ISSN 1 746-7233, England, UK
World Journal of Modelling and Simulation
Vol. 13 (2017) No. 1, pp. 12-18

Prediction of air pollution in Changchun based on OSR method*

Shuai Fu[1], Yong Jiang[2†], Shiqi Xu[3], Kai Zhao[1,2], Yi Jiang[1,2]

[1] Institute of Space Weather, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[2] School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[3] Jilin Climate Center, Changchun, 130062, China

(*Received May 14 2016, Accepted October 10 2016*)

Abstract. Applying the Optimal Subset Regression (OSR) method, the forecasting equations of air quality index (AQI) and pollutant ($PM_{2.5}$, PM_{10}, O_3, NO_2) concentrations are preliminary established for Changchun, China. Besides the simultaneous meteorological elements, adding the previous day's pollutant concentrations could make the regression equations more stable and accurate. However, deviation still exists between the forecasts and observations, especially in the extreme cases.

Keywords: optimal subset regression, prediction, error analysis

1 Introduction

With the high development of social economy and acute aggravation of city population, cities have been rapidly expanding, and the consumption of energy as well as the emission of pollutants are gradually growing. Serious air pollution are threatening our health[11]. How to prevent and control air pollution has become the focus of the general public. Air pollution prediction is a hot and difficult point in the field of environmental science[6].

The primary issue of preventing or controlling air pollution is the scientific understanding of it. Research work has confirmed meteorological elements (such as surface pressure, precipitation, wind speed and direction, temperature and so on) and atmospheric circulations could usually affect air quality[2–4, 9, 10, 12, 13, 15]. For example, the temperature inversion layer inhibits the diffusion of pollutants; rainfall has a role in the erosion of pollutants; wind speed affects the diffusion rate of pollutants and wind direction controls the influenced range.

Changchun is a famous national forest city of China, whose forest coverage rate is as high as 30.66%[7]. However, its air quality is getting worse in recent years, and it has boarded the "worst air quality ranking" for a few times[1, 5]. It is now urgent to control air pollution and provide effective air quality forecasting. To achieve this goal, we adopt OSR method and build forecasting equations of AQI and pollutant ($PM_{2.5}$, PM_{10}, O_3, NO_2) concentrations.

* National Natural Science Foundation of China (Grant No. 41404053, 41174165), Special Project for Meteo-scientific Research in the Public Interest (Grant No. GYHY201306073-2), Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China (Grant No. 14KJB170012)
† Corresponding author. *E-mail address*: jiang@nuist.edu.cn

2 Data and OSR

2.1 Data

(1) Daily observations from 10 automatic monitoring stations of environmental air quality, including AQI and pollutant concentrations ($PM_{2.5}$, PM_{10}, O_3, NO_2). The daily average is regarded as the representation of Changchun. AQI has 6 levels, from grade 1 to grade 6, respectively corresponding to excellent ($0 \leq AQI \leq 50$), good ($50 \leq P \leq 100$), slightly polluted ($100 \leq AQI < 150$), moderately polluted ($150 \leq AQI < 200$), heavily polluted ($250 \leq AQI < 300$) and severely polluted ($AQI \geq 300$). The greater AQI, the worse air quality. (2) Surface meteorological elements (pressure, precipitation, relative humidity, wind speed, mean-maximum-minimum temperature, temperature difference) obtained from Jilin Climate Center. Temperature difference is defined as the maximum minus minimum temperature. A way of coding daily precipitation is adopted, that is, no rain marked as 0, light rain marked as 1, moderate rain marked as 2, heavy rain marked as 3, and rainstorm and above marked as 4. Studies have indicated that such a processing is beneficial to the actual operation in conventional forecast[14].

2.2 Description of OSRs

Air quality tends to be affected by many factors as well as the interaction between each factor, such as meteorological elements referred above. In order to weaken such a interaction effect and then establish forecast equation reasonably, this paper applies the Optimal Subset Regression (OSR) method to filter and combine factors.

The principle of OSR is: assuming m independent variables, and the number of the arbitrary combination (out of the empty set) of the m independent variables should be:

$$\sum_{k=1}^{m} c_m^k = 2^m - 1. \tag{1}$$

The target of OSR is to decide the best regression equation from all the possible subsets. First of all, we utilize Furnial-Wilson algorithm to get all the possible subsets, and then determine the best one with the Couple Score Criterion (CSC). After getting the optimal equation, the fitting and prediction value is also calculated. Besides, to assess the accuracy of the equation, error analysis is quite necessary.

