Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

Mathematical Biosciences and Engineering, 16(4): 2266–2276
DOI: 10.3934/mbe.2019112
Received: 15 December 2018
Accepted: 25 Feburary 2019
Published: 15 March 2019

Research article

Comparative study of SARIMA and NARX models in predicting the incidence of schistosomiasis in China

Xinya Yu, Zhuang Chen and Longxing Qi*

School of Mathematical Sciences, Anhui University, Hefei, 230601, P.R.China

* **Correspondence:** Email:qilx@ahu.edu.cn.

Abstract: In this paper, based on the data of the incidence of schistosomiasis in China from January 2011 to May 2018 we establish SARIMA model and NARX model. These two models are used to predict the incidence of schistosomiasis in China from June 2018 to September 2018. By comparing the mean square error and the mean absolute error of two sets of predicted values, the results show that the NARX model is better and it has an effective forecasting precision to incidence of schistosomiasis. Then according to the results, a mixed model called NARX-SARIMA model is used to predict the incidence future trends and make a comparison with the two model. The mixed model has a better application based on its good fitting capability.

Keywords: schistosomiasis; incidence; autoregressive integrated moving average (ARIMA) model; NARX model

1. Introduction

Schistosomiasis is a zoonotic disease that endangers human health and is a parasitic disease closely related to biological and aquatic environments. It is distributed in many countries around the world, and China is one of these countries. With the diversification of control methods, the national schistosomiasis control and prevention work has achieved remarkable results, and the incidence and the number of death has gradually been controlled [1, 2, 3]. However, there is still a significant outbreak of schistosomiasis. Further prevention and control measures are needed. In order to understand the future trends of schistosomiasis, we consider using time series to predict the future incidence of schistosomiasis [4, 5].

Studying the incidence of schistosomiasis has been of interest to many scientists. It is well known that the incidence of schistosomiasis can be researched by differential equations, statistic model and so on. Qiyong Liu et.al [6] predicted the incidence of hemorrhagic fever with renal syndrome in China through the ARIMA model. Historical data showed that the ARIMA model is an effective method for predicting the future incidence of the disease. Shahid MA, et.al [7] studied the application of NARX model in heart disease prediction, and the results showed that the prediction results are in line with future trends. These scholars have applied time series or neural networks to the prediction of diseases, but neither of them has applied these two model to the prediction of schistosomiasis. We propose an optimal model through the comparison of ARIMA and NARX prediction results based on actual data from the incidence of schistosomiasis in China [8, 9].

2. Data and methods

2.1. The data

The number of schistosomiasis cases in China from January 2011 to September 2018 was collected from the website of the Chinese Center For Disease Control And Prevention (CDC, for short). The data are obtained from the web of *www.chinacdc.cn* which is recorded monthly. Because of the stability of the population, the annual data is used instead of the monthly data which comes from the website of the National Bureau of Statistics of China. The incidence was classified according to the date of these two sets (1/10000).

2.2. Methods

Autoregressive Integrated Moving Average (ARIMA, for short) model is a time series model which the autoregressive process uses its own regression variables. The model can be expressed as ARIMA (p,d,q) [10, 11]. Here AR represents autoregression, MA represents moving average, p and q are autoregressive terms, respectively, d is the number of regular differences in order to make the sequence to be smooth. It was observed that the intermediate host snail of schistosomiasis breed at a suitable temperature. The seasonal factors should be take into the model because of the schistosomiasis outbreak period. According to this, we consider a special ARIMA model, Seasonal Autoregressive Integrated Moving Average (SARIMA, for short) model which can be expressed as ARIMA$(p, d, q)(P, D, Q)^s$. P, Q represent the seasonal autoregressive and moving average orders, respectively, and D is the number of seasonal differences [12, 13]. The SARIMA model can be set up as follows:

$$\phi(B)U(B^s)\nabla^d\nabla_s^D x_t = \Theta(B)V(B^s)\varepsilon_t,$$

where B is the backward shift operator; B^s the seasonal shift operator; ε_t is the random disturbance term [14].

