

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 Predictive study of the tuberculosis incidence by time series
5
6 method combined with Elman neural network in Kashgar, China
7

8 Yanling Zheng¹, XueLiang Zhang^{1*}, XiJang Wang², Kai Wang¹, Yan Cui^{2*}

9
10 1. College of Medical Engineering and Technology, Xinjiang Medical University,
11 Urumqi 830054, People's Republic of China
12

13
14 2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,
15 Urumqi 830001, People's Republic of China
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

52 **Abstract:**
53
54

55
56 Yanling Zheng: 31372687@qq.com

57 XiJang Wang: 79057023@qq.com

58 Kai Wang: zhub168_math@sina.com,

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

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

1
2
3
4 **Objective:** Kashgar has the highest tuberculosis (TB) incidence in China. The task of
5 TB prevention and control in Kashgar is very serious. However, there is little
6 quantitative prediction study on the TB incidence in this area, therefore, it is urgent to
7 accurately predict the TB incidence in Kashgar, which can provide scientific reference
8 for the prevention and control of TB.
9
10

11
12
13 **Methods:**Based on the TB incidence in Kashgar, a single Box-Jenkins model and
14 Box-Jenkins model combined with Elman neural network model were used to do
15 prediction analysis of TB incidence in Kashgar. Root mean squared error (RMSE),
16 mean absolute error (MAE) was used to measure the prediction accuracy.
17
18
19

20
21 **Results:** The single AR((1, 2, 8)) model was established, the AIC, SC and R^2 of the
22 model were 6.53, 6.61 and 0.47 respectively. In order to improve the prediction
23 accuracy of AR ((1, 2, 8)) model, we used Elman neural network to further extract the
24 nonlinear information of AR ((1, 2, 8)) model, then, the AR ((1, 2, 8)) combined
25 Elman neural network model (AR-Elman) was established, its R^2 was 0.83. Based
26 on the AR-Elman model, the TB incidence were fitted and predicted, the fitting
27 RMSE and MAE were 3.78 and 3.38 respectively, the predicting RMSE and MAE
28 were 8.86 and 7.29 respectively.
29
30
31
32
33
34
35

36
37 **Conclusions:** Both the single AR((1,2,8)) model and AR-Elman model can be used to
38 predict the TB incidence in Kashgar, but the fitting and predicting RMSE and MAE
39 by AR-Elman model were smaller than that by AR((1,2,8)) model, and the R^2 of
40 AR-Elman model was higher than that of AR((1,2,8)) model, which indicated that
41 AR-Elman model was better than AR((1,2,8)) model. The AR-Elman model is used to
42 predict the future TB incidence in Kashgar, which can provide help for the planning
43 of health resources in this area in advance.
44
45
46
47
48
49

50 **Keywords:** Tuberculosis; Prediction; Box-Jenkins method; Elman neural network
51
52
53

54 **Strengths and limitations of this study:**

- 55 1. Kashgar has the highest tuberculosis (TB) incidence in China. However, there is
56 little quantitative prediction study on the TB incidence in this area, therefore, it is
57 urgent to do the prediction analysis of TB in Kashgar.
58
59
60

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
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.

21 **Introduction**

22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3
4 times higher in Xinjiang than that in China.

5
6 There are 14 Prefectural-Level cities in Xinjiang, China, among which Kashgar
7 has the highest TB incidence rate. From 2016 to 2017, the annual TB incidences in
8 Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7
9 times higher than that of the national level. Doing a good job in the prevention and
10 control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.
11
12

13
14
15 Mastering the changing law of the incidence of infectious diseases, using the
16 existing surveillance data to analyze, then, to predict the possible epidemic trend and
17 provide reference data for the prevention, can better control the occurrence and
18 epidemic of infectious diseases. Prediction of infectious diseases is to predict the
19 occurrence, development and epidemic trend of infectious diseases according to the
20 occurrence, development law and related factors of infectious diseases, combined
21 with analysis and judgment and mathematical model, etc. At present, the popular
22 analysis and prediction methods include time series analysis, neural network
23 prediction method and so on. Time series Box-Jenkins method has been widely used
24 in the prediction and analysis of infectious diseases in recent years⁵⁻¹⁶, it can control
25 the long-term trend, seasonality, periodicity and other factors. Neural network has
26 strong nonlinear mapping ability, in which Elman neural network is composed of
27 input layer, hidden layer, connection layer and output layer. The network has dynamic
28 memory function and is very suitable for time series prediction. At present, the
29 network is widely used in various fields, and has achieved successful prediction
30 results.¹⁷⁻²¹
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

46 The TB incidence in Kashgar, Xinjiang is almost the highest in China, and it is
47 urgent to do a good job in the prevention and control of TB in this area. Accurate
48 prediction of TB incidence is a prerequisite for prevention and control. In this study,
49 the popular Box-Jenkins time series method was used to found model for predicting
50 the TB incidences in Kashgar, in order to improve the prediction accuracy, the Elman
51 neural network with strong nonlinear information capture ability was used to
52 construct the combined model for prediction analysis.
53
54
55
56
57
58
59
60

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) model²²

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:

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

Where, X_t is the observed value at t, ϕ_1, \dots, ϕ_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:

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

Where, μ is the expected value of X_t . $\theta_1, \dots, \theta_p$ 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:

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

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 model²³

The Elman neural network is proposed by Elman in 1990. The model is generally divided into four layers: input layer, hidden layer, receiving layer and output layer. The characteristic of Elman 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, $1 \leq \alpha < 0$, $f(x)$ 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

1
2
3
4 similar, and the TB incidence from 2014 to 2017 were significantly higher than that of
5
6 previous years.

7
8 The data of TB incidences from January 2005 to December 2017 were divided
9
10 into two parts. The data from January 2005 to December 2016 were used to found
11
12 model and the data from January 2017 to December 2017 were used to test the model.

13 14 **Establishment of ARMA Model**

15
16 ARMA Modeling requires data stability, so, first of all, the stability of the
17
18 modeling part of the data is verified by ADF test. The results of ADF test showed
19
20 that Prob was less than 0.05(see Table 1), which indicated that the data was stable and
21
22 could be directly used to found the model. Secondly, the ACF and PACF diagrams of
23
24 the modeling part of the data were done (see Figure 3), It is obvious from the diagram
25
26 that the autocorrelation coefficients are trailing distribution, and the partial
27
28 correlation coefficients are almost a second-order truncated distribution, only at the
29
30 lag of 7, 8 and 9, the correlation coefficients are a little large. Based on this situation,
31
32 we considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and
33
34 AR((1,2,9)), The least square method was used to test the parameters of the four
35
36 models. The results of the test were shown in Table 2; we can see that, of the four
37
38 models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing
39
40 the two models, it was found that, the AR((1,2,8)) had smaller AIC and SC values,
41
42 and the R^2 values of AR((1,2,8))were larger than the R^2 values of AR (2), so the AR
43
44 (2) model was abandoned. The residual analysis of the AR((1,2,8)) model was carried
45
46 out. The autocorrelation and partial correlation coefficient of the residuals were
47
48 almost all within two times standard deviation, and only in the lag 5,6,12 order, they
49
50 were beyond the range of two times standard deviation (see Figure 4), which indicated
51
52 that the AR((1,2,8)) model could be used to roughly predict the TB incidences in
53
54 Kashgar. We used AR((1,2,8)) model to fit the TB incidence from September 2005 to
55
56 December 2016, the fitting RMSE and MAE were 6.15 and 4.33, respectively; we
57
58 used AR((1,2,8)) model to predict the TB incidence from January 2017 to December
59
60 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:

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

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

1
2
3
4 with the highest TB incidences in Xinjiang. Therefore, it is urgent to do a good job in
5 the prevention and control of TB in Kashgar.
6

7
8 The prediction and early warning of infectious diseases is an important link in
9 the prevention and control of infectious diseases.²⁴⁻²⁶ Therefore, this study carried
10 out research from the point of view of prediction to explore an accurate prediction
11 model and do prediction and analysis of TB incidence in Kashgar, so as to provide
12 scientific reference for the prevention and control of the disease in this area. The Box
13 - Jenkins method is a popular time series prediction method, this method has good
14 prediction performance and high prediction accuracy; Elman Neural network can
15 capture nonlinear information of time series data very well. In this study, the two
16 methods were combined to study the prediction model of TB incidence in Kashgar.
17 First of all, a single prediction model was established by Box-Jenkins method, which
18 requires the data to be stable. The probability value tested by ADF unit root was less
19 than 0.05, which indicated that the data was stable and could be directly used for
20 modeling. Based on the ACF and PACF diagrams of stationary data, we established
21 four models, then we tested the parameters of the four models by means of the least
22 square method and calculated the AIC and SC values of these models, finally, we
23 found that all parameters of AR((1,2,8)) model passed the test, and the AIC and SC of
24 AR((1,2,8)) model were the smallest, so the AR((1,2,8)) model was the best model, its
25 residual errors were basically within twice the standard deviation, which indicated
26 that the AR((1,2,8)) model was feasible. We used this model to fit the TB incidence,
27 the fitting RMSE was 6.15, the fitting MAE was 4.33, we used the model to predict
28 the TB incidence in Kashgar from January 2017 to December 2017, the predicting
29 RMSE was 10.88, the predicting MAE was 8.75. In order to improve the prediction
30 accuracy of the model, we tried to establish the AR-Elman combination model. The
31 TB incidences fitted by AR((1,2,8)) were used as the input data of Elman network,
32 and the actual incidences were used as the output data of Elman network. The Matlab
33 cyclic structure was used to find the optimal neuron from 1 to 20, when the number of
34 neurons was 6, the fitting error of the Elman neural network was the smallest;
35 therefore, the model was the best, we called the model AR-Elman combination model.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 YLZ analysed the data and wrote the manuscript. XLZ, XJW, KW and YC
4 wrote and revised the manuscript. All authors read and approved the final manuscript.
5
6

7 **Acknowledgements**

8
9 We would like to express our gratitude to anonymous peer reviewers for
10 carefully revising our manuscript and for his or her useful comments.
11
12

13 **References**

- 14
15 1. Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia:
16 prevalence, risk factors and its public health importance. *BMC Infectious Diseases*,
17 2019, 19(1).
18
19
- 20 2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity
21 pattern of *Mycobacterium tuberculosis* strains isolated from pulmonary
22 tuberculosis patients in and around Ambo Town, Central Ethiopia. *Plos One*, 2018,
23 13(2):e0193083-
24
25
- 26 3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control
27 in Debre Markos Town, Northwest Ethiopia. *Advances in Medicine*, 2018,
28 2018(2):1-8.
29
30
- 31 4. WHO. Global tuberculosis report 2019, [https://www.who.int/tb/publications/
32 global_report/en/](https://www.who.int/tb/publications/global_report/en/). (Accessed on 17 Oct 2019).
33
34
- 35 5. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of
36 tuberculosis cases reported at a tertiary hospital in Ghana: a time series model
37 approach. *BMC Public Health*.2018, 18(1):1292.
38
39
- 40 6. Tohidinik H R , Mohebalı M , Mansournia M A, et al. Forecasting zoonotic
41 cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran:
42 A SARIMA Analysis. *Tropical Medicine & International Health*, 2018,
43 23(8):860–869.
44
45
- 46 7. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in
47 Ouagadougou, Burkina Faso, 2011–2015. *Malaria Journal*, 2018, 17(1):138.
48
49
- 50 8. Martínezbello D A, Lópezquílez A, Torresprieto A. Bayesian dynamic modeling of
51 time series of dengue disease case counts. *Plos Neglected Tropical Diseases*, 2017,
52 11(7):e0005696.
53
54
55
56
57
58
59
60

- 1
2
3
4 9. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease
5 outbreak and primary healthcare delivery in Liberia: Time-series analyses for
6 2010–2016. *Plos Medicine*, 2018, 15(2):e1002508.
7
8
- 9
10 10. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth
11 disease in Nanjing, China: a time-series model based on weather variability.
12 *International Journal of Biometeorology*, 2017(6):1-10.
13
14
- 15
16 11. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in
17 Kumasi: Using ARIMA models to forecast future incidence. *Cogent Social*
18 *Sciences*, 2018, 4(1):1461544.
19
20
- 21
22 12. Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving
23 average model and generalised regression neural network model for prediction of
24 haemorrhagic fever with renal syndrome in China: a time-series study. *Bmj Open*,
25 2019, 9(6):e025773.
26
27
- 28
29 13. Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in
30 children younger than 48 months in Suzhou using a seasonal autoregressive
31 integrated moving average model. *Bmj Open*, 2019, 9(1).
32
33
- 34
35 14. Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable
36 incidence time series. *Epidemiology & Infection*, 2019, 147
37
38
- 39
40 15. Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand,
41 foot and mouth disease cases in China from 2008 to 2018. *Epidemiology &*
42 *Infection*, 2019, 147
43
44
- 45
46 16. Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in
47 modelling and forecasting emergency department visits in a medical centre in
48 Southern Taiwan[J]. *Bmj Open*, 2017, 7(11):e018628
49
50
- 51
52 17. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum
53 Analysis for wind speed forecasting based on Elman Neural Network. *Energy*
54 *Conversion & Management*, 2017, 148:895-904.
55
56
- 57
58 18. Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for
59 Forecasting China's Carbon Dioxide Emissions. *Communications in Computer*
60 *and Information Science*, 2018(951): 36-47.

- 1
2
3
4 19. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural
5
6 Network for Disease Prediction. *Arabian Journal for Science and Engineering*,
7
8 2019, 44(11): 9209–9220.
- 9
10 20. Mehrgini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman
11
12 artificial neural network. *Carbonates & Evaporites*, 2017(6):1-11.
- 13
14 21. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the
15
16 application of disease modifying therapy for sickle cell using classification
17
18 models. *Neurocomputing*, 2017, 228(C):154-164.
- 19
20 22. Box G E, Jenkins G M. *Time series analysis: forecasting and control rev. ed.*
21
22 Oakland, California, Holden-Day, 1976, 31(4):238-242.
- 23
24 23. Ming Cheng. *The principle and example of MATLAB neural network.* Tsinghua
25
26 University Press, 2013.
- 27
28 24. Huppert A, Katriel G. Mathematical modelling and prediction in infectious
29
30 disease epidemiology. *Clin Microbiol Infect.* 2013, 19(11):999-1005
- 31
32 25. Wearing HJ1, Rohani P, Keeling MJ. Appropriate models for the management of
33
34 infectious diseases. *PLoS Med.* 2005, 2(7):e174.
- 35
36 26. Neuberger A, Paul M, Nizar A, Raoult D. Modelling in infectious diseases:
37
38 between haphazard and hazard. *Clin Microbiol Infect.* 2013, 19(11):993-8.
- 39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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
	C	21.42	2.17	9.86	<0.01		
AR (2)	AR(1)	0.42	0.08	5.20	<0.01	6.55	6.62
	AR(2)	0.34	0.08	4.17	<0.01		
	C	22.93	3.73	6.14	<0.00		
AR((1, 2, 7))	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		

	C	23.53	4.56	5.16	<0.01		
AR((1, 2, 8))	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		
	C	29.07	15.53	1.87	0.06		
AR((1, 2, 9))	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		
	C						

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

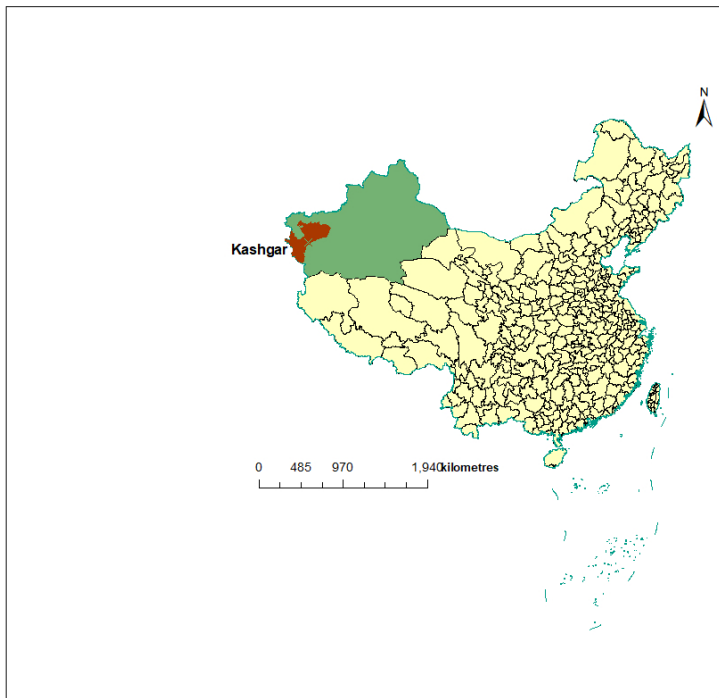


Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China.

