ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

Calibration of low-cost harmful gas sensors based on linear regression method

Yu Yan[1*], Quansen Zhao[2,a], Miaomiao Gu[3,b], Yizhou Mao[4,c]

[1]Nanjing University of Information Science & Technology, Nanjing, Jiangsu, China,

[2]University of Jinan, Jinan, Shandong, China

[3]Suzhou University of Science and Technology, Suzhou, Zhejiang, China

[4]Bishop garcia diego high school, Santa Barbara, CA, USA

*Corresponding author: 271412173@qq.com

[a]861613420@qq.com, [b]1109980920@qq.com, [c]jetmao@sina.com

These authors have contributed equally to this work

Abstract

With the aggravation of air pollution, low-cost sensors, as a kind of cheap and portable air detection equipment is being widely used in different situations, like in industry or at home. However, the precision of this sensor can be easily influenced by surroundings. In order to solve this problem, a kind of calibration way by building linear regression model is tested in detail in this work. Through choosing temperature and humidity as factors and comparing values from sensors with observation station, a linear model can be built to represent relations among sensors and different factors. This model can also be used to calibrate values from low-cost sensors to get more accurate results. Three kinds of linear regression methods are discussed and compared in this work: point gradient descent method, matrix gradient descent method and least square method. It is found that the results of these three methods are approximately the same, but each method is suitable for different conditions. Point gradient method is suitable for the situation when the number of factors is small, matrix gradient descent method takes the longest time among three method. The least square method is the most convenient one but need to determine the invertibility of the matrix before using.

Keywords

Linear regression; air pollution; point gradient descent method.

1. Introduction

The atmosphere is one of the indispensable environmental factors for human survival and development. It is a mixture of many components, such as clean air, water and suspended particles. The main components of clean air are nitrogen, oxygen and argon, which account for 99.96% of the total volume of the air. Other components such as carbon dioxide, neon, helium and ozone account for the rest. Although the moisture content in the atmosphere is relatively low, it is greatly affected by time, region and meteorological conditions, and plays an important role in the dry and wet conditions of the air. Suspended particles refer to the particles generated by natural factors or human activities, such as rock weathering, volcanic eruption, coal combustion, fireworks and firecrackers [1]. Air pollution usually refers to the phenomenon that a large number of substances enter the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

atmosphere due to human activities or natural evolution process, reaching a sufficient time and thus endangering human comfort, health and welfare or harming the environment.

However, with the continuous development of science and technology, people gradually put more resources and energy into the research and development of new technologies. Due to the rapid growth of population, all kinds of pollution are increasing. The inevitable waste of science and technology brought by the rapid

development of science and technology has been irreversibly damaged. One of the more serious pollution is that air pollution. The toxic substances contained in the polluted atmosphere will pose a great potential threat to human health. The London smog incident in Britain, the asthma incident in Japan's 4-day City, the smog incident in Los Angeles in the United States, and the gas leakage incident in Bhopal, India, all use their lives to warn human beings that harmful gases will not only seriously endanger the health of residents, but also cause hundreds or thousands of deaths. Therefore, scientists began to study the relevant equipment to test the air conditions in different weather regions and time periods, trying to find problems and find effective solutions.

At present, people usually use high-precision equipment such as electrochemical sensors to collect the data of harmful gas content in the atmosphere, and establish a fitting model based on the collected real value and the value to be corrected to correct the corresponding harmful gas data. Although electrochemical sensor is a mature technology with high selectivity, high sensitivity, good repeatability and accuracy, its limitations can’t be ignored. This kind of sensor is easy to be affected by temperature and humidity, which makes the collected data not accurate or achieve the expected effect. At the same time, even the same sensor of the same production batch in the same environment test value is also different. This will bring great difficulty and uncertainty to the experimental research, and greatly reduce the accuracy of the experiment.

Therefore, we need to establish the corresponding model to correct the collected values, so as to improve the accuracy, representativeness and usability of the data, so as to reduce the difficulty of the corresponding experiment, speed up the data processing, and make the experiment achieve the desired effect as far as possible.

The experimental project is based on the data of Beijing Wanliu monitoring station from July to October 2017 as the real value, and the linear regression algorithm model is used as the main correction method to correct the corresponding harmful gases in the data, so as to achieve the goal.

