TRACEABILITY OF SOYBEANS PRODUCED IN ARGENTINA BASED ON THEIR TRACE ELEMENT PROFILES

Short Title: Chemometric Tools for Classification of Soybean Grains.

Melisa J. Hidalgo[a*], Diana C. Fechner[a], Davide Ballabio[b], Eduardo J. Marchevsky[c], and Roberto G. Pellerano[a].

[a]Instituto de Química Básica y Aplicada del Nordeste Argentino (IQUIBA-NEA), UNNE-CONICET, Facultad de Ciencias Exactas y Naturales y Agrimensura, Av. Libertad 5400 (3400) Corrientes, Argentina.

[b] Milano Chemometrics y QSAR Research Group, Departamento de Ciencias Ambientales, Universidad de Milano-Bicocca, Piazza della Scienza 1 (20126) Milano, Italia.

[c]Instituto de Química San Luis (INQUISAL), UNSL-CONICET, Facultad de Química Bioquímica y Farmacia, Av. Ejército de los Andes 950 (5700) San Luis, Argentina.

* Corresponding author. Tel: +54-379-4457996; Fax: +54-379-4464793.

E-mail address: hidalgo.melisa@conicet.gov.ar (M.J. Hidalgo)

Summary

Soybean (Glycine max (L.) Merril) is a popular foodstuff and crop plant, used in human and animal food. In this work, multielement analysis of soybean grains samples in combination with chemometric tools was used to classify the geographical origins. For this purpose, 120 samples from three provinces of Argentina were analyzed for a panel of 20 trace elements by inductively coupled plasma mass spectrometry (ICP-MS). First, we used principal component analysis (PCA) for exploratory analysis. Then, supervised classification techniques such as support vector machine discriminant analysis (SVM-DA), random forest (RF), k- nearest neighbors (k-NN) and class-modeling techniques such as soft independent modeling of class analogy (SIMCA), potential functions (PF), and one class support vector machine (OC-SVM) were applied as tools to establish a model of origin of samples. The performance of the techniques was compare using global indexes. Among all the models tested, SVM and SIMCA showed the highest percentages in terms of prediction ability in cross-validation with average values of 99.3% for SVM-DA and a median value of balanced accuracy of 96.0%, 91.7%, 88.3% for the three origins using SIMCA. Results suggested that the developed methodology by chemometric techniques is robust and reliable for the geographical classification of soybean samples from Argentina.

Keywords: Soybean grains; Geographical origin; Class-modeling techniques.

1. Introduction

In recent years, the traceability of food products has become increasingly relevant for citizens of many countries mainly interested in food safety and quality. The traceability of a food means that movements can be traced back one step and one step forward anywhere in the supply chain. Most traceability systems are registration systems that document the path of a product from suppliers through intermediate steps to consumers. However, traceability systems mainly depend on the quality of the records and controls that are usually carried out by local food safety authorities[1]. Having a traceability system based on the chemical composition of food is a vital tool to guarantee the origin of food, especially for producing countries, such as Argentina.

Currently, Argentina is one of the main soybeans (*Glycine max* (L.) Merrit) exporters worldwide. However, this country does not currently offer chemical traceability systems of origin for the soybean produced. This fact acquires great relevance today, considering the soybean production often implies the indiscriminate deforestation of native forests that still serve as a home for native communities, among other environmental and health problems. Having an origin system of chemical traceability of this food would allow its consumers to guarantee that this product was produced respecting environmental and human issues. Today, major food producing companies and even the European Union itself are committed to eliminating deforestation from their global supply chains.

In this work, three important Argentine soybean producing areas were considered. The province of Córdoba, which stands out as the main producing region in the heart of Humid Pampas. The province of San Luis, in a border region of Humid Pampas with a lower production than Córdoba but of great importance at the national level. Finally, the

province of Chaco, questioned in the last years since it produces soy on the agricultural frontier occupying native forest regions.

The authenticity and traceability of the food product can be demonstrated by modern analytical techniques [2]. Reviews have been published about chemometric techniques applied on the geographical origin of foods [3–5]. For instance, the geographical origin of soybean seeds was characterized using X-ray fluorescence, showing differences between the trace element contents in soybean from different geographical regions [6]. These results showed that Mg, P, Cl, K, Mn, Cu, Br, and Ba were good parameters for constructing a discriminant model for geographical origin characterization between Japanese and imported samples (from Canada, China, and America). Beside, transgenic and non-transgenic soybean seeds was differentiated according to mineral content analyzed by ICP-MS [7]. The results show that transgenic and non-transgenic soybean seeds show differences in concentrations of Cu, Fe, and Sr.

