A Research on Mathematical Model and Prediction Method for Raw Tobacco Material Quality Based on Cluster Classification

Hao Shu\textsuperscript{1,2,3}, Song Gao\textsuperscript{1,2,3,*}, Bin Xiong\textsuperscript{1,2,3}, Qunshan Yan\textsuperscript{1,2,3}, Pei Wei\textsuperscript{1,2,3}, Ziwei Wang\textsuperscript{1,2,3}, Yuxing Tong\textsuperscript{1,2,3}, Xi Lei\textsuperscript{1,2,3}, Yizhen Chen\textsuperscript{1,2,3}, Zhichang Liu\textsuperscript{1,2,3}, Dezhi Chen\textsuperscript{4}, Jun Xu\textsuperscript{4}, Yi Wang\textsuperscript{4}

1China Tobacco Hubei Industrial Co., Ltd., Wuhan 430040, China
2Hubei Xinye Reconstituted Tobacco Development Co., Ltd., Wuhan 430056, China
3Applied Technology Research of Reconstituted Tobacco Hubei Province Key Laboratory, Wuhan, Hubei 430040
4State Key Laboratory of Coal Combustion, School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

*Corresponding authors: gaosong@hjl.hbtobacco.cn

Abstract. The chemical components of 189 kinds of Chinese tobacco materials were analyzed. There were obvious correlations between the total sugar content and reducing sugar content, total sugar content and total nitrogen content, reducing sugar content and nitrogen content, nicotine content and nitrogen content, respectively. There were certain correlations between chemical components and quality evaluation score of the tobacco materials. Total sugar and reducing sugar content showed similar correlations, while nitrogen, chloride ion and nicotine content showed similar correlations. The correlation between each single component and the quality evaluation score was not strong, and it was difficult to be used as a classification standard. Using the cluster classification method, the tobacco materials were well classified into three clusters. For each cluster, a reasonable relationship between the quality evaluation score and the chemical component contents of the raw tobacco material was set up. The method for rapid evaluation of tobacco quality score were established based on these results.

Keywords: Cluster classification, Tobacco quality, Rapid evaluation, Chemical components.

1. Introduction
China is the world’s largest tobacco production and consumption country, producing and consuming over one third of the world’s tobaccos. The quality of raw tobacco material is closely related to the processes of tobacco production and planting [1]. The quality and style characteristics of raw tobacco materials with different origins, grades and types are significantly different. After the tobacco materials entered in the industrial enterprises, the technical personnel will evaluate the grade of the tobacco materials by sensory evaluation, and the relevant physical and chemical components will be
also tested. The physical and chemical components of tobacco [2] are quite complex [3] and diverse [4]. The classification methods of the raw tobacco materials based on the single component or simple combinations of some components are hard to satisfy the needs of fine classification [5]. Due to the requirements of controlling the raw material grade and quality in tobacco production process [6], it is of great significance to further refine or establish the classification method of the raw tobacco materials.

Cluster classification is a statistical analysis method that uses mathematical methods to achieve classification [7, 8]. According to the similarity degree of things, cluster classification can intuitively show the differences and connections of classified objects [9]. It is an important data mining algorithm [10], and has been widely used to analyze the sample properties. At present, there have been some reports on cluster classification of some components of tobacco in China. These works mainly focused on chemical components, aroma substance contents and sensory index, such as total sugar, total nitrogen, nicotine, nonvolatile organic acid, amino acid content, etc. Generally, the samples can be divided into several categories according to the different characteristics, which can be used in tobacco material matching, quality sampling, authenticity identification and other applications. Although there have been studies on the cluster analysis of tobacco materials, most of them are relatively one-side, considering only one or just few factors, which lack a comprehensive consideration of the various evaluation indicators of tobacco materials. Besides, good correlations between the chemical components of tobacco and its tobacco quality score were not set up, which were very important for the rapid evaluation of tobacco quality. Cluster classification can be used to divide the tobacco materials into some clusters with similar properties, which is beneficial for setting up the quickly prediction models for the quality of tobacco materials by integrating multiple indicators of tobacco materials [11], to satisfy the requirements of rapid and accurate classification of raw tobacco materials with no need for manual testing. Therefore, it is very meaningful to further study the raw tobacco materials with the cluster classification method.

