Mathematical Biosciences and Engineering

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

MBE, 17(4): 4317–4327.
DOI: 10.3934/mbe.2020238
Received: 12 April 2020
Accepted: 12 June 2020
Published: 19 June 2020

Research article

The impact of air pollution on the transmission of pulmonary tuberculosis

Zuqin Ding[1,†], Yaxiao Li[1,2,†], Xiaomeng Wang[1,†], Huling Li[3,†], Yongli Cai[1], Bingxian Wang[1], Kai Wang[3,*] and Weiming Wang[1,*]

[1] School of Mathematics and Statistics, Huaiyin Normal University, Huaian 223300, China

[2] College of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, China

[3] Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830011, China

[†] These authors contributed equally.

[*] **Correspondence:** Email: wangkaimath@sina.com (K. Wang); weimingwang2003@163.com (W. Wang).

Abstract: In this paper, we investigate the relationship between the air pollution and tuberculosis cases and its prediction in Jiangsu, China by using the time-series analysis method, and find that the seasonal ARIMA$(1,1,0)\times(0,1,1)_{12}$ model is the preferred model for predicting the TB cases in Jiangsu, China. Furthermore, we evaluate the relationship between AQI, PM2.5, PM10 and the number of TB cases, and find that the prediction accuracy of the ARIMA model is improved by adding monthly PM2.5 with 0-month lag as an external variable, i.e., ARIMA$(1,1,0)\times(0,1,1)_{12}$+PM2.5. The results show that ARIMAX model can be a useful tool for predicting TB cases in Jiangsu, China, and it can provide a scientific basis for the prevention and treatment of TB.

Keywords: tuberculosis; air pollution; time series analysis; ARIMA model; prediction

1. Introduction

The tuberculosis (TB) is caused by the bacterium *Mycobacterium* tuberculosis. The bacteria usually attack the lungs, but TB bacteria can attack any part of the body such as the kidney, spine, and brain [1]. TB is contagious, which means the bacteria easily spread from an infected person to someone else. One can get TB by breathing in air droplets from a cough or sneeze of an infected person. Once infected, the individual is at the highest risk of developing TB disease within the first two years, and there is still no vaccine able to prevent pulmonary TB, the most common form of the disease [2–4]. Pulmonary TB is typical and infectious, which is responsible for 1.5 million deaths each year. Not all infected with

TB bacteria becomes sick. As a result, two TB-related conditions exist: latent TB infection (LTBI) and TB disease [1].

In mainland of China, although the governments do their best to control TB, China has the second highest number of TB cases in the world. According to the global TB report in 2018, the number of TB in China was 823,342, the death number of TB was 3149 [1]. Particularly, pulmonary TB in Jiangsu province, China, showed a slow oscillatory trend. In 2015, reported cases were 36,039 and deaths 91; 2016, cases 36,647, deaths 93; 2017, cases 28,402, deaths 97; 2018, cases 33,566, deaths 80; and 2019, cases 32,880, deaths 90 [5]. Therefore, TB is still a major infectious disease that needs to be controlled whether in China or in Jiangsu province, China.

In order to estimate the relationship between variables described disease dynamics, a classical statistical approach is the use of time series analysis of the incident cases. For example, Ekpenyong et al. [6] established an ARMA$(1, 0, 1) \times (1, 1, 2)_{12}$ model to analyze and predict the monthly TB cases in University of Calabar Teaching Hospital based on data from 1990–2015. Moosazadeh et al. [7] used monthly TB incidence data recorded in the Iranian National Tuberculosis Control Program for time series analysis and selected SARIMA$(0, 1, 1) \times (0, 1, 1)_{12}$ as the most adequate model for prediction. Li et al. [8] used hybrid ARIMA-EGARCH model to analyze the visceral leishmaniasis data in Kashgar, China and the visceral leishmaniasis cases were simulated by ARIMA$(2, 1, 2)(1, 1, 1)_{12}$-EGARCH$(1, 1)$ model, and found that the root-mean-square error was 7.23% in the validation phase, which offered a scientific basis to control visceral leishmaniasis spread in Kashgar prefecture of Xinjiang, China.

On the other hand, in China, there is experiencing challenge of public health caused by air pollution [9]. And numerous epidemiological studies showed that air pollution associate with risk of various disease [10–15]. Recently, Peng et al. [16] found that long-term exposure to PM2.5 increases the risk of death among TB patients, and claimed that the control of ambient air pollution may help decreasing the mortality of TB. Liu et al. [17] showed that the short-term exposure of PM10 and PM2.5 could significantly increase the risk of death of residents, and the increase of PM10 and PM2.5 in short-term was significantly correlated with the total mortality, cardiovascular death and respiratory death. Tang et al. [9] and He et al. [18, 19] focused on how air pollution affects the dynamics of epidemic models, and found that only taking sustained, long-term and high-intensity emission reduction measures can effectively reduce the air quality index and the number of respiratory cases.