Supervised pattern recognition techniques goal is to establish a classification model based on experimental data to assign unknown samples to a previously defined sample class based on its pattern of measured features[8]. For instance, the discriminant classification methods are appropriate when at least two classes are defined in the problem under study and allow to properly address only multi-class situations. The discriminant classification methods separate the hyperspace in as many regions as the number of sample groups. So, if a sample is matched in the region of space corresponding to a category, it is classified as that category. In reality, each sample is always assigned to one group, even if this sample is not from the studied categories. The most used discriminant classification methods in food chemistry include Linear Discriminant Analysis (LDA), k- Nearest Neighbors (k-NN), Support Vector Machine

discriminant analysis (SVM-DA), Random Forest (RF), Partial Least-Squares Discriminant Analysis (PLS-DA), among others[9].

A different approach to supervised pattern recognition are the class-modeling methods that are useful when the focus is on a single class [10]. The analogies among the elements of a class in each category are modeled separately. For this reason, the class-modeling methods can be used to study both one-class and multi-class problems. The samples in agreement with the model are assumed as a member of the class, while objects not in agreement are assumed as non-members.

When more than one class is modeled, three different situations are possible: each sample can be assigned to a single category or assigned to several categories or not assigned in any category. One of the main advantages offered by class modeling methods is the possibility of recovering samples that are not represented in any of the studied categories. As a consequence, these methods will be able to identify as "foreign" samples those problem samples that correspond to external observations or members of a new class not considered in the training stage. Another advantage of the method is that any additional class can be added without recalculating the already existing class models, as each category is modeled separately. In addition, all class-modeling methods can be used as discriminant tools, while the reverse is not always the case. The most used chemometric class-modeling techniques are Soft Independent Modeling of Class Analogy (SIMCA), Potential Functions (PF), One-Class Support Vector Machine (OC-SVM)[11].

Many studies in the literature compare the performance of different pattern recognition techniques in food. Hence, the aims of this work were: (1) to characterize the geographical origin of soybeans produced in three regions of Argentina using

chemometrics tools applied to trace element compositions, and (2) to compare the classification performance of three discriminant classification methods (k-NN, SVM-DA, RF) and three class-modeling methods (SIMCA, OC-SVM, PF).

2. Experimental

2.1. Reagents

Mono and multi-element standard solutions of trace analysis grade were purchased from Sigma-Aldrich and Agilent. Ultrapure grade 65% (m/m) HNO_3 and 30% (m/m) H_2O_2 was acquired from Sigma (St. Louis, MO, USA). Nitric acid was further purified by sub-boiling distillation. Water with a resistivity of 18.1 MΩ cm^{-1} was obtained from a Milli-Q Pluswater purification system Millipore (Molsheim, France). Indium solution 100 µg/L obtained from Agilent (Santa Clara, CA) was used. All the chemicals used were of the highest purity available and all the glass materials used were soaked in 10% (v/v) nitric acid and washed with deionized water.

2.2. Samples

A total of 120 samples were collected from six test fields located in the north-central region of Argentina, corresponding to the main soy producing region of this country. The test fields were located at: Almirante Brown (Chaco province, 26°40' S 60°54' W), San Justo (Córdoba province, 31°26’ S 62°04’ W) and General Pedernera (San Luis province, 33º37’ S, 65º19’W). Representative samples of soybeans were collected during the 2018-2019 campaign at different times to create composite samples labeled according to the sampling region. All the samples studied correspond to the botanical species *G. max* and were grown by direct sowing.

In the laboratory, soybean seeds were Soybeans were manually separated from pods. Then, allseeds were washed with tap water and rinsed with deionized water. After that, seeds were homogenized with a domestic mixer and stored at −20 °C in a freezer until analysis.

2.3. Sample preparation

Previous to multielemental determination in botanical samples organic matter should be eliminated (digested). For the digestion of the soybean samples a microwave digestion oven, Milestone® (Chicago, USA) model Ethos One was used. The microwave digestion program was: (1) 25-200 ºC for 15 min, (2) 200 ºC for 15 min and (3) 200-110 ºC for 15 min, followed by ventilation at room temperature for 20 min. After cooling to room temperature, the volume was made up to 25 mL with deionized water. Blank solutions and validation spiked samples were prepared in the same way. Prior to use, all plastic containers were soaked in 10% v/v sub-boiling HNO_3 for at least 24 h and then rinsed extensively with deionized water. The samples were measured in triplicate.

2.4. ICP-MS analysis

The measurements of trace elements concentrations have been carried out by using an Agilent 7700 x ICP-MS spectrometer (Agilent Technologies, Santa Clara, CA). The instrument is equipped with off-axis ion lens, a quadrupole mass analyzer and an electron multiplier detector. MicroMist glass concentric nebulizer combined with a cooled double-pass spray chamber made of quartz, an octopole collision/reaction system (ORS). The operating parameters for the instrument were described as follow: RF power (1350 W), plasma gas (14.0 L/min), the flow rate of auxiliary gas (0.9 L/min), carrier gas (1.0 L/ min). Indium was used as internal standard. The selected isotopes for

164 measurement were ^{107}Ag, ^{11}B, ^{137}Ba, ^{59}Co, ^{53}Cr, ^{63}Cu, ^{56}Fe, ^{7}Li, ^{55}Mn, ^{95}Mo, ^{60}Ni, ^{208}Pb,
165 ^{85}Rb, ^{121}Sb, ^{78}Se, ^{118}Sn, ^{88}Sr, ^{205}Tl, ^{51}V, ^{66}Zn.

