For office use only
T1 _____
T2 _____
T3 _____
T4 _____

Team Control Number

00004

Problem Chosen

A

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2019 Mathematical Contest in Modeling (MCM) Summary Sheet
(Attach a copy of this page to each copy of your solution paper.)

Metal smelting prediction problem based on random forest and multiple nonlinear regression

Summary

Temperature control and critical elemental content are important for achieving optimum performance of the target metal during metal smelting. We studied the relationship between the Kelvin temperature and the key element content, and analyzed the causes of the error and gave a way to control the error.

The first stage, we assume that more than 85% of the eigenvalues of the information can be used to replace all of the original optical information feature data. In order to avoid the influence of excessive factors on Kelvin temperature and prediction of key element content, the feature selection model was established by using principal component analysis combined with random forest, and the optical information characteristic data affecting Kelvin temperature and key element content in three processes were finally determined. The category is its corresponding contribution rate f_691,f_212,f_1197.

The second stage, we established a multivariate nonlinear regression model. The model first judges the relationship of the dependent variable and then selects the appropriate function to determine the relationship between Kelvin temperature and the content of key elements and the time and cumulative consumption of optical information feature data. The program yields a function of the dependent variable Kelvin temperature and the content of key elements in the first metal smelting process. The fitting effect is good.

The third stage, the function model established in question 2 is firstly cross-validated to the three processes, and the images of actual values and predicted values are drawn. The $RMSE$ is used to determine the error size, and the RMSE of each cross-validation is calculated. small. Then we study the type of error generation and find out the types of errors that can be controlled: model error and source data error. Then we study the two types of errors separately, and finally give four kinds of control methods of error.

Finally, because the model has a small standard mean square error in cross-validation, the above model can accurately predict the Kelvin temperature and the key element content, showing that the model has good applicability.

Contents

1 Introduction

1.1 Problem Background

Accurate control of temperature is a crucial factor in the metal smelting process. In order to obtain the best target performance of the objective metal, we need to precisely control the flame temperature in the furnace and the content of key elements in raw materials. In case of the same frequency, the greater the light intensity, the better the thermal effect of the flame. Therefore, we can predict the flame temperature and the content of key elements in raw materials in real time by detecting the intensity of the flame in furnace.

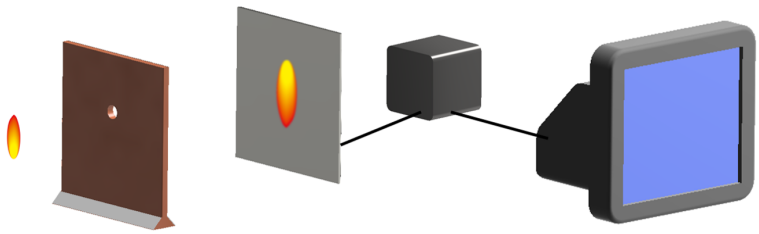

Figure 1: Optical data generation process at a certain time,designed with *Inventor 2017*

1.2 Previous Research

So far, many aspects of metal smelting have been studied. In the past, metal smelting was mainly to study the effects of smelting temperature and key element content on smelting effect.

Literature [1]mainly studies element contentHCA grouped mushrooms in three statistically significant clusters, while PCA indicated connection between analyzed metalsLiterature [2] uses inductively coupled plasma optical emission spectroscopy (ICP OES) to determine the content of nine elements in a substance. Literature [3] reviews thermodynamic data for the recovery of trace valuable elements from primary copper and secondary copper smelting, indicating that comprehensive elemental data is essential for smelting. Literature [4] selects 6 different classification algorithms for classification, compares the accuracy between them, and finally establishes a method that can automatically classify data with high precision.

The above studies have their own advantages, and most of them focus on the analysis of the content of the research elements, and do not take into account the influence of other factors. Therefore, we are concerned with the overall design factors.

With the support of theory and technology, we develop a mathematical model to determine the various influencing factors affecting metal smelting, and analyze the influence weights of various influencing factors.

1.3 Our Work

1. In order to reduce the complexity of model calculation by reducing the amount of data, we found several eigenvalues from 2048 light intensity data.

2. We establish a model to predict the Kelvin temperature T and key element content C and the relationship between them as well.

3. Based on the results of **1** and **2**, we designed a crossover experimental scheme to verify the error produced by model when predicting the key elements. In addition, we provide a feasible error control scheme on the basis of error analysis.

2 Analysis and Key Points

- Analysis of problem 1: If 2048 sets of light intensity data are used as input to make the flame temperature and key element content as outputs, the mathematical model's computational complexity will be huge, so our goal is to find one or more eigenvalues from the light intensity data. Represents most of the information in the original data. Then choose the method of principal component analysis to reduce the dimension of 2048 sets of data, and the original data is replaced by principal components with smaller dimensions and not related to each other. In order to reduce the loss caused by the decrease of dimension, random forest is used to combine with it. Therefore, the appropriate feature values are selected more accurately.

- Analysis of problem 2: In order to be able to predict the Kelvin temperature T and the key element content C using the optical information feature data λ, time t and cumulative consumption Q, we need to select an appropriate prediction model. Therefore, we first use the curve estimation function in SPSS to get the approximate relationship between the target and the corresponding variables. Finally, we establish a multivariate nonlinear regression model, and finally use the optical information, time, and accumulated consumption to predict the Kelvin temperature and key elements. The expression of the content.

- Analysis of problem 3: Based on the conclusion analysis, we design an error control scheme. In our study, the composition of the error may include model errors, source data errors, and random errors. However, the random error is uncontrollable, so it is not considered. So we analyze the model error and the cause of the source error, and use RMSE to determine the error. Finally, we give the error control method.

3 Assumptions and Justification

- Assumption 1: the characteristic value which represent more than 85% of information can be used to replace all the original light intensity data.

- Reason: In this way, the amount of calculation can be reduced, and the operation result can be quickly obtained.

- Assumption 2: The selected principal component is highly representative.

- Reason: This can reduce errors due to dimensionality reduction and improve the applicability of the model.

- Assumption 3: No more influencing factors on Kelvin temperature T and key The prediction of element content C has an impact.

