ARIMA and Multiple Linear Regression Additive Model for SO2 Monitoring Data's Calibration

XuYan[1,a], Lan Shuangting[2,b]*

[1]School of Humanities and Social Sciences, Guangzhou Civil Aviation College, Guangzhou, Guangdong
[2]School of Mathematics and Systems Science, Guangdong Polytechnic Normal University, Guangzhou, Guangdong

Abstract-SO2 is one of the main air pollutants produced by industrial waste gas, civil combustion and automobile exhaust. Real-time monitoring of the concentration of SO2 can grasp the air quality in time and take corresponding measures to the pollution sources. Monitoring data may be affected by the internal factors and the external factors. ARIMA was used for the internal factor as A. Meteorological factors were taken as external factors, and the difference of SO2 between the standard data and monitoring data was taken as dependent variable. Multivariate linear regression was modeled as B. Time series calibration model was obtained Y=A+B. The error analysis showed that the accuracy of SO2 was improved. The additive model could effectively calibrate SO2 monitoring data.

1 Introduction

Environmental pollution is becoming more and more serious with the development of industrialization and urbanization. Air pollution has brought a huge threat to human health. The 19th National Congress of the Communist Party of China emphasized that "Green water and green mountains are golden and silver mountains". In recent years, air quality in China has generally improved, but the situation is still grim, for example the treatment of acid rain is still very difficult. SO2 is one of the main pollutants in the air, and the main precursor to the formation of acid rain. It's also one of the three types of carcinogens published by International Cancer Research Institute of World Health Organization [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. Part2 was the exploratory analysis for the monitoring data of SDD and NCD. This part included statistical description and hypothesis testing. Part 3 was difference analysis between the two groups. This part included the correlation analysis and the autocorrelation analysis. Part 4 was the time series calibration model based on ARIMA and multiple linear regression. Part 5 was the error analysis. The relative errors were computed and analysed. Part 6 was the conclusion.

2 Exploratory analysis

In this part, We use statistical description, hypothesis test and other statistical research methods for the intercepted data in the same time period to preliminarily explore the

[a]10000583@caac.net; *Corresponding author e-mail: [b]28270031@qq.com

difference of SO2 monitoring data between NCD and SDD, as well as the relationship between SO2 and other possible influencing factors.

2.1 Statistical description

We computed the basic statistics of SO2 monitoring data by NCD (n=4130) and SDD (n=4200) in the same period.

Taking the day as the unit, we calculated the mean value of daily SO2 monitoring data to explore the seasonal variation. The trend of the daily mean value of SO2 was fluctuation (Figure 1). It was higher in autumn and winter, and lower in spring and summer. The difference between the maximum values and the minimum values were large in November, December and January. There was a high peak in January, and the values in other months were basically in a relatively low level and small fluctuation.

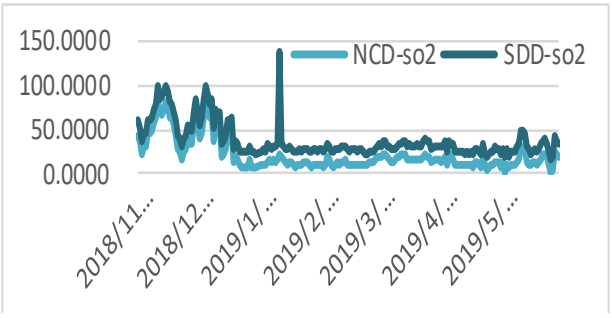

Figure 1 The daily mean value of SO2 by NCD and SDD

Taking per hour as the unit, we calculated the mean value of each whole time in hours to explore the variation in a day. The mean value per hour of SO2 showed a double-peak feature (Figure2.2). The hourly variation trend of SO2 could be seen that it was higher at 6 and 7 points in a day, then it slowly declined, the lowest at 13 and 14 points, then it slowly raised, the highest at 19 and 20 points, and then it declined again. The monitoring data of NCD was slightly higher than that of SDD.

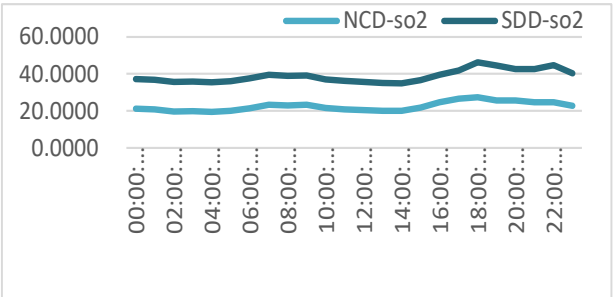

Figure 2 The mean value per hour of SO2 by NCD and SDD

2.2 Hypothesis testing

Paired t-test was used for SO between NCD and SDD. It should be noted that the data of SDD was not complete at the whole point. So, we used two methods of calculation (Table 1).