296x210mm (96 x 96 DPI)

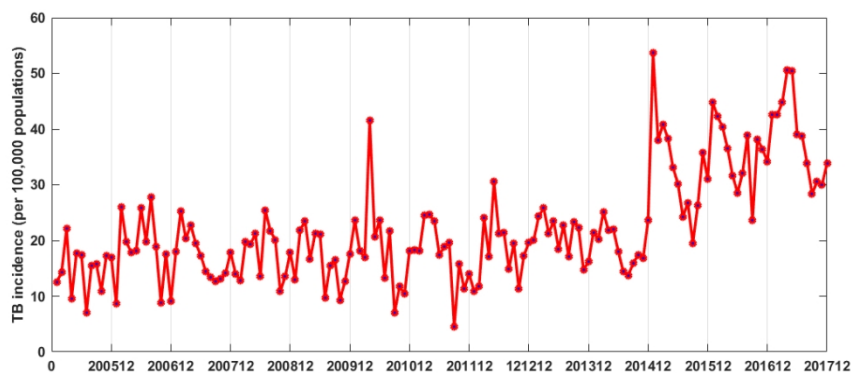


Figure 2. Graph of the TB incidences in Kashgar from January 2005 to December 2017

320x128mm (96 x 96 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

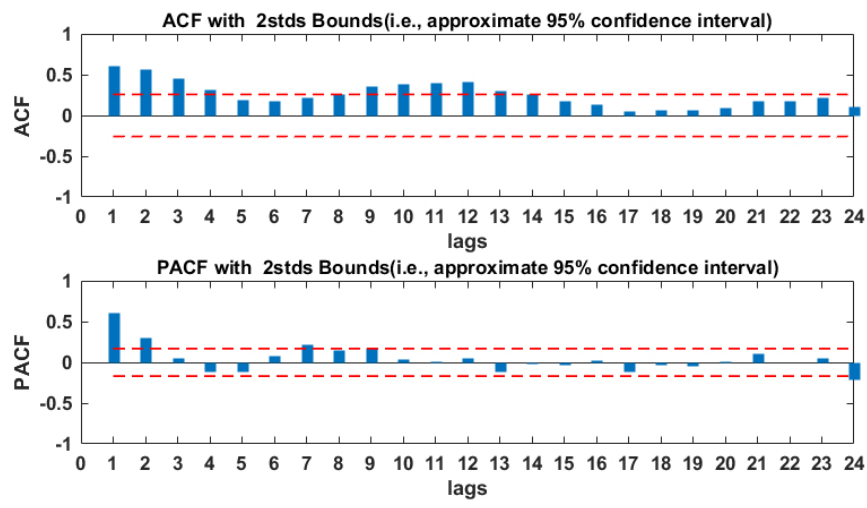


Figure 3. The ACF and PACF diagrams of modeling data

233x121mm (96 x 96 DPI)

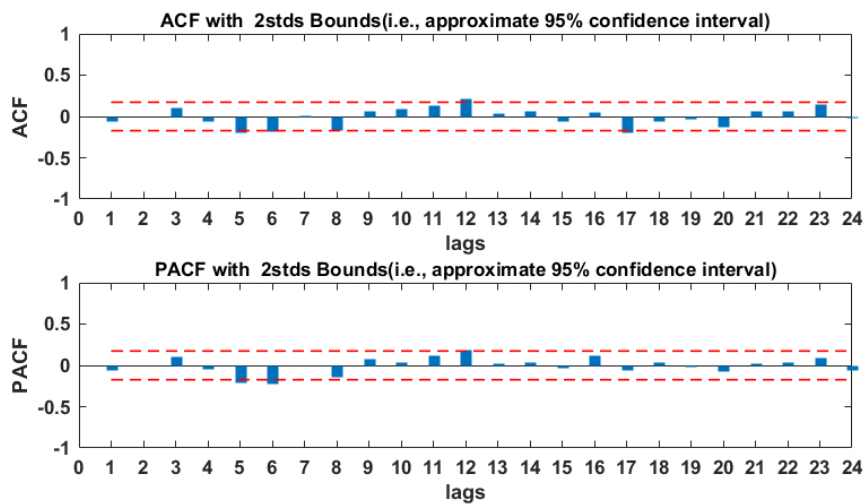


Figure 4. The ACF and PACF graphs of residual errors of AR ((1,2,8)) Model

233x122mm (96 x 96 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

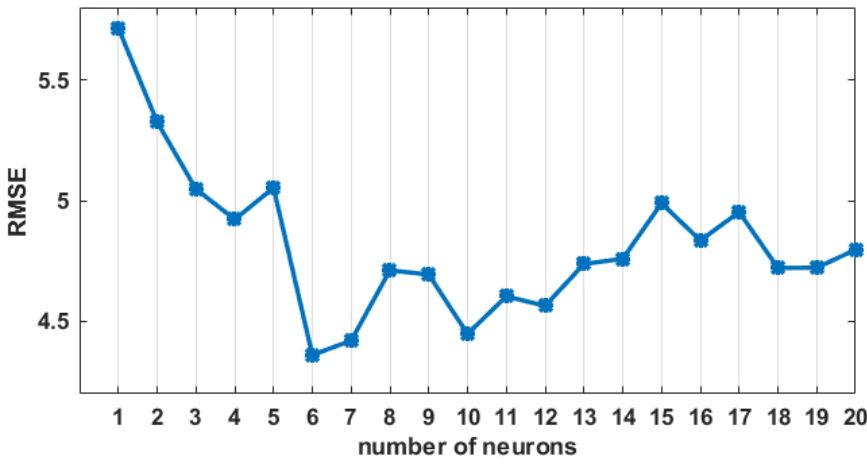


Figure 5. The numbers of neurons in AR-Elman Model and the corresponding RMSE Values

196x95mm (96 x 96 DPI)

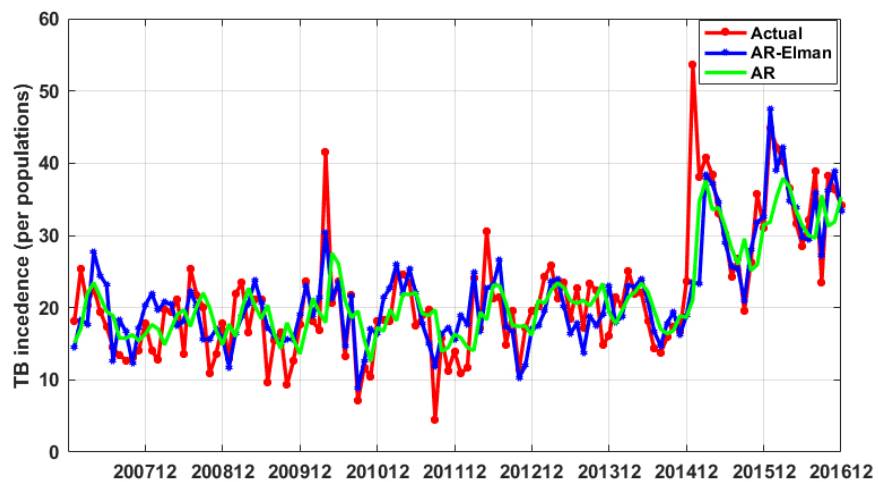


Figure 6. Fitting comparison Diagram of AR ((1,2,8)) Model and AR-Elman Model

233x123mm (96 x 96 DPI)

BMJ Open

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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 Predictive study of tuberculosis incidence by time series method
5
6
7 and Elman neural network in Kashgar, China

8
9 **Yanling Zheng¹, XueLiang Zhang^{1*}, XiJiang Wang², Kai Wang¹, Yan Cui^{2*}**

10
11 1. College of Medical Engineering and Technology, Xinjiang Medical University,
12
13 Urumqi 830054, People's Republic of China

14
15 2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,
16
17 Urumqi 830001, People's Republic of China

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

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

1 **Abstract:**

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

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

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

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

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

28 29 **Strengths and limitations of this study:**

- 1
2
3
4 30 1. The incidence of tuberculosis (TB) is very high in Kashgar, therefore, it is urgent to
5
6 31 do the prediction analysis of TB in Kashgar.
- 7 32 2. In the study, AR-Elman model with high prediction accuracy was established, it
8
9 33 can be used to predict the development of TB in Kashgar, China.
- 10 34 3. This study did not consider possible influencing factors related to TB, such as
11
12 35 climatic factors, environmental factors, demographic factors and political
13
14 36 issues,etc.

15 37 **Introduction**

18 38 Tuberculosis (TB) is still a major global public health problem and the ninth
19
20 39 leading cause of death in the world.¹⁻³ Because TB is contagious, patient resistance is
21
22 40 low, treatment time is long, patients' labor force is lost and their contribution to
23
24 41 society is reduced, so TB brings great burden to society and economy. All countries in
25
26 42 the world are working hard to fight tuberculosis. In 2018, the number of new reported
27
28 43 TB cases was about 10 million; this figure has remained relatively stable in recent
29
30 44 years. The latest treatment results showed that the global TB treatment success rate
31
32 45 was 83%. World Health Organization (WHO) has set targets for the stop TB strategy.
33
34 46 The targets mentioned that by 2030, on the basis of the work in 2015, TB deaths will
35
36 47 reduce by 90%, and annual new TB cases will reduce by 80%.⁴ In order to achieve
37
38 48 these goals, TB prevention and control services must be provided in the broad context
39
40 49 of universal health coverage, joint action must be taken to address the social and
41
42 50 economic consequences of TB, and technological breakthroughs should be achieved
43
44 51 by 2025 to make the TB incidence decline faster than that at any time in history.
45
46 52 According to the global tuberculosis report 2019⁴, China has the second highest
47
48 53 number of TB cases in the world.⁴ The TB incidence in western China was much
49
50 54 higher than that in eastern and central China. The province with the highest TB
51
52 55 incidence in the west is Xinjiang province. From 2016 to 2017, the annual TB
53
54 56 incidence in Xinjiang was 185.66/100,000 and 202.58/100,000, while the annual TB
55
56 57 incidence from 2016 to 2017 in China was 61/100,000 and 60.53/100,000,
57
58 58 respectively, it is nearly three times higher in Xinjiang than that national level.

59 59 There are 14 Prefectural-Level cities in Xinjiang, China, among which Kashgar
60

1
2
3
4 60 has a very high TB incidence rate. From 2016 to 2017, the annual TB incidences in
5
6 61 Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7
7
8 62 times higher than that of the national level. Doing a good job in the prevention and
9
10 63 control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.

11 64 Mastering the changing law of the incidence of infectious diseases, using the
12
13 65 existing surveillance data to analyze, then, to predict the possible epidemic trend and
14
15 66 provide reference data for the prevention, can better control the occurrence and
16
17 67 epidemic of infectious diseases. Prediction of infectious diseases is to predict the
18
19 68 occurrence, development and epidemic trend of infectious diseases according to the
20
21 69 occurrence, development law and related factors of infectious diseases, based on
22
23 70 analysis, judgment and mathematical model, etc.

24
25 71 For study of quantitative prediction of infectious diseases, there are many
26
27 72 methods, such as grey prediction method ⁵, exponential smoothing prediction method
28
29 73 ^{6,7}, dynamic model prediction method ⁸, Box-Jenkins method ⁹, neural network
30
31 74 method ¹⁰, etc., with the deepening of prediction research, more and more scholars
32
33 75 like to use the Box-Jenkins method ¹¹⁻²¹, there are many different models in this
34
35 76 method, and if appropriate models are established according to the characteristics of
36
37 77 time series, high prediction ability often can be obtained. Neural network has strong
38
39 78 nonlinear mapping ability, in which Elman neural network is composed of input layer,
40
41 79 hidden layer, connection layer and output layer. The Elman network has dynamic
42
43 80 memory function, and it is very suitable for time series prediction. At present, the
44
45 81 Elman network is widely used in various fields, and has achieved successful
46
47 82 prediction results. ²²⁻²⁶ Sometimes, the prediction effect of a single model is not ideal,
48
49 83 in order to further improve the prediction accuracy, many studies adopt the combined
50
51 84 model prediction method ²⁷⁻²⁹, the combined model can absorb the advantages of two
52
53 85 or more methods so as to achieve a higher prediction accuracy.

54 86 The TB incidence of Kashgar is very high, and it is urgent to do a good job in the
55
56 87 prevention and control of TB in this area. Accurate prediction of TB incidence is a
57
58 88 prerequisite for prevention and control, which can help advance resource planning and
59
60 89 policy formulation. In this study, the popular Box-Jenkins time series method was

1
2
3
4 90 used to build model for predicting the TB incidence in Kashgar, in order to improve
5
6 91 the prediction accuracy, the Elman neural network with strong nonlinear information
7
8 92 capture ability was used to construct the combined model for prediction analysis.

93 **Materials and Methods**

94 **Study area and data Sources**

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

102 **Patient and public involvement**

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

105 **Autoregressive moving average (ARMA) model**

106 ARMA model³⁰ is an important time series analysis and prediction model in
107 Box-Jenkins method, also known as auto-regression moving average model. The
108 ARMA (p,q) model is a model with autocorrelation order p and moving average order
109 q, p and q are judged by autocorrelation function (ACF) and partial autocorrelation
110 function (PACF) diagram of stationary data. If the original data is stable, the
111 autocorrelation coefficients are trailing, and the partial correlation coefficients are the
112 p-order truncated, then q=0, the ARMA(p,q) model is recorded as AR(p), and its
113 expression is as follows:

$$114 \quad X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t,$$

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

1
2
3
4 118 If the original data is stable, the autocorrelation coefficients are q-order truncated
5
6 119 and the partial correlation coefficients are trailing, then $p=0$ and the ARMA(p,q)
7
8 120 model becomes MA(q), its expression is as follows:

9
10 121
$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$

11
12 122 where, μ is the expected value of X_t . $\theta_1, \dots, \theta_p$ are model parameters.

13
14 123 If the original data is stable, the autocorrelation coefficients are trailing, and the
15
16 124 partial correlation coefficients are also trailing, then the expression of the ARMA (p, q)
17
18 125 model is as follows:

19
20 126
$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}.$$

21
22 127 There are four main steps in ARMA modeling:

23
24 128 First step. The prerequisite for ARMA modeling is the stationary of time series.
25
26 129 Check whether the data is stable by Augmented Dickey-Fuller (ADF) unit root test. In
27
28 130 this study, the significant level probability (p-value) is 0.05, and if the p-value is less
29
30 131 than 0.05, then, the data is stable. By observing the autocorrelation function (ACF)
31
32 132 and partial autocorrelation function (PACF) of the stable data, we can determine the
33
34 133 possible values of p and q and establish the possible ARMA (p, q) model.

35
36 134 Second step. The parameters of ARMA(p, q) model are estimated by maximum
37
38 135 likelihood estimation or least square estimation, and the model parameters are tested.
39
40 136 If p-value is less than 0.05, the parameters have statistical significance. The best
41
42 137 model is determined according to the value of Akaike information criterion (AIC),
43
44 138 Schwarz criterion (SC) and Goodness of Fit (R^2) of model. The smaller AIC and SC
45
46 139 are, the larger the R^2 is, and the better the model is.

47
48 140 Third step. To determine whether the established ARMA (p, q) model is suitable.
49
50 141 The residual sequence of a suitable model shall be the white noise process, and its
51
52 142 ACF and PACF coefficients should be within twice the standard deviation range,
53
54 143 otherwise, it is considered that the extraction information of the established model is
55
56 144 not sufficient, and it is necessary to consider improving the accuracy of the model.

57
58 145 Fourth step. Using the established model to do prediction and analysis.

59
60 146 **Elman neural network model**

1
2
3
4 147 The Elman neural network is proposed by Elman in 1990³¹. The model is
5
6 148 generally divided into four layers: input layer, hidden layer, receiving layer and output
7
8 149 layer. The characteristic of Elman network is that the output of the hidden layer is
9
10 150 connected to the input of the hidden layer through the delay and storage of the
11
12 151 receiving layer, which makes it sensitive to the data of the historical state. The
13
14 152 addition of the internal feedback network increases the ability of the network itself to
15
16 153 deal with dynamic information, thus achieving the purpose of dynamic modeling.

17 154 The mathematical structure of the Elman neural network is as follows:

18
19 155
$$x(k)=f(w^1x_c(k)+ w^2u(k-1))$$

20
21 156
$$x_c(k)= \alpha x_c(k-1)+ x(k-1)$$

22
23 157
$$y(k)=g(w^3x(k))$$

24
25 158 Where, w^1 is the connection weight matrix between the contact unit and the
26
27 159 hidden layer unit, w^2 is the connection weight matrix between the input unit and the
28
29 160 hidden layer unit, w^3 is the connection weight matrix between the hidden layer unit
30
31 161 and the output unit. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the
32
33 162 hidden layer unit, respectively, $y(k)$ represents the output of the output unit, α is a
34
35 163 self-connected feedback gain factor, $0 \leq \alpha < 1$, $f(x)$ often takes the sigmoid function

36 164 There are four main steps in Elman neural network modeling:

37 165 First step. Data standardization processing. Data standardization is scaling the
38
39 166 data to a small specific interval. In order to remove the unit limit of the data and
40
41 167 convert it into dimensionless pure value, it is convenient for the index of different
42
43 168 units or order of magnitude to be compared and weighted. In our study, we use
44
45 169 function package `mapminmax()` to standardize the data, standardized data is in [-1,1]
46
47 170 range.

48
49 171 Second step. Determine the input layer, the output layer. Generally, the input and
50
51 172 output layer are determined according to the characteristics of the data and the needs
52
53 173 of the analysis.

54
55 174 Third step. Set the parameters of the Elman model, such as training epochs and
56
57 175 goals, etc., in our study, training epochs and goals of Elman neural network were set
58
59 176 to 2000 and 0.00001, respectively. To determine the number of neurons in the hidden
60
177 layer, so that the error of the established Elman model is minimized. At present, there
178 is no ideal analytical expression for the number of neurons in the hidden layer. The
179 number of neurons has a great influence on the performance of the network. When the

180 number of neurons is too large, it will lead to that the network learning time is too
 181 long, the generalization performances are not good, and even the convergence failure,
 182 but when the number of neurons is too small, the fault-tolerant ability of the network
 183 is poor. In general, the number of neurons does not exceed 20.³² In this study, matlab
 184 cyclic structure was used to find the optimal number of neurons by comparing the
 185 RMSE values of Elman networks with neurons 1 to 20.

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

188 **Model comparison measures**

189 Three performance indexes, root mean square error (RMSE), mean absolute error
 190 (MAE) and mean absolute percentage error (MAPE) were used to assess the fitting
 191 and forecasting accuracy of two models. The smaller these three values are, the better
 192 the model is. Their expressions are as follows¹⁰:

$$193 \quad RMSE = \sqrt{\frac{\sum_{t=1}^n (X_t - \hat{X}_t)^2}{n}}$$

$$194 \quad MAE = \frac{\sum_{t=1}^n |X_t - \hat{X}_t|}{n},$$

$$195 \quad MAPE = \frac{\sum_{t=1}^n \left| \frac{X_t - \hat{X}_t}{X_t} \right| \times 100}{n},$$

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

198 **Statistical software**

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

200 **Results**

201 From January 2005 to December 2017, the number of reported TB cases was
 202 141984 in Kashgar, Xinjiang, the average annual TB cases were 7888 and the average
 203 annual incidence was 191.18. Figure 2 showed the time series graph of the TB
 204 incidence. It can be seen from the Figure 2 that the curve of TB incidence showed
 205 strong nonlinear characteristics from 2005 to 2014, and the TB incidence from 2015
 206 to 2017 were significantly higher than that of previous years.

