Analysis of PM2.5 Influencing Factors Based on Various Statistical Methods—A Case Study of Beijing in 2021

Qing Yu(✉)

School of Statistics, Beijing Normal University, Beijing 100875, China
yuqing1026@126.com

Abstracts. Beijing, the capital of China, is suffering from great pollution by PM2.5. In order to give suggestions to solve this problem, several studies have been conducted to explore the internal relationship between PM2.5 and other pollutants, showing different results. This paper compared different kinds of mainstream statistical methods and gave the convincing influence factors based on the AQI index and six indicators of Beijing in 2021. Firstly, the preparation work was done by detecting the possible problems with the data itself, constructing training set and testing set. Secondly, this study generalized models with explained variable PM2.5 and explaining variables PM10, SO_2, CO, NO_2 and O_3. Then, GLS, ridge regression, LASSO regression, PCA and RF are done, which are all calculated with test MSE to show the accuracy. Finally, the conclusion is that RF is the best among those statistical methods. All methods prove that the concentration of carbon monoxide plays a decisive role in PM2.5 concentration, which means reducing automobile exhaust emission may low down the PM2.5 content.

Keywords: PM2.5 content · statistical methods · influencing factors · Beijing

1 Introduction

PM2.5, which refers to particles with a diameter of no more than 2.5 microns in the ambient air, has a strong impact on the climate system through direct radiation and indirect radiation forcing in the troposphere [13]. As the capital of China, Beijing has suffered from serious PM2.5 pollution, which may cause economic losses [14] as well as health hazardsc. To deal with this problem, some people tend to discover internal relationships between PM2.5 and other pollutants, trying to find possible sources of PM2.5 that can be controlled artificially [3]. While, statistical methods to deal with those problems are quite different, so the results are also varied, and the lack of comparations among methods make them less convincing.

This article tries to solve the above problems by comparing the application effect - test Mean Square error - of current mainstream statistical analysis methods on this subject. The methods being compared are GLS (Generalized Least Squares) which typifies linear regression methods applied to time series data [16] ridge regression and LASSO

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 195–204, 2023.
https://doi.org/10.2991/978-94-6463-034-3_20

regression which typify linear regression regularization methods [5], PCA (Principal Component Analysis) which typifies dimension reduction model regression methods [6], and RF(Random Forest) which typifies machine learning methods [11]. Also, this article will present a more comprehensive results on the relationship between PM2.5 and other pollutants.

From this research, the best method with the minimum analysis error rate can be found adopting common statistical methods, which means the analysis of the relationship between PM2.5 and other pollutants will become more unified. Also, the main influence factors of PM2.5 based on the case of Beijing, 2021 can be discovered. In this way, suggestions are being made to low down the PM2.5 content according to the results.

2 Data Collection

This report adopts the 365 days air quality data of 2021 published by China air quality online analysis platform as the data to be analyzed. The information in a single day data set includes the specific date, air quality level, AQI ranking for the day, PM2, PM10, SO_2, NO_2, CO, and O_3. All values have no missing data, so they can be applied directly. Since the units of the six indicators (PM2.5, PM10, SO_2, NO_2, CO and O_3) used for analysis are the same, there is no need to standardize again. The data has already tested by the kappa function, implying that there is no explicit collinearity [9]. Moreover, heteroscedasticity is also tested by the residual diagram (Residuals versus Fitted) drawn under ordinary least squares regression demonstrated by Fig. 1, indicating the residual has no obvious trend and is randomly distributed on both sides of the 0division line [1], which means heteroscedasticity is not serious enough to be considered. Additionally, in order to judge PM2.5 for the internal consistency with the air quality rating, the significance of the difference between the two samples was tested by one-way ANOVA [4]. Since the analysis of variance aims to analyze the influence of qualitative variables on quantitative variables, it is necessary to treat the qualitative variable "air quality grade" as a dummy variable. The result shows that the air quality rating is very important for PM2.5. Last but not least, the data is finally equally divided into two parts—the training set and testing set. The whole analysis is done by R studio which is widely used in statistical analysis.

