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 were collected through the official website of the National Health Commission, and the number of newly confirmed and cured cases per week was sorted out. IBM SPSS 25.0 software was used to analyze the time series of newly diagnosed and cured COVID-19 cases every week from April 12, 2020 to December 5, 2021. The model was established through model identification, parameter estimation and model fitting. **Results** The number of reported cases of COVID-19 has no obvious seasonal characteristics. ARIMA (2,1,1) model well fitted the sequence of confirmed / cured COVID-19 cases every week, 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, $P > 0.05$. The number of cases and cures in the four weeks after December 5, 2021

were predicted. The measured values in the first week and the second week were 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

COVID-19 is an acute respiratory infectious disease that first appeared in Wuhan, China. The symptoms of COVID-19 is similar to the flu (Influenza), which can be fever, cough, shortness of breath, pneumonia and breathing difficulties [1].Due to its infectiousness and general susceptibility to the crowd, its transmission speed is relatively fast [2]. Most SARS-CoV-2 infected patients present with mild respiratory symptoms. However, some have developed various fatal complications including organ failure, septic shock, pulmonary oedema, severe pneumonia, and Acute Respiratory Distress Syndrome (ARDS) [3]. Much people infected with novel coronavirus have to be hospitalized and use respirators. Since the original outbreak, COVID-19 has spread to 221 countries and territories [4]. On 30th January 2020, the World Health Organization declared COVID-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. Since the COVID-19 epidemic swept the world in early 2020, novel coronavirus has been spreading among people. For the virus will replicate the genetic genome in the process of transmission, the virus genome changes due to its unavoidable replication errors, resulting in mutant strains, which brings new challenges to epidemic prevention. The World Health Organization has classified COVID-19 variant strains into two categories according to the degree of risk: worrying variant strains (VOC, variant of concern) and noteworthy variant strains (VOI, variant of interest). At present, VOC is the variant strain with the greatest impact on the epidemic and the greatest threat to the world, including alpha, beta, gamma and delta. The World Health Organization issued a statement saying that these variants 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 accurate prediction of the epidemic trend of COVID-19.

Many studies have predicted pandemic trends in different countries and regions, which can be broadly divided into three categories: statistical

modeling, mathematical pandemic modeling and deep learning methods [8]. The prediction methods for infectious diseases include the ANN algorithm, radial basis function (RBF), time delay neural networks (TDNN), the ARIMA model, and susceptible–infectious–recovered–susceptible (SIRS) [9~10]. Developing mathematical models can provide key information for policy makers to take preventive measures. Autoregressive integrated moving average (ARIMA) is a moving average autoregressive technique, which was created by box and Jenkins in the 1970s [11]. The ARIMA can describe the changes of time series by using mathematical methods. The researchers used ARIMA to predict prevalence, death, and recovery rates for 25 different countries. In [12], ARIMA was used to forecast COVID-19 duration, infections, and deaths. The researcher in [13] verified the accuracy of COVID-19 propagation prediction based on ARIMA. In this study, we use the autoregressive comprehensive moving average (ARIMA) model to predict the number of new COVID-19 cases and rehabilitation in China in the next four weeks.

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. Time series model refers to ARIMA model and some

expression form. 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 achieve idealized results in predicting the incidence or death of some diseases [15]. The researchers predicted the incidence rate of hepatitis A virus (HAV) [16], severe acute respiratory syndrome (SARS) [17], hemorrhagic fever with renal syndrome (HFRS) [18] and hantavirus lung syndrome (HPS) [19] using the regression comprehensive moving average (ARIMA) model.

David et al [20] used the autoregressive comprehensive moving average (ARIMA) time series model to confirmed rehabilitated 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 for modeling and predicting 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

2.1 Data source

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

2.2 Methods

Time series analysis was used to analyze the number of newly confirmed cases and cured cases of COVID-19 in China from April 12, 2020 to December 5, 2021, and 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, to test the prediction effect of the model.