$$\begin{aligned}\phi(B) &= 1 - \varphi_1 B - \varphi_2 B^2 - \cdots \varphi_p B^p,\\ \Theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \cdots \theta_q B^q,\\ U(B^s) &= 1 - u_1 B^s - u_2 B^{2s} - \cdots - u_P B^{Ps},\\ V(B^s) &= 1 - v_1 B^s - v_2 B^{2s} - \cdots - v_Q B^{Qs}.\end{aligned}$$

Artificial neural networks is an artificial computing system that mimics the interconnected network of animal brain neurons. This neural network model is not an algorithm rather than a framework model which is combined by many individual computing units with many different machine algorithms. Using artificial neural network models to solve problems does not require us to develop a program model for performing specific tasks. NARX model accomplish corresponding tasks through autonomous learning. Artificial neural networks rely on the connection of each artificial neurons. The connection of each unit corresponds to a synapse in the brain of the organism that is capable of passing signals from the first unit to the second unit. In our common artificial neural network, this signal is a real number and is calculated internally by each nonlinear function model [15]. Artificial neural networks was put forward to solve specific tasks in the same way as the human brain from the beginning. There are also many classifications in artificial neural networks such as Back Popagation (BP, for short) networks, Radial Basis Function (RBF, for short) networks, and Hopfield networks. This article mainly uses the NARX (Nonlinear Autoregressive) model in BP network for calculation. The NARX model is a widely used, well-fitting, and predictable model. It is well suited for modeling nonlinear system data, especially the time series models. Compared with other neural networks, we can find that it has the following important properties; (1) NARX model has better learning result than other neural networks. Compared with the general BP neural network, the NARX model adds certain sequence learning ability to the neural network. (2) The NARX model has a faster convergence rate and better popilarization. The fitting effect of many experimental data models can show that the NARX model can more accurately fit the correlation of long-term data models than the previous time series models.

3. Establishment of SARIMA model

3.1. Trends of schistosomiasis incidence in the country

The monthly national schistosomiasis incidence rate from January 2011 to May 2018 has a seasonal fluctuation for 12 months. It can be clearly observed from the Figure 1 that the peak value appeared between September to November every year, and it is also the period of schistosomiasis.

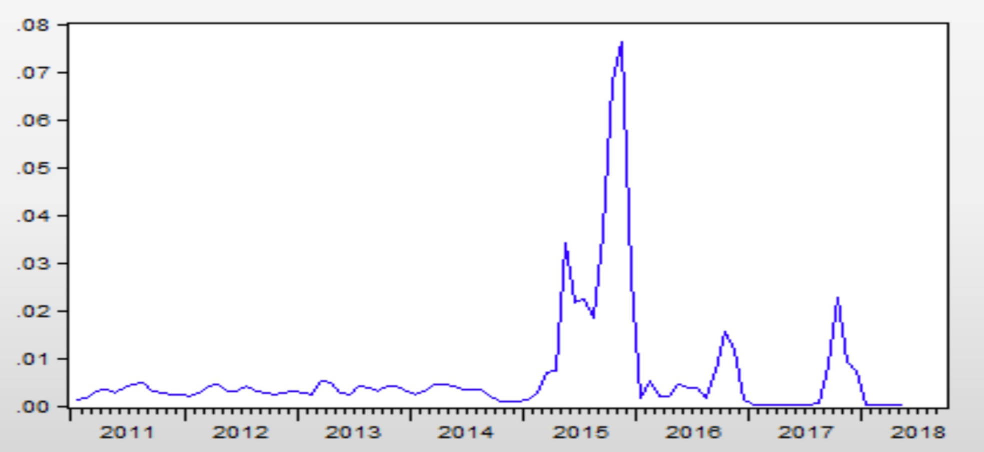


Figure 1. Timing diagram of the incidence of schistosomiasis in China.

