Use of near infrared spectroscopy for rapid estimation of sugar cane juice quality components

Paulo A. Costa Filho and Ronei J. Poppi
Universidade Estadual de Campinas, Instituto de Química, C.P. 6154, CEP 13083-970, Campinas-SP, Brazil.

Introduction

In the sugar cane industry there are a set of important quality parameters that are monitored to evaluate the raw material quality and the production process. Among these parameters it is fundamental to monitor the Brix and Pol.

Brix[1] is an important parameter in sugar cane quality control, because it furnaces the amount of soluble solids in 100 g of solution. The major solids are sugars (glucose, fructose and sucrose) and mineral salts. Polarimetry[2] (pol) is employed to determine the sucrose content of juice samples in the cane sugar industry. It is a measurement of optic rotation of molecules contained in the solution. The principle method for estimating the quality of sugar cane juice is based on the polarimetric reading difference before and after the sample hydrolysis.

Traditionally these determinations are performed using a refractometer and a saccharimeter, respectively, which are adequate under ideal conditions, but they can have serious limitations under circumstances in which optically-active substances other than sucrose are present. Moreover, these methods are time consuming and not suitable for on-line process monitoring.

The use of near infrared (NIR) spectroscopy has developed rapidly over last few decades, during which time it has been used increasingly in the field of food analysis, notably cereal products.[3,4] The appearance of instruments that allow fast recording of full NIR spectra and the utilisation of chemometric[5] methods to complex data treatment has extended its application. Now it is possible to develop methodologies of analysis that are fast, without reagent consumption and suitable for on line monitoring.[6] There has been little interest, up until now, for using NIR methods to monitor the quality of sugar cane, but it is a perfectly suitable technique for obtaining these measurments, particularly if on-line monitoring in the distilleries is required.

Usually, multivariate calibration[7] needs to be employed to establish the relationship between the full spectra (or parts of the spectra) and the analytes of interest. Partial least squares (PLS)[7] and principal component regression (PCR)[7] are factors based on methods that have been used in NIR[8] quantitative determinations and they can be considered, in addition to multiple linear regression,[9] as the patterns in multivariate calibration. Multiple linear regression was the first multivariate approach used in NIR determinations and until now is preferred by many users, since the model developed is easier to interpret and understand from a physical chemistry point of view. However, its application is limited by the use of few model variables, because the number of samples have to be equal or greater than the number of variables (absorbances in different wavelengths). Results obtained using multiple linear regression are similar or better than PLS,[10] when it is possible to select a set of few optimum variables.

Near Infrared Spectroscopy: Proceedings of the 9th International Conference

Based on this fact, several methods have been proposed for variable selection[10] and the genetic algorithm[11] is one of the most useful methods for this purpose.

Genetic algorithm is based on Darwin's theory of evolution and has been used for optimisation processes. The main advantage in using this algorithm for variable selection is that it can extract the best subset of variables from the whole set which approximate the absolute optimum. The criterion used for the optimum, is the minor value of the standard error value obtained after multivariate calibration predictions. This characteristic of the genetic algorithm is a result of its random (non selection criteria at the beginning) processing of variables and a parallel approach aimed at producing the best prediction results for a given subset of variables.

In this work, an investigation into the suitability of near infrared spectroscopy for the rapid estimation of sugar cane juice quality components, such as Brix and pol, was carried out by using multiple linear regression with variable selections by genetic algorithm. Results show that there are no significant differences between the values obtained by the proposed method and those obtained by conventional analysis for the quality parameters monitored.

Experimental

Data acquisition

For all quality parameters monitored in sugar cane juice, the spectra was acquired using a FEMTO PLS PLUS dispersive spectrophotometer with a quartz cell which has a 1 mm optical path. The spectral region monitored was in the range of 1200 to 2500 nm, with 2 nm of resolution.

Data set for Brix determination

307 samples of sugar cane juice were collected in different batches over a period of three months. Each sample was submitted to conventional analysis for Brix determination (by refractometer), followed by spectral acquisition in the near infrared region. Figure 1 shows the analysis of spectra of sugar cane juice.

For spectral baseline correction multiplicative scattering correction (MSC)[12] was applied, followed by the elimination of an intense band in the region of 1900 nm, due the water absorption. After data pretreatment, genetic algorithm was used to select the best wavelength sets for producing the best correlation with brix values using multiple linear regression.

