Microchemical Journal 129 (2016) 243–248

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Tracing the geographical origin of Argentinean lemon juices based on trace element profiles using advanced chemometric techniques

Jose E. Gaiad [a], Melisa J. Hidalgo [a], Roxana N. Villafañe [b], Eduardo J. Marchevsky [b,*], Roberto G. Pellerano [a,*]

[a] Instituto de Química Básica y Aplicada del Nordeste Argentino (IQUIBA-NEA), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste (UNNE), Av. Libertad 5470, Corrientes 3400, Argentina

[b] Instituto de Química San Luis (INQUISAL), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Facultad de Química, Bioquímica y Farmacia, Universidad Nacional de San Luis (UNSL), Av. Ejército de los Andes 950, San Luis 5700, Argentina

ARTICLE INFO

Article history:
Received 31 March 2016
Received in revised form 23 June 2016
Accepted 1 July 2016
Available online 02 July 2016

Keywords:
Citrus limon
Multi-element analysis
Geographical origin
ICP-MS
Chemometrics

ABSTRACT

This study examines the application of chemometric techniques associated with trace element concentrations for origin evaluation of lemon juice samples. Seventy-four lemon juice samples from three different provinces of Argentina were evaluated according to their microelement contents to identify differences in patterns of elements in the three provinces. Inductively coupled plasma mass spectrometry (ICP-MS) was used for the determination of twenty-five elements (Ag, Al, As, Ba, Bi, Co, Cr, Cu, Fe, Ga, In, La, Li, Mn, Mo, Ni, Rb, Sb, Sc, Se, Sn, Sr, Tl, V, and Zn). Once the analytical data were collected, supervised pattern recognition techniques were applied to construct classification/discrimination rules to predict the origin of samples on the basis of their profiles of trace elements. Namely, linear discriminant analysis (LDA), partial least square discriminant analysis (PLS-DA), k-nearest neighbors (k-NN), random forest (RF), and support vector machine with radial basis function Kernel (SVM). The results indicated that it was feasible to attribute unknown lemon juice samples to its geographical origin. SVM had better performance compared to RF, k-NN, LDA and PLS-DA, listed in descending order. Eventually, this study verifies that trace element pattern is a powerful geographical indicator when identifying the origin of lemon juice samples by analyzing trace element data with the help of SVM technique. This level of accuracy provides an interesting foundation to propose the combination of trace element contents with SVM technique as a valuable tool to evaluate the geographical origin of lemon juice samples produced in Argentina.

1. Introduction

Natural lemon (*Citrus limon* (L.) Burm.) juices are rich sources of antioxidant nutrients such as flavonoids, carotenoids, and vitamin C; essential elements, K, Cu, Fe, Mg, and Zn; and soluble as well as insoluble dietary fiber. Together these nutrients promote several health benefits and provide protection against several illnesses [1,2]. In general, lemon fruits can be commercialized as fresh fruits, juices or oil in the international market. Today, Argentina produces around 1500×10^3 metric tons a year, mainly of the Genova and Eureka varieties. Taking the world production (around 6000×10^3 metric tons/year), Argentina's share is 20% approximately [3]. The lemon fruit annual production can vary between countries because of climatic variations.

The most important producing regions of lemon fruits in Argentina are mainly concentrated in two principal regions: *Argentine northwest (NW)* that is a region formed by three provinces, Tucuman, Salta and Jujuy, which share their borders with Bolivia and Chile; and *Argentine Northeastern* (NE) formed by three provinces, namely Entre Rios, Corrientes, and Misiones, which share their borders with Brasil, Paraguay and Uruguay. The NW region presents important competitive advantages in relation to the NE, principally because this region presents very low incidence of some important botanical diseases, that are enabling exports of lemon juices and fruits to markets such as the European Community or United States, principally [4].

For these reasons, producers, traders and consumers are especially interested in correct labelling of origin and traceability in lemon juice products. Determination of geographical origin authenticity is an important issue for the growing food industry in quality control and safety of food [5]. In Argentina, there are no established methodology to determine the traceability or geographical origins of Argentinean lemon juices [6]. In this context, the use of multielemental analytical techniques has increasingly been used to determine the geographical origin

* Corresponding authors.
E-mail addresses: eduardo.marchev@gmail.com (E.J. Marchevsky), roberto.pellerano@comunidad.unne.edu.ar (R.G. Pellerano).

http://dx.doi.org/10.1016/j.microc.2016.07.002
0026-265X/

of foods [7]. In general, the techniques to obtain elemental fingerprint of food are those with multi-element detection capability, such as ICP-based techniques [8]. Chemometric analysis of the complex element composition data obtained by these instrumental methods provides a better interpretation and the possibility to acquire relevant information about genuineness of these foods [9].

