Analysis and Prediction of Pulmonary Tuberculosis Using an ARIMA Model in Shaanxi Province, China

Cong Yang1,*, Yali Yang1, Zhiwei Li1 and Yan Li2
1Department of Basic Sciences, Air Force Engineering University, Xi'an, Shaanxi 710051, China;
2Institute of Tuberculosis Prevention and Treatment in Shaanxi, Xi'an Shaanxi 710048, P.R. China

*Corresponding author email: yangc0729@126.com

Abstract. An analysis and prediction for the incidence of tuberculosis (TB) is particularly important since TB still has a high fatality rate in the world. However, this prediction is often influenced by inaccurate forecasting ways. We used data from 364,762 reported TB cases between January 2005 and December 2015 in Shaanxi Province, China. The known number of cases in 2016 was used to assess the accuracy of the model’s predictions. Through all aspects of analysis and comparison, the ARIMA (0, 1, 2) (0, 1, 1)12 were the most model. In the fitting dataset, for the ARIMA (0, 1, 2) (0, 1, 1)12 model, RMSE, MAPE, MAE and MER were 0.7667, 6.7810, 6.04944 and 0.06836, respectively; And in the forecasting dataset were 0.32808, 6.01834, 0.2899 and 0.0615, respectively. The model can predict the seasonal changes and trends of tuberculosis in the Shaanxi province’s population.

1. Introduction
Tuberculosis (TB) is a disease caused by bacterial infection, which mainly affects the lungs but also can affect any other organ. A normal model is autoregressive integrated moving average (ARIMA) model, this is a time series model first put forward in the 1970s. This model works by treating the order of data to be predicted as a random sequence. The model is simple in structure, internal parameter is needed, and the prediction accuracy is high. ARIMA model was used in earthquake prediction such diseases as malaria, influenza, hemorrhagic fever or hand, foot and mouth disease. So, the model has become a widely used method in fields such as biological sciences. In predicting the incidence of malaria in Afghanistan, Anwar et al. used the model ARIMA (4,1,1)(1,0,1)12. When using ARIMA model to predict the hemorrhagic fever, Li et al. used ARIMA(0,1,1)(2,1,0)12 model. In predicting smear-positive TB cases, Mahmood et al. constructed the ARIMA(0,1,1)(0,1,1)12 model to complete the prediction.

2. Construction of the ARIMA Model
ARIMA(p, d, q) model is included in autoregressive (AR), moving average (MA), and ARMA models. All of models are suitable for data in order to better analyze future points or prediction sequence (forecasting). The significance of parameters p, d and q corresponding to autoregressive order, variance degree and moving average order respectively. The ARIMA model is generally recognized as the best model for forecasting non-seasonal data. Since the data in this study are highly seasonal, we use the ARIMA (SARIMA [p, d, q] [P, D, Q] s) seasonal model. P, D and Q represent seasonal autoregression, seasonal integration and seasonal moving average, respectively. s represent seasonal cycle length. When
building an ARIMA model, it is generally divided into three steps: analysis, fitting and prediction. First, we get the stationary data. When the original data is unstable, we conduct differential processing on this data. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the model residual are drawn from the obtained stationary sequence. Second, we first determine the parameters of the seasonal parts (P and Q), and then define the Bayesian information criterion (BIC) according to the Akaike information criterion (AIC), where the minimum data obtained is the optimal model. The residuals of ACF and PACF and the white test were used to assess the fitness. Finally, we predict the incidence of every month in 2016. Then, a comparative analysis was conducted based on the known data in 2016 to verify the predictive power of this model. This model takes the dataset of previous years as the original data, performs data fitting, and predicts the data of 2016. The accuracy of the forecast was analyzed using RMSE, MAPE, MER, and MAE.

