Acta Scientiarum
http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v41i2.37382

Multi-product multivariate calibration: determination of quality parameters in soybean industrialized juices

Dayane Aparecida dos Santos, Karen Priscila de Lima, Marcilene Ferrari Barriquello Consolin, Nelson Consolin Filho, Paulo Henrique Março and Patrícia Valderrama*

Universidade Tecnológica Federal do Paraná, Via Rosalina Maria dos Santos, 1233, Cx. Postal 271, 87301-899, Campo Mourão, Paraná, Brazil. *Author for correspondence. E-mail: patriciav@utfpr.edu.br

ABSTRACT. Total acidity and vitamin C were determined by using ultraviolet spectroscopy and multi-product multivariate calibration alternately to the reference methods, the potentiometry and Tillman's, respectively. In the developed multi-products models, different products were included (industrialized juices based on soya of different flavors and light). The linear partial least squares (PLS) method was used in the model construction and the outlier samples were evaluated. The accuracy at the 99% level, represented by the root mean square error of calibration (RMSEC) and prediction (RMSEP), was confirmed through the confidence ellipse, whereas the residuals presented random behavior, which indicates that the data fit a linear model. Sensitivity and analytical sensitivity presented adequate results in the determination of vitamin C and acidity, considering the concentration range used 0.83-16.83 mg 100 mL^{-1} for vitamin C and 0.17-0.34 g 100 mL^{-1} for total acidity. The inverse of the analytical sensitivity shows that it is possible to distinguish samples with difference in vitamin C concentration of the order of 0.73 mg 100 mL^{-1}, and samples with difference in total acidity of the order of 6.1 x 10^{-3} g 100 mL^{-1}.The multi-product PLS model present limits of detection and quantification for vitamin C of 2.43 and 7.36 mg 100 mL^{-1}, respectively. For total acidity, the limits of detection and quantification achieved were 0.02 and 0.06 mg 100 mL^{-1}, respectively. The values for residual prediction deviation (RPD) presented results within the range of values, which classify the models as satisfactory. In addition, the multi-product calibration is fast, because it does not require reagents/solvents and does not generate toxic waste, being an alternative to the conventional methods and being in agreement with the requirements of green chemistry.

Keywords: acidity; vitamin C; ultraviolet spectroscopy; multi-product calibration; figures of merit; soybean juices.

Received on May 29, 2017.
Accepted on June 5, 2018.

Introduction

Partial Least Squares (PLS) is a linear multivariate regression method developed in the 1960s by H. Wold for the economics area. It was only in the early 1980's that his son, S. Wold, together with H. Martens, started applications in the chemistry field (Sanchez, 2017). Currently, the multivariate calibration from the PLS method is consolidated for first-order data, i.e. when a vector of instrumental responses is available for each sample.

The PLS regression is considered to have the least mathematical disadvantage compared to other multivariate regression methods such as Classical Least Squares (CLS), Multiple Linear Regression (MLR) or the Principal Components Regression (PCR). For instance: 1) For CLS application it is necessary to know the concentration of all species that contribute to the instrumental signal, which is most of the time impossible when working with complex matrices like food. 2) MLR contours the problem described by the CLS, however, for this regression method, it is necessary to have the number of samples larger than the number of variables. This is something difficult to access when working with spectroscopy, where many variables are considered in the development of the multivariate model. 3) PCR is a regression method that contours the problem presented by CLS and MLR. However, with PCR, no information about the reference method is employed in the dimensionality reduction of the instrumental matrix (Ferreira, Antunes, Melgo, & Volpe, 1999; Ferreira, 2015).

Multi-product multivariate calibration had its first scientific report in 1992 (Naes & Isaksson, 1992), and the second one in 1994 (Wang, Isaksson, & Kowalski, 1994). These works reported the possibility of

developing multivariate calibration and included, in the same model different types of products. These studies evaluated a set of different products that presented homogeneous responses. Furthermore, the main goal of these researches was to evaluate new algorithms to develop multivariate regression. A data set that did not present homogeneous responses was evaluated in a research performed in the year 2000 (Berzaghi, Shenk, & Westerhaus, 2000). However, in this latter work, the main objective was to evaluate the performance of the algorithm named Local.

Micklander, Kjeldahl, Egebo, and NØrgaard (2006) introduced the term multi-product calibration to the scientific world in the year 2006. The authors investigated the use of the PLS regression method, nonlinear regression using neural networks, and three variations of the Local algorithm in the development of multi-product multivariate calibration models. The PLS method presented larger prediction to errors, which could maybe be justified by an inconsistent sampling representativeness, or by the absence of outliers when evaluating the developed model.

The successful use of the PLS regression method in the development of multi-product multivariate calibration models is recent (Rambo, Amorim, & Ferreira, 2013; Santos, Março, & Valderrama, 2013; Santos, Lima, Março, & Valderrama, 2015; 2016). The use of the PLS method by the industrial sector has been growing and gaining more and more space. In this sense, the possibility of using this method of multivariate regression applied to different products becomes an interesting alternative in terms of time and practicality.

