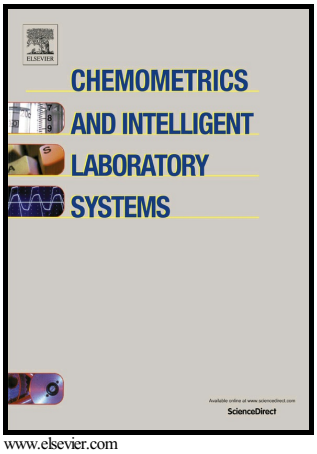
Optimization of NIR calibration models for multiple processes in the sugar industry

Iván Ramírez-Morales, Daniel Rivero, Enrique Fernández-Blanco, Alejandro Pazos



www.elsevier.com

PII: S0169-7439(16)30360-4
DOI: http://dx.doi.org/10.1016/j.chemolab.2016.10.003
Reference: CHEMOM3325

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received date: 1 July 2016
Revised date: 20 September 2016
Accepted date: 5 October 2016

Cite this article as: Iván Ramírez-Morales, Daniel Rivero, Enrique Fernández Blanco and Alejandro Pazos, Optimization of NIR calibration models fo multiple processes in the sugar industry, *Chemometrics and Intelligen Laboratory Systems,* http://dx.doi.org/10.1016/j.chemolab.2016.10.003

Optimization of NIR calibration models for multiple processes in the sugar industry

Iván Ramírez-Morales[*ab], Daniel Rivero[b], Enrique Fernández-Blanco[b] , Alejandro Pazos[b]

[a] **Universidad Técnica de Machala,** Faculty of Agricultural & Livestock Sciences. Address: 5.5 km Pan-American Av, Machala, El Oro, Ecuador.

[b] **Universidade A Coruña,** Department of Computer Science. Address: 15071 A Coruña (03082), España.

[*] Corresponding Author: iramirez@utmachala.edu.ec

Optimization of NIR calibration models for multiple processes in the sugar industry

Abstract

The measurements of Near-Infrared (NIR) Spectroscopy, combined with data analysis techniques, are widely used for quality control in food production processes.

This paper presents a methodology to optimize the calibration models of NIR spectra in four different stages in a sugar factory. The models were designed for quality monitoring, particularly °Brix and Sucrose, both common parameters in the sugar industry.

A three stage optimization methodology, including pre-processing selection, feature selection and support vector machines regression metaparameters tuning, were applied to the spectral data divided by repeated cross-validation. Global models were optimized while endeavoring to ensure they are able to estimate both quality parameters with a single calibration, for the four steps of the process.

The proposed models improve the prediction for the test set (unseen data) compared to previously published models, resulting in a more accurate quality assessment of the intermediate products of the process in the sugar industry.

Keywords: NIR; chemometrics; calibration models; machine learning; support vector machines; agro-industry;

1. Introduction

The production flow in the sugar industry encompasses several processes and subprocesses that need to be analyzed in order to maintain a quality standard [1]. The agro-industrial plants require cost-efficient and non-destructive systems to monitor the quality of their production process, food safety and compliance with the technical specifications [2].

One of these non-destructive systems aimed at ensuring quality is chemometrics, which has been developing since the 1970s as an interdisciplinary field of study. This field covers a

wide and varied range of mathematical and statistical techniques for analyzing the chemical composition of materials [3].

To analyze the quality of organic raw materials, a commonly-used technique is the Near-Infrared Reflectance (NIR) spectroscopy, associated with chemometrics; however, the relationship between the absorption in the spectral region of the near infrared and the analyte is frequently of a non-linear type [4].

The origin of these non-linear relationships is diverse and difficult to identify, in some cases due to the differences in viscosity, temperature, pH, particle size and the chemical nature of the analyte. For this reason, calibration is commonly performed using non-linear methods and multivariate analysis [5]. A proper selection of the variables aimed at gathering a small subgroup with lower sensitivity to non-linearities or at discarding the most pronounced wavelengths is usually effective to improve the performance of the models [6,7].

Chemometrics is an essential part of NIR spectroscopy in the food sector [28], recently, with the development of information technologies, the applications of NIR spectroscopy have become increasingly popular and arousing great interest among researchers, considering that the technique is able to detect analyte concentrations of 0.1% w/w [8].

Common chemometric techniques use multivariate analysis methods, such as PCA as a qualitative analysis technique of the spectral data, and PLS regression analysis as a technique for quantitative prediction of the parameters of interest in the sample [2]. Currently, the literature that applies machine learning techniques in chemometrics is in constant expansion [9–12], the use of artificial neural networks, support vector regression (SVR) is also reported [5,24,30] as these techniques are based on pattern recognition [11].

The food industry has widely used NIR spectroscopy[26] since it is a rapid, accurate, minimally invasive, non-destructive quality analysis technique [2]. NIR has been used to analyze quality of dairy products [13,14], oils [14], meat products [15], fish [16], cereals [17] and fruit [18].

In the sugarcane industry, studies were found which proved a good correlation between the NIR spectra and quality indicators of sugarcane [19]. The use of NIR spectra preprocessing techniques in sugar cane was analyzed [20], along with the selection of features [21] and chemometric algorithms in order to improve prediction of target analytes [22,23].

A recent study conducted by Tange *et al*. [24] has shown that the use of calibration models with Support Vector Machines (SVM) for regression is efficient in order to predict °Brix and Sucrose values, the quality parameters of the industrial process of sugar. The use of SVM improves in terms of *RMSE* compared to the technique of Partial Least Square (PLS); however, the proposed model uses the entire NIR spectrum, which leads us to believe that the optimization of the model is still possible by implementing an appropriate preprocessing technique, feature selection and optimization of the parameters of the machine support vectors.

The aim of this study is to optimize global calibration NIR models in order to improve quality control of the °Brix and Sucrose parameters. A global calibration NIR model is capable of predict a value in the four steps of the sugar production process.

2. Materials and methods

Generally, the sample processing consists of the following steps: acquisition of spectral data, preprocessing of the data to reduce the noise, thereby increasing the signal-to-noise ratio (S/N) [25], selection of relevant features, and development of the calibration model using a set of spectra from which the values of target analytes obtained by reference techniques are known, and, finally, the model validation using data different to those of calibration [8].

2.1 Data description

The database was published by Tange *et al*. [24], and the employed data were obtained in a Japanese sugar factory (Daito Togyo Co), where sugar cane is processed. The samples were obtained throughout three months in the harvest season, and during each of the process steps: after grinding (juice), after the process of evaporation (syrup), after crystallization (massecuite) and after centrifugation (molasses), three cycles of crystallization and centrifugation were carried out, resulting in a higher number of samples of massecuite and molasses than those obtained in the other two stages. Immediately after sampling, the NIR signals were extracted, along with the reference technique in relation to process temperature.

The °Brix quality parameter expresses all the dissolved solids (sugar and non-sugar) as a percentage of total weight, its scale reflecting the percentage of sucrose in pure solutions. In any sugar materials (juice, honey, etc.) the °Brix parameters are always higher than those of Sucrose, whereas in high-purity materials, such as spirits from a refinery, the difference between these indicators is minimal. For the current work, the °Brix quality parameter was measured using an Abbemat-WR refractometer, developed by Anton Paar GmbH in Germany.

The Sucrose quality parameter (POL) refers to the amount of sucrose contained in a solution, expressed as percent of the weight; in pure solutions, the POL percentage is equivalent to the percentage of Sucrose, while in other impure solutions, like cane juice and honey, there is a difference between these two values, the more impure is the solution, the higher the difference. For this reason, the POL value is internationally accepted as apparent sucrose. In the current paper, Sucrose was measured using a MCP500 polarimeter, developed by Anton Paar GmbH in Germany.

2.2 Near Infrared Reflectance

NIR spectra are obtained as a result of vibrational transitions primarily associated with chemical bonds containing hydrogen, C-H, N-H, S-H and O-H, and which are present in most organic compounds; the NIR spectrum region spans within the wavelength range between 780 nm and 2500 nm [2,28,29].

The NIR spectrum, at all wavelengths, may contain information regarding several analytes [30] and other physical features as temperature, viscosity, crystals and pH. This implies that the resulting spectrum is the consequence of the modifications carried out simultaneously in all the analytes in the sample, making the calibration process more complicated [5,28,31].

A total of 1797 NIR spectra ranging between 400.0 nm and 1888 nm, with an increase of 2 nm, were obtained using an NIR DS2500 spectrometer, developed by FOSS AB in Denmark. In the work presented by Tange *et al*. [24], the spectral signals with an absorbance value greater than two were excluded from the analysis; in contrast, the current study used the full dataset.

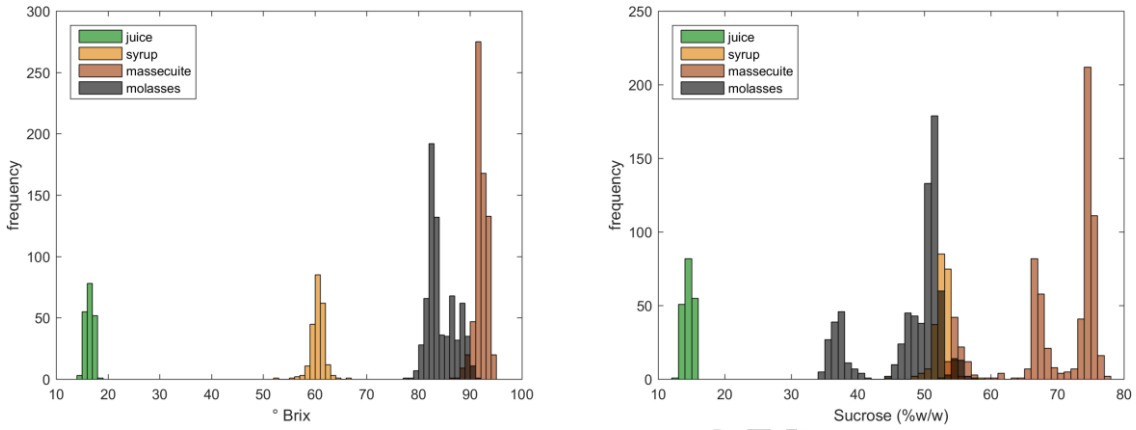


Figure 1. Histograms of °Brix and Sucrose content for each process step.

