Target: Journal of Agricultural and Food Chemistry

**Determination of geographical origin by multi-elemental profiling combined with machine learning techniques: a study on Chinese geographical indication (GI) rice**

Fei Xu1, Fanzhou Kong1, Hong Peng1, Shuofei Dong2, Weiyu Gao1, Guangtao Zhang1\*

1Mars Global Food Safety Center, Beijing 101407, China

2Agilent Technologies (China) Co., Ltd, Beijing 100102, China

\*Corresponding author: Tel.: +86 13331152657. E-Mail: [Guangtao.zhang@effem.com](mailto:Guangtao.zhang@effem.com)

**ABSTRACT**

The demand for geographical indication (GI) rice has increased amongst Chinese consumers, which potentially results in a high risk of adulteration due to its high value and limited production. This study aims to develop a novel strategy of non-targeted data analysis to determine geographical origins of Chinese GI rice based on multi-elemental profiling obtained by inductively coupled plasma mass spectrometry (ICP-MS). 131 samples from six types of Chinese GI rice were analyzed. Coupled with feature selection (relief algorithm), two machine learning based classifier, support vector machines (SVM) and random forest (RF) were utilized to predict the origins of GI rice; the results were validated through repeated grid-search cross-validation. For both SVM and RF, four elements (Na, Al, Cd, and Rb) only could enable the prediction with 100% accuracy. These results demonstrate the feasibility of ICP-MS combined with machine learning techniques as an effective strategy for authentication of GI rice in China.

KEYWORDS

TBC

**INTRODUCTION**

In the current world of globalization, geographical authentication of food materials has emerged as an important issue, which is crucial for both quality assurance and food safety1. The notion of geographical indication (GI) first originated from the 19th century in Europe, with the purpose of protecting industrial property rights2. Nowadays, product with GI certification generally possess a given quality, reputation or other characteristics attributable to geographical origins3, thus making it possible for them to be differentiated from competitors’ (giving them edge over their competitors). According to the European Union (EU) quality schemes, three different systems have been enforced to protect the origin integrity of agroproducts and food: protected designation of origin (PDO), protected geographical indication (PGI) and traditional specialties guaranteed (TSG)4. Currently in China, State Administration for Industry and Commerce/ the Trademark Office (SAIC/TMO), the General Administration of Quality Supervision, Inspection and Quarantine (AQSIQ), and the Ministry of Agriculture (MoA) are supervising and protecting GIs from different aspects or direction, at the administrative level5.

Rice (*Oryza sativa L.)* is among the world’s top three largest food crops and is a staple food for nearly 50% of the world population. China is leading in the rice paddy production in the world, with 214.4 million metric tons in 20176. With the increasing living standard, there is a growing demand domestically for rice with traceable origins7. For example, Wuchang rice, one of GI rice produced in Heilongjiang Province in China’s northeast region, is known for its superior quality and unique sensory characteristics. Due to the gap between the limited production and high market demand, the price of Wuchang rice is roughly twice than that of other domestic rice, making it vulnerable to adulteration such as partial substitution and fraudulent labeling8,9. Therefore, determination of geographical origins of rice is of great importance to protect the rights of farmers, retailers and consumers3.

Since it is challenging to accurately distinguish rice from different origins via visual inspection, multiple “fingerprinting”-based approaches have been developed in recent years. For example, vibrational spectroscopy-based techniques such as Raman spectroscopy10 and near-infrared (NIR) spectroscopy 11,12 have been utilized to differentiate rice with different origins. Notably, multi-elemental profiling by inductively coupled plasma mass spectrometry (ICP-MS) 13–15 and stable isotope ratio analysis have already been successfully employed to authenticate the geographical origins of rice. As been summarized in a recent review, principle component analysis (PCA) combined with linear discriminate analysis (LDA) is by far the dominant strategy for rice authentication, thanks to its simplicity and ability in detecting hidden information in data16. Furthermore, there is an emerging trend of utilizing machine learning techniques to achieve the goal of rice discrimination. Support vector machines (SVM)14, decisions trees, random forest (RF) and artificial neural network (ANN)17 are the most popular ones found in recent studies.