166 The accuracy and precision of the ICP-MS method were verified with one Standard
167 Reference Materials from the National Institute of Standards and Technology (NIST),
168 namely SRM 1573a tomato leaves. The precision of the proposed procedure was
169 evaluated by measuring the repeatability and reproducibility. In the repeatability test
170 (within-day precision), the SRM was analyzed three times within one day; and in the
171 reproducibility test (day-to-day precision), sample digestion and ICP MS analysis were
172 studied by triplicate analyses of three aliquots of the SRM on three days for a period of
173 three weeks. All concentration values were found to be in good agreement with the
174 reference values (Table 1). Limit of detection (LOD) and limit of quantification (LOQ)
175 were obtained according to IUPAC guidelines. LOD was calculate as $LOD = \bar{x}_b +$
176 $k.S_b$, where $\bar{x}_b$ is the mean of the blank measures, S_b is the standard deviation of the
177 blank measures, and k is a numerical factor chosen according to a confidence level. The
178 LOQ was defined as 3.3-fold the LOD.

179 TABLE 1

180 **2.5. Chemometric models**

181 **2.5.1. Exploratory data-analysis**

182 The results obtained were organized in a matrix with dimension 120 rows (soybeans
183 samples) and 20 columns (element concentrations). Prior to the exploratory analysis of
184 the data matrix, the concentration values of each element corresponding to each sample
185 were autoscaled (each value is subtracted by the mean and divided by the standard
186 deviation). This pretreatment method allows to avoid dimensionality problems between
187 the levels of the different trace elements in the samples.

188 A basic exploratory analysis was performed using principal component analysis
189 (PCA). PCA is a strategic technique that allows knowing relationships between
190 variables, between samples as well as between variables and samples. Characterized by
191 orthogonal linear combinations called principal components. The first components
192 retain the highest percentage of variability present in the initial data set[12].

193 **2.5.2. Supervised classification models**

194 As the objective of this work is to provide a classification model capable to predict
195 the geographical origin of soybean seeds from three principal production regions of
196 Argentina, we firstly perform a comparative study on the performance of three learning
197 classification algorithms. In addition, in a subsequent stage we compare the perform of
198 three class-modeling methods, in order to exploit their comparative advantage in terms
199 of class prediction of future unknown samples.

200 Three supervised classification algorithms were used to classify provenance of
201 soybean seeds. The supervised model uses pre-defined classes to learn through a
202 training phase how data is organized, making possible to predict unlabeled samples
203 based on the classification model. k nearest neighbor (k-NN), support vector machines
204 (SVM-DA) and random forest (RF) are three techniques which have yielded good
205 results in small rectangular data arrays in the literature[9].

206 k-NN is a distance based non-parametric procedure. The basic idea on which this
207 paradigm is based is that a new sample is going to be classified in the most frequent
208 class to which its k nearest neighbors belong. The value assumed by k is implicitly
209 related to the shape of the decision boundaries that separate the classes. In practice, the
210 optimal value of k is found by some validation procedure[13].

SVM is a supervised technique that produces linear boundaries among the objects of the groups in a transformed space. Three parameters affect the performance of this technique: penalty factor (C value), regularization parameter (ε) and the type of kernel function used. Radial basis function (RBF) kernel was used in this study. In the optimization of the parameters (C and ε) a grid-search and cross-validation were used to best fit the model and improve the accuracy results[14].

RF algorithm is an ensemble learning method. The idea of ensemble learning is to build and combine base learners to obtain a better classification capability. In this technique, multiple trees are generated. Each tree gives a classification (vote for a class). The result is the class with the highest number of votes in the whole forest. As the base learner, random forest uses the CART (classification and regression tree)[9].

Then, in order to propose robust predictive models, three class modeling methods were performed to classify soybean samples. Soft independent modeling by class analogy (SIMCA), one class SVM (OC-SVM) and potential functions (PF).

SIMCA was one of the first class-modeling technique introduced in the literature. A principal component analysis is generated for each of the classes present in the data set. The number of principal components that are retained by each class is generally obtained by cross-validation and the number of principal components retained may be different for each class[15].

OC-SVM consists in estimating the function that encloses training samples in a hypersphere with a reduced volume. This technique allows to classify only the objects of a class and distinguish them from other objects. RBF was the selected kernel function. This function allows to determine the radius of the hypersphere considering the parameter γ[16].

PF try to estimate the shape of the probability density distribution of the class as a sum of individual contributions of the samples of the class in the training phase. To define the contributions of the samples, different functions can be used[17].

2.5.3. Selection of a test set for external validation of models

In the classification and class modeling phase, the data matrix was random split in training (n = 84) and test (n = 36) sets. The random sampling occurred within each class and preserved the overall class distribution of the data. For discrimination models, the training set was used to tuning the parameters of k-nearest neighbors (k-NN), support vector machine (SVM-DA) and random forest (RF). Optimization of parameters was made using k-fold cross validation (k = 10). Testing set was used to compare the performance of each optimized method. To ensuring that the same resamples are used, internal parameters in R software was used. Finally, to compare the performance of optimized methods we resampled 50 iterations to avoid bias.

