Accepted Manuscript

Classification of cowpea beans using multielemental fingerprinting combined with supervised learning

Michael Pérez-Rodríguez, José E. Gaiad, Melisa J. Hidalgo, María V. Avanza, Roberto G. Pellerano

PII:	S0956-7135(18)30395-5
DOI:	10.1016/j.foodcont.2018.08.001
Reference:	JFCO 6265
To appear in:	*Food Control*
Received Date:	18 June 2018
Accepted Date:	01 August 2018

Please cite this article as: Michael Pérez-Rodríguez, José E. Gaiad, Melisa J. Hidalgo, María V. Avanza, Roberto G. Pellerano, Classification of cowpea beans using multielemental fingerprinting combined with supervised learning, *Food Control* (2018), doi: 10.1016/j.foodcont.2018.08.001

Classification of cowpea beans using multielemental fingerprinting combined with supervised learning

Michael Pérez-Rodríguez*, José E. Gaiad, Melisa J. Hidalgo, María V. Avanza, Roberto G. Pellerano

Institute of Basic and Applied Chemistry of the Northeast of Argentina (IQUIBA-NEA), National Scientific and Technical Research Council (CONICET), Faculty of Natural and Exact Sciences and Surveying, National University of the Northeast – UNNE, Av. Libertad 5470, Corrientes 3400, Argentina.

* Corresponding author. Phone: +54 379 445 7996

E-mail address: michaelpr1984@gmail.com (M. Pérez-Rodríguez).

ABSTRACT

Multielemental compositions (Ag, As, Ba, Be, Cd, Cs, Co, Cr, Cu, Mo, Ni, Pb, Sb, Se, Sn, Sr, Tl, Rb, V, and Zn) of 106 cowpea bean samples belonging to different varieties collected from the province of Corrientes in Argentina were determined using inductively coupled plasma mass spectrometry (ICP-MS). Based on the multielemental data, five supervised learning techniques, namely, linear discriminant analysis (LDA), partial least square discriminant analysis (PLS-DA), k nearest neighbors (k-NN), random forest (RF), and support vector machine (SVM) with radial basis function Kernel, were computed aiming at building classification models that allow one to predict the botanical variety of the samples based on their element profiles. The best classification performance was obtained by SVM with 93% accuracy rate. The model developed through this method enabled the correct separation of the samples into the five cowpea varieties investigated, where 100% sensitivity was achieved for most of the predicted classes. Thus, SVM was the algorithm selected for the classification of the cowpea beans according to their botanical variety. Multielemental determination coupled with supervised pattern recognition techniques have proved to be an interesting approach for differentiating a diverse range of cowpea genotypes. This study has contributed toward generalizing the use of multielemental fingerprinting as a promising tool for testing the authenticity of cowpea beans on a global scale.

Keywords: Cowpea bean; multielemental fingerprinting; genotype; supervised learning; ICP-MS; authenticity.

1. Introduction

Grain legumes are an important source of essential nutrients in several developing countries owing to nutritional and socioeconomic reasons. These edible plants are regarded as the main protein source for populations in Asia, Africa and Latin America (Kato, Fernandes, Bacchi, Sarriés, & Reyes, 2015; Santos, Santos, Fernandes, Castro, & Korn, 2013). Within the Leguminosae family, cowpea beans (*Vigna unguiculata* L. Walp.) contain good contents of vitamins, fibers, minerals, and other nutrients that are vital for the normal metabolic functioning of the human body (Alayande, Mustapha, Dabak, & Ubom, 2012). In addition, phenolic compounds, which are known for their use in the treatment and preventive action against various pathologies, such as diabetes (Asgar, 2013), arterial hypertension (Souza, Marcadenti, & Portal, 2017), and cardiovascular diseases (Scolaro, Kim, & Castro, 2018), are found to be present in the composition of cowpea beans (Moreira-Araújo et al., 2018).

The presence of multielemental contents in cowpea products may stem from several factors. The contents might come from the soil, environment, botanical variety (genotype) or may be introduced during the production process (crop cultivation, transport, storage, or preparation) (Kato et al., 2015; Santos, Gramacho, & Teixeira, 2008). Trace elements quantification in legumes may also reveal the presence of inorganic contaminants which are dangerous to human health (Zhou et al., 2016). In view of that, several international organizations have set tolerance levels for toxic metals in leguminous seeds (MERCOSUR 2012; Codex Alimentarius, 2011; European Community, 2006). Considering the presence of such harmful substances in grain legumes, performing an exhaustive quality control is essentially important to ensure human food safety.

A huge number of authors have published studies in the literature regarding the nutritional composition of different cowpea genotypes grown in South Africa (Belane &

Dakora, 2011), West Africa (Alayande, Mustapha, Dabak, & Ubom, 2012; Inobeme, Nlemadim, Obigwa, Ikechukwu, & Ajai, 2014), Argentina (Avanza, Acevedo, Chaves, & Añón, 2013), and Brazil (Santos et al., 2008, 2013, 2009, 2008). Predominantly, multielemental analysis techniques with high sensitivity and ability to determine the isotope composition of a sample were used in these studies. In this context, inductively coupled plasma mass spectrometry (ICP-MS) has proved to be a rapid and high-accuracy technique for simultaneous determination of several elements. The outstanding properties of this technique are reflected in its conceivably superior performance when it comes to conducting reliable detection and quantification of trace elements in a wide variety of samples (Ammann, 2007). Furthermore, the information obtained about the multielemental composition of samples can be analyzed by pattern recognition techniques that provide one with powerful tools for identifying geographical origin, quality grade, or genetic characteristics of foods (Liu, Xue, Wang, Li, Xue, & Xu, 2012).

The present work focuses on characterizing cowpeas using ICP-MS based multielemental fingerprinting (Ag, As, Ba, Be, Cd, Cs, Co, Cr, Cu, Mo, Ni, Pb, Sb, Se, Sn, Sr, Tl, Rb, V, and Zn) aiming at identifying differences among the cultivated varieties and check the authenticity of the beans through the application of machine learning techniques. To this end, the study applied some pattern recognition techniques, such as, principal component analysis (PCA), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), k nearest neighbors (k-NN), random forest (RF), and support vector machines (SVM), on multielemental data measured by ICP-MS for bean seeds produced in the northeastern region of Argentina (Corrientes Province). The combined benefits of analytical and chemometric techniques studied in this work were tested as a strategic tool for the differentiation of a wide range of cowpea genotypes, which exhibited the potential to meet the quality regulatory requirements for this food.

2. Materials and methods

2.1. Instrumentation

A Spex 6750 (Metuchen, NJ, USA) cryogenic mill was used to reduce the particle sizes of the bean samples, while decomposition was performed using a microwave digestion system, Ethos One (Milestone, Chicago, USA) equipped with programmable power control (1600 W maximum power) and HPR 1000/10s segmented rotor (operating conditions of 35 bar maximum pressure and 260 °C maximum temperature) with 10 reaction vessels.

Multielemental fingerprinting was determined using an Agilent 7700 Series (Agilent Technologies, Japan) inductively coupled plasma mass spectrometer. The instrument was equipped with a cooled double-pass quartz spray and a MicroMist glass concentric nebulizer. The Fassel-type ICP torch was constituted by three-cylinder assembly with injector diameter of 2.5 mm. Ni sampler and skimmer cones of 1.0 mm and 0.4 mm were used. The ICP-MS operating conditions are summarized in Table 1. Internal standards were used in all determinations in order to correct interferences stemming from the sample matrices. In order of mass number, the selected isotopes for measurement included the following: ^{9}Be, ^{51}V, ^{53}Cr, ^{59}Co, ^{60}Ni, ^{63}Cu, ^{66}Zn, ^{75}As, ^{78}Se, ^{85}Rb, ^{88}Sr, ^{98}Mo, ^{107}Ag, ^{111}Cd, ^{118}Sn, ^{121}Sb, ^{133}Cs, ^{137}Ba, ^{205}Tl, and ^{208}Pb.

