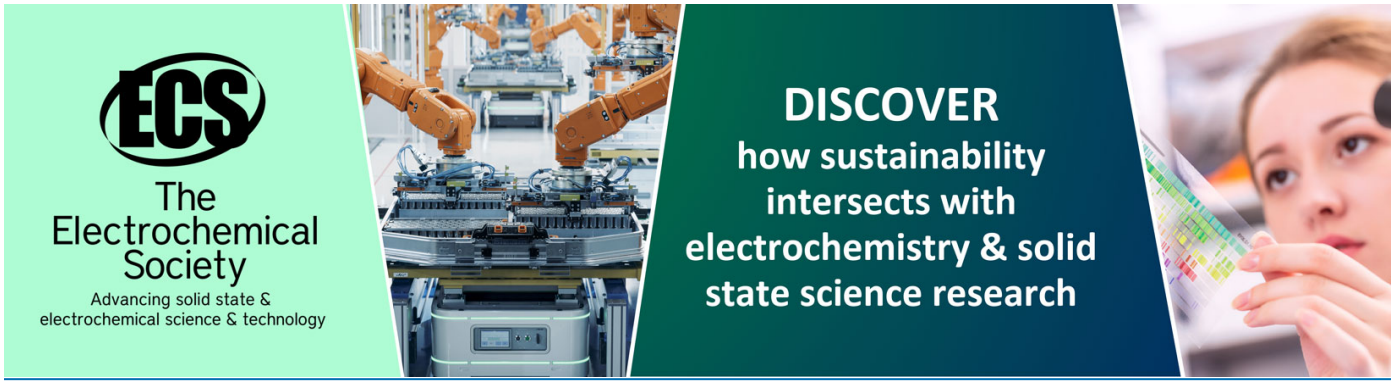
Multiple regression model for identification of material concentration and color reading



IOP Conf. Series: Materials Science and Engineering **612** (2019) 022086 doi:10.1088/1757-899X/612/2/022086

Multiple regression model for identification of material concentration and color reading

Hanping Zhang [a]

Computer Department, Wuhan Polytechnic, Wuhan 430074, China

[a]2811853540@qq.com

Abstract. According to the experimental data, the color dimensions of different concentrations of five substances were analyzed, and reasonable data was selected to analyze the correlation between concentration and color readings. The quantitative relationship model between material concentration and color reading was established by multiple regression. Considering the autocorrelation of the variables, a one-dimensional regression model of concentration and gray scale was established for sulfur dioxide. The two models were tested and were significant with minimal error. This method can be extended to measuring the concentration of the substance In practice.

Key words: Linear regression; RGB and HSV; Correlation analysis; Substance concentration

1. Introduction

Colorimetric method is a commonly used method for detecting the substance concentration. It refers to the reaction of a substance and a reagent, and then compared with a standard color chart to know the concentration of the substance, but the accuracy of method is greatly affected by the objective reasons such as sensitive differences and observation errors of each person, and the accuracy is not guaranteed. With the development of computer image processing technology and photographic technology, it has become possible to determine the substance concentration by color reading, and the key problem is to establish a quantitative model of color reading and substance concentration.

In image processing, the most common color spaces are the RGB model and the HSV model, both of which can determine a unique color. According to the tristimulus theory [1], our eyes perceive color by light stimulation of three visual pigments in the cone cells of the retina. Respectively, these three pigments are most sensitive to light having wavelengths of 630 nm (R), 530 nm (G), and 450 nm (B). By comparing the intensities in the light source, we feel the color of the light.This visual theory uses red, green, and blue as the color base color and the basis for displaying color on a video monitor, called the RGB color model. In addition, there is a set of representations called the HSV primary color model,which describes the colors more intuitively. To give a set of color descriptions, the users need to select a spectral color and add a certain amount of white and black to achieve different shades, color and hue. The color parameter values for this model are hue (H), color saturation (S), and lightness value (V). The three-dimensional table of the HSV model evolved from RGB cubes. RGB and HSV can be converted to each other, so we analyzed the problem mainly by using RGB to analyze the substance

concentration and test it with HSV.The formula for RBG and HSV conversion in the literature is as follows:

$$V = \max(R,\ G,\ B)$$

$$S = \begin{cases} V - \dfrac{\min(R,\ G,\ B)}{V} & \text{if } V \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$H = \begin{cases} \dfrac{60(G-B)}{V-\min(R,\ G,\ B)} & if\ V = R \\ 120 + \dfrac{60(B-R)}{(V-\min(R,G,B))} & if\ V = G \\ 240 + \dfrac{60(R-G)}{V-\min(R,G,B)} & if\ V = B \end{cases}$$

