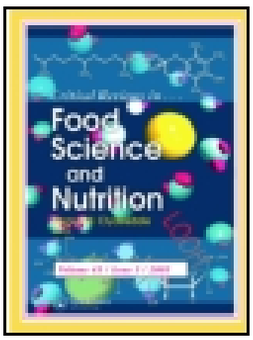

Critical Reviews in Food Science and Nutrition

ISSN: 1040-8398 (Print) 1549-7852 (Online) Journal homepage: http://www.tandfonline.com/loi/bfsn20

Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: a review

Camila Maione & Rommel Melgaço Barbosa

To cite this article: Camila Maione & Rommel Melgaço Barbosa (2018): Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: a review, Critical Reviews in Food Science and Nutrition, DOI: 10.1080/10408398.2018.1431763

To link to this article: https://doi.org/10.1080/10408398.2018.1431763

Accepted author version posted online: 24 Jan 2018.

Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: a review

Camila Maione[1], Rommel Melgaço Barbosa[1,*]

[1] *Instituto de Informática, Universidade Federal de Goiás, Brazil.*

[*]*Corresponding author. Email: rommel@inf.ufg.br.*

ABSTRACT

Rice is one of the most important staple foods around the world. Authentication of rice is one of the most addressed concerns in the present literature, which includes recognition of its geographical origin and variety, certification of organic rice and many other issues. Good results have been achieved by multivariate data analysis and data mining techniques when combined with specific parameters for ascertaining authenticity and many other useful characteristics of rice, such as quality, yield and others. This paper brings a review of the recent research projects on discrimination and authentication of rice using multivariate data analysis and data mining techniques. We found that data obtained from image processing, molecular and atomic spectroscopy, elemental fingerprinting, genetic markers, molecular content and others are promising sources of information regarding geographical origin, variety and other aspects of rice, being widely used combined with multivariate data analysis techniques. Principal component analysis and linear discriminant analysis are the preferred methods, but several other data classification techniques such as support vector machines, artificial neural networks and others are also frequently present in some studies and show high performance for discrimination of rice.

Keywords *– rice, authenticity, geographical origin, variety recognition, data mining, multivariate data analysis*

1. INTRODUCTION

Rice (*Oryza sativa L.*) is one of the most important staple foods for people around the world. In several countries, rice is also the food which is most consumed by its population. Although the excessive consumption of rice can increase the risk of developing type 2 diabetes (Hu et al. 2012) and is related to exposure to toxic elements such as arsenic (Batista et al. 2011; Gilbert-Diamond et al. 2011; Davis et al. 2012), different types of rice still provide many essential elements, nutrients, fibers and vitamins that are beneficial to humans (Batres-Marquez et al. 2009).

Due to the high preference of the world population for this food, rice is constantly being the subject of several studies. Authentication of rice is one of the most addressed concerns in the current literature, and is usually concerned with the recognition of its geographical origin. Identification of the rice variety is also a popular problem in the recent literature. Moreover, the recent growth of the organic industry and the preference of some consumers for organic food has prompted the need for methods capable of certificating organic rice.

Overall discrimination of rice has been widely performed in the recent literature with the aid of multivariate data analysis and data mining techniques. Both processes offer powerful methods capable of performing statistical and predictive analysis over data sets described by many variables. In the case of rice, such variables can be chemical elements, physicochemical properties, climate parameters and many others. One of the main questions regarding the study and differentiation of rice is which variables to analyze, since different parameters may contain more or less information regarding the geographical origin, variety, type of rice or any other label. Another question is regarding what multivariate data analysis method to use combined

with the chosen parameters, since the performance of these methods are highly influenced by the types variables, the number of available samples and the quality of the data set (presence of noises and unspecified values).

In this review, we gather and discuss the recent applications of multivariate data analysis and data mining techniques applied to the discrimination, authentication and study of rice. We highlight the most used parameters for the discrimination and also the preferred algorithms for the differentiation.

2. OVERVIEW OF THE RECENT METHODS

Table 1 summarizes the recent works found regarding discrimination of rice using multivariate data analysis techniques. Aim of the study, data set analyzed, descriptive parameters and multivariate data analysis methods employed are described for each work.

Regarding goals, the majority of the studies aim for recognition of geographical origin and variety of rice, which tackle mainly the authenticity issue. Most of the works on recognition of geographical origin focus on differentiating rice samples from various producing regions of Asia, especially China, followed by Korea and Thailand, which is expected since approximately 90% of world's rice is produced in Asian countries (Muthayya et al. 2012) and rice consumption is among the highest in especially poor populations from this continent (Muthayya et al. 2014). There are also a few papers addressing the geographical origin of Italian, Brazilian, Indonesian, Indian, Japanese and Iranian rice.

We also observe some works focused on predicting specific contents of rice such as nitrogen levels (Shao et al. 2012) and adulterants (Feng et al. 2013; Lim et al. 2017). Culinary quality of rice and the impact of climate parameters to rice yield are also investigated by (Zhang et al.

2010) and (Chen et al. 2016), respectively. Moreover, with the recent growth and popularization of organic and transgenic industries worldwide, the study and authentication of organic and transgenic rice has also become a solid recent concern for rice researchers (Borges et al. 2015; Barbosa et al. 2016; Liu et al. 2016).

2.1 Common statistical and machine learning methods

As we can see in Table 1, principal component analysis (PCA) and linear discriminant analysis (LDA) are clearly the most employed multivariate data analysis techniques for the discrimination of rice. PCA uses a vector space transformation to reduce the dimensionality of the dataset into a smaller number of variables called principal components (Wold et al. 1987), which are linear combinations of the original variables. The principal components are extracted in order of contribution to the total variance of the data, and by investigating the loadings of the variables in the first components, it is possible to measure the relevance of variables. PCA is mostly used as an initial step of the analysis before classification, since it is a useful technique to spot hidden patterns in the data. LDA is a classification technique which aims for maximizing the ratio of between-class variance to the within-class variance in order to achieve maximal separability. The decision boundary created by LDA is called a discriminant function, and is a linear combination of the variables which best separate the classes. Due to its simplicity and empirical success in spotting hidden trends in data, PCA combined with LDA is widely used for classification of data from several knowledge fields and for various purposes.

Machine learning techniques are also present in many of the papers regarding rice discrimination listed in this review. Support vector machines (SVM), k-nearest neighbor (KNN), artificial neural networks (ANN) and hierarchical cluster analysis (HCA) are among the most commonly

used. SVM, ANN and KNN are classification models that are developed through supervised learning, i.e., a set of labeled training data (samples) is used to compute a function capable of predicting the label of new and unknown data.

