Pollution Study 2021, 2(1), 1976.
https://doi.org/10.54517/ps.v2i1.1976

Article

Analysis of air quality pollution in Chengdu utilizing logistic regression for multi-class classification based on the distance metric between classes

Ting-Ting Li[1], Wang-Yong Lv[1,2,*], Jiao Zhou[1], Shi-Jing Zeng[1]

[1] Institute of Mathematics Science, Sichuan Normal University, Chengdu 610068, China
[2] Visual Computing and Virtual Reality Key Laboratory of Sichuan Province, Chengdu 610068, China
*** Corresponding author:** Wang-Yong Lv, lvwangy@163. com

CITATION

Li TT, Lv WY, Zhou J, Zeng SJ. Analysis of air quality pollution in Chengdu utilizing logistic regression for multi-class classification based on the distance metric between classes. Pollution Study. 2021; 2(1): 1976. https://doi.org/10.54517/ps.v2i1.1976

ARTICLE INFO

Received: 1 July 2021
Accepted: 25 July 2021
Available online: 9 August 2021

Abstract: The health of individuals is intimately linked to air quality, making it crucial to investigate the pollutants that influence it. A sequential Logistic multi-classification approach, which utilizes the distance between classes, was applied to analyze the air quality data of Chengdu spanning from May 2019 to April 2020. By leveraging the inter-class distance metric, the multi-class classification challenge was converted into a series of binary classification tasks. Using a sequential strategy, binary Logistic regression was then employed for each task. The accuracy rate post-stepwise regression was utilized to assess the impact of various pollutants on air quality. The findings indicate that $PM_{2.5}$, PM_{10}, NO_2, and O_3 are the four primary pollutants with the most significant collective influence on Chengdu's air quality. Consequently, the government should enhance the joint monitoring of these pollutants and develop targeted policies to mitigate their levels.

Keywords: air pollutants; distance between classes; sequential; logistic multiple classification

With the continuous development of social economy and the gradual increase of human activities, various kinds of pollution have been produced, which has caused great harm to the environment, so the environmental problems are becoming more and more serious. Among them, air pollution is one of the more serious environmental problems. The quality of air is closely related to people's life, so controlling air pollution is one of the important decisions to protect the environment.

To control air pollution, first of all, we need to understand the types of pollutants in the air, the causes of these air pollutants and the impact of major air pollutants on air quality. These issues have always been the research focus of Chinese and foreign scholars, and have achieved remarkable results. Zhang et al. [1] used remote sensing images and ground observation data to study the impact of dust weather on the atmospheric environment and air pollutant concentration in Xilingol League, a grassland city in Inner Mongolia, and provided a theoretical basis for the environmental impact assessment of dust weather; Liu et al. [2] used the surface meteorological observation data of the Yangtze River Delta urban agglomeration to analyze the meteorological causes and pollutant paths of a heavy pollution weather process in the Yangtze River Delta urban agglomeration, and concluded that the atmospheric chemical model WRF Chem can better simulate the change process of $PM_{2.5}$ concentration, which can be used as the business model of heavy pollution weather forecast in the Yangtze River Delta urban agglomeration; Xu et al. [3] studied the change trend of urban air quality and six pollutant concentrations in different periods after the implementation of the new ambient air quality standard in

Jinan by using statistical analysis, trend test and correlation analysis, and identified that Jinan is a composite air pollution characterized by PM_{10}, $PM_{2.5}$ and O_3 pollution.

And there are many ways to study these problems. Scholars use different methods to explore these problems. In order to study the concentration and source of polycyclic aromatic hydrocarbons (pahs) in $PM_{2.5}$ during heating and non heating periods in Changchun, Li et al. [4] combined the ratio method to analyze the source of pahs, and conducted health risk assessment through the health risk assessment model of the national environmental protection agency, so as to provide a scientific basis for the comprehensive prevention and control of air pollution and environmental management in Changchun; Xie et al. [5] used the multivariate linear model to predict the concentration of air pollutant $PM_{2.5}$ in Nanning, and the results showed that the prediction model was applicable to the monitoring and prediction of air pollutant $PM_{2.5}$ in Nanning; Zhang et al. [6] used statistical regression analysis to analyze the particle size and quantity of atmospheric particles, draw the corresponding particle size change histogram, infer the chemical composition and source of atmospheric particles according to the inspection results of atmospheric particles, and put forward effective strategies to control the pollution of atmospheric particles $PM_{2.5}$ and PM_{10}.

Air quality is one of the important factors related to people's quality of life and health. Therefore, a sequential logistic multi classification algorithm based on distance between classes (SMLG) is proposed to analyze the main pollutants affecting air quality. In this method, the data are clustered by the distance between classes, and the multi classification problem is transformed into multiple two classification problems, and then the two classification logistic is used for classification. Using this method, the air quality data of Chengdu from April 2019 to may 2020 are classified, and then the air quality data affecting Chengdu are regressed step by step. Finally, the pollutants that have the greatest impact on air quality are obtained according to the classification accuracy of SMLG.