Based on the discussions above, in this study, we apply time-series approach to analyze the impact of air quantity on the spreading of TB in Jiangsu province, China during the years 2015 to 2019, and predict the trend of TB epidemic in 2020.

2. Materials and methods

2.1. Data collection

The required information was collected as two parts: Information regarding TB cases: The number of TB cases was collected from the Jiangsu Disease Control Center from January 2015 to December 2019 (60 months) [5] (see Figure 1). Information regarding air pollution including air quality index (AQI), particulate matter $< 2.5\mu m$ in diameter (PM2.5) and particulate matter $< 10\mu m$ in diameter (PM10), was gathered from the National Meteorological Information Center [20] (See Figure 2 for details).

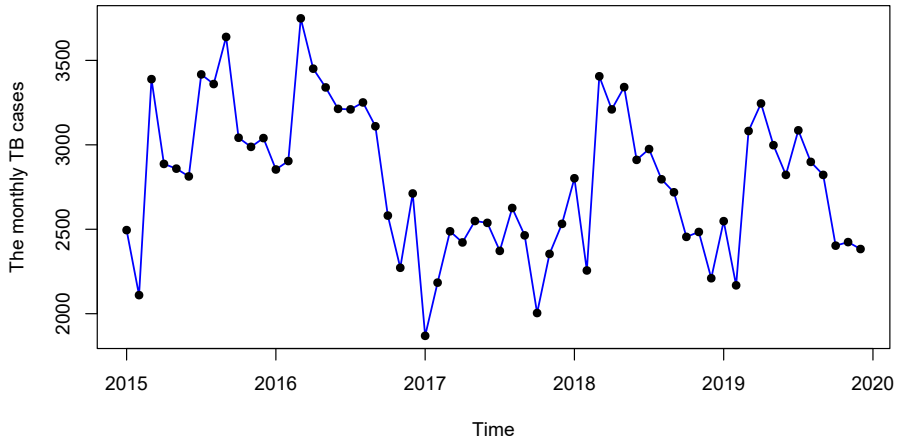


Figure 1. Time series of the monthly reported TB cases from January 2015 to December 2019.

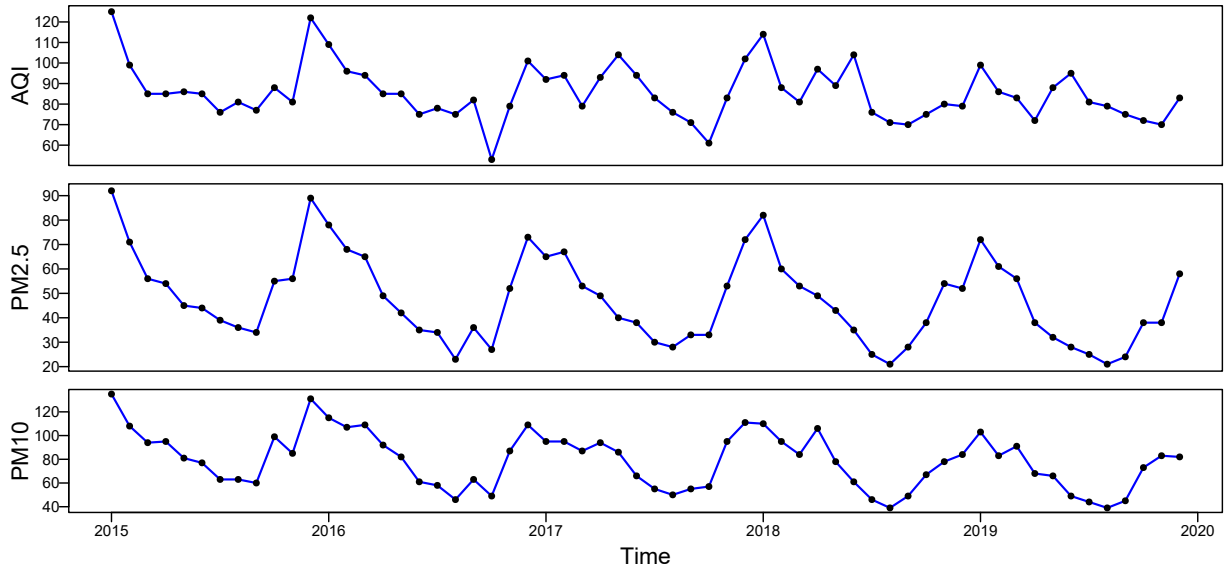


Figure 2. Time series of the monthly AQI, PM2.5 and PM10 in Jiangsu, China from January 2015 to December 2019.

