ISSN: 2688-8653

Forecasting the Trend of New Crown Epidemic based on ARIMA Model

Peiqi Lu, Miaomiao Chen, Darui Chen, Bai Yi, Chengxin Zhang

Qingdao Huanghai University Qingdao, Shandong 266427, China

Abstract

In this paper, the cumulative number of new coronary pneumonia diagnoses in China from January 1, 2022 to April 29, 2022 was selected as the sample data. Firstly, the data were pre-processed; secondly, the model was tested; then, the model was optimized and a time series model was established to predict the trend of the number of confirmed cases, which provides a scientific basis for the government to formulate relevant epidemic prevention policies. The results showed that the accuracy of the predicted number of confirmed cases in the next five days by fitting ARIMA(1,2,0) was excellent, and the prediction accuracy almost approximated the actual cumulative number of confirmed cases. The predicted cumulative number of confirmed cases for four days differed from the actual cumulative number of confirmed cases by no more than 10, and the predicted value for one day was exactly the same as the actual cumulative number of confirmed cases without deviation.

Keywords

ARIMA Model; Predictive Model; Second Order Difference.

1. Introduction

The 2019 outbreak of Newcastle Pneumonia has made the peaceful and calm Chinese New Year pervasive with an air of danger. The growing number of confirmed cases and the climbing number of deaths have touched people across the country. In the face of the successive mutations of the New Coronavirus, the number of infected patients is increasing day by day. In order to deal with the huge impact of NCCV, the government needs to make timely responses and decisions. The most important and critical part of this process is to build an effective crisis management and emergency response mechanism. In order for government departments to generate, prepare and deploy early warnings for epidemic prevention and control. To this end, this paper takes the epidemic development as the research object and builds a model to predict the future development trend of the epidemic.

2. Research Methodology

ARIMA model for differential integrated moving average autoregressive model for differential steady-state series fitting to analyze seasonal trends and non-stationary time series.

Structure of the ARIMA(p,d,q) model.

where p is the number of autoregressive terms, q is the number of moving average terms, and d is the number of differences made when the time series becomes stationary. $\{\varepsilon_t\}$ is the white noise series when the expectation value is 0. $\nabla^d = (1-B)^d, \Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the autoregressive part of the ARMA(p,d) model; $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$, is the sliding average part of the ARMA(p,d) model.

ISSN: 2688-8653

$$\begin{cases} \Phi(B)\nabla^{\mathrm{d}} x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_S) = 0, s \neq t \\ Ex_s \varepsilon_t = 0, \forall_s < t \end{cases}$$

$$\nabla^{\mathrm{d}} = (1-B)^d$$

$$\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

$$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$

3. ARIMA Modeling Steps

(1) Data pre-processing : Due to the large size of the collected data, it is necessary to take the logarithm and then use EViews software to perform the smoothness test. After passing the smoothness test, the R-Score and R2 values were observed to determine the method used to effectively improve the fitting accuracy and provide its reference basis for the subsequent study.

(2) Smoothing of the data: If the smoothness test in the first step does not pass, the time series needs to be differenced, and the differenced time series needs to be tested for smoothness again, and then the ARIMA model is fitted after passing the test.

(3) Fitting of ARIMA model: The autocorrelation and partial autocorrelation plots are plotted on the time axis using EViews software, and the trailing and truncated tails of the autocorrelation and partial autocorrelation plots are observed, and then the estimated p and q values are selected.

(4) White noise test of ARIMA model: The purpose is to check whether the residuals of the model are consistent with the white noise series, and the residuals are white noise indicating a good fit of the model, and the residuals are not white noise then the model needs to be re-fitted.

(5) Optimization of ARIMA model: In the third step, more than one model can be passed, and the optimization of the model is to select a more accurate model, and to compare the passed model with AIC and BIC to choose the best model for prediction.

(6) Forecasting: Use the optimal model to forecast past and future data, and then analyze and summarize the forecast results.

4. Data Pre-processing

If the selected data values span a large range, the direct time series plot of the original data fluctuates greatly, and the data that change in a smaller range are not displayed accurately. Therefore, the need for data pre-processing, commonly used logarithmic transformation method. Logarithmic transformation is based on the range of the definition of the logarithmic function, the definition of the domain is a monotonically enhanced function, after taking the logarithm does not change the corresponding relationship of the data, but also to reduce the amplitude of vibration of the data, so that its prediction of the linear law is more significant. The logarithmic transformation needs to satisfy the data is greater than 0.

The data in this paper were obtained from the cumulative number of new confirmed cases of new coronary pneumonia nationwide from January 1, 2022 at 24:00 to April 29, 2022 at 24:00 as a sample data published by the National Health and Wellness Commission of the People's Republic of China.