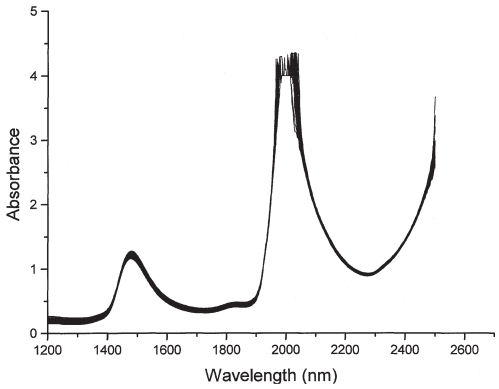

Figure 1. Absorbance spectra of sugar cane juice.

Data set for Pol determination

235 samples of sugar cane juice collected over a period of three months in different batchesand used for determining pol. Each sample was first analysed by conventional methods (by saccharimeter), followed by spectral acquisition in the near infrared region.

For spectral baseline correction multiplicative scattering correction (MSC) was applied. A previous study has shown that the best results for Pol determination were in the region in the range of 1310–1870 nm. After preprocessing genetic algorithm was used for selecting the best wavelength sets.

Data arrangement for modelling

The sets for Brix and Pol determination were split in three groups: one group was used for calibration model development, a second group was used to test the variables selected by genetic algorithm and a third group was used to validate the final model developed. The validation set was only employed after the final variable selection by genetic algorithm, to evaluate the generalisation capability of the model and to certify that the developed model based on variables selected was able to produce good predictions for new samples. This sample selection was based on scores similar to the first and second principal components,[5] as shown for Brix determination in Figure 2.

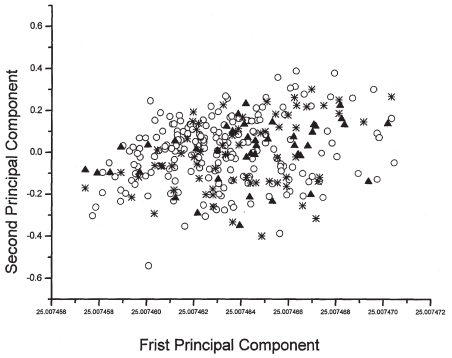

Figure 2. Chart showing the first vs. second principal component for brix determination. O **calibration samples; * test samples; ▲ validation samples.**

The calibration set was chosen for having a wide concentration range for both analytes, while the validation and test sets were prepared to have the analytes in concentrations encompassed by that range.

For the determination of Brix, 195 spectra were chosen for calibration, 68 spectra for testing the variables selected by genetic algorithm and 44 spectra for the final validation. For determining pol, 148 samples were selected for calibration, 50 samples for testing thee variables selected by genetic algorithm and 37 samples for the validation.

Partial least squares modelling

Partial least squares (PLS) modelling was also employed for determining Brix and pol and the results were compared with GA/MLR. The PLS was based on NIPALS algorithm and is similar to that described elsewhere.[7]

The same sample sets were employed to build the PLS model and because there is no distinction between validation and test sets, only the validation set was used to compare with GA/MLR results. The number of factors used in the PLS modelling was determined by cross-validation,[7] resulting in the choice of eight factors for Brix and seven for Pol analysis.

Selection of variables by genetic algorithm

The variable selection by GA was made using a maximum number of individuals per generation equal to 80 and a finalisation criteria of 2000 generations or fitness smaller than 0.01. The cross-over probability was set to 90% and the mutation probability was set at 1%. The maximum number of selected variables was never higher than 25.

Computation programs

The genetic algorithm (GA) and the multiple linear regression (MLR) programs were written in MATLAB, version 5.2 for Windows. The PLS program utilised was from the PLS Toolbox for use with MATLAB, version 2.0.[13] The programs were run on a Pentium II 300 IBM compatible, 64 Mbytes RAM, microcomputer.

Results and discussion

The first step in this work was the data preprocessing, as well as the selection of the major information spectral region, following the reduction in the number of independent variables (absorbance values at different wavelengths).

Preprocessing was performed using multivariate scattering correction (MSC) to correct the base line variation in the whole data set of sugar cane juice. This procedure was adopted because the spectra were acquired on different days, resulting in different experimental conditions, causing variations in the spectral baseline. The spectra region selection was performed to minimise noise effects and to eliminate redundant information into the model developed, even before the genetic algorithm had been applied. For the Brix determination the number of independent variables were reduced from 626 to 568 and for Pol determination the number of variables were reduced from 626 to 292.