2.2. Time series analysis

The ARIMA model is one of the most important and basic models. According to whether the model contains seasonal components, it can be divided into continuous ARIMA(p, d, q) model, seasonal ARIMA$(P, D, Q)_S$ model and product seasonal ARIMA$(p, d, q) \times (P, D, Q)_S$ model. p, d, q and P, D, Q are the order values of continuous and seasonal autoregressive (AR), difference (I), and moving average (MA), respectively. s represents a seasonal period. Briefly, the ARIMA univariate analysis models consist of 3 sub-processes: model identification, parameter estimation and model diagnosis. By repeating these three steps, the optimal prediction model is screened out [21].

The following steps are used to fit the model:

- Firstly, the stationarity of the original sequence is tested by using the disease sequence diagram and Augmented Dickey-Fuller (ADF) test. If the sequence is non-stationary, in order to eliminate the trend and seasonality of the sequence, the first-order ordinary difference ($d = 1$) and the first-order seasonal difference ($D = 1$) are applied to make it stable. We further analyze the stationary series.
- Secondly, we examine the autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs to identify the parameters in the model, p and q, respectively. Then, the maximum likelihood estimation (MLE) method is used to estimate the parameters in the model. In order to evaluate the suitability of the established ARIMA model, the parameters and residual of the model are tested respectively, and Ljung-Box (Q) test is applied to check whether the residual of the model is white noise.
- Finally, if several models satisfy the condition that the parameters are significant and the residual sequence of the model is white noise, then the optimal univariate model can be selected by using Akaike information criterion (AIC), Schwarz Bayesian information criterion (SBC) and mean absolute percentage error (MAPE) and root mean square error (RMSE) indicators of the model.

In order to establish the optimal multivariate model, we consider the air quality variables as regression variables into the model to test whether they can improve the prediction performance of ARIMA model. The cross-correlation function (CCF) analysis is performed on the number of TB cases and climate data to find the best predictor and its optimal lag period to be included in the final model. In order to eliminate the trending and seasonal characteristics of each meteorological variable sequence, we differentially process each meteorological variable to achieve stability. Next, we perform a pre-whitening process to establish an optimal ARIMA model for each individual meteorological variable, which is used as a filter to filter the input and the output sequence. And the cross-correlation coefficient of the filtered output and the input sequence is calculated by the CCF, so the pre-whitening process is completed. By means of the Cross-Correlation diagram to judge the hysteresis relationship between the input and the output sequence, the climatic variables ($P < 0.05$) which are significantly correlated with the number of TB cases are included in the multivariate ARIMA model.

Simply speaking, the ARIMA model with input variables is called a dynamic regression model, abbreviated as ARIMAX. The optimal selection criteria for the ARIMAX model are still AIC and MAPE. The MAPE is calculated for training and validation data to assess the predictive validity of the models. Smaller the values of this metric indication, the better the predictive performance. The MAPE equals to:

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

where X_t is the actual value and $\hat{X}_t$ the forecast value. The difference between X_t and $\hat{X}_t$ is divided by the actual value X_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying by 100% makes it a percentage error.

All data are analyzed by using packages `tseries, fUnitRoots, zoo, forecast` and `TSA` of the software R (version 3.6.3). The ARIMA and ARIMAX models are constructed by using the processed data.

3. Results

3.1. ARIMA model

From January 2015 to December 2019, there were 167,534 TB cases in Jiangsu, China. The annual TB cases were 36,039, 36,647, 28,402, 33,566, 32,880, respectively [5]. Figure 3 shows that the number of TB cases has seasonal fluctuations with an annual cycle. During the period of March, the seasonal index (or called season exponent, which is reflects a stable relationship between the average number of newly TB cases and the average number of total newly TB) is the largest.

