Doctoral Dissertation
Doctoral Program in Chemical Engineering (37th Cycle)

Chemometrics applied to food industry: improving quality, safety and sustainability using multivariate data analysis

Mattia Sozzi
* * * * * *

Supervisors
Prof. Francesco Savorani, Supervisor
Prof. Francesco Geobaldo, Co-Supervisor
Dr. Nicola Cavallini, Co-Supervisor

Politecnico di Torino
March 28th, 2025

Summary

This thesis explores the potential of chemometrics in the food industry and its broader applications in chemical engineering. By leveraging multivariate data analysis, chemometrics enables the extraction of valuable information from complex datasets, facilitating process optimization, quality control and product characterization. This interdisciplinary approach can integrate advanced statistical and mathematical methods with analytical techniques such as Nuclear Magnetic Resonance and similar spectroscopies, Liquid Chromatography and Mass Spectrometry, or even microscopy techniques like Field Emission Scanning Electron Microscopy, demonstrating its versatility in addressing challenges related to food quality, safety, and sustainability.

The research is organized around different case studies that progressively increase in complexity. In the first case study, Principal Component Analysis (PCA) is employed to explore and optimize industrial food treatments. By analyzing ^{1}H–NMR spectra, the study evaluates process parameters and experimental conditions that enhance resource efficiency. The second case study leverages a data fusion approach to improve food authenticity and traceability. Here, datasets obtained from NMR and LC–MS are combined and analysed with multivariate classification techniques such as Partial Least Squares Discriminant Analysis (PLS–DA), leading to robust models that discriminate hazelnut varieties based on geographical origin and cultivar. A further case study focuses on challenges in wine analysis using ^{1}H–NMR, like the ethanol quantification in alcoholic beverages. Through a systematic experimental design, an optimization of acquisition parameters was performed testing both direct and indirect quantification methods, advancing the accuracy and reliability of NMR ethanol measurements in wine. Lastly, the final case study introduces an innovative algorithm for image analysis, specifically targeting the characterization of rice kernels. By automatically extracting morphological features

from FESEM images and correlating these with the glycaemic index and other biochemical properties, the study offers a new tool for rapid and objective quality assessment in agrifood applications.

In addition to the case studies, the thesis provides a comprehensive review of chemometric methodologies, including both unsupervised and supervised approaches. It also addresses critical aspects of data preprocessing, model validation, and the application of Design of Experiments (DoE) for process optimization. These discussions underscore the importance of rigorous data treatment and methodological transparency in achieving reproducible and reliable outcomes.

Overall, the findings illustrate that the integration of chemometrics with advanced analytical techniques not only enhances the evaluation and optimization of food processes but also offers substantial improvements in product authentication and safety. By reducing experimental complexity and resource consumption, the approaches developed in this research contribute to the sustainable evolution of the food industry. Moreover, the versatility of chemometric tools employed demonstrates their broader applicability in various sectors of chemical engineering, suggesting a valuable direction for future research and industrial innovation.

1. Introduction 1

1.1 Thesis organization 4

2. Materials and methods 5

2.1 Metabolomics and Foodomics 5
2.2 Analytical techniques 6

2.2.1 Nuclear Magnetic Resonance (NMR) 6
2.2.2 Field Emission Scanning Electron Microscopy 8

2.3 Chemometric methods 8

2.3.1 Introduction to chemometric concepts 8
2.3.2 How to correctly treat the data: dataset organization and preprocessing methods 10
2.3.3 Unsupervised approaches: exploratory analysis and Principal Component Analysis 11
2.3.4 Supervised approaches: Regression and Classification methods 13
2.3.5 Data fusion approaches: fundamentals and advantages 17
2.3.6 Multivariate Curve Resolution (MCR) 18
2.3.7 Process optimization and Design of Experiment 19
2.3.8 Model validation 21

References | Chapter 2 25

3. Case Study #1: Optimization of industrial food treatments 33

3.1 Project overview 33
3.2 Experimental design and laboratory analysis 34
3.3 NMR data acquisition 35
3.4 Data processing 35
3.5 Data exploration and process optimization by PCA 35

3.5.1 PCA for exploring clusters, trends and outliers 35
3.5.2 PCA for time-trend evaluation and process optimization 41

3.6 Project conclusions and future perspectives 44

References | Chapter 3 45

4. Case study #2: Data fusion to improve food traceability e authenticity 46

4.1 Project overview 46

4.1.1 Employed techniques 47

4.2 Sample collection, data acquisition and processing 48
4.2.1 Samples collection and preparation.....48
4.2.2 ^{1}H-NMR analysis 50
4.2.3 Liquid Chromatography coupled with Mass Spectrometry.....53
4.3 PCA and PLS-DA on independent datasets.....53
4.3.1 Results for the NMR dataset.....54
4.3.2 Results for the LC-MS dataset.....57
4.4 Data fusion approaches and compatibility between techniques.....59
4.4.1 Data Fusion applied to Piedmont TGT hazelnuts case study 59
4.4.2 Fused dataset: exploratory analysis.....59
4.4.3 Fused dataset: classification analysis 61
4.5 Project conclusions and future perspectives 65
References | Chapter 4.....67
5. Case study #3: challenges in wine analysis using NMR 71
5.1 Project overview and aim of the work.....71
5.2 Sample collection and data acquisition: ^{1}H-NMR.....72
5.3 The problem of measuring ethanol content with ^{1}H-NMR.....77
5.3.1 Experimental design.....78
5.3.2 Sample preparation and data acquisition (^{1}H-NMR) 79
5.4 Ethanol spectra and conversion factor for satellite signals.....80
5.5 NMR signal area calculation and optimal conditions.....82
5.5.1 Combinations of different parameters.....83
5.6 Indirect ethanol quantification 84
5.6.1 An FT-IR based method: WineScan 84
5.6.2 PLS regression for NMR spectra 85
5.7 Case study evaluation: real wine samples.....87
5.7.1 Direct ethanol quantification using optimal conditions 87
5.7.2 Parameters evaluation using ASCA 88
5.7.3 Ethanol quantification using an external standard 90
5.8 Conclusions and future perspectives 92
References | Chapter 5.....93

6. Case study #4: image analysis in agrifood 95

6.1 Project overview.....95

6.1.1 Image analysis to automatically extract objects' features 95
6.1.2 A case study on rice kernels and glycaemic index.....96

6.2 Sample collection and data acquisition.....98

6.2.1 Plant materials treatment 98
6.2.2 Glycaemic Index evaluation 100
6.2.3 Biochemical analysis.....102
6.2.4 FESEM analysis.....103

6.3 Algorithm development.....104

6.3.1 Image processing.....106
6.3.2 Morphological features calculation.....111

6.4 Case study results 114

6.4.1 Algorithm outputs evaluation 114
6.4.2 Multivariate exploratory analysis.....121

6.5 Design of Experiment to evaluate the effect of the algorithm input parameters.....124

6.5.1 Experimental design.....124
6.5.2 Exploratory analysis.....126

6.6 Algorithm evaluation.....128

6.6.1 Evaluation on the extracted features calculation using ASCA 128
6.6.2 Evaluation of the total processing time using MLR.....130

6.7 Project conclusions.....131

References | Chapter 6.....133

7. General Conclusions.....138

Chapter 1

Introduction

The food industry represents one of the most dynamic and complex sectors of the global economy, as it is driven by the constant need to ensure high standards of quality, safety, and sustainability. In an historical moment marked by growing consumer awareness, stringent regulatory frameworks, and the increasing importance of evaluating the environmental impact, the food industry is continuously facing new challenges. Dealing with these complex difficulties requires innovative approaches capable of exploring intricate production processes, complying with rigorous regulations, and adapting to a series of evolving market conditions. Among many available tools, chemometrics has emerged as a powerful and innovative methodology. By leveraging advanced mathematical and statistical techniques to analyse chemical data, chemometrics enables the extraction of valuable insights from complex datasets, offering innovative solutions that meet the needs of the modern food industry.

This PhD project, entitled "Chemometrics Applied to the Food Industry: Improving Quality, Safety, and Sustainability Using Multivariate Data Analysis", elaborates the potential of chemometric approaches in revolutionizing food industry practices. The research is structured in different chapters that correspond to different projects carried out during the PhD period, ordered according to the complexity of the chemometric techniques employed. These projects correspond to multiple case studies, each demonstrating the application of different chemometric tools to real-world challenges.

The broader implications of this research extend beyond the food industry. Chemometric approaches are widely applicable across numerous domains within chemical engineering, including process optimization, environmental monitoring, and materials science. The ability to efficiently process and interpret complex datasets is fundamental for progressing these fields, leading to more accurate predictions, improved product development, and enhanced regulatory compliance.

Furthermore, the integration of chemometrics with analytical technologies, such as Nuclear Magnetic Resonance (NMR) spectroscopy and Field Emission Scanning Electron Microscopy (FESEM), synergically contribute to drive innovation. In this respect, the current Ph.D. project demonstrates how the combination of robust statistical methodologies with widely applied analytical techniques can provide novel insights, streamline workflows, and ultimately contribute to the development of more sustainable and reliable industrial processes.

The thesis begins with some theoretical foundations, introducing the core principles of chemometrics and emphasizing the pivotal role of multivariate data analysis in understanding the complexity of chemical datasets. When assessing the processes and products of the food industry, whose analytical approach is traditionally characterized by a multitude of techniques, it is very common to face very heterogeneous data, derived from diverse sources as, for example, spectroscopy, chromatography, and imaging techniques. These datasets often exhibit peculiar instrumental bias and high dimensionality, necessitating appropriate pre-processing and modelling methods (based on multivariate statistics) able to reveal hidden patterns, correlations, and underlying trends.

After the first chapter, the thesis extends beyond theoretical discussions, presenting a series of case studies that illustrate practical applications of different chemometric tools across real and current challenges faced by the food industry. These case studies correspond to projects carried out in collaboration either with other research groups from academia or industrial companies.

The first case study (case study #1 - Chapter 3) focuses on the optimization of industrial food processes performed using ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy coupled with Principal Component Analysis, which is considered the simplest and most applied chemometric technique, at least from the point of view of exploratory data analysis. The study demonstrates the feasibility of exploring, monitoring and optimizing an industrial process in the food industry context. This approach not only identifies the key parameters influencing the composition of the final product but also allows to obtain optimal experimental conditions, leading to save both time and resources.

The following case study (case study #2 - Chapter 4) indicates how to use chemometric tools to enhance food authenticity and traceability for fighting food fraud and protecting regional food excellences like the Piedmont TGT hazelnut. By employing a combination of NMR spectroscopy and liquid chromatography coupled with mass spectrometry (LC-MS) and utilising advanced multivariate classification techniques such as Partial Least Squares-Discriminant Analysis (PLS-DA), the study proposes different multivariate classification models able to discriminate hazelnuts based on their geographical origin and cultivar. This research highlights the importance of using chemometric tools to combine datasets originating from different analytical techniques, extracting relevant information from all sources in a synergic way. Data-fusion approaches, coupled with classification models can be fundamental in safeguarding the authenticity of high-value food products, protecting consumers, and supporting the reputation of food industries dealing with traditional products.

Chemometrics, coupled with spectroscopic techniques, can also be very useful to extract significant information from a single large dataset to develop research lines with diverse goals. Chapter 5 (case study #3) of this thesis is focused on a metabolomic research project based on a large wine dataset. Aside from the

metabolomic investigation, this project also explores the analytical challenges encountered when quantifying ethanol in alcoholic beverages using ^{1}H-NMR. Indeed, this task is complicated by the high concentration of ethanol in such matrices and such problem is reported in literature with some non-definitive work around. The study presented in this Ph.D. thesis, for the first time systematically investigates experimental parameters, including NMR pulse sequences, the use of deuterated solvents, and different quantification methods, to optimize accuracy and precision of direct ethanol quantification by means of ^{1}H-NMR. This meticulous approach highlights the importance of procedures performed both during sample preparation and data acquisition to obtain reliable measurements in complex systems. The choice of optimal experimental parameters is fundamental for ensuring product quality and regulatory compliance.

Lastly, probably the most innovative aspect of the thesis (case study #4 - Chapter 6) involves the development of a novel algorithm for analysing images obtained from Field Emission Scanning Electron Microscopy (FESEM). The algorithm was developed and applied to a large study on rice kernels with the aim of automatically extracting morphological features such as area, perimeter, eccentricity, and porosity of starch granules from the acquired FESEM standardized images. These features provide critical insights into the internal structure of rice kernels and their relationship to the glycaemic index, which is a parameter of growing interest in the context of health and nutrition. By offering a more efficient and objective method for analysing FESEM images, this approach facilitates the characterization of rice varieties and enhances the understanding of their suitability for various dietary and culinary applications. In addition, to confirm the advantages that chemometric tools can bring to food industry related topics, a Design of Experiment (DoE) approach was employed to monitor the effect of different parameters on the algorithm performance, for optimizing the algorithm input selection.

Through these case studies, the thesis demonstrates the versatility and efficacy of chemometrics as a powerful tool for addressing a series of pressing challenges in the food industry. Among these, chemometrics enables the detection of adulterations, the monitoring of process parameters, and the improvement of food safety, authenticity, and quality. Furthermore, this work explores concepts such as data fusion, which involves integrating information from multiple analytical techniques to achieve a more comprehensive understanding of food systems. By combining datasets from different analytical methods such as spectroscopy, chromatography, and also imaging techniques, data fusion holds the potential to provide holistic insights into food composition, covering also food processing and quality control.

In summary, this thesis not only reinforces the importance of chemometrics in chemical engineering but also justifies the need for constant research in this field. By providing a systematic approach to data analysis, chemometrics enhances the accuracy, efficiency, and interpretability of complex chemical datasets. The findings presented in this work emphasize its transformative potential, encouraging

future advancements in both academic research and industrial applications with the aim of improving the food industry towards safer, more sustainable, and higher-quality practices.

1.1 Thesis organization

The thesis is structured in seven chapters, including the Introduction (Chapter 1) and the General Conclusions (Chapter 7). A brief summary of each chapter is reported in the following list:

Chapter 2: Metabolomics and foodomics concepts, techniques employed and chemometric methods applied in the data analysis steps.

Chapter 3: Case study #1 focused on the optimization of industrial food treatments. In this case study principal component analysis is used to explore, evaluate and optimize an industrial treatment applied to a sample of lentil flour.

Chapter 4: Case study #2 focused on the application of data fusion approaches to improve authenticity and traceability. In this study, a data fusion approach is used to merge data obtain with Nuclear Magnetic Resonance spectroscopy and Liquid Chromatography coupled with Mass Spectrometry analysis with the aim of improving authenticity and traceability concepts related to the Piedmont TGT hazelnut variety.

Chapter 5: Case study #3 focused on new challenges in wine metabolomic analysis performed using ^{1}H-NMR. In this study, an experimental design was developed before acquiring a large dataset with more than 200 wine samples. Starting from this dataset, different methods to quantify the ethanol content in alcoholic beverages were explored.

Chapter 6: Case study #4 focused on the development of a new algorithm able to automatically perform image analysis on greyscale images. This algorithm was applied in a project developed to explore correlations among structural properties, biochemical traits and glycaemic index from more than 50 rice varieties. In addition, a design of experiment approach was employed to evaluate and improve the algorithm functioning.

Chapter 2

Materials and methods

2.1 Metabolomics and Foodomics

Metabolomics [37–40] and foodomics [41,42] are two consolidated scientific approaches that combine chemistry, biology and data science by offering innovative insights into the composition and functionality of biological systems.
Metabolomics focuses on the comprehensive analysis of metabolites, which are the small molecules involved in metabolic processes within organisms. Their presence and relative amounts provide information about biochemical activities and are crucial indicators of health, safety, and nutritional value in food systems. Foodomics expands on this concept, including also the study of food and nutrition through the integration of other omics technologies like genomics and proteomics [43]. It aims to explore the relationship between food components and their effects on human health, emphasizing personalized nutrition and the development of functional foods [44]. This multidisciplinary approach enables a deeper understanding of food at molecular, cellular, and systemic levels.
The application of metabolomics and foodomics in the food industry is vast and varied. They are often employed to detect contaminants [45,46], authenticate food products [47,48], prevent frauds [49,50], but also monitor processes like fermentations [51] and other treatments [52]. Advanced analytical techniques, such as Nuclear Magnetic Resonance (NMR) spectroscopy [53–56], Liquid or Gas Chromatography (LC and GC)[57–59] and Mass Spectrometry [60–62](MS) generate high-dimensional datasets that require sophisticated data analysis and chemometric tools to be interpreted.

Both fields heavily count on robust data analysis and model validation to ensure accuracy and reliability. The data analysis steps of preprocessing, feature selection, and statistical modelling are fundamental for transforming complex datasets into meaningful information. Focusing on food science and health, metabolomics and foodomics offer innovative pathways to improve food quality, safety and sustainability, thus contributing to the well-being of consumers and the evolution of the food industry.[63,64]

2.2 Analytical techniques

2.2.1 Nuclear Magnetic Resonance (NMR)

Nuclear Magnetic Resonance (NMR) spectroscopy is an analytical technique largely used in the field of metabolomics [65,66] and food science [67]. NMR spectroscopy, due to the nature of its output (i.e., the NMR spectra) and to the large amount of information contained in the resulting data (i.e. spectra), can be used as a fingerprint technique, which is very appealing also for foodomics studies. Some examples of NMR applications in food science [68] include the elucidation of complex mixtures chemical composition [69–73], its employment for detecting and preventing frauds [49,74], the authentication of valuable food products [75,76], and it is also utilized for quality control purposes [77–79].

The main advantages of NMR spectroscopy are that it only requires little sample preparation and, since it is a non-destructive technique, it is possible to perform more than one measurement on the same sample. Speaking about the physical principles upon which this spectroscopic technique is based, NMR exploits the magnetic properties of certain active nuclei to provide detailed information about the chemical composition of complex mixtures and to investigate for the compositional and spatial structure of target molecules. The NMR signal is produced by the excitation, by means of generated radio frequencies, of the sample's active nuclei exposed to the effect of a strong external static magnetic field. The interaction between this magnetic field and the nuclei is possible only when the spin number of the studied element is different from zero: for this reason, only nuclei such as ^{1}H, ^{13}C, ^{15}N, ^{31}P, (and some others) are suitable for NMR analysis. Nuclei with this characteristic magnetic property, when placed in a strong magnetic field, increasing the energy distance between the distinct quantized levels; when a radiofrequency pulse, set at the nucleus resonance frequency, is applied, the nuclei start to resonate with the magnetic field and generate a signal. The generated signal, called "free induction decay" (FID) is a time domain resonating signal that is then transformed, using Fourier Transform, into a spectrum in the frequency domain, generating the NMR signals that are characteristic of the investigated nuclei of the molecules present in the analysed sample. The frequency at which the nuclei resonate is strictly related to their chemical environment, and, for this reason, only truly equivalent nuclei generate the same specific signals, while same nuclei that occupy different positions within the molecule structure, will generate different signals. Since the resonance frequency is characteristic to individual nuclei, each signal can be associated to a specific kind of nucleus in a specific position within the different functional groups composing the molecule, allowing the extraction of detailed structural information from the NMR spectra.

Among the many kinds of information that an NMR spectrum can provide, the first one to evaluate is the chemical shift (δ), which indicates the signal position in the NMR spectrum, and it is strictly related to the chemical environment affected by

the electrons that surround each nucleus (magnetic shield). Chemical shifts are conventionally reported in ppm (parts per million) and the values of this scale are measured relatively to the signal of a reference compound, allowing comparability among spectra generated from different NMR spectrometers. For example, tetramethylsilane (TMS) and 3-(trimethylsilyl)propanoic acid (TSP) are two of the most commonly used reference compounds for ^{1}H-NMR analysis and their protonic signal is set to represent the zero value in the ppm scale.

Together with chemical shift, the signal's multiplicity is fundamental to decode and assign groups of signals to functional groups of the studied molecules. The signal multiplicity results from spin-spin coupling interactions and generate doublets, triplets and more complex patterns. An example of an NMR signal is given in Figure 2.1.

Another relevant information carried by the spectral output is the signal intensity or the signal area, which is directly proportional to the number of identical nuclei contributing to generate the signal. Thanks to this feature, NMR spectroscopy is a primary method for quantification and therefore suitable for straight quantitative analysis and also for determining ratios between different components. In addition, the NMR outputs provide other relevant piece of information such as the coupling constants (J), i.e. the distance in ppm between peaks of the same multiple signal which carry information about bonds, angle and stereochemistry, and the relaxation times, measured by means of the FID in the time domain, which are related to molecular dynamics and interactions with the environment.

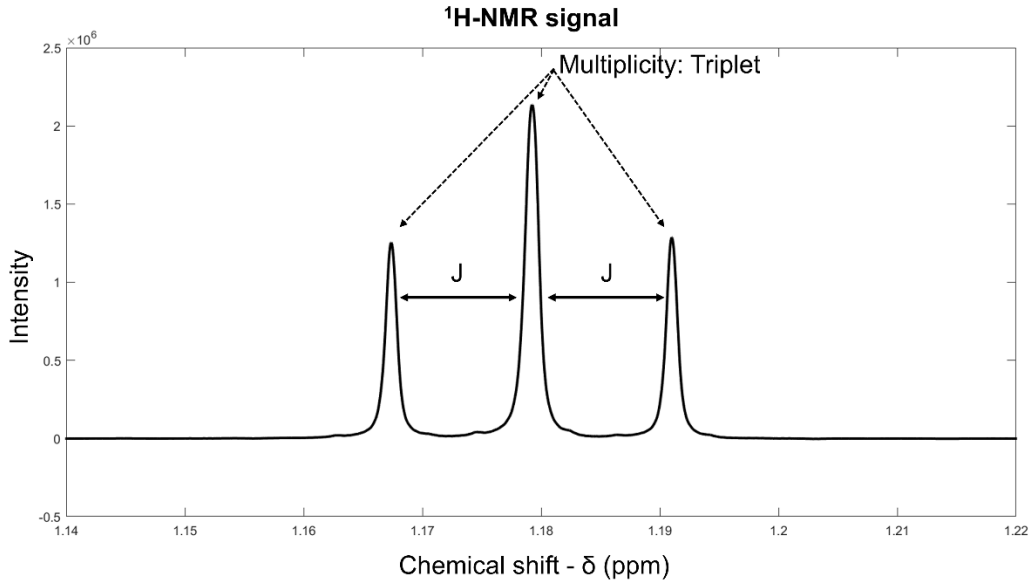

Figure 2.1. Example of a typical ^{1}H-NMR signal with its chemical shift, intensity, multiplicity and coupling constant (J). The reported signal belongs to the CH_3 functional group of the ethanol molecule.

2.2.2 Field Emission Scanning Electron Microscopy

Field Emission Scanning Electron Microscopy (FESEM) is an advanced imaging technique, mainly applied in materials science and nanotechnology to investigate microstructures or surface morphology with a very high resolution. In general, a scanning electron microscope (SEM) uses a focused beam of electrons to scan the sample surface: when the electrons interact with the sample, they produce signals (like secondary electrons or backscattered electrons) which are detected to form high-resolution images and provide compositional information.

With respect to the classical SEM, the FESEM has a different electron source. In brief, it uses a "field emission gun" that offers a highly focused electron beam with much higher brightness and resolution. As a result, it can achieve nanometer-scale imaging also for non-conductive samples.

The signals generated by the samples are then collected in two main detector types: the secondary electron detector that, reveals surface details like texture and morphology, and the backscattered electron detector, which is more precise for compositional analyses and differentiation between materials. Some details need to be added about these electron types: while the primary electrons are the one emitted by the field emission gun to interact with the samples, the secondary electrons are the ones ejected from the sample due to inelastic scattering interactions with primary electrons. These secondary electrons are defined as low energy electrons; the backscattered electrons are instead defined as high energy electrons, and they are elastically scattered back from the sample without a significant loss of energy.

2.3 Chemometric methods

2.3.1 Introduction to chemometric concepts

The term "Chemometrics" was first introduced by Svante Wold in 1971. A general definition of Chemometrics was given by Wold in "Chemometrics and Intelligent Laboratory Systems", the leading journal in the field: *"the chemical discipline that uses mathematical, statistical, and other methods to design or select optimal measurement procedures and experiments, but also to provide maximum relevant chemical information by analysing chemical data"* [1,2].
In brief, chemometrics is a field that integrates chemistry, mathematics and statistical methods to extract meaningful information from complex chemical data [3]. In modern chemical industries, the application of chemometrics has become essential due to the increasing demand for high-quality [4,5], safe [6], and sustainable processes [7–9]. By using multivariate data analysis, chemometrics enables researchers to reveal patterns, make predictions, optimize processes [10,11] and improve decision-making and efficiency [12,13].

Multivariate analysis is a core component of chemometrics. Chemical engineering data often involve multiple correlated variable that must be analysed simultaneously. Differently from univariate approaches, that consider one variable at a time, multivariate methods allow for the exploration of complex interactions and correlations among variables, resulting in a more comprehensive understanding of the studied system. These techniques are particularly useful when variables exhibit collinearity, meaning that they are highly correlated and carry redundant information.

The ability of chemometrics techniques to handle large datasets characterized by high dimensionality and complex relationships among variables, make these approaches suitable for many research problems faced by industries. Dimensionality reduction, key variables identification and samples classification are just a small part of the goals that can be achieved using chemometrics. Many tools result to be particularly suited for the food industry [14–16], where data often arise from diverse sources, including spectroscopy, chromatography, sensory analysis, and environmental monitoring.

For what it concerns more specifically the context of food industry, chemometrics plays a pivotal role in challenges like detecting adulteration [17–19], monitoring process parameters [20], and ensuring compliance with regulatory standards [21,22]. Furthermore, chemometrics contributes to economic and environmental sustainability by saving resources and reducing waste, aligning with global efforts to create more environmentally friendly food systems [23].

Nowadays chemometrics methods can be used to face a large variety of data types and challenges. Data analysis involves a series of sequential steps that allow obtaining robust and reliable statistical models. These steps involve an initial preprocessing phase [24,25], like scaling and normalization that ensure that data are suitable for analysis. Depending on the scope, the chemometrician may need to build *unsupervised* models [26,27], when the aim is to explore the data, or *supervised* methods [28] when the need is to model a response, like a chemical or physical property or like belonging to a specific class Unsupervised methods, e.g. Principal Component Analysis (PCA) [29,30], are generally used for data exploration to obtain a first overview of the information contained in the data. Supervised methods are usually more complex and developed with specific targets. In these cases, there is a predictive goal, both in terms of predicting a specific response, like in regression [31], or predicting a class of belonging by applying classification methods [32,33]. Generally, a common approach to data analysis involves a first step of *unsupervised* data modelling, to both gain knowledge about the data and/or the process generating the data and to decide further direction to take [34]. Lastly, it is fundamental to mention model validation [35,36], which is a critical step aimed at guaranteeing that both the findings and the developed models are robust, reliable, and generalizable.

2.3.2 How to correctly treat the data: dataset organization and preprocessing methods

A well-structured dataset is the foundation of a robust and reliable data analysis process. Typically, data should be organized in a matrix format where rows correspond to individual samples, and columns represent measured variables. These variables can be spectral intensities with specific wavelengths, chromatographic peaks, or other chemical descriptors and properties. Often, for each samples other information need to be reported. For example, an additional column vector that contains class labels can be included to indicate the category to which each sample belongs. If each sample has generated also a response variable, these variables should be stored as another column vector that contains all the responses sorted following the sample position (the rows) in the dataset matrix. Structuring data in this way ensures compatibility with different multivariate analysis techniques and allows for a clear distinction between samples, variables and responses or class labels, facilitating further data analysis steps and models development.

After the dataset organization, the first step of data analysis should always be data preprocessing [24]. With preprocessing the aim is to improve the quality of the data and, as a consequence, the performances of any developed models. Raw data, such as spectral measurements, often contains noise and variability not useful for the problem faced. By applying appropriate preprocessing techniques, the data are prepared for subsequent analysis focusing on meaningful chemical information, while minimizing the influence of irrelevant variables and unwanted source of variance. There are two main families of preprocessing methods: vertical, that involve operations applied to the single variables, and horizontal, that focus on the single sample and are generally aimed at ensuring comparability among all the samples of the whole set [80].

Examples of vertical preprocessing methods are mean centering [81] and scaling [82] that are used to better compare the different variables. Mean centering is almost always used with continuous data (i.e., spectral profile) and it consists in centering the data around zero; autoscaling, the most used scaling method, is used with discrete data and each variable is mean centered and divided by the standard deviation of its own column. Other examples of vertical methods are baseline corrections, like polynomial fits used to adjust baseline drifts [83,84], or derivatives, that eliminate baseline shifts and better highlight peaks [85,86].

Speaking about horizontal methods, very common examples are the normalization tools [87], which are used to scale each entire sample individually, with the aim of reducing variability among the samples and remove unwanted systematic biases. When working with spectroscopic data, smoothing methods [88] allow removing noise in the spectral resolution, while peak alignment methods improve the comparability among spectra of different samples. One of the most used alignment methods in spectroscopic and chromatographic applications is "*i*coshift" [89,90], which is a quick and intuitive tool for correcting horizontal shifts in signals,

particularly common in data such as Nuclear Magnetic Resonance spectra or chromatograms. This method ensures that comparable features across samples (i.e., peaks or signals) are aligned to the same position, which is essential for accurate multivariate analysis. In brief, "*i*coshift" allows aligning the spectra either globally or by dividing them into user-defined intervals, which are then aligned independently. This flexibility helps addressing local shifts without affecting other regions of the spectrum/chromatogram.

2.3.3 Unsupervised approaches: exploratory analysis and Principal Component Analysis

Unsupervised approaches are data analysis techniques used to identify patterns, structures or relationships within datasets without considering any external responses or additional labelled information. In chemometrics, unsupervised methods are employed to analyse multivariate datasets to generate hypotheses, identify outliers, and improve knowledge on the data to employ supervised methods with a deeper awareness in further modelling steps.

Unsupervised methods are often used for exploratory analysis [91], which is commonly employed in the initial phase of data analysis to better understand complex datasets, reduce their dimensionality and identify patterns, trends and even anomalies or outliers. In brief, the aim of exploratory analysis is to reveal hidden and unknown information in a specific dataset. Some examples of unsupervised methods are Multivariate Curve Resolution (MCR [92]), the already mentioned Principal Component Analysis (PCA, [93]) and clustering techniques like k-means [94] or hierarchical clustering [95].

Data exploration is of fundamental importance to obtain an immediate, direct and easy visual representation of the information contained in the data. Especially with complex and multivariate data, the exploratory analysis is the best way to start the data interpretation phase, with the aim of looking for patterns among the samples and evaluate variables effects, assess the overall quality of a dataset and reduce noise by removing unwanted variance.

Among the exploratory methods, the most utilized is by far Principal Component Analysis (PCA) [93,96–99]. PCA is a multivariate statistic approach able to highlight the correlations among the properties used to describe the data (i.e., the variables), and which also allows to explore the samples distribution. It is a decomposition method that is used to reduce dataset dimensionality by transforming the original dataset (organized in samples and variables) into a new set of orthogonal and therefore uncorrelated summary variables, called Principal Components (PCs), that represent the sources of maximum variance. The results of this approach allow visualizing the input data in a new set of coordinates identified as the Principal Components space.

The PCs are obtained by linear combination of the original variables, and they are ordered by the amount of explained variance, meaning that the first PC describes the absolute largest source of variance in the data, the second PC describes the second largest source of residual variance in the data, and so on. Because of their nature, while modelling the data, it is important to define the correct number of PCs to be considered when modelling the data because, by including all the possible PCs, the resulting model would contain not only the significant variance but also the noisy unwanted one. A commonly used approach to define how many PCs should be considered to model a dataset is to look at the total explained variance of the different PCs, and selecting PCs until the total explained variance reaches 80–90 %. This number is strictly related to the nature of the original dataset. To better evaluate the contribution of each PC, a so-called "scree plot" can be used which allows visualizing the total explained variance and the one explained by each PC. In brief, the X axis represents the principal components (ordered from first to last), while the Y axis represents the explained variance for each PC. With this graph, the contribution of each PC to the total explained variance can be easily visualized, making the choice simpler of how many PCs should be selected for the exploration of the dataset.

The mathematics behind PCA can be summarized in a bilinear decomposition described by Equation 2.1:

$$X = T \cdot P^T + E = \hat{X} + E \qquad (2.1)$$

In the above equation, for which the visual representation is reported in Figure 2.2, X is the original dataset containing I objects described by J variables; T is the scores matrix that has as many rows as the sample number (I) and as many columns as the number of the selected PCs (F); P^T is the loadings matrix that has the same number of column of the original dataset (J, the number of variables) and as many rows as the selected PCs (F); finally, E is the residuals matrix containing the unmodelled part of the data. The second form of the equation is commonly used to distinguish the modelled part of the dataset $\hat{X}$ and the unmodelled one E. In theory, in a good PCA model, the whole structured variance should be included in $\hat{X}$, while all the random and uninformative variance should be retained in the residuals matrix E.

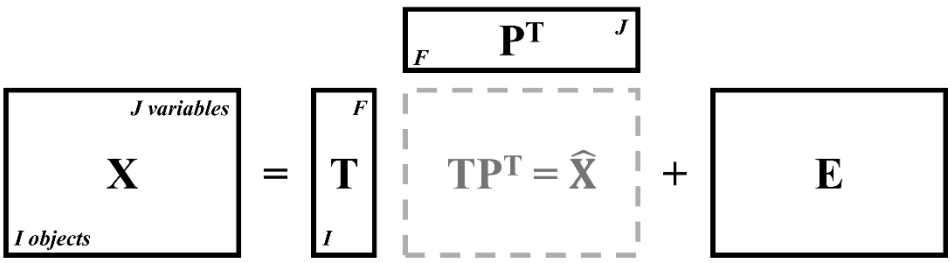

Figure 2.2. Graphical representation of a PCA model.

To visually inspect the samples' distribution and clustering tendencies, the scores matrix T is commonly plotted one PC scores against another PC scores values in the so-called "scores plot". The scores are the coordinates of the samples in the new PC space, and they are used to describe how the samples distribute in the PC space. Furthermore, the loadings matrix P^T contains the loadings, which are instead the coefficients of the original variables in the linear combinations that determine the different PCs. From this last matrix, another important plot can be generated: the so–called "loadings plot".

The scores plots allow finding clusters of similar samples, trends related to specific properties and even possible outliers. In addition, the distance between samples in the PC space indicates similarity or dissimilarity with respect to the explored PCs. The loadings plots instead, allow inspecting the variables' distribution and help interpreting the correlations among them. In brief, it is mainly used to see the influence of the different variables on the samples with respect to the explored PCs. Since they describe complementary information, the scores and loadings plots can be interpreted together by plotting the same pairs of principal components. In this way, the same portion of the PCs space is fully inspected, both from the point of view of the samples and of the original variables in a single plot that is commonly known as "biplot".

A third informative visualization that can be useful for data interpretation, and particularly to inspect the sample effect on the model, is the residuals plot. This plot contains information about the distance of samples from the developed model and it is obtained by plotting two quantities one against the other: the Q residuals, which define how the sample is well-described and well-represented by the model (high Q residual means that the sample is not well-described by the model), and T^2, which defines the effect of the sample on the model creation (high T^2 means that the sample has a strong contribution on the model development).

2.3.4 Supervised approaches: Regression and Classification methods

Differently from unsupervised approaches, for which the main aim is to explore the data with no *a priori* hypothesis and without using quantitative responses or class information, supervised methods are used to build models able to predict a quantitative response or a qualitative class depending on the aim of the research and require labelled data and/or reference values.

There are mainly two types of supervised approaches: regression methods and classification methods. Regression is a statistical approach used to model the relationship between one or more independent variables and a response [100,101]. The aim of these methods is to create a model, based on the above-mentioned relationships, able to predict the response from new samples or inputs with

unknown response. Regression methods are widely applied to predict, for example, chemical concentrations and other properties for which a quantitative response can be obtained. The most used regression method is Partial Least Squares (PLS) regression that will be better explained in the next chapters.

Classification methods instead differ from regression methods since the response they model is of qualitative nature, like for example belonging to a specific class [9,102]. The idea of classification analysis is to model the relationship between variables and a categorical response, with the aim of creating models able to predict classes of unknown samples. Classification methods are commonly used for tasks such as identifying the origin of a product [103–105], detecting food varieties [106,107] and adulterations [108,109], or classifying materials [110,111] based on spectroscopic data.

Among classification methods two main families can be identified, namely, class modelling [112] and discriminant analysis [113]. The former focuses on modelling a single class of interest and distinguishing it from all the other data. An example is Soft Independent Modelling of Class Analogy (SIMCA) [114]. Instead, the discriminant methods directly discriminate between two or more classes by developing decision thresholds that are valid on the whole samples' domain. With this approach all samples are always assigned to one of the predicted classes and the most used method is Partial Least Squares Discriminant Analysis (PLS-DA) [102] which will be better explained in the next chapters.

Partial Least Squares Regression (PLS)

PLS regression is a statistical method used to model the relationship between a set of independent variables and one or more dependent variables (responses) [100]. With respect to the PCA theory (Section 2.4), where the aim is to search for the maximum variance within the X data matrix, the aim of PLS is to maximize the covariance between X (data matrix) and Y (response). The simplest mathematical representation of the PLS theory is described in Equation 2.2 where X is the data matrix to model, Y is the column vector of the response (which is a matrix in case of multiple responses) and B contains the regression coefficients that are modelled to maximize the covariance between X and Y.

$$Y = XB \qquad (2.2)$$

In brief, this equation allows to predict a Y response starting from an X matrix by using a series of regression coefficients (B). So, PLS, as well as other regression tools, can be used for building predictive models that relate measured spectra or other analytical signals to chemical or physical properties of samples. Since the response is quantitative, these models can be used to quantify concentrations (like in our project), classify compounds, or predict sample characteristics.

Similar to PCA Equation 2.1, PLS regression can be represented with two equations that explain how the data decomposition is performed for X matrix (Equation 2.3) and Y response (Equation 2.4).

$$X = T \cdot P^T + E = X \cdot W + E \qquad (2.3)$$

$$Y = U \cdot Q^T + E \qquad (2.4)$$

In particular, *T* and *U* are the matrices of scores for *X* and *Y* respectively, while *P* and *Q* are the corresponding loadings matrices. In addition, *W* is the weights matrix, and it defines the directions in *X* that maximize the covariance between *X* and *Y*. Differently from PCA, the *T* scores matrix is calculated as linear combinations of original variables using the weights matrix *W*. Lastly, *E*, as for the PCA equation, represents the residual matrix that contains the unexplained variance. To completely understand the PLS regression workflow, a further equation needs to be explained. Equation 2.5 describes how the coefficients matrix *B* is computed starting from loadings *P* and *Q* and from weights *W*.

$$B = W \cdot (P^T \cdot W)^{-1} \cdot Q^T \qquad (2.5)$$

An important step in the model creation is the choice of how many components use (called "latent variables" for PLS models). The correct choice of the number of latent variables to use must be driven by the aim of obtaining models with good prediction ability both on the data use to calibrate the model, but also to external data unknown with respect to the calibration dataset. The focus must be to avoid underfitting and overfitting that will generate non-robust and non-reliable models.

The main graphical outputs of a PLS model are the prediction plot and the leverage plot. The first is useful to visualize the experimental response (real) plotted against the response predicted by the PLS model. The possibility of visualizing how the samples are predicted with respect to their real value allows also to identify and evaluate potential outliers or misclassified samples. The leverage plot instead shows how each sample is influencing the model and it can be used to spot both samples important for the model construction and samples that can badly influence the model development.

Two other main outputs of a PLS model are two parameters used to assess the quality of the regression fit and of its prediction ability: the squared correlation coefficient (R^2) and the Root Mean Squared Error (RMSE). The R^2 coefficient defines the correlation of the data with respect to the regression response; it ranges from 0 to 1 and a high value of R^2 indicates a high correlation of the data with the response. RMSE instead defines how much the predicted responses are far (high RMSE) or close (low RMSE) with respect to the real response.

Partial Least Squares Discriminant Analysis (PLS-DA)

Partial Least Squares Discriminant Analysis (PLS-DA) [33,102] is one of the most used statistical methods for classification analysis. Classification tools in chemometrics are essential for classifying samples based on their chemical or physical properties, typically derived from spectral or analytical data. These tools help identify the class or category to which a sample belongs, facilitating applications like quality control, authenticity testing, and compound identification.

PLS-DA is a multivariate supervised method based on the application of Partial Least Squares (PLS) algorithm that allow to build a model able to predict a class information (coded in the matrix Y) from a set of data (X). PLS-DA looks at the covariance between X and Y (the class belongings) and allows to find and model the information of X most related to the class assignments reported in Y. The model equation (Equation 2.6) is used to predict a Y class information starting from an X matrix (defined by a set of variables) by using a series of regression coefficients (B).

$$Y = XB \qquad (2.6)$$

In PLS-DA the response is categorical, and the class information is described by a "dummy matrix", which is a binary matrix that indicates class membership of each sample. Among the different classification methods, PLS-DA is recommended with large datasets, and it is particularly suitable when only two classes are considered.

The development of a PLS-DA model allows the generation of different outputs, both graphical and numerical. Starting from the first, the classification plot helps in the sample visualization with respect to the thresholds defined to separate the classes. Another important graphical output is the leverage plot, which shows how much each sample is influencing the model development and can be used to spot potential outliers and important samples. Other important outputs are the classification parameters which are related to the model classification ability; the most useful are the error rate, that define the percentage of misclassified samples with respect to the total number of considered samples, and the accuracy, which is instead used to estimate the model error. Other classification parameters can be calculated starting from the confusion matrix, which is a matrix that contains the class predictions and allows to compare the predicted class of a sample with respect to its real class [115].

Speaking about the variables instead, the main outputs of PLS-DA are the regression coefficients and the variables importance in projection (VIP) scores. In brief, they are used to interpret the model in relation to the original variables and help in understanding which variables are the most important for the classification [116].

2.3.5 Data fusion approaches: fundamentals and advantages

Data Fusion methods are strategies that combines data originating from different analytical techniques to enhance the extraction of meaningful information and improve decision making processes. In chemistry and chemical engineering, data fusion is a widely used chemometric approach especially where complex systems are analysed using techniques that generate heterogeneous data types such as spectroscopic and chromatographic datasets.

Data fusion methods are usually divided into three families, low-, mid- and high-level methods [117–119] as represented in Figure 2.3.

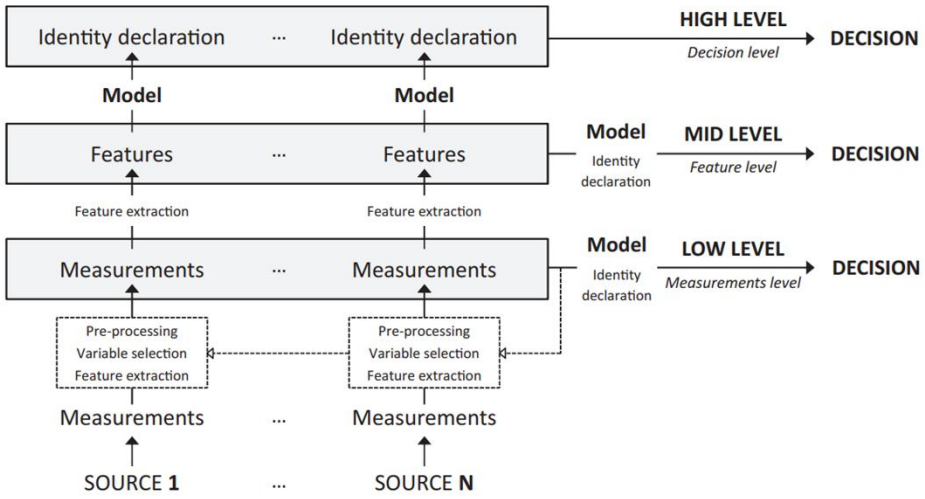

Figure 2.3. Graphical visualisation of different data fusion strategies. (from Borràs et al. [117])

In low-level data fusion, two or more blocks of data are simply joined together by concatenating raw data matrices obtained from different sources on the same samples. The resulting matrix has the same number of rows of the raw data matrices, corresponding to the number of samples, and as many columns as the total sum of the number of columns of each single data matrix, corresponding to the sum of the number of variables considered for all the employed techniques. Practically speaking, this approach preserves the original dataset structure but usually requires a carefully weighted data preprocessing to ensure comparability among datasets.

For mid-level data fusion approaches [118,120] a series of relevant features separately extracted from the individual datasets are merged together to form a new data matrix. These methods combine data reduction and extraction of relevant information to improve and simplify the modelling phase. These extracted features can be for example PCA scores, or also scores obtained from regression and classification models [120].

High-level data fusion approaches are more complex since, at this level, decisions/predictions derived from different analyses/models are combined. These analyses usually are supervised models, and for this reason they are also named "decision level fusion". These approaches are particularly useful with more complex systems, with data showing completely different structures.

By combining information coming from different and compatible techniques, data fusion approaches lead to improve model performance and robustness, and to obtain a more comprehensive understanding of the system under study. The use of the correct data fusion method can also improve detection of anomalies, in the case of process monitoring problems, and classification, by highlighting small differences among samples of different classes.

Previous works suggested the great potential of data fusion coupled with chemometrics and multivariate statistical analysis in food authenticity for characterizing the quality of different foodstuffs, including olive oil, honey, fish and meat [121–124].

2.3.6 Multivariate Curve Resolution (MCR)

Multivariate Curve Resolution (MCR [125,126]) is a chemometric technique used to decompose complex data, such as overlapping signals, into their pure contributions (e.g., pure spectra or concentration profiles). From a mathematical point of view, MCR is related to PCA, but it does not require orthogonality among the components. However, MCR often incorporates constraints (e.g., non-negativity, closure) to reduce rotational ambiguity. MCR decomposition can be represented as in Figure 2.4 and summarized in Equation 2.7, where X is the data matrix, C is the concentration matrix (pure component concentration), S is the spectral matrix (pure spectra profile) and E is the residuals matrix containing the unmodelled variance.

$$X = CS^T + E = \hat{X} + E \qquad (2.7)$$

Similar to PCA, the data matrix (X) is decomposed into two matrices C and S^T. The concentration matrix (C) is the analogue of the PCA scores, while the spectral matrix (S^T) is the analogue of the PCA loadings.

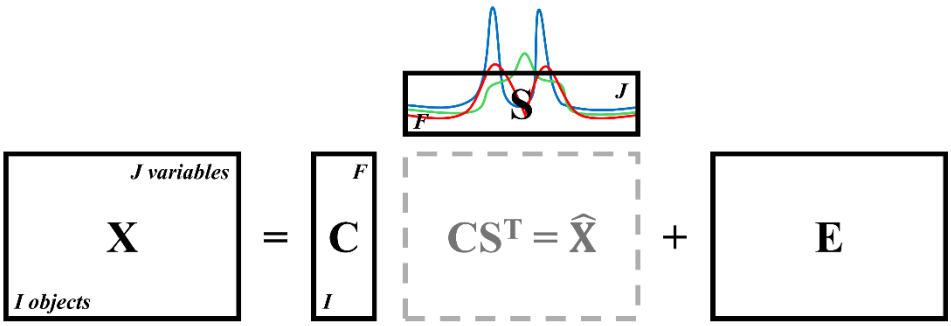

Figure 2.4. Graphical representation of MCR decomposition

Given that pure signals are obtained, MCR can also be used as a method for filtering the data, since undesired sources of variability (e.g., noise or background effects) can be removed, and end up in the residual matrix (*E*).

In NMR applications, MCR is particularly suitable for resolving overlapped signals in manually selected intervals where a few signals are present and can be more easily extracted by the algorithm as MCR components. Each extracted component is characterized by a pure spectral profile, making it easier to match it with known chemical compound profiles. In addition, metabolites quantification from complex ^{1}H-NMR spectra using MCR has proven a valuable approach for the unbiased quantitative screening of hundreds of spectra simultaneously [127].

Using the MCR approach, the complex ^{1}H-NMR signals can be decomposed into a set of relative concentrations and resolved pure spectral profiles (in our case, the ^{1}H-NMR resonances). For each signal, the obtained relative concentration can be considered as the signal area, and they can be used for quantitative purposes.

2.3.7 Process optimization and Design of Experiment

The optimization of industrial processes is a critical task in the field of chemical engineering. This is mainly due to the high number of variables influencing the process, such as efficiency, employed resources and industrial scalability. Traditional approaches, like the classical "Trial-and-error" method are often inefficient and inappropriate, most often leading to non-optimal procedural conditions. In the context of multivariate data analysis, the Design of Experiments (DoE) approach offers more reliable solutions, overcoming the several limitations of trial-and-error methods, when working with process optimization problems [128,129].

DoE is a statistical approach that systematically studies the effects of multiple factors on one or more response variables, enabling an efficient exploration of the experimental space of the studied problem. The main advantage of this approach

with respect to trial-and-error methods is that DoE allows to minimize the number of experiments required to obtain statistically significant insights and to reduce both resource consumption and experimental time required.

DoE methodologies are particularly suitable for chemical engineering problems since its basic concepts enable the systematic simultaneous investigation of multiple factors influencing entire chemical processes [130]. On the basis of these concepts, DoE can be at first used to plan the experiments to be performed [131–133] by specifying in advance the total number of experiments and which combinations of factors to perform. In a second instance, starting from the information obtained from the previously selected experiments, and employing other multivariate statistical methods like ANOVA Simultaneous Component Analysis (ASCA) [134] or Multi-Linear Regression (MLR) [135], a series of indications about the optimization of the studied process can be obtained [136].

When dealing with DoE, a correct terminology must be introduced, since terms like "Factors" and "Levels" are of fundamental importance, as well as a clear specification of one or more response variables to study. The term "Factor" is used to represent the variables that potentially have an effect on the studied response and that can mostly influence the output of the process. The term "Level" is referred to the different factor values employed during the problem evaluation. Selecting appropriate levels ensures a correct cover of the experimental space and leads to reliable solutions. The choice of which factors and how many levels to explore is often based on literature, preliminary exploratory studies or previous knowledge and experience.

ANOVA Simultaneous Component Analysis (ASCA)

ASCA [137,138] is a statistical method that combines Analysis of Variance (ANOVA) and Principal Component Analysis (PCA). The aim of this approach is to explore multivariate data structured according to an experimental design in order to simultaneously evaluate the effect of different factors (i.e., experimental conditions). With respect to the traditional ANOVA, where the factors effects are analysed in relation to a single response variable, ASCA extends this to multivariate data, using as a response a matrix with multiple variables.