3.2. Data preprocessing

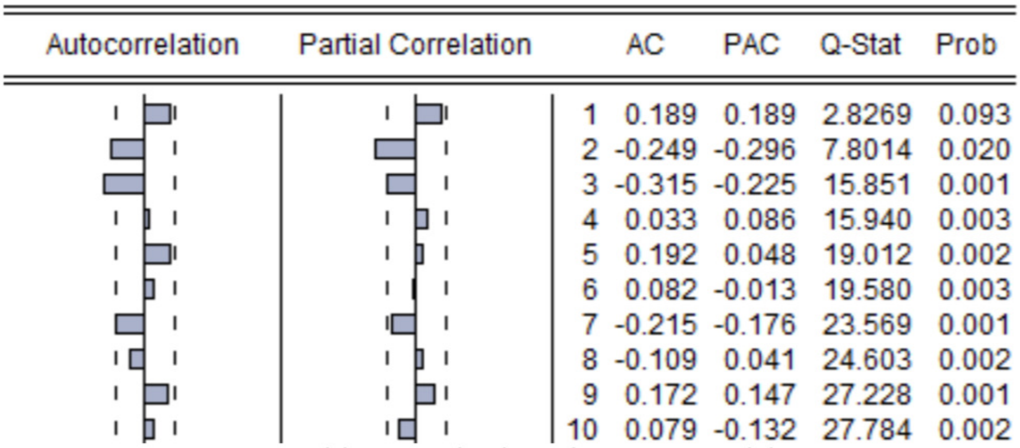
In order to smooth the data and reduce the fluctuations, the data needs to be differenced. Since the data has seasonal factors, we need consider seasonal differences and eliminate seasonal factors. We use the Augmented Dickey-Fuller test to test the null hypothesis that a unit root is present in a time series sample which is showed in Table 1. The sequence is smooth after the first-order difference and the first-order seasonal difference.

Table 1. Unit root test after difference.

	t-Statistic	Prob*
Augmented Dickey-Fuller test statistic	-7.275894	0.0000
Test critical values: 1% level	-3.521579	
5% level	-2.901217	
10% level	-2.587981	

3.3. Model Identification

Since the data has been made a first-order difference and a seasonal difference, the parameter d=1, D=1, and because s represents the length of the season, s=12.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.189	0.189	2.8269	0.093
		2	-0.249	-0.296	7.8014	0.020
		3	-0.315	-0.225	15.851	0.001
		4	0.033	0.086	15.940	0.003
		5	0.192	0.048	19.012	0.002
		6	0.082	-0.013	19.580	0.003
		7	-0.215	-0.176	23.569	0.001
		8	-0.109	0.041	24.603	0.002
		9	0.172	0.147	27.228	0.001
		10	0.079	-0.132	27.784	0.002

Figure 2. ACF and PACF diagram.

From Figure 2, autocorrelation coefficient tailing, partial autocorrelation coefficient censored and the order is 3 , we judge p = 3, q = 0. In order to further validate our judgment, we make some different group by q= 0, p= (0,1,2,3), and let P and Q be (0,1,2) respectively to experiment. The optimal model was selected based on the p-value and the Akaike information criterion. When $p = 1$, the p-value is over 0.05, so p can't be 1. According to the software results, Q must evaluate to 0 to make sure the p-value corresponding to the variable is less than 0.05. Let $Q = 0, P \neq 0$, the p-value corresponding to the variable also has a value greater than 0.05. So $P = 0$. It is based on the concept of entropy and can weigh the complexity of the estimated model and the superiority of the model fitting data. AIC provides a means for model selection. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model. The AIC of ARIMA(2,1,0)$(0,1,0)^{12}$ is -6.200048.The model is validated as ARIMA(2,1,0)$(0,1,0)^{12}$ (Table 2).

Table 2. Selection of SARIMA Model.

