Multivariate calibration of metal content using FTIR spectra of flour peel-mesocarp ashes of *oenocarpus-jessenia* complex fruits

Rojas, J.M.*; Bruns, R. E.**,;Serruya***, H.; Bentes***, M. H.
*Facultad de Ingeniería Química-Universidad Nacional de la Amazonía
Peruana, Iquitos-Perú;
**Instituto de Química da Universidade Estadual de Campinas,
Campinas, SP-Brasil.
H*** Pós-graduação de Química da Universidade Federal de Belém,
Pará-Brasil
Calle Dos de Mayo 1046 –Iquitos: ttilexx@yahoo.com

Abstract

An indirect method of multivariate calibration (MC) has been applied to samples of *oenocarpus* and *jessenia* palms from brazilian and peruvian amazon regions for prediction of metal contents of their flour peel+mesocarp ashes. Metal contents were determined by energy-dispersive x-ray fluorescence (EDXRF). Then, Fourier transform infrared (FTIR) spectra of the same ashes were obtained. MC was performed on the samples in order to develop calibration models for predicting metal contents using mid-infrared spectra as independent variables and EDXRF metal contents as dependent variables. Calibration models providing approximate metal contents in ashes were obtained. K and Cu gave the best models, followed by Fe, Mn, P, Ca and Zn in decreasing order of quality.

Keywords: *oenocarpus-jessenia*, FTIR spectroscopy, multivariate calibration

Introduction

Principal component regression (PCR) and partial least-squares (PLS) regression methods based on principal component analysis (PCA), are widely used for multivariate data analysis in spectroscopic calibration model building[1]. Multivariate calibration (MC) is the process of constructing a mathematical model to relate the output (multiple responses) of an instrument to a property or properties of samples[2,3]. This process involves two steps. First it relates the matrix of independent variables (**X**) and the vector or matrix of dependent variables (**Y)** by an equation **Y** =f(**X**) or **Y** =**X**β, where β is a regression vector. The second step is to predict properties (**Y**$_{prev}$) of new samples, i.e. samples not included in the calibration procedure, given an instrument output (**X**$_{prev}$) and the equation

$$\mathbf{Y}_{prev} = \mathbf{X}_{prev}\beta.$$

The number of principal components (PCs) for a regression model, is generally much less than the number of originally measured variables. The PCs are incorporated into each regression model in decreasing order of the percentage of total variance they explain until some pre-set criterion has been satisfied.

MC is applicable to the determination of major constituents as well as microcomponents and other qualities or properties, and for a very wide range of instrument types. Using near infrared spectroscopy (NIRS) the prediction of protein content in intact, whole wheat kernels[4], the determination of the octane number of gasoline, the characterization of filaments of acrylic fibres[5] and the chemical composition and the energy value of compound feeds for cattle[6] were successfully modelled. Also Fourier transform infrared (FTIR) transmission spectrometry has been applied to the direct determination of glucose in whole blood without any sample preparation[7]

The goal of this paper is to investigate the capability of mid-infrared transmission spectrometry for the indirect determination of metal content in the flour peel+mesocarp ashes of *oenocarpus* and *jessenia* palms samples. The metal content was determined by using energy-

dispersive x-ray fluorescence (EDXRF). Then, the FTIR spectra were obtained. Those spectra were used to develop calibration models for the two genera above, having as dependent variables the metal contents. If these models predict with reasonable accuracy it will be possible to cross information between equipments for constructing calibration models for other systems.

Metodology

Apparatus

Determinations of the metal contents were performed by energy-dispersive x-ray fluorescence using Espectrae-mod. 510 spectrophotometer. The operational instrumental conditions for EDXRF were the followings: Tube voltage:15 KV, Tube current :0.02mA, Filter name: Cellulose, Lifetime:100 s, Max. Energy:20 KeV, Atmosphere: Vacuum, Number of channels :1024.

Infrared spectra were scanned on a BOMEM MB-Series B-100 FTIR spectrometer. The infrared measurements were carry out in the 4000-400 cm^{-1} region. Gain was selected automatically. An apodization cosine was applied, and the interferometer mirror speed was set at 1,9272 cm s^{-1}. The number of scans was 20, resolution 4 cm^{-1}.