The main objectives of this work were the determination of trace element contents of lemon juice samples and the use of that chemical information to obtain adequate classification models to authenticate Argentinean lemon juice samples. Accordingly, an ICP-MS method has been proposed and the contents of 25 trace elements (Ag, Al, As, Ba, Bi, Co, Cr, Cu, Fe, Ga, In, La, Li, Mn, Mo, Ni, Rb, Sb, Sc, Se, Sn, Sr, Tl, V, and Zn) have been determined in lemon juice obtained from fruits cultivated in Argentina and derived from the three best-known lemon producing Argentinean regions: Northwest (Jujuy and Tucumán provinces) and Northeast (Corrientes province). In order to distinguish the geographical origin of the considered lemon juice samples, pattern recognition techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS DA), k nearest neighbors (kNN), support vector machines (SVM), and random forest (RF) have been applied.

2. Material and methods

2.1. Reagents

All the chemicals used were of the highest purity available and all the glass materials used were washed with nitric acid and rinsed with ultrapure water. Ultrapure deionized water with a resistivity of 18.1 MΩ cm^{-1} was used exclusively. Ultrapure grade 65% (m/m) HNO_3 was acquired from Sigma (St. Louis, MO, USA). Nitric acid was further purified by sub-boiling distillation. Mono and multi-element standard solutions of trace analysis grade were purchased from Sigma-Aldrich and Agilent.

2.2. Apparatus

A high-performance microwave digestion oven, Milestone® (Chicago, USA) model Ethos One was used to digest the samples. The trace elements concentrations in digested samples has been carried out by Agilent 7700 cx (Agilent Technologies, Santa Clara, CA) ICP-MS spectrometer powered by a 27.12 MHZ radiofrequency solid-state generator at 1500 W. The ICP torch was a Fassel-type torch. The ICP torch consists of a three-cylinder assembly, with injector diameter 2.5 mm. Ni sampler and skimmer cones of 1.0 mm and 0.4 mm were used. This instrument was equipped with a MicroMist glass concentric nebulizer combined with a cooled double-pass spray chamber made of quartz. To suppress polyatomic interferences originating from sample matrix, octopole reaction system (ORS) with 5 mL/min He as collision gas and kinetic energy discrimination mode was used (collision mode). The equipment is provided with off-axis ion lens, a quadrupole mass analyzer and an electron multiplier detector. All instrument parameters were optimized daily while aspirating the tuning solution. The selected isotopes for measurement were ^{107}Ag, ^{27}Al, ^{75}As, ^{137}Ba, ^{209}Bi, ^{59}Co, ^{53}Cr, ^{63}Cu, ^{56}Fe, ^{71}Ga, ^{115}In, ^{139}La, ^{7}Li, ^{55}Mn, ^{95}Mo, ^{60}Ni, ^{85}Rb, ^{121}Sb, ^{45}Sc, ^{78}Se, ^{118}Sn, ^{88}Sr, ^{205}Tl, ^{51}V, and ^{66}Zn.

2.3. Sample collection

In this work, we analyzed 74 fresh lemon juice samples derived from Argentinean mature fruits collected from different agricultural cooperatives and producers during 2014/2015. The lemon fruits obtained correspond to three botanical varieties: Eureka ($n = 25$), Lisboa ($n = 30$) and Genova ($n = 19$). Four different locations in the north region of Argentina were selected for fruit collection (Fig. 1). Three different sites located in two provinces corresponding to NW region: Tucumán (TN-I: Tafí Viejo and TN-II: Famaillá), and Jujuy (Jy: Santa Clara) provinces and one site corresponding to NE region: Corrientes (CTE: Bella Vista) province. In order to maintain homogeneity with respect to the collection season all samples were simultaneously collected in the different sites. Finally, it is important to emphasize that all fruit considered in this study were produced under the recommendations formulated by INTA (Instituto Nacional de Tecnología Agropecuaria, Argentina) for the application of agrochemicals for this crop.

Once in the laboratory, the fruits were cleaned and washed with deionized water. Juice was extracted with a domestic plastic reamer and strained to remove seeds. The samples were not centrifuged or filtered, other than remove large particles from fresh juices. All samples were freeze-dried for a minimum of 48 h at a chamber pressure of 0.05 mbar, homogenized and stored in labelled polyethylene zipper bags.

