Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/86/20240829

Forecasting Trends in the Number of Tuberculosis Cases in China Using the SARIMA Model

Kaiwen Huang[1,a,*]

[1]*Yanjing Medical College, Capital Medical University, Dadong Lu, Beijing, China*
a. calvin_h430@mail.ccmu.edu.cn
**corresponding author*

Abstract: Development of a seasonal ARIMA model to forecast the trend and number of tuberculosis cases in China. It is useful for analysing tuberculosis prevention and control, treatment and the detection of related external variables from public health dimension. SARIMA model were used to modelling data on tuberculosis incidence cases in China from 2004 to 2018, and comparing and analysing the fitted values of the model with the actual values. The ARIMA (4, 1, 0) (0, 1, 1) [12] function has a MAPE of only 4.07 and a value of 0.89 for the R-squared, which is able to fit the 2004-2018 TB case data well. The error rate in January, the peak month of incidence, was 2.81%. The error rate in October, the turning point of the incidence rate, was 0.08 %. As a predictive model, this is a relatively good fit goodness of fit reference value. This result can provide time-scale suggestions for the development of TB prevention and control measures, the arrangement of medical resources, and the production of relevant drugs and medical devices.

Keywords: Time series analysis, SARIMA model, Tuberculosis, Forecast

1. Introduction

1.1. Research Background and Motivation

Tuberculosis (TB) is caused by bacteria and it most often affects the lungs. It is one of the 13th leading causes of death globally and the second leading cause of death from a single source of infection today after novel coronavirus pneumonia, posing a major threat to the lives and health of all human beings. China is one of the countries with a high burden of tuberculosis [1,2]. Based on this, the research question of this paper is how to establish a mathematical model to better predict the future development trend of tuberculosis cases in China.

1.2. Literature Review

Previous studies used ARIMA models of modeling TB infection data in different regions and over different time spans to obtain region-specific ARIMA models. There are not current answers for the TB forecasting model in the range of China [3-7]. The feasibility of the ARIMA model as a famous and influential time series prediction model has been demonstrated in studies related to the prediction of the number of cases of tuberculosis, and the author can likewise utilize the model for model fitting and prediction from different perspectives.

DOI: 10.54254/2754-1169/86/20240829

Previous studies have focused on smaller areas, such as a province or a city. Previous studies used ARIMA model or SARIMA model to explain the influencing factors of the number of cases of pulmonary tuberculosis, but it was rarely used in the range of entire nation and forecasting problems based on R-language (programming language). This paper expanded the topic to the entire China, and used SARIMA model based on R to explain and predict, to some extent, making up for the shortcomings of current research.

1.3. Research Contents

The main study is to build a prediction model using the SARIMA seasonal model after analysing data on tuberculosis cases in China from 2004 to 2018. This will provide theoretical and data support for the prevention and control of tuberculosis in China.

2. Methodology

Seasonal Autoregressive moving average model (ARIMA) is a well-known time series forecasting method, and the SARIMA model is generally adopted for time series with seasonal periodicity.

The model can comprehensively consider the effects of seasonal period, trend changes and random disturbances of the data, and is mostly used for infectious disease forecasting in the medical field. In this paper, the SARIMA model is used to forecast the incidence of Tuberculosis in China.

Examine and clean the data, for example, dealing with missing values, outliers, etc. Determine the frequency of the time series and convert the data into time series objects. Use Autocorrelation Function and Partial Autocorrelation Function plot to analyse data for trends and seasonality. Differentiate the data to ensure the stationary of the time series. The research target of this paper is the data of tuberculosis epidemic in China from 2004 to 2018, The data source is China Public Health Science Data Centre (https://www.phsciencedata.cn/Share/index.jsp), and the data indexes are the number of total tuberculosis cases in each month from January 2004 to December 2018.which where data from January 2004-December 2016 were used to fit the time series model and data from January 2017-December 2018 were used to test the model prediction effects.

The ARIMA model expression is ARIMA (p, d, q) (P, D, Q) $^{[s]}$. Parameters p, P and q, Q denote autoregressive and sliding average orders, d, D denote the number of differences, and s denotes the cycle length. Since d (D), q (Q) is generally not more than 2, the patchwork method is used to take the values 0, 1, 2 test and compare the models according to the overall significance of the model, goodness of fit and other indicators. Steps to establish ARIMA product seasonal model:

(1) Time series stationary: through the difference and (or) seasonal difference, making the original series to meet the ARIMA modelling stationary requirements.

