Study on the preparation of C4 olefins by ethanol coupling based on multiple regression analysis

Yizhou Xu [1, *, #], Jianxun Liu [2, #], Rui Zhou [3, #]

[1] Institute of Mechanical and Transport Engineering, Taiyuan University of Technology, Taiyuan, China, 030000

[2] College of Science, Minzu University of China, Beijing, China, 100081

[3] College of Mathematics and Statistics, Lingnan Normal University, Zhanjiang, Guangdong, 524048

* Corresponding author: 996971265@qq.com

[#]These authors contributed equally.

Abstract. With the continuous development of the current chemical industry, the demand for olefins as raw materials for the production of chemical products and pharmaceutical intermediates is increasing. Compared with traditional fossil energy, ethanol is an excellent green raw material for the production of C4 olefins. In the preparation process, the selectivity of C4 olefins and the conversion rate of ethanol depend to a certain extent on the reaction temperature and the catalyst of the reaction, so the reasonable selection of catalyst and temperature becomes very important. In order to study the effects of different catalyst combinations and temperatures on ethanol conversion and C4 olefin selectivity, a multiple regression model was established and its endogeneity was tested by Monte Carlo simulation. The stepwise regression method was used to select the appropriate independent variables, and then the model was optimized to reduce the dimension, and the first optimized linear regression equation was obtained. Finally, in order to test and eliminate the heteroscedasticity of the model, this paper uses the combination of White test and Wls least squares estimation and robust variance error to obtain the second optimized linear regression equation.

Keywords: multiple regression model, stepwise regression, white test.

1. Introduction

C4 olefin is an important chemical raw material, which is widely used and is the basis of petrochemical industry.

Therefore, the catalytic preparation of C4 olefins has been widely concerned.For example, in 2008, Wang, Zhou et al.published a study on the catalytic cracking of mixed C4 hydrocarbons to light olefins over rare earth modified HZSM-5 zeolites[1],Hu Chao's research on catalytic dehydrogenation of mixed C4 alkanes to mixed C4 olefins was published in 2011[2].As a raw material for the production of C4 olefins, ethanol has the advantages of low pollution and wide source. Therefore, the catalytic coupling of ethanol to olefins has entered people 's field of vision and has received extensive attention from the academic community.For example, Cao Jing published a study on the preparation of C4 olefins by ethanol coupling in 2022[3].In the process of preparing C4 olefins, the selectivity of C4 olefins and the yield of C4 olefins are affected by the combination of different catalysts and temperature settings. Therefore, it is of great significance to design catalyst combinations and explore optimal process conditions. There have been some studies on the problem of C4 olefins, for example,a study on the coupling of C4 olefin catalytic cracking and methanol to olefins reaction over SAPO-34 catalyst was published in 2013[4],Gao, Zhao et al. published in 2022 the use of tantalum pentafluorooxide anion column hybrid microporous materials for efficient separation of warm wine C4 olefins[5],etc.

In this field, domestic and foreign scholars are uneven, mainly divided into two parts. The first type is to carry out chemical experiments directly and determine the results. This method is relatively backward compared with today, and the operation is cumbersome and dangerous. The second type optimizes the problem by establishing a model and draws conclusions. For example, Jiang and Wang will use the RBF neural network method to study production parameters in 2022[6]. Bai Qianqian and Wu Yajun et al. used the genetic algorithm method to analyze the process conditions in 2022[7]. Li and Guo used multiple linear regression analysis to process the conversion rate data[8].Preparation of C4 olefins by Coupling Ethanol Based on Fuzzy Analytic Hierarchy Process by Guo and Zou[9].Study on Preparation of C4 olefins by Coupling Ethanol Based on Grey Relational Analysis[10].

This paper aims to investigate the effects of different catalyst combinations and temperatures on ethanol conversion and C4 olefin selectivity. Firstly, we calculated the correlation coefficient and obtained the correlation coefficient between temperature and ethanol conversion and C4 olefin selectivity under different catalysts. Then, we established a multiple linear regression model of different catalyst combinations and temperatures on ethanol conversion and C4 olefin selectivity. And we tested their endogeneity, multicollinearity and heteroscedasticity.