Figure 1. shows the histograms of spectra according to their content of °Brix and Sucrose content for each process step, for the °Brix data, a pronounced separation in the frequency distributions is observed, with a slight overlapping between massecuite and molasses in the range around 90 °Brix. On the other hand, the sucrose measurements show overlapping between syrup, massecuite and molasses, in the range around 50% w/w, of sucrose.

2.3 Preprocessing techniques

The near-infrared (NIR) spectral data preprocessing is an integral part of chemometric modeling. The purpose of preprocessing is to remove physical phenomena from the spectra [28,30,32].

The NIR spectra of solid samples are influenced by the physical properties of the samples, making preprocessing increasingly important to minimize the contributions which contain irrelevant information, therefore, simpler and more robust models could be developed [28,30]. The proper choice of the preprocessing technique is difficult to assess before the validation of the model, therefore, NIR spectra preprocessing is still carried out through trial and error [32,33].

NIR spectroscopy has led to both a greater number and diversity of preprocessing techniques, primarily because the spectra may be significantly influenced by the non-linearities introduced by the light scattering [32]. This study applies four basic techniques for

the preprocessing combinations of NIR spectra: the Beer-Lambert law, the First Spectral Derivative, Standard Normal Variate, and detrending.

2.3.1 Beer-Lambert law: is empirical for NIR spectra and suggests a linear relationship between the absorbance of spectra and concentration(s) of the constituent; this law is valid only for systems of pure transmittance without scattering; in reflectance measurements, the law can be expressed as follows [32]:

$$A\lambda = -\log_{10}(R) \cong \epsilon\lambda \times l \times c$$

Where *Aλ* is the absorbance dependent on wavelength, R is the reflectance detected, ελ is the molar absorptivity dependent on wavelength, *l* is the effective length of the light path through the sample matrix, and *c* is the concentration of the component of interest.

According to this law, the spectra is processed by taking the base 10 negative logarithmic calculation of the sample's reflectance, which results in the linear relationship to the concentration of the analyte.

2.3.2 First Spectral Derivative (FSD): have the ability to remove both additive and multiplicative effects from the spectra and have been used in analytical spectroscopy for decades[28]. The most basic method for derivation is finite differences [32]; the first order derivative is calculated as the difference between two subsequent spectral measurement points:

$$X_{i,fsd} = X_i - X_{i-1}$$

Where $X_{i,fsd}$ denotes the first order derivative at wavelength *i*. This technique removes only the baseline of the spectra.

2.3.3 Standard Normal Variate (SNV): is possibly the second most applied method for scattering correction of NIR data [34]. The basic format for SNV, correction and normalization is as follows:

$$X_{i,snv} = \frac{X_i - \bar{x}}{S}$$

Where $X_{i,snv}$ denotes the SNV at a wavelength *i*, $\bar{x}$ is the spectrum average of the sample to be corrected, and *S* is the standard deviation of the spectrum sample.

2.3.4 Detrending technique: is applied to the spectra to remove the effects of the baseline and curvilinearity changes, it is typical of NIR spectra to which the pre-processing technique based on the Beer-Lambert law was applied. This effect is generally linear [35]. The method consists of modeling the baseline as a linear function of the wavelength and this function is subsequently subtracted from each spectrum value independently. The expression is as follows:

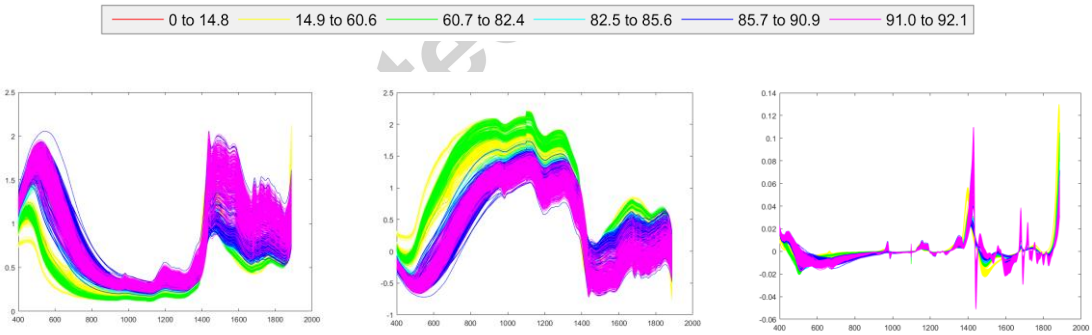
$$X_{i,dt} = X_i - b_i$$

*W*here $X_{i,dt}$ denotes the detrended spectrum at a wavelength *i*, b_i is the linear model baseline of the spectrum in the wavelength *i*.

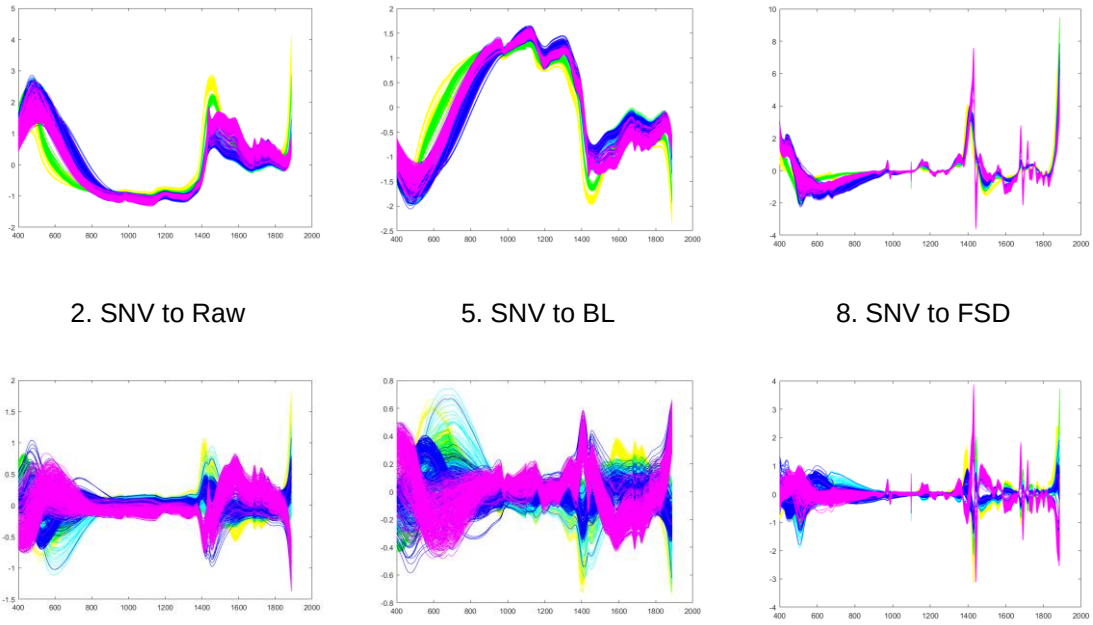
In general, the combination of techniques is not advisable, and, as a minimum requirement, preprocessing should maintain or reduce the complexity of the effective model [32]. In contrast to the above-stated criteria, Xu *et al*. [33] showed that the combination of preprocessing methods improved the stability of the models and the results in terms of *RMSE*, as it takes advantage of the complementary information given by each preprocessing method, therefore, the stability of the models and the results are improved in terms of *RMSE*. Other recent studies have obtained good results by combining preprocessing techniques [36–39]. In our work, preliminary tests with commonly used preprocessing methods were performed, and based on the results of these tests, nine combinations were defined, as explained below:

1. Raw: unprocessed reflectance signal (as extracted from the instrument).
2. SVN to Raw: the Standard Normal Variate was calculated for the raw signal.
3. SNN to Raw & detrend: a technique of detrending is applied to the processed signal in Combination 2.
4. Beer Lambert (BL): the Beer-Lambert law is applied to the raw signal.
5. SNV to BL: calculation of SNV for the processed signal in Combination 4.
6. SNV to BL & detrend: a technique of detrending is applied to the processed signal in Combination 5.
7. First spectral derivative (FSD): the calculation of the first spectral derivative from the raw signal.
8. SNV to FSD: calculation of SNV for the processed signal in Combination 7.

SNV to FSD & detrend: a technique of detrending is applied to the processed signal in Combination 8.



1. Raw 4. Beer-Lambert law 7. First spectral derivative



2. SNV to Raw | 5. SNV to BL | 8. SNV to FSD

3. SNV to Raw & detrend | 6. SNV to BL & detrend | 9 SNV to FSD & detrend

Figure 2. Changes in the NIR spectra when applying the preprocessing techniques

Figure 2 shows nine alternative NIR spectra transformations that result from combining the above-mentioned basic preprocessing techniques, which will be evaluated in the first stage. In order to better illustrate this section for the reader, spectra were colored according to their respective °Brix values; it should also be noted how the differences are softened and accentuated between the different shades in the NIR spectrum, when applying preprocessing techniques.