Multivariate models maintenance can be laborious. Thus, the multi-product multivariate model has the advantages of saving time, robustness and practicality, considering the terms of keeping its maintenance. Moreover, another disadvantage is a large number of steps that the quality control analyst needs to perform with a single model. For example, in each analysis performed in a laboratory routine, a specific model is used for a small population of samples (a single product), and each sample should then be carefully identified, as well as the correct and specific model, in order for that product be properly chosen (Santos et al., 2013).

Acidity and vitamin C are quality parameters, responsible for aroma, flavor, sensory and nutritional characteristics, as well as for the state of conservation of food (Venâncio & Martins, 2012). These parameters are used by the juice industry in the quality control of the final product.

Industrialized juices have been gaining consumer preference because of their practicality. In this sense, fruit nectar – which is defined as an unfermented drink ready for consumption, that is obtained from the edible part of the fruit diluted in potable water, and that may or may not be added with sugars, acids (Santos et al., 2015) or soy. The soy juices preserve the desirable sensory characteristics of fruits, along with the functional properties of soybeans, such as the presence of bioactive compounds such as isoflavones. The isoflavones have beneficial effects to human health, such as: estrogenic, antiestrogenic activity (especially on the symptoms of the climacteric syndrome and osteoporosis), hypocholesteremic and anticarcinogenic activities (Lui, Aguiar, Alencar, Scamparini, & Park, 2003; Torrezan et al., 2004; Abreu, Pinheiro, Maia, Carvalho, & Sousa, 2007).

Although these quality parameters have already been evaluated from multivariate multi-product calibration for fruit nectar, soybean industrialized juices present very different physicochemical aspects (opacity for example), which justifies an investigation into the determination of these parameters for this type of food sample. Therefore, the objective of this study was to propose the development of multivariate calibration models based on ultraviolet (UV) spectroscopy for the determination of the total acidity and vitamin C in soybean industrialized juices of different flavors, also including the light type.

Material and methods

Samples and reagents

One hundred and twenty-six samples were acquired in the Campo Mourão – PR marketplaces: pineapple (21 samples), grape (18 samples), orange (9 samples), peach and apples (15 samples for each flavor), strawberry, passion fruit, light apple, light grape and light peach (6 samples for each flavor), tangerine, pomegranate, mango, papaya, lemon and light orange (3 samples for each flavor).

Sodium hydroxide (Synth) and hydrochloric acid (Synth) were used to determine the total acidity. For vitamin C determination, we used ascorbic acid (Impex), 2,6-dichlorophenol indophenol, indigo carmine and 1% phenolphthalein (Sigma-Aldrich), metaphosphoric acid, glacial acetic acid and hydrochloric acid (Vetec), sodium bicarbonate, sodium hydroxide and potassium b (Alphatec).

Methods

The total acidity (mg 100 mL^{-1}) was determined, in triplicate, according to the Federation International des Producteurs de Jus de Fruit (2005) and to the methodology described by Santos et al. (2015).

The vitamin C (mg 100 mL^{-1}) was determined, in triplicate, according to the Association of Official Analytical Chemists (AOCS) and to the Tilmman's Method (Latimer, 1990). Santos et al. (2016) are the responsible for describing this methodology in detail. The samples were previously diluted (300 µL sample: 10 mL distilled water) and UV spectra (200 – 350 nm, steps of 1nm, Ocean Optics, model USB-650-UV-VIS) were obtained by using a 1mm quartz cuvette.

The multi-product multivariate calibration was performed by using the Matlab R2007b and PLS-Toolbox 5.2 (Eigenvector Research Inc.). The regression method used in the development of multi-product multivariate calibration was PLS. In the PLS, the X matrix contains the instrumental responses (UV spectra in this case) and the y vector contains the results for acidity and vitamin C (which is obtained by the reference methods). These ones are decomposed into two matrix products, a score matrix, and a loadings matrix. A least squares regression was obtained from the scores and loadings from X matrix against the scores from y vector. More detailed information on the PLS regression method, including a mathematical step-by-step, can be obtained in Ferreira (2015).

The outliers were evaluated according to ASTM E-1655-05 (American Society for Testing and Materials [ASTM], 2005) during the model development. Outliers were identified based on leverage, unmodeled residuals in spectra and unmodeled residuals on the dependent variable (residual in y).

Multi-product models were validated by calculating the parameters of merit: accuracy, Residual Prediction Deviation (RPD), sensitivity, inverse of analytical sensitivity (analytical $sensitivity^{-1}$), limits of detection and quantification, according to the equations shown in Table 1 (Valderrama, Braga, & Poppi, 2009; Santos et al., 2016).

Results and discussion

Multi-product models were developed based on PLS regression method. For this, the UV spectra of soybean juice samples were organized into a matrix. Figure 1 shows the spectra in the UV region for all analyzed samples. It was verified the need to apply the first derivative preprocessing to the spectra. This occurred probably due to the opaque color of the soybean juice samples, even after its dilution.

Table 1. Equations for the parameters of merit.

