Parameters Optimization in Threshing and Redrying Process Based on Random Forest and Orthogonal Experiment Design

Xiaoyan Huang¹, Guangren Zhao¹, Yurui Zheng¹, Jing Huang¹, Junli Xie² and Zhiwei Jia²*¹

¹Chenzhou Redrying Factory, Hunan tobacco redrying Co., Ltd, Zhenzhou, China
²College of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha, China

*Corresponding author email: Jiayege@csust.edu.cn

Abstract. To improve the quality of threshing and re-drying production, a method fused by random forest and orthogonal experiment scheme are presented to optimal the process parameters. Random forest performed regression modelling on the historical data of process parameters and moisture value, and ranked the process parameters. The model is verified by field test and the gap between the predict results and reality is less than 5%, which satisfied the on-site demands. To achieve optimal parameter combination with limited numbers of time of experiments, orthogonal experiment scheme according to the order of importance is designed. The optimal parameter combination was calculated according to the results and verified by the random forest model and field test. Results of our optimal combination in filed test showed it was better than the historical ones. This fusion method illustrates its potentials to promote the decision of process of threshing and re-drying.

Keywords: Threshing and re-drying; Random forest; Orthogonal experiment design; Parameters optimization.

1. Introduction

In the process of threshing and re-drying, stabilizing the quality of tobacco threshing and re-drying has always been an important issue faced by enterprises, setting reasonable process parameters is an important and effective method to this problem [1-8]. The main difficulties in optimal this process lie in two aspects: too many parameters concerned in this process and too many disturb factors affect the results. The common way to set these parameters depends on the experience of the on-site technicians and the temporary situation on the spot. And the final tobacco leaf moisture is difficult to meet an orderly standard since the uneven experience and responsibility of the field staff. Hence, there are huge historical data and small number of effective independent data for this parameter optimal problem.

To find out the most important parameters and optimize them, lots of valuable researches have been done. Yang et al [2] used Analytic Hierarchy Process (AHP) and orthogonal experiment to test the designated tobacco leaves. By setting different key process parameters of moistening leaves, Xu et al [3] found the frequency of circulating fan and the steam volume is very important. According to the experience of three kinds of producing areas, 11 groups of different parameter combinations were set by Wu [6]. they found that the quality will be greatly improved if the appropriate parameters are set according to the raw materials of tobacco leaves from different producing areas. Liu et al [7] established the control models of the two indexes and adjustment parameters in the moisturizing section of threshing and re-drying through uniform experiment and multiple linear regression analysis. And the results show that the moisturizing effect of the moisturizing section can be accurately controlled and optimized.
The above parameters optimization researches are based on the experience of field personnel or enterprises. Even if rare models are built and analysed, the data also depends on the combination of indicators and process parameters obtained through experience settings. The huge historical data are ignored since too many disturb factors involved and could not simply remove. To solve this problem, random forest are adopted to dig the useful information from historical data, and the process parameters are ranked by the export moisture index. The orthogonal experiment scheme is proposed by based on the importance rank of parameters and the optimal parameter combination is presented. Optimal combinations are predicted by random forest model and verified by field test.

2. Materials and Methods

2.1. System Overview

The algorithm design is shown in Figure 1. The original historical data is trained by random forest algorithm after pre-processing and build a random forest model, which can forecast the output by input parameters combinations. The parameter feature importance could be achieved and orthogonal experiment scheme is designed based on it. orthogonal experiment scheme is executed in field test and predicted by random forest model, the different of the two are analysed to verify and improve the random forest model.

![Algorithm flow chart of the presented method.](image)

An optimal parameter combination is given from the range analysis on results of orthogonal experiment scheme. and the corresponding index is predicted by the improved random forest model and verified. More details are given in the sections below.

