WILSON JUNIOR CARDOSO

QUANTIFICATION OF CARBOHYDRATES AND FIBER FROM DEHYDRATE SUGARCANE JUICE USING NEAR-INFRARED SPECTROSCOPY AND CHEMOMETRICS

Dissertation submitted to the Agrochemistry Graduate Program of the Universidade Federal de Viçosa as part of the requirements for the *Magister Scientiae* degree.

VIÇOSA
MINAS GERAIS – BRASIL
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
C268q
2019

Cardoso, Wilson Júnior, 1992-
Quantification of carbohydrates and fiber from dehydrate sugarcane juice using near-infrared spectroscopy and chemometrics / Wilson Júnior Cardoso. – Viçosa, MG, 2019.
xiv, 63f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Orientador: Reinaldo Francisco Teófilo.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.58-63.

1. Espectrometria de infravermelho próximo. 2. Carboidratos. 3. Cana-de-açúcar - Seleção. 4. Cana-de-açúcar - Genética. I. Universidade Federal de Viçosa. Departamento de Química. Programa de Pós-Graduação em Agroquímica. II. Título.

CDD 22 ed. 543.55

WILSON JUNIOR CARDOSO

QUANTIFICATION OF CARBOHYDRATES AND FIBER FROM DEHYDRATE SUGARCANE JUICE USING NEAR-INFRARED SPECTROSCOPY AND CHEMOMETRICS

Dissertation submitted to the Agrochemistry Graduate Program of the Universidade Federal de Viçosa as part of the requirements for the *Magister Scientiae* degree.

APPROVED: February 22th, 2019.

Márcio Henrique Pereira Barbosa

Luiz Alexandre Peternelli

Reinaldo Francisco Teófilo
(Advisor)

I dedicate my achievements

to my dearest parents, Elenice and Wilson

to my fearless sister, Luana and

to Marco for all the support and patience.

Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something and that this thing must be attained.

(Marie Curie)

Acknowledgements

I would like to thank God for guiding my steps through the academic path.

I am particularly grateful to my parents and sister for all the support and incentive.

I would like to thank Marco for his support and encouragement throughout my study years.

I would like to express my deepest gratitude to Professor Teófilo, my research advisor, for providing me this opportunity. I appreciate his patient guidance, enthusiastic encouragement, and useful critiques.

I would like to thank Professor Barbosa for supporting and providing resources to my research.

I would like to thank the Sugarcane Breeding Program of Federal University of Viçosa (PMGCA-UFV) and to RIDESA (Inter-University Network for the Development of Sugarcane Industry) for providing the sugarcane samples.

I would like to thank Professor Luiz Antonio dos Santos Dias for making available the near-infrared instrument.

I would like to thank my fellow peers from the MCDA Laboratory, Wesley, Larissa, Nathália, Danilo, and Helder. Special mention to Jussara for all the assistance and for being present.

I would like to thank Rafael, Cássio, and Nathália for the friendship and laughs and Mel, the best doggie ever, for all the moments and unconditional love, may she rest in peace.

I am very thankful to the Universidade Federal de Viçosa, specially the Chemistry Department and the Department of Crop Science, for providing the resources.

I am very thankful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance code 001 for granting the scholarship.

Last of all, I express my sincere appreciation to all those who have contributed to this work and supported me in one way or the other.

Summary

Abbreviations

A	Accuracy
ANOVA	Analysis of variance
BRIX	Soluble solids
CCD	Central composite design
CLA	Bright Paper Card (Kraft side)
CONAB	Companhia Nacional de Abastecimento
CV	Cross-validation
DOE	Design of Experiments
E	Error
ELSD	The evaporative light scattering detector
ESC	Dark Paper Card (Kraft side)
EVA	Ethylene vinyl acetate
FAR	Far-infrared
FIN	Thin couche paper
FN	False negative
FP	False positive
FRU	Fructose
GCS	Glass cover slips
GLU	Glucose
GRO	Thick couche paper
HPD	Heating plate
HPLC	High performance liquid chromatography
LOD	Limit of detection
LOQ	Limit of quantification
LV	Latent variable
MCDA Lab	Multivariate chemical data analysis laboratory
MID	Mid-infrared
MSC	Multiplicative scatter/signal correction
NAS	Net analytical signal
NIR	Near-infrared spectroscopy
OSC	Orthogonal signal correction
OVD	Oven
PAR	Paper kraft
PLS	Partial-least squares
PMGCA-UFV	Sugarcane Breeding Program of Federal University of Viçosa
POL	Saccharimetric reading based on dextro-rotatory substances
R	Correlation coefficient
R^2	Coefficient of determination
Rc	Correlation coefficient of calibration
Rc^2	Quadratic correlation coefficient of calibration
Rcv	Correlation coefficient of cross-validation
Rcv^2	Determination coefficient of cross-validation
RIDESA	Inter-University Network for the Development of Sugarcane Industry
RMSE	Root mean square error
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross-validation
RMSEP	Root mean square error of prediction
Rp	Correlation coefficient of prediction
RSD	Relative standard deviation
RSM	Response surface methodologies
S	Sensitivity

SEL	Selectivity
SEN	Sensibility
SIL	Silicon
SNV	Standard normal variate scaling
Sp	Specificity
SUC	Sucrose
TEF	Teflon
TN	True negative
TP	True positive

Figures

Tables

Abstract

CARDOSO, Wilson Junior, M.Sc., Universidade Federal de Viçosa, February, 2019. **Quantification of Carbohydrates and Fiber from Dehydrate Sugarcane Juice using Near-Infrared Spectroscopy and Chemometrics.** Advisor: Reinaldo Francisco Teófilo.

The main goal of this work was to propose a new method to collect NIR spectrum from dehydrated sugarcane juice to predict soluble solids (BRIX) and dextro-rotatory substances (POL), fiber, sucrose, glucose, and fructose, using multivariate regression models. A total of 133 different genotypes harvested in different periods were analyzed. The BRIX, POL and fiber values ranged from 5.4 to 23.6; 8.3 to 89.8; and 8.4 to 22.5 %, respectively. Sucrose (SUC), glucose (GLU), and fructose (FRU) were quantified using high performance liquid chromatography (HPLC). The SUC, GLU and FRU values ranged from 25.7 to 215.7; 2.8 to 25.5, and 3.2 to 21.7 mg/mL. As the main constituent of the sugarcane juice, water absorbs strongly the NIR radiation and hinders the detection of other chemical species. For that reason, two new methods were proposed to remove water: one using an oven (OVD) and another using a heating plate (HPD). Among the 9 support materials studied, thick couche paper (GRO) was the best support for the OVD method and glass cover slips (GCS) was the support chosen for the HPD method. For the OVD method, the temperature, time, and sample volume were optimized using a central composite design (CCD). The optimum parameters were, 113.6 ºC, 19.55 minutes, and 51.6 µL. For the HPD method, once there is more control of the drying process, the temperature was fixed at 120 ºC and the time at 2.5 minutes. NIR spectra were acquired in the range from 1000 to 2500 nm. PLS models were built for OVD and HPD methods using two ranges of NIR spectra, *i.e.*, the full, from 1000 to 2500 nm and the short, from 1000 to 1700 nm. The short one is normally used in portable NIR instruments. The root mean square error of prediction (*RMSEP*) values for SUC, GLU, FRU and fiber using the OVD method were respectively 17.33; 3.98; 1.85; and 1.95 for full spectra and 18.52; 4.37; 3.04; and 2.20 for short spectra. The RMSEP for SUC, GLU, FRU and fiber using the HPD method were respectively 19.91; 3.77; 1.97; and 2.58 for full spectra and 26.86; 3.89, 2.60; and 3.56 for short spectra. BRIX and POL models were not presented. For OVD and HPD methods , the analytical frequency were 2 samples per min and 1.2 samples per min, respectively. For NIR, the analytical frequency was 1 sample per min. The best sucrose model presented to be superior to the traditional methods (BRIX and POL) when selecting genotypes.

Resumo

CARDOSO, Wilson Junior, M.Sc., Universidade Federal de Viçosa, fevereiro de 2019. **Quantificação de carboidratos e fibra no caldo de cana-de-açúcar desidratado usando espectroscopia no infravermelho próximo e quimiometria.** Orientador: Reinaldo Francisco Teófilo.

O principal objetivo do presente trabalho foi propor um novo método para obtenção de espectros NIR a partir do caldo de cana-de-açúcar desidratado para prever os valores de BRIX, POL, fibra, sacarose, glicose e frutose usando regressão multivariada. Foram utilizados 133 diferentes genótipos coletados em diferentes épocas. Os valores de BRIX, POL e fibra variaram de 5,4 a 23.6, 8,3 a 89,8 e 8,4 a 22,5 %, respectivamente. Sacarose (SUC), glicose (GLU) e frutose (FRU) foram quantificadas usando cromatografia líquida de alta eficiência (HPLC). Os valores de SUC, GLU e FRU variaram de 25,7 a 215,7, 2,8 a 25,5 e 3,2 a 21,7 mg/mL, respectivamente. A água, como principal constituinte do caldo de cana-de-áçucar, absorve fortemente a radiação NIR e, assim, impede a detecção de outras espécies químicas. Sendo assim, dois novos métodos foram pospostos para remoção da água, um usando uma estufa (OVD) a outro utilizando uma chapa de aquecimento (HPD). Entre os 9 suportes estudados, o papel couchê grosso (GRO) se mostrou o melhor suporte para o método OVD e a lâminula de vidro foi o suporte escolhido para o método HPD. Para o método OVD, a temperature, tempo e volume de amostra foram otimizados utilizando um planejamento compost central (CCD). Os parametros ótimos foram, 113,6 ºC, 19,55 minutos e 51,6 µL. Para o método HPD, uma vez que há mais controle do processo de secagem, a temperature foi fixada em 120 ºC e o tempo em 2,5 minutos. Os espectros NIR foram obtidos na faixa de 1000 a 2500 nm. Modelos PLS foram construídos para os métodos OVD e HPD usando duas faixas de espectros NIR, ou seja, completa, de 1000 a 2500 nm, e curta, de 1000 a 1700 nm. A faixa curta normalmente é utilizada em instrumentos NIR portáteis. A raiz do erro quadrático médio de validação (*RMSEP*) para SUC, GLU, FRU e fibra usando o método OVD foram 17,33, 3,98, 1,85 e 2,58 para o espectro complete e 18,52, 4,37, 3,04 e 2,33 para o espectro curto. O RMSEP para SUC, GLU, FRU e fibra usando o método HPD foram 19,91, 3,77, 1,97 e 3,52 para o espectro complete e 26,86, 3,89, 2,60 e 3,18 para o espectro curto. Modelos para BRIX e POL não foram apresentados. Para os métodos OVD e HPD, a frequência analítica foi de 2 amostras por minuto e 1,2 amostras por minuto, respectivamente. Para o NIR, a frequência analítica foi de 1 amostra por minuto. O melhor

modelo de sacarose apresentado foi superior aos métodos tradicionais (BRIX e POL) na seleção de genótipos.

1. Introduction

Renewable energy, such as hydropower, solar, wind, and bioenergy, are potential energy sources that have been slowly replacing fossil fuels throughout the years [1]. Among the bioenergy sources, biomass, such as corn and sugarcane, is one of the most promising energy source [2,3].

Brazil is the largest sugarcane producer, leading the production of both ethanol and sugar. This leadership came from years of research in genetic breeding, pest management, better cultivation techniques, and improvement of the ethanol and sugar production techniques. Furthermore, the sugarcane economic interest still ascending due to the increasing demand for sustainable energy sources [4]. According to the Companhia Nacional de Abastecimento (CONAB), the sugarcane production estimative is of 626 millions of tons for the 2018/2019 harvest [5].

Genetic breeding consists in the selection and crossing of genotypes with desirable agronomic properties in order to obtain cultivars with superior properties in comparison with their genitors[4]. This process is laborious, taking more than a decade to obtain a commercial cultivar. As of 250,000 new genotypes obtained from crossings, one commercial genotype is obtained. These superior genotypes, having as main characteristic high sucrose content, have high economic value and are responsible in large part for the increasing in the productivity of sugarcane and industrial yield [6,7]. Consequently, sucrose content is one of the main characteristics that must be quantified both in the genetic breeding to select new genotypes and in the sugarcane harvest to assess the feedstock quality.

In genetic breeding as well as in the industry, sucrose content is estimated through BRIX and POL measurements. These conventional methods are lengthy since the juice needs to be clarified a priori, which makes these analyses expensive and time consuming. Although BRIX and POL are associated to sucrose content, their values may contain inherent errors, leading to an incorrect estimation of sucrose content. High performance liquid chromatography (HPLC) is one of the most accurate techniques for sugar quantification [8], but it is an onerous method to be applied routinely. However, this setback can be solved by means of spectroscopic methods coupled with chemometrics. Among the spectroscopic analysis techniques, one of the most versatile is the near-infrared spectroscopy (NIR).

NIR uses the region of the electromagnetic spectrum (from 780 nm to 2500 nm) and brings information based on bonds vibrations. There are bench instruments that investigates full NIR range and portable instruments that investigates a short range of the spectrum. It is a fast, non-destructive technique capable of analyzing molecules containing C-H, N-H, O-H and S-H bonds. Despite the high overlap of spectral information, when coupled with chemometrics methods, such as multivariate regression, NIR may return qualitative and quantitative information about the samples. Partial-least squares (PLS) regression is one of the most common regression methods used to build models using NIR spectra. A properly calibrated multivariate model is a simple, inexpensive, and fast method capable of replacing routine analyzes applied in the industry and genetic breeding programs. A well-built and validated model may be able to predict technological properties such as BRIX, POL, fiber, and sugars, such as sucrose, glucose, and fructose using a simple spectrum quickly collected from any part of the plant. Therefore, NIR coupled with chemometric methods has potential to replace conventional analyzes used both in genetic breeding sites as well as in the industry [8,9].

The main goal of this work is to propose a new method to collect NIR spectrum from sugarcane juice to predict BRIX, POL, fiber, sucrose, glucose, and fructose, using multivariate models.

2. Literature Review

2.1. Sugarcane

2.1.1. Overview

The sugarcane is a perennial tropical grass belonging to the *Poaceae* and *Saccharum* family and genera, respectively. It is composed mostly by water (75-82%), sugars (15-24%), and fiber (8-14%), as can be seen in **Figure 1** below [10]. Moreover, it comprises several species, such as *Saccharum officinarum* L., *Saccharum spontaneum* L., *Saccharum robustum* J., *Saccharum sinnensis* R., and *Saccharum barberi* J. [11]. Although its certain place of origin is not known, it is believed that the sugarcane is native from southeast Asia.

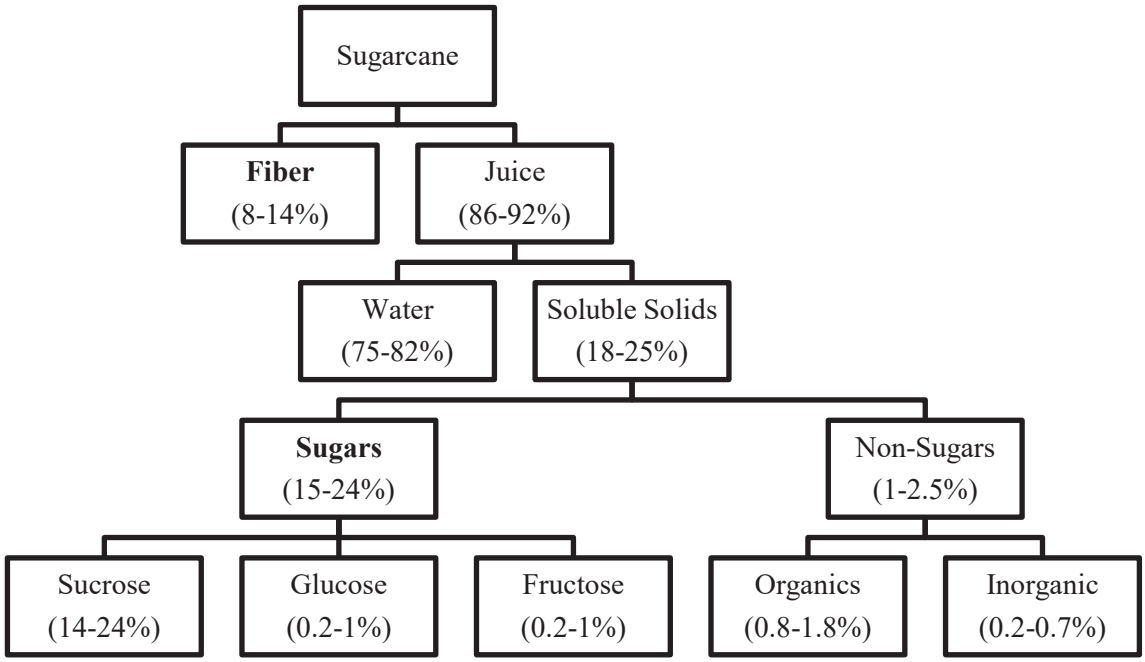

Figure 1. Sugarcane composition.

The sugarcane was introduced in Brazil in the beginning of 14th century with the colonization process. The promising climate and fertile soil made possible the expansion of the cultivar all over the country. At that time, the most important species were *S. barberi*, *S. officinarum*, and their hybrids, because of their high sugar content. From the beginning, the sugarcane, and therefore its byproducts, such as sugar, molasses, brown sugar, and spirits, contributed to the development of the country [12].

With the expansion of the sugarcane cultivation, several diseases started affecting the crops. In Brazil, phytosanitary health issues relating to the sugarcane crops started in 1863 with the rot of cuttings, then other sugarcane varieties resistant to this disease began being cultivated [13]. In 1920, the mosaic started affecting the crops in Brazil, spreading

fast to all producing regions, causing major productivity problems in the canebrakes. At that point, the search for control of these illnesses started in several science fields, but efficient results appeared only through breeding [14,15]. Breeding programs made possible the creation of new cultivars less sensible to the mosaic virus and other diseases. Thus, the ascension of the genetic breeding programs started all around the world.

Plant breeding is the science of changing and improving the heredity of plants. It is believed that genetic improvements of crop plants began thousands years ago, when the best plants were selected from the wild for domestication, and crossings by chance, natural mutation, and arrival of new types from other lands played an important role in this process [16]. Deliberate breeding did not begin until the 20th century, but there are evidences of breeding efforts for apple and pear in the 17th century and for cereals and potatoes in the 18th and 19th century [17].

Gregor Mendel was knows as the first genetic engineer when he crossed true breeding lines of garden pea and described the inheritance of parental traits [18]. From then on, plant breeding evolved from being a random process to being deliberate and directed [19].

Among the plant breeding techniques, crossbreeding plays a major role. Crossbreeding process aim to produce more productive, tolerant to stresses (*i.e.*, diseases, insect damage, and weather) crop plants through crossings of plants with superior characteristics [20]. On that note, breeders search, create, and select genetically mixed populations of plants with desirable characteristics aiming commercial distribution. Before the breeder finds a superior strain, it is necessary to make many crosses and to grow thousands of experimental strains. It is a very complex, costly, time-consuming process. Therefore, to make the breeding programs more efficient, breeders must develop careful and accurate crossing and selection techniques [21].

Sugarcane specific genetic breeding programs resulted from the advent of numerous diseases, the main concern at that time. The first hybridizations were accomplished in Java and Barbados in 1888, following observations that the sugarcane was able to produce viable seeds [22]. Since then, interspecific crossings of species such as *S. officinarum*, *S. spontaneum*, *S. barberi*, and *S. sinense* and their hybrids were made to acquire cultivars with superior characteristics [23–25]. The sugarcane breeding programs were adapting themselves accordingly to the new demands, aiming to the medium and long-term [26].

Nowadays, the main goal of the sugarcane genetic breeding programs is to provide new strains tolerant to specific diseases and insects and capable of increasing the productivity of the sugarcane or energycane, enhancing its sugar and fiber contents. In addition, the high productivity of the Brazilian's genetic breeding programs made possible the cultivation of several commercial sugarcane, adapted to specific soil, weather, and harvesting season conditions.

Most current cultivated sugarcane are multi-species hybrids, taxonomically known as *Saccharum* spp. [27,28]. These cultivars present high sugar content, low or medium fiber content, and high tolerance to most existing diseases. The sugarcane quality is assessed through a series of intrinsic characteristics, such as sugar and fiber contents and tolerance/resistance to disease. Thus, considering the industrial concept, the economic yield will be assured only through maximization of sugar and ethanol productivity by acreage, *i.e.*, maximization of the sucrose content.

2.1.2. Quality parameters of sugarcane

2.1.2.1. Carbohydrates

The soluble solids from sugarcane juice are composed mainly by sucrose, glucose, and fructose, the first being the most abundant, from 14 to 24%. Sucrose is a disaccharide formed by two monosaccharides, glucose and fructose, linked by a glycosidic bond. Under acidic or enzymatic condition, this glycosidic bond can be broken down releasing the glucose and fructose, also known as inverted sugars. Sucrose is the most important sugarcane quality parameter, and it is usually estimated through soluble solids (BRIX) and dextro-rotatory substances (POL) measurements [29,30]. **Figure 2** shows the structures of sucrose, glucose, and fructose, respectively.

