MARIA LUÍSA SOARES DE ABREU

PREDICTION OF SUGARS IN RIPE BANANA PULP FROM SPECTRA OF UNRIPE FRUIT PEEL USING NEAR-INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

Dissertation submitted to the Agrochemistry Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

Adviser: Reinaldo Francisco Teófilo

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
A162p
2024

Abreu, Maria Luísa Soares de, 1998-
Prediction of sugars in ripe banana pulp from spectra of unripe fruit peel using near-infrared spectroscopy and multivariate calibration / Maria Luísa Soares de Abreu. – Viçosa, MG, 2024.
1 dissertação eletrônica (59 f.): il. (algumas color.).

Texto em inglês.
Orientador: Reinaldo Francisco Teófilo.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Química, 2024.
Referências bibliográficas: f. 54-59.
DOI: https://doi.org/10.47328/ufvbbt.2024.251
Modo de acesso: World Wide Web.

1. Açúcares (Química orgânica). 2. Banana. 3. Espectroscopia de infravermelho próximo. 4. Quimiometria. 5. Cromatografia líquida de alta pressão. 6. Sacarose. 7. Glicose. 8. Frutose. I. Teófilo, Reinaldo Francisco, 1978-. II. Universidade Federal de Viçosa. Departamento de Química. Programa de Pós-Graduação em Agroquímica. III. Título.

CDD 22. ed. 547.78

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

MARIA LUISA SOARES DE ABREU

PREDICTION OF SUGARS IN RIPE BANANA PULP FROM SPECTRA OF UNRIPE FRUIT PEEL USING NEAR-INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

Dissertation submitted to the Agrochemistry Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

APPROVED: February 23, 2024.

Assent:

Documento assinado digitalmente
gov.br
MARIA LUISA SOARES DE ABREU
Data: 29/07/2024 18:48:53-0300
Verifique em https://validar.iti.gov.br

Maria Luísa Soares de Abreu
Author

Documento assinado digitalmente
gov.br
REINALDO FRANCISCO TEOFILO
Data: 29/07/2024 20:34:18-0300
Verifique em https://validar.iti.gov.br

Reinaldo Francisco Teófilo
Adviser

To my mom.

AGRADECIMENTOS

Quero expressar minha profunda gratidão à minha maior fonte de inspiração, minha querida mãe, Paula. Agradeço por estar sempre ao meu lado, pelo apoio constante e por acreditar em mim mais do que eu mesma. Eu te amo! Agradeço também ao meu pai, José Roberto, pelo apoio.

Aos meus amigos, Thiago, Pedro, Mateus, Bia, Rafaela e Nelson, obrigada pela amizade, apoio e companheirismo.

Desejo expressar meus sinceros agradecimentos ao meu orientador, o Prof. Reinaldo. Agradeço pela oportunidade concedida, pela confiança depositada e pelo empenho, dedicação e cumplicidade demonstrados ao longo de toda a minha trajetória na UFV.

Agradeço de todo coração aos colegas e estagiários do laboratório LQA pelo incrível suporte e disposição em me auxiliar. Um agradecimento especial à Juliana, Igor, Helder e Wilson, pelos momentos preciosos de ensinamentos e insights químicos que compartilhamos. Não posso deixar de mencionar que a Juliana não foi apenas uma colega, mas também se transformou em uma grande amiga ao longo desse período. Quero expressar minha profunda gratidão por todas as palavras de apoio e incentivo que trocamos, tornando essa jornada acadêmica ainda mais significativa.

Também expresso meu agradecimento ao Prof. Márcio e Prof. Reinaldo pela disponibilidade em conceder o carro para a colheita das bananas, ao Prof. Jackson pelo fornecimento das bananas, e um agradecimento especial ao Sr. Nadil por toda a ajuda na colheita.

Além disso, expresso minha gratidão à Universidade Federal de Viçosa pela oportunidade de realizar o mestrado em um local conceituado e incrível. Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Agradeço à CAPES pela bolsa de estudos, que tornou possível a realização deste mestrado.

"Without the love of research, mere knowledge and intelligence cannot make a scientist."

(Irène Joliot-Curie)

ABSTRACT

ABREU, Maria L. Soares, M.Sc., Universidade Federal de Viçosa, February 2024. **Prediction of sugars in ripe banana pulp from spectra of unripe fruit peel using near-infrared spectroscopy and multivariate calibration.** Adviser: Reinaldo Francisco Teófilo.

The prediction of sucrose, glucose, and fructose concentrations in ripe banana pulp from spectra of both unripe and ripe fruit peel employing near-infrared spectroscopy (NIR) and multivariate calibration is the aim of this work. 320 unripe bananas were harvested, and NIR spectra were collected on the peel using diffuse reflectance. The sugars, i.e., sucrose, glucose, and fructose, were extracted from the pulp using an adapted procedure presented in this work. Subsequently, the extract was analyzed by high-performance liquid chromatography (HPLC) with an evaporative light-scattering detector. The sucrose, glucose, and fructose concentrations ranged from 45.10-207.20, 23.81-78.18, and 21.60-85.26 mg/g, respectively. Multivariate calibration models using partial least squares regression (PLS) were built to predict sugars from NIR spectra. Ordered predictors selection (OPS) was employed to select the most informative variables. The PLS models from the unripe peel for sucrose, glucose, and fructose presented root mean square error of prediction (RMSEP) values of 15.18, 3.41, and 8.7 mg/g, respectively. For the dataset using the ripe banana peel, the RMSEP values were 19.29, 5.29, and 6.65 mg/mL, respectively.

Keywords: Banana; NIR; Chemometrics; PLS; HPLC; Sucrose; Glucose; Fructose; Early prediction

RESUMO

ABREU, Maria L. Soares, M.Sc., Universidade Federal de Viçosa, fevereiro de 2024. **Previsão de açúcares na polpa da banana madura a partir de espectros da casca da fruta não madura usando espectroscopia no infravermelho próximo e calibração multivariada.** Orientador: Reinaldo Francisco Teófilo.

O objetivo deste trabalho é a previsão das concentrações de sacarose, glicose e frutose na polpa de banana madura a partir de espectros da casca de frutas tanto verdes quanto maduras, empregando espectroscopia no infravermelho próximo (NIR) e calibração multivariada. Foram colhidas 320 bananas verdes, e os espectros NIR foram coletados na casca usando reflectância difusa. Os açúcares, ou seja, sacarose, glicose e frutose, foram extraídos da polpa usando um procedimento adaptado apresentado neste trabalho. Posteriormente, o extrato foi analisado por cromatografia líquida de alta eficiência (HPLC) com um detector de espalhamento de luz evaporativo. As concentrações de sacarose, glicose e frutose variaram de 45,10 a 207,20; 23,81 a 78,18 e 21,60 a 85,26 mg/g, respectivamente. Modelos de calibração multivariada utilizando regressão por quadrados mínimos parciais (PLS) foram construídos para prever os açúcares a partir dos espectros NIR. A seleção de preditores ordenados (OPS) foi empregada para selecionar as variáveis mais informativas. Os modelos PLS da casca verde para sacarose, glicose e frutose apresentaram valores de erro quadrático médio de previsão (RMSEP) de 15,18; 3,41; 8,7 mg/g, respectivamente. Para o conjunto de dados usando a casca de banana madura, os valores de RMSEP foram 19,29; 5,29 e 6,65, mg/g respectivamente.

Palavras-chave: Banana; NIR; Quimiometria; PLS; HPLC; Sacarose; Glicose; Frutose; Previsão precoce.

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

CE	Capillary Electrophoresis
ELSD	Evaporative Light Scattering Detector
EMBRAPA	*Empresa Brasileira de Pesquisa Agropecuária*
FAO	Food and Agriculture Organization of the United Nations
GC	Gas Chromatography
HPLS	High-Performance Liquid Chromatography
MSC	Multiplicative scatter/signal correction
NAS	Net analytical signal
NIR	Near-Infrared Spectrometry
PLS	PartialLeast Squares
R	Correlation coefficient
Rc	Correlation coefficient of calibration
Rcv	Correlation coefficient of cross-validation
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 validation
Rp	Correlation coefficient of validation
SG	Savitzky-Golay
SNV	Standard normal variate scaling

LIST OF SYMBOLS

ºC	Degree Celsius
%	Percentage
mg	Milligram
g	gram
mL	Milliliter
t	tons

SUMMARY

1. INTRODUCTION

Bananas are one of the most consumed fresh fruits in the world, and their global market is worth around US$5 billion annually. The highlight is not without reason; it is one of the main sources of nutritious food, perpetuated in different cultures throughout history. [1,2]

In Brazil, around 6.854 million tons are produced in eight Brazilian states. Family farming accounts for half of this amount, generating an average of 500 thousand jobs. Globally, banana consumption reaches around 100 million people.[2,3]

Despite being a major producer and consumer of fruit, Brazil has an insignificant share (approximately 1% of total production) in the international market. One of the limiting factors is the loss of phytosanitary quality of the fruit, meaning factors such as pest control, diseases, and disorders are crucial to gaining new markets.[2]

Therefore, it is essential to understand the properties of the fruit that contribute to its quality. Fruit quality and maturity are typically determined by size, color, or texture. Visual inspection based on peel color suggests the fruit's sweetness. While widely used among producers and consumers, this method is subjective and inaccurate.[4] For example, between two ripe bananas, it is impossible to visually tell which one is sweeter just from the color of the peel.

In this regard, one of the main characteristics of bananas is that they are climacteric fruits, meaning even after harvest, the fruit will undergo physiological changes. Harvested while still green, during storage, there is an increase in respiratory rate due to the release of ethylene gas, which is responsible for fruit ripening and is suitable for the shelf life maintained for consumers.[4]

In the case of bananas, errors may occur in assessing their maturity based on peel color, changing from green to yellow due to chlorophyll degradation. Lina and colleagues (2016) reported that changes in peel color are not always consistent with sweetness inside the fruit, as the proteins involved in chlorophyll degradation differ from those present in starch degradation in the banana's cell wall.[5]

Despite the author's[5] emphasis that peel color is unrelated to fruit sweetness, some compounds may be present in the green peel directly or indirectly related to the sweetness of the fruit. They know in advance whether unripe bananas will be sweet when ripe is of interest to the market and scientists who study plant breeding. For the market, the consumer will be

satisfied with purchasing a fruit that pleases the palate when consumed. For agricultural scientists, knowing carefully about the sweetness of the fruit helps them select the best plants.

Another challenge in banana cultivation is determining the exact harvest time. The current methods are based on farmer experience and can negatively impact the banana supply chain. The longer the banana is attached to the tree, the higher its sugar concentration will be when ripe. However, if the banana is harvested prematurely, its sweetness will undoubtedly be low.

Therefore, this work proposes to use Near-Infrared Spectroscopy (NIR) to predict the sugar content in the ripe pulp from unripe banana peels. This early prediction type will help plant breeding and define the exact harvest time.

The group explored the early prediction method, and the results were satisfactory and promising. Oliveira et al. showed that it is possible to predict the amount of oil produced by ripe macaw based on NIR spectra of the unripe fruit's peel.[6] Porto et al. showed that it is possible to classify whether sugarcane will be more resistant or susceptible to attack by the sugarcane borer based on NIR spectra and Laser-Induced Breakdown Spectra (LIBS) before the plant is attacked.[7] These results show the potential of early prediction or non-obvious prediction for agriculture and its great applicability in plant breeding.

This study aimed to determine the optimal time to harvest ripe bananas with peak sugar levels, a major challenge farmers and agronomists face. The study proposed a method that uses Near-Infrared Spectroscopy (NIR) to predict sucrose, glucose, and fructose concentration in the pulp of ripe bananas. This method utilizes the NIR spectra of both unripe and ripe banana peels through multivariate calibration. This approach aimed to provide a reliable alternative for farmers, agronomists, and the market to determine the best time to harvest bananas. Implementing this approach would produce sweeter fruits after ripening and ensure the market gets high-quality produce.

NIR spectroscopy was chosen in this work because it is an environmentally friendly, non-destructive, non-invasive, fast, and relatively low-cost technique. Furthermore, the natural ripening process for bananas, which takes 13 to 22 days, was chosen. This long period justifies the early prediction of sugar concentration in ripe bananas from unripe ones.

2. LITERATURE REVIEW

2.1 The banana fruit

*2.1.1. **Overview***

Bananas are a diverse fruit, with approximately a thousand distinct types found worldwide. They are scientifically classified under the genus *Musa*. Bananas have a long history in Southeast Asia and the Western Pacific. They are believed to have been cultivated more than four thousand years ago in regions like India, Malaysia, and the Philippines.[8]

Indigenous communities in Brazilian territory cultivated bananas before the Portuguese colonization in 1500. The registration of agricultural activities in Brazil formally began in 1820. Brazil is among the world's top four banana producers, along with India, China, and Indonesia.[1,2,9]

In this regard, the fruit assumes paramount significance in society's dietary habits owing to its inherent nutritive qualities and delightful flavor. Consequently, it substantially impacts the economy by fostering growth in numerous sectors that derive their livelihood from its cultivation.[2]

The ubiquity of banana consumption transcends cultural boundaries, making it a prevailing dietary preference across diverse global cultures. Regarding market value, the banana trade has surged to unprecedented levels recently, marked by an estimated export volume of 21 million tons in 2019. This remarkable growth can be attributed to the substantial expansion of supply in the leading export nations, Ecuador and the Philippines, as well as a substantial uptick in import demand, notably from China and the European Union.[10]

Despite being a significant banana producer, Brazil's involvement in the global market remains relatively modest. Most of the country's banana yield is directed towards domestic consumption, establishing Brazil as the leading consumer of this fruit worldwide.