1. Sequential logistic multiple score based on distance between classes

There is a data set *B* with a sample *m* size of. The $B_1, B_2, \cdots, B_k$ data *k* is composed of such data. The sample size of each type n_i of *B* data is. For classification, the dependent *Y* variable is $1, 2, \cdots, k$ taken as and $X = (X_1, X_2, \cdots, X_n)$ the *n* independent variable is a dimension variable. According *Y* to the value of order or disorder, logistic multiple classification can be divided into two categories, one is ordered logistic multiple classification [7] (omlg), and the other is unordered multi class logistic [8] (UOMLG).

1.1. Ordered logistic multiple classification

When the dependent variable *y* is an ordered variable, the logistic multiple classification should adopt the ordered logistic multiple classification. The principle of classification is to regard as the first category, the remaining categories as the second category, and then use binary logistic regression to classify. Let *X* be given, the probability of y taking J $P_j = P(Y = j|X)$ is $P_1 + P_2 + \cdots + P_k = 1$ 1, and. $j = 1,2,\cdots,k$. Now, the k classes are $\{1, 2, \cdots, j\}$ divided $\{j+1, j+2, \cdots, k\}$ into and, of $j = 1, 2, \cdots, k-1$ which. Then the two classification logistic regression model is used for classification, and the ordered multi classification logistic regression model is transformed into multiple two classification logistic regression models. A total of $K-1$ two classification logistic regression equations are fitted, and the formula is:

$$\begin{cases} f_j = \ln \dfrac{\sum_{i=1}^{j} P(Y = j \mid X)}{1 - \sum_{i=1}^{j} P(Y = j \mid X)} = \beta_{0j} + \sum_{i=1}^{n} \beta_{ij} X_i \\ j = 1, 2, \cdots, k-1 \end{cases} \tag{1}$$

In Equation (1), β_{0j} , β_{ij} and j are the intercept and partial regression coefficients of the second classification logistic model respectively. According to the maximum likelihood estimation, the estimated value β_{0j} of β_{ij} the sum can $\hat{\beta}_{0i}, \hat{\beta}_{ij}$ be obtained, and the value can be obtained by substituting it into $P(Y = j| X), j = 1, 2, \cdots, k$ Equation (1). Then, according to the principle of maximum probability, the values of $P(Y = j \mid X), j = 1,2,\cdots,k$ each are compared.

1.2. Unordered logistic multiple classification

When the dependent variable y is an unordered variable, unordered logistic multiple classification should be used. The principle of this method is to set a certain class as the main class, then carry out binary logistic regression with other classes, and establish $k-1$ binary logistic regression model. Let the probability that y takes J under the condition of given $P_j(Y = j \mid X), j = 1,2,\cdots,k$ X $P_1 + P_2 + \cdots + P_k = 1$ be, and. Take class k in B as the main category, and then perform binary logistic regression on the other $K-1$ categories and the selected main category respectively. A total of K1 independent binary logistic regression models are fitted. The formula is:

$$\begin{cases} g_j = \ln \dfrac{P(Y = j \mid X)}{P(Y = k \mid X)} = \beta_{0j} + \sum_{i=1}^{n} \beta_{ij} X_i \\ j = 1, 2, \cdots, k-1 \end{cases} \tag{2}$$

Similarly, according to the maximum likelihood estimation, and are $\hat{\beta}_{0j}$ substituted $\hat{\beta}_{ij}$ into Equation (2), and then the inverse solution of Equation (2) can $P(Y = j \mid X), j = 1,2,\cdots,k-1$ be obtained. And because $P_1 + P_2 + \cdots + P_k = 1$ of, it is $P(Y = k \mid X)$ available.

Finally, by deriving the calculated $P(Y = j \mid X), j = 1,2,\cdots,k$ probability, the values of each are compared according to the principle $P(Y = j \mid X), j = 1,2,\cdots,k$ of maximum probability.

1.3. SMLG method

Due to the problems of poor classification accuracy and low operation efficiency of big data in both ordered and disordered multi classification. Based on this, the SMLG method is proposed. It uses the idea of fast clustering in clustering analysis to cluster according to the distance between classes (the distance between classes). So, we need to solve the distance between classes first.

1.3.1. Distance between classes

The data B set $B_1, B_2, \cdots, B_k$ is k composed of such data. The sample size of each type of data is $n_i, i = 1,2,\cdots,k$. The distance between any two types of and in B the data set can B_p be B_q calculated by using $D_{pq}(p, q \in 1, 2, \cdots, k$ Equation $p < q)$ (3) and the formula is:

$$D_{pq} = \frac{1}{n_p n_q} \sum_{\substack{x_i \in B_p \\ x_j \in B_q}} d_{ij}^2 \tag{3}$$

In Equation (3): x_i, x_j is the i-th $(i \in 1, 2, \cdots n_p)$ and the j-th $(j \in 1, 2, \cdots, n_q)$ n dimension vectors of B_p, B_q, respectively; d_{ij}^2 is the distance between x_i, x_j.

