International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 5, No. 1, 2024

Prediction Analysis of the Number of Patients with Respiratory Diseases based on SVR

Xiaotian Ma [1], Yinghua Li [2, *]

[1] School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China
[2] School of Resource and Environmental Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China

*** Corresponding author:** Yinghua Li

Abstract: The rapid development of industrialization is accompanied by a further increase in air pollution. Serious air pollution will bring a variety of diseases to the human body, which has been a basic consensus on a global scale. It is of great significance to further explore the specific diseases related to air pollution. In this paper, air pollutant emission data and respiratory disease data of public data sets are collected to analyze the changes between them and the relationship between them. Further, the SVR machine learning analysis model was established to analyze the impact of air pollutant emission on the number of patients with respiratory system diseases. Meanwhile, the data of the number of patients with respiratory system diseases was predicted based on the data of air pollutant emission, and the experimental effect was satisfactory. It lays a foundation for further research on the relationship between various air pollutants and human respiratory diseases.

Keywords: Air Pollution; SVR; Respiratory Disease; Smoke and Dust Emission.

1. Introduction

With the continuous improvement of China's industrial development level, air pollution problems are also accompanied. Studies in countries around the world have shown that air pollution causes a high incidence of human respiratory diseases [1].The emission of all kinds of air harmful air pollutants is very harmful to various organs of the human body, which will lead to various human diseases, such as respiratory diseases, cardiovascular diseases and various cancers [2]. All kinds of air pollutants discharged into the atmosphere in the industrial production process, when people breathe into the body, although some of the pollutants will not immediately make the human body sick, but these pollutants enter the human body, along with the blood vessels in the lungs into the human circulatory system, and then slowly form a variety of potential disease threats to the human body [3]. Excessive air pollution and PM2.5 exposure can have adverse effects on both pregnant women and their fetuses, increasing the risk of ADHD in the fetus or newborn baby [4]. Short-term exposure to air containing PM2.5, a common atmospheric pollutant, resulted in significant changes in urinary metabolic data, which further suggests that exposed individuals are at risk of developing urinary infections from inhaling air containing PM2.5 [5]. When polluting gases or dust substances are discharged into the atmosphere, for people who work outdoors for a long time, they will increase the risk of skin diseases caused by air pollution. The pollutants in the air will cause changes in our skin metabolism, which may cause a series of skin diseases or inflammation [6]. To sum up, it is necessary to further study the effects of air pollutants on human health. In recent years, with the development of machine learning methods, more and more researchers have begun to use various kinds of machine learning methods to analyze the relationship and influence between air pollutants and various diseases of the human body from various angles. Razavi-Termeh et al. [7] established a support vector machine regression model to analyze the relationship between air pollution and asthma in Tehran, Iran and its influencing factors, and obtained good research results. Ravindra [8] et al. established and analyzed the relationship between different air pollutants and acute respiratory diseases by using machine learning methods. Through research, they found that the application of machine learning methods to analyze related problems has considerable potential and research value. Yang [9] et al. analyzed the correlation between meteorological factors and various air pollution factors on respiratory diseases, and further established support vector machine regression model to study the impact of meteorological factors and air pollution on the incidence of respiratory diseases, and found that support vector machine regression model has certain analytical potential in this aspect.

Machine learning method has become an effective research and analysis method in analyzing the relationship between various air pollution factors and various diseases in the human body, and it also shows the research potential of various machine learning methods in different fields. Therefore, in this study, support vector machine (SVM) machine learning method will be adopted to establish a regression model to analyze the relationship and influence between air pollutant emission and the number of patients with respiratory diseases in the entire mainland of China over the years, so as to provide research direction and data analysis basis for further preventing the outbreak of large-scale respiratory diseases caused by air pollution in advance.

