WILSON JUNIOR CARDOSO

IMPROVING THE ACCURACY OF MULTIVARIATE MODELS: A STUDY OF SAMPLE DEHYDRATION AND DATA PREPROCESSING OPTIMIZATION

Thesis submitted to the Agrochemistry Graduate Program of the Universidade Federal de Viçosa in partial fulfilment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Reinaldo Francisco Teófilo

Co-Advisor: Jussara Valente Roque

VIÇOSA - MINAS GERAIS
2023

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

T
C268i
2023

Cardoso, Wilson Júnior, 1992-
Improving the accuracy of multivariate models: a study of sample dehydration and data preprocessing optimization / Wilson Júnior Cardoso. – Viçosa, MG, 2023.
1 tese eletrônica (93 f.): il. (algumas color.).

Texto em inglês.
Orientador: Reinaldo Francisco Teófilo.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Química, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.297
Modo de acesso: World Wide Web.

1. Espectroscopia de infravermelho próximo. 2. Quimiometria - Processamento de dados. 3. Análise multivariada. 4. Desidratação. 5. Cana-de-açúcar. I. Teófilo, Reinaldo Francisco, 1978-. II. Universidade Federal de Viçosa. Departamento de Química. Programa de Pós-Graduação em Agroquímica. III. Título.

CDD 22. ed. 543.5

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

WILSON JUNIOR CARDOSO

IMPROVING THE ACCURACY OF MULTIVARIATE MODELS: A STUDY OF SAMPLE DEHYDRATION AND DATA PREPROCESSING OPTIMIZATION

Thesis submitted to the Agrochemistry Graduate Program of the Universidade Federal de Viçosa in partial fulfilment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: March 9th, 2023

Assent:

Wilson Júnior Cardoso
Author

Documento assinado digitalmente
gov.br WILSON JUNIOR CARDOSO
Data: 13/08/2023 19:13:29-0300
Verifique em https://validar.iti.gov.br

Reinaldo Francisco Teófilo
Adviser

Documento assinado digitalmente
gov.br REINALDO FRANCISCO TEOFILO
Data: 13/08/2023 16:21:19-0300
Verifique em https://validar.iti.gov.br

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.

ACKNOWLEDGEMENTS

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

I am particularly grateful to my parents, Elenice and Wilson, and sister, Luana, for all the support and incentive.

I would like to thank Marco for his support and encouragement.

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

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

I would like to thank the Sugarcane Breeding Program of the 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, particularly Helder and Igor for all the discussions and fun moments.

I would like to thank Jussara for all the assistance and discussions.

I would like to thank Cássio and Nathália for their friendship.

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

I am very thankful to Professor Jansen for having me at the Institute for Molecules and Materials at Radboud University, Nijmegen, The Netherlands, for the internship period.

I would like to thank Dr. Teng for all the help and discussion during my stay at Radboud University.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

To the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), for granting the scholarship.

To the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), to 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.

“Don’t aim for success if you want it; just do what you love and believe in, and it will come naturally.”

(David Frost)

ABSTRACT

CARDOSO, Wilson Júnior, D.Sc., Universidade Federal de Viçosa, March, 2023. **Improving the accuracy of multivariate models: a study of sample dehydration and data preprocessing optimization.** Adviser: Reinaldo Francisco Teófilo. Co-Adviser: Jussara Valente Roque.

The aim of this thesis is to study approaches to improve the accuracy of multivariate models. Two approaches were considered, one relating to sample preparation and the other related to data preprocessing. The first chapter aimed to study sample dehydration to improve the prediction of sucrose, glucose, and fructose in sugarcane juice using near-infrared (NIR) spectroscopy and partial least squares (PLS) regression models. Models using the NIR spectra acquired using the liquid (LSJ) and dehydrated sugarcane juice (DSJ) were compared. In addition, the NIR spectra were acquired using a benchtop and a portable instrument. Ordered predictors selection (OPS) was applied to select the most informative variable. The results indicated better predictions for all sugars using the DSJ for both instruments, being the benchtop statistically better than the portable instrument. To sum up, the dehydration approach showed to be a great technique to improve the predictability of PLS-OPS models for sugars in sugarcane juice using NIR spectra by removing the water and concentrating the analytes. The second chapter presented an algorithm that automatically searches for the best preprocessing strategy without fixing their order based on the artifact they fix, i.e., baseline correction, scatter correction, noise removal, and scaling. The number of preprocessing methods in each strategy and their hyperparameters were evaluated. The algorithm was compared with methods presented in the literature by Gerretzen et al. (2015) and Jiao et al. (2020). A fair, extensive, and comprehensive study was carried out, evaluating 67 different calibration datasets. This work demonstrated that not fixing the order in which the preprocessing is applied was essential to find the best models with a significant reduction in the RMSEP values when compared with the other methods, therefore presenting a comprehensive insight into data preprocessing. These results showed that a proper sample preparation and a proper optimization of the data preprocessing strategy are fundamental to build the best models.

Keywords: Chemometrics. Sample Preparation. Water Removal. Data Preprocessing.

RESUMO

CARDOSO, Wilson Júnior, D.Sc., Universidade Federal de Viçosa, março de 2023. **Melhorando a eficiência de modelos multivariados usando a desidratação de amostra e otimizando o pré-processamento de dados.** Orientador: Reinaldo Francisco Teófilo. Coorientadora: Jussara Valente Roque.

O objetivo desta tese é estudar diferentes metodologias para melhorar a acurácia de modelos multivariados. Duas abordagens foram consideradas, uma relativa ao preparo de amostra e outra relacionada ao pré-processamento dos dados. O objetivo do primeiro capítulo foi estudar a desidratação como forma de melhorar a predição da concentração de sacarose, glicose e frutose no caldo de cana-de-açúcar usando espectroscopia de infravermelho próximo (NIR) e regressão por quadrados mínimos parciais (PLS). Os modelos utilizando os espectros NIR adquiridos a partir do caldo líquido (LSJ) e desidratado (DSJ) foram comparados. Além disso, os espectros NIR foram adquiridos usando um instrumento de bancada e um instrumento portátil. A seleção de preditores ordenados (OPS) foi aplicada para selecionar as variáveis mais informativas. Os resultados indicaram melhores predições para todos os açúcares utilizando o DSJ para ambos os instrumentos, sendo o de bancada estatisticamente melhor que o instrumento portátil. Em suma, a desidratação da amostra mostrou ser uma ótima técnica para melhorar a acurácia dos modelos, removendo a água e concentrando os analitos. O objetivo do segundo capítulo foi apresentar um algoritmo que busca a melhor estratégia de pré-processamento sem fixar sua ordem com base no artefato que eles corrigem, ou seja, correção de linha de base, correção de dispersão, remoção de ruído e dimensionamento. O número de métodos de pré-processamento em cada estratégia e seus hiper-parâmetros foram avaliados. O algoritmo foi comparado com métodos apresentados na literatura por Gerretzen et al. (2015) e Jiao et al. (2020). Um estudo imparcial, extenso e abrangente foi realizado neste trabalho, avaliando 67 conjuntos de dados de calibração diferentes. Este trabalho demonstrou que não fixar a ordem de aplicação do pré-processamento foi essencial para encontrar os melhores modelos com redução significativa nos valores de RMSEP quando comparados com os outros métodos, apresentando, portanto, uma visão abrangente sobre o pré-processamento de dados. Esses resultados mostraram que uma preparação adequada da amostra e uma otimização adequada da estratégia de pré-processamento de dados são fundamentais para construir os melhores modelos.

Palavras-chave: Quimiometria. Preparo de Amostra. Remoção de Água. Pré-processamento de dados.

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
AS	Autoscaling
AsLs	Asymmetric Least Squares
BO	Baseline Offset
Brix	Soluble Solids Content
CCD	Central Composite Design
D	Detrending
DSJ	Dehydrated Sugarcane Juice
ELSD	Evaporative Light Scattering Detector
EVA	Ethylene-Vinyl Acetate
FIN	Thin Couche Paper
GC	Gas Chromatography
GRO	Thick Couche Paper
GSC	Glass Slips Cover
HPLC	High Performance Liquid Chromatography
LB	Linear Baseline
LOD	Limit of Detection
LS	Level Scaling
LSJ	Liquid Sugarcane Juice
LV	Latent Variable
MC	Mean Centering
MIR	Mid-Infrared
MSC	Multiplicative Scatter Correction
nas	Absolute Value of the Net Analytical Signal
NAS	Net Analytical Signal
NIR	Near-Infrared
Norm	Normalization
OPS	Ordered Predictors Selection
PAR	Kraft Paper
PD	Pairwise Detrending
PEI	Prediction Error Index
PLS	Partial Least Squares
PLS-OPS	Partial Least Squares - Ordered Predictors Selection
Pol	Polarimetric Value
POS	Poisson Scaling
PPO	Preprocessing Optimizer
PS	Pareto Scaling
R^2	Coefficient of Determination
Rc	Correlation Coefficient of Calibration
Rcv	Correlation Coefficient of Cross-Validation
RMSEC	Root Mean Squared Error of Calibration
RMSECV	Root Mean Squared Error of Cross-Validation
RMSEP	Root Mean Squared Error of Prediction
RNV	Robust Normal Variate

Rp	Correlation Coefficient of Prediction
RS	Range Scaling
SEL	Selectivity
SG	Smoothing
SG1D	First Derivative
SG2D	Second Derivative
SIL	Silicone
SNV	Standard Normal Variate

LIST OF ILLUSTRATIONS

Chapter I

Chapter II

LIST OF TABLES

Chapter I

Chapter II

SUMMARY

OVERVIEW

This document is divided into two chapters, and it is organized in the following manner:

Chapter I reports an algorithm that automatically optimizes the data preprocessing strategy and their hyperparameters without fixing their order based on the artifact they fix, i.e., baseline correction, scatter correction, noise removal, and scaling. The influence of the number of preprocessing methods in the preprocessing strategy was evaluated. The algorithm was compared with the ones presented in the literature.

Chapter II reports a dehydration method to improve the accuracy of the NIR-based models to predict the contents of sucrose, glucose, and fructose in the sugarcane juice. The models using the dehydration method were compared with the models using the liquid juice. All models were built using a benchtop and portable instrument. In addition, the algorithm developed in Chapter I was used to optimize the preprocessing strategy of the models.

The two chapters, although independent, are complementary as they cover methods to improve the accuracy of the models.

GENERAL INTRODUCTION

High-dimensional data has become widely used in many fields as it can replace time-consuming analysis when combined with chemometrics methods [1–3]. However, the quality of the data is crucial to obtain accurate results. Data can be affected by instrument or analytical artifacts, so data preprocessing plays a significant role in removing these artifacts and ensuring that the resulting model accurately represents the real system [4]. Therefore, preprocessing methods are mainly used to remove artifacts, such as noise, baseline shift, slope, non-informative variables, and light scattering, generating a more interpretive and robust model [5–9].

Near-infrared spectroscopy (NIR) is a versatile analytical technology that can be used as an alternative to expensive, time-consuming routine analysis. It is fast, inexpensive, and requires little or no sample preparation [10–14]. However, the predictive accuracy of NIR-based models can be directly affected by the sample preparation.

Although NIR does not require an extensive sample preparation, studies have shown that proper sample preparation should be considered to ensure meaningful and reliable analytical results [15–22]. In aqueous samples, water highly absorbs NIR radiation, making it difficult to access the overlapping information, thus making the models less predictive [23–26]. Usually, these inferences are removed by selecting wavelength ranges or by preprocessing the data, but these methods are not always effective [27–30].

However, selecting an optimal preprocessing method is not an easy task and is commonly referred to as one of the main bottlenecks in data modeling [7,31–34]. Preprocessing selection is generally chosen based on trial and error, visual inspection, and past experiences [35,36]. There are thousands of different preprocessing possibilities to try, and it is difficult to determine the optimum preprocessing by just looking at the data. Each dataset has its peculiarities causing the most suitable preprocessing strategy to differ. Therefore, the tedious procedure of manual preprocessing must be repeated for different datasets.

Several preprocessing optimization strategies have been introduced in the literature, including the genetic algorithm, design of experiments, and algorithm evaluation [5,37–40]. Some of the optimization methods presented in the literature require significant computational resources and may be specific to certain types of data. Also, the methods that use experimental designs focus on a search space limited by the number of levels of the design. Furthermore, these works concluded that the effect of preprocessing is highly data-dependent, and therefore the preprocessing strategy should be optimized for each dataset individually.

Therefore, the general goals of this work are: (i) to develop a simple algorithm to optimize data preprocessing strategies, (ii) to study the influence of water removal and data preprocessing on NIR-based models to predict sugarcane carbohydrates aiming at improving the prediction power of the multivariate models.

REFERENCES

(1) Cortés, V.; Blasco, J.; Aleixos, N.; Cubero, S.; Talens, P. Monitoring Strategies for Quality Control of Agricultural Products Using Visible and Near-Infrared Spectroscopy: A Review. *Trends Food Sci Technol* **2019**, *85* (October 2018), 138–148. https://doi.org/10.1016/j.tifs.2019.01.015.

(2) Horta, A.; Malone, B.; Stockmann, U.; Minasny, B.; Bishop, T. F. A.; McBratney, A. B.; Pallasser, R.; Pozza, L. Potential of Integrated Field Spectroscopy and Spatial Analysis for Enhanced Assessment of Soil Contamination: A Prospective Review. *Geoderma* **2015**, *241–242*, 180–209. https://doi.org/10.1016/j.geoderma.2014.11.024.

(3) Wang, H.; Peng, J.; Xie, C.; Bao, Y.; He, Y. Fruit Quality Evaluation Using Spectroscopy Technology: A Review. *Sensors* **2015**, *15* (5), 11889–11927. https://doi.org/10.3390/s150511889.

(4) Rinnan, Å.; Berg, F. van den; Engelsen, S. B. Review of the Most Common Pre-Processing Techniques for near-Infrared Spectra. *TrAC - Trends in Analytical Chemistry* **2009**, *28* (10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

(5) Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Bart, J.; van Manen, H.-J.; van den Heuvel, E. R.; Buydens, L. M. C. Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments. *Anal Chem* **2015**, *87* (24), 12096–12103. https://doi.org/10.1021/acs.analchem.5b02832.

(6) Xu, L.; Zhou, Y.; Tang, L.; Wu, H.; Jiang, J.; Shen, G.; Yu, R. Ensemble Preprocessing of Near-Infrared (NIR) Spectra for Multivariate Calibration. **2008**, *6*, 138–143. https://doi.org/10.1016/j.aca.2008.04.031.

(7) Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; de la Pezuela, C. Effect of Data Preprocessing Methods in Near-Infrared Diffuse Reflectance Spectroscopy for the Determination of the Active Compound in a Pharmaceutical Preparation. *Appl Spectrosc* **1997**, *51* (2), 240–246. https://doi.org/10.1366/0003702971939947.

(8) Verboven, S.; Hubert, M.; Goos, P. Robust Preprocessing and Model Selection for Spectral Data. *J Chemom* **2012**, *26* (6), 282–289. https://doi.org/10.1002/cem.2446.

(9) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. C. Breaking with Trends in Pre-Processing? *TrAC - Trends in Analytical Chemistry* **2013**, *50*, 96–106. https://doi.org/10.1016/j.trac.2013.04.015.

(10) Grassi, S.; Alamprese, C. Advances in NIR Spectroscopy Applied to Process Analytical Technology in Food Industries. *Curr Opin Food Sci* **2018**, *22*, 17–21. https://doi.org/10.1016/j.cofs.2017.12.008.

(11) Arendse, E.; Fawole, O. A.; Magwaza, L. S.; Opara, U. L. Non-Destructive Prediction of Internal and External Quality Attributes of Fruit with Thick Rind: A Review. *J Food Eng* **2018**, *217*, 11–23. https://doi.org/10.1016/j.jfoodeng.2017.08.009.

(12) El-Mesery, H.; Mao, H.; Abomohra, A. Applications of Non-Destructive Technologies for Agricultural and Food Products Quality Inspection. *Sensors* **2019**, *19* (4), 846. https://doi.org/10.3390/s19040846.

(13) Saeys, W.; Nguyen Do Trong, N.; Van Beers, R.; Nicolaï, B. M. Multivariate Calibration of Spectroscopic Sensors for Postharvest Quality Evaluation: A Review. *Postharvest Biol Technol* **2019**, *158* (August), 110981. https://doi.org/10.1016/j.postharvbio.2019.110981.

(14) Zhang, J.; Huang, Y.; Pu, R.; Gonzalez-Moreno, P.; Yuan, L.; Wu, K.; Huang, W. Monitoring Plant Diseases and Pests through Remote Sensing Technology: A Review. *Comput Electron Agric* **2019**, *165* (June), 104943. https://doi.org/10.1016/j.compag.2019.104943.

(15) Brunet, D.; Barthès, B. G.; Chotte, J.-L.; Feller, C. Determination of Carbon and Nitrogen Contents in Alfisols, Oxisols and Ultisols from Africa and Brazil Using NIRS Analysis: Effects of Sample Grinding and Set Heterogeneity. *Geoderma* **2007**, *139* (1–2), 106–117. https://doi.org/10.1016/j.geoderma.2007.01.007.

(16) Gherardi Hein, P. R.; Lima, J. T.; Chaix, G. Effects of Sample Preparation on NIR Spectroscopic Estimation of Chemical Properties of Eucalyptus Urophylla S.T. Blake Wood. *Holzforschung* **2010**, *64* (1), 45–54. https://doi.org/10.1515/hf.2010.011.

(17) Finzi, A.; Oberti, R.; Negri, A. S.; Perazzolo, F.; Cocolo, G.; Tambone, F.; Cabassi, G.; Provolo, G. Effects of Measurement Technique and Sample Preparation on NIR Spectroscopy Analysis of Livestock Slurry and Digestates. *Biosyst Eng* **2015**, *134*, 42–54. https://doi.org/10.1016/j.biosystemseng.2015.03.015.

(18) Arndt, M.; Rurik, M.; Drees, A.; Bigdowski, K.; Kohlbacher, O.; Fischer, M. Comparison of Different Sample Preparation Techniques for NIR Screening and Their Influence on the Geographical Origin Determination of Almonds (Prunus Dulcis MILL.). *Food Control* **2020**, *115* (February), 107302. https://doi.org/10.1016/j.foodcont.2020.107302.

(19) Nawi, N. M.; Chen, G.; Jensen, T. Prediction of Sugarcane Quality from Juice Samples Using Portable Spectroscopy. *JOURNAL OF MECHANICAL ENGINEERING AND SCIENCES* **2014**, *7* (December), 1219–1226. https://doi.org/10.15282/jmes.7.2014.21.0119.

(20) Maraphum, K.; Chuan-Udom, S.; Saengprachatanarug, K.; Wongpichet, S.; Posom, J.; Phuphaphud, A.; Taira, E. Effect of Waxy Material and Measurement Position of a Sugarcane Stalk on the Rapid Determination of Pol Value Using a Portable near Infrared Instrument. *J Near Infrared Spectrosc* **2018**, *26* (5), 287–296. https://doi.org/10.1177/0967033518795810.

(21) Phuphaphud, A.; Saengprachatanarug, K.; Posom, J.; Wongpichet, S.; Maraphum, K.; Taira, E. Effects of Waxy Types of a Sugarcane Stalk Surface on the Spectral Characteristics of Visible-Shortwave near Infrared Measurement. *Engineering Journal* **2019**, *23* (1), 13–24. https://doi.org/10.4186/ej.2019.23.1.13.

(22) Corrêdo, L. de P.; Maldaner, L. F.; Bazame, H. C.; Molin, J. P. Evaluation of Minimum Preparation Sampling Strategies for Sugarcane Quality Prediction by Vis-NIR Spectroscopy. *Sensors* **2021**, *21* (6), 2195. https://doi.org/10.3390/s21062195.

(23) Alfaro, G.; Meurens, M.; Birth, G. S. Liquid Analysis by Dry-Extract Near-Infrared Reflectance on Fiberglass. *Appl Spectrosc* **1990**, *44* (6), 979–986. https://doi.org/10.1366/0003702904086687.

(24) Fischer, W. B.; Eysel, H. H.; Nielsen, O. F.; Bertie, J. E. Corrections to the Baseline Distortions in the OH-Stretch Region of Aqueous Solutions. *Appl Spectrosc* **1994**, *48* (1), 107–112. https://doi.org/10.1366/0003702944027525.

(25) Li, W.; Goovaerts, P.; Meurens, M. Quantitative Analysis of Individual Sugars and Acids in Orange Juices by Near-Infrared Spectroscopy of Dry Extract. *J Agric Food Chem* **1996**, *44* (8), 2252–2259. https://doi.org/10.1021/jf9500750.

(26) Nicolaï, B. M.; Beullens, K.; Bobelyn, E.; Peirs, A.; Saeys, W.; Theron, K. I.; Lammertyn, J. Nondestructive Measurement of Fruit and Vegetable Quality by Means of NIR Spectroscopy: A Review. *Postharvest Biol Technol* **2007**, *46* (2), 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024.

(27) DU, Y. P.; LIANG, Y. Z.; KASEMSUMRAN, S.; MARUO, K.; OZAKI, Y. Removal of Interference Signals Due to Water from in Vivo Near-Infrared (NIR) Spectra of Blood Glucose by Region Orthogonal Signal Correction (ROSC). *Analytical Sciences* **2004**, *20* (9), 1339–1345. https://doi.org/10.2116/analsci.20.1339.

(28) Devaux, M. F.; Bertrand, D.; Robert, P.; Qannari, M. Application of Principal Component Analysis on NIR Spectral Collection after Elimination of Interference by a Least-Squares Procedure. *Appl Spectrosc* **1988**, *42* (6), 1020–1023. https://doi.org/10.1366/0003702884430443.

(29) Chen, D.; Shao, X.; Hu, B.; Su, Q. A Background and Noise Elimination Method for Quantitative Calibration of near Infrared Spectra. *Anal Chim Acta* **2004**, *511* (1), 37–45. https://doi.org/10.1016/j.aca.2004.01.042.

(30) Rambla, F. J.; Garrigues, S.; Guardia, M. De. PLS-NIR Determination of Total Sugar , Glucose , Fructose and Sucrose in Aqueous Solutions of Fruit Juices. **1997**, *2670* (97).

(31) Zahri, S.; Moubarik, A.; Charrier, F.; Chaix, G.; Baillères, H.; Nepveu, G.; Charrier, B. Quantitative Assessment of Total Phenol Contents of European Oak (Quercus Petraea and Quercus Robur) by Diffuse Reflectance NIR Spectroscopy on Solid Wood Surfaces. *Holzforschung* **2008**, *62* (6), 679–687. https://doi.org/10.1515/HF.2008.114.

(32) Vrtiška, D.; Šimáček, P. Prediction of 2-EHN Content in Diesel/Biodiesel Blends Using FTIR and Chemometrics. *Talanta* **2018**, *178* (July 2017), 987–991. https://doi.org/10.1016/j.talanta.2017.09.003.

(33) Storey, E. E.; Helmy, A. S. Optimized Preprocessing and Machine Learning for Quantitative Raman Spectroscopy in Biology. *Journal of Raman Spectroscopy* **2019**, *50* (7), 958–968. https://doi.org/10.1002/jrs.5608.

(34) Yu, H.-D.; Zuo, S.-M.; Xia, G.; Liu, X.; Yun, Y.-H.; Zhang, C. Rapid and Nondestructive Freshness Determination of Tilapia Fillets by a Portable Near-Infrared Spectrometer Combined with Chemometrics Methods. *Food Anal Methods* **2020**, *13* (10), 1918–1928. https://doi.org/10.1007/s12161-020-01816-1.

(35) Butler, H. J.; Smith, B. R.; Fritzsch, R.; Radhakrishnan, P.; Palmer, D. S.; Baker, M. J. Optimised Spectral Pre-Processing for Discrimination of Biofluids via ATR-FTIR Spectroscopy. *Analyst* **2018**, *143* (24), 6121–6134. https://doi.org/10.1039/c8an01384e.

(36) Lee, L. C.; Liong, C. Y.; Jemain, A. A. A Contemporary Review on Data Preprocessing (DP) Practice Strategy in ATR-FTIR Spectrum. *Chemometrics and Intelligent Laboratory Systems* **2017**, *163* (December 2016), 64–75. https://doi.org/10.1016/j.chemolab.2017.02.008.

(37) Jarvis, R. M.; Goodacre, R. Genetic Algorithm Optimization for Pre-Processing and Variable Selection of Spectroscopic Data. *Bioinformatics* **2005**, *21* (7), 860–868. https://doi.org/10.1093/bioinformatics/bti102.

(38) Devos, O.; Duponchel, L. Parallel Genetic Algorithm Co-Optimization of Spectral Pre-Processing and Wavelength Selection for PLS Regression. *Chemometrics and Intelligent Laboratory Systems* **2011**, *107* (1), 50–58. https://doi.org/10.1016/j.chemolab.2011.01.008.

