ISSN 2616-5767 Vol.4, Issue 8: 43-47, DOI: 10.25236/AJETS.2021.040806

# Study on the Loss Process of Gasoline Octane Number Based on Principal Component Analysis

**Yuqing Yang, Zhengfu Li*, Liping Wu**

*School of Computer Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu, 222005, China*
**Corresponding author*

***Abstract:*** *This paper mainly takes the octane number loss in the gasoline refining process as the research object, and studies the main variables that affect the octane number loss and the prediction of octane number loss. In this paper, the data are screened out based on the 3 σ principle, and then the main variables are determined based on the principal component analysis (PCA). The sample size of the original data is 325 and the variables are 367. First of all, the descriptive statistics of the data samples are carried out to get the approximate range of the data. According to the different conditions satisfied by the data, Pearson correlation coefficient and Spelman rank correlation coefficient are selected to calculate and analyze the correlation between the two variables. Then the principal component analysis method was used to select the main variables affecting the octane number loss, and the index of 367 variables was reduced to 17 independent and representative variables, and the cumulative contribution rate of 17 variables reached 80.7%. Taking into account the continuous expansion of data samples in the future, the solution method of factor analysis is provided.*

***Keywords:*** *Correlation coefficient, principal component analysis, factor analysis*

## 1. Introduction

Gasoline is the main fuel for small vehicles, and the exhaust emission from gasoline combustion has an important impact on the atmosphere. To this end, countries around the world have developed increasingly strict gasoline quality standards. The focus of gasoline cleaning is to reduce the sulfur and olefin content in gasoline, while keeping the octane number as far as possible.

China's crude oil dependence is more than 70%, and most of it is sulfur and high-sulfur crude oil from the Middle East. Heavy oil typically makes up 40-60% of crude oil, which is difficult to use directly (sulfur is also high in impurities).In order to make effective use of heavy oil resources, China has vigorously developed heavy oil and light technology with catalytic cracking as the core to convert heavy oil into gasoline, diesel and low olefins. More than 70% of gasoline is produced by catalytic cracking, so more than 95% of sulfur and olefins in finished gasoline come from CATALYTIC cracking gasoline. Therefore, FCC gasoline must be refined to meet the quality requirements of gasoline. In this paper, based on the collected samples of FCC gasoline refining unit, the prediction model of gasoline octane number (RON) loss is established based on data mining technology, and the optimal operating conditions of each sample are obtained.

## 2. Correlation analysis

### *2.1. Pearson correlation coefficient [1]*

Suppose that we have two sets of data: $X:\{X_1, X_2, \cdots, X_n\}, Y:\{Y_1, Y_2, \cdots, Y_n\}$,

Calculate the sample mean: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

Sample covariance: $\operatorname{cov}(X, Y) = \frac{\sum_{i=1}^{n} \left( X_i - \bar{X} \right) \left( Y_i - \bar{Y} \right)}{n-1}$

ISSN 2616-5767 Vol.4, Issue 8: 43-47, DOI: 10.25236/AJETS.2021.040806

Pearson correlation coefficient of samples: $r_{XY} = \frac{\operatorname{cov}(X,Y)}{S_X S_Y}$

### *2.2. Spearman correlation coefficient*

Definition: X and Y are two sets of variables, and their Spearman (rank) correlation coefficients are:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \tag{1}$$

$d_i$ is the grade difference of $X_i$ and $Y_i$. And it turns out to be between minus 1 and 1. Meanwhile, the correlation is explained in the following table

*Table 1: Correlation size interpretation*

| Correlation | Negative | Positive |
|---|---|---|
| No correlation | 0.09-0.0 | 0.0-0.09 |
| Weak correlation | 0.3-0.1 | 0.1-0.3 |
| Moderate correlation | 0.5-0.3 | 0.3-0.5 |
| Strong correlation | 1.0-0.5 | 0.5-1.0 |

Before calculating Pearson's correlation coefficient, a scatter plot must be made to see whether there is a linear relationship between two groups of variables. Here, SPSS software is used to draw a matrix scatter plot.

