Team Number :	1857

Problem Chosen :	A

2016 APMCM summary sheet

Temperature and key element content prediction based on optical information data

Summary

Based on optical information data in the metal smelting process, through principal component analysis (PCA) to extract characteristic index, making a multiple multivariate regression analysis model to reflect the relationship between temperature ***T***, the key elements ***C*** and various influence factors.

Aiming at the question one, using the PCA to simplify the 2048 optical information data of annex 1 for six characteristic indexes, This six characteristic indexes is the linear combination of the original data. Its cumulative variance contribution rates can be achieved 84.14%, It contains the main information of the original data, so we can research it instead of the original data, The first principal component characteristic index coefficient for annex 1 is (-0.0001,-0.0021,-0.0012,$\cdots$, 0.0241). The detailed data of characteristic indexes can be found in Body of the text.

Aiming at the question two, making standardized processing of data in annex 1 to get it unified in common numerical characteristics. Using characteristic indexes extracted from annex 1, the time interval, cumulative consumption of combustion gas as independent variable and the temperature, the key elements as dependent variable, building a multiple multivariate regression analysis model which can reflect the relationship between them. Drawing residual plot with MATLAB, manual removal of outliers to reduce errors. Residual sum of temperature ***T*** and the key elements ***C*** is $1.9993\times10^{-16}$and -5.0603×10^{-15}, correlation coefficient is 1 and 0.9998, which prove that model fitting effect is good. The function relation between temperature and various influence factors is T=0.0018+0.2787Q and the relation between key elements and various influence factors is C=0.0232+8.6824Q-0.0716λ_1+0.0096λ_2-0.0142λ_5 -0.0247λ_6.

Aiming at the question three, design the crossover experiment, cross-validate the error based on the model built in question two. Fitting forecast curve and actual data curve by MATLAB, we can find that the three model can be better self-prediction through analysis graph. But there are error when cross-validate between the model of process 1 and the model of process 2, 3 because of the difference of time. So we develop error control model. Adding a constant in the model to eliminating errors. Using the model of process 1 to predict the process 2, 3, figure out that the "compensation constant" of temperature and the key elements is 0.00072 and 0.0100. The prediction results is consistent with the original data based on the improved model.

Contents

1 Introduce

1.1 Restatement of the problem

Light contains energy, and it can be converted into heat in some certain conditions. Burning is an example as common phenomenon which can give both light and heat. Light and heat are usually concomitant, and generally speaking, the higher light intensity, the higher the temperature.

Metal smelting is a process that metal material thrown into a given melting pot will be smelt under a certain temperature in order to control or eliminate inferior element and remain or increase the proportion of superior element. Thus seek to acquire the best performance of objective metal. So the key factor of metal smelting lies in the control of temperature and the content of key elements. In the process of metal smelting, the furnace produces flame, then the optical information emitted by the flame is projected onto the photo detector through the theory of pinhole imaging. Next, according to the discrete frequency, the light intensity data of the flame is recorded at every 0.5s by the photo detector. Optical information data generation process at a certain time:

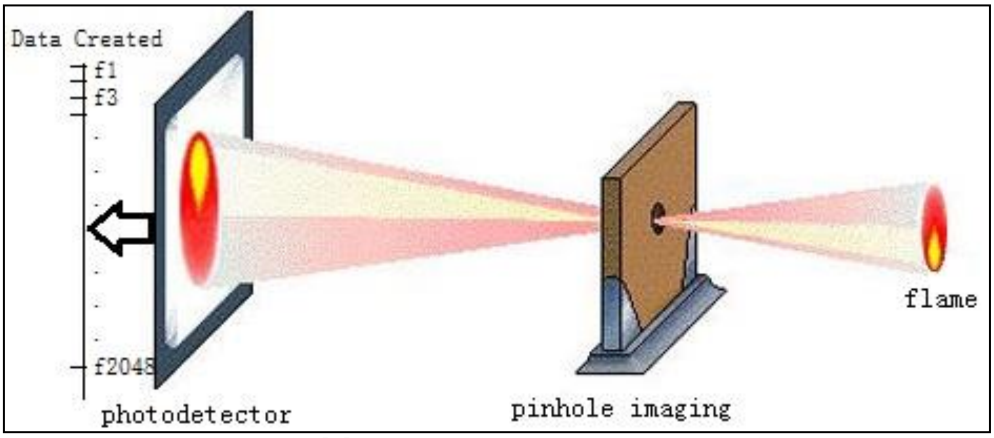

Fig. Optical data generation process at a certain time

From the image above, the amount of data received by the photo detector is 2048 during every 0.5 seconds. It is required to use the values of these 2048 light intensity to predict the flame temperature and the content of the key elements in the raw materials in real time.

Through the above background, explore the following three questions:

1)The mathematical model or algorithm model should be established by the data given in the annex to find the characteristic λ of the optical information data and extract the features (λ1, λ2,…, λn) into the Excel table. The data files name are: feature_output_1.xlsx、feature_output_2.xlsx and feature_output_3.xlsx.

2)The mathematical model or algorithm model should be established by using the optical information characteristic data extracted in 1) and the data given in the data table, such as the time t and the accumulated consumption Q of the combustion-supporting gas, to predict the Kelvin temperature T and key element content C and to explore the relationship.

