Hindawi
Journal of Food Quality
Volume 2022, Article ID 7730427, 9 pages
https://doi.org/10.1155/2022/7730427

Research Article

Continuous Monitoring Analysis of Rice Quality in Southern China Based on Random Forest

Lin Lu,[1] **Bo Zeng**,[2] **Shihua Yang**,[1] **Mingxue Chen**,[1] **and Yonghong Yu**[1]

[1]*China National Rice Research Institute, Hangzhou 310006, China*
[2]*National Agricultural Technology Extension and Service Center, Beijing 100125, China*

Correspondence should be addressed to Lin Lu; luzi0522@163.com and Yonghong Yu; yongh_yu@hotmail.com

Received 17 September 2022; Revised 29 November 2022; Accepted 1 December 2022; Published 13 December 2022

Academic Editor: Shudong He

Rice quality has received more attention, so monitoring and analysis are of great significance to rice quality. General quality indexes of rice in southern China from 2011 to 2020 were determined, including processing quality (brown rice yield, milled rice recovery, head rice yield), appearance quality (grain length, length-width ratio, chalky rice percentage, chalkiness degree, transparency), and cooking quality (alkali spreading value, gel consistency, amylose). Principal component analysis was used to distinguish the regional quality of southern rice. The results showed that amylose and chalkiness were the main contributory quality indexes of rice in South China, the upper reaches of the Yangtze River, and the middle and lower reaches of the Yangtze River. In the past decade, the total high-quality rate of rice in the South has improved. The random forest was used to determine the important influence index of rice quality. The results showed that chalkiness degree, alkali spreading value, and gel consistency were important indexes affecting the quality of southern rice, and random forest could be used as an effective approach for continuous monitoring and analysis of rice quality.

1. Introduction

Rice is one of the main food crops, and more than half of the world's population takes it as their staple food. With the standard of living rising, people pay more attention to the quality of rice as well as the yield of rice. The improvement of rice quality has become the key to alleviating the contradiction between supply and demand, enhancing market competitiveness, developing the local economy, and increasing 'farmers' income.

Rice quality mainly includes processing quality, appearance quality, and cooking quality. In the rice processing quality, brown rice yield is the ratio of brown rice to the weight of the rice, milled rice recovery is the ratio of the milled rice to the weight of the rice, and head rice yield is one of the good qualities required for high-quality rice. Appearance quality depends on grain shape, chalkiness, and transparency [1, 2]. Chalkiness is the white and opaque part of the rice endosperm, which is caused by the change in light transmission due to the gap between starch grains in the endosperm. The chalkiness of rice is high, and its transparency is low, so its appearance quality is poor. Grain shape is closely related to the yield of rice and has a certain effect on the processing quality. Amylose content, alkali spreading value, and gel consistency are the main indexes of rice cooking quality. Cooking quality is the key to affecting the taste of rice, and it has become an important factor in meeting the consumption demand for high-quality rice and affecting the domestic and international rice markets.

Dozens of new varieties of rice are planted in southern China every year. Monitoring the conventional quality of cultivated rice is an important way to study the quality of rice, and it needs to find a new method to analyze the monitoring data. In view of the excellent classification accuracy and processing efficiency, the random forest algorithm is becoming more widely used [3–5]. Li et al. [6] applied the random forest to the recommendation system and proposed a multidimensional context-aware recommendation method based on the improved random forest algorithm. The results showed that it could reduce the

average absolute error and root mean square error. Jin et al. [7] used the random forest algorithm to identify rice varieties. de Santana et al. [8] used random forest and infrared spectra to detect food adulteration. This paper intends to monitor and analyze rice quality in southern China by using the selection and ranking abilities of random forests.

2. Materials and Methods

2.1. Experimental Materials. The rice in this paper was all from the southern rice region of China, including South China, the upper reaches of the Yangtze River, and the middle and lower reaches of the Yangtze River. Rice includes indica rice and japonica rice and can also be divided into early rice, semilate rice, and late rice. The number of monitored rice varieties from 2011 to 2020 was 261, 269, 271, 307, 304, 307, 293, 284, 269, and 241, respectively.

2.2. Rice Quality Determination. According to the agricultural industry standard of China, NY/T 83-2017, 140 g of rice were taken and hulled into brown rice with a rice huller (Model THU35B, Japan), then the brown rice was milled into fine rice with a rice milling machine (Model 7132, China), and brown rice yield and milled rice recovery were calculated by weighing.

