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. However, due to the gap between the limited production and high market demand, GI rice become volunerable 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.

Recent years, inductively coupled plasma mass spectrometry (ICP-MS) analysis with the aid of multivariate analysis (MVA) and machine learning (ML) techniques have been developed to address the issue of geographical authentication of rice11–13. Being able to simultaneously detect both metal and non-metal elements, ICP-MS has the advantage of high throughput measurement with wide dynamic range and relateviely simple sample preparation14. Beyond the advance of ICP-MS, it is also of great importance to ensure the large volumn of data generated can be properly processed and interpreted15. As summarized in a recent review16, 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 demand17. In the past decade, ML has demonstrated its capability to process complex problems, particularly in the domain of ecology18, biomedical19,20 , astronomy21 and bioinformatics22. Particularly, some widely used ML algorithms such as support vector machines (SVM) and random forest (RF), have been reported to outperform traditional MVA23 and lead to predition models with increased reliability and robustness12,24. By far, only a few studies have explored the application of ML in geological authentication12,25

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 applied to 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 factories in 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 standard (part# 5190-8578) and 103Rh (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. To avoid cross contamination, all materials including 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.

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 with Agilent 7900 ICP-MS (Agilent technologies, Santa Clara, CA, USA). The instrumental setting and operative conditions were adopted from a published method26 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. An online internal standard solution of 103Rh (10 mg · L-1) was used to correct matrix effects and compensate for possible instrumental deviations. CRM were digested and analyzed following the same procedure above to verify the accuracy of the ICP-MS analysis. The digestion and analysis for each rice sample were repeated in duplicate.

*Statistical analysis*

To compare the levels of elements in the GI rice samples, the original dataset was analyzed by one-way analysis of variance (ANOVA) coupled with Tukey’s test (*p* ≤ 0.05). 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 Breiman27 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 classes28. Feature selection is a data processing technique for data mining, aiming to identify pertinent features while discarding irrelevant ones, which are not informative but contribute to the overall dimensionality of the problem space29. In our study, ReliefF30,31 was utilized as the feature selector by assigning relative importance to features basing on a calculated proxy statistic31. 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. Spec, After which, stepwise forward selection32 was conducted that pre-ranked features were added to the model subsequently until the model performance cannot be improved any further. With all possible combinations of hyperparametes (supplement material) were tested, the model performance was evaluated with a 10-fold grid-search cross-validation and optimal classifiers were selecte d based on the mean accuracy.
      * 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)34, factoextra (R)35, FSelector (R)35, sklearn (Python)36, skrebate (Python)37, numpy (Python)38 and pandas (Python)39 .

**RESULTS AND DISCUSSION**

*Elemental concentrations*

As shown in Table S1, the measured concentrations of elements in CRM 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 kernals11.

*Determination of geographical origins*

High quality sampling is fundamental to achieve reliable results from multivariate model building40, while sample scaricity along with unrepresentative samplingare major reasons leading to unreliable classification41. In this study, rather than sampling from the market, we obtained all GI rice samples from reliable sources, which ensured 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 predictbility42. 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 ,feature selection was only applied to the training set43 . Letting feature selection look at the data before building and validating the model would usually lead to over-optimistic prediction, and results in a worsening in performance as the model applied to the future data (reference 1). The study presented by Ambroise et al. showed the effect of the over-optimistic issue (same reference with 43). With the same classification and feature selection algorithms on the identical dataset, Guyon el al. obtained a 100% internal cross-validation accuracy which was incurred with selection bias (Reference 2); while Ambroise et al. corrected the selection bias by using external cross-validation and resulted in a true cross-validation accuracy of 82.5%. The size of selection bias was profound thus we carefully implemented the feature selection in our workflow to avoid such issue for constructing reliable models.