2. Methods

2.1. Data Set

In the research covered in this paper, The main data falls into two broad categories, both of which are publicly available data sets, One is the statistical data of pollutant emissions in the exhaust emissions of Chinese Mainland over the years; The other category is the statistics of the total number of people suffering from respiratory and tuberculosis diseases in Chinese Mainland over the years. First of all, the statistical data of pollutant emissions in the exhaust emissions

of Chinese Mainland over the years are from the National Statistical Yearbook issued by the China National Bureau of Statistics over the years [10]. The statistical data on the emission of various pollutants in the dataset covers the period from 1998 to 2019, and a total of 21 years of pollutant emission data are collected, mainly including the annual total emission data of sulfur dioxide and smoke (powder) dust in the past years.

Another public data set comes from the Institute for Health Metrics and Evaluation (IHME) [11], which has statistical data on the total number of patients with respiratory diseases and tuberculosis diseases in Chinese Mainland for 21 years in total from 1998 to 2019. This public data set comes from the Institute for Health Metrics and Evaluation in Seattle, the United States, Its public dataset website can retrieve research data on disease burden worldwide [11]. This study mainly used the research data on the burden of respiratory diseases over the years in Chinese Mainland as a whole.

2.2. Research Method

This study mainly uses the support vector machine regression model in machine learning to model the research data. Support vector machine was first proposed by Vapnik [12] and his colleagues in 1995, and then Drucker H [13] proposed support vector regression machine based on the support vector theory proposed by Vapnik in 1997. Since then, support vector machine theory can be used to solve various classification problems and related regression prediction problems.

Support vector machine (SVM) is a supervised generalized linear classifier for binary classification of data. The whole process of SVM can be converted into a quadratic convex optimization problem. When dealing with classification and regression optimization problems, if the whole problem is linearly separable, then we will find the optimal hyperplane which can separate the categories in the same dimensional space; If the problem model is nonlinear, then we will choose to add relaxation variables and introduce nonlinear mapping to map the low-dimensional input space to the high-dimensional space, and then transform the problem to be linearly separable in the high-dimensional space, and then find the optimal hyperplane in the high-dimensional space. Based on this principle, support vector machine method can solve the general classification and regression prediction and other related problems.

2.3. Evaluation Index

In this study, the mean squared error (MSE), and the coefficient of determination (R2), the mean absolute error (MAE) was selected as the evaluation indexes of the final model effect. Their specific calculation formula is shown in equation (1)-(3):

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y}_i)^2} \tag{2}$$

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{3}$$

MSE is a method to measure the error between the predicted value and the actual value of the model. It is more sensitive to outliers, and the value range is generally positive. The closer the final value is to 0, the better the fitting effect of the model is. R^2, known as the coefficient of determination, is a measure of a model's ability to explain the variability of dependent variables. The value is between 0 and 1, and the closer the final result is to 1, the better the final fitting effect of the model is, that is, the model can explain most of the dependent variable variation. MAE is an average measure of the error between the predicted value and the actual value. The sensitivity to outliers is generally low, and a positive number is usually taken. The closer the final value is to 0, the better the model fitting effect is.