In the study of genetic algorithm application for selectioning wavelengths, for Brix determination, using multiple linear regression (MLR),20 wavelengths (1296, 1337, 1412, 1502, 1512, 1554, 1662, 1689, 1749, 1787, 1789, 1812, 1861, 1938, 2059, 2124, 2284, 2327, 2371 and 2384 nm) were necessary. Most of the wavelengths selected could be attributed to overtones and combination bands of –CH, $–CH_2$, $–CH_3$ and –OH bonds. These wavelengths were correlated with the sugars (glucose, fructose and sucrose) dissolved in solution. In the literature,[14] several bands of sugar cane juice found in the NIR spectrum were identified and their attribution was realised. There is a band around the 2500 nm region attributed to the combination of C–H, C–C, C–O–C stretchings. There are several bands in the region of 2280–2330 nm from the combination of C–H stretching and CH_2 deformation. There is a band at 2100 nm which resulted from the combination of O–H stretching and H–O–H deformation. Finally, in the 1450 nm region a band appeared from the first overtone of O–H stretching.

After variable selection by genetic algorithm, there is a significant improvement in the results for Brix, as showed in Table 1, where the standard error of prediction (*SEP*),[7] using PLS with all the variables, is compared with the MLR/GA.

Figure 3 is presents a graph with brix values determined by conventional methods using a refractometer against the values found by NIR procedurse using MLR/GA. An excellent correlation is obtained (0.9828), showing that the proposed method does not present important deviations from conventional methods. In this graph the results for validation and test sets are

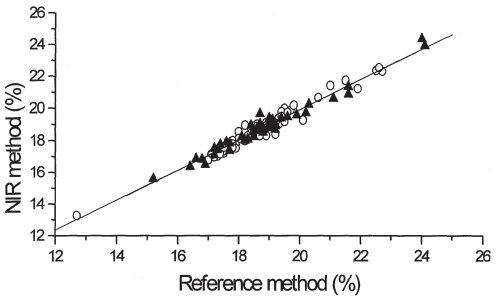

Figure 3. Reference vs. NIR method for brix determination.
▲ validation set ; O **test set**

Table 1. Standard error of prediction for PLS and MLR in brix determination.

	Test set (%)	Validation set (%)
Partial Least Squares (PLS)	0.57	0.53
Multiple Linear Regression (MLR/GA)	0.32	0.32

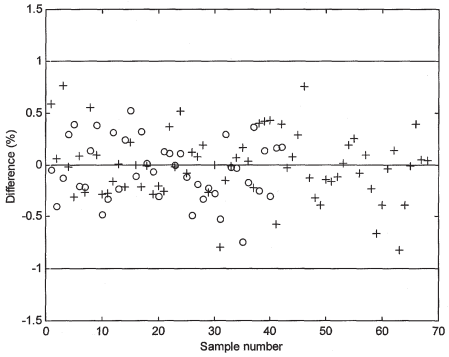

Figure 4. Difference between reference value and NIR prediction value for each sample in brix determination.
O **validation set;** + **test set.**

plotted, where it is possible to observe that there is no difference between the two curves denoting a robust model. This feature is also observed in Table 1 where the *SEP* values for both test and validation sets are identical.

Figure 4 shows a picture where the difference (in %) between the reference method and NIR results are plotted for all samples in validation and test sets. For all samples, a differences of less than 1% were obtained. These results are significant because in order to replace a reference method of analysis for routine monitoring in distilleries, the difference must not be higher than 1%.

The second quality parameter determined by NIR was the Pol. For data modelling, partial least squares and linear multiple regression with variable selected by genetic algorithm were used. In the variable selection, 11 wavelengths were chosen : 1312, 1316, 1368, 1423, 1443, 1533, 1570, 1612, 1614, 1699 and 1739 nm. These selected wavelengths were also correlated to overtones and combina-

Table 2. Standard error of prediction for PLS and MLR in pol determination.

	Test set (%)	Validation set (%)
Partial least squares (PLS)	0.65	0.63
Multiple linear regression (MLR/GA)	0.53	0.63

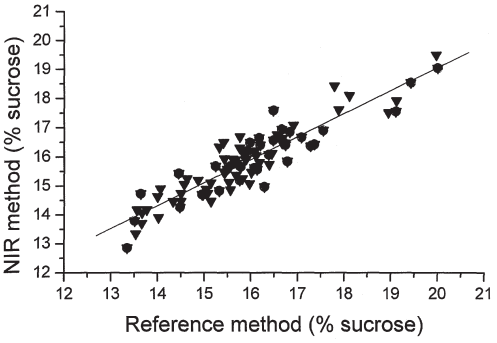

Figure 5. Reference vs. NIR method for pol determination.
▲ **validation set ;** ● **test set**