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 model17 since it is crippled in nature by omitting the critical sampling variance from future data. Rao et al. showed cross-valition usually underestimates error since algorithms effectively overfits in cross-validation space, and cease to be a efficient estimate of generalization for the algorithms (reference). The independent validation with the testing set, is the one and only valid paradigm44 for model assessment. Besides proper chosen validation process, the algorithms used to contruct the prediction model are pivotal for high performance models. Machine learning algoritms had rapidly developed in the past decade and have been reported to have higher performance than traditional discriminant analysis methods (MVA?) (reference) for complex datasets in various research areas. Hence, we have incorporateed 2 most well-known machine learning algoritms, RF and SVM into our workflow and tested the feasibility of using the machine learning algoritms with feature selector to construct classification models. 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 with the information from only 4 elements selected. The results indicated that the workflow we have established have successfully identified key biomarkers and constructed reliable models, not only within the training set, but also account for unknown data, the test set. (Need to state our workflow was successful… Coincidently, in a similar study 12114Cd, 85Rb, 12Mg and, 19K as the most relevant elements for the differientiation between rice samples obtained from two geological orgins in Brazil. To further visualize the elemental pattern differences among 6 type of GI rices, we plotted their relative median concentrations. As shown in Fig. 5 , the elemental profile in each GI rice showed significantly different pattern. Interestingly, the level of 27Al varied significantly between PJ-1 and PJ-2 that while PJ-1 had the highest level of 27Al among all six types of GI rice, PJ-2 had the lowest. And 23Na, 85Rb, and 10B, all showed considerably different composition in PJ-1 &2 even they were obtained from the same geological location. Such observation agreed with previous findings that cultivar types also majorly impact on the elemental composition in rice45,46. Overall, it remains a challenging task to elucidate the rationale for why the four elements showed 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 crops47,48.

Beyond 27Al, 85Rb, 10B, and 23Na, the element of 114Cd also drew our attention. In the study on Brazilian rice, it was found that the level of 114Cd alone can be used to differentiate rice from two geological origins. The author further pointed out that it was the difference in irrigation methods applied to paddy filed that resulted in the variance of 114Cd composition in rice 12. According to our results, 114Cd was found in all six types of GI rice, with the GG rice, originated from Guangxi province, having the highest level (Table 1). Therefore, we further applied the ML-based workflow established earlier by first regrouping the original dataset as GG rice and non-GG rice. According to the result of feature selection, it was found that 114Cd was identified as the element with the most differentiation power over the two subgroups. Fig. 6 visualized the significant difference of 114Cd concentration between GG and non-GG rice samples (*P* < 0.05). Ultimately, the independent validation by testing set confirmed that 114Cd along can be used to differentiate GG rice and non-GG rice. A previous national scale study showed that the concentration of 114Cd in paddy soils from different Chinese regions varied significantly, majorly due to the combination of geogenic factors and anthropogenic pollutions (e.g. mining)49. Particularly, the highest level of 114Cd can be found in southeast coastal regions including Hunan and Guangxi province49. With this, we concluded that among the six GI rice in the studt, 114Cd can be used as a biomarker to differentiate GG rice from the others. This provided GI rice from potential fraudulent activities, we would introduce “positive” samples into the classification. One common solution is to dilute GI rice samples with serial does of highly “look-alikes”50 ..

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. Rodriguez, L., Hall, B., Avenue, S. G., Hall, G. & Street, S. W. Social trust and risk knowledge , perception and behaviours resulting from a rice tampering scandal Sela Sar. **5**, 80–96 (2014).

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

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. 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).

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

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. 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).

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

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

20. 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).

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

22. 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).

23. 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).

24. 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).

25. Richter, B., Rurik, M., Gurk, S., Kohlbacher, O. & Fischer, M. Food monitoring: Screening of the geographical origin of white asparagus using FT-NIR and machine learning. *Food Control* **104**, 318–325 (2019).

26. 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).

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

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

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

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

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

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

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

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

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

37. 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).

38. 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).

39. 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).

40. 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).

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

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

43. 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).

44. 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

45. 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).

46. 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).

47. 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).

48. 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

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).