3) Through the exploration of question 1) and question 2), design the crossover experiment scheme, cross-validate the error generated by the prediction target, and provide the error control scheme on the basis of the error analysis.

1.2 Analysis

During the metal smelting, the light intensity data of the flame is recorded at every 0.5s by the photo detector.at the same time, the burning gas's consumption was known in the additional contents. The question is analyzing those known data to predict how the light intensity and the every burning gas's consumption affection affect the smelting tempera true and the smelting metal's key elements, then check it.

In the first question, the data volume of annex is large, the calculation will be very large with direct analysis, and the results will be not accurate.so we need to develop a mathematical model, and extract the main features from the give optical information data. The main feature should be able to better reflect the whole, and reduce the complexity of the model calculation, and enhance the applicability of the model with accuracy. The PCA method is used for modeling, the PCA can simplify the original data to several new composite indicators and retain the vast major information of original data. We can use PCA method to give the 2048 sets of data to reduce the dimension of the integration, analyzing the main characteristic indicator. What's more, we can use the same way to attach the coefficients to annex 2 and 3, then get the principal component characteristic of the annex 2 and 3.

For question two, in the base of the extracted principal component feature λ, I add the time t and the gas consumption Q as independent variables Kelvin Temperature T and metal key elements C, so we can develop a model to predict their relationship. Because the dimension of each variable data and the order of magnitude may be different, we take the normalization to standard to reduce the errors. Analysis can be known, either the independent or dependent variables, they have multiple inputs and multiple responses. So we consider taking multiple regression analysis model to develop it. Test it with Residual analysis.

For the question three, we are asked to design cross-over experiments, the meaning is using each different process model to predict each data, then compared with actual data to analyze error. In the question one and two, the multiple regression analysis model of temperature T and the key element C and the various influencing factors have been developed, across-over test the three models and data. For example, using the model of process one to predict the three processes separately and test separately, and with the help of MATLAB, drawing the predict data curve and the actual data curve, observing the curves' coincidence degree, then analyzing the cause of errors and reducing it.

2 Assumptions and Notations

2.1 Assumptions

- Flame temperature ***T*** is not affected by other environmental factors.
- The optical information data given is accurate.
- Flame of optical information through holes imaging after it won't change or loss.
- Temperature ***T*** and key elements of ***C*** content is only related to optical information，time *t*， the cumulative consumption of combustion gas ***Q***.

2.2 Notations

Symbol	Meaning
λ	the characteristic indexes
x	the original data indexes
R	correlation coefficient matrix
γ	the correlation coefficient matrix eigenvalues
H	the correlation coefficient matrix eigenvalue vector
f	contribution rate
F	cumulative contribution rate
b_{ij}	coefficient of the original data values
Σ	covariance matrix
β_T	kelvin temperature T of the regression coefficient
β_C	content of key element C of the regression coefficient
b	error Compensation constant

3 The Models

3.1 The developing and solving model of question 1

There are too many indexes which can influence the Kelvin temperature ***T*** and key element content ***C***. the quantity of data is too large and the feasibility is low if analyze them one by one. So we develop the model by the PCA method, then solve the data and extract characteristic indexes from it.

Suppose the original indexes as $x_1, x_2, \cdots, x_n$, and we can get the characteristic indexes $\lambda_1, \lambda_2, \cdots, \lambda_k$ by the PCA. The characteristic indexes are linear combinations of the original indexes. The Model of Principal Components Analysis are the followings:

- Develop matrix of the characteristic indexes

Suppose there are n indexes, m groups of data, and develop the characteristic indexes matrix as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

- Standardized treatment

$$x_{ij}^* = (x_{ij} - \overline{x}_i) / \delta_i \quad (i = 1, 2, \cdots, n, \ \ j = 1, 2, \cdots, m)$$

The x_{ij}^* is the number i index in the formula and the standardized value of the number j group data. The x_{ij} is the number i index in the formula and the value of the number j group data. The x_i is the sample mean and the δ_i is the standard deviation of the number i index.

- Compute the correlation coefficient matrix

$$R = (r_{ij})_{n \times n}$$

$$r_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) / \delta_i \delta_j$$

- Compute the eigenvalue and eigenvector of R

We can get the eigenvalue γ_1, $\gamma_{2,}, \cdots$, γ_n through the characteristic equation $\left| R - \gamma I \right| = 0$ and arrange them in order from lowest value to highest value.as the same way, we can get the eigenvector $H_1, H_2, \cdots, H_n$.

- Compute the contribution rate and the accumulating contribution rate

The contribution rate:

$$f_k = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i}$$

The accumulating contribution rate

$$F_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

In general, in the computing of models, the accumulating contribution rate can be principal component if it reaches 84%.