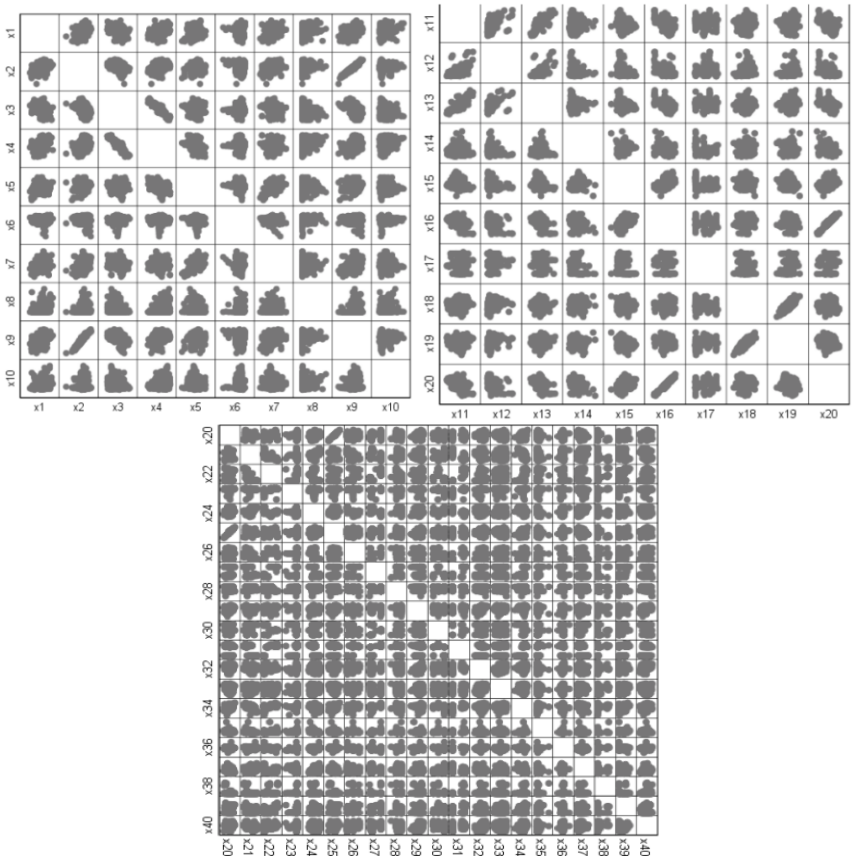

*Figure 1: Matrix scatter diagram between X1 and X40*

As can be clearly seen in the above figure, there is a strong positive correlation between x2 and x9, x16 and x20, x18 and x19, x20 and x25, and a strong negative correlation between x3 and x4.

At the same time, in order to verify the reliability of the correlation between variables, the Spelman rank correlation coefficient matrix is calculated and compared. It can be found that the correlation

ISSN 2616-5767 Vol.4, Issue 8: 43-47, DOI: 10.25236/AJETS.2021.040806

between variables is basically the same, thus proving the reliability of the above methods. In fact, we tend to pay more attention to significance, that is, hypothesis testing, than the size of the correlation coefficient.

## 3. Conditions of Pearson correlation coefficient hypothesis testing

The experimental data are usually assumed to be paired from the population of normal distribution and the gap between the related data is not too large, and each group of samples are sampled independently.

Jarque−Beratest (JB) test was used. For a random variable, assuming its skewness is S and kurtosis is K, we can construct the JB statistic:

$$JB = \frac{n}{6}\left[ S^2 + \frac{(K-3)^2}{4} \right] \tag{2}$$

And it can be proved that if it is a normal distribution, then in the case of large samples (chi-square distribution with 2 degrees of freedom), and the skewness of the normal distribution is 0, and the kurtosis is 3.

