UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Química

Hélio Milito Martins de Amorim Neto

CHEMOMETRIC AUTHENTICATION OF CONILON COFFEE (*COFFEA CANEPHORA)* FROM MINAS GERAIS: Untargeted Strategies Using Differential Scanning Calorimetry, Near Infrared Spectroscopy and Data Fusion

Belo Horizonte
2025

UFMG/ICEX/DQ. 1.675
D. 915

Hélio Milito Martins de Amorim Neto

CHEMOMETRIC AUTHENTICATION OF CONILON COFFEE (*COFFEA CANEPHORA)* FROM MINAS GERAIS: Untargeted Strategies Using Differential Scanning Calorimetry, Near Infrared Spectroscopy and Data Fusion

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Química.

Orientador: Marcelo Martins de Sena

Coorientadora: Elionai Cassiana de Lima Gomes

Belo Horizonte
2025

Ficha Catalográfica

Amorim Neto, Hélio Milito Martins de.
A524c 2025 D
Chemometric authentication of conilon coffee (Coffea canephora) from Minas Gerais [manuscrito] : untargeted strategies using differential scanning calorimetry, near infrared spectroscopy and data fusion / Hélio Milito Martins de Amorim Neto. 2025.
90 f. : il., gráfs., tabs.

Orientador: Marcelo Martins de Sena.
Coorientadora: Elionai Cassiana de Lima Gomes.

Dissertação (mestrado) – Universidade Federal de Minas Gerais – Departamento de Química.
Bibliografia: f. 58-65.
Apêndices: f. 66-90.

1. Química analítica – Teses. 2. Quimiometria – Teses. 3. Café – Análise – Teses. 4. Alimentos – Controle de qualidade – Teses. 5. Calorimetria – Teses. 6. Espectroscopia de infravermelho – Teses. 7. Análise discriminante – Teses. I. Sena, Marcelo Martins de, Orientador. II. Gomes, Elionai Cassiana de Lima, Coorientadora. III. Título.

CDU 043

Elaborada por Márcia Cavalcanti Moreira – Bibliotecária – CRB6-3847.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

UFMG
Programa de Pós-Graduação em Química
Departamento de Química - ICEx

química
pós-graduação-ufmg

"Chemometric Authentication Of Conilon Coffee (*coffea Canephora*) From Minas Gerais: Untargeted Strategies Using Differential Scanning Calorimetry, Near Infrared Spectroscopy And Data Fusion"

Hélio Milito Martins de Amorim Neto

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. Marcelo Martins de Sena - Orientador
UFMG

Profa. Elionai Cassiana de Lima Gomes - Coorientadora
UFMG

Prof. José Germano Veras Neto
Universidade Estadual da Paraíba

Prof. Bruno Gonçalves Botelho
UFMG

Belo Horizonte, 28 de julho de 2025.

Documento assinado eletronicamente por **Jose Germano Veras Neto**, **Usuário Externo**, em 30/07/2025, às 12:55, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Goncalves Botelho**, **Professor do Magistério Superior**, em 20/08/2025, às 13:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Elionai Cassiana de Lima Gomes**, **Usuária Externa**, em 21/08/2025, às 18:13, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcelo Martins de Sena**, **Professor do Magistério Superior**, em 26/08/2025, às 13:00, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4408519** e o código CRC **0CB0A888**.

Referência: Processo nº 23072.245416/2025-73 SEI nº 4408519

This dissertation is dedicated to my parents, who have always supported me throughout my education.

AKNOWLEDGEMENTS

I would like to acknowledge the funding agencies that supported this research: the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and the Instituto Nacional de Ciência e Tecnologia de Bioanalítica (INCTBio) as their financial support was essential for the development and execution of this work. I also extend my thanks to the Universidade Federal de Minas Gerais (UFMG) for providing the institutional and academic environment necessary for this research. In particular, I am grateful to the Department of Chemistry of the Institute of Exact Sciences for offering the infrastructure, resources, and support that made this dissertation possible. I would also like to thank the coffee producers who provided the samples used in this study, especially those from the Associação da Agricultura Familiar de Humaitá (AAFH) of Mutum, whose contributions were invaluable to this research.

Additionally, I am deeply grateful to my supervisors, Marcelo and Elionai, for their guidance, constant support, and encouragement throughout the development of this work. Their scientific expertise, constructive feedback, and dedication were fundamental not only to the completion of this dissertation but also to my academic and professional growth. I am sincerely thankful for their patience, availability, and for the opportunity to learn and collaborate under their supervision.

I would like to express my sincere gratitude to all my colleagues from the laboratory, whose companionship, collaboration, and support were fundamental throughout this journey. In particular, I am especially thankful to Ana Fulgêncio and Enrico for their constant willingness to help, for the valuable discussions, and for sharing their knowledge and experience.

I am profoundly grateful to my mother, Denise, for her unconditional love, strength, and unwavering support throughout every stage of my life. Her dedication, encouragement, and example of resilience have been fundamental in shaping who I am and in making this achievement possible. I also dedicate this work to the memory of my father, Petrônio, whose values, teachings, and inspiration continue to guide me. Although he

is no longer physically present, his influence remains alive in my choices and in the pursuit of my dreams.

Lastly, but not least, I am deeply thankful to my friends, who have been an essential source of encouragement, companionship, and support throughout this journey. My heartfelt gratitude goes especially to Bianca, Ana Clara, Bricia, Henrique, Túlio, Júlia, Larissa, Nayara, Larissa, Luan, Victor, Gustavo, Fabio, Matheus, Filipe and Felipe. Each of you, in your own way, contributed to making this path lighter, more meaningful, and filled with moments of joy and motivation. Your friendship, whether through words of encouragement, shared laughter, or simply being present, has been invaluable. I am truly grateful for your constant support, understanding, and for always believing in me.

“It is not our part to master all the tides of the world, but to do what is in us for the succour of those years wherein we are set, uprooting the evil in the fields that we know, so that those who live after may have clean earth to till. What weather they shall have is not ours to rule.”

J. R. R. Tolkien

ABSTRACT

This work aims to chemometrically authenticate coffee samples from a new production region of specialty conilon coffee located in the city of Mutum, in the Brazilian State of Minas Gerais. Canephora coffee has historically commanded lower prices than its Arabica counterpart, therefore, the establishment of high production standards has become a key objective for new producers seeking to deliver higher-quality products. In this study, commercial samples provided by the producers from Mutum, as well as from three other Brazilian States, were analyzed by differential scanning calorimetry and portable near infrared spectroscopy. The data obtained was used to construct exploratory models, discriminant analysis, and one-class classification chemometric models aimed at distinguishing the specialty coffees from Minas Gerais. Although it was not possible to define a chemical profile capable of fully discriminating the target class from other producing regions, the multivariate models, particularly those developed by combining calorimetry and spectroscopy data through a data fusion strategy, proved effective in identifying and distinguishing coffee samples from Minas Gerais.

Keywords: food control; supervised classification; protected designation of origin; one-class modelling; discriminant analysis.

RESUMO

Este trabalho tem como objetivo autenticar quimiometricamente amostras de café provenientes de uma nova região produtora de café conilon especial, localizada no município de Mutum, no Estado de Minas Gerais, Brasil. O café da espécie *Coffea canephora* historicamente apresenta preços inferiores em relação ao seu equivalente arábica, de modo que a busca por melhores padrões de produção é um objetivo dos novos produtores que almejam um produto de maior qualidade. Neste estudo, amostras comerciais fornecidas pelos produtores de Mutum, bem como de outros três estados brasileiros, foram analisadas por calorimetria exploratória diferencial e espectroscopia no infravermelho próximo portátil. Os dados obtidos foram utilizados para a construção de modelos quimiométricos exploratórios, de análise discriminante e de modelagem de classe, com o objetivo de diferenciar os cafés especiais mineiros. Embora não tenha sido possível determinar um perfil químico que discrimine a classe alvo das demais regiões produtoras, os modelos multivariados, especialmente aqueles construídos associando os dados de calorimetria e espectroscopia por meio de uma estratégia de fusão de dados, foram bem-sucedidos na identificação e separação das amostras de café provenientes de Minas Gerais.

Palavras-chave: controle de alimentos; classificação supervisionada; denominação de origem protegida; modelagem de classe; análise discriminante.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ACC — Accuracy

DD-SIMCA — Data Driven Soft Independent Modelling of Class Analogies

DSC — Differential Scanning Calorimetry

EFR — Efficiency Rate

FOM — Figure of Merit

FN — False Negative

FP — False Positive

GI — Geographical Indication

IUPAC — International Union of Pure and Applied Chemistry

MC — Mean Center

MIR — Mid Infrared

NIR — Near Infrared

NIRS — Near Infrared Spectroscopy

PC — Principal Component

PCA — Principal Component Analysis

PGI — Protected Geographical Indication

PLS — Partial Leas Squares

PLS-DA — Partial Least Squares Discriminant Analysis

RMSECV — Root Mean Squared Error of Cross-Validation

SG — Savitzky-Golay

SIMCA — Soft Independent Modelling of Class Analogies

SNV — Standard Normal Variate

STR — Sensitivity Rate

SPR — Specificity Rate

TN — True Negative

TP — True Positive

VIP — Variable Importance in Projection

SUMMARY

1 INTRODUCTION

Coffee is a globally consumed commodity, prized both as a stimulant and a symbol of hospitality. Although the most common species, *Coffea arabica*, is named after Arabia, the plant is not native to that region. The earliest recorded references to coffee plants date back to the fifth century, recounting the tale of an Ethiopian goat herder who noticed the invigorating effects of a fruit eaten by his animals and concocted an elixir with that fruit.[1–3]

In 1450, the beverage was introduced in Mecca, where it was called *qahwa*, an old Arabic word for wine. Due to its stimulating effects, the fruit was mistakenly considered a hallucinogen and condemned as contrary to Islam. Travelers returning from the Orient brought detailed descriptions of the beverage to Europe, which led to coffee being introduced to the continent via Venice in 1615, as the city maintained commercial relations with the Ottoman Empire. By the eighteenth century, countries such as the Dutch Republic, France, and Britain had begun cultivating coffee in their colonies.[2–4]

The first recorded introduction of coffee to Brazil dates back to 1727 in the state of Pará, when a Portuguese officer, sent on a diplomatic mission to French Guiana, secretly returned with coffee seeds, allegedly obtained by seducing the wife of the governor of the region. Throughout the nineteenth and twentieth centuries, coffee became Brazil's most significant agricultural commodity, symbolizing national economic growth. Its significance was such that coffee was featured on Brazil’s first national flag and on the logo of the Brazilian Football Confederation during the 1982 World Cup.[3–7]

In the last three decades of the nineteenth century, a plague called the coffee rust emerged as a threat to global coffee production. Coffee producers, particularly in the most affected regions of Asia, began searching for alternative species that could be more resistant than *Coffea arabica*. The solution came with *Coffea canephora*, known as robusta coffee, a species native from Congo. Robusta is naturally resistant to rust and better adapted to higher temperatures, greater humidity and lower altitudes than Arabica, making it more suitable for regions such as Southeast Asia. By the 1930s, coffee production in Asia and the Pacific region had more than doubled from before

the rust crisis. Today, robusta accounts for over 40% of global coffee production. Brazil is the second-largest producer, behind only Vietnam, contributing more than 20% of global output in the 2024 harvest.[4,8,9]

The pursuit of specialty coffees in Brazil began in the 1990s, driven by an Italian producer seeking unique Arabica beans for their brand blends. Robusta coffee, however, has traditionally been associated with a more bitter flavor, resulting in lower market prices and its predominant use in blends and instant coffee products. By the late 2000s, however, the first specialty canephora coffees of the conilon variety were produced in the Brazilian State of Espírito Santo, resulting in higher market values.[3]

To ensure quality and origin, certification of provenance was introduced, requiring analytical methods to verify product authenticity. In the 2020s, the states of Espírito Santo and Rondônia achieved protected geographical indication (PGI) certificates for their Conilon and Robusta coffees, respectively.[10–12]

Traditionally, coffee analysis has employed techniques such as chromatography and capillary electrophoresis, which require extensive sample preparation. In contrast, thermal analysis methods demand little to no sample manipulation. Another technique increasingly utilized for differentiating coffee samples is near-infrared (NIR) spectroscopy. Although less common than mid-infrared (MIR) spectroscopy, near-infrared is gaining traction due to advancements in equipment miniaturization and the development of portable devices that can be used on-site by producers and inspection agents.[1,13,14]

Chemometric methods are extensively applied for the multivariate classification of food samples using untargeted strategies. By analyzing the entire dataset rather than focusing on a single signal, it becomes possible to study complex matrices without prior knowledge of all their components. Techniques such as thermal analysis and near-infrared spectroscopy have been proved effective for analyzing complex samples including food, pharmaceuticals, fuel, and more. Additionally, data fusion strategies offer a robust approach for integrating different types of chemical information within the models.[15–19]

This study aims to develop a method for differentiating and classifying conilon coffee beans produced in the region of the city of Mutum, Minas Gerais, from those grown in other regions of Brazil. Exploratory analysis, discriminant analysis and one-class modeling approaches were applied using differential scanning calorimetry and near-infrared spectroscopy, both individually and in combination through a data fusion strategy. The developed supervised methods were then compared and validated using appropriate qualitative figures of merit.

2 OBJECTIVES

2.1 General Objective

To build chemometric models to discriminate conilon coffee samples from the city of Mutum, Minas Gerais, from samples from other producer regions. This city was chosen as it is family farming region starting their specialty conilon coffee production, aiming for higher producing standards and product quality.

2.2 Specific Objectives

I. Differential scanning calorimetry analysis of the ground coffee samples.
II. Near infrared spectroscopy analysis of the ground coffee samples.
III. Chemometric modeling with exploratory and supervised analysis for discrimination and authentication of the producer regions using DSC curves, NIR spectra and data fusion strategies.
IV. Qualitative validation of the built models.

3 BIBLIOGRAPHIC REVISION

3.1 Coffee

Coffee is one of the most consumed beverages in the world, prepared from the roasted beans of the plants of the *Coffea* genus. Out of approximately 500 species within this genus, only two hold significant economic importance: *Coffea arabica* and *Coffea canephora*. Canephora coffees, also called robusta, generally command lower market prices than arabica, with a price difference of approximately 30% in the 2024/2025 harvest. But this difference in value does not mean that robusta coffee has a lower quality and studies on producing standards resulted in two geographical indications (GIs) in Brazil, one for conilon coffees from the State of Espirito Santo and another for robusta coffees from the State of Rondônia.[9,20,21]

GI certifications require reliable authentication methods for the products, and many studies are being conducted utilizing instrumental analysis, such as spectroscopy and chromatography, allied with chemometric tools for this objective. As coffee beans have complex chemical composition, including a great variety of organic and inorganic compounds, it is not always possible to identify specific compounds that serve as a fingerprint from a producer region. In such cases, untargeted strategies may be applied, with a total screening of the samples and use of mathematical tools to identify initial trends prior to more in-depth investigations.[21–24]

3.2 Differential Scanning Calorimetry (DSC)

Differential scanning calorimetry (DSC) was first introduced and patented by the Perkin-Elmer Corporation in the 1960s. The related instrument is capable of measuring the energy involved in endothermic and exothermic processes by a process called power compensation DSC. In this technique a sample and a reference are placed in separated ovens and the difference in energy applied between the sample and the reference to keep both at the same temperature generates the analytical signal. Another version of DSC, called heat-flux DSC, consists in both sample and reference

being placed in the same oven and difference in heat flow between the sample and the reference while the oven is heated generates the analytical signal [25–27]

In food science, DSC has been widely used to evaluate the thermal properties of food products during processing and storage. Among the advantages of this technique are the small sample size required, minimal sample preparation, and high repeatability, making it a robust analytical tool. In coffee analysis, DSC has primarily been employed to monitor the roasting process of coffee beans and to detect adulteration of grounded coffee.[1,28–30]

3.3 Near Infrared Spectroscopy (NIRS)

The discovery of the near infrared (NIR) region of the light spectrum was documented in the literature in 1800 by Frederick William Herschel and earliest studies on near infrared spectroscopy (NIRS) date back to 1905 in a publication by William Weber Coblentz. NIRS is defined as a technique that analyses the interaction between electromagnetic radiation and matter, specifically within the spectral range of 780 nm to 2500 nm, according to the International Union of Pure and Applied Chemistry (IUPAC).[31–34]

Due to the complexity of NIR spectra, traditional methods of spectral interpretation are not always an option, particularly when dealing with food matrices. Therefore, multivariate approaches are used alongside the spectral data for analytical purposes. With the development and advances of chemometrics in the second half of the twentieth century, NIRS emerged as a powerful complementary technique for obtaining both qualitative and quantitative results. In the context of food quality and authenticity, NIRS has proven to be a promising tool, offering shorter analysis times, minimal sample preparation, and the possibility of onsite measurements through the use of portable sensors, making it an attractive alternative to conventional instrumental analyses.[35,36]

3.4 Chemometrics

According to the International Union of Pure and Applied Chemistry (IUPAC), chemometrics, a term first coined by Svante Wold in 1971, is defined as the “science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods”. It was developed as a solution for the increasing volume of data generated by modern analytical instruments and the parallel advancement of computer technology. In simpler terms, chemometric analysis is the use of multivariate statistical techniques as tools to extract meaningful chemical information from complex data matrices that is easier to chemically interpret.[37,38]

3.4.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is an exploratory analysis technique first presented by Karl Pearson in 1901, even before the advent of chemometrics. PCA is a statistical method that reduces the dimensionality of large datasets, while aiming to preserve the most variance possible. It calculates new variables, principal components (PCs), linearly independent of one another, but that are linear combinations of the original variables. This new and simplified information can then be visualized in a plane or a higher dimensional space for easier analysis.[39,40]

In a study with the objective of discriminating samples in an unsupervised manner, PCA can be used by plotting the scores, which are the projections of the original samples onto the new PCs axes, of each sample on the relevant PCs and looking for clusters of samples that can be identified as similar. For chemically interpreting this clustering effect, it is important to study the variable’s loadings, vectors that indicate the contribution of each original variable to a given PC, therefore, it is possible to identify which variables are most strongly associated with a group of samples.[41]

3.4.2 Partial Least Squares Discriminant Analysis (PLS-DA)

The first use of partial least squares (PLS) for discriminating between samples from two classes occurred in 1987. In this study by Lars Stahle and Svante Wold, an algorithm was developed and validated by a Monte Carlo study, using the multivariate regression method and cross-validation to discriminate between a control and a test group.[42] It was only in 2003 that the partial least squares discriminant analysis (PLS-DA) method was formalized.[43]

Discriminant analysis is a technique used to classify elements of a group based on predefined information about the different possible subgroups. PLS-DA uses PLS regression and a vector of dummy variables for creating rules of classification for each class, utilizing Bayesian theory for determining a threshold between the two classes. One limitation of this method is the requirement for prior information about all samples to ensure good discrimination.[44,45]

To interpret the model, it is essential to examine its informative vectors. Two of the most used vectors for this purpose are the Variable Importance in Projection (VIP) scores and the Regression Vector. VIP Scores identify the variables that contribute most significantly to the model in absolute terms, while the Regression Vector reflects how variables are correlated with each class, positively or negatively, in the dummy variable vector. By checking these vectors, it is possible to identify which variables are most relevant for discriminating the class of interest.[46–48]