Parameters of merit	Equation
Accuracy	$RMSEP = \sqrt{\frac{\sum_{i=1}^{nv}(y_i - \hat{y}_i)^2}{nv}}$
	$RMSEC = \sqrt{\frac{\sum_{i=1}^{nc}(y_i - \hat{y}_i)^2}{nc - nVL + 1}}$
RPD	$RPD_{cal} = \frac{DP_{cal}}{RMSECV}$
	$RPD_{val} = \frac{DP_{val}}{RMSEP}$
Sensitivity	$Sensitivity = \frac{1}{\|b\|}$
Analytical Sensitivity	$AnalyticalSensitivity = \frac{Sensitivity}{\|\delta x\|}$
Analytical $Sensitivity^{-1}$	$AnalyticalSensitivity^{-1} = \frac{1}{AnalyticalSensitivity}$
Limit of detection	$Limit of detection = 3.3\delta x \frac{1}{Sensitivity}$
Limit of quantification	$Limit of quantification = 10\delta x \frac{1}{Sensitivity}$

nv is the number of samples in the validation set, yi is the reference value for the samples and $\hat{y}$ is the value predicted by the model for the sample i, nc is the number of samples in the calibration set, nVL is the number of latene variables, DPca l is the standard deviation of the reference values in the calibration set, DPval is the standard deviation of the reference values in the validation set, RMSECV is the Root Mean Square Error for Cross Validation, RMSEC is the Root Mean Square Error for Calibration, RMSEP is the Root Mean Square Error for Prediction, **b** is the regression coefficient vector obtained from the model, ⊠x is the instrumental noise estimation. On the RMSEC equation, the '+1' is added when the pre-processing is the mean center.

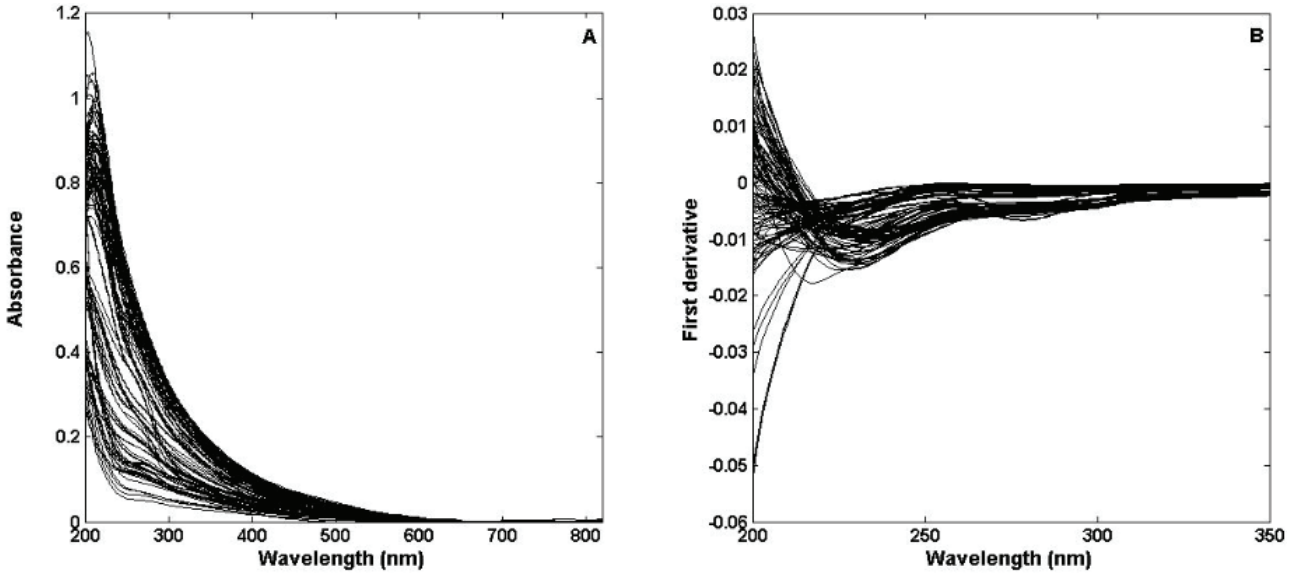

Figure 1. UV spectra of soybean juice samples. (A) Raw spectra. (B) Spectra after the first derivative.

The calibration and validation data sets were composed by 94 and 32 samples, respectively, selected by the *kenston* algorithm (Kennard & Stone, 1969). The next step in the model development was the outlier detection, in order to improve the model's quality. The outliers were identified based on data with extreme leverage, unmodeled residuals in spectral data and unmodeled residuals in the response obtained by the reference method. This procedure resulted in 80 and 76 calibration samples and, in 21 and 23 validation samples for models in the determination of total acidity and vitamin C, respectively. A detailed description of the samples identified as outliers, as well as the acidity and vitamin C values obtained through the reference methods can be seen in Appendix 1 and 2.

Models were developed with mean center pre-processing and 10 latent variables (LVs), which were determined through the Root Mean Square Error for Cross-Validation in the contiguous block of nine samples. The accuracy of the models was evaluated by the Root Mean Square Error of calibration (RMSEC) and the Prediction (RMSEP), as shown in Table 2.