H0: The random variable is normally distributed, H0: The random variable does not follow a normal distribution

Because the Spelman spearman rank correlation coefficient does not need the data to satisfy the condition of normal distribution, the Spelman rank correlation coefficient method is used to analyze it. Hypothesis testing of Spelman correlation coefficient: in the case of large samples, statistics $r_s\sqrt{n-1} \sim N(0,1)$

$$H_0 : r_s = 0, H_1 : r_s \neq 0 \tag{3}$$

The test value is calculated: $r_s\sqrt{n-1}$ and the corresponding P value is compared with 0.05. When p value is greater than 0.05, the null hypothesis is accepted; otherwise, the null hypothesis is rejected.

## 4. Construction of principal component analysis model

### *4.1. Identify the main variables for modeling*

In this paper, a sample matrix of size is constructed [2].

$$x = \begin{bmatrix} x_{1\cdot 1} & x_{1\cdot 2} & \cdots & x_{1\cdot 367} \\ x_{2\cdot 1} & x_{2\cdot 2} & \cdots & x_{2\cdot 367} \\ \vdots & \vdots & \ddots & \vdots \\ x_{325\cdot 1} & x_{325\cdot 2} & \cdots & x_{325\cdot 367} \end{bmatrix} = (x_1, x_2, \ldots, x_{367}) \tag{4}$$

First, the sample matrix is processed with Z standardization:

To calculate the mean and standard deviation of each column, the formula is as follows [3]:

$$\bar{x}_j = \frac{1}{325}\sum_{i=1}^{325} x_{ij} \quad \text{and} \quad S_j = \sqrt{\frac{\sum_{i=1}^{325}\left( x_{ij} - \bar{x}_j \right)^2}{325-1}} \tag{5}$$

Calculate the covariance matrix of the standardized sample:

ISSN 2616-5767 Vol.4, Issue 8: 43-47, DOI: 10.25236/AJETS.2021.040806

$$R=\begin{bmatrix} r_{1\cdot1} & r_{1\cdot2} & \cdots & r_{1\cdot367} \\ r_{2\cdot1} & r_{2\cdot2} & \cdots & r_{2\cdot367} \\ \vdots & \vdots & \ddots & \vdots \\ r_{367\cdot1} & r_{367\cdot2} & \cdots & r_{367\cdot367} \end{bmatrix} \tag{6}$$

$$r_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}\left(X_{ki}-\bar{X}_i\right)\left(X_{kj}-\bar{X}_j\right) \tag{7}$$

Calculate the eigenvalues and eigenvectors of the covariance matrix:

The R matrix satisfies the semidefinite matrix, $tr(R)=\sum_{k=1}^{367}\lambda_k=367$ and $r_{ij}=r_{ji}$

Its characteristic values: $\lambda_1\geq\lambda_2\geq\lambda_3\ldots\geq\lambda_{367}\geq 0$

Calculate principal component contribution rate and cumulative contribution rate:

$c_i=\frac{\lambda_i}{\sum_{k=1}^{367}\lambda_k}(i=1,2,\cdots,367)$ ( $c_i$ represents the contribution rate of the first eigenvalue)

$cc_i=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{367}\lambda_k}(i=1,2,\cdots,367)$ ( $cc_i$ Represents the cumulative contribution rate to the first eigenvalue)

Generally, the first, second, and... corresponding to the characteristic values whose cumulative contribution rate exceeds 80% are taken. And the m (m≤ P) principal component.