3. Results
Monthly bulletin sequence charts show a downward trend and seasonal fluctuations (see Figure 1). The trough of the disease mostly occurs in November and December (s=12). The time series analysis model needs the data of stable sequence, but the original data is not stable enough. We have to make a difference to the initial data. Usually we choose first order or second order, in which the sequence after the first difference is stable (d=1, D=1). The ADF test is important in the time series, indicating that it is a stationary sequence(P<0.001). Figure 2(A) shows the ACF and PACF diagrams of the static notification series. At the seasonal part, ACF graph and PACF graph have important peaks at the lag of 12, while ACF graph and PACF graph have no important peaks at the lag of 24(P=0, Q=1). For the non-seasonal part, after preliminary analysis, we summarize the feasibility of 8 models, since ACF and PACF graphs don’t performance obvious patterns. Table 1 lists the AIC and BIC values of eight paradoxical models. Because its AIC and BIC values are the minimum, we selected ARIMA (0,1,2) (0,1,1)12 which is considered to be the optimal model. All of the model’s parameters were clear (P<0.001). All of the basic parameters of the model are shown in table 2. As we see in figure 2(B), the remaining ACF and PACF graphs are all white noise because the coefficients were not significantly correlated. Ljung-Box test proves that the residual is noise, not data error (P>0.05). Although both ACF and PACF exceeded the rated value, since it only happened once in the 24 lag times, the probability of occurrence is very small, so we can still regard it as an accidental situation. Therefore, the monthly incidence of TB in 2016 was predicted using the ARIMA (0,1,2) (0,1,1)12 model by fitting the previous data, which are shown in table 3. The prediction accuracy of the ARIMA model can meet the needs of normal prediction, and the relative error of the predicted value is less than 10% per month. The number of fitting data from 2005 to 2015 is 36,423; the number of forecasts for 2016 is 3,419.

![Figure 1. Monthly TB notification rate in Shaanxi province, China](image-url)
Figure 2. The ACF and PACF of the TB notification sequence after a non-seasonal and a seasonal difference in Figure 2(A), the residuals of ARIMA (0,1,2) (0,1,1)12 models in Figure 2(B).

Table 1. The AIC and BIC values of the model that meet the first step requirements.

| Type                          | AIC               | BIC               |
|-------------------------------|-------------------|-------------------|
| ARIMA (0,1,1) (0,1,1)12       | 61.55212447926002 | 69.83864628365329 |
| ARIMA (0,1,2) (0,1,1)12       | 58.26856281284353 | 69.28292357726899 |
| ARIMA (1,1,0) (0,1,1)12       | 73.79520523824802 | 82.107259111645   |
| ARIMA (1,1,1) (0,1,1)12       | 58.22557018109034 | 69.27426592028137 |
| ARIMA (1,1,2) (0,1,1)12       | 58.54934041342724 | 72.3172913689591  |
| ARIMA (2,1,0) (0,1,1)12       | 64.76815894410252 | 75.85089744196519 |
| ARIMA (2,1,1) (0,1,1)12       | 61.22498036588064 | 75.03585003986942 |
| ARIMA (2,1,2) (0,1,1)12       | 60.51648273892411 | 77.038023855623   |

Table 2. Parameter estimation of ARIMA (0,1,2) (0,1,1)12 model.

| Parameter               | Coefficient | Standard error | t       | P      |
|-------------------------|-------------|----------------|---------|--------|
| Moving average, lag 1   | -0.4013     | 0.076          | -5.295  | 0.000  |
| Moving average, lag 2   | -0.2869     | 0.081          | -3.548  | 0.000  |
| Seasonal moving         | -0.3068     | 0.076          | -4.058  | 0.000  |

4. Discussion
The fatality rate of this disease is relatively low, considering the large population in our country, the annual death toll is still not optimistic, China has a high incidence of the disease in the world. In order to get the purpose of “ending tuberculosis”, precise prediction of tuberculosis has important biological value for meaningful control of tuberculosis. Based on the forecast datatest, we will implement pertinence prevention and regulatory measures and effectively distribution public resources. So far, many kinds of prediction models have been developed⁰. This paper studies the application of ARIMA model in the prediction of tuberculosis in Shaanxi province, China, and analyzes the prediction error. The ARIMA model predicts future data by fitting predictions based on past and present data⁰. According to whether the target sequence has seasonality or not, ARIMA model have two different series, which is more widely used than the general time series model and breaks the conventional limitations. The model is optimized to overcome the process of repeated identification, analysis, fitting and diagnosis⁰.
ARIMA model can meet the requirements of different time series analysis, which is a commonly used model at present. However, this model still has some disadvantages, which are generally applicable to the prediction in a short time, and generally applicable to the capture of linear relations. Because the development of tuberculosis is affected by some unknown factors, the trends in morbidity of tuberculosis is often non-linear, which cannot be effectively solved by ARIMA model.