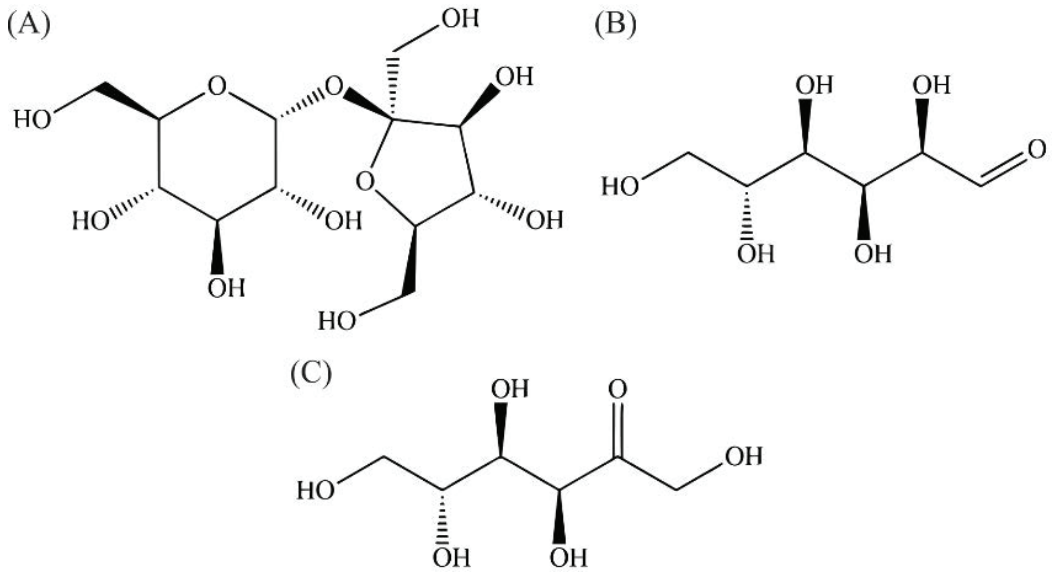

Figure 2. Chemical structure of sucrose (A), glucose (B), and fructose (C).

2.1.2.2. POL

POL, also known as saccharimetric reading, represents the estimated sucrose contained in a sugar solution and it is determined by saccharimetric methods, such polarimeters or saccharimeters. These instruments consist of an optical device that measures light deviation based on the presence of dextrorotatory (deviation to the right) and levorotatory (deviation to the left). Sucrose and glucose are dextrorotatory compounds while fructose is levorotatory. Therefore, if glucose and fructose contents are different, the instrument would not indicate the exact sucrose content but an overestimated or underestimate value [10,29].

2.1.2.3. BRIX

BRIX, also known as dry substance, is the most utilized parameter in the industry. It expresses the weight percentage of soluble solids in a pure solution of sucrose. Consequently, BRIX is an estimative of soluble solids, *i.e.*, sucrose, in impure solutions, such as sugarcane juice. It can be determined by refractometry, measuring the light deviated by the solution in function of its density. Some factors may affect its accuracy, such as temperature [10,29].

2.1.2.4. Fiber

Fiber is the insoluble solid of the sugarcane and is constituted mostly of holocellulose (71.39%), cellulose (54.87%), hemicelluloses (16.52%), and Klason

lignin (23.33%) [31]. Its content is associated with the resistance of the plant to physical distresses and some insects. In addition, it is important for the energetic balance of the sugar and alcohol industry, once the bagasse is burned to cogenerate electricity [10,29]. Fiber content is determined by the hydraulic press method [32] or estimated through a linear regression, **Equation 1**, that correlates fiber content with the wet bagasse weight [33].

$$\mathbf{F = 0.08 \times PBU + 0.8760} \qquad \textbf{Eq. 1}$$

2.2. High Performance Liquid Chromatography

High performance liquid chromatography (HPLC) is the most versatile and employed elution chromatography. It offers a fast, automated, and accurate method to separate, determine, and quantify compounds in a wide variety of organic, inorganic, and biological materials [34].

The HPLC system is composed of a mobile phase, a liquid where the solutes are dissolved, and a stationary phase, a solid or liquid covalently bonded to the column walls. The separation of the solutes occurs when the mobile phase, containing the solutes, is pumped through the column, containing the stationary phase. The stationary phase retains some compounds longer than others, consequently, the solutes are separated based on the polarity of the solutes and the polarity of the mobile and stationary phases. The chromatogram obtained at the end of the analysis displays the variation of the detector as a function of the time, where each peak corresponds to a different solute and the retention time is the time needed to a specific solute to reach the detector [35].

The most important aspect of the HPLC is the development of smaller particles, which resulted in columns more packed and, consequently, more efficient. In addition, these columns demand the use of pumps to pressurize the system and facilitate the flow of the mobile phase through the column. As of its operation, the mobile phase is pumped to the column, the sample is then introduced to the mobile phase at the head of the column by an injection valve and it is then carried to the column, where the separation occurs. The effluent of the column is directed to a detector and the chromatogram is generated by a suitable software based on the signal acquired from the detector [36].

Among the detectors commercially available, the refractive index is the most used for carbohydrates detection. It is a universal detector and its principle involves the measurement of the change in the refractive index of the column effluent [37,38]. Another detector used for carbohydrates is the evaporative light scattering detector (ELSD). It is

a universal detector suitable to detect most compounds less volatile than the mobile phase, such as sugars, not relying on the optical properties of the analyte. It works by spraying the eluate to remove the mobile phase by evaporation, then light is irradiated on the solute, and the light scattered is detected. In comparison with other detector, it presents stable baseline and high accuracy [39–41].

In the literature, HPLC was used to quantify sugars in sugarcane molasses [42], in potato juice [43], in bayberry juice [44], and in sweet sorghum juice [8].

2.3. Design of Experiments

One of the main challenges when carrying out experiments is to determine the influence of the factors, such as time, temperature, volume, also known as independent variables, over a response of interest, such as a reaction yield, for instance, known as dependent variable [45].

Design of experiments (DOE) is a multivariate statistical approach introduced for the first time in the 1920s. It can be applied to screen for the effects of the individual and interaction factors, optimize the system, save time, and obtain a model that explains the system. In addition, it could be used to predict the response of interest using different variable levels not studied [46,47]. DOE method requires the user to define minimum (-1) and maximum (+1) levels to study the system. Those levels must be carefully defined to generate suitable information [48].

There are several DOE techniques established in the literature. The most used techniques are the two-level full and fractional factorial designs, used when a linear relationship describes the system, and response surface methodologies (RSM), such as central composite design (CCD), Box-Behnken Design, and Doehlert design, used to describe the linear and quadratic behaviors of the system. For RSM methodologies, quadratic terms are added along with the minimum and maximum levels. Normally, central points (0) are added to circumvent the need to repeat all the design runs [49].

2.4. Near-infrared Spectroscopy

Figure 3 shows the electromagnetic spectrum, accentuating the ultraviolet, visible, NIR, mid-infrared (MID), far-infrared (FAR) regions. NIR region starts after the visible region and comprehends the initial portion of the infrared region, ranging from 780 to 2500 nm. MID region ranges from 2500 to 25000 nm and FAR from up to 1,000,000 nm [50].

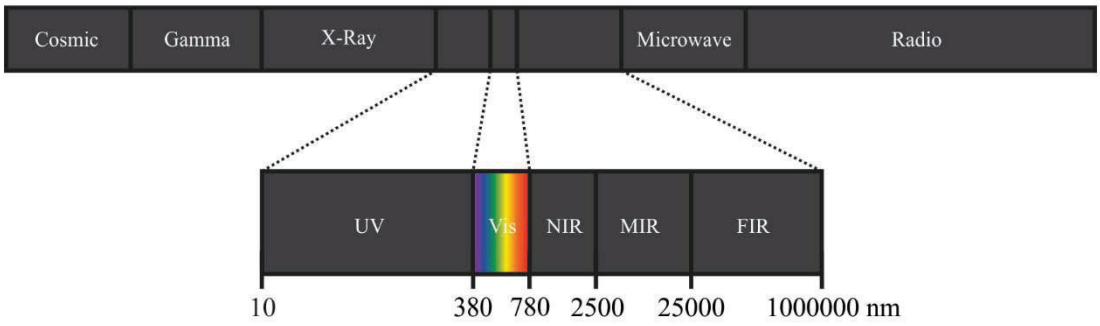

Figure 3. Electromagnetic spectrum highlighting the UV-Vis-NIR-MID-FAR infrared regions.

Figure 4 shows the molecular vibrational modes that absorbs the NIR radiation. Each vibration has a unique frequency and the location of the frequency is associated to a molecular structure. Stretching modes is the variation of the bond's length, it can be symmetric or asymmetric. Bending modes are the variation or change in the bond angle, it can be symmetric, rocking and wagging, and asymmetric, scissoring and twisting [51].

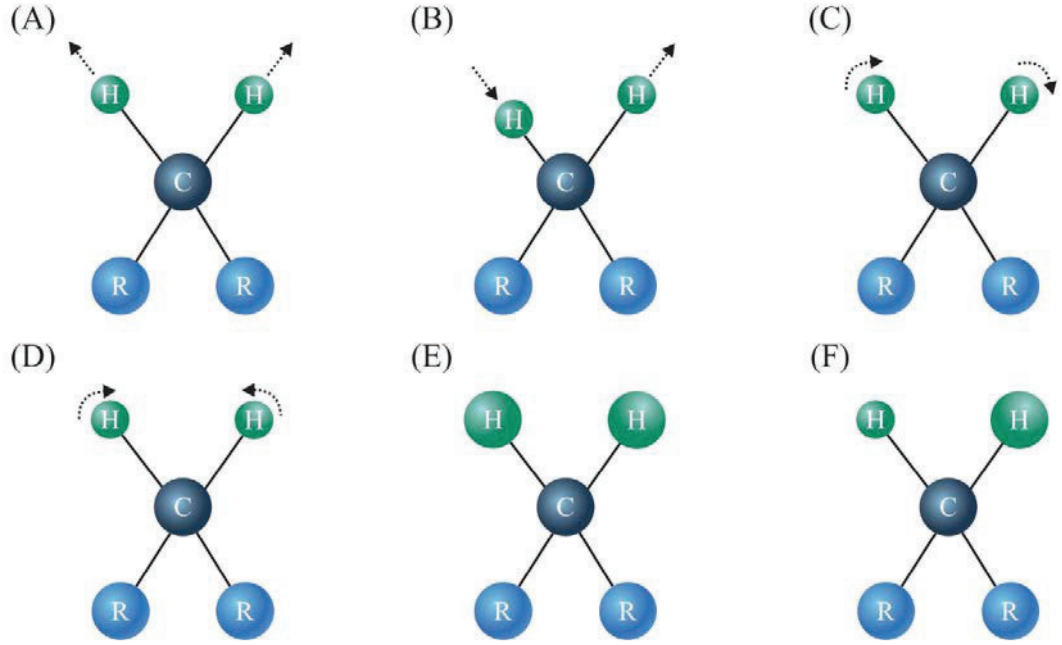

Figure 4. Vibrational modes that absorb NIR radiation, symmetrical (A) and asymmetrical (B) stretching and the bending modes, rocking (C), scissoring (D), out of plane wagging (E), and out of plane twisting (F).

Vibrational spectroscopy based on the idea that, when exposed to a light source, halogen lamp (~380 to ~2500 nm), the molecule is promoted from the fundamental state to a higher state, more energetic, which increases the atom-to-atom bond vibrations, especially stretching and bending. NIR spectrum is comprised between the wavelengths from 780 to 2500 nm and is composed by overtone and combinations (addition or subtraction of the energies) of two or more vibrations. The overtones and combinations regions from the stretching modes of O-H, C-H, S-H, and N-H can be observed in **Figure 5** [51,52].

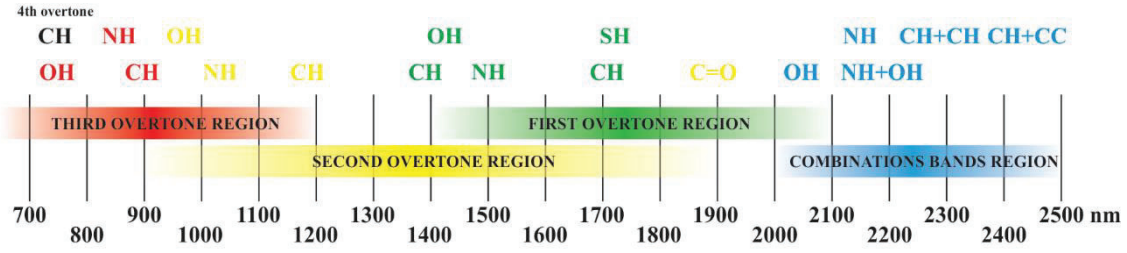

Figure 5. Overtones and combinations spectral regions.

As the bands observed in the NIR region are broad and overlap one another, making the spectra very complex, assigning these bands are near impossible, therefore NIR could not be used to compound identification. However, NIR together with multivariate tools, such as multivariate regression, could be used to quantify chemical and physical properties in complex samples [50,53,54].

NIR is broadly used to determine chemical and physical properties of foods, soils, drugs, plants, and others. It is a fast and nondestructive method and involves little or no sample preparation [55,56]. It has been applied to determine sugars in sweet sorghum juice [8], BRIX and POL values in sugarcane byproducts [57–59], sugars in bayberry juice [44], sugars in the mandarin fruit rind [60], sugar content in potato tubers [43], and sucrose in Arabica coffee beverages [61].

2.5. Multivariate Regression

Multivariate regression is one of the most important chemometric methods. It is also known as multivariate calibration and establishes a relationship between a set of independent variables, such as spectra or chromatograms, and one or more properties (dependent variables), such as concentrations [48]. This data analysis is a powerful method for building empirical models capable to predict, indirectly, important properties. It is often used to determine the concentration of one or more analytes in a complex mixture [62]. The aim of multivariate regression is to replace a slow and expensive method by a rapid and easy approach [46].

Normally, inverse regression is used to solve a linear system, as represented in **Equation 2**. Where **X** is a matrix $n \times m$ and n is the number of rows (samples) and m is the number of columns (variables). The **X** matrix represents the independent variables and is normally built, in chemistry field, using instrumental responses, as shown in **Figure 6**.

$$\mathbf{y} = \mathbf{Xb} \qquad \textbf{Eq. 2}$$

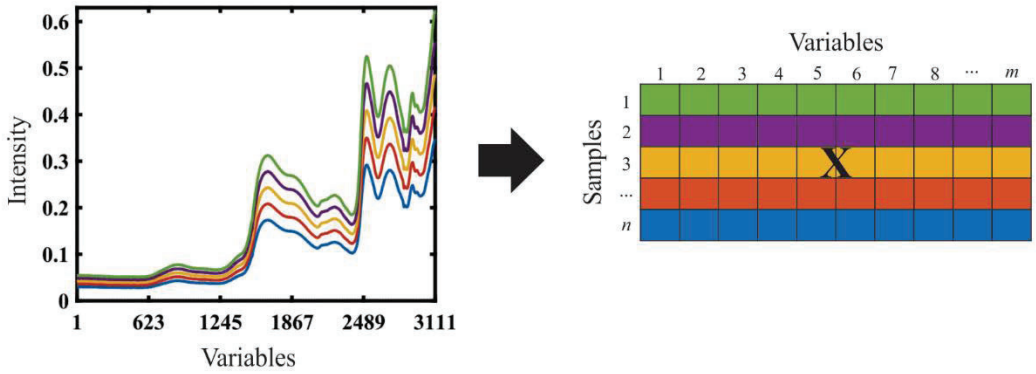

Figure 6. Representation of **X** matrix obtained from the instrumental response (spectra).

The dependent variable is the interest property for each sample (*e.g.,* concentration, biological activity, etc.) and is represented by the **y** vector. The vector **b** is the solution of the linear system and it is also known as regression coefficient. **Figure 7** exemplifies the independent and dependent variables and the regression coefficient.

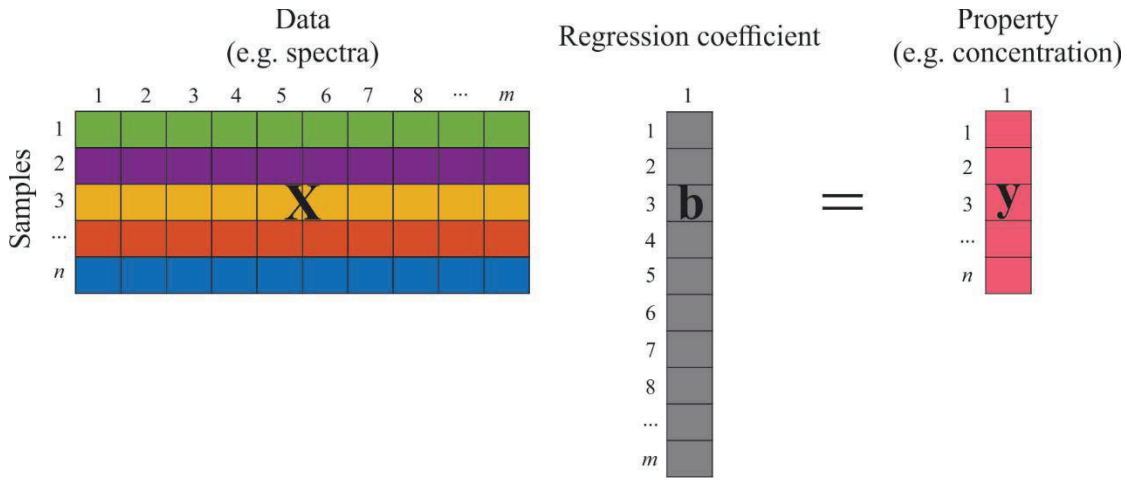

Figure 7. Representation of the independent variables (X) and the dependent variable (y).

The regression model consists of two steps. Initially, the regression coefficient, $\hat{\mathbf{b}}$, is calculated using the calibration set. Then a validation or external set, is used to assess the performance of the regression model. **Equation 3** shows the values predicted, $\hat{\mathbf{y}}$, using the external set, **Xp**, and the regression coefficient, $\hat{\mathbf{b}}$, estimated previously. To assess the performance, the predicted parameter, $\hat{\mathbf{y}}_\mathbf{p}$, is compared with the experimental parameter measured a priori, $\mathbf{y}_\mathbf{p}$, and therefore the quality parameters can be calculated. The model validation is normally evaluated using the root mean square error, *RMSE* (**Equation 4**), and the correlation coefficient, *R* **(Equation 5**). *RMSE* must be as low as possible, at least 10 times smaller than the minimum value analyzed, without overfitting the model and *R* must be as close to 1 as possible [63–67].

$$\hat{\mathbf{y}}_p = \mathbf{Xp} \cdot \hat{\mathbf{b}} \qquad \textbf{Eq. 3}$$

where $\hat{\mathbf{y}}_p$ is the values predicted, $\mathbf{Xp}$ is the matrix of independent variables used to validation/prediction, and $\hat{\mathbf{b}}$ is the regression coefficient.

$$RMSE = \sqrt{\frac{\sum(\mathbf{y} - \hat{\mathbf{y}})^2}{I}} \qquad \textbf{Eq. 4}$$

where $\mathbf{y}$ is the values observed experimentally, $\hat{\mathbf{y}}$ is the values predicted by the model, and I is the number of observations.

$$R = \frac{\sum(\mathbf{y}_o - \bar{\mathbf{y}}_o)(\mathbf{y}_p - \bar{\mathbf{y}}_p)}{\sqrt{\sum(\mathbf{y}_o - \bar{\mathbf{y}}_o)^2}\sqrt{\sum(\mathbf{y}_p - \bar{\mathbf{y}}_p)^2}} \qquad \textbf{Eq. 5}$$

where $\mathbf{y}_o$ is the value observed, $\bar{\mathbf{y}}_o$ is the average value observed, $\mathbf{y}_p$ is the value predicted by the model, and $\bar{\mathbf{y}}_p$ is the average values predicted.

The data acquisition is an important step and must be done carefully. Before building the model, the data are usually pretreated. This step is crucial to eliminate variations acquired during data acquisition, such as noise, and increase signal to noise ratio [67].

There are several multivariate regression methods. Among them, partial least squares regression (PLS) plays a major role as one of the most utilized modeling methods for chemical data [68].

2.6. Partial Least Squares

The PLS were first proposed in the 1960s by the statistician Hermann Wold for economic predictions. During the 1970s, Wold's son, Svante Wold, and Harald Martens expanded the application of PLS to chemical data.

There are two different PLS methods available, PLS1 and PLS2. In PLS1, a model is calibrated for each property of interest. With PLS2, one calibration model is built for all properties, all together [69]. In this work, PLS1 was used and from now on it will be named only PLS. In this method, modelling the concentration information is as important as modelling the experimental information, maximizing the covariance between the two blocks and minimizing their errors [46,70].

Several PLS algorithms are available in the literature and chemometric softwares used to solve multivariate systems, such as NIPALS [64], modified NIPALS [71], the Kernel [72], bidiagonal PLS [73,74], and SIMPLS [75]. All algorithms reach the same results,

however, the bidiagonal PLS requires the least amount of computational time [76]. Therefore, the bidiagonal PLS algorithm was chosen to be used in this work.

To solve the system represented before in **Equation 2**, the matrix **X** is decomposed in three other matrixes considering the information from **y**, as shown in the **Equation 6,** below.

$$\mathbf{y} \rightarrow \mathbf{X} = \mathbf{URV^t} \qquad \textbf{Eq. 6}$$

where **UR** is the matrix of scores and **V** is the matrix of loadings.