(39) Zheng, H.; Cai, A.; Zhou, Q.; Xu, P.; Zhao, L.; Li, C.; Dong, B.; Gao, H. Optimal Preprocessing of Serum and Urine Metabolomic Data Fusion for Staging Prostate Cancer through Design of Experiment. *Anal Chim Acta* **2017**, *991*, 68–75. https://doi.org/10.1016/j.aca.2017.09.019.

(40) Jiao, Y.; Li, Z.; Chen, X.; Fei, S. Preprocessing Methods for Near-Infrared Spectrum Calibration. *J Chemom* **2020**, *34* (11), 1–19. https://doi.org/10.1002/cem.3306.

CHAPTER I

Sequence Combination Preprocessing Optimization with Constraint Matrix: A Comparison of 67 Different Datasets

---

The work presented in this chapter was carried out during the research internship at the Institute for Molecules and Materials at Radboud University, Nijmegen, The Netherlands (09/2021 to 02/2022) under the co-advisement of Prof. Jeroen J. Jansen.

ABSTRACT

The aim of this work was to present an algorithm that automatically searches for the best preprocessing strategy without fixing their sequence based on the artifact they fix. The number of preprocessing methods in each strategy and their hyperparameters were evaluated. The algorithm was compared with methods presented in the literature by Gerretzen et al. (2015) and Jiao et al. (2020). A fair, extensive, and comprehensive study was carried out, evaluating 67 different calibration datasets. The results showed that using a combination of three preprocessing methods in the preprocessing strategy resulted in an average root mean squared error of prediction (RMSEP) reduction of 31.2%, while using the combination of two preprocessing resulted in an average RMSEP reduction of 27.3% and using only one preprocessing method resulted in an average RMSEP reduction of 18.4%. The comparison with methods presented in the literature showed that the method proposed resulted in an average RMSEP reduction of 31.7%. In comparison, Gerretzen et al. method presented an average RMSEP reduction of 16.1%, and the Jiao et al. method presented an average RMSEP reduction of 20.6%. Overall, this work demonstrated that not fixing the sequence that the preprocessing is applied and permutating all methods were essential to find the best models with a significant reduction in the RMSEP values when compared with the other methods, therefore presenting a comprehensive insight on preprocessing.

Keywords: Chemometrics; Calibration; Partial Least Squares.

1 - INTRODUCTION

Nowadays, high-dimensional data analysis has become indispensable due to the rapid and continuous development of analytical technologies and its capability to replace laborious analysis [1–3]. The interpretation of such complex data requires a meticulous design of a proper modelling pipeline. In a typical modelling pipeline, data collection governs the initial quality of the data for interpretation. However, the quality of the data may be highly constrained due to hardware limitations. Data preprocessing therefore plays an essential role in the modeling process [4] to remove instrumental and other analytical artifacts and enhance the chemical information of interest, so that quantitative or qualitative models (from partial least squares to deep neural networks) may focus on the actual chemistry represented in the data.

In data analysis, preprocessing methods are mainly used to remove known artifacts with a known representation in the data, such as noise, baseline shift, slope, non-informative variables, and light scattering through the sample. These factors seriously decrease the signal-to-noise ratio and need to be eliminated to generate interpretative and robust models [5]. Each dataset will contain several artifacts originating from the instrument or sample and therefore requires a multifactorial specific and optimal preprocessing strategy to reconstruct the specific relationship of the spectra with the sample property of interest. Designing, optimizing, and selecting such optimal preprocessing is currently one of the main bottlenecks in data modeling [6–10].

Based on the literature, preprocessing is generally chosen by combining different preprocessing techniques based on trial-and-error, visual inspection, and past experiences [11,12]. Researchers may combine techniques that previously worked for themselves or were presented in the literature, without evaluating optimally in the currently studied dataset. This empirical approach typically cannot guarantee an optimal preprocessing selection. Manual preprocessing selection also has crucial challenges, which includes: (i) there are thousands of different preprocessing possibilities to try, (ii) it is difficult to determine the optimum preprocessing by just looking at the data, and (iii) past experiences might work only for similar data, and preprocessing needs to be updated in time even for the same analyzed system [13]. Despite everything, each data has its own peculiarities causing the most suitable preprocessing strategy to differ, which means that the tedious procedure of manual preprocessing must be repeated for different datasets. This strongly underpins the need for automated and evidence-based preprocessing optimization and search.

A few preprocessing optimization strategies have been introduced in the literature. Jarvis and Goodacre (2005)[14] presented an application of the genetic algorithm to optimize data preprocessing. The method includes thirteen different preprocessing methods and variations of their hyperparameters, resulting in 50 different preprocessing variations. The approach assessed approximately 100.000 preprocessing combinations and a computation time of nearly 5 days. Devos and Duponchel (2011)[15] presented a similar approach, which used a parallel genetic algorithm to optimize preprocessing strategy for support vector machines (SVM) and partial least squares - discrimination analysis (PLS-DA) models. Gerretzen et al. (2015)[16] presented an optimization method using design of experiments (DoE) to determine the best preprocessing strategy. This method was popularly used due to its simplicity and near-to-optimal performance. The DoE evaluated eighteen different preprocessing methods, and it was able of reducing the computation time from a day to less than an hour, finding similar results evaluating only a small fraction of the 4900 possible preprocessing combinations. Although the method presented satisfactory results, it is limited to a few preprocessings, according to the number of factors used in the DoE. The DoE requires previous knowledge about the effects of the preprocessing and the order of the preprocessing must be fixed to calculate the individual and combined effects. Zheng et al. (2017)[17] applied a similar DoE-based approach to optimize the preprocessing strategy. Another practical and effective preprocessing optimization strategy was proposed by Jiao et al. (2020)[18]. In this work, the authors created an algorithm that evaluates 108 preprocessing strategies commonly used in the literature. These works concluded that the effect of preprocessing is highly data-dependent, and therefore the preprocessing strategy should be optimized for each dataset individually.

There are a few automated preprocessing optimization methodologies proposed in the literature, however these methodologies lack a more comprehensive assessment of the preprocessing strategies. These methods evaluate too few preprocessing strategies [18] or are limited by the design factors (fixed order) and levels studied when using a DoE approach [16]. Although these methods were capable of find a preprocessing strategy that improved the performance when compared to using no preprocessing, the best preprocessing combination among the preprocessing methods evaluated might not been found due to the limited searching space imposed by the methodology used.

This work presents an approach that automatically searches for the best preprocessing strategy by combining the preprocessing methods assessed without fixing the sequence that they are applied. The approach was called Preprocessing Optimizer (PPO). The PPO was compared with the methods presented in the literature by Gerretzen et al. [16] and Jiao et al. [18]. In

addition, the number of preprocessing methods applied in the pipeline and the effect of preprocessing hyperparameters optimization were evaluated. The proposed method reduces the time and effort involved in this step, aiding unexperienced users, thus facilitating the achievement of the analysis purpose, choosing the best preprocessing strategy with its hyperparameters, therefore presenting a comprehensive insight on preprocessing.

2 - METHODS

2.1 - DATASETS

Preprocessing optimization analysis can often be biased when tested on too little datasets. To carry out a fair, extensive, and comprehensive study for data preprocessing optimization, 67 different calibration cases were evaluated. The different calibration cases were obtained from 14 publicly available data sources. Some of the datasets contain samples obtained using different instruments or different properties, leading to 67 calibration cases. A detailed description of each dataset can be seen in Table I-1.

Table I-1. Calibration cases used to evaluate the preprocessing optimization algorithm.

Case	Matrix	Source	Instrument	Property	Samples	Variables	Range
1	Mixture [19]	Fluorescence	-	Catechol µmol L^{-1}	38	221	270 - 380 nm
2				Hydroquinone µmol L^{-1}	38	221	270 - 380 nm
3				Phenol µmol L^{-1}	38	221	270 - 380 nm
4	Roasted coffee [20]	GC	-	Quality Score	159	860	2.1 - 19.2 min
5	Corn [21]	NIR	M5	Moisture wt %	80	700	1100 - 2498 nm
6				Oil wt %	80	700	1100 - 2498 nm
7				Protein wt %	80	700	1100 - 2498 nm
8				Starch wt %	80	700	1100 - 2498 nm
9			MP5	Moisture wt %	80	700	1100 - 2498 nm
10				Oil wt %	80	700	1100 - 2498 nm
11				Protein wt %	80	700	1100 - 2498 nm
12				Starch wt %	80	700	1100 - 2498 nm
13			MP6	Moisture wt %	80	700	1100 - 2498 nm
14				Oil wt %	80	700	1100 - 2498 nm
15				Protein wt %	80	700	1100 - 2498 nm
16				Starch wt %	80	700	1100 - 2498 nm
17	Diesel [21]	NIR	-	50 % boiling point ºC	395	401	750 - 1550 nm
18				Cetane Number	381	401	750 - 1550 nm
19				Density g mL^{-1}	395	401	750 - 1550 nm
20				Flash ºC	395	401	750 - 1550 nm
21				Freezing point ºC	395	401	750 - 1550 nm

Case	Matrix	Source	Instrument	Property	Samples	Variables	Range
22				Total aromatics wt %	395	401	750 - 1550 nm
23				Viscosity cSt	395	401	750 - 1550 nm
24	Grain [21]	NIR	-	Casein wt %	231	117	1104 - 2496 nm
25				Glucose wt %	231	117	1104 - 2496 nm
26				Lactate wt %	231	117	1104 - 2496 nm
27				Moisture wt %	231	117	1104 - 2496 nm
28	Marzipan [22]	NIR	Bomen MB 160 Diffusir	Moisture wt %	32	664	850 - 2443 nm
29				Sugar wt %	32	664	851 - 2443 nm
30		MID	PerkinElmer System 2000	Moisture wt %	32	950	3800 - 650 cm^{-1}
31				Sugar wt %	32	950	3801 - 650 cm^{-1}
32		NIR	Infraprover II	Moisture wt %	32	406	1050 - 2147 nm
33				Sugar wt %	32	406	1050 - 2147 nm
34		Vis-NIR	NIRSystems 6500	Moisture wt %	32	1000	450 - 2448 nm
35				Sugar wt %	32	1000	450 - 2448 nm
36		NIR	NIRSystems 6500	Moisture wt %	32	600	850 - 2048 nm
37				Sugar wt %	32	600	850 - 2048 nm
38		NIR	Infratex 1255	Moisture wt %	32	100	850 - 1048 nm
39				Sugar wt%	32	100	850 - 1048 nm
40	Gasoline [23]	NIR	-	Octane Number	60	401	850 - 1048 nm
41	Soil [24]	Vis-NIR	-	Ergosterol mg g^{-1}	108	1050	400 - 2498 nm
42				Organic Matter %	108	1050	400 - 2498 nm
43	Tablets [25]	NIR	-	Active compound wt %	310	404	10507 - 7400 cm^{-1}

Case	Matrix	Source	Instrument	Property	Samples	Variables	Range
44	Wheat [26]	NIR	-	Protein wt %	523	100	850 - 1048 nm
45	Pork Fat [27]	Raman	-	C14	105	5667	1900 - 200 cm^{-1}
46				C16	105	5667	1900 - 200 cm^{-1}
47				C17	105	5667	1900 - 200 cm^{-1}
48				C18	105	5667	1900 - 200 cm^{-1}
49				C20	105	5667	1900 - 200 cm^{-1}
50				Iodine Value wt %	105	5667	1900 - 200 cm^{-1}
51				MUFA %	105	5667	1900 - 200 cm^{-1}
52				PUFA %	105	5667	1900 - 200 cm^{-1}
53				SFA %	105	5667	1900 - 200 cm^{-1}
54				UFA %	105	5667	1900 - 200 cm^{-1}
55	Tablets [21]	Raman	-	Active compound %	120	3401	3600 - 200 cm^{-1}
56	Mixture [28]	NMR	-	Butanol %	231	1077	0.64 - 3.84 ppm
57				Pentanol %	231	1077	0.64 - 3.84 ppm
58				Propanol %	231	1077	0.64 - 3.84 ppm
59	Tablets [21]	Vis-NIR	Instrument 1	Assay	655	650	600 - 1898 nm
60				Hardness	655	650	600 - 1898 nm
61				Weight g	655	650	600 - 1898 nm
62			Instrument 2	Assay	655	650	600 - 1898 nm
63				Hardness	655	650	600 - 1898 nm
64				Weight g	655	650	600 - 1898 nm
65	Mixture [19]	Voltammetry	-	Ascorbic acid µmol L^{-1}	32	525	0.01 - 1.5 V

Case	Matrix	Source	Instrument	Property	Samples	Variables	Range
66				Uric acid μmol L^{-1}	32	525	0.01 - 1.5 V
67				Dopamine μmol L^{-1}	32	525	0.01 - 1.5 V

GC: gas chromatography; NIR: near-infrared spectroscopy; MID: mid-infrared spectroscopy; Vis-NIR: visible and near-infrared spectroscopy; and NMR: nuclear magnetic resonance.

2.2 - PREPROCESSING METHODS

Each data can be affected by instrumental and sampling artifacts, originated from the instrument used, such as noise, baseline deviation, slope, and light scattering, or from the sample, such as the scale of the variables. Each artifact, or their combination, can be corrected using a preprocessing strategy. There is no rule about the order in which the preprocessing methods should be applied. Therefore, the need to try different combinations to find the suitable one. The proposed algorithm includes twenty-three different preprocessing methods, 1) smoothing (SG)[29], 2) first derivative (SG1D)[29], 3) second derivative (SG2D)[29], 4) asymmetric least squares (AsLs)[30], 5) baseline offset (BO)[5,16], 6) detrending (D)[5,16], 7) linear baseline (LB)[5,16], 8) multiplicative scatter correction (MSC)[31], 9) net analytical signal (NAS)[32], 10) normalization (Norm), 11) pairwise detrending (PD)[18], 12) standard normal variate (SNV)[33], 13) robust normal variate (RNV)[34], 14) autoscaling (AS), 15) level scaling (LS)[5,16], 16) mean centering (MC), 17) mean scaling[5,16], 18) median scaling[5,16], 19) minimum scaling[5,16], 20) Pareto scaling (PS)[5,16], 21) Poisson scaling (POS)[5,16], 22) power scaling[5,16], and 23) range scaling (RS)[5,16]. A more detail description can be found in Table I-2. All preprocessing methods were selected based on literature review and previous experience.

Table I-2. Preprocessing methods included in the algorithm.

Method	Equation	Description
Smoothing [29] First Derivative [29] Second Derivative [29]	$x_t = \frac{1}{h}\left(\sum_{i=-\frac{w-1}{2}}^{\frac{w-1}{2}} a_i x_{t+i}\right)$	Where x represents the input value, w represents the filter size, a represents the polynomial coefficient, and h is the normalizing factor
AsLs [30]	$Q = \sum_i (1-p)(y_i - f_i)^2 + \lambda \sum_i (\Delta^2 f_i)^2$	Where y represents the input data, p is asymmetric parameter, λ smoothing factor, f is a smooth trend, and Δ is the difference operator.
Baseline Offset [5,16]	$x_i = x_i - \min(x_i)$	Where x_i represents the i^{th} row of X
Detrending [5,16]	$x_i = \sum_0^n \beta_n (x_i)^n$	Where β_n represents the coefficients of polynomial of order n that fits x_i against the number of variables in x_i
Linear Baseline [5,16]	$x_i = x_i - (\frac{x_{ij} - x_{i1}}{j} - x_{i1})$	Where x_{ij} represents the last variable and x_{i1} the first variable in x_i, and j is the total number of variables
MSC [31]	$x_i = \beta_0 + \beta_1 x_i$	Where β_0 and β_1 are the coefficients of the polynomial of order 1 that fits x_i against the mean $\bar{x}$
NAS [32]	$x_k^* = [I - (X_{-k})^+ X_{-k}]x$	Where I is the identity matrix, X_{-k} is the matrix representing the space spanned by the spectra of all other analytes except k, X_{-k}^+ is its pseudoinverse.
Normalization	$x_i = \frac{x_i}{\|x_i\|}$	Where $\|x_i\|$ can be the norm to unit area, norm to unit length, or norm to maximum value
Pairwise Detrending [18]	$x_i = \sum_0^n \beta_n (x_i)^n$	Where β_n represents the coefficients of the polynomial of order n that fits x_i against $\bar{x}$
SNV [33]	$x_i = \frac{x_i - \bar{x}_i}{\text{std}(x_i)}$	Where x_i represents the i^{th} row of X
RNV [34]	$x_i = \frac{x_i - \text{percentile}(x_i, k)}{\text{std}(x_i \le \text{percentile}(x_i, k))}$	Where percentile(x_i,k) is the k^{th} percentile of x
Autoscaling	$x_j = \frac{x_j - \bar{x}_j}{\text{std}(x_j)}$	Where x_j represents the j^{th} column of X

Method	Equation	Description
Level Scaling [5,16]	$x_j = \frac{x_j - \bar{x}_j}{\bar{x}_j}$	Where x_j represents the j^{th} column of X
Mean Centering	$x_j = x_j - \bar{x}_j$	Where x_j represents the j^{th} column of X
Mean Scaling [5,16]	$x_i = x_i - \bar{x}_i$	Where x_i represents the i^{th} row of X and $\bar{x}$ is the average of x_i
Median Scaling [5,16]	$x_i = x_i - \hat{x}_i$	Where x_i represents the i^{th} row of X and $\hat{x}$ is the median of x_i
Minimum Scaling	$x_{i,j} = x_{i,j} - \min(x_j)$ $x_{i,j} = x_{i,j} + \bar{x}_j$	Where i represents the i^{th} row and j the j^{th} column of X and $\bar{x}$ is the average of x_j
Pareto Scaling [5,16]	$x_j = \frac{x_j - \bar{x}_j}{\sqrt{\text{std}(x_j)}}$	Where x_j represents the j^{th} column of X
Poisson Scaling [5,16]	$x_j = \frac{x_j - \bar{x}_j}{\sqrt{\bar{x}_j}}$	Where x_j represents the j^{th} column of X
Power Scaling [5,16]	$x_j = \sqrt{x_j} - \text{mean}\left(\sqrt{x_j}\right)$	Where x_j represents the j^{th} column of X
Range Scaling [5,16]	$x_j = \frac{x_j - \bar{x}_j}{\max(x_j) - \min(x_j)}$	Where x_j represents the j^{th} column of X

AsLs: asymmetric least squares; MSC: multiplicative scatter correction; NAS: net analytical signal; SNV: standard normal variate; and RNV: robust normal variate.

2.3 - OPTIMIZATION ALGORITHM AND CONSTRAINT MATRIX

The PPO combines the preprocessing methods selected, up to three preprocessing methods in the preprocessing strategy, without fixing the sequence in which the preprocessing is performed. As some preprocessing methods has hyperparameters, i.e., parameters that need to be set before executing the preprocessing method, such as: i) as filter size and polynomial order for Savitzky-Golay derivatives; ii) smoothing factor and asymmetric parameter for AsLs; iii) the number of components for NAS; iv) the type of norm for normalizing; v) the order for pairwise detrend, and vi) the percentile for RNV, the PPO also optimizes the preprocessing methods' hyperparameters prior to the search.

The main advantage of the PPO is that it searches for the best preprocessing strategy in a larger number of possibilities and considers the optimizations of the preprocessing methods' hyperparameters when compared with Gerretzen et al.[16] and Jiao et al.[18]. The hyperparameters for Gerretzen et al.[16] algorithm must the set a priori and Jiao et al.[18] optimizes the hyperparameters a prior, but both searches for the best preprocessing strategy considering a limited searching space. Figure I-1 shows a flowchart of the proposed method and the methods presented by Gerretzen et al.[16] and Jiao et al.[18]. Gerretzen et al.[16] methodology is based on design of experiments, therefore the searching space is limited to the factors and levels defined for the design, in addition, a preprocessing sequence needs to be defined to apply the preprocessing methods and also the hyperparameters values set a priori. Consequently, considering a factorial design, only 16 preprocessing strategies are evaluated. Jiao et al.[18] method is a compilation of 108 preprocessing strategies commonly used in the literature, also the hyperparameters are optimized prior to the search.

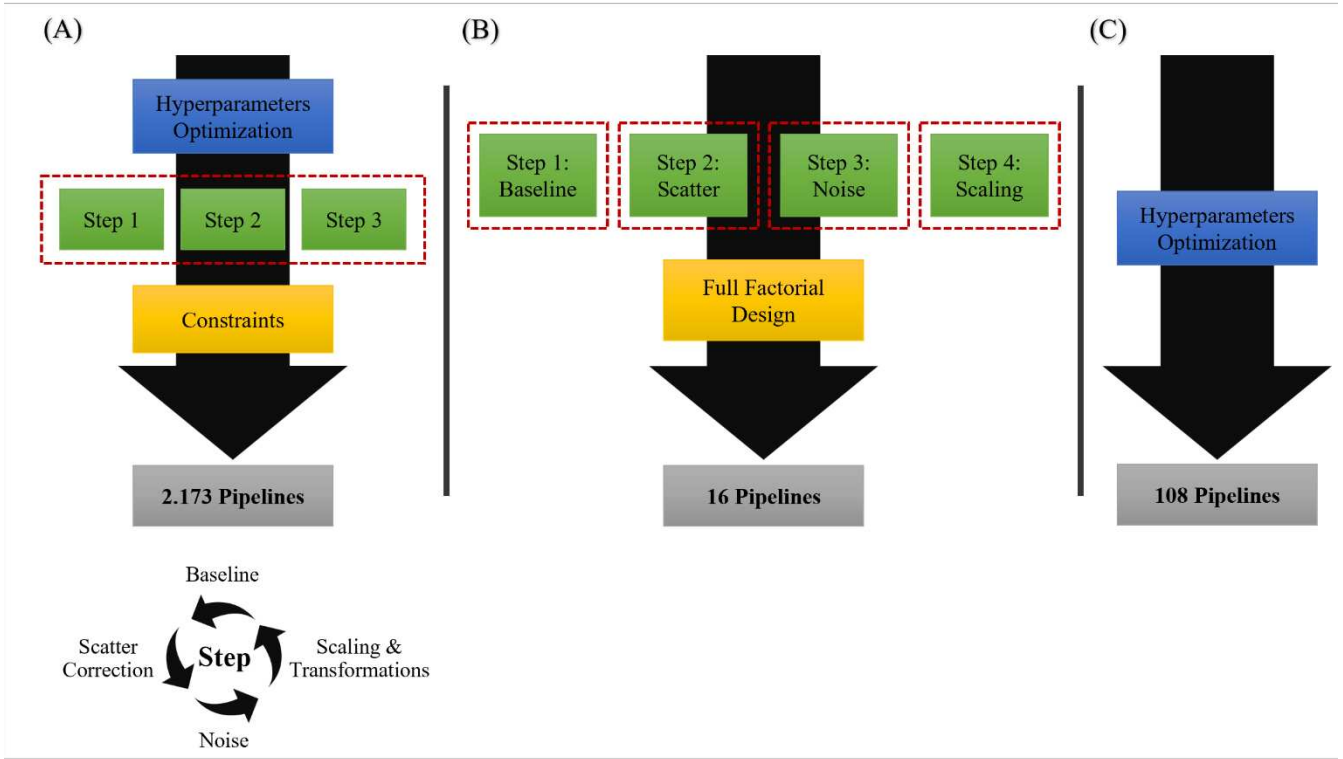


Figure I-1. Flowchart of (A) the proposed optimization method, (B) Gerretzen et al. [16] method, and (C) Jiao et al. [18] method.

The PPO algorithm evaluates the preprocessing strategy by combining different preprocessing methods. As some preprocessing methods have hyperparameters, with a combination of more than 3 preprocessing methods, the number of different strategies is too large, which is time-consuming and could take days depending on the data size. For instance, with a maximum combination of 3 preprocessing methods, considering all preprocessing methods available and their hyperparameters, the algorithm would assess 1,955,028 different preprocessing strategies, which is not practicable. Therefore, to reduce the number of preprocessing strategies to be evaluated, the hyperparameter of each preprocessing method is optimized first, and then the preprocessing combinations are evaluated. Each preprocessing method with hyperparameter is evaluated individually and the hyperparameter that returns the smallest root mean square error of cross-validation (RMSECV) is selected and set for subsequent search for the best preprocessing strategy. This approach reduced the number of preprocessing combinations to 12,721. The hyperparameters values of the preprocessing methods assessed in this work can be found in Table I-3. The effects of hyperparameters optimization were evaluated comparing the prior optimization of each hyperparameters and the use of their default values. The default values were defined based on values most used in other chemometric softwares.

Table I-3. Hyperparameters values set for each preprocessing method.