3. Results and Discussion

3.1. Data Feature Analysis

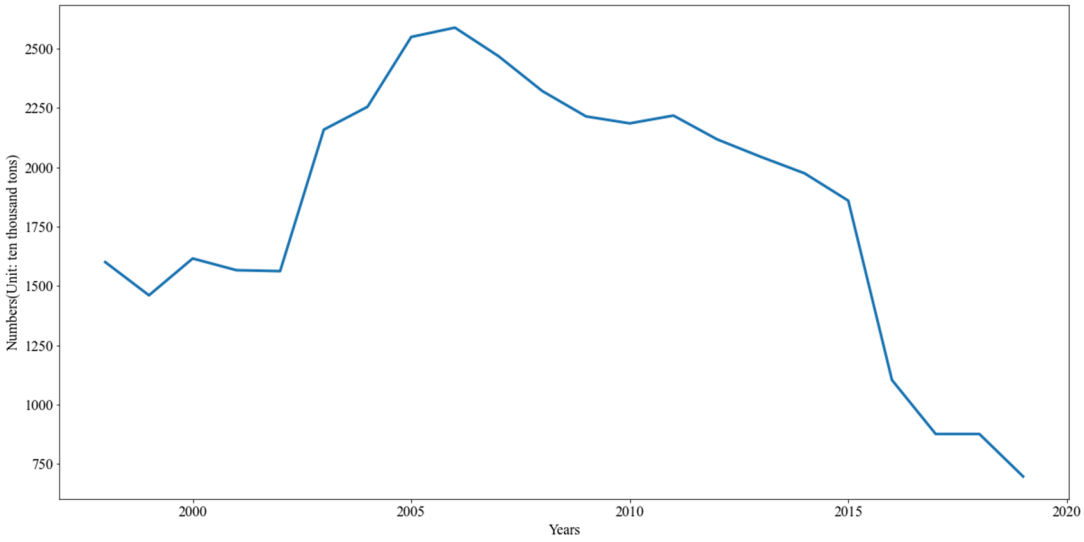

Figure 1. Numbers of SO_2 emissions over the past 21 years

The data involved in this study spans from 1998 to 2019, a total of 21 years. In the process of building the model, the basic data characteristics and the overall trend of the data were firstly sorted out for all kinds of data. For the air pollutant emission data over the past years, which will be input into the model as the characteristic data, mainly including the annual emission data of SO_2 and smoke (powder) dust, the trend of the main data in the past 21 years is shown in Figure 1 and Figure 2. From the overall data trend of Figure 1 and Figure 2, it can be clearly found that with the passage

of time and the continuous recommendation and improvement of the development process of China's industrial system, the exhaust gas emissions in the early years of China showed an increasing trend of realization year by year. After reaching a certain peak, with the continuous management and improvement of environmental protection related issues by national policies, At the same time, with the maturity of the national industrial system and the continuous promotion of industrial upgrading and other related policies in recent years, the emissions of various types of air pollutants have maintained a good overall downward trend in recent years and have been at a relatively low level, which can be seen that China's environmental protection policies in recent years have played a very important role and effect on the improvement of the overall natural environment in China.