Then, a new truncated matrix $\hat{\mathbf{X}}$ is built, as shown in **Equation 7**, considering the number of latent variables, *LV*, which are variables inferred based on the variables observed that describes the covariance of the data.

$$\hat{\mathbf{X}} = \mathbf{U}_{LV}\mathbf{R}_{LV}\mathbf{V}^{\mathbf{t}}_{LV} \qquad \textbf{Eq. 7}$$

After that, the Moore-Penrose pseudoinverse can be estimated, and the regression coefficient can be found through the PLS solution, **Equation 8**.

$$\hat{\mathbf{b}} = \mathbf{U}_{LV}\mathbf{R}^{-1}_{LV}\mathbf{V}^{\mathbf{t}}_{LV}\mathbf{y} \qquad \textbf{Eq. 8}$$

Choosing the *LV* is a crucial step to avoid underfitting or overfitting of the model. In general, the cross-validation step (CV) is applied for choosing the ideal number of latent variables. In the CV, considering all samples used to calibrate de model, a new model is built without a certain number of samples that are latter predicted by the model, this step is repeated up until all the samples used to calibrate are left out and predicted. Thereafter, the *RMSE* of cross-validation (*RMSECV*), is calculated using **Equation 9** for each latent variable and then the latent variable corresponding to the minimum error is chosen.

$$RMSECV = \sqrt{\frac{\sum(\mathbf{y} - \hat{\mathbf{y}})^2}{I_{cv}}} \qquad \textbf{Eq. 9}$$

where, **y** is the observed value (experimental), $\hat{\mathbf{y}}$ is the estimated value (predicted by the model), and I_{cv} is the number of samples used in the cross-validation.

There are four CV methods commonly used, leave-one-out, continuous blocks, venetian blinds, and random blocks. Among them, random blocks is the most rigorous method, randomly separating the calibration set. For this work, random blocks method was chosen as the CV method.

3. Method

3.1. Support Material Selection

A new method to predict sugars and technological properties using sugarcane juice is proposed. This new method consisted of drying the sugarcane juice sample on a support and then acquiring the NIR spectrum of the dried sample-support. Two methods were proposed, one drying the sample on an oven and another were the sample would be dried using a heating plate.

For the method using an oven, a total of nine supports were tested. **Table 1** shows the name and the description of the supports assessed.

Table 1. Description of the selected supports for NIR assessment.

Support material	Description
GRO	Thick Couche Paper
FIN	Thin Couche Paper
SIL	Silicon
EVA	Ethylene Vinyl Acetate
TEF	Teflon
PAR	Paper Kraft
ESC	Dark Paper Card (Kraft side)
CLA	Bright Paper Card (Kraft side)
GCS	Glass Cover Slips

For the method using a heating plate, only glass cover slips (GCS) were used.

Figure 8 shows the supports assessed for the methods using the oven. All supports had a radius of 25 mm, except TEF with a radius of 13 mm. The supports were assessed considering whether the support absorbs or not the sample, its stability under temperatures above 50º Celsius, and its homogeneity. **Figure 9** shows the glass cover slip assessed as a support for the methods using the heating plate. The GCS had 24 mm of width and 24 mm of height.

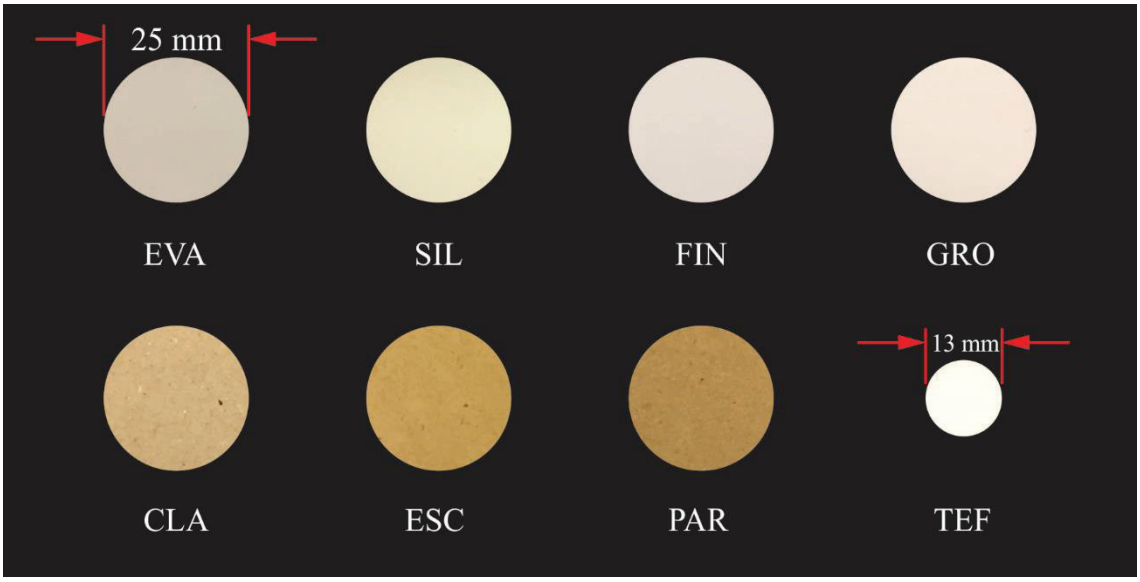

Figure 8. Supports assessed for the methods using the oven.

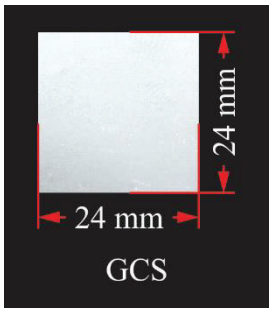

Figure 9. Support assessed for the method using the oven and heating plate.

3.2. Sample Selection

The samples utilized in this study belongs to the Sugarcane Breeding Program of Universidade Federal de Viçosa (PMGCA-UFV) and RIDESA (Inter-University Network for the Development of Sugarcane Industry) germplasm banks localized in the experimental zone at Universidade Federal de Viçosa (UFV), Viçosa, state of Minas Gerais, Brazil.

The sugarcane stalks were harvested and milled. An aliquot of 500 g was pressed by means of a hydraulic press and the juice extracted was collected. The conventional technological analyses were carried out and the remaining juice were kept in tubes and stored in a minus 80º Celsius freezer for posterior use. **Figure 10** illustrates the process for obtaining the sugarcane juice.

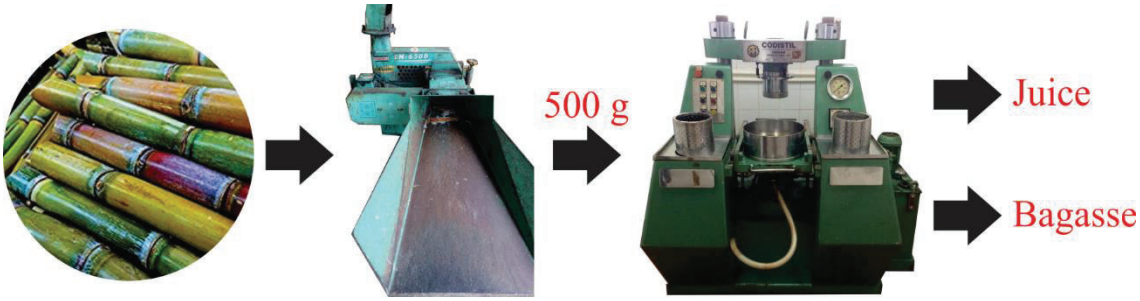

Figure 10. Process for obtaining sugarcane juice.

For this work, 200 sugarcane juice samples were selected considering their BRIX and POL values among 627 samples available using the Kennard-Stone algorithm [77].

3.3. Technological Properties

The technological properties, BRIX, POL, and fiber content were acquired a priori after the harvesting. The analyses were carried out at the PMGCA-UFV site.

BRIX values were acquired using a refractometer (HANNA Instruments) and the POL values using a saccharimeter (ACATEC SDA 2500). For POL analysis, Octapol® was used to clarify the juice. **Figure 11** shows the instruments utilized.

For BRIX analyses, the refractometer was tared using water and then an aliquot of the juice was put on the sampling windows and the BRIX value was read.

For POL analyses, an aliquot of the clarified juice was introduced in the saccharimeter sample input and the POL values were acquired.

Figure 11. Saccharimeter (A) and refractometer (B).

3.4. Drying

3.4.1. Oven Method

On the NIR spectral range, water absorbs radiation around 1400 and 2100 nm, regions were the sugars also absorbs. Since water accounts for about 86% of the sugarcane juice constitution, it acts as a strong interferent, masking the sugars absorption bands. Therefore, the proposal was to remove the water from the sample.

A CCD was used to optimize the drying process. Three factors (independent variables) were assessed, volume of sample, drying temperature, and time. Mass loss was chosen as the dependent variable. Three samples, containing high, medium, and low sucrose content using BRIX, as shown in **Table 2**, were chosen to assess the drying parameters and optimize the drying process to best represent all samples. The CCD factors and levels are shown in **Table 3**.

Table 2. Samples used to assess the drying parameters.

Sample	BRIX value
62	5.40
19	11.50
78	20.73

Table 3. Factors and levels used in the central composite design.

Factors	Levels				
	-α	**-1**	**0**	**1**	**+α**
Temperature (ºC)	46.36	60	80	100	113.64
Time (min)	19.55	40	70	100	120.45
Volume (µL)	43.18	50	60	70	76.82

A method to dry the sample was proposed, where the sugarcane juice would be dried on top of a support. To find the mass loss, the support was weighted (*P0*), then the balance with the support was tared, the support plus sample was weighted (*P1*), the support was dried in the oven, and then the support was weight after drying (*P2*). The mass loss (*ML*) was calculated as show in **Equation 10**. **Figure 12** shows a schematic of the drying method to calculate the sample mass loss used as dependent variable in the CCD.

$$ML = \frac{P1 - (P2 - P0)}{P1} \times 100 \qquad \textbf{Eq. 10}$$

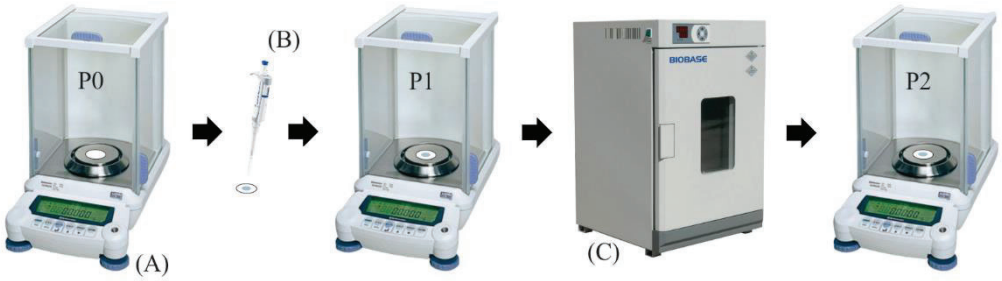

Figure 12. Drying method to calculate the sample moisture used as dependent variable in CCD. (A) Analytical balance, (B) automatic pipette, and (C) air circulating oven.

The codified and decodified levels utilized for the optimization of the method using the oven are presented in **Table 4**.

Table 4. CCD for optimization of the drying method using the oven.

Run	Temperature (ºC)		Time (min)		Volume (µL)	
	Codified	Decodified	Codified	Decodified	Codified	Decodified
1	-1	60	-1	40	-1	50
2	-1	60	-1	40	+1	70
3	-1	60	+1	100	-1	50
4	-1	60	+1	100	+1	70
5	+1	100	-1	40	-1	50
6	+1	100	-1	40	+1	70
7	+1	100	+1	100	-1	50
8	+1	100	+1	100	+1	70
9	-α	46.3641	0	70	0	60
10	+α	113.636	0	70	0	60
11	0	80	-α	19.5462	0	60
12	0	80	+α	120.454	0	60
13	0	80	0	70	-α	43.1821
14	0	80	0	70	+α	76.8179
15 (C)	0	80	0	70	0	60
16 (C)	0	80	0	70	0	60
17 (C)	0	80	0	70	0	60
18 (C)	0	80	0	70	0	60
19 (C)	0	80	0	70	0	60

Figure 13 shows the schematics of the process developed for the oven method. The samples were pipetted to the supports placed in an iron tray, and them placed into the oven. For this method, 60 samples were dried at a time.

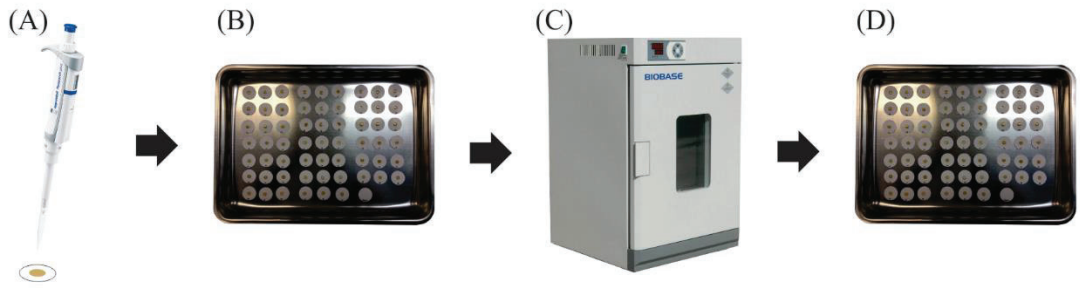

Figure 13. Schematics of the oven method. (A) Automatic pipette, (B) iron tray before drying, (C) air circulating oven, and (D) iron tray after drying.

3.4.2. Heating Plate Method

The method to dry the samples on the heating plate consisted of placing the GCS support on the heating plate, then pipetting 60 µL of sample to the support, and drying it until there were no water, visually. Several temperatures were tested to find the proper one. **Figure 14** shows the schematics of the process developed for the heating plate method. For this method, three samples were dried at a time.

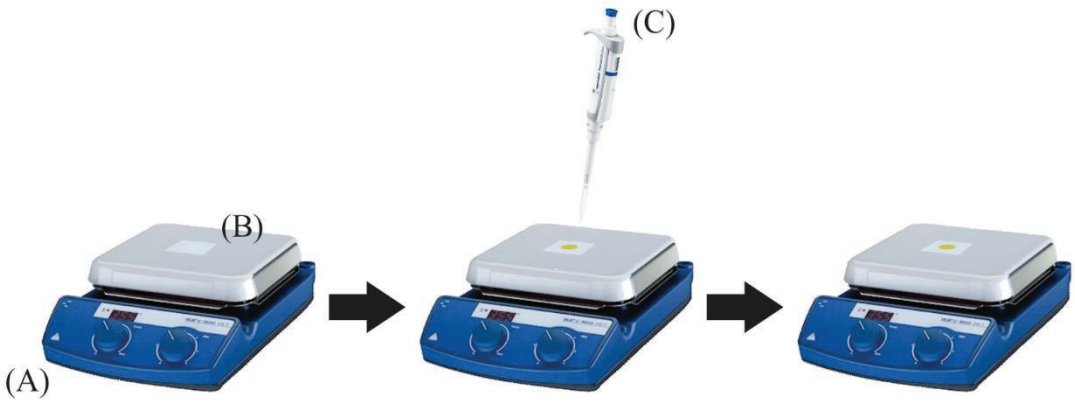

Figure 14. Schematics of the heating plate method. (A) Heating plate, (B) glass cover slip, and (C) automatic pipette.

3.5. High Performance Liquid Chromatography

3.5.1. Method

The sugarcane juice samples were centrifuged, diluted 12 times, and then filtered through a nylon syringe filter (pore size 0.45 µm and 25 mm diameter) to vials. The determination of the sugars content was made using a Shimadzu 20AT Prominence instrument and the software LabSolutions. An HPX-87P (Bio-Rad® Aminex) column (9 µm particle size, pH ranges from 5 to 9, 300 mm height, and 7.8 mm width) was used to separate the sugars. The mobile phase was constituted of 20% acetonitrile:water. The

separation parameters consisted of temperature at 80 ºC, flow of 0.850 mL/min, and injection volume of 10 µL. The ELSD was used as sugar detector.

3.5.2. Univariate calibration

To quantify the sugars in the samples a univariate calibration (analytical curve) was carried out using standard solution of sucrose, glucose, and fructose. The univariate quadratic regressions were built using twelve levels of concentration from 0.84 to 17.86 mg mL^{-1} for sucrose; ten levels of concentration from 0.14 to 4.06 mg mL^{-1} for glucose and, ten levels of concentration from 0.34 to 4.06 mg mL^{-1} for fructose. Each level was performed in triplicate to assess the accuracy and precision of the calibration. The sugar content in the solutions were separated and analyzed by HPLC as described previously.

The univariate quadratic regressions were built through inverse regression, where the model is built with the concentration values on the ordinate axis (y axis) and the area measured values on the abscissa's axis (x axis). The model is represented in **Equation 11**.

$$\hat{y} = b_0 + b_1\mathbf{x} + b_2\mathbf{x}^2 \qquad \textbf{Eq. 11}$$

where $\hat{\mathbf{y}}$ is the area of the peaks, b_0, b_1, and b_2 are the coefficients, and $\mathbf{x}$ is the sugar concentration (mg mL^{-1}).

The sugars content in the samples were obtained considering the dilution factor of 12 times.

3.5.3. Validation

The methods were validated considering the linearity, accuracy, limit of detection (*LOD*), and limit of quantification (*LOQ*) [78]. The linearity was assessed through *R* value of the analytical curve built. The precision was assessed with replications of measurements under the similar conditions, it was expressed as the relative standard deviation (*RSD*). The *RSD* was calculated according to **Equation 12**. Where ∂ is the standard deviation of the measurements and $\bar{x}$ is the average measurement.

$$RSD = \left(\frac{\partial}{\bar{x}}\right) \times 100 \qquad \textbf{Eq. 12}$$

LOD is the least amount of an analyte that can be differentiated from the noise, *i.e.*, the least amount detectable. Normally, *LOD is* calculated considering the concentration that generates a signal three times larger than standard deviation of the

noise measured using a blank (∂) over the linear regression coefficient (b_1), as show in **Equation 13**.

$$LOD = \frac{3\partial}{b_1} \qquad \textbf{Eq. 13}$$

LOQ is the least amount of an analyte that can be detected accurately. It can be also estimated using a blank injection. Usually, *LOD* is a value 10 times larger than the standard deviation of a signal generated by a blank, as shown in **Equation 14**.

$$LOQ = \frac{10\partial}{b_1} \qquad \textbf{Eq. 14}$$

3.6. Near-infrared Spectroscopy

The NIR analyses were carried out by means of a Thermo Scientific Antaris II with Fourier Transform spectrometer and integration sphere. The range from 1000 to 2500 nm with increment of 0.48 nm was investigated. The spectra were collected using the TQ Analysis software. The mean of 32 scans performed for each sample was stored. The spectra were acquired in reflectance mode as $\mathbf{log}\left(\frac{\mathbf{1}}{Ref}\right)$, where *Ref* is the reflectance. The support (GRO or GCS) without sample and a gold support (blank) was used as reference (total reflectance), for this prior to each NIR analysis, the instrument background was set as the blank NIR spectrum. The spectra of the dried sample-support were acquired in triplicates, resulting in 600 samples.

3.7. Multivariate Regression

The NIR spectra were acquired through the instrument software and imported to MATLAB R2019a (MathWorks, Natick, USA) using algorithms written in the Multivariate Chemical Data Analysis Laboratory (MCDA Lab). A matrix **X** (independent variable), was built using the spectra, were the matrix rows and columns represents the samples and wavelengths analyzed, respectively. A vector **y** was built for each property analyzed, *i.e.*, sucrose, glucose, fructose, BRIX, POL, and fiber. The vector **y** contains the same number of rows as the matrix **X**. The matrix **X** and all the vectors **y** are represented in the **Figure 15**. To build the regressions models, the PLS regression was used. The algorithms used to build and validate the models were written at the MCDA Lab in the MATLAB environment.

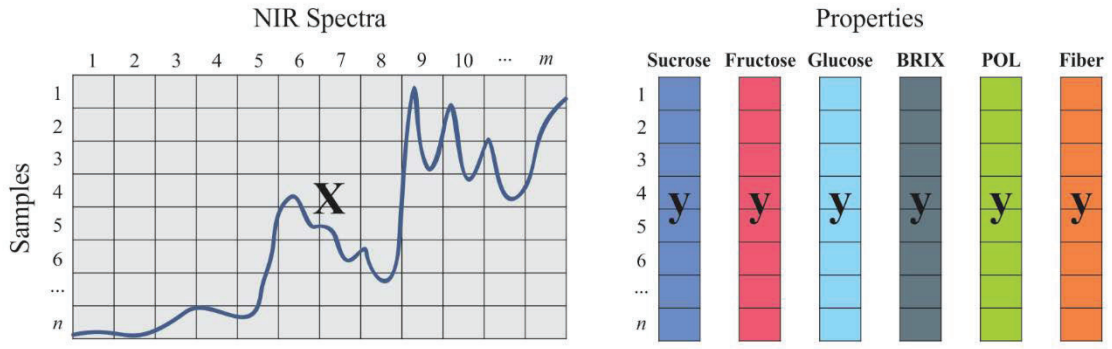

Figure 15. Matrix **X** built using the NIR spectra and the vectors y built using the properties of interest, sucrose, fructose, glucose, BRIX, POL, and fiber.

By using the Kennard-Stone algorithm [77]the matrix **X** were split into two new matrices, calibration matrix (**Xc**) and validation matrix (**Xp**). The regression model was built using the matrix **Xc** and the matrix **Xp** was used to validate the model.