In this study, the chemical components of 189 kinds of Chinese tobacco materials were analyzed. 34 kinds of typical tobacco materials were selected for further analysis by the cluster classification method. The relationships between the chemical components and the tobacco quality were discussed. The method for rapid evaluation of tobacco quality score was also tried to establish.

2. Materials and methods

189 kinds of raw tobacco materials were provided by Hubei Xinye Reconstituted Tobacco Development Co, Ltd. These samples were from different producing area at different years, and had different grades. In order to study the effects of sample collection time on the tobacco properties, the same tobacco sample was repeatedly collected at different sampling dates. 40 tobacco grades including tobacco fragments, tobacco leaves, stems and sticks were selected in 11 different dates. Each sample of the raw tobacco material was divided into two parts. One was used for chemical component analysis and another one was further used for quality evaluation. In order to reduce the influence of moisture content in the sample on the results, all tobacco materials were dried before the analysis and tests.

The chemical components of tobacco materials are the key parameters to determine its quality. In this study, the chemical components of tobacco materials were detected by continuous flow method on a scalar san continuous flow analyzer. The contents of water-soluble sugar (total sugar and reducing sugar), chlorine, nitrogen and nicotine were measured according to tobacco industry standards of YC/T 159-2002 [12], YC/T162-2011[13], YC/T 161-2002 [14] and YC/T 468-2013 [15].

34 kinds of the raw tobacco materials were further selected for sensory evaluation. The evaluators were professional engineers from China Tobacco Hubei Industrial Co, Ltd. The evaluation descriptors included flavor type, aroma quality, aroma quantity, penetrability, impurity, strength, concentration, fineness, clustering, irritation, dryness, cleanliness, and sweetness. The quality evaluation score of each tobacco material was obtained by summarizing all descriptor scores. SPSS software was used to
cluster the data and establish a fitting correlation between the tobacco quality evaluation score and the contents of chemical components of tobacco samples.

3. Results and discussion

3.1. Characteristics of chemical components in tobacco materials

The content characteristics of each chemical component are shown in Fig. 1, in which the horizontal axis is the content of each component, and the vertical axis is the percentage of the content range. It can be seen that the distribution of total sugar and nitrogen content of 189 samples is basically the normal distribution, the maximum probability content of total sugar is about 10%, and the nitrogen is about 2.0%, there is a few kinds of tobacco distributed in the large value (total sugar content > 20%, nitrogen content > 3.5%). The distribution of reducing sugar content deviates from the normal distribution slightly. The distribution of reducing sugar content between 4% and 14% is relatively uniform. When the content is higher than 15%, the distribution is few and the distribution is fewer at 20%. The distribution of chloride ion content in the samples presents two aggregation regions, which are 0-1.2% and 1.2%-2.2%, respectively. Although the chloride ion content in the samples is mainly distributed in the range of 0-1.2%, the high chloride ion content in a few kinds of sample should be paid attention to. The nicotine content in the samples distributes between 0-3.5%, and the distribution is relatively uniform. Except that the samples with a nicotine content of about 1% are few, and large quantity samples with a nicotine content of about 0.5%, the samples with other nicotine content are almost uniform distribution.

In order to obtain the statistical data of chemical components of the samples, the maximum, minimum, average, standard deviation and variance of each component of the samples were calculated and shown in Table 1. Combined with Fig. 1 and Table 1, the contents of total sugar, reducing sugar, nicotine, chloride ion and total nitrogen of the samples were distributed between 2.45-28.48%, 0.54-26.78%, 0.17-3.33, 0.23-2.19 and 1.03-4.10%, respectively. The average contents were 11.42%, 8.85%, 1.66%, 1.08% and 2.02%, respectively.

| Component      | Minimum | Maximum | Average | Standard deviation | Variance |
|----------------|---------|---------|---------|--------------------|----------|
| Total suger (wt%) | 2.45    | 28.48   | 11.42   | 5.17               | 26.78    |
| Reducing suger (wt%) | 0.54    | 26.78   | 8.85    | 4.79               | 22.95    |
| Nicotine (wt%)    | 0.17    | 3.33    | 1.66    | 0.86               | 0.73     |
| Chloride ion (wt%) | 0.23    | 2.19    | 1.08    | 0.53               | 0.28     |
| Nitrogen (wt%)    | 1.03    | 4.10    | 2.02    | 0.42               | 0.17     |
Fig. 1. Content distribution of chemical components in tobacco samples.