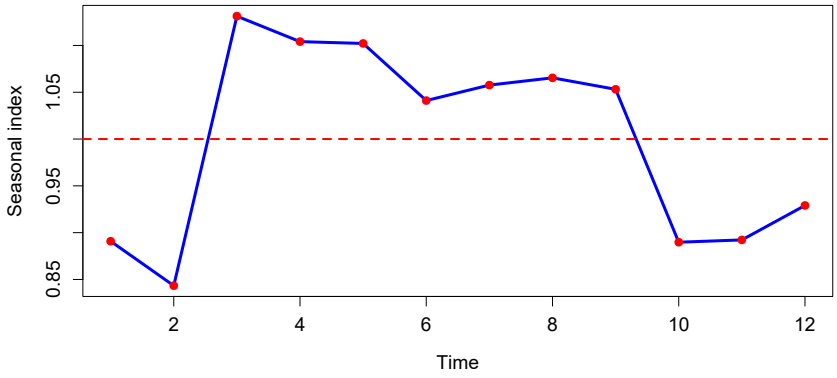


Figure 3. Season index of TB cases from January 2015 to December 2019.

Therefore, the multiplicative seasonal ARIMA model is used. The ADF test of the original sequence shows $P = 0.523 > 5\%$, indicating that the sequence of TB cases is not a stationary sequence. To eliminate the trend and seasonality, the original sequence is pre-processed using first-order ordinary difference ($d = 1$) and first-order seasonal difference ($D = 1$), then we can obtain a stationary sequence (ADF Test, $P = 0.01$) (see Figure 4).

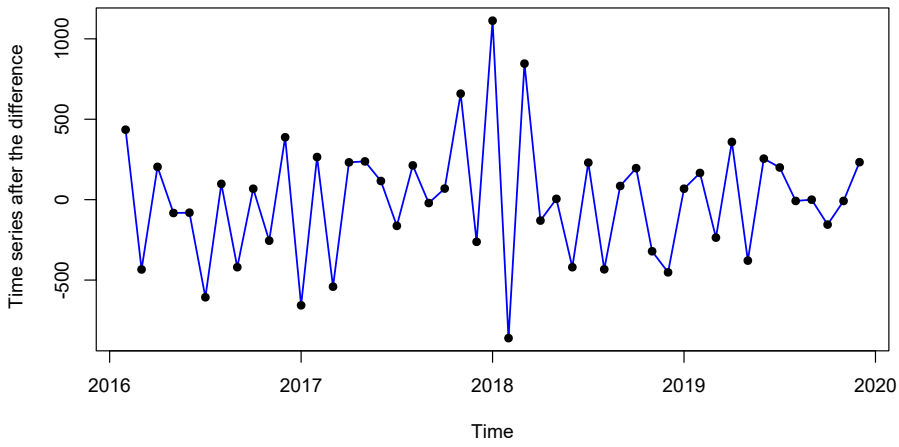


Figure 4. Time series of stationary sequence obtained from first-order ordinary difference ($d = 1$) and first-order seasonal difference ($D = 1$).

Next, autocorrelation function (ACF) and partial autocorrelation function (PACF) analyses are performed. ACF shows $q = 0$, 1 or 2, and PACF shows $p = 0$ or 1. Considering the seasonal autocorrelation, since the data are collected monthly, S is equal to 12 (See Figure 5).

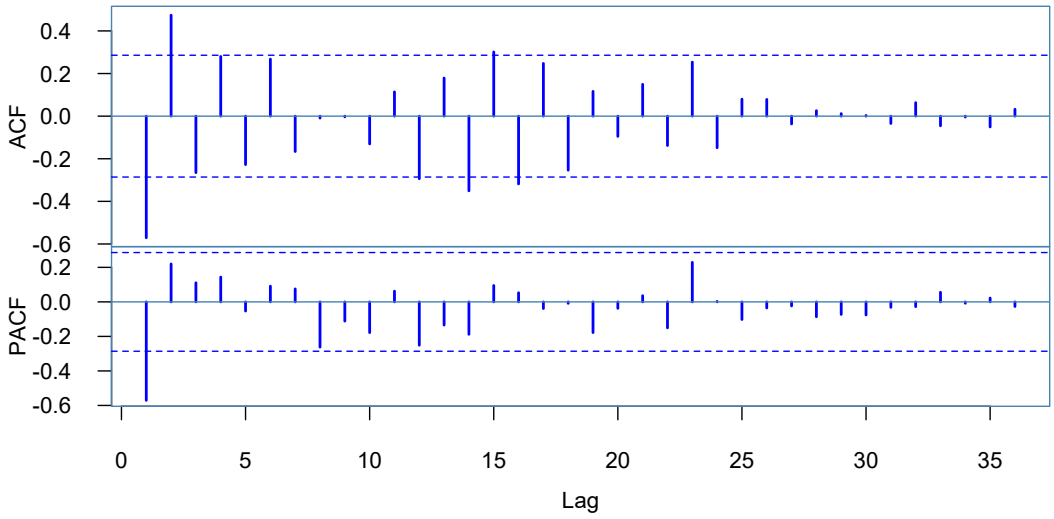


Figure 5. ACF and PACF figures after differencing.

