Target: Journal of Agricultural and Food Chemistry (IF: 3.571)

**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 makes the GI rice at high risk of adulteration due to the high value and limited production. This study aimed to develop a novel strategy to determine geographical origins of Chinese GI rice, which was non-targeted data analysis based on multi-elemental profiling using inductively coupled plasma mass spectrometry (ICP-MS). One hundred and thirty-one samples from six types of Chinese GI rice were analyzed, and 80 % and 20 % of the dataset were used as training set and testing set respectively. Two machine learning algorithms, support vector machines (SVM) and random forest (RF), along with feature slection (reliefF algorithm) were implemented to build classificaition models. For both SVM and RF, four elements (Al, Rb, B, and Na) only could enable the prediction of geographical origins with 100% accuracy. These results demonstrate that using ICP-MS combined with machine learning techniques is an effective strategy forauthenticating GI rice in China.

KEYWORDS

rice, ICP-MS, Geographical Indication, machine learning, feature selection, chemometrics

**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, products with GI certification generally possess given quality, reputations or other characteristics attributable to geographical origins3, thus making it possible for them to be differentiated from 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, three government sectors supervise and protect GIs from different aspects at the administrative level5, including the 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).

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 population6. China is leading in the rice paddy production in the world, with 220 million metric tons in 20187.With the improvement of people’s living standard, there is a growing domestic demand for rice with traceable origins8. For example, Wuchang rice, one 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 the high market demand, the price of Wuchang rice is roughly twice that of other domestic ones, making WuChang rice vulnerable to be aduleterered such as partial substitution and fraudulent labeling9,10. Therefore, determination of geographical origins of rice is of great importance for protecting the rights of farmers, retailers and consumers3. Yet to our knowledge, there is by far still a lack of universally recongnized techniques for the protection Chinese GI rice.

Recent years, various analytical methods have been developed to address the issue of geographical authentication, among which inductively coupled plasma mass spectrometry (ICP-MS) analysis with xx has drawn more and more attentions. ICP-MS can simutltaneously detect both metal and non-metal elements, and is characterized by the high throughput measurement with wide dynamic range and relateviely simple sample preparation11. Based on the scientific proof that element composition in plants are vestly determinded by factors such as soil characteristics and agricultural practice12,13, multi-elemental profiling based on ICP-MS analysis has been successfuly applied to determine geographical origins of rice14–16.

As summarized in a recent review18, MVA such as principal component analysis (PCA) and linear discriminate analysis (LDA) are by far the dominant methods for data processing, due to their simplicity in spotting hidden trend embedded in the dataset. However, with the complexity and volume of data increasing, more advanced models based on pattern recognition are in urgent demand19. In the past decade, ML has demonstrated its capability to process complex problems, particularly in the domain of ecology20, biomedical21,22 , astronomy23 and bioinformatics24. Only a few studies have explored the application of ML in food authentication19,where support vector machines (SVM) and random forest (RF) have been reported to outperform traditional MVA25 and lead to predition models with increased reliability and robustness.15,26

The aim of this study was to develop a novel ML-based workflow for the determination of geographical origins of six types of Chinese GI rice. SVM and RF were utilized to uncover the hidden information in the elemental profiling obtained by ICP-MS, and thereby construct reliable predition models. Furthermore, feature selection was also appliedto identify biomarkers that contributed the most to the differentiation between GI rices.

**MATERIALS AND METHODS**

*Rice samples*

In this study, a total of one hundred and thirty-one Chinese GI rice samples were directly collected from credible rice processing factoriesin five provinces in China (Heilongjiang, Liaoning, Jiangsu, Hubei and Guangxi). For simplicity’s sake, we named those samples as WC, PJ-1, PJ-2, SY, JS, and GG. An overview of the geographical information of samples was shown in Fig. 1.

*Reagents and standards*

Nitric acid (69%, part# 100441) was 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) and 4 (part# 8500-6942), environmental calibration standard (part# 5183-4688), 45Sc (concentration? part# 5190-8578) and 103Rh (10 mg L-1, part# 8500-6945) standards were purchased from Agilent Technologies (Santa Clara, CA, USA). The certified reference material (CRM) of rice flour (1568b) was purchased from the National Institute of Standards and Technology (Gaithersburg, MD, USA).