Insert Table 1

2.2. Chemicals and standard solutions

All chemical reagents employed were of ultrapure analytical grade. Nitric acid (65% m/m) and hydrogen peroxide (30% m/m) were purchased from Merck (Darmstadt, Germany). Nitric acid was further purified by sub-boiling distillation for later use. All standard and working solutions were prepared in deionized water (18.0 MΩ cm at 25 °C) obtained from a Milli-Q Plus Water purification system (Millipore Corp., Molsheim, France). Multielemental calibration solutions were prepared from multi-element standard solution, TraceCERT® CRM, purchased from Sigma-Aldrich (St. Louis, MO, USA). For assessing the accuracy of analytical method, a plant based SRM tomato leaves (NIST® SRM 1573a) purchased from Sigma-Aldrich (St. Louis, MO, USA) was used.

Materials, including plastic containers, polyethylene flasks, pipette tips, and PFA Teflon digestion vessels, were constantly checked for contaminations by blank quality control tests, and glassware were strictly avoided.

2.3. Bean samples

A total of 106 cowpea bean samples were analyzed so as to determine their mineral contents. Different agricultural cooperatives from the province of Corrientes (Argentina) provided us with cowpeas belonging to five botanical varieties, namely: Alarcon (ALA), California Black (CBK), Carmevalia Ensiformis (CES), Moro Riach (MRH), and San Miguel (SML). The samples were obtained between September and October 2016–2017. After their collection, the samples were immediately stored in zipped bags in a vacuum desiccator cabinet until analysis.

2.4. Sample preparation

The bean samples were homogenized in a cryogenic mill, and approximately 500 mg of dry samples were placed in closed vessels in a microwave oven, where 5 mL HNO_3 and 2 mL H_2O_2, both in concentrated solution, were added later. The procedure was completed by using the following temperature program: first stage: 25–200 °C for 10 min; second stage: 200 ºC for 15 min; and third stage: 200–110 ºC for 10 min; followed immediately by ventilation at room temperature for 10 min. Finally, the digested samples were diluted to 10 mL with deionized water and prepared for analysis by ICP-MS. Blank solutions were prepared based on the same procedure applied for the samples. To correct any instrument signal drift during analysis and suppress mistakes in the analyte quantification, an internal standard solution (Sigma-Aldrich, St. Louis, MO, USA) containing ^{72}Ge and ^{115}In at 100 μg L^{-1} was added to all samples prior to the digestion stage. All measurements were performed in triplicate.

2.5. Analytical performance parameters

The analytical features of the proposed ICP-MS method were measured through performance parameters, such as coefficient of determination, limit of quantification (LOQ), accuracy, precision and overall recovery.

To construct the multi-element calibration curves, five different concentration levels in triplicate were used, and the calibration ranges were modified according to the expected mineral concentration ranges. Linear regression analysis by the least squares method was used to calculate the coefficients of determination (r^2). The limits of quantification (LOQ) were determined as ten times the standard deviation of measurements of 10 blank solutions, divided by the slope of the calibration curve, according to the IUPAC recommendations (Thompson, Ellison, & Wood, 2002).

Precision assays were conducted under conditions of repeatability and intermediate precision over a period of three weeks, as such the samples under study were digested and analyzed in triplicate in different days. The variability of the measurement was expressed as the relative standard deviation (%RSD).

The accuracy of the method was assessed using replicate analysis of 5 different samples of standard reference materials (SRM 1573a). In addition, recovery studies were carried out by fortifying randomly selected bean samples with multi-element standard solutions at concentrations of 10 and 100 μg kg^{-1}. The solutions, including the blank solutions, the fortified solutions, and the certified material solutions, were all prepared following the same procedure as the samples. One limitation of this study is that the contents of some elements such as, Ag, Ba, Be, Cs, Mo, Pb, Sn, Sr and Tl were not certified in the NIST SRM 1573a. Therefore, only fortification results for these elements are available.

2.6. Multivariate data analysis

The data matrix consisted of 106 rows, which corresponded to different cowpea bean samples, and 20 columns containing the mineral concentration values (Ag, As, Ba, Be, Cd, Cs, Co, Cr, Cu, Mo, Ni, Pb, Sb, Se, Sn, Sr, Tl, Rb, V, and Zn). Multivariate data processing was carried out via machine learning techniques using caret package in R-project software version 3.3.3 (R Core Team, 2017).

Principal component analysis (PCA) was used for exploratory data analysis aiming at visualizing the natural distribution of samples in a reduced dimensional space. The dimension reduction in PCA allows one to accurately represent high-dimensional data in lower dimensional space. It is worth noting that this unsupervised learning method also verifies the existence of relationships between the variables in a multidimensional space (Bro & Smilde, 2014; Varmuza & Filzmoser, 2009). Unsupervised learning focuses on describing the associations and patterns among a set of input measures, revealing properties of the data density (Hastie, Tibshirani, & Friedman, 2008).

After that, five supervised learning techniques were applied to the dataset for the predictive modeling of botanical varieties of cowpea beans. To achieve this goal, two linear methods, including linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLS-DA), and three non-linear models, namely, k nearest neighbors (k-NN), support vector machine (SVM) and random forest (RF) were compared according to their classification performance. Supervised learning focuses on predicting the value of an outcome measure based on a number of input measures. The predictions are based on the training samples of previously solved cases (samples), where the joint values of all of the variables are known. Thus, a model is trained to generate reasonable classifications for new data (Hastie et al., 2008).

During the training step, the parameters needed to build optimal classifiers were optimized by ten-fold cross-validation so as to avoid the occurrence of bias. For this purpose, the data matrix was randomly split into 10 mutually exclusive subsets, and each classifier was trained and tested 10 times, so that the cross-validation estimate of metrics is over-tested 10 times. This procedure is repeated five times, so each subset is used to test the model developed at least one time. The parameters that were subjected to optimization included the number of significant components (ncomp) for PLS-DA, number of k neighbors for k-NN, number of trees (nt) and number of variables tried at each split (mtry) for RF, in addition to penalty factor (C), γ of the γ-insensitive loss function and kernel type for SVM.

After selecting the optimal values for each model, some basic measures derived from the confusion matrix were considered in order to evaluate the classification obtained using the supervised algorithms under investigation (LDA, PLS-DA, k-NN, RF and SVM). The metrics calculated per class included the following: sensitivity (correct positive predictions divided by the number of positive cases), specificity (correct negative predictions divided by the number of negative cases), and overall accuracy (all correct predictions divided by the total number of cases examined) (Lantz, 2015).

LDA classification model is used for differentiating classes of samples by minimizing the variance within classes and maximizing the variance among classes. This method is based on the estimation of several canonical or discriminant functions, which are linear combinations of the original variables in order to optimize the separation (Moncayo, Manzoor, & Caceres, 2015).

PLS-DA is a linear classification algorithm based on partial least squares regression used for conducting predictive modeling. This technique allows class distinction, which enables one to find latent variables from dependent variables with maximum covariance (Brereton & Lloyd, 2014).

K-NN is a distance-based non-parametric discriminant technique. This method uses the distance between objects to assign one of them to the most common class among the k-nearest neighbors. The optimal size of neighbor k must be optimized (Bevilacqua et al., 2013).

RF is a supervised learning method based on a set of tree decision predictors for classification, which was created as an extension of bagging. In this method, a random vector is generated independently on the input vector, and each tree casts a vote meant for classifying an input vector (Archer & Kimes, 2008; Hernández-Pereira, Álvarez-Estévez, & Moret-Bonillo, 2015).

SVM is a machine learning technique that performs a mapping of the training data in a high-dimensional space directed toward building a classifier in that space (Batista et al., 2012). This method classifies the data by constructing a separate hyperplane in n-dimensional space, which maximizes the margin between classes (Bona et al., 2017).

Aside from the calculated metrics, Kappa coefficient of agreement was likewise considered to enable us to compare the performance of the proposed classification models, since the cowpea bean dataset is unbalanced in terms of the sample number per group (variety). This robust statistic is commonly used either for evaluating inter-classifier or intra-classifier reliability (McHugh, 2012). In addition, Kappa statistical measure is a form of correlation coefficient, so that squaring the correlation value facilitates its interpretation. Squared Kappa is referred to as coefficient of determination; it is defined as the amount of variation in the dependent variable that can be explained by the independent variable (Stephens & Diesing, 2014; Tang, Hu, Zhang, Wu, & He, 2015). A detailed description regarding the calculation of the kappa statistic can be found in the work published by McHugh, 2012 (McHugh, 2012).