(1) The nearest point of the whole point time as the whole point monitoring data.

(2) The mean of half an hour before and after the whole point time as the whole point monitoring data.

Table 1 Paired t test for SO2 between NCD and SDD

	mean	SD	95%CI		t	P
(1)	5.7944	30.5027	4.8649	6.7239	12.22	<0.0001
(2)	5.9264	26.5621	5.1173	6.7354	14.36	<0.0001

The paired t-test showed that there were significant differences between the two groups ($P < 0.05$).

3 Difference Analysis

In this part, 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.

3.1 Correlation analysis

Correlation analysis showed that SO2 between NCD and SDD was correlated (r=0.38392, P<0.0001). They were correlated between SO2 of SDD and the meteorological factors (Table 2. 1, P<0.0001). They were negative correlations with wind, precipitation and temperature, positive correlation with pressure and humidity.

Table 2 Correlation analysis between SO2 and meteorological factors（N=234717）

	Wind	Pressure	Precipitation	Temperature	Humidity
r	-0.4901	0.2533	-0.5428	-0.4800	0.3277
P	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

3.2 Autocorrelation analysis

Autocorrelation and partial autocorrelation showed that the autocorrelation of CO was high significance (Figure 3, P<0.0001). Therefore, it was considered that CO monitoring data belong to time series data, and time series analysis methods could be used in the following study.

Autocorrelation analysis of white noise									
Lags	χ2	df	P	Autocorrelation					
6	9999.99	6	<.0001	0.960	0.918	0.880	0.845	0.818	0.798
12	9999.99	12	<.0001	0.782	0.772	0.766	0.762	0.759	0.758
18	9999.99	18	<.0001	0.759	0.761	0.762	0.764	0.765	0.768
24	9999.99	24	<.0001	0.774	0.779	0.787	0.793	0.795	0.793

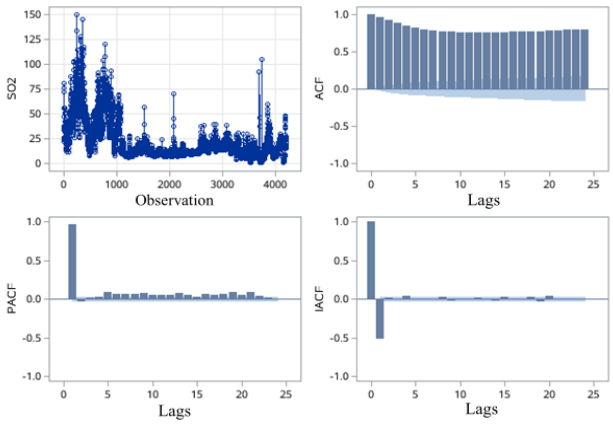

Figure 3 Trend and autocorrelation analysis of SO2

4 Time series calibration model

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.

$$Y = A + B$$

SO2 of SDD was considered as time series data. So, A was the predicted value of SO2 of SDD based on ARIMA. Considering external meteorology factors, the difference between NCD and SDD was the dependent variable (Δ=NCD-SDD), and meteorology factors were the independent variables (COL1~COL5, i.e., wind, pressure, precipitation, temperature, humidity). B was modeled based on multiple linear regression.

$$B = \Delta = \beta_0 + \beta_1 COL_1 + \beta_2 COL_2 + \beta_3 COL_3 + \beta_4 COL_4 + \beta_5 COL_5$$

4.1 A 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) : $\Delta x_t = \sum_{i=1}^{p} \varphi_i x_{t-i}$.

MA (Moving-Average) : $\Delta x_t = \mu_t + \sum_{i=1}^{q} \theta_i x_{t-i}$

ARMA: $\Delta x_t = \mu_t + \sum_{i=1}^{q} \theta_i x_{t-i} + \sum_{i=1}^{p} \varphi_i x_{t-i}$

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. To prevent it, 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].

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 (=-1.2471) of ARIMA model when p=0 and q=1. So, ARIMA (0, 1, 1) was finally used to predict SO2 of SDD. Parameter estimation was shown in Table 3, and model prediction was shown in Figure 4.

