ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

Research on the copyrolysis of biomass and coal based on statistical test and machine learning

Yuanji Ren[#,*], Xu Sun[#]

School of Automation, Shenyang Aerospace University, Shenyang, 110136, China
[*]Corresponding author: renyuanji@stu.sau.edu.cn
[#]These authors contributed equally.

Abstract: *At present, the problem of environmental pollution is still serious. At the same time, in the context of increasing global demand for energy, finding a sustainable energy has become the common goal of the world. As a renewable and low-carbon energy carrier, biomass has a high energy conversion rate and a green environment when coheating with coal. In this paper, the influence of INS on pyrolysis changed. In the pyrolysis experiment, there is an interactive effect between INS and mixing ratio on pyrolysis product yield. In this paper, a linear regression model is constructed to evaluate the horizontal effect of different INS and mixing ratio on the yield of the three main pyrolysis products: Tar, Water and Char. Analysis of the three pyrolysis products of each combination by paired samples t-test indicates a significant difference between theoretical and experimental values at a specific mixing ratio. This study reveals the bias between them by comparing experimental values with theoretical predicted values.*

Keywords: *Biomass, coal, co-pyrolysis, pyrolysis products*

1. Introduction

Coal plays an irreplaceable role in heating, thermal power generation and metal smelting. In 2020, coal consumption accounted for 56.8% of China's total energy consumption, making it the main energy source in China. Coal thermal processing has a history of thousands of years, and it is still the main way of coal utilization [1]. The process of coal producing oil and gas resources by isolating air is called coal pyrolysis technology [2]. China is rich in biomass resources, its thermochemical transformation and utilization can not only improve the raw materials is not easy to transport, low calorific value, complex composition, but also can obtain chemical raw materials such as biological oil, biochar and combustible gas, the obtained products have the potential of alternative energy, with the dual economic and environmental benefits [3]. As one of the most important and most potential renewable energy sources in the world, biomass has abundant reserves and has the characteristics of hydrogen-carbon ratio and oxygen-carbon ratio. Copyrolysis with coal can effectively improve the coal pyrolysis conversion rate and tar quality [4-5]. This paper details whether there is an interactive effect between the mixing ratio of INS and INS and coal and biomass, and evaluates the effect of different INS and mixing ratios on the yield of the three main pyrolysis products, Tar, Water and Char.

2. Model building

2.1 Correlation analysis

Correlation analysis is a statistical method that is often used to study the strength of the relationship between two or more variables. Simultaneous correlation analysis aims to quantify the degree of association between two or more variables. This association can be either positive or negative or uncorrelated. In mathematical modeling, this correlation analysis approach can be used to help understand the interaction relationships between data. Statistics and analysis of correlation is a commonly used method in economics. Correlation is when there is a link between two factors, a typical showing that one variable changes with the other. The correlation will be divided into a positive correlation and negative correlation.

The Pearson correlation coefficient is used to measure the linear correlation between positive INS and Tar, Water and Char.

ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

Calculation formula of the correlation coefficient:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} \tag{1}$$

where X is INS; Y1 is Tar; Y2 is Water; Y3 is Char

The correlation coefficient is a value between -1 and 1. The closer the correlation coefficient is to 1, the stronger the positive correlation, the closer the correlation coefficient is to -1, the stronger the negative correlation, and the closer the correlation coefficient is to 0, the less there is a linear correlation. Therefore, the magnitude of the correlation coefficient can be used to determine the degree of influence of hexane insoluble on the three yields.

2.2 Scatter plot model

A scatter map (Scatter Plot), also known as a scatter map table, is a graphical tool used to show the relationship between two variables. In the scatter plot, each dot represents an observed value with the abscissa and ordinate corresponding to the values of the two variables. Such graphs are widely used in statistical analysis, data mining, and machine learning, and are especially suitable for exploring data relevance and trends. Scatter plots can show not only linear relationships, but also non-linear relationships or chaotic correlations. In some cases, the overall trend is more clearly visible by adding a fitting line (e.g. a linear regression line). Moreover, more dimensions can be displayed in the same scatterplot by means of color or shape changes.

