

## ORIGINAL ARTICLE

## Integrated Food Science

# Classification and quantification of sucrose from sugar beet and sugarcane using optical spectroscopy and chemometrics

Hilmi Eriklioglu<sup>1</sup> | Esmanur Ilhan<sup>1</sup> | Mikhail Khodasevich<sup>2</sup> | Darya Korolko<sup>2</sup> |  
Marena Manley<sup>3</sup>  | Rosario Castillo<sup>4</sup> | Mecit Halil Oztop<sup>1</sup> 

<sup>1</sup>Department of Food Engineering, Middle East Technical University, Ankara, Turkey

<sup>2</sup>B.I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus, Minsk, Belarus

<sup>3</sup>Department of Food Science, Stellenbosch University, Stellenbosch, Western Cape, South Africa

<sup>4</sup>Biotechnology Center and Faculty of Pharmacy Concepción, University of Concepción, Concepción, Chile

**Correspondence**

Marena Manley, Department of Food Science, Stellenbosch University, Stellenbosch, Western Cape, South Africa.

Email: [mman@sun.ac.za](mailto:mman@sun.ac.za)

Mecit Halil Oztop, Department of Food Engineering, Middle East Technical University, Ankara, Turkey.

Email: [mecit@metu.edu.tr](mailto:mecit@metu.edu.tr)

**Funding information**

This paper has been part of the H2020 MSCA RISE Consortium “SuChAQuality.” The project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement 101008228.

**Abstract**

Sucrose, obtained from either sugar beet or sugarcane, is one of the main ingredients used in the food industry. Due to the same molecular structure, chemical methods cannot distinguish sucrose from both sources. More practical and affordable methods would be valuable. Sucrose samples (cane and beet) were collected from nine countries, 25% (w/w) aqueous solutions were prepared and their absorbances recorded from 200 to 1380 nm. Spectral differences were observable in the ultraviolet–visible (UV–Vis) region from 200 to 600 nm due to impurities in sugar. Linear discriminant analysis (LDA), classification and regression trees, and soft independent modeling of class analogy were tested for the UV–Vis region. All methods showed high performance accuracies. LDA, after selection of five wavelengths, gave 100% correct classification with a simple interpretation. In addition, binary mixtures of the sugar samples were prepared for quantitative analysis by means of partial least squares regression and multiple linear regression (MLR). MLR with first derivative Savitzky–Golay were most acceptable with root mean square error of cross-validation, prediction, and the ratio of (standard error of) prediction to (standard) deviation values of 3.92%, 3.28%, and 9.46, respectively. Using UV–Vis spectra and chemometrics, the results show promise to distinguish between the two different sources of sucrose. An affordable and quick analysis method to differentiate between sugars, produced from either sugar beet or sugarcane, is suggested. This method does not involve complex chemical analysis or high-level experts and can be used in research or by industry to detect the source of the sugar which is important for some countries’ agricultural policies.

**KEYWORDS**

multivariate data analysis, sucrose, sugar beet, sugarcane, UV–Vis spectroscopy

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Journal of Food Science* published by Wiley Periodicals LLC on behalf of Institute of Food Technologists.

## 1 | INTRODUCTION

Sugar is one of the crucial components of the confectionery industry that provides unique multifunctional properties, such as fermentation substrate, structure, texture, color, taste, and preservation. Sucrose can be extracted from both sugar beet and sugarcane. These sources of sucrose are most affected by geographical location (Sefaoglu et al., 2016). Although sugarcane (*Saccharum officinarum*) grows in a tropical climate zone, sugar beet (*Beta vulgaris saccharifera* L.) grows in different climate zones and specifically regions located between 60° north and 30° south latitudes. The price of beet and cane sugar depends on season, year and country of production and countries from which sugar is imported. Two to four times more sucrose can be obtained from one decare (100 m<sup>2</sup>) of sugarcane compared to sugar beet. Due to this higher yield, the raw material and production cost of cane sugar are much lower than that of beet sugar. This makes importing cane profitable for some countries in the short term.

The climatic conditions prevent the cultivation of sugarcane in some countries. Turkey, Russia, Ukraine, Belarus, and the EU produce sugar from sugar beet, whereas countries, such as USA, Japan, and China, produce sugar from both sources. Brazil, India, Thailand, Mexico, Pakistan, Australia, and several other countries in the middle latitudes produce sugar from sugarcane (Thow et al., 2021). Also, in some countries, for example, Turkey, there are governmental policies and regulations that require sugar production from sugar beet and limits the amount of sugar to be produced (Sugar Act, 2001). It might not be economical to maintain sugar production using sugar beet, but there are other factors to consider such as foreign dependency, providing employment opportunities to local factory workers and income opportunities for farmers. Due to the requirements by policies and regulations, it is important to know the source of the sugar present in the market. In addition, sugar beet producing countries may consider importing cane sugar rather than producing beet sugar due to strategic reasons. Conflicts among countries, war, sanctions, and foreign dependency may become important and as local people are often providing their families with beet cane crops, it is important to prevent illegal imports of cane sugar. The need therefore exists for a rapid, easy to perform, and economical method to distinguish between sugars produced from either sugar beet or sugarcane.

As beet sugar and cane sugar are completely sucrose-based, it can be difficult and expensive to detect the source of the sugar using chemical (Bubnik et al., 1995) and/or sensory methods (Urbanus et al., 2014). Both these methods are expensive, require the use of chemicals, extensive labor, or experts such analytical chemists or trained sensory panelists. When long term and continuous analysis

are considered, development of an alternative method is important. Even though beet sugar and cane sugar are both essentially sucrose, there are some detectable differences depending on the source and manufacturing process. The total amount of polysaccharides present in sugarcane was reported as 169 ppm (solids), whereas 77 ppm was reported for sugar beet samples. The initial amounts were 8238 and 4067 ppm for cane and beet raw juices, respectively (Godshall et al., 2002). Similarly, sugarcane juice contained 5% (w/w) non-sugar compounds, whereas sugar beet juice contained 2.5% (w/w). Finally, the presence of fibers was reported as 5% (w/w) and 10% for sugar beet and -cane, respectively (Asadi, 2007). Other markers that can be used for sugar beet and -cane differentiation is the presence of raffinose and theandrose. Theandrose is only present in sugarcane and is considered a natural constituent (Moreldu Boil, 1996). Raffinose is present in both sugar beet and -cane; however, raffinose levels are higher in sugar beet sugar compared to -cane (Morel du Boil, 1997; Vaccari & Mantovani, 1995). Despite the difference in micro-impurities in sugars, the purpose of this work is not to determine the composition and concentrations of these impurities, but to classify beet and cane sugars while considering the diversity of their plant sources and the features of production technologies.

Optical spectroscopy has been successfully used as an analytical tool that can largely be attributed to its ability to provide rapid qualitative and quantitative analysis of multicomponents in single samples (Manley, 2014). It can efficiently provide access to various physical, chemical, and structural properties such as particle size, protein, starch, moisture, fat, ash content, soluble solids, or acidity in food samples if calibrated appropriately (Bahrami et al., 2020). Lately, ultraviolet-visible (UV-Vis) spectra, which correspond to the wavelengths 200–800 nm in the electromagnetic spectrum, have gained increasing interest among food scientists. It has been utilized for food analysis purposes because of its easy application, relatively low equipment costs, and minimum sample preparation requirements. Methods based on UV spectra have been used for authentication of food materials and to detect adulteration in several studies (Boggia et al., 2017; Dankowska & Kowalewski, 2019; Fanelli et al., 2021). Moreover, implementation of UV spectroscopy for routine analysis is also possible (Suhandy & Yulia, 2021). Therefore, optical spectroscopy appears to be a feasible analytical tool to investigate beet and cane sugar identification.

