Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 2, No. 3, 2022

Prediction Modeling and Research on the Relationship Between Urban Air Pollutants and Respiratory Diseases

Yang Zhu[1, a]

[1]School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou 730070, China
[a]1466627344@qq.com

Abstract: In recent years, with the development of urbanization, the use of traditional fuels such as oil and coal is increasing, and air pollution is also becoming increasingly serious. In recent years, people have paid more attention to health issues, and the relationship between air pollution and health has gradually become a research hotspot. Based on the data of air pollutant concentration and respiratory diseases in Shijiazhuang, China, this paper analyzes the correlation between air pollutants and respiratory diseases, and finds that a variety of air pollutants will increase the prevalence of acute respiratory diseases, influenza and pneumonia, and acute upper respiratory infection. Then, this paper uses ARIMA model to predict the data of six air pollutants, and uses ridge regression model, Using the predicted air pollutant data, the number of respiratory diseases in urban population is predicted. Finally, this paper provides suggestions on how to prevent diseases for urban residents in the future.

Keywords: Air pollutants, Respiratory diseases, Prediction, Arima.

1. Introduction

In recent years, with the rapid development of cities, China's urban population has become more and more dense. With the development of urbanization, the use of traditional fuels such as oil and coal has also increased. Although China has taken the development of new materials and new energy as a strategic task, clean energy such as wind energy, solar energy and nuclear energy still need to be developed to replace traditional fuels [1]. With the use of traditional fuels, air pollution is also becoming increasingly serious. According to statistics, the average volume of air that an adult breathes every day is about 15 cubic meters [2]. When the concentration of pollutants in the air rises, it will be harmful to human health after inhalation. In recent years, people have paid more attention to health issues, and the relationship between air pollution and health has gradually become a research hotspot.

As the system with the most frequent contact between the human body and the external environment and the largest contact area, when the respiratory system exchanges gas with outdoor air, in addition to inhaling a certain amount of oxygen, particles or harmful gases in the air may also enter different parts of the human respiratory tract and lung tissue with the breathing process, and then induce respiratory symptoms such as cough, dyspnea, nasal congestion and runny nose. With the delay of time, gradually cause pneumonia, asthma and other diseases [3]. In recent years, China has emphasized the theory that "green water and green mountains are golden mountains and silver mountains", and the state is also focusing on environmental issues [4]. When the air quality improves, people's exposure to air pollutants will be reduced correspondingly, and the risk of illness will be greatly reduced. Therefore, under this background, this paper studies and analyzes the relationship between air pollutants and respiratory diseases.

In addition, according to the data of air pollutants and the historical data of hospital morbidity, it is found that the increase of pollutants is often accompanied by the increase of respiratory diseases. Therefore, it is necessary to study and predict the concentration of air pollutants in a city, explore the relationship between air pollution and respiratory diseases, take effective measures in advance to prevent abnormal changes of air pollutants, and encourage hospitals to take effective measures against the number of patients, it will reduce the additional burden on the local medical and health undertakings.

2. Method

2.1. Ridge Regression Model of Respiratory Diseases Under Air Pollutants

Ridge regression is a multivariate data analysis method proposed to solve the problem of multicollinearity in multivariate regression analysis. This method was proposed by American statistician Hoerl in 1970 [5]. When establishing a linear regression model, due to the correlation between independent variables, multiple collinearity problems will occur in the model, making the model parameters unstable. Some commonly used methods will reduce the interpretability of the model to a certain extent. Several studies have shown that ridge regression can compress the estimation of model parameters by calculating the ridge parameter k when there is a multicollinearity problem among multiple independent variables, so that the parameter values approach the real values, thus making the parameter estimation more accurate and realistic.