207 The data of TB incidence from January 2005 to December 2017 were divided

208 into two parts. The data from January 2005 to December 2016 were used to build
209 model and the data from January 2017 to December 2017 were used to test the model.

210 **Establishment of ARMA Model**

211 ARMA Modeling requires data stability, so, first of all, the stability of the
212 modeling part of the data was verified by ADF test. The results of ADF test showed
213 that p-value was less than 0.05(see Table 1), which indicated that the data was stable
214 and could be directly used to build the model. Secondly, the ACF and PACF graphs
215 of the modeling data were plotted (see Figure 3), it was obvious from Figure 3 that the
216 autocorrelation coefficients were trailing distribution, and the partial correlation
217 coefficients were almost a second-order truncated distribution, only at the lag 7, 8 and
218 9, the correlation coefficients were a little large. Based on this situation, we
219 considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and
220 AR((1,2,9)). The least square method was used to test the parameters of the four
221 models. The results of the test were shown in Table 2; we can see that, of the four
222 models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing the
223 two models, it was found that, the AR((1,2,8)) model had smaller AIC and SC values,
224 and the R^2 values of AR((1,2,8)) were larger than the R^2 values of AR (2), so the AR
225 (2) model was abandoned. Then, the residual analysis of the AR((1,2,8)) model was
226 carried out, the autocorrelation and partial correlation coefficient of the residuals were
227 almost all within two times standard deviation, and only in the lag 5,6,12, they were
228 beyond the range of two times standard deviation (see Figure 4), which indicated that
229 the AR((1,2,8)) model could be used to roughly predict the TB incidence in Kashgar.
230 We used AR((1,2,8)) model to fit the TB incidence from September 2005 to
231 December 2016, the fitting RMSE ,MAE and MAPE were 6.15, 4.33 and 0.2858,
232 respectively; we used AR((1,2,8)) model to predict the TB incidence from January
233 2017 to December 2017, the prediction RMSE, MAE and MAPE were 10.88 , 8.75
234 and 0.2029, respectively.

235 **Establishment of AR-Elman Model**

236 In order to improve the prediction accuracy of the AR((1,2,8)) model, we tried

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

$$244 \quad \text{input matrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_i \\ x_2 & x_3 & \dots & x_{i+1} \\ \dots & \dots & \dots & \dots \\ x_N & x_{N+1} & \dots & \dots \end{bmatrix}, \text{output matrix} = [x_{N+1} \quad x_{N+2} \quad \dots \quad x_{N+i}]$$

245 We selected twelve as the number of input layers of AR-Elman network and
 246 one as the number of output layers representing the forecast value. By the matlab
 247 cyclic structure, we selected the optimal number of neurons between 1 and 20, and
 248 finally we found when the number of neurons was 6 (see Figure 5), the RMSE was the
 249 smallest, and the AR-Elman was optimal. We used the AR-Elman model to fit the
 250 training data, RMSE was 3.78, MAE was 3.38, MAPE was 0.1837, and the R^2 of the
 251 model was 0.83; we used the AR-Elman model to predict the TB incidence from
 252 January 2017 to December 2017, RMSE was 8.86, MAE was 7.29, and MAPE was
 253 0.2006. The fitting graph of AR((1,2,8)) model and AR-Elman model was shown in
 254 Figure 6. Comparison results of the AR((1,2,8)) model and AR-Elman model was
 255 shown in Table 3, both the fitting RMSE, MAE and MAPE and the predicting RMSE,
 256 MAE and MAPE of the AR-Elman model were smaller than those of the single
 257 AR((1,2,8)) model, which indicated that the AR-Elman combined model established
 258 in this study was more suitable for predicting the TB incidence in Kashgar.

259 Discussion

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

1
2
3
4 268 the high TB incidence in Xinjiang. Therefore, it is urgent to do a good job in the
5
6 269 prevention and control of TB in Kashgar.

7
8 270 The prediction and early warning of infectious diseases is an important link in
9
10 271 the prevention and control of infectious diseases.³²⁻³⁴ Therefore, this study carried out
11
12 272 research from the point of view of prediction to explore an accurate prediction model
13
14 273 and do prediction analysis of TB incidence in Kashgar, so as to provide scientific
15
16 274 reference for the prevention and control of the disease in this area. The Box -Jenkins
17
18 275 method is a popular time series prediction method, this method has good prediction
19
20 276 performance and high prediction accuracy; Elman Neural network can capture
21
22 277 nonlinear information of time series data very well. In this study, the two methods
23
24 278 were combined to study the prediction model of TB incidence in Kashgar.

25
26 279 Many studies have found that Box-Jenkins method has a good ability of fitting
27
28 280 and forecasting. For stationary time series that do not contain seasonality, it is more
29
30 281 suitable to use the ARMA model of the Box-Jenkins method to do prediction analysis
31
32 282 ³⁵, for non-stationary time series of infectious diseases with obvious seasonality, it is
33
34 283 more suitable to use seasonal autoregressive integrated moving average (SARIMA)
35
36 284 model of the Box-Jenkins method for prediction analysis.⁹⁻¹² In our study, from Figure
37
38 285 2, we could see that the seasonality of the TB incidence in Kashgar from 2005 to 2014
39
40 286 was not obvious, there was only a certain seasonality from 2015 to 2017, and we
41
42 287 found that the time series of TB incidence was stable by ADF unit root test, and the
43
44 288 autocorrelation and partial correlation coefficients of modeling data at lag 12, 24 were
45
46 289 not obviously large, therefore, for our research data, we used ARMA model to do
47
48 290 forecast analysis, finally, we established AR((1,2,8)) model of the Box-Jenkins
49
50 291 method, it has a good performance to fit and predict the TB incidence of Kashgar in
51
52 292 Xinjiang. From Figure 2, we can also see that the time series of TB incidence has
53
54 293 strong non-linear, the established AR((1,2,8)) model mainly extracted the linear
55
56 294 information of data, considering that the neural network can capture the non-linear
57
58 295 information of data well, in order to improve the prediction accuracy of TB incidence
59
60 296 rate in Kashgar, we used AR((1,2,8)) model and Elman neural network model to
297 establish AR-Elman hybrid model. Many studies have found that the combination
298 model can improve the accuracy of prediction, such as, Wang et al.²⁸ found that
299 SARIMA-NAR hybrid model has an outstanding ability to improve the prediction
300 accuracy relative to SARIMA model and nonlinear autoregressive network (NAR)

1
2
3
4 301 model when they were used to predict pertussis incidence in China. Li et al.²⁷ found
5 302 ARIMA-GRNN hybrid model was shown to be superior to the single ARIMA model
6
7 303 in predicting the short-term TB incidence in the Chinese population. Our research was
8
9 304 consistent with these literatures that our AR-Elman hybrid model was more accurate
10 305 than the single AR((1,2,8)) model.

11
12 306 The incidence of tuberculosis in Kashgar, Xinjiang is very high, and the relevant
13 307 departments of disease prevention and control in Xinjiang have also done a lot of
14 308 effective work. Our research was mainly to build a high-precision prediction model to
15 309 help early warning of tuberculosis in Kashgar. Finally, we established the AR-Elman
16 310 hybrid model, which had high fitting and prediction accuracy of TB incidence in
17 311 Kashgar, Xinjiang. Because the development of any event may be affected by many
18 312 factors, such as social, economic, political, demographic factors, so the long-term
19 313 forecast accuracy of the AR-Elman hybrid model may decline.

24 314 **Conclusions**

25
26
27 315 Kashgar has a very high TB incidence, in order to provide some help for the
28 316 prevention and control of this disease in this area, the prediction problem of the TB
29 317 incidence was studied. Firstly, a single AR((1,2,8)) prediction model was established
30 318 by using Box-Jenkins method, and its fitting and prediction performance were good,
31 319 secondly, in order to improve the prediction accuracy of the single AR((1,2,8)) model,
32 320 we used single AR((1,2,8)) and Elman neural network with strong ability to capture
33 321 nonlinear information to establish AR-Elman hybrid model. The fitting and prediction
34 322 accuracy of the hybrid model was higher than that of the single AR((1,2,8)) model.
35 323 The AR-Elman hybrid model can provide scientific help for predicting and warning
36 324 the TB incidence in Kashgar, Xinjiang. This study also has some limitations, we did
37 325 not consider possible influencing factors related to TB, such as climatic factors and
38 326 political issues, etc, in further studies, we will consider these factors.

37 327 **Competing interests**

38 328 The authors declare no conflict of interest.

39 329 **Funding**

40 330 This work was supported by Major projects of science and technology in
41 331 Xinjiang Autonomous region (Grant No.2017A03006), China, and National Natural
42 332 Science Foundation of China Regional Science Foundation Project (Grant No.
43 333 72064036, 82060609),and State Key Laboratory of Pathogenesis, Prevention and
44 334 Treatment of High Incidence Diseases in Central Asia Fund (Grant No.

1
2
3
4 335 SKL-HIDCA-2020-9)

5 336 **Data sharing statement**

6
7 337 No additional data are available.

8 338 **Author contributions**

9
10 339 YLZ analyzed the data and wrote the manuscript. XLZ, XJW, KW and YC wrote
11 and revised the manuscript. All authors read and approved the final manuscript.

12 340
13 341 **Acknowledgements**

14
15 342 We would like to express our gratitude to anonymous peer reviewers for
16 343 carefully revising our manuscript and for his or her useful comments.

17
18 344 **References**

19
20 345 1. Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia:
21 prevalence, risk factors and its public health importance. BMC Infectious Diseases,
22 346 2019; 19(1).
23 347

24
25 348 2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity
26 pattern of Mycobacterium tuberculosis strains isolated from pulmonary
27 349 tuberculosis patients in and around Ambo Town, Central Ethiopia. Plos One, 2018;
28 350 13(2):e0193083-.
29 351

30
31 352 3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control
32 in Debre Markos Town, Northwest Ethiopia. Advances in Medicine,
33 353 2018;2018(2):1-8.
34 354

35 355 4. WHO. Global tuberculosis report 2019, [https://www.who.int/tb/publications/
36 global_report/en/](https://www.who.int/tb/publications/global_report/en/). (Accessed on 17 Oct 2019).
37 356

38
39 357 5. Zhao YF, Shou MH, Wang ZX. Prediction of the Number of Patients Infected with
40 COVID-19 Based on Rolling Grey Verhulst Models. Int J Environ Res Public
41 358 Health. 2020; 17(12):4582.
42 359

43
44 360 6. Guan P, Wu W, Huang D. Trends of reported human brucellosis cases in mainland
45 China from 2007 to 2017: an exponential smoothing time series analysis. Environ
46 361 Health Prev Med. 2018; 23(1):23.
47 362

48
49 363 7. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, Xu JY, Shi J, Sun DY.
50 Analysis and predication of tuberculosis registration rates in Henan Province,
51 364 China: an exponential smoothing model study. Infect Dis Poverty. 2020;9(1):123.
52 365

- 1
2
3
4 366 8. Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling
5 367 of time series of dengue disease case counts. *PLoS Negl Trop Dis*.
6 368 2017;11(7):e0005696.
- 9 369 9. Wang Y, Xu C, Zhang S, Wang Z, Zhu Y, Yuan J. Temporal trends analysis of
11 370 human brucellosis incidence in mainland China from 2004 to 2018. *Sci Rep*.
13 371 2018;8(1):15901.
- 15 372 10. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S. An Advanced Data-Driven
17 373 Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series
19 374 Forecasting in Qinghai Province, China. *Infect Drug Resist*. 2020;13:867-880.
- 21 375 11. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of
23 376 tuberculosis cases reported at a tertiary hospital in Ghana: a time series model
25 377 approach. *BMC Public Health*.2018;18(1):1292.
- 27 378 12. Tohidinik H R , Moheballi M , Mansournia M A, et al. Forecasting zoonotic
29 379 cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran:
31 380 A SARIMA Analysis. *Tropical Medicine & International Health*,
33 381 2018;23(8):860–869.
- 35 382 13. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in
37 383 Ouagadougou, Burkina Faso, 2011–2015. *Malaria Journal*, 2018;17(1):138.
- 39 384 14. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease
41 385 outbreak and primary healthcare delivery in Liberia: Time-series analyses for
43 386 2010–2016. *Plos Medicine*, 2018; 15(2):e1002508.
- 45 387 15. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth
47 388 disease in Nanjing, China: a time-series model based on weather variability.
49 389 *International Journal of Biometeorology*, 2017;(6):1-10.
- 51 390 16. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in
53 391 Kumasi: Using ARIMA models to forecast future incidence. *Cogent Social*
55 392 *Sciences*, 2018; 4(1):1461544.
- 57 393 17. Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving
59 394 average model and generalised regression neural network model for prediction of
61 395 haemorrhagic fever with renal syndrome in China: a time-series study. *BMJ Open*,

- 1
2
3
4 396 2019; 9(6):e025773.
- 5
6 397 18.Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in
7
8 398 children younger than 48 months in Suzhou using a seasonal autoregressive
9
10 399 integrated moving average model. *BMJ Open*, 2019; 9(1).
- 11
12 400 19.Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable
13
14 401 incidence time series. *Epidemiology & Infection*, 2019; 147.
- 15
16 402 20.Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand,
17
18 403 foot and mouth disease cases in China from 2008 to 2018. *Epidemiology &*
19
20 404 *Infection*, 2019; 147.
- 21
22 405 21.Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in
23
24 406 modelling and forecasting emergency department visits in a medical centre in
25
26 407 Southern Taiwan. *Bmj Open*, 2017; 7(11):e018628.
- 27
28 408 22. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum
29
30 409 Analysis for wind speed forecasting based on Elman Neural Network. *Energy*
31
32 410 *Conversion & Management*, 2017; 148:895-904.
- 33
34 411 23.Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for
35
36 412 Forecasting China's Carbon Dioxide Emissions. *Communications in Computer*
37
38 413 *and Information Science*, 2018;(951): 36-47.
- 39
40 414 24. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural
41
42 415 Network for Disease Prediction. *Arabian Journal for Science and Engineering*,
43
44 416 2019;44(11): 9209–9220.
- 45
46 417 25. Mehrgini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman
47
48 418 artificial neural network. *Carbonates & Evaporites*, 2017;(6):1-11.
- 49
50 419 26. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the
51
52 420 application of disease modifying therapy for sickle cell using classification
53
54 421 models. *Neurocomputing*, 2017; 228(C):154-164.
- 55
56 422 27. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, Ji Y, Xu D, Wang J. Application of a
57
58 423 hybrid model in predicting the incidence of tuberculosis in a Chinese population.
59
60 424 *Infect Drug Resist.* 2019;12:1011-1020.
- 425 28. Wang Y, Xu C, Wang Z, Zhang S, Zhu Y, Yuan J. Time series modeling of

- 1
2
3
4 426 pertussis incidence in China from 2004 to 2018 with a novel wavelet based
5
6 427 SARIMA-NAR hybrid model. PLoS One. 2018;13(12):e0208404.
7
8 428 29. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, Yuan J. Temporal trends
9
10 429 analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a
11
12 430 new SARIMA-NARNNX hybrid model. BMJ Open. 2019;9(7):e024409.
13
14 431 30. Box G E, Jenkins G M. Time series analysis: forecasting and control rev. ed.
15
16 432 Oakland, California, Holden-Day, 1976; 31(4):238-242.
17
18 433 31. Ming Cheng. The principle and example of MATLAB neural network. Tsinghua
19
20 434 University Press, 2013.
21
22 435 32. Huppert A, Katriel G. Mathematical modelling and prediction in infectious
23
24 436 disease epidemiology. Clin Microbiol Infect. 2013;19(11):999-1005
25
26 437 33. Wearing HJ1, Rohani P, Keeling MJ. Appropriate models for the management of
27
28 438 infectious diseases. PLoS Med. 2005; 2(7):e174.
29
30 439 34. Neuberger A, Paul M, Nizar A, Raoult D. Modelling in infectious diseases:
31
32 440 between haphazard and hazard. Clin Microbiol Infect. 2013;19(11):993-8.
33
34 441 35. Tipirneni-Sajja A, Krafft AJ, Loeffler RB, Song R, Bahrami A, Hankins JS,
35
36 442 Hillenbrand CM. Autoregressive moving average modeling for hepatic iron
37
38 443 quantification in the presence of fat. J Magn Reson Imaging.
39
40 444 2019;50(5):1620-1632.
41
42 445

446 **Figures**

- 447 **Figure 1.** The red part of this picture is the location of kashgar in Xinjiang, China.
448 Kashgar is located in the south of Xinjiang, and it has a very high incidence of
449 tuberculosis.
450
451 **Figure 2.** Graph of the tuberculosis (TB) incidence in Kashgar from January 2005 to
452 December 2017. The curve of TB incidence showed strong nonlinear characteristics
453 from 2005 to 2014, and the TB incidence increased significantly from 2015 to 2017.

1
2
3
4 454

5 455 **Figure 3.** Autocorrelation function (ACF) and partial autocorrelation function (PACF)
6
7 456 graphs of modeling data. As the delay of the lag order, the autocorrelation coefficients
8
9 457 were trailing and the partial correlation coefficients were truncated, so it was suitable
10
11 458 to establish the AR model.

12
13 459

14
15 460 **Figure 4.** Autocorrelation function (ACF) and partial autocorrelation function (PACF)
16
17 461 graphs of residuals of AR((1,2,8)) model. Autocorrelation coefficients and partial
18
19 462 correlation coefficients were almost in 95% confidence interval, so AR ((1,2,8))
20
21 463 model could extract the information of original data well.

22
23 464

24
25 465 **Figure 5.** The numbers of neurons in AR-Elman model and the corresponding root
26
27 466 mean square error (RMSE). When the number of neuron was 6, the RMSE was the
28
29 467 smallest, and the AR-Elman model fitting ability was the strongest.

30
31 468

32
33 469 **Figure 6.** Fitting comparison graph of AR ((1,2,8)) model and AR-Elman model.

34
35 470 Red line stands for the original tuberculosis (TB) incidence curve, green line stands
36
37 471 for AR((1,2,8)) model fitting curve, blue line stands for AR-Elman model fitting
38
39 472 curve. The fitting ability of AR-Elman hybrid model was slightly better than that of
40
41 473 the single AR((1,2,8)).