3.4.3 Soft Independent Modelling of Class Analogies (SIMCA) and Data Driven Soft Independent Modelling of Class Analogies (DD-SIMCA)

In the search for a classification method capable of empirically describing whether objects belong to a known or unknown class, the soft independent modelling of class analogies (SIMCA) was developed and introduced in 1976 by Svante Wold. This method assumes that the information of a set of similar samples follows a consistent and modellable distribution according to the Bayes theorem.[44,49]

The one-class SIMCA model uses PCA to define an enclosed class space constructed solely from the samples of the target class in the training set, eliminating the need for

representative samples from external classes. This strategy eliminates the possibility of the overlapping of two classes, in which a sample can be classified as being assigned to more than one class, and also the possibility that a sample is not located within the acceptance values of any of the modeled classes as each sample is classified as either belonging or not belonging to the target class.[44,49,50]

Another alternative for addressing problems requiring the determination of only one class was presented in data driven soft independent modelling of class analogies (DD-SIMCA) by Yuri Zontov in 2017 as a modification of the original SIMCA algorithm. In DD-SIMCA, PCA is used to define acceptance thresholds and samples will be classified by their score distances and orthogonal distances as regular, external or outliers.[51,52]

For chemically interpreting the SIMCA and DD-SIMCA models, the informative vector modelling power can be evaluated. This involves analyzing the residual variance for each variable with its corresponding total variance to determine which variables are most influential in modeling the target class space. With this information, it is possible to identify the chemical properties that differentiate the target samples from others.[53]

3.5 Data Fusion

When studying complex matrices such as coffee, a single instrumental analysis may not be able to provide fully accurate authentication of the sample. In such cases, a viable strategy is to combine multiple analytical techniques that provide complementary chemical or physical information about the sample. The data obtained from these techniques can be integrated into a single dataset to construct multivariate models that improve classification performance compared to models based on a single analytical technique.[54,55]

These strategies of data fusion can be classified at three levels: low, mid or high level, depending on the type of data processing performed before combining the matrices. In low-level data fusion, the raw data matrices from each technique are directly concatenated without prior transformation (except for preprocessing and variable selection). In mid-level or intermediate-level fusion only the most relevant variables

from each data source are selected and included in the final data matrix. In high-level data fusion separate chemometric models are built for each analysis and then their outputs are combined to form a single dataset. Each strategy has its own strengths and limitations and may be better suited for the specific type of applications at hand.[54]

3.6 Figures of Merit

When constructing multivariate models for food authentication, it is essential to validate the models to ensure their effectiveness. For qualitative methods, binary figures of merit (FOM) related to accuracy (ACC) are the most used for this validation. Among these FOMs, the European Commission defines sensitivity rate (STR) and specificity rate (SPR) as the most important ones. STR is a FOM related to the false negative rate and measures the proportion of actual positive samples that are correctly identified by the model, while SPR is related to the false positive rate and quantifies the proportion of negative samples that are correctly classified. Another FOM used for accuracy evaluation is the efficiency rate (EFR), defined as 1 - (STR + SPR), which offers a balanced overview of the model's accuracy across both classes.[56,57]

To calculate these metrics, a confusion matrix is typically employed. In this matrix, each column represents the true class assignment of the samples, while each row corresponds to the predicted class assignment. In an ideal model, all samples are correctly classified, forming a diagonal matrix. The numerical values of the matrix can then be used to determine the model's trueness and to calculate the corresponding accuracy rates.[44,56]

4 METHODOLOGY

4.1 Samples

The medium roasted commercial coffee samples were obtained from producers in four different states: Minas Gerais, Espírito Santo, Rondônia, and Amazonas, with 30 samples collected from each state. The samples were grounded using a Hamilton Beach 80393-BZ127 domestic coffee grinder and stored under refrigeration until analysis. The grinder was cleaned and dried between each sample to prevent cross contamination. For each state, it was assigned a two-letter code, combining the coffee variety with the state's demonym, along with a symbol for use in the chemometric models, as shown in Table 1.

Table 1. Samples codes and symbols.

State	Variety	Demonym	Code	Samples	Symbol
Minas Gerais	Conilon	Mineiro	CM	30	▲
Espírito Santo	Conilon	Capixaba	CC	30	★
Rondônia	Robusta	Rondoniense	RR	30	■
Amazonas	Robusta	Amazonense	RA	30	●

4.2 Differential Scanning Calorimetry (DSC)

DSC curves were obtained with a DSC-60 from Shimadzu. The analyses were performed using samples masses of about 1.75 mg (±0.25 mg) in aluminum crucible with heating rate of 10 °C min^{-1}, under nitrogen atmosphere at flow 50 mL min^{-1} and at the temperature range between 40 ºC and 400 °C. These conditions of heating rate

and temperature range were defined as optimal for the time of analysis and resolution of peaks after preliminary tests realized.

4.3 Near Infrared Spectroscopy (NIR)

NIR spectra were obtained using a portable MicroNIR® (Viavi), in the wavelength between 908 nm and 1676 nm, with a resolution of 6 nm. The analyses were performed inside a 2 mL glass vial held by a 3D printed support designed by the author and his research group. The spectra were acquired using diffuse reflectance mode and the equipment lamp was maintained at temperature between 40 °C and 50 °C. For each sample, 20 scans were obtained with a time of 1 second between scans.

4.4 Chemometric Modeling

Chemometric analyses were performed using MATLAB version 7.11.0 (MathWorks Inc), PLS_Toolbox version 5.2.2 (Eigenvector Co.) and DD-SIMCA toolbox version 1.2 (Zontov, et al. 2017).[51]

4.4.1 Data Preprocessing

The DSC curves and NIR spectra were imported to MATLAB using routines written by the author, presented in Appendix A1 and A2, respectively.

The DSC curves were normalized by the sample mass. Increments of 0.1 °C were selected and missing points were completed with the mean of the values before and after. The points between 200 °C and 400 °C were selected for the model, totalizing 2001 variables. The initial temperature utilized was chosen since no peaks were observed before 200 ºC, as it is the commercial roasting temperature of coffee. The curves were then pre-processed with Savitzky-Golay (SG) filter, standard normal variate (SNV) and mean center (MC).

The data obtained from the 20 scans of the NIR spectra was converted from transmittance values to absorbance using Equation 1. The data obtained was then pre-processed with Savitzky-Golay (SG) filter, standard normal variate (SNV) and mean center (MC).

$$A = 2 - \log(R) \quad \text{Eq. (1)}$$

4.4.2 Data Fusion

For data fusion models, the low-level approach was selected. Data from each analysis was preprocessed individually and then all variables from both techniques were concatenated in a single matrix and autoscaled with the DSC variables first, ranging for variable scale from 1 to 2001 and NIR variables second, ranging from 2002 to 2126.

4.4.3 Principal Component Analysis (PCA)

Exploratory analyses were performed using PCA to differentiate between the conilon coffee from Mutum and the canephora coffee samples from other Brazilian States analyzed.

The datasets for the PCA models were built containing all the 120 samples, divided into four classes, one for each state. The optimal number of PCs for PCA models was chosen based on the captured variance.

To further elucidate the separation of the classes, PCA loadings were evaluated to determine the contribution of each variable to the models.

4.4.4 Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA models were built for discriminating between the samples from the city of Mutum, in the Brazilian State of Minas Gerais and the other producer regions.

The datasets for the PLS-DA models were built using the routine written by the author, presented in Appendix A3. Each class was divided between training and test set using the Kennard-Stone algorithm with two thirds of the samples being selected for the training set.[58] A non-binary approach, PLS2-DA was chosen for these analyses, modeling representative samples from each class on the training set. The results are displayed, however, on a binary form, with samples predicted by the model as being from Minas Gerais being placed above the threshold. The optimal number of latent variables was chosen by random subsets cross-validation procedure.

To further interpretate the discrimination by the models, the variable importance in projection (VIP) scores and the regression vectors were plotted. VIP scores identify the variables that contribute most significantly to the model, in absolute terms, while regression vector reflects how variables are correlated with each class.

4.4.5 Soft Independent Modelling of Class Analogies (SIMCA) and Data Driven Soft Independent Modelling of Class Analogies (DD-SIMCA)

SIMCA and DD-SIMCA models were built for authenticating the samples from the city of Mutum, Minas Gerais against the other producer regions.

The datasets for the SIMCA and DD-SIMCA models were built using the routines written by the author, presented in Appendix A4 and A5, respectively, following a rigorous approach, in which only samples from the target class are used for the modelling in the training set, in opposition to the compliant approach, in which representative samples from both target and non-target classes are utilized in this step. The samples from Minas Gerais were divided between training and test set using the Kennard-Stone algorithm with two thirds of the samples being selected for the training set.[58] All samples from the other states were placed in the test set. The optimal number of PCs was chosen by random subsets cross-validation.

To further interpretate the models, the modelling power of the variables were plotted using the routine written by the author, presented in Appendix A6. The modelling power determines which variables are the most influential in modeling the target class space.

4.4.6 Outlier detection

The outlier detection was performed by selecting the samples with Hotelling's T^2 and Q residuals above the 95% confidence limit. For each model, up to three rounds of outlier detection were conducted and a maximum of 22.2% of the samples of each class was removed.[59]

4.4.7 Qualitative Validation

For each supervised model built, FOMs based on the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) rates were estimated. These FOMs include the sensitivity rate (STR), specificity rate (SPR) and efficiency rate (EFR) and are related to the accuracy (ACC) of the qualitative models.

5 RESULTS AND DISCUSSION

5.1 Differential Scanning Calorimetry (DSC)

5.1.1 DSC Curves

The grounded coffee samples were analyzed by DSC and the curves were normalized by the mass of each sample and preprocessed with Savitsky-Golay filter, using a 11-point filter width and SNV. The DSC curves of the coffee samples before and after preprocessing can be seen in Figure 1 and 2, respectively. Figure 3 shows average DSC curves of the coffee samples from each different Brazilian State. The curves were also mean centered for building chemometric models.

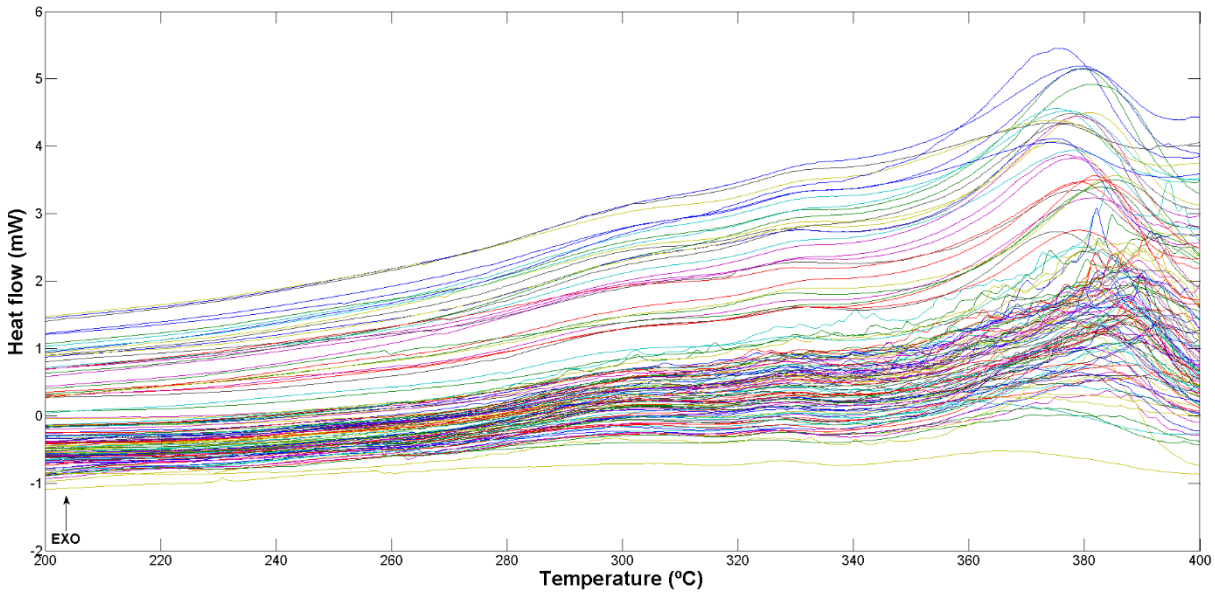

Figure 1. DSC curves of coffee samples before preprocessing.

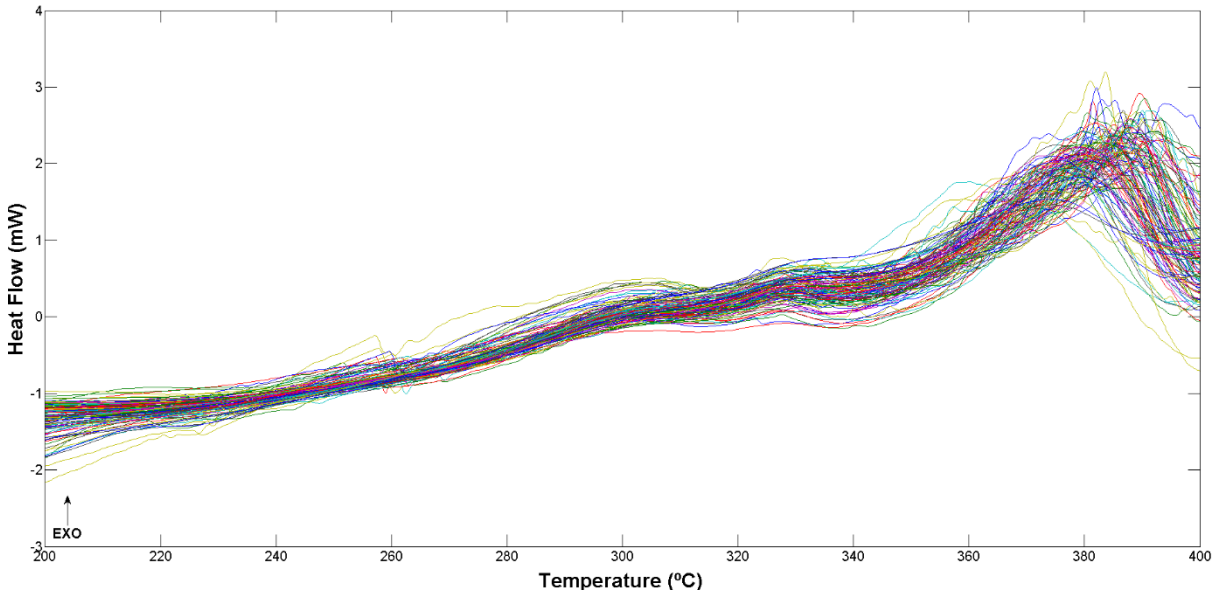

Figure 2. DSC curves of coffee samples after preprocessing.

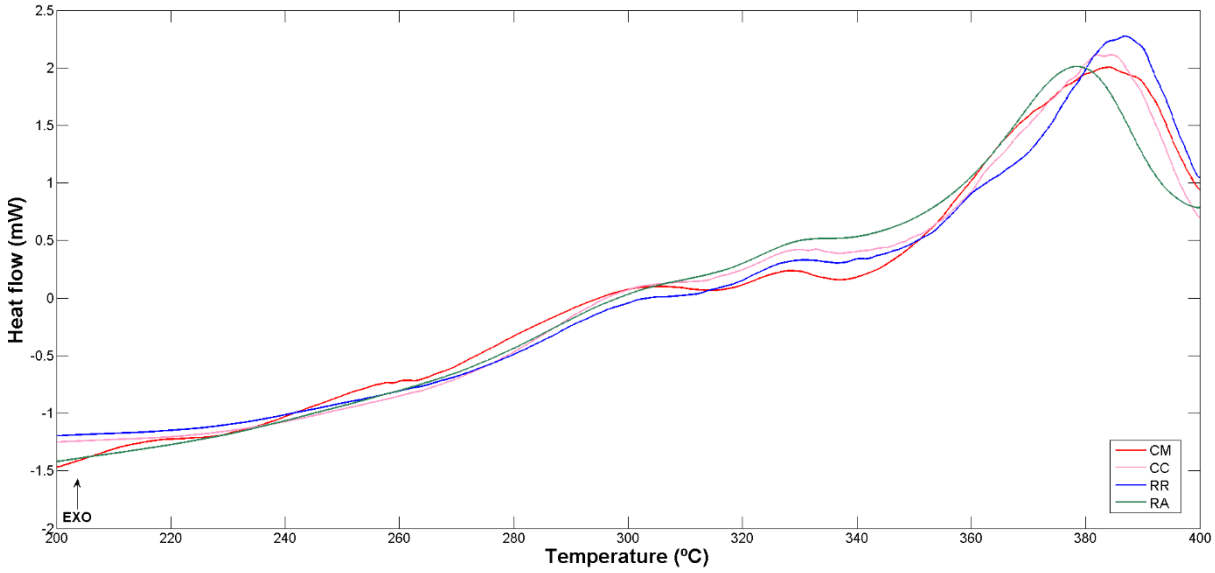

Figure 3. Average DSC curve of coffee samples from each different Brazilian State.

After analyzing the DSC curves of the coffee samples, it is possible to identify three main regions of exothermic events, one between 250 °C and 310 °C, another between 310 °C and 350 °C and the third between 350 °C and 400 °C. These peaks can be associated with the decomposition of the sample and the release of volatile components, but, as the coffee samples are complex matrixes, it is not possible to identify specific compounds that generate each event seen in the DSC curves without

the use of a coupled technique, such as infrared spectroscopy or mass spectrometry, used as a detector.[1]

The use of DSC is justified, however, particularly because the analysis is performed at temperatures exceeding the typical roasting temperatures of commercial coffees, thereby minimizing the influence of the manufacturing process on sample discrimination.

5.1.2 Principal Component Analysis (PCA)

The PCA scores plot of PC1 *versus* PC3 for the DSC analysis can be seen in Figure 4. The model was built with 3 PCs and 88.00% accumulated variance captured. PC2 was not deemed relevant for the differentiation between coffees from Mutum and other regions and its scores plot can be seen in Appendix B.

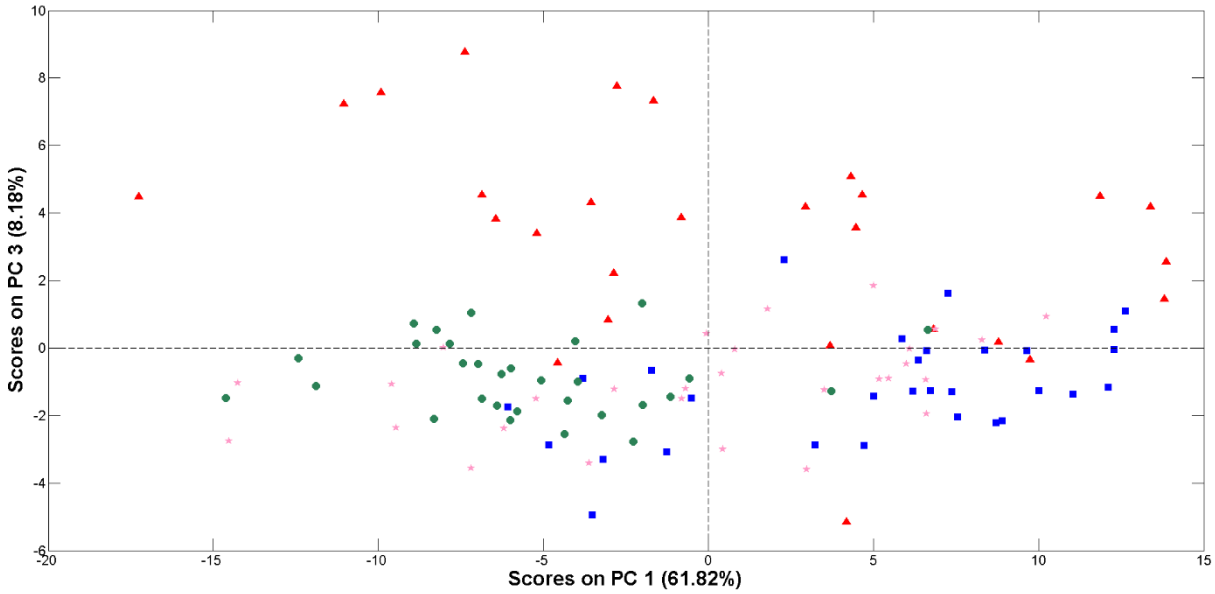

Figure 4. PCA scores of PC1 *versus* PC3 for the canephora coffee samples analyzed with DSC. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.