Head rice yield, grain length, length-width ratio, chalky rice percentage, chalkiness degree, and transparency were determined according to the agricultural industry standard NY/T 2334-2013 using the appearance tester and analysis software. The alkali spreading value was analyzed according to NY/T 83-2017. Several completely milled rice grains were added to a alkaline solution, and after constant temperature incubation, the digestion of rice grains was observed one by one and the classification was judged.

An appropriate amount of milled rice was taken and ground into rice flour by cyclone grinding (Foss Tecator, Sweden). Then rice flour was passed through a 0.15 mm sieve for the determination of amylose and gel consistency according to the methods of Lu and Zhu [9].

2.3. Analysis Algorithm

2.3.1. Principal Component Analysis. Principal component analysis (PCA) projects the original data into the simplified hyperspace defined by the principal components, which are linear combinations of the original variables. The first principal component has the largest variance, the second principal component has the second-largest variance, and so on. The multidimensional data is thinned into low-dimensional approximations, and the interpretation of the data by the first two or three principal components in two or three dimensions is simplified. Therefore, PCA can reduce the dimension of data and retain as much effective information as possible [10]. The specific calculation steps are as follows:

(1) Standardization of raw data

There are m original data: $X_1, X_2, \ldots, X_m$, which are converted into standardized values of $x_1, x_2, \ldots, x_m$.

$$x_i = \frac{X_i - \overline{X}_i}{s_i} \ (i = 1, 2, \ldots, m), \tag{1}$$

where $\overline{X}_i$ is the sample mean; s_i is the standard deviation.

(2) Calculating correlation coefficient matrix

$$R = \left(r_{ij}\right)m \times m,$$
$$r_{ij} = \frac{\sum_{k=1}^{n} x_{ki}x_{kj}}{n-1} \ (i, j = 1, 2, \ldots, m), \tag{2}$$

where $r_{ij} = r_{ji}, r_{ij}$ is the correlation coefficient between the ith variable and the jth variable.

(3) Calculating eigenvalues and eigenvectors

Solving characteristic equation $|\lambda I - R| = 0$, and finding the eigenvalue of the correlation coefficient matrix R, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$, the corresponding eigenvector is $e_1, e_2, \ldots, e_m$, where $e_j = (e_{1j}, e_{2j}, \ldots, e_{nj})^T$, and m new index variables are composed of the following eigenvectors:

$$\begin{cases} y_1 = e_{11}x_1 + e_{21}x_2 + \cdots + e_{n1}x_n \\ y_2 = e_{12}x_1 + e_{22}x_2 + \cdots + e_{n2}x_n \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m = e_{1m}x_1 + e_{2m}x_2 + \cdots + e_{nm}x_n \end{cases}, \tag{3}$$

where y_m is the mth principal component.

(4) Calculating the principal component contribution rate and cumulative contribution rate:

Contribution rate of the principal component y_j

$$\alpha_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} \ (j = 1, 2, \cdots, m). \tag{4}$$

Cumulative contribution rate of the principal components $y_1, y_2, \ldots, y_w$

$$\beta_w = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}. \tag{5}$$

When βw is close to 1, the first w principal components can replace m original data.

(5) Obtaining the principal component score

$$Z = \sum_{j=1}^{w} \alpha_j y_j, \tag{6}$$

where α_j is the contribution rate of the jth principal component y_j.

In this study, the rice quality indexes were used as the raw data of PCA, and the scores of principal component 1 (PC1), principal component 2 (PC2), and principal component 3 (PC3) and the corresponding contribution rates were acquired as well as the load matrix.

2.3.2. Random Forest. The random forest algorithm clearly shows the processing of forming a forest composed of multiple decision trees in a random manner, which is a machine learning algorithm [11]. When an unknown sample enters the constructed forest as the input, each decision tree in the forest will judge separately to identify the category to which the sample belongs, and then predict this sample as the category that is judged the most.

The construction process of a random forest is as follows: first, the subdataset is constructed. Random sampling of samples is carried out from the original data set through the sampling method with some samples being put back. Second, a decision tree is constructed using subdataset. Suppose a subset has x attributes. When each node of the decision tree needs to be split, y attributes are randomly selected from these attributes. Also, select one of the Y attributes as the split attribute of the node in some way. Repeat this step until it can no longer split. Following the abovementioned two steps, a large number of subdecision trees that will form a random forest are built. Finally, the dataset is input into different subdecision trees, and then different judgment results can be obtained. The result which is judged the most is the best classification scheme of random forest.