Preprocessing of the experimental data (Question B of the 9th "Digital dimension Cup" College Students Mathematical Modeling Competition). As shown in Table 1

Table 1: Preprocessing of the experimental data

Time	sample	ratio	weight	Char	Water	INS	Tar	Water	Char	HEX
20131206	HN	100	10.5737	1.3179	0.58	0.0000	0.1246	0.0579	0.7599	0.0000
20131206	HN	100	10.2179	1.2832	0.59	0.4566	0.1256	0.0579	0.7587	0.0809
20140312	SM	100	10.3176	1.0082	0.94	0.0000	0.0977	0.0914	0.7365	0.0000
20140312	SM	100	10.1371	1.0754	0.93	0.0000	0.1061	0.0914	0.7281	0.0000
20140312	SM	100	9.2990	0.9803	0.85	0.0000	0.1054	0.0914	0.7269	0.0000
20140312	SM	100	8.3511	0.8995	0.76	0.2990	0.1077	0.0914	0.7254	0.0000
20140105	NM	100	10.1726	0.4244	1.70	0.0000	0.4244	0.1671	0.6244	0.0719
20140105	NM	100	10.1743	0.3613	1.70	0.0249	0.0355	0.1671	0.6288	0.0331
20140105	NM	100	10.1018	0.3670	1.69	0.0000	0.3670	0.1671	0.6270	0.0000
20151015	HS	100	8.7366	0.7328	0.79	0.0000	0.0839	0.0904	0.7348	0.0000
20151015	HS	100	8.8229	0.7379	0.80	0.0775	0.0836	0.0904	0.7353	0.0749

2.3 Linear Regression Model

Linear regression model is a statistical method widely used in data analysis. It is used to study the linear relationship between one or more independent variables (predictor variable) and the dependent variable (response variable). The model describes this relationship by fitting a line (simple linear regression) or a hyperplane (multiple linear regression), where the slope of the line (or the coefficient of the hyperplane) represents the influence of the independent variable on the dependent variable. The validity of the linear regression models is based on a series of assumptions, including a linear relationship, homoscedasticity, independence, and a normal distribution of the error terms. These assumptions ensure that the model can accurately reflect the true relationships between the variables. In practice, linear regression models are favored for their simplicity and ease of interpretability. It can be used not only to predict and interpret the relationship between variables, but also as a basis for other complex models.

In this paper, it is assumed that the pyrolysis reaction is conducted stably under certain conditions, and the yield is only affected by the composition of the raw material and the operating conditions.

For Tar, Char, and Water, you can describe the relationship with the INS through a linear regression model:

$$J_{tar} = \beta_0 + \beta_1 x + e \tag{2}$$

ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

$$J_{\text{char}} = \alpha_0 + \alpha_1 x + e \tag{3}$$

$$J_{\text{water}} = \gamma_0 + \gamma_1 x + e \tag{4}$$

where S is the mass ratio of INS; J is the yield of pyrolysis product; β0 β1 γ0 γ1 α0 α1 is the model parameter and should be obtained by fitting the data; e represents the model error.

The paper uses the least squares method to estimate the parameters in the linear model. This requires the sum of squares minimizing the error (e). The coefficient of determination (R2) was used to evaluate the determination of the model, and the t-test was used to evaluate the significance of each parameter model.

2.4 Thermal map mode

Heatmap is a chart with a color representation of data density or intensity. It shows the distribution and variation of the data by mapping the data to different colors. The design principle of heat map is based on the sensitivity of the human eye to color, and the brightness of different colors can convey different data density or intensity information. Through the intuitive display of colors, users can quickly understand the patterns and relationships of data. Heatmap Can accommodate a relatively large amount of data, help to find the relationship between the data, find out the extreme value, and depict the overall appearance of the data.