3 Results and Analysis

3.1 GLS Regression

GLS not only effectively give weight to each residual when it is uncertain whether it is heteroscedasticity, but also it is convinced that GLS model aims at the possible endogeneity, that is, autocorrelation and other problems [15]. So, GLS is better than OLS [8]. Now, we use GLS regression and calculate the test MSE. The output regression model is

$$\mathrm{PM2.5} = 0.139\mathrm{PM10} + 0.520\mathrm{SO_2} + 0.784\mathrm{NO_2} + 65.336\mathrm{CO} + 0.034\mathrm{O_3} - 37.860 \quad (1)$$

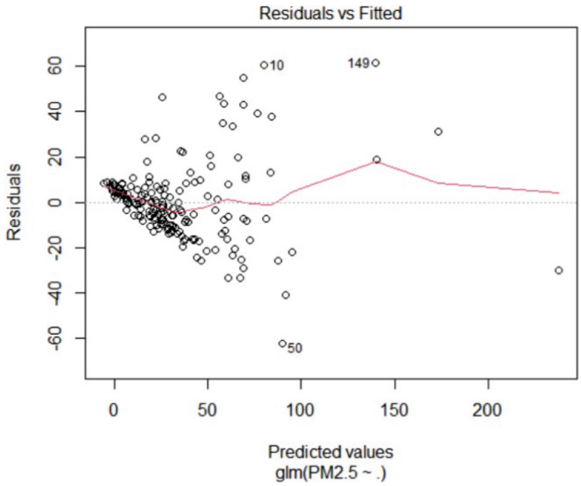

Fig. 1. Residual diagram (drawn by the author)

and AIC is 1544.8 indicating an unsatisfactory fitting result. Also, 40% of the independent variables did not pass the 0.05 significance test, so we need to find a better model.

3.2 Ridge Regression

Ridge regression estimation is based on least square estimation, which changes the final regression result by changing the standardized matrix of independent variable matrix X. For the estimation formula with parameters of multiple linear regression.

$$\hat{\beta} = (X^T X)' X^T Y \tag{2}$$

the ridge regression processes the parameters of this part and obtains.

$$\hat{\beta}(k) = \left(X^T X + kI\right)' X^T Y \tag{3}$$

It can be seen that the size of K has a great impact on the regression results. In R, the optimal K can be determined through CV (cross validation) (as shown in Fig. 3), which is also an ergodic method. By using ridge regression, the model comes to.

$$\mathrm{PM2.5} = 0.175\mathrm{PM10} - 1.076\mathrm{SO}_2 + 0.462\mathrm{NO}_2 + 63.021\mathrm{CO} - 0.021\mathrm{O}_3 - 23.897 \tag{4}$$

Which was built by the best $\lambda = 2.55$.

3.3 LASSO Regression

Lasso (Least absolute shrinkage and selection operator) is a kind of compression estimation. It is a more refined model that is obtained by constructing a penalty function (L1-penalized). The construction method of penalty function is:

$$L = \|\beta \mathrm{X} - \mathrm{Y}\|^2 + \lambda \|\beta\|_1 \tag{5}$$

To calculate the parameter β, we find the partial derivative of this formula, which results in:

$$\frac{\partial L}{\partial \beta} = X^T\left(Y - X\hat{\beta}\right) + \frac{\partial(\lambda\|\beta\|_1)}{\partial \beta} \tag{6}$$

$$\hat{\beta}_j = sign\left(\frac{1}{n}X^T Y\right)\left(\left|\frac{1}{n}X^T Y\right| - \frac{\lambda}{2}\right)\left(\left|\frac{1}{n}X^T Y\right| - \frac{\lambda}{2}\right)_+ \tag{7}$$

In this case, the model is

$$\mathrm{PM2.5} = 0.1657\mathrm{PM10} + 0.376\mathrm{NO_2} + 63.431\mathrm{CO} - 25.151 \tag{8}$$

It is worth noting that lasso has a heavier penalty on the number of variables, so the number of independent variables retained is less than that of ridge regression. SO_2 and O_3 are compressed close to 0 (< 0.0000001).

3.4 PCA

Firstly, fa.parallel () function is used to do simple principal component analysis. This function can automatically find the appropriate number of principal components and output the eigenvalues of principal components. This method is a parallel test method, that is, generating a matrix of a group of random data. These matrices have the same number of variables and subjects as the real case data matrix. The average eigenvalue of this group of random data matrix is calculated. By comparing the gravel diagram of eigenvalues in the real data and the curve of the average eigenvalue of this group of random matrix, the intersection of the two characteristic curves can be found. If the eigenvalue of real data is lower than that of random data, there will be no retained value. As can be seen from Fig. 2, the first and second principal components have great retention value, and the third principal component is located at the boundary and is temporarily retained for further exploration.

Secondly, we initially do not specify the number of principal components. At this time, all principal components are output which can be noticed in Fig. 3. By observing the cumulative contribution rate of principal components, it is found that the cumulative contribution rate of variance of the first three principal components is 83.82%, and the corresponding eigenvalues of the first three principal components are basically greater than 1. The cumulative contribution rate of variance of the first two principal components is less than 65%, which indicates that they cannot well retain the original variable information, so the first three principal components should be taken. To support our choice, a scree plot is drawn in Fig. 4.