2.4. Analytical procedures

Approximately 1.0 g of dry samples were placed into a microwave-closed vessel, added to each flasks 2.0 mL of 30% (m/m) H_2O_2 and 6.0 mL of sub-boiling HNO_3 65% (m/m) and stood for 10 min. The microwave digestion program applied included the next temperature stages:

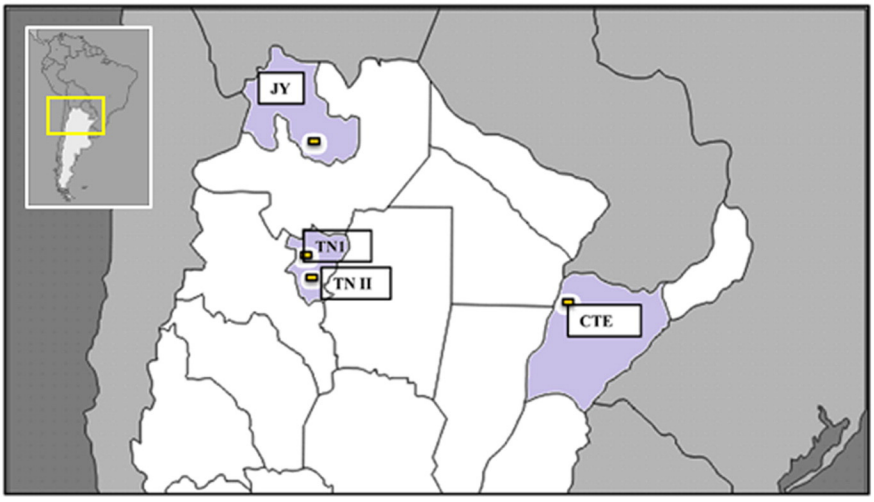

Fig. 1. Production areas of Argentinean lemon samples analyzed in this work: TN-I (Tafí Viejo, Tucumán), TN-II (Famaillá, Tucumán), JY (Santa Clara, Jujuy), and CTE (Bella Vista, Corrientes).

(1) 25–200 °C for 15 min, (2) 200 °C for 15 min and (3) 200–110 °C for 15 min, followed immediately by ventilation at room temperature (20 min). Then, the acid digested samples were diluted to a final volume of 10 mL with 1 M HNO_3 [10]. Blank solutions were prepared in the same way as the samples.

The digested samples and the blank solution were measured by direct nebulization. The selection of isotopes for each target element was carried out by checking the absence of polyatomic, isobaric and physical interferences. The calibration curves were obtained at five different concentration levels in triplicate. The coefficient of determination (R^2) values of such linear fit ranged from 0.9980 to 0.9997. An internal standard solution of 100 μg/L ^{115}In and ^{80}Y, obtained from Agilent (Santa Clara, CA), was added to each sample in order to correct for any instrument drift during analysis and avoid mistakes in the quantification of different analytes.

2.5. Quality assurance

Several parameters were evaluated for the validation of the analytical method followed, for determination of trace elements in lemon juices. Accuracy was measured analyzing at least one certified reference material (CRM) in each analytical batch. Sample concentrations for each certified element were recalculated on the base of the obtained recoveries, ranging between 94 and 107% using SRM 1570a for Al, As, Co, Cu, Mn, Ni, Se, Sr, V and Zn. Because not all studied elements were considered in the certified material, recovery studies in spiked samples were carried out in randomly selected lemon juice solutions at concentrations of 10 and 100 μg/kg. The solutions were analyzed under robust conditions in triplicate. Good recovery values depicted in Table 1, were in the range of 92–104%. The obtained recoveries confirmed that no significant element losses occurred during the digestion procedure.

The limits of detection (LOD) for all selected elements, calculated as a concentration that yields three times the standard deviation of blank signal, are shown in Table 1. As shown, the calculated LODs allowed the determination of all trace elements at the required levels.

Inter-day ($n = 7$) and intra-day precision data were obtained for all the analyzed elements. The inter-day data were generated on different days using new calibrations curves and new instrument tuning. The relative standard deviations were lower than 7.3% and 5.2% for all the elements for inter- and intra-day data, respectively.

Table 1
Limits of detection (LOD), limits of quantification (LOQ), coefficient of determination (R^2), precision and recovery of samples spiking by ICP-MS.