Palm samples and chemicals

Twenty three samples of fruits from plants belonging to the *oenocarpus* and *jessenia* genus as specified in Table 1 were used. Fourteen samples (1-14) were of the *oenocarpus* genus and nine samples (15-23) from *jessenia*. KBr for infrared spectroscopy (Merk) was used.

Analytical procedures

From ripe fruits, peel+mescocarp were separated and dried at 103°C[8] until they attained constant weight. Later these samples were ground and powdered. Before being analyzed they were previously mineralized[9] by treating the assay portions of 5 g of powdered sample. Homogenized ash samples were subjected to EDXRF analysis and metal contents expressed as weight

percentages, were determined by this analytical technique.

Ash samples after homogenization with KBr in a 100 mg KBr/1 mg ash ratio with a precision of 0,1 mg were converted into a thin disk after applying a pressure of 5,85 MPa for 1 min. Infrared spectra were scanned on a FTIR spectrometer. Each spectrum was an average of 16 scans over the 4000-400 cm^{-1} wavelength range.

Table 1: Analysed *oenocarpus* and *jessenia* palms samples from Brazilian and Peruvian Amazony .

Sample code	Origin	Genus
1(O.min) [a]	Abaetetuba (Br)	
Oenocarpus		
2(O.min)	Abaetetuba (Br)	"
3(O.map) [d]	Abaetetuba (Br)	"
4(O.min)	Abaetetuba (Br)	"
5(O.bac) [b]	Abaetetuba (Br)	"
6(O.min)	Abaetetuba (Br)	"
7(O.min)	Igarapé Mirim (Br)	"
8(O.min)	Igarapé Mirim (Br)	"
9(O.min)	Igarapé Mirim (Br)	"
10(O.min)	Igarapé Mirim (Br)	"
11(O.map)	Carr Quistococha Km 13 (Perú)	"
12(O.min)	Carr Quistococha Km 13 (Perú)	"
13(O.min)	Carr Iq-N Km 7 F. Pizarro (Perú)	"
14(O.map)	Suni Mirano R-Napo (Perú)	"
15(J.B) [c]	AlpH Carr Iq-N Km 20 (Perú)	*Jessenia*
16(J.B)	AlpH Carr Iq-N Km 20 (Perú)	"
17(J.B)	Carr Iq-N Km 7 F. Pizarro (Perú)	"

18(J.B)	AlpH Carr Iq-N Km 20 (Perú)	"
19(J.B)	AlpH Carr Iq-N Km 20 (Perú)	"
20(J.B)	AplH Carr Iq-N Km 20 (Perú)	"
21(J.B)	AlpH Carr Iq-N Km 20 (Perú)	"
22(J.B)	AlpH Carr Iq-N Km 20 (Perú)	"
23(J.B)	AlpH Carr Iq-N Km 20 (Perú)	"

[a] *Oenocarpus minor;*   Br, Brazil.

[b] *Oenocarpus bacaba;*

[c] *Jessenia bataua;*

[d] *Oenocarpus mapora;*

Data analysis

Seven metals have been determined for each *oenocarpus* and *jessenia* sample. These seven metals were considered as dependent variables: P, K, Ca, Mn, Fe, Cu and Zn (Table 2). Twenty three FTIR spectra (Fig. 1) were scanned and the numerical responses considered as independent variables. The data matrix has 23 sample rows and 1868 spectroscopic response columns with absorbance values (23x1868 matrix).

PCR[10] and PLS[11,12] as multivariate regression methods were used. The computational programs PIROUETTE 2.02[13] and UNSCRAMBLE 6.0[14], furnished by the Chemistry Institute of UNICAMP, Campinas, SP, Brazil were used.

Table 2: Metal content as dependent variables of oenocarpus and jessenia samples by EDXRF.

Sample	P	K	Ca	Mn	Fe	Zn	Cu
1(O.min)	0,289	5,681	0,626	0,018	0,019	0,008	0,004
2(O.min)	0,197	5,504	0,636	0,008	0,019	0,006	0,003
3(O.map)	0,350	6,727	1,041	0,015	0,026	0,023	0,006