To evaluate the interaction effect between INS and the ratio of biomass and coal, the regression model was:

$$Q = \mathrm{a}_0 + \mathrm{a}_1 INS\_g + \mathrm{a}_2 \mathrm{Ratio} + \mathrm{a}_3 (INS\_g \times \mathrm{Ratio}) + b \tag{5}$$

where Q is the yield of pyrolysis product; INS \ -g is the content of INS; Ratio is the mixing ratio of biomass and coal; a0a1a2a3 is the model parameters; b is the error terms

2.5 Box plot analysis model

A boxplot is a graphical representation of the distribution of the data by drawing the quartiles of the data. It visually shows the location of the center, the range of spread, and outliers of the data. The boxplot clearly shows the central trend (median), the degree of dispersion (interquartile distance), and the distribution range of the data (minimum, maximum). Boxplots are able to effectively identify and highlight outliers in the data, those data points far from the primary data distribution.

2.6 T Test

T test, also known as Student's t-test, is a statistical method widely used in data analysis, mainly used to compare whether the means of two data groups differ significantly. T-test is a hypothesis test method to infer whether the means of the two data groups are significant by calculating t-values. It is based on the T distribution theory and uses the sample data to infer the differences between the population parameters. The T-test is usually used for normal distributed data with a small sample content (e. g. n <30) and the overall standard deviation σ unknown.

The paired sample t-test is used to analyze whether there are significant differences between the experimental values and the theoretical values of each product. The paired sample t-test (matched samples t-test) is mainly used to compare whether the observed values of the same group were significantly different at two different times or under two different conditions. The underlying assumption of this approach is that if the two treatments actually have no difference, the overall mean of the difference should be 0; if the two treatments are different, the overall mean of the difference should be far from 0.

The paired sample t-test assuming the difference (r) is from an approximately normal population. The purpose of this test was to determine whether the difference between these two paired samples was significantly greater than zero.

Assuming two samples, each composed of n observations, the difference is calculated for each pair:

$$x_i = P_i - Q_i \tag{6}$$

ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

where Sample 1 is P1, P2, P3,... Pn; Sample 2 is Q1, Q2, Q3,... Qn

The mean and standard deviation of the differences for all samples were calculated separately:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} \tag{8}$$

The T values are used to measure the magnitude of the difference between the means of the sample relative to the dispersion of these differences in the sample. The t-value statistic is calculated as follows:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{9}$$

The degree of freedom of the paired t-test is n-1, where n is the number of difference pairs, and the value of the t statistic is then compared to the t distribution to determine the value of P, the probability of observing this or more extreme result if the null hypothesis (assuming no difference between two samples) is true.

3. Interpretation of Result

3.1 Analysis of the results of the correlation analysis

The correlation coefficients between INS and Tar, Water and Char were calculated by correlation analysis:

The correlation coefficient between the INS and Tar is 0.209. Through the coefficient, there is a positive correlation between the two, but the correlation is not strong. The correlation coefficient between the INS and Water is 0.084. Through the coefficient, there is a certain positive correlation between the two, but the correlation is not strong. The correlation coefficient between the INS and Char is -0.199. Through the coefficient. There is a certain negative correlation between the two, but the correlation is not strong.

3.2 Analysis of the scatter plot model results

In order to better grasp the data of the whole, this paper draws the following scatter chart. The scatterplots are shown in Figure 1

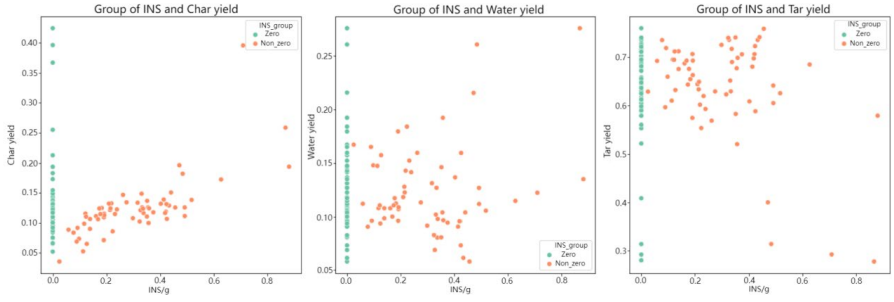

Figure 1: Association of INS with individual yields

From the scatter plot

The distribution of data points between INS and Tar is relatively scattered, but it can be seen that as the number of INS increases, Tar also tends to increase. The relationship between Tar and INS is less obvious, and the data points are very scattered. There is a certain trend of negative correlation between Char and INS, that is, with the increase of INS number, the number of Char has decreased.

ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

3.3 Analysis of the linear regression model

The yield model of the Tar

R2=0.044 indicates that the model is able to explain 4.4% of the variation in tar yield, which is a relatively low value and shows that the hexane insoluble has limited ability to explain the tar yield. Coefficient=0.0611 the positive coefficient can indicate that the n-hexane insoluble material and the tar yield are positively correlated, that is, when the n-hexane insoluble content increases. P =0.015 this data indicates that hexane insoluble material have a significant effect on tar yield.

A model of Water

R2=0.007 indicates that this value is relatively low, meaning that n-hexane insoluble has little effect on the aquatic rate. Coefficient =0.0243: The positive coefficient can indicate a positive correlation between hexane insoluble material and the aquatic rate, that is, when the number of hexane insoluble material increases, the aquatic rate has a trend to increase. P=0.412: According to non-significant statistics, it means that the effect of hexane insoluble material on aquatic rate is not significant.

Model of the Char

R2=0.040 indicates that the model can explain 4.0% of the variation of coke residue yield, with relatively low interpretation power. Coefficient = -0.0989: The negative coefficient indicates a negative correlation between n-hexane insoluble material and the yield of coke residue, that is, when the number of n-hexane insoluble material increases, the yield of coke residue tends to decrease. P=0.020: Statistically significant, which means that hexane insoluble has a significant effect on the coke residue yield.

From the coefficient of linear regression, although hexane tar yield and coke residue yield data of R2 value is small, but hexane insoluble material tar yield and coke residue yield has a more significant degree, the size of the P-value also shows the statistically significant, hexane insoluble content on the three effects of the yield is statistically important.

3.4 Analysis of the results of the heat map model

Through the data calculation of formula 5, the main effect of the INS and the ratio and its interaction response can be evaluated. The significant interaction response shows that the proportion of the INS will affect the effect of the ratio on the yield of the pyrolysis products. In order to showcase the interaction effects intuitively, the heat map is drawn to show the change of pyrolysis products under different INS and ratios. The Tar heat map is shown in Figure 2, The Char heat map is shown in Figure 3, The Water heat map is shown in Figure 4.