3. Results and discussion

3.1. Analytical features of the ICP-MS method under investigation

Table 2 summarizes the analytical features of the method studied in this work. The calibration curves obtained exhibited good linearity in the selected concentration range for each element, with coefficients of determination (r^2) greater than 0.9985. The LOQs ranged from 2 to 545 µg kg^{-1}. These values indicated that the method proposed in this work was sufficiently sensitive for determining trace elements in cowpea beans.

Insert Table 2

The intra-day and inter-day precisions for the elements measured were within the ranges of 0.2–4.8% and 1.5–6.9%, respectively. Accuracy assays performed by analyzing SRM 1573a showed good agreement between the results obtained by the present method and the certified values. The average recoveries of the elements ranged from 93% to 107%, with RSD values being less than 12% (Table 2). Furthermore, overall recoveries in fortified samples were in the range of 86.2–109.2%. According to the Codex Alimentarius recommendations (Codex Alimentarius, 2009), these results are within the acceptable criteria for recovery (70–110 %) and intermediate precision (≤15% RSD). In effect, the results clearly demonstrate that the performance of the method proposed here was satisfactory and that the microwave digestion step was essentially effective in contributing toward obtaining reliable results with adequate accuracy and precision.

3.2. Multielemental fingerprinting of cowpea bean samples

The applicability of the present ICP-MS method was checked through the determination of multi-elements in *Vigna unguiculata* L. Walp. samples. Table 3 shows the element concentrations obtained for the different cowpea bean varieties under study. The results are expressed as average values of three measurements with the corresponding standard deviation (SD).

Insert Table 3

Based on the results obtained, Ba and Zn were found to be the most abundant elements in the cowpea beans investigated, with values ranging from 19.80–96.60 and 17.34–60.27 mg kg^{-1}, respectively. Ba concentrations were higher than the reported values for cowpea beans (0.2–12 mg kg^{-1}) (Santos et al., 2013). Zn presented values similar to those found in different cowpea genotypes (44–65 mg kg^{-1}) (Belane & Dakora, 2011).

In terms of abundance, the elements that came next included Cu, Sr, and Mo. Cu content ranged from 2.60 to 7.90 mg kg^{-1}, whereas Sr exhibited concentrations ranging from 1.50 to 8.12 mg kg^{-1}. These elements showed relatively greater variability compared to those found in other cowpea varieties (3.1–5.8 mg kg^{-1} for Cu and 3.3–4.8 mg kg^{-1} for Sr) (Santos et al., 2013). With regard to Mo, MRH presented the highest average content followed by CBK, SML, and CES; while ALA exhibited an average value that was farthest from the rest of the varieties. In general, Mo contents in the beans investigated here were relatively lower than those observed in common beans (*Phaseolus vulgaris* species) by other authors (1.3–5.4 mg kg^{-1}) (Kato et al., 2015).

The third group that came next in terms of abundance consisted of Ni, Pb, Sn, Be, Rb, Cr, V, and Tl, in decreasing concentration order, with average values ranging from 0.09–0.67 mg kg^{-1}. The highest Ni concentration was found in the SML variety, whereas the lowest was obtained in the ALA variety. The concentrations of this element were lower compared to those reported by Santos et al. in common beans (0.75–6.7 mg kg^{-1}) and cowpeas (2.9–3.4 mg kg^{-1}) (Santos et al., 2013). Pb concentrations ranged from 0.14 to 0.52 mg kg^{-1} for the cowpea varieties investigated here, with the highest values observed in beans from CBK and lowest values in beans from MRH. The overall Pb concentration of 0.28 mg kg^{-1} obtained was slightly above the tolerance limit in dried legumes, which is 0.20 mg kg^{-1}, set by the MERCOSUR (MERCOSUR, 2012), FAO/WHO (Codex Alimentarius, 2011), and European Commission (European Community, 2006). The concentrations of Sn and Be were found to be lowest in beans from the ALA variety, while being highest for the MRH variety; the average concentration of Sn (0.26 mg kg^{-1}) was found to be well below the allowable maximum limit in canned foods according to the regulations of MERCOSUR (250 mg kg^{-1}) (MERCOSUR, 2012) and the European Community (200 mg kg^{-1}) (European Community, 2006). The overall concentration values found for Cr and V ranged between 0.12–0.21 and 0.11–0.15 mg kg^{-1}, respectively. The concentration of Rb ranged from 0.05 to 0.45 mg kg^{-1}, being lower than those determined in different common bean varieties collected from Hungary (2.6–18.2 mg kg^{-1}) (Kato et al., 2015). With regard to Tl, similar concentration profiles were observed between samples from different varieties of cowpeas investigated, and their mean contents were lower than those found in vegetables (0.02–0.3 mg kg^{-1}) (Karbowska, 2016).

The trace elements at concentrations lower than 0.1 mg kg^{-1} can be arranged in the following order: Se > Sb > Co > Ag > As > Cd > Cs. The last two elements were found at ultra-trace levels varying between 5 and 25 μg kg^{-1}. Co concentrations (0.02–0.08 mg kg^{-1}) were below those found in cowpea beans (0.7–2.3 mg kg^{-1}) (Santos et al., 2013). Sb contents were in the range of 0.02–0.14 mg kg^{-1}, with an overall average of 80 μg kg^{-1}. However, this element is not commonly found in legumes. On the other hand, As and Cd elements exhibited concentration values of 30 and 17 μg kg^{-1}, respectively; these values were strictly in compliance with the regulatory limits of 100 μg kg^{-1} for each element in dried legumes imposed by MERCOSUR (MERCOSUR, 2012) and FAO/WHO (Codex Alimentarius, 2011).

3.3. Exploratory statistical analysis

Firstly, the data matrix that consisted of concentrations of 20 elements from 106 cowpea bean samples was autoscaled, and PCA was applied for exploratory analysis. This helped ensure an equal contribution of variables to the results. PCA reduces the number of variables used in data description, and is the most frequently applied method for computing components (linear latent variables) (Bro & Smilde, 2014; Varmuza & Filzmoser, 2009). The first three principal components (PCs) extracted according to Kaiser Criterion represented 16.2%, 11.1%, and 9.1%, respectively, of the variability of the system.

Fig. 1 shows the score-plot for PC2 vs. PC1, where one can observe a clear overlap among the scores corresponding to the different cowpea samples identified according to their botanical variety. Samples from the ALA variety showed negative scores in PC1, and were more easily differentiated from the SML and MRH varieties which exhibited positive scores in PC1. However, samples from the CBK and CES varieties presented a combination of positive and negative scores in PC2; the CBK group exhibited a yet distant sample, which tended to complicate the distinction between them. The orientation of the variables in the PC2–PC1 plane is observed from the loading plot shown in Fig. 2. PC1 was strongly influenced by the values of Sn, Mo, Ba, Ni and Zn with positive contributions and by those of V and Sr with negative contributions. The dominant variables in PC2 included Cr, Ni, Tl, Se, Rb, Be and As.

Insert Fig. 1

Insert Fig. 2

In short, the results obtained by PCA showed that the samples from the ALA variety can be distinguished from those of the SML and MRH varieties according to their multielemental fingerprinting. Nonetheless, the last two groups would be difficult to differentiate from one another due to the overlap of their scores. In the same way, although one could differentiate between the samples from the CBK and CES varieties, these varieties cannot be differentiated from the rest of the samples. Through the application of this unsupervised method of pattern recognition, one notices the natural grouping of the 106 cowpea bean samples in the original data matrix, indicating a slight tendency to group some samples. Remarkably, the systematic separation of samples is not clear. Hence, the application of supervised pattern recognition methods for the development of classification models is required to enable one to distinguish the cowpea bean samples from each other according to their botanical varieties.

3.4. Predictive modeling for sample classification

Upon the completion of the exploratory analysis, several machine learning techniques were applied to the data matrix aiming at the predictive modeling of the cowpea bean varieties. To construct the different classification models, the dataset was divided into two subsets randomly; these included a training set with known class memberships used to calculate the classifiers, and a test set containing samples that were not included in the training and which also had known class memberships that served to validate the models constructed. Considering the different cowpea varieties, a random sampling of the samples was performed in order to balance the group distributions within the splits (stratified sampling). The samples included in each set were randomly changed for each model that was replicated. The training set was formed by 70% of the total samples (n = 75), while the remaining 30% (n = 31) constituted the testing set. The parameters requiring optimization were calculated during the training stage via the cross-validation technique described above; the maximum accuracy was selected as the best criterion. The samples included in the test set were used to evaluate the performance of the methods developed here against an unknown set of samples.

