JUSTC

Article

http://justc.ustc.edu.cn

Received: April 23, 2021; Accepted: September 05, 2021

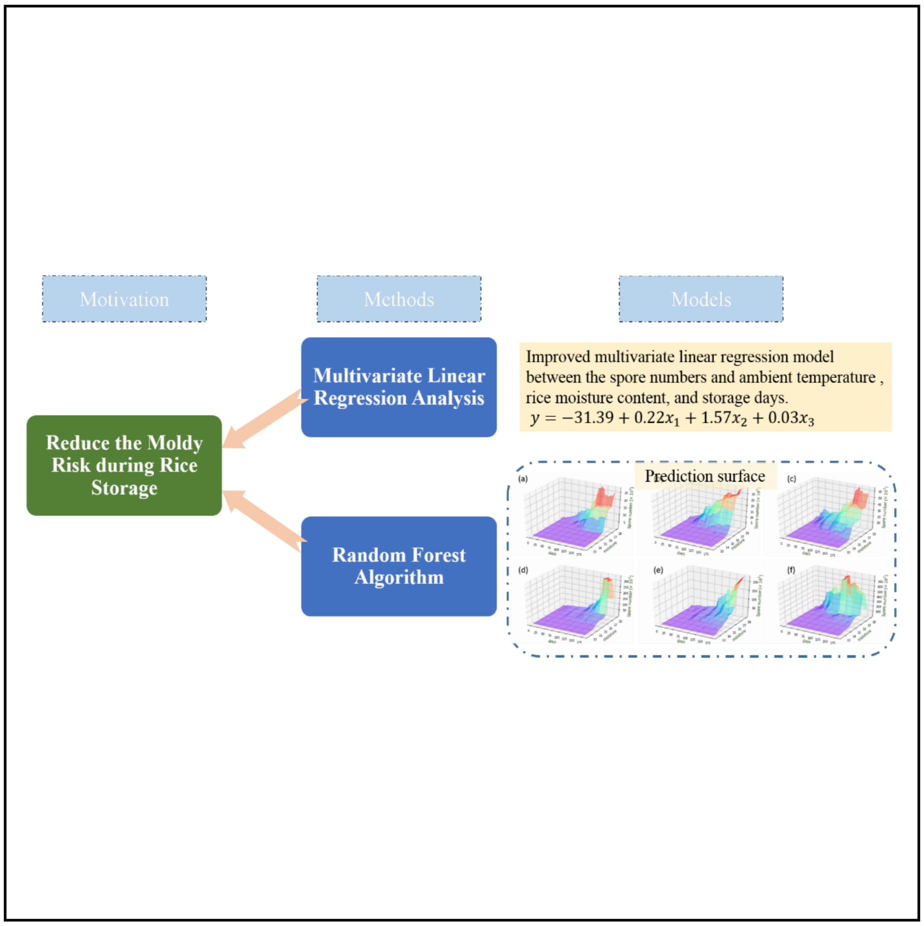
The control of moldy risk during rice storage based on multivariate linear regression analysis and random forest algorithm

Yurui Deng[1], Xudong Cheng[1], Fang Tang[2], and Yong Zhou[1] ✉

[1]State Key Laboratory of Fire Science, University of Science and Technology of China, Hefei 230027, China;
[2]Academy of National Food and Strategic Reserves Administration, Beijing 100037, China

✉Correspondence: Yong Zhou, E-mail: yongz@ustc.edu.cn

Graphical abstract



Motivation

Methods

Models

Multivariate Linear Regression Analysis

Reduce the Moldy Risk during Rice Storage

Random Forest Algorithm

Improved multivariate linear regression model between the spore numbers and ambient temperature , rice moisture content, and storage days.
$y = -31.39 + 0.22x_1 + 1.57x_2 + 0.03x_3$

Prediction surface

Public summary

- A multivariate linear regression model among several important factors, such as spore number and ambient temperature, rice moisture content and storage days, were developed based on the experimental data.

- Random forest algorithm was introduced into fungal spore prediction during grain storage.

- The established regression models can be used to predict the spore number under different ambient temperature, rice moisture content and storage days during the storage process.

Citation: Deng Y R, Cheng X D, Tang F, et al. The control of moldy risk during rice storage based on multivariate linear regression analysis and random forest algorithm. *JUSTC*, **2022**, 52(1): 6. DOI: 10.52396/JUSTC-2021-0118

JUSTC
Article
http://justc.ustc.edu.cn
Received: April 23, 2021; Accepted: September 05, 2021

The control of moldy risk during rice storage based on multivariate linear regression analysis and random forest algorithm

Yurui Deng[1], Xudong Cheng[1], Fang Tang[2], and Yong Zhou[1] ✉

[1]State Key Laboratory of Fire Science, University of Science and Technology of China, Hefei 230027, China;
[2]Academy of National Food and Strategic Reserves Administration, Beijing 100037, China

✉Correspondence: Yong Zhou, E-mail: yongz@ustc.edu.cn

Cite This: *JUSTC*, **2022**, 52(1): 6 (8pp)   Read Online   SI Supporting Information

Abstract: Clarifying the mechanism of fungi growth is of great significance for maintaining the quality during grain storage. Among the factors that affect the growth of fungi spores, the most important factors are temperature, moisture content and storage time. Therefore, through this study, a multivariate linear regression model among several important factors, such as the spore number and ambient temperature, rice moisture content and storage days, were developed based on the experimental data. In order to build a more accurate model, we introduce a random forest algorithm into the fungal spore prediction during grain storage. The established regression models can be used to predict the spore number under different ambient temperature, rice moisture content and storage days during the storage process. For the random forest model, it could control the predicted value to be of the same order of magnitude as the actual value for 99% of the original data, which have a high accuracy to predict the spore number during the storage process. Furthermore, we plot the prediction surface graph to help practitioners to control the storage environment within the conditions in the low risk region.