The **y** variable was mean centered for all the properties. Different pretreatments were studied for the matrix **X**. Besides the raw data, were performed two preprocessing 1) mean center and 2) autoscale, and nine transformations: 1) smoothing, 2) 1st derivative, 3) 2nd derivative, 4) multiplicative scatter/signal correction (MSC), 5) orthogonal signal correction (OSC), 6) detrend, 7) normalize, 8) baseline, and 9) standard normal variate scaling (SNV). The combinations of two and three pretreatments were also studied. Smoothing, 1st derivative, and 2nd derivative was applied using the odd widths from 3 to 23. MSC was applied using the least squares regression between each spectrum and the reference spectrum and using the median of the ratio between a spectrum and the reference. Normalize was applied normalizing the data to unit area (norm 1), to unit length (norm 2), and to maximum value (norm infinite). OSC was applied using 1 to 25 components. The models were built in MATLAB R2019a (MathWorks, Natick, USA) environment using PLS regression. Random cross-validation was used with 10 splits.

On total, 11046 different pretreatments combinations were applied. **Figure 16** shows the pretreatments applied to the X matrix and their combinations. The models were built using latent variables from 1 to 10, which resulted in 110460 different models. An algorithm was written to perform all calculations automatically. The model that presented the lowest *RMSECV* value was chosen for each property and drying method.

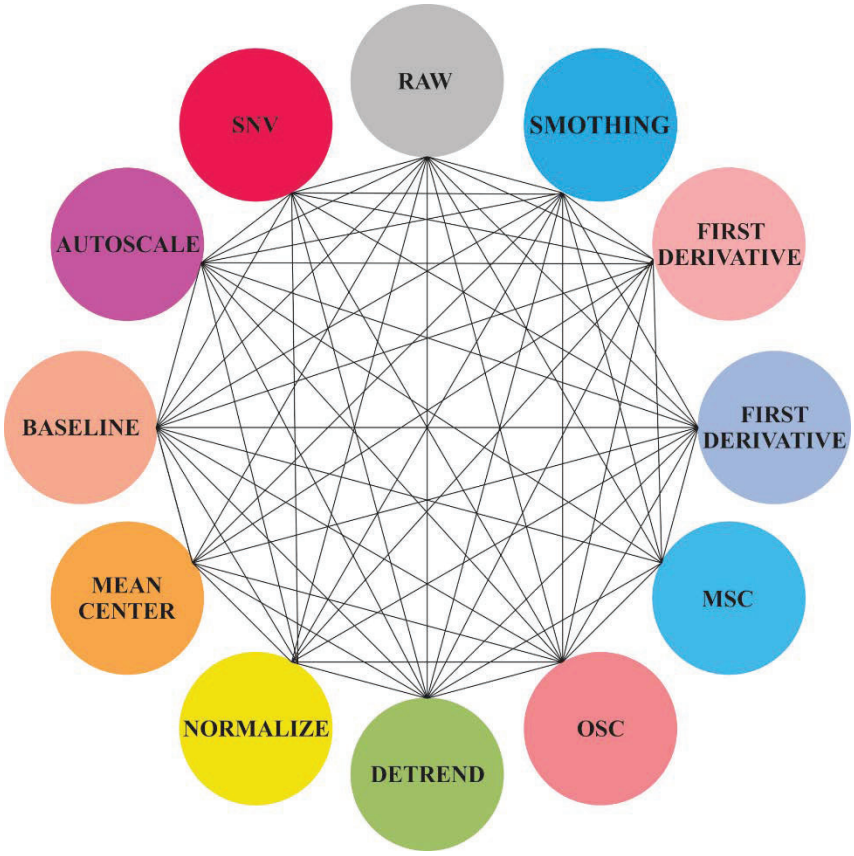

Figure 16. Scheme of the pretreatment combinations applied to the matrix **X**.

The quality of the regression models was evaluated considering the *RMSE* and *R* of calibration (*RMSEC, Rc*), cross-validation (*RMSECV, Rcv*), and validation (*RMSEP, Rp*), sensibility (*SEN*), selectivity (*SEL*) and *LOD* related to the concept of net analyte signal (**NAS**), defined as the analytical signal orthogonal to the interferents presents in the samples [79].

SEN is defined as the signal fraction that increases one unity of concentration to the analyte, *i.e.*, how strong the analyte responds. *SEN* is calculated using the **Equation 15** [80]. Where $\|\mathbf{b}\|$ is the Euclidean norm of the regression coefficient.

$$SEN = \frac{\mathbf{1}}{\|\mathbf{b}\|} \qquad \textbf{Eq. 15}$$

The analytical sensibility (γ) is the sensibility of the method estimated by means of the standard deviation (∂) of the NAS for the reference spectra, as shown in the **Equation 16**.

$$\gamma = \frac{SEN}{\|\partial\|} \qquad \textbf{Eq. 16}$$

The inverse of the analytical sensibility (γ^{-1}) establishes the lowest concentration that can be differentiated by the method.

SEL is the capacity to assess the analyte in the presence of interferents. This parameter varies from 0 to 1, where 0 means the analytical signal contain information of the interferent and 1 means no interferents present in the sample. The selectivity is defined by the net analytical signal (**NAS**) and instrumental response (**X**) ratio, as shown in **Equation 17** [81]. The final selectivity (*SEL*) is the average value of the **SEL** calculated for each sample ***i*.**

$$\mathbf{SEL}_i = \frac{\mathbf{NAS}_i}{\|\mathbf{X}_i\|} \qquad \textbf{Eq. 17}$$

LOD is the least amount of analyte that can be detected and quantified. *LOD* can be calculated using **Equation 18** [82].

$$LOD = \frac{\hat{a} + (3\partial)}{\hat{b}} \qquad \textbf{Eq. 18}$$

where ∂ is the standard deviation of the reference signal, $\hat{a}$ is the intercept of the regression equation, and $\hat{b}$ is the regression coefficient.

Each model was verified for chance correlation [83,84]. The **y** vector (dependent variables) was randomized 100,000 times and models were built using these randomizations. The correlation of the models built were evaluated. If the pearson correlation parameter of the model built using the authentic **y** is isolated from those of the models using the randomized **y**, then the model not occurred by chance.

3.8. Analytical Frequency

The analytical frequency of the procedures was estimated based on the time spend to pipette the sample to the support, dry the sample, and collect the NIR spectra.

3.9. Ranking

Targeting the application for genotypes selection, the performances of BRIX, POL, and sucrose predicted by the model were evaluated using the sucrose quantifyed via HPLC as reference. For that, the same 100 samples were selected for each property and then ranked according to their respective values. The top 30 samples for each property were selected. Next, the samples chosen for each property were then compared with the samples chosen using the reference sucrose content. True positive (*TP*) was assigned to the samples that belong to the top 30 of the reference sucrose and to the top 30 of the property being analyzed. False positive (*FP*) was assigned to the samples that do not belong to the top 30 of the reference sucrose but belong in the top 30 of the property being

analyzed. True negative (*TN*) was assigned to the samples that do not belong to the top 30 of either the reference sucrose and the property being analyzed. False negative (*FN*) was assigned to the samples that belong to the top 30 of the reference sucrose but do not belong to the top 30 of the property being analyzed.

With the *TP*, *FP*, *TN*, and *FN* values, the efficacy of each property was assessed using the error (*E*), sensitivity (*S*), specificity (*Sp*), and accuracy (*A*), calculated by means of the **Equations 19, 20, 21, and 22**.

$$E = \left(\frac{FP + FN}{TP + FP + TN + FN} \right) \times 100 \qquad \textbf{Eq. 19}$$

$$S = \left(\frac{TP}{TP + FN} \right) \times 100 \qquad \textbf{Eq. 20}$$

$$Sp = \left(\frac{TN}{TN + FP} \right) \times 100 \qquad \textbf{Eq. 21}$$

$$A = (1 - E) \times 100 \qquad \textbf{Eq. 22}$$

4. Results and Discussion

4.1. Samples Selection

Figure 17 shows the graph BRIX vs. POL representing the selection of samples. It can be observed that the selected samples (red spheres) are representative of the total samples available (black spheres).

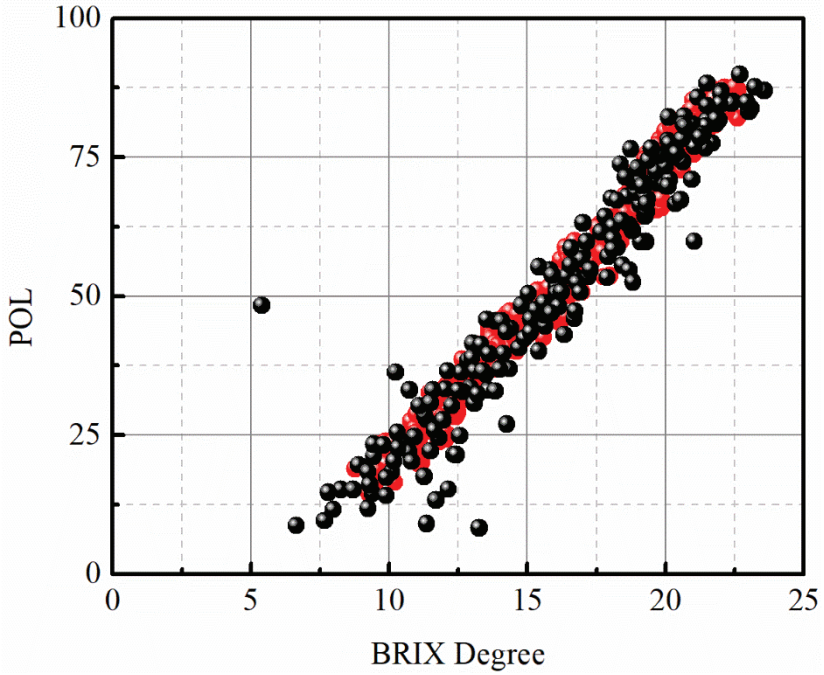

Figure 17. Graph BRIX vs. POL representing the samples selection where the red circles represent the total samples (627 samples) and the black circles represent the samples selected (200 samples).

Table 5 shows the information about the selected samples. From the 200 samples used to build the model, 69 samples, consisting of 62 genotypes, were collected in May, 2013; 84 samples, consisting of 78 genotypes, were collected in August, 2013; 14 samples, consisting of 14 genotypes, were collected in August, 2015; 15 samples, consisting of 15 genotypes, were collected in September, 2015; 9 samples, consisting of 9 genotypes, were collected in October, 2015; 9 samples, consisting of 9 genotypes, were collected in November, 2015. On total, considering all the samples, 133 different genotypes were used, 105 from 2013 and 28 from 2015.

Table 6 shows the samples selected, indicating the genotype and its harvesting period, used to build the calibration model. The model was built using 105 genotypes harvested in May and August of 2013 and 28 genotypes harvested in August, September, October, and November of 2015.

Table 5. Information about the samples.

Month	Period	Samples	Genotypes (Month)	Genotypes (Period)
May	**2013**	69	62	105
August		84	78	
August	**2015**	14	14	28
September		15	15	
October		9	9	
November		9	9	
Total		**200**		**133**

Table 6. Samples identification indicating the genotype and the harvest.

Sample	Genotype	Harvest	Sample	Genotype	Harvest	Sample	Genotype	Harvest	Sample	Genotype	Harvest
1	301	May 2013	**25**	329	May 2013	**49**	332	May 2013	**73**	320	August 2013
2	304	May 2013	**26**	336	May 2013	**50**	438	May 2013	**74**	323	August 2013
3	313	May 2013	**27**	410	May 2013	**51**	446	May 2013	**75**	316	August 2013
4	316	May 2013	**28**	P1	May 2013	**52**	424	May 2013	**76**	313	August 2013
5	313	May 2013	**29**	375	May 2013	**53**	348	May 2013	**77**	302	August 2013
6	302	May 2013	**30**	366	May 2013	**54**	386	May 2013	**78**	318	August 2013
7	319	May 2013	**31**	P1	May 2013	**55**	352	May 2013	**79**	315	August 2013
8	312	May 2013	**32**	387	May 2013	**56**	419	May 2013	**80**	310	August 2013
9	317	May 2013	**33**	430	May 2013	**57**	406	May 2013	**81**	317	August 2013
10	308	May 2013	**34**	416	May 2013	**58**	P1	May 2013	**82**	305	August 2013
11	325	May 2013	**35**	421	May 2013	**59**	404	May 2013	**83**	325	August 2013
12	339	May 2013	**36**	365	May 2013	**60**	464	May 2013	**84**	334	August 2013
13	359	May 2013	**37**	341	May 2013	**61**	467	May 2013	**85**	322	August 2013
14	322	May 2013	**38**	344	May 2013	**62**	465	May 2013	**86**	392	August 2013
15	342	May 2013	**39**	362	May 2013	**63**	455	May 2013	**87**	342	August 2013
16	P2	May 2013	**40**	439	May 2013	**64**	463	May 2013	**88**	P2	August 2013
17	P2	May 2013	**41**	355	May 2013	**65**	367	May 2013	**89**	331	August 2013
18	331	May 2013	**42**	349	May 2013	**66**	459	May 2013	**90**	340	August 2013
19	P2	May 2013	**43**	357	May 2013	**67**	456	May 2013	**91**	P2	August 2013
20	340	May 2013	**44**	P2	May 2013	**68**	P2	May 2013	**92**	389	August 2013
21	345	May 2013	**45**	414	May 2013	**69**	468	May 2013	**93**	345	August 2013
22	374	May 2013	**46**	371	May 2013	**70**	301	August 2013	**94**	328	August 2013
23	327	May 2013	**47**	368	May 2013	**71**	303	August 2013	**95**	327	August 2013
24	382	May 2013	**48**	354	May 2013	**72**	304	August 2013	**96**	370	August 2013

Table 6. (Continuation).

Sample	Genotype	Harvest	Sample	Genotype	Harvest	Sample	Genotype	Harvest	Sample	Genotype	Harvest
97	382	August 2013	**123**	332	August 2013	**149**	P2	August 2013	**175**	RB057145	Sep 2015
98	329	August 2013	**124**	384	August 2013	**150**	447	August 2013	**176**	RB987955	Sep 2015
99	330	August 2013	**125**	431	August 2013	**151**	469	August 2013	**177**	RB037017	Sep 2015
100	400	August 2013	**126**	440	August 2013	**152**	470	August 2013	**178**	RB037194	Sep 2015
101	385	August 2013	**127**	424	August 2013	**153**	P2	August 2013	**179**	RB987917	Sep 2015
102	366	August 2013	**128**	356	August 2013	**154**	RB967526	August 2015	**180**	RB047110	Sep 2015
103	379	August 2013	**129**	360	August 2013	**155**	RB977450	August 2015	**181**	RB047137	Sep 2015
104	430	August 2013	**130**	432	August 2013	**156**	RB987932	August 2015	**182**	RB985823	Sep 2015
105	417	August 2013	**131**	399	August 2013	**157**	RB975950	August 2015	**183**	RB016916	October 2015
106	369	August 2013	**132**	398	August 2013	**158**	RB006996	August 2015	**184**	RB987932	October 2015
107	341	August 2013	**133**	442	August 2013	**159**	RB93509	August 2015	**185**	RB01474	October 2015
108	362	August 2013	**134**	401	August 2013	**160**	RB988079	August 2015	**186**	RB961536	October 2015
109	367	August 2013	**135**	P1	August 2013	**161**	RB987955	August 2015	**187**	RB975157	October 2015
110	388	August 2013	**136**	433	August 2013	**162**	RB037017	August 2015	**188**	RB985476	October 2015
111	415	August 2013	**137**	P1	August 2013	**163**	RB037055	August 2015	**189**	RB93509	October 2015
112	397	August 2013	**138**	404	August 2013	**164**	RB988499	August 2015	**190**	RB988499	October 2015
113	355	August 2013	**139**	423	August 2013	**165**	RB966928	August 2015	**191**	RB047110	October 2015
114	349	August 2013	**140**	457	August 2013	**166**	RB987917	August 2015	**192**	RB016916	Nov 2015
115	391	August 2013	**141**	464	August 2013	**167**	RB027040	August 2015	**193**	RB977450	Nov 2015
116	377	August 2013	**142**	467	August 2013	**168**	RB037223	Sep 2015	**194**	RB987932	Nov 2015
117	357	August 2013	**143**	465	August 2013	**169**	RB975950	Sep 2015	**195**	RB961536	Nov 2015
118	411	August 2013	**144**	455	August 2013	**170**	RB961536	Sep 2015	**196**	RB975932	Nov 2015
119	P2	August 2013	**145**	444	August 2013	**171**	RB961552	Sep 2015	**197**	RB037055	Nov 2015
120	353	August 2013	**146**	367	August 2013	**172**	RB975157	Sep 2015	**198**	RB985596	Nov 2015
121	358	August 2013	**147**	459	August 2013	**173**	RB93509	Sep 2015	**199**	RB988499	Nov 2015
122	371	August 2013	**148**	456	August 2013	**174**	RB988079	Sep 2015	**200**	RB047110	Nov 2015

4.2. Support Material Selection

Ethylene vinyl acetate (EVA) support was not stable under temperatures over to 50 ºC. Silicon (SIL) support was too smooth to hold the sample. Bright paper card (CLA) and dark paper card (ESC) supports were not homogenous and absorbed the samples. Teflon (TEF) support surface was too small to hold the samples volume. Thin couche paper (FIN) absorbed the samples. Paper kraft (PAR) and thick couche paper (GRO) presented good performances. Visually, PAR composition was not homogeneous, as small particles could be seen in the support. GRO was homogenous and had high thickness, holding the sample longer without absorbing it. Thus, GRO support was chosen as a suitable support for the method using the oven.

GCS was not suitable for the method using the oven since the sample spread over its surface. However, when previously heated, the sample adhered to the glass without spread, for that reason GCS was the support chosen for the method using the heating plate.

4.3. Drying

4.3.1. Oven Method

Table 7 presents the runs of the CCD, the decodified factors (temperature, time, and volume), and the mass loss for samples 62, 19, and 78 for each run of the CCD.

Table 7. Mass loss of samples 62, 19, and 78 for each run of the CCD.

Run	Temperature (ºC)	Time (min)	Volume (µL)	Mass Loss of Sample		
				62	19	78
1	60	40	50	86.9%	84.0%	76.9%
2	60	40	70	86.8%	83.0%	76.1%
3	60	100	50	87.9%	85.3%	77.4%
4	60	100	70	87.3%	84.8%	77.4%
5	100	40	50	91.7%	89.4%	82.8%
6	100	40	70	89.6%	87.3%	80.5%
7	100	100	50	91.0%	89.4%	83.4%
8	100	100	70	89.4%	86.9%	80.3%
9	46.36	70	60	90.1%	87.2%	79.5%
10	113.63	70	60	94.8%	91.4%	84.1%
11	80	19.54	60	90.2%	87.2%	78.1%
12	80	120.45	60	90.8%	84.2%	78.5%

Run	Temperature (ºC)	Time (min)	Volume (µL)	Mass Loss of Sample		
				62	19	78
13	80	70	43.18	86.8%	85.6%	78.7%
14	80	70	76.81	88.3%	84.4%	77.8%
15 (C)	80	70	60	88.5%	85.4%	77.6%
16 (C)	80	70	60	88.1%	85.6%	78.2%
17 (C)	80	70	60	88.3%	84.8%	77.2%
18 (C)	80	70	60	87.6%	85.0%	77.7%
19 (C)	80	70	60	88.2%	85.2%	77.8%

The CCD pareto chart, **Figure 18**, considering the mean square residual error and significance level (α) of 0.05, indicated that for the three samples studied the linear and quadratic temperature factors were significative. For samples 19 and 78, the linear volume factor was significative, but much less than the temperature. Hence, the temperature has more influence over the mass loss.

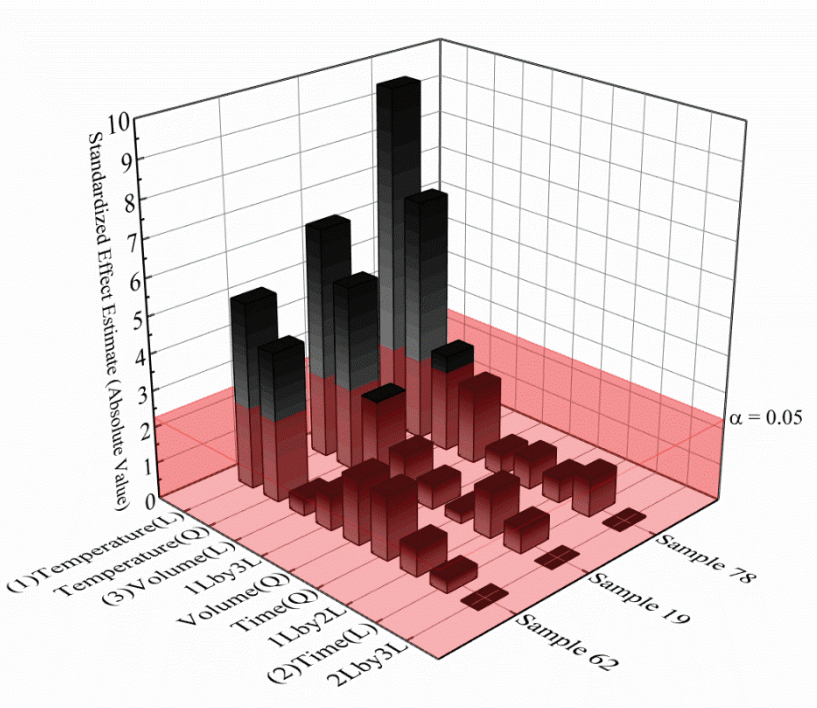

Figure 18. Pareto chart in three dimensions for the drying factors using the mean square residual and 0.05 of significance level.

Table 8 shows the estimated effects for linear and quadratic temperature, time, volume, and their interactions. Temperature varies linearly and quadratically and their effects for the three samples are positive, which means that increasing the temperature the mass loss increases. Volume varies linearly, for samples 19 and 78 only, which means that decreasing the volume would increase the mass loss.