It is possible to see a good separation between the samples from Minas Gerais, in the positive region of PC3 (8.18%), and the samples from other regions, in the negative region.

The PCA loadings for PC1 and PC3 can be seen in Figure 5 below.

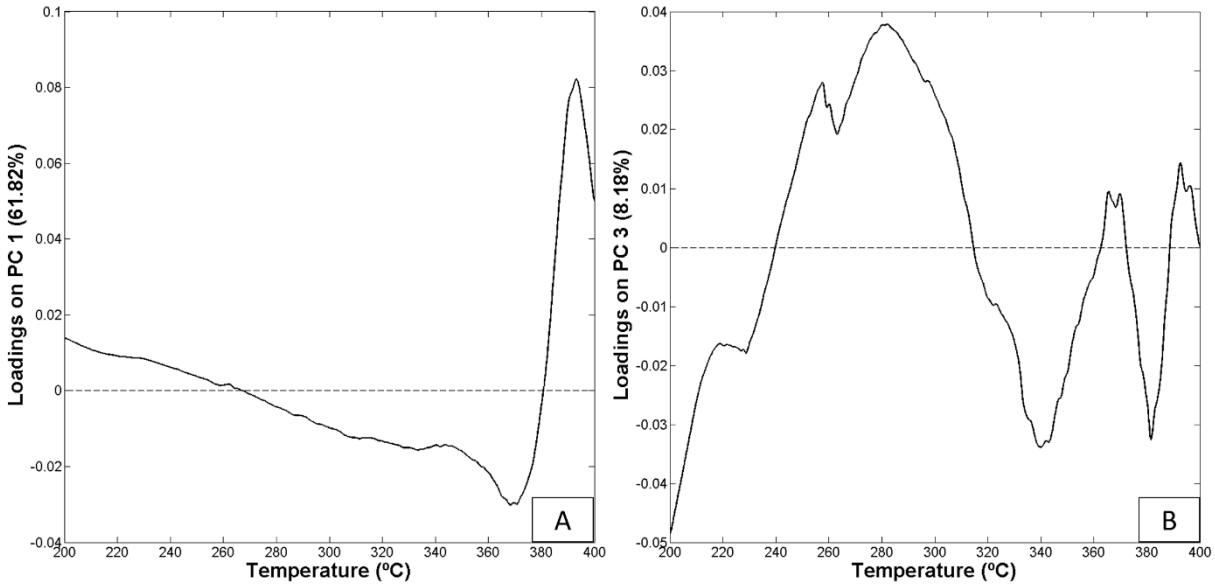

Figure 5. Loadings in PC1 (A) and PC3 (B) for the PCA model built with DSC data.

The loadings of PC1, seen in Figure 5A, show that the region of events between 350 °C and 400 °C is the most relevant for PC1, contributing mostly for the negative part of the PC, while the end of the curve contributes positively. The loadings of PC3, seen in Figure 5B shows that the region between 250 °C and 310 °C contributes positively for the PC3, while the region between 310 °C and 350 °C contributes negatively. The region between 350 °C and 400 °C is split between the two parts of the PC. These contributions can be associated with the separation between the samples, with the events occurring between 250 °C and 310 °C being more relevant for the coffee samples from Minas Gerais.

5.1.3 Partial Least Squares Discriminant Analysis (PLS-DA)

The results for the PLS-DA model built using DSC data can be seen in Figure 6. The model was built with 7 LVs, capturing 96.77% accumulated variance in the X block and 61.30% accumulated variance in the Y block.

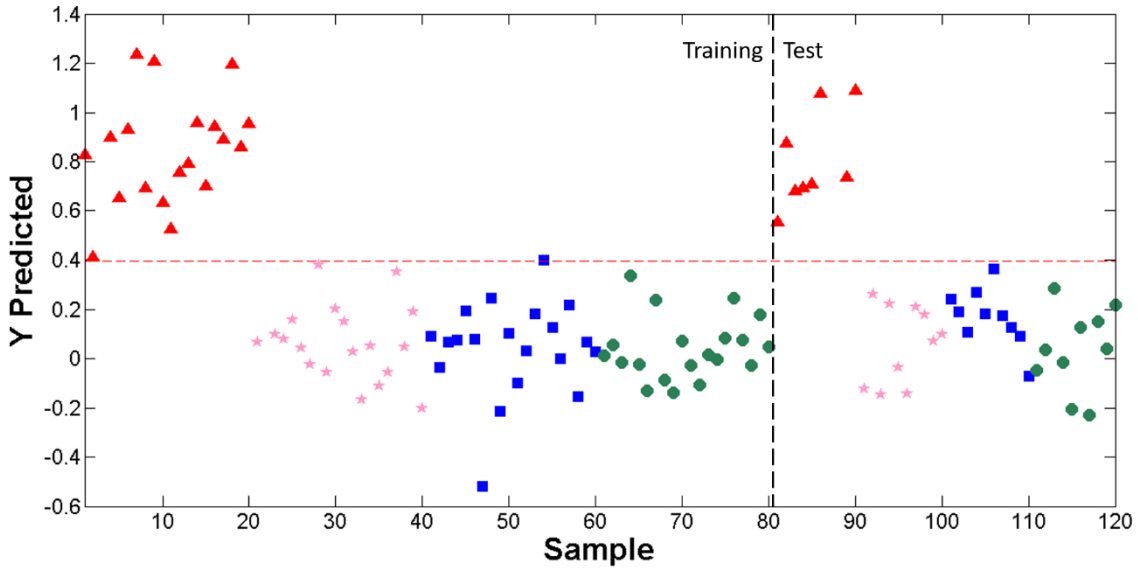

Figure 6. PLS2-DA predicted classes for samples of different Brazilian States analyzed with DSC. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA. The horizontal dashed line indicates the estimated threshold for predictions. The vertical dashed line indicates the separation between training and test samples.

The model presented a good discrimination, with only one false positive on the training set and no error in the test set. The VIP scores and the regression vector can be seen in Figure 7.

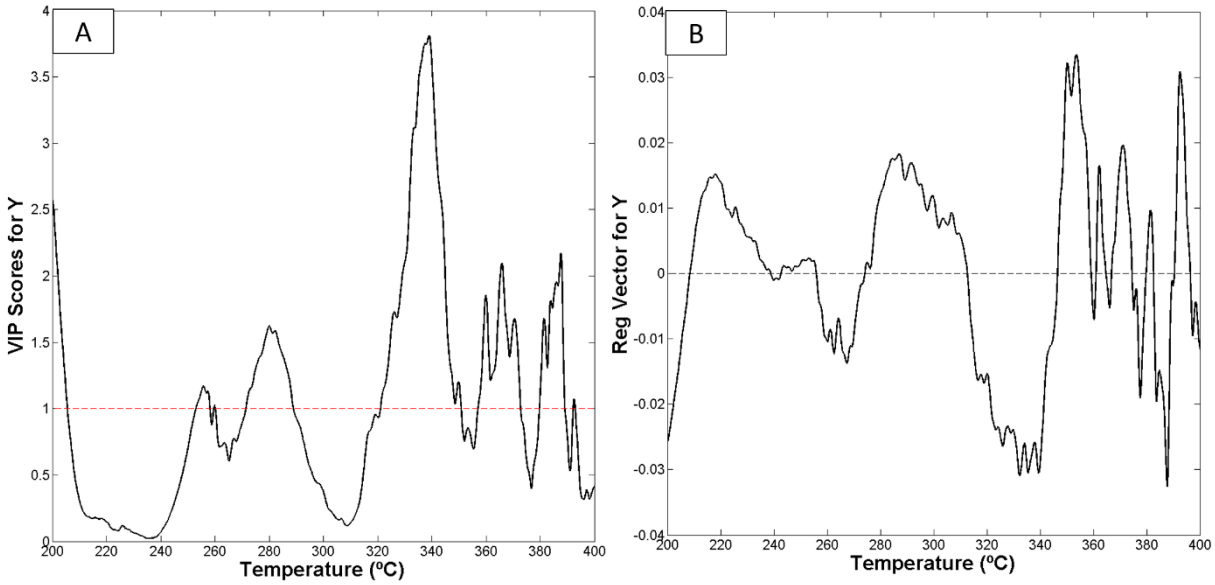

Figure 7. Informative vectors VIP scores (A) and regression vector (B) for the PLS2-DA model built with DSC data.

The VIP scores plot, seen in Figure 7A, shows that the region of events between 310 °C and 350 °C is the most important for the modeling of coffees from Minas Gerais, and the regions between 250 °C and 310 °C and between 350 °C and 400 °C have similar importances. The regression vector, seen in Figure 7B, shows that the region between 250 °C and 310 °C contributes especially to the samples above the threshold, while the region between 310 °C and 350 °C contributes for the samples below the threshold. The region between 350 °C and 400 °C has an intermediate behavior between these two previous regions.

5.1.4 Soft Independent Modelling of Class Analogies (SIMCA)

The results for the SIMCA model built using DSC data can be seen in Figure 8. The model was built with 7 PCs and accounted for 98.87% of the total variance.

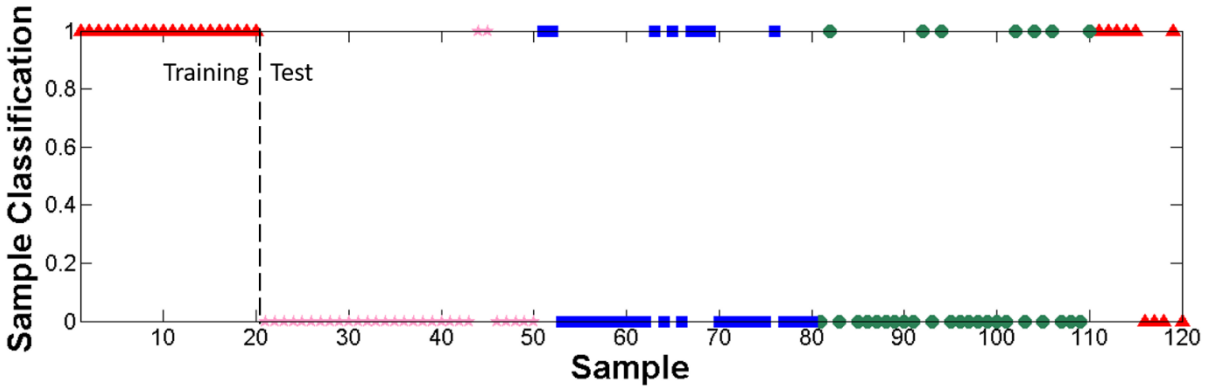

Figure 8. Class predictions for SIMCA model built with samples of different Brazilian States analyzed with DSC. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA. The vertical dashed line indicates the separation between training and test samples.

The training set of the model presented no errors, while the test set presented 17 false positives and 4 false negatives. The modelling power vector of the variables is presented in Figure 9.

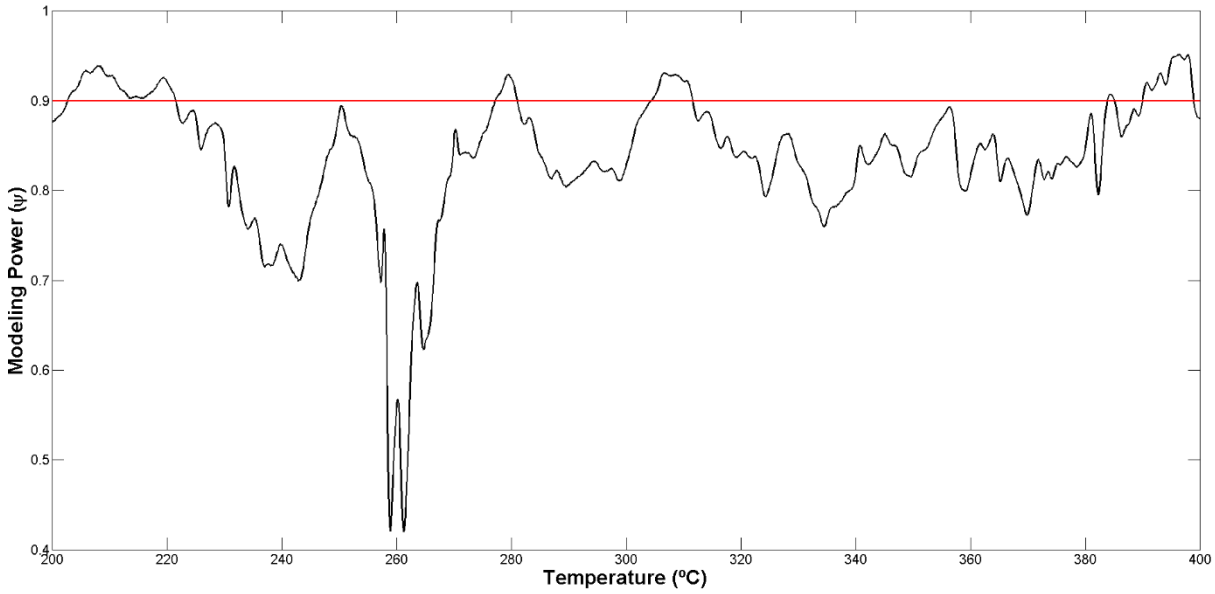

Figure 9. Modeling power vector for the SIMCA model built with DSC data.

The modelling power of the variables shows that the events in 280 °C, 310 °C and 395 °C were considered by the model as the most important ones for the separation between classes.

5.1.5 Data Driven Soft Independent Modelling of Class Analogies (DD-SIMCA)

The acceptance plots for the DD-SIMCA model built using the DSC data can be seen in Figure 10. The model was built with 7 PCs and accounted for 98.87% of the total variance.

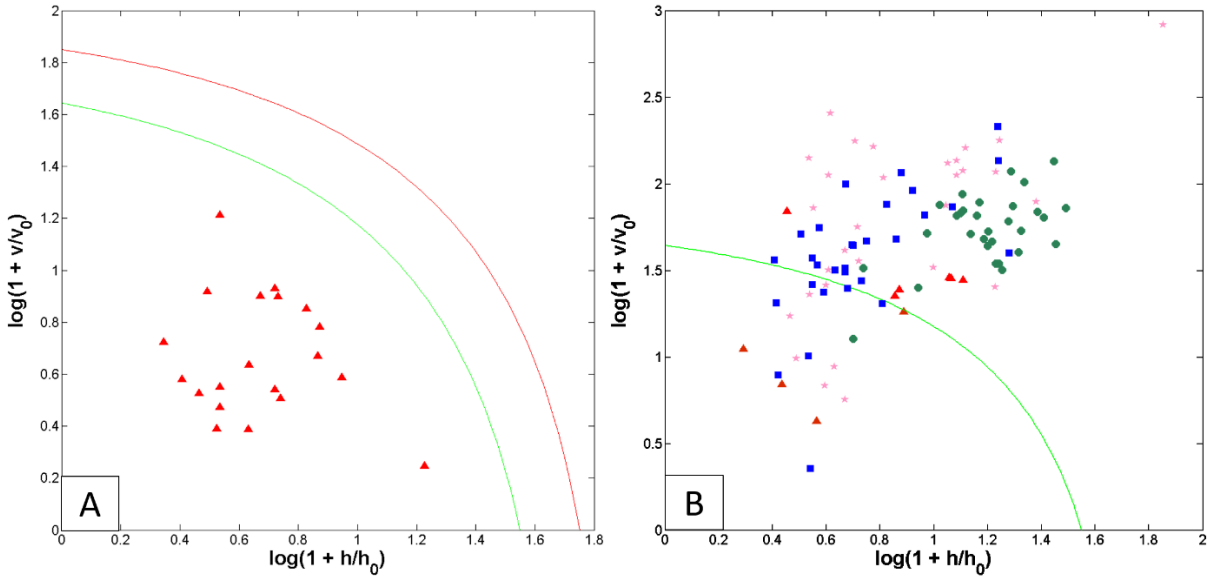

Figure 10. Acceptance plot for training (A) and test sets (B) of the DD-SIMCA model built with samples of different Brazilian States analyzed with DSC. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.

The training set of the model presented no errors, while the test set presented 9 false positives and 4 false negatives, a result similar to that obtained with the SIMCA model. The modelling power vector is presented in Figure 11.

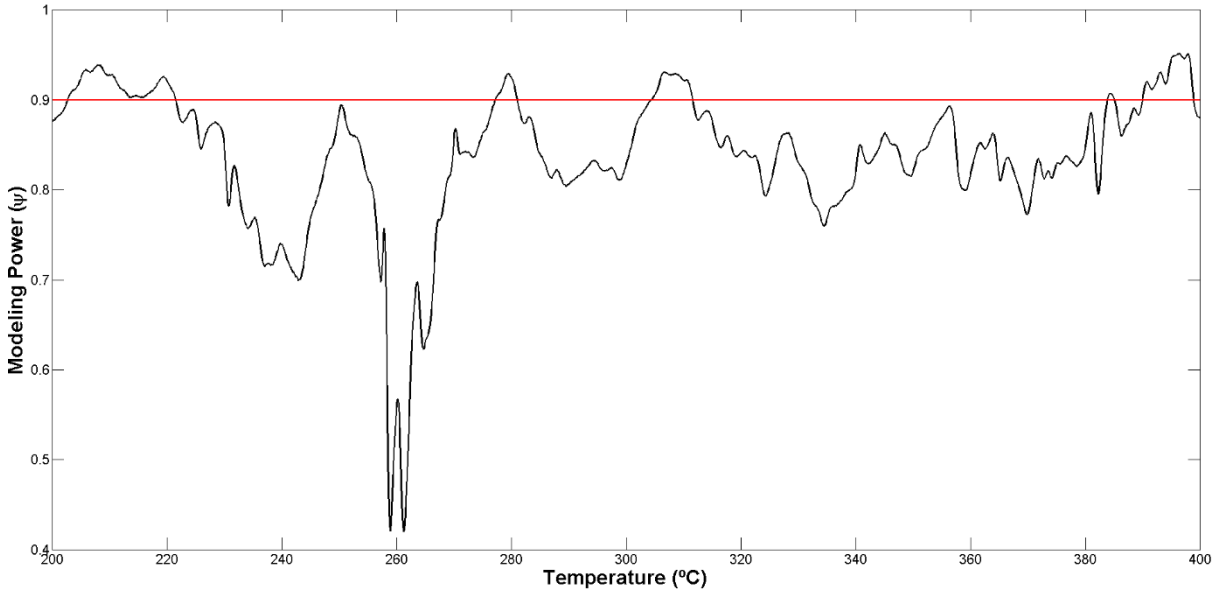

Figure 11. Modeling power vector for the DD-SIMCA model built with DSC data.