Fig. 2. Correlation among chemical components in tobacco samples.

3.2. Correlation among chemical components
Fig. 2 shows the correlations among the five chemical components of tobacco materials. Matrix scattering method was used to analyze the data. It can be seen that there are obvious correlations
between the total sugar content and reducing sugar content, total sugar content and nitrogen content, reducing sugar content and nitrogen content, nicotine content and nitrogen content respectively.

![Fig 3. Correlations between total sugar and reducing sugar, total sugar and nitrogen, nitrogen and reducing sugar, nicotine and nitrogen of the raw tobacco materials.](image)

In order to further quantify the correlation among these components, the correlations of these components were further analyzed and shown in Fig. 3. It can be seen that the correlation between reducing sugar and water-soluble total sugar is the best one among all the chemical components of tobacco materials. When the linear function was used for fitting the correlation, the correlation coefficient could reach 0.92, which indicated that the proportion of reducing sugar content in total sugar was basically unchanged. In addition, with the increase of total sugar content, the nitrogen content decreased rapidly at first, then remained unchanged, which indicated that there was a certain correlation between nitrogen and total sugar content in tobacco material. The corresponding nitrogen content also decreased with the increase of reducing sugar, showing a similar law with the change of total sugar. There was a strong correlation between nicotine and nitrogen content. With the increase of nicotine content, the nitrogen content increased.
### 3.3. Relationship between the chemical components and the tobacco quality evaluation score.

**Table 2.** The contents of chemical components and score of quality evaluation.

| Number | Total sugar (wt. %) | Reducing sugar (wt. %) | Nicotine (wt. %) | Chloride ion (wt. %) | Nitrogen (wt. %) | Score of quality evaluation |
|--------|---------------------|------------------------|------------------|----------------------|-----------------|-----------------------------|
| 1      | 12.86               | 4.83                   | 1.09             | 0.74                 | 2.31            | 83.50                       |
| 2      | 11.55               | 5.77                   | 1.19             | 0.66                 | 2.16            | 83.50                       |
| 3      | 4.98                | 2.71                   | 1.88             | 0.76                 | 2.26            | 72.00                       |
| 4      | 2.87                | 0.87                   | 2.02             | 0.45                 | 2.48            | 80.25                       |
| 5      | 8.16                | 6.61                   | 2.38             | 0.76                 | 2.28            | 74.25                       |
| 6      | 19.23               | 15.56                  | 1.75             | 0.65                 | 1.87            | 81.00                       |
| 7      | 2.45                | 0.79                   | 2.66             | 1.93                 | 2.60            | 64.25                       |
| 8      | 11.09               | 9.21                   | 2.87             | 0.61                 | 2.49            | 67.50                       |
| 9      | 5.48                | 3.28                   | 0.50             | 1.27                 | 2.51            | 80.50                       |
| 10     | 14.52               | 13.13                  | 2.34             | 0.80                 | 2.23            | 81.00                       |
| 11     | 18.34               | 15.44                  | 2.27             | 0.65                 | 2.63            | 83.50                       |
| 12     | 14.95               | 10.41                  | 1.35             | 0.68                 | 2.37            | 83.50                       |
| 13     | 5.50                | 4.21                   | 3.12             | 1.72                 | 2.68            | 64.25                       |
| 14     | 10.53               | 8.38                   | 2.22             | 0.63                 | 2.49            | 80.25                       |
| 15     | 10.29               | 8.71                   | 2.34             | 0.60                 | 2.40            | 81.00                       |
| 16     | 8.18                | 6.49                   | 2.45             | 0.63                 | 2.53            | 75.75                       |
| 17     | 24.11               | 21.79                  | 1.95             | 0.61                 | 1.90            | 68.25                       |
| 18     | 19.26               | 18.71                  | 2.63             | 0.50                 | 2.42            | 80.50                       |
| 19     | 4.62                | 3.04                   | 2.37             | 0.97                 | 2.72            | 77.75                       |
| 20     | 5.54                | 3.77                   | 2.09             | 0.70                 | 2.99            | 78.75                       |
| 21     | 13.61               | 12.67                  | 2.94             | 0.71                 | 2.64            | 81.25                       |
| 22     | 5.48                | 3.28                   | 0.50             | 1.27                 | 2.51            | 80.50                       |
| 23     | 22.87               | 20.92                  | 1.84             | 0.87                 | 2.08            | 68.50                       |
| 24     | 14.10               | 11.06                  | 1.56             | 0.92                 | 2.37            | 75.50                       |
| 25     | 10.61               | 8.78                   | 2.15             | 0.89                 | 4.10            | 73.50                       |
| 26     | 11.71               | 7.33                   | 2.00             | 0.58                 | 2.22            | 77.25                       |
| 27     | 8.38                | 7.14                   | 2.58             | 1.08                 | 2.09            | 61.25                       |
| 28     | 15.03               | 13.36                  | 2.25             | 0.65                 | 1.77            | 69.00                       |
| 29     | 6.66                | 4.29                   | 1.34             | 0.37                 | 1.80            | 75.25                       |
| 30     | 11.72               | 8.01                   | 1.22             | 0.45                 | 1.59            | 81.00                       |
| 31     | 6.35                | 3.79                   | 1.49             | 0.57                 | 2.11            | 77.00                       |
| 32     | 8.87                | 8.11                   | 2.82             | 1.40                 | 2.29            | 61.25                       |
| 33     | 4.89                | 4.31                   | 1.67             | 0.68                 | 2.05            | 71.50                       |
| 34     | 5.75                | 4.82                   | 2.96             | 1.70                 | 2.39            | 65.00                       |