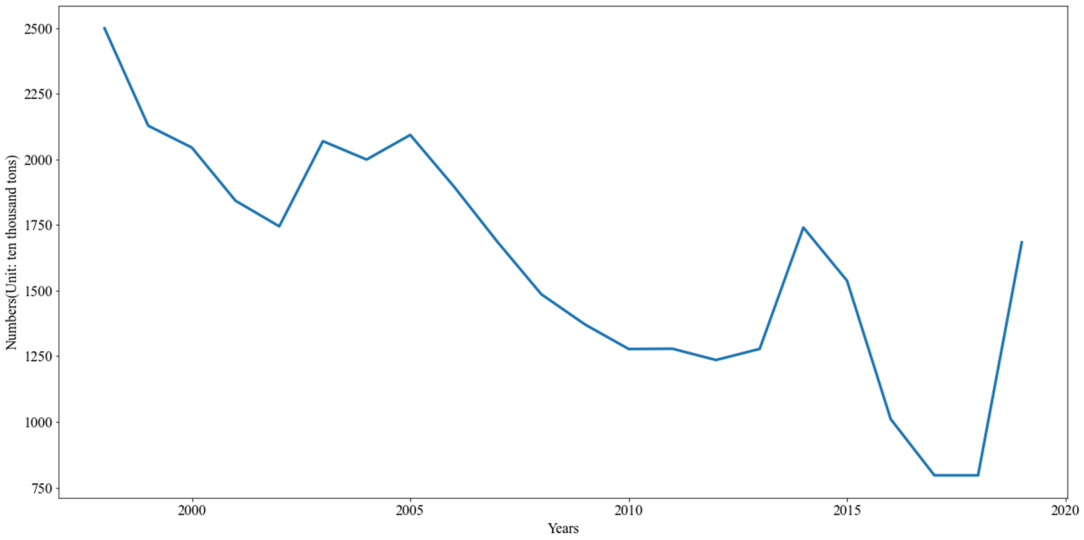

Figure 2. Numbers of smoke (dust) emission over the past 21 years

At the same time, as the data of the number of patients with respiratory diseases in Chinese Mainland over the years, which was finally set as the prediction target, the main data change trend over time is shown in Figure 3. From Figure 3, it can be clearly observed that from 1998 to 2019, the number of people suffering from respiratory diseases in Chinese Mainland is decreasing year by year, and there is no fluctuation in the number of people suffering from respiratory diseases year by year. At the same time, it can be seen that before 2005, the annual reduction rate of the number of patients was relatively fast, but after 2005, the reduction rate began to slow down year by year, but the overall trend is still the number of patients decreasing year by year. The overall change trend of the number of respiratory diseases in Chinese Mainland over the years is basically consistent with the overall change trend of air pollutant emissions, showing a relatively obvious downward trend.

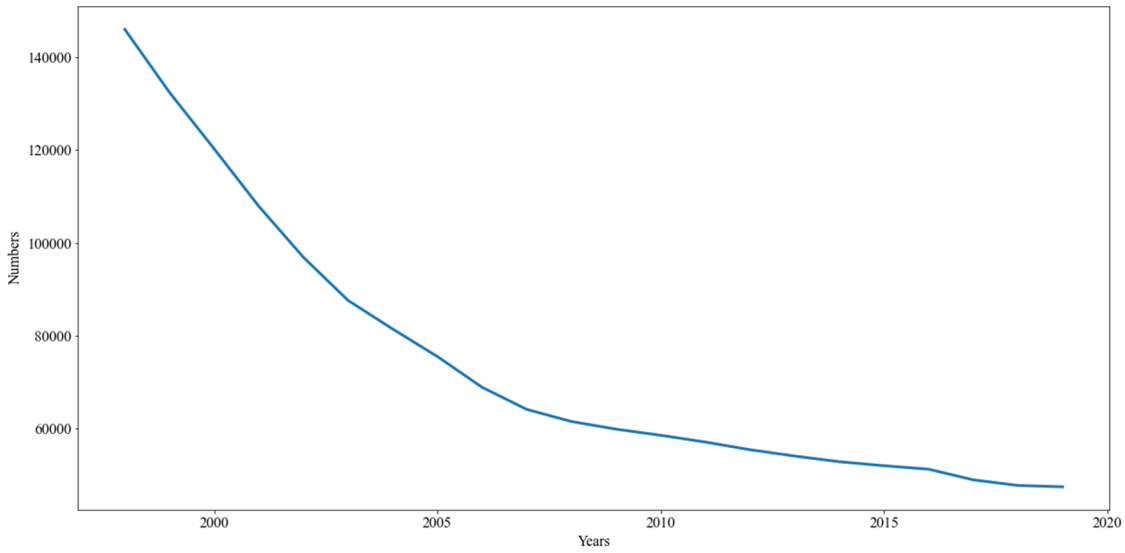

Figure 3. The change trend of the number of patients with respiratory diseases over the past years

3.2. Model Results Analysis

In this study, it is necessary to standardize experimental data before establishing the model. For a model such as support vector machine, it is very necessary to standardize before inputting characteristic data into the model. The specific mathematical expression of standardization processing is shown in equation (4).

$$X' = \frac{X-\mu}{\sigma} \tag{4}$$

Where X' represents the eigenvector after normalization, X represents the original eigenvector, μ represents the mean value of each original eigenvector, and σ represents the standard deviation of each original eigenvector.

In the process of SVR model construction, grid search method [14] was adopted in this study to determine the kernel function kernel, penalty parameter C and kernel function parameter gamma involved in the process of model construction. In this study, the final search results show that when the kernel function is Gaussian (RBF), the penalty parameter C is 2, and the kernel function parameter gamma is scale, the overall performance of the model is the best. In the final model verification stage, the study adopts the method of 10-fold verification to evaluate the overall prediction performance of the model. Table 1 shows the prediction effect of the SVR model constructed in this study under the relevant evaluation indicators. Where MSE=0.22, R^2 =0.7063, MAE=0.3734. Figure 4 shows the fitting between the predicted value of the final model and the test value in the actual data set. It can be seen that there is a good fitting effect between the predicted value of the model and the actual test value, except for some extreme values that cannot be thoroughly fitted. From the final results, SVR model has a good application effect in analyzing the relationship between air pollution emissions and the number of respiratory diseases and forecasting.