Both the prediction tendency and quantity are taken into consideration by CSC. It is composed of two parts (trend score and quantity score). Assuming that a subset contains k predictors, and CSCk is calculated as follows.

$$CSC_k = S_1 + S_2, \tag{2}$$

$$S_1 = nR^2 = n(1 - \frac{Q_k}{Q_y}), \tag{3}$$

$$S_2 = 2I = 2[\sum_{i=1}^{I}\sum_{j=1}^{I} n_{ij} + n\ln n - (\sum_{i=1}^{I} n_i \ln n_i + \sum_{j=1}^{I} n_j \ln n_j)], \tag{4}$$

where, S_1 and S_2 respectively represents quantity score (fine score) and trend score (raw score); n is the sample length; I is the forecast trend type.

$$Q_k = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2, \tag{5}$$

$$Q_y = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y}_t)^2, \tag{6}$$

$$n_{,j} = \sum_{i=1}^{I} n_{ij}, \tag{7}$$

$$n_{i,} = \sum_{j=1}^{I} n_{ij}, \tag{8}$$

among them, Q_k is the residual sum of squares; Q_y is the climatological forecast; n_{ij} is the numbers of contingency tables.

CSC is designed to realize better fitting and accurate trend forecast. Obviously, when the CSC_k reaches its maximum, the corresponding subset is exactly the optimal one[8].

3 Characteristics of air pollution in changchun

3.1 Air quality situation

Of all the observed data, 91 excellent air quality days (12.5%), 403 good air quality days (55.2%), 141 slightly polluted days (19.3%), 53 moderately polluted days (7.2%), 32 heavily polluted days (4.4%) and 10 severely polluted days (1.4%) are included. In general, its air quality is good, and the probability of heavy and severe pollution is rather low (only 5.8%).

A significant feature of AQI and pollutant concentration variations is the semi-annual variations (see Fig. 1), which is usually higher in the winter (from October to March of the following year) half year than that in the summer half year (from April to September) (O_3 is inverted). Seen from Tab. 1, all the heavily polluted days take place in winter half year; and 10 severely polluted days mainly concentrate in October, November and December (only 1 day in May); the air quality below grade 2 (i.e. grade 1 and 2) is 176 and 318 days for the half year of winter and summer, respectively. Obviously, the air quality of Changchun is better in the summer half year. As a result, we separately establish forecast models for the winter and summer half year.

Table 1: Statistics of monthly air quality (unit: day)

Air quality	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Excellent	0	2	1	2	17	11	13	19	19	5	1	1
Good	23	31	42	36	36	43	40	41	41	20	24	26
Slightly polluted	20	9	14	21	6	6	7	2	0	13	22	21
Moderately polluted	10	8	4	1	2	0	2	0	0	12	6	8
Heavily polluted	9	6	1	0	0	0	0	0	0	7	4	5
Severely polluted	0	0	0	0	1	0	0	0	0	5	3	1

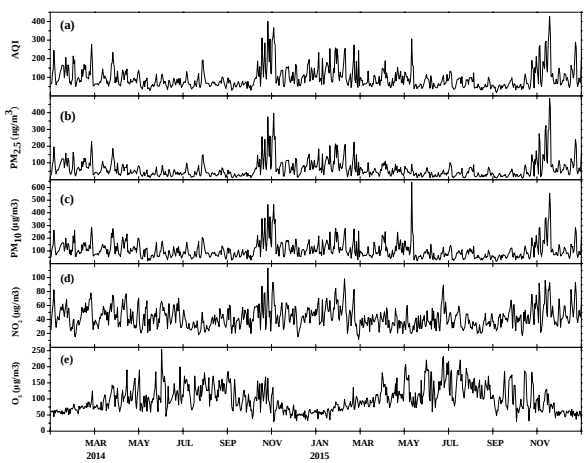

Fig. 1: Daily variation of air pollution in Changchun

3.2 Primary pollutant analysis

Daily primary pollutant is also investigated. Results show that, of all the primary pollutants days, $PM_{2.5}$, PM_{10}, O_3 and NO_2 accounts for 44.5%, 31.3%, 17.0% and 5.2%, respectively. $PM_{2.5}$ is the dominant pollutant of Changchun. Besides, it has a significant positive correlation between the concentration of $PM_{2.5}$, PM_{10}, NO_2 and AQI (through the significant test at 0.001 level) (see Tab. 2). And the correlation between the concentration of $PM_{2.5}$ and AQI is the maximum (0.952). From the primary pollutant point, the influence of $PM_{2.5}$ should be fully considered during atmospheric control and prevention.

