Time Series and Multiple Linear Regression Calibration Model for CO Monitoring Data

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: CO is a kind of air pollutant with the largest amount and the widest distribution in the atmosphere produced by combustion of carbon containing substances. Real-time monitoring of the concentration of CO 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 CO 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 CO was improved. The additive model could effectively calibrate CO monitoring data.

1 Introduction

CO is a kind of air pollutant with the largest amount and the widest distribution in the atmosphere. It is the product of incomplete combustion of carbon containing substances such as industrial waste gas, civil combustion and automobile exhaust. Due to the continuous development of transportation, industry and mining enterprises in the world, the consumption of coal, oil and other fuels continues to grow, and the emission of CO also increases. In recent years, the trend of atmospheric warming is obvious, and the meteorological situation in winter is more stable and less changeable. CO combines with haemoglobin to form carboxyl hemoglobin. However, carboxyl haemoglobin does not carry oxygen itself. Its existence also affects the dissociation of oxygenated haemoglobin, which leads to hypoxia and carbon dioxide retention, and causes symptoms of poisoning [1]. It seriously threatens the health and safety of human beings and animal.

CO monitoring plays an important role in environmental protection, electric power, chemical industry, medical treatment, coating and other fields, and has become an essential protective link in industrial safety production. Real-time and accurate monitoring for CO could effectively control the pollution. 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 [2]. The self-developed micro air monitor has huge market value because of its flexible and economic type.

CO monitoring depends on electrochemical sensors, which 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 [3]. 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 CO by NCD and SDD. Five meteorology factors, i.e. wind, pressure, precipitation, temperature, and humidity were also given. It was found that CO 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

*Corresponding author: Lan Shuangting
[a]10000583@caac.net; [b*]28270031@qq.com

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

2.1 Statistical description

We computed the basic statistics of CO 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 CO monitoring data to explore the seasonal variation. The trend of the daily mean value of CO was fluctuation. It was higher in autumn and winter and lower in spring and summer (Figure 1).

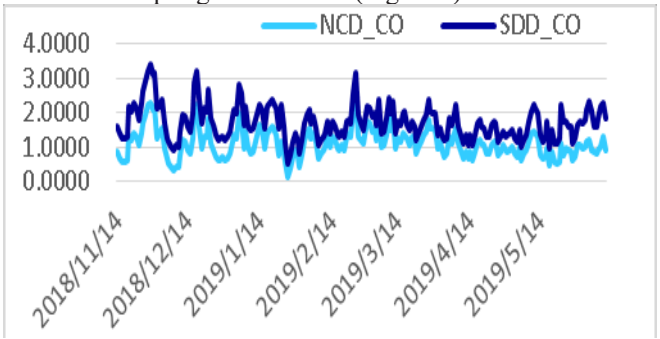

Figure 1 The daily mean value of CO by NCD and SDD

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 CO showed a double-peak feature. The hourly variation trend of CO could be seen that it was higher at 7 and 8 points in a day, then it slowly declined, the lowest at 13 and 14 points, then it slowly raised, the highest at 18 and 21 points, and then it declined again. The monitoring data of SDD was slightly higher than that of NCD (Figure 2).

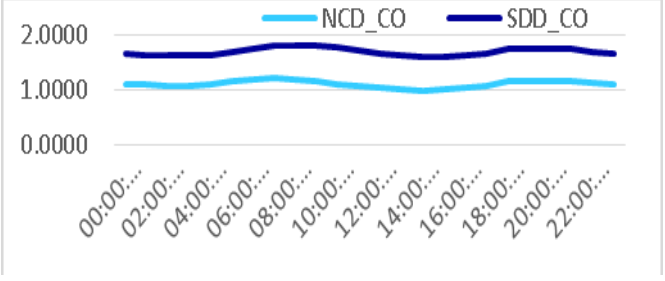

Figure 2 The mean value per hour of CO by NCD and SDD

2.2 Hypothesis testing

Paired t-test was used for CO 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 CO between NCD and SDD

	mean	SD	95%CI		t	P
(1)	0.5017	0.4683	0.4874	0.5159	-54.93	<0.0001
(2)	0.5027	0.4639	0.4885	0.5168	69.75	<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 CO between NCD and SDD, and the correlations between CO and the five meteorological factors. Then, we studied the autocorrelation of CO.

3.1 Correlation analysis

Correlation analysis showed that CO between NCD and SDD was correlated (r=0.33573, P<0.0001). They were correlated between CO of SDD and the meteorological factors (Table 2, P<0.0001). They were negative correlations with wind pressure and humidity, positive correlation with precipitation and temperature.