For class-modeling, internal cross validation, venetian blinds with 5 cross validation groups has been used with training samples to select model parameters such as number of PCs for Soft Independent Modeling of Class Analogy (SIMCA), kernel for Potential Functions (PF) and One-Class Support Vector Machine (OC-SVM). The results were achieved on the 100 iterations for each class with each classifier.

2.5.4. Software

Exploratory and supervised classification analysis were performed using R software[18], version 3.5. For class-modeling methods the Classification toolbox[19] for MATLAB® was used.

3. Results and discussion

3.1. Trace elements in soybean samples

Table 1 shows the median, minimum and maximum concentration of the elements detected in soybean samples from the three geographical origins. Fe and Mn were the most abundant elements in all samples, followed by Zn, Cu, Rb, V, Li, Ti, Sr, Ba, B, Mo and Cr at levels above 1 μg kg^{-1}, in decrescent order. The concentrations obtained for the 120 samples are provided in the Supplementary Information (SI-1).

The non-parametric Kruskal-Wallis test was applied to evaluate differences between the means population of three origins. Cr, Li, Mn and Zn concentrations were significantly different among three pairs of origins ($p < 0.01$) (Table 1). Thirteen out of 20 elements exhibited significant differences between the mean ranks of at least one pair of origins ($p < 0.05$), demonstrating that soybean samples from different regions have a characteristic elemental profile. There was no evidence of variation in the concentrations of Ag, Pb, or Se ($p > 0.05$) between any pair of origins.

3.2. Exploratory data-analysis by PCA

PCA was performed based on the concentration of 20 elements determined by ICP-MS in samples of soybean (*G. max)* grains from three geographical origins of Argentina. The first two principal components (PCs) accounted for 51.6% of the total variance. The PC1 summarized 33.4% and the second 18.2% of the variance present in the multielemental results of analyzed samples. As can be seen in the loading-plot obtained from PCA (Figure 1a), PC1 presented a strong positive correlation with the contents of Co, Rb and Sb, and in the direction of the negative values on the x-axis with the contents of Fe and Sr. The representation of mathematical space defined by the first

two PCs is completed with the PC2, which shows a strong negative correlation with the concentrations of B, Ni and Cu mainly.

FIGURE 1

Fig. 1b shows the distribution of samples in the space of the two first computed PCs (score plot). No clear separation is achieved by the samples from Córdoba and San Luis, although some trends can be observed. On the other hand, the Chaco samples appear in negative values of PC1 clearly differentiated from the previous groups, indicating that there are particular characteristics in the multielemental compositions of samples of this group. These differentiation trends are in accordance with the results of the previous Kruskal–Wallis multiple comparison test. As can be seen, the five variables showing statistical differences, such as Fe, Sr, Co, Rb and Sb, also appear as most contributing to the PC1, which is able to group samples in two principal groups.

These results of the exploratory analysis by PCA indicate that the contents of trace elements in the samples studied would be useful for the proposal of supervised classification models of geographical origin of soybeans produced in Argentina.

3.3. Supervised classification methods

In a first stage, we begin comparing the yield to correctly classify soybean samples according to their geographical region of origin, applying supervised classification methods. The algorithms selected were k-NN, SVM and RF. These methods were selected due to their great ability to achieve high correct classification rates, especially when only a small number of samples are available.

The dataset (120 × 20) was splitting up into training and testing sets, in a ratio 70/30. The partition of data matrix was carried out in a stratified form by random

sampling to create balanced splits of the data. The random sampling was performed within each class to preserve the overall class distribution of the data. Training set was used to tuning the hyperparameters of the algorithms k-NN, RF and SVM-DA. Optimization of parameters was made using k-fold CV (k = 10). The split of data was repeated 50 times. Finally, testing set was used to compare the performance of each optimized method. Table 2 shows the results of the classification metrics achieved by each technique in the data matrix.

TABLE 2

As can be seen, Table 2 shows the results achieved by each optimized algorithm in terms of average accuracy, sensitivity and specificity. Sensitivity (also called the true positive rate) describes the positive test samples of each group correctly classified as such, and specificity (also called the true negative rate) measures the ratio of negative test samples belonging to a different group which have been correctly predicted as such. The results of optimized SVM-DA model were the highest, followed by RF and k-NN in this order. This result is consistent with the findings from samples from Brazil[20], since it is expected that non-linear techniques (SVM-DA) have greater flexibility in solving non-linear systems. In addition, other commonly used supervised techniques such as LDA or PLS-DA (results not shown) were simultaneously tested with worse results.