The test result is also very promising, our group elaborate the linear regression from three aspects: matrix, dot mode, and least square method. We extract four figures naming them theta1 to theta4 and find the loss function and running time respectively. The result under the three ways of linear regression are different with each other. According to the distinctions between each set of the data, we could find out the most accurate data.

2. Method Introduction:Gradient Descent

The gradient descent method or steepest descent method is one of the most commonly used methods to solve unconstrained optimization problems, and has the advantage of simple implementation[2]. The gradient descent method is an iterative algorithm, each step needs to solve the gradient vector of the objective function. The hypothesis $f(x)$ is a function R^n with a first-order continuous partial derivative[3]. The unconstrained problem to be solved is:

$$min f(x), x \in R^n$$

The gradient descent method is an iterative algorithm. Choose appropriate initial value $x^{(0)}$,Keep iterating,Update x’s value， The objective function is minimized until convergence. Since the negative gradient direction is the direction that makes the function value drop the fastest, at each step of the iteration, the value is updated in the negative gradient direction to achieve the purpose of reducing the function x’s value[5-7]. Since $f(x)$ has a first-order continuous partial derivative, if the value of the k iteration is $x^{(k)}$, the first-order Taylor expansion of $f(x)$ can be performed near $x^{(k)}$:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

$$f(x) = f(x^{(k)}) + {g_k}^T(x - x^{(k)}) + \sigma(x - x^{(k)})^2 \tag{1}$$

Here, $g_k = g(x^{(k)}) = \nabla f(x^{(k)})$ is $f(x)$ in $x^{(k)}$'s gradient.

Find the k+1 iteration value $x^{(k+1)}$:

$x^{(k+1)} \leftarrow x^{(k)} + \lambda_k p_k$

Where p_k is the search direction, taking the negative gradient direction $p_k = -\nabla f(x^{(k)})$, is the step size, λ_x is determined by one-dimensional search, that is, λ_x makes:

$$f(x^{(k)} + \lambda_k p_k) = \min_{\lambda>=0} f(x^{(k)} + \lambda p_k) \tag{2}$$

Given m training examples described by n features (variables)，Linear model tries to learn a model (function) that predicts through a linear combination of features[4]:

$$h(\boldsymbol{x}) = h_\theta(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \tag{3}$$

This model contains n+1 parameters and n features. In order to simplify its form and introduce virtual features $x_0 = 1$, the model can be transformed into:

$$h_\theta(\boldsymbol{x}) = \sum_{i=0}^{n} \theta_i x_i = \boldsymbol{\theta}^T \boldsymbol{x} \tag{4}$$

At this time, the parameter in the model is an n+1-dimensional vector, and any training instance is also an n+1-dimensional vector. Construct a loss function J, which represents the sum of squares of all prediction errors, namely

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(\boldsymbol{x}^{(i)}) - y^{(i)})^2 \tag{5}$$

Find out the parameter θ when the loss function takes the minimum value, and the prediction result can be optimized[3]. The following uses the gradient descent algorithm to find the minimum value of $J(\theta)$:

2.1 Calculate the gradient of the objective function :

$$\nabla J(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ \vdots \\ \frac{\partial J}{\partial \theta_n} \end{pmatrix} = \begin{pmatrix} \frac{1}{m} \sum_{i=1}^{m} (h_\theta(\boldsymbol{x}^{(i)}) - y^{(i)}) x_1^{(i)} \\ \frac{1}{m} \sum_{i=1}^{m} (h_\theta(\boldsymbol{x}^{(i)}) - y^{(i)}) x_2^{(i)} \\ \vdots \\ \frac{1}{m} \sum_{i=1}^{m} (h_\theta(\boldsymbol{x}^{(i)}) - y^{(i)}) x_n^{(i)} \end{pmatrix} \tag{6}$$

2.2 Calculate the Hesse matrix of the objective function $J(\theta)$:

$$J_{ij} = \frac{\partial^2 J}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_j}\left(\frac{\partial J}{\partial \theta_i}\right) = \frac{\partial}{\partial \theta_j}\left[\frac{1}{m} \sum_{k=1}^{m} (h_\theta(\boldsymbol{x}^{(k)}) - y^{(k)}) x_i^{(k)}\right] = \frac{1}{m} \sum_{k=1}^{m} x_i^{(k)} x_j^{(k)}$$

$$\boldsymbol{H}_J(\boldsymbol{\theta}) = (J_{ij})_{n\times n} = \begin{pmatrix} J_{11} & J_{12} & \cdots & J_{1n} \\ J_{21} & J_{22} & \cdots & J_{2n} \\ \vdots & \vdots & & \vdots \\ J_{n1} & & & J_{nn} \end{pmatrix} \tag{7}$$