42
43 474

45 475 **Tables**

46
47 476 **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	

48
49
50
51
52
53
54
55
56
57
58
59 477
60

478

479 **Table 2.** Parameter estimates of the tentative models with their Akaike information
 480 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		

481

482 Table 3. Comparison results of in-sample fitting and out-of-sample forecasting
 483 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

484

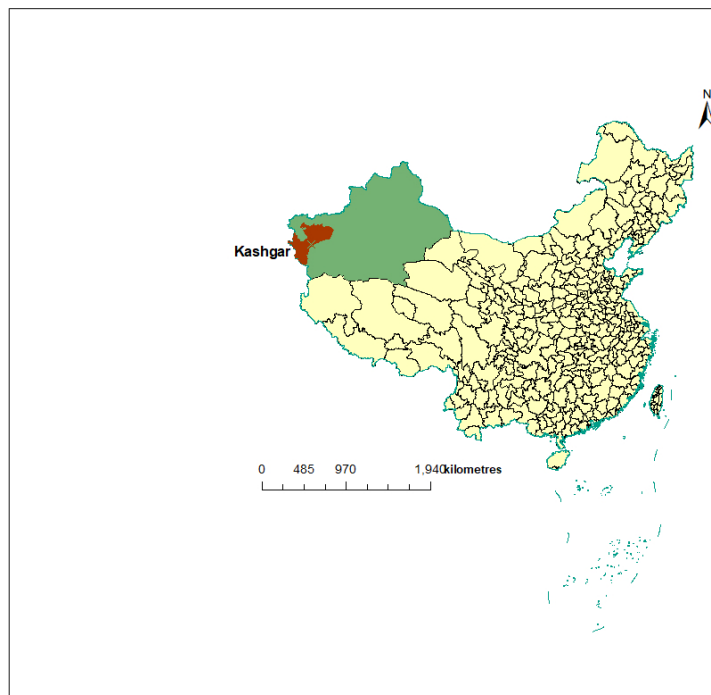


Figure 1. The red part of this picture is the location of kashgar in Xinjiang, China.

296x210mm (96 x 96 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

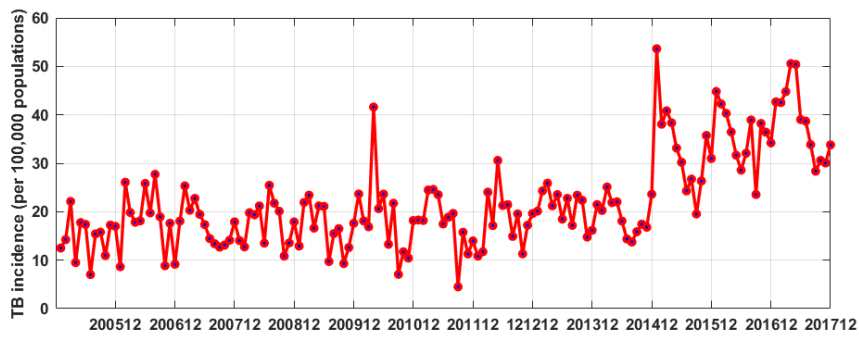


Figure 2

285x99mm (96 x 96 DPI)

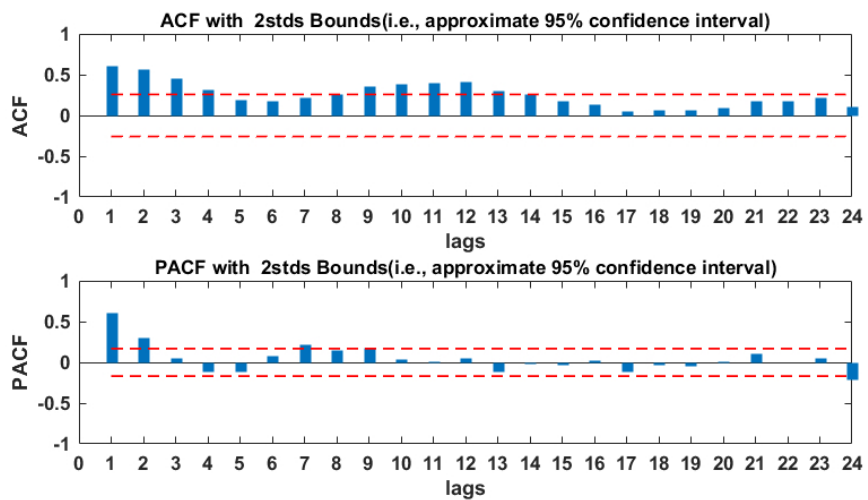


Figure 3. The ACF and PACF diagrams of modeling data

233x121mm (96 x 96 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

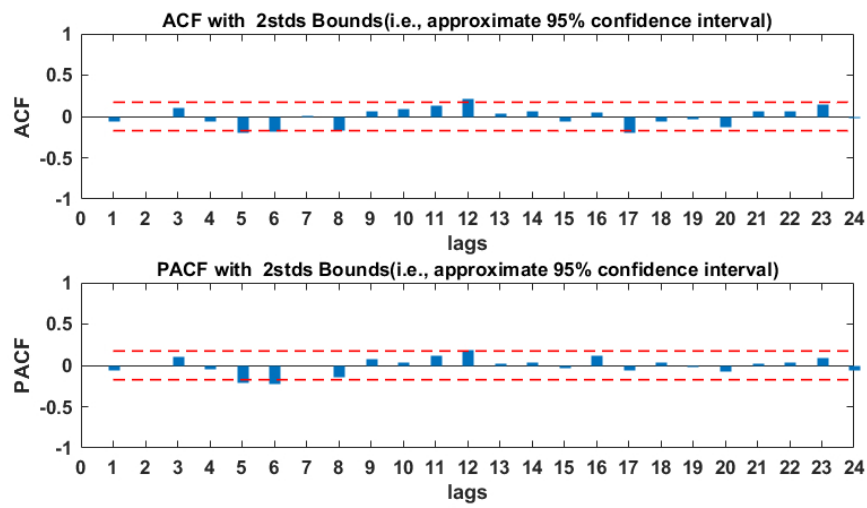


Figure 4. The ACF and PACF graphs of residual errors of AR ((1,2,8)) Model

233x122mm (96 x 96 DPI)

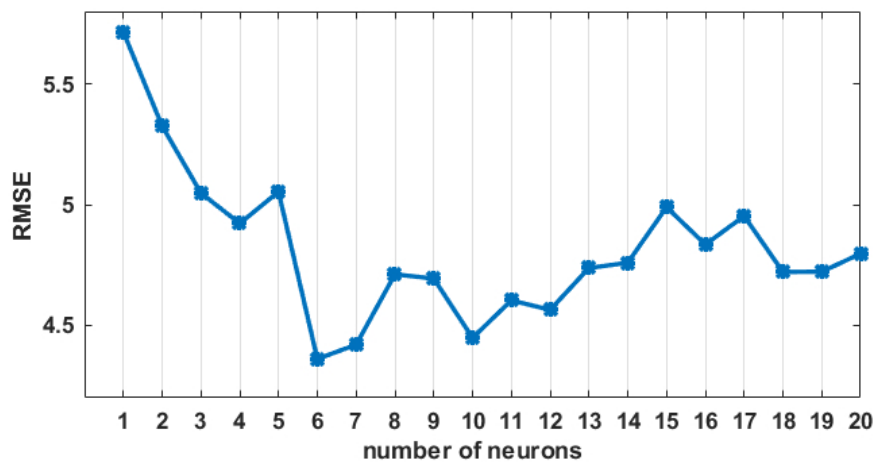


Figure 5. The numbers of neurons in AR-Elman Model and the corresponding RMSE Values

196x95mm (96 x 96 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

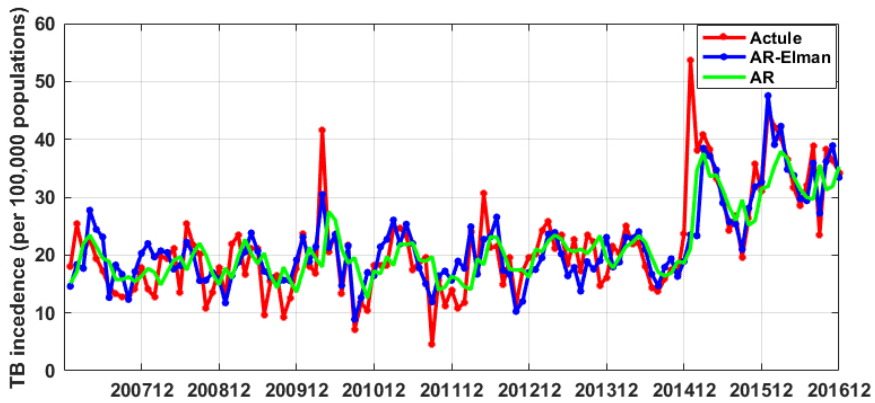


Figure 6

234x99mm (96 x 96 DPI)

BMJ Open

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

Journal:	<i>BMJ Open</i>
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



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 Predictive study of tuberculosis incidence by time series method
5
6
7 and Elman neural network in Kashgar, China

8
9 **Yanling Zheng¹, XueLiang Zhang^{1*}, XiJiang Wang², Kai Wang¹, Yan Cui^{2*}**

10
11 1. College of Medical Engineering and Technology, Xinjiang Medical University,
12
13 Urumqi 830054, People's Republic of China

14
15 2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,
16
17 Urumqi 830001, People's Republic of China

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

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

60 **Yan Cui :^{*} Corresponding author, cuiyan_jk@sina.com**

1
2
3
4 **Abstract:**

5
6 **Objective:** The incidence of tuberculosis (TB) in Kashgar, China, is very high, so the
7
8 prevention and control of TB is arduous. However, there is little quantitative
9
10 prediction study on the TB incidence, therefore, it is urgent to do prediction analysis
11
12 of TB incidence in Kashgar, which can provide scientific reference for the prevention
13
14 and control of TB.

15
16 **Methods:** Based on the monthly TB incidence, a single Box-Jenkins method and a
17
18 Box-Jenkins and Elman neural network (ElmanNN) hybrid method were used to do
19
20 prediction analysis of TB incidence in Kashgar. Root mean square error (RMSE),
21
22 mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to
23
24 measure the prediction accuracy.

25
26 **Results:** After careful analysis, the single AR((1,2,8)) model and
27
28 AR((1,2,8))-ElmanNN (AR-Elman) hybrid model were established, and the optimal
29
30 neurons value of the AR-Elman hybrid model was 6. For the fitting dataset, the
31
32 RMSE, MAE and MAPE were 6.15, 4.33 and 0.2858, respectively, for the AR((1,2,8))
33
34 model, and 3.78, 3.38 and 0.1837, respectively, for the AR-Elman hybrid model. For
35
36 the forecasting dataset, the RMSE, MAE and MAPE were 10.88, 8.75 and 0.2029,
37
38 respectively, for the AR((1,2,8)) model, and 8.86, 7.29 and 0.2006, respectively, for
39
40 the AR-Elman hybrid model.

41
42 **Conclusions:** Both the single AR((1,2,8)) model and AR-Elman model could be used
43
44 to predict the TB incidence in Kashgar, but the modeling and validation
45
46 scale-dependent measures (RMSE, MAE and MAPE) in the AR((1,2,8)) model were
47
48 inferior to those in the AR-Elman hybrid model, which indicated that AR-Elman
49
50 hybrid model was better than AR((1,2,8)) model. The Box-Jenkins and Elman neural
51
52 network hybrid method can be highlighted in predicting the temporal trends of TB
53
54 incidence in Kashgar, which may act as the potential for far-reaching implications for
55
56 prevention and control of TB.

57 **Keywords:** Tuberculosis; Prediction; Box-Jenkins method; Elman neural network
58
59
60

30 **Strengths and limitations of this study:**

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

36 **Introduction**

37 Tuberculosis (TB) is still a major global public health problem and the ninth
38 leading cause of death in the world.¹⁻³ Because TB is contagious, patient resistance is
39 low, treatment time is long, patients' labor force is lost and their contribution to
40 society is reduced, so TB brings great burden to society and economy. All countries in
41 the world are working hard to fight tuberculosis. In 2018, the number of new reported
42 TB cases was about 10 million; this figure has remained relatively stable in recent
43 years. The latest treatment results showed that the global TB treatment success rate
44 was 83%. World Health Organization (WHO) has set targets for the stop TB strategy.
45 The targets mentioned that by 2030, on the basis of the work in 2015, TB deaths will
46 reduce by 90%, and annual new TB cases will reduce by 80%.⁴ In order to achieve
47 these goals, TB prevention and control services must be provided in the broad context
48 of universal health coverage, joint action must be taken to address the social and
49 economic consequences of TB, and technological breakthroughs should be achieved
50 by 2025 to make the TB incidence decline faster than that at any time in history.
51 According to the global tuberculosis report 2019⁴, China has the second highest
52 number of TB cases in the world.⁴ The TB incidence in western China was much
53 higher than that in eastern and central China. The province with the highest TB
54 incidence in the west is Xinjiang province. From 2016 to 2017, the annual TB
55 incidence in Xinjiang was 185.66/100,000 and 202.58/100,000, while the annual TB
56 incidence from 2016 to 2017 in China was 61/100,000 and 60.53/100,000,
57 respectively, it is nearly three times higher in Xinjiang than that national level.

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

1
2
3
4 60 Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7
5
6 61 times higher than that of the national level. Doing a good job in the prevention and
7
8 62 control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.

9
10 63 Mastering the changing law of the incidence of infectious diseases, using the
11
12 64 existing surveillance data to analyze, then, to predict the possible epidemic trend and
13
14 65 provide reference data for the prevention, can better control the occurrence and
15
16 66 epidemic of infectious diseases. Prediction of infectious diseases is to predict the
17
18 67 occurrence, development and epidemic trend of infectious diseases according to the
19
20 68 occurrence, development law and related factors of infectious diseases, based on
21
22 69 analysis, judgment and mathematical model, etc.

23
24 70 For study of quantitative prediction of infectious diseases, there are many
25
26 71 methods, such as grey prediction method ⁵, exponential smoothing prediction method
27
28 72 ^{6,7}, dynamic model prediction method ⁸, Box-Jenkins method ⁹, neural network
29
30 73 method ¹⁰, etc., with the deepening of prediction research, more and more scholars
31
32 74 like to use the Box-Jenkins method ¹¹⁻²¹, there are many different models in this
33
34 75 method, and if appropriate models are established according to the characteristics of
35
36 76 time series, high prediction ability often can be obtained. Neural network has strong
37
38 77 nonlinear mapping ability, in which Elman neural network is composed of input layer,
39
40 78 hidden layer, connection layer and output layer. The Elman network has dynamic
41
42 79 memory function, and it is very suitable for time series prediction. At present, the
43
44 80 Elman network is widely used in various fields, and has achieved successful
45
46 81 prediction results. ²²⁻²⁶ Sometimes, the prediction effect of a single model is not ideal,
47
48 82 in order to further improve the prediction accuracy, many studies adopt the combined
49
50 83 model prediction method ²⁷⁻²⁹, the combined model can absorb the advantages of two
51
52 84 or more methods so as to achieve a higher prediction accuracy.

53
54 85 The TB incidence of Kashgar is very high, and it is urgent to do a good job in the
55
56 86 prevention and control of TB in this area. Accurate prediction of TB incidence is a
57
58 87 prerequisite for prevention and control, which can help advance resource planning and
59
60 88 policy formulation. In this study, the popular Box-Jenkins time series method was
89 89 used to build model for predicting the TB incidence in Kashgar, in order to improve

1
2
3
4 90 the prediction accuracy, the Elman neural network with strong nonlinear information
5
6 91 capture ability was used to construct the combined model for prediction analysis.

7 8 92 **Materials and Methods**

9 10 93 **Study area and data Sources**

11
12 94 We selected Kashgar as the study site (see Figure 1). This area is located in the
13
14 95 south of Xinjiang province in China with an area of approximately 16.2 thousand
15
16 96 square kilometers and a permanent population of 4.64 million in 2018. TB case data
17
18 97 from January 2005 to December 2017 were obtained from the Center for Disease
19
20 98 Control and Prevention of Xinjiang Uygur Autonomous Region. Population data were
21
22 99 obtained from Xinjiang Bureau of Statistics. Based on the population data and TB
23
24 100 case data, we calculated the incidence data of TB.

25 26 101 **Patient and public involvement**

27
28 102 Patients were not involved in the design of this study as it involved only
29
30 103 observational analysis of an anonymised, pre-existing, routinely collected dataset.

31 32 104 **Autoregressive moving average (ARMA) model**

33
34 105 ARMA model³⁰ is an important time series analysis and prediction model in
35
36 106 Box-Jenkins method, also known as auto-regression moving average model. The
37
38 107 ARMA (p,q) model is a model with autocorrelation order p and moving average order
39
40 108 q, p and q are judged by autocorrelation function (ACF) and partial autocorrelation
41
42 109 function (PACF) diagram of stationary data. If the original data is stable, the
43
44 110 autocorrelation coefficients are trailing, and the partial correlation coefficients are the
45
46 111 p-order truncated, then q=0, the ARMA(p,q) model is recorded as AR(p), and its
47
48 112 expression is as follows:

$$39 113 X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t,$$

40 114 where, X_t is the observed value at t, ϕ_1, \dots, ϕ_p are model parameters, c is a constant, if
41
42 115 only ϕ_1, ϕ_2, ϕ_p are not zeros, then, The AR(p) model becomes a sparse model, which
43
44 116 can be recorded as AR((1,2,p)).

45
46 117 If the original data is stable, the autocorrelation coefficients are q-order truncated

1
2
3
4 118 and the partial correlation coefficients are trailing, then $p=0$ and the ARMA(p,q)
5
6 119 model becomes MA(q), its expression is as follows:

7
8 120
$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$

9
10 121 where, μ is the expected value of X_t . $\theta_1, \dots, \theta_p$ are model parameters.