Preprocessing Method	Hyperparameter	Values of the Hyperparameter	Default Values
Smoothing	Filter size	3 5 7 9 11 13 15 17 19 21 23	13
First derivative	Filter size	3 5 7 9 11 13 15 17 19 21 23	13
	Polynomial order	1 2 3 4	2
Second derivative	Filter size	3 5 7 9 11 13 15 17 19 21 23	13
	Polynomial order	1 2 3 4	2
Normalize	Type	1 2 Inf	3
Detrend	Order	1 2 3	2
AsLs	Smoothing factor	1×10^{2} 1×10^{3} 1×10^{4} 1×10^{5} 1×10^{6} 1×10^{7} 1×10^{8} 1×10^{9}	1×10^{9}
	Asymmetric parameter	1×10^{-3} 1×10^{-2} 1×10^{-1}	1×10^{-3}
RNV	Percentile	10 20 30 40 50	30
Pairwise detrend	Order	1 2 3	2
NAS	Component	1 2 3	2

AsLs: asymmetric least squares; NAS: net analytical signal; and RNV: robust normal variate

In addition, some constraints (Figure I-2) were created to avoid unwanted combinations of preprocessing methods and save computation time, reducing the number of possible preprocessing strategies to 2,173, which could be executed in suitable time.

It is important when choosing the preprocessings methods that the users use the chemical intuition and critical evaluation of the data to choose the preprocessing methods suitable for each data. The PPO methodology gives the users total autonomy to choose which preprocessing methods they want to evaluate, the maximum number of preprocessing methods in the strategy, and the hyperparameters for each dataset. For this work, all preprocessing methods available were used, and the maximum number of preprocessing methods in the strategy was set to three. Our purpose in this work is to evaluate the algorithm capability to select the best preprocessing strategy without intervention of the user. The PPO has potential to assist nonexperience users in the selection of preprocessing methods.

The PPO algorithm was implemented in MATLAB 2019a (Math Works, Natick, USA). Gerretzen et al.[16] and Jiao et al.[18] was also executed in MATLAB environment.

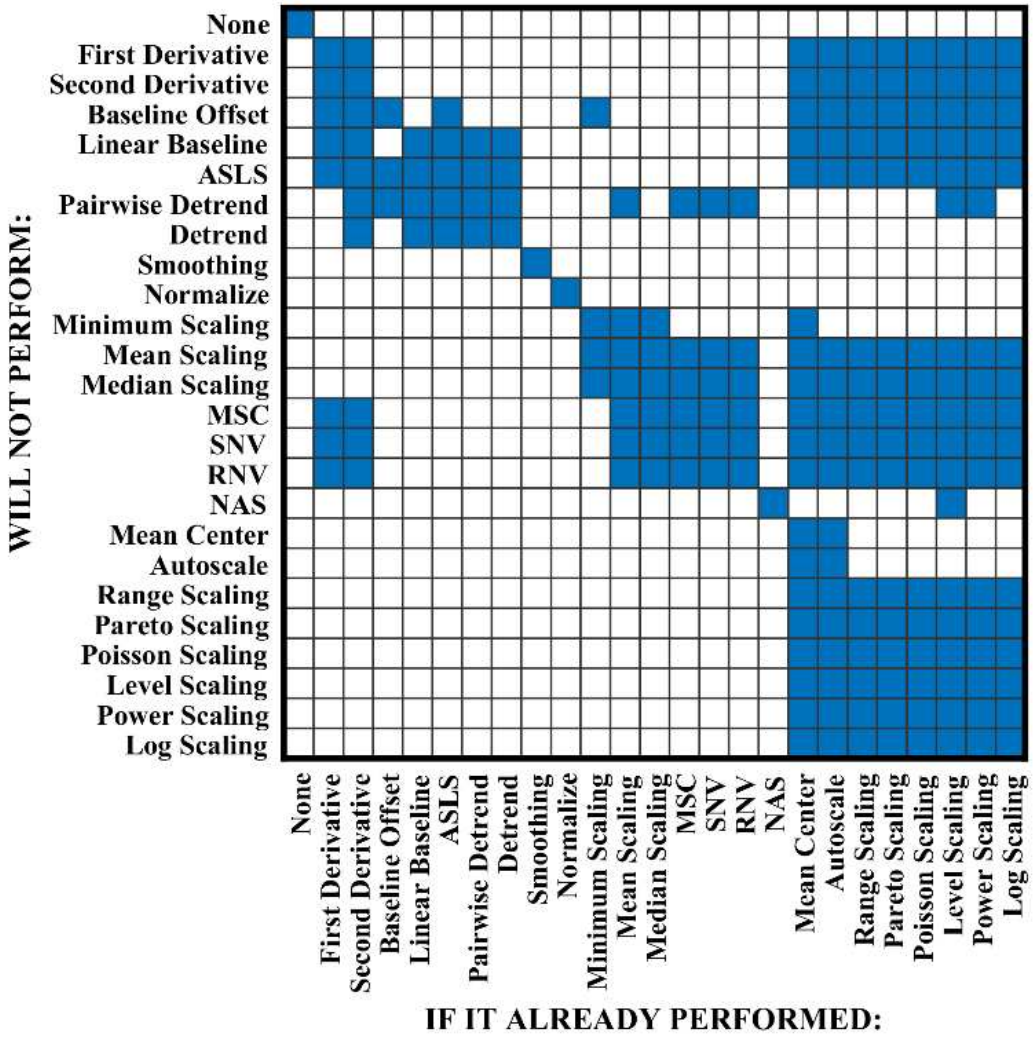


Figure I-2. Constraints of the optimization algorithm. The blue boxes indicate that the preprocessing on Y-axis will not be performed if the preprocessing on X-axis has already been performed.

2.4 - ASSESSMENT OF THE PREPROCESSING SEQUENCE

The influence of the sequence in which the preprocessing methods are applied in the strategy was assessed by comparing the results of all the models created using the PPO methodology, i.e., without fixing the preprocessing sequence, with the performance of the models fixing the preprocessing sequence, i.e., first baseline/slope correction, followed by scattering correction, then smoothing, and finally scaling.

2.5 - VALIDATION OF THE OPTIMIZATION ALGORITHM

Each dataset evaluated was split into a calibration set consisting of 70% of the data and a test set composed of 30% using the Kennard-Stone algorithm [35]. Partial least squares (PLS) regression was used. The dependent variables were always mean-centered. The number of latent variables and the best preprocessing strategy was selected using the root mean square error of cross-validation (RMSECV). Ten-fold venetian blinds cross-validation was used, and 1 to 15 latent variables (LV) were evaluated. The number of LV for each model was selected considering the RMSECV value using LV and LV+1, if the difference between the RMSECV using LV and LV+1 is less than 5%, LV is selected. After selecting the best model, the root mean square error of prediction (RMSEP) was calculated for the test set and used as an accuracy parameter in this work. In this work, the best model is the model that returned the smallest RMSECV value.

To evaluate the number of preprocessing methods that should be applied to the data, the accuracy when using no preprocessing, one preprocessing method, two preprocessing methods, and three preprocessing methods were compared.

The PPO algorithm was compared with the ones available in the literature presented by Gerretzen et al.[16] and Jiao et al.[18]. These methodologies were chosen as they are simple and can be executed within a reasonable time when compared with other methods, such as the one using genetic algorithm that could take days to be executed.

For Gerretzen et al.[16] method the number of repeats for validation was set to one, the number of repeats for optimization was set to 3, and the fraction of data to include in the test set in validation and optimization was set to 30 %. For Jiao et al.[18] method the number of repeats was set to one.

The nonparametric Wilcoxon Rank Sum test [36] with 95% confidence was used to evaluate if there are any statistically significant differences between the different methodologies used. As the properties of interest presented in this work for each dataset has different

magnitudes, a new index called prediction error index (PEI) was created to facilitate visualization and comparison. For PEI calculation the RMSEP values of the preprocessing strategies were normalized according to the RMSEP using the data with no preprocessing, as follows:

$$PEI = \left(\frac{RMSEP_{strategy}}{RMSEP_{raw}}\right) \times 100 \quad \text{Eq. (1)}$$

where $RMSEP_{strategy}$ is the RMSEP value of the optimum preprocessing strategy, $RMSEP_{raw}$ is the RMSEP value using no preprocessing, and PEI is the normalized RMSEP value. The PEI shows how much the RMSEP value increased or decreased compared to the RMSEP using no preprocessing. Therefore, the PEI using no preprocessing is 100% for all calibration cases.

3 - RESULTS AND DISCUSSION

The results and discussion section are divided in four parts: i) effects of the hyperparameters optimization, ii) effects of sequence permutation, iii) effects of the number of preprocessing in the strategy, and iv) comparison with other preprocessing selection methods.

3.1 - EFFECTS OF HYPERPARAMETERS OPTIMIZATION

The effects of hyperparameter optimization were evaluated comparing the results of the models using the default hyperparameters values and optimizing the hyperparameters a priori (Table I-3).

Figure I-3 shows the PEI values of the calibration cases studied using the default and optimizing the hyperparameters values. Considering all the calibrations cases investigated, using the default values resulted in the smallest PEI values for 32 calibration cases, with an average PEI reduction of 32.9%, while optimizing the hyperparameters resulted in the smallest PEI values for 39 calibration cases, with an average PEI reduction also of 32.7%. No statistical difference was found between using the default values and optimizing the hyperparameters (*p*-value 0.433). Although no statistical difference was found, the optimizing options was used, as it resulted in smallest PEI values for 39 calibration cases and this step is very fast.

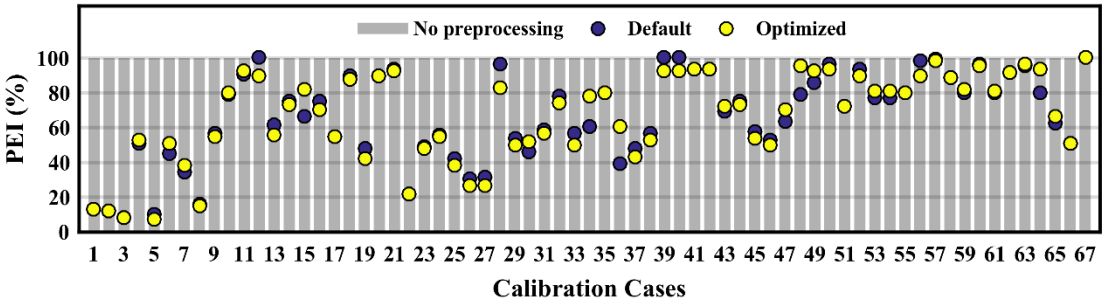


Figure I-3. PEI values of the calibration cases using the default hyperparameters values (blue circles) and optimizing the hyperparameters values (yellow circles). The gray bar represents the models using no preprocessing, i.e., PEI equal to 100%.

3.2 - EFFECTS OF SEQUENCE PERMUTATION

The idea of the proposed algorithm is to evaluate the combination of the preprocessing methods available without fixing their sequence based on the artifact they fix. There is not an agreement about the sequence that the preprocessings methods should be applied, however, Gerretzen et al.[16], based on literature review, proposed that the preprocessing methods should be applied in the following fixed sequence, first baseline/slope correction, followed by scattering correction, then smoothing, and finally scaling. Figure I-4 shows the PEI values of

the calibration cases using the methodology presented in this work without fixing the preprocessing sequence and fixing the sequence. Not fixing the sequence, resulted in the smallest PEI values for 46 calibrations cases, with an average PEI reduction of 32.7%, while fixing the preprocessing sequence resulted in the smallest PEI reduction for 22 calibration cases, with and average PEI reduction of 28.4%. Our methodology presented a significative PEI reduction compared with the method fixing the preprocessing sequence (p-value 6.9×10^{-5}) with an average PEI gain of 4.2%.

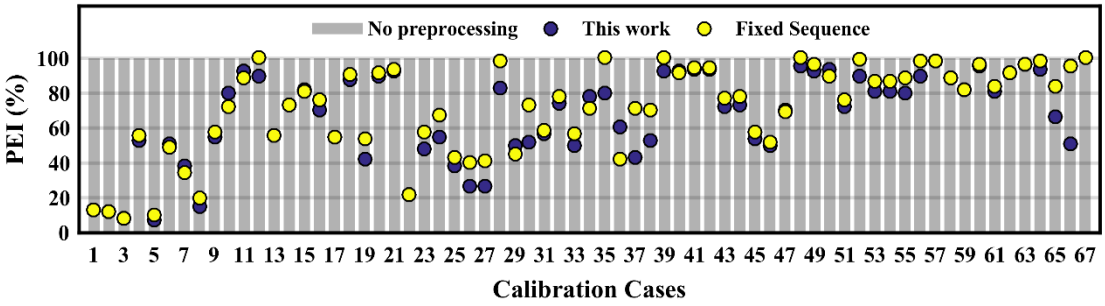


Figure I-4. PEI values of the calibration cases using the methodology presented in this work (blue circle) and fixing the sequence which the preprocessings were applied (yellow circles). The gray bar represents the models using no preprocessing, i.e., PEI equal to 100%.

3.3 - COMBINATION OF PREPROCESSING METHODS

To analyze the combination of preprocessing methods, the best models for each regression case using one, two, and three preprocessing methods in the preprocessing strategy were compared. According to Figure I-5, increasing the number of preprocessing in the strategy also increased the accuracy of most of the models, i.e., reduced the PEI values. Using a preprocessing strategy consisting of three different preprocessing methods resulted in the best models for 52 calibration cases, with an average PEI reduction of 31.2%. In contrast, using a preprocessing strategy consisting of two different preprocessing methods resulted in an average PEI reduction of 27.3%, while using only one preprocessing resulted in an average PEI reduction of 18.7%. The preprocessing strategy using a combination of three preprocessing methods presented a significative PEI reduction compared to the preprocessing strategy using two preprocessing methods (p-value 3.6×10^{-6}) and the preprocessing strategy using two preprocessing methods presented a significative PEI reduction compared to the preprocessing strategy using only one preprocessing method (p-value 8×10^{-8}). Therefore, using more than one preprocessing method is recommended.

These results demonstrated the importance of having an algorithm capable of performing an automatic preprocessing strategy search, as manual search based on trial-and-

error is not practicable due to the large number of preprocessing strategy possibilities. Besides assessing the raw spectra, our methodology performs 24 preprocessing strategies consisting of one preprocessing, 146 strategies consisting of two preprocessing, and 2002 strategies consisting of three preprocessing, resulting in 2,173 different preprocessing strategies.

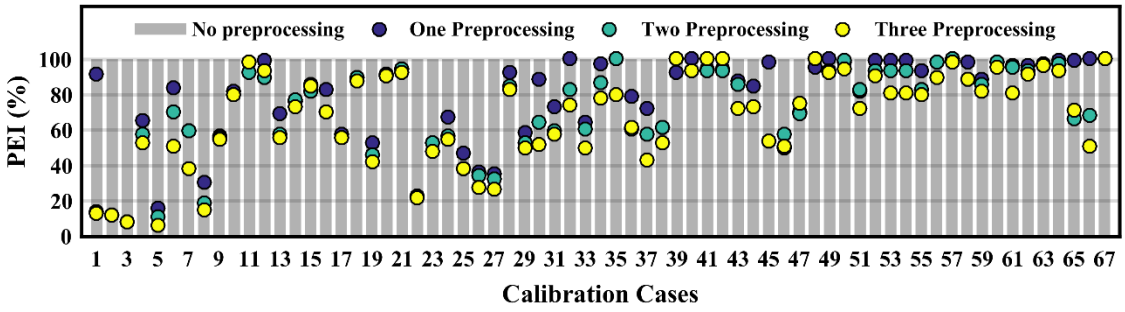


Figure I-5. PEI values of the calibration cases using one preprocessing method in the pipeline (blue circles), two preprocessing methods in the pipeline (aqua green circles), and three preprocessing methods in the pipeline (yellow circles). The gray bar represents the models using no preprocessing, i.e., PEI equal to 100%.

3.4 - COMPARISON WITH LITERATURE

The PPO was compared with the ones demonstrated in the literature by Gerretzen et al.[16] and Jiao et al.[18]. The main advantage of our methodology is that it tries out combinations of the preprocessing methods available without fixing the preprocessing sequence. Considering the 67 calibration cases, our methodology presented an average PEI reduction of 31.7%, while Gerretzen et al.[16] and Jiao et al.[18] methods presented an average PEI reduction of 20.6% and 16.1%, respectively. The methodology proposed in this work presented a significant PEI reduction compared to Gerretzen et al.[16] (p-value 8.2×10^{-8}) and Jiao et al.[18] (p-value 8.7×10^{-7}). Figure I-6 shows the PEI values for the 67 calibration cases using the three preprocessing optimization methodologies. The method proposed in this work resulted in the smallest PEI values for 48 calibration cases, while Gerretzen et al.[16] method resulted in the smallest PEI values for 8 calibration cases, and Jiao et al.[18] method for 12 calibration cases. In addition, our methodology reduced the number of LV used in the models. Comparing the three methodologies, the number of LV were the smallest in 62 calibrations cases when using our method, for 5 calibration cases using Gerretzen et al.[16] method, and 5 calibration cases using Jiao et al.[18] method. These results support the idea that a more versatile searching methodology is more robust and crucial in searching for the best preprocessing strategy while also decreasing the complexity of the models.

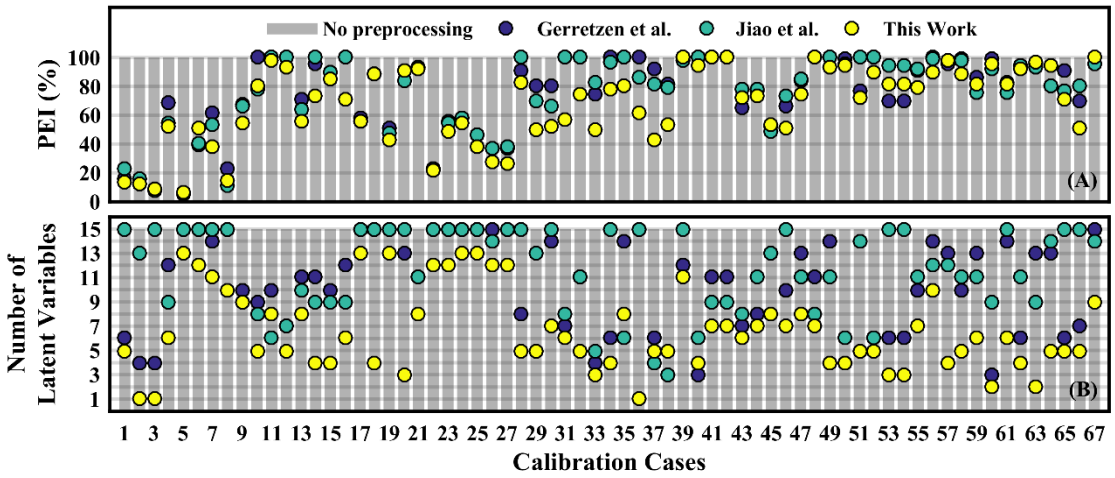


Figure I-6. PEI values (A) and number of latent variables (B) for the methodology presented in this work (yellow circles), presented by Gerretzen et al.[16] (blue circles), and presented by Jiao et al.[18] (aqua green circles) for the 67 calibration cases. The gray bar represents the models using no preprocessing, i.e., PEI equal to 100%.

Figure I-7 shows the preprocessed spectra and regression coefficients for the best preprocessing strategy found using the Gerretzen et al.[16] method (PEI 66.5%), Jiao et al.[18] method (PEI 73.1%), the method proposed in this work (PEI 51.2%), and using no preprocessing (PEI 100%) for calibration case 46. In Figure I-7E, it can be observed that as the PEI values decrease, the absolute weights of the variables marked in red increased, while the absolute weights of the variables marked white decreased. These results show that allowing the algorithm to combine the preprocessing methods is an advantage as a larger searching space is used. Besides reducing the PEI value, our method also reduced the number of LV used, resulting in a more parsimonious model. Our method selected 7 LV, while Gerretzen et al.[16] and Jiao et al.[18] selected 10 and 15 LV. The best processing strategies, number of latent variables, and PEI values found using the three methodologies for the 67 calibration cases can be found in Table I-4. It can be observed that the PPO increase the accuracy of the models and reduced the complexity of the models.

The number of LV in a PLS model can be influenced by various factors, including data preprocessing. Different preprocessing techniques can impact the number of LV needed to explain the relationship between the dependent and independent variables. Therefore, the number of LV was not fixed to make the comparison as each method has its own built-in LV and preprocessing selection approach. The goal of this work is to evaluate and compare the ability of each preprocessing optimization method to return a suitable preprocessing strategy and LV automatically.

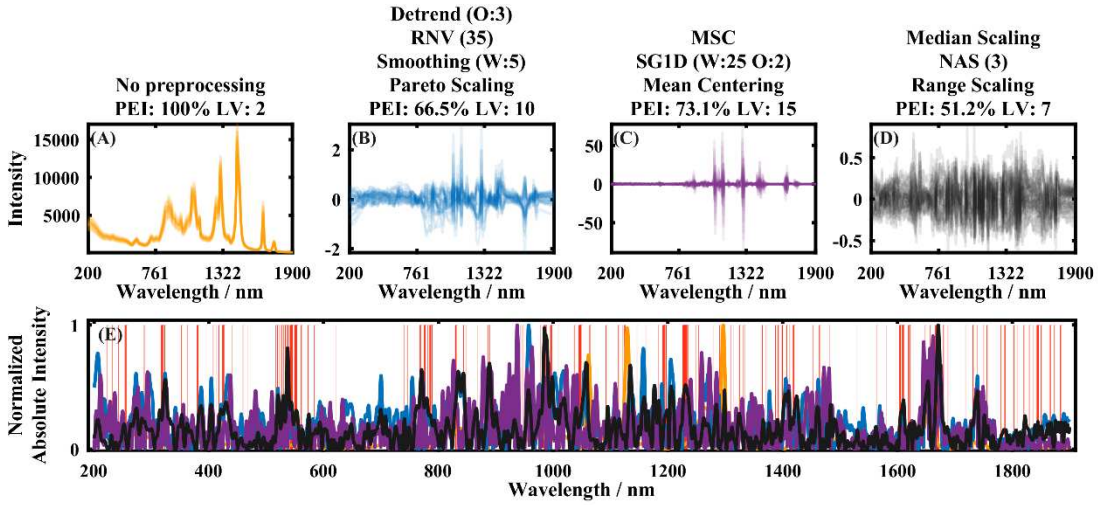


Figure I-7. Preprocessed spectra for calibration case 46 using no preprocessing (A), Gerretzen et al.[16] method (B), Jiao et al.[18] method (C), and the method proposed in this work (D), and their normalized absolute regression coefficients (E). The red regions indicate the weights that had increased and the white regions the weights that had decreased.

Table I-4. Best preprocessing strategy, latent variable, and PEI value of each calibration case using the methodology proposed, Gerretzen et al.[16], and Jiao et al.[18] methods.

