BMJ Open

Predictive study of the tuberculosis incidence by time series method combined with Elman neural network in Kashgar, China

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041040
Article Type:	Original research
Date Submitted by the Author:	30-May-2020
Complete List of Authors:	Zheng, Yanling; Xinjiang Medical University, College of Medical Engineering and Technology Zhang, Xueliang; Xinjiang Medical University, College of Medical Engineering and Technology, Xinjiang Medical University Wang, Xijiang; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People's Republic of China Wang, Kai; Xinjiang Medical University, College of Medical Engineering and Technology,Urumqi 830054, People's Republic of China Cui, Yan; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,Urumqi 830001, People's Republic of China
Keywords:	International health services < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Tuberculosis < INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS

Predictive study of the tuberculosis incidence by time series method combined with Elman neural network in Kashgar, China

Yanling Zheng [1] , XueLiang Zhang [1*], XiJang Wang [2] , Kai Wang [1] , Yan Cui [2*]

1. College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830054, People's Republic of China

2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People's Republic of China

Abstract:

Yanling Zheng：31372687@qq.com

XiJang Wang: 79057023@qq.com

Kai Wang: zhub168_math@sina.com,

Xueliang Zhang :* Corresponding author, zhengyl_math@sina.cn

Yan Cui :* Corresponding author, cuiyan_jk@sina.com

Objective: Kashgar has the highest tuberculosis (TB) incidence in China. The task of TB prevention and control in Kashgar is very serious. However, there is little quantitative prediction study on the TB incidence in this area, therefore, it is urgent to accurately predict the TB incidence in Kashgar, which can provide scientific reference for the prevention and control of TB.

Methods: Based on the TB incidence in Kashgar, a single Box-Jenkins model and Box-Jenkins model combined with Elman neural network model were used to do prediction analysis of TB incidence in Kashgar. Root mean squared error (RMSE), mean absolute error (MAE) was used to measure the prediction accuracy.

Results: The single AR((1, 2, 8)) model was established, the AIC, SC and R^2 of the model were 6.53, 6.61 and 0.47 respectively. In order to improve the prediction accuracy of AR ((1, 2, 8)) model, we used Elman neural network to further extract the nonlinear information of AR ((1, 2, 8)) model, then, the AR ((1, 2, 8)) combined Elman neural network model（AR-Elman） was established, its R^2 was 0.83. Based on the AR-Elman model, the TB incidence were fitted and predicted, the fitting RMSE and MAE were 3.78 and 3.38 respectively, the predicting RMSE and MAE were 8.86 and 7.29 respectively.

Conclusions: Both the single AR((1,2,8)) model and AR-Elman model can be used to predict the TB incidence in Kashgar, but the fitting and predicting RMSE and MAE by AR-Elman model were smaller than that by AR((1,2,8)) model, and the R^2 of AR-Elman model was higher than that of AR((1,2,8)) model, which indicated that AR-Elman model was better than AR((1,2,8)) model. The AR-Elman model is used to predict the future TB incidence in Kashgar, which can provide help for the planning of health resources in this area in advance.

Keywords: Tuberculosis; Prediction; Box-Jenkins method; Elman neural network

Strengths and limitations of this study:

1. Kashgar has the highest tuberculosis (TB) incidence in China. However, there is little quantitative prediction study on the TB incidence in this area, therefore, it is urgent to do the prediction analysis of TB in Kashgar.

2. In the study, the feasibility of predicting TB incidence in Guangxi by Box-Jenkins model and Box-Jenkins model combined with Elman neural network model were discussed respectively, finally, AR-Elman model with high prediction accuracy was established, it can be used to predict the development of TB so as to prevent its outbreak and provide help for the planning of health resources in advance in Kashgar.
3. This study did not consider possible influencing factors related to TB, such as Climatic factors, environmental factors, demographic factors and political issues,etc.

Introduction

Tuberculosis (TB) is still a major global public health problem and the ninth leading cause of death in the world. [1-3] Because TB is contagious, patient resistance is low, treatment time is long, patients' labor force is lost and their contribution to society is reduced, so TB brings great burden to society and economy. All countries in the world are working hard to fight tuberculosis. In 2018, the number of new reported TB cases was about 10 million; this figure has remained relatively stable in recent years. The latest treatment results showed that the global TB treatment success rate was 83%.WHO has set targets for the stop TB strategy. The targets mentioned that by 2030, on the basis of the work in 2015, TB deaths will reduce by 90%, and TB incidences will reduce (annual new cases) by 80%.[4] In order to achieve these goals, TB prevention and control services must be provided in the broad context of universal health coverage, joint action must be taken to address the social and economic consequences of TB, and technological breakthroughs should be achieved by 2025 to make the TB incidence decline faster than that at any time in history. According to the latest WHO report, China ranks third in the world after India and Indonesia in the number of new TB cases. [4] The TB incidence in western China was much higher than that in eastern and central China. The province with the highest TB incidence in the west is Xinjiang province. From 2016 to 2017, the annual TB incidence in Xinjiang was 185.66/100,000 and 202.58/100,000, while the annual TB incidence from 2016 to 2017 in China was 61/100,000 and 60.53/100,000, respectively, it is nearly three times higher in Xinjiang than that in China.

There are 14 Prefectural-Level cities in Xinjiang, China, among which Kashgar has the highest TB incidence rate. From 2016 to 2017, the annual TB incidences in Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7 times higher than that of the national level. Doing a good job in the prevention and control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.

Mastering the changing law of the incidence of infectious diseases, using the existing surveillance data to analyze, then, to predict the possible epidemic trend and provide reference data for the prevention, can better control the occurrence and epidemic of infectious diseases. Prediction of infectious diseases is to predict the occurrence, development and epidemic trend of infectious diseases according to the occurrence, development law and related factors of infectious diseases, combined with analysis and judgment and mathematical model, etc. At present, the popular analysis and prediction methods include time series analysis, neural network prediction method and so on. Time series Box-Jenkins method has been widely used in the prediction and analysis of infectious diseases in recent years [5-16], it can control the long-term trend, seasonality, periodicity and other factors. Neural network has strong nonlinear mapping ability, in which Elman neural network is composed of input layer, hidden layer, connection layer and output layer. The network has dynamic memory function and is very suitable for time series prediction. At present, the network is widely used in various fields, and has achieved successful prediction results. [17-21]

The TB incidence in Kashgar, Xinjiang is almost the highest in China, and it is urgent to do a good job in the prevention and control of TB in this area. Accurate prediction of TB incidence is a prerequisite for prevention and control. In this study, the popular Box-Jenkins time series method was used to found model for predicting the TB incidences in Kashgar, in order to improve the prediction accuracy, the Elman neural network with strong nonlinear information capture ability was used to construct the combined model for prediction analysis.

Materials and Methods

Study area and data Sources

We selected Kashgar as the study site (see Figure 1). This city is located in the south of Xinjiang province in China with an area of approximately 16.2 thousand square kilometers and a permanent population of 4.64 million in 2018. TB case data from January 2005 to December 2017 were obtained from the Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region. Population data were obtained from Xinjiang Bureau of Statistics. Based on the population data and TB case data, we calculated the incidence data of TB.

Patient and public involvement

Patients were not involved in the design of this study as it involved only observational analysis of an anonymised, pre- existing, routinely collected dataset.

Autoregressive moving average (ARMA) model [22]

ARMA model is an important time series analysis and prediction model in Box-Jenkins method, also known as auto-regression moving average model. The ARMA (p,q) model is a model with autocorrelation order p and moving average order q, p and q are judged by autocorrelation function (ACF) and partial autocorrelation function (PACF) diagram of stationary data. If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are the p-order truncated, then q=0, the ARMA(p,q) model is recorded as AR(p), and its expression is as follows:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

Where, X_t is the observed value at t, $\phi_1, \ldots, \phi_p$ are model parameters, C is a constant, if only ϕ_1, ϕ_2, ϕ_p are not zeros, then, The AR(p) model becomes a sparse model, which can be recorded as AR((1,2, p)).

If the original data is stable, the autocorrelation coefficients are q-order truncated and the partial correlation coefficients are trailing, then p=0 and the ARMA(p,q) model becomes MA(q), its expression is as follows:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Where, μ is the expected value of X_t. $\theta_1, \ldots, \theta_p$ are model parameters.

If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are also trailing, then the expression of the ARMA (p, q) model is as follows:

$$X_t = c + \phi_1 X_{t-1} + \phi_1 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

There are four main steps in ARMA modeling:

First step. The prerequisite for ARMA modeling is the stationary of time series. Check whether the data is stable by Augmented Dickey-Fuller (ADF) unit root test, in this study, the significant level probability (Prob) is 0.05, and if the Prob is less than 0.05, then, the data is stable. By observing the autocorrelation function (ACF) and partial function (PACF) of the stable data, we can determine the possible values of p and q and establish the possible ARMA (p, q) model.

Second step. The parameters of ARMA(p, q) model are estimated by maximum likelihood estimation or least square estimation, and the model parameters are tested. If Prob is less than 0.05, the parameters have statistical significance. The best model is determined according to the value of AIC, SC and R^2 of model. The smaller AIC and SC are, the larger the R square is, and the better the model is.

Third step. To determine whether the established ARMA (p, q) model is suitable. The residual sequence of a suitable model shall be the white noise process, and its ACF and PACF coefficients should be within twice the standard deviation range, otherwise, it is considered that the extraction information of the established model is not sufficient, and it is necessary to consider improving the accuracy of the model.

Fourth step. Using the established model for prediction and analysis

Elman neural network model[23]

The E1man neural network is proposed by E1man in 1990. The model is generally divided into four layers: input layer, hidden layer, receiving layer and output layer. The characteristic of E1man network is that the output of the hidden layer is connected to the input of the hidden layer through the delay and storage of the receiving layer, which makes it sensitive to the data of the historical state. The addition of the internal feedback network increases the ability of the network itself to deal with dynamic information, thus achieving the purpose of dynamic modeling.

The mathematical structure of the Elman neural network is as follows:

$$x(k) = f(w^1 x_c(k) + w^2 u(k-1))$$

$$x_c(k) = \alpha x_c(k-1) + x(k-1)$$

$$y(k) = g(w^3 x(k))$$

Where, w^1 is the connection weight matrix between the contact unit and the hidden layer unit, w^2 is the connection weight matrix between the input unit and the hidden layer unit, w^3 is the connection weight matrix between the hidden layer unit and the output unit. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, α is a self-connected feedback gain factor, $1 \le \alpha < 0$, $f(x)$ often takes the sigmoid function

There are four main steps in Elman neural network modeling:

First step. Data standardization processing.

Second step. Determine the input layer, the output layer.

Third step. To determine the number of neurons in the hidden layer, so that the error of the established Elman model is minimized. At present, there is no ideal analytical expression for the number of neurons in the hidden layer. The number of neurons has a great influence on the performance of the network. When the number of neurons is too large, it will lead to that the network learning time is too long, the generalization performances are not good, and even the convergence failure, but when the number of neurons is too small, the fault-tolerant ability of the network is poor. In general, the number of neurons does not exceed 20. [23] In this study, matlab cyclic structure was used to find the optimal number of neurons by comparing the RMSE values of Elman networks with neurons 1 to 20.

Fourth step. According to the optimal number of neurons in the hidden layer, the Elman model is constructed, and then the prediction and analysis are made.

Model comparison

Two performance indexes, root mean squared error (RMSE), mean absolute error (MAE) were used to assess the fitting and forecasting accuracy of two models. The smaller these two values are, the better the model is.

Statistical software

All data analyses were conducted using Eviews7, matlab2015b, Arcmap10.1.

Results

From January 2005 to December 2017, the number of reported TB cases is 141984 in Kashgar, Xinjiang, the average annual TB cases were 7888 and the average annual incidence was 191.18. Figure 2 shows the time series map of the TB incidences. It can be seen from the figure that the trend of the annual TB incidence is similar, and the TB incidence from 2014 to 2017 were significantly higher than that of previous years.

The data of TB incidences from January 2005 to December 2017 were divided into two parts. The data from January 2005 to December 2016 were used to found model and the data from January 2017 to December 2017 were used to test the model.

Establishment of ARMA Model

ARMA Modeling requires data stability, so, first of all, the stability of the modeling part of the data is verified by ADF test. The results of ADF test showed that Prob was less than 0.05(see Table 1), which indicated that the data was stable and could be directly used to found the model. Secondly, the ACF and PACF diagrams of the modeling part of the data were done (see Figure 3), It is obvious from the diagram that the autocorrelation coefficients are trailing distribution, and the partial correlation coefficients are almost a second-order truncated distribution, only at the lag of 7, 8 and 9, the correlation coefficients are a little large. Based on this situation, we considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and AR((1,2,9)), The least square method was used to test the parameters of the four models. The results of the test were shown in Table 2; we can see that, of the four models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing the two models, it was found that, the AR((1,2,8)) had smaller AIC and SC values, and the R^2 values of AR((1,2,8))were larger than the R^2 values of AR (2), so the AR (2) model was abandoned. The residual analysis of the AR((1,2,8)) model was carried out. The autocorrelation and partial correlation coefficient of the residuals were almost all within two times standard deviation, and only in the lag 5,6,12 order, they were beyond the range of two times standard deviation (see Figure 4), which indicated that the AR((1,2,8)) model could be used to roughly predict the TB incidences in Kashgar. We used AR((1,2,8)) model to fit the TB incidence from September 2005 to December 2016, the fitting RMSE and MAE were 6.15 and 4.33, respectively; we used AR((1,2,8)) model to predict the TB incidence from January 2017 to December 2017, the prediction RMSE and MAE were 10.88 and 8.75 respectively.

Establishment of AR-Elman Model

In order to improve the prediction accuracy of the AR((1,2,8)) model, we tried to establish the AR-Elman model. The fitting sequence of AR((1,2,8)) model was used as input variable, and the actual TB incidence was used as output variable. Due to the similarity of the annual trend of TB incidence in Kashgar (see Figure 1), therefore, we created twelve time-lagged variables as input features. Supposing that x_t represented the TB incidence at time t, and then the input matrix and the output matrix of the modeling data set used in this study were written as follows:

$$input\ matrix=\begin{bmatrix} x_1 & x_2 & \cdots & x_{12} \\ x_2 & x_3 & \cdots & x_{13} \\ \vdots & \vdots & \vdots & x_{14} \\ x_{t-12} & x_{t-11} & \cdots & x_{t-1} \end{bmatrix},\ output\ matrix=\begin{bmatrix} x_{13} \\ x_{14} \\ \vdots \\ x_t \end{bmatrix}$$

We selected twelve as the number of input layers of Elman and one as the number of output layers representing the forecast value. Through the matlab cyclic structure, we selected the optimal number of neurons between 1 and 20, and finally determined that the number of neurons was 6 (see Figure 5), the RMSE was the smallest, and the AR-Elman was optimal. We used the AR-Elman model to fit the training data, the RMSE was 3.78, the MAE was 3.38, and the R^2 of the model was 0.83; we used the AR-Elman model to predict the TB incidence from January 2017 to December 2017, the RMSE was 8.86, and the MAE was 7.29. The fitting diagram of AR((1,2,8)) model and AR-Elman model was shown in Figure 6.

Discussion

According to the WHO 2019 Global Tuberculosis report, around the world, TB mortality was down about 3% every year, the incidence was down about 2% every year, 16% of TB patients died of the disease. [4] But the rate of decline has not reached the pace of the stop Tuberculosis Strategy plan. Therefore, it is necessary to strengthen the prevention and control of tuberculosis. In order to significantly narrow these gaps, greater progress must be made in a group of countries with a high burden of tuberculosis. The burden of TB in China ranks second in the world, and Xinjiang is the province with the highest incidence of TB in China, and Kashgar is the area with the highest TB incidences in Xinjiang. Therefore, it is urgent to do a good job in the prevention and control of TB in Kashgar.

The prediction and early warning of infectious diseases is an important link in the prevention and control of infectious diseases. [24-26] Therefore, this study carried out research from the point of view of prediction to explore an accurate prediction model and do prediction and analysis of TB incidence in Kashgar, so as to provide scientific reference for the prevention and control of the disease in this area. The Box - Jenkins method is a popular time series prediction method, this method has good prediction performance and high prediction accuracy; Elman Neural network can capture nonlinear information of time series data very well. In this study, the two methods were combined to study the prediction model of TB incidence in Kashgar. First of all, a single prediction model was established by Box-Jenkins method, which requires the data to be stable. The probability value tested by ADF unit root was less than 0.05, which indicated that the data was stable and could be directly used for modeling. Based on the ACF and PACF diagrams of stationary data, we established four models, then we tested the parameters of the four models by means of the least square method and calculated the AIC and SC values of these models, finally, we found that all parameters of AR((1,2,8)) model passed the test, and the AIC and SC of AR((1,2,8)) model were the smallest, so the AR((1,2,8)) model was the best model, its residual errors were basically within twice the standard deviation, which indicated that the AR((1,2,8)) model was feasible. We used this model to fit the TB incidence, the fitting RMSE was 6.15, the fitting MAE was 4.33, we used the model to predict the TB incidence in Kashgar from January 2017 to December 2017, the predicting RMSE was 10.88, the predicting MAE was 8.75. In order to improve the prediction accuracy of the model, we tried to establish the AR-Elman combination model. The TB incidences fitted by AR((1,2,8)) were used as the input data of Elman network, and the actual incidences were used as the output data of Elman network. The Matlab cyclic structure was used to find the optimal neuron from 1 to 20, when the number of neurons was 6, the fitting error of the Elman neural network was the smallest; therefore, the model was the best, we called the model AR-Elman combination model.

We used the AR-Elman model to fit the TB incidence, the fitting RMSE was 3.78, the fitting MAE was 3.38, we used the AR-Elman model to predict the TB incidence in Kashgar from January 2017 to December 2017, the predicting RMSE was 8.86, the predicting MAE was 7.29. Both the fitting RMSE and MAE and the predicting RMSE and MAE of the AR-Elman model were smaller than those of the single AR((1,2,8)) model, which indicated that the combined model established in this study was more suitable for predicting the TB incidence in Kashgar.

Conclusions

The TB incidence in Kashgar, Xinjiang is almost the highest in China, in order to provide some help for the prevention and control of this disease in this area, in this paper, the prediction problem of the TB incidence was studied. Firstly, a single AR((1,2,8)) prediction model was established by using Box-Jenkins method, and its fitting and prediction performance were good, secondly, In order to improve the prediction accuracy of the single AR((1,2,8)) model, we combined single AR((1,2,8)) with Elman neural network with strong ability to capture nonlinear information to establish AR-Elman combination model. The fitting and prediction accuracy of the combined model was higher than that of the single AR((1,2,8)) model. The AR-Elman combined model can provide scientific help for predicting and warning the TB incidence in Kashgar, Xinjiang. This study also has some limitations, we did not consider possible influencing factors related to TB, such as climatic factors and political issues, etc, in further studies, we will consider these factors.