In ordinary least squares estimation

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1}$$

Among them, when calculating $X^T X$, due to the ill conditioned nature of the data, the result may appear singular matrix or approximate singular matrix. Therefore, in order to prevent this situation, kI is added to ridge regression in the process of finding the inverse matrix. The improved estimate is:

$$\hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y \tag{2}$$

When $k=0$, $E(\hat{\beta}(k))=\beta$. When $k\neq 0$, $\hat{\beta}(k)$ is a biased estimate of β. From this we can get:

$$\hat{\beta}(k)=(X^TX+kI)^{-1}X^TX\hat{\beta} \tag{3}$$

In the above formula, k is the ridge regression constant. In the least square method, when $k=0$ and $k>0$, I is the unit matrix. The decrease of collinearity is accompanied by the increase of k value, but the prediction variance will also increase. The stable k value should ensure the stability of the ridge trace and the small absolute value of itself. When analyzing the correlation between air pollutants and incidence rate of respiratory diseases, ridge regression is applicable to the above cases because of the small number of factors and the existence of multicollinearity. Thus, the value of parameter k is determined by ridge regression analysis, and the regression coefficient of each air pollutant is obtained, and then the regression equation between the number of patients with respiratory diseases and the influencing factors is calculated.

2.2. Prevalence Prediction Based on Time Series

ARIMA model is a time series model, which is widely used in finance, medicine, environmental monitoring and other fields [6]. ARIMA model is also known as differential moving average autoregressive model. Its main role is to mine the change and development rules of data by studying the existing time series data, and use the discovered rules to predict and study its future.

Before modeling ARIMA model, it is first necessary to ensure that the time series is stable. In actual use, wide stationary time is widely used. The conditions that need to be met are: (1) Any $t\in T$, with $EX_t^2<\infty$; (2) Any $t\in T$, with $EX_t^2=\mu$, μ Is a constant; (3) Take any t, s, $k\in T$, and $k+s-t\in T$, have $\gamma(t,s)=\gamma(k,k+s-t)$. Based on the above premise, the structure of ARIMA (p, d, q) model is:

$$\begin{cases}(1-\varphi_1B-\cdots-\varphi_pB^p)(1-B)^dx_t=(1-\theta_1B-\cdots-\theta_qB^q)\varepsilon_t\\ E(\varepsilon_t)=0,\ VAR(\varepsilon_t)=\sigma^2,\ E(\varepsilon_t\varepsilon_s)=0,\ s\neq t\\ E(x_s\varepsilon_t)=0,\ \forall s<t\end{cases} \tag{4}$$

When $d=q=0$ and $p\neq 0$, ARIMA$(p,0,0)$ is an autoregressive AR(p) model; When $p=d=0$ and $q\neq 0$ ARIMA$(0,0,q)$ is a moving average MA(q) model; When $d=0$ and $p\neq 0$, ARIMA$(p,0,q)$ is an autoregressive moving average ARMA(p,q) model.

3. Experiments

3.1. Data Source and Introduction

The respiratory disease data in this paper comes from the first hospital of Shijiazhuang city and the first hospital of Hebei Medical University [7]. The statistical range is the daily average outpatient visits of respiratory diseases from January 2014 to December 2016. The air quality data comes from the China air quality online monitoring and analysis platform. Respiratory diseases are divided into four categories: acute upper respiratory tract infection, acute lower respiratory tract infection, influenza, pneumonia and acute respiratory diseases. Air pollutants include six indicators: fine particulate matter ($PM_{2.5}$) carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), inhalable particulate matter (SO_2) and nitrogen dioxide (NO_2).

3.2. Data Correlation Analysis

Correlation analysis is a method proposed to measure the closeness between two or more variable factors [8]. In this paper, the correlation of six air pollutants is preliminarily judged by the scatter plot matrix, as shown in Figure 1.

