Study on the Impact of Catalyst Factors on Ethanol Conversion and C4 Olefin Selectivity Based on Random Forest Regression

Jiandong Wei *

School of Mathematics and Statistics, Ningbo University, Ningbo, China, 315211

* Corresponding Author Email: 19518322709@163.com

Abstract. Optimizing catalyst combinations for efficient Ethanol Conversion and selective generation of C4 Olefins is important for resource conservation and environmental protection. Specifically, this study started from the temperature data in the catalysts. Firstly, the data were preprocessed to unify the temperature intervals using Hermite interpolation, and then Pearson correlation coefficients were used to further determine the linear relationship between the Ethanol Conversion rate and the C4 Olefin selectivity, respectively, and the temperature. On this basis, a random forest regression model was established to explore the importance of each catalytic factor on ethanol conversion rate and C4 olefin selectivity. Finally, multifactor ANOVA was used in this paper to explore the degree of influence of each catalytic factor on ethanol conversion and C4 olefin selectivity. In this study, it was found that for ethanol conversion rate, the F-value of ethanol concentration was 594.66, which was the largest compared to the other catalyst factors, indicating that it had the greatest influence on ethanol conversion rate.

Keywords: Random Forest Regression, Catalyst Portfolio, Ethanol Conversion, C4 Olefin.

1. Introduction

Ethanol conversion and C4 olefin selectivity are important research directions in petrochemical and catalytic fields. As an important renewable resource, ethanol has a wide range of applications in energy, chemical raw materials, and environmental protection due to its low pollution and renewable characteristics. The optimization of catalyst combinations to achieve efficient ethanol conversion as well as the selective generation of C4 olefins not only helps to save resources but also reduces the impact on the environment. Therefore, an in-depth study of the relationship between ethanol conversion and C4 olefin selectivity is of great practical significance.

In this research field, factors such as temperature and catalyst composition all have significant effects on ethanol conversion and C4 olefin selectivity. However, due to the complex relationship between these factors and the nonlinear characteristics, it is difficult for the traditional linear analysis methods to fully reveal the patterns. To explore the influence of catalyst factors on reaction selectivity more precisely, machine learning-based models have been gradually applied to catalytic research in recent years. In particular, the random forest regression method has been more widely used in the chemical industry. Zheng et al applied this method for gasoline octane number prediction [1]. Ren et al predicted the compositional distribution of thiophene sulfide in decompressed distillate based on the random forest algorithm [2]. Jin et al used this model to predict the structure of the oil layer, and this paper applies the random forest algorithm to the effect of catalyst combinations on C4 olefin preparation from ethanol, which is feasible [3]. Qin et al applied Random Forest to the near-infrared quantitative analysis of gasoline, and Yi et al used the random forest model to establish a risk model for drilling project assessment to play the purpose of risk avoidance, and the model is more effective. All these show that the Random Forest model is widely used in the field of materials and chemicals, so it can also be applied to the research field of this paper [4-5].

In this work, this study firstly interpolates the data to unify the temperature interval, then studies the relationship between ethanol conversion rate, C4 olefin selectivity, and temperature through correlation analysis, and finally establishes the Random Forest model to calculate the characteristic importance of different catalyst factors on the ethanol conversion rate and the C4 olefin selectivity, and based on which conducts the multifactorial analysis of variance (ANOVA) and the multiple comparisons test to deeply explore the influence of the catalyst factors in the preparation process. The effect of catalyst factors in the preparation process was investigated in depth by multi-factor ANOVA and multiple comparison tests.

2. Data Preprocessing and Correlation Analysis

2.1. Data Preprocessing

The data in this article comes from https://www.mcm.edu.cn.

Data preprocessing is a series of processing operations on raw data aimed at improving data quality, model performance, and saving computational resources. Matlab software was first used to perform the necessary preprocessing for the ethanol conversion, C4 olefin selectivity, and temperature of the 21 catalyst combinations provided in the data.

The individual data with temperatures outside this range were excluded by Hermite interpolation between 250 degrees Celsius and 400 degrees Celsius for each combination of experiments with a uniform temperature interval of 25 degrees Celsius. The Hermite interpolating polynomials were as follows.

$$H(x) = \sum_{i=0}^{n} y_i \, \alpha_i(x) + \sum_{i=0}^{n} y_i{}' \, \beta_i(x) \tag{1}$$

y_i and $y_i{}'$ are at the x_i node function values and derivative values, respectively.