Competing interests

The authors declare no conflict of interest.

Funding

This work was supported by Major projects of science and technology in Xinjiang Autonomous region (2017A03006), China.

Author contributions

YLZ analysed the data and wrote the manuscript. XLZ, XJW, KW and YC wrote and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to express our gratitude to anonymous peer reviewers for carefully revising our manuscript and for his or her useful comments.

References

1.Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia: prevalence, risk factors and its public health importance. BMC Infectious Diseases, 2019, 19(1).

2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity pattern of Mycobacterium tuberculosis strains isolated from pulmonary tuberculosis patients in and around Ambo Town, Central Ethiopia. Plos One, 2018, 13(2):e0193083-.

3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control in Debre Markos Town, Northwest Ethiopia. Advances in Medicine, 2018, 2018(2):1-8.

4. WHO. Global tuberculosis report 2019, https://www.who.int/tb/publications/ global_report/en/. (Accessed on 17 Oct 2019).

5. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of tuberculosis cases reported at a tertiary hospital in Ghana: a time series model approach. BMC Public Health.2018, 18(1):1292.

6. Tohidinik H R , Mohebali M , Mansournia M A, et al. Forecasting zoonotic cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran: A SARIMA Analysis. Tropical Medicine & International Health, 2018, 23(8):860–869.

7. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in Ouagadougou, Burkina Faso, 2011–2015. Malaria Journal, 2018, 17(1):138.

8. Martínezbello D A, Lópezquílez A, Torresprieto A. Bayesian dynamic modeling of time series of dengue disease case counts. Plos Neglected Tropical Diseases, 2017, 11(7):e0005696.

9. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease outbreak and primary healthcare delivery in Liberia: Time-series analyses for 2010–2016. Plos Medicine, 2018, 15(2):e1002508.

10. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: a time-series model based on weather variability. International Journal of Biometeorology, 2017(6):1-10.

11. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. Cogent Social Sciences, 2018, 4(1):1461544.

12. Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving average model and generalised regression neural network model for prediction of haemorrhagic fever with renal syndrome in China: a time-series study. Bmj Open, 2019, 9(6):e025773.

13. Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in children younger than 48 months in Suzhou using a seasonal autoregressive integrated moving average model. Bmj Open, 2019, 9(1).

14. Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable incidence time series. Epidemiology & Infection, 2019, 147

15. Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand, foot and mouth disease cases in China from 2008 to 2018. Epidemiology & Infection, 2019, 147

16. Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan[J]. Bmj Open, 2017, 7(11):e018628

17. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum Analysis for wind speed forecasting based on Elman Neural Network. Energy Conversion & Management, 2017, 148:895-904.

18. Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for Forecasting China's Carbon Dioxide Emissions. Communications in Computer and Information Science, 2018(951): 36-47.

19. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural Network for Disease Prediction. Arabian Journal for Science and Engineering, 2019, 44(11): 9209–9220.

20. Mehrgini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman artificial neural network. Carbonates & Evaporites, 2017(6):1-11.

21. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. Neurocomputing, 2017, 228(C):154-164.

22. Box G E, Jenkins G M. Time series analysis: forecasting and control rev. ed. Oakland, California, Holden-Day, 1976, 31(4):238-242.

23. Ming Cheng.The principle and example of MATLAB neural network. Tsinghua University Press, 2013.

24. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clin Microbiol Infect. 2013, 19(11):999-1005

25.Wearing HJ1, Rohani P,Keeling MJ. Appropriate models for the management of infectious diseases. PLoS Med. 2005, 2(7):e174.

26. Neuberger A, Paul M, Nizar A, Raoult D. Modelling in infectious diseases: between haphazard and hazard. Clin Microbiol Infect. 2013,19(11):993-8.

Figures

Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China.

Figure 2. Graph of the TB incidences in Kashgar from January 2005 to December 2017

Figure 3. The ACF and PACF diagrams of modeling data

Figure 4. The ACF and PACF graphs of residual errors of AR((1,2,8)) Model

Figure 5. The numbers of neurons in AR-Elman Model and the corresponding RMSE Values

Figure 6. Fitting comparison Diagram of AR((1,2,8))Model and AR-Elman Model

Tables

Table 1. The ADF test of the training data

		t-Statistic	Prob-value
Augmented Dickey-Fuller test statistic		**-3.47**	**0.01**
Test critical values:	**1% level**	**-3.48**	
	5% level	**-2.88**	
	10% level	**-2.58**	

Table 2 Parameter estimates of the tentative models with their AIC and SC

Model	Variable	Coefficient	Std. Error	t-Statistic	Prob	AIC	SC
AR (2)	C	21.42	2.17	9.86	<0.01		
	AR(1)	0.42	0.08	5.20	<0.01	6.55	6.62
	AR(2)	0.34	0.08	4.17	<0.01		
AR((1,2,7))	C	22.93	3.73	6.14	<0.00		
	AR(1)	0.41	0.08	5.10	<0.00	2.87	3.00
	AR(2)	0.32	0.08	3.88	<0.00		
	AR (7)	0.12	0.007	1.74	0.08		

AR((1, 2, 8))	C	23.53	4.56	5.16	<0.01		
	AR(1)	0.40	0.08	4.84	<0.01	6.53	6.61
	AR(2)	0.32	0.08	3.96	<0.01		
	AR (8)	0.15	0.07	2.17	0.03		
AR((1, 2, 9))	C	29.07	15.53	1.87	0.06		
	AR(1)	0.37	0.08	4.64	<0.01	6.46	6.55
	AR (2)	0.31	0.08	3.93	<0.01		
	AR (9)	0.26	0.07	3.80	<0.01		

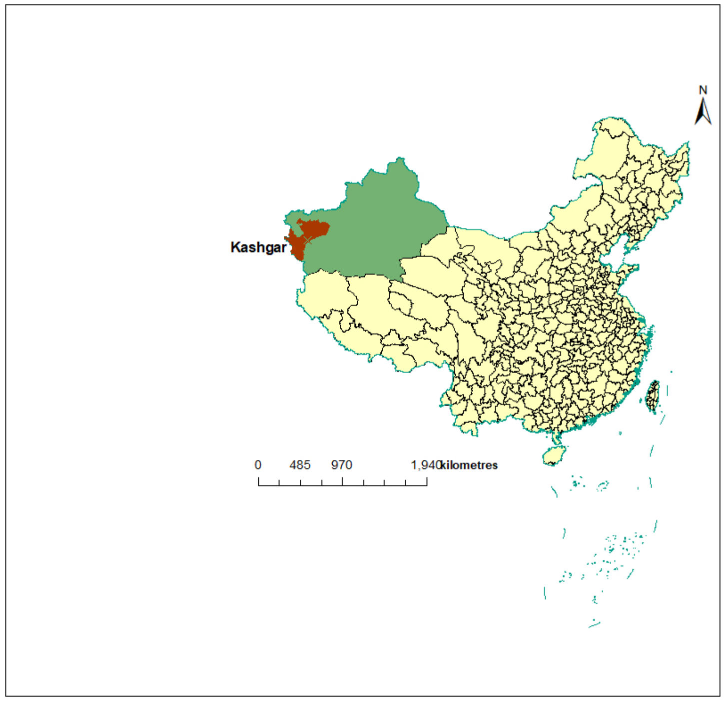

Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China.

296x210mm (96 x 96 DPI)

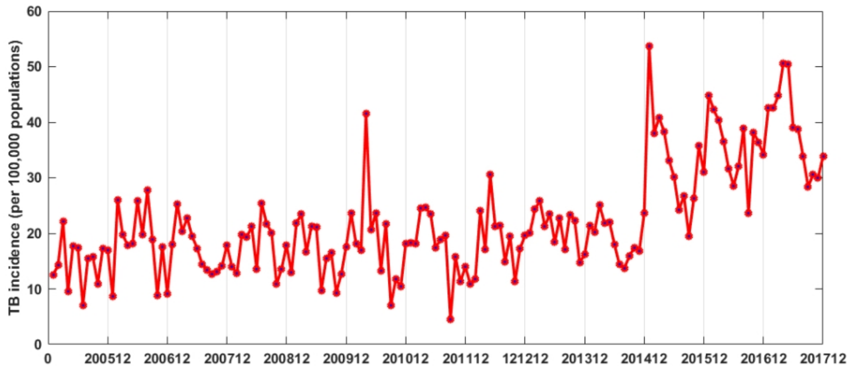

Figure 2. Graph of the TB incidences in Kashgar from January 2005 to December 2017

320x128mm (96 x 96 DPI)

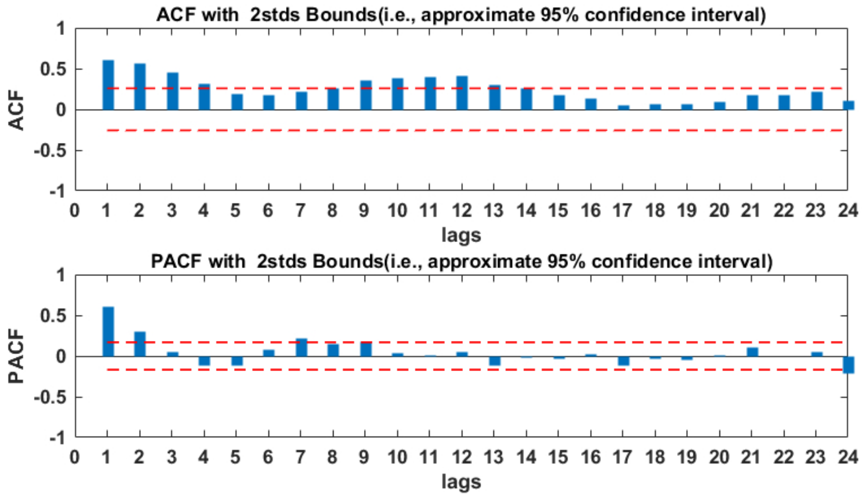

Figure 3. The ACF and PACF diagrams of modeling data

233x121mm (96 x 96 DPI)

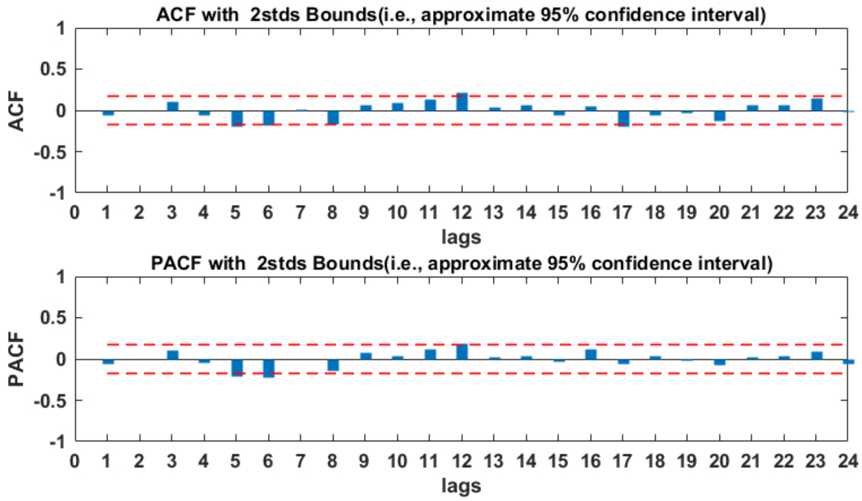

Figure 4. The ACF and PACF graphs of residual errors of AR ((1,2,8)) Model

233x122mm (96 x 96 DPI)

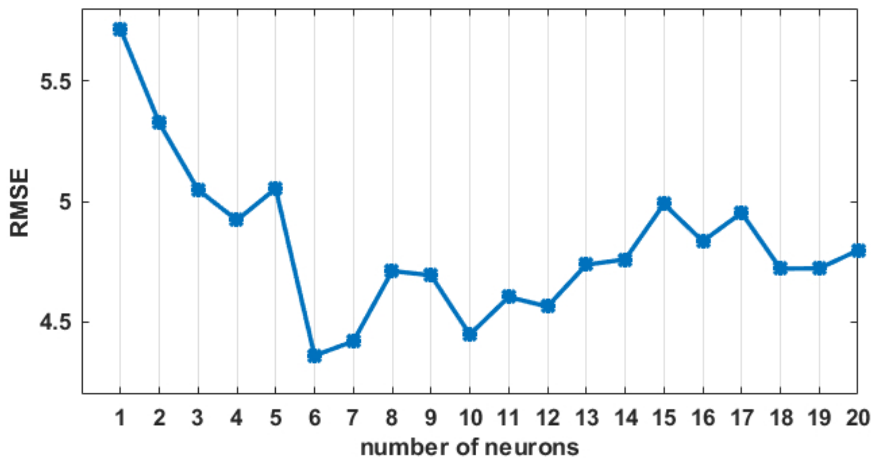

Figure 5. The numbers of neurons in AR-Elman Model and the corresponding RMSE Values

196x95mm (96 x 96 DPI)

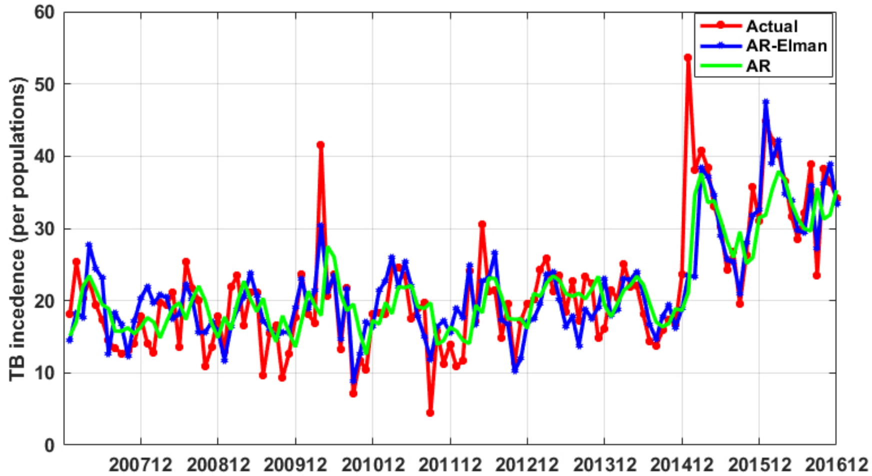

Figure 6. Fitting comparison Diagram of AR ((1,2,8)) Model and AR-Elman Model

233x123mm (96 x 96 DPI)

Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041040.R1
Article Type:	Original research
Date Submitted by the Author:	06-Oct-2020
Complete List of Authors:	Zheng, Yanling; Xinjiang Medical University, College of Medical Engineering and Technology Zhang, Xueliang; Xinjiang Medical University, College of Medical Engineering and Technology, Xinjiang Medical University Wang, Xijiang; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People's Republic of China Wang, Kai; Xinjiang Medical University, College of Medical Engineering and Technology,Urumqi 830054, People's Republic of China Cui, Yan; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,Urumqi 830001, People's Republic of China
Primary Subject Heading:	Health services research
Secondary Subject Heading:	Health services research, Infectious diseases
Keywords:	International health services < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Tuberculosis < INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS

Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China

Yanling Zheng [1] , XueLiang Zhang[1*], XiJang Wang[2] , Kai Wang[1] , Yan Cui[2*]

1. College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830054, People's Republic of China

2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People's Republic of China

Xueliang Zhang :* Corresponding author, zhengyl_math@sina.cn

Yan Cui :* Corresponding author, cuiyan_jk@sina.com

Abstract:

Objective: The incidence of tuberculosis (TB) in Kashgar, China, is very high, so the prevention and control of TB is arduous. However, there is little quantitative prediction study on the TB incidence in Kashgar, therefore, it is urgent to do prediction analysis of TB incidence in Kashgar, which can provide scientific reference for the prevention and control of TB.

Methods: Based on the monthly TB incidence in Kashgar, a single Box-Jenkins method and a Box-Jenkins and Elman neural network (ElmanNN) hybrid method were used to do prediction analysis of TB incidence in Kashgar. Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to measure the prediction accuracy.

Results: After careful analysis, the single AR((1,2,8)) model and AR((1,2,8))-ElmanNN (AR-Elman) hybrid model were established, and the optimal neurons value of the AR-Elman hybrid model was 6. For the fitting dataset, the RMSE, MAE and MAPE were 6.15, 4.33 and 0.2858, respectively, for the AR((1,2,8)) model, and 3.78, 3.38 and 0.1837, respectively, for the AR-Elman hybrid model. For the forecasting dataset, the RMSE, MAE and MAPE were 10.88, 8.75 and 0.2029, respectively, for the AR((1,2,8)) model, and 8.86, 7.29 and 0.2006, respectively, for the AR-Elman hybrid model.

Conclusions: Both the single AR((1,2,8)) model and AR-Elman model could be used to predict the TB incidence in Kashgar, but the modeling and validation scale-dependent measures (RMSE,MAE and MAPE) in the AR((1,2,8)) model were inferior to those in the AR-Elman hybrid model, which indicated that AR-Elman hybrid model was better than AR((1,2,8)) model. The AR-Elman hybrid model can be highlighted in predicting the temporal trends of TB incidence in Kashgar, which may act as the potential for far-reaching implications for prevention and control of TB.

Keywords: Tuberculosis; Prediction; Box-Jenkins method; Elman neural network

Strengths and limitations of this study:

1. The incidence of tuberculosis (TB) is very high in Kashgar, therefore, it is urgent to do the prediction analysis of TB in Kashgar.
2. In the study, AR-Elman model with high prediction accuracy was established, it can be used to predict the development of TB in Kashgar, China.
3. This study did not consider possible influencing factors related to TB, such as climatic factors, environmental factors, demographic factors and political issues,etc.

Introduction

Tuberculosis (TB) is still a major global public health problem and the ninth leading cause of death in the world. [1-3] Because TB is contagious, patient resistance is low, treatment time is long, patients' labor force is lost and their contribution to society is reduced, so TB brings great burden to society and economy. All countries in the world are working hard to fight tuberculosis. In 2018, the number of new reported TB cases was about 10 million; this figure has remained relatively stable in recent years. The latest treatment results showed that the global TB treatment success rate was 83%. World Health Organization (WHO) has set targets for the stop TB strategy. The targets mentioned that by 2030, on the basis of the work in 2015, TB deaths will reduce by 90%, and annual new TB cases will reduce by 80%.[4] In order to achieve these goals, TB prevention and control services must be provided in the broad context of universal health coverage, joint action must be taken to address the social and economic consequences of TB, and technological breakthroughs should be achieved by 2025 to make the TB incidence decline faster than that at any time in history. According to the global tuberculosis report 2019 [4], China has the second highest number of TB cases in the world. [4] The TB incidence in western China was much higher than that in eastern and central China. The province with the highest TB incidence in the west is Xinjiang province. From 2016 to 2017, the annual TB incidence in Xinjiang was 185.66/100,000 and 202.58/100,000, while the annual TB incidence from 2016 to 2017 in China was 61/100,000 and 60.53/100,000, respectively, it is nearly three times higher in Xinjiang than that national level.

