Trends in Social Sciences and Humanities Research
ISSN: 2959-9938
DOI: 10.61784/tsshr3004

ANALYSIS AND PREDICTION OF AIR QUALITY INFLUENCE FACTORS IN CHANGSHA CITY

WenHui Zeng
School of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Email: 2541290358@qq. com

Abstract: With the continuous advancement of modernization, the problem of air quality is becoming more and more serious. In this paper, LASSO regression analysis is used to screen out some variables that have little impact on the research object, so as to construct the variable system of the research question. Then the multivariate linear regression analysis is carried out according to the variable system. Finally, the main pollutants affecting the air quality of Changsha City are analyzed, which can effectively predict the air quality of Changsha City and take corresponding measures to improve the air quality.
Keywords: LASSO regression; Multiple linear regression; Air quality; Influence factor

1 INTRODUCTION

With the continuous advancement of China's industrialization process and rapid economic development, while bringing a lot of convenience to people's lives, it has also caused a lot of environmental problems, among which air pollution is especially prominent[1-2]. The economy of our country has entered a period of high-quality development, while ensuring the economic development, we must also pay attention to the protection of the environment[3-4]. Today's air quality assessment standards are too simple, only by calculating several air pollutant index to determine the quality of air quality is good or bad, in the current complex air pollution, it is obviously not enough[5]. By constructing a more comprehensive air quality evaluation system and improving analysis methods, this paper can better evaluate air quality, so as to grasp the key points affecting air quality and carry out atmospheric protection more quickly and efficiently[6].

2 DATA AND METHODS

2.1 Data Collection and Source

2.1.1 Data source
National bureau of statistics;
China statistics yearbook;
China environmental air monitoring analysis platform;
China meteorological administration.

2.1.2 Data content
1) Air quality monitoring data
The daily data set of air quality in changsha city is from September 2021 to November 2022. The data summary of the atmosphere daily data is also included in China's latest concentration data, which records the data of typical atmospheric pollutants such as $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2 and volatile product O_3.
2) Historical weather forecast data
Changsha Weather Forecast Daily data set, the data span from September 2021 to November 2022. The daily data set includes meteorological elements such as air temperature (maximum and minimum temperatures), weather, wind direction, and wind power.

2.2 Data Processing

Data symbols are described in Table 1:

Table 1 Description of symbols

Symbol	Explain
y	AQI
x_1	$PM_{2.5}$
x_2	PM_{10}
x_3	SO_2

x_4	CO
x_5	NO_2
x_6	O_3
x_7	minat
x_8	wc
x_9	wd
x_{10}	wp

2.3 Variable Selection and Prediction Methods

2.3.1 Lasso filter variable method

The steps to preliminarily screen variables using Lasso are as follows:
1) Feature standardization.
2) Lasso linear regression model was established.
3) Select the best adjustment parameters.
4) Screen out important variables.

2.3.2 Multiple linear regression analysis

The so-called regression analysis is to analyze the relationship between the dependent variable and the independent variable according to some indicators, and compare the analyzed dependent variable with the true value, and judge the pros and cons of the regression according to the error or goodness of fit between them. It can also study the changes of the dependent variable according to the independent variable obtained from the provided data, which can be used for predictive analysis. There are many factors affecting air quality in this paper, so multiple linear regression is selected. Construct random variables y and independent variables $x_1, x_2, ..., x_p$, between the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

In the equation, $\beta_0, \beta_1, ..., \beta_p$ is unknown parameter, β_0 called the regression constant term, $\beta_1, ..., \beta_p$, called the regression coefficient.

3 RESULT AND ANALYSIS

3.1 Lasso Model Screening Variable Analysis

Table 2 Results of LASSO regression coefficients

The characteristics of LASSO regression	LASSO regression coefficient
$PM_{2.5}$	1. 194
PM_{10}	-0. 027
SO_2	-0. 817
CO	-0. 631
NO_2	-0. 060
O_3	0. 390
maxat	0
minat	0. 188
wc	-1. 475
wd	-0. 395
wp	-0. 940

As can be seen from the results in Table 2 above, the regression coefficient of LASSO regression model is returned. It can be seen that only maxat is compressed to 0 among the regression coefficients of various features. This suggests that maxat has no significant effect on air quality. Among the features that have a significant impact on air quality, the features that have a positive impact on air quality are: $PM_{2.5}$, O_3 and minat, totaling 3 features; the features that have a negative impact on air quality are: PM_{10}, SO_2, CO, NO_2, wc, wd and wp, totaling 7 features.

3.2 Multiple Linear Regression Modeling

3.2.1 Variable selection

The samples were selected by LASSO variables based on AQI data of time series and $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3,

minat, wc and wp.