The first two techniques work around optimization problems. While SVM locates the decision boundary with the largest margin possible which separates the data, ANN tries to reduce the generalization error of the decision function by constantly adjusting the weights associated with each input variable. Both techniques are among the most popular in the present data classification literature due to their proven advantages and empirical success. SVM is developed through a convex optimization problem, has a regularization parameter which works against overfitting and can efficiently handle non-linearly separable data through the kernel trick. ANN is known for its high capability of solving complex problems such as image processing, simple and easy structure development and also the vast number of implementations available. However, both models have parameters that must be properly tuned in order to produce good results, and also generate decision functions that are black boxes, difficult to interpret and to provide information about the individual impact of the variables. While SVM can perform well with a relatively small number of samples, ANN's performance depends on the amount of training samples available while often requiring a large amount of time to be trained when analyzing large data sets. On the other hand, SVM is a binary classifier which basically uses a one-versus-all approach to handle multiclass problems, while ANN can naturally solve multiclass problems. Also, a proper kernel function must be chosen in order to produce an accurate SVM model.

KNN is a classification method which uses the Euclidean distance to compute the *k* samples (neighbors) that are nearest to the test sample in the feature space, and then sets its class label as the most frequent class label occurring in the found neighbors. It is a very simple, easy and low-cost algorithm which works well with small data sets and multiclass problems. However, KNN models are also difficult to interpret and the time required to find the *k* nearest neighbors is highly influenced by the size of the data set. Moreover, the estimated number of training samples required to produce a good model is directly proportional to the number of input variables.

Hierarchical cluster analysis implements unsupervised learning since it handles unlabeled data. The goal of cluster analysis is to divide the data into distinct groups (clusters) in such a way that samples associated with the same cluster are considered similar according to the pattern found, while samples associated with different clusters are as dissimilar as possible. Clustering algorithms are especially useful for uncovering hidden patterns in unlabeled data.

There are several algorithms for data classification and analysis available in the recent literature. Several of them can be found in the studies on discrimination of rice listed in this review, such as partial least square discriminant analysis (PLS-DA) and its orthogonal variant (OPLS-DA), random forests (RF), least square support vector machines (LS-SVM), decision trees, and others.

3. ANALYZED PARAMETERS AND MULTIVARIATE DATA ANALYSIS METHODS USED

3.1 Image processing

All the works listed in this section use as descriptive variables morphological traits and other features such as color, texture, and shape extracted by image processing from images of the rice grains. These features are a promising source of information for data analysis and discrimination.

For instance, color is a powerful descriptor that simplifies object identification from scenes since it offers thousands of different shades and intensities in comparison with shades of grey, and texture is a connected set of pixels that occur repeatedly in an image and provides information about the variation in the intensity of a surface (Singh and Chaudhury 2016).

A methodology for discrimination of Basmati rice grain variety by means of image processing is proposed by (Kambo and Yerpude 2014). Images of the rice grains were captured by a mobile camera and then pre-processed, smoothed and segmented in order to obtain morphological traits. Area, major and minor axis lengths, eccentricity and perimeter of the grains were determined by feature extraction from the stored images and used as input variables for the PCA and KNN methods. The proposed methodology achieved 70% classification accuracy for classic rice, 75% for rozana rice, and 80% for mini rice. The overall accuracy of the model for discriminating basmati rice was 79%.

(Singh and Chaudhury 2016) employed color, texture and wavelet features to discriminate four varieties of bulk rice grain images using back-propagation neural network (BPNN), whose performance is later compared to SVM and KNN. Different lighting conditions and quality were considered in the analyzed images of the rice grains. Eighteen color features, 27 texture features using gray-level co-occurrence matrix, 24 wavelet features and 45 combined features were obtained from images taken from the rice grains. Three different data sets were analyzed, each data set composed of images taken from different cameras and in different lighting conditions and quality (resolution and DPI). Two color models were used for the colour feature extraction in this project, namely RGB and HSI, which provided red, blue and green components along with hue, saturation and intensity values for each image. The co-occurrence matrix method aided in

determining the texture features. Databases of features were created for each data set using 100 images of each rice type. The BPNN model correctly classified 99.5% of the images with low quality and 96.25% of the images with high quality when all the 69 features were used for training.

Morphological, color and textural traits of rice grain images are also used by (Kuo et al. 2016) as descriptive variables to classify 1500 rice grains from 30 different varieties. These images were obtained by a digital camera and microscope, and later processed in order to extract the mentioned traits. The classification algorithm chosen, the sparse-representation-based classification (SRC), encodes the variables of training samples as the atoms in a dictionary, and query samples provided are coded as a sparse combination of atoms, and finally assigned to the class that yields the fewest coding errors (Kuo et al. 2016). This algorithm achieved approximately 89.1% correct classifications through 10-fold cross validation. The performance was compared to the SVM, which achieved a similar accuracy (92.8%). However, the greatest advantage of using SRC over SVM is that it develops a dictionary for each variety which contains essential trait information.

Finally, (Zareiforoush et al. 2016) discriminates milled Hashemi rice grains into four quality classes based on color, shape, and texture information extracted from processed and fragmented images. The four quality classes are low-processed sound grains, low-processed broken grains, high-processed sound grains and high-processed broken grains. Machine learning algorithms were employed for the discrimination task, and SVM, neural networks, Bayesian networks and decision trees achieved 98.48%, 98.72%, 96.89% and 97.5% classification accuracy, respectively.

3.2 Molecular and atomic spectroscopy

3.2.1 Chemical composition and elemental fingerprinting

The use of the chemical composition features combined with multivariate data analysis techniques has been popularized over the last years and is widely used for the differentiation of types, varieties and geographical origins of several foods (Batista et al. 2012; D'Haen et al. 2013; Barbosa et al. 2014, 2015, Maione et al. 2016a, b, 2017). The study of the elemental composition of agricultural products is especially promising for certifying their geographical origin since their elemental fingerprinting are highly affected by the solubility of the minerals in the soil, and therefore agricultural products have unique and principle multi elemental patterns which can reflect the geochemistry of the soil of origin (Chung et al. 2015).

One of the most popular tools for determination of the elemental fingerprinting and concentration levels of chemical elements in food samples and other substances is inductively coupled plasma mass spectrometry (ICP-MS). ICP-MS is capable of discriminating metals and non-metal content at very low concentrations (ng L^{-1}) by using inductively coupled plasma to ionize the samples, and the generated ions are separated and quantified by the mass spectrometer. ICP-MS present several desirable features and advantages over other tools, such as high sensitivity, multi-element capability, wide linear dynamic range, high sample throughput and the ability to discriminate between isotopes (Cubadda 2007). The use of ICP-MS to study the elemental fingerprinting of food samples is very popular, and this tool was employed in practically all studies we found which aimed to discriminate rice samples based on their chemical composition through the use of multivariate data analysis techniques. We found reports on the authentication of Brazilian rice samples (Borges et al. 2015; Barbosa et al. 2016),

discrimination of the geographic origin, variety of Thai rice (Cheajesadagul et al. 2013; Promchan et al. 2016) and also geographic origin of Spanish (Gonzalvez et al. 2011) and Asian (Li et al. 2012; Chung et al. 2015, 2018) rice.