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

	Wind	Pressure	Precipitation	Temperature	Humidity
r	-0.0843	-0.1841	0.17511	0.30370	-0.05919
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	0.937	0.860	0.799	0.749	0.708	0.672
12	9999.99	12	<.0001	0.647	0.627	0.616	0.617	0.614	0.606
18	9999.99	18	<.0001	0.591	0.571	0.552	0.534	0.524	0.515
24	9999.99	24	<.0001	0.509	0.507	0.506	0.508	0.504	0.495

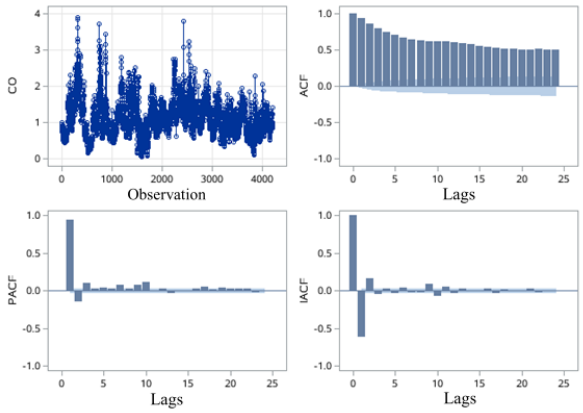

Figure 3 Trend and autocorrelation analysis of CO

4 Time series calibration model

In this part, the data from NCD was considered as the standard data. We remodeled CO 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.

$$Y = A + B$$

CO of SDD was considered as time series data. So, A was the predicted value of CO 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 CO 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 (=-6.47914) of ARIMA model when p=0 and q=1. So, ARIMA (0, 1, 1) was finally used to predict CO 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	0.00003893	0.000233	0.17	0.8673	0
MA1,1	0.73479	0.01519	48.36	<0.0001	1

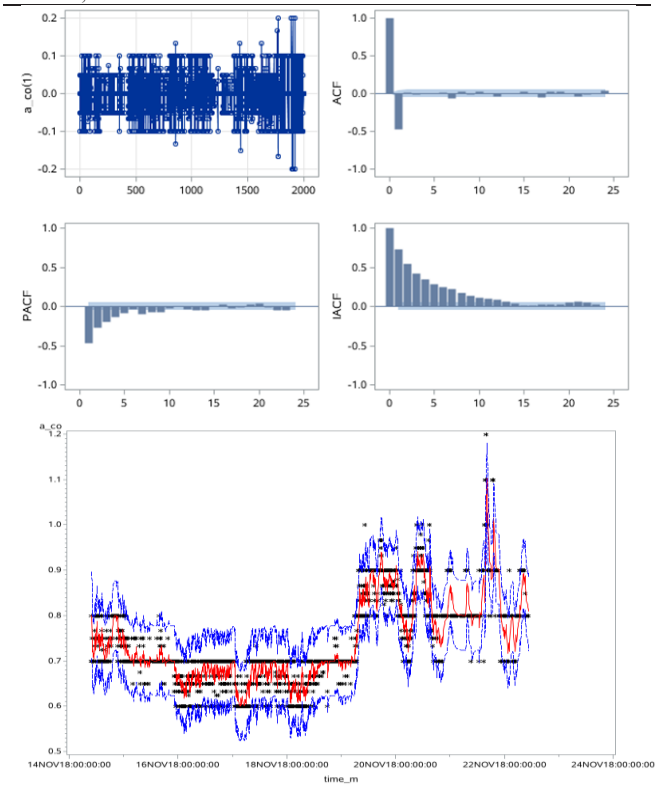

Figure 4 ARIMA model of CO

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. 5106) . The MLR function was as follows.

$$B = -34.67 - 207.74 COL_1 + 35.07 COL_2 + 4.62 COL_3 - 12.51 COL_4 + 4.23 COL_5$$

Table 4 MLR model of CO

Parameter Estimate					
Variable	df	estimate	SD	t	P
Intercept	1	-34.67465	2.15513	-16.09	<0.001
COL1	1	-207.74636	18.60812	-11.16	<0.001
COL2	1	35.07134	2.14798	16.33	<0.001
COL3	1	4.62252	0.30749	15.03	<0.001
COL4	1	-12.51179	1.08527	-11.53	<0.001
COL5	1	4.22735	0.42814	9.87	<0.001

Table 5 ANOVA of MLR model

ANOVA					
Variation	df	SS	MS	F	P
Model	5	117.79157	23.55381	120.90	<0.001
Errors	3405	664.29543	0.19487		
Total	3414	782.08699			

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 CO by SDD, ARIMA, and additive calibration model

Variable	SDD	ARIMA	$Y_i = A_i + B_i$
CO	0.6312	0.4964	0.4511

6 Conclusion

Through the exploratory analysis of CO 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 CO 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 CO 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, 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 [5]. 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. Brendan P McDonnell BA MB MRCPI. Smoking in pregnancy: pathophysiology of harm and current evidence for monitoring and cessation [J] The Obstetrician & Gynecologist, 2019, 10: 169-175.
2. Zhang Lingwei. Precision and accuracy analysis of air automatic monitoring instrument [J] environmental science guide, 2019, 38 (2): 123-129.
3. WEI Peng, REN Zhenhai, SU Fuqing. Seasonal Distribution and Cause Analysis of NO2 in China, Research of Environmental Sciences, 2011, 24: 155-161.
4. Forecast comparison based on ARIMA model, grey model and regression model [J] statistics and decision, 2019, 23: 38-41.
5. Gao Geng. Estimation drift of multiple linear regression and its determination method [J] statistics and decision, 2018, 14: 31-34.