2.3 Gradient descent algorithm：

1)Select the initial point arbitrarily, give the allowable error, and let;k←0；

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

2)Calculate $p^k = \nabla J(\theta^{(k)})$

3)Check whether the convergence criterion is met:

$\| \boldsymbol{p}^{(k)} \| < \varepsilon$

If the criterion is met, stop the iteration and get the approximate optimal solution, otherwise proceed to 4);

4)The optimal solution of univariate extreme value problem α_k :

$$\min_{\alpha \geq 0} J\left(\boldsymbol{\theta}^{(k)} - \alpha \boldsymbol{p}^{(k)}\right) = J\left(\boldsymbol{\theta}^{(k)} - \alpha_k \boldsymbol{p}^{(k)}\right) \tag{8}$$

Solve to get the optimal step size :

$$\alpha_k = \frac{[\boldsymbol{p}^{(k)}]^{\mathrm{T}} \boldsymbol{E} [\boldsymbol{p}^{(k)}]}{[\boldsymbol{p}^{(k)}]^{\mathrm{T}} \boldsymbol{H}_J(\boldsymbol{\theta}^{(k)}) [\boldsymbol{p}^{(k)}]} \tag{9}$$

5) Let $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha_k \boldsymbol{p}^{(k)}$, $k \leftarrow k+1$, back to 2):

In this experiment, the step size algorithm does not need to obtain the optimal step size, only the number of iterations and the learning rate need to be substituted into the formula to obtain it.

Example implementation[4]

By collecting the data detected by the Wanliu sensor in Beijing, CO data is extracted, and the temperature and humidity are normalized, and the linear regression result is obtained through the combined processing of the sensor and the Wanliu data.

The experimental question asked:

Aiming at the CO data set of various gases in the region, it contains a total of training samples, denoted as (x_i, y_i), which x_i represents temperature and humidity, y_i represents the CO data of Wanliu, and establishes a linear regression model.

Suppose the fitted model is as follows:

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 \tag{10}$$

Brief description of the experiment

The operating environment is Python, and the gradient descent method is used to minimize the loss function of the linear regression model. The learning rate of the batch gradient descent method is α=0.07, and the number of running iterations is 1500.

By type :

$$J(\theta) = 1/2m \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{11}$$

3. Related Work

Sensor calibration refers to the use of different algorithms to give appropriate parameters to multiple dependent variables affecting sensor readings, so that the output of the sensor is closer to the measured amount. Linear regression is a regression analysis that uses the least square function of the linear regression equation to model the relationship between one or more variables. By establishing a linear model for factors x_i, the function,

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T X \tag{12}$$

could be obtained. By building the loss function

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{13}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

and finding out the parameters θ when the loss function reaches the minimum value can make the prediction result reach the optimal value. Here are three ways to calculate the parameter θ.

3.1 Point gradient descent method

Gradient descent is an iterative algorithm that minimizes the loss function by keeping iterating the value of θ. The point gradient descent method is to solve each θ value separately. The formula is

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta), \tag{14}$$

where α is the step length. By further calculation, the formula can be rewritten as

$$\theta_j := \theta_j - \alpha\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\cdot x_j^{(i)}, \tag{15}$$

from which the value of each θ can be calculated in the form of points.

3.2 Matrix gradient descent method

The matrix gradient descent method has the same principle as the point gradient descent method, but the difference is that the matrix one expresses several values in the form of matrix, and calculates all values at once, which is suitable for the case of large sample size. The derivation process is as follows:

$$\frac{\partial}{\partial\theta_j}J(\theta) = \sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\cdot x_j^{(i)} \tag{16}$$

