Study on the relationship between PM2.5 and the interaction between air pressure and temperture based on GAM model

Weiwei Xiao[1,*], *Meixia Fan*[2]

[1]School of science, North China University of Technology, Shijingshan district, Beijing 100043, China
[2]School of science, North China University of Technology, Shijingshan district, Beijing 100043, China

Abstract. This paper studies the influence of the interaction of two factors on the important air quality monitoring index PM2.5. Specifically, the GAM model based on the interaction of air pressure and temperature and the PM2.5 value is used to obtain the nonlinear relationship between air pressure, temperature and PM2.5. The GAM model has a high degree of fit, that is, the interaction of air pressure and temperature has a greater impact on the PM2.5 value. Therefore, the interaction between pressure and temperature can be used to predict the response variable PM2.5 accurately.
Keywords: GAM model;interaction;PM2.5

1 Introduction

The culprit of smog is PM2.5, and the pressure and temperature have a certain effect on the concentration of PM2.5. According to common sense, when the temperature in winter is low, the PM2.5 concentration is high, and when the temperature is high in summer, the PM2.5 concentration is relatively low, and it can be known from the analysis of the change characteristics of PM2.5 and the meteorological conditions during the pollution process. Higher air pressure is conducive to PM2.5 accumulation to form polluted weather, so the PM2.5 concentration increases [1], but during the maintenance of high concentration, it shows very different results. The formation of favorable diffusion conditions at high altitude ground pressure can quickly remove PM2.5, thereby the PM2.5 concentration reduces.

The main content of this article is to build a GAM model based on the pressure, temperature and PM2.5 in model based on the pressure, temperature and PM2.5 in the meteorological data of 45 cities. The last 4 cities use the established GAM model to make predictions, find the residuals, and use R language [2- 3] to intuitively compare the difference between the predicted value and the true value, so as to quantitatively study the influence of the interaction of influencing factors on the change of PM2.5 concentration [4].

2 The GAM model

We assume that the scalar response y_i , $i=1\cdots n$ is independent of each other, and $y_i's$ expected value is μ_i ,

$$\mu_i=\eta_i=\beta_0+\int x_{i1}(s)\xi_1(s)ds+\int x_{i2}(t)\xi_2(t)dt$$
$$+\iint x_{i1}(s)x_{i2}(t)\beta(s,t)dsdt \quad (1)$$

Among them, β_0 is the intercept term, and x_{i1} and x_{i2} are the two influenced covariates or influence variables, D and E are the interval range of two discrete observation points, the covariate value x_{i1} is the value observed in the discrete observation point $\{s_1,s_2\cdots s_j\}\subset D$, the covariate value x_{i2} is The values observed in the discrete observation point $\{t_1,t_2\cdots t_k\}\subset E$,and $\xi_1(s),\xi_2(t)$ and $\beta(s,t)$ are the regression functions corresponding to the two covariates and interaction terms.

In the linear case $y_i=\mu_i+\varepsilon_i$, we assume ε_i is an independent normal distribution with zero mean and σ^2 variance. According to Wood (2011), the sum of products is used to approximate the integral of the model (1). This model can be expressed as

$$\mu_i\approx\beta_0+h_1\sum_{j=1}^{J}x_{i1}(s_j)\xi_1(s_j)+h_2\sum_{k=1}^{K}x_{i2}(t_k)\xi_1(t_k)+$$
$$h_1h_2\sum_{j=1}^{J}\sum_{k=1}^{K}x_{i1}(s_j)x_{i2}(t_k)\beta(s_j,t_k) \quad (2)$$

Among them, h_1,h_2 is the length of the two observation points in the interval D,E , assuming that there is a regular observation grid between the two observation points.

The two main functions of the model (2) can be expanded in the spline bases, and the interaction term can be represented by the tensor product of two univariate spline bases, so the estimation method used in this paper is the smooth spline base.