The autoregressive moving average model (ARIMA model) considers 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. The historical data of the time series reveals the law of the change of the target variable with time, and predicts the state at a certain time in the future according to the law. When the observation value is a stationary time series,

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

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 Q is the order of moving average.

2.3 Modeling steps

Time series analysis and modeling mainly includes four steps: check the stationarity of time series: if it is an unstable series, make the series stable through difference and then analyze; 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; ③ 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; ④ 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.

2.4 Statistical 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

By drawing the time series diagram of the number of newly confirmed cases and cured cases per week from April 12, 2020 to December 5, 2021 (Fig. 1and Fig. 2), the number of confirmed cases in August 2020 and January and August 2021 increased significantly, with the highest number of newly confirmed cases in a week of 800, with different weekly increases and no obvious seasonal trend. The number of cured patients changes with the development trend of the number of confirmed patients.

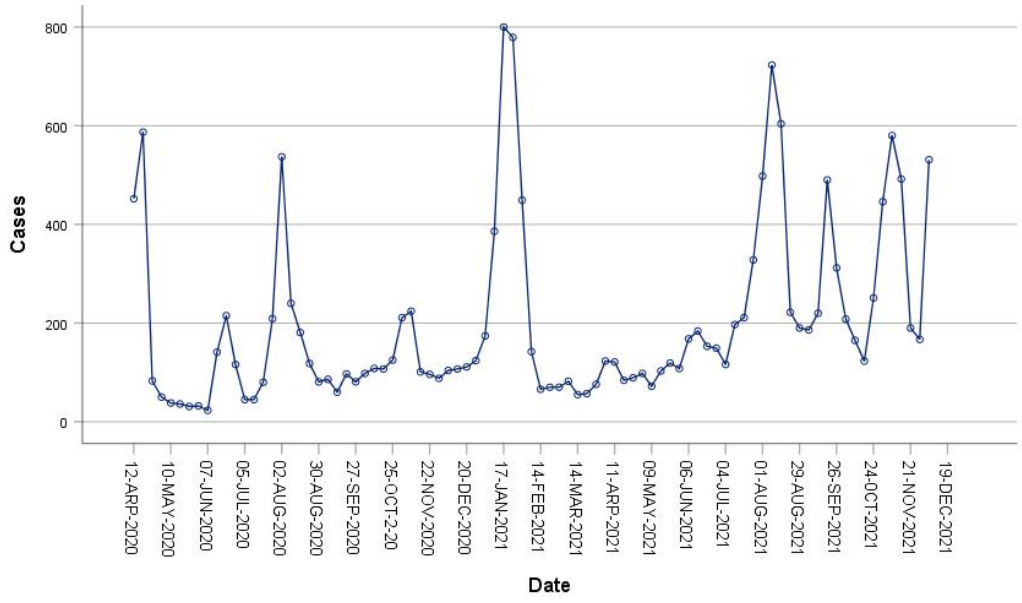

Fig. 1 Time series of confirmed cases of COVID-19 every week from April 12, 2020 to December 5, 2021