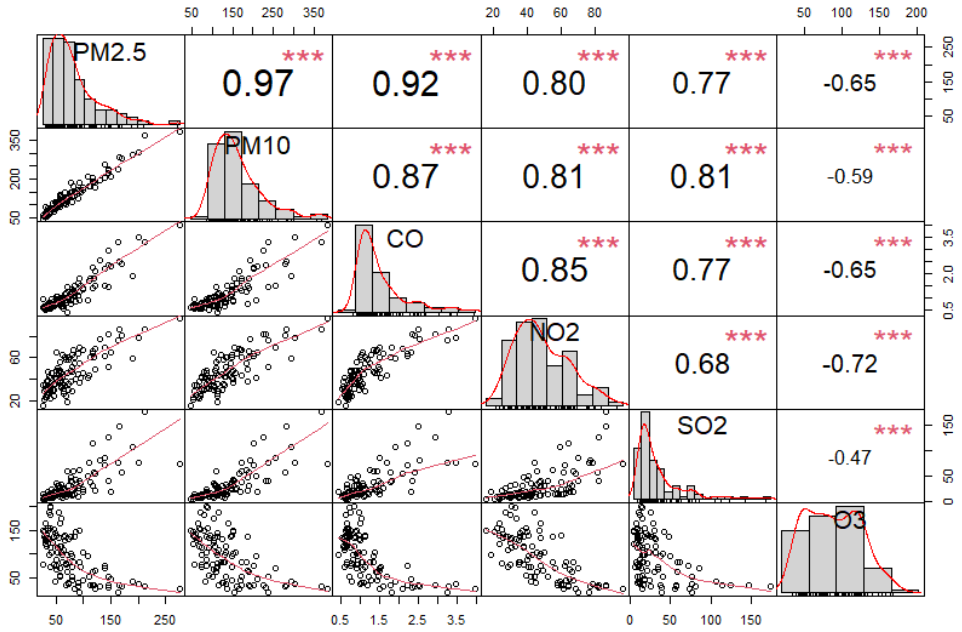

Figure 1. Correlation coefficient results of six indicators

As can be roughly seen from Fig. 1, except for so_ 2 and o_ 3, most of the correlation coefficients between other indicators are greater than 0.7, showing a high correlation, indicating that there may be multiple collinearity between air pollutants.

3.3. Ridge Regression Model of Respiratory Diseases Under Air Pollutants

According to the analysis results of the previous step, Taking Shijiazhuang City, Hebei Province, which is seriously polluted, as an example, the air pollution data is the monthly average value of that month. Respiratory diseases are the average daily outpatient volume of that month. Since the autocorrelation coefficient of some independent variables is found to be high in the correlation analysis, the multicollinearity analysis [9] is carried out for six air pollutants, totaling four models. Because the independent variables of each model are the same, the multicollinearity Vif value is also the same. The results are shown in Table 1:

Table 1. Variance expansion factor test

variable	$PM_{2.5}$	PM_{10}	CO	NO_2	SO_2	O_3
VIF	32.197	24.035	12.74	7.411	3.553	1.472

According to the variance expansion factor test, the VIF values of the independent variables $PM_{2.5}$, PM_{10} and CO are all greater than 10. It can be seen from table 1 that the VIF values of these variables are all high. Although the VIF value of NO_2 is less than 10, it still does not meet the requirements. It shows that there is a serious multicollinearity problem in the above independent variables, so ridge regression method is adopted in this paper for data analysis.

When the value of k is 0.4, the coefficients of each independent variable in the four ridge regression equations gradually tend to be stable. According to the determination principle of k, the variance expansion factor method is used to determine k. when k is equal to 0.188, the ridge regression has the best effect. At this time, the regression coefficients for respiratory diseases are:

Table 2. Ridge regression coefficient

Normalization coefficient	$PM_{2.5}$	PM_{10}	CO	NO_2	SO_2	O_3
Acute respiratory disease	-0.027	-0.2	0.541	0.309	0.137	0.128
Acute lower respiratory tract infection	-0.031	0.045	0.392	0.04	0.277	-0.12
Influenza and pneumonia	-0.014	-0.138	0.553	0.183	0.239	0.074
Acute upper respiratory tract infection	-0.03	-0.255	0.42	0.468	-0.114	0.26