2. Solution of Pearson Correlation Coefficient

2.1. model explanation

Due to these data exist strong normality, we explore the correlation between ethanol conversion, C4 olefin selectivity and temperature under these 21 catalyst combinations by calculating the Pearson correlation coefficient. In formula *I*, *a* represents temperature and *b* represents ethanol conversion and C4 olefin selectivity, respectively.

$$R_{ab}=\frac{l_{ab}}{\sqrt{l_{aa}l_{bb}}}\frac{\sum(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum(a_i-\bar{a})^2\sum(b_i-\bar{b})^2}}=\sqrt{\frac{SSR}{SST}} \tag{1}$$

Perform hypothesis testing at the same time H_0: $\rho = 0$.

$$|R|=\sqrt{\frac{F}{F+(n-2)}}, F\sim F(1,n-2) \tag{2}$$

When $|R|\le R_\alpha$, there is a significant linear relationship between the two variables.

2.2. Model solving

The calculation results are as table 1. The Pearson coefficient of 19 groups in 21 groups of data is greater than 0.9. It can be seen that under the conditions of different catalyst combinations, temperature is strongly correlated with ethanol conversion and C4 olefin conversion data.

Then, drawing Heat Map of A1 and B1 with Matlab as figure 1.

3. Establishment of multiple linear regression model

3.1. Analysis of variance

It can be seen from analysis of the relationship that there is a linear relationship between the effects of various catalyst combinations on the formation of reactants. The effects of different catalyst combinations and temperatures on ethanol conversion and C4 olefin yield were determined by variance analysis.

Table 1. Correlation coefficients of ethanol conversion, C4 olefin selectivity and temperature under different catalyst combinations

Catalysts combination	Ethanol conversion(%)	C4 olefin selectivity(%)	Catalysts combination	Ethanol conversion(%)	C4 olefin selectivity(%)
A1	0.965	0.943	A12	0.963	0.924
A2	0.995	0.892	A13	0.936	0.973
A3	0.883	0.982	A14	0.964	0.928
A4	0.998	0.953	B1	0.993	0.928
A5	0.939	0.935	B2	0.929	0.903
A6	0.672	0.984	B3	0.890	0.986
A7	0.999	0.907	B4	0.899	0.894
A8	0.891	0.977	B5	0.912	0.918
A9	0.989	0.954	B6	0.882	0.94
A10	0.923	0.695	B7	0.954	0.936
A11	0.903	0.968	/	/	/

First, the variance analysis of the variables is performed to obtain the significance index of the homogeneous test as table 2.