In brief, ASCA separates the variability in the dataset into components related to the different factors and their interactions (using ANOVA). Secondly, once the variability associated with each factor is isolated, PCA is applied to each component to identify the overall effect of each factor. The main output of an ASCA model is the factor effect, expressed with the corresponding p-value calculated with a permutation test to assess the statistical significance of the observed factors effects. In addition, scores and loadings plots can be obtained for each factor with the possibility of visualizing also the levels tested to evaluate samples distribution and

variable contribution. Like in PCA, also the residual variance in the dataset can be obtained.

Multi Linear Regression (MLR)

When applying Design of Experiments approach, a commonly used way to evaluate combinations of tested factors and levels is the employment of Multi Linear Regression methods (MLR,[101]). This approach allows to model an experimental domain (i.e., the set of tested combinations of factors) by defining a mathematical function to describe the experimental data. The MLR function is generally expressed as:

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_1X_2 + B_4X_1^2 \ldots + B_nX_n + R \qquad (2.8)$$

In Equation 2.8 Y represents the response, so the dependent variable; X_1, X_2, X_n are the independent variables (i.e., the factors); B_0 is the intercept; B_1, B_2, B_n are the coefficients representing the effect of each factor (linear terms, interaction terms and quadratic terms) and R is the residual error.

The strong point of MLR is the possibility of describing both the possible effects of the factors and their interactions: this is done by including different terms in the model's mathematical equation. Each term describes a contribution to the modelled response, as related to a factor or an interaction between factors. By inspecting how the response behaves across the experimental domain it is possible to identify interesting sets of experimental conditions, especially in the perspective of response optimization (i.e., maximize or minimize the response).

One of the most important outputs to be inspected in an MLR model are the regression coefficients β, which are generally displayed as a bar plot in which each bar clearly describes the magnitude and the sign of the different modelled terms (related to individual factors or to pairwise factors' interactions). The coefficients are strictly linked to the modelled response, and their global and additive relationship with the response is to be inspected with the response surface plots, which allow to visualize the response relative to the experimental domain. In this study, also the processing time of the algorithm was modelled as a response.

2.3.8 Model validation

When working with chemometrics and data analysis, the main outcome of every study is one or more multivariate statistical models. These models, to be robust and reliable, must be adaptable to data and samples that differ from the one used to develop/train them, and the best way to ensure model adaptability and robustness is to perform accurate model validation [35,36].

In general, model validation is used to assess the ability of a model to make accurate predictions on data obtained from new and unknown samplings. In chemical engineering, this is of fundamental importance in contexts like quality control, process monitoring or decision-making, since new data are continuously generated and submitted to existing model. Every time a new model is created, validation is applied also to avoid overfitting and underfitting situations. The first situation is faced when a model is too complex and able to perfectly fit only the data used to develop it, resulting in a model unable to perform predictions on unseen data. The second situation is instead exactly the opposite, and it corresponds to a model which is too simple, and its predictions are therefore not reliable.

There are different ways to validate a chemometric model but, to keep this chapter short and to avoid introducing too many theoretical concepts, only three validation methods will be explained: test set-validation, external validation and cross-validation.

The test set-validation method is usually employed when enough samples are available (usually more than 50). It consists of splitting the original dataset into two smaller datasets: the calibration set and the test set. These two sets contain samples, and their respective responses or class information, as taken from the initial data. Usually, the set dimensions, with respect to the original dataset, are around 70 % for the calibration set and around 30 % for the test set. This "splitting phase" is very important because the calibration set must contain enough information to build a reliable model, and, at the same time, the test set must be representative of the studied population. It is of fundamental importance that the test set is never used for calibrating the model: to trust the predictions obtained for this set, the model must be completely independent from it. The way the split is performed depends on the amount of available data and their nature but can be also based on human experience. Speaking about which samples to include in the calibration set and which one to include in the test set, there are many algorithms that allow performing an optimized set separation. Some examples are the Kennard-Stone [139], the Duplex and the D-optimal approaches [140,141]. In addition, if several samples are available, another common approach is based on a random sample selection, which avoids introducing biases. Indeed, when the modelling phases involve a variable selection step, approaches like Kennard-Stone and D-optimal may generate biases since slitting is done with a set of variables which is different with respect to those used for the final supervised modelling. Once the splitting is performed, the calibration set is employed to build a model. When the model development is completed, the test set is projected onto the optimized model to test its performances by comparing the results predicted by the model with the real responses of the test set. If the predictions performed by the model agree with the real test set responses, the model can be defined as validated.

A further way to perform model validation using a similar approach is the external validation. This method is considered more rigorous than test set validation because it involves testing the model on an entirely independent dataset. Unlike test set

validation, where the dataset is split into calibration and test sets, external validation requires a completely new set of samples with their responses. To validate the model, this new set (validation set) is projected onto the model and the model predictions are then compared with the real validation set responses. If the external validation results are consistent with the predictions, the model can be considered validated for future applications. A visual representation of the test set validation and external validation is reported in Figure 2.5.

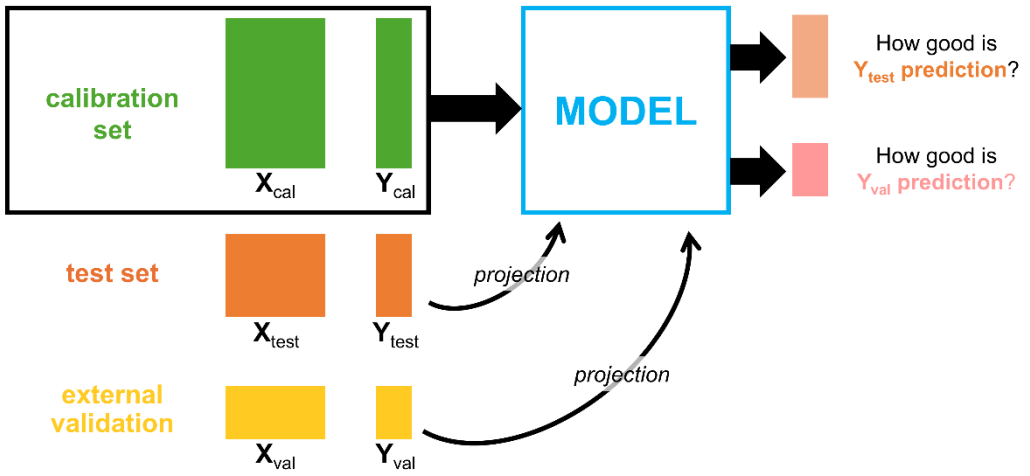

Figure 2.5. Visual representation of test set validation and external validation.

A common alternative to these approaches, performed especially when few samples are available, is the cross-validation [142–144]. This is an "internal validation" approach, since the model performances are evaluated directly from the original dataset without splitting it at the beginning, or without using external data. Cross-validation principles are not that different from those of test set-validation: the original dataset is divided into multiple subsets and all the subsets but one are used to build a model. The only subset not included in the modelling phase is then used to evaluate the model (i.e., this subset has the same function of a test set, for the specific sub-model). Following an iterative process, cross-validation continues until all the subsets are used as a test set to evaluate the models built on the other joined subsets. At the end of the iterative process, the model performance parameters, such as the root mean square error (RMSE) and the squared correlation coefficient (R^2), are calculated for each iteration and the average values provide an overall estimate of the model performances. So, basically, the cross-validation is nothing more than doing a test-set validation, but many times using all (sub-)parts of the original dataset. Figure 2.6 [145] reports a graphical representation of how cross-validation is structured and of how the evaluation of the optimal number of components to be selected is performed, while developing, for example, a PLS model.

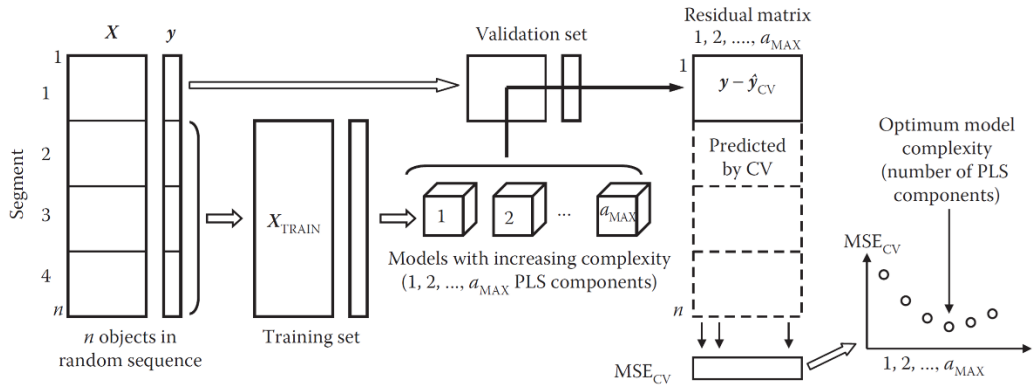

Figure 2.6. Graphical representation of cross-validation performed with four subsets to estimate mean error associated with each number of selected latent variables, from [145].

References | Chapter 2

[1] K. Héberger, Chemoinformatics—multivariate mathematical-statistical methods for data evaluation, Medical Applications of Mass Spectrometry (2007) 141–169. https://doi.org/10.1016/B978-044451980-1.50009-4.

[2] P. Oliveri, C. Malegori, M. Casale, Chemometrics: multivariate analysis of chemical data, Chemical Analysis of Food: Techniques and Applications, Second Edition (2020) 33–76. https://doi.org/10.1016/B978-0-12-813266-1.00002-4.

[3] S. Wold, Chemometrics; what do we mean with it, and what do we want from it?, Chemometrics and Intelligent Laboratory Systems 30 (1995) 109–115. https://doi.org/10.1016/0169-7439(95)00042-9.

[4] N. Islam, Chemometrics in Nondestructive Quality Evaluation, Nondestructive Quality Assessment Techniques for Fresh Fruits and Vegetables (2022) 331–355. https://doi.org/10.1007/978-981-19-5422-1_14.

[5] M. Efenberger-Szmechtyk, A. Nowak, D. Kregiel, Implementation of chemometrics in quality evaluation of food and beverages, Crit Rev Food Sci Nutr 58 (2018) 1747–1766. https://doi.org/10.1080/10408398.2016.1276883.

[6] M. Mohd Ali, N. Hashim, S.A. Aziz, O. Lasekan, Emerging non-destructive thermal imaging technique coupled with chemometrics on quality and safety inspection in food and agriculture, Trends Food Sci Technol 105 (2020) 176–185. https://doi.org/10.1016/j.tifs.2020.09.003.

[7] K. Kalinowska, M. Bystrzanowska, M. Tobiszewski, Chemometrics approaches to green analytical chemistry procedure development, Curr Opin Green Sustain Chem 30 (2021) 100498. https://doi.org/10.1016/J.COGSC.2021.100498.

[8] A. Astel, G. Głosińska, T. Sobczyński, L. Boszke, V. Simeonov, J. Siepak, Chemometrics in the assessment of the sustainable development rule implementation, Central European Journal of Chemistry 4 (2006) 543–564. https://doi.org/10.2478/S11532-006-0021-5/MACHINEREADABLECITATION/RIS.

[9] M. Bystrzanowska, M. Tobiszewski, Chemometrics for Selection, Prediction, and Classification of Sustainable Solutions for Green Chemistry—A Review, Symmetry 2020, Vol. 12, Page 2055 12 (2020) 2055. https://doi.org/10.3390/SYM12122055.

[10] A.S. Rathore, N. Bhushan, S. Hadpe, Chemometrics applications in biotech processes: A review, Biotechnol Prog 27 (2011) 307–315. https://doi.org/10.1002/BTPR.561.

[11] C.D. Kappatou, J. Odgers, S. García-Muñoz, R. Misener, An Optimization Approach Coupling Preprocessing with Model Regression for Enhanced Chemometrics, Ind Eng Chem Res 62 (2022) 6196–6213. https://doi.org/10.1021/ACS.IECR.2C04583/ASSET/IMAGES/LARGE/IE2C04583_0011.JPEG.

[12] S.M. Sarsam, Reinforcing the decision-making process in chemometrics: Feature selection and algorithm optimization, ACM International Conference Proceeding Series Part F147956 (2019) 11–16. https://doi.org/10.1145/3316615.3316644.

[13] W.A.S. Khalil, A. Goonetilleke, S. Kokot, S. Carroll, Use of chemometrics methods and multicriteria decision-making for site selection for sustainable on-site sewage effluent disposal, Anal Chim Acta 506 (2004) 41–56. https://doi.org/10.1016/J.ACA.2003.11.003.

[14] J.J. Roberts, D. Cozzolino, An Overview on the Application of Chemometrics in Food Science and Technology—An Approach to Quantitative Data Analysis, Food Anal Methods 9 (2016) 3258–3267. https://doi.org/10.1007/S12161-016-0574-7/TABLES/1.

[15] J.L. Aleixandre-Tudo, L. Castello-Cogollos, J.L. Aleixandre, R. Aleixandre-Benavent, Chemometrics in food science and technology: A bibliometric study, Chemometrics and Intelligent Laboratory Systems 222 (2022) 104514. https://doi.org/10.1016/J.CHEMOLAB.2022.104514.

[16] D. Granato, P. Putnik, D.B. Kovačević, J.S. Santos, V. Calado, R.S. Rocha, A.G. Da Cruz, B. Jarvis, O.Y. Rodionova, A. Pomerantsev, Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing, Compr Rev Food Sci Food Saf 17 (2018) 663–677. https://doi.org/10.1111/1541-4337.12341.

[17] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283–293. https://doi.org/10.1016/j.foodcont.2017.11.034.

[18] A. Aguzzoni, F. Scandellari, The geographical origin of fresh horticultural products: analytical methods to prevent food frauds, Italus Hortus (2017) 41–57. https://doi.org/10.26353/j.itahort/2017.1.4157.

[19] H.X. Mac, T.T. Pham, N.T.T. Ha, L.L.P. Nguyen, L. Baranyai, L. Friedrich, Current Techniques for Fruit Juice and Wine Adulterant Detection and Authentication, Beverages 2023, Vol. 9, Page 84 9 (2023) 84. https://doi.org/10.3390/BEVERAGES9040084.

[20] C.M. Andre, C. Soukoulis, Food Quality Assessed by Chemometrics, Foods 2020, Vol. 9, Page 897 9 (2020) 897. https://doi.org/10.3390/FOODS9070897.

[21] A. Inobeme, V. Nayak, T.J. Mathew, S. Okonkwo, L. Ekwoba, A.I. Ajai, E. Bernard, J. Inobeme, M. Mariam Agbugui, K.R. Singh, Chemometric approach in environmental pollution analysis: A critical review, J Environ Manage 309 (2022) 114653. https://doi.org/10.1016/J.JENVMAN.2022.114653.

[22] S. Mas, A. de Juan, R. Tauler, A.C. Olivieri, G.M. Escandar, Application of chemometric methods to environmental analysis of organic pollutants: A review, Talanta 80 (2010) 1052–1067. https://doi.org/10.1016/J.TALANTA.2009.09.044.

[23] B. Khakimov, G. Gürdeniz, S.B. Engelse, Trends in the application of chemometrics to foodomics studies, Acta Aliment 44 (2015) 4–31. https://doi.org/10.1556/AALIM.44.2015.1.1.

[24] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing?, TrAC - Trends in Analytical Chemistry 50 (2013) 96–106. https://doi.org/10.1016/j.trac.2013.04.015.

[25] A. Famili, W.M. Shen, R. Weber, E. Simoudis, Data Preprocessing and Intelligent Data Analysis, Intelligent Data Analysis 1 (1997) 3–23. https://doi.org/10.3233/IDA-1997-1102.

[26] M. Weber, M. Welling, P. Perona, Unsupervised Learning of Models for Recognition, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 1842 (2000) 18–32. https://doi.org/10.1007/3-540-45054-8_2.

[27] Z. Ghahramani, Unsupervised Learning, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3176 (2004) 72–112. https://doi.org/10.1007/978-3-540-28650-9_5.

[28] T. Jiang, J.L. Gradus, A.J. Rosellini, Supervised Machine Learning: A Brief Primer, Behav Ther 51 (2020) 675–687. https://doi.org/10.1016/J.BETH.2020.05.002.

[29] I.T. Jollife, J. Cadima, Principal component analysis: A review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2016). https://doi.org/10.1098/rsta.2015.0202.

[30] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[31] M. Sergent, D. Mathieu, R. Phan-Tan-Luu, G. Drava, Chemometrics and intelligent laboratory systems Tutorial Correct and incorrect use of multilinear regression, 1995.

[32] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, J Chemom 26 (2012) 42–51. https://doi.org/10.1002/CEM.1418.

[33] M. Cocchi, A. Biancolillo, F. Marini, Chemometric Methods for Classification and Feature Selection, Comprehensive Analytical Chemistry 82 (2018) 265–299. https://doi.org/10.1016/BS.COAC.2018.08.006.

[34] M. Kharbach, M. Alaoui Mansouri, M. Taabouz, H. Yu, Current Application of Advancing Spectroscopy Techniques in Food Analysis: Data Handling with Chemometric Approaches, Foods 2023, Vol. 12, Page 2753 12 (2023) 2753. https://doi.org/10.3390/FOODS12142753.

[35] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part II: modeling, validation, and applications, Anal Bioanal Chem 410 (2018) 6691–6704. https://doi.org/10.1007/s00216-018-1283-4.

[36] F. Westad, F. Marini, Validation of chemometric models - A tutorial, Anal Chim Acta 893 (2015) 14–24. https://doi.org/10.1016/j.aca.2015.06.056.

[37] X. Liu, J.W. Locasale, Metabolomics: A Primer, Trends Biochem Sci 42 (2017) 274–284. https://doi.org/10.1016/J.TIBS.2017.01.004/ASSET/C0A01118-37AD-4829-B5CC-57C8F1BBD471/MAIN.ASSETS/GR3.JPG.

[38] A. Zhang, H. Sun, P. Wang, Y. Han, X. Wang, Modern analytical techniques in metabolomics analysis, Analyst 137 (2011) 293–300. https://doi.org/10.1039/C1AN15605E.

[39] U. Roessner, J. Bowne, What is metabolomics all about?, Biotechniques 46 (2009) 363–365. https://doi.org/10.2144/000113133.

[40] B. Worley, R. Powers, Multivariate Analysis in Metabolomics, Curr Metabolomics 1 (2013) 92–107. https://doi.org/10.2174/2213235X11301010092.

[41] A. Valdés, G. Álvarez-Rivera, B. Socas-Rodríguez, M. Herrero, E. Ibáñez, A. Cifuentes, Foodomics: Analytical Opportunities and Challenges, Anal Chem 94 (2022) 366–381. https://doi.org/10.1021/ACS.ANALCHEM.1C04678/ASSET/IMAGES/LARGE/AC1C04678_0005.JPEG.

[42] P. Balkir, K. Kemahlioglu, U. Yucel, Foodomics: A new approach in food quality and safety, Trends Food Sci Technol 108 (2021) 49–57. https://doi.org/10.1016/J.TIFS.2020.11.028.

[43] L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, S. Wold, Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm), Anal Bioanal Chem 380 (2004) 419–429. https://doi.org/10.1007/S00216-004-2783-Y/FIGURES/10.

[44] F. Li, J. Zhang, Y. Wang, Vibrational Spectroscopy Combined with Chemometrics in Authentication of Functional Foods, Crit Rev Anal Chem 54 (2024) 333–354. https://doi.org/10.1080/10408347.2022.2073433.

[45] R. López-Ruiz, R. Romero-González, A. Garrido Frenich, Metabolomics approaches for the determination of multiple contaminants in food, Curr Opin Food Sci 28 (2019) 49–57. https://doi.org/10.1016/J.COFS.2019.08.006.

[46] E.K. Matich, N.G. Chavez Soria, D.S. Aga, G.E. Atilla-Gokcumen, Applications of metabolomics in assessing ecological effects of emerging contaminants and pollutants on plants, J Hazard Mater 373 (2019) 527–535. https://doi.org/10.1016/J.JHAZMAT.2019.02.084.

[47] E. Cubero-Leon, R. Peñalver, A. Maquet, Review on metabolomics for food authentication, Food Research International 60 (2014) 95–107. https://doi.org/10.1016/J.FOODRES.2013.11.041.

[48] J. Selamat, N.A.A. Rozani, S. Murugesu, Application of the Metabolomics Approach in Food Authentication, Molecules 2021, Vol. 26, Page 7565 26 (2021) 7565. https://doi.org/10.3390/MOLECULES26247565.

[49] H.R. Tan, W. Zhou, Metabolomics for tea authentication and fraud detection: Recent applications and future directions, Trends Food Sci Technol 149 (2024) 104558. https://doi.org/10.1016/J.TIFS.2024.104558.

[50] N. Mialon, B. Roig, E. Capodanno, A. Cadiere, Untargeted metabolomic approaches in food authenticity: A review that showcases biomarkers, Food Chem 398 (2023) 133856. https://doi.org/10.1016/J.FOODCHEM.2022.133856.

[51] Y. Gao, L. Hou, J. Gao, D. Li, Z. Tian, B. Fan, F. Wang, S. Li, Metabolomics Approaches for the Comprehensive Evaluation of Fermented Foods: A Review, Foods 2021, Vol. 10, Page 2294 10 (2021) 2294. https://doi.org/10.3390/FOODS10102294.

[52] S. Li, Y. Tian, P. Jiang, Y. Lin, X. Liu, H. Yang, Recent advances in the application of metabolomics for food safety control and food quality analyses, Crit Rev Food Sci Nutr 61 (2021) 1448–1469. https://doi.org/10.1080/10408398.2020.1761287.

[53] A.H. Emwas, R. Roy, R.T. McKay, L. Tenori, E. Saccenti, G.A. Nagana Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, D.S. Wishart, Nmr spectroscopy for metabolomics research, Metabolites 9 (2019). https://doi.org/10.3390/metabo9070123.

[54] D.S. Wishart, Quantitative metabolomics using NMR, TrAC Trends in Analytical Chemistry 27 (2008) 228–237. https://doi.org/10.1016/J.TRAC.2007.12.001.

[55] N. V. Reo, NMR-BASED METABOLOMICS, Drug Chem Toxicol 25 (2002) 375–382. https://doi.org/10.1081/DCT-120014789.

[56] J.L. Markley, R. Brüschweiler, A.S. Edison, H.R. Eghbalnia, R. Powers, D. Raftery, D.S. Wishart, The future of NMR-based metabolomics, Curr Opin Biotechnol 43 (2017) 34–40. https://doi.org/10.1016/J.COPBIO.2016.08.001.

[57] H.G. Gika, G.A. Theodoridis, R.S. Plumb, I.D. Wilson, Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics, J Pharm Biomed Anal 87 (2014) 12–25. https://doi.org/10.1016/J.JPBA.2013.06.032.

[58] M.M. Koek, R.H. Jellema, J. van der Greef, A.C. Tas, T. Hankemeier, Quantitative metabolomics based on gas chromatography mass spectrometry: Status and perspectives, Metabolomics 7 (2011) 307–328. https://doi.org/10.1007/S11306-010-0254-3/FIGURES/1.

[59] S.T. Ovbude, S. Sharmeen, I. Kyei, H. Olupathage, J. Jones, R.J. Bell, R. Powers, D.S. Hage, Applications of chromatographic methods in metabolomics: A review, Journal of Chromatography B 1239 (2024) 124124. https://doi.org/10.1016/J.JCHROMB.2024.124124.

[60] K. Dettmer, P.A. Aronov, B.D. Hammock, Mass spectrometry-based metabolomics, Mass Spectrom Rev 26 (2007) 51–78. https://doi.org/10.1002/MAS.20108.

[61] G.A. Nagana Gowda, D. Djukovic, Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges, Methods in Molecular Biology 1198 (2014) 3–12. https://doi.org/10.1007/978-1-4939-1258-2_1.

[62] Z. Lei, D. V. Huhman, L.W. Sumner, Mass spectrometry strategies in metabolomics, Journal of Biological Chemistry 286 (2011) 25435–25442. https://doi.org/10.1074/JBC.R111.238691/ASSET/280A3D88-EF42-40F7-8D81-4EBB058636C6/MAIN.ASSETS/GR4.JPG.

[63] J. Liu, H. Zhao, Z. Yin, H. Dong, X. Chu, X. Meng, Y. Li, X. Ding, Application and prospect of metabolomics-related technologies in food inspection, Food Research International 171 (2023) 113071. https://doi.org/10.1016/J.FOODRES.2023.113071.

[64] W. Wu, L. Zhang, X. Zheng, Q. Huang, M.A. Farag, R. Zhu, C. Zhao, Emerging applications of metabolomics in food science and future trends, Food Chem X 16 (2022) 100500. https://doi.org/10.1016/J.FOCHX.2022.100500.

[65] A. Tomassini, G. Capuani, M. Delfini, A. Micchelli, NMR-Based Metabolomics in Food Quality Control, Data Handling in Science and Technology 28 (2013) 411–447. https://doi.org/10.1016/B978-0-444-59528-7.00011-9.

[66] A. Sussulini, ed., Metabolomics: From Fundamentals to Clinical Applications, 965 (2017). https://doi.org/10.1007/978-3-319-47656-8.

[67] A. Trimigno, F.C. Marincola, N. Dellarosa, G. Picone, L. Laghi, Definition of food quality by NMR-based foodomics, Curr Opin Food Sci 4 (2015) 99–104. https://doi.org/10.1016/J.COFS.2015.06.008.

[68] A.P. Sobolev, C. Ingallina, M. Spano, G. Di Matteo, L. Mannina, NMR-Based Approaches in the Study of Foods, Molecules 2022, Vol. 27, Page 7906 27 (2022) 7906. https://doi.org/10.3390/MOLECULES27227906.

[69] A.M. Gil, I.F. Duarte, M. Godejohann, U. Braumann, M. Maraschin, M. Spraul, Characterization of the aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection, Anal Chim Acta 488 (2003) 35–51. https://doi.org/10.1016/S0003-2670(03)00579-8.

[70] I.F. Duarte, M. Godejohann, U. Braumann, M. Spraul, A.M. Gil, Application of NMR spectroscopy and LC-NMR/MS to the identification of carbohydrates in beer, ACS PublicationsIF Duarte, M Godejohann, U Braumann, M Spraul, AM GilJournal of Agricultural and Food Chemistry, 2003•ACS Publications 51 (2003) 4847–4852. https://doi.org/10.1021/jf030097j.

[71] C. Almeida, I.F. Duarte, A. Barros, J. Rodrigues, M. Spraul, A.M. Gil, Composition of beer by 1H NMR spectroscopy: Effects of brewing site and date of production, J Agric Food Chem 54 (2006) 700–706. https://doi.org/10.1021/JF0526947/ASSET/IMAGES/LARGE/JF0526947F00006.JPEG.

[72] L.I. Nord, P. Vaag, J. Duus, Quantification of organic and amino acids in beer by 1H NMR spectroscopy, Anal Chem 76 (2004) 4790–4798. https://doi.org/10.1021/AC0496852.

[73] J.E.A. Rodrigues, G.L. Erny, A.S. Barros, V.I. Esteves, T. Brandão, A.A. Ferreira, E. Cabrita, A.M. Gil, Quantification of organic acids in beer by nuclear magnetic resonance (NMR)-based methods, Anal Chim Acta 674 (2010) 166–175. https://doi.org/10.1016/J.ACA.2010.06.029.

[74] A.P. Sobolev, F. Thomas, J. Donarski, C. Ingallina, S. Circi, F. Cesare Marincola, D. Capitani, L. Mannina, Use of NMR applications to tackle future food fraud issues, Trends Food Sci Technol 91 (2019) 347–353. https://doi.org/10.1016/J.TIFS.2019.07.035.

[75] L. Mannina, F. Marini, R. Antiochia, S. Cesa, A. Magrì, D. Capitani, A.P. Sobolev, Tracing the origin of beer samples by NMR and chemometrics: Trappist beers as a case study, Electrophoresis 37 (2016) 2710–2719. https://doi.org/10.1002/ELPS.201600082.

[76] T. Kuballa, T.S. Brunner, T. Thongpanchang, S.G. Walch, D.W. Lachenmeier, Application of NMR for authentication of honey, beer and spices, Curr Opin Food Sci 19 (2018) 57–62. https://doi.org/10.1016/J.COFS.2018.01.007.

[77] C.A. Hughey, C.M. McMinn, J. Phung, Beeromics: from quality control to identification of differentially expressed compounds in beer, Metabolomics 12 (2016) 1–13. https://doi.org/10.1007/S11306-015-0885-5.

[78] I.F. Duarte, A. Barros, C. Almeida, M. Spraul, A.M. Gil, Multivariate analysis of NMR and FTIR data as a potential tool for the quality control of beer, J Agric Food Chem 52 (2004) 1031–1038. https://doi.org/10.1021/JF030659Z.

[79] D.W. Lachenmeier, W. Frank, E. Humpfer, H. Schäfer, S. Keller, M. Mörtter, M. Spraul, Quality control of beer using high-resolution nuclear magnetic resonance spectroscopy and multivariate analysis, European Food Research and Technology 220 (2005) 215–221. https://doi.org/10.1007/S00217-004-1070-7/TABLES/3.

[80] S. Verboven, M. Hubert, P. Goos, Robust preprocessing and model selection for spectral data, J Chemom 26 (2012) 282–289. https://doi.org/10.1002/CEM.2446.

[81] I. Karaman, Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis, Adv Exp Med Biol 965 (2017) 145–161. https://doi.org/10.1007/978-3-319-47656-8_6.

[82] C. Fan, M. Chen, X. Wang, J. Wang, B. Huang, A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data, Front Energy Res 9 (2021) 652801. https://doi.org/10.3389/FENRG.2021.652801/BIBTEX.

[83] F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, Chemometrics and Intelligent Laboratory Systems 82 (2006) 59–65. https://doi.org/10.1016/J.CHEMOLAB.2005.08.009.

[84] K.H. Liland, B.-H. Mevik, T. Almøy, Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra, Applied Spectroscopy, Vol. 64, Issue 9, Pp. 1007-1016 64 (2010) 1007–1016. https://opg.optica.org/abstract.cfm?uri=as-64-9-1007 (accessed February 14, 2025).

[85] L. Vega-Montoto, C.D. Brown, P.D. Wentzell, Derivative Preprocessing and Optimal Corrections for Baseline Drift in Multivariate Calibration, Applied Spectroscopy, Vol. 54, Issue 7, Pp. 1055-1068 54 (2000) 1055–1068. https://opg.optica.org/abstract.cfm?uri=as-54-7-1055 (accessed February 14, 2025).

[86] A.G. Ryder, M.N. Leger, Comparison of Derivative Preprocessing and Automated Polynomial Baseline Correction Method for Classification and Quantification of Narcotics in Solid Mixtures, Applied Spectroscopy, Vol. 60, Issue 2, Pp. 182-193 60 (2006) 182–193. https://opg.optica.org/abstract.cfm?uri=as-60-2-182 (accessed February 14, 2025).

[87] S. Gopal, K. Patro, K. Kumar Sahu, Normalization: A Preprocessing Stage, IARJSET (2015) 20–22. https://doi.org/10.17148/iarjset.2015.2305.

[88] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal Chem 36 (1964) 1627–1639. https://doi.org/10.1021/AC60214A047/ASSET/AC60214A047.FP.PNG_V03.

[89] F. Savorani, G. Tomasi, S. Engelsen, Alignment of 1D NMR Data using the iCoshift Tool: A Tutorial, in: 2013: pp. 14–24. https://doi.org/10.1039/9781849737531-00014.

[90] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, Journal of Magnetic Resonance 202 (2010) 190–202. https://doi.org/10.1016/j.jmr.2009.11.012.

[91] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, Data Handling in Science and Technology 28 (2013) 55–126. https://doi.org/10.1016/B978-0-444-59528-7.00003-X.

[92] S.C. Rutan, A. de Juan, R. Tauler, Introduction to Multivariate Curve Resolution, Comprehensive Chemometrics 2 (2009) 249–259. https://doi.org/10.1016/B978-044452701-1.00046-6.

[93] R. Bro, A.S.-A. methods, undefined 2014, Principal component analysis, Pubs.Rsc.OrgR Bro, AK SmildeAnalytical Methods, 2014•pubs.Rsc.Org (2014). https://doi.org/10.1039/c3ay41907j.

[94] M. Ahmed, R. Seraj, S.M.S. Islam, The k-means Algorithm: A Comprehensive Survey and Performance Evaluation, Electronics 2020, Vol. 9, Page 1295 9 (2020) 1295. https://doi.org/10.3390/ELECTRONICS9081295.

[95] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, Wiley Interdiscip Rev Data Min Knowl Discov 2 (2012) 86–97. https://doi.org/10.1002/WIDM.53.

[96] M. Ringnér, What is principal component analysis?, Nature Biotechnology 2008 26:3 26 (2008) 303–304. https://doi.org/10.1038/nbt0308-303.

[97] T. Kurita, Principal Component Analysis (PCA), Computer Vision (2021) 1013–1016. https://doi.org/10.1007/978-3-030-63416-2_649.

[98] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, n.d.

[99] S. Karamizadeh, S.M. Abdullah, A.A. Manaf, M. Zamani, A. Hooman, S. Karamizadeh, S.M. Abdullah, A.A. Manaf, M. Zamani, A. Hooman, An Overview of Principal Component Analysis, Journal of Signal and Information Processing 4 (2013) 173–175. https://doi.org/10.4236/JSIP.2013.43B031.

[100] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[101] M. Sergent, D. Mathieu, R. Phan-Tan-Luu, G. Drava, Chemometrics and intelligent laboratory systems Tutorial Correct and incorrect use of multilinear regression, 1995.

[102] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, Analytical Methods 5 (2013) 3790–3798. https://doi.org/10.1039/c3ay40582f.

[103] O. Masetti, A. Sorbo, L. Nisini, Nmr tracing of food geographical origin: The impact of seasonality, cultivar and production year on data analysis, Separations 8 (2021). https://doi.org/10.3390/separations8120230.

[104] S. Ghisoni, L. Lucini, F. Angilletta, G. Rocchetti, D. Farinelli, S. Tombesi, M. Trevisan, Discrimination of extra-virgin-olive oils from different cultivars and geographical origins by untargeted metabolomics, Food Research International 121 (2019) 746–753. https://doi.org/10.1016/j.foodres.2018.12.052.

[105] R. Bachmann, S. Klockmann, J. Haerdter, M. Fischer, T. Hackl, 1H NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts, J Agric Food Chem 66 (2018) 11873–11879. https://doi.org/10.1021/acs.jafc.8b03724.

[106] X. Zhang, F. Liu, Y. He, X. Li, Application of Hyperspectral Imaging and Chemometric Calibrations for Variety Discrimination of Maize Seeds, Sensors 2012, Vol. 12, Pages 17234-17246 12 (2012) 17234–17246. https://doi.org/10.3390/S121217234.

[107] I. Feher, D.A. Magdas, A. Dehelean, C. Sârbu, Characterization and classification of wines according to geographical origin, vintage and specific variety based on elemental content: a new chemometric approach, J Food Sci Technol 56 (2019) 5225–5233. https://doi.org/10.1007/S13197-019-03991-4/TABLES/1.

[108] K. Kucharska-Ambrożej, J. Karpinska, The application of spectroscopic techniques in combination with chemometrics for detection adulteration of some herbs and spices, Microchemical Journal 153 (2020) 104278. https://doi.org/10.1016/J.MICROC.2019.104278.

[109] S. Grassi, M. Tarapoulouzi, A. D'Alessandro, S. Agriopoulou, L. Strani, T. Varzakas, How Chemometrics Can Fight Milk Adulteration, Foods 2023, Vol. 12, Page 139 12 (2022) 139. https://doi.org/10.3390/FOODS12010139.

[110] K.J. Siebert, Using chemometrics to classify samples and detect misrepresentation, ACS Symposium Series 1081 (2011) 39–65. https://doi.org/10.1021/BK-2011-1081.CH004.

[111] W. Fortunato de Carvalho Rocha, M.M. Schantz, D.A. Sheen, P.M. Chu, K.A. Lippa, Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data, Fuel 197 (2017) 248–258. https://doi.org/10.1016/J.FUEL.2017.02.025.

[112] P. Oliveri, Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial, Anal Chim Acta 982 (2017) 9–19. https://doi.org/10.1016/j.aca.2017.05.013.
[113] M.D. Sorochan Armstrong, A.P. de la Mata, J.J. Harynuk, Review of Variable Selection Methods for Discriminant-Type Problems in Chemometrics, Frontiers in Analytical Science 2 (2022) 867938. https://doi.org/10.3389/FRANS.2022.867938.
[114] B. Stumpe, T. Engel, B. Steinweg, B. Marschner, Application of PCA and SIMCA statistical analysis of FT-IR spectra for the classification and identification of different slag types with environmental origin, Environ Sci Technol 46 (2012) 3964–3972. https://doi.org/10.1021/ES204187R/SUPPL_FILE/ES204187R_SI_001.PDF.
[115] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemometrics and Intelligent Laboratory Systems 174 (2018) 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004.
[116] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J Chemom 29 (2015) 528–536. https://doi.org/10.1002/CEM.2736.
[117] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal Chim Acta 891 (2015) 1–14. https://doi.org/10.1016/J.ACA.2015.04.042.
[118] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines, Chemometrics and Intelligent Laboratory Systems 137 (2014) 181–189. https://doi.org/10.1016/J.CHEMOLAB.2014.06.012.
[119] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, Data Fusion for Food Authentication. Combining near and Mid Infrared to Trace the Origin of Extra Virgin Olive Oils, Http://Dx.Doi.Org/10.1255/Nirn.1355 24 (2013) 12–15. https://doi.org/10.1255/NIRN.1355.
[120] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal Chim Acta 820 (2014) 23–31. https://doi.org/10.1016/J.ACA.2014.02.024.
[121] M. Ottavian, L. Fasolato, L. Serva, P. Facco, M. Barolo, Data Fusion for Food Authentication: Fresh/Frozen-Thawed Discrimination in West African Goatfish (Pseudupeneus prayensis) Fillets, Food Bioproc Tech 7 (2014) 1025–1036. https://doi.org/10.1007/s11947-013-1157-x.
[122] S. Schwolow, N. Gerhardt, S. Rohn, P. Weller, Data fusion of GC-IMS data and FT-MIR spectra for the authentication of olive oils and honeys—is it worth to go the extra mile?, Anal Bioanal Chem 411 (2019) 6005–6019. https://doi.org/10.1007/s00216-019-01978-w.
[123] C. Robert, W. Jessep, J.J. Sutton, T.M. Hicks, M. Loeffen, M. Farouk, J.F. Ward, W.E. Bain, C.R. Craigie, S.J. Fraser-Miller, K.C. Gordon, Evaluating low- mid- and high-level fusion strategies for combining Raman and infrared spectroscopy for quality assessment of red meat, Food Chem 361 (2021). https://doi.org/10.1016/j.foodchem.2021.130154.
[124] Y. Hong, N. Birse, B. Quinn, Y. Li, W. Jia, P. McCarron, D. Wu, G.R. da Silva, L. Vanhaecke, S. van Ruth, C.T. Elliott, Data fusion and multivariate analysis for food authenticity analysis, Nat Commun 14 (2023). https://doi.org/10.1038/s41467-023-38382-z.
[125] S.C. Rutan, A. de Juan, R. Tauler, Introduction to Multivariate Curve Resolution, Comprehensive Chemometrics 2 (2009) 249–259. https://doi.org/10.1016/B978-044452701-1.00046-6.
[126] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Crit Rev Anal Chem 36 (2006) 163–176. https://doi.org/10.1080/10408340600970005.
[127] B. Khakimov, N. Mobaraki, A. Trimigno, V. Aru, S.B. Engelsen, Signature Mapping (SigMa): An efficient approach for processing complex human urine 1H NMR metabolomics data, Anal Chim Acta 1108 (2020) 142–151. https://doi.org/10.1016/J.ACA.2020.02.025.
[128] R. Leardi, Experimental design in chemistry: A tutorial, Anal Chim Acta 652 (2009) 161–172. https://doi.org/10.1016/j.aca.2009.06.015.
[129] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and

data analysis tools, Anal Bioanal Chem 409 (2017) 5891–5899. https://doi.org/10.1007/s00216-017-0517-1.

[130] Ž.R. Lazić, Design of Experiments in Chemical Engineering, Design of Experiments in Chemical Engineering (2004). https://doi.org/10.1002/3527604162.

[131] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, Anal Bioanal Chem 409 (2017) 5891–5899. https://doi.org/10.1007/S00216-017-0517-1/FIGURES/1.

[132] R. Leardi, Experimental design in chemistry: A tutorial, Anal Chim Acta 652 (2009) 161–172. https://doi.org/10.1016/J.ACA.2009.06.015.

[133] H. Ebrahimi-Najafabadi, R. Leardi, M. Jalali-Heravi, Experimental Design in Analytical Chemistry—Part I: Theory, J AOAC Int 97 (2014) 3–11. https://doi.org/10.5740/JAOACINT.SGEEBRAHIMI1.

[134] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048. https://doi.org/10.1093/BIOINFORMATICS/BTI476.

[135] M. Sergent, D. Mathieu, R. Phan-Tan-Luu, G. Drava, Correct and incorrect use of multilinear regression, Chemometrics and Intelligent Laboratory Systems 27 (1995) 153–162. https://doi.org/10.1016/0169-7439(95)80020-A.

[136] S.A. Weissman, N.G. Anderson, Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications, Org Process Res Dev 19 (2015) 1605–1633. https://doi.org/10.1021/OP500169M/ASSET/IMAGES/LARGE/OP-2014-00169M_0027.JPEG.

[137] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048. https://doi.org/10.1093/BIOINFORMATICS/BTI476.

[138] C. Bertinetto, J. Engel, J. Jansen, ANOVA simultaneous component analysis: A tutorial review, Anal Chim Acta X 6 (2020). https://doi.org/10.1016/j.acax.2020.100061.

[139] C.L.M. Morais, M.C.D. Santos, K.M.G. Lima, F.L. Martin, Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach, Bioinformatics 35 (2019) 5257–5263. https://doi.org/10.1093/BIOINFORMATICS/BTZ421.

[140] N.K. Nguyen, A.J. Miller, A review of some exchange algorithms for constructing discrete D-optimal designs, Comput Stat Data Anal 14 (1992) 489–498. https://doi.org/10.1016/0167-9473(92)90064-M.

[141] P.F. de Aguiar, B. Bourguignon, M.S. Khots, D.L. Massart, R. Phan-Than-Luu, D-optimal designs, Chemometrics and Intelligent Laboratory Systems 30 (1995) 199–210. https://doi.org/10.1016/0169-7439(94)00076-X.

[142] R. Bro, K. Kjeldahl, A.K. Smilde, H.A.L. Kiers, Cross-validation of component models: A critical look at current methods, Anal Bioanal Chem 390 (2008) 1241–1251. https://doi.org/10.1007/S00216-007-1790-1/FIGURES/8.

[143] M.W. Browne, Cross-Validation Methods, J Math Psychol 44 (2000) 108–132. https://doi.org/10.1006/JMPS.1999.1279.

[144] M. Stone, Cross-validation:a review 2, Statistics: A Journal of Theoretical and Applied Statistics 9 (1978) 127–139. https://doi.org/10.1080/02331887808801414.

[145] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, Introduction to Multivariate Statistical Analysis in Chemometrics (2016). https://doi.org/10.1201/9781420059496.

Chapter 3

Case Study #1: Optimization of industrial food treatments

3.1 Project overview

In the past years, the interest towards bioeconomy concepts has been considerably growing. In particular, the development of sustainable and renewable bio-based technologies for food production is becoming increasingly important. One of the most interesting applications of bioeconomy in the "food" area is the use of enzymes for the transformation of food ingredients, waste or by-products, to improve food safety and optimize the overall food treatment process [1].

In this perspective, the present study is focused on the optimization of the parameters used for an industrial treatment performed using a protease enzyme on red lentil flour. In particular, red lentil flour is recognized as a valuable "functional food" in the field of food supplements due to its high protein content and healthy properties [2,3]. After an initial extraction process performed at a fixed pH and temperature, the red lentil flour samples were hydrolysed using a protease enzyme changing the most critical process control parameters in accordance with a simple experimental design: addition or not of the enzyme, two different stirring rates and different treatment times. A total of 32 different samples were obtained and analysed using ^{1}H-NMR spectroscopy. Samples were collected at specific timepoints of the process and immediately frozen to avoid chemical changes before analysis.

All the collected spectra were then imported into MATLAB environment for multivariate analysis. Principal component analysis (PCA) [4] was used to explore the data, with the specific aim of looking for time-dependent trends and relations among the different factors' levels. The aim of this project was to evaluate the ability of PCA with respect to process monitoring, possibly leading to the optimization of agrifood industrial treatments just by using one of the simplest tools in multivariate statistical analysis. Its efficiency in the analysis of complex datasets helps improving decision-making and overall efficiency. Recently, PCA has been utilized to develop indicators for monitoring biological processes like anaerobic digestions [5] or drying treatments [6]. In this project, a similar approach was used both to investigate and to optimize an extraction and hydrolyzation process.

3.2 Experimental design and laboratory analysis

In this project, the laboratory analysis procedure started from one single specimen of red lentils flour that was treated with different experimental conditions according to an experimental design. This design was developed to monitor a combination of different parameters with the aim of evaluating the effect of each parameter tested on the overall treatment process [7–10]. Three different conditions were evaluated: the presence of an enzyme for hydrolysis, the stirring rate of the centrifugation process, and the effect of the time on the treated samples.

To evaluate all these conditions (summarized in table 3.1), the starting amount of red lentil flour was divided into two batches: the first one was treated with a protein extraction process, while the second one was treated with a combination of protein extraction and hydrolyzation with a specific protease enzyme (covered by a non-disclosure agreement). In both cases, the processes were conducted in basic conditions controlled by additions of $CaOH_2$, also to keep pH constant at 8.00. Both batches were stirred at two fixed RPM (revolution per minute) values to evaluate also the effect of the stirring rate. The lower rate was set at 60 RPM, while the more energic one was set at 120 RPM. The batch that was treated only with the extraction procedure was maintained at room temperature, while the hydrolysed batch was treated at 60 °C with the enzyme content fixed at 0.2 % with respect to the water content. The last monitored parameter was the treatment time, that was followed applying a sampling procedure according to a designed timetable. For the batch in which the lentil flour was only extracted, 6 sequential sampling times were decided and an aliquot of about 20 ml of liquid was withdrawn at 0, 1, 15, 30, 45 and 60 minutes from both the batches stirred at 60 and 120 RPM. For the batches in which also the hydrolysis was performed, stirred at 60 and 120 RPM, 10 aliquots of about 20 ml per batch were withdrawn at 0, 1, 15, 30, 45, 60, 120, 180, 240 and 300 minutes. The 32 acquired samples were quenched by quickly lowering the pH to 5.0 using an HCl solution (0.1 M) and immediately frozen in single tubes at -20 °C to ensure proper preservation.

Table 3.1. Summary of parameters evaluated (factors) and their values (levels)

Factors	Enzyme	Stirring rate [RPM]	Times [min]
Levels	Hydrolase (-1)	60 (-1), 120 (+1)	0, 1, 15, 30, 45, 60
	No enzyme (+1)	60 (-1), 120 (+1)	0, 1, 15, 30, 45, 60, 120, 180, 240, 300

3.3 NMR data acquisition

For the spectroscopic analysis, the samples were thawed at room temperature and 600 µl per samples were transferred into an NMR tube with an addition of D_2O deuterated solvent (10 %) and of 3-(Trimethylsilyl)propionic acid (TSP) (5 mM) used as internal standard. Two replicates were prepared per each sample.

The ^{1}H-NMR analysis for the samples collected in this study was performed on a Jeol ECZR 600 spectrometer (JEOL Ltd., Akishima, Tokyo, Japan) operating at 600.17 MHz for protons. The spectra were collected at a fixed temperature of 298 K by acquiring 32768 points and performing 128 scans for each sample, using a 3 s relaxation delay. A solvent suppression procedure ("Watergate" Jeol nomenclature) was applied to remove the water signal [11].

3.4 Data processing

The spectra were automatically baseline- and phase-corrected using the DELTA processing tool offered by JEOL Ltd. The raw ^{1}H-NMR spectra were then imported and processed under the MATLAB environment (2021b, Mathworks, Natick, MA, USA). For all spectra, the ppm scale was referenced to the TSP signal at 0.00 ppm. The spectral range was corrected to include only signals between -0.1 ppm and 10.0 ppm to remove useless and noisy areas. To avoid interferences, water signal residue (from 4.55 and 4.85 ppm) was also cut out from the spectra.

To increase the comparability among spectra composing the spectral dataset, the "*i*coshift" [12,13] tool was applied to horizontally align the most important signals located inside specific intervals, accurately manually defined.

After signal's alignment, the spectra were mean centered and then normalized with respect to the total sum of the signal areas [14]. After data pre-processing, an exploratory multivariate data analysis was performed, by means of PCA, using the PLS_Toolbox software package (version 8.9, Eigenvector Research Inc., Manson, WA, USA) working under the MATLAB environment.

3.5 Data exploration and process optimization by PCA

3.5.1 PCA for exploring clusters, trends and outliers

After the data processing step, the samples were organized to be explored using PCA [4]. A first PCA was performed considering the spectra obtained for all the samples, for this model only two principal components (PC) were selected since the total explained variance was above 90 % (85.48 % for PC1 and 4.61 % for PC2, respectively). Looking at the scores plot obtained with respect to PC1 and PC2

(Figure 3.1), two large clusters and a third less populated one can be observed. The less populated cluster contains the samples collected at time zero, no matter the presence of the enzyme for hydrolysis. The other two clusters, instead, are respectively composed by the hydrolysed samples and by the samples extracted without enzyme. Interestingly, one minute of treatment is enough to generate significative differences among the samples.

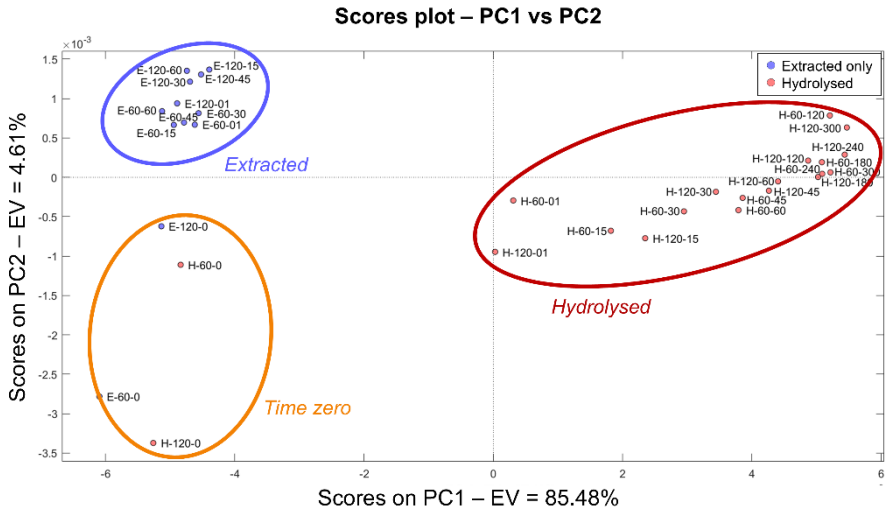

Figure 3.1. Scores plot of PC1 vs PC2 obtained from a PCA performed considering the spectra obtained from all the samples. Three main clusters are highlighted corresponding to extracted samples (blue), hydrolysed samples (red) and zero-time samples (orange).