11
12 122 If the original data is stable, the autocorrelation coefficients are trailing, and the
13
14 123 partial correlation coefficients are also trailing, then the expression of the ARMA (p, q)
15
16 124 model is as follows:

17
18 125
$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}.$$

19
20 126 There are four main steps in ARMA modeling:

21
22 127 First step. The prerequisite for ARMA modeling is the stationary of time series.
23
24 128 Check whether the data is stable by Augmented Dickey-Fuller (ADF) unit root test. In
25
26 129 this study, the significant level probability (p-value) is 0.05, and if the p-value is less
27
28 130 than 0.05, then, the data is stable. By observing the autocorrelation function (ACF)
29
30 131 and partial autocorrelation function (PACF) of the stable data, we can determine the
31
32 132 possible values of p and q and establish the possible ARMA (p, q) model.

33
34 133 Second step. The parameters of ARMA(p, q) model are estimated by maximum
35
36 134 likelihood estimation or least square estimation, and the model parameters are tested.
37
38 135 If p-value is less than 0.05, the parameters have statistical significance. The best
39
40 136 model is determined according to the value of Akaike information criterion (AIC),
41
42 137 Schwarz criterion (SC) and Goodness of Fit (R^2) of model. The smaller AIC and SC
43
44 138 are, the larger the R^2 is, and the better the model is.

45
46 139 Third step. To determine whether the established ARMA (p, q) model is suitable.
47
48 140 The residual sequence of a suitable model shall be the white noise process, and its
49
50 141 ACF and PACF coefficients should be within twice the standard deviation range,
51
52 142 otherwise, it is considered that the extraction information of the established model is
53
54 143 not sufficient, and it is necessary to consider improving the accuracy of the model.

54
55 144 Fourth step. Using the established model to do prediction and analysis.

56
57 145 **Elman neural network model**

58
59 146 The Elman neural network (see Figure 2) is proposed by Elman in 1990³¹. The
60

1
2
3
4 147 model is generally divided into four layers: input layer, hidden layer, receiving layer
5
6 148 and output layer. The characteristic of Elman network is that the output of the hidden
7
8 149 layer is connected to the input of the hidden layer through the delay and storage of the
9
10 150 receiving layer, which makes it sensitive to the data of the historical state. The
11
12 151 addition of the internal feedback network increases the ability of the network itself to
13
14 152 deal with dynamic information, thus achieving the purpose of dynamic modeling.

15 153 The mathematical structure of the Elman neural network is as follows:

16
17 154
$$x(k)=f(w^1x_c(k)+ w^2u(k-1))$$

18
19 155
$$x_c(k)= \alpha x_c(k-1)+ x(k-1)$$

20
21 156
$$y(k)=g(w^3x(k))$$

22
23 157 Where, w^1 is the connection weight matrix between the contact unit and the
24
25 158 hidden layer unit, w^2 is the connection weight matrix between the input unit and the
26
27 159 hidden layer unit, w^3 is the connection weight matrix between the hidden layer unit
28
29 160 and the output unit. $x_c(k)$ and $x(k)$ represent the output of the contact unit and the
30
31 161 hidden layer unit, respectively, $y(k)$ represents the output of the output unit, α is a
32
33 162 self-connected feedback gain factor, $0 \leq \alpha < 1$, $f(x)$ often takes the sigmoid function.

34 163 There are four main steps in Elman neural network modeling:

35 164 First step. Data standardization processing. Data standardization is scaling the
36
37 165 data to a small specific interval. In order to remove the unit limit of the data and
38
39 166 convert it into dimensionless pure value, it is convenient for the index of different
40
41 167 units or order of magnitude to be compared and weighted. In our study, we use
42
43 168 function package `mapminmax()` to standardize the data, standardized data is in $[-1,1]$
44
45 169 range.

46
47 170 Second step. Determine the input layer, the output layer. Generally, the input and
48
49 171 output layer are determined according to the characteristics of the data and the needs
50
51 172 of the analysis.

52
53 173 Third step. Set the parameters of the Elman model, such as training epochs and
54
55 174 goals, etc., in our study, training epochs and goals of Elman neural network were set
56
57 175 to 2000 and 0.00001, respectively. To determine the number of neurons in the hidden
58
59 176 layer, so that the error of the established Elman model is minimized. At present, there
60
177 is no ideal analytical expression for the number of neurons in the hidden layer. The
178 number of neurons has a great influence on the performance of the network. When the
179 number of neurons is too large, it will lead to that the network learning time is too

180 long, the generalization performances are not good, and even the convergence failure,
 181 but when the number of neurons is too small, the fault-tolerant ability of the network
 182 is poor. In general, the number of neurons does not exceed 20.³² In this study, matlab
 183 cyclic structure was used to find the optimal number of neurons by comparing the
 184 RMSE values of Elman networks with neurons 1 to 20.

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

187 **Model comparison measures**

188 Three performance indexes, root mean square error (RMSE), mean absolute error
 189 (MAE) and mean absolute percentage error (MAPE) were used to assess the fitting
 190 and forecasting accuracy of two models. The smaller these three values are, the better
 191 the model is. Their expressions are as follows¹⁰:

$$192 \quad RMSE = \sqrt{\frac{\sum_{t=1}^n (X_t - \hat{X}_t)^2}{n}},$$

$$193 \quad MAE = \frac{\sum_{t=1}^n |X_t - \hat{X}_t|}{n},$$

$$194 \quad MAPE = \frac{\sum_{t=1}^n \left| \frac{X_t - \hat{X}_t}{X_t} \right| \times 100}{n},$$

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

197 **Statistical software**

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

199 **Results**

200 From January 2005 to December 2017, the number of reported TB cases was
 201 141984 in Kashgar, Xinjiang, the average annual TB cases were 7888 and the average
 202 annual incidence was 191.18. Figure 3 showed the time series graph of the TB
 203 incidence. It can be seen from the Figure 3 that the curve of TB incidence showed
 204 strong nonlinear characteristics from 2005 to 2014, and the TB incidence from 2015
 205 to 2017 were significantly higher than that of previous years.

206 The data of TB incidence from January 2005 to December 2017 were divided
 207 into two parts. The data from January 2005 to December 2016 were used to build

208 model and the data from January 2017 to December 2017 were used to test the model.

209 **Establishment of ARMA Model**

210 ARMA Modeling requires data stability, so, first of all, the stability of the
211 modeling part of the data was verified by ADF test. The results of ADF test showed
212 that p-value was less than 0.05(see Table 1), which indicated that the data was stable
213 and could be directly used to build the model. Secondly, the ACF and PACF graphs
214 of the modeling data were plotted (see Figure 4), it was obvious from Figure 4 that the
215 autocorrelation coefficients were trailing distribution, and the partial correlation
216 coefficients were almost a second-order truncated distribution, only at the lag 7, 8 and
217 9, the correlation coefficients were a little large. Based on this situation, we
218 considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and
219 AR((1,2,9)). The least square method was used to test the parameters of the four
220 models. The results of the test were shown in Table 2; we can see that, of the four
221 models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing the
222 two models, it was found that, the AR((1,2,8)) model had smaller AIC and SC values,
223 and the R^2 values of AR((1,2,8)) were larger than the R^2 values of AR (2), so the AR
224 (2) model was abandoned. Then, the residual analysis of the AR((1,2,8)) model was
225 carried out, the autocorrelation and partial correlation coefficient of the residuals were
226 almost all within two times standard deviation, and only in the lag 5,6,12, they were
227 beyond the range of two times standard deviation (see Figure 5), which indicated that
228 the AR((1,2,8)) model could be used to roughly predict the TB incidence in Kashgar.
229 We used AR((1,2,8)) model to fit the TB incidence from September 2005 to
230 December 2016, the fitting RMSE ,MAE and MAPE were 6.15, 4.33 and 0.2858,
231 respectively; we used AR((1,2,8)) model to predict the TB incidence from January
232 2017 to December 2017, the prediction RMSE, MAE and MAPE were 10.88 , 8.75
233 and 0.2029, respectively.

234 **Establishment of AR-Elman Model**

235 In order to improve the prediction accuracy of the AR((1,2,8)) model, we tried
236 to establish the AR((1,2,8))-Elman hybrid model. The fitting sequence of AR((1,2,8))

237 model was used as input variable, and the actual TB incidence was used as output
 238 variable. Due to the a little similarity of the annual trend of TB incidence in Kashgar
 239 (see Figure 3), therefore, we created twelve time-lagged variables as input features.
 240 Supposing that x_t represented the TB incidence at time t, and then the input matrix and
 241 the output matrix of modeling data set used in this study were designed as follows
 242 ($N=12$):

$$243 \quad \text{input matrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_i \\ x_2 & x_3 & \dots & x_{i+1} \\ \dots & \dots & \dots & \dots \\ x_N & x_{N+1} & \dots & \dots \end{bmatrix}, \text{output matrix} = [x_{N+1} \quad x_{N+2} \quad \dots \quad x_{N+i}]$$

244 We selected twelve as the number of input layers of AR-Elman network and
 245 one as the number of output layers representing the forecast value. By the matlab
 246 cyclic structure, we selected the optimal number of neurons between 1 and 20, and
 247 finally we found when the number of neurons was 6 (see Figure 6), the RMSE was the
 248 smallest, and the AR-Elman was optimal. We used the AR-Elman model to fit the
 249 training data, RMSE was 3.78, MAE was 3.38, MAPE was 0.1837, and the R^2 of the
 250 model was 0.83; we used the AR-Elman model to predict the TB incidence from
 251 January 2017 to December 2017, RMSE was 8.86, MAE was 7.29, and MAPE was
 252 0.2006. The fitting curves of AR((1,2,8)) model and AR-Elman model, and the
 253 prediction curve of AR-Elman model were shown in Figure 7. Comparison results of
 254 the AR((1,2,8)) model and AR-Elman model were shown in Table 3, both the fitting
 255 RMSE, MAE and MAPE and the predicting RMSE, MAE and MAPE of the
 256 AR-Elman model were smaller than those of the single AR((1,2,8)) model, which
 257 indicated that the AR-Elman combined model established in this study was more
 258 suitable for predicting the TB incidence in Kashgar.

259 Discussion

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

1
2
3
4 268 the high TB incidence in Xinjiang. Therefore, it is urgent to do a good job in the
5
6 269 prevention and control of TB in Kashgar.

7
8 270 The prediction and early warning of infectious diseases is an important link in
9
10 271 the prevention and control of infectious diseases.³²⁻³⁴ Therefore, this study carried out
11
12 272 research from the point of view of prediction to explore an accurate prediction model
13
14 273 and do prediction analysis of TB incidence in Kashgar, so as to provide scientific
15
16 274 reference for the prevention and control of the disease in this area. The Box -Jenkins
17
18 275 method is a popular time series prediction method, this method has good prediction
19
20 276 performance and high prediction accuracy; Elman Neural network can capture
21
22 277 nonlinear information of time series data very well. In this study, the two methods
23
24 278 were combined to study the prediction model of TB incidence in Kashgar.

25
26 279 Many studies have found that Box-Jenkins method has a good ability of fitting
27
28 280 and forecasting. For stationary time series that do not contain seasonality, it is more
29
30 281 suitable to use the ARMA model of the Box-Jenkins method to do prediction analysis
31
32 282 ³⁵, for non-stationary time series of infectious diseases with obvious seasonality, it is
33
34 283 more suitable to use seasonal autoregressive integrated moving average (SARIMA)
35
36 284 model of the Box-Jenkins method for prediction analysis.⁹⁻¹² In our study, from Figure
37
38 285 3, we could see that the seasonality of the TB incidence in Kashgar from 2005 to 2014
39
40 286 was not obvious, there was only a certain seasonality from 2015 to 2017, and we
41
42 287 found that the time series of TB incidence was stable by ADF unit root test, and the
43
44 288 autocorrelation and partial correlation coefficients of modeling data at lag 12, 24 were
45
46 289 not obviously large, therefore, for our research data, we used ARMA model to do
47
48 290 forecast analysis, finally, we established AR((1,2,8)) model of the Box-Jenkins
49
50 291 method, it has a good performance to fit and predict the TB incidence of Kashgar in
51
52 292 Xinjiang. From Figure 3, we can also see that the time series of TB incidence has
53
54 293 strong non-linear, the established AR((1,2,8)) model mainly extracted the linear
55
56 294 information of data, considering that the neural network can capture the non-linear
57
58 295 information of data well, in order to improve the prediction accuracy of TB incidence
59
60 296 rate in Kashgar, we used AR((1,2,8)) model and Elman neural network model to
297 establish AR-Elman hybrid model. Many studies have found that the combination
298 model can improve the accuracy of prediction, such as, Wang et al.²⁸ found that
299 SARIMA-NAR hybrid model has an outstanding ability to improve the prediction
300 accuracy relative to SARIMA model and nonlinear autoregressive network (NAR)

1
2
3
4 301 model when they were used to predict pertussis incidence in China. Li et al.²⁷ found
5 302 ARIMA-GRNN hybrid model was shown to be superior to the single ARIMA model
6
7 303 in predicting the short-term TB incidence in the Chinese population. Our research was
8
9 304 consistent with these literatures that our AR-Elman hybrid model was more accurate
10 305 than the single AR((1,2,8)) model.

11
12 306 In the past few years, Xinjiang's economic development was relatively backward,
13
14 307 medical resources were scarce, diagnosis and treatment were delayed, the continuous
15 308 spread of TB has become a difficult problem in the control of TB in Xinjiang. In
16
17 309 recent years, Xinjiang has introduced many new policies to increase investment in TB
18
19 310 prevention and control, and the relevant departments of disease prevention and control
20
21 311 in Xinjiang have also done a lot of effective work, which has helped to control
22 312 effectively the rapid increase of the TB incidence in Xinjiang. In order to do a good
23
24 313 job in the prevention and control of TB in Xinjiang, many departments need to make
25
26 314 joint efforts. Our research was mainly to build a high-precision prediction model to
27 315 help early warning and prediction analysis of tuberculosis in Kashgar. Finally, we
28
29 316 established the AR-Elman hybrid model, which had high fitting and prediction
30
31 317 accuracy of TB incidence in Kashgar, Xinjiang.

32
33 318 Our study found that Box-Jenkins and Elman neural network hybrid method was
34
35 319 an effective method for predicting the incidence of TB in Kashgar, it could provide a
36
37 320 scientific reference for prediction analysis of TB incidence. However, our study also
38
39 321 has some limitations: our method is only suitable for short-term prediction, long-term
40
41 322 prediction performance will decline, two main reasons: first, our model was based on
42
43 323 historical data characteristics; second, climatic factors, environmental factors,
44
45 324 demographic factors and political issues may have certain impacts on the change of
46
47 325 incidence. Therefore, if the established model becomes old and people want to obtain
48
49 326 more accurate prediction results, it will be needed to adjust the model parameters,
50
51 327 update the model based on the new modeling sample data, and then to do prediction
52
53 328 analysis.

50 329 **Conclusions**

51
52 330 Kashgar has a very high TB incidence, in order to provide some help for the
53
54 331 prevention and control of this disease, the prediction problem of the TB incidence was
55
56 332 studied. Firstly, a single AR((1,2,8)) prediction model was established by using
57
58 333 Box-Jenkins method, and its fitting and prediction performance were good, secondly,
59
60 334 in order to improve the prediction accuracy of the single AR((1,2,8)) model, we used

1
2
3
4 335 single AR((1,2,8)) and Elman neural network with strong ability to capture nonlinear
5 336 information to establish AR-Elman hybrid model. The fitting and prediction accuracy
6
7 337 of the hybrid model was higher than that of the single AR((1,2,8)) model. The
8
9 338 AR-Elman hybrid model can provide scientific reference for predicting and warning
10
11 339 the TB incidence in Kashgar, Xinjiang. This study also has some limitations,
12
13 340 long-term prediction accuracy of the AR-Elman hybrid model will decline, after the
14
15 341 model becomes old, it will be needed to adjust the model parameters and establish
16
17 342 new Box-Jenkins and Elman neural network hybrid model to do prediction analysis
18
19 343 based on new sample data.

344 **Competing interests**

345 The authors declare no conflict of interest.

346 **Funding**

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

352 **Data sharing statement**

353 No additional data are available.

354 **Author contributions**

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

358 **Acknowledgements**

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

361 **References**

- 362 1.Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia:
363 prevalence, risk factors and its public health importance. BMC Infectious Diseases,
364 2019; 19(1).
- 365 2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity
366 pattern of Mycobacterium tuberculosis strains isolated from pulmonary
367 tuberculosis patients in and around Ambo Town, Central Ethiopia. Plos One, 2018;

- 1
2
3
4 368 13(2):e0193083-.
- 5
6 369 3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control
7
8 370 in Debre Markos Town, Northwest Ethiopia. *Advances in Medicine*,
9
10 371 2018;2018(2):1-8.
- 11
12 372 4. WHO. Global tuberculosis report 2019, [https://www.who.int/tb/publications/
13
14 373 global_report/en/](https://www.who.int/tb/publications/global_report/en/). (Accessed on 17 Oct 2019).
- 15
16 374 5. Zhao YF, Shou MH, Wang ZX. Prediction of the Number of Patients Infected with
17
18 375 COVID-19 Based on Rolling Grey Verhulst Models. *Int J Environ Res Public
19
20 376 Health*. 2020; 17(12):4582.
- 21
22 377 6. Guan P, Wu W, Huang D. Trends of reported human brucellosis cases in mainland
23
24 378 China from 2007 to 2017: an exponential smoothing time series analysis. *Environ
25
26 379 Health Prev Med*. 2018; 23(1):23.
- 27
28 380 7. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, Xu JY, Shi J, Sun DY.
29
30 381 Analysis and predication of tuberculosis registration rates in Henan Province,
31
32 382 China: an exponential smoothing model study. *Infect Dis Poverty*. 2020;9(1):123.
- 33
34 383 8. Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling
35
36 384 of time series of dengue disease case counts. *PLoS Negl Trop Dis*.
37
38 385 2017;11(7):e0005696.
- 39
40 386 9. Wang Y, Xu C, Zhang S, Wang Z, Zhu Y, Yuan J. Temporal trends analysis of
41
42 387 human brucellosis incidence in mainland China from 2004 to 2018. *Sci Rep*.
43
44 388 2018;8(1):15901.
- 45
46 389 10. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S. An Advanced Data-Driven
47
48 390 Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series
49
50 391 Forecasting in Qinghai Province, China. *Infect Drug Resist*. 2020;13:867-880.
- 51
52 392 11. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of
53
54 393 tuberculosis cases reported at a tertiary hospital in Ghana: a time series model
55
56 394 approach. *BMC Public Health*.2018;18(1):1292.
- 57
58 395 12. Tohidinik H R , Moheballi M , Mansournia M A, et al. Forecasting zoonotic
59
60 396 cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran:
397 A SARIMA Analysis. *Tropical Medicine & International Health*,

- 1
2
3
4 398 2018;23(8):860–869.
- 5
6 399 13. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in
7
8 400 Ouagadougou, Burkina Faso, 2011–2015. *Malaria Journal*, 2018;17(1):138.
- 9
10 401 14. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease
11
12 402 outbreak and primary healthcare delivery in Liberia: Time-series analyses for
13
14 403 2010–2016. *Plos Medicine*, 2018; 15(2):e1002508.
- 15
16 404 15. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth
17
18 405 disease in Nanjing, China: a time-series model based on weather variability.
19
20 406 *International Journal of Biometeorology*, 2017;(6):1-10.
- 21
22 407 16. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in
23
24 408 Kumasi: Using ARIMA models to forecast future incidence. *Cogent Social*
25
26 409 *Sciences*, 2018; 4(1):1461544.
- 27
28 410 17. Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving
29
30 411 average model and generalised regression neural network model for prediction of
31
32 412 haemorrhagic fever with renal syndrome in China: a time-series study. *BMJ Open*,
33
34 413 2019; 9(6):e025773.
- 35
36 414 18. Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in
37
38 415 children younger than 48 months in Suzhou using a seasonal autoregressive
39
40 416 integrated moving average model. *BMJ Open*, 2019; 9(1).
- 41
42 417 19. Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable
43
44 418 incidence time series. *Epidemiology & Infection*, 2019; 147.
- 45
46 419 20. Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand,
47
48 420 foot and mouth disease cases in China from 2008 to 2018. *Epidemiology &*
49
50 421 *Infection*, 2019; 147.
- 51
52 422 21. Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in
53
54 423 modelling and forecasting emergency department visits in a medical centre in
55
56 424 Southern Taiwan. *Bmj Open*, 2017; 7(11):e018628.
- 57
58 425 22. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum
59
60 426 Analysis for wind speed forecasting based on Elman Neural Network. *Energy*
427 *Conversion & Management*, 2017; 148:895-904.