The data from the Food and Agriculture Organization of the United Nations (FAO) and EMBRAPA (2023) illustrate, in Figure 1, the participation of eight Brazilian states in banana production. The average production of the eight states is 6,811.000 tons. This average highlights the significance of national banana production, reinforcing its status as the most consumed fruit globally.[9]

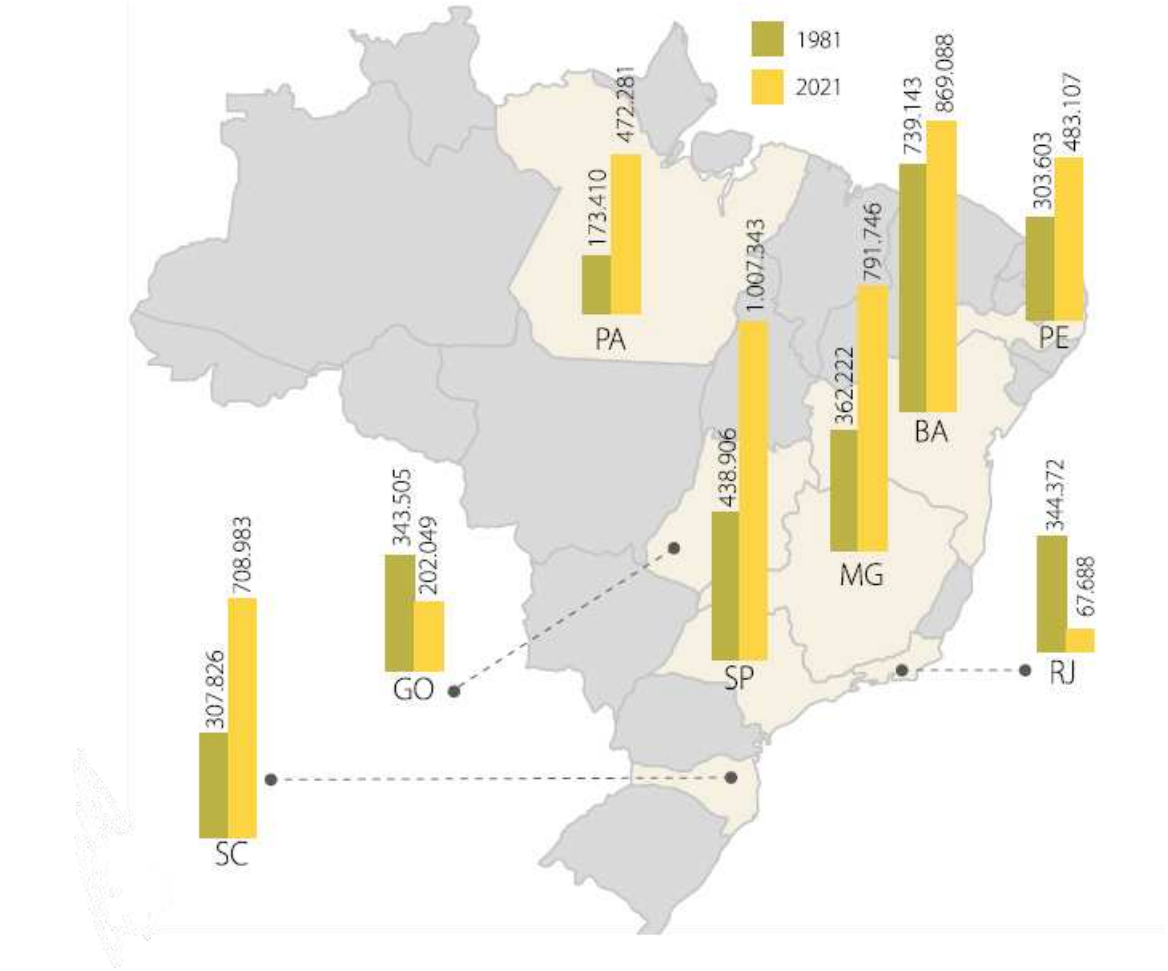

Figure 1. Tons of bananas were produced in Brazil in 1981 and 2021. (EMBRAPA, 2023)

2.1.2. Chemical changes during ripening

During the ripening process of a banana, some biochemical transformations occur. Von Loesecke (1950) highlighted the most noticeable change: the hydrolysis of starch and the accumulation of sugars. This process involves a reduction in the amount of starch in the fruit and a simultaneous increase in the concentration of sugars. Sensory-wise, the banana becomes yellow and develops a sweet taste. The fruit's acidity also increases until it reaches its peak when the banana is yellow, and its pH ranges between 5 and 5.6 when it is unripe. When the fruit is ripe, its pH is approximately 4. [11,12]

Regarding the green peel of the banana, its changes are related to the firmness of the fruit, which decreases until a yellowish color appears. These alterations occur due to the degradation of chlorophyll and the synthesis of carotenoids. [13]

2.2 High-Performance Liquid Chromatography

High-Performance Liquid Chromatography (HPLC) is a highly versatile and commonly used method of elution chromatography used for separation. Professionals from various fields use this technique to separate and identify different species in various matrices, including organic, inorganic, and biological substances. [14]

In HPLC, the mobile phase consists of a liquid solvent that carries the analytes dissolved in this phase. The analytes are separated due to their specific interactions with the stationary phase. The stationary phase, fixed within the chromatographic column, may exist in solid, liquid, or polymetric form. It plays a crucial role in separating sample components, interacting based on the sample's polarity and the nature of the stationary phase. The interaction between the analytes and the stationary phase results in the retention of analytes at various points throughout the chromatographic column. A chromatogram is obtained at the end of a chromatographic run. This graph represents the exit of analyte components over time or eluent volume. The chromatogram displays separate peaks corresponding to the different analytes of the sample that were separated during the chromatographic process.[15] After exiting the column, the analytes are detected using a specific, universal, or semi-universal detector. Choosing a suitable detector is essential to identify the components retained by the column.

Since carbohydrates do not absorb UV/Vis radiation, a semi-universal detector with an adequate detection limit is the evaporative light scattering detector (ELSD).[14,15] In the literature, some studies employ ELSD in HPLC to detect sugars in fruits, vegetables, cereals, plants, and processed foods.[16–20]

The operation of the ELSD is based on vaporization or nebulization of the mobile phase, followed by detecting light scattered by the resulting particles formed by the analyte. Therefore, the analyte must have a boiling point higher than the mobile phase to use ELSD.

The flow exiting the column is turned into an aerosol using a carrier gas, typically nitrogen. The aerosol containing droplets of the mobile phase with the analyte is carried by nitrogen or compressed air into a heated tube. The mobile phase is volatilized, and the analyte agglomerates form particles. The light scattered by these particles is captured by a photomultiplier positioned at a 90-degree angle relative to the incident light source (LED lamp). Adjusting the flow rate and temperature of the carrier gas according to the mobile phase is necessary. The detector's response is directly related to the concentration of the analyte.[21,22]

2.3 Near-infrared Vibrational Spectroscopy

Historically, when discussing the electromagnetic spectrum, it is essential to highlight that the infrared region was first discovered by William Herschel in 1800.[23] However, it required collaborative efforts after approximately 150 years to achieve the initial analytical applications. Initially, these applications were related to studies involving water in gelatin [24] and determining hydrocarbons in petroleum.

Pasquini (2018) attributed the delay to (a) the lack of commercial instruments for the near-infrared (NIR) region at the time to enable studies in the NIR spectral region, (b) the NIR spectra displaying broad and overlapping absorption bands, complicating their interpretation and assignment, and (c) the limited availability of essential mathematical resources to extract chemical information, representing a significant limitation.[25]

NIR corresponds to a part of the electromagnetic spectrum with vibrational energy of overtones and combinations of functional groups, covering the wavelength range of 750 to 2500 nm (10,000- 4000 cm^{-1}). In the NIR, organic compounds exhibit distinctive spectral features due to the relatively intense absorption of overtones and combination modes related to various functional groups. The main groups responding in the NIR are CH (aliphatic), CH (aromatic), C=O (carboxyl), OH (hydroxyl), and NH (amine and amide), which are frequently present in organic compounds (Figure 2). [26,27]

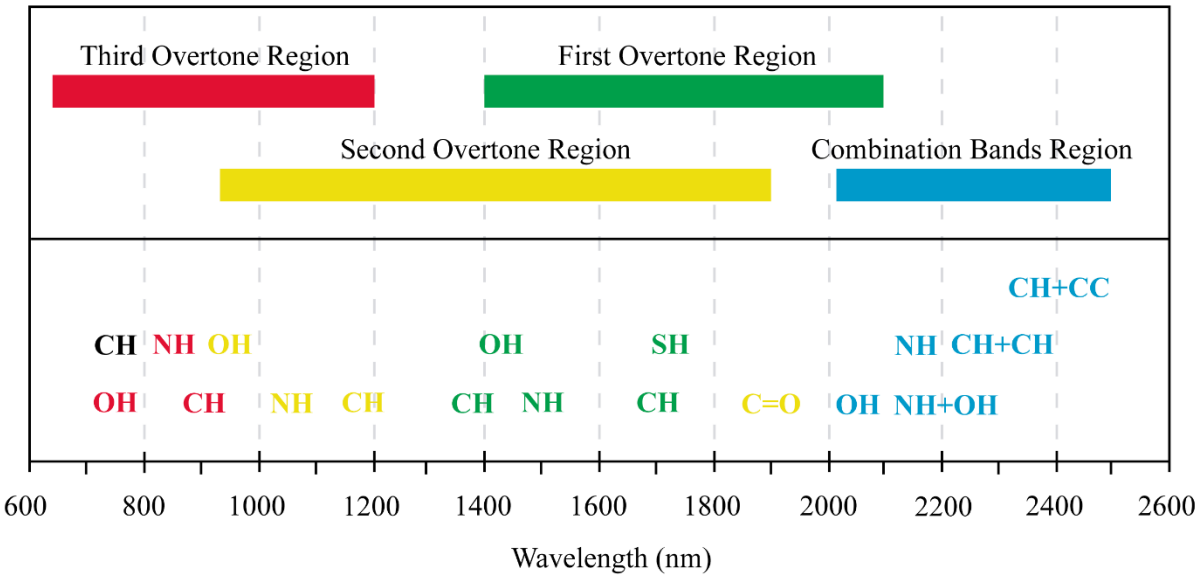

Figure 2. NIR overtones and combination spectral regions.

These overtones, combined with other effects such as neighboring groups, hydrogen bonding, crystallinity, phase separation, and thermal and mechanical influences, make analyses non-trivial.[28,29]

Since the NIR spectrum has a lot of overlapping signals, it is widely used together with chemometrics for quantitative and qualitative analysis. The NIR is generally chosen for its speed, low cost, and non-destructive characteristics towards the analyzed sample.[30] In the literature, studies can be found in fields such as agriculture [31,32], food chemistry [33,34], and the oil industry [6,35,36].

2.4 Chemometrics

2.4.1. Definition and Applications

Chemistry has strengthened itself with societal and technological advancements and means of communication. These advancements have enabled scientists to explore uncharted territories and expand their research methods. Scientific thinking now involves multivariate reasoning, thanks to the remarkable technological strides of the last century. [37–40] Data analysis with multivariate chemical information is called chemometrics.

Chemometrics is the science of relating multivariate measurements made on a chemical system to the system's state via mathematical or statistical algorithms.[41]

Based on Figure 3, the scope of chemometrics extends to various and complementary areas. On the one hand, it caters to fields such as engineering and computing, while on the other hand, it pertains to analytical chemistry, organic chemistry, and theoretical physical chemistry. One significant similarity among these areas is the importance of statistics.[42]

Chemometrics takes a data-driven perspective, unlike fundamental principles such as scientific laws and natural rules. This approach enables the creation of empirical statistical models that can predict values of important properties, even if they are not directly related. [43]

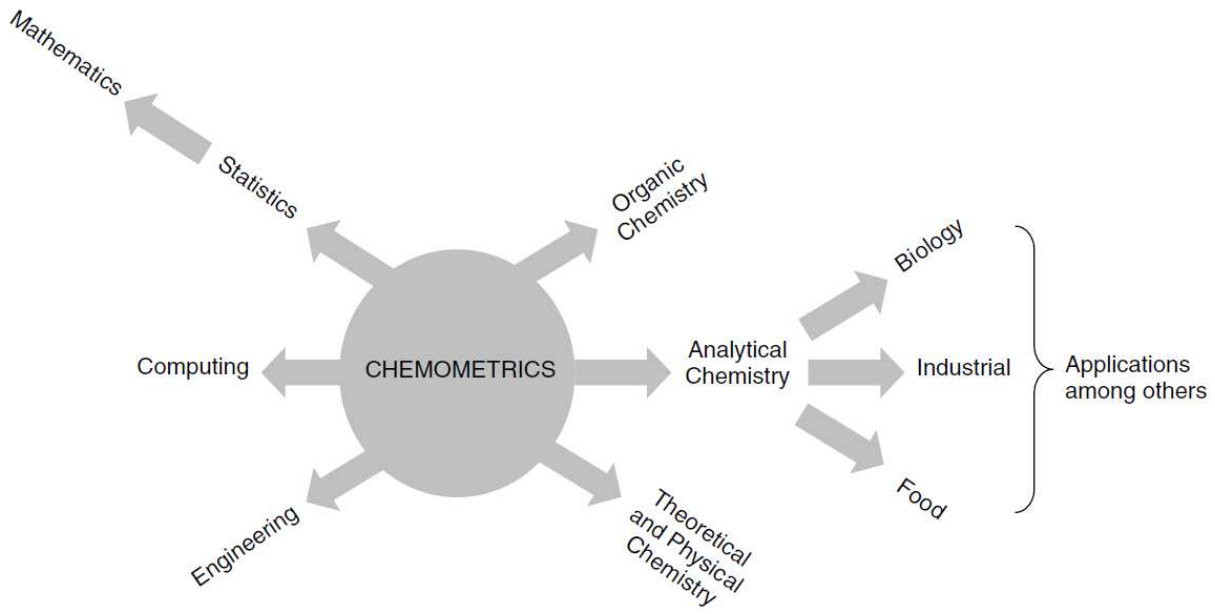

Figure 3. Fields where chemometrics is applied. (BRERETON, 2003)

2.4.2. Organization and Preparation of Data

For example, chemical data obtained from instruments, such as spectra, chromatograms, voltammograms, or diffractograms, can be transformed into digital information and presented as scalars, vectors, or matrices.[40]

In NIR spectroscopy, the band's positions are measured by the wavenumber, which is considered an independent variable. This process creates a dataset defined as a matrix **X** (*i*x*j*). The **X** matrix has *I* rows (samples) and *J* columns (variables), as illustrated in Figure 4. [37,40]

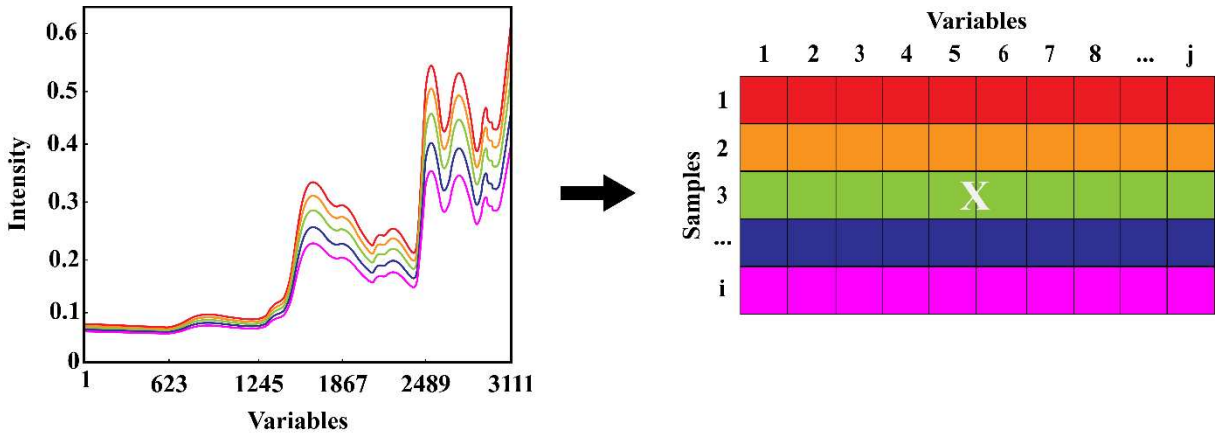

Figure 4. Representation of X matrix from instrument response. (Adapted from Cardoso, 2019)

Once data has been organized into a matrix, it is crucial to pretreat it before conducting chemometric analysis. Pretreatment is necessary to minimize undesirable variations, such as noise, that were not eliminated during data acquisition and cannot be addressed during analysis. Pretreatments can be divided into two types: the first, called transformations, is applied to the rows of the **X** matrix (sample), and the second, called preprocessing, is applied to the variables, which are the columns of **X**.[37]

It is essential to consider the data's composition during data transformation, as they may exhibit physical phenomena that can interfere with the multivariate regression step. Transformations such as baseline correction, smoothing, 1st derivative, and others are available to address this.