|							
4(O.min)	0,201	6,364	0,275	0,005	0,013	0,005	0,004
5(O.bac)	0,417	5,370	0,224	0,009	0,016	0,005	0,004
6(O.min)	0,215	6,282	0,266	0,008	0,015	0,006	0,004
7(O.min)	0,527	7,039	0,795	0,023	0,037	0,014	0,006
8(O.min)	0,300	3,708	0,529	0,007	0,020	0,290	0,005
9(O.min)	0,372	5,863	0,470	0,014	0,020	0,008	0,004
10(O.min)	0,346	6,553	0,553	0,009	0,016	0,008	0,003
11(O.map)	0,416	7,925	1,473	0,078	0,031	0,029	0,010
12(O.min)	0,114	1,627	0,156	0,025	0,009	0,000	0,001
13(O.min)	0,602	7,715	0,571	0,015	0,032	0,018	0,004
14(O.map)	0,463	7,784	1,137	0,189	0,045	0,036	0,007
15(J.B)	0,130	7,093	0,767	0,018	0,031	0,554	0,006
16(J.B)	0,075	3,919	0,857	0,025	0,029	0,007	0,006
17(J.B)*	0,283	5,524	1,291	0,042	0,033	0,016	0,003
18(J.B)	0,158	3,564	2,601	0,090	0,033	0,007	0,006
19(J.B)	0,134	2,899	0,577	0,008	0,014	0,001	0,003
20(J.B)	0,167	2,861	1,136	0,024	0,026	0,005	0,006
21(J.B)	0,222	3,150	2,031	0,120	0,060	0,006	0,004
22(J.B)	0,101	7,642	0,820	0,053	0,045	0,046	0,008
23(J.B)	0,174	4,372	2,135	0,069	0,040	0,050	0,004
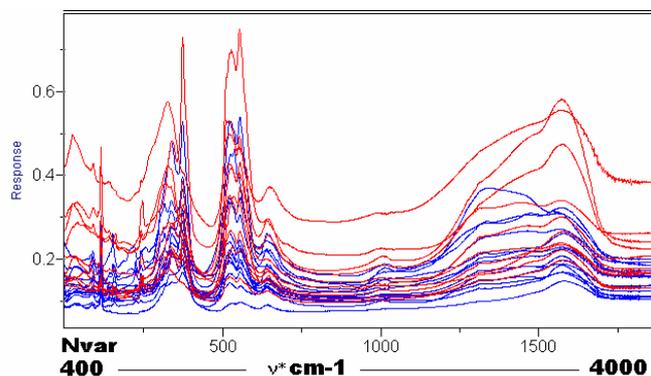
Fig. 1: Infrared spectra plot for *oenocarpus* (blue) and *jessenia* (red) samples.

Results and discussion

Fig. 1 shows the mid-infrared spectra for 23 samples of peel+mesocarp flour ashes in absorbance vs wavenumber. Extensive preprocessing tools were performed on the data. The spectra were normalized after the second derivative was applied to remove systematic variations like the effects of KBr-ash tablet preparation. Multiplicative scatter correction (MSC) for reducing low-frequency sources of variation that are not related to the chemistry under investigation, especially for light scattering problems in reflectance spectroscopy were also applied to the spectra [11].

The data set was randomly divided into two parts, calibration samples (nineteen samples) and test samples (2, 3, 9, and 18) to perform the role of unknown samples, PCR and PLS methods were applied for constructing separate models for each dependent variable.

In the first step all the 1868 independent variables and the seven dependent variables were used. Variables were mean centered, samples were normalised and the second derivative applied. It was necessary to estimate the optimal number of PCs for each PCR and PLS models in order to obtain accuracy and confiability in the predictions. One criterion used is the magnitude of the corresponding percentage variance accounted by each PC, compared with a

chosen stopping criteria in terms of percentage total variance. Another approach is estimating model size and involves validation, the process of evaluating the model's predicitive ability. This is done when cross-validation is applied for PCR or PLS. Then, it is necessary to evaluate the Prediction Residual Error Sum of Squares (PRESS), Standard Error of Validation (SEV), Standard Error of Prediction (SEP). In cross-validation, one sample from the calibration set is temporarily left out and a

$$PRESS = \sum \left(c_i - \hat{c}_i\right)^2 \qquad (1)$$

where, c_i is the true value for the dependent variable of the validation sample and $\hat{c}_i$ the prediction obtained for each temporary validation sample from the built model.

$$SEV = (PRESS/n)^{1/2} \qquad (2)$$

where n is the number of samples in the validation set.

$$SEP = (PRESS/m)^{1/2} \qquad (3)$$

model created from the remaining samples. From this model, a prediction of the left-out sample's dependent variable is made and its residual recorded. The left-out sample is then returned to the calibration set, another sample is excluded, a new model made, and a new prediction and residual generated. This process is repeated until every sample has been left out once. The PRESS calculated from the residual is related to SEV and SEP by the equations:

where m is the number of samples in the test set.[15,16]

$$SEP^* = \{PRESS/(n+m)\}\}^{1/2} \qquad (4)$$

SEP* is used when n and m samples take part in calculating the error.

