Prediction Method Based on Time Series

Jie Zheng[1,*], Shunping Ouyang[2], Jian Kang[3], Youcun Xiao[4]

[1] School of Information Engineering, College of Arts and Sciences・Kunming, Kunming, China, 651705

[2] College of Mathematics and Computer Science, Yunnan Minzu University, Kunming, China, 650031

[3] Yangzhou Jiangdu District Yucai Middle School, Yangzhou, China, 225200

[4] College of Humanities, College of Arts and Sciences・Kunming, Kunming, China, 651705

* Corresponding Author Email: zjhunter_bachelor@163.com

Abstract. In order to predict the development trend of a set of time series data in Yunnan Province in 2023, we collected the daily number of confirmed cases in Yunnan Province from 2020 to 2022. Firstly, according to the data analysis of the current situation of prevention and control in Yunnan Province, ARIMA model is used to predict in SPSS system. The final results showed that the number of confirmed cases in Yunnan Province in each month of 2023 showed a slow upward trend, eventually reaching more than 400. It is recommended to pay attention to the publicity of prevention and control in each state and the popularization of medical equipment.

Keywords: ARIMA model, forecast, slow rise, SPSS system.

1. Introduction

Yunnan Province is a major tourism province in China, attracting a large number of Chinese and foreign tourists every year. Therefore, time series analysis of hazard data is particularly important.

So far. Guo Jing, Liu Hanqing, Anonymous et al. Used ARIMA model to predict the total population, imported minerals and hand-foot-mouth disease in Yunnan Province [1-3]. Xia Ruyu predicted Chongqing GDP with ARIAM model [4]. Shou Yanfang, Wang Yujie, Yan Xiangxiang, Wang Yingwei and others used ARIMA model to predict a set of time series data [5-10]. However, the predecessors have never predicted the epidemic risk in Yunnan Province.

2. Data sources

The data used in this paper include the number of confirmed virus-positive cases (unit: person) per day in each city and autonomous prefecture of Yunnan Province in 2020,2021 and 2022, a total of 18 tables. This data is from the official data released by the National Health Commission, and is searched and collected through the csmar database. The data of Diqing Tibetan Autonomous Prefecture in 2020 and 2021 are from the Global Portal and Huaxia. This paper uses the number of tertiary hospitals in cities and states of Yunnan Province, which are all from the 99 hospital library.

3. Data processing

A set of time series data of Yunnan Province in recent three years is processed. Firstly, the data of 18 cities and states in Yunnan Province were extracted, and then the sum of the number of confirmed cases per unit time in each region was extracted every six months : N (unit : person), and then the average number of confirmed cases in each city and state in the past three years was calculated : $n = N / 6$ (unit : person), and compared. The number of tertiary hospitals in Yunnan Province was counted : m (unit : home), and it was found that the value of n was proportional to the virus hazard index u in the region ; the m value is inversely proportional to the hazard index u in the area.

4. Analysis of the current epidemic situation in Yunnan Province

According to the results of data processing, the average number of confirmed cases in each half of the past three years in various regions of Yunnan Province was obtained in Table 1.

According to the numerical size of each region, the GIS technology is used to draw the risk map of Yunnan Province so far. The higher the value, the deeper the color, and the lower the value, the lighter the color in Figure 1.

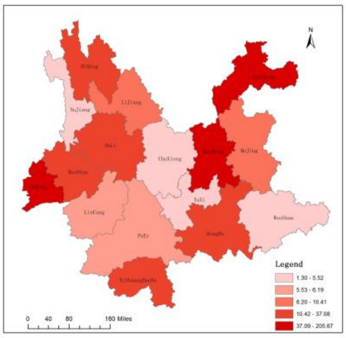

Figure 1. Coronavirus risk maps in various regions of Yunnan

From the map, we can directly divide Yunnan Province into high-risk areas, medium-risk areas and low-risk areas. Then, the number of tertiary hospitals in various regions of Yunnan Province will be collected to establish a statistical table of tertiary hospitals.