Table 2: The correlation between the pollutants

Pollutants	AQI	$PM_{2.5}$	PM_{10}	O_3	NO_2
AQI	1	0.952^{**}	0.946^{**}	-0.054	0.692^{**}
$PM_{2.5}$	–	1	0.893^{**}	-0.113^{*}	0.709^{**}
PM_{10}	–	–	1	0.012	0.634^{**}
O_3	–	–	–	1	-0.016

Note: * passes the significant test at 0.01 level, ** passes the significant test at 0.001 level.

4 Prediction models for air pollution

Various pollutants into the atmosphere are bound to be affected by the atmospheric turbulence, turbulent diffusion and atmospheric turbulence. And atmospheric pollutants could be transported, mixed and diluted. As a result, meteorological condition has an important role in determining the air quality. Here, 8 conventional ground meteorological elements (i.e., pressure, precipitation, relative humidity, wind speed, mean-maximum-minimum temperature, and temperature difference) are chosen to examine their correlations with AQI and air pollutants and further establish prediction model.

Table 3: The correlation coefficients between AQI, pollutant concentrations and the surface meteorological elements

Period		Pressure (X_1)	Precipitation (X_2)	Relative humidity (X_3)	Wind speed (X_4)	Mean temperature (X_5)	Maximum temperature (X_6)	Minimum temperature (X_7)	Temperature difference (X_8)
Winter	AQI	-0.083^{*}	-0.007	0.191^{**}	-0.115^{*}	0.135^{**}	0.133^{**}	0.127^{*}	0.047
	$PM_{2.5}$	-0.229^{**}	-0.012	0.202^{**}	-0.174^{**}	0.126^{*}	0.114^{*}	0.122^{*}	0.002
	PM_{10}	-0.015	-0.024	0.07	-0.028	0.245^{**}	0.244^{**}	0.235^{**}	0.076
	O_3	-0.250^{**}	-0.17^{**}	0.345^{**}	0.223^{**}	0.651^{**}	0.670^{**}	0.598^{**}	0.345^{**}
	NO_2	-0.248^{**}	-0.161^{**}	0.141^{**}	-0.450^{**}	0.101^{*}	0.108^{*}	0.051	0.186^{**}
Summer	AQI	0.029	-0.269^{**}	-0.461^{**}	-0.219^{**}	0.086^{*}	0.125^{*}	-0.004	0.217^{**}
	$PM_{2.5}$	0.069	-0.17^{**}	-0.258^{**}	-0.281^{**}	0.092^{*}	0.102^{*}	0.045	0.088^{*}
	PM_{10}	0.07	-0.26^{**}	-0.482^{**}	-0.22^{**}	0.003	0.052	-0.084^{*}	0.241^{**}
	O_3	-0.128^{*}	-0.269^{**}	-0.235^{**}	-0.228^{**}	0.608^{**}	0.626^{**}	0.484^{**}	0.156^{**}
	NO_2	0.243^{**}	-0.25^{**}	-0.216^{**}	-0.486^{**}	0.082^{*}	0.164^{**}	-0.089^{*}	0.438^{**}

As shown in Tab. 3, in the winter half year, AQI, air pollutants have positive correlations with relative humidity and temperature (including mean, maximum, minimum and difference temperature), but negative correlations with pressure, precipitation, wind speed. Considering the physical process, when the air temperature and humidity are low, surface pressure field is mainly controlled by the huge clod high pressure and divergent airflow benefits the spread of pollutants and degradation. Besides, flushing action of precipitation could further dilute pollutants. On the other hand, with the higher temperature and humidity in the winter half year, surface pressure field is mainly controlled by the warm low pressure and convergent airflow inhibits the spread of pollutants. For the summer half year, AQI and air pollutants have stable negative correlations with precipitation, relative humidity, wind speed, but positive correlations with pressure and temperature in most conditions.

Analysis above shows that atmospheric condition indeed has significant influences on air quality. Using OSR method and setting the 8 meteorological elements as the forecast factors, the winter (from October to March of the following year) and summer (from April to September) half year forecasting equations for AQI, $PM_{2.5}$, PM_{10}, O_3, NO_2 concentration are preliminary built, respectively. Of all the data, observations from 1 October 2014 to 31 March 2015 (1 April to 31 September 2014) are selected as the fitting samples for the winter (summer) half year, and others will be used in the prediction. To assess the accuracy of the established equations, error between the forecasts and observations is also calculated investigated.

