Additive Calibration Model for NO2 Based on Linear Interpolation

Yan Xu, Enhua Xu*

Guangzhou Civil Aviation College, Guangzhou 510403, China
*Corresponding author, 10000232@caac.net

Abstract. The paper proposed the additive model for NO2 considering the influence of internal and external factors. Linear interpolation filling the missing values could be effectively solved the problem of data missing and improved the effect of the additive model of ARIMA and multivariate linear regression. The additive calibration model by ARIMA and Multiple linear regression for NO2 was reconstructed based on linear interpolation filling. The error analysis showed that the accuracy of NO2 was improved. The prediction effect was also improved by considering the interaction effect.

1. Introduction

NO2 is the most important form of nitrogen oxides, and one of the main air pollutants. NO2 in the air is mainly caused by the emission of petroleum products and coal combustion. The air quality could be grasped in time and the corresponding measures could be taken for the pollution source by the real-time monitoring data. Due to the insufficiency and diseconomy of the real-time monitoring data by national controlled detector (abbreviated as NCD in the following), the self-developed micro air quality detector (abbreviated as SDD in the following) would have great practical value in application for its excellent timeliness and economy. The monitoring data of SDD may be influenced by meteorological factors such as temperature, and the monitoring errors may be occurred through the prolongation of the using time of the equipment itself, too. Therefore, we need to check and calibrate the monitoring data to improve its accuracy.

It was found that the monitoring data was mainly affected by internal and external factors based on the analysis of the monitoring data. It showed an additive relationship so that an additive calibration model was established. At the same time, the processing of missing monitoring data would also affect the accuracy. We considered two filling methods, namely mean value and linear interpolation.

The data was from the mathematical modeling competition of college students in 2019. It included the monitoring data of NO2 by NCD and SDD. Five meteorology factors, i.e. wind, pressure, precipitation, temperature, and humidity were also given. It was found that NO2 conformed to time series. ARIMA model could be used to describe the trend before and after its own data. For the influence of the meteorology factors, multiple linear regression could be used to describe the influence of the meteorology factors.

The paper was structured as follows. Part 2 was the exploratory analysis for the additive model based on ARIMA and multiple linear regression. Part 3 was the model based on mean filling. Part 4 was the model based on linear interpolation filling. Part 5 was the error analysis. The relative errors were computed and analysed. Part 6 was the conclusion.
2. Exploratory Analysis

In this part, the data from NCD was considered as the standard data. We remodeled NO2 of SDD combined with meteorological factors. We divided the variation of the dependent variable (Y) into two parts. Its internal factor (A) and the external factor (B). The internal factor was caused by its autocorrelation. The external factor was caused by meteorological factors. The two parts were additive.

\[ Y = A + B \]  

(1)

2.1. A based on ARIMA

A was the predicted value of NO2 of SDD based on ARIMA. ARIMA model was a famous time series model proposed by Box and Jenkins. It mainly included the following three forms [4].

**AR (Auto-regressive)**

\[ \Delta x_t = \sum_{j=1}^{p} \varphi_j x_{t-j} \]  

(2)

**MA (Moving-Average)**

\[ \Delta x_t = \mu + \sum_{j=1}^{q} \theta_j x_{t-j} \]  

(3)

**ARMA**

\[ \Delta x_t = \mu + \sum_{j=1}^{p} \varphi_j x_{t-j} + \sum_{j=1}^{q} \theta_j x_{t-j} \]  

(4)

Since the time interval of the monitoring data of SDD was inconsistent and the lowest common multiple was huge, it was considered that it may lead to higher bias of the model if the huge time interval was ignored.

In order to prevent this kind of situation, we adopt two kinds of missing value filling methods, namely mean filling and linear interpolation filling.

2.2. B based on multiple linear regression

Considering external meteorology factors, the difference between NCD and SDD was the dependent variable (\( \Delta = \text{NCD-SDD} \)), and meteorology factors were the independent variables (\( \text{VAR1~VAR5, i.e., wind, pressure, precipitation, temperature, humidity} \)). B was based on multiple linear regression.

We considered the simple linear regression and interactive regression model.

\[ B = \Delta = \beta_0 + \beta_1 \text{VAR1} + \beta_2 \text{VAR2} + \beta_3 \text{VAR3} + \beta_4 \text{VAR4} + \beta_5 \text{VAR5} + \beta_6 \text{VAR12} + \cdots \]  

(5)

\( \text{VAR12=VAR1*VAR2, namely the interactive effect between VAR1 and VAR2, and so on.} \)

3. Model based on mean filling

In this part, we took every five minutes of the time point as the observation point from the whole point on. The mean of the value within every five minutes was computed as the observation value of this point. Finally, 2000 time points were obtained as the samples for modeling. It was a week continuous time series data. The parameters of model were estimated by the maximum likelihood method [4].