By focusing only on PC1 and plotting its scores values against the samples' number (Figure 3.2A) and the loadings values against the variables' number (Figure 3.2B), some interesting observations can be made. The two clusters of hydrolysed and extracted only samples are still visible, and in addition, by exploring the loadings, the extracted only samples seem to be characterized by more intense signals in the spectral area between 3.0 and 5.0 ppm, meaning that the enzyme activity probably leads to a reduction of the metabolites that generate signals in that ppm range. A possible explanation of this behaviour is that the proteases activity, acting on peptide bonds, causes changes in the sample protein content. These modifications are particularly visible in the range around 4.0 ppm, which can correspond to protons on amino acids' side chains and alpha-protons adjacent to peptide bonds. Another interesting outcome from the PC1 scores plot is that, among the hydrolysed samples, four treatments appear different from the other hydrolysed ones: these samples are the two withdrawn at time zero and the one withdrawn after one minute. In particular, the two samples collected before the treatment (T=0) have the same scores values of the extracted only samples. This is not unexpected since at time zero the enzyme activity has not started yet, making these samples similar to the one without the enzyme. The other two samples instead, were withdrawn after one minute (T=1), so the hydrolysis process is already begun, but the differences

generated inside those samples after only one minute are not comparable with the effect of a longer enzymatic activity. This behaviour makes these two treatments (T=1) different from treatments withdrawn at T=0, but also from samples treated for longer times (T>1).

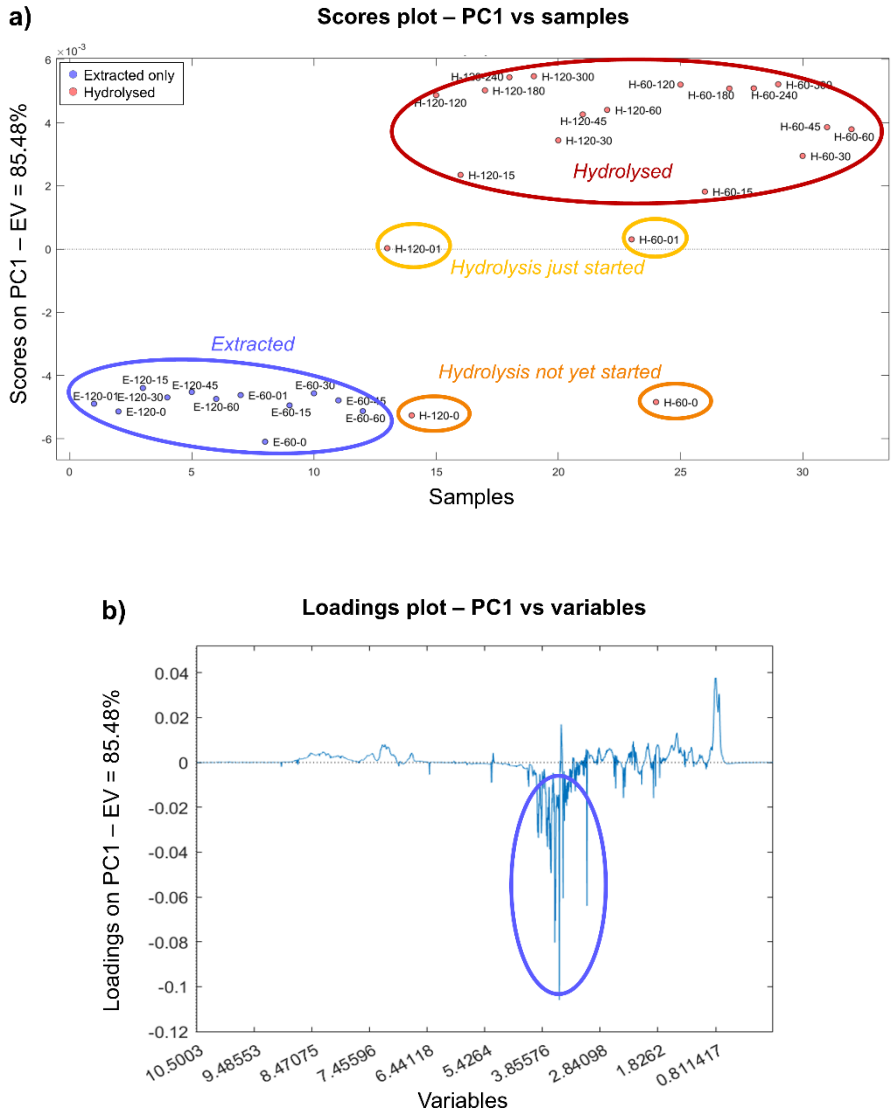

Figure 3.2. Scores plot of PC1 vs samples (A) and loadings plot of PC1 vs variables (B) obtained from a PCA performed considering the spectra obtained from all the samples. In the scores plot groupings related to the treatment are associated with different colours. In the loadings plot an interesting spectral area is highlighted.

Exploring also PC2 in the same way and looking at the corresponding scores plot (Figure 3.3A), the time effect previously observed among the hydrolysed samples is still visible. In addition, two samples appear very far from the others and further inspection need to be done to assess if they can be considered as outliers. The residuals plot (Figure 3.3B), obtained by plotting the Q residuals and Hotelling T^2 values, confirmed the suspect about those samples revealing particularly high T^2 values and slightly high Q residual values. The exploration of the corresponding ^{1}H-NMR spectra (Figure 3.4) highlighted a few small differences with respect to

the mean spectra of all the samples, revealing a possible instrumental effect during the acquisition or a possible human effect during the sample preparation phase.

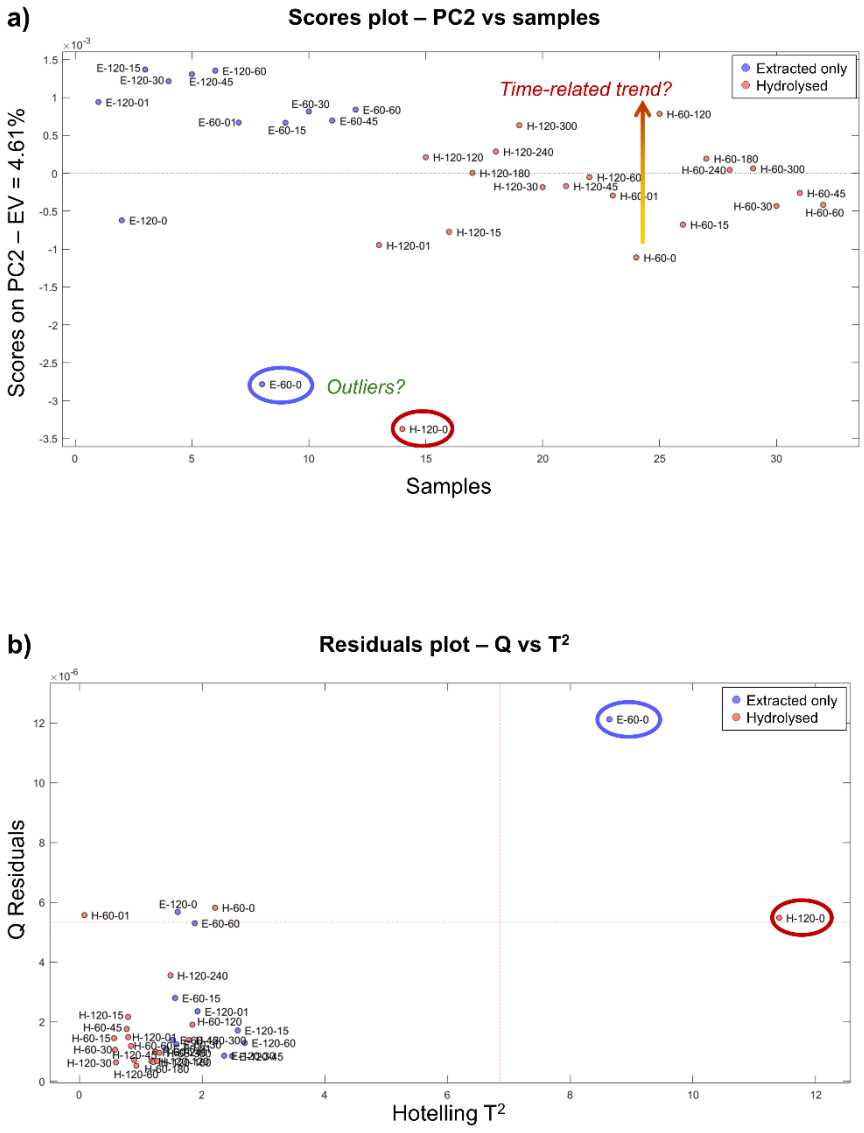

Figure 3.3. Scores plot of PC2 vs samples (A) obtained from a PCA performed considering the spectra obtained from all the samples and residuals plot (B) obtained by plotting Q residuals vs Hotelling T^2. In the scores plot a time-related trend among the hydrolysed samples is highlighted as well as two possible outliers. In the residual plot, the two possible outliers are highlighted.

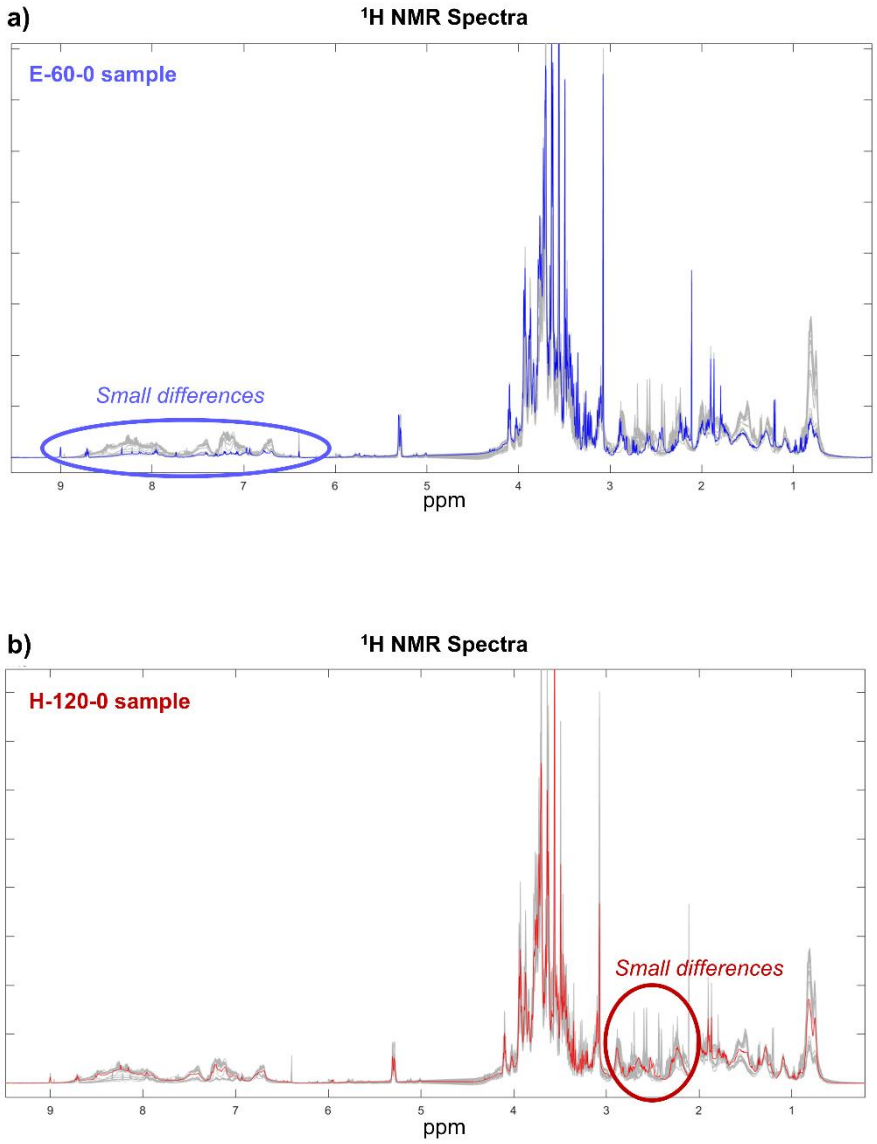

Figure 3.4. ¹H-NMR spectra of the two samples hypothesized as outliers, E-60-0 in blue (A) and H-120-0 in red (B). Both spectra show small differences with respect to the other overlapped spectra (grey)

To explore possible additional treatment effects among the samples, a second PCA was performed considering only the samples treated without the enzyme. From the scores plot of PC1 and PC2 (Figure 3.5), accounting a total explained variance of 77 %, a clear separation between the zero-time samples and the other ones was found. To better explore the extracted only samples, the two zero-time samples were removed, and a new PCA model was developed. By exploring the PC2 vs PC3 scores plot (Figure 3.6A), all samples but one resulted separated in two clusters on the basis of the stirring rate, demonstrating that PCA can be used to highlight differences related to the stirring rate. By exploring the only sample that was not correctly placed, it reveals a very high Q residual value, making it a potential outlier (Figure 3.6B).

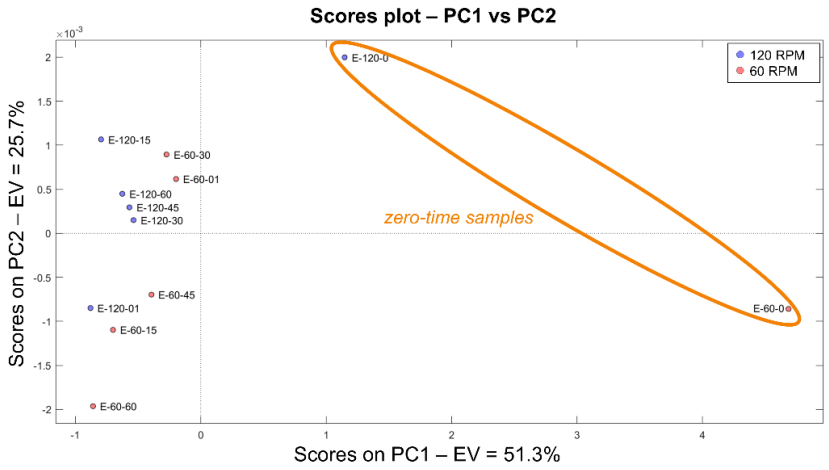

Figure 3.5. Scores plot of PC1 vs PC2 obtained from a PCA performed considering the spectra obtained only from the samples treated with the extraction process without the hydrolysis enzyme. Zero-time samples revealed to be different from other samples.

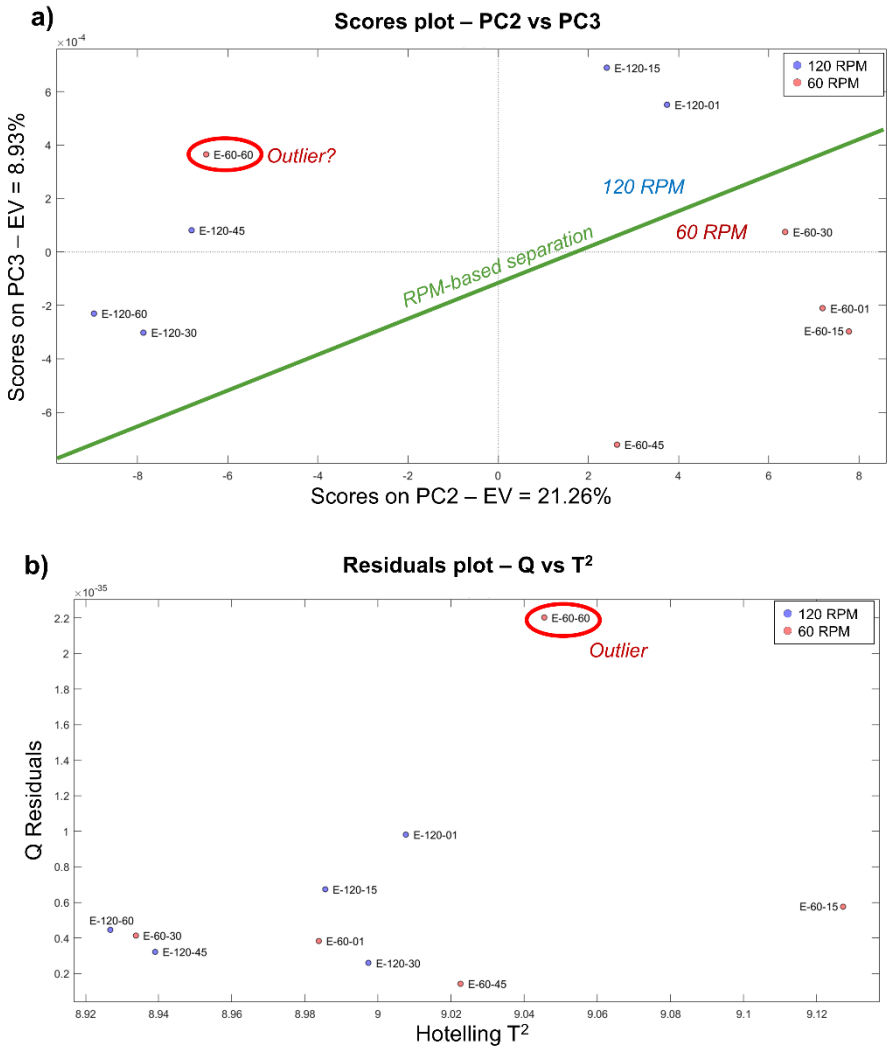

Figure 3.6. Scores plot of PC2 vs PC3 (a) obtained from a PCA performed considering the spectra obtained only from the samples treated with the extraction process without the hydrolysis enzyme to highlight an RPM-based separation. In the residuals plot (b), the only sample placed wrong is highlighted.

3.5.2 PCA for time-trend evaluation and process optimization

Since the aim of the project was to monitor and optimize the industrial treatment of lentil flour hydrolysation, the core part of the data analysis was focused specifically on those samples for which the enzyme was added, so on the 20 hydrolysed samples.

To start with data exploration, a PCA model was computed to inspect for spontaneous clusters, trends and potential outliers. As expected, by looking at the scores plot of PC1 (Figure 3.7), containing the largest part of the information in the data (85.11 % of the total explained variance), it results clear that this single PC is enough to separate the samples withdrawn at time zero and time one from the other ones. Once again, to better investigate potential trends among the other hydrolysed samples, the time-zero and time-one samples were removed from the data and a new PCA model was computed.

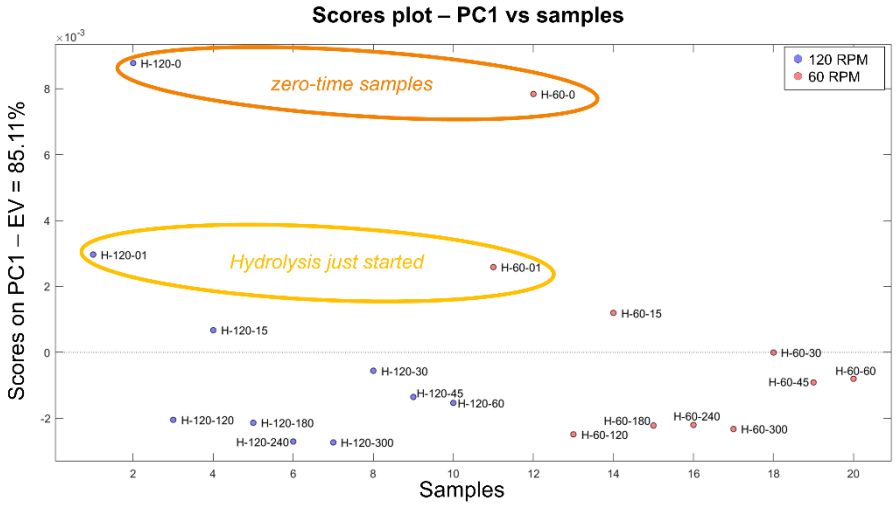

Figure 3.7. Scores plot of PC1 vs samples obtained from a PCA performed considering the spectra obtained only from the samples treated with the hydrolysis enzyme. Zero-time samples and samples collected after one minute of treatment are highlighted.

From this second PCA, by exploring the scores plot of the first principal component that describes 65.6 % of the total variance (Figure 3.8A), a trend related to the time of the enzymatic activity was spotted. Interestingly, by looking at the loadings plot (Figure 3.8B), some signals seem to be related to this particular trend, meaning that the concentration of the compounds they generate from changes during treatment.

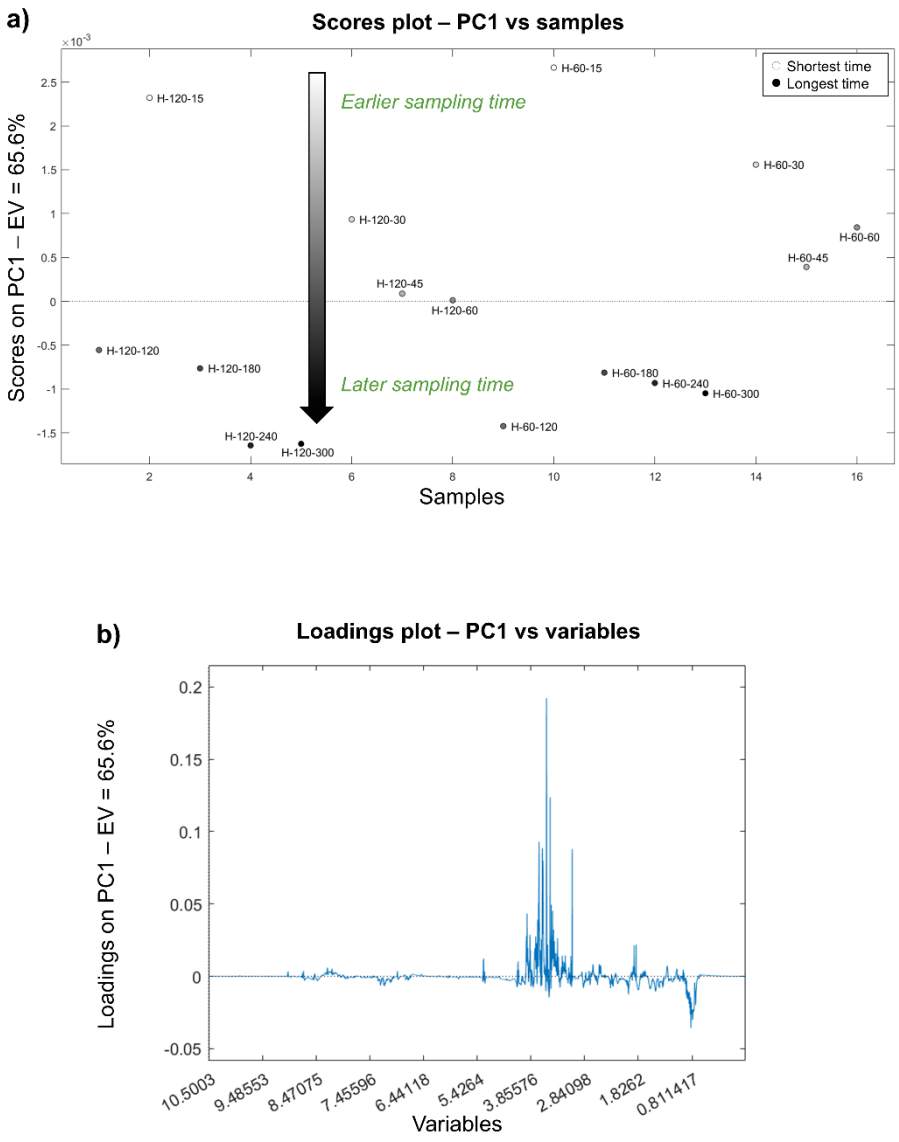

Figure 3.8. Scores plot of PC1 vs samples (A) and loadings plot (B) obtained from a PCA performed considering the spectra obtained only from the samples treated with the hydrolysis enzyme for more than one minute. In the scores plot samples are coloured according to the treatment time, a time-related trend is spotted along PC1. Loadings plot allow to understand which NMR signals mostly characterize the samples.

To better evaluate the spotted time trend, the scores values of these samples were extracted, and two different plots were created for the two set stirring rates by plotting the scores against the time at which the samples were withdrawn and the hydrolysis was stopped. Figure 3.9A is the plot obtained from the samples treated at 60 RPM, while Figure 3.9B is the plot obtained from those treated at 120 RPM. For both plots a sort of plateau seems to be reached at some point, very likely corresponding to the end of the hydrolysis process. In theory, since at the first point of the plateau the hydrolysis should be considered complete, this timepoint could correspond to the optimal processing time. From the two plots, the two hypothetical optimal conditions are reached at 60 RPM for 120 minutes or at 120 RPM for 240

minutes., respectively. In addition, to confirm that these two experimental conditions really lead to obtain comparable hydrolysation results, the PC1 vs PC2 scores plot (Figure 3.10) with all the hydrolysed samples was explored. Interestingly, the two samples corresponding to the two potential optimal experimental conditions are located very close each other, meaning that they actually are very similar, and their experimental conditions can be directly compared. Therefore, the best experimental conditions, based only on these explored possibilities, are achieved when treating the samples for 120 minutes at a 60 RPM stirring rate. This choice, rather than the other hypothesized "optimal" set of conditions, allow working with a lower stirring rate and for a shorter time, thus saving both time and energy resources.

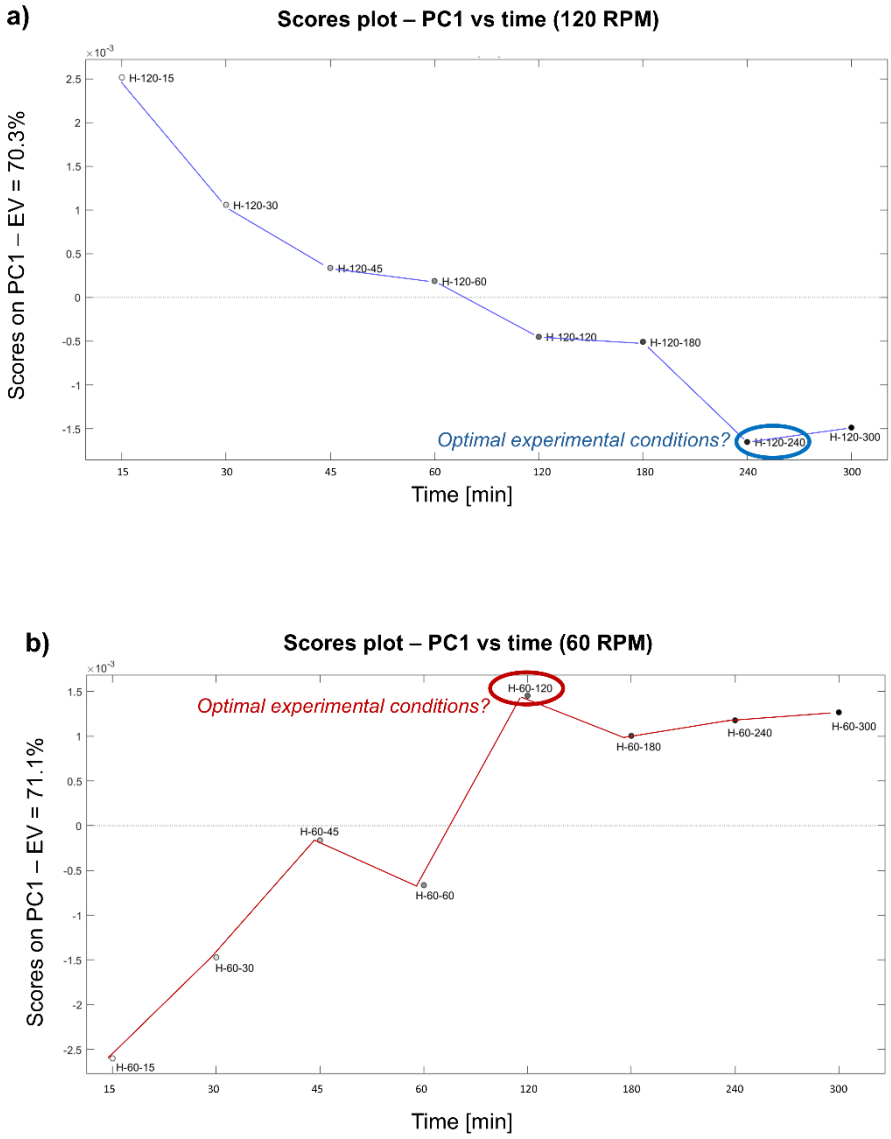

Figure 3.9. Scores plot of PC1 vs treatment time obtained from a PCA performed on the samples treated at 120 RPM (A) and from PCA performed on the samples treated at 60 RPM. Highlighted samples (H-120-240 in blue and H-60-120 in red) are the samples that could correspond to the optimal experimental conditions.

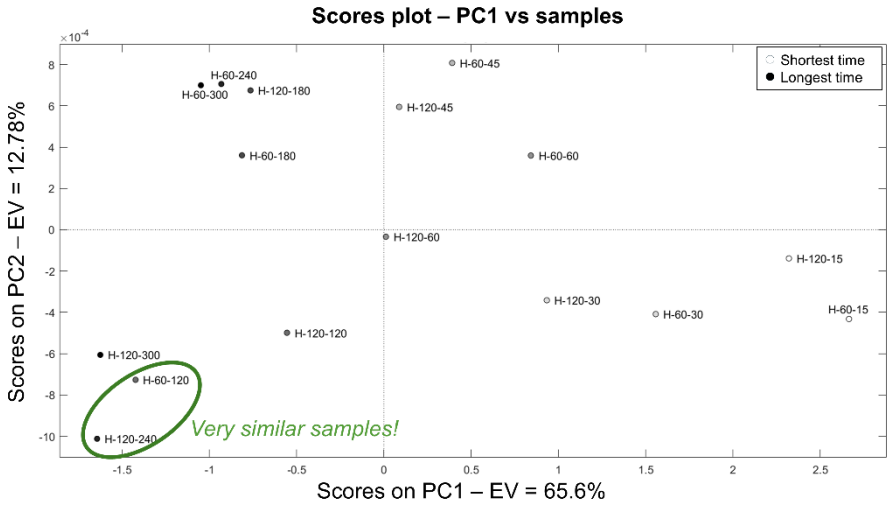

Figure 3.10. Scores plot of PC1 vs PC2 obtained from a PCA performed considering the spectra obtained only from the samples treated with the hydrolysis enzyme for more than one minute. The two highlighted samples are the one obtained with the hypothesized optimal conditions, since their score values are similar, they are similar samples.

3.6 Project conclusions and future perspectives

This explorative study demonstrates the potential of using ^{1}H-NMR spectroscopy combined with multivariate data analysis methods like PCA to monitor and optimize industrial food treatments. By focusing on extraction and hydrolyzation process of lentil flour as a case study, the results confirmed the ability of PCA to identify key factors influencing the process, such as enzyme activity, stirring rate, and treatment time. The analysis highlighted significant trends and separations among samples, providing insights into how enzymatic activity and experimental conditions affect the sample distribution in the PCA graphical outputs (i.e., scores plot). Additionally, PCA successfully identified outliers and potential artifacts in the data, emphasizing its suitability for quality control in experimental workflows. These findings highlight the potential of multivariate chemometric tools to enhance agrifood processes, improving efficiency and sustainability in bioeconomy applications.

To conclude, the future of this project can follow different directions since many more parameters can be evaluated with this approach. For example, different enzymes and experimental conditions can be explored by implementing a structured experimental design, paving the way for an optimized industrial application of this treatment. Taking advantage form the information carried by the NMR data, a metabolite identification step, performed by assigning to the spectral signals to specific molecules, can be useful to better understand which compounds are involved in the extraction and in the hydrolyzation. This approach will help understanding which metabolic pathway the samples differences, highlighted in the PCA scores plots previously obtained, derive from.

References | Chapter 3

[1] O.L. Tavano, Protein hydrolysis using proteases: An important tool for food biotechnology, J Mol Catal B Enzym 90 (2013) 1–11. https://doi.org/10.1016/j.molcatb.2013.01.011.

[2] M. Jarpa-Parra, Lentil protein: a review of functional properties and food application. An overview of lentil protein functionality, Int J Food Sci Technol 53 (2018) 892–903. https://doi.org/10.1111/IJFS.13685.

[3] P. Morales, J.D.J. Berrios, A. Varela, C. Burbano, C. Cuadrado, M. Muzquiz, M.M. Pedrosa, Novel fiber-rich lentil flours as snack-type functional foods: an extrusion cooking effect on bioactive compounds, Food Funct 6 (2015) 3135–3143. https://doi.org/10.1039/C5FO00729A.

[4] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[5] R. Jia, Y.C. Song, Z. An, K. Kim, C.Y. Lee, B.U. Bae, A New Comprehensive Indicator for Monitoring Anaerobic Digestion: A Principal Component Analysis Approach, Processes 2024, Vol. 12, Page 59 12 (2023) 59. https://doi.org/10.3390/PR12010059.

[6] X. Liu, X. Chen, W. Wu, Y. Zhang, Process control based on principal component analysis for maize drying, Food Control 17 (2006) 894–899. https://doi.org/10.1016/J.FOODCONT.2005.06.008.

[7] R. Leardi, Experimental design in chemistry: A tutorial, Anal Chim Acta 652 (2009) 161–172. https://doi.org/10.1016/j.aca.2009.06.015.

[8] Ž.R. Lazić, Design of Experiments in Chemical Engineering, Design of Experiments in Chemical Engineering (2004). https://doi.org/10.1002/3527604162.

[9] H. Ebrahimi-Najafabadi, R. Leardi, M. Jalali-Heravi, Experimental Design in Analytical Chemistry—Part I: Theory, J AOAC Int 97 (2014) 3–11. https://doi.org/10.5740/JAOACINT.SGEEBRAHIMI1.

[10] S.A. Weissman, N.G. Anderson, Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications, Org Process Res Dev 19 (2015) 1605–1633. https://doi.org/10.1021/OP500169M/ASSET/IMAGES/LARGE/OP-2014-00169M_0027.JPEG.

[11] M. Liu, X.A. Mao, C. Ye, H. Huang, J.K. Nicholson, J.C. Lindon, Improved WATERGATE Pulse Sequences for Solvent Suppression in NMR Spectroscopy, Journal of Magnetic Resonance 132 (1998) 125–129. https://doi.org/10.1006/JMRE.1998.1405.

[12] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, Journal of Magnetic Resonance 202 (2010) 190–202. https://doi.org/10.1016/j.jmr.2009.11.012.

[13] F. Savorani, G. Tomasi, S. Engelsen, Alignment of 1D NMR Data using the iCoshift Tool: A Tutorial, in: 2013: pp. 14–24. https://doi.org/10.1039/9781849737531-00014.

[14] P. Giraudeau, I. Tea, G.S. Remaud, S. Akoka, Reference and normalization methods: Essential tools for the intercomparison of NMR spectra, J Pharm Biomed Anal 93 (2014) 3–16. https://doi.org/10.1016/J.JPBA.2013.07.020.

Chapter 4

Case study #2: Data fusion to improve food traceability e authenticity

4.1 Project overview

Protection of food authenticity is an increasingly hot topic in many countries worldwide. Different international organizations, like the European Union, have issued several documents and regulations to ensure food quality. This activity requires to develop accurate methods able to detect food adulteration or mislabelling. Several analytical techniques have demonstrated to be suitable to determine food authenticity. However, since each analytical technique has specific advantages and disadvantages, the recent literature has highlighted the benefits of using multi-technique data fusion approaches [1] as they can provide more robust results than those obtained independently from each individual dataset [2]. In fact, these approaches enhance the extracted information by leveraging the strengths of different analytical techniques. Thanks to the possibility of combining different characterization techniques, this approach can be adapted to different food matrices and conditions according to the needs [3,4].

Italy is the second largest producer of hazelnuts (*Corylus avellana L.*) worldwide (INC International Nuts & Dried Fruit), and one of the principal importers of fresh nuts as well as exporter of hazelnut-based products [5]. Italian hazelnut production is mostly certified by the Protected Designation of Origin (PDO), the Protected Geographical Indication (PGI) (Regulation (EU) No 1151/2012), and the Traditional Speciality Guaranteed (TSG) certifications (Council Regulation (EC) No 509/2006). In this regard, the concept of "terroir", linking primary productions with a geographical origin, is being extended beyond the wine sector [6] and applies to hazelnuts.

The richness of hazelnuts in secondary metabolites is mirrored in well-recognised sensory quality traits making them an excellent ingredient in the confectionery industry, chocolate spreads, and processed foods. In particular, the cultivar "Tonda Gentile Trilobata" (TGT) from Piedmont (Italy) is highly appreciated for its organoleptic properties and is considered a "gold standard" for product quality [7].In 1993, it also obtained the PGI collective label from the European Union under

the name of “Nocciola Piemonte” (“Piemonte” being the Italian term for “Piedmont”). Due to its limited production, the high-value TGT is at risk of fraud practices.

In this project, different multivariate statistical methods were applied to data coming from different analytical techniques combined through a data fusion approach. The chosen techniques, Nuclear Magnetic Resonance (NMR) and Liquid Chromatography coupled with Mass Spectrometry (LC-MS), have already been proven suitable to be used for origin traceability [8,9]. The potential of this approach was tested through a research case study focused on hazelnuts of the TGT cultivar collected in Piedmont (Italy) compared with hazelnuts of other origins and cultivars.

4.1.1 Employed techniques

NMR is one of the most employed techniques in food characterisation and authentication due to its non-destructivity, rapidity, and simultaneous detection of all the major organic classes of compounds [10]. The high reproducibility of NMR, combined with its low difficulty in the automatization of the entire analytical process, even with a very high number of samples, makes it suitable for high-throughput analyses [11]. This technique, coupled with chemometrics and multivariate statistical analysis, was successfully used for classification and authentication of many food products [12] such as extra virgin olive oils [13], parmesan [14], tomato [15], honey [16], coffee [17], fruit juices [18], beers [19], wines [20], and balsamic vinegar [21].

Regarding hazelnuts, Caligiani *et al.* discriminated between TGT and “Tonda Gentile Romana” varieties from Turkey blend cultivars according to their NMR profiles [22], while Bachmann *et al.* discriminated among countries and harvest seasons [23]. The discriminating ability of NMR is due to its capacity to detect specific metabolites and markers that are characteristics of a particular geographical origin [24], cultivar [25] or that characterize one specific production year [26]. These factors have an impact on the metabolite composition of hazelnuts (but, of course, this is valid for food and natural products in general), resulting in unique NMR profiles; even if not all the signals can always be identified, the signals pattern in the spectrum can be considered a “fingerprint” of the specific sample, leading to NMR-based discriminations.

LC-MS is also widely applied to guarantee food safety, quality, traceability, and authenticity, especially concerning high-resolution untargeted metabolomics approaches [27]. Differently from NMR, LC-MS shows remarkable sensitivity and excellent identification capabilities [28]. In the last years, this approach has been applied to a broad range of diverse high-value food matrices such as extra-virgin olive oils [29,30], milk [31], and PDO cheeses [32], saffron [33], rice [34] and black

rice [35], nuts [36], and maize [37]. For what it concerns hazelnuts, in a work by Lelli *et al.* [38] the authors discriminated Tonda gentile Romana and Tonda Giffoni with the same origin and identified the relative marker metabolites, mainly belonging to polyphenols.

Compared to NMR, LC-MS has much higher sensitivity and higher dynamic range reliability [37] despite having lower reproducibility and being more limited in absolute quantification. It is worth mentioning that it is a destructive technique, requiring a higher number of representative samples. Noteworthy, based on the respective compounds' coverage profiles, these two techniques can be considered as complementary analytical methods [39].

The datasets obtained with ^{1}H-NMR and LC-MS were firstly inspected individually by PCA [40] to search for information related to harvest year, geographical origin, and hazelnut cultivar. To better investigate the information related to Piedmont's geographical origin and TGT cultivar, the datasets were also investigated using supervised approaches (PLS-DA) and data fusion methods [41,42].

4.2 Sample collection, data acquisition and processing

4.2.1 Samples collection and preparation

In this study, a total of 54 hazelnut samples were analysed. Of these, 44 samples were from different Italian regions, 4 samples from Turkey, 2 samples from Romania, 2 samples from Chile, 1 sample from Georgia, and 1 sample from Bulgaria; 29 were of the cultivar Tonda Gentile Trilobata, 6 Giffoni, 2 Nocchione, 1 Nostrale, and 4 of other minor cultivars. Sixteen samples were collected in 2020 and 38 were collected in 2021. More detailed information about the origin and cultivar is provided in Table 4.1.

The sample preparation was the same for both the techniques. Coarsely ground samples were dispersed in an 80:20 methanol:water solution using a Ultra-Turrax blender. The samples were then centrifuged and filtered (cellulose membrane, 0.22 μm) into vials for analysis [36]. The LC-MS and NMR analyses were performed on the methanolic extract. To discuss the potential benefits of the combined multi-omics metabolomic profiling of hazelnuts, it is important to point out how the data were collected. In particular, a single preparation protocol was used for homogenization and extraction, starting from the same raw materials, without dedicated workflows.

*Table 4.1. Hazelnuts samples with origin, variety and harvest year (*late harvested batch)*

ID	Geographical Origin			Variety	Harvest year
	District	**Region**	**Country**		
5	/	/	Bulgaria	Unknown	2020
16	/	/	Cile	Tonda Gentile	2020
21	Matelica	Marche	Italy	Unknown	2020
2	Alessandria	Piedmont	Italy	Tonda Gentile	2020
3	Asti	Piedmont	Italy	Tonda Gentile	2020
11	Camerana	Piedmont	Italy	Tonda Gentile	2020
12	Camerana	Piedmont	Italy	Tonda Gentile	2020
17	Cortemilia	Piedmont	Italy	Tonda Gentile	2020
46	Torre Bormida	Piedmont	Italy	Tonda Gentile	2020
50	Verduno	Piedmont	Italy	Tonda Gentile	2020
37	Palermo	Sicily	Italy	Nostrale (Sicily)	2020
39	/	Piedmont	Italy	Tonda Gentile	2020
40	/	Piedmont	Italy	Tonda Gentile	2020
49	/	Umbria	Italy	Giffoni	2020
44	/	/	Romania	Tonda Gentile	2020
48	/	/	Turkey	Unknown	2020
15	/	/	Cile	Tonda Gentile	2021
18	/	/	Georgia	Tonda Gentile	2021
36	Olivetto Citra	Campania	Italy	Giffoni	2021
42	Pompei	Campania	Italy	Giffoni	2021
45	Terzigno	Campania	Italy	Giffoni	2021
22	Bardi	Emilia	Italy	Tonda Gentile	2021
23	Bardi	Emilia	Italy	Tonda Gentile	2021
24	Bardi	Emilia	Italy	Tonda Gentile	2021
25	C. S. Pietro Terme	Emilia	Italy	Tonda Gentile	2021
26	C. S. Pietro Terme	Emilia	Italy	Tonda Gentile	2021
27	C. S. Pietro Terme	Emilia	Italy	Tonda Gentile	2021
28	Grizzana Morandi	Emilia	Italy	Unknown	2021
29	Sesto Imolese	Emilia	Italy	Nocchione	2021
30	Sesto Imolese	Emilia	Italy	Nocchione	2021
31	Sesto Imolese	Emilia	Italy	Giffoni	2021
32	Sesto Imolese	Emilia	Italy	Giffoni	2021
33	Sesto Imolese	Emilia	Italy	Giffoni	2021
34	Sesto Imolese	Emilia	Italy	Giffoni	2021
4	Borgorose	Lazio	Italy	Giffoni	2021
1	Barge	Piedmont	Italy	Tonda Gentile	2021*
6	Carentino	Piedmont	Italy	Tonda Gentile	2021*
7	Feisoglio	Piedmont	Italy	Tonda Gentile	2021*
8	Piagera	Piedmont	Italy	Tonda Gentile	2021*
9	Murazzano	Piedmont	Italy	Tonda Gentile	2021*
10	Castino	Piedmont	Italy	Tonda Gentile	2021*
13	Camerana	Piedmont	Italy	Tonda Gentile	2021
20	Levice	Piedmont	Italy	Tonda Gentile	2021
38	Pezzolo	Piedmont	Italy	Tonda Gentile	2021

47	Torre Bormida	Piedmont	Italy	Tonda Gentile	2021
51	Verduno	Piedmont	Italy	Tonda Gentile	2021
41	Palermo	Sicily	Italy	Nostrale (Sicily)	2021
14	Caposele	Campania	Italy	Giffoni	2021
35	/	Lazio	Italy	Nocchione	2021
19	/	Piedmont	Italy	Tonda Gentile	2021
43	/	/	Romania	Tonda Gentile	2021
52	/	/	Turkey (A)	Unknown	2021
53	/	/	Turkey (B)	Unknown	2021
54	/	/	Turkey (C)	Unknown	2021

4.2.2 ^{1}H-NMR analysis

For proton nuclear magnetic resonance analysis, the methanol solvent was removed by under-vacuum evaporation, and the metabolic mixtures were redissolved in deuterated methanol. Each ^{1}H-NMR tube was prepared at least in duplicate by adding 600 μL of the sample solution and 10 μL of trimethylsilylpropanoic-d4 (TSP-d4) acid standard solution 0.005 M.

Also for this study, the ^{1}H-NMR analysis was performed on a Jeol ECZR 600 spectrometer (JEOL Ltd., Akishima, Tokyo, Japan) operating at 600.17 MHz for protons. The spectra were collected at a fixed temperature of 298 K by acquiring 32768 points and performing 256 scans for each sample, using a 30 s relaxation delay. A solvent suppression procedure (Dante pre-saturation) was applied to remove the water signal [43]. The spectra were baseline- and phase-corrected using the DELTA processing tool offered by JEOL Ltd. The raw ^{1}H-NMR spectra were imported and processed under MATLAB environment (R2021b, Mathworks, Natick, MA, USA). For all the spectra, the ppm scale was referenced to the TSP peak (0.00 ppm) (Figure 4.1). To increase the comparability among spectra composing the spectral dataset, the "*i*coshift" [44,45] tool was applied to horizontally align the most important signals located inside specific intervals, accurately manually defined. The spectra width was corrected to include only signals between -0.1 ppm and 9.3 ppm to remove unwanted and noisy areas. To avoid interferences, methanol signals (from 3.26 ppm to 3.38 ppm) and water residues (from 4.55 and 5.05 ppm) were removed from the spectra.

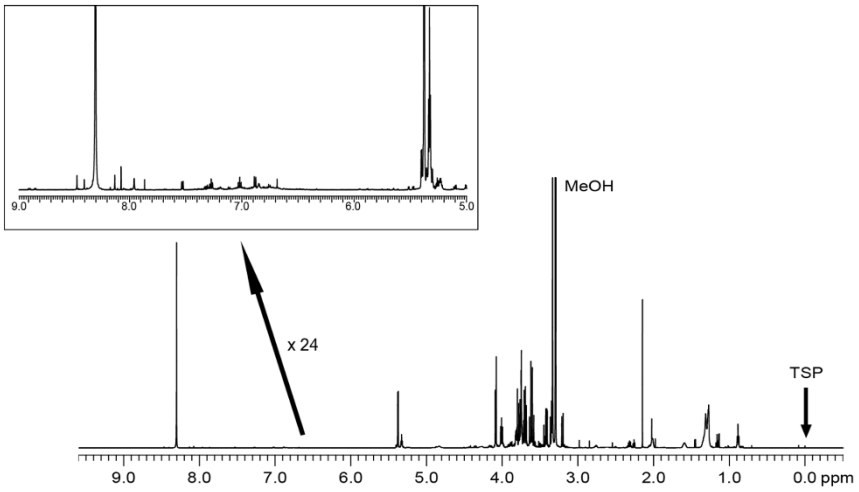

Figure 4.1. Example of ^{1}H-NMR spectra obtained from hazelnut analysis (sample n°50 from Table 4.1)

After data collecting, the NMR signals were assigned in order to identify the corresponding metabolites. Signal identification allows taking one step further towards clarity and interpretability: each signal, or group of signals is directly related to a specific molecule. Assigning a chemical name to the variables makes the model interpretation much easier [46]. To better understand and evaluate the information captured with the ^{1}H-NMR analyses, a tentative identification of the metabolites was performed on the ^{1}H-NMR spectra. The attributions given to the resolved signals were gathered starting from the signals' shape, position and multiplicity. By comparison with assignments from literature [22,23,47] it was possible to label most of the unique signals. In this respect, other sources of information were the reference library from the Chenomx NMR Suite (Version 10.0), which is a software dedicated to the interpretation of NMR spectra, and the Human Metabolome Database (HMDB) [48], which, among the others, contains metabolites also found in hazelnuts. The 28 identified metabolites are shown in Table 4.2, that also contains information related to the chemical shift and the multiplicity of the signals assigned to specific metabolites. Nevertheless, some signals remained unassigned because no match was found in these metabolite repositories.

Table 4.2. List of the assigned metabolites with tentative names, chemical shifts (δ, ppm) and signal multiplicity.

