Time Series Analysis on Incidence of Pulmonary Tuberculosis With Weather Factors During 2004-2017 in Guangdong Province, China

Tianyu Qin
Beijing University of Chinese Medicine https://orcid.org/0000-0001-7600-0566
Yu Hao
Beijing University of Chinese Medicine
Juan He (hejuan6428@sina.com)
Beijing University of Chinese Medicine

Research Article

Keywords: tuberculosis, weather factors, time series, Guangdong

Posted Date: February 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-181418/v1

Abstract

Background: Although the occurrence of some infectious diseases including TB was found to be associated with specific weather factors, few studies have incorporated weather factors into the model to predict the incidence of tuberculosis (TB). We aimed to establish an accurate forecasting model using TB data in Guangdong Province, incorporating local weather factors.

Methods: Data of sixteen meteorological variables (2003-2016) and the TB incidence data (2004-2016) of Guangdong were collected. Seasonal autoregressive integrated moving average (SARIMA) model was constructed based on the data. SARIMA model with weather factors as explanatory variables (SARIMAX) was performed to fit and predict TB incidence in 2017.

Results: Maximum temperature, maximum daily rainfall, minimum relative humidity, mean vapor pressure, extreme wind speed, maximum atmospheric pressure, mean atmospheric pressure and illumination duration were significantly associated with log(TB incidence). After fitting the SARIMAX model, maximum pressure at lag 6 (β= -0.007, $P < 0.05$, 95% confidence interval (CI): -0.011, -0.002, mean square error (MSE): 0.279) was negatively associated with log(TB incidence), while extreme wind speed at lag 5 (β=0.009, $P < 0.05$, 95% CI: 0.005, 0.013, MSE: 0.143) was positively associated. SARIMAX (1, 1, 1) $(0, 1, 1)_{12}$ with extreme wind speed at lag 5 was the best predictive model with lower Akaike information criterion (AIC) and MSE. The predicted monthly TB incidence all fall within the confidence intervals using this model.

Conclusions: Weather factors have different effects on TB incidence in Guangdong. Incorporating meteorological factors into the model increased the accuracy of prediction.

Background

As a chronic infectious disease caused by Mycobacterium tuberculosis, Tuberculosis (TB) is mainly transmitted by respiratory tract and threatens human health, ranking first among the top ten death causes in the world. The World Health Organization (WHO) estimated that ranging 8.9-11.0 million people fell ill with TB around the world in 2019 [1]. To end the world's top infectious disease killer, the WHO set a goal of reducing the morbidity and mortality of TB by 90% and 95%, respectively, between 2015 and 2035. However the incidence has just been declining very slowly in the past few years. In China, the incidence and death of TB always rank first among class A and B level infectious diseases. In recent years, due to effective prevention and control, the epidemic situation of TB in China has been declining at an annual rate of 3% between 2005 and 2017 [2]. Meanwhile, the infectivity of TB has become stronger and stronger since 2010 [3]. Since the population base of TB infection is large, China is still one of the 22 countries with high TB burden in the world, with approximately 866 000 new cases identified in 2018, second only to India [1].

Chemotherapy is the main treatment of TB while there is no major breakthrough in new medication for treating TB in the past 40 years. The chemotherapy of TB is time consuming, and the chemotherapy regimen is complex. The only vaccine in use is Bacillus Calmette Guerin (BCG), which was developed almost 100 years ago and is not applicable to adults. At the same time, the double infection of Mycobacterium tuberculosis and HIV, and the emergence of drug-resistant TB cause the current TB prevention and treatment in a dilemma [4]. In addition to endeavoring to care and treat after infection, targeted publicity, prevention and control measures or policies help to cut off its spread from the source, reduce the incidence of TB effectively, and further alleviate the burden of

individuals and burden of subsequent treatment. Based on this, accurately predicting the trend of this epidemic is helpful to foresee the possible peaks and provide a reference for the prevention and control of TB.

Climate change has been a contributing factor in the transmission of various of infectious diseases, especially water-borne and vector-borne infectious diseases [5]. In some recent studies the incidence of some infectious diseases including TB was associated with specific weather factors. In one study, the relationship between the spread of SARS-Cov-2 and meteorological variables was analyzed. The result showed that the average temperature was negatively correlated with the number of SARS-Cov-2 infection cases. The precipitation was positively correlated with the spread of SARS-Cov-2. Countries with higher rainfall showed an increase in disease transmission [6]. In another study, distributed lag nonlinear model (DLNM) was used to analyze TB incidence in Jinghong, a city in Yunnan Province, and found that the average temperature was negatively correlated with the incidence of TB, with lag period 2 months; the total precipitation and the lowest relative humidity were negatively correlated, with lag period 3 months and 4 months respectively [7]. In another study, geographically weighted regression (GWR) model was used to analyze the relationship between the incidence of TB in 2005-2015 in different districts of China and local meteorological factors. It was found that the average temperature was positively correlated with the incidence of TB, while the mean relative humidity and the mean wind speed were negatively correlated [8].

Previous studies have also explored various models, such as autoregressive integrated moving average (ARIMA) [2, 9], X12-ARIMA [9], ARIMA-generalized regression neural network (GRNN), DLMN [7] and GWR [8] model in predicting TB [2], but most were conducted without incorporating meteorological factors. In a study performed on three cities in Jiangsu province, the ARIMAX model was found to be superior to the ARIMA and RNN models in predicting PTB when taking weather factors into consideration [10]. The result of a study on the dengue in Guangzhou showed that an ARIMA model with imported cases and minimum temperature as input variables was superior to a single ARIMA model in forecasting dengue transmission [11]. Guangdong is the province with the largest population and high TB mobility in China, in which the incidence of TB ranks near the top in 31 provinces in the country. But to our knowledge, there hasn't been any study exploring the TB incidence pattern of this province. In the current study, we added weather factors into the SARIMAX model, explored correlations between weather factors and the incidence of TB in Guangdong, and endeavored to establish an accurate model for estimating epidemic trends pertaining to TB.

Methods

Study area

Guangdong Province is located in the coast of South China (20°13'-25°31'N, 109°39'-117°19'E), with an area of 179725 square kilometer (Fig 1). As the most populous and largest economic province province in China at present, Guangdong governed 21 prefectures and had a permanent population of 115.21 million at the end of 2019. It is situated in a zone of subtropical monsoon climate with abundant rainfall. Spring is warm, autumn and winter are relatively sunny, while summer is hot and rainy.

Data collection

Data of TB

In mainland China, the TB Information Management System has been established and operated by CCDC (Chinese Centre for Disease Control and Prevention) since 2004. It is mandatory to report every single TB case through this on-line system. Monthly TB data from January 2004 to December 2017 were obtained from the Chinese public health science data centre. The TB data used used in this study was TB monthly incidence, including TB cases that was sputum smear positive, culture positive, bacteria negative and without sputum examination. The TB cases were diagnosed according to the criteria used by the National Health and Family Planning Commission of the People's Republic of China.