Firstly, the classical LDA method was applied to the whole set of beans samples, where a centroid was created which represented the mean position of all points in all directions. The prediction of results was carried out by projecting the new samples in accordance with the minimal distance to the centroid of each class. Here, a success rate of 87% was recorded in the test set while better sensitivity was obtained for the ALA, MRH and SML varieties.

Fig. 3 shows the distribution patterns of cowpea bean samples according to their botanical variety in the plot defined by the first two canonical discriminant functions (DFs). As can be observed, this figure depicts a good discrimination between the four main groups formed by samples from ALA (negative scores on DF2), MRH (positive scores on DF2), SML (majority of positive scores with a negative score farthest on DF2), and CBK + CES (positive and negative scores on DF2). The latter group appears at the center of the graph, exhibiting some degree of difficulty to successfully distinguish the varieties by which it is constituted.

Insert Fig. 3

Secondly, the cowpea data matrix was analyzed using PLS-DA aiming at combining the properties of partial least squares regression with the discriminative ability of a classification technique for predictive modeling (Ballabio & Consonni, 2013). Accordingly, the number of significant components for the PLS regression was optimized (ncomp = 2). Through the combination of these chemometric tools, one is able to obtain additional information regarding the importance of the variables in the classification model, thus enabling the selection of variables and the reduction of noise. Fig. 4 shows the importance of variables for the PLS-DA model. From what can be noted, Ag and Cr were the variables that exerted the highest influence on the discrimination of the CES and CBK groups, while the ALA and MRH groups were mostly influenced by Sn and Mo. For the SML group, none of the variables were found to have stood out among the lots. Surprisingly though, a global accuracy of 77% was obtained despite the fact that the independent variables of the model did not appear to be highly collinear. In view of that, supervised non-linear methods were needed to solve the prediction of varieties in cowpea bean samples as a result of the difficulties encountered in discriminating some sample groups.

Insert Fig. 4

For the construction of the k-NN model, the optimal size of neighbor k was optimized, so that when k = 21 was applied, the highest average accuracy was obtained. The parameter k indicates the number of neighbors which are considered to predict the unknown sample classes by majority. This method stores all cases and classifies new samples by projection into the multivariate space and attributing these samples to the class of their closest neighbor in the training set. The balanced accuracies for all 106 samples were 63% for ALA samples, 50% for CBK samples, 70% for CES samples, 79% for MRH samples and 50% for SML samples; and the overall classification accuracy was 50%. The K-NN model failed to resolve the problem of sample classification; the linear methods studied presented better results.

In the RF model, 500 trees were constructed, and an mtry = 2 was obtained during the training stage. Each tree was grown using a bootstrapped sample from the original learning sample. However, at each node of the tree, a set of variables were randomly selected while a size was defined. This random selection of features at each node decreases the correlation between the trees in the forest, thereby resulting in a decline in the forest error rate (Archer & Kimes, 2008). Based on the RF classifier, the samples from ALA, CES and MRH groups were classified correctly according to their element profiles; here, the overall classification accuracy obtained was 86% as the samples belonging to the CBK and SML varieties were difficult to differentiate. It is worth pointing out that some works published in the literature have shown an improved accuracy of RF in comparison to other supervised learning methods (Canizo et al., 2017; Villafañe et al., 2017). Oddly enough, the classification performance obtained here was clearly not as expected.

By contrast, the SVM classifies the samples by constructing a separate hyperplane in n-dimensional space, maximizing the margin between the classes. This model uses an iterative algorithm, learning the distribution of samples at the boundaries of each of the classes considered (Bona et al., 2017).

From the classification results obtained by the LDA and PLS-DA models, a non-linear relation between classes and objects can be considered. In these cases, the use of Kernel trick is recommended as it enables the transformation of the input data in a linearly separable higher-dimensional feature space (Moncayo et al., 2015). In this work, several types of kernel functions were used to carry out the classification: linear kernel, polynomial kernel, gaussian kernel, radial basis function (RBF) and sigmoid kernel. The best results were obtained when the RBF kernel was used. In general, this function is run first due to its ability to handle multivariate data.

The complexity of the SVM algorithm is controlled by a penalty error function (C value), which is meant to improve the prediction result and prevent over-fitting (Lantz, 2015). Here, γ is the RBF kernel free parameter. Upon the completion of training, the C and γ were selected, and the model was verified with a set of tests leading to a prediction result. The hyperparameters, $C = 0.5$ and $\gamma = 0.096$, were the best fit for the classification model. Table 4 shows the SVM model achieve 100% of sensitivity for prediction of four varieties of cowpea beans (ALA, CBK, CES and MRH). The SML variety reached only 67% which is still considered a good result considering the great similarity between the different groups. Finally, the overall classification accuracy was 93%. Thus, this was the most suitable supervised learning method for predicting the variety of cowpea beans from their multielemental fingerprinting.

3.5. Comparing classification performance obtained by the chemometric methods under investigation

The classification performance of the supervised learning methods investigated in this work was compared in terms of sensitivity, accuracy and kappa metrics. Clearly, Kappa value becomes greatly important in cases involving classifiers with unbalanced datasets (Lantz, 2015), as is the case of this work. This statistic provides a more robust measure of agreement than accuracy, since it takes into account the expected agreement by random chance (Stephens & Diesing, 2014).

Table 4 summarizes the results obtained for the different algorithms. The order of successful predictions for the models was as follows: SVM > LDA > RF > PLS-DA > k-NN. In general, the samples from the ALA and MRH varieties can be classified by the different models with 100% sensitivity, with the exception of K-NN. LDA and RF models presented similar performance in terms of overall accuracy and sensitivity per class. A further observation that merits mentioning is that the ALA, SML and MRH groups were correctly classified by the LDA method as predicted in Fig. 3. The classification by PLS-DA exhibited an excellent performance for the ALA, CES and MRH samples, but it failed to separate the samples of CBK and SML. When the K-NN method was used, the problem related to the classification of the samples was not possible to be solved. This algorithm failed to differentiate all the cowpeas varieties investigated here, obtaining a very low overall accuracy. The SVM with RBF kernel function was the model that attained the best success rate in the test set, with 100% sensitivity in most of the groups considered.

Insert Table 4

On the other hand, according to Cohen (McHugh, 2012), the Kappa statistic can be more easily interpreted if we rely on Table 5. For instance, any kappa value below 0.60 indicates that the confidence intervals obtained for this statistical measure are wide enough; this implies that about half of the data may be incorrect. Such a situation can be noted in the case of the K-NN model, where only 10% (kappa = 0.32) of the data were analyzed correctly, which may have even included anything from good to poor agreement. Thus, the results obtained by this classifier are found to be unreliable. The PLS-DA model showed an agreement of 46% (kappa = 0.68), indicating a moderate level of agreement, since 54% of the data analyzed were erroneous. The RF (Kappa = 0.81) and LDA (Kappa = 0.82) models were within the data reliability range of 64–81%, which indicates a strong agreement of the classifiers. Finally, the SVM model showed a kappa value equal to 0.91, indicating an almost perfect agreement between the predicted model values and the actual values. Furthermore, this classifier obtained the highest overall accuracy among the lots. The model created by SVM was, thus, selected for predicting the botanical variety of cowpea bean samples according to their element profiles.

Insert Table 5

Based on the estimated metrics per class, good classification results were obtained by a nonlinear model probably due to the flexibility and ability of the SVM algorithm to create a generalized model (Gaiad et al., 2016). An SVM classifier generally finds a linear or non-linear decision surface in a feature space that separates the training classes with the largest distance between objects on the boundaries (Moncayo et al., 2015).

In general, multivariate analysis applied to the cowpea data matrix demonstrated that the concentration of traces elements investigated here varied among the varieties of the same bean species; thus enabling the distinction and authentication of bean samples of different varieties based on their multielemental fingerprinting.