- Suppose $\lambda_1, \lambda_2, \cdots, \lambda_k$ separately represents the number 1, number 2, … , number k principal component, search for the linear combination of the original index according to the principle of the principal component. and $\lambda_1, \lambda_2, \cdots, \lambda_k$ can be showed as the followings:

$$\begin{cases} \lambda_1 = b_{11}x_1 + b_{12}x_2 + \cdots + b_{1j}x_j + \cdots + b_{1n}x_n \\ \lambda_2 = b_{21}x_1 + b_{22}x_2 + \cdots + b_{2j}x_j + \cdots + b_{2n}x_n \\ \cdots \\ \lambda_i = b_{i1}x_1 + b_{i2}x_2 + \cdots + b_{ij}x_j + \cdots + b_{1n}x_n \\ \cdots \\ \lambda_k = b_{k1}x_1 + b_{k2}x_2 + \cdots + b_{kj}x_j + \cdots + b_{kn} \end{cases}$$

b_{ij} is the Pearson Correlations of the number *j* original index under the number *i* principal component, the coefficient sum of square of every principal component is 1:

$$b_{i1}^2 + b_{i2}^2 + \cdots + b_{in}^2 = 1$$

There are no overlapping messages, and the principal component is independent between each one:

$$\mathrm{cov}(\lambda_i, \lambda_j) = 0, \quad i \neq j, i, j = 1, 2, \cdots, k$$

The variances of the principal component and the importance are lowered accordingly:

$$\mathrm{var}(\lambda_1) \geq \mathrm{var}(\lambda_2) \geq \cdots \geq \mathrm{var}(\lambda_p)$$

$\lambda_1, \lambda_2, \cdots, \lambda_k$ are separately called the number 1,number 2,…,number *k* principal component of the original variables.

The data has 2048 indexes in schedule 1, and every index has 404 data. After taking standardized treatment to the data, then use the Model of Principal Components Analysis developed above. through MATLAB program, we can get eigenvalue， eigenvector, contribution rate and accumulating contribution rate of the correlation coefficient matrix and 6 characteristic index. And relational data is showed in table 1:

Table 1 the principal component data of the contribution rate

Principal component	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
Eigenvalue	1674.0630	23.3358	14.9123	4.1208	3.4986	3.3431
Contribution rate	0.8174	0.0114	0.0073	0.0020	0.0017	0.0016
Accumulating contribution rate	0.8174	0.8288	0.8361	0.8381	0.8398	0.8414

We can know the accumulating contribution rate of the 6 characteristic indexes reaches 84.14%, including a lot of messages of light intensity data. The coefficients are showed in table 2.

Table 2 coefficients of the characteristic index

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
x_1	-0.0001	0.0048	0.0214	-0.4458	0.0260	-0.0195
x_2	-0.0021	0.0030	-0.0182	-0.0654	0.0039	-0.0051
x_3	-0.0012	-0.0042	0.0264	-0.0151	-0.0097	-0.0499
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{2048}	0.0241	-0.0289	-0.0081	-0.0009	0.0011	-0.0013

Substitute the coefficient b_{ij} of the characteristic indexes λ_i and the original variables x_i into the formula:

$$\lambda_i = b_{i1}x_1 + b_{i2}x_2 + \cdots + b_{ij}x_j + \cdots + b_{1n}x_n \quad (i = 1,2,\cdots,6;\ j = 1,2,\cdots 2048)$$

Then we can get the characteristic indexes. And the indexes are showed in table 3(partly).the whole data can be seen in the annex in excel table.

Table 3 the data of the characteristic indexes of annex1 (partly)

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0.00	237317.2	-22502.2	-6297.4	748.8	-1244.2	-439.9
0.50	175327.9	-20358.5	-6630.8	152.0	-1494.2	562.0
1.00	199399.9	-21150.5	-6573.5	335.2	-1377.6	191.8
…	…	…	…	…	…	…
143.50	37832.5	1393.0	-3802.6	-832.2	-2352.7	3001.7

Substitute coefficients of the characteristic indexes from annex 1 into annex 2 and 3, then we can get the coefficient value of the characteristic indexes of annex 2 and 3, as showed in the schedule.

Table 4 the data of the characteristic indexes of annex2 (partly)

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0.00	152236.4	-25559.0	-8637.8	-292.9	-1837.6	1485.1
0.50	145816.8	-24469.1	-8606.1	-256.9	-1989.4	1529.9
1.00	152528.3	-19370.8	-7227.9	-183.4	-1938.4	1404.2
…	…	…	…	…	…	…
143.50	71539.8	-2265.3	-7001.6	-853.8	-2373.8	2802.2

Table 5 the data of the characteristic indexes of annex3 (partly)

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0.00	88571.3	-9602.9	-6358.5	-694.5	-2110.1	2451.6
0.50	69354.5	-5251.3	-5177.8	-892.6	-2155.0	2812.6
1.00	138729.3	-22482.0	-7910.5	-344.6	-1797.2	1631.2
…	…	…	…	…	…	…
143.50	66491.8	-3391.5	-21720.7	-1000.6	-2257.5	2970.6

3.2 The developing and solving model of question 2

We put forward 6 characteristic indexes from the given data of every annex through the question 1. Question 2 demands we combine the extracted characteristic indexes, think about the time interval t and the cumulative consumption of combustion gas $\boldsymbol{Q}$, and research the relationship between Kelvin temperature ***T***, content of key element ***C*** and them. By analyzing, we can know that there are multivariate input and multiresponse. For this, we develop multiple multivariate regression analysis model to research.

- data pretreatment

The dimensions of characteristic indexes and magnitude given by the subject may not be the same.it may influence the accuracy of the results. So it is necessary to take pretreatment to the data. We take standardized treatment to make every index value in the same intrinsic numeric area.

$$Y_{ij} = \frac{x_{ij} - \bar{x}_j}{\delta_j} \quad (i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, w)$$

Y_{ij} is standardized value in the formula, $\bar{x}_j$ is the mean of the number j index, δ_j is the standard deviation of the number j index.

