Analysis of AIDS Transmission Based on ARIMA Model

Yicheng Han [1], Nianhao Li [2, *], Leyu Qian [3] and Qinlin Yu [4]

[1] Nanchang No.2 High School, Nanchang, China

[2] Queen's University, Kingston, Ontario, Canada

[3] School of XI'AN TIE YI High school, Shan'xi, China

[4] Basis Park Lane Harbour, Huizhou, China

* Corresponding Author Email: 19nl20@queensu.ca

Abstract. In response to the ongoing challenge of infectious diseases like AIDS, infectious disease experts have turned to mathematical modeling. One such model, the ARIMA (Auto Regressive Integrated Moving Average) model, has proven effective in predicting disease spread. ARIMA relies on historical data to forecast future transmission rates, enabling proactive measures to be taken. This study utilizes the ARIMA model to predict the future trajectory of AIDS cases in Guangdong Province, China, based on historical data. Initial data analysis reveals a non-linear growth pattern in AIDS cases, emphasizing the need for a more sophisticated modeling approach. Through the application of the ARIMA model with parameter selection guided by the Bayesian Information Criterion (BIC), we achieve a robust fit to historical data. The model's predictions closely align with observed data, offering valuable insights into the potential course of the disease in the region.

Keywords: ARIMA model; predict; AIDS; influences.

1. Introduction

The rampant global infectious diseases in recent years have led people's attention to the prevention and control of infectious diseases. People have gained a more firsthand understanding of infectious diseases during the years of the pandemic. Before this pandemic happened, people often underestimate the impact of infectious diseases. What goes beyond people's understanding is COVID-19 not only has a significant impact on the global economy but also has a negative impact on people's mental health [1]. Therefore, we need to pay more attention to infectious diseases in order to reduce the harm they bring. This paer response to infectious diseases is often passive, so for infectious diseases, prevention and control of effective drugs are more important [2]. In today's COVID-19 era, which has almost ended, people still need to be vigilant and make efforts to prevent and control infectious diseases to make better judgments in the next pandemic. Therefore, we need to conduct some research on similar infectious diseases and find models for the speed of transmission to prevent and control infectious diseases in advance.

The Acquired Immune Deficiency Syndrome (AIDS) is one of the most prevalent infectious diseases in the world. It is not the deadliest infectious disease. Every AIDS patient has a different life span according to their own situation [3], but in the past 20 years, it is estimated that 36 million people have been infected with HIV and about 20 million have died [4]. AIDS was first discovered in 1981 and became notorious because of its widespread spread and high mortality in Africa. In some southern African countries, there is even a prevalence rate of over 15% [5]. AIDS is mainly transmitted through mucosa, blood, and mother to child [6]. Due to the difficulty in detecting the disease in the early stages, it is difficult to prevent and control it. As AIDS is a disease aimed at the human immune system, it cannot be cured. The two drugs for AIDS, PREP and PEP, are not universal and difficult to promote [7]. Its difficulty in prevention and control, combined with its inability to cure, has led to it becoming an infectious disease that has been circulating to this day and cannot be completely restricted. Therefore, the best way to limit the spread of AIDS is to improve people's awareness of AIDS and establish a model to make observations.

In the past decade or so, infectious disease experts have used mathematical modeling for some infectious diseases [8]. The use of mathematical models to model infectious diseases has been developed for many years, and there are now many mature models to study infectious diseases. These models can predict the spread of infectious diseases under certain conditions. Among many models, the time series analysis model, also known as the ARIMA model, has a good performance in predicting the spread of infectious diseases. Its main function is to determine the number of people infected with infectious diseases in the future based on the number of people infected with infectious diseases in the past, thereby predicting the speed of infectious disease transmission [9]. After predicting the future spread rate of infectious diseases, we can take certain measures to better respond to them.

The spread of infectious diseases involves many factors that cannot be fully considered, so it is necessary to control the regions and times involved in the data to control variables and better fit the model. In order to facilitate data collection, data on AIDS patients in Jiangsu Province, China, from 2014 to 2019 were selected, including the number of people suffering from AIDS and the number of deaths [10].