We studied the correlation of NO2 between NCD and SDD, and the correlations between NO2 and the five meteorological factors. Then, we studied the autocorrelation of NO2.

The ACF and the PACF of NO2 showed that it was basically stable by first-order difference. So, the difference order was set as \( d=1 \). By comparing the BIC values, we got the minimum BIC (1, 2) = -3.623846 of ARIMA model when \( p=1 \) and \( q=2 \). So, ARIMA (1, 1, 2) was finally used to predict NO2 of SDD.

| Variable | estimate | SD    | F      | P       |
|----------|----------|-------|--------|---------|
| Intercept| 1654.59721| 117.47805| 14.08  | <0.0001 |
| VAR2     | -1617.98847| 117.05374| -13.82 | <0.0001 |
| VAR3     | -384.00418| 16.74956 | -22.93 | <0.0001 |
| VAR4     | -125.69277| 59.00481 | -2.13  | 0.0332  |
| VAR5     | 530.25672 | 22.76150 | -23.30 | <0.0001 |
Table 2 Parameter estimate by interactive regression for NO2

| Variable | estimate | SD     | F      | P      |
|----------|----------|--------|--------|--------|
| Intercept| 3092.68507 | 207.62961 | 221.87 | <0.0001 |
| VAR1     | -17.22810  | 3.14599  | 29.99  | <0.0001 |
| VAR2     | -2.95955   | 0.20085  | 217.12 | <0.0001 |
| VAR3     | -11.45839  | 1.34821  | 72.23  | <0.0001 |
| VAR4     | 29.47755   | 5.43094  | 29.46  | <0.0001 |
| VAR5     | -0.47346   | 0.06149  | 59.28  | <0.0001 |
| VAR13    | 0.06282    | 0.01054  | 35.55  | <0.0001 |
| VAR15    | 0.13086    | 0.03984  | 10.79  | 0.0010  |
| VAR23    | 0.01083    | 0.00130  | 69.85  | <0.0001 |
| VAR24    | -0.03103   | 0.00538  | 33.29  | <0.0001 |
| VAR34    | 0.01659    | 0.00158  | 109.84 | <0.0001 |
| VAR35    | 0.00120    | 0.00029076 | 17.10 | <0.0001 |
| VAR45    | -0.03989   | 0.00216  | 340.66 | <0.0001 |

Table 3 ANOVA

| Variation | df | SS  | MS   | F      | P      |
|-----------|----|-----|------|--------|--------|
| Model     | 12 | 1338969 | 111581 | 228.28 | <0.0001 |
| Errors    | 3402 | 1662859 | 488.7857 |        |        |
| Total     | 3414 | 3001828 | 488.7857 |        |        |

4. Model based on linear interpolation filling
In this part, we used linear interpolation to fill the missing \[5\].

\[
y = \frac{y_2 - y_1}{t_2 - t_1} (t - t_1) + y_1
\]  

(6)

Where, \(y\) and \(t\) referred to the variable value and time to be filled respectively. \(y_2\) and \(t_2\) referred to the observation value and time of monkey at a recorded time point. \(y_1\) and \(t_1\) referred to the observation value and time of the previous recorded time point.

The ACF and the PACF of NO2 showed that it was basically stable by first-order difference. So, the difference order was set as \(d=1\). By comparing the BIC values, we got the minimum BIC \((6, 4) = 5.832408\) of ARIMA model when \(p=6\) and \(q=4\). So, ARIMA \((1, 6, 4)\) was finally used to predict NO2 of SDD.

### Table 4 Maximum Likelihood Estimation for NO2

| Parameter | estimate | SD   | t    | P    | Lags |
|-----------|----------|------|------|------|------|
| MA1,1     | 0.299999 | 0.14174 | 2.12 | 0.0343 | 1    |
| MA1,2     | -0.12840 | 0.16643 | -0.77 | 0.4404 | 2    |
| MA1,3     | -0.20748 | 0.16174 | -1.28 | 0.1996 | 3    |
| MA1,4     | 0.73799 | 0.11110 | 6.64 | <0.0001 | 4    |
| AR1,1     | -0.19121 | 0.14389 | -1.33 | 0.1839 | 1    |
| AR1,2     | -0.54014 | 0.10485 | -5.15 | <0.0001 | 2    |
| AR1,3     | -0.59552 | 0.15756 | -3.78 | 0.0002 | 3    |
| AR1,4     | 0.29481 | 0.05920 | 4.98 | <0.0001 | 4    |
| AR1,5     | -0.12624 | 0.03293 | -3.83 | 0.0001 | 5    |
| AR1,6     | 0.0025022 | 0.03441 | 0.07 | 0.9420 | 6    |