In order to establish the quantitative relationship between material concentration and color reading, is it to choose one of the independent RGB models and HSV models, or is it to choose other color dimensions that matter is color sensitive? This is the first problem that our model has to solve. After determining the color index of the substance concentration, the quantitative relationship between the concentration and the color reading is the key to solving the problem. Finally, the model checking is also an indispensable link.

2. Materials and methods

2.1. The data source and data processing

This paper selects the data of the 2017 National College Students Mathematical Modeling Contest C title [2]. Annex 1 gives corresponding color readings of five substances: histamine, potassium bromate, industrial alkali, aluminum sulfate potassium and milk urea in different concentrations of R, G, B, H, S. Annex 2 gives five sets of color readings of sulfur dioxide in different concentrations of R, G, B, H, S. The number of experimental data sets of the five substances in Annex 1 is different, including 10 groups of histamine, 10 groups of bromine potassium acid, 7 groups of industrial alkali, 35 groups of potassium aluminum sulfate, and 15 groups of urea in milk. The same substance data are arranged in ascending order of concentration. If there are multiple sets of color reading data at the same concentration, the average value should be taken. Taking histamine data as an example, the following Table 1 is organized, and other data are treated the same.

In Annex 2, 25 sets of color readings of sulfur dioxide at 7 concentrations are provided. We observed that the data of H and S may be incorrect. We use the formula to convert the R, G, and B values into H and S values, and find that the calculated S value approximates the H value in the raw data of Annex 2, and the calculated H value is about twice the S value in the original data. The data of the two groups of H and S are just opposite, which may be due to the error in the data recording of the experiment. We correct the data and substitute the data R, G, and B into the formula (1) to calculate S, H to compare with the adjusted data. The data consistency was 98.83% and 99.62% respectively, and the data was accurate. The five groups of data of R, G, B, S and H at different concentrations of sulfur dioxide were averaged to obtain modeling data.

Table 1. Experimental data on histamine concentration and color readings

Histamine	Concentration	B	G	R	H	S
Group 1	0	68	110	121	23	111
	12.5	66	102	118	20	112
	25	62	99	120	19	122
	50	46	87	117	16	155
	100	37	66	110	12	169
Group 2	0	65	110	120	24	115
	12.5	64	101	118	20	115
	25	60	99	120	19	126
	50	46	87	118	16	153
	100	35	64	109	11	172
Average value	0	66.5	110	120.5	23.5	113
	12.5	67.5	111	121.5	24.5	114
	25	68.5	112	122.5	25.5	115
	50	69.5	113	123.5	26.5	116
	100	70.5	114	124.5	27.5	117

3. Method for establishing quantitative model of material concentration and color reading

3.1. Analyze correlation of material concentration and color reading

Since the color reading consists of RGB values, firstly use MATLAB software to observe the relationship between the color and concentration of each group in the experimental data visually, then draw a line chart between the concentration and color R, G, B, H, and S respectively. Finally, the data is imported into SPSS for correlation analysis to find the relationship between concentration and color reading, to establish a mathematical model and test it.

Make a line chart of material concentration W and color reading R, G, B, H, S (Figure 1).

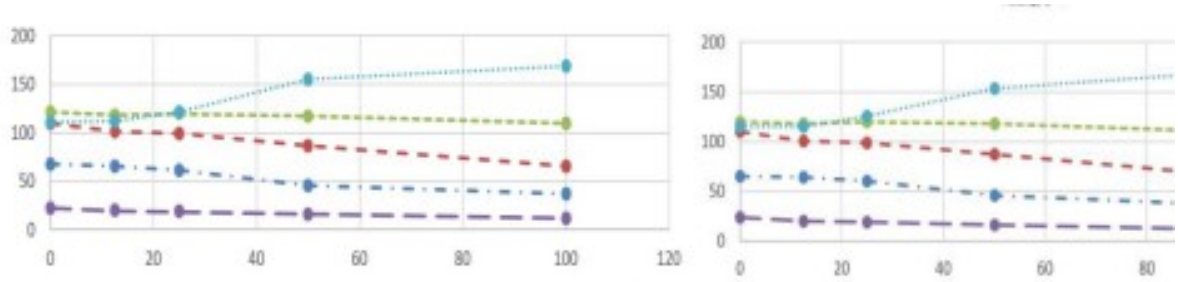


Figure 1. The line chart of histamine concentration and color reading

Similarly, a line graph of concentration and color readings for potassium bromate, industrial alkali, potassium aluminum sulfate, and urea in milk can be obtained, which can show the correlation between concentration and color reading visually.