Case	This Work			Gerretzen et al.			Jiao et al.		
	LV	PEI (%)	Preprocessing Strategy	LV	PEI (%)	Preprocessing Strategy	LV	PEI (%)	Preprocessing Strategy
01	5	13.3	SG (W:3) - PD (O:3) - SNV	6	15.4	D (4) + SG (W:11 O:4) + MC	15	23.4	SG1D(W:11 O:3) + AS
02	1	12.3	PD (O:1) - NAS (3) - SG1D (W:7 O:1)	4	12.5	SG (W:11 O:3) + MC	13	15.6	AS
03	1	8.4	MC - NAS (3) - PD (O:1)	4	8.5	MC	15	8.1	MC
04	6	52.6	AsLs (L:100 p:0.1) - NAS (2) - AS	12	68.6	D (4) + SG (W:5 O:4) + MC	9	54.6	AS
05	13	6.7	MC - NAS (3) - D (O:3)	15	4.7	SG (W:11 O:2) + MC	15	4.7	MC
06	12	50.8	SG (W:3) - SG1D (W:7 O:3) - LS	15	39.6	D (4) + SG (W:11 O:4) + Log Scaling	15	40.8	SG1D(W:19 O:3) + AS
07	11	38.2	SG2D (W:23 O:2) - SG (W:3) - LS	14	61.1	MC	15	54.0	SG2D(W:21 O:2) + AS
08	10	14.7	NAS (3) - SG1D (W:13 O:4) - LS	15	23.1	MC	15	10.8	SG1D(W:9 O:1) + AS
09	9	54.8	PD (O:3) - Min Scaling - SNV	10	67.6	SG (W:11 O:3) + PS	9	66.3	MC
10	5	79.9	AS - NAS (3)	9	110.1	Level Scaling	8	78.2	AS + SG2D(W:13 O:2)
11	8	98.3	D (O:1) - Min Scaling - LS	10	110.0	SG (W:5 O:4) + MC	6	105.1	SG2D(W:25 O:2) + MC
12	5	93.9	SG (W:23) - NAS (3) - SG2D (W:21 O:1)	7	106.3	D (3) + RNV (35) + SG (W:11 O:3) + MC	7	105.6	VN + MC
13	8	56.2	SG1D (W:13 O:4) - SG (W:3) - AS	11	71.5	SG (W:5 O:4) + MC	10	64.5	SG1D(W:5 O:3) + AS
14	4	72.9	RNV(30) - NAS (3) - AS	11	95.1	Log Scaling	9	112.7	SG2D(W:25 O:3) + MC
15	4	84.9	PD (O:1) - NAS (3) - SG1D (W:5 O:3)	10	89.7	MC	9	89.3	SG1D(W:23 O:2) + AS
16	6	70.5	BO - Median Scaling - SG2D (W:23 O:4)	12	100.5	PS	9	108.4	AS + SG1D(W:25 O:3)
17	13	55.4	PD (O:1) - Norm (1) - RS	15	58.2	MC	15	56.3	SG1D(W:11 O:3) + SNV
18	4	88.2	AsLs (L:100 p:0.1) - SG (W:7) - AS	15	88.3	SG (W:9 O:4) + Log Scaling	15	88.1	MC + SG(W:3 O:2)
19	13	42.5	AS - NAS (3) - D (O:2)	15	51.0	SG (W:9 O:4) + PS	15	47.3	AS
20	3	90.4	RNV(10) - NAS (3) - SG1D (W:5 O:1)	13	91.0	SG (W:11 O:2) + Log Scaling	15	84.0	SG2D(W:19 O:2) + MMN
21	8	92.5	Min Scaling - NAS (3) - SG2D (W:19 O:2)	11	93.1	D (4) + MC	11	93.2	SG2D(W:7 O:2) + MC
22	12	22.2	AsLs (L:100 p:0.01) - NAS (2) - PS	15	22.7	SG (W:5 O:2) + MC	15	22.8	SG2D(W:9 O:3) + RNV (25)
23	12	48.2	AS - NAS (3) - SG (W:3)	15	56.2	D (2) + MC	15	54.4	SG2D(W:21 O:3) + MC
24	13	54.6	RS - NAS (3) - AS	15	57.9	SG (W:5 O:4) + AS	15	57.9	AS
25	13	37.9	Mean Scaling - NAS (3) - SG2D (W:3 O:1)	15	45.9	AS	15	46.6	RNV50
26	12	27.2	Norm (Inf) - SG2D (W:7 O:4) - NAS (3)	15	37.0	AS	14	37.4	SNV + MC

Case	This Work			Gerretzen et al.			Jiao et al.		
	LV	PEI (%)	Preprocessing Strategy	LV	PEI (%)	Preprocessing Strategy	LV	PEI (%)	Preprocessing Strategy
27	12	26.8	Norm (Inf) - SG2D (W:7 O:4) - NAS (3)	15	36.5	SG (W:5 O:4) + AS	15	37.6	VN + MC
28	5	83.2	RNV(40) - Norm (Inf) - NAS (2)	8	91.3	D (4) + Max Scaling + SG (W:11 O:4) + MC	15	109.5	MSC + SG1D(W:25 O:3) + MC
29	5	50.4	MC - NAS (3) - PD (O:1)	5	80.2	LB + RNV (35) + SG (W:5 O:4) + Poisson Scaling	13	69.5	SG1D(W:9 O:3) + AS
30	7	52.1	NAS (3) - AS - SG (W:13)	14	80.6	D (4) + RNV (25) + RS	15	66.3	AS
31	6	57.4	SG2D (W:13 O:1) - Norm (Inf) - AS	7	102.5	LB + BO + RNV (25) + SG (W:11 O:3) + MC	8	154.1	SG1D(W:25 O:1) + RNV (25)
32	5	74.4	PD (O:1) - Min Scaling - Norm (Inf)	11	101.5	RNV (25) + Log Scaling	11	104.1	MSC + SG(W:3 O:2) + MC
33	3	49.7	NAS (3) - Min Scaling - SG1D (W:15 O:2)	4	75.0	D (2) + MSC + MC	5	82.6	MSC + SG1D(W:19 O:3) + MC
34	4	78.0	NAS (3) - SG2D (W:11 O:3) - PS	6	133.8	D (3) + RNV (35) + Log Scaling	15	96.7	MSC + SG2D(W:9 O:3) + MC
35	8	80.3	LB - Norm (Inf) - Min Scaling	14	277.6	D (4) + Mean Scaling + MC	6	113.5	SG1D(W:3 O:1) + RNV (50)
36	1	61.3	NAS (3) - SG2D (W:23 O:1) - LS	15	187.7	D (4) + SG (W:5 O:4) + Level Scaling	15	86.2	SG2D(W:11 O:3) + AS
37	5	42.8	SNV - NAS (3) - LS	6	91.8	Mean Scaling + SG (W:5 O:4) + Level Scaling	4	81.4	MSC + SG2D(W:21 O:3) + MC
38	5	52.9	Norm (Inf) - SG2D (W:7 O:4) - Poisson Scaling	3	81.6	Max Scaling + SG (W:9 O:3) + MC	3	78.8	AN + SG(W:3 O:2) + MC
39	11	115.6	SG1D (W:23 O:2) - NAS (3) - Poisson Scaling	12	111.5	LB + SG (W:5 O:4) + MC	15	98.6	AN + SG2D(W:3 O:2) + MC
40	4	94.0	LB - NAS (3) - LS	3	118.6	D (2) + PS	6	123.3	SG(W:15 O:3) + MC
41	7	108.9	NAS (3) - SG (W:3) - LS	11	130.9	D (3) + SG (W:11 O:3) + MC	9	123.4	SG2D(W:23 O:2) + AS
42	7	108.9	NAS (3) - SG (W:3) - LS	11	130.9	D (3) + SG (W:11 O:3) + MC	9	123.4	SG2D(W:23 O:2) + AS
43	6	72.3	SG (W:13) - BO - Power Scaling	7	65.3	D (3) + RNV (15) + SG (W:5 O:4) + AS	8	78.1	SG(W:19 O:1) + MMN
44	7	73.2	AsLs (L:1000 p:0.01) - Norm (1) - NAS (2)	8	76.4	D (3) + RNV (25) + SG (W:11 O:3) + MC	11	78.0	SG2D(W:9 O:2) + RNV (50)
45	8	54.0	PS - NAS (3) - MC	13	51.9	MSC + Poisson Scaling	13	48.9	MSC + SG(W:15 O:3) + MC
46	7	51.2	Median Scaling - NAS (3) - RS	10	66.5	D (3) + RNV (35) + SG (W:5 O:4) + PS	15	73.1	MSC + SG1D(W:25 O:2) + MC
47	8	74.8	Norm (2) - AS - NAS (2)	13	84.2	LB + BO + RNV (25) + SG (W:11 O:4) + Poisson Scaling	11	85.3	AsLs
48	7	113.0	AsLs (L:100000000 p:0.1) - Norm (1) - PS	11	143.8	D (4) + Poisson Scaling	8	115.9	SG1D(W:23 O:1) + RNV (25)
49	4	92.8	RNV(30) - NAS (3) - D (O:2)	14	108.5	D (2) + Mean Scaling + SG (W:11 O:4) + PS	11	102.7	AsLs + SNV
50	4	94.5	NAS (1) - BO - Min Scaling	4	98.7	D (2) + RNV (15) + MC	6	95.9	SG2D(W:25 O:3) + AN
51	5	72.2	Norm (1) - AsLs (L:1000000 p:0.01) - PS	5	77.2	D (3) + RNV (25) + Poisson Scaling	14	109.8	AN + SG2D(W:3 O:2) + MC
52	5	90.3	NAS (2) - AsLs (L:100000000 p:0.1) - PS	6	103.4	D (4) + RNV (25) + SG (W:5 O:4) + AS	6	112.2	SG1D(W:25 O:1) + MMN
53	3	81.2	Norm (1) - Power Scaling - NAS (1)	6	69.6	D (2) + RNV (15) + SG (W:11 O:3) + MC	15	94.4	AN + SG2D(W:13 O:3) + MC
54	3	81.2	Norm (1) - Power Scaling - NAS (1)	6	69.6	D (2) + RNV (15) + MC	15	94.4	AN + SG2D(W:13 O:2) + MC

Case	This Work			Gerretzen et al.			Jiao et al.		
	LV	PEI (%)	Preprocessing Strategy	LV	PEI (%)	Preprocessing Strategy	LV	PEI (%)	Preprocessing Strategy
55	7	79.8	NAS (2) - LB - SNV	10	90.4	Log Scaling	11	91.7	AS
56	10	90.2	Min Scaling - D (O:3) - RNV(40)	14	103.5	D (4) + MC	12	99.5	SG2D(W:11 O:3) + MC
57	4	98.0	SG (W:7) - Poisson Scaling - NAS (3)	13	95.9	D (4) + Poisson Scaling	12	97.5	MC + SG1D(W:15 O:3)
58	5	88.5	MC - SG (W:13) - NAS (2)	10	99.2	Poisson Scaling	11	97.8	AsLs + SG(W:15 O:3)
59	6	81.8	AsLs (L:1000 p:0.1) - SG (W:13) - Mean Scaling	13	86.4	D (3) + RNV (35) + MC	11	75.3	SG(W:23 O:3) + AN + MC
60	2	95.5	NAS (1) - AsLs (L:10000000 p:0.01) - Poisson Scaling	3	99.2	D (3) + RNV (25) + MC	9	92.5	SG2D(W:21 O:3) + AS
61	6	81.0	Norm (2) - SG2D (W:19 O:1) - AS	14	83.1	D (4) + RNV (35) + MC	15	75.6	SG2D(W:25 O:2) + AS
62	4	91.7	Norm (2) - Min Scaling - Power Scaling	6	94.1	D (4) + RNV (25) + MC	11	94.2	SG1D(W:21 O:3) + AsLs
63	2	96.6	NAS (1) - AsLs (L:1000000 p:0.1) - PS	13	95.7	D (2) + MC	9	93.0	MSC + SG1D(W:25 O:1) + MC
64	5	94.0	NAS (1) - AsLs (L:1000000 p:0.1) - RS	13	80.6	D (4) + RNV (25) + MC	14	80.9	SG2D(W:25 O:2) + AS
65	5	71.5	AsLs (L:1000 p:0.001) - SG (W:13) - NAS (3)	6	91.4	D (4) + SG (W:11 O:3) + MC	15	76.6	MC + SG2D(W:21 O:3)
66	5	51.0	MC - NAS (3) - SG (W:21)	7	70.3	MC	15	80.8	SG1D(W:15 O:2) + MC
67	9	100.0	PD + NAS (3) + AS	15	95.8	MC	14	95.9	MC

AN: area normalization; AS: autoscale; AsLs: asymmetric least squares; BO: baseline offset; D: detrend; LB: linear baseline; LS: level scaling; MC: mean center; MSC: multiplicative scatter correction; NAS: net analytical signal; Norm: normalize; O: derivative order; PD: pairwise detrend; PS: pareto scaling; RNV: robust normal variate; RS: range scaling; SG: smoothing; SG1D: first derivative; SG2D; second derivative; SNV: standard normal variate; W: derivative window; VN: vector normalization; MMN: Maximum Minimum Normalization; PEI: Prediction Error Index.

The optimization method presented performed well for most calibration cases studied, including different data sources, such as fluorescence, gas chromatography, Vis-NIR, NIR, MID, Raman, NMR, and voltammetry. Figure I-8 shows the computation time for all methods previously discussed for datasets of different sizes. The computation times were measured using an Intel Core i5-3317U CPU @ 1.70GHz 8GB RAM. Regarding the computation time, overall, the proposed method performed similar to the other two methodologies, however, for larger datasets, a dataset containing 655 samples and 650 variables, for example, the PPO algorithm took under 30 min to perform the search, which is very satisfactory, outperforming the other two methodologies. Although the methodologies presented by Gerretzen et al.[16] and Jiao et al.[18] try out fewer preprocessing strategies, there are other particularities on each algorithm that increases the executing time, such as the number of iterations or the way the code was written, for example.

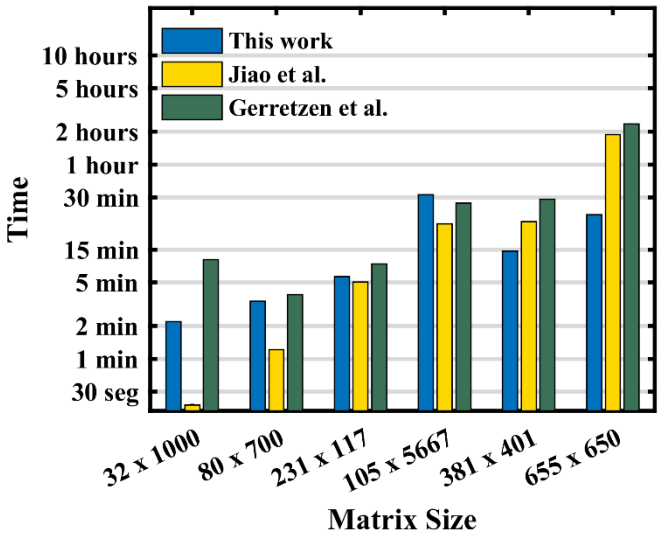


Figure I-8. Computation times for different matrix sizes (samples x variables) using the proposed method, Gerretzen et al.[16] and Jiao et al.[18] methods.

The proposed algorithm is robust as it can evaluate how much preprocessing improves the model. The strategy redefines the trial-and-error strategy by performing several combinations in a suitable time. Some combinations are unlikely, but the algorithm is objective, seeking the smallest RMSECV, and therefore, RMSEP values. The added value of this approach is furthermore demonstrated on a repository of 67 calibration cases, these consistently show that optimizing the hyperparameters a priori did not show significative improvement when compared with using the default hyperparameters' values, not fixing the sequence in which the preprocessing methods are applied increased the average accuracy in 4.2%, considering the combination of more than one preprocessing method increased the average accuracy in over

10%, and searching for the best preprocessing strategy in a larger space resulted in the best models.

3.5 - GRAPHICAL USER INTERFACE

A graphical user interface (GUI) was build using Matlab App Designer package. Creating a GUI for the preprocessing optimization code can offer several advantages. Firstly, it can make the optimization process more user-friendly by allowing users to interact with the code visually. This can reduce the potential errors caused by manually entering code. Ultimately, implementing a GUI can improve efficiency, accuracy, and usability, making it an essential tool for the end users. Figure I-9 show the GUI for preprocessing optimization.

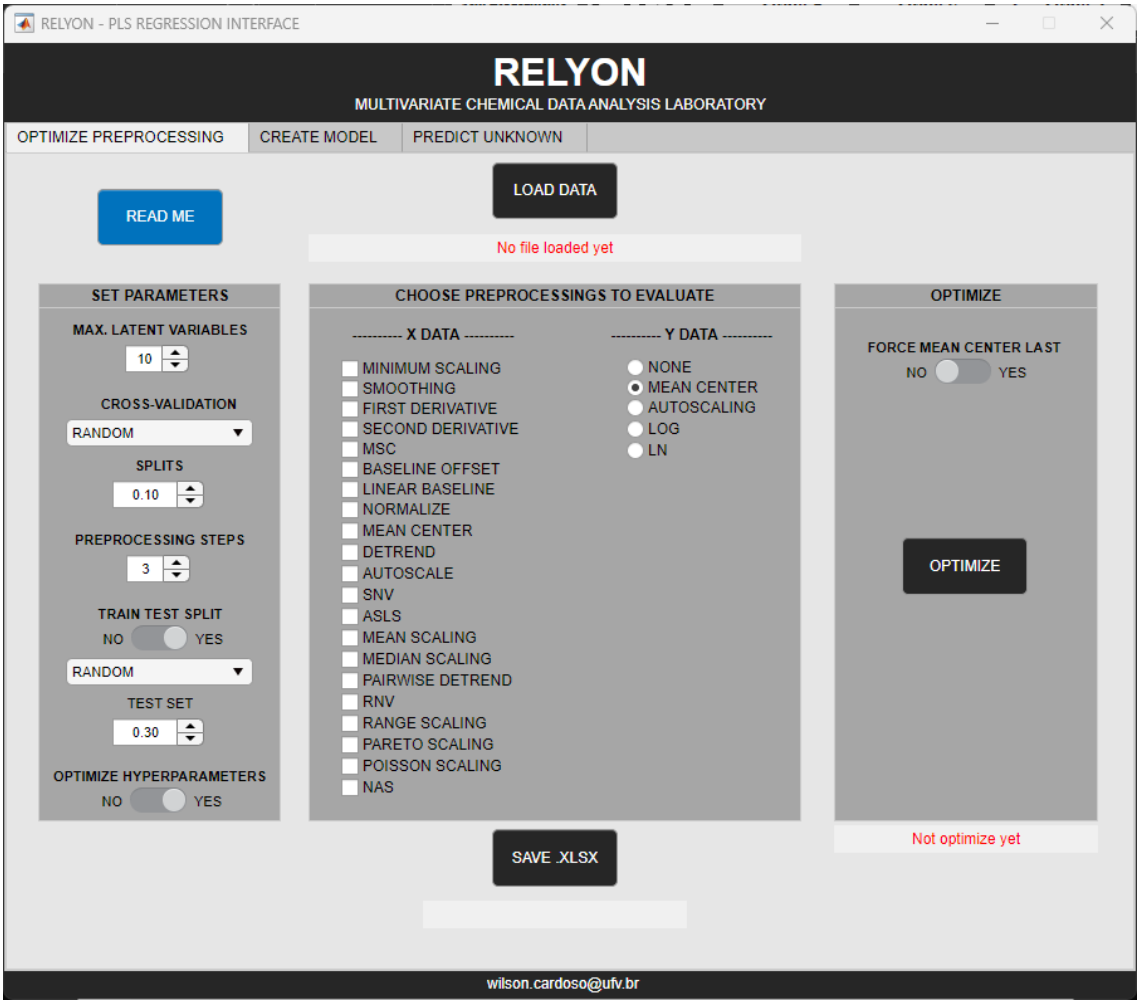


Figure I-9. Graphical User Interface (GUI) for the preprocessing optimization.

The input for GUI can be .mat or a .xlsx file. If the file is a .mat file, the independent variable must be named Spectra, the dependent variables must be named Lab, and the names of the samples must be named Samples, as shown in Figure I-10.

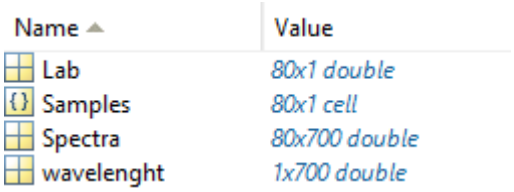


Figure I-10. Variables names for .mat file.

If the file is a .xlsx file the independent variable must be on a sheet named Spectra and the dependent variables must be on a sheet named Lab, as shown in Figure I-11. Both variables must be on the same order based on the sample named. For either .mat or .xlsx file the data must be cleaned up a prior, i.e., not numbers (nan), infinite (-inf +inf), and blank cells must be removed.

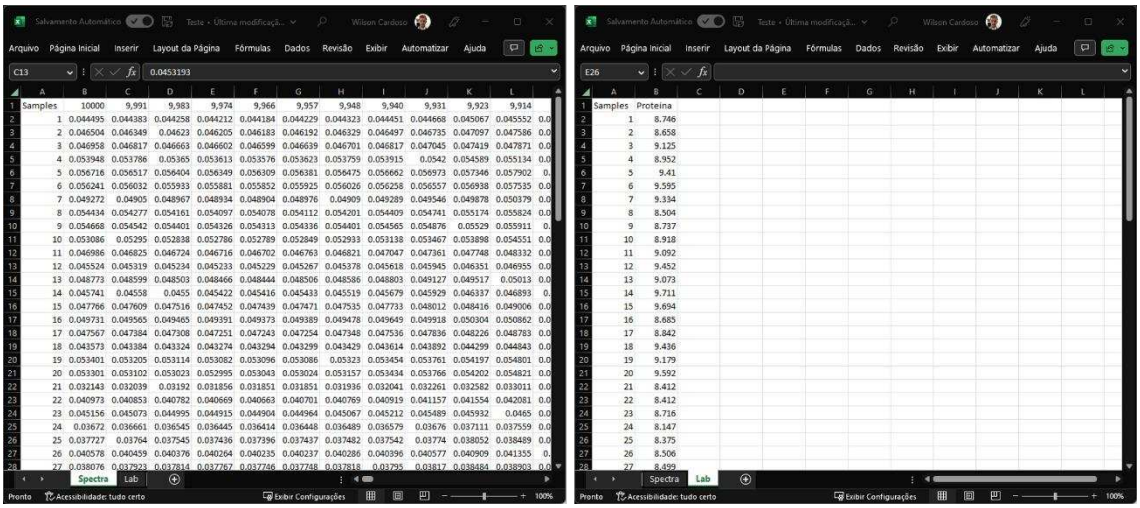
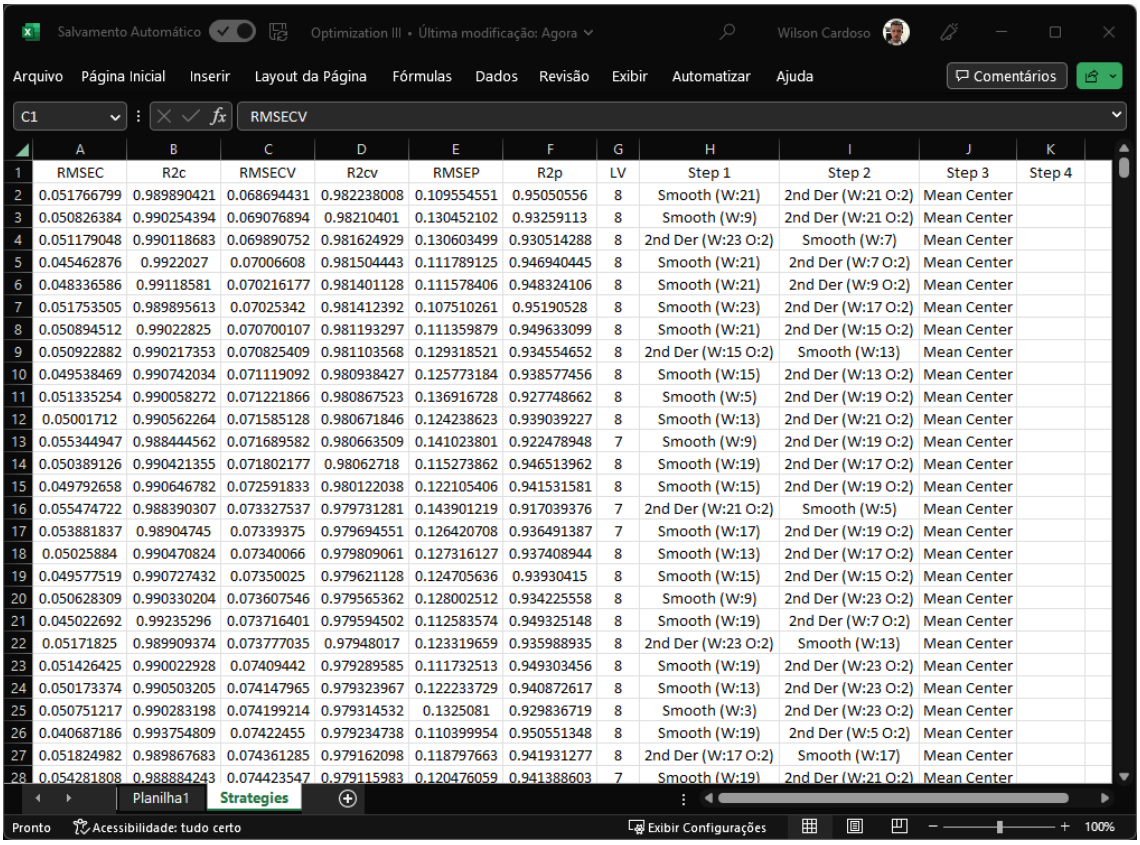


Figure I-11. Data arrangement for .xlsx file

After setting the parameters and optimizing the preprocessing strategy, the result can be saved in a .xlsx file, as show in Figure I-12. The spreadsheet contains all preprocessing strategies assessed and the model parameters, the user can choose which preprocessing strategy is most fitted based on the RMSECV values and the preprocessing in the strategy. Once the preprocessing strategy has been chosen, the user can build a model for the chosen preprocessing using the tab ‘Create Model’ and later use the model to predict unknown samples using the ‘Predict Unknown’ tab.



