Calibration of Air Quality Data Based on Multiple Regression Model

Hanping Zhang
Computer Department, Wuhan Polytechnic, Wuhan 430074, China

Abstract. According to the data of question D’s Annex 1 and Annex 2 of the National University Student Modeling Contest 2019, the correlation between the 6 air quality data errors and 5 meteorological factors at the self-built station is obtained by using the gray correlation method. Using the multiple linear regression method, and making the difference between the air quality data of the self-built station and the national control station as the dependent variable, and the meteorological data of the self-built station as the independent variables, six multiple regression equations were established to correct the air quality data of the self-built station.

1. Introduction
Although the data of national control stations is more accurate, due to the problems of less control, high cost, and time lag of data release, it cannot meet people's timely understanding air quality requirements. The self-built station can monitor the local air quality in real time, and can monitor temperature, humidity, wind speed, barometric pressure, precipitation and other meteorological parameters simultaneously. However, due to various factors, the self-built station may not detect accurate data, so the data from the national control station needs to be used to calibrate the data given by the self-built station. In this paper, the data of the National University Student Modeling Contest 2019 Question [D] [1] is taken as an example (where Annex 1 is the data of the national control station and Annex 2 is the data of the self-built station), and the calibration method of the self-built station data is given.

2. Materials and methods

2.1. Data sources
This article selects the data of the attachment to the question D of the 2019 National College Students Mathematical Modeling Contest. Attachment 1 shows part of concentration data of the national control station somewhere starting at 10 o'clock on November 14, 2018 and ending at 3 p.m. on June 11, 2019, including PM2.5, PM10, CO, NO2, SO2, and O3 concentrations in the air per hour nearly 7 months. The data in Annex 2 are from the self-built station and the data volume and the information is large. Due to the convenient testing, the measurements is frequent. In general, the measurement is performed every 5 minutes. Each measurement has not only real-time data of “two dusts and four gases”, but also measurement data of temperature, humidity, wind speed, barometric pressure, precipitation and other meteorological parameters at the same time.
2.2. Data quality analysis and processing

2.2.1. Outlier analysis. Taking PM2.5 and PM10 as examples, we draw Scatter diagram of air pollution monitoring data at national control stations and self-built stations, and observe the changes and trends of the data (Figure 1, Figure 2).

![Figure 1. PM2.5 Scatter diagram of national control station](image1)

![Figure 2. PM10 Scatter diagram of national control station](image2)

It can be seen from the figure that the abnormal value of PM2.5 at the national control station is not obvious, and the abnormal value of PM10 at the national control station is obvious. Using the quartile method to find the abnormal value is as follows.

1. The range \( R \) is the difference between the maximum and minimum values of the sample data. When the data is more dispersed, the range is greater.

\[
R = x_n - x_i
\]  

(\( x_i \) is the smallest value in a set of numbers, \( x_n \) is the maximum value in a set of numbers)

2. To calculate the cut-off point of outliers, average a set of numbers into four parts, and the 25% and 75% segmentation points constitute the lower and upper quartiles. The lower and upper quartiles are numerical features that measure the dispersion of samples. Data with outliers play an important role in the analysis of robustness data.

\[
R_{on} = Q_3 + 1.5R
\]  

(Q3 is the upper quartile)

\[
R_{below} = Q_1 - 1.5R
\]  

(Q1 is the lower quartile)

3. Remove outliers. The data below R or above R are all outliers. Use MATLAB software to find out and remove the outliers. The Scatter diagrams of the PM2.5 and PM10 data at the national control stations after cleaning are shown below (Figure 3, Figure 4).
Similarly, remove the abnormal values of the other 4 air quality indexes including CO, NO₂, SO₂, O₃ of the national control station and 11 indexes data of the self-built stations. Finally, the data dimension of Annex 1 is 4137*6, and the data dimension of the 6 air quality indexes and 5 meteorological index data from the self-built station is 197339*11 after removing outliers.