It can be seen from table 2 that in different equations, the standardized coefficient values of some independent variables are small, which indicates that these independent variables have a small impact on the average daily outpatient volume of diseases in the current regression model. And some independent variables, such as $PM_{2.5}$, are always negative, which means that the higher the $PM_{2.5}$, the lower the prevalence rate, which is inconsistent with the reality. Therefore, variable selection is carried out according to the ridge regression variable selection standard and the ridge trace map. To sum up, according to the relationship between diseases and air pollutants obtained after variable deletion and ridge regression analysis:

$$\text{Acute respiratory disease} = 124.571 + 101.616 \times CO + 2.116 \times NO_2 + 0.21 \times SO_2 \quad (5)$$

$$\text{Acute lower respiratory tract infection} = 16.365 + 18.543 \times CO + 0.383 \times SO_2 \quad (6)$$

$$\text{Influenza and pneumonia} = 5.285 + 51.043 \times CO + 0.505 \times NO_2 + 0.49 \times SO_2 \quad (7)$$

$$\text{Acute upper respiratory tract infection} = -2.063 + 28.269 \times CO + 1.821 \times NO_2 + 0.899 \times O_3 \quad (8)$$

Taking acute respiratory diseases as an example, when NO_2 and SO_2 are unchanged, the number of residents seeking medical treatment for acute respiratory diseases will increase by 101.616 on average every mg/m^3 increase in atmospheric CO concentration, but the actual content of CO in the atmosphere has always remained below 5 mg/m^3, and the change range is usually around 0.1 mg/m^3. When SO_2 and CO remain unchanged, every 1 ug/m^3 increase in NO_2 will increase the average number of residents seeking medical treatment for acute respiratory diseases by 2.116. When NO_2 and CO remain unchanged, every 1 ug/m^3 increase in SO_2 will increase the average number of residents seeking medical treatment for acute respiratory diseases by 0.21. The analysis of the other three air pollutants is the same as above and will not be repeated.

3.4. Prediction of Respiratory Disease Prevalence in Shijiazhuang

For the time series of air pollutant $PM_{2.5}$, it is found that it does not have stationarity through direct observation, so the stationarity ADF test is carried out. After the test, P = 0.803 > 0.05 is obtained, and the difference is not statistically significant, which indicates that the $PM_{2.5}$ time series is non-stationary. After the first-order difference processing, the autocorrelation and partial autocorrelation coefficients of the sequence are calculated. By observing the coefficients of ACF and $PACF$, different P and Q combination values are selected for model fitting and modeling. Finally, the best model suitable for $PM_{2.5}$ is selected through comprehensive comparison according to AIC criteria [10].

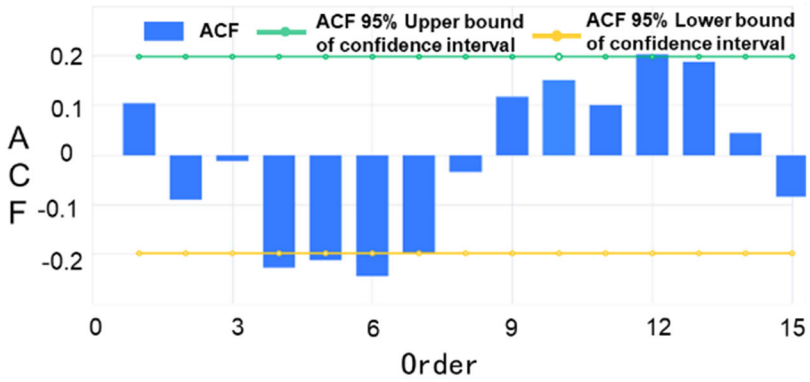

Figure 2 $PM_{2.5}$ autocorrelation graph (ACF)