RMSEC and RMSEP values are close and suggest that the number of LVs was properly chosen, i.e. it did not present overfit or even lack of fit. RMSEC values decrease with the increase in the LVs number. This occurs due to errors in spectra and concentrations included in the model adjust. In contrast, RMSECV and RMSEP occasionally increase when more LVs are included in the model. However, new samples that were not present in the calibration step will have a different behavior of random errors. Therefore, the calibration model does not 'fit' these errors to the same degree as the errors in samples employed in the calibration. In practice, obtaining the same values for these parameters is not easy and it is better than the RMSEC presents values slightly higher than the RMSEP, which suggests that this model is suitable for the random errors present in the samples that were not part of step calibration (Santos et al., 2013).

The RMSEC and RMSEP are global parameters and they incorporate random and bias errors. Therefore, it is interesting to evaluate these results along with other accuracy indicators, such as the fit of the reference values against the predicted ones (correlation coefficient – Table 1). Also, the elliptical joint confidence regions (Valderrama et al., 2009) shown in Figure 2. It is observed that the ellipse contains the ideal point (1.0) for slope and intercept, respectively, which shows that the reference values and the PLS model are not significantly differenced at the 99% confidence level. It is possible to conclude also that the values for these parameters determined by titration (potentiometric or oxidation-reduction), and the values of total acidity and vitamin C determined by the multi-product PLS model do not present significant difference with 99% confidence.

Correlation coefficient to the fit of the multiproduct model, presented by plotting the reference values against the estimated values, was 0.7188 for vitamin C and 0.7435 for total acidity. These values were considered satisfactory since previous research reported coefficient values around 0.7, when the reference method was the titration method (Valderrama, Braga, & Poppi, 2007a; 2007b; Ferreira, Pallone, & Poppi, 2013; Santos et al., 2015; 2016).

Table 2. Multi-product model's parameters of merit.

Parameters of merit		Vitamin C	Total Acidity
Accuracy	RMSEC[a]	2.5332	0.0218
	RMSEP[a]	1.6973	0.0199
Correlation coefficient		0.7188	0.7435
RPD_{cal}		1.3	1.5
RPD_{val}		2.6	1.7
Analytical sensitivity^{-1} [a]		0.7356	0.0061
Limit of detection[a]		2.4275	0.0200
Limit of quantification[a]		7.3561	0.0606

[a](mg 100 mL^{-1}).

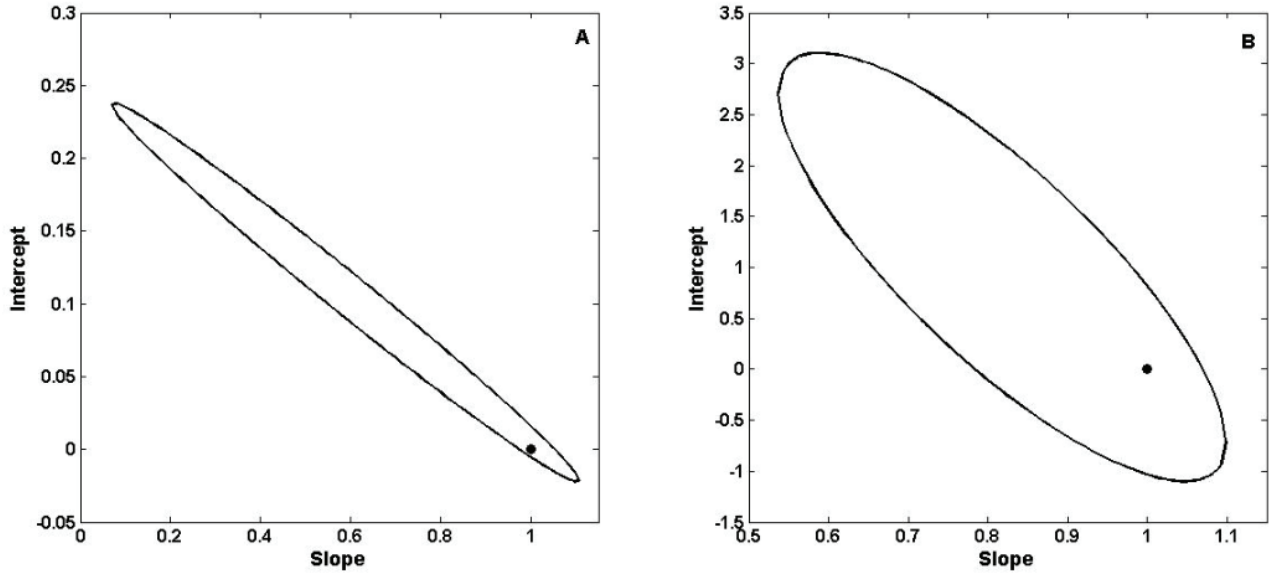

Figure 2. Elliptical joint confidence regions at 99% for the slope and intercept of the regression of predicted concentrations versus reference experimental values using ordinary least squares. (A) Total acidity. (B) Vitamin C. (•) Point where the intercept is zero and the slope is one.