The modelling power vector shows that the same regions of events as those previously indicated for the SIMCA model were considered by DD-SIMCA as the most important ones for the authentication of the coffee from Minas Gerais.

5.2 Near Infrared Spectroscopy (NIR)

5.2.1 NIR Spectra

The grounded coffee samples were analyzed by NIR spectroscopy and the spectra were preprocessed by Savitsky-Golay filter, using a 11-point filter width, and SNV. The NIR spectra of coffee samples before and after preprocessing can be seen in Figure 12 andFigure 13, respectively. Figure 14 shows the average NIR spectrum of the coffee samples from each different Brazilian State. The spectra were also mean centered for the use in the chemometric models.

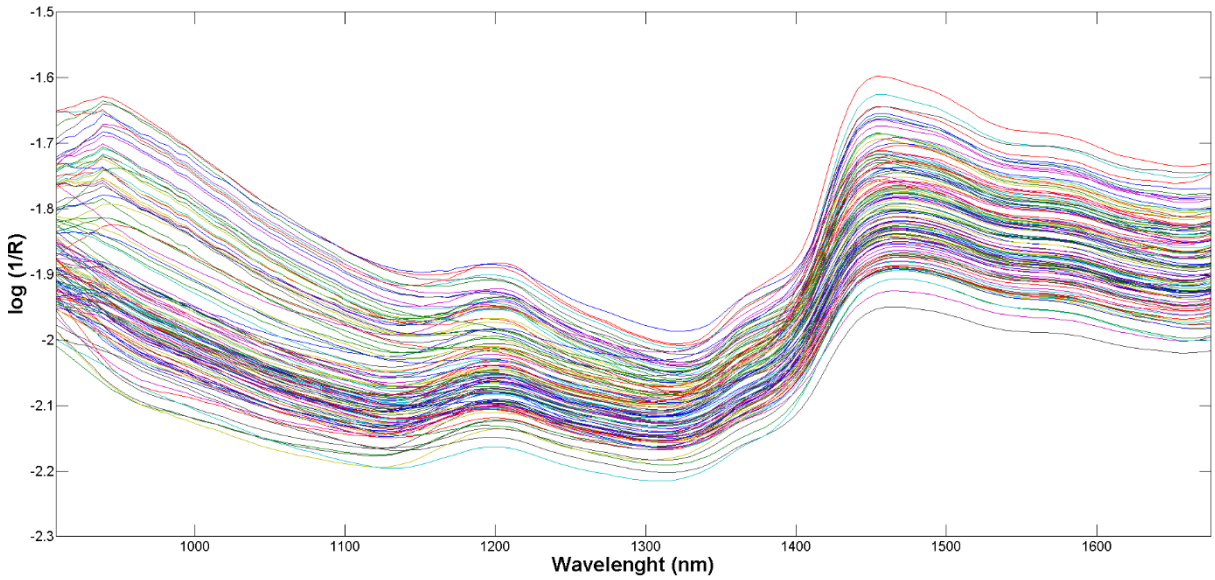

Figure 12. NIR spectra from coffee samples before preprocessing.

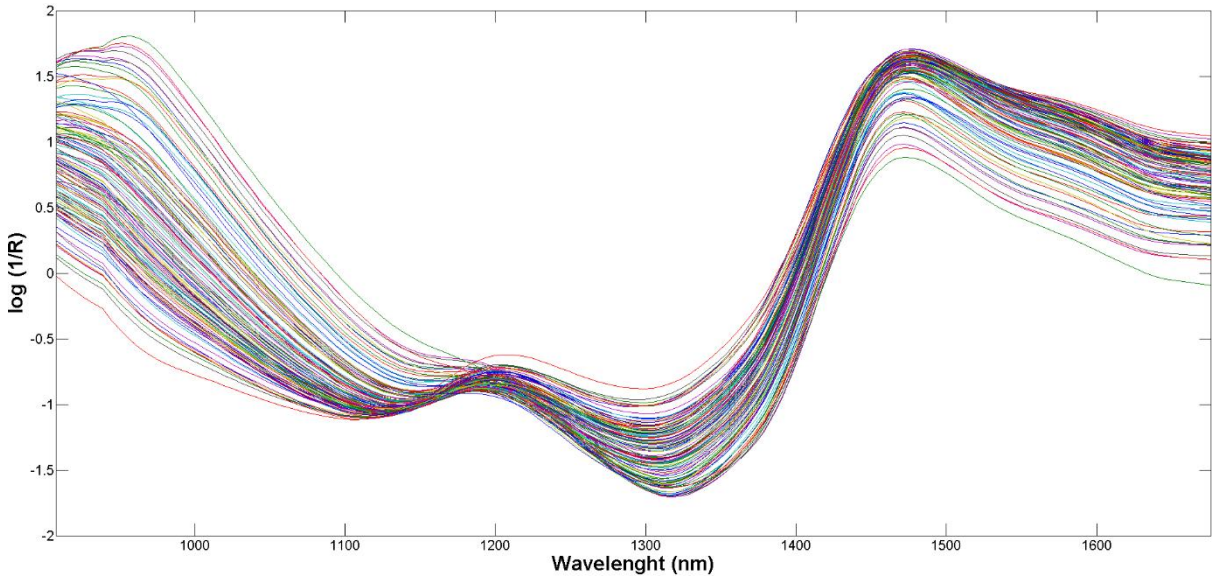

Figure 13. NIR spectra from coffee samples after preprocessing.

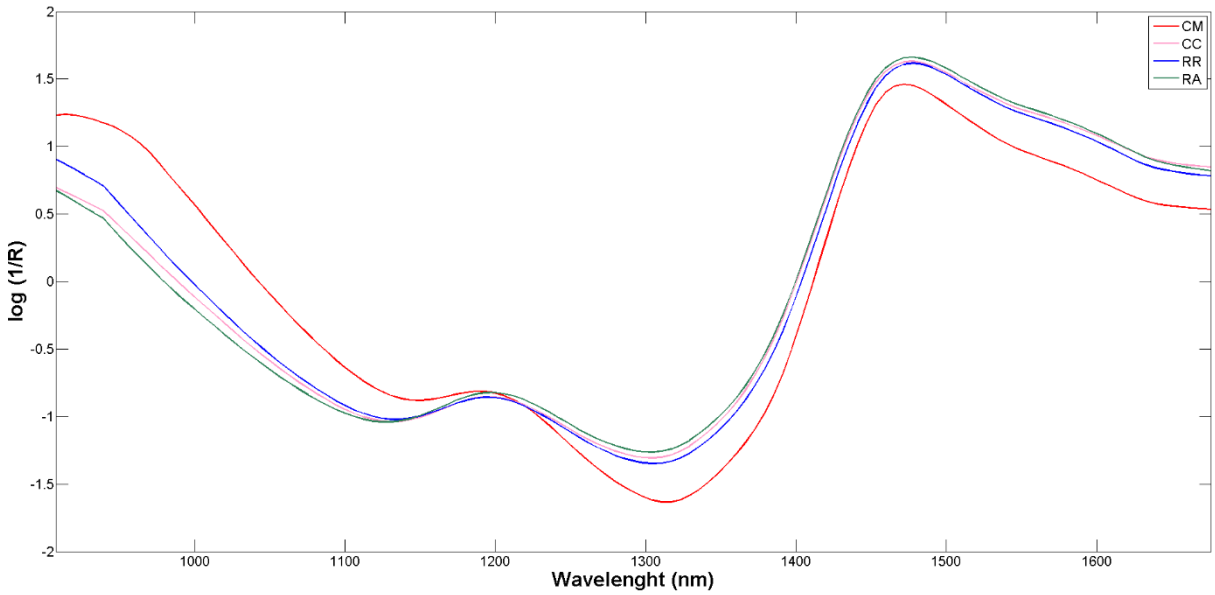

Figure 14. Average NIR spectra of coffee samples from each different Brazilian State.

By observing NIR spectra of the coffee samples, three absorption bands can be identified around 920 nm, 1200 nm and 1450 nm. The band around 920 nm can be correlated to the third overtone of H_2O, the 1200 nm spectral band can be associated with the second overtone of CH stretching and the 1450 nm band can be assigned to the second overtone of ROH and H_2O[60]. NIR spectroscopy is a technique with limited selectivity, and these spectral attributions cannot be directly associated with any particular component of the coffee samples like caffeine or chlorogenic acids.

Figure 15 shows the average spectra of coffee samples from Mutum with different roast levels.

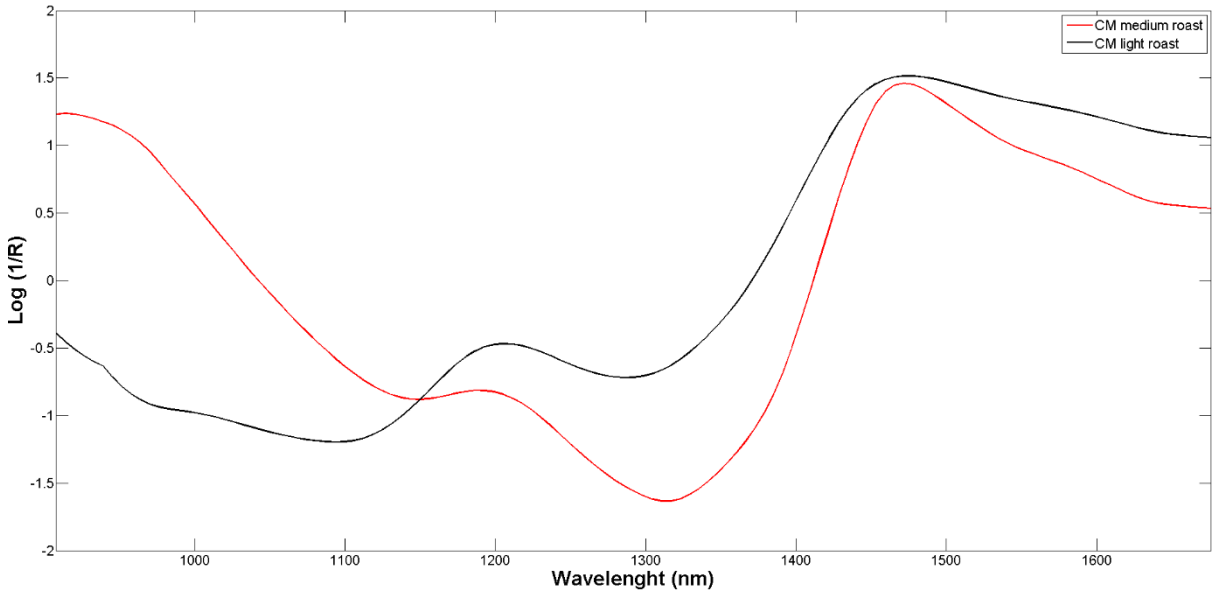

Figure 15. Difference between the NIR spectra of the coffee samples with different roast levels.

This figure indicates that the light roast samples exhibit a relatively lower absorption band around 920 nm and a higher absorption band around 1200 nm compared to the medium roast samples. This observation demonstrates that the NIR spectra is influenced by the roast level of the coffee, suggesting that the manufacturing process can affect the spectral characteristics and, consequently, the discrimination or the authentication of the samples.

5.2.2 Principal Component Analysis (PCA)

The PCA scores plot of samples *versus* PC1 for the NIR analysis can be seen in Figure 16. The model was built with only one PCs and 96.63% accumulated variance captured.

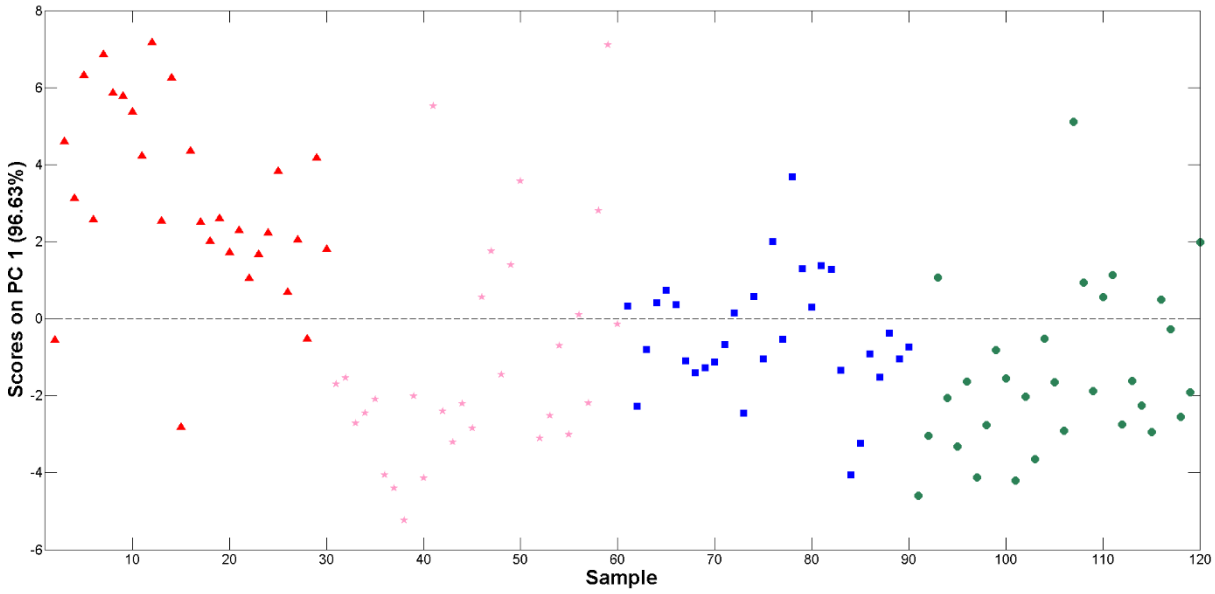

Figure 16. PCA scores of PC1 *versus* sample for the canephora coffee samples analyzed with NIR. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.

It is possible to see a tendency of separation between samples from Minas Gerais, especially in the positive region of PC1, and the samples from other regions, especially on the negative region of PC1. The loadings for PC1 can be seen in Figure 17 below.

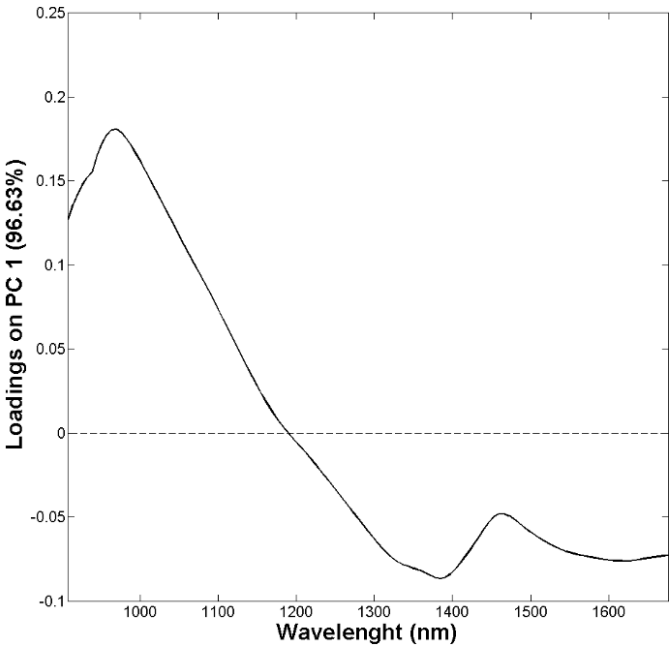

Figure 17. Loadings in PC1 for the PCA model built with NIR data.

The loadings of PC1, seen in Figure 17A, show that the third overtone of H_2O band contributes positively for PC1, while the spectral bands of the second overtones of CH, ROH and H_2O contribute negatively for this PC.

5.2.3 Partial Least Squares Discriminant Analysis (PLS-DA)

The results for the PLS-DA model built using the NIR data can be seen in Figure 18. The model was built with 6 LVs, capturing 99.97% accumulated variance in the X block and 55.94% accumulated variance in the Y block.

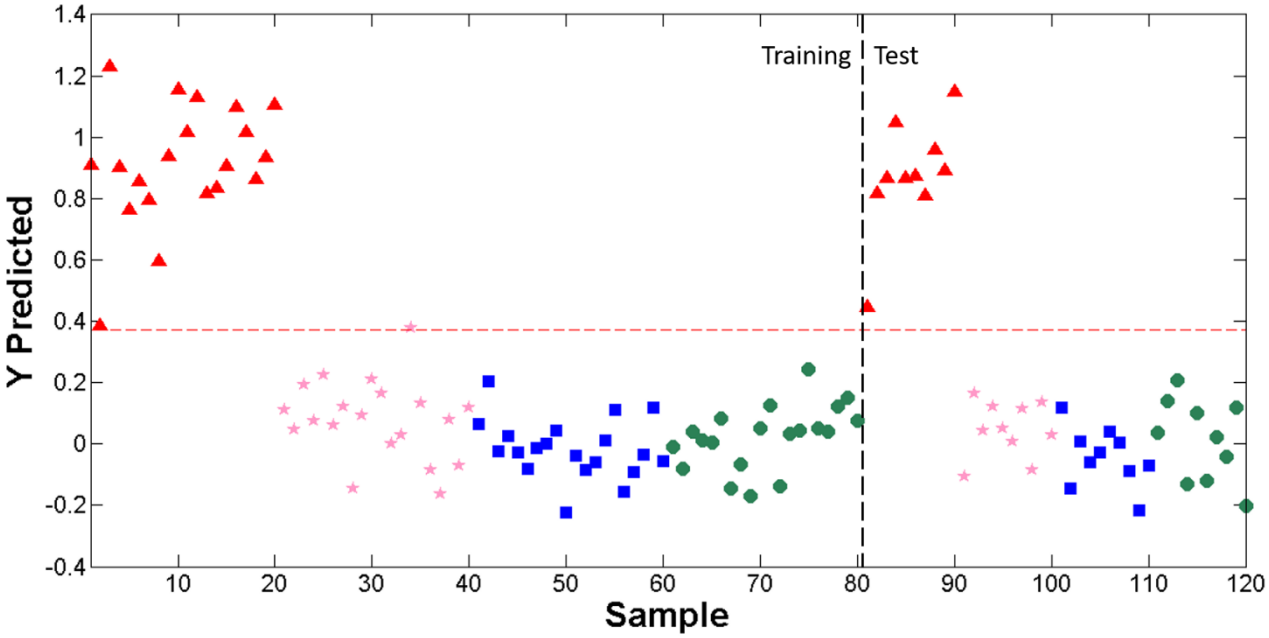

Figure 18. PLS2-DA predicted classes for samples of different Brazilian States analyzed by NIR spectroscopy. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA. The horizontal dashed line indicates the estimated threshold for predictions. The vertical dashed line indicates the separation between training and test samples.

The model presented almost perfect discrimination, with only one false positive in the training set and no error in the test set.