2. Methods

After determining the research object, Guangdong Province, China, was selected as the specific research object. According to research [10], Guangdong Province is prone to the spread of infectious diseases due to its geographical location, and in data collection, the data in Guangdong Province is relatively stable. As a result, Guangdong Province will be chosen as the data source for establishing the ARIMA model.

2.1. Time Series

As mentioned earlier, data from 2004 onwards and 250 months later were selected, sourced from the National Population Health Data Center. It includes the number of AIDS cases, deaths, incidence rate and mortality. According to the scatter chart in Fig. 1, it can be seen that the number of AIDS patients has increased over time. So, we hope to use this data to predict the number of future illnesses through the ARIMA model.

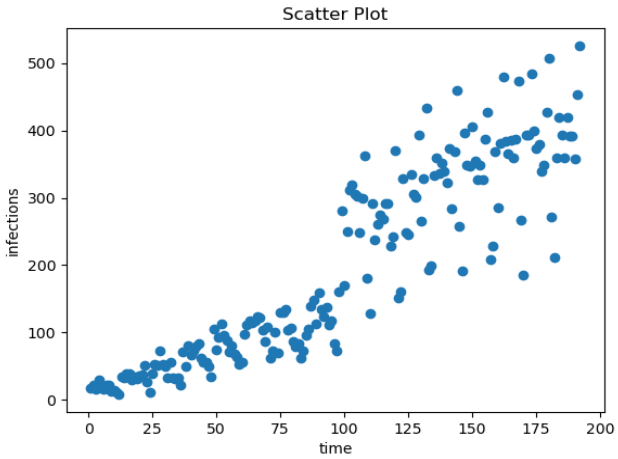

Fig. 1 Scatter plot of constructing ARIMA model.

2.2. Model Principle

The ARIMA model is suitable for non-stationary time series data, and an appropriate difference can make it a stationary sequence. The AR model is a linear model, and the general expression for the p-order autoregressive model is:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t \quad (1)$$

Among them, $\{\varepsilon_t\}$ is a white noise sequence.

The MA (q) model is called the moving average model, and a q-order moving average model can be mathematically expressed as:

$$x_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_p\varepsilon_{t-q} \tag{2}$$

When combining the AR (p) model with the MA (q) model, we obtain the ARMA (p, q) model as follows:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_p\varepsilon_{t-q} \tag{3}$$

Compared with the MA and AR models, the overall model is more complete and more versatile. Firstly, we use some order determination models to determine. After determining the order, all equation coefficients can be updated using the least squares maximum likelihood estimation or gradient descent method. Through the expression of the model, it can be concluded that continuous iteration can complete infinite prediction.

In order to determine the values of p and q in the model, a wide range of tools are used, such as AIC and BIC. AIC is a standard for evaluating the complexity of statistical models and measuring the goodness of fit of data in statistical models. BIC refers to the subjective probability estimation of partially unknown states under incomplete information, followed by Bayesian formula correction of occurrence probability, and finally using expected values and correction probability to make optimal decisions. The AIC criterion and BIC criterion are a trade-off between the likelihood function and the number of parameters. In a model, it is desirable to have larger likelihood parameters and fewer parameters. Let k be the number of parameters, the following is a method of using formulas to express the definitions of AIC and BIC:

$$AIC = -2\ln(L) + 2k \tag{4}$$

$$BIC = -2\ln(L) + kln(n) \tag{5}$$

The above n is the sequence width. When the order p and q increase, 2ln (L) will increase, and 2k will also increase. Therefore, there is an optimal value for AIC and BIC. When searching for the optimal order, it is to find the order p, q that maximizes AIC and BIC. Once the order is determined, the model can be established.

3. Results and Discussion

3.1. Preliminary Work

In order to have a general understanding of the number of AIDS patients in Guangdong Province, China, we selected 185 months of case data from 2004 onwards, including the number of patients, deaths, mortality, and morbidity. And it was made into a bar chart to visualize it in Fig. 2.