Optical spectroscopic methods of analyses produce spectra between 190 and 1100 nm and provide large amounts of data and information. The obtained data can be exploited using chemometric techniques, for example, principal component analysis (PCA) which is an unsupervised pattern recognition technique can be used as a first step for

explorative data analysis, outlier detection, graphical clustering, and classification (Cortés et al., 2019; Esbensen & Geladi, 2009). Depending on the objectives of the study, qualitative and/or quantitative data analysis approaches are selected. Using pattern recognition techniques, classifying samples based on their spectra is possible (Roggo et al., 2007). A training set with known categories is used to create a classification model, which is then tested on a test set of unknown samples. In this study, several qualitative and quantitative techniques were used, that is, linear discriminant analysis (LDA) (Baranowski et al., 2012), classification and regression trees (CART) (Barbosa et al., 2014), and soft independent modeling of class analogy (SIMCA) (Souto et al., 2015). The models' performance is usually evaluated by means of sensitivity, specificity, precision, and accuracy. On the other hand, partial least squares regression (PLSR) and multiple linear regression (MLR) (Dankowska et al., 2017) were used for quantitative analysis. The performance of quantitative models is usually evaluated by means of the root mean square error of cross-validation (RMSECV), root mean square error of prediction (RMSEP), and the ratio of (standard error of) prediction to (standard) deviation (RPD) (Williams, 2014) values as well as the coefficient of determination ( $R^2$ ). Generally, a good model should achieve a low root mean square error (RMSE) and a high  $R^2$ . Additionally, a satisfactory model should have an RPD value of more than 3.1, a value above 6.5 being very good (Williams, 2014). This study aimed to investigate the ability of optical spectroscopy in association with chemometrics to differentiate between sucrose samples produced from either sugar beet or sugarcane.

## 2 | MATERIALS AND METHODS

### 2.1 | Materials and chemicals

Different sucrose samples (comprising 23 different brands) originating from sugarcane and sugar beet plants were collected from 9 different countries, including Pakistan, Portugal, Poland, Romania, Italy, Serbia, Belarus, Ukraine, and Colombia. For classification purposes, only known source samples and white sugars were used. Brown sugars were not included since their absorbances were not comparable with that of white sugars.

### 2.2 | Preparation of sucrose solutions for spectral analysis

For qualification, from every sucrose bag, two or three different 25% (w/w) sucrose solutions were prepared with

deionized water and five replicates from each were taken for further analysis ( $n = 235$ ). A 25% aqueous solution was selected to match the dynamic range of the spectrophotometer used.

For quantification, selected sucrose samples originating from beet and cane sugar were mixed at different concentrations from 0% (w/w) beet sugar to 100% (w/w) beet sugar with 5% (w/w) increments. In total, 21 samples were used to obtain binary mixtures that had a final concentration of 25% (w/w) sucrose. All samples were mixed properly before adding water because spectroscopic analyses require homogenous samples. After the addition of water, the samples were stirred for about 10 min in glass beakers and then placed and scanned in quartz cuvettes (10 mm path length).

### 2.3 | Spectral data acquisition

Absorbance data were recorded with a UV-Vis-NIR scanning spectrophotometer UV-3101 PC (Shimadzu, Inc., Nakagyo-ku, Kyoto 604-8511, Japan) that covered a spectral range from 190 to 3200 nm. For the qualification of beet and cane sucrose samples, wavelengths ranging from 200 to 1300 nm were used with 1 nm spectral interval. All measurements were conducted at a slow scan speed, with 1 nm spectral slit width and 1 nm spectral interval. Each measurement was made after the spectrum of an air reference was obtained. Spectra were collected in 10 mm path length quartz cuvettes, and each cuvette was measured only once.

For quantification, the wavelength range was 200–600 nm and the spectra recorded with the same spectral settings as for the qualitative analysis. However, for quantification a cuvette filled with distilled water was placed in the second beam of the two-beam spectrophotometer as the reference to remove the effect of the solvent.

### 2.4 | Multivariate data analysis

#### 2.4.1 | Preprocessing, exploratory analysis, and selection of training and test sets

During quantitative model development, different preprocessing methods, such as mean centering, standard normal variate (SNV), normalization, Gaussian smoothing, and Savitzky–Golay first (1st Der) and second (2nd Der) derivatives, were tested. The UV-Vis raw spectra were preprocessed with the different methods to remove unwanted variation and artifacts that could be present in the data. PCA was mainly used for exploratory purposes to interpret and visualize the differences between the

samples in the multivariate space, with only mean centering applied.

Furthermore, the samples were divided into training and test sets for both qualitative and quantitative model development. Selections of samples were performed manually by considering representativeness and leverages. Subsequently, the brown sugar and all but two powdered sugars were excluded from the dataset. In total, 124 spectra (50 cane, 74 beet sugar) were used for classification, with 85 spectra in the training and 39 spectra in the test set, replicates of each sample were always kept together. For all classification chemometric analysis, fivefold cross-validation (CV) was applied to determine model abilities followed by validation with the independent test set. For quantification, leave-one-out CV was performed on the calibration set ( $n = 15$ ) to determine model parameters followed by independent validation ( $n = 6$ ). The preprocessing and data analysis performed are shown in a chemometric analysis flowchart in Figure 1.

#### 2.4.2 | Qualitative analysis

For classification purposes, the performance of LDA, CART, and SIMCA was investigated by means of sensitivity, specificity, precision, and accuracy.

*Linear discriminant analysis (LDA)*: LDA looks for linear combinations of variables which best explain the difference between the classes of data. Because LDA works well with a specific ratio between sample number and variable number, first, PCA was applied to decrease the number of factors. This was followed by selection of five wavelengths based on maximum classification performance.

*Classification and regression tree (CART)*: Decision trees provide structural mapping that consists of binary selection (Kotsiantis, 2013). By selecting the variables from numerous input data, algorithm grows treelike shapes with root nodes. Any root that is added to the algorithm is based on an appointed value for one variable, also called univariate split. These splits are basically threshold values selected from variables, which are used to differentiate between samples. The main aim of the algorithm is to improve the model performance by adding one split with the least split numbers possible. For this study, only two roots were applied on PC1 and PC2 scores.

*Soft independent modeling of class analogy (SIMCA)*: SIMCA is a method that works with PCA. As SIMCA operates by applying PCA to the classes separately, this approach gives more information related to the classes with reference to separation measures and relation

between different variables (van den Branden & Hubert, 2005). While training each class specific model, residual distributions of classes are generated. With class specific distribution, according to their probabilities, observations are assigned to the mentioned classes. The algorithm tested different number of PCs and shows which one gave the maximum sensitivity, specificity, and minimum error.