The VIP scores and the regression vectors of the variables can be seen in Figure 19.

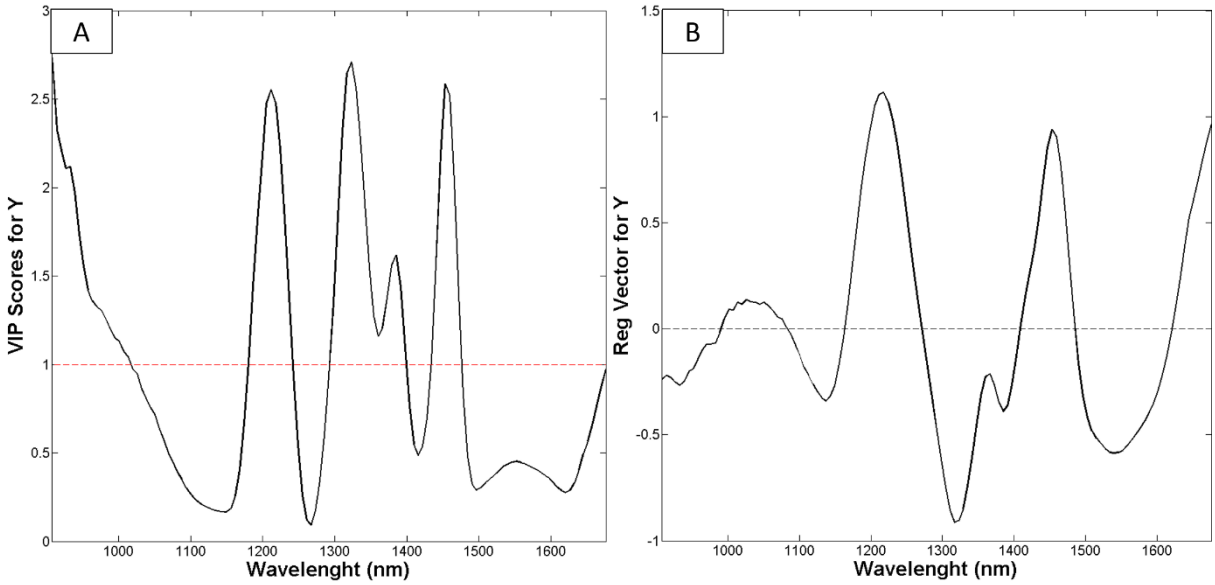

Figure 19. Informative vectors VIP Scores (A) and Regression coefficients (B) for the PLS2-DA model built with NIR data.

The VIP Scores plot, seen in Figure 19A shows that all three spectral bands at 920 nm, 1200 nm and 1450 nm have similar importance for the discriminant power of the model. The regression vector, seen in Figure 19B, shows that the third overtone of H_2O band contributes to the samples from other states (below the threshold), while bands assigned to second overtones of CH, ROH and H_2O contribute for the target samples, predicted above the threshold.

5.2.4 Soft Independent Modelling of Class Analogies (SIMCA)

The SIMCA model built using NIR data can be seen in Figure 20. The model was built with 4 PCs and accounted for 99.91% of the total variance.

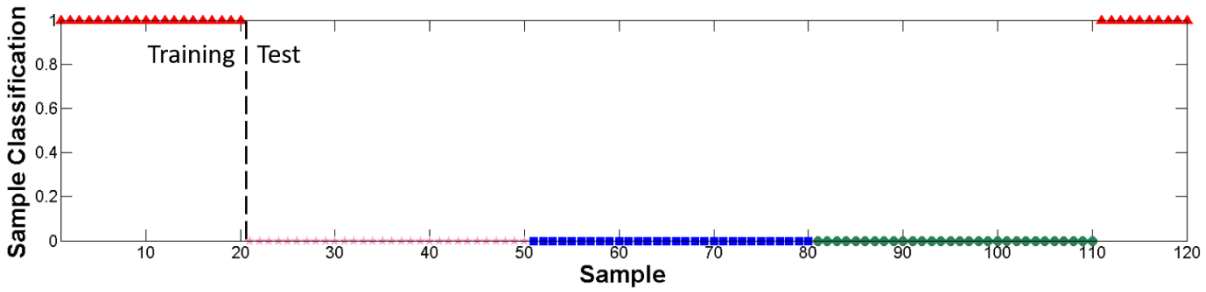

Figure 20. Class predictions for SIMCA model built with samples of different Brazilian States analyzed by NIR spectroscopy. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA. The vertical dashed line indicates the separation between training and test samples.

Both training and validation sets presented no false positive or false negative errors. The modelling power vector is presented in Figure 21.

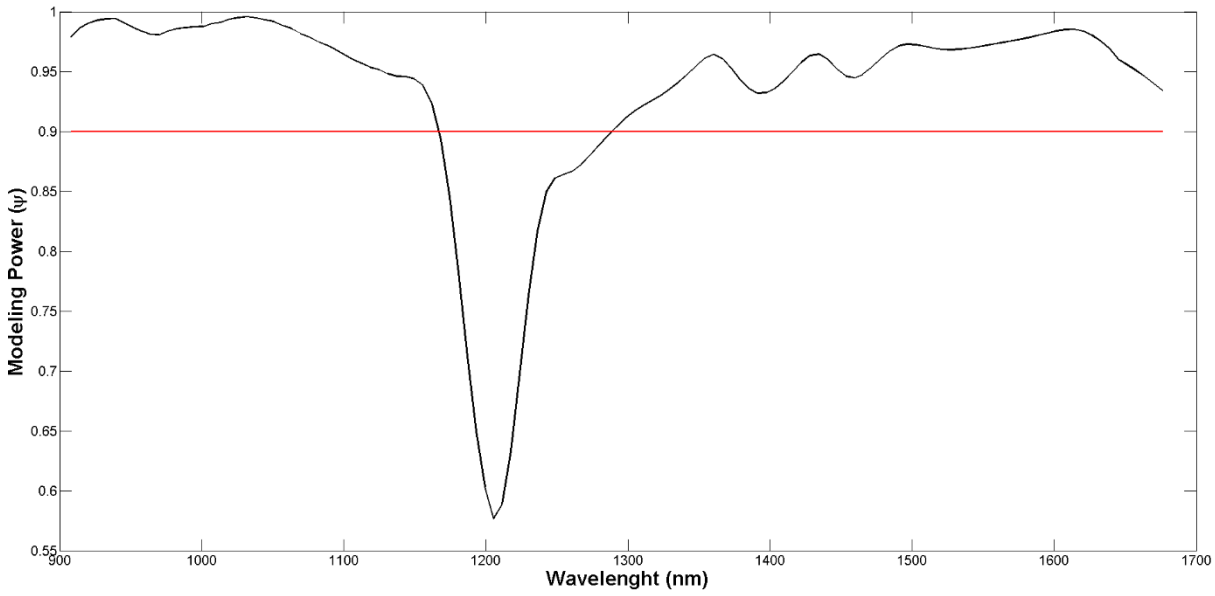

Figure 21. Modeling Power for the SIMCA model built with NIR data.

The modelling power of the variables shows that NIR band assigned to second overtone of CH was not considered relevant by the model for the authentication, while the bands of the third overtone of H_2O and second overtone of H_2O and ROH were considered important.

5.2.5 Data Driven Soft Independent Modelling of Class Analogies (DD-SIMCA)

The acceptance plots for DD-SIMCA model built using NIR data can be seen in Figure 22. The model was built with 4 PCs and accounted for 99.91% of the total variance.

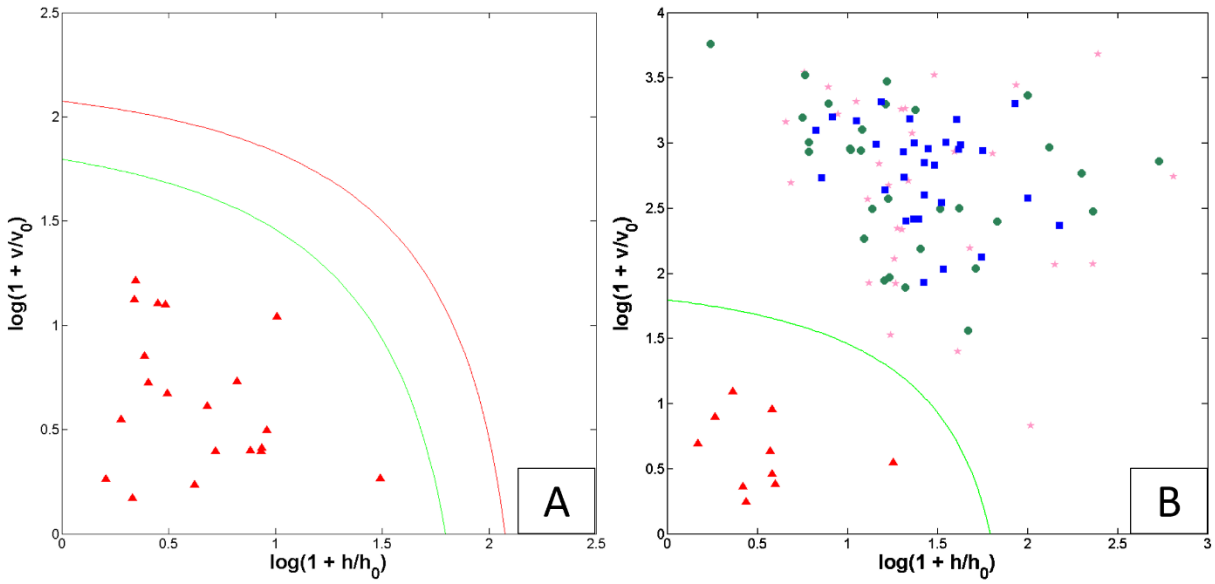

Figure 22. Acceptance plot for training (A) and test sets (B) of DD-SIMCA model built with samples of different Brazilian States analyzed by NIR spectroscopy. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.

Both training and validation sets presented no false positive or false negative errors, a result identical to that obtained with the SIMCA model. The modelling power vector is presented in Figure 23.

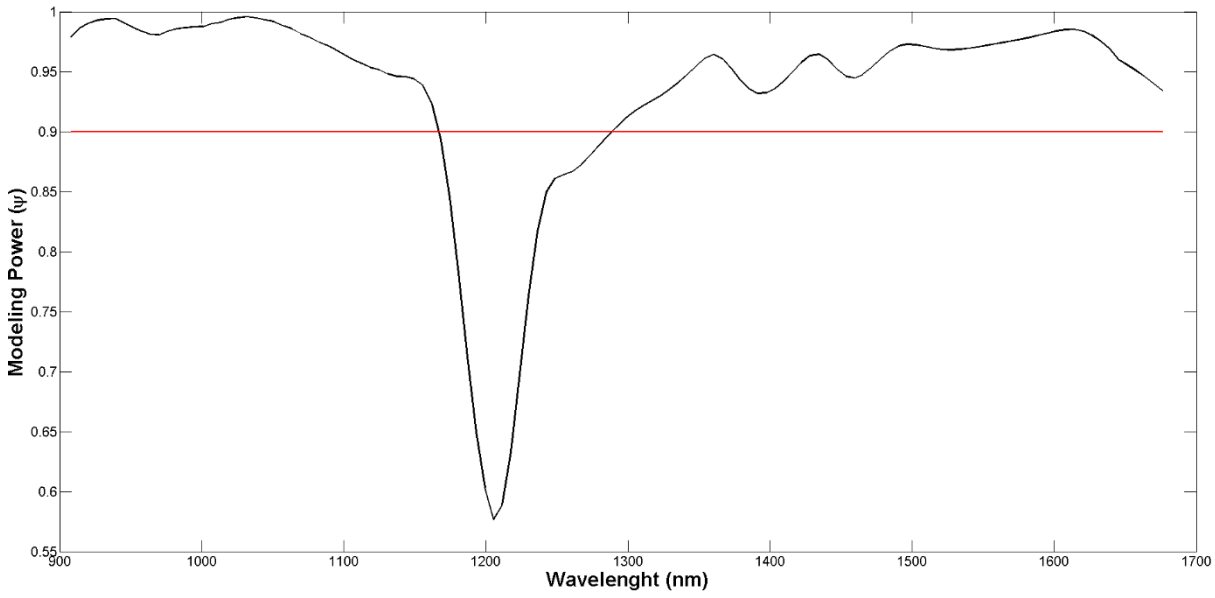

Figure 23. Modeling power vector for the SIMCA model built with NIR data.

The modelling power of the variables shows that the band of the second overtone of CH was not considered relevant by the model for authentication, while bands assigned to the third overtone of H_2O, to the second overtone of H_2O and to ROH were considered important. This result is similar to that obtained with SIMCA model.

5.3 Data Fusion

5.3.1 Principal Component Analysis (PCA)

The PCA scores plot of PC3 *versus* PC4 for fused DSC and NIR data can be seen in Figure 24. The model was built with 4 PCs and 92.73% accumulated variance captured. PC1 and PC2 were not deemed relevant for the differentiation between coffees from Mutum and other regions and the respective scores plot can be seen in Appendix B.

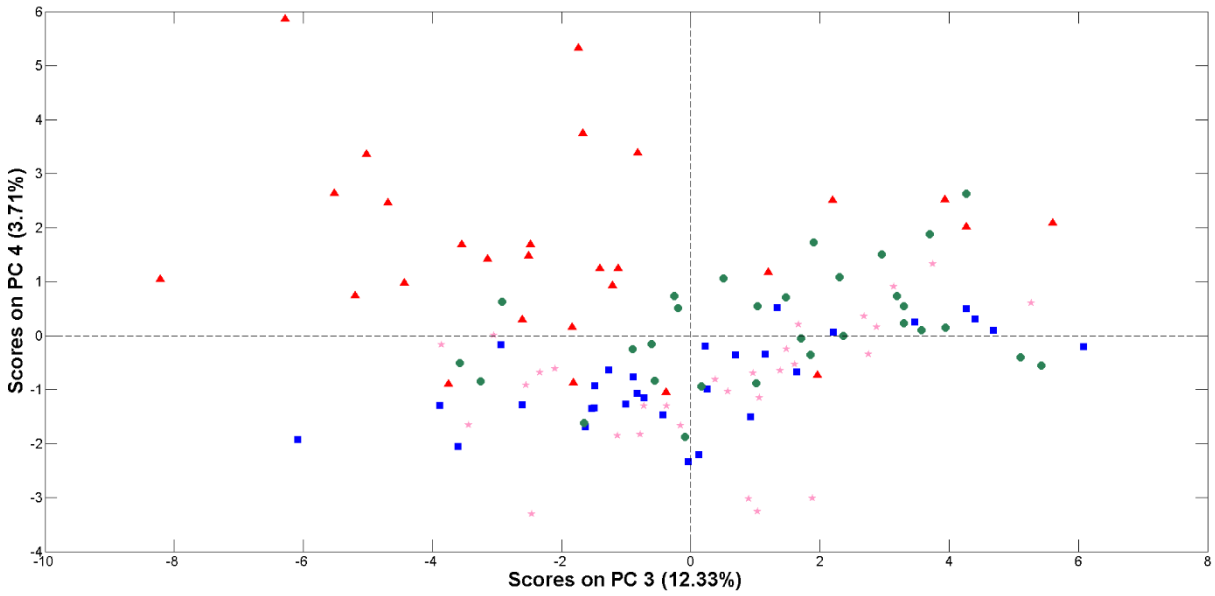

Figure 24. PCA scores plot of PC3 *versus* PC4 for the canephora coffee samples analyzed by the data fusion model DSC-NIR. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.

It is possible to see a good separation between the Mineiro samples, especially in the positive region of PC4 and negative region of PC3, while the samples of other States

are dispersed in other regions of the plot. The loadings for PC3 and PC4 can be seen in Figure 25 below.

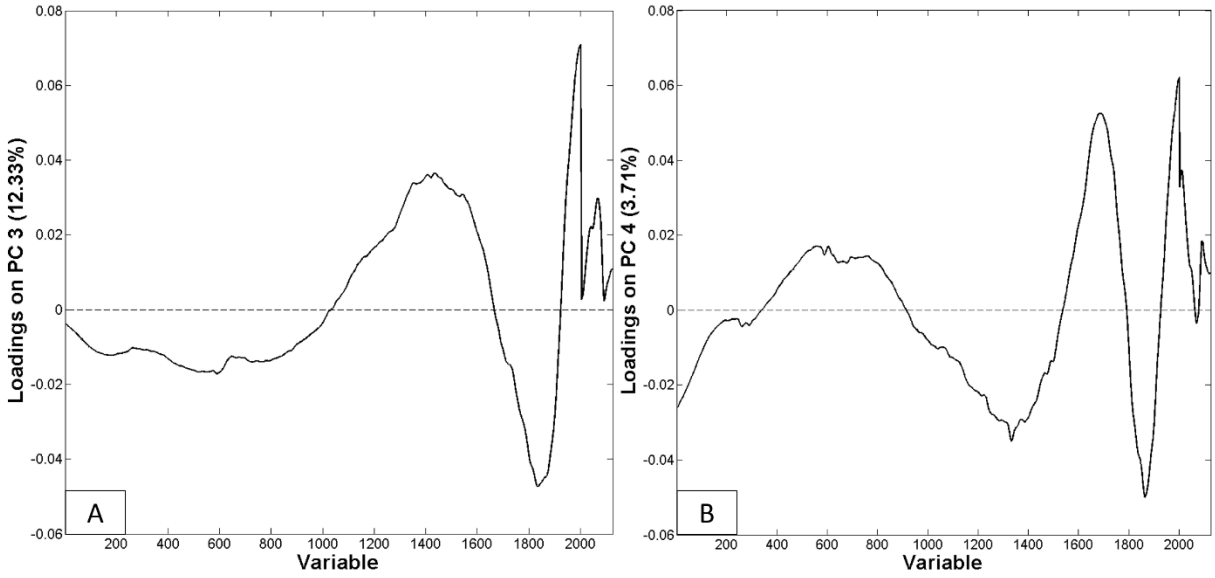

Figure 25. Loadings of PC3 (A) and PC4 (B) for the PCA model built with fused DSC and NIR data.

The loadings of PC3, seen in Figure 25A, show that the regions of DSC events between 250 °C and 310 °C and 310 °C and 350 °C, and the NIR bands of the third overtone of H_2O and second overtones of CH, ROH and H_2O contribute positively for this PC, while the region of DSC events between 350 °C and 400 °C contributes negatively for the PC. The loadings of PC4, seen in Figure 25B show that the regions of DSC events between 250 °C and 310 °C and 310 °C and 350 °C contributes negatively for this PC, while the region of DSC events between 350 °C and 400 °C is divided between both regions of this PC. The NIR bands assigned to third overtone of H_2O and second overtones of H_2O and ROH contribute positively for the PC, while the band of second overtone of CH contributes negatively. These contributions indicate that the region of DSC events between 350 °C and 400 °C and the bands associated to H_2O and ROH are more correlated with the class of coffees from Minas Gerais.

5.3.2 Partial Least Squares Discriminant Analysis (PLS-DA)

The PLS-DA model built using the fused DSC and NIR data can be seen in Figure 26. The model was built with 8 LVs, capturing 97.11% accumulated variance in the X block and 66.09% accumulated variance in the Y block.