The results presented in Table 2 and Figure 2 show that the multi-product PLS model results in more 'dispersed' results for vitamin C, which may be justified by the fact that the vitamin C is oxidized quickly, when in the presence of oxygen. In addition, the titration method shows a color turning point that may be difficult to identify, especially in colored and cloudy samples, such as soybean juices.

Figure 3 shows the residuals plot of the calibration and validation samples. The residuals distribution seems to present a random behavior, which reinforces that the data fit a linear model.

RPD value of the calibration model for vitamin C showed close value to what is considered satisfactory and may be considered adequate in relation to the RPD value for the validation of this parameter. In the model to determine acidity, the RPD can be considered satisfactory for calibration and validation. According to the literature (Botelho, Mendes, & Sena, 2013), multivariate models are considered good models when they show values for RPD above 2.4. Models with RPD values between 2.4 and 1.5 are also satisfactory.

The sensitivity and analytical sensitivity showed satisfactory results, taking into account the analytical range of the models, 0.83-16.83 mg 100 mL^{-1}for vitamin C and 0.17-0.34 mg 100 mL^{-1} for total acidity. The analytical sensitivity^{-1} allows one to establish a minimum concentration difference that is discernible by the multi-product model. Thus, it is possible to distinguish samples with vitamin C concentration in the order of 0.73 mg 100 mL^{-1} and samples with total acidity in the order of 6.1 x 10^{-3} mg 100 mL^{-1}.

Detection limit shows the lowest concentration of vitamin C and total acidity that can be detected but not necessarily accurately quantified. On the other hand, the limit of quantification shows the lowest concentration of vitamin C and total acidity that can be quantified with accuracy. In the multi-product model for vitamin C determination, the results indicate that the proposed multi-product model cannot accurately detect and quantify samples with vitamin C concentration below 2.43 and 7.36 mg 100 mL^{-1}, respectively.

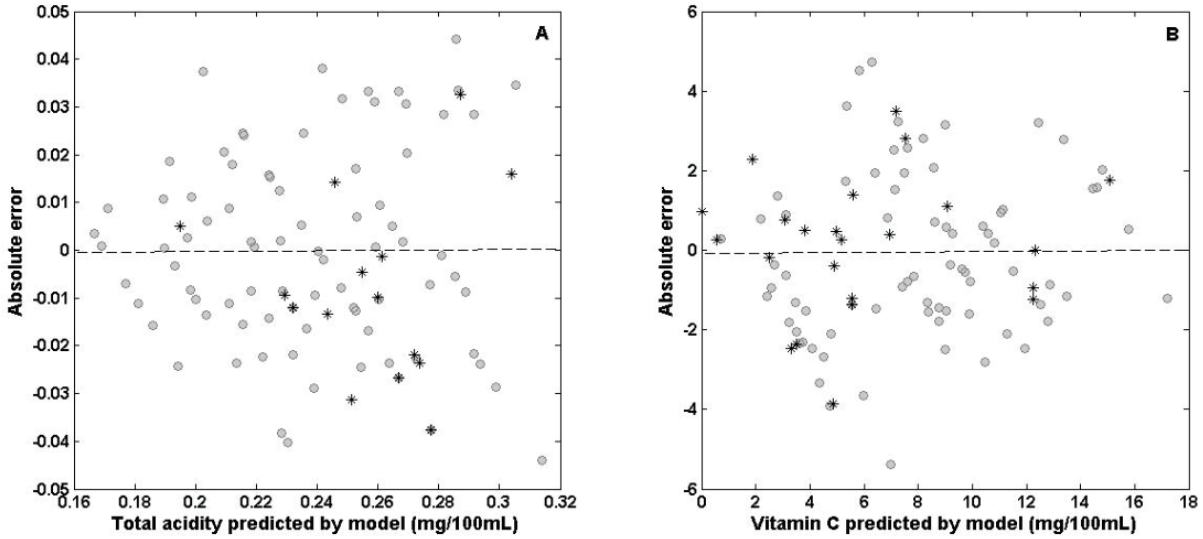

Figure 3. Residuals for the multi-product models. (A) Total acidity. (B) Vitamin C. (•) Calibration samples. (*) Validation samples.

Conclusion

The results show that there is a favorable possibility of using a PLS model in the evaluation of the total acidity and vitamin C in different products (soybean juices with different flavors, and the light type of juice) simultaneously. Therefore, UV spectroscopy coupled with the PLS regression method allows the construction of multi-product calibration models. In addition, the multi-product models allow rapid quantification of the total acidity and vitamin C content and does not require the use of reagents/solvents. Thus, it does not generate toxic residues, which is an alternative to the conventional methods based on titration and being in accordance with the requirements of the green chemistry. However, we point out that the methodology could be improved (perhaps evaluating other spectral pre-processing types or different sample of dilutions) in order to obtain lower prediction errors.

AcknowledgementS

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes).

References

Abreu, C. R. A., Pinheiro, A. M., Maia, G. A., Carvalho, J. M., & Sousa, P. H. M. (2007). Avaliação química e físico-química de bebidas de soja com frutas tropicais. *Brazilian Journal of Food and Nutrition, 18*(3), 291-296.