2.4.3. Multivariate Calibration Model

Multivariate regression or calibration is commonly used to build models and perform predictions of interest properties. Eq. 1 shows the linear system to be solved by the multivariate regression. The system presents the matrix X, the regression coefficient b, and the associated error term e, with y representing the dependent variable, which signifies the property interest for each compound, as exemplified in Figure 5. [31,42]

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \qquad \textbf{Eq.1}$$

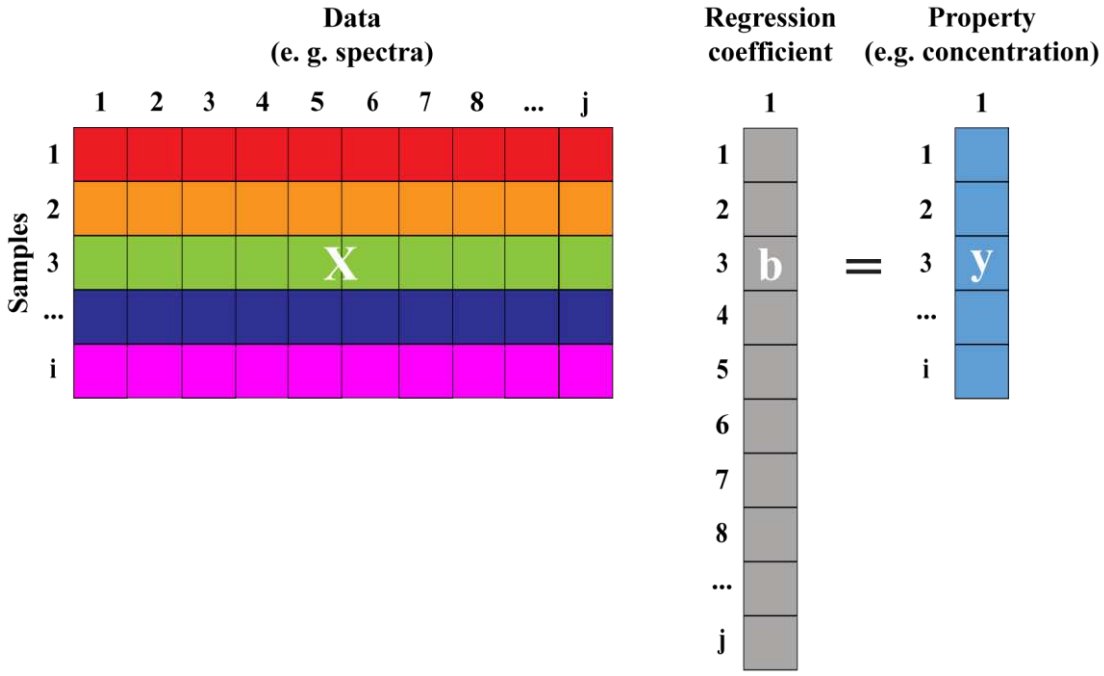

Figure 5. Representation of X matrix from the instrumental response. (Adapted from Cardoso, 2019).

The matrix **X** represents a data set showing the measured instrumental responses, i.e., the independent variables. Meanwhile, the vector **y** represents the dependent variables, such as concentration or biological activities. When solving the linear system shown in Eq.1 using the least squares with the Moore-Penrose inverse method, vector **b** is estimated. The vector **b** is the calibration model.

Subsequently, validation becomes a crucial step to ensure accuracy in predictions. When the model is validated, you can rely on it to make predictions.

2.4.4. Partial Least Squares (PLS)

The traditional modeling of the relationship between dependent variables (**y**) and independent variables (**X**) employs multiple linear regression (MLR), which is effective when there are more samples than variables and when the variables are linearly independent. However, advances in modern instrumentation have led to numerous and highly correlated **X** variable sets. Furthermore, the number of variables is enormous, i.e., 200 to 10000, making it challenging to obtain more samples than variables. Most of the time, these two reasons limit the use of MLR for instrumental chemical analysis data.

Data compression methods were used to solve the linear system where the matrix **X** has more columns than rows, and the columns of **X** are collinear. Data compression methods use principal components (PCs) or latent variables (LVs) to explain the variance of the space

generated by the columns of matrix **X**. Typically, fewer than 10 PCs or LVs are needed to explain the information contained in matrix **X**. Furthermore, the built PCs or LVs are orthogonal.

In regression, the vector **y** is projected onto the built PCs or LVs and not onto the original variables of **X**. This type of linear system resolution is known as Moore-Penrose pseudo inverse since the matrix **X** is not explicitly inverted, but the PCs or LVs are. Inverting the PCs or LVs is simple. Since they are orthogonal, just transpose them.

The most used data compression-based regressions in chemometrics are Partial Least Squares (PLS) and Principal Component Regression (PCR).

The PLS stands out in handling collinear **X** variables, enabling the exploration of complex issues and a more realistic data analysis. [44,45] Predictions using PLS present more accurate results than PCR since, during the building of the LVs, in addition to the decomposition of the **X** matrix, the **y** vector is used to increase the correlation and improve the prediction. So, the objective of PLS is to choose a set of latent variables that explain the variation in both **X** and **y**. [38,44]

2.4.5. Ordered Predictors Selection

PLS regression is useful for handling highly correlated independent variables, band overlaps, and experimental noise. However, the predictions using the model built with all variables may not always be satisfactory. The use of all variables may not accurately represent variations in the concentration of the target analyte in the sample. Therefore, it is crucial to select variables strategically to improve the model's efficiency and enable more accurate interpretations. In this context, a decisive step in multivariate regression is the variable selection process. Variable selection avoids redundant, irrelevant, or noise-representing variables to build models with enhanced predictive quality.[44,46,47]

Figure 6 displays a comprehensive scheme that explains the functionality of the OPS algorithm. It offers a detailed insight into each step involved in the process, from acquiring informative vectors to selecting the optimal subset of variables. Each phase is distinctly outlined.

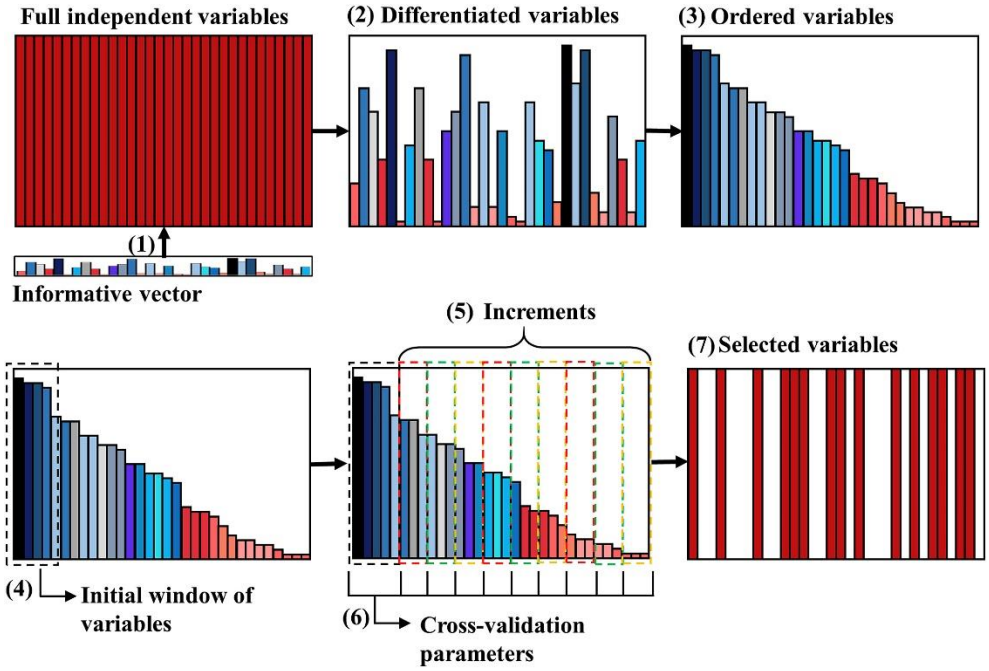

Figure 6. General scheme of variable selection steps using the PLS Algorithm. In (1), informative vectors are acquired; (2) The data matrix is differentiated based on the corresponding absolute values of the elements of the informative vector; (3) The differentiated variables are ranked in descending order; (4) A window is established for the building and evaluation of the first model; (5) The subset is gradually expanded by the addition of a fixed number of variables until all, or a portion of the variables are considered; (6) Cross-validation parameters are calculated for each model, and (7) The variable subsets are compared using the quality parameters obtained during the validation, and the best subset of variables is determined. (ROQUE J.V *et al*. 2019)

The OPS method uses an informative vector to capture the optimal locations of independent variables for prediction (Figure 6, step 1). Both independent and dependent variables are considered to compute these vectors, and the length of the vector is aligned with the number of independent variables. The original method for vector computation involves various techniques such as regression coefficients (REG), correlation assessments between each column of the **X** matrix and **y** (COR), residual data extracted from matrix reconstruction using a specified number of latent variables (SQR), variable importance in projection (VIP), net analyte signal (NAS), and covariance procedures (COV). [47,48]

Initially, the informative vector must be selected. From there, the independent variables are sorted based on their absolute values concerning the informative vector, as illustrated in Figure 6, step 2. The significance of each independent variable is determined by its highest absolute value, resulting in a descending order of magnitude in step 3 (Figure 6). Next, multivariate regression models are built and assessed using a cross-validation strategy in step 4. During this process, a subset of variables, or window, is selected to build and evaluate the

initial model. The matrix is then expanded by incrementally adding a fixed number of variables, culminating in building a new model. This expansion cycle continues until all or a portion of the variables are considered. [48]

Finally, the sets of evaluated variables are compared using the same parameters as the validation process. The model with the best quality parameters will contain the most effective predictive models, thereby selecting the variables. [47,48]

There are four approaches when applying OPS variable selection: autoOPS, feedOPS, autoiOPS, and feediOPS, each with its peculiarities. AutoOPS automatically performs all variable selection steps using various informative vectors. FeedOPS aims to provide feedback to the OPS algorithm, where the initially selected variables return for a new selection until more specific criteria are met. This strategy allows for looping until established rules are reached, with RMSE as a reference parameter. Lastly, feediOPS presents an approach that applies OPS to complete matrix subdivisions, called OPS intervals, selecting variables through specific matrix intervals.[47]

3. MATERIALS AND METHODS

3.1 Banana samples

The banana fruits utilized in this study originated from the collection of cultivars established in the Fundão Orchard, integrated into the UEPE (Teaching, Research, and Extension Unit) Tropical and Subtropical Fruit Culture of the Department of Agronomy (DAA) at the Universidade Federal de Viçosa (UFV), Viçosa, Minas Gerais State. 320 bananas, representing 19 varieties (Table 1), were harvested between May and October 2023.

Table 1. *Information about the type of cultivar harvested.*

Genomic group	Subgroup or type	Cultivar	Harvest	DtM[a]	ST / H[b]
AA	-	Ouro	Agosto	16	23.5 ºC /58%
		Ouro de Colatina	Outubro	12	23 ºC /45%
AAA	Cavendish	Nanica	Maio	15	23 ºC /59%
		Nanicão	Maio	20	23 ºC /59%
	Caru	Caru Roxa	Setembro	14	24 ºC /50%
		Caru Verde	Outubro	20	23 ºC /45%
			Outubro	22	23 ºC /45%
	-	Caipira	Agosto	16	23.5 ºC /58%
			Outubro	20	23 ºC /45%
AAB	Prata	Prata	Setembro	14	23.2 ºC /50%
			Outubro	15	23 ºC /45%
		Prata-anã	Agosto	16	23.5 ºC /58%
			Setembro	14	24 ºC /50%
		Prata Catarina	Agosto	16	23.5 ºC /58%
		Prata Gorutuba	Outubro	20	23 ºC / 45%
		Pacovan	Agosto	16	23.5 ºC /58%
			Setembro	15	24 ºC /50%
			Outubro	15	23 ºC /45%
	Terra	Terrinha	Julho	13	23 ºC /59%
			Setembro	15	24 ºC /50%
	Maçã	Maçã (verde)	Julho	21	23 ºC /59%
			Outubro	20	23 ºC /45%
		Maçã (roxa)	Julho	13	23 ºC /59%
			Setembro	15	23.2 ºC /50%
			Outubro	13	23 ºC /45%
	Mysore	Mysore	Agosto	16	23.5 ºC /58%
			Outubro	15	23 ºC /45%
AAAB	Tipo Pacovan	BRS Japira	Setembro	14	23.2 ºC /50%
	Tipo Prata	Prata Graúda	Outubro	15	23 ºC /45%

a. Days to maturity, b. Storage temperature/humidity

3.2 Spectral analysis (independent variables)

To aid in understanding the process of obtaining bananas from the bunch, Figure 7 summarizes the process from acquiring the bunch to harvesting and separating the hands and individual bananas.

