EnPress Publisher

Application of Logistic Regression Model in the Prediction of Air Quality Level in Zibo City

Ting Fan

College of Geography and Environmental Science, Zhejiang Normal University, Jinhua 321021, China.

Abstract: Objectively evaluating urban ambient air quality and analyzing its influencing factors are of great significance for understanding the current status of air quality and controlling pollution sources. In this paper, logistic regression analysis is carried out for 366 days of air quality data from January to December 2020 in Zibo City, Shandong Province, with air quality class as the categorical variable, and six variables, $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3, are selected as pollution indicators affecting the air quality in Zibo City, and the stepwise regression method is utilized to establish a model and determine the weights of each pollution indicator. The established model is used to predict the samples of Zibo City, and the predicted data and actual data are compared to test the fit of the model. The results show that the logistic regression model fits well, and SO2 is the strongest factor affecting air quality, and the probability of air pollution is 1.035 times higher for every unit it increases and other variables remain unchanged, providing a basis for controlling the emissions of the primary pollution factors.

Keywords: Logistic; SPSS; Ambient air quality

1. Introduction

With the development of social industrialization and information technology, people's lives are becoming more and more convenient, but the environmental problems faced by mankind are also becoming more and more serious, of which air pollution is one of the topics worthy of people's attention.[1-2]. The level of environmental management in a city can reflect the state of ambient air quality, which directly affects the level of urban economic development and the quality of life of the people. With the sustained socio-economic development and the increasing intensity of regional development, the environmental pollution situation is worsening; and there are some cities where the degree of pollution is very obvious, thus affecting to a certain extent the speed of the city's economic development[3-5]. The increasing consumption of various types of energy and some other pollution generated by the daily production and life of residents have aggravated pollution and reduced the carrying capacity of the environment. Environmental pollution is not caused overnight, and the improvement of environmental quality also requires long-term governance and investment, which must start from multiple perspectives, and many aspects of joint efforts to form a long-term mechanism. Thus, quantitative and qualitative study of the ambient air quality situation in Zibo City, and analyze the influencing factors of ambient air quality, to understand the current situation of air pollution, looking for the source of pollution is of great significance, so as to provide guidance for the development of prevention and control measures and policies to effectively reduce pollution.

In recent years, Shandong Province is vigorously promoting the development of resource transformation, Zibo City, as an old industrial base in Shandong Province in the protection of environmental air quality has taken a series of strong measures to achieve significant achievements. Fan Xingxing[6] Characteristics, sources and impacts on human health of black carbon (BC) aerosol in Zibo in 2020-2022 were analyzed, and the results showed that BC was highest in winter, and the main sources were traffic emissions and coal combustion, and the impacts on children's lung function were high, but the study only selected one observation site and did not analyze different types of monitoring sites. Fang Bin.[7] According to the AQI and various pollutants data and corresponding meteorological data of Zibo City from 2013 to 2015, the air quality characteristics and its relationship with various meteorological elements were analyzed, and the results showed that $PM_{2.5}$ pollution was the most serious in Zibo City, and the high concentration values mostly appeared in winter. Hao Yujiao[8] The road mobile sources in Zibo were analyzed to study the pollutant emission characteristics of mobile sources with different fuel types, car models, emission stages and the concentration share of mobile source pollutants at the air monitoring stations, and the results showed that light-duty buses, heavy-duty diesel vehicles and off-road mobile machinery are the mobile sources with more pollutant emissions. This paper uses Logistic regression method to analyze the air pollution indicators in the city, which can objectively understand the current situation of ambient air quality in Zibo

City, and put forward relevant suggestions for improving the ambient air quality in Zibo City as well as provide a theoretical basis for managing air quality.

2. Data sources

In this article, we have collected 366 sets of air quality monitoring data of Zibo city from January to December 2020 from China Weather and the official website of China Environmental Monitoring General Station.[9] The data are complete and accurate. To get the quality level grading, you have to first determine the pollution level by the size of the AQI index. A total of six pollutants are involved in the calculation, namely: O_3 for 8 hours, PM_{10} for 24 hours, PM2.5 for 24 hours, CO for 8 hours, SO_2 for 24 hours, and NO_2 for 24 hours, and the unit of CO is mg/m^3, and the unit of the other five variables is ug/m^3 . Six variables, namely, $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, and O_3, are designated as dependent variables, which are summarized in Table 2-1.