Compound name	Chemical shift (δ, ppm)	Multiplicity	Reference
TSP	0	s	
Beta sitosterol	0.707	s	Caligiani *et al.* (2014)
Isoleucine	0.87	t	Caligiani *et al.* (2014)
	1.06	d	Bachmann *et al.* (2018); Schmitt *et al.* (2020)
Leucine	0.9	t	Caligiani *et al.* (2014); Schmitt *et al.* (2020)
Valine	1.015	d	Caligiani *et al.* (2014)
	1.037	d	Bachmann et al. (2018); Schmitt et al. (2020)
Threonine	1.3	d	Bachmann et al. (2018); Schmitt et al. (2020)
Alanine	1.46	d	Caligiani et al. (2014); Schmitt et al. (2020)
Arginine	1.72	m	Caligiani et al. (2014); Schmitt et al. (2020)
Acetic acid	1.932	s	Bachmann et al. (2018); Schmitt et al. (2020)
Acetyl glutamate	2.015	s	
	2.27	t	
Glutamic acid	2.16	s	Caligiani et al. (2014); Schmitt et al. (2020)
Malate	2.34	m	Bachmann et al. 2018); Schmitt et al. (2020)
Succinic acid	2.555	s	Caligiani et al. (2014)
Malic acid	2.775	dd	Caligiani et al. (2014)
Asparagine	2.9	dd	Caligiani et al. (2014)
Choline derivate	3. 179	s	Caligiani et al. (2014)
Choline	3. 201	s	Caligiani et al. (2014); Schmitt et al. (2020)
Sucrose	3.44	dd	Caligiani et al. (2014)
	3.615	m	Bachmann et al. (2018)
	3.76	m	Schmitt et al. (2020)
	4.01	t	
	4.08	s	
	4.11	s	
	5.377	d	
Glycerol	3.505	m	Caligiani et al. (2014)
α-hydroxy acid	4.3	m	Caligiani et al. (2014)
Ribose nucleotides	4.485	m	Caligiani et al. (2014)
Oligosaccharides	5.399	m	Caligiani et al. (2014)
	5.41	m	
	5.515	m	
Fumarate	6.65	s	Bachmann et al. (2018)
Tyrosine	6.773	d	Caligiani et al. (2014)
	7.128	m	Bachmann et al. (2018)
Indole derivate	6.9	d	
	7.035	dt	
	7.283	dt	
	7.53	d	
Tryptophan	7.384	m	Caligiani et al. (2014)
	7.69	d	
Trigonelline	8.043	dd	Caligiani et al. (2014)
	8.841	d	
	8.889	d	
	9.178	s	
Adenosine	8.14	s	Caligiani et al. (2014)
Formic acid	8.48	s	Bachmann et al. (2018); Schmitt et al. (2020)

4.2.3 Liquid Chromatography coupled with Mass Spectrometry

Liquid Chromatography coupled with Mass Spectrometry (LC-MS) analysis was performed by the research group of Professor Luigi Lucini at the Catholic University of Sacred Hearth in Piacenza. They employed an ultra-high-pressure liquid chromatography (1290 series, Agilent Technologies, Santa Clara, CA, USA) coupled to a quadrupole-time-of-flight mass spectrometer (6550 iFunnel, Agilent Technologies), as reported previously [33]. The mass spectrometer worked in full-scan mode, with a positive ionisation (ESI +), to acquire accurate masses in the 100–1200 m/z range. The chromatographic separation was performed on an Agilent Zorbax Eclipse plus C18 analytical column (50 × 2.1 mm, 1.8 μm), using water-acetonitrile gradient elution (from 6 % to 94 % organic in 34 min). The injection volume was 6 μL per analysis; three replicates for each sample were analysed. The variables were aligned for mass (5 ppm accuracy) and retention time (0.05 min) and annotated according to the "find by-formula" algorithm using the Agilent Profinder B.07 software against the database FooDB. To this aim, the whole isotopic pattern of molecular features (accurate monoisotopic mass, isotope spacing, and ratio) was used as previously described [9]. Data filtering was also carried out in Profinder B.07, retaining only the compounds identified within 100 % of replications in at least one treatment. The Agilent Mass Profiler Professional B.12.06 software was finally used as post-acquisition pre-processing, as previously reported [49]. Therein, compounds were filtered by abundance considering only those compounds with an area >5000 counts, normalised at the 75th percentile and baselined to their median in the dataset.

4.3 PCA and PLS-DA on independent datasets

The datasets obtained from ^{1}H-NMR and LC-MS were firstly analysed and treated separately. Mean centering was used to preprocess both ^{1}H-NMR and LC-MS data. Principal Component Analysis (PCA) [50,51] was performed using a MATLAB toolbox specific for exploratory analysis [52]. Different PCA models were created using different datasets. In addition, to visualize groupings and potential outliers, the samples in the scores' plots were coloured according to the different features inspected (i.e., harvest year, geographical origin, and hazelnut cultivar).

In addition, a PLS-DA [53] modelling approach was also used to investigate the discriminating ability with respect to some classes of samples (i.e., Piedmont hazelnuts and TGT cultivar). For each dataset, two models were created, one to observe how the samples with Piedmont geographical origin were separated from the other having different origins, and the second to evaluate the samples' classification with respect to the TGT cultivar. Due to the low number of total samples, the approach used to validate PLS-DA models was cross validation for all the PLS-DA models built in this phase.

4.3.1 Results for the NMR dataset

The ^{1}H-NMR spectral dataset is structured in 54 samples and 10508 variables. With a first PCA model performed including all samples, two major clusters associated with the harvest year were observed in the scores plot of the first two PCs (Figure 4.2). In addition, by exploring the samples, a cluster related to the late-harvested hazelnuts was identified, later confirmed also by the analysis of LC-MS data.

The largest part of the information in the dataset resulted related to the harvest year (Figure 4.2). To evaluate the information related to the origin and the cultivar, the data corresponding to the two harvest years were split in two datasets; however, because of the higher number of samples and a better distribution in terms of geographical origins and cultivars, only data from 2021 were included in this exploratory analysis. A PCA model was created to evaluate the information about the origin and the cultivar. Different clusters related to the geographical origin were found in the scores plot of the first two PCs. In particular, three groups were identified for hazelnuts from Piedmont, Emilia, and from outside Italy (Figure 4.3). As previously obtained by Bachmann *et al.* [23], the ^{1}H-NMR analysis allowed to distinguishing Italian hazelnuts from foreign samples. However, the variance of a single sample group is sometimes larger than the distance between two groups, especially for samples coming from the same country. Finally, looking at the samples' cultivar, from the scores plot of the first two PCs, it was possible to identify clusters related to TGT and Giffoni cultivars (Figure 4.4).

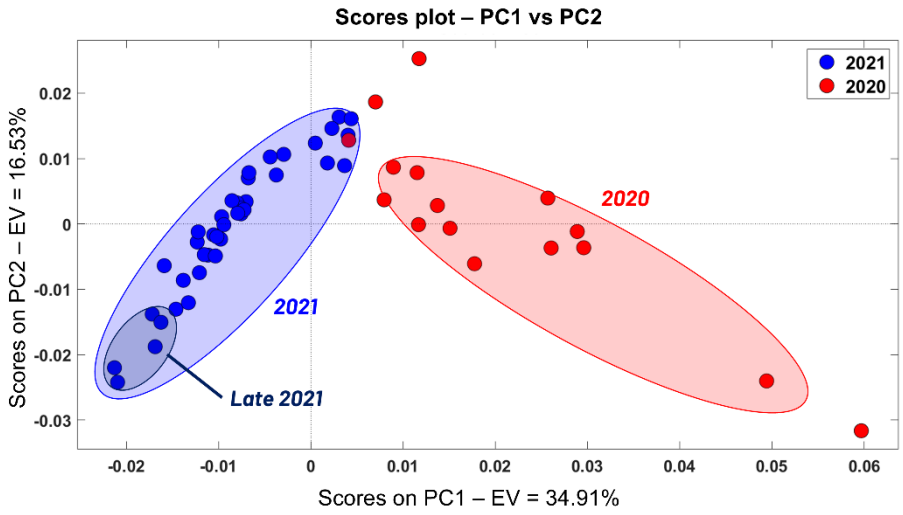

Figure 4.2. Scores plot with groupings obtained from a PCA model on ^{1}H-NMR dataset with the samples coloured according to the harvest year.

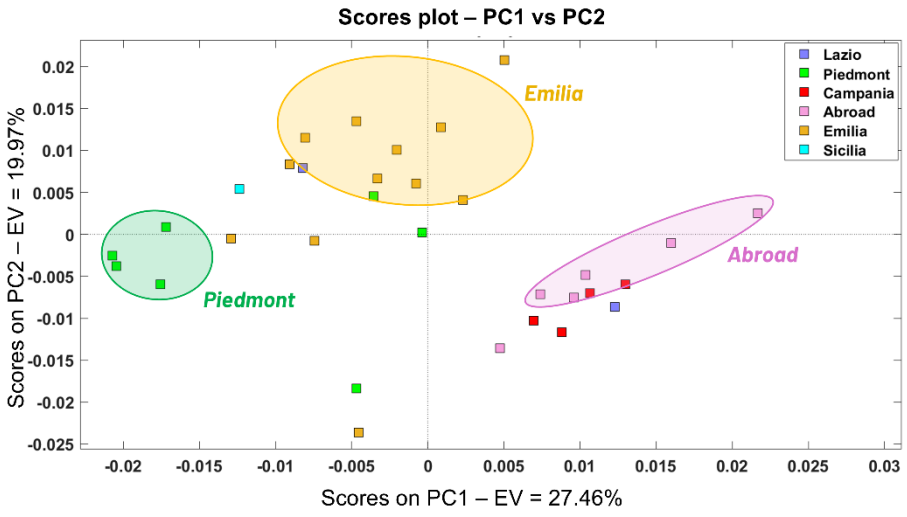

Figure 4.3. Scores plot obtained from a PCA model on ^{1}H-NMR dataset of the samples harvested in 2021. The samples are coloured according to the geographical origin (region).

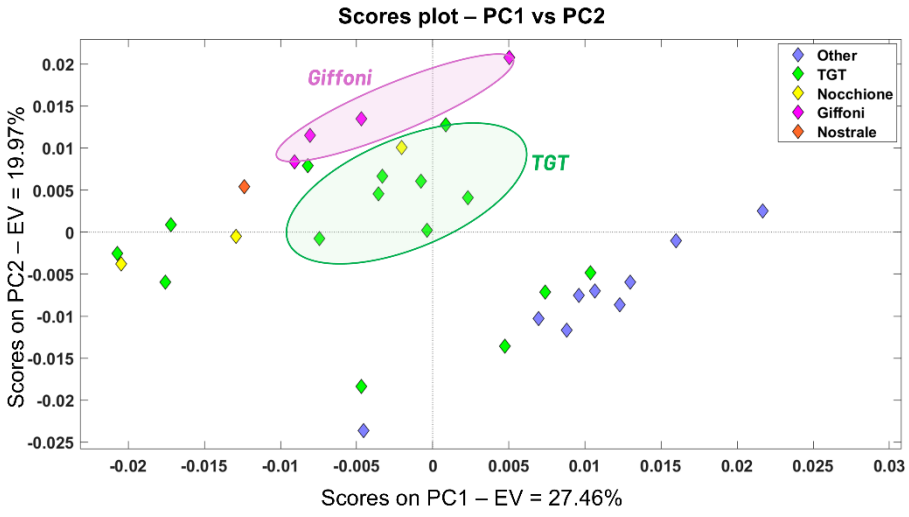

Figure 4.4. Scores plot obtained from a PCA model on ^{1}H NMR dataset of the samples harvested in 2021. The samples are coloured according to the hazelnut variety.

Two distinct supervised models were created to better evaluate the possibility of discriminating the TGT hazelnuts from those cultivated abroad and the Piedmont hazelnuts from other geographical origins. A first PLS-DA model was developed to identify the TGT cultivar among the other cultivars from the 2021 samples. By looking at the scores plot of the first two Latent Variables (LV), a cluster related only to the TGT cultivar was spotted (Figure 4.5).

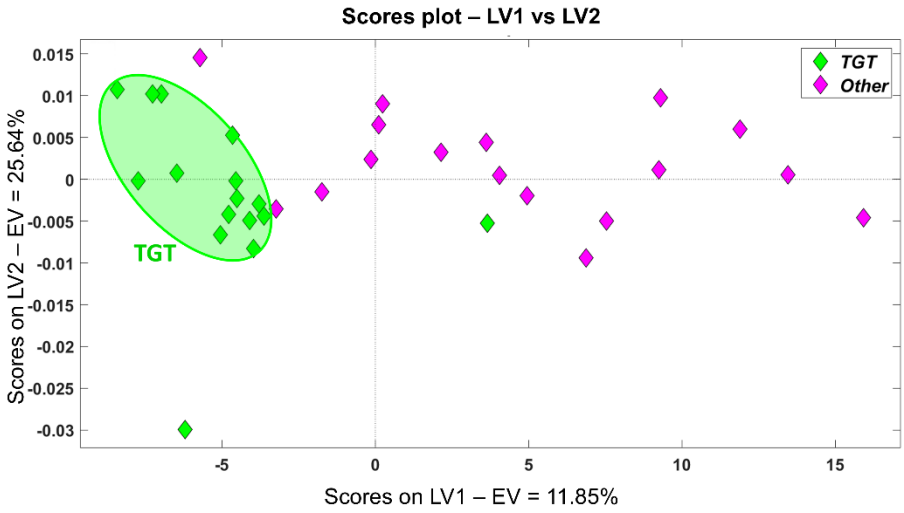

Figure 4.5. Scores plot obtained from a PLS-DA model built with the 2021 samples from ^{1}H-NMR dataset where the two classes considered were hazelnuts of the TGT cultivar and other cultivars.

The same approach was used to evaluate a possible separation between Piedmont hazelnuts and all the others. By exploring the results, a slight separation based on geographical origin was spotted (Figure 4.6). These two pieces of evidence confirmed the results previously highlighted by Bachmann *et al.* (2018), which successfully differentiated hazelnut samples according to geographical origin, and improved the results obtained by Caligiani *et al.* (2014) about the TGT hazelnut characterization [22]. This first exploration suggests that, for ^{1}H-NMR spectroscopy, even if most of the information allows us to distinguish the samples according to the harvest year, it is possible to find information strictly related to the geographical origin and to the cultivar of the samples.

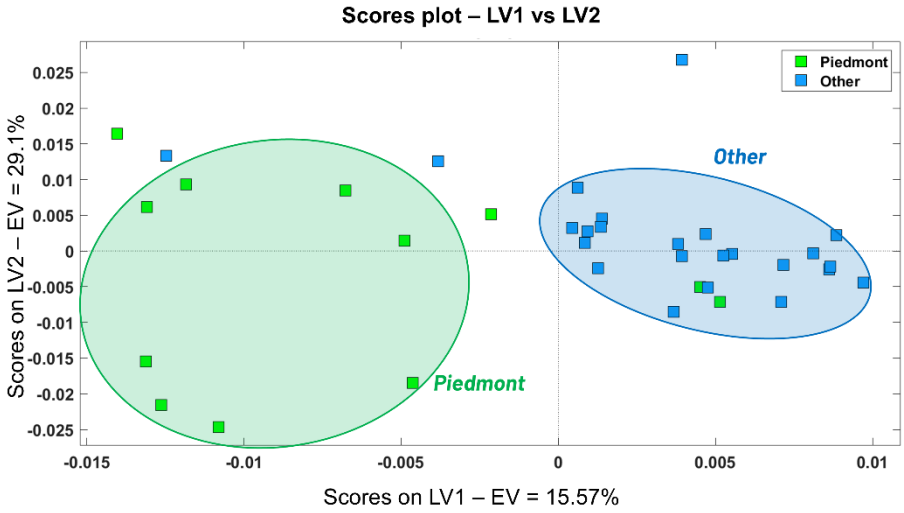

Figure 4.6. Scores plot obtained from a PLS-DA model on ^{1}H-NMR dataset of the samples harvested in 2021. The samples were separated in two distinct classes according to Piedmont geographical origin or other geographical origin.

4.3.2 Results for the LC-MS dataset

The dataset obtained from LC-MS was explored through different PCA models, where the samples were labelled according to the harvest year, the geographical origin, and the cultivar. From the first PCA, performed considering all samples, three major clusters, associated with different harvest years, were found. PC2 mainly allows separating the sample from seasons 2020 and 2021, while PC1 also discriminates the late samples from 2021 from the other 2021 hazelnuts (Figure 4.7).

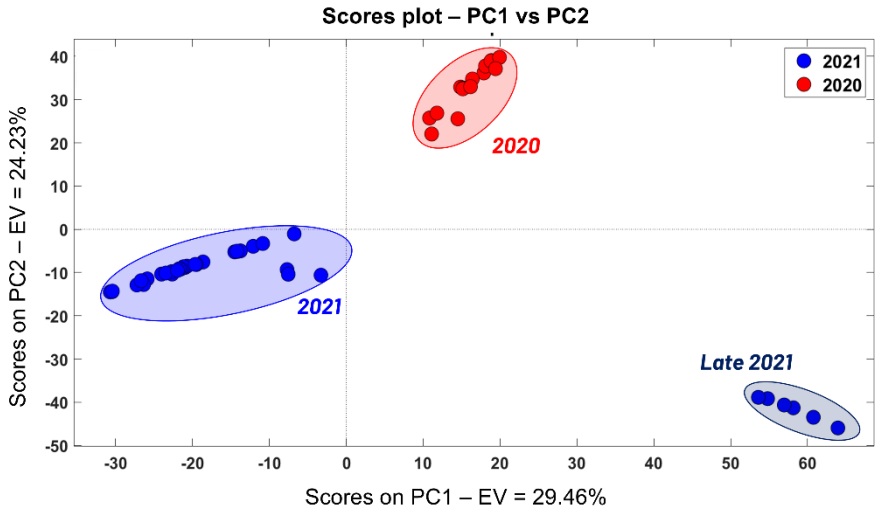

Figure 4.7. LC-MS dataset: scores plot of PC1 and PC2 with clear groupings related to the harvest year. Samples are coloured according to the harvest year.

As for ^{1}H-NMR, most of the information in the dataset consists of clusters related to the year of harvesting, while no grouping related to other features can be found. Accordingly, further models were developed for the two years separately to look for possible groupings based on geographical origin and cultivar. As a result, a series of clusters related to different geographical origins and/or cultivar combinations could be spotted. Once evidence related to these aspects was found, a supervised approach was developed using all the samples.

A PLS-DA model was built to distinguish Piedmont hazelnuts from samples with different geographical origins; this separation was confirmed by evaluating the model response plot (Figure 4.8). A second PLS-DA model then highlighted the possibility to discriminate the TGT hazelnuts from other cultivars as confirmed by the model response shown in Figure 4.9. These outcomes confirmed the possibility of finding information related to origin and cultivar using LC-MS analysis, even if

the largest part of the information in the obtained dataset separates the samples according to the harvest year.

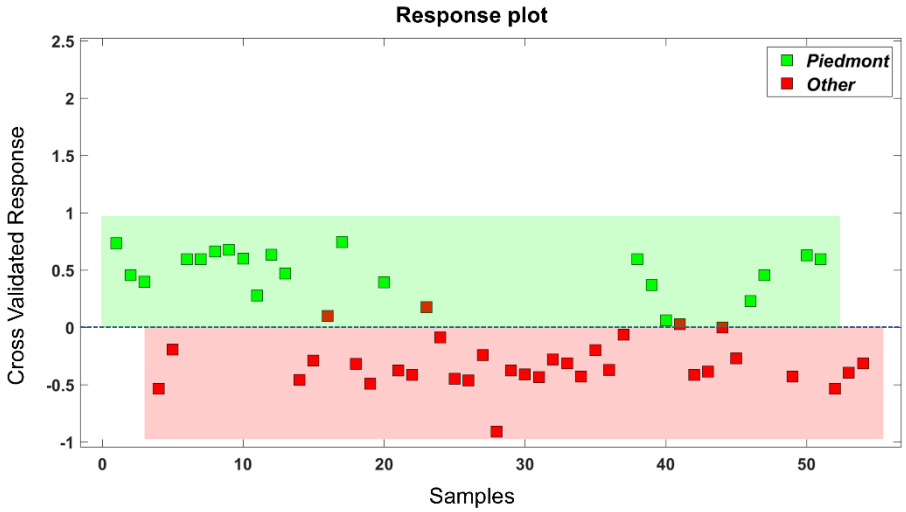

Figure 4.8. Response plot obtained from a PLS-DA model built using LC-MS data of the hazelnuts harvested in 2021. The two considered classes were "Piedmont" and "Other" to try to highlight the discrimination based on the geographical origin.

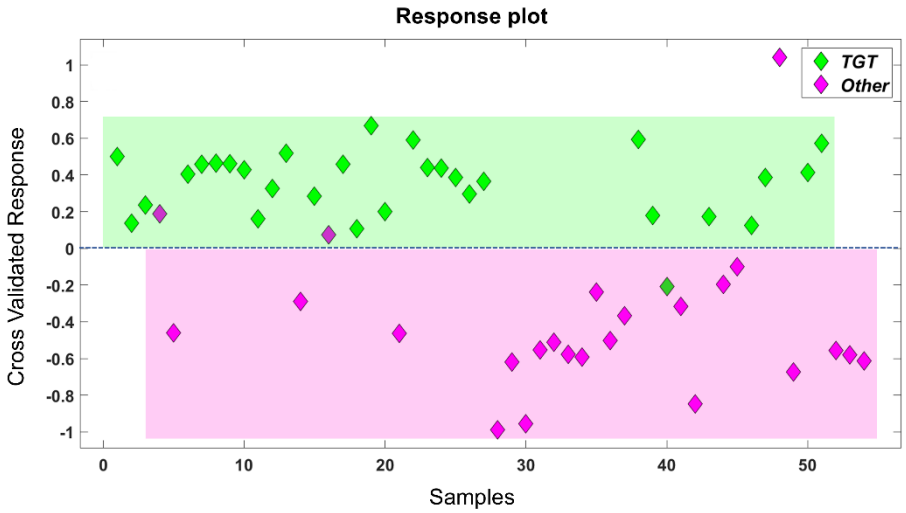

Figure 4.9. Response plot obtained from a PLS-DA model built using LC-MS data of the hazelnuts harvested in 2021. The two considered classes were "TGT" and "Other" to try to highlight the discrimination based on hazelnuts cultivar.

4.4 Data fusion approaches and compatibility between techniques

4.4.1 Data Fusion applied to Piedmont TGT hazelnuts case study

In this project, the ^{1}H-NMR and the LC-MS datasets were merged, and a PLS-DA classification model was chosen to evaluate the complementarity of the two techniques, and also to understand which variables better characterize the geographical origin and variety of hazelnuts.

A mid-level data fusion approach was applied to combine the two datasets, and it consisted of merging the scores extracted from two individual PCA models, built independently on the two datasets. From the NMR dataset the scores from the first 7 PCs were selected, while from the LC-MS dataset only the scores from the first 4 PCs were extracted. The choice of a mid-level approach was based on the different scales used to describe the variables in the two techniques, and to the non-comparable number of variables, around 26000 for NMR and around 2400 for LC-MS. The fused dataset contained only 11 variables with respect to the initial total of more than 28000. In addition, these 11 variables, representing the most relevant PCs' score values of both techniques, contained the 81.37 % and the 86.31 % of the total explained variance of the original NMR and LC-MS datasets, respectively.

All the statistical analyses performed on the fused dataset were developed using the PLS_Toolbox (version 8.9.2, Eigenvector Research Inc., Manson, WA, USA). To improve the comparability among the variables of the fused dataset, autoscaling was applied since these variables consist in PCA scores values.

4.4.2 Fused dataset: exploratory analysis

After the application of a mid-level data fusion approach, the obtained fused dataset was firstly explored using PCA and considering all the samples and all the 11 variables of the new datasets. Focusing on PC1 vs PC2 scores and loadings plots (Figure 4.10), a clear separation related to the different harvest times was spotted again in the scores plot. Looking at the loadings plot, a remarkable contribution to this separation seems to be related to PC1 (x %) from the NMR dataset and to PC1 (y %) and PC2 (z %) from the LC-MS dataset, meaning that the largest part of information contained in both the datasets allow to discriminate samples according to the harvest time.

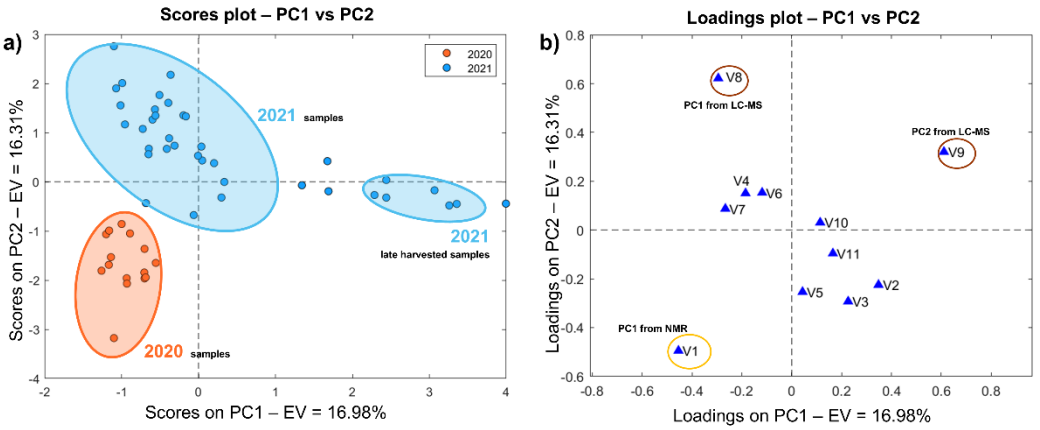

Figure 4.10. a) Scores plot of PC1 vs PC2 from a PCA model obtained from the fused dataset. Groupings based on harvested year are highlighted in orange (2020 samples) and blue (2021 samples). b) Loadings plot of PC1 vs PC2 from a PCA model obtained from the fused dataset. Important variables and NMR and LC-MS contributions are highlighted.

By focusing instead on PC2 vs PC3 and colouring the samples according to their geographical origin, in the scores plot (Figure 4.11a) two main groupings can be spotted, one containing hazelnut samples from Emilia Romagna and another one containing hazelnut samples from Piedmont. Speaking about variable contributions, by exploring the loadings plot (Figure 4.11b), the NMR dataset contributes to this separation with the first and the fifth principal components, while for the LC-MS dataset the first and the fourth principal components give a significative contribution.

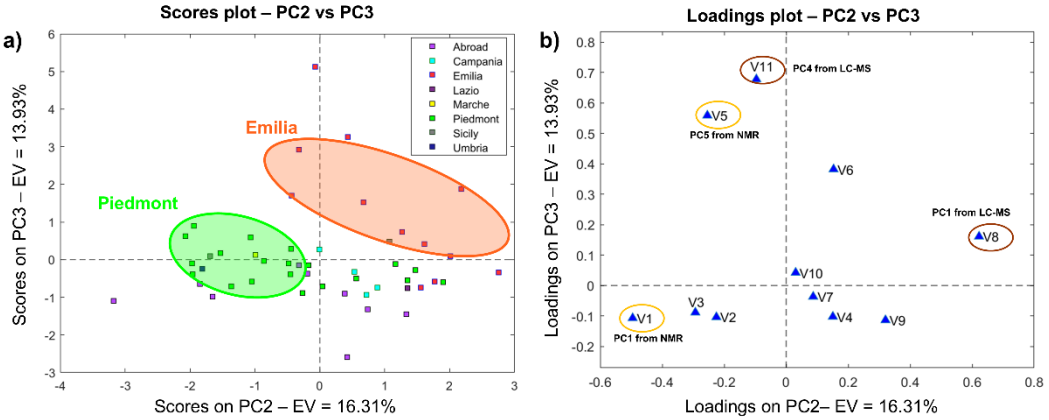

Figure 4.11. a) Scores plot of PC2 vs PC3 from a PCA model obtained from the fused dataset. Groupings based on geographical origin are highlighted in orange (samples from Emilia) and green (samples from Piedmont). b) Loadings plot of PC2 vs PC3 from a PCA model obtained from the fused dataset. Important variables and NMR and LC-MS contributions are highlighted.

A third group of clusters related to hazelnut variety was also spotted in the scores plot exploration by plotting PC1 and PC3 (Figure 4.12a) and colouring the samples according to their cultivar type. Two main clusters were spotted, both containing only samples of the TGT cultivar, the one on which this project is mainly focused. Interestingly, the smaller of the two clusters is composed only by the hazelnut late harvested in 2021, confirming again the strong effect related to the harvest time spotted also in the previously described analyses. The corresponding loadings plot (Figure 4.12b) revealed different contributions from PCs extracted from the NMR dataset and also contributions coming from PC1, PC2 and PC4 from the LC-MS dataset.

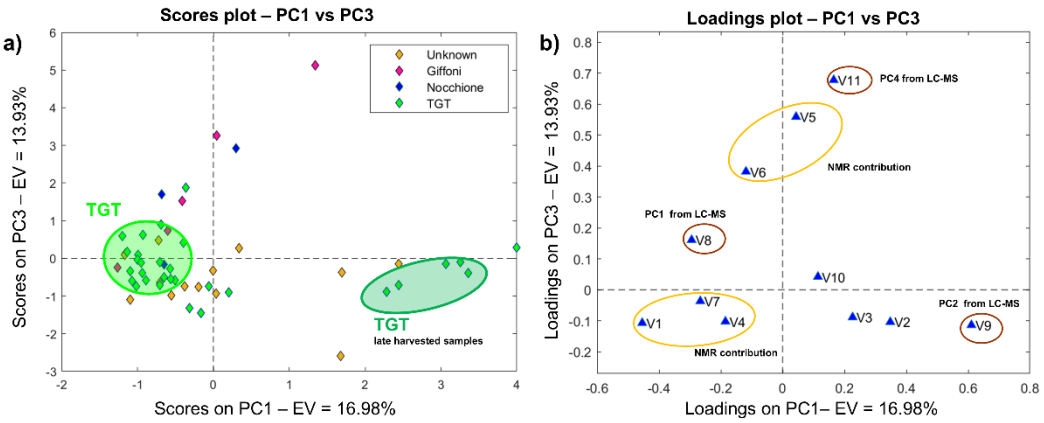

Figure 4.12. a) Scores plot of PC1 vs PC3 from a PCA model obtained from the fused dataset. Groupings based on TGT cultivar are highlighted in green. b) Loadings plot of PC1 vs PC3 from a PCA model obtained from the fused dataset. Important variables and NMR and LC-MS contributions are highlighted.

4.4.3 Fused dataset: classification analysis

Since the exploratory analysis confirmed the presence of information useful to discriminate the hazelnut samples according to their geographical origin, their cultivar type and their harvested year, a series of classification models were developed using PLS-DA on the fused dataset. In particular, three different PLS-DA models were developed: the first one with the aim of classifying the samples according to the harvest year; the second focused instead on geographical origin, developed with the aim of distinguishing Piedmont hazelnuts; and the last one focused on hazelnut variety, to try to distinguish the TGT cultivar hazelnuts among the other cultivars. Due to the low total number of samples available, also for the fused dataset all the classification models were validated using only cross validation. Unfortunately, the number of samples and the class distribution make this dataset not suitable for splitting it into calibration and test sets.

Before proceeding with the results obtained from these analyses, a short explanation about the outputs interpretation needs to be introduced. When multivariate classification models such as PLS-DA are built on chemical datasets (spectroscopic or chromatographic, for instance), the interpretation of the most relevant variables

is straightforward, as the actual modelled columns correspond to measurements with a physical or chemical meaning. With the application of mid-level data fusion, the modelled variables generally correspond to extracted features or "summary" variables, such as the scores of a principal component, therefore interpretation can be less straightforward. For example, for the loadings and coefficients plots (and all the model's outputs involving the variables), after identifying which principal component to inspect, a step back to the corresponding source PCA model (either NMR or LC-MS, in this case) must be done to inspect the relationship of that feature with the original variables.

For example, to clarify the concept, inspecting the coefficient plot of a PLS-DA model performed on a fused dataset obtained by fusing scores using a mid-level data fusion approach, an important contribution of PC1 of the first dataset was highlighted; to better understand the contribution of that PC1 to the system, their original scores and loadings plots need to be explored to evaluate which original variables contribute to the information described by that PC1.

Classification analysis based on harvest year: 2020 vs 2021

Starting now with the PLS-DA results evaluation, focusing on harvest year, two classes were considered: hazelnuts from 2020 and from 2021. Using PLS-DA, 1 latent variable was selected leading to a model with only one misclassified sample (Figure 4.13a) and describing 86.01 % of Y-block variance. This result is not unexpected since a clear separation between the two harvest years was already visible in PCA, so this result can also be taken as a confirmation of the successful transfer of information from the source dataset and the fused one. By exploring then the coefficients plot (Figure 4.13b), PC1 from NMR dataset and PC1 and PC2 from LC-MS dataset revealed to be particularly informative. These PCs were further explored to understand which variables mostly contribute to the variance they are describing.

By focusing on V1 from NMR dataset, its contribution was explored and, from the loadings plots, NMR signals belonging to glycerol and choline metabolites revealed to be characteristic for the samples of 2021 class, while signals belonging to glutamic acid and its derivates seems to be more associated with samples from 2020 class. Unfortunately, some potentially important peaks were not assigned to any metabolite yet, and so they are still unknown markers.

Moving now to the LC-MS data contribution, by inspecting V8 and V9 separately, the loadings plots revealed that metabolites like asparagynil glycine and dodecandioic acid are more present in samples of 2020 class, while geranoic acid and calystegine are more abundant in samples of 2021 class. Interestingly, some indole derivates seem to characterize the late harvested samples: this last outcome can be particularly useful to try to explain the reason why the harvest time has a so strong contribution even within the same harvesting year.

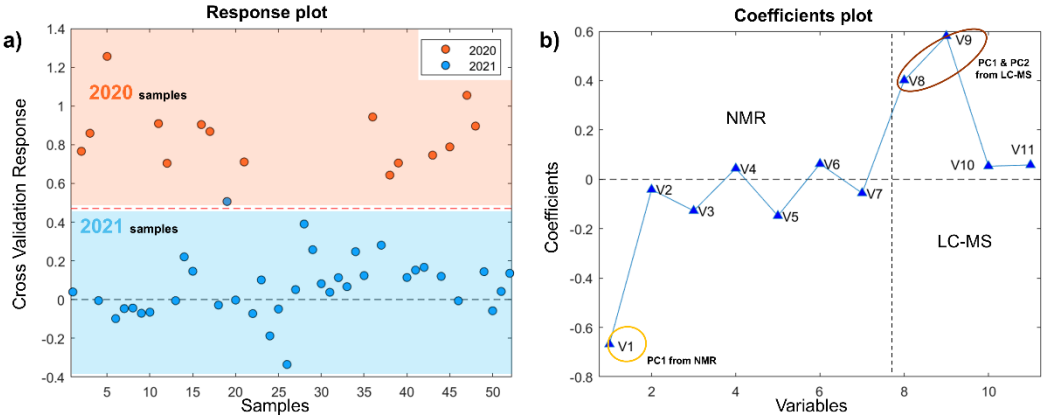

Figure 4.13. a) Response plot from a PLS-DA model obtained from the fused dataset considering two classes: 2020 and 2021 samples. b) Coefficients plot from a PLS-DA model obtained from the fused dataset considering two classes: 2020 and 2021 samples. Important variables from NMR and LC-MS datasets are highlighted.

Classification analysis based on geographical origin: Piedmont vs Other

Due to the functioning of the PLS-DA algorithm, which works better with few classes to be modelled, the model developed to classify hazelnuts according to their geographical origin was based on two main classes: hazelnuts from Piedmont and hazelnuts from other regions, both Italian and from other countries. The PLS-DA model was developed considering 1 latent variable, describing a Y-block variance of 35.98 % and allow to correctly classify 39 of the 52 samples included in this analysis. In particular, the confusion matrix reported in Table 4.3, revealed that 14 of the 20 Piedmont samples were correctly identified as Piedmont hazelnuts, while only 6 Piedmont samples were misclassified. The model overall error rate in cross-validation is of 25 % with a consequent accuracy of 75 %. With respect to the Piedmont class, a sensitivity of 0.70 and a specificity of 0.78 were obtained, confirming the previously observed PCA outcomes that these techniques allow to measure information related to geographical origin of hazelnuts samples (with a particular focus on those coming from Piedmont).

Table 4.3. Confusion matrix obtained from a PLS-DA model developed considering two classes: Piedmont hazelnuts and hazelnuts with other geographical origin.

Real / Predicted	**Piedmont**	**Other**
Piedmont	14	7
Other	6	25

By evaluating the response plot (Figure 4.14a), three of the six misclassified Piedmont hazelnuts (highlighted in green) are close to the class threshold, while other three appear to be not well-modelled. Speaking about hazelnuts marked as "*Other*", all the misclassified ones are close to the delimiter. In general, some samples are particularly far from their real class, and they probably need to be

further inspected to evaluate if they can be considered as outliers. From the coefficient plots evaluation (Figure 4.14b) different variables (i.e., the merged scores of different principal components) show a remarkable contribution. From the NMR dataset PC3, PC4 and PC7 are the most informative, speaking about LC-MS instead, PC1 and PC3 are the most informative. The exploration of the loadings plots of these original principal components, using the same approach employed for the previously described year-based PLS-DA model, highlighted that metabolites like glucopyranoside derivates, indole derivates and choline derivates mostly contribute to the correct classification of Piedmont hazelnuts, leading to become potentially interesting biomarkers for this class.

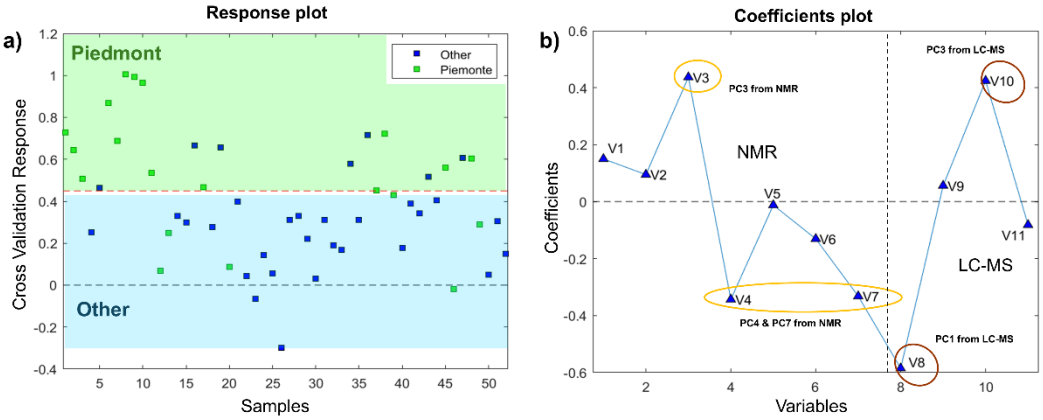

Figure 4.14. a) Response plot from a PLS-DA model obtained from the fused dataset considering two classes: Piedmont hazelnuts and hazelnuts with other geographical origin. b) Coefficients plot from a PLS-DA model obtained from the fused dataset considering two classes: Piedmont hazelnuts and hazelnuts with other geographical origin. Important variables from NMR and LC-MS datasets are highlighted.

Classification analysis based on hazelnut cultivar: TGT vs Other

Lastly, a third PLS-DA model was developed to distinguish among cultivar types. For this purpose, two classes were considered: TGT cultivar and cultivars of other types and so with different properties and characteristics. The model was developed using 1 latent variable, describing a Y-block variance of 36.21 % and allow to correctly classify 37 of the 52 hazelnut samples considered for this analysis, leading to an overall accuracy of 72 % and an error rate in cross validation of 28 %. According to the confusion matrix showed in Table 4.4, among the 30 samples of TGT cultivar, 7 were misclassified, resulting in a sensitivity of 0.77 and in a specificity of 0.64 for the TGT class. The results obtained from this classification models highlighted a larger difficulty in classifying samples according to their cultivar. The major effect previously observed for harvest time, and partially also for geographical origin, contribute to make this variety-based classification less precise. However, considering also the low amount of samples available and the explorative nature of this study, the two employed techniques confirmed the

possibility of using them to find information related to the cultivar type among different hazelnut samples (with a particular focus on TGT variety).

Table 4.4. Confusion matrix obtained from a PLS-DA model developed considering two classes: TGT hazelnuts and hazelnuts of other cultivars.

Real / Predicted	**TGT**	**Other**
TGT	23	8
Other	7	14

By exploring the graphical outputs, from the response plot (Figure 4.15a), considering the misclassified samples, many samples marked as "Other" are close to the threshold, while 4 TGT samples appear to be far from the class delimiter, meaning that a variety-based classification is probably more difficult than the previously explored origin-based discrimination. Although, a series of potential biomarkers for the TGT cultivar can be hypothesized by exploring the coefficients plot (Figure 4.15b). In this plot, a predominant effect of PC3 and PC5 from NMR dataset and PC3 and PC4 from LC-MS dataset was found. By evaluating loadings plots of these principal components, interesting metabolites like geranate, laurine derivates and adenosine-monopshosphate derivates are pointed out.

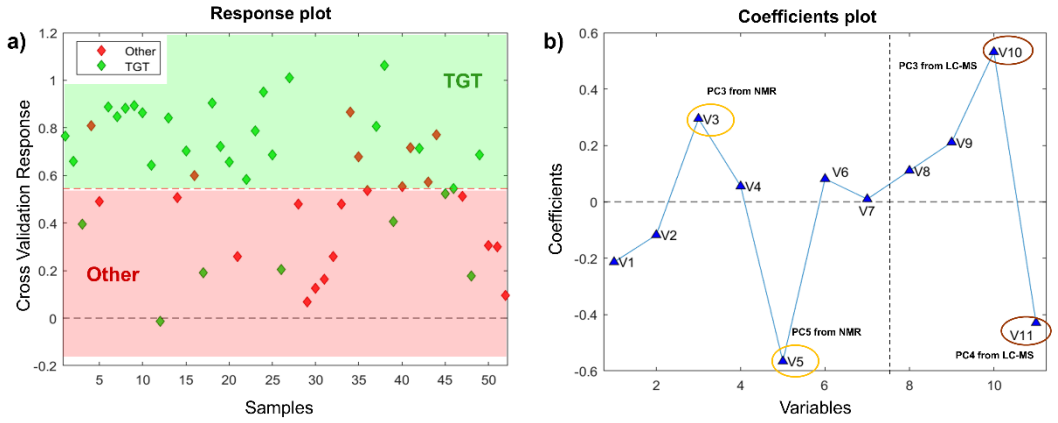

Figure 4.15. a) Response plot from a PLS-DA model obtained from the fused dataset considering two classes: TGT hazelnuts and hazelnuts of other cultivars. b) Coefficients plot from a PLS-DA model obtained from the fused dataset considering two classes: TGT hazelnuts and hazelnuts of other cultivars. Important variables from NMR and LC-MS datasets are highlighted.

4.5 Project conclusions and future perspectives

A multi-technique approach was used after the evaluation of each individual analytical technique, to highlight the advantages of a multi-omics approach in ensuring the integrity of food products using hazelnuts as a case study. To our knowledge, this approach, combining ^{1}H-NMR and LC-MS techniques, has never been applied to hazelnut samples before. Thus, the identification and

characterisation of hazelnuts potential biomarkers based on the metabolomics profile obtained from two different complementary techniques was taken advantage of to confirm the origins and the varieties of hazelnuts.

In more detail, this study focuses on hazelnuts of the Tonda Gentile Trilobata (TGT) cultivar grown in Piedmont (Italy) using two complementary analytical techniques and highlights the differences in chemical signatures when the effects of cultivar, origin, and harvest year overlap. Although the selected features were primarily affected by the harvest year, the results within each season were comparable. This result, being based on only two subsequent harvests, clearly cannot by generalized to the whole lifetime of a hazelnut grove, but it demonstrates that the multi-technique data-fusion approach can provide information even in the presence of such variability. The analysis of the individual datasets pointed out similar conclusions. In particular, both the techniques highlighted differences between the 2020 and the 2021 harvests and, though with a different level of confidence, were able to distinguish hazelnuts grown in Piedmont from those harvested outside the region and could also differentiate the TGT variety from other cultivars.

These results indicate that data fusion increases class separations with respect to individual results, despite the limitations in relation to robustness due to the small sample size. Nevertheless, the chemometric outputs obtained by the fusion of LC-MS and ^{1}H-NMR data allowed for the identification of the most relevant features distinguishing the geographical origin and the cultivar, even considering different harvest years of hazelnut samples. In conclusion, combining techniques with complementary information associated with proper data fusion approaches, can significantly improve the identification of the geographical origin and cultivar variety of hazelnuts.

A further step of this research could be focused on the creation of new classification models based only on the biomarkers assigned and selected from the variable interpretation performed on the fused dataset. In addition, to perform a deeper evaluation of the studied variables, variable selection approaches can be applied to the fused dataset. Variable selection approaches, like Variable Importance in Projection (VIP) scores evaluation or Selectivity Ratio (SR) [54], allow to evaluate variables contribution using *ad hoc* developed algorithms. This kind of approaches extract a subset containing a reduced number of variables with respect to the original one; these variables are selected thanks to their strong contribution to the discriminant ability of the model and so they can be considered more important than other variables for classification purposes.

Furthermore, the dataset could be further tested by adding samples (also from different origins and cultivars) to test the robustness of data interpretation in the presence of an increasing variability of the sample. This information can be of high interest when aiming at officialising these techniques for food security.

References | Chapter 4

[1] J. Van De Steene, J. Ruyssinck, J.A. Fernandez-Pierna, L. Vandermeersch, A. Maes, H. Van Langenhove, C. Walgraeve, K. Demeestere, B. De Meulenaer, L. Jacxsens, B. Miserez, Fingerprinting methods for origin and variety assessment of rice: development, validation and data fusion experiments, Food Control 151 (2023). https://doi.org/10.1016/j.foodcont.2023.109780.

[2] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283–293. https://doi.org/10.1016/j.foodcont.2017.11.034.

[3] S. Hassani, U. Dackermann, M. Mousavi, J. Li, A systematic review of data fusion techniques for optimized structural health monitoring, Information Fusion 103 (2024). https://doi.org/10.1016/j.inffus.2023.102136.

[4] X. Lin, S. Chao, D. Yan, L. Guo, Y. Liu, L. Li, Multi-Sensor Data Fusion Method Based on Self-Attention Mechanism, Applied Sciences (Switzerland) 13 (2023). https://doi.org/10.3390/app132111992.

[5] C. Zinnanti, E. Schimmenti, V. Borsellino, G. Paolini, S. Severini, Economic performance and risk of farming systems specialized in perennial crops: An analysis of Italian hazelnut production, Agric Syst 176 (2019). https://doi.org/10.1016/j.agsy.2019.102645.

[6] L. Lucini, G. Rocchetti, M. Trevisan, Extending the concept of terroir from grapes to other agricultural commodities: an overview, Curr Opin Food Sci 31 (2020) 88–95. https://doi.org/10.1016/j.cofs.2020.03.007.

[7] F. Ortega-Gavilán, S. Squara, C. Cordero, L. Cuadros-Rodríguez, M.G. Bagur-González, Application of chemometric tools combined with instrument-agnostic GC-fingerprinting for hazelnut quality assessment, Journal of Food Composition and Analysis 115 (2023). https://doi.org/10.1016/j.jfca.2022.104904.

[8] A.P. Sobolev, F. Thomas, J. Donarski, C. Ingallina, S. Circi, F. Cesare Marincola, D. Capitani, L. Mannina, Use of NMR applications to tackle future food fraud issues, Trends Food Sci Technol 91 (2019) 347–353. https://doi.org/10.1016/j.tifs.2019.07.035.

[9] B. Senizza, G. Rocchetti, S. Ghisoni, M. Busconi, M. De Los Mozos Pascual, J.A. Fernandez, L. Lucini, M. Trevisan, Identification of phenolic markers for saffron authenticity and origin: An untargeted metabolomics approach, Food Research International 126 (2019). https://doi.org/10.1016/j.foodres.2019.108584.

[10] Q. Qu, L. Jin, Application of nuclear magnetic resonance in food analysis, Food Science and Technology (Brazil) 42 (2022). https://doi.org/10.1590/fst.43622.

[11] A.H. Emwas, R. Roy, R.T. McKay, L. Tenori, E. Saccenti, G.A. Nagana Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, D.S. Wishart, Nmr spectroscopy for metabolomics research, Metabolites 9 (2019). https://doi.org/10.3390/metabo9070123.

[12] L. Mannina, A.P. Sobolev, S. Viel, Liquid state 1H high field NMR in food analysis, Prog Nucl Magn Reson Spectrosc 66 (2012) 1–39. https://doi.org/10.1016/j.pnmrs.2012.02.001.

[13] V. Maestrello, P. Solovyev, L. Bontempo, L. Mannina, F. Camin, Nuclear magnetic resonance spectroscopy in extra virgin olive oil authentication, Compr Rev Food Sci Food Saf 21 (2022) 4056–4075. https://doi.org/10.1111/1541-4337.13005.

[14] N. Cavallini, L. Strani, P.P. Becchi, V. Pizzamiglio, S. Michelini, F. Savorani, M. Cocchi, C. Durante, Tracing the identity of Parmigiano Reggiano “Prodotto di Montagna - Progetto Territorio” cheese using NMR spectroscopy and multivariate data analysis, Anal Chim Acta 1278 (2023). https://doi.org/10.1016/j.aca.2023.341761.

[15] A.P. Sobolev, A. Segre, R. Lamanna, Proton high-field NMR study of tomato juice, Magnetic Resonance in Chemistry 41 (2003) 237–245. https://doi.org/10.1002/mrc.1176.

[16] R. Consonni, L.R. Cagliani, Geographical characterization of polyfloral and acacia honeys by nuclear magnetic resonance and chemometrics, J Agric Food Chem 56 (2008) 6873–6880. https://doi.org/10.1021/jf801332r.

[17] M. Bosco, R. Toffanin, D. De Palo, L. Zatti, A. Segre, High-resolution 1H NMR investigation of coffee, J Sci Food Agric 79 (1999) 869–878. https://doi.org/10.1002/(SICI)1097-0010(19990501)79:6<869::AID-JSFA302>3.0.CO;2-6.

[18] M. Spraul, B. Schütz, P. Rinke, S. Koswig, E. Humpfer, H. Schäfer, M. Mörtter, F. Fang, U.C. Marx, A. Minoja, NMR-based multi parametric quality control of fruit juices: SGF profiling, Nutrients 1 (2009) 148–155. https://doi.org/10.3390/nu1020148.

[19] N. Cavallini, F. Savorani, R. Bro, M. Cocchi, A Metabolomic Approach to Beer Characterization, Molecules 26 (2021). https://doi.org/10.3390/molecules26051472.

[20] M. Amargianitaki, A. Spyros, NMR-based metabolomics in wine quality control and authentication, Chemical and Biological Technologies in Agriculture 4 (2017). https://doi.org/10.1186/s40538-017-0092-x.

[21] R. Consonni, L.R. Cagliani, The potentiality of NMR-based metabolomics in food science and food authentication assessment, Magnetic Resonance in Chemistry 57 (2019) 558–578. https://doi.org/10.1002/mrc.4807.

[22] A. Caligiani, J.D. Coisson, F. Travaglia, D. Acquotti, G. Palla, L. Palla, M. Arlorio, Application of 1H NMR for the characterisation and authentication of "Tonda Gentile Trilobata" hazelnuts from Piedmont (Italy), Food Chem 148 (2014) 77–85. https://doi.org/10.1016/j.foodchem.2013.10.001.