- The developing of the Multiple multivariate regression analysis model

1) Multiple multivariate regression parameter estimation

Suppose there are *m* independent variables and *n* groups of observations of the *p* independent variables, apart showed by data matrix X, Y:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix},\ Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

Suppose B as regression coefficient and E as residue, they can be showed like this:

$$\beta = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{m1} & \beta_{m2} & \cdots & \beta_{mp} \end{bmatrix} = \begin{bmatrix} \beta'(0) \\ \beta'(1) \\ \vdots \\ \beta'(m) \end{bmatrix},\ E = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{np} \end{bmatrix} = \begin{bmatrix} e'(1) \\ e'(2) \\ \vdots \\ e'(n) \end{bmatrix}$$

The multiple linear regression model can be expressed as

$$\begin{cases} Y = (\ln X)\beta + E = C\beta + E \\ E(i) \sim N_p(0, \Sigma) \quad i = 1, 2, \cdots, n; \text{相互独立} \end{cases}$$

Σ is Variance - covariance matrix of residual $e(i)$, C is a design matrix of $n \times (m+1)$. According to the least squares method, the least squares estimate of β is:

$$\hat{\beta} = (\mathrm{C}^T C)^{-1} C^T Y$$

- The test of Multiple multivariate regression model

Regression model test is an important work of regression analysis. Usually first intuitive residual analysis, get a preliminary judgment of whether the model is correct or not, Then make significance test. Among it, visual inspection mainly includes the test of residual normal concept map and the test of the relation between residuals and predicting variables. When make significance test of multiple multivariate regression model, we need consider if any independent variable $x(i)$ have a significant impact on the dependent variable Y or not. That is to say, we should have a test on regression coefficient $\beta_1 = (\beta_{(1)}, \beta_{(2)}, \cdots, \beta_{(i)}, \cdots, \beta_{(m)})$:

$$H_0 = \beta_{(1)} = \beta_{(2)} = \cdots = \beta_{(m)} = 0$$

The corresponding test statistics is:

$$\frac{MSS / f_{MSS}}{ESS / f_{ESS}} \sim E(f_{MSS}, f_{ESS})$$

MSS is regression sum of squares, ESS is Residual sum of square, f_{MSS}, f_{ESS} is the corresponding degrees of freedom.

- solution method for model

According to multiple linear regression model, using function "regress" in the MATLAB to calculate the residual and confidence interval; using "rocoplot" draw the residuals analysis graph like the figure 1.

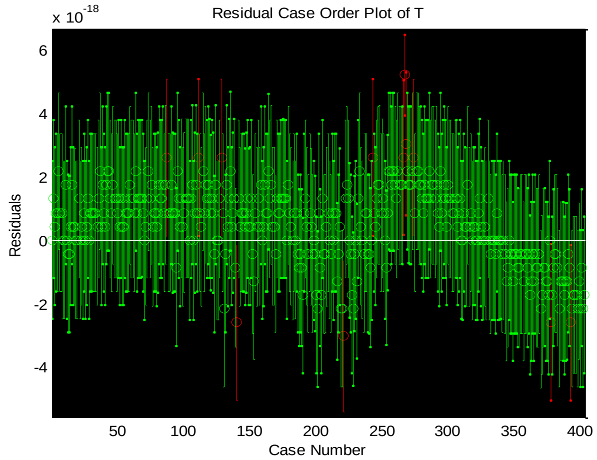

Figure 1 The residuals analysis graph of temperature ***T***

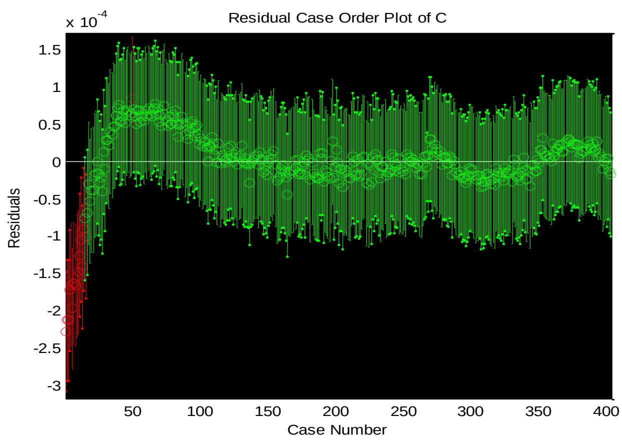

Figure 2 The residuals analysis graph of the key elements ***C***

Seen from the picture above, there are some outliers in the observation data, in order to get more accurate regression coefficient, manual removal of outliers, then make regression analysis again to get regression coefficient of ***T*** and ***C***:

Table 6 regression coefficient of ***T*** and ***C***

β_{T_0}	β_{T_1}	β_{T_2}	β_{T_3}	β_{T_4}	β_{T_5}	β_{T_6}	β_{T_7}	β_{T_8}
0.0018	0	0.2787	0	0	0	0	0	0
β_{C_0}	β_{C_1}	β_{C_2}	β_{C_3}	β_{C_4}	β_{C_5}	β_{C_6}	β_{C_7}	β_{C_8}
0.0232	0.0016	−8.6824	−0.0716	0.0096	0.0039	−0.0018	−0.0142	−0.0247

We can get the expression of temperature ***T*** and the key elements ***C***:

$$T = 0.0018 + 0.2787Q$$

$$C = 0.0232 + 0.0016t - 8.6824Q - 0.0716\lambda_1 + 0.0096\lambda_2$$
$$+ 0.0039\lambda_3 - 0.0018\lambda_4 - 0.0142\lambda_5 - 0.0247\lambda_6$$

At the same time, we can programming to calculate residual sum, coefficient of association, λ statistical magnitude and the probability of the corresponding. They can be seen in table 7.

