Research Article

Predictive Study of Tuberculosis Incidence by ARMA Model Combined with Air Pollution Variables

Yanling Zheng

College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830011, China

Correspondence should be addressed to Yanling Zheng; zhengyl_math@sina.cn

Received 14 December 2019; Revised 5 April 2020; Accepted 9 April 2020; Published 4 May 2020

Academic Editor: Rosa M. Lopez Gutierrez

Copyright © 2020 Yanling Zheng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

China has the second largest number of tuberculosis (TB) cases in the world, and the Xinjiang province has the highest TB incidence in China. Urumqi is the capital city of Xinjiang; good TB prevention and control in Urumqi can provide an example for other parts of Xinjiang, considering that predicting the TB incidence is the prerequisite of prevention and control; therefore, it is necessary to do a prediction study on TB incidence in Urumqi. In this paper, based on the data of TB incidence and air pollution variables ($PM_{2.5}$, PM_{10} , SO_2 , CO , NO_2 , O_3) in Urumqi, the ARMA (1, (1, 3)) + model was established by time series ARMA model method, cross-correlation analysis, and principal component regression method, and its predictive performance was superior to that of the ARMA (1, (1, 3)) model based on TB historical data. The research idea of this paper was good, which can provide a reference for other researchers. The prediction of the ARMA (1, (1, 3)) + model can provide scientific help for TB prevention and control in Urumqi, China. During the analysis, it was found that the higher the concentration of O_3 , the higher the incidence of TB. This study suggests that people in Urumqi should pay more attention to the hazards of O_3 and do a good job of personal protection.

1. Introduction

Tuberculosis (TB) is a chronic infectious disease with serious long-term health hazards; it is also called consumption disease and is caused by *Mycobacterium tuberculosis*, which mainly affects the lungs of the human body. If this disease is not found in time and the treatment is not thorough, it will cause serious harm to health, even cause respiratory failure and death, and bring heavy economic burden to patients and families [1]. China is one of the 22 countries with high burden of TB, and the burden of TB ranks second in the world [2].

Xinjiang Uygur Autonomous region is a province in northwest China, located in the center of Eurasia, latitude $34^{\circ}22' \sim 49^{\circ}33'N$, longitude $73^{\circ}41' \sim 96^{\circ}18'E$ [3]. In recent years, the annual TB incidence in Xinjiang has always been the highest in China, and is much higher than 2 to 3 times the national level. Urumqi is the capital city of Xinjiang (its position is shown in the red part of Figure 1); its average annual TB incidence is 64.33 from 2014 to 2017.

Due to the better per capita economic, medical, and environmental health conditions, the prevention and control of TB in Urumqi are much better than those of other areas of Xinjiang; therefore, the TB incidence is relatively low, but the prevention and control work of TB in Urumqi cannot be relaxed.

The prediction of infectious diseases is an important health work, which can detect the development trend of diseases as early as possible and increase the predictability of epidemic prevention work and help making health decisions [4]. ARMA model method is a popular prediction method in recent years, it can reveal the quantitative relationship between the research object and other objects with the development and change of time, this method has been widely used in the prediction of infectious diseases and has achieved successful results [5–8]. Although the ARMA model has good prediction performance, it only forecasts according to the historical incidence data and does not take into account some factors that may affect the incidence of the disease. Many studies have found that the infectious disease is related

FIGURE 1: Geographic location of Urumqi, Xinjiang, China.

to air pollution variables [9–13]. Proper consideration of air quality factors can not only improve the accuracy of prediction, but also improve the scientific nature of prediction. General linear regression model (GLM) can take into account the influencing factors of prediction variables [14], so as to improve the accuracy and scientific nature of prediction. In regression analysis, if there is a high correlation between independent variables, it will lead to pseudoregression phenomenon. In order to avoid the false regression problem, principal component regression is a good choice; it can effectively avoid the correlation between variables, establish a real multiple regression model, and thus give a scientific prediction [15–18].

In this study, first, the ARMA model method was used to establish the prediction model of TB incidence in Urumqi based on the data of monthly TB incidence from January 2014 to December 2017; second, considering that the TB incidence may be affected by air-quality factors, cross-

correlation analysis was used to analyze the effect of air pollution variables on TB incidence and to improve the prediction accuracy of the ARMA model, some of the variables that affected the TB incidence were incorporated into the ARMA model. Then, considering the high correlation between the independent variables of the new model, this study used principal component regression to do analysis, which can avoid pseudoregression, and finally, a prediction model with air pollution variables and high prediction accuracy was established.