Element	LOD (μg/kg)	LOQ (μg/kg)	R^2	RSD (%)	Spiked concentrations (μg/kg)	Recovery (%)
Ag	3.1	9	0.9997	1.4	10	99.0
Al	185	560	0.9987	2.5	100	98.6
As	2.8	8	0.9995	0.9	10	97.2
Ba	21.8	66	0.9985	1.5	100	94.0
Bi	0.9	2	0.9980	2.1	10	100.1
Co	4.4	13	0.9981	2.0	10	99.2
Cr	7.8	23	0.9987	1.8	10	97.5
Cu	44.7	135	0.9992	1.2	100	96.3
Fe	11.3	34	0.9986	0.8	100	102.1
Ga	4.7	14	0.9992	2.4	10	98.3
In	2.3	7	0.9891	2.3	10	104.0
La	11.7	35	0.9998	1.1	10	101.1
Li	18.2	55	0.9986	3.1	10	98.2
Mn	22.6	68	0.9980	1.2	100	96.8
Mo	10.3	31	0.9997	0.7	10	97.3
Ni	20.8	63	0.9988	2.7	100	99.4
Rb	30.4	92	0.9992	1.5	100	106.3
Sb	2.1	6	0.9980	0.9	10	98.7
Sc	1.5	4	0.9995	1.8	10	103.2
Se	17.7	53	0.9990	1.5	10	99.9
Sn	1.8	5	0.9991	0.8	10	101.0
Sr	23.6	71	0.9986	2.1	100	101.4
Tl	12.3	37	0.9997	1.8	10	103.7
V	9.1	27	0.9984	0.9	10	100.0
Zn	157	475	0.9982	0.6	100	97.7

RSD: Relative standard deviation ($n = 3$).

2.6. Chemometrics

Basic multivariate characterization of the investigated lemon juice samples was made by principal component analysis (PCA). This unsupervised technique shows the natural grouping of the studied samples as well as the variables in a multidimensional space. Additionally, PCA reduce the number of variables used to describe data [11]. After that, five classification techniques such as linear discriminant analysis (LDA), k-nearest neighbor (K-NN), partial least square discriminant analysis (PLS-DA), random forest (RF), and support vector machine (SVM) were investigated for correct identification of provenance of lemon juice.

2.6.1. LDA

The purpose of LDA is to discriminate between classes by maximizing the variance between classes and minimizing the variance within each class. LDA defines canonical or discriminant functions that are linear combinations of the original variables that optimize that separation. The model creates a centroid that is the mean position of all points in all directions. The prediction result for the test set is obtained by the projection of the new samples according the minimal distance to the centroid of each class [12].

2.6.2. PLS-DA

PLS-DA is a linear classification method based on partial least squares regression (PLS) algorithm for constructing predictive models when the factors (independent variables) are many and highly collinear. PLS algorithm searches for latent variables with maximum covariance with the dependent variables which represent class membership. As such, PLS takes into account the dependent variable when defining latent variables. This technique become an established tool in chemometric modeling, because it is often possible to interpret the extracted factor in terms of the underlying physical system. In general, the PLS-DA method often is reported to work well in practice [13].

2.6.3. k-NN

K-nearest neighbor is a non-linear discriminant technique focusing on distances between objects, and in particular on the closest objects. This method classifies unknown samples by projection into the multivariate space and assigning to the class of its closest neighbor in the training set. k is a parameter to be optimized and represents the number of neighbor is taken into account to decide by majority vote the class of unknown samples [14].

2.6.4. SVM

SVM is a learning machine method that creates a mapping of the training data into a high-dimensional space, SVM compute an optimal separation hyperplane by means of an iterative algorithm learning the sample distribution in the boundaries of each considered class [15]. The complexity (C value) of the model is controlled by a penalty error function in order to avoid over-fitting. In this study, we used radial basis function (RBF) kernel was used for classification. Radial basis function (RBF) kernel was selected because of its effectiveness and speed in training process. In addition a grid-search and cross-validation were used to optimize the parameters C-value and ε, that best fit the model and improve the accuracy results [16].

2.6.5. RF

Random Forest is an ensemble learning method that combines a bootstrap aggregation (bagging) to form sub-samples sets and tree decision predictors for classification. This method uses the averaging in the response to improve the predictive accuracy and control over-fitting

Table 2
Trace element concentrations in 74 lemon juice samples from Argentina (dry weight) analyzed by ICP-MS.

Element	Mean (μg/g)	Std Dev (μg/g)	Element	Mean (μg/g)	Std Dev (μg/g)
Al	3.2	0.02	Mo	0.25	0.01
Ba	5.4	0.64	Ni	1.4	0.24
Bi	0.15	0.05	Rb	10.2	1.57
Co	0.025	0.15	Sc	0.02	0.01
Cr	0.35	0.05	Se	0.35	0.05
Cu	9.4	1.22	Sn	0.2	0.02
Fe	15.6	1.18	Sr	8.7	0.84
La	0.4	0.02	V	0.15	0.02
Li	0.3	0.01	Zn	11.6	0.23
Mn	7.5	0.44			

[17]. One of the main advantages of this method is that generally achieves high levels of accuracy, usually much higher than the accuracy obtained with a single decision tree [18]. In addition, this method requires low computational cost for large datasets and gives estimates of what variables are important in the classification.