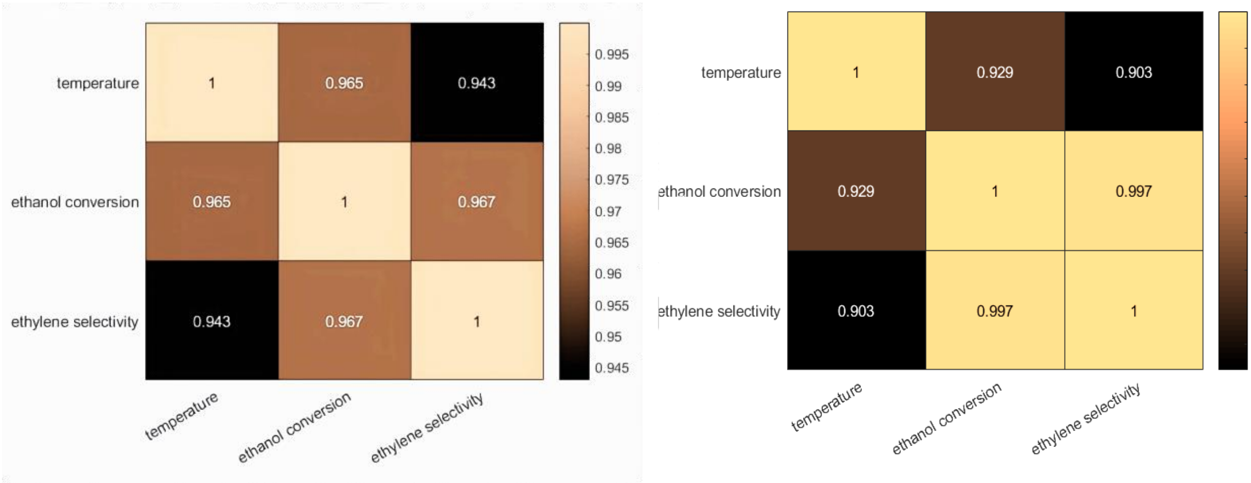

Figure 1. Correlation Coefficient Heat Map of A1 and B1

Analysis of the table shows that in addition to Co load and HAP quality significance of about 0.05, the remaining independent variables are greater than 0.05, it can be considered that the analysis of variance neat. Further bivariate analysis of variance, as shown in the table 3.

On the independent part of the control variables, it can be concluded that the most contribution to the dependent variable, that is, the most influential independent variable is temperature, Co load minimal impact. This can be used to build regression models.

Table 2. The significance index of the homogeneous

Independent variate	Ethanol conversion	C4 olefin selectivity
	significance	
Co/SiO2 quality	0.123	0.123
Co loading	0.048	0.043
ethanol concentration	0.056	0.056
temperature	0.078	0.078
HAP quality	0.043	0.043

3.2. model solving

Suppose the dependent variable is y and k independent variables are $x_1, x_2, \ldots, x_k$ respectively. The equation describing how the dependent variable y depends on the independent variables $x_1, x_2, \ldots, x_k$

and the error term ε is called the multiple regression model. Its general form can be expressed as formula (3).

$$y = C_0 + C_1 x_1 + C_2 x_2 + C_3 x_3 + ... + C_k x_k + \varepsilon \qquad \varepsilon \sim N(0, \alpha^2) \tag{3}$$

Table 3. Inter-subject effect test

source	dependent variable	degree of freedom	Mean square	F	independent variate
modified model	ethanolconversion	98	599.151	48.554	0.000
	C4 olefin selectivity	98	200.461	2.875	0.012
intercept	ethanol conversion	1	39000.888	3160.570	0.000
	C4 olefin selectivity	1	8858.081	127.053	0.000
HAP quality	ethanol conversion	1	138.358	11.212	0.004
	C4 olefin selectivity	1	725.798	10.410	0.006
Ethanol concentration	ethanol conversion	3	858.272	69.553	0.000
	C4 olefin selectivity	3	352.571	5.057	0.013
Co loading	ethanol conversion	3	391.136	31.697	0.000
	C4 olefin selectivity	3	1046.561	15.011	0.000
temperature	ethanol conversion	6	4411.247	357.480	0.000
	C4 olefin selectivity	6	1159.335	16.629	0.000

Where C_0, C_1, C_2,..., C_k are parameters of the model ; ε is the random error term.

Assuming that there is a linear relationship between the dependent variable and the respective variable, the overall linear regression model between them can be expressed as

$$\begin{cases} y_1 = C_0 + C_1 x_{11} + C_2 x_{12} + C_3 x_{13} + C_4 x_{14} + C_5 x_{15} + \varepsilon_1 \\ y_2 = C_0 + C_1 x_{21} + C_2 x_{22} + C_3 x_{23} + C_4 x_{24} + C_5 x_{25} + \varepsilon_2 \end{cases} \tag{4}$$

Where ε_1 and ε_2 are random error terms.