2.2.2. Timeline Consistency Analysis of a Single Monitoring Station. Before comparing and analyzing the national control station data and self-built station data, we perform quality inspection on the data of Annex 1 and Annex 2 firstly and find that the data has good consistency and there are no invalid values or missing values. Considering that the national control stations and self-built stations are interrupted due to some reasons, the data quality check is mainly aimed at monitoring whether the recording time of the data is continuous. Taking the monitoring data in Annex 1 as an example, the normal interval of monitoring data at the national control station is once an hour. Once it is found that there are two adjacent data recording intervals longer than one hour, it indicates that there is a discontinuity in the monitoring data recording.

Similarly, there are many “discontinuities of hours” (shown by dashed circles), which correspond to the records of Articles 333, 351, 382, and 502 in Annex 1. Similarly, there are some discontinuities of monitoring time in Annex 2. The normal interval of monitoring data of national control stations is once within 5 minutes at least. For example, the data of Article 9073 is a record of 18:51 on December 1, 2018, and no monitoring records were generated within 19: 00 ~ 20: 00 on that day.

It should be pointed out that due to the limited number of missing values recorded at the time point, its impact is not considered in the data analysis, because the analysis is based on large sample statistics.
2.2.3. **Time stamp consistency analysis and data processing of two monitoring stations.** To quantitatively characterize the difference between self-built station data and national control station data, the difference between them needs to be calculated. Therefore, the calculation method of the difference directly determines whether the second model can be established and solved. For the difference between the two data, the comparability of national control station data and monitoring station data can only be achieved by calculating the difference value based on time alignment.

The observation stations of the national control station data in Annex 1 are all hourly time records. The observation data of the self-built stations in Annex 2 are records that are separated by a few minutes and most of them are not integer points, and the time stamps of the two are different. We need to use the time point in Annex 1 as a reference firstly to find the records in Annex 2 that have a time relationship, or use the timestamp in Annex 2 as a reference to find the corresponding monitoring records in Annex 1, so align national control station data and self-built station data in time to analyze the error of the self-built station. In this paper, the previous time alignment method and cubic spline difference method are used. The matlabde interp1 command is used to obtain the observation value of the self-built station on the national control station time-stamp. The basic principle of cubic spline interpolation is as follows.

Nodes \( a = x_0 < x_1 < \cdots < x_n = b \) on given interval \([a, b]\) and function values at these nodes:

\[
f(x_i) = y_i \quad (i = 0, 1, \ldots, n)
\]  \hspace{1cm} (4)

If \( S(x) \) satisfies,

\[
S(x_i) = y_i \quad (i = 0, 1, 2, \ldots, n);
\]  \hspace{1cm} (5)

It is a cubic polynomial at most in each interval \([a, b]\).

\( S(x), S'(x), S''(x) \) is continuous on \([a, b]\).

Then \( S(x) \) is called \( x_0, x_1, \ldots, x_n \) cubic spline interpolation function of function \( f(x) \) in respect of nodes.

After time alignment processing, the monitoring data of the national control station is recorded as NValue1, and the dimension is 4137*6. The corresponding time field is in the format of a 4-dimensional array of year, month, day, and hour. The concentration of "two dusts and four gases", and meteorological data such as wind speed, barometric pressure, precipitation, temperature, and humidity are interpolated in columns of self-built station. It is recorded as NValue2 that compression of 19,739 pieces of data into 4,137 pieces and the dimension is 12.

2.3. **Correlation analysis between difference value and meteorological factors**

2.3.1. **Grey relational analysis.** Grey correlation analysis [2] is a commonly used statistical analysis method, which means that the relative strength of the variable of interest in a gray system is affected by other variables. It is simple and easy to use grey correlation analysis, but it has strong subjectivity at the same time.

2.3.2. **Grey correlation analysis of air pollutant difference value and meteorological factors.** Considering the data that affects the measurement error of the self-built stations includes wind speed, barometric pressure, precipitation, temperature, and humidity. After 2.2.3 interpolation processing, the timestamps are unified. On this basis, the gray correlation method is applied to analyze the influencing factors causing measurement errors of PM2.5, PM10, CO, NO2, SO2, and O3, and the gray correlation between each measurement error and each influencing factor is shown in Table 1.
Table 1. Correlation between the difference of 6 air quality indexes and 5 meteorological factors

| Correlation  | Wind speed | Barometric pressure | Precipitation | Temperature | Humidity |
|--------------|------------|---------------------|---------------|-------------|----------|
| PM2.5 difference | 0.9511     | 0.5214              | 0.9485        | 0.9986      | 0.3387   |
| PM10 difference     | 0.9942     | 0.5374              | 0.9961        | 0.9448      | 0.3454   |
| CO difference        | 0.5168     | 0.9236              | 0.5176        | 0.5031      | 0.5091   |
| NO2 difference       | 0.9971     | 0.5360              | 0.9977        | 0.9491      | 0.3448   |
| SO2 difference       | 0.9581     | 0.5229              | 0.9555        | 0.9910      | 0.3391   |
| O3 difference        | 0.3448     | 0.4920              | 0.3452        | 0.3387      | 0.9989   |