Keywords: rice storage; fungi growth; spore number; multivariate linear regression; random forest algorithm

CLC number: X826 **Document code:** A

1 Introduction

Grain mildew is an important cause of grain loss during grain. The condition of grain mildew can be quickly determined by detecting the number of fungi spores[1]. Therefore, clarifying the mechanism of fungi growth is of great significance for maintaining the quality during grain storage. Rice as one of the main crops has a wide range of cultivation in China. Its storage environment and conditions are especially important to ensure an edible quality. For example, rice with high moisture content is very easy to be heated and molded during storage. If it cannot be controlled appropriately, it could spread rapidly inside the grain heap, resulting in the damage of rice even food accidents. The storage of rice could be dependent on many influencing factors, such as ambient temperature, the moisture content of rice, and storage days. Although it is commonly known that these factors affect the storage of rice, the related quantitative analysis is limited to address the behind mechanism to benefit the related maintenance of appropriate storage environment.

Many studies have built a solid basis for the rice storage research, but these studies have been much focused on one or two influencing factors, lacking a general analysis to address those main influencing factors. For example, Yin et al.[2] investigated the rice mycoflora in China and found that the main fungi at the early stage of the new grain are field fungi. With the increase of storage time, the field fungi decreased which was gradually replaced by storage fungi. Purushtham et al.[3] addressed the relationship between the change of storage quality and the growth of fungi. Laca et al.[4] indicated that the mold is mainly distributed at the surface of the grain. Genkawa et al.[5] investigated the changes of fatty acid value, germination rate and mold activity during rice storage under the effects of moisture contents. Soponronnarit et al.[6] found that the quality of the rice with high initial moisture content is more likely to change. Zhou et al.[7, 8] studied the exchanges of rice fungal mycoflora under different storage conditions, revealing that the ambient temperature and rice moisture content can directly affect the growth of rice storage fungi during the storage process. Therefore, it is quite critical to address the behind mechanism of rice storages under the effects of the main influencing factors.

Biological methods have been widely used previously, which have focused on revealing the relationship between the growth mechanism of fungi and environmental factors such as rice moisture content, and temperature. However, the related mechanism on fungal growth is often complicated due to the interaction of multiple influencing factors. It is difficult to fully understand the inherent causal relationship between these factors and fungal growth and establish an accurate mathematic model based on mechanism analysis. The usual approach is to collect a large amount of data and build a mathematical model based on statistical analysis of these data, but very few people carry out research work in this area due to the long-term process of the experimental tests.

In this study, based on the experimental data of rice storage experiments carried out by the Tang Fang research group at the Academy of National Food and Strategic Reserves Administration, a multivariate linear regression model and ran-

DOI: 10.52396/JUSTC-2021-0118

dom forest model were developed between spore number of rice and those main influencing factors, containing ambient temperature, rice moisture content, and storage days. Firstly, Regression analysis, which is a statistical model widely used to address the behind correlation, was conducted to establish a linear relationship model between spore number of rice and those main influencing factors. After building the model, significant tests on the multivariate linear regression model were also performed to improve the model, like the F test, the t-test, and the residual analysis. However, even though the linear regression model could deduce a clear function for the spores number of stored rice, it do not have sufficient precision for a prediction model.

Random forest, a relatively new algorithm in the AI field, is an ensemble of multiple decision trees and a nonparametric statistical technique to enhance the prediction accuracy, which has been widely used to address engineering problems[9–11]. As a result, we introduce decision tree and random forest algorithm into fungal spore prediction during grain storage. For the random forest model, it could control the predicted value to be of the same order of magnitude as the actual value for 99% of the original data, which have a high accuracy to predict the spore number during the storage process. Furthermore, the prediction surface graph can be plotted to help practitioner to control the storage environment within the conditions in the low risk region. The results have suggested that the random forest model showed a higher predictive ability in classification and regression of this issue, which benefit to understand the growth pattern of fungal spores and assist managers in taking measures in advance.

2 Materials and methods

2.1 Experimental details and treatments

Experimental studies on the mechanism of fungal growth have shown that there are three main factors affecting the growth of rice spores, namely ambient temperature, rice moisture content, and storage time. As a result, relevant experiments were conducted in the laboratory to simulate the rice storage environment.

The rice sample was obtained from Heilongjiang province in the north of China. After determining the initial moisture content, deionized water was added into the samples to adjust the sample moisture to the target moisture, and then the samples were sealed and refrigerated at 4 ℃ to balance the moisture for 30～60 days until the moisture content of the rice reaches the target moisture. Then, rice with 10 moisture contents, including 13%, 13.5%, 14%, 14.5%, 15%, 15.5%, 16%, 16.5%, 17%, 18%, were sealed and placed in different storage temperatures (10, 15, 20, 25, 30, 35 ℃) in biochemical culture to simulate storage for 180 days. So there were a total of 10 × 6 = 60 experimental scenarios. 2 samples from each of all the 60 scenarios were taken every 10 days to detect moisture and the number of spores by microscope, and the average of the spore number of the two samples was used for further analysis. For example, by counting the numbers under the microscope, the grain with a moisture content of 16% shows a spores number of 4650000 at 30 ℃ after 180 days of storage. After done the whole experiments, a total of 18 × 60 = 1080 groups of data were obtained.