The most commonly used strategy is absolute majority voting. Assuming that the set of categories is $\{C_1, C_2, \cdots, C_n\}$, the predicted output of G_i on sample x is expressed as an n-dimensional vector $(G_i^1(x), G_i^2(x), \cdots, G_i^n(x))^T$, where $h_i^j(x)$ represents the output of G_i on category C_j.

$$G(x) = \begin{cases} C_j, & \sum_{i=1}^{T} G_i^j(x) > 0.5 \sum_{k=1}^{n} \sum_{i=1}^{T} G_i^k(x), \\ \text{reject}, & \text{else}. \end{cases} \tag{7}$$

The selection of the optimal parameters is the premise for obtaining optimal results. In this study, random forest was used to rank and analyze the important influence degree of variables.

3. Results and Discussion

3.1. Analysis of Quality of Rice Varieties in Southern China. Monitoring of rice regular quality has been carried out on rice varieties in southern China for ten consecutive years, including brown rice yield (BRY), milled rice recovery (MRR), head rice yield (HRY), grain length (GL), length-width ratio (LWR), chalky rice percentage (CRP), chalkiness degree (CD), translucency (TC), alkali spreading value (ASV), gel consistency (GC), and amylose (AS). There is mainly indica rice in the southern rice region, but there is still about 5% japonica rice. BRY and MRR of indica rice were slightly lower than those of japonica rice, while HRY was much lower. Therefore, the processing quality of indica rice was overall lower than that of japonica rice. As for appearance, the GL of indica rice is larger than that of japonica rice. There was no significant difference in CRP, CD, and TC between the two. There was no significant difference in GC, but other cooking qualities were slightly different. The ASV of indica rice was lower than that of japonica rice, while AS was the opposite. It can be seen from the trend of quality indexes in the past ten years (Figure 1) that CRP has the largest change, which is decreasing year by year, and the decrease of indica rice is greater than that of japonica rice. In addition, CDs for the two were all decreased, which was consistent with the result of the previous report [12]. In the past ten years, HRY of indica rice first decreased and then increased, reaching the highest value of 61.9% in 2019, while HRY of japonica rice decreased as a whole.

The processing qualities of early, semilate, and late rice were different; the difference in MRR was the smallest, but the difference in MRR is larger. The order of HRY was late rice > semilate rice > early rice, and the order of BRY was late rice > early rice > semilate rice. The difference in length-width ratio among early, semilate, and late rice was small, but the GL of semilate rice was slightly longer than that of early and late rice. The appearance qualities, namely CRP, GC, and TC, of the three kinds of rice all appeared in the same order: late rice > semilate rice > early rice. As for cooking quality, the ASV of semilate and late rice was higher than that of early rice, the GC of semilate rice is slightly higher than that of early and late rice, and the AS of early rice is more than 18%, which is generally higher than that of semilate and late rice. As seen in Figure 1, the chalkiness of three rice showed a decreasing trend, and CRP of early rice decreased the most. GC of late rice increased obviously.

3.2. Quality Differentiation Analysis of Southern Regions. The rice region in southern China can be divided into several subrice regions, including South China, the upper reaches of the Yangtze River, and the middle and lower reaches of the Yangtze River. PCA was used to distinguish and analyze rice quality in different regions in the past ten years. As an important dimension reduction analysis method in multivariate statistical analysis, PCA transforms highly correlated variables into mutually independent or uncorrelated variables, whose main purpose is to use fewer variables, i.e., principal components, to explain the comprehensive indicators of the original variables.

BRY and MRR of rice in South China and the middle and lower reaches of the Yangtze River were slightly higher than those in the upper reaches of the Yangtze River. In terms of HRY, those in the upper reaches of the Yangtze River were higher in the first five years and decreased in the latter five years, while those in the middle and lower reaches of the Yangtze River were relatively high in the last five years. CRP and CD of rice in the upper reaches of the Yangtze River were higher, indicating that the appearance quality of rice in this area was relatively low. From 2011 to 2016, ASV in South China was very low and then increased, reaching the average value in the upper reaches, and the middle and lower reaches of the Yangtze River. GC and AS in the upper reaches of the Yangtze River were higher than those in South China and the middle and lower reaches of the Yangtze River.