3.2.2 Model building

This study conducted computational modeling based on R language software, and selected data from September 2021 to November 2022. According to the output results of R language, the multiple linear regression model of air quality is as follows:

$$y = 2.981 + 1.201x_1 - 0.077x_2 - 0.967x_3 - 0.724x_4 + 0.006x_5 + 0.370x_6 + 0.265x_7 - 1.361x_8 - 0.188x_9 - 0.888x_{10}$$

3.2.3 Model test

For the multiple linear regression model of this equation, the linear goodness test of the linear fitting model, the sum of the significance test and (F test) of the equation's overall multilinear regression model, the sum of the significance test of the variable linear regression model and (t test) were respectively analyzed. The statistical significance of the results is as follows.

1) Goodness of fit test

From the goodness of fit of the equation, R=0. 9541, R^2=0. 9104, and nearly equal to 1, indicating that the fit is good, and the independent variable can effectively explain 91. 04% of the change of the dependent variable.

2) Significance test of the overall linearity of the equation (F test)

The probability value of F statistics is 0. 00, because 0. 00<0. 01, when the independent variable is introduced, the possibility of its significance is much smaller than 0. 01, so we can well exclude the original assumption of global regression coefficient 0. 0, indicating that there is an obvious linear relationship between the independent variable and the dependent variable.

3.4.3 Prediction of multiple linear regression model

The data from December 1 to December 5, 2022 are predicted according to the regression equation model, as shown in Table 3.

Table 3 Comparison of predicted and actual values

Date	True value	Multiple linear regression predicted value	Multivariate forecast relative error /%
2022/12/1	34	33. 0315	0. 97
2022/12/2	43	42. 5354	0. 99
2022/12/3	57	59. 1668	1. 04
2022/12/4	77	73. 1888	0. 95
2022/12/5	63	63. 7064	1. 01

As can be seen from Table 1, there is little difference between the predicted value of multiple linear regression and the true value, indicating that the prediction accuracy of multiple linear regression model is high.

4 DISCUSSION

In this paper, 10 variables with significant impact on air quality in Changsha City were selected through LASSO regression analysis, and an index system for evaluating air quality in Changsha City was established. Due to the accuracy of LASSO regression analysis, this evaluation system can provide a good basis for the next model construction. Then, through the introduction and establishment of multiple linear regression model, combined with the example of Changsha City air quality, the model is studied and analyzed in many aspects. The research results show that the multiple linear regression model is more accurate in the prediction of Changsha City air quality.

According to the established multiple linear regression equation, after data standardization, $PM_{2.5}$, NO_2, O_3 and minat are positively correlated with air quality. PM_{10}, SO_2, CO, wc, wd and wp were all negatively correlated with air quality. Among them, $PM_{2.5}$ has a greater and positive impact on air quality. $PM_{2.5}$ is an important indicator to evaluate air quality in the world today. The higher the concentration of $PM_{2.5}$ in the air, the worse the air quality and the more serious the pollution. Weather conditions have a large and negative impact on air quality, good weather is good air quality. Therefore, it can be explained that the multiple linear regression analysis is reasonable.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Kai Su, Zhongshan Peng, Dan Zhu, Ruiqian Liu, Qin Wang, Rong Cao, Jun He. Water quality evaluation based on water quality index and multiple linear regression: A research on Hanyuan Lake in southern Sichuan Province, China. Water environment research: a research publication of the Water Environment Federation, 2024, 96(6): e11055-e11055.

[2] Mahmuda Akter, Elias Khalil, Md. Haris Uddin, Md. Kamrul Hassan Chowdhury, Shah Md. Maruf Hasan.

Artificial neural network and multiple linear regression modeling for predicting thermal transmittance of plain-woven cotton fabric. Textile Research Journal, 2024, 94(11-12): 1279-1296.

[3] Mahmoudi Mohammadreza, Toufigh Vahab, Ghaemian Mohsen. A Novel Multiple Linear Regression Approach for Predicting the Unconfined Compressive Strength of Soil. International Journal of Geomechanics, 2024, 24(8).

[4] Sharaf AlKheder. Experimental road safety study of the actual driver reaction to the street ads using eye tracking, multiple linear regression and decision trees methods. Expert Systems With Applications, 2024, 252(PA): 124222.

[5] Zhong H, Hu H, Hou N, et al. Study on Abnormal Pattern Detection Method for In-Service Bridge Based on Lasso Regression. Applied Sciences, 2024, 14(7).

[6] Jiang J, Li Y, Kleeman M. Air quality and public health effects of dairy digesters in California. Atmospheric Environment, 2024, 331120588.