Table 1. SVR model prediction results

Model	Assessment criteria		
	MSE	R^2	MAE
SVR	0.22	0.7063	0.3734

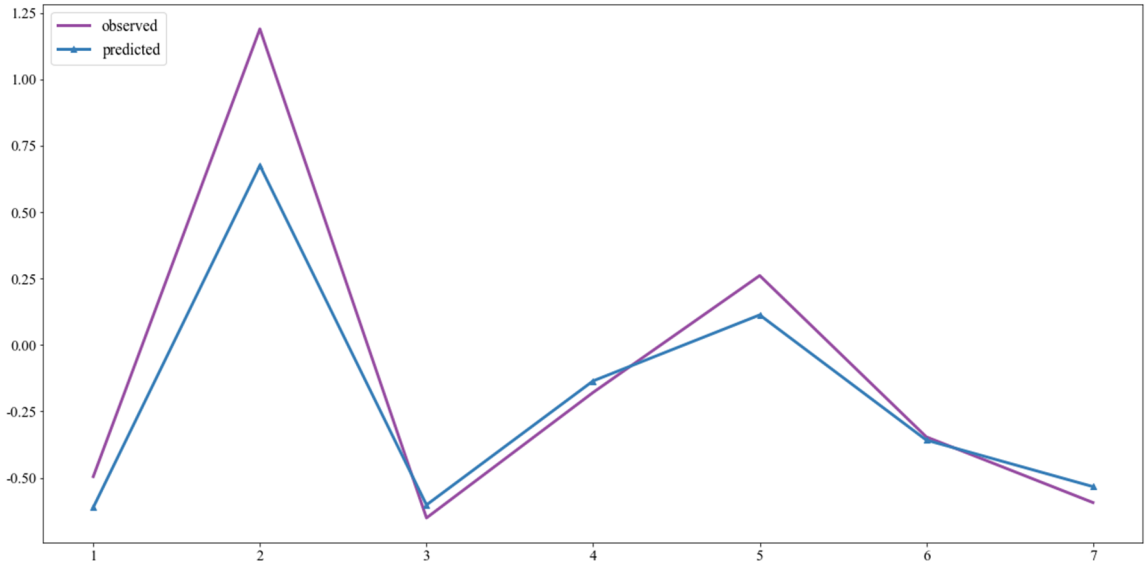

Figure 4. Fitting diagram of SVR model prediction effect

3.3. Discussion

Serious air pollution will cause various degrees of damage and persecution to various organs of the human body, and even threaten people's life and health. Therefore, it has become very important and urgent to further explore the relationship and influence between various air pollutants and various diseases of the human body. In order to facilitate relevant departments and enterprises to carry out more efficient and more targeted air pollution prevention and control work. Based on the above situation, this study collected the data of pollutant emissions read from Chinese Mainland during the 21 years from 1998 to 2019 and the number of patients with respiratory diseases over the years, and carried out further data processing and data analysis. It was found that the pollutant emissions in Chinese Mainland showed a trend of first increasing and then decreasing year by year, but the overall change trend was toward the direction of decreasing pollutant emissions year by year, There is a certain similarity between the same period and its changing patterns, as well as a decreasing trend in the number of respiratory disease patients over the years. At the same time, on the basis of basic data analysis, the prediction model of the number of patients with respiratory system diseases based on SVR is further established, and the data of the number of patients with respiratory system diseases is further calculated and predicted according to the data of air pollutant discharge. The final model has obtained reasonable and ideal prediction results, and established a good research direction and theoretical and experimental basis for further detailed study of the relationship and influencing factors between the two. At the same time, it also provides a reference direction for relevant departments to study the relationship between them.