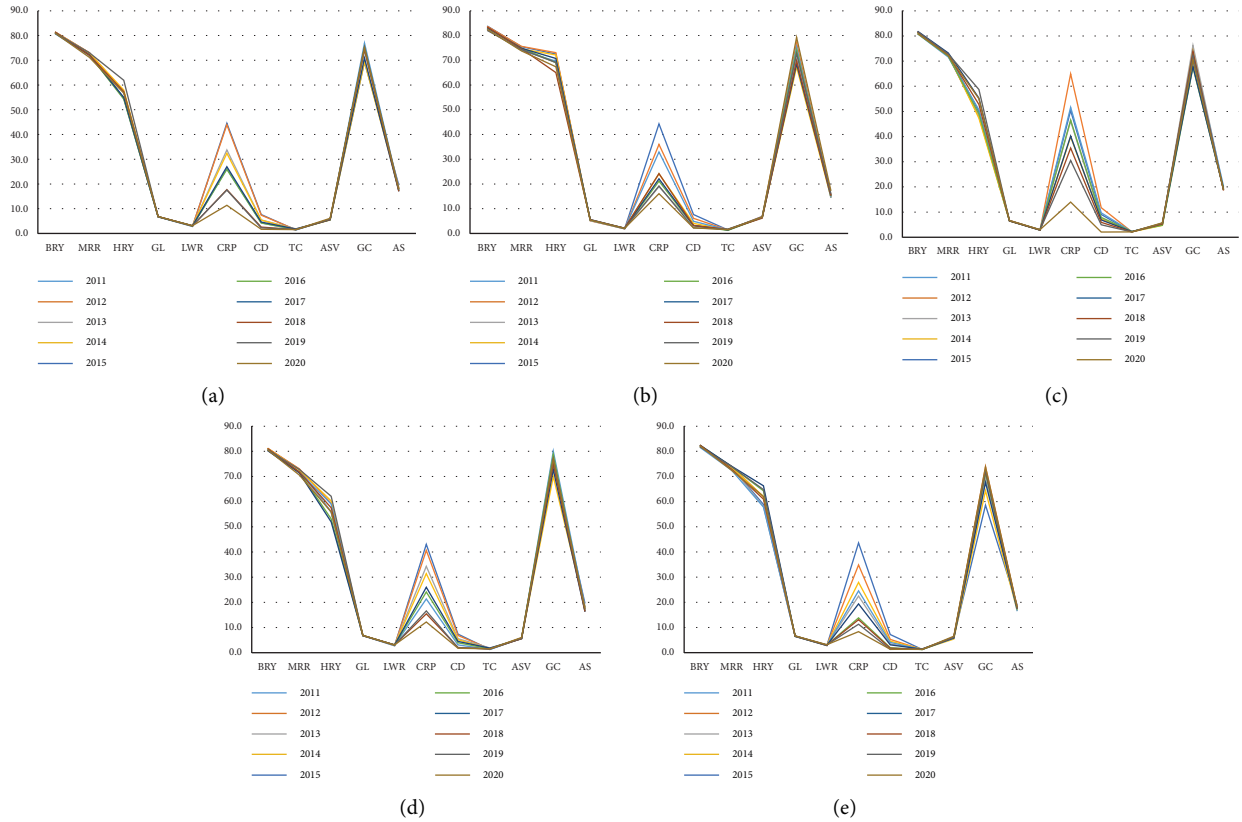

Figure 1: Rice quality indexes of indica rice (a), japonica rice (b), early rice (c), semilate rice (d), and late rice (e) in 2011–2020.

From the PCA chart of the three regions (Figure 2(a)), it could be seen that the three-dimensional points in the upper reaches of the Yangtze River were completely different from those in South China and the middle and lower reaches of the Yangtze River, while the three-dimensional points in South China and the middle and lower reaches of the Yangtze River partially overlapped, indicating that the overall quality of rice in the upper reaches of the Yangtze River could be significantly different from that in the other two regions. As seen in Figures 2(b) and 2(c), in the first five years, the three-dimensional points of the three regions were completely distinguished, while in the latter five years, the three-dimensional points of South China and the middle and lower reaches of the Yangtze River partially overlapped, indicating that the discrimination degree of rice quality in the two regions was reduced. The reason might be related to the popularity of rice varieties in the south, and the same or similar varieties were planted in different regions. According to the load matrix score in PCA, the contribution index of the overall quality of rice in the three regions could be judged. It was seen from Figure 3 that in PC1 of the three regions in the first five years, AS had the largest positive load, followed by ASV. For PC2, BRY had the largest positive load. In PC1 and PC2 of the latter five years, AS and ASV had corresponding maximum positive loads. This result was consistent with the result from the ten-year overall load matrix score chart. AS was the maximum positive load of PC1, and CRY and CD were the maximum positive loads of PC2, illustrating that amylose and chalkiness were the main contributory indexes to distinguish the quality of rice in the three regions.

