 ACF diagram of differential sequence of cured cases every week

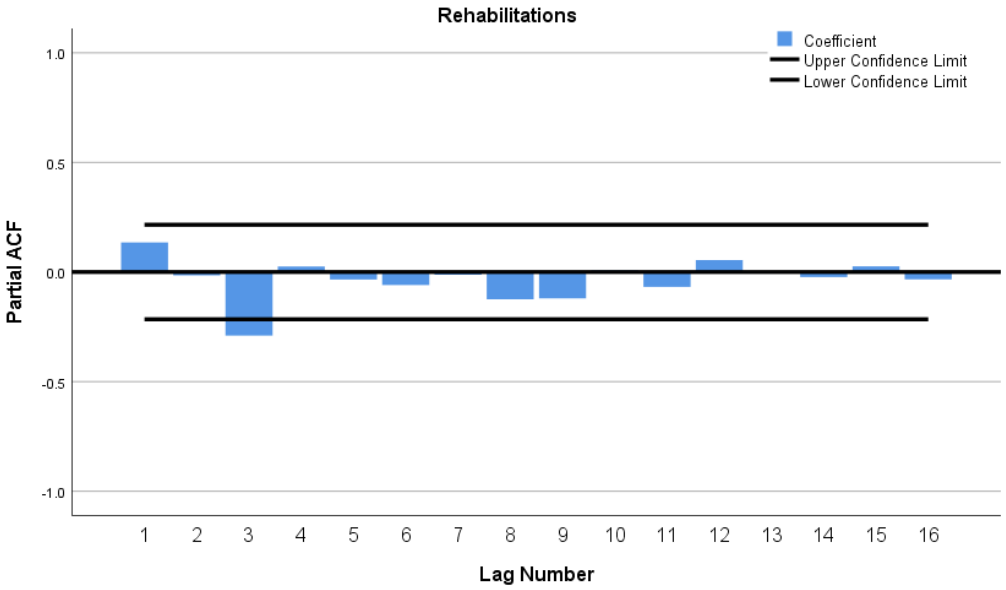

Fig. 6 PACF diagram of differential sequence of cured cases every week

4.2.2 Parameter estimation and model diagnosis

The parameters of the model were tried one by one from low to high, and the models with different order combinations were obtained. After screening, the model with the smallest normalized BIC was selected as the

best model. The parameter values of ARIMA (2, 1, 1) with relatively high R^2were significant, R^2 = 0.542/0.617, and the normalized BIC value = 9.917 /9.415 werethe smallest among the alternative models (Table 1 and Table 2). Ljung box q = 9.095/9.651, P> 0.05 passed the white noise test. The ACF and PACF of the residual sequence were shown in Figures 7 and 8. The auto-correlation coefficient and partial autocorrelation coefficient of the residual sequence fell into 95% CI, which proved that the residual sequence was a random error and indicated that the model had extracted the information contained in the time series.

Table 1 Parameter estimation results and fitting statistics of ARIMA models of confirmed cases

Model	Normalized BIC	R^2	*Ljung-Box Q*	*P*
ARIMA (2,1,2)	9.984	0.540	10.377	0.734
ARIMA (2,1,1)	9.917	0.542	9.095	0.872
ARIMA (2,1,0)	10.012	0.463	14.840	0.536
ARIMA (1,1,2)	9.936	0.533	12.213	0.663
ARIMA (1,1,1)	10.065	0.433	21.546	0.158
ARIMA (1,1,0)	10.035	0.414	21.432	0.211
ARIMA (0,1,2)	9.977	0.481	20.034	0.210
ARIMA (0,1,1)	10.010	0.428	17.334	0.432
ARIMA (0,1,0)	10.001	0.396	27.448	0.071

Table 2 Parameter estimation results and fitting statistics of ARIMA models of cured cases