The aim of this study was to determine the geographical origins of six different Chinese GI rice basing on their multi-elemental profiling obtained by ICP-MS. Particularly, a novel machine learning based workflow was implemented by combining feature selection (relief algorithm) and supervised classification (i.e. SVM/ RF algorithms).

**MATERIALS AND METHODS**

*Rice samples*

Proper sampling is fundamental for achieving reliable results from multivariate model building18. In this study, a total of one hundred and thirty-one Chinese GI rice samples were directly collected from paddy fields from five provinces, i.e. Heilongjiang, Liaoning, Jiangsu, Hubei and Guangxi. The name of each GI rice consists with two parts: geographical location and commodity name: Wuchang Wuyoudao (WC), Guigang Dongjinxi (GG), Panjin Yanfeng46 (PJ-1), Panjin Liaoxinyihao (PJ-2), Jingshan Qiaomi537 (JS), and Sheyang? (SY) (See Fig.1).

*Reagents and chemicals*

Nitric acid (HNO3, 69%, part# 100441) were purchased from Merck Millipore (Darmstadt, Germany). Deionized water (18.3MΩ cm) was obtained from a Milli-Q system (Millipore, MA, USA). Multi-element calibration standard 2A (part# 8500-6940), 4 (part# 8500-6942), Environmental calibration standard (part# 5183-4688) and Scandium standard (part# 5190-8578) were purchased from Agilent Technologies (Santa Clara, CA, USA).

*Sample digestion and ICP-MS analysis*

In order to minimize the impact of unexpected contamination, the Teflon digestion vessels were soaked in 30% (v/v) nitric solution for 24h, and then rinsed with deionized water for three times before use. For pre-digestion procedure, 0.5 g of rice grains was directly digested in duplicate in digestion vessels with 6mL of concentrated HNO3 and left in fume hood overnight. The vessels were then placed in microwave oven (Anton Paar, Austria) next day. The digestion procedure was programed so that the temperature will gradually reach 180 °C in 15 min, and then kept at such temperature for 20 min. The digestion was completed when the sample solution mixtures turned into clear liquid. All solutions were cooled down at room temperature and diluted into 50mL metal-free plastic tubes. An Agilent 7900 ICP-MS (Agilent technologies, Santa Clara, CA, USA) was utilized for multi-elemental profiling. The instrumental setting and operative conditions were adopted from a previous study19 with modifications (RF power of 1550 W, RF matching of 1.85 V, , and carrier gas flow rate of 1.05 L·min-1). The concentrations of 30 elements (10B, 23Na, 24Mg, 27Al, 39K, 43Ca, 45Sc, 48Ti, 51V, 52Cr, 55Mn, 56Fe, 59Co, 60Ni, 65Cu, 66Zn, 70Ga, 73Ge, 75As, 78Se, 85Rb, 86Sr, 93Nb, 98Mo, 107Ag, 114Cd, 133Cs, 138Ba, 201Hg, 208Pb) in rice was then detected. The internal standard solution of 103Rh (10 mg· L-1, part# 8500-6945) was also obtained from Agilent Technologies. One certified reference material of rice flour (1568b) from the National Institute of Standards and Technology (Gaithersburg, MD, USA) was used to verify the accuracy of the analysis method.