There are 14 Prefectural-Level cities in Xinjiang, China, among which Kashgar has a very high TB incidence rate. From 2016 to 2017, the annual TB incidences in Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7 times higher than that of the national level. Doing a good job in the prevention and control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.

Mastering the changing law of the incidence of infectious diseases, using the existing surveillance data to analyze, then, to predict the possible epidemic trend and provide reference data for the prevention, can better control the occurrence and epidemic of infectious diseases. Prediction of infectious diseases is to predict the occurrence, development and epidemic trend of infectious diseases according to the occurrence, development law and related factors of infectious diseases, based on analysis, judgment and mathematical model, etc.

For study of quantitative prediction of infectious diseases, there are many methods, such as grey prediction method [5], exponential smoothing prediction method [6,7], dynamic model prediction method [8], Box-Jenkins method [9], neural network method [10], etc., with the deepening of prediction research, more and more scholars like to use the Box-Jenkins method [11-21], there are many different models in this method, and if appropriate models are established according to the characteristics of time series, high prediction ability often can be obtained. Neural network has strong nonlinear mapping ability, in which Elman neural network is composed of input layer, hidden layer, connection layer and output layer. The Elman network has dynamic memory function, and it is very suitable for time series prediction. At present, the Elman network is widely used in various fields, and has achieved successful prediction results. [22-26] Sometimes, the prediction effect of a single model is not ideal, in order to further improve the prediction accuracy, many studies adopt the combined model prediction method [27-29], the combined model can absorb the advantages of two or more methods so as to achieve a higher prediction accuracy.

The TB incidence of Kashgar is very high, and it is urgent to do a good job in the prevention and control of TB in this area. Accurate prediction of TB incidence is a prerequisite for prevention and control, which can help advance resource planning and policy formulation. In this study, the popular Box-Jenkins time series method was used to build model for predicting the TB incidence in Kashgar, in order to improve the prediction accuracy, the Elman neural network with strong nonlinear information capture ability was used to construct the combined model for prediction analysis.

Materials and Methods

Study area and data Sources

We selected Kashgar as the study site (see Figure 1). This area is located in the south of Xinjiang province in China with an area of approximately 16.2 thousand square kilometers and a permanent population of 4.64 million in 2018. TB case data from January 2005 to December 2017 were obtained from the Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region. Population data were obtained from Xinjiang Bureau of Statistics. Based on the population data and TB case data, we calculated the incidence data of TB.

Patient and public involvement

Patients were not involved in the design of this study as it involved only observational analysis of an anonymised, pre- existing, routinely collected dataset.

Autoregressive moving average (ARMA) model

ARMA model [30] is an important time series analysis and prediction model in Box-Jenkins method, also known as auto-regression moving average model. The ARMA (p,q) model is a model with autocorrelation order p and moving average order q, p and q are judged by autocorrelation function (ACF) and partial autocorrelation function (PACF) diagram of stationary data. If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are the p-order truncated, then q=0, the ARMA(p,q) model is recorded as AR(p), and its expression is as follows:

$$X_t=c+\phi_1X_{t-1}+\phi_2X_{t-2}+\cdots+\phi_pX_{t-p}+\varepsilon_t,$$

where, X_t is the observed value at t, $\phi_1,\ldots,\phi_p$ are model parameters, c is a constant, if only ϕ_1, ϕ_2, ϕ_p are not zeros, then, The AR(p) model becomes a sparse model, which can be recorded as AR((1,2,p)).

If the original data is stable, the autocorrelation coefficients are q-order truncated and the partial correlation coefficients are trailing, then p=0 and the ARMA(p,q) model becomes MA(q), its expression is as follows:

$$X_t=\mu+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q},$$

where, μ is the expected value of X_t. $\theta_1,\ldots,\theta_p$ are model parameters.

If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are also trailing, then the expression of the ARMA (p, q) model is as follows:

$$X_t=c+\phi_1X_{t-1}+\phi_2X_{t-2}+\cdots+\phi_pX_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q}.$$

There are four main steps in ARMA modeling:

First step. The prerequisite for ARMA modeling is the stationary of time series. Check whether the data is stable by Augmented Dickey-Fuller (ADF) unit root test. In this study, the significant level probability (p-value) is 0.05, and if the p-value is less than 0.05, then, the data is stable. By observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the stable data, we can determine the possible values of p and q and establish the possible ARMA (p, q) model.

Second step. The parameters of ARMA(p, q) model are estimated by maximum likelihood estimation or least square estimation, and the model parameters are tested. If p-value is less than 0.05, the parameters have statistical significance. The best model is determined according to the value of Akaike information criterion (AIC), Schwarz criterion (SC) and Goodness of Fit (R^2) of model. The smaller AIC and SC are, the larger the R^2 is, and the better the model is.

Third step. To determine whether the established ARMA (p, q) model is suitable. The residual sequence of a suitable model shall be the white noise process, and its ACF and PACF coefficients should be within twice the standard deviation range, otherwise, it is considered that the extraction information of the established model is not sufficient, and it is necessary to consider improving the accuracy of the model.

Fourth step. Using the established model to do prediction and analysis.

Elman neural network model

The E1man neural network is proposed by E1man in 1990 [31]. The model is generally divided into four layers: input layer, hidden layer, receiving layer and output layer. The characteristic of E1man network is that the output of the hidden layer is connected to the input of the hidden layer through the delay and storage of the receiving layer, which makes it sensitive to the data of the historical state. The addition of the internal feedback network increases the ability of the network itself to deal with dynamic information, thus achieving the purpose of dynamic modeling.

The mathematical structure of the Elman neural network is as follows:

$$x(k)=f(w^1 x_c(k)+ w^2 u(k\text{-}1))$$

$$x_c(k)= \alpha\, x_c(k\text{-}1)+ x(k\text{-}1)$$

$$y(k)=g(w^3 x(k))$$

Where, w^1 is the connection weight matrix between the contact unit and the hidden layer unit, w^2 is the connection weight matrix between the input unit and the hidden layer unit, w^3 is the connection weight matrix between the hidden layer unit and the output unit. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, α is a self-connected feedback gain fac**tor,** $0 \leqslant \alpha < 1$**,** $f(x)$ often takes the sigmoid function

There are four main steps in Elman neural network modeling:

First step. Data standardization processing. Data standardization is scaling the data to a small specific interval. In order to remove the unit limit of the data and convert it into dimensionless pure value, it is convenient for the index of different units or order of magnitude to be compared and weighted. In our study, we use function package mapminmax() to standardize the data, standardized data is in [-1,1] range.

Second step. Determine the input layer, the output layer. Generally, the input and output layer are determined according to the characteristics of the data and the needs of the analysis.

Third step. Set the parameters of the Elman model, such as training epochs and goals, etc., in our study, training epochs and goals of Elman neural network were set to 2000 and 0.00001, respectively. To determine the number of neurons in the hidden layer, so that the error of the established Elman model is minimized. At present, there is no ideal analytical expression for the number of neurons in the hidden layer. The number of neurons has a great influence on the performance of the network. When the number of neurons is too large, it will lead to that the network learning time is too long, the generalization performances are not good, and even the convergence failure, but when the number of neurons is too small, the fault-tolerant ability of the network is poor. In general, the number of neurons does not exceed 20. [32] In this study, matlab cyclic structure was used to find the optimal number of neurons by comparing the RMSE values of Elman networks with neurons 1 to 20.

Fourth step. According to the optimal number of neurons in the hidden layer, the Elman model is constructed, and then the prediction and analysis can be made.

Model comparison measures

Three performance indexes, root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to assess the fitting and forecasting accuracy of two models. The smaller these three values are, the better the model is. Their expressions are as follows [10]:

$$RMSE=\sqrt{\frac{\sum_{t=1}^{n}(X_t - \hat{X}_t)^2}{n}},$$

$$MAE=\frac{\sum_{t=1}^{n}|X_t - \hat{X}_t|}{n},$$

$$MAPE=\frac{\sum_{t=1}^{n}\left|\frac{X_t - \hat{X}_t}{X_t}\right| \times 100}{n},$$

where $\hat{X}_t$ is the simulating and forecasting values, X_t is the actual values and n is the number of observations.

Statistical software

All data analyses were conducted using Eviews7, matlab2015b, Arcmap10.1.

Results

From January 2005 to December 2017, the number of reported TB cases was 141984 in Kashgar, Xinjiang, the average annual TB cases were 7888 and the average annual incidence was 191.18. Figure 2 showed the time series graph of the TB incidence. It can be seen from the Figure 2 that the curve of TB incidence showed strong nonlinear characteristics from 2005 to 2014, and the TB incidence from 2015 to 2017 were significantly higher than that of previous years.

The data of TB incidence from January 2005 to December 2017 were divided into two parts. The data from January 2005 to December 2016 were used to build model and the data from January 2017 to December 2017 were used to test the model.

Establishment of ARMA Model

ARMA Modeling requires data stability, so, first of all, the stability of the modeling part of the data was verified by ADF test. The results of ADF test showed that p-value was less than 0.05(see Table 1), which indicated that the data was stable and could be directly used to build the model. Secondly, the ACF and PACF graphs of the modeling data were plotted (see Figure 3), it was obvious from Figure 3 that the autocorrelation coefficients were trailing distribution, and the partial correlation coefficients were almost a second-order truncated distribution, only at the lag 7, 8 and 9, the correlation coefficients were a little large. Based on this situation, we considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and AR((1,2,9)). The least square method was used to test the parameters of the four models. The results of the test were shown in Table 2; we can see that, of the four models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing the two models, it was found that, the AR((1,2,8)) model had smaller AIC and SC values, and the R^2 values of AR((1,2,8)) were larger than the R^2 values of AR (2), so the AR (2) model was abandoned. Then, the residual analysis of the AR((1,2,8)) model was carried out, the autocorrelation and partial correlation coefficient of the residuals were almost all within two times standard deviation, and only in the lag 5,6,12, they were beyond the range of two times standard deviation (see Figure 4), which indicated that the AR((1,2,8)) model could be used to roughly predict the TB incidence in Kashgar. We used AR((1,2,8)) model to fit the TB incidence from September 2005 to December 2016, the fitting RMSE ,MAE and MAPE were 6.15, 4.33 and 0.2858, respectively; we used AR((1,2,8)) model to predict the TB incidence from January 2017 to December 2017, the prediction RMSE, MAE and MAPE were 10.88 , 8.75 and 0.2029, respectively.

Establishment of AR-Elman Model

In order to improve the prediction accuracy of the AR((1,2,8)) model, we tried to establish the AR((1,2,8))-Elman hybrid model. The fitting sequence of AR((1,2,8)) model was used as input variable, and the actual TB incidence was used as output variable. Due to the a little similarity of the annual trend of TB incidence in Kashgar (see Figure 2), therefore, we created twelve time-lagged variables as input features. Supposing that x_t represented the TB incidence at time t, and then the input matrix and the output matrix of modeling data set used in this study were designed as follows (N=12):

$$input\ matrix=\begin{bmatrix} x_1 & x_2 & \dots & x_i \\ x_2 & x_3 & \dots & x_{i+1} \\ & \dots\dots & & \\ x_N & x_{N+1} & \dots & \dots \end{bmatrix},\ output\ matrix=[x_{N+1}\quad x_{N+2}\quad \dots\quad x_{N+i}]$$

We selected twelve as the number of input layers of AR-Elman network and one as the number of output layers representing the forecast value. By the matlab cyclic structure, we selected the optimal number of neurons between 1 and 20, and finally we found when the number of neurons was 6 (see Figure 5), the RMSE was the smallest, and the AR-Elman was optimal. We used the AR-Elman model to fit the training data, RMSE was 3.78, MAE was 3.38, MAPE was 0.1837, and the R^2 of the model was 0.83; we used the AR-Elman model to predict the TB incidence from January 2017 to December 2017, RMSE was 8.86, MAE was 7.29, and MAPE was 0.2006. The fitting graph of AR((1,2,8)) model and AR-Elman model was shown in Figure 6. Comparison results of the AR((1,2,8)) model and AR-Elman model was shown in Table 3, both the fitting RMSE, MAE and MAPE and the predicting RMSE , MAE and MAPE of the AR-Elman model were smaller than those of the single AR((1,2,8)) model, which indicated that the AR-Elman combined model established in this study was more suitable for predicting the TB incidence in Kashgar.

Discussion

According to the WHO 2019 Global Tuberculosis report [4], around the world, TB mortality was down about 3% every year, the incidence was down about 2% every year, 16% of TB patients died of the disease. [4] But the rate of decline has not reached the pace of the stop Tuberculosis Strategy Plan. Therefore, it is necessary to strengthen the prevention and control of tuberculosis. In order to significantly narrow these gaps, greater progress must be made in a group of countries with a high burden of tuberculosis. The burden of TB in China ranks second in the world, and Xinjiang is the province with the highest incidence of TB in China, and Kashgar is the area with the high TB incidence in Xinjiang. Therefore, it is urgent to do a good job in the prevention and control of TB in Kashgar.

The prediction and early warning of infectious diseases is an important link in the prevention and control of infectious diseases. [32-34] Therefore, this study carried out research from the point of view of prediction to explore an accurate prediction model and do prediction analysis of TB incidence in Kashgar, so as to provide scientific reference for the prevention and control of the disease in this area. The Box -Jenkins method is a popular time series prediction method, this method has good prediction performance and high prediction accuracy; Elman Neural network can capture nonlinear information of time series data very well. In this study, the two methods were combined to study the prediction model of TB incidence in Kashgar.

Many studies have found that Box-Jenkins method has a good ability of fitting and forecasting. For stationary time series that do not contain seasonality, it is more suitable to use the ARMA model of the Box-Jenkins method to do prediction analysis [35], for non-stationary time series of infectious diseases with obvious seasonality, it is more suitable to use seasonal autoregressive integrated moving average (SARIMA) model of the Box-Jenkins method for prediction analysis.[9-12] In our study, from Figure 2, we could see that the seasonality of the TB incidence in Kashgar from 2005 to 2014 was not obvious, there was only a certain seasonality from 2015 to 2017, and we found that the time series of TB incidence was stable by ADF unit root test, and the autocorrelation and partial correlation coefficients of modeling data at lag 12, 24 were not obviously large, therefore, for our research data, we used ARMA model to do forecast analysis, finally, we established AR((1,2,8)) model of the Box-Jenkins method, it has a good performance to fit and predict the TB incidence of Kashgar in Xinjiang. From Figure 2, we can also see that the time series of TB incidence has strong non-linear, the established AR((1,2,8)) model mainly extracted the linear information of data, considering that the neural network can capture the non-linear information of data well, in order to improve the prediction accuracy of TB incidence rate in Kashgar, we used AR((1,2,8)) model and Elman neural network model to establish AR-Elman hybrid model. Many studies have found that the combination model can improve the accuracy of prediction, such as, Wang et al.[28] found that SARIMA-NAR hybrid model has an outstanding ability to improve the prediction accuracy relative to SARIMA model and nonlinear autoregressive network (NAR) model when they were used to predict pertussis incidence in China. Li et al.[27] found ARIMA-GRNN hybrid model was shown to be superior to the single ARIMA model in predicting the short-term TB incidence in the Chinese population. Our research was consistent with these literatures that our AR-Elman hybrid model was more accurate than the single AR((1,2,8)) model.

The incidence of tuberculosis in Kashgar, Xinjiang is very high, and the relevant departments of disease prevention and control in Xinjiang have also done a lot of effective work. Our research was mainly to build a high-precision prediction model to help early warning of tuberculosis in Kashgar. Finally, we established the AR-Elman hybrid model, which had high fitting and prediction accuracy of TB incidence in Kashgar, Xinjiang. Because the development of any event may be affected by many factors, such as social, economic, political, demographic factors, so the long-term forecast accuracy of the AR-Elman hybrid model may decline.

Conclusions

Kashgar has a very high TB incidence, in order to provide some help for the prevention and control of this disease in this area, the prediction problem of the TB incidence was studied. Firstly, a single AR((1,2,8)) prediction model was established by using Box-Jenkins method, and its fitting and prediction performance were good, secondly, in order to improve the prediction accuracy of the single AR((1,2,8)) model, we used single AR((1,2,8)) and Elman neural network with strong ability to capture nonlinear information to establish AR-Elman hybrid model. The fitting and prediction accuracy of the hybrid model was higher than that of the single AR((1,2,8)) model. The AR-Elman hybrid model can provide scientific help for predicting and warning the TB incidence in Kashgar, Xinjiang. This study also has some limitations, we did not consider possible influencing factors related to TB, such as climatic factors and political issues, etc, in further studies, we will consider these factors.

Competing interests

The authors declare no conflict of interest.

Funding

This work was supported by Major projects of science and technology in Xinjiang Autonomous region (Grant No.2017A03006), China, and National Natural Science Foundation of China Regional Science Foundation Project (Grant No. 72064036, 82060609),and State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia Fund (Grant No.

SKL-HIDCA-2020-9)

Data sharing statement

No additional data are available.

Author contributions

YLZ analyzed the data and wrote the manuscript. XLZ, XJW, KW and YC wrote and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to express our gratitude to anonymous peer reviewers for carefully revising our manuscript and for his or her useful comments.

References

1.Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia: prevalence, risk factors and its public health importance. BMC Infectious Diseases, 2019; 19(1).

2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity pattern of Mycobacterium tuberculosis strains isolated from pulmonary tuberculosis patients in and around Ambo Town, Central Ethiopia. Plos One, 2018; 13(2):e0193083-.

3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control in Debre Markos Town, Northwest Ethiopia. Advances in Medicine, 2018;2018(2):1-8.

4. WHO. Global tuberculosis report 2019, https://www.who.int/tb/publications/ global_report/en/. (Accessed on 17 Oct 2019).

5. Zhao YF, Shou MH, Wang ZX. Prediction of the Number of Patients Infected with COVID-19 Based on Rolling Grey Verhulst Models. Int J Environ Res Public Health. 2020; 17(12):4582.

6. Guan P, Wu W, Huang D. Trends of reported human brucellosis cases in mainland China from 2007 to 2017: an exponential smoothing time series analysis. Environ Health Prev Med. 2018; 23(1):23.

7. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, Xu JY, Shi J, Sun DY. Analysis and predication of tuberculosis registration rates in Henan Province, China: an exponential smoothing model study. Infect Dis Poverty. 2020;9(1):123.

8. Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts. PLoS Negl Trop Dis. 2017;11(7):e0005696.

9. Wang Y, Xu C, Zhang S, Wang Z, Zhu Y, Yuan J. Temporal trends analysis of human brucellosis incidence in mainland China from 2004 to 2018. Sci Rep. 2018;8(1):15901.

10. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S. An Advanced Data-Driven Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series Forecasting in Qinghai Province, China. Infect Drug Resist. 2020;13:867-880.

11. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of tuberculosis cases reported at a tertiary hospital in Ghana: a time series model approach. BMC Public Health.2018;18(1):1292.

12. Tohidinik H R , Mohebali M , Mansournia M A, et al. Forecasting zoonotic cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran: A SARIMA Analysis. Tropical Medicine & International Health, 2018;23(8):860–869.

13. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in Ouagadougou, Burkina Faso, 2011–2015. Malaria Journal, 2018;17(1):138.

14. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease outbreak and primary healthcare delivery in Liberia: Time-series analyses for 2010–2016. Plos Medicine, 2018; 15(2):e1002508.

15. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: a time-series model based on weather variability. International Journal of Biometeorology, 2017;(6):1-10.

16. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. Cogent Social Sciences, 2018; 4(1):1461544.

17.Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving average model and generalised regression neural network model for prediction of haemorrhagic fever with renal syndrome in China: a time-series study. BMJ Open, 2019; 9(6):e025773.

18.Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in children younger than 48 months in Suzhou using a seasonal autoregressive integrated moving average model. BMJ Open, 2019; 9(1).

19.Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable incidence time series. Epidemiology & Infection, 2019; 147.

20.Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand, foot and mouth disease cases in China from 2008 to 2018. Epidemiology & Infection, 2019; 147.

21.Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. Bmj Open, 2017； 7(11):e018628.

22. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum Analysis for wind speed forecasting based on Elman Neural Network. Energy Conversion & Management, 2017; 148:895-904.

23.Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for Forecasting China's Carbon Dioxide Emissions. Communications in Computer and Information Science, 2018;(951): 36-47.

24. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural Network for Disease Prediction. Arabian Journal for Science and Engineering, 2019;44(11): 9209–9220.

25. Mehrgini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman artificial neural network. Carbonates & Evaporites, 2017;(6):1-11.

26. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. Neurocomputing, 2017; 228(C):154-164.

27. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, Ji Y, Xu D, Wang J. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infect Drug Resist. 2019;12:1011-1020.

28. Wang Y, Xu C, Wang Z, Zhang S, Zhu Y, Yuan J. Time series modeling of

pertussis incidence in China from 2004 to 2018 with a novel wavelet based SARIMA-NAR hybrid model. PLoS One. 2018;13(12):e0208404.

29. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, Yuan J. Temporal trends analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a new SARIMA-NARNNX hybrid model. BMJ Open. 2019;9(7):e024409.

30. Box G E, Jenkins G M. Time series analysis: forecasting and control rev. ed. Oakland, California, Holden-Day, 1976;31(4):238-242.

31. Ming Cheng.The principle and example of MATLAB neural network. Tsinghua University Press, 2013.

32. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clin Microbiol Infect. 2013;19(11):999-1005

33.Wearing HJ1, Rohani P,Keeling MJ. Appropriate models for the management of infectious diseases. PLoS Med. 2005; 2(7):e174.

34. Neuberger A, Paul M, Nizar A, Raoult D. Modelling in infectious diseases: between haphazard and hazard. Clin Microbiol Infect. 2013;19(11):993-8.

35. Tipirneni-Sajja A, Krafft AJ, Loeffler RB, Song R, Bahrami A, Hankins JS, Hillenbrand CM. Autoregressive moving average modeling for hepatic iron quantification in the presence of fat. J Magn Reson Imaging. 2019;50(5):1620-1632.

Figures

Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China. Kashgar is located in the south of Xinjiang, and it has a very high incidence of tuberculosis.

Figure 2. Graph of the tuberculosis (TB) incidence in Kashgar from January 2005 to December 2017. The curve of TB incidence showed strong nonlinear characteristics from 2005 to 2014, and the TB incidence increased significantly from 2015 to 2017.

Figure 3. Autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of modeling data. As the delay of the lag order, the autocorrelation coefficients were trailing and the partial correlation coefficients were truncated, so it was suitable to establish the AR model.

Figure 4. Autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of residuals of AR((1,2,8)) model. Autocorrelation coefficients and partial correlation coefficients were almost in 95% confidence interval, so AR ((1,2,8)) model could extract the information of original data well.

Figure 5. The numbers of neurons in AR-Elman model and the corresponding root mean square error (RMSE). When the number of neuron was 6, the RMSE was the smallest, and the AR-Elman model fitting ability was the strongest.

Figure 6. Fitting comparison graph of AR ((1,2,8)) model and AR-Elman model. Red line stands for the original tuberculosis (TB) incidence curve, green line stands for AR((1,2,8)) model fitting curve, blue line stands for AR-Elman model fitting curve. The fitting ability of AR-Elman hybrid model was slightly better than that of the single AR((1,2,8)).

Tables

Table 1. The Augmented Dickey-Fuller (ADF) test of the training data

		t-Statistics	p-value
Augmented Dickey-Fuller test statistic		-3.47	0.01
Test critical values:	1% level	-3.48	
	5% level	-2.88	
	10% level	-2.58	

Table 2. Parameter estimates of the tentative models with their Akaike information criterion (AIC) and Schwarz criterion (SC) values.

Models	Variables	Coefficients	Std. Errors	t	p-values	AIC	SC
AR (2)	C	21.42	2.17	9.86	<0.01	6.55	6.62
	AR(1)	0.42	0.08	5.20	<0.01		
	AR(2)	0.34	0.08	4.17	<0.01		
AR((1, 2, 7))	C	22.93	3.73	6.14	<0.00	6.53	6.62
	AR(1)	0.41	0.08	5.10	<0.00		
	AR(2)	0.32	0.08	3.88	<0.00		
	AR (7)	0.12	0.007	1.74	0.08		
AR((1, 2, 8))	C	23.53	4.56	5.16	<0.01	6.53	6.61
	AR(1)	0.40	0.08	4.84	<0.01		
	AR(2)	0.32	0.08	3.96	<0.01		
	AR (8)	0.15	0.07	2.17	0.03		
AR((1, 2, 9))	C	29.07	15.53	1.87	0.06	6.46	6.55
	AR(1)	0.37	0.08	4.64	<0.01		
	AR (2)	0.31	0.08	3.93	<0.01		
	AR (9)	0.26	0.07	3.80	<0.01		

Table 3. Comparison results of in-sample fitting and out-of-sample forecasting performance for the AR((1,2,8)) model and AR-Elman model.

Models	Fitted efficacy			Models	Forecasted efficacy		
	RMSE	MAE	MAPE		RMSE	MAE	MAPE
AR((1,2,8))	6.15	4.33	0.2585	AR((1,2,8))	10.88	8.75	0.2029
AR-Elman	3.78	3.38	0.1837	AR-Elman	8.86	7.29	0.2006

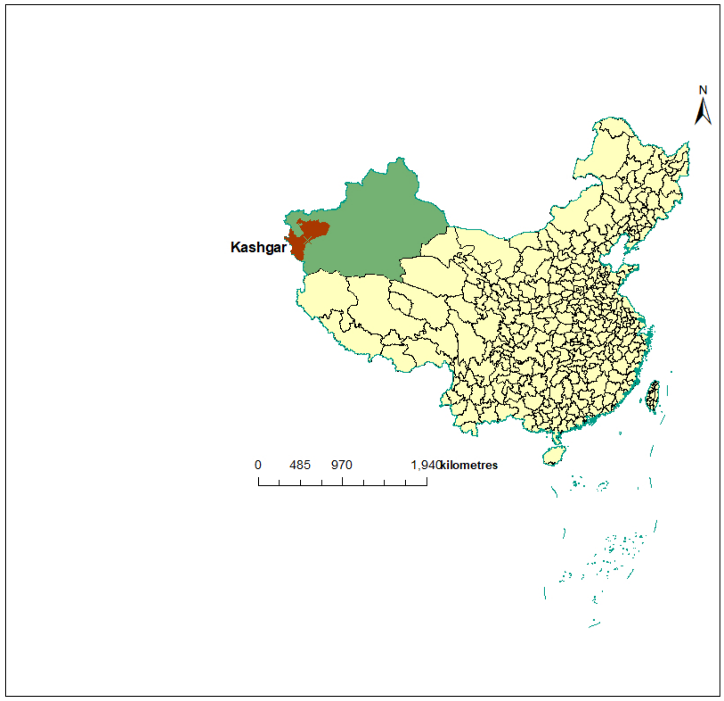

Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China.

296x210mm (96 x 96 DPI)

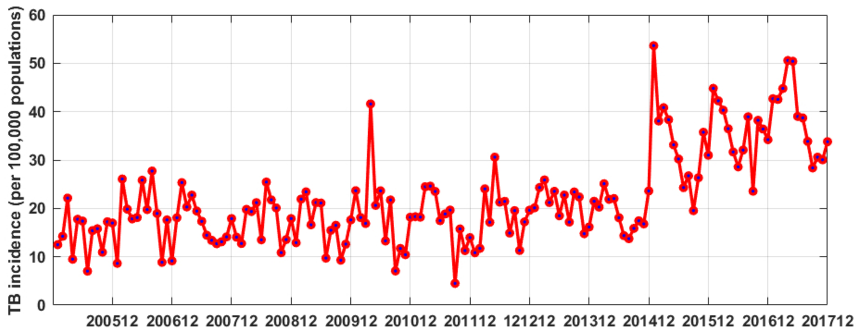

Figure 2

285x99mm (96 x 96 DPI)

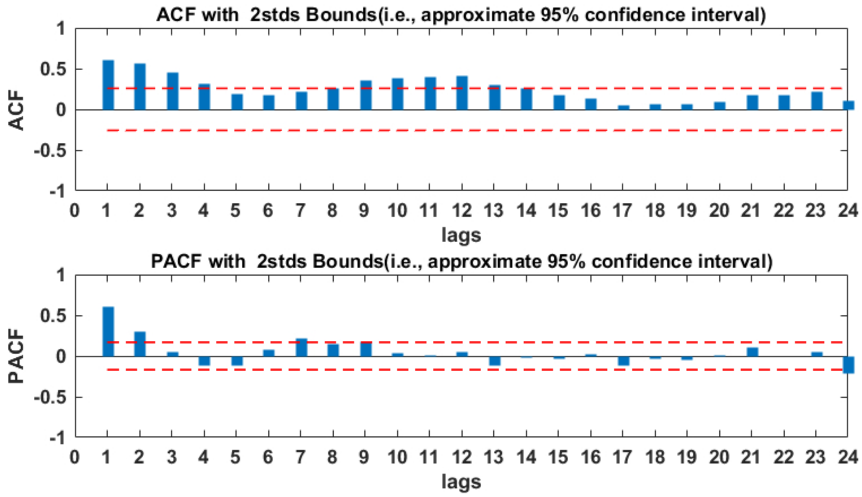

Figure 3. The ACF and PACF diagrams of modeling data

233x121mm (96 x 96 DPI)

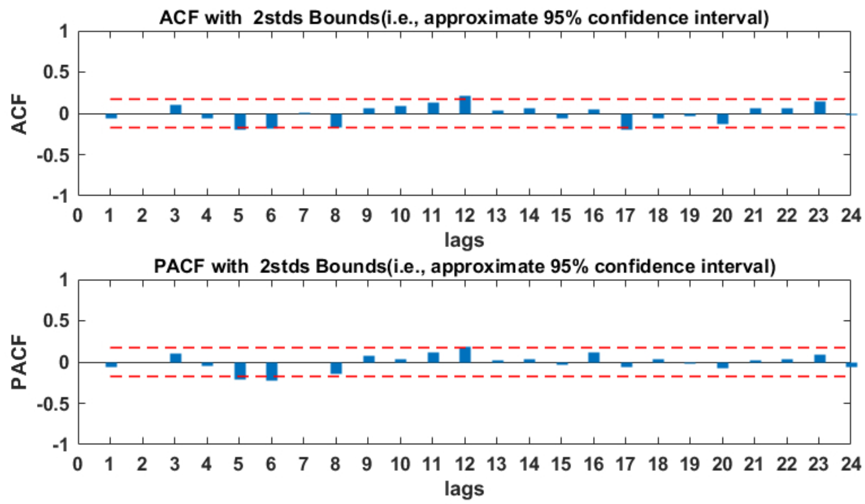

Figure 4. The ACF and PACF graphs of residual errors of AR ((1,2,8)) Model

233x122mm (96 x 96 DPI)

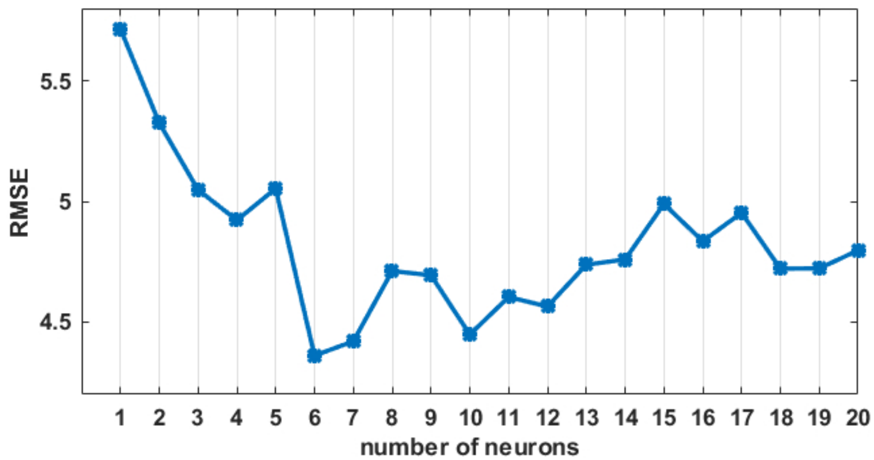

Figure 5. The numbers of neurons in AR-Elman Model and the corresponding RMSE Values

196x95mm (96 x 96 DPI)

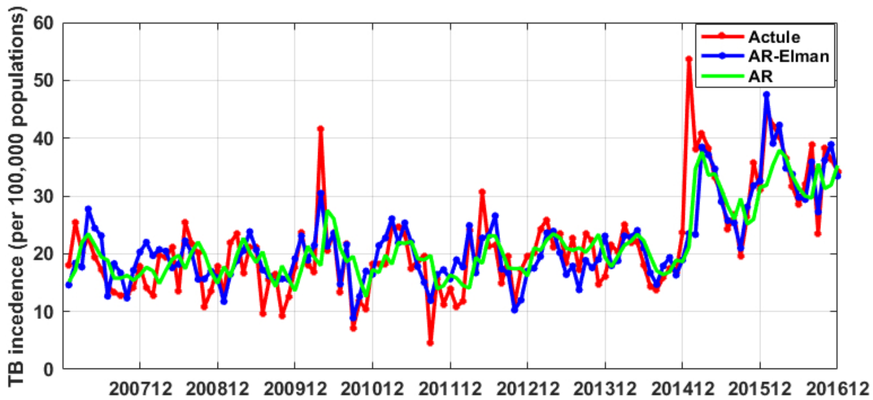

Figure 6

234x99mm (96 x 96 DPI)

Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041040.R2
Article Type:	Original research
Date Submitted by the Author:	18-Nov-2020
Complete List of Authors:	Zheng, Yanling; Xinjiang Medical University, College of Medical Engineering and Technology Zhang, Xueliang; Xinjiang Medical University, College of Medical Engineering and Technology, Xinjiang Medical University Wang, Xijiang; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People's Republic of China Wang, Kai; Xinjiang Medical University, College of Medical Engineering and Technology,Urumqi 830054, People's Republic of China Cui, Yan; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,Urumqi 830001, People's Republic of China
Primary Subject Heading:	Health services research
Secondary Subject Heading:	Health services research, Infectious diseases
Keywords:	International health services < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Tuberculosis < INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS

SCHOLARONE™
Manuscripts

Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China

Yanling Zheng [1] , XueLiang Zhang[1*], XiJang Wang[2] , Kai Wang[1] , Yan Cui[2*]

1. College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830054, People's Republic of China
2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People's Republic of China

Xueliang Zhang :* Corresponding author, zhengyl_math@sina.cn
Yan Cui :* Corresponding author, cuiyan_jk@sina.com

Abstract:

Objective: The incidence of tuberculosis (TB) in Kashgar, China, is very high, so the prevention and control of TB is arduous. However, there is little quantitative prediction study on the TB incidence, therefore, it is urgent to do prediction analysis of TB incidence in Kashgar, which can provide scientific reference for the prevention and control of TB.

Methods: Based on the monthly TB incidence, a single Box-Jenkins method and a Box-Jenkins and Elman neural network (ElmanNN) hybrid method were used to do prediction analysis of TB incidence in Kashgar. Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to measure the prediction accuracy.

Results: After careful analysis, the single AR((1,2,8)) model and AR((1,2,8))-ElmanNN (AR-Elman) hybrid model were established, and the optimal neurons value of the AR-Elman hybrid model was 6. For the fitting dataset, the RMSE, MAE and MAPE were 6.15, 4.33 and 0.2858, respectively, for the AR((1,2,8)) model, and 3.78, 3.38 and 0.1837, respectively, for the AR-Elman hybrid model. For the forecasting dataset, the RMSE, MAE and MAPE were 10.88, 8.75 and 0.2029, respectively, for the AR((1,2,8)) model, and 8.86, 7.29 and 0.2006, respectively, for the AR-Elman hybrid model.

Conclusions: Both the single AR((1,2,8)) model and AR-Elman model could be used to predict the TB incidence in Kashgar, but the modeling and validation scale-dependent measures (RMSE,MAE and MAPE) in the AR((1,2,8)) model were inferior to those in the AR-Elman hybrid model, which indicated that AR-Elman hybrid model was better than AR((1,2,8)) model. The Box-Jenkins and Elman neural network hybrid method can be highlighted in predicting the temporal trends of TB incidence in Kashgar, which may act as the potential for far-reaching implications for prevention and control of TB.

Keywords: Tuberculosis; Prediction; Box-Jenkins method; Elman neural network

Strengths and limitations of this study:

1. The incidence of tuberculosis (TB) is very high in Kashgar, therefore, it is urgent to do the prediction analysis of TB.
2. In the study, AR-Elman model with high prediction accuracy was established, it can be used to predict the development of TB in Kashgar, China.
3. The long-term prediction accuracy of AR-Elman hybrid model will decline.

Introduction

Tuberculosis (TB) is still a major global public health problem and the ninth leading cause of death in the world. [1-3] Because TB is contagious, patient resistance is low, treatment time is long, patients' labor force is lost and their contribution to society is reduced, so TB brings great burden to society and economy. All countries in the world are working hard to fight tuberculosis. In 2018, the number of new reported TB cases was about 10 million; this figure has remained relatively stable in recent years. The latest treatment results showed that the global TB treatment success rate was 83%. World Health Organization (WHO) has set targets for the stop TB strategy. The targets mentioned that by 2030, on the basis of the work in 2015, TB deaths will reduce by 90%, and annual new TB cases will reduce by 80%.[4] In order to achieve these goals, TB prevention and control services must be provided in the broad context of universal health coverage, joint action must be taken to address the social and economic consequences of TB, and technological breakthroughs should be achieved by 2025 to make the TB incidence decline faster than that at any time in history. According to the global tuberculosis report 2019 [4], China has the second highest number of TB cases in the world. [4] The TB incidence in western China was much higher than that in eastern and central China. The province with the highest TB incidence in the west is Xinjiang province. From 2016 to 2017, the annual TB incidence in Xinjiang was 185.66/100,000 and 202.58/100,000, while the annual TB incidence from 2016 to 2017 in China was 61/100,000 and 60.53/100,000, respectively, it is nearly three times higher in Xinjiang than that national level.