Model	Normalized BIC	R^2	*Ljung-Box Q*	*P*
ARIMA (2,1,2)	9.554	0.587	11.769	0.608
ARIMA (2,1,1)	9.415	0.617	9.651	0.841
ARIMA (2,1,0)	9.442	0.580	12.454	0.712
ARIMA (1,1,2)	9.429	0.612	10.025	0.818
ARIMA (1,1,1)	9.443	0.580	12.391	0.717
ARIMA (1,1,0)	9.379	0.580	12.361	0.778
ARIMA (0,1,2)	9.436	0.583	10.607	0.833
ARIMA (0,1,1)	9.380	0.580	12.507	0.769
ARIMA (0,1,0)	9.333	0.357	14.498	0.696

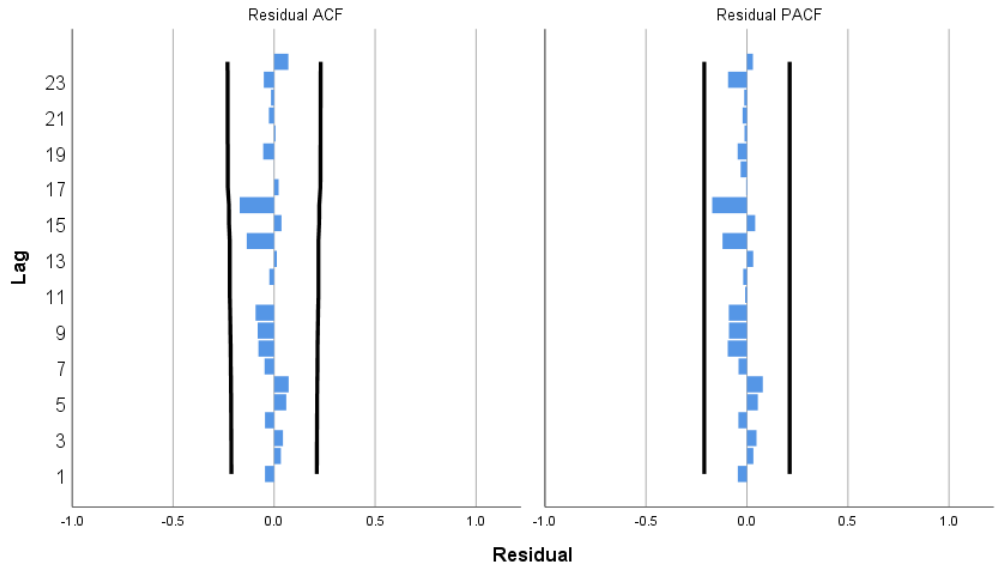

Fig. 7 ACF and PACF diagram of residual sequence of confirmed cases

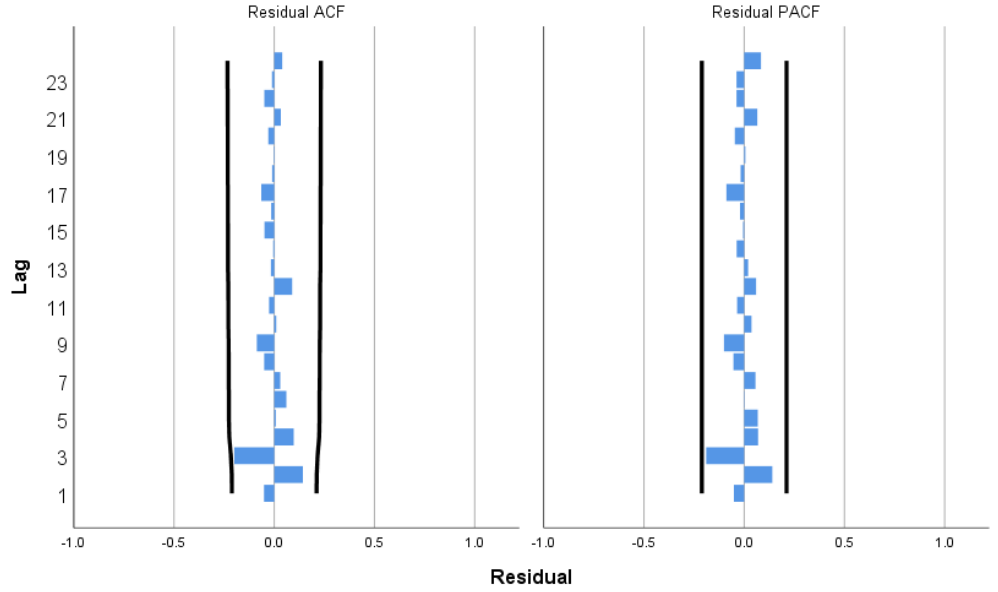

Fig. 8 ACF and PACF diagram of residual sequence of cured cases

4.2.3 Prediction analysis

ARIMA(2,1,1) was used to fit the number of confirmedand cured cases of COVID-19 every week from April 12, 2020 to November 5, 2021. The results were shown in Fig. 9 and Fig. 10whichcould be seen that the model fitted well.The overall dynamic trend predicted by the model was

