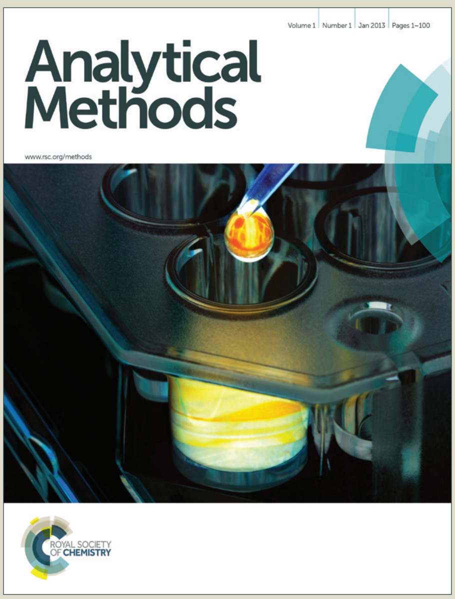
Analytical Methods

Accepted Manuscript



Journal Name RSCPublishing

ARTICLE

Methods of multivariate analysis of NIR reflectance spectra for classification of yerba mate

Cite this: DOI: 10.1039/x0xx00000x

Marcelo Caetano Alexandre Marcelo[a], Camila Alves Martins[a], Dirce Pozebon[a]*, Marco Flôres Ferrão [a,b]

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

The present article is about a method of classification of yerba mate (*Ilex paraguariensis*) from South America. Yerba mate samples were ground in cryogenic mill and the reflectance of milled samples in near infrared (NIR) was directly measured. Hierarchical cluster analysis (HCA), principal components analysis (PCA), k-nearest neighbour (kNN), soft independent modelling class analogy (SIMCA), partial least square discriminant analysis (PLS-DA), support vector machine discriminant analysis (SVM-DA) were used for multivariate analysis of the NIR reflectance spectra. Fifty four brands of yerba mate from Argentina, Brazil, Paraguay and Uruguay were analyzed in order to classify the commercialized product according to the country of origin. For all intervals of the NIR reflectance spectra evaluated (4,435-4,318 cm^{-1}; 4,358-4,200 cm^{-1}; 4,436-4,200 cm^{-1}; and 4,673-4,200 cm^{-1}), the classification of all brands was 100% correct by SVM-DA. For KNN, classification was not 100% correct in any interval. Classification by PCA, HCA and SIMCA was 100% correct for the 4,435-4,318 cm^{-1} interval; for PLS-DA it was for the 4,358-4,200 cm^{-1} and 4,435 - 4,318 cm^{-1} intervals.

1 Introduction

Near infrared spectroscopy (NIR) together with chemometrics have been used for identification of origin of food products such as vegetable oils [1-3], wine and honey bee [4,5] among other applications.[6,7] The NIR spectrum is a kind of "fingerprint" of a given product[8] and through multivariate analysis of the spectrum it is possible to identify the product origin.[9] The main advantages of NIR are the cheaper and faster information acquired in a non-destructive way and potential for routine analysis.[10,11]

Ilex paraguariensis (named yerba mate or simply mate) has medicinal properties and infusions of yerba mate leaves have been widely consumed for centuries in southern Latin America (mainly in northeastern Argentina, Uruguay, Paraguay and southern Brazil).[12] Moreover, in the last decade yerba mate has also been cultivated for other uses, including addition to cosmetic, beverages, and food.[13] This way, yerba mate is now used worldwide.

Yerba mate contains bioactive compounds such as polyphenols (chlorogenic acids), alkaloids (caffeine and theobromine), flavonoids (rutin and luteolin), and saponins (metasaponins). These compounds have been associated with antioxidant, anticarcinogenic, antiallergic, diuretic, and hypocholesterolaemic properties of the yerba mate.[14-17]

The levels of the bioactive compounds and others in the final product are linked to the mode of processing of yerba mate. The overall processing technique can be divided in five main stages: harvesting (cutting thin branches of the trees), blanching (toasting at wood flame for 3 to 5 minutes) to inactivate enzymes and prevent bad taste, drying (for 30 minutes in furnace at 100 ^{0}C), grinding and stationing for several days.[18] Subsequently, the product is minced and separated into leaves and twigs. According to product regulations and specifications, a proper mixture of leaves and twigs is undertaken and then the product is packaged. The duration of each stage and the temperature varies from country to country, giving a different feature to the commercialized yerba mate, as desired by consumers.