2. Materials and Methods

2.1. Data Source. The data of the monthly TB incidence (per 100,000 populations) in Urumqi from January 2014 to December 2017 were obtained from center for Disease Control and Prevention (CDC), Urumqi, China. The data of the air pollution variables in Urumqi from January 2014 to

June 2018 were obtained from Urumqi Meteorological Bureau, the six air pollution variables are PM_{2.5}, PM₁₀, SO₂, CO, NO₂, and O₃.

2.2. Model Descriptions

2.2.1. ARMA Model Description. The auto regression moving average (ARMA) model, also known as Box–Jenkins model, includes two different parts, the auto regression part, and the moving average part [19]. The auto regression moving average process can be expressed as ARMA (p, q). The specific expressions is as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}, \quad (1)$$

where ε_t is white noise, c is a constant, $\phi_1, \phi_2, \dots, \phi_p$ are the coefficients of the autoregressive model, $\theta_1, \theta_2, \dots, \theta_q$ are the coefficients of moving average model.

If some values of autocorrelation coefficients and moving average coefficients in the ARMA (p, q) model are 0, then the model is called sparse model; for example, if θ_2 is 0 in ARMA (1, 3) model, then, the ARMA (1, 3) model becomes sparse model ARMA (1, (1, 3)).

In general, ARMA modeling includes five steps.

First step: ARMA modeling needs data stability, Augmented Dickey–Fuller Test (ADF) is used to test the stability of the data, and the test level P is 0.05.

Second step: by the analysis of autocorrelation and partial autocorrelation functions (ACF and PACF), the model can be identified, and the values of tentative p, q are determined.

Third step is to estimate and test the parameters of the model and calculate the AIC and SC values of the model; the smaller the AIC and SC, the better the performance of the model.

Fourth step is to do model residual analysis by Box–Jenkins Q test, if the residual sequence of the model is white noise, the prediction stage begins; otherwise, it is shown that the information extraction is not sufficient, and some methods need to be considered to improve the accuracy of the model.

Fifth step is using the established model to predict the future change trend of the sequence.

2.2.2. Principal Component Regression Prediction. General linear regression models (GLMs) are also often used for prediction; when the correlation between independent variables is high, the GLM model will appear pseudoregression phenomenon [20–23]. One of the commonly used methods to deal with this phenomenon is principal component regression; its modeling steps are as follows.

First step is to analyze the correlation between independent variables; if the correlation is significant, move on to the second step.

Second step is to extract principal components (PCs) from all independent variables; the number of PCs is determined by the cumulative contribution rate of eigenvalues.

Generally, the cumulative contribution rate needs to be more than 85%, and its corresponding eigenvalues are extracted. The correlation between the PCs is zero.

Third step is using the reserved PCs as independent variables to build the GLM model and then to estimate and test the parameters of the model. Taking $P < 0.05$ as the parameter has significant statistical significance. R^2 is used to measure the goodness of fit of the model.

Fourth step is to fit and predict dependent variables based on established models.

2.3. The Indexes of Assessing Forecast Accuracy. In this study, we use mean absolute error (MAE) to measure the prediction accuracy. The smaller the MAE is, the better the model is, and the higher the prediction accuracy is.

The MAE is defined as

$$MAE = \frac{\sum_{t=1}^n |y_t - \hat{y}_t|}{n}, \quad (2)$$

where $\hat{y}_t$ is the predicted value, y_t is the observed value at t , and n is the number of observed values.

2.4. Data Processing and Analysis. All analyses were performed using Arcmap10.1, Eviews 7.2, SPSS17.0, and Matlab2016b.

3. Results

From January 2014 to December 2017, there were 8923 TB cases in Urumqi, Xinjiang, and average annual TB cases were 2231, and annual TB incidence was 63.44 (per 100,000 populations). The change of TB incidence is shown in Figure 2. The TB incidence increased significantly in 2015 compared with 2014, and it changed little from 2015 to 2017, only slightly with the season.

Considering the relationship between the TB incidence and air pollution, we collected relevant air pollution data in Urumqi from January 2014 to December 2017, including PM_{2.5}, PM₁₀, SO₂, CO, NO₂, and O₃. The six indicators changed over time as shown in Figure 3, and their statistical descriptions are shown in Table 1.