*ICP-MS analysis*

Firstly, 0.5 g of rice grains was weighed in a Teflon digestion vessel and mixed with 6mL of nitric acid. The vessel was placed in a fume hood overnight for pre-digestion and then transferred to the microwave oven (Anton Paar, Austria) for acid digestion. The digestion temperature of 180 °C was gradually reached in 15 min, and held for 20 min. Following the digestion, the solution was cooled down to room temperature and diluted to 50mL with dionized water, in metal-free plastic tubes. Before use, the digestion vessels were soaked in a nitric acid solution (30% in water, *v*/*v*) for 24h and rinsed with deionized water for three times, thus to avoid cross contamination.

Xx mL was injected into an Agilent 7900 ICP-MS (Agilent technologies, Santa Clara, CA, USA) for multi-elemental profiling. The instrumental setting and operative conditions were adopted from a published method27 with some modifications: radio frequency power of 1550 W, radio frequency 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) were measured. The standard solution of 103Rh and CRM were analyzed to verify the stability and accuracy of the ICP-MS method.

Sample replicate here

*Statistical analysis*

The original dataset was analyzed by one-way analysis of variance (ANOVA) coupled with Tukey’s test (*p* ≤ 0.05), to compare the levels of elements in the GI rice samples. The dataset was then scaled by taking logarithmic transformation and subjected to unsupervised PCA for initial visualization. Two machine learning algorithms, RF and SVM were implemented to construct classifiers. Fig. 2 showed the workflow for the training of classifiers and the validation of the classification models:

* + - * The scaled dataset was randomly splitted into a training set (sample number: 104) and a testing set (sample number: 27) in a stratified fashion (80:20).
      * The ReliefF algorithm was applied only to the training set. Following this, a 10-fold grid-search cross-validation33 was conducted and optimal classifiers were selected based on the mean cross-validation accuracy. Forward selection34 was conducted to select pre-ranked features, and stopped when the adding of features made no contribution to the improvement of the prediction accuracy. Meanwhile, all possible combinations of hyperparametes were tested.
      * The optimal classifiers were independently validated on the testing set, and the prediction accuracy was evaluated.

All data analysis were carried out by R (R Core Team, 2019), RStudio (v3.5.1, Boston, MA, USA) and Python (v3.7, Python Core Team) with additional packages: dplyr (R)35, factoextra (R)36, FSelector (R)36, sklearn (Python)37, skrebate (Python)38, numpy (Python)39 and pandas (Python)40 .

**RESULTS AND DISCUSSION**

*Elemental concentrations*

As shown in Table S1, the measured concentrations of elements in SRM agreed well with the certified values, indicating the high accuracy of the ICP-MS analysis. Table 1 shows the measured concentrations of 30 elements in the six typies of Chinese GI rice. Overall, except for 208Pb, significant differences could be observed among all elements across all types of rice. However, it was not obvious which element contributed the most to differentiate all types of rice.

*PCA analysis*

As shown in Fig. 3a, the 1st and 2nd principal component (PC) were accountable for 60.7% of the total variance, and a clear separation was observed among PJ-1, GG and the rest types. While for JS, PJ-2, SY and WC, no satisfactory separation could be achieved. The loading plot (Fig. 3b) 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. Notably, PJ-1 and PJ-2 could be clearly separated, despite that they were from the same geological origin (Fig. 3a). A possible explaination could be that cultivar types could also contribute to the elemental composition in rice kernals14.

*Determination of geographical origins*

Xx Sampling is fundamental to achieve reliable results from multivariate model building41, while sample scaricity along with unrepresentative samplingare major reasons leading to unreliable classification42. In this study, rather than sampling from the market, we obtained all GI rice samples from reliable sources, which enesurd the authencitity of samples and minimized the risk of modeling with “contaminated” dataset.