The discrimination of Thai rice based on elemental fingerprinting was performed by (Promchan et al. 2016) and (Cheajesadagul et al. 2013), although their main goals were different. (Promchan et al. 2016) employ the LDA technique to the recognition of the geographical origin of Thai rice between Northeast and South Thailand, and also between white, black, red and yellow types of Thai rice. The LDA obtained a good separation between samples from the two regions of production, with an overall correct classification of 93.8%, and also correctly classified all samples of different types, i.e., 100% accuracy. (Cheajesadagul et al. 2013) aimed for the discrimination of Thai rice samples among three regions of production and also five foreign countries (France, Japan, Italy, India and Pakistan) using PCA and LDA. LDA was much more effective than PCA for classification of rice, achieving a 96.83% prediction accuracy to determine the geographical origin of all the samples using the mean concentration of 12 elements. They also used 7 elements to classify Thai rice regarding its region, achieving a correct classification ratio of 95.22%. The two studies found similar numbers of chemical elements in the rice samples (20 and 21, respectively) with 12 elements in common, which were Al, As, Cd, Co, Cu, Fe, Mg, Mn, Mo, Pb, Rb and Zn.

The authenticity of Brazilian organic rice by means of its elemental composition was investigated by (Barbosa et al. 2016) and (Borges et al. 2015). The first employed PCA, SIMCA, hierarchical cluster analysis and KNN for the differentiation of the rice samples between organic and conventional, while the latter opted for SVM. The two studies found almost identical

elements in the samples, with the only difference being Zn, which was found only in the samples studied by (Borges et al. 2015). The other elements found in the samples from both studies were As, B, Ba, Ca, Cd, Ce, Co, Cr, Cu, Fe, K, La, Mg, Mn, Mo, P, Pb, Rb and Se. (Borges et al. 2015) worked with 18 organic rice samples and 32 conventional rice samples, and all methods were capable of differentiating the organic samples from the conventional ones using only 14 elements, with KNN and SIMCA in particular achieving a 100% prediction accuracy. In this study, Ba, Co and P were the most discriminative elements. (Barbosa et al. 2016) worked with 17 organic rice samples and 33 conventional rice samples of 50 different brands obtained from various Brazilian states, and SVM correctly classified 98% of the samples. The most discriminative elements for the classification in this study were Ca and Cd, and interestingly the SVM achieved a 96% prediction accuracy when only these two chemical components were observed.

(Gonzalvez et al. 2011) employed the LDA technique to discriminate the geographical origin of rice from several PDOs recognized in Spain, such as “Arrós de Valencia” (67 samples), “Arrós del Delta del Elbro” (22 samples) and “Arrós de Calasparra” (11 samples). Uncertified rice samples from specific regions in Spain were also analyzed, such as Tarragona (8 samples) and Extramadura (19 samples), and the remaining 26 samples were obtained from Japan, Brazil and India. For the analysis, the data set was divided in 107 training samples and 46 test samples. LDA was applied to the concentrations of the 32 elements delete and complete separation of the Spanish rice samples from samples from other countries was observed with a correct classification ratio of 91.56%. Lanthanides, Cd and Co have been identified as the most

influential indicators of geographical origin of rice samples due to their different concentrations in soils and their effective uptake by plants.

Studies of the recognition of the geographical origin of Asian rice based on elemental fingerprinting were carried out by (Li et al. 2012), (Chung et al. 2015) and (Chung et al. 2018). (Li et al. 2012) was the first one to employ Fibonacci Index Analysis (FIA) to the discrimination of rice samples. Along with FIA, the authors also used PCA and discriminant function analysis (DFA) to differentiate the Chinese rice produced in nine different regions, namely Fuzhou, Longyan, Nanping, Ningde, Putian, Quanzhou, Sanming, Xiamen and Zhangzhou. Seven, 23, 17, 12, 36, 35, 5 and 39 samples were obtained from each region for analysis, respectively. Thirteen chemical elements were determined in the samples via ICP-MS/OES, which were Ca, K, Mg, P, B, Mn, Fe, Ni, Cu, As, Se, Mo and Cd. While PCA did not present satisfactory performance, DFA presented good performance in classifying samples from two regions. FIA performed well for all regions. The results determined Ca, Ni, Fe and Cd as the most meaningful elements. Rice samples from Korea, China and Philippines were distinguished in (Chung et al. 2015) by PCA and PLS-DA using the concentrations of 25 elements determined by ICP-AES. Again, PCA could not distinguish samples from Korea and China in a satisfactory manner, so PLS-DA was carried out in search of a better separation. PLS-DA was indeed able to differentiate the samples from the three countries with a good level of accuracy. The weighted sum of squares of the PLS weight revealed eleven elements (Cu, Ag, Zn, Cr, Ca, Ba, Cd, Bi, K, Pb and In) as the most meaningful for the differentiation. (Chung et al. 2018) differentiated rice samples from six countries, Cambodia (14 samples), China (6 samples), Japan (10 samples), Korea (12 samples), Philippines (13 samples) and Thailand (4 samples), all of which were harvested in 2015. Along

with the concentrations of the 25 chemical elements determined in the samples by ICP-MS, the authors also analyze the isotope ratios of light-elements (C, N, O, S) determined by IRMS. The first principal components of PCA were able to explain 37.8% and 24.7% of the total variance within the data, respectively, and successfully separated all the geographic origins except for Japan and Philippines. OPSL-DA was used to differentiate between samples from Korea and those of other countries, and obtained a clear separation. The variables with the largest contribution were determined by S-plot: $\delta^{34}S$, Mn and Mg.

3.2.2 Raman spectroscopy and other spectra

Raman spectroscopy is used by (Feng et al. 2013) and (Hwang et al. 2012) to determine the geographical origin of Chinese and Korean rice, respectively. The former also investigates the presence of paraffin in adulterated rice samples and use several multivariate data analysis techniques to perform discrimination based on Raman spectra, such as PCA, SIMCA, SVM, KNN and PLS-DA, while the latter employ the classic PCA combined with LDA methodology. (Feng et al. 2013) analyzed 42 rice samples from indica and japonica varieties obtained from various regions of China, and PCA clearly separated both types of rice. Regarding the recognition of the geographical origin, SVM and KNN performed above 90% accuracy, while SIMCA and PLS-DA achieved a relatively greater error. (Hwang et al. 2012) analyzed 30 imported and 30 domestic polished rice samples provided by the National Agricultural Products Quality Management Service in Korea, and imported rice samples came mostly from diverse regions in China. Two data sets were analyzed, comprising back-scattering and transmission measurements, respectively, obtained from enhanced Raman spectroscopy. Initially, 48 and 12 samples were randomly assigned to calibration and validation sets and PCA-LDA was performed

200 times. The separation of the two groups was more visible when the transmission measurements were used, and discrimination errors were 9.97% and 1.61% for back-scattering and transmission measurements, respectively.