ISSN: 2688-8653

5. Establishment of ARIMA Model

5.1. Stability Check

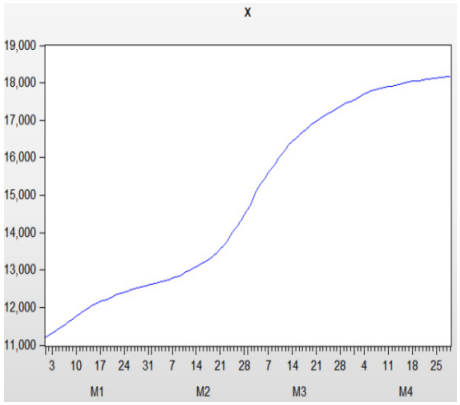


Figure 1. Original sequence x

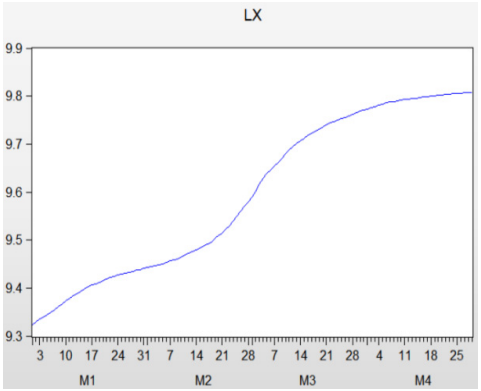


Figure 2. Logarithmically processed sequence dx

As shown in Figure 1 and Figure 2, the series has a significant trend and is a typical non-stationary series. The time series is periodic or has a significant trend, so it is necessary to perform the difference operation on the original series data.4.2 Difference processing

According to the results of the smoothness test for non-stationary time series, non-stationary time series with a random trend differencing, usually first-order differencing or second-order differencing can be carried out, until it is turned into a smooth time series. Then the differenced time series is subjected to unit root test. If the test is passed, the ARIMA model can be fitted and modeled.

From the test results, we can see that the series is not transformed into a smooth series after the first-order difference, so the second-order difference is carried out.

Table 1. dx2 unit root test results

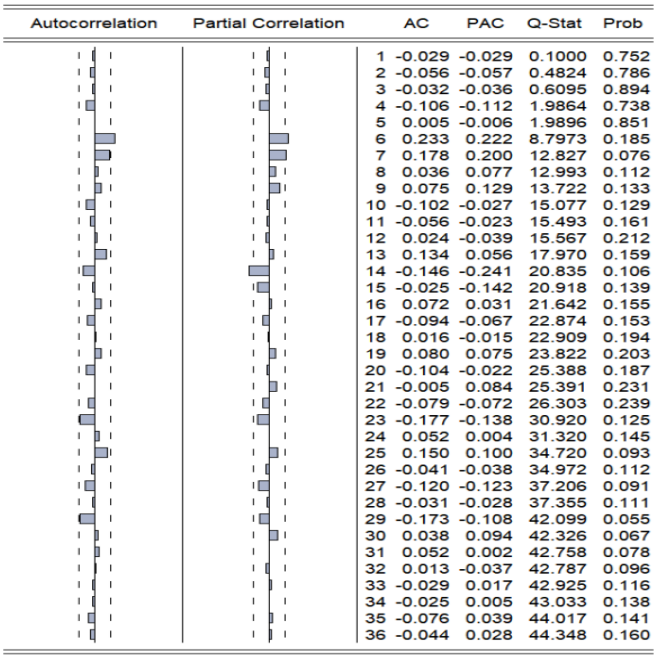
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-16.94143	0.0000
Test critical values:	1% level	-3.487550	
	5% level	-2.886509	
	10% level	-2.580163	

As shown in Figure 3 and Table 1, the dx2 series is basically smooth in the subjective observation, and the p-value is less than 0.05, which means that dx2 is a smooth series.

5.2. White Noise Test

The residual series was tested by white noise, and the results of the sequence randomness test are shown in Figure 4. Under the condition that the significance level of the test is 0.05, the p-value of the test statistic is greater than 0.05, and the residual series can be considered to contain a strong correlation at the 95% confidence level, so the residual series is a smooth white noise series, i.e. the significance test of the model is passed.