*Statistical analysis*

One-way analysis of variance (ANOVA) was carried out to access the statistically significant differences in the element contents of rice different GI rice (P ≤ 0.05). Raw data was first preprocessed by taking log -transformation and then, subjected to unsupervised PCA, as an initial step to uncover hidden information before classification. Two machine learning based techniques, RF and SVM were implemented for the training of classifiers with all six types of GI rice. RF was first introduced by Breiman20 and it is made of a collection of “tree”-structured classifiers (i.e. decision trees), which are generated from original dataset using bootstrap partition. The central idea of SVM is to project the input vectors into a high dimensional space, and find a hyperplane that could separate different classes21. In addition, radar plots analysis was also applied to … Moreover, in this study, a feature selection algorithm, ReliefF 24 was utilized to determine how each feature is contributing to the overall prediction accuracy. It works by assigning relative importance to features basing on a calculated proxy statistic25. However, as Damjan et al. stated, studies without correctly conducted cross-validation and feature selection are flawed in nature, we have designed our workflow carefully to avoid pitfalls stated in Damjan’s articles. First of all, The 131 samples were split into training set and testing set with ratio of 80%:20% (n = 104 in training set, n = 27 in testing set). ~~Currently, opposite views exist regarding when feature selection shall be conducted during the process of classifier training. While it is usually recognized as an important data preprocessing step to remove feature variables with low or null discriminating power for the samples~~~~14~~~~; there is also the concern that classification results could be severely biased, if selection of features is done prior to the cross-validation~~~~26~~~~.~~ The feature selection algorithm is applied only on the training dataset to avoid selection bias. In order to provide a fair assessment of the classifiers (i.e. RF and SVM) while minimizing bias, we implemented repeated grid-search cross-validation (RGSCV) for feature selection and ~~parameter tuning~~ classifier building23. For SVM, a linear kernel was chosen and the C values (Cost) were tested. As for RF, max number of levels in each decision tree (max\_depth), max number of features considered for splitting a node (max\_features) and number of trees in the forest (n\_estimators) were optimized. ~~However, since there is no clear guideline for selecting features after they were ranked, the selection is s usually arbitrary~~27 Given the relative low numbers of features in this study, we combined ranking technique with greedy search strategy adapted from previous studies28,29 See Figure XX for the pipeline for building of classifier and validation.

1. ~~Conduct Relief algorithm and rank all features basing on their relative importance;~~
2. ~~Construct 30 subsets of features as below: the 1~~~~st~~ ~~subset containing only the highest ranked feature, the 2~~~~nd~~ ~~subset containing top 2 highest ranked features. Similar practice will be repeated until the 30~~~~th~~ ~~subset, which ultimately contains all 30 features;~~
3. ~~Construct a 2-dimension matrix with feature subsets on one dimension and hyperparameters on the other;~~
4. ~~Repeat 10-fold cross-validation 10 time for all possible grids and record average accuracy for each grid;~~
5. ~~Choose the best hyperparameter-feature subset-combination in terms of average accuracy~~

All analyses were carried out in R (R Core Team, 2019), using RStudio (version 3.5.1, Boston, MA, USA) with additional packages: e107130, caret31, randomForest32, dplyr33, factoextra34, and FSelector34.

**RESULTS AND DISCUSSION**

*Elemental concentration in Chinese GI rice*

Table 1 shows the mean values and standard deviations of 30 targeted elements from six GI rice in this study. Particularly, ANOVA and Tukey HSD test were conducted to determine the statistical significance. The validation of accuracy was conducted with SRM (1568b), and measured concentration agreed well with the certified values.

Overall, except for 208Pb, significant differences could be observed among all elements. The GG rice, which are harvested from Guangxi Zhuang Autonomous region (southwest China, were leading in the levels of heavy metals such as 107Ag, 114Cd and 201Hg. A possible explanation for this is that since the pH in rice paddies variances from different regions in China (weakly alkaline in the north and weakly-acidic in the south), the bioavailability of heavy metal elements is generally higher in rice paddies grown in the south35. Besides, SY rice has the most abundance of macro elements such as 23Na and 39K. PJ-1 rice has significant higher levels of 27Al, 45Sc, 48Ti, 51V, 56Fe, 70Ga, 86Sr and 93Nb than others. PJ-1 and PJ-2, harvested in the geological location, have similar levels of 24Mg, 52Cr, 60Ni, 73Ge, 78Se, 114Cd, 133Cs, and 208Pb.

*Principle component analysis (PCA)*

In order to get an initial overview of the entire dataset, an unsupervised PCA after log-scaling of original dataset was conducted (95% confident ellipses included). As shown in Fig 2a, there was a clear separation pattern among PJ-1, GG and the rest of GI rice. while for JS, PJ-2, SY and WC, no satisfactory separation could be achieved based only on the 1st and 2nd principle component (PC). The loading plot (Fig 2c) showed that 27Al, 70Ga, 93Nb,51V, and 48Ti primarily contributed to the variations on PC1, while 23Na, 45Sc, 85Rb, 133Cs, and 114Cd contributed to both PC1 and PC2. For PJ-1 and PJ-2, even though from the same geological origin, they can still be clearly separated apart, with 27Al, 70Ga, 51V, and 45Sc showed significant difference among the two (Fig 2a& 2c). This may be related to the notion that rice discrimination remains a complex issue, since that not only geographical conditions but the cultivar type may play important roles13. In general, the first two PCs explained 60.7 % of the entire variances; by including the 3rd and 4th PC, nearly 85% of the total variances can be explained then (Fig 2b).