Table 1. The average number of confirmed cases per half year in each region

City	Mean value (Unit: People)	City	Mean value (Unit: People)
Zhaotong	86.3	Lijiang	9.17
Baoshan	11.67	Lincang	5.83
Chuxiong	5.3	Yuxi	5
Dali	24.5	Nujiang	1.3
Dehong	87.67	Puer	6.17
Diqing	18	Xishuangbanna	20.5
Honghe	30.83	Qujing	9.3
Kunming	205.67	Wenshan	4.67

Table 2. Number of tertiary hospitals in each province

City	Number of hospitals (Unit: Home)	City	Number of hospitals (Unit: Home)
Zhaotong	2	Lijiang	1
Baoshan	4	Lincang	3
Chuxiong	2	Yuxi	2
Dali	3	Nujiang	1
Dehong	2	Puer	2
Diqing	1	Xishuangbanna	1
Honghe	6	Qujing	6
Kunming	22	Wenshan	2

From Table 2, it can be seen that the medical conditions in Yunnan Province so far are relatively backward compared with the whole country. The distribution of tertiary hospitals is very sparse, and the medical conditions in other cities and states except Kunming, the capital city, are quite limited. The data in Table 1 and Table 2 are clustered by K-means, and the K-means hazard level clustering diagram is obtained in Figure 2. Three centroids are specified. According to the Euclidean distance, the data are roughly divided into three categories and the final classification results of the data are observed.

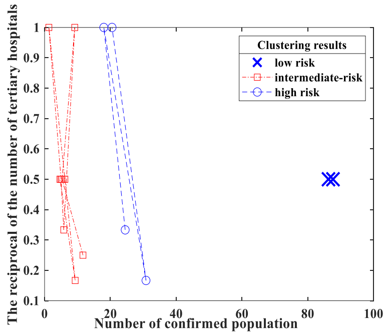

Figure 2. K-means clustering diagram of hazard level in Yunnan Province

5. The basic principle of ARIMA model

5.1. Model background

ARIMA model is an important part of time series analysis, it is an important branch of statistical science. It mainly predicts the future development trend of things by studying the data law in the process of development and change of things over time. In addition, the ARIMA model is composed of three parts: AR (autoregressive model), I (difference) and MA (moving average model).

5.2. AR (Auto Regression) model

AR is called autoregressive model, which describes the relationship between the current value and its own historical value. Its essence is to predict its future trend with the historical time data of the variable itself. The mathematical expression of the general p-order autoregressive model is:

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \mu_t \tag{1}$$

where X_t is the current value, α_i is the weight parameter, and μ_t is the random disturbance term. The formula shows that the value of a certain moment can be linearly expressed by several current moment values and corresponding weight parameters. If μ_t is a white noise $\mu_t=\varepsilon_t$, the formula is:

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t \tag{2}$$

The autoregressive model first needs to determine an order p, indicating the use of past p-period data.

5.3. MA (Moving Average) model

In the AR model, if μ_t is not a white noise, we consider it as a q-order moving average. That is:

$$\mu_t = \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \tag{3}$$

Here ε_t denotes a white noise sequence. If when $X_t = \mu_t$, the current value of the time series has nothing to do with the historical value, and only depends on the linear combination of historical white noise, the MA model is obtained:

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \beta_i \, \varepsilon_{t-1} \tag{4}$$

5.4. ARMA model

An autoregressive moving average model ARMA (p, q) is obtained by combining the AR model and the MA model:

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \, \varepsilon_{t-i} \tag{5}$$

The above equation shows that a time series can be represented by an autoregressive moving average model, which can be represented by its own past value and random disturbance term.

5.5. Difference method

5.5.1 Stationary sequence

If a series of data floats in a fixed interval, it will not increase or decrease all the time, then this series of data is a stationary sequence, otherwise it is called a non-stationary sequence.

5.5.2 Difference method

The method needed to transform a series of non-stationary sequences into stationary sequences is the difference method. A difference is made on the non-stationary sequence, and the result of the value $\boldsymbol{x_i}$ of the previous item in the sequence minus the value x_{i+1} of the latter item: $(x_i - x_{i+1})$ is used to replace x_i, which is iterated until the last item. Observe whether the final sequence is a stationary sequence. If not, continue the second difference until a stationary sequence can be obtained.