Let $e^{(i)} = h_\theta\left(x^{(i)}\right) - y^{(i)}$, then

$$\frac{\partial}{\partial\theta_j}J(\theta)=\sum_{i=1}^{m} e^{(i)}\cdot x_j^{(i)} = M_j = [x_j^{(1)}\ x_j^{(2)}\ \cdots x_j^{(m)}]\begin{bmatrix} e^{(1)} \\ e^{(2)} \\ \vdots \\ e^{(m)} \end{bmatrix}, \tag{17}$$

which means

$$\begin{bmatrix} M_0 \\ M_1 \\ \vdots \\ M_n \end{bmatrix} = \begin{bmatrix} x_0^{(1)} & \cdots & x_0^{(m)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \cdots & x_n^{(m)} \end{bmatrix}\begin{bmatrix} e^{(1)} \\ e^{(2)} \\ \vdots \\ e^{(m)} \end{bmatrix} \tag{18}$$

Let $X=\begin{bmatrix} x_0^{(1)} & \cdots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \cdots & x_n^{(m)} \end{bmatrix}$, then

$$\theta = \theta - \alpha\cdot X^T\cdot e \tag{19}$$

So the θ can be solved in matrix form.

3.3 Least square method

By establishing linear function model, it can be obtained that

$$h_\theta(x) = \theta^T X \tag{20}$$

Ideally, it can be written as

$$y=X\cdot\theta,\ \text{so}\ \theta=(X^TX)^{-1}X^Ty \tag{21}$$

The value of θ can be gotten from this equation. However, while using this method, it is necessary to judge whether the matrix is invertible first. Under normal circumstances, irreversible matrix is easy to be generated when the number of samples is less than or equal to the character numbers.

3.4 Comparison

By using these three methods respectively, it's found that there is no significant difference in the fitting effect of these three methods. All three methods can be used when the numbers of factors and samples are small. However, when the number of factors is large, point gradient descent method is not as convenient as matrix gradient descent method or least square method. When the number of

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

samples is large, the matrix gradient descent method takes longer time and is less efficient, while the least square method shows the fastest speed. It can be seen from Table 1 that in this experiment, the time required by the matrix gradient descent method is as much as 400 times that of the least square method. However, it should be noted that the invertibility of the matrix needs to be determined before using the least square method. Therefore, different methods can be selected to establish linear regression model for fitting according to different sample conditions.

4. Result

In this study, linear regression model based on gradient descent method is used to calibrate the results from low-cost sensors. It is developed for the data set of CO in various gases in the area, including training samples, which represent temperature and humidity, and the CO data of Wanliu.

The results show that linear regression algorithm can effectively correct the deviations of the sensor measurements and control the measurement error cost function to below 0.05 in most cases. By analyzing the results obtained from the three evaluation methods, it’s found that the least squares method is faster and more efficient, and the matrix gradient descent method is more general but less efficient.

5. Discussion

In this study, a machine learning method is used to study the calibration of low-cost sensors based on a linear regression model to minimize the influence of environmental factors and get close to the real value.

By collecting the data detected by the Wanliu sensor in Beijing, CO data is extracted, the temperature and humidity are normalized, and the linear regression result is obtained through the combined processing of the sensor and the Wanliu data.

Nowadays, with the development of industry, high attention is paid to the problem of environmental pollution, a complete network of harmful gases detection is urgently need to be built. Low cost sensors, as a kind of cheap and portable gas detector is gradually coming to the attention. However, the accuracy of the sensor can be affected by temperature and humidity[8-9], even if the manufacturer adjusts the sensor before using, the accuracy of the sensor will still have great changes under different environments. Moreover, as the performance of each sensor is unique, it will cost a lot of manpower and material resources to calibrate each sensor separately, which is not practicable for flow-line production. The traditional mechanical calibration method can’t meet the needs of production anymore, it’s in urgent need of a fast and universal sensor calibration method in order to make the sensor meet the data quality requirements of scientific research. The linear regression model perfectly solves this problem by constructing the fitting model after processing the data with higher efficiency and lower cost.