Data of weather factors

Monthly local weather factors in 2003 to 2017 were obtained from the China Meterorological Science Data Center. Sixteen meteorological characteristics (minimum relative humidity, minimum atmospheric pressure, maximum atmospheric pressure, mean atmospheric pressure, mean vapor pressure, maximum wind speed, extreme wind speed, mean 2-minute wind speed, mean temperature, minimum temperature, maximum temperature, mean minimum temperature, mean maximum temperature, illumination duration, 8pm-8pm rainfall, maximum daily rainfall) were included. Means of monthly values of these meteorological characteristics were calculated from 6 meteorology station (Station number: 59082, station name: Shaoguan; station number: 59287, station name: Guangzhou; station number: 59293, station name: Dongyuan; station number: 59316, station name: Shantou; station number: 59501, station name: Shanwei; station number: 59663, station name: Yangjiang). Data in the same month and the same station during other year were averaged to fill the missing data. The used data in this study are presented in Additional Files.

Construction of SARIMA Model

The well-known ARIMA model is widely used in time series analysis for describing and predicting epidemic prevalence for its accuracy and practicality [12-14]. The seasonal ARIMA (SARIMA) model that is developed from the ARIMA model, incorporates seasonal period and performs better in the presence of an obvious seasonal pattern. Hence SARIMA model is optimal in this study because both seasonal and non-seasonal trends were observed. A SARIMA (p, d, q)(P, D, Q)$_s$ model has 7 parameters, in which the on-seasonal parameters includes autoregressive model order (p), number of differences (d), and moving average model order (q), and seasonal parameters includes seasonal autoregressive model order (P), number of seasonal differences (D), and seasonal moving average model order (Q). Also, the parameter s (s=12 in this study) indicates the length of the periodic pattern. An SARIMA model can be described as the following formula:

$$Z_t = \frac{\Theta(B)\theta(B)\varepsilon_t}{\Phi(B)\phi(B)(1-B)^d(1-B^s)^D}$$

In which Z_t representing the value of time series at time t, and ε_t a white noise series. B here refers to a backward shift operator (e.g. $BZ_t = Z_{t-1}$). $\phi(B) = 1 - \phi_1 B - \ldots \phi_p B^p$ and $\Phi(B) = 1 - \Phi_1 B^s - \ldots \Phi_p B^{Ps}$ denote the general and seasonal auto-regressive operators respectively. $\theta(B) = 1 - \theta_1 B - \theta_q B^q$ and $\Theta(B) = 1 - \Theta 1Bs - \Theta_Q B_{Qs}$ stand for the general and seasonal moving average operators respectively.

Based on the monthly number of PTBs during 2004-2016, we constructed an SARIMA model for Guangdong province. The steps of the model construction are described below.

The first step was stabilization processing of the sequence. The data were processed by ordinary difference and logarithmic transformation when the sequence was unstable or by seasonal difference when the sequence had seasonal distribution. The second step was model identification. To establish optimal SARIMA models, the values of parameters q and Q was initially identified by referring to the plots of the autocorrelation function (ACF) while the order of p and P was determined by referring to the plots of the partial autocorrelation function (PACF) of the stabilized series. The parameter q could not exceed the lag order at which the ACF cut off and the parameter p did not exceed the lag order at which the PACF cut off. Also, the upper limits of seasonal parameter Q and P were the number lag order (where the sample ACF and PACF values exceeded the critical values respectively) divided by 12. Third, the model parameters were estimated and verified. The maximum likelihood method were used to estimate the coefficients in the model and t-tests was then applied to test the significance of coefficients. Akaike Information Criterion (AIC) values were then used to measure the model fit (smaller AIC values indicated better model fit). Fourthly, model diagnostics were performed. The Ljung-Box Q test was conducted to ascertain the degree to which the residual series of the model was demonstrated to be white noise. Only *P* value more than 0.05 suggested that the residual series was white noise and that all its information was adequately extracted. Finally, the optimal SARIMA model was determined through the above steps.

Construction of SARIMAX Model and Prediction of TB incidence

SARIMAX extends the capability of the SARIMA model by integrating external factors, such as temperature, rainfall, and other meteorological factors, into the time series model. The SARIMAX model was built to evaluate the relationship between monthly TB incidence and meteorological factors, and then to forecast the TB incidence. An SARIMAX model can be described as follows:

$$Y_t = \sum_{i=m1}^{m2} \beta_i X_{t-i} + Z_t$$

In this model, Y_t represents the value of time series at time t and X_t is a covariant time series that, hopefully, could help explain or forecast Y_t. Z_t satisfies Eq (1) and β_i refers to the coefficient. The expressions m_1 and m_2 are the lower and upper limits of the lag respectively.

An SARIMAX model was constructed based on meteorological variables and the optimal SARIMA model built above. The steps of the model construction are described below. Firstly, since it was difficult to assess the dependence between the two processes with strongly autocorrelation, data of monthly TB incidence and monthly meteorological variables were prewhitened by the fitted optimal SARIMA model in the previous section to separate the linear associations from their autocorrelation for both incidence data and meteorological data. On this basis, cross-correlation function (CCF) plots were then used to evaluate the relationship between the TB incidence and weather factors at different lag times. Weather factors which were significantly cross-correlated with incidence at specific lags were possible covariant for the model. Secondly, we incorporated the covariant into the model and repeated steps three (parameters estimation and verification) and step four (model diagnostics) in the previous section to build the best SARIMAX model. Only the covariant which had significant parameter estimates and random residual series and lowered AIC value were selected. We used the 2004-2016 data to establish the best SARIMA model and SARIMAX model. We then used the 2017 data to validate the predictive effects of the SARIMAX model.

The package “TSA” in R 4.0.3 (https://www.r-project.org/) was used to construct the SARIMAX model. The significance level was set to be 0.05.

Results

Descriptive analysis

The annual averaged TB incidence between 2004 and 2017 of Guangdong was 86.85/100 000. As shown in Figure 2, the peak of incidence occurred in 2009 with annual incidence of 109.70/100 000, and the lowest incidence occurred in 2016 with annual incidence of 71.82/100 000 with a decreasing TB trend during the years included in this study. The seasonal variation is obvious in Figure 3. There was a peak of incidence in the spring (March to May) and a trough in winter (November to December). More descriptive statistics about the weather factors and TB incidence were shown in Table 1.