	A	B	C	D	E	F	G	H	I	J	K
1	RMSEC	R2c	RMSECV	R2cv	RMSEP	R2p	LV	Step 1	Step 2	Step 3	Step 4
2	0.051766799	0.989890421	0.068694431	0.982238008	0.109554551	0.95050556	8	Smooth (W:21)	2nd Der (W:21 O:2)	Mean Center	
3	0.050826384	0.990254394	0.069076894	0.98210401	0.130452102	0.93259113	8	Smooth (W:9)	2nd Der (W:21 O:2)	Mean Center	
4	0.051179048	0.990118683	0.069890752	0.981624929	0.130603499	0.930514288	8	2nd Der (W:23 O:2)	Smooth (W:7)	Mean Center	
5	0.045462876	0.9922027	0.07006608	0.981504443	0.111789125	0.946940445	8	Smooth (W:21)	2nd Der (W:7 O:2)	Mean Center	
6	0.048336586	0.99118581	0.070216177	0.981401128	0.111578406	0.948324106	8	Smooth (W:21)	2nd Der (W:9 O:2)	Mean Center	
7	0.051753505	0.989895613	0.07025342	0.981412392	0.107510261	0.95190528	8	Smooth (W:23)	2nd Der (W:17 O:2)	Mean Center	
8	0.050894512	0.99022825	0.070700107	0.981193297	0.111359879	0.949633099	8	Smooth (W:21)	2nd Der (W:15 O:2)	Mean Center	
9	0.050922882	0.990217353	0.070825409	0.981103568	0.129318521	0.934554652	8	2nd Der (W:15 O:2)	Smooth (W:13)	Mean Center	
10	0.049538469	0.990742034	0.071119092	0.980938427	0.125773184	0.938577456	8	Smooth (W:15)	2nd Der (W:13 O:2)	Mean Center	
11	0.051335254	0.990058272	0.071221866	0.980867523	0.136916728	0.927748662	8	Smooth (W:5)	2nd Der (W:19 O:2)	Mean Center	
12	0.05001712	0.990562264	0.071585128	0.980671846	0.124238623	0.939039227	8	Smooth (W:13)	2nd Der (W:21 O:2)	Mean Center	
13	0.055344947	0.988444562	0.071689582	0.980663509	0.141023801	0.922478948	7	Smooth (W:9)	2nd Der (W:19 O:2)	Mean Center	
14	0.050389126	0.990421355	0.071802177	0.98062718	0.115273862	0.946513962	8	Smooth (W:19)	2nd Der (W:17 O:2)	Mean Center	
15	0.049792658	0.990646782	0.072591833	0.980122038	0.122105406	0.941531581	8	Smooth (W:15)	2nd Der (W:19 O:2)	Mean Center	
16	0.055474722	0.988390307	0.073327537	0.979731281	0.143901219	0.917039376	7	2nd Der (W:21 O:2)	Smooth (W:5)	Mean Center	
17	0.053881837	0.98904745	0.07339375	0.979694551	0.126420708	0.936491387	7	Smooth (W:17)	2nd Der (W:19 O:2)	Mean Center	
18	0.05025884	0.990470824	0.07340066	0.979809061	0.127316127	0.937408944	8	Smooth (W:13)	2nd Der (W:17 O:2)	Mean Center	
19	0.049577519	0.990727432	0.07350025	0.979621128	0.124705636	0.93930415	8	Smooth (W:15)	2nd Der (W:15 O:2)	Mean Center	
20	0.050628309	0.990330204	0.073607546	0.979565362	0.128002512	0.934225558	8	Smooth (W:9)	2nd Der (W:23 O:2)	Mean Center	
21	0.045022692	0.99235296	0.073716401	0.979594502	0.112583574	0.949325148	8	Smooth (W:19)	2nd Der (W:7 O:2)	Mean Center	
22	0.05171825	0.989909374	0.073777035	0.97948017	0.123319659	0.935988935	8	2nd Der (W:23 O:2)	Smooth (W:13)	Mean Center	
23	0.051426425	0.990022928	0.07409442	0.979289585	0.111732513	0.949303456	8	Smooth (W:19)	2nd Der (W:23 O:2)	Mean Center	
24	0.050173374	0.990503205	0.074147965	0.979323967	0.122233729	0.940872617	8	Smooth (W:13)	2nd Der (W:23 O:2)	Mean Center	
25	0.050751217	0.990283198	0.074199214	0.979314532	0.1325081	0.929836719	8	Smooth (W:3)	2nd Der (W:23 O:2)	Mean Center	
26	0.040687186	0.993754809	0.07422455	0.979234738	0.110399954	0.950551348	8	Smooth (W:19)	2nd Der (W:5 O:2)	Mean Center	
27	0.051824982	0.989867683	0.074361285	0.979162098	0.118797663	0.941931277	8	2nd Der (W:17 O:2)	Smooth (W:17)	Mean Center	
28	0.054281808	0.988884243	0.074423547	0.979115983	0.120476059	0.941388603	7	Smooth (W:19)	2nd Der (W:21 O:2)	Mean Center	

Figure I-12. Excel spreadsheet for the preprocessing optimization.

4 - CONCLUSION

This work presented a new approach for the optimization of data preprocessing methodologies and their hyperparameters. The results obtained by this new strategy outperformed the other evaluated methods. The methodology proposed in this work resulted in the smallest RMSEP values for 48 out of the 67 calibration cases, while the other two optimization strategies (from literature) resulted in the smallest RMSEP for 20 calibration cases combined. In addition, the proposed method resulted in more parsimonious models for 62 calibration cases. This study highlighted that not fixing the sequence which the preprocessing methods are applied and combining the preprocessing methods available were essential to achieve the best preprocessing strategy. Furthermore, the proposed method has comparable computation time when compared to the preprocessing optimization methods already presented in the literature, reducing the possibility of obtaining a suboptimal preprocessing strategy. The proposed method is universal and can be applied to datasets from different spectroscopies and other analytical method.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (Project: CEX - APQ-02254-15). We are thankful to Prof. Jeroen Jansen and Dr. Sin Yong Teng from the Department of Analytical Chemistry & Chemometrics at Radboud University, Nijmegen, the Netherlands for welcoming me and helping me in the development of this work.

REFERENCES

(1) CONAB. Companhia Nacional de Abastecimento. *Acompanhamento da Safra Brasileira: Cana-de-açúcar* **2019**, 58.

(2) Zabed, H.; Faruq, G.; Sahu, J. N.; Azirun, M. S.; Hashim, R.; Nasrulhaq Boyce, A. Bioethanol Production from Fermentable Sugar Juice. *The Scientific World Journal* **2014**, *2014*, 1–11. https://doi.org/10.1155/2014/957102.

(3) Cavalett, O.; Junqueira, T. L.; Dias, M. O. S.; Jesus, C. D. F.; Mantelatto, P. E.; Cunha, M. P.; Franco, H. C. J.; Cardoso, T. F.; Maciel Filho, R.; Rossell, C. E. V.; Bonomi, A. Environmental and Economic Assessment of Sugarcane First Generation Biorefineries in Brazil. *Clean Technol Environ Policy* **2012**, *14* (3), 399–410. https://doi.org/10.1007/s10098-011-0424-7.

(4) Silva, L. A.; Gasparini, K.; Assis, C.; Ramos, R.; Kist, V.; Barbosa, M. H. P.; Teófilo, R. F.; Bhering, L. L. Selection Strategy for Indication of Crosses between Potential Sugarcane Genotypes Aiming at the Production of Bioenergy. *Ind Crops Prod* **2017**, *104* (December 2016), 62–67. https://doi.org/10.1016/j.indcrop.2017.04.025.

(5) Kandel, R.; Yang, X.; Song, J.; Wang, J. Potentials, Challenges, and Genetic and Genomic Resources for Sugarcane Biomass Improvement. *Front Plant Sci* **2018**, *9* (February), 1–14. https://doi.org/10.3389/fpls.2018.00151.

(6) Barbosa, M. H. P.; Resende, M. D. V.; Dias, L. A. dos S.; Barbosa, G. V. de S.; Oliveira, R. A. de; Peternelli, L. A.; Daros, E. Genetic Improvement of Sugar Cane for Bioenergy: The Brazilian Experience in Network Research with RIDESA. *Crop Breeding and Applied Biotechnology* **2012**, *12* (spe), 87–98. https://doi.org/10.1590/S1984-70332012000500010.

(7) Lavanholi, M. G. D. Qualidade Da Cana-de-Açúcar Como Matéria-Prima Para Produção de Açúcar e Álcool. In *Cana-de-açúcar*; Instituto Agronômico, 2010; p 882. https://doi.org/doi.org/10.3738/nucleus.v5i2.102.

(8) Fernandes, A. C. Cálculos Na Agroindústria Da Cana-de-Açúcar. *STAB: Piracicaba, SP, Brasil* **2003**, No. 2º.

(9) Tai, P. Y. P.; Miller, J. D. Germplasm Diversity among Four Sugarcane Species for Sugar Composition. *Crop Sci* **2002**, *42* (3), 958–964. https://doi.org/10.2135/cropsci2002.9580.

(10) Rein, P.; others. *Cane Sugar Engineering.*; Verlag Dr. Albert Bartens KG, 2016.

(11) Al-Mhanna, N. M.; Huebner, H.; Buchholz, R. Analysis of the Sugar Content in Food Products by Using Gas Chromatography Mass Spectrometry and Enzymatic Methods. *Foods* **2018**, *7* (11). https://doi.org/10.3390/foods7110185.

(12) Zhao, D.; MacKown, C. T.; Starks, P. J.; Kindiger, B. K. Rapid Analysis of Nonstructural Carbohydrate Components in Grass Forage Using Microplate Enzymatic Assays. *Crop Sci* **2010**, *50* (4), 1537–1545. https://doi.org/10.2135/cropsci2009.09.0521.

(13) Liu, W.; Wu, H.; Li, B.; Dong, C.; Choi, M. M. F.; Shuang, S. Immobilization of Platinum Nanoparticles and Glucose Oxidase on Eggshell Membrane for Glucose Detection. *Analytical Methods* **2013**, *5* (19), 5154–5160. https://doi.org/10.1039/c3ay40327k.

(14) Teixeira, A. I.; Ribeiro, L. F.; Rezende, S. T.; Barros, E. G.; Moreira, M. A. Development of a Method to Quantify Sucrose in Soybean Grains. *Food Chem* **2012**, *130* (4), 1134–1136. https://doi.org/10.1016/j.foodchem.2011.07.128.

(15) Filip, M.; 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. *Food Chem* **2016**, *199*, 653–659. https://doi.org/10.1016/j.foodchem.2015.12.060.

(16) Simeone, M. L. F.; Parrella, R. A. C.; Schaffert, R. E.; Damasceno, C. M. B.; Leal, M. C. B.; Pasquini, C. Near Infrared Spectroscopy Determination of Sucrose, Glucose and Fructose in Sweet Sorghum Juice. *Microchemical Journal* **2017**, *134*, 125–130. https://doi.org/10.1016/j.microc.2017.05.020.

(17) Shanmugavelan, P.; Kim, S. Y.; Kim, J. B.; Kim, H. W.; Cho, S. M.; Kim, S. N.; Kim, S. Y.; Cho, Y. S.; Kim, H. R. Evaluation of Sugar Content and Composition in Commonly Consumed Korean Vegetables, Fruits, Cereals, Seed Plants, and Leaves by HPLC-ELSD. *Carbohydr Res* **2013**, *380*, 112–117. https://doi.org/10.1016/j.carres.2013.06.024.

(18) 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* **2014**, *145*, 784–788. https://doi.org/10.1016/j.foodchem.2013.08.135.

(19) Montesano, D.; Cossignani, L.; Giua, L.; Urbani, E.; Simonetti, M. S.; Blasi, F. A Simple HPLC-ELSD Method for Sugar Analysis in Goji Berry. *J Chem* **2016**, *2016*, 1–5. https://doi.org/10.1155/2016/6271808.

(20) Dugalic, K.; Sudar, R.; Viljevac, M.; Josipovic, M.; Cupic, T. Sorbitol and Sugar Composition in Plum Fruits Influenced by Climatic Conditions. *Journal of Agricultural Science and Technology* **2014**, *16* (5), 1145–1155.

(21) Moral, A.; Hernández, M. D.; Tijero, A.; González, Z.; García, J.; De La Torre, M. J. NIRS Determination of Carbohydrates from Hydrothermal-Treated Rice Straw. *Tappi J* **2012**, *11* (4), 27–32. https://doi.org/10.32964/tj11.4.27.

(22) García Asprilla, I. D. S.; Ramírez-Navas, J. S. Near-Infrared Spectroscopy: A Rapid Alternative Technique to Reducing Sugars Determination in Juice of Sugarcane (Saccharum Officinarum L.). *J Pharm Pharmacogn Res* **2018**, *6* (5), 392–401.

(23) Rácz, A.; Héberger, K.; Fodor, M. Quantitative Determination and Classification of Energy Drinks Using Near-Infrared Spectroscopy. *Anal Bioanal Chem* **2016**, *408* (23), 6403–6411. https://doi.org/10.1007/s00216-016-9757-8.

(24) Caporaso, N.; Whitworth, M. B.; Grebby, S.; Fisk, I. D. Non-Destructive Analysis of Sucrose, Caffeine and Trigonelline on Single Green Coffee Beans by Hyperspectral Imaging. *Food Research International* **2018**, *106* (November 2017), 193–203. https://doi.org/10.1016/j.foodres.2017.12.031.

(25) Ilaslan, K.; Boyaci, I. H.; Topcu, A. Rapid Analysis of Glucose, Fructose and Sucrose Contents of Commercial Soft Drinks Using Raman Spectroscopy. *Food Control* **2015**, *48*, 56–61. https://doi.org/10.1016/j.foodcont.2014.01.001.

(26) Özbalci, B.; Boyaci, I. H.; Topcu, A.; Kadilar, C.; Tamer, U. Rapid Analysis of Sugars in Honey by Processing Raman Spectrum Using Chemometric Methods and Artificial Neural Networks. *Food Chem* **2013**, *136* (3–4), 1444–1452. https://doi.org/10.1016/j.foodchem.2012.09.064.

(27) Huang, Y.; Carragher, J.; Cozzolino, D. Measurement of Fructose, Glucose, Maltose and Sucrose in Barley Malt Using Attenuated Total Reflectance Mid-Infrared Spectroscopy. *Food Anal Methods* **2016**, *9* (4), 1079–1085. https://doi.org/10.1007/s12161-015-0286-4.

(28) Leopold, L. F.; Leopold, N.; Diehl, H. A.; Socaciu, C. Quantification of Carbohydrates in Fruit Juices Using FTIR Spectroscopy and Multivariate Analysis. *Spectroscopy* **2011**, *26* (2), 93–104. https://doi.org/10.3233/SPE-2011-0529.

(29) Pasquini, C. Near Infrared Spectroscopy: A Mature Analytical Technique with New Perspectives – A Review. *Anal Chim Acta* **2018**, *1026*, 8–36. https://doi.org/10.1016/j.aca.2018.04.004.

(30) Pasquini, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. *J Braz Chem Soc* **2003**, *14* (2), 198–219. https://doi.org/10.1590/S0103-50532003000200006.

(31) Osborne, B. G. Near-Infrared Spectroscopy in Food Analysis. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons, Ltd: Chichester, UK, 2000; pp 1–14. https://doi.org/10.1002/9780470027318.a1018.

(32) Cen, H.; He, Y.; Huang, M. Measurement of Soluble Solids Contents and PH in Orange Juice Using Chemometrics and Vis-NIRS. *J Agric Food Chem* **2006**, *54* (20), 7437–7443. https://doi.org/10.1021/jf061689f.

(33) Liu, C.; Yang, S. X.; Deng, L. Determination of Internal Qualities of Newhall Navel Oranges Based on NIR Spectroscopy Using Machine Learning. *J Food Eng* **2015**, *161*, 16–23. https://doi.org/10.1016/j.jfoodeng.2015.03.022.

(34) Rodriguez-Saona, L. E.; Fry, F. S.; McLaughlin, M. A.; Calvey, E. M. Rapid Analysis of Sugars in Fruit Juices by FT-NIR Spectroscopy. *Carbohydr Res* **2001**, *336* (1), 63–74. https://doi.org/10.1016/S0008-6215(01)00244-0.

(35) Cayuela, J. A.; Weiland, C. Intact Orange Quality Prediction with Two Portable NIR Spectrometers. *Postharvest Biol Technol* **2010**, *58* (2), 113–120. https://doi.org/10.1016/j.postharvbio.2010.06.001.

(36) Oliveira-Folador, G.; Bicudo, M. de O.; de Andrade, E. F.; Renard, C. M. G. C.; Bureau, S.; de Castilhos, F. Quality Traits Prediction of the Passion Fruit Pulp Using NIR and MIR Spectroscopy. *Lwt* **2018**, *95* (April), 172–178. https://doi.org/10.1016/j.lwt.2018.04.078.

(37) Alamar, P. D.; Caramês, E. T. S.; Poppi, R. J.; Pallone, J. A. L. Quality Evaluation of Frozen Guava and Yellow Passion Fruit Pulps by NIR Spectroscopy and Chemometrics. *Food Research International* **2016**, *85*, 209–214. https://doi.org/10.1016/j.foodres.2016.04.027.

(38) Shao, Y.; He, Y. Nondestructive Measurement of the Internal Quality of Bayberry Juice Using Vis/NIR Spectroscopy. *J Food Eng* **2007**, *79* (3), 1015–1019. https://doi.org/10.1016/j.jfoodeng.2006.04.006.

(39) Taira, E.; Ueno, M.; Saengprachatanarug, K.; Kawamitsu, Y. Direct Sugar Content Analysis for Whole Stalk Sugarcane Using a Portable near Infrared Instrument. *J Near Infrared Spectrosc* **2013**, *21* (4), 281–287. https://doi.org/10.1255/jnirs.1064.

(40) Taira, E.; Ueno, M.; Kawamitsu, Y. Automated Quality Evaluation System for Net and Gross Sugarcane Samples Using near Infrared Spectroscopy. *J Near Infrared Spectrosc* **2010**, *18* (3), 209–215. https://doi.org/10.1255/jnirs.884.

(41) Tewari, J.; Mehrotra, R.; Irudayaraj, J. Direct near Infrared Analysis of Sugar Cane Clear Juice Using a Fibre-Optic Transmittance Probe. *J Near Infrared Spectrosc* **2003**, *11* (5), 351–356. https://doi.org/10.1255/jnirs.386.

(42) Sorol, N.; Arancibia, E.; Bortolato, S. A.; Olivieri, A. C. Visible/near Infrared-Partial Least-Squares Analysis of Brix in Sugar Cane Juice. *Chemometrics and Intelligent Laboratory Systems* **2010**, *102* (2), 100–109. https://doi.org/10.1016/j.chemolab.2010.04.009.

(43) 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* **2007**, *18* (2), 259–266. https://doi.org/10.1590/S0103-50532007000200003.

(44) Maraphum, K.; Chuan-Udom, S.; Saengprachatanarug, K.; Wongpichet, S.; Posom, J.; Phuphaphud, A.; Taira, E. Effect of Waxy Material and Measurement Position of a Sugarcane Stalk on the Rapid Determination of Pol Value Using a Portable near Infrared Instrument. *J Near Infrared Spectrosc* **2018**, *26* (5), 287–296. https://doi.org/10.1177/0967033518795810.

(45) Corrêdo, L. de P.; Maldaner, L. F.; Bazame, H. C.; Molin, J. P. Evaluation of Minimum Preparation Sampling Strategies for Sugarcane Quality Prediction by Vis-NIR Spectroscopy. *Sensors* **2021**, *21* (6), 2195. https://doi.org/10.3390/s21062195.

(46) Phetpan, K.; Udompetaikul, V.; Sirisomboon, P. An Online Visible and Near-Infrared Spectroscopic Technique for the Real-Time Evaluation of the Soluble Solids Content of Sugarcane Billets on an Elevator Conveyor. *Comput Electron Agric* **2018**, *154* (October), 460–466. https://doi.org/10.1016/j.compag.2018.09.033.

(47) DU, Y. P.; LIANG, Y. Z.; KASEMSUMRAN, S.; MARUO, K.; OZAKI, Y. Removal of Interference Signals Due to Water from in Vivo Near-Infrared (NIR) Spectra of Blood Glucose by Region Orthogonal Signal Correction (ROSC). *Analytical Sciences* **2004**, *20* (9), 1339–1345. https://doi.org/10.2116/analsci.20.1339.

(48) Devaux, M. F.; Bertrand, D.; Robert, P.; Qannari, M. Application of Principal Component Analysis on NIR Spectral Collection after Elimination of Interference by a Least-Squares Procedure. *Appl Spectrosc* **1988**, *42* (6), 1020–1023. https://doi.org/10.1366/0003702884430443.

(49) Chen, D.; Shao, X.; Hu, B.; Su, Q. A Background and Noise Elimination Method for Quantitative Calibration of near Infrared Spectra. *Anal Chim Acta* **2004**, *511* (1), 37–45. https://doi.org/10.1016/j.aca.2004.01.042.

(50) Rambla, F. J.; Garrigues, S.; Guardia, M. De. PLS-NIR Determination of Total Sugar , Glucose , Fructose and Sucrose in Aqueous Solutions of Fruit Juices. **1997**, *2670* (97).

(51) Golic, M.; Walsh, K.; Lawson, P. Short-Wavelength Near-Infrared Spectra of Sucrose, Glucose, and Fructose with Respect to Sugar Concentration and Temperature. *Appl Spectrosc* **2003**, *57* (2), 139–145. https://doi.org/10.1366/000370203321535033.

(52) Valderrama, P.; Braga, J. W. B.; Poppi, R. J. Estado Da Arte de Figuras de Mérito Em Calibração Multivariada. *Quim Nova* **2009**, *32* (5), 1278–1287. https://doi.org/10.1590/S0100-40422009000500034.

(53) Ortiz, M. C.; Sarabia, L. A.; Herrero, A.; Sánchez, M. S.; Sanz, M. B.; Rueda, M. E.; Giménez, D.; Meléndez, M. E. Capability of Detection of an Analytical Method Evaluating False Positive and False Negative (ISO 11843) with Partial Least Squares. *Chemometrics and Intelligent Laboratory Systems* **2003**, *69* (1–2), 21–33. https://doi.org/10.1016/S0169-7439(03)00110-2.

(54) Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C. C. Sorting Variables by Using Informative Vectors as a Strategy for Feature Selection in Multivariate Regression. *J Chemom* **2009**, *23* (1), 32–48. https://doi.org/10.1002/cem.1192.

(55) Roque, J. V.; Cardoso, W.; Peternelli, L. A.; Teófilo, R. F. Comprehensive New Approaches for Variable Selection Using Ordered Predictors Selection. *Anal Chim Acta* **2019**, *1075*, 57–70. https://doi.org/10.1016/j.aca.2019.05.039.

(56) Buijs, K.; Choppin, G. R. Near-Infrared Studies of the Structure of Water. I. Pure Water. *J Chem Phys* **1963**, *39* (8), 2035–2041. https://doi.org/10.1063/1.1734579.

(57) Segtnan, V. H.; Šašić, Š.; Isaksson, T.; Ozaki, Y. Studies on the Structure of Water Using Two-Dimensional Near-Infrared Correlation Spectroscopy and Principal Component Analysis. *Anal Chem* **2001**, *73* (13), 3153–3161. https://doi.org/10.1021/ac010102n.

(58) 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* **2019**, *24* (20), 3696. https://doi.org/10.3390/molecules24203696.

(59) Beganović, A.; Beć, K. B.; Grabska, J.; Stanzl, M. T.; Brunner, M. E.; Huck, C. W. Vibrational Coupling to Hydration Shell – Mechanism to Performance Enhancement of

Qualitative Analysis in NIR Spectroscopy of Carbohydrates in Aqueous Environment. *Spectrochim Acta A Mol Biomol Spectrosc* **2020**, *237*, 118359. https://doi.org/10.1016/j.saa.2020.118359.

CHAPTER II

Dehydration as a Tool to Improve Predictability of Sugarcane Juice Carbohydrates Using Near-Infrared Spectroscopy Based PLS Models

The contents of this chapter have been adapted from:

W.J. Cardoso, J.G.R. Gomes, J. V. Roque, M.H.P. Barbosa, R.F. Teófilo, Dehydration as a Tool to improve predictability of sugarcane juice carbohydrates using near-infrared spectroscopy based PLS models, Chemom. Intell. Lab. Syst. 220 (2022) 104459. https://doi.org/10.1016/j.chemolab.2021.104459.

ABSTRACT

The aim of this work was to study dehydration as a way to improve the prediction of sucrose, glucose, and fructose in sugarcane juice using near-infrared (NIR) spectroscopy and partial least squares (PLS) regression models. The temperature, time, and sample volume involved in the dehydration process were optimized using design of experiments. Six different sample supports were assessed, being the thick couche paper the best support. NIR spectra from liquid (LSJ) and dehydrated sugarcane juice (DSJ) were obtained. Sucrose, glucose, and fructose in LSJ were analyzed using high-performance liquid chromatography with an evaporative light scattering detector (HPLC-ELSD). Sucrose, glucose, and fructose ranged from 99.29 to 249.27 mg/mL, 5.96 to 14.94 mg/mL and 3.99 to 16.10 mg/mL. PLS models were built using the sugars content and NIR spectra collected from a benchtop and a portable instrument. Ordered predictors selection (OPS) was applied to select the most informative variable. The results indicated better predictions for all sugars using the DSJ for both instruments, being the benchtop statistically better than the portable instrument. On the benchtop instrument, the PLS-OPS models presented root mean square error of prediction (RMSEP) respectively for sucrose, glucose, and fructose 7.98, 0.82, and 1.00 mg/mL using the DSJ against 12.75, 1.00, and 1.35 mg/mL using the LSJ. For the portable instrument, the RMSEP were respectively 15.90, 1.18, and 1.65 mg/mL using DSJ against 23.23, 1.40, and 2.08 mg/mL using LSJ. To sum up, the dehydration approach showed to be a great technique to improve the predictability of PLS-OPS models for sugarcane juice sugars using NIR spectra by removing the water and concentrating the analytes.