Table 2-1 Data on Air Quality Categories and Their Associated Variables in Zibo City, 2019

dates	AQI	PM2.5	PM10	NO2	SO2	CO	O3_8h
2020/1/1	82	103	20	119	0.1	106	66
2020/1/2	162	63	23	2	1.6	51	28
2020/1/3	239	86	103	16	1.1	37	117
2020/1/4	292	86	21	36	0.3	12	23
2020/1/5	273	69	60	61	1.9	114	51
2020/1/6	90	111	28	45	1.5	88	70
2020/1/7	49	5	9	109	1.7	2	60
2020/1/8	72	17	31	22	0.1	49	92
2020/1/9	79	104	96	66	0.3	36	82
2020/1/10	137	44	81	34	1.3	96	109
2020/1/11	133	89	42	5	0.4	60	22
2020/1/12	84	8	79	67	0	115	112
2020/1/13	145	16	53	15	1.7	108	85
2020/1/14	85	59	97	57	0.4	110	82
2020/1/15	104	78	84	4	0	10	117
2020/1/16	158	84	81	33	0.5	85	53
2020/1/17	225	85	7	102	1.3	25	65
2020/1/18	212	3	54	114	0.2	72	107
2020/1/19	183	90	41	24	1.5	112	75
2020/1/20	86	72	41	5	0.4	3	64
2020/12/20	125	84	71	75	0.6	52	84
2020/12/21	137	59	77	50	1.6	70	45
2020/12/22	129	33	4	55	0.2	51	23
2020/12/23	125	93	56	76	0.8	43	114
2020/12/24	109	89	15	43	0	81	85
2020/12/25	99	0	23	53	1.4	5	44
2020/12/26	158	110	58	53	1.3	28	103
2020/12/27	204	51	34	3	1.4	68	36
2020/12/28	267	79	8	25	1.8	111	5
2020/12/29	43	46	117	47	1.8	95	38
2020/12/30	50	118	32	18	1	49	96
2020/12/31	58	19	62	4	1.9	53	97

3. Logistic regression model

Logistic regression model as a probabilistic nonlinear regression model, with the probability of occurrence of an event as the dependent variable and the influencing factors as the independent variables, is suitable for the case where the dependent variable is a categorical variable, and has more applications in the fields of society and economy, etc.[10] . Let an effect outcome indicator y be a dichotomous variable that takes the value of y = l to indicate that an effect occurs, and y = 0 to indicate that an effect does not occur. The risk factors affecting the effect outcome y are covariates (i.e., independent variables, also known as explanatory or forecast variables, which can be either continuous or discrete), and there are x_1. x_2.xm There are a total of m. The probability that the dependent variable will take 1, P(y=1 ||x) is the object to be studied, called Logistic linear regression model, which is a generalized linear model, and the method of linear modeling can be applied systematically.

In this study, we used the daily air quality level of whether the air quality level is polluted or not as a dichotomous variable and transformed it into a dichotomous dummy variable, where the air quality level of "excellent", "good" is transformed into 0, "mild pollution", "moderate pollution", "heavy pollution" and "severe pollution" are transformed into 1, i.e., those assigned a value of 0 are ", "moderately polluted", "heavily polluted" and "severely polluted" is converted to 1, i.e., the value of 0 is assigned to the value of "not polluted", and the value of 1 is "polluted", using SPSS software for binary logistic regression analysis.

$$\ln\frac{P}{1-P} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \quad (1)$$

where β_i (i=0,1,2,3, ..., m) is the regression coefficient.

where $\frac{P}{1-P}$ denotes the dominance ratio (the ratio of the probability of the event occurring to the probability of the event not occurring), the parameter to be estimated a, theβ_1, .β_2, ..., andβ_m reflects the change in the dominance ratio. If β_i is positive, its opposition value (index) must be greater than 1, then the dominance ratio will increase; conversely, if β_i is negative, the dominance ratio decreases. The formula for the probability P is:

::

$$P = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_m x_m)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_m x_m)} \quad (2)$$

4. Analysis of results

The starting block 0 is the model with only constant terms for the independent variables, which can be referred to as the baseline or null model. The model is given by considering variables that are selected into but not in the equation, and the null hypothesis (H0) is that no improvement will occur if a variable is included in the current model.

The data in Table 2-1 were imported into SPSS software to obtain the correlation statistics in the logistic regression equation (Table 4-1), and the results showed that the significance test values (Sig. values) for testing PM10, SO2 , CO and O3_8h- were all less than 0.05, which indicated that there was an improvement in including them in the model, but the Sig. values for PM2.5 and NO2 were greater than 0.05, which means that PM2.5 and NO2 will not improve the model. So remove PM2.5 and NO2 and proceed to step 2 of the process.