American Society for Testing and Materials [ASTM]. (2005). *ASTM E1655-05. Standards practices for infrared, multivariate, quantitative analysis*. Harrisburg, PA: ASTM.

Berzaghi, P., Shenk, J. S., & Westerhaus, M. O. (2000). LOCAL prediction with near infrared multi-product databases. *Journal of Near Infrared Spectroscopy, 8*(1), 1-9. doi: 10.1255/jnirs.258

Botelho, B. G., Mendes, B. A. P., & Sena, M. M. (2013). Implementação de um método robusto para o controle fiscal de umidade em queijo minas artesanal. Abordagem metrológica multivariada. *Química Nova, 36*(9), 1416-1422. doi: 10.1590/S0100-40422013000900023

Federation International des Producteurs de Jus de Fruits. (2005). *Determination of titrable acidity.* Paris, FR: IFUMA03.

Ferreira, D. S., Pallone, J. A. L., & Poppi, R. J. (2013). Fourier transform near-infrared spectroscopy (FT-NIRS) application to estimate Brazilian soybean [*Glycine max* (L.) Merril] composition. *Food Research International, 51*(1), 53-58. doi: 10.1016/j.foodres.2012.09.015

Ferreira, M. M. C. (2015). *Quimiometria – conceitos, métodos e aplicações*. Campinas, SP: Unicamp.

Ferreira, M. M. C., Antunes, A. M., Melgo, M. S., & Volpe, P. L. O. (1999). Quimiometria I: calibração multivariada, um tutorial. *Química Nova, 22*(5), 724-731. doi: 10.1590/S0100-40421999000500016

Kennard, R. W., & Stone, L. A. (1969). Computer aided desing of experiments. *Technometrics, 11*(1), 137-148. doi: 10.2307/1266770

Latimer, G. W. (1990). *Official methods of analysis of the association of official analytical chemistry* (15th ed.). Arlington, TX: AOAC.

Lui, M. C. Y., Aguiar, C. L., Alencar, S. M., Scamparini, A. R. P., & Park, Y. K. (2003). Isoflavonas em isolados e concentrados protéicos de soja. *Ciência e Tecnologia de Alimentos, 23*(supl.), 206-212. doi: 10.1590/S0101-20612003000400038

Micklander, E., Kjeldahl, K., Egebo, M., & NØrgaard, L. (2006). Multi-product calibration models of near infrared spectra of foods. *Journal of Near Infrared Spectroscopy, 14*(6), 395-402. doi: 10.1255/jnirs.659

Naes, T., & Isaksson, T. (1992). Locally weighted regression in diffuse near-infrared transmittance spectroscopy. *Applied Spectroscopy, 46*(1), 34-43. doi: 10.1366/0003702924444344

Rambo, M. K. D., Amorim, E. P., & Ferreira, M. M. C. (2013). Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. *Analytica Chimica Acta, 775*, 41-49. doi: 10.1016/j.aca.2013.03.015

Sanchez, G. (2017). *The saga of PLS*. Retrieved from http://sagaofpls.github.io

Santos, D. A., Lima, K. P., Março, P. H., & Valderrama, P. (2015). UV spectroscopy and multi-product multivariate calibration in the determination of the total acidity in industrialized juices. *Revista Brasileira de Pesquisa em Alimentos, 6*(1), 1-8. doi: 10.14685/rebrapa.v6i1.177

Santos, D. A., Lima, K. P., Março, P. H., & Valderrama, P. (2016). Vitamin C determination by ultraviolet spectroscopy and multiproduct calibration. *Journal of the Brazilian Chemical Society, 27*(10), 1912-1917. doi: 10.5935/0103-5053.20160071

Santos, D. A., Março, P. H., & Valderrama, P. (2013). Multi-product calibration: preliminar studies to determine quality parameters in industrialized juices based on ultravioleta spectroscopy. *Brazilian Journal of Analytical Chemistry, 12*(3), 495-498.

Torrezan, R., Ceccato, C. M., Barreto, A. C. S., Silva, V. S., Caratin, C., Pereira, C. G., ... Cardello, H. M. A. B. (2004). Avaliação do perfil sensorial de alimento com soja sabor laranja. *Boletim do CEPPA, 22*(2), 199-216. doi: 10.5380/cep.v22i2.1190

Valderrama, P., Braga, J. W. B., & Poppi, R. J. (2007a). Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy. *Journal of the Brazilian Chemical Society, 18*(2), 259-266. doi: 10.1590/S0103-50532007000200003

Valderrama, P., Braga, J. W. B., & Poppi, R. J. (2007b). Variable selection, outlier detection, and figures of merit estimation in a Partial Least-Squares regression multivariate calibration model. A case Study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry, 55*(21), 8331-8338. doi: 10.1021/jf071538s

Valderrama, P., Braga, J. W. B., & Poppi, R. J. (2009). Estado da Arte de figuras de mérito em calibração multivariada. *Química Nova, 32*(5), 1278-1287. doi: 10.1590/S0100-40422009000500034

Venâncio, A. A., & Martins, O. A. (2012). Chemical analysis of different brands of nectars and juice orange sold in the city of Cerqueira César – São Paulo. *Revista Eletrônica de Educação e Ciência, 2*(3), 45-50.