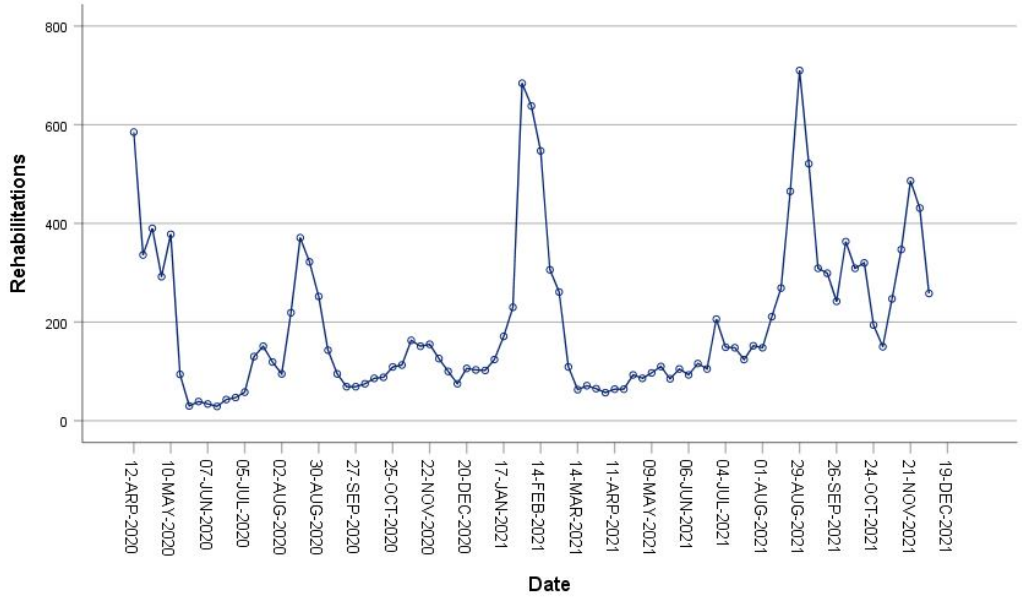

Fig. 2 Time series of COVID-19 cured cases every week from April 12, 2020 to December 5, 2021

3.2 Model identification and construction

3.2.1 The characteristic description of the stationarity trend of the data shows that the two time series are non-stationary series. In order to eliminate the influence of the trend, observe the difference of the time series of diagnosed cases and cured cases and observe the ACF and PACF

charts, as shown in Fig. 3, Fig. 4, Fig. 5, Fig. 6. The model form is preliminarily determined as ARIMA (p, 1, q), and p and q are the autoregressive 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.