Keywords: Sugars; Water Removal; Chemometrics; Multivariate Regression; Genetic Breeding;

1 - INTRODUCTION

Sugarcane (*Saccharum* spp.) stands out among the bioenergy crops, and it is one of the most important Brazilian commodities. 2019/2020 Brazil's production is estimated at 615.978,9 thousand tons of sugarcane, and most of its volume supplies the sugar and ethanol industry.[1–3] In addition to industrial production, research in sugarcane breeding plays an important role in the research for more productive varieties. As a result, the quality control of sugarcane is essential for decision-making both in the industry and breeding programs.[4–6]

The sugarcane is composed mainly of juice (86 to 92%), which is composed of water (75 to 82%) and soluble solids (18 to 25%). The soluble solids fraction comprises sugars (15 to 24%) and non-sugars (1 to 2.5%). The quality control of sugarcane is mainly based on its juice sugar content. Sucrose is the primary sugar present in the juice (14 to 24%), and it can be estimated through soluble solids content (Brix), measured by a refractometer. The apparent sugar content can also be estimated through polarimetry (Pol), which measures the light deviated to the right caused by dextrorotatory compounds. Sucrose and glucose are dextrorotatory compounds, while fructose is levorotatory.[7,8] Thus, due to the presence of glucose and fructose, and other soluble solids, these instruments do not indicate the exact sucrose content but an overestimated or underestimated value. However, as sucrose is the main component of the juice, the correlation between these parameters and actual sucrose content is still high. In addition, these parameters are fast and straightforward to measure, which justifies their use. The principal drawback of the Pol method is that the juice needs to be clarified using an aluminum-based compound to produce a more reliable measurement, eliminating interferences, which increases the analysis cost and time. Besides, Brix and Pol methods do not provide information about glucose and fructose, only sucrose.[9,10] Therefore, there is a demand for a fast, low-cost, selective, and reliable method for sugarcane quality control.

Enzymatic sugars quantification methods have been applied for date juice,[11] grass forage,[12] eggshell membranes,[13] and soybean seeds[14]. Another method widely applied is high-performance liquid chromatography (HPLC); it has been applied for apple leaf and peel,[15] sweet sorghum,[16] Korean vegetables, fruits, cereals, seed plants, leaves,[17] and fruits.[18–20] Despite being accurate, these methodologies are high-priced, time-consuming, require specific knowledge or training to be applied for routine analysis, making its application unfeasible in routine analysis for quality control of sugarcane. Given the previous, near-infrared spectroscopy (NIR) combined with chemometric methods can be an alternative to these methodologies.[21–24] Other spectroscopic methods combined with chemometrics have been applied to quantify

sugars, such as Raman in soft drinks and honey,[25,26] and mid-infrared spectroscopy (MIR) in barley malt and fruit juice. [27,28]

NIR is a fast, low-priced, and simple method, requiring little or no sample preparation. Its region (750 to 2500 nm) contains information related to the overtones and combinations bands of the fundamental vibrations of C-H, O-H, N-H, and S-H bonds, most present in every organic molecule. Alone, the NIR spectrum from complex samples is not highly informative, as it presents broad, highly overlapped bands. However, when allied to chemometrics methods, such as partial least squares regression (PLS), it can be used to create models capable of predicting properties of interest, such as sugar content. Several applications of NIR combined with chemometrics are described in the literature for different fields, mainly food control.[29–31] NIR has been applied to predict sugar content in orange juice[32–35], passion fruit pulp[36,37], bayberry juice[38]. Regarding sugarcane, NIR has been used to predict the apparent sucrose content (soluble solids or polarimetric reading) using stalks[39,40] and juice[41–43]. Although presenting good results, these methodologies do not provide information about other sugars, such as glucose and fructose, and are more susceptible to inaccuracies as they do not use standard solutions to calibrate the instrument.

Nowadays, emphasis has been given to the use of portable NIR instruments. The use of such instruments expanded the applications of NIR, as the technique could be applied on field. Typically, a portable NIR comprises a narrow region (1000 to 1700 nm) that corresponds mostly to overtones of C-H and O-H stretching. Portable NIR instruments has been successfully applied to measure sugarcane properties such as Pol, soluble solids, and fiber content.[44–46] One of the main advantages of NIR is that it require little or no sample preparation, however, some works demonstrated that modeling sugar content using the sugarcane juice NIR spectra lead to a better prediction than using the stalk NIR spectra.[39,45]

Despite the PLS capability of modeling interferents present in the sample, water interference in the NIR spectra is still a concern when water is the main constituent of the sample. The NIR spectrum of a water-rich sample is similar to the spectrum of pure water, making it difficult to access the overlapping information and thus make the models less predictive. Usually, these inferences are removed by selecting wavelength ranges or using preprocessing, but these methods are not always effective.[47–51]

This chapter aims to study the advantages and disadvantages of a dehydration step of sugarcane juice to improve the prediction of sucrose, glucose, and fructose content in sugarcane juice using NIR spectroscopy and PLS regression. The goal is to propose a more accurate

methodology as an alternative to the reference methods used nowadays in the genetic breeding sites.

To the best of our knowledge, it is the first time a dehydration method to concentrate sugarcane juice samples for NIR spectroscopy is presented.

2 - MATERIALS AND METHODS

2.1 - SAMPLING

For this work, 282 samples, comprising 22 different genotypes (RB027052, RB027058, RB037017, RB037223, RB047108, RB077000, RB077210, RB077223, RB077227, RB077300, RB087202, RB087234, RB087242, RB087251, RB087842, RB097203, RB097217, RB107262, RB117000, RB127825, RB867515, and RB966928), were harvested every month from April to November of 2019 from the Sugarcane Breeding Program of Federal University of Viçosa (PMGCA-UFV) and Inter-University Network for the Development of Sugarcane Industry (RIDESA). The germplasm bank is located at the Federal University of Viçosa (UFV) experimental zone, Viçosa city, state of Minas Gerais, Brazil. The sugarcane stalks were harvested and grinded. An aliquot of 500 g was pressed employing a hydraulic press, and the juice extracted was collected. Approximately 10 mL of juice were transferred to polypropylene tubes (Falcon®) and stored at minus 80 ºC for posterior use.

2.2 - SUPPORT ASSESSMENT

The dehydrating method consists of depositing an aliquot of the sugarcane juice on a support for the dehydrating step. Afterwards, obtaining the NIR spectrum of the dehydrated sample. Six different supports were evaluated considering: the material capillarity, homogeneity, and stability at temperatures above 50º C. Figure II-1 show the supports considered in this work. The supports were cut using a manual press cutting, show in Figure II-2, so the supports size was standardized to a circle with 20mm of diameter.

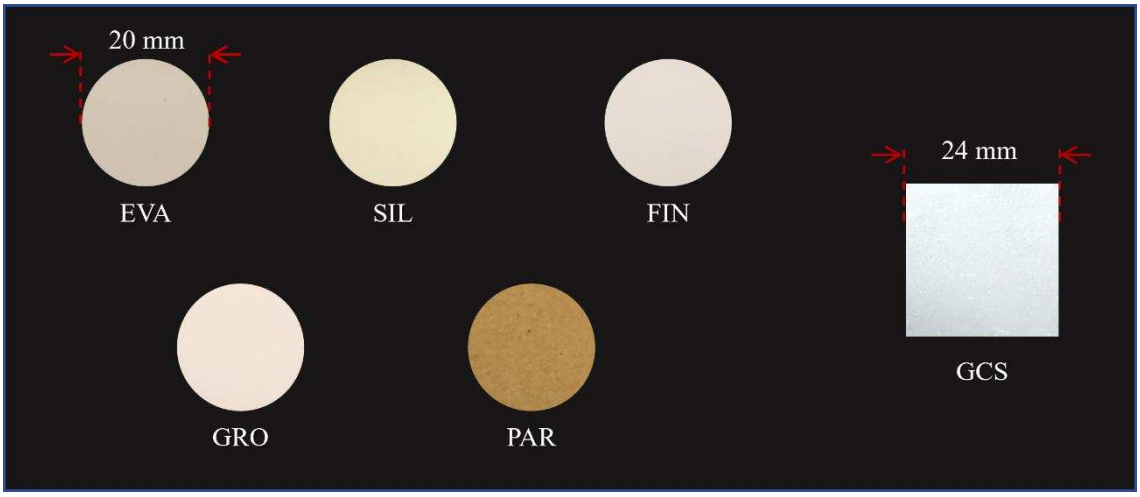


Figure II-1. Supports assessed in this study. Ethylene-vinyl acetate (EVA), silicone (SIL), thin couche paper (FIN), thick couche paper (GRO), thin couche paper (FIN), kraft paper (PAR), and glass slips cover (GSC).

Figure II-2. Manual press cutting.

2.3 - DEHYDRATION

The method consists of pipetting a specific sample volume on top of the support taken into the oven to dehydrate, as depicted in Figure II-3. A central composite design (CCD) was used to optimize the dehydration process. Three factors were studied: volume of sample, oven temperature, and dehydration time. Mass loss was chosen as the dependent variable. Three contents of sucrose in samples were chosen for this study, i.e., 93.64 (Sample I), 153.29 (Sample II), and 203.65 mg/L (Sample III). The objective was to verify and account for the influence of sucrose concentration in the dehydration process. The CCD factors and levels studied are showed in Table II-1.

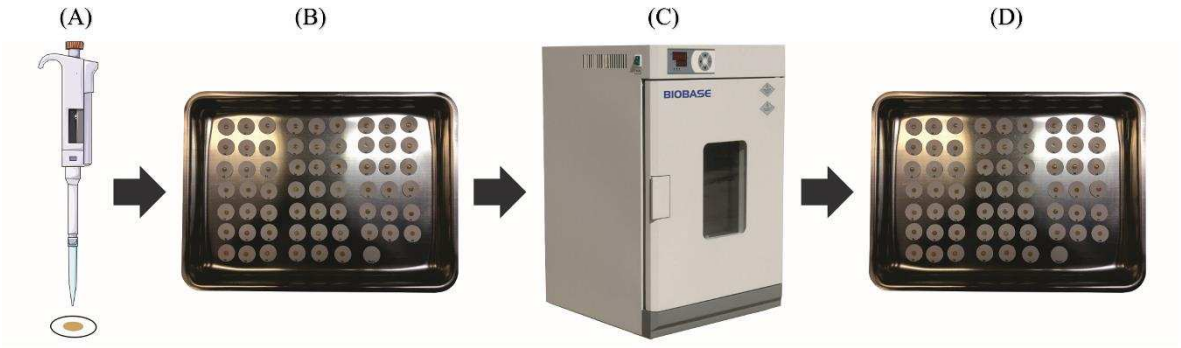


Figure II-3. Scheme of the dehydration method. An aliquot of the sugarcane juice is pipetted to the sample support (A), which is then placed on an aluminum tray (B), then taken to a preheated oven for a fixed time (C) and then the dehydrated samples are acquired (D).

Table II-1. Factors and levels used in the central composite design.

Factors	Levels				
	-α	-1	0	+1	+α
Temperature (ºC)	46.3	60	80	100	113.6
Time (min)	19.5	40	70	100	120.45
Volume (µL)	43.18	50	60	70	76.82

2.4 - REFERENCE ANALYSIS

A method based on high-performance liquid chromatography (HPLC) was used as a reference to quantify the sucrose, glucose, and fructose contents. The sugarcane juice samples were diluted 20 times with deionized water, homogenized, and then filtered through a nylon syringe filter (pore size 0.45 µm and 25 mm diameter) to vials. A Shimadzu 20AT Prominence instrument and the software LabSolutions were used. An RPM-Monosaccharide column (Phenomenex® Rezex) with 8 µm particle size, 300 × 7.8 mm was used. The mobile phase constituted 15% acetonitrile and 85% water. The separation parameters consisted of temperature at 80 ºC, the flow of 0.850 mL/min, and injection volume of 10 µL. An evaporative light scattering detector (ELSD) was used.[19] Standard calibration curves were built for sucrose, glucose, and fructose.

2.5 - NEAR-INFRARED SPECTROSCOPY

Spectra of the samples were obtained on two different instruments. The first instrument consisted of a benchtop Antaris II with Fourier Transform spectrometer with integration sphere (Thermo Scientific), ranging from 1000 to 2500 nm with an increment of 0.48 nm. The spectra were collected using the TQ Analysis software. The mean of 32 scans performed for each sample was stored. The second instrument consisted of a portable DLP® NIRscan Nano™

EVM (Texas Instruments), ranging from 900 to 1700 nm with increments of 1.32 nm, operating with the Hadamard scan method. The mean of 50 scans performed for each sample was stored. The spectra were acquired in two different ways, LSJ and DSJ, on both instruments. The first one, to collect the LSJ spectra, approximately 1 mL of sample was placed on a glass cup (internal diameter of 16.8 mm) with a transflectance accessory (diameter of 16 mm and an optical path of 1 mm). The second one, to collect the DSJ spectra, the sample-support was placed directly on the instrument window, the spectra were acquired in reflectance mode. For both instruments, the spectra were acquired as absorbance, 1/log(R), where R is either the reflectance or transflectance. Figure II-4 shows the benchtop and portables instruments.

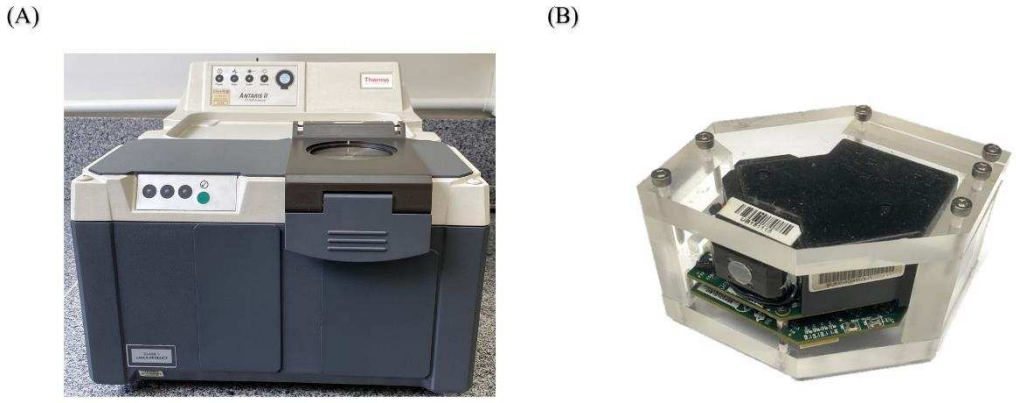


Figure II-4. Benchtop Antaris II (A) and portable DLP® NIRscan Nano™ EVM (B) instruments.

2.6 - MODELING

The NIR spectra were imported to MATLAB R2019a (MathWorks, Natick, USA). The PLS regression was used to build the models. All algorithms used were written at the Multivariate Chemical Data Analysis Laboratory (MCDA Lab). The samples were randomly divided into calibration set (70%) and external validation set (30%). All models have the same samples in the calibration and external validation set so that for each property, a comparison can be made. Different preprocessing methods were studied. Besides the raw data, one preprocessing, mean center, and seven transformations were assessed, smoothing, first derivative, second derivative, standard normal variate scaling, multiplicative scatter/signal correction, normalize, and baseline. The combinations of two and three preprocessing steps were also studied. Smoothing and first and second derivatives were applied using the odd widths of 7, 13, and 21. Multiplicative scatter correction was applied using the least-squares regression between each spectrum and the reference spectrum. Normalization was applied to normalize the data to a unit area (norm 1), to unit length (norm 2), and maximum value (norm infinite). Random cross-validation was used with 10 splits. Random cross-validation was

chosen so the same calibration and validations samples could be selected for all the models to facilitate comparison. In total, 2,032 different preprocessing combinations were applied. The models were built using latent variables from 1 to 8, which resulted in 16,256 different models. The algorithm presented in Chapter I was used.

The quality of the regression models was evaluated considering the root mean square error (RMSE), Eq. I-1, and correlation coefficient (R), Eq. I-2, of calibration (RMSEC, Rc), cross-validation (RMSECV, Rcv), and validation (RMSEP, Rp). The model parameters were optimized utilizing RMSECV, where the model with the lowest RMSECV was chosen. Outliers were removed considering the Student's residue and leverage and principal component analysis first and second scores.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad \text{(Eq. I-1)}$$

$$R = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2 (y_i - \bar{y})^2}} \quad \text{(Eq. I-2)}$$

where y is the reference value, $\bar{y}$ is the mean of the reference values, $\hat{y}$ is the predicted values, $\bar{\hat{y}}$ is the mean of the predicted values, and N is the number of samples.

Selectivity (SEL) shows the models capacity to determine an analyte without the interference of other compounds in the matrix and is calculated by Eq. I-3.[52]

$$SEL = \frac{\sum_{i=1}^{N}\left(\frac{nas_i}{\|x_i\|}\right)}{N} \quad \text{(Eq. I-3)}$$

where nas_i is the absolute value of the net analytical signal and $\|x_i\|$ represents the Euclidean norm of the instrument response vector, and N is the number of samples.

Limit of detection (LOD) is the minimum detectable value of concentration and is calculated by Eq. I-4.[53]

$$LOD = \frac{\Delta(\alpha,\beta) w_0 \hat{\sigma}}{\beta_0} \quad \text{(Eq. I-4)}$$

where $\Delta(\alpha,\beta)$ is a parameter of non-centrality as a function of the probability of type I error (α) and type II error (β), $w_0\hat{\sigma}$ is the standard deviation, and β_0 is the constant of the linear regression.

The inverse of analytical sensitivity (γ^{-1}) is the minimum difference of concentration that can be determined by the model and is calculated by Eq. I-5.[52]

$$\gamma^{-1} = \|\delta_x\| \times \|b\| \qquad \text{(Eq I-5)}$$

where $\|\delta_x\|$ is the Euclidean norm of the standard deviation of the reference signal and $\|b\|$ is the Euclidean norm of the regression coefficient.

Feature selection was performed using the new ordered predictors selection (OPS) algorithm (available at www.deq.ufv.br/chemometrics). It selects variables by sorting the variables according to informative vectors and systematically investigating the regression models to identify the most relevant set of variables. OPS was applied using all approaches available, all vectors were investigated, and the variables were searched using a window of 50 and an increment of 10. A more detailed explanation of the OPS algorithms can be found elsewhere.[54,55] The purpose is to find the best model possible for each property, so the particularities of each instrument, properties, and spectral analysis could be considered, and then compare the results.

Each model was verified for chance correlation. The dependent variables (y vector) were randomized 5.000 times and a model was built for each randomization and for the authentic y vector. If the Pearson correlation parameter of the authentic y vector is isolated from the randomized ones, then the model did not occur by chance.

2.7 - STATISTICAL ANALYSIS

An analysis of variance (ANOVA) was carried out to compare the best PLS models for sucrose, glucose, and fructose, respecting the assumptions of normality, homoscedasticity, and homogeneity of the variances. For each property, the models were compared based on the instrument (benchtop and portable instrument) and the spectral analysis (LSJ and DSJ). ANOVA was performed in the R environment (R Studio, 2020, R Development Core Team, 2020). The best models for each property, instrument and spectral analysis were randomly split 100 different times into calibration set (70%) and external validation set (30%) to perform the ANOVA. So, there are 100 different replications for each model. RMSEP was used to compare the models. The means were compared using the Tukey test with 95% confidence.

3 - RESULTS AND DISCUSSION

3.1 - SAMPLING

The 282 sugarcane samples were harvested every month from April to November of 2019. Thus, the variance due to the maturation of each genotype could be assessed. Table II-2 presents the descriptive statistics of the sucrose, glucose, and fructose that were quantified using HPLC.

Table II-2. Samples descriptive statistics.

Property	Min.*	Mean*	Max.*	STD	CV (%)
Sucrose	99.29	186.37	249.27	30.29	16.2
Glucose	5.96	9.57	14.94	1.88	19.6
Fructose	3.99	8.81	16.10	2.80	31.8

* mg/mL; STD: Standard Deviation; Min.: Minimum; Max.: Maximum, CV: coefficient of variation.

3.2 - SUPPORT ASSESSMENT

Six different supports were assessed: thick couche paper (GRO), thin couche paper (FIN), silicone (SIL), ethylene-vinyl acetate (EVA), kraft paper (PAR), and glass slips cover (GSC). EVA support was not stable at temperatures over 50 ºC. SIL support was too smooth to hold the sample. FIN absorbed the sample. GCS was not suitable as the sample spread over its surface. PAR and 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 novel dehydration methodology, as it did not absorb the sample and allowed a greater analytical frequency.

3.3 - DEHYDRATION

A CCD design was performed using as response the mass loss, which reads as the water removed from the sample. The analysis of variance showed that the quadratic regression adjusted the data with a coefficient of determination (R^2) of 0.85, 0.89, and 0.94 and pure error of $1x10^{-5}$, $9.5x10^{-6}$, and $1.4x10^{-5}$ for Sample I, II, and III, respectively. Figure II-5 presents the Pareto chart of the standardized effects for Sample I, II, and III. The most important effects for the three samples were the linear terms for temperature and volume and the quadratic term for temperature. Temperature effects were positive, which means that higher temperatures increased the mass loss for the samples. The volume effect was negative, which means that a lower volume increased the mass loss of the samples. The interaction between temperature and

volume was also significant, presenting a positive effect for Sample I and a negative effect for Sample II and III.

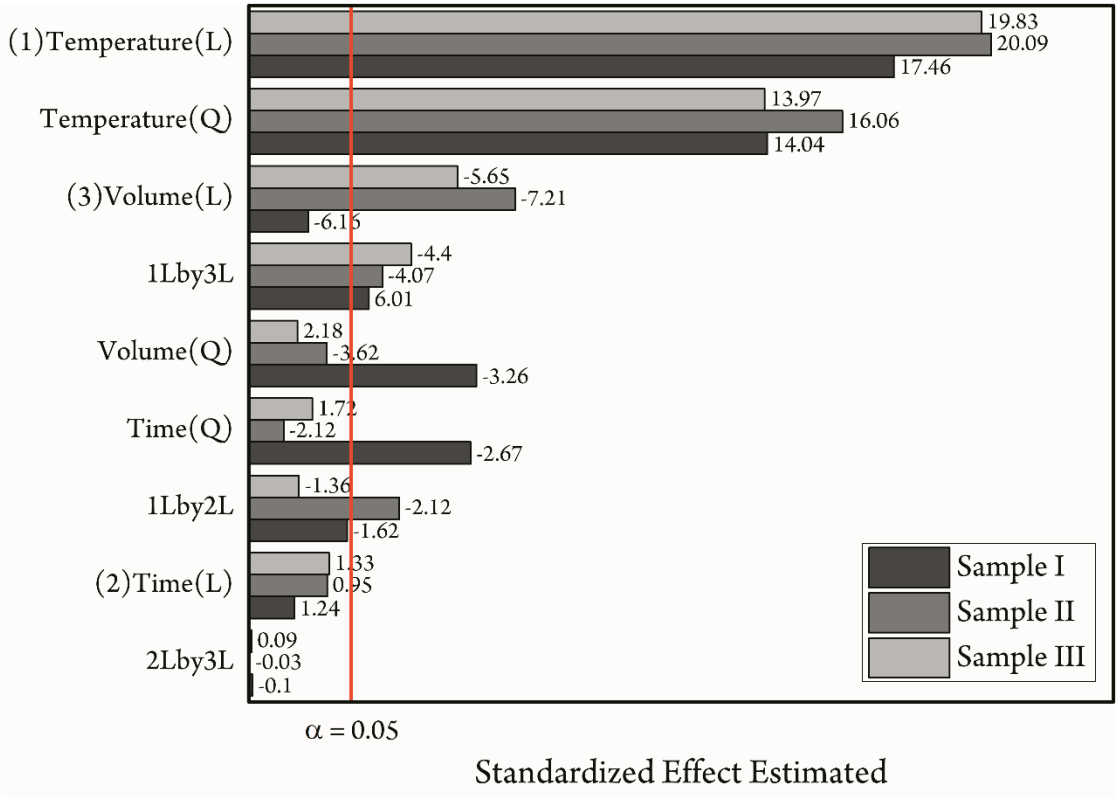


Figure II-5. Pareto chart of standardized effects for Sample I, II and III.

Figure II-6 presents the response surfaces as a function of temperature and volume for Samples I, II, and III. The time was fixed at 19.5 minutes. A desirability profile was made for each sample (I, II, and III) and the factors studied (temperature, time, and volume). It was observed that the greater the mass loss more significant the water removal. So, the maximum mass removal was set as high desirability. Therefore, for temperature, time, and volume, the optimal parameters were defined as 113.6 ºC, 19.5 min, and 51.6 µL, respectively.