$(p,d,q)(P,D,Q)^{12}$	variable	P-value	AIC
$(1,1,0)(0,1,0)^{12}$	AR(1)	0.1018	
$(2,1,0)(0,1,0)^{12}$	AR(1)	0.0328	-6.200048
	AR(2)	0.0106	
$(2,1,0)(1,1,0)^{12}$	AR(1)	0.1557	-6.209163
	AR(2)	0.0259	
	SAR(1)	0.0002	
$(2,1,0)(0,1,1)^{12}$	AR(1)	0.1343	-6.742250
	AR(2)	0.0172	
	SMA(1)	0.0000	

3.4. Parameter estimation

Table 3. Correlation coefficient.

Variable	Coefficient	Std.Error	t-Statistic	Prob
AR(1)	0.245011	0.112587	2.176190	0.0328
AR(2)	-0.295527	0.112586	-2.624890	0.0106

Based on ARIMA(2,1,0)$(0,1,0)^{12}$and correlation coefficient, we rewrite the model as follow

$$(1-B)(1-B^{12})(1-0.2450B+0.2956B^2)X_t = e_t. \tag{3.1}$$

According the model, BX_t represents the value of X_t lag one phase , that is the value of X_{t-1}, and B^2X_t represents the value of lag two phase. We use existing data based on the expansion of the model to estimate unknown data (Table 3).

3.5. Model diagnosis

The unit root test is performed on the residual sequence. And the p-value is 0.0000 and less than 0.05. It can be seen that there is no autocorrelation of the residual sequence, which can be seen as a white noise sequence .So the selected model is reasonable.

4. Establishment of NARX model

4.1. NARX model

In time series models, nonlinear autoregressive models are a class of nonlinear regression models with external inputs, which means that the current values of such models are related to the past values of the sequence and the external input drive values of past values. The model can be expressed as:

$$y_t = F(y_{t-1}, y_{t-2}, \cdots, u_t, u_{t-1}, u_{t-2}, u_{t-3}, \cdots) + \varepsilon_t.$$

Here y is the research variable and u is the external input variable. In this model, the value of u can both help calculate the present value of y and also help predict the future value of y. So y can be seen as a function which is defined by x_t and t. ε_t is an error term (Figure 3). Nonlinear Auto Regressive exogenous Neural Network is a neural network model based on nonlinear autoregressive model. It has feedback and memory functions. The output of each moment is based on the dynamic synthesis result of the system before the current moment. It has significant advantages in the modeling and simulation of time series dynamic changes. In the past research, many scholars used the NARX model as an external input nonlinear autoregressive neural network [16]. It is a general recurrent neural network that introduces a delay module and feedback based on a static multilayer perceptron. The main feature of this network is the addition of Tapped Delay Lines delay unit and output feedback at the input layer level, which enables better recording of past data input and output states, combined with the nonlinear mapping characteristics of multi-layer perceptron (MLP). Autoregressive artificial neural network can predict the next time status value through past status [17].

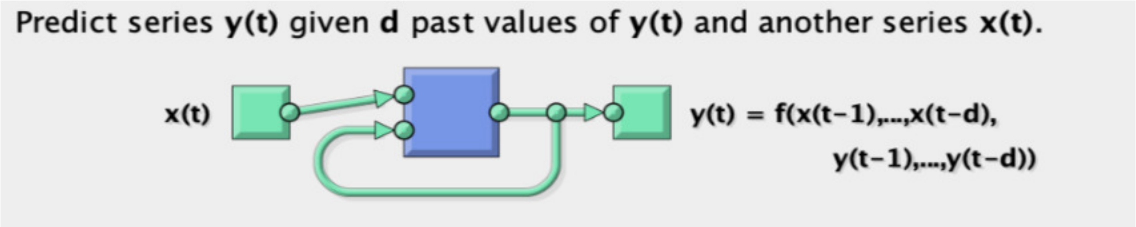

Predict series y(t) given d past values of y(t) and another series x(t).
x(t)
y(t) = f(x(t−1),...,x(t−d),
y(t−1),...,y(t−d))

Figure 3. NARX model structure.