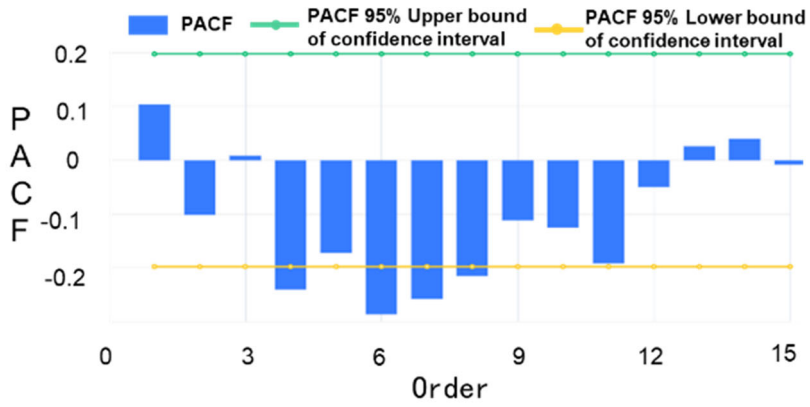

Figure 3 $PM_{2.5}$ partial autocorrelation graph (PACF)

The optimal parameters are automatically found based on AIC information criterion, and the time series model is $ARIMA$ (5,1,2). Next, the validity of the model is diagnosed to judge whether the residual of the model has the property of white noise, that is, whether it has three characteristics of zero mean, normality and independence. The method used is: Q statistic test, which is used to test whether a series of observations in a certain period are random independent observations. When the p value of the test result is less than 0.05, the sequence is generally considered as non white noise sequence, and the results are shown in Table 4.

Table 4. ARIMA model (5,1,2) test table

Q6(p value)	Q12(p value)	AIC
0.154(0.695)	3.963(0.682)	945.44

From the analysis of the results of Q statistics, it can be concluded that Q6 is not significant horizontally, and the assumption that the residual of the model is a white noise sequence cannot be rejected, and the model meets the requirement that the residual is a white noise. The predicted results of air pollutant $PM_{2.5}$ in 12 months from March 2022 to February 2023 are shown in Figure 4.

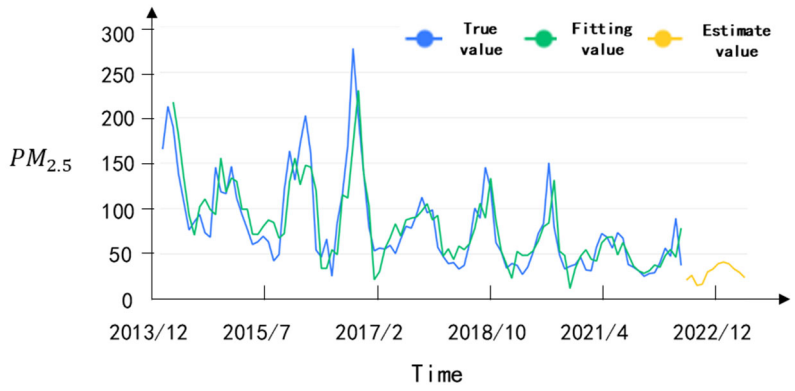

Figure 4. ARIMA model fitting and prediction of $PM_{2.5}$

It can be seen from Fig. 4 that $PM_{2.5}$ has generally shown a downward trend since 2013, with the peak from December of each year to February of the next year. In the next year from March 2022 to February 2023, the forecast of $PM_{2.5}$ will remain below 50 ug/m_3 , which shows that the environmental department has made certain achievements in maintaining air quality. Then, according to the above analysis process, the five air pollutants PM_{10}, CO, NO_2, SO_2 and O_3 are predicted in turn. The ARIMA prediction model is:

Table 5. Model inspection table

Air pollutants	ARIMA	Q6(p value)	Q12(p value)	AIC
PM_{10}	ARIMA(7,1,5)	0.174(0.677)	1.346(0.969)	1004.714
CO	ARIMA(8,1,0)	0.629(0.428)	2.946(0.816)	159.911
NO_2	ARIMA(8,1,0)	0.903(0.342)	2.798(0.834)	776.75
SO_2	ARIMA(7,2,1)	0.029(0.865)	2.626(0.854)	811.744
O_3	ARIMA(2,1,3)	3.638(0.056)	5.55(0.475)	865.322

Based on the above prediction data, the data of air pollutants in the next year are substituted into the ridge regression model of respiratory diseases, and the prediction is made according to the prediction results and air concentration. The visualization results are shown in Figure 5 to Figure 8.