4. Conclusions

The quantitative cowpea fingerprints consisting of 20 trace elements were obtained by ICP-MS. This technique has been proved to be a fast and reliable tool for generating highly chemical information-rich multielemental fingerprinting. The essential elements, including Co, Cr, Cu, Mo, V, and especially Zn, exhibited good nutritional contribution in the bean samples. For non-essential elements, the concentration levels of As, Cd and Pb showed no significant danger to human health. By projecting multielemental data on computing platform of five supervised learning techniques, the study demonstrated that there are significant differences among the concentrations of some elements across different cowpea varieties. The most relevant variations observed were associated with Ag, As, Ba, Cr, Mo, Ni, Rb, Sb, Sn and V, which enabled the successful classification of cowpea samples according to their botanical variety. Based on the models proposed using the LDA, PLS-DA, k-NN, SVM, and RF algorithms, the classification performance of SVM was found to be the best, as it attained a success rate of 93% during the test step, with 100% sensitivity for most of the predicted classes. In view of that, the SVM model was regarded the most robust classification algorithm for predicting genotype of bean samples from their element profiles. Multielemental determination combined with supervised pattern recognition techniques have proved to be an essentially promising tool for differentiating cowpea varieties. This work unfolds the great potential of multielemental fingerprinting when it comes to evaluating the authenticity of foods, such as cowpea beans.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgements

The authors would like to express their sincerest gratitude to the National University of the Northeast (SGCyT-UNNE) and the National Scientific and Technical Research Council (File LH: 172645 CONICET) for granting a postdoctoral scholarship to M. Pérez-Rodríguez.

References

Alayande, L. B., Mustapha, K. B., Dabak, J. D., & Ubom, G. A. (2012). Comparison of nutritional values of brown and white beans in Jos North Local Government markets. *African Journal of Biotechnology*, *11*, 10135–10140.

Ammann, A. A. (2007). Inductively coupled plasma mass spectrometry (ICP MS): a versatile tool. *Journal of Mass Spectrometry*, *42*, 419–427.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*, 2249–2260.

Asgar, M. A. (2013). Anti-diabetic potential of phenolic compounds: A review. *International Journal of Food Properties*, *16*, 91–103.

Avanza, M., Acevedo, B., Chaves, M., & Añón, M. (2013). Nutritional and anti-nutritional components of four cowpea varieties under thermal treatments: Principal component analysis. *LWT - Food Science and Technology*, *51*, 148–157.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*, 3790–3798.

Batista, B. L., Silva, L. R. S., Rocha, B. A., Rodrigues, J. L., Berretta-silva, A. A., Bonates, T. O., Gomes, V. S. D., Barbosa, R. M., & Barbosa, F. (2012). Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. *Food Research International*, *49*, 209–215.

Belane, A. K., & Dakora, F. D. (2011). Levels of nutritionally-important trace elements and macronutrients in edible leaves and grain of 27 nodulated cowpea (Vigna unguiculata L. Walp.) genotypes grown in the Upper West Region of Ghana. *Food Chemistry*, *125*, 99–105.

Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L., Nescatelli, R., & Marini, F. (2013). Data handling in science and technology, Chapter 5 - Classification and class-modelling, Vol. 28, pp. 171–233, Oxford: Elsevier.

Bona, E., Marquetti, I., Varaschim, J., Yasuo, G., Makimori, F., Arca, C., Guimar, L., Brígida, M., Valderrama, P., & Poppi, R. J. (2017). Support vector machines in tandem with infrared spectroscopy for geographical classi fi cation of green arabica coffee. *LWT - Food Science and Technology*, *76*, 330–336.

Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis : taking the magic away. *Journal of Chemometrics*, *28*, 213–225.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*, 2812–2831.

Canizo, B. V, Escudero, L. B., Pérez, M. B., Pellerano, R. G., & Wuilloud, R. G. (2018). Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi- elemental analysis and chemometrics tools. *Food Chemistry*, *242*, 272–278.

Codex Alimentarius. (2009). Guidelines for the design and implementation of national regulatory food safety assurance program associated with the use of veterinary drugs in food producing animals. CAC/GL–71, p. 22.

Codex Alimentarius. (2011). Joint FAO/WHO Food Standards Programme. Codex Committee on contaminants in foods. Fifth Session, The Hague, The Netherlands, 21-25/03/2011.

Código Alimentario Argentino. (2012). Reglamento Técnico MERCOSUR sobre "Límites máximos de contaminantes inorgánicos en alimentos". Resolución Conjunta 116/2012 y 356/2012 Modificación. Bs. As., 18/7/2012.

European Community. (2006). Commission Regulation No. 1881/2006. Setting maximum levels for certain contaminants in foodstuffs. DO L 364 20/12/2006, p. 5.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The elements of statistical learning: Data mining, inference, and prediction, Second Edition, Stanford University, Springer.

Hernández-Pereira, E. M., Álvarez-Estévez, D., & Moret-Bonillo, V. (2015). Automatic classification of respiratory patterns involving missing data imputation techniques. *Biosystems Engineering*, *138*, 65–76.

Inobeme, A., Nlemadim, A. B., Obigwa, P. A., Ikechukwu, G., & Ajai, A. I. (2014). Determination of proximate and mineral compositions of white cowpea beans (Vigna Unguiculata) collected from markets in Minna, Nigeria. *International Journal of Scientific & Engineering Research*, *5*, 502–504.

Karbowska, B. (2016). Presence of thallium in the environment: Sources of contaminations, distribution and monitoring methods. *Environmental Monitoring and Assessment*, *188*, 640–659.

Kato, L. S., Fernandes, E. A. D. N., Bacchi, M. A., Sarriés, G. A., & Reyes, A. E. L. (2015).

Elemental characterization of Brazilian beans using neutron activation analysis. *Journal of Radioanalytical and Nuclear Chemistry*, *306*, 701–706.

Lantz, B. (2015). Machine Learning with R. Second Edition, Packt Publishing Ltd., Birmingham.

Liu, X., Xue, C., Wang, Y., Li, Z., Xue, Y., & Xu, J. (2012). The classification of sea cucumber (Apostichopus japonicus) according to region of origin using multi-element analysis and pattern recognition techniques. *Food Control*, *23*, 522–527.

McHugh, M. L. (2012). Interrater reliability : the kappa statistic. *Biochemia Medica*, *22*, 276–282.

Moncayo, S., Manzoor, S., & Caceres, J. O. (2015). Chemometrics and Intelligent Laboratory Systems Evaluation of supervised chemometric methods for sample classi fi cation by Laser Induced Breakdown Spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *146*, 354–364.

Moreira-Araújo, R. S. R., Sampaio, G. R., Soares, R. A. M., Silva, C. P., Araújo, M. A. M., & Arêas, J. A. G. (2018). Identification and quantification of phenolic compounds and antioxidant activity in cowpeas of brs xiquexique cultivar. *Revista Caatinga*, *31*, 209–216.

Santos, W. P. C., Gramacho, D. R., & Teixeira, A. P. (2008). Use of Doehlert design for optimizing the digestion of beans for multi-element determination by inductively coupled plasma optical emission spectrometry. *Journal of the Brazilian Chemical Society*, *19*, 1–10.

Santos, W. P. C., Hatje, V., Lima, L. N., Trignano, S. V., Barros, F., Castro, J. T., & Korn, M. G. A. (2008). Evaluation of sample preparation (grinding and sieving) of bivalves, coffee and cowpea beans for multi-element analysis. *Microchemical Journal*, *89*, 123–130.

Santos, W. P. C., Santos, D. C. M. B., Fernandes, A. P., Castro, J. T., & Korn, M. G. A. (2013). Geographical characterization of beans based on trace elements after microwave-assisted digestion using diluted nitric acid. *Food Analytical Methods*, *6*, 1133–1143.

Santos, W. P. C., Teixeira, J., Almeida, M., Pires, A., Luis, S., Ferreira, C., Graças, M., Korn, A. (2009). Application of multivariate optimization in the development of an ultrasound-assisted extraction procedure for multielemental determination in bean seeds samples using ICP OES. *Microchemical Journal*, *91*, 153–158.

Scolaro, B., Kim, H. S. J., & Castro, I. A. (2018). Bioactive compounds as an alternative for drug co-therapy: Overcoming challenges in cardiovascular disease prevention. *Critical Reviews in Food Science and Nutrition*, *58*, 958–971.

Souza, P. A. L., Marcadenti, A., & Portal, V. L. (2017). Effects of olive oil phenolic compounds on inflammation in the prevention and treatment of coronary artery disease. *Nutrients*, *9*, 1087–1109.