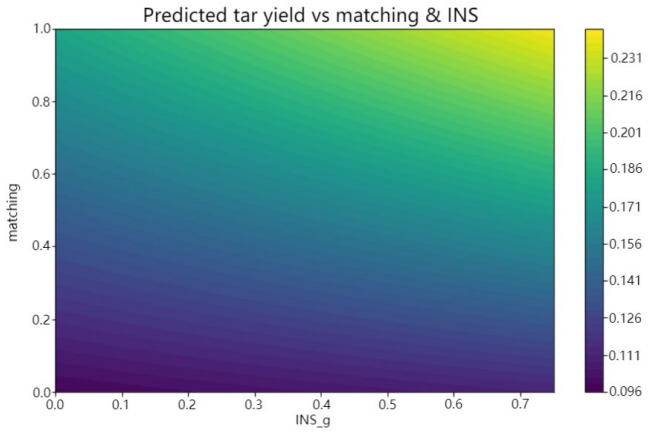

Figure 2: The heat map of the Tar

ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

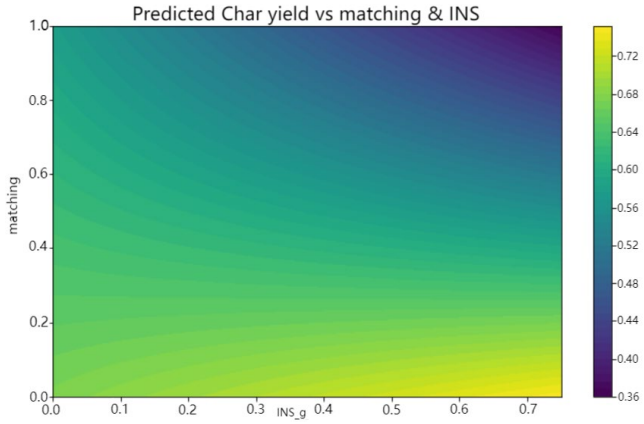

Figure 3: The heat map of the Char

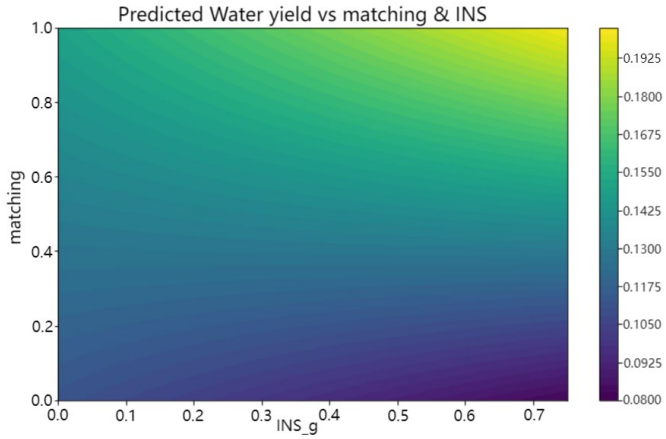

Figure 4: The heat map of the Water

The following conclusions can be obtained via the Heatmap:

The positive correlation between Tar and Water and mixture proportion, that is, the larger the mixing ratio, Tar and Water, the higher the mixing ratio, which is more significant at low INS concentration. The negative correlation between Char and the mixing proportion is, that is, the larger the mixing proportion, the lower the Char, which is more significant at the higher INS concentration.

3.5 Box plot analysis model results analysis

To better visualize the results, the following boxplots are plotted. The boxplot model is shown in Figure 5

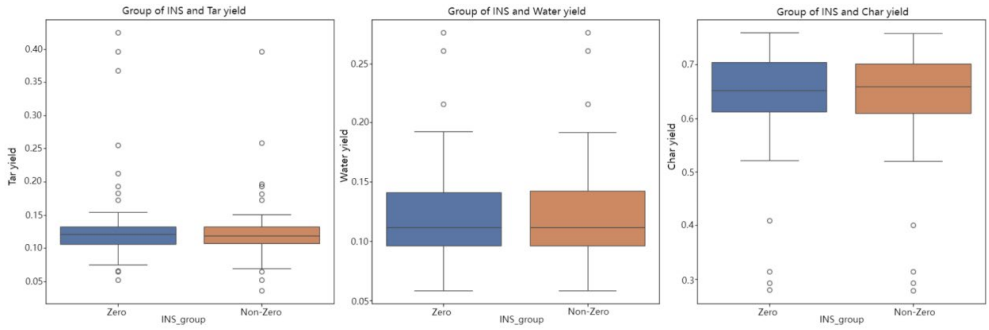

Figure 5: Box line diagram

The conclusions can be drawn from the box plot:

Tar has a wider distribution of Tar in the non-zero INS group, with a relatively high median, indicating that INS may increase Tar. The difference between the two groups of Water is not obvious, which is consistent with the results of previous statistical analysis that the effect of INS on Water is not

ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

significant. Lower range of Char in the non-zero INS group, the distribution range of Char is relatively low and the median is relatively low, and the results are consistent with the finding that the effect of INS on Char is negatively correlated.