The regression coefficient table is listed as table 4.The radar charts drawn from the regression coefficient are as figure 2.The obtained multiple linear regression model can be expressed as

$$y_1 = -81.519 + 0.077 x_{11} + 1.312 x_{12} - 9.058 x_{13} + 0.333 x_{14} + 0.025 x_{15} \tag{5}$$

$$y_2 = -47.084 - 0.142 x_{21} + 0.181 x_{22} + 1.439 x_{23} + 0.232 x_{24} - 3.373 x_{25} \tag{6}$$

Table 4. Regression coefficient table of ethanol conversion rate and C4 olefin selectivity

variable name	non-normalized coefficients		variable name	non-normalized coefficients	
	B	standard error		B	standard error
constant	-81.519	7.374	constant	-47.084	5.268
x_{11}:Co/SiO2 quality	0.077	0.078	x_{21}:HAP quality	-0.142	0.056
x_{12}:Co loading	1.312	1.277	x_{22}:temperature	0.181	0.014
x_{13}:ethanol concentration	-9.058	2.048	x_{23}:ethanol concentration	1.439	1.463
x_{14}:temperature	0.333	0.019	x_{24}:Co/SiO2 quality	0.232	0.055
x_{15}:HAP quality	0.025	0.078	x_{25}:Co loading	-3.373	0.913

4. Test of three characteristics

4.1. Endogeneity

Endogeneity is the correlation between the independent variable x and the perturbation term ε ,that is $cov(x,\varepsilon) \neq 0$. The existence of endogeneity will lead to regression analysis can not be unbiased estimates, making the conclusion unreliable. If the error term ε is not correlated with all the independent variables x , the regression model is called exogenous.

A B

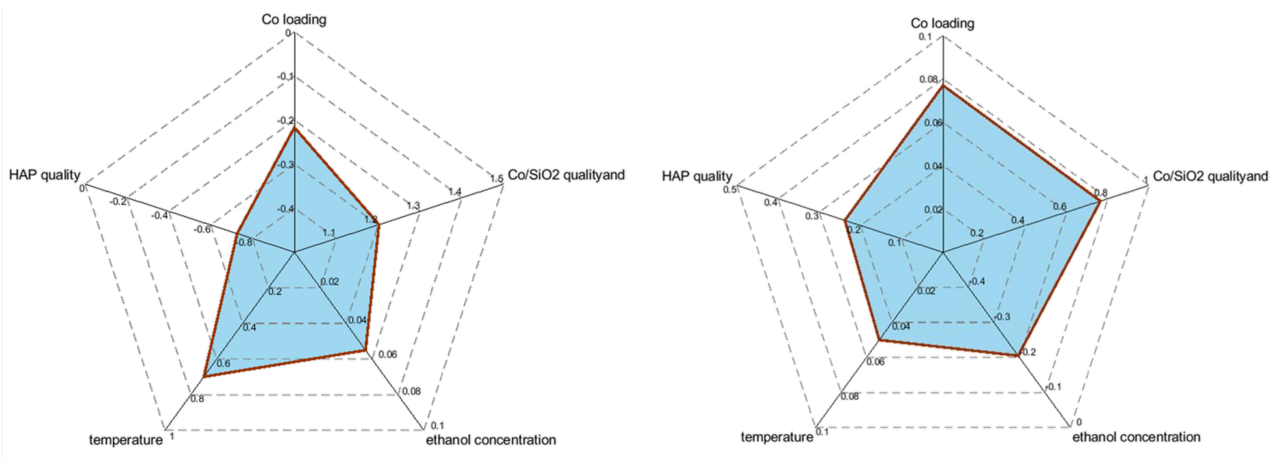

Figure 2. Radar chart drawn from the regression coefficient
(A: ethanol conversion rate, B:C4 olefin selectivity.)

When evaluating the results, if missing the variables which have strong correlation with the results, then the variables will be reflected in the disturbance term, resulting in endogeneity. So we selected five indicators were analyzed to determine whether they have a strong correlation with the results.Therefore, Monte Carlo simulation is used to test the endogeneity of variables.