Table 7 the data of residual sum and so on

	residual sum	coefficient of association	F coefficient	P
T	1.9993×10^{-16}	1	1.6118×10^{29}	0
C	-5.0603×10^{-15}	0.9998	2.6957×10^{5}	0

From the table, we can find that the residual sum of ***T*** and ***C*** is very small, so the model can reflect the relationship between the independent variable and dependent variable accurately.

3.3 The developing and solving model of question 3

Based on modeling of the problem one and two, we can get the relation between temperature ***T***, the key elements ***C*** and various influence factors. And design cross-over experiment to predicting test.

- The solution of relational expression of annex 2

We have get the coefficient of characteristic indexes in question 1 and extract the characteristic index from annex 2, considering time *t* and the key elements ***C***, working out the relation between ***T*** and ***C*** of annex 2.

$$T_2 = 0.0026 + 0.2594Q$$

$$C_2 = -0.0139 - 0.0181t + 7.0815Q + 0.2051\lambda_1 - 0.0174\lambda_2$$
$$- 0.0075\lambda_3 - 0.0091\lambda_4 + 0.0331\lambda_5 - 0.0300\lambda_6$$

- The design of the cross experiment

In order to test the universal applicability and accuracy of the model, design cross experiment, using the model on different annexes data, Compared prediction curve with accurate curve to analysis the effect of model, and test error. Listing the error analysis table of the cross experiment at first. As shown in table 8.

Table8 The error analysis table of the cross experiment

Data Model	Prediction model based on 1 process	Prediction model based on 2 process	Prediction model based on 3 process
1 process data	$Err11$	$Err12$	$Err13$
2 process data	$Err21$	$Err22$	$Err23$
3 process data	$Err31$	$Err32$	$Err33$

Err11 represents the self test error based on the prediction model of 1 process, Err21 represents the crossover inspection error based on the prediction model of 1 process and the data of 2 process, and so on.

Using the model of process 1 to predict three process, Using ***T*** and ***C*** as ordinate, influence factor as abscissa, through MATLAB fitting prediction curve and actual data curve, the fitting result of process 1 can be seen in figure 3, 4.

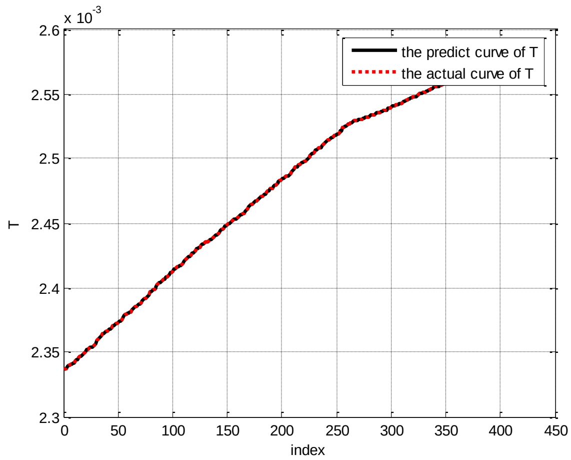

Figure 3 Self test curve of the temperature T of process 1

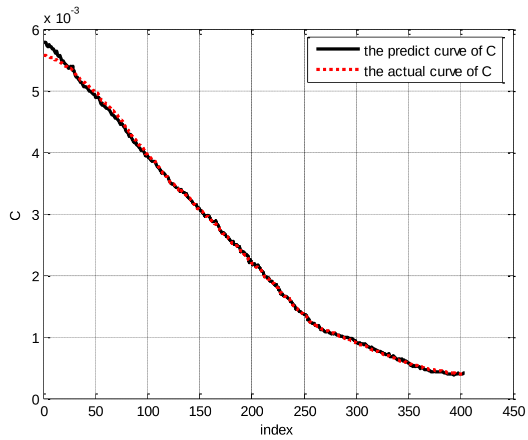

Figure 4 Self test curve of the key elements of process 1

From the figure we can find that predict data of ***T*** and ***C*** are consistent with actual data, so the model have high accuracy, the error of self test is small. In like manner, make self test of process 2 and 3, the result is satisfactory. The self test provide that the model have high accuracy. Now we need make cross test to analysis the general applicability of the model, Using the model of process 1 to predict process 2, figure out the predict curve through MATLAB. As figure 5, 6 shown.