2.2. Data Pre-processing

Materials: The annual production process parameters and tobacco index data accumulated by CHENZHOU Re-drying factory in 2019, including tobacco leaf grades such as B3F and C3F (B3F and C3F are the leaf grades, where B and C indicate that the growth position of leaf are the top and the middle respectively, 3 is the quality grade of leaf and F indicates that the colour of leaf is orange). They are totally 119472 process parameters records for the whole year, every record includes 8 parameters of moistening leaves sectors, 5 parameters of threshing sectors,13 parameters of re-drying section, Equivalent to 26 features of threshing and re-drying parameters in total.

The process parameters, as well as the data of physical and chemical indexes of tobacco leaf, show the real-time monitoring of tobacco leaf in the process of threshing and re-drying. However, since this production is composed by multiple production links with sequential sequence, the data pre-processing stage should eliminate the abnormal and missing values, so as to provide high-quality data for later model training and mining the knowledge contained in the data.

According to the process parameters, i.e. the input value is \( x_1, x_2, \cdots, x_p \), where \( x_j \in X_j (j = 1, 2, \cdots, p) \) form a \( p (p = 26) \) dimensional input vector \( x \). Base on input vector \( x \), we construct an input space \( X \), the value of which equals \( X = x_1, x_2, \cdots, x_p \). The quality index of threshing and re-drying is mainly determined by the moisture value of output tobacco leaf, which is calculated as follow:

\[
Y = \left( 1 - \frac{\text{Outlet moisture} - \text{Average moisture}}{\text{Average moisture}} \right) \times 100\% \quad (1)
\]
Generally, the processing target is $Y = 12\%$.

2.3. Random Regression Forest

The random forest algorithm is a supervised learning algorithm proposed by Breiman in 2001 \[9\]. The core idea is integrating the optimal classification and regression of the judgment results of many random decision trees, which explain the effect of the input variables $X_1, X_2, \cdots, X_n$ on the explanatory variable $Y$. Decision tree is the core of random forest, which can be generated optimally according to conditional entropy and information gain. The algorithm goes as follow:

a) Various variables and different number of samples are selected from data set $D$ using Bagging method. These combinations could use to construct numerous Random decision trees and the corresponding classification node.

b) Calculate the information entropy of data set $D$ ($H(D)$), the joint entropy between $D$ and variable $X_i$ ($H(D, X_i)$) as well as the conditional entropy between the variables $X_i$ and dataset $D$ ($H(D | X_i)$):

$$H(D) = E[- \log D_i] = \sum_{i=1}^{n} D_i \log_2 D_i$$  \hspace{1cm} (2)

$$H(D, X_i) = - \sum_{i=1}^{n} \sum_{j=1}^{n} p(D, x_i) \log(D, x_i)$$  \hspace{1cm} (3)

$$H(D | X_i) = H(D, X_i) - H(X_i)$$  \hspace{1cm} (4)

Where $D_i$ is the probability of the data $i$ in the data set $D$.

c) Calculate the information gain $g(D, X_i)$ of each variable $X_i$ relative to the data set $D$:

$$g(D, X_i) = H(D) - H(D | X_i)$$  \hspace{1cm} (5)

d) Taking information entropy as a measure, the tree with the fastest entropy decline is generated. Meanwhile, the variables with the largest information gain rate and Gini coefficient are selected as variables of split nodes of decision tree. Where Gini coefficient is defined as follows:

$$Gini(p) = 1 - \sum_{k=1}^{n} \frac{|C_k|}{|D|}^2$$  \hspace{1cm} (6)

e) A large number of random decision trees which can be processed in parallel are generated according to the parameter combination randomly selected from dataset $D$. These random decision trees make up a random decision forest together.

Random forest studies the effect of several features $(X_1, X_2, \cdots, X_P)$ on the dependent variable $Y$, setting the dependent variable $Y$ as the export quality score index, and giving a mapping $Y = Y_1, Y_2, \cdots, Y_n$ of threshing and re-drying data $(X_1, X_2, \cdots, X_P)$. In the process of continuous bagging $B$ times, select a random sample in the training set to replace the process data of threshing and re-drying:

For $b = 1, 2, \cdots, B$ : (a) Using $X, y$ instead of data sample $X_B, X_B$; (b) training $X_B$.