There are 14 Prefectural-Level cities in Xinjiang, China, among which Kashgar has a very high TB incidence rate. From 2016 to 2017, the annual TB incidences in

Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7 times higher than that of the national level. Doing a good job in the prevention and control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.

Mastering the changing law of the incidence of infectious diseases, using the existing surveillance data to analyze, then, to predict the possible epidemic trend and provide reference data for the prevention, can better control the occurrence and epidemic of infectious diseases. Prediction of infectious diseases is to predict the occurrence, development and epidemic trend of infectious diseases according to the occurrence, development law and related factors of infectious diseases, based on analysis, judgment and mathematical model, etc.

For study of quantitative prediction of infectious diseases, there are many methods, such as grey prediction method [5], exponential smoothing prediction method [6,7], dynamic model prediction method [8], Box-Jenkins method [9], neural network method [10], etc., with the deepening of prediction research, more and more scholars like to use the Box-Jenkins method [11-21], there are many different models in this method, and if appropriate models are established according to the characteristics of time series, high prediction ability often can be obtained. Neural network has strong nonlinear mapping ability, in which Elman neural network is composed of input layer, hidden layer, connection layer and output layer. The Elman network has dynamic memory function, and it is very suitable for time series prediction. At present, the Elman network is widely used in various fields, and has achieved successful prediction results. [22-26] Sometimes, the prediction effect of a single model is not ideal, in order to further improve the prediction accuracy, many studies adopt the combined model prediction method [27-29], the combined model can absorb the advantages of two or more methods so as to achieve a higher prediction accuracy.

The TB incidence of Kashgar is very high, and it is urgent to do a good job in the prevention and control of TB in this area. Accurate prediction of TB incidence is a prerequisite for prevention and control, which can help advance resource planning and policy formulation. In this study, the popular Box-Jenkins time series method was used to build model for predicting the TB incidence in Kashgar, in order to improve the prediction accuracy, the Elman neural network with strong nonlinear information capture ability was used to construct the combined model for prediction analysis.

Materials and Methods

Study area and data Sources

We selected Kashgar as the study site (see Figure 1). This area is located in the south of Xinjiang province in China with an area of approximately 16.2 thousand square kilometers and a permanent population of 4.64 million in 2018. TB case data from January 2005 to December 2017 were obtained from the Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region. Population data were obtained from Xinjiang Bureau of Statistics. Based on the population data and TB case data, we calculated the incidence data of TB.

Patient and public involvement

Patients were not involved in the design of this study as it involved only observational analysis of an anonymised, pre- existing, routinely collected dataset.

Autoregressive moving average (ARMA) model

ARMA model [30] is an important time series analysis and prediction model in Box-Jenkins method, also known as auto-regression moving average model. The ARMA (p,q) model is a model with autocorrelation order p and moving average order q, p and q are judged by autocorrelation function (ACF) and partial autocorrelation function (PACF) diagram of stationary data. If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are the p-order truncated, then q=0, the ARMA(p,q) model is recorded as AR(p), and its expression is as follows:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

where, X_t is the observed value at t, $\phi_1, \dots, \phi_p$ are model parameters, c is a constant, if only ϕ_1, ϕ_2, ϕ_p are not zeros, then, The AR(p) model becomes a sparse model, which can be recorded as AR((1,2,p)).

If the original data is stable, the autocorrelation coefficients are q-order truncated and the partial correlation coefficients are trailing, then p=0 and the ARMA(p,q) model becomes MA(q), its expression is as follows:

$$X_t=\mu+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q},$$

where, μ is the expected value of X_t. $\theta_1,\ldots,\theta_p$ are model parameters.

If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are also trailing, then the expression of the ARMA (p, q) model is as follows:

$$X_t=c+\phi_1X_{t-1}+\phi_2X_{t-2}+\cdots+\phi_pX_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q}.$$

There are four main steps in ARMA modeling:

First step. The prerequisite for ARMA modeling is the stationary of time series. Check whether the data is stable by Augmented Dickey-Fuller (ADF) unit root test. In this study, the significant level probability (p-value) is 0.05, and if the p-value is less than 0.05, then, the data is stable. By observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the stable data, we can determine the possible values of p and q and establish the possible ARMA (p, q) model.

Second step. The parameters of ARMA(p, q) model are estimated by maximum likelihood estimation or least square estimation, and the model parameters are tested. If p-value is less than 0.05, the parameters have statistical significance. The best model is determined according to the value of Akaike information criterion (AIC), Schwarz criterion (SC) and Goodness of Fit (R^2) of model. The smaller AIC and SC are, the larger the R^2 is, and the better the model is.

Third step. To determine whether the established ARMA (p, q) model is suitable. The residual sequence of a suitable model shall be the white noise process, and its ACF and PACF coefficients should be within twice the standard deviation range, otherwise, it is considered that the extraction information of the established model is not sufficient, and it is necessary to consider improving the accuracy of the model.

Fourth step. Using the established model to do prediction and analysis.

Elman neural network model

The E1man neural network (see Figure 2) is proposed by E1man in 1990 [31]. The model is generally divided into four layers: input layer, hidden layer, receiving layer and output layer. The characteristic of E1man network is that the output of the hidden layer is connected to the input of the hidden layer through the delay and storage of the receiving layer, which makes it sensitive to the data of the historical state. The addition of the internal feedback network increases the ability of the network itself to deal with dynamic information, thus achieving the purpose of dynamic modeling.

The mathematical structure of the Elman neural network is as follows:

$$x(k)=f(w^1x_c(k)+ w^2u(k\text{-}1))$$

$$x_c(k)= \alpha\, x_c(k\text{-}1)+ x(k\text{-}1)$$

$$y(k)=g(w^3x(k))$$

Where, w^1 is the connection weight matrix between the contact unit and the hidden layer unit, w^2 is the connection weight matrix between the input unit and the hidden layer unit, w^3 is the connection weight matrix between the hidden layer unit and the output unit. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, α is a self-connected feedback gain factor, $0\leqslant\alpha<1$, $f(x)$ often takes the sigmoid function.

There are four main steps in Elman neural network modeling:

First step. Data standardization processing. Data standardization is scaling the data to a small specific interval. In order to remove the unit limit of the data and convert it into dimensionless pure value, it is convenient for the index of different units or order of magnitude to be compared and weighted. In our study, we use function package mapminmax() to standardize the data, standardized data is in [-1,1] range.

Second step. Determine the input layer, the output layer. Generally, the input and output layer are determined according to the characteristics of the data and the needs of the analysis.

Third step. Set the parameters of the Elman model, such as training epochs and goals, etc., in our study, training epochs and goals of Elman neural network were set to 2000 and 0.00001, respectively. To determine the number of neurons in the hidden layer, so that the error of the established Elman model is minimized. At present, there is no ideal analytical expression for the number of neurons in the hidden layer. The number of neurons has a great influence on the performance of the network. When the number of neurons is too large, it will lead to that the network learning time is too long, the generalization performances are not good, and even the convergence failure, but when the number of neurons is too small, the fault-tolerant ability of the network is poor. In general, the number of neurons does not exceed 20. [32] In this study, matlab cyclic structure was used to find the optimal number of neurons by comparing the RMSE values of Elman networks with neurons 1 to 20.

Fourth step. According to the optimal number of neurons in the hidden layer, the Elman model is constructed, and then the prediction and analysis can be made.

Model comparison measures

Three performance indexes, root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to assess the fitting and forecasting accuracy of two models. The smaller these three values are, the better the model is. Their expressions are as follows [10]:

$$RMSE=\sqrt{\frac{\sum_{t=1}^{n}(X_t - \hat{X}_t)^2}{n}},$$

$$MAE=\frac{\sum_{t=1}^{n}|X_t - \hat{X}_t|}{n},$$

$$MAPE=\frac{\sum_{t=1}^{n}\left|\frac{X_t - \hat{X}_t}{X_t}\right| \times 100}{n},$$

where $\hat{X}_t$ is the simulating and forecasting values, X_t is the actual values and n is the number of observations.

Statistical software

All data analyses were conducted using Eviews7, matlab2015b, Arcmap10.1.

Results

From January 2005 to December 2017, the number of reported TB cases was 141984 in Kashgar, Xinjiang, the average annual TB cases were 7888 and the average annual incidence was 191.18. Figure 3 showed the time series graph of the TB incidence. It can be seen from the Figure 3 that the curve of TB incidence showed strong nonlinear characteristics from 2005 to 2014, and the TB incidence from 2015 to 2017 were significantly higher than that of previous years.

The data of TB incidence from January 2005 to December 2017 were divided into two parts. The data from January 2005 to December 2016 were used to build model and the data from January 2017 to December 2017 were used to test the model.

Establishment of ARMA Model

ARMA Modeling requires data stability, so, first of all, the stability of the modeling part of the data was verified by ADF test. The results of ADF test showed that p-value was less than 0.05(see Table 1), which indicated that the data was stable and could be directly used to build the model. Secondly, the ACF and PACF graphs of the modeling data were plotted (see Figure 4), it was obvious from Figure 4 that the autocorrelation coefficients were trailing distribution, and the partial correlation coefficients were almost a second-order truncated distribution, only at the lag 7, 8 and 9, the correlation coefficients were a little large. Based on this situation, we considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and AR((1,2,9)). The least square method was used to test the parameters of the four models. The results of the test were shown in Table 2; we can see that, of the four models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing the two models, it was found that, the AR((1,2,8)) model had smaller AIC and SC values, and the R^2 values of AR((1,2,8)) were larger than the R^2 values of AR (2), so the AR (2) model was abandoned. Then, the residual analysis of the AR((1,2,8)) model was carried out, the autocorrelation and partial correlation coefficient of the residuals were almost all within two times standard deviation, and only in the lag 5,6,12, they were beyond the range of two times standard deviation (see Figure 5), which indicated that the AR((1,2,8)) model could be used to roughly predict the TB incidence in Kashgar. We used AR((1,2,8)) model to fit the TB incidence from September 2005 to December 2016, the fitting RMSE ,MAE and MAPE were 6.15, 4.33 and 0.2858, respectively; we used AR((1,2,8)) model to predict the TB incidence from January 2017 to December 2017, the prediction RMSE, MAE and MAPE were 10.88 , 8.75 and 0.2029, respectively.

Establishment of AR-Elman Model

In order to improve the prediction accuracy of the AR((1,2,8)) model, we tried to establish the AR((1,2,8))-Elman hybrid model. The fitting sequence of AR((1,2,8)) model was used as input variable, and the actual TB incidence was used as output variable. Due to the a little similarity of the annual trend of TB incidence in Kashgar (see Figure 3), therefore, we created twelve time-lagged variables as input features. Supposing that x_t represented the TB incidence at time t, and then the input matrix and the output matrix of modeling data set used in this study were designed as follows (N=12):

$$input\ matrix=\begin{bmatrix} x_1 & x_2 & \dots & x_i \\ x_2 & x_3 & \dots & x_{i+1} \\ & \dots\dots & & \\ x_N & x_{N+1} & \dots & \dots \end{bmatrix},\ output\ matrix=[x_{N+1}\ \ x_{N+2}\ \ \dots\ \ x_{N+i}]$$

We selected twelve as the number of input layers of AR-Elman network and one as the number of output layers representing the forecast value. By the matlab cyclic structure, we selected the optimal number of neurons between 1 and 20, and finally we found when the number of neurons was 6 (see Figure 6), the RMSE was the smallest, and the AR-Elman was optimal. We used the AR-Elman model to fit the training data, RMSE was 3.78, MAE was 3.38, MAPE was 0.1837, and the R^2 of the model was 0.83; we used the AR-Elman model to predict the TB incidence from January 2017 to December 2017, RMSE was 8.86, MAE was 7.29, and MAPE was 0.2006. The fitting curves of AR((1,2,8)) model and AR-Elman model, and the prediction curve of AR-Elman model were shown in Figure 7. Comparison results of the AR((1,2,8)) model and AR-Elman model were shown in Table 3, both the fitting RMSE, MAE and MAPE and the predicting RMSE , MAE and MAPE of the AR-Elman model were smaller than those of the single AR((1,2,8)) model, which indicated that the AR-Elman combined model established in this study was more suitable for predicting the TB incidence in Kashgar.

Discussion

According to the WHO 2019 Global Tuberculosis report [4], around the world, TB mortality was down about 3% every year, the incidence was down about 2% every year, 16% of TB patients died of the disease. [4] But the rate of decline has not reached the pace of the stop Tuberculosis Strategy Plan. Therefore, it is necessary to strengthen the prevention and control of tuberculosis. In order to significantly narrow these gaps, greater progress must be made in a group of countries with a high burden of tuberculosis. The burden of TB in China ranks second in the world, and Xinjiang is the province with the highest incidence of TB in China, and Kashgar is the area with the high TB incidence in Xinjiang. Therefore, it is urgent to do a good job in the prevention and control of TB in Kashgar.

The prediction and early warning of infectious diseases is an important link in the prevention and control of infectious diseases. [32-34] Therefore, this study carried out research from the point of view of prediction to explore an accurate prediction model and do prediction analysis of TB incidence in Kashgar, so as to provide scientific reference for the prevention and control of the disease in this area. The Box -Jenkins method is a popular time series prediction method, this method has good prediction performance and high prediction accuracy; Elman Neural network can capture nonlinear information of time series data very well. In this study, the two methods were combined to study the prediction model of TB incidence in Kashgar.

Many studies have found that Box-Jenkins method has a good ability of fitting and forecasting. For stationary time series that do not contain seasonality, it is more suitable to use the ARMA model of the Box-Jenkins method to do prediction analysis [35], for non-stationary time series of infectious diseases with obvious seasonality, it is more suitable to use seasonal autoregressive integrated moving average (SARIMA) model of the Box-Jenkins method for prediction analysis.[9-12] In our study, from Figure 3, we could see that the seasonality of the TB incidence in Kashgar from 2005 to 2014 was not obvious, there was only a certain seasonality from 2015 to 2017, and we found that the time series of TB incidence was stable by ADF unit root test, and the autocorrelation and partial correlation coefficients of modeling data at lag 12, 24 were not obviously large, therefore, for our research data, we used ARMA model to do forecast analysis, finally, we established AR((1,2,8)) model of the Box-Jenkins method, it has a good performance to fit and predict the TB incidence of Kashgar in Xinjiang. From Figure 3, we can also see that the time series of TB incidence has strong non-linear, the established AR((1,2,8)) model mainly extracted the linear information of data, considering that the neural network can capture the non-linear information of data well, in order to improve the prediction accuracy of TB incidence rate in Kashgar, we used AR((1,2,8)) model and Elman neural network model to establish AR-Elman hybrid model. Many studies have found that the combination model can improve the accuracy of prediction, such as, Wang et al.[28] found that SARIMA-NAR hybrid model has an outstanding ability to improve the prediction accuracy relative to SARIMA model and nonlinear autoregressive network (NAR) model when they were used to predict pertussis incidence in China. Li et al.[27] found ARIMA-GRNN hybrid model was shown to be superior to the single ARIMA model in predicting the short-term TB incidence in the Chinese population. Our research was consistent with these literatures that our AR-Elman hybrid model was more accurate than the single AR((1,2,8)) model.

In the past few years, Xinjiang's economic development was relatively backward, medical resources were scarce, diagnosis and treatment were delayed, the continuous spread of TB has become a difficult problem in the control of TB in Xinjiang. In recent years, Xinjiang has introduced many new policies to increase investment in TB prevention and control, and the relevant departments of disease prevention and control in Xinjiang have also done a lot of effective work, which has helped to control effectively the rapid increase of the TB incidence in Xinjiang. In order to do a good job in the prevention and control of TB in Xinjiang, many departments need to make joint efforts. Our research was mainly to build a high-precision prediction model to help early warning and prediction analysis of tuberculosis in Kashgar. Finally, we established the AR-Elman hybrid model, which had high fitting and prediction accuracy of TB incidence in Kashgar, Xinjiang.

Our study found that Box-Jenkins and Elman neural network hybrid method was an effective method for predicting the incidence of TB in Kashgar, it could provide a scientific reference for prediction analysis of TB incidence. However, our study also has some limitations: our method is only suitable for short-term prediction, long-term prediction performance will decline, two main reasons: first, our model was based on historical data characteristics; second, climatic factors, environmental factors, demographic factors and political issues may have certain impacts on the change of incidence. Therefore, if the established model becomes old and people want to obtain more accurate prediction results, it will be needed to adjust the model parameters, update the model based on the new modeling sample data, and then to do prediction analysis.

Conclusions

Kashgar has a very high TB incidence, in order to provide some help for the prevention and control of this disease, the prediction problem of the TB incidence was studied. Firstly, a single AR((1,2,8)) prediction model was established by using Box-Jenkins method, and its fitting and prediction performance were good, secondly, in order to improve the prediction accuracy of the single AR((1,2,8)) model, we used single AR((1,2,8)) and Elman neural network with strong ability to capture nonlinear information to establish AR-Elman hybrid model. The fitting and prediction accuracy of the hybrid model was higher than that of the single AR((1,2,8)) model. The AR-Elman hybrid model can provide scientific reference for predicting and warning the TB incidence in Kashgar, Xinjiang. This study also has some limitations, long-term prediction accuracy of the AR-Elman hybrid model will decline, after the model becomes old, it will be needed to adjust the model parameters and establish new Box-Jenkins and Elman neural network hybrid model to do prediction analysis based on new sample data.

Competing interests

The authors declare no conflict of interest.

Funding

This work was supported by Major projects of science and technology in Xinjiang Autonomous region (Grant No.2017A03006), China, and National Natural Science Foundation Project of China (Grant No. 72064036, 82060609),and State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia Fund (Grant No. SKL-HIDCA-2020-9)

Data sharing statement

No additional data are available.

Author contributions

YLZ and XLZ analyzed the data and wrote the manuscript. XLZ, XJW, KW and YC wrote and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to express our gratitude to anonymous peer reviewers for carefully revising our manuscript and for his or her useful comments.

References

1.Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia: prevalence, risk factors and its public health importance. BMC Infectious Diseases, 2019; 19(1).

2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity pattern of Mycobacterium tuberculosis strains isolated from pulmonary tuberculosis patients in and around Ambo Town, Central Ethiopia. Plos One, 2018;

13(2):e0193083-.

3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control in Debre Markos Town, Northwest Ethiopia. Advances in Medicine, 2018;2018(2):1-8.