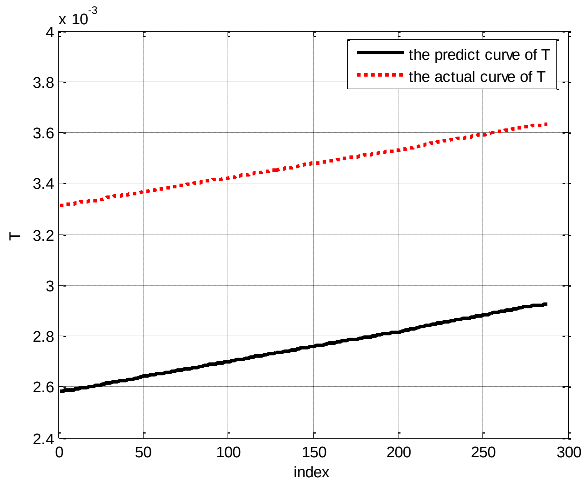

Figure 5 predict curve of the ***T*** of process 2 based on model of process 1

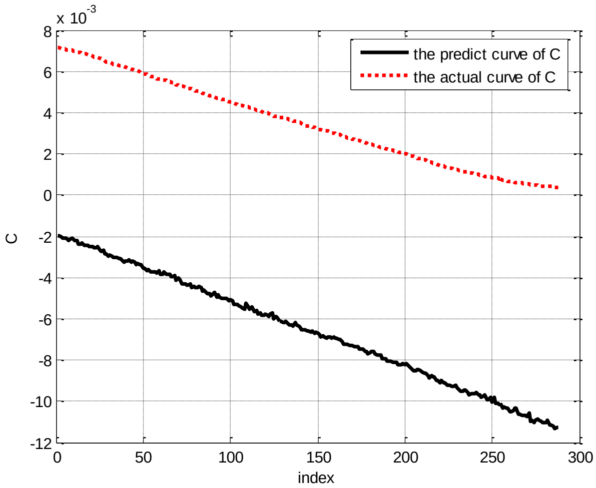

Figure 6　Predict curve of the ***C*** of process 2 based on model of process 1

As the picture 5 and 6 show, the variation tendency of the estimated curve and the original data curve are basically the same, so we can easily get that the difference between the estimated data from the model and the real data is a constant. Compare the data through process 1 and 2, we can find that the difference between the two process is the length of time. So we get the inference that the reason of the existing error may be the time.by further observation, the length of the time in process 2 and 3 is basically the same. Then use the model of the process 2 to estimate the data of process 3. The fitting results of the estimated and the real curves are as the followings:

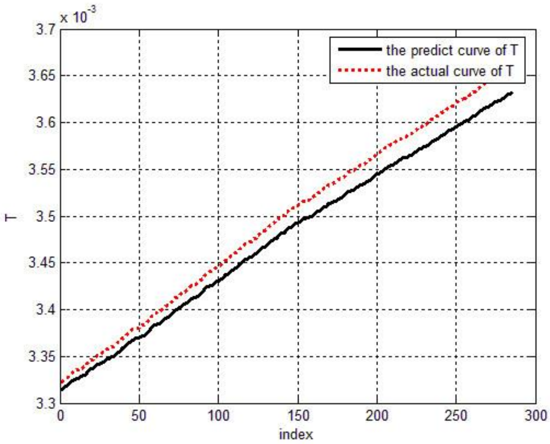

Figure 7　Predict curve of the ***T*** of process 3 based on model of process 2

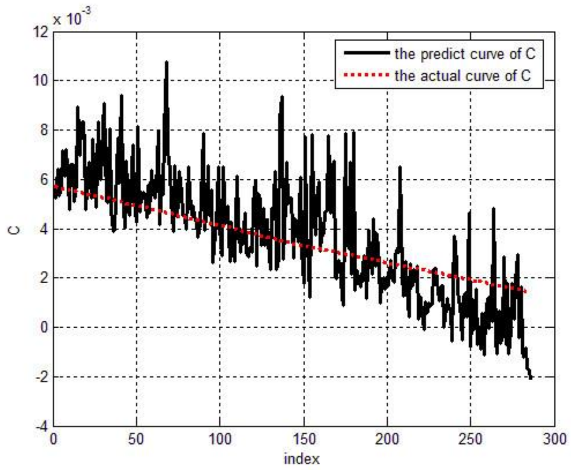

Figure 8 Predict curve of the ***C*** of process 3 based on model of process 2

The estimated curve and the original data curve in the picture are basically the same, the process 2 can better estimate the process 3. Then use the model from the process 3 to estimate the process 2, which turns out to be good. On the contrary, we can find that the difference between the estimated data from the model and the real data is a constant by using the models from process 2 and 3 to separately estimate the process 1. These groups of the crossover experiment confirm that the error is caused by the different time of the process. For this, we have to design a control scheme to improve the models and the universality of the models.

- The model of the error control scheme

For the difference between the estimated data from the model and the real data is a constant, we need to improve the control scheme. Suppose the estimated value of the number i data is m_i, and the real value is s_i, and there are n data, cause

$$b = \frac{\sum_{i=1}^{n}(s_i - m_i)}{n} \quad (i = 1, 2, \cdots, \ n)$$

The b in the formula is error compensation constant.