basically consistent with the actual situation, and the model made a good prediction of the trend of the number of cases in the near future.

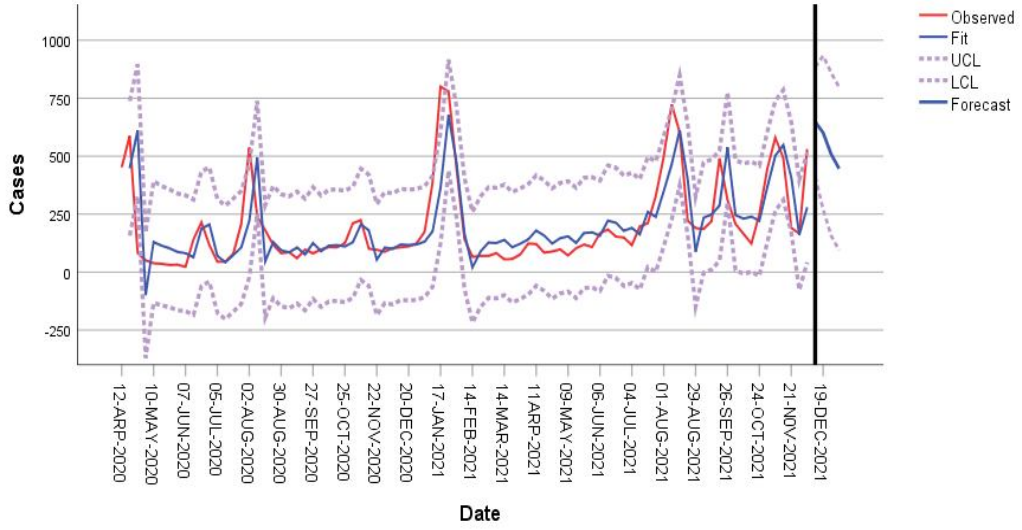

Fig. 9 Forecast of the number of confirmed cases of COVID-19 per week

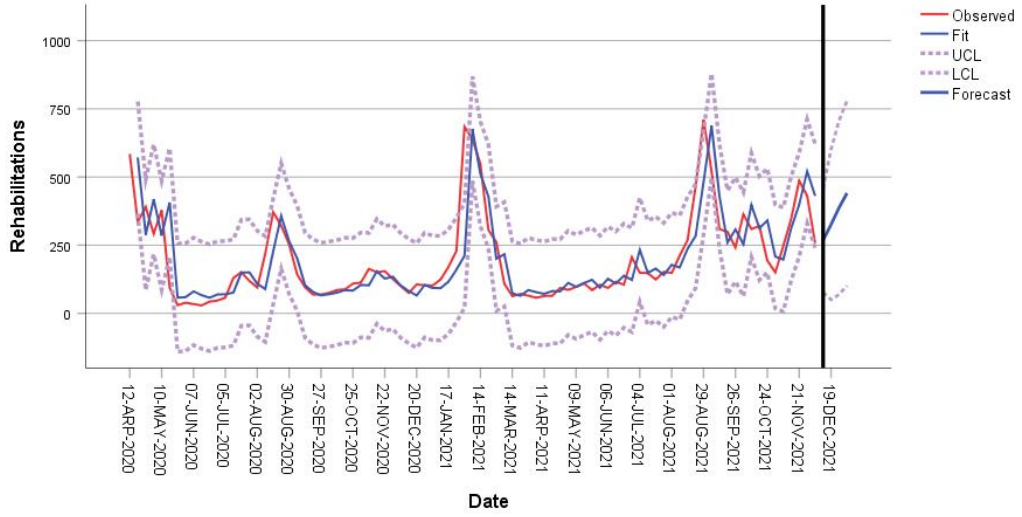

Fig. 10 Forecast of the number of cured cases of COVID-19 per week

ARIMA(2,1,1) was used to predict confirmed cases of COVID-19 four weeks after December 5, 2021 (Table 3).The number of confirmed

cases observed in the first and second weeks after December 5, 2021 were within the 95% confidence interval of the predicted value. However, the number of confirmed cases observed in the third and fourth weeks were not within the 95% confidence interval of the predicted value.