Table 2 shows the content of chemical components and the score of quality evaluation of 34 kinds of tobacco materials. The scatter diagram of the content of each chemical component and the scores of quality evaluation of tobacco material samples are provided (as shown in Fig.4). According to the scatter diagrams, it can be judged that there were certain correlations between the chemical components and quality evaluation score. Total sugar and reducing sugar content showed similar correlations, while nitrogen, chloride ion and nicotine content showed similar correlations. Besides, it
could be seen that the correlation between single component and quality evaluation score was not
strong enough to be a factor that can predict the quality evaluation score of the raw tobacco materials.

3.4. Systematic cluster classification

Based on the correlations between the content of each component and the score of quality evaluation, we took the sum of the content of total sugar and reducing sugar as variable 1, and took the sum of the content of nicotine, chloride ion, and nitrogen as variable 2. We further took the score of quality evaluation as a case labeling, and then select the square Euclidean distance for cluster classification. Before clustering classification, in order to eliminate the error caused by the order of magnitude between the values, the data should be standardized first. The dendrogram of cluster classification is shown in Fig.5 According to the results of the dendrogram, 34 tobacco materials can be divided into three clusters, as shown in Fig. 6. Among them, 11 samples numbered 1, 2, 3, 4, 9, 22, 26, 29, 30, 31 and 33 were classified into the first cluster, and 14 samples numbered 5, 7, 8, 13, 14, 15, 16, 19, 20, 21, 25, 27, 32 and 34 were classified into the second cluster. The nine samples numbered 6, 10, 11, 12, 17, 18, 23, 24 and 28 were classified into the third cluster.

According to the cluster classification results, variable 1 and 2 were used as independent variables, and the score of quality evaluation was used as the dependent variable. The nonlinear regression analysis was carried out for the three clusters of samples, and a rapid prediction model for the score of quality evaluation of tobacco materials was established.

For the first cluster of the raw tobacco materials:

\[ Z = 0.153x^2 - 3.756x - 6.741y^2 + 51.959y \]  

(1)

For the second cluster of the raw tobacco materials:

\[ Z = 0.007x^3 - 0.231x^2 + 1.967x + 1.318y^3 - 19.6y^2 + 81.475y \]  

(2)

For the third cluster of the raw tobacco materials:

\[ Z = -0.074x^2 + 4.772x + 0.23y^2 - 0.45y \]  

(3)

Where Z represents the score of quality evaluation, X represents the sum of total sugar and reducing sugar content, and Y represents the sum of nicotine, chloride ion and nitrogen content.

According to the calculation results of the fitting formula, the average error percentages of the first, second and third clusters are respectively 2.5%, 5.5% and 4.4%, which indicates that the rapid prediction model can give a reasonable result about the score of the quality evaluation of tobacco materials with a very fast rate when the chemical components are known. For the tobacco material to be tested, if we get the content of each chemical component by experiment, and the cluster of the sample can be determined according to the distance between the sample point and the three cluster centers. And then, the score of the tobacco material can be calculated by the rapid prediction model of the tobacco material based on cluster classification. In needs to be pointed out that this method can be further optimized by increasing the number of the tobacco materials in the database, and improve the prediction accuracy of tobacco properties.
Fig 4. Relationship between chemical components and the score of quality evaluation.