ISSN: 2688-8653



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.029	-0.029	0.1000	0.752
		2	-0.056	-0.057	0.4824	0.786
		3	-0.032	-0.036	0.6095	0.894
		4	-0.106	-0.112	1.9864	0.738
		5	0.005	-0.006	1.9896	0.851
		6	0.233	0.222	8.7973	0.185
		7	0.178	0.200	12.827	0.076
		8	0.036	0.077	12.993	0.112
		9	0.075	0.129	13.722	0.133
		10	-0.102	-0.027	15.077	0.129
		11	-0.056	-0.023	15.493	0.161
		12	0.024	-0.039	15.567	0.212
		13	0.134	0.056	17.970	0.159
		14	-0.146	-0.241	20.835	0.106
		15	-0.025	-0.142	20.918	0.139
		16	0.072	0.031	21.642	0.155
		17	-0.094	-0.067	22.874	0.153
		18	0.016	-0.015	22.909	0.194
		19	0.080	0.075	23.822	0.203
		20	-0.104	-0.022	25.388	0.187
		21	-0.005	0.084	25.391	0.231
		22	-0.079	-0.072	26.303	0.239
		23	-0.177	-0.138	30.920	0.125
		24	0.052	0.004	31.320	0.145
		25	0.150	0.100	34.720	0.093
		26	-0.041	-0.038	34.972	0.112
		27	-0.120	-0.123	37.206	0.091
		28	-0.031	-0.028	37.355	0.111
		29	-0.173	-0.108	42.099	0.055
		30	0.038	0.094	42.326	0.067
		31	0.052	0.002	42.758	0.078
		32	0.013	-0.037	42.787	0.096
		33	-0.029	0.017	42.925	0.116
		34	-0.025	0.005	43.033	0.138
		35	-0.076	0.039	44.017	0.141
		36	-0.044	0.028	44.348	0.160

Figure 3. Graph of the results of the white noise test

5.3. Model Optimization

From the parameter estimation, it can be seen that the parameter significance tests of MA(1) and AR(1) pass. By comparing the values of AIC and Schwarz criterion, it can be seen that the AR(1) model is optimal according to the AIC and SBC information criterion, and ARIMA(1,2,0) should be selected as the relatively optimal model.

The ARIMA(1,2,0) model expression is.

$$\forall^2 X_t = \frac{\varepsilon_t}{2 - 0.426963B}$$

5.4. Model Prediction

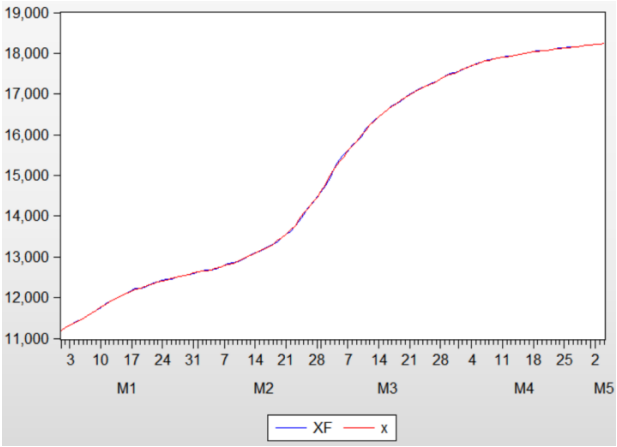


Figure 4. Projection Chart

From the prediction graph in Figure 4, it can be seen that the two lines of predicted and true values almost overlap, which shows that the model has a good prediction.

ISSN: 2688-8653

6. Conclusion

The data during the outbreak of New Crown pneumonia fit the requirements and resemble a time-series model. Therefore, it can be studied and analyzed with the time-series ARIMAX model to find out the pattern of the occurrence of New Crown pneumonia and provide a theoretical basis for prevention and control. This paper discusses the problem mainly from two aspects: theory and empirical evidence. The actual observations are influenced by seasonal changes, environmental changes and climate change, and the relationship between various elements and the interaction of these elements is statistically called correlation, and it is not easy to describe the interrelationship between the elements by mathematical models, and there is also an interdependence between the observations, and this relationship is an important property of the target object, and the dynamic model based on the data change pattern is an effective method. Therefore, the time series prediction model is a more effective tool when analyzing the data of NCCP. In conclusion, the ARIMA(1,2,0) model has a very good fit for the development trend of the novel coronary pneumonia epidemic in China, and the short-term prediction is remarkable. By accurately predicting the epidemic in the next few days, the spread of novel coronavirus was stopped, providing theoretical basis and data support for epidemic intervention decisions and prevention and control policies.

Acknowledgments

Fund Project: 2021 Innovation and Entrepreneurship Training Program for College students of Qingdao Huanghai University "Prediction and prevention and control based on Internet + smart medical care"(202113320152).

References

[1] Liu Zhongdian, Li Yanning. Prediction of the development trend of novel coronavirus pneumonia epidemic in Guangxi based on ARIMA model[J]. Journal of Guangxi Medical University, 2021, 38 (12): 2367-2371.

[2] Bai L, Guo Peiwen,Fan Jinrong. Modeling and prediction analysis of the number of confirmed cases of new crown pneumonia in Hubei Province[J]. Journal of Inspection and Quarantine, 2020,30 (02): 10-12.

[3] YANG Zhenzhen, XIE Yanqiu, JIN Xudong, ZHUANG Guimin. Prediction of infectious disease development trend based on ARIMA time series model - COVID-19 as an example[J]. China Science and Technology Information, 2021(Z1):70-72.