6. Parameter declaration

(1) AR is autoregressive, p is the autoregressive term, p means that from the moment forward backward p moments, you can see the autocorrelation graph to estimate.

(2) MA is the moving average, q is the moving average number of terms, which can be estimated by partial correlation diagram.

(3) d is the number of differences made when the time series becomes stationary.

7. Application of Examples

7.1. Data pre-processing

The data of each city and state in Yunnan Province in the past three years were classified and integrated. First, integrate the monthly data of each city and state in 3 years ; then, the number of confirmed cases in each city and state in each month is added to the corresponding position ; finally, we get the statistical table of the number of Yunnan Province in the past 3 years (a total of 36 months), each month.

7.2. Operation flow

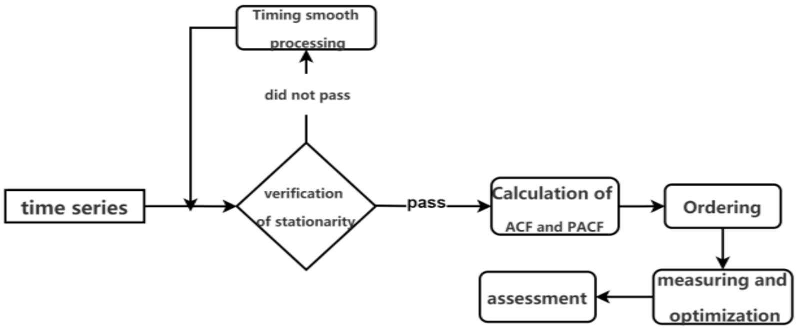

Figure 3. Operation flow chart of ARIMA algorithm

Figure 3 is operation flow chart of ARIMA algorithm.

7.3. Data stabilization

7.3.1 Current sequence diagram

First, the data is imported into SPSS, and the sequence diagram of the current data is obtained directly.

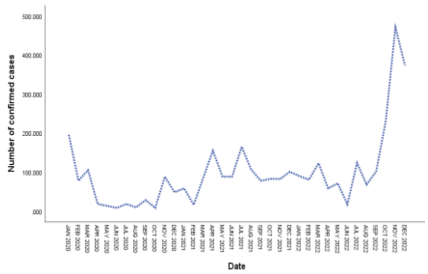

Figure 4. Sequence diagram of current data

From Figure 4, we can see that the current data has a significant trend, so we next make the first difference.

7.3.2 Perform the first difference

After the first-order difference of the data, the sequence diagram is drawn in Figure 5.

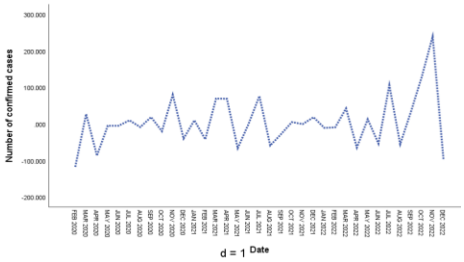

Figure 5. Data sequence diagram after first-order difference

7.4. Calculate p and q values

7.4.1 Autocorrelation of initial data

We use SPSS to analyze the autocorrelation of the initial data and draw the autocorrelation graph (ACF) and partial autocorrelation graph (PACF).

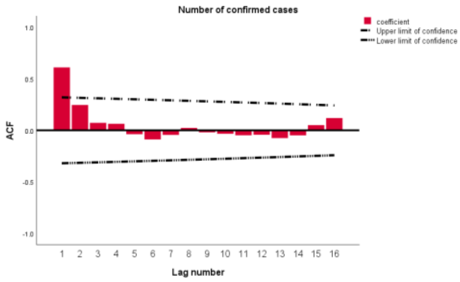

Figure 6. Autocorrelation graph

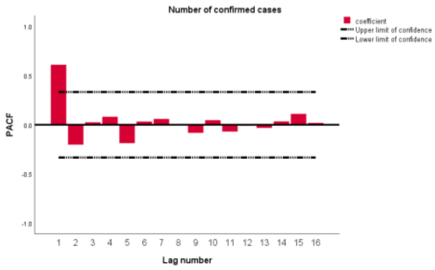

Figure 7. Partial autocorrelation graph

By observing Figure 6 and Figure 7, we find that ACF and PACF are generally ideal, but the first column of ACF and PACF exceeds the upper confidence limit. Therefore, a difference is made until the autocorrelation graph is all within the confidence limit.