- Reason: It can reduce the unfavorable error caused by other factors and improve the accuracy of the model.

4 Symbols and Definitions

In the section,we use some symbols for constructing the model as follows:

Table 1: Symbols and Definitions

Symbols	Meanings
R^2	Goodness of fit determination coefficient
λ_i	Characteristic value
S_i	Unit characteristic vector
u	Mean
a_k	Variance contributionrate
$\sum_{i=1}^{p} a_i$	Cumulative variance contributionrate

P.s.Other symbols instructions will be given in the text.

5 The Model

Before delving into the specific modeling steps, briefly explain the whole idea and analyze the mind map shown in **Fig.2** :

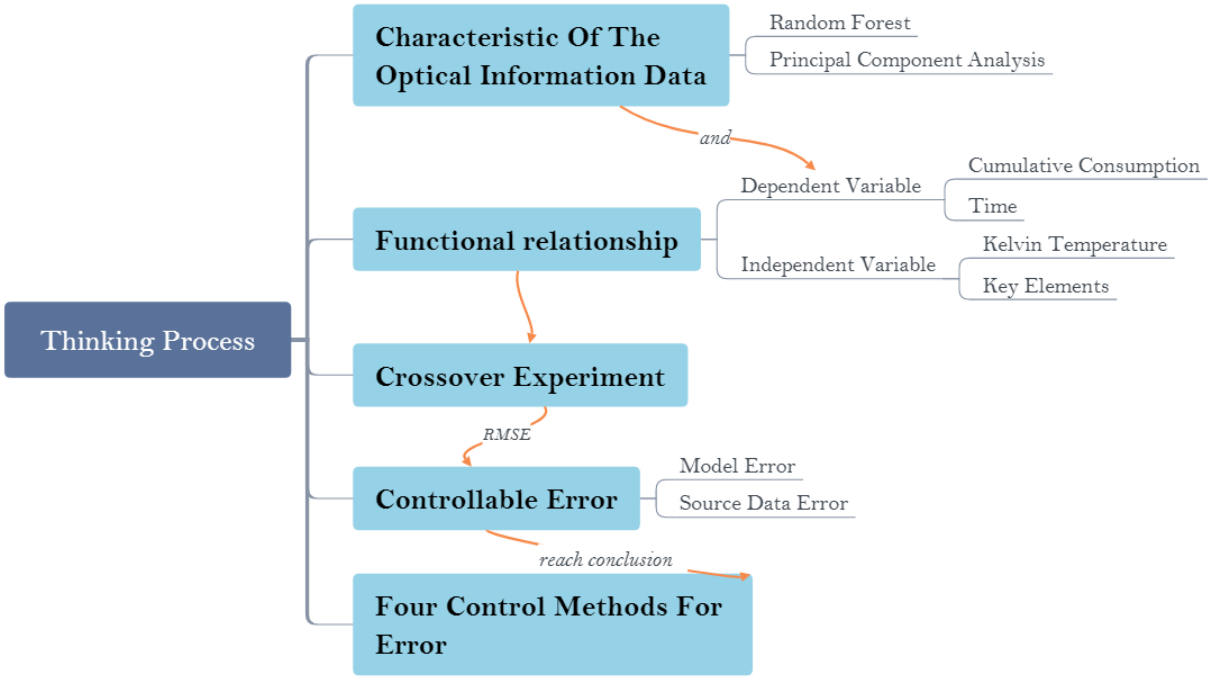

Figure 2: Thinking Process

5.1 Model I:Extraction of optical information data features

5.1.1 Modeling Ideas

We use the random forest feature selection method to extract the features (λ_1 , λ_2 ,$\cdots$, λ_n) from optical information. However, due to the large scale of data, the time complexity O of the algorithm directly using the random forest method to solve the problem can be written as $O(mnlogn)$,where m is the known optical information category, which is 2,048. n is the sample size in annex iii, because the sample size is the smallest relative to the sample size of the other two metals, namely 286. From this expression, we know that it is very complicated to directly solve the problem by random forest feature selection method for the given data.

Therefore, we need to roughly extract some feature quantities first, and then use the random forest feature selection algorithm to extract optical information data features for these feature quantities.

5.1.2 Supplementary Assumptions and Justification

Considering the time complexity of subsequent calculations and in order to simplify the solution difficulty, we extract 30 sets of data for some features in the first stage.

The key assumption that we use feature extraction techniques is that there is no loss of information when features contained in the dataset are removed, i.e., there is no obvious correlation between two or more features. Considering the following situation, when one feature A is strongly correlated with another feature B, the removal of feature A will not affect the contribution and influence of feature B to the result.

5.1.3 Extraction of the Data Feature

Principal component analysis (PCA) classifies some disordered variables according to their correlation, and reduces the variables with high correlation to new, independent and unrelated factors through orthogonal transformation. Each factor is an independent factor, each factor is a principal component, used to represent, explain multiple indicators, forming a new variable. We use this method to reduce the dimensionality of a large number of data, that is, 2048 light intensity values. Since there is no unit difference between the original data, it can be directly used for subsequent analysis. The calculation process is as follows.

We establish a correlation coefficient matrix R for standardized data.

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

where:

$$r_{ij} = cov(x_i, x_j) = \frac{\sum_{k=1}^{n}(x_i - \overline{x_i})(x_j - \overline{x_j})}{n-1}, n > 1 \tag{2}$$

Calculate the eigenvalues λ_{10} , λ_{20} ,$\cdots$, λ_{n0} and eigenvectors a_i can be represented as:

$$a_i = (a_{i1}, ai2, \cdots, a_{in}), (i = 1, 2, \cdots n) \tag{3}$$

and determine the principal component and calculate the corresponding contribution rate C_r:

$$C_r = \frac{\lambda_{i0}}{\sum_{k=1}^{n}(\lambda_{i0})} \tag{4}$$

We standardized the given data, and check the reliability and validity of samples. Then we use Kaiser-Meyer-Olkin (KMO) test to investigate the partial correlation between variables. KMO statistics are evaluated between 0 and 1. When the KMO statistic is greater than 0.7, we conduct dimensional reduction analysis on the variables.