Nuclear magnetic resonance based metabolomics were used by (Huo et al. 2017) for the discrimination of Chinese rice produced in nine provinces. Metabolites data obtained from ^{1}H NMR spectroscopy, with and without sugar, were analyzed by PCA, which was able to separate the samples from each region, combined with LDA, which was used to identify the most meaningful variables for the differentiation. Sucrose, fructose and glucose were the most important variables to differentiate the samples with respect to their geographical origin when sugar were considered in the analysis, and succinate, polyphenols, trigonelline and asparagine otherwise. LDA achieved 100% correct classifications via leave-one-out cross validation in both data sets.

Determination of nitrogen levels in rice via classification of soil plant analysis development value (SPAD) based on near infrared spectra was performed by (Shao et al. 2012). Canopy spectral reflectance data of 64 rice samples were extracted from visible and NIR spectroscopy and used as input data for the LS-SVM, PLS and ANN models in order to predict the SPAD value. Independent component analysis (ICA) was used to determine the sensitive wavelengths and more meaningful variables for the classification. The best model achieved was the LS-SVM trained with ten variables determined by ICA, which presented values for correlation coefficient r_p of 0.9421, root mean square error of 0.2586 and bias of -1.012e-06.

(Yang et al. 2017) investigate laser-induced fluorescence spectroscopy combined with PCA and SVM to classify different varieties of indica rice. Six different varieties of paddy rice were

analyzed: victory indica, manley indica, cucumber indica, guangchang ai, ii-youg 838 and yangdao 2, all of them cultivated in 2016 in Huanggang County in the province of Hubei, China. The three principal components explained 96.58% of the total variance within the fluorescence spectra. The factor scores computed from the three PCs were employed as input parameters for the SVM model, trained by 3-fold cross validation, which presented 91.36% accuracy for predicting the rice variety.

(Liu et al. 2016) propose terahertz spectroscopy imaging combined with several machine learning techniques to discriminate transgenic and conventional rice seeds. The genotype data from terahertz spectroscopy imaging was obtained from 200 rice seed samples from both types obtained from the Institute of Rice Research, China. After pre-treatment with the first and second derivative and standard normal variate transformation (SNV), the data was used as input variables for PCA, LS-SVM, PCA combined with back propagation neural networks and random forests. The best model achieved was the random forest combined with first derivative pre-treatment, which presented a 96.67% accuracy.

3.3 Genetic parameters

(Xia et al. 2010) proposed a method based on SVM and CGR for the prediction of protein in rice that is resistant against *Xoo* (*Xanthomas oryzae* pv. *oryzae*). The motivation for this study was the yield loss in several producing areas due to the increasing occurrence of diseases, and *Xoo* is known to cause one of the most devastating bacterial blights in rice-producing regions. For the analysis, the authors obtained 13 proteins (closed genes) in rice from NCBI that are resistant to *Xoo*, and 48 proteins (selective genes) obtained from KOME that have no evidence of being *Xoo* resistant. SVM classifiers were employed to differentiate Xoo-resistant proteins from non-

resistant. The analyzed variables were 20 residues, 400 dipeptides and 24 other features obtained via chaos game representation (CGR), which processes a given DNA sequence into a picture with a fractal structure that visually reveals previously unknown structures. The SVM models trained with different combinations of features and leave-one-out cross validation achieved a mean accuracy of 95% for predicting proteins resistant to *Xoo*, proving the success of the SVM-CGR methodology for this purpose.

(Zhang et al. 2010) brings a study of good eating quality of japonica rice based on genetic similarity matrix detected by simple sequence repeat (SSR) markers, which is an effective tool for identifying genetic variation of germplasms. Phenotypic values of agronomic and taste traits were also used to evaluate the genetic diversity of the rice varieties, which can be a useful source of information for breeding rice varieties with good eating quality. The analyzed data comprised 60 high-quality conventional japonica rice accessions in China and abroad. Analysis of genetic basis of the accessions based on SSR markers helped to determine their inter-relations. Genetic similarity coefficients among the accessions were calculated, and the values obtained suggest that most of the current japonica materials with good eating quality have a high genetic similarity and a narrow genetic basis. The rice germplasm resource survey and the recorded standard in China aided in acquiring agronomic traits. For the taste evaluation, a sensory score test was conducted with 22 persons of age between 20 to 60 from 11 provinces of China. Cluster analysis based on taste traits provided valuable information for taste research and breeding. Four groups were identified: medium taste, good taste, lower taste, and the last group comprised samples that had outstanding performance in appraisal indicators of sensory tests. Cluster analysis based on agronomic traits showed high consistency with their geographical origins.

3.4 Cell and molecular contents and features

A method for detection of adulterated admixtures of Korean white rice using targeted lipidomics and machine learning techniques was proposed by (Lim et al. 2017). The analyzed data were comprised of batches of white rice from Korea, China, and adulterated rice originated from mixing rice from both countries. Seventeen lysoglycerophospholipids (lysoGPLs) available in the rice samples, being 6 lysophosphatidylcholines, 7 lysophosphatidylethanolamines and 4 lysophosphatidylglycerols, were used as input variables. The machine learning techniques used for the discrimination of the adulterated rice were SVM, KNN, random forests, ANN and C5.0 based decision trees. Random forests and SVM presented the best results for differentiating Korean, Chinese and mixed rice samples, while KNN showed the worst performance.

(Lee et al. 2012) performed discrimination of rice cultivars from Korea by the application of cluster analysis techniques on starch processability indicators such as hydration and pasting parameters, and amylose content. Rice grains from 12 different cultivars grown at the National Institute of Crop Science in Korea were used for the analysis. Cluster analysis was also employed by (Pramai and Jiamyangyuen 2016) for the discrimination of white, red and black varieties of pigmented rice based on color parameters and antioxidants such as total anthocyanins, phenolic and flavonoid contents, DPPH radical-scavenging ability, ferric antioxidant power, trolox equivalent antioxidant capacity, α-tocopherol, and γ-oryzanol.

Finally, (Sabir et al. 2017) differentiated red and white Indonesian rice brans based on compounds determined in rice by HPLC chromatography and fingerprint analysis with diode array detector (DAD): cycloartenol ferulate, cyclobranol ferulate, campesterol ferulate and β-sitosterol ferulate. PCA did not show a clear separation of the red and the white rice brans,

possibly due to similar chemical compounds contained in them according to the chromatogram profiles. Cycloartenol ferulate and β-sitosterol ferulate are among the most important variables according to the standardized discriminant function coefficients. Discriminant analysis achieved correct classifications of 100% when evaluated by leave-one-out cross validation.