4.2. *Determination of structural parameters in the model*

When building a suitable NARX model, it is important to determine the model structure. This includes the number of hidden layers, the number of hidden layer nodes, and the input layer delay order. If these parameters are too small, the accuracy of the model is not sufficient and the parameters of the engine cannot be correctly reflected. Correspondingly, if the parameters of each model are too large, the model will be over-fitting, which will reduce the functionality of the model and increase the calculation time and complexity. Therefore, it is necessary to select various model structural parameters reasonablely (Figure 4).

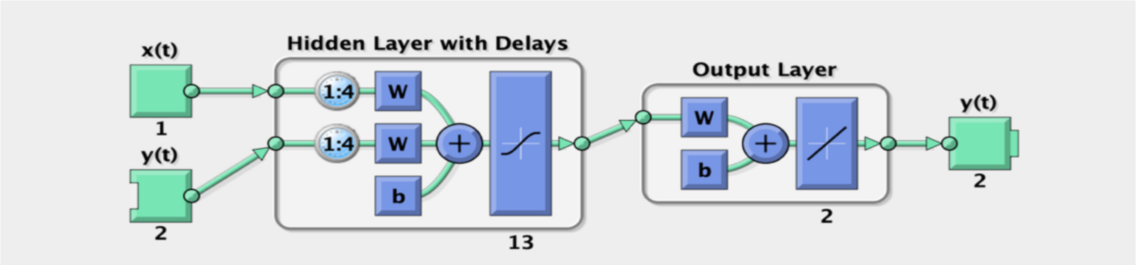


Figure 4. Location of the structure parameters of the NARX model.

We can identify the nonlinear characteristics of the model by using the Kolmogorov theorem. We choose one hidden layer. We use the loss function J-based method based on the system identification order theory to determine the order to be 12 multiple times. For the nonlinear difference equation, there is no definite decision theorem. We use the experimental method to determine the number of hidden layer nodes of the model is 3.

4.3. Modeling and Simulation

Using the parameters determined above, we built the NARX model and entered the data for fitting. The software is Matlab R2017b Version. And we use the network toolbox package to complete this work. The fitted image was obtained as follow. In this figure, yellow line means the error distance from the predict value from the true value. It can be seen from the image that the error between the predicted value and the true value is small (Figure 5).