We used ARIMA (2,1,1) to predict the cured cases of COVID-19 four weeks after December 5, 2021 (Table 4). It was found that the cured cases observed 4 weeks after December 5, 2021 were within the 95% confidence interval of the predicted value.

Table 3 Comparison of predicted and observed values of confirmed cases of COVID-19 in the four weeks after December 5, 2021

Model		Week 1	Week 2	Week 3	Week 4
ARIMA (2,1,1)	Observed value	577	606	891	1389
	Estimate	648	601	509	446
	UCL	885	932	860	798
	LCL	411	270	157	95

Table 4 Comparison of predicted and observed values of cured cases of COVID-19 in the four weeks after December 5, 2021

Model		Week 1	Week 2	Week 3	Week 4
ARIMA (2,1,1)	Observed value	256	251	469	420
	Estimate	268	326	388	441
	UCL	458	602	707	780
	LCL	77	50	69	102

5. CONCLUSION

The premise of doing a good job in prevention and control is to fully grasp the development trend and laws of the COVID-19 epidemic. This study analyzed the time series of confirmed cases and cured cases of COVID-19 from April 12, 2020 to December 5, 2021. The results showed that COVID-19 had no obvious seasonal change trend, which was

different from many other respiratory infectious diseases such as influenza.This study attemptedto search for optimal ARIMA model that could fit the times series and predict confirmed and curedcases of COVID-19 in China.

The study utilized data on the confirmed, cured due to COVID-19 in China from April 12, 2020 to December 5,2021. The analysis results showed that ARIMA (2, 1, 1) could better fit the time series of confirmed cases and cured cases in this period. This model could better extract the information in the time series, and the fitting value was basically consistent with the measured value. Therefore, ARIMA model can be used to predict the COVID-19 epidemic situation. The comparative analysis of measured and predicted values of confirmed cases showed that the measured values in the first and second weeks after December 5, 2021 fell within the prediction range, and the measured values in the third and fourth weeks exceeded the prediction range. However, the observed values of cured cases in the four weeks after December 5, 2021 were all within the 95% CI range of the predicted value of ARIMA (2, 1, 1) model.

The advantage of ARIMA model for time series analysis was that it was convenient to obtain data, and it can extract information and model through the self change law of time series, without considering other relevant factors analysis, and modeling can be realized in the IBM SPSS and other softwares. The modeling points of ARIMA model could follow: (1) analyze the stationarity of time series, and make difference for non-stationary series to make it stable; (2) According to the autocorrelation

coefficient (ACF diagram) and partial autocorrelation coefficient (PACF diagram) of the stationary time series, the lag order of the model is preliminarily determined and the preselected model is determined; (3) Determine the best model according to the BIC value and R2 of the preselected model and whether it passes the white noise test; (4) The best model is used for prediction and analysis, and compared with the measured value. The disadvantage of ARIMA model is that it only can make short-term prediction. The established model cannot be used as a permanent prediction tool. New actual values should be added continuously to correct or refit the better model.

In short, the time series model can provide references for the prediction of the epidemic situation of COVID-19 and other unknown infectious diseases in the future, and for the formulation of relevant prevention and control measures. However, as a data processing method, it can not truly reflect the development trend of the disease. In reality, the impact of other factors on the prediction results must be considered.

LIMITATIONS

As a new infectious disease, COVID-19 has a short observation period. The inadequacy of this study is that it only includes the time series of the number of new cases per week from April 2020 to December 2021. It is not possible to obtain more information about the legal characteristics of COVID-19 from this shorter time series. The follow-up study will further improve the time series through continuous data collection, and it is planned to use a variety of methods for modeling, and select models with

small fitting and prediction errors to analyze the series, so as to more deeply reflect the internal laws and future trends of the time series of COVID-19.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

Ethical Considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors. There are no ethical issues in this manuscript.