According to the criterion of minimum information, the model ARIMA$(1, 1, 0) \times (0, 1, 1)_{12}$ has the minimum value of AIC= 670.203, AICc= 670.761, SBC =675.753 in two candidate models which are the optimal model (see Table 1).

The parameter estimation results of the model and white noise test results are shown in Table 2 and Table 3. All parameters in the ARIMA$(1, 1, 0) \times (0, 1, 1)_{12}$ model are statistically significant.

Table 1. Goodness of fits for plausible ARIMA models.

Model	AIC	AICc	SBC
ARIMA$(1, 1, 0) \times (0, 1, 1)_{12}$	670.203	670.761	675.753
ARIMA$(1, 1, 0) \times (1, 1, 0)_{12}$	673.439	673.997	678.989

Table 2. Parameters estimation for ARIMA$(1, 1, 0) \times (0, 1, 1)_{12}$.

Parameter	Coefficient	Standard error	T-value	P-value
AR(1)	-0.556	0.121	-4.584	<0.001
SMA(1)	-0.638	0.283	-2.255	0.014

Table 3. White noise test results of residual sequences.

Model	Lag	χ^2	DF	P-value
ARIMA$(1, 1, 0) \times (0, 1, 1)_{12}$	6	5.176	5	0.395
	12	11.634	11	0.392
	18	18.150	17	0.379

Finally, ARIMA$(1,1,0)\times(0,1,1)_{12}$ model is employed for fitting TB cases from January 2015 to December 2019. The fitting and forecasting results are shown in Figure 6. And in the case of predicting in 2020, MAPE= 6.243% < 10%. Almost the statistic data are located in the confidence interval of 95%. Hence we can use ARIMA$(1,1,0)\times(0,1,1)_{12}$ model to predict new TB cases in Jiangsu in short term.

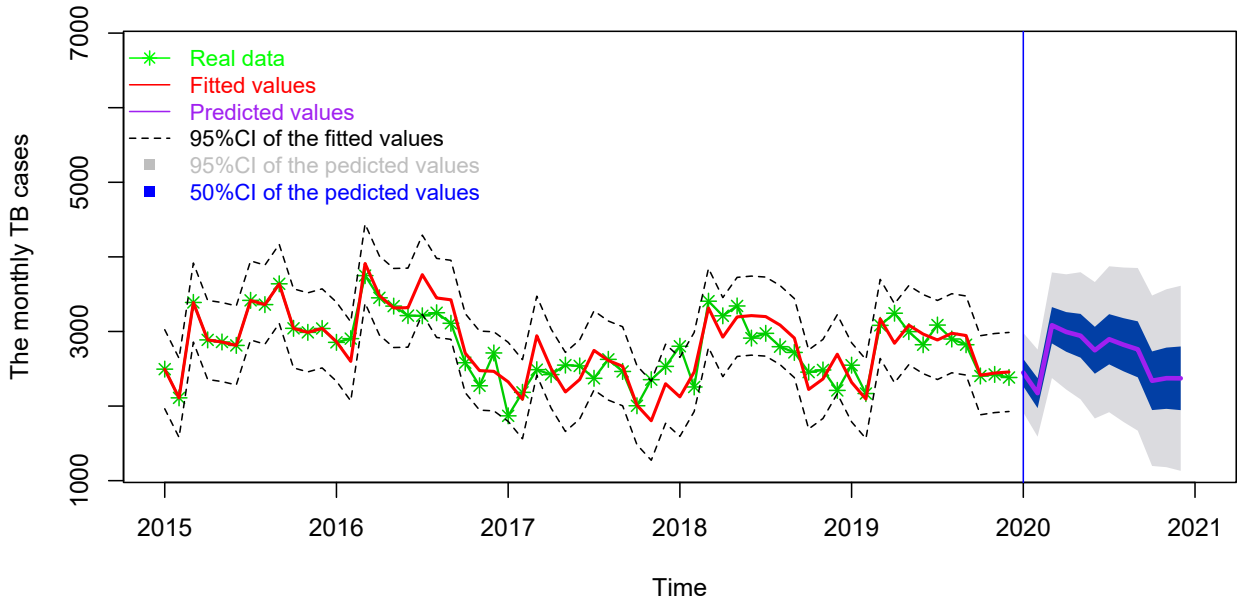


Figure 6. The prediction results of ARIMA$(1,1,0)\times(0,1,1)_{12}$ model.