*Table 2: Eigenvalue and contribution rate (part)*

| Principal components | Eigenvalue | Contribution | Cumulative contribution rate |
|---|---|---|---|
| 1 | 118.253 | 0.322 | 0.322 |
| 2 | 39.852 | 0109 | 0.431 |
| 3 | 24.230 | 0.066 | 0.497 |
| 4 | 20.830 | 0.057 | 0.554 |
| 5 | 14.738 | 0.040 | 0.594 |

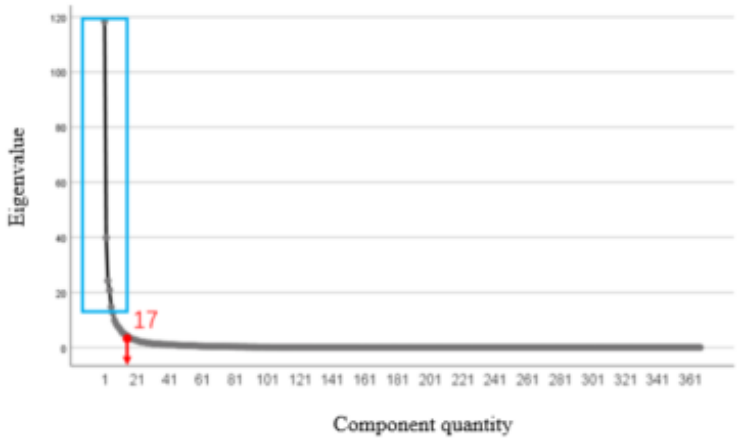


*Figure 2: Gravel figure*

The first 17 components were extracted and analyzed as principal components. The cumulative contribution rate of the first 17 principal components is 81% [4], so we can consider only the first 17 principal components, which can well summarize the original variables.

ISSN 2616-5767 Vol.4, Issue 8: 43-47, DOI: 10.25236/AJETS.2021.040806

### *4.2. Factor analysis*

To build the factor model, we need to estimate the factor loading matrix $A_{p\times m}=\left(a_{ij}\right)$ and the variance matrix D,

Factor analysis is to express variables as linear combinations of common factors and special factors. In addition, we can conversely express the common factors as linear combinations of the original variables to obtain factor scores.

$$\begin{cases} x_1=u_1+a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1 \\ x_2=u_2+a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2 \\ \cdots \\ x_p=u_p+a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p \end{cases} \Rightarrow \begin{cases} f_1=b_{11}x_1+b_{12}x_2+\cdots+b_{1p}x_p \\ f_2=b_{21}x_1+b_{22}x_2+\cdots+b_{2p}x_p \\ \cdots \\ f_m=b_{m1}x_1+b_{m2}x_2+\cdots+b_{mp}x_p \end{cases} \tag{8}$$

The $i_{th}$ factor score can be written as $f_i=b_{i1}x_1+b_{i2}x_2+\cdots+b_{ip}x_p \quad (i=1,2,\cdots,m)$.

After calculating the coefficients of the factor scoring function, all factor scores can be calculated.

## 5. Conclusion

This paper studies the loss of octane number in the process of gasoline refining. After data processing, principal component analysis is used to study, first of all, to determine the main variables. Based on the original data sample size of 325 and variables of 367, and then descriptive statistics of the data samples, the approximate range of the data is obtained. Pearson correlation coefficient and Spelman rank correlation coefficient are selected to calculate and analyze the correlation between the two variables. Then the principal component analysis method is used to select the main variables that affect the octane number loss, and after considering the continuous expansion of data samples, the solution method of factor analysis is provided.

## References

*[1] Zhang Xiaopeng, Ma Lijing, Empirical Research on Employment Rate Factors in China, Modern Trade and Industry, 20(5): 29-30, 2008.*
*[2] Wang Xuemin, Application of Multivariate Statistical Analysis [M], Shanghai, Shanghai University of Finance and Economics Press, 209-244, 2017.*
*[3] Fan Tongda, Jiang Bing, Multiple linear regression model of water Consumption based on principal component Analysis: A case study of Anhui Province, Infrastructure Optimization, 28 (2): 53-55, 2007.*
*[4] Lv Qiongshuai. Optimization and Research of BP Neural Network [D]. Zhengzhou University, 2011.*