$\alpha_i(x)$ and $\beta_i(x)$ are basis functions, respectively, which are composed of variants of the Lagrange interpolating polynomials used to ensure that the interpolating polynomials satisfy the conditions for function values and derivative values at the nodes x_i.

2.2. Correlation Analysis

This question requires the study of the relationship between ethanol conversion rate, C4 olefin selectivity, and temperature, respectively. To make a preliminary judgment and analysis of the relationship between them, this study will first categorize the data according to their characteristics and obtain labeled integrated data. Then, Pearson's correlation coefficient was introduced to visualize and intuit the relationship between the target data at a certain level.

To avoid the complexity of the data from interfering with the analysis, this paper chose Pearson's correlation coefficient, which is more adaptable and precise, and analyzed the simple linear relationship between ethanol conversion, C4 olefin selectivity, and temperature for each group according to the formula described below [6].

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{2}$$

Where r represents the correlation coefficient, n is the number of samples, and X_i and Y_i represent the two sets of attribute values of the i sample, respectively. When $r = 1$ says complete correlation, at this time X_i has a linear relationship with Y_i. Using Matlab software, the results of its correlation are obtained, as shown in Table 1.

Table 1. Correlation coefficient with temperature

Catalyst portfolio	Ethanol Conversion with Temperature	C4 Olefin Selectivity with Temperature
A1	0.965	0.887
A2	0.995	0.914
A3	0.982	0.952
A4	0.997	0.965
A5	0.939	0.972
A6	0.986	0.855
A7	0.999	0.965
A8	0.975	0.992
A9	0.907	0.997
A10	0.910	0.833
A11	0.885	0.987
A12	0.958	0.982
A13	0.927	0.988
A14	0.959	0.954
B1	0.957	0.985
B2	0.917	0.984
B3	0.895	0.959
B4	0.900	0.913
B5	0.919	0.981
B6	0.950	0.985
B7	0.946	0.995

As a whole, it was found that the correlation coefficients of ethanol conversion, C4 olefin selectivity, and temperature for each catalyst combination were all greater than 0.85, which showed a highly positive correlation. The calculation of the correlation coefficients provides a preliminary judgment basis and basic support for the following analysis of the relationship between the target data.

3. Analysis of Catalyst Factors on Ethanol Conversion and C4 Olefin Selectivity

3.1. Characteristic Importance Analysis

Based on the above data preprocessing, then the features are screened to get the important influence features. Because of the large number of data features in this question, there is a certain degree of difficulty in the interpretation, this study chooses the random forest regression model with high accuracy for feature screening [7].

The importance of each catalytic factor on ethanol conversion and C4 olefin selectivity was assessed according to the Random Forest algorithm, as shown in Figure 1.