#### 2.4.3 | Quantitative analysis

*Partial least squares regression (PLSR)*: Similar to PCA, PLSR also works with components but in this case, they are called latent variables/factors. Also, in PLSR, while applying regression for the chosen dataset,  $X$  matrix decomposition is guided by the variance in the  $y$  vector which are target values. Thus, the main purpose is to increase the covariation between  $y$  and  $X$ . The basic linear model can be shown as

$$y = Xb + e \quad (1)$$

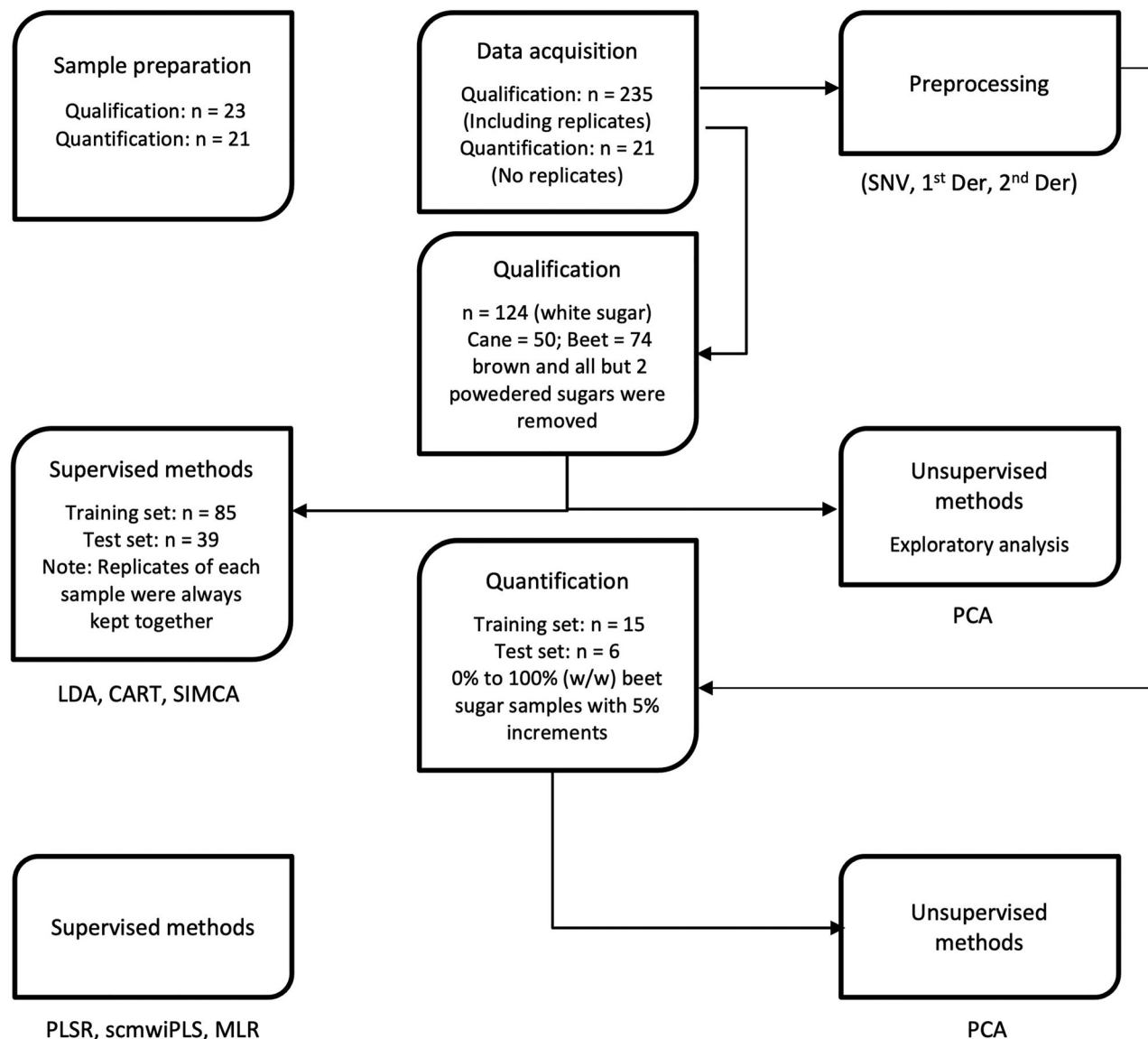
where  $e$  is the residual, and  $b$  is the vector containing coefficients of regression obtained after model calibration. For this study, different numbers of latent variables (LV) were evaluated to determine the least number of LVs that would maximize the models' ability.

While performing PLSR, the number of LV which cover the adequate variance of data was kept as low as possible to achieve high performing predictions. To calculate the performance of the models, the dataset was split into two subgroups, that is, training and test sets. Preprocessing methods, spectral regions, and number of factors were selected by considering minimum RMSECV on calibration dataset. First, a simple moving window variable selection was performed manually by considering loading values then the following method was applied for further selection.

*Searching combination moving window interval' PLS (scmwiPLS)*: Variable selection was done, to eliminate unnecessary variables, using scmwiPLS (Du et al., 2004). In this method, after selecting the size of the windows (number of variables in one window), the algorithm determines the best combination of windows with the lowest RMSE. Furthermore, different numbers of windows were compared and selected according to their performance result, that is, lowest RMSE. In this study, scmwiPLS was used to find the combination of informative bands to increase prediction capability of the PLS model.

*Multiple linear regression (MLR)*: MLR was used to build a quantification model. Ridge regularization was





**FIGURE 1** Chemometric analysis flowchart. 1st Der, first derivative; 2nd Der, second derivative; CART, classification, and regression trees; LDA, linear discriminant analysis; MLR, multiple linear regression; PCA, principal component analysis; PLSR, partial least squares regression; scmwPLS, searching combination moving window interval PLS; SIMCA, soft independent modeling of class analogy; SNV, standard normal variate.

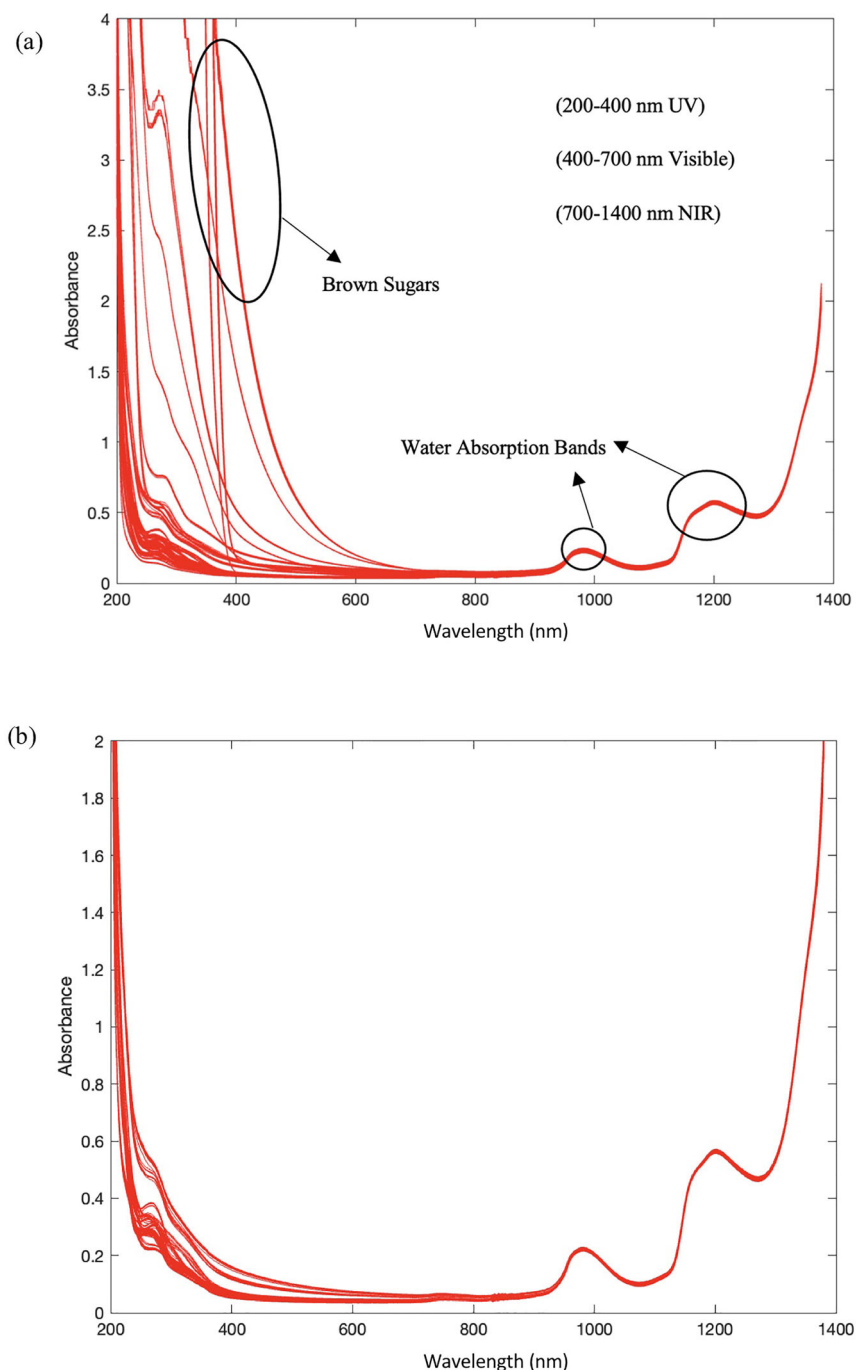
**TABLE 1** Classification results for the training set with fivefold cross-validation and prediction results of the independent test set.