(2) Model identification: draw ACF plot and PACF plot according to the stationary time series, initially determine the model parameters d, D, and S, and then determine the optimal order value by combining with the Bayesian information criterion minimum principle.

(3) Model diagnosis: Based on the parameters selected in (2), carry out model construction, model accuracy test, determine whether the statistical indicators of the selected parameters of the model are consistent with the model residuals, and further carry out diagnostic analysis to determine whether the model residuals are white noise sequences.

(4) Forecast evaluation: Use the selected optimal model to forecast the future value and sequence trend, and use the forecast value to compare with the actual value to evaluate the model prediction effect [8].

Draw a trend chart of the sequence using original data in time series format, as shown in Figure 1.

DOI: 10.54254/2754-1169/86/20240829

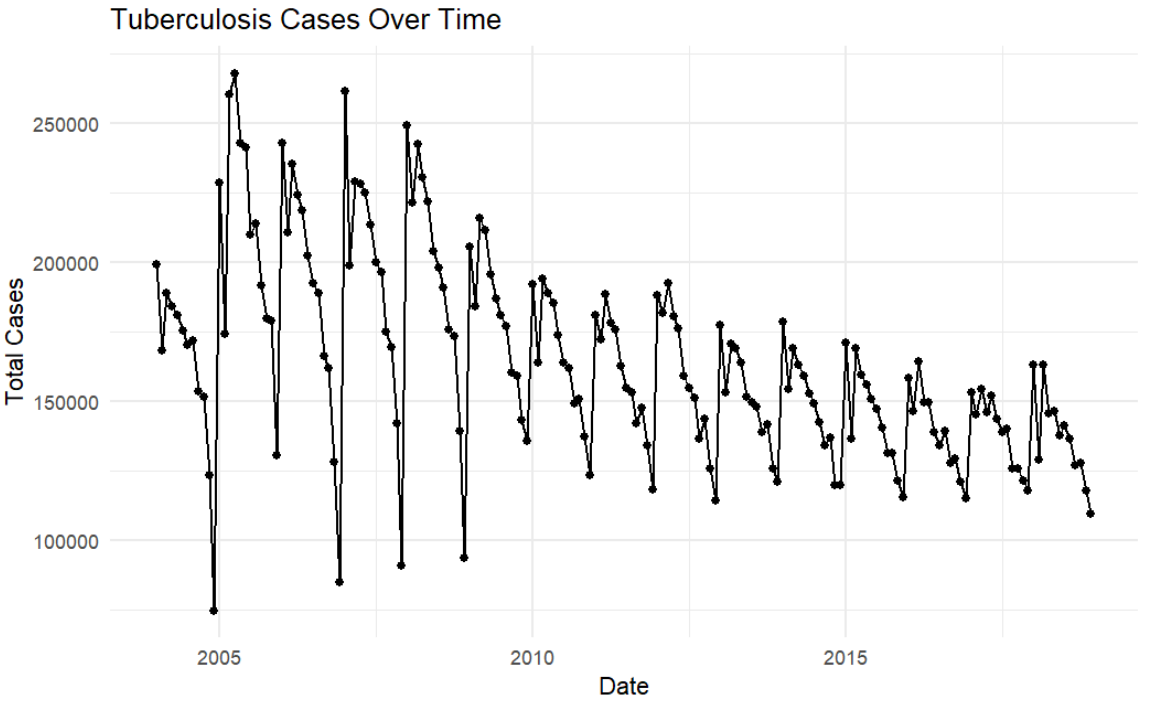

Figure 1: Tuberculosis cases over time from 2004-2018

The incidence of tuberculosis in China is clearly seasonal. Seasonal decomposition of the original time series shows that the incidence of tuberculosis is at its trough in February each year; the number of cases rises sharply to reach the peak in March, then declines slowly to reach the second trough in October; the overall change in the incidence of tuberculosis from October to February is small in magnitude and the curve is relatively stable.

3. Empirical Results Analysis

3.1. Unit Root Test

This article adopts the ADF unit root test, and the test results are shown in Table 1.

Table 1: Result of ADF test

Subject	Dicky-Fuller	Lag order	P-value
Value	-7.2881	5	0.01

As can be seen from Table 1, the P-value is less that 0.05, lower than the significance level, it indicates that the series is stationary.