In order to evaluate numerically the model quality of fit the multiple R-squared value, R^2, called determination coefficient of the model can be used. The maximum R^2 value is 1. The closer the R^2 value is to 1, the

better the model fit to the observed responses[17,18].

$$R^2 = \frac{\sum(\widehat{c}_i - \overline{c})^2}{\sum(c_i - \overline{c})^2}$$

$$= \frac{PRESS}{\sum(c_i - \overline{c})^2}$$

where $\overline{c}$ is the mean of the responses.

Seven PCR and PLS models were obtained. Table 3 shows the experimental values for the dependent variables (EDXRF values) of the test set. Tables 4 and 5 show the predicted values of the test set samples using the PCR and PLS models, including the SEV, PRESS, SEP*, R^2, number of PCs for each model, and the percentage variance values.

Table 3: Metal content (%) by EDXRF for the test set

Sample\c_i	P	K	Ca	Mn	Fe	Zn	Cu
2	0.197	5.504	0.636	0.008	0.019	0.006	0.003
3	0.350	6.727	1.041	0.015	0.026	0.023	0.006
9	0.372	5.863	0.470	0.014	0.02	0.008	0.004
18	0.283	5.524	1.291	0.042	0.033	0.016	0.003

Table 4: Metal content (%) for the test set predicted by PCR models.

Sample\$\widehat{c}_i$	P	K	Ca	Mn	Fe	Cu
2	0.192	5.190	0.727	0.014	0.016	0.005
3	0.327	6.146	1.162	0.087	0.036	0.007
9	0.288	6.886	0.662	0.032	0.021	0.006

18	0.251	5.712	1.276	0.073	0.034	0.007
SEV	0.138	1.466	0.667	0.050	0.011	0.002
SEP*	0.115	0.838	0.495	0.035	0.009	0.0015
PRESS	0.302	16.135	5.624	0.027	0.002	0.00005
R^2	0.33	0.80	0.35	0.51	0.47	0.61
PCs	2	8	3	4	3	6
%Var	59.0	93.0	70.0	77.0	70.0	87

Table 5: Metal content (%) for the test set predicted by PLS

Sample\$\hat{c}_i$	P	K	Ca	Mn	Fe	Cu
2	0.226	5.166	0.665	0.010	0.013	0.005
3	0.303	6.144	1.188	0.094	0.038	0.006
9	0.321	6.735	0.689	0.033	0.022	0.006
18	0.234	5.577	1.403	0.078	0.035	0.007
SEV	0.140	1.610	0.700	0.051	0.012	0.002
SEP*	0.107	0.820	0.391	0.033	0.008	0.00136
PRESS	0.265	15.482	3.518	0.025	0.002	0.00004
R^2	0.41	0.80	0.59	0.59	0.58	0.70
PCs	1	3	2	2	2	3
%Var	30.0	62.0	46.0	31.0	46.0	54.0

According to the results presented in Tables 4 and 5, the PLS models normally use less components than the PCR models. This effect is well known and happens because PCR uses the PCA scores obtained only from the independent variables **X** matrix**.** These values are used in regression with the values of the dependent variable **Y** matrix**.** On the other hand, PLS decomposes the **X** matrix while interchanging information with the **Y** matrix and conversely. These

scores values from **X** and **Y** matrices are used in the regression for obtaining the Y values with more accuracy. It can be observed that SEV values of PLS are slightly higher than the SEV values of PCR, but all SEP* values of PLS are slightly lower than the PCR SEP* values.