4. WHO. Global tuberculosis report 2019, https://www.who.int/tb/publications/global_report/en/. (Accessed on 17 Oct 2019).

5. Zhao YF, Shou MH, Wang ZX. Prediction of the Number of Patients Infected with COVID-19 Based on Rolling Grey Verhulst Models. Int J Environ Res Public Health. 2020; 17(12):4582.

6. Guan P, Wu W, Huang D. Trends of reported human brucellosis cases in mainland China from 2007 to 2017: an exponential smoothing time series analysis. Environ Health Prev Med. 2018; 23(1):23.

7. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, Xu JY, Shi J, Sun DY. Analysis and predication of tuberculosis registration rates in Henan Province, China: an exponential smoothing model study. Infect Dis Poverty. 2020;9(1):123.

8. Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts. PLoS Negl Trop Dis. 2017;11(7):e0005696.

9. Wang Y, Xu C, Zhang S, Wang Z, Zhu Y, Yuan J. Temporal trends analysis of human brucellosis incidence in mainland China from 2004 to 2018. Sci Rep. 2018;8(1):15901.

10. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S. An Advanced Data-Driven Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series Forecasting in Qinghai Province, China. Infect Drug Resist. 2020;13:867-880.

11. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of tuberculosis cases reported at a tertiary hospital in Ghana: a time series model approach. BMC Public Health.2018;18(1):1292.

12. Tohidinik H R , Mohebali M , Mansournia M A, et al. Forecasting zoonotic cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran: A SARIMA Analysis. Tropical Medicine & International Health,

2018;23(8):860–869.

13. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in Ouagadougou, Burkina Faso, 2011–2015. Malaria Journal, 2018;17(1):138.

14. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease outbreak and primary healthcare delivery in Liberia: Time-series analyses for 2010–2016. Plos Medicine, 2018; 15(2):e1002508.

15. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: a time-series model based on weather variability. International Journal of Biometeorology, 2017;(6):1-10.

16. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. Cogent Social Sciences, 2018; 4(1):1461544.

17.Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving average model and generalised regression neural network model for prediction of haemorrhagic fever with renal syndrome in China: a time-series study. BMJ Open, 2019; 9(6):e025773.

18.Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in children younger than 48 months in Suzhou using a seasonal autoregressive integrated moving average model. BMJ Open, 2019; 9(1).

19.Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable incidence time series. Epidemiology & Infection, 2019; 147.

20.Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand, foot and mouth disease cases in China from 2008 to 2018. Epidemiology & Infection, 2019; 147.

21.Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. Bmj Open, 2017； 7(11):e018628.

22. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum Analysis for wind speed forecasting based on Elman Neural Network. Energy Conversion & Management, 2017; 148:895-904.

23.Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for Forecasting China's Carbon Dioxide Emissions. Communications in Computer and Information Science, 2018;(951): 36-47.

24. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural Network for Disease Prediction. Arabian Journal for Science and Engineering, 2019;44(11): 9209–9220.

25. Mehrgini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman artificial neural network. Carbonates & Evaporites, 2017;(6):1-11.

26. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. Neurocomputing, 2017; 228(C):154-164.

27. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, Ji Y, Xu D, Wang J. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infect Drug Resist. 2019;12:1011-1020.

28. Wang Y, Xu C, Wang Z, Zhang S, Zhu Y, Yuan J. Time series modeling of pertussis incidence in China from 2004 to 2018 with a novel wavelet based SARIMA-NAR hybrid model. PLoS One. 2018;13(12):e0208404.

29. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, Yuan J. Temporal trends analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a new SARIMA-NARNNX hybrid model. BMJ Open. 2019;9(7):e024409.

30. Box G E, Jenkins G M. Time series analysis: forecasting and control rev. ed. Oakland, California, Holden-Day, 1976;31(4):238-242.

31. Ming Cheng.The principle and example of MATLAB neural network. Tsinghua University Press, 2013.

32. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clin Microbiol Infect. 2013;19(11):999-1005

33.Wearing HJ1, Rohani P,Keeling MJ. Appropriate models for the management of infectious diseases. PLoS Med. 2005; 2(7):e174.

34. Neuberger A, Paul M, Nizar A, Raoult D. Modelling in infectious diseases:

between haphazard and hazard. Clin Microbiol Infect. 2013;19(11):993-8.

35. Tipirneni-Sajja A, Krafft AJ, Loeffler RB, Song R, Bahrami A, Hankins JS, Hillenbrand CM. Autoregressive moving average modeling for hepatic iron quantification in the presence of fat. J Magn Reson Imaging. 2019;50(5):1620-1632.

Figures

Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China. Kashgar is located in the south of Xinjiang, and it has a very high incidence of tuberculosis.

Figure 2. The Structure diagram of Elman neural network. w^1, w^2 and w^3 are the connection weight matrixes. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, $u(k\text{-}1)$ represents the input of the input unit.

Figure 3. Graph of the tuberculosis (TB) incidence in Kashgar from January 2005 to December 2017. The curve of TB incidence showed strong nonlinear characteristics from 2005 to 2014, and the TB incidence increased significantly from 2015 to 2017.

Figure 4. Autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of modeling data. As the delay of the lag order, the autocorrelation coefficients were trailing and the partial correlation coefficients were truncated, so it was suitable to establish the AR model.

Figure 5. Autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of residuals of AR((1,2,8)) model. Autocorrelation coefficients and partial correlation coefficients were almost in 95% confidence interval, so AR ((1,2,8)) model could extract the information of original data well.

Figure 6. The numbers of neurons in AR-Elman model and the corresponding root mean square error (RMSE). When the number of neuron was 6, the RMSE was the smallest, and the AR-Elman model fitting ability was the strongest.

Figure 7. The fitting curves of AR((1,2,8)) model and AR-Elman model, and the prediction curve of AR-Elman model. Red line stands for the original tuberculosis (TB) incidence curve, green line stands for AR((1,2,8)) model fitting curve, blue line stands for AR-Elman model fitting curve. Blue dotted line stands for prediction curve of AR-Elman model, black dotted line stands for predicted curve of confidence intervals. The fitting ability of AR-Elman hybrid model was slightly better than that of the single AR((1,2,8)).

Tables

Table 1. The Augmented Dickey-Fuller (ADF) test of the training data

		t-Statistics	p-value
Augmented Dickey-Fuller test statistic		-3.47	0.01
Test critical values:	1% level	-3.48	
	5% level	-2.88	

10% level	-2.58

Table 2. Parameter estimates of the tentative models with their Akaike information criterion (AIC) and Schwarz criterion (SC) values.

Models	Variables	Coefficients	Std. Errors	t	p-values	AIC	SC
AR (2)	C	21.42	2.17	9.86	<0.01	6.55	6.62
	AR(1)	0.42	0.08	5.20	<0.01		
	AR(2)	0.34	0.08	4.17	<0.01		
AR((1, 2, 7))	C	22.93	3.73	6.14	<0.00	6.53	6.62
	AR(1)	0.41	0.08	5.10	<0.00		
	AR(2)	0.32	0.08	3.88	<0.00		
	AR (7)	0.12	0.007	1.74	0.08		
AR((1, 2, 8))	C	23.53	4.56	5.16	<0.01	6.53	6.61
	AR(1)	0.40	0.08	4.84	<0.01		
	AR(2)	0.32	0.08	3.96	<0.01		
	AR (8)	0.15	0.07	2.17	0.03		
AR((1, 2, 9))	C	29.07	15.53	1.87	0.06	6.46	6.55
	AR(1)	0.37	0.08	4.64	<0.01		
	AR (2)	0.31	0.08	3.93	<0.01		
	AR (9)	0.26	0.07	3.80	<0.01		

Table 3. Comparison results of in-sample fitting and out-of-sample forecasting performance for the AR((1,2,8)) model and AR-Elman model.

Models	Fitted efficacy			Models	Forecasted efficacy		
	RMSE	MAE	MAPE		RMSE	MAE	MAPE
AR((1,2,8))	6.15	4.33	0.2585	AR((1,2,8))	10.88	8.75	0.2029
AR-Elman	3.78	3.38	0.1837	AR-Elman	8.86	7.29	0.2006

Figure 1

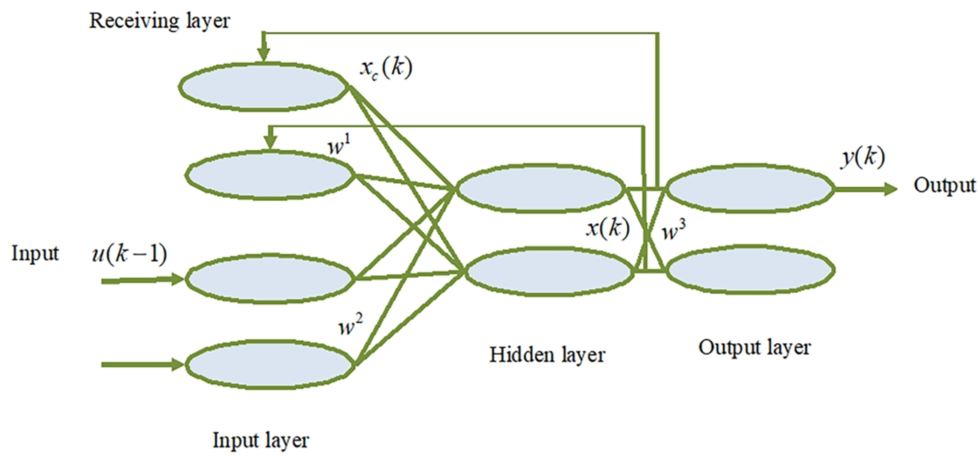
Receiving layer
$x_c(k)$
w^1
$y(k)$
Output
$x(k)$
w^3
Input
$u(k-1)$
w^2
Hidden layer
Output layer
Input layer

Figure 2

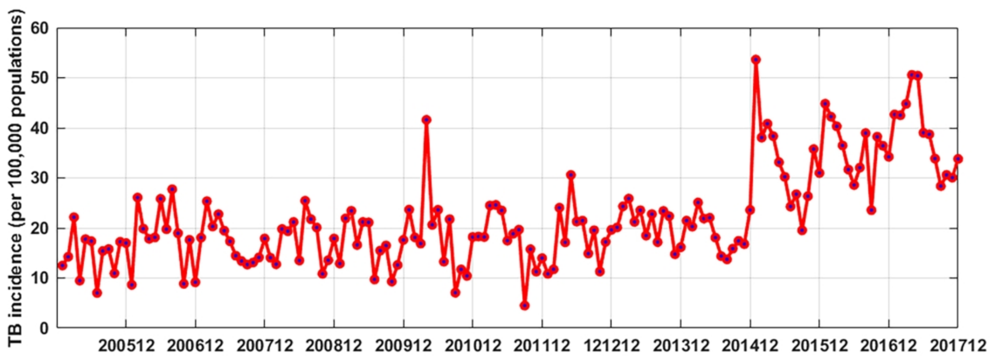
TB incidence (per 100,000 populations)
0
10
20
30
40
50
60
200512
200612
200712
200812
200912
201012
201112
121212
201312
201412
201512
201612
201712

Figure 3

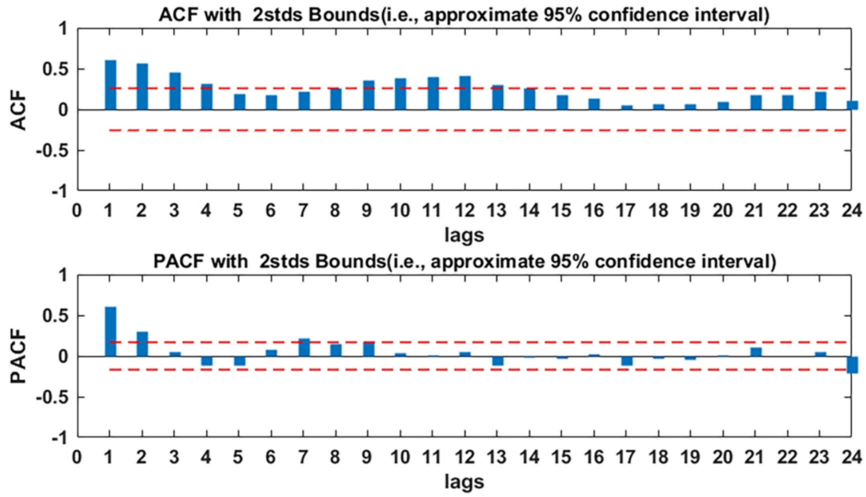
ACF with 2stds Bounds(i.e., approximate 95% confidence interval)
ACF
lags
PACF with 2stds Bounds(i.e., approximate 95% confidence interval)
PACF
lags

Figure 4

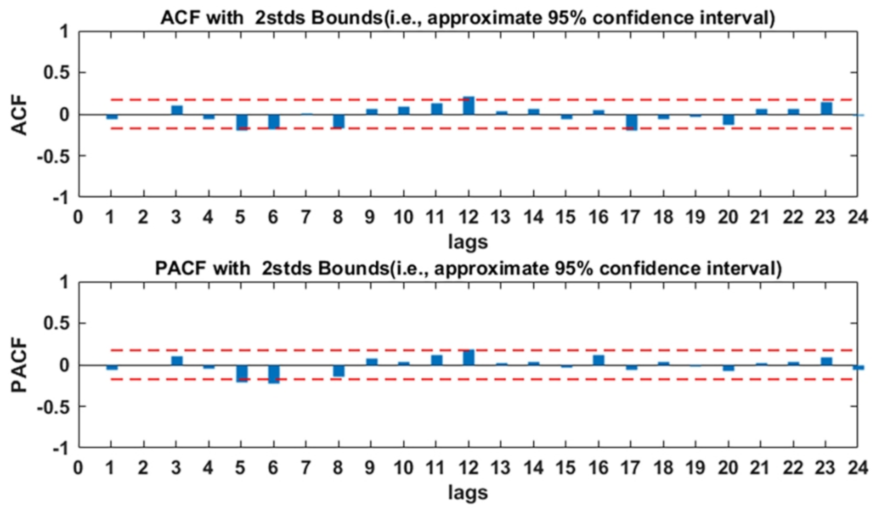
ACF with 2stds Bounds(i.e., approximate 95% confidence interval)
ACF
1
0.5
0
-0.5
-1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
lags
PACF with 2stds Bounds(i.e., approximate 95% confidence interval)
PACF
1
0.5
0
-0.5
-1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
lags

Figure 5

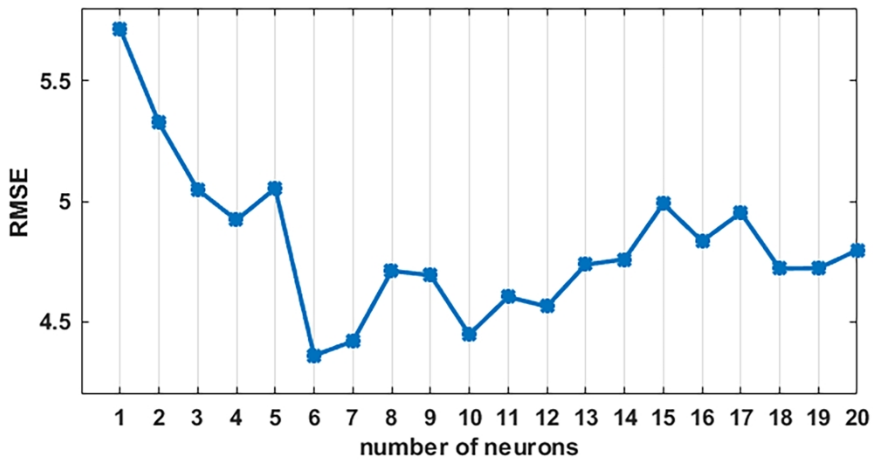
5.5
5
4.5
RMSE
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
number of neurons

Figure 6

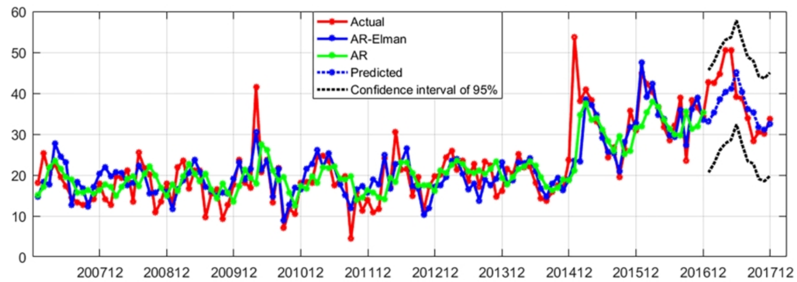
Actual
AR-Elman
AR
Predicted
Confidence interval of 95%
60
50
40
30
20
10
0
200712
200812
200912
201012
201112
201212
201312
201412
201512
201612
201712

Figure 7

BMJ Open

Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041040.R3
Article Type:	Original research
Date Submitted by the Author:	07-Dec-2020

Complete List of Authors:	Zheng, Yanling; State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, College of Medical Engineering and Technology,Xinjiang Medical University,Urumqi 830054, People’s Republic of China Zhang, Xueliang; State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, College of Medical Engineering and Technology,Xinjiang Medical University,Urumqi 830054, People’s Republic of China Wang, Xijiang; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People’s Republic of China Wang, Kai; State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, College of Medical Engineering and Technology,Xinjiang Medical University,Urumqi 830054, People’s Republic of China Cui, Yan; Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,Urumqi 830001, People’s Republic of China

Primary Subject Heading:	Health services research
Secondary Subject Heading:	Health services research, Infectious diseases
Keywords:	International health services < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Tuberculosis < INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS

Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China

Yanling Zheng [1], XueLiang Zhang[1*], XiJang Wang[2], Kai Wang[1] , Yan Cui[2*]

1. State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830054, People's Republic of China
2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830001, People's Republic of China

Xueliang Zhang :* Corresponding author, zhengyl_math@sina.cn
Yan Cui :* Corresponding author, cuiyan_jk@sina.com

Abstract

Objective: The incidence of tuberculosis (TB) in Kashgar, China, is very high, so the prevention and control of TB is arduous. However, there is little quantitative prediction study on the TB incidence, therefore, it is urgent to do prediction analysis of TB incidence in Kashgar, which can provide scientific reference for the prevention and control of TB.

Methods: Based on the monthly TB incidence, a single Box-Jenkins method and a Box-Jenkins and Elman neural network (ElmanNN) hybrid method were used to do prediction analysis of TB incidence in Kashgar. Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to measure the prediction accuracy.

Results: After careful analysis, the single AR((1,2,8)) model and AR((1,2,8))-ElmanNN (AR-Elman) hybrid model were established, and the optimal neurons value of the AR-Elman hybrid model was 6. For the fitting dataset, the RMSE, MAE and MAPE were 6.15, 4.33 and 0.2858, respectively, for the AR((1,2,8)) model, and 3.78, 3.38 and 0.1837, respectively, for the AR-Elman hybrid model. For the forecasting dataset, the RMSE, MAE and MAPE were 10.88, 8.75 and 0.2029, respectively, for the AR((1,2,8)) model, and 8.86, 7.29 and 0.2006, respectively, for the AR-Elman hybrid model.