Table 1

Description of monthly average meteorological factors (2003-2016) and monthly incidence (2004-2016) of TB case in Guangdong

Variables	Min	Max	Mean±SD	Median
Minimum relative humidity(%)	88.33333333	55	75.83±6.62	77.75
Minimum atmospheric pressure(hPa)	1012.066667	972.1333333	997.67±6.95	998.5333333
Maximum atmospheric pressure(hPa)	1031.716667	1001.983333	1015.07±7.4	1015.233333
Mean atmospheric pressure(hPa)	1019.133333	995.55	1007.02±6.05	1007.575
Mean vapor pressure(hPa)	32.41333333	7.533333333	21.36±7.35	20.85
Maximum wind speed(m/s)	16.6	6.266666667	8.96±1.49	8.891666667
Extreme wind speed(m/s)	26.06666667	10.7	14.94±2.51	14.46666667
Mean 2-minute wind speed(m/s)	3.3	1.85	2.35±0.24	2.316666667
Mean temperature(°C)	30.05	9.916666667	22.1±5.62	23.24583333
Minimum temperature(°C)	24.41666667	1.033333333	14.34±7.34	15.575
Maximum temperature(°C)	38.5	19.91666667	31.45±4.07	32.26666667
Mean minimum temperature(°C)	26.91666667	7.283333333	19.16±5.71	20.13333333
Mean maximum temperature(°C)	34.26666667	13.96666667	26.32±5.46	27.05
Illumination duration (hours)	289.6666667	28.65	148.22±55.49	148.45
Rainfall 20-20h(mm)	816.4666667	0.433333333	154.92±142	122.2833333
Maximum daily rainfall(mm)	193.6833333	1.7	49.99±36.09	45.99166667
Monthly morbidity(per 100,000 people)	10.78	3.37	7.24±1.55	6.95
Annual morbidity(per 100,000 people)	109.7045402	71.82208916	86.85±13.78	80.69748196

SARIMA model analysis

A SARIMA model for TB incidence from 2004 to 2016 was first developed with 156 monthly data points without any covariant. Obviously, the TB incidence time series fluctuated within a large range (Figure 2). A logarithmic transformation of the time series of TB incidence was done to stabilize fluctuations in the time series data. As the time series plot of the logarithm of TB incidence showed a time-dependent trend and an obvious seasonal distribution, 1-step non-seasonal and 1-step seasonal differences were applied separately. In this case, the values for d and D were 1. As shown in Figure 4, the ACF values of lag 1, 11 and 12 exceeded the critical value. The ACF value of lag 11, the neighbor of seasonal lag 12, was caused by the cross effect of the seasonal and non-seasonal autocorrelation. Therefore, the maximum values of the seasonal parameter, Q, and the non-seasonal parameter, q, were 1 and 1, respectively. Similarly, the PACF values were significant at lag 1, 2, 11, 12 and 37 (Figure 4), so the maximum values of the seasonal parameter, P, and the non-seasonal parameter, p, were 3 and 2, respectively. We assumed that the maximum values of P was 1 to make the model concise. We searched all 24 SARIMA models that satisfied the conditions $p \leq 2$, $P \leq 1$, $q \leq 1$, and $Q \leq 1$ to find the most suitable model. The parameters of only five models were significant. Table 2 shows the results for these five models: Model A (SARIMA (0, 1, 1) (0, 1, 1)$_{12}$), Model B (SARIMA (0, 1, 1) (1, 1, 0)$_{12}$), Model C (SARIMA (1, 1, 0) (0, 1, 1)$_{12}$), Model D (SARIMA (2, 1, 0) (1, 1, 0)$_{12}$), Model E (SARIMA (2, 1, 0) (0, 1, 1)$_{12}$). The AIC value of Model A (-63.96) was lower than the AIC value for other models. Therefore, the SARIMA (0, 1, 1) (0, 1, 1)$_{12}$ model fit the data better.

Table 2

Comparison of SARIMA models with significant parameters

Model	Coefficient	β	SE(β)	T	P-value	Log likelihood	df	P-value of Ljung-Box Q test	AIC
SARIMA(0,1,1)$(0,1,1)_{12}$	AR1	/	/	/	/	145.34	10	0.756	-286.68
	MA1	-0.5859	0.0786	-7.4542	<0.001				
	SAR1	/	/	/	/				
	SMA1	-0.5779	0.0682	-8.4736	<0.001				
SARIMA(0,1,1)$(1,1,0)_{12}$	AR1	/	/	/	/	142.36	10	0.486	-280.72
	MA1	-0.5961	0.0819	-7.2784	<0.001				
	SAR1	-0.5106	0.078	-6.5462	<0.001				
	SMA1	/	/	/	/				
SARIMA(1,1,0)$(0,1,1)_{12}$	AR1	-0.3761	0.0774	-4.8592	<0.001	139.07	10	0.064	-274.14
	MA1	/	/	/	/				
	SAR1	/	/	/	/				
	SMA1	-0.5779	0.0692	-8.3512	<0.001				
SARIMA(2,1,0)$(1,1,0)_{12}$	AR1	-0.428	0.0813	-5.2645	<0.001	139.66	9	0.139	-273.33
	AR2	-0.24	0.0815	-2.9448	<0.001				
	SAR1	-0.5178	0.0783	-6.6130	<0.001				
	SMA1	/	/	/	/				
SARIMA(2,1,0)$(0,1,1)_{12}$	AR1	-0.4625	0.0817	-5.6610	<0.001	142.8	9	0.303	-279.6
	AR2	-0.2259	0.0816	-2.7684	<0.001				
	SAR1	/	/	/	/				
	SMA1	-0.5816	0.0675	-8.6163	<0.001				

SARIMAX model analysis and prediction

The CCF was used to explore the relationship between weather factors and TB incidence. The fitted SARIMA model was applied to prewhiten the data of monthly TB incidence and the monthly averaged values of the weather factors. Figure 5 shows the cross-correlation between the prewhitened weather variables and log(TB incidence) at lags of 0 to 6 months. Only positive lags would be considered because the positive value indicated that meteorological factors could affect TB incidence a certain period of time later. The weather factors, maximum

temperature, maximum daily rainfall, minimum relative humidity, mean vapor pressure, extreme wind speed, maximum atmospheric pressure, mean atmospheric pressure and illumination duration, at different lags were significantly associated with log(TB incidence). For example, the CCF for minimum relative humidity and log(TB incidence) was significant at lag 1 and lag2. The CCF for maximum daily rainfall and log(TB incidence) was significant at lag 5 and lag 6. Those significant weather factors were added as covariant into the SARIMA model to establish the SARIMAX model. First, the model was tested with single lagged weather factor. As is shown in Table 3, two SARIMAX models with covariant had significant parameters. Afterwards these two parameters were added together as covariant to build the SARIMAX model with multiple independent variables. The result indicated that maximum atmospheric pressure at lag 6, extreme wind speed at lag 5 as well as their combination affected log(TB incidence) after fitting the time series regression model, but their combination failed to have significant parameters. Maximum atmospheric pressure at lag 6 (β= -0.007, $P < 0.05$, 95% confidence interval (CI): -0.011, -0.002, mean square error (MSE): 0.279) was negatively associated with log(TB incidence), while extreme wind speed at lag 5 (β= 0.009, $P < 0.05$, 95% CI: 0.005, 0.013, MSE: 0.143) was positively associated. SARIMAX (1, 1, 1) $(0, 1, 1)_{12}$ extreme wind speed at lag 5 was the optimal model with the lowest AIC value and MSE (Table 4).

Based on the SARIMAX model constructed above, this study attempted to predict TB incidence from January 2017 to December 2017 in Guangdong. The estimated and predicted results are shown in Figure 6. The predicted monthly numbers of TB incidence all fall within the confidence intervals.