4. Conclusion

By collecting and sorting out the air pollutant emission data and respiratory disease data over the years in Chinese Mainland, and sorting out and analyzing the relevant data, the following basic conclusions are obtained:

(1) As can be seen from the basic change trend of the data, the air pollutant emission data shows a trend of first increasing and then decreasing year by year with the change of the year. The overall change law of this trend is basically consistent with the approach of China's industrial development, and also reflects that the rapid development of the industrial system will inevitably be accompanied by the worsening of air pollution to a certain extent. However, from the changes in the data in recent years, it is clear that the emissions of air pollutants are decreasing year by year, and the air quality is also improving year by year.

(2) As can be seen from the overall change trend of the number of patients with respiratory diseases, as time goes by, the number of patients shows a decreasing trend year by year, and the change trend of the overall data is basically consistent with the change trend of air pollutant discharge. From the data changes, it can be intuitively seen that there is a certain correlation between air pollutant emissions and the number of people suffering from respiratory diseases.

(3) Further, through the establishment of the prediction and analysis model of the number of patients with respiratory diseases, the correlation between air pollutant discharge and the number of patients with respiratory diseases and the interaction between them were better interpreted. Finally, through continuous optimization of a series of relevant parameters of the model, the ideal experimental results are achieved. Under the ten-fold cross-validation model verification criteria, the final model achieved MSE=0.22, R^2=0.7063 and MAE=0.3734 error effects.

Through the above conclusions and a series of data analysis and research basis carried out in this paper, we can better explore the specific impact of air pollutant emission on the number of patients with respiratory diseases, and provide new ideas and research direction for the analysis of the impact of air pollution on other diseases of the human body.

References

[1] Qiang Z ,Xujia J ,Dan T , et al.Transboundary health impacts of transported global air pollution and international trade.[J]. Nature,2017,543(7647):705-709.

[2] Landrigan J P ,Fuller R ,Acosta R J N , et al.The Lancet Commission on pollution and health[J].The Lancet,2018, 391 (10119):462-512.

[3] D R B ,Sanjay R ,Arden C P , et al.Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. [J]. Circulation, 2010,121(21):2331-78.

[4] Chang Y C, Chen W T, Su S H, et al. PM2. 5 exposure and incident attention-deficit/hyperactivity disorder during the prenatal and postnatal periods: A birth cohort study[J]. Environmental Research, 2022, 214: 113769.

[5] Chen C, Li H, Niu Y, et al. Impact of short-term exposure to fine particulate matter air pollution on urinary metabolome: a randomized, double-blind, crossover trial[J]. Environment international, 2019, 130: 104878.

[6] Araviiskaia E, Berardesca E, Bieber T, et al. The impact of airborne pollution on skin[J]. Journal of the European Academy of Dermatology and Venereology, 2019, 33(8): 1496-1505.

[7] Razavi-Termeh S V, Sadeghi-Niaraki A, Choi S M. Spatio-temporal modelling of asthma-prone areas using a machine learning optimized with metaheuristic algorithms[J]. Geocarto International, 2022, 37(25): 9917-9942.

[8] Ravindra K, Bahadur S S, Katoch V, et al. Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections[J]. Science of The Total Environment, 2023, 858: 159509.

[9] Yang J, Xu X, Ma X, et al. Application of machine learning to predict hospital visits for respiratory diseases using meteorological and air pollution factors in Linyi, China[J]. Environmental Science and Pollution Research, 2023, 30(38): 88431-88443.

[10] National Bureau of Statistics. China Statistical Yearbook[M]. Zhong guo tong ji chu ban she, 1998-2019.

[11] Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (GBD 2019) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2020. Available from https://vizhub. healthdata. org/gbd-results/.

[12] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20: 273-297.

[13] Drucker H, Burges C J, Kaufman L, et al. Support vector regression machines[J]. Advances in neural information processing systems, 1996, 9.

[14] Syarif I, Prugel-Bennett A, Wills G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2016, 14(4): 1502-1509.