*Feature selection and Repeated grid-search cross-validation for model assessment.*

For the training of classifier (a.k.a. construction of classification model), feature selection coupled with supervised machine learning algorithms were introduced. Currently, opposite views exist regarding when feature selection shall be conducted during the process of classifier training (and validation?): a very popular view sees feature selection as an important data preprocessing step to remove feature variables with low or null discriminating power for the samples14. While on the other hand, there is concern that classification results could be severely biased (i.e. over-optimistic) if selection of features is done prior to the cross-validation26. In this study, we first ranked all the elements basing on their relative importance assigned by ReliefF algorithms. Fig 3 shows the relative importance assigned to each feature (top xx features?). And high ranking features such as x, y, z, were deemed to have higher discriminating power over the classification. Following feature ranking, multiple subsets of elements were then constructed. As shown in Table 2, the 1st subset is made of solely the most important element (i.e. 23Na), the 2nd subset will then include both 23Na and 27Al, the top two elements. Eventually, the 30th subset will include all 30 elements in this study. (此处如何与下文衔接？如何接下来说明我们要做classifier training? 我的建议是先说classifier training (how you fixed the parameter?) 再说如何做的validation 并提供相应结果。)

(这一段应该开始说validation的结果) Krastajic et al. demonstrated that choosing a set of fixed hyperparameters for cross-validation may not render optimal model performance23 .In our study, repeated grit search cross validation (RGSCV) was implemented to avoid pitfalls mentioned above. A two-dimensional matrix was shown in Fig 4 to demonstrate how the “grid-search” was conducted. Particularly, each grid represents a specific subset-hyperparameter combination. The average classification accuracy in each grid were generated after 10-fold cross-validation for 10 times.

After RGSCV, model performance via SVM and RF algorithms were compared side by side. As shown in Fig 5, by using only the top ranked element 23Na, RF achieved 74.76% classification accuracy, while for SVM the accuracy reached 67.60%. After including the second element Al, both RF and SVM achieved very satisfactory performance of 99.14% and 89.42% respectively.

Overall, SVM and RF had comparable performance in terms of the classification accuracy, with RF being slightly better within only top three features (Fig 5). The accuracy got improved with more top-ranking elements been added, and with only four elements (Na, Al, Cd and Rb), both SVM and RF led to satisfactory classification models with 100% accuracy. Add in radar plot and some explanation

This opens the door for future study on whether measuring only a handful of elements could lead to reliable rice classification; particularly, in this case we developed a step-by-step scheme for…

The establishment of simplified analysis may potentially boost the application of easy, cost-effective i~~n-field authenticatio~~n; even However, challenge remains as the elemental profile of crops may be influenced by multiple factors simultaneously, for example, genotype, soil type, climate, and agricultural practice climate37,38. As a result, rather than exploring a universal solution (e.g. a number) for all rice types, a carefully conducted validation should be conducted to ensure the reliability and robustness of classification model.

Given the possibility that there may be some correlation among the concentrations of different elements, traditional univariate data analysis methods was not suitable for discrimination37, 39. Instead, machine learning based multivariate data analysis methods will provide the unique power of processing complex data. It is quite clear to us that sample scarcity along with lack of sample representativeness are of the major reasons leading poor or unreliable classification40. For this study, only 131 samples from six different GI rice were collected within one year of harvest. A larger dataset and further model refining shall be introduced to assess whether different harvest year would have any impact on model training. In this study, multielement profiling with ICP-MS was combined with machine learning data analysis, in order to distinguish six GI rice in China. Followed by feature selection, SVM and RF models were developed, and their performance was compared side by side. With only four elements, both SVM and RF achieved satisfying performance with 100% classification accuracy. In conclusion, the workflow we have established in this study proved to be a feasible way for GI rice authentication and will therefore protect farmer, supplier and consumer from potentially fraudulent activities.

**ACKNOWLEDGMENT**

We want to thank Di Wu, Si Lin for their precious support for sample acquisition and preparation. We also appreciate Peter Markwell for valuable suggestion on the preparation of the manuscript.

