Study on Haze Pollution in Northeast China Based on Data Mining Method

Yi Guan and Lei Tan

Dalian Neusoft University of Information, Dalian

guanyi@neusoft.edu.cn; tanlei@neusoft.edu.cn

Keywords: Haze pollution; Data mining methed; Big data

Abstract. In recent years, the severe fog and haze weather in many areas of our country has made a tremendous impact on people's lives. According to the air quality and meteorological data of Beijing in recent years on the Internet, the occurrence of haze is predicted, and the role of various factors in haze prediction is analyzed. A variety of classification models including BP neural network are established, and the model is trained by cross-validation and the prediction results are obtained. Select different attribute groups for classification, combined with ROC curve, accuracy and other evaluation criteria, analyze the impact of different attribute groups on haze weather, and get the relationship between heating, transportation and haze weather. This work can provide theoretical support for haze prevention and control.

In recent years, haze has seriously affected people's daily life. As we all know, the formation of fog and haze weather is closely related to not only meteorological reasons, but also pollution gas emissions, topography and other factors. As far as Beijing is concerned, the consumption of coal-fired heating in winter is huge, and the number of motor vehicles keeps rising, which is an important source of pollution. Chai Jing pointed out in "Under the Sky" in 2015 that "the biggest source of pollution in Beijing is motor vehicles", and a research group of the Institute of Atmospheric Physics of the Chinese Academy of Sciences published a set of data: the three main sources of PM2.5 are soil dust (15%), coal combustion (18%) and biomass combustion (12%), but only motor vehicle exhaust gas. 4%. It has been a controversial topic how much pollution factors play a role in haze generation, especially the impact of vehicle exhaust and heating on haze. In recent years, many scholars have used a variety of non-linear models to predict and analyze haze weather, but there is no systematic analysis and Research on the above controversial topics. Based on network data and data mining methods, this paper analyses and evaluates the causes of haze, especially the effects of vehicle exhaust and heating pollution on the formation of haze weather [1].

This paper presents a new method to forecast and analyze haze meteorological data based on various classification algorithms. By calculating the classification accuracy of different attribute groups and ROC (Receiver Operating Characteristic Curve) curves, the different causes of haze are evaluated, and the effects of vehicle exhaust and heating exhaust on haze weather are analyzed. The influence of formation. This method chooses three attribute groups which are related to vehicle exhaust, heating and weather. First, BP neural network is used as classifier to forecast haze weather through cross validation, and ROC curves of different attribute groups are drawn to evaluate the impact of each attribute group on classification[2]. Then, C4.5, RIPPER and K nearest neighbors are used to evaluate the impact of each attribute group on classification. Various classification algorithms, such as SVM, random forest and so on, are used to classify and forecast, and the accuracy of classification results and the area of ROC curve are analyzed. Through the above process analysis, it is concluded that the influence of vehicle exhaust and coal-fired heating on haze weather in Beijing is greater, and the effect is comparable.

Research Status at Home and Abroad

In recent years, many scholars have used non-linear analysis and prediction methods to predict haze weather. Among them, the neural network method is used to predict haze and build a model. This

  DOI: 10.25236/ciais.2019.001

method is suitable for the analysis and prediction of non-linear feature objects, and the network has the characteristics of self-learning ability and good robustness. Ai Hongfu and Shi Ying used BP neural network to modify the connection weight and threshold of the hidden layer in the network, so that the accuracy of network prediction and analysis can be maintained even when the haze weather index is relatively single. Ma Chu-yan, Zu Jian, Fu Qing-wan and Luo Lingxiao designed BP neural network optimization based on genetic algorithm. The problem of local minimization and flat area during network training improves the effectiveness of haze air visibility prediction model.

Analytical Methods

Structural Equation Modeling (SEM) model is also called structural equation model. Its essence is to use generalized linear equation to express the statistical method of causality between latent variables and explicit variables.

$$Lv(T_1) = w_{1i}\sum_{i=1}^{3} A_{1i} f_{1i} + w_{3i}\sum_{i=1}^{3} A_{3i} f_{3i}$$

$$Lv(T_2) = w_{2i}\sum_{i=1}^{3} A_{2i} f_{2i} + w_{3i}\sum_{i=1}^{3} A_{3i} f_{3i}$$

$$Lv(T_3) = w_{1i}\sum_{i=1}^{3} A_{1i} f_{1i} + w_{2i}\sum_{i=1}^{3} A_{2i} f_{2i} + w_{3i}\sum_{i=1}^{3} A_{3i} f_{3i}$$

Among them, explicit variables can also be called measurable variables, which often represent the objective essential factors of the object of analysis; latent variables can also be called unmeasurable variables, because many indicators in social and economic life can not be quantified, but they exist objectively. According to these properties, the concept of latent variables is introduced to specify the variables of unmeasurable indicators. To express the potential factors of the object of analysis is in line with the objectivity of the realistic economic and social research, and to improve the accuracy and comprehensiveness of the actual analysis[3]. The advantage of structural equation model is that it introduces the concepts of latent variable and explicit variable, can analyze the model of unmeasurable variable, and can study the endogeneity of variables and the possible path dependence. It is very suitable for the study and analysis of haze pollution, a multifaceted and complex natural phenomenon.