Training set	Sensitivity	Specificity	Precision	Error rate	Accuracy
Cane	0.98	0.97	0.96	0.02	0.98
Beet	0.97	0.98	0.99	0.02	0.98
Test set					
Cane	1.00	1.00	1.00	0.00	1.00
Beet	1.00	1.00	1.00	0.00	1.00

*Note:* Sensitivity is the probability of a truly positive test result [ $= TP/(TP + FN)$ ]; Specificity is the probability of a truly negative test result [ $= TN/(TN + FP)$ ]; Precision is the consistency of the results when measurement is repeated [ $= TP/(TP + FP)$ ]; Error rate is the ratio of the number of erroneous units of data to the total number of units; and Accuracy is the closeness of a measurement to the true value [ $= TP + TN/(TP + FP + FN + TN)$ ].

Abbreviations: FN, sugar beet identified as sugarcane; FP, sugarcane identified as sugar beet; TN, sugar beet identified as sugar beet; TP, sugarcane identified as sugarcane.

**FIGURE 2** Raw absorbance spectra of (a) the 25 sucrose samples from beet and cane sugar and (b) after brown sugar samples and outliers were excluded.



applied to address the multicollinearity problem, since MLR gives biased prediction results if independent variables are highly correlated, also selected wavelengths were tested for model improvement.

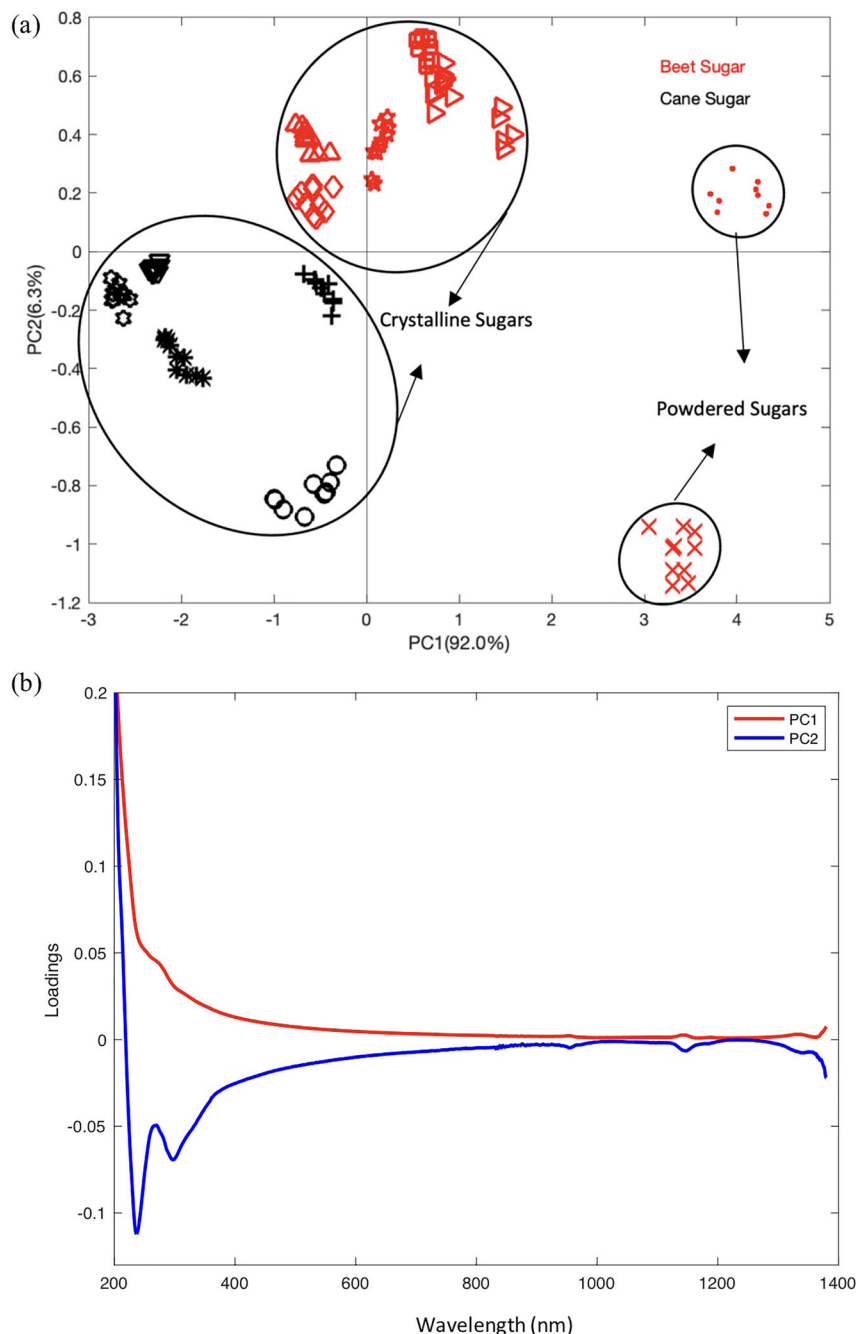
## 2.5 | Software and algorithms

PCA was applied using MATLAB® Release 2022a (The MathWorks, Inc., Natick, MA, USA). PLS without variable selection and MLR with and without Savitzky–Golay

1st Der and other preprocessing methods were conducted using Orange (Demšar et al., 2013). ScmwPLS and all classification methods were performed with MATLAB. When performing classification, the ‘classification toolbox’ of MATLAB was used (Ballabio & Consonni, 2013).

## 3 | RESULTS AND DISCUSSION

Firstly, raw spectra of the sucrose samples from beet and cane sugar were visually examined, followed by



**FIGURE 3** Principal component analysis (PCA) (a) score plot and (b) loading spectra for principal component (PC) 1 and PC 2 of the different brands of crystalline cane and beet sugar samples as well as two powdered samples.

PCA and classification with different chemometric techniques. Then, quantitative regression analysis was conducted on binary mixtures of 25% sucrose solutions from sugar beet and sugarcane. The results, obtained after the application of different multivariate analysis methods, were given and evaluated as follows: (1) absorbances at different wavelengths in the raw spectra, (2) sensitivity and specificity for classification of beet and cane sucrose samples, and (3) RMSEC, RMSECV, RMSEP,  $R^2$ , and RPD values for quantitative regression analysis.