3.3. Analysis of Influence Index of Rice Quality in Southern China. According to the annual average values of rice quality indexes in southern China over the past ten years (Figure 4), LWR had a small increase trend, but CRP and CD decreased year by year. From 2011 to 2020, the total high-quality rate (total HQR) of rice declined and then increased, exceeding 50% in 2018 and reaching the highest value of 56.4% in 2020. According to the previous reports on rice quality indexes [13, 14], correlation analysis, principal component analysis, and cluster analysis were generally used to identify the differences among rice indexes as common analysis methods, but they could not link the rice quality rate with the rice indexes. In this part, a random forest was utilized to obtain the link. The performances of the quality index and the total high-quality rate were shown in Figure 5 using random forest, and the importance of each index could be obtained to determine the most important impact of rice quality.

The parameters of the random forest were determined by the minimum variance, which was as follows: the method was "regression;" the number of decision trees was 10 or 20;

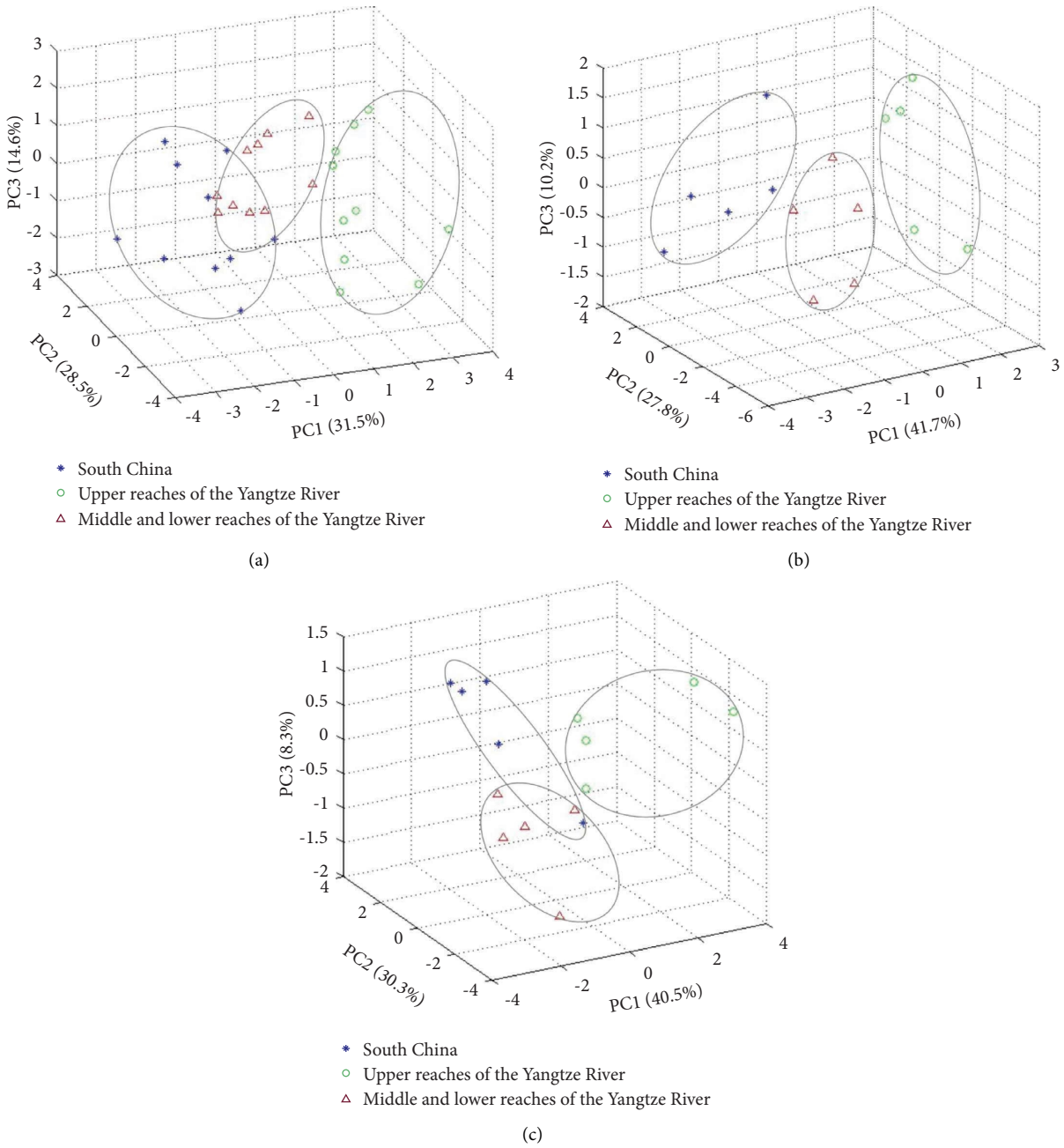

Figure 2: PCA chart of rice from the three regions, including South China, the upper reaches of the Yangtze River, and the middle and lower reaches of the Yangtze River. (a) Ten years; (b) the first five years; (c) the latter five years.

the minimum leaf node was 5; "Oobvarimp" and "surrogate" were both "on;" Fboot was 1. According to the results of the random forest, the rankings of HRY, CD, ASV, GC, and AS were relatively high. The abovementioned five indexes were input into the random forest again. The result showed that CD, ASV, and GC were more important. Therefore, the chalkiness degree, alkali spreading value, and gel consistency were important indexes affecting the quality of southern rice.

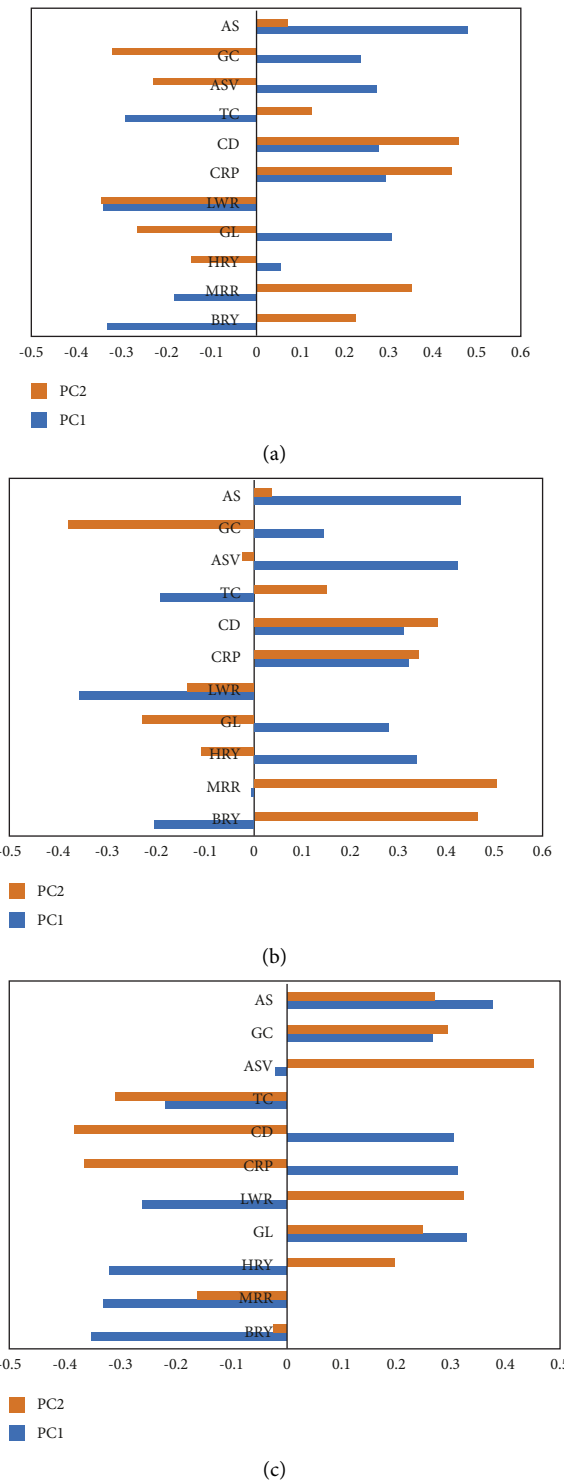

Figure 3: The load values of PCA for the regional quality of rice. (a) Ten years; (b) the first five years; (c) the latter five years.