Table 3 Maximum Likelihood Estimation

Parameter	estimate	SD	t	P	Lags
MU	0.000604	0.00256	0.24	0.8132	0
MA1,1	0.78633	0.01388	56.66	<0.0001	1

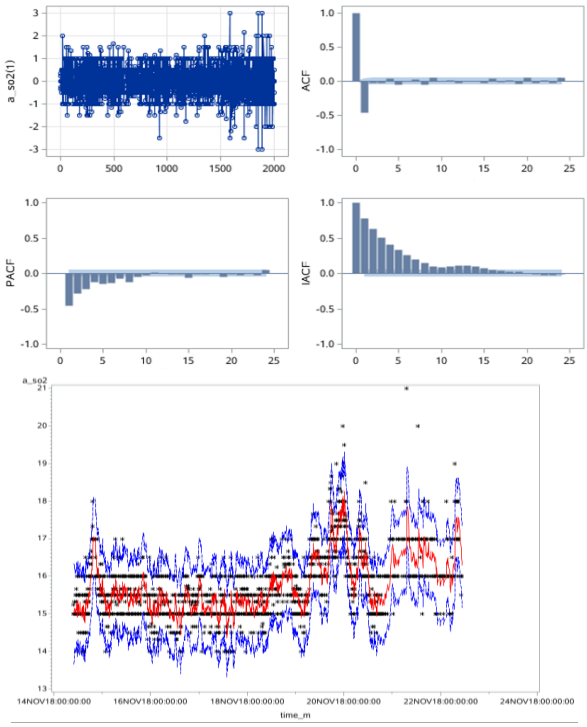

Figure 4. ARIMA model of SO2

4.2 B based on multiple linear regression

The parameter estimations of Multiple Linear Regression were shown in Table 4. The ANOVA results were shown in Table 5 (R^2=0. 932) . The MLR function was as follows.

$$B = -1573.16 - 6552.41 COL_1 + 1565.93 COL_2 + 304.55 COL_3 - 240.27 COL_4 + 62.06 COL_5$$

Table 4 MLR model of SO2

Parameter Estimate					
Variable	df	estimate	SD	t	P
Intercept	1	-1573.0598	146.4220	-10.74	<0.001
COL1	1	-6552.4119	1264.2564	-5.18	0.5692
COL2	1	1565.9314	145.9361	10.73	<0.001
COL3	1	304.5493	20.8915	14.58	<0.001
COL4	1	-204.2688	73.7342	-3.26	0.0011
COL5	1	62.0629	29.0882	2.13	0.0329

Table 5 ANOVA of MLR model

ANOVA					
Variation	df	SS	MS	F	P
Model	5	315082	63016	70.06	<0.001
Errors	3405	3066384	899.49661		
Total	3414	3381466			

5 Discussion

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.

$$\text{Average relative error} = \frac{|PV - SV|}{SV * n}$$

We got the predictive values by the additive calibration models and the ARIMA models. We also compared the monitoring data of SDD. The average relative errors were computed as follows (Table 6). The average relative error of ARIMA was the highest, and $Y = A + B$ the lowest. The accuracy has been improved.

Table 6 Average relative errors of SO2 by SDD, ARIMA, and additive calibration model

Variable	SDD	ARIMA	$Y_i = A_i + B_i$
SO2	0.7087	0.7289	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 suggested that SO2 monitoring data might be affected by the internal factors and the external factors. ARIMA was used for the internal factors. Meteorological factors were taken as external factors, and the difference of SO2 between the standard data and monitoring data was taken as dependent variable. Multivariate linear regression was modeled as B. Time series calibration model was obtained Y=A+B. The prediction precision the validity of additive calibration model was verified.

Our model still had some shortcomings to be improved. First of all, due to the lack of data, we only calibrated from the data point of view. 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]. Secondly, the interaction factors were not considered in the construction of multiple linear regression. That was where our model should be improved in the future.

Acknowledgment

This research is supported by Guangdong universities characteristic innovation project "prediction and analysis of regional differences and evolution of air emissions".

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.
2. Zhang Lingwei. Precision and accuracy analysis of air automatic monitoring instrument [J] environmental science guide, 2019, 38 (2): 123-129.
3. Su Hang, Wang Guimei, Zhang Zhenxing, et al. Optical path research of atmospheric SO2 detection module [J] laser technology, 2019, 7:91-97.
4. WEI Peng, REN Zhenhai, SU Fuqing. Seasonal Distribution and Cause Analysis of NO2 in China, Research of Environmental Sciences, 2011, 24: 155-161.
5. Yang Pengshi, Ding Hui, Chen Tong, et al. Prediction of urban bus emission energy consumption based on local weighted linear regression [J] Journal of Sun Yat sen University, Natural Science Edition, 2019, 06:111-118.
6. Gao Geng. Estimation drift of multiple linear regression and its determination method [J] statistics and decision, 2018, 14: 31-34.