First of all, the TB incidences data were analyzed so as to establish the ARMA prediction model; the ARMA model needs the data to be stable; for this reason, the ADF test was used to test the stability of the original data, and the test results are shown in Table 2, the P value was 0.044 and less than 0.05, which showed that the data was stable and could be used directly to found ARMA model. The ACF and PACF diagram of the data are shown in Figure 4, based on Figure 4, it was preliminarily judged that p took 1, q might have taken 1, 2, or 3, and the various tentative ARMA models were founded; these models with their parameter test results and AIC and SC are shown in Table 3. From the results of parameter test, it could be seen that only two models could pass all the parameter tests, namely, ARMA (1, (1, 3)) model and ARMA (1, (3)) model. The values of AIC and SC of ARMA (1, (1, 3)) model were smaller than those of ARMA (1, (3)) model, and the R^2 value of ARMA (1, (1, 3)) model was

FIGURE 2: The TB incidences in Urumqi from January 2014 to December 2017.

FIGURE 3: Time series chart of data on six air pollution variables from January 2014 to December 2017.

TABLE 1: Descriptive statistics results of air pollution data.

Variables	Range	Mean $\pm$ s.d
PM _{2.5} ($\mu\text{g}/\text{m}^3$)	18–226	67.65 $\pm$ 52.93
PM ₁₀ ($\mu\text{g}/\text{m}^3$)	54–287	127.08 $\pm$ 55.53
CO (ppm)	0.53–3.45	1.43 $\pm$ 0.88
SO ₂ (ppb)	6–50	17.65 $\pm$ 11.42
NO ₂ (ppb)	32–90	52.15 $\pm$ 16.30
O ₃ (ppm)	14–124	61.85 $\pm$ 33.48

TABLE 2: The ADF test of the training data.

	t-statistic	P
Augmented Dickey–Fuller test statistic	–2.98	0.044
Test critical values	1% level	–3.58
	5% level	–2.93
	10% level	–2.60

FIGURE 4: The ACF and PACF graphs of TB incidence. (a) ACF with 2stds Bounds (i.e., approximate 95% confidence interval). (b) PACF with 2stds Bounds (i.e., approximate 95% confidence interval).

TABLE 3: Parameter estimates of the tentative models with their AIC and SC.

Model	Variable	Coefficient	Std. error	t-statistic	P	AIC	SC
ARMA (1, 1)	C	5.41	0.46	11.64	<0.01	2.57	2.69
	AR (1)	0.79	0.12	6.51	<0.01		
	MA (1)	-0.23	0.20	-1.14	0.26		
ARMA (1, 2)	C	5.35	0.45	11.82	<0.01	2.56	2.71
	AR (1)	0.72	0.16	4.53	<0.01		
	MA (1)	-0.21	0.20	-1.05	0.29		
	MA (2)	0.25	0.17	1.45	0.16		
ARMA (1, (2))	C	5.32	0.39	13.69	<0.01	2.53	2.65
	AR (1)	0.602	0.13	4.69	<0.01		
	MA (2)	0.28	0.15	1.82	0.08		
ARMA (1, 3)	C	5.37	0.46	11.68	<0.01	2.60	2.80
	AR (1)	0.74	0.216	3.48	<0.01		
	MA (1)	-0.214	0.266	-0.81	0.42		
	MA (2)	0.234	0.186	1.291	0.20		
	MA (3)	-0.06	0.21	-0.27	0.79		
ARMA (1, (3))	C	5.72	0.15	37.39	<0.01	2.46	2.58
	AR (1)	0.802	0.07	11.62	<0.01		
	MA (3)	-0.93	0.046	-20.18	<0.01		
ARMA (1, (1, 3))	C	5.79	0.15	38.68	<0.01	2.39	2.55
	AR (1)	0.86	0.05	17.20	<0.01		
	MA (1)	-0.48	0.11	-4.38	<0.01		
	MA (3)	-0.48	0.11	-4.41	<0.01		
ARMA (1, (2, 3))	C	5.36	0.39	13.50	<0.01	2.57	2.73
	AR (1)	0.60	0.13	4.45	<0.01		
	MA (2)	0.28	0.16	1.80	0.08		
	MA (3)	0.01	0.15	0.08	0.92		

also larger than that of ARMA (1, (3)) model; therefore, ARMA (1, (1, 3)) model was the best model.

The mathematical expression of the ARMA (1, (1, 3)) model was as follows:

$$y_t = 5.79 + 0.86y_{t-1} + \varepsilon_t + 0.479\varepsilon_{t-1} + 0.482\varepsilon_{t-3}. \quad (3)$$

Using this model to fit the historical data of TB incidence, the MAE was 0.633.