However, when the KMO statistic is less than 0.5, which means that the validity of the sample set is not good so we wont consider it.

The reliability and validity test results of sample data are shown in **Table 2**:

Table 2: KMO and Bartletts Test

Kaiser-Meyer-Olkin Measure of sampling Adequacy	0.915
Bartlett's Test of the value Sig	0.000
Alpha	0.947

As can be seen from **Table X**, the KMO measurement accuracy of the given data is 0.897> 0.5. The result of Bartlett sphere test is 0.0000, statistic significantly, Alpha value of 0.924, indicating that each variable is highly correlated.

The results show that the given sample set is suitable for dimension reduction analysis.

5.1.4 Model Calculation and Result Analysis

1.Algorithm thought

Firstly, we use SPSS (Statistical Product and Service Solutions) software to conduct preliminary screening to obtain part of the data, so as to use the random forest feature selection method to select features in the next step. After that, we carry out the matrix operation and sorting with MATLAB software. Finally, we use Python to get further extraction of the key features.

2.Algorithm steps

Step1: According to the principal component analysis method, analyze the component matrix and variance matrix of the main characteristic quantities by SPSS.

Step2: Import the metal composition matrix and variance analysis matrix into MATLAB.

Step3: To determine the weights of each optical information data on the flame temperature and the key elements in the raw materials, we combine the two matrices obtained in **step 2** and sequence the weights in descending order.

Step4: Select the top30 of weight data as L_n(n=1,2,$\cdots$) in ***step3***. To obtain the characteristic value Y_n(n=1,2$\cdots$) corresponding to optical information data of these 30 rates by comparing with the given data in annex.

Step5: Import Y_n and the corresponding flame temperature and raw material values into Python and use the random forest algorithm to derive the percentage of impact.

3.Model result

According to the influence percentage result calculated by Python, we extract the

first five data with the largest value from the influence percentage of Metal i, Metal ii and Metal iii respectively and we obtained **Fig.3** by the data visualization software Tableau. The characteristic values of optical information data corresponding to each bar graph are shown in **Table 3**.

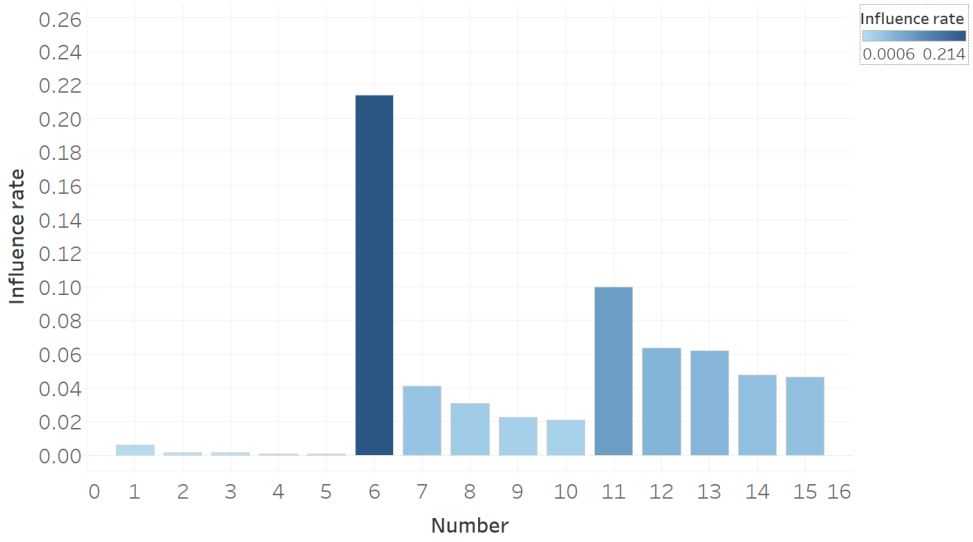

Figure 3: Influence rate of optical information data of three metals

As indicated in **Table 3**, for Metal 2, its optical information data f_212 has the largest impact on flame temperature and key elements in raw materials, accounting for 21.4%. For Metal i, compared with Metal ii and Metal iii, the optical information data has the least influence on flame temperature and key elements in raw materials, and the maximum value is only 0.62%.

5.2 Model II: Establish a relationship model

Through practical research and the conclusions drawn in question one, the paper finds that the reasons affecting the Kelvin temperature T and the key element T content are inseparable from the special composition of time t , gas consumption Q and optical information data.

5.2.1 Modeling Ideas

Using SPSS software, the three influencing factors of time t , gas consumption Q and optical information data extracted in question 1 are used as independent variables, and the relationship between the three independent variables and the dependent variable Kelvin temperature T and the key element C content is respectively carried out. Curve fitting, establishing the curve equation. Then using the method of multivariate nonlinear regression analysis, a prediction model of Kelvin temperature T and key element C content is established.

Table 3: Selection of eigenvalues

Metal	Influence rate	Characteristic values of light intensity data
Metal i	0.0015	f_313
	0.0015	f_313
	0.0015	f_320
	0.0007	f_321
	0.0006	f_305
Metal ii	0.2140	f_212
	0.0015	f_690
	0.0015	f_691
	0.0007	f_306
	0.0006	f_313
Metal iii	0.0998	f_1197
	0.0637	f_1197
	0.0637	f_1017
	0.0620	f_1027
	0.0463	f_1051

5.2.2 Supplementary Assumptions and Justification

In this problem, in order to simplify the solution process, we make the following assumptions: the independent variables used for prediction do not have measurement errors; at the same time, the respective variables do not affect each other.

5.2.3 Multiple Nonlinear Regression Model

1. Curve estimation

The curve estimation theory is mainly applicable to the case of using one variable to predict another variable. The mathematical model that can be used is shown in **Table 4**. When it is not possible to quickly determine an optimal model based on observation, the curve estimation method can be used to select a regression model with the best fitting effect among the regression model by regression model test.