Table 3

Comparison of SARIMA models with and without covariant

Model	Meteorological factors						AIC	MSE
	Variables	Lag	β	SE(β)	T	P-value		
SARIMA(0,1,1)$(0,1,1)_{12}$	-	-	-	-	-	-	-	0.161
SARIMA(0,1,1)(0,1,1)12	maximum atmospheric pressure	0	0.005	0.004	1.195	0.234	-286.1	-
SARIMA(0,1,1)(0,1,1)12	minimum relative humidity	0	-0.002	0.001	-1.643	0.103	-287.45	-
SARIMA(0,1,1)(0,1,1)12	minimum relative humidity	1	0.002	0.001	1.357	0.177	-286.56	-
SARIMA(0,1,1)(0,1,1)12	minimum relative humidity	2	-0.002	0.001	-1.357	0.177	-286.54	-
SARIMA(0,1,1)(0,1,1)12	maximum daily rainfall	5	0.000	0.000	-1.000	0.319	-285.74	-
SARIMA(0,1,1)(0,1,1)12	maximum daily rainfall	6	0.000	0.000	0.500	0.618	-284.82	-
SARIMA(0,1,1)(0,1,1)12	maximum atmospheric pressure	6	-0.007	0.003	-2.276	0.025*	-289.77	0.279
SARIMA(0,1,1)(0,1,1)12	maximum temperature	0	-0.004	0.005	-0.774	0.441	-285.27	-
SARIMA(0,1,1)(0,1,1)12	extreme wind speed	5	0.009	0.003	3.333	0.001*	-294.90	0.143
SARIMA(0,1,1)(0,1,1)12	illumination duration	5	0.000	0.000	0.500	0.618	-285.05	-
SARIMA(0,1,1)(0,1,1)12	illumination duration	6	0.000	0.000	0.000	1.000	-284.68	-
SARIMA(0,1,1)(0,1,1)12	mean vapor pressure	0	-0.003	0.004	-0.892	0.374	-285.48	-
SARIMA(0,1,1)$(0,1,1)_{12}$	maximum atmospheric pressure	6	-0.005	0.003	-1.893	0.061	-296.64	-
	extreme wind speed	5	0.008	0.003	3.037	0.003*		-

*: *P* value<0.05

Table 4

Description of SARIMAX model with extreme wind speed at lag 5

	β	SE(β)	T	P-value
MA1	-0.529	0.085	-6.221	<0.001
SMA1	-0.570	0.069	-8.278	<0.001
Lag5 extreme wind speed	0.009	0.003	3.333	0.001
Log likelihood	150.45			
df	10			
P-value of Ljung-Box Q test	0.624			
AIC	-294.90			

Discussion

Seasonality fluctuation is a common phenomenon in the incidence of many infectious diseases including TB. In this study, a clear seasonality in the time series of TB incidence in Guangdong was found. TB incidence peaked in the spring (March to May). This result is consistent with what was found in the study by Wang et al. that TB incidence from January 1997 to August 2019 of China predominantly peaks in spring and early summer [15]. In another study of the epidemiology of TB in Xinjiang by Wubuli et al., big peaks and trough of TB were found in March and in October respectively [16]. Despite common peak and trough observed, there were also slight variations in seasonality pattern between different studies on various regions, which might arise from meteorological pattern diversity of regions involved and study periods discrepancy involved between study. The mechanisms underlying seasonal periodicity remain poorly understood, while the oscillatory changes in infectiousness, contact patterns, pathogen survival, host susceptibility, population behaviors and meteorological factors may contribute to this phenomenon [17]. For TB, strong wind and increased outdoor activities in spring may play significant role in this seasonality pattern.

This study examined the association between weather variables and TB. We found that the weather factors including maximum temperature, maximum daily rainfall, minimum relative humidity, mean vapor pressure, extreme wind speed, maximum atmospheric pressure, mean atmospheric pressure and illumination duration were significantly associated with log(TB incidence). Additionally, extreme wind speed and maximum atmospheric pressure were fitted into the model. Extreme wind speed at lag 5 was positively associated with log(TB incidence), and maximum atmospheric pressure at lag 6 was positively associated. SARIMAX (0, 1, 1) $(0, 1, 1)_{12}$ with extreme wind speed at lag 5 as covariant was the optimal model with lower AIC and highest prediction accuracy (lower MSE than model without weather factors as covariant or model with maximum atmospheric pressure as covariant), which overcame the assumption of linear dependence of variables in traditional time series model and improved the accuracy of the prediction.

Our results are similar to several previous studies on the effects of meteorological factors on TB in China. Xiao et al. [7] used DLNM to analyze the 10-year TB surveillance data in Jinghong, a city in Yunnan Province. After controlling the autocorrelation, the average temperature was negatively correlated with the incidence of TB, with lag period of 2 months; the total precipitation and the lowest relative humidity were negatively correlated, with lag

period of 3 months and 4 months respectively, and there was no lag in the effect of the mean wind speed and total sunshine hours on TB incidence. Zhang et al. [8] used GWR model to analyze the incidence of TB in 2005-2015 in different districts of the country and local meteorological factors. It was found that the average temperature was positively correlated with the incidence of TB, while the mean relative humidity and the mean wind speed were negatively correlated. The different lag effects of weather variables in other studies probably resulted from the differences between study locations involved.

Different from the normal mean value of weather factors which was found to be correlated with TB incidence in other studies, the maximum value of weather factors represents the occurrence of more abnormal weather. Abnormal weather will lead to the decline of host resistance to pathogen, which is closely related to the occurrence of infectious diseases. The possible link between TB and weather factors may be attributable to the following reasons:

We found extreme wind speed at lag 5 was positively associated with log(TB incidence). The higher the wind speed, the greater the possibility of disease transmission through respiratory droplets [18]. This result is consistent with the characteristic that TB incidence was high in spring when wind speed was high in Guangdong.

We also found maximum atmospheric pressure at lag 6 was positively associated with TB incidence. Airflow usually occurs from high-pressure areas to low-pressure areas, so the correlation between TB and atmospheric pressure may be related to wind speed. However, the mechanism by which pressure affects the transmission of TB virus is poorly understood. Additional studies are warranted to further delineate the underlying mechanisms.

In present study, maximum temperature was also found to be significantly associated with log(TB incidence). As for temperature, it can affect the indoor and outdoor activities of TB patients and other susceptible people. For example, temperature was positively associated with the number of individuals walking on the track except extreme high temperature [19, 20]. Frequent outdoor activities may increase the risk of infection with TB.

Except for the above, minimum relative humidity, mean vapor pressure and maximum daily rainfall and illumination duration of different lags were also found significantly associated with log(TB incidence). The survival of viruses depends partially on levels of relative humidity. Viruses with lipid envelopes will tend to survive longer at lower (20–30%) RHs [21]. Continuous exposure to dry air may reduce the production of protective mucus on the surface of respiratory tract, thus weakening its resistance to the pathogen [22]. In one study, precipitation, atmospheric pressure, and relative humidity were found to have negative effects on TB incidence by indirectly lowering the concentrations of inhalable particulate matter and sulfur dioxide. And TB incidence was found to be negatively correlated with the concentration of inhalable particulate matter, sulfur dioxide, or nitrogen dioxide [23]. Also, the large amount of ultraviolet light provided by long-term sunshine not only restricts the growth of M. tuberculosis but also promotes the synthesis of vitamin D, which can protect people from TB to some extent [24].