The residual sequence of the ARMA (1, (1, 3)) model was tested by Box–Jenkins Q test. The test results are shown in Figure 5, it can be seen that the correlation coefficient of the residual sequence at lag 6 was significantly greater than twice the standard deviation, which indicated that the residual sequence was not white noise sequence, and the accuracy of the ARMA (1, (1, 3)) model needed to be further improved.

Considering that the TB incidence may be affected by air pollution variables, the spearman correlation analysis between the TB incidence and the air pollution variables was carried out, and the results of the analysis are shown in Table 4. From the results of Table 4, it could be seen that only PM₁₀ and SO₂ were associated with TB incidence, further; cross-correlation analysis was carried out considering the lag effect of air pollution variables on tuberculosis incidence. Figure 6 shows the identified highest associated time lags TB incidence to each air pollution variable, and these highest lag correlation coefficients are shown in Table 5.

Then, on the basis of ARMA (1, (1, 3)) model, the time lead variables of air pollution factors, such as PM_{2.5} (lead 3), PM₁₀ (lead 3), SO₂ (lead 1), CO (lead 3) NO₂ (lead 2), and O₃ (lead 3) were included in new model for further analysis. To found the GLM model with the TB incidence as dependent variable, AR (1), MA (1), MA (3), PM_{2.5} (lead 3), PM₁₀ (lead 3), SO₂ (lead 1), CO (lead 3), NO₂ (lead 2), and O₃ (lead 3) as independent variables. In order to avoid pseudo regression, first of all, Pearson's correlation analysis was performed on these independent variables, and the results showed that PM_{2.5} (lead 3) was highly correlated with PM₁₀ (lead 3) and CO (lead 3) (correlation coefficient is more than 0.8), and PM₁₀ (lead 3) was highly correlated with CO (lead 3) (correlation coefficient is more than 0.8), and PM_{2.5} (lead 3) was significantly correlated with NO₂ (lead 2) and O₃ (lead 3), and CO (lead 3) was significantly correlated with NO₂ (lead 2) and O₃ (lead 3), and MA (1) was significantly correlated with AR (1) (see Table 6); therefore, GLM could not be established directly, and the principal component regression was used to establish the model. Firstly, for AR (1), MA (1), MA (3), PM_{2.5} (lead 3), PM₁₀ (lead 3), SO₂ (lead 1), CO (lead 3), NO₂ (lead 2), and O₃ (lead 3), their principal components (PCs) were extracted. Since the cumulative contribution rate of the first six eigenvalues had reached 96.88 (see Table 7), it was enough to extract the six PCs for modeling. The six PCs were independent of each other, and the scores of the six principal components were calculated, to found the ARMA (1, (1, 3)) + model using six PCs as independent variables and TB incidence as dependent variable; finally, ARMA (1, (1, 3)) + model was established by stepwise regression method. The test results of ARMA (1, (1, 3)) + model showed that the ARMA (1, (1, 3)) + model was statistic significant, the R^2 of

the model was 0.851, and the MAE was 0.36, the parameter test results are shown in Table 8. The ARMA (1, (1, 3)) + model represented by principal components was as follows:

$$y_t = 5.3 - 0.336PC_1 - 0.189PC_2 + 0.267PC_3 - 0.615PC_4 + 0.543PC_6. \quad (4)$$

After converting PCs into standardized variables, the ARMA (1, (1, 3)) + model was as follows:

$$y_t = 5.3 + 0.650Z_{y_{t-1}} + 0.153Z_{\varepsilon_{t-1}} + 0.189Z_{\varepsilon_{t-3}} - 0.059Z_{PM_{25}} - 0.194Z_{PM_{10}} - 0.062Z_{CO} - 0.123Z_{SO_2} - 0.056Z_{NO_2} + 0.137Z_{O_3}. \quad (5)$$

Based on ARMA (1, (1, 3)) + model, the monthly TB incidence in Urumqi from October 2017 to March 2018 was predicted, which were 5.15, 4.81, 4.89, 5.56, 5.75 and 6.56, respectively. The fitting and prediction diagram of the ARMA (1, (1, 3)) + model is shown in Figure 7, from which it could be seen intuitively that the fitting and prediction effect were good.