3.2. ARIMAX model

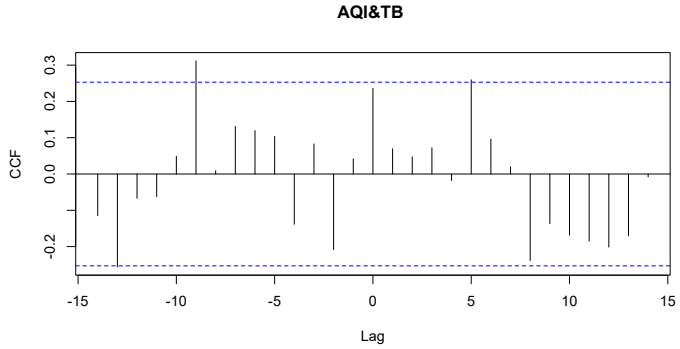
In order to find the best multivariate model, we consider air pollution variables (AQI, PM2.5 and PM10) as regression variables into the model, namely, ARIMAX model. By calculating the cross-correlation (CCF) between TB cases and air quality series, the best predictor and its lag order are found and finally incorporated into the model. Table 4 lists the optimal models for each meteorological sequence in pre-whitening. The *P*-values of the residual sequences are significantly greater than 0.05. Parameters are significantly passed. Significant air quality variables independently associated with TB cases by cross-correlations are shown in Figure 7.

Table 4. Comparisons of ARIMA models.

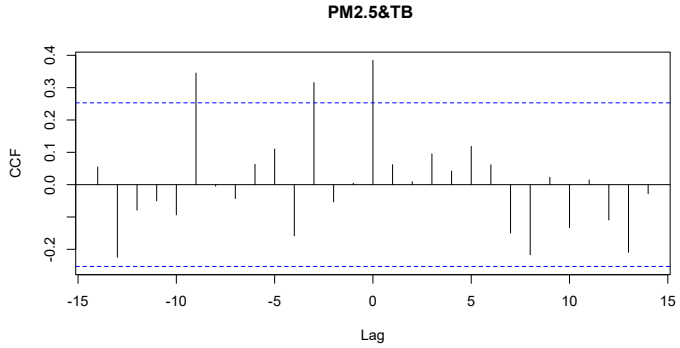
Air quality	Optimization model	AIC	AICs	SBC
AQI	ARIMA$(0,1,1)\times(0,1,1)_{12}$	368.496	369.054	374.047
PM2.5	ARIMA$(0,1,1)\times(1,1,0)_{12}$	330.791	331.349	336.342
PM10	ARIMA$(0,1,1)\times(1,1,0)_{12}$	379.408	379.966	384.958

From Figure 7, we can know that the monthly average AQI at a lag of 5 months, PM2.5 at a lag of 0 month and PM10 at a lag of 0 month or 13 months are significantly related to the number of TB cases, which can be included in the multivariate ARIMAX model. And hence, there are four ARIMAX

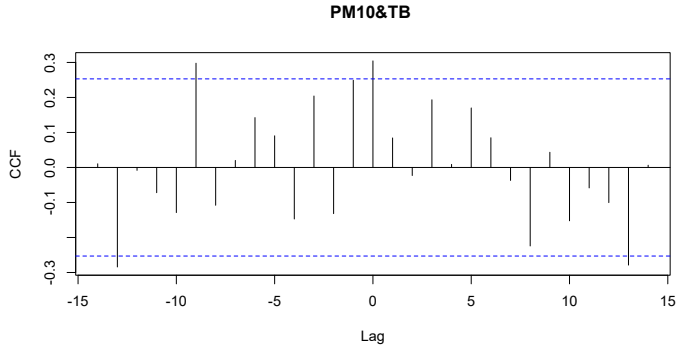
models, but only two of which pass the parameter test, i.e., ARIMA$(1,1,0)\times(0,1,1)_{12}$+PM2.5 and ARIMA$(1,1,0)\times(0,1,1)_{12}$+PM10 (see Table 5). And ARIMA$(1,1,0)\times(0,1,1)_{12}$+PM2.5 with 0-month lag has the smallest AIC = 664.066 and MAPE = 5.891%, which is the best model. And the numerical prediction results of ARIMA$(1,1,0)\times(0,1,1)_{12}$+PM2.5 model are shown in Figure 8.