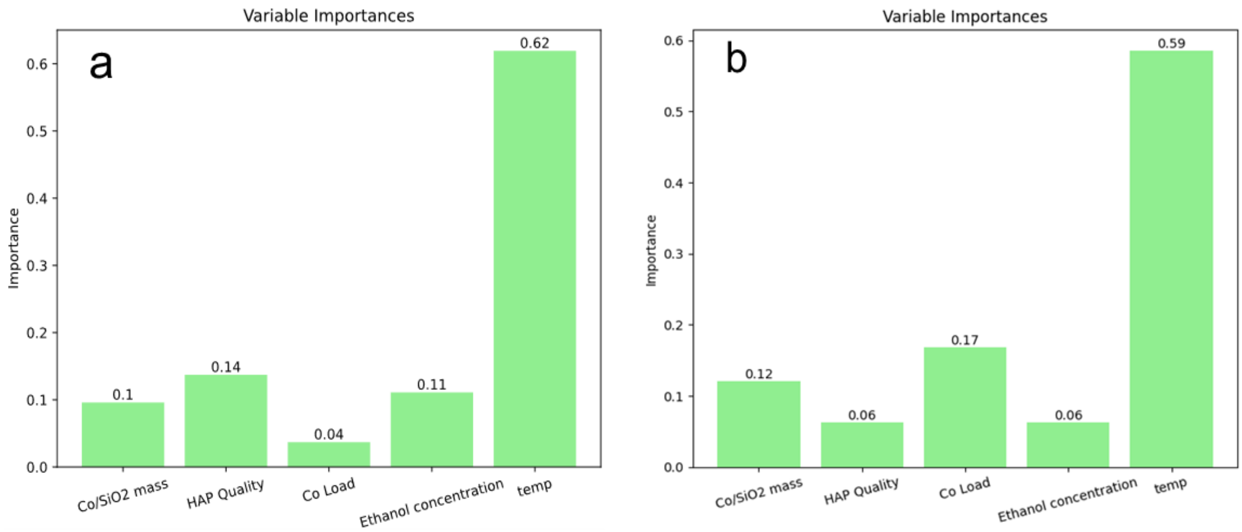

Figure 1. Quantification of influencing factors. (a) Effect on ethanol conversion. (b) Effect on C4 olefin selectivity.

Combining the information shown in Fig. 1 and using the absolute value of importance as a basis for comparison, it can tentatively conclude that temperature is the most important catalytic factor for both ethanol conversion and C4 olefin selectivity. Combined with the previously analyzed relationships between ethanol conversion, C4 olefin selectivity, and temperature, this study further confirms that an increase in temperature has the strongest positive effect on ethanol conversion and C4 olefin selectivity, with the highest correlation. On the other hand, among all the catalytic factors on the importance of ethanol conversion rate, Co loading had a relatively small effect on the ethanol conversion rate, and the effects of the other three catalysts except temperature were roughly equal. Among the importance of C4 olefin selectivity, the characteristic importance of HAP mass and ethanol concentration is relatively small, while the importance of Co loading on C4 olefin selectivity is relatively high. Overall, different catalyst factors have an impact on ethanol conversion as well as C4 olefin selectivity, and these effects should be included in further considerations afterward.

3.2. Analysis of variance

3.2.1. variance chi-square test

Combined with the previously obtained conclusions, this paper determined that the effect of each catalyst combination on the final reflected product constitutes some kind of non-linear relationship. All catalytic factors are now analyzed to examine the magnitude of their effects on ethanol conversion and C4 olefin selectivity.

The variables were analyzed by ANOVA using SPSS software and p-value indicators for the chi-square test of variance were obtained as shown in Tables 2 and 3[8].

Table 2. Ethanol conversion rate test table

independent variable	p-value
Co/SiO2 mass	0.984
HAP Quality	0.984
Co load	0.905
Ethanol concentration	0.131
Temp.	0.972

Table 3. C4 olefin selectivity rate test table

independent variable	p-value
Co/SiO2 mass	0.644
HAP Quality	0.644
Co load	0.898
Ethanol concentration	0.664
Temp.	0.940

Analyzing Tables 2 and 3, it was learned that the p-values of the variances were all greater than 0.05, and the null hypothesis that the variances of all comparison groups are not equal and equal was not rejected, so the data can be considered to have passed the test of chi-square.

3.2.2. multifactor analysis of variance

With the results of the ANOVA chi-square test, this study then conducted a multifactorial ANOVA for further judgment to investigate the extent of the influence of each catalytic factor on the ethanol conversion rate and C4 olefin selectivity [9].

For the multifactor ANOVA, this paper mainly considered the following indicators:

(1) Interaction sums of squares: indicates the variation caused by the interaction between two factors.

$$SSAB = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(a_{ij} - \bar{a}_i - b_j + \bar{b})^2 \tag{3}$$

(2) Interaction mean square: represents the interaction sum of squares divided by the interaction degrees of freedom.

$$MSAB = \frac{SSAB}{dfAB} \tag{4}$$

(3) Interaction F statistic: interaction mean square divided by the within-group mean square.

$$FAB = \frac{MSAB}{MSW} \tag{5}$$

Based on the above indicators, using SPSS software, a multifactor ANOVA table was obtained and the results of the analysis are shown in Tables 4 and 5.