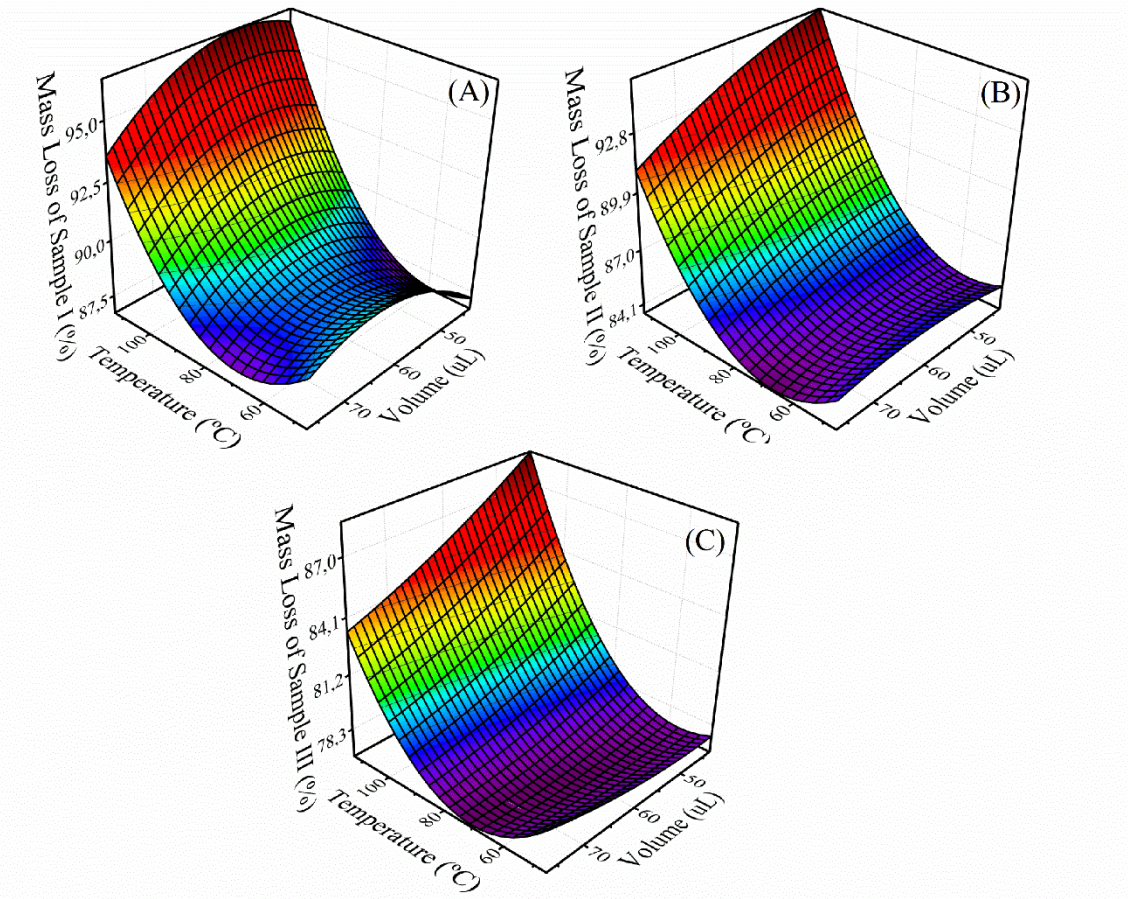


Figure II-6. Response surface of mass loss as a function of temperature and volume for Sample I (A), II (B), and (C). The time was fixed at 19.55 minutes.

3.4 - NEAR-INFRARED SPECTROSCOPY

Figure II-7 shows the NIR profiles acquired for the sugarcane juice, dehydrated juice, water, GRO support, and pure solid sucrose, glucose, and fructose. It can be observed that the spectra of the pure sugars (sucrose, glucose, and fructose) are similar, although presenting some differences around 1400 to 1600 nm and 2000 to 2300 nm. Considering the LSJ spectra, it can be observed that the spectra are similar to the water NIR spectra, as water is the main constituent of the juice and have a strong absorbance in the NIR. The pure sucrose, glucose, and fructose spectra and the LSJ spectra present very distinct profiles.

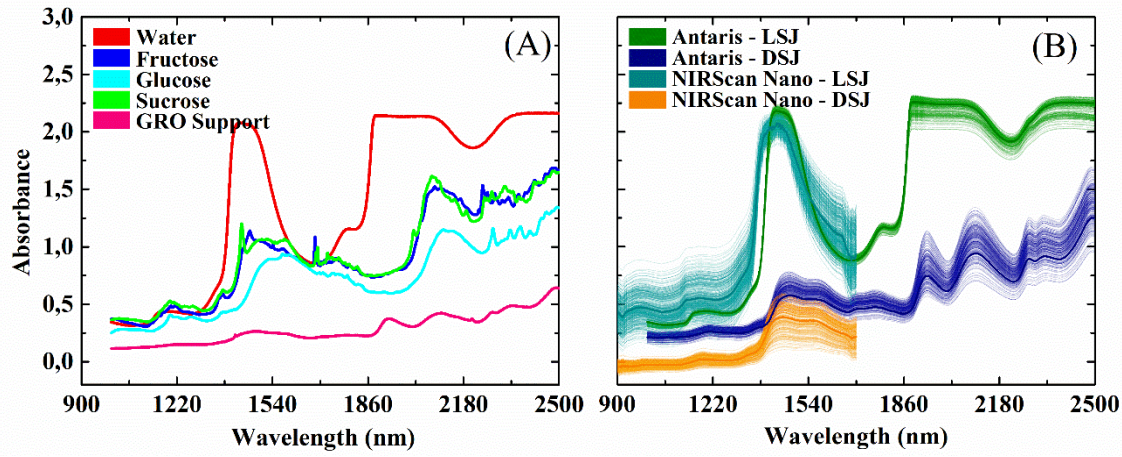


Figure II-7. Spectra of pure water, sucrose, glucose, fructose, and GRO support (A), spectra of the samples acquired using the LSJ and DSJ methods on Antaris and NIRScan Nano instrument (B).

On the other hand, the spectra of pure sugars and DSJ spectra present similar profiles, as observed in 1400 to 1700 nm and 2000 to 2300 nm regions. The dehydrated juice spectrum differs from the water spectrum, evidencing that the drying process removes the water influence. About the GRO support NIR spectrum, it can be observed that it has some similarities to the DSJ spectra, it may be due as its main constituent is cellulose, which resembles the sugars structures. It presents a low-intensity NIR profile compared with the DSJ spectra, as shown in Figure I-4; therefore, it does not interfere with the modeling process.

It can be observed in Figure II-7 that NIR spectra of the juice samples on both instruments present a strong absorbance band between 1300 and 1500 nm due to O-H stretch bands of water. In addition, saturated bands can be seen in the Antaris NIR spectra from 1800 to 2200 nm and 2300 to 2500 nm due to O-H water combinations bands.[56–58] No association between the juice NIR spectra and the sugars absorbance bands can be made due to the strong water absorbance bands. Otherwise, the dehydrated juice NIR spectra present a more interpretative profile. Bands corresponding to O-H and C-H stretch first overtones can be seen from 1400 to 1600 nm e 1650 to 1800 nm. The second overtone of C=O stretches bands from 1900 to 2000 nm. Combinations bands of O-H bands from 2000 to 2200 nm and C-C and C-H combinations bands from 2200 to 2500 nm.[31,51,59] Although presenting similar structures and NIR profiles, sucrose, glucose, and fructose may be individually modeled using NIR due to their different absorptivity related to O-H, C-H, and C=O bands.

3.5 - MODELING

Before modeling, the raw spectra were preprocessed to improve the signal-to-noise ratio, remove interference such as multiplicative shifts due to light scattering and baseline shifts.

Several preprocessing methods were assessed. The optimal preprocess for each property can be seen in Table 3. The spectra acquired in the Antaris instruments consisted of 3112 variables, and the one acquired in the NIRscan Nano consisted of 605 variables. The edges of the NIRscan Nano spectra were removed, as they exhibited a noisy and irregular shape, so the models were built using 551 variables. The models built using all variables will be referred to as PLS, and the models using the variables selected using OPS as PLS-OPS.

The results from Table II-3 show that the OPS method improved most models, reducing the RMSE and increasing the R values. The improvement could also be observed in reducing LOD values when comparing the PLS and PLS-OPS models. For all models, low γ^{-1} values can be observed. A decrease in their values is noticed when using selected variables, which may be due to more predictive variables. As expected, low SEL values were observed for all models, as sugarcane juice is a complex sample, and its spectrum also contains information about other components.

The DSJ method resulted in better predictions for all sugars using both instruments, being the benchtop statistically better than the portable instrument. On the benchtop instrument, the PLS-OPS models presented RMSEP values for sucrose, glucose, and fructose 7.98, 0.82, and 1.00 mg/mL, respectively, using the DSJ method against 12.75, 1.00, and 1.35 mg/mL using the LSJ method. For the portable instrument, the RMSEP values of PLS-OPS models were and 1.65 mg/mL using the DSJ method against 23.23, 1.40, and 2.08 mg/mL using the LSJ method. Therefore, the best models for sucrose, glucose, and fructose were achieved using the DSJ and Antaris instruments.

Table II-3. Modeling parameters of the sucrose, glucose, and fructose NIR-based PLS models using the Antaris and NIRScan Nano instruments and liquid sugarcane juice (LSJ) and dehydrated sugarcane juice (DSJ).

Instrument	Analysis	Property	Method	Variables	RMSEC	Rc	RMSECV	Rcv	RMSEP	Rp	LV	SEL	γ^{-1}	LOD	Preprocessment
Antaris	LSJ	Fructose	PLS	3112	1.32	0.89	1.59	0.83	1.54	0.84	10	0.009	$3.9x10^{-5}$	1.65	S - SNV - MC
			PLS-OPS	250	1.25	0.90	1.38	0.87	1.35	0.87	9	0.003	$2.7x10^{-4}$	1.47	
		Glucose	PLS	3112	0.94	0.87	1.13	0.80	1.12	0.80	10	0.010	$2x10^{-5}$	1.96	S - SNV - MC
			PLS-OPS	1320	0.90	0.88	1.01	0.85	1.00	0.84	10	0.004	$8.1x10^{-5}$	1.81	
		Sucrose	PLS	3112	14.24	0.88	16.67	0.84	15.00	0.87	10	2.06E-04	0.001	34.89	S - SNV
			PLS-OPS	40	12.42	0.91	13.93	0.89	12.75	0.91	10	4.17E-05	0.055	28.50	
	DSJ	Fructose	PLS	3112	0.89	0.95	1.20	0.91	1.00	0.93	9	0.040	0.001	0.97	N - FD - MC
			PLS-OPS	680	0.56	0.98	1.12	0.92	1.01	0.92	9	0.035	0.002	0.81	
		Glucose	PLS	3112	0.66	0.94	0.94	0.87	0.82	0.90	9	0.025	0.004	1.18	FD - MC
			PLS-OPS	1420	0.42	0.97	0.90	0.88	0.80	0.89	9	0.025	0.009	0.93	
		Sucrose	PLS	3112	10.33	0.94	13.26	0.90	10.78	0.93	9	0.005	0.031	22.33	FD - SNV - N
			PLS-OPS	90	6.83	0.97	9.48	0.95	7.98	0.96	9	0.004	0.164	13.76	
NIRscan Nano	LSJ	Fructose	PLS	551	1.79	0.77	2.30	0.61	2.08	0.69	10	0.026	0.437	2.93	FD - N - MC
			PLS-OPS	210	1.81	0.77	2.23	0.64	2.09	0.68	10	0.022	1.608	2.95	
		Glucose	PLS	551	1.37	0.69	1.57	0.57	1.43	0.65	10	0.002	0.243	4.52	S - FD - SNV
			PLS-OPS	170	1.34	0.71	1.47	0.63	1.40	0.68	10	0.002	0.348	4.21	
		Sucrose	PLS	551	21.53	0.71	25.04	0.58	23.23	0.64	6	0.006	2.702	82.67	FD - MSC - S
			PLS-OPS	30	21.11	0.72	22.92	0.66	23.33	0.63	6	0.005	14.224	79.72	
	DSJ	Fructose	PLS	551	1.44	0.86	1.90	0.75	1.69	0.79	9	0.006	0.111	1.90	MC
			PLS-OPS	220	1.21	0.90	1.81	0.77	1.65	0.81	9	0.006	0.098	1.87	
		Glucose	PLS	551	0.86	0.89	1.33	0.72	1.27	0.74	10	0.003	0.108	1.70	MC
			PLS-OPS	320	0.86	0.89	1.35	0.71	1.18	0.77	10	0.003	0.200	1.89	
		Sucrose	PLS	551	9.71	0.95	14.65	0.88	15.90	0.85	10	0.005	4.256	20.68	MC
			PLS-OPS	420	9.52	0.95	14.86	0.87	15.76	0.85	10	0.005	4.256	20.68	

*RMSE and LOD in mg/mL; LSJ: liquid sugarcane juice; DSJ: dehydrated sugarcane juice; RMSEC: root mean square error of calibration; Rc: correlation coefficient of calibration; RMSEP: root mean square error of prediction; Rp: correlation coefficient of prediction; LV: latent variables; S: smoothing; SNV: standard normal variate scaling; N: normalize; MC: mean center; FD: first derivative; MSC: multiplicative scattering correction; SD: second derivative

The LSJ models using the portable instrument did not exhibit satisfactory results. It may be due to water interference since the main band present in the region from 900 to 1700 corresponds to water absorbance. The plots of the reference values and the content predicted for sucrose, glucose, and fructose using both instruments and spectral analysis methods can be seen in Figure II-8. Using the DSJ method significantly improved the results using the portable instrument. This result highlights the effect of the dehydration step in the model predictive power for the portable instrument, which could increase its use. The loss in predictive power when compared with the benchtop instrument is compensated with the possibility of using the instrument on-site.

The implementation of a dehydration step increased the predictive power of the sucrose, glucose, and fructose models using either the benchtop or portable instrument. These increasing in accuracy could justify its implementation in the genetic breeding sites and in the industry. When creating a NIR based method, the quality of the reference analysis is very important. For that reason, the proposed dehydration methodology could also be used as reference methodology, instead of the HPLC that is very expensive and time-consuming and the polarimetry or soluble solids methodology that are indirect measurements only for sucrose content and do not give any information about glucose and fructose contents, to create other models with less sample preparation, such as using the stalk, for instance.

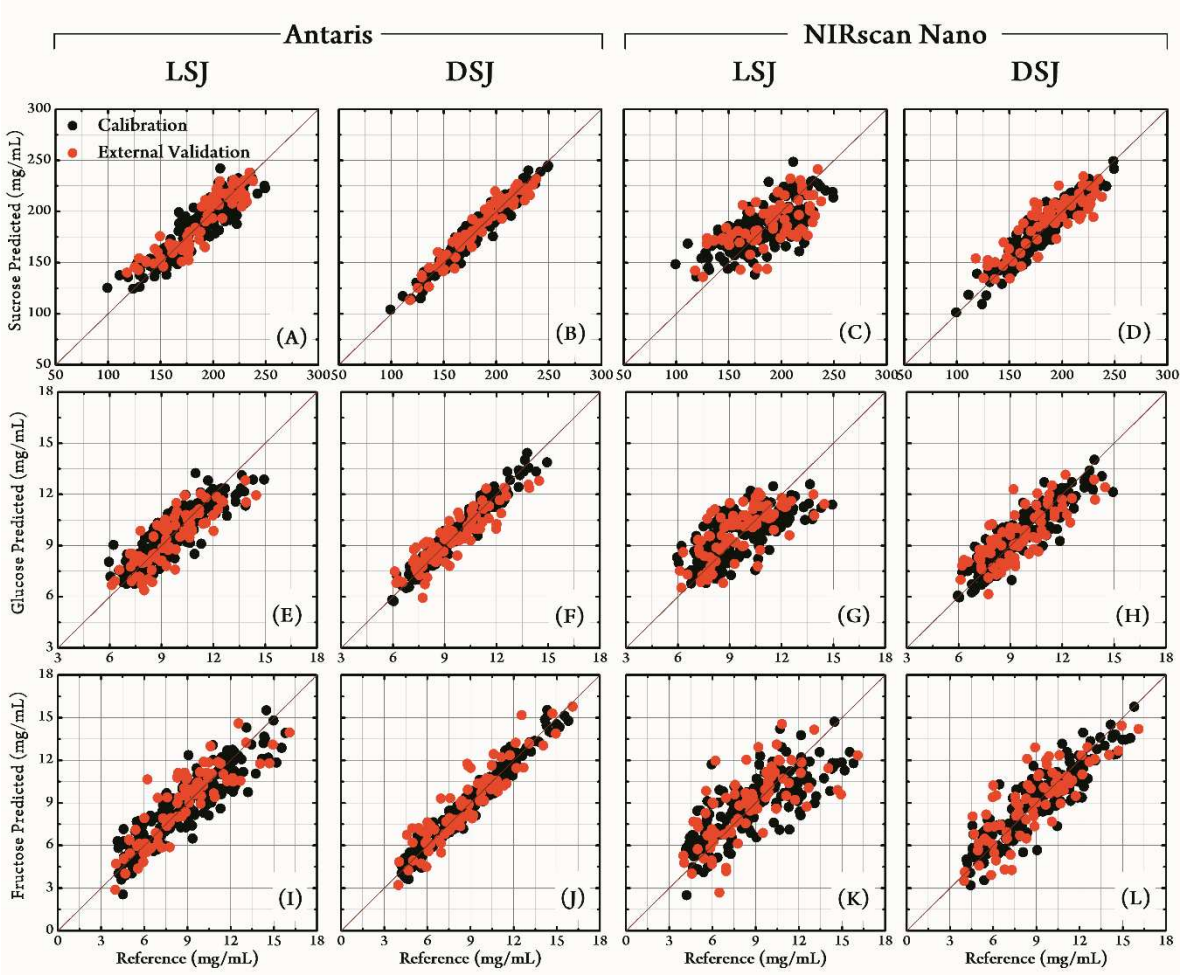


Figure II-8. HPLC reference analysis versus NIR predicted values for sucrose (A-D), glucose (E-H), and fructose (I-L) using the Antaris and NIRScan Nano instruments.

Figure II-9 shows the relative errors of the validation set. For all properties evaluated, the model built using the Antaris instrument and the DSJ resulted in the smallest average absolute relative error. For fructose, the average absolute relative error were 12.7%, 10.1%, 20.3%, and 18% for LSJ and DSJ using Antaris and LSJ and DSJ using NIRScan Nano. For glucose, the average absolute relative error were 8.6%, 7%, 12%, and 11.3% for LSJ and DSJ using Antaris and LSJ and DSJ using NIRScan Nano. For sucrose, the average absolute relative error were 5.8%, 3.4%, 10.3%, and 7.3% for LSJ and DSJ using Antaris and LSJ and DSJ using NIRScan Nano.

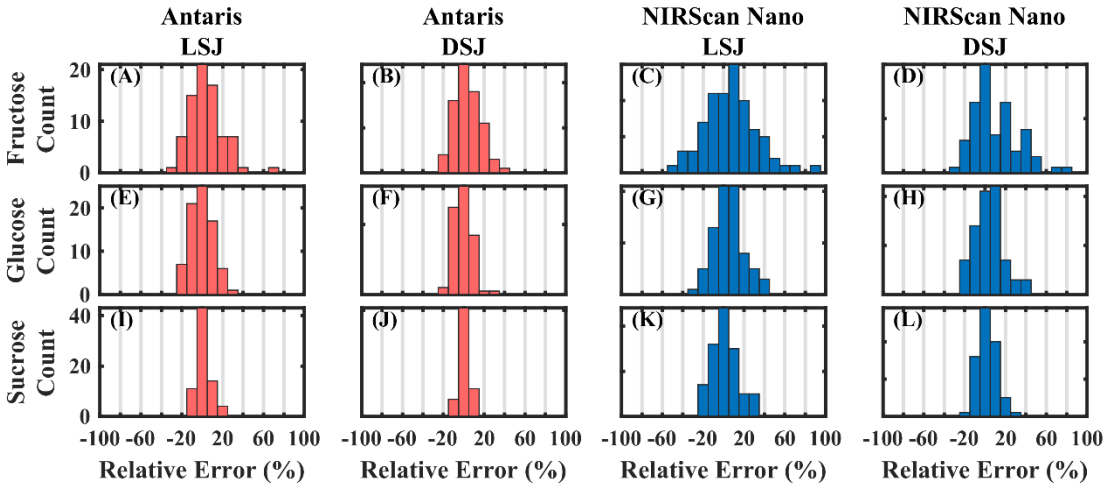


Figure II-9. Relative error for the validation set of the models for fructose (A-D), glucose (E-H), and sucrose (I-L) using the LSJ and DSJ analysis and Antaris and NIRScan Nano instruments.

The results acquired using the Antaris instrument were superior to the ones using the NIRscan Nano. It may be due to the broader wavelength range, higher resolution, and less noisy spectra of the Antaris instrument, which results in more information, and therefore, more predictive power. Figure II-10 shows the variables selected using OPS for DSJ and LSJ models using the Antaris and NIRScan Nano instruments. Considering the models using the Antaris, it can be observed that for the model using the LSJ (Figure II-10A), the OPS did not select variables in the water-related regions, 1400 to 1550 nm, and above 1900 nm. On the other hand, for the model using the DSJ (Figure II-10B), the OPS selected variables in this region, corroborating that the dehydration process can remove the water influence. Although pure sucrose, glucose, and fructose present similar NIR profiles, it can be observed that different regions were selected for each sugar model, i.e., for sucrose, the OPS selected regions from 1000 to 1300, 1400 to 1500, around 1700, 1950, 2100, and above 2300 nm. For glucose, the OPS selected regions from 1000 to 1850, around 2000, 2100, and 2250 nm. For fructose, the OPS selected regions from 1100 to 1400 and 1550 to 1850 nm. These selected regions can be associated with specific regions of each sugar, as shown in Figure II-7A. Considering the models using the NIRScan Nano, Figures II-10C and II-10D, the OPS could not improve the models.

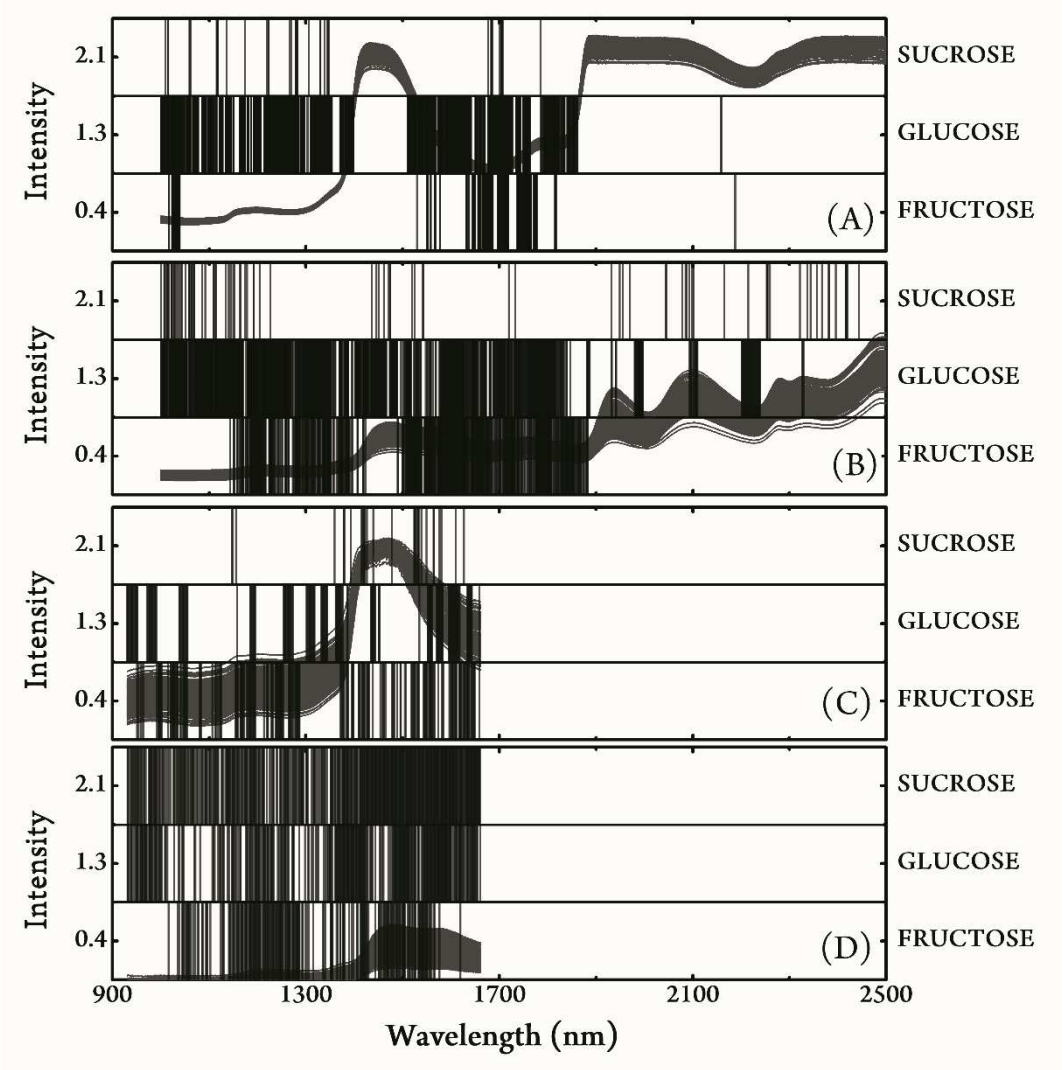


Figure II-10. Variables selected by OPS for the sucrose, glucose, and fructose models using the LSJ (A) and DSJ (B) using the Antaris instrument and LSJ (C) and DSJ (D) using the NIRScan Nano instrument.

