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 of six GI rice with 100% accuracy. These results demonstrated that the feasibility of using ICP-MS combined with machine learning techniques as an effective strategy for authentication of 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 demand domestically for rice with traceable origins8. For example, Wuchang rice, one GI rice produced in Heilongjiang province in China’s northeast region, is known for the 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 than that of other domestic ones, making WuChang rice vulnerable to adulteration 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.

In recent years, various analytical approaches with the aid of multivariate analysis (MVA) and machine learning (ML) techniques are in rapid development. Among all the analytical approaches, multi-elemental profiling based on inductively coupled plasma mass spectrometry (ICP-MS) analysis, has already been proved to be a promising information souces to determine the origins of rice11–13.

Beyond the advance of modern analytical appoarches, another cornerstone for the success of geographical origination, is to ensure the large volumn of data generated can be properly processed and interpreted (FZ:ref). As been summarized in a recent review14, MVA, such as principal component analysis (PCA), and linear discriminate analysis (LDA) is by far the dominant method for data processing, due to their simplicity and success in spotting hidden trend embedded in the dataset. However, as the variety and the amount of data generated by various modern analytical instructments have been booming recently, more advanced pattern recognition models are essential to make good use of this gigantic and complex dataset (reference). In the past decade, ML has demonstrated its capabilities to process complex problems15, particularly in the domain of ecology (reference), medicine (reference), astronomy (reference), and bioinformatics (reference), while only being implemented by few researchers in the area of food authentication. Quite interestingly, some of the most widely adopted techniques such as support vector machines (SVM) and random forest (RF) have been reported to outperform traditional MVA16 , and lead to predition models with increased reliability and robustness.12,17

The aim of the present 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 from the elemental profiling obtained by ICP-MS, and thereby construct reliable predition models. Furthermore, feature selection was also applied, with the aim of identifying key biomarkers that contribute the most to the difference 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 factories, from five provinces in China (Heilongjiang, Liaoning, Jiangsu, Hubei and Guangxi). For simplicity’s sake, we named thos 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), 4 (part# 8500-6942), Environmental calibration standard (part# 5183-4688), 45Sc standard (part# 5190-8578), and 103Rh (10 mg· L-1, part# 8500-6945)were purchased from Agilent Technologies (Santa Clara, CA, USA). One certified reference material (CRM) of rice flour (1568b) was purchased from the National Institute of Standards and Technology (Gaithersburg, MD, USA)

*ICP-MS analysis*

First of all, a 30% (v/v) nitric solution was prepared by diluting nitric acid with dionized water . Before use, the Teflon digestion vessels were soaked in nitric solution for 24h and then rinsed with deionized water for three times, thus to avoid cross contamination. For pre-digestion procedure, 0.5 g of rice grains was directly digested using 6mL of nitric acid in a digestion vessel, in duplicate. The vessel was placed in a fume hood overnight and then transferred to the microwave oven (Anton Paar, Austria). The digestion temperature of 180 °C was gradually reached in 15 min, and held for 20 min. Following the digestion, all solutions were cooled to room temperature and diluted to 50mL, with dionized water, in 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 published method18 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 internal standard solution of 103Rh and SRM was used to verify the stability and accuracy of the analyzing method.

*Statistical analysis*

One-way analysis of variance (ANOVA) coupled with Tukey’s test (*p* ≤ 0.05) was carried out on the original dataset for the comparison of elements’ levels in six GI rice.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. RF was first introduced by Breiman20 and it is made of an ensemble of decision trees, which are generated from original dataset using bootstrap partition; SVM makes classifications by projecting the input vectors into a high dimensional space, and finding a hyperplane that could separate different classes21. For feature selection, ReliefF19,20 was utilized to determine how much each feature contributed to the overall classification by assigning relative importance to features basing on a calculated proxy statistic25. Fig. 2 demonstrated the workflow we used for the training of classifiers and the validation of the classification models:

* + - * The entire scaled dataset generated from 131 samples was randomly splitted into a training set (n=104) and a testing set (n=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-validation22 was conducted on the training set and the mean cross-validation accuracy was reported as the metric of selecting optimal classifiers. Forward selection23 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 tested22.
      * The optimal classifiers generated were independently validated on the testing set.

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)24, factoextra (R)25, FSelector (R)25, sklearn (Python)26, skrebate (Python)27, numpy (Python)28 and pandas (Python)29 .

**RESULTS AND DISCUSSION**

*Elemental concentrations in Chinese GI rice*

Sampling is fundamental to achieve reliable results from multivariate model building30The measured concentrations of SRM is showed in Table S1, which agreed well with the certified values indicating xx. Table 1 shows the measured concentrations of 30 elements in the six typies of Chinese GI rice. ~~The statistical significance was determined by ANOVA and Tukey HSD test~~. Overall, except 208Pb, significant differences could be observed among all elements across all types of rice. However, based on ANOVA, no lelement was identified which could directly differentiate all types of rice.

