Using ARIMA and BP neural network to analyse incidence rate of AIDS in China

Qiuying Yang[1,*], *Xingyuan* Gao[2], *Jinwang* Zhang[1], and *Hongli* Ma[3]

[1]School of Biomedical Engineering, Capital Medical University. Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Beijing, China
[2]Beijing Robot Design Office, Beijing, China
[3]Beijing Fuxing Hospital, Capital Medical University, Beijing, China

Abstract. To analyse the characteristics of AIDS transmission from incidence, we used ARIMA and BP neural networks to model the incidence of AIDS and predict them based on modelling. When the sequence is a small sample sequence and instability, the input of the BP neural network can use raw data or stationary sequence in the ARIMA. When using the stationary sequences of incidence as the input of the BP neural network, we can obtain the output corresponding to raw data by matrix operations. Results show that raw data combined with the stationary sequences as the input of the BP neural network can get better modelling results. Moreover, all the predicted values fall within the 95% CI of the ARIMA model. Although there was also a study (reference 14) using BP to predict the incidence of AIDS, it is the original used stationary series as the input of BP in this study.

1 Introduction

The acquired immune deficiency syndrome (AIDS) is a condition caused by infection with the human immunodeficiency virus (HIV) [1, 2]. As time goes on, People with AIDS have a growing risk of developing various viral-induced cancers due to progressive failure of the immune system [3, 4]. The first reported case of AIDS was a gay man in 1981 in the United States [5]. The first case of AIDS was found in China in 1985 [6].

According to *CONFRONTING INEQUALITIES – Lessons for pandemic responses from 40 years of AIDS* issued by UNAIDS in July 2021, the number of people living with HIV went up to about 37.7 million [30.2 million to 45.1 million] in 2020. About 1.5 million [1 million to 2 million] people were newly infected in 2020, and about 680,000 [480,000 to 1 million] died from AIDS-related diseases [7]. In 2021, there were about 61.03 thousand new HIV-infected people and about 20.26 thousand deaths in China, accounting for 89.61 percent of the total deaths from infectious diseases [8].

AIDS is seriously threatening human health, and the Current Highly Active Antiretroviral Therapy (HATTY) method still has limitations. Not only can HATTY not remove the virus in the body, but it also requires lifelong treatment [9]. Taking medicine for a long time may lead to drug resistance. The treatment is so expensive. According to

* Corresponding author: y_yangqy@163.com

CONFRONTING INEQUALITIES – Lessons for pandemic responses from 40 years of AIDS, by the end of 2020, low-and-middle-income Countries had spent about $21.5 billion on measures to control HIV/AIDS, 61% of which came from domestic funds [7].

Analysis epidemic situation of AIDS is one of the vital branches of the epidemiologic study. Currently, various mathematical models are used to forecast the incidence rate of infectious diseases. However, due to the complexity and variability of practical matters, it becomes more difficult to find an appropriate prediction model [10,11]. The objective of the present study is to develop an Autoregressive Integrated Moving Average (ARIMA) model and Back Propagation (BP) neural network model for the analysis of the AIDS epidemic from aspects of Incidence by analyzing actual data and combining with the characteristics of AIDS transmission.

Applications of artificial neural networks include waveform analysis of biomedical signals, medical image analyses, and outcome prediction, biochemical data and heart sound for valve diagnostics, eye tracking, diagnosis of myocardial infarction, automatic detection of diabetic retinopathy, nephritis, and heart disease [12], and so on. In most studies, the researchers use the raw data as the input in the neural network. In this study, we use the stationary sequences in the ARIMA model as the input of the BP neural network. The results indicate that raw data combined with the stationary sequences of raw data input can get better modeling results.

2 Methods

2.1 ARIMA model [13,14]

ARIMA model, that is, the Box-Jenkins model. According to whether the data show stationary in the different parts of regression, the ARIMA model has three basic types: Moving Average (MA), Auto-Regressive (AR), and ARIMA.

A non-seasonal ARIMA model is generally denoted Arima (*p*,*d*,*q*). If the series is the stationary series, the ARIMA model can expressed as:

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} \delta_j \varepsilon_{t-j} \tag{1}$$

where *p* is the number of autoregressive terms, *q* is the number of lagged forecast errors in the prediction equation, and *y* is the estimated parameter. If the series is non-stationary, *d* is the number of non-seasonal differences needed for stationary. Construction of the ARIMA model includes four steps: data stabilization, model identification, parameter estimation, diagnostic test, and model prediction results analysis and evaluation.