Table 8. Effects of the factors estimated for each sample.

	Sample 62		Sample 19		Sample 78	
	Effect	*p*	Effect	*p*	Effect	*p*
Mean/Intercept	***0.8821***	0.0000	***0.8523***	0.0000	***0.7772***	0.0000
(1) Temperature (L)	***0.0302***	0.0006	***0.0335***	0.0001	***0.0397***	0.0000
Temperature (Q)	***0.0243***	0.0026	***0.0268***	0.0006	***0.0279***	0.0001
(2) Time (L)	0.0021	0.7257	-0.0035	0.5175	0.0044	0.3221
Time (Q)	0.0104	0.1127	0.0016	0.7696	0.0034	0.4302
(3) Volume (L)	-0.0028	0.6477	***-0.0120***	0.0479	***-0.0113***	0.0238
Volume (Q)	-0.0107	0.1049	-0.0035	0.5182	0.0027	0.5402
1L by 2L	-0.0060	0.4551	-0.0089	0.2282	-0.0036	0.5287
1L by 3L	-0.0074	0.3656	-0.0079	0.2803	-0.0115	0.0637
2L by 3L	-0.0002	0.9767	-0.0001	0.9928	0.0002	0.9681
MS Residual Error	1.19×10^{-4}		0.94×10^{-4}		0.59×10^{-4}	
R^2	0.85		0.89		0.94	
R^2 Adjusted	0.71		0.79		0.88	

Bold italics values indicate the significative effects.

The analysis of variance (ANOVA) showed that the quadratic regression adjusted the data with a coefficient of determination (R^2) of 0.85, 0.89, and 0.94 and an adjusted R^2 of 0.71, 0.79, and 0.88 for samples 62, 19, and 78, respectively.

Figure 19 presents the desirability profiles considering the CCD models for each sample and the factors studied (temperature, time, and volume). Considering that as the mass loss increases the water removal increases, the maximum mass removal was set as high desirability (set as 1). Therefore, for temperature, time, and volume, the optimal parameters, considering the levels studied, were defined as 113.64 ºC, 19.55 min, and 51.6 µL, respectively.

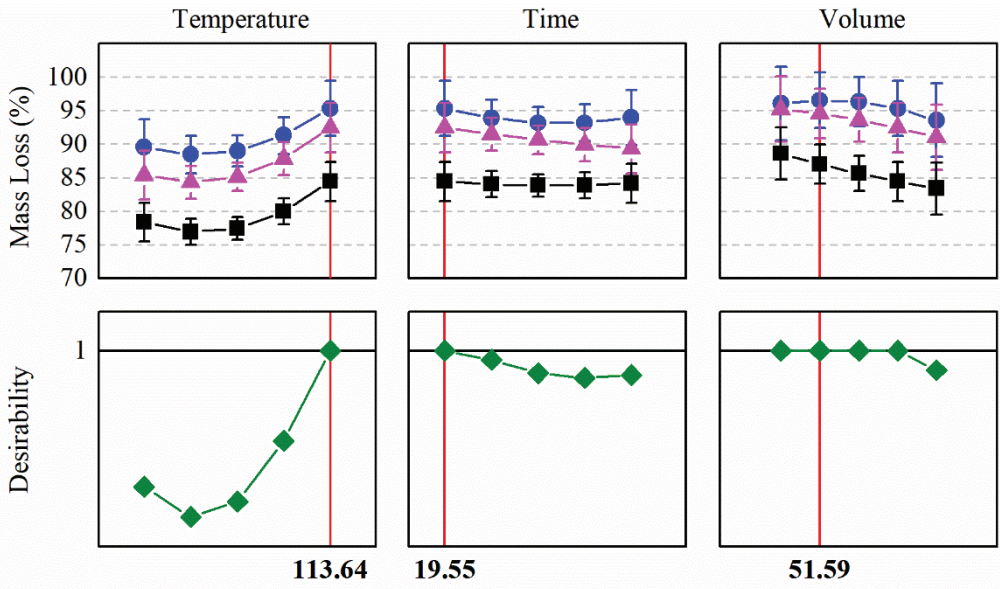

Figure 19. Profiles for predicted values and desirability for optimization of the moisture. Blue line represents sample 62, pink line represents sample 19, and black line represents sample 78.

It can be observed that the composition of the samples affects the drying process, as the sample with low BRIX lost more mass and the one with high BRIX lost less mass in the same period. In addition, after drying, the samples with low BRIX were solid and the ones with high BRIX resembled a tick molasses, indicating the water was not completely removed.

The temperature versus volume response surfaces for samples 62, 19, and 78 are presented in **Figure 20**. Their profiles confirm the previous findings. It can be observed that as the temperature increases, and the volume decreases, the mass loss increases.

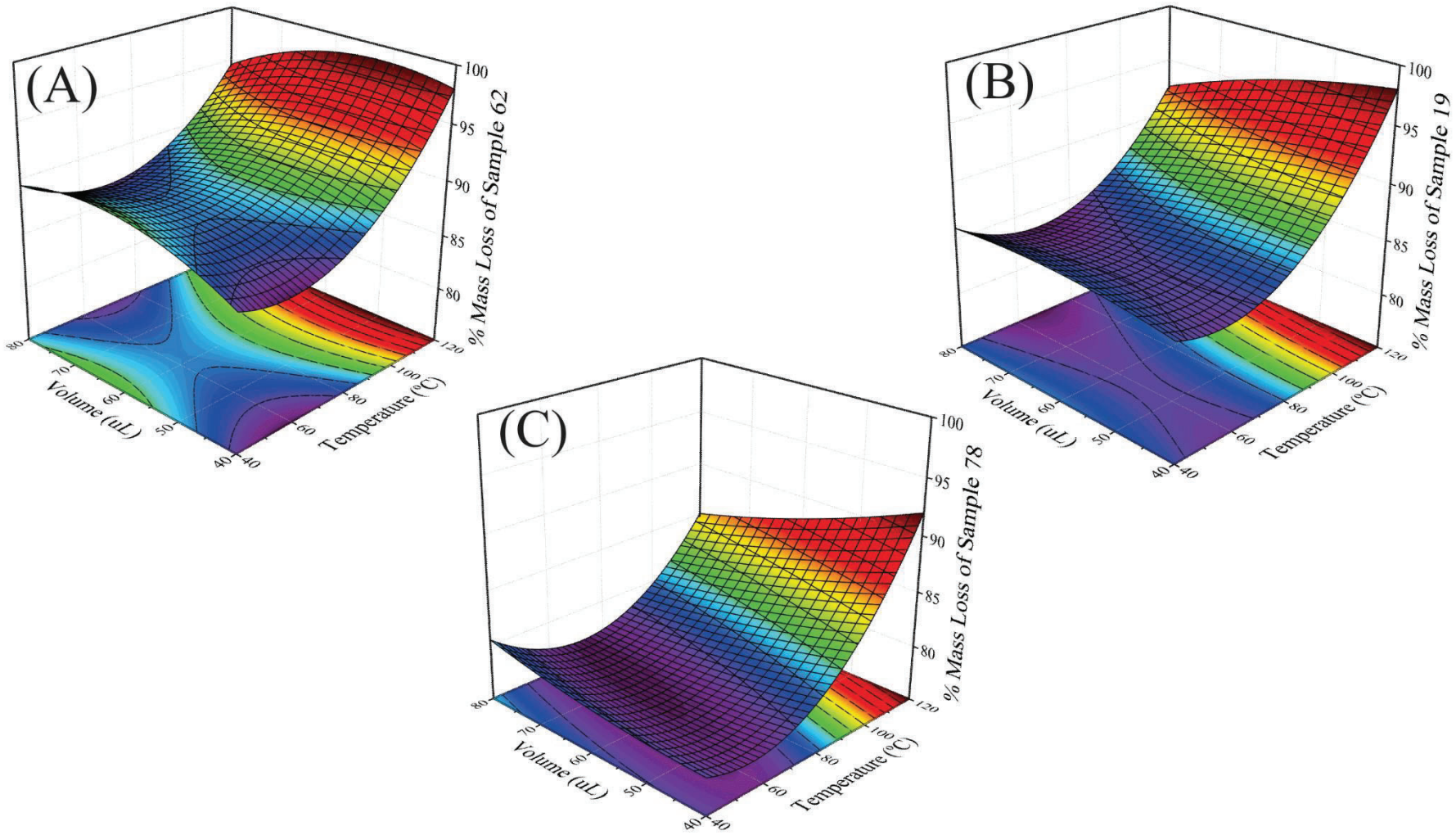

Figure 20. Temperature versus volume surface responses for mass loss of sample 62 (A), sample 19 (B), and sample 78 (C). Time were fixed in 19.55 minutes.

4.3.2. Heating Plate Method

Using the heating plate, the drying process is more controlled. Consequently, after a few experiments, the optimum condition was achieved. The method to dry the sample on the heating plate consisted of placing the GCS support on the heating plate, set at 120 ºC, for approximately 30 seconds, then pipetting 60 µL of sample to the support, and drying it, which took approximately 2.50 minutes.

4.4. Sugars Quantification

Figure 21 shows the analytical curves for quantification of sucrose, glucose, and fructose in sugarcane juice. All the analytical curves present a quadratic fitting, and the models were well adjusted, presenting coefficient of determination of 0.9998, 0.9995, and 0.9997 for sucrose, glucose, and fructose curves, respectively.

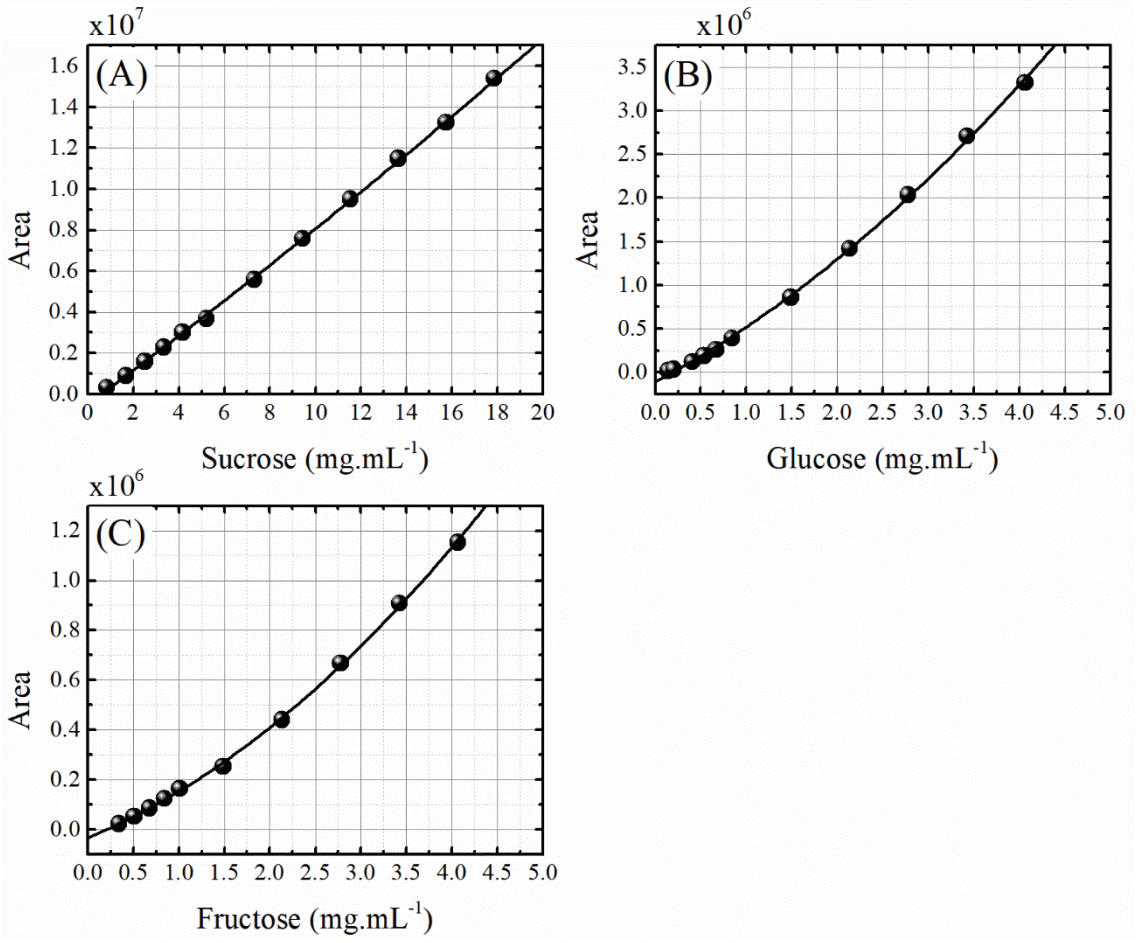

Figure 21. Analytical curve for quantification of sucrose (A), glucose (B), and fructose (C) in the sugarcane juice.

The **Equation 23, 24, and 25** are the model for sucrose, glucose, and fructose, respectively.

$$\mathbf{Area} = \left(\mathbf{0.04} \times [\mathbf{Suc}]^2 + \mathbf{8.17} \times [\mathbf{Suc}] - \mathbf{4.80}\right) \times \mathbf{10}^5 \qquad \textbf{Eq. 23}$$

$$\mathbf{Area} = \left(\mathbf{0.74} \times [\mathbf{Glu}]^2 + \mathbf{5.55} \times [\mathbf{Glu}] - \mathbf{1.08}\right) \times \mathbf{10}^5 \qquad \textbf{Eq. 24}$$

$$\mathbf{Area} = \left(\mathbf{0.35} \times [\mathbf{Fru}]^2 + \mathbf{1.51} \times [\mathbf{Fru}] - \mathbf{0.33}\right) \times \mathbf{10}^5 \qquad \textbf{Eq. 25}$$

The figures of merit, RSD, LOD, and LOQ, used to validate the analytical curves are presented in **Table 9**.

Table 9. Analytical parameters for validation of the chromatographic methods.

Sugar	RSD (%)	LOD (mg mL^{-1})	LOQ (mg mL^{-1})
Sucrose	0.48	0.013	0.026
Glucose	0.94	0.009	0.018
Fructose	2.86	0.030	0.061

RSD – Relative standard deviation; LOD – Limit of detection; LOQ – Limit of quantification.

Figure 22 shows the chromatograms of the sugarcane samples. The sucrose peak can be observed around 8 minutes, glucose peak around 10 minutes, and fructose peak around 14.5 minutes. The samples were quantified using the analytical curves presented previously.

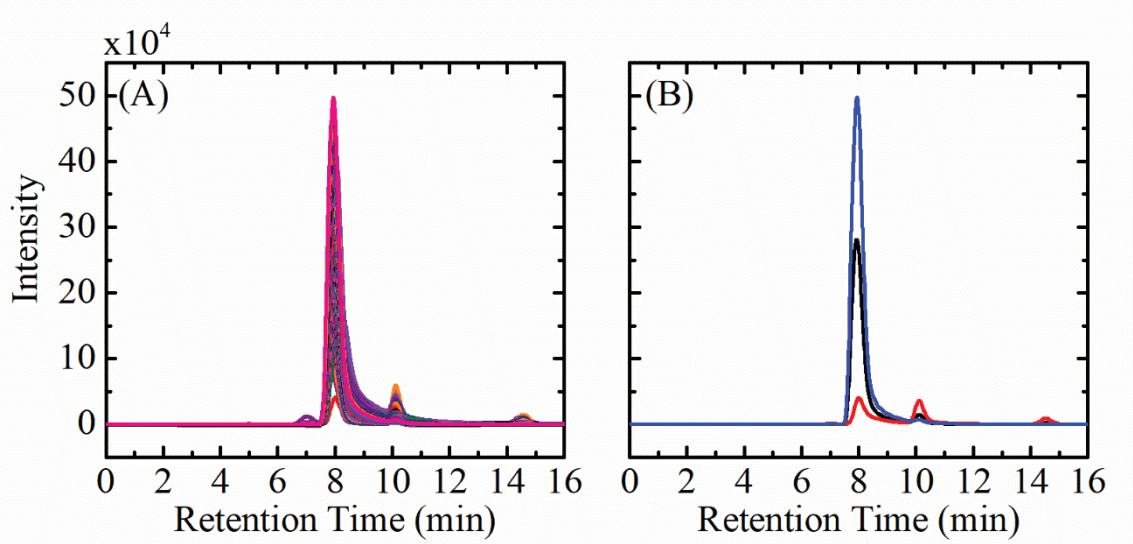

Figure 22. Chromatograms of the sugarcane samples (A) and high (blue line), medium (black line), and low (red line) sucrose content (B).

4.5. Independent Variables

Figure 23 shows the NIR spectra of the samples for the oven and heating plate methods ranging from 1000 to 2500 nm (full) and from 1000 to 1700 nm (short). The

region from 1000 to 1700 nm will be studied to simulate the region provided by most portable NIR instruments.

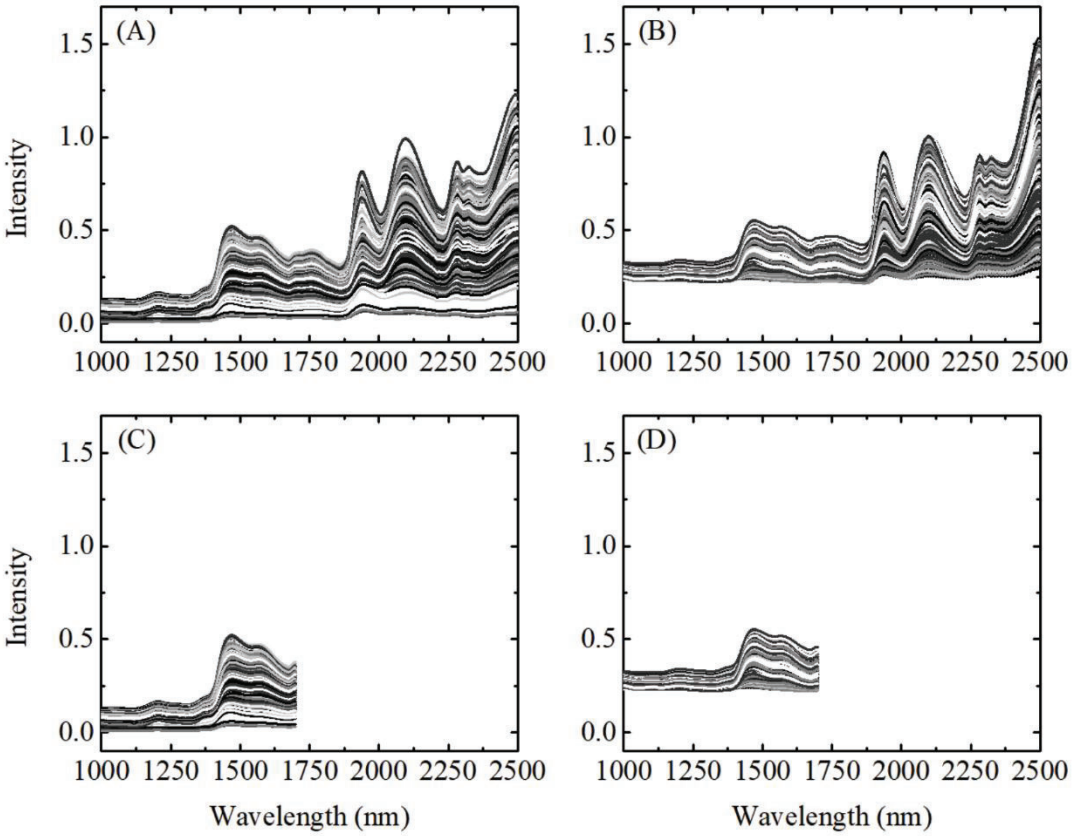

Figure 23. NIR spectra ranging from 1000 to 2500 nm (A) and 1000 to 1700 nm (C) of the sugarcane juice on GRO using the oven method and from 1000 to 2500 nm (B) and 1000 to 1700 nm (D) on GCS using the heating plate method.

Although the sugars present similar absorption peaks, they may be differentiated by their absorption magnitude due to the different numbers of O-H groups, and slight changes in the O-H and C-H absorption band positions caused by inter-molecular hydrogen bonds. The absorption bands due to O-H and C-H groups in sugars in which sucrose contains eight O-H functional groups and glucose and fructose contain five groups each largely influence the spectral variation although they may still be differentiated. In addition, glucose and fructose present C=O bonds, not presented in sucrose. Spectral ranges between 1100 and 1300 nm may be attributed to C-H second overtone, between 1400 and 1600 nm to O-H first overtone, between 1650 and 1800 nm to C-H first overtone, between 1900 and 2000 nm to C=O second overtone, between 2000 and 2200 nm to O-H combinations, between 2250 and 2350 nm to C-H-C-H combinations, and between 2400 to 2500 nm to C-H-O-H combinations.

Figure 24 presents the spectra of the pure solid and dehydrated sucrose, glucose, and fructose, the dehydrated sample, sample solution, and water. It can be observed the spectra of both solid and dehydrated compounds are very similar. The spectrum of the

juice is very similar to the water spectra, which means the water, as the main compound of the juice, strongly influences the juice spectra. The dehydrated juice spectrum differs from the water spectrum, proving the drying process removes the water influence.