The ARIMA model, also known as the Box-Jenkins model, can analyze various types of time series data and is a commonly used model in time series analysis [25-27]. Unlike the ARIMA model, which is a univariate time series model, the ARIMAX model can deal with multivariate time series data. It adds other variables related to the target series as input variables to improve the prediction accuracy. Previous studies have explored various models, such as ARIMA [2,9], X12-ARIMA [9], ARIMA-generalized regression neural network (GRNN), DLMN [7] and DWR [8] in predicting TB [2]. However, few models have considered seasonal variation characteristics and meteorological factors [28-30]. In a study performed on three cities in Jiangsu province, the ARIMAX model was found to be

superior to the ARIMA and RNN models in predicting PTB when taking meteorological factors into consideration [10]. A time series study in Guangzhou, China, showed that an ARIMA model with imported cases and minimum temperature as input variables was superior to a single ARIMA model in forecasting dengue transmission [11]. Another time series study in Abidjan, Coted' Ivoire, also indicated that including rainfall as an input variable can increase the accuracy of the ARIMA model in predicting infuenza [31]. In the present study, we found the addition of meteorological factors decreased MSE of SARIMA model without covariant and improved the prediction accuracy (Table 4). Both ARIMA and ARIMAX are linear regression models. Considering that this relationship may be nonlinear and possess the lag time, distributed lag nonlinear model and long-term prediction model might be applied in the future studies .

This study is not free from limitations. One limitation was that only monthly TB incidence data were available. Weekly or daily incidence data may decrease the accuracy of lagged time estimation. Secondly, this model was established based on TB incidence and meteorological data in one single study area Guangdong during 13 years and the prediction was conducted for only one year period. Therefore, it is only suitable to predict the overall trends in Guangdong. Then, other than local climate conditions alone, the result might be affected by other potentially confounding variables such as vaccine usage, improvement in medical care, population growth, economic development, populations, and ecological characteristics. However, these data were not available for assessment in the current study. Therefore, comparisons among different places were necessary to attain results free from confounding factors. Extension of studying term, extension of prediction term, continuous data collection, acquisition and filtering of confounding variables, as well as prediction method improvement and update are necessary to enhance the fit degree of the model and verify the prediction accuracy. These hopefully can be fulfilled in future studies. Lastly, the long-term effect of weather factors on TB is only calculated by mathematical methods, and its biological mechanism is not clear. It is hoped that future studies will reveal the mechanism of this delayed impact and help to properly interpret the results.

Conclusions

In conclusion, SARIMAX model was successfully applied to the time series data on TB incidence in Guangdong. To our knowledge, this is the first time weather factors were added to the SARIMAX model for prediction of TB in Guangdong. The effectiveness of the model can be reflected in the fitting and predicting accuracy. In this sense, health practitioners and other relative departments might benefit from this research and take full consideration of weather factors as well as the lag effects so as to rationally allocate health resources, intervene in possible TB epidemics, and formulate corresponding healthcare strategies. With this study, earlier public health measures can be taken to effectively prevent and control the occurrence of TB. Additionally, our results suggested that it is possible to expand the application of SARIMAX model in TB and other infectious diseases prediction in different regions in the future.

Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Availability of Data and Materials

The more detailed data used to support the findings of this study are available in the supplementary information.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding Statement

This research was supported by Natural Science Foundation of Beijing Municipality (CN) (general project) (No.: 7182094) and 2020 Major Infectious Disease Prevention and Control Emergency Project of Beijing University of Chinese medicine (No.: 2020-yjgg_08)

Authors' Contributions

Conceptualization: H.Y.; data analysis: Q.T.; writing—original draft preparation and editing: Q.T.; supervision: H.J. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We are grateful to Dr. Long Yan for assistance with the code design. We also appreciate the help from Dr. Lingsheng Shi for meteorological data acquisition.

References

1. Who. Global tuberculosis report 2020[Z]. 2020:2020.
2. Wang H, Tian C W, Wang W M, et al. Time-series analysis of tuberculosis from 2005 to 2017 in China[J]. Epidemiol Infect,2018,146(8):935-939.
3. Guo Z, Xiao D, Wang X, et al. Epidemiological characteristics of pulmonary tuberculosis in mainland China from 2004 to 2015: a model-based analysis[J]. BMC Public Health,2019,19(1):219.
4. Knight G M, Mcquaid C F, Dodd P J, et al. Global burden of latent multidrug-resistant tuberculosis: trends and estimates based on mathematical modelling[J]. Lancet Infect Dis,2019,19(8):903-912.
5. Ek S. Global climate change and infectious diseases[J]. The international journal of occupational and environmental medicine,2011.
6. Sobral M, Duarte G B, Da P S A, et al. Association between climate variables and global transmission oF SARS-CoV-2[J]. Sci Total Environ,2020,729:138997.
7. Xiao Y, He L, Chen Y, et al. The influence of meteorological factors on tuberculosis incidence in Southwest China from 2006 to 2015[J]. Sci Rep,2018,8(1):10053.
8. Zhang Y, Liu M, Wu S S, et al. Spatial distribution of tuberculosis and its association with meteorological factors in mainland China[J]. BMC Infect Dis,2019,19(1):379.
9. Liao Z, Zhang X, Zhang Y, et al. Seasonality and Trend Forecasting of Tuberculosis Incidence in Chongqing, China[J]. Interdiscip Sci,2019,11(1):77-85.

10. Li Z Q, Pan H Q, Liu Q, et al. Comparing the performance of time series models with or without meteorological factors in predicting incident pulmonary tuberculosis in eastern China[J]. Infect Dis Poverty,2020,9(1):151.
11. Jing Q L, Cheng Q, Marshall J M, et al. Imported cases and minimum temperature drive dengue transmission in Guangzhou, China: evidence from ARIMAX model[J]. Epidemiol Infect,2018,146(10):1226-1235.
12. Cryer Jd C K. Time Series Analysis With Applications in R. Second Edition ed.[M]. New York:Springer,2008.
13. Hao Y, Wang R R, Han L, et al. Time series analysis of mumps and meteorological factors in Beijing, China[J]. BMC Infect Dis,2019,19(1):435.
14. Yan L, Wang H, Zhang X, et al. Impact of meteorological factors on the incidence of bacillary dysentery in Beijing, China: A time series analysis (1970-2012)[J]. PLoS One,2017,12(8):e0182937.
15. Wang Y, Xu C, Ren J, et al. Secular Seasonality and Trend Forecasting of Tuberculosis Incidence Rate in China Using the Advanced Error-Trend-Seasonal Framework[J]. Infect Drug Resist,2020,13:733-747.
16. Wubuli A, Li Y, Xue F, et al. Seasonality of active tuberculosis notification from 2005 to 2014 in Xinjiang, China[J]. PLoS One,2017,12(7):e0180226.
17. Fisman D. Seasonality of viral infections: mechanisms and unknowns[J]. Clin Microbiol Infect,2012,18(10):946-54.
18. Li R, Lin H, Liang Y, et al. The short-term association between meteorological factors and mumps in Jining, China[J]. Sci Total Environ,2016,568:1069-1075.
19. Liao C M, Hsieh N H, Huang T L, et al. Assessing trends and predictors of tuberculosis in Taiwan[J]. BMC Public Health,2012,12:29.
20. Suminski R R, Poston W C, Market P, et al. Meteorological conditions are associated with physical activities performed in open-air settings[J]. Int J Biometeorol,2008,52(3):189-97.
21. Tang J W. The effect of environmental parameters on the survival of airborne infectious agents[J]. J R Soc Interface,2009,6 Suppl 6:S737-46.
22. Kudo E, Song E, Yockey L J, et al. Low ambient humidity impairs barrier function and innate resistance against influenza infection[J]. Proc Natl Acad Sci U S A,2019,116(22):10905-10910.
23. Zhang C Y, Zhang A. Climate and air pollution alter incidence of tuberculosis in Beijing, China[J]. Ann Epidemiol,2019,37:71-76.
24. Zhang Y, Liu M, Wu S S, et al. Spatial distribution of tuberculosis and its association with meteorological factors in mainland China[J]. BMC Infect Dis,2019,19(1):379.
25. Wang Y W, Shen Z Z, Jiang Y. Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China[J]. PLoS One,2018,13(9):e0201987.
26. Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset[J]. Data Brief,2020,29:105340.
27. Liu L, Luan R S, Yin F, et al. Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model[J]. Epidemiol Infect,2016,144(1):144-51.
28. Kumar V, Singh A, Adhikary M, et al. Seasonality of tuberculosis in delhi, India: a time series analysis[J]. Tuberc Res Treat,2014,2014:514093.
29. Wang K W, Deng C, Li J P, et al. Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network[J]. Epidemiol Infect,2017,145(6):1118-1129.