4. Discussion

Dynamic estimation and prediction of disease epidemic is an important link in the prevention and control of infectious diseases, and it is the main basis for health management departments to formulate prevention and control countermeasures and allocate resources [19, 24]. However, due to the influence of many uncertain factors on the prevalence of TB, it is difficult to identify early, which leads to the lag of the corresponding prevention and control measures. How to warn TB epidemic timely and effectively is an important part of disease prevention and control.

It is generally believed that air pollution factors affect the prevalence of TB by affecting the survival of the virus, population activities, individual behavior, and so on [25]. The effect of air pollution variables on TB epidemic also suggests that TB prediction should include some meaningful explanatory variables, and it should not be limited to historical data of TB incidence.

In this study, firstly, the single ARMA (1, (1, 3)) model based on the TB incidence in Urumqi from January 2014 to December 2017 was established. The AIC and SC of the ARMA (1, (1, 3)) model were minimal in the seven tentative ARMA models, and all the parameters test of the ARMA (1, (1, 3)) model were statistically significant, R^2 of the model was 0.59 and the MAE was 0.633. The only deficiency of the ARMA (1, (1, 3)) model is that residual sequence of the model was not white noise, which showed that the ARMA (1, (1, 3)) model was not sufficient for information extraction. Taking into account the effect of air pollution variables (PM_{2.5}, PM₁₀, SO₂, CO, NO₂, and O₃) on the TB incidence, this study performed spearman correlation analysis between air pollution variables and the TB incidence, and the results showed that only PM₁₀ and SO₂ were significantly correlated with the TB incidence, some studies suggest that the effect of

FIGURE 5: The ACF and PACF graphs of residual sequences of ARMA (1, (1, 3)) model. (a) ACF with 2stds Bounds (i.e., approximate 95% confidence interval). (b) PACF with 2stds Bounds (i.e., approximate 95% confidence interval).

TABLE 4: The results of spearman correlation test between air pollution variables and TB incidence.

Variables	Coefficient	<i>P</i>
pm _{2.5}	-0.176	0.231
pm ₁₀	-0.286*	0.049
SO ₂	-0.474**	0.001
CO	-0.069	0.64
NO ₂	-0.21	0.151
O ₃	0.225	0.124

P* < 0.05, and *P* < 0.01.

air pollution variables and weather factors on infectious disease is delayed [26–29], so cross-correlation analysis was done for studying the delayed effect of six air pollution variables on the TB incidence; it was found that the three-month delay in the TB incidence was most associated with PM_{2.5}, PM₁₀, CO, and O₃, the one-month delay in the TB incidence was the most correlated with SO₂, and the two-month delay in the TB incidence was the most correlated with NO₂; in other words, PM_{2.5}, PM₁₀, SO₂, Co, NO₂, and O₃ had a time-leading effect on the TB incidence, and their leading orders were, 3, 3, 1, 3, 2, and 3, respectively; to this end, we considered establishing a new model based on ARMA (1, (1, 3)) model combined with six variables, such as PM_{2.5} (lead3), PM₁₀ (lead3), SO₂ (lead1), CO (lead3), NO₂ (lead2), and O₃ (lead3), to predict TB incidence, which should improve the prediction accuracy, so the TB incidence as dependent variables and AR (1), MA (1), MA (3), PM_{2.5} (lead 3), PM₁₀ (lead 3), SO₂ (lead 1), CO (lead 3), NO₂ (lead 2),

and O₃ (lead 3) as independent variables were used to establish a regression model; however, after careful analysis, it was found that there were high correlations (correlation coefficient greater than 0.8) and significant correlations (*P* < 0.001) between independent variables, which might lead to pseudo-regression, in order to avoid this situation, the principal component regression method was used to further establish the model; then, six PCs were extracted from nine independent variables (AR (1), MA (1), MA (3), PM_{2.5} (lead 3), PM₁₀ (lead 3), SO₂ (lead 1), CO (lead 3), NO₂ (lead 2), and O₃ (lead 3)), and the cumulative contribution rate reached 96.88%, which indicated that the information extraction was sufficient. Using six PCs as independent variables and TB incidence as dependent variables, we established ARMA (1, (1, 3)) + model, its *R*² (0.851) was more than that of ARMA (1, (1, 3)), its MAE (0.36) was less than that of ARMA (1, (1, 3)), which suggested that ARMA (1, (1, 3)) + model was better than ARMA (1, (1, 3)) model. By ARMA (1, (1, 3)) + model, the TB incidences from October 2017 to March 2018 were predicted, and the predicted values were 5.15, 4.81, 4.89, 5.56, 5.75, and 6.56, respectively, which was able to capture the trend pattern of TB in Urumqi.