After training, the missing $X_i$ can also be predicted by averaging predictions from all individual regression trees on the $X_i$:

$$f = \frac{1}{B} \sum_{b=1}^{B} f_b(x_i)$$  \hspace{1cm} (7)

Similarly, the prediction standard deviation of all individual regression trees on $X_i$ is taken as the evaluation of prediction uncertainty. Many independent decision trees can be trained in the training set, which together form a strong correlation stochastic forest model. Generally, the number of trees in the forest depends on the size and nature of the training set \[10\].

2.4. Data Processing

For the physical and chemical indexes of process parameters and tobacco export quality scores accumulated before Chenzhou re-drying plant in 2019, 70% and 30% of them are selected as the training set and test set of cart decision tree and random regression forest, using the complete process parameter characteristics to map the tobacco export quality score index data. The decision tree regression model
and random forest regression model, which are the sharpest weapons in machine learning, are used to fit the training data set.

In order to better evaluate the performance of regression model, root mean square error (RMSE) and mean absolute error (MAE) are used as model indexes, which are defined as follows [10]:

\[
RMSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{predicted})^2
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_{predicted}|
\]

\[
Accuracy = \frac{1}{n} \sum_{i=1}^{n} \left(100 - \frac{|y_i - \hat{y}_{predicted}|}{y_i}\right)
\]

Among them, root mean square error (RMSE) is very sensitive to the extremely large and small errors in regression, which can well reflect the accuracy of regression, while mean absolute error (MAE) can better describe the degree of dispersion between predicted value and actual value.

### 3. Algorithm Model Evaluation

The RMSE and MAE of actual and predicted values after optimization of different hyper parameters of random regression forest, and the accuracy of outlet moisture are shown. For the model results of setting different hyper parameters of random regression forest, the same indexes are RMSE, Mae and accuracy, and the results are shown in Table 1. According to the analysis of Table 1, for the multi-dimensional input sample x, the best accuracy the random forest after optimization reaches 98.74%. In general, Random forest has better learning ability in establishing regression model of multidimensional process parameters of threshing and re-drying.

Table 1. Model verification of Random forest algorithm.

| Hyper parameters setting | n_etrs=50   | n_etrs=75   | n_etrs=100  | n_etrs=200  |
|---------------------------|-------------|-------------|-------------|-------------|
| Max_feature=all           | RMSE:2.471  | RMSE:2.61   | RMSE:2.69   | RMSE:2.71   |
|                           | MAE:1.22    | MAE:1.27    | MAE:1.27    | MAE:1.28    |
|                           | Accuracy:96.58% | Accuracy:96.54% | Accuracy:96.12% | Accuracy:96.51% |
| Max_feature=log2(all)     | RMSE:3.4058 | RMSE:3.577  | RMSE:3.5262 | RMSE:3.455  |
|                           | MAE:1.41    | MAE:1.454   | MAE:1.445   | MAE:1.43    |
|                           | Accuracy:97.36% | Accuracy:98.31% | Accuracy:96.32% | Accuracy:95.34% |
| Max_feature=sqrt(all)     | RMSE:1.69   | RMSE:1.750  | RMSE:1.755  | RMSE:1.73   |
|                           | MAE:1.08    | MAE:1.098   | MAE:1.101   | MAE:1.09    |
|                           | Accuracy:98.74% | Accuracy:97.72% | Accuracy:97.71% | Accuracy:96.72% |

Table 2 is the ranking of the characteristic importance of the random forest model (n_etrs=50, max_feature=sqrt(all)), which indicates the ranking of the process parameters with relatively high correlation.
between the export quality score index and the impact of the two models. It can provide reference for the factor level design of the next orthogonal experimental design. Since there are 26 different variables concerned with the index, and the value range of every variable are relatively large, so, it is possible to optimal this problem using traversal methods. According to the demands for on-site, an orthogonal experiment scheme is designed to verify the random forest model and propose an optimal parameter combination.