Table 4. Independent Effects Components of Ethanol Conversion Rate Between-Subjects Effects Tests

Source	Class III sum of squares	Degrees of Freedom	Mean Square	F-value	Significance
Modified Model	69.731[a]	125	.558	294.728	.000
intercept	35.663	1	35.663	18841.783	.000
Co/SiO2 mass	.000	0	.	.	.
HAP Quality	.000	0	.	.	.
Co load	.418	3	.139	73.647	.000
Ethanol concentration	3.377	3	1.126	594.660	.000
Temp.	23.172	6	3.862	2040.422	.000

Table 5. Independent Effects Components of C4 olefin Selectivity Between-Subjects Effects Tests

Source	Class III sum of squares	Degrees of Freedom	Mean Square	F-value	Significance
Modified Model	27.006[a]	124	.218	7.613	.000
intercept	64.709	1	64.709	2261.994	.000
Co/SiO2 mass	.000	0	.	.	.
HAP Quality	.000	0	.	.	.
Co load	4.493	3	1.498	52.354	.000
Ethanol concentration	.716	3	.239	8.341	.002
Temp.	10.458	6	1.743	60.929	.000

Based on the independent action part of multiple factors in Tables 4 and 5, the F-value and significance of five catalytic factors, namely, Co/SiO2 mass, HAP mass, Co loading, ethanol concentration, and temperature, were mainly examined. In this study, it was found that the magnitude of the influence of each factor on the dependent variable was measured by the F size, and it was concluded that the degree of their influence on the ethanol selectivity could be arranged in the order of temperature, ethanol concentration, Co loading, Co/SiO2 mass, and HAP mass in descending order, e.g., the temperature as a factor had the highest degree of influence on the conversion rate of ethanol; in the same way, the degree of influence of the five catalytic factors on the selectivity of the C4 olefins could be arranged in the order of temperature, Co loading, ethanol concentration, HAP mass, and Co/SiO2 mass in descending order.

3.2.3. Multiple Comparison Test

Next, differentiated from the comparative analysis among the factors, the analysis is carried out for different value levels in each factor. In this section, based on the fundamental purpose of efficient production of C4 olefin, the research object is targeted to C4 olefin selectivity. According to the previous analysis of the influence of each factor, temperature, Co loading, ethanol concentration has a greater influence on the C4 olefin selectivity, this paper selectively multiple comparison tests on the C4 olefin selectivity with the temperature, ethanol concentration amount of the factors, respectively, and get the test results shown in Table 6 and Table 7 [10].

Table 6. Multiple Comparison Analysis of Table-C4 Olefin Selectivity Against Temperature Factors (partial)

(I) x5	(J) x5	Difference in Mean Values (I-J)	Standard Error	Standard Error	95% confidence interval	
					lower limit	Upper limit
250.00	275.00	-.1928*	.05418	.003	-.3090	-.0766
	300.00	-.3707*	.05418	.000	-.4869	-.2545
	325.00	-.5558*	.05418	.000	-.6720	-.4395
	350.00	-.7251*	.05418	.000	-.8413	-.6089
	375.00	-.8831*	.05418	.000	-.9993	-.7669
	400.00	-1.0129*	.05418	.000	-1.1292	-.8967
. . .	. . .	. . .	. . .	. . .	. . .	. . .
400.00	250.00	1.0129*	.05418	.000	.8967	1.1292
	275.00	.8201*	.05349	.000	.7054	.9348
	300.00	.6422*	.05349	.000	.5275	.7570
	325.00	.4572*	.05349	.000	.3425	.5719
	350.00	.2879*	.05349	.000	.1731	.4026
	375.00	.1299*	.05349	.000	.0151	.2446

Table 7. Multiple Comparison Analysis of Table-C4 Olefin Selectivity Against Ethanol Concentration Factors (partial)

(I) x4	(J) x4	Difference in Mean Values (I-J)	Standard Error	Standard Error	95% confidence interval	
					lower limit	Upper limit
.30	.90	-.1922*	.05836	.005	-.3173	-.0670
	1.68	-.0771	.04887	.137	-.1819	.0277
	2.10	.1706*	.05836	.011	.0455	.2958
.90	.30	.1922*	.05836	.005	.0670	.3173
	1.68	.1150*	.04131	.015	.0264	.2037
	2.10	.3628*	.05220	.000	.2509	.4748
1.68	.30	.0771	.04887	.137	-.0277	.1819
	.90	-.1150*	.04131	.015	-.2037	-.0264
	2.10	.2478*	.04131	.000	.1592	.3364
2.10	.30	-.1706*	.05836	.011	-.2958	-.0455
	.90	-.3628*	.05220	.000	-.4748	-.2509
	1.68	-.2478*	.04131	.000	-.3364	-.1592

Based on the data obtained from the above table, the following analysis was done. Firstly, the difference in the mean values and the corresponding significance at a total of seven temperature levels from 250 degrees Celsius to 400 degrees Celsius were compared. The comparison reveals that at 400 degrees Celsius, the mean values are greater than the other six groups of data and the p-values are all 0, indicating that the C4 olefin selectivity at 400 degrees Celsius is generally higher than that of the other groups and is in a more favorable condition. The differences in means at a total of four ethanol concentrations from 0.30 ml/min to 2.10 ml/min were then compared along with the corresponding significance p. The comparison found that the difference in means at an ethanol concentration of 2.10 ml/min was less than 0 compared to the other levels, suggesting that the C4 olefin selectivity at an ethanol concentration of 2.10 ml/min was lower than the other concentration levels.