It has been observed that the concentration of sugar decreased after blanching whereas those of polyphenols and antioxidants increased.[19] The form of cultivation and age of the yerba mate tree also influence on the concentration level of several compounds. For example, the incidence of light can affect the concentration of theobromine, caffeine, methylxanthine, [20] vitamin E, and stigmasterol.[21] If yerba mate fruits are processed along with the leaves, the saponins concentration increases because they are higher in the fruit.[22] Addition of leaves of other South American *Ilex* species changes pharmacological activities and taste because those leaves do not contain caffeine and the saponins concentration is different.[23, 24]

Foods and beverages are usually related to their place of origin, which has become increasingly important, bearing in mind the increased interest for products with quality and origin identification.[25] Producers directly associate the brand with the product provenance in order to obtain market recognition, better prices and other commercial advantages. In conjunction with regulatory authorities, producers have interest in ensuring correct product labelling as well as methodologies for authentication.[26]

The growth conditions of the yerba mate plant, harvesting and leaves processing as well as adulteration can affect the final product. In this sense, the aim of this study is to develop a method for classification of commercialized yerba mate according to country in South America. To this end, multivariate analysis of the NIR reflectance spectra is used.

2 Experimental Part

2.1 Instrumentation and Sample Preparation

The NIR spectra were obtained using a PerkinElmer 400 IR spectrometer equipped with integrating sphere and indium gallium-arsenic (In-Ga-As) detector. A Spectralon disc was used for the background registration. The reflectance was measured in the range of 10,000-4,000 cm^{-1} whereas the resolution was 4 cm^{-1}. Thirty two scans were performed for each sample; all of them were run in random order and in triplicate. An aliquot (50 g) of each yerba mate brand was ground in cryogenic mill (Spex Certiprep, 6750 Freezer Mill, USA) and directly analyzed. The cryogenic grinding was carried out in argon atmosphere where the sample was frozen for two minutes and then ground for two minutes under 20 beats per second. Preliminary tests revealed that particles size was of paramount importance. Any trend of classification was not observed if the packaged yerba mate was directly analyzed or if it were milled only in agate mortar, ball mill, or knife mill.

2.2 Samples

Fifty-four brands (19 from Brazil, 14 from Paraguay, 14 from Argentina and 7 from Uruguay) of commercial yerba mate were analyzed. One package (0.5 or 1 kg) of each brand was acquired in local markets of the countries. The number of brands available for each country was different. The geographic origin and additional information (use of pesticides, addition of sugar etc.) were taken from the label of the packages.

2.3 Multivariate Analysis

The software Matlab 7.11 (MathWorks Inc., USA), PLS-TOOLBOX 6.2.1 (Eigenvector Research Inc., USA) and iToolbox (http://www.models.kvl.dk) were used. For multivariate analysis, the NIR spectra were normalized and derivatized at first order (second degree polynomial and 15 points were used). Hierarchic cluster analysis (HCA),[28,29] principal components analysis (PCA),[30] k-nearest neighbour (kNN),[28] soft independent modelling class analogy (SIMCA),[31] partial least square discriminant analysis (PLS-DA),[32] and support vector machine discriminant analysis (SVM-DA)[33] were used for multivariate analysis of the reflectance spectra. The Kennard-Stone algorithm[34] was employed for the selection of training and test sets.

3 Results and Discussion

3.1 Exploratory Analysis

The8,000 - 4,200 cm^{-1} spectral region was selected due to excessive noise observed for the other NIR regions. For each sample, the reflectance values were normalized to the maximum value observed for the sample in order to correct multiplicative scaling effects on the spectrum. As a consequence, all samples have impacted the discriminant analysis model similarly. For each sample, the data were then derivatized at first order using the Savitzy-Golay algorithm (2nd order polynomial and 15 point per window). The derivatization of the data removed the samples baseline signals and emphasized differences among the samples. Low noise was observed in the interval selected of the spectra and it was not necessary to smooth the spectra. In the following step, the data related to all samples together were mean centred for each wavelength of the spectra. The last treatment facilitated the subsequent exploratory analysis and classification model[35]. Fig. 1(a) shows the raw spectra (each spectrum is the mean of 3 spectra) the mean of all samples (54 brands) and Fig 1(b) does the mean spectra after data pre-treatment. Other methods of pre-treatment were also evaluated: a) standard normal variation (SNV) - this weighted normalization method considers that all samples deviate from the mean following a normal distribution;[35] b) SNV and derivation to first order; and c) normalization. However, it was observed that these pre-treatments were not satisfactory for classification.