4. Orthogonal Experiment Design Based on Feature Importance

4.1. Principle of Orthogonal Experiment Design
Orthogonal experiment design is a kind of experiment design method which studies many factors and levels. According to the orthogonality, some representative points are selected from the comprehensive test, which are uniform and comparable. Orthogonal table is determined by interfering factors and the corresponding levels. According to the orthogonality of the table, the effect of all combinations in the table could be represent by some representative combinations which of the number are much less. The results of all combinations test and the special representative combinations test are approximately equivalent in theory [8]. Therefore, orthogonal table design is an efficient, fast and economic multi-factor test design method. The range analysis method is generally used in the analysis of the results of orthogonal test design.

4.2. Orthogonal Experiment Design
After using the random forest model to obtain all the feature importance rankings of the leaf re-drying process parameters, due to the limited number of tobacco leaf tests on the production site, the level of each factor is determined based on the practical experience of the production line technicians. 3 levels, so take the 9 factors and 3 levels of orthogonal design table L27 (39) for orthogonal experimental design, and orthogonal experimental design for the process parameters with the highest feature importance illustrated in Table 1. Furthermore, Table 3 gives the orthogonal experimental factor level table of random Forest. According to the above table, the orthogonal experiment of random forest can be designed. The specific parameter scheme is shown in Table 4. Meanwhile, the corresponding moisture value of outlet leaf are calculated using random forest model and listed in Table4.

| Combination | A | B | C | D | E | F | G | H | I | Prediction |
|-------------|---|---|---|---|---|---|---|---|---|------------|
| 1           | 1 | 1 | 2 | 2 | 1 | 3 | 2 | 1 | 3 | 14.11      |
| 2           | 1 | 1 | 1 | 3 | 3 | 2 | 2 | 1 | 1 | 12.20      |
| 3           | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 12.07      |
| ...         |   |   |   |   |   |   |   |   |   | ...        |
| 27          | 3 | 3 | 1 | 1 | 3 | 1 | 2 | 1 | 3 | 11.71      |

Table 3. Factors and level of Orthogonal experiment.

| Fact or level | A | B | C | D | E | F | G | H | I |
|---------------|---|---|---|---|---|---|---|---|---|
| G₁(℃)        | 63| 72| 69| 67| 64| 34| 32| 29| 26|
| G₂(℃)        | 68| 75| 72| 71| 69| 38| 35| 32| 29|
| G₃(℃)        | 73| 78| 75| 73| 72| 42| 37| 35| 32|

Table 4. Orthogonal experimental design scheme (levels) and responding predicted moisture value by model.
5. Results and Analysis
Orthogonal experiment scheme is executed, and the moisture value from field test are illustrated and compared with the predicted one (Figure 2). The gap between these two are less than 5%. This gap may mainly cause by the yearly different of leaf quality, and it could be accepted by the on-site demands. It verifies the feasibility of the presented random forest model.

![Figure 2. The outlet moisture of scheme from filed test and predict.](image)

The specific results are shown in Table 6. The orthogonal optimal parameter combination obtains a moisture value of 12.44 predicted by random forest model, while the field test value is 11.93, the gap of which is 0.51, equals 4.27%, satisfied the on-site demands. Compared with the most recent work in literature [5-8], the proposed method has two main advantages. Firstly, the influence factors in this method comes from abundant historical data instead of experience, which are more comprehensive and independent. Secondly, the forecast ability of the proposed random forest, which are verified by field test, could easily give a reasonable predict moisture and could apply in different type of tobacco leaf. Since the historical data came from the normal process status, and the corresponding parameters setting always determined by on-spot experience, the range of which are always limited and incomplete. The generalization ability of presented model is doubtable, especially these parameter combinations which is far from the on-spot experience parameter setting, the different between the predict results by our method and the field test could be relatively big. To avoid this problem, enlarge the historical data range is a suitable way to improve the proposed method.