- 1
2
3
4 428 23. Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for
5 429 Forecasting China's Carbon Dioxide Emissions. *Communications in Computer*
6 and Information Science, 2018;(951): 36-47.
7
8 430
9 431 24. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural
10 432 Network for Disease Prediction. *Arabian Journal for Science and Engineering*,
11 433 2019;44(11): 9209–9220.
12
13 434 25. Mehrgini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman
14 435 artificial neural network. *Carbonates & Evaporites*, 2017;(6):1-11.
15
16 436 26. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the
17 437 application of disease modifying therapy for sickle cell using classification
18 438 models. *Neurocomputing*, 2017; 228(C):154-164.
19
20 439 27. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, Ji Y, Xu D, Wang J. Application of a
21 440 hybrid model in predicting the incidence of tuberculosis in a Chinese population.
22 441 *Infect Drug Resist*. 2019;12:1011-1020.
23
24 442 28. Wang Y, Xu C, Wang Z, Zhang S, Zhu Y, Yuan J. Time series modeling of
25 443 pertussis incidence in China from 2004 to 2018 with a novel wavelet based
26 444 SARIMA-NAR hybrid model. *PLoS One*. 2018;13(12):e0208404.
27
28 445 29. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, Yuan J. Temporal trends
29 446 analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a
30 447 new SARIMA-NARNNX hybrid model. *BMJ Open*. 2019;9(7):e024409.
31
32 448 30. Box G E, Jenkins G M. *Time series analysis: forecasting and control* rev. ed.
33 449 Oakland, California, Holden-Day, 1976; 31(4):238-242.
34
35 450 31. Ming Cheng. *The principle and example of MATLAB neural network*. Tsinghua
36 451 University Press, 2013.
37
38 452 32. Huppert A, Katriel G. Mathematical modelling and prediction in infectious
39 453 disease epidemiology. *Clin Microbiol Infect*. 2013;19(11):999-1005
40
41 454 33. Wearing HJ1, Rohani P, Keeling MJ. Appropriate models for the management of
42 455 infectious diseases. *PLoS Med*. 2005; 2(7):e174.
43
44 456 34. Neuberger A, Paul M, Nizar A, Raoult D. *Modelling in infectious diseases*:
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 457 between haphazard and hazard. Clin Microbiol Infect. 2013;19(11):993-8.
5
6 458 35. Tipirneni-Sajja A, Krafft AJ, Loeffler RB, Song R, Bahrami A, Hankins JS,
7
8 459 Hillenbrand CM. Autoregressive moving average modeling for hepatic iron
9
10 460 quantification in the presence of fat. J Magn Reson Imaging.
11
12 461 2019;50(5):1620-1632.

13
14 462

16 463 **Figures**

17
18 464 **Figure 1.** The red part of this picture is the location of kashgar in Xinjiang, China.
19
20 465 Kashgar is located in the south of Xinjiang, and it has a very high incidence of
21
22 466 tuberculosis.

23
24
25 467

26 468 **Figure 2.** The Structure diagram of Elman neural network. w^1 , w^2 and w^3 are the
27
28 469 connection weight matrixes. $x_c(k)$ and $x(k)$ represent the output of the contact unit and
29
30 470 the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, $u(k-1)$
31
32 471 represents the input of the input unit.

33
34
35 472

36 473 **Figure 3.** Graph of the tuberculosis (TB) incidence in Kashgar from January 2005 to
37
38 474 December 2017. The curve of TB incidence showed strong nonlinear characteristics
39
40 475 from 2005 to 2014, and the TB incidence increased significantly from 2015 to 2017.

41
42
43 476

44 477 **Figure 4.** Autocorrelation function (ACF) and partial autocorrelation function (PACF)
45
46 478 graphs of modeling data. As the delay of the lag order, the autocorrelation coefficients
47
48 479 were trailing and the partial correlation coefficients were truncated, so it was suitable
49
50 480 to establish the AR model.

51
52
53 481

54 482 **Figure 5.** Autocorrelation function (ACF) and partial autocorrelation function (PACF)
55
56 483 graphs of residuals of AR((1,2,8)) model. Autocorrelation coefficients and partial
57
58 484 correlation coefficients were almost in 95% confidence interval, so AR ((1,2,8))
59
60 485 model could extract the information of original data well.

1
2
3
4 486

5 487 **Figure 6.** The numbers of neurons in AR-Elman model and the corresponding root
6
7 488 mean square error (RMSE). When the number of neuron was 6, the RMSE was the
8
9 489 smallest, and the AR-Elman model fitting ability was the strongest.

10
11 490

12
13 491 **Figure 7.** The fitting curves of AR((1,2,8)) model and AR-Elman model, and the
14
15 492 prediction curve of AR-Elman model. Red line stands for the original tuberculosis
16
17 493 (TB) incidence curve, green line stands for AR((1,2,8)) model fitting curve, blue line
18
19 494 stands for AR-Elman model fitting curve. Blue dotted line stands for prediction curve
20
21 495 of AR-Elman model, black dotted line stands for predicted curve of confidence
22
23 496 intervals. The fitting ability of AR-Elman hybrid model was slightly better than that
24
25 497 of the single AR((1,2,8)).

26
27 498

28
29
30 499

31
32
33 500

34
35 501

36
37 502

38
39 503

40
41 504

42
43 505

44
45
46
47
48 506 **Tables**

49
50 507 **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

508

509

510 **Table 2.** Parameter estimates of the tentative models with their Akaike information
 511 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		

512

513 **Table 3.** Comparison results of in-sample fitting and out-of-sample forecasting
 514 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

515

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



Figure 1

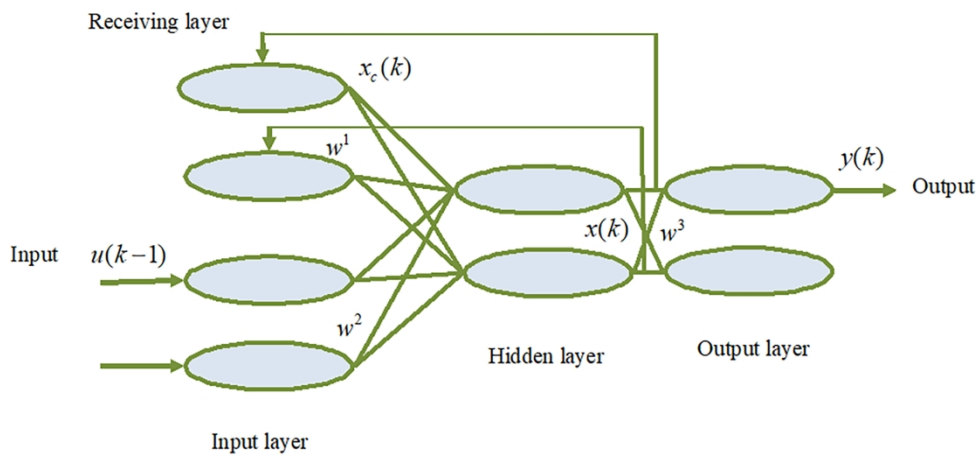


Figure 2

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

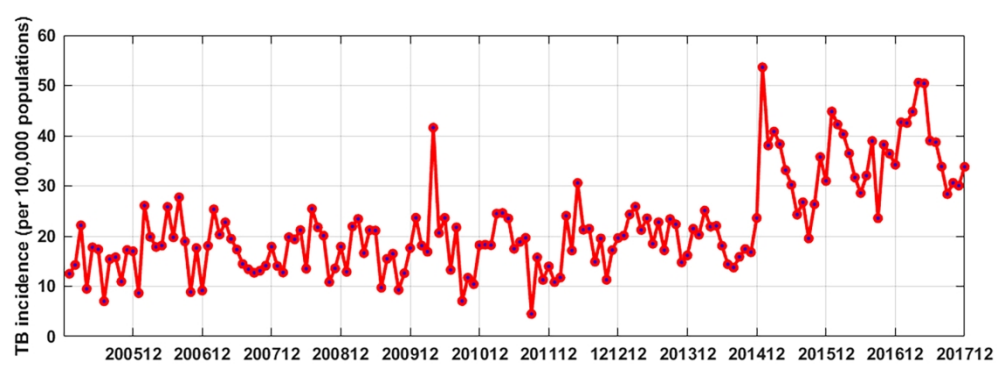


Figure 3

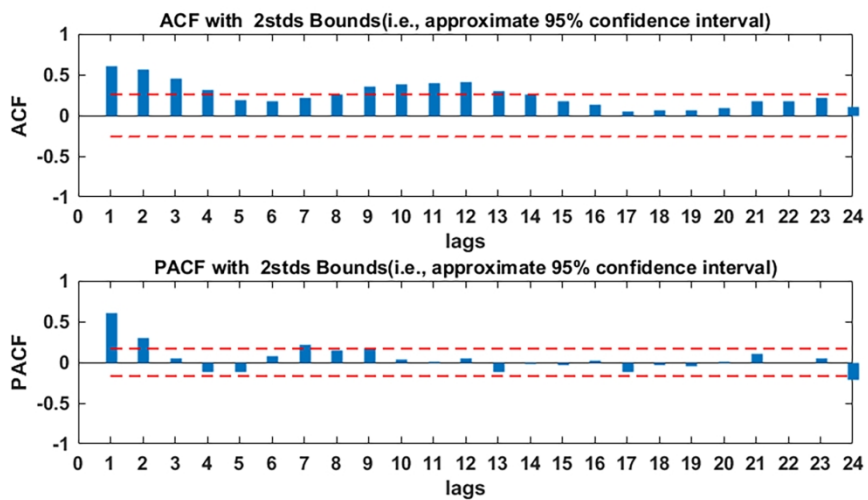


Figure 4

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

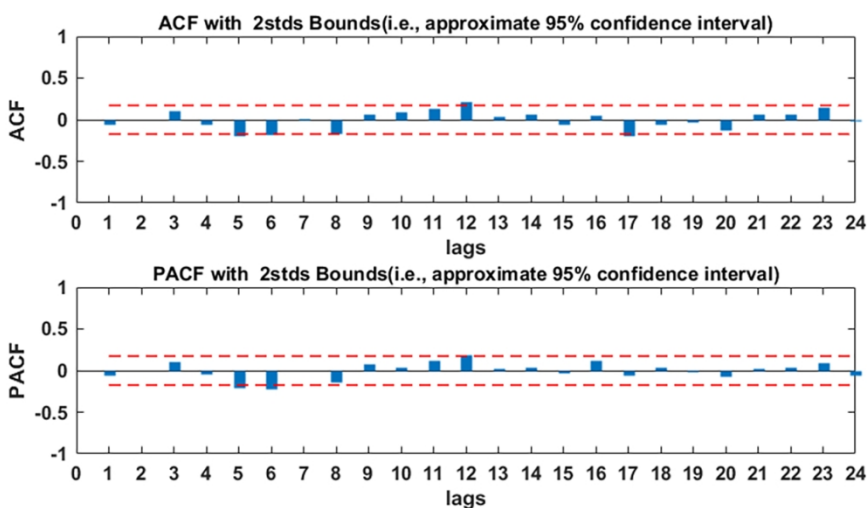


Figure 5

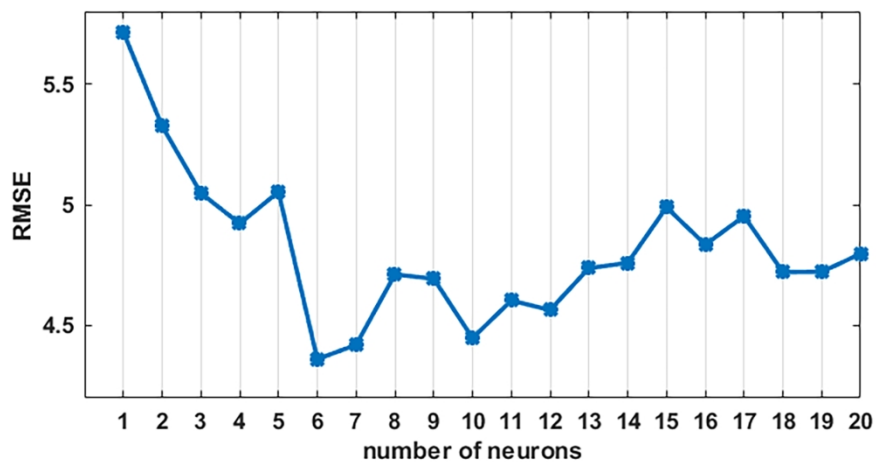


Figure 6

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

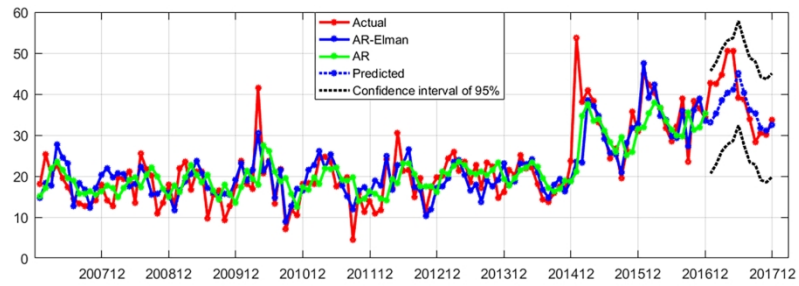


Figure 7

BMJ Open

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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 Predictive study of tuberculosis incidence by time series method
5
6
7 and Elman neural network in Kashgar, China

8 **Yanling Zheng¹, XueLiang Zhang^{1*}, XiJiang Wang², Kai Wang¹, Yan Cui^{2*}**

- 9
10
11 1. State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence
12 Diseases in Central Asia, College of Medical Engineering and Technology,
13 Xinjiang Medical University, Urumqi 830054, People's Republic of China
14
15
16 2. Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region,
17 Urumqi 830001, People's Republic of China
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57

58
59 **Xueliang Zhang** : * Corresponding author, zhengyl_math@sina.cn
60 **Yan Cui** : * Corresponding author, cuiyan_jk@sina.com

1 Abstract

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

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

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

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

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

29

30 **Strengths and limitations of this study:**

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

36 **Introduction**

37 Tuberculosis (TB) is still a major global public health problem and the ninth
38 leading cause of death in the world.¹⁻³ Because TB is contagious, patient resistance is
39 low, treatment time is long, patients' labor force is lost and their contribution to
40 society is reduced, so TB brings great burden to society and economy. All countries in
41 the world are working hard to fight tuberculosis. In 2018, the number of new reported
42 TB cases was about 10 million; this figure has remained relatively stable in recent
43 years. The latest treatment results showed that the global TB treatment success rate
44 was 83%. World Health Organization (WHO) has set targets for the stop TB strategy.
45 The targets mentioned that by 2030, on the basis of the work in 2015, TB deaths will
46 reduce by 90%, and annual new TB cases will reduce by 80%.⁴ In order to achieve
47 these goals, TB prevention and control services must be provided in the broad context
48 of universal health coverage, joint action must be taken to address the social and
49 economic consequences of TB, and technological breakthroughs should be achieved
50 by 2025 to make the TB incidence decline faster than that at any time in history.
51 According to the global tuberculosis report 2019⁴, China has the second highest
52 number of TB cases in the world.⁴ The TB incidence in western China was much
53 higher than that in eastern and central China. From 2016 to 2017, the annual TB
54 incidence in Xinjiang was 185.66/100,000 and 202.58/100,000, while the annual TB
55 incidence from 2016 to 2017 in China was 61/100,000 and 60.53/100,000,
56 respectively, it is nearly three times higher in Xinjiang than that national level.

57 There are 14 Prefectural-Level cities in Xinjiang, China, among which Kashgar
58 has a very high TB incidence rate. From 2016 to 2017, the annual TB incidences in
59 Kashgar were 427.44/100,000 and 465.33/100,000, respectively, which were nearly 7

1
2
3
4 60 times higher than that of the national level. Doing a good job in the prevention and
5
6 61 control of TB in Kashgar is an important link to reduce the TB incidence in Xinjiang.