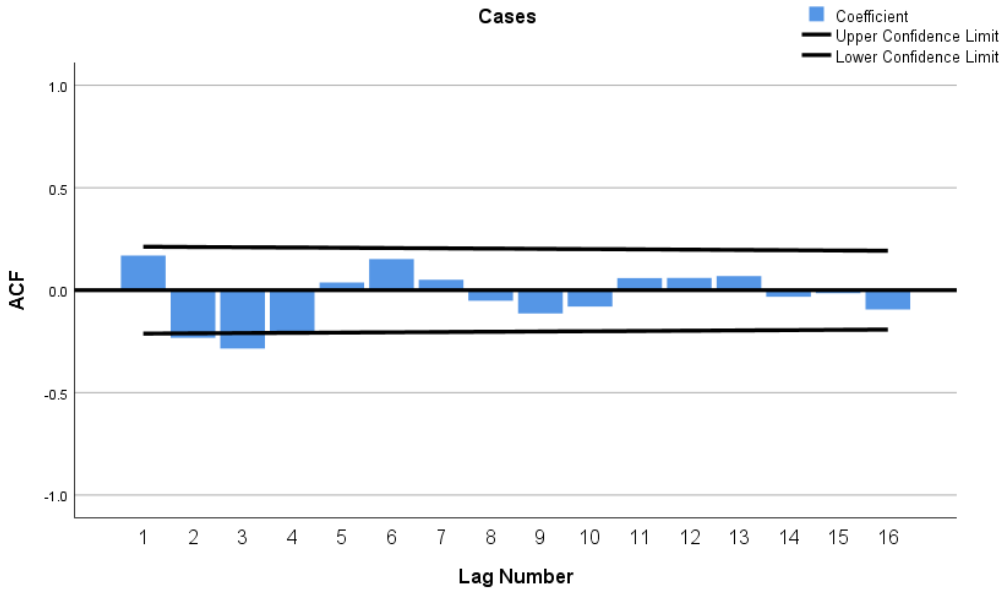

Fig. 3 ACF diagram of differential sequence of confirmed cases every week

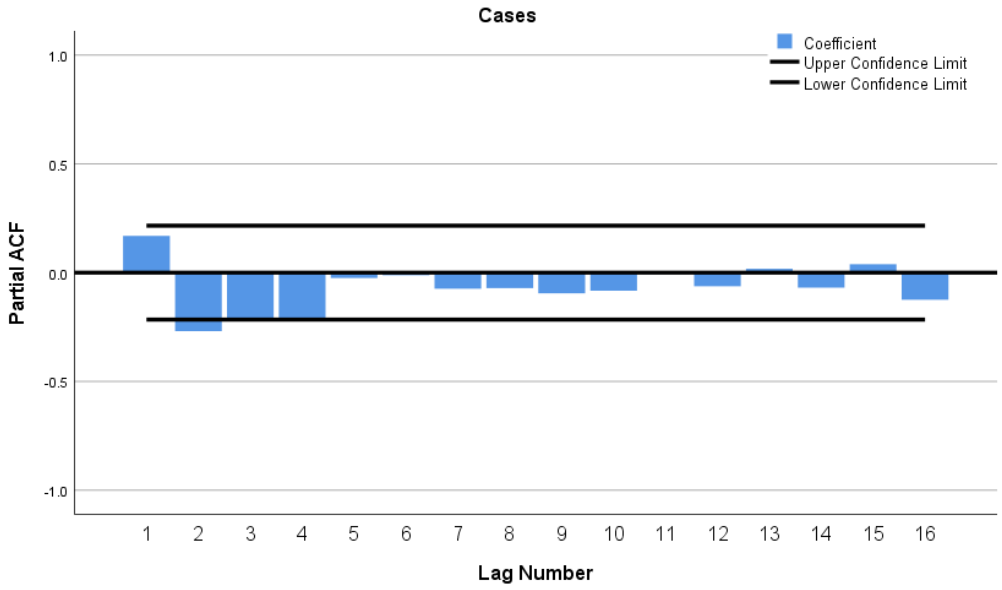

Fig. 4 PACF diagram of differential sequence of confirmed cases every week

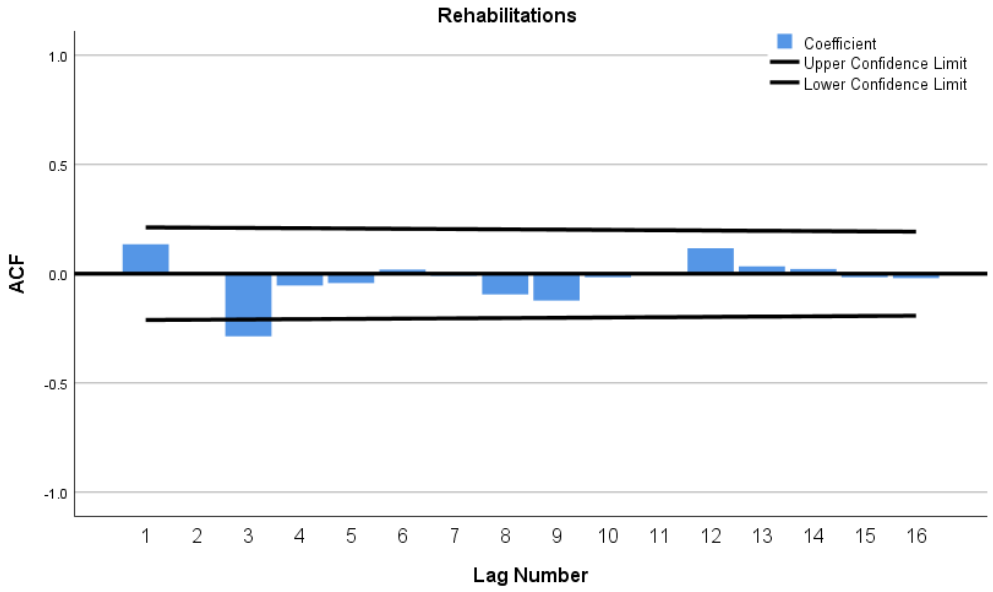

Fig. 5 ACF diagram of differential sequence of cured cases every week

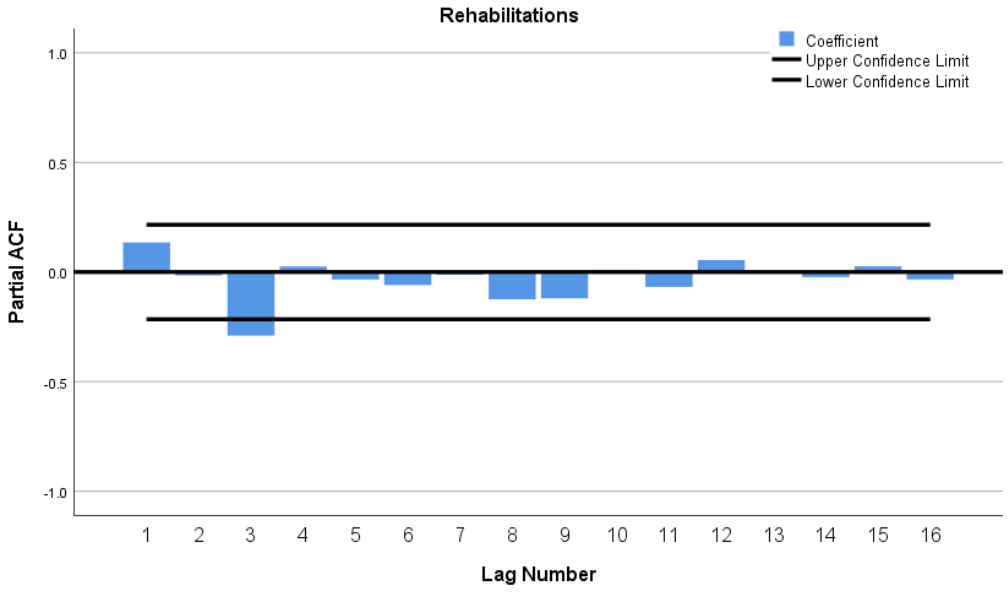

Fig. 6 PACF diagram of differential sequence of cured cases every week

4.2.2 Parameter estimation and model diagnosis

The parameters of the model are tried one by one from low to high, and the models with different order combinations are 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^2 are significant, R^2 = 0.542/0.617, and the normalized BIC value = 9.917 /9.415 are the smallest among the alternative models (Table 1 and Table 2). Ljung box q = 9.095/9.651, $P > 0.05$ pass the white noise test. Make ACF and PACF diagrams of the residual sequence (Fig.7 and Fig.8). The autocorrelation coefficient and partial autocorrelation coefficient of the residual sequence fall into 95% CI, which proves that the residual sequence is a random error and indicates that the model has 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