Figure 4: Annual average values of rice quality indexes and the total high-quality rate (total HQR) in southern China from 2011 to 2020. (a) BRY (brown rice yield); (b) MRR (milled rice recovery); (c) HRY (head rice yield); (d) GL (grain length); (e) LWR (length-width ratio); (f) CRP (chalky rice percentage); (g) CD (chalkiness degree); (h) TC (translucency); (i) ASV (alkali spreading value); (j) GC (gel consistency); (k) AS (amylose).

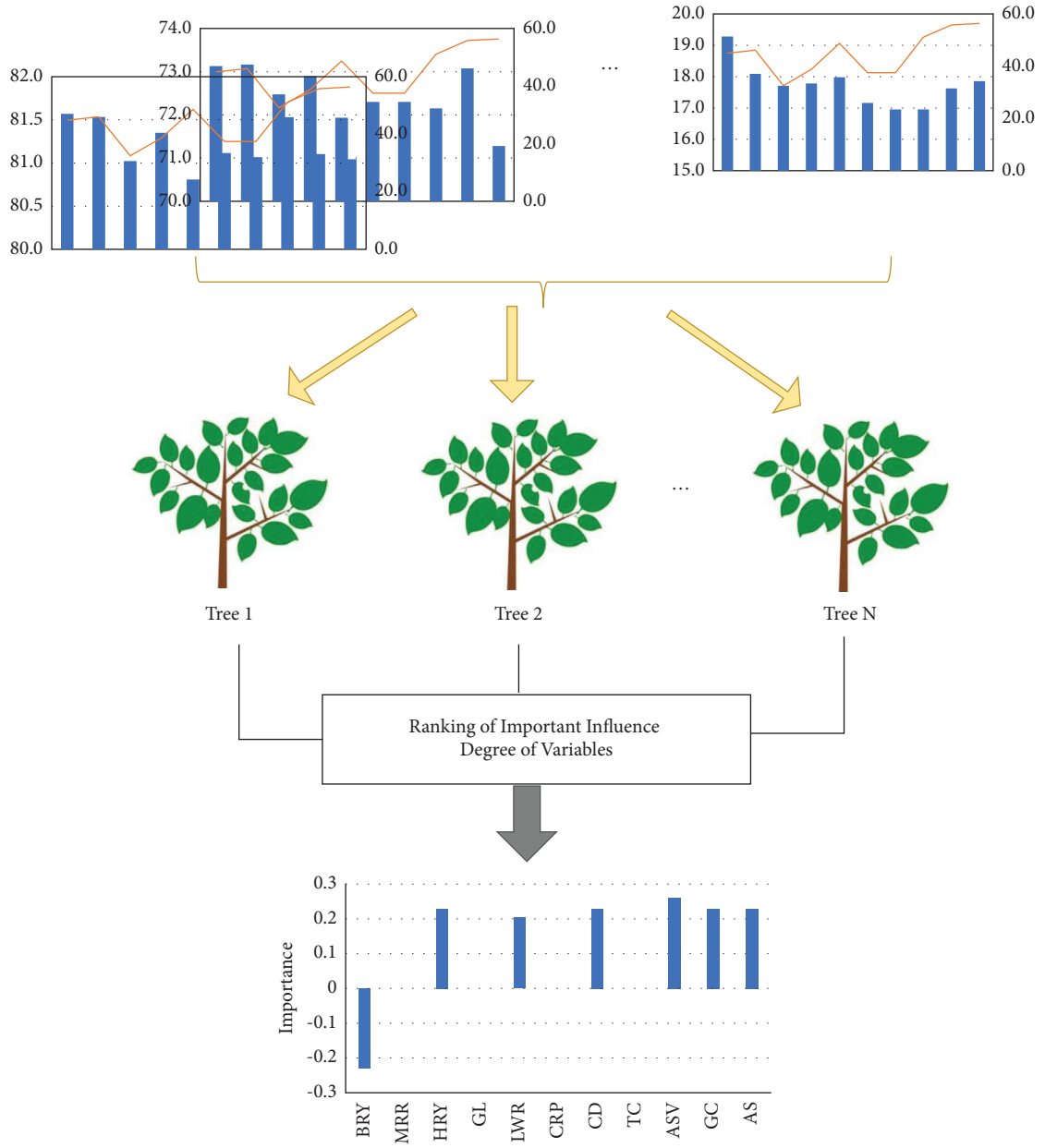

Figure 5: Schematic diagram of a random forest for continuous monitoring and analysis of rice quality.