2.4 Feature Selection Techniques

Due to the large amount of spectral information provided by NIR spectrophotometers, a substantial reduction of the number of samples needed to build the classification and calibration models is required [30]. Over the past decade, in the construction of models, feature selection (FS) has passed from illustrative examples regarding its operation to becoming a requirement, particularly due to the nature of the high-dimensionality problems, such as microarray and spectral analyses [7].

Considering that many of the pattern recognition techniques were not originally designed to deal with large amounts of information with little relevance, the application of FS techniques has currently become a necessity in many applications [40]. Its application prevents overfitting of the model, improves performance and reduces the computation time, obtaining a deeper understanding of the data [7,41,42]. FS does not alter the original representation of the variables, since it simply consists of selecting a subset of the best features, preserving their original nature, allowing them to be easily interpretable by an expert in the field [7].

A group of techniques, commonly used for FS, are filters which evaluate the relevance of a feature, analyzing the intrinsic properties of the data; generally, a score of relevance is

calculated and the characteristics with lower score are eliminated [7,43]. There are two types, univariate and multivariate filters [44]. The univariate filter is a simple but efficient paradigm, which is fast, scalable and independent of the classification/regression technique [45]. A threshold method is usually defined to select the filters which meet a condition above or below the threshold. Filters work regardless of the model and use the intrinsic properties of the data [43].

The univariate filter method can be applied using a t-test [46], an F-test [47] or the Wilcoxon signed-rank test [43], by calculating a p-value which represents the statistical significance of each variable in the model, thus variables are organized depending upon their p-value [7].

2.5 Support Vector Regression

There are two types of machine learning algorithms: supervised and unsupervised; the former are used when there is prior knowledge about the desired outputs of the model, while the latter generate similar groups without having prior information [48]. Once a machine learning algorithm is trained, it is able to transfer what it learned to new data [49], that is, to new data that are not used during training, this set allows evaluating the generalization error of the final model chosen.

Support vector machines (SVM) are supervised machine learning algorithms [50], based on the structural risk minimization, and can be used in classification and regression (SVR) problems. Their operation starts from a set of training patterns whose outputs are known, and which allow making predictions about new patterns [51,52].

The principles of SVM were developed by Vapnik and Chervonenkins in 1963, in a study on statistical learning theories that aimed to reduce the error generalization according to the complexity of the search space [53]. The current standard of SVM was proposed by Cortes and Vapnik [54]. The purpose of SVM is to obtain models which structurally present little risk of error with respect to future data. Although originally designed to solve binary (two classes) classification problems, their application was extended to regression, multiclass classification, clustering and other tasks [52].

A version of an SVM for regression was proposed in 1997 by [55]. This method is called Support Vector Regression (SVR). The model depends only on a subset of data (support vectors), because the cost function for the construction of the model does not consider the points that are beyond the margin; in addition, the cost function ignores any data that are close to the prediction model, within a threshold ε [51].

SVRs have been applied in several fields, such as time series [56], finances [57], engineering approaches in complex analyses [58], and convex quadratic programming [59], among others [51].

SVRs are learning methods based on kernel, which makes it possible to perform a transformation of data space, so that data could be described by linear models and the problem could be simplified [60]. The kernel is one of the most important SVR parameters; in the current study, some preliminary tests were carried out by trial and error using the most

common kernels for SVR, in accordance with [61]. However, only the radial basis function kernel generated acceptable results, thus it was used to optimize its parameters.

Searching for optimal parameters of an SVR is essential in building a prediction model that is accurate and stable [62,63]. Kernel parameters are adjustable in the SVRs to control the complexity of the resulting hypothesis and to avoid the overfitting of the model [64,65]. The optimization of the parameters *C* (regularization parameter), γ (gamma RBF kernel parameter) and ε (epsilon parameter) is a key step in an SVR, since their combined values determine the complexity of the limits and therefore the performance of the model. Different techniques may be used to optimize these parameters [65,66].

2.6 Experimental optimization

The spectral data were divided into training (calibration) and test subsets using repeated cross-validation [67,68]. A 10-fold repeated cross-validation technique was chosen, in which data were divided into 10 groups, 9 being used as calibration sets and the remaining one as a test set; the test set was changed until all groups have been tested. Cross-validation was repeated 100 times. Support Vector Regression (SVR) was used with a Radial Basis Function (RBF) kernel [69].

There are many approaches to optimize simultaneously the pre-processing and the feature selection in order to improve the prediction ability, one of them is genetic algorithm approach. Devos and Duponchel, [70] found that all cooptimizations converge to the same kind of solutions; another new strategy is described for the combined implementation of variable, pre-processing and sample selection using algorithms like ant optimization colony and genetic algorithms [71]. The evaluation of the parameters in this work was performed using a t-test univariate filter [46], and a grid search method [60,72]. The evaluation was conducted in three stages:

First stage: The aim of this phase was to define what processing technique is most appropriate, and which wavelengths (features) should be selected. To this end, the percentile p-value threshold used in the FS process was evaluated for the nine preprocessing techniques, with fixed values of parameter C equal to 100, and both γ and ε parameters equal to 0.1.

Second stage: The objective of this phase was to find an optimal combination of the C and γ parameters. To achieve this, both were simultaneously optimized, using the grid search method [60,72] on a logarithmic scale. Parameter C was evaluated within the range from 10^1 to 10^6 and γ within the range of 10^{-1} to 10^{-3}, whereas the evaluation interval was 0.25 in the exponent, with a fixed value for ε equal to 0.1.

Grid search is a brute-force approach for parameter optimization which is widely used in machine learning techniques. Once the ranks and measures of the parameters are defined, each combination of parameters is tested in order to find the best one, based on a performance measure [72].

Third stage: The objective of this phase was to tune the performance of SVR. For this, the ε parameter of SVR was optimized within the range of 0 to 1 on linear scale, whereas the evaluation interval was 0.01.

At each stage, 100 repetitions for each CV were performed. In total, 478,000 SVRs were evaluated in order to determine the optimal configuration of the calibration model.

All models were evaluated using as performance measures the r-squared coefficient of determination (R^2) and root mean square error (*RMSE*) in the cross-validation test data as indicators of the accuracy of the model based on NIR [28,73]. The equation for *RMSE* is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2}{n}}$$

Where y_i is the prediction value of the i-th observation, $\hat{y}_i$ is the measured value of the observation, and n is the number of observations.

The *RMSE* consists of the differences between the values predicted by a model and the values actually observed. This technique is often preferred over others, since its interpretation is on the same scale as the data, and it gained popularity because of its theoretical relevance in statistical modeling [74,75]. Mean and standard deviation of 100 repetitions in 10 fold cross-validation, were calculated in order to present the results in graphics and tables.

3. Results

This study presents a methodology which optimizes NIR calibration models for multiple processes in the sugar industry. This consists of a sequence of steps described below:

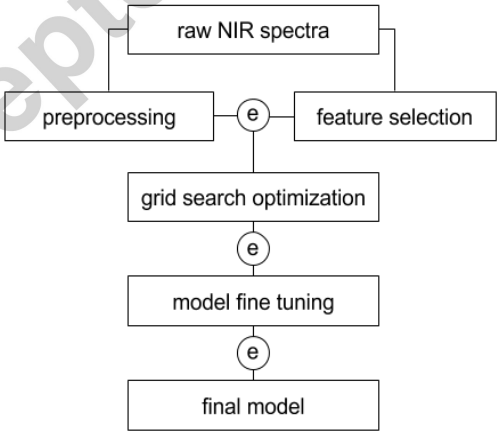


Figure 3. Methodology for the optimization of NIR calibration models used in the current study.

As shown in Figure 3, the choice of the most appropriate preprocessing technique and the feature selection which best models the non-linearities of the spectrum is performed in

parallel. At this point, the model is evaluated and its optimal values are set, in order that the C and γ parameters of SVR can be optimized using a grid search technique. Using these optimized values, a tuning of the parameter ε of SVR which defines the final model is carried out, and each evaluation (e) point corresponds to the evaluation phases outlined in section 3.

3.1 First stage: processing technique and feature selection

The nine alternative preprocessing techniques proposed were evaluated at different intensities of feature selection, which was performed by changing the values of threshold percentile for the p-value of the T-test technique for feature selection.

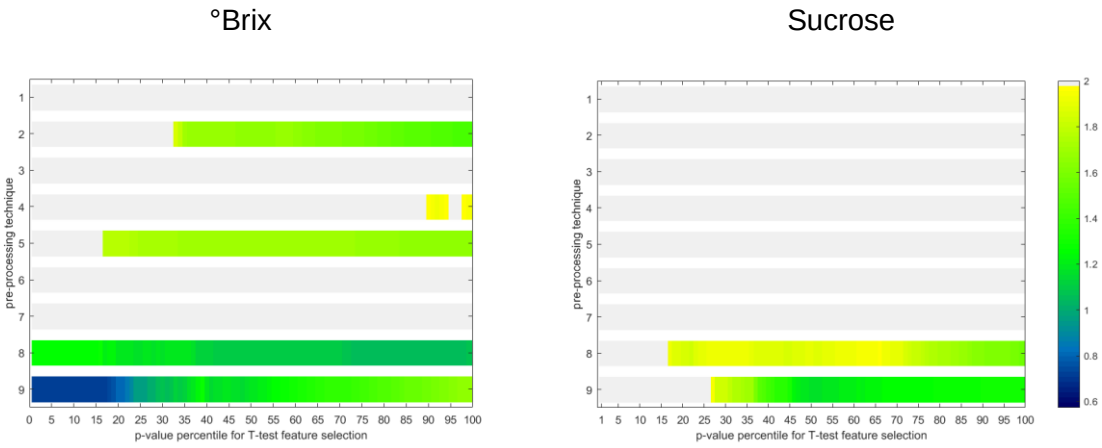


Figure 4: RMSE in CV obtained with the nine preprocessing techniques at different values of percentile p-value for the T-test feature selection.