3.5 Other parameters

In addition to the parameters discussed, several other types of features were evaluated in combination with multivariate data analysis techniques in recent studies for the discrimination of rice, such as physical properties (Marini et al. 2004), climate parameters (Chen et al. 2016), sensor data from electronic nose and tongue (Lu et al. 2015) and others.

(Chen et al. 2016) applied machine learning techniques to six climate variables to study their relative importance in rice yield. According to the authors, climate indicators play critical roles in crop growth and development and they can provide valuable information on adopting proper strategies in crop planting and management under climate change condition. The climate variables analyzed were temperature, daily temperature range, relative humidity, rainfall, rainy days and sunshine hours during rice growth. The analyzed data comprised statistical values of rice yield during the period of 1985-2012 obtained from Chongqing Statistical Yearbook. The classification techniques adopted were SVM and ANN. Three kernel functions were used to develop the SVM models: linear, polynomial and radial basis function. The employed parameters for SVM models were $C=100$, $\varepsilon=10^{-5}$, $d=1$ for the polynomial kernel and $\gamma=1.5$ for the radial basis function. Regarding the ANN models, the Levenberg-Marquadt algorithm combined with the Newtonian gradient descent algorithm was used to adjust the connection weights and biases, and the number of hidden neurons varied from 3 to 10. Leave-one-out cross validation was

applied to evaluate the performance of the models. SVM outperformed the ANN and linear regression models, presenting the best performance measures (MAE = 0.39 t ha^{-1}, MRAE = 5.97%, RMSE = 0.47 t ha^{-1}, RRMSE = 7.19%, highest R^2 of 0.56). This study also found out that sunshine hours and daily temperature range were the variables that played critical roles in rice yield variability in the study area. The most relevant climatic factors in descending order for the rice yield prediction are sunshine hours > daily temperature range > rainfall > relative humidity > mean temperature > rainy days.

(Marini et al. 2004) use PCA and counterpropagation artificial neural networks (CP-ANN) to classify the variety of Italian rice samples based on 8 physical features regarding shape, uniformity and processing quality of the grains. The analyzed data set was composed of 1779 rice samples from 11 varieties harvested in Northern Italy. The first two components of PCA explained 66.12% of the total variance within the data, however the authors did not consider the achieved separations satisfactory. On the other hand, CP-ANN was able to correctly predict between 91 to 99% of the samples during several tests with different variations in the parameters until the optimal model was found. The results were compared to LDA, which presented an accuracy of 89.5% on the training set and 87.5% on the test set, hence showing lower performance than CP-ANN, which is as interesting fact since the LDA is a popular technique for the classification of rice samples for various purposes.

(Lu et al. 2015) used sensor data obtained from electronic nose and electronic tongue to classify conventional and hybrid indica rice. Analyzed samples, 60 conventional and 60 hybrid, were obtained in 2012 from three southern cities in China. PCA and locally linear embedding (LLE) were used to preprocess the data, while SVM and KNN were used as classification models. The

best model achieved was a LLE-SVM combined model: the accuracies achieved by the models on the calibration set composed of data obtained from electronic tongue, electronic nose and the combination of both were 91%, 98% and 98%, respectively, and those of the prediction set were 65%, 75% and 80%.

4. DISCUSSION

Multivariate data analysis and machine learning techniques are emerging solutions in recent literature for authentication and recognition of geographical origin of food and other products. Regarding the discrimination of rice, a few methods are gaining popularity due to their proven capability of yielding good results when combined with spectroscopy, image processing and other tools.

As we observed in the discussed academic studies, the majority of the recent works use molecular and atomic spectra as parameters for the discrimination of rice to achieve various goals. The elemental fingerprint of rice, which can be precisely determined by ICP-MS and other tools, is an especially promising source of information about the origin, variety, type and other features of rice, hence attracting special attention from researches that want to perform rice discrimination. Also, PCA and discriminant analysis algorithms are the most widely employed algorithms for the discrimination of rice. Although the goal of PCA is not exactly to perform data classification, it has proved to be an effective tool for visualizing the natural groupings inside the data and to compute the impact of the variables in their differentiation. Discriminant analysis produced very satisfactory results in most of the listed studies and considered parameters. These considerations lead us to conclude that, even in light of newer approaches and technologies for data classification in the recent literature, this methodology is still reliable and

adequate for the classification of rice, especially when combined with molecular and atomic spectroscopy and other molecular and cell features of rice.

Regarding newer solutions for data classification, only a few of the reviewed studies used machine learning techniques, which add artificial intelligence and mathematical optimization concepts in the processes of data classification and pattern discovery. The articles listed in this review show that machine learning techniques such as SVM, ANN, RF and others are capable of discriminating the geographical origin of rice with high accuracy based on its chemical composition (CITAR), targeted lipidomics (CITAR), Raman and terahertz spectra (CITAR) and other parameters. These algorithms were also capable of discriminating varieties of rice with high performance when the analyzed variables were physical features of the grains (CITAR), morphological traits of the grains obtained from image processing (CITAR), laser-induced fluorescence spectra (CITAR), sensor data from electronic tongue and nose (CITAR) and others. These methods also presented high performance for differentiating organic (CITAR) and transgenic rice (CITAR) from its conventional version. In addition to the studies discussed in this paper, there are several reports in the recent literature of machine learning and data mining algorithms being employed to analyze elemental fingerprints, spectra and other chemical data of other products, producing very good results (Fernández Pierna et al. 2004; Widjaja et al. 2008; Balabin et al. 2011; Alcázar et al. 2012; Reidy et al. 2013; Barbosa et al. 2014, 2015; Zain et al. 2016; Maione et al. 2016a, b, c). Hence, we expect that these techniques will receive wider consideration from researchers in forthcoming studies regarding the discrimination of rice.

The quality of the data set is another aspect to take into consideration in order to improve the results obtained from the application of multivariate data analysis and machine learning

techniques for any problem. Some studies work with very small data sets, with a very small number of samples available for each class, and it would be interesting to simulate the same methodologies with larger numbers of samples in order to improve the learning curve of the proposed classification functions. Moreover, many works discussed perform multiclass classification, especially those which address the discrimination of the varieties and geographical origin of rice. A similar amount of samples from each class label (in this case, the label which represent the variety or geographical origin) is required in order for the model to produce reliable generalizations. Unbalanced data sets tend to produce biased classifiers which can present high predictive accuracy for the majority class but poor predictive accuracy for the smaller class (Elkan and Noto 2008). A possible way to overcome this problem is to employ techniques such as Synthetic Minority Oversampling Technique (SMOTE) which can fill the data set with additional samples of the minority class until both classes are equally represented (Chawla et al. 2002).