4.1 Meteorological factors only

In this part, we just put meteorological elements as the predictor. After filtering and combining the 8 elements by OSR, we establish the fitting equations. As shown in Tab. 4, all the multiple correlation coefficients distribute between 0.50 and 0.68. By comparison, the multiple correlation coefficients in winter half year tens to be greater than that in summer half year. Table 6 assesses the fitting and prediction ability, and Type 1 refers to the equations considering meteorological factors only. It is not hard to find that both the root mean square errors and mean absolute errors are larger in the winter half year than that in the summer half year, which means that the summer equations have slightly better and more stable prediction effect and might be more conducive to the use of actual business.

Table 4: The equations between AQI, pollutant concentrations and the surface meteorological elements established by OSR

Period	AQI and pollutants (Y)	Regression equations	Multiple correlation coefficients
Winter	AQI	$Y = -1279.580 - 1.337X_1 + 1.767X_3 - 11.298X_4 + 3.472X_6$	0.66
	$PM_{2.5}$	$Y = -1561.472 - 1.628X_1 + 1.573X_3 - 15.411X_4 + 3.539X_5$	0.6
	PM_{10}	$Y = -1238.717 - 1.317X_1 + 1.541X_3 - 10.370X_4 + 4.722X_6$	0.68
	O_3	$Y = 68.732 - 12.527X_2 + 1.694X_5 + 2.001X_8$	0.54
	NO_2	$Y = -265.672 - 0.320X_1 - 5.732X_2 + 0.299X_3 - 7.863X_4 + 0.706X_5 + 0.114X_7 + 0.472X_8$	0.64
Summer	AQI	$Y = -763.894 + 0.870X_1 - 0.694X_3 - 3.063X_4 + 1.705X_5$	0.65
	$PM_{2.5}$	$Y = -606.953 + 0.683X_1 - 0.499X_3 - 2.343X_4 - 0.276X_6 + 1.454X_7$	0.5
	PM_{10}	$Y = -1388.438 + 1.548X_1 - 1.244X_3 - 5.311X_4 + 2.452X_5$	0.61
	O_3	$Y = -1088.187 + 1.145X_1 - 0.447X_3 + 5.920X_4 + 5.438X_5 - 0.651X_7$	0.52
	NO_2	$Y = 80.033 - 0.423X_3 - 8.345X_4 + 0.208X_6$	0.61

4.2 Adding previous pollutant concentration

Since the development of anything is connected with its past, the past actions not only affect the present, but even the future. It also applies to the changes of air quality. As a result, the above scheme only considering meteorological elements is apparently not enough. In this paragraph, we add previous day's pollutant concentration as a predictor and establish new equations. The correlation coefficients of AQI, $PM_{2.5}$, PM_{10}, O_3, NO_2, and their own previous day's value are 0.67, 0.69, 0.62, 0.74, 0.63, respectively. A very good self-correlation is shown. Table 5 presents the new equations adding concentration term (X_9). Obviously, all multiple correlation coefficients have been improved significantly comparing to Tab. 5. For example, the multiple correlation coefficient of AQI in the winter (summer) half year increases from 0.66 (0.65) to 0.75 (0.74). It is surprisingly found that X_9 is only the factor introduced by all equations. In Table 6, Type 2 represents the errors of new equations. All the root mean square errors and mean absolute errors are obviously reduced. Therefore, adding the previous day's pollutant condition makes the forecast more stable, and it is also quite reasonable from the sense of physical process.

Here, we present the prediction of AQI and $PM_{2.5}$ for example. In Figs. 2 and 3 , the winter period refers to January to March in 2014 and October to December in 2015, and the summer period covers April to September in 2015. Type 1 and Type 2 have the same meaning as before. For AQI, in the winter half year,