3.2. Correlation Test

The author uses ACF plot and PACF plot of total cases to ensure the lag of ARIMA model, the results are shown in Figures 2-3.

DOI: 10.54254/2754-1169/86/20240829

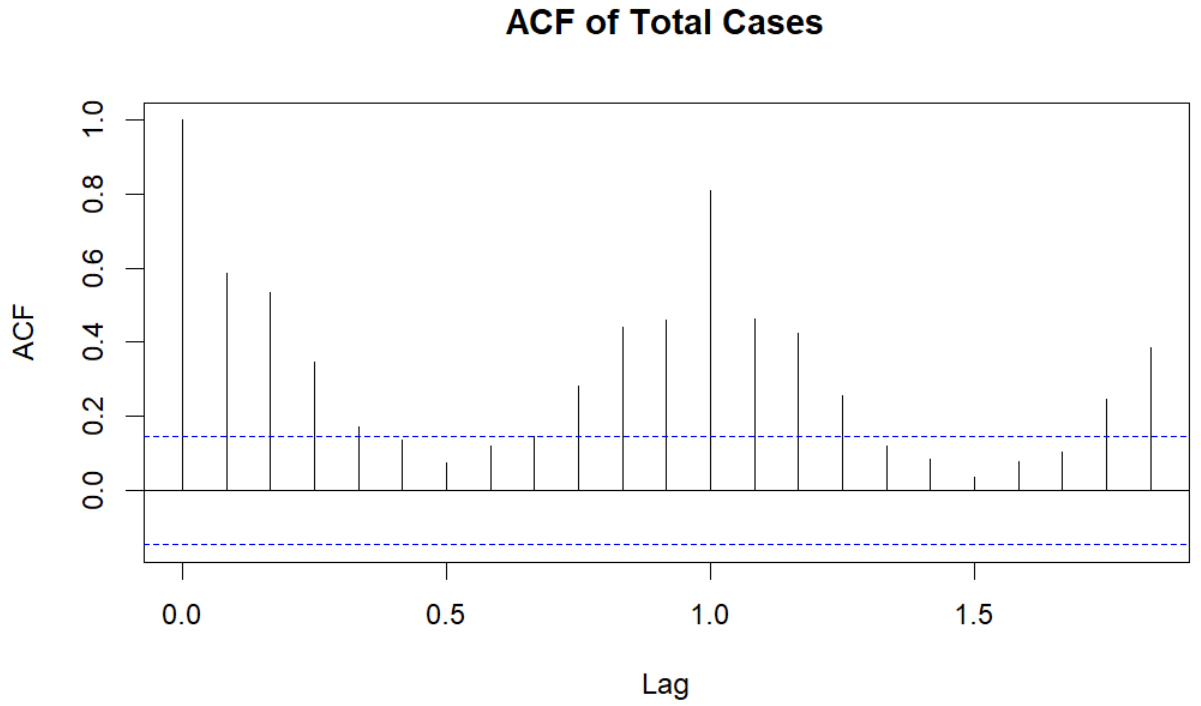

Figure 2: ACF plot for total cases

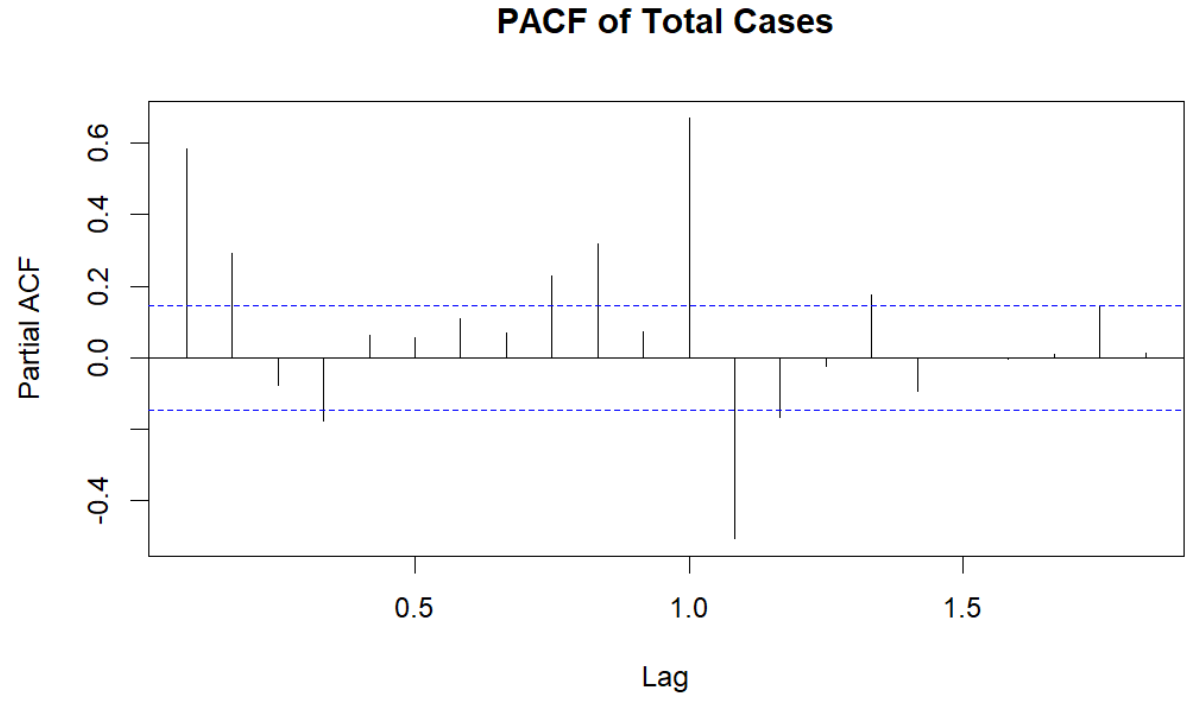

Figure 3: PACF plot of total cases

The decomposition and stationary process of the data series shows that the original data is a seasonal time series with a period of 12 months. The p-value of the original ADF test is less than 0.05, which indicates that the time series is already stationary and but in order to eliminate the seasonal pattern well one differencing process is required. The model parameters d=1, D=1 can be determined. p and q values need to be selected in conjunction with the ACF and PACF analyses of the

DOI: 10.54254/2754-1169/86/20240829

preprocessed series from Figure 2 and Figure 3 , and after preliminary observations and analyses, the values in p=l, 2, 3, 4, and q=0, 1 are selected. It is difficult to determine the parameter P and Q values because they rarely exceed the second-order difference, so they can be filtered among 0, 1 and 2. Taking into account the root mean squared error, minimum information criterion, mean absolute error, mean absolute scaled error, residual series and other relevant indicators, the author can compare and judge multiple models.

3.3. Model Estimation Results Comparisons