7.4.2 Autocorrelation after first-order difference

We use SPSS to study the word autocorrelation after the first-order difference of the initial data, and draw the autocorrelation graph (ACF) and partial autocorrelation graph (PACF) in Figure 8 and Figure 9.

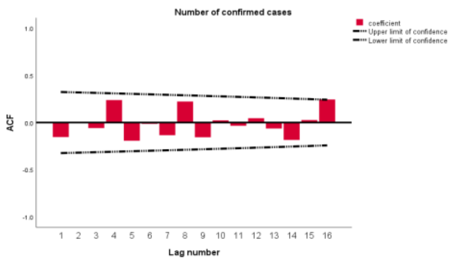

Figure 8. Autocorrelation graph

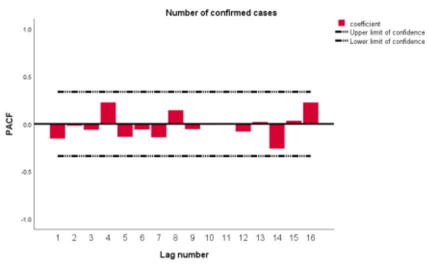

Figure 9. Partial autocorrelation graph

From Figure 8, we can see that each column from the first column is within the confidence limit. The trend is gradually decaying, and the value is always not equal to zero, so the ACF after the first-order difference is trailing, and it is zero-order trailing. Similarly, the PACF after the first-order difference in Figure 9 is also trailing, and it is a zero-order trailing.

7.4.3 Gets the values of p and q

Since the autocorrelation graphs after the first-order difference are all zero-order tailing graphs, we conclude that the value of p is 0; the value of q is 0.

7.4.4 statistical model

Table 3. Model statistical table

statistical model			
goodness-of-fit		Young-Box Q (18)	
stationary	R square	statistics	DF
0.345	0.654	20.584	18

From Table 3, we can see that the value of R square is 0.654, which has exceeded six layers, indicating that the model fitting results are better.

7.4.5 residual analysis

The residual plots of ACF and PACF in our observations.

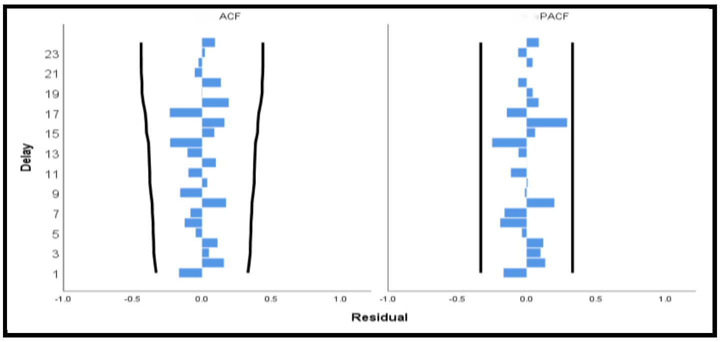

Figure 10. Autocorrelation and non-autocorrelation residuals

The results in Figure 10 at this time show that the residual has been white noise, and the residual values of ACF and PACF fall within the limit. That is to say, the accuracy of the predicted results is high.