Wang, Z., Isaksson, T., & Kowalski, B. R. (1994). New approach for distance measurement in locally weighted regression. *Analytical Chemistry, 66*(2), 249-260. doi: 10.1021/ac00074a012

Appendix 1. Outliers identification and total acid values obtained through reference method.

Nº	Flavor	Total acidity (mg 100 mL^{-1})-Potentiometry	Samples in calibration set	Samples in validation set	Outliers based on
1	light Peach	0.28	X		Residual in y
2	light Peach	0.29	X		
3	light Peach	0.29	X		
4	Peach	0.22	X		
5	Peach	0.22	X		
6	Peach	0.22		X	
7	Peach	0.18	X		
8	Peach	0.19	X		Leverage
9	Peach	0.19	X		
10	Peach	0.28	X		
11	Peach	0.28	X		
12	Peach	0.28	X		
13	Peach	0.25	X		
14	Peach	0.25		X	
15	Peach	0.25		X	
16	Pineapple	0.22	X		
17	Pineapple	0.22		X	
18	Pineapple	0.23		X	
19	Pineapple	0.21	X		Residual in y
20	Pineapple	0.21	X		
21	Pineapple	0.21	X		
22	Pineapple	0.26		X	
23	Pineapple	0.26	X		
24	Pineapple	0.26	X		
25	Pineapple	0.19	X		
26	Pineapple	0.19	X		
27	Pineapple	0.19		X	Residual in y
28	Pineapple	0.27	X		
29	Pineapple	0.27	X		
30	Pineapple	0.27	X		
31	Pineapple	0.14	X		
32	Pineapple	0.13		X	Residual in y
33	Pineapple	0.14		X	Residual in y
34	Pineapple	0.32		X	
35	Pineapple	0.32	X		
36	Pineapple	0.32		X	Residual in y
37	Apple	0.24	X		
38	Apple	0.24	X		
39	Apple	0.24	X		
40	Apple	0.20	X		
41	Apple	0.20		X	
42	Apple	0.20	X		
43	Apple	0.15		X	Residual in y
44	Apple	0.15		X	Residual in y
45	Apple	0.15	X		Residual in y
46	Apple	0.32	X		
47	Apple	0.32	X		
48	Apple	0.32		X	
49	Apple	0.25	X		
50	Apple	0.25		X	
51	Apple	0.24	X		
52	light Apple	0.24	X		
53	light Apple	0.24	X		
54	light Apple	0.24	X		
55	Apple	0.24		X	Residual in y
56	Apple	0.24		X	
57	Apple	0.24		X	
58	Lemon	0.21	X		
59	Lemon	0.22	X		Residual in y
60	Lemon	0.22	X		Residual in y
61	Pomegranate	0.27	X		
62	Pomegranate	0.27	X		
63	Pomegranate	0.27	X		

64	Strawberry	0.27	X		Residual in y
65	Strawberry	0.27	X		
66	Strawberry	0.27	X		
67	Strawberry	0.21	X		
68	Strawberry	0.21	X		
69	Strawberry	0.21	X		
70	Grape	0.30	X		
71	Grape	0.30	X		
72	Grape	0.30		X	Residual in y
73	Grape	0.25		X	
74	Grape	0.26		X	
75	Grape	0.26	X		
76	light Grape	0.16	X		
77	light Grape	0.16		X	Residual in y
78	light Grape	0.17		X	Residual in y
79	Grape	0.24	X		
80	Grape	0.24	X		
81	Grape	0.24	X		
82	Grape	0.19	X		Residual in y
83	Grape	0.19	X		
84	Grape	0.19	X		
85	Grape	0.36	X		
86	Grape	0.36		X	Residual in y
87	Grape	0.35	X		Residual in y
88	Grape	0.17	X		
89	Grape	0.17	X		
90	Grape	0.17	X		
91	light Grape	0.23	X		
92	light Grape	0.23	X		
93	light Grape	0.23	X		
94	light Orange	0.28	X		
95	light Orange	0.29		X	Residual in y
96	light Orange	0.29	X		
97	Orange	0.19	X		
98	Orange	0.19		X	Leverage
99	Orange	0.19		X	Leverage
100	Orange	0.34	X		
101	Orange	0.33	X		
102	Orange	0.34	X		
103	Orange	0.19		X	Residual in y
104	Orange	0.19	X		
105	Orange	0.19	X		
106	Passion fruit	0.17	X		
107	Passion fruit	0.17	X		
108	Passion fruit	0.17	X		
109	Passion fruit	0.23	X		
110	Passion fruit	0.23	X		
111	Passion fruit	0.24	X		
112	Peach	0.20	X		
113	Peach	0.20	X		
114	Peach	0.20	X		
115	light Peach	0.24	X		
116	light Peach	0.24	X		
117	light Peach	0.24	X		Residual in y
118	Papaya	0.28	X		
119	Papaya	0.28	X		
120	Papaya	0.28		X	Residual in y
121	Mango	0.24	X		
122	Mango	0.24	X		Residual in y
123	Mango	0.24	X		
124	Tangerine	0.22	X		
125	Tangerine	0.22		X	
126	Tangerine	0.22		X	

Appendix 2. Outliers identification and vitamin C values obtained through reference method.