(a)



(b)



(c)

Figure 7. Cross-correlations between the pre-whitened TB cases and air quality factors from 2015 to 2019.

Table 5. Parameters estimation for ARIMA$(1,1,0)\times(0,1,1)_{12}$ with different air qualities.

Air quanlity	Lag	Coefficient	Standard error	T-value	P-value	AIC	MAPE
PM2.5	0	12.223	4.760	2.568	0.006	664.066	5.891%
PM10	0	6.193	3.120	1.985	0.026	666.414	5.957%

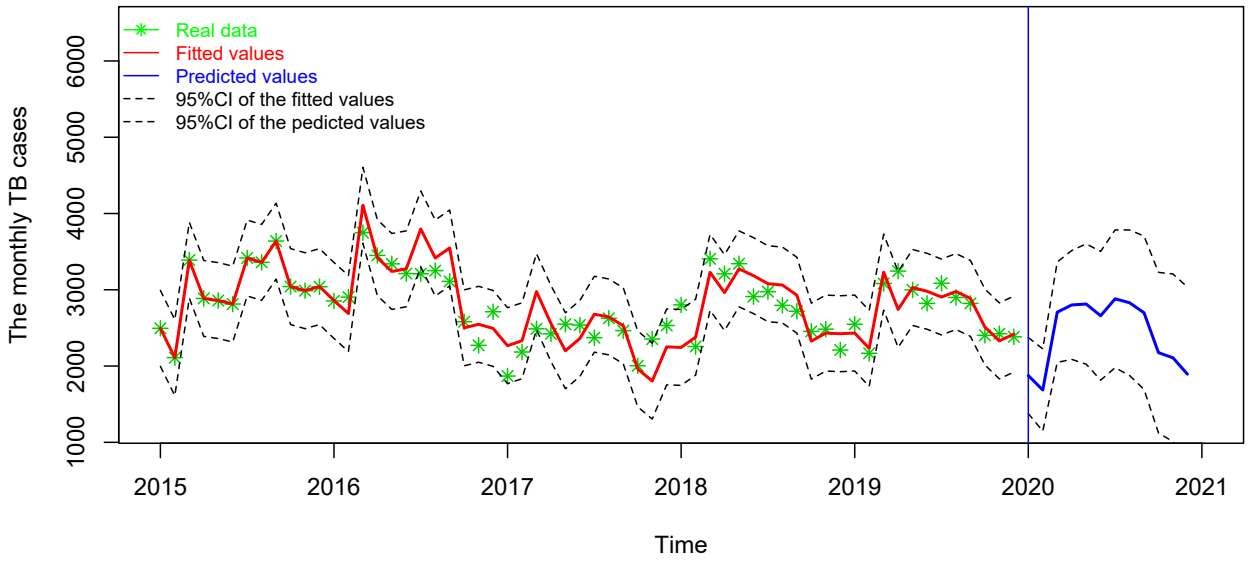


Figure 8. The prediction results of ARIMA$(1,1,0)\times(0,1,1)_{12}$+PM2.5 model.

4. Conclusions and discussions

TB is a chronic infectious disease that seriously endangers people's health. In the present paper, based on the reported TB cases, we establish ARIMA and ARIMAX models to study the trend of TB epidemic in Jiangsu, China by using the method of time series analysis.

Although the research of time series data has developed rapidly in recent years, most of it focuses on one-dimensional time series, and there are few studies on multi-dimensional time series. Our ARIMAX model of multiple time series of air quantity and the reported TB cases is a useful attempt.

It is worthy to note that, in [16], the authors found that long-term exposure to PM2.5 increases the risk of death from TB and other diseases among TB patients. And in the present paper, we investigate the impact of AQI, PM2.5 and PM10 on the spreading of TB in Jiangsu, China, and find that long-term exposure to PM2.5 is closed related to the spreading of TB. More precisely, when the monthly PM2.5 with 0-month lag is introduced into the ARIMA model, the results reveal that ARIMA$(1,1,0)\times(0,1,1)_{12}$+PM2.5 with 0-month lag model can improve the predictive performance of the ARIMA model. These results can be seen as supplements of the results in [16], and may provide a scientific basis for the prevention and control of TB.

Acknowledgments

The authors would like to thank the anonymous referees for very helpful suggestions and comments which led to improvement of our original manuscript. This research was supported by the National Natural Science Foundation of China (Grant No. 61672013, 11601179, 61772017 and 11961071), Research project of philosophy and Social Sciences in Jiangsu Province, China (2019SJA1671), Huaian Key Laboratory for Infectious Diseases Control and Prevention, China (HAP201704), and Innovation and Entrepreneurship Training Program for College Students in Jiangsu Province (201710323027Y and 201810323013Z), National Innovation Training Program for College Students, China (201810323007).