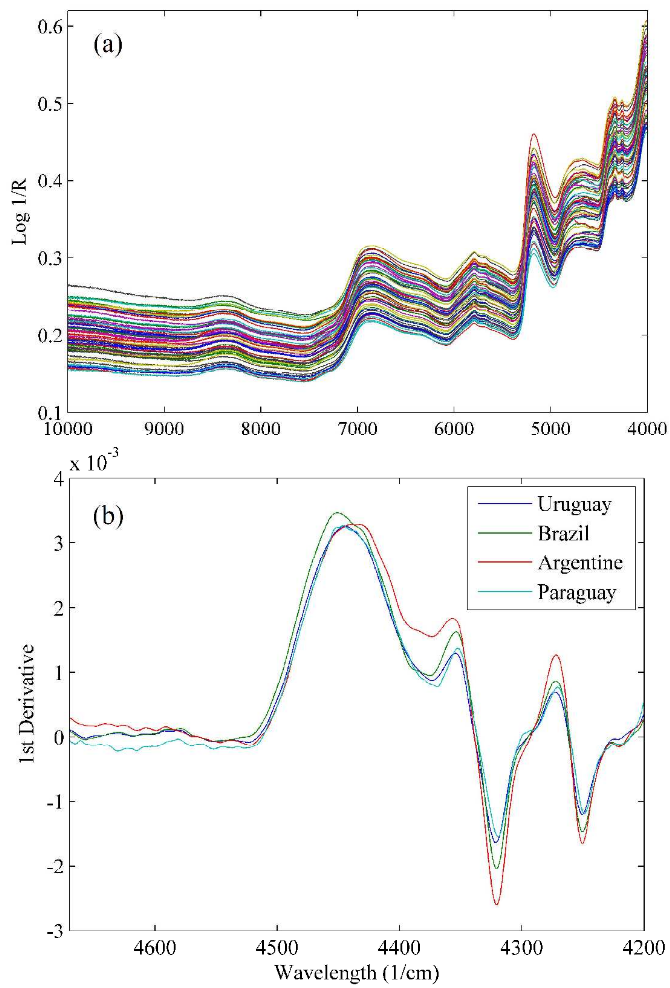


Fig. 1 Raw spectra of NIR reflectance (R) of the 54 yerba mate samples (a) and mean spectra (4,200 - 4,670 cm^{-1}) after normalization and derivation to first order (b).

3.2 Unsupervised Methods (HCA and PCA)

In order to find the region of the spectra that would provide the correct classification of yerba mate and eliminate irrelevant information and noise, i-PCA was performed. This algorithm divides the considered interval of the spectrum into i intervals. That is, for i = 2, the NIR spectra were divided into two intervals, each interval with the same number of variables. The PCA was then carried out for each interval separately. For all samples, i was 1 (global PCA), 2, 4, 8, 12, 16, 24, and 32. Thereby, the classification of the samples according to the country of origin was better for the 4,435-4,318 cm^{-1} interval.

Fig. 2 shows the dendrogram obtained by HCA of the NIR spectral region considered, for all 54 brands of yerba mate. The linkage Ward's method and Mahalanobis distance were used to construct the dendrogram. Four clusters are revealed by HCA, with correct classification of all brands according to the country of origin. The dendrogram shows that Brazilian and Uruguayan samples have similarity, and the same is observed among the Argentine and Paraguayan samples.

The scores of PCA of the three first components for the 4,435-4,318 cm^{-1} region of the NIR spectra are shown in Fig. 3. For the separation of all groups (countries), three principal components (PCs) were necessary. The PC1, accounting for 79.64% of the variance, separated the Uruguayan and Brazilian samples of those from Paraguay. The PC2, accounting for 8.78% of the variance, separated the samples in three groups: Brazil, Argentina and a third group composed by Paraguay and Uruguay. Nevertheless, only PC1 and PC2 did not separate the countries and, therefore, an additional component was necessary. Then, PC3, with 4.95% of the total variance, enhanced the separation.