2.2 Statistical analysis

2.2.1 *Multivariate linear regression model*

Theory Multivariate linear regression model as an important and popularly adopted method to address the relationship of multiple influence factors was utilized in this study. This method was briefly introduced here, more details are presented by Chen[12]. The multivariate linear regression model assumes that the interpreted variable (dependent variable) y is affected by n explanatory independent variables following a linear relationship. The model can be expressed as

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n \tag{1}$$

where b_0, b_1, b_2, b_3, …, b_n are regression coefficients based on experimental data.

It is assumed that there is a linear relationship between the dependent and independent variables. The rationality of this assumption will be determined later by a significance test. The significance test can be further divided into an F test (the significance test for the regression model as a whole) and T-test (the significance test for each independent variable)[12].

Residual analysis The multivariate linear regression method mentioned above is one of the popularly adopted regression method for engineering uses. In the practical applications, abnormal data may exist due to the interference of accidental factors during data acquisition. Even if the significance test is passed, some abnormal data cannot be completely excluded. The residual $e_i = y_i - \hat{y}_i (i = 1, 2, \cdots, m)$ is the difference between the observed values of each data group in the dataset and the fitted values of the regression model, which is usually introduced for the residual analysis to eliminate abnormal data. If the confidence interval of the residual contains zero, it indicates that the regression model can well match the original data, otherwise, it can be regarded as an abnormal point. After finding and eliminating the abnormal points, the regression model is re-established with the remaining data to improve the accuracy of the regression model.

2.2.2 *Random forest algorithm.*

Theory Decision tree is a common method for various machine learning tasks, but this method can occur with low precision and over-fitting. Therefore, many researchers combine other models with decision trees to improve the accuracy of the classifier or regression models. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners.

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest, which was suggested by Breiman[13]. After the random forest algorithm was proposed, it has been widely used for classification and regression in various fields. Vladimir Svetnik et al.[14]introduced a new method for predicting a compound's quantitative or categorical biological activity based on random forest. Sandra Oliveira et al.[15] investigated the potential applicability of the Random Forest method in the assessment of fire occurrence. Classification and regression trees or CART for short is a term introduced to decision tree algorithms by Leo Breiman[16] that can be used for classification or regression predictive modeling problems. Creating a CART model involves selecting input variables and split points on those variables until a suitable tree is constructed. A greedy approach is used to divide the space called recursive binary splitting[17]. The greedy algorithm means to select the input variables and the specific split or cut points to minimize the cost function. For regression predictive modeling problems, the objective is to minimize the cost function of the chosen split points across all training samples that fall within the rectangle:

$$J = \frac{1}{m}\sum_{j=1}^{m} L(y_i, \hat{y}_{R_j}),$$

$\hat{y}_{R_j}$ is the average prediction value of the training samples in the rectangle R_j[18]. The tree construction ends with a predefined stopping criterion, for example, if the split of a node results in a child node whose node size is less than the user-specified minimum child node size value, the node will not be split[19]. In addition, many improved algorithms have been proposed to enhance the prediction accuracy[20–22].

Principle of random forest algorithm Given a training set $X=x_1, x_2, \ldots, x_n$ with responses $Y=y_1, y_2, \ldots, y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set. For b from 1 to B, the training examples from X, Y are named X_b, Y_b and the corresponding decision tree f_b is trained by X_b, Y_b. Then, the predictions of multiple decision trees are combined and the final prediction results are obtained by voting. A large number of theoretical and empirical studies have proved that RF has a high prediction accuracy and is not prone to overfitting. overfitting.

Random forest is generated by the method as following:

(Ⅰ) Generate n samples from the original set by means of bootstrap aggregating method;

(Ⅱ) Assuming that the number of sample features is a, select k features in a for n samples, and obtain the optimal segmentation point by establishing a decision tree;

(Ⅲ) Repeat m times to generate m decision trees $\{h_1(X), h_2(X), h_3(X)\cdots\}$;

(Ⅳ) Use Bagging's strategy to make predictions:

$$H(X) = \arg\max_Y \sum_{i=1}^{k} I(h_i(x) = Y).$$

3 Results and discussion

3.1 Multivariate linear regression analysis

3.1.1 Results of multivariate linear regression model of the spores number of stored rice

During the data analysis, those experimental data were imported into Matlab to draw the scatter plots of spore number along with ambient temperature, rice moisture content, and storage days. The related outputs can be seen in Fig. 1. To simplify calculation, define y as "spore number" and$y = N \times 10^{-5}$, where N represent the number of spores.

After the regression analysis, the regression coefficient vector $\widehat{B} = \left[\widehat{b_0}\ \widehat{b}\ \widehat{b_2}\ \widehat{b_3}\right]^{\mathrm{T}} = [-377.35\ \ 3.44\ \ 19.70\ \ 0.30]^{\mathrm{T}}$can be obtained, and the multivariate linear regression model between the spore number and ambient temperature, rice moisture content and storage days can be expressed by

$$y = -377.35 + 3.44x_1 + 19.70x_2 + 0.30x_3 \tag{2}$$

where y is the spore number; x_1 is the ambient temperature; x_2 is the moisture content of the rice samples; and x_3 is the storage days.