Taking x_5 as an example, we assume that the variable is missing:

$$y = C'_0 + C'_1 x_1 + C'_2 x_2 + C'_3 x_3 + C'_4 x_4 + \varepsilon' \tag{7}$$

Explore the relationship between the size of the estimated regression coefficient C and the correlation coefficient $\rho_{x1,\varepsilon'}$ as figure 3.

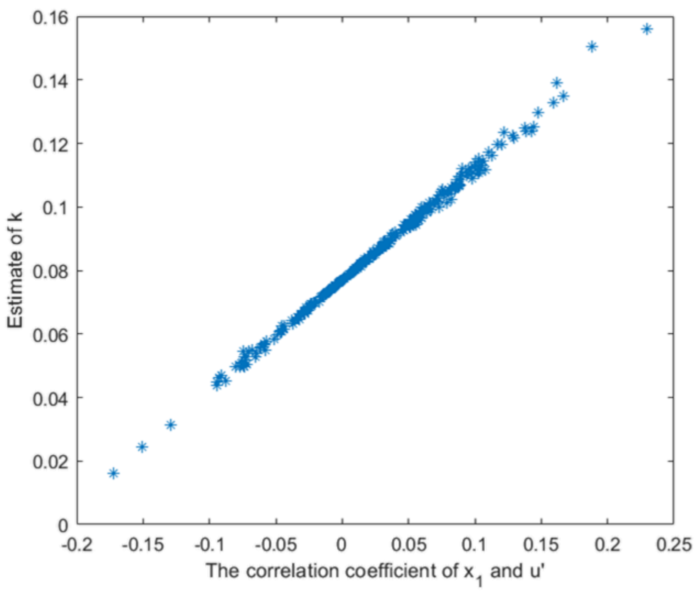

Figure 3. The relationship between regression coefficient and correlation coefficient

When the absolute value of the relative coefficient is larger, the endogeneity of the model will be greater. Comparing the data, we find that when variable x_5 is missing, the model is endogenous. Similarly, verify the correlation between the remaining four variables and the disturbance term.

It can be obtained that the five variables we selected are strongly correlated with the results, and the loss of one of them will lead to the endogeneity of the model, indicating that the variables selected by the model meet the requirements and have a significant impact on the results.

4.2. Multicollinearity and its solutions

Data in table 5 can be obtained from table 4.

Table 5. The significance and VIF value

	constant	temperature	Co/SiO2	Co loading	Ethanol addition	HAP
significance	0.015	0.000	0.637	0.882	0.000	0.044
VIF	1.002	1.002	25.235	1.102	1.159	25.358

According to the VIF value, the VIF of most variables is greater than 2, that is, there is a serious multicollinearity problem. Therefore, most variables are not significant is likely to be caused by multicollinearity, we use a stepwise regression approach to solve the problem of multicollinearity.

After stepwise regression of ethanol conversion rate, the final multiple regression equation is shown in table 6.

Table 6. The final multiple regression equation coefficients of ethanol conversion rate

	constant	temperature	HAP	Ethanol addition
regression coefficient	-80.762	0.333	0.110	-8.680
significance	0.000	0.000	0.000	0.000
VIF	\	1.002	1.100	1.102

Similarly, the final multiple regression equation after stepwise regression of c4 olefin selectivity is shown in table 7.

Table 7. The final multiple regression equation coefficients of c4 olefin selectivity

	constant	temperature	HAP	Co loading
regression coefficient	-44.187	0.180	0.083	-290.762
significance	0.000	0.000	0.000	0.000
VIF	\	1.000	1.1048	1.048

Therefore, the improved stepwise regression model is

$$y_1 = -80.762 - 8.680 x_{13} + 0.333 x_{14} + 0.110 x_{15} \tag{8}$$

$$y_2 = -44.187 + 0.083 x_{21} + 0.180 x_{22} - 290.762 x_{25} \tag{9}$$

4.3. Heteroscedasticity and its solutions

For a multivariate linear regression mode

$$y = C_0 + C_1 x_1 + C_2 x_2 + C_3 x_3 + \ldots + C_k x_k + \varepsilon \tag{10}$$

In general, we default to the homoscedasticity assumption when modeling
$Var(\varepsilon_i | X_{i1}, X_{i2}, \ldots, X_{ik}) = \sigma^2, i = 1,2,\ldots,n$

But for this problem, if for different sample points, the variance of the random interference term is no longer constant, but different from each other, that is, the heteroscedasticity, the accuracy of the model is greatly affected, so we need to test the heteroscedasticity of the residual.