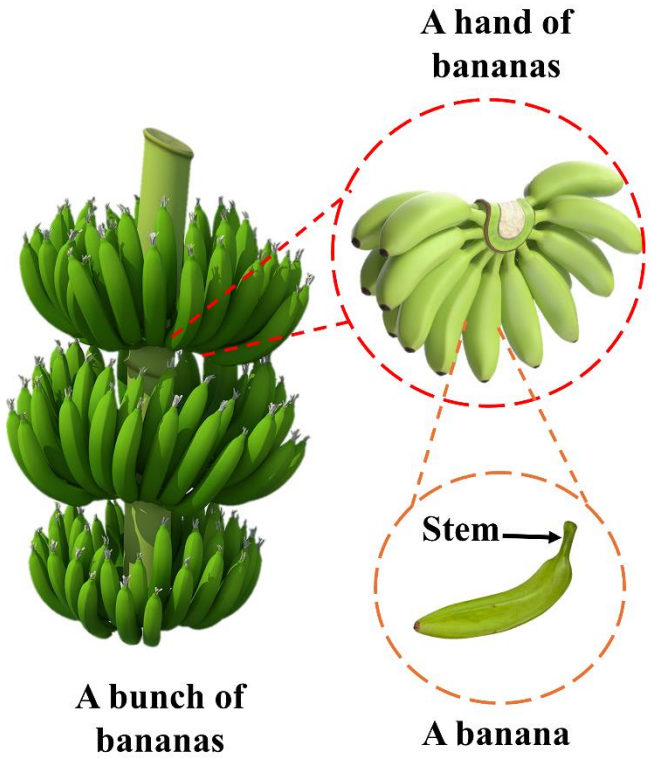

Figure 7. Components of banana harvest.

The banana harvest involves removing the entire bunch from the banana plant when it is still unripe but has reached the appropriate stage for harvesting. Each bunch contains more than five hands, each containing several bananas. In this study, three bananas were harvested from four hands belonging to a bunch of nineteen different cultivars.

These bananas were harvested while keeping the stem of each hand intact, and the three bananas represent a triplicate from each hand.

The spectra were obtained from the stem, middle, and tip for both unripe and ripe bananas, as presented in Figures 8B and C. The ripe bananas were allowed to ripen naturally without external intervention until they turned yellow with brown spots. The spectra of the three parts were then averaged for unripe and ripe bananas, according to Figure 8.

The NIR spectra were obtained using a Thermo Scientific Antaris II with Fourier Transform spectrometer and integration sphere. The range investigated was 10,000- 4000 cm^{-}

[1]. The spectra were obtained using the TQ Analysis software and stored as log (1/R), where R is the reflectance collected. For each NIR spectrum (stem, middle, and tip) acquired in the instrument, 32 scans were performed to increase the signal-to-noise ratio. The average of these scans was stored.

The harvesting of banana bunches involved excluding the first hand and collecting four hands, represented as 1 and 2 in Figure 6A.

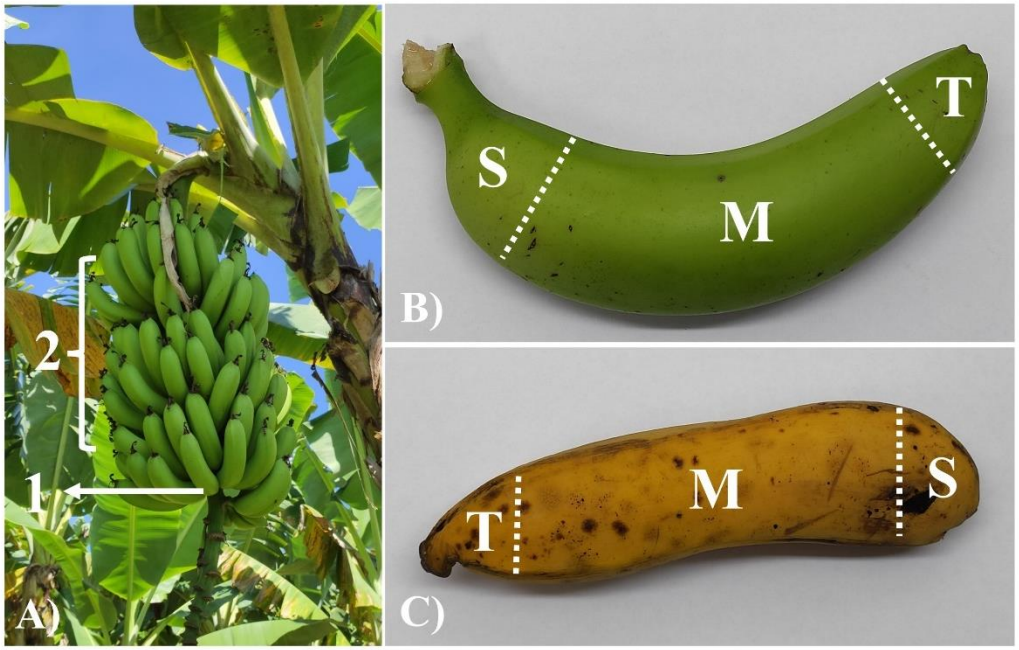

Figure 8. Bunch harvesting involves excluding the first hand represented by 1 and harvesting the other hands by 2 (A). The division was performed for NIR analysis on the unripe banana, with S representing stem, M representing medium, and T representing tip (B), The same division was carried out on the ripe banana (C).

Figure 9 illustrates NIR spectra acquisition in unripe and ripe bananas without sample preparation.

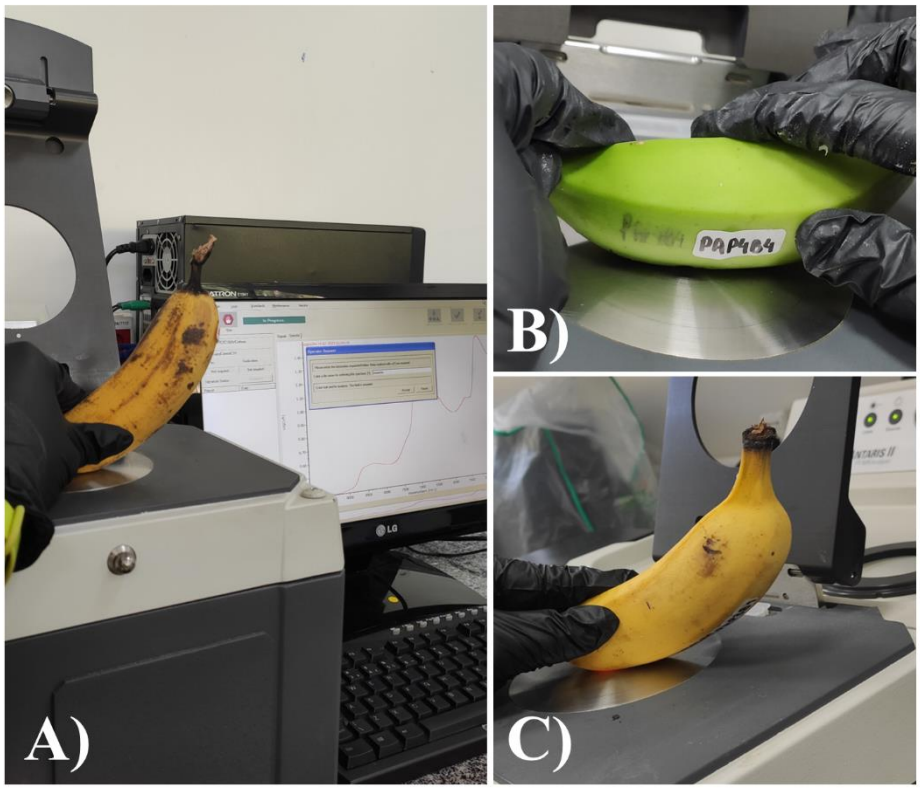

Figure 9. Spectra taken from the tip with ripe bananas (A) and (C). Spectra taken from the middle in unripe banana (B).

3.3 Extraction of soluble sugars from bananas

Adapted methods were employed to extract the soluble sugars from banana pulp.[49–51] This involved mixing approximately 1.0 grams of previously triturated banana pulp with approximately 400 mg of calcium carbonate in 3 mL of distilled water under agitation for 20 minutes, followed by centrifugation for 15 minutes at 4000 RPM. The process was repeated three times, combining the obtained supernatants for subsequent dilution in a 25 mL volumetric flask. Before being transferred to vials, the sample was pre-filtered using a nylon syringe filter with 0.45 µm pores and a diameter of 13 mm.

The inhibition of the invertase enzyme in sugar quantification is essential in this analysis since sugar concentrations can be changed during the analytical procedure. In the presence of this enzyme, the sucrose is broken down into fructose and glucose. Hence, calcium carbonate was employed for this purpose. The enzyme is inactivated by raising the mixture's pH with calcium carbonate. Calcium carbonate was also used to adjust the pH to the conditions required by the chromatographic column.

3.4 Validation of the extraction method

Validation of the extraction method involved conducting accuracy/recovery tests, in which sample triplicates were analyzed before and after fortification with sucrose, glucose, and fructose calibration standards at three concentration levels: 2, 10, and 20 mg/mL.

When recovery is achieved through fortification of the blank matrix, the recovery factor can be calculated using Eq. 2. So, the average recovery factor and the coefficient of variation (Eq. 3) can be calculated. [52]

$$f_{rec} = \frac{C_f - C_{nf}}{C_{ad}} \; x \; 100 \qquad \textbf{Eq.2}$$

$\boldsymbol{C_f}$ represents the measured concentration after fortifying the blank matrix, $\boldsymbol{C_{nf}}$ denotes the measured concentration in the non-fortified blank matrix, and $\boldsymbol{C_{ad}}$ means the concentration of the pure analyte added to the blank matrix.

$$CV = \frac{s}{\bar{x}} \; x \; 100 \qquad \textbf{Eq.3}$$

where *s* denotes the standard deviation, representing the data dispersion measure, and $\overline{x}$ represents the arithmetic means of the data.

3.5 HPLC (dependent variables)

The analysis of sugars in bananas was performed using a Shimadzu HPLC, model 20AT Prominence, with LabSolutions software. The HPX-87P column (Bio-Rad® Aminex) was employed for sugar separation, featuring a particle size of 9 µm, a pH range of 5 to 9, a length of 300 mm, and a diameter of 7.8 mm. The mobile phase consisted of 20% acetonitrile and 80% ultrapure water in isocratic mode. Separation parameters included a constant temperature of 80 ºC, a flow rate of 0.850 mL/min, and an injection volume of 10 µL. The ELSD was used to detect sugars in the sample.

3.6 Analytical curve

To quantify sugars via HPLC, a calibration curve was built at ten concentration levels, ranging from 2 to 20 mg/mL, for sucrose, glucose, and fructose. Each preparation of the analytical curve levels was carried out in triplicate to ensure accuracy and precision.

The calibration curve was quadratic, with inverse regression applied, where the x-axis (horizontal axis) represents the area, and the y-axis (vertical axis) represents concentration values. The model is represented by Eq. 4**.**

$$y = b_0 + b_1 x + b_2 x^2 \qquad \textbf{Eq. 4}$$

where y is the concentration of sugars, b_0, b_1, and b_2 are the coefficients, and x is the peak area in mg/mL.

3.7 Multivariate regression models

As shown in Figure 10, the models were developed based on two matrices of spectra from distinct sources. One of these matrices was generated from spectra of unripe banana peels with dimensions of 110 rows (samples) and 3112 columns (variables), while the other originated from spectra of ripe bananas, comprising 101 rows and 3112 columns. Calibration models were built for each sugar, i.e., sucrose (SUC), glucose (GLU), and fructose (FRU), based on the corresponding spectra of unripe and ripe banana peels. It is important to note that the dataset comprising all harvested bananas was analyzed, and in pursuit of better results, the decision was made to utilize the average of the triplicates for each hand.

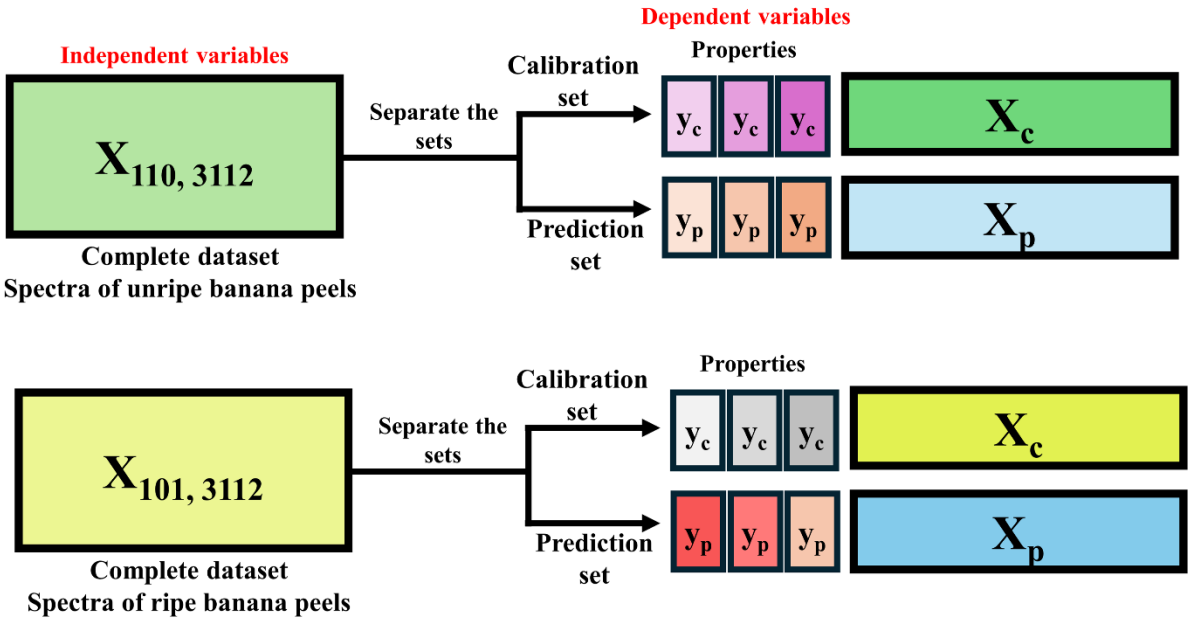

Figure 10. Two datasets containing spectra obtained from unripe and ripe banana peels were separated into two sets, one for calibration and the other for prediction, along with the vectors y representation sucrose, glucose, and fructose were modeled.

Each value is used to construct vector y with the average of triplicates of bananas from each hand, representing the dependent variables SUC, GLU, and FRU quantified through HPLC.