**FUNDING SOURCES**

**TBC**

**REFERENCE**

1. Drivelos, S. A. & Georgiou, C. A. Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. *TrAC - Trends Anal. Chem.* **40**, 38–51 (2012).

2. Summary of the Paris Convention for the Protection of Industrial Property (1883). Available at: https://www.wipo.int/treaties/en/ip/paris/summary\_paris.html.

3. Luykx, D. M. A. M. & Ruth, S. M. Van. Food Chemistry An overview of analytical methods for determining the geographical origin of food products. **107**, 897–911 (2008).

4. PDO and PGI Agricultral products:a 14.2 billion euro turnover for over 800 products. Available at: https://ec.europa.eu/agriculture/quality/schemes/newsletter-2010\_en.pdf.

5. Li, Y. Protection of Geographical Indications in China. (2017). Available at: https://www.niuyie.com/protection-of-geographical-indications-in-china/.

6. FAOSTAT 2017. Available at: http://www.fao.org/faostat/en/?#data/QC.

7. Jin, S., Zhang, Y. & Xu, Y. Amount of information and the willingness of consumers to pay for food traceability in China. *Food Control* (2017). doi:10.1016/j.foodcont.2017.02.012

8. Veeck, G. Food Safety Concerns and Rice Imports in China : 1998 - 2016. (2017).

9. Qian, L. *et al.* Determination of geographical origin of wuchang rice with the geographical indicator by multielement analysis. *J. Food Qual.* **2019**, (2019).

10. Zhu, L. *et al.* Identification of rice varieties and determination of their geographical origin in China using Raman spectroscopy. *J. Cereal Sci.* **82**, (2018).

11. Mertens, B. J. A. & Thompson, M. The authentication of Basmati rice using near infrared spectroscopy. (1993). doi:10.1255/jnirs.8

12. Pettinger, B., Ren, B., Picardi, G., Schuster, R. & Ertl, G. Tip-enhanced Raman spectroscopy (TERS) of malachite green isothiocyanate at Au(111): Bleaching behavior under the influence of high electromagnetic fields. *J. Raman Spectrosc.* **36**, 541–550 (2005).

13. Cheajesadagul, P., Arnaudguilhem, C., Shiowatana, J., Siripinyanond, A. & Szpunar, J. Discrimination of geographical origin of rice based on multi-element fingerprinting by high resolution inductively coupled plasma mass spectrometry. *Food Chem.* **141**, 3504–3509 (2013).

14. Maione, C., Batista, B. L., Campiglia, A. D., Barbosa, F. & Barbosa, R. M. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. *Comput. Electron. Agric.* **121**, 101–107 (2016).

15. Wang, X. & Harrington, P. de B. Differentiating Rice Varieties by Inductively Coupled Plasma Mass Spectrometry Chemical Profiling with Singular Value Decomposition Background Correction. *J. Anal. Test.* **2**, 138–148 (2018).

16. Maione, C. & Barbosa, R. M. Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: A review. *Crit. Rev. Food Sci. Nutr.* **8398**, 1–12 (2018).

17. Barbosa, R. M., Nacano, L. R., Freitas, R., Batista, B. L. & Barbosa, F. The Use of Decision Trees and Na??ve Bayes Algorithms and Trace Element Patterns for Controlling the Authenticity of Free-Range-Pastured Hens’ Eggs. *Journal of Food Science* **79**, C1672–C1677 (2014).

18. Brereton, R. G. *et al.* Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. *Anal. Bioanal. Chem.* **409**, 5891–5899 (2017).

19. Hopfer, H., Nelson, J., Collins, T. S., Heymann, H. & Ebeler, S. E. The combined impact of vineyard origin and processing winery on the elemental profile of red wines. *Food Chem.* **172**, 486–496 (2015).

20. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

21. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).

22. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).

23. Krstajic, D., Buturovic, L. J., Leahy, D. E. & Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **6**, 10 (2014).

24. Dietterich, T. Machine Learning Research: Four Current Directions. *AI Mag.* **18**, (2000).

25. Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* **85**, 189–203 (2018).

26. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci.* **99**, 6562–6566 (2002).