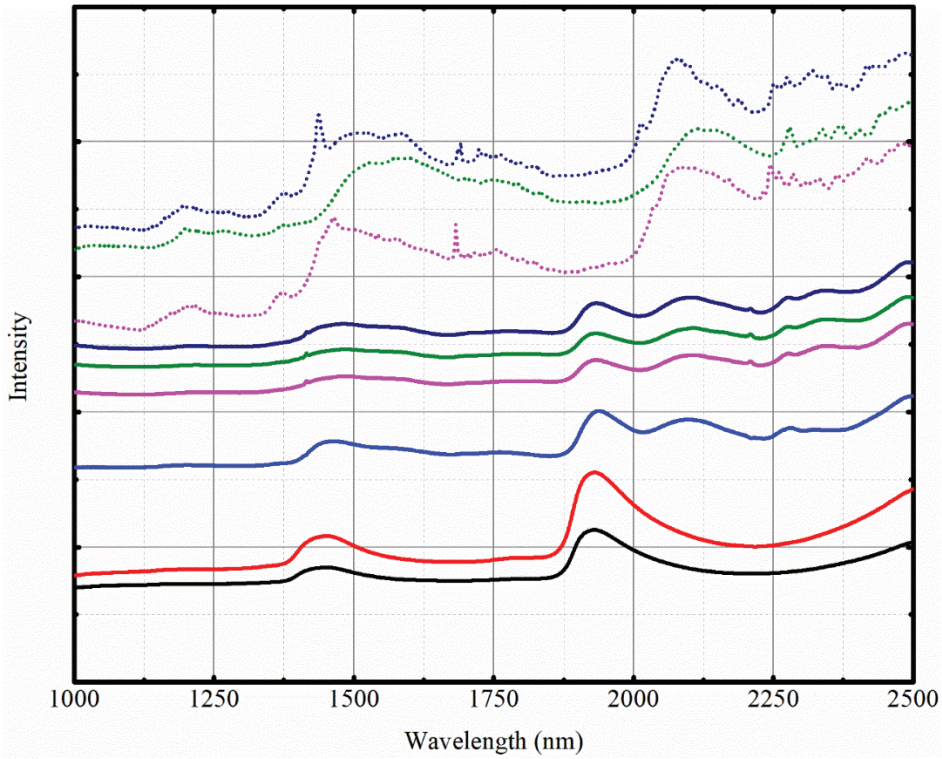

Figure 24. Stalked spectra of pure solid sucrose (navy short dot), pure solid glucose (olive short dot), pure solid fructose (magenta short dot), pure dehydrated sucrose (navy line), pure dehydrated glucose (olive line), pure dehydrated fructose (magenta line), dehydrated juice (blue line), juice (red line), and water (black line).

4.6. Dependent Variables

Table 10 shows the distribution of the dependent variables, sucrose, glucose, fructose, BRIX, fiber, and POL, used in this work.

Table 10. Descriptive statistics of the dependent variables.

Variable	Unit	Min	25%	50%	75%	Max	Mean	*Std*
Sucrose	mg/mL	25.7	105.9	142.3	172.2	215.8	138.1	42.6
Glucose	mg/mL	0.0	6.9	10.6	15.4	25.6	11.3	5.6
Fructose	mg/mL	0.0	3.8	5.3	7.7	21.7	6.2	4.0
BRIX	%	5.4	12.8	16.7	19.7	23.6	16.2	4.2
POL	ºS	8.3	33.1	53.5	71.8	89.8	52.2	22.5
Fiber	%	8.4	12.9	14.6	16.9	22.5	14.9	2.9

Min: Minimum; Max: Maximum; *Std*: Standard deviation.

Figure 25 shows the dispersion of the contents of all samples and the histogram for sucrose, glucose, fructose, BRIX, fiber, and POL. For BRIX and POL, was observed a tendency, *i.e.*, the content increases with the time, since the samples are organized in

chronological order. It may be attributed to instruments badly calibrated or experimental errors.

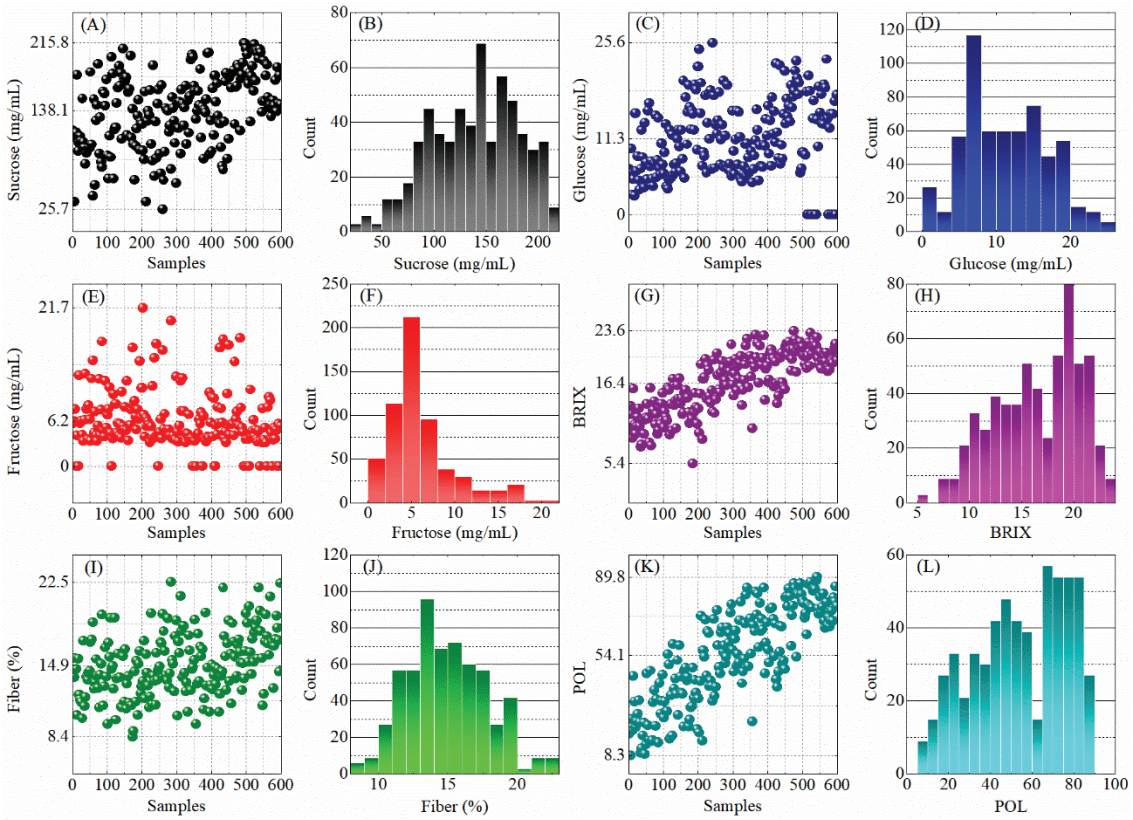

Figure 25. Samples versus content of sucrose (A), glucose (C), fructose (E), BRIX (G), fiber (I), and POL (K). Histogram for sucrose (B), glucose (D), fructose (F), BRIX (H), fiber (J), and POL (L).

Figure 26 present the pearson correlation chart for the dependent variables. It can be observed that only BRIX and POL presented a high correlation (0.95). BRIX and POL should be also highly correlated to sucrose content, since they are used to estimate the sucrose content. In the literature, POL presents correlation higher than 0.90 with sucrose [29]. However, the correlations with sucrose were low, 0.71 for BRIX and 0.75 for POL. This result also suggests that the BRIX and POL determinations for these samples could not be used to effectively choose sugarcane genotypes. Probably the instruments used were not well calibrated or sample preparation was performed incorrectly. Thus, choosing sugarcane genotypes based on the estimation of the sucrose content through BRIX and POL could jeopardize the genotypes assessment. For that reason, BRIX and POL models were not built, since they are not suitable to choose genotypes.

It can be observed that glucose and fructose present correlation of 0.65. Also, BRIX and POL correlates with fructose, -0.39 and -0-48, respectively, with does not happen with glucose. Thus, fructose content may be interfering in the BRIX and POL analyzes.

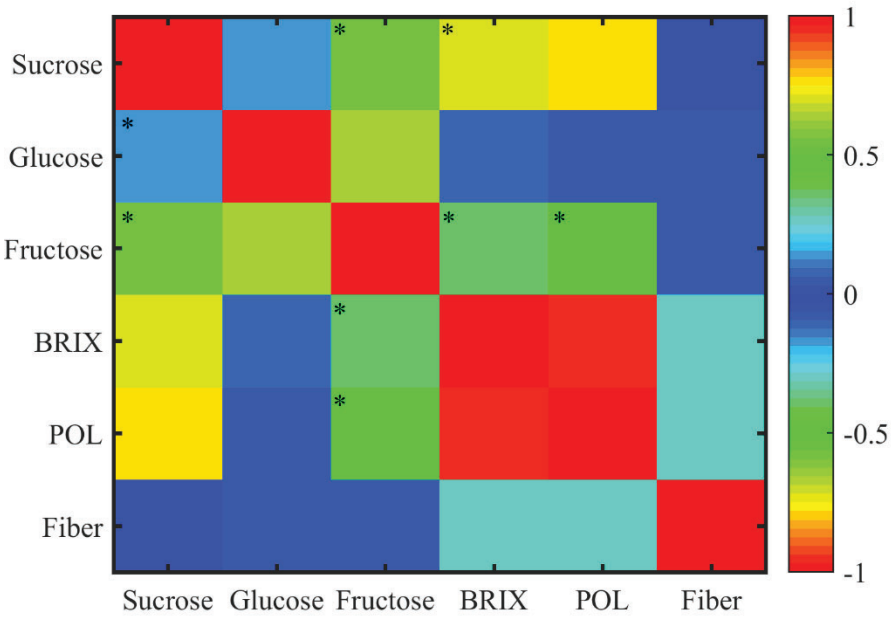

Figure 26. Correlation grid of the dependent variables, sucrose, glucose, fructose, BRIX, POL, and fiber. Cells marked with * identifies negative correlations.

Figure 27 shows the correlation between the sucrose and BRIX and POL. Although the correlations between sucrose and BRIX and sucrose and POL were 0.71 and 0.75, respectively, it can be observed that the samples presented a scattered profile. It can be observed a data scattering, most related to the samples from May 2013.

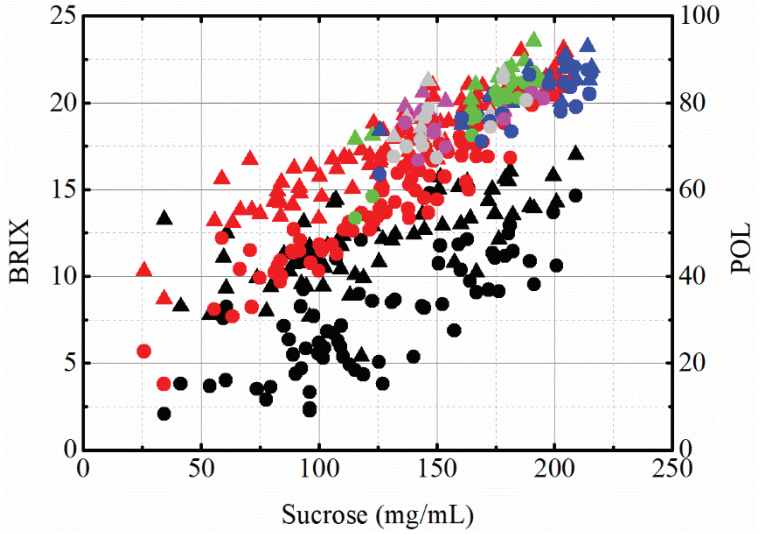

Figure 27. Sucrose versus BRIX (spheres) and POL (triangles) for the samples harvested on May (black) and August 2013 (red) and August (green), September (blue), October (magenta), and November 2015 (gray).

4.7. Multivariate Regression

Table 11 shows the descriptive statistics, minimum, median, maximum, 25%, 50%, and 75% percentiles, mean, and the standard deviation (*Std*) of the properties, sucrose, glucose, fructose, and fiber.

Table 11. Descriptive statistics of the properties modeled before outliers removal.

	Minimum	25%	50%	75%	Maximum	Mean	*Std*
Sucrose (mg/mL)							
Calibration	25.71	104.57	141.50	168.98	215.75	136.52	42.16
Validation	34.19	114.16	150.05	181.73	215.75	145.85	44.02
Glucose (mg/mL)							
Calibration	2.84	7.30	11.27	15.43	25.57	11.94	5.15
Validation	2.84	6.83	10.18	15.19	24.63	11.26	5.48
Fructose (mg/mL)							
Calibration	3.26	4.15	5.65	7.99	21.70	6.91	3.77
Validation	3.26	3.99	5.05	6.43	21.70	5.98	3.21
Fiber (%)							
Calibration	8.38	12.92	14.58	16.90	22.53	14.88	2.91
Validation	10.04	12.69	15.21	16.66	22.02	14.88	2.62

25% - first quartile; 50% - second quartile; 75% - third quartile; *Std* – Standard deviation

It can be seen in **Table 11** that sucrose is at least ten times more concentrated than glucose and fructose, both in the high and low concentrations. This information indicates that NIR detection and the PLS model building for this property will certainly be facilitated.

Table 12 shows the drying method, range, number of samples used to calibrate and validate the model, the preprocessing and transformation applied to the raw spectra, and the number of outliers removed from each of the 16 models. The number of samples used to validate the model was around 20% of the total samples. The full set consisted of 3112 variables ranging from 1000 to 2500 nm, and short set consisted of 2135 variables ranging from 1000 to 1700 nm.

It is possible to observe that the preprocessing and transformations used were very different for each set. The execution order also has been considered and can make a substantial difference in the model quality. As an example, it can be noted that for oven drying , mean centering was applied differently in the full and short sets. The short set used smoothing and the full did not. Another relevant result is the use of autoscaling for NIR data. Using this preprocessing is not indicated in the literature for spectroscopies. However, it was observed that the results using this preprocessing were better than using

mean centering. This work did not explore the details about the preprocessing and transformations used and nor on their application order. However, more in-depth studies will be carried out to better understand the observed results.

Figure 28 shows raw full (1000 to 2500 nm) and short (1000 to 1700 nm) spectra collected using the oven method and heating plate method and the pretreated spectra for each model of sucrose, glucose, fructose, and fiber. It can be observed that the data structure changed for each model, reiterating the idea that the best preprocessment combination must be achieved individually for each set.

Table 12. Information about the sucrose, glucose, fructose, and fiber models.

Drying Method	Range (nm)	Calibration (*N*)	Validation (*N*)	Preprocessing* or Transformations	Outliers (*N*)
			Sucrose		
Oven	1000 to 2500	500	100	Normalize, mean center*, OSC	25
Oven	1000 to 1700	500	100	Mean center*, OSC, smoothing	39
Heating Plate	1000 to 2500	500	100	Autoscale*, baseline, OSC	26
Heating Plate	1000 to 1700	500	100	First derivative, autoscale*, OSC	10
			Glucose		
Oven	1000 to 2500	479	94	Second derivative, autoscale*, OSC	23
Oven	1000 to 1700	480	93	Detrend, autoscale*, OSC	20
Heating Plate	1000 to 2500	475	98	First derivative, autoscale*, OSC	29
Heating Plate	1000 to 1700	475	98	Autoscale*, first derivative, OSC	32
			Fructose		
Oven	1000 to 2500	459	90	Baseline, OSC, mean center*	24
Oven	1000 to 1700	459	90	First derivative, baseline, OSC	31
Heating Plate	1000 to 2500	455	94	Autoscale*, OSC, smoothing	34
Heating Plate	1000 to 1700	459	90	Autoscale*, detrend, OSC	38
			Fiber		
Oven	1000 to 2500	500	100	Autoscale*, First derivative, OSC	16
Oven	1000 to 1700	500	100	Autoscale*, OSC, smoothing	14
Heating Plate	1000 to 2500	500	100	Autoscale*, OSC, normalize	25
Heating Plate	1000 to 1700	500	100	Autoscale*, first derivative, OSC	15

N – Number of samples;

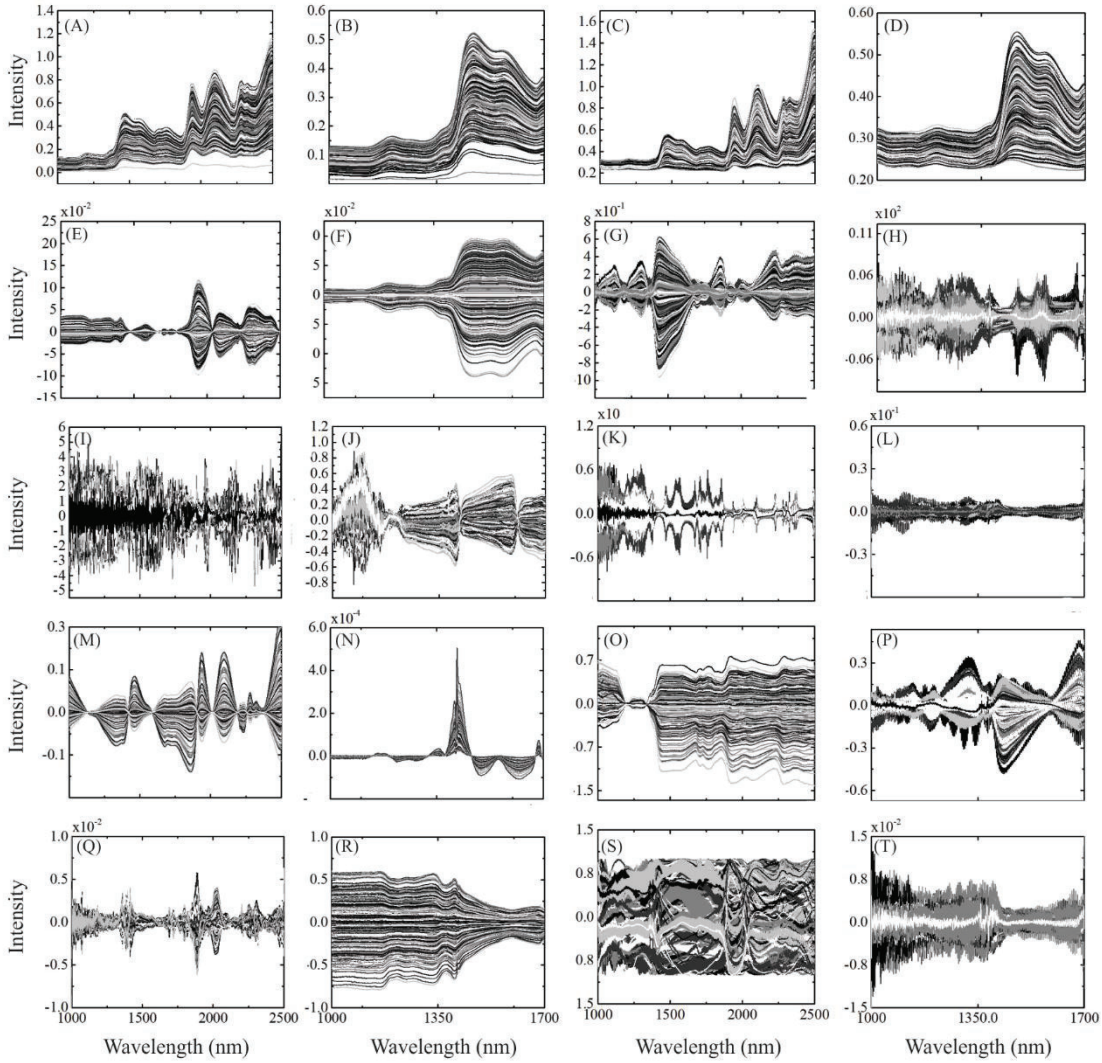

Figure 28. Raw full spectra (1000 to 2500 nm) collected using the oven method (A) and heating plate method (C) and raw short spectra (1000 to 1700 nm) collected using the oven method (B) and heating plate method (D). Pretreated full spectra using the oven method for the sucrose (E), glucose (I), fructose (M), and fiber model (Q). Pretreated short spectra using the oven method for the sucrose (F), glucose (J), fructose (N), and fiber model (R). Pretreated full spectra using the heating plate method for the sucrose (G), glucose (K), fructose (O), and fiber model (S). Pretreated short spectra using the heating plate method for the sucrose (H), glucose (L), fructose (P), and fiber model (T).

Table 13 shows the quality parameters for the sucrose, glucose, fructose, and fiber models. Most of the models were built using one latent variable, making them simpler models. It can be observed that sucrose models presented high SEL and low SEN in comparison to the other properties, reiterating that sucrose modeling is facilitated due to its high concentration. LOD for sucrose and fructose models were close to the minimum content used to calibrate de model. For the other properties, the LOD were higher than the minimum value, thus the samples less concentrated of glucose and fiber could not be predicted well.

Table 13. Quality parameters for the sucrose, glucose, fructose, and fiber model.