All data analyzes were performed using R software version 3.2.1 [19] with caret package for classification models training and testing [20].

3. Results and discussion

3.1. Trace element contents in Argentinean lemon juices

The results of total concentrations of trace elements are presented in Table 2. They are expressed as mean values for three replicates with standard deviations (SDs). The concentrations of six trace elements (Ag, As, Ga, In, Sb, and Tl) are not showed in Table 2 because they were below the limits of detection (LODs) in all samples.

Similar concentration profiles between samples can be observed in all samples. The trace element concentrations (above 1.0 μg/g) can be arranged in the following order: Fe > Zn > Rb > Cu > Sr > Mn > Ba > Al > Ni. On the other hand, highest discrepancies between ultra-trace element (concentrations below 1.0 μg/g) were observed for La, Cr, Se, Li, Mo, Co, Sn, Sc, V, and Bi in different samples. Considering the average concentration of all samples, the most abundant trace element was Fe followed by Zn and Rb, which showed average contents higher than 10 μg/g. Aluminium, Ba, Cu, Mn, Ni and Sr had content ranging from 1 μg/g to 10 μg/g, whereas Bi, Co, Cr, La, Li, Mo, Sc, Se, Sn and V were all present at concentrations at ultra-trace levels lower than 1 μg/g.

The concentration levels found for Co, Cr, Fe, La, Sc, and Zn were in accordance with data reported in the literature for natural lemon juices [3,6,21]. In comparison with other citric fruit juices, such as oranges or grapefruits, the reported contents of Al, Co, Cu, Fe, Li, Mn, Ni, Sr, V, and Zn were at similar levels [22,23]. In general, the contents of Rb were low while Ba, and Mo were at high levels than reported for fresh Australian orange juice [22].

Several factors may contribute to variations in trace elements levels in lemon juice according to their geographical origin. First, is the availability of the element in the soil for uptake by the plant. This availability depends principally on the soils cation exchange capacity which can vary between soil types, pH and mineral matrix composition. Other factors, such as agricultural practices, fertilizer applications, artificial irrigation, or fruit maturity at harvest, can influence the trace element concentrations among other reasons [24].

3.2. Geographical differentiation of Argentinean lemon juices

As preliminary stage, before classification modeling, PCA was carried out for exploratory data analysis. PCA was applied to the autoscaled data matrix, to provide a data structure study in a reduced dimension, keeping the maximum amount of variability present in data. Four principal components with eigenvalues exceeding one were extracted explaining 56.3% of the total variance (25.3%, 14.9%, 9.2% and 6.9%, respectively). Fig. 2 shows the results obtained in the space formed by the two first principal components (PCs).

The results given in Fig. 2a showed a tendency of groupings between samples of the same origin. This figure clearly shows the systematic separation of samples in two principal groups according to PC1, the first group consisting of CTE samples (NE region), and the second group of JY, TN-I and TN-II samples (NW region). This distribution can be interpreted from the loading plot (Fig. 2b) that indicated Fe, La, V, Cu and Zn concentrations are higher for samples from CTE. Extra information can be obtained of these graphs, which suggest that Mo and Cr have highest values for TN samples, while JY samples appear subdivided in two groups, one with high overlap with TN samples and positive projections on PC2, and other group better resolved (low overlapping) by negative scores in PC2, that correspond to low concentrations of Ni and Li. In summary, PCA showed the presence of natural grouping in the samples according to their geographical origin. So, in order to obtain suitable and