Big data refers to data sets with large capacity, multiple types, fast access speed and high application value. Big data processing technology mainly includes large data collection, integration, mining and analysis, visual analysis and prediction analysis.

Table 1 Structural equation model index of haze pollution degree

Indirect Influencing Factors	Potential Variables	Variable Names	Observation Variables	Variable Names	Units
	Regional Development	LV_1	Regional Gross Domestic Product	A_{11}	billion yuan
			Additional value of secondary industry	A_{12}	billion yuan
			Urban area	A_{13}	km2
	Resident Life	LV_2	Resident Electricity Consumption	A_{21}	billion KWh
			Number of Civil Vehicles	A_{22}	Ten thousand vehicles
			Regional population	A_{23}	ten thousand people
Direct Impact Factors	Air Pollution	LV_3	Carbon Dioxide	A_{31}	ten thousand tons
			Nitrogen oxides	A_{32}	ten thousand tons
			Smoke (powder) dust	A_{33}	ten thousand tons

Construction of Index System

In order to meet the actual needs of haze pollution analysis in China, and in combination with the situation in Northeast China, this paper establishes a scientific index system based on the main

components and social factors of haze. The research shows that in the process of evaluating haze pollution, the influence factors of haze include not only measurable indicators, but also unmeasurable indicators. Therefore, on the basis of previous studies, the research should be improved by using potential variable indicators and observation variable indicators[4]. In the process of establishing haze pollution indicators system, science must be followed. Principles of sex, systematicness, comparability and operability.

Model Setting. In this paper, the index system of structural equation model of haze pollution is grouped. The purpose of the evaluation is "the impact of regional development on haze pollution", "the impact of residents'lives on haze pollution" and "the comprehensive impact". There are corresponding explicit and latent variables. The results are shown in Table 2.

Table 2 Objective of haze pollution assessment

Evaluation Purpose	Explicit Variables	Latent Variables
Impact of regional development on haze pollution	A_{11},A_{12},A_{13}.A_{31}.A_{32}.A_{33}	LV_1.LV_3
Impact of Residents'Life on Haze Pollution	A_{21},A_{22},A_{23}..A_{31}.A_{32}.A_{33}	LV_2.LV_3
Comprehensive impact	A_{11},A_{12},$A_{13.}$ A_{21},A_{22},A_{23}..A_{31}.A_{32}.A_{33}	LV_1. LV_2.LV_3

Based on the above assumptions and the purpose of evaluation, the following models are established to analyze the haze pollution level:

$$x=\Lambda x\xi+\delta \quad (1)$$

$$y=\Lambda y\eta+\varepsilon \quad (2)$$

$$\eta=B\eta+\Gamma\xi+\zeta \quad (3)$$

Formula (1) denotes the measurement equation of exogenous variables of haze pollution level, x is an explicit set of variables consisting of "SO2", "NOX" and "smoke (powder) dust". Zeta denotes the vector set of "air quality". _x denotes the factor load matrix of X on zeta, and delta is the error of the measurement equation of exogenous variables. Formula (2) represents the measurement equation of endogenous variables of haze pollution level. y is an explicit variable set consisting of "gross domestic product", "added value of secondary industry", "urban area", "residential electricity consumption", "number of civil vehicles" and "regional population". _represents "regional development", "residential life". ” The set of vectors consisting of Y represents the factor load matrix of Y on_, and is the error of the measurement equation of endogenous variables. Formula (3) describes the causal relationship between endogenous latent variables and exogenous latent variables in the structural equation model[5]. According to the above conditions, is a relational matrix representing the interaction between endogenous latent variables_, while_is a relational matrix representing the influence of exogenous latent variables_on endogenous latent variables_. _can not be explained by the equation. The part.

Model Fitting Results. To sum up the impact of regional development on haze pollution, the structural equation model of the impact of regional development on haze pollution is established, and the results are shown in Figure 1.

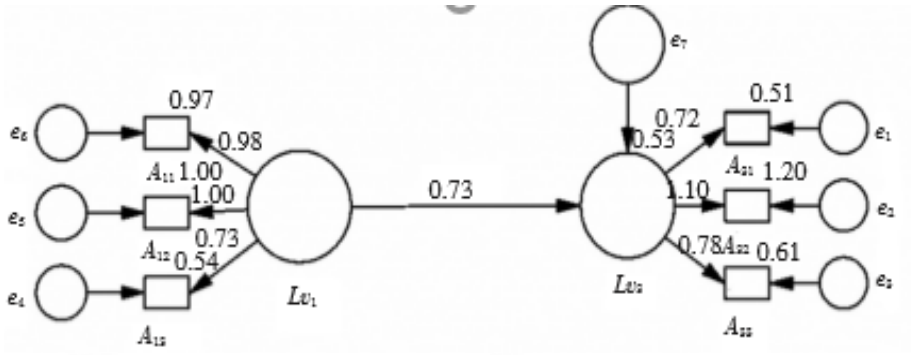

Figure 1. Structural Equation Model of the Impact of Regional Development on Haze

Conclusion

The haze pollution caused by residents'life is mainly divided into three aspects:

The higher the resident's demand for electricity, the higher the requirement for regional power generation level. In Northeast China, thermal power is still the main power generation, and a large number of electricity demand will cause serious pollution to the atmosphere.