4. Conclusion

In this study, this paper first preprocessed the data, applied Elmit interpolation to unify the temperature intervals, and calculated Pearson correlation coefficients, after analyzing the data, this study found that both the ethanol conversion rate and the C4 olefin selectivity had a strong correlation with temperature. Then, a random forest model was used to evaluate the feature importance to obtain the feature importance histogram, and the temperature was found to have a significant effect on the ethanol conversion rate and C4 olefin selectivity visually. After that, a multifactor ANOVA model was further developed, and the ANOVA showed that the five catalytic factors had significant effects on the ethanol conversion rate and C4 olefin selectivity, and the temperature had the greatest effect. Finally, the multiple comparison tests of C4 olefin selectivity against temperature and ethanol concentration showed that the C4 olefin selectivity was generally higher at 400 degrees Celsius than at other temperature levels, and the C4 olefin selectivity was lower at an ethanol concentration of 2.10 ml/min than other concentration levels.

This paper provides a research idea and framework applied to materials science-related fields, using the random forest algorithm to determine the effect of catalyst factors on the conversion rate of ethanol and the selectivity of C4 olefins and obtaining more effective conclusions, which proves the feasibility of the method. While this study demonstrates the effectiveness of using random forest regression in analyzing catalyst factors, future research could explore more complex models incorporating interaction terms to assess the synergistic effects of different factors. Additionally, investigating alternative catalyst systems and optimizing reaction parameters like pressure and reaction time could further enhance the efficiency and selectivity of ethanol conversion.

References

[1] Zheng Bin, SUN Hongxia, WANG Weimin. Octane prediction of gasoline research method based on random forest regression [J]. Petroleum Refining and Chemical Industry, 2020, 51 (12): 69 - 75.

[2] REN Xiaotian, CHU Xiaoli, TIAN Songbai. Prediction of thiophene sulfide composition distribution in decompression distillate based on random forest algorithm [J]. Journal of Petroleum (Petroleum Processing), 2020, 36 (05): 995 - 1002.

[3] JIN Li-Ping, DENG Jin-Gen. Oil formation fouling prediction method based on differential evolutionary optimization random forest model[J]. Energy and Chemical Industry, 2022, 43 (06): 28 - 32.

[4] QIN Qiang, CHEN Huazhou, LI Linghui, et al. Multidimensional scaling and random forest applied to near-infrared quantitative analysis ofgasoline [J]. Science Technology and Engineering, 2014, 14 (36): 137 - 140.

[5] Yi Siqi, Wei Kai. Random forest stuck drill risk assessment method based on SMOTE under sampling [J]. Petroleum Geology and Engineering,2023, 37 (04): 100 - 103.

[6] Liu Lirui, Li Peiqiang, Chen Jiayu, et al. Hybrid energy storage capacity allocation based on Pearson correlation analysis and T-tFD under VMD decomposition [J]. China Electric Power, 2024, 57 (07): 82 - 97.

[7] Sun Shaocong, Xu Yang, Cao Bin. Aluminum output prediction from BiLSTM electrolyzer based on random forest feature selection[J]. Light Metals, 2023, (10): 30 - 36.

[8] HU Chunyan, HU Liangping. How to use ANOVA correctly--covariance chi-square test and SAS implementation [J]. Sichuan Mental Health, 2022, 35 (01): 11 - 15.

[9] Wang R. Research on work order scheduling based on multifactor ANOVA method [J]. Telecommunications Engineering Technology and Standardization, 2023, 36 (08): 65 - 68.

[10] Qiu Nianwei, Liu Qian, Liu Hui. Multiple comparison method in statistical analysis of experimental data in biology [J]. Journal of Biomathematics, 2015, 30 (03): 535 - 541.