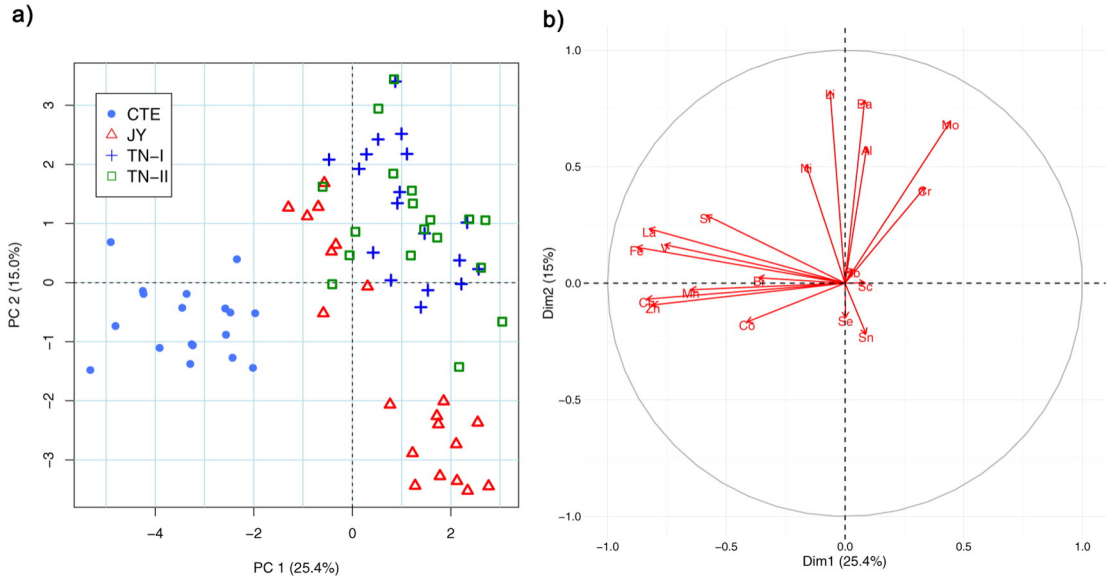

Fig. 2. Score and loading plots of PC1 vs PC2. TN-I (Tafí Viejo, Tucumán), TN-II (Famaillá, Tucumán), JY (Santa Clara, Jujuy), and CTE (Bella Vista, Corrientes).

Table 3
Statistics by class with the discrimination results of different models for the test sets.

Groups	Number of samples		LDA		PLS-DA ($ncomp = 2$)[a]		k-NN ($k = 9$)[b]		SVM ($C = 32$; $\varepsilon = 0.038$)[c]		RF ($nt = 500$; $mtry = 7$)[d]	
	Training set	Test set	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)
CTE	13	5	100	100	100	87	100	100	**100**	**100**	100	100
JY	14	6	100	93	67	87	83	93	**100**	**100**	100	80
TN-I	13	5	60	75	80	80	60	81	**60**	**81**	60	87
TN-II	13	5	–	87	20	100	20	81	**40**	**87.5**	20	93
Mean accuracy (%)			**66.7**		**66.7**		**66.7**		**76.2**		**71.4**	

Best results for each class are indicated in bold.
[a] *ncomp*: number of significant components.
[b] *k*: number of k neighbors.
[c] *C*: penalty factor; ε: ε-insensitive loss function.
[d] *nt:* number of trees; *mtry:* number of variables tried at each split.

effective classification models for differentiate lemon juice samples, supervised learning pattern recognition methods were applied.

In order to build the different models, data matrix was divided in a training set (70% of the cases) to build the models and a test set (30%) to calculate their classification performance. Then, we used *k*-fold cross-validation (repeated *n* times) on training set to build the different classifiers. In *k*-fold cross-validation a set of data is divided in *k* subsets of equal size. One of the subsets is selected for testing and the others are used for training. This procedure is repeated *n* times, so each subset is used for testing at least one time, here we used $k = 10$ and $n = 5$. The random sampling was done according to the geographical origin of samples in order to balance the class distributions within the splits (stratified sampling). The cases included in each set are randomly changed for each reproduced model.

In this work, we tested five different models (LDA, PLS-DA, k-NN, SVM and RF) in order to classify lemon juice according to their geographical origin. PLS-DA, kNN, RF, and SVM need to optimize several parameters in such a way that a suitable number of parameters are selected to build the model. In this work, the choices of number of significant components (*ncomp*) for PLS-DA; number of neighbor *k* for kNN; number of variables tried at each split (*mtry*) and number of trees (*nt*) for RF; penalty factor C, ε of the ε-insensitive loss function and kernel type for SVM, were calculated by using the cross-validation technique described before by which maximum accuracy rate was selected as criterion. Once selected the optimal values for each model, the sensitivity (samples belonging to the class and classified correctly in this class), specificity (samples not belonging to the modeled class and correctly classified as not belonging), and overall accuracy rate were considered for evaluation of the classification achieved with the supervised methods on testing set [25]. Table 3 summarizes the results obtained after the application of the different classification models.

As can be seen, the order of successful identification rates was as follows: SVM > RF > LDA = PLS-DA = kNN. Linear models, such as LDA and PLS-DA, together with k-NN presented similar performance from the overall accuracy point of view, however do not solve the classification problem. In general, the samples from the NE region (CTE) can be classified correctly by the five proposed models, except by PLS-DA (that misclassified two JY samples as CTE). RF presented a good performance to classify CTE and JY samples, but it fails to resolve the samples from TN-I and TN-II. The best classification for lemon juice samples was achieved using the SVM method with RBF kernel function, the cost and gamma functions were 32 and 0.038, respectively. The overall classification accuracy was 76.2%, because the similarities between TN-I (3 samples misclassified) and TN-II (2 samples misclassified) groups in terms of trace element contents. The classification was 100% correct for the other samples (CTE and JY groups). The better classification achieved by SVM, a nonlinear model, generally is due to the flexibility and ability of the algorithm for creating a generalized model, even for small training groups.