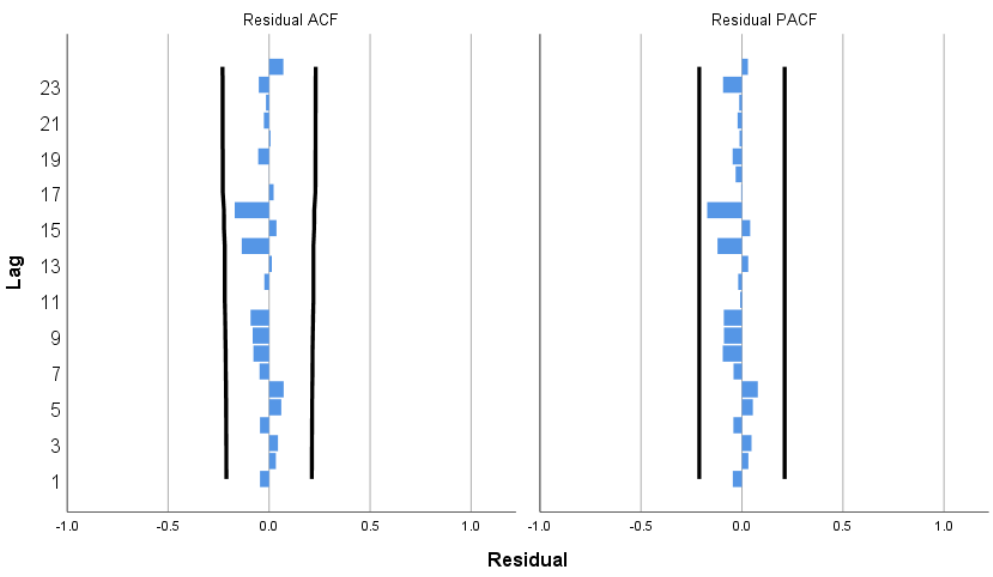

Fig. 7 ACF and PACF diagram of residual sequence of confirmed cases

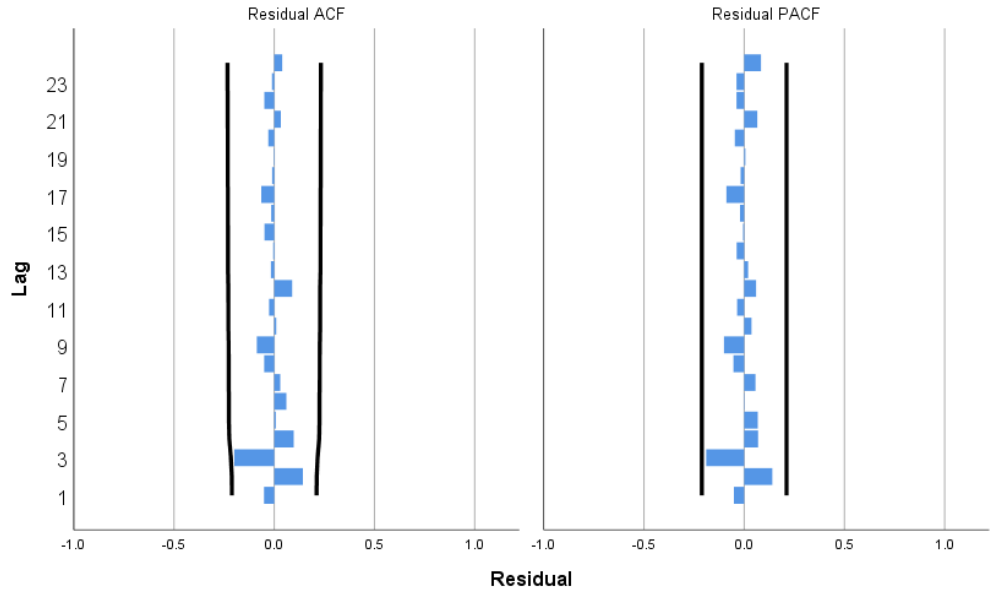

Fig. 8 ACF and PACF diagram of residual sequence of cured cases

4.2.3 Prediction analysis

The ARIMA model is used to fit the number of confirmed cases and cured cases of COVID-19 every week from April 12, 2020 to November 5, 2021. The results are shown in Fig. 9 and Fig. 10. It can be seen that the model fits well.