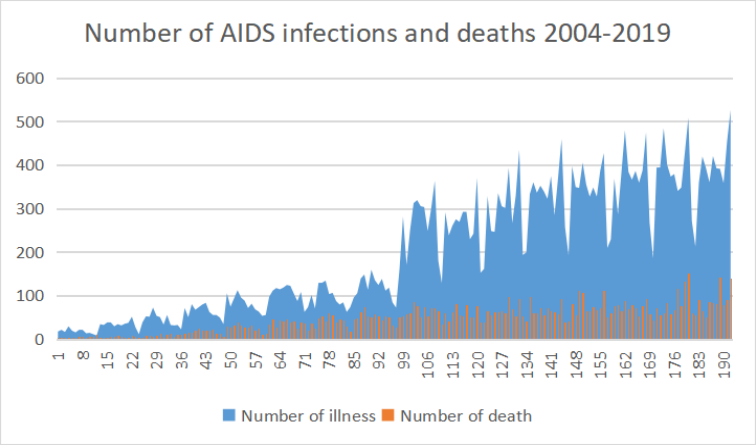

Fig. 2 Approximate ratio of death and illnesses in AIDS cases

Through visualization, we have gained a certain understanding of the characteristics of the research object's data. We found that although the number of infected individuals fluctuates slightly over time, possibly due to seasonal influences, on a large scale, the number of infected individuals is proportional to time. Moreover, the fluctuation in the number of deaths is not significant, so we will model the changes in the number of infections and time. Then, we performed differential processing on the data in Fig. 3.

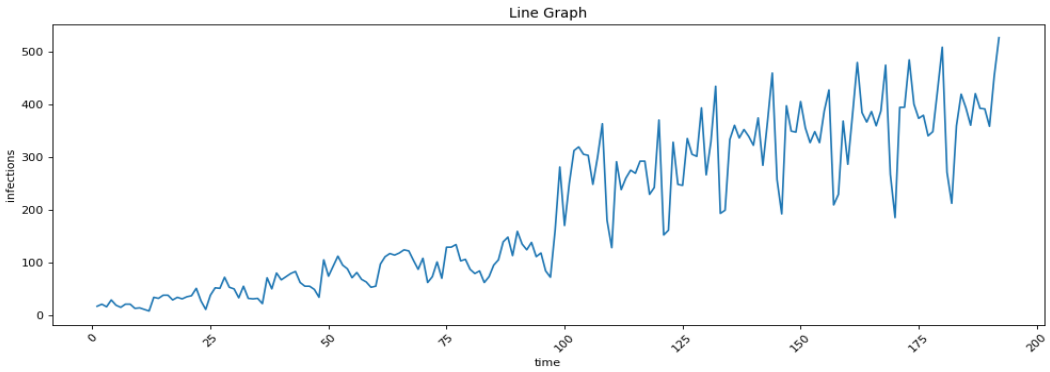

Fig. 3 Growth rate of AIDS patients in 200 months

According to the trend, we can see that the growth rate is not obvious in the first 100 months, but after 100 months, the number of AIDS patients has increased significantly, which is irregular and does not conform to linear growth. So, we will use the ARIMA model to make a certain degree of prediction. We will use the BIC matrix to determine the values of p and q, in order to establish the model (Fig. 4).

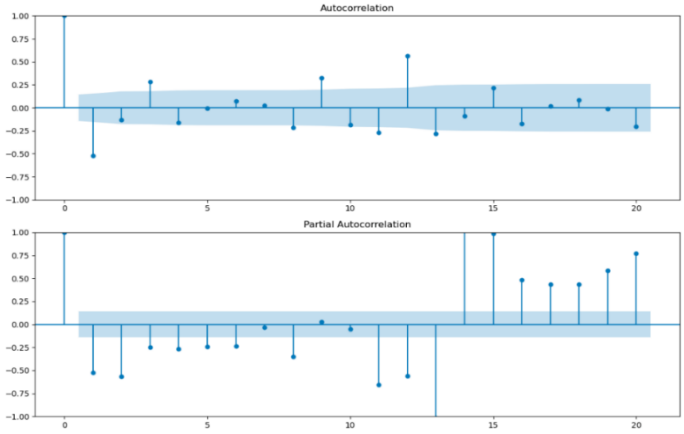

Fig. 4 ACF, PACF plots

Fig. 5 BIC matrix

After using this BIC matrix and obtaining p=3 and q=4 in shown in Fig. 5, After determining the parameters of the ARIMA model through these methods, we will fit the model and ultimately obtain the results in Fig. 6.