Figure 4 shows the results of *RMSE* in CV obtained with nine preprocessing techniques at different values of percentile p-value for the T-test feature selection, in which it is observed that both in the case of °Brix and Sucrose, the preprocessing technique number 9 (listed in Figure 2) is the one providing the best results. The lowest values of *RMSE* are recorded within the percentile range between 1 and 17 for °Brix, while for Sucrose, the lowest *RMSE* values were recorded within the percentile range between 45 and 55. The preprocessing technique number 9 consists of combining the calculation of the first derivative, which is subsequently normalized with the SNV technique and finally the trend is extracted.

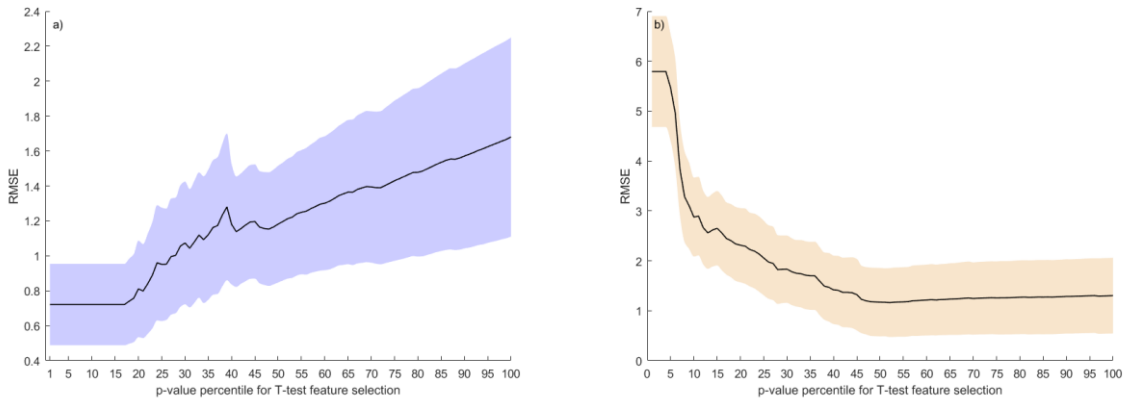


Figure 5: *RMSE* in CV obtained in the model a) °Brix and b) Sucrose, with the preprocessing technique number 9, at different values of percentile p-value for the T-test feature selection.

Figure 5 shows the mean and standard deviation *RMSE* in CV obtained with the preprocessing technique number 9, at different values of percentile p-value for the T-test feature selection. Note that in the case of the °Brix model, the optimal values of percentile p-value for the T-test feature selection are between 1 and 17, a value equal to 10 is selected, since it has a good mean value of *RMSE*; the Sucrose model has an optimal value of percentile p-value for the T-test feature selection equal to 52.

The selected features with a percentile p-value for the T-test feature selection equal to 10 are shown in Figure 6, where the most relevant features for the °Brix model are highlighted in vertical, shaded bands.

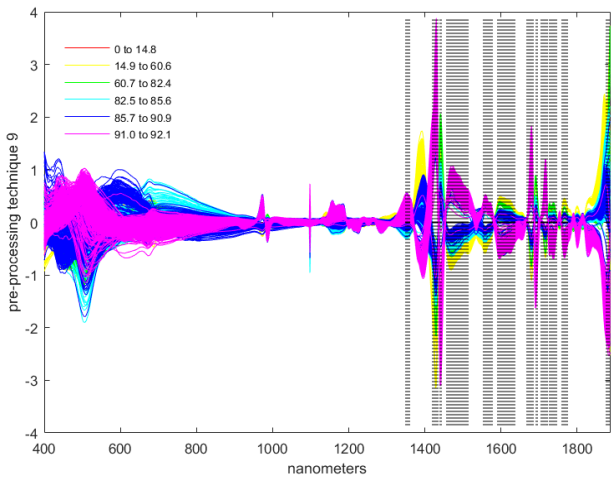


Figure 6: Spectral bands (features) selected to build the prediction model for °Brix.

The selected features with a percentile p-value for the T-test feature selection equal to 52 are shown in Figure 7, where the most relevant features for the Sucrose model are highlighted in vertical, shaded bands.

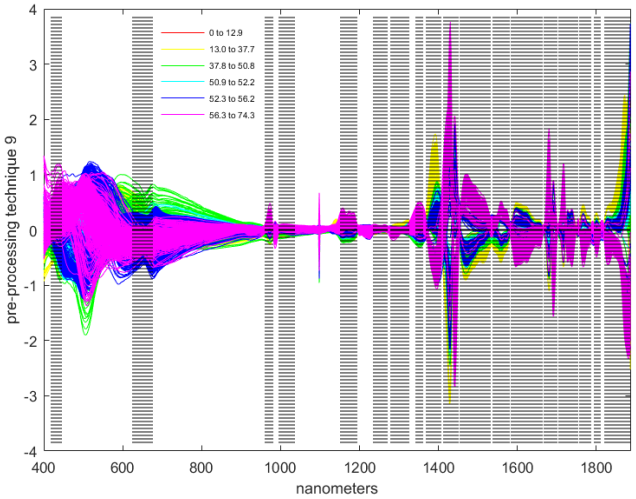


Figure 7: Spectral bands (features) selected to build the prediction model for Sucrose.

3.2 Second stage: Optimization of C and γ parameters of SVR

After applying the preprocessing technique and selecting the wavelengths (features), the calibration models were evaluated in order to find the optimal combination of parameters C and γ, for a fixed ε value equal to 0.1.

The grid search method [44,51] was applied in logarithmic scale, parameter C was evaluated within the range from 10^1 to 10^1, γ within the range from 10^{-1} to 10^{-3}, and the evaluation interval was 0.25 in the exponent.

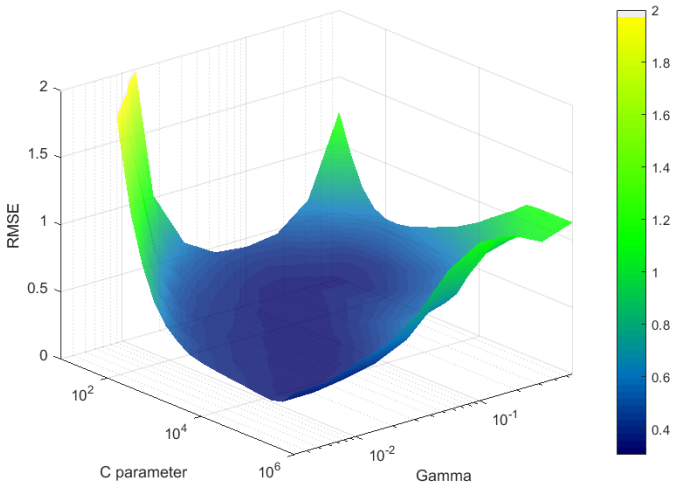


Figure 8: Surface chart of *RMSE* in cross-validation of the prediction model for °Brix according to the values of parameters *C* and γ in all steps of the process.

Figure 8 shows a surface chart with the *RMSE* values obtained by the global model of °Brix for different combinations of parameters *C* and γ; the optimal parameters are 10e+2.75 and

10e-1.25 respectively, with those that together generate the lowest *RMSE* being evaluated in the average test results of the repeated cross-validation.

Figure 9 details the performance of the global model at each step of the process, and it confirms that the optimal parameter values are those presented above.

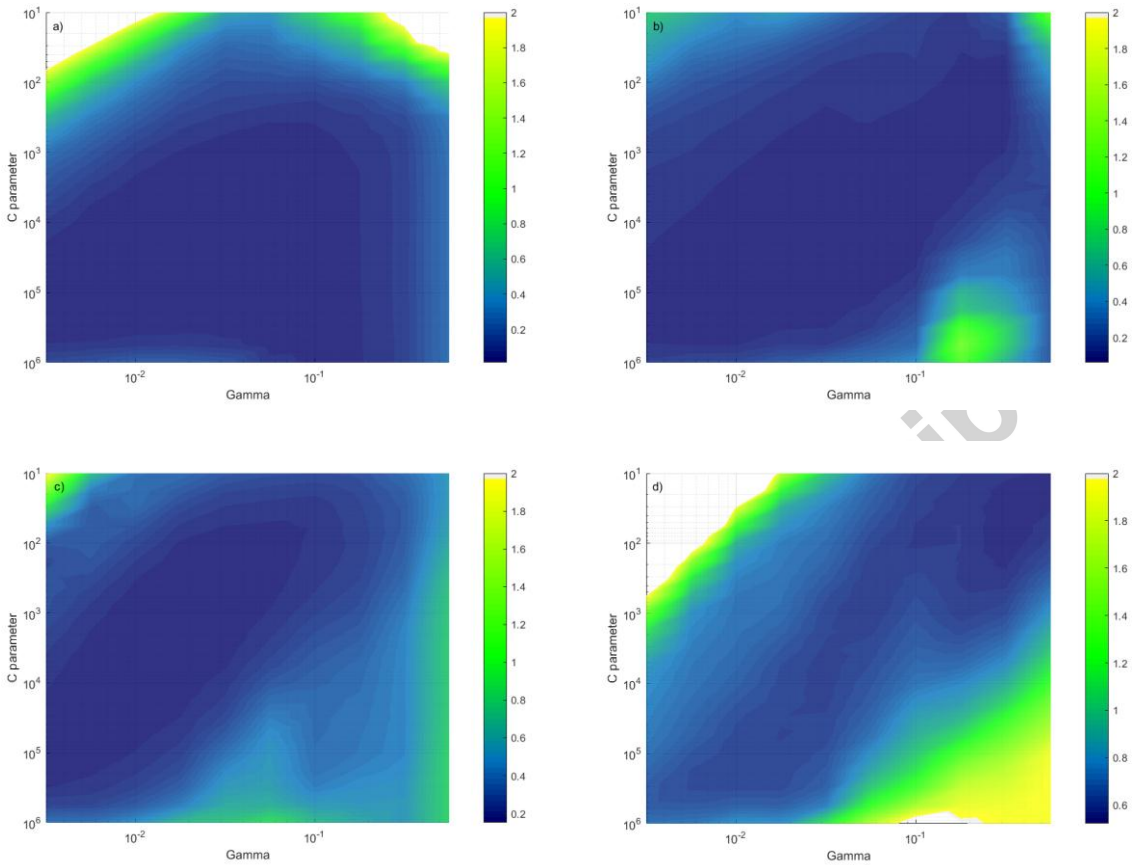


Figure 9. Heat maps of *RMSE* in cross-validation of the prediction model for °Brix according to the values of parameters C and γ in the 4 steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.

