Scientific Research Publishing

Journal of Geoscience and Environment Protection, 2017, 5, 37-41
http://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Study on Probability Estimation of Haze in Beijing Based Logistic Regression Model

Yonghua Zhu, Tian Zhang, Chenglin Chen

North China Electric Power University, School of Mathematics and Physics, Beijing, China
Email: zter2015@126.com

How to cite this paper: Zhu, Y.H., Zhang, T. and Chen, C.L. (2017) Study on Probability Estimation of Haze in Beijing Based Logistic Regression Model. *Journal of Geoscience and Environment Protection*, **5**, 37-41.
https://doi.org/10.4236/gep.2017.56005

Received: May 10, 2017
Accepted: June 9, 2017
Published: June 12, 2017

Abstract

The Logistic Regression Model of two categories is used to explore the relationship between haze and season, various meteorological factors such as air pressure, temperature, relative humidity, precipitation, wind direction and so on. Among all the factors, the relative humidity is best related to haze and season is in the second place. The odds of haze in winter are 17.87 times bigger than that in summer, 3.99 times bigger than that in spring. The odds of haze would increase by 48 percent averagely when the relative humidity increase by 10 percent.

Keywords

Haze, Logistic Regression, Season, Relative Humidity

1. Introduction

In many areas, haze has been reported as one of the disastrous weather. Fog is an aerosol system that consists of a large number of tiny droplets or ice crystals suspending in near-surface air. Haze refers to the deterioration of visibility caused by aerosol contamination. The external cause of the severe haze episodes was the unusual atmospheric circulation, the depression of strong cold air activities and the very unfavorable dispersion due to geographical and meteorological conditions. However, the internal cause was the quick secondary transformation of primary gaseous pollutants to secondary aerosols, which contributed to the "explosive growth" and "sustained growth" of PM2.5 [1]. Increased attention has been paid to PM 2.5 pollutants in China [2]. The haze weather has a direct impact on many aspects, especially the public health and the visibility. So far, the study of PM2.5 is focused on its composition, source and its infection in climate etc. The methods used are correlation analysis, linear regression model, and principal component analysis.

The dependent variable must be continuous in linear regression model. As Haze here is classification variable, the Logistic Regression Model of two categories is applied, which is a special form of logarithmic linear model [3]. Logistic models are applied in many fields, such as psychology, education, sociology and finance.

2. Data and Variable

2.1. Data Acquisition

The concentration of PM2.5 in Beijing comes from the official website of the US Embassy in China (http://www.stateair.net/web/historical/1/html). The data of meteorological factors is collected from China Meteorological Data Service Center (http://data.cma.cn/data/index/6d1b5efbdcbf9a58.html). The Dataset called Data of Specific Synoptic Hours from Global Surface Weather Stations includes the daily routine observations such as pressure, temperature, humidity, wind direction, wind speed, precipitation amounts at the synoptic hours from Chinese Surface Stations for global exchange acquired by the National Meteorological Information Centre (NMIC) through the domestic telecommunication system and those from foreign surface stations for global exchange acquired via GTS.

According to Ambient air quality standards, the 24-hour average primary concentration limitation of PM2.5 is 35 micrograms per cubic meter, and the secondary limitation is 75 micrograms per cubic meter. When the PM2.5 concentration is no more than 35 micrograms per cubic meter, the air quality is defined as good. When the PM2.5 concentration is between 35 and 75 micrograms per cubic meter, the air quality is defined as moderate.

This paper selected the PM2.5 concentration and meteorological data for every three hours from January 1 to December 31, 2016. After deleting the missing value, a total of 2526 sample cases left.

2.2. Variable Selection

If the PM2.5 concentration is no more than 75 micrograms per cubic meter, haze, which is the dependent variable, equals 1, else equals 0.

Air pollution is closely related to meteorological factors. Meteorological factors have an important impact on atmospheric environmental pollution. When the weather conditions are different, the concentration of ground pollutants caused by the same source of pollution may vary by several times or even hundreds of times. The main factors influencing the diffusion of air pollutants are wind and turbulence, atmospheric boundary structure, precipitation, inverse temperature, humidity, etc. The long-lasting fog and haze event occurred in a high pressure weather system and calm wind condition. The stable boundary-layer structure resulted from temperature inversions that were built by warm advection and radiation cooling provided a favorable condition for the accumulation of polluted aerosols and the formation and development of the fog and haze event [4]. This paper considers the relationship between the probability of occurrence of haze and various meteorological conditions and seasonal factors.