[23] R. Bachmann, S. Klockmann, J. Haerdter, M. Fischer, T. Hackl, 1H NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts, J Agric Food Chem 66 (2018) 11873–11879. https://doi.org/10.1021/acs.jafc.8b03724.

[24] R. Popescu, R.E. Ionete, O.R. Botoran, D. Costinel, F. Bucura, E.I. Geana, Y.F.J. Alabedallat, M. Botu, 1H-NMR profiling and carbon isotope discrimination as tools for the comparative assessment of walnut (Juglans regia L.) cultivars with various geographical and genetic origins-a preliminary study, Molecules 24 (2019). https://doi.org/10.3390/molecules24071378.

[25] P. Bambina, A. Spinella, G. Lo Papa, D.F. Chillura Martino, P. Lo Meo, L. Cinquanta, P. Conte, 1H-NMR Spectroscopy Coupled with Chemometrics to Classify Wines According to Different Grape Varieties and Different Terroirs, Agriculture (Switzerland) 14 (2024). https://doi.org/10.3390/agriculture14050749.

[26] O. Masetti, A. Sorbo, L. Nisini, Nmr tracing of food geographical origin: The impact of seasonality, cultivar and production year on data analysis, Separations 8 (2021). https://doi.org/10.3390/separations8120230.

[27] S. Li, Y. Tian, P. Jiang, Y. Lin, X. Liu, H. Yang, Recent advances in the application of metabolomics for food safety control and food quality analyses, Crit Rev Food Sci Nutr 61 (2021) 1448–1469. https://doi.org/10.1080/10408398.2020.1761287.

[28] M. Castro-Puyana, M. Herrero, Metabolomics approaches based on mass spectrometry for food safety, quality and traceability, TrAC - Trends in Analytical Chemistry 52 (2013) 74–87. https://doi.org/10.1016/j.trac.2013.05.016.

[29] S. Ghisoni, L. Lucini, F. Angilletta, G. Rocchetti, D. Farinelli, S. Tombesi, M. Trevisan, Discrimination of extra-virgin-olive oils from different cultivars and geographical origins by untargeted metabolomics, Food Research International 121 (2019) 746–753. https://doi.org/10.1016/j.foodres.2018.12.052.

[30] M. Ben Mohamed, G. Rocchetti, D. Montesano, S. Ben Ali, F. Guasmi, N. Grati-Kamoun, L. Lucini, Discrimination of Tunisian and Italian extra-virgin olive oils according to their phenolic and sterolic fingerprints, Food Research International 106 (2018) 920–927. https://doi.org/10.1016/j.foodres.2018.02.010.

[31] P. Bellassi, G. Rocchetti, M. Nocetti, L. Lucini, F. Masoero, L. Morelli, A combined metabolomic and metagenomic approach to discriminate raw milk for the production of hard cheese, Foods 10 (2021). https://doi.org/10.3390/foods10010109.

[32] G. Rocchetti, L. Lucini, A. Gallo, F. Masoero, M. Trevisan, G. Giuberti, Untargeted metabolomics reveals differences in chemical fingerprints between PDO and non-PDO Grana Padano cheeses, Food Research International 113 (2018) 407–413. https://doi.org/10.1016/j.foodres.2018.07.029.

[33] B. Senizza, P. Ganugi, M. Trevisan, L. Lucini, Combining untargeted profiling of phenolics and sterols, supervised multivariate class modelling and artificial neural networks for the origin and

authenticity of extra-virgin olive oil: A case study on Taggiasca Ligure, Food Chem 404 (2023). https://doi.org/10.1016/j.foodchem.2022.134543.

[34] R. Xiao, Y. Ma, D. Zhang, L. Qian, Discrimination of conventional and organic rice using untargeted LC-MS-based metabolomics, J Cereal Sci 82 (2018) 73–81. https://doi.org/10.1016/j.jcs.2018.05.012.

[35] C.L. Dittgen, J.F. Hoffmann, F.C. Chaves, C.V. Rombaldi, J.M.C. Filho, N.L. Vanier, Discrimination of genotype and geographical origin of black rice grown in Brazil by LC-MS analysis of phenolics, Food Chem 288 (2019) 297–305. https://doi.org/10.1016/j.foodchem.2019.03.006.

[36] S. Ghisoni, L. Lucini, G. Rocchetti, G. Chiodelli, D. Farinelli, S. Tombesi, M. Trevisan, Untargeted metabolomics with multivariate analysis to discriminate hazelnut (Corylus avellana L.) cultivars and their geographical origin, J Sci Food Agric 100 (2020) 500–508. https://doi.org/10.1002/jsfa.9998.

[37] D. Schütz, E. Achten, M. Creydt, J. Riedl, M. Fischer, Non-targeted LC-MS metabolomics approach towards an authentication of the geographical origin of grain maize (Zea mays L.) samples, Foods 10 (2021). https://doi.org/10.3390/foods10092160.

[38] V. Lelli, R. Molinari, N. Merendino, A.M. Timperio, Detection and comparison of bioactive compounds in different extracts of two hazelnut skin varieties, tonda gentile romana and tonda di giffoni, using a metabolomics approach, Metabolites 11 (2021). https://doi.org/10.3390/metabo11050296.

[39] R.M. Boiteau, D.W. Hoyt, C.D. Nicora, H.A. Kinmonth-Schultz, J.K. Ward, K. Bingol, Structure elucidation of unknown metabolites in metabolomics by combined NMR and MS/MS prediction, Metabolites 8 (2018). https://doi.org/10.3390/metabo8010008.

[40] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[41] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, Analytical Methods 5 (2013) 3790–3798. https://doi.org/10.1039/c3ay40582f.

[42] Y. Hong, N. Birse, B. Quinn, Y. Li, W. Jia, P. McCarron, D. Wu, G.R. da Silva, L. Vanhaecke, S. van Ruth, C.T. Elliott, Data fusion and multivariate analysis for food authenticity analysis, Nat Commun 14 (2023). https://doi.org/10.1038/s41467-023-38382-z.

[43] G. Zheng, W.S. Price, Solvent signal suppression in NMR, Prog Nucl Magn Reson Spectrosc 56 (2010) 267–288. https://doi.org/10.1016/J.PNMRS.2010.01.001.

[44] F. Savorani, G. Tomasi, S. Engelsen, Alignment of 1D NMR Data using the iCoshift Tool: A Tutorial, in: 2013: pp. 14–24. https://doi.org/10.1039/9781849737531-00014.

[45] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, Journal of Magnetic Resonance 202 (2010) 190–202. https://doi.org/10.1016/j.jmr.2009.11.012.

[46] A.C. Dona, M. Kyriakides, F. Scott, E.A. Shephard, D. Varshavi, K. Veselkov, J.R. Everett, A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments, Comput Struct Biotechnol J 14 (2016) 135–153. https://doi.org/10.1016/J.CSBJ.2016.02.005.

[47] C. Schmitt, T. Schneider, L. Rumask, M. Fischer, T. Hackl, Food Profiling: Determination of the Geographical Origin of Walnuts by 1H NMR Spectroscopy Using the Polar Extract, J Agric Food Chem 68 (2020) 15526–15534. https://doi.org/10.1021/acs.jafc.0c05827.

[48] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, HMDB 3.0—The Human Metabolome Database in 2013, Nucleic Acids Res 41 (2013) D801–D807. https://doi.org/10.1093/NAR/GKS1065.

[49] G. Rocchetti, G. Chiodelli, G. Giuberti, F. Masoero, M. Trevisan, L. Lucini, Evaluation of phenolic profile and antioxidant capacity in gluten-free flours, Food Chem 228 (2017) 367–373. https://doi.org/10.1016/j.foodchem.2017.01.142.

[50] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, n.d.

[51] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[52] D. Ballabio, A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure, Chemometrics and Intelligent Laboratory Systems 149 (2015) 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003.

[53] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, Analytical Methods 5 (2013) 3790–3798. https://doi.org/10.1039/c3ay40582f.

[54] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J Chemom 29 (2015) 528–536. https://doi.org/10.1002/CEM.2736.

Chapter 5

Case study #3: challenges in wine analysis using NMR

5.1 Project overview and aim of the work

Wine is a product of remarkable complexity, influenced by a multitude of factors including grape variety, terroir, climate, and vinification processes. These elements contribute to its unique chemical composition, making it an ideal subject for metabolomic studies. Several analytical techniques are routinely employed for the quantification of chemical compounds in foods, from cumbersome classical wet chemistry methods, over chromatographic methods, to "green" spectroscopic techniques, such as infrared (IR) and near-infrared (NIR) spectroscopies [1–3]. Proton Nuclear Magnetic Resonance (^{1}H-NMR) spectroscopy is an inherently quantitative method, which is routinely used for the analysis of diverse biological samples in a wide range of research and industrial fields, including food quality control. The collected data holds the potential to explore topics such as traceability, impact of geographical origin, and varietal differentiation; but also topics more related to the behaviour and the nature of ^{1}H-NMR spectroscopy in wine analysis. For example, the NMR signal intensity is directly proportional to the number of protons giving rise to that signal, and absolute quantification by ^{1}H-NMR can thus be achieved using an internal/external reference compound of known concentration or using a calibrated artificial signal [4–7]. When performing absolute quantitative ^{1}H-NMR, parameters optimization, including calibration of the 90° pulse (p1) and recycle delay (d1), are essential to obtain reliable quantifications [8,9]. However, despite careful instrumental calibration, absolute quantification of metabolites in complex mixtures characterized by high concentration solvents still represents a challenge for ^{1}H-NMR spectroscopy due to physical phenomena, including radiation damping (RD) [10].

In this context, two main parallel projects were developed, one more focused on the previously defined *metabolomics topics* like traceability, identification and quality assessment; and another instead more technical and focused on the evaluation of factors that can affect the quantification ability of the NMR spectroscopy. Due to the complexity of the developed project, and due to the large amount of data to be analysed, the metabolomics part of the project is still ongoing. However, since the data acquisition and exploration phases were of fundamental importance to give rise to the second parallel project, the wine sample collection and data acquisition

phase is reported in Section 5.2, while the rest of the chapter will be more focused on the evaluation of the different parameters that are involved in the quantification ability of proton NMR spectroscopy in a very complex natural mixture like wine.

5.2 Sample collection and data acquisition: ^{1}H-NMR

A total of 233 table wines (225 red wines and 8 white wines) was selected and purchased from local shops in Denmark to be included in the analysis. The wine sample set was selected to include wines from different grapes, geographical origin, processing, and vintage. The selected wines span an ethanol range of 12–18 %. An overview of the wines included in the method validation experiment is given in Table 5.1.

The wine samples were prepared according to the method proposed by Aru *et al.* [11]. Briefly, from each bottle 2 mL of wine were withdrawn through the cork cap using a syringe-based system replacing the wine headspace with argon (Coravin). For each sample, two replicates of a solution containing 700 µL of wine and 300 µL of 1M KH_2PO_4 buffer in D_2O (4:1 v/v, pH = 3.50 ± 0.02) were prepared. Subsequently, 600 µL of the solution were transferred into a 5 mm (O.D.) SampleJet NMR tubes (Bruker BioSpin, Ettlingen, Germany).

Deuterium oxide (D_2O, 99.9 %), potassium phosphate monobasic (KH_2PO_4) and sodium 3-trimethylsilyl-propionate-2,2,3,3-d4 (TSP) were purchased from Sigma-Aldrich (Darmstadt, Germany). The water used throughout the study was purified using a Millipore lab water system (Merck KGaA, Darmstadt, Germany) equipped with a 0.22 µm filter membrane.

^{1}H-NMR spectra were recorded on a Bruker Avance III 600 operating at a proton Larmor's frequency of 600.13 MHz and equipped with a 5-mm broadband inverse (BBI) probe. Data acquisition and processing were carried out using the TopSpin software (version 4.2, Bruker, Rheinstetten, Germany). After temperature equilibration (5 min) the ^{1}H-NMR spectra were measured at 298 K using two different experiments: one with a standard pulse sequence for presaturation of the water signal (zgcppr pulse program, Bruker nomenclature), and one experiment without the solvent suppression procedure (zg pulse program, Bruker nomenclature). The sweep width was of 12019 Hz (20 ppm), a 90° pulse, and an acquisition time of 3 s. The relaxation delay was set to 40 s. Spectral data were collected into 64 k data points, after 32 scans. The receiver gain was fixed for all the experiments at 0.25, according to a standard operating procedure (Aru et al., 2018). All spectra were acquired in automation using iconNMR™ (Bruker Biospin, Rheinstetten, Germany) and the SampleJet™ system (Bruker BioSpin, Ettlingen, Germany). Phase and baseline correction were performed in the TopSpin software. All analyses of spectral data were carried out using homemade scripts in MATLAB (R2021b, Mathworks, Natick, MA, USA). First, for all the acquired spectra (both

ethanol solutions and wines) a signal alignment preprocessing method was applied using the *i*coshift tool. This method allows to align the NMR signals in manually selected intervals in order to facilitate comparison among different samples. In our study, the selected intervals were the regions with the most intense signals corresponding to ethanol, and to the TSP signal used to set the chemical shift scale at 0.0 ppm.

Table 5.1. Overview of the wine samples included in the study. Samples marked with # are the one used for the case study.

ID	Wine Type	Origin	Variety	Vintage	EtOH %
1	Red	USA	Pinot nero	2019	13.5
2	Red	USA	Pinot nero	2020	13.5
3	Red	Italy	*Blend**	2018	15.0
4	Red	Italy	*Blend**	2020	15.0
5	Red	Italy	Corvina	2017	15.0
6	Red	Italy	Corvina	2018	15.0
7	Red	Italy	Sangiovese	2017	14.0
8	Red	Italy	Sangiovese	2019	14.0
9	Red	Italy	*Blend**	2016	14.0
10	Red	Italy	Sangiovese	2016	14.0
11	Red	Italy	*Blend**	2017	14.0
12	Red	Australia	*Blend**	2016	14.0
13	Red	Italy	Sangiovese	2017	14.5
14	Red	Italy	*Blend**	2019	14.0
15	Red	USA	Pinot nero	/	13.5
16	Red	USA	Primitivo	2014	14.5
17	Red	Italy	Aglianico	2017	14.5
18	Red	Italy	Montepulciano	2020	14.0
19	Red	Italy	Primitivo	2019	14.5
20	Red	Italy	Primitivo	2020	14.5
21	Red	Italy	Primitivo	2018	16.0
22	Red	France	Pinot nero	2020	13.0
23	Red	France	Merlot	2018	15.0
24	Red	France	*Blend**	2018	14.5
25	Red	Spain	Tempranillo	2018	14.0
26	Red	Italy	Corvina	2018	14.0
27	Red	USA	Pinot nero	2020	13.5
28	Red	Australia	Merlot	2019	14.0
29	Red	France	*Blend**	2020	15.0
30	Red	Italy	Merlot	2020	14.0
31	Red	Australia	Syrah	2018	14.5
32	Red	Italy	*Blend**	2018	14.5
33	Red	Italy	Negroamaro	2018	14.0
34	Red	Spain	Tempranillo	2012	14.5
35	Red	Italy	Primitivo	2017	13.5
36	Red	Italy	Montepulciano	2018	13.5
37	Red	Spain	Mataro	2020	14.5
38	Red	Italy	*Blend**	2018	13.5
39	Red	Italy	*Blend**	2017	13.5

40 #	White	Italy	*Blend**	2020	12.5
41	Red	Spain	Tempranillo	2018	14.5
42	Red	Italy	*Blend**	2020	13.5
43	Red	France	*Blend**	2019	16.0
44	Red	Italy	Primitivo	2020	13.5
45 #	White	Chile	Chardonnay	2020	13.5
46	Red	Spain	Mencia	2019	14.5
47	Red	Spain	*Blend**	2015	15.0
48	Red	France	*Blend**	2019	14.5
49	Red	France	*Blend**	2012	14.5
50	Red	USA	Pinot nero	2018	13.5
51	Red	Italy	Negroamaro	2021	14.0
52	Red	France	Merlot	2018	12.5
53	Red	Italy	*Blend**	2018	13.5
54	Red	Spain	Tempranillo	2017	15.0
55	Red	Italy	Primitivo	2020	14.0
56	Red	Italy	Sangiovese	2017	14.0
57	Red	France	*Blend**	2021	14.0
58	Red	Spain	*Blend**	2015	14.5
59	Red	Spain	Tempranillo	2020	14.0
60	Red	France	Pinot nero	2020	13.0
61	Red	South Africa	*Blend**	2019	14.5
62	Red	Australia	*Blend**	2019	14.5
63	Red	Italy	*Blend**	2020	14.0
64	Red	Italy	Sangiovese	2018	14.0
65	Red	Italy	*Blend**	2017	13.5
66	Red	USA	Pinot nero	2019	13.5
67	Red	Spain	*Blend**	2019	14.0
68	Red	Spain	Tinta de Toro	2012	15.5
69	Red	Spain	*Blend**	/	14.0
70	Red	Spain	*Blend**	2018	14.5
71	Red	Italy	Sangiovese	2017	14.0
72	Red	Spain	Cannonau	2018	14.5
73	Red	Italy	*Blend**	/	14.5
74	Red	France	*Blend**	/	13.5
75	Red	Italy	Nebbiolo	2015	14.5
76	Red	Spain	Tempranillo	2020	14.5
77	Red	Spain	Tempranillo	2017	15.0
78	Red	France	*Blend**	2017	15.5
79	Red	Spain	Tempranillo	2018	14.0
80	Red	Spain	Tempranillo	2014	14.0
81	Red	Italy	Nero d'Avola	2020	14.5
82	Red	Italy	Nero d'Avola	2019	14.5
83	Red	France	*Blend**	2010	15.0
84	Red	France	*Blend**	2016	14.0
85	Red	Portugal	Tempranillo	2020	14.0
86	Red	France	*Blend**	2017	13.5
87	Red	Italy	Sangiovese	2017	13.5
88	Red	Spain	Tempranillo	2019	13.0
89	Red	Italy	*Blend**	2018	14.5
90	Red	Spain	*Blend**	2016	14.0
91	Red	France	*Blend**	2018	13.0
92	Red	France	*Blend**	2019	13.5
93	Red	Italy	*Blend**	2018	14.0

94	Red	France	*Blend**	2019	14.0
95	Red	Chile	*Blend**	2018	14.5
96	Red	Italy	Primitivo	2020	14.5
97	Red	USA	Primitivo	2019	14.5
98	Red	Italy	Negroamaro	2020	14.5
99	Red	Italy	Primitivo	2018	14.0
100	Red	Italy	Sangiovese	2018	13.5
101	Red	Italy	Barbera	2019	14.5
102	Red	Italy	Syrah	2017	14.5
103	Red	Italy	Merlot	2020	14.0
104	Red	Italy	Sangiovese	2017	14.5
105	Red	Italy	*Blend**	2018	14.5
106	Red	Italy	*Blend**	/	14.0
107	Red	Italy	*Blend**	2018	13.5
108	Red	Italy	Primitivo	2018	14.5
109	Red	Spain	Tempranillo	2017	15.0
110	Red	Italy	Brachetto	2018	6.5
111	Red	Italy	Primitivo	2019	14.5
112	Red	Spain	*Blend**	2016	13.5
113	Red	Spain	*Blend**	2017	13.5
114	Red	Spain	*Blend**	/	14.0
115	Red	Italy	*Blend**	2018	14.0
116	Red	France	*Blend**	2010	13.0
117 #	White	Italy	Pecorino	2021	13.5
118	Red	Spain	*Blend**	/	13.0
119	Red	USA	*Blend**	2020	12.5
120	Red	Italy	Primitivo	2020	14.0
121	Red	France	*Blend**	2016	13.0
122	Red	Spain	*Blend**	2019	15.0
123	Red	Italy	*Blend**	2013	14.0
124	Red	Italy	Sangiovese	2020	15.0
125	Red	Italy	*Blend**	2018	13.5
126	Red	Italy	Corvina	2015	16.5
127	Red	Italy	*Blend**	/	14.0
128	Red	Italy	Primitivo	2016	14.5
129	Red	Italy	Aglianico	2016	14.5
130	Red	Italy	Primitivo	2019	17.0
131	Red	Italy	Sangiovese	2019	14.0
132	Red	Italy	Primitivo	2015	14.5
133	Red	Italy	Primitivo	2019	14.0
134	Red	Italy	*Blend**	2016	13.5
135	Red	Italy	*Blend**	2018	14.5
136	Red	Portugal	Touriga	2020	13.5
137	Red	Portugal	Touriga	2020	13.5
138	Red	France	*Blend**	2018	14.5
139	Red	USA	Cabernet S.	2018	14.5
140	Red	France	*Blend**	2020	14.5
141	Red	France	*Blend**	2018	14.0
142	Red	France	*Blend**	2018	14.0
143	Red	France	*Blend**	2019	14.0
144	Red	France	*Blend**	2018	14.0
145	Red	Spain	*Blend**	/	14.0
146	Red	France	Cannonau	2019	15.0
147	Red	Italy	*Blend**	2020	16.0

148	Red	Australia	Syrah	2019	14.5
149	Red	Spain	*Blend**	/	14.0
150	Red	Italy	Primitivo	2021	13.0
151	Red	France	*Blend**	2016	13.0
152	Red	France	*Blend**	2018	13.0
153	Red	France	*Blend**	2019	14.5
154	Red	France	*Blend**	2015	14.0
155	Red	France	*Blend**	2018	14.0
156	Red	France	*Blend**	2016	14.5
157	Red	Italy	Negroamaro	2020	15.0
158	Red	Italy	Montepulciano	/	14.5
159	Red	Italy	Primitivo	2020	14.0
160	Red	Italy	Sangiovese	2018	13.0
161	Red	Spain	Tempranillo	2019	14.0
162	Red	Spain	*Blend**	2016	15.0
163	Red	France	*Blend**	2018	14.5
164	Red	France	*Blend**	2016	13.0
165	Red	Spain	*Blend**	2016	13.5
166	Red	Spain	*Blend**	2017	14.0
167	Red	Spain	Tempranillo	2010	14.0
168	Red	France	*Blend**	2019	14.5
169	Red	Portugal	Castelao	2020	15.0
170	Red	Italy	*Blend**	2015	15.0
171	Red	Italy	Sangiovese	2016	15.0
172	Red	France	*Blend**	2010	13.0
173	Red	USA	Primitivo	2021	13.5
174	Red	Spain	Syrah	2021	14.0
175	Red	France	*Blend**	2018	13.5
176	Red	France	Merlot	2016	13.5
177	Red	France	*Blend**	2020	14.0
178	Red	France	*Blend**	2018	14.0
179	Red	France	*Blend**	2015	14.0
180	Red	France	*Blend**	2019	13.5
181 #	Red	France	*Blend**	2021	14.0
182 #	Red	France	*Blend**	2020	15.0
183 #	Red	Argentina	*Blend**	2015	16.0
184 #	Red	France	*Blend**	2010	13.0
185 #	Red	France	*Blend**	2015	15.0
186 #	Red	France	Merlot	2015	14.5
187 #	Red	Spain	*Blend**	2019	15.0
188 #	Red	France	*Blend**	2018	14.0
189 #	Red	Spain	Tempranillo	2018	14.0
190 #	Red	Italy	Primitivo	2018	16.0
191 #	Red	Italy	Primitivo	2017	13.5
192 #	Red	Spain	Tempranillo	2014	14.0
193	Red	Italy	Primitivo	2017	14.5
194	Red	Italy	*Blend**	2019	13.5
195	Red	Portugal	*Blend**	2015	14.5
196	Red	Australia	Syrah	2020	14.5
197	Red	Australia	Syrah	2019	14.0
198	Red	Australia	Syrah	2018	14.0
199	Red	Italy	Barbera	2019	14.5
200	Red	Italy	*Blend**	2016	13.5
201	Red	Italy	Dolcetto	2021	12.5

202	Red	Italy	Barbera	2020	12.5
203	Red	USA	Primitivo	2015	14.0
204	Red	Argentina	Malbec	2015	14.5
205	Red	USA	Pinot nero	2019	13.5
206	Red	Italy	Bonarda	2019	18.0
207	Red	Argentina	Pinot nero	2020	14.3
208	Red	Italy	Sangiovese	2012	14.5
209	Red	France	Cabernet S.	2016	13.5
210	Red	Italy	Nebbiolo	2015	14.5
211	Red	France	*Blend**	2017	15.0
212	Red	Spain	*Blend**	2019	14.0
213	Red	Italy	*Blend**	2020	13.5
214	Red	France	Cannonau	2017	15.0
215	Red	Argentina	Cabernet Franc	2018	15.0
216	Red	Italy	Primitivo	2021	14.0
217	Red	USA	Primitivo	2017	14.5
218	Red	Italy	Primitivo	2019	14.0
219	Red	USA	Primitivo	2014	14.5
220	Red	Italy	Primitivo	2020	14.0
221	Red	Italy	*Blend**	2019	14.0
222	Red	Italy	*Blend**	2017	14.0
223	Red	Spain	Mataro	2020	15.0
224	Red	France	*Blend**	2020	14.5
225	Red	Italy	*Blend**	2020	14.5
226	Red	France	*Blend**	2019	14.0
227 #	White	France	Sauvignon	2020	12.5
228	Red	Australia	*Blend**	2020	13.5
229	Red	Denmark	Pinot Noir	2022	12.0
230 #	White	Denmark	Solaris	2022	12.0
231 #	White	Poland	Solaris	2021	13.0
232 #	White	Denmark	Solaris	2022	12.0
233 #	White	Denmark	Solaris	2022	13.0

**Blend details not included in the tables due to lack of space*

5.3 The problem of measuring ethanol content with ^{1}H-NMR

The absolute quantification of ethanol in alcoholic beverages can be measured with high accuracy by Fourier-Transform Infra-Red (FT- IR) spectroscopy, with prediction errors in the order of ±0.01 %(v/v) [12] despite overlapped signals. Together with water, ethanol is a main component in a wide range of beverages, including beers, spirits and wines. Such beverages are characterized by a very complex and diverse chemical composition with a high dynamic range spanning concentrations from 0.1 to 500 mg/L [13].

Recent applications include quality control [14], authentication/fraud [15],[14], traceability [16], but also process monitoring such as fermentation and aging effects [17]. Despite the numerous and obvious analytical advantages, to the best of our

knowledge, only few studies have focused on the possibility of using ^{1}H-NMR for the absolute quantification of high concentration solvents, such as ethanol [18,19]. Lopez *et al.* developed a qNMR method for ethanol quantification in wines using its main and satellite resonances [19]. Method validation was based on the approved reference method namely distillation, and by a prediction method based on FT-IR (WineScan). In contrast, Caleja-Ballesteros *et al.* focused on the development of a novel qNMR method for ethanol quantification in distilled spirits using non-deuterated solvents [18]. In this case, method validation was performed by gas chromatography with flame ionization.

In this part of our study, we propose a deep evaluation of different parameters that can potentially affect the ethanol absolute quantification in alcoholic beverages obtained using ^{1}H-NMR spectroscopy. Our study consists of two parts: first, the investigation of the impact of different parameters (i.e., pulse sequence, d-solvent, NMR signal and quantification method) on the quantification of ethanol by measuring 10 pure ethanol solutions at different concentrations (from 0 to 40 % v/v); second, the application and evaluation of the best operating conditions on a real case study based on wine samples (taken from the list of 233 wine spectra previously acquired).

5.3.1 Experimental design

The experiments of the first part were planned according to a Design of Experiment (DoE) approach based on four factors, each with two levels [20]. To test the quantification capability of ^{1}H-NMR spectroscopy with respect to increasing ethanol concentrations, ten samples with ethanol percentages ranging from 0 to 40 % (v/v) were prepared in triplicates (Figure 5.1a). Samples for ^{1}H-NMR analysis were prepared using D_2O and DMSO as deuterated solvents, giving a total of 60 samples (10 samples × 3 replicates × 2 d-solvents). The ethanol solutions were then analysed by ^{1}H-NMR spectroscopy following two experimental setups: in the first, water suppression was employed, while in the second set up samples were measured with no water suppression. A total of 120 spectra of pure ethanol solutions were recorded. Ethanol concentrations were calculated using two different quantification methods ("Raw Sum" vs. MCR [21,22]), leading to a total of 240 quantifications for each signal monitored in the ^{1}H-NMR spectra. According to our preliminary results, two different NMR signals were tested (i.e., ethanol triplet signal and ethanol triplet right ^{13}C satellite signal) leading to a total of 480 quantifications.

For the second part, a total of 20 table wines (12 red wines and 8 white wines) were selected from the previously analysed 233 table wines to be included in the analysis. The wine sample set was defined to include wines from different grapes, geographical origin, processing, and vintage. An overview of the wines included in the method validation experiment is given in Table 5.1 (wines highlighted with #). The ^{1}H-NMR spectra acquisition parameters of the wines are summarized in

Section 5.2 and they are exactly the same used for the ethanol-in-water solutions. According to our results, ethanol NMR quantifications were performed on the ethanol triplet signal using the Raw Sum approach. To evaluate the NMR quantification ability, the WineScan IR instrument was used as a reference for the ethanol measurements (Figure 5.1b).

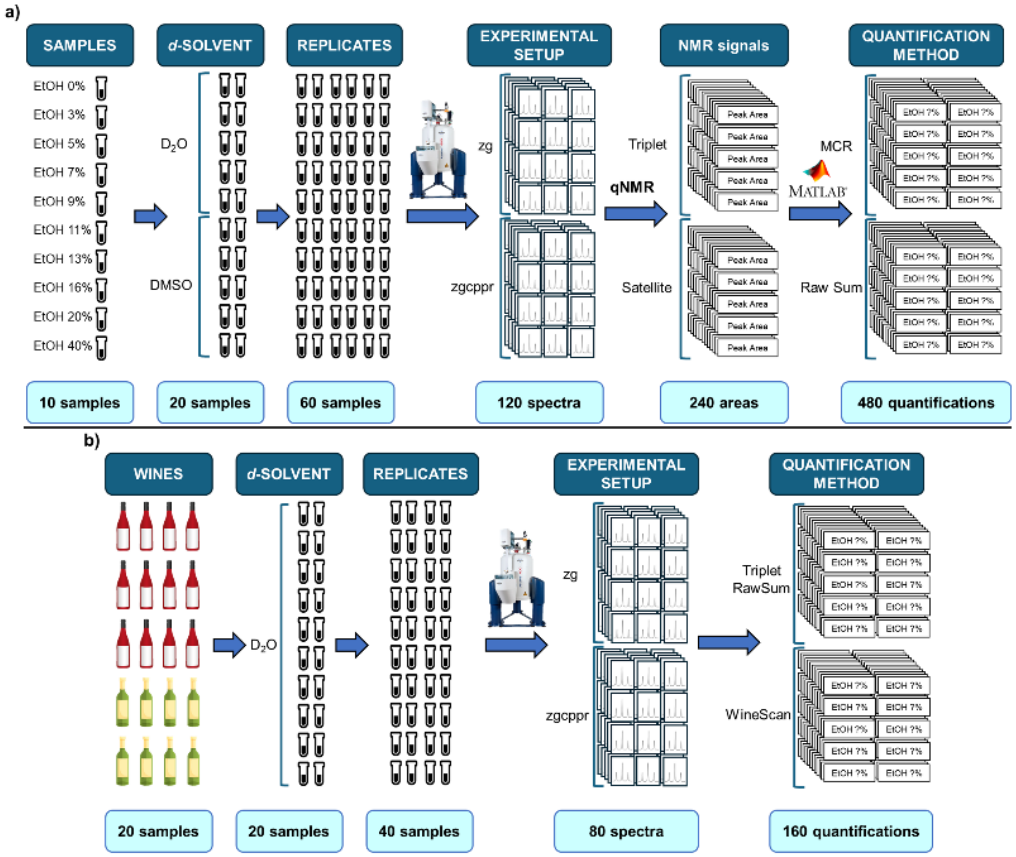

Figure 5.1. Visual representation of the experimental design for the ethanol pure solution study (a) and for the wine case study (b). a) Ten concentrations of ethanol in water are prepared using two different deuterated solvents and measured in three replicates using two different water suppression pulse sequences. In the data analysis step two different NMR signals (the triplet and its satellite) were monitored and two approaches for signal quantification were used. In total 480 quantifications were made. b) Application of the optimal parameters of (a) on a real case study on wines. 20 table wines (12 red wines and 8 white wines) were prepared in two replicates with D2O and measured using two pulse sequences. Ethanol quantification was performed on the ethanol triplet signal using the Raw Sum approach and using the WineScan IR instrument.

5.3.2 Sample preparation and data acquisition (^{1}H-NMR)

The ethanol solutions were prepared at different concentrations, spanning the range 0–40 %(v/v). For each sample, aliquots of 1 mL were prepared by mixing the specific volumes of ethanol and ultra-pure water. Deuterated solvents – D_2O or DMSO – were added to each sample (10% v/v) and the 3-(Trimethylsilyl)propionic acid (TSP) (5 mM) was used as internal standard. Three technical replicates of 1.0 ml were prepared for each sample and 600 µL were positioned in a 5 mm (O.D.) SampleJet NMR tubes (Bruker BioSpin, Ettlingen, Germany).

Deuterium oxide (D_2O, 99.9%), dimethyl sulfoxide-d6 (DMSO), sodium 3-trimethylsilyl-propionate-2,2,3,3-d4 (TSP), and ethanol (99.8%) were purchased from Sigma-Aldrich (Darmstadt, Germany). The water used throughout the study was purified using a Millipore lab water system (Merck KGaA, Darmstadt, Germany) equipped with a 0.22 μm filter membrane.

^{1}H-NMR spectra were recorded on a Bruker Avance III 600 operating at a proton Larmor's frequency of 600.13 MHz and equipped with a 5-mm broadband inverse (BBI) probe. Data acquisition and processing were carried out using the TopSpin software (version 4.2, Bruker, Rheinstetten, Germany) following exactly the same procedure applied for wine samples and described in Section 5.2. After temperature equilibration (5 min) the ^{1}H-NMR spectra were measured at 298 K using two different experiments: zgcppr and zg. Phase and baseline correction were performed in the TopSpin software. *i*coshift tool was used to align ethanol and TSP signals [23,24].

5.4 Ethanol spectra and conversion factor for satellite signals

In order to find the best approach for the quantification of ethanol, we focused on the evaluation of the different ^{1}H-NMR signals associated with the ethanol molecule. Figure 5.2 shows a representative ^{1}H-NMR spectrum of a red wine in which the two ethanol signals are clearly visible: the triplet (Figure 5.2a) at 1.180 ppm generated by the coupling of the methyl ($–CH_3$) protons with the two protons of the methylene ($–CH_2–$) moiety; and the quartet (Figure 5.22b) at 3.687 ppm generated by the coupling of the methylene ($–CH_2–$) protons with the three methyl ($–CH_3$) protons. In addition, close to those signals, the ^{13}C-satellites signal can be observed, which are clearly visible especially for the intense triplet signal (Figure 5.2c). The generation of these small signals is related to the coupling of the protons with the ^{13}C isotopes of the adjacent carbon atoms. Their signal shape is the same as the main proton signal and their lower intensity is due to the low natural abundance of the ^{13}C isotope, which is about 1 % of the total carbon presence.

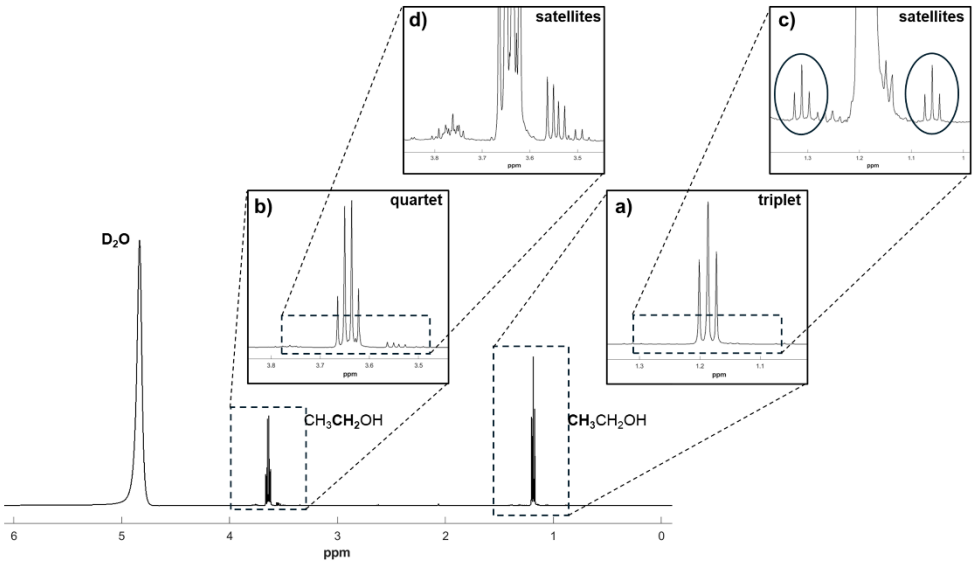

Figure 5.2. A representative [1]H-NMR spectrum of a red wine sample where the ethanol resonances are highlighted in the figure insets. The triplet (a) is related to the -CH3 protons; the quartet (b) is related to the -CH2- protons; the satellites (c and d) are generated from the coupling of the [13]C nuclei with the corresponding [1]H nuclei. Their intensity is low due to the low natural abundance of the [13]C nuclei.

Since compound quantification by ^{1}H-NMR is related to the area or to the intensity of the signals generated by the molecule (in our case the target compound is ethanol), the simultaneous presence of signals generated by other molecules (interferences) in the same spectral region can affect the quantification accuracy. In this work, the ethanol quartet was discarded due to the heavy overlap with the signals generated by glycerol at 3.448 ppm and other minor resonances in the same spectral region. For the same reason, the quartet ^{13}C-satellite signal was not included in the further analysis. Thus, ethanol quantification was performed only on the ethanol triplet signal or its satellite.

The inherent quantitative nature of ^{1}H-NMR spectroscopy is well known, and it is associated to the signals generated by proton-proton coupling. However, literature studies show the possibility of using the ^{13}C-satellite signals for quantification purposes [19]. Since the quantitative nature of the ^{1}H-NMR signals is due to the fact that the area under each NMR signal is proportional to the number of equivalent contributing protons, only the ^{1}H signals (i.e., the triplet at 1.18 ppm) are inherently quantitative. So, the ^{13}C-satellite signals (generated by the presence of ^{13}C nuclei and not directly from the ^{1}H nuclei) must be scaled up to their main quantitative ^{1}H-NMR signals to obtain quantitative areas. In our work, a correction factor k was calculated from the ^{1}H-NMR spectra of ethanol solutions with ethanol content ranging from 7% v/v to 16% v/v, which is similar to the % v/v ethanol content in wines. The ratio between the raw sum area of the main triplet and the raw sum area of its corresponding satellite signal was calculated. Then, for all the four

combinations of deuterated solvent and experimental setup, the mean value of the ratios was calculated (Figure 5.3). The selected correction factor, k=178.91, was the one obtained following the standard operating procedure for wine analysis (D_2O as a deuterated solvent and water suppression applied in the experimental setup). Ethanol areas suitable for absolute concentrations were calculated from the ^{13}C-satellite after multiplication by the correction factor.

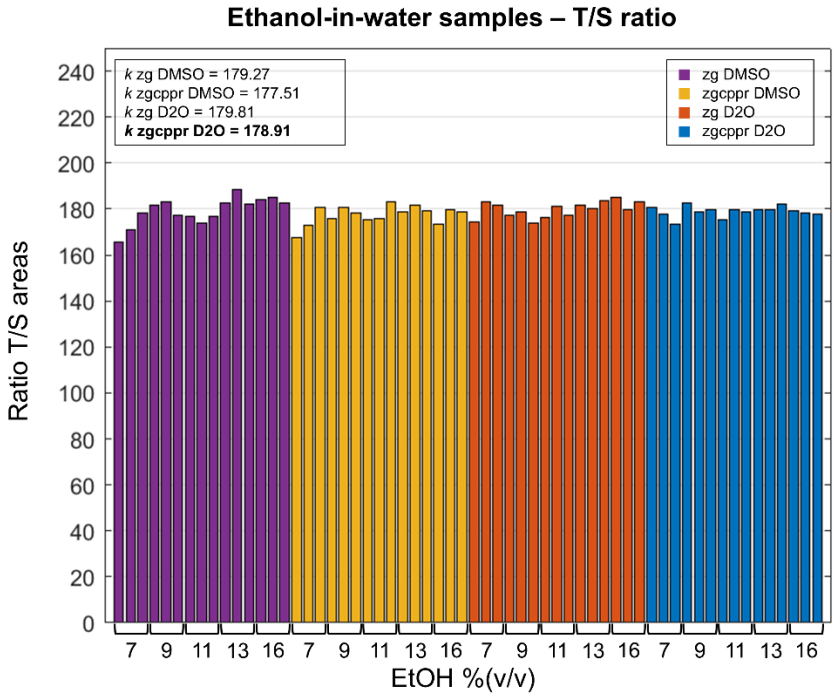

Figure 5.3. Bar plot representing the triplet/satellite (T/S) raw sum areas ratio. The bars are divided in four groups and coloured according to the combination of deuterated solvent and experimental setup. In each group the samples (the bars) are ordered according to the EtOH % (v/v). For each group a mean value (k) was calculated and reported in the top left corner. The samples selected to obtain the correction factor k=178.91 (zgcppr and D2O) are coloured in blue.

5.5 NMR signal area calculation and optimal conditions

The alcohol content in the ethanol-in-water solutions was quantified from the ^{1}H-NMR spectra by processing the different ethanol signals (triplet and its right ^{13}C-satellites), using two different integration methods. First, the "Raw Sum" method was applied: the intensities of all the data points associated with a specific signal are summed up to obtain the signal area. To minimize the systematic errors, an offset correction was applied to each signal interval to correct the baseline before calculating the area. This correction was performed on all the acquired spectra, and it consists in subtracting the lowest point of each spectrum (in terms of intensity) from all the NMR data points of that portion of spectrum.

The second integration method is based on Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [21,25,26]. The optimal number of components for MCR was set to one due to the simplicity of the system (just individual ethanol signals were processed with MCR, and only signals related to the ethanol molecule were observed and included in the modelled interval). The non-negativity constraint was applied to both the concentrations and resolved profiles directions.

The output of this approach is a series of relative concentration values that can be treated as signal areas (MCR allows to obtain the same output type of the Raw Sum approach) and just need to be scaled to provide absolute quantifications.

5.5.1 Combinations of different parameters

In order to evaluate the effect of the different tested experimental (i.e., the deuterated solvent) and instrumental (i.e., the pulse sequence) conditions, and to evaluate the suitability of the inspected ethanol signals (i.e., main triplet or carbon-related satellite) and of the employed quantification methods (i.e., Raw Sum or MCR), the NMR signal areas obtained for all the possible parameters combinations (2 possibilities × 4 conditions, 24 = 16 different combinations) were plotted against the corresponding theoretical (by design) ethanol content. To evaluate the quality of the calculated areas, and to hypothesize the best operating conditions employable in alcoholic beverages with ethanol content ranging from 3 to 40 % v/v, the correlation coefficient R^2 was calculated for the areas obtained using the 16 different combinations. So, after collecting such areas for all the combination of parameters tested (10 samples × 2 deuterated solvents × 3 replicates × 2 pulse sequences × 2 NMR signals × 2 quantification methods = 480 calculated areas), they were plotted against the real ethanol content (by design) to estimate the optimal experimental conditions. For each combination, the corresponding R^2 was calculated and all the correlation coefficients plotted together in Figure 5.4, where the case corresponding to the standard operating procedure proposed by Aru et al. [11] is highlighted in green.

The main outcome from the observation of Figure 5.4a concerns the quantification method. Due to the high concentration of ethanol, in some samples the triplet signal is not well-resolved and the MCR approach results less accurate and less consistent with respect to the "more classical" Raw Sum approach. Except for the MCR approach applied on the triplet signal, all other combinations resulted in an R^2 higher than 0.97.

Focusing only on the combinations with R^2 values higher than 0.99 (Figure 5.4b), it can be pointed out that there are not many differences in terms of consistency between the pulse sequences used, while some effects related to the deuterated solvent can be spotted. In particular, DMSO seems to be more consistent, even

changing either the pulse sequence or the monitored signal or also the quantification method. Curiously, MCR seems to work very well if applied to the satellite signal acquired using DMSO.

Focusing on the standard operating procedure highlighted in green in Figure 5.4, its accuracy is comparable with the best results obtained in DMSO (R^2 of 0.995 vs R^2 of 0.998), and since wine samples are commonly analysed with ^{1}H-NMR using D_2O, it was decided focus the following investigation only on samples using this solvent and proceed with real case study on wine samples whose results are described in Section 5.7.

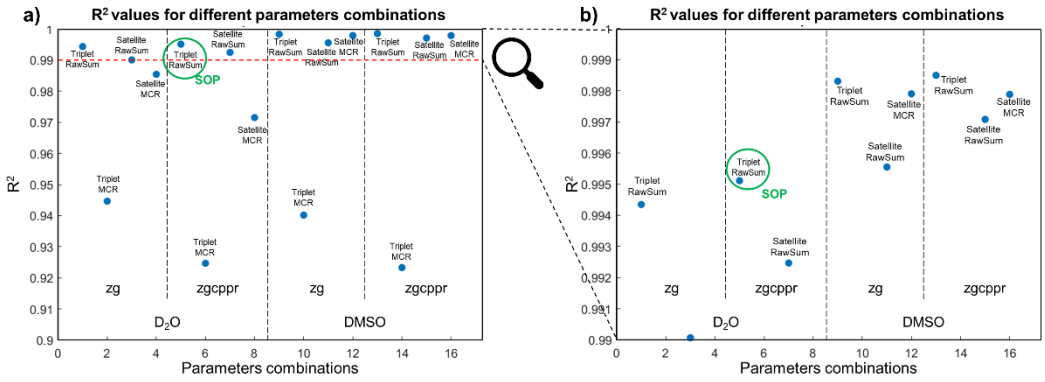

Figure 5.4. Visualization of the correlation coefficients (R^2) obtained for the different parameters combinations (a) and focus only on the combinations with an R^2 higher than 0.99 (b).

5.6 Indirect ethanol quantification

5.6.1 An FT-IR based method: WineScan

The Infrared (IR) spectroscopy, even if it is largely used for qualitative analysis, can be effectively used also for quantification purposes. In brief, since the intensity of infrared absorption bands correlates with the concentration of the absorbing species, the signals obtained at specific wavelengths can be referred to the concentration of chemical compounds in the sample. However, this type of spectroscopy is not inherently quantitative, but it requires the development of a reliable calibration curve, resulting in an indirect quantification method. The calibration is needed since the equation that correlates absorbance (A) and concentration (C) is the Lambert-Beer law (Equation 5.1). In particular, the terms ε (molar absorptivity) and l (path length of the sample cell) need to be estimated using a calibration model specific developed for each compound. In practice, the calibration step consists in measuring the absorbance of standards at different known concentrations to create a mathematical model able to quantify future samples with unknown concentrations. The developed calibration method can be

linear or non-linear, using methods such as the partial least square regression (PLS), or even artificial neural networks (ANN) can be used for more complex situations.

$$A = \varepsilon \cdot c \cdot l \qquad (5.1)$$

In this project, the ethanol content in wine samples was calculated by FT-IR spectroscopy using the WineScan instrument (WineScan FT 120, FOSS A/S, Hillerød, Denmark). This is a Fourier Transform interferometer equipped with a 37 µm transmission measurement cell with CaF_2 windows. The scanning range is from 929 to 5011 cm^{-1} and 12 scans are averaged to produce the final spectrum.

Before WineScan measurements, the samples were centrifuged at 4000 rpm for 5 min. The WineScan instrument collects information about the fundamental molecular vibrations, and the spectra have been carefully calibrated by the producer on a long list of wine and grape juice parameters using advanced multivariate regression techniques. The wines' ethanol content measured with the WineScan was used as a reference in our work to compare its results with the NMR quantifications.

5.6.2 PLS regression for NMR spectra

An alternative approach for the indirect ethanol quantification using the proton NMR spectroscopy can be the application of PLS regression to create a calibration model suitable for quantifying the ethanol content from the NMR interval of a selected ethanol signal. This approach follows the same idea used by the WineScan instrument, where chemometric methods (like PLS) are employed to create calibration models to obtain reliable indirect quantifications.

The PLS approach was applied in our project for the indirect quantification of ethanol, where the two applied area calculation methods (Raw Sum and MCR) were compared and evaluated to find which one had better EtOH prediction ability when applied on the ethanol ^{13}C-satellite triplet signal. Two PLS regression models [27] were built using as a calibration dataset the ^{13}C-satellites signal interval and the Raw Sum or the MCR quantifications as a response vector, separately. The number of latent variables was set to 4 respectively describing a cumulative X-variance of 99.57 % and a cumulative Y-variance of 99.72 % for the Raw Sum response and 99.88 % for the MCR response.

Even if the spectra were suitable for the use of only one component due to the dominant presence of the ethanol signals, a higher number of components was selected to improve the comparison among spectra overcoming small signal misalignments (i.e., small but potentially impactful horizontal shifts).

The models' performances were evaluated by comparing the Root Mean Square Error (RMSE) and the R^2 value. The RMSE comes from the difference between the designed ethanol content and the predicted ethanol content and represents the

overall error in prediction. The R^2 value represents the overall quality of the regression fit. All PLS models were also validated by means of cross-validation, with the venetian blinds selection scheme taking care of the sample replicates in the dataset.

By inspecting the prediction plots obtained from the two models (Figure 5.5), the quantification methods prediction abilities are both accurate, with $RMSE_{CV}$ = 0.69 % (v/v) for the raw sum method and $RMSE_{CV}$ = 0.49 % (v/v) for the MCR-based method. The R^2 in Cross Validation was R^2 = 0.996 for the raw sum-based model and R^2 = 0.998 for the MCR-based model. For both models, the samples with 40 %(v/v) EtOH content were the one with the worst prediction accuracy, however, these samples have high residual values and have a strong influence on the model.