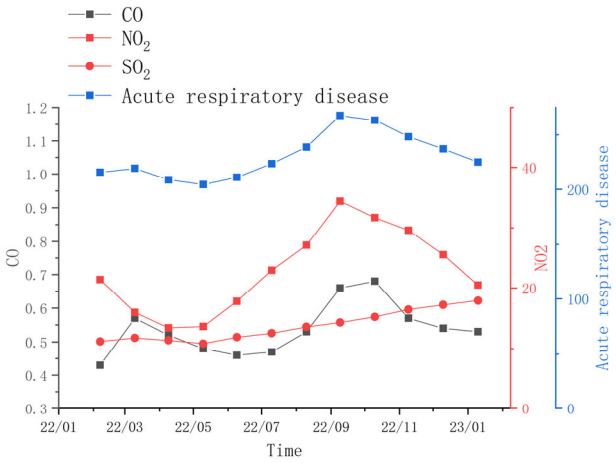

Figure 5. Forecast results of outpatient volume of acute respiratory disease

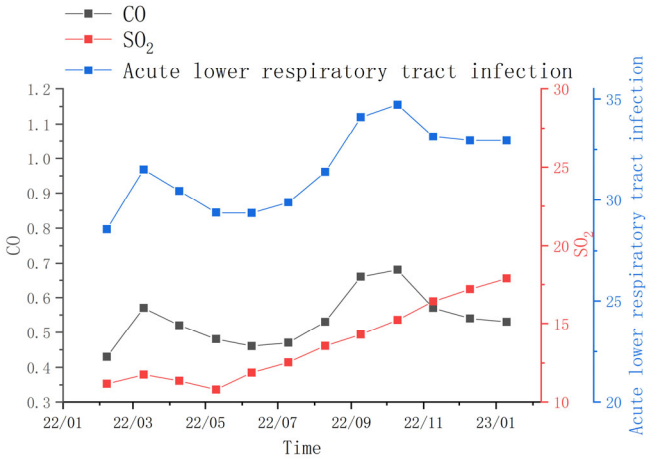

Figure 6. Forecast results of outpatient volume of acute lower respiratory infection

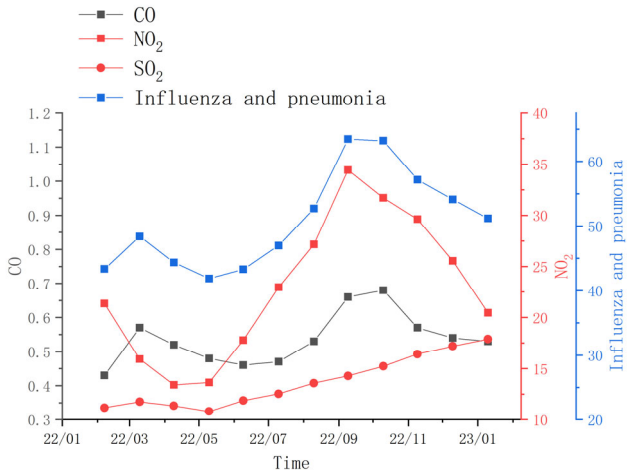

Figure 7. Forecast results of outpatient volume of influenza and pneumonia

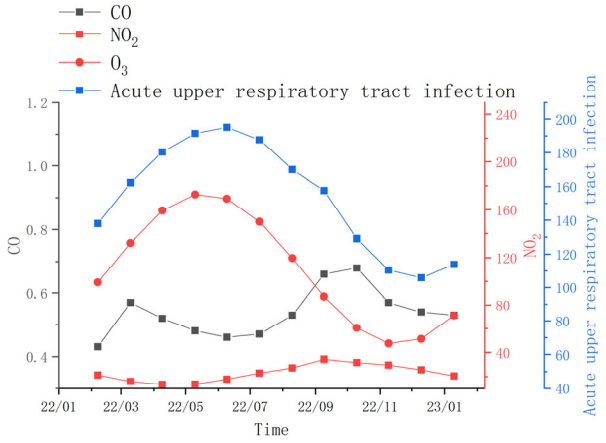

Figure 8. Forecast results of outpatient volume of acute upper respiratory tract infection