3.3. Comparison of proposed models for classification

Finally, we additionally compare the performance of the five proposed classification models by using the repeated one-third holdout estimator considering only the provinces of origin as grouping factor (i.e. three groups: CTE, JY and TN: TN-I + TN-II). This method reserves 33% of samples for testing and uses the remainder for training. Since the result will depend on the initial choice of the test data, the procedure is repeated 50 times. This gives a distribution for the overall accuracy, in order to compare the performance of each method at optimal parameter configuration found for each, the accuracy results from each of the

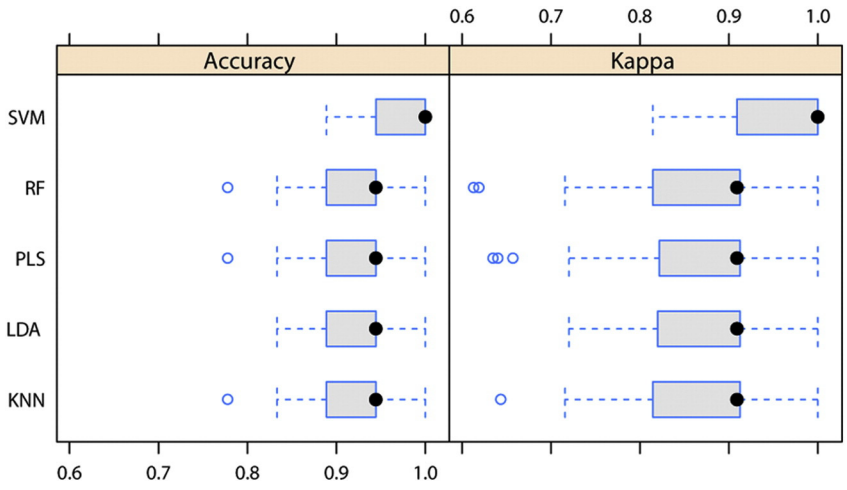

Fig. 3. Box and whisker plot comparing model results.

best models were collected. The distributions obtained are shown in Fig. 3 represented by Box and Whisker plots.

The limits around the mean represent the widths of the confidence intervals for the means. The results obtained confirm the best method for this data set was SVM, with a median accuracy of 100% and kappa value of 98.7%.

4. Conclusions

In the present study, the levels of 25 trace elements were measured by ICP-MS in 74 lemon juice samples. Ag, As, Ga, In, Sb, and Tl were not detected in any of the analyzed samples (they were below the LOD). The exploratory analysis showed the existence of patterns to separate the samples according to their geographical origin. In addition, the results obtained from applying five different classification techniques: SVM, RF, K-NN, LDA and PLS-DA, were compared using holdout cross-validation with stratified sampling of test subset. Each method was optimized by repeated 10-fold cross-validation. The performance of SVM resulted to be the best, with a success rate of 76% in the tested set, followed by RF with 71%. The samples from the NE region were correctly classified with respect to the NW samples. The samples collected from the NW region presented difficulties in their resolution in regard to the site of origin. The lemon juice samples from the two sites of the province of Tucuman were confused in the discrimination process. These samples could be influenced by the geographical or environmental similarities. The best results are achieved when considering the samples province of origin as a classification factor. Our results demonstrate the potential of SVMs as a chemometric pattern recognition tool for the origin identification of lemon juice in Argentina.

Acknowledgements

The authors are grateful to Universidad Nacional del Nordeste (Project: F012/2014 SGCyT-UNNE) and to Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), for financial support. EJ Marchevsky and RG Pellerano are members of research career, and JE Gaiad, MJ Hidalgo and RN Villafañe are fellow of CONICET.