30. de Castro D B, de Seixas M E, Sadahiro M, et al. Tuberculosis incidence inequalities and its social determinants in Manaus from 2007 to 2016[J]. Int J Equity Health,2018,17(1):187.
31. N'Gattia A K, Coulibaly D, Nzussouo N T, et al. Effects of climatological parameters in modeling and forecasting seasonal influenza transmission in Abidjan, Cote d'Ivoire[J]. BMC Public Health,2016,16:972.

Figures

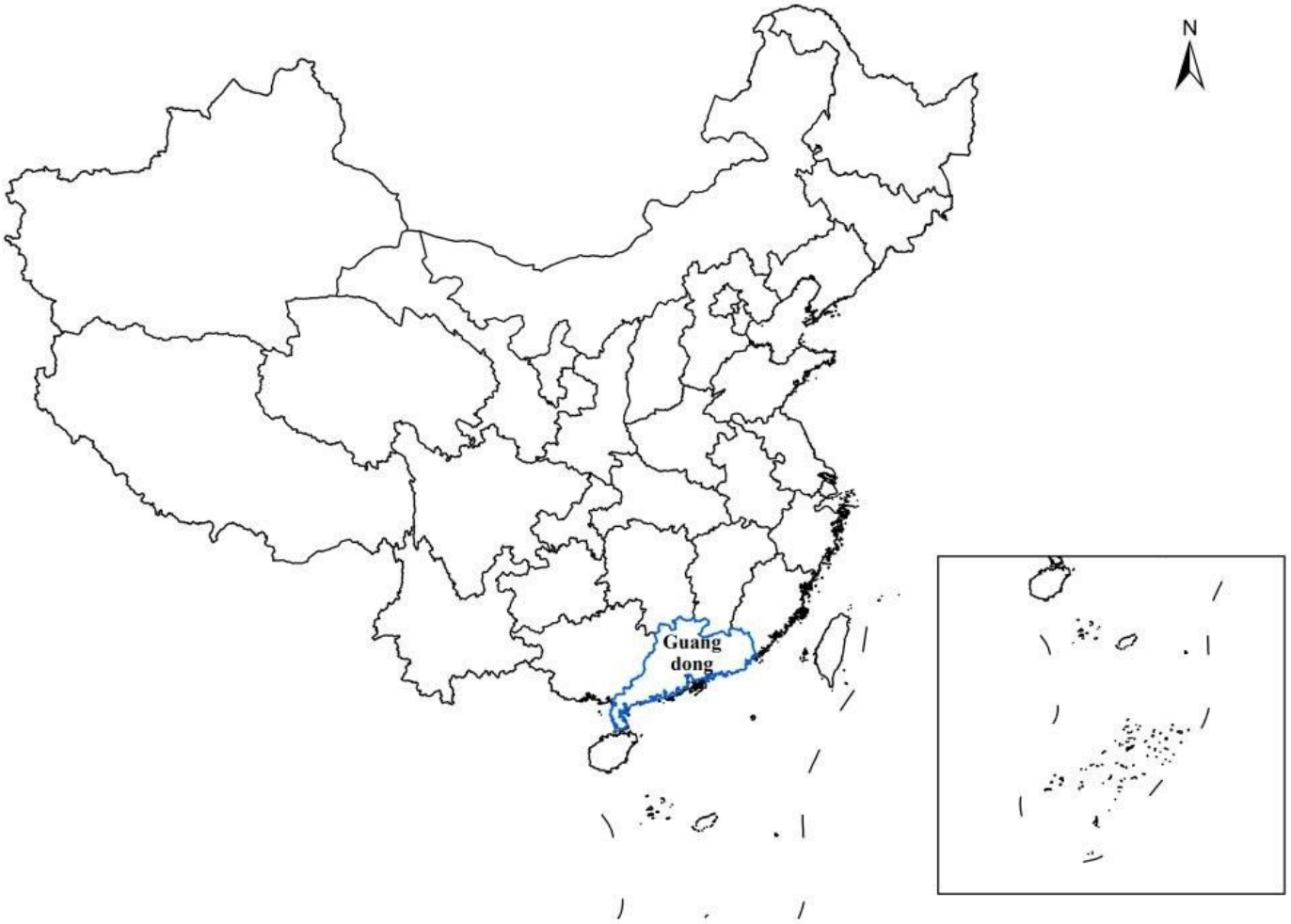

Figure 1

Location of the study area in China. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

