ARIMA Based on Linear Interpolation for SO2 Monitoring Data’s Calibration

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

Abstract. 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 addictive calibration model by ARIMA and Multiple linear regression for SO2 was reconstructed based on linear interpolation filling. The error analysis showed that the accuracy of SO2 was improved. The prediction effect was also improved by considering the interaction effect.

1. Introduction
SO2 is the most common irritant sulfur oxide, one of the main pollutants in the air, and the main precursor to the formation of acid rain [1]. Industrial waste gas, civil combustion, automobile exhaust and so on will produce a large number of SO2 emissions. Real-time and accurate monitoring can effectively control and control pollution [2].

However, due to the restriction of economy and other factors, the setting of national measurement points often cannot meet the requirements of accurate fixed-point, real-time, accurate and economic monitoring. The self-developed micro air monitor has huge market value because of its flexible and economic type. The basic principle of SO2 monitoring is based on the ultraviolet light to excite SO2 molecules through filters, and produce fluorescence when it attenuates back to the basic state. The fluorescence intensity is proportional to the SO2 concentration. The SO2 concentration in the air can be calculated by measuring the fluorescence intensity after amplification by photomultiplier tube [3]. The micro air monitor has certain requirements for environmental conditions. The calibration of the instrument is often completed under the standard environmental conditions such as constant temperature and humidity. In the actual application of natural climate environment, the change of temperature, humidity, wind speed, pressure and precipitation and other meteorological factors will have a certain impact on the accuracy of its monitoring data [2]. Therefore, we need to analyse and calibrate the real environment monitoring data.

The data was from the mathematical modeling competition of college students in 2019. It included the monitoring data of SO2 by NCD and SDD. Five meteorology factors, i.e. wind, pressure, precipitation, temperature, and humidity were also given. It was found that SO2 conformed to time series. ARIMA model could be used to describe the trend before and after its own data. For the influence of the meteorological factors, multiple linear regression models could be used to describe the influence of the meteorology factors.

Our paper was structured as follows. Part 2 was the exploratory analysis for the additive model based on ARIMA and multiple linear regressions. 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 SO2 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 SO2 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) :  
\[ \Delta x_t = \mu_t + \sum_{j=1}^{q} \theta_j x_{t-j} \]  

(3)

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

(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 (\( \Delta = NCD - SDD \)), and meteorology factors were the independent variables (VAR1~VAR5, i.e., wind, pressure, precipitation, temperature, humidity). \( B \) was based on multiple linear regression.

\[ B = \Delta = \beta_0 + \beta_1 \text{VAR1} + \beta_2 \text{VAR2} + \beta_3 \text{VAR3} + \beta_4 \text{VAR4} + \beta_5 \text{VAR5} + \beta_6 \text{VAR12} + \beta_7 \text{VAR13} + \cdots \]  

(5)

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 SO2 between NCD and SDD, and the correlations between SO2 and the five meteorological factors. Then, we studied the autocorrelation of SO2.

The ACF and the PACF of SO2 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 (0, 1) = -1.2471 of ARIMA model when \( p=0 \) and \( q=1 \). So, ARIMA (0, 0, 1) was finally used to predict SO2 of SDD.

| Variable | estimate | SD   | F    | P       |
|----------|----------|------|------|---------|
| Intercept| -1573.0598| 146.4220| -10.74| <0.0001|
| VAR1     | -6552.4119| 1264.2564| -5.18 | 0.5692 |
| VAR2     | 1565.9314 | 145.9361 | 10.73 | <0.0001|
| VAR3     | 304.5493  | 20.8915  | 14.58 | <0.0001|
VAR4   -204.2688  73.7342  -3.26  0.0011
VAR5   62.0629  29.0882  2.13  0.0329

Table 2 Parameter estimate by interactive regression for SO2

| Variable | estimate      | SD      | F    | P    |
|----------|---------------|---------|------|------|
| Intercept| -1126.14871   | 156.75821| 51.61| 0.0106|
| VAR2     | 1.06752       | 0.15221 | 49.19| 0.0135|
| VAR4     | -52.72882     | 7.10628 | 55.06| <0.0001|
| VAR13    | -0.05365      | 0.00743 | 52.16| <0.0001|
| VAR23    | 0.00018256    | 0.00002329 | 61.43| 0.0237|
| VAR24    | 0.05306       | 0.00705 | 56.70| <0.0001|
| VAR25    | 0.0021574     | 0.00006612 | 10.65| 0.0001|
| VAR35    | -0.00112      | 0.00029557 | 14.32| <0.0001|
| VAR45    | 0.01173       | 0.00284 | 17.06| 0.0199|