To refine the classificatory results obtained by SVM-DA, we further optimize the hyperparameters of the algorithm performing 10-fold-cross validation (repeated 50 times). Radial basis function (RBF) kernel SVM was selected because of its speed and great capacity to obtain good results in complex systems. The hyperparameters C = 8 and $\gamma = 0.039$ were the best to obtain the minimal performance error in training setting.

Using these optimized hyperparameters, the SVM-DA algorithm achieved a classification rate with a range of 97.3% - 100% of global accuracy. A perfect classification rate was obtained for samples from the provinces of Chaco and San Luis, while only one sample from Cordoba could not be classified correctly. These results indicate that the SVM-DA method is suitable for the geographical classification of soybean samples, being able to differentiate even the samples coming from neighboring provinces (SLS and COR), which as could be observed in the exploratory analysis showed a high degree of similarity.

3.4. Class-modeling methods

Three different class-modeling algorithms (one class classifiers) were used to model trace element compositions of soybean produced in Argentina: SIMCA, PF (Gaussian Kernels) and OC-SVM. Being class modeling methods, they model one class at a time, we have considered the three classes (CHC, COR, SLS) separately.

For each class and for each type of classifiers, we used the following validation protocol. We made a double cross-validation with the following protocol, which have been repeated 100 times (iterations): I) Random split samples in training (70%) and test (30%) sets; II) Use the training samples to calibrate the model; internal cross validation has been used with training samples to select model parameters (such as number of PCs for SIMCA, kernel for Potential Functions and one-class SVM); III) Predict the test samples and calculate sensitivity, specificity and their average (balanced accuracy). Thus, test samples do not participate in the model optimization along each iteration.

Figure 2 shows the distribution of balanced accuracy obtained for each modelled class obtained on the 100 validation iterations summarized in box plots. Boxplots are a standardized way of displaying the distribution of data based on a five parameters

summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

FIGURE 2

As is shown in Fig. 2, CHC was best modelled class with a median of 96.0% for the three methods studied. However, the OC-SVM method shows a higher density of results around 100% success, indicating a better performance for modelling the samples of this group. It is also observed that class modelling methods have greater difficulties in differentiating samples from COR and SLS provinces, being these classes associated to lower average balanced accuracies (91.7% and 88.3% respectively) and greater dispersion of results over the validation iterations with respect to CHC. The best method for the COR group was SIMCA with a higher average accuracy and a lower dispersion of results. The samples of the SLS group presented greater difficulties to be modelled correctly, showing similar performances between OC-SVM and PF methods. These results, while compatible with the performances achieved by the supervised classification methods, indicate that the recommendation of class modelling techniques for CHC samples is appropriate, with their respective advantages. As an example, Fig. 3 shows the Hotelling T2 versus residual Q plot based on 1 PC SIMCA model for CHC class, where the separation of samples from this group is clearly observed.

FIGURE 3

4. Conclusion

In this study, ICP-MS in combination with chemometric tools was successfully used to classify soybean grains samples of three different geographical origins from Argentina. Supervised classification techniques (RF, SVM-DA, k-NN) and class-

modeling techniques (SIMCA, OC-SVM and PF) were performed on trace element compositions of samples. Among all the models tested, SVM-DA and SIMCA showed the highest percentages in terms of prediction ability in cross-validation with average values of 99.3% for SVM-DA and a median value of balanced accuracy of 96.0% for CHC, 92.0% for COR, 88.0% for SLS using SIMCA. It is important to highlight that although the average accuracy for SIMCA reached lower values, the main advantage of this method is that it has a high capacity to identify unmodelled samples, which in the case of supervised classification techniques are unable to detect. For future studies, we expect that some limitations found in our present research can be addressed, such as the expansion of soybean data from other regions.

5. References

1. Kemsley EK, Defernez M, Marini F. Multivariate statistics: Considerations and confidences in food authenticity problems. *Food Control*. 2019;**105**:102-112. doi:10.1016/J.FOODCONT.2019.05.021.

2. Wadood SA, Boli G, Xiaowen Z, Hussain I, Yimin W. Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchem J*. 2020;**152**:104295. doi:10.1016/j.microc.2019.104295.

3. Badia-Melis R, Mishra P, Ruiz-García L. Food traceability: New trends and recent advances. A review. *Food Control*. 2015;**57**:393-401. doi:10.1016/j.foodcont.2015.05.005.

4. Borràs E, Ferré J, Boqué R, Mestres M, Aceña L, Busto O. Data fusion methodologies for food and beverage authentication and quality assessment – A

review. *Anal Chim Acta*. 2015;**891**:1-14. doi:10.1016/j.aca.2015.04.042.

5. Granato D, Putnik P, Kovačević DB, et al. Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing. *Compr Rev Food Sci Food Saf*. 2018;**17**(3):663-677. doi:10.1111/1541-4337.12341.

6. Otaka A, Hokura A, Nakai I. Determination of trace elements in soybean by X-ray fluorescence analysis and its application to identification of their production areas. *Food Chem*. 2014;**147**:318-326. doi:10.1016/j.foodchem.2013.09.142.