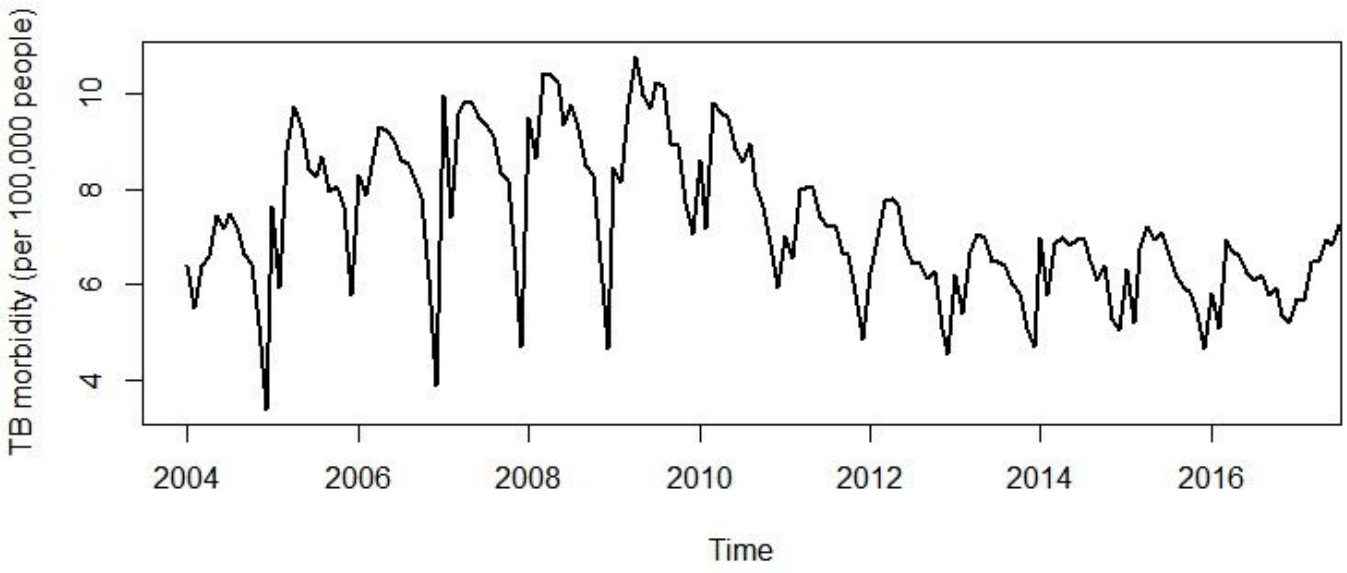

Figure 2

Time series plot of the TB incidence in Guangdong, 2004-2017

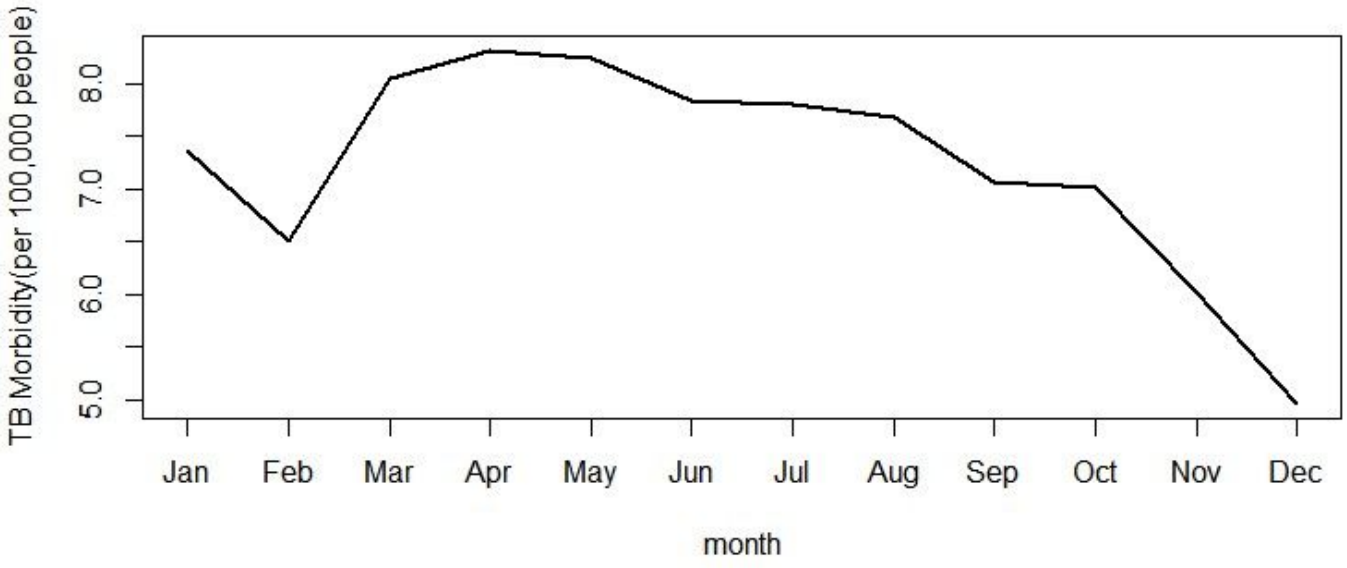

Figure 3

Monthly incidence of TB in Guangdong, 2004-2017

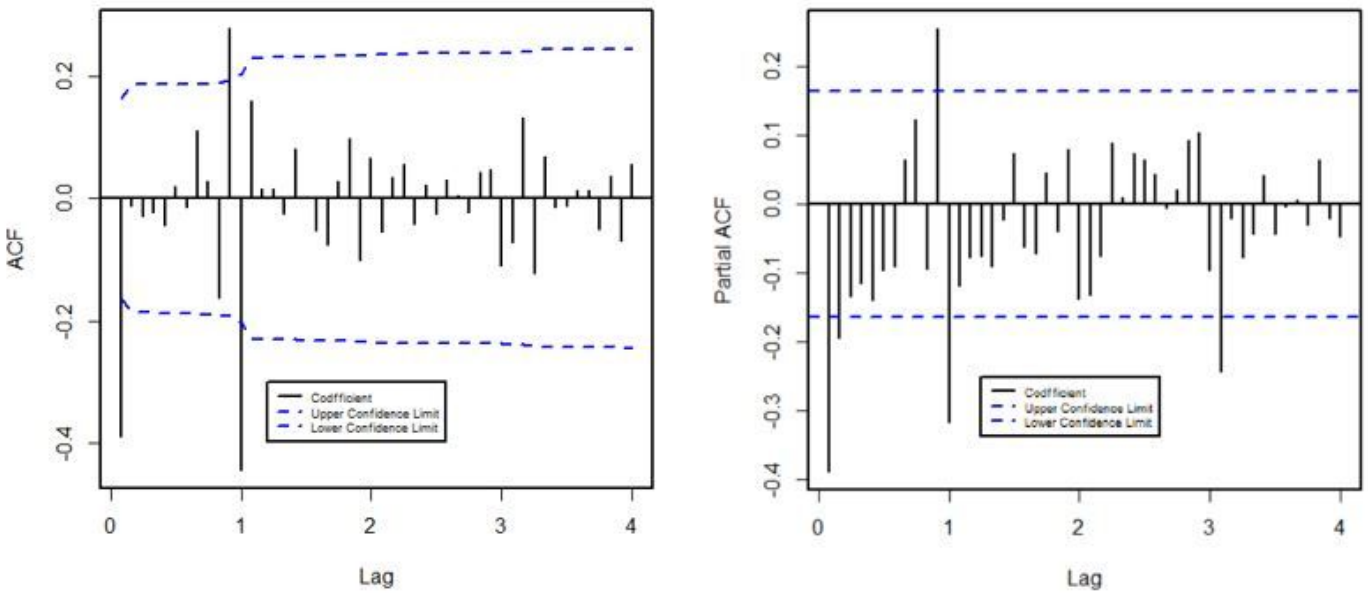

Figure 4

Results of ACF and PACF for time series analysis