The total sum of squares T = 8535399.78, the regression sum of squares U = 2164591.45, and the residual sum of squares $Q = 6370808.33, F = \frac{U/n}{Q/(m-n-1)} = 121.86, F_{0.05}(3,1076) =$ 2.61.

It is also known that $F > F_{0.05}(3,1076)$, so it can be considered that there is a multivariate linear relationship between the spore number and ambient temperature, rice moisture content and storage days. But the multiple correlation coefficient $R = \sqrt{U/T} = 0.5036$, $R^2 = 0.2536$, which indicates that this linear relationship is not that good and further analysis is needed.

From the scatter plot shown in Fig. 1, it can be seen that the spore number is approximately exponential with the ambient temperature, the rice moisture content, and the storage days. Therefore, the multivariate linear regression could be performed after taking a natural logarithm of the spore number. The scatter plot after taking the natural logarithm is shown in Fig. 2.

Based on the multivariate linear regression, the regression coefficient vector $\widehat{B} = \left[\widehat{b_1}\ \ \widehat{b_2}\ \ \widehat{b_3}\ \ \widehat{b_4}\right]^{\mathrm{T}} = [-32.50\ \ \ 0.20\ \ 1.65\ \ 0.03]^{\mathrm{T}}$ can be obtained accordingly, and the multivariate linear regression model is given as

$$y = -32.50 + 0.20x_1 + 1.65x_2 + 0.03x_3 \tag{3}$$

where y is the spore number after taking the natural logarithm.

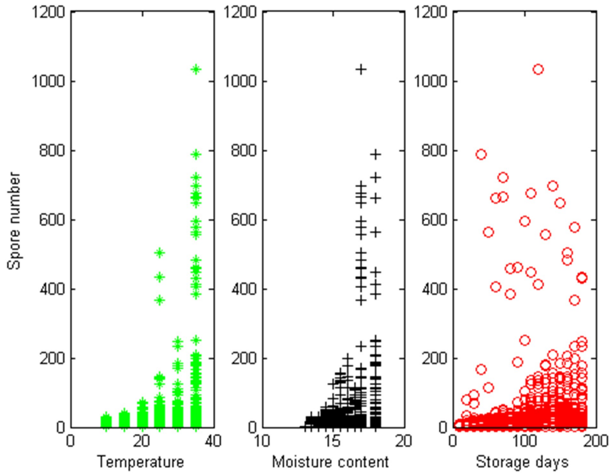


Fig. 1. Spore number along with ambient temperature, rice moisture content and storage days.

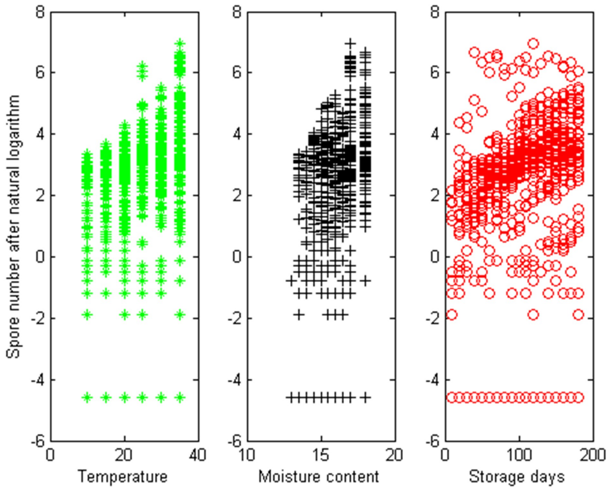


Fig. 2. Scatter plot after taking a natural logarithm of the spore number.

***F*-test** After taking natural logarithm, the total sum of squares $T = 15833.98$, the regression sum of squares $U = 12044.03$, and the residual sum of squares Q = 3789.95. And it is also obtained that $F=\frac{U/n}{Q/(m-n-1)}=1139.80, F_{0.05}(3, 1076)=2.61$.

It is also known that $F > F_{0.05}(3,1076)$, and the multiple correlation coefficient $R = \sqrt{U/T} = 0.8721$, $R^2 = 0.7606$. The results indicated that there is an ideal multivariate linear relationship between the spore number and ambient temperature, rice moisture content and storage days after taking the natural logarithm.

***T*-test** According to $t_i = \frac{\widehat{b_i}/\sqrt{c_{ii}}}{Q/(m-n-1)}(i = 1, 2, \ldots, n)$, the statistics variables t_1, t_2, t_3 corresponding to ambient tempera the ture, trice moisture content, and storage days are obtained as $[t_1\ t_2\ t_3]^{\mathrm{T}}$ = [30.36 43.91 23.86]$^{\mathrm{T}}$, $|t_1|$, $|t_2|$, $|t_3|$ are all greater than $t_{0.025}(1076)$ = 1.96. Under the circumstance, it is indicated that all of these three independent variables show important effect on the dependent variable, namely the spore number after taking the natural logarithm.

Residual analysis Take the residual as the ordinate and the case number of observed values as the abscissa to plot the residual analysis, the outputs can be seen in Fig. 3.