Stephens, D., & Diesing, M. (2014). A comparison of supervised classification methods for the prediction of substrate type using multibeam acoustic and legacy grain-size data. *PLoS One*, *9*, e93950.

Tang, W., Hu, J., Zhang, H., Wu, P., & He, H. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai Archives of Psychiatry*, *27*, 62–67.

Thompson, M., Ellison, S. L. R., & Wood, R. (2002). Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC Technical Report). *Pure and Applied Chemistry*, *74*, 835–855.

Varmuza, Kurt & Filzmoser, P., (2009). Introduction to multivariate statistical analysis in chemometrics, CRC Press, Taylor & Francis Group.

Zhou, H., Yang, W. T., Zhou, X., Liu, L., Gu, J.F., Wang, W. L., Zou, J. L., Tian, T., Peng, P.

Q., & Liao, B. H. (2016). Accumulation of heavy metals in vegetable species planted in contaminated soils and the health risk assessment. *International Journal of Environmental Research and Public Health*, *13*, 289–301.

Figure captions:

Fig. 1. Score plot of the first principal component (PC1) versus the second principal component (PC2).

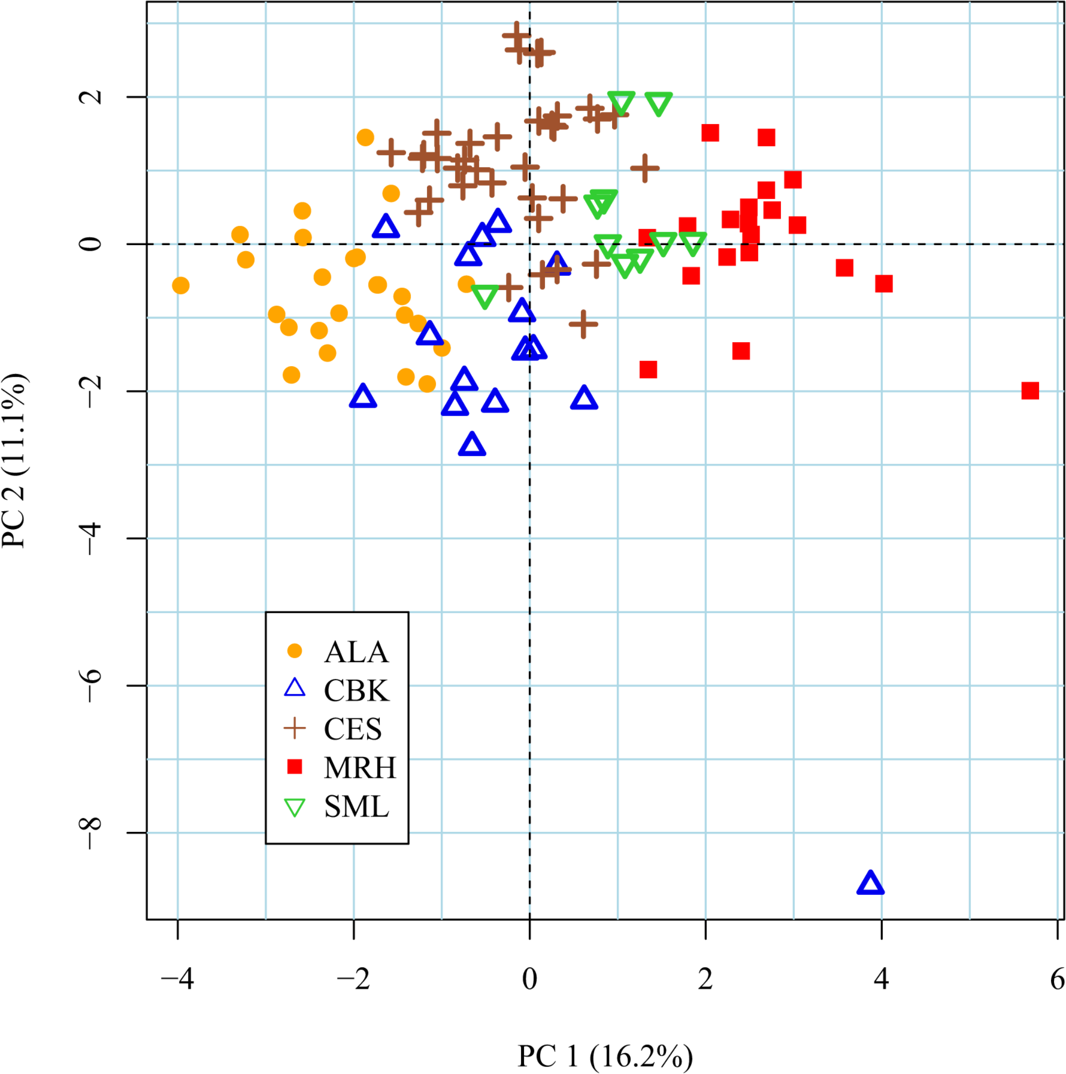
Fig. 2. Loading plot for the original variables in the first two principal components (PCs).

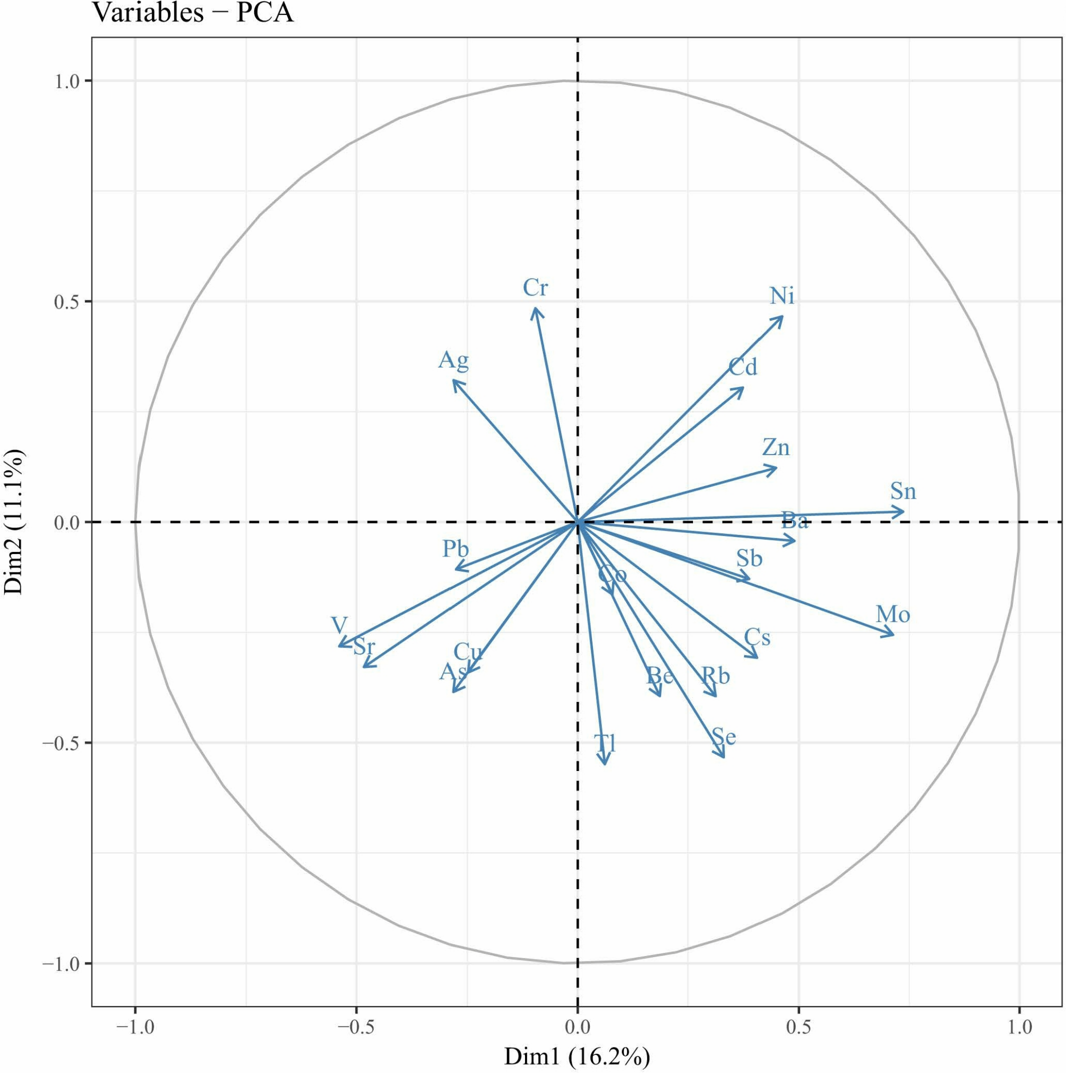
Fig. 3. Scatter plot of the first two discriminant functions obtained from linear discriminant analysis of cowpea beans according to their botanical variety.