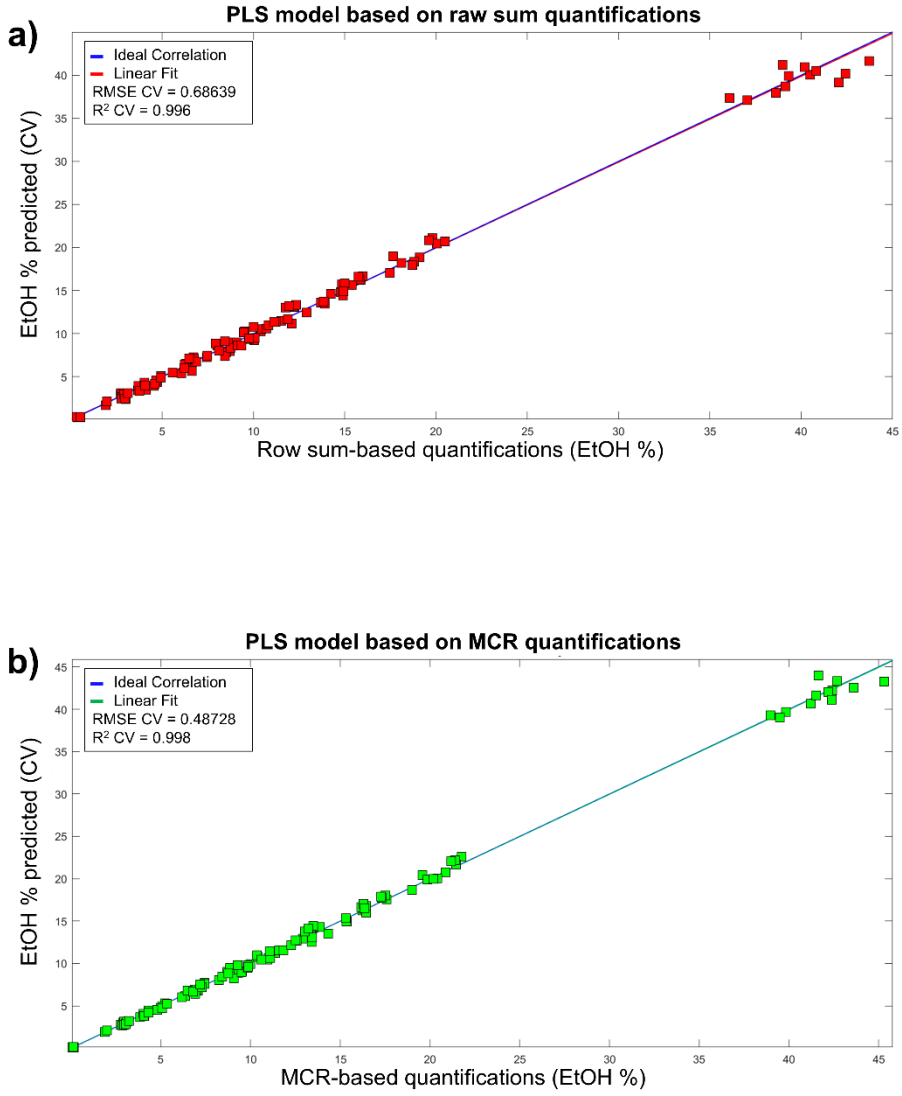

Figure 5.5. Prediction plots obtained from the two PLS models built using the satellite signal intervals as a calibration dataset. The red samples (a) are obtained using the row sum quantifications as response vector. The green samples (b) are obtained using instead the MCR quantifications as response vector.

To briefly summarise the results obtained from the NMR indirect ethanol quantification based on PLS models: this approach, even if it was developed on pure ethanol solutions, seems to be affordable and robust. As expected, the performances obtained using the ^{1}H-NMR spectroscopy are comparable with the one achieved by FT-IR spectroscopy. For this reason, in our project, we decided to take advantage of this NMR characteristic and to avoid the need for calibration models by trying to develop a method able to perform direct ethanol quantifications from NMR spectra both of ethanol-in-water solutions and wine samples.

5.7 Case study evaluation: real wine samples

5.7.1 Direct ethanol quantification using optimal conditions

The ethanol content in wine samples was quantified from the NMR spectra using the Raw Sum approach applied also for the ethanol-in-water solution (see Section 5.5.2). In addition, using the TSP signal as internal standard, the ethanol percentage by volume (EtOH % v/v) was calculated. The signal areas obtained from "Raw Sum" method were firstly converted into molar concentrations using Equation 5.2:

$$C_{EtOH} = (I_{EtOH}/I_{TSP}) * (N_{TSP}/N_{EtOH}) * C_{TSP} * D \qquad (5.2)$$

Where C_{EtOH} is the ethanol concentration in mM, I_{EtOH} is the integral of the ethanol signal, I_{TSP} is the integral of the reference compound TSP, N_{TSP} is the number of protons giving rise to the TSP signal, N_{EtOH} is the number of protons giving rise to the ethanol resonance, C_{TSP} is the known TSP concentration (1.5 mM), and D is a correction factor associated with the sample dilution. In this work its value is 1.0/0.9 because the amount of deuterated solvent was 10 %(v/v) per each sample.

To obtain values in EtOH % v/v unit of measurement, the ethanol molecular weight (MW_{EtOH} = 46.07 g/mol) and the solution density (set according to the ethanol percentage in the different solutions) were considered, and the values expressed in EtOH %(v/v) were obtained by converting the molar concentrations using Equation 5.3:

$$\%_{EtOH} = ((C_{EtOH} * MW_{EtOH})/d/1000) * 100 \qquad (5.3)$$

Where $\%_{EtOH}$ is the ethanol concentration in %(v/v), C_{EtOH} is the previously calculated ethanol concentration in mM, MW_{EtOH} is the molecular weight of ethanol, d is ethanol density and 1000 was used to assess the correct unit of measurement.

For the case study, a total of 20 table wines (12 red and 8 white, Table 5.1 sample) were analysed with ^{1}H-NMR spectroscopy following the previously selected standard operating procedure using D_2O as deuterated solvent. In addition, both pulse sequences ("zg" and "zgcppr") were used also for the wine samples NMR acquisitions. The ethanol content was then quantified from the triplet signal in the

NMR spectra using the Raw Sum approach. The absolute designed ethanol content was calculated both in molar concentrations and in EtOH % (v/v) using Equations 5.2 and 5.3.

The contents of EtOH %(v/v) were plotted against the WineScan measurements (used as a reference) for all the samples (Figure 5.6). Unexpectedly, these quantifications resulted in a lower correlation coefficient ($R^2 = 0.909$) with respect to the one obtained from the previous study on ethanol solutions ($R^2 = 0.995$). The ethanol percentage obtained with NMR highlighted a general underestimation of the real ethanol content that lead us to explore also other possibilities such as the application of an external standard approach (see section 5.7.3).

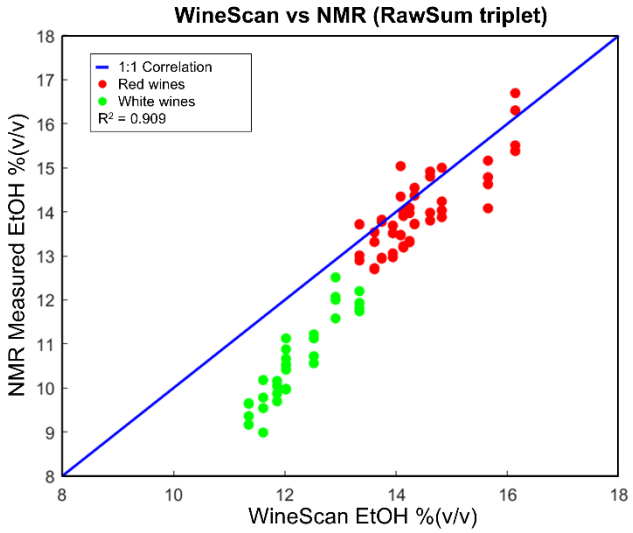

Figure 5.6. EtOH quantifications obtained from NMR spectra using the Raw Sum approach on the ethanol triplet signal vs EtOH quantifications obtained from the WineScan instrument used as a reference. Red wine sample are coloured in red, white wines in green.

5.7.2 Parameters evaluation using ASCA

To evaluate the effect of the parameters on the ethanol quantification in wine samples, ANOVA Simultaneous Component Analysis (ASCA) [28,29] was applied using the PLS_Toolbox (version 8.9.2, Eigenvector Research Inc., Manson, WA, USA).

Two different ASCA models were explored. The first was developed using as a response the ethanol triplet area obtained from the spectra of the case study on wines. The second was developed instead using the corresponding TSP area as a response. For both models a DoE coded matrix was built considering three factors: the pulse sequence used during NMR spectra acquisition, the wine type (i.e., red or white) and the number of sample replicates, to evaluate also the experimental reproducibility. A fourth factor representing the ethanol content in the sample was coded in the DoE matrix used for the model with the TSP area set as the response.

For all models, to assess the significance of the ASCA outcomes, the corresponding p-value was calculated setting the number of permutations to 1000.

To better understand the reasons behind the lower precision in terms of ethanol quantification observed from the ^{1}H-NMR spectra acquired on the wine samples, ASCA was used to explore if any parameters would affect the areas of the ethanol triplet signal and of the internal standard TSP signal used for the quantification. The first ASCA model was developed to explore the effects of the pulse sequence, of the wine type and of the replicates on the ethanol triplet signal area. The results reported in Table 5.2 (with the corresponding p-value) revealed a strong effect of the wine type, as expected and mainly due to the different ethanol content in red and white wines; a rather reduced effect associated with the pulse sequence was also found. At the same time, no effect was observed for the replicates, thus excluding the presence of potential instrumental errors.

Table 5.2. Results obtained from ASCA performed using the ethanol triplet signal area as a response. The monitored factors are the pulse sequence (F1), the wine type (F2) and the replicates (F3)

Factor	PCs	Effect	p-value
F1: Pulse sequence	1	1.67	0.043
F2: Wine type	1	64.25	0.001
F3: Replicates	1	0.07	0.710
F1 × F2	1	0.01	0.893
F1 × F3	1	0.01	0.889
F2 × F3	1	0.00	0.931
Residuals	/	33.99	/

The second model was developed to explore the same effects monitored in the first model, plus a fourth parameter related to the ethanol content in the wine samples. For this model the response was the signal area of the internal standard TSP. The results are reported in Table 5.3 with the corresponding p-value. A small effect was again observed for the pulse sequence used and again no effects related to the replicates were observed. However, surprisingly enough, a large effect of 59.69 and a significant effect of 14.48 were observed respectively for the wine type and for the ethanol content, meaning that the TSP signal area is changing across the wine samples even if they were prepared all with the same TSP concentration and following the same standard operating procedure. This unexpected result is strongly affecting the previously calculated ethanol % v/v and justifies the lower precision obtained with respect to the ethanol solution study.

Table 5.3. Results obtained from ASCA performed using the TSP signal area as a response. The monitored factors are the pulse sequence (F1), the wine type (F2), the replicates (F3) and the ethanol content (F4).

Factor	PCs	Effect	p-value
F1: Pulse sequence	1	1.96	0.001
F2: Wine type	1	59.69	0.001
F3: Replicates	1	0.04	0.462
F4: EtOH content	1	14.48	0.001
F1 × F2	1	0.05	0.525
F1 × F3	1	0.04	0.583
F1 × F4	1	0.20	0.984
F2 × F3	1	0.07	0.465
F2 × F4	1	15.54	0.001
F3 × F4	1	0.90	0.464
Residuals	/	7.01	/

5.7.3 Ethanol quantification using an external standard

To improve the ethanol quantification in wine samples, and to compare different approaches, a method described in literature using the TSP signal as an external standard [30–32]. In short, two spectra were selected from the samples of the experimental design on ethanol solutions: the first one ("external standard") is the one containing only the signals of TSP, water and D_2O (with EtOH 0.0 % v/v); the second one ("ethanol reference") is obtained from the ethanol solution with EtOH 13.0 % v/v, corresponding to an average alcoholic content in wines. The pulse sequence used for these two samples was the "zg" (no water suppression) to prevent adding variations due to the water suppression pulse sequence. All the other acquisition parameters were kept identical for all the experiments.

Starting from these two samples, a coefficient (k_{ext}) was calculated using Equation 5.4:

$$k_{ext} = A_{TSP}/A_{EtOH} * C_{EtOH}/C_{TSP} * N_{EtOH}/N_{TSP} \qquad (5.4)$$

A_{TSP} is the area of the TSP signal in the external standard sample, A_{EtOH} is the area of the ethanol triplet signal in the ethanol reference sample, C_{EtOH} is the ethanol molar concentration in ethanol reference sample corresponding to 13 % v/v, C_{TSP} is the TSP molar concentration of the external standard sample, N_{EtOH} is the number of protons described by the ethanol triplet signal (3) and N_{TSP} is the number of protons described by the TSP signal (9).

The k_{ext} coefficient was calculated in triplicate considering the three replicates of the external standard sample and of the ethanol reference sample. The final obtained mean value (k_{ext} = 1.8681) was used as a correction factor to obtain the molar concentration of all the wine samples analysed during the case study starting from

the area of the ethanol triplet signal applying, for all samples, the formula shown in Equation 5.5

$$C_{EtOH} = k_{ext} * A_{EtOH} * N_{TSP} * C_{TSP} / A_{TSP} * N_{EtOH} \qquad (5.5)$$

When applying this formula, to obtain reliable quantifications, C_{TSP}, A_{TSP}, N_{TSP} and N_{EtOH} must have the same values as the ones used in Equation 5.4. While A_{EtOH} is the area of the signal to convert in molar concetration.

All the ethanol quantifications of the case study on wines were performed starting from the ethanol triplet signal area calculated with Raw Sum. The areas were converted into molar concentrations using Equation 5.2 and then they were converted into EtOH % v/v using Equation 5.3. The obtained NMR-based quantifications were plotted against the corresponding WineScan measurements used as a reference (Figure 5.7). Because of the small effect showed by the pulse sequence on the areas, only the spectra acquired with the "zgcppr" pulse sequence were considered.

Figure 5.7 clearly shows that the external standard approach allows to drastically reduce the previously highlighted problems related to the TSP area. In particular, the strong effect of the wine type is now made neglectable, demonstrating that it was mainly generated by a matrix effect affecting the TSP signal area. The small underestimation still present in some samples can be referred to different causes. The first can be that the ethanol content is varying from 11 % v/v to 16 % v/v in wine samples (while k_{ext} constant was calculated from a sample with EtOH 13 % v/v). Secondly, FT-IR based instruments (i.e., the WineScan) are not directly quantitative, but they need to be calibrated using regression models while NMR is proven to be suitable for direct quantifications.

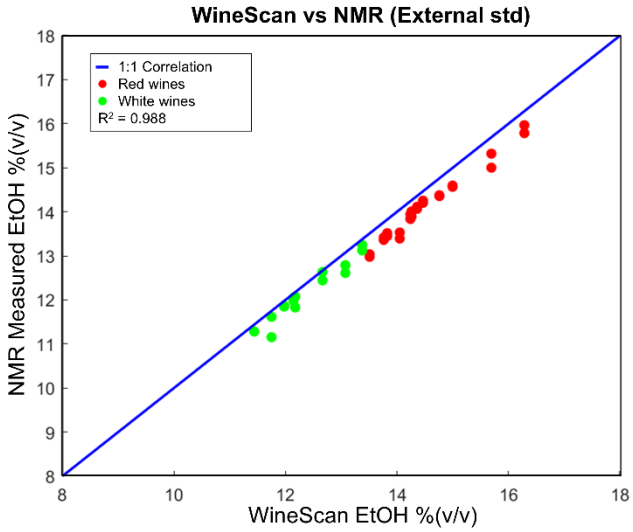

Figure 5.7. EtOH quantifications obtained from NMR spectra using the external standard approach vs EtOH quantifications obtained from the WineScan instrument used as a reference. Red wine sample are coloured in red, white wines in green.

5.8 Conclusions and future perspectives

In this work, 233 wines were analysed using ^{1}H-NMR spectroscopy to create a dataset to be used as a starting point from which build upon projects with different aims. The first main project is focused on metabolomics topics, and it is mainly based on traceability identification and quality assessment. Unfortunately, due to the large amount of data collected, this part of the project is still ongoing, and it was not included in this thesis. However, the wine dataset was used as a starting point for a more technical project, focused on the development of a detailed experimental design, performed to explore the possibility of using ^{1}H-NMR spectroscopy as a method for ethanol quantification in alcoholic beverages. Different experimental parameters and data processing approaches were evaluated on a series of samples of pure ethanol solutions with concentrations ranging from 0 to 40 %(v/v). As expected, a good linearity between the main ^{1}H triplet and the ^{13}C triplet satellite integrals was observed. A conversion factor k was calculated, which was used to scale the integral of the ^{13}C triplet satellite prior to calculating ethanol absolute concentrations.

The different experimental parameters used in the experimental design were evaluated looking at the correlation coefficients R^2 obtained from the comparison between the triplet areas for the different experimental combinations and the designed ethanol content, whose correlation was also visually investigated. The results show that, even if DMSO solvent seems slightly more consistent across different experimental conditions and quantification methods, also the commonly used standard operating procedure (D_2O as a solvent, and quantification performed with Raw Sum on ethanol triplet signal) provided comparable results.

Finally, the obtained optimal conditions were applied to a real case study on red and white wines. Unexpectedly, the area of the TSP internal standard was found to significantly vary among the samples due to different wine type and ethanol content. To overcome these problems, a parallel approach based on the use of an external standard was explored. The ethanol quantification obtained from this approach gave interesting results both in terms of precision and consistency, overcoming the unexpected matrix effects related to the TSP signal area.

To conclude, this study evaluated different experimental conditions using an experimental design developed to inspect the problem of measuring the ethanol content in alcoholic beverages. Among the tested parameters, the use of water suppression pulse sequence together with the use of D_2O as a solvent was confirmed to be a correct choice in terms of experimental procedure to follow. However, the use of TSP as an internal standard revealed some problems, especially while working with different ethanol concentrations. On the other hand, the use of the TSP as an external standard gave promising results, due to the application of a correction factor *k* able to mitigate/reduce/remove the previously described matrix effect.

References | Chapter 5

[1] W. Alahmad, S.I. Kaya, A. Cetinkaya, P. Varanusupakul, S.A.A. Ozkan, Green chemistry methods for food analysis: Overview of sample preparation and determination, Advances in Sample Preparation 5 (2023) 100053. https://doi.org/10.1016/J.SAMPRE.2023.100053.

[2] A. Hassoun, S. Jagtap, G. Garcia-Garcia, H. Trollman, M. Pateiro, J.M. Lorenzo, M. Trif, A.V. Rusu, R.M. Aadil, V. Šimat, J. Cropotova, J.S. Câmara, Food quality 4.0: From traditional approaches to digitalized automated analysis, J Food Eng 337 (2023) 111216. https://doi.org/10.1016/J.JFOODENG.2022.111216.

[3] J.A.L. Pallone, E.T. dos S. Caramês, P.D. Alamar, Green analytical chemistry applied in food analysis: alternative techniques, Curr Opin Food Sci 22 (2018) 115–121. https://doi.org/10.1016/J.COFS.2018.01.009.

[4] I.W. Burton, M.A. Quilliam, J.A. Walter, Quantitative 1H NMR with external standards: Use in preparation of calibration solutions for algal toxins and other natural products, Anal Chem 77 (2005) 3123–3131. https://doi.org/10.1021/AC048385H/SUPPL_FILE/AC048385HSI20050128_044400.PDF.

[5] R.D. Farrant, J.C. Hollerton, S.M. Lynn, S. Provera, P.J. Sidebottom, R.J. Upton, NMR quantification using an artificial signal, Magnetic Resonance in Chemistry 48 (2010) 753–762. https://doi.org/10.1002/MRC.2647.

[6] G. Wider, L. Dreier, Measuring protein concentrations by NMR spectroscopy, J Am Chem Soc 128 (2006) 2571–2576. https://doi.org/10.1021/JA055336T/ASSET/IMAGES/MEDIUM/JA055336TN00001.GIF.

[7] P. Giraudeau, I. Tea, G.S. Remaud, S. Akoka, Reference and normalization methods: Essential tools for the intercomparison of NMR spectra, J Pharm Biomed Anal 93 (2014) 3–16. https://doi.org/10.1016/J.JPBA.2013.07.020.

[8] P. Giraudeau, Challenges and perspectives in quantitative NMR, Magnetic Resonance in Chemistry 55 (2017) 61–69. https://doi.org/10.1002/MRC.4475.

[9] A.M. Torres, W.S. Price, Common problems and artifacts encountered in solution-state NMR experiments, Concepts Magn Reson Part A Bridg Educ Res 45A (2016). https://doi.org/10.1002/CMR.A.21387.

[10] V. V. Krishnan, N. Murali, Radiation damping in modern NMR experiments: Progress and challenges, Prog Nucl Magn Reson Spectrosc 68 (2013) 41–57. https://doi.org/10.1016/J.PNMRS.2012.06.001.

[11] V. Aru, K.M. Sørensen, B. Khakimov, T.B. Toldam-Andersen, S.B. Engelsen, Cool-Climate Red Wines—Chemical Composition and Comparison of Two Protocols for 1H–NMR Analysis, Molecules 2018, Vol. 23, Page 160 23 (2018) 160. https://doi.org/10.3390/MOLECULES23010160.

[12] S. Armenta, S. Garrigues, M. de la Guardia, Recent developments in flow-analysis vibrational spectroscopy, TrAC Trends in Analytical Chemistry 26 (2007) 775–787. https://doi.org/10.1016/J.TRAC.2007.06.003.

[13] Chemistry of Foods and Beverages: Recent Developments - Google Libri, (n.d.). https://books.google.it/books?hl=it&lr=&id=eiaX6BMnHLgC&oi=fnd&pg=PP1&dq=Charalambous,+G.,+2012.+Chemistry+of+foods+and+beverages:+recent+developments.&ots=jKyz11lhL2&sig=SJtbZ40eheazIUo5vNbz06QoF4w&redir_esc=y#v=onepage&q=Charalambous%2C%20G.%2C%202012.%20Chemistry%20of%20foods%20and%20beverages%3A%20recent%20developments.&f=false (accessed February 18, 2025).

[14] J.C. Teipel, T. Hausler, K. Sommerfeld, A. Scharinger, S.G. Walch, D.W. Lachenmeier, T. Kuballa, Application of 1H Nuclear Magnetic Resonance Spectroscopy as Spirit Drinks Screener for Quality and Authenticity Control, Foods 2020, Vol. 9, Page 1355 9 (2020) 1355. https://doi.org/10.3390/FOODS9101355.

[15] R. Preti, Progress in Beverages Authentication by the Application of Analytical Techniques and Chemometrics, Quality Control in the Beverage Industry: Volume 17: The Science of Beverages (2019) 85–121. https://doi.org/10.1016/B978-0-12-816681-9.00003-5.

[16] A. Palmioli, D. Alberici, C. Ciaramelli, C. Airoldi, Metabolomic profiling of beers: Combining 1H NMR spectroscopy and chemometric approaches to discriminate craft and industrial products, Food Chem 327 (2020) 127025. https://doi.org/10.1016/J.FOODCHEM.2020.127025.

[17] E. López-Rituerto, K.M. Sørensen, F. Savorani, S.B. Engelsen, A. Avenoza, J.M. Peregrina, J.H. Busto, Monitoring of the Rioja red wine production process by 1H-NMR spectroscopy, J Sci Food Agric 102 (2022) 3808–3816. https://doi.org/10.1002/JSFA.11729.

[18] H.J.R. Caleja-Ballesteros, J.I. Ballesteros, M.C. Villena, No-D quantitative 1H Nuclear Magnetic Resonance spectroscopy method for the determination of ethanol in distilled spirits, Microchemical Journal 164 (2021) 105999. https://doi.org/10.1016/J.MICROC.2021.105999.

[19] E. López-Rituerto, S. Cabredo, M. López, A. Avenoza, J.H. Busto, J.M. Peregrina, A Thorough Study on the Use of Quantitative 1H NMR in Rioja Red Wine Fermentation Processes, J Agric Food Chem 57 (2009) 2112–2118. https://doi.org/10.1021/JF803245R.

[20] R. Leardi, Experimental design in chemistry: A tutorial, Anal Chim Acta 652 (2009) 161–172. https://doi.org/10.1016/j.aca.2009.06.015.

[21] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Crit Rev Anal Chem 36 (2006) 163–176. https://doi.org/10.1080/10408340600970005.

[22] S.C. Rutan, A. de Juan, R. Tauler, Introduction to Multivariate Curve Resolution, Comprehensive Chemometrics 2 (2009) 249–259. https://doi.org/10.1016/B978-044452701-1.00046-6.

[23] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, Journal of Magnetic Resonance 202 (2010) 190–202. https://doi.org/10.1016/j.jmr.2009.11.012.

[24] F. Savorani, G. Tomasi, S. Engelsen, Alignment of 1D NMR Data using the iCoshift Tool: A Tutorial, in: 2013: pp. 14–24. https://doi.org/10.1039/9781849737531-00014.

[25] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems 30 (1995) 133–146. https://doi.org/10.1016/0169-7439(95)00047-X.

[26] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review, Anal Chim Acta 1145 (2021) 59–78. https://doi.org/10.1016/j.aca.2020.10.051.

[27] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[28] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048. https://doi.org/10.1093/BIOINFORMATICS/BTI476.

[29] C. Bertinetto, J. Engel, J. Jansen, ANOVA simultaneous component analysis: A tutorial review, Anal Chim Acta X 6 (2020). https://doi.org/10.1016/j.acax.2020.100061.

[30] A. Avenoza, J.H. Busto, N. Canal, J.M. Peregrina, Time course of the evolution of malic and lactic acids in the alcoholic and malolactic fermentation of grape must by quantitative 1H NMR (qHNMR) spectroscopy, J Agric Food Chem 54 (2006) 4715–4720. https://doi.org/10.1021/JF060778P/ASSET/IMAGES/LARGE/JF060778PF00007.JPEG.

[31] E. López-Rituerto, S. Cabredo, M. López, A. Avenoza, J.H. Busto, J.M. Peregrina, A thorough study on the use of quantitative 1H NMR in Rioja red wine fermentation processes, J Agric Food Chem 57 (2009) 2112–2118. https://doi.org/10.1021/JF803245R/ASSET/IMAGES/LARGE/JF-2008-03245R_0002.JPEG.

[32] E. López-Rituerto, K.M. Sørensen, F. Savorani, S.B. Engelsen, A. Avenoza, J.M. Peregrina, J.H. Busto, Monitoring of the Rioja red wine production process by 1H-NMR spectroscopy, J Sci Food Agric 102 (2022) 3808–3816. https://doi.org/10.1002/JSFA.11729.

Chapter 6

Case study #4: image analysis in agrifood

6.1 Project overview

6.1.1 Image analysis to automatically extract objects' features

In the past years, the interest towards approaches based on automatic image analysis has been considerably growing. The possibility of a rapid automated extraction of qualitative and quantitative information from greyscale and/or coloured images changed the way researchers have been performing their studies and their understanding of image-related scientific issues. Image analysis is generally non-invasive and non-destructive [1], and it provides accurate information employing methods that can be used in a wide variety of fields such as materials science [2–4], chemistry [5–8] engineering [9–11] and even biomedicine [12–14].

Several analytical techniques generate images and, in many cases, regular shaped objects with a wide variety of dimensions and dispositions are represented [15,16]. Researchers often face the problem of measuring particles and objects present in the images and many approaches, both automatic and manual, have been developed and fruitfully applied [17,18]. Among image analysis and characterization methodologies, Field Emission Scanning Electron Microscopy (FESEM) is gaining more and more momentum. Its application in engineering ranges from empty spaces quantification [19] to particle size distribution measurement [20,21] and morphology exploration [22]. It is also used to characterize grains of different materials like sand [23] to evaluate structural morphology [24] and damages [25], and in general also to characterize materials with different aims like carbon materials for energy storage [26] or nanotubes for biomedical approaches [27].

In this context, the idea was to develop a new algorithm able to perform high-throughput analyses of large sets of FESEM images with the aim of automatically extracting important morphological features (e.g., area, perimeter, circularity, etc.) of similarly shaped objects. Indeed, in addition to the previously mentioned research applications, image analysis approaches have been becoming popular also in the food industry [28] and, more generally, for food-related topics [29]. To mention some examples: food quality assessment and evaluation [30,31], texture and structure analyses [32,33], and characterization and measurements of particles of different nature [34,35].

Furthermore, since the algorithm requires some parameters to be manually set as the model inputs, an experimental design approach was applied to evaluate the effect of these parameters on the morphological features calculation, but also on the overall processing time and performances of the algorithm.

6.1.2 A case study on rice kernels and glycaemic index

To show the applicability and the effectiveness of the previously described approach, but also to support the development of the new algorithm and of the experimental design, a case study about rice kernels characterization, both morphological and biochemical, was investigated [36]. The aim of the case study was, in addition to developing the algorithm, also on the evaluation of the correlations between the glycaemic index and the grain physical structure of a considerable number of different rice varieties.

Rice is probably the most important staple food, especially in low- and middle-income countries where it serves as the primary source of livelihood for nearly 4 billion people worldwide. In addition, the global demand for rice is likely to rise and reach 536 million tons (milled rice) by 2030. In 2020 rice was cultivated on approximately 164.2 million hectares, accounting for 22.3 % of the total cereals area and 11.8 % of world's arable land. This cultivation yielded 756.7 million tons paddy rice, representing nearly 25.3 % of the global cereal production (FAOSTAT Database, *December 2022*).

The glycaemic index (GI) concept, first introduced in 1981 [37], classify foods based on their postprandial glycaemic response relative to a reference carbohydrate source [38]. GI values are typically expressed as a percentage compared to glucose or white bread (International Standards Organization, 2010 – ISO 26642), with the latter having a GI of 70 in relation to the former [39]. The consumption of high-GI foods causes a rapid rise in blood-glucose level and many studies have shown correlations between this popular practice and the development of type 2 diabetes [40–42], cardiovascular diseases [43–45], and cancer [46,47].

In Europe, according to International Diabetes Federation, the number of diabetic subjects between the ages of 20 and 79 years is about 61.4 million, corresponding to the 9.2 % of the entire population ("IDF Diabetes Atlas, 10th edition", *2021*). By focusing only on Italy, more than 3.5 million people are affected by diabetes, accounting for 5.8 % of the population (Istat, *2017*). Given the increasing number of people affected by diabetes all over the world and given the important role that such a widely consumed and valued staple food could play in enhancing human nutrition [48], the use of a careful diet based on low-GI rice varieties could be an effective strategy for diabetes prevention and management. The use of this approach can also reduce the progression of prediabetes, prevent hyperglycaemia and reduce the risk of complications such as cardiovascular diseases [49,50]. For these reasons, it can be promising to identify and develop rice varieties with a low

GI, while maintaining grain quality and organoleptic characteristics to better meet the needs of consumers.

Generally, rice is classified as a high-GI food [51,52] and is not particularly suitable for people suffering from diabetes or other metabolic disorders. However, this is not always accurate as the GI value can vary significantly among different varieties [39,53–55]. In addition, several factors, such as cooking method and time, presence of antioxidant components, grain degree of polishing and industrial treatments, can affect the GI value in rice [39,42,48,55–60]. Recently, a wide study proposed by the International Rice Research Institute (IRRI) and the Commonwealth Scientific and Industrial Research Organisation's (CSIRO) Food Futures Flagship on 235 rice varieties from all over the world, conducted while working under the same controlled conditions, highlighted the presence of a strong genetic variation and a great variability in GI [54]. This is mainly due to the fact that the genetic basis of GI in rice is complex and involves multiple factors [61] that increase the difficulties in performing breeding programs for the development of low-GI varieties, more suitable for diabetic subjects. Among the different factors affecting rice GI, the Waxy gene (Wx) has been suggested to play a main role [54]. The Wx gene is of fundamental importance in the synthesis of amylose in endosperm of rice [62] and the amylose, which is a relatively short polymer of glucose units linked by α 1→4 bonds and characterized by a mainly linear structure, is the main component of the rice starch together with amylopectin, which is a longer polymer where glucose units are arranged linearly by α 1→6 bonds occurring every 24 to 30 glucose units, that is characterized by a branched structure [39,61,63,64]. As a consequence, starch, that represents the main form of energy storage in plants, as well as the main component of rice grain endosperm, has a key role in determining rice postprandial glycaemic response, and consequently its GI.

From a structural point of view, differing from other main cereals, rice endosperm starch granules are usually 10–20 μm in diameter [65]. Each compound granule is composed of smaller granules with an average size between 2–3 and 7–10 μm, as reported in several publications [65–69]. The amylose content in rice cultivars is greatly variable, resulting practically absent in glutinous or "waxy" varieties and reaching 30 % or even more in some non-waxy ones and up to 40 % in amylose extender mutants [59,69–71]. Rice varieties can then be classified according to their amylose content as follows: waxy (0–2 %), very low (2–12 %), low (12–20 %), medium (20–25 %) and high (25–33 %). Varieties having a high amylose content usually show a lower GI, therefore the correlation observed between the two factors suggests that amylose represents the most determining component for rice GI, although not the only one [54,60,72].

In this context, a case study about GI evaluation and other biochemical properties of 25 Italian rice genotypes was developed concerning also the morphological characterization of endosperm inner structure of 36 Italian genotypes and other 18 non-Italian rice varieties for a total number of 54 genotypes. Those data could represent the starting point for future breeding programs aimed to develop rice lines

more suitable for diabetic consumers, having also the most appropriate characteristics for each destination of use.

6.2 Sample collection and data acquisition

6.2.1 Plant materials treatment

In this study a total of 54 rice varieties were evaluated from different points of view (GI, biochemical traits, inner structure of the grain). Among them 36 were Italian (Table 6.1) and 18 were of non-Italian origin (Table 6.2). These varieties were chosen to represent the different product groups, including varieties that are still cultivated on the national territory and varieties currently no longer in cultivation, and also some recently selected lines. The list of Italian varieties includes both recent and traditional varieties, with different types of grain (rice with completely crystalline or partially chalky grains, waxy rice, aromatic rice) and belonging to different product groups: Round grain, Medium grain, Long A grain for the National market, Long A grain for parboilization and Long B grain. These types were selected with the aim of representing almost all the types present on the Italian market. Many famous and appreciated Italian traditional cultivars, such as: Arborio, Baldo, Carnaroli and S. Andrea, considered among the most suitable for cooking preparations, were also included in this study.

With regard to the non-Italian varieties, the list of the considered ones was driven by two main factors: choosing varieties whose GI was already known and reported in literature and which the International Rice Research Institute (IRRI, Los Baños, Philippines), could provide us for research purposes. Those considered in this study are therefore in most cases well-known varieties, already registered and cultivated, and for this reason they present characteristics of distinguishability, uniformity and stability and show recognizable and hereditary traits.

Table 6.1. Italian rice varieties and grain type provided by Ente Nazionale Risi.

Denomination	Grain type
Arborio	Long A national market
Argo	Medium
Baldo	Long A national market
Carnaroli	Long A national market
Castelmochi	Round (waxy)
CL 71	Long B
CL12	Round
CL18	Round
CL31	Long A for parboilization
CL35	Long A for parboilization
CL388	Long A national market
CL510	Long A national market

CL80	Long B
Cripto	Medium
CRLB1	Long B
CRW3	Round (waxy)
Dedalo	Long B
Drago	Long A for parboilization
Duilio	Medium
Elio	Round
Enr-18126	Long B
Enr-18215	Round
Enr-18328	Long B
Enr-18433	Long A for parboilization
Europa	Long A for parboilization
Iarim	Long B
Italmochi	Medium (waxy)
Lince	Long A for parboilization
Padano	Medium
Pegaso	Long B
Prometeo	Round
Puma	Long A for parboilization
S. Andrea	Long A national market
Selenio	Round
Tiberio	Long A for parboilization
Valente	Long A for parboilization

Table 6.2. Rice varieties and genotype provided by International Rice Research Institute

Denomination	**Genotype ID** (https://gringlobal.irri.org)
Cisokan	IRGC 76981
Cypress	IRGC 117282
Doongara	IRGC 78392
Fedearroz 50	IRGC 117395
Hetadawee	IRGC 12084
Iac 165	IRGC 82775
IR 42	IRGC 39291
IR 4630-22-2-5-1-3	IRGC 72958
IR 50	IRGC 53433
IR 6	IRGC 51504
IR 64	IRGC 66970
Kahawanu	IRGC 36263
Kaluheenati	IRGC 7750
Mahsuri	IRGC 10929
Pajam	IRGC 25910
Sinandomeng	IRGC 78627
Swarna	IRGC 122258
Taichung Sen 17	IRGC 78210

All the rice samples analysed in this study were produced in 2020 at the Rice Research Center of Ente Nazionale Risi, located in Castello D'Agogna (27030, PV, Italy; 45° 14' 53" N, 8° 42' 00" E), using "Basic" category seed for sowing. Paddy rice samples were harvested manually at the end of the season and dried until the optimal storage humidity limit (14 %) was reached. Samples produced in the same year and in the same place were used, to minimize possible environmental effects, and thus bring out the variability linked to the differences among varieties.

Biochemical analyses and GI evaluation were carried out on 25 Italian rice genotypes, deemed the most interesting and promising among the 36 considered, using white rice samples processed at the Rice Research Center. The white rice samples obtained showed high uniformity both in milling degree and grain size and were therefore comparable to commonly marketed rice available to the final consumer. A small aliquot of the rice samples was sent to the Chelab laboratory in Resana (TV, Italy) to carry out biochemical analyses.

The study on the inner structure of the grain, carried out by FESEM on paddy rice samples, involved 54 genotypes: 36 Italian varieties and 18 genotypes belonging to the IRRI gene bank. Paddy rice grains of each variety were carefully chosen to avoid distorted or damaged grains. It should be noted that the term "paddy rice" refers to the rice grain wrapped in the husks, as it is harvested in the field. "Brown rice" refers to the rice grain after without husks, obtained from paddy rice after the dehulling process, while the term "white rice" refers to rice grains subjected to the whitening process, and thus deprived of the germ and of the outermost layers of the grain which are rich in fibres, lipids and proteins; for this reason, milled rice consists of the endosperm of the grain and is composed almost exclusively of starch.

6.2.2 Glycaemic Index evaluation

The study was conducted following the standard ISO 26642 method for determining GI (International Standards Organization, 2010). The GI was calculated evaluating the blood glucose response curve of 10 volunteers [73]. A glucose solution was used as reference food for GI evaluation. The detailed procedure followed by Rondanelli *et al.* at the Dietetic and Metabolic Unit of the Santa Margherita Institute, University of Pavia, Italy, for the glycaemic index measurements is fully available in reference [36] and is not reported in this thesis due to its limited relevance to the main topics of this PhD research study.

The GI value of the rice was calculated based on the method described by ISO 26642-2010 as the incremental area under the curve (IAUC) of a 50 g carbohydrate portion of the test food expressed as a percent of the response to the same amount of carbohydrate from a reference food taken by the same subject. The mean values for all subjects were considered the GI of a given test food. The GI values were categorized into low, medium, or high glycaemic response. The cut-off for GI

values was: ≤ 55, 56–69, ≥ 70, respectively [38]. Data were thus obtained following a widely shared and used methodology for estimating GI, whose protocol requires the determination of the glycaemic curves in ten healthy volunteers to then calculate GI based on the average values. Consequently, the volunteers themselves represent the replicates, according to this methodology.

GIs were calculated *in vivo* for each of the considered 25 Italian rice genotypes, with respect to glucose, used as standard. Ten genotypes showed a high GI, overcoming 70; in particular four of them (Arborio, Lince, Duilio and Castelmochi) exceeded a value of 85, with Arborio showing the highest value (92.31 ± 8.35). Other ten genotypes were found to show a medium GI (between 56 and 69), and five genotypes showed a low GI (≤ 55); among the last, we found the cultivars Argo and Selenio, and the Enr-18215, Enr-18328 and Enr-18433 advanced lines, with Enr-18433 and Selenio showing the lowest values (49.21 ± 6.59 and 49.15 ± 6.55 respectively).

The Carnaroli cultivar, often considered the best Italian variety for cooking preparations, was found to have a 64.17 ± 6.50 GI value, in agreement with the value previously reported (64 ± 11) referring to a commercial sample of rice belonging to the "Carnaroli" group [74].

Several studies have calculated the GI of rice varieties using different methods, both *in vivo*, on the basis of the glycaemic curves detected in volunteers subjected to tests, and "in vitro" (simulating digestion) [48,54,55,57,71]. However, currently available data are still limited compared to the thousands of rice varieties grown all over the world, especially as regards varieties cultivated in Italian and European rice area. Implementing knowledge on the GI of commercially available temperate Japonica rice varieties is essential to provide clear information to consumers, thus protecting their health.

GI values calculated in this study vary between 49 and 92, the same range reported by Fitzgerald *et al.* that analysed a wider number of genotypes [54]. Accordingly, 2 rice varieties and 3 selected lines among the 25 Japonica Italian rice genotypes analysed are therefore suitable for feeding diabetic subjects and subjects with impaired fasting glycemia, thanks to their low GI. The finding that 20 % of the analysed varieties have a low GI is an interesting result, suggesting that many of the over 250 rice varieties in the Italian National Register may also have a low GI. These results show for the first time that, even among the European temperate Japonica varieties, there are some with low GI and that this characteristic is not strictly related to a very high amylose content. In addition, a plus is given by the fact that the round grain variety Selenio, showing the lowest GI, has always been among the most cultivated varieties in Italy in recent years, despite its release dates back to 1987.

6.2.3 Biochemical analysis

Biochemical analyses on the same rice milled samples used for GI evaluation (Table 6.3) were performed in outsourcing by an accredited laboratory (Chelab, Mérieux NutriSciences Corporation, Resana, Treviso province, Italy). According to standard operation procedures the following properties were measured: protein content (g/100 g), total fat substances (g/100 g), ashes (g/100 g), carbohydrates (g/100 g), energy value (Kcal/100 g). The amylose content (g/100 g d.m.) was determined by the accredited Merceological and Chemical Laboratory of Ente Nazionale Risi, situated in the Rice Research Centre (Castello D'Agogna, 27030 PV, Italy). The analytical determinations were conducted in duplicate as indicated in ISO 6647-1:2020 protocol.

Protein content (g/100 g) was found to vary from 5.64 ± 0.35, in Puma, to 7.78 ± 0.48, in Duilio. The highest contents of total fat substances were registered in the variety Argo (also characterized by the highest average porosity and showing one of the lowest GI values), reaching 1.08 ± 0.073 g/100 g milled rice, and in the waxy variety Castelmochi, reaching 1.44 ± 0.093 g/100 g. Instead, the lowest content of total fat substances (0.57 ± 0.048 g/100 g) was registered by the Arborio variety that is also characterized by the highest GI value.

The lowest content of ashes (0.21 ± 0.04 g/100 g milled rice) was observed in the Selenio variety (showing the lowest GI), followed by Enr-18126 selected line (0.22 ± 0.04 g/100 g), Lince (0.34 ± 0.04 g/100 g) and Enr-18328 (0.35 ± 0.04); hence, three of four varieties characterized by a low ashes content, are also characterized by a low – medium GI. The highest ashes content was observed in the variety Valente (1.08 ± 0.07 g/100 g), followed by the waxy variety Castelmochi (0.66 ± 0.05 g/100 g), and by Puma (0.61 ± 0.05 g/100 g), Padano (0.6 ± 0.05 g/100 g), Duilio (0.6 ± 0.05 g/100 g), CL35 (0.6 ± 0.05 g/100 g), Iarim (0.59 ± 0.05 g/100 g) and CL71 (0.58 ± 0.05 g/100 g), all characterized by a high GI value with the only exceptions of Iarim and Valente.

Although at most of the low-GI varieties known in literature corresponds a high amylose content [54,71], the presented results suggest that this is not necessarily a discriminating trait, since the Selenio variety, for example, showed a low GI value despite having a medium-low amylose content, thus signifying that many other factors could be involved in determining GI. Other factors could also be involved in GI variations: for example, Indica non-waxy starch has been reported to contain about three times more protein bodies than Japonica rice starch having a similar amylose content [75].

Table 6.3. Glycaemic Index (GI) evaluation and biochemical analysis results of 25 Italian rice genotypes. Values ± standard deviation are reported.

Variety	Protein Content (g/100 g)	Ashes (g/100 g)	Total fat substances (g/100 g)	Amylose content (%)	GI index (% glucose)	GI content
Arborio	6.7 ± 0.4	0.4 ± 0.0	0.5 ± 0.0	14.1 ± 3.0	92.3 ± 8.35	High
Lince	5.9 ± 0.3	0.3 ± 0.0	0.6 ± 0.0	17.5 ± 3.5	88.9 ± 9.22	High
Duilio	7.7 ± 0.4	0.6 ± 0.0	0.7 ± 0.0	9.6 ± 2.3	86.2 ± 10.1	High
Castelmochi	7.5 ± 0.4	0.6 ± 0.0	1.4 ± 0.0	< 5	84.7 ± 10.6	High
Padano	6.6 ± 0.4	0.6 ± 0.0	0.9 ± 0.0	12.7 ± 2.8	73.6 ± 10.6	High
CL18	6.8 ± 0.4	0.5 ± 0.0	0.8 ± 0.0	11.6 ± 2.6	73.0 ± 8.17	High
Puma	5.6 ± 0.3	0.6 ± 0.0	0.6 ± 0.0	14.0 ± 3.0	73.0 ± 7.05	High
Baldo	7.2 ± 0.4	0.4 ± 0.0	0.7 ± 0.0	14.2 ± 3.0	71.4 ± 6.44	High
CL71	7.4 ± 0.4	0.5 ± 0.0	0.9 ± 0.0	20.7 ± 4.0	71.2 ± 7.14	High
CL35	5.7 ± 0.3	0.6 ± 0.0	0.9 ± 0.0	14.4 ± 3.0	71.0 ± 9.76	High
CL12	6.7 ± 0.4	0.5 ± 0.0	0.9 ± 0.0	13.0 ± 2.8	68.6 ± 7.03	Medium
S. Andrea	5.9 ± 0.3	0.5 ± 0.0	0.7 ± 0.0	15.2 ± 3.1	66.4 ± 7.64	Medium
Valente	7.1 ± 0.4	1.0 ± 0.0	1.7 ± 0.0	12.1 ± 2.7	66.1 ± 6.86	Medium
Carnaroli	6.6 ± 0.4	0.4 ± 0.0	0.7 ± 0.0	20.7 ± 4.0	64.1 ± 6.50	Medium
CL388	7.4 ± 0.4	0.4 ± 0.0	0.6 ± 0.0	12.6 ± 2.7	62.5 ± 8.67	Medium
Tiberio	6.6 ± 0.4	0.5 ± 0.0	0.9 ± 0.0	23.9 ± 4.5	61.7 ± 5.99	Medium
CRLB1	7.5 ± 0.4	0.5 ± 0.0	0.7 ± 0.0	21.8 ± 4.1	61.0 ± 3.73	Medium
Elio	5.9 ± 0.3	0.4 ± 0.0	0.7 ± 0.0	22.9 ± 4.3	60.3 ± 5.87	Medium
Enr-18126	7.3 ± 0.4	0.2 ± 0.0	0.7 ± 0.0	17.7 ± 3.5	58.4 ± 5.83	Medium
Iarim	7.3 ± 0.4	0.5 ± 0.0	0.9 ± 0.0	24.4 ± 4.5	58.0 ± 9.29	Medium
Enr-18215	6.6 ± 0.4	0.4 ± 0.0	0.8 ± 0.0	12.0 ± 2.7	12.0 ± 6.79	Low
Enr-18328	6.9 ± 0.4	0.3 ± 0.0	0.7 ± 0.0	23.4 ± 4.4	23.4 ± 5.01	Low
Argo	7.3 ± 0.4	0.4 ± 0.0	1.0 ± 0.0	20.3 ± 3.9	20.3 ± 7.17	Low
Enr-18433	6.0 ± 0.3	0.5 ± 0.0	0.8 ± 0.0	18.7 ± 3.7	18.7 ± 5.59	Low
Selenio	6.3 ± 0.3	0.2 ± 0.0	0.6 ± 0.0	14.6 ± 3.0	14.6 ± 6.55	Low

6.2.4 FESEM analysis

The paddy rice samples of 54 different varieties were analysed by FESEM to observe the characteristics of their internal structure, determining starch granules dimensions and the porosity of the structure, understood as the percentage of empty spaces area among the starch granules. To this aim, the rice samples of 36 varieties and selected lines (Table 6.1), together with 18 rice samples of IRRI genotypes (Table 6.2), were prepared as follows.

Rice grains were cut longitudinally (for the entire length) by incision of the surface with a scalpel. Since rice is a non-conductive material, the halves of the grains were then fixed on a microscope stub with conductive tape and subsequently metallized with a thin layer of Pt by DC sputtering in Ar atmosphere (Q150T-ES, Quorun Technologies) with a sputtering current of 30 mA for 25 s.

A Zeiss SUPRA 40 (Zeiss SMT, Oberkoechen, Germany) FESEM, equipped with a detector for secondary electrons and a detector for backscattered electrons, was

used to acquire the images. Micrographs were acquired with an accelerating voltage of 5 kV and a 30 µm opening using a secondary electron detector. After fine-tuning the sample preparation technique, 3 grains for each variety were analysed to obtain 4 images for each grain (1 at 100× and 3 at 5000× magnification). The 5000× images were acquired in the most representative areas of the grain section, chosen after carefully observing the 100× images, specifically looking for regions in which the cut did not alter the starch granule structure, to better capture and describe the granules' characteristics and arrangement.

6.3 Algorithm development

An ad hoc algorithm was simultaneously developed for image analysis and processing, using the MATLAB software development environment (version R2021b, Mathworks, Natick, MA, USA). To better explain how the algorithm works, a supporting flowchart containing all the image analysis steps developed for processing the images can be seen in Figure 6.1. The algorithm is designed to perform a fully automated analysis of the provided images, but requires five parameters that the user needs to set before starting the analysis and that will be applied to all the images in the study.