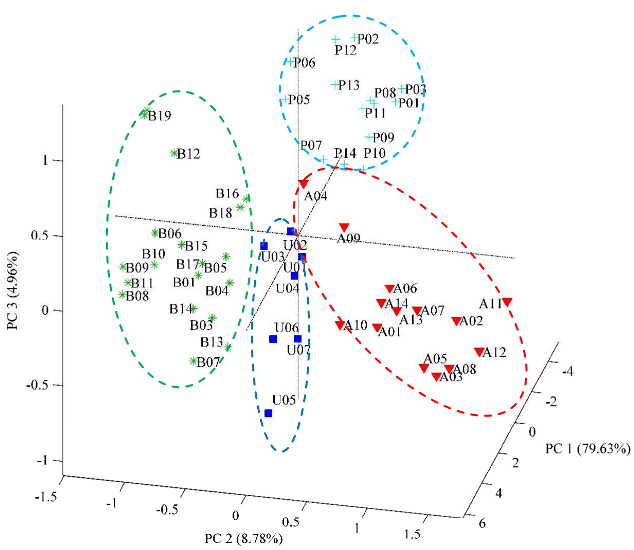


Fig.3 Scores of the three first principal components in the space, for the NIR reflectance spectra (4,435 a 4,318 cm^{-1} interval); A = Argentina, B = Brazil, P = Paraguay and U = Uruguay.

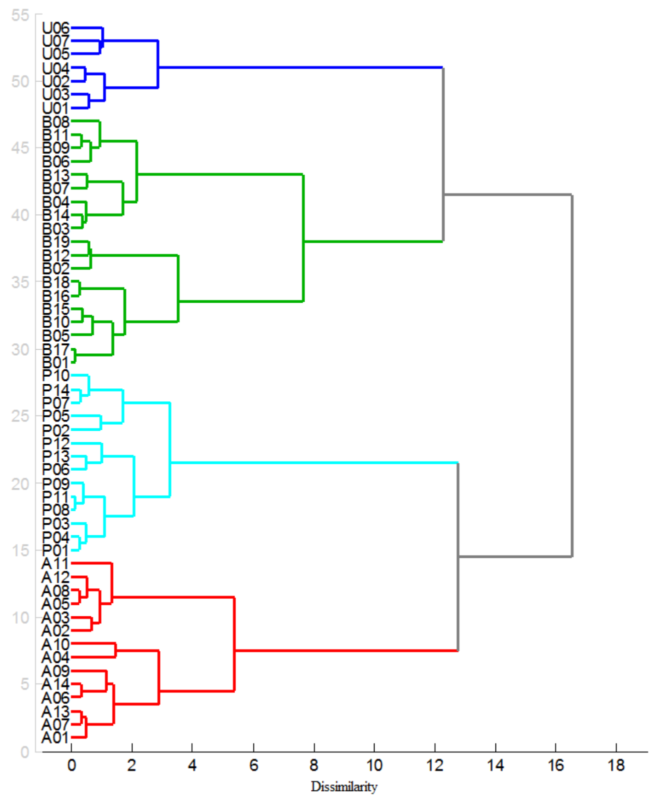


Fig. 2 Dendrogram of the NIR reflectance spectra of the yerba mate (4,435 a 4,318 cm^{-1} interval); A = Argentina, B = Brazil, P = Paraguay e U = Uruguay.

The results obtained by HCA and PCA (unsupervised multivariate analysis methods) revealed that both methods can be applied for yerba mate classification. Nevertheless, in order to confirm the results obtained, other methods of multivariate analysis were employed for additional intervals of the NIR spectra.

3.3 Supervised Methods (kNN, SIMCA, PLS-DA and SVM-DA)

A classification model was created whereby it was possible to predict the origin of the brand. To this end, all 54 samples were grouped according to the Kennard-Stone algorithm, which separated the brands into two sets (training and test) by similarity. The training set was composed of 35 brands whereas the test set was of the remaining 19 brands. These two sets were used for supervised multivariate analysis of the 4,435-4,318 cm^{-1} (i = 32), 4,358-4,200 cm^{-1} (i = 24), 4,436-4,200 cm^{-1} (i = 16) and 4,673-4,200 cm^{-1} (i = 8) regions of the NIR spectra. The IR signals in these regions are due to combinations of functional groups CH, CH_2 and CH_3[36] of the various compounds present in the yerba mate, such as xanthines, theobromines and saponins.

The results obtained for kNN, SIMCA, PLS-DA and SVM-DA are summarized Table 1. Sensitivity (brands belonging to the class and classified correctly in this class) and specificity (brands not belonging to the modeled class and correctly classified as not belonging) were used in order to evaluate the classification.