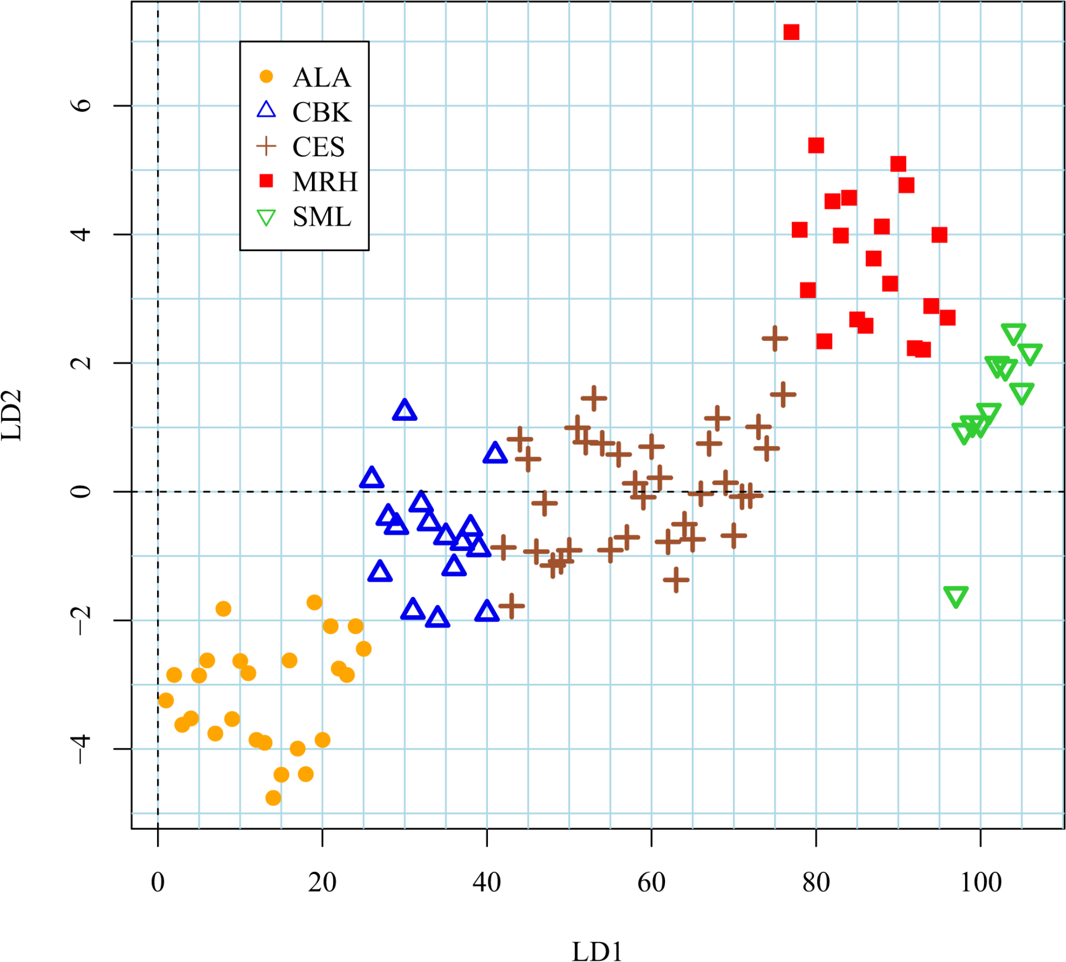
Fig. 4. Importance of variables for predicting cowpea genotype according to the PLS-DA model.

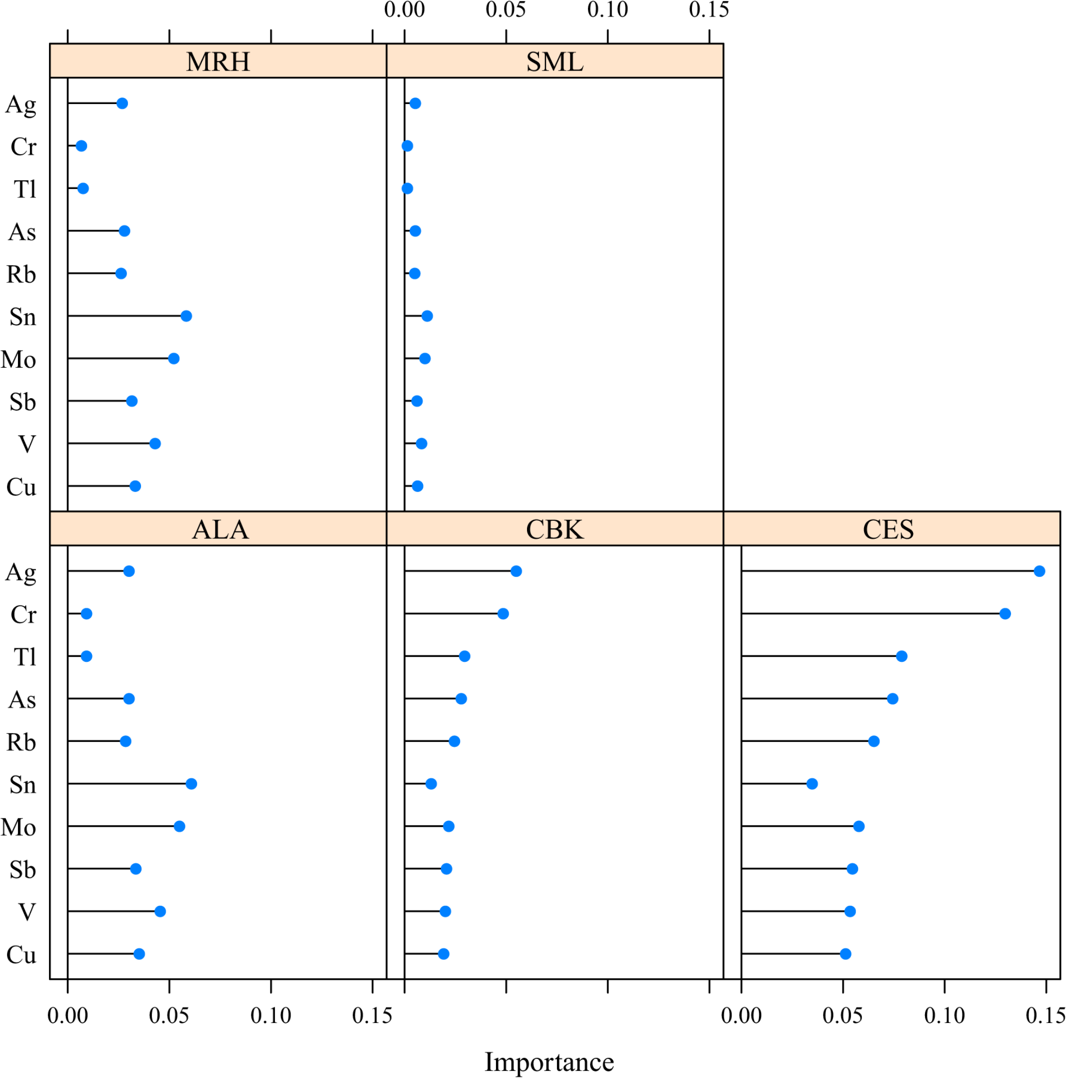
HIGHLIGHTS

- Multielemental fingerprinting was used to verify the authenticity of cowpea beans.
- LDA, PLS-DA, k-NN, RF and SVM were applied to differentiate cowpea genotypes.
- The best performance for predicting cowpea varieties was achieved by SVM classifier.