| combination | A | B | C | D | E | F | G | H | I |
|-------------|---|---|---|---|---|---|---|---|---|
| K1 mean     | 13.10 | 12.74 | 12.38 | 12.72 | 13.01 | 12.18 | 12.80 | 12.45 | 12.82 |
| K2 mean     | 13.09 | 13.39 | 13.59 | 12.75 | 13.09 | 12.97 | 12.58 | 13.04 | 12.84 |
| K3 mean     | 12.04 | 12.10 | 12.26 | 12.76 | 12.14 | 13.08 | 12.85 | 12.74 | 12.57 |
| Optimal level | A3 | B3 | C3 | D1 | E3 | F1 | G2 | H1 | I3 |
| Rj          | 1.05 | 1.29 | 1.32 | 0.01 | 0.94 | 0.10 | 0.26 | 0.29 | 0.27 |

6. Conclusions
In this paper, random forest is use to map the process parameters and the quality score index of threshing and re-drying. Multiple process parameters are treated as independent variables, and the leaf moisture value of outlet is taken as corresponding variable to form this mapping. This moisture value is predicted
by random forest model, and the super parameters of the model are optimized. The importance ranking of process parameters can be obtained by using the priority of the characteristic importance of the model. To search an optimal parameter combination under limited available number of times of experiments, the orthogonal experiment scheme is designed based on the parameter importance ranking. An optimal parameter combination, obtained from this scheme, is predicted by random forest model and verified by field test. By using the result of orthogonal experiment design, and the combination of process parameters is realized the result of field process parameter verification and model prediction shows that the error between the real production result and prediction result is less than 5%. The experiment shows that the prediction effect of process parameters is good, which can provide good help for the decision-making of process parameters before threshing and re-drying.

Acknowledgements
This work is supported by Special science and technology project of Hunan Branch of China Tobacco Corporation (No. HN2020KJ17, HN2020KJ16, HN2020KJ15, HN2020KJ14) and the Foundation of Hunan Educational Committee (No.19B016).

References
[1] Tang Ying. Discussion on Optimization of process parameters of threshing and Re-drying[J]. Public Communication of Science & Technology, 2013, 5(16):134-135.
[2] Yang Yang et al, Parameter Optimization of Threshing and Re-drying Process in Tobacco Processing[J]. Agricultural Engineering, 2018, 8(08):83-85.
[3] Xu Guojin. Effect of different parameters of foliage equipment on the quality of foliage [J]. Agriculture and Technology, 2017, 37(06):99.
[4] Zhou Xuejuan, Fu Qixing. Effect of re-drying temperature on the quality of tobacco [J]. YUNNAN Science and Technology, 2013, 26(05):66-69.
[5] Zhan Lei, Xiao Mingli, Effects of re-drying temperature on quality of fresh flavor style tobacco in southwest of China [J]. Journal of HUNAN Agricultural University (Natural Sciences), 2018, 44(02):136-139+150.
[6] Wu Fang, et al. The relationship between the setting of different technological parameters and the quality of threshing and Re-drying[J]. Technology and enterprise, 2012(14):324.
[7] Liu Yanling et al. Establishment of Control Model of Threshing and Re-drying Condition Leaf Section with Tobacco Water and Temperature [J]. Anhui Agri. Sci. Bull. 2017, 23(22):109-111.
[8] Lu Youxiang, Xu Qimin, Yang Bo et al. Study on Optimization of Southern ANHUI Tobacco Leaf Threshing and Re-drying Technical Parameters Based on Orthogonal Design. Journal of ANHUI Agri. Sci, 2015, 43(27):218-219+224.
[9] Leo Breiman. Random Forests. Machine Learning [J]. 2001, 45, 5–32
[10] Su Zhitong, Wang Wujun. Research and Analysis of Calcination Process Parameters Based on Random Forest[J]. COMPUTER ENGINEERING & SOFTWARE, 2018, 39(04):148-150.
[11] Hu Hongtian, Li Xiaohong et al. Quality inspection of leaf beating tobacco: YC/T147-2010[S]. BEIJING: Standards Press of china, 2011