Linear condition: d_{ij}^2 is Euclidean distance when the data set is linear. Then $d_{ij}^2 = \|x_i - x_j\|^2 = \sum_{l=1}^{n}(x_{il} - x_{jl})^2$, where x_{il}, x_{jl} are the l-th component of n dimension vectors x_i and x_j, respectively.

The nonlinear change is transformed into a linear problem in another space. $\varphi(x)$ is the corresponding vector to map the original space vector x to the high-dimensional space, then $d_{ij}^2 = \|\varphi(x_i) - \varphi(x_j)\|^2 = K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)$, where $K(x_i, x_j) = \varphi(x_i)\varphi(x_j)$ is the kernel function. Common kernel functions are shown in **Table 1**.

Table 1. Common kernel functions.

Kernel function	**Expression form**	**Parameter**
Liner kernel	$K(x_1, x_2) = x_1 x_2 + c$	c is constant
Polynomial kernel	$K(x_1, x_2) = (x_1 x_2 + c)^d$	c is constant, d >0
Sigmoid kernel	$K(x_1, x_2) = \tan h[\beta(x_1 x_2) + c]$	c is a constant, $\beta > 0$
Gaussian kernel	$K(x_1, x_2) = \exp\left(-\frac{\lvert x_1 - x_2 \rvert 2}{2\sigma^2}\right)$ or $\exp(-\gamma \lvert x_1 - x_2 \rvert^2), \gamma = \frac{1}{2\sigma^2}$	$\Sigma > 0$
Laplace kernel	$K(x_1, x_2) = \exp\left(-\frac{\lvert x_1 - x_2 \rvert}{\sigma}\right)$	$\Sigma > 0$

1.3.2. SMLG method

SMLG method is proposed to solve the problems of poor accuracy, low efficiency and limited use of traditional logistic multi classification. The method first calculates the distance between classes and selects the minimum distance. Assuming that the distance B between B_i the $B_j (i, j \in 1, 2, \cdots, k)$ sum in the dataset is the smallest, B_i the B_j sum can be aggregated into a new B_{ij} category 'Regard B_{ij} as the first category, then regard other $k - 2$ categories as the second category, and then conduct binary logistic regression. In this $Y_1 = \{0$ case, $1\}, X_1 = (X_{11}, X_{21}, \cdots, X_{n1})$

it n is a dimensional variable. The following is the derivation of the binary logistic regression model.

Under given X conditions, the sum $Y_1 = 0$ of $Y_1 = 1$ probabilities of and is 1, i.e.,

$$P(Y_1 = 0 \mid X_1) = 1 - P(Y_1 = 1 \mid X_1) \tag{4}$$

After introducing the logistic function,

$$\begin{aligned} \frac{P(Y = 0 \mid X_1)}{P(Y = 1 \mid X_1)} &= \frac{P(Y = 0 \mid X_1)}{1 - P(Y = 1 \mid X_1)} \\ &= \frac{\dfrac{1}{1 + e^{f_1}}}{1 - \dfrac{1}{1 + e^{f_1}}} \\ &= e^{-f_1} \end{aligned} \tag{5}$$

In Equation (5), and $f_1 = \beta_{01} + \sum_{i=1}^{n} \beta_{i1} X_{i1}$ the logarithm of the left and right sides of Equation (5) can be obtained.

$$\ln \frac{P(Y_1 = 1| \ X_1)}{P(Y_1 = 0| \ X_1)} = f_1 \tag{6}$$

Because $f_1 = \beta_{01} + \sum_{i=1}^{n} \beta_{i1} X_{i1}$ of $\beta_{01}, \beta_{11}, \cdots, \beta_{n1}$ the unknown, the estimated value can be obtained by maximum likelihood estimation. Existing, $P(Y = 0| \ X_1) = h_\beta(X_1), P(Y = 1| \ X_1) = 1 \ h_\beta(X_1)$ consolidated as:

$$\begin{aligned} P(Y_1 \mid X_1, \beta) = [\hbar_\beta(X_1)]^{Y_1} \\ [1 - h_\beta(X_1)]^{1-Y_1} \end{aligned} \tag{7}$$

According to Equation (7), the likelihood function is:

$$\begin{aligned} L(\beta) = \prod_{i=1}^{m} \left\{h_\beta \left[X_1^{(i)}\right]\right\}^{Y_1^{(i)}} \\ \left\{1 - h_\beta \left[X_1^{(i)}\right]\right\}^{1-Y_1^{(i)}} \end{aligned} \tag{8}$$