In order to clarify the degree of correlation, a matrix of the relationship between the substance concentration and the color reading is obtained. The correlation coefficient is the amount of linear correlation between the variables studied. It is usually expressed by the letter r and is used to measure the linear relationship between the two variables. Equation (2) is as follows.

$$r = \frac{\sum_{i=}^{n} X_i - \bar{X} y_i - \bar{y}}{\sqrt{\sum_{i=}^{n} X_i - \bar{X}} \sqrt{\sum_{i=}^{n} y_i - \bar{y}}} \tag{2}$$

Among them $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

In general, after r takes the absolute value, 0-0.09 is not correlated, 0.1-0.3 is weak correlation, 0.3-0.5 is medium correlation, and 0.5-1.0 is strong correlation.

Correlation tests were carried out on the concentration and color dimensions of the five substances by Matlab software. The data are as follows:

Table 2. Correlation between concentrations and color readings of five substance

w	R	G	B	H	S
Histamine	-0.9357	-0.9975	-0.9760	-0.9808	0.9561
Potassium bromate	-0.1672	-0.8723	-0.9564	0.7023	0.9529
Industrial alkali	-0.6241	-0.6640	-0.4907	0.7084	0.6583
Potassium aluminum sulfate	-0.6776	-0.7013	0.5350	0.3452	0.6544
Urea in milk	-0.2611	0.1970	-0.9653	0.7740	0.9744

It can be determined from the above table that R, G, B, H, and S are strongly correlated with the concentration in the color reading of histamine. B, S have a higher correlation with the concentration in the color reading of potassium bromate. There is a correlation between the B and S of milk urea and aluminum. R, G, and S are not strongly correlated with the concentration in the color reading of potassium sulfate. And the concentration of industrial alkali has a strong correlation with R, G, H, and S.

3.2. Multiple linear regression models and tests for different substance concentrations and color readings

The formula of the linear regression model of the dependent variable Y and the K-factor variable X is as follows:

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

It is called Gauss-Markov linear model (k-ary linear regression model) and is abbreviated as $(Y, X\beta, \sigma^2 I_n)$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3}$$

$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ is called Regression plane equation.

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

Based on the idea of correlation, when a dependent variable has multiple independent variables that have a correlation, we can remove some of the weaker independent variables and retain the more relevant independent variables. We use r=0.5 as the threshold for the strength of correlation.

According to Table 1, the correlation coefficient between histamine concentration and the five color dimensions is greater than 0.5. This is a 5-element linear regression model of concentration W1 and color dimensions R, G, B, H, and S established by using MATLAB.

$$W_1 = \begin{bmatrix} 100 \\ 50 \\ 25 \\ 12.5 \\ 0 \end{bmatrix}, \quad X_1 = \begin{bmatrix} 1 & 109.5 & 65 & 36 & 11.5 & 170.5 \\ 1 & 117.5 & 87 & 46 & 16 & 154 \\ 1 & 120 & 99 & 61 & 19 & 124 \\ 1 & 118 & 101.5 & 65 & 20 & 113.5 \\ 1 & 120.5 & 110 & 66.5 & 23.5 & 113 \end{bmatrix}$$

$$\beta = \begin{matrix} 0 & 4.7962 & -5.5956 & -0.6929 & 6.7929 & -0.6724 \end{matrix}$$

We can get a 5-element linear regression equation:

W1 = 4.7962*R-5.5956*G-0.6929*B+6.7929*H-0.6725*S. MATLAB shows that the regression parameters: stats =1.0000, NaN, NaN, NaN, which passed R and F tests.

Similarly, we can obtain the linear regression equations of the other four substance concentrations of potassium bromate concentration W2, industrial alkali concentration W3, aluminum sulfate potassium concentration W4, milk urea concentration W5 and color reading which all passed the R and F tests.