The kNN method generally leads to good results when there are small differences among the samples of the same group. However, it provides little information about the structure of the classes and variables that are responsible for the classification.[37] In the present study, k = 3 was used for kNN because the number of Uruguayan brands was low (7). The classification for the training group was unsatisfactory for i = 32, i = 24 and i = 16 due to the low number of samples. However, even with a small value of k the classification was correct for almost all brands in the test group (see Table 1).

For SIMCA, the leave-one-out method was used for cross-validation. The number of PCs was selected aiming the best classification with the lowest possible number of PCs. For i = 32, all Brazilian and Paraguayan brands were modeled with 3 PCs whereas the Argentine and Uruguayan were with 2 PCs. As can be seen in Table 1, for i = 32 all brands were correctly classified. For i = 24, all Uruguayan and Paraguayan brands were modeled with 3 PCs whereas the Argentine were with 2 PCs and the Brazilian were with 4 PCs. In that way, only one sample was misclassified (B17 was classified as Uruguayan); for i = 16, the Argentine, Brazilian, Paraguayan and Uruguayan brands were modeled with 4, 3, 2, 3 PCs, respectively, whereas for i = 8 they were with 3, 5, 2, 3 PCs, respectively. Unlike KNN, modeling a larger part of the spectra by means of SIMCA worsened the classification (see Table 1).

The leave-one-out method for cross-validation was also used for PLS-DA. The classification by PLS-DA is shown in Fig. 4 for i = 32, with 5 latent variables (LVs). The threshold (see the red line in Fig. 4) for each class was obtained by using the Bayes theorem[38] and the respective data. The classification was 100% correct for all brands in the training and test sets, for i = 32 and i = 24 (4 LVs). For i = 16 (6 LVs), sample B10 was a false positive in the Uruguayan model whereas samples B10 and A12 were for i = 8 (5 LVs) in the same model. Likewise SIMCA, classification was worse when a larger part of the spectra was modeled using PLS-DA (see Table 1).

All brands were correctly classified (100%) by SVM-DA in the training and test sets for all intervals (Table 1). For all models, the kernel function was the radial basis function; for i = 32, 24, 16 and 8, the cost function values were 31.62, 10000, 10000 and 100, respectively, while the gamma function were 100, 1000, 100, 316.22, respectively. The best classification obtained using SVM-DA was not only due to the ability of the model to generalize, even when the training set is small, like the Uruguayan class, but also the non-linear modeling and the flexibility of the model. In addition, the selection of the appropriate kernel function makes the classification model more robust and more efficient.[39]

The usefulness of NIR spectra combined with multivariate data analysis to classify yerba mate according to their geographical origin was previously observed in a preliminary study conducted by Cozzolino *et al*. [27] However, they analyzed only 5 brands of yerba mate from Uruguay, Argentina and Brazil. The yerba mate brands were correctly classified by linear discriminant analysis (LDA), but were not by PLS-DA. However, the authors stated that further development was necessary.

The worst classification by PLS-DA could be due to the low number of brands analyzed, which would not represent the population. In the present study, a larger number of brands were analyzed (54) and an additional country (Paraguay) was included. The better classification achieved in the present study by using PLS-DA as well as good classification by other methods of multivariate analysis employed can be attributed to the larger number of brands analyzed, lower particles size, and appropriate variables selection. The variables selection is very important, allowing to separation of the relevant information that is necessary for the classification. In this way, interfering such as water content can be eliminated - water content influence was not observed in the intervals of spectra evaluated. In the study conducted by Cozzolino *et al*, the yerba mate was dried at 100 ^{o}C for 24 h but in the present study the yerba mate was simply milled and the reflectance spectra directly obtained.