Table 5: Same as Table 4, but adds pollutant concentration of previous day

Period	AQI and pollutants (Y)	Regression equations	Multiple correlation coefficients
Winter	AQI	$Y = -299.496 + 0.319X_1 - 14.473X_2 + 1.181X_3 - 13.852X_4 + -3.374X_5 + 5.439X_6 + 0.471X_9$	0.75
	$PM_{2.5}$	$Y = -566.256 + 0.612X_1 - 12.550X_2 + 0.902X_3 - 17.593X_4 + 2.091X_6 + 0.497X_9$	0.68
	PM_{10}	$Y = 33.197 - 17.366X_2 + 1.088X_3 - 14.152X_4 - 3.308X_5 + 6.127X_6 + 0.447X_9$	0.72
	O_3	$Y = 21.050 - 12.073X_2 + 2.374X_4 + 2.713X_5 + 0.728X_6 - 2.829X_7 + 0.497X_9$	0.63
	NO_2	$Y = -58.807 + 0.102X_1 - 5.899X_2 + 0.055X3 - 7.614X_4 + 0.249X_7 + 0.391X8 + 0.487X_9$	0.7
Summer	AQI	$Y = -296.827 + 0.346X_1 - 2.229X_3 - 5.007X4 + 2.863X_5 - 2.046X_6 + 1.349X_8 + 0.588X_9$	0.74
	$PM_{2.5}$	$Y = -300.738 + 0.338X_1 - 0.258X_3 - 3.923X_4 + 1.257X_5 - 0.620X6 + 0.626X_9$	0.63
	PM_{10}	$Y = -675.300 + 0.748X_1 - 0.610X_3 - 8.051X_4 + 3.512X_5 - 2.076X7 + 0.535X_9$	0.66
	O_3	$Y = -794.801 + 0.825X_1 - 0.336X_3 + 4.914X4 + 5.770X_5 - 1.798X_7 + 0.216X_9$	0.61
	NO_2	$Y = 50.138 - 0.244X3 - 7.096X_4 + 0.695X_5 + 0.307X_6 - 0.637X_7 + 0.417X_9$	0.72

Table 6: : Error analysis of fitting and predicted samples

Period	AQI and pollutants	Fitting samples				Predicted samples			
		Root mean square error		Mean absolute error		Root mean square error		Mean absolute error	
		Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
Winter	AQI	59.988	52.206	44.083	37.839	68.977	53.484	50.491	38.166
	$PM_{2.5}$ ($\mu g/m^3$)	54.782	46.589	38.121	32.858	69.762	54.457	47.513	35.418
	PM_{10} ($\mu g/m^3$)	67.982	60.371	49.493	43.181	82.076	64.952	57.028	43.999
	O_3 ($\mu g/m^3$)	21.513	18.067	17.01	13.473	19.627	17.014	14.753	12.269
	NO_2 ($\mu g/m^3$)	12.263	9.947	9.4904	7.7764	14.682	11.0197	11.455	8.5202
Summer	AQI	22.492	17.574	16.087	12.359	32.278	27.023	22.512	17.204
	$PM_{2.5}$ ($\mu g/m^3$)	20.314	15.606	14.401	10.82	20.314	16.084	16.445	11.753
	PM_{10} ($\mu g/m^3$)	33.334	27.275	25.428	20.161	58.443	51.923	35.985	26.941
	O_3 ($\mu g/m^3$)	23.472	12.312	19.014	14.104	30.561	21.521	24.238	18.167
	NO_2 ($\mu g/m^3$)	8.5272	7.0307	6.6673	5.6781	9.6774	8.0292	7.7277	6.3159

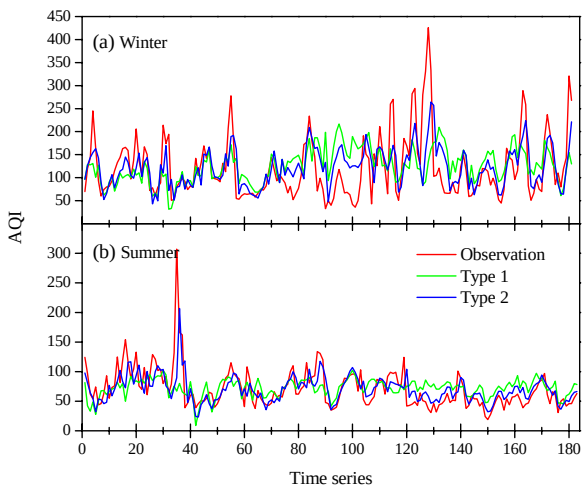

Fig. 2: Predicted and observational AQI for the winter half year (a) and the summer half year (b)