Take logarithm of Equation (8) to obtain:

$$\begin{aligned} \ln L(\beta) &= \sum_{i=1}^{m} \Big(Y_1^{(i)} \ln h_\beta [X_1^{(i)}] + [1 - Y_1^{(i)}] \times \\ &\quad \ln\{1 - h_\beta [X_1^{(i)}]\} \Big) \\ &= \sum_{i=1}^{m} \Big(Y_1^{(i)} \ln \frac{h_\beta [X_1^{(i)}]}{1 - h_\beta [X_1^{(i)}]} + \\ &\quad \ln\{1 - h_\beta [X_1^{(i)}]\} \Big) \qquad (\end{aligned} \tag{9}$$

Deriving from Equation (9),

$$\frac{\partial \ln L(\beta)}{\partial \beta} = 0 \tag{10}$$

The $\beta_{01}, \beta_{11}, \cdots, \beta_{n1}$ resulting $\hat{\beta}_{01}, \hat{\beta}_{11}, \cdots$ estimate, $\hat{\beta}_{n1}$ In the $\hat{\beta}_{01}, \hat{\beta}_{11}, \cdots, \hat{\beta}_{n1}$ Equation (6) of the derivation, the logistic regression function is obtained, and the probability value of the given sample is calculated after it is substituted into the logistic regression function. By comparing the size with 0.5, the class of the sample is judged. In this way, the binary logistic regression is used to separate B_{ij} ““ and B_{k-2} ““. Next, binary logistic regression is used to separate ““ and ““ again, B_i and ““ B_j is regarded B_i as the first class ““ B_j is regarded as the second class, $Y_2 =$ $\{0,1\}$ then, $X_2 = \left(X_{12}, X_{22}, \cdots X_{n2}\right)$.

The second logistic regression function is:

$$\begin{aligned} \frac{P(Y_2 = 0 \mid X_2)}{P(Y_2 = 1 \mid X_2)} &= \frac{P(Y_2 = 0 \mid X_2)}{1 - P(Y_2 = 0 \mid X_2)} \\ &= \frac{\frac{1}{1 + e^{f_2}}}{1 - \frac{1}{1 + e^{f_2}}} \\ &= e^{-f_2} \end{aligned} \tag{11}$$

The logarithm of Equation (11) can be obtained at the same time.

$$\ln \frac{P\left(Y_2 = 1 \mid X_2\right)}{P\left(Y_2 = 0 \mid X_2\right)} = f_2 \tag{12}$$

In Equation $f_2 = \beta_{02} + \sum_{i=1}^{n} \beta_{i2} X_{i2}$ (12), the maximum likelihood estimation is also used $\hat{\beta}_{02}, \hat{\beta}_{12}, \cdots, \hat{\beta}_{n2}$ here.

Through the above method, and are B_i separated, B_j and the above $k-2$ method is repeated in the remaining classes. In this way, the $k-$classification can be completed through a binary logistic regression. The regression model is:

$$P\left(Y_i \mid X_i\right) = \frac{1}{1 + e^{f_i}}, i = 1, 2, \cdots, k-1 \tag{13}$$

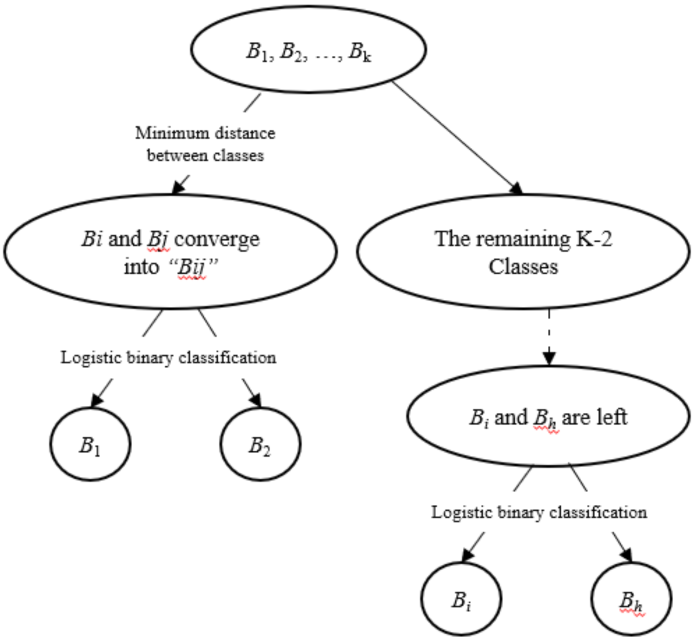
$B_1, B_2, \ldots, B_k$
Minimum distance between classes
Bi and *Bj* converge into *“Bij”*
The remaining K-2 Classes
Logistic binary classification
B_1
B_2
B_i and B_h are left
Logistic binary classification
B_i
B_h

Figure 1. Classification model diagram. $B_1, B_2, \ldots, B_k$

Finally, according to the regression results, the accuracy of the method can be counted, and the classification model of the method is shown in **Figure 1**.