4. Conclusion

Continuous monitoring was realized by analyzing the change in rice quality in southern China from 2011 to 2020. The processing quality of indica rice was lower than that of japonica rice. The processing qualities of early, semilate, and late rice were different, and their appearance qualities showed the same order. As for the cooking quality, the alkali spreading values of semilate and late rice were higher than those of early rice, the gel consistency of middle rice was slightly higher than that of early and late rice, and the amylose of early rice was generally higher than that of semilate and late rice. Principal component analysis was used to distinguish the regional quality of southern rice. The results showed that the overall quality of rice in the upper reaches of the Yangtze River was significantly different from that in South China and the middle and lower reaches of the Yangtze River. Amylose and chalkiness were the main contributory indexes to distinguish the rice quality in the three regions. In the past ten years, the total high-quality rate of rice in southern China has increased, reaching the highest value in 2020. The random forest was used to determine the important influence index of rice quality. The results showed that chalkiness degree, alkali spreading value, and gel consistency were important indexes affecting the quality of southern rice.

Data Availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Matsue, K. Takasaki, and J. Abe, "Water management for improvement of rice yield, appearance quality and palatability with high temperature during ripening period," *Rice Science*, vol. 28, no. 4, pp. 409–416, 2021.

[2] W. Cuili, G. Wen, H. Peisong, W. Xiangjin, T. Shaoqing, and J. Guiai, "Differences of physicochemical properties between chalky and translucent parts of rice grains," *Rice Science*, vol. 29, no. 6, pp. 577–588, 2022.

[3] M. Belgiu and L. Dragut, "Random forest in remote sensing: a review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[4] P. Du, A. Samat, B. Waske, S. Liu, and Z. Li, "Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, pp. 38–53, 2015.

[5] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for landcover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012.

[6] X. Li, Z. J. Wang, L. Y. Wang, R. L. Hu, and Q. Y. Zhu, "A multi-dimensionalcontext-aware recommendation approach based on improved random forest algorithm," *IEEE Access*, vol. 6, pp. 45071–45085, 2018.

[7] B. C. Jin, C. Zhang, L. Q. Jia et al., "Identification of rice seed varieties based on near-infrared hyperspectral imaging technology combined with deep learning," *ACS Omega*, vol. 7, no. 6, pp. 4735–4749, 2022.

[8] F. B. de Santana, W. Borges Neto, and R. J. Poppi, "Random forest as one-class classifier and infrared spectroscopy for food adulteration detection," *Food Chemistry*, vol. 293, pp. 323–332, 2019.

[9] L. Lu and Z. W. Zhu, "Prediction model for eating property of indica rice," *Journal of Food Quality*, vol. 37, no. 4, pp. 274–280, 2014.

[10] L. Lu, Z. Q. Hu, C. Y. Fang, X. Q. Hu, and S. Y. Tian, "Improvement on the identification and discrimination ability for rice of electronic tongue multi-sensor array based on information entropy," *Journal of the Electrochemical Society*, vol. 169, no. 3, Article ID 037524, 2022.

[11] L. Breiman, "Random forest, machine learning 45," *Journal of Clinical Microbiology*, vol. 2, pp. 199–228, 2001.

[12] B. Y. Qian, H. Y. Zhou, S. Q. Dai, M. Xu, and X. L. Wu, "Variation trend of chalkiness of indica rice in southern China rice regions and its improvement countermeasures," *Modern Agricultural Science and Technology*, vol. 16, pp. 20–27, 2022.

[13] A. Buhaliqem, J. Yuan, Y. H. Zhang et al., "Analysis of rice quality traits of different japonica rice varieties (lines)," *Xinjiang Agricultural Sciences*, vol. 59, pp. 1347–1355, 2022.

[14] X. J. He, J. C. Zhang, C. Yang, X. Q. Yu, Z. Song, and L. J. Zhou, "Quality Characteristics and population structure of glutinous rice landrace resources from Guizhou province," *Seeds*, vol. 41, pp. 68–73, 2022.