3.6 Analysis of the results from the T-test

For the combinations with significant differences, the subgroups are analyzed to determine in which way the differences between the experimental values and the theoretical calculated values are more pronounced. This includes the calculation of experimentally derived values for each significantly different product, and evaluates the magnitude of these differences. Considering the most obvious differences in practical applications to facilitate optimization and adjustment, the top three with the largest differences in each mixture ratio may be more practical and intuitive. In this way, it is more rapid to identify which specific proportions require the focus on attention and possible adjustment. In this paper, the proportion of the first three mixtures with the largest difference for each combination is extracted and visually displayed. Through this method, it is more intuitive to see the experimental and theoretical values under conditions. To the results shown in Figure 6

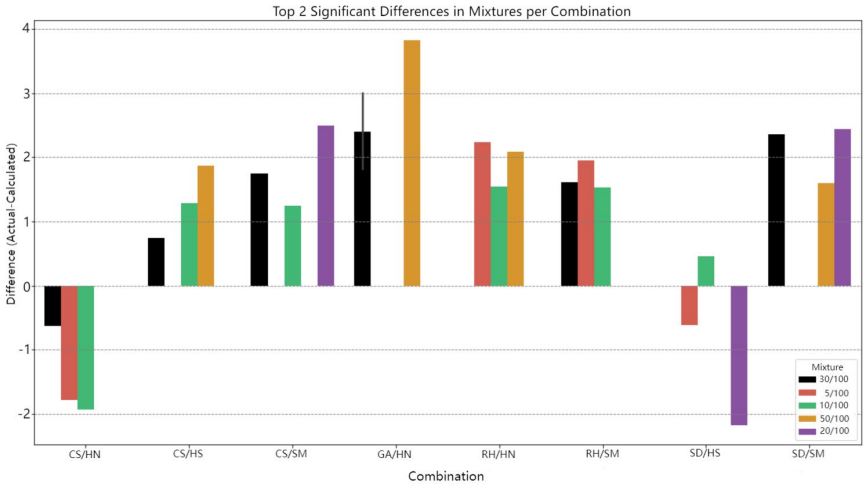

Figure 6: Plot of subgroup analysis results

4. Conclusion

Significant synergies between biomass and coal during copyrolysis.This synergistic effect can beverified by statistical tests (e. g. analysis of variance, T test, etc.). The experimental results show that the yield and quality of biomass copyrolysis products with coal are generally better than those of pyrolysis alone, especially under certain specific conditions, the synergistic effect is more significant. Different kinds of biomass and coal with different coal steps show different characteristics during the copyrolysis. By learning the model, these properties can be identified and suitable biomass and coal are selected for copyrolysis according to the actual requirements. Through statistical test and optimization algorithm, the optimal biomass to coal mixing ratio can be found. This ratio optimizes the yield and quality of the copyrolysis products.

The research on biomass and copyrolysis of coal based on statistical tests and machine learning. By deepening statistical test methods, optimizing experimental design and data analysis, constructing and optimizing machine learning model, and interdisciplinary integration and technological innovation, the internal law of coppyolysis process can be further revealed, the product quality and production efficiency can be improved, and the industrial application of copyrolysis technology can be promoted. In the future, with the continuous deepening of related research and the continuous progress of technology, biomass and coal copyrolysis technology is expected to play a greater role in the energy field.

References

[1] Wang Yanshun, Xiao Ruirui, Cong Xingshun, et al. Characteristics of copyrolysis of biomass and coal [J]. Guangdong Chemical Industry, 2023, 50 (05): 20-21 + 16.

ISSN 2616-5872 Vol.6, Issue 5: 37-44, DOI: 10.25236/AJEE.2024.060506

[2] Liu Junjie, Wu Ruirui, Yuan Yue, et al. Current situation and development trend of coal pyrolysis process [J]. Chemical Technology and Development, 2020, 49 (12): 23-27.
[3] Xu Qing, Peng Liming, Ling Changming, et al. Research on the technical progress of cothermolysis of biomass with four different organics [J]. Journal of Guangzhou Navigation University, 2019, 27 (02): 73-78 + 69.
[4] Lu Bo, Ma Mingming, Su Xiaoping, et al. Progress in biomass copyrolysis [J]. Chemical Technology, 2021, 29 (06): 54-58.
[5] Dai Chongyang, Tian Yishui, Hu Erfeng, et al. Research on the copyrolysis characteristics of biomass and low-rank coal and its technical progress [J]. Journal of Solar Energy, 2021, 42 (12): 326-333.