Fig 5. Dendrogram of the cluster classification of the 34 kinds of the tobacco materials.
Fig 6. Three clusters of the tobacco material sample

4. Conclusions
The content of total sugar, reducing sugar, nicotine, chloride ion and nitrogen in the tobacco materials ranged from 2.45% to 28.48%, 0.54 to 26.78%, 0.17 to 3.33%, 0.23 to 2.19% and 1.03 to 4.10%, respectively, with an average content of 11.42%, 8.85%, 1.66%, 1.08% and 2.02%, correspondingly. Among them, the contents of total sugar and reducing sugar, total sugar and total nitrogen, reducing sugar and nitrogen, nicotine and total nitrogen were significantly correlated, respectively. There were certain correlations between chemical components and sensory evaluation score of the raw tobacco materials. Total sugar and reducing sugar content showed similar correlations, while nitrogen, chloride ion and nicotine content showed similar correlations. The correlation between each single component and the quality evaluation score was not strong, and it was difficult to be used as a classification standard.

According to the results, the tobacco materials can be classified into three clusters by cluster classification. A rapid and reasonable prediction model for the score of quality evaluation of tobacco materials was established. It can accurately predict the score of tobacco material. This rapid prediction method can meet the needs of enterprises for rapid and refined classification of tobacco materials and avoid the trouble of manual testing.

References
[1] X.Q Hu, M.Y Xu, Y.Q He, et al. Effects of future climate change on climatic suitability of flue-cured tobacco plantation in Yunnan, China. Chinese Journal of Applied Ecology. 27 (2016) 1241-1247.
[2] Z Tang, L Chen, Z Chen, et al. Climatic factors determine the yield and quality of Honghe flue-cured tobacco. Scientific reports. 10 (2020) 19868.
[3] T.A Perfetti, A Rodgman. The Complexity of Tobacco and Tobacco Smoke. Beiträge zur Tabakforschung / Contributions to Tobacco Research., 24 (2011) 215-232.
[4] T.A Perfetti, A Rodgman. The Chemical Components of Tobacco and Tobacco Smoke. CRC Press, Boca Raton, Fl, USA. 2013.
[5] F Li, B Lu, L Yang, et al. Effects of Partial Replacement of Chemical Fertilizer by Bio-organic Fertilizer Kunyijian on Yield and Quality of Tobacco Leaves. Asian agricultural research. 12 (2020) 42-46.
[6] Y Bi, S Li, L Zhang, et al. Quality evaluation of flue-cured tobacco by near infrared spectroscopy and spectral similarity method. Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy. 215 (2019) 398-404.
[7] G Pfeiffer, K Kunze. PS-33-3 Towards “objective” EMG-differential-diagnosis. Performance of discriminant and cluster classification in myositis. Electroencephalography and clinical neurophysiology / electromyography and motor control. 97 (1995) S167-S168.

[8] K Woolnough, C Newby, M Richardson, et al. P75 Cluster classification as a predictor of adverse radiological outcomes in allergic fungal airways disease (AFAD). Thorax. 70 (2015) A113.

[9] A Fielding. Cluster and classification techniques for the biosciences. Cambridge, UK; New York: Cambridge University Press. (2007)

[10] W Wei, E.A Huerta, B.C Whitmore, et al. Deep Transfer Learning for Star Cluster Classification: I. Application to the PHANGS-HST Survey. Monthly notices of the Royal Astronomical Society. 493 (2020) 3178-3193.

[11] M.V Gil, M.P González-Vázquez, R García, et al. Assessing the influence of biomass properties on the gasification process using multivariate data analysis. Energy conversion and management. 184 (2019) 649-660.

[12] YC /T 159-2002, Tobacco and tobacco products-Determination of water soluble sugars-Continuous flow method.

[13] YC /T 162-2011, Tobacco and tobacco products-Determination of chloride-Continuous flow method.

[14] YC /T 161-2002, Tobacco and tobacco products-Determination of total nitrogen-Continuous flow method.

[15] YC/T 468-2013, Tobacco and tobacco products-Determination of total alkaloids-Continuous flow (potassium thiocyanate) method.