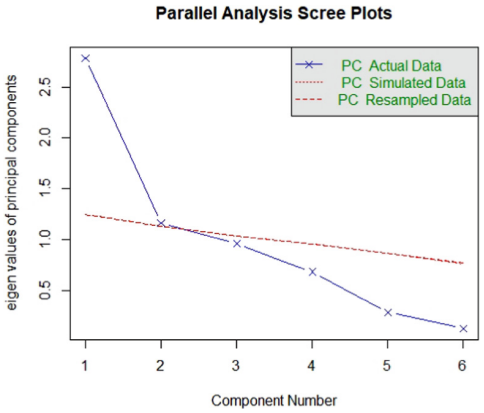

Fig. 2. Parallel analysis scree plots (drawn by the author)

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.4669232 1.0287819 0.9903358 0.7484553 0.49879224
Proportion of Variance 0.4303727 0.2116784 0.1961530 0.1120371 0.04975874
Cumulative Proportion  0.4303727 0.6420512 0.8382042 0.9502413 1.00000000

Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
PM10  0.120  0.234  0.963
So2   0.419  0.562 -0.159 -0.678  0.157
No2   0.614 -0.202                0.762
Co    0.519  0.286 -0.184  0.688  0.376
O3   -0.404  0.712 -0.117  0.252 -0.502
```

Fig. 3. Principal components result (drawn by the author)

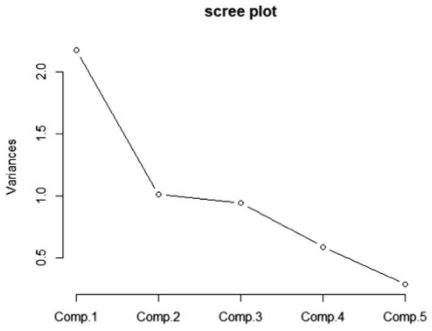

Fig. 4. Scree plot (drawn by the author)

Next, visualize the load of principal component analysis and draw the scatter diagram (as shown in Fig. 5) with 1, 2, 3 and 4 loads. It can be observed that PM10 has a large load on the third principal component, CO has a large load on the first principal component, NO_2 has a large load on the first principal component, O_3 has a large load on the second principal component, and SO_2 has a larger load on the second component compared to the first component, but none of them are big enough to allow SO_2 to be involved. The principal component score is shown in Fig. 6. It can be seen that the difference between

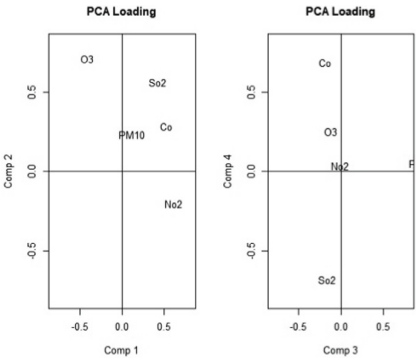

Fig. 5. Scatter diagram (drawn by the author)

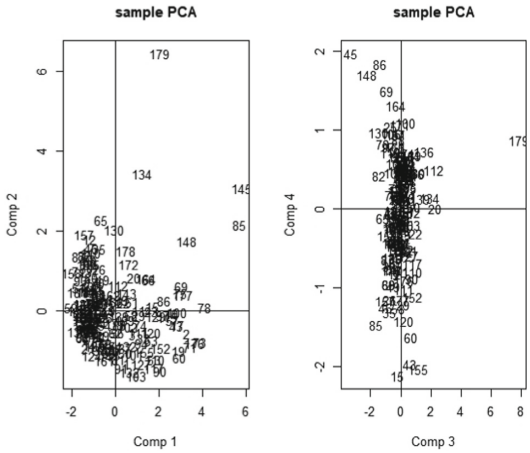

Fig. 6. Principal component score plots (drawn by the author)

the groups is significantly greater and the difference within the group is smaller after the dimensionality reduction of the first and second principal components.

Finally, the original five prediction variables are changed into three principal components after dimensionality reduction, and the principal components are used for regression. The model is.

$$\mathrm{PM2.5} = 16.548\mathrm{Comp1} + 8.306\mathrm{Comp2} + 14.965\mathrm{Comp3} + 37.478 \quad (9)$$

which indicates that Comp1 refers to CO and NO_2, Comp2 refers to O_3, Comp3 refers to PM10, and all variables are significant.

Consequently, the first principal component is 'automobile exhaust' [7], the second principal component is 'urban light pollution' [2], and the third principal component is 'dust pollution' [10].

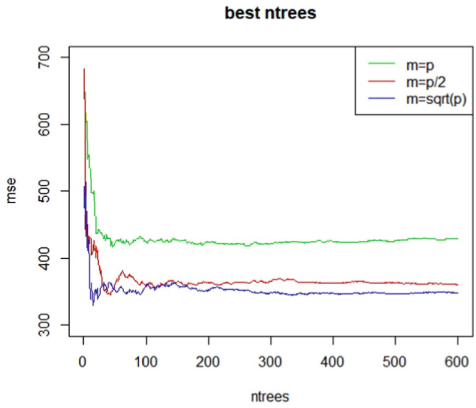

Fig. 7. Number decision diagram of tree (drawn by the author)

3.5 RF

In machine learning, RF (Random Forest) is a classifier containing multiple decision trees, and its output category is determined by the mode of the category output by individual trees. This method combines the ideas of 'bootstrap aggregating' and 'random subspace method' to build a set of decision trees.

Firstly, it is well-known that the number of trees in the random forest is not the more the better because too many trees will cause over fitting problems. Therefore, it is very important to select the optimal number of trees. The judgment standard is that the minimum number of trees with basically stable error is the optimal number of trees.

Secondly, it is also very important to judge the optimal number of segmentation points. Here we select three commonly used mtry values and compare the mtry value with the smallest MSE.As shown in Fig. 7, the optimal mtry is $m = \sqrt{P}$ ($P = 5$, $m \approx 2$), which is basically stable when the number of trees is greater than 350, so $nTree = 350$.