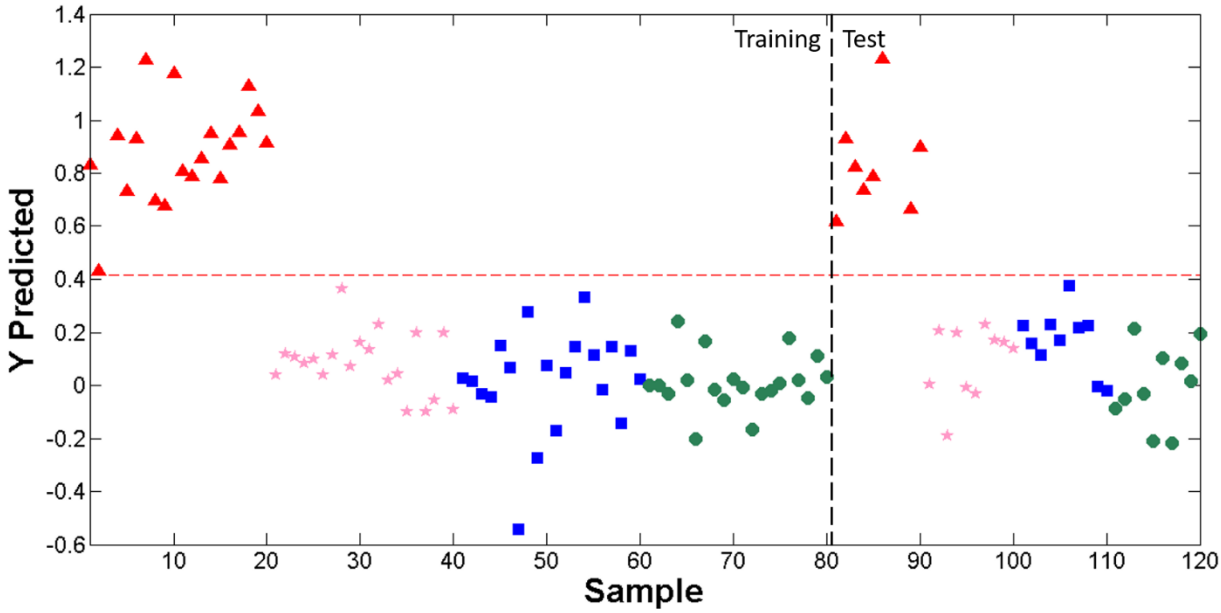

Figure 26. PLS2-DA predicted classes for samples of different Brazilian States analyzed by the DSC and NIR data fusion model. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA. The horizontal dashed line indicates the estimated threshold for predictions. The vertical dashed line indicates the separation between training and test samples.

The model presented perfect discrimination, with no false positive or negative errors on the training or the test set. The VIP scores and the regression vectors can be seen in Figure 27.

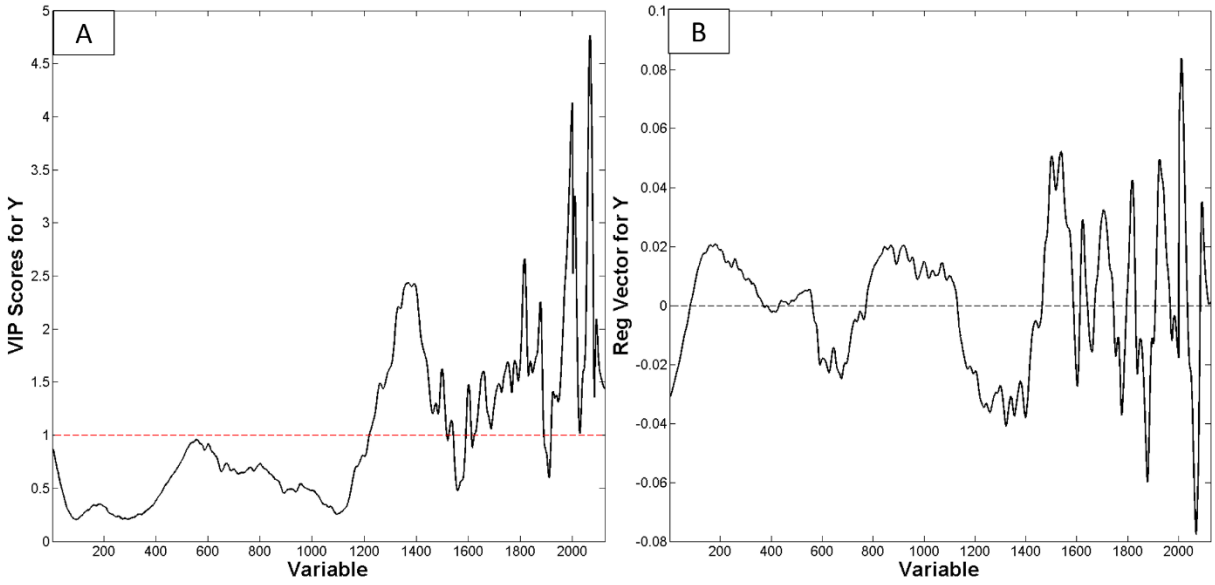

Figure 27. Informative vectors VIP Scores (A) and Regression coefficients (B) for the PLS2-DA model built with fused DSC and NIR data.

The VIP Scores plot, seen in Figure 27A, shows that the regions of DSC events between 310 °C and 350 °C and between 350 °C and 400 °C, as well as the NIR spectral bands assigned to the third overtone of H_2O and to second overtones of CH, ROH and H_2O were considered relevant for the model. The regression vector, seen in Figure 27B, shows that the region of DSC events between 310 °C and 350 °C contributes for the samples below the threshold, while the region of DSC events between 350 °C and 400 °C contributes to the samples above the threshold. The NIR spectral bands are split between the two regions, with the bands assigned to the third overtone of H_2O and second overtones of H_2O and ROH contributing to authentic samples (above the threshold), while the band assigned to the second overtone of CH contributes to the other samples, below the threshold.

5.3.3 Soft Independent Modelling of Class Analogies (SIMCA)

The SIMCA model built using fused DSC and NIR data can be seen in Figure 28. The model was built with 7 PCs and accounted for 98.90% of the total variance.

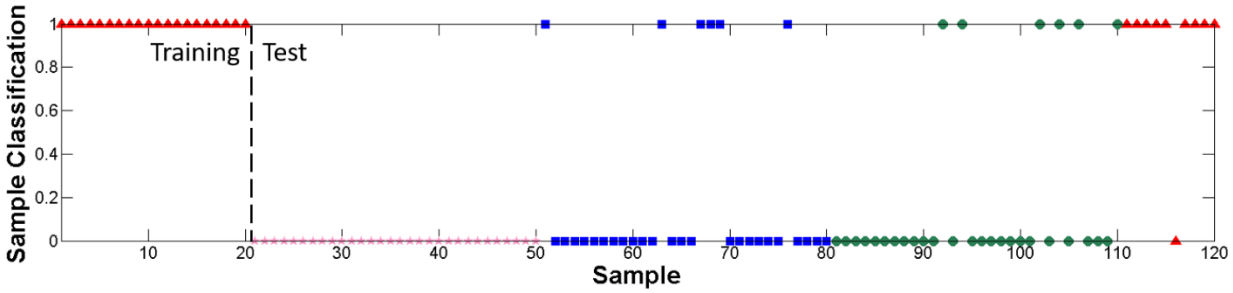

Figure 28. Class predictions for SIMCA model built with samples of different Brazilian States analyzed by DSC and NIR after data fusion. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA. The vertical dashed line indicates the separation between training and test samples.

The training set presented no errors, while the test set presented 11 false positives and 1 false negative. The modelling power of the variables is presented in Figure 29.

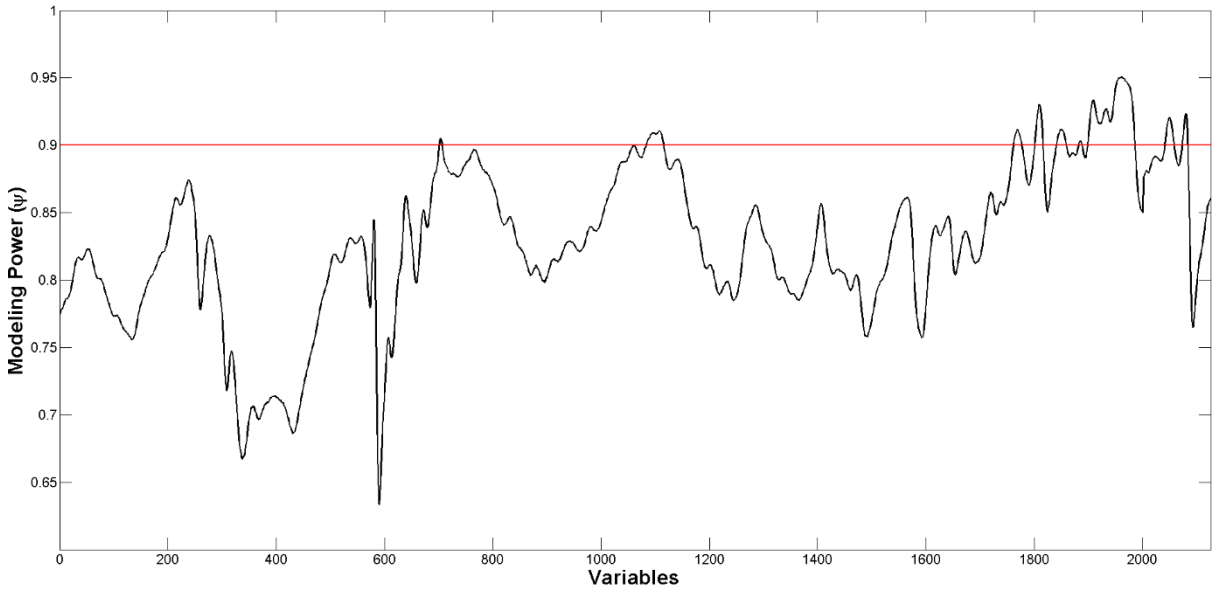

Figure 29. Modeling power vector for the SIMCA model built with fused DSC and NIR data.

The modelling power vector shows that the region of DSC events between 350 °C and 400 °C and the spectral bands assigned to the third overtone of H_2O and to the second overtone of CH were considered by the SIMCA model as the most important variables for the authentication of the target class.

5.3.4 Data Driven Soft Independent Modelling of Class Analogies (DD-SIMCA)

The acceptance plots for the DD-SIMCA model built using fused DSC and NIR data can be seen in Figure 30. The model was built with 7 PCs accounted for 98.90% of the total variance.

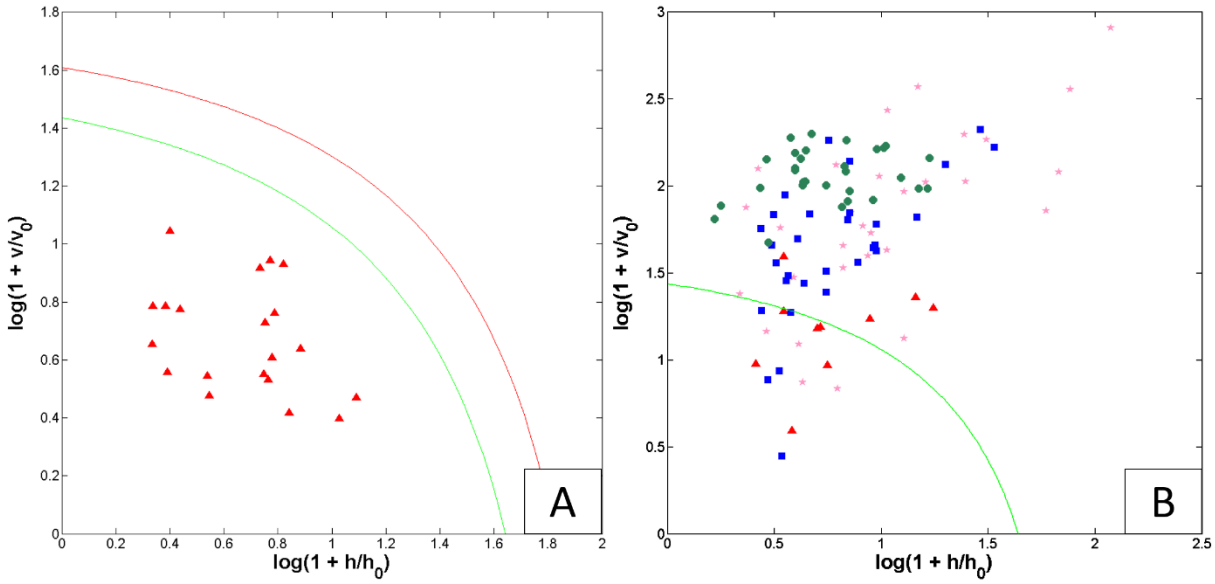

Figure 30. Acceptance plot for training (A) and test sets (B) of DD-SIMCA model built with samples of different Brazilian States analyzed by DSC-NIR data fusion model. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.

The training set presented no errors, while the test set presented 16 false positives and 6 false negatives, a result similar to that obtained with the SIMCA model. The modelling power vector is presented in Figure 31.

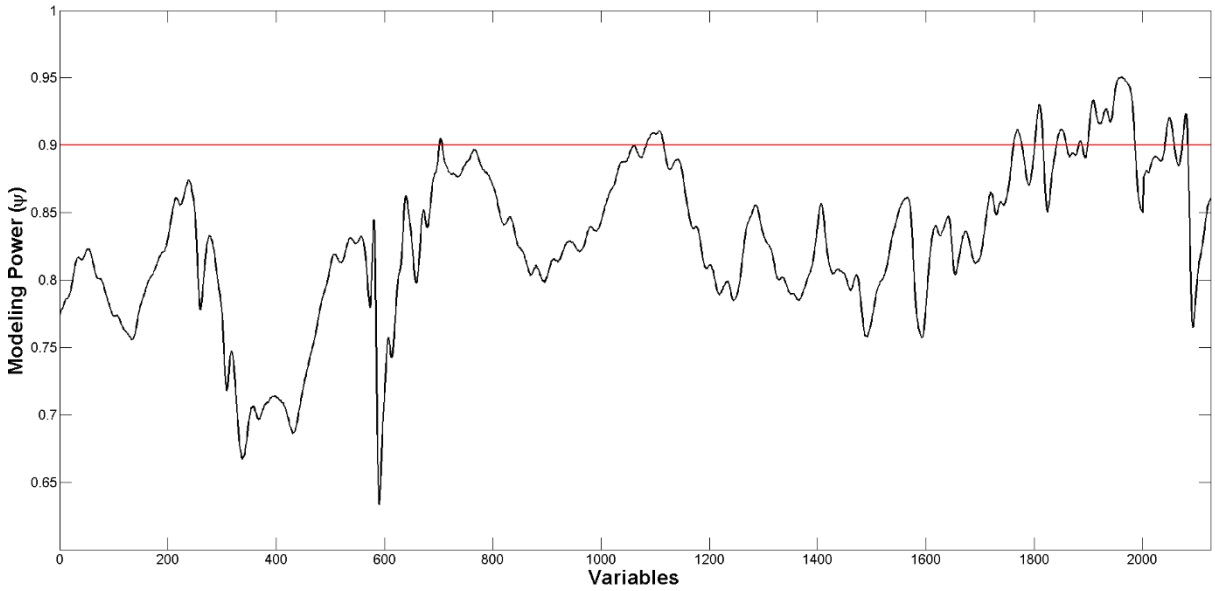

Figure 31. Modeling power vector for the SIMCA model built with fused DSC and NIR data.

The modelling power vector shows that the same variables previously highlighted by SIMCA were considered by the model as the most important ones for authentication.

5.4 Qualitative Validation

For validation and comparison between the built models, figures of merit (FOM) associated with accuracy rates were estimated. Those FOMs are presented in Table 2 presented below.

Table 2. Figures of merit for supervised classification models.

Analysis	Model	LVs/PCs	Training Set			Test Set		
			Sensitivity	Specificity	Efficiency	Sensitivity	Specificity	Efficiency
DSC	PLS-DA	7	100.00%	98.31%	99.15%	100.00%	100.00%	100.00%
	SIMCA	7	100.00%	–	–	60.00%	80.90%	69.67%
	DD-SIMCA	7	100.00%	–	–	40.00%	82.22%	57.35%
NIR	PLS-DA	6	100.00%	98.33%	99.16%	100.00%	100.00%	100.00%
	SIMCA	4	100.00%	–	–	100.00%	100.00%	100.00%
	DD-SIMCA	4	100.00%	–	–	100.00%	100.00%	100.00%
Data Fusion	PLS-DA	8	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	SIMCA	7	100.00%	–	–	90.00%	87.64%	88.81%
	DD-SIMCA	7	100.00%	–	–	60.00%	90.00%	73.48%

All the models built obtained global efficiency values above 50% for the authentication of discrimination between coffees from Minas Gerais and from other Brazilian States, especially PLS-DA and one-class models built with NIR data.

The models built using NIR data generally obtained better performance, but NIR spectra are sensitive to the roasting process of each sample, as the color and the volatile composition of the coffee can be varied during this process, what may affect the model if the roasting level was not controlled. Differential scanning calorimetry can be used as an alternative for this problem, as the analysis was performed above the commercial roasting temperature of coffee, mitigating the variance derived from the roasting process. Combining both techniques in a data fusion strategy is presented as an option for providing more accurate and confident results than those obtained with individual techniques, while increasing the robustness of the analysis regarding the roasting level of the samples.

Although PLS-DA models provided the best results among all the supervised classification methods used, it may not be able to predict samples from regions outside those modeled in the training set, requiring representative samples from all modeled classes. The possibility of generating a model with low robustness regarding other producer regions should be avoided. One class methods, such as SIMCA and DD-SIMCA, on the other hand, require only samples from the target class and will classify the samples as belonging or not to that class, generating a more robust model more suitable coffee authentication tool.

6 CONCLUSION

The instrumental analysis conducted in this study proved to be fast, with simple sample preparation methods and in accordance with the green chemistry principles for the analysis of coffee samples, but it was not possible to define specifical chemical compounds to be characterized as markers for differentiating between the coffee samples from the city of Mutum and the other producer regions. NIR spectroscopy has been a technique vastly used in association with chemometric methods, while the use of thermal analysis has still been rarely used in this context. NIR bands associated with H_2O and ROH compounds and decomposition events detected by DSC at temperatures from 250 °C to 310 °C and from 350 °C to 400 °C were evaluated as the most most important for characterizing coffee samples from Mutum.

The exploratory, discriminant analysis and one-class modeling models built using individually DSC and NIR spectroscopy and data fusion strategy provided satisfactory performance results for the classification of the conilon coffees produced in Mutum, Minas Gerais. Models based on NIR spectroscopy provided better performance, but they may not be robust to samples obtained at different roasting. Models based on DSC are more robust regarding this variance of the manufacturing process, but they provided lower efficiency rates. The data fusion strategy proved to be a good alternative, incrementing the performance results of the DSC models and increasing the robustness of the NIR models when dealing with samples of unknown manufacturing process.

The developed discriminant models presented better classification than those from one-class modelling, but may generate local models that need information about representative samples of all the classes studied. The one-class methods, however, presented more robust models that can be used with information solely about the target class, being more suitable for the intended authentication.

This study provided fast and simple strategies for the authentication of canephora coffee samples produced in the city of Mutum, in the State of Minas Gerais, using traditional and less common instrumental analysis allied with chemometric tools.