7. Mataveli LRV, Pohl P, Mounicou S, Arruda MAZ, Szpunar J. A comparative study of element concentrations and binding in transgenic and non-transgenic soybean seeds. *Metallomics*. 2010;**2**(12):800-805. doi:10.1039/c0mt00040j.

8. Berrueta LA, Alonso-Salces RM, Héberger K. Supervised pattern recognition in food analysis. *J Chromatogr A*. 2007;**1158**(1-2):196-214. doi:10.1016/J.CHROMA.2007.05.024.

9. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Springer; 2013.

10. Oliveri P, Downey G. Multivariate class modeling for the verification of food-authenticity claims. *TrAC Trends Anal Chem*. 2012;**35**:74-86. doi:10.1016/j.trac.2012.02.005.

11. Oliveri P. Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial. *Anal Chim Acta*. 2017;**982**:9-19. doi:10.1016/J.ACA.2017.05.013.

12. Bro R, Smilde AK. Principal component analysis. *Anal Methods*. 2014;**6**(9):2812-2831. doi:10.1039/C3AY41907J.

13. Gemperline P. *Practical Guide to Chemometrics*. CRC press; 2006.

14. Varmuza K, Filzmoser P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC press; 2016.

15. Mees C, Souard F, Delporte C, et al. Identification of coffee leaves using FT-NIR spectroscopy and SIMCA. *Talanta*. 2018;**177**:4-11. doi:10.1016/J.TALANTA.2017.09.056.

16. Guerbai Y, Chibani Y, Hadjadji B. The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters. *Pattern Recognit*. 2015;**48**(1):103-113. doi:10.1016/J.PATCOG.2014.07.016.

17. Marini F. *Chemometrics in Food Chemistry*. Vol 28. Newnes; 2013.

18. Team RC. R: A language and environment for statistical computing. 2015.

19. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal Methods*. 2013;**5**(16):3790-3798. doi:10.1039/C3AY40582F.

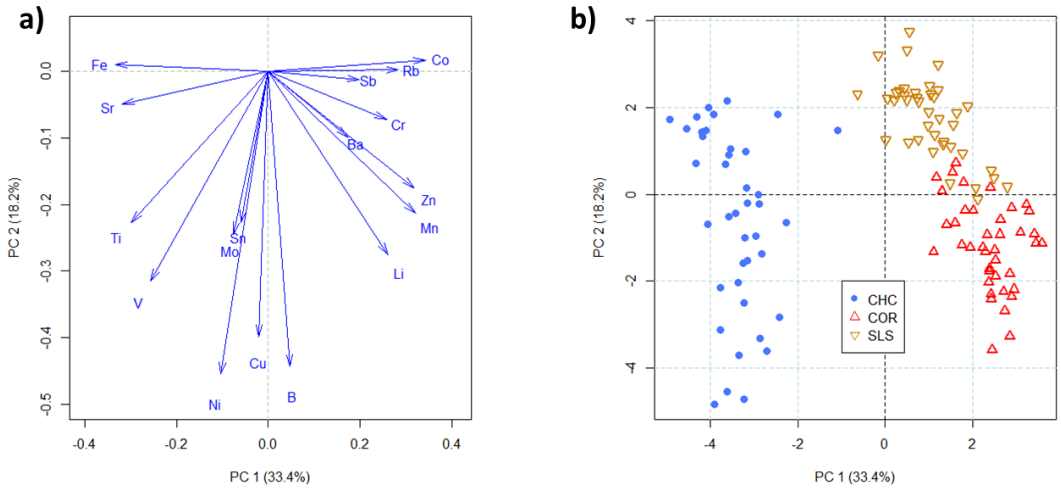
20. Barbosa RM, de Paula ES, Paulelli AC, et al. Recognition of organic rice samples based on trace elements and support vector machines. *J Food Compos Anal*. 2016;**45**:95-100. doi:10.1016/j.jfca.2015.09.010.

Figure captions

Fig. 1. PCA results: (a) Loading plot of PC1 vs PC2 of PCA performed using all the determined trace element concentrations; (b) the corresponding score plot of PC1 vs PC2 with the scores identified according to their geographical origin: CHC Chaco, COR Córdoba and SLS San Luis.

Fig. 2. Box plot comparing the supervised classification chemometrics models applied for soybean seeds classification.