*PCA analysis*

In order to get an initial overview of the entire dataset, an unsupervised PCA was conducted (95% confident ellipses included). As shown in Fig. 3a, based on the 1st and 2nd principle component (PC), there was a clear separation among PJ-1, GG and other 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. Sepecially, PJ-1 and PJ-2 could be clearly separated, even though they were from the same geological origin, with 27Al, 70Ga, 51V, and 45Sc showed significant difference (Fig. 3a& 3b). This may due to that rice discrimination is a complex issue, where not only geographical origins but cultivar types play an important role11. In general, 60.7 % of the entire variances could be explained by the first two PCs, and nearly 83% by including the 3rd and 4th PC (Fig. 3c).

*Determination of geographical origins of Chinise GI rice*

Fig. 4 and Fig. 5 show the relative importance of featuresand the results of cross-validation, respectively. With only one selected feature, the mean cross-validation accuracy of 48% was achieved by RF, while 63% by SVM. The performance of both RF and SVM boosted dramatcially with more features been added. Eventually, with only four features (Al, Rb, B, and Na), the accuracy of 100% was obtained by both RF and SVM along with optimal hyperparameters were applied. The result of independent validation using the testing set is shown in table 2, including accuracy and kappa coefficient which is a statistic for testing the interrater reliability31. Both classifiers could predicit the geographical originis of all types of GI rice with 100% accuracy. The result indicated that the information from the four features had a significant power of differentiation to enable the classification. By far, it is still challenging to elucidate the rationale why these four elements are showing such strong differentiation power in this study. The complexity here, is at least partially due to the fact that we covered samples from all three dominate rice producing regions in China: the Northeast China plain (WC, PJ-1, and PJ-2), Yangtze River Basin (SY, JS), and southeast coastal region (GG). The wide geological sampling scope would potentially brought along significant diversities in factors such as soil characteristics, agricultural practices, etc, which are all closely related to the elemental profile of crops32,33

*Elemental profiling of Chinise GI rice*The relative median concentrations of the four features were shown in Fig. 6 using radar plot. It’s obvious that each type of GI rice possessed its unique elemental profiling. Specifically, PJ-1 and PJ-2, which were from a same geographical region but with genotypic difference, showed significantly different elemental profiling. For example, PJ-1 had the highest level of Al among all the six types of GI rice, whilst PJ-2 had the lowest. The results demonstrated that the genotype also played an important role on the accumulation of metals in rice, which have been reported in muiliple stuides34,35.

*Biomaker?*

In our study, 114 Cd which is a well known carcinogenic contaminant in rice, was detectd in all six types of GI rice, although the concentrations were xx. Particaulty, the concentration of Cd in GG which was sampled from southeast costal region of China, was significantly higher than in other types. The result was consistent with the previous national scale study, which revealed that the concentration of Cd in paddy soils from different Chinese regions varied significantly, with the higheset level in southeast coastal regions (e.g. Hunan, Guangxi) 36. In a recent study conducted by Maione et al., it was reported that Cd alone could be used to differentiate rice from two Brazilian regions with satisfied accuracy12. Therefore, in our study, the feasility of using Cd as a biomarker to recognize rice from a specific region was evaluated. For better visualization, the kernel density estimation (KDE) plot37 was constructed to estimate the probability density of Cd (Fig. 7). There was a clear cutoff of xx at around 7 between GG and other types, indicating that Cd itself was sufficient to differentiate GG rice from others.

Overall, our study demonstrated that multi-elemental profiling using ICP-MS coupled with machine learning techniques, could differentiate six types of Chinese GI rice with extremely high accuracy. Particularly, we identified four elements with the most differentiation power, which opens the door to a reliable rice classification using only a handful of elements. As been pointed out by other researchers, sample scarcity along with lack of sample representativeness are major reasons leading to poor or unreliable classification38. In this study, only 131 samples from six different GI rice within one year of harvest were analyzed. Therefore, a larger dataset consists of samples from multiple harvest years shall be introduced in the future, which will increase the robustness of the classification model. In addition, considering the ultimate goal is to protect high value GI rice from potential fraudulent activities, it is of great importantce that we also introduce “positive” samples into the classification. One common solution is to dilute GI rice samples with serial does of highly “look-alikes”39 . Given the possibility that there may be certain correlation between the concentrations of different elements, traditional methods of univariate data analysis was not suitable for discrimination32, 40. This study is by far one of few attempts of applying machine learning techniques to process multi-elemental data, and therefore constructed classification models for rice samples. F

**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 for her 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. 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).

12. 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).

13. 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).

14. 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).

15. 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).

16. 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).

17. 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).

18. 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).

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

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

21. 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).

22. 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).

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

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

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

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

27. 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).

28. 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).

29. 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).

30. 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).

31. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. medica* **22**, 276–282 (2012).

32. 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).

33. 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

34. 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).

35. 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).

36. 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).

37. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **27**, 832–837 (1956).

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

39. 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).

40. 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).