Heteroscedasticity can be roughly divided into three types :

(1) Monotone increasing type: increases with *X*

(2) Monotone decreasing type: decreases with increasing *X*

(3) Complex: Complex form with *X* changes

Here we take C4 selectivity as an example.Using Eviews to draw residual plot as figure 4.

Through Residual square figure and scatter diagram ,we can see that with the increase of x, the y value is more dispersed or denser(Errors caused by individual data not considered), which indicates that the multiple regression model has increasing or decreasing variance, so there is a need for heteroscedasticity test. Here we use the commonly used heteroscedasticity test ,White test.

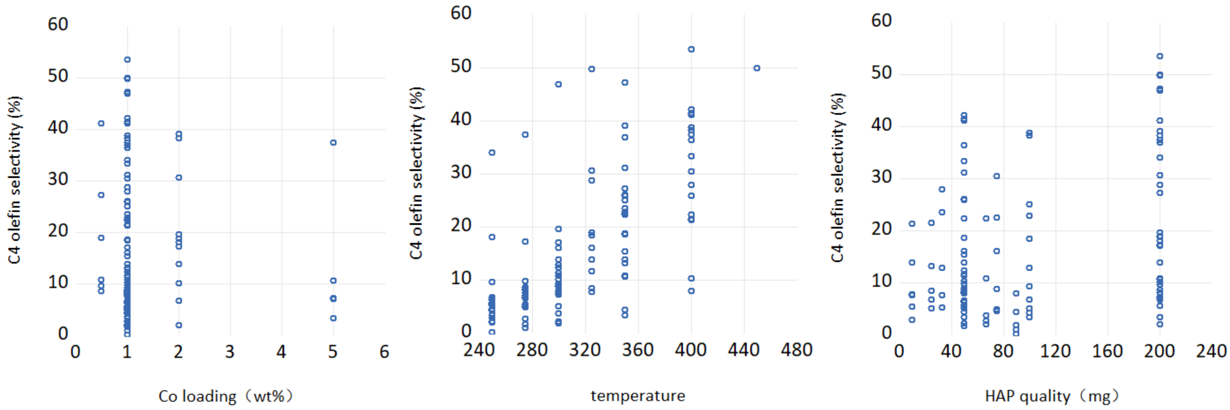

Figure 4. Residual square figure

Here is a test for heteroscedasticity at a 95 % confidence interval and a significance level of 0.05, as shown in the table 8 and table 9.

Table 8. Test for heteroscedasticity of C4 olefin selectivity

chi-square	degree of freedom	significance
48.102	18	0.000

Table 9. Test for heteroscedasticity of ethanol conversion rate

chi-square	degree of freedom	significance
49.310	18	0.000

We see that the significance is less than 0.05.Therefore, we reject the null hypothesis.

Here use Wls least squares estimation and robust variance error to correct heteroscedasticity

Wls least squares estimation,as shown in the table 10,table 11,table 12 and table 13.