This method is not only effective in gas sensor calibration, but can also be applied in other fields. For example, in this study, only the calibration of outdoor low-cost sensors by linear regression method is discussed, but this method can also be used in the detection and early warning of indoor harmful gases. Indoor harmful gas monitoring equipment can be widely used in factories or homes. Through simple linear regression model, sensor data can be calibrated to give early warning when indoor harmful gas content exceeds the standard value, which effectively protect people's health and provide better living and

 Volume 7 Issue 11, 2021
DOI: 10.6919/ICJE.202111_7(11).0064

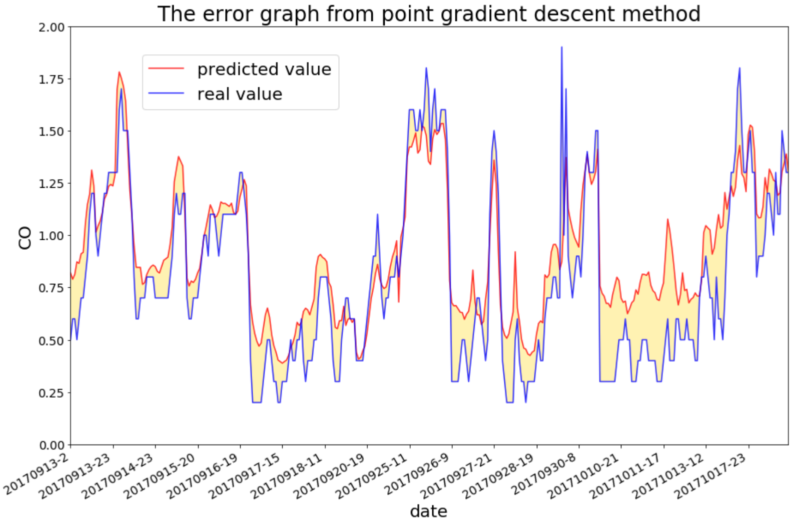

Fig1. The error graph from point gradient descent method

working environment. In addition to gas monitoring, this method can also be used for calibration of water pollution detector. The linear regression model can be established for each factor affecting the water detector, and the real value can be fitted[10].

However, there are still a lot of shortages needed to be improved in this study. It's known that the performance of sensor can be influenced by many environmental conditions like the surroundings, type of gas it test and sensor itself. However, in this survey, only CO is detected to test the accuracy of linear regression model and only temperature and humidity are concerned as influence factor while establishing the linear model.

Although the results show that linear regression model has obvious effects in sensor calibration, it's uncertain that the linear regression method will be as effective in the detection of other harmful gases. Moreover, in this research, we regard the data from Wanliu monitoring station as the actual content of harmful gas in the air and build the fitting model based on these data, which ignore the deviation of the data of Wanliu monitoring station itself. These errors might affect the results and cause the limitation of this research.

Another shortage comes from the limitation of the model itself. The essence of linear regression method is to establish a linear model with a high fitting degree to the real data for