### Table 5 Parameter estimate by simple linear regression for NO2

| Variable | estimate | SD     | F     | P     |
|----------|----------|--------|-------|-------|
| Intercept| 1807.60053 | 105.61760 | 292.91 | 0.0106 |
| VAR2     | -1.70875 | 0.10160 | 282.86 | <0.0001 |
| VAR3     | -0.11553 | 0.00451 | 657.17 | <0.0001 |
| VAR4     | -2.14288 | 0.11688 | 336.13 | 0.0237 |
| VAR5     | -0.74107 | 0.02437 | 924.61 | <0.0001 |

### Table 6 Parameter estimate by interactive regression for NO2

| Variable | estimate | SD     | F     | P     |
|----------|----------|--------|-------|-------|
| Intercept| 3213.83287 | 202.93329 | 250.81 | <0.0001 |
| VAR1     | -11.21872 | 2.43391 | 21.25 | <0.0001 |
| VAR2     | -3.08521 | 0.19625 | 247.13 | <0.0001 |
| VAR3     | -10.88441 | 1.27541 | 72.83 | <0.0001 |
5. Discussions

In this part, we mainly focused on the prediction validity of the model. After removing the samples for the modeling, the remaining samples were used to test the prediction precision. We compared the predictive values (PV) and the standard values (SV), and calculated the average relative error to evaluate the calibration effects.

\[ \text{Average relative error} = \frac{|PV - SV|}{SV \times n} \]  

(7)
We got the predictive values by the additive calibration models based on mean filling and linear interpolation filling. We also compared the monitoring data of SDD and ARIMA. The results were showed in Table 8.

The prediction effect of the additive calibration models was higher than that of SDD and ARIMA models. The prediction effect was improved by considering the interaction effect.

By using different filling methods, it could be found that the effect of linear interpolation filling was higher than that of mean filling.

| Model                  | NO2  |
|------------------------|------|
| Linear Interpolation Filling |      |
| Y=A+B (interactive)   | 0.2373 |
| SDD                    | 0.8108 |
| ARIMA                  | 0.8385 |
| Y=A+B (simple)         | 0.3303 |
| Mean Filling           |      |
| Y=A+B (interactive)   | 0.2407 |
| SDD                    | 0.8548 |
| ARIMA                  | 0.8440 |
| Y=A+B (simple)         | 0.3894 |

6. Conclusion
The paper proposed the additive model for NO2 considering the influence of internal and external factors. At the same time, the lack of monitoring data would lead to the poor fitting effect of the model. The paper proposed mean and linear interpolation to fill the missing data, reconstruct the model, and calculated the average relative errors.

The prediction effect of additive calibration model based linear interpolation filling for NO2 was the best. The prediction effect was improved by considering the interaction effect.

Our model still had some shortcomings to be improved. First of all, we did not explore the long-term and short-term differences of ARIMA model in predicting the observation values of SDD, and did not consider the timing of national control points. Secondly, there was no quantitative analysis and discussion on the physical factors such as zero drift and range drift of the electrochemical gas sensor that will be used for a long time [6]. That was where our study should be improved in the future.

Acknowledgments
This work was financially supported by Guangdong universities characteristic innovation project "prediction and analysis of regional differences and evolution of air emissions".(2019GKTSCX008)

References
[1] SUN Zhuling. Time series and multivariate statistical analysis, Shanghai: Shanghai Jiaotong University Press, 2016(in Chinese).
[2] Zhang Lingwei. Precision and accuracy analysis of air automatic monitoring instrument [J] environmental science guide, 2019, 38 (2): 123-129(in Chinese).
[3] Christoph Lotteraner, Martin Piringer. Mixing-Height Time Series from Operational Ceilometer Aerosol-Layer Heights [J] Boundary-Layer Meteorology, 2016,161 (2) : 265-287(in English).
[4] JIANG Jianfang, HOU Lili, WANG Xinlong. Spatial Analysis of Ground-level NO2 in China
and Its Socio-economic Factors, Ecology and Environmental Sciences, 2019, 28: 1632-1641.

[5] SONG W Z, JIA H F, et al. Detecting urban land-use configuration effects on NO2 and NO variations using geographically weighted land use regression [J]. Atmospheric Environment, 2019, 197: 166-176(in English).

[6] Gao Geng. Estimation drift of multiple linear regression and its determination method [J] statistics and decision, 2018, 14: 31-34(in Chinese).