Table 10. Wls least squares estimation analysis of ethanol conversion rate

	Constant	Temperature	Ethanol addition	HAP
regression coefficient	-77.459	0.318	-7.536	0.146
VIF	\	1.106	1.111	1.035

Table 11. Wls least squares estimation analysis of C4 olefin selectivity

	Constant	Temperature	Co loading	HAP
regression coefficient	-43.007	0.179	-300.386	0.090
VIF	\	1.098	1.578	1.507

The corrected model is

$$y_1 = -77.459 - 7.536x_{13} + 0.318x_{14} + 0.146x_{15} \tag{11}$$

$$y_2 = -43.007 + 0.090x_{21} + 0.179x_{22} - 300.386x_{25} \tag{12}$$

Table 12. Robust variance error analysis of ethanol conversion rate

	Constant	Temperature	HAP	Ethanol addition
regression coefficient	-80.762	0.333	0.110	8.680
VIF	0.000	0.000	0.000	0.002

Table 13. Robust variance error analysis of C4 olefin selectivity

	Constant	Co loading	HAP	Temperature
regression coefficient	-44.178	-290.762	0.083	0.180
VIF	0.000	0.000	0.000	0.002

The corrected model is

$$y_1 = -80.762 + 8.680x_{13} + 0.333x_{14} + 0.110x_{15} \tag{13}$$

$$y_2 = -44.178 + 0.083x_{21} + 0.180x_{22} - 290.762x_{25} \tag{14}$$

Visible robust variance error correction effect is not very ideal.

5. Conclusions

Through a series of test analysis, it is found that the model we established can solve the related chemical experiment problems, and can also obtain the key variables that affect the results in the experiment and the influence degree of different variables on the results. Multiple regression model in solving this type of problem has a strong feasibility.

From a horizontal perspective, the model not only considers the impact of different variables on the results, but also optimizes the model. It can not only be used to study the effects of different catalyst combinations and temperatures on ethanol conversion and C4 olefin selectivity, but also to study the effects of different influencing factors on the experimental results in similar experiments.

From the vertical point of view, other factors that affect the experiment can also be further analyzed, from multiple perspectives to analyze ethanol conversion and C4 olefin selectivity factors, better analysis of ethanol conversion or C4 olefin selectivity when the best conditions. Not only that, but also can be extended to industrial production, and on this basis, the establishment of planning model, to develop the most suitable for industrial production programs.

References

[1] Wang Xiaoning, Zhou Xinyu, Jiang Guiyuan, et al. Study on catalytic cracking of mixed C4 hydrocarbons to light olefins over rare earth modified HZSM-5 zeolite [J]. Rare earth, 2008, 29 (5).

[2] Hu Chao. Study on catalytic dehydrogenation of mixed C4 alkanes to mixed C4 olefins [D]. Shandong: China University of Petroleum (East China), 2012.

[3] Cao. Ethanol coupling to prepare C4 olefins [J]. Modern salt chemical industry, 2022,49 (5).

[4] Coupling catalytic cracking of C4 olefins with methanol conversion to olefins over SAPO-34 catalyst [J]. Chemical reaction engineering and process, 2013, 29 (2).

[5] Gao Bin, Zhang Zhaoqiang, Hu Jianbo, Cui Jiyu, Chen Liyuan, Cui Xili, Xing Huabin. Efficient separation of C4 olefins using tantalum pentafluor oxide anion-pillared hybrid microporous material [J]. Chinese Journal of Chemical Engineering, 2022, 42.

[6] Jiang Tianhao, Wang Xu, Hu Shizhao. C4 olefin synthesis based on neural network and genetic algorithm [J]. Scientific and technological innovation, 2022 (28).

[7] Bai Qianqian, Wu Yajun, Mu Hai. Study on the preparation of C4 olefins by ethanol coupling based on genetic algorithm [J]. Computer programming skills and maintenance, 2022 (08).

[8] Li Haiqin, Guo Juan. Optimization of experimental parameters for C4 olefin preparation under multiple linear regression model [J]. Yunnan Chemical, 2022, 49 (08).

[9] Haotian Hou,Han Shi. Research on the Optimum Preparation of C4 Olefin by Ethanol Coupling[J]. Academic Journal of Engineering and Technology Science, 2022, 5.0 (1.0).

[10] Li Sanjie, Hu Yulin, Wang Taihang. Research on the preparation of C4 olefins by ethanol coupling based on grey correlation analysis [J]. Science and Technology Information, 2022, 20 (14).