* Corresponding author: 1456522632@qq.com

$$\mu_i \approx \beta_0 + h_1\sum_{j=1}^{J}\sum_{f=1}^{F} x_{i1}(s_j)b_{1f}\phi_{1f}(s_j) + h_2\sum_{k=1}^{K}\sum_{g=1}^{G} x_{i2}(t_k)$$
$$b_{2g}\phi_{2g}(t_k) + h_1h_2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} x_{i1}(s_j)x_{i2}(t_k)c_{lm}\phi_{3l}(s_j)\phi_{4m}(t_k)$$
$$(\Phi_4 \otimes \Phi_3)vec(C) \quad (3)$$

so $\xi_1(s) \approx \sum_{f=1}^{F} b_{1f}\phi_{1f}(s)$ 、 $\xi_2(t) \approx \sum_{g=1}^{G} b_{2g}\phi_{2g}(t)$ 、 $\beta(s,t) \approx \sum_{l=1}^{L}\sum_{m=1}^{M} c_{lm}\phi_{3l}(s)\phi_{4m}(t)$ and $\phi_{1f},\phi_{2g},\phi_{3l},\phi_{4m}$ are an appropriate basis function and b_{1f},b_{2g},c_{lm} is an unknown coefficient [5].

The $\boldsymbol{x}_{i1}=(x_{i1}(s_1),\ldots,x_{i1}(s_J))^T$, $\boldsymbol{x}_{i2}=(x_{i2}(t_1),\ldots,x_{i2}(t_K))^T$, $\boldsymbol{\Phi_1}=(\phi_{1f}(s_j))_{j=1,\ldots,J;f=1,\ldots,F}$, Φ_2, Φ_3, Φ_4 (similarly established) is a matrix of basis function calculations, $\boldsymbol{b_1}=(b_{11},\ldots,b_{1F})^T$, $\boldsymbol{b_2}=(b_{21},\ldots,b_{2G})^T$ $\otimes$ means the Kronecker product which is also called tensor product, $\boldsymbol{c}=(c_{11},\ldots,c_{L1},\ldots,c_{1M},\ldots,c_{LM})^T$ is column vector.

In addition, the confidence interval can be calculated as the estimated parameter function twice the estimated standard error, for example, $\beta(s_j,t_k)'s$ the confidence interval

$$\hat{\beta}(s_j,t_k) \pm 2\hat{sd}\left(\hat{\beta}(s_j,t_k)\right), \hat{sd}\left(\hat{\beta}(s_j,t_k)\right) =$$
$$\sqrt{\left(\Phi_4(t_k)\otimes\Phi_3(s_j)\right)\hat{\Sigma}_{\beta}\left({\Phi_4}^T(t_k)\otimes{\Phi_3}^T(s_j)\right)} \quad (4)$$

$\hat{\Sigma}_{\beta}$ is $\hat{c}'s$ Bayesian post-covariance matrix.

3 Modeling the interaction between PM2.5 and air pressure or temperature

Taking the time series of PM2.5 concentration, pressure and temperature in 49 cities from January 1st, 2017 to December 31st, 2017 as the research objects, the daily pressure, temperature and annual average PM2.5 concentration value in the meteorological data of 45 cities are used to establish the GAM model, the last 4 cities are used to test the GAM model, so as to study the relationship between air pressure, temperature and PM2.5 concentration change [6].

3.1 Determine whether the pressure and temperature are related to the PM2.5 concentration and whether there is a linear relationship

The GAM model was established by responding to the daily average air pressure and temperature values in 45 cities in 2017 and corresponding annual average PM2.5 concentration values [7], and using B-spline basis to obtain a smooth regression function of air pressure and temperature , and get the effect diagram of the influence of air pressure and temperature on PM2.5 concentration, at is to establish mod1 and mod2

$$\text{mod1:}\quad \mu_i=\beta_0+\int x_{i1}(s)\xi_1(s)ds \quad (5)$$
$$\text{mod2:}\quad \mu_i = \beta_0 + \int x_{i2}(t)\xi_2(t)dt \quad (6)$$