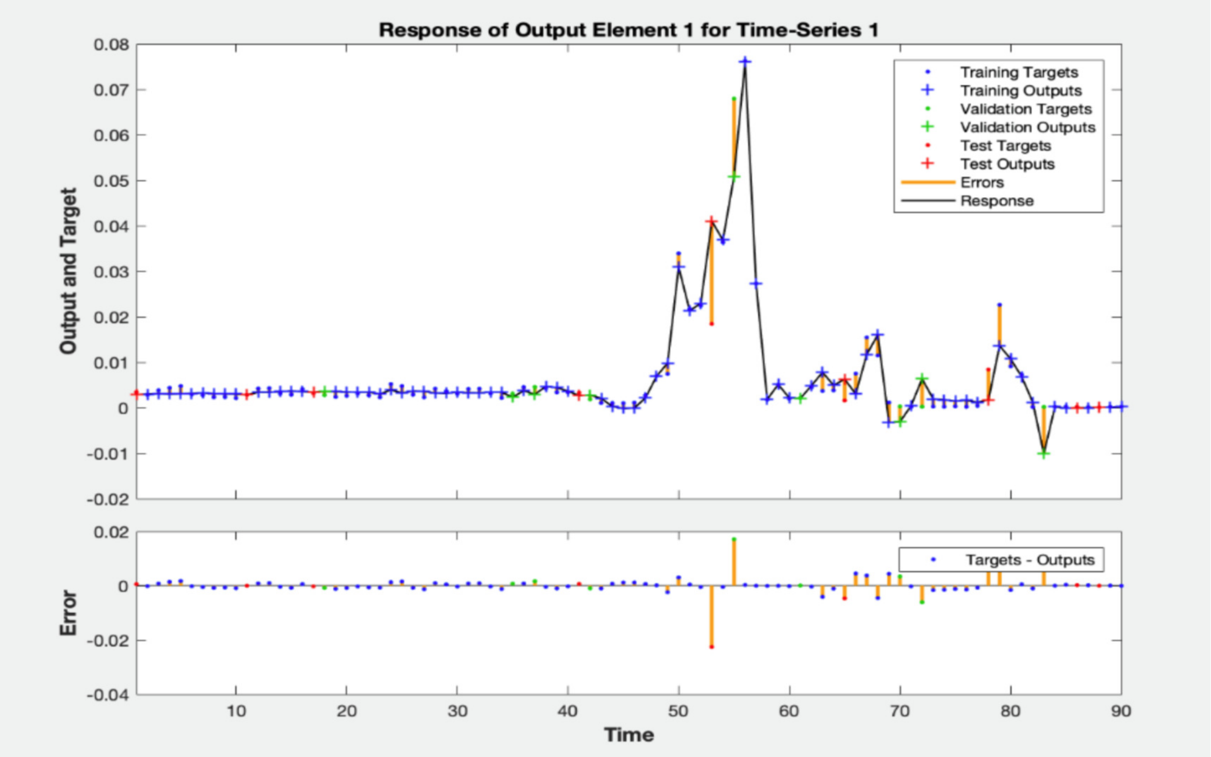


Figure 5. NARX model predicted value and true value error map.

It can be seen from the error autocorrelation coefficient graph in Figure 6 that only the confidence interval of the error autocorrelation coefficient exceeds 90 when the delay is 0. And blue bar is between these two dotted lines which means all the correlations is under the cofidence limit. The other autocorrelation coefficients are within the confidence interval and fluctuate around 0, indicating that the model is reasonable (Figure 6).

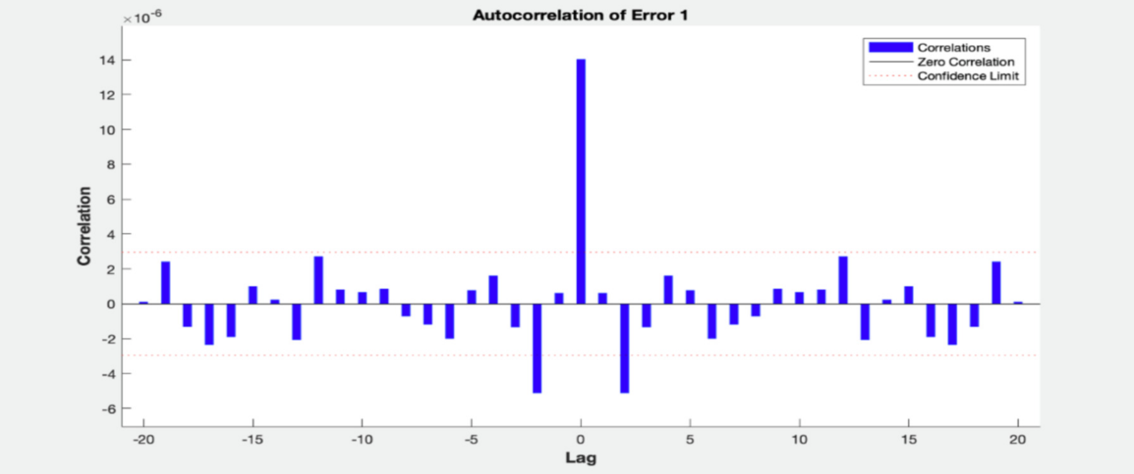


Figure 6. Error autocorrelation coefficient graph.

5. Comparison of two model predictions and results

Based on data from January 2011 to May 18, 2011, the incidence of schistosomiasis from June 2018 to September 2018 can be predicted by software. Table 4 shows the predicted values and true values obtained by the two models. The comparison shows that the predicted values which was obtained by the NARX model are closer to the true values than the predicted values obtained by SARIMA. To further compare the validity of the two models, the mean square error can be calculated (Table 4).