Firstly, data stabilization, if necessary, can be obtained by difference.

Then, we calculate the autocorrelations function (ACF) and partial autocorrelations function (PACF) of the d-order difference sequence. PACF and ACF plots determine the *p* and *q* of the ARIMA model.

Thirdly, check the residual diagnostics of the model, particularly the residual ACF and PACF plots.

Finally, patterns that remain in the ACF and PACF may suggest the need for additional AR or MA terms.

ARIMA model construction is performed using the SPSS for Windows software package (ver.24.0, IBM).

2.2 BP neural network [14]

BP neural network is known as widely applied neural network models. The model is a feed-forward neural network. The main characteristic of the model is that the signals are transmitted forward, and the errors are transmitted in reverse. In the learning process, backpropagation is used to update the weights and thresholds of the network to achieve the minimum error sum of square. Hecht-Nielsen proved that three layers of a feed-forward network with one hidden layer can be used to learn and store the relationship between the input and output.

{X_1, X_2, ⋯ , Xn} are the input values of the BP neural network. Xi preserves the coordinates of feature points on the preoperative model. {Y_1, Y_2, ⋯ , Yn} are the forecasted values. Yi preserves the displacements of corresponding feature points from the preoperative to postoperative model. {T_1, T_2, ⋯ , Tn} are the desired outputs, which are not the same values in samples of different classes. ω ij and ω jk are the weights of the BP neural network. The topology of the BP neural network and Error Back Propagation (BP) are shown in Figure 1.

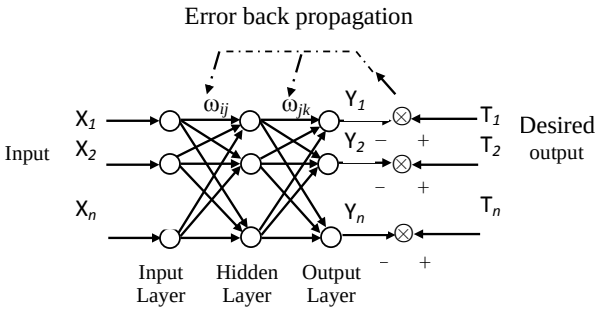

Fig. 1. The topology of BP neural network and Error Back Propagation.

2.3 Construction of BP neural network

BP neural network model construction was performed using the tool of Neural Networks Toolbox for MATLAB from Math Works, Inc. [15].

For the present, the architectures of many BP neural networks are shown as Figure 2(a). The inputs of BP neural network are raw data (R_1, R_2, R_3,…, Rn). The outputs are RY_i(i=1,2,3,…,n). In this study, the architectures of BP neural network are shown as in Figure 2(a) and 2(b) simultaneously. If raw data are non-stationary series, FD_i(i=2,3,…,n) stands for the first difference of raw data. If the first difference is non-stationary series, second difference is required, and so on, until the S-Order difference (SOD_i, i=S+1, S+2,…, n) is stationary. Then S-Order difference is used as the input of BP neural network. The output is SY'_i(i=S+1,S+2,…,n).

In this study, the maximum order of difference is 2. In Figure 2(b), FD_i(i=2,3,…,n) is the first difference of the raw data. When S=2, SOD_i(i=2,3,…,n) is the second difference of the raw data. The corresponding relationships of FD_i(i=2,3,…,n), SOD_i(i=3,4,…,n) and R_i(i=1,2,…,n) are shown as follows.

FD_i(i=2,3,…,n) is the first difference of the raw data:

$$FD_i = R_i - R_{i-1} \tag{2}$$

Using FD_i(i=2,3,…,n), R_i(i=2,3,…,n) can be represented as:

$$R_i = R_1 + FD_2 + FD_3 + \cdots + FD_i \tag{3}$$

Using a matrix to describe formula 3 as :

$$\begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_i \\ \vdots \\ R_n \end{bmatrix} = [R_1 \;\; FD_2 \;\; \cdots \;\; FD_i \;\; \cdots \;\; FD_n] \begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 & \cdots & 1 \\ & & & \vdots & & \\ 0 & 0 & \cdots & 1 & \cdots & 1 \\ & & & \vdots & & \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \tag{4}$$