So the improved expression of the Kelvin temperature ***T*** is

$$T = \beta_{T_0} + b_T + \beta_{T_1} t + \beta_{T_2} Q + \beta_{T_3} \lambda_1 + \cdots + \beta_{T_8} \lambda_6$$

The improved expression of the content of the key element ***C*** is

$$C = \beta_{c_0} + b_c + \beta_{c_1} t + \beta_{c_2} Q + \beta_{c_3} \lambda_1 + \cdots + \beta_{c_8} \lambda_6$$

The b_T in the formula presents the compensation constant of the Kelvin temperature ***T***, and the b_c present compensation constant of the key element ***C***. Using the improved models basically from process 1 to estimate the process 2. Through the model of the control scheme, we can count the compensation constant b_T of the Kelvin temperature ***T*** is 0.00072, the compensation constant b_c of the key element ***C*** is 0.0100. By putting the value of b_T and b_c into the expression of the Kelvin temperature ***T*** and the content of the key element ***C***, we can get the improved estimated curve, as showing in picture

9 and 10

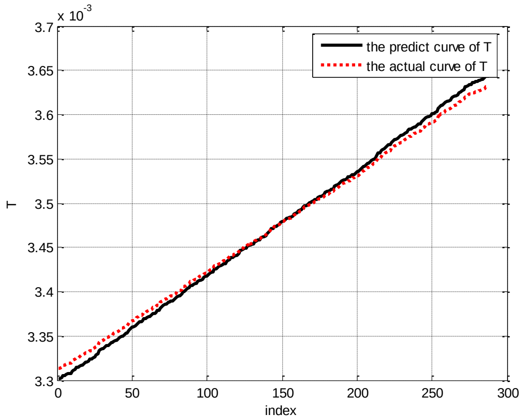

Figure 9 Predict curve of the ***T*** of process 2 based on improved model of process 1

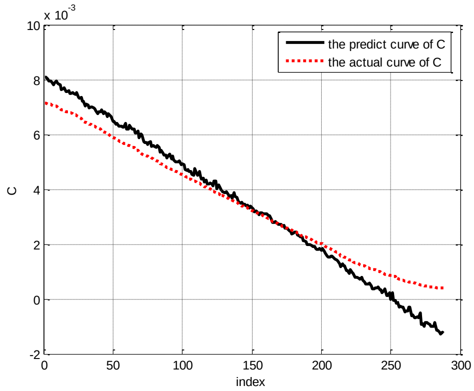

Figure 10 Predict curve of the ***C*** of process 2 based on improved model of process 1

By analyzing the pictures, we can know that the improved models can greatly estimate the real data and improve the universality of models.

4 Model evaluation

4.1 Advantages of model

- Question 1 using the method of principal component analysis in 2048 optical information extracted six characteristic indexes, greatly reduce the number of input data and reduce the complexity of the model, to improve the applicability of the model.
- Question 2 set up a multiple linear regression model to solve the relational expression of the temperature ***T*** and the key element ***C*** and the each various. Through the fitting of expression, the curve of the prediction was consistent with actual data curve, fully shows the accuracy of the model.
- Question 3 was design the crossover experiment scheme, pass one of the attachment data to establish the model to simulate all the attachment data, repass proper error control, the resulting forecasts also base is consistent with the actual data,

to prove the universality of the model.

4.2 Disadvantage of the model

- In the regression model of question 2 ， think about temperature T and the key element C and the various is linear relation，have certain blindness, the results may be to have influence.

References

[1] SUN Cheng-bao , LI Yun-li, The Principal Components Analysis and the Global Principal Components Analysis in the Environmental Quality Evaluation，natural science，Vol.20 No.2：2006.
[2] YUE lihua，DUAN Wuduo，probability and mathematical statistics，Nanchang：Jiangxi university press，2015.
[3] ZHANG Jinchun, WU Chao， Coke Quality Prediction Model Based on Multivariate Regression，safety engineering，Changsha：2010.
[4] YANG xiuhua, DING wangxian，Using the regression analysis opposite coking production forecast and control，Coal Quality Technology，2004（6）：25-26.

Appendices

CODE of QUESTION1

```
myload.m
clear,clc
A1 = xlsread('.\2016 APMCM Problem\Problem A\1.xlsx',    'D3:BZW406');
tq1 = xlsread('.\2016 APMCM Problem\Problem A\1.xlsx',    'A3:B406');
tc1 = xlsread('.\2016 APMCM Problem\Problem A\1.xlsx',    'BZX3:BZY406');
tqtc1=[tq1 tc1];
A2 = xlsread('.\2016 APMCM Problem\Problem A\2.xlsx',    'D3:BZW290');
tq2 = xlsread('.\2016 APMCM Problem\Problem A\2.xlsx',    'A3:B290');
tc2 = xlsread('.\2016 APMCM Problem\Problem A\2.xlsx',    'BZX3:BZY290');
tqtc2=[tq2 tc2];
A3 = xlsread('.\2016 APMCM Problem\Problem A\3.xlsx',    'D3:BZW288');
tq3 = xlsread('.\2016 APMCM Problem\Problem A\3.xlsx',    'A3:B288');
tc3 = xlsread('.\2016 APMCM Problem\Problem A\3.xlsx',    'BZX3:BZY288');
tqtc3=[tq3 tc3];

save('A1.mat','A1');
save('A2.mat','A2');
save('A3.mat','A3');
```

```
save('tqtc1.mat','tqtc1');
save('tqtc2.mat','tqtc2');
save('tqtc3.mat','tqtc3');

table1.m
clear,clc
load A1;
cor_A1=corrcoef(A1);
[m,n]=size(cor_A1);
[var,val]=eig(cor_A1);
newval=diag(val);
[sort_val,index]=sort(newval,'descend');
rate=[];
num=0;
while true
        num=num+1;
        rate(num)=sort_val(num)/sum(sort_val);
        if sum(rate)>0.84
                break;
        end
end

F1=[];
xishu=[];
for i=1:num
        xishu(:,i)=var(:,index(i));
      F1(:,i)= A1*var(:,index(i));
end
furture=newval(index(1:num))';
accum_rate(1)=rate(1);

for i=2:num
      accum_rate(i)=accum_rate(i-1)+rate(i);
end
zong=[furture;rate;accum_rate]';
save('F1.mat','F1');
save('xishu.mat','xishu');
xlswrite('.\feature_output_1.xls',F1);


table2.m
clear,clc
load A2
load xishu
F2=A2*xishu;
save('F2.mat','F2');
xlswrite('.\feature_output_2.xls',F2);

table3.m
clear
```

```
clc
load A3
load xishu
F3=A3*xishu;
save('F3.mat','F3');
xlswrite('.\feature_output_3.xls',F3);
```