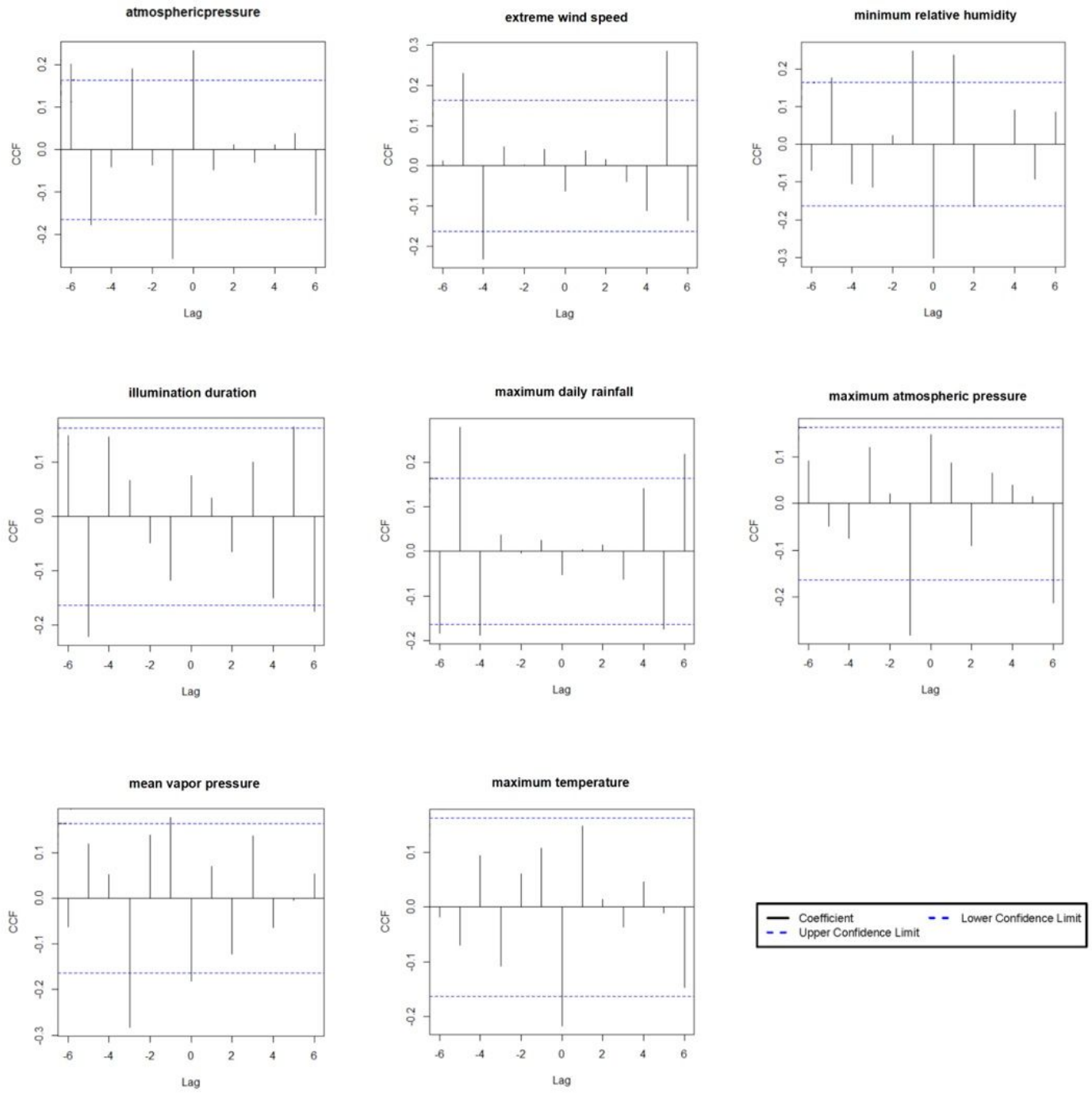

Figure 5

The cross-correlation between the prewhitened TB incidence and weather variables

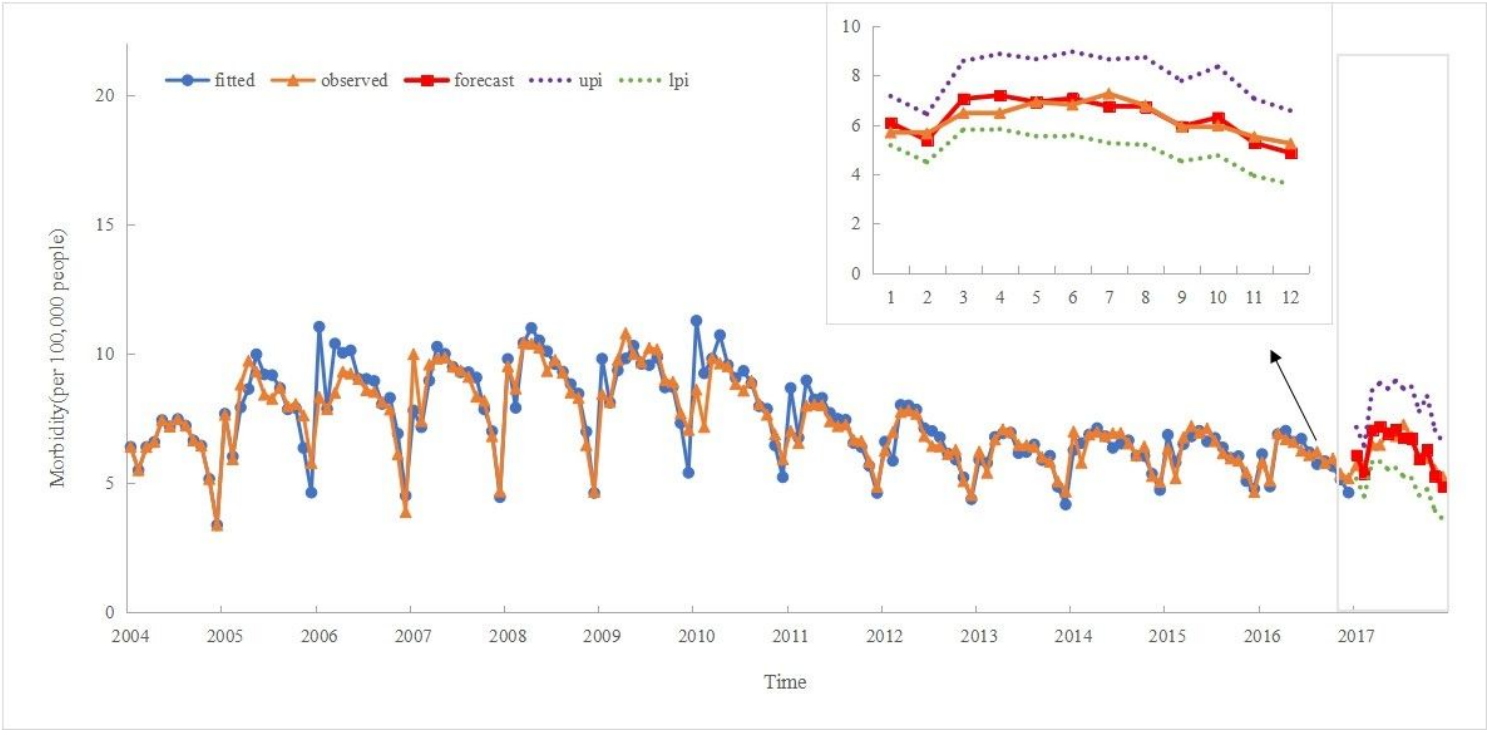

Figure 6

Prediction by the SARIMAX model

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryfile.xlsx
- graphicalabstract.tif