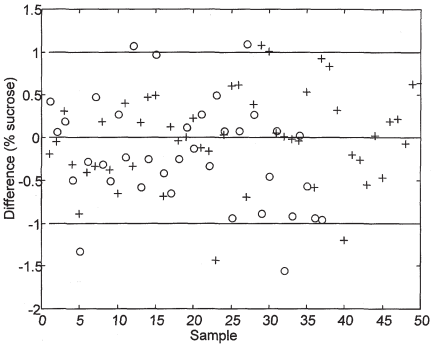

Figure 6. Difference between reference and NIR prediction value for each sample in pol determination.
O **validation set ;** + **test set**

tion bands of –CH, $–CH_2$, $–CH_3$ and –OH bonds, because pol indicates a concentration of sucrose in the solution.

Table 2 shows the results, in terms of SEP, for the two models studied. In this case, the results for PLS and MLR/GA are quite similar, but the MLR/GA has the advantage that only 11 variables are used and the model is simpler to interpret.

Figure 5 presents a plot of Pol (corrected for % sucrose) determined by conventional polarimetric methods against the NIR results. A good correlation coefficient is obtained (0.9184), showing that the two methods are very similar. Again the values of the results for test and validation sets are plotted together to verify robustness of the model.

Figure 6 shows the differences between determinations using the conventional and NIR methods for all samples in the test and validation sets. These differences are never less than 1.5 %, denoting that these procedures can be implemented in the routine monitoring sugar cane juice.

Conclusions

The GA approach to variable selection of NIR spectral data has again proved to be a useful tool. The improved results, like those obtained for Brix determination, obtained by MLR modelling procedures, agree with previous studies.[15] Even the results for determining pol are quite similar to PLS, but the model is simpler to understand and to be implemented in routine controls.

There was no significant difference between values obtained by the proposed NIR methodology and those obtained by conventional analysis for both quality parameters analysed, indicating that NIR can be used in sugar cane juice quality control . The next step, on line monitoring using near infrared spectroscopy, will be evaluated for automatic distillery control.

Acknowledgement

The authors are grateful to the FAPESP proc. 98/01919-6 for P.A. Costa Filho's fellowship and to FEMTO Ind. Com. Instrumentos Ltda (Brazil) for the use of an NIR spectrophotometer.

References

1. F.L. Hart and H.J. Fisher, *Modern Food Analysis.* Springer-Verlag, New York, USA (1971).
2. G.P. Meade and J.C.P. Chen, *Sugar Cane Handbook.* 11th edn. Wiley, New York, USA (1985).
3. P. William and K. Norris, *Near Infrared Technology in agricultural and foods industry.* American Association of Cereal Chemists, St. Paul, Minnesota, USA (1987).
4. B.G. Osborne and T. Fearn, *Near Infrared Spectroscopy in Food Analysis.* Wiley, New York, USA (1986).
5. M.J. Adams, *Chemometrics in Analytical Spectroscopy.* The Royal Society of Chemistry, Wolverhampton, UK (1995).
6. M. Martin, A. Garrison, M. Roberts, P. Hall and C. Moore, *Process Control Qual.* **5,** 187 (1993).
7. H. Martens and T. Næs, *Multivariate Calibration*. John Wiley & Sons, Chichester, UK (1989).
8. J.J. Workman, Jr, P.R. Mobley, B. Kowalski and R. Bro, *Appl. Spectrosc. Rev.* **31,** 73 (1996).
9. N. Draper and H. Smith, *Applied Regression Analysis,* 2nd edn. Wiley, New York, USA (1981).
10. D. Juan-Rimbaud, D.L. Massart, R. Leardi and O.E. de Noord, *Anal. Chem.* **67,** 4295 (1995).
11. D.B. Hibbert, *Chemometrics Intell. Lab. Syst.* **19,** 277 (1993).
12. T. Isaksson and B.R. Kowalski, *Appl. Spectrosc.* **47,** 702 (1993).
13. B.M. Wise and N.B. Gallagher, *PLS_Toolbox for use with Matlab*, ver. 1.5.1. Eigenvector Technologies, Manson (1995).
14. J.J. Workman Jr, *Appl. Spectrosc. Rev.* **31,** 251 (1996).
15. R. Guchardi, P.A. da Costa Filho, R.J. Poppi and C. Pasquini, *J. Near Infrared Spectrosc.* **6,** 333 (1998).