CODE of QUESTION2

Standardization.m

```
function [standF standTqtc]=Standardization(F,tqtc)
for i=1:size(F,2)
    F(:,i)=F(:,i)/sum(F(:,i));
end

for i=2:4
    tqtc(:,i)=tqtc(:,i)/sum(tqtc(:,i));
end
tqtc(:,1)=tqtc(:,1)/162.5;

standF=F;
standTqtc=tqtc;
end
```

Picture.m

```
function []=Picture(T,T_y,C,C_y)
figure
plot(1:length(T_y),T_y,'k-','LineWidth',2.5);
hold on
plot(1:length(T),T,'r:','LineWidth',2.5);
xlabel('index');
ylabel('T')
legend('the predict curve of T ','the actual curve of T');
grid on



figure
plot(1:length(C_y),C_y,'k-','LineWidth',2.5);
hold on
plot(1:length(C),C,'r:','LineWidth',2.5);
xlabel('index');
ylabel('C')
legend('the predict curve of C ','the actual curve of C');
grid on
end
```

CrossoverExperiment.m

```
function []=CrossoverExperiment(T_b,C_b,F,tqtc)

[standF standTqtc]=Standardization(F,tqtc);

x=[ones(size(standF,1),1) standTqtc(:,1:2) standF];
T=standTqtc(:,end-1);
C=standTqtc(:,end);

T_y=x*T_b;
C_y=x*C_b;

Picture(T,T_y,C,C_y);
end
```

question2.m

```
clear,clc
load F1
load tqtc1
[standF1 standTqtc1]=Standardization(F1,tqtc1);

x1=[ones(size(standF1,1),1) standTqtc1(:,1:2) standF1];
T1=standTqtc1(:,end-1);
C1=standTqtc1(:,end);

[T_b1,T_bint1,T_r1,T_rint1,T_stats1]=regress(T1,x1);
[C_b1,C_bint1,C_r1,C_rint1,C_stats1]=regress(C1,x1);

figure(1)
rcoplot(T_r1,T_rint1);
figure(2)
rcoplot(C_r1,C_rint1);

index=find(abs(C_r1)>=0.055*10^-3);
x1(index,:)=[];
T1(index,:)=[];
C1(index,:)=[];
[C_b1,C_bint1,C_r1,C_rint1,C_stats1]=regress(C1,x1);

T_y1=x1*T_b1;
C_y1=x1*C_b1;

Picture(T1,T_y1,C1,C_y1)

save('T_b1.mat','T_b1');
save('C_b1.mat','C_b1');
```

CODE of QUESTION3

question3_1.m
```
clear,clc
load F2
load tqtc2
[standF2 standTqtc2]=Standardization(F2,tqtc2);

x2=[ones(size(standF2,1),1) standTqtc2(:,1:2) standF2];
T2=standTqtc2(:,end-1);
C2=standTqtc2(:,end);

[T_b2,T_bint2,T_r2,T_rint2,T_stats2]=regress(T2,x2);
[C_b2,C_bint2,C_r2,C_rint2,C_stats2]=regress(C2,x2);

save('T_b2.mat','T_b2');
save('C_b2.mat','C_b2');
```

question3_2.m
```
clear,clc
load F2
load tqtc2
load F3
load tqtc3
load T_b1
load C_b1
load T_b2
load C_b2
CrossoverExperiment(T_b1,C_b1,F2,tqtc2);
CrossoverExperiment(T_b1,C_b1,F3,tqtc3);
CrossoverExperiment(T_b2,C_b2,F3,tqtc3);
```

question3_3.m
```
clear,clc
load F2
load tqtc2
load F3
load tqtc3
load T_b1
load C_b1

[standF2 standTqtc2]=Standardization(F2,tqtc2);

x2=[ones(size(standF2,1),1) standTqtc2(:,1:2) standF2];
T2=standTqtc2(:,end-1);
C2=standTqtc2(:,end);

T_y2=x2*T_b1;
```

```
C_y2=x2*C_b1;

T_constant=mean(T2-T_y2)
C_constant=mean(C2-C_y2)

T_b1(1)=T_b1(1)+T_constant;
C_b1(1)=C_b1(1)+C_constant;

CrossoverExperiment(T_b1,C_b1,F2,tqtc2);
CrossoverExperiment(T_b1,C_b1,F3,tqtc3)
```