Green lines in Fig. 3 show the confidence interval of the residual crossing zero. Red lines are the confidence interval of the residual, which do not cross zero, namely the abnormal point. The original data corresponding to these abnormal points are the observations obtained by using different spore counting standard during the experiments, and those show significant differences with the others. After eliminating these abnormal points, the multivariate linear regression is re-performed on the remaining data, and we can get the new regression coefficients vector $\widehat{B}=\left[\widehat{b_1}\ \widehat{b_2}\ \widehat{b_3}\ \widehat{b_4}\right]=[-31.39\ 0.22\ 1.57\ \ 0.03]$. The new multivariate linear regression model can be given as

$$y=-31.39+0.22x_1+1.57x_2+0.03x_3 \tag{4}$$

It is known that F=1654.73, R=0.9155, R^2=0.8381, $[t_1\ \ t_2\ \ t_3]^{\mathrm{T}}$ = [39.07 48.47 28.20]$^{\mathrm{T}}$. All the above values have been improved, so the regression model becomes more accurate. From the regression coefficients, it can be seen that the order of these three independent variables' influences on the dependent variable can be concluded as $x_2>x_1>x_3$. In other words, the effect of rice moisture content is the biggest, ambient temperature is the second, and storage days show the least impact on the spore number of rice.

3.1.2 Conclusions of multivariate linear regression model of the spores number of stored rice

Based on a long-term experimental test on rice storage, the multivariate linear regression model was developed between the spore number and the main influencing factors, such as ambient temperature, rice moisture content, and storage days. Several conclusions can be addressed:

(Ⅰ) Multivariate linear regression based on the natural logarithm shows its advantage on the spore number of rice during the storage. During the process, both the statistical variable F and the multiple correlation coefficient R are significantly improved;

(Ⅱ) It is known from the T-test that all of the three explan-

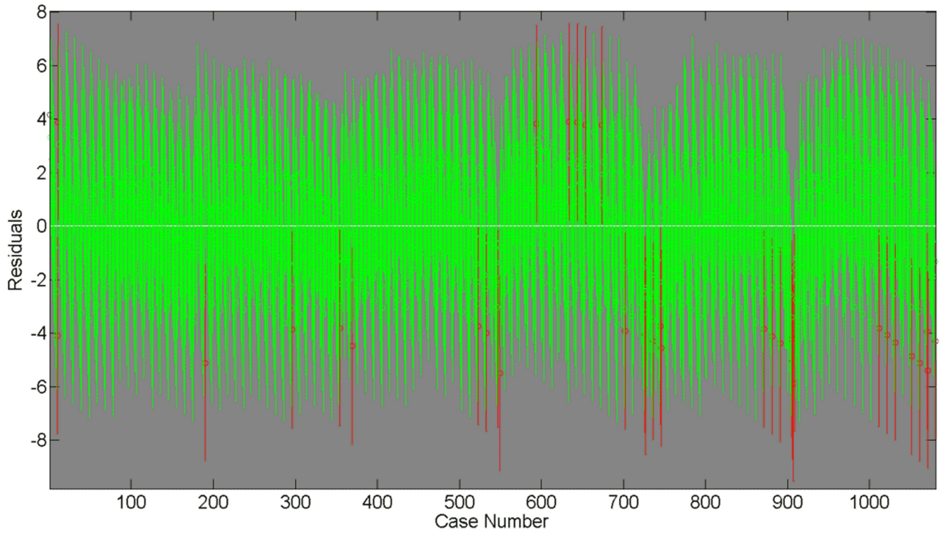


Fig. 3. Residual case order plot.

DOI: 10.52396/JUSTC-2021-0118

atory variables (e.g. ambient temperature, rice moisture content and storage days) play important role in the spore number of rice during its storage. A multivariate linear regression model was also developed to address the mechanism, which indicates that the growth of fungal spores is exponentially related to ambient temperature, rice moisture content, and storage days;

(Ⅲ) Through the residual analysis, the improved multivariate linear regression model y=−31.39+0.22x_1+1.57x_2+ 0.03x_3 is obtained after eliminating the abnormal data, where y, x_1, x_2, x_3 are the spore number after taking a natural logarithm, ambient temperature, rice moisture content and storage days respectively. Statistics variable F = 1654.73, the multiple correlation coefficient R = 0.9155, R^2 = 0.8381, the statistics variables t_1, t_2, t_3 corresponding to ambient temperature, rice moisture content, and storage days, can be obtained as 39.07, 48.47, 28.20;

(Ⅳ) The multivariate linear regression model obtained above can predict the spore number of rice under different ambient temperature, rice moisture content and storage days during the storage process, so as to judge whether mildew could occur in rice;

(Ⅴ) It can be seen from the regression coefficients that the number of fungal spores is most closely related to the moisture content of rice, followed by ambient temperature, and then the storage days.

3.2 Prediction model of the spores number of stored rice based on random forest algorithm

3.2.1 Data processing

The experimental data is the same as the second chapter, containing the number of spores and the three main affecting factors including ambient temperature, rice moisture content, and storage time. For the number of spores (N) in the original data ploted in Fig. 4 a shows no apparent distribution, we plot logarithm spore number data for easy analysis. Since many sample raw data is 0.01, after taking the logarithm, only the data of the part of x>0 is concerned, and it can be observed that this part of the data basically conforms to the normal distribution.

3.2.2 Model construction

For our issue, the goal of the training algorithm is to reduce the regression error.The model construction containing the following procedures:

(Ⅰ) From the training data of n samples, draw a bootstrap sample.