2. Regression model test

(1) Goodness of fit test

In this paper, when the goodness-of-fit test is performed on the fitting result of the selected curve equation, the sample $R^2 \in [0, 1]$ is used as the basis for determining the

Table 4: Table Form of curve estimation model

Model Name	Model expression
Linear function	$y = b_0 + b_1 x$
Logarithmic function	$y = b_0 + b_1 x + b_2 x^2$
Quadratic function	$y = b_0 (b_1)^x$
Cubic function	$y = e^{(b_0 + b_1 x)}$
Composite function	$y = b_0 + b_1 \ln x$
Power function	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
System function	$y = e^{(b_0 + b_1 / x)}$
Growth function	$y = b_0 e^{b_1 x}$
Exponential function	$y = b_0 x^{b1}$
Logistic function	$y = (1/u + b_0 b_1 x)^{-1}$

goodness of fit, which is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\text{SSE}}{SST} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \overline{y})^2} \tag{5}$$

In the formula, *SSR* is the sum of squares of regression, *SSE* is the sum of squares of residuals, and *SST* is the sum of squares of total deviations. In many regression models, the decision coefficient $R^2 \in [0, 1]$ is a comprehensive measure of the degree of fit to the regression model. The larger $R^2 \in [0, 1]$ is, the better the fit of the model is; the smaller $R^2 \in [0, 1]$ is, the worse the fitting effect is. That is, the closer $R^2 \in [0, 1]$ is to 1, the higher the degree of fit.

(2) Significance test of regression equation

In this paper, the *F* test method is used to test the significance of the selected regression equation. The *F* statistic utilizes the ratio of the two data, which is the mean of the sum of the squares of the regression, and the average of the sum of the squares of the residuals:

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{\sum(\hat{y} - \overline{y})^2/k}{\sum(y - y)^2/(n-k-1)} \tag{6}$$

Where: *n* is the number of samples, *k* is the number of independent variables, and *F* is the *F* distribution of the statistic obeying the first degree of freedom *k* and the second degree of freedom *n-k-1*, that is, $F^{\sim}(k, n-k-1)$. It can be seen from the definition of *F* statistic that if the *F* value is large, it can explain the change of the dependent variable caused by the independent variable, which is greater than the influence of the random factor on the dependent variable. Therefore, if the statistic *F* is more significant, the goodness of fit of the regression equation is also higher.

5.2.4 Model Calculation and Result Analysis

According to the multivariate nonlinear regression analysis theory described above, curve estimation is performed on the time t, the cumulative gas consumption amount Q, and the optical information data characteristics and the Kelvin temperature T and the key element C content, respectively.

1. Curve fitting of time and Kelvin temperature

The time t is fitted to the Kelvin temperature T and the results are shown in **Table 5** below:

Table 5: Curve fitting result of f_691 and Kelvin temperature

Equation	R^2	F	df_1	df_1	sig	$Constant$	b_1
Linear function	0.98	19578.4	1	402	0	1766.288	0.923
Logarithmic function	0.752	1943.2	1	402	0	4.23	0.738
Quadratic function	0.998	122201.7	2	401	0	1749.883	1.413
Cubic function	1	278775.3	3	400	0	1754.71	1.123
Composite function	0.977	16985.5	1	402	0	1767.464	1
Power function	0.963	13824.2	1	402	0	1724.34	0.913
System function	0.963	13824.2	1	402	0	1724.34	0.913
Growth function	0.977	16985.5	1	402	0	7.477	0
Exponential function	0.977	16985.5	1	402	0	1767.464	0
Logistic	0.977	16985.5	1	402	0	0.001	1

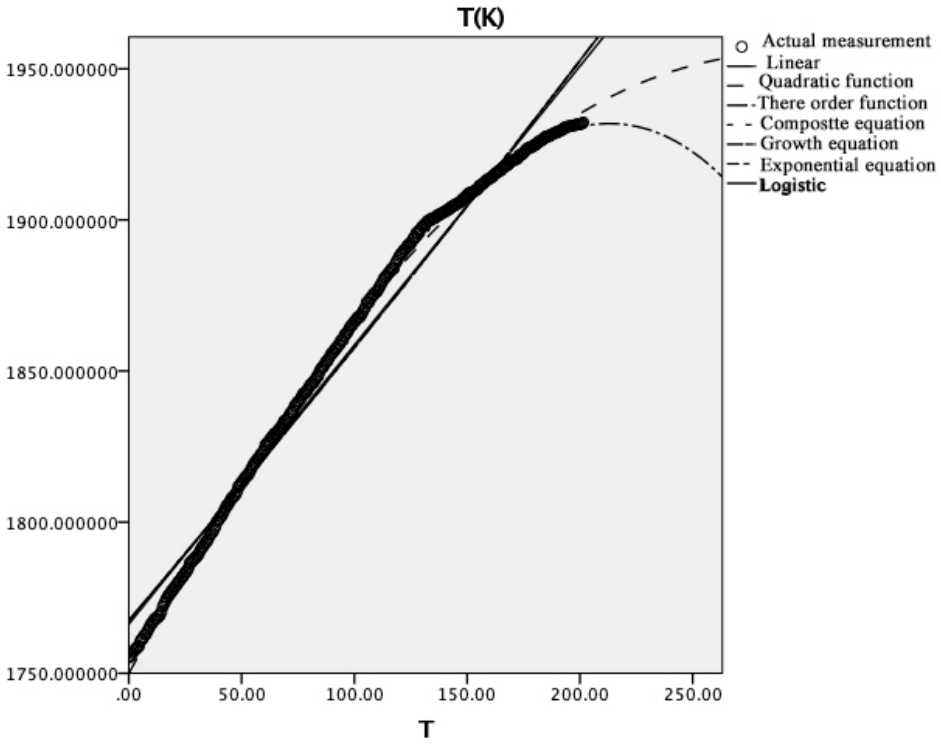

Figure 4: Curve fitting of time and Kelvin temperature

From the results of the fitting, it can be seen that when the time t is fitted to the Kelvin temperature T curve, the $R^2 \in [0,1]$ = 1.000 of the cubic equation, and the significance of the F value is 0. 000, which is less than the significance level of 0.05. It is indicated that the cubic curve fitting is feasible, and the gas consumption Q is selected as the independent variable of the Kelvin temperature T prediction model.