Finally, we observed that only a few studies directed efforts to identifying the most important variables for the discrimination of the analyzed rice samples. This process is significant since analyzed data sets can often contain irrelevant or redundant variables, which can harm the performance of classification functions. Redundant variables present high dependence to other variables and the information contained in them can be expressed by fewer variables, and irrelevant variables do not hold information which can contribute to the generation of hypotheses regarding the samples with relation to their class labels. Discarding these variables can yield advantages such as improvement of prediction accuracy, dimensionality reduction, reduction of the time needed to build and run classification models, and others (Dash and Liu 1997; Guyon

and Elisseeff 2003). Feature evaluation methods such as Correlation Based Feature Selection (CFS) (Hall 2000), chi-square, F-score (Chen and Lin 2006) and others could be employed by forthcoming studies in order to check the most significant variables and also eliminate the least important ones.

5. CONCLUSION

In this study, we gather and discuss several recent works on the application of multivariate data analysis techniques to the discrimination of rice. The recognition of geographical origin and differentiation of varieties of rice are the main goals of the majority of the studies found, which highlights the global concern about authentication issues. Most of the analyzed rice samples were produced in Asian countries, especially China, Korea and Thailand, what is expected since about 90% of the world rice production is concentrated in this continent.

Among the most considered parameters for the analysis of rice, we catalogue data obtained from image processing, molecular and atomic spectroscopy, elemental fingerprinting, genetic markers, molecular content and others. PCA and LDA are the preferred techniques for analysis of rice by the majority of the researches, but several other data classification techniques, such as SVM, ANN, KNN, cluster analysis and others, are also present in some studies and show high performance for discrimination of rice. With the ascension of data mining and machine learning techniques, which are known for producing good results for classifying data from several knowledge areas, including discrimination of many types of food, we expect to see more of these algorithms employed in future studies regarding the discrimination of rice.

REFERENCES

Alcázar Á, Jurado JM, Palacios-Morillo A, et al (2012) Recognition of the geographical origin of

beer based on support vector machines applied to chemical descriptors. Food Control 23:258–262 . doi: 10.1016/j.foodcont.2011.07.029

Balabin RM, Safieva RZ, Lomakina EI (2011) Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines. Microchem J 98:121–128 . doi: 10.1016/j.microc.2010.12.007

Barbosa RM, Batista BL, Barião C V., et al (2015) A simple and practical control of the authenticity of organic sugarcane samples based on the use of machine-learning algorithms and trace elements determination by inductively coupled plasma mass spectrometry. Food Chem 184:154–159 . doi: 10.1016/j.foodchem.2015.02.146

Barbosa RM, Batista BL, Varrique RM, et al (2014) The use of advanced chemometric techniques and trace element levels for controlling the authenticity of organic coffee. Food Res Int 61:246–251 . doi: 10.1016/j.foodres.2013.07.060

Barbosa RM, Paula ES de, Paulelli AC, et al (2016) Recognition of organic rice samples based on trace elements and support vector machines. J Food Compos Anal 45:95–100

Batista BL, da Silva LRS, Rocha BA, et al (2012) Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. Food Res Int 49:209–215 . doi: 10.1016/j.foodres.2012.07.015

Batista BL, Souza JMO, Souza SS de, Junior FB (2011) Speciation of arsenic in rice and estimation of daily intake of different arsenic species by Brazilians through rice consumption. J Hazard Mater 191:342–348

Batres-Marquez SP, Jensen HH, Upton J (2009) Rice Consumption in the United States: Recent

Evidence from Food Consumption Surveys. J Am Diet Assoc 109:1719–1727

Borges EM, Gelinski JMLN, Souza VC de O, et al (2015) Monitoring the authenticity of organic rice via chemometric analysis of elemental data. Food Res Int 77:299–309

Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 16:321–357

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

Chen H, Wu W, Liu H-B (2016) Assessing the relative importance of climate variables to rice yield variation using support vector machines. Theor Appl Climatol 126:105–111

Chen Y-W, Lin C-J (2006) Combining SVMs with Various Feature Selection Strategies. In: Feature Extraction. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 315–324

Chung I-M, Kim J-K, Lee J-K, Kim S-H (2015) Discrimination of geographical origin of rice (Oryza sativa L.) by multielement analysis using inductively coupled plasma atomic emission spectroscopy and multivariate analysis. J Cereal Sci 65:252–259

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

Cubadda F (2007) Chapter 19 – Inductively coupled plasma mass spectrometry. In: Picó Y (ed) Food Toxicants Analysis: Techniques, Strategies and Developments, 1st edn. Elsevier, pp 697–751

D'Haen J, Van den Poel D, Thorleuchter D (2013) Predicting customer profitability during

acquisition: Finding the optimal combination of data source and data mining technique. Expert Syst Appl 40:2007–2012 . doi: 10.1016/j.eswa.2012.10.023

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156 . doi: 10.1016/S1088-467X(97)00008-5

Davis MA, Mackenzie TA, Cottingham KL, et al (2012) Rice Consumption and Urinary Arsenic Concentrations in U.S. Children. Environ Health Perspect 120:1418–1424

Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada, USA, pp 213–220

Feng X, Zhang Q, Cong P, Zhu Z (2013) Preliminary study on classification of rice and detection of paraffin in the adulterated samples by Raman spectroscopy combined with multivariate analysis. Talanta 115:548–555

Fernández Pierna JA, Baeten V, Renier AM, et al (2004) Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. J Chemom 18:341–349 . doi: 10.1002/cem.877

Gilbert-Diamond D, Cottingham KL, Gruber JF, et al (2011) Rice consumption contributes to arsenic exposure in US women. Proc Natl Acad Sci U S A 108:20656–20660

Gonzalvez A, Armenta S, Guardia M de la (2011) Geographical traceability of "Arròs de Valencia" rice grain based on mineral element composition. Food Chem 126:1254–1260

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hall MA (2000) Correlation-based Feature Selection for Discrete and Numeric Class Machine

Learning. In: Pat Langley (ed) Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000). Morgan Kaufmann Publishers Inc., Stanford, CA, USA, pp 359–366

Hu EA, Pan A, Malik V, Sun Q (2012) White rice consumption and risk of type 2 diabetes: meta-analysis and systematic review. BMJ 344:

Huo Y, Kamal GM, Wang J, et al (2017) 1H NMR-based metabolomics for discrimination of rice from different geographical origins of China. J Cereal Sci 76:243–252

Hwang J, Kang S, Lee K, Chung H (2012) Enhanced Raman spectroscopic discrimination of the geographical origins of rice samples via transmission spectral collection through packed grains. Talanta 101:488–494

Kambo R, Yerpude A (2014) Classification of Basmati Rice Grain Variety using Image Processing and Principal Component Analysis. Int J Comput Trends Technol 11:2893–2900

Kuo T-Y, Chung C-L, Chen S-Y, et al (2016) Identifying rice grains using image analysis and sparse-representation-based classification. Comput Electron Agric 127:716–725

Lee I, We GJ, Kim DE, et al (2012) Classification of rice cultivars based on cluster analysis of hydration and pasting properties of their starches. LWT - Food Sci Technol 48:164–168