8. Analysis and evaluation of the results

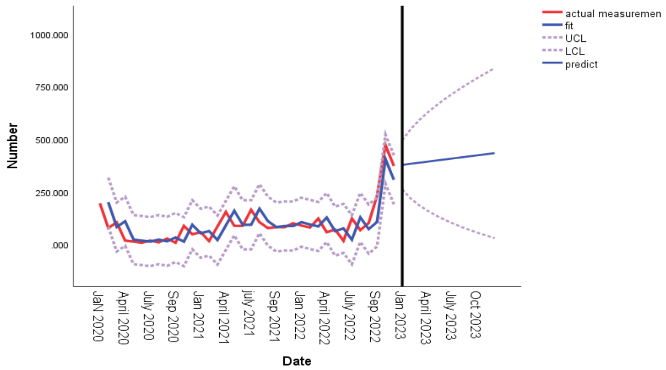

Figure 11. Prediction results

From the result Figure 11, we can clearly see that the prediction results of the ARIMA model show that the development trend of the time series of Yunnan Province will show a slow upward trend in the next 2023. The red line is the measured value. In January 2023, there is a numerical black dividing line. The blue line on the left of the dividing line is the actual fitting value, and the blue line on the right of the dividing line is the predicted value.

Table 4. Forecasts for each month of 2023

Time	Predicted value	Time	Predicted value
2023.1	378.065	2023.7	408.456
2023.2	383.130	2023.8	413.521
2023.3	388.195	2023.9	418.568
2023.4	393.260	2023.10	423.651

2023.5	398.325	2023.11	428.716
2023.6	403.391	2023.12	433.781

Table 4 is the prediction result of ARIMA. From the table, we can clearly see that Yunnan Province will reach a peak in December 2023, and the number of confirmed cases in the whole month is close to 434.

9. Conclusion

We use Origin software to fit the data predicted by SPSS, observe the trend of the data, and infer the law of development in 2023 and beyond.

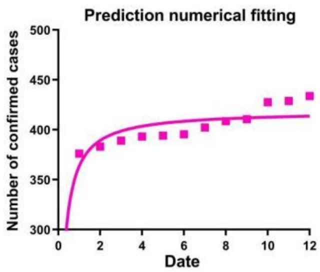

Figure12. Future trend fitting diagram

Through Figure12 observation, it is found that with the passage of time, the number of people diagnosed in each month shows a slow increasing trend and gradually tends to be gentle.

With the change of prevention and control policies, China has entered the moment of full opening and unblocking. It is reasonable that the number of confirmed cases in various regions of the country has increased. Yunnan Province should strengthen the publicity of grassroots prevention and control and the popularization of medical equipment in cities and states.

Reference

[1] Guo Jing, Zhang Yinxiang. Comparative study on the prediction model of total population in Yunnan Province [J]. Journal of Chuxiong Normal University, 2021, 36(3): 8.

[2] Dou Bin, Li Jing, Liu Hanqing. Prediction of imported mineral quality problems in Yunnan based on ARIMA model [J]. Technology and Industry, 2020, 20(4): 6.

[3] Yi Ming. Prediction and early warning model of hand-foot-mouth disease epidemic in Yunnan Province based on complex network theory [J]. Chinese Journal of Disease Control, 2015, 9(3): 3.

[4] Xia Ruyu, Wang Ziqiao. Analysis of GDP forecast of Chongqing based on ARIMA model [J]. China Storage and Transportation, 2022, 19(8): 2.

[5] Xu Jianmin, Shou Yanfang. Dynamic programming algorithm based on ARIMA model and K-means clustering analysis [J]. Journal of Chongqing Jiaotong University: Natural Science Edition, 2022, 41 (7): 5.

[6] Wang Yujie, Bai Li, Zhao Banghong. Prediction of garlic price in China based on ARIMA model [J]. Vegetables, 2021, 17(12): .5

[7] Yan Xiangxiang. Using ARIMA model to predict the park green area [J] Computer Science, 2020, 47(S02): 5.

[8] Zhao Peng, Li Lu. Research on the prediction of urban rail transit arrival volume based on ARIMA model [J]. Journal of Chongqing Jiaotong University: Natural Science Edition, 2020, 39 (1): 5.

[9] Wang Yingwei, Ma Shucai. Time series prediction based on ARIMA and LSTM hybrid model [J]. Computer application and software, 2021, 31(2): 3-4.

[10] Li Lisha. Prediction of air passenger traffic volume based on ARIMA model [J]. Scientific and technological innovation, 2022, 23(4): 9-12.