After obtaining the NIR spectra with the instrument software, the data were exported to Matlab R2019a (MathWorks, Natick, USA). Subsequently, the data was exported to Excel (Microsoft Co., USA) to be imported into the Relyon software (https://github.com/wjcardoso/Relyon).

The Relyon software was used in the pretreatment steps, data division into test and training sets, variable selection using the OPS, calibration model building, and prediction.

Twenty-one data pretreatment methods were studied using Relyon software. After optimization, the best preprocessing combination was chosen.

The division into test and training sets was executed using the Kennard and Stone algorithm, with 25% separated for the test set. Random cross-validation with 10 splits was applied. The maximum number of selected latent variables was set to 10.

Four OPS methods were performed to select the most predictive variables: auto OPS, feed OPS, autoiOPS, and feediOPS. A window of 20 and an increment of 5 were utilized. All vectors available in the Relyon software, including the regression vector, were used.

3.8 Models' validation

The model's quality was assessed by examining two key parameters: root mean square error (RMSE) and correlation coefficient (R). It is expected that the RMSE value to be less than the smallest value found in **y**, i.e., min(y). When RMSE is smaller than the smallest value found in **y** divided by 10 (RMSE < (min(**y**)/10)), it indicates an excellent model fit without overfitting. However, it is expected that at least RMSE be less than the smallest value of **y**. On the other hand, the R-value should be closer to 1. The RMSE and R were calculated using Eq. 5 and Eq. 6, respectively.

From Equations 5 and 6, **y** and $\overline{\boldsymbol{y}}$ represent the observed values and the mean of the observed values, respectively. *N* represents the number of samples. $\widehat{\boldsymbol{y}}\mathbf{i}$ and $\overline{\widehat{\boldsymbol{y}}}$ denote the estimated value and mean of estimated values, respectively. In the cross-validation (CV) context, *N* means the number of samples in the cross-validation set. The corresponding error and correlation coefficient are named root mean square error of cross-validation (RMSECV) and correlation coefficient of cross-validation (R_{CV}), respectively. For external validation, *N* means the number of samples predicted (P). In this scenario, the error and correlation

coefficients are denoted as correlation root mean square error of prediction (RMSEP) and correlation coefficients of prediction (R_P), respectively.[6]

$$RMSE = \frac{\sqrt{\sum_i^N (yi - \widehat{yi}}}{N} \quad \textbf{Eq.5}$$

$$R = \frac{\sum_{i=1}^N (\widehat{yi} - \widehat{\bar{y}})\,(yi - \bar{y})}{\sqrt{\sum_{i=1}^N (\widehat{yi} - \widehat{\bar{y}})^2}\sqrt{\sum_{i=1}^N (yi - \bar{y})^2}} \quad \textbf{Eq.6}$$

4. RESULTS AND DISCUSSION

4.1 Dependent variables

A quadratic fit was observed to provide the best adjustment to the models for the three sugars in the sucrose, glucose, and fructose quantification using HPLC-ELSD. This choice was based on identifying a trend in the regression plots, better determination coefficient values, and lower deviations from the repetitions used to construct the curve.

The models' coefficients of determination (R^2) were 0.998, 0.999, and 0.998 for sucrose, glucose, and fructose, respectively. The models for the respective sugars are presented in Eq. 7, 8, and 9.

$$C = (1.46 + (2.76\ x\ 10^{-6}) \times\ [Area] + (-\ 6.06\ x\ 10^{-14}\ ) \times\ [Area]^2) \quad \textbf{Eq. 7}$$

$$C = (1.37 + (1.52\ x\ 10^{-6}) \times [Area] + (-\ 1.87\ x\ 10^{-14}) \times [Area]^2) \quad \textbf{Eq. 8}$$

$$C\ = (1.62 + (4.45\ x\ 10^{-6}) \times [Area] + (-\ 1.84\ x\ 10^{-13}) \times\ [Area]^2) \quad \textbf{Eq. 9}$$

The range of individual sugars in all banana samples varied as follows in Table 2:

Table 2. *Descriptive statistics of the dependent variables.*

Sugars	Min. (mg/ g)	Max. (mg/ g)	Mean (mg/ g)	Std
Sucrose	45.10	207.20	123.74	37.48
Glucose	23.81	78.18	53.27	10.04
Fructose	21.60	85.26	49.90	14.02

"Min" represents minimum, "Max" represents maximum, and Std represents standard deviation.

Figure 11 presents the chromatograms of soluble sugars extracted from ripe banana pulp and the sugar standard. It can be observed that the peak of sucrose appears around 8 minutes, glucose around 10 minutes, and fructose around 14.5 minutes.

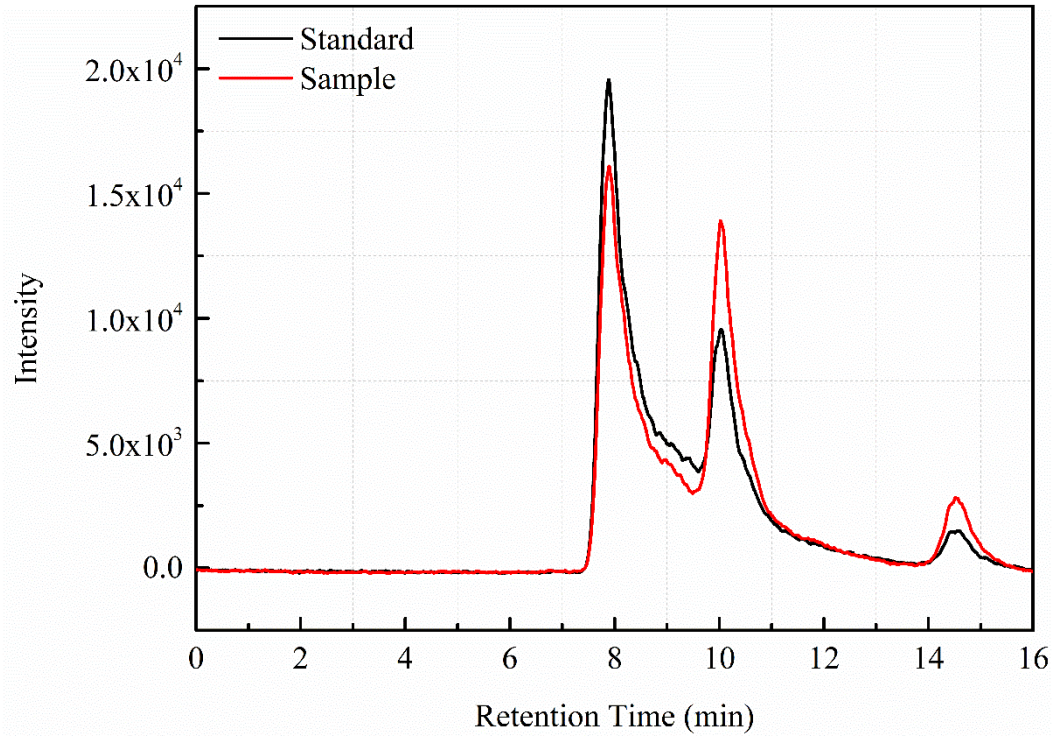

Figure 11. Chromatograms of the soluble sugars extracted from ripe banana pulp (red line) and the sugar standard (black line).

Figure 12 presents the Pearson Correlation matrix for the dependent variables. It is possible to observe that only glucose and fructose presented a high positive correlation (0.85). Sucrose presented a low correlation with glucose and fructose, i.e., -0.62 and -0.30 respectively. The negative correlation is related to the conversion of sucrose into glucose and fructose during the ripening process of bananas.

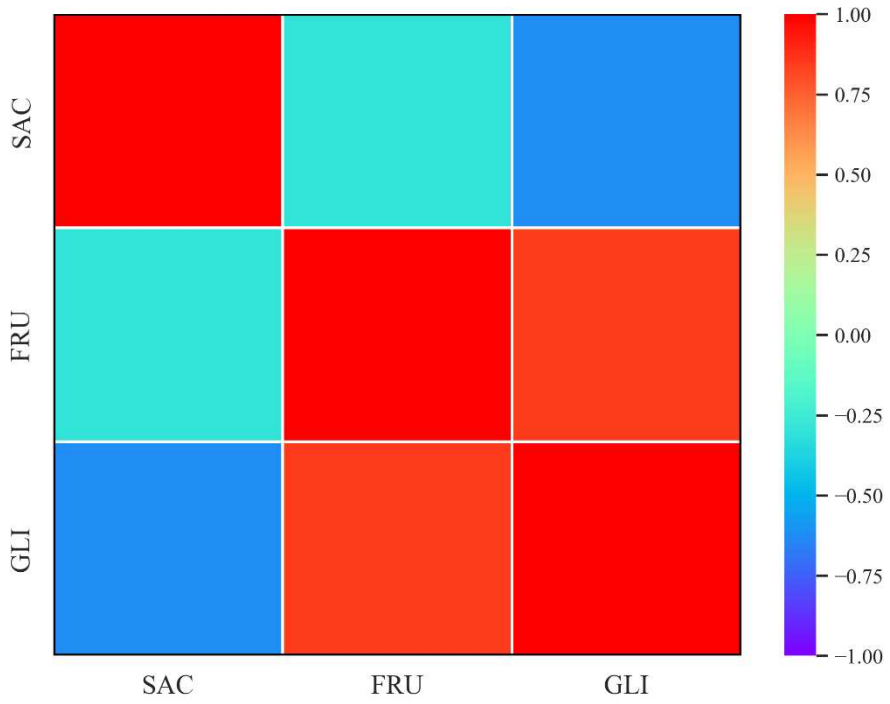

Figure 12. Correlation matrix of the dependent variables sucrose (SAC), glucose (GLI), and fructose (FRU).

4.2 Evaluation of the Extraction Method

The recovery parameter was assessed at three concentration levels to evaluate the accuracy of sugar extraction from bananas. Additionally, the average of the three recovery levels was computed and presented in Table 3.

Table 3. *Average of the recovery factors from the extraction and their coefficient of variation.*

Sugars	f_{rec} (%) **1st level**	f_{rec} (%) **2nd level**	f_{rec} (%) **3rd level**	**CV (%)**
Sucrose	88	84	83	3%
Glucose	90	89	91	1%
Fructose	92	93	90	2%

* f_{rec} *is the recovery factor and is calculated using the Eq.2.*

Table 3 presents the results related to one of the main parameters for evaluating the effectiveness of an extraction method. For the extraction of sucrose, glucose, and fructose from banana pulp, the recovery factor values were positive and indicative of nearly complete extraction of the sugars, with the recovery factor reaching close to 90% for all three concentration levels used.

The complexity of the matrix worked on is highlighted due to the presence and action of bio enzymes catalyzing the process of converting sucrose into glucose and fructose. Therefore, the use of calcium carbonate was essential to inhibit this action. This method proved to be effective for the accurate extraction of sugars from banana pulp.

4.3 Spectral interpretation

Figure 13 displays the pure spectra of water, sucrose, glucose, and fructose. These spectra reveal a subtle similarity among the sugars and a difference in the absorption intensities of water.

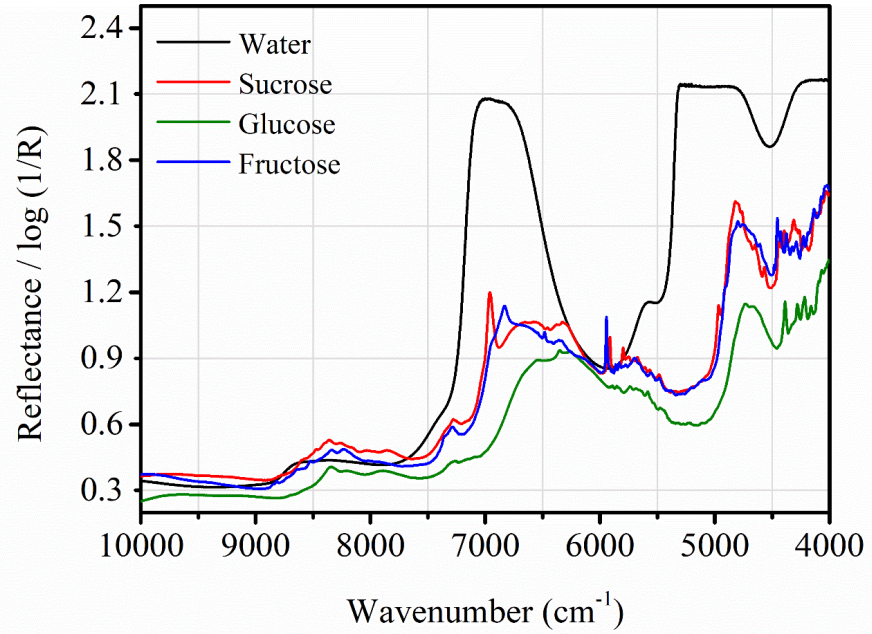

Figure 13. Spectra of pure water, sucrose, glucose, and fructose.

Despite the similarity among the characteristics of sucrose, glucose, and fructose, they exhibit differences around 7000-6000 cm^{-1} and 5000-4000 cm^{-1} due to variations in absorbance related to the OH, CH, and CO bands. Conversely, water, which accounts for approximately 73% of bananas [49], shows high absorbance in the NIR, displaying characteristic OH bands around 7000 cm^{-1}. This significant presence of water in the spectra justifies its detection and contribution to the spectral characteristics.[53]

The spectra obtained from the peel of both unripe and ripe bananas are presented in Figure 14. They represent the average spectra from three parts where the spectra were taken: the peel, the stem, and the tips of the bananas, both ripe and unripe.