W2 = 1867.6375+10.6250*G-23.0750*B-2.9249*H-12.1750*S

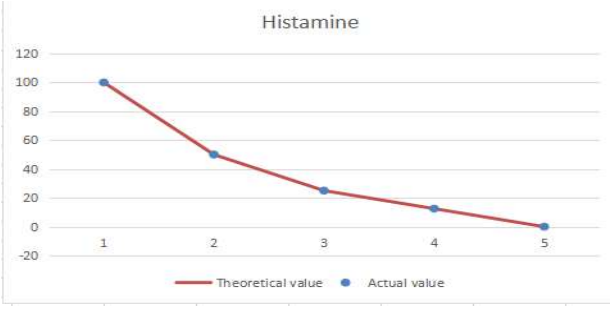
W3 = 585.3064-10.1742*R+7.5728*G-9.3144*H+0.9700*S

W4 = -67.2445+10.2728*R-6.8485*G—3.0072*B+4.7273*S

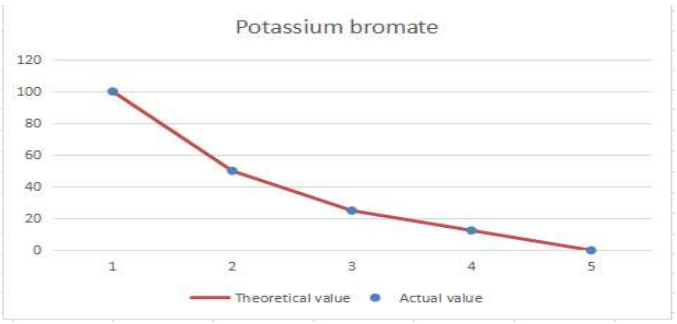
W5 = -16033.7619+104.1308*B-43.8576*H+135.2230*S

In order to test the accuracy of the regression model we obtained, the actual concentration value is compared with the theoretical concentration value. The relative error is small, and the correctness of the model is verified.

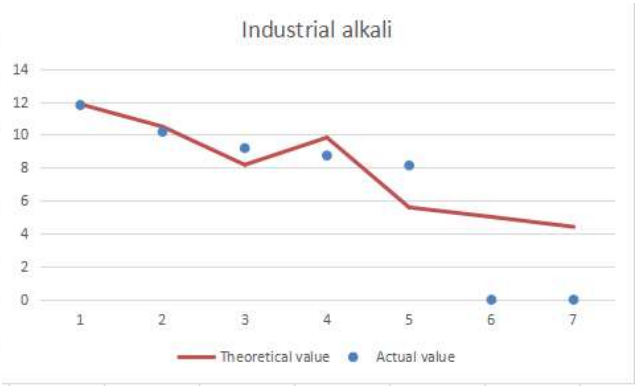
Histamine	
Actual value	Theoretical value
100	99.9826
50	49.9843
25	24.9878
12.5	12.489
0	-0.0111



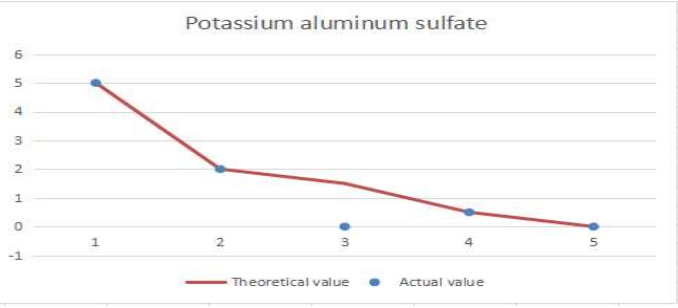
Potassium aluminum sulfate	
Actual value	Theoretical value
5	5
2	1.9999
1.5	1.4995
0.5	0.4997
0	-0.0003



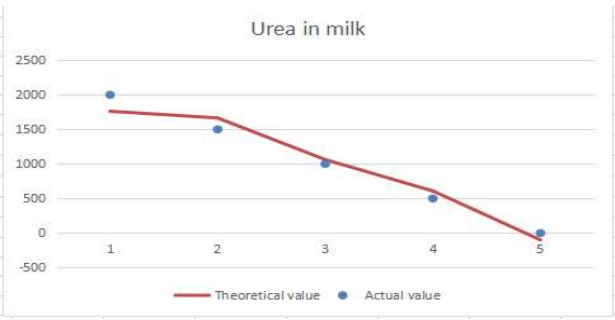
Urea in milk	
Actual value	Theoretical value
2000	1762.76
1500	1665.28
1000	1062.79
500	608.836
0	-99.687