Multiple models were computed separately to obtain ARIMA(4, 1, 0), (0, 1, 1) and ARIMA(2, 1, 0), (2, 1, 1). Two alternative models, model parameter estimation tests and fitting results.

Table 2: Estimation results of parameters

Model Parameters	SARIMA (4,1,0)(0,1,1)[12]	SARIMA (2,1,0)(2,1,1)[12]
Stationary of residuals	Stationary	Stationary
Normality of residuals	Normality	Normality
BIC	3650.34	3657.65
RMSE	11782.65	12020.20
MAE	7010.77	7051.06
Ljung-Box test	0.4596	0.4902
R2	0.89	0.88
MAPE	4.07	4.09

From Table 2, it can be seen that both models conformed to the Ljung--Box test P>0.05, and it was concluded that the sequence residuals all conformed to the random sequence distribution. From the aspect of goodness of fit, the fitted coefficients R-squared values are high and the models are well fitted. From the t-test of the model coefficients, the coefficients of the model ARIMA(4, 1, 0)(0, 1, 1)[12] all pass the t-test (P<0.05), and the coefficients of SAR(1), SAR(2), and SMA(1) of the model ARIMA(2, 1, 0)(2, 1, 1) [12] do not pass the f-test (P> 0.05), so the model ARIMA(2. 1, 0) (2, 1, 1) [12] was discarded. The white noise Q statistic test for the residual series of model ARIMA(4, 1, 0)(0, 1, 1) [12], the P-value of the Q statistic under lag order 12, 18, 24, 30 are all > 0.05, so the residual series of the model are all white noise series. The model ARIMA(4, 1, 0)(0, 1, 1) is determined to be the optimal model.