REFERENCES

1. BRONDI, A. M., TORRES, C., GARCIA, J. S., TREVISAN, M. G. Differential scanning calorimetry and infrared spectroscopy combined with chemometric analysis to the determination of coffee adulteration by corn. **Journal of the Brazilian Chemical Society,** 28, 1308–1314, 2017.

2. CRAWFORD, J. History of Coffee. **Journal of the Statistical Society of London**, 15, 1852.

3. CONCETTA, M., COUTO, C. **Sou Barista**. São Paulo: Editora Senac São Paulo, 2018.

4. MORRIS, J. **Coffee: A Global History**. London: Reaktion Books, 2019.

5. PENDERGRAST, M. **Uncommon Grounds: The History of Coffee and How It Transformed Our World**. New York: Hachette Book Group, 2019.

6. CONSELHO NACIONAL DO MINISTÉRIO PÚBLICO. **Ministério Público em Defesa do Estado Laico**. Brasilia: CNNP, 2014.

7. WILKSON, A. O dia em que CBF driblou a Fifa e mudou a camisa da seleção por US$ 3 mi. UOL. 2014. Available at: <https://copadomundo.uol.com.br/noticias/redacao/2014/04/01/o-dia-em-que-cbf-driblou-a-fifa-e-mudou-a-camisa-da-selecao-por-us-3-mi.htm?cmpid=copiaecola> Acessed in: 14 July 2025.

8. MCCOOK, S. G. **Coffee Is Not Forever: A Global History of the Coffee Leaf Rust**. Athens, Ohio University Press, 2019.

9. MAPA – Ministério da Agricultura e Pecuária. **Sumário Executivo Café**. 2025.

10. INPI - Instituto Nacional da Propriedade Industrial. **Certificado de Registro de Indicação Geográfica BR 412020000004-0**. June, 2021.

11. INPI - Instituto Nacional da Propriedade Industrial. **Certificado de Registro de Indicação Geográfica BR 402020000002-7**. May, 2022.

12. DOS SANTOS, L. B., TARABAL, J., SENA, M. M., ALMEIDA, M. R. UV-Vis spectroscopy and one-class modeling for the authentication of the geographical origin of green coffee beans from Cerrado Mineiro, Brazil. **Journal of Food Composition and Analysis,** 123, 105555, 2023.

13. BAQUETA, M. R., VALDERRAMA, P., ALVES, E. A., PALLONE, J. A. L., MARINI, F. Discrimination of Robusta Amazônico coffee farmed by indigenous and non-indigenous people in Amazon: comparing benchtop and portable NIR using ComDim and duplex. **Analyst,** 148, 1524–1533, 2023.

14. ZHU, C., FU, X., ZHANG, J., QIN, K., WU, C. Review of portable near infrared spectrometers: Current status and new techniques. **Journal of Near Infrared Spectroscopy,** 30, 2, 51–66, 2022.

15. DANIEL, J. S. P., CRUZ, J. C., CATELANI, T. A., GARCIA, J. S., TREVISAN, M. G. Erythromycin-excipients compatibility studies using the thermal analysis and dynamic thermal infrared spectroscopy coupled with chemometrics. **Journal of Thermal Analysis and Calorimetry,** 143, 3127–3135, 2021.

16. FURTADO, W. L., CORGOZINHO, C. N. C., TAULER, R., SENA, M. M. Monitoring biodiesel and its intermediates in transesterification reactions with multivariate curve resolution alternating least squares calibration models. **Fuel**, 283, 119275, 2021.

17. FERNANDES, D. D. S., SANTANA, C., P., FERNANDES, F. H. A., RAMOS, H. A., MEDEIROS, A. C. D., VERAS, G. One-Class Classification Models for the Authentication of Analgesic Tablet Reference Medicine Using Differential Scanning

Calorimetry and Visible-Near Infrared Spectroscopy. **Journal of the Brazilian Chemical Society,** 34, 213–219, 2023.

18. PEREIRA, L. H., CATELANI, T. A., COSTA, E. D. M., GARCIA, J. S., TREVISAN, M. G. Coffee adulterant quantification by derivative thermogravimetry and chemometrics analysis. **Journal of Thermal Analysis and Calorimetry,** 147, 7353–7362, 2022.

19. NUNES, K. M., ANDRADE, M. V. O., SANTOS FILHO, A. M. P., LASMAR, M. C., SENA, M. M. Detection and characterisation of frauds in bovine meat in natura by non-meat ingredient additions using data fusion of chemical parameters and ATR-FTIR spectroscopy. **Food Chemistry** 205, 14–22, 2016.

20. JESZKA-SKOWRON, M., ZGOŁA-GRZEŚKOWIAK, A., GRZEŚKOWIAK, T. Analytical methods applied for the characterization and the determination of bioactive compounds in coffee. **European Food Research and Technology**, 240, 19–31, 2014.

21. BAQUETA, M. R., ALVES, E. A., VALDERRAMA, P., PALLONE, J. A. L. Brazilian Canephora coffee evaluation using NIR spectroscopy and discriminant chemometric techniques. **Journal of Food Composition and Analysis,** 116, 105065, 2023.

22. YEAGER, S. E., BATALI, M. E., GUINARD, J. X., RISTENPART, W. D. Acids in coffee: A review of sensory measurements and meta-analysis of chemical composition. **Critical Reviews in Food Science and Nutrition**. 63, 1010–1036, 2021.

23. AURUM, F. S., IMAIZUMI, T., MANASIKAN, T., PRASEPTIANGGA, D., NAKANO, K. Coffee Origin Determination Based on Analytical and Nondestructive Approaches –A Systematic Literature Review. **Reviews in Agricultural Science,** 10, 257–287, 2022.

24. GARCÍA-PÉREZ, P., BECCHI, P. P., ZHANG, L., ROCCHETTI, G., LUCINI, L. Metabolomics and chemometrics: The next-generation analytical toolkit for the evaluation of food quality and authenticity. **Trends in Food Science and Technology**, 147, 104481, 2024.

25. WATSON, E. S., O'NEILL, M. J., JUSTIN, J., BRENNER, N. A Differential Scanning Calorimeter for Quantitative Differential Thermal Analysis. **Analytical Chemistry,** 36, 1233–1238, 1964.

26. WATSON, E. S., O'NEILL, M. J. **Differential Microcalorimeter**. The Perkin-Elmer Corporation. 3,263,484. Deposit: 4 Apr. 1962. Patented: 2 Aug. 1966.

27. IONASHIRO, M. **Giolito. Fundamentos Da Termogravimetria, Análise Térmica Diferencial e Calorimetria Exploratória Diferencial.** São Paulo: Giz Editorial, 2004.

28. IACCHERI, E., RAGNI, L., CEVOLI, C., ROMANI, S., ROSA, M. D., ROCCULI, P. Glass transition of green and roasted coffee investigated by calorimetric and dielectric techniques. **Food Chemistry,** 301, 125187, 2019.

29. MUTOVKINA, E. A., BREDIKHIN, S. A. Analysis of coffee thermophysical changes during roasting using differential scanning colorimetry. **Food Science and Technology**, 43, 119722, 2023.

30. PARNIAKOV, O., BALS, O., BARBA, F. J., MYKHAILYK, V., LEBOVKA, N., VOROBIEV, E. Application of differential scanning calorimetry to estimate quality and nutritional properties of food products. **Critical Reviews in Food Science and Nutrition**, 58, 362–385, 2018.

31. HERSCHEL, W. XIV. Experiments on the refrangibility of the invisible rays of the sun. **Philosophical Transactions of the Royal Society of London,** 90, 284–292, 1800.

32. MCCLURE, W. F. 204 years of near infrared technology: 1800-2003. **Journl of Near Infrared Spectroscopy,** 11, 487–518, 2003.

33. DAVIES, T. The history of near infrared spectroscopic analysis: Past, present and future - From sleeping technique to the morning star of spectroscopy. **Analusis,** 26, 17-19, 1998.

34. INFANTE, H. G., WARREN, J., CHALMERS, J., DENT, G., TODOLI, J. L. COLLINGWOOD, J., TELLING, N., RESANO, M., LIMBECK, A., SCHOENBERGER, T., HIBBERT, D. B. LEGRESLEY, A., ADAMS, K., CRASTON, D. Glossary of methods and terms used in analytical spectroscopy (IUPAC Recommendations 2019). **Pure and Applied Chemistry,** 93, 647–776, 2021.

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

36. QU, J. H., LIU, D., CHENG, J., SUN, D., MA, J., PU, H., ZENG, X. Applications of Near-infrared Spectroscopy in Food Safety Evaluation and Control: A Review of Recent Research Advances. **Critical Reviews in Food Science and Nutrition**, 55, 1939–1954, 2015.

37. BRERETON, R. G., JANSEN, J., LOPES, J., MARINI, F., POMERANTSEV, A., RODIONOVA, O., ROGER, J. M., WALCZAK, B., TAULER, R. Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. **Analytical and Bioanalytical Chemistry**. 409. 25. 5891–5899. 2017.

38. HIBBERT, D. B. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). **Pure and Applied Chemistry,** 88, 407–443, 2016.

39. PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, 2, 559–572, 1901.

40. JOLLIFE, I. T., CADIMA, J. Principal component analysis: A review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, 374, 2065, 2016.

41. BRO, R., SMILDE, A. K. Principal component analysis. **Analytical Methods**. 6, 2812–2831, 2014.

42. STAHLE, L., WOLD, S. PARTIAL LEAST SQUARES ANALYSIS WITH A MONTE CARL0 STUDY CROSS-VALIDATION FOR THE TWO-CLASS PROBLEM. **Journal of Chemometrics**, 1, 185-196, 1987.

43. BARKER, M., RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics,** 17, 166–173, 2003.

44. FERREIRA, M. M. C. **QUIMIOMETRIA: Conceitos, Métodos e Aplicações**. Campinas: Editora da Unicamp, 2015.

45. MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada**. Belo Horizonte: Editora UFMG, 2005.

46. KUBINYI, H., FOLKERS, G., MARTIN, Y. C. **3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity**. London: Kluwer Academic Publishers, 1998.

47. BRERETON, R. G., LLOYD, G. R. Partial least squares discriminant analysis: Taking the magic away. **Journal of Chemometrics,** 28, 213–225, 2014.

48. WOLD, S., SJOSTROM, M., ERIKSSON, L. PLS-Regression: A Basic Tool of Chemometrics. **Chemometrics and Intelligent Laboratory Systems**, 58, 109-130, 2001.

49. WOLD, S. PATTERN RECOGNITION BY MEANS OF DISJOINT PRINCIPAL COMPONENTS MODELS. **Pattern Recognition**, 8, 127-139, 1976.

50. OLIVERI, P., DOWNEY, G. Multivariate class modeling for the verification of food-authenticity claims. **TrAC - Trends in Analytical Chemistry**, 35, 74–86, 2012.

51. ZONTOV, Y. V., RODIONOVA, O. Y., KUCHERYAVSKIY, S. V., POMERANTSEV, A. L. DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach. **Chemometrics and Intelligent Laboratory Systems,** 167, 23–28, 2017.

52. POMERANTSEV, A. L. Acceptance areas for multivariate classification derived by projection methods. **Journal of Chemometrics,** 22, 601-609, 2008.

53. WOLD, S., SJÖSTRÖM, M. **SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy**. In: KOWALSKI, B. R. Chemometrics: Theory and Application. Washington: American Chemical Society, 1977.

54. BORRÀS, E., FERRÉ, J., BOQUÉ, R., MESTRES, M., ACEÑA, L., BUSTO, O. Data fusion methodologies for food and beverage authentication and quality assessment - A review. **Analytica Chimica Acta**, 891, 1–14, 2015.

55. FORSHED, J., STOLT, R., IDBORG, H., JACOBSSON, S. P. Enhanced multivariate analysis by correlation scaling and fusion of LC/MS and 1H NMR data. **Chemometrics and Intelligent Laboratory Systems,** 85, 179–185, 2007.

56. POMERANTSEV, A. L., RODIONOVA, O. Y. New trends in qualitative analysis: Performance, optimization, and validation of multi-class and soft models. **TrAC - Trends in Analytical Chemistry,** 143, 116372, 2021.

57. EUROPEAN UNION. Commission Decision of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results. Available at: < http://data.europa.eu/eli/dec/2002/657/oj>. Access in 14 July 2025.

58. KENNARD, R. W., STONE, L. A. Computer Aided Design of Experiments. **Technometrics,** 11, 137-148, 1969.

59. HORWITZ, W. Protocol for the design, conduct and interpretation of method-performance studies: Revised 1994 (Technical Report). **Pure and Applied Chemistry,** 67, 331–343, 1995.

60. METROHM NIRSYSTEMS. **A Guide to Near-Infrared Spectroscopic Analysis of Industrial Manufacturing Processes**. Herisau: Metrohm, 2014.

APPENDICES

APPENDIX A – MATLAB 2010B ROUTINES

A1. DSC Curves Import to Matlab

```
function importdsc(X, Y)
% ------------------------------------------------------------------------
% Function:
% importdsc(X, Y)
% ------------------------------------------------------------------------
% Usage:
% This function imports data from .txt files of DSC analyses in the current folder,
% processes them, and adds them to the MATLAB workspace.
% ------------------------------------------------------------------------
% Parameters:
%   X: initial value of the scale range
%   Y: final value of the scale range
% ------------------------------------------------------------------------
% The function performs the following steps:
% 1. Identifies all .txt files in the current folder.
% 2. Initializes the output matrices Scale and Data.
% 3. For each .txt file:
```

```
%    a. Opens the file and reads its content.
%    b. Identifies the line containing the sample weight.
%    c. Locates the [Data] section and extracts the data columns.
%    d. Processes the data line by line and interpolates the DSC mW values.
%    e. Adds the processed data to the Data matrix.
% 4. Imports a vector with the sample labels.
% 5. Adds the variables Scale, Data, and Label to the base MATLAB workspace.
% ------------------------------------------------------------------------
% Written by:
% Hélio Milito Martins de Amorim Neto
% Department of Chemistry, Institute of Exact Sciences
% Federal University of Minas Gerais
% January 2025
% ------------------------------------------------------------------------
% Identify .txt files in the current folder
asc = dir('*.txt');
% Initialize output matrices
Scale = X:0.1:Y; % Vector for the range specified by the user
Data = []; % Imported data for all samples
for i = 1:length(asc)
% Open the current file
```

```
fileID = fopen(asc(i).name, 'r');
content = textscan(fileID, '%s', 'Delimiter', '\n');
content = content{1};
fclose(fileID);
% Find sample weight
weightLine = false(size(content));
for j = 1:length(content)
if ~isempty(strfind(content{j}, 'Sample Weight:'))
weightLine(j) = true;
end
end
weightStr = regexp(content{weightLine}, '\d+\.\d+', 'match');
if isempty(weightStr)
error('Could not find sample weight in file %s.', asc(i).name);
end
sampleWeight = str2double(weightStr{1});
% Locate the [Data] section and extract columns
dataStart = find(~cellfun('isempty', strfind(content, '[Data]'))) + 2;
if isempty(dataStart)
error('[Data] section not found in file %s.', asc(i).name);
end
```

```
rawData = content(dataStart:end);

% Process data line by line

dataArray = [];

for j = 1:length(rawData)

lineData = sscanf(rawData{j}, '%f');

if length(lineData) == 3 % Ensure there are 3 columns

dataArray = [dataArray; lineData'];

end

end

% Check if expected columns are present

if size(dataArray, 2) < 3

error('Insufficient data in file %s. Check if columns are complete.', asc(i).name);

end

% Extract and process columns

tempC = round(dataArray(:, 2) * 10) / 10; % Round Temp C to 1 decimal place

dscMw = dataArray(:, 3) / sampleWeight; % Divide DSC mW by sample weight

% Create temporary matrix for interpolation

tempMatrix = [tempC, dscMw];

% Remove duplicates and interpolate missing values

[uniqueTemp, ia, ~] = unique(tempMatrix(:, 1));

uniqueDSC = tempMatrix(ia, 2);
```

```
interpolatedDSC = interp1(uniqueTemp, uniqueDSC, Scale, 'linear', 'extrap');
% Fill in missing values as specified
for j = 1:length(interpolatedDSC)
if isnan(interpolatedDSC(j))
prev = find(~isnan(interpolatedDSC(1:j-1)), 1, 'last');
next = find(~isnan(interpolatedDSC(j+1:end)), 1, 'first') + j;
if ~isempty(prev) && ~isempty(next)
interpolatedDSC(j) = mean([interpolatedDSC(prev), interpolatedDSC(next)]);
elseif ~isempty(prev)
interpolatedDSC(j) = interpolatedDSC(prev);
elseif ~isempty(next)
interpolatedDSC(j) = interpolatedDSC(next);
else
interpolatedDSC(j) = 0; % Default value if none available
end
end
end
% Add processed data column
Data = [Data, interpolatedDSC'];
end
% Import vector with sample labels
```

```
asc = dir('*.txt');

file_names = {asc.name}';

Label = cellfun(@(x) strtok(x, '.'), file_names, 'UniformOutput', false);

assignin('base', 'Label', Label);

clear file_names asc

% Add variables Scale, Data, and Label to the base workspace

assignin('base', 'Scale', Scale'); % Transposed Scale

assignin('base', 'Data', Data');   % Transposed Data

assignin('base', 'Label', Label);  % Label vector

end
```

A2. NIR Spectra Import to Matlab

```
% ------------------------------------------------------------------

% Function:

% importnir

% ------------------------------------------------------------------

% Usage:

% This function imports data from .csv files of NIR analyses in the current folder,

% processes them, and adds them to the MATLAB workspace.

% ------------------------------------------------------------------

% The function performs the following steps:
```

```
% 1. Identifies all .csv files in the current folder.

% 2. Initializes the output matrices Scale and Data.

% 3. For each .csv file:

%    a. Opens the file and reads its content.

%    b. Checks if the decimals are separated by a dot.

%    c. If the decimals are separated by a comma, replaces it with a dot.

%    d. Performs the logarithmic conversion of the spectra.

%    e. Adds the processed data to the X matrix.

% 4. Imports a vector with the sample labels.

% 5. Adds the variables Scale, X, and Label to the base MATLAB workspace.

% ------------------------------------------------------------------------

% Written by:

% Hélio Milito Martins de Amorim Neto

% Department of Chemistry, Institute of Exact Sciences

% Federal University of Minas Gerais

% January 2025

% ------------------------------------------------------------------------

% 1st: List all CSV files in the current folder

asc = dir('*.csv');

numfile = length(asc);

% Initialize processed data matrix
```

```
dadosX = cell(numfile, 1);

% Loop to process each CSV file

for k = 1:numfile

arquivo = asc(k).name;

% Open the file for reading

fid = fopen(arquivo, 'r');

if fid == -1

error('Could not open the file: %s', arquivo);

end

% Read all lines from the file

linhas = textscan(fid, '%s', 'Delimiter', '\n');

linhas = linhas{1};

fclose(fid);

% Check the number of commas in the second line (ignoring the header)

num_virgulas = length(strfind(linhas{2}, ','));

% If there are 41 commas, replace commas in odd positions with dots

if num_virgulas == 41

for i = 2:length(linhas) % Start from the second line (ignore header)

% Find positions of all commas in the line

virgula_posicoes = strfind(linhas{i}, ',');

% Replace commas in odd positions with dots
```

```
for j = 1:2:length(virgula_posicoes)

posicao = virgula_posicoes(j); % Position of the comma to be replaced

linhas{i}(posicao) = '.'; % Replace comma with dot

end

end

% Save the modified file

novo_arquivo = strrep(arquivo, '.csv', '_modificado.csv');

fid = fopen(novo_arquivo, 'w');

if fid == -1

error('Could not create the modified file: %s', novo_arquivo);

end

for i = 1:length(linhas)

fprintf(fid, '%s\n', linhas{i});

end

fclose(fid);

% Update the file name to the modified file

arquivo = novo_arquivo;

elseif num_virgulas ~= 20

% If the number of commas is neither 20 nor 41, display a warning

warning('The file %s has %d commas. No action was taken.', arquivo, num_virgulas);

end
```

```
% 4th: Apply the data processing routine
dadosX{k} = importdata(arquivo);
% Check if the data was imported correctly and has enough columns
if isfield(dadosX{k}, 'data') && size(dadosX{k}.data, 2) >= 21
M = mean(dadosX{k}.data(:, 2:21)'); % Mean of columns 2 to 21
X(k, :) = M;
else
error('The file %s does not contain enough data (at least 21 columns).', arquivo);
end
end
% Apply the transformation 2 - log(X)
X = 2 - log(X);
% Extract labels from file names
nomes_arquivos = {asc.name}';
Label = cellfun(@(x) strtok(x, '.'), nomes_arquivos, 'UniformOutput', false);
% Extract the scale (first column of the data)
Scale = dadosX{1}.data(:, 1);
% Delete "_modificado.csv" files
modificados = dir('*_modificado.csv');
for k = 1:length(modificados)
delete(modificados(k).name);
```

```
end

% Clear unnecessary variables

clear nomes_arquivos asc M dadosX k numfile fid linhas i j novo_arquivo num_virgulas
virgula_posicoes posicao modificados ans arquivo;
```