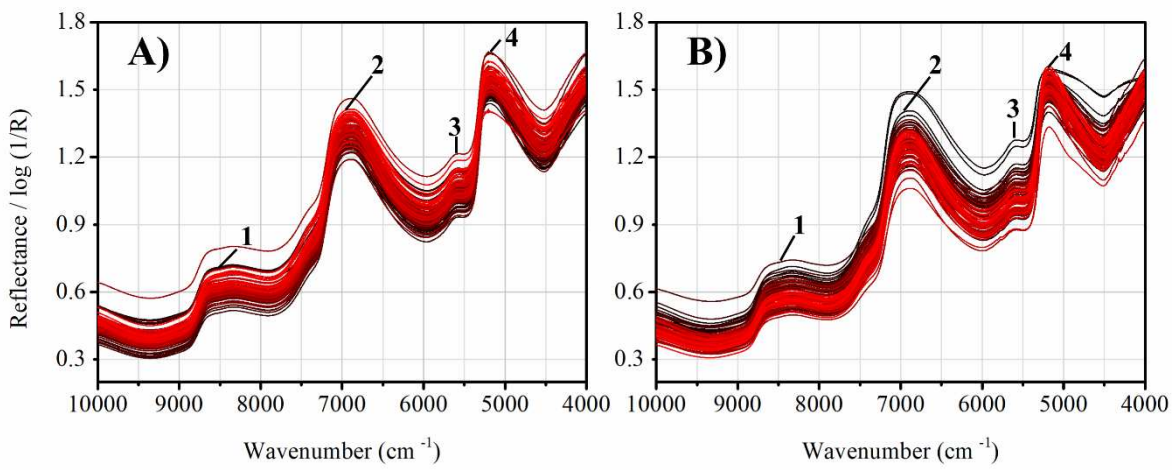

Figure 14. Spectra of green peel (unripe) (A), spectra of the yellow peel with brown spots (ripe) (B). The spectra of both ripe and unripe bananas, as shown in Figure 10, exhibit significant similarity with minor changes in reflectance near 5000 cm $^{-1}$.

The characteristic peak of water in Figure 13 is present in the spectra of Figure 14, where characteristic water bands are visible around 7200 cm^{-1} to 6500 cm^{-1} (region 2) and around 5200 cm^{-1} (region 4), as depicted in Figure 14 (A) and (B). [6,54]

The spectrum shows the areas where overtones and combinations of functional groups can be found, as illustrated in Figure 14 (region 1). The second region, which contains overtones of C-H and O-H groups, can be observed between 9000 to 8000 cm^{-1}. In Region 3, the first overtone of the CH group and CO is visible around 5900 to 5400 cm^{-1}. Lastly, the region around 4500 cm^{-1} signifies the combination of functional groups CH+CH; all of these functional groups can be correlated with the structures of sugars. [55]

4.4 Modeling

Figure 14 displays the preprocessing and transformations applied to the spectra dataset from unripe and ripe banana peel. The transformations presented were performed using the Relyon software.

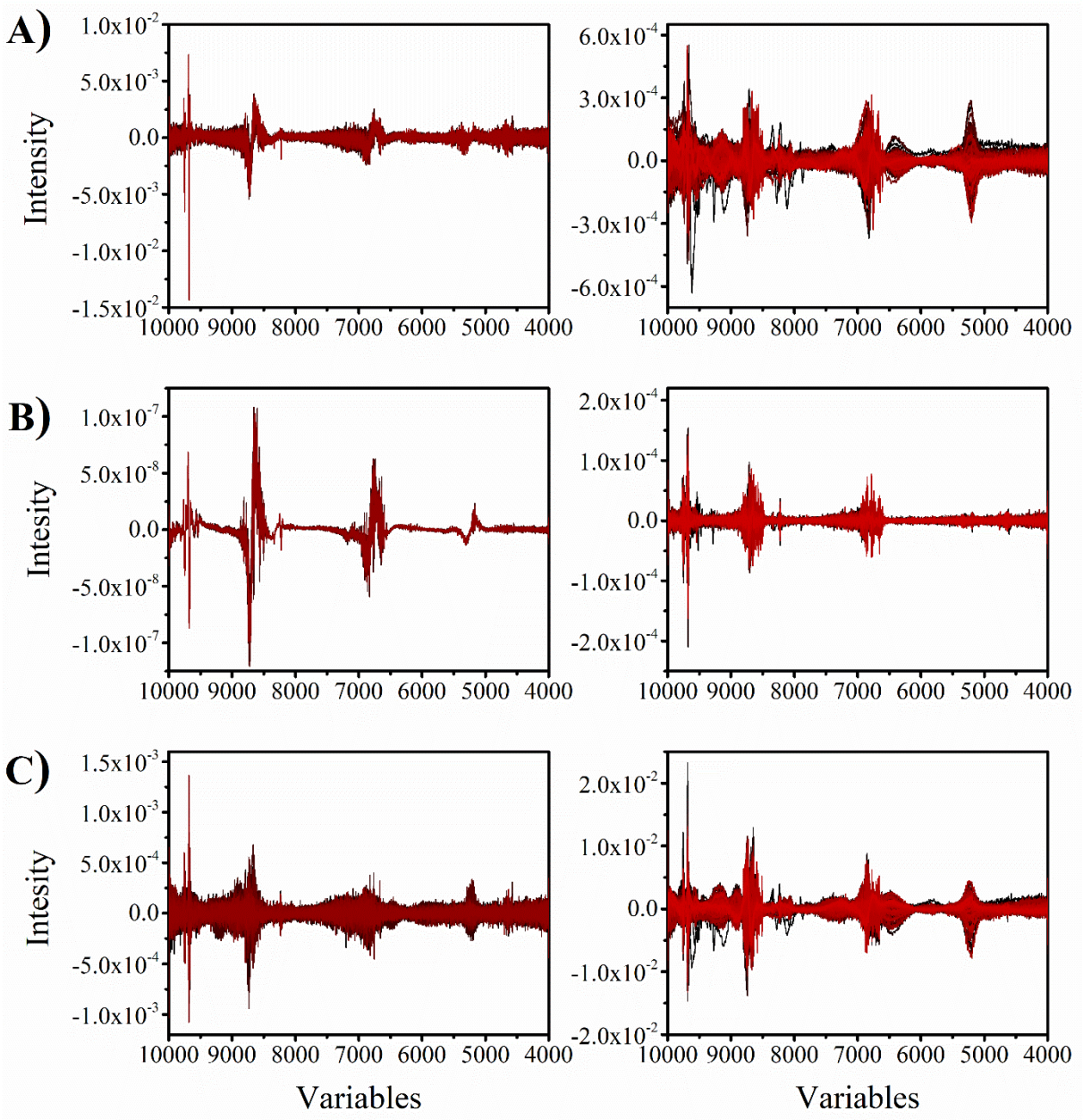

Figure 15. Different preprocessing strategies were applied to the spectra of unripe and ripe banana peels for sucrose (A), glucose (B), and fructose (C) analysis. Unripe banana peels underwent the second derivative, NAS, and normalization treatments for sucrose (A), while ripe banana peels were subjected to the first derivative, smoothing, and NAS treatments. For glucose (B), unripe bananas were normalized, subjected to MSC, and underwent second derivative treatment, whereas ripe bananas were treated with NAS, second derivative, and mean centering. Regarding fructose (C), unripe banana peels were processed with first derivative, NAS, and detrend treatments, while ripe banana peels were treated with first derivative, normalization, and NAS strategies.

Extensive optimizations of preprocessing techniques and transformations were combined and used on the original spectra of ripe and unripe bananas. Each optimization resulted in around 1756 combinations. Each dataset was focused on sugars and was closely analyzed to

find the best combination of transformation and preprocessing. The goal was to achieve the smallest RMSECV and high Rcv. Once the optimal combination was identified, it was applied to the spectra of banana peels, both ripe and unripe, using distinct transformations and preprocessing steps.

Figure 14 presents the preprocessing and transformations used for the sucrose dataset in the unripe peel. The transformation and preprocessing steps, which were Second Derivate, Net Analytical Signal (NAS), and Normalization, were applied. On the other hand, transformations such as MSC, Second Derivate, and Normalization were used for the dataset of glucose in the unripe peel. Transformations including NAS, First Derivative, and Detrend were applied to the dataset of fructose in the unripe peel.

Specific transformations and preprocessing steps were also employed for the datasets of ripe bananas. For sucrose, NAS, Smooth, and First Derivative were used. The Second Derivative, NAS, and Mean Center were applied for glucose. First Derivate, Normalization, and NAS were used for fructose.

Table 4 shows the statistical parameters for the models built using PLS with variable selection, employing AutoiOPS, FeediOPS, AutoiOPS, FeediOPS, and using all variables (Full). After comparing the models for sugars present in unripe and ripe banana peels, it was found that the best parameters were achieved when using the FeediOPS method for variable selection and AutoiOPS for only fructose in ripe bananas.

Table 4. Statistical parameters for the PLS models with all variables (Full) and variables selected using AutoiOPS, FeediOPS, AutoOPS, and FeedOPS.

Analysis	Constituent	Method	Vars	RMSEC	Rc	RMSECV	Rcv	RMSEP	Rp	LV	Strategy
Unripe	Sucrose	**Full**	**3112**	**13.03**	**0.94**	**31.89**	**0.56**	**39.13**	**0.35**	4	
		AutoiOPS	610	6.96	0.98	18.88	0.89	22.07	0.87	4	
		FeediOPS	**155**	**5.53**	**0.99**	**13.41**	**0.95**	**15.18**	**0.93**	**4**	2nd Der (W:9 O:2) + NAS (1) + Norm (1)
		AutoOPS	1210	5.87	0.99	19.57	0.89	21.92	0.86	4	
		FeedOPS	160	6.02	0.99	12.43	0.95	15.33	0.94	4	
	Glucose	**Full**	**3112**	**3.76**	**0.94**	**8.23**	**0.68**	**5.69**	**0.90**	**7**	
		AutoiOPS	155	2.41	0.98	4.83	0.90	4.12	0.96	7	
		FeediOPS	**145**	**2.30**	**0.98**	**4.17**	**0.93**	**3.41**	**0.97**	**7**	Norm (1) + MSC + 2nd Der (W:21 O:2)
		AutoOPS	300	2.33	0.98	4.93	0.90	4.10	0.96	7	
		FeedOPS	140	1.97	0.98	4.29	0.93	3.19	0.97	7	
	Fructose	**Full**	**3112**	**3.83**	**0.96**	**7.22**	**0.85**	**9.85**	**0.83**	**4**	
		AutoiOPS	365	3.53	0.97	6.03	0.90	8.75	0.86	4	
		FeediOPS	**50**	**3.93**	**0.96**	**5.75**	**0.91**	**8.70**	**0.85**	**4**	1st Der (W:5 O:2) + NAS (3) + Detrend (O:3)
		AutoOPS	375	4.24	0.95	5.78	0.91	9.01	0.86	3	
		FeedOPS	300	3.40	0.97	6.16	0.89	8.31	0.86	4	
Ripe	Sucrose	**Full**	**3112**	**18.44**	**0.88**	**24.97**	**0.78**	**22.86**	**0.76**	**5**	
		AutoiOPS	1065	14.95	0.93	22.38	0.82	20.39	0.82	5	
		FeediOPS	**250**	**14.58**	**0.93**	**21.42**	**0.84**	**19.29**	**0.85**	**5**	1st Der (W:3 O:2) + Smooth (W:13) + NAS (3)
		AutoOPS	1160	14.12	0.93	24.44	0.79	19.93	0.82	5	
		FeedOPS	865	15.74	0.92	22.04	0.83	21.70	0.76	5	
	Glucose	**Full**	**3112**	**3.87**	**0.91**	**6.43**	**0.74**	**6.99**	**0.64**	**5**	
		AutoiOPS	675	1.92	0.98	5.47	0.81	5.19	0.82	5	
		FeediOPS	**265**	**2.31**	**0.97**	**5.36**	**0.82**	**5.29**	**0.82**	**5**	NAS (3) + 2nd Der (W:17 O:2) + Mean Center
		AutoOPS	860	2.00	0.98	6.07	0.77	5.60	0.79	5	
		FeedOPS	515	2.43	0.97	5.53	0.81	5.36	0.82	5	
	Fructose	**Full**	**3112**	**4.67**	**0.95**	**7.45**	**0.86**	**8.13**	**0.80**	**7**	
		AutoiOPS	740	3.05	0.98	6.74	0.88	7.14	0.85	7	
		FeediOPS	550	3.41	0.97	7.45	0.86	6.42	0.88	7	1st Der (W:9 O:2) + Norm (2) + NAS (2)
		AutoOPS	**2350**	**3.46**	**0.97**	**6.81**	**0.88**	**6.65**	**0.86**	**7**	
		FeedOPS	1805	3.31	0.97	6.94	0.88	6.92	0.85	7	

Rc: correlation coefficient of calibration; Rcv: correlation coefficient of cross-validation; Rp: correlation coefficient of prediction; RMSEC: root mean square error of calibration; root mean square error of cross-validation; RMSEP: root mean square error of prediction; LV: Latent variable.

In Table 4, the values highlighted in red, using the FeediOPS and AutoOPS methods for variable selection, led to a significant decrease in mean squared errors for all built models. The selection of a few latent variables in the PLS model contributed to this substantial reduction in root mean square error, indicating that the model efficiently captures the underlying data structure with minimal complexity, simplifying model interpretation.

The red highlights display the statistical parameters in model building. For the model using the unripe banana peel, the investigated property of sucrose exhibited RMSECV and RMSEP values of 13.41 and 15.18 mg/g, respectively, with cross-validation and prediction coefficients of 0.95 and 0.93. When using ripe peel, the values for sucrose were 21.42, 19.29 mg/mL, 0.84, and 0.85, respectively. For the unripe peel model, 4 latent variables and 5 latent variables about sucrose were used for the ripe peel model.

In the case of glucose in unripe banana peel, the model showed RMSECV and RMSEP values of 4.17 and 3.41 mg/g, respectively, with cross-validation and prediction coefficients of 0.93 and 0.97. The values for glucose in ripe peel were 5.36, 5.29 mg/mL, 0.82, and 0.82, respectively, utilizing 7 and 5 latent variables for unripe and ripe peel, respectively.

For fructose in unripe banana peel, the model exhibited RMSECV and RMSEP values of 5.75 and 8.70 mg/mL, respectively, with cross-validation and prediction coefficients of 0.88 and 0.86. The values for fructose in ripe peel were 6.81, 6.65 mg/mL, 0.88, and 0.86, respectively, utilizing 4 and 7 latent variables for unripe and ripe peel, respectively.

When analyzing the complex data generated from spectra obtained from the banana peel, we observed a positive trend in the model results. Notably, there was a significant reduction in mean squared error values compared to the minimum values of the dependent variables. Additionally, the cross-validation and prediction correlation coefficients suggest a strong agreement between the observed data and the values predicted by the models.

The models built for unripe banana peel exhibit excellent results concerning a significant reduction in mean squared errors and determination and prediction coefficients close to one. The models can be used for prediction with a certain level of accuracy. Notably, the models' results for ripe peel show potential for utilization; the coefficients and errors demonstrate the ability to discriminate quantities of sucrose, glucose, and fructose. For all models, there was no need to remove samples representing outliers, indicating the quality of the model constructed with few samples.

In Figure 16, the original spectra are presented along with the selected variables that would contribute the most to improving the models.