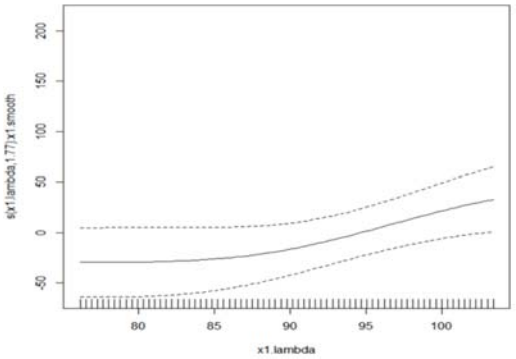

Figure 1. The effect of atmospheric pressure on PM2.5 concentration changes

Note 1: The abscissa represents the true value interval of air pressure, and the ordinate represents the smooth fitting value of air pressure to PM2.5 (smoothing the true value), the number in parentheses is the estimated degree of freedom, and the dotted line is the upper and lower limits of the confidence interval, the solid line represents the smooth fitting curve of PM2.5 concentration.

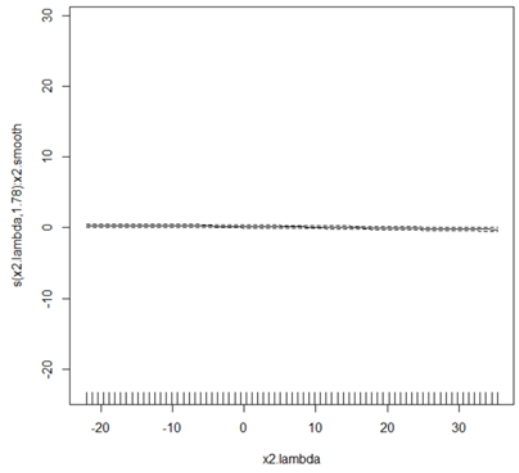

Figure 2. The effect of temperature on the PM2.5 concentration change

Note 2: The abscissa represents the true value interval of temperature, and the ordinate is the smooth fitting value of temperature to PM2.5 (smoothing the true value), the number in parentheses is the estimated degree of freedom,and the dotted line is the upper and lower limits of the confidence interval,the solid line represents the smooth fitting curve of PM2.5 concentration.

When the degree of freedom is 1, the function is a linear equation, indicating that there is a linear relationship between the influencing factors and the response variable PM2.5; when the degree of freedom is greater than 1, it means that the function is a nonlinear curve equation, and the influencing factors and the PM2.5 concentration changes. There is some kind of nonlinear

relationship, and the value is larger , the nonlinear relationship is more significant.

The results in Figure 1 and 2 show that the estimated degrees of freedom on the vertical axis are all greater than 1,so the correlation between air pressure and the PM2.5 concentration shows a nonlinear relationship [8]. When the air pressure is greater than 75kPa and less than 85kPa, the PM2.5 concentration shows a steady trend. When the air pressure is greater than 85kPa and less than 105kPa, the PM2.5 concentration shows an increasing trend. The change shows that the temperature and the PM2.5 concentration has a non-linear relationship. When the temperature gradually increases in the range of -20 to 30, the PM2.5 concentration change is small and tends to be stable. This conclusion also shows the PM2.5 concentration not only affects by temperature, but also by other factors.

3.2 Establish the GAM model of multiple influencing factors and PM2.5 response variables, the GAM model of the interaction of influencing factors and the PM2.5, and compare and test their models

3.2.1 Establish the GAM model

$$\text{mod3}\ \mu_i = \beta_0 + \int x_{i1}(s)\xi_1(s)ds + \int x_{i2}(t)\xi_2(t)dt \quad (7)$$

$$\text{mod4}\quad \mu_i = \iint x_{i1}(s)x_{i2}(t)\beta(s,t)\,ds\,dt \quad (8)$$

$$\text{mod5}\quad \mu_i = \beta_0 + \int x_{i1}(s)\xi_1(s)ds + \int x_{i2}(t)\xi_2(t)dt + \iint x_{i1}(s)x_{i2}(t)\beta(s,t)ds\,dt \quad (9)$$

```
Approximate significance of smooth terms:
                           edf  Ref.df     F p-value
s(x1.lambda):x1.smooth 0.0005409 0.001068 0.331   0.985
s(x2.lambda):x2.smooth 1.4044330 1.824071 3.022   0.125

Rank: 9/11
R-sq.(adj) =  0.0964   Deviance explained = 12.3%
-REML = 237.51  Scale est. = 1490.9    n = 49
```