Potassium aluminum sulfate	
Actual value	Theoretical value
5	5
2	1.9999
1.5	1.4995
0.5	0.4997
0	-0.0003



Urea in milk	
Actual value	Theoretical value
2000	1762.76
1500	1665.28
1000	1062.79
500	608.836
0	-99.687



3.3. Multiple linear regression model and test of the same substance concentration and color reading

Because both the RGB model and the HSV model can represent colors independently, the data in the attachment is indeed V, and the RGB model is selected to represent the color reading. We found that the material color readings R, G, and B are related to the concentration. But when the color readings R, G, and B have autocorrelation, we cannot use the three variables R, G, and B as the independent variable to analyze with concentration directly. After the data processed in Annex II is imported into MATLAB, the autocorrelation analysis is performed on R, G and B, as shown in the following table.

Table 3. Correlation coefficient between color readings RGB of sulfur dioxide

Correlation coefficient	R	G	B
R		0.988	-0.895
G	0.988		-0.916
B	-0.895	-0.916	

It is found that R, G, and B are strongly autocorrelated, and other independent variables should be selected for regression analysis. In graphics, grayscale has a high degree of recognition, and in medicine, CT has always used grayscale photos because of its high recognition. Therefore, we use the gray value instead of the R, G, and B readings. According to the literature [3], we get the gray scale calculation formula as follows:

$$HD=0.2989R+0.587G+0.114B \quad (4)$$

Import concentration and gray scale into MATLAB for linear regression, and get the model as: f(x) = -3.612*x+515.3. The R value is 0.742, indicating that the equation is significant, and the function image is as shown below:

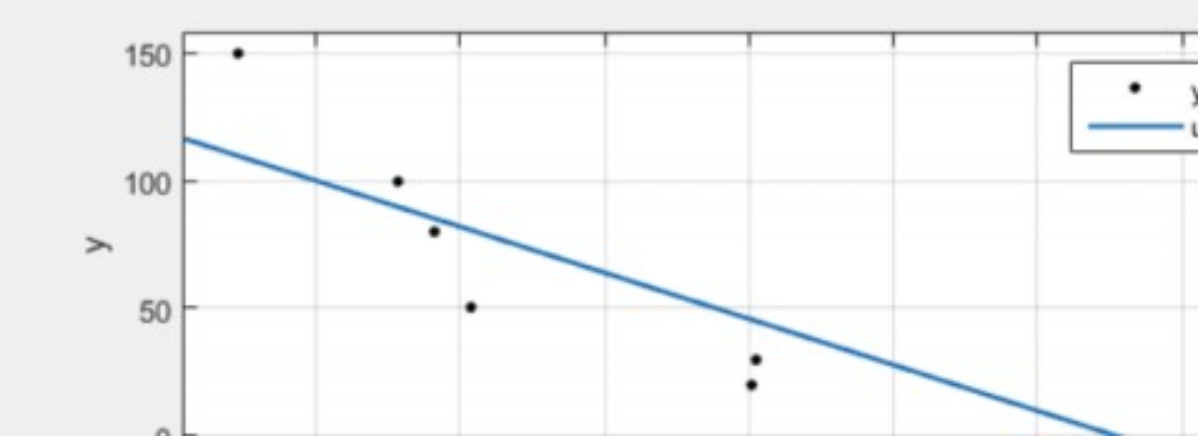


Figure 2. Scatter plot and linear regression image of sulfur dioxide concentration and gray scale

At the same time, the significance of the model is tested. The coefficient significance is 0.013, which showed the model is established. And then the linear regression model obtained by error-checking showed the maximum error is less than 3 times the RMSE, so the model error is small.