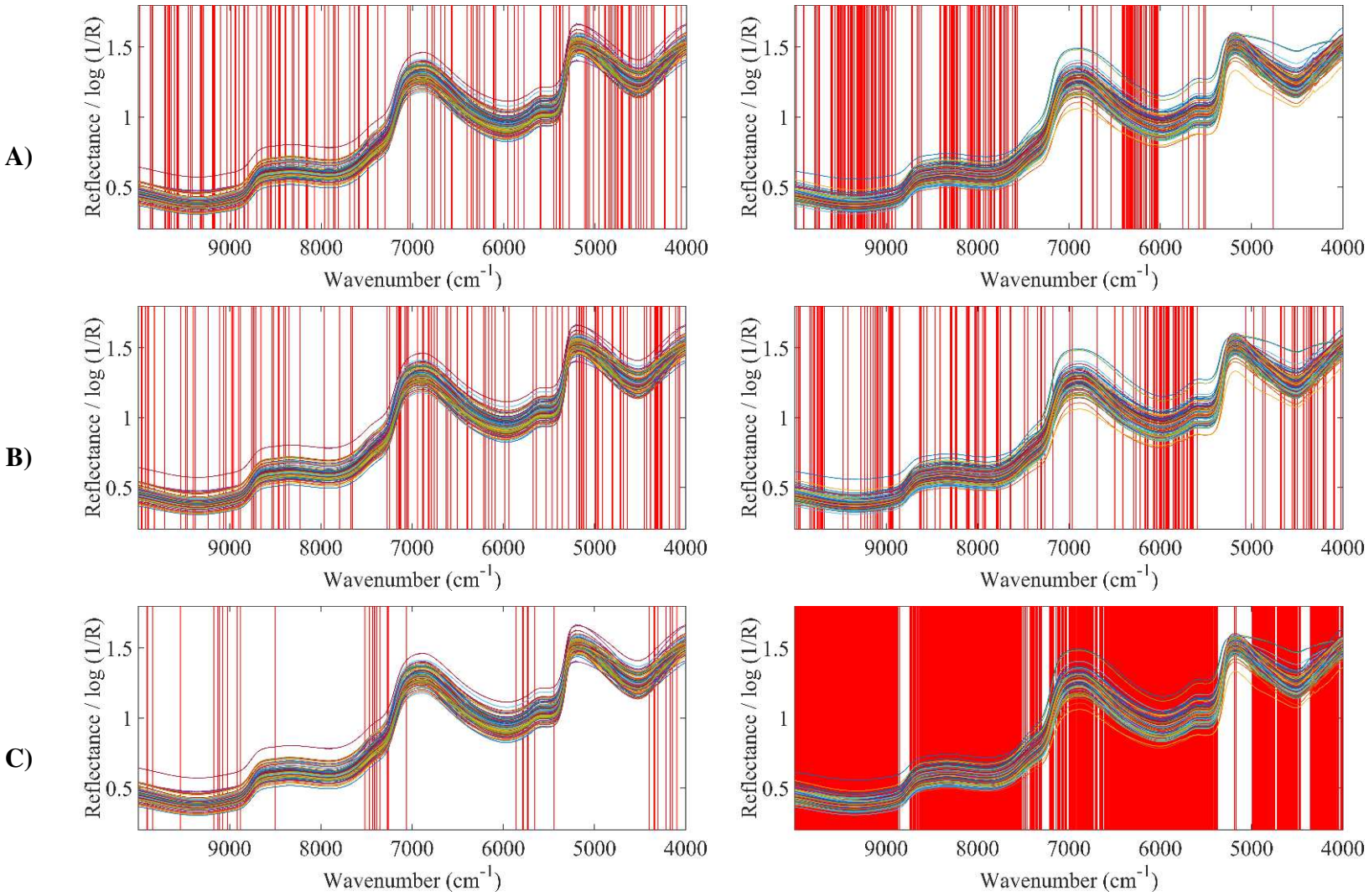

Figure 16. Variables were selected using FeediOPS and AutoiOPS in unripe banana on the left and ripe banana on the right for sucrose (A), glucose (B), and fructose (C).

Figure 16 presents the selected variables for sucrose, fructose, and glucose using spectra from both unripe and ripe banana peels. After analyzing all the spectra, it is evident that the selection of variables is closely connected to the absorption bands of CH, OH, and CO, which are linked to the molecular composition of sugars.

The variable selection process proved effective, narrowing down from a total of 3112 variables (as shown in Table 4) to 155, 145, and 50 for unripe banana data regarding sucrose, glucose, and fructose, respectively, and 250, 265, and 2350 for ripe banana data regarding sucrose, glucose, and fructose, respectively. These selected variables contribute to building a more predictive model by selecting the most important variables.

In Figure 17, the Predicted vs. Measured graphs are presented, a crucial parameter in assessing the quality and accuracy of the built models.

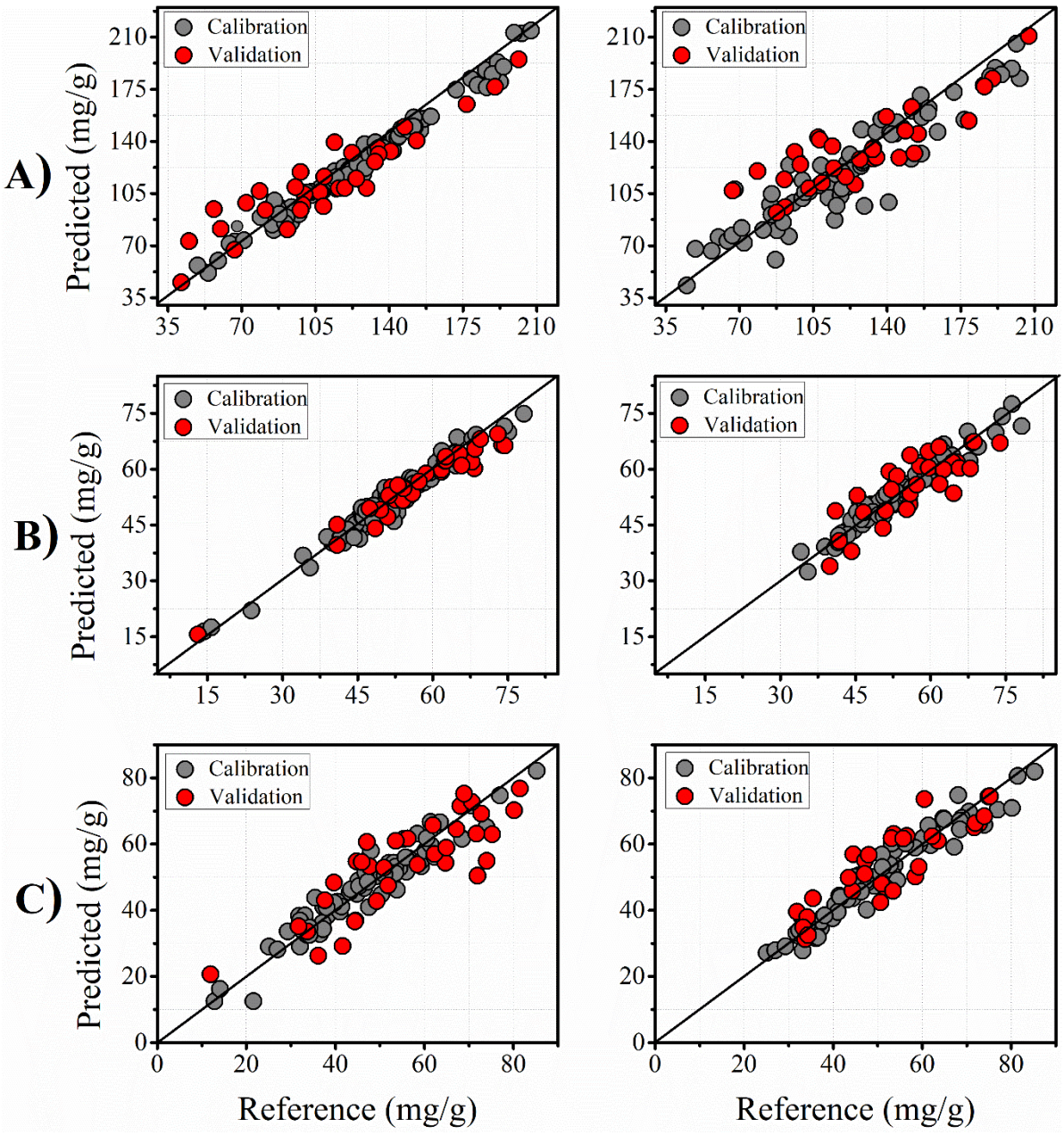

Figure 17. Reference and predicted sugars content in unripe banana on the left and ripe banana on the right for sucrose (A), glucose (B), and fructose (C).

The graphs presented in Figure 16 represent the dataset for each sucrose, glucose, and fructose prediction model. On the x-axis (abscissa), we have the measured or reference values, while on the y-axis (ordinate), we have the values predicted by the model. Ideally, the models should align on a diagonal 45-degree line, indicating an exact prediction by the model.

In Figure 17, it is noticeable that for all models built, the dispersion of the calibration dataset approaches the 45-degree diagonal line with minimal variation. It can be observed that there is an excellent linear fit for all models, and a linear trend is observed, indicating good accuracy among the results.

Figure 17 presents the relative errors of the models concerning the calibration and validation datasets used for model building.

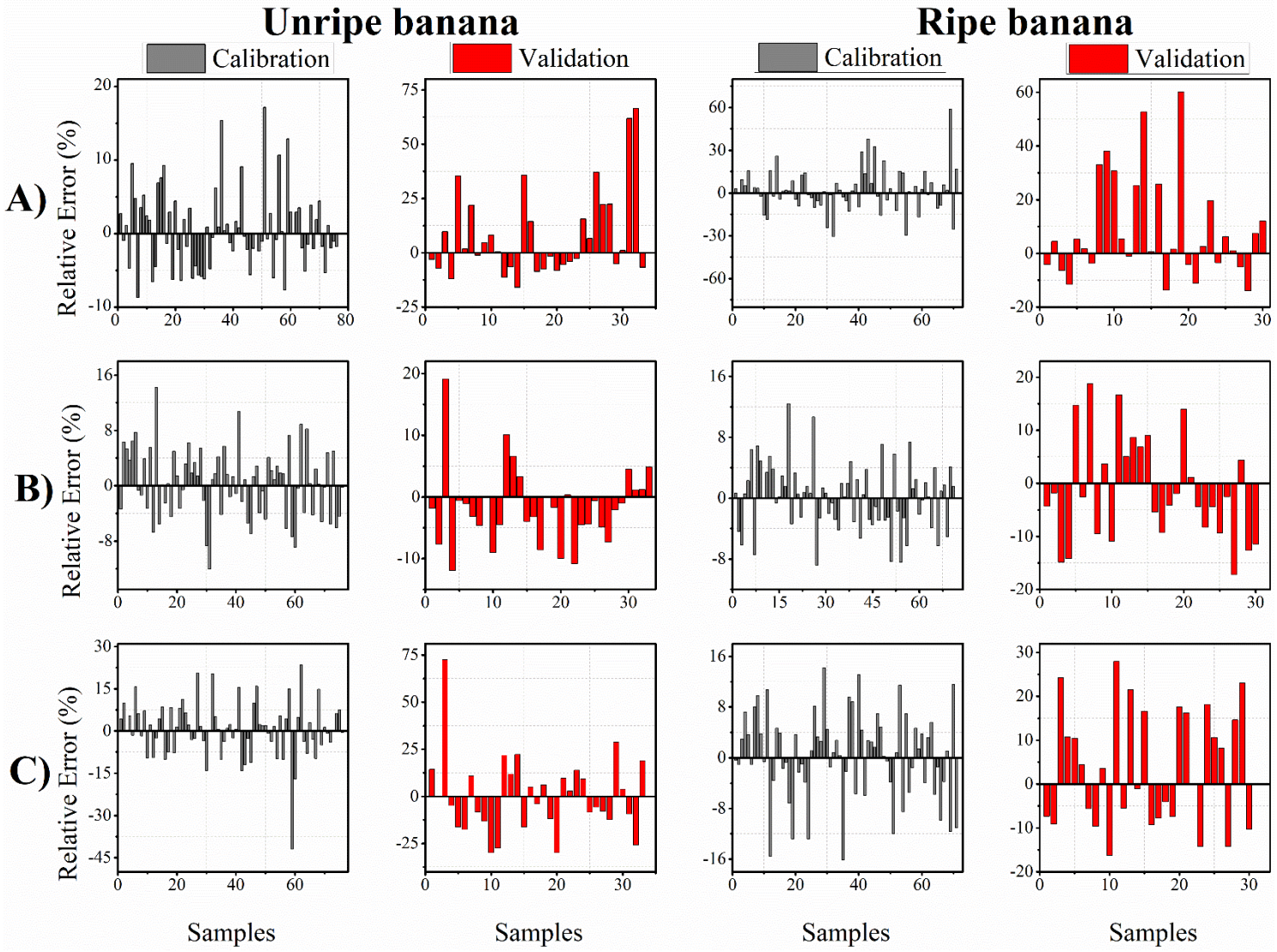

Figure 18. Relative error for calibration and validation set in unripe banana on the left and ripe banana on the right. For Sucrose (A), Glucose (B) and Fructose (C).

Figure 18 shows the relative errors of both the calibration and validation datasets. Most of the relative errors were below 10%, indicating that the models are suitable for prediction.

The calibration data showed mostly minor errors, while larger errors were observed among the validation datasets. The discrepant errors could be due to systematic errors in some samples, such as sample preparation or spectrum acquisition. To enhance the models, adding more samples could provide a more comprehensive representation of the data distribution and potentially detect patterns of errors associated with model predictions.

This work confirms previous results [5,6] that early predictions are possible and reliable with adequate design in obtaining the calibration set.

Figure 21 displays the Pearson correlation charts for the constructed models. The red spheres represent the authentic y vector, while the black spheres represent the randomized y vectors.