A3. PLS-DA Dataset construction

```
function [ModelPLSDA, TestPLSDA] = DS_PLSDA(varargin)

% DS_PLSDA splits multiple datasets using Kennard-Stone and merges them.

% ------------------------------------------------------------------------

% Usage:

%   [ModelPLSDA, TestPLSDA] = DS_PLSDA(S1, LabelS1, S2, LabelS2, ...)

% ------------------------------------------------------------------------

% Inputs:

%   - S1, S2, ...     : Numeric datasets (matrices)

%   - LabelS1, LabelS2, ... : Corresponding labels (categorical, numeric, or cell array)

% ------------------------------------------------------------------------

% Outputs:

%   - ModelPLSDA : Dataset containing all merged model sets

%   - TestPLSDA  : Dataset containing all merged test sets

% ------------------------------------------------------------------------

% Example:

%   [ModelPLSDA, TestPLSDA] = DS_PLSDA(data1, labels1, data2, labels2);
```

```
% ------------------------------------------------------------------------

% Written by:

% Hélio Milito Martins de Amorim Neto

% Department of Chemistry, Institute of Exact Sciences

% Federal University of Minas Gerais

% January 2025

% ------------------------------------------------------------------------

numInputs = length(varargin);

numDatasets = numInputs / 2; % Number of datasets (each dataset has a corresponding label)

if mod(numInputs, 2) ~= 0

error('Each dataset must have a corresponding label vector.');

end

modelSets = cell(1, numDatasets);

testSets = cell(1, numDatasets);

modelLabels = cell(1, numDatasets);

testLabels = cell(1, numDatasets);

for i = 1:numDatasets

X = varargin{2*i-1}; % Dataset

labels = varargin{2*i}; % Corresponding labels

% Validate dataset and labels
```

```
if ~isnumeric(X)
error('Dataset %d must be a numeric matrix.', i);
end
if length(labels) ~= size(X, 1)
error('Label vector %d must have the same number of rows as its dataset.', i);
end
numRows = size(X, 1);
k = round(2/3 * numRows); % Select 2/3 of the rows for the model set
% Apply the Kennard-Stone function
[model, test] = kenstone(X, k);
% Store model and test sets
modelSets{i} = X(model, :);
testSets{i} = X(test, :);
% Store corresponding labels
modelLabels{i} = labels(model, :);
testLabels{i} = labels(test, :);
end
% Merge all model sets, test sets, and their labels
ModelPLSDA = dataset(cat(1, modelSets{:}));
TestPLSDA = dataset(cat(1, testSets{:}));
LabelModel = cat(1, modelLabels{:});
```

```
LabelTest = cat(1, testLabels{:});

ModelPLSDA.label{1}={LabelModel};

TestPLSDA.label{1}={LabelTest};

end
```

A4. SIMCA Dataset Construction

```
function [ModelSIMCA, TestSIMCA] = DS_SIMCA(varargin)

% DS_SIMCA applies Kennard-Stone only to the first dataset (S1), while all others are test data.

% ------------------------------------------------------------------

% Usage:

%   [ModelSIMCA, TestSIMCA] = DS_SIMCA(S1, LabelS1, S2, LabelS2, ...)

% ------------------------------------------------------------------

% Inputs:

%   - S1, LabelS1: First dataset and labels (used for training/testing split)

%   - S2, LabelS2, ...: Other datasets and labels (used entirely for testing)

% ------------------------------------------------------------------

% Outputs:

%   - ModelSIMCA : Dataset containing training data (from S1)

%   - TestSIMCA  : Dataset containing test data (from S1 and all other datasets)

% ------------------------------------------------------------------

% Example:
```

```
%    [ModelSIMCA, TestSIMCA] = DS_SIMCA(data1, labels1, data2, labels2, data3, labels3);
% ------------------------------------------------------------------------
% Written by:
% Hélio Milito Martins de Amorim Neto
% Department of Chemistry, Institute of Exact Sciences
% Federal University of Minas Gerais
% January 2025
% ------------------------------------------------------------------------
numInputs = length(varargin);
if mod(numInputs, 2) ~= 0
error('Each dataset must have a corresponding label vector.');
end
numDatasets = numInputs / 2; % Number of dataset-label pairs
% Extract the first dataset and its labels
S1 = varargin{1};
LabelS1 = varargin{2};
% Validate the first dataset
if ~isnumeric(S1)
error('The first dataset (S1) must be a numeric matrix.');
end
```

```
if length(LabelS1) ~= size(S1, 1)

error('LabelS1 must have the same number of rows as S1.');

end

numRows = size(S1, 1);

k = round(2/3 * numRows); % Select 2/3 of the rows for the training set

% Apply Kennard-Stone only to S1

[modelIdx, testIdx] = kenstone(S1, k);

% Training data (from S1)

ModelSIMCA = dataset(S1(modelIdx, :));

LabelModel = LabelS1(modelIdx, :);

% Test data (S1 test + all other datasets)

TestSIMCA = dataset(S1(testIdx, :));

LabelTest = LabelS1(testIdx, :);

% Process additional datasets (S2, S3, ...)

for i = 2:numDatasets

X = varargin{2*i-1}; % Dataset

labels = varargin{2*i}; % Corresponding labels

% Validate dataset

if ~isnumeric(X)

error('Dataset %d must be a numeric matrix.', i);

end
```

```
if length(labels) ~= size(X, 1)

error('Label vector %d must have the same number of rows as its dataset.', i);

end

% Add the entire dataset to the test set

TestSIMCA = dataset(cat(1, X, double(TestSIMCA))); % Convert to double to avoid errors

LabelTest = cat(1, labels, LabelTest);

ModelSIMCA.label{1}={LabelModel};

TestSIMCA.label{1}={LabelTest};

end

end
```

A5. DD-SIMCA Dataset Construction

```
function [ModelDDSIMCA, TestDDSIMCA, LabelModel, LabelTest] = DS_DDSIMCA(varargin)

% DS_DDSIMCA applies Kennard-Stone only to the first dataset (S1), while all others are test data.

% ------------------------------------------------------------------

% Usage:

%   [ModelDDSIMCA, TestDDSIMCA, LabelModel, LabelTest] = DS_DDSIMCA(S1, LabelS1, S2, LabelS2, ...)

% ------------------------------------------------------------------

% Inputs:
```

```
%   - S1, LabelS1: First dataset and labels (used for training/testing split)
%   - S2, LabelS2, ...: Other datasets and labels (used entirely for testing)
% ------------------------------------------------------------------------
% Outputs:
%   - ModelDDSIMCA : Dataset containing training data (from S1)
%   - TestDDSIMCA  : Dataset containing test data (from S1 and all other datasets)
%   - LabelModel : Labels for ModelDDSIMCA
%   - LabelTest  : Labels for TestDDSIMCA
% ------------------------------------------------------------------------
% Example:
% [ModelDDSIMCA, TestDDSIMCA, LabelModel, LabelTest] = DS_DDSIMCA(data1, labels1, data2, labels2, data3, labels3);
% ------------------------------------------------------------------------
% Written by:
% Hélio Milito Martins de Amorim Neto
% Department of Chemistry, Institute of Exact Sciences
% Federal University of Minas Gerais
% January 2025
% ------------------------------------------------------------------------
numInputs = length(varargin);
if mod(numInputs, 2) ~= 0
```

```
error('Each dataset must have a corresponding label vector.');
end
numDatasets = numInputs / 2; % Number of dataset-label pairs
% Extract the first dataset and its labels
S1 = varargin{1};
LabelS1 = varargin{2};
% Validate the first dataset
if ~isnumeric(S1)
error('The first dataset (S1) must be a numeric matrix.');
end
if length(LabelS1) ~= size(S1, 1)
error('LabelS1 must have the same number of rows as S1.');
end
numRows = size(S1, 1);
k = round(2/3 * numRows); % Select 2/3 of the rows for the training set
% Apply Kennard-Stone only to S1
[modelIdx, testIdx] = kenstone(S1, k);
% Training data (from S1)
ModelDDSIMCA = S1(modelIdx, :);
LabelModel = LabelS1(modelIdx, :);
% Test data (S1 test + all other datasets)
```

```
TestDDSIMCA = S1(testIdx, :);

LabelTest = LabelS1(testIdx, :);

% Process additional datasets (S2, S3, ...)

for i = 2:numDatasets

X = varargin{2*i-1}; % Dataset

labels = varargin{2*i}; % Corresponding labels

% Validate dataset

if ~isnumeric(X)

error('Dataset %d must be a numeric matrix.', i);

end

if length(labels) ~= size(X, 1)

error('The label vector %d must have the same number of rows as its dataset.', i);

end

% Add the entire dataset to the test set

TestDDSIMCA = cat(1, double(TestDDSIMCA), X); % Convert to double to avoid errors

LabelTest = cat(1, LabelTest, labels);

end

end
```

A6. SIMCA and DD-SIMCA Modeling Power

```
% ------------------------------------------------------------------
```

```
% Function:
% [Power] = ModelPower(X,T,P,E,L,C)
% ------------------------------------------------------------------------
% Objective:
% Calculate and plot the modeling power for SIMCA and DD-SIMCA models
% ------------------------------------------------------------------------
% Input:
% X, data matrix, samples in rows and variables in columns
% T, score matrix, samples in rows and PCs in columns
% P, loading matrix, PCs in rows and variables in columns
% E, variable vector, variables in row
% L, threshold (optional, if empty L = 0.9)
% C, number of classes (optional, if empty C = 1)
% ------------------------------------------------------------------------
% Output:
% Power, Modeling Power
% ------------------------------------------------------------------------
% Written by:
% Pedro Micael de Castro Caputo, Hélio Milito Martins de Amorim Neto
% Department of Chemistry, Institute of Exact Sciences
% Federal University of Minas Gerais
```

```
% May 2024
% ------------------------------------------------------------------------
% Reference:
% WOLD, Svante; SJÖSTRÖM, Michael. SIMCA: a method for analyzing chemical data in terms of similarity and analogy.
% DOI: 10.1021/bk-1977-0052.ch012
function [Power] = ModelPower(X,T,P,E,L,C)
Data_Matrix_X = X;
Score_Matrix_T = T;
Loading_Matrix_P = P;
if ~exist('E');
E = [1:size(Data_Matrix_X,2)]';
end
scale_vector = E;
PC_components = size(Loading_Matrix_P,1);
N_training_samples = size(Data_Matrix_X,1);
V_model_variables = size(Data_Matrix_X,2);
if ~exist('L');
L = 0.9;
end
Threshold = L;
```

```
if ~exist('C');

C = 1;

end

C_training_classes = C;

% Compute TxP, X_TxP, Eki2, Eik2, NA1

Matrix_TxP = Score_Matrix_T * Loading_Matrix_P;

Matrix_X_TxP = Data_Matrix_X - Matrix_TxP;

Matrix_Eki2 = Matrix_X_TxP .^ 2;

Matrix_Eik2 = Matrix_Eki2';

NA1_value = N_training_samples - PC_components - 1;

% Compute Div, Q, sumSi, multSi, Si

Matrix_Div = Matrix_Eik2 / NA1_value;

Q_value = (1 / C_training_classes) * (V_model_variables) / (V_model_variables - PC_components);

sum_Si = sum(Matrix_Div, 2);

mult_Si = Q_value * sum_Si;

Matrix_Si = sqrt(mult_Si);

% Compute Ym, Ym_rep, Xt_Ym_rep, quadSiy, N_1_Siy, Siy

Vector_Ym = (sum((Data_Matrix_X'), 2)) / N_training_samples;

Matrix_Ym_rep = repmat(Vector_Ym, 1, size(Data_Matrix_X', 2));

Matrix_Xt_Ym_rep = Data_Matrix_X' - Matrix_Ym_rep;
```

```
Matrix_quadSiy = Matrix_Xt_Ym_rep.^2;

Matrix_N_1_Siy = Matrix_quadSiy / (N_training_samples - 1);

Matrix_Siy = sqrt(sum(Matrix_N_1_Siy, 2));

% Compute Statistical Power

Power = 1 - (Matrix_Si ./ Matrix_Siy);

% Plot the modeling power graph

plot(scale_vector, Power, 'k-', 'LineWidth', 2);

hold on

plot(scale_vector, Threshold * ones(size(scale_vector)), 'r-', 'LineWidth', 2);

xlabel('Variables');

ylabel('Modeling Power (\psi)');
```

APPENDIX B – PCA SCORES PLOT

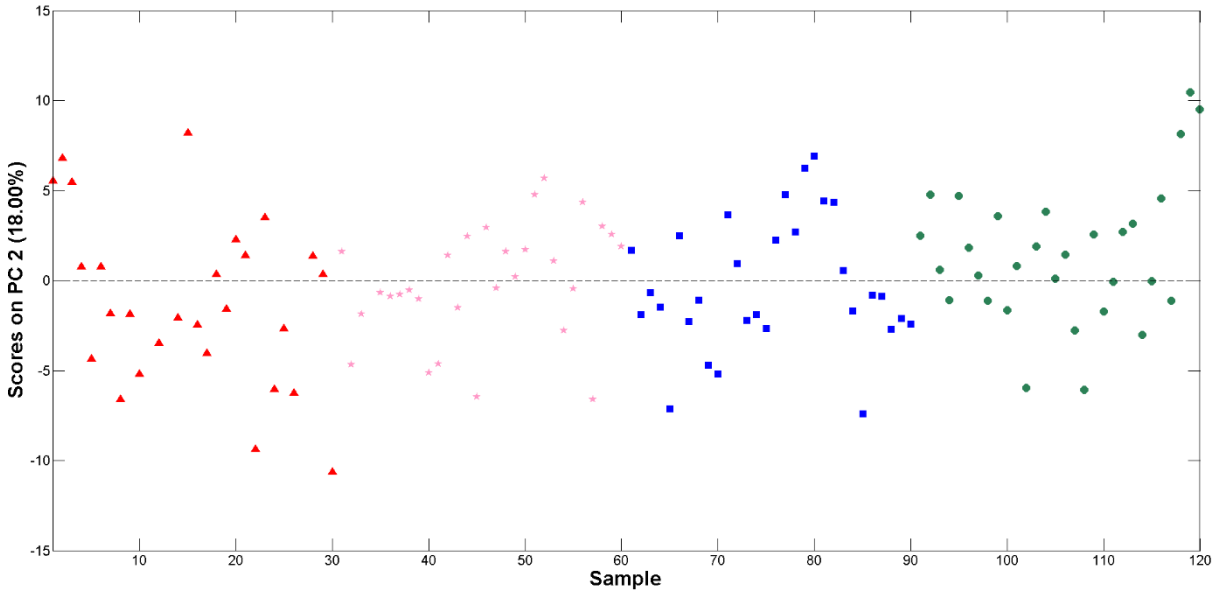

Figure B1. PCA scores of sample *versus* PC2 for the canephora coffee samples analyzed with DSC. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.

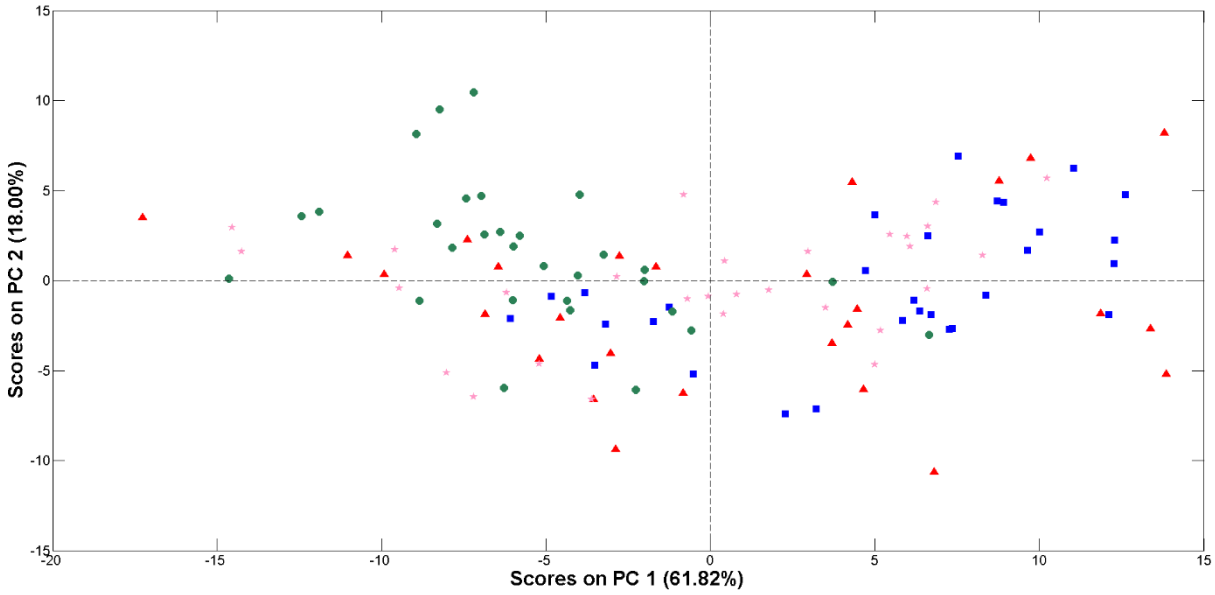

Figure B2. PCA scores plot of PC1 *versus* PC2 for the canephora coffee samples analyzed by the data fusion model DSC-NIR. Red triangles: CM; pink stars: CC; blue squares: RR; green circles: RA.