Journal Name RSCPublishing

ARTICLE

Table 1 Sensitivity and specificity of classification using different methods of multivariate analysis of the NIR reflectance spectra of yerba mate

				PLS-DA			SVM-DA			SIMCA			kNN		
Interval (cm^{-1}), i	Country	Number of Samples (n) in the Sets		Sensitivity (%)		Specificity (%)	Sensitivity (%)		Specificity (%)	Sensitivity (%)		Specificity (%)	Sensitivity (%)		Specificity (%)
		Training	Test	Training	Test		Training	Test		Training	Test		Training	Test	
4,435-4,318 i = 32	Argentina	8	6	100	100	100	100	100	100	100	100	100	88	100	98
	Brazil	14	5	100	100	100	100	100	100	100	100	100	85	100	97
	Paraguay	8	6	100	100	100	100	100	100	100	100	100	88	100	98
	Uruguay	5	2	100	100	100	100	100	100	100	100	100	80	100	98
4,358-4,200 i = 24	Argentina	8	6	100	100	100	100	100	100	100	100	100	88	100	98
	Brazil	14	5	100	100	100	100	100	100	100	100	100	85	100	94
	Paraguay	8	6	100	100	100	100	100	100	100	100	100	100	83	98
	Uruguay	5	2	100	100	100	100	100	100	100	98	100	80	100	98
4,436-4,200 i = 16	Argentina	8	6	100	100	100	100	100	100	100	100	100	88	100	98
	Brazil	14	5	100	100	100	100	100	100	100	100	94	93	100	97
	Paraguay	8	6	100	100	100	100	100	100	100	83	100	88	100	100
	Uruguay	5	2	100	100	98	100	100	100	100	100	100	80	100	95
4,673-4,200 i = 8	Argentina	8	6	100	100	100	100	100	100	100	83	100	100	100	100
	Brazil	14	5	100	100	100	100	100	100	100	100	91	100	100	97
	Paraguay	8	6	100	100	100	100	100	100	100	66	94	100	100	100
	Uruguay	5	2	100	100	96	100	100	100	100	100	100	80	100	100