### 3.1 | Visual interpretation of spectra

As can be seen from Figure 2a, most of the differences between the sucrose samples were observed in the UV region of the spectra. However, for some samples the UV absorbance values were too high for the spectrometer to read as the detector was saturated. Those were identified as the brown sugar samples, and it can be explained by the fact that in the UV region, colored compounds are highly absorbed. Parameters such as color, which are considerably affecting the spectral signatures

**TABLE 2** Regression results.

Model	Number of variables	Factors	Calibration set		Prediction set		RPD
			Rc <sup>2</sup>	RMSEC (%)	RMSECV (%)	RMSEP (%)	
PLSR	Full	2	0.98	4.19	5.23	6.46	4.81
PLSR (SG)	Full	3	0.98	3.18	4.36	4.44	6.99
PLSR (SG + 1st Der)	60	2	0.99	2.46	3.30	1.90	16.35
scmwiPLSR	21	2	0.99	2.32	4.11	1.88	16.50
scmwiPLSR (SG + 1st Der)	24	3	0.99	1.80	3.46	0.33	94.01
Linear regression	Full	–	0.99	1.57	4.77	3.78	8.20
Linear regression (SG)	Full	–	0.99	2.38	4.68	3.68	8.43
Linear regression (SG + 1st Der)	6	–	0.99	3.38	3.92	3.28	9.46

Abbreviations: PLSR, partial least squares regression; RMSEC, root mean square error of calibration; RMSECV, root mean square error of cross-validation; RMSEP, root mean square error of prediction; RPD, ratio of (standard error of) prediction to (standard) deviation; scmwiPLS, searching combination moving window interval PLS; SG, Savitzky–Golay.

in the UV region, might also hide valuable information required to perform accurate classification or regression analysis. It was decided that such high color deviations should be excluded from the spectral dataset to differentiate white sucrose samples, from beet and cane sugar, more effectively. Following exclusion of the brown color sugar samples, the spectra were evaluated again. There were still some outliers (Figure 2b), and these were identified as the finely powdered sugars such as “icing sugar.” During the production of powdered sugars, different ingredients such as starch are often added for anticaking purposes, which may cause hazy sugar solutions (Hollenbach et al., 1982). The presence of the starch made the solution turbid and even after waiting a day for starch to precipitate, results were not promising for most of the samples as the solutions were still hazy. Only two powdered sugars gave acceptable absorbance spectra after precipitation and could be kept in the experimental dataset.

### 3.2 | Classification of beet and cane sugar

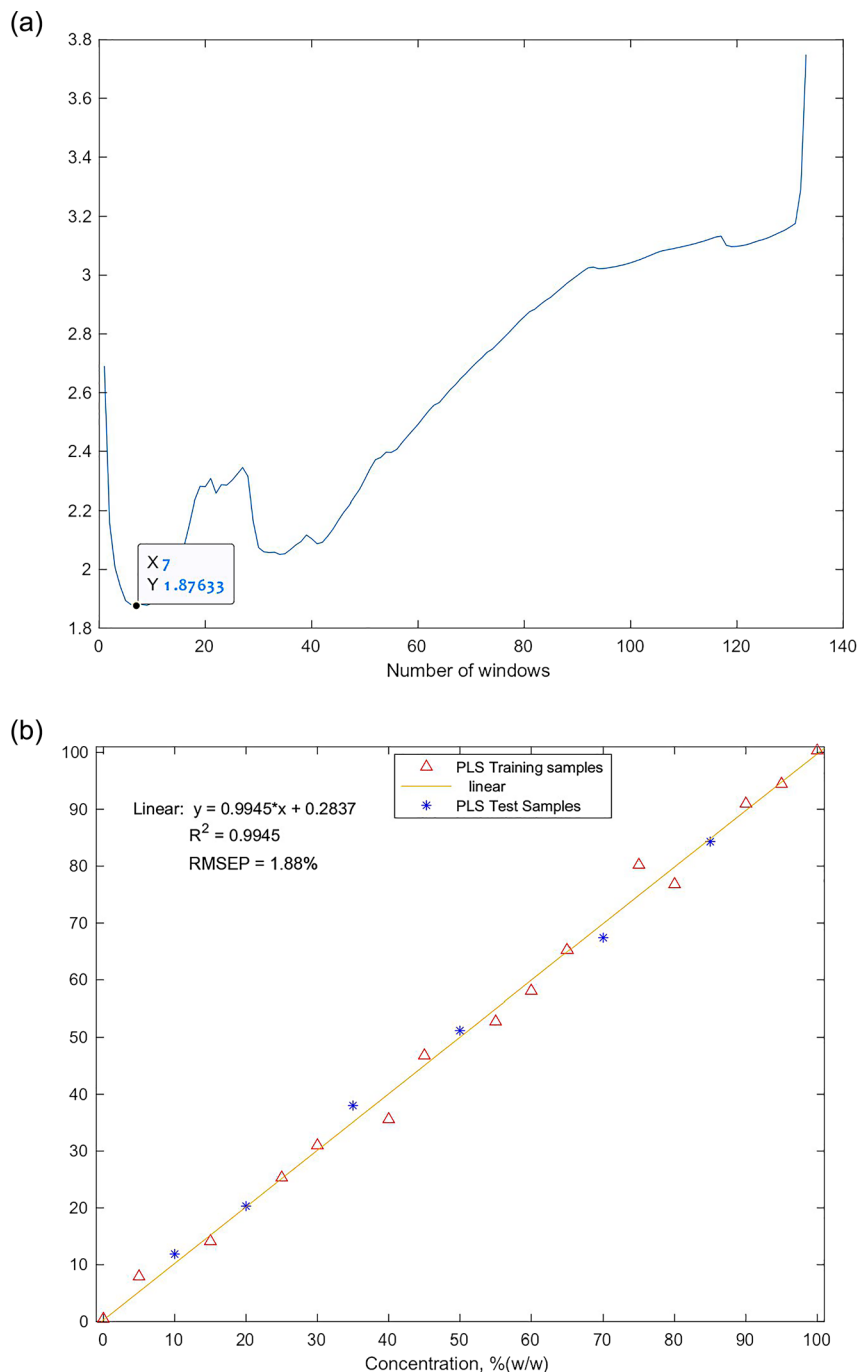
At first, it was expected to observe differences in the NIR region (800–1400), since there are two significant water absorption bands (980 and 1200 nm) in this region (Palmer & Williams, 1974). Moreover, from a theoretical point different processing steps and strategies when extracting and purifying sucrose from their sources might have resulted in impurities or some molecular changes due to the interaction with water. This could cause changes in the water bands in the NIR region, resulting in spectral signature differences. However, the resultant spectra showed that there was a noticeable difference in the UV range (220–340 nm),

much smaller differences in the visible range, and no clear differences in the NIR region. It is possible that in the NIR region, the concentration and type of impurities were not adequate to be observed in the presence of the solvent (water) signal, neither did it affect the water bands.

Since beet and cane sugar are chemically similar, and absorption of impurities present in sugar seemed to be too low to be detected in the NIR region, measuring from 800 to 1400 nm did not give information which can contribute to classification. It has been shown that even minor impurities, and processing differences could contribute to differentiation between beet and cane sugar (Lu et al., 2017). In this study, however, it was noted that impurities could only be observed in the UV region which diversified the spectra. Even small amounts of non-sugar compounds such as starch in powdered sugar or color, due to Maillard reaction, in brown sugar caused significant spectral shifts in the UV region (Figures 2a and 3a). The cause of spectral differences is most likely due to non-sugar compounds present in sugar beet or -cane. These could be polysaccharides (Godshall et al., 2002), fibers (Asadi, 2007), raffinose, and/or theandrose. The amounts of polysaccharides and fibers found in beet and cane sugar are different. Theandrose is only present in cane sugar, and it is considered a natural constituent (Moreldu Boil, 1996). In beet sugar, raffinose levels are higher compared to cane (Moreldu Boil, 1997; Vaccari & Mantovani, 1995).