Table 4-1 Variables not in the equation

			Score	degrees of freedom	significance
Step 0	variant	PM2.5	2.573	1	.109
		PM10	4.669	1	.031
		NO2	.425	1	.514
		SO2	5.729	1	.017
		CO	6.126	1	.013
		O3_8h	8.124	1	.004
	Overall statistical information		33.015	6	.000

Table 4-2 shows the results of two Hosmer tests, P1=0.454>0.05, P2=0.895>0.05, accepting the null hypothesis (the null hypothesis is that the observed data and the regression model fit well, the alternative hypothesis is that the fit is not good), which means that the binary logistic regression model established by the current data and the real data fit well.

Table 4-2 Hosmer and Lemeshow tests

Step length (T)	chi-square	degrees of freedom	significance
1	7.793	8	.454
2	3.913	8	.865

Table 4-3 shows the predicted values of the real data using the binary logistic regression fitting equation. In the first regression processing, the correct rate of prediction using the binary logistic regression fitting equation is 63.4%, and after the processing in step 2, the final correct rate of prediction is 64.8%, which indicates that the binary logistic regression fitting equation is more effective.

Table 4-3 Classification Tablea

observed value			projected value		
			quality level		Percentage correct
			non-contamination	contamination	
Step 1	quality level	non-contamination	174	44	79.8
		contamination	90	58	39.2
	Overall percentage				63.4
Step 2	quality level	non-contamination	176	42	80.7
		contamination	87	61	41.2
	Overall percentage				64.8

a. A cut-off value of .500

Checking Table 4-4, at the end of Step 2, the coefficients of the remaining five variables are suggested to be significant by the Wald test P-value; further checking their corresponding coefficients, the B-values are all greater than 0, which indicates that PM2.5, PM10, SO2, and O3_8h have a positive effect on the AQI, and the larger the AQI the more serious the air pollution condition is, therefore, the larger the concentration of PM2.5, PM10, CO, and O3_8h, the Therefore, the larger the concentrations of PM2, PM, CO, O3_8h are, the worse the air quality is. The B-value of NO is -0.049, which indicates that it is negatively correlated with the air quality level. The “exp (B)” value in the table is the OR value, which is specifically explained as follows: PM2.5 concentration increases by one unit, the possibility of air pollution is 1.107 times; PM10 concentration increases by one unit, the possibility of air pollution is 1.030 times; SO2 concentration increases by one unit, the possibility of air pollution is 1.31 times; SO2 concentration increases by one unit, the possibility of air pollution is 1.3 times; SO2 concentration increases by one unit, the possibility of air pollution is 1.3 times. The possibility of air pollution is 1.318 times higher for each unit increase in SO2 concentration; NO2 is a protective factor, indicating that the air treatment of nitrogen oxides has been effective. The possibility of air pollution increases by 6.2% for each unit increase in O3_8h concentration.

Table 4-4 Variables in the program

		B	S.E.	Wald	degrees of freedom	significance	Exp(B)	95% C.I. for EXP(B)	
								lower limit	limit
Step 1	PM2.5	.006	.003	2.813	1	.094	1.006	.999	1.012
	PM10	.006	.003	3.159	1	.075	1.006	.999	1.013
	NO2	.005	.003	2.012	1	.156	1.005	.998	1.012
	SO2	.039	.018	4.766	1	.029	1.040	1.004	1.077
	CO	.012	.003	13.503	1	.000	1.013	1.006	1.019
	O3_8h	.008	.003	7.753	1	.005	1.008	1.002	1.013
	constant	-2.626	.467	31.577	1	.000	.072		
Step 2	PM10	.006	.003	3.196	1	.074	1.006	.999	1.013
	SO2	.035	.017	3.976	1	.046	1.035	1.001	1.071
	CO	.013	.003	14.679	1	.000	1.013	1.006	1.020
	O3_8h	.007	.003	6.881	1	.009	1.007	1.002	1.013
	constant	-2.354	.422	31.168	1	.000	.095		