SOD_i(i=3,4,…,n) is the second difference of the raw data. Using FD_i(i=2,3,…,n), SOD_i(i=3,4,…,n) can be represented as:

$$SOD_i = FD_i - FD_{i-1} \tag{5}$$

Using R_1, FD_i(i=2,3,…,n) and SOD_i(i=3,4,…,n), R_i(i=3,4,…,n)can be represented as:

$$R_i = R_1 + (n-1)FD_2 + (i-2)SOD_3 + (i-3)SOD_4 \cdots + FD_i \tag{6}$$

Using a matrix to describe formula 6 as :

$$\begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_i \\ \vdots \\ R_n \end{bmatrix} = [R_1 \;\; FD_2 \;\; SOD_3 \;\; \cdots \;\; FD_i \;\; \cdots \;\; FD_n] \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & \cdots 1 \\ 0 & 1 & 1 & \cdots & n-1 & \cdots 1 \\ 0 & 0 & 1 & \cdots & i-2 & \cdots 1 \\ & & & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & \cdots 1 \\ & & & & \vdots & \\ 0 & 0 & \cdots & & 0 & \cdots 1 \end{bmatrix} \tag{7}$$

If raw data is non-stationary series and FD_i(i=2,3,…,n) is stationary, FD_i(i=2,3,…,n) is input to BP neural network. SY'_i(i=2,3,…,n) is output corresponding to FD_i(i=2,3,…,n). The output corresponding to raw data was obtained by addition and subtraction using SY'_i(i=2,3,…,n) in formula (3) instead of FD_i(i=2,3,…,n). If FD_i(i=2,3,…,n) is still non-stationary series and SOD_i(i=3,4,…,n) is stationary, SOD_i(i=3,4,…,n) is input to BP neural network. SY'_i(i=3,4,…,n) is output corresponding to SOD_i(i=3,4,…,n). The output corresponding to raw data was obtained by addition, subtraction and simple multiplication using SY'_i(i=3,4,…,n) in formula (6) instead of SOD_i(i=3,4,…,n). In brief, if the first difference is stationary, using output of BP neural network, the first data of raw data and the first difference, the output corresponding to raw data was obtained by addition, subtraction(using the matrix of formula(4)). If the first difference is non-stationary, the second difference is stationary, using output of BP neural network, the first data of raw data, the first data of the first difference and the second difference, the output corresponding to raw data was obtained by addition, subtraction and simple multiplication(using the matrix of formula(7)).

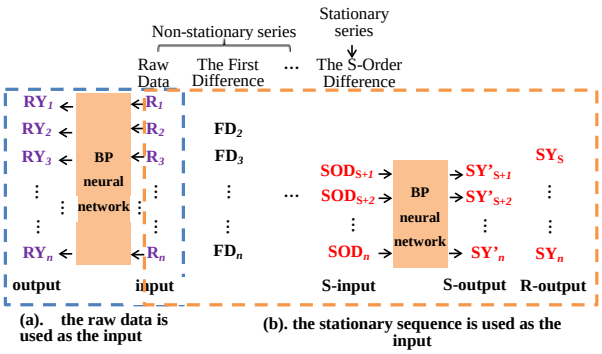

Fig. 2. Construction of BP neural network.

3 Results

Although in China, people found the first case of AIDS was in 1985, the official AIDS records of incidence rate were found until 1992. Five years from 1992 to 1996, the incidence of AIDS were 0.

3.1 Data description

In China, from 2001 to 2020, the incidence rate (1/100000) of AIDS is shown in Figure 3, which appears the upward trend and data sequence instability.

Apply the first 70% of the data sequence to modeling, use the built model to forecast the remaining 30% of data, and use the parameter Mean Absolute Error (MAE) as the evaluation criterion. MAE is expressed as formula 8:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_i| \tag{8}$$

where x_i is the actual value at some time point, $\bar{x}_i$ is the forecasted value at the same time point, and n is the number of forecast data.