It can be found that there is a clear correlation between the wind speed and the errors of PM2.5, PM10, NO2, SO2, and the gray correlation is greater than 0.9. The error of barometric pressure and CO has a higher gray correlation. The grey correlation of the precipitation error to PM2.5 and NO2 is higher. Temperature will cause errors in the measurement of PM2.5, PM10, NO2 and SO2. Humidity is only related to the error value of O3.

2.4. Multivariate linear regression model of air quality data difference and meteorological factors

2.4.1. Establishment of multiple linear regression model. The basic steps of multiple regression analysis [3] are as follows:
1) Obtain data of independent and dependent variables as sample data.
2) Determine the model as a multiple regression linear model based on the relationship between the independent and dependent variables.
3) Use the sample data of the independent and dependent variables to fit the coefficients of the regression mathematical model.
4) Evaluate the pros and cons of the model through the degree of significance and fit of the model.

The difference between the data of the self-built station and the national control station on the concentrations of PM2.5, PM10, CO, NO2, SO2, and O3 is the dependent variable respectively. The wind speed, barometric pressure, precipitation, temperature, and humidity of five meteorological data is used as independent variable. The model of multiple linear regression analysis established on this basis is as follows.

The linear regression model of the dependent variable $Y$ and the K-variable $X$ is as follows,

$$
\begin{cases}
Y = X \beta + \varepsilon \\
E(\varepsilon) = 0, \text{COV}(\varepsilon, \varepsilon) = \sigma^2 I_n
\end{cases}
$$

Which is called Gaussian Markov linear model (k-element linear regression model), and abbreviated as $(Y, X\beta, \sigma^2 I_n)$

$$
Y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix},
X = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix},
\beta = \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_k
\end{bmatrix},
\varepsilon = \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
$$

$$
y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k
$$

is called regression plane equation,

Where $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ (in this model, $k = 5$) is an unknown parameter that has nothing to do with $x_1, x_2, x_k$ (in this model, $k = 5$).
\( y_i (i = 1, 2, \cdots, 6) \) is the difference between the concentration of air pollutants at the national control station and the self-built station, which corresponds to PM2.5, PM10, CO, NO\(_2\), SO\(_2\), O\(_3\) in order, which means that \( x_1, x_2, \cdots, x_5 \) is the difference value impact model based on multiple linear regression. It also needs to be run 6 times to calculate the influence of various meteorological factors on the difference value of the "two dusts and four gases" monitoring data.

The model needs to pass R and F tests, where the closer R is to 1 and the probability P corresponding to F is less than 0.05, so that the linear relationship holds.

2.4.2. Solving and testing multiple regression models. The running environment for this model solution is Intel Core i7 processor, 8G memory and Matlab2012a. The following model solutions are all run in this environment. Regression model of air quality index difference and meteorological factors was obtained by using the stepwise regression method in Matlab2012a fitting toolbox (Table 1). The \( y_1, y_2, \cdots, y_6 \) represents the difference between PM2.5, PM10, CO, NO\(_2\), SO\(_2\), and O\(_3\) at the two observation stations respectively. The \( x_1, x_2, \cdots, x_5 \) is the value of meteorological index including wind speed, barometric pressure, precipitation, temperature and humidity. The F is the F-test value of the model, and p is the p-value corresponding to the F-test, and RMSE is the root mean square error.