Li G, Nunes L, Wang Y, et al (2012) Profiling the ionome of rice and its use in discriminating geographical origins at the regional scale, China. J Environ Sci 25:144–154

Lim DK, Long NP, Mo C, et al (2017) Combination of mass spectrometry-based targeted lipidomics and supervised machine learning algorithms in detecting adulterated admixtures of white rice. Food Res Int 100:814–821

Liu W, Liu C, Hu X, et al (2016) Application of terahertz spectroscopy imaging for

discrimination of transgenic rice seeds with chemometrics. Food Chem 210:415–421

Lu L, Deng S, Zhu Z, Tian S (2015) Classification of Rice by Combining Electronic Tongue and Nose. Food Anal Methods 8:1893–1902

Maione C, De Paula ES, Gallimberti M, et al (2016a) Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. Expert Syst Appl 49:60–73 . doi: 10.1016/j.eswa.2015.11.024

Maione C, Lemos B, Dobal A, et al (2016b) Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. Comput Electron Agric 121:101–107 . doi: 10.1016/j.compag.2015.11.009

Maione C, Souza VC de O, Togni LR, et al (2016c) Establishing chemical profiling for ecstasy tablets based on trace element levels and support vector machine. Neural Comput Appl 1–9 . doi: 10.1007/s00521-016-2736-3

Maione C, Turra C, Fernandes EADN, et al (2017) Finding the Most Significant Elements for the Classification of Organic Orange Leaves: A Data Mining Approach. Anal Lett 50:2292–2307

Marini F, Zupan J, Magrì AL (2004) On the use of counterpropagation artificial neural networks to characterize Italian rice varieties. Anal Chim Acta 510:231–240

Muthayya S, Hall J, Bagriansky J, et al (2012) Rice fortification: an emerging opportunity to contribute to the elimination of vitamin and mineral deficiency worldwide. Food Nutr Bull 33:296–307

Muthayya S, Sugimoto JD, Montgomery S, Maberly GF (2014) An overview of global rice production, supply, trade, and consumption. Ann N Y Acad Sci 1324:7–14

Pramai P, Jiamyangyuen S (2016) Chemometric classification of pigmented rice varietiesbased on antioxidative properties in relation to color. Songklanakarin J Sci Technol 38:463–472

Promchan J, Günther D, Siripinyanond A, Shiowatana J (2016) Elemental imaging and classifying rice grains by using laser ablation inductively coupled plasma mass spectrometry and linear discriminant analysis. J Cereal Sci 71:198–203

Reidy L, Bu K, Godfrey M, Cizdziel J V. (2013) Elemental fingerprinting of soils using ICP-MS and multivariate statistics: A study for and by forensic chemistry majors. Forensic Sci Int 233:37–44

Sabir A, Raf M, Darusman LK (2017) Discrimination of red and white rice bran from Indonesia using HPLC fingerprint analysis combined with chemometrics. Food Chem 221:1717–1722

Shao Y, Zhao C, Bao Y, He Y (2012) Quantification of Nitrogen Status in Rice by Least Squares Support Vector Machines and Reflectance Spectroscopy. Food Bioprocess Technol 5:100–107

Singh KR, Chaudhury S (2016) Efficient technique for rice grain classification using back-propagation neural network and wavelet decomposition. IET Comput Vis 10:780–787

Widjaja E, Zheng W, Huang Z (2008) Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. Int J Oncol. doi: 10.3892/ijo.32.3.653

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2:37–52

Xia J, Hu X, Shi F, et al (2010) Support vector machine method on predicting resistance gene against Xanthomonas oryzae pv. oryzae in rice. Expert Syst Appl 37:5946–5950

Yang J, Sun J, Du L, Gong W (2017) Monitoring of Paddy Rice Varieties Based on the

Combination of the Laser-Induced Fluorescence and Multivariate Analysis. Food Anal Methods 10:2398–2403

Zain SM, Behkami S, Bakirdere S, Koki IB (2016) Milk authentication and discrimination via metal content clustering – A case of comparing milk from Malaysia and selected countries of the world. Food Control 66:306–314 . doi: 10.1016/j.foodcont.2016.02.015

Zareiforoush H, Minaei S, Alizadeh MR, Banakar A (2016) Qualitative classification of milled rice grains using computer vision and metaheuristic techniques. J Food Sci Technol 53:118–131

Zhang C, Li J, Zhu Z, et al (2010) Cluster Analysis on Japonica Rice (Oryza sativa L.) with Good Eating Quality Based on SSR Markers and Phenotypic Traits. Rice Sci 17:111–121

TABLES

Table 1. Summary of recent studies on discrimination of rice using multivariate data analysis and machine learning techniques. Aim of the study, properties analyzed and methods used are detailed.

Aim of the study	Samples analyzed	Properties analyzed	Multiv
Study of good quality of japonica rice.	Sixty high-quality conventional japonica rice accessions obtained in China and abroad	Genetic similarity matrix detected by simple sequence repeat (SSR) markers, 26 phenotypic values of agronomic traits, and 13 taste phenotypic traits.	Cluster a unweigh arithmeti method.

Prediction of proteins resistant to Xoo in rice.	Thirteen proteins (cloned genes) in rice for positive data, obtained from NCBI, and 48 proteins (selective genes) in rice for negative data, obtained from KOME.	Twenty residues, 400 dipeptides and 24 other features obtained from chaos game representation (CGR).	SVM
Differentiation of conventional and hybrid indica rice.	One hundred and twenty indica rice samples (60 conventional, 60 hybrid) harvested in 2012 from three southern cities in China. A hundred samples were used to train/calibrate, 20 were used to test.	Sensor data obtained from electronic tongue and electronic nose.	PCA and the data, discrimi
Discrimination of varieties of Italian rice.	1779 rice samples from 11 varieties harvested during the period of 1995-1998 from northern Italy.	Eight physical features regarding shape, uniformity and processing quality of the grains: average length, average width, average length to width ratio, coefficient of variation of length, coefficient of variation of width, coefficient of variation of length to width ratio, total milling yield and whole-grain milling yield.	PCA and ANN.
Study of the impact of climate variables to rice yield.	Rice yield data obtained from the Chongqing Statistical Yearbook which is published by the Chongqing Bureau of Statistics.	Six climate parameters: temperature, daily temperature range, relative humidity, rainfall, rainy days, sunshine hours.	SVM, A
Authentication of Brazilian organic rice	Eighteen organic rice samples and 32 conventional rice samples obtained from Brazilian markets.	Concentration levels of 20 chemical elements determined in the rice samples using ICP-MS: As, B, Ba, Ca, Cd, Ce, Co, Cr, Cu, Fe, K, La, Mg, Mn, Mo, P, Pb, Rb, Se and Zn.	PCA, SI KNN.
Authentication of Brazilian organic rice	Seventeen organic rice samples and 33 conventional rice samples of 50 different brands obtained from stores in RS, SC, MG, GO and TO states, Brazil.	Concentration levels of 19 chemical elements determined in the samples using q-ICP-MS: As, B, Ba, Ca, Cd, Ce, Cr, Co, Cu, Fe, La, Mg, Mn, Mo, P, Pb, Rb, Se and Zn.	SVM