2. Empirical analysis

2.1. Method validation

The effectiveness of SMLG method is verified by data simulation. The experimental data used are the balance scale dataset (hereinafter referred to as the balance dataset), datauser dataset and grisoni in UCI (University of California in Irvine) [9] machine learning database_ et _ al_ 2016 _ Envint88 data set (hereinafter referred to as Gris data set), seam data set, crowdsourced mapping data set, available data set and Beijing data set in the data network [10] (which contains the air quality data of Beijing scenic spots) were tested with R software.

The datasets shown in Table 2 are all data of Category 3, among which the balance datasets, datause datasets and Gris datasets are all non-linear datasets. Through comparison, it is found that when the data is nonlinear, the accuracy of the linear distance SMLG method is lower than that of the nonlinear distance SMLG method. Therefore, the advantages of this method cannot be demonstrated, so the selection of distance is very important when using SMLG method.

Table 2 shows the running time and accuracy of different data sets using SMLG method and UOMLG method. Through comparison, it can be found that for small data, the running time of UOMLG method is shorter than that of SMLG method, because SMLG will classify and cluster the data, resulting in an increase in running time. However, for big data, it can be found from **Table 2** that the running time of SMLG method is more saved than that of unordered multi classification. This is because when the data set is large, SMLG method will classify and re cluster the data, so that big data will become multiple small data, and the running time of small data is less. Even if classification and re clustering are added, the total running time is less than that of UOMLG method.

Table 2. Comparison of the classification accuracy of three data linear distance and nonlinear distance SMLG methods.

Data set	UOMLG method/%	SMLG method using nonlinear distance/%	SMLG method using linear distance/%
Balance	88.98	9164	8967
Datause	84.03	8570	8341
Gris	73.50	7380	7088

Note that the Beijing dataset in **Table 3** adopts the ordered logistic multi classification method, so the UOMLG accuracy in the Beijing dataset is actually the UOMLG accuracy. By showing the classification accuracy of SMLG method and UOMLG method, it shows that the accuracy of classification is improved when using SMLG method for multi classification.

Table 3. Comparison of running time and accuracy of SMLG method and UOMLG method for multiple data.

Data set	Category	Data volume	SMLG method		UOMLG method	
			Time/s	**Accuracy rate/%**	**Time/s**	**Accuracy rate/%**
Balance	3	624	0.694392	91.64	0.177954	88.98
Datause	3	234	0.886043	85.70	0.158663	83.94
Gris	3	779	0.748206	91.64	0.187465	73.50
Seane	4	732	1.287105	66.68	0.177831	67.30
Crowdsourced mapping	6	10845	2.790425	86.91	3.016118	86.77
Beijing	4	237	0.140644	77.68	1.326987	74.51

2.2. Empirical analysis

The air quality data of Chengdu from 1 May 2019 to 30 April 2020 were collected from the post weather website [11], with a total air quality of 334 days. The data includes the air quality index (AQI) value and the concentrations of $PM_{2.5}$, PM_{10}, O_3, sO_2, nO_2 and Co. AQI is divided into six grades according to the concentration limit of air pollutants in the technical regulations for ambient air quality index (AQI) (Trial) (hj633–2012), namely excellent, good, light pollution, medium pollution, heavy pollution and serious pollution, which are represented by numbers 1, 2, 3, 4, 5 and 6. See **Table 4** for details. Since there are only four kinds of air quality data in this data set, i.e., Excellent, good, slightly polluted and moderately polluted, the AQI values of this data set are 1, 2, 3 and 4.

Table 4. AQI size value and corresponding level.

Air quality class	Numerical grade	AQI limits
Excellent	1	0~50
Good	2	51–100
Light pollution	3	101~150
Moderate pollution	4	151–200
Heavy pollution	5	201–300
Serious pollution	6	>300

Step 1: carry out correlation analysis on Chengdu air quality data set, and draw a scatter diagram using R software, as shown in **Figure 2**. The scatter diagram between variables can reflect the linear relationship between the two variables. For example, the third sub figure in Row 2 in **Figure 2** shows the scatter diagram of $PM_{2.5}$ and PM_{10}. It can be observed that the distribution of scatter points is similar to a straight line, so $PM_{2.5}$ and PM_{10} have a strong linear relationship. Therefore, it can be seen that $PM_{2.5}$ and PM_{10} have obvious linear correlation, while $PM_{2.5}$ and Co, PM_{10} and Co, PM_{10} and NO_2 have no obvious linear correlation. Therefore, the data are generally nonlinear.

Step 2: Randomly divide the Chengdu air quality data set into training set and test set according to the proportion (the proportion is 7:3), and then calculate the distance between various types in the training set. Since the data is nonlinear, the nonlinear distance formula should be adopted, in which the Gaussian kernel is selected as the kernel function:. $K(x_1,x_2)=\exp\left(-\frac{|x_1-x_2|^2}{2\sigma^2}\right),\sigma=0.02$ Finally, SMLG method is used to classify the data set and judge the air quality level of each $Y=\{1,2,3,4\}$ data, $X=\{PM_{2.5}\, PM_{10}, O_3, SO_2, NO_2, CO\}$. The distance between categories of the data is shown in **Figure 3**. The shortest distance between categories 1 and 2 can be regarded as 12, and categories 3 and 4 as 34. In this way, we can first separate 12 and 34 with binary logistic, then separate category 1 and Category 3 with binary logistic, and finally separate category 3 and category 4 with binary logistic. As shown in **Figure 3**, the training data can be separated by three binary logistic regression, and three logistic classifiers can be trained.