Figure 3. Test results of model 3

```
Approximate significance of smooth terms:
                                                edf Ref.df    F  p-value
te(x1.lambda.matrix,x2.lambda.matrix):weights.x1.x2 7.65  8.613 16.4 9.54e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 25/26
R-sq.(adj) =  0.746   Deviance explained = 78.6%
-REML = 230.18  Scale est. = 419.79    n = 49
```

Figure 4. Test results of model 4

```
Approximate significance of smooth terms:
                                                  edf    Ref.df     F p-value
s(x1.lambda):x1.smooth                      2.070e-05 3.814e-05 0.256   0.998
s(x2.lambda):x2.smooth                      5.342e-05 9.442e-05 0.064   0.998
te(x1.lambda.matrix,x2.lambda.matrix):weights.x1.x2 7.537e+00 8.502e+00 16.320 1.51e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 33/36
R-sq.(adj) =  0.743   Deviance explained = 78.3%
-REML = 222.25  Scale est. = 424.37    n = 49
```

Figure 5. Test results of model 5

After a series of analysis of the established models, R^2 is 0.272 in model 3, but the parameters have not passed the significance test (p value is greater than 0.01), while are statistically significant, R^2 is 0.733 in model 5, and the interpretation rate is 76.9%, but it fails the significance test, so model 4 is more beneficial to study the interaction of influencing factors on the effect of the PM2.5 concentration changes.

3.2.2 Test model

Use AIC criteria to test the fit of the model

```
> AIC(mod1,mod2,mod3,mod4,mod5)
             df      AIC
mod1  3.996255 499.7219
mod2  3.833053 502.2721
mod3  3.825139 502.2898
mod4 10.612943 446.7129
mod5 10.502598 447.1605
```