In summary, the curve estimation equation of time t and Kelvin temperature T is a cubic equation, and the expression is:

$$T_t = at + bt^2 + ct^3 + m_1 \tag{7}$$

Where: m_1 is a constant term, and a, b and c are cubic equation coefficients.

2. Curve fitting of acumulated consumption of the combustion-supporting gas and Kelvin temperature

The gas consumption Q is fitted to the Kelvin temperature T, and the results are shown in **Table 6** below:

From the fitting results, we can see that when using the gas cumulative consumption Q and the Kelvin temperature T curve, the exponential function $R^2 \in [0,1]$ is 1.999, and the F value significant probability is 0. 000, less than the significance. Level 0. 05. It is indicated that the exponential function fitting is feasible, and the gas consumption Q is selected as the independent variable of the Kelvin temperature T prediction model.

In summary, the curve estimation equation of the gas consumption Q and the Kelvin

Table 6: Curve fitting result of accumulated consumption of the combustion-supporting gas and Kelvin temperature

Equation	R^2	F	df_1	df_1	sig	$Constant$	b_1
Linear function	0.98	19578.45	1	402	0	1766.288	0.923
Logarithmic function	0.998	199203.1	1	402	0	-2215.14	504.63
Quadratic function	0.998	122201.7	2	401	0	1749.883	1.413
Cubic function	1	278775.3	3	400	0	1754.71	1.123
Composite function	0.988	2721956	1	402	0	1404.352	1
Power function	0.977	374390.2	1	402	0	205.005	0.273
System function	0.994	2721956	1	402	0	7.247	-849.6
Growth function	0.977	2721956	1	402	0	7.477	8.60E-05
Exponential function	0.999	2721956	1	402	0.000	1404.352	8.6E-5
Logistic	0.977	2721956	1	402	0	0.001	1

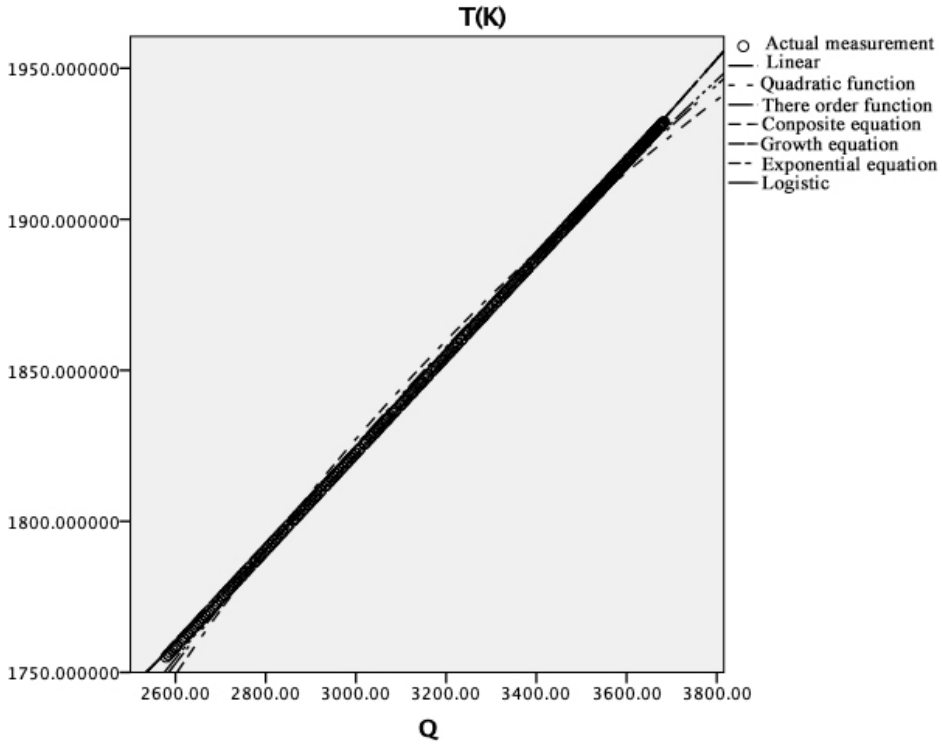

Figure 5: Curve fitting of accumulated consumption of the combustion-supporting gas and Kelvin temperature

temperature T is an exponential function, and the expression is:

$$T_Q = me^{nQ} \tag{8}$$

Where: m, n are exponential function coefficients.

3. Curve fitting of f_691 and Kelvin temperature

In order to simplify the solution process, we consider the optical information data feature quantity solved by the model one and use it as a representative to fit. Using the optical information data feature quantity f_691 of metal one, f_691 is fitted to the Kelvin temperature T, and the results are shown in **Table 7** below:

Table 7: Curve fitting result of f_691 and Kelvin temperature

Equation	R^2	F	df_1	df_1	sig	$Constant$	b_1
Linear function	0.383	249.169	1	402	0	1913.953	-0.017
Logarithmic function	0.526	445.886	1	402	0	-2215.14	-52.31
Quadratic function	0.613	317.297	2	401	0	1749.883	1.413
Cubic function	0.792	224.184	3	400	0	1995.414	-1
Composite function	0.376	241.973	1	402	0	1913.737	1
Power function	0.518	431.256	1	402	0	2313.757	-0.028
System function	0.567	525.338	1	402	0	7.498	54.614
Growth function	0.977	2721956	1	402	0	7.557	-9E-6
Exponential function	0.376	241.973	1	402	0	1913.737	-9E-6
Logistic	0.376	241.973	1	402	0	0.001	1

From the fitting results, it can be seen that when f_691 is fitted to the Kelvin temperature T, the R^2 of the cubic equation is 0.792, and the significance of the F value is 0. 000, which is less than the significance level of 0.05, indicating the cubic equation. The fitting is feasible, and f_691 is selected as the independent variable of the Kelvin temperature T prediction model.

In summary, the curve estimation equation of f_691 and Kelvin temperature T is a cubic equation, and the expression is:

$$\boldsymbol{T_V = eV + fV^2 + gV^3 + m_{2+}} \tag{9}$$

Where: V represents the numerical value of f_691, $\boldsymbol{m_2}$ is a constant term, and e, f, g are cubic equation coefficients.