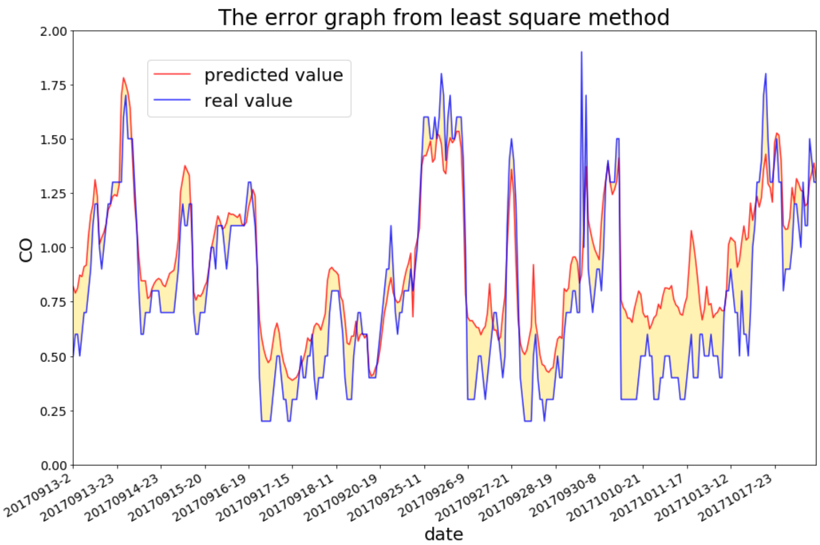

Fig2. The error graph from least square method

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

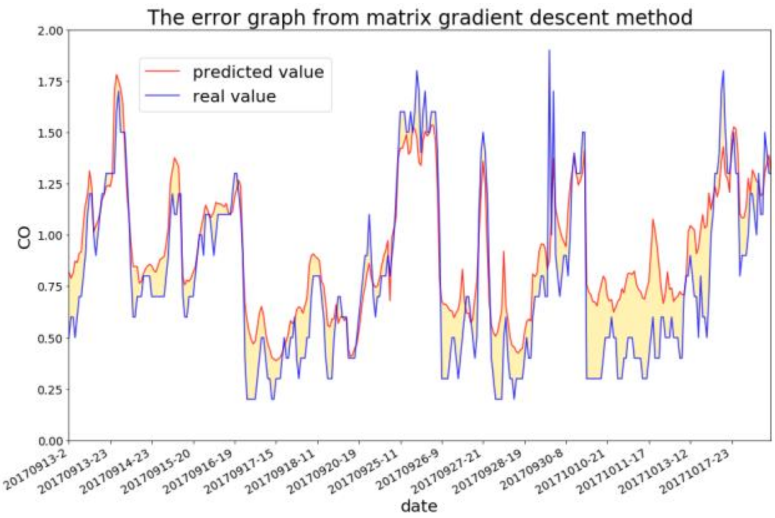

Fig3. The error graph from matrix gradient descent method

Table1. The results and time used in three methods

Method	Point gradient decent method	Matrix gradient decent method	Least square method
Theta0	-0.8523	-0.8523	-0.8523
Theta1	0.7575	0.7575	0.7575
Theta2	0.0506	0.0506	0.0506
Theta3	-0.1015	-0.1015	-0.1015
Loss	0.0498	0.0498	0.0498
Time (seconds)	45.3434	140.4919	0.2963

correction. However, when the real data presents a nonlinear distribution or the relationship between the influence factors is relatively complex, the fitting model established by using the linear regression can’t match the real data very well. Therefore, when the data distribution is complex, linear regression method can’t show great effect in sensor calibration.

The advantages of linear regression method are also obvious. It is concise and easy to understand and use. It can also quickly model, calculate and correct the data in the case of large data volume. Moreover, the results of the linear regression method have a good interpretability, which can clearly show the influence proportion of each influencing factor on the results, which is conducive to the decision analysis. However, the method used in this study only establishes a linear model to fit the real data for correction, which fails to achieve a good fitting effect in the case of relatively complex sample conditions. The linear model is established on the basis of determining the relationship among the influencing factors, which ignores the uncertainty of some factors and may leads to the error of the calculation results.

Therefore, it is suggested that when other data correction methods are studied, emphasis could be placed on how to consider the influence of the characteristics of influencing factors on the results in the process of modeling, or how to establish a suitable fitting model for nonlinear samples.

Acknowledgment

The authors wish to acknowledge the support from professor Li Shuangshuang for helping in programming parts and providing data needed in this research.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0064

References

[1] [Z.Shuguang, Research on ecological Countermeasures for air pollution control. New Countryside, 2011(19)

[2] S.Yanan, L.Wenbin. Application of gradient descent method in machine Learning [J]. Journal of Suzhou University of Science and Technology (Natural Science edition), 2008,35(02):26-31.

[3] L.Hang. Statistical Learning Methods [M]. Beijing: Tsinghua University Press, 2012:78-93.

[4] Z.Zhihua. Machine Learning [M]. Beijing: Tsinghua University Press, 2016:53-60.

[5] N.Z.Shor, K.C.Kiwiel, and A.Ruszcaynski. Minimization methods for ndifferentiable functions[M]. Springer-Verlag, 1985:885-888.

[6] N.Z.Shor. Minimization methods for non-differentiable functions[M].volume 3.SpringerScience & Business Media,2012:93-146.

[7] Q.Ning. On the momentum term in gradient descent learning algorithms ， Neuralnetworks:the official journal of the International Neural Network Society,12(1);145-151, 199.

[8] Aleixandre, M., Gerboles, M.. Review of small commercial sensors for indicative monitoring of ambient gas. Chem. Eng. Trans. , 2012(30):169–174.

[9] Mead, M.I., Popoola, O.A.M., Stewart, G.B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J.J., McLeod, M.W., Hodgson, T.F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J.R., Jones, R.L.. The use of electrochemical sensors for monitoring urban air quality in lowcost, high-density networks. Atmos. Environ. , 2013(70):186–203.

[10] G.Dingxi. Application of multiple Linear regression method in the prediction of lake Water Pollution [J]. Environmental Science and Technology,1984(04):62-72.