7
8 62 Mastering the changing law of the incidence of infectious diseases, using the
9
10 63 existing surveillance data to analyze, then, to predict the possible epidemic trend and
11
12 64 provide reference data for the prevention, can better control the occurrence and
13
14 65 epidemic of infectious diseases. Prediction of infectious diseases is to predict the
15
16 66 occurrence, development and epidemic trend of infectious diseases according to the
17
18 67 occurrence, development law and related factors of infectious diseases, based on
19
20 68 analysis, judgment and mathematical model, etc.

21
22 69 For study of quantitative prediction of infectious diseases, there are many
23
24 70 methods, such as grey prediction method ⁵, exponential smoothing prediction method
25
26 71 ^{6,7}, dynamic model prediction method ⁸, Box-Jenkins method ⁹, neural network
27
28 72 method ¹⁰, etc., with the deepening of prediction research, more and more scholars
29
30 73 like to use the Box-Jenkins method ¹¹⁻²¹, there are many different models in this
31
32 74 method, and if appropriate models are established according to the characteristics of
33
34 75 time series, high prediction ability often can be obtained. Neural network has strong
35
36 76 nonlinear mapping ability, in which Elman neural network is composed of input layer,
37
38 77 hidden layer, connection layer and output layer. The Elman network has dynamic
39
40 78 memory function, and it is very suitable for time series prediction. At present, the
41
42 79 Elman network is widely used in various fields, and has achieved successful
43
44 80 prediction results. ²²⁻²⁶ Sometimes, the prediction effect of a single model is not ideal,
45
46 81 in order to further improve the prediction accuracy, many studies adopted combined
47
48 82 model prediction method ²⁷⁻²⁹, the combined model can absorb the advantages of two
49
50 83 or more methods so as to achieve a higher prediction accuracy.

51
52 84 The TB incidence of Kashgar is very high, and it is urgent to do a good job in the
53
54 85 prevention and control of TB in this area. Accurate prediction of TB incidence is a
55
56 86 prerequisite for prevention and control, which can help advance resource planning and
57
58 87 policy formulation. In this study, the popular Box-Jenkins time series method was
59
60 88 used to build model for predicting the TB incidence in Kashgar, in order to improve
89 89 the prediction accuracy, the Elman neural network with strong nonlinear information

1
2
3
4 90 capture ability was used to construct the combined model for prediction analysis.
5

6 91 **Materials and Methods**

7 8 92 **Study area and data Sources**

9
10 93 We selected Kashgar as the study site (see Figure 1). This area is located in the
11
12 94 south of Xinjiang province in China with an area of approximately 16.2 thousand
13
14 95 square kilometers and a permanent population of 4.64 million in 2018. TB case data
15
16 96 from January 2005 to December 2017 were obtained from the Center for Disease
17
18 97 Control and Prevention (CDC) of Xinjiang Uygur Autonomous Region, all TB cases
19
20 98 in Xinjiang must be reported to the CDC through the infectious disease surveillance
21
22 99 system within 24 hours. The TB cases datasets used need permission of the CDC.
23
24 100 Population data were obtained from the website of Xinjiang Bureau of Statistics
25
26 101 (http://tjj.xinjiang.gov.cn/tjj/tjfw/list_tjfw.shtml). Based on the population data and
27
28 102 TB case data, we calculated the incidence data of TB.

29 30 103 **Patient and public involvement**

31
32 104 Patients were not involved in the design of this study as it involved only
33
34 105 observational analysis of an anonymised, pre-existing, routinely collected dataset.

35 36 106 **Ethics approval**

37
38 107 Since no primary data collection was undertaken, no patient or public was
39
40 108 involved, no formal ethical assessment or informed consent was required.

41 42 109 **Autoregressive moving average (ARMA) model**

43
44 110 ARMA model³⁰ is an important time series analysis and prediction model in
45
46 111 Box-Jenkins method, also known as auto-regression moving average model. The
47
48 112 ARMA (p,q) model is a model with autocorrelation order p and moving average order
49
50 113 q, p and q are judged by autocorrelation function (ACF) and partial autocorrelation
51
52 114 function (PACF) diagram of stationary data. If the original data is stable, the
53
54 115 autocorrelation coefficients are trailing, and the partial correlation coefficients are the
55
56 116 p-order truncated, then $q=0$, the ARMA(p,q) model is recorded as AR(p), and its
57
58 117 expression is as follows:
59
60

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

where, X_t is the observed value at t , ϕ_1, \dots, ϕ_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:

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

where, μ is the expected value of X_t . $\theta_1, \dots, \theta_p$ 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:

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

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

148 not sufficient, and it is necessary to consider improving the accuracy of the model.

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

150 **Elman neural network model**

151 The Elman neural network (see Figure 2) is proposed by Elman in 1990³¹. The
152 model is generally divided into four layers: input layer, hidden layer, receiving layer
153 and output layer. The characteristic of Elman network is that the output of the hidden
154 layer is connected to the input of the hidden layer through the delay and storage of the
155 receiving layer, which makes it sensitive to the data of the historical state. The
156 addition of the internal feedback network increases the ability of the network itself to
157 deal with dynamic information, thus achieving the purpose of dynamic modeling.

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

$$159 \quad x(k) = f(w^1 x_c(k) + w^2 u(k-1))$$

$$160 \quad x_c(k) = \alpha x_c(k-1) + x(k-1)$$

$$161 \quad y(k) = g(w^3 x(k))$$

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

168 There are four main steps in Elman neural network modeling:

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

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

178 Third step. Set the parameters of the Elman model, such as training epochs and
179 goals, etc., in our study, training epochs and goals of Elman neural network were set

180 to 2000 and 0.00001, respectively. To determine the number of neurons in the hidden
 181 layer, so that the error of the established Elman model is minimized. At present, there
 182 is no ideal analytical expression for the number of neurons in the hidden layer. The
 183 number of neurons has a great influence on the performance of the network. When the
 184 number of neurons is too large, it will lead to that network learning time is too long,
 185 generalization performances are not good, and even the convergence failure, but when
 186 the number of neurons is too small, fault-tolerant ability of the network is poor. In
 187 general, the number of neurons does not exceed 20.³² In this study, matlab cyclic
 188 structure was used to find the optimal number of neurons by comparing the RMSE
 189 values of Elman networks with neurons 1 to 20.

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

192 Model comparison measures

193 Three performance indexes, root mean square error (RMSE), mean absolute error
 194 (MAE) and mean absolute percentage error (MAPE) were used to assess the fitting
 195 and forecasting accuracy of two models. The smaller these three values are, the better
 196 the model is. Their expressions are as follows¹⁰:

$$197 \quad RMSE = \sqrt{\frac{\sum_{t=1}^n (X_t - \hat{X}_t)^2}{n}},$$

$$198 \quad MAE = \frac{\sum_{t=1}^n |X_t - \hat{X}_t|}{n},$$

$$199 \quad MAPE = \frac{\sum_{t=1}^n \left| \frac{X_t - \hat{X}_t}{X_t} \right| \times 100}{n},$$

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

202 Statistical software

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

204 Results

205 From January 2005 to December 2017, the number of reported TB cases was
 206 141984 in Kashgar, Xinjiang, the average annual TB cases were 7888 and the average
 207 annual incidence was 191.18. Figure 3 showed the time series graph of the TB

1
2
3
4 208 incidence. It can be seen from the Figure 3 that the curve of TB incidence showed
5 209 strong nonlinear characteristics from 2005 to 2014, and the TB incidence from 2015
6 210 to 2017 were significantly higher than that of previous years.

7
8 211 The data of TB incidence from January 2005 to December 2017 were divided
9 212 into two parts. The data from January 2005 to December 2016 were used to build
10 213 model and the data from January 2017 to December 2017 were used to test the model.

14 214 **Establishment of ARMA Model**

15
16
17 215 ARMA Modeling requires data stability, so, first of all, the stability of the
18 216 modeling part of the data was verified by ADF test. The results of ADF test showed
19 217 that p-value was less than 0.05(see Table 1), which indicated that the data was stable
20 218 and could be directly used to build the model. Secondly, the ACF and PACF graphs
21 219 of the modeling data were plotted (see Figure 4), it was obvious from Figure 4 that the
22 220 autocorrelation coefficients were trailing distribution, and the partial correlation
23 221 coefficients were almost a second-order truncated distribution, only at the lag 7, 8 and
24 222 9, the correlation coefficients were a little large. Based on this situation, we
25 223 considered establishing four models, such as AR(2), AR((1,2,7)), AR((1,2,8)) and
26 224 AR((1,2,9)). The least square method was used to test the parameters of the four
27 225 models. The results of the test were shown in Table 2; we can see that, of the four
28 226 models, only AR (2) model and AR((1,2,8)) passed the parameter test. Comparing the
29 227 two models, it was found that, the AR((1,2,8)) model had smaller AIC and SC values,
30 228 and the R^2 values of AR((1,2,8)) were larger than the R^2 values of AR (2), so the AR
31 229 (2) model was abandoned. Then, the residual analysis of the AR((1,2,8)) model was
32 230 carried out, the autocorrelation and partial correlation coefficient of the residuals were
33 231 almost all within two times standard deviation, and only in the lag 5,6,12, they were
34 232 beyond the range of two times standard deviation (see Figure 5), which indicated that
35 233 the AR((1,2,8)) model could be used to roughly predict the TB incidence in Kashgar.
36 234 We used AR((1,2,8)) model to fit the TB incidence from September 2005 to
37 235 December 2016, the fitting RMSE ,MAE and MAPE were 6.15, 4.33 and 0.2858,
38 236 respectively; we used AR((1,2,8)) model to predict the TB incidence from January
39 237 2017 to December 2017, the prediction RMSE, MAE and MAPE were 10.88 , 8.75

238 and 0.2029, respectively.

239 **Establishment of AR-Elman Model**

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

$$248 \quad \text{input matrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_i \\ x_2 & x_3 & \dots & x_{i+1} \\ \dots & \dots & \dots & \dots \\ x_N & x_{N+1} & \dots & \dots \end{bmatrix}, \text{output matrix} = [x_{N+1} \quad x_{N+2} \quad \dots \quad x_{N+i}]$$

249 We selected twelve as the number of input layers of AR-Elman network and
 250 one as the number of output layers representing the forecast value. By the matlab
 251 cyclic structure, we selected the optimal number of neurons between 1 and 20, and
 252 finally we found when the number of neurons was 6 (see Figure 6), the RMSE was the
 253 smallest, and the AR-Elman was optimal. We used the AR-Elman model to fit the
 254 training data, RMSE was 3.78, MAE was 3.38, MAPE was 0.1837, and the R^2 of the
 255 model was 0.83; we used the AR-Elman model to predict the TB incidence from
 256 January 2017 to December 2017, RMSE was 8.86, MAE was 7.29, and MAPE was
 257 0.2006. The fitting curves of AR((1,2,8)) model and AR-Elman model, and the
 258 prediction curve of AR-Elman model were shown in Figure 7. Comparison results of
 259 the AR((1,2,8)) model and AR-Elman model were shown in Table 3, both the fitting
 260 RMSE, MAE and MAPE and the predicting RMSE, MAE and MAPE of the
 261 AR-Elman model were smaller than those of the single AR((1,2,8)) model, which
 262 indicated that the AR-Elman combined model established in this study was more
 263 suitable for predicting the TB incidence in Kashgar.

264 **Discussion**

265 According to the WHO 2019 Global Tuberculosis report ⁴, around the world, TB
 266 mortality was down about 3% every year, the incidence was down about 2% every
 267 year, 16% of TB patients died of the disease. ⁴ But the rate of decline has not reached
 268 the pace of the stop Tuberculosis Strategy Plan. Therefore, it is necessary to

1
2
3
4 269 strengthen the prevention and control of tuberculosis. In order to significantly narrow
5
6 270 these gaps, greater progress must be made in a group of countries with a high burden
7
8 271 of tuberculosis. The burden of TB in China ranks second in the world, and Xinjiang is
9
10 272 the province with high incidence of TB in China, and Kashgar is the area with the
11
12 273 high TB incidence in Xinjiang. Therefore, it is urgent to do a good job in the
13
14 274 prevention and control of TB in Kashgar.

15
16 275 The prediction and early warning of infectious diseases is an important link in
17
18 276 the prevention and control of infectious diseases.³²⁻³⁴ Therefore, this study carried out
19
20 277 research from the point of view of prediction to explore an accurate prediction model
21
22 278 and do prediction analysis of TB incidence in Kashgar, so as to provide scientific
23
24 279 reference for the prevention and control of the disease in this area. The Box -Jenkins
25
26 280 method is a popular time series prediction method, this method has good prediction
27
28 281 performance and high prediction accuracy; Elman Neural network can capture
29
30 282 nonlinear information of time series data very well. In this study, the two methods
31
32 283 were combined to study the prediction model of TB incidence in Kashgar.

33
34 284 Many studies have found that Box-Jenkins method has a good ability of fitting
35
36 285 and forecasting. For stationary time series that do not contain seasonality, it is more
37
38 286 suitable to use the ARMA model of the Box-Jenkins method to do prediction analysis
39
40 287 ³⁵, for non-stationary time series of infectious diseases with obvious seasonality, it is
41
42 288 more suitable to use seasonal autoregressive integrated moving average (SARIMA)
43
44 289 model of the Box-Jenkins method for prediction analysis.⁹⁻¹² In our study, from Figure
45
46 290 3, we could see that the seasonality of the TB incidence in Kashgar from 2005 to 2014
47
48 291 was not obvious, there was only a certain seasonality from 2015 to 2017, and we
49
50 292 found that the time series of TB incidence was stable by ADF unit root test, and the
51
52 293 autocorrelation and partial correlation coefficients of modeling data at lag 12, 24 were
53
54 294 not obviously large, therefore, for our research data, we used ARMA model to do
55
56 295 forecast analysis, finally, we established AR((1,2,8)) model of the Box-Jenkins
57
58 296 method, it has a good performance to fit and predict the TB incidence of Kashgar in
59
60 297 Xinjiang. From Figure 3, we can also see that the time series of TB incidence has
298 strong non-linear, the established AR((1,2,8)) model mainly extracted the linear
299 information of data, considering that the neural network can capture the non-linear
300 information of data well, in order to improve the prediction accuracy of TB incidence

1
2
3
4 301 in Kashgar, we used AR((1,2,8)) model and Elman neural network model to establish
5 302 AR-Elman hybrid model. Many studies have found that the combination model could
6
7 303 improve accuracy of prediction, such as, Wang et al.²⁸ found that SARIMA-NAR
8
9 304 hybrid model had an outstanding ability to improve the prediction accuracy relative to
10
11 305 SARIMA model and nonlinear autoregressive network (NAR) model when they were
12
13 306 used to predict pertussis incidence in China. Li et al.²⁷ found ARIMA-GRNN hybrid
14
15 307 model was superior to single ARIMA model in predicting the short-term TB
16
17 308 incidence in Chinese population. Our research was consistent with these literatures
18
19 309 that our AR-Elman hybrid model was more accurate than the single AR((1,2,8))
20
21 310 model.

22
23 311 In the past few years, Xinjiang's economic development was relatively backward,
24
25 312 medical resources were scarce, diagnosis and treatment were delayed, the continuous
26
27 313 spread of TB has become a difficult problem in the control of TB in Xinjiang. In
28
29 314 recent years, Xinjiang has introduced many new policies to increase investment in TB
30
31 315 prevention and control, and the relevant departments of disease prevention and control
32
33 316 in Xinjiang have also done a lot of effective work, which has helped to control
34
35 317 effectively rapid increase of the TB incidence in Xinjiang. In order to do a good job in
36
37 318 the prevention and control of TB in Xinjiang, many departments need to make joint
38
39 319 efforts. Our research was mainly to build a high-precision prediction model to help
40
41 320 early warning and prediction analysis of tuberculosis in Kashgar. Finally, we
42
43 321 established the AR-Elman hybrid model, which had high fitting and prediction
44
45 322 accuracy of TB incidence in Kashgar, Xinjiang.

46
47 323 Our study found that Box-Jenkins and Elman neural network hybrid method was
48
49 324 an effective method for predicting the incidence of TB in Kashgar, it could provide a
50
51 325 scientific reference for prediction analysis of TB incidence. However, our study also
52
53 326 has some limitations: our method was only suitable for short-term prediction,
54
55 327 long-term prediction performance may decline, two main reasons: first, our model
56
57 328 was based on historical data characteristics; second, climatic factors, environmental
58
59 329 factors, demographic factors and political issues may have certain impacts on the
60
61 330 change of incidence. Therefore, if the established model becomes old and people will
62
63 331 want to obtain more accurate prediction results, it will be needed to adjust the model
64
65 332 parameters, update the model based on the new modeling sample data, and then to do
66
67 333 prediction analysis.

334 **Conclusions**

1
2
3
4 335 Kashgar has a very high TB incidence, in order to provide some help for the
5 336 prevention and control of this disease, the prediction problem of the TB incidence was
6
7 337 studied. Firstly, a single AR((1,2,8)) prediction model was established by using
8
9 338 Box-Jenkins method, and its fitting and prediction performance were good, secondly,
10
11 339 in order to improve the prediction accuracy of the single AR((1,2,8)) model, we used
12
13 340 single AR((1,2,8)) and Elman neural network with strong ability to capture nonlinear
14
15 341 information to establish AR-Elman hybrid model. The fitting and prediction accuracy
16
17 342 of the hybrid model was higher than that of the single AR((1,2,8)) model. The
18
19 343 AR-Elman hybrid model can provide scientific reference for predicting and warning
20
21 344 the TB incidence in Kashgar, Xinjiang.

22 345

23 346 **Acknowledgements**

24 347 We would like to express our gratitude to peer reviewers for carefully revising
25
26 348 our manuscript and for their very useful comments.

27 349 **Contributions**

28
29 350 YLZ and XLZ analyzed the data and wrote the manuscript. XLZ, XJW, KW and
30
31 351 YC wrote and revised the manuscript. All authors read and approved the final
32
33 352 manuscript.

34 353 **Funding**

35
36 354 This work was supported by Major projects of science and technology in
37
38 355 Xinjiang Autonomous region (Grant No.2017A03006), China, and National Natural
39
40 356 Science Foundation Project of China (Grant No. 72064036, 82060609),and State Key
41
42 357 Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in
43
44 358 Central Asia Fund (Grant No. SKL-HIDCA-2020-9)

45 359 **Competing interests**

46 360 None declared.