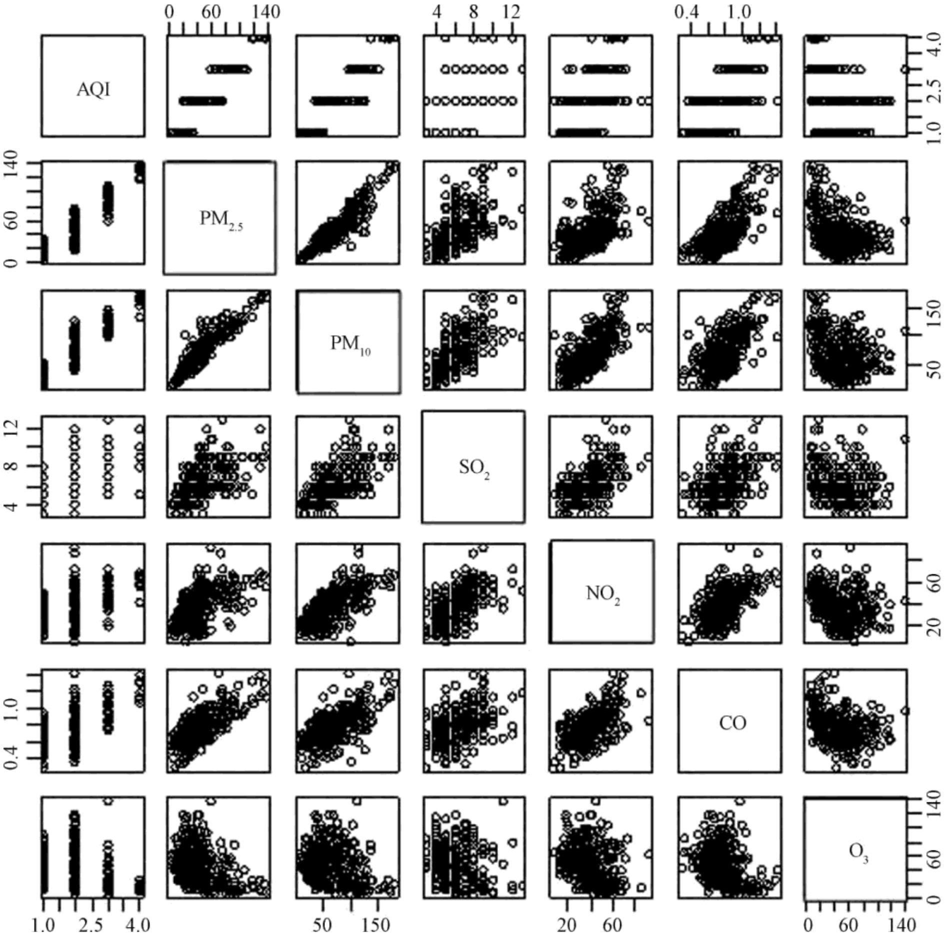

Figure 2. Correlation scattered matrix diagram of Chengdu meteorological data.

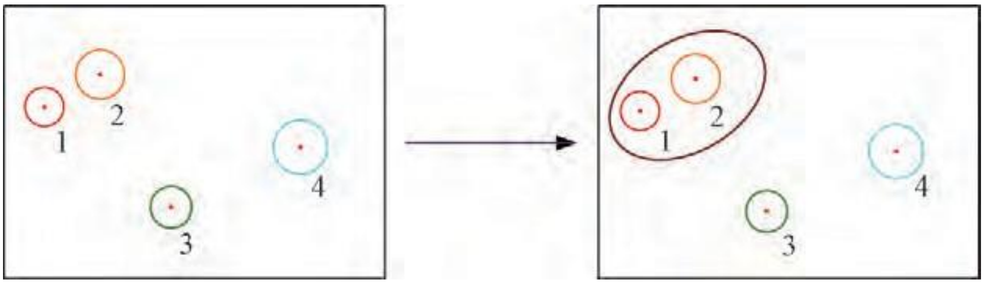

Figure 3. The map of SMLG classification idea.

Step 3: put the test set into the trained classifier, repeat the above steps for 100 times, and finally get the average value of the correct rate of 100 times, which is used as the correct rate of SMLG. The accuracy of UOMLG method and SMLG method for Chengdu air quality data is compared, as shown in **Table 5**.

Table 5. Classification accuracy of air quality data in Chengdu based on SMLG and UOMLG.