4. Establishment of Kelvin Temperature T Model

Through the above curve estimation, the expressions of the respective variables relative to the dependent variable are obtained. With the Kelvin temperature T as the

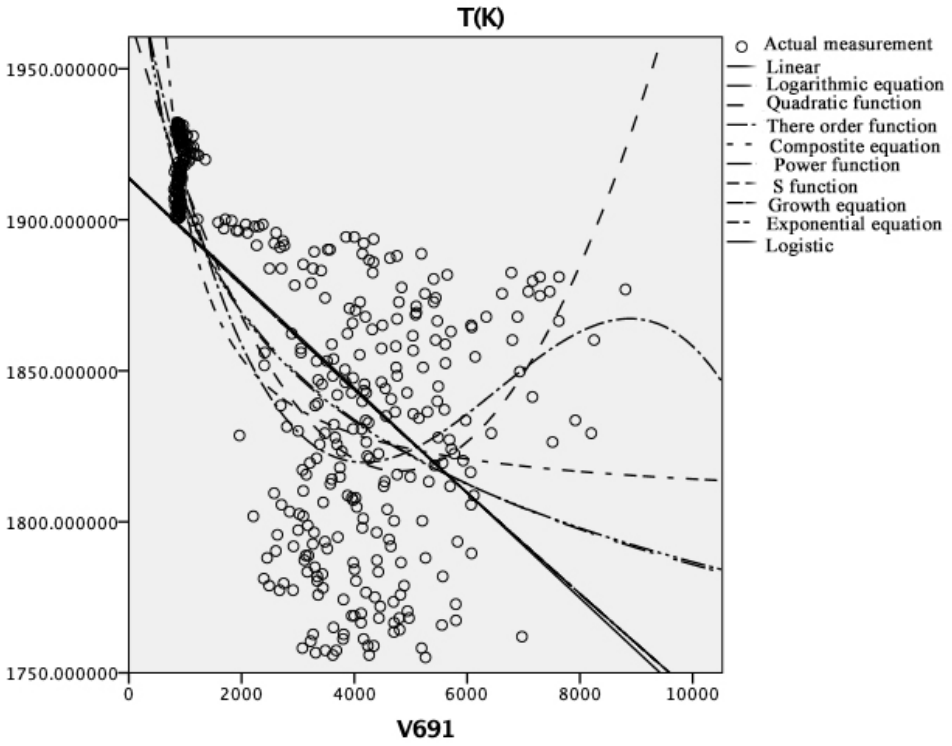

Figure 6: Curve fitting of f_691 Kelvin temperature

dependent variable, the time t, the gas consumption Q and the f_691 as the independent variables, the following equation is constructed:

$$T = a_1 * t + a_2 * t^2 + a_3 * t^3 + a_4 * e^{a_5+Q} + a_6 * V + a_7 * V^2 + a_8 * V^3 + m \tag{10}$$

Where: a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8 are the coefficients of the independent variables in each equation, V represents the numerical value of f_691, and m is a constant term.

Using SPSS nonlinear regression analysis, the Kelvin temperature T is obtained as the coefficient between the dependent variable and each related variable. See **Table 8** and **Table 9**.

Therefore, a nonlinear regression model is constructed with the Kelvin temperature T as the dependent variable, time t, gas consumption Q and f_691 as independent variables:

$$T = 1.101t + 0.001t^2 + 0.003V + 1750.969 \tag{11}$$

Where t represents time, increments every 0.5 seconds, and V represents the numerical value of f_691.

Repeat the above steps to obtain a nonlinear regression model with the key element C content as the dependent variable, time t, gas consumption Q and f_691 as independent variables:

$$c = 0.523t - 0.001t^2 - 0.185Q + 610.449 \tag{12}$$

In summary, the relationship between the Kelvin temperature T and the key element C content, time T, gas consumption Q and V values of the three metals is shown in the **Table 10**:

Table 8: Estimation results of nonlinear regression parameters

Parameter	Estimate	Standard error	Lower limit	Upper limit
a1	1.101	0	1.1	1.101
a2	0.001	0	0.001	0.001
a3	-0.0000127	0	0	0
a4	-10000	0	-10000	-10000
a5	-10000	0	-10000	-10000
a6	0.003	0	0.003	0.003
a7	-0.000000696	0	-0.000000911	-0.000000482
a8	4.62E-11	0.01	-0.021	0.021
a9	1750.969	0.556	1749.876	1752.062

Table 9: Variance analysis table

Source	Sum of square	*df*	Mean square
regress	1.398E+9	9	15530848
Residual	500.269	345	1.267
Total before correction	1.398E+5	404	
Total after correction	1194115	403	
R^2		0.99	

Table 10: Curve fitting result of f_691 and Kelvin temperature

Metal	T	C
Metal i	$T = 1.101t + 0.001t^2 + 0.003V + 1750.969$	$C = 0.523t - 0.001t^2 - 0.185Q + 610.449$
Metal ii	$T = -8.73t + 0.14Q + 0.31V + 1419.526$	$C = 0.432t + 0.1534Q + 0.051V + 547.64$
Metal iiI	$T = 0.612t + 0.1495Q + 0.19V + 1621.147$	$C = 0.167t - 0.1052Q + 0.064V + 527.49$

5. Result analysis

It can be seen from **Table 8** that the nonlinear fitting equation R^2 value is 0.99, and the fitting degree is very high. According to the actual results, the dependent variable and the independent variable result roughly show the above relationship, so the result is reasonable.

5.3 Model III:Cross-Validation and Error Analysis Model

5.3.1 Modeling idea

Based on question one or two, we obtained the experimental results, and then based on the crossover experiments designed in the problem, the predicted and actual values of the key elements of the data in the metal smelting process are shown in the figure.

5.3.2 Cross-validation analysis model

Figure err11 compare in C showed the predicted and actual values of the key element content based on the prediction model of 1 process and the data of 1 process, figure err21 compare in C showed the predicted and actual values of the key element content based on the prediction model of 1 process and the data of 2 process, and so on. Due to the limited space, the error analysis chart for T is given in the appendix.