Figure 10 shows a surface chart with the *RMSE* values obtained by the global model of Sucrose for different combinations of parameters *C* and γ; the optimal parameters are 10e+3 and 10e-1.75 respectively, with those that together generate the lowest *RMSE* being evaluated in the average test results of the repeated cross-validation.

Figure 11 details the performance of the global model at each step of the process, and it confirms that the optimal parameter values are those presented above.

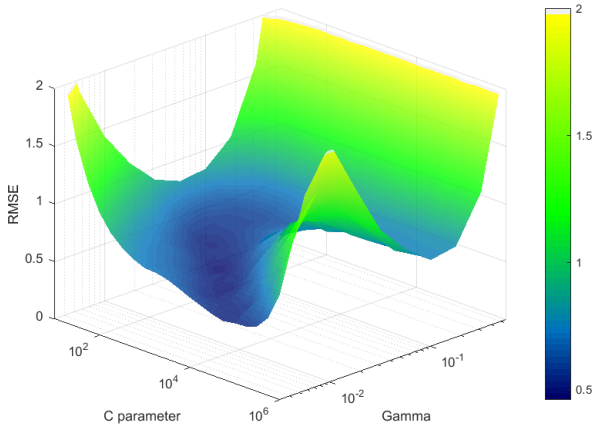


Figure 10: Surface chart of *RMSE* in cross-validation of the prediction model for Sucrose according to the values of parameters C and γ in all steps of the process.

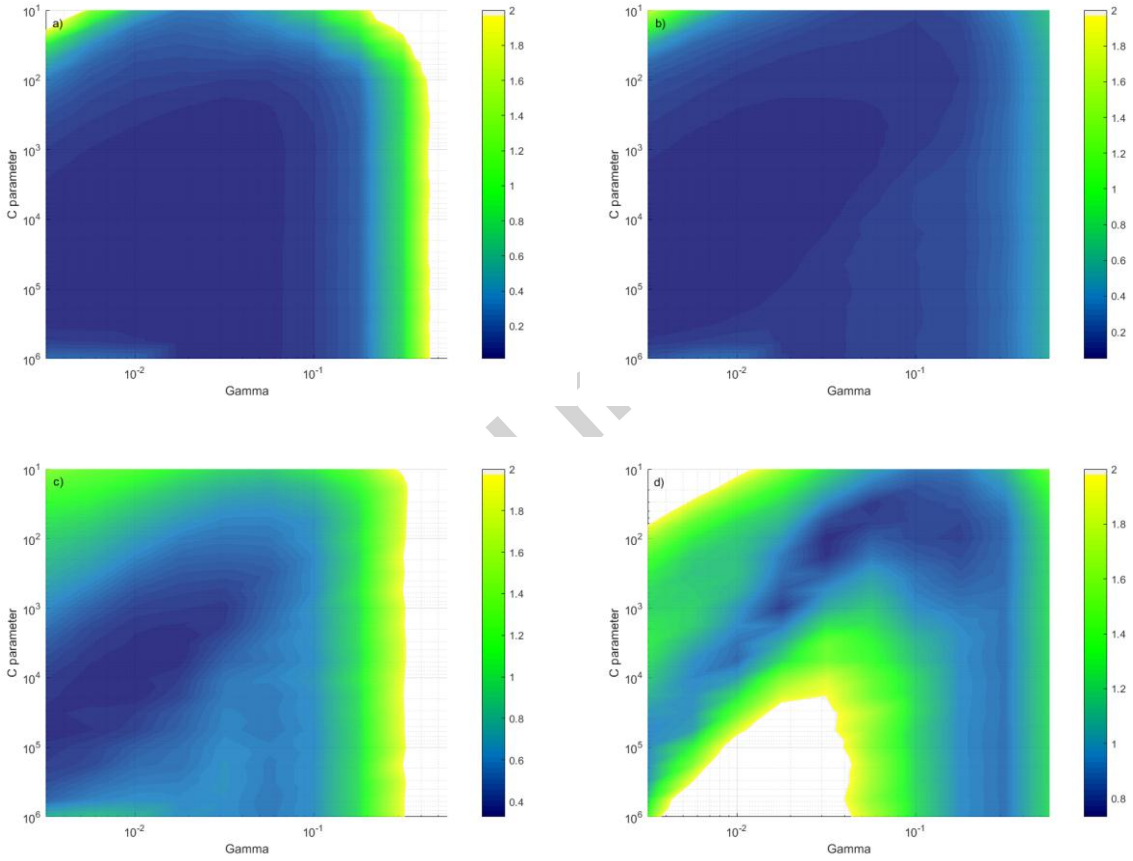


Figure 11. Heat maps of *RMSE* in cross-validation of the prediction model for Sucrose according to the values of parameters C and γ in the 4 steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.

3.3 Third phase: Optimization of parameter ε

Once the preprocessing technique, the wavelengths (features) and the optimal combination of parameters C and γ were selected, the parameter ε of SVR was evaluated within the range of 0 to 1 in the linear scale, and the evaluation interval was 0.01.

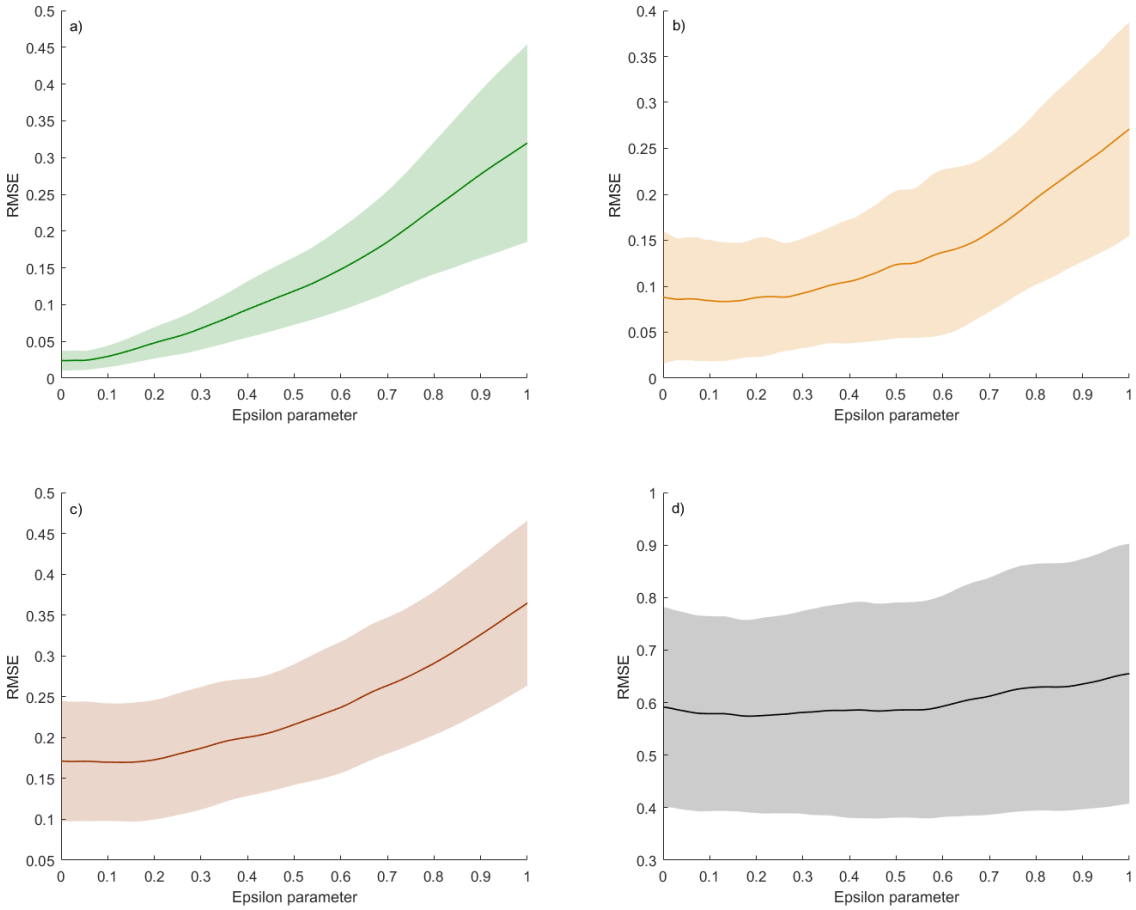
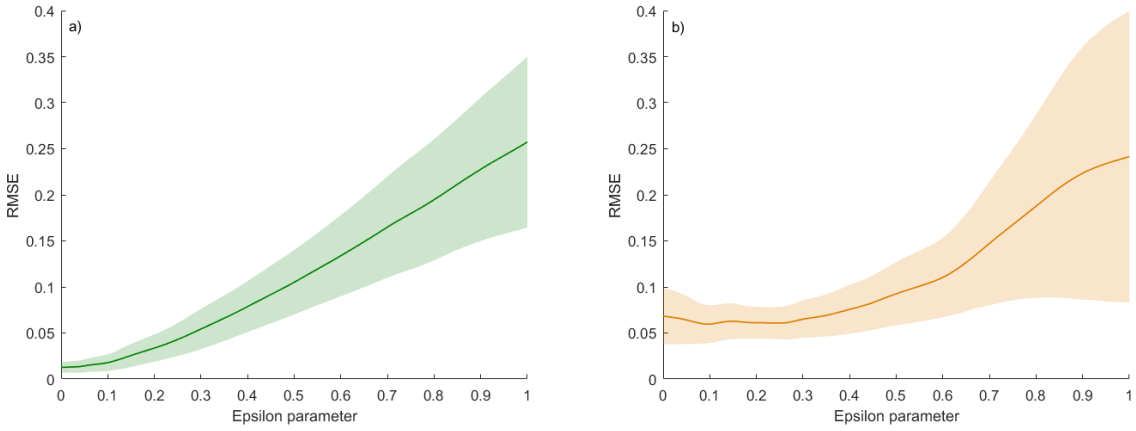


Figure 12. Curves and bands of a standard deviation of *RMSE* in cross-validation of the prediction model for °Brix according to the value of ε in the 4 steps of the process.