Nº	Flavor	Vitamin C (mg 100 mL^{-1}) Titration	Samples in calibration set	Samples in validation set	Outliers based on
1	light Peach	9,60	X		
2	light Peach	9,44	X		
3	light Peach	9,60	X		
4	Peach	12,32	X		
5	Peach	12,16	X		
6	Peach	12,32		X	
7	Peach	26,24	X		Leverage/ Residual in y
8	Peach	26,56	X		Leverage
9	Peach	26,72	X		
10	Peach	1,60	X		
11	Peach	1,60	X		Residual in spectrum
12	Peach	1,60	X		
13	Peach	4,16	X		
14	Peach	3,84		X	
15	Peach	4,32		X	
16	Pineapple	4,96	X		
17	Pineapple	5,44		X	
18	Pineapple	5,44		X	
19	Pineapple	1,76	X		
20	Pineapple	1,60	X		Residual in y
21	Pineapple	1,60	X		
22	Pineapple	11,33		X	
23	Pineapple	11,00	X		
24	Pineapple	11,00	X		
25	Pineapple	15,67	X		
26	Pineapple	16,00	X		
27	Pineapple	15,50		X	Residual in y
28	Pineapple	1,83	X		
29	Pineapple	2,33	X		
30	Pineapple	2,17	X		
31	Pineapple	4,00	X		Residual in y
32	Pineapple	4,50		X	
33	Pineapple	4,17		X	
34	Pineapple	10,17		X	
35	Pineapple	10,67	X		
36	Pineapple	10,67		X	
37	Apple	6,50	X		
38	Apple	7,67	X		
39	Apple	7,67	X		
40	Apple	10,50	X		
41	Apple	11,00		X	
42	Apple	11,00	X		Residual in y
43	Apple	0,83		X	
44	Apple	1,00		X	
45	Apple	1,00	X		
46	Apple	9,50	X		
47	Apple	9,33	X		
48	Apple	9,33		X	Residual in y
49	Apple	2,33	X		
50	Apple	2,33		X	
51	Apple	2,50	X		
52	light Apple	1,00	X		
53	light Apple	0,83	X		
54	light Apple	1,00	X		
55	Apple	0,83		X	Residual in y
56	Apple	1,00		X	Residual in y
57	Apple	1,00		X	Residual in y
58	Lemon	7,50	X		
59	Lemon	7,17	X		Residual in y
60	Lemon	7,67	X		
61	Pomegranate	2,33	X		
62	Pomegranate	3,00	X		
63	Pomegranate	2,67	X		

64	Strawberry	9,33	X			
65	Strawberry	9,67	X			
66	Strawberry	9,17	X			
67	Strawberry	16,83	X			Residual in y
68	Strawberry	17,17	X			Residual in spectrum/ Residual in y
69	Strawberry	16,83	X			
70	Grape	10,33	X			
71	Grape	10,17	X			
72	Grape	10,33		X		
73	Grape	0,83		X		
74	Grape	1,00		X		
75	Grape	0,83	X			
76	light Grape	12,16	X			
77	light Grape	11,84		X		Residual in y
78	light Grape	12,16		X		Residual in y
79	Grape	8,33	X			
80	Grape	8,83	X			
81	Grape	8,67	X			
82	Grape	9,50	X			
83	Grape	9,00	X			
84	Grape	9,17	X			
85	Grape	7,04	X			
86	Grape	7,36		X		
87	Grape	7,04	X			
88	Grape	8,32	X			Residual in y
89	Grape	9,12	X			
90	Grape	9,12	X			
91	light Grape	1,44	X			
92	light Grape	1,44	X			
93	light Grape	1,28	X			
94	light Orange	1,44	X			
95	light Orange	1,12		X		
96	light Orange	1,28	X			
97	Orange	28,50	X			Residual in y
98	Orange	28,67		X		Leverage/ Residual in y
99	Orange	28,50		X		Leverage/ Residual in y
100	Orange	6,83	X			
101	Orange	6,50	X			Residual in y
102	Orange	6,83	X			Residual in y
103	Orange	16,83		X		
104	Orange	16,33	X			
105	Orange	16,83	X			
106	Passion fruit	11,00	X			
107	Passion fruit	10,83	X			Residual in y
108	Passion fruit	11,00	X			
109	Passion fruit	16,17	X			
110	Passion fruit	16,17	X			
111	Passion fruit	16,00	X			
112	Peach	12,00	X			Residual in spectrum
113	Peach	12,00	X			
114	Peach	12,17	X			Residual in y
115	light Peach	44,33	X			Residual in y
116	light Peach	44,50	X			
117	light Peach	44,33	X			
118	Papaya	7,00	X			
119	Papaya	7,33	X			
120	Papaya	7,00		X		
121	Mango	11,00	X			
122	Mango	11,00	X			
123	Mango	11,17	X			
124	Tangerine	4,17	X			Residual in y
125	Tangerine	4,33		X		
126	Tangerine	4,17		X		