In this study, it was found that air pollution variables had a significant time lag effect to TB incidence in Urumqi; from the results of cross-correlation analysis and ARMA (1, (1, 3)) + model analysis, it is found that the effect of PM_{2.5}, PM₁₀, SO₂, CO, and NO₂ on the TB incidence was negative, which seemed to be contrary to some research conclusions [30, 31], but after careful analysis, it was found that there was no contradiction, and the reason was that people often

FIGURE 6: Cross-correlation analysis of TB incidence to air pollution variables.

TABLE 5: The highest associated time lags and maximum correlation coefficients TB incidence to each pollution variable.

Pollution variables	PM _{2.5}	PM ₁₀	SO ₂	CO	NO ₂	O ₃
Correlation	-0.37**	-0.54**	-0.53**	-0.36**	-0.45**	0.53**
Coefficient (lag)	(lag3)	(lag3)	(lag1)	(lag3)	(lag2)	(lag3)

** $P < 0.01$.

TABLE 6: The results of Pearson's correlation analysis between independent variables.

		PM ₂₅ (lead3)	PM ₁₀ (lead3)	SO ₂ (lead1)	CO (lead3)	NO ₂ (lead2)	O ₃ (lead3)	AR (1)	MA (1)
PM ₂₅ (lead3)	Coefficient	1							
	P								
PM ₁₀ (lead3)	Coefficient	0.869	1						
	P	<0.001							
SO ₂ (lead1)	Coefficient	0.115	0.303	1					
	P	0.451	0.043						
CO (lead3)	Coefficient	0.97	0.821	0.101	1				
	P	<0.001	<0.001	0.51					
NO ₂ (lead2)	Coefficient	0.614	0.553	0.563	0.607	1			
	P	<0.001	<0.001	<0.001	<0.001				
O ₃ (lead3)	Coefficient	-0.704	-0.621	-0.266	-0.751	-0.52	1		
	P	<0.001	<0.001	0.077	<0.001	<0.001			
AR (1)	Coefficient	-0.334	-0.508	-0.436	-0.349	-0.313	0.541	1	
	P	0.025	<0.001	0.002	0.019	0.034	<0.001		
MA (1)	Coefficient	-0.367	-0.318	-0.067	-0.426	-0.14	0.449	0.652	1
	P	0.013	0.033	0.656	0.004	0.353	0.002	<0.001	
MA (3)	Coefficient	0.009	-0.024	0.282	-0.1	0.238	0.119	0.108	0.022
	P	0.953	0.876	0.061	0.95	0.116	0.438	0.482	0.884

TABLE 7: Total variance explained.

Component	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	4.50	50.01	50.01	4.50	50.01	50.01
2	1.47	16.35	66.36	1.47	16.35	66.36
3	1.24	13.75	80.11	1.24	13.75	80.11
4	0.81	8.94	89.06	0.81	8.94	89.06
5	0.40	4.47	93.52	0.40	4.47	93.52
6	0.30	3.36	96.88	0.30	3.36	96.88
7	0.19	2.15	99.03			
8	0.06	0.73	99.76			
9	0.02	0.24	100.00			

TABLE 8: The test results of principal component regression.

Variables	coefficient	Std. error	t-statistic	Prob
C	5.300	0.100	53.001	<0.001
PC1	-0.336	0.048	-7.058	<0.001
PC2	-0.189	0.083	-2.269	0.029
PC3	0.267	0.091	2.936	0.006
PC4	-0.615	0.113	-5.459	<0.001
PC6	0.543	0.184	2.952	0.005

FIGURE 7: The comparison of fitting and prediction TB effects of ARMA (1, (1, 3)) + model.

looked over the weather forecast before they planned to go out or participate in activities in Urumqi; when the weather forecast showed that the air pollution was serious, and there was heavy haze, people tended to go out less or participate less in some activities; this greatly reduced the chance of close contact between people, thus reducing the probability of being infected with TB. The effect of O_3 on the TB incidence was positive; that is, the higher the concentration of O_3 , the higher the incidence of TB. The main reasons were that on one hand, on sunny and high-temperature days, O_3 concentration was relatively high, and in this kind of weather, people in Urumqi liked to go out to party or participate in activities, which

increased the chance of close contact, thus increasing the probability of being infected with TB and that on the other hand, O_3 at high concentration will strongly stimulate the respiratory tract, causing sore throat, chest tightness, and cough, causing bronchitis and emphysema, thus reducing the immunity of people infected with TB.