Figure 12 shows the curves of mean and standard deviation of *RMSE* for the optimization of the ε parameter in the prediction model for °Brix; it can be observed that a value of ε equal to 0.16 optimizes *RMSE* in the four steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.



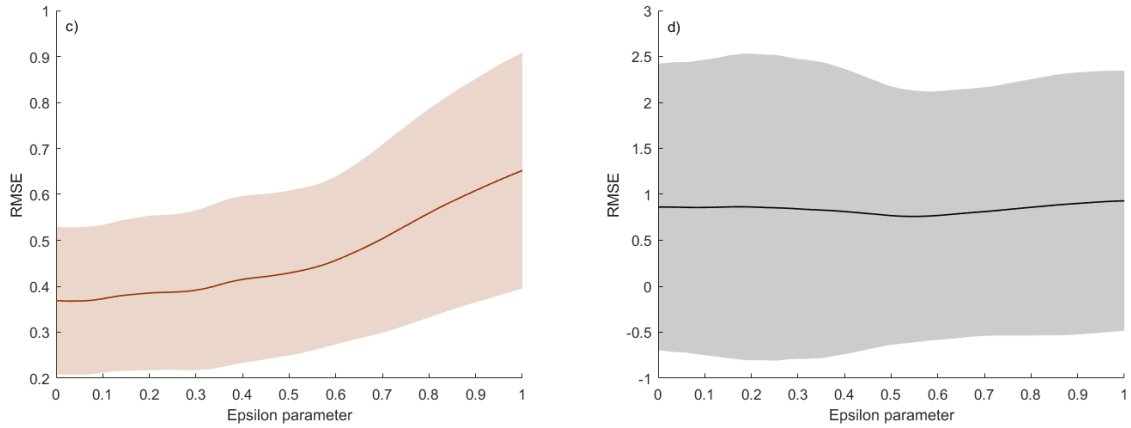


Figure 13. Curves and bands of a standard deviation of *RMSE* in cross-validation of the prediction model for Sucrose according to the value of ε in the 4 steps of the process: a) juice, b) syrup, c) massecuite, d) molasses.

Figure 13 shows the curves of mean and standard deviation for the optimization of the parameter ε in the prediction model for Sucrose; it can be observed that the optimal ε value is 0.07. However, in the case of molasses, the optimal ε value is 0.51; for the discussion section, the models optimized with both values will be evaluated, to consider an improvement in the estimation of Sucrose in the process step of molasses.

Finally, to statistically evaluate the performance of models whose parameters were optimized, R-squared, Adjusted R-squared and p-value were calculated. The model for °Brix was analyzed with its optimal parameters, in the case of Sucrose, the model was analyzed with the option of ε equal to 0.07.

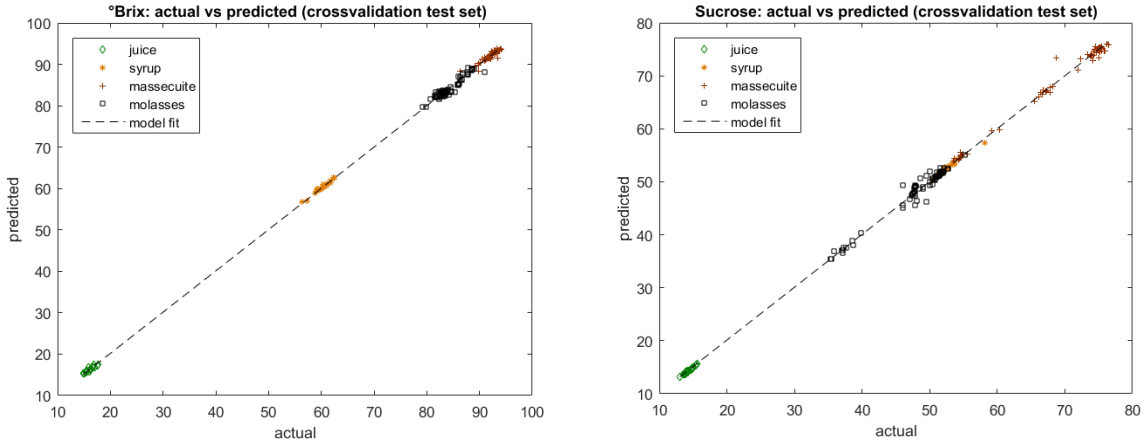


Figure 14. Regression plot (actual vs predicted) in cross-validation global models for °Brix and Sucrose.

Figure 14 shows two regression graphs (actual vs predicted); the calibration model for °Brix obtained an R-squared of 0.99, an Adjusted R-Squared of 0.99 with a p-value<0.01, whereas the calibration model for Sucrose obtained an R-squared of 0.99, an Adjusted R-Squared of 0.99 with a p-value<0.01. For both models, it was determined that there was a highly statistically significant correlation between the actual and the predicted values, strengthening the importance of our findings.

4. Discussion

The results of this study were compared to those published by Tange *et al*. [24], who used SVR to carry out the model calibration for °Brix and Sucrose in the manufacturing process of sugar cane. The results presented herein obtained lower *RMSE* values than Tange *et al*. [24], who presented more accurate estimates of the quality parameters. This is explained by the fact that in the present work techniques of preprocessing and selection of features were used, and therefore noise spectra were removed, along with the wavelengths which did not contribute significantly to the model.

Rinnan *et al*. [32] argued that performing various preprocessing stages was not advisable in NIR spectra; however, the results of this study showed an improvement with the preprocessing technique number 9, which consisted of calculating the first spectral derivative, then apply SNV and finally extract the trend. Other recent studies on NIR spectroscopy have combined several preprocessing techniques with very good results [33,36–39]. Xu *et al*. [33] stated that by combining preprocessing techniques, the model takes advantage of the complementary information given by each preprocessing method, therefore, the stability of the models and the results are improved in terms of *RMSE*.

In the proposed model, the T-test feature selection technique allowed recognizing the most relevant wavelengths, removing noise generated by the other wavelengths and obtaining more accurate estimates, which is consistent with other research studies in which T-test feature selection was used to remove the irrelevant variables, proving its efficacy in obtaining more accurate results [42,76–79].

The optimization of parameters of SVR proved its importance to obtain minimum *RMSE* for the model. In our case, it was performed by using a search grid technique, providing the optimal combination of parameters C, γ and ε, which is consistent with the results obtained by Jeng [66] and by Devos *et al*. [65] who claimed that the combined values of the parameters of SVM determined the complexity of the limits and therefore the performance of the model.

In this regard, one can deduce that the proposed model is accurate and stable, due to the parameter optimization, which is consistent with the results obtained by Cristianini and Shawe [64] and Devos *et al*. [65] who stated that the adjustment of the SVM kernel parameters controlled the complexity of the resulting hypothesis and avoided the overfitting of the model.

The evaluation of the models was performed using the repeated cross-validation technique, which, according to Garcia and Filzmoser [68] leads to a suitable method aimed at choosing the best model to analyze the mean and standard deviation of the results of repetitions; these results correspond to the test data set, that is, they are data which were not used for calibration, which allows estimating how the model would behave in the future with new data [9,67].

Table 1 shows the results of the proposed global model for °Brix compared to those reported by Tange *et al*. [24]; note that the proposed global model has a *RMSE* of 0.305, whereas the previously published global model reaches 0.59, showing an optimization of the model. The

same applies to the four steps of the process, in which the proposed global model improves the previously published global model, and also the four published local (individual) models.

Table 1: Comparative results of *RMSE* of the models referred* by Tange et al. [24] with an optimized prediction model for °Brix proposed by the authors.

Model	Juice	Syrup	Massecuite	Molasses	Global
Reference *	0.1	0.2	0.5	0.5	0.5
Local SVM *	0.08	0.22	0.39	0.75	-
Global SVM *	0.16	0.25	0.47	0.79	0.59
Optimized SVM	0.040±0.018	0.084±0.063	01702±0.073	0.576±0.183	0.305±0.076

Table 2 shows the results of the proposed global model for Sucrose, with two alternative values that ε may take, Optimized SVM 1 with a value of 0.07 and Optimized SVM 2 with a value of 0.51, which are compared to those published by Tange *et al*. [24]. The proposed global model has a *RMSE* of 0.486 (Optimized SVM 1) and 0.485 (Optimized SVM 2), whereas the published global model reaches 0.64, showing an optimization of the model with both values of ε.