From the results of the crossover experiment, the cross-simulation curve between the T and C groups produced a certain degree of deviation. More often, the tail of the curve showed a bifurcation comparison in err12.

5.3.3 Modeling Calculation and Result Analysis

With the help of SPSS software, we calculated the NMSE of each unit in the cross-test, which accurately describes the difference between the predicted data and the given data. The results of the analysis are shown in the **table 11**.

$$NMSE = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2} \tag{13}$$

Table 11: Normalized Mean Square Error of C

Model	Process 1	Process 2	Process 3
Model 1	0.0131	8.7133	26.3482
Model 2	9.7245	0.0142	12.2319
Model 3	15.6232	14.3247	0.0101

The error of the model often comes from many aspects, and the error analysis will help to improve the general applicability of the model. The error analysis will be given

in the next section.

5.3.4 Error source analysis

The error of the model can be divided into several aspects: the error generated by the modeling method, the error in the data measurement, and the random error. Since the degree of autonomous control of random error is not as good as the other two, we only analyze the error of the modeling method and the error of the data itself.

1. Principal component analysis error

In the data processing method we use, the error of principal component analysis mainly comes from the choice of the dimensionality of the data reduction and the contribution rate of the cumulative variance. The error caused by the data reduction is caused by the method itself, which is an uncontrollable factor, and the cumulative variance contribution rate is determined by the person's choice.

(1) Dimension reduction error

Assuming that the target is from m samples to n-samples of n samples, this process should ensure that this dimensionality reduction does not result in the loss of important information. In other words, we need to project n samples from m-dimensional space to k-dimensional. For each sample point, we can use the following formula to represent the projection process:

$$Z = A^T X \tag{14}$$

X is an m-dimensional sample point; Z is a K-dimensional sample point obtained after projection; A is an $m * k - dimensional$ matrix. In principal component analysis, we first need to find the average of the samples:

$$u = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{15}$$

$$S = \sum_{i=1}^{n} (X_i - u)(X_i - u)^T \tag{16}$$

$$A = [S_1, S_2 \cdots S_k] \tag{17}$$

Taking a set of n sample point data as an example, when we use PCA to reduce it to 1 dimension, we can express the loss of this set of sample points by the following formula:

$$L = \sum_{l} \left\| X_i - AA^T X_i \right\|^2 \tag{18}$$

The meaning of this formula is the sum of the distances from high-dimensional space to low-dimensional space for each sample point. In order to achieve the purpose of dimensionality reduction, the error is uncontrollable. This part of the error cannot

be eliminated or reduced unless a more reliable dimensionality reduction method is found.

(2) Wrong choice of parameters

The parameter here refers to the cumulative variance contribution rate. In the principal component algorithm. The variance contribution rate of the principal component α_k is defined by:

$$\alpha_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}(i = 1, 2 \cdots p) \tag{19}$$

$$\sum_{i=1}^{m} a_j = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}(m < p) \tag{20}$$

In the actual analysis, the value of m is determined according to the cumulative variance contribution rate. The greater the cumulative variance contribution rate, the stronger the variance of the random vector, interpreted by a small number of selected principal components. However, there are contradictions here. If we choose a larger cumulative variance contribution rate, although it can reduce the loss of data information, it will lead to the main, increased choice, and ultimately lead to the deviation from the target dimension. If the cumulative variance contribution rate is too small, although it will greatly reduce the number of principal components, it will result in insufficient data representation and the meaning of modeling will be lost.

In the principal component analysis model, we use SPSS software to generate a gravel map and a total variance interpretation table. After analysis and comparison, we selected the main components with cumulative variance contribution rate greater than 85%. As we have seen from the gravel diagram, not only is the choice of principal components suitable for quantity, but the loss of data information is also controlled within an acceptable range. Let us take the total variance interpretation table generated by the principal component analysis of the first set of eigenvalues as an example.

2. Data source error

According to the analysis of the source data, the fluctuation of the optical characteristic value is very intense, and there is no rule to follow, but the characteristic value found after the principal component analysis can be linearly distributed with the temperature. Therefore, in the regression analysis factor f_691, its linear characteristics should be considered, and linear regression is the best choice. However, we found that when combined with the results of principal component analysis of optical eigenvalues, the linear regression model does not fit well, and the results in the fitted graph and the goodness of the fit table indicate:

Table 12: Selection of eigenvalues

R	R^2	Adjusted R^2	Estimated standard error
0.619	0.383	0.381	42.823

Ingredient	Total	Initial eigenvalue variance percentage	Accumulation (%)	Total	Percentage of variance	Accumulat (%)
1	1674.063	81.741	81.741	1674.063	81.741	81.741
2	23.336	1.139	82.881	23.336	1.139	82.881
3	14.912	.728	83.609	14.912	.728	83.609
4	4.121	.201	83.810	4.121	.201	83.810
5	3.499	.171	83.981	3.499	.171	83.981
6	3.343	.163	84.144	3.343	.163	84.144
7	3.323	.162	84.307	3.323	.162	84.307
8	3.264	.159	84.466	3.264	.159	84.466
9	3.204	.156	84.622	3.204	.156	84.622
10	3.181	.155	84.778	3.181	.155	84.778
11	3.156	.154	84.932	3.156	.154	84.932
12	3.122	.152	85.084	3.122	.152	85.084
……	……	……	……			
2040	-9.174E-15	-4.480E-16	100.000			
2041	-9.515E-15	-4.646E-16	100.000			
2042	-9.546E-15	-4.661E-16	100.000			
2043	-9.972E-15	-4.869E-16	100.000			
2044	-1.022E-14	-4.989E-16	100.000			
2045	-1.359E-14	-6.635E-16	100.000			
2046	-1.838E-14	-8.974E-16	100.000			