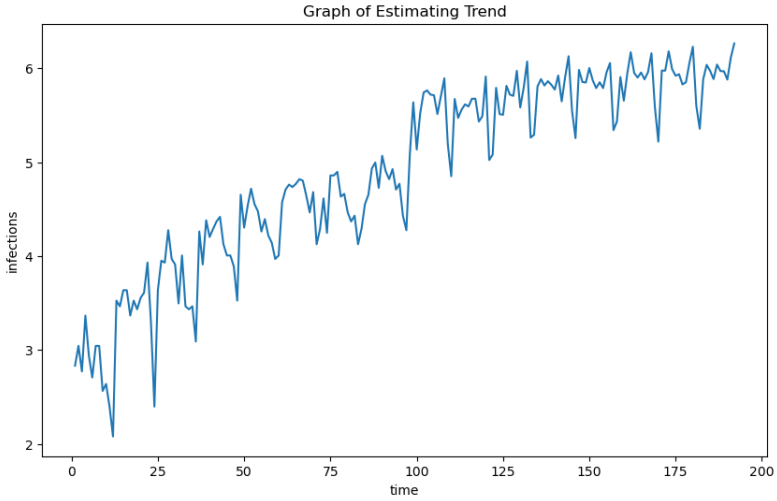

Fig. 6 graph of estimating trend

After we know the probability of an HIV infection in 200 months by the test we analyze it by combining the second-order difference. Then, we brought p=3, q=4, d=2 into the model and performed the first difference to obtain the results in shown in Fig. 7. With first-order differencing, it is clear that the jumps between values are not as pronounced as before, except in isolated places. Therefore, we do two differencing with the aim of mitigating the irregular fluctuations between the predicted data and making the fluctuation curves smoother.

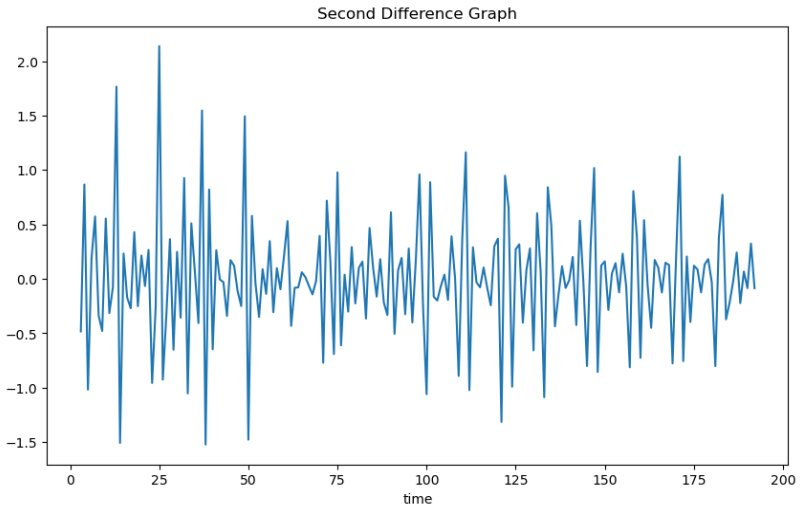

Fig. 7 Data series of second-order difference

3.2. Model Results

After establishing the fitting degree of the model and ensuring that the ARIMA model has an accurate fitting degree to the actual data, we will predict the number of AIDS cases in the future and use the actual data to verify the feasibility of prediction (Fig. 8).