	Sucrose				Glucose				Fructose				Fiber			
	Oven		Heating Plate		Oven		Heating Plate		Oven		Heating Plate		Oven		Heating Plate	
	Full	Short	Full	Short	Full	Short	Full	Short	Full	Short	Full	Short	Full	Short	Full	Short
LV	1	8	1	1	1	1	1	1	10	1	10	1	2	7	10	1
Calibration																
RMSEC	14.86	12.47	15.77	15.10	2.31	2.37	2.30	2.73	1.48	2.50	1.34	1.79	1.75	1.64	2.00	1.66
Rc	0.93	0.95	0.92	0.93	0.89	0.88	0.89	0.84	0.92	0.88	0.92	0.86	0.79	0.80	0.66	0.81
SEL	0.99	0.0011	0.94	0.52	0.15	0.75	0.26	0.41	0.0022	0.99	0.006	0.87	0.40	0.004	0.01	0.20
SEN	0.02	1.9×10^{-5}	0.19	0.51	1.71	1.18	1.57	0.004	0.0006	0.02	0.013	0.85	0.007	0.008	0.20	0.003
γ	0.36	6.6×10^{-5}	0.94	0.33	0.48	1.15	5.07	0.97	0.0036	6.96	0.01	1.48	80.2	0.002	0.13	5.56
γ^{-1}	2.71	15008	1.06	2.99	2.07	0.87	0.20	1.02	275.42	0.14	73.68	0.67	0.01	392	7.7	0.18
LOD	23.9	16.7	24.9	22.4	3.8	4.0	3.6	5.8	1.8	109.8	1.55	3.0	10.7	10.0	22.0	9.45
Cross-Validation																
RMSECV	14.97	14.84	15.94	15.26	2.48	2.38	2.39	2.75	1.72	2.50	1.74	1.80	1.83	1.87	2.29	1.73
Rcv	0.93	0.93	0.92	0.92	0.87	0.88	0.89	0.84	0.89	0.87	0.88	0.86	0.77	0.74	0.54	0.79
Validation																
RMSEP	17.33	18.52	19.91	26.86	3.98	4.37	3.77	3.89	1.85	3.04	1.97	2.60	1.95	2.20	2.58	3.56
Rp	0.92	0.92	0.87	0.82	0.69	0.68	0.77	0.71	0.84	0.85	0.85	0.74	0.67	0.69	0.43	0.46

LV – Latent variable; *RMSEC* – Root mean square error of calibration; *Rc* – Correlation coefficient of calibration; *SEL* – Selectivity; *SEN* – Sensitivity; γ – Analytical sensibility; γ^{-1} – Inverse of analytical sensibility; *LOD* – Limit of detection; *RMSECV* – Root mean square error of cross-validation; *Rcv* – Correlation coefficient of cross-validation; *RMSEP* – Root mean square error of validation.

Overall, sucrose content was well modelled using all the methods, however the models built using the oven method presented higher quality than those built using the heating plate method.

From the information presented, the best sucrose model was the one using the oven method and full spectra, presenting *RMSECV* and *RMSEP* values of 14.97 and 17.33, respectively, and high correlation coefficients, 0.93 for cross-validation and 0.92 for validation. The model presented high SEL value, 0.99, indicating the full spectra is more suitable for sucrose modeling, and low SEN, 0.02. In addition, the γ^{-1} for the model was 2.71. The best glucose model was the one using the heating plate method and full spectra, presenting *RMSECV* and *Rcv* of 2.39 and 0.89, and *RMSEP* and *Rp* of 3.77 and 0.77, respectively. This model presented the smallest γ^{-1} value. The best fructose model was the one using the oven method and full spectra, *RMSECV, Rcv, RMSEP,* and *Rp* of 1.72, 0.89, 1.85, and 0.84. The best fiber model was the one using the oven method and full spectra, presenting *RMSECV* and *Rcv* of 1.83 and 0.77 and *RMSEP* and *Rp* of 1.95 and 0.67. It can be assumed that even though fiber being the insoluble part of the sugarcane, there is something soluble in the sugarcane juice that correlates to the fiber content, hence the fiber content is indirectly modeled.

The model built for sucrose could be used for sucrose prediction with certain accuracy. The models built for glucose, fructose, and fiber are not suitable for prediction, however they could be used to discriminate low and high content of the respective properties.

Figure 29 shows the values observed (black spheres) and predicted (red spheres) by the model built for the properties, sucrose, glucose, fructose, and fiber using the oven and heating plate method and full and short spectra. It can be observed that there is a good linear fitting for the sucrose models and a linear tendency can be observed for most of the models, except for the fructose models using the oven method and short spectra (**Figure 29**-J).

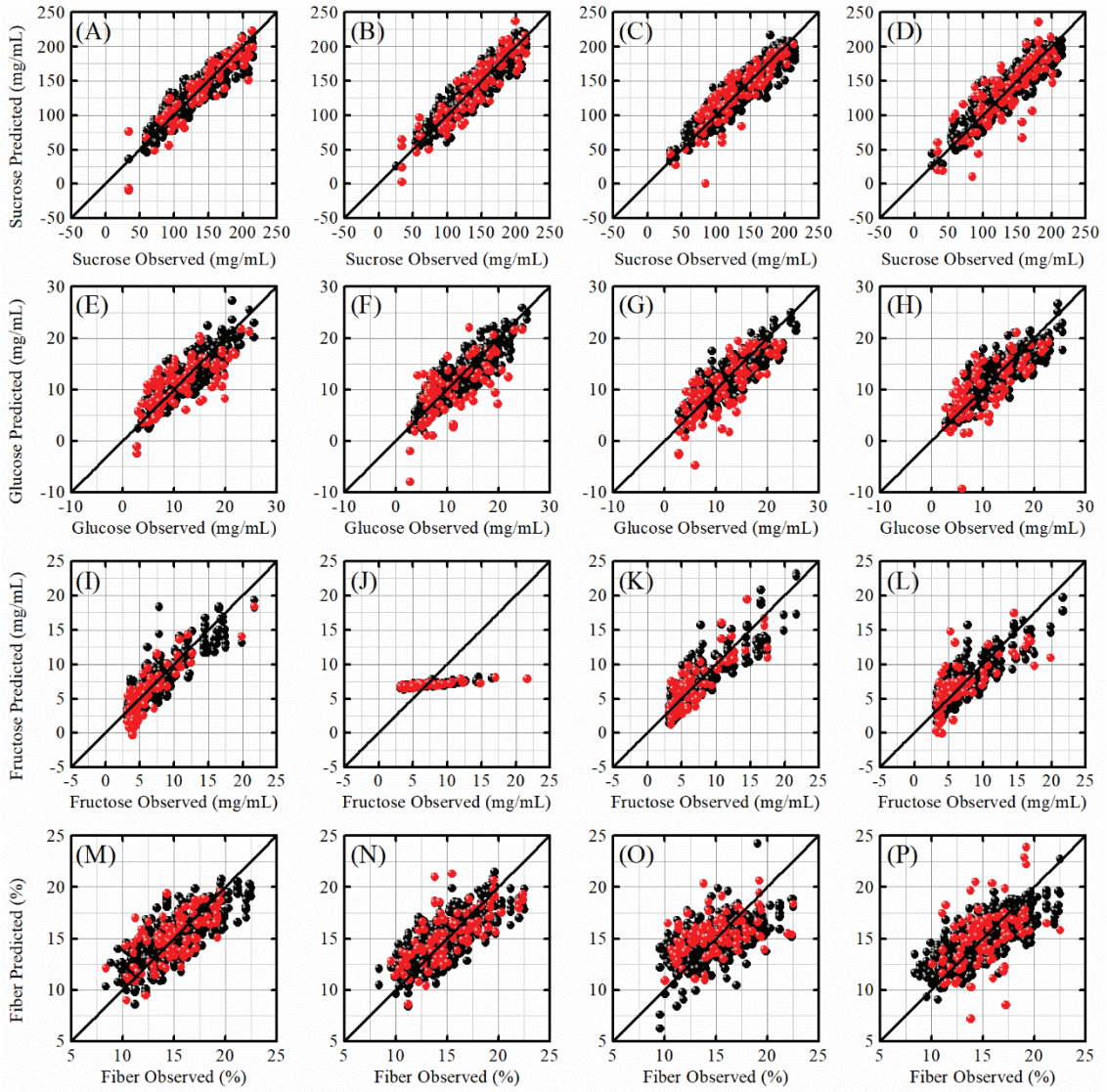

Figure 29. Observed versus Predicted for the sucrose models using the oven method and full spectra (A) and short spectra (B) and the heating plate method and full spectra (C) and short spectra (D); for the glucose models using the oven method and full spectra (E) and short spectra (F) and the heating plate method and full spectra (G) and short spectra (H); for the fructose models using the oven method and full spectra (I) and short spectra (J) and the heating plate method and full spectra (K) and short spectra (L); and for the fiber models using the oven method and full spectra (M) and short spectra (N) and the heating plate method and full spectra (O) and short spectra (P). The black spheres represent the cross-validation samples and the red spheres the validation samples.

Figure 30 shows the relative errors of the validation sets and the distribution of the errors for the properties modeled. For sucrose, most of the samples were predicted with less than 10% of error, and the model using the oven method and full spectra presented the smallest errors in comparison to the other sucrose models. For glucose and fructose, most of the samples presented relative errors smallest than 50%, and the best model was the one using the heating plate method and full spectra. For fiber, most samples presented relative errors smallest than 40%, and the best fiber model was the one using the oven method and short spectra.

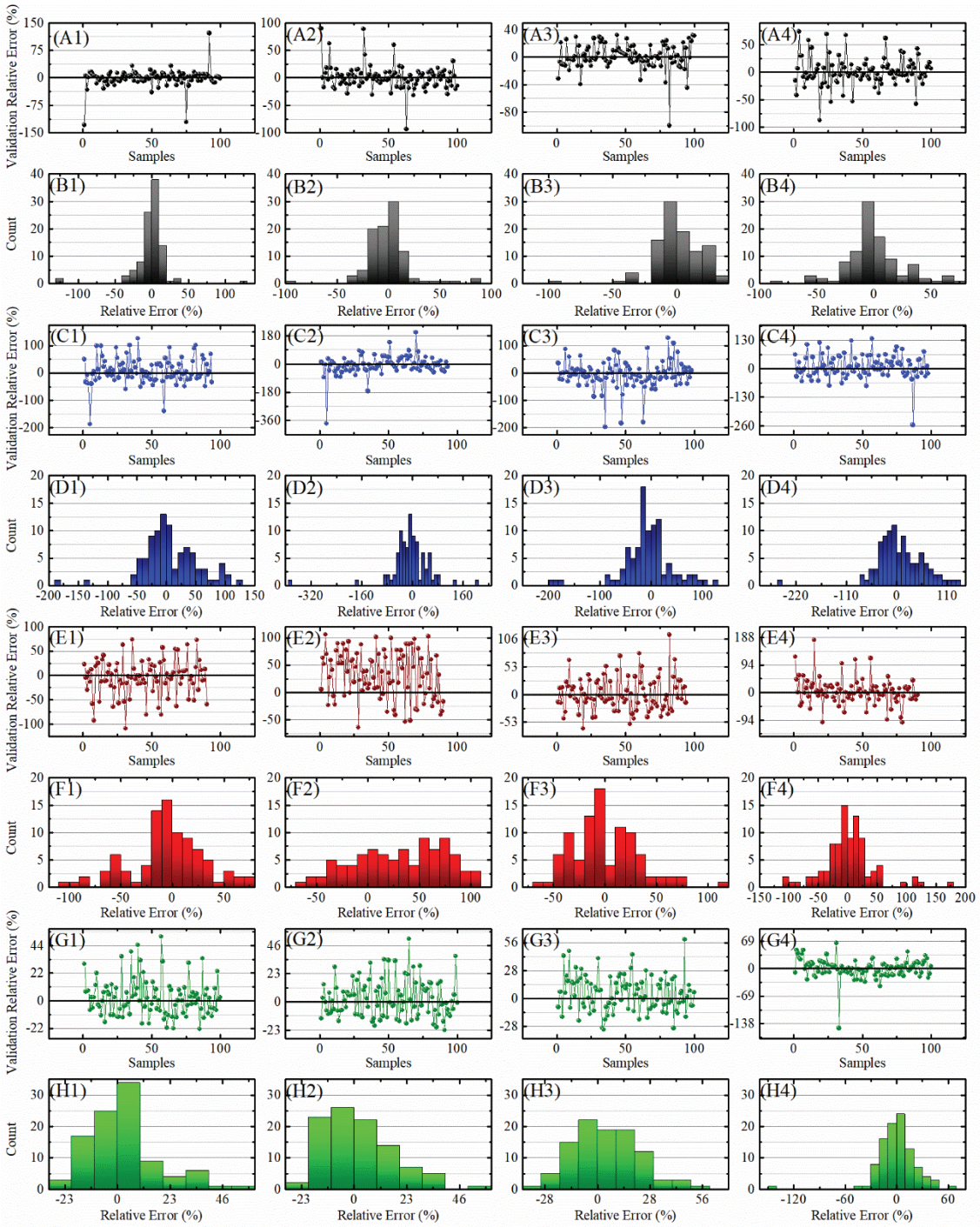

Figure 30. Relative errors for the validation sets (A, C, E, and G) and histogram of errors (B, D, F, and H) using the oven method full set (column 1) and short set (column 2) and using the heating plate method full set (column 3) and short set (column 4) for sucrose (black color A1, A2, A3, A4, B1, B2, B3, and B4, respectively), glucose (blue color C1, C2, C3, C4, D1, D2, D3, and D4, respectively), fructose (red color E1, E2, E3, E4, F1, F2, F3, and F4, respectively), and fiber (green color G1, G2, G3, G4, H1, H2, H3, and H4, respectively).

Figure 31 shows the pearson correlation charts for the models built. It can be observed that for all models built using the authentic **y** vector (red sphere) the Rcv^2 and Rc^2 differ from those using the randomized **y** vectors (black spheres), except the model for fructose using the oven method and short spectra (**Figure 31**-J).

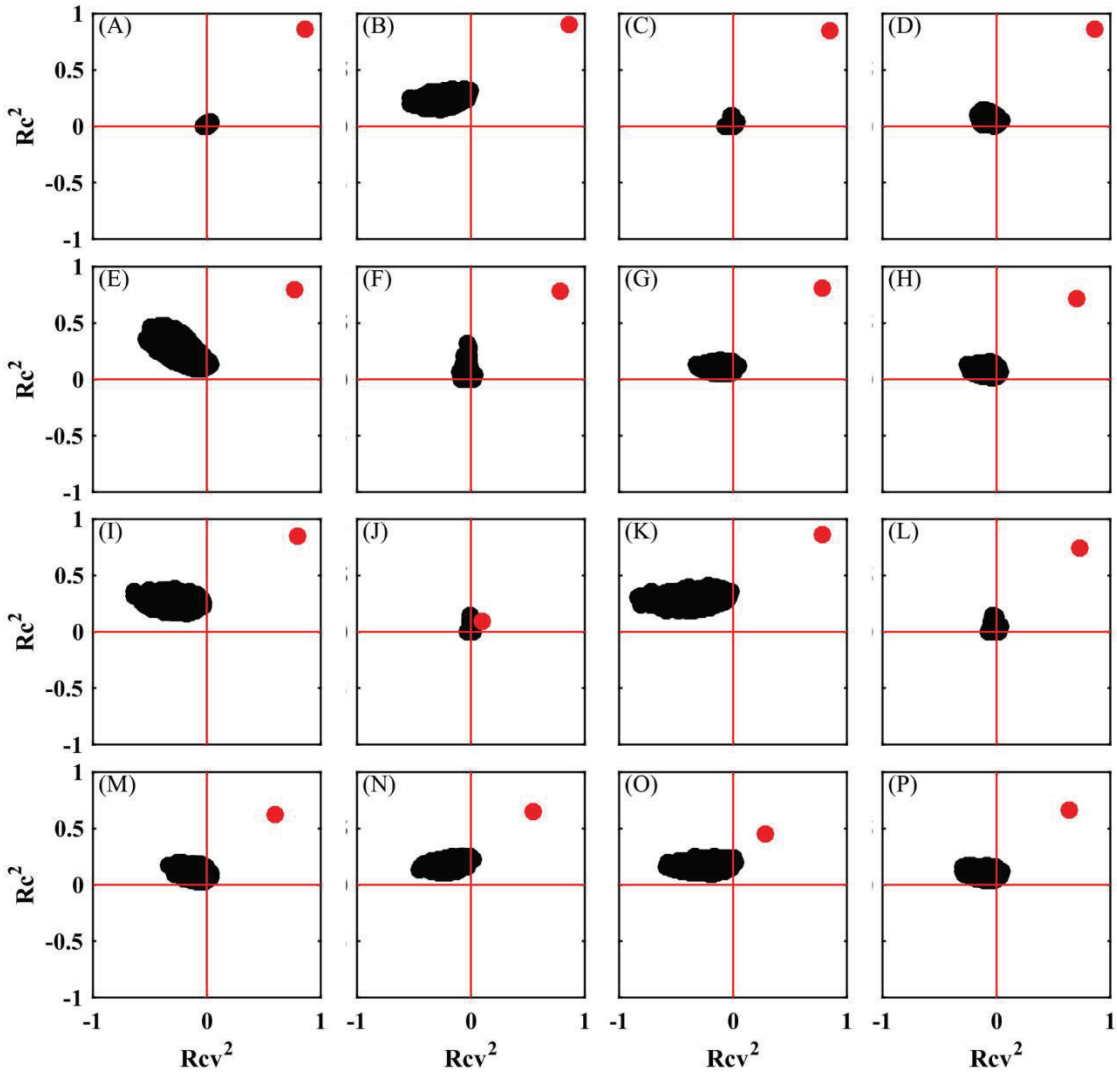

Figure 31. Determination coefficient of cross-validation (Rcv^2) versus determination coefficient of calibration (Rc^2) (pearson correlation) for sucrose (A), glucose (E), fructose (I), and fiber (M) using full spectra and oven method; for sucrose (B), glucose (F), fructose (J), and fiber (N) using short spectra and oven method; for sucrose (C), glucose (G), fructose (K), and fiber (O) using full spectra and heating plate method; and for sucrose (D), glucose (H), fructose (L), and fiber (P) using short spectra and heating plate method. Black spheres represent the randomized models and the red sphere the original model.

4.8. Analytical Frequency

For the oven, the analytical frequency achieved in this work was of approximately 2 samples per min. For the heating plate, the analytical frequency achieved in this work was of 1.2 samples per min for oven. For NIR analysis, the analytical frequency achieved in this work was of 1 sample per minute.

For the heating plate, the frequency was smaller and, besides that, it is necessary be presented near the heating plate to add the samples to the support, which is a disadvantage, as for the oven method, the samples are placed on the iron tray and introduced into the oven.

4.9. Ranking

Table 14 show the parameters calculated to assess the efficacy of the sample selection using each property, BRIX, POL, and sucrose predicted by the PLS model using the oven method and full spectra. It can be observed that the selection using the sucrose predicted by the NIR and PLS model was better than using BRIX or POL, with error, sensitivity, specificity, and accuracy of 9, 84, 94, and 91%, respectively.

Table 14. Parameters to assess the efficacy of the sample selection.

Property	*TP*	*FP*	*FN*	*TN*	*E* (%)	*S* (%)	*Sp* (%)	*A* (%)
Sucrose (HPLC)	30	0	0	70	0	100	100	100
BRIX	16	14	15	55	29	52	80	71
POL	15	15	15	55	30	50	79	70
Sucrose (NIR-PLS)	26	4	5	65	9	84	94	91

TP – True positive; *FP* – False positive; *FN* – False negative; *TN* – True negative; *E* – Error; *S* – Sensibility; *Sp* – Specificity; *A* – Accuracy.

Table 15 shows the sample ranking using the reference sucrose content, BRIX, POL, and sucrose predicted by PLS model using the full NIR spectra collected from juice dehydrated in oven.

Table 15. Ranking of genotypes using the reference sucrose (HPLC), BRIX, POL, and sucrose modeled by PLS. Bold red samples represent the TOP 30 samples for each property.

Sample	Reference (HPLC)		Technological			Technological			PLS model		
	Sucrose (mg/mL)	Rank	BRIX	Rank	Class	POL	Rank	Class	Sucrose (mg/mL)	Rank	Class
1	115.32	75	10.10	94	TN	18.37	97	TN	90.78	91	TN
6	34.34	98	12.52	85	TN	31.29	86	TN	7.02	99	TN
11	113.00	76	13.90	78	TN	39.37	80	TN	100.29	87	TN
17	92.33	87	9.44	98	TN	21.20	94	TN	120.16	72	TN
18	92.33	87	10.82	91	TN	24.34	90	TN	99.24	88	TN
21	101.69	81	12.94	83	TN	33.57	84	TN	101.23	85	TN
27	103.54	79	16.20	60	TN	56.26	54	TN	116.31	78	TN
29	152.36	46	12.76	84	TN	32.23	85	TN	155.73	45	TN
31	102.40	80	10.92	90	TN	25.40	89	TN	109.85	79	TN
38	77.60	95	9.24	99	TN	19.14	96	TN	96.12	89	TN
45	94.43	85	12.28	86	TN	34.80	83	TN	126.34	69	TN
50	96.01	84	11.50	88	TN	22.15	93	TN	75.75	94	TN
57	99.70	82	10.20	93	TN	16.50	99	TN	89.75	92	TN
69	127.03	69	15.22	67	TN	44.92	72	TN	118.66	74	TN
79	130.86	66	11.72	87	TN	31.26	87	TN	143.81	60	TN
97	173.49	33	11.24	89	TN	25.66	88	TN	142.73	63	TN
102	87.23	91	17.00	55	TN	57.40	53	TN	100.69	86	TN
108	109.48	77	15.48	64	TN	49.88	61	TN	95.53	90	TN
109	85.10	92	15.02	68	TN	45.72	70	TN	101.87	84	TN
110	85.10	92	14.28	73	TN	40.74	79	TN	104.93	82	TN

Table 15. (Continuation).