(Ⅱ) For each bootstrap sample, tree grows by choosing the split at each node by minimizing the cost function: $J = \sum_{j=1}^{J}\sum_{i\in R_j}|y_i-\hat{y}_{R_j}|$, where $\hat{y}_{R_j}$ is the average prediction value of the training samples in the chosen district R_j. We choose mean absolute error not mean square error because it is not sensitive to outliers.

(Ⅲ) A single node is formed by dividing continuous or discrete attributes (for example, temperature > 20), and a tree is generated by different nodes as a model to determine the final regression result. The focus point of the algorithm is to determine whether a node needs to continue to increase, in other words, to judge the gain of a node to the entire model. The node will stop to split when the current tree depth or the node size reaches the maximum limit valued or when splitting the node could not reduce the value of the cost function.

(Ⅳ) Repeat the above steps until sufficient trees from different random training data are obtained

(Ⅴ) For the given condition,the average value of different trees is finally used as the final prediction result.

3.2.3 Out of bag (OOB) score and mean absolute error

Out of bag (OOB) score is a way of validating the random forest model. After creating the classifiers (S trees), for each (x_i, y_i) in the original training set i.e. (T), select all (T_i) which does not include (x_i, y_i). This set is called out-of-bag examples. There are n such subsets (one for each data record in original dataset T). OOB classifier is the aggregation of votes only over (T_k) such that it does not contain (x_i, y_i). Out-of-bag estimate for the generalization error is the error rate of the out-of-bag classifier on the training set (compare it with known y_i)[23].

As shown in Fig. 5, the OOB error rate decrease dramatically as the tree number increase from 1 to 6. After the number of trees reaches 10, the error rate does not decline significantly. We also calculate the mean absolute error (MAE =

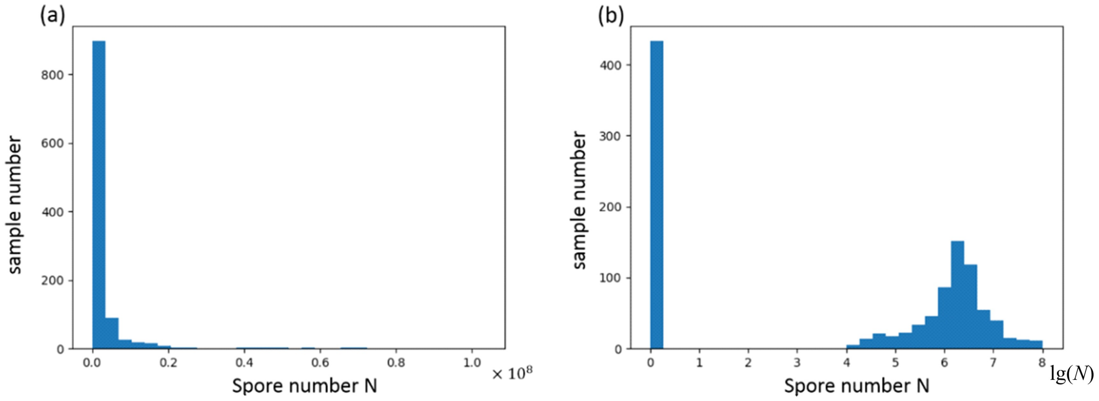


Fig. 4. The distribution of spore number.

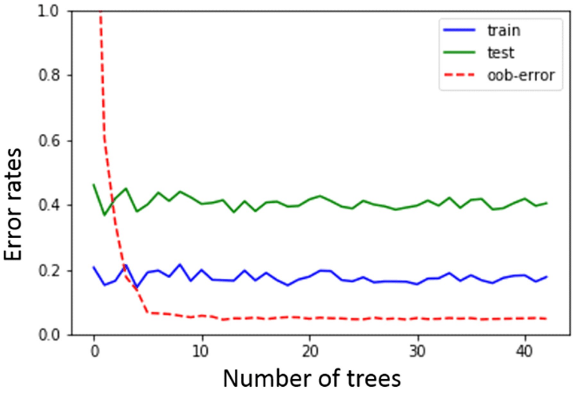


Fig. 5. Comparison of training, out-of-bag, and independent test set error rates for random forest as the number of trees increases.

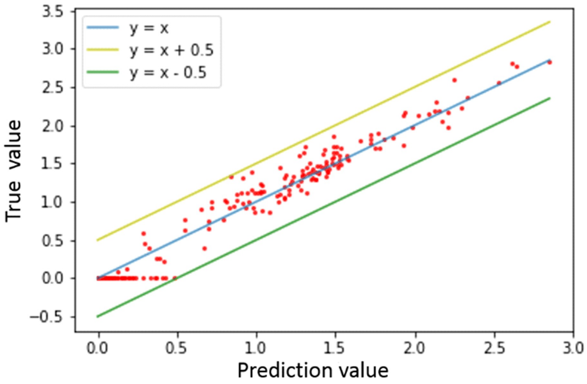


Fig. 7. The Scatter plot of prediction value and true value.