Table 3 ANOVA

| Variation | df  | SS      | MS     | F      | P      |
|-----------|-----|---------|--------|--------|--------|
| Model     | 8   | 398448  | 49806  | 56.87  | <0.001 |
| Errors    | 3406| 2983019 | 857.81 | 286    |        |
| Total     | 3414| 3381466 |        |        |        |

Fig 1 Trend and autocorrelation analysis

Fig 2 Forecast based on ARIMA

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 SO2 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 \((2, 2) = \text{1.744715}\) of ARIMA model when \(p=2\) and \(q=2\). So, ARIMA \((2, 2, 1)\) was finally used to predict SO2 of SDD.

| Parameter | estimate | SD  | t    | P     | Lags |
|-----------|----------|-----|------|-------|------|
| MA1,1     | 0.82340  | 0.22889 | 3.60 | 0.0003 | 1    |
| MA1,2     | 0.15260  | 0.22584 | 0.68 | 0.4992 | 2    |
| AR1,1     | 0.9252   | 0.22907 | 1.28 | 0.2016 | 1    |
| AR1,2     | -0.04527 | 0.10207 | -0.44| 0.6574 | 2    |

| Parameter | estimate | SD  | F    | P     |
|-----------|----------|-----|------|-------|
| Intercept | -1446.15963 | 129.11332 | 125.46 | 0.0106 |
| VAR1      | -4.53911  | 0.94371 | 23.13 | 0.0135 |
| VAR2      | 1.38926   | 0.12426 | 125.00 | <0.0001 |
| VAR3      | 0.06010   | 0.00551 | 119.08 | <0.0001 |
| VAR4      | 1.44391   | 0.14273 | 102.34 | 0.0237 |
| VAR5      | 0.21265   | 0.03013 | 49.81  | <0.0001 |

| Parameter | estimate | SD  | F    | P     |
|-----------|----------|-----|------|-------|
| Intercept | -1139.66844 | 445.63328 | 6.54  | 0.0106 |
| VAR2      | 1.07023   | 0.43289 | 6.11  | 0.0135 |
| VAR3      | 6.53776   | 1.63305 | 16.03 | <0.0001 |
| VAR4      | -41.97867 | 6.57248 | 40.79 | <0.0001 |
| VAR5      | -11.78914 | 5.20858 | 5.12  | 0.0237 |
| VAR13     | -0.03538  | 0.00582 | 36.92 | <0.0001 |
| VAR23     | -0.00610  | 0.00157 | 15.14 | 0.0001 |

Table 4 Maximum Likelihood Estimation

Table 5 Simple Linear Regression Model of SO2

Table 6 Interactive Regression Models model of SO2
VAR24  0.04285  0.00647  43.92  <0.0001
VAR25  0.01182  0.00507  5.43   0.0199
VAR34  -0.00845  0.00191  19.66  <0.0001
VAR35  -0.00199  0.00036793 29.17  <0.0001
VAR45  0.01787  0.00465  14.78  0.0001

Table 7: ANOVA

| Variation | df  | SS     | MS    | F     | P         |
|-----------|-----|--------|-------|-------|-----------|
| Model     | 11  | 403581 | 36689 | 43.91 | <0.0001   |
| Errors    | 4122| 3443864| 835.48378|
| Total     | 4133| 3847445|

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.

\[
\text{Average relative error} = \frac{|PV - SV|}{SV} \tag{7}
\]

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.
Table 8 Average relative errors for different models

| Model                      | SO2  |
|----------------------------|------|
| Linear Interpolation Filling |      |
| Y=A+B (interactive)         | 0.3480|
| SDD                        | 0.5455|
| ARIMA                      | 0.5508|
| Y=A+B (simple)             | 0.3808|
| Mean Filling               |      |
| Y=A+B (interactive)         | 0.3186|
| SDD                        | 0.5412|
| ARIMA                      | 0.5438|
| Y=A+B (simple)             | 0.5372|

6. Conclusion
Through the exploratory analysis of SO2 monitoring data, it was found that the observation variables have certain timing and autocorrelation. At the same time, through the correlation analysis, it was also found that some correlations between the observation variables and other influencing factors.

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 SO2 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] Liu Lei, Wan ziqianhong. The impact of central environmental performance assessment on local sulfur dioxide emissions: Based on the inspection during the 11th Five Year Plan and the 12th Five Year Plan [J] China environmental management, 2019, 5:113-118 (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] Su Hang, Wang Guimei, Zhang Zhenxing, et al. Optical path research of atmospheric SO2 detection module [J] laser technology, 2019, 7:91-97 (in Chinese).
[4] WEI Peng, REN Zhenhai, SU Fuqing. Seasonal Distribution and Cause Analysis of NO2 in China, Research of Environmental Sciences, 2011, 24: 155-161 (in Chinese).
[5] 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).
[6] Gao Geng. Estimation drift of multiple linear regression and its determination method [J] statistics and decision, 2018, 14: 31-34 (in Chinese).