**Principal component analysis (PCA):** For qualitative analysis, first PCA was applied to the dataset to detect if any outliers were present and to see if any data clusters existed. Principal component (PC) 1 and PC2 explained 98.4% of variance in the dataset which was considered high. PC3 was not included since its contribution was





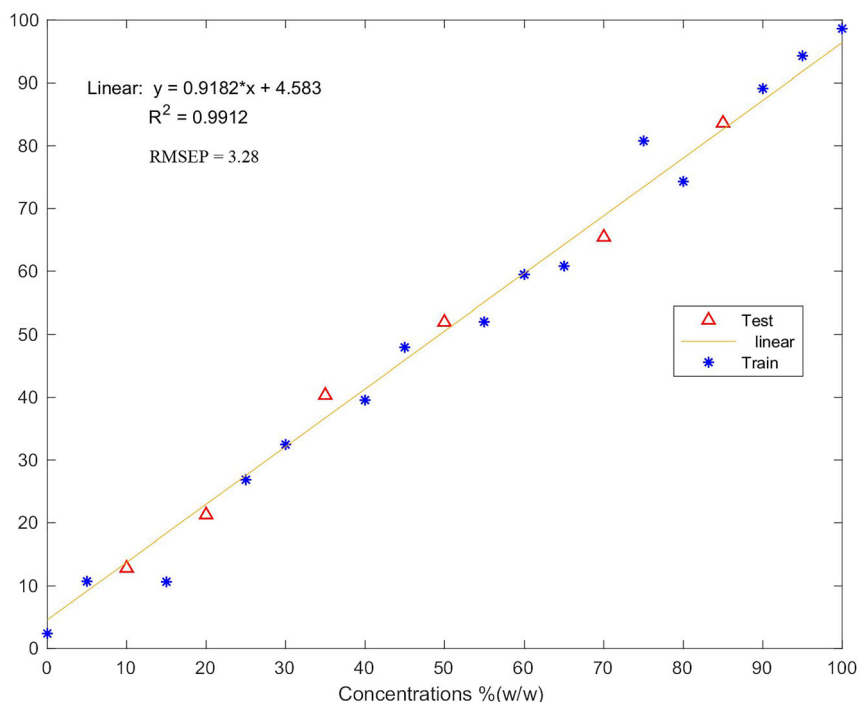
**FIGURE 4** (a) Plot of number of windows, as selected with searching combination moving window interval' PLS (scmwiPLS), versus root mean square error of prediction (RMSEP) showing lowest RMSEP (1.88%) with 7 windows and (b) actual sugar beet sugarcane binary concentration versus prediction plot with reduced variables.

considerably smaller when compared to that of PC1 and PC2 and did not provide valuable information in terms of further classification of the samples. As can be seen from Figure 3, the PCA scatter plot of PC1 versus PC2 showed differences between sucrose samples from beet and cane sugar, regions, and brands. Samples collected from different countries and brands formed small clusters without a specific pattern about their country of origin. However, samples which belong to the same brand (each brand has a different shaped marker in the figure) showed similar PC scores. It was expected since raw materials that were used for the production of the same

brand most likely underwent similar processing steps with similar equipment, and the plants sources were harvested from geographically close regions. Small distance between the PC scores for similar brand samples also showed that sampling process was done effectively, since otherwise some large deviations would have been observed.

Another observation from the PCA score plot was the easy detection of the powdered sugars. They were separated from crystalline samples even if from the same source (beet and cane sugar). As stated earlier powdered sugars caused spectral shifts because of the additives present for

**FIGURE 5** Multiple linear regression actual sugar beet sugarcane binary concentration versus prediction plot with Savitzky–Golay first derivative for six moving window selected wavelengths.



anticaking purposes. This outcome is promising since it seems likely that one can detect the presence of impurities or foreign materials added during production steps from UV absorbance spectra. On the other hand, if the goal is beet or cane differentiation, it can also be a challenge because if contamination level is high, as it can also mask the differences caused by sugar source and then separation by plant type is compromised. However, it is a problem that can be solved by filtering and decoloring to remove impurities that might affect the UV absorbance spectra drastically. Moreover, the difference in particle size could have contributed, but this effect can be reduced with appropriate preprocessing techniques.

PCA loading plots explained the level of importance of the variables to PC scores. For this study, loadings are basically linear combination of wavelengths represented by different PCs (Bro & Smilde, 2014). Here in Figure 3, for PC1, the most important wavelengths start at 200 nm, and as the wavelengths increase the loadings decrease. PC3 was not included as the goal of this study was already fulfilled with PC1 and PC2.

**Linear discriminant analysis (LDA):** As the first approach, LDA was selected, since it is one of the simplest chemometric methods that could be applied to a multivariate dataset. As the name implies, it works with a linear approach. If it is applied with two variables as in this case, it is quite intuitive to make comments on the classification. By applying LDA it was clear that differentiation between groups of sugars was possible (Figure S1). However, it had some challenges since LDA cannot be applied for datasets which have a greater number of variables than the number of samples. Thus, in this study

first it was applied by using score values of the first 2 PCs as variables (de Luca et al., 2012). Then, it was applied for five selected wavelengths, that is, 230, 250, 255, 270, and 320 nm to see if it could easily be applied in industry. Wavelengths were selected by considering the loadings given in Figure 3 and classification power results given in (Figure S2). If one can build a system with less variables, in this case fewer wavelengths, equipment costs decrease significantly. For the LDA method, the sensitivity and specificity values were both 1.00 for the model with five selected wavelengths.

**Classification and regression trees (CART):** In CART, an algorithm determines a threshold that can separate between two groups and continues in that way. CART was also tested for all variables (without applying PCA) with fivefold CV and the results are shown in Table 1. In addition, two of the beet sucrose samples were classified as cane and only one of the cane sucrose samples were classified as beet. There was a small error, and it was the only error for all trials in this study. The results for prediction on the external set were successful for all test samples, as 100% correct classification.

**Soft independent modeling of class analogy (SIMCA):** SIMCA operates well, even with high number of variables. The reason that is that SIMCA algorithm applies PCA to the classes separately thus provides a dimensionality reduction. All samples were classified correctly with specificity and sensitivity of one with fivefold cross validation and separate test set. Figure S2 shows the class distances of the model and there is a successful separation for all classes. Those results are also complementary to the findings with PCA. Classification power of SIMCA was

also complementary with the PCA loadings. The UV wavelengths have higher discrimination ability when compared with visible and NIR regions (Figure S3).

### 3.3 | Quantification of beet and cane sugar

*Raw spectra:* For the quantitative regressions, the selected wavelength range was from 200 to 600 nm since the spectral differences in the NIR region were not observable. Also, it was observed that after 380 nm, which is the start of the visible region, there were not observable differences compared to the shorter UV wavelength regions.

At approximately 270 nm there is a band which was observed also in the classification measurement (Figure S3). On the other hand, at approximately 220 nm, spectral signatures differ in an observable manner. Quantification conducted by considering the stated differences with spectral preprocessing and wavelength selection methods. Contributions of UV wavelengths to the first PC, which explained 99.8% of the variance in the dataset, were the highest. In the following sections some manual wavelength selections were applied considering the findings in Figure S3.