Figure II-11 shows the Pearson correlation charts for the models. It can be observed that all the models built using the authentic y (red sphere) presented a Rc and Rcv different and higher than those using the randomized y vectors, which means the models did not occur by chance.

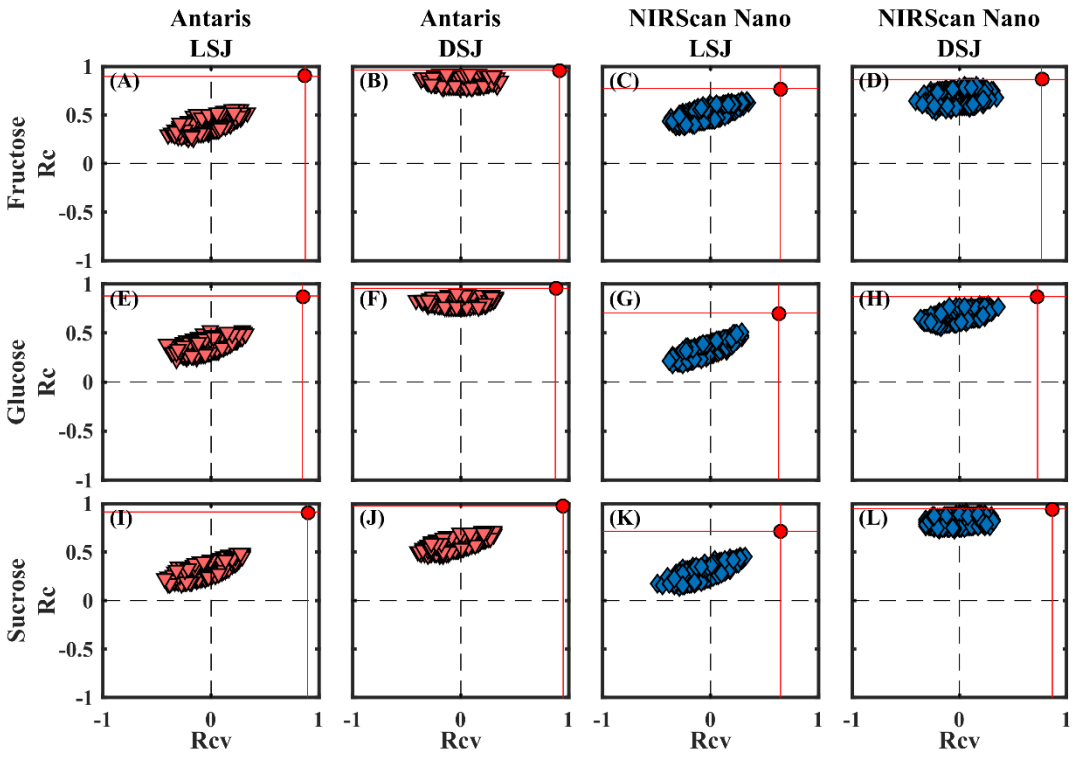


Figure II-11. Pearson correlation coefficient of calibration (Rc) versus cross-validation (Rcv) for fructose (A-D), glucose (E-H), and sucrose (I-L) using the Antaris instruments and NIRScan Nano and the LSJ and DJS analysis. The red sphere represents the model using the authentic y vector.

3.6 - STATISTICAL COMPARISON OF MODELS

Tukey's test with 95% confidence was performed to compare the model's means. Figure II-12 shows the average RMSEP values for the sucrose, glucose, ad fructose models. Tukey's test results confirmed the idea that the water removal could improve the models' predictive power. The predictive power was improved for all properties and instruments when dehydrating the sample (DSJ models). In particular, the results for the NIRscan Nano instrument were significantly improved once the models using the sugarcane juice did not present enough capability to predict the properties. The Tukey's test corroborated with the results in Table II-3 that the best models were acquired using the DSJ method and the Antaris instrument, followed by the LSJ method using the Antaris instrument, and then the DSJ using the NIRScan Nano. The LSJ using the NIRScan Nano did not present satisfactory predictions.

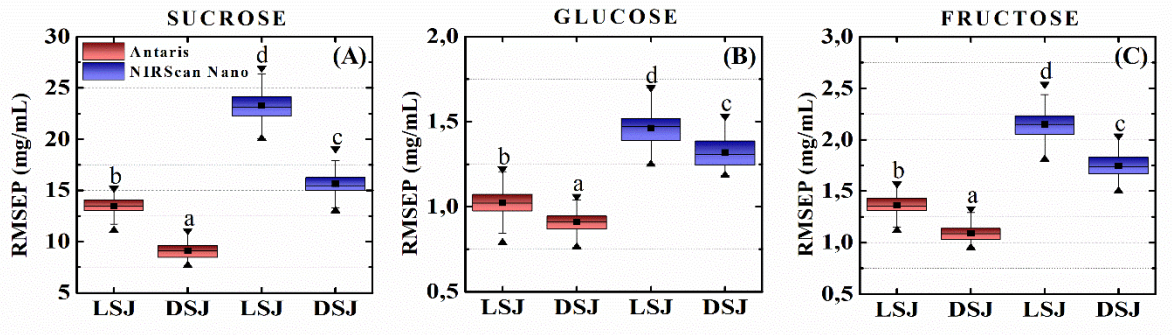


Figure II-12. Boxplot of the RMSEP values of the models using the Antaris and NIRScan Nano for sucrose (A), glucose (B), and fructose (C). The low case letters indicate the Tukey's test results with 95% confidence, where different letters mean the models are significantly different.

4 - CONCLUSIONS

The proposed dehydration methodology showed to be a great technique to improve the predictive power of PLS models by removing the water interference and concentrating the analytes. The dehydration process is quite simple and presents a high analytical frequency of 36 samples per hour (including dehydration and NIR analysis). In addition, it presents a lower cost and faster methodology than HPLC, which has an analytical frequency of 3 samples per hour. Using the novel approach resulted in more predictive models for sucrose, glucose, and fructose than liquid sugarcane juice.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (Project: CEX - APQ-02254-15). We are thankful to Prof. Luiz A.S. Dias from the Plant Science Department of the UFV for providing the NIR spectrometer.

REFERENCES

(1) CONAB. Companhia Nacional de Abastecimento. *Acompanhamento da Safra Brasileira: Cana-de-açúcar* **2019**, 58.

(2) Zabed, H.; Faruq, G.; Sahu, J. N.; Azirun, M. S.; Hashim, R.; Nasrulhaq Boyce, A. Bioethanol Production from Fermentable Sugar Juice. *The Scientific World Journal* **2014**, *2014*, 1–11. https://doi.org/10.1155/2014/957102.

(3) Cavalett, O.; Junqueira, T. L.; Dias, M. O. S.; Jesus, C. D. F.; Mantelatto, P. E.; Cunha, M. P.; Franco, H. C. J.; Cardoso, T. F.; Maciel Filho, R.; Rossell, C. E. V.; Bonomi, A. Environmental and Economic Assessment of Sugarcane First Generation Biorefineries in Brazil. *Clean Technol Environ Policy* **2012**, *14* (3), 399–410. https://doi.org/10.1007/s10098-011-0424-7.

(4) Silva, L. A.; Gasparini, K.; Assis, C.; Ramos, R.; Kist, V.; Barbosa, M. H. P.; Teófilo, R. F.; Bhering, L. L. Selection Strategy for Indication of Crosses between Potential Sugarcane Genotypes Aiming at the Production of Bioenergy. *Ind Crops Prod* **2017**, *104* (December 2016), 62–67. https://doi.org/10.1016/j.indcrop.2017.04.025.

(5) Kandel, R.; Yang, X.; Song, J.; Wang, J. Potentials, Challenges, and Genetic and Genomic Resources for Sugarcane Biomass Improvement. *Front Plant Sci* **2018**, *9* (February), 1–14. https://doi.org/10.3389/fpls.2018.00151.

(6) Barbosa, M. H. P.; Resende, M. D. V.; Dias, L. A. dos S.; Barbosa, G. V. de S.; Oliveira, R. A. de; Peternelli, L. A.; Daros, E. Genetic Improvement of Sugar Cane for Bioenergy: The Brazilian Experience in Network Research with RIDESA. *Crop Breeding and Applied Biotechnology* **2012**, *12* (spe), 87–98. https://doi.org/10.1590/S1984-70332012000500010.

(7) Lavanholi, M. G. D. Qualidade Da Cana-de-Açúcar Como Matéria-Prima Para Produção de Açúcar e Álcool. In *Cana-de-açúcar*; Instituto Agronômico, 2010; p 882. https://doi.org/doi.org/10.3738/nucleus.v5i2.102.

(8) Fernandes, A. C. Cálculos Na Agroindústria Da Cana-de-Açúcar. *STAB: Piracicaba, SP, Brasil* **2003**, No. 2º.

(9) Tai, P. Y. P.; Miller, J. D. Germplasm Diversity among Four Sugarcane Species for Sugar Composition. *Crop Sci* **2002**, *42* (3), 958–964. https://doi.org/10.2135/cropsci2002.9580.

(10) Rein, P.; others. *Cane Sugar Engineering.*; Verlag Dr. Albert Bartens KG, 2016.

(11) Al-Mhanna, N. M.; Huebner, H.; Buchholz, R. Analysis of the Sugar Content in Food Products by Using Gas Chromatography Mass Spectrometry and Enzymatic Methods. *Foods* **2018**, *7* (11). https://doi.org/10.3390/foods7110185.

(12) Zhao, D.; MacKown, C. T.; Starks, P. J.; Kindiger, B. K. Rapid Analysis of Nonstructural Carbohydrate Components in Grass Forage Using Microplate Enzymatic Assays. *Crop Sci* **2010**, *50* (4), 1537–1545. https://doi.org/10.2135/cropsci2009.09.0521.

(13) Liu, W.; Wu, H.; Li, B.; Dong, C.; Choi, M. M. F.; Shuang, S. Immobilization of Platinum Nanoparticles and Glucose Oxidase on Eggshell Membrane for Glucose Detection. *Analytical Methods* **2013**, *5* (19), 5154–5160. https://doi.org/10.1039/c3ay40327k.

(14) Teixeira, A. I.; Ribeiro, L. F.; Rezende, S. T.; Barros, E. G.; Moreira, M. A. Development of a Method to Quantify Sucrose in Soybean Grains. *Food Chem* **2012**, *130* (4), 1134–1136. https://doi.org/10.1016/j.foodchem.2011.07.128.

(15) Filip, M.; 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. *Food Chem* **2016**, *199*, 653–659. https://doi.org/10.1016/j.foodchem.2015.12.060.

(16) Simeone, M. L. F.; Parrella, R. A. C.; Schaffert, R. E.; Damasceno, C. M. B.; Leal, M. C. B.; Pasquini, C. Near Infrared Spectroscopy Determination of Sucrose, Glucose and Fructose in Sweet Sorghum Juice. *Microchemical Journal* **2017**, *134*, 125–130. https://doi.org/10.1016/j.microc.2017.05.020.

(17) Shanmugavelan, P.; Kim, S. Y.; Kim, J. B.; Kim, H. W.; Cho, S. M.; Kim, S. N.; Kim, S. Y.; Cho, Y. S.; Kim, H. R. Evaluation of Sugar Content and Composition in Commonly Consumed Korean Vegetables, Fruits, Cereals, Seed Plants, and Leaves by HPLC-ELSD. *Carbohydr Res* **2013**, *380*, 112–117. https://doi.org/10.1016/j.carres.2013.06.024.

(18) 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* **2014**, *145*, 784–788. https://doi.org/10.1016/j.foodchem.2013.08.135.

(19) Montesano, D.; Cossignani, L.; Giua, L.; Urbani, E.; Simonetti, M. S.; Blasi, F. A Simple HPLC-ELSD Method for Sugar Analysis in Goji Berry. *J Chem* **2016**, *2016*, 1–5. https://doi.org/10.1155/2016/6271808.

(20) Dugalic, K.; Sudar, R.; Viljevac, M.; Josipovic, M.; Cupic, T. Sorbitol and Sugar Composition in Plum Fruits Influenced by Climatic Conditions. *Journal of Agricultural Science and Technology* **2014**, *16* (5), 1145–1155.

(21) Moral, A.; Hernández, M. D.; Tijero, A.; González, Z.; García, J.; De La Torre, M. J. NIRS Determination of Carbohydrates from Hydrothermal-Treated Rice Straw. *Tappi J* **2012**, *11* (4), 27–32. https://doi.org/10.32964/tj11.4.27.

(22) García Asprilla, I. D. S.; Ramírez-Navas, J. S. Near-Infrared Spectroscopy: A Rapid Alternative Technique to Reducing Sugars Determination in Juice of Sugarcane (Saccharum Officinarum L.). *J Pharm Pharmacogn Res* **2018**, *6* (5), 392–401.

(23) Rácz, A.; Héberger, K.; Fodor, M. Quantitative Determination and Classification of Energy Drinks Using Near-Infrared Spectroscopy. *Anal Bioanal Chem* **2016**, *408* (23), 6403–6411. https://doi.org/10.1007/s00216-016-9757-8.

(24) Caporaso, N.; Whitworth, M. B.; Grebby, S.; Fisk, I. D. Non-Destructive Analysis of Sucrose, Caffeine and Trigonelline on Single Green Coffee Beans by Hyperspectral Imaging. *Food Research International* **2018**, *106* (November 2017), 193–203. https://doi.org/10.1016/j.foodres.2017.12.031.

(25) Ilaslan, K.; Boyaci, I. H.; Topcu, A. Rapid Analysis of Glucose, Fructose and Sucrose Contents of Commercial Soft Drinks Using Raman Spectroscopy. *Food Control* **2015**, *48*, 56–61. https://doi.org/10.1016/j.foodcont.2014.01.001.

(26) Özbalci, B.; Boyaci, I. H.; Topcu, A.; Kadilar, C.; Tamer, U. Rapid Analysis of Sugars in Honey by Processing Raman Spectrum Using Chemometric Methods and Artificial Neural Networks. *Food Chem* **2013**, *136* (3–4), 1444–1452. https://doi.org/10.1016/j.foodchem.2012.09.064.

(27) Huang, Y.; Carragher, J.; Cozzolino, D. Measurement of Fructose, Glucose, Maltose and Sucrose in Barley Malt Using Attenuated Total Reflectance Mid-Infrared Spectroscopy. *Food Anal Methods* **2016**, *9* (4), 1079–1085. https://doi.org/10.1007/s12161-015-0286-4.

(28) Leopold, L. F.; Leopold, N.; Diehl, H. A.; Socaciu, C. Quantification of Carbohydrates in Fruit Juices Using FTIR Spectroscopy and Multivariate Analysis. *Spectroscopy* **2011**, *26* (2), 93–104. https://doi.org/10.3233/SPE-2011-0529.

(29) Pasquini, C. Near Infrared Spectroscopy: A Mature Analytical Technique with New Perspectives – A Review. *Anal Chim Acta* **2018**, *1026*, 8–36. https://doi.org/10.1016/j.aca.2018.04.004.

(30) Pasquini, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. *J Braz Chem Soc* **2003**, *14* (2), 198–219. https://doi.org/10.1590/S0103-50532003000200006.

(31) Osborne, B. G. Near-Infrared Spectroscopy in Food Analysis. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons, Ltd: Chichester, UK, 2000; pp 1–14. https://doi.org/10.1002/9780470027318.a1018.

(32) Cen, H.; He, Y.; Huang, M. Measurement of Soluble Solids Contents and PH in Orange Juice Using Chemometrics and Vis-NIRS. *J Agric Food Chem* **2006**, *54* (20), 7437–7443. https://doi.org/10.1021/jf061689f.

(33) Liu, C.; Yang, S. X.; Deng, L. Determination of Internal Qualities of Newhall Navel Oranges Based on NIR Spectroscopy Using Machine Learning. *J Food Eng* **2015**, *161*, 16–23. https://doi.org/10.1016/j.jfoodeng.2015.03.022.

(34) Rodriguez-Saona, L. E.; Fry, F. S.; McLaughlin, M. A.; Calvey, E. M. Rapid Analysis of Sugars in Fruit Juices by FT-NIR Spectroscopy. *Carbohydr Res* **2001**, *336* (1), 63–74. https://doi.org/10.1016/S0008-6215(01)00244-0.

(35) Cayuela, J. A.; Weiland, C. Intact Orange Quality Prediction with Two Portable NIR Spectrometers. *Postharvest Biol Technol* **2010**, *58* (2), 113–120. https://doi.org/10.1016/j.postharvbio.2010.06.001.

(36) Oliveira-Folador, G.; Bicudo, M. de O.; de Andrade, E. F.; Renard, C. M. G. C.; Bureau, S.; de Castilhos, F. Quality Traits Prediction of the Passion Fruit Pulp Using NIR and MIR Spectroscopy. *Lwt* **2018**, *95* (April), 172–178. https://doi.org/10.1016/j.lwt.2018.04.078.

(37) Alamar, P. D.; Caramês, E. T. S.; Poppi, R. J.; Pallone, J. A. L. Quality Evaluation of Frozen Guava and Yellow Passion Fruit Pulps by NIR Spectroscopy and Chemometrics. *Food Research International* **2016**, *85*, 209–214. https://doi.org/10.1016/j.foodres.2016.04.027.

(38) Shao, Y.; He, Y. Nondestructive Measurement of the Internal Quality of Bayberry Juice Using Vis/NIR Spectroscopy. *J Food Eng* **2007**, *79* (3), 1015–1019. https://doi.org/10.1016/j.jfoodeng.2006.04.006.

(39) Taira, E.; Ueno, M.; Saengprachatanarug, K.; Kawamitsu, Y. Direct Sugar Content Analysis for Whole Stalk Sugarcane Using a Portable near Infrared Instrument. *J Near Infrared Spectrosc* **2013**, *21* (4), 281–287. https://doi.org/10.1255/jnirs.1064.

(40) Taira, E.; Ueno, M.; Kawamitsu, Y. Automated Quality Evaluation System for Net and Gross Sugarcane Samples Using near Infrared Spectroscopy. *J Near Infrared Spectrosc* **2010**, *18* (3), 209–215. https://doi.org/10.1255/jnirs.884.

(41) Tewari, J.; Mehrotra, R.; Irudayaraj, J. Direct near Infrared Analysis of Sugar Cane Clear Juice Using a Fibre-Optic Transmittance Probe. *J Near Infrared Spectrosc* **2003**, *11* (5), 351–356. https://doi.org/10.1255/jnirs.386.

(42) Sorol, N.; Arancibia, E.; Bortolato, S. A.; Olivieri, A. C. Visible/near Infrared-Partial Least-Squares Analysis of Brix in Sugar Cane Juice. *Chemometrics and Intelligent Laboratory Systems* **2010**, *102* (2), 100–109. https://doi.org/10.1016/j.chemolab.2010.04.009.

(43) 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* **2007**, *18* (2), 259–266. https://doi.org/10.1590/S0103-50532007000200003.

(44) Maraphum, K.; Chuan-Udom, S.; Saengprachatanarug, K.; Wongpichet, S.; Posom, J.; Phuphaphud, A.; Taira, E. Effect of Waxy Material and Measurement Position of a Sugarcane Stalk on the Rapid Determination of Pol Value Using a Portable near Infrared Instrument. *J Near Infrared Spectrosc* **2018**, *26* (5), 287–296. https://doi.org/10.1177/0967033518795810.

(45) Corrêdo, L. de P.; Maldaner, L. F.; Bazame, H. C.; Molin, J. P. Evaluation of Minimum Preparation Sampling Strategies for Sugarcane Quality Prediction by Vis-NIR Spectroscopy. *Sensors* **2021**, *21* (6), 2195. https://doi.org/10.3390/s21062195.

(46) Phetpan, K.; Udompetaikul, V.; Sirisomboon, P. An Online Visible and Near-Infrared Spectroscopic Technique for the Real-Time Evaluation of the Soluble Solids Content of Sugarcane Billets on an Elevator Conveyor. *Comput Electron Agric* **2018**, *154* (October), 460–466. https://doi.org/10.1016/j.compag.2018.09.033.

(47) DU, Y. P.; LIANG, Y. Z.; KASEMSUMRAN, S.; MARUO, K.; OZAKI, Y. Removal of Interference Signals Due to Water from in Vivo Near-Infrared (NIR) Spectra of Blood

Glucose by Region Orthogonal Signal Correction (ROSC). *Analytical Sciences* **2004**, *20* (9), 1339–1345. https://doi.org/10.2116/analsci.20.1339.

(48) Devaux, M. F.; Bertrand, D.; Robert, P.; Qannari, M. Application of Principal Component Analysis on NIR Spectral Collection after Elimination of Interference by a Least-Squares Procedure. *Appl Spectrosc* **1988**, *42* (6), 1020–1023. https://doi.org/10.1366/0003702884430443.

(49) Chen, D.; Shao, X.; Hu, B.; Su, Q. A Background and Noise Elimination Method for Quantitative Calibration of near Infrared Spectra. *Anal Chim Acta* **2004**, *511* (1), 37–45. https://doi.org/10.1016/j.aca.2004.01.042.

(50) Rambla, F. J.; Garrigues, S.; Guardia, M. De. PLS-NIR Determination of Total Sugar , Glucose , Fructose and Sucrose in Aqueous Solutions of Fruit Juices. **1997**, *2670* (97).

(51) Golic, M.; Walsh, K.; Lawson, P. Short-Wavelength Near-Infrared Spectra of Sucrose, Glucose, and Fructose with Respect to Sugar Concentration and Temperature. *Appl Spectrosc* **2003**, *57* (2), 139–145. https://doi.org/10.1366/000370203321535033.

(52) Valderrama, P.; Braga, J. W. B.; Poppi, R. J. Estado Da Arte de Figuras de Mérito Em Calibração Multivariada. *Quim Nova* **2009**, *32* (5), 1278–1287. https://doi.org/10.1590/S0100-40422009000500034.

(53) Ortiz, M. C.; Sarabia, L. A.; Herrero, A.; Sánchez, M. S.; Sanz, M. B.; Rueda, M. E.; Giménez, D.; Meléndez, M. E. Capability of Detection of an Analytical Method Evaluating False Positive and False Negative (ISO 11843) with Partial Least Squares. *Chemometrics and Intelligent Laboratory Systems* **2003**, *69* (1–2), 21–33. https://doi.org/10.1016/S0169-7439(03)00110-2.

(54) Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C. C. Sorting Variables by Using Informative Vectors as a Strategy for Feature Selection in Multivariate Regression. *J Chemom* **2009**, *23* (1), 32–48. https://doi.org/10.1002/cem.1192.

(55) Roque, J. V.; Cardoso, W.; Peternelli, L. A.; Teófilo, R. F. Comprehensive New Approaches for Variable Selection Using Ordered Predictors Selection. *Anal Chim Acta* **2019**, *1075*, 57–70. https://doi.org/10.1016/j.aca.2019.05.039.

(56) Buijs, K.; Choppin, G. R. Near-Infrared Studies of the Structure of Water. I. Pure Water. *J Chem Phys* **1963**, *39* (8), 2035–2041. https://doi.org/10.1063/1.1734579.

(57) Segtnan, V. H.; Šašić, Š.; Isaksson, T.; Ozaki, Y. Studies on the Structure of Water Using Two-Dimensional Near-Infrared Correlation Spectroscopy and Principal Component Analysis. *Anal Chem* **2001**, *73* (13), 3153–3161. https://doi.org/10.1021/ac010102n.

(58) 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* **2019**, *24* (20), 3696. https://doi.org/10.3390/molecules24203696.

(59) Beganović, A.; Beć, K. B.; Grabska, J.; Stanzl, M. T.; Brunner, M. E.; Huck, C. W. Vibrational Coupling to Hydration Shell – Mechanism to Performance Enhancement of Qualitative Analysis in NIR Spectroscopy of Carbohydrates in Aqueous Environment. *Spectrochim Acta A Mol Biomol Spectrosc* **2020**, *237*, 118359. https://doi.org/10.1016/j.saa.2020.118359.

GENERAL CONCLUSIONS

In general, a proper sample preparation and data preprocessing strategy showed to be very important to improve the accuracy of the models. A dehydration step prior to acquiring the NIR spectra of the sugarcane juice significantly improved the accuracy of the models used to predict sucrose, glucose, and fructose contents of sugarcane juice. In addition, the data preprocessing step is very important when creating models. A comprehensive analysis was carried out by studying 67 calibration cases. Not fixing the preprocessing order based on the artifact they fix was fundamental to build the most accurate models. The proposed optimization method is universal and can be applied to different datasets sources.