47 361 **Patient consent for publication**

48 362 Not required.

49 363 **Provenance and peer review**

50 364 Not commissioned; externally peer reviewed.

51 365 **Data availability statement**

52 366 The data used in this study are available from the corresponding authors on
53
54 367 reasonable request.

55 368 **References**

56
57
58
59
60

- 1
2
3
4 369 1. Kemal J, Sibhat B, Abraham A, et al. Bovine tuberculosis in eastern Ethiopia:
5
6 370 prevalence, risk factors and its public health importance. *BMC Infectious Diseases*,
7
8 371 2019; 19(1).
- 9
10 372 2. Tilahun M, Ameni G, Desta K, et al. Molecular epidemiology and drug sensitivity
11
12 373 pattern of *Mycobacterium tuberculosis* strains isolated from pulmonary
13
14 374 tuberculosis patients in and around Ambo Town, Central Ethiopia. *Plos One*, 2018;
15
16 375 13(2):e0193083-.
- 17
18 376 3. Reta A, Simachew A. The Role of Private Health Sector for Tuberculosis Control
19
20 377 in Debre Markos Town, Northwest Ethiopia. *Advances in Medicine*,
21
22 378 2018;2018(2):1-8.
- 23
24 379 4. WHO. Global tuberculosis report 2019, [https://www.who.int/tb/publications/
25
26 380 global_report/en/](https://www.who.int/tb/publications/global_report/en/). (Accessed on 17 Oct 2019).
- 27
28 381 5. Zhao YF, Shou MH, Wang ZX. Prediction of the Number of Patients Infected with
29
30 382 COVID-19 Based on Rolling Grey Verhulst Models. *Int J Environ Res Public
31
32 383 Health*. 2020; 17(12):4582.
- 33
34 384 6. Guan P, Wu W, Huang D. Trends of reported human brucellosis cases in mainland
35
36 385 China from 2007 to 2017: an exponential smoothing time series analysis. *Environ
37
38 386 Health Prev Med*. 2018; 23(1):23.
- 39
40 387 7. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, Xu JY, Shi J, Sun DY.
41
42 388 Analysis and predication of tuberculosis registration rates in Henan Province,
43
44 389 China: an exponential smoothing model study. *Infect Dis Poverty*. 2020;9(1):123.
- 45
46 390 8. Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling
47
48 391 of time series of dengue disease case counts. *PLoS Negl Trop Dis*.
49
50 392 2017;11(7):e0005696.
- 51
52 393 9. Wang Y, Xu C, Zhang S, Wang Z, Zhu Y, Yuan J. Temporal trends analysis of
53
54 394 human brucellosis incidence in mainland China from 2004 to 2018. *Sci Rep*.
55
56 395 2018;8(1):15901.
- 57
58 396 10. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S. An Advanced Data-Driven
59
60 397 Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series
398 Forecasting in Qinghai Province, China. *Infect Drug Resist*. 2020;13:867-880.

- 1
2
3
4 399 11. Aryee, Kwarteng , Essuman, Nkansa Agyei, et al. Estimating the incidence of
5
6 400 tuberculosis cases reported at a tertiary hospital in Ghana: a time series model
7
8 401 approach. BMC Public Health.2018;18(1):1292.
- 9
10 402 12. Tohidinik H R , Mohebalı M , Mansournia M A, et al. Forecasting zoonotic
11
12 403 cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran:
13
14 404 A SARIMA Analysis. Tropical Medicine & International Health,
15
16 405 2018;23(8):860–869.
- 17
18 406 13. Ouedraogo B, Inoue Y, Kambiré A, et al. Spatio-temporal dynamic of malaria in
19
20 407 Ouagadougou, Burkina Faso, 2011–2015. Malaria Journal, 2018;17(1):138.
- 21
22 408 14. Wagenaar B H, Augusto O, Beste J, et al. The 2014–2015 Ebola virus disease
23
24 409 outbreak and primary healthcare delivery in Liberia: Time-series analyses for
25
26 410 2010–2016. Plos Medicine, 2018; 15(2):e1002508.
- 27
28 411 15. Liu S, Chen J, Wang J, et al. Predicting the outbreak of hand, foot, and mouth
29
30 412 disease in Nanjing, China: a time-series model based on weather variability.
31
32 413 International Journal of Biometeorology, 2017;(6):1-10.
- 33
34 414 16. Anokye R, Acheampong E, Owusu I, et al. Time series analysis of malaria in
35
36 415 Kumasi: Using ARIMA models to forecast future incidence. Cogent Social
37
38 416 Sciences, 2018; 4(1):1461544.
- 39
40 417 17. Wang Y W , Shen Z Z , Jiang Y . Comparison of autoregressive integrated moving
41
42 418 average model and generalised regression neural network model for prediction of
43
44 419 haemorrhagic fever with renal syndrome in China: a time-series study. BMJ Open,
45
46 420 2019; 9(6):e025773.
- 47
48 421 18. Guo W L , Geng J , Zhan Y , et al. Forecasting and predicting intussusception in
49
50 422 children younger than 48 months in Suzhou using a seasonal autoregressive
51
52 423 integrated moving average model. BMJ Open, 2019; 9(1).
- 53
54 424 19. Gabriel A F B , Alencar A P , Miraglia S G E K . Dengue outbreaks: unpredictable
55
56 425 incidence time series. Epidemiology & Infection, 2019; 147.
- 57
58 426 20. Tian C W , Wang H , Luo X M . Time-series modelling and forecasting of hand,
59
60 427 foot and mouth disease cases in China from 2008 to 2018. Epidemiology &
428 Infection, 2019; 147.

- 1
2
3
4 429 21. Juang W C , Huang S J , Huang F D , et al. Application of time series analysis in
5 430 modelling and forecasting emergency department visits in a medical centre in
6 431 Southern Taiwan. *Bmj Open*, 2017; 7(11):e018628.
- 9 432 22. Yu C, Li Y, Zhang M. An improved Wavelet Transform using Singular Spectrum
11 433 Analysis for wind speed forecasting based on Elman Neural Network. *Energy*
13 434 *Conversion & Management*, 2017; 148:895-904.
- 15 435 23. Huang Y, Shen L. Elman Neural Network Optimized by Firefly Algorithm for
17 436 Forecasting China's Carbon Dioxide Emissions. *Communications in Computer*
19 437 *and Information Science*, 2018;(951): 36-47.
- 21 438 24. Alkhasawneh M S. Hybrid Cascade Forward Neural Network with Elman Neural
23 439 Network for Disease Prediction. *Arabian Journal for Science and Engineering*,
25 440 2019;44(11): 9209–9220.
- 27 441 25. Mehrhini B, Izadi H, Memarian H. Shear wave velocity prediction using Elman
29 442 artificial neural network. *Carbonates & Evaporites*, 2017;(6):1-11.
- 31 443 26. Khalaf M, Hussain A J, Keight R, et al. Machine learning approaches to the
33 444 application of disease modifying therapy for sickle cell using classification
35 445 models. *Neurocomputing*, 2017; 228(C):154-164.
- 37 446 27. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, Ji Y, Xu D, Wang J. Application of a
39 447 hybrid model in predicting the incidence of tuberculosis in a Chinese population.
41 448 *Infect Drug Resist.* 2019;12:1011-1020.
- 43 449 28. Wang Y, Xu C, Wang Z, Zhang S, Zhu Y, Yuan J. Time series modeling of
45 450 pertussis incidence in China from 2004 to 2018 with a novel wavelet based
47 451 SARIMA-NAR hybrid model. *PLoS One*. 2018;13(12):e0208404.
- 49 452 29. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, Yuan J. Temporal trends
51 453 analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a
53 454 new SARIMA-NARNNX hybrid model. *BMJ Open*. 2019;9(7):e024409.
- 55 455 30. Box G E, Jenkins G M. *Time series analysis: forecasting and control rev. ed.*
57 456 *Oakland, California, Holden-Day*, 1976; 31(4):238-242.

- 1
2
3
4 457 31. Ming Cheng. The principle and example of MATLAB neural network. Tsinghua
5 458 University Press, 2013.
- 6
7
8 459 32. Huppert A, Katriel G. Mathematical modelling and prediction in infectious
9 460 disease epidemiology. Clin Microbiol Infect. 2013;19(11):999-1005
- 10
11 461 33. Wearing HJ1, Rohani P, Keeling MJ. Appropriate models for the management of
12 462 infectious diseases. PLoS Med. 2005; 2(7):e174.
- 13
14 463 34. Neuberger A, Paul M, Nizar A, Raoult D. Modelling in infectious diseases:
15 464 between haphazard and hazard. Clin Microbiol Infect. 2013;19(11):993-8.
- 16
17 465 35. Tipirneni-Sajja A, Krafft AJ, Loeffler RB, Song R, Bahrami A, Hankins JS,
18 466 Hillenbrand CM. Autoregressive moving average modeling for hepatic iron
19 467 quantification in the presence of fat. J Magn Reson Imaging.
20 468 2019;50(5):1620-1632.
- 21
22
23
24
25
26
27
28 469

30 470 **Figures**

- 31
32 471 **Figure 1.** The red part of this picture is the location of Kashgar in Xinjiang, China.
33
34 472 Kashgar is located in the south of Xinjiang, and it has a very high incidence of
35 473 tuberculosis.
36
37
38 474
- 39
40 475 **Figure 2.** The Structure diagram of Elman neural network. w^1 , w^2 and w^3 are the
41 476 connection weight matrixes. $x_c(k)$ and $x(k)$ represent the output of the contact unit and
42 477 the hidden layer unit, respectively, $y(k)$ represents the output of the output unit, $u(k-1)$
43 478 represents the input of the input unit.
44
45
46
47
48 479
- 49
50 480 **Figure 3.** Graph of the tuberculosis (TB) incidence in Kashgar from January 2005 to
51 481 December 2017. The curve of TB incidence showed strong nonlinear characteristics
52 482 from 2005 to 2014, and the TB incidence increased significantly from 2015 to 2017.
53
54
55 483
- 56
57 484 **Figure 4.** Autocorrelation function (ACF) and partial autocorrelation function (PACF)
58 485 graphs of modeling data. As the delay of the lag order, the autocorrelation coefficients

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

488

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

493

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

497

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

505

506

507 Tables

508 **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	

509

510

511 **Table 2.** Parameter estimates of the tentative models with their Akaike information
 512 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		

513

514 Table 3. Comparison results of in-sample fitting and out-of-sample forecasting
 515 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

516

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



Figure 1

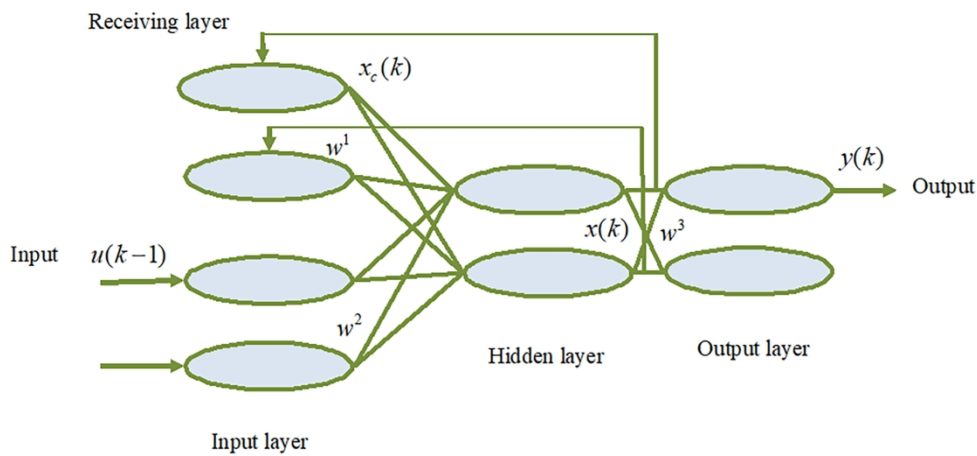


Figure 2

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

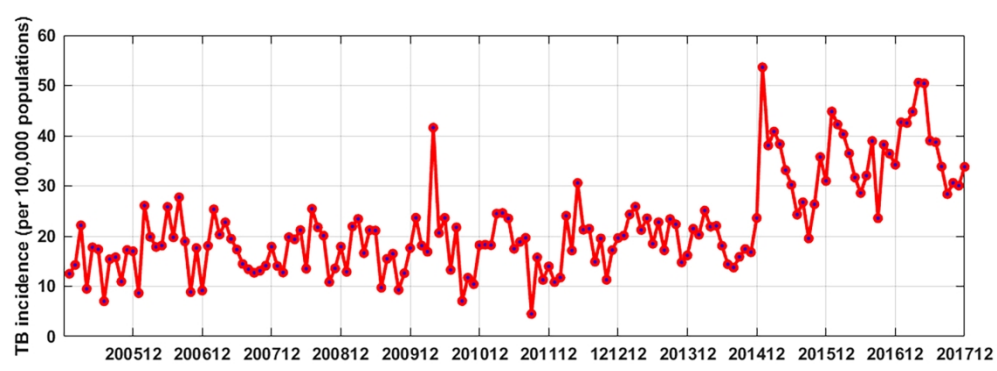


Figure 3

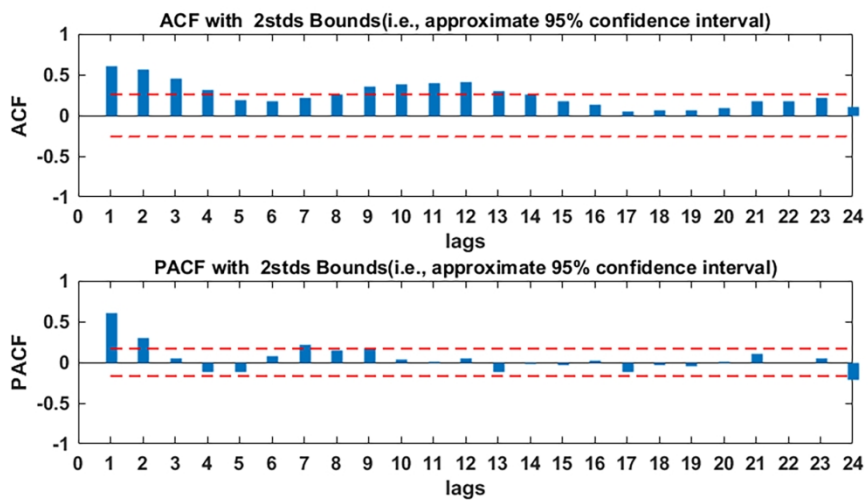


Figure 4

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

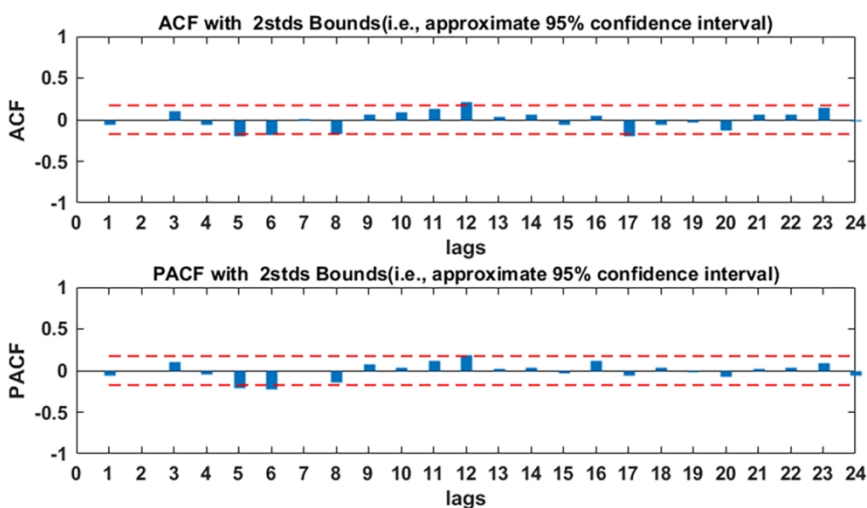


Figure 5

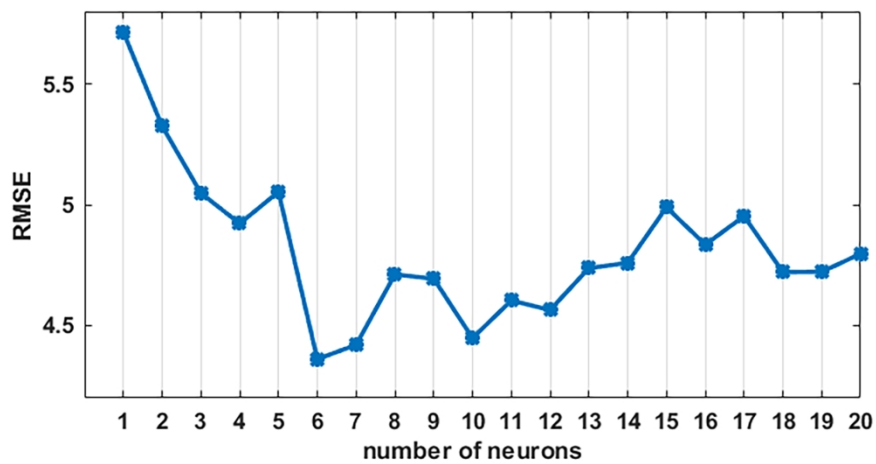


Figure 6

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

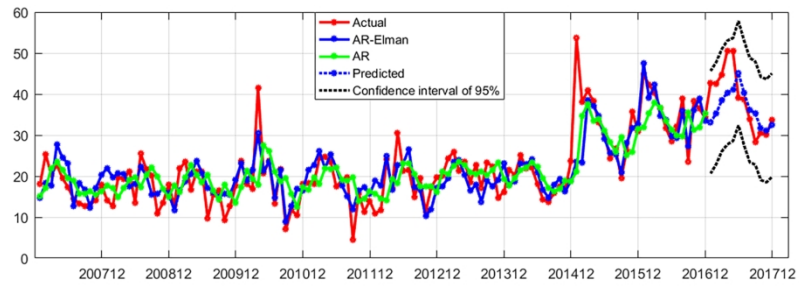


Figure 7