*Partial least squares regression:* Before building PLSR model, 15 samples were assigned as training and the remaining 6 a test set by considering leverage points and reference data distributions. The numbers of LV and other parameters were decided by assessing leave-one-out CV applied to the calibration set. And RMSECV (CV on training data), RMSEP, and RPD were calculated as 5.23%, 6.46%, and 4.81, respectively. Even though the results were promising, they can be enhanced by applying data preprocessing and variable selection methods of preprocessing included SNV, normalization, mean centering, Gaussian smoothing, Savitzky–Golay 1st and 2nd Der were tested to enhance the model quality. The best outcome was obtained with Savitzky–Golay 1st Der, thus only these results are discussed.

Application of Savitzky–Golay 1st Der (5 window gap, second polynomial order) smoothed the data and removed spectral noise. Results were 4.36%, 4.44%, and 6.99 for RMSECV, RMSEP, and RPD, respectively. With preprocessing, errors were decreased and RPD was increased (Table 2). Moreover, the difference between the errors of calibration and prediction became smaller which contributes to the robustness of the model.

After applying preprocessing, wavelength selection could be a good option since wavelengths which are not related with the target outcome can cause incorrect predictions. After selection of 60 wavelengths, using a moving window based on loadings, between 240 and 300 nm the results were 3.30%, 1.90%, and 16.35 for RMSECV,

RMSEP, and RPD, respectively. The RMSECV and RMSEP decreased and the RPD increased, which showed that preprocessing and wavelength selection strategies increased the model capabilities. However, difference between the calibration and prediction errors was larger. The  $R^2$ -values and all RMSE- and RPD-values of the abovementioned models are given in (Table 2).

When it comes to wavelength selection methods, scmwiPLS is one of the novel ones. The number of windows selections can be seen in Figure 4. For unprocessed data, 7 windows were used with 21 wavelengths as window size and the model was calculated with 2 factors. For the preprocessed data, 6 windows were used with 24 wavelengths as window size, and the model was calculated with 3 factors. It is a very successful automatized method since one cannot try all combinations by hand and assess all the results in such short times. To apply the method, first a PLS without any wavelength selection was applied, to determine the latent structure number, which gives minimum RMSECV and then window length was selected by adding one to the component number. The reason was that after several trials from different datasets, this application gave the lowest errors and highest RPDs. scmwiPLSR was successful in terms of predicting the test samples with an RMSEP of 1.88% and RPD of 16.50 and with RMSECV of 4.11%. The actual versus prediction plot is shown Figure 4, showing the good results obtained with scmwiPLSR.

Results obtained from scmwiPLS can be enhanced with preprocessing methods. After applying Savitzky–Golay 1st Der, results were 3.46%, 0.33%, and 94.01 for RMSECV, RMSEP, and RPD, respectively (Table 2). With preprocessing methods, both errors were decreased. Even though abilities of recent wavelength selection method were observed, for this study difference between prediction and calibration errors was high compared to MLR, which will be discussed in the next section.

*Multiple linear regression (MLR):* Linear regression is a method which works well under well-posed datasets which have less variables than number of samples and if multicollinearity does not exist. Regularization (ridge regression) and wavelength selection methods were applied to solve ill-posed data problem. Also in this method, samples were split as validation and calibration groups then leave-one-out cross validation was applied on calibration set to decide model parameters. Evaluation results were 4.77%, 3.78%, and 8.20 for RMSECV, RMSEP, and RPD, respectively (Table 2). With the application of Savitzky–Golay 1st Der, the results were slightly better with 4.68%, 3.68%, and 8.43 for RMSECV, RMSEP, and RPD, respectively.

Finally, after applying Savitzky–Golay 1st Der, six wavelengths (226, 227, 228, 229, 230, and 231 nm) were selected from the UV spectra. The wavelengths were selected based

on PCA loading results and using moving window method and for the purpose of easier application of the method. The results obtained were 3.92%, 3.28%, and 9.46 for RMSECV, RMSEP, and RPD, respectively. Moreover, actual concentration versus predicted concentration plot can be seen in Figure 5. Results for this narrow wavelength range seemed promising for industrial application. As shown above, by applying preprocessing and wavelength selection methods, error difference between test and train datasets can be decreased since noise and unnecessary spectra can be excluded from dataset.

## 4 | CONCLUSION

Optical spectroscopy with chemometric methods provided promising results for many studies in differentiating origins of food materials. However, to our knowledge, UV spectroscopy was not studied to test the authenticity of sucrose sources (sugar beet and sugarcane), which made current work relevant and important. This study showed that the UV region of the electromagnetic spectrum was highly sensitive for impurities that could be used to diversify the sources of sucrose. All supervised classification methods, including LDA, CART, and SIMCA, showed high performance to authenticate the source of the sucrose. Data clusters were obtained for same branded sugars but not for the same country of origin. In addition to that, LDA with only five selected wavelengths provided 100% classification with the simplest interpretation. For regression analysis, even though PLS gave the highest RPD and lowest prediction errors, MLR with Savitzky–Golay 1st Der preprocessing and the least number of variables gave the most applicable results with RMSECV, RMSEP, and RPD as 3.92%, 3.28%, and 9.46, respectively. This indicates the potential of a simple and easy to use industry approach. The obtained results seemed promising that the plant source of sucrose can be differentiated using UV spectra in association with chemometric methods.

## AUTHOR CONTRIBUTIONS

**Hilmi Eriklioglu:** Conceptualization; methodology; formal analysis; investigation; writing—original draft. **Esmanur Ilhan:** Conceptualization; methodology; writing—review and editing. **Mikhail Khodasevich:** Supervision; methodology; writing—review and editing. **Darya Korolko:** Formal analysis; writing—review and editing. **Marena Manley:** Conceptualization; methodology; supervision; writing—review and editing. **Rosario Castillo:** Methodology; supervision; writing—review and editing. **Mecit Halil Oztop:** Conceptualization; methodology; supervision; resources; writing—review and editing; funding acquisition.



## ACKNOWLEDGMENTS

The authors express their deepest gratitude to Assist. Dr. Ali Can Karaca for his guidance, advice, criticism, encouragements, and insight throughout the research.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## ORCID

Marena Manley  <https://orcid.org/0000-0001-7581-7208>  
Mecit Halil Oztop  <https://orcid.org/0000-0001-6414-8942>

## REFERENCES

- Asadi, M. (2007). *Beet-sugar handbook*. Wiley-Interscience.
- Bahrami, M. E., Honarvar, M., Ansari, K., & Jamshidi, B. (2020). Measurement of quality parameters of sugar beet juices using near-infrared spectroscopy and chemometrics. *Journal of Food Engineering*, 271, 109775. <https://doi.org/10.1016/j.jfoodeng.2019.109775>
- Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods*, 5(16), 3790–3798. <https://doi.org/10.1039/c3ay40582f>
- Baranowski, P., Mazurek, W., Wozniak, J., & Majewska, U. (2012). Detection of early bruises in apples using hyperspectral data and thermal imaging. *Journal of Food Engineering*, 110(3), 345–355. <https://doi.org/10.1016/j.jfoodeng.2011.12.038>
- Barbosa, R. M., Nacano, L. R., Freitas, R., Batista, B. L., & Barbosa, F. (2014). The use of decision trees and naïve bayes algorithms and trace element patterns for controlling the authenticity of free-range-pastured hens' eggs. *Journal of Food Science*, 79(9), C1672–C1677. <https://doi.org/10.1111/1750-3841.12577>
- Boggia, R., Turrini, F., Anselmo, M., Zunin, P., Donno, D., & Beccaro, G. L. (2017). Feasibility of UV–VIS–Fluorescence spectroscopy combined with pattern recognition techniques to authenticate a new category of plant food supplements. *Journal of Food Science and Technology*, 54(8), 2422–2432. <https://doi.org/10.1007/s13197-017-2684-7>
- Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9), 2812–2831. <https://doi.org/10.1039/c3ay41907j>
- Bubnik, Z., Kadlec, P., Urban, D., & Bruhns, M. (1995). Chemical and physical data for sugar manufacturers and users. In: *Sugar technologists manual*, (8th ed) (pp. 417). Bartens Pub. Co.
- Cortés, V., Blasco, J., Aleixos, N., Cubero, S., & Talens, P. (2019). Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review. *Trends in Food Science and Technology*, 85, 138–148. <https://doi.org/10.1016/j.tifs.2019.01.015>
- Dankowska, A., Domagała, A., & Kowalewski, W. (2017). Quantification of *Coffea arabica* and *Coffea canephora* var. robusta concentration in blends by means of synchronous fluorescence and UV-Vis spectroscopies. *Talanta*, 172(January), 215–220. <https://doi.org/10.1016/j.talanta.2017.05.036>
- Dankowska, A., & Kowalewski, W. (2019). Tea types classification with data fusion of UV–Vis, synchronous fluorescence and NIR spectroscopies and chemometric analysis. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 211, 195–202. <https://doi.org/10.1016/j.saa.2018.11.063>