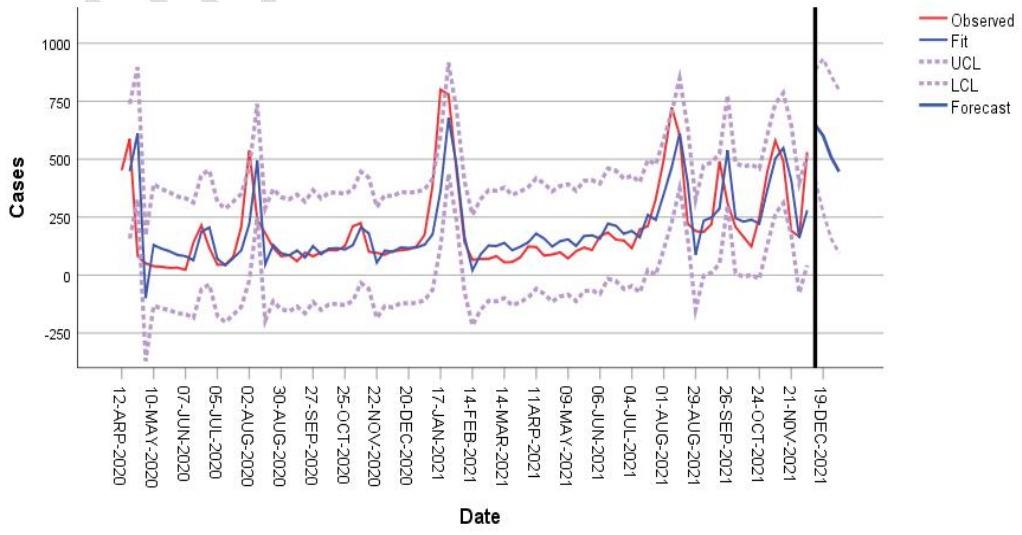

Fig. 9 Forecast of the number of confirmed cases of COVID-19 per week

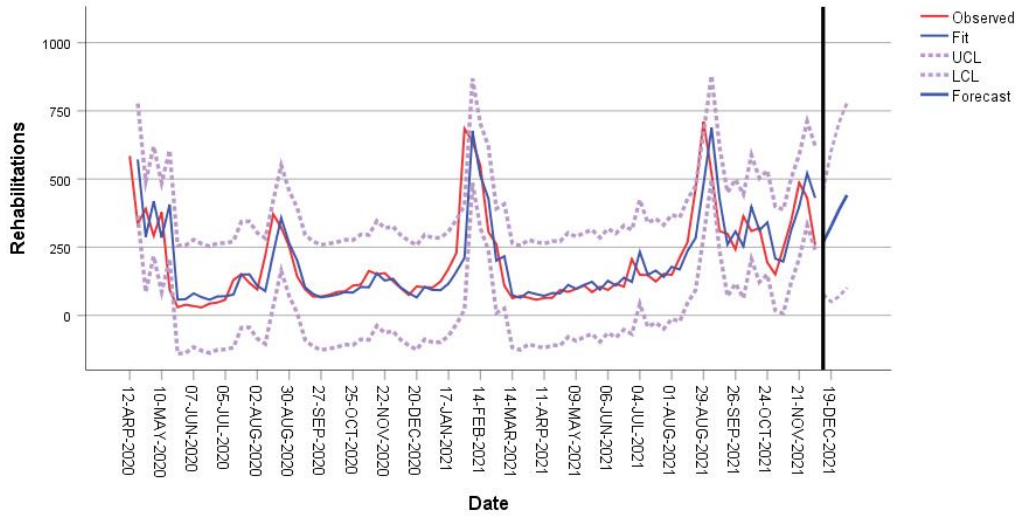

Fig. 10 Forecast of the number of cured cases of COVID-19 per week

The selected model is used to predict the number of confirmed cases and cured cases of COVID-19 per week in the four weeks after December 5, 2021(Table 3 and Table4). The observed values of confirmed in the first and second weeks after December 5, 2021 do not exceed the 95% confidence interval. The observed values of the confirmed cases in the third and fourth weeks after December 5, 2021 are not within the 95% CI range of the predicted value. However, the number of cures in the four weeks after December 5, 2021 are within the 95% CI range 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

After the outbreak of infectious diseases, how to predict the epidemic trend timely and effectively is one of the focuses of infectious disease prevention and control. As a new infectious disease, there are still many difficulties in early prediction and early warning of novel coronavirus pneumonia, mainly due to the impact of various factors after the outbreak of the disease, and the great fluctuations of epidemic data.

The occurrence and development of infectious diseases are affected by many factors, and show seasonal, periodic and trend characteristics in the transmission process. Novel coronavirus pneumonia is a new infectious disease. With the extension of the epidemic cycle, the virus mutants are constantly updated iteratively, and show the characteristics of high virus load, fast transmission speed and short incubation period. The prevention and control of the epidemic is still the main way to deal with the spread of the epidemic. 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 observed the number of confirmed cases and cured cases of COVID-19 per week from April 12, 2020 to December 5, 2021. No obvious seasonal trend was found, but there was a significant increase in August 2020 and August 2021. The time series needs to be further extended to verify the characteristics of COVID-19.