Table 4. Predicted values based on two models.

Moon	Actual value	SARIMA Pridicted value	NARX Predicted value
June, 2018	0.00015065	0.000235014	0.000057225
July, 2018	0.00009326	0.000140524	0.000143334
August, 2018	0.00017934	0.000403713	0.000135688
September,2018	0.00021521	0.008352088	0.000155781

For compare these two models, we use the value of the mean square error and the mean absolute error.The mean square error can be calculated by this formula

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(f_i - y_i)^2.$$

And the mean absolute error can be calculated by this formula

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|(f_i - y_i)|.$$

f_i represents predictive value. y_i represents the true value. By calculating the mean square error(MSE, for short) and the mean absolute error(MAE, for short) of the two models, the MSE and MAE of the NARX model are smaller than those of the SARIMA model. Therefore, the NARX model is more suitable for predicting the incidence of schistosomiasis. However, without using software for prediction, NARX does not have a defined model expression to calculate unknown data, and SARIMA has specific model expressions, and the SARIMA model can be considered for calculation (Table 5).

Table 5. Comparison of the results of the two models.

	MSE	MAE
SARIMA	0.00001656713	0.002123222
NARX	0.00000000417	0.000036606

6. Establishment of NARX-SARIMA model and discussion

This model separates incidence time series to decisive sequence and error sequence. We build NARX model to fit the decisive sequence and SARIMA model to fit the error sequence. The models fitted interval (May, 2011-May, 2018) and predicted values interval (June,18-September,2018) are presented in Figure 7.

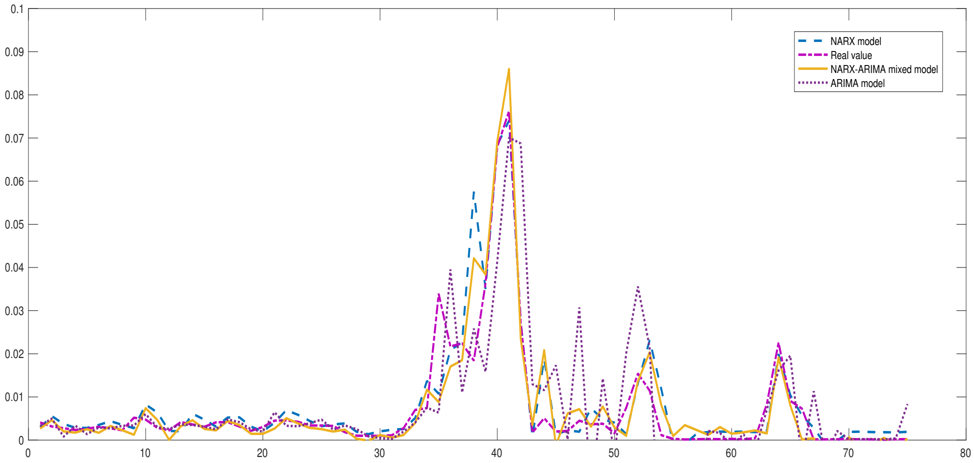


Figure 7. Comparison of predicted and actual values of three models.

In the figure, the incidence is the vertical axis and month order is horizontal axis. From the Figure 7 we can find out that, the fixed model can also fit the real value. Time series analysis of surveillance data on incidence of various schistosomiasis is very helpful in developing hypotheses to explain and anticipate the dynamics of the observed phenomena and subsequently in the establishment of a quality control system and reallocation of resources. SARIMA model is one of the most widely used time-series forecasting techniques because of its structured modeling basis and acceptable forecasting performance.