Fig. 4 showed the key results obtained from the training of classifiers. Beyond the function of dimention reduction, feature selection is also capable to identify biomarkers with high predictbility44. The relative importance of each of the 30 elements is shown in Fig. 4a, indicating 27Al, 85Rb, 10B, 23Na, and 86Sr were the leading elements that contributed the most to the differentiation of all types of GI rice. With only one selected feature (27Al), the mean cross-validation accuracy of 48% and 63% were achieved for RF and SVM, respectively (Fig. 4b). The performance of both RF and SVM boosted significantly with more features been added. Eventually, with only four features (27Al, 85Rb, 10B, and 23Na), the mean cross-validation accuracy of 100% was obtained by both RF and SVM with optimal hyperparameters applied (Fig. 4b). Notably, to ensure the integrity of the validation process and to avoid selection bias such as over-optimistic prediction, feature selection was only applied to the training set45 .

While cross-validation was applied to assess the goodness-of-fit of modeling within the training set, it is not enough to validate the classification model19. The independent validation with the testing set, is the one and only valid paradigm46 for model assessment. The result of independent validation using the testing set is shown in table 2.According to the result, both classifiers could predicit the geographical originis of all six types of GI rice with 100% accuracy. The results indicated that the information from the four features (27Al, 85Rb, 10B, and 23Na) were sufficient to be used to predict the geological origins. A similar approach was applied by Maione et al.15, who has identified 114Cd, 85Rb, 12Mg and, 19K as the most relevant elements for the differientiation between rice samples obtained from two geological orgins in Brazil. To further study why the four features played such a critical role in the differentiation, we plotted their relative median concentrations. As shown in Fig. 5 , each GI rice had a unique elemental profile. Interestingly, while PJ-1 had the highest level of 27Al among all six types of GI rice, PJ-2 had the lowest. And for GI rice obtained from the same geological location, PJ-1& 2 differed in the composition of 23Na, 85Rb, and 10B. Such observation agreed with previous findings that cultivar types also majorly impact on the elemental composition in rice47,48. Overall, it remains a challenging task to elucidate the rationale for why the four elements were showing strong differentiation power. The complexity here, shall partially attribute to the sample diversity. In this study, we collected samples from all three dominate rice producing regions in China, including the northeast China plain (WC, PJ-1, and PJ-2), Yangtze River Basin (SY, JS), and southeast coastal region (GG). Such wide geological sampling scope, introduced multi-layers of complexity (e.g. soil characteristics, agricultural practices, and genotype variation), which are all closely related to the elemental profile of crops12,13.

Beyond 27Al, 85Rb, 10B, and 23Na, the elemental of 114Cd, which is regarded as major contaminant in paddy field, also drew our attention. In the study on Brazilian rice mentioned earlier, it was found that the level of Cd alone can be used to differentiate rice from two geological origins. The author further pointed out that it was the difference in irrigation methods that resulted in the variance of 114Cd composition15. According to our results, 114Cd was found in all six types of GI rice, with the concentrations all below the China national stardard of 0.2 ppm (Table 1). Particularly for GG rice, which was sampled from Guangxi province, had the highest level of 114Cd. This agreed with a previous national scale study, which showed that the concentration of 114Cd in paddy soils from different Chinese regions varied significantly, with the higheset level found in southeast coastal regions (e.g. Hunan and Guangxi province)49. With this, we further evaluated the feasility of using 114Cd as a biomarker to differentiate GG rice from the others. ~~.~~ After reconstruction, a binary dataset consists of GG rice and non-GG rice was generated. Fig. 6 was used to visualize the significant difference of 114Cd concentration between GG and non-GG rice samples (*P* < 0.05). Following the same workflow we have established earlier, we have further confirmed that classifiers built with only 114Cd could lead to 100% differentiation between GG and non-GG rice samples.

Our study demonstrated that multi-elemental profiling using ICP-MS coupled with ML techniques, could differentiate six types of Chinese GI rice with extremely high accuracy. Particularly, we identified four elements with the most differentiation power. This opens the door for future study on the development of reliable rice classification with only a handful of elements.

**ACKNOWLEDGMENT**

We want to thank Dr. Di Wu from Yangtze Delta Region Institute of Tsinghua University for his tremendous support on sampling. We thank Ms. Si Lin and Hongwei Qiao for their industrious work on experimentation and documentation.