Journal Name RSCPublishing

ARTICLE

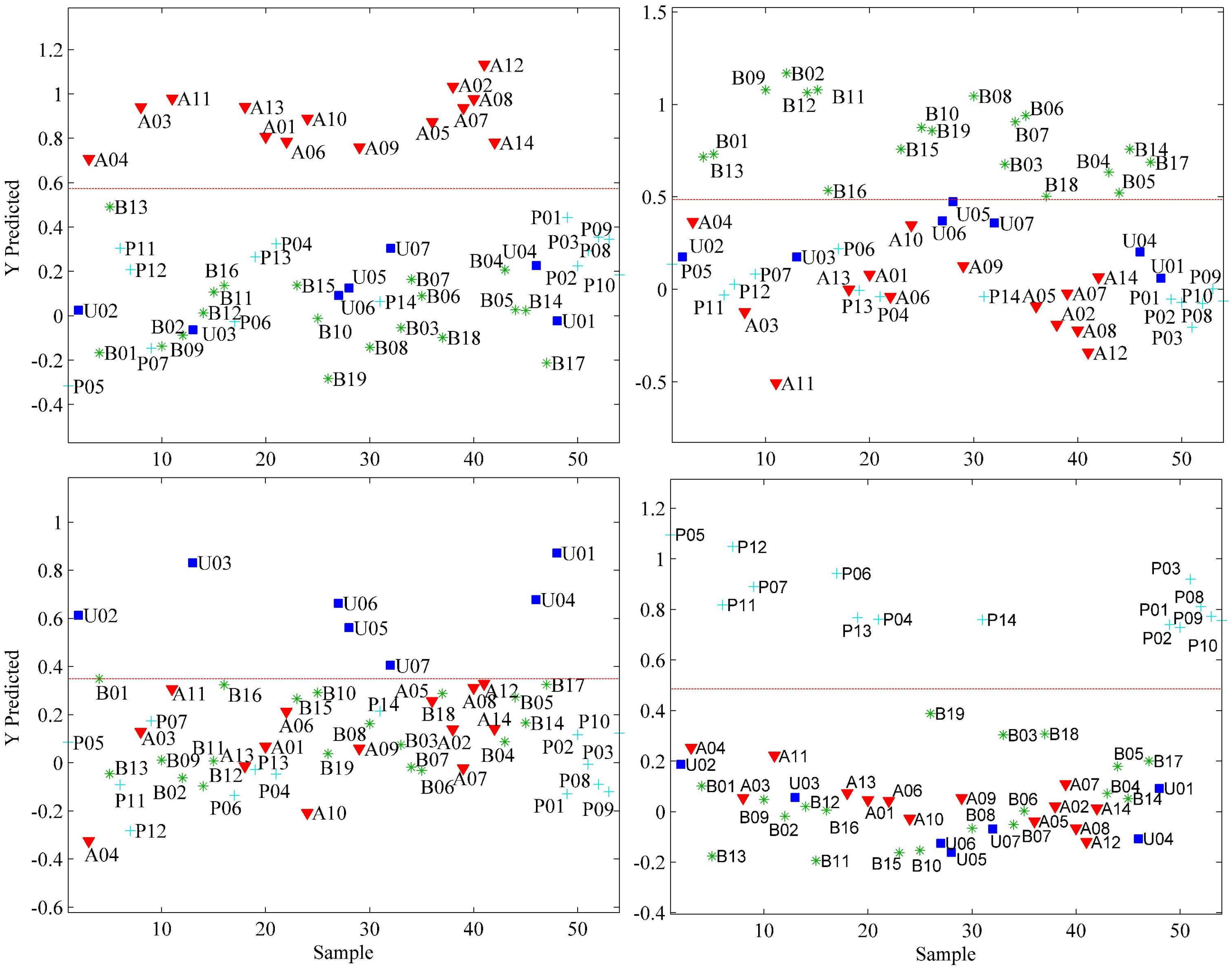


Fig. 4 Yerba mate classification by PLS-DA of the NIR reflectance spectra (4,435 - 4,318 cm^{-1} interval); A = Argentina, B = Brazil, P = Paraguay and U = Uruguay. The red horizontal line (threshold) separates the evaluated class.

Journal Name RSCPublishing

ARTICLE

4 Conclusions

A methodology of yerba mate classification was developed using multivariate analysis of the reflectance NIR spectrum. Through unsupervised multivariate methods, it was possible to classify all brands according to the country of origin correctly. Through supervised multivariate methods, the classification was 100% correct by SVM-DA for all spectral ranges modelled, but, in the case of PLS-DA, the classification was 100% correct for two evaluated ranges of the spectra (4,358-4,200 and 4,435-4,318). Classification by SIMCA was also 100% correct for a smaller range of the spectra (4,435-4,318). The worst results were obtained by kNN where classification was not 100% correct for any spectral range modelled.

Grinding (particles size must be ≤ 100 µm) the sample is the only preparation required for yerba mate, which makes the method simpler and "green" (without use of reagents and/or solvents). The cultivation and, principally, the processing of yerba mate, which is different among the countries, are responsible for the differentiation.

From the results obtained we can conclude that the developed method can be useful for quality control of yerba mate and correct identification. Fraud of yerba mate (*i.e.* mixing with other herbs) could also be detected.

Acknowledgements

The authors are grateful to CNPq for financial support, CAPES and CNPq for scholarships.

Notes and references

[a] Instituto de Química, Universidade Federal do Rio Grande do Sul, Avenida Bento Gonçalves, 9500, 91501-970, Porto Alegre, RS, Brazil.

[b] Instituto Nacional de Ciência e Tecnologia de Bioanalítica, 13083-970, Campinas, SP, Brazil

e-mail: dircepoz@iq.ufrgs.br

The authors intend to publish this article in the themed issue "Analytical Sciences in Brazil".