Figure 7: Total variance interpretationpartial selection

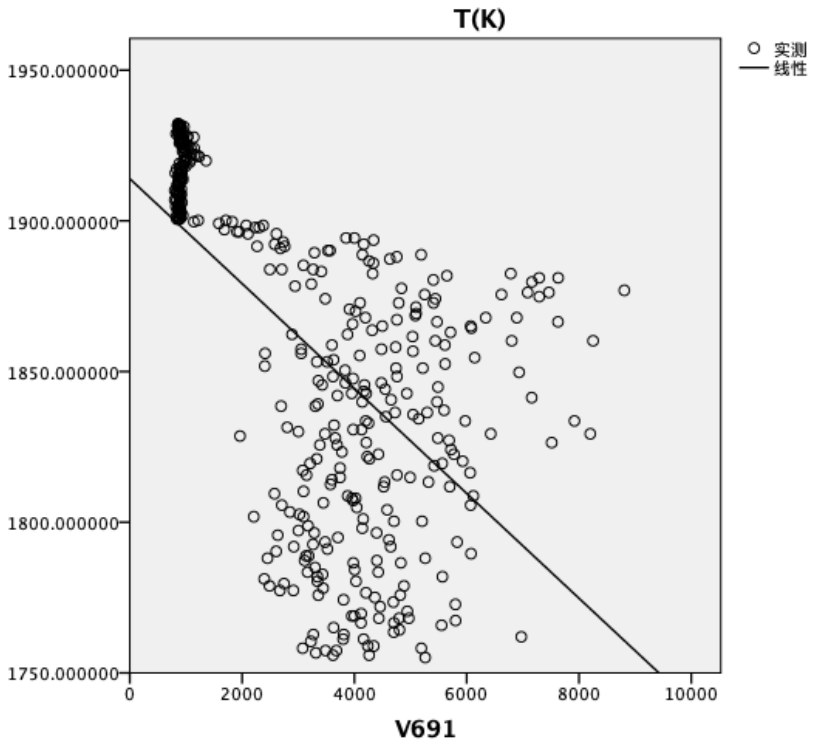

Figure 8: Curve fitting map

It can be seen from the table that the R^2 after linear fitting is only 0.383, indicating that the linear fitting effect is poor, indicating that there is an error in the process of collecting the source data.

3. Error control method

By analyzing the source of the error, errors can be divided into controllable errors and uncontrollable errors. From the source, process and results, reducing the negative impact of controllable errors can improve the versatility of the model. Based on the above analysis, we propose a method of controlling the error:

(1) Improve the metal smelting process technology and improve the measurement and recording accuracy of optical information data.

(2) For the volatility of the initial data, a smooth curve fit should be used. Through the design of reasonable algorithms, the abnormal fluctuation values are filtered to reduce the negative impact on the accuracy of the model..

(3) In the principal component algorithm, choose the larger cumulative variance contribution rate to improve the data loss in the dimension reduction process, so as to control the error reasonably and improve the versatility of the model.

(4) The influence of random errors can be reduced by multiple measurements of the same process.

6 Future Improvements

When applying the principal component method, we should try to select more features to avoid reducing the loss of information. In the case of multivariate nonlinear regression, more data should be used to design more crossover experiments, improve the accuracy of the model, and make the model more universal. In addition, in the process of detecting light intensity data, a more accurate method should be selected to reduce the error caused by the source data.

7 Strengths and Weaknesses

7.1 Strengths

- Combining principal component analysis with random forest can reduce the error caused by dimensionality reduction, and the model is more accurate, which not only improves the operating efficiency, but also improves the applicability of the model.
- Regression analysis can accurately measure the degree of correlation between various factors and the degree of regression fitting, and improve the effect of prediction equations. Multivariate nonlinear regression analysis is more suitable for practical problems and is used when combined with multiple factors. Before the

regression analysis, the curve is used to select the most suitable curve to improve the accuracy of the model prediction.

- Random forests can process higher-dimensional data and identify the relationship between factors, and the algorithm can maintain its accuracy when many features are lost.
- Principal component analysis can reduce the workload of selecting indicators, avoiding the adverse effects caused by excessive indicators in this paper, and can eliminate the mutual influence between the indicators.

7.2 Weaknesses

- Principal component analysis can reduce the workload of selecting indicators, avoiding the adverse effects caused by excessive indicators in this paper, and can eliminate the mutual influence between the indicators.

References

[1] Marija V. Dimitrijevic, Violeta D. Mitic,Jelena S. Cvetkovic,Vesna P. Stankov Jovanovic,Jelena J. Mutic,Snezana D. Nikolic Mandic.Update on element content profiles in eleven wild edible mushrooms from family Boletaceae [J]. European Food Research and Technology,2015(1):1-10.

[2] Emanuela dos Santos Silva,Erik Galvao Parahosda Silva,Danielen dos Santos Silva,Cleber Galvao Novaes, Fabio Alan Carqueija Amorim,Marcio Jose Silva dos Santos,Marcos Almeida Bezerra. Evaluation of macro and micronutrient elements content from soft drinks using principal component analysis and Kohonen self-organizing maps [J].Food Chemistry,2019(273):9-14.

[3] M. A. H. Shuva, M. A. Rhamdhani. Thermodynamics data of valuable elements relevant to e-waste processing through primary and secondary copper production: a review [J]. Metallurgical and Materials Transactions B,2016(1):317-327

[4] Musa Peker,Aye Arslan,Baha en,Fatih V. Çelebi,Abdulkadir But.A Novel Hybrid Method for Determining the Depth of Anesthesia Level: Combining ReliefF Feature Selection and Random Forest Algorithm (ReliefF+RF) [C].Madrid:IEEEPress,2015,pp.27-34.

[5] Duan Mengmeng, Tang Boming, Liu Tangzhi, Hu Yixin. Accident Prediction Model of Freeway with High Ratioof Bridges and Tunnels Based on Multivariate Nonlinear Regression [J]. Highway Engineering,2018(6):122-126.

[6] M. Peker, A. Arslan, B. en, F. V. Çelebi and A. But, "A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF)," 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA), Madrid, 2015, pp. 1-8.

[7] P. Mohana Chelvan and K. Perumal, "A comparative analysis of feature selection stability measures," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 124-128.

[8] S. S. Kumar and T. Shaikh, "Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest," 2017 International Conference on Computer and Applications (ICCA), Doha, 2017, pp. 227-231.