27. Chen, Y.-W. & Lin, C.-J. Combining SVMs with Various Feature Selection Strategies. in *Feature Extraction: Foundations and Applications* (eds. Guyon, I., Nikravesh, M., Gunn, S. & Zadeh, L. A.) 315–324 (Springer Berlin Heidelberg, 2006). doi:10.1007/978-3-540-35488-8\_13

28. Costa, N. L., Llobodanin, L. A. G., Castro, I. A. & Barbosa, R. Finding the most important sensory descriptors to differentiate some Vitis vinifera L. South American wines using support vector machines. *Eur. Food Res. Technol.* **245**, 1207–1228 (2019).

29. Cordella, L., De Stefano, C., Fontanella, F. & Marrocco, C. A Feature Selection Algorithm for Handwritten Character Recognition. *Proc. - Int. Conf. Pattern Recognit.* 1–4 (2009). doi:10.1109/ICPR.2008.4761834

30. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. (2019).

31. Jed Wing, M. K. C. *et al.* caret: Classification and Regression Training. (2019).

32. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).

33. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A Grammar of Data Manipulation. (2019).

34. Mundt, A. K. and F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. (2017).

35. Egli, M., Fitze, P. & Oswald, M. Changes in heavy metal contents in an acidic forest soil a  ected by depletion of soil organic matter within the time span 1969 ± 93. **105**, 367–379 (1999).

36. Borges, E. M. *et al.* Monitoring the authenticity of organic grape juice via chemometric analysis of elemental data. *Food Anal. Methods* **9**, 362–369 (2016).

37. Chung, I. M. *et al.* Geographic authentication of Asian rice (Oryza sativa L.) using multi-elemental and stable isotopic data combined with multivariate analysis. *Food Chem.* **240**, 840–849 (2018).

38. Zhang, Y., Song, Q., Yan, J. & Tang, J. Mineral element concentrations in grains of Chinese wheat cultivars. 303–313 (2010). doi:10.1007/s10681-009-0082-6

39. Liu, Z. *et al.* Assuring food safety and traceability of polished rice from different production regions in China and Southeast Asia using chemometric models. *Food Control* **99**, 1–10 (2019).

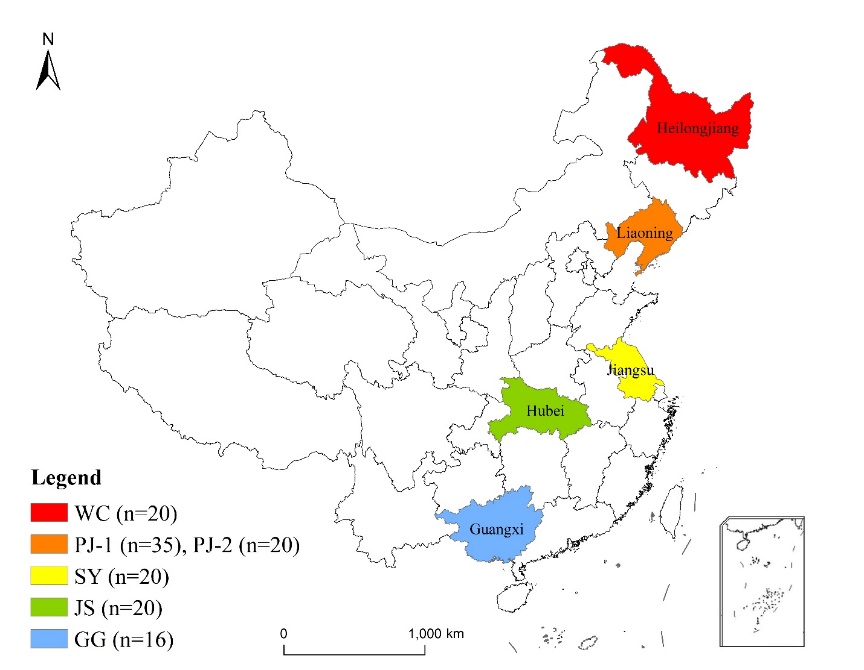
40. Filzmoser, P. & Group, F. *Intro. to multivariate statistical Analysis in Chemometrics*. (2008).