Method	**SMLG**	**OMLG**
Accuracy rate/%	9308	9183

2.3. Analysis of pollutants affecting AQI

The air quality data of Chengdu from 1 May 2019 to 30 April 2020 are classified by SMLG method, and their accuracy is recorded. Through the method of step-by-step classification, the pollutant factors that affect air quality are analyzed, and the types of pollutants that have a greater impact on air quality are obtained. The data is composed of AQI value and the concentrations of $PM_{2.5}$, PM_{10}, O_3, SO_2, NO_2 and CO. After gradual classification, it is concluded that the most important pollutants affecting Chengdu’s air quality are $PM_{2.5}$, PM_{10}, nO_2 and O_3, with a classification accuracy of 9334%. As the stepwise classification has been carried out 63 times using SMLG method, only part of the stepwise classification results are shown, as shown in **Table 6**.

Table 6. Analysis of pollutants affecting AQI.

$PM_{2.5}$, PM_{10}, NO_2, O_3	93.34
$PM_{2.5}$, PM_{10}, NO_2, CO	91.59
$PM_{2.5}$, PM_{10}, CO, O_3	92.54
$PM_{2.5}$, PM_{10}, NO_2, CO, O_3	92.32
PM_{10}, CO, SO_2	8451

It can be seen from **Table 6** that after the pollutants $PM_{2.5}$, PM_{10}, nO_2 and O_3 are classified by SMLG again, the classification accuracy is the highest compared with that of other pollutant factors through SMLG. Therefore, it can be concluded that the pollutants that have a great comprehensive impact on Chengdu’s air quality are $PM_{2.5}$, PM_{10}, nO_2 and O_3.

Cause analysis of important pollutants

As an important economic transportation hub and cultural and Art Center in western cities, Chengdu is densely populated and prosperous in industry. However, the large number of motor vehicles and human activities caused by industrial production have an important impact on air quality. In recent years, with the gradual improvement of urban development and the gradual maturity of industry in Chengdu, however, the development of industrialization has a huge demand for energy, and the environmental pollution caused by energy consumption is inevitable. As the government continues to promote the concepts of “green development”, “green water and green mountains" and "sustainable development", since 2015, although the use of coal, crude oil, diesel and other energy has been decreasing year by year, it still accounts for the largest proportion in the overall energy consumption. It is precisely because of the use of these energy sources in industrial construction that they have become an important source of sulfur dioxide and smoke (powder) dust emissions in the air. For example, in 2018, coal and crude oil accounted for 82% of the total energy used [12], as shown in Figure 4. At the same time, with the increase of the number of motor vehicles in Chengdu, the number of motor vehicles in Chengdu has exceeded 5million by 2019, becoming the second largest city in China [13]. The emission of sulfur compounds in automobile exhaust has a great impact on the air quality of Chengdu. Although the emission of sulfur dioxide and smoke (powder) dust in the air is gradually decreasing from 2015 to 2019, its overall proportion is still the first in the exhaust gas content.

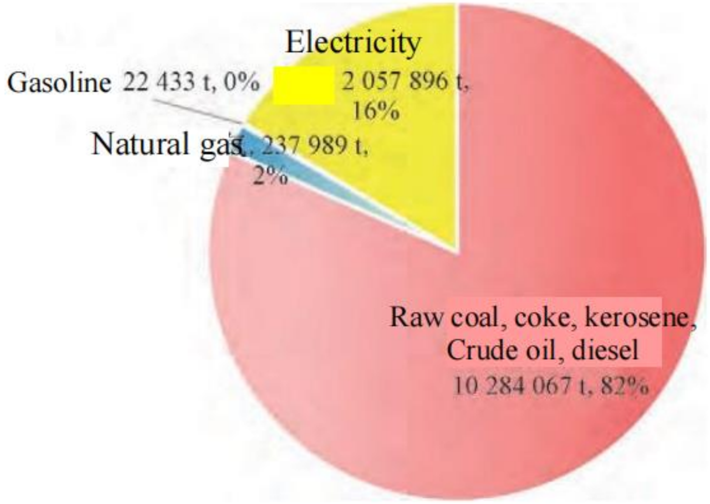

Figure 4. Proportion of Chengdu's energy consumption in 2018.

At the same time, the climate will also have a certain impact on the air quality. The Chengdu Plain has four distinct seasons and little sunshine. It is mainly affected by the warm and humid subtropical Pacific southeast monsoon. The climate is humid and warm. Humid climate will lead to moisture absorption of $PM_{2.5}$, making the secondary conversion of $PM_{2.5}$ faster, leading to the increase of its concentration and aggravating the degree of air pollution [14].