DOI: 10.54254/2754-1169/86/20240829

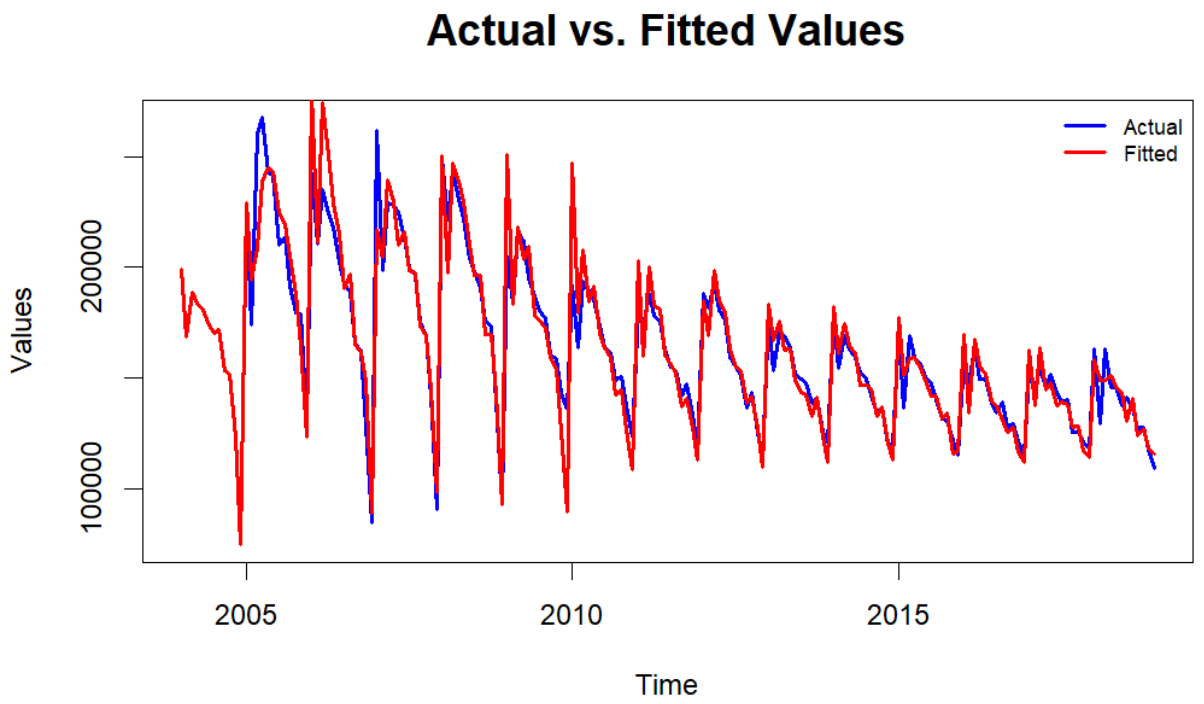

Figure 4: The plot of actual values and the fitted values by the model

Figure 4 shows the fitting curve graph of the model ARIMA(4, 1, 0)(0, 1, 1)[12] and the fitted value fits well with the measured value, which can better simulate the fluctuation pattern and seasonal characteristics of the original time series. The predicted values and actual values are shown in Table 3.

Table 3: The values of actual and fitted

Date	Actual values	Fitted values	E	ER (%)
2018-01	162819	158243	4576	2.810482806
2018-02	129026	148804.9	19778.9	15.32939098
2018-03	162829	148869.8	13959.2	8.572920057
2018-04	145371	150815.6	5444.6	3.745313715
2018-05	146360	146297.9	62.1	0.042429626
2018-06	137684	143675	5991	4.351268121
2018-07	141187	130139.3	11047.7	7.824870562
2018-08	136459	140867.3	4408.3	3.230494141
2018-09	126995	123926.4	3068.6	2.416315603
2018-10	127443	127339.7	103.3	0.081055845
2018-11	117685	118619.5	934.5	0.794068913
2018-12	109268	115460.1	6192.1	5.666892411

The model ARIMA(4, 1, 0)(0, 1, 1) [12]was used to predict the number of episodes in 2018, and the mean absolute percentage error MAPE of the model was 4.07 %, indicating that the overall prediction of the model was more satisfactory.