The independent variables contained in the model obtained in the second step are PM_{10}, SO_2, CO and O_3, a total of five variables, and the Sig. value of SO_2, CO and O_3 is less than 0.05 (indicating statistical significance) and the significance is also less than 0.05; further check the corresponding coefficients, the value of B is greater than 0, which indicates that they all have a positive effect on the AQI, and the larger the AQI, the more serious the air pollution situation is; therefore, the larger the concentration of PM, SO_2, CO and O_3_8h, the worse the air quality is; the column B in the table is the constant and the coefficient of the independent variable (weights). The larger the air pollution is, the more serious it is, so the larger the concentrations of PM_{10}, SO_2, CO, O_3_8h are, the worse the air quality is; the column of B in the table is an estimate of the constant and the coefficients of the independent variables (weights), and Exp(B) is the dominance ratio, which is generally referred to as the odds ratio (OR), and it can be viewed as a ratio of the occurrence category of the independent variables to the occurrence category of the independent variable for each increase of a unit, while the other variables are kept unchanged. It can be viewed as the multiple of the change in the odds ratio of the occurrence category caused by each unit increase in the independent variable when all other variables are held constant (when the probability of occurrence of an event is not very large, it can be approximated as the multiple of the change in the probability of occurrence of the event caused by each unit increase in the independent variable). In this final model, we can see that the coefficient of SO_2 is the largest, so SO_2 is the strongest influence on air quality, it increases by one unit, other variables remain unchanged, the probability of air pollution is 1.035 times the original; similarly, CO increases by one unit, the probability of air pollution is 1.013 times the original, O_3 increases by one unit, the probability of air pollution is 1.007 times the original. 1.007 times the original.

Using the logistic regression model, the final predictive model is as follows:

$$P = \frac{\exp(0.006x_1 + 0.35x_2 + 0.13x_3 + 0.07x_4 - 2.354)}{1 + \exp(0.006x_1 + 0.35x_2 + 0.13x_3 + 0.07x_4 - 2.354)}$$

5. Conclusions and recommendations

In this study, according to the categorization characteristics of whether the air is polluted in the air quality grade, based on the binary logistic regression method to screen the pollution indicators, we established the air quality grade model of Zibo City, determined the factors affecting the air quality in Zibo City as SO_2, CO, O_3 and PM_{10}, and derived the weight and advantage ratio of the pollution indicators, of which SO_2 has the greatest impact on the air quality. It is applied to the evaluation and prediction of the actual air quality, and a better prediction effect is achieved.

Zibo city government should further accelerate the conversion of old and new kinetic energy and industrial upgrading, especially for enterprises that will produce sulfide focus on governance, optimize the structure of energy consumption and the structure of import and export commodities trade, accelerate the management of the number of small and heavy polluting enterprises, air pollution is serious disorder,

the transformation of large and heavy polluting enterprises, and further improve their energy utilization and recycling rate, reduce the pollution of the air.

References

[1] Song Hong, Sun Yajie, Chen Dengke. Evaluation of the effect of governmental air pollution control--an empirical study from the construction of “low-carbon cities” in China[J]. Management World,2019,35(06):95-108+195.

[2] Zhou Q, Li X, Hu J, et al. Dynamics and optimal control for a spatial heterogeneity model describing respiratory infectious diseases affected by air pollution[J]. Mathematics and Computers in Simulation,2024,220276-295.

[3] Li Weibing, Zhang Kaixia. The impact of air pollution on firm productivity - Evidence from Chinese industrial firms[J]. Management World, 2019, 35(10): 95-112+119.DOI:10.19744/j.cnki.11-1235/f.2019.0134.

[4] Ghaffarpasand O, Okure D, Green P, et al. The impact of urban mobility on air pollution in Kampala, an exemplar sub-Saharan African city[J].Atmospheric Pollution Research, 2024, 15(4):102057.

[5] Wang Min, Huang Ying. Environmental pollution and economic growth in China[J]. Economics(Quarterly),2015,14(02):557-578.

[6] Fan Xingxing. Source analysis and health risk evaluation of black carbon aerosol in Zibo[D]. Tianjin University of Technology, 2023.

[7] Fang Bin, Liu Houfeng. Characteristics of ambient air quality and relationship with meteorological conditions in Zibo[J]. Green Science and Technology, 2017, (24): 26-2.

[8] Hao Yujiao. Research on Emission Reduction Countermeasures of Mobile Pollution Sources with Multi-source Data Fusion[D]. Shandong University of Technology, 2022.

[9] Ma Xinhua. Causes of ambient air pollution in Boshan District, Zibo City, Shandong Province and countermeasures against it[J]. Qinghai Environment, 2001.

[10] Zibo Municipal Bureau of Statistics, National Bureau of Statistics Zibo Investigation Team.2020 Zibo City National Economic and Social Development Statistical Yearbook [M]. Beijing:China Statistics Press, 2020.

[11] OHLMACHER GC, DAVIS JC. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA[J]. Engineering Geology, 2003, 69(3/4):331-343.