Chengdu is located in the west of Sichuan Basin, surrounded by plateau mountains, Longquan Mountain in the southeast and Longmen Mountain in the northwest. Its unique geographical environment has caused special ozone pollution in Chengdu, which is characterized by the compound pollution of high concentration O_3 and $PM_{2.5}$ [15]. At the same time, Sichuan Basin is located near the Qinghai Tibet Plateau. Due to the influence of air flow, temperature change and air pressure change, it often has a sunny, less cloudy, high temperature and low humidity environment in summer, which is conducive to the photochemical reaction of ozone. In addition, Chengdu is represented as the middle and lower atmosphere [16], and the downdraft establishes a static weather situation, resulting in the accumulation of ozone pollutants in the near ground layer and the formation of continuous pollution.

3. Conclusions and suggestions

The air quality data of Chengdu from May 2019 to April 2020 are classified according to the air quality index AQI. The classification method is based on sequential multiple classification logistic regression (SMLG), and the feasibility of the model is analyzed. Then SMLG is used to classify the air quality data of Chengdu and calculate the classification accuracy. Finally, stepwise regression is used to analyze the pollutant factors affecting AQI.

The results show that the pollution $PM_{2.5}$, PM_{10}, NO_2 and O_3 have the strongest comprehensive influence on AQI. Therefore, the four important pollutants that affect the air quality in Chengdu are $PM_{2.5}$, PM_{10}, nO_2 and O_3. In fact, the factors that affect the air quality in Chengdu are complex. Although the special geographical location and climatic conditions will lead to changes in the air quality to some extent, the greater impact on the air quality is human activities. Only by actively responding to and strictly supervising, can the environmental air quality in Chengdu be improved.

Chengdu meteorological monitoring department should strengthen the supervision of exhaust emission in industrial production. At the same time, the government should also strengthen the promotion of the use of clean energy, strengthen the treatment of automobile use, prevent dust, and create a healthy living environment for people.

Funding: National Natural Science Foundation of China Youth Science Fund (11601357); application foundation project of Sichuan Provincial Department of science and Technology (2017jy0159); Sichuan Key Laboratory Project of visual computing and virtual reality (SCVCVR2018. 08VS).

Conflict of interest: The authors declare no conflict of interest.

References

1. Zhang B, Li J, Zhao J, et al. Effects of aeolian dust on urban air quality in Inner Mongolia grassland. Science Technology and Engineering. 2018; 18(3): 123-131.
2. Liu L, Zhao Q, Wang J, et al. Analysis of a winter-time heavily polluted weather process in the Yangtze river Delta urban agglomeration. Science Technology and Engineering. 2019; 19(26): 376-383.
3. Xu M, Zhang L, Zheng N, et al. Analysis on the change trend and pollution characteristics of urban air quality after revision of ambient air quality standards: take Jinan City as an example. Science Technology and Engineering. 2020; 20(13): 5422-5428.
4. Li N, Wei X, Zhou Y, et al. Source analysis and health risk assessment of polycyclic aromatic hydrocarbons in atmospheric environment $PM_{2.5}$ in Changchun City. Science Technology and Engineering. 2021; 21(1): 410-416.
5. Xie S, Zhou Z, Li G. Relationship between PM2.5 concentration and meteo-logical factors in Nanning. Science Technology and Engineering. 2020; 20(1): 460-466.
6. Zhang X, Yu L, Sun M, et al. Distribution and elemental analysis of particulate matter $PM_{2.5}$, PM10 in the atmosphere of spring in Jinan City. Science Technology and Engineering. 2018; 18(25): 278-285.
7. Tian H. Linear regression model estimation method with ordinal multi-category explanatory variables and its application research. Chengdu: Southwest Jiaotong University; 2019.
8. Li J. Multi-classification research based on Logistic regression model and support vector machine (SVM) model. Wuhan: Central China Normal University; 2014.
9. Murphy PM, Ahad W. UCI repository of machine learning database. Available online: https://archive.ics.uci.edu/ (accessed on 13 April 2020).

10. Wang Y. PM2.5 air quality data, historical data and meteorological data - data network. Available online: http:// www.shu-ju.net/ (accessed on 13 April 2020).
11. Wu Y. Weather report. Available online: http://www.tianqihoubao.com (accessed on 13 April 2020).
12. Chengdu Bureau of Statistics. Chengdu statistical yearbook 2018. Beijing: China Statistics Press; 2018.
13. Urban Planning Newsletter. Baidu Maps released the China Urban transport research report for the Second Quarter of 2018. Urban Planning Newsletter. 2018(14): 14.
14. Cao Y, Wang C, Zhao X, et al. Analysis of the characteristics of PM2.5 pollution in Chengdu and its relationship with surface meteorological elements. Mid-low Latitude Mountain Meteorology. 2020; 44(4): 59-64.
15. Li P, Luo B, Zhang W, et al. Analysis of temporal and spatial distribution characteristics and pollution characteristics of ozone in Sichuan Province. Environmental Science and Technology. 2018; 41(S1): 293-298.
16. Yang X, Yi J, Lü Y, et al. Analysis of the causes of severe ozone pollution in Chengdu and surrounding areas. China Environmental Science. 2020; 40(5): 2000-2009.