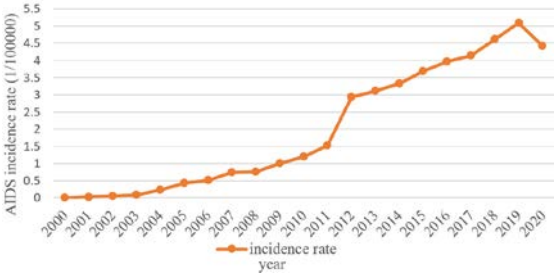

Fig. 3. AIDS incidence rate in China, 2001-2020**.**

3.2 ARIMA model

ARIMA and BP neural network model of incidence rate in China are shown in Figure 4. The 1st difference and 2nd difference of the first 70% incidence rate data are shown in Figure 4(a). The 2nd difference tends to be stationary. However, in the 1st difference, data from 2011 to 2012 are special [16]. The ACF and PACF of the 2nd difference are shown in Figure 4(b). The incidence rate is modelled with ARIMA (0,2,0). Residual plots of ACF and PACF are shown in Figure 4(c). In figure 4(d), the forecast remaining 30% data corresponding to the incidence rate forecasted by the model built based on the first 70% data. The MAE is 0.278. LCL (Lower Confidence Limits) indicates that the ARIMA model forecasts the lower 95% confidence interval (CI). UCL (Upper Confidence Limits) represents that the ARIMA model forecasts the upper 95% CI. The incidence rate in 2021 based on the established ARIMA(0,2,0) is 5.15.

3.3 BP neural network

We used the raw data and its second difference as inputs to the BP neural network separately. The raw data, modeling, and predicting results of the ARIMA model and BP neural network are shown in Figure 4(d). When the input is raw data and the second difference, the predicted value of the BP neural network all fall within the 95% CI of the ARIMA (0,2,0) model. The incidence rate in 2021 based on the BP neural network using

raw data is, and its second difference is 5.53 and 5.24. The MAE based on the BP neural network using raw data is, and its second difference is 0.673 and 0.226.

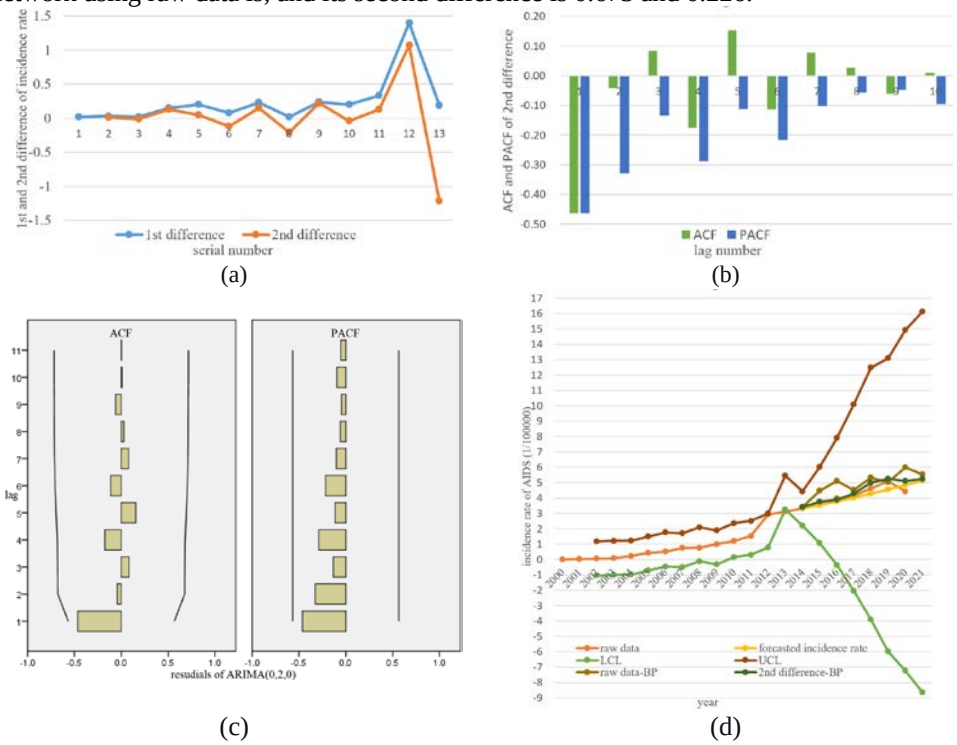

Fig. 4. ARIMA and BP neural network model of incidence rate.