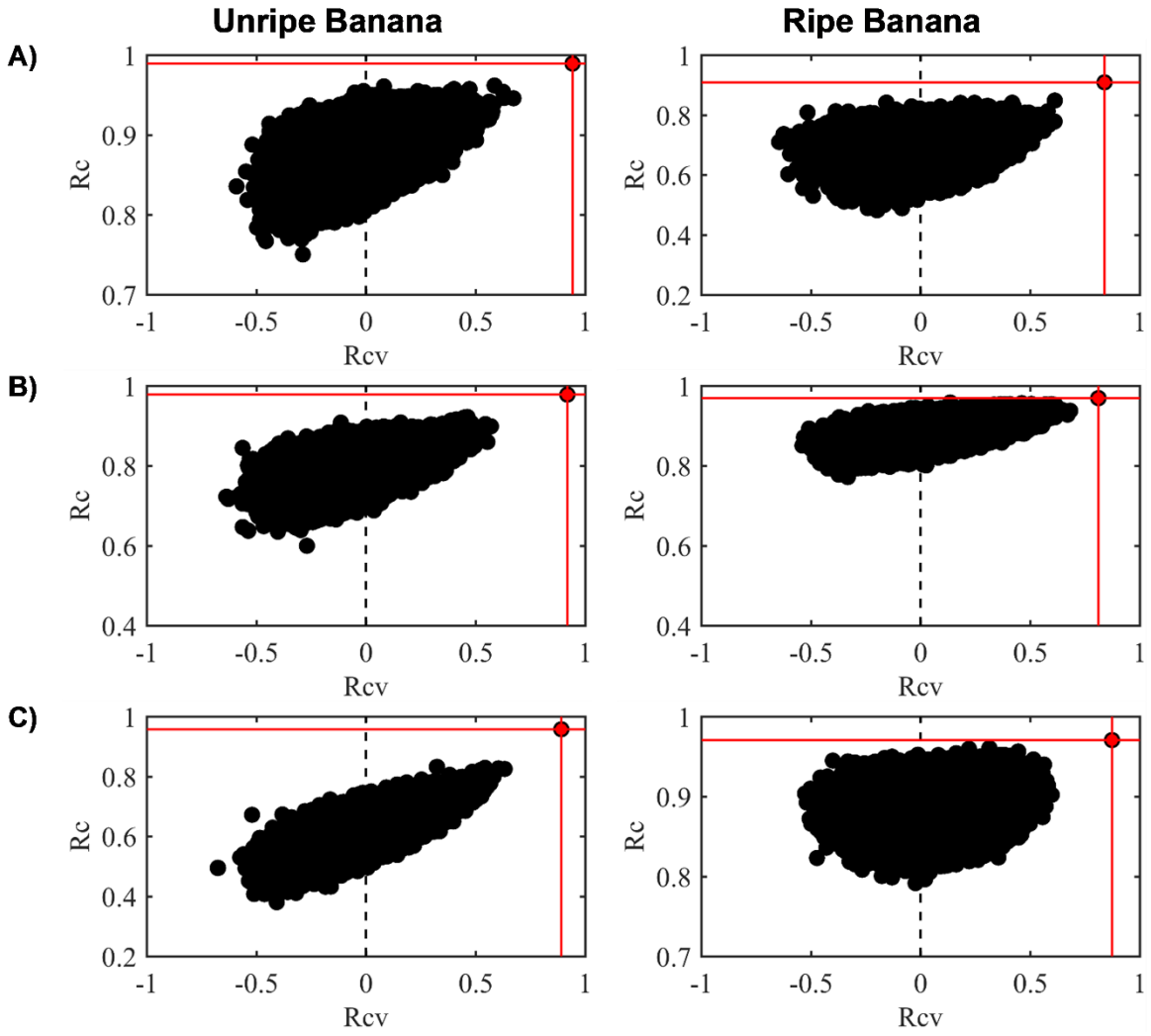

Figure 19. The chance correlation plot. Determination coefficient of cross-validation (Rcv) versus determination coefficient of calibration of calibration (Rc – person correlation) for sucrose (A), glucose (B), and fructose (C). The black spheres show the randomized models and the red sphere is the original model for both unripe (left side) and ripe (right side) banana models.

The developed models were evaluated to investigate the possibility of correlation by chance. Figure 19 shows that the models generated and represented by the red spheres are distinguishable from those produced by randomizing the y vector. It is concluded that the models were not obtained by chance.

The implementation of accurate early prediction methods, such as the models developed in this study, not only benefits farmers involved in plant selection and plant breeding but also promotes efficiency and profitability throughout the banana supply chain by providing consumers with the best fruit based on its sweetness. Additionally, using NIR as a technique

that requires no sample preparation, it stands out as an environmentally friendly, cost-effective, and non-destructive approach for selecting the best fruits, contributing to the efficiency and sustainability of the entire banana cultivation chain.

5. CONCLUSION

Predicting specific properties of plants that enhance their market value is highly beneficial for the entire agribusiness chain, especially when determining the ideal time for harvest remains subjective. Through studying the requirements of bananas within the production chain, successful prediction of sugar content in banana pulp was achieved using models constructed from its unripe (green) peel. This method has proven effective in identifying the optimal harvest stage, genetic enhancements, and environmental factors before fruit distribution. Near-infrared spectroscopy is emerging as a portable, non-invasive, and widely accepted tool in this domain. Early predictions of sucrose, glucose, and fructose concentrations in ripe banana pulp derived from NIR spectra on unripe banana peel yielded remarkable results. Furthermore, it can be inferred that sugar predictions in ripe banana pulp from NIR spectra on unripe banana peel outperformed those from ripe peels, underscoring the efficacy of this methodology in optimizing harvest timing and enhancing fruit quality before market entry.

6. REFERENCES

1. Gomes, J. F. S., Vieira, R. R. & Leta, F. R. Colorimetric indicator for classification of bananas during ripening. *Scientia Horticulturae* **150**, 201–205 (2013).
2. EMBRAPA. *Banana Fitossanidade* 1–121 (2000).
3. FAO. https://www.fao.org/faostat/en/#data/QCL. *Crops and livestook products* (2024).
4. Ma, L. *et al.* Prediction of banana maturity based on the sweetness and color values of different segments during ripening. *Current Research in Food Science* **5**, 1808–1817 (2022).
5. Du, L. *et al.* Proteome changes in banana fruit peel tissue in response to ethylene and high-temperature treatments. *Hortic Res* **3**, 16012 (2016).
6. Oliveira, U. F. *et al.* Predicting oil content in ripe Macaw fruits (Acrocomia aculeata) from unripe ones by near infrared spectroscopy and PLS regression. *Food Chemistry* **351**, 129314 (2021).
7. Porto, N. D. A. *et al.* Early prediction of sugarcane genotypes susceptible and resistant to Diatraea saccharalis using spectroscopies and classification techniques. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **218**, 69–75 (2019).
8. Loesecke, V. & W, H. *Bananas.* (Interscience Publishers, New York, 1949).
9. EMBRAPA. *Brasil em 50 alimentos*. vol. 1 (Brasília, DF, 2023).
10. FAO. Markets and Trade Bananas. https://www.fao.org/markets-and-trade/commodities/bananas/en/ (2023).
11. Barnell, H. R. Studies in Tropical Fruits: XI. Carbohydrate Metabolism of the Banana Fruit during Ripening under Tropical Conditions. **Annals of Botany, New Series, Vol. 5, No. 18 , pp. 217–247**, (1941).
12. MOTA, R. V. da, LAJOLO, F. M. & CORDENUNSI, B. R. Composição em carboidratos de alguns cultivares de banana (Musa spp.) durante o amaduremcimento. vol. 17 94–97 (1997).

13. Botrel, N., Freire Junior, M., Vasconcelos, R. M. D. & Barbosa, H. T. G. Inibição do amadurecimento da banana-'Prata-Anã' com a aplicação do 1-metilciclopropeno. *Rev. Bras. Frutic.* **24**, 53–56 (2002).

14. Skoog, D. A., West, D. M. & Crouch. *Fundamentos de química analítica*. (Editora Thomson, São Paulo - SP, Brasil, 2006).

15. Collins, C. H., Braga, G. L. & Bonato, P. S. *Fundamentos da cromatografia*. (Editora Unicamp, Campinas, 2006).

16. 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 Chemistry* **145**, 784–788 (2014).

17. Guo, S. *et al.* Characterization of Triterpenic Acids in Fruits of Ziziphus Species by HPLC-ELSD-MS. *J. Agric. Food Chem.* **58**, 6285–6289 (2010).

18. Guo, S. *et al.* Content variations of triterpenic acid, nucleoside, nucleobase, and sugar in jujube (Ziziphus jujuba) fruit during ripening. *Food Chemistry* **167**, 468–474 (2015).

19. Wasik, A., McCourt, J. & Buchgraber, M. Simultaneous determination of nine intense sweeteners in foodstuffs by high performance liquid chromatography and evaporative light scattering detection—Development and single-laboratory validation. *Journal of Chromatography A* **1157**, 187–196 (2007).

20. Shanmugavelan, P. *et al.* Evaluation of sugar content and composition in commonly consumed Korean vegetables, fruits, cereals, seed plants, and leaves by HPLC-ELSD. *Carbohydrate Research* **380**, 112–117 (2013).

21. Swartz, M. HPLC DETECTORS: A BRIEF REVIEW. *Journal of Liquid Chromatography & Related Technologies* **33**, 1130–1150 (2010).

22. Megoulas, N. C. & Koupparis, M. A. Twenty Years of Evaporative Light Scattering Detection. *Critical Reviews in Analytical Chemistry* **35**, 301–316 (2005).

23. Hershel, F. Experiments on the Refrangibility of the Invisible Rays of the Sun, Philosophical Transactions of the Royal Academy. 284–292 (1800).

24. Ellis, J. W. & Bath, J. Alterations in the Infrared Absorption Spectrum of Water in Gelatin. *The Journal of Chemical Physics* **6**, 108–108 (1938).

25. Pasquini, C. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. *Analytica Chimica Acta* **1026**, 8–36 (2018).

26. Li, J., Sun, D. & Cheng, J. Recent Advances in Nondestructive Analytical Techniques for Determining the Total Soluble Solids in Fruits: A Review. *Comp Rev Food Sci Food Safe* **15**, 897–911 (2016).

27. Xiaobo, Z., Jiewen, Z., Povey, M. J. W., Holmes, M. & Hanpin, M. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta* **667**, 14–32 (2010).

28. Blanco, M. & Villarroya, I. NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry* **21**, 240–250 (2002).

29. Miller, C. E. *Near-Infrared Technology in the Agricultural and Food Industries, American Society of Cereal Chemists*. (St. Paul, Minessota, 2001).

30. Norris, K. H. *Infrared Spectroscopy*. vol. 4 (1996).

31. Cardoso, W. J. Quantification of Carbohydrates and Fiber from Dehydrate Sugarcane juice using Near-Infrares Spectroscopy and Chemometrics. (Universidade Federal de Viçosa, Viçosa, MG, 2019).

32. Cozzolino, D. & Morón, A. Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions. *Soil and Tillage Research* **85**, 78–85 (2006).

33. Fernández Pierna, J. A. *et al.* NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed. *Chemometrics and Intelligent Laboratory Systems* **117**, 233–239 (2012).

34. Biancolillo, A., Marini, F., Ruckebusch, C. & Vitale, R. Chemometric Strategies for Spectroscopy-Based Food Authentication. *Applied Sciences* **10**, 6544 (2020).

35. Lin, P., Chen, Y. & He, Y. Identification of Geographical Origin of Olive Oil Using Visible and Near-Infrared Spectroscopy Technique Combined with Chemometrics. *Food Bioprocess Technol* **5**, 235–242 (2012).

36. Casale, M. & Simonetti, R. Review: Near Infrared Spectroscopy for Analysing Olive Oils. *Journal of Near Infrared Spectroscopy* **22**, 59–80 (2014).

37. Ferreira, M. M. de C. *QUIMIOMETRIA - Conceitos, Métodos e Aplicações*. (Editora Unicamp, Campinas, 2015).

38. FERREIRA, Márcia et al. Quimiometria I: Calibração Multivariada, Um Tutorial. vol. 22 724–731 (1999).

39. Rencher, A. C. *Methods of Multivariate Analysis*. (Wiley, New York, 2002).

40. Teófilo, R. F. Chemometric methods in the electrochemical studies of phenols on boron-doped diamond films. (Universidade Estadual de Campinas, Campinas, 2007).

41. Prasada, R. & Talasila & Biju, V. M. *SPECTROPHOTOMETRY | Organic Compounds. 10.1016/B0-12-369397-7/00721-4.* (2005).

42. Brereton R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant.* (John Wiley & Sons Inc, Chinchester, 2003).

43. Varmuza, K. & Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. (CRC Press, Boca Raton, 2009).

44. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**, 109–130 (2001).

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

46. Assis, C., Oliveira, L. S. & Sena, M. M. Variable Selection Applied to the Development of a Robust Method for the Quantification of Coffee Blends Using Mid Infrared Spectroscopy. *Food Anal. Methods* **11**, 578–588 (2018).

47. Roque, J. V., Cardoso, W., Peternelli, L. A. & Teofilo, R. F. Comprehensive new approaches for variable selection using ordered predictors selection. *Analytica Chimica Acta* 57–70 (2019).

48. Teófilo, R. F., Martins, J. P. A. & Ferreira, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *Journal of Chemometrics* **23**, 32–48 (2009).

49. Ribeiro, L. R., Oliveira, L. M. D., Silva, S. D. O. E. & Borges, A. L. Caracterização física e química de bananas produzidas em sistemas de cultivo convencional e orgânico. *Rev. Bras. Frutic.* **34**, 774–782 (2012).

50. Silva, A. P. F. B. Perfil de carboidratos de diferentes cultivares de manga durante o desenvolvimento e amadurecimento. (Universidade de São Paulo, São Paulo, 2000). doi:10.11606/D.9.2000.tde-11022015-081839.

51. Hattenhauer, S. K. & Carvalho, R. I. N. D. Caracterização física e química da banana 'Nanica' em função da época de colheita e do diâmetro do fruto em Corupá, SC. *RAC* **29**, 80–83 (2016).

52. *Guia de validação e controle de qualidade analítica: fármacos em produtos para alimentação e medicamento veterinários*. (Mapa, 2011).

53. Cardoso, W. J., Gomes, J. G. R., Roque, J. V., Barbosa, M. H. P. & Teófilo, R. F. Dehydration as a Tool to improve predictability of sugarcane juice carbohydrates using near-infrared spectroscopy based PLS models. *Chemometrics and Intelligent Laboratory Systems* **220**, 104459 (2022).

54. Beganović, A., Moll, V. & Huck, C. W. Comparison of Multivariate Regression Models Based on Water- and Carbohydrate-Related Spectral Regions in the Near-Infrared for Aqueous Solutions of Glucose. *Molecules* **24**, 3696 (2019).

55. Workman, Jr., Jerry & Weyer, L. *Practical Guide to Interpretive Near-Infrared Spectroscopy*. (CRC Press, 2007). doi:10.1201/9781420018318.