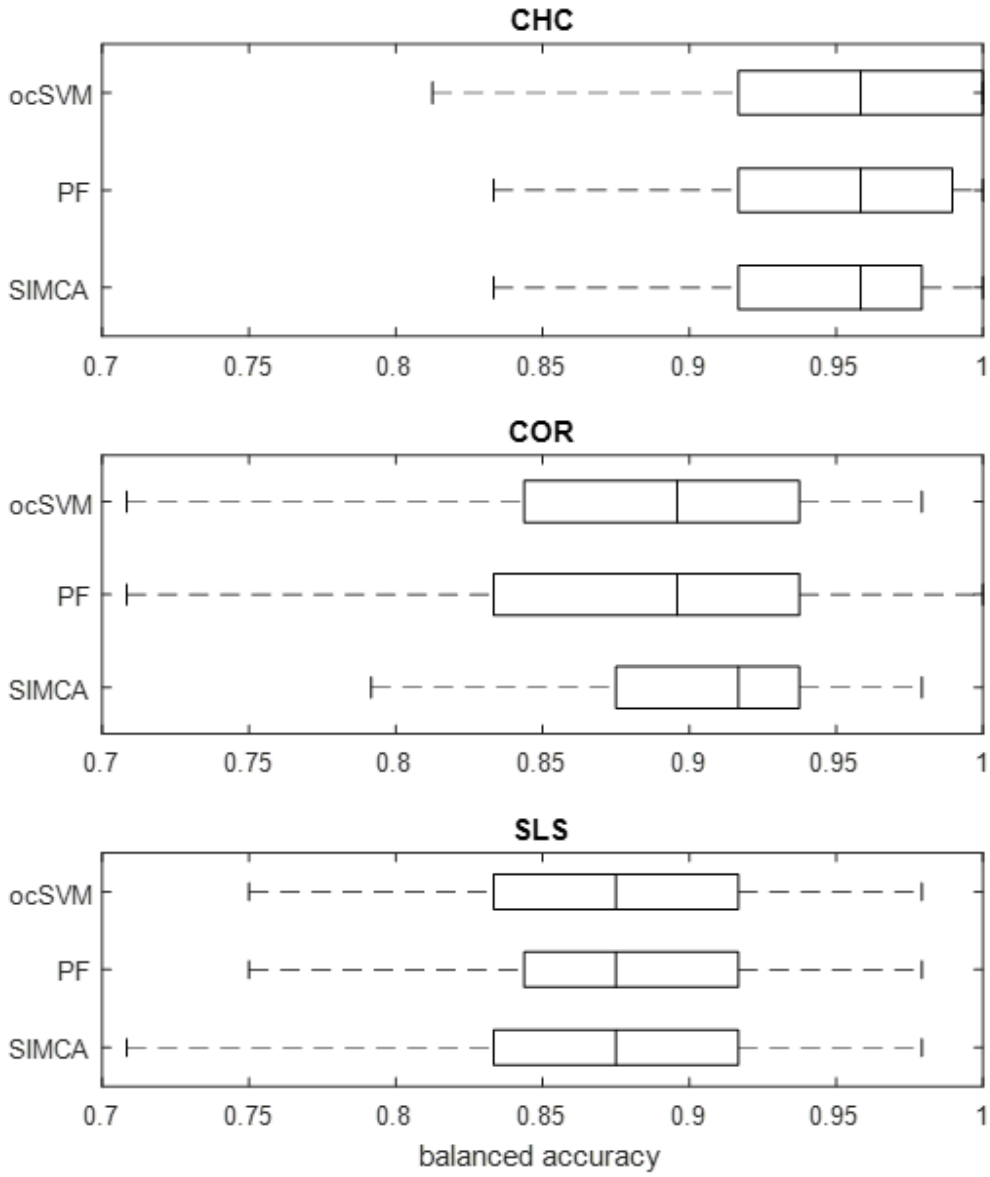
Fig. 3. Hotelling T2 versus Q residuals for the simples from Chaco (CHC) SIMCA model.