Conclusions: Both the single AR((1,2,8)) model and AR-Elman model could be used to predict the TB incidence in Kashgar, but the modeling and validation scale-dependent measures (RMSE,MAE and MAPE) in the AR((1,2,8)) model were inferior to those in the AR-Elman hybrid model, which indicated that AR-Elman hybrid model was better than AR((1,2,8)) model. The Box-Jenkins and Elman neural network hybrid method can be highlighted in predicting the temporal trends of TB incidence in Kashgar, which may act as the potential for far-reaching implications for prevention and control of TB.

Keywords: Tuberculosis; Prediction; Box-Jenkins method; Elman neural network

Strengths and limitations of this study:

1. The incidence of tuberculosis (TB) is very high in Kashgar, therefore, it is urgent to do the prediction analysis of TB.
2. In the study, AR-Elman model with high prediction accuracy was established, it can be used to predict the development of TB in Kashgar, China.
3. The long-term prediction accuracy of AR-Elman hybrid model will decline.

Introduction

Tuberculosis (TB) is still a major global public health problem and the ninth leading cause of death in the world. [1-3] Because TB is contagious, patient resistance is low, treatment time is long, patients' labor force is lost and their contribution to society is reduced, so TB brings great burden to society and economy. All countries in the world are working hard to fight tuberculosis. In 2018, the number of new reported TB cases was about 10 million; this figure has remained relatively stable in recent years. The latest treatment results showed that the global TB treatment success rate was 83%. World Health Organization (WHO) has set targets for the stop TB strategy. The targets mentioned that by 2030, on the basis of the work in 2015, TB deaths will reduce by 90%, and annual new TB cases will reduce by 80%.[4] In order to achieve these goals, TB prevention and control services must be provided in the broad context of universal health coverage, joint action must be taken to address the social and economic consequences of TB, and technological breakthroughs should be achieved by 2025 to make the TB incidence decline faster than that at any time in history. According to the global tuberculosis report 2019 [4], China has the second highest number of TB cases in the world. [4] The TB incidence in western China was much higher than that in eastern and central China. From 2016 to 2017, the annual TB incidence in Xinjiang was 185.66/100,000 and 202.58/100,000, while the annual TB incidence from 2016 to 2017 in China was 61/100,000 and 60.53/100,000, respectively, it is nearly three times higher in Xinjiang than that national level.

There are 14 Prefectural-Level cities in Xinjiang, China, among which Kashgar has a very high TB incidence rate. From 2016 to 2017, the annual TB incidences in Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7 times higher than that of the national level. Doing a good job in the prevention and control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.

Mastering the changing law of the incidence of infectious diseases, using the existing surveillance data to analyze, then, to predict the possible epidemic trend and provide reference data for the prevention, can better control the occurrence and epidemic of infectious diseases. Prediction of infectious diseases is to predict the occurrence, development and epidemic trend of infectious diseases according to the occurrence, development law and related factors of infectious diseases, based on analysis, judgment and mathematical model, etc.

For study of quantitative prediction of infectious diseases, there are many methods, such as grey prediction method [5], exponential smoothing prediction method [6,7], dynamic model prediction method [8], Box-Jenkins method [9], neural network method [10], etc., with the deepening of prediction research, more and more scholars like to use the Box-Jenkins method [11-21], there are many different models in this method, and if appropriate models are established according to the characteristics of time series, high prediction ability often can be obtained. Neural network has strong nonlinear mapping ability, in which Elman neural network is composed of input layer, hidden layer, connection layer and output layer. The Elman network has dynamic memory function, and it is very suitable for time series prediction. At present, the Elman network is widely used in various fields, and has achieved successful prediction results. [22-26] Sometimes, the prediction effect of a single model is not ideal, in order to further improve the prediction accuracy, many studies adopted combined model prediction method [27-29], the combined model can absorb the advantages of two or more methods so as to achieve a higher prediction accuracy.

The TB incidence of Kashgar is very high, and it is urgent to do a good job in the prevention and control of TB in this area. Accurate prediction of TB incidence is a prerequisite for prevention and control, which can help advance resource planning and policy formulation. In this study, the popular Box-Jenkins time series method was used to build model for predicting the TB incidence in Kashgar, in order to improve the prediction accuracy, the Elman neural network with strong nonlinear information capture ability was used to construct the combined model for prediction analysis.

Materials and Methods

Study area and data Sources

We selected Kashgar as the study site (see Figure 1). This area is located in the south of Xinjiang province in China with an area of approximately 16.2 thousand square kilometers and a permanent population of 4.64 million in 2018. TB case data from January 2005 to December 2017 were obtained from the Center for Disease Control and Prevention (CDC) of Xinjiang Uygur Autonomous Region, all TB cases in Xinjiang must be reported to the CDC through the infectious disease surveillance system within 24 hours. The TB cases datasets used need permission of the CDC. Population data were obtained from the website of Xinjiang Bureau of Statistics (http://tjj.xinjiang.gov.cn/tjj/tjfw/list_tjfw.shtml). Based on the population data and TB case data, we calculated the incidence data of TB.

Patient and public involvement

Patients were not involved in the design of this study as it involved only observational analysis of an anonymised, pre- existing, routinely collected dataset.

Ethics approval

Since no primary data collection was undertaken, no patient or public was involved, no formal ethical assessment or informed consent was required.

Autoregressive moving average (ARMA) model

ARMA model [30] is an important time series analysis and prediction model in Box-Jenkins method, also known as auto-regression moving average model. The ARMA (p,q) model is a model with autocorrelation order p and moving average order q, p and q are judged by autocorrelation function (ACF) and partial autocorrelation function (PACF) diagram of stationary data. If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are the p-order truncated, then q=0, the ARMA(p,q) model is recorded as AR(p), and its expression is as follows:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

where, X_t is the observed value at t, $\phi_1, \ldots, \phi_p$ are model parameters, c is a constant, if only ϕ_1, ϕ_2, ϕ_p are not zeros, then, The AR(p) model becomes a sparse model, which can be recorded as AR((1,2,p)).

If the original data is stable, the autocorrelation coefficients are q-order truncated and the partial correlation coefficients are trailing, then p=0 and the ARMA(p,q) model becomes MA(q), its expression is as follows:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

where, μ is the expected value of X_t. $\theta_1, \ldots, \theta_p$ are model parameters.

If the original data is stable, the autocorrelation coefficients are trailing, and the partial correlation coefficients are also trailing, then the expression of the ARMA (p, q) model is as follows:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}.$$

There are four main steps in ARMA modeling:

First step. The prerequisite for ARMA modeling is the stationary of time series. Check whether the data is stable by Augmented Dickey-Fuller (ADF) unit root test. In this study, the significant level probability (p-value) is 0.05, and if the p-value is less than 0.05, then, the data is stable. By observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the stable data, we can determine the possible values of p and q and establish the possible ARMA (p, q) model.

Second step. The parameters of ARMA(p, q) model are estimated by maximum likelihood estimation or least square estimation, and the model parameters are tested. If p-value is less than 0.05, the parameters have statistical significance. The best model is determined according to the value of Akaike information criterion (AIC), Schwarz criterion (SC) and Goodness of Fit (R^2) of model. The smaller AIC and SC are, the larger the R^2 is, and the better the model is.

Third step. To determine whether the established ARMA (p, q) model is suitable. The residual sequence of a suitable model shall be the white noise process, and its ACF and PACF coefficients should be within twice the standard deviation range, otherwise, it is considered that the extraction information of the established model is not sufficient, and it is necessary to consider improving the accuracy of the model.

Fourth step. Using the established model to do prediction and analysis.

Elman neural network model

The E1man neural network (see Figure 2) is proposed by E1man in 1990 [31]. The model is generally divided into four layers: input layer, hidden layer, receiving layer and output layer. The characteristic of E1man network is that the output of the hidden layer is connected to the input of the hidden layer through the delay and storage of the receiving layer, which makes it sensitive to the data of the historical state. The addition of the internal feedback network increases the ability of the network itself to deal with dynamic information, thus achieving the purpose of dynamic modeling.

The mathematical structure of the Elman neural network is as follows:

$$x(k)=f(w^1x_c(k)+ w^2u(k-1))$$

$$x_c(k)= \alpha x_c(k-1)+ x(k-1)$$

$$y(k)=g(w^3x(k))$$

Where, w^1 is the connection weight matrix between the contact unit and the hidden layer unit, w^2 is the connection weight matrix between the input unit and the hidden layer unit, w^3 is the connection weight matrix between the hidden layer unit and the output unit. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, α is a self-connected feedback gain factor, $0\leqslant\alpha<1$, $f(x)$ often takes the sigmoid function.

There are four main steps in Elman neural network modeling:

First step. Data standardization processing. Data standardization is scaling the data to a small specific interval. In order to remove the unit limit of the data and convert it into dimensionless pure value, it is convenient for the index of different units or order of magnitude to be compared and weighted. In our study, we used function package mapminmax() to standardize the data, standardized data was in [-1,1] range.

Second step. Determine the input layer, the output layer. Generally, the input and output layer are determined according to the characteristics of data and the needs of the analysis.

Third step. Set the parameters of the Elman model, such as training epochs and goals, etc., in our study, training epochs and goals of Elman neural network were set to 2000 and 0.00001, respectively. To determine the number of neurons in the hidden layer, so that the error of the established Elman model is minimized. At present, there is no ideal analytical expression for the number of neurons in the hidden layer. The number of neurons has a great influence on the performance of the network. When the number of neurons is too large, it will lead to that network learning time is too long, generalization performances are not good, and even the convergence failure, but when the number of neurons is too small, fault-tolerant ability of the network is poor. In general, the number of neurons does not exceed 20. [32] In this study, matlab cyclic structure was used to find the optimal number of neurons by comparing the RMSE values of Elman networks with neurons 1 to 20.

Fourth step. According to the optimal number of neurons in the hidden layer, the Elman model is constructed, and then the prediction and analysis can be made.

Model comparison measures

Three performance indexes, root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to assess the fitting and forecasting accuracy of two models. The smaller these three values are, the better the model is. Their expressions are as follows [10]:

$$RMSE=\sqrt{\frac{\sum_{t=1}^{n}(X_t-\hat{X}_t)^2}{n}},$$

$$MAE=\frac{\sum_{t=1}^{n}|X_t-\hat{X}_t|}{n},$$

$$MAPE=\frac{\sum_{t=1}^{n}\left|\frac{X_t-\hat{X}_t}{X_t}\right|\times 100}{n},$$

where, $\hat{X}_t$ is the simulating and forecasting values, X_t is the actual values and n is the number of observations.

Statistical software

All data analyses were conducted using Eviews7, matlab2015b, Arcmap10.1.

Results

From January 2005 to December 2017, the number of reported TB cases was 141984 in Kashgar, Xinjiang, the average annual TB cases were 7888 and the average annual incidence was 191.18. Figure 3 showed the time series graph of the TB incidence. It can be seen from the Figure 3 that the curve of TB incidence showed strong nonlinear characteristics from 2005 to 2014, and the TB incidence from 2015 to 2017 were significantly higher than that of previous years.

The data of TB incidence from January 2005 to December 2017 were divided into two parts. The data from January 2005 to December 2016 were used to build model and the data from January 2017 to December 2017 were used to test the model.

Establishment of ARMA Model

ARMA Modeling requires data stability, so, first of all, the stability of the modeling part of the data was verified by ADF test. The results of ADF test showed that p-value was less than 0.05(see Table 1), which indicated that the data was stable and could be directly used to build the model. Secondly, the ACF and PACF graphs of the modeling data were plotted (see Figure 4), it was obvious from Figure 4 that the autocorrelation coefficients were trailing distribution, and the partial correlation coefficients were almost a second-order truncated distribution, only at the lag 7, 8 and 9, the correlation coefficients were a little large. Based on this situation, we considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and AR((1,2,9)). The least square method was used to test the parameters of the four models. The results of the test were shown in Table 2; we can see that, of the four models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing the two models, it was found that, the AR((1,2,8)) model had smaller AIC and SC values, and the R^2 values of AR((1,2,8)) were larger than the R^2 values of AR (2), so the AR (2) model was abandoned. Then, the residual analysis of the AR((1,2,8)) model was carried out, the autocorrelation and partial correlation coefficient of the residuals were almost all within two times standard deviation, and only in the lag 5,6,12, they were beyond the range of two times standard deviation (see Figure 5), which indicated that the AR((1,2,8)) model could be used to roughly predict the TB incidence in Kashgar. We used AR((1,2,8)) model to fit the TB incidence from September 2005 to December 2016, the fitting RMSE ,MAE and MAPE were 6.15, 4.33 and 0.2858, respectively; we used AR((1,2,8)) model to predict the TB incidence from January 2017 to December 2017, the prediction RMSE, MAE and MAPE were 10.88 , 8.75

and 0.2029, respectively.

Establishment of AR-Elman Model

In order to improve the prediction accuracy of the AR((1,2,8)) model, we tried to establish the AR((1,2,8))-Elman hybrid model. The fitting sequence of AR((1,2,8)) model was used as input variable, and the actual TB incidence was used as output variable. Due to a little similarity of the annual trend of TB incidence in Kashgar (see Figure 3), therefore, we created twelve time-lagged variables as input features. Supposing that x_t represented the TB incidence at time t, and then the input matrix and the output matrix of modeling data set used in this study were designed as follows (N=12):

$$\textit{input matrix}=\begin{bmatrix} x_1 & x_2 & \cdots & x_i \\ x_2 & x_3 & \cdots & x_i+1 \\ & \cdots\cdots & & \\ x_N & x_N+1 & \cdots & \cdots \end{bmatrix}, \textit{output matrix}=[x_{N+1} \quad x_{N+2} \quad \cdots \quad x_{N+i}]$$

We selected twelve as the number of input layers of AR-Elman network and one as the number of output layers representing the forecast value. By the matlab cyclic structure, we selected the optimal number of neurons between 1 and 20, and finally we found when the number of neurons was 6 (see Figure 6), the RMSE was the smallest, and the AR-Elman was optimal. We used the AR-Elman model to fit the training data, RMSE was 3.78, MAE was 3.38, MAPE was 0.1837, and the R^2 of the model was 0.83; we used the AR-Elman model to predict the TB incidence from January 2017 to December 2017, RMSE was 8.86, MAE was 7.29, and MAPE was 0.2006. The fitting curves of AR((1,2,8)) model and AR-Elman model, and the prediction curve of AR-Elman model were shown in Figure 7. Comparison results of the AR((1,2,8)) model and AR-Elman model were shown in Table 3, both the fitting RMSE, MAE and MAPE and the predicting RMSE , MAE and MAPE of the AR-Elman model were smaller than those of the single AR((1,2,8)) model, which indicated that the AR-Elman combined model established in this study was more suitable for predicting the TB incidence in Kashgar.

Discussion

According to the WHO 2019 Global Tuberculosis report [4], around the world, TB mortality was down about 3% every year, the incidence was down about 2% every year, 16% of TB patients died of the disease. [4] But the rate of decline has not reached the pace of the stop Tuberculosis Strategy Plan. Therefore, it is necessary to strengthen the prevention and control of tuberculosis. In order to significantly narrow these gaps, greater progress must be made in a group of countries with a high burden of tuberculosis. The burden of TB in China ranks second in the world, and Xinjiang is the province with high incidence of TB in China, and Kashgar is the area with the high TB incidence in Xinjiang. Therefore, it is urgent to do a good job in the prevention and control of TB in Kashgar.

The prediction and early warning of infectious diseases is an important link in the prevention and control of infectious diseases.[32-34] Therefore, this study carried out research from the point of view of prediction to explore an accurate prediction model and do prediction analysis of TB incidence in Kashgar, so as to provide scientific reference for the prevention and control of the disease in this area. The Box -Jenkins method is a popular time series prediction method, this method has good prediction performance and high prediction accuracy; Elman Neural network can capture nonlinear information of time series data very well. In this study, the two methods were combined to study the prediction model of TB incidence in Kashgar.

Many studies have found that Box-Jenkins method has a good ability of fitting and forecasting. For stationary time series that do not contain seasonality, it is more suitable to use the ARMA model of the Box-Jenkins method to do prediction analysis [35], for non-stationary time series of infectious diseases with obvious seasonality, it is more suitable to use seasonal autoregressive integrated moving average (SARIMA) model of the Box-Jenkins method for prediction analysis.[9-12] In our study, from Figure 3, we could see that the seasonality of the TB incidence in Kashgar from 2005 to 2014 was not obvious, there was only a certain seasonality from 2015 to 2017, and we found that the time series of TB incidence was stable by ADF unit root test, and the autocorrelation and partial correlation coefficients of modeling data at lag 12, 24 were not obviously large, therefore, for our research data, we used ARMA model to do forecast analysis, finally, we established AR((1,2,8)) model of the Box-Jenkins method, it has a good performance to fit and predict the TB incidence of Kashgar in Xinjiang. From Figure 3, we can also see that the time series of TB incidence has strong non-linear, the established AR((1,2,8)) model mainly extracted the linear information of data, considering that the neural network can capture the non-linear information of data well, in order to improve the prediction accuracy of TB incidence in Kashgar, we used AR((1,2,8)) model and Elman neural network model to establish AR-Elman hybrid model. Many studies have found that the combination model could improve accuracy of prediction, such as, Wang et al.[28] found that SARIMA-NAR hybrid model had an outstanding ability to improve the prediction accuracy relative to SARIMA model and nonlinear autoregressive network (NAR) model when they were used to predict pertussis incidence in China. Li et al.[27] found ARIMA-GRNN hybrid model was superior to single ARIMA model in predicting the short-term TB incidence in Chinese population. Our research was consistent with these literatures that our AR-Elman hybrid model was more accurate than the single AR((1,2,8)) model.

In the past few years, Xinjiang's economic development was relatively backward, medical resources were scarce, diagnosis and treatment were delayed, the continuous spread of TB has become a difficult problem in the control of TB in Xinjiang. In recent years, Xinjiang has introduced many new policies to increase investment in TB prevention and control, and the relevant departments of disease prevention and control in Xinjiang have also done a lot of effective work, which has helped to control effectively rapid increase of the TB incidence in Xinjiang. In order to do a good job in the prevention and control of TB in Xinjiang, many departments need to make joint efforts. Our research was mainly to build a high-precision prediction model to help early warning and prediction analysis of tuberculosis in Kashgar. Finally, we established the AR-Elman hybrid model, which had high fitting and prediction accuracy of TB incidence in Kashgar, Xinjiang.