- De Luca, M., Terouzi, W., Kzaiber, F., Ioele, G., Oussama, A., & Ragno, G. (2012). Classification of moroccan olive cultivars by linear discriminant analysis applied to ATR-FTIR spectra of endocarps. *International Journal of Food Science and Technology*, 47(6), 1286–1292. <https://doi.org/10.1111/j.1365-2621.2012.02972.x>
- Demšar, J., Erjavec, A., Hočevár, T., Milutinovič, M., Možina, M., Toplak, M., Umek, L., Zbontar, J., & Zupan, B. (2013). Orange: Data mining toolbox in python Tomaz Curk Matija Polajnar Laň Zagar. *Journal of Machine Learning Research*, 14(2013), 2349–2353.
- Du, Y. P., Liang, Y. Z., Jiang, J. H., Berry, R. J., & Ozaki, Y. (2004). Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Analytica Chimica Acta*, 501(2), 183–191. <https://doi.org/10.1016/j.aca.2003.09.041>
- Esbensen, K. H., & Geladi, P. (2009). Principal component analysis: Concept, geometrical interpretation, mathematical background, algorithms, history, practice. *Comprehensive Chemometrics*, 2, 211–226. <https://doi.org/10.1016/B978-0-44452701-1.00043-0>
- Fanelli, V., Mascio, I., Miazzi, M. M., Savoia, M. A., De Giovanni, C., & Montemurro, C. (2021). Molecular approaches to agri-food traceability and authentication: An updated review. *Foods*, 10(7), 1644. <https://doi.org/10.3390/foods10071644>
- Godshall, M. A., Vercellotti, J. R., & Triche, R. (2002). Comparison of cane and beet sugar macromolecules in processing. *International Sugar Journal*, 103(1241), 228–233.
- Hollenbach, A. M., Peleg, M., & Rufner, R. (1982). Effect of four anticaking agents on the bulk characteristics of ground sugar. *Journal of Food Science*, 47(2), 538–544. <https://doi.org/10.1111/j.1365-2621.1982.tb10119.x>
- Khodasevich, M. A., Trofimova, D. V., & Nezalzova, E. I. (2010). Principal component analysis of UV-VIS-NIR transmission spectra of Moldavian matured wine distillates. In *LAT 2010: International Conference on Lasers, Applications, and Technologies*, (Vol. 7994, pp. 79941F). <https://doi.org/10.1117/12.881593>
- Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, 39(4), 261–283. <https://doi.org/10.1007/s10462-011-9272-4>
- Lu, Y., Thomas, L., & Schmidt, S. (2017). Differences in the thermal behavior of beet and cane sucrose sources. *Journal of Food Engineering*, 201, 57–70. <https://doi.org/10.1016/j.jfoodeng.2017.01.005>
- Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews*, 43(24), 8200–8214. <https://doi.org/10.1039/c4cs00062e>
- Morel du Boil, P. G. (1996). Theanderose—A characteristic of cane sugar crystals. *South African Sugar Technology Association*, 70140–70144.
- Morel du Boil, P. G. (1997). Theanderose—Distinguishing cane and beet sugars. *International Sugar Journal*, 99(1179), 102–106.
- Palmer, K. F., & Williams, D. (1974). Optical properties of water in the near infrared. *Journal of the Optical Society of America*, 64(8), 1107–1110. <https://doi.org/10.1364/JOSA.64.001107>
- Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, 44, 683–700. <https://doi.org/10.1016/j.jpba.2007.03.023>
- Sefaoğlu, F., Kaya, C., & Karakuş, A. (2016). Farklı Tarihlerde Hasat Edilen Şeker Pancarı Genotiplerinin Verim ve Verim Unsurlarının Belirlenmesi. *Tarla Bitkileri Merkez Araştırma Enstitüsü Dergisi*, 25(ÖZEL SAYI-2), 61. <https://doi.org/10.21566/tarbitderg.281846>
- Souto, U. T. D. C. P., Barbosa, M. F., Dantas, H. V., De Pontes, A. S., Lyra, W. D. S., Diniz, P. H. G. D., De Araújo, M. C. U., & Da Silva, E. C. (2015). Identification of adulteration in ground roasted coffees using UV-Vis spectroscopy and SPA-LDA. *LWT - Food Science and Technology*, 63(2), 1037–1041. <https://doi.org/10.1016/j.lwt.2015.04.003>
- Suga Act (2001). Sugar Act, Act No. 4634, T.C. Official Gazette (24378, 19 March 2021). <https://articlereview.pubmate.in/#/?templateID=08b417d8a2604c688f4c730be866b411202256849>
- Suhandy, D., & Yulia, M. (2021). The use of UV spectroscopy and SIMCA for the authentication of Indonesian honeys according to botanical, entomological and geographical origins. *Molecules*, 26(4), 915. <https://doi.org/10.3390/molecules26040915>
- Thow, A. M., Lencucha, R. A., Rooney, K., Colagiuri, S., & Lenzen, M. (2021). Implications for farmers of measures to reduce sugars consumption. *Bulletin of the World Health Organization*, 99(1), 41–49. <https://doi.org/10.2471/BLT.19.249177>
- Urbanus, B. L., Cox, G. O., Eklund, E. J., Ickes, C. M., Schmidt, S. J., & Lee, S.-Y. (2014). Sensory differences between Beet and cane sugar sources. *Journal of Food Science*, 79(9), S1763–S1768. <https://doi.org/10.1111/1750-3841.12558>
- Vaccari, G., & Mantovani, G. (1995). Sucrose crystallization. In: Mathlouthi, M., Reiser, P. (Eds.), *Sucrose properties and applications*, (1st ed.) (pp. 33–72). Blackie Academic and Professional.
- Vanden Branden, K., & Hubert, M. (2005). Robust classification in high dimensions based on the SIMCA Method. *Chemometrics and Intelligent Laboratory Systems*, 79(1–2), 10–21. <https://doi.org/10.1016/j.chemolab.2005.03.002>
- Williams, P. (2014). The RPD statistic: A tutorial note. *NIR News*, 25(1), 22–26. <https://doi.org/10.1255/nirn.1419>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Eriklioglu, H., Ilhan, E., Khodasevich, M., Korolko, D., Manley, M., Castillo, R., & Oztop, M. H. (2023). Classification and quantification of sucrose from sugar beet and sugarcane using optical spectroscopy and chemometrics. *Journal of Food Science*, 88, 3274–3286. <https://doi.org/10.1111/1750-3841.16674>