With the continuous improvement of the quality of life of residents in Northeast China, the use of civilian cars is also increasing, which not only causes the urban traffic burden, but also causes serious haze pollution.

With the influx of a large number of people into the economically developed areas, not only a large number of social problems have arisen, but also a greater burden on the environment of the region. Therefore, the environmental pollution caused by residents'lives can not be ignored.

Suggestions for Improvement

Use Big Data to Monitor and Crack down on Violations of Laws and Regulations. Big data monitoring is mainly through the establishment of a complete air quality monitoring system, integrating the information of relevant departments such as transportation, industry and commerce, quality supervision and other departments about pollution source enterprises, combining with the information reported by the public, and then using certain analysis methods to analyze and model these data intelligently, to find out the key points of haze pollution, and accordingly put forward. Some effective measures to control haze pollution.

At present, China's air quality monitoring system has been initially formed, but there is still much room for improvement. Generally speaking, the air quality monitoring system can monitor the air pollution situation from multiple dimensions, so as to have a thorough understanding of the air quality of the whole society, develop more measures to help control the haze pollution, and make the haze pollution control more efficient. The main applications include the following three aspects[6].

From the point of view of pollutant discharge, we use big data to measure the total amount of pollutant discharge in different regions, and then formulate the strategy of haze control.

From the point of view of the industry, the emission standards of air pollutants in various industries should be formulated.

From the point of view of a single enterprise, we should deal with illegal and irregular acts.

Use Big Data to Forecast and Control Haze Pertinently. Large data can be used to predict air quality more accurately in the future. Large data analysis and prediction is to continuously superimpose the velocity and composition of pollutants in the atmosphere on the pollutant emission curve. According to the superposition results of pollutant emission curve, the air quality in the next few days can be predicted, and the severe polluted weather that may occur in the future can be more accurately identified. Moreover, the use of large data analysis can determine which pollutants are

the most important sources of pollution. Furthermore, it can also determine which industries the pollutants mainly come from. In view of the main sources of pollution, the most effective prevention and control measures are put forward to reduce the emission of major pollutants and gradually alleviate air pollution before the emergence of heavily polluted weather. Effectively prevent the arrival of heavily polluted weather. If the pollutants mainly come from the transport industry, the restriction policy of motor vehicles can be temporarily changed to reduce emissions; if the pollutants mainly come from the iron and steel industry, the Annex steel plant can be forced to shut down for a few days or shut down enterprises that exceed the emission standards.

Use Traffic Information Provided by Big Data to Travel Intelligently and Reduce Environmental Pressure. Literature studies show that vehicle exhaust emissions are one of the important sources of haze pollution. During the rush hour of commuting in big cities, the emission of vehicle exhaust caused by the increase of driving vehicles and congestion is very huge, because the traffic demand is inevitable, so the emission of vehicle exhaust can only be reduced by alleviating traffic congestion. Big data can provide real-time traffic information. Residents can get this information through the Internet, choose intelligent travel, and avoid congestion in time[7]. Big data can also be used to estimate the speed, flow and emission of each road in history, predict the speed, flow and emission of each road at a certain time in the future through machine learning and data mining. It can not only provide reference for residents'travel choice, but also provide reference for relevant departments to deal with traffic congestion and environmental pollution. For example, to optimize the urban road network, establish tidal lanes, adjust restriction policies and so on. Further, restriction policies can be specific to a certain route or region.

References

[1] F.H.Ai and Y.Shi. Prediction of haze weather based on BP artificial neural network,J.Computer simulation.32(2015) 402-405.

[2]Z.J. Ma, J.Zu, Q.P.Fu and L.X.Luo. Air visibility prediction based on genetic neural network model, J. Journal of Environmental Engineering.4(2015)1905-1910.

[3]Q.H.Hou and H.Yang. Haze weather analysis and prediction based on cubic exponential smoothing model,J.Environmental Protection Science.2(2014)73-77.

[4] W.G.Yang, L.H.Lin and L.Q.Tian. Analysis and prediction of haze weather based on wavelet analysis,J. Journal of Shaanxi University of Science and Technology (Natural Science Edition).34(2016) 166-170.

[5]W.Z.Zhou. Big data provide a technological platform and path for realizing the modernization of national governance,J. China Economic and Trade Guide, 16(2016)36.

[6]http://news.163.com/special/dbdwm/

[7]http://finance.cnr.cn/jjpl/20151117/t20151117_520524845.shtml