Table 3. The ARIMA model was used to predict the monthly TB notification rate in 2016.

| Month    | Actual rate (1/100,000) | Predicted rate (1/100,000) | Relative error (%) |
|----------|-------------------------|-----------------------------|--------------------|
| January  | 5.24752                 | 5.62395                     | 7.1734             |
| February | 4.88196                 | 4.5305                      | 7.1992             |
| March    | 5.98921                 | 6.06697                     | 1.2983             |
| April    | 4.96673                 | 5.21335                     | 4.9644             |
| May      | 4.81309                 | 5.19226                     | 7.8779             |
| June     | 4.60118                 | 4.40752                     | 4.3478             |
| July     | 4.47668                 | 4.50129                     | 0.6711             |
| August   | 4.68065                 | 4.47466                     | 4.4871             |
| September| 4.22503                 | 4.5482                      | 7.6449             |
| October  | 4.11907                 | 3.87752                     | 5.8752             |
| November | 4.13762                 | 4.1153                      | 0.5394             |
| December | 4.429                   | 4.29215                     | 3.0932             |

5. Conclusion
The ARIMA model could be predicted trends in the incidence of TB in the Shaanxi Province, which has a high accuracy in predicting morbidity. At present, there is no unified model to predict. Although it is possible to achieve more accurate modelling by considering the way in which data should be counted locally. As a result, the model remains the dominant method for predicting TB incidence.

Acknowledgments
Supported by the Graduate Scientific Research Foundation of Department of Basic Sciences.

References
[1] Lin Y, Chen M, Chen G, Wu X, Lin T. Application of an autoregressive integrated moving average model for predicting injury mortality in Xiamen, China. BMJ Open. 2015;5(12): e008491. doi:10.1136/bmjopen-2015-008491
[2] Anwar MY, Lewnard JA, Parikh S, Pitzer VE. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. Malar J. 2016;15(1):566.
[3] He Z, Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: a nine-year retrospective study. Int J Infect Dis. 2018; 74:61–70.
[4] Li Q, Guo NN, Han ZY, et al. Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome. Am J Trop Med Hyg. 2012;87(2):364–370. doi:10.4269/ajtmh.2012.11-0472
[5] Liu L, Luan RS, Yin F, Zhu XP, Lu Q. Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model. Epidemiol Infect. 2016;144(1):144–151. doi:10.1017/S0950268815001144
[6] Wang CH, Mo LR, Lin RC, Kuo JJ, Chang KK, Wu JJ. Artificial neural network model is superior to logistic regression model in predicting treatment outcomes of interferon-based combination therapy in patients with chronic hepatitis C. Intervirology. 2008;51 (1):14–20. doi:10.1159/000118791
[7] Baxt WG. Application of artificial neural networks to clinical medicine. Lancet. 1995;346(8983):1135–1138. doi:10.1016/s0140-6736 (95)91804-3
[8] Khan MT, Kaushik AC, Ji L, Malik SI, Ali S, Wei DQ. Artificial neural networks for prediction of tuberculosis disease. Front Microbiol. 2019; 10:395. doi:10.3389/fmicb.2019.00395

[9] Anwar MY, Lewnard JA, Parikh S, Pitzer VE. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. Malar J. 2016;15(1):566.

[10] Li Q, Guo NN, Han ZY, et al. Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome. Am J Trop Med Hyg.2012;87(2):364–370. doi:10.4269/ajtmh.2012.11-0472

[11] MOOSAZADEH M, KHANJANI N, NASEHI M, BAHRAMPOURA. Predicting the incidence of smear positive tuberculosis cases in Iran using time series analysis. Iran J Public Health. 2015;44 (11):1526–1534.

[12] Song X, Xiao J, Deng J, Kang Q, Zhang Y, Xu J. Time series analysis of influenza incidence in Chinese provinces from 2004 to 2011. Medicine (Baltimore)2016;95(26)e3929.

[13] WHO. Global tuberculosis report 2018. Available from: https://www.who.int/tb/publications/global_report/en/. Accessed December 18, 2018.

[14] Li Z, Wang Z, Song H, et al. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infect Drug Resist. 2019; 12:1011–1020. doi:10.2147/IDR.S190418.

[15] Brockwell PJ, Davis RA. Time series: theory and methods. Technometrics. 1989; 31(1):121. doi:10.1080/00401706.1989.10488 491.