Recognition of geographical origin of Chinese rice.	Two hundred and six paddy rice samples obtained from 9 different regions of China: Fuzhou (7), Longyan (23), Nanping (31), Ningde (17), Putian (12), Quanzhou (36), Sanming (36), Xiamen (5), Zhangzhou (39).	Concentration levels of 13 chemical elements determined in the rice samples via ICP-MS/OES.	PCA, DF index an
Recognition of the geographical origin of Spanish rice from various protected designations of origins.	One hundred fifty-three rice samples obtained during 2007-2009 from 4 different countries. One hundred twenty-seven samples were Spanish, being 67 from Valencia ("Arrós de Valencia"), 30 from Tarragona (22 from which certified as "Arrós del Delta del Elbro"), 11 from Murcia ("Arrós de Calasparra"), 19 from Extramadura. The remaining samples were produced in Japan, Brazil and India.	Concentration levels of 32 chemical elements determined by ICP-OES: Al, As, Ba, Bi, Cd, Ca, Cr, Co, Cu, Fe, Pb, Li, Mg, Mn, Mo, Ni, K, Se, Na, Sr, Tl, Ti, Zn, La, Ce, Pr, Nd, Sm, Eu, Ho, Er and Yb.	LDA
Recognition of the geographical origin of Thai rice.	Thirty-one Thai rice samples obtained from three regions of Thailand and 5 rice samples obtained from foreign countries (France, Japan, Italy, India and Pakistan).	Concentration levels of 21 chemical elements determined by high resolution ICP-MS.	PCA and
Elemental imaging and recognition of the geographical origin and type of rice	For elemental imaging: white rice samples from Northeast of Thailand. For classification: 8 white rice samples from Northeast Thailand, and 5 white, 1 black, 1 red and 1 yellow rice samples from South Thailand.	Concentration levels of 20 chemical elements determined in the rice samples using ICP-MS: Al, As, Br, Ca, Cd, Cl, Co, Cu, Fe, Hg, K, Mg, Mn, Mo, Na, P, Pb, Rb, S, Zn.	LDA
Recognition of the geographical origin of Asian rice.	Fifty-nine rice samples obtained from Cambodia (n=14), China (n=6), Japan (n=10), Korea (n=12), Philippines (n=13) and Thailand (n=4) in 2015.	Concentration levels of 25 chemical elements determined by ICP-MS and light-element (C, N, O, S) isotope ratios determined by IRMS.	PCA and

Classification of 12 different rice cultivars from Korea	Rice grains from the cultivars grown at the National Institute of Crop Science, Korea, in 2009.	Starch processability indicators such as hydration and pasting parameters and amylose content.	HCA via
Discrimination of varieties of Thai pigmented rice.	Twenty rice samples including white, red and black varieties, obtained between June and October 2011 from different producing locations in northern Thailand.	Color parameters and antioxidants (total anthocyanins, phenolic and flavonoid contents, DPPH radical-scavenging ability, ferric antioxidant power, trolox equivalent antioxidant capacity, α-tocopherol, γ-oryzanol).	PCA and
Detection of adulterated admixtures of Korea white rice	Three batches of white rice from Korea, China and seven mixing ratios.	Seventeen lysoGPLs (lysoglycerophospholipids) available in the samples: 6 lysophosphatidylcholines, 7 lysophosphatidylethanolamines and 4 lysophosphatidylglycerols.	SVM, K decision
Discrimination of red and white rice from Indonesia.	Twenty-nine rice brans consisted of 9 red rice brans and 20 white rice brans obtained from several districts in South Sulawesi and West Java, Indonesia, in 2015.	Compounds determined by HPLC chromatography and diode array detector fingerprint such as cycloartenol ferulate, cyclobranol ferulate, campesterol ferulate and β-sitosterol ferulate.	PCA and
Discrimination of the Basmati rice grain variety	Rice grains obtained from supermarket	Morphological data obtained by image processing: area, major and minor axis lengths, eccentricity and perimeter.	PCA and
Discrimination of four varieties of rice	Four hundred images obtained from rice grains of the four varieties in different lighting conditions and quality of the camera.	Eighteen color features, 27 texture features, 24 wavelet features and 45 combined features (color and texture).	Back-pro and KNN
Discrimination of rice varieties	One thousand five hundred rice grains from 30 different varieties obtained from Genetic Stocks Oryza germplasm collection.	12 morphological traits, 9 color traits, 7 texture traits and 20 Fourier descriptors obtained from image analysis of the grains.	SRC (Sp Based cl
Discrimination of Hashemi rice grains into four quality grades	Images of milled rice grains of Hashemi variety obtained from the North region of Iran.	Five color and shape traits, 4 textural traits, and 48 variables based on color information, obtained from image segmentation.	Decision Bayesian selection CFS.

Recognition of the geographical origin of Korean rice samples.	Thrity imported and 30 domestic polished rice samples provided by the National Agricultural Products Quality Management Service, Seoul, Korea. Imported rice samples came mostly from diverse regions in China.	Back-scattering and transmission measurements obtained from enhanced Raman spectroscopy.	PCA and
Recognition of the geographical origin of Chinese rice and detection of paraffin in adulterated samples	Forty-two rice samples from indica and japonica varieties obtained from various regions of China.	Data obtained from Raman spectroscopy.	PCA, SI PLS-DA
Recognition of the geographical origin of Chinese rice.	One hundred six rice samples produced in 2015 in nine provinces of China: Guangxi, Heilongjiang, Hunan, Hainan, Liaoning, Ningxia, Xinjiang, Yunnan, Zhejiang.	Data obtained from 1H NMR spectroscopy.	PCA and
Determination of nitrogen levels in rice via classification of soil plant analysis development value (SPAD).	Sixty-four rice samples. Source and variety of the samples were not mentioned by the authors.	Canopy spectral reflectance data extracted from visible and near infrared spectroscopy.	LS-SVM Indepen analysis determin meaning
Discrimination of six different varieties of indica rice.	Paddy rice samples cultivated in 2016 in Huanggang County, China, from six different varieties: victory indica, manley indica, cucumber indica, guangchang ai, ii-youg 838, yangdao 2.	Data obtained from laser-induced fluorescence spectroscopy.	PCA and
Authentication of transgenic Chinese rice seeds	Two hundred rice seed samples, including transgenic and conventional, obtained from Institute of Rice Research, China.	Genotype data obtained from terahertz spectroscopy imaging, pre-treated with the first and second derivative and standard normal variate transformation.	PCA, LS and rand