4 Discussion and conclusions

AIDS has become a chronic rather than an acutely fatal disease. AIDS threatens human health and life and brings heaven economic burden to the country and the individual. This study confirmed that the artificial neural network combined ARIMA model can be used to analyze the AIDS epidemic from incidence in China. We used the raw data and the stationary sequence of AIDS incidence as inputs of the BP neural network separately. The results show that when the sequence is a small sample sequence and sequence instability, raw data, we can use the stationary sequence as the input of the BP neural network. If the input of the BP neural network is raw data and the result is unreasonable, we can use the stationary sequence as input to the BP neural network to see whether the output meets the requirements.

For the present, the input of the BP neural network is usually raw data. In this study, the study participants used raw data and stationary sequence as the input of the BP neural network. Through the difference disposal, we can obtain the data stabilization sequence. The maximum order of difference is 2. If the first difference is stationary, using the output of the BP neural network, the first data of raw data and the first difference, the output corresponding to raw data was obtained by addition, subtraction(using the matrix of formula(4)). If the first difference is non-stationary, the second difference is stationary, using the output of the BP neural network, the first data of raw data, the first data of the first difference, and the second difference, we can obtain the output corresponding to raw data by addition, subtraction, and simple multiplication(using the matrix of formula(7)).

Using raw data and the stationary sequences of the raw data input BP neural network separately, enable all the predicted values to fall within the 95% CI of the ARIMA model.

This work is supported by Beijing Natural Science Foundation (7202016).

References

1. Kent A, Sepkowitz, M. D. AIDS-The First 20 Years. The New England Journal of Medicine **344(23)**, 1764-1772(2001)
2. Daniel C. Douek, Mario Roederer, Richard A. Koup. Emerging Concepts in the Immunopathogenesis of AIDS. Annual Review Medicine. **60(60)**, 471-484(2009)
3. UNAIDS, WHO. AIDS epidemic update December, 3-10(2017)
4. Vogel M, Schwarzezander C, Wasmuth JC, et al. The Treatment of Patients With HIV. Deutsches Ärzteblatt International. **107 (28–29)**, 507–516(2010)
5. Michael S.Gottlieb. Pneumocystis pneumonia –Los Angeles. 1981. American Journal of Public Health. **96(6)**: 980-983(2006)
6. He N, Detels R. The HIV epidemic in China: history, response, and challenge. Cell Research. 5(11-12), 825-832(2005)
7. UNAIDS Joint United Nations Programme on HIV/AIDS. CONFRONTINH INEQUALITIES-Lessons for pandemic responses from 40 years of AIDS. GLOBAL AIDS UPDATE 2021. 1-386|(2021)
8. National Health commission of the People's Republic of China[EB/OL]. Overview of the National Epidemic Situation of Notifiable Infectious Diseases from January to December 2021. http://www.nhc.gov.cn/jkj/s7923/new_list.shtml(2021)
9. NCAIDS, NCSTD, China CDC. Update on the AIDS/STD epidemic in china in the first quarter of 2017. Chin J AIDS STD, **23(5)**, 371(2017)
10. Liu F, Zhu N, Qiu L,Wang JJ, et al. Application of R- based multiple seasonal ARIMA model, in predicting the incidence of hand, foot and mouth disease in Shaanxi province. Chinese Journal of Epidemiology. **37(8)**, 1117-1120(2016).
11. Amato F, Lopez A, Pena-Mendez EM, et al. Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine. **11(2)**, 47–58(2013)
12. Ferhat O, Vilarino F. Low cost eye tracking: the current panorama. Computational Intelligence and Neuroscience. 1–14(2016)
13. Haneen Alabdulrazzaq[1], Mohammed N Alenezi[1], Yasmeen Rawajfih[2] et al. On the accuracy of ARIMA based prediction of COVID-19 spread. Results Phys. (2021). doi: 10.1016/j.rinp.2021.104509
14. Li Z, Li Y. A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. BMC Med Inform Decis Mak. **20(1)**,143 (2020). doi: 10.1186/s12911-020-01157-3
15. Beale MH, Hagan MT, Demuth HB. Neural Networks Toolbox. User's Guide for MATLAB R2020b. Natrick: The Math Works (2020)
16. Ministry of Health of the People's Republic of China, Joint United Nations Programme on HIV/AIDS, World Health Organization. Estimation of AIDS Epidemic in China in 2011.11(2011)