This study attempts to search for optimal ARIMA models that will fit and predict weekly cases of confirmed and cured of COVID-19 in China. The study utilized data on the confirmed, cured due to COVID-19 in China from 12/04/2020-05/12/2021. The data from 12/04/2020-05/12/2021 were used for model building while 4 week observations after from 05/12/2021 were used for training and forecast evaluations.

In this paper, through model identification and parameter estimation, the optimal model is ARIMA (2,1,1). This model can better extract the information in the time series, and the fitting value is basically consistent with the measured value. Therefore, ARIMA model can be used to predict the COVID-19 epidemic situation. Based on the analysis of measured values and predicted values, only the measured values in the first and second weeks after December 5, 2021 fall within the predicted range, and the measured values in the third and fourth weeks are beyond the predicted range. 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 is that it is 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 Eviews and other software. The modeling points of ARIMA model are as follows: (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 makes short-term prediction. The established model cannot be used as a permanent prediction tool. New actual values should be added continuously to correct or re fit 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.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

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. J Virol Methods. 2022 Mar; 301:114433. doi: 10.1016/j.jviromet.2021.114433.

[2] Yang Q, Wang J, Ma H, Wang X. Research on COVID-19 based on ARIMA modelΔ-Taking Hubei, China as an example to see the

epidemic in Italy. J Infect Public Health. 2020 Oct;13(10):1415-1418. doi: 10.1016/j.jiph.2020.06.019.

[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 doi: 10.1016/S0140-6736(20)30211-7.

[4] Sun J. Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models. Comput Methods Programs Biomed Update. 2021; 1:100029. doi: 10.1016/j.cmpbup.2021.100029.

[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). Int J Surg. 2020 Apr; 76:71-76. doi: 10.1016/j.ijsu.2020.02.034.

[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. Comput Methods Programs Biomed. 2022 Sep; 224:106981. doi: 10.1016/j.cmpb.2022.106981.

[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. doi: 10.1016/j.eswa.2005.09.002

[10] Shaman J, Karspeck A. Forecasting seasonal outbreaks of inflfluenza. Proc Natl Acad Sci USA. (2012) 109:20425–30. doi: 10.1073/pnas.1208772109

[11] 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 Phys. 2021 Aug; 27:104509. doi: 10.1016/j.rinp.2021.104509.

[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. J Risk Financial Manag. 2020;13(4):66. doi: 10.3390/jrfm13040066.

[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 Phys. 2021 Aug; 27:104509. doi: 10.1016/j.rinp.2021.104509.

[14] Awan TM, Aslam F. Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. J Public Health Res. 2020 Jul 8;9(3):1765. doi: 10.4081/jphr.2020.1765.

[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 Jun 3;11(6): e0156768. doi: 10.1371/journal.pone.0156768.

[16] Guan P., Huang D.S., Zhou B. Sen Forecasting model for the incidence of hepatitis A based on artificial neural network. World J. Gastroenterol. 2004;10:3579–3582. doi: 10.3748/wjg.v10.i24.3579.

[17] Nsoesie E.O., Beckman R.J., Shashaani S., Nagaraj K.S., Marathe M.V. A Simulation Optimization Approach to Epidemic Forecasting. PLoSONE. 2013;8:e67164.doi: 10.1371/journal.pone.0 067164.

[18] Liu Q., Liu X., Jiang B., Yang W. Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. BMC Infect. Dis. 2011; 11:218. doi: 10.1186/1471-2334-11-218.

[19] Nsoesie E.O., Beckman R.J., Shashaani S., Nagaraj K.S., Marathe M.V. A Simulation Optimization Approach to Epidemic Forecasting.

PLoS ONE. 2013;8: e67164. doi: 10.1371/journal.pone.0067164.

[20] Kuhe, D. A., & Atsua Ikughur, J. (2021). A Time Series Model on the Occurrence of COVID-19 Pandemic in Nigeria. Asian Journal of Research in Infectious Diseases, 8(4), 66-80. https://doi.org/10.9734/ ajrid/2021/ v8i430251.

[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 Aug 26;8(4):598-613. doi: 10.3934/publichealth.2021048.