Conflict of interest

The authors declare that they have no competing interests.

References

1. WHO, Global tuberculosis report 2018. Geneva: World health organization, 2018. Available from: http://www.who.int/tb/publications/global_report/en/.
2. S. Basu, J. R. Andrews, E. M. Poolman, N. R. Gandhi, N. S. Shah, A. Moll, et al., Prevention of nosocomial transmission of extensively drug-resistant tuberculosis in rural south african district hospitals: an epidemiological modelling study, *Lancet*, **370** (2007), 1500–1507.
3. S. D. Lawn, A. I. Zumla, Tuberculosis, *Lancet*, **378** (2011), 57–72.
4. B. I. Restrepo, Convergence of the tuberculosis and diabetes epidemics: renewal of old acquaintances, *Clin. Infect. Dis.*, **45** (2007), 436–438.
5. The reported tuberculosis cases in jiangsu province, 2018. Available from: http://www.jshealth.com/.
6. B. O. Ekpenyong, ARMA type modeling of certain non-stationary time series in calabar, *Am. J. Appl. Math. Stat.*, **4** (2016), 118–125.
7. M. Moosazadeh, N. Khanjani, M. Nasehi, A. Bahrampour, Predicting the incidence of smear positive tuberculosis cases in iran using time series analysis, *Iran. J. Publ. Health*, **44** (2015), 1526–1534.
8. H. Li, R. Zheng, Q. Zheng, W. Jiang, X. Zhang, W. M. Wang, et al., Predicting the number of visceral leishmaniasis cases in Kashgar, Xinjiang, China using the ARIMA-EGARCH model, *Asian Pac. J. Trop. Med.*, **13** (2020), 81–90.
9. S. Tang, Q. Yan, W. Shi, X. Wang, X. Sun, P. Yu, et al., Measuring the impact of air pollution on respiratory infection risk in china, *Environ. Pollut.*, **232** (2018), 1–10.
10. Y. Alyousifi, N. Masseran, K. Ibrahim, Modeling the stochastic dependence of air pollution index data, *Stoch. Env. Res. Risk.*, **26** (2018), 1603–1611.
11. S. Chauhan, S. Bhatia, S. Gupta, Effect of pollution on dynamics of sir model with treatment, *Int. J. Biomath.*, **8** (2015), 1550083.

12. M. Laeremans, E. Dons, I. Avila-Palencia, G. Carrasco-Turigas, Short-term effects of physical activity, air pollution and their interaction on the cardiovascular and respiratory system, *Environ. Int.*, **117** (2018), 82–90.
13. P. M. Mannucci, Airborne pollution and cardiovascular disease: burden and causes of an epidemic, *Eur. Heart. J.*, **34** (2013), 1251–1253.
14. G. Polezer, Y. Tadano, H. Siqueira, A. Godoi, C. Yamamoto, Assessing the impact of pm 2.5 on respiratory disease using artificial neural networks, *Environ. Pollut.*, **235** (2018), 394–403.
15. C. Sun, Y. Xiang, Y. Xin, Social acceptance towards the air pollution in china: Evidence from public's willingness to pay for smog mitigation, *Energy Policy*, **92** (2016), 313–324.
16. Z. Peng, C. Liu, B. Xu, H. Kan, W. Wang, Long-term exposure to ambient air pollution and mortality in a chinese tuberculosis cohort, *Sci. Total Environ.*, **580** (2017), 1483–1488.
17. C. Liu, R. Chen, F. Sera, A. M. Vicedo-Cabrera, Y. Guo, S. Tong, et al., Ambient particulate air pollution and daily mortality in 652 cities, *N. Engl. J. Med.*, **381** (2019), 705–715.
18. S. He, S. Tang, W. M. Wang, A stochastic SIS model driven by random diffusion of air pollutants, *Phys. A*, **532** (2019), 121759.
19. S. He, S. Tang, Y. Xiao, R. A. Cheke, Stochastic modelling of air pollution impacts on respiratory infection risk, *Bull. Math. Biol.*, **80** (2018), 3127–3153.
20. National meteorological information center. Available from: http://data.cma.cn/.
21. S. Chadsuthi, C. Modchang, Y. Lenbury, S. Iamsirithaworn, W. Triampo, Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses, *Asian Pac. J. Trop. Med.*, **5** (2012), 539–546.

AIMS Press