5 References

1. O. Galtier, N. Dupuy, Y. Le Dreau, D. Ollivier, C. Pinatec, J. Kister, J. Artaud, *Anal. Chim. Acta*, 2007, **595**, 136.
2. D. L. Garcia-Gonzalez, N. Tena, R. Aparicio, *Grasas Y Aceites*, 2012, **63**, 26.
3. S. Tao, J. Li, J. Tang, J. Mi, L. Zhao, *Comput. Comput. Technol. Agric.*, 2012, **369**, 478.
4. K. Ruoff W. Luginbuhl, R. Kunzli,, M. T. Iglesias S. Bogdanov, J. O. Bosset, K. von der Ohe, W. von der Ohe, R. Amado, *Agric. Food Chem.*, 2006, **54**, 6873.
5. F. R. S. Bentlin, F. H. Pulgati, V. L. Dressler, D. Pozebon, *J. Braz. Chem. Soc.*, 2011, **22**, 327.
6. T. Gallardo-Velazquez, G. Osorio-Revilla, M. Zuniga-de Loa Y. Rivera-Espinoza, *Food Res. Int.*, 2009, **42**, 313.
7. F. Liu, Y. He, *Food Res. Int.*, 2008 **41**, 562.
8. C. A. Nunes, *Food Res. Int.*, 2013, in press, doi: org/10.1016/j.foodres.2013.08.041
9. D. Cozzolino, *Food Res. Int.*, 2013, in press, doi: org/10.1016/j.foodres.2013.08.034
10. S. Armenta, S. Garrigues, M. de la Guardia, *Trends Anal. Chem.*, 2008, **27**, 497.
11. C. Pasquini, *J. Braz. Chem. Soc.*, 2003, **14**, 198.
12. N. Bracesco, A. G. Sanchez, V. Contreras, T. Menini, A. Gugliucci, *J. Ethnopharmacol.*, 2011, **136**, 378.
13. C. I. Heck, E. G. De Mejia, *J. Food Sci.*, 2007, **72**, R138.
14. T. Andersen, J. Fogh, *J. Hum. Nutr. Diet.*, 2001, **14**, 243.
15. R. Filip, S. B. Lotito, G. Ferraro, C. G. Fraga, *Nutr. Res.*, 2000, **20**, 1437.
16. G. Schinella, J. C. Fantinelli, S. M. Mosca, Clin. Nutr., 2005, 24, 360.
17. S. Gorzalczany, R. Filip, M. R. Alonso, J. Miño, G. E. Ferraro, C. Acevedo, *J. Ethnopharmacol.*, 2001, 75, 291.
18. M. E. Schmalko, S. M. Alzamora, *Dry. Technol.*, 2001, **19**, 599.
19. N. Dartora, L. M. de Souza, A. P. Santana, M. Iacomini, A. T. Valduga, P. A. J. Gorin, G. L. Sassaki, *Food Chem.*, 2011, **129**, 1453.
20. G. C. Coelho, M. F. G. Rachwal, R. A. Dedecek, G. R. Curcio, K. Nietsche, E. P. Schenkel, *Biochem. Syst. Ecol.*, 2007, **35**, 75.
21. A. A. Esmelindro, J. D. Girardi, A. Mossi, R. A. Jacques, C. Dariva, *J. Agric. Food Chem.*, 2004, **52**, 1990.
22. A. T. C. Taketa, E. Breitmaier, E. P. Schenkel, *J. Braz. Chem. Soc.*, **2004**, 15, 205.
23. V. S. Pires, D. Guillaume, G. Gosmann, E. P, Schenkel, *J. Agric. Food Chem.*, 1997, **45**, 1027.
24. F. H. Reginatto, M. L. Athayde, G. Gosmann, E. P Schenkel, *J. Braz. Chem. Soc.*, 1999, **10**, 443
25. D. Luykx, S. M. Van Ruth, *Food Chem.*, 2008, **107**, 897.
26. R. Karoui, J. De Baerdemaeker, *Food Chem.*, 2007, **102**, 621.
27. D. Cozzolino, E. Restaino, A. Fassio, *Sens. & Instrumen. Food Qual.*, 2010, **4**, 67.
28. R. G. Brereton, in *Chemometrics*, John Wiley & Sons Ltd, England, 2003.
29. M. A. Morgano, S. C. N. Queiroz, M. M. C. Ferreira, Braz. *J. Food Technol.*, 1999, **2**, 73.
30. S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37.
31. M. Monfreda, L. Gobbi, A. Grippa, *Food Chem.*, 2012, **134**, 2283.
32. R. Vitale, M. Bevilacqua, R. Bucci, A. D. Magri, A. L. Magri, F Marini, *Chemom. Intell. Lab. Syst.*, 2013, **121**, 90.
33. A. S. Luna, A. P. da Silva, J. Ferré, R Boqué, *Spectrochim. Acta A*, 2013, **100**, 109.
34. R. W. Kennard, L. A. Stone, *Technometrics*, 1969, **11**, 137.
35. B. M. Wise, J. M. Shaver, N. B. Gallagher, W. Windig, R. Bro, S. Koch, in *PLS_toolbox 4.0 for use with MATLAB™*, Eigenvector Research Inc. USA, 2006.
36. X. B. Zou, J. W. Zhao, M. J. W. Povey, M. Holmes, H. P. Mao, *Anal. Chim. Acta*, 2010, **667**, 14.
37. J. W. Zhao, H. Lin, Q. S. Chen, X. Y. Huang, Z. B. Sun, F. Zhou, *J. of Food Eng.*, 2010, **98**, 408.
38. D. B. Hibbert, N. Armstrong, *Chemom. Intell. Lab. Syst.*, 2009, **97**, 211.
39. A. I. Belousov, S. A. Verzakov, J. von Frese, *Chemom. Intell. Lab. Syst.*, 2002, **64**, 15.