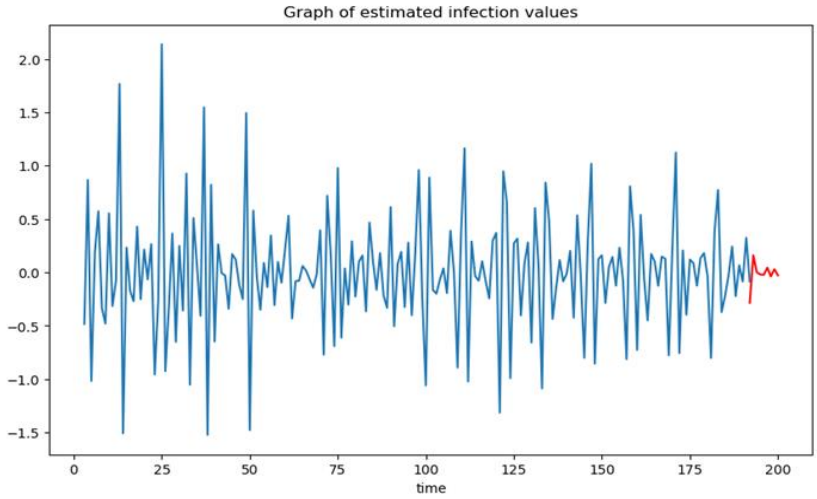

Fig. 8 Graph of estimated infection values

4. Conclusion

The analysis of the ARIMA model for predicting the future number of AIDS cases in Guangdong Province, China, reveals several important insights. Firstly, the initial data analysis showed a relatively stable trend in the number of AIDS patients during the first 100 months. However, after this period, a significant increase was observed, indicating a non-linear and irregular growth pattern. This non-linear behavior makes it challenging to predict future cases using simple linear models.

To address this complexity, this paper applied the ARIMA model and used the Bayesian Information Criterion (BIC) matrix to determine the appropriate values of p and q, which were found to be p=3 and q=4. Additionally, a differencing parameter d=2 was applied to stabilize the data. The model, with these selected parameters, demonstrated a relatively accurate fit to the historical data. This suggests that the ARIMA model can capture the underlying patterns and fluctuations in the number of AIDS cases, including the irregular growth observed in recent years. It is important to note that while the ARIMA model provides valuable insights, it may not account for unforeseen factors or sudden changes in the epidemiological landscape. Therefore, the predictions should be interpreted with caution. In the prediction analysis, the ARIMA model demonstrated its ability to capture the underlying trends and fluctuations in AIDS cases. The predictions closely followed the observed data, suggesting that the model can offer valuable insights into the potential trajectory of the disease in Guangdong Province.

In conclusion, the ARIMA model, with parameters p=3, q=4, and d=2, provides a valuable tool for predicting the future number of AIDS cases in Guangdong Province. While the model's predictions closely align with historical data, it is essential to continuously monitor and adapt prevention and control strategies to address the evolving nature of infectious diseases. Mathematical models like ARIMA can be instrumental in assisting healthcare authorities in making informed decisions to combat the spread of diseases like AIDS.

5. Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Peng Zhixing, et al. Progress in research on methods of AIDS epidemic estimation and prediction. Chinese Journal of Epidemiology, 2009, (3): 4.

[2] Peng Zhixing, et al. AIDS epidemic model in Asia and its application in the prediction of AIDS epidemic in China. Chinese Journal of Preventive Medicine, 2010, 44(2): 4.

[3] Liu Li. Estimation and prediction of AIDS epidemic in Jiangsu Province. Nanjing Medical University, 2010.

[4] Piot P, et al. The global impact of HIV/AIDS. Nature, 2021, 410(6831): 968-973.

[5] Lewthwaite P, Wilkins E. Natural history of HIV/AIDS. Medicine, 2009, 37(7): 333-337.

[6] Hermann D H. Liability related to diagnosis and transmission of AIDS. Law, Medicine and Healthcare, 1987, 15(1): 36-45.

[7] Chowell G, et al. Mathematical models to characterize early epidemic growth: A review. Physics of life reviews, 2016, 18: 66-97.

[8] Allard R. Use of time-series analysis in infectious disease surveillance. Bulletin of the World Health Organization, 1998, 76(4): 327.

[9] Burnham K P, Anderson D R. Model selection and multimodel inference: a practical information-theoretic approach. New York Springer, 2002.

[10] Meng Xiaojun. Estimation and Prediction of AIDS Epidemic Situation in Jilin Province. China Center for Disease Control and Prevention, 2012.