DOI: 10.54254/2754-1169/86/20240829

From the absolute error rate of each month, the prediction error rate of February was more than 10%, and the error rates of March and July were relatively high. Combined with the time series chart of TB incidence, February and March were the trough of incidence, and July was the slowdown of incidence reduction, which had less impact on the decision of prevention and control of TB, and the average error rate of prediction for other 9 months was 2.57%, which was the prediction accuracy to satisfy the overall assessment of the incidence of TB. Prediction accuracy meets the overall forecast of TB incidence. The error rate in January, the peak month of incidence, was 2.81%. The error rate in October, the turning point of the incidence rate, was 0.08 %, which can provide relatively accurate data to support the stockpiling of materials and staffing arrangements for specific prevention and control measures.

4. Discussion

The higher accuracy of the model's predictions in particular months, such as May, October, and November, may be due to the fact that these months are less affected by other external factors. Similarly, in months such as February, March, and July, the prediction errors were higher, possibly due to the impact of air quality and weather conditions on patients and TB infection conditions. In summary, without changing the model itself, the error rate of the model in predicting the number of TB cases is higher than 5 per cent only in February, March, July and December, and in the rest of the months, the accuracy of the model can provide data and theoretical support for the formulation of policies on prevention and control of TB, rationing of supplies, and mobilisation of personnel [9,10].

5. Conclusion

Seasonal Autoregressive Integrated Moving Average Model can better eliminate the disturbances of time series trends, seasons and other related factors, The SARIMA (4,1,0)(0,1,1)[12] model fitted well and it can achieve satisfactory results in forecasting the trend of tuberculosis incidence analysis. The number of cases of tuberculosis is not only seasonally related, but is also related to geographic location, income level of the population, population movement, and relevant policies.So the model will have relatively large errors between fitted and actual values in some months. For example, there has been a significant decrease in the number of tuberculosis cases since 2009, which may be attributed to the implementation of the national policy on tuberculosis prevention and control. In the future, external factors such as climatic conditions, public health policies, population mobility, and changes in the disease itself could be included based on external variables related to TB incidence. These external variables could be added to the existing SARIMA forecasting model to increase the accuracy of the model's predictions.

References

[1] *Lu Chunrong, Fange Hongxia, Lu Puxuan, et al. (2021). WHO Global Tuberculosis Report 2021: analysis of global and Chinese key data[J]. Electronic Journal of Emerging Infectious Diseases, 6(4):368-372.*

[2] *World Health Organization. (2023). Global tuberculosis report 2022.Geneva. Retrieved from www.who.org/.*

[3] *Xun Mengjun, Li Jinlan, Huang Aiju, et al. (2023). Application of ARIMA model and Holt-Winters exponential smoothing method in predicting the incidence of tuberculosis in Guizhou province. Chinese Journal of Preventive Medicine, 24(7):678-682.*

[4] *Jiang Jianguo, Sun Dingyong, Zhang Yanqiu, et al. (2023). Forecasting and analysing the epidemic trend of tuberculosis in Henan Province by applying ARIMA model. Modern Disease Prevention and Control, 34(7):495-499,563.*

[5] *Su Yanping, Sun Xiaowei, Gao Hanqing, et al. (2023). Evaluation of the effect of exponential smoothing method model and ARIMA model in the prediction of tuberculosis epidemic trend in Tongzhou District, Beijing. Medical Animal Defence, 39(1):8-12.*

DOI: 10.54254/2754-1169/86/20240829

[6] *Ren Jiahao, Xu Jie, Yang Haiyan. (2022). Application of ARIMA and Holt-Winters exponential smoothing model in the prediction of tuberculosis epidemic trend in Henan Province. Journal of Zhengzhou University-Medical Edition,57(6):756-760.*

[7] *Nie Yanwu, Yang Zhen, Sun Yahong, et al. (2022). Application of SARIMA-SVR combined model in the prediction of tuberculosis epidemic trend in Shanghai. Medical Animal Defence, 38(9):817-821.*

[8] *Chen T. (2015). Comparative study of ARIMA model and BP neural network model in the application of HIV incidence prediction. Guilin: Guangxi Medical University.*

[9] *Zhou Meiyan, Huang Ying, Yan Yulong, et al. (2022). Application of exponential smoothing method and ARIMA model in the prediction of tuberculosis in students. Practical Preventive Medicine, 29(1):18-22.*

[10] *Yang Meitao, Wang Yanding, Li Zhiqiang, et al. (2023). Application of ARIMA-SVM combined model in the prediction of tuberculosis incidence trend. Modern Preventive Medicine, 50(11):1921-1926.*