References

[1] J. Lorente, S. Vegara, N. Martí, A. Ibarz, L. Coll, J. Hernández, M. Valero, D. Saura, Chemical guide parameters for Spanish lemon (*Citrus limon* (L.) Burm.) juices, Food Chem. 162 (2014) 186–191.
[2] A.C. Matheyambath, P. Padmanabhan, G. Paliyath, Citrus Fruits, Academic Press, Oxford, Encyclopedia of Food and Health, 2016 136–140.
[3] United State Department of Agriculture, Citrus Fruit | USDA Foreign Agricultural Service, http://www.fas.usda.gov/commodities/citrus-fruit2016 (Accessed: 2016/03/15).
[4] Federcitrus, La Actividad Citrícola Argentina, Federcitrus, 2015 http://www.federcitrus.org/noticias/upload/informes/La_Actividad_Citricola_2015.pdf (Accessed: 2016/03/25).
[5] M. Amenta, G. Ballistreri, S. Fabroni, F.V. Romeo, A. Spina, P. Rapisarda, Qualitative and nutraceutical aspects of lemon fruits grown on the mountainsides of the Mount Etna: a first step for a protected designation of origin or protected geographical indication application of the brand name 'Limone dell'Etna', Food Res. Int. 74 (2015) 250–259.
[6] R.G. Pellerano, S.S. Mazza, R.A. Marigliano, E.J. Marchevsky, Multielement analysis of Argentinean lemon juices by instrumental neutronic activation analysis and their classification according to geographical origin, J. Agric. Food Chem. 56 (13) (2008) 5222–5225.
[7] S. Kelly, K. Heaton, J. Hoogewerff, Tracing the geographical origin of food: the application of multi-element and multi-isotope analysis, Trends Food Sci. Technol. 16 (12) (2005) 555–567.
[8] A. Gonzalvez, S. Armenta, M. de la Guardia, Trace-element composition and stable-isotope ratio for discrimination of foods with protected designation of origin, TrAC Trends Anal. Chem. 28 (11) (2009) 1295–1311.
[9] M. Forina, M. Casale, P. Oliveri, 4.04 - Application of Chemometrics to Food Chemistry, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford 2009, pp. 75–128.
[10] H. Altundag, M. Tuzen, Comparison of dry, wet and microwave digestion methods for the multi element determination in some dried fruit samples by ICP-OES, Food Chem. Toxicol. 49 (11) (2011) 2800–2807.
[11] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (9) (2014) 2812–2831.
[12] S. Moncayo, S. Manzoor, F. Navarro-Villoslada, J.O. Caceres, Evaluation of supervised chemometric methods for sample classification by laser induced breakdown spectroscopy, Chemom. Intell. Lab. Syst. 146 (2015) 354–364.
[13] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (16) (2013) 3790.
[14] X. Huang, E. Teye, J.D. Owusu-Sekyere, J. Takrama, L.K. Sam-Amoah, L. Yao, C.K. Firempong, Simultaneous measurement of titratable acidity and fermentation index in cocoa beans by electronic tongue together with linear and non-linear multivariate technique, Food Anal. Methods 7 (10) (2014) 2137–2144.
[15] H. Li, Y. Liang, Q. Xu, Support vector machines and its applications in chemistry, Chemom. Intell. Lab. Syst. 95 (2) (2009) 188–198.
[16] H. Zheng, H. Lu, A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (*Mangifera indica* L.), Comput. Electron. Agric. 83 (2012) 47–51.
[17] M. Kuhn, K. Johnson, Nonlinear classification models, in: M. Kuhn, K. Johnson (Eds.), Applied Predictive Modeling, Springer New York, New York, NY 2013, pp. 329–367.
[18] B.L. Batista, L.R.S. da Silva, B.A. Rocha, J.L. Rodrigues, A.A. Berretta-Silva, T.O. Bonates, V.S.D. Gomes, R.M. Barbosa, F. Barbosa, 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 (1) (2012) 209–215.
[19] R Core Team, The R Project for Statistical Computing, R: A Language and Environment for Statistical Computing, https://www.r-project.org2014 (Accessed: 03/10/2016).
[20] M. Kuhn, Caret package, J. Stat. Softw. 28 (5) (2008).
[21] J. Tufour, J. Bentum, D. Essumang, J. Koranteng-Addo, Analysis of heavy metals in citrus juice from the Abura-Asebu-Kwamankese District, Ghana, J. Chem. Pharm. Res. 3 (2) (2011) 397–402.
[22] W.A. Simpkins, H. Louie, M. Wu, M. Harrison, D. Goldberg, Trace elements in Australian orange juice and other products, Food Chem. 71 (4) (2000) 423–433.
[23] A. Szymczycha-Madeja, M. Welna, Evaluation of a simple and fast method for the multi-elemental analysis in commercial fruit juice samples using atomic emission spectrometry, Food Chem. 141 (4) (2013) 3466–3472.
[24] A. Szymczycha-Madeja, M. Welna, D. Jedryczko, P. Pohl, Developments and strategies in the spectrochemical elemental analysis of fruit juices, TrAC Trends Anal. Chem. 55 (2014) 68–80.
[25] M.C.A. Marcelo, C.A. Martins, D. Pozebon, V.L. Dressler, M.F. Ferrão, Classification of yerba mate (*Ilex paraguariensis*) according to the country of origin based on element concentrations, Microchem. J. 117 (September 2015) (2014) 164–171.