The proposed model for Sucrose exceeds the one published in three out of the four steps of the process; in the process step of Molasses, the proposed model is similar to the one published, however it does not exceed it; this is due to the fact that in the set of Molasses spectra, values with absorbance over two were found, which, according to Tange *et al*., [24] in their publication were removed from the original dataset. However, in the opinion of the authors, these spectra are not considered outliers, thus in the current study they were maintained to provide more robustness to the model.

Table 2: Comparative results of *RMSE* of the models referred* by Tange *et al*. [24] with the optimized prediction models for Sucrose proposed by the authors.

Model	Juice	Syrup	Massecuite	Molasses	Global
Reference *	0.1	0.2	0.5	0.5	0.5
Local SVM *	0.11	0.22	0.56	0.62	-
Global SVM *	0.20	0.24	0.72	0.72	0.64
Optimized SVM	0.016±0.008	0.062±0.023	0.369±0.161	0.858±1.586	0.485±0.631
Optimized SVM 2	0.108±0.036	0.094±0.035	0.431±0.180	0.765±1.396	0.486±0.559

5. Conclusions

This study evaluates the application of feature selection techniques and the determination of the optimal configuration of the parameters of a chemometric calibration model based on support vector regression, a technique commonly used in machine learning. Compared to the published models, the models proposed herein were able to better estimate the non-linearities caused by the combination of the NIR spectra from multiple stages of the manufacturing process of sugar cane.

The proposed models for Brix and Sucrose were improved compared to those published by Tange *et al*. [24] in the four steps of the process, except for the prediction of Sucrose in molasses, which is similar to the published model, although in the present work no spectrum was separated from the original dataset.

Calculating the first spectral derivative from the raw signal, performing SNV and finally extracting the trend was the best combination of preprocessing techniques for the case study. Its implementation improved the stability of the models and the results in terms of *RMSE* by taking supplementary information from each individual technique.

Feature selection reduced the number of wavelengths selected for the calibration of the models, which simplifies the final model, with the corresponding calculation reduction needed to estimate the quality parameters in the sugar industry.

The use of global models with a lower *RMSE* allows a better estimate of the quality parameters, with a single calibration process and, therefore, a simpler and more effective monitoring of the process in the sugar industry. Relying on a methodology that allows a more accurate quality control paves the way for the detection of substances which are found in lower concentrations using NIR spectroscopy.

6. Future developments

Future work focuses on the application of these feature selection techniques such as multivariate filters and wrappers, in addition to the identification of redundant wavelengths by means of analysis of covariance to simplify the models. Other algorithms, such as decision tree regression, artificial neural networks, optimized PLS, and knn regression, can be studied comparatively in order to determine their performance measures. Additionally, authors consider it is important to study the origin of noises are generated in raw spectra.

The proposed methodology can be applied to other models in the food industry, such as ingredients for feed, whose chemical composition could be determined by applying a similar process. Additionally, it could be applied to other signals, such as voltammetry, Raman spectroscopy or electrical impedance, whereby employing a supervised machine learning approach, application models beneficial to the industry can be calibrated.

7. Acknowledgements

The authors would like to express their gratitude to Professor Ramus Bro and the University of Copenhagen, who kindly provided the spectral dataset for this research study. They also gratefully acknowledge the support from CEDIA National Research and Education Network, and CESGA Supercomputing Center. This work is part of DINTA-UTMACH and RNASA-UDC research groups. Our special thanks to the three anonymous reviewers whose suggestions helped to improve and clarify this manuscript.

8. References

[1] L.S. Polanco, D.F. Day, S. Savoie, S. Bergeron, T. Charlet, B.L. Legendre, Improvements of raw sugar quality using double purge of c-massecuites performance comparison, in: LSU AgCenter Audubon Sugar Institute Factory Operations Seminar, lsuagcenter.com, 2014: p. 46.

[2] C. Kumaravelu, A. Gopal, A review on the applications of Near-Infrared spectrometer and Chemometrics for the agro-food processing industries, (2015) 8–12.

[3] B.R. Kowalski, Chemometrics, Anal. Chem. 52 (1980) 112R–122R.

[4] E. Bertran, M. Blanco, S. Maspoch, M.C. Ortiz, M.S. Sánchez, L.A. Sarabia, Handling intrinsic non-linearity in near-infrared reflectance spectroscopy, Chemometrics Intellig. Lab. Syst. 49 (1999) 215–224.

[5] H. Martens, T. Naes, Multivariate Calibration, Wiley, 1992.
[6] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemom. 6 (1992) 267–281.
[7] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics. 23 (2007) 2507–2517.
[8] H. Cen, Y. He, Theory and application of near infrared reflectance spectroscopy in determination of food quality, Trends Food Sci. Technol. 18 (2007/2) 72–83.
[9] J.B.O. Mitchell, Machine learning methods in chemoinformatics, Wiley Interdiscip. Rev. Comput. Mol. Sci. 4 (2014) 468–481.
[10] P. Torrione, L.M. Collins, K.D. Morton, Multivariate analysis, chemometrics, and machine learning in laser spectroscopy.
[11] R.G. Brereton, Pattern recognition in chemometrics, Chemometrics Intellig. Lab. Syst. 149, Part B (2015) 90–96.
[12] M.M. Tomazzoli, R.D. Pai Neto, R. Moresco, L. Westphal, A.R.S. Zeggio, L. Specht, C. Costa, M. Rocha, M. Maraschin, Discrimination of Brazilian propolis according to the seasoning using chemometrics and machine learning based on UV-Vis scanning data, J. Integr. Bioinform. 12 (2015) 279.
[13] M. Tajammal Munir, W. Yu, B.R. Young, D.I. Wilson, The current status of process analytical technologies in the dairy industry, Trends Food Sci. Technol. 43 (2015/6) 205–218.
[14] L. Wang, D.-W. Sun, H. Pu, J.-H. Cheng, Quality Analysis and Classification and Authentication of Liquid Foods by Near-infrared Spectroscopy: A Review of Recent Research Developments, Crit. Rev. Food Sci. Nutr. (2016) 0.
[15] E. Zamora-Rojas, D. Pérez-Marín, E. De Pedro-Sanz, J.E. Guerrero-Ginel, A. Garrido-Varo, Handheld NIRS analysis for routine meat quality control: Database transfer from at-line instruments, Chemometrics Intellig. Lab. Syst. 114 (2012) 30–35.
[16] H.-J. He, D. Wu, D.-W. Sun, Rapid and non-destructive determination of drip loss and pH distribution in farmed Atlantic salmon (Salmo salar) fillets using visible and near-infrared (Vis–NIR) hyperspectral imaging, Food Chem. 156 (2014) 394–401.
[17] R. Henry, P. Kettlewell, Cereal Grain Quality, Springer Netherlands, 2012.
[18] L.S. Magwaza, U.L. Opara, H. Nieuwoudt, P.J.R. Cronje, W. Saeys, B. Nicolaï, NIR Spectroscopy Applications for Internal and External Quality Analysis of Citrus Fruit—A Review, Food Bioprocess Technol. 5 (2011) 425–444.
[19] P. Valderrama, J.W.B. Braga, R.J. Poppi, Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy, J. Braz. Chem. Soc. 18 (2007) 259–266.
[20] E.P. Zayas-Ruiz, M. Lorenzo-Izquierdo, F.O. Fragoso-Concepción, La quimiometría y la industria del azúcar y sus derivados, ICIDCA. Sobre Los Derivados de La Caña de Azúcar. 49 (2015) 31–33.
[21] N. Sorol, E. Arancibia, S.A. Bortolato, A.C. Olivieri, Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice: A test field for variable selection methods, Chemometrics Intellig. Lab. Syst. 102 (2010) 100–109.
[22] X. Wang, H.-J. Ye, Q.-T. Li, J.-C. Xie, J.-J. Lu, A.-L. Xia, J. Wang, Determination of Brix and POL in Sugar Cane Juice by Using Near Infrared Spectroscopy Coupled with BP-ANN, Spectroscopy and Spectral Analysis. 30 (2010) 1759–1762.
[23] H.Z. Chen, J.B. Wen, J.C. Chen, L.H. Li, Y.J. Huo, Near-infrared spectroscopic modeling optimization for quantitative determination of sugar brix in sugarcane initial-pressure juice, Int. J. Tech. Res. Applic. 2 (2014) 6.
[24] R.I. Tange, M.A. Rasmussen, E. Taira, R. Bro, Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps, J. Near Infrared Spectrosc. (2015). http://www.forskningsdatabasen.dk/en/catalog/2266200601.
[25] C.D. Brown, P.D. Wentzell, Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration, J. Chemom. 13 (1999) 133–152.
[26] E. Stark, Near infrared spectroscopy past and future, Near Infrared Spectroscopy The Future Waves. (1996) 701–713.
[27] A.M. Rady, D.E. Guyer, Evaluation of sugar content in potatoes using NIR reflectance and wavelength selection techniques, Postharvest Biol. Technol. 103 (2015/5) 17–26.
[28] C.A.T. dos Santos, M. Lopo, R.N.M.J. Páscoa, J.A. Lopes, A review on the applications of portable near-infrared spectrometers in the agro-food industry, Appl. Spectrosc. 67 (2013) 1215–1233.