Fig. 2 and 3 show the experimental vs predicted K metal content plots of calibration and test set samples using the PCR and PLS models. In these figures the points belonging to calibration set are in the 1,627%-7,925% K range for the experimental values and 1,497%-7,6745% of K for predicted values. The predicted values range from   -   % for PCR and   -   % for PLS. However the test set K contents occupy only a small portion of this range, 5.50-6.73% K.
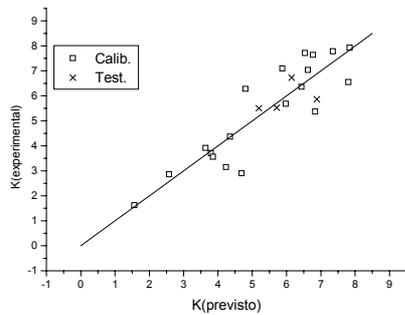
Fig. 2: Experimental vs predicted metal contents of K plot for calibration (□) and test (X) set samples with the PCR model.
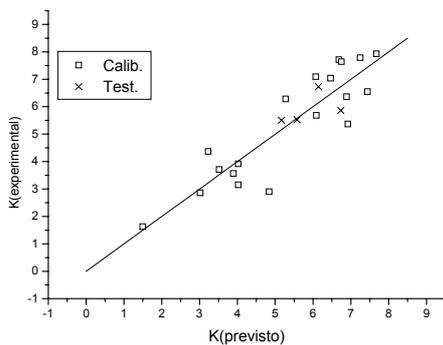
Fig. 3: : Experimental vs predicted metal contents of K plot for calibration (□) and test (X) set samples for the PLS model.

After observing figures for the other metal contents that are analogous to those of Figs. 2 and 3 the best calibration models were found to be K and Cu followed by Mn, Fe, Ca, and P in decreasing order of quality. The Zn model (Fig 4) was not included in tables 4 and 5 because it was built without samples 8 and 15 of the calibration set. These samples have the higher values of Zn content than the other samples, presenting high leverages and student residuals. As such they were considered outliers.
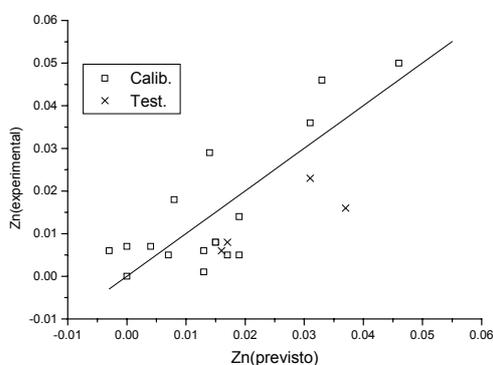

Fig. 4: Experimental vs predicted metal content of Zn plot for calibration (□) and test (X) set samples with PLS model.

In order to reduce the number of variables and possibly improve the models the spectra (Fig 1) and loadings and regression vector plots (Fig. 5) were examined. It was found that extensive ranges of the spectra do not have relevant absorptions and little useful information for calibration. Fig. 5 illustrates the important spectral regions for calibration and the ones with high absolute loadings and regression coefficients. Spectral regions were excluded from calibration if they contain **both** small absolute loading and regression coefficients.

Fig. 5 (a) shows the loading plot for two PCs for 400-4000 cm^{-1} spectral range. In Fig 5(b) one can observe the coefficients of vector regression plot for the first two PCs vs wave number in this spectral range. These plots have

appearances similar to those found for the other metals. The chosen ranges for the reduced variable PCR and PLS regression models of all metals were : 400-1227 cm^{-1}, 1364-1767 cm^{-1} and 3466-3888 cm^{-1}. 860 of the initial 1868

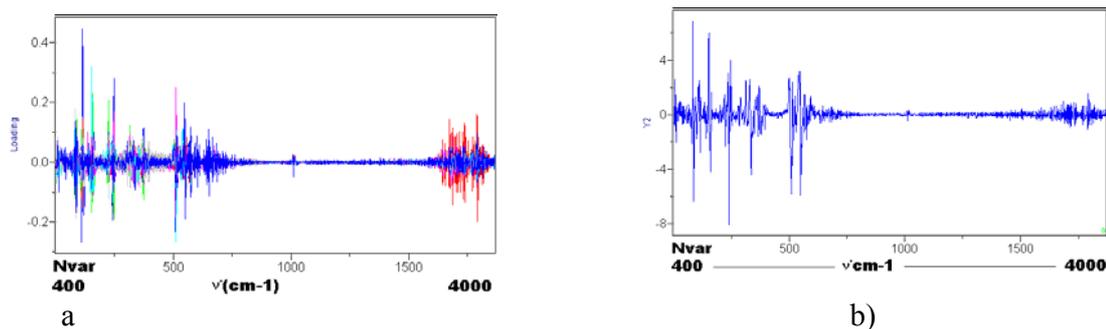variables were retained and the calculations were carried out in the same way as for all the variables.

a

b)

Fig. 5: (a) Loading plot and (b) Vector regression coefficient vs wave number for calibration set of K content PCR model.