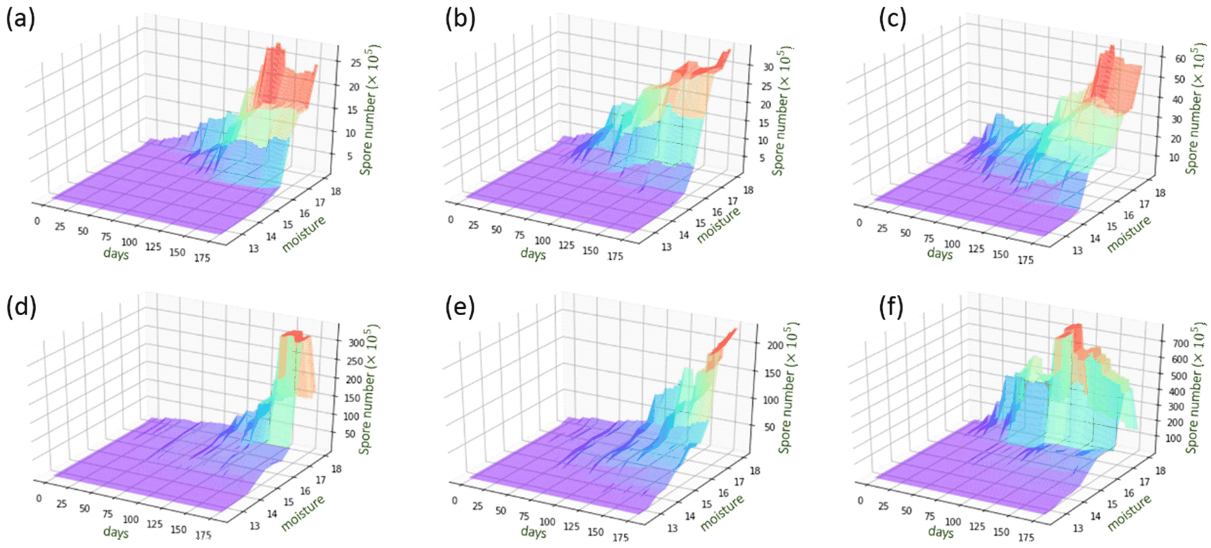


Fig. 6. Prediction surface graph in different temperature of 10 ℃ (a), 15 ℃ (b), 20 ℃ (c), 25 ℃ (d), 30 ℃ (e), 35 ℃(f).

$\frac{1}{n}\sum_{i=1}^{n}|(y_i-\widehat{y_i})|)$ of the model for both training data and test data, which fluctuate around 0.2 and 0.4 respectively.

3.2.4 R^2 score of the multiple linear regression model and random forest model

In order to compare the advantages and disadvantages of the random forest model and the multiple linear regression method, here we use R^2 score to evaluate the models.

$$R^2(y,\widehat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i-\widehat{y_i})^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i-\overline{y})^2}.$$

$y_1, y_2, \ldots, y_n$ is the real value of the samples, $\widehat{y_1}, \widehat{y_2}, \cdots, \widehat{y_n}$ is the predicted value, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and the R^2 score of the two models is shown in Table 1.

The R^2 score for test data of RF model is 0.95, which is much higher than the MLR model, thus it prove that the random forest prediction model can play a better role in predicting the number of mold spores grown in rice.

3.2.5 The prediction model

Fig. 6 displays the variation of spore number with temperature, moisture and storage days intuitively. The number of spores grows with the addition of moisture and the extension of storage time. As shown in the figure, the moisture and storage days in the purple district means the grain would not mildew and it is a safe district for grain storage. Therefore, it is significant to control the storage environment within the conditions in the low risk region.

The Fig. 7 was plotted by using the predicted value as the abscissa and the actual value as the ordinate. It can be found in the figure that the scatter points basically fall between the

Table 1. The R^2 score of the linear regression model and random forest model.

Model	Training data	Test data
Random forest	0.99	0.95
Linear regression	0.71	0.68

DOI: 10.52396/JUSTC-2021-0118
JUSTC, **2022**, 52(1): 6

two lines of $y = x - 0.5$ and $y = x + 0.5$, which suggest that the prediction model is accurate. After conversion, it was found that for 99% of the original data, the predicted value could be controlled to be of the same order of magnitude as the actual value. All the analysis above suggest that the random forest model have a high prediction accuracy.

3.2.6 The feature importance

Random forests can calculate the importance of each feature, or its contribution to the decision tree. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. We calculate the importance for the moisture, temperature and storage time by the following formula:

$$f_{i_i} = \frac{\sum_{j:\text{ node } j \text{ splits on feature } i} n_{ij}}{\sum_{k\in\text{all nodes}} n_{ik}}.$$

And then these can be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values.

$$\text{norm } f_{i_i} = \frac{f_{i_i}}{\sum_{j\in\text{all nodes}} f_{ij}}.$$

The final feature importance, at the Random Forest level, is the average of the feature's importance value on each trees. As a result, the importance for the three features of moisture, temperature and storage time is calculated to be 0.539, 0.246 and 0.215, respectively. Therefore, for the rice mildew, moisture is the most important factor, followed by temperature, again storage time, which is consistent with the result of the multivariate linear regression analysis.

4 Conclusions

In this work,based on the experimental data of rice storage experiments, a multivariate linear regression model and random forest model were developed between spore number of rice and main influencing factors, containing ambient temperature, rice moisture content, and storage days.