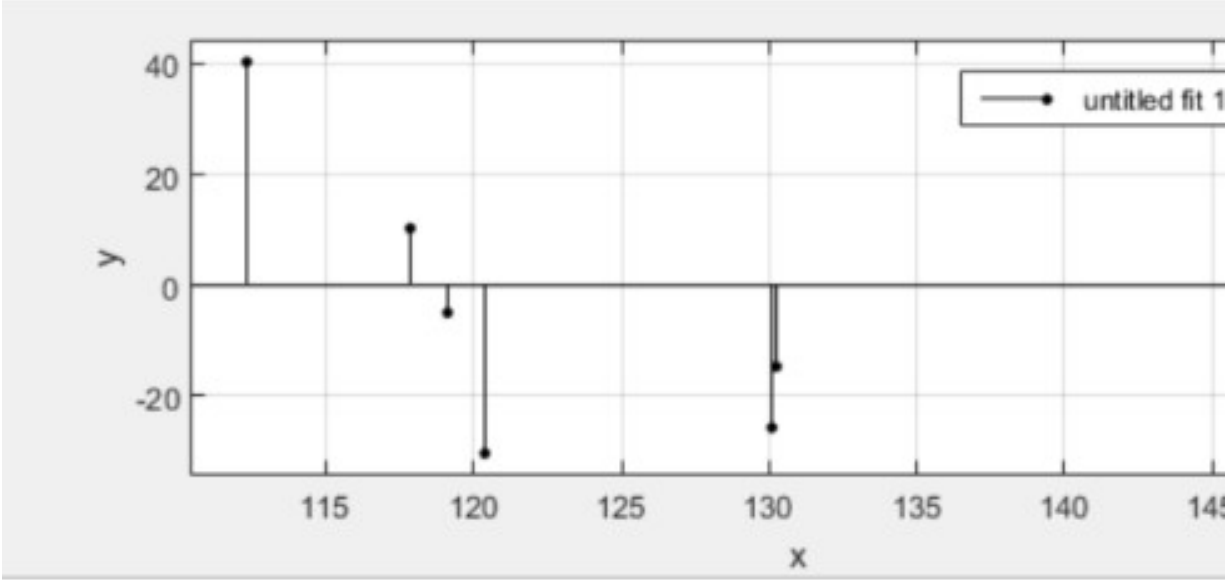


Figure 3. Error test of sulfur dioxide concentration and gray regression model

4. Results and discussion

4.1. The effect of color dimension on the model

R, G, B, S, H, and V are readings of six color dimensions, and three sets of data can form a color system. RGB and SH are readings for two different color systems. The data of the color dimension can be verified between different systems mutually. The five color dimensions and the data of two different color systems are given in the annex. The data of the SH system is incomplete, which brings great inconvenience to the mutual verification of the data of the two groups of systems, and makes the accuracy of the data lower,so the reliability of the returned model is not guaranteed.

4.2. The impact of data volume on the model

The amount of data has an impact on model building and testing. The industrial alkali with the least amount of data and the Potassium aluminum sulfate with the largest amount of data were selected for analysis in the attachment.

Industrial alkali data is only one group of 7 kinds of data, which are color readings of 7 different concentration. It is impossible to test the data under the same concentration, and it can not prove the reliability of the data. If abnormal values are found, it is difficult to find and process. Regression like this less data will result in an ideal model, but it will also reduce the reliability and prediction range of the model. There are 35 data of aluminum aluminum sulfate, and there is sufficient data for cross-validation between data, which improves the reliability and prediction range of the regression model, but at the same time.But the increase of data quantity also increases the possibility of occurrence of abnormal values, and single value anomalies can have a large impact on the overall analysis.

So too much or too little quantity of data will have different degrees of impact on the model. How to verify the accuracy of the data and deal with the outliers in the data will have a great impact on the model of data regression.

5. Conclusion

According to the color given in the question, drawing a line chart of the relationship between color and concentration, we can find the relationship between the two visually, and use the correlation coefficient method to judge whether the color reading of each substance has a correlation with the concentration. When the color reading RGB has autocorrelation, it is not allowed to use with RGB three variables as independent variables to make regression analysis with concentrations directly. To convert the RGB value into a gray scale with high recognition for regression analysis, the model was tested by R, F and errors.

References

[1] [United States] Hearn (D.), Baker (MP), Carithers (WR); Cai Shijie, Yang Ruoyu. "Computer Graphics: 4th Edition". Electronics Industry Publishing house, 2014.
[2] National College Students Mathematical Modeling Network http: // www. shumo. Com /home /html/ 3532.html.
[3] Baidu Library. From RGB color to gray color algorithm. https:/ /wenku. baidu. com/ view/ 0da4374549649b6648d747b2.html?from=searc h, 2015.9.11