Many studies have found that both air pollution factors and meteorological factors affect the TB incidence of tuberculosis. Due to the limitation of data collection, in our study, only air pollution factors were included into the prediction of TB incidence; in further research, we will also consider meteorological factors.

5. Conclusions

Most predictive models of TB incidence were based only on historical data of TB incidence and did not consider the effect of air pollution on TB incidence, in this study, based on the data of historical TB incidence and air pollution variables, the ARMA (1, (1, 3)) + prediction model of TB incidence in Urumqi was established, its prediction accuracy was higher than that of the ARMA (1, (1, 3)) model considering only historical TB incidence data. Based on this ARMA (1, (1, 3)) + model, we predicted the TB incidence from October 2017 to March 2018, which were consistent with the variation of TB incidence data in Urumqi. The prediction method used in this study was good, it can provide some reference ideas for other researchers to improve the accuracy of prediction model, and the ARMA (1, (1, 3)) + model can also provide prediction for the TB incidence in Urumqi, which can provide scientific reference for the prevention and control of the TB in Urumqi, China. In the modeling process, the effect of air pollution variables on the TB incidence was analyzed, and there is one thing to note: the higher the O₃ concentration, the higher the TB incidence; on sunny, high-temperature days, O₃ levels exceed the standard, and people in Urumqi preferred to go out to party or participate in activities in such weather, ignoring the dangers posed by O₃. This study suggests that people in Urumqi pay more attention to O₃ hazards and do a good job of personal protection.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by Provincial ministry co-construction of State Key Laboratory Project on the Causes and Control of High Incidence of Diseases in Central Asia, China (grant no. SKL-HIDCA-2017-12).

References

- [1] H. A. Mahmud, H. Seo, S. Kim et al., "Thymoquinone (TQ) inhibits the replication of intracellular Mycobacterium tuberculosis in macrophages and modulates nitric oxide production," *BMC Complementary & Alternative Medicine*, vol. 17, no. 1, p. 279, 2017.
- [2] WHO, *Global Tuberculosis Report*, WHO, Geneva, Switzerland, 2019.
- [3] Y. Bin, "A boost for Xinjiang," *Beijing Review*, vol. 53, no. 23, p. 2, 2010.
- [4] E. L. Ray and N. G. Reich, "Prediction of infectious disease epidemics via weighted density ensembles," *PLoS Computational Biology*, vol. 14, no. 2, Article ID e1005910, 2018.
- [5] E. L. Ray, K. Sakrejda, S. A. Lauer et al., "Infectious disease prediction with kernel conditional density estimation," *Statistics in Medicine*, vol. 36, no. 18, 2017.
- [6] Q. Xu, R. Li, Y. Liu et al., "Forecasting the incidence of mumps in Zibo city based on a SARIMA model," *International Journal of Environmental Research and Public Health*, vol. 14, no. 8, p. 925, 2017.
- [7] K. T. Han, H. K. Jeong, S. S. Do et al., "Forecasting respiratory infectious outbreaks using ED-based syndromic surveillance for febrile ED visits in a Metropolitan City," *The American Journal of Emergency Medicine*, vol. 37, 2019.
- [8] M. A. Daw, "The epidemiology of hepatitis c virus infection in African immigrants; bridging the gaps," *Travel Medicine and Infectious Disease*, vol. 27, 2019.
- [9] S. Selmane, "Dynamic relationship between climate factors and the incidence of cutaneous leishmaniasis in Biskra Province in Algeria," *Annals of Saudi Medicine*, vol. 35, no. 6, pp. 445–449, 2015.
- [10] N. Watts, M. Amann, N. Arnell et al., "The 2018 report of the Lancet countdown on health and climate change: shaping the health of nations for centuries to come," *Lancet*, vol. 392, no. 10163, pp. 2479–2514, 2018.
- [11] G. Arikawa, Y. Fujii, M. Abe et al., "Meteorological factors affecting the risk of transmission of HPAI in Miyazaki, Japan," *Veterinary Record Open*, vol. 6, no. 1, Article ID e000341, 2019.
- [12] N. Raymond, O. Jimmy, S. M. Mwalili et al., "Effect of climate on incidence of respiratory syncytial virus infections in a refugee camp in Kenya: a non-Gaussian time-series analysis," *PLoS One*, vol. 12, no. 6, Article ID e0178323, 2017.
- [13] H. H. Hussien, F. H. Eissa, and K. E. Awadalla, "Statistical methods for predicting malaria incidences using data from Sudan," *Malaria Research and Treatment*, vol. 2017, no. 4, 9 pages, 2017.
- [14] N. J. Aquilina, D. S. J. Maria, B. Stefano et al., "Comparison of machine learning approaches with a general linear model to predict personal exposure to benzene," *Environmental Science & Technology*, vol. 52, 2018.
- [15] X. Wang, Y. Liang, Q. Wang, and Z. Zhang, "Empirical models for tool forces prediction of drag-typed picks based on principal component regression and ridge regression methods," *Tunnelling and Underground Space Technology*, vol. 62, pp. 75–95, 2017.
- [16] W. M. T. Thupeng, T. Mothupi, B. Mokgweetsi et al., "A principal component regression model, for forecasting daily peak ambient ground level ozone concentrations," in *The Presence of Multicollinearity Amongst Precursor Air Pollutants and Local Meteorological Conditions: A Case Study of Maun*, Social Science Electronic Publishing, Rochester, NY, USA, 2018.
- [17] S. Alalawi, S. Abdulwahab, and C. Bakheit, "Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone," *Environmental Modelling & Software*, vol. 23, no. 4, pp. 396–403, 2008.
- [18] R. B. Solanki, H. D. Kulkarni, S. Singh, A. K. Verma, and P. V. Varde, "Optimization of regression model using principal component regression method in passive system reliability assessment," *Progress in Nuclear Energy*, vol. 103, pp. 126–134, 2018.
- [19] G. E. P. Box and G. M. Jenkins, "Time series analysis: forecasting and control," *Journal of Time*, vol. 31, no. 4, p. 303, 2010.
- [20] L. M. Surhone, M. T. Timpledon, S. F. Marseken et al., *Principal Component Regression*, Betascript Publishing, Port Louis, Africa, 2013.