Our study found that Box-Jenkins and Elman neural network hybrid method was an effective method for predicting the incidence of TB in Kashgar, it could provide a scientific reference for prediction analysis of TB incidence. However, our study also has some limitations: our method was only suitable for short-term prediction, long-term prediction performance may decline, two main reasons: first, our model was based on historical data characteristics; second, climatic factors, environmental factors, demographic factors and political issues may have certain impacts on the change of incidence. Therefore, if the established model becomes old and people will want to obtain more accurate prediction results, it will be needed to adjust the model parameters, update the model based on the new modeling sample data, and then to do prediction analysis.

Conclusions

Kashgar has a very high TB incidence, in order to provide some help for the prevention and control of this disease, the prediction problem of the TB incidence was studied. Firstly, a single AR((1,2,8)) prediction model was established by using Box-Jenkins method, and its fitting and prediction performance were good, secondly, in order to improve the prediction accuracy of the single AR((1,2,8)) model, we used single AR((1,2,8)) and Elman neural network with strong ability to capture nonlinear information to establish AR-Elman hybrid model. The fitting and prediction accuracy of the hybrid model was higher than that of the single AR((1,2,8)) model. The AR-Elman hybrid model can provide scientific reference for predicting and warning the TB incidence in Kashgar, Xinjiang.

Acknowledgements

We would like to express our gratitude to peer reviewers for carefully revising our manuscript and for their very useful comments.

Contributions

YLZ and XLZ analyzed the data and wrote the manuscript. XLZ, XJW, KW and YC wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by Major projects of science and technology in Xinjiang Autonomous region (Grant No.2017A03006), China, and National Natural Science Foundation Project of China (Grant No. 72064036, 82060609),and State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia Fund (Grant No. SKL-HIDCA-2020-9)

Competing interests

None declared.

Patient consent for publication

Not required.

Provenance and peer review

Not commissioned; externally peer reviewed.

Data availability statement

The data used in this study are available from the corresponding authors on reasonable request.

References

1. Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia: prevalence, risk factors and its public health importance. BMC Infectious Diseases, 2019; 19(1).

2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity pattern of Mycobacterium tuberculosis strains isolated from pulmonary tuberculosis patients in and around Ambo Town, Central Ethiopia. Plos One, 2018; 13(2):e0193083-.

3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control in Debre Markos Town, Northwest Ethiopia. Advances in Medicine, 2018;2018(2):1-8.

4. WHO. Global tuberculosis report 2019, https://www.who.int/tb/publications/global_report/en/. (Accessed on 17 Oct 2019).

5. Zhao YF, Shou MH, Wang ZX. Prediction of the Number of Patients Infected with COVID-19 Based on Rolling Grey Verhulst Models. Int J Environ Res Public Health. 2020; 17(12):4582.

6. Guan P, Wu W, Huang D. Trends of reported human brucellosis cases in mainland China from 2007 to 2017: an exponential smoothing time series analysis. Environ Health Prev Med. 2018; 23(1):23.

7. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, Xu JY, Shi J, Sun DY. Analysis and predication of tuberculosis registration rates in Henan Province, China: an exponential smoothing model study. Infect Dis Poverty. 2020;9(1):123.

8. Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts. PLoS Negl Trop Dis. 2017;11(7):e0005696.

9. Wang Y, Xu C, Zhang S, Wang Z, Zhu Y, Yuan J. Temporal trends analysis of human brucellosis incidence in mainland China from 2004 to 2018. Sci Rep. 2018;8(1):15901.

10. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S. An Advanced Data-Driven Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series Forecasting in Qinghai Province, China. Infect Drug Resist. 2020;13:867-880.

11. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of tuberculosis cases reported at a tertiary hospital in Ghana: a time series model approach. BMC Public Health.2018;18(1):1292.

12. Tohidinik H R , Mohebali M , Mansournia M A, et al. Forecasting zoonotic cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran: A SARIMA Analysis. Tropical Medicine & International Health, 2018;23(8):860–869.

13. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in Ouagadougou, Burkina Faso, 2011–2015. Malaria Journal, 2018;17(1):138.

14. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease outbreak and primary healthcare delivery in Liberia: Time-series analyses for 2010–2016. Plos Medicine, 2018; 15(2):e1002508.

15. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: a time-series model based on weather variability. International Journal of Biometeorology, 2017;(6):1-10.

16. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. Cogent Social Sciences, 2018; 4(1):1461544.

17.Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving average model and generalised regression neural network model for prediction of haemorrhagic fever with renal syndrome in China: a time-series study. BMJ Open, 2019; 9(6):e025773.

18.Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in children younger than 48 months in Suzhou using a seasonal autoregressive integrated moving average model. BMJ Open, 2019; 9(1).

19.Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable incidence time series. Epidemiology & Infection, 2019; 147.

20.Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand, foot and mouth disease cases in China from 2008 to 2018. Epidemiology & Infection, 2019; 147.

21.Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. Bmj Open, 2017； 7(11):e018628.

22. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum Analysis for wind speed forecasting based on Elman Neural Network. Energy Conversion & Management, 2017; 148:895-904.

23.Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for Forecasting China's Carbon Dioxide Emissions. Communications in Computer and Information Science, 2018;(951): 36-47.

24. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural Network for Disease Prediction. Arabian Journal for Science and Engineering, 2019;44(11): 9209–9220.

25. Mehrgini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman artificial neural network. Carbonates & Evaporites, 2017;(6):1-11.

26. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. Neurocomputing, 2017; 228(C):154-164.

27. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, Ji Y, Xu D, Wang J. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infect Drug Resist. 2019;12:1011-1020.

28. Wang Y, Xu C, Wang Z, Zhang S, Zhu Y, Yuan J. Time series modeling of pertussis incidence in China from 2004 to 2018 with a novel wavelet based SARIMA-NAR hybrid model. PLoS One. 2018;13(12):e0208404.

29. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, Yuan J. Temporal trends analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a new SARIMA-NARNNX hybrid model. BMJ Open. 2019;9(7):e024409.

30. Box G E, Jenkins G M. Time series analysis: forecasting and control rev. ed. Oakland, California, Holden-Day, 1976；31(4):238-242.

31. Ming Cheng.The principle and example of MATLAB neural network. Tsinghua University Press, 2013.

32. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clin Microbiol Infect. 2013;19(11):999-1005

33.Wearing HJ1, Rohani P,Keeling MJ. Appropriate models for the management of infectious diseases. PLoS Med. 2005; 2(7):e174.

34. Neuberger A, Paul M, Nizar A, Raoult D. Modelling in infectious diseases: between haphazard and hazard. Clin Microbiol Infect. 2013;19(11):993-8.

35. Tipirneni-Sajja A, Krafft AJ, Loeffler RB, Song R, Bahrami A, Hankins JS, Hillenbrand CM. Autoregressive moving average modeling for hepatic iron quantification in the presence of fat. J Magn Reson Imaging. 2019;50(5):1620-1632.

Figures

Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China. Kashgar is located in the south of Xinjiang, and it has a very high incidence of tuberculosis.

Figure 2. The Structure diagram of Elman neural network. w^1, w^2 and w^3 are the connection weight matrixes. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, $u(k\text{-}1)$ represents the input of the input unit.

Figure 3. Graph of the tuberculosis (TB) incidence in Kashgar from January 2005 to December 2017. The curve of TB incidence showed strong nonlinear characteristics from 2005 to 2014, and the TB incidence increased significantly from 2015 to 2017.

Figure 4. Autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of modeling data. As the delay of the lag order, the autocorrelation coefficients

were trailing and the partial correlation coefficients were truncated, so it was suitable to establish the AR model.

Figure 5. Autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of residuals of AR((1,2,8)) model. Autocorrelation coefficients and partial correlation coefficients were almost in 95% confidence interval, so AR ((1,2,8)) model could extract the information of original data well.

Figure 6. The numbers of neurons in AR-Elman model and the corresponding root mean square error (RMSE). When the number of neuron was 6, the RMSE was the smallest, and the AR-Elman model fitting ability was the strongest.

Figure 7. The fitting curves of AR((1,2,8)) model and AR-Elman model, and the prediction curve of AR-Elman model. Red line stands for the original tuberculosis (TB) incidence curve, green line stands for AR((1,2,8)) model fitting curve, blue line stands for AR-Elman model fitting curve. Blue dotted line stands for prediction curve of AR-Elman model, black dotted line stands for predicted curve of confidence intervals. The fitting ability of AR-Elman hybrid model was slightly better than that of the single AR((1,2,8)).

Tables

Table 1. The Augmented Dickey-Fuller (ADF) test of the training data

		t-Statistics	p-value
Augmented Dickey-Fuller test statistic		-3.47	0.01
Test critical values:	1% level	-3.48	
	5% level	-2.88	
	10% level	-2.58	

Table 2. Parameter estimates of the tentative models with their Akaike information criterion (AIC) and Schwarz criterion (SC) values.

Models	Variables	Coefficients	Std. Errors	t	p-values	AIC	SC
AR (2)	C	21.42	2.17	9.86	<0.01	6.55	6.62
	AR(1)	0.42	0.08	5.20	<0.01		
	AR(2)	0.34	0.08	4.17	<0.01		
AR((1, 2, 7))	C	22.93	3.73	6.14	<0.00	6.53	6.62
	AR(1)	0.41	0.08	5.10	<0.00		
	AR(2)	0.32	0.08	3.88	<0.00		
	AR (7)	0.12	0.007	1.74	0.08		
AR((1, 2, 8))	C	23.53	4.56	5.16	<0.01	6.53	6.61
	AR(1)	0.40	0.08	4.84	<0.01		
	AR(2)	0.32	0.08	3.96	<0.01		
	AR (8)	0.15	0.07	2.17	0.03		
AR((1, 2, 9))	C	29.07	15.53	1.87	0.06	6.46	6.55
	AR(1)	0.37	0.08	4.64	<0.01		
	AR (2)	0.31	0.08	3.93	<0.01		
	AR (9)	0.26	0.07	3.80	<0.01		

Table 3. Comparison results of in-sample fitting and out-of-sample forecasting performance for the AR((1,2,8)) model and AR-Elman model.

Models	Fitted efficacy			Models	Forecasted efficacy		
	RMSE	MAE	MAPE		RMSE	MAE	MAPE
AR((1,2,8))	6.15	4.33	0.2585	AR((1,2,8))	10.88	8.75	0.2029
AR-Elman	3.78	3.38	0.1837	AR-Elman	8.86	7.29	0.2006

Figure 1

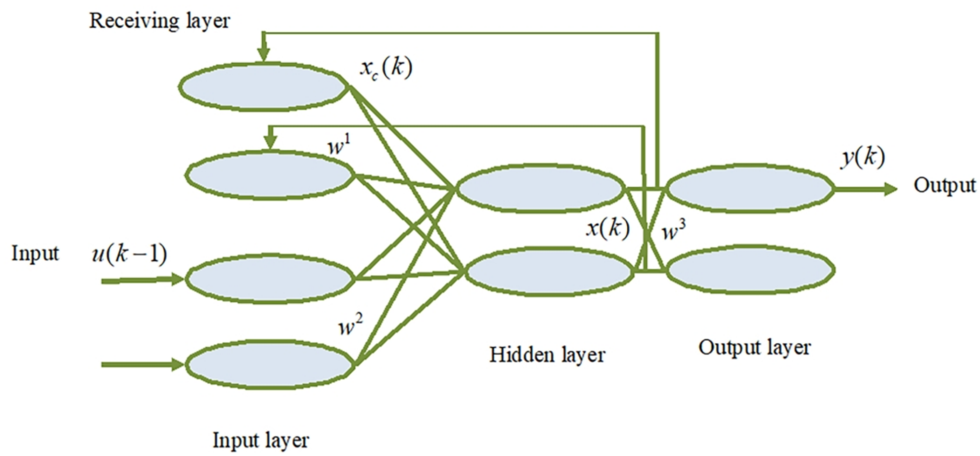
Receiving layer
$x_c(k)$
w^1
$y(k)$
Output
$x(k)$
w^3
Input
$u(k-1)$
w^2
Hidden layer
Output layer
Input layer

Figure 2

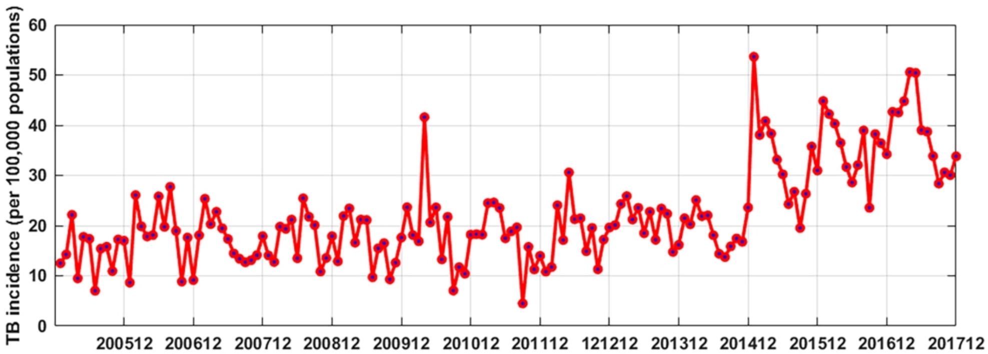
TB incidence (per 100,000 populations)
60
50
40
30
20
10
0
200512
200612
200712
200812
200912
201012
201112
121212
201312
201412
201512
201612
201712

Figure 3

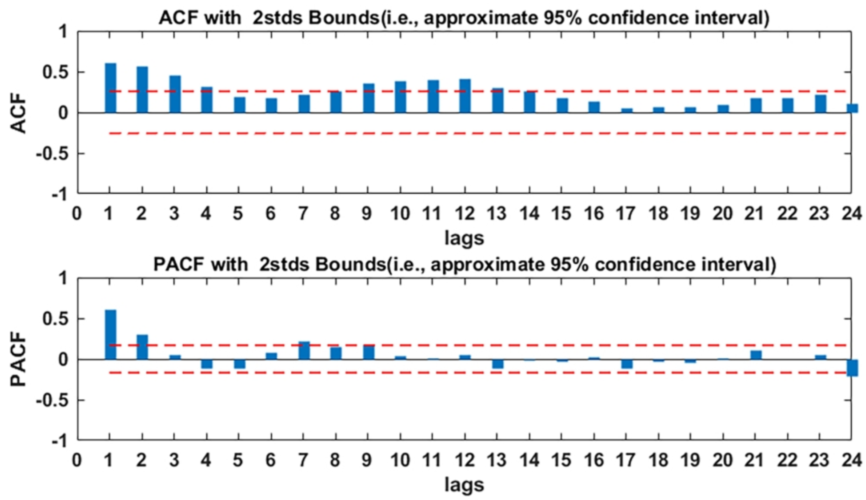
ACF with 2stds Bounds(i.e., approximate 95% confidence interval)
ACF
lags
PACF with 2stds Bounds(i.e., approximate 95% confidence interval)
PACF
lags

Figure 4

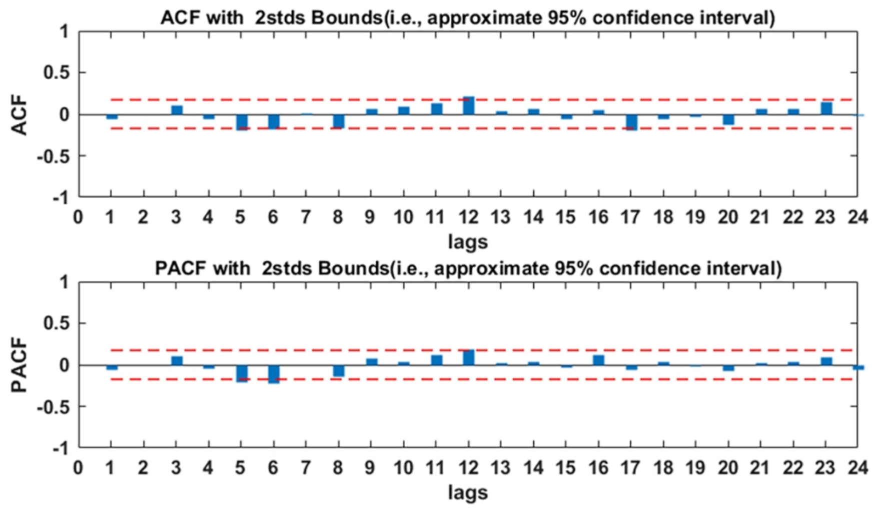
ACF with 2stds Bounds(i.e., approximate 95% confidence interval)
ACF
1
0.5
0
-0.5
-1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
lags
PACF with 2stds Bounds(i.e., approximate 95% confidence interval)
PACF
1
0.5
0
-0.5
-1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
lags

Figure 5

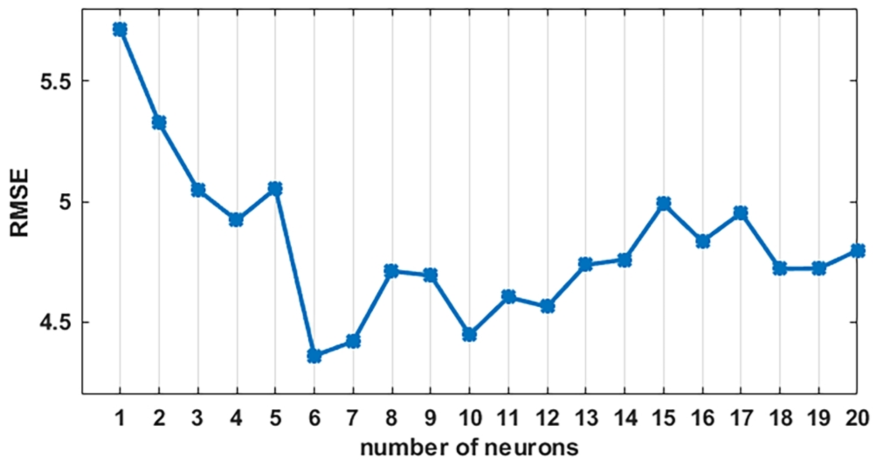
5.5
5
4.5
RMSE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
number of neurons

Figure 6

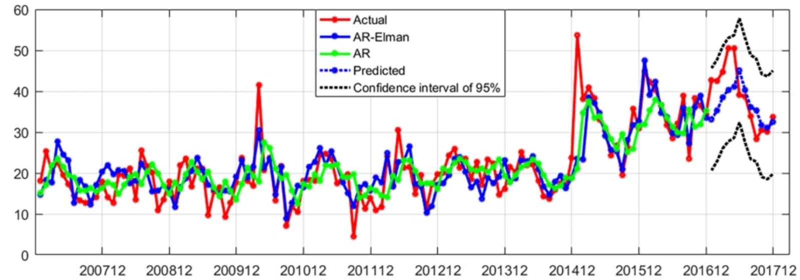
Actual
AR-Elman
AR
Predicted
Confidence interval of 95%
60
50
40
30
20
10
0
200712
200812
200912
201012
201112
201212
201312
201412
201512
201612
201712

Figure 7