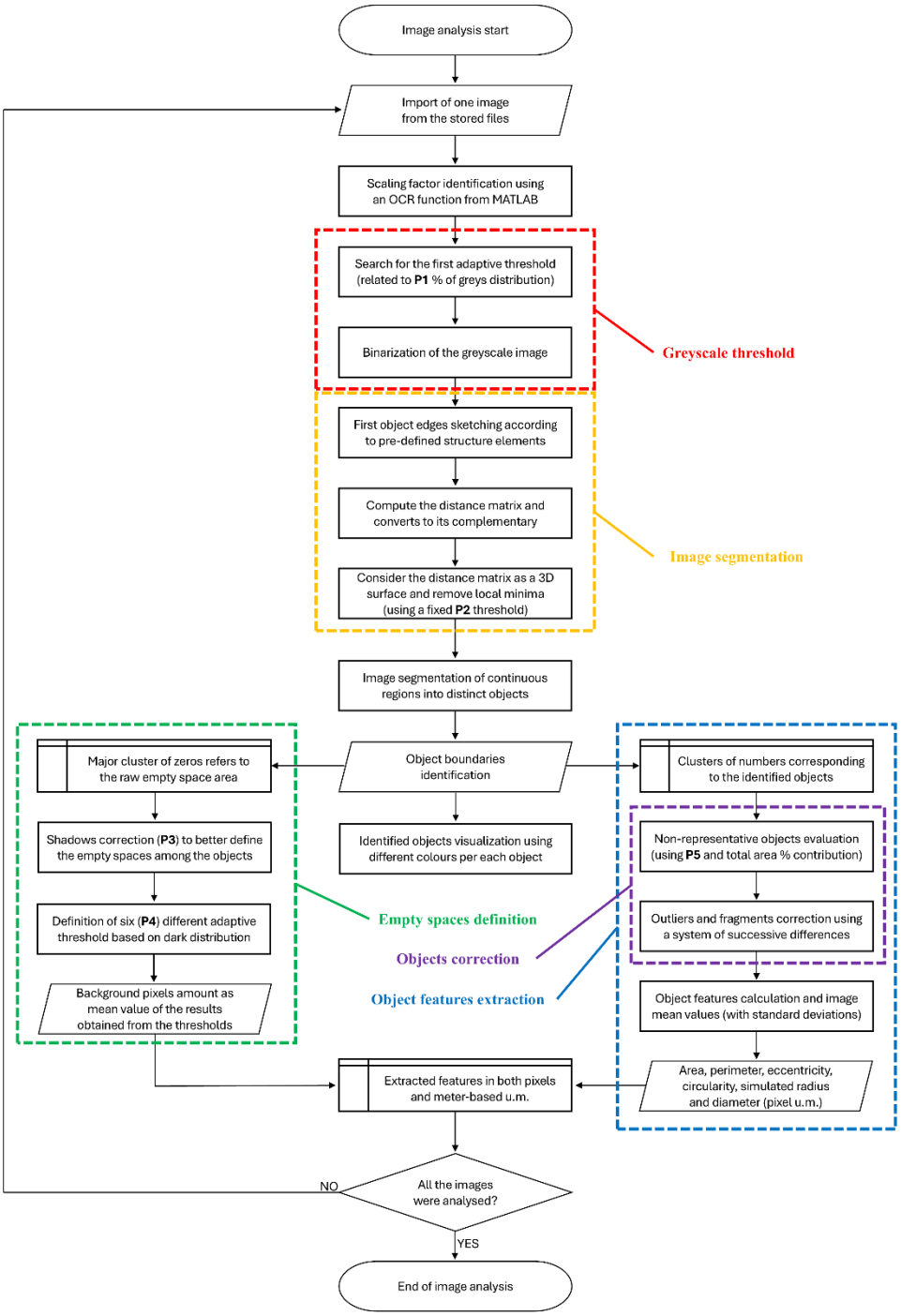

Figure 6.1. Flowchart summarizing the algorithm functioning. The red box refers to the definition of the grey distribution threshold. The orange box represents the image segmentation steps. The green box corresponds to the definition of empty spaces. The blue box is focused on the object features extraction and contains also the purple box which is focus on the optional object correction step.

6.3.1 Image processing

Scaling factor

To work with the images using the pixels as measure units instead of meter-based units, a scaling factor needs to be defined. To extract such factor directly from the image, a piece of code based on an OCR function able to read the instrumental information printed in the image itself was developed. Pairing this scaling factor with each image becomes fundamental to refer the automatically measured morphological features back to a meter-based reference. The scaling factor is calculated as the ratio between the extracted meter-based unit number (e.g., 10 µm) and the length of the scale reference bar, expressed as the corresponding number of pixels (Figure 6.2).

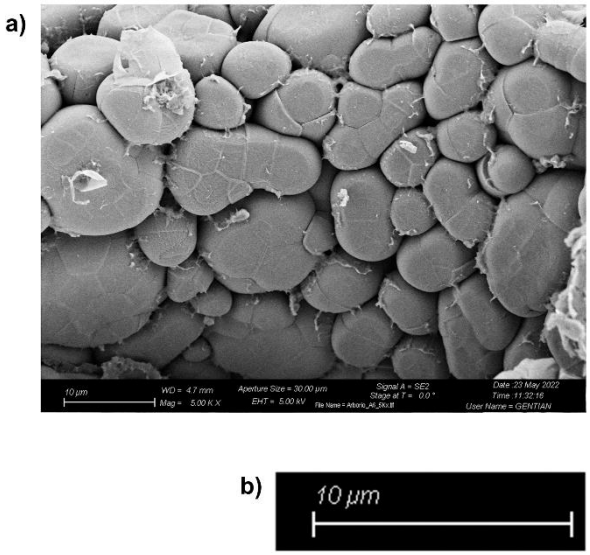

Figure 6.2. a) example of a FESEM image containing also the instrumental information. b) Focus on the scale reference bar and on the meter-based unit number used to define the scaling factor in pixels. These outputs are obtained from the third step of the flowchart reported in Figure 6.1.

Size and resolution check

After storing the image's instrumental information, the image sizes and resolutions must be checked: to compare different images, sizes and resolutions need to be made consistent. First, a "*resize*" function was used to ensure that all analysed images would have the same pixel dimensions, thus being directly comparable among each other. This step was coded to work with inputs either fixed, manually selected or automatically derived from an image chosen as a reference.

Brightness adjustment

Once all images have been made consistent, the first image editing step entails brightness correction. FESEM images are acquired in grayscale and, depending on the instrumental set-up, they could globally be darker or brighter, also within the same sample scanned area. MATLAB offers a series of functions able to adjust the brightness to facilitate the comparison among images. Simple functions like *imlocalbrighten* just brighten low-light areas, while more complex ones allow to use the greyscale distribution and the brightness of a selected image as a reference or allow manual tuning the light contrast and the brightness.

Description of the five manual inputs

The script performs fast analysis of large number of images, but, because of the complexity of the task, it is only able to process one image at a time. Since the algorithm requires some inputs to be manually selected at the beginning of the process, to better understand in which part of the algorithm these five inputs (or parameters) are applied along the manuscript. The parameters were named P1, P2, P3, P4 and P5. Their purpose is explained in the following paragraphs where the values shown are the one optimized for a rice kernels study recently published [36].

Adaptive threshold definition

One of the first critical steps of the algorithm, that allows to start defining the object edges, is the choice of a threshold to distinguish the dark grey shades of the empty spaces among the objects from the lighter grey shades belonging to the objects to be investigated. The idea behind using a simple threshold is based on the fact that, with objects depicted in a bidimensional manner (as a FESEM image, for instance) and lit from above, the empty spaces among them would be mostly corresponding to darker areas (i.e., areas not reflecting electrons). To make the algorithm versatile for a wide variety of applications, an adaptive threshold instead of a fixed one was developed and applied (Figure 6.1 – Red box).

Image segmentation: the binarization step

Image binarization is a crucial step for both the definition of the empty spaces and the identification of the particles from the image. This step is aimed at dividing the pixels of the image roughly into two sets: the first set refers to the empty spaces while the second refers to the particles. The second set is further processed to obtain the refined identifications of both the empty spaces and the particles.

In our algorithm, binarization is performed based on the grey distribution of the image. An adaptive threshold is applied to this distribution: all pixels whose value

is below the threshold constitute the initial empty space areas of the image, while all pixels whose value is above the threshold are identified as the raw areas containing the particles ("raw" because in the next step the measured areas will be better refined through different correction steps). The value of P1 regulates the shade of grey to which the threshold is going to be set. The threshold is adaptive since it is defined as a "percentage of dark grey" and it is applied individually to the grey distributions of each image. In this way, the same percentage threshold value would correspond to a different shade of dark grey from image to image, and therefore potential differences in brightness can be mitigated across the set of images. The threshold computation is based on a cumulative sum over the pixels from the darker end of the distribution, up to the desired threshold value. For our real case of rice images, the threshold value P1 was set to 17 %, but of course for other applications the threshold can be set according to needs.

Because of the different shapes and the tridimensionality of the sample structure, the set of pixels corresponding to the particles needs to be further refined to clearly distinguish and correctly defining the shapes of the particles (Figure 6.1 – Yellow box). First, a smoothing function (*activecontour*) allows to sketch the object boundaries using the active contour algorithm, that superimposes the binarized image over the original one. After that, the first objects identification (Figure 6.3) is performed: using the *strel* function, the script creates a structure element with a user-selected shape and applies it with the *imclose* function. The result of this procedure is a new binarized image where a first correction of the areas recognized as empty spaces and as particles is performed. This correction step prepares the image for the following segmentation step. Depending on the type and nature of the objects described by the images, the user can manually choose the approximate object shape to be used by the *strel* function. For example, in the case of round-shaped objects the best results can be obtained by using the 'disk' option; instead, in the case of more squared ones, the 'rectangle' or the 'square' options prove better suited, whereas for linear and elongated ones the 'line' option can be chosen.

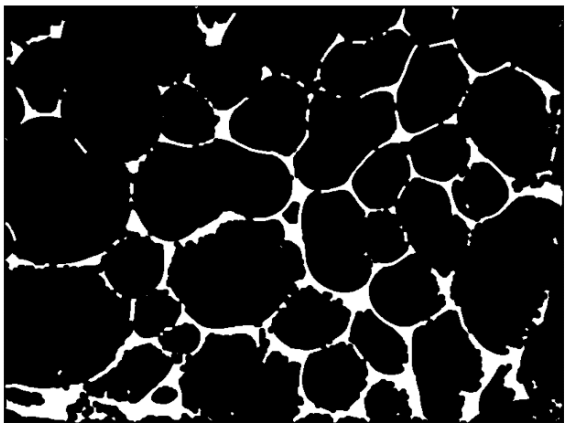

Figure 6.3. Binarized image obtained after the application of the user-selected shape structure elements applied after the binarization step: in black the pixels belonging to the raw particles set, in white the pixels belonging to the empty spaces set. This output is obtained from the red box part of the flowchart reported in Figure 6.1.

Image segmentation: finding the distinct objects

The binarized image contains an initial rough description of the particles of the image, and it needs to be further refined to actually identify the distinct objects. To do so, a distance transform matrix (based on Euclidean distance) is computed using the *bwdist* function. This function computes, for each pixel, its distance from the nearest "non-assigned" pixel (i.e., the parts in white of Figure 6.3, not recognized as objects). As a result, a numerical matrix ("distance matrix") with the same dimensions of the original image is obtained and can be interpreted as follows: the values that are far from zero correspond to the centre of the objects, while the values become closer to zero moving towards the boundaries of the object, and become equal to zero when pixels represent the empty spaces among the objects (Figure 6.4a).

To clearly identify the edges of the particles, the complementary of the previously obtained "distance matrix" is computed by subtracting the value of each pixel from the absolute maximum value of the distance transform matrix in order to facilitate the next steps of the algorithm functioning (Figure 6.4b). This matrix is then treated like a three-dimensional surface where light pixels, represented by "surface minima", correspond to objects, while dark pixels correspond to empty spaces. Then, by applying the *imhmin* function on the image, all the minima in the matrix whose distance to zero is less than a fixed optimized value (**P2 input**) are suppressed to better define boundaries. After that, the final image segmentation is obtained using the *watershed* transform, which is used to segment continuous regions of interest into distinct objects, again by treating them like a three-dimensional surface (Figure 6.5).

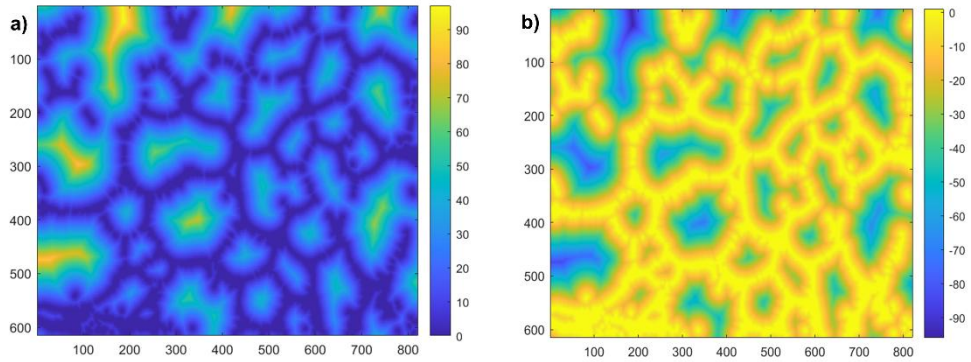

Figure 6.4. a) Image representation of the distance transform matrix computed using the Euclidean distance. b) Image representation of the complementary of the distance transform matrix. These outputs are obtained from the second step of the yellow box part of the flowchart reported in Figure 6.1.

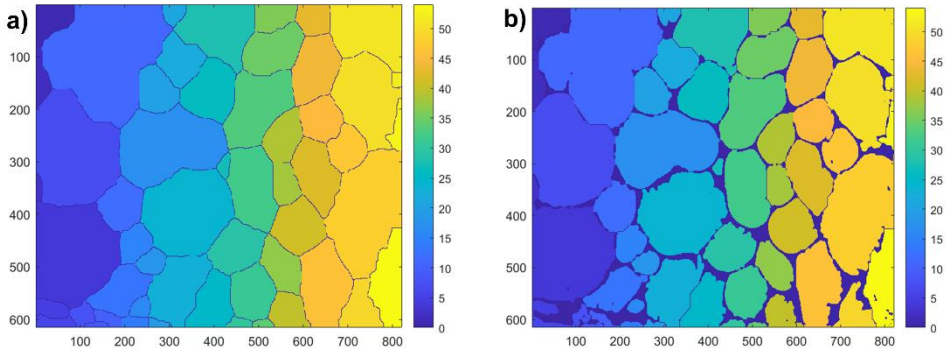

Figure 6.5. a) Representation of the first segmentation of the continuous region of interest into distinct objects. b) Image segmentation obtained after recognizing the empty spaces among the particles. These outputs are obtained from the three steps executed after the yellow box part of the flowchart reported in Figure 6.1.

This approach allows to obtain an image where the identified objects are completely distinguished from the empty spaces among them. In addition, with the *labeloverlay* function, it is also possible to visualize the identified objects by highlighting them with different colours in contrast to the black-coloured empty spaces. Figure 6.6 shows an example of such visualization.

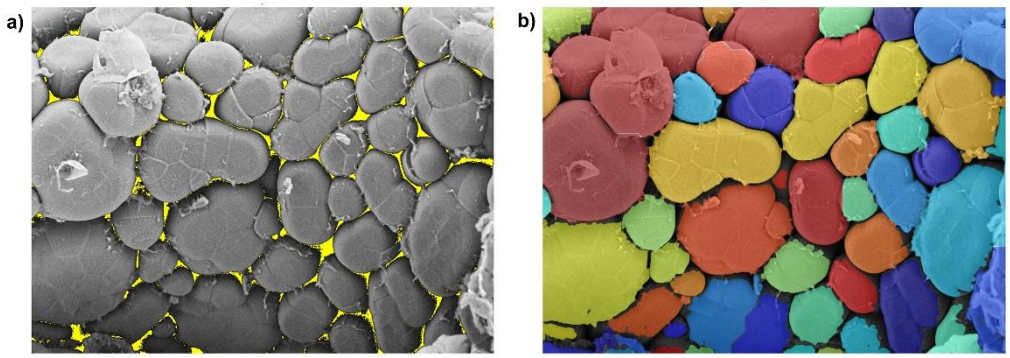

Figure 6.6. Images produced by the algorithm during the feature extraction process. a) Representation of the empty spaces (in yellow) found in the image that correspond to the sample porosity. b) Representation of the starch particles identified by the algorithm and coloured differently to visualize the result obtained before the objects correction step.

6.3.2 Morphological features calculation

The image-segmentation step produces a list of objects in which each object is described by the pixels belonging to the original image. All objects are from now on treated independently by the algorithm, to calculate their morphological features. The script allows the user to choose which features the algorithm will extract from the images and, more importantly, also to select either if the interest is focused only on the objects' features or if also the empty spaces among the objects must be computed.

Empty spaces definition

The first morphological feature that the algorithm can compute is the area of the space non-assigned as object (i.e., non-object space), that in our rice case study corresponds to the amount of empty spaces among the objects. Since the analysed samples are three-dimensional, the images represent a non-planar surface characterized by holes, cavities and superpositions that, all combined, generate shadows. In greyscale images, the darkest shadows can be easily confused with the dark areas corresponding to the "non-object space". To better estimate the real amount of empty space, an algorithmic step involving a set of adaptive thresholds was developed to process the marginal shadows, with a "soft approach" (Figure 6.1 – Green box).

Working on the "non-object" pixels greyscale values, an initial selection is performed removing all pixels lighter than a chosen adaptive threshold (**P3 input**): those pixels correspond to "lighter shadows". The remaining darker pixels are then processed by a piece of code which calculates the mean value (M) and the standard deviation (S) of their distribution. Then, six different increasing adaptive thresholds are applied to the distribution (the number of thresholds is the value of P4), by addition to the mean value M. The spacing of these increasing thresholds is

regulated by a coefficient (K) ranging from 1 to 6 (the maximum value of K corresponds to P4). The thresholds are calculated with Equation 6.1:

$$thr_{(K)} = M + ((K - 1) * 0.1 * S) \qquad (6.1)$$

For each threshold value a different estimation of the empty space area is obtained. These areas correspond to the number of pixels with greyscale value lower than the corresponding threshold. The final number of pixels identified as the image "non-object area" is the mean of the six areas obtained with the different thresholds. The amount of empty spaces is finally expressed both in terms of area (in pixel) and in terms of percentage with respect to the total image area (Figure 6.7).

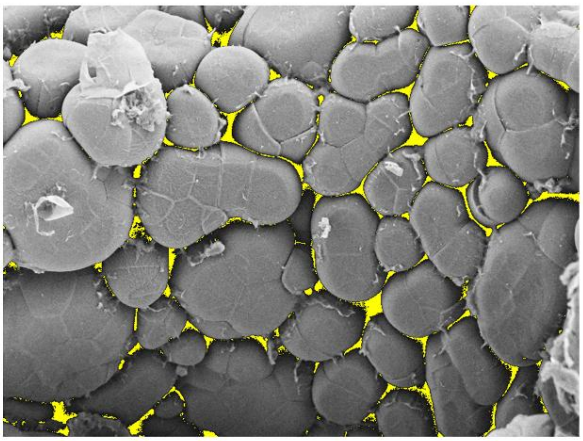

Figure 6.7. Visualization of the identified empty spaces among the particles. The empty spaces found are coloured in yellow. This output is obtained from the green box part of the flowchart reported in Figure 6.1.

Object features calculation

Before proceeding with the objects' features calculation (Figure 6.1 – Blue box), a step of object quality assessment is performed. The aim is to recognize and remove most of the defective or incomplete ones (e.g., objects cut by image edges or partially hidden because of shadows and overlaps), so that the morphological features can be computed only from the most representative objects, concerning the real situation captured by the images (Figure 6.1 – Purple box).

Two consecutive steps are applied to remove the non-representative objects: in the first one the percentage contribution of each object's area to the total area is computed and used to separate larger objects from smaller ones; the second step is operated only on the set of small objects defined in the first step, and it is based on the computation of subsequent differences of the particles' dimensions aimed at finding a significant leap, ideally separating the fragments and the smallest objects. More in detail, the first step starts calculating the area of all objects found by segmentation; then, all computed areas are sorted from the smallest to the largest and the percentage contribution to the total object area is computed. The separation

between small and large objects is then performed selecting, from the smallest value up, all objects until the cumulative percentage contributions sum reaches 8 % (**P5 input**) of the total object area. The second step is then performed calculating the difference between the areas of two consecutive sorted objects, considering only the previously defined set of smallest objects. This set is finally divided into "small objects" and "fragments" (or "defective" objects) and the threshold between them is fixed at the highest computed difference (i.e. the significant leap). All objects below this threshold are then excluded from the morphological features calculation (Figure 6.8).

Once this correction is performed, each object can be processed to extract the desired morphological features, the most common being the area (number of pixels composing the object), and the perimeter (number of pixels of the object's boundary). Starting from these two features the values of radius and diameter can then be derived assuming a perfect circular shape object.

An additional feature that can be requested by the user is eccentricity, a property commonly used for describing elliptical objects: it describes how different the object's shape is from a perfect circle (0 in the case of a perfect circle, and 1 in the case of the degeneration of the shape to a segment). Its mathematical definition is described in Equation 6.2 and corresponds to the ratio between the distance of the focus from the centre and the length of the semimajor axis (a).

$$E = \sqrt{a^2 - b^2}/a \qquad (6.2)$$

In addition, also circularity can be obtained, again related to the object similarity to a perfect circle, but it is based on the area (A) and perimeter (p) values as described by Equation 6.3:

$$C = 4\pi A/p^2 \qquad (6.3)$$

All requested morphological values are calculated and stored for each object in the image, along with the image mean value of each feature and the corresponding standard deviation, as more general information related to the whole image. The last step of the features extraction process consists in the application of the previously identified scaling factor, to convert the features values from pixels unit to a meter-based unit of measurement.

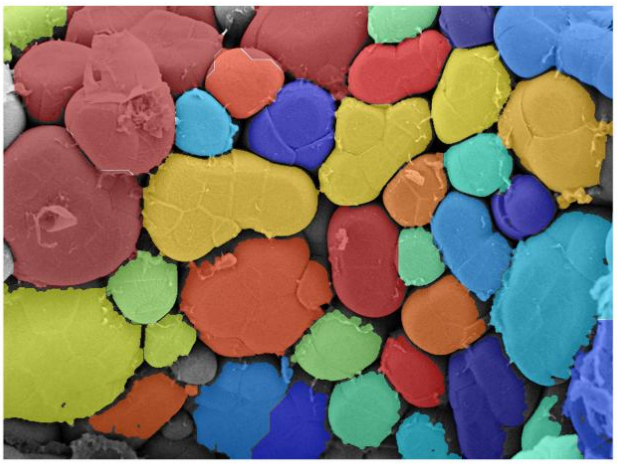

Figure 6.8. Visualization of the objects considered after the object correction step in order to remove the small fragments from the found objects. This output is obtained from the violet box part of the flowchart reported in Figure 6.1.

6.4 Case study results

The algorithm described in Section 6.3 allows the recognition of the areas represented by the starch granules and the gaps among them, distinguishing those that are actual empty spaces from those that are shadows created by the image capture perspective. The average percentage of porosity of the grain was thus automatically calculated from the obtained images together with the relative standard deviation. The average size of the starch granules (μm^2) was calculated based on the number of pixels of each grain, and also for this parameter the standard deviation was calculated.

To estimate the porosity percentage of the endosperm and the average area of the starch granules, the mean of the values obtained from the three 5000X images of each grain and of the 3 grains analysed for each variety was calculated. Furthermore, since the chalky and crystalline fractions of the grain, presenting completely diverse morphological characteristics, have different extents, a weighted average was calculated for each variety, which considers the extent of the chalky part of the endosperm, as a percentage.

6.4.1 Algorithm outputs evaluation

After the application of the image analysis algorithm to the rice kernels analysed as a case study, starch granules diameters were estimated from their average area, considering them approximately round, finding values between 5.83 µm (in Prometeo and Pajam), and 12.53 µm (in Lince). The average percentage of endosperm porosity resulted greatly variable, showing the lowest values in CL80

(0.40 ± 0.26) and Selenio (0.40 ± 0.28), and the highest in Argo (6.72 ± 0.39). The peculiar endosperm structure of some varieties made it impossible to calculate starch granules size.

Once the varieties characterized by a crystalline (or almost crystalline) endosperm (Figure 6.9a, 6.9b, 6.9c) were separated from those whose average chalky fraction was estimated equal or higher than 10 % (Figure 6.9d, 6.9e, 6.9f), we observed that the crystalline ones seemed to show a lower average percentage of porosity (2.65 ± 2.26) and a larger average dimension of starch granules (62.39 ± 27.33 μm^2, corresponding to 8.7 µm in diameter). Instead, the ones having a more extent chalky fraction of the grain, showed a higher average percentage of porosity (5.15 ± 1.53) and a smaller average starch granules size (45.23 ± 15.01 μm^2, corresponding to 7.5 µm in diameter).

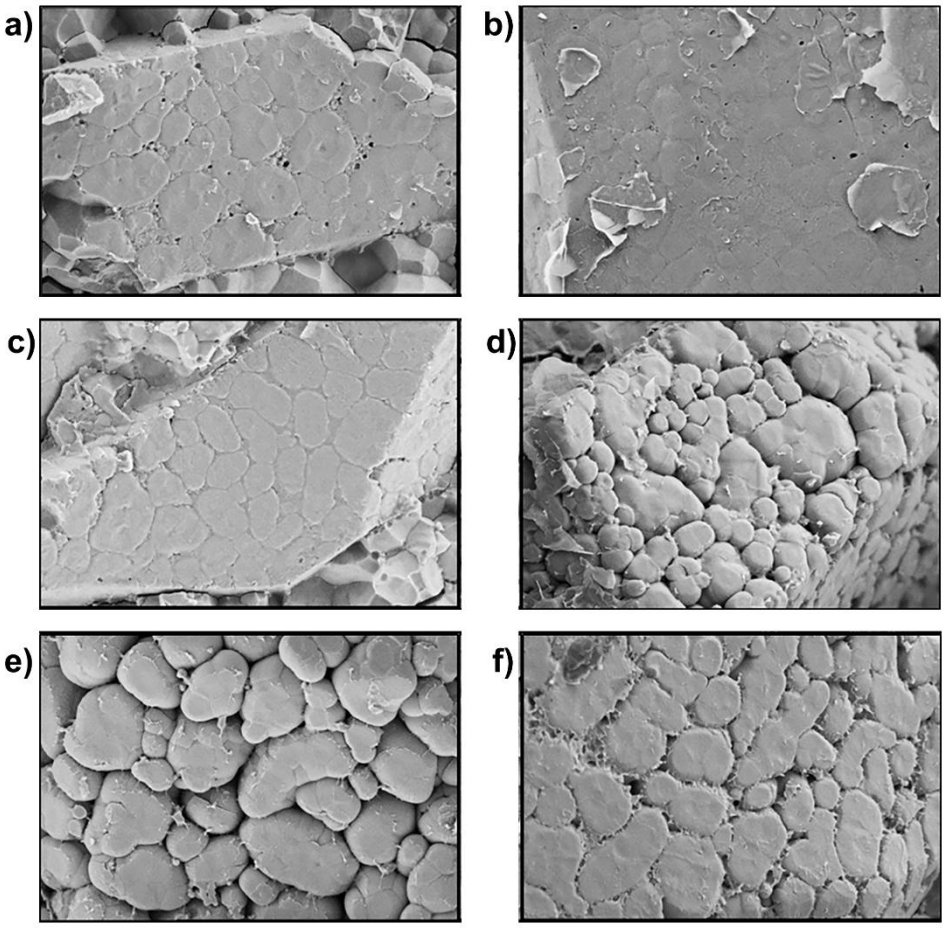

Figure 6.9. Micrographs obtained at a 5000X magnification from grains of crystalline rice varieties Selenio (A), Valente (B) and CL80 (C), and chalky rice varieties Argo (D), Carnaroli (E) and CL388 (F).

No differences in average eccentricity (0.72) and average circularity (0.4) of starch granules were observed between the two groups. Nevertheless, at a visual observation of micrographs, the starch granules of the varieties with a more

compact structure and with fewer empty spaces appear generally to show a more polyhedral shape, with sharp edges, while those with greater porosity present granules with a less polygonal shape. As regards the abovementioned subdivision of the varieties between crystalline (or almost crystalline) and those with a chalky fraction greater than 10 %, the average GI values resulted similar for the two groups: respectively 64.39 ± 12.48 and 68.39 ± 10.87.

Concerning the waxy rice varieties Castelmochi, Italmochi and CRW3, both Castelmochi and Italmochi were characterized by a low average percentage of porosity (respectively 0.83 ± 0.23 and 1.09 ± 0.29), showing a peculiar endosperm structure where starch granules appeared almost indistinguishable and fused together in an uneven mass, which made it impossible to automatically determine their size (Figure 6.10a). On the contrary, CRW3 showed a completely different structure, with a higher percentage of porosity, equal to 6.35 ± 0.54, comparable to that of many varieties with a chalky endosperm, and its starch granules appeared more easier distinguishable, having an average dimension of 35.61 ± 18.79 μm^2 (6.74 μm in diameter) (Figure 6.10b). Even if starch granules morphology was not determined for the variety Castelmochi, starch granules of Italmochi and CRW3 showed average circularity values comparable to those observed in the non-waxy varieties (0.41 and 0.40 respectively), but the values of average eccentricity (0.68 for both) resulted the lowest observed.

A peculiar structure, clearly different from all the others, was observed in micrographs obtained from a grain of the variety Dedalo at 5000X magnification, with the presence of big roundish and smooth granules and many smaller round structures (Figure 6.10c).

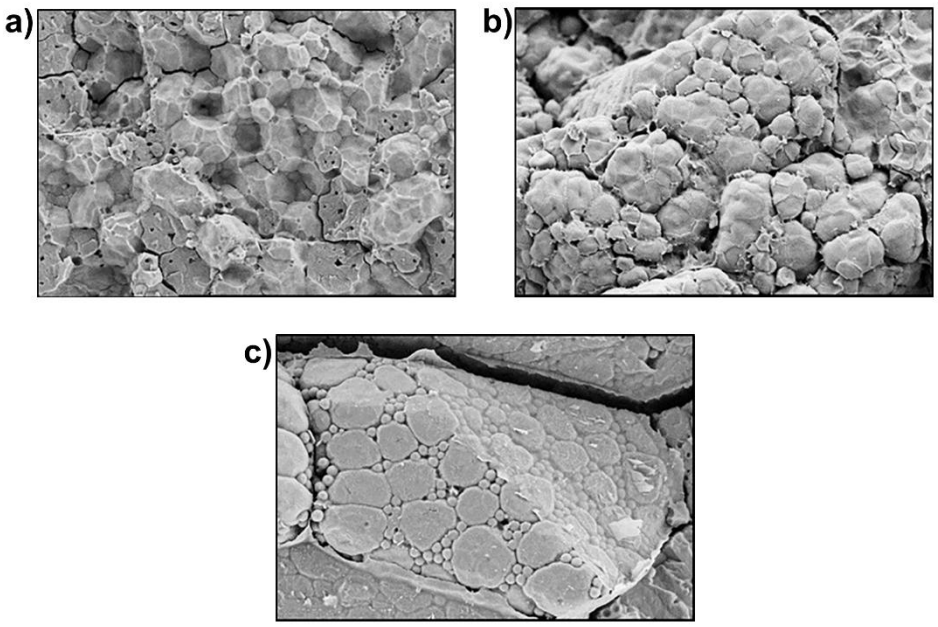

Figure 6.10. Micrographs obtained at a 5000X magnification from grains of rice varieties characterized by an inner structure different from all the others observed: Italmochi (A), CRW3 (B) and Dedalo (C).

The observation of the grain structure gave us information concerning the existing variability among genotypes in starch granules morphology and distribution, as well as in endosperm average porosity (Table 6.4). Hence, this study adds another brick to the knowledge of rice grain microstructure by assessing the existence of relevant differences in morphological features among different rice genotypes.

Table 6.4. Characterization of 11 Italian rice varieties and 18 IRRI genotypes through FESEM imaging. Values ± standard deviation are reported when possible.

Variety	Area (μm^2)	Eccentricity (0 < E < 1)	Circularity (0 < C < 1)	Diameter (μm^2)	Porosity (%)	Origin
CL31	34.48 ± 10.37	0.72	0.39	6.63	5.64 ± 1.85	Italy
CL80	/	0.73	0.38	/	0.40 ± 0.26	Italy
CL510	50.55 ± 0.65	0.74	0.39	8.03	5.30 ± 0.71	Italy
Cripto	33.24 ± 2.15	/	/	6.51	6.15 ± 0.59	Italy
CRW3	35.61 ± 18.79	0.68	0.40	6.74	6.35 ± 0.54	Italy
Dedalo	64.81 ± 25.09	0.73	0.38	9.09	2.14 ± 2.86	Italy
Drago	75.02 ± 32.28	0.71	0.39	9.78	6.35 ± 0.20	Italy
Europa	30.75 ± 1.26	0.73	0.38	6.26	6.28 ± 0.27	Italy
Italmochi	/	0.68	0.41	/	1.09 ± 0.29	Italy
Pegaso	69.83 ± 38.27	0.71	0.41	9.43	1.33 ± 1.41	Italy
Prometeo	26.69 ± 0.46	0.71	0.41	5.83	0.92 ± 0.22	Italy
Kaluheenati	29.39 ± 4.86	0.74	0.37	6.12	5.66 ± 1.52	IRRI
Mashuri	28.22 ± 7.01	/	/	6.00	5.74 ± 1.10	IRRI
Hetadawee	53.45 ± 13.04	/	/	8.25	6.52 ± 0.67	IRRI
Pajam	26.70 ± 0.72	0.73	0.40	5.83	5.97 ± 0.24	IRRI
Kahawanu	/	0.72	0.36	/	1.66 ± 0.61	IRRI
IR42	54.75 ± 19.47	0.74	0.40	8.35	0.69 ± 0.13	IRRI
IR6	66.95 ± 41.65	0.73	0.42	9.24	1.64 ± 1.87	IRRI
IR50	87.63 ± 9.13	0.74	0.40	10.57	6.31 ± 0.34	IRRI
IR64	75.89 ± 41.41	0.73	0.39	9.83	4.16 ± 3.52	IRRI
IR4630-22…	21.04 ± 11.69	0.72	0.40	5.18	5.18 ± 0.55	IRRI
Cisokan	/	0.72	0.40	/	0.52 ± 0.01	IRRI
TaichungSen	32.22 ± 1.77	0.71	0.39	6.41	6.13 ± 0.19	IRRI
Doongara	29.00 ± 2.17	0.71	0.39	6.08	6.36 ± 0.05	IRRI
Sinandomeng	/	/	/	/	0.99 ± 0.80	IRRI
Iac 165	52.87 ± 29.26	0.74	0.39	8.21	4.27 ± 3.14	IRRI
Fedearroz 50	89.17 ± 1.99	0.72	0.42	10.66	0.79 ± 0.39	IRRI
Cypress	84.12 ± 7.16	0.71	0.39	10.35	0.89 ± 0.21	IRRI
Swarna	49.62 ± 29.61	/	/	7.95	4.75 ± 3.37	IRRI

By comparing micrographs obtained from different samples it is possible to observe compound starch granules, appearing as rounded structures, made up of "smaller wedges". Small starch granules, composing the amyloplast, are separated by cross-wall (inner envelope membrane) and each compound starch granule is enveloped by an outer membrane, at whose synthesis contribute plastid division proteins, having a critical role also in amyloplast division [76]. These membranes were found to be visible by FESEM in some of the micrographs obtained at a 5000X magnification. In some cases, these membranes were more evident and create a sort of film over the starch granules which made it difficult to distinguish them, especially in varieties having a more compact structure.

The average diameter of starch granules, estimated from their average size (area) considering them approximately as round, resulted 8.15 ± 1.78 μm, in agreement with measurements made by Jane *at al.* (1994) on purified rice starch, who found polygonal shaped granules with diameters of 3–8 μm, much smaller than those from maize (5–20 μm) and other cereal sources [66]. On the other hand, the extraction of morphological parameters of starch granules resulted impossible for some genotypes: nine rice varieties showed an extremely compact structure, and the starch granules resulted indistinguishable, making it impossible to calculate their average size. It was however still possible to obtain porosity values (very low) for these varieties.

The characterized Italian varieties can be broadly distinguished in three main groups, based on their average porosity: a group of varieties characterized by a very low porosity value (< 1.7 %), including mainly those with a crystalline endosperm (Figure 6.9a, 6.9b, 6.9c).; a second group characterized by a high porosity value (> 5 %) including the most of Long A varieties for the National market among those considered (e.g., Arborio, Carnaroli, Padano, S.Andrea and the more recently released variety CL388) (Figure 6.9d, 6.9e, 6.9f); and a third group with intermediate porosity values between 2.1 % and 4.2 % including the Baldo variety and other more recent varieties with intermediate characteristics. Porosity seems to be correlated with the extension of the chalky fraction of the endosperm and with grain width in Italian genotypes, thus confirming being a main characteristic of Long A grain varieties for cooking preparations.

The structure of the chalky area of the endosperm, typical characteristic of many Italian Short and Long A grain varieties for national consumption, was generally more regular and better defined, with well separated starch granules, thus resulting more adapt to an automatic processing of the images, if compared to not chalky areas (Figure 6.9d, 6.9e, 6.9f). The crystalline part of the endosperm, which in some cases represents the entire grain section, appears morphologically completely different, with starch granules joined together into larger structures (Figure 6.9a, 6.9b, 6.9c).

It is known that starch granules development begins from the innermost cells of the endosperm and spreads to the outer cells. Hence, the peripheral cells, located near the aleurone layer, are the last to be filled during grain development, approximately one month after the process begins [64,77]. Matsushima *et al.* suggested that the molecular mechanisms responsible for starch granules formation could change during rice caryopsis filling, giving rise to differences in starch granule morphology and distribution between the inner and outer parts of the grain [64]. Even if those mechanisms are still unknow, this could explain the formation of the chalky part of the endosperm, which characterizes many of the Italian rice varieties.

Italian rice varieties are characterized by the ability to absorb seasonings and release starch during cooking; this specific characteristic could be linked to a less compact internal structure of the grains, being characterized by a higher average endosperm

porosity with more distant starch granules having a smoother surface, particularly in the chalky portion of the grain, compared to varieties characterized by a crystalline endosperm, such as those most appreciated in northern-European countries, which show a compact structure with larger starch granules and a reduced average porosity.

No differences were observed in the average eccentricity of the starch granules (0.72), nor in the mean circularity (0.4) between crystalline (or almost crystalline) varieties and those with a larger chalky fraction. The average amylose content, as well as GI values, resulted similar for the two groups, indicating that these differences in grain inner structure are not primarily due to the total amylose content. In rice amylose extender mutants (characterized by a particularly high amylose content, up to 40 %), starch granules have been reported to be small, round and loosely packed [78]; these characteristics are similar to those observed in the chalky fraction of the grain of Italian rice varieties. So, despite the overall content of the grain, the chalky fraction of the endosperm could contain a higher amount of amylose than the outer part. As a consequence, the amylose distribution within the grain could be more effective in determining the physical structure of rice endosperm with respect to the total amylose content.

Glutinous (waxy) varieties are characterized by a very low amylose content. Their inner structure appears compact, with hardly distinguishable starch granules, often fused together in clusters, resulting more uneven and irregular than in crystalline varieties. In agreement with this, Jane *et al.* observed that starch granules of waxy rice are more irregular in shape and appear compound or fused [66]. A high porosity was expected to explain the completely white appearance of the milled grains of waxy rice varieties, nevertheless, endosperm porosity was found to be greatly variable in these varieties. Micrographs showed a similar inner structure for the Castelmochi and Italmochi varieties, with low porosity values (0.89 ± 0.23 and 1.09 ± 0.29, respectively), comparable to that of crystalline varieties such as CRLB1 (0.83 ± 0.16); while the waxy CRW3 variety showed a totally different morphological structure with a high porosity value (6.35 ± 0.54), comparable to that of chalky rice varieties. These differences could depend on the different genetic background of these varieties, which probably affects other factors involved in starch granules morphology and distribution. The irregular and confused structure of waxy starch granules, whose contours do not appear clearly distinct by the enveloping membranes as in non-waxy varieties, allowed to measure their average size only in CRW3, showing no differences compared to non-glutinous rice varieties and confirming literature evidence that amylose content does not directly determine starch granules size [65].

Small round structures were observed among the compound starch granules in several images obtained at 5000X magnification, being more evident in some varieties such as Dedalo, Swarna and Doongara, than in others. Those structures (about 2–3 μm in diameter) could correspond to smaller starch granules originated

from amyloplasts division [68], and their presence could be due to the incomplete maturation of the grains, as observed in immature waxy maize seeds [66].

In the present study, several compositional and structural features of rice grain were considered, leading to observe that at a desired low GI or a higher amylose content does not correspond a specific endosperm structure in temperate Japonica rice varieties. Nevertheless, biochemical analysis pointed out some peculiar traits characterizing varieties showing a particularly high or low GI or characterized by specific features in endosperm structure. Among non-waxy rice varieties, Argo (also characterized by the highest average porosity and by one of the lowest GI values), showed the highest content of total fat substances (1.08 ± 0.07~~3~~ g/100 g milled rice). Instead, the lowest content of total fat substances (0.57 ± 0.05~~48~~ g/100 g) was registered by the Arborio variety that is also characterized by the highest GI value.

Surprisingly, ashes content seems to be related to different endosperm morphological traits and in lesser extent to GI: several varieties characterized by a low to medium GI were reported to show also a particularly low ashes content, with Selenio showing 0.21 ± 0.04 g ashes /100 g milled rice; on the contrary the most of the varieties showing a high ashes content were characterized by a high GI value with the only exceptions of Iarim and Valente. Hence, a putative role of ashes content with respect to other parameters, including GI, emerges, which could be explained by differences in starch granules density and nanostructure, such as those in chains length and arrangement of amylose and amylopectin molecules, or in the composition of the amyloplast membranes that could affect the mineral content.

Although the GI values of the 18 IRRI genotypes considered in this study are reported in literature, we chose to avoid taking them into consideration with respect to other evaluated parameters, as these values were calculated with different methodologies, both in vivo and in vitro. The values reported in the literature can give a general indication regarding whether the GI is high or low, but they are in any case poorly comparable with each other and with those obtained from this work. Because of this, to avoid confusion, although the GI values reported in other publications were available, they were not considered and reported in this paper.

Even though genotype plays a major role in determining starch properties and structure, environmental factors are also able to influence starch qualities, protein and lipids synthesis, as well as starch granules size distribution [69,70]; particular growing condition can also impact on the amylose content, that increases in excess of nutrients of zinc and potassium as well as in flooding conditions, and decreases in salinity conditions. Hence, the same variety harvested in different seasons and in different locations cannot guarantee the same quality, thus making particularly difficult to precisely define the characteristics of each variety and to conduct breeding programs aimed to improve rice grain quality. Efforts should be made to identify genotypes characterized by more stable quality traits and to test selected lines in different cultivation areas to better assess their features. Despite this, the

obtained results showed a common endosperm structure for similar varieties, with great differences among varieties belonging to different market groups or destined to different purposes, suggesting that the internal structure, albeit with a certain variability among grains, can be considered characteristic of each variety, or at least of a variety group, representing a sort of fingerprint.

6.4.2 Multivariate exploratory analysis

From the previously described analytical techniques (Section 6.2) two different datasets were obtained: one gathers the different biochemical traits (carbohydrates, nitrogen compounds, proteins, total fatty compounds, raw lipids, ashes and energy content) belonging to the 25 Italian varieties, and the other collects the morphological features (mean eccentricity, mean perimeter, mean area, mean circularity and porosity percentage) extracted from the 54 varieties treated with the ad hoc image analysis algorithm. For nine varieties the algorithm was not able to extract the features related to starch granules, because of the compact nature of the internal structure and starch spatial disposition.

An exploratory multivariate data analysis by PCA was performed using the PLS_Toolbox (operating under MATLAB environment, version 8.9, Eigenvector Research Inc., Manson, WA, USA) software package. PCA was applied to explore the information content of the biochemical data and other assessed characteristics of the studied varieties.

The biochemical and morphological data were imported into MATLAB to be pre-processed and inspected with PCA. After pre-processing the data with autoscaling (also called “unit variance scaling”) [79], two different PCAs were carried out, only for the Italian varieties, by separately looking at biochemical traits and morphological features, with the aim of relating the two computed and acquired datasets with the GI values. In addition, another PCA was performed considering the morphological features calculated for both Italian and non-Italian varieties for which all morphological features could be extracted (45 out of 54).

In total, three PCA models were built. The first model was obtained from the biochemical properties data concerning the 25 Italian varieties whose GI was evaluated (Figure 6.11), finding no relations between GI and the considered properties.

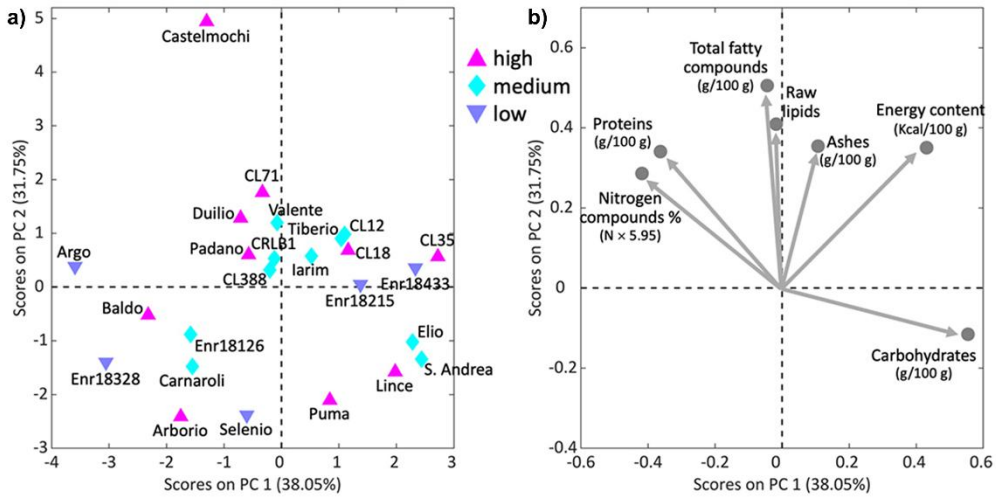

Figure 6.11. Scores plot (A) with the distribution of the 25 Italian genotypes analysed by principal component analysis (PCA) on the results of the biochemical analyses performed on milled rice, coloured on the basis of their high, medium or low GI, and corresponding loadings plot (B) with biochemical traits.

The second PCA model was built on the internal structure features data resulting from FESEM image analysis and considering the additional information about GI for 21 of the characterized Italian varieties. Castelmochi, CL12, CL18 and Enr-18433 were not included in this PCA model due to the impossibility of automatically extracting features related to the average size of starch granules. From the evaluation of the loadings plot (Figure 6.12b) some interesting trend in the variables disposition could be noticed. Two anti-correlation patterns seem to exist: between eccentricity and circularity and between mean starch granules area and porosity.

Looking at the scores plot, a trend related to the GI can be observed in Figure 6.12a, as most of the samples with low GI (namely Enr-18215, Enr-18328, Enr-18433 and Selenio) are in the top-right part of the plot. By inspecting the PCA loadings (Figure 6.12b), it can be seen that these low-GI samples are characterized by high mean area, i.e., they tend to have larger starch granules, while at the same time they exhibit low porosity. On the contrary, the medium- and high-GI samples are located diagonally in the opposite direction, thus having higher porosity and lower mean area, i.e., generally smaller starch granules. Selenio, the variety with the lowest GI, was found to show one of the lowest porosity values (0.40 ± 0.28), coherently with its position, which is opposite to the direction of the porosity loading (Figure 6.12b). The Argo variety, also characterized by a low GI, showed instead the highest porosity value 6.72 ± 0.39, and it is the only low-GI sample not following the detected trend. However, this is not unexpected because GI is indeed a multiparametric value that could be related to other chemical, biological and/or morphological aspects.

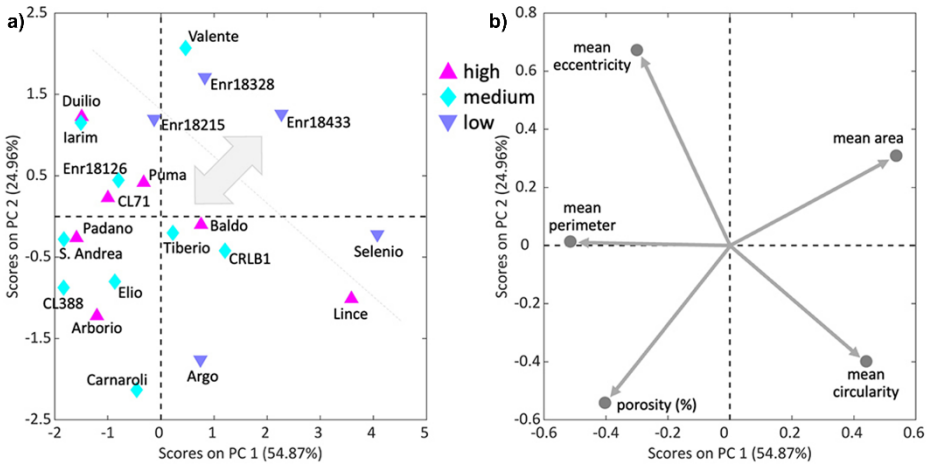

Figure 6.12. Scores plot (A) with the distribution of 21 of the 25 Italian genotypes analysed by PCA on starch granules characteristics and endosperm average porosity, coloured on the basis of their high, medium or low GI, and corresponding loadings plot (B) with starch granules features. The physical structure of 4 of the 25 varieties could not be analysed by the ad hoc developed algorithm, due to the strongly prevalent presence of compact structures, thus causing some missing values in the data table.

The third PCA was made considering 45 of 54 Italian rice varieties and IRRI genotypes, and the morphological features of their internal structure. Nine varieties were excluded because of the impossibility of extracting the average starch granules size value. Looking at the scores and loadings plots (Figure 6.13), eccentricity and circularity appear to be anti-correlated (Figure 6.13b); the samples in the scores plot (Figure 6.13a) are distributed along this direction in relation to how round their starch granules are. Varieties like IAC165, Duilio and Iarim show lower circularity and higher eccentricity, while varieties like Lince and CRLB1 show higher circularity values (round shaped grains). From the same PCA results, focusing again on the loadings plot, also the anti-correlation between porosity and average starch granules area could be spotted.

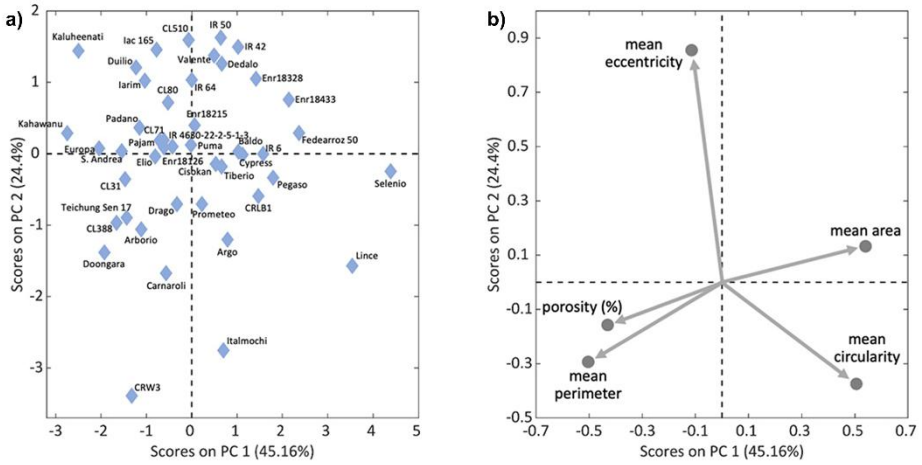

Figure 6.13. Scores plot (A) with the distribution of 45 of the 54 Italian varieties and IRRI genotypes analysed by PCA on starch granules characteristics and endosperm average porosity, and the corresponding loadings plot (B) with starch granules features. The physical structure of nine of the 54 varieties could not be analyzed by the ad hoc developed algorithm, due to the strongly prevalent presence of compact structures, thus causing some missing values in the data table.