TABLES

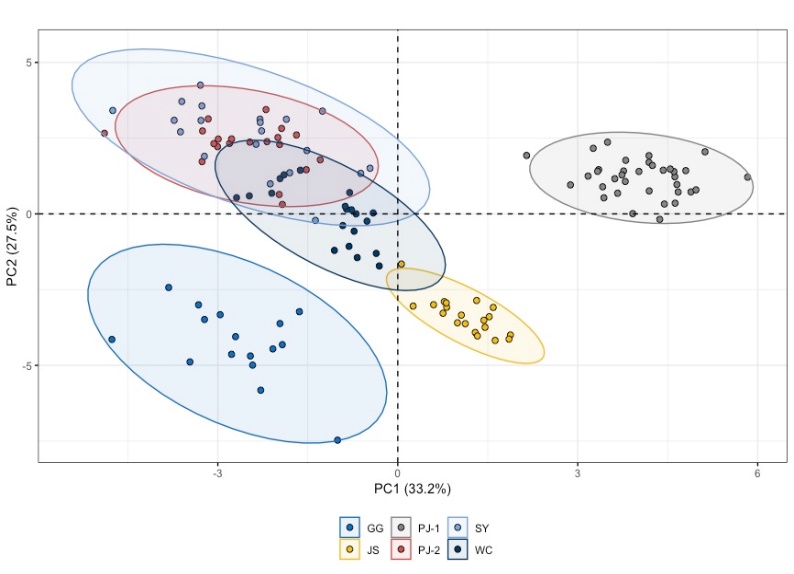




FIGURES



(a)



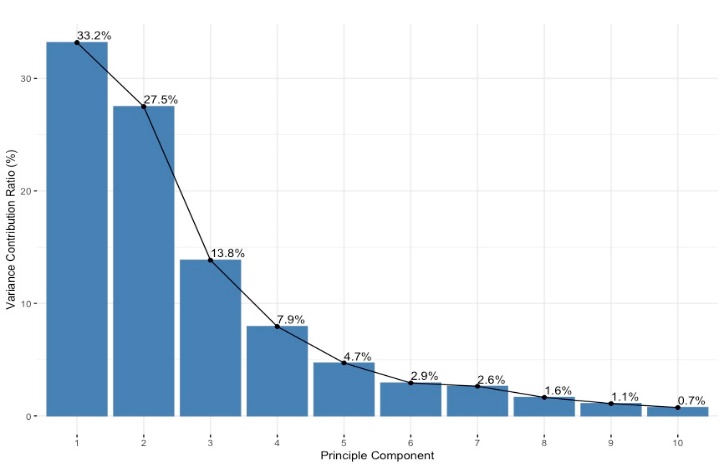
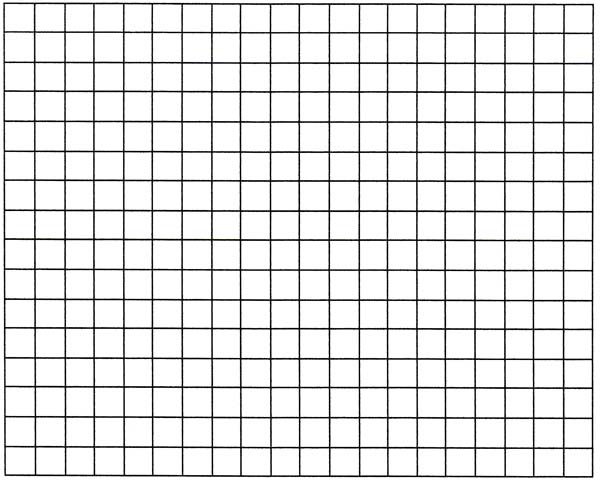
(b)(c)



Fig.2.PCA on elemental concentration based on 30 variables in six GI rice: (a) Score plot of PC1 and PC2, with 95% confidence interval eclipse; (b) Distribution of variance contribution ratio of first ten PC (c) Loading plot of all variables on first two PCs.

Fig. 3. Relative variable importance based on Relief algorithm .



S1

S3

S2

Sn

H1

H2

H3

Hn

Fig. 4. The construction of Two-dimensional matrix for Grid-search. Each grid represent different feature subset-hyperparameter combination.

Fig. 5. RGSCV results on Chinese GI rice dataset using RF, and SVM, with different number of selected features