**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. Chauhan, B. S., Jabran, K. & Mahajan, G. *Rice Production Worldwide*.

7. Paddy rice production worldwide 2017-2018, by country,. *Statista* (2019). Available at: https://www.statista.com/statistics/255937/leading-rice-producers-worldwide/.

8. 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

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

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

11. Wadood, S. A., Boli, G., Xiaowen, Z., Hussain, I. & Yimin, W. Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchem. J.* **152**, 104295 (2020).

12. 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).

13. 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

14. 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).

15. 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).

16. 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).

17. Bevilacqua, M. *et al.* Recent chemometrics advances for foodomics. *TrAC Trends Anal. Chem.* **96**, 42–51 (2017).

18. 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).

19. Jiménez-Carvelo, A. M., González-Casado, A., Bagur-González, M. G. & Cuadros-Rodríguez, L. *Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review*. *Food Research International* **122**, (Elsevier Ltd, 2019).

20. Cutler, D. R. *et al.* RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology* **88**, 2783–2792 (2007).

21. Hu, W. Identifying predictive markers of chemosensitivity of breast cancer with random forests. *J. Biomed. Sci. Eng.* **3**, (2010).

22. Ram, M., Najafi, A. & Shakeri, M. T. Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest. *Iran. J. Pathol.* **12**, 339–347 (2017).

23. Gao, D., Zhang, Y. & Zhao, Y. Random forest algorithm for classification of multiwavelength data. *Res. Astron. Astrophys.* **9**, 220 (2009).

24. Boulesteix, A.-L., Janitza, S., Kruppa, J. & König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Min. Knowl. Discov.* **2**, 493–507 (2012).

25. Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* **879**, 10–23 (2015).

26. Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. & Huvenne, J. Chemometrics and Intelligent Laboratory Systems Support vector machines ( SVM ) in near infrared ( NIR ) spectroscopy : Focus on parameters optimization and model interpretation ☆. *Chemom. Intell. Lab. Syst.* **96**, 27–33 (2009).

27. 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).

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

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

30. Urbanowicz, R., Meeker, M., LaCava, W., Olson, R. & Moore, J. Relief-Based Feature Selection: Introduction and Review. *J. Biomed. Inform.* **85**, (2017).

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

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

33. 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).

34. Data Handling in Science and Technology. in *Wavelets in Chemistry* (ed. Walczak, B.) **22**, ii (Elsevier, 2000).

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

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

37. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

38. Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M. & Moore, J. H. Benchmarking relief-based feature selection methods for bioinformatics data mining. *J. Biomed. Inform.* **85**, 168–188 (2018).

39. Walt, S. van der, Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

40. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference* (eds. van der Walt, S. & Millman, J.) 51–56 (2010).

41. 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).

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

43. Guyon, I., Andr, #233 & Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).

44. Grissa, D. *et al.* Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. *Front. Mol. Biosci.* **3**, 30 (2016).

45. 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).

46. Esbensen, K. H. & Geladi, P. Principles of Proper Validation : use and abuse of re-sampling for validation. 168–187 (2010). doi:10.1002/cem.1310

47. Li, Z., Li, L., Pan, G. & Chen, J. Bioavailability of Cd in a soil-rice system in China: Soil type versus genotype effects. *Plant Soil* **271**, 165–173 (2005).

48. Wang-da, C., Guo-ping, Z., Hai-gen, Y. A. O., Wei, W. U. & Min, X. U. Genotypic and environmental variation in cadmium , chromium , arsenic , nickel , and lead concentrations in rice grains \*. **7**, 565–571 (2006).

49. Liu, X., Tian, G., Jiang, D., Zhang, C. & Kong, L. Cadmium (Cd) distribution and contamination in Chinese paddy soils on national scale. *Environ. Sci. Pollut. Res.* **23**, 17941–17952 (2016).

50. Yu, Y. *et al.* Accuracy and stability improvement in detecting Wuchang rice adulteration by piece-wise multiplicative scatter correction in the hyperspectral imaging system. *Anal. Methods* **10**, 3224–3231 (2018).