450

451 Figure 1

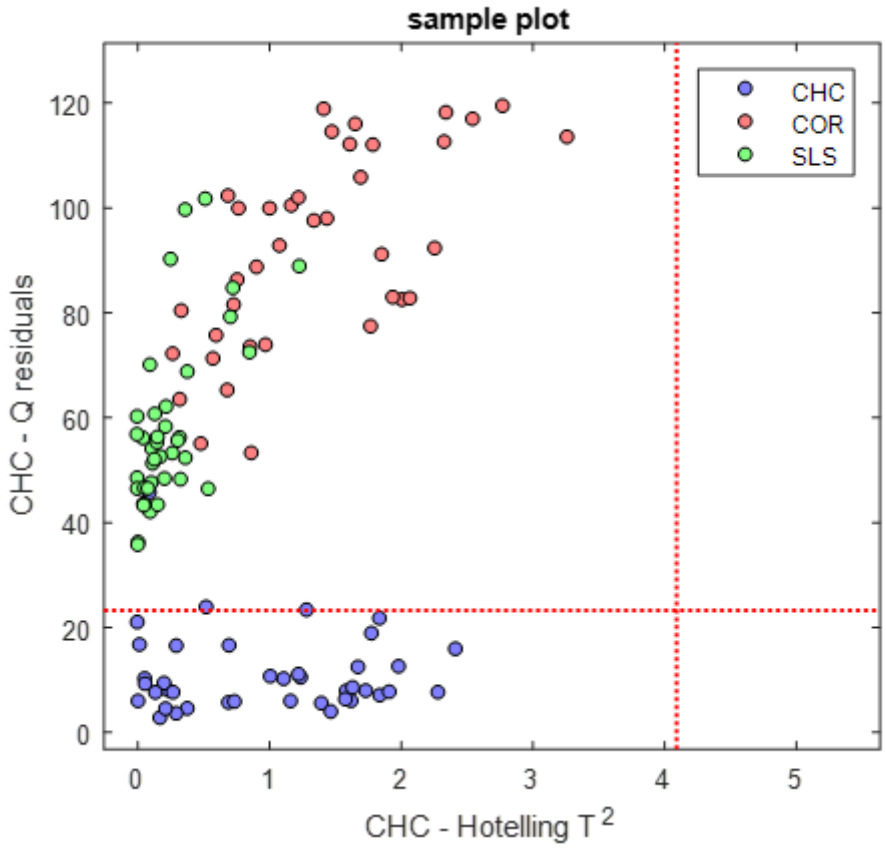
452



453

454 Figure 2

455



456

457 Figure 3

458

459 Table I. Trace element composition of soybean seeds according to their geographical origin.

Element	Certified Values (µg g^{-1})	Recovery Percentage (%)	Geographical origin (µg kg^{-1}) Chaco *n* = 40	Córdoba *n* = 40	San Luis *n* = 40	*P* value
Ag	-	-	0.05 (nd – 0.15)	0.05 (nd – 0.15)	0.05 (nd – 0.10)	ns
B	33.1	98.0	3.3 a (0.3 – 4.6)	3.4 a (1.7 – 4.8)	1.5 b (0.5 – 2.8)	***
Ba	-	-	2.0 a (0.6 – 4.9)	4.8 b (0.4 – 7.2)	3.4 b (2.0 – 6.4)	**
Co	0.58	97.7	1.1 a (0.8 – 1.4)	1.4 b (1.3 – 1.5)	1.4 b (1.3 – 1.4)	**
Cr	1.99	100.1	1.2 a (0.8 – 3.4)	3.2 b (1.3 – 3.9)	1.8 c (0.9 – 3.8)	***
Cu	4.70	98.2	20.2 a (2.8 – 44.1)	18.3 b (3.9 – 27.8)	5.1 b (1.9 – 18.4)	***
Fe	367.5	97.9	162 a (140 – 190)	135 b (120 – 148)	138 b (120 – 150)	**
Li	-	-	3.3 a (0.9 – 6.8)	10.6 b (3.0 – 16.7)	4.9 c (0.9 – 14.8)	***
Mn	246.3	104.0	24 a (10 – 69)	98 b (56 – 132)	64 c (33 – 109)	***
Mo	-	-	2.4 a (1.1 – 5.3)	2.6 b (0.5 – 4.6)	0.9 b (0.2 – 3.7)	***
Ni	1.58	100.8	4.2 a (1.3 – 10.8)	4.1 a (1.9 – 5.2)	1.5 b (0.6 – 3.7)	***
Pb	-	-	0.05 (nd – 0.10)	0.05 (nd – 0.08)	0.05 (nd – 0.10)	ns
Rb	14.8	99.3	8.8 a (5.2 – 16.5)	15.4 b (9.2 – 16.3)	14.3 b (6.4 – 16.3)	***
Sb	-	-	0.06 a (nd – 0.06)	0.07 b (nd – 0.07)	0.06 a (nd – 0.07)	**
Se	0.05	98.5	0.12 (nd – 0.12)	0.12 (nd – 0.12)	0.12 (nd – 0.15)	ns
Sn	-	-	0.8 a (nd – 1.4)	2.5 b (nd – 3.2)	1.8 b (nd – 2.7)	***
Sr	-	-	5.1 a (2.1 – 6.2)	3.1 a (2.1 – 3.9)	3.3 b (2.1 – 5.4)	***
Ti	-	-	6.4 a (2.0 – 7.9)	3.3 b (2.2 – 6.1)	2.8 b (2.1 – 3.8)	**
V	0.83	97.6	11.2 a (3.9 – 17.4)	2.4 b (1.7 – 2.5)	2.2 b (2.0 – 3.8)	**
Zn	30.9	99.9	11.3 a (10.5 – 20.4)	30.6 b (20.8 – 31.4)	20.5 c (12.2 – 31.2)	***

460 Nonparametric Kruskal-Wallis test was applied: ns. not significant at $p > 0.05$; *. $p < 0.05$; **. p
461 < 0.01; ***. $p < 0.001$. Pairwise comparison, different letters a, b or c, in the same row indicate
462 significant differences ($p < 0.05$).
463
464

465

Table II. Classification results achieved with the different chemometrics models.

Method	Number of samples		Classification metrics		
	Training set	Testing set	Balanced accuracy (%)	Sensitivity (%)	Precision (%)
k-NN[a]	28	12	83.4	83.4	83.4
SVM-DA[b]	**28**	**12**	**91.7**	**91.7**	**91.7**
RF[c]	28	12	83.4	83.4	91.7

[a] *k*: number of neighbors = 5

[b] C: penalty factor = 16; Gamma: intensive loss function: 0.039

[c] nt: number of trees = 500; mtry: number of variables tried in each split = 7