Finally, randomforest package is used with $mtry = 2$ and $nTree = 350$. Importance function is applied to explore the importance of variables and observe the % IncMse of each prediction variable. The result is demonstrated in Fig. 8, showing the variables with greater importance are PM10 and CO.

3.6 Results

After comparing the test errors of each model (The comparing information can be seen in Table 1), it is judged that the random forest model is the optimal model, and the model judges the impact of CO and PM10 on PM2.5 concentration has a great influence.

Additionally, several models have proved that the concentration of carbon monoxide plays a decisive role in PM2.5 concentration (The comparing information can be seen in Table 2). Also, through principal component analysis, it can be found that the first principal component is carbon monoxide and nitrogen dioxide.

In addition, urban light pollution (O3) and dust pollution (PM10) also affect PM2.5 content to some extent. However, it is worth noting that sulfur dioxide concentration is

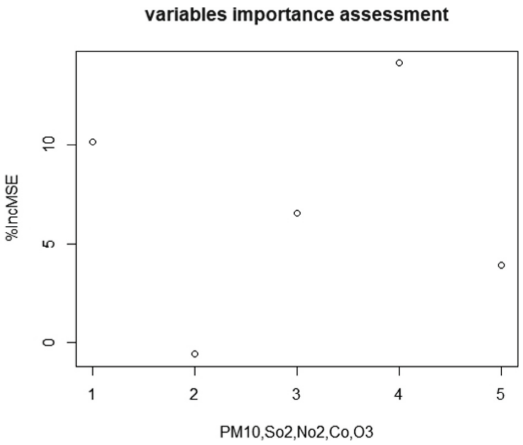

Fig. 8. Variables importance assessment plots (drawn by the author)

Table 1. Illustration of test MSE of different methods

Test MSE of each method	
GLS	313.786
Ridge regression	2080.596
LASSO regression	2016.848
PCA	2033.5
RF	157.5329

Table 2. Illustration of decisive factors of PM2.5 in each model

Decisive factors of PM2.5 in each model	
GLS	CO
Ridge regression	CO
LASSO regression	CO
PCA	CO, NO_2
RF	CO, PM10

negatively correlated with PM2.5 in some models, which means that industrial pollution and automobile exhaust pollution have a negative correlation as the two main factors of urban pollution.

4 Conclusion

Firstly, through this survey of relationships between PM2.5 and other crucial pollutants using a variety of statistical analysis methods, the best method 'random forest model' is discovered by comparing to test errors of other different methods. By applying ANOVA and several factors to convincing the data, it can be proved that PM2.5 can reflect air quality.

Secondly, the results indicate that CO and PM10 play essential roles in PM2.5 content, especially the CO, which is consistent with the scientific common sense that PM2.5 is inhalable particulate matter. It can also be found that the PCA shows the most decisive influencing factors are carbon monoxide and nitrogen dioxide, which are related to automobile exhaust emission. Therefore, controlling automobile exhaust emission is very important for controlling PM2.5 content, and this conclusion is practical for governments taking actions to deal with the troublesome problem. Also, it is more persuasive because the comparation between various mainstream statistical methods.

Furthermore, this article only discusses the influence factors of PM2.5 among other commonly tested pollutant from public datasets, the relationship between PM2.5 and other pollutants must not be restricted in them and more influence factors are being found in the future.

To predict the future research trend of this problem, it will be concentrated on finding the one or more reasons for PM2.5 formation in essence, and more environmental analysis will be taken instead of indirect statistical analysis.

References