Figure 6 AIC values of model 1-5

```
  Resid. Df Resid. Dev       Df Deviance Pr(>Chi)
1    45.781      65463
2    45.753      69421  0.027853    -3958
3    45.755      69468 -0.001899      -47  0.00223 **
4    38.424      16939  7.330429    52530  < 2e-16 ***
5    38.532      17171 -0.107408     -233  0.05210 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7 ANOVA analysis of the model 1-5

The AIC criterion and ANOVA analysis are both standards for measuring the goodness of model fitting. The smaller the AIC value and the residual error, the better the model fit. Therefore, according to Figures 6 and 7, in the model 4 ,the AIC value is smaller , and the residuals in the ANOVA analysis are also small, which is more favorable to show that the interpretation rate and fitting degree of Model 4 are higher, that is, the interaction of influencing factors has a higher interpretation rate for the PM2.5 concentration changes.

3.3 Use model 4 to predict the remaining 4 cities and compare them with their true values and find the residual

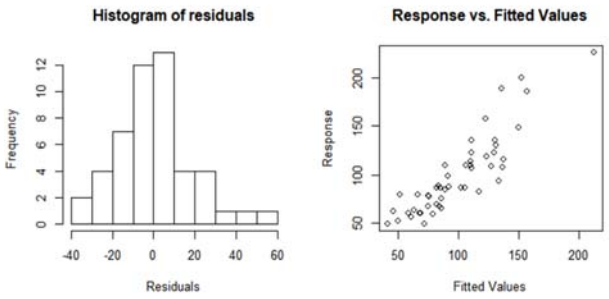

Figure 8. The relationship between the predicted value and the true value of model 4

Note 3: The left picture is the residual histogram, the right picture is the relationship between the fitted value and the true value

It can be seen from the left figure that the frequency of the residual error is between -5 and 5, and the fitted value and the true value are almost in a straight line according to the right figure. In summary, the model 4 has a high degree of fit, and the interaction between the temperature and pressure has a great influence on the PM2.5 concentration.

3.4 Perform visual drawing to get the estimated regression parameters in model 4

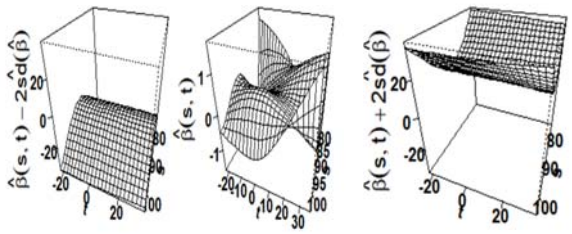

Figure 9 $\hat{\beta}(s,t)$ 3D visualization in model 4

Note 4: The left picture is $\hat{\beta}(s,t)-2\hat{sd}\left(\hat{\beta}(s,t)\right)$, the middle picture is $\hat{\beta}(s,t)$, the right picture is $\hat{\beta}(s,t)+2\hat{sd}\left(\hat{\beta}(s,t)\right)$.

Figure 9 shows $\hat{\beta}(s,t)$ and the confidence interval changes characteristics of in different dimensions. It can be seen from the middle graph that when t is in a different range, $\hat{\beta}(s,t)$ changes differently as s increases. When $t\in[-25,-15]$, as s increases, $\hat{\beta}(s,t)$ first decreases and then rises, then decreases and then rises; when $t\in[-15,10]$, as s increases, it first decreases and then rises and then decreases; when $t\in[10,35]$, as s increases, $\hat{\beta}(s,t)$ first decreases and then rises.

4 Summary

(1) This article uses the GAM model to fit each influencing factor in air pressure and temperature and the PM2.5 concentration value, and then obtains that the air pressure, temperature and PM2.5 have a non-linear relationship.

(2) By establishing the GAM model of multiple influencing factors and PM2.5 response variables, the GAM model of the interaction of influencing factors and the PM2.5 concentration and comparing and testing, it is obtained that in the model 4 R^2 is as high as 0.736, and the interpretation rate is 77.2%, The interaction of influencing factors has a higher interpretation rate for the PM2.5 concentration changes, and model 4 has a better fit, and significantly affects the PM2.5 concentration value at the level of P<0.01 (or P<0.05).

Therefore, by constructing the GAM model through the interaction of influencing factors and the PM2.5 concentration values, we can quantitatively predict and analyze the influence of the interaction between the influencing factors on the change of PM2.5 concentration.

Acknowledgement

Special fund for talents of North China University of Technology (207051360020XN140 / 004)

References

1. Yin Yanzhen, Wang Miao, Wang Jingyuan, Wang Jing, Fan Qingsheng. *PM_(10), PM_(2.5) pollution characteristics and their relationship with meteorological factors in Nanyang City*[J]. Drought Environment Monitoring， **32**(1) :12-18(2018).
2. BrianDennis. *A Beginner's Guide to R Language* [M]. Beijing: People's Posts and Telecommunications Press(2016).
3. RobertL. Kabacoff. *R language combat* [M]. (Second Edition). Beijing: People’s Posts and Telecommunications Press(2016).
4. He Xiang, Lin Zhenshan. *Based on GAM model to analyze the influence of the interaction of influencing factors on the change of PM2.5 concentration* [J]. Environmental Science, **38**(01): 22-32(2017).
5. KarenFuchs, FabianScheipl, SonjaGreven. *Penalized scalar-on-functions regression with interaction term*[J]. Computational Statistics and Data Analysis.,**81**:38–51(2015).
6. Wang Cuilian, Zhang Jun, Zheng Yao, Zhao Tongqian, Lou Yamin, Zheng Hua. *Variation characteristics of PM10 and PM2.5 mass concentration in Zhengzhou urban area and their response to meteorological factors*[J]. Environmental Protection Science, **45**(06) :76-83(2019).
7. ZhangBE,JiaoLM,XuG. *Influences of wind and precipitation on different-sized particulate matter concentrations(PM2.5,PM10,PM2. 5-10)*[J].Meteorology and Atmospheric Physics.,**130**(3) :383-392(2018).
8. Nanyang, Zhang Qianqian, Zhang Bihui. *Based on GAM model to analyze the influencing factors of long-term changes of gridded PM2.5 in typical regions of China* [J]. Environmental Science, **41**(2): 499-509(2020).