The independent variables chosen are

1) Pressure (Prs). Pressure results from the interaction of ground features and other meteorological parameters. To a certain extent, the variation of air pressure can reflect the comprehensive characteristics of this area.

2) Wind Direction (Win_d)

3) Wind Speed (Win_s)

4) Temperature (Tem). The temperature plays a leading role in determining the diffusion and dilution of air pollutants. The temperature also has effect in the transformation between different pollutants and the formation of secondary pollutants. The diffusion activity of the pol particles is closely related to the temperature.

5) Relative Humidity (Rhu). The relative humidity, expressed as a percentage, is the ratio of the water vapor density actually contained in the unit volume of air to the saturated water vapor density at the same temperature [5]. Lots of studies have shown that relative humidity is an effective meteorological factor which determines atmospheric particulate matter concentrations.

6) Precipitation_1h (Pre).

7) Season (Sea). Spring (Spr), summer (Sum), autumn (Aut) and winter (Win) are divided differently in different times and different areas. Now the method widely used for dividing four seasons, is proposed by Chinese scholars Zhang Baokun in 1934. Every five days are divided into a group, which are divided to spring when their average temperature rises from 10 to 22 degrees Celsius, divided to summer when the average temperature exceeds 22 degrees Celsius, divided to autumn when the average temperature decreases from 22 to 10 degrees Celsius [6]. Located in the northwest edge of the North China Plain, Beijing is affected by the northern temperate monsoon climate, which appears hot and humid in summer, dry and cold in winter. Besides, winter lasts longest among four seasons. In 2016, According to Zhang Baokun, four seasons in Beijing start from March 26, May 21, September 21, November 11 in turn.

3. Model Establishment and Test

On the basis of forward stepwise selection method, the model includes pressure, temperature, humidity, wind direction, wind speed and intercept. To estimate the regression coefficients, we use the maximum likelihood method. The estimates and significance tests of the regression coefficients are shown in the following **Table 1**.

The logit model is

$$\ln\frac{p}{1-p} = -0.53 + 3.94Rhu - 2.96Spr - 5.18Sum - 2.18Aut - 0.002Win\_d + 0.09Win\_s + 0.09Tem - 0.07Prs \quad (1)$$

Therefore, the logistic regression model of haze in Beijing is

$$p = \frac{\exp(-0.53 + 3.94Rhu - 2.96Spr - 5.18Sum - 2.18Aut - 0.002Win\_d + 0.09Win\_s + 0.09Tem - 0.07Prs)}{1 + \exp(-0.53 + 3.94Rhu - 2.96Spr - 5.18Sum - 2.18Aut - 0.002Win\_d + 0.09Win\_s + 0.09Tem - 0.07Prs)} \quad (2)$$

Table 1. Model estimation and Wald test.

	Prs	Win_d	Win_s	Tem	Rhu	Sea	Spr	Sum	Aut	Cons
B	−0.07	−0.002	0.09	0.09	3.94		−2.96	−5.18	−2.18	−0.53
S.E.	0.01	0.00	0.04	0.01	0.29		0.23	0.30	0.22	0.35
Wald	42.66	12.15	3.83	54.19	184.1	305.4	163.8	292.2	101.8	2.31
df	1	1	1	1	1	3	1	1	1	1
Sig	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.13
exp(B)	0.93	1.00	1.09	1.09	51.32		0.052	0.006	0.113	0.591

3.1. The Fit Goodness Test of the Model

As the number of independent variables rises, many covariates have only a few cases of observation especially when continuous independent variables are included in the model [3]. As the model contains five continuous independent variables, the Hosmer-Lemeshow index is chosen to test the model's goodness of fit. The p value of the test is 0.044, which shows that the model fits well for the haze.