Sample	Reference (HPLC)		Technological			Technological			PLS model		
	Sucrose (mg/mL)	Rank	BRIX	Rank	Class	POL	Rank	Class	Sucrose (mg/mL)	Rank	Class
113	199.41	**13**	13.92	77	FN	36.82	82	FN	185.65	**24**	TP
115	180.36	**28**	9.90	96	FN	17.38	98	FN	176.81	34	FN
133	189.54	**19**	13.62	80	FN	43.29	75	FN	186.10	**23**	TP
134	189.54	**19**	13.60	81	FN	44.32	74	FN	186.52	**21**	TP
137	166.78	39	17.52	51	TN	59.29	51	TN	165.26	43	TN
139	181.12	**27**	15.44	65	FN	47.73	64	FN	177.26	33	FN
143	106.12	78	9.88	97	TN	19.86	95	TN	120.71	71	TN
145	208.99	**4**	14.20	75	FN	41.38	78	FN	152.19	48	FN
165	162.83	40	14.38	72	TN	47.24	65	TN	176.29	35	TN
179	59.39	97	10.00	95	TN	23.17	91	TN	70.63	95	TN
189	132.34	65	14.90	69	TN	42.41	77	TN	139.39	65	TN
190	150.80	47	17.67	47	TN	59.67	50	TN	165.44	42	TN
192	150.80	47	8.70	100	TN	15.15	100	TN	172.78	38	TN
196	125.44	73	17.21	53	TN	53.48	56	TN	130.18	67	TN
201	122.66	74	14.22	74	TN	44.89	73	TN	146.63	56	TN
203	92.53	86	18.78	39	TN	65.20	41	TN	69.55	96	TN
211	34.19	99	13.69	79	TN	42.46	76	TN	62.26	98	TN
212	34.19	99	16.71	57	TN	59.95	49	TN	-12.06	100	TN
215	126.20	70	17.63	48	TN	62.91	44	TN	147.72	54	TN

Table 15. (Continuation).

Sample	Reference (HPLC)		Technological			Technological			PLS model		
	Sucrose (mg/mL)	Rank	BRIX	Rank	Class	POL	Rank	Class	Sucrose (mg/mL)	Rank	Class
220	133.26	64	17.62	49	TN	61.10	47	TN	148.78	53	TN
230	81.07	94	10.33	92	TN	22.71	92	TN	107.23	80	TN
251	88.98	90	16.10	63	TN	53.42	58	TN	105.12	81	TN
265	91.56	89	19.62	33	TN	70.13	37	TN	103.79	83	TN
271	138.58	57	16.46	58	TN	53.36	59	TN	146.07	57	TN
273	138.58	57	17.02	54	TN	63.18	43	TN	141.47	64	TN
274	137.83	59	14.67	71	TN	46.56	67	TN	118.01	76	TN
275	137.83	59	17.80	45	TN	60.88	48	TN	123.90	70	TN
276	137.83	59	17.28	52	TN	54.58	55	TN	118.01	75	TN
283	71.50	96	15.36	66	TN	49.19	62	TN	62.65	97	TN
301	146.90	54	16.73	56	TN	57.68	52	TN	144.35	59	TN
315	162.50	41	17.76	46	TN	53.45	57	TN	126.37	68	TN
326	166.86	37	18.01	43	TN	62.53	45	TN	150.71	50	TN
327	166.86	37	16.18	61	TN	45.21	71	TN	158.93	44	TN
331	150.05	49	13.45	82	TN	38.73	81	TN	145.33	58	TN
332	150.05	49	21.48	**6**	FP	78.09	**19**	FP	151.93	49	TN
333	150.05	49	16.18	61	TN	46.65	66	TN	152.29	47	TN
338	154.87	43	18.85	38	TN	70.59	36	TN	169.82	39	TN
339	154.87	43	19.85	31	TN	65.84	39	TN	177.86	32	TN

Table 15. (Continuation).

Sample	Reference (HPLC)		Technological			Technological			PLS model		
	Sucrose (mg/mL)	Rank	BRIX	Rank	Class	POL	Rank	Class	Sucrose (mg/mL)	Rank	Class
344	202.07	**10**	14.09	76	FN	45.88	69	FN	174.80	36	FN
346	199.48	**11**	18.72	40	FN	69.16	38	FN	212.65	**2**	TP
347	199.48	**11**	14.78	70	FN	48.39	63	FN	210.70	**3**	TP
352	160.42	42	21.01	**13**	FP	75.50	**30**	FP	178.80	31	TN
368	185.61	**23**	17.53	50	FN	62.42	46	FN	173.93	37	FN
389	203.78	**9**	16.39	59	FN	46.45	68	FN	184.70	**26**	TP
391	204.48	**7**	21.60	**3**	TP	84.53	**2**	TP	187.21	**18**	TP
392	204.48	**7**	21.84	**2**	TP	81.01	**13**	TP	186.91	**20**	TP
398	127.10	67	21.07	**12**	FP	81.76	**9**	FP	117.84	77	TN
399	127.10	67	21.29	**8**	FP	82.51	**8**	FP	119.55	73	TN
428	96.52	83	17.89	44	TN	53.35	60	TN	88.35	93	TN
452	176.07	32	19.03	36	TN	71.15	35	TN	192.03	**13**	FP
460	181.73	**25**	18.41	42	FN	63.55	42	FN	181.66	**27**	TP
462	181.73	**25**	21.54	**5**	TP	83.37	**4**	TP	189.58	**15**	TP
463	184.41	**24**	21.56	**4**	TP	82.61	**7**	TP	191.69	**14**	TP
477	191.04	**16**	18.99	37	FN	65.25	40	TN	204.36	**4**	TP
479	187.10	**22**	21.32	**7**	TP	79.79	**16**	TP	192.87	**12**	TP
490	192.84	**15**	20.04	**28**	TP	76.63	**26**	TP	187.07	**19**	TP
495	215.75	**1**	20.88	**16**	TP	83.26	**5**	TP	200.95	**7**	TP

Table 15. (Continuation).

Sample	Reference (HPLC)		Technological			Technological			PLS model		
	Sucrose (mg/mL)	Rank	BRIX	Rank	Class	POL	Rank	Class	Sucrose (mg/mL)	Rank	Class
500	209.06	3	21.16	10	TP	82.91	6	TP	186.17	22	TP
506	168.98	36	19.62	32	TN	73.85	33	TN	179.22	30	FP
511	125.83	71	20.15	27	FP	78.08	20	FP	149.58	51	TN
512	125.83	71	20.49	24	FP	77.59	21	FP	143.58	61	TN
514	177.59	30	20.74	20	TP	77.34	23	TP	189.45	16	TP
516	177.59	30	20.84	17	TP	81.30	12	TP	185.15	25	TP
519	189.45	21	21.22	9	TP	83.38	3	TP	199.68	8	TP
524	213.89	2	21.08	11	TP	81.33	11	TP	222.84	1	TP
531	206.66	6	20.60	22	TP	78.28	18	TP	187.56	17	TP
539	208.36	5	18.56	41	FN	71.45	34	FN	202.21	5	TP
544	136.52	62	20.25	26	FP	76.87	25	FP	147.04	55	TN
546	136.52	62	20.53	23	FP	77.44	22	FP	149.33	52	TN
552	195.43	14	19.90	30	TP	75.42	31	FN	201.36	6	TP
554	148.81	53	19.99	29	FP	76.51	28	FP	168.88	40	TN
557	178.51	29	20.69	21	TP	79.51	17	TP	197.56	10	TP
560	190.37	17	19.38	35	FN	74.63	32	FN	194.50	11	TP
561	190.37	17	20.35	25	TP	75.79	29	TP	199.35	9	TP
564	153.74	45	21.00	15	FP	80.31	15	FP	168.80	41	TN
570	144.01	55	21.01	13	FP	81.35	10	FP	130.37	66	TN

Table 15. (Continuation).

Sample	Reference (HPLC)		Technological			Technological			PLS model		
	Sucrose (mg/mL)	Rank	BRIX	Rank	Class	POL	Rank	Class	Sucrose (mg/mL)	Rank	Class
578	149.94	52	20.82	**18**	FP	77.02	**24**	FP	153.30	46	TN
580	143.89	56	19.48	34	TN	76.60	**27**	FP	143.25	62	TN
583	172.78	34	20.75	**19**	FP	80.96	**14**	FP	181.58	**28**	FP
584	172.78	34	22.51	**1**	FP	87.67	**1**	FP	180.43	**29**	FP

5. Conclusion

The two new methods, oven and heating plate, to collect NIR spectra from dehydrated sugarcane juice were successfully presented, studied and compared. NIR spectroscopy and PLS regression were capable to predict sucrose, glucose, fructose, and fiber content directly from the sugarcane juice dehydrated in oven or heating plate. The best material supports for juice were thick couche paper for oven and glass cover slips for heating plate. The analytical frequency of the procedure was higher for the oven than heating plate. The predictions for all properties were better when the juice was dehydrated using the oven and full spectra, however, the predictions for fructose were better when the heating plate and full spectra were used. The sucrose model using the short spectra could also be used to predict the sucrose content, enabling the use of portable instruments in the field. The model built for sucrose using the oven method and full spectra presented good capability to select sugarcane genotypes, being superior to the traditional methods used nowadays, BRIX and POL.

6. References

1. Goldemberg, J. The promise of clean energy. *Energy Policy* **34**, 2185–2190 (2006).

2. Baños, R. *et al.* Optimization methods applied to renewable and sustainable energy: A review. *Renew. Sustain. Energy Rev.* **15**, 1753–1766 (2011).

3. Scharlemann, J. P. W. & Laurance, W. F. How Green Are Biofuels? *Science (80-.).* **319**, 43–44 (2008).

4. Cheavegatti-Gianotto, A. *et al.* Sugarcane (Saccharum X officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil. *Trop. Plant Biol.* **4**, 62–89 (2011).

5. CONAB. Acompanhamento da safra brasileira de cana-de-açúcar. (2018).

6. Matsuoka, S., Garcia, A. A. F. & Calheiros, G. C. Hibridação em cana-de-açúcar. *Hibridacao Artif. Plantas. Vicosa Ed. UFV* 221–254 (1999).

7. Jackson, P. A. Breeding for improved sugar content in sugarcane. *F. Crop. Res.* **92**, 277–290 (2005).

8. Simeone, M. L. F. *et al.* Near infrared spectroscopy determination of sucrose, glucose and fructose in sweet sorghum juice. *Microchem. J.* **134**, 125–130 (2017).

9. Pasquini, C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *J. Braz. Chem. Soc.* **14**, 198–219 (2003).

10. Lavanholi, M. Qualidade da cana-de-açúcar como matéria-prima para produção de açúcar e álcool. *Sugarcane'.(Eds LL Dinardo-Miranda, ACM Vascon. MGA Landell)(Instituto Agronômico Campinas)* (2008). doi:doi.org/10.3738/nucleus.v5i2.102

11. Clayton, W. D. & Daniels, C. A. Geographical, historical and cultural aspecto f origin of the Indian and Chinese sugarcane S. barberi and S. sinensis. *ISSCT Sug. Breed. Newslater* **36**, 4–23 (1975).

12. Figueiredo, P. Breve história da cana-de-açúcar e do papel do Instituto Agronômico no seu estabelecimento no Brasil. *Cana-de-açúcar. Campinas Inst. Agronômico* 31–44 (2008).

13. Dantas, B. *Contribuição para a história da" Gomose" da cana de açucar: em Pernambuco e no Brasil*. (Instituto Agronômico do Nordeste, 1960).

14. Bremer, G. A cytological investigation of some species and species hybrids within the genus saccharum. *Genetica* **5**, 97–148 (1923).

15. Deerr, N. *The history of sugar*. **1**, (Chapman and Hall, 1949).

16. Renfrew, J. M. *Palaeoethnobotany. The prehistoric food plants of the Near East and Europe.* (1973).

17. Ng, N. Q., Smartt, J. & Simmonds, N. W. Evolution of crop plants. (1995).

18. Mendel, G. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines Brunn 4 3* **44**, (1866).

19. Forster, B. P. *et al.* Genetic engineering of crop plants: from genome to gene. *Exp. Agric.* **33**, 15–33 (1997).

20. Schwartz, H. F. & Singh, S. P. Breeding Common Bean for Resistance to White Mold: A Review. *Crop Sci.* **53**, 1832 (2013).

21. Poehlman, J. M., Sleper, D. A., Rudd, J. & Edition, T. *Breeding field crops*. **378**, (Springer, 1995).

22. Stevenson, G. C. Genetics and breeding of sugar cane. *Genet. Breed. sugar cane.* (1965).

23. Edgerton, C. W. Sugarcane and its diseases. *Sugarcane its Dis.* (1955).

24. Ming, R. *et al.* Sugarcane improvement through breeding and biotechnology. *Plant Breed. Rev.* **27**, 15 (2006).

25. Daniels, J. & Roach, B. T. *Taxonomy and evolution*. *Developments in Crop Science* **11**, (Elsevier B.V., 1987).

26. Landell, M. G. de A. *Variedades de cana-de-açúcar para o Centro-Sul do Brasil*. (Instituto Agronômico, 2005).

27. D'Hont, A. *et al.* Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *MGG Mol. Gen. Genet.* **250**, 405–413 (1996).

28. Landell, M. G. de A. & Bressiani, J. A. Melhoramento genético, caracterização e manejo varietal. *Cana-de-açúcar. Campinas Inst. Agronômico* 101–155 (2008).

29. Fernandes, A. C. Cálculos na agroindústria da cana-de-açúcar. *STAB Piracicaba, SP, Bras.* (2003).

30. Silva, L. A. *et al.* Selection strategy for indication of crosses between potential sugarcane genotypes aiming at the production of bioenergy. *Ind. Crops Prod.* **104**, 62–67 (2017).

31. Guimarães, J. L., Frollini, E., Silva, C. G., Wypych, F. & Satyanarayana, K. G. Characterization of banana , sugarcane bagasse and sponge gourd fibers of Brazil. **30**, 407–415 (2009).

32. Tanimoto, T. The press method of cane analysis. *Hawaiian Plant. Rec.* **57**, 133–150 (1964).

33. CONSECANA. *Manual de Instruções*. (2001).

34. Skoog, D. A., West, D. M. & Holler, F. J. *Fundamentos de química analítica*. **2**, (Reverté, 1997).

35. Harris, D. C. *Explorando a Química Analítica*. (LTC, 2011). doi:9788521618034

36. Lanças, F. M. *Cromatografia líquida moderna: HPLC/CLAE*. *Campinas, SP* (2009).

37. Vlassa, M., Coman, V. & Halmagyi, A. Simultaneous determination of glucose , fructose , sucrose and sorbitol in the leaf and fruit peel of different apple cultivars by the HPLC – RI optimized method Miut. **199**, 653–659 (2016).

38. Givry, S., Bliard, C. & Duchiron, F. Selective ketopentose analysis in concentrate carbohydrate syrups by HPLC. **342**, 859–864 (2007).

39. Clement, A., Yong, D. & Brechet, C. Simultaneous Identification of Sugars by HPLC Using Evaporative Light Scattering Detection (ELSD) and Refractive Index Detection (RI). Application to Plant Tissues. *J. Liq. Chromatogr.* **15**, 805–817 (1992).

40. Liu, X. *et al.* Quantification of glucose, xylose, arabinose, furfural, and HMF in corncob hydrolysate by HPLC-PDA–ELSD. *Carbohydr. Res.* **353**, 111–114 (2012).

41. Ma, C., Sun, Z., Chen, C., Zhang, L. & Zhu, S. Simultaneous separation and determination of fructose, sorbitol, glucose and sucrose in fruits by HPLC–ELSD. *Food Chem.* **145**, 784–788 (2014).

42. Xu, W., Liang, L. & Zhu, M. Determination of sugars in molasses by HPLC following solid-phase extraction. *Int. J. Food Prop.* **18**, 547–557 (2015).

43. Yu Chen, J., Zhang, H., Miao, Y. & Asakura, M. Nondestructive determination of sugar content in potato tubers using visible and near infrared spectroscopy. *Japan J. Food Eng.* **11**, 59–64 (2010).

44. Xie, L., Ye, X., Liu, D. & Ying, Y. Quantification of glucose, fructose and sucrose in bayberry juice by NIR and PLS. *Food Chem.* **114**, 1135–1140 (2009).

45. Neto, B. B., Scarminio, I. S. & Bruns, R. E. Como fazer experimentos. *Unicamp, Campinas* (2001).

46. Brereton, R. G. *Applied chemometrics for scientists*. (John Wiley & Sons, 2007).

47. Mullen, K. & Ennis, D. M. Rotatable designs in product development. *Food Technol.* (1979).

48. Brereton, R. G. *Chemometrics: data analysis for the laboratory and chemical plant*. (John Wiley & Sons, 2003).

49. Ebrahimi-Najafabadi, H., Leardi, R., Jalali-Heravi, M., Guest, S. & Section, E. Experimental design in analytical chemistry—part I: theory. *J. AOAC Int.* **97**, 3–11 (2014).

50. Porep, J. U., Kammerer, D. R. & Carle, R. On-line application of near infrared (NIR) spectroscopy in food production. *Trends Food Sci. Technol.* (2015). doi:doi.org/10.1016/j.tifs.2015.10.002

51. Weyer, L. & Workman Jr, J. *Practical guide to interpretive near-infrared*

spectroscopy. (CRC press, 2007).

52. Ciurczak, E. W. & Burns, D. A. *Handbook of near-infrared analysis*. (CRC Press, 2001).

53. Taylor, P., Drennen, J. K., Kraemer, E. G. & Lodder, R. A. Critical Reviews in Analytical Chemistry Advances and Perspectives in Near-Infrared Spectrophotometry Advances and Perspectives in Near-lnf rared Spectrophotometry. 37–41 doi:10.1080/10408349108051642

54. Manley, M. & Baeten, V. *Spectroscopic Technique : Near Infrared (NIR) Spectroscopy*. *Modern Techniques for Food Authentication* (Elsevier Inc., 2018). doi:10.1016/B978-0-12-814264-6.00003-7

55. Ozaki, Y., McClure, W. F. & Christy, A. A. *Near-infrared spectroscopy in food science and technology*. (John Wiley & Sons, 2006).

56. Igne, B. & Ciurczak, E. W. *Pharmaceutical and medical applications of near-infrared spectroscopy*. (CRC Press, 2014).

57. Ramírez-Morales, I., Rivero, D., Fernández-Blanco, E. & Pazos, A. Optimization of NIR calibration models for multiple processes in the sugar industry. *Chemom. Intell. Lab. Syst.* **159**, 45–57 (2016).

58. de Almeida, V. E. *et al.* Vis-NIR spectrometric determination of Brix and sucrose in sugar production samples using kernel partial least squares with interval selection based on the successive projections algorithm. *Talanta* **181**, 38–43 (2018).

59. Valderrama, P., Braga, J. W. B. & Poppi, R. J. Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy. *J. Braz. Chem. Soc.* **18**, 259–266 (2007).

60. Magwaza, L. S. *et al.* The use of Vis/NIRS and chemometric analysis to predict fruit defects and postharvest behaviour of 'Nules Clementine' mandarin fruit. *Food Chem.* **163**, 267–274 (2014).

61. Ribeiro, J. S., Ferreira, M. M. C. & Salva, T. J. G. Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy. *Talanta* **83**, 1352–1358 (2011).

62. Varmuza, K. & Filzmoser, P. *Introduction to multivariate statistical analysis in chemometrics*. (CRC press, 2016).

63. Ferreira, M. M. C., Antunes, A. M., Melgo, M. S. & Volpe, P. L. O. Chemometrics I: multivariate calibration, a tutorial. *Quim. Nova* **22**, 724–731 (1999).

64. Geladi, P. & Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **185**, 1–17 (1986).

65. Wold, S., Sjostrom, M., Sjöström, M. & Eriksson, L. PLS-regression, a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).

66. Wold, S., Trygg, J., Berglund, A. & Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **58**, 131–150 (2001).

67. Ferreira, M. M. C. Quimiometria: conceitos, métodos e aplicações. *Campinas, SP Ed. da Unicamp* **496**, (2015).

68. Brereton, R. G. Introduction to multivariate calibration in analytical chemistry. *Analyst* **125**, 2125–2154 (2000).

69. Gemperline, P. *Practical guide to chemometrics*. (CRC press, 2006).

70. Kramer, R. & others. *Chemometric techniques for quantitative analysis*. (Marcel Dekker New York, 1998).

71. Dayal, B. S. & Macgregor, J. F. Improved PLS algorithms. *J. Chemom.* **11**, 73–85 (1997).

72. Lindgren, F., Geladi, P. & Wold, S. The kernel algorithm for PLS. *J. Chemom.* **7**, 45–59 (1993).

73. Manne, R. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemom. Intell. Lab. Syst.* **2**, 187–197 (1987).

74. Wu, W. & Manne, R. Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications. *Chemom. Intell. Lab. Syst.* **51**, 145–161 (2000).

75. de Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **18**, 251–263 (1993).

76. Martins, J. P. A., Teófilo, R. F. & Ferreira, M. M. C. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. *J. Chemom.* n/a-n/a (2010). doi:10.1002/cem.1309

77. Kennard, R. W. & Stone, L. A. Computer aided design of experiments. *Technometrics* **11**, 137–148 (1969).

78. IUPAC. Units and their usage in spectrochemical analysis-ii. *Spectrochim. Acta B* **33**, 242 (1978).

79. Teofilo, R. F. Métodos quimiométricos em estudos eletroquímicos de fenóis sobre filmes de diamante dopado com boro. (2007).

80. Booksh, K. S. & Kowalski, B. R. Theory of analytical chemistry. *Anal. Chem.* **66**, 782A--791A (1994).

81. Lorber, A., Faber, K. & Kowalski, B. R. Net analyte signal calculation in multivariate calibration. *Anal. Chem.* **69**, 1620–1626 (1997).

82. Cuadrado, J. A. P. & Forn, M. P. Validación de métodos analíticos. *Barcelona Asoc. Española Farm. La Ind. AEFI* (2001).

83. Golbraikh, A. & Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **20**, 269–276 (2002).

84. Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **22**, 69–77 (2003).