6.5 Design of Experiment to evaluate the effect of the algorithm input parameters

6.5.1 Experimental design

The developed image analysis algorithm requires the user to input five parameters. To evaluate the influence of these parameters on the whole image processing and features extraction process, a Design of Experiment (DoE, [80]) approach was developed to systematically explore and model their individual and combined effects (Figure 6.14).

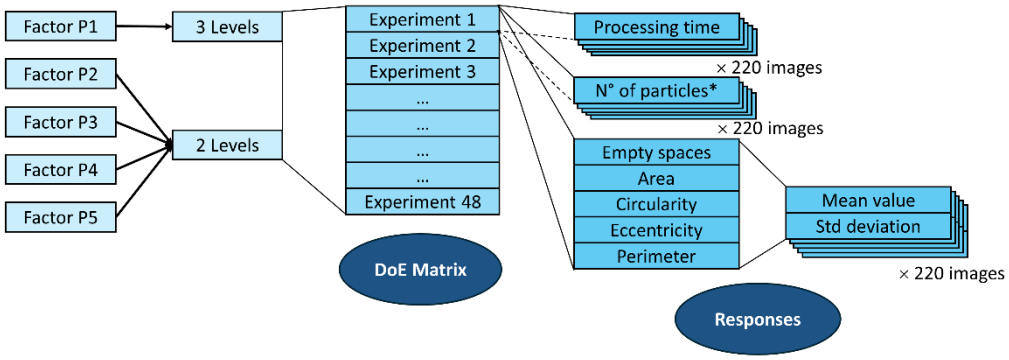

Figure 6.14. Experimental design developed to evaluate the 5 manual parameters (here called also "factors") to be set for using the image analysis algorithm. Schematic representation of the combination of factors and levels that generates 48 experiments (DoE Matrix). The different responses calculated per each experiment were also reported.

In a DoE approach, the five parameters can be treated as "factors" (from P1 to P5). Two or three values (levels) per factor were tested. The values to test were selected based on our knowledge about the algorithm functioning, taking care of setting reasonable values for all the parameters. Due to the computational nature of the problem at hand, all the possible combinations of the parameters were considered and calculated. This means that all features related to the 220 images of the rice starch particles case study were computed for each experimental point of the experimental design (i.e., the specific combination of five parameters). Their values were then modelled as the responses of the experimental design. In addition to these responses, also the total number of objects found, and the calculation time were considered. In addition, both the results before and after the quality assessment step were considered to estimate the influence of this "correction step".

In Table 6.5 is reported a description of the 5 parameters, with the tested levels and the section of the algorithm where they are applied. The three equally spaced levels tested for the P1 factor were 9 %, 18 % and 27 %. Factor P2 is involved in the image segmentation and two different levels were set: 1 and 13. For factor P3 the two considered levels were 60 and 140. For factor P4 the two levels selected were 4 and 8. Finally, for factor P5 the two considered levels were 5 % and 14 %.

Table 6.5. List of the factors (parameters) evaluated using DoE. The tested values, the section of the algorithm where they are computed, and a concise description of the parameters effect are reported for each factor.

Factor	Levels	Algorithm section	Explanation
P1	9, 18, 27	Greyscale threshold	Threshold percentage for adaptive binarization
P2	1, 13	Image segmentation	Threshold to suppress minima lower than P2
P3	60, 140	Empty spaces definition	Threshold to suppress lighter greys
P4	4, 8	Empty spaces definition	Number of thresholds to estimate empty spaces
P5	5, 14	Objects correction	Percentage of fragments to leave out

6.5.2 Exploratory analysis

An exploratory analysis of the results obtained running the algorithm according to the experimental design (48 combinations of the five factors) was performed by PCA.

The calculated morphological features (obtained before and after the correction step), the empty spaces values (expressed as a percentage with respect to the total space) and the number of identified objects obtained from the 220 images analysed with all the 48 combinations of factors were inspected using PCA. The data were pre-processed with autoscaling.

To explore the performances of the developed method a PCA was computed on a dataset consisting of the calculated features (for the empty spaces, corresponding in this case study to the measured porosity, only the value expressed in percentage was considered) and the number of identified objects. It was also decided to exclude the time measurements from this exploratory analysis to better focus on the time effect in a second moment.

To calculate the model, three principal components were kept, corresponding to a cumulative explained variance of 85.35 %. The scores plots could also be coloured according to the different levels set for each factor, thus allowing for a direct interpretation of possible samples' clusters. In the scores plot of PC1 (64.22 %, Figure 6.15a), by colouring the samples according to the two levels of factor P2, a clear separation can be observed. When P2 is set to its high value (P2 = 13) the values of area, eccentricity and perimeter tend to be higher (Figure 6.15b), while the circularity and the number of detected particles tend to be lower. The empty space area does not influence this clear samples separation, confirming that PC1 mostly refers to the "particle shape-related" features.

No clear separations could be found exploring PC2 and PC3, but some interesting trends were spotted. In the scores plot of PC2 (Figure 6.15c), colouring the samples according to the three levels of factor P1, two combined tendencies are visible. First, the samples seem to be organized into three blocks, with those corresponding to the low level of P1 (= 9) characterized by lower PC2 values, then the middle and high levels of P1 correspond to blocks with, on average, increasing PC2 values. By exploring the corresponding loadings plot (Figure 6.15d), the samples with high P1 value shows higher values for eccentricity and empty spaces, while shows lower value for area and circularity.

The second trend is somehow "nested" within each block, and it becomes more visible with the second and third block (Figure 6.15c, samples in green and blue): within each block four sub-blocks can be spotted. This particular behaviour can be connected to the slight trend potentially related to a combination of two factors that is visible when looking at the scores plot of PC3 (Figure 6.15e). Indeed, an increasing separation between the two levels of P3 can be spotted colouring the samples according to the two levels of P3 and sorting the samples (horizontally)

according to the three increasing levels of P1. By looking at the loadings plot (Figure 6.15f), the empty space value is the main responsible for this slight separation, while samples with a higher value of eccentricity (corrected) and number of objects (corrected) tend to show a peculiar low value of PC3 scores.

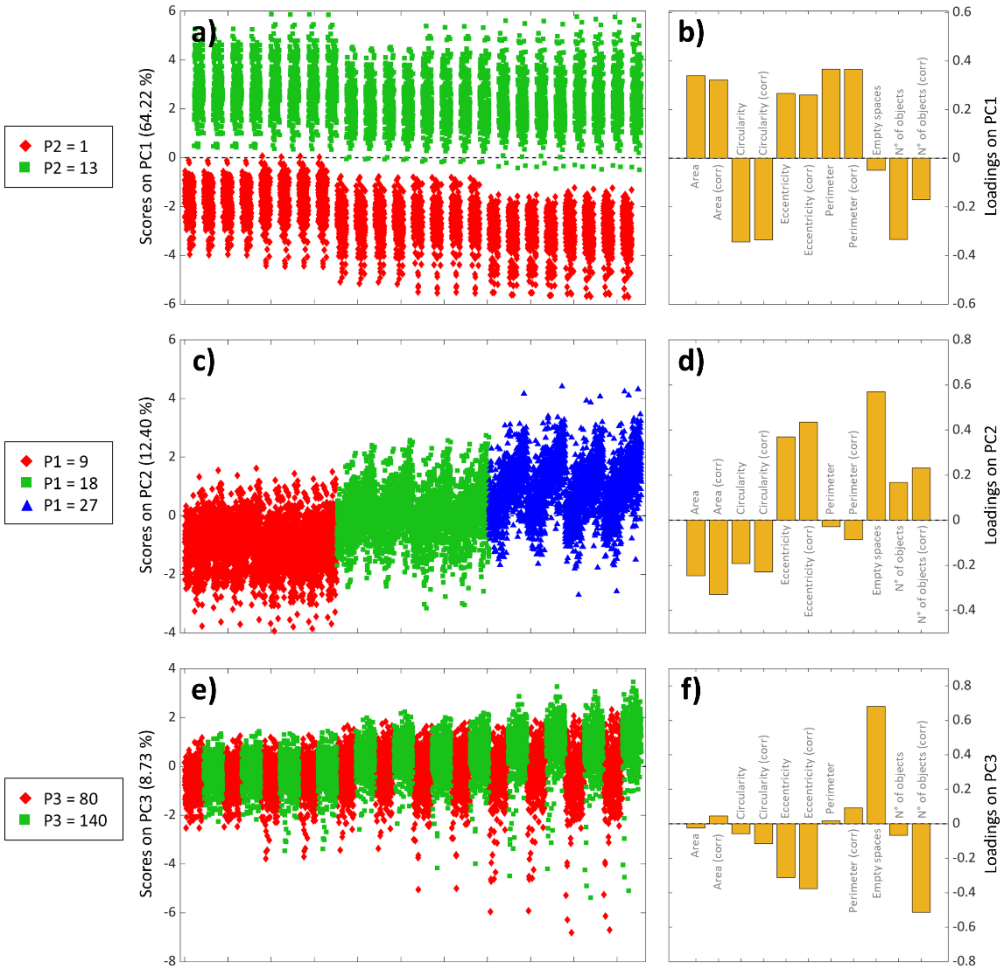

Figure 6.15. PCA results. In the central part the scores plots are reported: (a) PC1 coloured according to the two levels of P2 factor, (c) PC2 coloured according to the three levels of P1 factor, and (e) PC3 coloured according to the two levels of P3 factor. On the right side, the loadings plots of (b) PC1, (d) PC2, and (f) PC3.

The most interesting piece of information from this PCA is that the choice of P1 has an influence on the effect on the choice of the other parameters. This behaviour can be associated to the fact that P1 is correlated with one of the first step of the entire image analysis algorithm (binarization step). Interestingly, with a low P1 value, the choice of the subsequent parameters seems to have a lower contribution on the scores value for PC2 and PC3. On the other hand, a higher P1 value result in a larger dispersion of the scores value for PC2 and PC3, that can be clearly visible in the corresponding scores plot (Figure 6.15c and 6.15e). However, to better understand the origin of these separations spotted in PC2 and PC3 scores plots, even

if part of the detected variability seems to be related to the calculated empty spaces value, also the other features show different contributions to the samples' distribution, for this reason the ASCA method was applied.

6.6 Algorithm evaluation

In this project, following the typical approach of Design of Experiments, ASCA was explored using the PLS_Toolbox software package to better understand and quantify the effect of the parameters on the algorithm performances. In addition, the combinations of tested parameters were also analysed with MLR to understand the contribution of each factor on the overall processing time of the algorithm.

6.6.1 Evaluation on the extracted features calculation using ASCA

ASCA was carried out using a coded DoE matrix (built associating -1 to the lowest value and +1 to the highest value) to represent all the possible parameters combinations and using as a response the matrix containing the 11 extracted features. According to the ASCA results reported in Table 6.6, the largest effect on the features calculation is related to P2 (suppression threshold of local minima in the image segmentation step).

Table 6.6. Summary of the results obtained using ASCA with respect to the extracted features. Five factors and their interaction terms were evaluated. Per each term are reported the number of principal components used, the overall effect (%) and the p-value obtained after a permutation test.

Term	PCs	Effect	P-value
P1	2	8.20	0.01
P2	2	55.69	0.01
P3	1	2.56	0.01
P4	1	0.12	0.01
P5	1	0.44	0.01
P1 × P2	2	1.66	0.01
P1 × P3	1	0.79	0.01
P1 × P4	1	0.02	0.01
P1 × P5	2	0.01	0.01
P2 × P3	1	0.01	0.04
P2 × P4	1	0.00	1.00
P2 × P5	1	0.03	0.01
P3 × P4	1	0.01	0.01
P3 × P5	2	0.00	1.00
P4 × P5	1	0.00	1.00
Residuals	/	30.46	/

From the scores plot of PC1 in ASCA, the separation related to the two levels of P2 is clearly visible (Figure 6.16c), and this is very coherent with the PCA outcomes previously reported. Again, the loadings plot (Figure 6.16d) show that the

separation seems to be related to the particle shape features, while porosity seems to be less related to the effect of P2. The second factor that mostly affects the features calculation is P1, with a lower effect (8.20) than P2 (55.96). By exploring the scores and loadings plots (Figure 6.16a and 6.16b) it becomes apparent that the samples with lower P1 values are more correlated with features like area, circularity and perimeter; on the other hand, the samples with a higher P1value, are more correlated with porosity.

Factors P4 and P5 seem to have very low contributions to the features calculation, while a slightly higher effect was found for P3 (2.56). From the loadings plot, a correlation between porosity and factor P3 can be observed (Figure 6.16f). Interestingly, the scores plot of PC1 (P3) shows a separation between the two P3 levels only for the samples with the lowest value of P1, while the samples with higher P1 values have a more disperse distribution (Figure 6.16e).

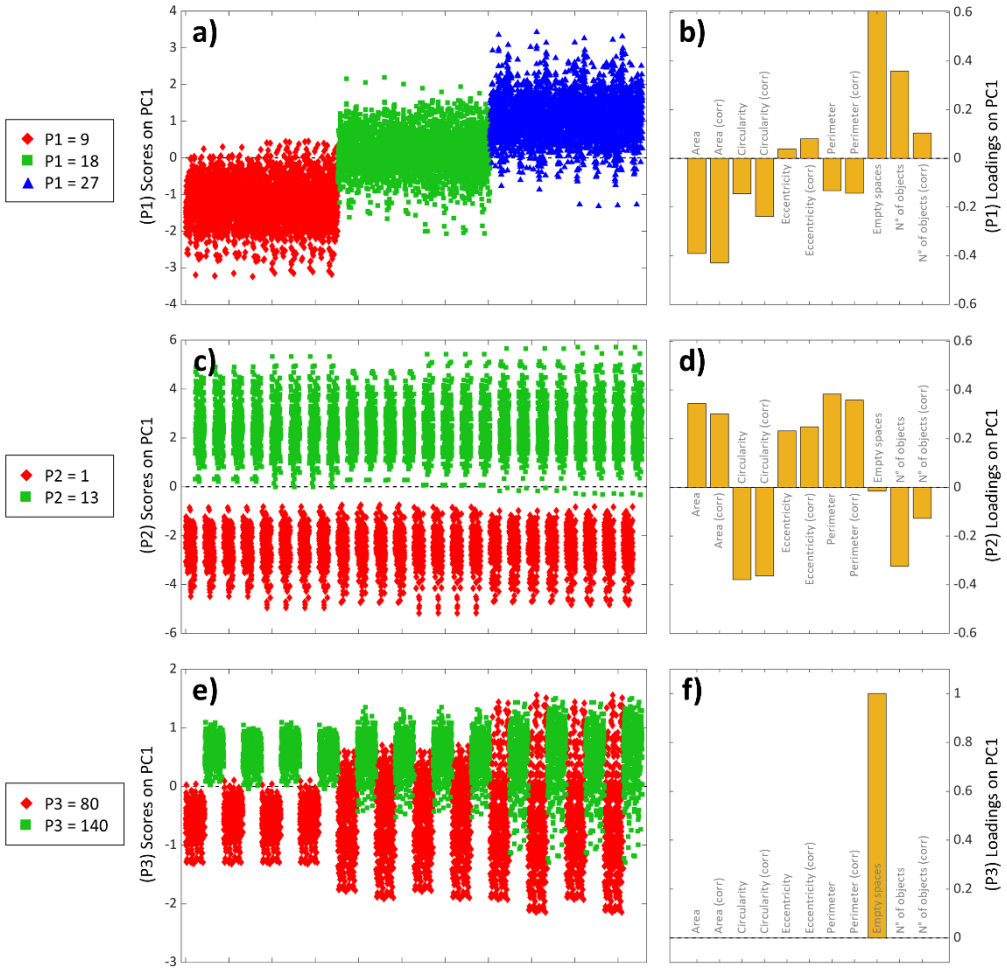

Figure 6.16. ASCA results obtained using the extracted features as a response. In the central part, the scores plots of (a) PC1 (for factor P1) coloured according to the three levels of P1 factor, (c) PC1 (for factor P2) coloured according to the two levels of P2 factor, and (e) PC1 (for factor P3) coloured according to the two levels of P3 factor. On the right side, the loadings plots of (b) PC1 (P1), (d) PC1 (P2), and (f) PC1 (P3).

6.6.2 Evaluation of the total processing time using MLR

While performing the tests based on the DoE approach, it was noticed that the total processing time also changed. Therefore, it was decided to explore the effect of the parameters also in relation to the total processing time of the algorithm. In particular, trying to gather more detailed information about the effect's directions of the different parameters, an MLR model was computed using the same DoE coded matrix used for ASCA and the overall algorithm processing runtime as a response.

MLR modelling was then applied to explore how the choice of factors' levels influences the overall processing time. The MLR coefficient plots (Figure 6.17a) revealed that the largest contribution on processing time comes from the choice of P1. Interestingly, also P3 and P4 showed a significant contribution, as well as the interaction terms of P1 with P3 and P1 with P4. In addition, by looking at the bars' orientation in Figure 6.17a, it can be noticed that at higher values of P1, P3 and P4 generally correspond to longer processing times, while P2 has an opposite behaviour but also a much-reduced relevance. P5 seems to have no effect on the processing time, strengthening our previous findings.

From these general considerations about the regression coefficients, it is also important to consider that different factors interactions and the quadratic term of P1 appear to be relevant, so the response surfaces of the MLR model were explored for visualizing the contribution of all these coefficients together (Figure 6.17). Given that only two factors at a time can be selected for visualizing the response surfaces (as the third dimension of the plot is assigned to the response values), from the outcomes of the regression coefficients plot (Figure 6.17a) it was decided to evaluate only P1, P3 and P4 while fixing the value of P2 at its high level (corresponding to the tested value of 13.0) and the value of P5 at its central level (corresponding to the average of the two tested values). By exploring the response surfaces relative to P1 and P3 (Figure 6.17b), it becomes clear that when P1 (i.e., the factor with the strongest effect) is set to a low value, if the value of P3 is set to higher values, a shorter processing time is obtained, due to the strong P1-P3 interaction term. On the contrary, from the response surface of P1 and P4 (Figure 6.17c), it is clear that the value set for P4 has a low impact on the total processing time, but that setting it to a low value can be generally preferred. These two pieces of evidence are also confirmed by the response surface of P3 and P4 when setting P1 to its lowest level (Figure 6.17d). In this plot, even if the parameters' contribution is small compared to the P1 effect, the lowest processing time corresponds to a high value for P3 and a lower value for P4. Interestingly, the second lowest region of the surface corresponds to setting a low value of P3 and a high value of P4, while keeping both P3 and P4 high or low results in longer processing times.

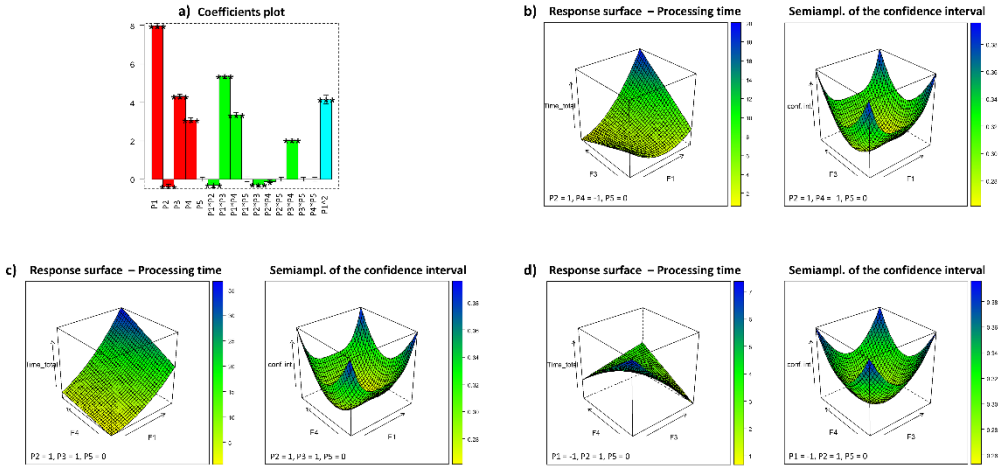

Figure 6.17. Results obtained from the MLR model built using the processing time as the response to model. (a) Plot of the regression coefficients obtained for the linear (red) and interaction (green) terms of the 5 factors, and for the quadratic (cyan) term of P1 factor. (b) Response surface and corresponding confidence interval obtained by observing P1, P3 and the total processing time. P2, P4 and P5 were fixed at levels 1, -1 and 0 respectively. (c) Response surface and corresponding confidence interval obtained by observing P1, P4 and the total processing time. P2, P3 and P5 were fixed at levels 1, 1 and 0 respectively. (d) Response surface and corresponding confidence interval obtained by observing P3, P4 and the total processing time. P1, P2 and P5 were fixed at levels -1, 1 and 0 respectively.

To resume, according to the results obtained from ASCA and MLR, the values chosen for P3 and P4 seem to slightly affect the values of the calculated features, but they have a remarkable effect on the overall processing time. On the contrary, the choice of P2 has a strong effect on the values obtained for the features, but no effect on the algorithm processing time. Interestingly, P1 is the only parameter that highlighted a strong effect both on the features calculation and on the algorithm running time. The effect of P5 appears to be negligible both on time and features extraction. This is not unexpected, as P5 regulates the percentage of fragments to be left out while selecting the objects based on their size distribution, which is, in any case, a simple computation to be performed.

6.7 Project conclusions

This project highlights the existence of a large variability among the analysed rice genotypes. For these samples, the GI was determined and the internal structure of the grain in a large set of varieties was investigated. Five of them, due to their particularly low GI, resulted to be already suitable for people affected by diabetes or other metabolic disorders.

When evaluating the endosperm inner structure, with particular reference to starch granules size and arrangement, it results that crystalline varieties are characterized by low porosity, having bigger and more tightly packed starch granules, compared

to chalky short-grain and Long A-grain varieties destined to the Italian national market, which are characterized, in average, by a higher porosity of the grains. Glutinous varieties, on the other hand, are not characterized by a defined level of porosity.

The use of a multivariate exploratory approach based on PCA also allowed inspecting the relationships among the morphological parameters algorithmically extracted from the SEM images: an interesting trend related to the GI could be observed, showing that most of the samples with low GI tend to have larger starch granules, while at the same time they exhibit lower porosity values. On the contrary, the medium- and high-GI samples showed the opposite features (i.e., higher porosity and lower mean area).

All the information obtained in the study will be useful for rice breeding programs for developing varieties with the best characteristics according to their target purpose, focusing on grain quality to fully satisfy consumers. Furthermore, the low GI genotypes already identified, will represent the starting point for carrying out specific crossbreeding programs to develop new low GI rice lines, thus meeting the needs of diabetic consumers.

In this context, to obtain specific information about the rice kernel inner structure, a new versatile method to perform image analysis was developed by extracting a series of morphological features related to similar-shaped objects from greyscale images. As a result, different morphological features (i.e., area, perimeter, eccentricity, circularity, radius, diameter), related to the starch particles that define the internal kernel structure, were obtained. In addition, also the total area of the empty spaces among the particles was computed. In addition, for evaluating the effect of the choice of the input parameters on the algorithm result, a design of experiment approach was applied, and the resulting outcomes were thoroughly analysed using PCA, ASCA and MLR.

The outcomes of this case study demonstrate the ability of the algorithm to recognize similar-shaped objects (particles) and to automatically calculate their related morphological features. The opportunity of realizing high-throughput image analysis and the possibility of varying the input options depending on the different targets, combined with the adaptability of many crucial steps of the developed algorithm, makes this new (image analysis) method extremely versatile for a wide variety of images (not only FESEM acquisitions) and suitable for different research fields.

References | Chapter 6

[1] M. Mohd Ali, N. Hashim, S.A. Aziz, O. Lasekan, Emerging non-destructive thermal imaging technique coupled with chemometrics on quality and safety inspection in food and agriculture, Trends Food Sci Technol 105 (2020) 176–185. https://doi.org/10.1016/j.tifs.2020.09.003.

[2] L. Zhang, S. Shao, Image-based machine learning for materials science, J Appl Phys 132 (2022). https://doi.org/10.1063/5.0087381.

[3] M. Ragone, R. Shahabazian-Yassar, F. Mashayek, V. Yurkiv, Deep learning modeling in microscopy imaging: A review of materials science applications, Prog Mater Sci 138 (2023). https://doi.org/10.1016/j.pmatsci.2023.101165.

[4] M. Ge, F. Su, Z. Zhao, D. Su, Deep learning analysis on microscopic imaging in materials science, Mater Today Nano 11 (2020). https://doi.org/10.1016/j.mtnano.2020.100087.

[5] J.M. Prats-Montalbán, A. de Juan, A. Ferrer, Multivariate image analysis: A review with applications, Chemometrics and Intelligent Laboratory Systems 107 (2011) 1–23. https://doi.org/10.1016/j.chemolab.2011.03.002.

[6] L.F. Capitán-Vallvey, N. López-Ruiz, A. Martínez-Olmos, M.M. Erenas, A.J. Palma, Recent developments in computer vision-based analytical chemistry: A tutorial review, Anal Chim Acta 899 (2015) 23–56. https://doi.org/10.1016/j.aca.2015.10.009.

[7] M. Rezazadeh, S. Seidi, M. Lid, S. Pedersen-Bjergaard, Y. Yamini, The modern role of smartphones in analytical chemistry, TrAC - Trends in Analytical Chemistry 118 (2019) 548–555. https://doi.org/10.1016/j.trac.2019.06.019.

[8] H.L. Zhai, B.Q. Li, J. Chen, X. Wang, M.L. Xu, J.J. Liu, S.H. Lu, Chemical image moments and their applications, TrAC - Trends in Analytical Chemistry 103 (2018) 119–125. https://doi.org/10.1016/j.trac.2018.03.017.

[9] J. Zhang, C. Li, M.M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang, M. Grzegorzek, A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches, Artif Intell Rev 55 (2022) 2875–2944. https://doi.org/10.1007/s10462-021-10082-4.

[10] A.A. Appel, M.A. Anastasio, J.C. Larson, E.M. Brey, Imaging challenges in biomaterials and tissue engineering, Biomaterials 34 (2013) 6615–6630. https://doi.org/10.1016/j.biomaterials.2013.05.033.

[11] D. Li, C. Li, Y. Yao, M. Li, L. Liu, Modern imaging techniques in plant nutrition analysis: A review, Comput Electron Agric 174 (2020). https://doi.org/10.1016/j.compag.2020.105459.

[12] S.K. Jung, A Review of Image Analysis in Biochemical Engineering, Biotechnology and Bioprocess Engineering 24 (2019) 65–75. https://doi.org/10.1007/s12257-018-0372-8.

[13] T.W. Nattkemper, Multivariate image analysis in biomedicine, J Biomed Inform 37 (2004) 380–391. https://doi.org/10.1016/j.jbi.2004.07.010.

[14] Y. Skaf, R. Laubenbacher, Topological data analysis in biomedicine: A review, J Biomed Inform 130 (2022). https://doi.org/10.1016/j.jbi.2022.104082.

[15] R. Murphy, A. Turcott, L. Banuelos, E. Dowey, B. Goodwin, K.O.H. Cardinal, SIMPoly: A Matlab-Based Image Analysis Tool to Measure Electrospun Polymer Scaffold Fiber Diameter, Tissue Eng Part C Methods 26 (2020) 628–636. https://doi.org/10.1089/ten.tec.2020.0304.

[16] A.Q. Lu, S.Z. Zhang, Q. Tian, Matlab image processing technique and application in pore structure characterization of hardened cement pastes, in: Adv Mat Res, 2013: pp. 1374–1379. https://doi.org/10.4028/www.scientific.net/AMR.785-786.1374.

[17] J.P. Kim, U. Nowostawska, K.A. Hunter, Comparison of particle size spectrum determination from images made using manual and automated image analysis, Environ Technol 29 (2008) 1191–1198. https://doi.org/10.1080/09593330802217773.

[18] S. Maaß, J. Rojahn, R. Hänsch, M. Kraume, Automated drop detection using image analysis for online particle size monitoring in multiphase systems, Comput Chem Eng 45 (2012) 27–37. https://doi.org/10.1016/j.compchemeng.2012.05.014.

[19] R.S. Edwin, M. Mushthofa, E. Gruyaert, N. De Belie, Quantitative analysis on porosity of reactive powder concrete based on automated analysis of back-scattered-electron images, Cem Concr Compos 96 (2019) 1–10. https://doi.org/10.1016/j.cemconcomp.2018.10.019.

[20] R. Ziel, A. Haus, A. Tulke, Quantification of the pore size distribution (porosity profiles) in microfiltration membranes by SEM, TEM and computer image analysis, J Memb Sci 323 (2008) 241–246. https://doi.org/10.1016/j.memsci.2008.05.057.

[21] S. Diamond, M.E. Leeman, PORE SIZE DISTRIBUTIONS IN HARDENED CEMENT PASTE BY SEM IMAGE ANALYSIS, n.d.

[22] H. Kim, J. Han, T.Y.J. Han, Machine vision-driven automatic recognition of particle size and morphology in SEM images, Nanoscale 12 (2020) 19461–19469. https://doi.org/10.1039/d0nr04140h.

[23] N. Prakongkep, A. Suddhiprakarn, I. Kheoruenromne, R.J. Gilkes, SEM image analysis for characterization of sand grains in Thai paddy soils, Geoderma 156 (2010) 20–31. https://doi.org/10.1016/j.geoderma.2010.01.003.

[24] B. Gaël, T. Christelle, E. Gilles, G. Sandrine, S.F. Tristan, Determination of the proportion of anhydrous cement using SEM image analysis, Constr Build Mater 126 (2016) 157–164. https://doi.org/10.1016/j.conbuildmat.2016.09.037.

[25] G.N. Barrera, G. Calderón-Domínguez, J. Chanona-Pérez, G.F. Gutiérrez-López, A.E. León, P.D. Ribotta, Evaluation of the mechanical damage on wheat starch granules by SEM, ESEM, AFM and texture image analysis, Carbohydr Polym 98 (2013) 1449–1457. https://doi.org/10.1016/j.carbpol.2013.07.056.

[26] M. Shanmuga Priya, P. Divya, R. Rajalakshmi, A review status on characterization and electrochemical behaviour of biomass derived carbon materials for energy storage supercapacitors, Sustain Chem Pharm 16 (2020). https://doi.org/10.1016/j.scp.2020.100243.

[27] M.K.A. Mohammed, M.R. Mohammad, M.S. Jabir, D.S. Ahmed, Functionalization, characterization, and antibacterial activity of single wall and multi wall carbon nanotubes, in: IOP Conf Ser Mater Sci Eng, Institute of Physics Publishing, 2020. https://doi.org/10.1088/1757-899X/757/1/012028.

[28] W.D. Niemeyer, SEM/EDS analysis for problem solving in the food industry, in: Scanning Microscopies 2015, SPIE, 2015: p. 96360G. https://doi.org/10.1117/12.2196962.

[29] J.C. Russ, Image Analysis of Foods, J Food Sci 80 (2015) 1974–1987. https://doi.org/10.1111/1750-3841.12987.

[30] V. Sharma, A. Bhardwaj, Scanning electron microscopy (SEM) in food quality evaluation, in: Evaluation Technologies for Food Quality, Elsevier, 2019: pp. 743–761. https://doi.org/10.1016/B978-0-12-814217-2.00029-9.

[31] A. Baiano, Applications of hyperspectral imaging for quality assessment of liquid based and semi-liquid food products: A review, J Food Eng 214 (2017) 10–15. https://doi.org/10.1016/j.jfoodeng.2017.06.012.

[32] F. Pieniazek, V. Messina, Texture Analysis of Freeze Dried Banana Applying Scanning Electron Microscopy Combined with Image Analysis Techniques, ETP International Journal of Food Engineering (2018) 127–131. https://doi.org/10.18178/ijfe.4.2.127-131.

[33] J. Tan, H. Zhang, X. Gao, SEM image processing for food structure analysis, J Texture Stud 28 (1997) 657–672. https://doi.org/10.1111/j.1745-4603.1997.tb00145.x.

[34] J. Tan, B.M. Balasubramanian, Particle size measurements and scanning electron microscopy (SEM) of cocoa particles refined/conched by conical and cylindrical roller stone melangers, J Food Eng 212 (2017) 146–153. https://doi.org/10.1016/j.jfoodeng.2017.05.033.

[35] R. Peters, G. ten Dam, H. Bouwmeester, H. Helsper, G. Allmaier, F. vd Kammer, R. Ramsch, C. Solans, M. Tomaniová, J. Hajslova, S. Weigel, Identification and characterization of organic nanoparticles in food, TrAC - Trends in Analytical Chemistry 30 (2011) 100–112. https://doi.org/10.1016/j.trac.2010.10.004.

[36] F. Haxhari, F. Savorani, M. Rondanelli, E. Cantaluppi, L. Campanini, E. Magnani, C. Simonelli, G. Gavoci, A. Chiadò, M. Sozzi, N. Cavallini, A. Chiodoni, C. Gasparri, G.C. Barrile, A. Cavioni, F. Mansueto, G. Mazzola, A. Moroni, Z. Patelli, M. Pirola, A. Tartara, D. Guido, S. Perna, R.

Magnaghi, Endosperm structure and Glycemic Index of Japonica Italian rice varieties, Front Plant Sci 14 (2023). https://doi.org/10.3389/fpls.2023.1303771.

[37] D.J.A. Jenkins, T.M.S. Wolever, R.H. Taylor, H. Barker, H. Fielden, J.M. Baldwin, A.C. Bowling, H.C. Newman, D. V. Goff, Glycemic index of foods: A physiological basis for carbohydrate exchange, American Journal of Clinical Nutrition 34 (1981) 362–366. https://doi.org/10.1093/AJCN/34.3.362.

[38] F.S. Atkinson, K. Foster-Powell, J.C. Brand-Miller, International tables of glycemic index and glycemic load values: 2008, Diabetes Care 31 (2008) 2281–2283. https://doi.org/10.2337/DC08-1239.

[39] J. Hallfrisch, K.M. Behall, Mechanisms of the Effects of Grains on Insulin and Glucose Responses, J Am Coll Nutr 19 (2000) 320S-325S. https://doi.org/10.1080/07315724.2000.10718967.

[40] G. Livesey, R. Taylor, H. Livesey, S. Liu, Is there a dose-response relation of dietary glycemic load to risk of type 2 diabetes? Meta-analysis of prospective cohort studies1-3, American Journal of Clinical Nutrition 97 (2013) 584–596. https://doi.org/10.3945/AJCN.112.041467.

[41] S.N. Bhupathiraju, D.K. Tobias, V.S. Malik, A. Pan, A. Hruby, J.E. Manson, W.C. Willett, F.B. Hu, Glycemic index, glycemic load, and risk of type 2 diabetes: Results from 3 large US cohorts and an updated meta-analysis, American Journal of Clinical Nutrition 100 (2014) 218–232. https://doi.org/10.3945/AJCN.113.079533.

[42] F. Yao, C. Li, J. Li, G. Chang, Y. Wang, R. Campardelli, P. Perego, C. Cai, Effects of Different Cooking Methods on Glycemic Index, Physicochemical Indexes, and Digestive Characteristics of Two Kinds of Rice, Processes 11 (2023). https://doi.org/10.3390/PR11072167.

[43] W.H.H. Sheu, A. Rosman, A. Mithal, N. Chung, Y.T. Lim, C. Deerochanawong, P. Soewondo, M.K. Lee, K.H. Yoon, O. Schnell, Addressing the burden of type 2 diabetes and cardiovascular disease through the management of postprandial hyperglycaemia: An Asian-Pacific perspective and expert recommendations, Diabetes Res Clin Pract 92 (2011) 312–321. https://doi.org/10.1016/J.DIABRES.2011.04.019.

[44] X. yu Ma, J. ping Liu, Z. yuan Song, Glycemic load, glycemic index and risk of cardiovascular diseases: Meta-analyses of prospective studies, Atherosclerosis 223 (2012) 491–496. https://doi.org/10.1016/J.ATHEROSCLEROSIS.2012.05.028.

[45] A. Mirrahimi, R.J. de Souza, L. Chiavaroli, J.L. Sievenpiper, J. Beyene, A.J. Hanley, L.S.A. Augustin, C.W.C. Kendall, D.J.A. Jenkins, Associations of glycemic index and load with coronary heart disease events: a systematic review and meta-analysis of prospective cohorts., J Am Heart Assoc 1 (2012). https://doi.org/10.1161/JAHA.112.000752.

[46] Y. Choi, E. Giovannucci, J.E. Lee, Glycaemic index and glycaemic load in relation to risk of diabetes-related cancers: A meta-analysis, British Journal of Nutrition 108 (2012) 1934–1947. https://doi.org/10.1017/S0007114512003984.

[47] F. Turati, C. Galeone, S. Gandini, L.S. Augustin, D.J.A. Jenkins, C. Pelucchi, C. La Vecchia, High glycemic index and glycemic load are associated with moderately increased cancer risk, Mol Nutr Food Res 59 (2015) 1384–1394. https://doi.org/10.1002/MNFR.201400594.

[48] K. Srikaeo, Application of a Rapid In Vitro Method Based on Glucometer for Determination of Starch Digestibility and Estimated Glycemic Index in Rice, Starch/Staerke 75 (2023). https://doi.org/10.1002/STAR.202200174.

[49] A. Hanna-Moussa, M.J. Gardner, L. Romayne Kurukulasuriya, J.R. Sowers, Dysglycemia/prediabetes and cardiovascular risk factors, Rev Cardiovasc Med 10 (2009) 202–208. https://doi.org/10.3909/RICM0474.

[50] E.E. Blaak, J.M. Antoine, D. Benton, I. Björck, L. Bozzetto, F. Brouns, M. Diamant, L. Dye, T. Hulshof, J.J. Holst, D.J. Lamport, M. Laville, C.L. Lawton, A. Meheust, A. Nilson, S. Normand, A.A. Rivellese, S. Theis, S.S. Torekov, S. Vinoy, Impact of postprandial glycaemia on health and prevention of disease, Obesity Reviews 13 (2012) 923–984. https://doi.org/10.1111/J.1467-789X.2012.01011.X.

[51] I. Goñi, A. Garcia-Alonso, F. Saura-Calixto, A starch hydrolysis procedure to estimate glycemic index, Nutrition Research 17 (1997) 427–437. https://doi.org/10.1016/S0271-5317(97)00010-9.

[52] A. Lovegrove, O. Kosik, E. Bandonill, R. Abilgos-Ramos, M. Romero, N. Sreenivasulu, P. Shewry, Improving rice dietary fibre content and composition for human health, J Nutr Sci Vitaminol (Tokyo) 65 (2019) S48–S50. https://doi.org/10.3177/JNSV.65.S48.

[53] Z.H. Howlader, Screening for Nutritionally Rich and Low Glycemic Index Bangladeshi Rice Varieties, (2009). https://www.researchgate.net/publication/328164434 (accessed October 9, 2024).

[54] M.A. Fitzgerald, S. Rahman, A.P. Resurreccion, J. Concepcion, V.D. Daygon, S.S. Dipti, K.A. Kabir, B. Klingner, M.K. Morell, A.R. Bird, Identification of a major genetic determinant of glycaemic index in rice, Rice 4 (2011) 66–74. https://doi.org/10.1007/S12284-011-9073-Z.

[55] Y. Dwiningsih, J. Alkahtani, Potential of Pigmented Rice Variety Cempo Ireng in Rice Breeding Program for Improving Food Sustainability, (2023). https://doi.org/10.20944/PREPRINTS202302.0039.V1.

[56] H.N. Larsen, O.W. Rasmussen, P.H. Rasmussen, K.K. Alstrup, S.K. Biswas, I. Tetens, S.H. Thilsted, K. Hermansen, Glycaemic index of parboiled rice depends on the severity of processing: Study in type 2 diabetic subjects, Eur J Clin Nutr 54 (2000) 380–385. https://doi.org/10.1038/SJ.EJCN.1600969.

[57] P. Hettiarachchi, M.T. Jiffry, E.R. Jansz, A.R. Wickramasinghe, D.J. Fernando, Glycaemic indices of different varieties of rice grown in Sri Lanka., Ceylon Med J 46 (2001) 11–14. https://doi.org/10.4038/CMJ.V46I1.6516.

[58] Q. Sun, D. Spiegelman, R.M. Van Dam, M.D. Holmes, V.S. Malik, W.C. Willett, F.B. Hu, White rice, brown rice, and risk of type 2 diabetes in US men and women, Arch Intern Med 170 (2010) 961–969. https://doi.org/10.1001/ARCHINTERNMED.2010.109.

[59] H.M. Boers, J. Seijen Ten Hoorn, D.J. Mela, A systematic review of the influence of rice characteristics and processing methods on postprandial glycaemic and insulinaemic responses, British Journal of Nutrition 114 (2015) 1035–1045. https://doi.org/10.1017/S0007114515001841.

[60] T. Van Ngo, K. Kunyanee, N. Luangsakul, Insights into Recent Updates on Factors and Technologies That Modulate the Glycemic Index of Rice and Its Products, Foods 12 (2023). https://doi.org/10.3390/FOODS12193659.

[61] P. Rathinasabapathi, N. Purushothaman, R. Vl, M. Parani, Whole genome sequencing and analysis of Swarna, a widely cultivated indica rice variety with low glycemic index, Sci Rep 5 (2015). https://doi.org/10.1038/SREP11303.

[62] Z.-Y. Wang, F.-Q. Zheng, G.-Z. Shen, J.-P. Gao, D.P. Snustad, M.-G. Li, J.-L. Zhang, M.-M. Hong, The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene, The Plant Journal 7 (1995) 613–622. https://doi.org/10.1046/J.1365-313X.1995.7040613.X.

[63] G.E. Vandeputte, J.A. Delcour, From sucrose to starch granule to starch physical behaviour: A focus on rice starch, Carbohydr Polym 58 (2004) 245–266. https://doi.org/10.1016/J.CARBPOL.2004.06.003.

[64] R. Matsushima, M. Maekawa, N. Fujita, W. Sakamoto, A rapid, direct observation method to isolate mutants with defects in starch grain morphology in rice, Plant Cell Physiol 51 (2010) 728–741. https://doi.org/10.1093/PCP/PCQ040.

[65] R. Matsushima, Morphological variations of starch grains, Starch: Metabolism and Structure (2015) 425–441. https://doi.org/10.1007/978-4-431-55495-0_13.

[66] J. -L Jane, T. Kasemsuwan, S. Leas, H. Zobel, J.F. Robyt, Anthology of Starch Granule Morphology by Scanning Electron Microscopy, Starch - Stärke 46 (1994) 121–129. https://doi.org/10.1002/STAR.19940460402.

[67] N. Lindeboom, P.R. Chang, R.T. Tyler, Analytical, biochemical and physicochemical aspects of starch granule size, with emphasis on small granule starches: A review, Starch/Staerke 56 (2004) 89–99. https://doi.org/10.1002/STAR.200300218.

[68] M.S. Yun, Y. Kawagoe, Amyloplast division progresses simultaneously at multiple sites in the endosperm of rice., Plant Cell Physiol 50 (2009) 1617–1626. https://doi.org/10.1093/PCP/PCP104.

[69] J.A. Patindol, T.J. Siebenmorgen, Y.J. Wang, Impact of environmental factors on rice starch structure: A review, Starch/Staerke 67 (2015) 42–54. https://doi.org/10.1002/STAR.201400174.

[70] N. Kongseree, B.O. Juliano, N. Kongseree, Physicochemical Properties of Rice Grain and Starch from Lines Differing Amylose Content and Gelatinization Temperature, J Agric Food Chem 20 (1972) 714–718. https://doi.org/10.1021/JF60181A020.

[71] J.B. Miller, E. Pang, L. Bramall, Rice: A high or low glycemic index food?, American Journal of Clinical Nutrition 56 (1992) 1034–1036. https://doi.org/10.1093/AJCN/56.6.1034.

[72] L.N. Panlasigui, L.U. Thompson, B.O. Juliano, C.M. Perez, S.H. Yiu, G.R. Greenberg, Rice varieties with similar amylose content differ in starch digestibility and glycemic response in humans, American Journal of Clinical Nutrition 54 (1991) 871–877. https://doi.org/10.1093/AJCN/54.5.871.

[73] M. Rondanelli, F. Haxhari, C. Gasparri, G.C. Barrile, A. Cavioni, D. Guido, F. Mansueto, M. Zese, G. Mazzola, A. Moroni, Z. Patelli, G. Peroni, M. Pirola, A. Tartara, E. Cantaluppi, L. Campanini, E. Magnani, S. Feccia, S. Perna, Glycemic Index and Amylose Content of 25 Japonica Rice Italian Cultivar, Starch/Staerke 75 (2023). https://doi.org/10.1002/STAR.202300031.

[74] F. Scazzina, M. Dall'Asta, M.C. Casiraghi, S. Sieri, D. Del Rio, N. Pellegrini, F. Brighenti, Glycemic index and glycemic load of commercial Italian foods, Nutrition, Metabolism and Cardiovascular Diseases 26 (2016) 419–429. https://doi.org/10.1016/J.NUMECD.2016.02.013.

[75] C.P. Villareal, B.O. Juliano, Comparative Levels of Waxy Gene Product of Endosperm Starch Granules of Different Rice Ecotypes, Starch - Stärke 41 (1989) 369–371. https://doi.org/10.1002/STAR.19890411002.

[76] M.S. Yun, Y. Kawagoe, Septum formation in amyloplasts produces compound granules in the rice endosperm and is regulated by plastid division proteins, Plant Cell Physiol 51 (2010) 1469–1479. https://doi.org/10.1093/PCP/PCQ116.

[77] K. Hoshikawa, R. Sasaki, K. Hasebe, Development and Rooting Capacity of Rice Nursling Seedlings Grown under Different Raising Conditions, Japanese Journal of Crop Science 64 (1995) 328–332. https://doi.org/10.1626/JCS.64.328.

[78] M. Yano, K. Okuno, J. Kawakami, H. Satoh, T. Omura, High amylose mutants of rice, Oryza sativa L., Theoretical and Applied Genetics 69 (1985) 253–257. https://doi.org/10.1007/BF00662436.

[79] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: Improving the biological information content of metabolomics data, BMC Genomics 7 (2006). https://doi.org/10.1186/1471-2164-7-142.

[80] R. Leardi, Experimental design in chemistry: A tutorial, Anal Chim Acta 652 (2009) 161–172. https://doi.org/10.1016/j.aca.2009.06.015.

Chapter 7

General Conclusions

This thesis presents and condensate the results obtained during a three-year PhD project focused on the critical role of chemometrics in addressing key challenges within the food industry, highlighting its potential to contribute improving quality, safety, and sustainability through advanced data-driven approaches. By leveraging multivariate data analysis, this work has provided innovative solutions to optimize industrial processes, enhance product authenticity, and extract meaningful insights from complex chemical datasets in contexts relevant for different chemical engineering research fields.

Through a series of case studies, the thesis presents the versatility and the advantages of different chemometric approaches in addressing various challenges faced by companies and researchers. Particular attention was placed on applying different chemometric techniques in real-world scenarios, highlighting the practical benefits of these innovative methodologies. The first case study (case study #1 - Chapter 3) explored the optimization of industrial food processes using ^{1}H-NMR spectroscopy coupled with PCA. The results demonstrated how chemometric tools can streamline process efficiency, reduce resource consumption, and enhance product quality. The second case study (case study #2 - Chapter 4) focused on food traceability and authenticity, showcasing how a data fusion approach combining NMR and LC-MS can accurately classify hazelnuts based on geographical origin and cultivar. This highlights the importance of multivariate classification techniques such as PLS-DA in preventing food fraud and protecting regional food excellences like the Piedmont TGT hazelnuts. The third case study (case study #3 - Chapter 5) addressed analytical challenges in wine research, particularly in the quantification of ethanol in alcoholic beverages using NMR spectroscopy. The study systematically investigated experimental parameters, sample preparation protocols and processing strategies, optimizing the accuracy and precision of direct and indirect ethanol quantification. This study demonstrates how chemometric methodologies can contribute to improving analytical protocols in the food and beverage industry. The last explored case study (case study #4 - Chapter 6) explored a highly innovative approach in the field of image analysis, where a novel algorithm was developed to extract morphological features from FESEM images of rice kernels. This technique, combined with Design of Experiments (DoE) methodologies, provided insights into the correlation between starch granule structure and the glycaemic index, emphasizing the potential impact of chemometric tools in nutrition and health research. In addition, after coding the algorithm, DoE approaches, coupled with chemometric techniques, were employed

to evaluate the algorithm performances and to optimize the choice of input parameters with the aim of further improving automatic image analysis.

A key strength of this work can be found in its interdisciplinary impact, demonstrating how chemometric techniques can be applied across food industry, analytical chemistry, chemical engineering, and data science. The methodologies employed are highly scalable and can be implemented in industrial settings for quality control, process optimization, and product authentication. The integration of data fusion strategies further highlights the potential of combining information from multiple techniques, leading to more comprehensive and reliable decision-making tools. The ability to process large and heterogeneous datasets with high accuracy is essential for modern industries seeking to implement data-driven strategies for efficiency, sustainability, and regulatory compliance.

About the further advancements of this research, particular attention goes to areas such as machine learning integration, model validation, and automated data analysis. The increasing adoption of high-throughput analytical technologies will require even more advanced data processing techniques, making chemometrics a necessary component of future research and industrial innovation. Potential future developments related to these topics can include:

- Refinement of model validation to ensure robustness and reproducibility.
- Enhancing data fusion methods to effectively integrate other analytical techniques.
- Development of optimized chemometric models to facilitate real-time monitoring and to improve decision-making in industrial processes.
- Expanding these food industry applications to other industry-related fields and chemical engineering topics.

In conclusion, this thesis highlighted the versatile role of chemometrics in modern chemical engineering, demonstrating its ability to enhance data interpretation, process optimization and decision-making. By integrating theoretical advancements with practical applications, this work contributes to scientific knowledge providing concrete tools for industries aiming to improve quality, authenticity, and sustainability. The methodologies developed here represent a step forward in the application of chemometric tools, offering scalable and impactful solutions for both academic research and industrial innovation.