3.2. The Accuracy of Prediction

The classification table is a frequency table where the observed cases are divided into occurrence or not. The classification table can be used to test the predictive accuracy of the Logistic Regression Model. As 0.5 is used as the probability limit, an observed case will be predicted occurrence when the predicted probability exceeds 0.5, else it will be predicted not occurrence. There are 1955 observed cases predicted correctly, which account for 77.4% in all the 2526 sample cases.

4. Conclusion Analysis and Prospect

According to the partial contribution to the occurrence of haze, in the six independent variables included in the model, season, relative humidity, temperature, pressure, wind direction, wind speed are in descending order.

1) According to the estimation of the coefficient of the season factor, the occurrence ratio of haze weather in winter, 44.67%, is the largest. And the ratio in summer, 22.80% is the least.

From the frequency polygon, in Beijing the PM2.5 in winter was significantly higher than that in summer. On one hand, the weather in winter in the Beijing area is dry, and the traffic is causing a lot of particulates that are difficult to settle. On the other hand, coal-fired heating usually lasts for four months from November 15 to March 15 in the next year, which results in a sharp increase of the precursor pollutants like SO_2, NO_x and others.

2) In terms of controlling other factors, the odds of haze increased by 48 percent averagely every time the relative humidity increased by 10 percent. The concentration of PM2.5 rises significantly when relative humidity increase from a low start. This is mainly because when the relative humidity decreases, the

surface adsorption force of the pollutant becomes smaller, it is not easy to coagulate or settle, the particles that have been condensed under the force of wind can break down.

3) This paper uses the two-classification Logistic regression model to carry out the study of haze weather in Beijing. The dependent variable is the occurrence of the haze. According to China's current environmental air quality standards, the air quality can be divided into six grades, which are good, moderate, light pollution, moderate pollution, serve pollution and hazardous in turn. The cumulative logistic model can be used to analyze and compare the probability of the occurrence of every kind of haze weather.

4) The variables chosen in this article are season factors and various meteorological factors. Besides, the formation of fine particulate matter is affected by the precursor pollutants and human activities such as coal and motor vehicle exhaust. Restricted to data acquisition, the independent variables selected are relatively limited, which need to be more comprehensive.

5) The monitoring work has been operated actively in recent years, though the available monitoring data is still not perfect as the study of PM2.5 pollution in China start relatively late. Here, the observed data in the whole year of 2016 is applied, however; the two classifications of the Logistic Regression Model can analyze the occurrence probability of the haze weather more accurately if it is extended and detailed.

References

[1] Wang, Y.S., Yao, L., Wang, L.L., Liu, Z.R., Ji, D.S. and Tang, G.Q. (2014) Mechanism for the Formation of the January 2013 Heavy Haze Pollution Episode over Central and Eastern China. *Science China (Earth Sciences)*, **1**, 14-25. https://doi.org/10.1007/s11430-013-4773-4

[2] Sun, W. and Sun, J.Y. (2016) Daily PM 2.5 Concentration Prediction Based on Principal Component Analysis and LSSVM Optimized by Cuckoo Search Algorithm. *Journal of Environmental Management*, **188**, 144-152. https://doi.org/10.1016/j.jenvman.2016.12.011

[3] Wang, J.C. and Guo, Z.G. (2001) Logistic Regression Model-Methods and Application. Higher Education Press, Beijing.

[4] Guo, L.J., Guo, X.L., Fang, C.G. and Zhu, S.C. (2015) Observation Analysis on Characteristics of Formation, Evolution and Transition of a Long-Lasting Severe Fog and Haze Episode in North China. *Science China Earth Sciences*, **3**, 329-344. https://doi.org/10.1007/s11430-014-4924-2

[5] Jayamurugan, R., Kumaravel, B., Palanivelraja, S., Chockalingam, M.P. and Mavroidis, I. (2013) Influence of Temperature, Relative Humidity and Seasonal Variability on Ambient Air Quality in a Coastal Urban Area. *International Journal of Atmospheric Sciences*.

[6] Zhang, Y.J., Guo, Y.M., Li, G.X., Zhou, J., Jin, X.B., Wang, W.Y. and Pan, X.C. (2012) The Spatial Characteristics of Ambient Particulate Matter and Daily Mortality in the Urban Area of Beijing, China. *Science of the Total Environment*, **435-436**, 14–20. https://doi.org/10.1016/j.scitotenv.2012.06.092