[29] E. Teye, X.-Y. Huang, N. Afoakwa, Review on the potential use of near infrared spectroscopy (NIRS) for the measurement of chemical residues in food, American Journal of Food Science and Technology. 1 (2013) 1–8.
[30] M. Blanco, I. Villarroya, NIR spectroscopy: a rapid-response analytical tool, Trends Analyt. Chem. 21 (2002/4) 240–250.
[31] W.J. Florkowski, S.E. Prussia, R.L. Shewfelt, B. Brueckner, Postharvest Handling: A Systems Approach, Elsevier Science, 2009.
[32] Å. Rinnan, F. van D. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trends Analyt. Chem. 28 (2009) 1201–1222.
[33] L. Xu, Y.-P. Zhou, L.-J. Tang, H.-L. Wu, J.-H. Jiang, G.-L. Shen, R.-Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta. 616 (2008) 138–143.
[34] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, Appl. Spectrosc., AS. 43 (1989) 772–777.
[35] J. Luypaert, S. Heuerding, S. de Jong, D.L. Massart, An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream, J. Pharm. Biomed. Anal. 30 (2002) 453–466.
[36] M.J. Martelo-Vidal, M. Vázquez, Evaluation of ultraviolet, visible, and near infrared spectroscopy for the analysis of wine compounds, Czech J. Food Sci. 32 (2014) 37.
[37] L. Xie, X. He, B. Duan, S. Tang, J. Luo, G. Jiao, G. Shao, X. Wei, Z. Sheng, P. Hu, Optimization of Near-Infrared Reflectance Model in Measuring Gelatinization Characteristics of Rice Flour with a Rapid Viscosity Analyzer (RVA) and Differential Scanning Calorimeter (DSC), Cereal Chem. 92 (2015) 522–528.
[38] X. Pan, Y. Li, Z. Wu, Q. Zhang, Z. Zheng, X. Shi, Y. Qiao, A online NIR sensor for the pilot-scale extraction process in Fructus aurantii coupled with single and ensemble methods, Sensors . 15 (2015) 8749–8763.
[39] G.M. Hadad, A.S. Ra, M.M. Elkhoudarya, Simultaneous Determination of Clarithromycin, Tinidazole and Omeprazole in Helicure Tablets Using Reflectance Near-Infrared Spectroscopy with the Aid of Chemometry, Pharmaceutica Analytica Acta 2015.
[40] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[41] S. Keleş, M. van der Laan, M.B. Eisen, Identification of regulatory elements using a feature selection method, Bioinformatics. 18 (2002) 1167–1175.
[42] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Springer Berlin Heidelberg, 2008.
[43] E. Szymańska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet, L.M.C. Buydens, Chemometrics and qualitative analysis have a vibrant relationship, Trends Analyt. Chem. 69 (2015/6) 34–51.
[44] I. Guyon, A. Elisseeff, An Introduction to Feature Extraction, in: I. Guyon, M. Nikravesh, S. Gunn, L.A. Zadeh (Eds.), Feature Extraction, Springer Berlin Heidelberg, 2006: pp. 1–25.
[45] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer US, 2012.
[46] P. Jafari, F. Azuaje, An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors, BMC Med. Inform. Decis. Mak. 6 (2006) 27.
[47] G. Bhanot, G. Alexe, B. Venkataraghavan, A.J. Levine, A robust meta-classification strategy for cancer detection from MS data, Proteomics. 6 (2006) 592–604.
[48] A. Mucherino, G. Ruß, Recent Developments in Data Mining and Agriculture, in: Industrial Conference on Data Mining- …, 2011: pp. 1–9.
[49] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer New York, New York, NY, 2009.
[50] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, Data Mining in Agriculture, Springer New York, New York, NY, 2009.
[51] D. Basak, S. Pal, D.C. Patranabis, Support vector regression, Neural Information Processing-Letters and Reviews. 11 (2007) 203–224.
[52] J. Palma, R. Marín, Inteligencia artificial. Técnicas, métodos y aplicaciones, McGraw Hill, Murcia, 2013.
[53] B.E. Boser, I.M. Guyon, V.N. Vapnik, A Training Algorithm for Optimal Margin Classifiers, Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. (1992) 144–152.
[54] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[55] V. Vapnik, S.E. Golowich, A.J. Smola, Support Vector Method for Function Approximation, Regression Estimation and Signal Processing, in: M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems 9, MIT Press, 1997: pp. 281–287.
[56] L. Zhang, W.-D. Zhou, P.-C. Chang, J.-W. Yang, F.-Z. Li, Iterated time series prediction with multiple support vector regression models, Neurocomputing. 99 (2013) 411–422.
[57] J. Wu, J. Wei, Combining ICA with SVR for prediction of finance time series, in: 2007 IEEE International Conference on Automation and Logistics, ieeexplore.ieee.org, 2007: pp. 95–100.
[58] U.A. Acar, B. Hudson, G.L. Miller, T. Phillips, SVR: Practical Engineering of a Fast 3D Meshing Algorithm*, in: M.L. Brewer, D. Marcum (Eds.), Proceedings of the 16th International Meshing Roundtable, Springer Berlin Heidelberg, 2008: pp. 45–62.
[59] Y. Quan, J. Yang, L.-X. Yao, C.-Z. Ye, Successive overrelaxation for support vector regression, J. Softw. Maint. Evol.: Res. Pract. 15 (2004) 200–206.
[60] P. Koch, B. Bischl, O. Flasch, T. Bartz-Beielstein, C. Weihs, W. Konen, Tuning and evolution of support vector kernels, Evol. Intell. 5 (2012) 153–170.
[61] K. Mollazade, M. Omid, A. Arefi, Comparing data mining classifiers for grading raisins based on visual features, Comput. Electron. Agric. 84 (2012) 124–131.
[62] C.-H. Wu, G.-H. Tzeng, R.-H. Lin, A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression, Expert Syst. Appl. 36 (2009/4) 4725–4735.
[63] R.K. Prasoona, A. Jyoti, Y. Mukesh, S. Nishant, N.S. Anuraj, J. Shobha, Optimization of Gaussian Kernel Function in Support Vector Machine aided QSAR studies of C-aryl glucoside SGLT2 inhibitors, Interdiscip. Sci. 5 (2013) 45–52.
[64] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines (Cambridge, (2000).
[65] O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, J.-P. Huvenne, Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation, Chemometrics Intellig. Lab. Syst. 96 (2009) 27–33.
[66] J.-T. Jeng, Hybrid approach of selecting hyperparameters of support vector machine for regression, IEEE Trans. Syst. Man Cybern. B Cybern. 36 (2006) 699–709.
[67] M. Kuhn, K. Johnson, Applied Predictive Modeling:, Springer New York, 2013.
[68] H. Garcia, P. Filzmoser, Multivariate Statistical Analysis using the R package chemometrics, Vienna, Austria. (2015). ftp://155.232.191.229/cran/web/packages/chemometrics/vignettes/chemometrics-vignette.pdf.
[69] S.R. Gunn, Support vector machines for classification and regression, ISIS Technical Report. 14 (1998). http://ce.sharif.ir/courses/85-86/2/ce725/resources/root/LECTURES/SVM.pdf.
[70] O. Devos, L. Duponchel, Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression, Chemometrics Intellig. Lab. Syst. 107 (2011) 50–58.
[71] F. Allegrini, A.C. Olivieri, An integrated approach to the simultaneous selection of variables, mathematical pre-processing and calibration samples in partial least-squares multivariate calibration, Talanta. 115 (2013) 755–760.
[72] X. Ma, Y. Zhang, Y. Wang, Performance evaluation of kernel functions based on grid search for support vector regression, in: 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), ieeexplore.ieee.org, 2015: pp. 283–288.
[73] R.A. Viscarra Rossel, ParLeS: Software for chemometric analysis of spectroscopic data, Chemometrics Intellig. Lab. Syst. 90 (2008) 72–83.
[74] J.S. Armstrong, F. Collopy, Error measures for generalizing about forecasting methods: Empirical comparisons, Int. J. Forecast. 8 (1992) 69–80.
[75] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688.
[76] C. Christin, H.C.J. Hoefsloot, A.K. Smilde, B. Hoekman, F. Suits, R. Bischoff, P. Horvatovich, A critical assessment of feature selection methods for biomarker discovery in clinical proteomics, Mol. Cell. Proteomics. 12 (2013) 263–276.
[77] N. Erho, A. Crisan, I.A. Vergara, A.P. Mitra, M. Ghadessi, C. Buerki, E.J. Bergstralh, T. Kollmeyer, S. Fink, Z. Haddad, B. Zimmermann, T. Sierocinski, K.V. Ballman, T.J. Triche, P.C. Black, R.J. Karnes, G. Klee, E. Davicioni, R.B. Jenkins, Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy, PLoS One. 8 (2013) e66855.
[78] D. Zhu, K. Li, D.P. Terry, A.N. Puente, L. Wang, D. Shen, L.S. Miller, T. Liu, Connectome-scale assessments of structural and functional connectivity in MCI, Hum. Brain Mapp. 35 (2014) 2911–

2923.
[79] N.-F. Jie, M.-H. Zhu, X.-Y. Ma, E.A. Osuch, M. Wammes, J. Théberge, H.-D. Li, Y. Zhang, T.-Z. Jiang, J. Sui, V.D. Calhoun, Discriminating Bipolar Disorder From Major Depression Based on SVM-FoBa: Efficient Feature Selection With Multimodal Brain Imaging Data, IEEE Trans. Auton. Ment. Dev. 7 (2015) 320–331.

Highlights

Single calibration models for multiple process in the sugar industry were developed.

Feature selection techniques and Support Vector Regression were used to develop the models.

A total of 1797 NIR spectra ranging between 400.0 nm and 1888 nm, were analyzed.

The proposed models for Brix and Sucrose performed better in test set data compared to those published.

A R-squared of 0.99 and a RMSE of 0.305 were achieved for °Brix model. A R-squared of 0.99, RMSE of 0.486 for were achieved for Sucrose model.