For multivariate linear regression analysis, through the residual analysis, the improved multivariate linear regression model has been developed and it furtherly prove that moisture is the most important factor, followed by temperature, again storage time for the rice mildew.For the random forest model, it could control the predicted value to be of the same order of magnitude as the actual value for 99% of the original data, which have a high accuracy to predict the spore number under different ambient temperature, rice moisture content and storage days during the storage process. Furthermore, we plot the prediction surface graph to help practitioner to control the storage environment within the conditions in the low risk region. Therefore, we believe that our results can be applied in practice and reduce loss during grain storage.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2017YFC0805903).

Conflict of interest

Yuhong Li

Biographies

Yurui Deng PhD, research interest: Process safety.

References

[1] Cheng S F, Tang F, Wu S L. Study on the early detection method of stored grain fungus damage. *Journal of the Chinese Cereals and Oils Association,* **2011**, *26* (4): 85–88.

[2] Yin W S, Zhang Y D. A survey of paddy fungus flora in China and some researches in it' s evolutional laws. *Journal of Zhengzhou Grain College*, **1986** (3): 3−17. https://en.cnki.com.cn/Article_en/CJFDTotal-ZZLS198603002.htm.

[3] Purushtham S P, Shetty H S. Storage fungal invasion and deterioration of nutritional quality of rice. *Mycol Pl Pathol,* **2010**, *40* (4): 581–585.

[4] Adriana L, Zoe M. Distribution of microbial contamination within cereal grains. *Journal of Food Engineering,* **2006**, *72* (4): 332–338.

[5] Genkawa T, Uchino T. Development of a low-moisture-content storage system for brown rice: Storability at decreased moisture contents. *Biosystems Engineering,* **2008**, *99* (4): 515–522.

[6] Soponronnarit S, Chiawwet M. Comparative study of physicochemical properties of accelerated and naturally aged rice. *Journal of Food Engineering,* **2008**, *85* (2): 268–276.

[7] Zhou J X, Ju X R. Succession of mould flora for paddy in different storage conditions. *Journal of the Chinese Cereals and Oils Association*, **2008**, 23 (5): 133−136(Chinese). http://cqvip.53yu.com/qk/96663x/200805/28325464.html.

[8] Zhou J X, Zhang R. Temperature influence on microorganism flora and fatty acid value of stored paddy under high humidity. *Journal of the Chinese Cereals and Oils Association*, **2011**, 26(1): 92−95 (Chinese).https://en.cnki.com.cn/Article_en/CJFDTotal-ZLYX201101022.htm.

[9] Zhou J, Shi X Z, Du K, et al. Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel. *International Journal of Geomechanics,* **2017**, *17* (6): 04016129.

[10] Zhou J, Asteris P G, Armaghani D J, et al. Prediction of ground vibration induced by blasting operations through the use of the Bayesian network and random forest models. *Soil Dynamics and Earthquake Engineering,* **2020**, *139*: 106390.

[11] Qiu Y, Zhou J, Khandelwal M, et al. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. Engineering with Computers, 2021. https://doi.org/10.1007/s00366-021-01393-9.

[12] Breiman L. Random forests. *Machine Learning,* **2001**, *45*: 5–32.

[13] Semenick Doug C S. Tests and measurements. *National Strength and Conditioning Association Journal,* **1990**, *12* (1): 36–37.

[14] Svetnik V, Liaw A, Tong C, et al. Random forest: A classification and regression tool for compound classification and QSAR

modeling. *Journal of Chemical Information and Computer Sciences,* **2013**, *43* (6): 1947–1958.

[15] Oliveira S, Oehler F, San-Miguel-Ayanz J, et al. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. *Forest Ecology and Management,* **2012**, *275*: 117–129.

[16] Chen X R. Probability Theory and Mathematical Statistics. Hefei: University of Science and Technology of China Press, 2009: 281-325. https://www.taylorfrancis.com/books/mono/10.1201/9781482267761/probability-theory-mathematical-statistics-engineers-paolo-gatti.

[17] Breiman L, Friedman J, Stone C J, et al. Classification and Regression Trees. Belmont, USA: Wadsworth, 1984. https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-richard-olshen-charles-stone.

[18] Esmeir S, Markovitch S. Anytime learning of decision trees. *Journal of Machine Learning Research*, **2007,** *8*: 891-933. https://www.jmlr.org/papers/volume8/esmeir07a/esmeir07a.pdf.

[19] Yu Z, Shi X, Zhou J, et al. Effective assessment of blast-induced ground vibration using an optimized random forest model based on a Harris Hawks optimization algorithm. *Applied Sciences,* **2020**, *10* (4): 1403.

[20] Chandra B, Kuppili V B. Heterogeneous node split measure for decision tree construction. International Conference on Systems, Man, and Cybernetics. Anchorage, AK: IEEE, 2011: 872-877. https://ieeexploreieee.53yu.com/abstract/document/6083761..

[21] Zhou J, Qiu Y, Armaghani D J, et al. Predicting TBM penetration rate in hard rock condition: A comparative study among six XGB-based metaheuristic techniques. *Geoscience Frontiers,* **2021**, *12* (3): 101091.

[22] Yu Z, Shi X, Qiu X, et al. Effective assessment of blast-induced ground vibration using an optimized random forest model based on a Harris Hawks optimization algorithm. *Engineering Optimization,* **2021**, *53*: 1467–1482.

[23] Mitchell M W. Bias of the random forest out-of-bag (OOB) error for certain input parameters. *Open Journal of Statistics,* **2011**, *1* (3): 205–211.

DOI: 10.52396/JUSTC-2021-0118