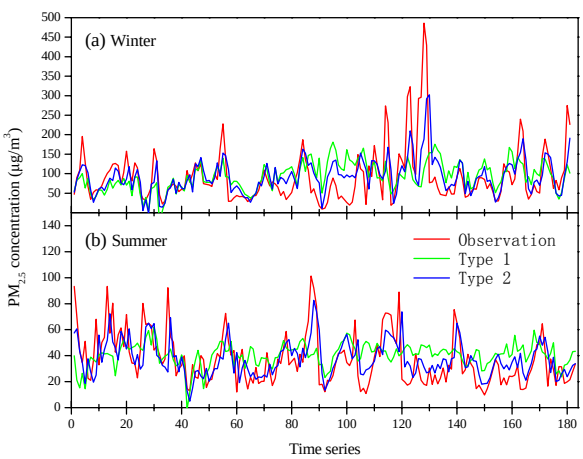

Fig. 3: Same as Fig.2, but for $PM_{2.5}$

the correlation coefficients between the observed curve and prediction curves are 0.25 and 0.61 for the type 1 and type 2, respectively; in the summer half year, the corresponding correlation coefficients are separately 0.29 and 0.59. For $PM_{2.5}$, the corresponding correlation coefficients are 0.28 and 0.64 in the winter, and 0.19 and 0.59 in the summer. Type 2 has a better correlation with the real observations. In addition, the trends of the three curves in each panels are quite consistent, but the predicting of extreme cases still exists great

difference. Anyhow, it is not to deny that adding the previous concentration makes the forecast more stable and more accurate.

5 Conclusions

Based on the daily AQI and pollutant concentrations data, as well as the simultaneous ground meteorological elements from January 2014 to December 2015, we have studied the current situation of air pollution in Changchun, China. Using 8 meteorological elements and previous pollutant concentrations as predictors, we build prediction models for AQI and pollutant concentration with OSR method. The winter and summer half year are considered separately. Results show that, previous concentration is an important predictor for the following forecast, besides meteorological elements. Adding the previous concentration makes the forecast more stable and more accurate. The trends of AQI and pollutant concentrations are well predicted, but the predicting of extreme cases still exist great difference.

References

[1] J. Ben. *The correlation analysis between the air pollution status and meteorological conditions in Changchun*. Ph.D. Thesis, Jilin University, Changchun, 2012.

[2] M. A. Cohen, S. D. Adar, R. W. Allen, E. Avol, C. L. Curl, T. Gould, D. Hardie, A. Ho, P. Kinney, T. V. Larson. Approach to estimating participant pollutant exposures in the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Environmental Science & Technology*, 2009, **43**(13): 4687–93.

[3] X. Meng, Y. U. Yu, et al. Preliminary study of the dense fog and haze events' formation over beijing-tianjin-and-hebei region in january of 2013. *Environmental Science & Technology*, 2014.

[4] Y. Meng. An analysis of air pollution and weather conditions during heavy-fog days in beijing area. *Meteorological Monthly*, 2000.

[5] Q. Zhou, S. Zhang, W. Chen. Pollution characteristics and sources of so_2, o_3 and no_x in changchun. *Research of Environmental Sciences*, 2014, **27**(7): 768–774.

[6] F. Shu. Forecasting air pollution based on the key meteorological elements and typical weather patterns in guangzhou. *Environmental Chemistry*, 2012, **31**(8): 1157–1164.

[7] W. Song. *Research on the development of tourism market in Changchun*. Ph.D. Thesis, Northeast Normal University, Haerbin, 2008.

[8] F. Wei. *Statistical diagnosis and prediction technology of modern climate (Second Edition)*. Beijing: China Meteorological Press, 2007.

[9] Y. Yang, G. Tang, et al. Effects of local circulation on atmospheric pollutants in beijing-tianjin-hebei region during summer. *Chinese Journal of Environmental Engineering*, 2015, **9**(5): 2359–2367.

[10] Z. Q. yun, W. Zhang, W. S. gong. A study on air pollution, visibility and general circulation feature. *Plateau Meteorology*, 2003.

[11] J. Zhang, K. Li, et al. Meteorological element analysis of four severe pollution processes in beijing-tianjin-hebei region. *Meteorological & Environmental Sciences*, 2016.

[12] W. Zhang, F. You, et al. Meteorological characteristics analysis of severe haze weather processes in beijing in january 2013. *Meteorological and Environmental Sciences*, 2016, **39**(2): 46–54.

[13] Y. Zhang, X. Li, et al. Analysis of pm2.5 pollution process and weather situation in beijing in february 2014. *Meteorological and Environmental Sciences*, 2016, **39**(2): 55–62.

[14] J. I. Zhong-Ping, S. B. Luo, et al. Variation characteristics and prediction of air pollution in guangzhou. *Journal of Tropical Meteorology*, 2006, **22**(6): 574–581.

[15] L. Zhou, X. Xu. The correlation factors and pollution forecast model for pm2.5 concentration in beijing area. *Acta Meteorologica Sinica*, 2003.