Applying MATLAB fitting toolbox, six models are obtained and the test results are as follows (Table 2).

| Air quality index | Multiple regression model of air pollutant difference and meteorological factors | F    | P     | RMSE |
|-------------------|---------------------------------------------------------------------------------|------|-------|------|
| PM2.5             | \( \hat{Y}_1 = 0.358201 x_1 - 0.0123077 x_2 + 0.357518 x_3 - 371.132 \)          | 93.70| 7.20e-57 | 19.25|
| PM10              | \( \hat{Y}_2 = 2.15413 x_1 + 1.2086 x_2 + 1.5117 x_3 - 2270.72 \)               | 206.37| 3.91e-117 | 50.47|
| CO                | \( \hat{Y}_4 = 0.0427413 x_1 + 0.0228121 x_2 - 44.4281 \)                       | 212.42| 1.17e-84 | 4.67 |
| NO\(_2\)          | \( \hat{Y}_6 = 3.34286 x_1 - 0.0252453 x_2 + 5.01137 x_3 + 0.994137 x_4 - 3051.45 \) | 78.10| 8.80e-62 | 44.51|
| SO\(_2\)          | \( \hat{Y}_6 = 1.42398 x_1 - 0.0294511 x_2 - 2.42315 x_3 - 0.264783 x_4 + 1493.2 \) | 65.10| 6.36e-52 | 25.87|
| O\(_3\)           | \( \hat{Y}_6 = 1.17884 x_1 + 1.87663 x_2 + 0.038635 x_3 + 1.97676 x_4 + 0.667597 x_5 - 1967.31 \) | 57.53| 3.46e-56 | 38.03|

The regression model obtained the influence of various meteorological parameters on the monitoring differences of PM2.5, PM10, NO\(_2\), CO, SO\(_2\) and O\(_3\) quantitatively. The P value indicates that all models are significant as a whole, and the F value indicates that the model fits well.

3. Results and discussion

3.1. Analysis for the influence of meteorological factors on the error of air data by models
It can be seen from the analysis of the regression equations that the constant terms will lead to changes in the monitoring difference, and the zero point shift and range shift can be regarded as that the constant terms of the equation changed. The barometric pressure and humidity have a positive effect on the PM2.5 data difference (the coefficients are 0.358201, and 0.357518), and the precipitation has a negative effect on the PM2.5 data difference (the coefficient is -0.00123077). Similarly, according to Table 2, the influence degree of meteorological factors on other air data errors can be analyzed, and we will not repeat them here.

3.2. Model test
Randomly extract data from 30 time-stamps, and use the average relative error to test the 3.2.2 model. The relative error calculation formula is:
\[ m = \frac{\sum_{i=1}^{n} \left| v_i - u_i \right|}{n} \]  

(8)

\( v_i \) and \( u_i \) are the actual observations and model calculations respectively, and \( n \) is the number of selected time-stamps, and the relative errors of the six models are shown in Table 3.

Table 3. Mean relative errors of calibration models

| Air pollutant | Pm25 | PM10 | CO   | NO₂  | SO₂  | O₃  |
|---------------|------|------|------|------|------|------|
| Mean relative error | 0.07949 | 0.08609 | 0.1088 | 0.10518 | 0.11642 | 0.07907 |

The relative error values of these six air quality indexes are all less than 15%, which indicates that the calibration data is credible in self-built station.

4. Conclusion

From the perspective of correlation analysis and multiple regression analysis, the factors that cause the difference between self-built station data and national control station data were examined, including influence factors that caused the difference value to be generated, and measuring the correlation between the difference and meteorological factors, and showing the significance degree, effect direction and influence degree of each influencing factor on the difference value.

The self-built station data after calibration is in good agreement with the national control station data. The model has passed the F test, and the method is clear. The calculation process is clear and operable. The model established in this paper can be applied to use the data of national control stations to calibrate the data of self-built stations, and then grasp the air quality to take corresponding measures, which is beneficial to maintaining the ecological environment and promoting human health development.

References

[1] National College Students Mathematical Modeling Contest. 2019 National College Student Mathematical Modeling Competition [EB/OL] (2019-09-12). http://www.Mcm.edu.cn/html_cn/node/b0ae8510b9ec0cc0deb2266d2de19ecb.html.
[2] Si Shoukui, Sun Xidi Mathematical modeling methods and applications [M]. Beijing: National Defense Industry Press, 2011.
[3] Jiang Qiuyuan, Xie Jinxing, Ye Jun Mathematical models [M]. Beijing: Higher Education Press, 2018.