Acknowledgement

This study was financially supported by 2020 Science and Technology Fund Project of Guizhou Provincial Health Commission.

REFERENCES

[1]Chyon FA, Suman MNH, Fahim MRI, Ahmmed MS. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. Journal of Virological Methods. 2022; 301.

[2] Yang Q, Wang J, Ma H, Wang X. Research on COVID-19 based on ARIMA model-Taking Hubei, China as an example to see the epidemic in Italy. Journal of Infection and Public Health. 2020;13(10):1415-1418.

[3] Chen N., Zhou M., Dong X. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in

Wuhan, China: a descriptive study. Lancet. 2020.

[4] Sun J. Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models. Computer Methods and Programs in Bio-medicine Update. 2021.

[5] Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery. 2020;76:71-76.

[6] World Health Organization. 2020. Novel Coronavirus(2019-nCoV) Situation Report-12.

[7] "Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern."https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern (accessed Jul. 12, 2022).

[8] Liao Z, Song Y, Ren S, Song X, Fan X, Liao Z. VOC-DL: Deep learning prediction model for COVID-19 based on VOC virus variants. Computer Methods and Programs in Biomedicine. 2022; 224.

[9] Ture M, Kurt I. Comparison of four different time series methods to forecast hepatitis A virus infection. Expert Systems Applications. 2006;31:41–6.

[10] Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences of the United States of America.2012;109(50):20425–20430.

[11] Ilie OD, Ciobica A, Doroftei B. Testing the Accuracy of the ARIMA Models in Forecasting the Spreading of COVID-19 and the Associated Mortality Rate. Medicina (Kaunas). 2020;56(11):566.

[12] Yue X.-G., Shao X.-F., Li R.Y.M., Crabbe M.J.C., Mi L., Hu S. Risk prediction and assessment: Duration, infections, and death toll of the COVID-19 and its impact on China's economy. Journal of Risk and Financial Management. 2020;13(4):66.

[13] Alabdulrazzaq H, Alenezi MN, Rawajfih Y, Alghannam BA, Al-Hassan AA, Al-Anzi FS. On the accuracy of ARIMA based prediction of COVID-19 spread. Results in Physics. 2021; 27:104509.

[14] Awan TM, Aslam F. Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. Journal of Public Health Research. 2020;9(3):1765.

[15] Wei W, Jiang J, Liang H, Gao L, Liang B, Huang J, Zang N, Liao Y, Yu J, Lai J, Qin F, Su J, Ye L, Chen H. Application of a Combined Model with Autoregressive Integrated Moving Average (ARIMA) and Generalized Regression Neural Network (GRNN) in Forecasting Hepatitis Incidence in Heng County, China. PLoS One. 2016;11(6): e0156768.

[16] Guan P, Huang DS, Zhou BS. Forecasting model for the incidence of hepatitis A based on artificial neural network. World Journal of Gastroenterology. 2004;10(24):3579–3582.

[17] Tan CV, Singh S, Lai CH, et al. Forecasting COVID-19 Case Trends Using SARIMA Models during the Third Wave of COVID-19 in Malaysia. Int J Environ Res Public Health. 2022;19(3):1504.

[18] Liu Q, Liu X, Jiang B, Yang W.Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model.BMC Infectious Diseases. 2011;11:218.

[19] Nsoesie EO, Beckman RJ, Shashaani S, Nagaraj KS, Marathe MV. A Simulation Optimization Approach to Epidemic Forecasting. PLoS ONE. 2013;8: e67164.

[20] KuheDA, Atsua IkughurJ. A Time Series Model on the Occurrence of COVID-19 Pandemic in Nigeria. Asian Journal of Research in Infectious Diseases, 2021;8(4), 66-80.

[21] Abolmaali S, Shirzaei S. A comparative study of SIR Model, Linear Regression, Logistic Function and ARIMA Model for forecasting COVID-19 cases. AIMS Public Health. 2021;8(4):598-613.