The PCR and PLS models for the reduced variable set are almost the same as those for all 1868 variables and for this reason their results have not been reproduced here. The most striking differences are that the PCR and PLS models for Mn and Ca, respectively, have one less significant component with the reduced variable set than with the 1868 variable set. However these differences are not really significant since it is very difficult to choose the correct number of components to include in calibration

models. Different selection criteria can lead to different numbers of components to be used. As always one does not want to leave out important factors since useful information about calibration will be missing and the models will provide unreliable predictions. Including too many factors also leads to inaccurate results since random error contributions will be built into the model. The SEV, SEP and R^2 values are hardly changed on reducing the spectral variables. For example the SEV values of 0.140, 1.541,

0.711, 0.051, 0.012 and 0.002 for the P, K, Ca, Mn, Fe and Cu PLS models are very similar to the corresponding values in Table 5 for the complete variable set.

Calibration models were also constructed for the 10 most relevant

Conclusions

FTIR spectra in the 4000-400cm^{-1} region can used to provide approximate values of metal contents in ashes of palms samples from *oenocarpus-jessenia* complex using PCR and PLS models. K

absorptions at 447, 573, 619, 1026, 1028, 1119, 1418, 1464, 2341 e 3431 cm^{-1}. Also different pre-processing tecniques like mean-centering and MSC were applied to the spectra. Prediction results changed only very slightly.

and Cu gave the best models, followed by Fe, Mn, P, Ca and Zn in decreasing order of quality. Although the absorption bands in the FTIR spectra important for calibration do not correspond to metal atom vibrations

References

[1] Kalivas, J. H.; Lang, P.M. Mathematical Analysis of spectral Orthogonality, Marcel Dekker, New York, 1994.,
[2] Martens, H.; Naes, T. Multivariate Calibration; Jhon Wiley & Sons Ltd; Chichester, 1996.
[3] Beebe, K. ; Pell, P.; Seasholtz, M. B. Chemometrics: A Practical Guide; John Wiley; New York, 1998
[4] Watson, C. A. Near-Infrared Refletance Spectrophotometric Analysis of Agricultural Productos, Anal. Chem., 1977, 49, 835A.
[5] Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Pagès, J. Anal. Chim. Acta 384, 1999, 207.
[6] De Boever, J. I.; Cottyn, B. G.; Vanacker, J. M.; Boucqué, Ch. V. Animal Feed Science and Technology 51, 1995, 243.
[7] Vonach, R.; Buschmaann, J.; Falkowski, R.; Schinder, R.; Lendl, B.; Kellner, R. Applied Spectrocopy 52, 1998, 820.

[8] Instituto Adolfo Lutz. Normas Analíticas do Instituto Adolfo Lutz. Métodos físicos e químicos para análise de alimentos; São Paulo, Vol 1, 1985.

[9] A.O C.S. Official Methods of Analysis. American Oil Chemists Society, Volume 1 e 2, 1989.

[10] Martens, H.; Naes, T. Multivariate Calibration; Jhon Wiley & Sons Ltd; Chichester, 1996.

[11] Geladi, P.; Kowalski, B. R. Anal. Chim. Acta, 185 (1986), 1-17

[12] Beebe, K. ; Pell, P.; Seasholtz, M. B. Chemometrics: A Practical Guide; John Wiley; New York, 1998.

[13] Infometrix, Inc. Pirouette Version 2.02; Woodinville, Washington, 1996.

[14] CAMO-Computer Aided Modelling A/S. The Unscrambler 6.0; Trondheim, Norway, 1986-1996.

[15] Xie, Y. L.; Kalivas, J. H. Anal. Chim Acta 348, 1997, 19.

[16] Xie, Y. L.; Kalivas, J. H. Anal. Chim Acta 348, 1997, 29.

[17] Barros, N. B.; Spacino, S. I.; Bruns, R. E. Como fazer experimentos: pesquisa e desenvolvimento na ciência e na industria; Editora da Unicamp, Campinas, SP, 2001.

[18] Vu Nguyen-Cong; Bernd M. R. *J. Chem. Inf. Comp. Sci.* 1996, 36, 114-117