According to the ridge regression model of acute respiratory disease, acute lower respiratory infection, influenza and pneumonia, it is shown that the pollution degree of CO, NO_2, SO_2 and O_3 is the most core pathogenic factor causing respiratory diseases. In particular, when CO pollution is serious, it can be suggested that the relevant population exposed to the outdoor for a long time. Pay attention to the possibility of inducing respiratory diseases, and take certain defensive measures. For acute upper respiratory tract infection, the main influencing factor is O_3. Ozone, as a strong oxidant, can induce the peroxidation of unsaturated fatty acids in human respiratory epithelial cells, thus causing oxidative stress damage to lung respiratory epithelial cells. Through time series analysis, it is found that O_3 has shown a small upward trend in recent years. Therefore, the environment and transportation departments should control the number of urban motor vehicles, improve the emission standards of vehicles on the road, and take measures such as real-time traffic control in local areas to reduce the generation of O_3 from the source.

4. Conclusion

In this paper, Shijiazhuang, a representative city with poor air quality, was used to analyze different air pollutants and respiratory diseases. Through the analysis of the model, it is found that CO has a very significant pathogenic effect on four respiratory diseases and is a significant pathogenic factor. In addition, the three air pollutants above NO_2, SO_2 and O_3 will increase the probability of corresponding types of diseases. For example, NO2 will increase the probability of acute respiratory disease, influenza, pneumonia and acute upper respiratory infection, SO_2 will increase the probability of acute respiratory disease, influenza, pneumonia and acute lower respiratory infection, and O_3 will increase the probability of acute upper respiratory infection.

In addition, by using the time series model, this paper also predicted the monthly average concentration of six air pollutants in Shijiazhuang from March 2022 to February 2023, and the predicted air data in the next year. The peaks of the average daily outpatient volume of the number of people seeking medical treatment for acute respiratory diseases, acute lower respiratory tract infections, and influenza and pneumonia are in April 2022 and November 2022 to January 2023. And the incidence rate of diseases in winter is significantly higher than that in summer. The peak of the daily average outpatient volume of acute upper respiratory tract infection occurs from June to August 2022. Local governments should pay attention to the work of ozone pollution control, improve the ozone monitoring and early warning system, and inform the public how to deal with ozone pollution, so that ozone prevention and control can be implemented practically.

References

[1] Xu B, Lin B. Do we really understand the development of China's new energy industry?[J]. Energy economics, 2018, 74: 733-745.

[2] Brown J S, Gordon T, Price O, et al. Thoracic and respirable particle definitions for human health risk assessment[J]. Particle and fibre toxicology, 2013, 10(1): 1-12.

[3] Liu J, Yin H, Tang X, et al. Transition in air pollution, disease burden and health cost in China: A comparative study of long-term and short-term exposure[J]. Environmental Pollution, 2021, 277: 116770.

[4] Feng S. Analysis on the Effect of Green Finance on Sustainable Development[J]. International Journal of Science, 2021, 8(9).

[5] Hoerl A E, Kennard R W. Ridge regression: applications to nonorthogonal problems[J]. Technometrics, 1970, 12(1): 69-82.

[6] Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset[J]. Data in brief, 2020, 29: 105340.

[7] Xu Jian Study on the correlation between outpatient visits and air pollution in Tianjin [D]. University of science and technology of China, 2011

[8] Dai Y H, Zhou W X. Temporal and spatial correlation patterns of air pollutants in Chinese cities[J]. PloS one, 2017, 12(8): e0182724.

[9] Akinwande M O, Dikko H G, Samson A. Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis[J]. Open Journal of Statistics, 2015, 5(07): 754.

[10] Mondal P, Shit L, Goswami S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices[J]. International Journal of Computer Science, Engineering and Applications, 2014, 4(2): 13.