In this study, the establishment of SARIMA model is simple and the fitting effect is good at a short time. NARX has a good simulating effect and forecasting precision. But it takes a lot of time to calculate and compare each values to choose the best one. In order to improve the model, updating the forecasts is our first goal. We should concern seasonal factor when we want to bulid a model to predict the incidence of schistosomiasis. The NARX model and the mixed model have a better pridict effection of the incidence than simple SARIMA model. A simple NARX model or NARX-SARIMA model has a general applicability for different data. These three model have their own advantage and disadvantage. NARX model and NARX-SARIMA model have higher accuracy than SARIMA model. NARX model has the best accuracy after several parameter adjustment. But much work is needed for the process of parameter selection and operation. Compared with NARX model, NARX-SARIMA model can save much time and get sufficient accuracy.

Acknowledgement

The authors greatly appreciate the anonymous referees for their valuable comments which help to improve the paper. This research is supported by National Natural Science Foundation of China (11401002), Natural Science Fund for Colleges and Universities in Anhui Province(KJ2018A0029) and Quality Engineering Project of Anhui University (ZLTS2016065).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. L.X. Qi, M. Xue and J.A. Cui, et al., Schistosomiasis Transmission Model and its Control in Anhui Province, *B. Math Biol.*, **80** (2018), 2435–2451.
2. L.X. Qi, Y.W. Tang and S.J. Tian, Parameter estimation of modeling schistosomiasis transmission for four provinces in China, *Math. Biosci. Eng.*, **16** (2019), 1005–1020.
3. L.X Qi, S.J Tian and J.A Cui, et al., Multiple infection leads to backward bifurcation for a schistosomiasis model, *Math. Biosci. Eng.*, **16** (2019), 701–712.
4. L.G. Song, X.Y. Wu and M. Sacko, et al., History of schistosomiasis epidemiology, current status, and challenges in China: on the road to schistosomiasis elimination, *Parasitol. Res.*, **115** (2016), 4071–4081.
5. X.N. Zhou, L.Y. Wang and M.G Chen, et al., The public health significance and control of schistosomiasis in China–then and now, *Acta. Trop.*, **96** (2005), 97–105.
6. Q.Y. Liu, X.D. Liu and B.F Jiang, et al., Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model, *BMC. Infect. Dis.*, **11** (2011), 1–7.
7. S.M. Awan, M.U. Riaz and A.G. Khan, Prediction of heart disease using artificial neural network, *VFAST.*, **13** (2018), 102–112.
8. M.E. Banihabib, A. Ahmadian and F.S Jamali, Hybrid DARIMA-NARX model for forecasting long-term daily inflow to Dez reservoir using the North Atlantic Oscillation (NAO) and rainfall data, *GeoResJ.*, **13** (2017), 9–16.
9. G.E.P. Box, G.M. Jenkins and G.C. Reinsel, Time series analysis: Forecasting and control, *Prentice Hall.*, 1994.

10. M.Y. Anwar, J.A. Lewnard and S. Parikh, et al., Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence, *Malaria. J.*, **15** (2016), 566–576.

11. A. Earnest, M.I. Chen and D. Ng, et al., Using autoregressive integrated moving average(ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore, *Bmc. Health. Serv. Res.*, **5** (2005), 36–44.

12. F. Cortes, C.M.T. Martelli and R.A.A. Ximenes, et al., Time series analysis of dengue surveillance data in two Brazilian cities, *Acta. Trop.*, **182** (2018), 190–197.

13. F.F. Nobre, A.B.S. Monteiro and P.B. Telles, et al., Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology, *Stat. Med.*, **20** (2001), 3051–3069.

14. R. Allard, Use of time-series analysis in infectious disease surveillance, *B. World. Health. Organ.*, **76** (1998), 327–333.

15. J.M.P.M. Júnior and G.A. Barreto, Long-term time series prediction with the NARX network: An empirical evaluation, *Neurocomputing.*, **71** (2008), 3335–3343.

16. E. Diaconescu, The use of NARX Neural Networks to predict Chaotic Time Series, *WSEAS.*, **3** (2008), 182–191.

17. M. Valipour, M.E. Banihabib and S.M.R Behbahani, Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir, *J. Hydrol.*, **476** (2013), 433–441.