- [21] R. B. Keithley, R. Mark Wightman, and M. L. Heien, "Multivariate concentration determination using principal component regression with residual analysis," *Trac Trends in Analytical Chemistry*, vol. 28, no. 9, pp. 1127–1136, 2009.
- [22] R. Mo and D. M. Straus, "Statistical dynamical seasonal prediction based on principal component regression of GCM ensemble integrations," *Monthly Weather Review*, vol. 130, no. 9, pp. 449–481, 2002.
- [23] S. Eslamian, M. Ghasemizadeh, M. Biabanaki, and M. Talebizadeh, "A principal component regression method for estimating low flow index," *Water Resources Management*, vol. 24, no. 11, pp. 2553–2566, 2010.
- [24] Z. Li, T. Liu, G. Zhu et al., "Dengue Baidu search index data can improve the prediction of local dengue epidemic: a case study in Guangzhou, China," *PLoS Neglected Tropical Diseases*, vol. 11, no. 3, Article ID e0005354, 2017.
- [25] W. J. Kim and C. Y. Lee, "Environmental exposures and chronic obstructive pulmonary disease," *Molecular & Cellular Toxicology*, vol. 13, no. 3, pp. 251–255, 2017.
- [26] S. Liu, J. Chen, J. Wang et al., "Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: a time-series model based on weather variability," *International Journal of Biometeorology*, vol. 62, 2018.
- [27] D. Zhao, L. Wang, J. Cheng et al., "Impact of weather factors on hand, foot and mouth disease, and its role in short-term incidence trend forecast in Huainan City, Anhui Province," *International Journal of Biometeorology*, vol. 61, no. 3, pp. 1–9, 2016.
- [28] Z. Du, L. Xu, W. Zhang et al., "Predicting the hand, foot, and mouth disease incidence using search engine query data and climate variables: an ecological study in Guangdong, China," *BMJ Open*, vol. 7, no. 10, Article ID e016263, 2017.
- [29] T. M. Carvajal, K. M. Viacrusis, L. F. T. Hernandez et al., "Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines," *BMC Infectious Diseases*, vol. 18, no. 1, p. 183, 2018.
- [30] J. Gibson, "Air pollution, climate change, and health," *The Lancet Oncology*, vol. 16, no. 6, p. e269, 2015.
- [31] D. O. Lipp, "Climate change, air pollution, and infectious diseases: a new epidemiological scenario in Argentina," in *Climate Change and Air Pollution*, Springer, Berlin, Germany, 2018.