ALA
CBK
CES
MRH
SML
PC 1 (16.2%)
PC 2 (11.1%)
−4
−2
0
2
4
6
−8
−6
−4
−2
0
2


Variables − PCA
Cr
Ni
Ag
Cd
Zn
Sn
Ba
Pb
Sb
Co
Mo
V
Sr
Cs
Cu
As
Be
Rb
Se
Tl
1.0
0.5
0.0
−0.5
−1.0
Dim2 (11.1%)
Dim1 (16.2%)


ALA
CBK
CES
MRH
SML
LD2
LD1
6
4
2
0
−2
−4
0
20
40
60
80
100



MRH
SML
ALA
CBK
CES
Ag
Cr
Tl
As
Rb
Sn
Mo
Sb
V
Cu
0.00
0.05
0.10
0.15
Importance

Table 1
Instrumental parameters for ICP-MS determinations.

Radiofrequency power (W)	1321
Sampling depth (mm)	10.0
Gas flow rate (L min^{-1})	
Plasma gas (Ar)	13.88
Nebulizer gas (Ar)	0.95
Carrier gas (Ar)	0.85
Auxiliary gas (Ar)	0.80
Sampling cone (Ni, mm)	1.0
Skimmer cone (Ni, mm)	0.4
Collection points/unit mass	3
Internal standard	^{72}Ge, ^{115}In

Table 2
Coefficients of determination (r^2), limits of quantification (LOQ), intra-day and inter-day precisions and overall recoveries obtained for the determination of multi-elements in bean samples by ICP-MS.

Element	r^2	LOQ (mg kg^{-1})	Intra-day (RSD%)	Inter-day (RSD%)	Certified values (mg kg^{-1})	Recovery (%) (n = 5)	
						CRM	Fortified samples
Ag	0.9989	0.002	4.1	5.5			101.5
As	0.9998	0.010	0.2	2.7	0.112 ± 0.004	97.1	109.2
Ba	0.9993	0.020	0.7	1.8			95.8
Be	0.9996	0.092	1.3	4.8			98.7
Cd	0.9987	0.011	0.5	3.9	1.52 ± 0.04	94.5	98.1
Cs	0.9986	0.003	0.6	5.7			95.3
Co	0.9995	0.011	1.7	1.9	0.57 ± 0.02	93.0	97.3
Cr	0.9997	0.079	1.4	4.9	1.99 ± 0.06	104.5	97.5
Cu	0.9992	0.545	2.6	6.5	4.70 ± 0.14	102.7	100.5
Mo	0.9988	0.205	4.8	6.9			96.0
Ni	0.9994	0.050	3.2	2.7	1.59 ± 0.07	94.2	89.8
Pb	0.9990	0.125	0.5	3.3			86.2
Rb	0.9988	0.015	2.0	3.8	14.89 ± 0.27	95.6	101.8
Sb	0.9998	0.018	0.7	2.8	0.063 ± 0.006	107.0	101.0
Se	0.9993	0.031	4.2	5.9	0.054 ± 0.003	98.4	103.7
Sn	0.9991	0.078	2.2	2.3			108.7
Sr	0.9989	0.007	1.2	7.2			92.7
Tl	0.9995	0.020	0.7	2.4			90.5
V	0.9997	0.010	2.8	1.5	0.835 ± 0.010	97.3	107.5
Zn	0.9985	0.063	1.0	3.7	30.9 ± 0.7	101.5	98.1

Table 3
Elemental composition of cowpea bean samples of different varieties analyzed by ICP-MS.

Element (mg kg^{-1})	Sampling varieties (average ± SD)					
	ALA (n = 25)	CBK (n = 16)	CES (n = 35)	MRH (n = 20)	SML (n = 10)	Overall range
Ag	0.04 ± 0.04	0.02 ± 0.03	0.07 ± 0.01	0.02 ± 0.01	0.03 ± 0.02	0.005 – 0.098
As	0.05 ± 0.02	0.03 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.04 ± 0.01	0.008 – 0.082
Ba	40.74 ± 16.91	58.76 ± 15.92	53.04 ± 14.38	66.57 ± 16.98	52.26 ± 15.66	19.8 – 96.6
Be	0.22 ± 0.02	0.22 ± 0.03	0.22 ± 0.02	0.23 ± 0.02	0.23 ± 0.02	0.17 – 0.30
Cd	0.015 ± 0.002	0.013 ± 0.005	0.018 ± 0.004	0.020 ± 0.002	0.020 ± 0.003	0.010 – 0.025
Cs	0.010 ± 0.002	0.010 ± 0.003	0.010 ± 0.001	0.011 ± 0.004	0.011 ± 0.003	0.005 – 0.025
Co	0.05 ± 0.01	0.05 ± 0.02	0.05 ± 0.01	0.05 ± 0.01	0.05 ± 0.02	0.02 – 0.08
Cr	0.17 ± 0.04	0.12 ± 0.04	0.21 ± 0.05	0.16 ± 0.04	0.14 ± 0.04	0.07 – 0.36
Cu	5.98 ± 1.05	4.74 ± 1.29	4.77 ± 1.11	4.95 ± 0.99	4.69 ± 1.44	2.60 – 7.90
Mo	0.69 ± 0.49	1.58 ± 0.71	1.02 ± 0.03	1.85 ± 0.45	1.36 ± 0.37	0.26 – 3.05
Ni	0.28 ± 0.07	0.32 ± 0.13	0.42 ± 0.11	0.46 ± 0.10	0.67 ± 0.17	0.15 – 0.96
Pb	0.28 ± 0.07	0.35 ± 0.11	0.28 ± 0.06	0.24 ± 0.09	0.25 ± 0.04	0.14 – 0.52
Sb	0.06 ± 0.02	0.09 ± 0.03	0.09 ± 0.01	0.08 ± 0.01	0.07 ± 0.01	0.02 – 0.14
Se	0.08 ± 0.01	0.09 ± 0.04	0.08 ± 0.01	0.08 ± 0.01	0.08 ± 0.01	0.06 – 0.24
Sn	0.16 ± 0.04	0.20 ± 0.04	0.23 ± 0.04	0.48 ± 0.16	0.23 ± 0.11	0.10 – 0.99
Sr	4.35 ± 1.45	5.94 ± 1.26	3.91 ± 0.99	2.55 ± 0.88	3.32 ± 1.43	1.50 – 8.12
Tl	0.11 ± 0.02	0.10 ± 0.02	0.09 ± 0.03	0.10 ± 0.02	0.11 ± 0.02	0.05 – 0.17
Rb	0.17 ± 0.04	0.26 ± 0.09	0.17 ± 0.05	0.23 ± 0.04	0.19 ± 0.09	0.05 – 0.45
V	0.15 ± 0.03	0.14 ± 0.02	0.13 ± 0.02	0.11 ± 0.02	0.11 ± 0.02	0.08 – 0.20
Zn	32.56 ± 8.46	32.45 ± 4.74	35.82 ± 8.13	42.12 ± 8.40	33.19 ± 5.39	17.34 – 60.27

Note: n indicates the number of samples per bean variety.

Table 4
Performance evaluation measurements for different classification models.

Varieties	LDA		PLS-DA (*ncomp* = 2)[a]		K-NN (*k* = 21)[b]		FR (*nt* = 500; *mtry* = 2)[c]		SVM (*C* = 0.5; γ = 0.096)[d]	
	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)
ALA	100	91	100	83	57	70	100	94	100	96
CBK	75	100	–	100	–	100	67	100	100	100
CES	67	100	100	100	70	70	100	93	100	100
MRH	100	96	100	88	67	92	100	94	100	96
SML	100	96	–	100	–	100	–	100	67	100
Accuracy	87		77		50		86		**93**	
Kappa	0.82		0.68		0.32		0.81		**0.91**	

[a] *ncomp*: number of significant components.
[b] *k*: number of k neighbors.
[c] *nt*: number of trees; *mtry*: number of variables tried at each split.
[d] *C*: penalty factor; γ: γ-insensitive loss function.

Table 5
Interpretation of Kappa statistic.

Kappa value	Agreement level	% of data reliability[a]
0–0.20	None	0–4%
0.21–0.39	Minimal	4–15%
0.40–0.59	Weak	16–35%
0.60–0.79	Moderate	36–62%
0.80–0.90	Strong	64–81%
0.91–1.00	Almost perfect to perfect	82–100%

[a] % agreement or squared kappa value