1. Astivia, Oscar L. Olvera and Zumbo, Bruno D. "Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS," Practical Assessment, Research, and Evaluation: Vol. 24, Article 1, 2019. https://doi.org/10.7275/q5xr-fr95
2. Cerón-Bretón, J.G., Cerón-Bretón, R.M., Kahl, J.D.W. et al. Diurnal and seasonal variation of BTEX in the air of Monterrey, Mexico: preliminary study of sources and photochemical ozone pollution. Air Qual Atmos Health 8,2015, 469–482.
3. https://doi.org/10.1007/s11869-014-0296-1
4. Cong Lin, sun Deshan, Zou Cunli, Zhang Lei Study on relevant factors of PM2.5 in Beijing [J] Economic mathematics,, 2017, 34(04): 26–29 https://doi.org/10.16339/j.cnki.hdjjsx16339/j.cnki
5. Connelly, Lynne M. Medsurg Nursing; Pitman Vol. 30, Iss. 3, (May/Jun 2021): 218.
6. Fang Jiajia, Li Yang, Zheng Zemin Network connection data variable selection based on ADMM algorithm [J] Computer system application, 2022,31 (01): 11–20 https://doi.org/10.15888/j.cnkicsa. eight thousand two hundred and forty-seven
7. Liu Shipeng. Research on intelligent campus network intrusion remote detection based on Internet of things [J] Automation technology and application, 2022,41 (02): 64–68
8. Magnus Lenner, Nitrogen dioxide in exhaust emissions from motor vehicles, Atmospheric Environment (1967), Volume 21, Issue 1, 1987, Pages 37–43, ISSN 0004-6981.
9. https://doi.org/10.1016/0004-6981(87)90268-X.
10. Ma Hanyang. Social responsibility effect and governance of major infrastructure projects [D] Shanghai Jiaotong University, 2018 https://doi.org/10.27307/d.cnki.gsjtu. 2018.000502.

11. M. Benchoufi, E. Matzner-Lober, N. Molinari, A.-S. Jannot, P. Soyer, Interobserver agreement issues in radiology, Diagnostic and Interventional Imaging, Volume 101, Issue 10, 2020, Pages: 639–641, ISSN: 2211–5684.
12. https://doi.org/10.1016/j.diii.2020.09.001.
13. Stojiljkovic A , Kauhaniemi M , Kukkonen J , et al. The impact of measures to reduce ambient air PM10 concentrations originating from road dust, evaluated for a street canyon in Helsinki[J]. Atmospheric Chemistry and Physics, 2019, 19(17):11199–11212.
14. Yuan Jinrong, Li Ling, Liu Aihong, Zhao Wenya Design and Research on quantitative risk assessment platform for fall monitoring [J] Nursing research, 2019,33 (15): 2615–2618
15. Yang Xiaozhe, Feng Lin, Zhang Yannan, Shi Yanfeng, Duan Junchao, sun Zhiwei, Effects of PM2.5 exposure on cardiac function and its mechanism[C] //Proceedings of the 2019 National Symposium on respiratory toxicology and health toxicology [publisher unknown], 2019:91–92
16. YANG Yingchuan, GE Baozhu, HAO Saiyu, XU Danhui, LIU Ying, GAN Lu, WANG Zifa. Inversion of PM2.5 Concentration in Beijing Based on Visibility and AOD Data[J]. Climatic and Environmental Research (in Chinese],2020, 25(5):521–530.
17. Wang J , Zhang L , Niu X , et al. Effects of PM2.5 on health and economic loss: Evidence from Beijing-Tianjin-Hebei region of China[J]. Journal of Cleaner Production, 2020, 257(3):120605.
18. Wang Qing, Zeng Fu, LV Xinmeng Will anti-dumping curb the export of digital trade—— On the regulating effect of economic internal circulation [J] Price monthly, 2022 (02): 45–53 https://doi.org/10.14076/j.issn. 1006–2025.
19. Zhang Xuhui, Zhu Jingying, Zhu Xun, Wang Lin, Zhang Qi Risk assessment of excess mortality of PM2.5, the main air pollutant in Wuxi [J] Jiangsu preventive medicine, 2022,33 (01): 17–20 https://doi.org/10.13668/j.issn. 1006–9070.2022.01.005.